POLITECNICO DI TORINO

Corso di Laurea Magistrale in Ingegneria Gestionale

Tesi di Laurea Magistrale

Data Analytics exploitation for Growth Synergies in Multi-Business Companies

Relatore:
Prof. Tania Cerquitelli

Candidato:
Alberto Alfonzo

Tutor Aziendale:
Dott. Vincenzo Scinicariello

Ottobre 2019

Summary

Introduction

This document is a master's degree thesis that crowns my studies in Management Engineering at Politecnico di Torino. Written entirely by the undersigned, it draws inspiration from the advice of Professor Tania Cerquitelli, thesis supervisor and professor of "Business Intelligence for Big Data" course, and above all from the collaboration as a six months internship with the Company "Mediamente Consulting s.r.l.", operating in the field of Data Analysis and Business Intelligence. The Company has made all this work possible as it has provided me with the data and the right tools and motivation to deal with it to the best of my ability, as well as having contributed to teaching notions and techniques just like a degree course. In this work I have tried to encompass the greatest possible number of skills and competences learned in these years of University, starting from the engineering approach to work and from the areas of research touched: business management, economy, and information technology.

The thesis is aimed at multi-business companies and how they can exploit historical data through advanced analytics techniques to find and solve problems and increase the performance by exploiting growth synergies between the lines of business they own.

The idea of this research arises in conjunction with the Business Intelligence course, starting from the internship experience: a client has been offered an analysis of its Data Warehouse and, actively participating in it, I found a real problem regarding the failure to exploit the synergies of the businesses involved in the analysis. Therefore, I decided to undertake research on the state of the art of diversification and business synergies, and to combine the theory learned with data analysis, in order to create a practical model to suggest to managers how exploiting synergies through Data Mining and Machine Learning algorithms.

This document is divided into four chapters:

The first chapter concerns the world of companies that have more than one business (multi-business), the way in which they diversify and economic and organizational-strategic values and risks to which they are subject.

The second chapter deals in a theoretical way with the synergies between the businesses, that is, it has the objective of outlining their characteristics and classifying them. Then, only the class of growth synergies is deepened, because it is the protagonist of the business case of chapter four.

The third chapter attempts to give a concise and effective overview of the main data analysis techniques, underlining how important it is for companies that decision makers make strategies based on data, that is better if historicized in functional structures for this purpose.

The last chapter, chapter four, deals with the real business case: a company with two businesses (plant nursery and supermarket) reveals problems concerning customer loyalty and failure to exploit synergies between the two businesses. It is shown, therefore, a strategic way to synergistically increase sales (growth synergies) through data analysis techniques.

I want to emphasize that there is no resolution of a business case, and therefore no demonstration of ex-post results. Instead, an attempt is made to define a practical model from which the decision makers of the companies, with a much broader vision of the company and the environment, can find ideas and opportunities for their own strategic lines.

Chapter 1

MULTI-BUSINESS CONTEXT: STRATEGY AND ORGANIZATION

This first chapter deals with multi-business companies and the concept of diversification, exploring both strategic and organizational implications. In particular, the first sub-chapter focuses on multi-business and diversification strategies, describing ways to diversify, classifying diversification into different types and presenting advantages and disadvantages generally encountered when companies undertake these strategies. The second sub-chapter, on the other hand, focuses purely on the managerial and organizational implications of the companies that have more than one line of business.

1.1 Multi-business companies and diversification strategies

Today many companies cover sectors that are diametrically opposed to each other, come out and continually enter from different businesses based on the profitability obtained or expected. Understanding “In which business do it works?” is the basis to outline a firm’s identity, but delineate boundaries of bigger multi-business companies is usually difficult, in fact also the companies define themselves broadly and tend to keep generic slogans.

A multi-business company, also known as diversified company, is a type of firm that includes a group of businesses managed jointly. “Diversification comes from the word *diverse* which means *different* or *varied* therefore for an organization, it refers to variation or differences in the performed activities of a firm”[1] .

The challenge these companies face is to produce different products or provide different services maintaining high competitiveness on all fronts. In other words, wide

variety of business investments are mixed within a portfolio, which will be managed at corporate level. In this context, a business is usually an organizational identity, self-contained but not free-standing, with its own profit and loss statement, its planning, focus on specific products or customers, and with its own operational-level strategy. From an information-oriented point of view, multi-business firms are distributed knowledge systems in which business units have their own network and their own information transfer processes. The knowledge is then channelled to the corporate level and exploited for high-level strategic lines.

So, the idea behind diversification is to get bigger by not expanding actual businesses, but by entering in new ones, more attractive for a strategic point of view or which seems more profitable in a specific market period. For profit reasons, a purpose of an enterprise may change over time. Between 1950s and 1980s diversification was been used a lot by companies for corporate growth in the most industrialized nations, since they had the economic possibility of growing at a historical moment when economies of scale were beginning to become fundamental, with a boom in 1970s, for the companies need to reorganize themselves in the form of conglomerates. The thrust came from America and UK by *ITT*, *Textron*, *Hanson*, *Slater-Walker*, …. From 1980s onwards the diversification trend was reversed, because companies began to disinvest on unprofitable non-core business, focusing on core activities. In these years, moreover, there were many acquisitions involving companies operating in the same sector. This specialization trend was the result of emphasis on shareholder value, more attention to transaction costs, and simply for management thinking trends. During the time, most companies have totally transformed their business, for example *Nokia*, supplier of paper, became one of the best telephone-maker in the world in 1990s (see Par 1.1.2 - Unrelated Diversification), and other have only diversified in the same fields of their core-business, like *Microsoft*, expanded in videogames, networking software, information services, … (see Par 1.1.2 - Related Diversification).

It is important not to confuse, as many do, product differentiation with product diversification: the first refer to producing a new product or brand on the same business (ex. *iPhone X, X Plus,…- Smartphones* for *Apple*), the second refer to entering on new business (ex. *Apple Watch – Smartwatches*) related or not related to already existing businesses.

1.1.1 How to diversify: entry, mergers and acquisitions

To explain how companies can diversify, the distinction between internal and external diversification concepts should be introduced. So, *internal diversification* occurs

when a company develop a new line of business internally, while *external diversification* occurs when a company enters in a new business by purchasing another company or single business units. The first concept leads to Greenfield entry method, the second leads to Merger & Acquisition methods. Other strategies, like Joint Venture, Partnerships and Alliances are hybrids between the two categories just described. These forms of diversification are described below.

- **Greenfields entry**

The firm expands upon entering in new markets with new products or services. Since they are "new", therefore new challenges in which the company has no experience, these markets are called *Greenfields:* the company must manage a new business, usually a new customer base, sometimes new geographical contexts, new strategic implications and new managerial needs. So, in these terms, “green fields” seem to be an excellent metaphor. Generally, exploring new market by themselves takes place by exploiting already consistent facilities in the company, as the use of already existing distribution channels or the introduction of cross-marketing actions, above all to bring the first customers. This type of diversification method requires significant costs, starting from those in research and development, heavy advertising, organization restructuring, … For these reasons, companies rarely launch into new markets by this way, seeing it as too risky. However, are an exception all those companies that always push to this type of growth, as they are known exactly for jumping well into new markets, thus having the unconditional trust of customers. Anyway, the investment and the probability of failure are much greater than other diversification methods, but this risk is dumped if diversification is done in areas of key competency and capabilities, and when the company has such financial stability that covers the new business with older ones (known as *cash cow*). So, this method is used specially by larger companies that have the capacity of large investments and strong R&D exploration.

According to Dyer, Godfrey, Jensen & Bryce [2] , when considering Greenfields entry diversification, attention must be paid to:

- Brands: if it is already strong in old businesses, less advertising is required;
- Technology compatibility: the new processes and systems must fit well with older ones;
- Customers: better if similar to those already owned;
- Entry rate: neither too fast nor too slow;
- Required resources and capability: they must exist or must be easily learned;
- Distribution channels: must be easily accessible by the organization.

- **Mergers & Acquisitions**

The company enters a new market by purchasing or joining another company or business unit that operates in this business; therefore, the company diversifies externally, and it looks outside its current operation by hunting other companies' know-how and resources. So, the access to new products or markets is obtained by union or bought. Particularly, we distinguish two types of external diversification: merger and acquisition.

A *Merger* is a corporate strategy usually done between two or more companies whereby the acquiring firm and the acquired firm stands on a merger agreement. Moreover, "In a merger, the boards of directors for two companies approve the combination and seek shareholders' approval. After the merger, the acquired company ceases to exist and becomes part of the acquiring company" [3] . So, mergers occur when two or more firms combine activities to compose one corporation, a common organization, sometimes with a new name, sometimes under the name of one (usually the biggest). In an ideal context, a new ownership and management structure is created. The principal goal of merger is to achieve management synergy (see Par. 2.2.3) by creating a unique united stronger management team.

Acquisition, a second form of external growth, occurs when the purchased corporation loses its independency, being absorbed by the acquiring company and its assets remain intact as an independent subsidiary or absorbed into a business unit. "In a simple acquisition, the acquiring company obtains the majority stake in the acquired firm, which does not change its name or legal structure" [4]. So, a new company does not emerge, but one disappears (usually, a larger firm purchase a smaller company).

In common thought, merger and acquisition distinguishing for its friendly (merger) and unfriendly (acquisition) operation mode, although the latter may sometimes not be, that is when the purchased firm is receptive to acquisition.

Merger and acquisition, but above all the second, are very choice solutions when a company desire to make quick entry to the market and the market favours large corporation.

Again, according to Dyer, Godfrey, Jensen & Bryce [2], attention must be paid to:

- Brand: requires strong advertisement and promotion if it is not well known yet;
- Distribution channels: can be used the old channel if the new is limited or expensive;
- Customers: can be different from the old ones;
- Technology compatibility: difficult to integrate with old processes and systems;

- Entry rate: fast is usually better;
- Scope economies: a learning effect of the new business is possible, so economies of scopes can be exploited.

So, by merger and acquisitions, can be relatively easily obtained greater market power, innovative capabilities thus reducing greenfield method risks, economies of scale and scope maximization. Reorganize in an excellent way the company remains a great challenge.

However, in merger and acquisition, the difficulty lies in the fact that it is not enough just to study the profitability of the new business, but it is necessary to study intensively the company to purchase. In detail, the due diligence checklist requires: an understanding of the business, a study of target company organization and employees, data about stocks and its ownership, and an examination of the company in terms of all products that offers and all industries in which operates.

- **Joint Venture, Partnerships, Alliances and other forms**

These are also milder forms of diversification, which however allows to invest to other businesses, risking less and exploiting the know-how of other companies by collaborating, and not by acquiring them. Some of these are Joint Venture, Partnerships and Alliances. Without going into too much detail, we can say that:

- A *Joint Venture* is a business arrangement in which two or more companies agree to join their resources to perform a specific project or any other business activity. In other words, Joint Venture combines companies of any size to take on one or more projects or deals, with or without continuing purpose. The venture is a separate entity, but the participants is responsible for costs and profits/losses.
- A *Partnership* is an arrangement by two or more participants to manage and operate a business and divide its profits/losses. The parts of liability and participation in profits can change between the parties and depend on the registered agreement.
- A *Strategic Alliance* is a long-term agreement in which two or more companies decide to share resources to specifics projects, to work toward common or correlating goals. It's less blinding than a joint venture and Partnership because companies maintain their autonomy in profits/losses.

Then there are many other ways of collaborating with other companies entering new businesses, like *Consortium*, *Association*, … and they also depend heavily on national laws and treaties between nations. But listing them and describing them one by one is not a thesis's goal. It is only important to know that there are many alternative methods of diversifying, more or less risky.

However, it could be interesting to mention *Licensing* as a border line type of diversification: a firm can exploit proprietary technology and resources by licensing it to other firms. This represent an alternative to direct investment.

1.1.2 Types of diversification

Once we have described how to diversify, we now go on describing what types of diversification exist, that is, classifying diversification concepts according to the target business and those already owned.

- **Vertical and horizontal diversification**

Vertical and horizontal diversification is a distinction that considers which production stage the new business belongs to. Production stages are the steps that a product goes through in being transformed during its entire production cycle (from raw material to finished product).

A company do *vertical diversification* when integrates vertically: so, a firm expanded beginning to operate in other production stages of an existing business supply chain. When a firm diversifies towards raw materials, it is doing *backward* vertical integration, while diversifies towards distribution/retail/service (or more simply towards the customer), means doing *forward* vertical integration. The first strategy is advisable when organization's suppliers are expensive and with high profit margins, not appropriate or with high bargaining power; moreover, is advisable when the company needs to quickly acquire resources or obviously has capital and human resources to manage the new and old business together. The second strategy is advisable when organization's clients (which are other companies) are expensive, not appropriate or with high bargaining power, or when the expansion is needed for stable production and demand prediction. Some firms employ vertical integration strategies to acquire the "margins of the middleman".

Horizontal diversification is the expansion of the company's activity to products, processes and know-how related or un-related to the same already technological-productive stages of a chain owned. This strategy is advisable when customers are loyal to the company, the expected gains are high, and the organization need growth by exploiting economies of scale, therefore with a positive effect on the old businesses.

- **Related and unrelated diversification**

The purpose of diversification, it has been said, is to allow the company to enter lines of business that are different from current operations. But the question is: how much different? When the new field is strategically related to the existing lines of business,

it is called *related* or *concentric diversification* (for example a pc maker enters smartphones business). Contrariwise, an *unrelated* or *conglomerate diversification* occurs when there is no fit or relationship between new and old businesses. In details, we are in front of *related diversification* when the value chain of the new business possesses valuable cross-business strategic fits with at least one already existing business. So, this involved businesses that have comparable value chains.

First, the concept of relatedness needs to be clarified: that is when are present sharing common resources and capabilities (resource and management commonalities), not the similarity of products or process technologies (operational commonalities). Therefore, the correlated diversification allows the company to obtain greater efficiency in terms of allocated resources and cost/revenue benefits. Common knowledge advises concentric diversification when:

- The industry does not grow or grow very slowly (the continuous search of more profitability associated to related diversification makes this strategy prevalent in developed economies);
- The product obtained by correlation has a good chance of increasing the sales of the old ones, moreover it covers the seasonal instabilities of the other products and can be offered to the market at competitive price;
- The management team has the right capabilities to manage effectively the new business (transferring expertise, technology, marketing, …) or needs to form more others to be competitive in the future;
- A cross-business resource sharing is possible.

Unrelated diversification, on the contrary, occurs when the difference between the activities and the value chains of new and old businesses are so dissimilar that no competitively cross-business relationship is present. Companies take this way to find new opportunities and attractive investments in other types of business, if areas in which they operate are limited. The major disadvantage of this diversification strategy is the advent of managerial problems because managers of completely different businesses may be unable to work together efficiently and indeed, they could even fight for the allocation of essential resources creating problems of rivalry within the same company. Common knowledge advises the conglomerate diversification strategy when:

- The company want to grow and its existing businesses are in too competitive areas, in no-growth industries, or is experiencing declining annual profits;
- The product obtained by entering in new business has a good chance of increasing the sales of the old ones, moreover it covers the seasonal instabilities of the other products and can be offered to the market at competitive price;

- The organization can compete successfully in the new business because has the appropriate capital and talented managers;
- The company want to spread market risk.

Related and unrelated diversification concepts introducing and are closely linked to the concept of cross-business synergy, primary objective of the thesis and topic to which the entire next chapter is dedicated. Synergy is the ability of an organization to achieve greater results from more business working together than the sum of the result of the business taken independently. Obviously related diversification conducts to synergy, or, better, exploiting the synergy between businesses is the main reason for the adoption of this strategy: resource sharing, relatively easy administration of the new business (because is similar to existing one), the transposition of consolidated customers to new products, and the possibility of using the technical know-how to gain some advantages. Exploiting synergies between businesses is more difficult for an unrelated diversification strategy and certainly not a primary reason. However, even in this case it is possible to detect cross-business synergies in finance or marketing, but not in combining operating or management efforts (see Chapter 2).

This possibility of exploiting scope economies suggest that concentric diversification is an even better strategy than the opposite. However, although R. Rumelt discovered that companies that diversify into related business should be more profitable than those that diversify into unrelated business, some more recent studies confirm the opposite: the impact of spreading risks is so positive that allows companies to make more profit than others. In conclusion, the situation is not so clear, as it is not always clear the distinction between related and unrelated business. What is unequivocal is that related diversification is more successful and popular in developed and highly competitive economies, and unrelated diversification, instead, in developing economies.

RELATED DIVERSIFICATION	PEPSICO	JONSON & JOHNSON	GILLETTE	PROCTER & GAMBLE
	Soft drinks	Baby products	Blades and Razors	Hair care products
	Fruit Juices	Medical devices	Tooth Brush	Household cleaning
	Snack foods	Contact lenses	Hair dryers	Beauty care products
	Sports drinks	Personal care products	Shavers	Laundry products
UNRELATED DIVERSIFICATION	WIPRO	LG	RELIANCE	TATA GROUPS
	Electrical appliances	Mobile Phones	Chemical products	Home appliances
	Computer accessories	Home appliances	Textiles	Watches
	Baby care products	Lamps	Construction	Tea products
	Toilet soap	Television	Mobile Phones	Info. technology

Figure 1.1. Some examples of related and unrelated diversification of known companies.
Source: Author, based on data from: [5]

1.1.3 Value and risks of diversification

Diversification is a largely used strategy for other purposes besides growth, for example to obtain competitive advantage, to compensate obsolescence, to distribute risks, to improve other existing business through synergy, to utilize excess capacity or many other reasons. Diversification strategies are adopted as they potentially bring much value to the company. However, they often result in defeats when not well implemented and managed. Below are the reasons why diversification brings value, then the risks and negativities that may arise.

- **Values of diversification**

According to Calori and Harvatopoulos [6], diversification rational reasons are two, relates to the nature of the objective: defensive or offensive. "Defensive" refers to spreading the risk of market contraction or to growing when current markets seems to provide no further opportunities. "Offensive" relates to conquering new positions that seems to have good profitability, usually reinvesting extra-profits. In another classification, there are five reasons why managers may consider diversification a valid strategy: a) growth, profits and market power; b) risk reduction; c) synergies exploitation. Let's see these in detail.

a) Growth, profits and market power

Diversification is part of growth strategies, that involve increase in performance objectives: just think of the consideration that companies have to sales KPI: more is almost always better, even if profits remain stable. Larger companies, in effect, have some advantages over smaller firms: the possibility of exploiting economies of scale and scope, increase bargaining power towards suppliers and customers, exploiting geographic differences placing multiple plants in location providing the lowest cost, and sharing of information between businesses.

In fact, diversification is one of the four main growth strategies defined by Igor Ansoff's Matrix Product/Market, shown in *Figure 1.2*.

While the other three strategies are implemented with the same old resources and capabilities (technical, financial, …), diversification usually need a company to acquire few or many new skills, resources and markets knowledge. This expose organization to high risk (see below).

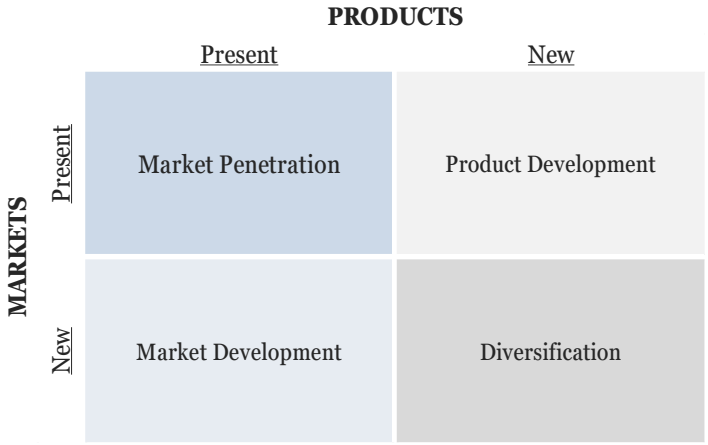


Figure 1.2. Ansoff Matrix. Source: author

Moreover, without diversification, firms are prisoners of their business. When the business owned is close to death, diversification is certainly a useful weapon to bring profits back to the company. In other words, whenever companies see stagnant sales and profits of their important business, it is an indication of the need to diversify. In general, begin to operate in new business is a choice which must be studied carefully. The famous 5 Porter's forces, to which the explanation in this document is omitted, can help to understand if the entry can offer profits or not.

So, the potential for diversification is also to enhance profitability by increasing a firm's market power. For this reason, diversification strategies are well monitored by antitrust authorities. In fact, large diversified companies can suppress competition and exercise market power through many mechanisms, of which the main ones are:

- *Predatory pricing*: multi-business companies can wipe away competitors from one business by cutting prices even sometimes under their respective costs, exploiting the fact of being economically covered by other owned businesses (cross-subsidization). In this way they go to create a price war that other competitors without coverage are unable to face, so they leave the market, and the company can re-increase price. This anticompetitive mechanism is also used when it seems that another company is about to enter a market, in order to make it desist.
- *Bundling*: a multi-business firm can bundle two or more products in order to force the purchase of one, usually subject to market competition, by offering it only together with another, usually in a monopoly market.
- *Reciprocal dealing*: when two diversified company operates in two or more same or related businesses, simultaneous agreements could arise such as "if you buy from me this, I buy from you the other".

- *Mutual forbearance*: game theory shows that in a multimarket competition two or more companies do not tend to compete strongly (maintaining high prices) for fear that others can start parallel price wars on other businesses.

As already said, these systems are all fined by the antitrust authority, but certainly they are not so easily to discover, especially if in tacit form.

b) Risk reduction

Often managers diversify because are guided by the desire to spread risk. This is certainly better if the diversification is unrelated rather than related, so the businesses are independent and don't follow similar trends of market's profits/losses. In other words, with conglomerate diversification the variance of cashflows of the combined businesses is less than the average of the separate businesses, therefore diversification reduces risk. In fact, it seems that diversification is a good strategy against bankruptcy when profits fluctuate a lot. Then, we can ask if this risk reduction brings value to shareholders: the answer is "no". The reason about this will be explained a little further on. However, risks spreading may benefit other stakeholders: in addition to coverage on the performance fluctuation, could also be an advantage the reallocation of employees in the most important businesses in certain periods.

c) Exploiting synergies

One of the main reasons why a diversification strategy should be developed is the presence of synergies. Before delving into the concept in Chapter 3, we can define a general overview. It has already been said that cross-business synergies develop more under related diversification, so we can affirm that unrelated diversification is to be preferred when spreading risk is needed, while related diversification has the main purpose of exploiting synergies between businesses. "Exploiting synergies" means sharing resources and capabilities between two or more business in such a way that, with the same result, the effort to manage more business together is less than the sum of the management efforts of the business taken individually. So, tangible resources (technology systems, sales forces, R&D laboratories, …) and intangible resources (brands, reputation, technology, …) could be shared in more businesses giving a cost, revenue and management advantage. Moreover, some business functions are managed at corporate level, as accounting, legal services, government relations, … without need to be duplicated. Also, capabilities are important in synergies: when a company, and therefore all its employees and managers, already knows how to do something well, it is good to jump into businesses that require the same similar skills and offer good profits. In this way the company does not start from zero, but it has a solid base on which building a winning business.

Moreover, successful diversification goes through the installation of appropriate performance measures, effective incentives and corporate culture alignment.

In history there are many good examples of successful diversification, for sure everyone knows the story of Apple, that moved from PCs to smartphones.

- **Risks of diversification**

Many executives consider diversification risky and sometimes destructive. Below are the main reasons of this type of evaluation: d) Implement new resources and capabilities; e) Loss of focus on the core business; f) No value for shareholders

d) Implement new resources and capabilities

It has been seen that diversification is one of those presented in the Ansoff matrix. Of the four strategies, however, it is certainly considered the riskiest because it is the only one that requires the acquisition of new resources and capacities in a large or small scale. For this reason, in addition to a careful business analysis, it is also important to understand how much the organization is able to compete effectively in a new field and understanding a priori if it is possible to acquire and reform new skills. Because of these high risks, many companies attempting to diversify have led to failure.

e) Loss of focus on the core business

Diversification strategy needs an expansion of human and financial resources, which may detract focus and commitment in core businesses. For this reason, it is advisable to start this operation only after careful market analysis and once it has seen that existing business no longer offers growth opportunities. In fact, According to Dawid and Reimann [7] , diversification compromises the quality of the products offered and is a brake on innovation.

f) No value for shareholder

Does risk reduction create values for shareholders? This consequent is actually not yet an established point, but the answer is probably “no”. The reasoning is as follows: if investors can get diversified portfolios, what advantage would they have if the companies diversified for them? The only way for advantage is that firms can diversify at lower costs than individual investors, but this is impossible because, while diversifying, companies face transaction costs (control premium, legal advisers, banks, linking functions, ...) much larger than investors. The CAPM (Capital Asset Pricing Model) formalizes this concept (for further information, see related studies). Empirical studies confirm this theory, especially for conglomerate diversification. Moreover, although diversified companies can avoid making use of the external capital market by maintaining a balanced portfolio that generates and use cash at zero cost, it seems

instead that this internal capital market tends to cross-subsidize the underperforming divisions and waste resources in internal competition among businesses, including the reluctance to transfer cash-flows to divisions with the best profit prospects.

1.2 Strategy and organization: managing multi-business context

Resource and capabilities developed by the company are key success factors for competing in more than one business. Many of the most important, that is those that most influence the success of the diversification strategy, are general management capabilities. In fact, as documented by Marlin, Lamont and Geiger study [8], performance of a new business after diversification strategy is related to the strength of its top management team members. For example, the success of a merger does not only depend on how well the businesses work together, but also how well-suited are managers to manage that operation. Compared to single business, in multi-businesses companies, objectives, strategic processes, tools, and the management role, should consider allocating resources to reach goals of the whole organization, promoting cooperation between business units to create synergies and value. This often means change functions structure, operations, and management systems, in order to improve long-term performance, capture synergies, and steer corporate resources into attractive projects. Moreover, if diversification includes foreign businesses (multinational diversification), managers must interface and integrate with different cultures. Given this complexity, capable top management has a key role.

In a single business company, corporate strategy and business strategy coincide. In a multi-business company, instead, corporate (top-level) strategy, implemented by top management, is the driver of other business (low-level) strategies (this concept was remarked by Porter in 1987, which defines corporate strategy as “wide strategy” and business strategies as “competitive strategies”), and for this reason is crucial for company success. The corporate of a multi-business should: formulate the objectives of the entire Group, constantly assess the potential of the sectors and the competitive position of the business units, improve the results of the corporate through the efficiency of the management of each business or with further diversification or disinvestment. So, corporate strategy, implemented by macro-management, is less specific, conceptual and value oriented. Sekulic define corporate strategy as “a pattern of decisions by an organization that determine and guide its corporate objectives, generate policies and determine the scope and nature as well as activities of the

organization" [9] . Dyer, Godfrey, Jensen, and Bryce [2], said that corporate strategy aims to increase stakeholder value and competitive advantage in multiple businesses. The corporate strategy therefore lays the foundation and directives for all business, that are semi-autonomous organizational entities which have independent missions and objectives. Business strategy, implemented by micro-management, deals with achieving objectives outlined by the corporate strategy, which is converted into execution and implementation. Dyer, Godfrey, Jensen and Bryce, affirm that the development of competitive advantage in a single business or market is generally referred to business unit strategies. Furthermore, all people, regardless of the organization level in which they work, must be aware of the corporate strategies in order to be coordinated with the whole group, to improve efficiency, information flows, and achieve high level objectives.

In a multi-business context, moreover, it is essential to manage the brands and logos that you want for products (a brand is a name, term, sign, symbol/design or a combination of these elements). In fact, generally, managers try to maintain strong and solid brands in order to have a good corporate image. Let's go into detail with the distinction between corporate brand and product brand: the corporate brand focuses on the entire organization and considers many stakeholders, "is about people values, practices and processes"[10] , and tends to differentiate, as a whole, the products of the company from those of its competitors. The product brand, on the contrary, focuses on the product and the customer, and is subject to short/medium term marketing actions. The decision to use a corporate brand strategy in multi-business companies is a double-edged sword: in fact, may have positive or negative implications, based on how the branding strategy is implemented and how fit with the organization. Corporate branding strategy can add value to the company by creating a distinctive position on the market, creating unique identity and therefore decreasing competition. In fact, customers' perception of all the products is transferred to other products under the same corporate brand. The risk of putting a company under the wing of a single brand is that a big mistake on a small product could bring down the entire image of the company.

Chapter 2

CROSS-BUSINESS SYNERGIES

In this second chapter of the paper, we deal with the synergies between businesses owned by a multi-business company, underlining the characteristics, the advantages, and the difficulties in exploiting them. In detail, in the first sub-chapter we define the concept of synergy in business management, in the second sub-chapter a detailed classification of the different types of synergies is carried out and, in the last sub-chapter, we focus only on one type of synergy: growth synergies. These are abundantly deepened in all their facets, both for their undoubted importance, and because they are the protagonists of the business case of chapter four.

2.1 Characteristics of cross-business synergies

Synergy derived from the Greek word *synergos*, which means working together. So, synergy means "combined action", and refers to the effect that the result of the aggregate is greater than the sum of the individual results. In this document the subjects are the individual businesses in a multi-business company: they exploit synergies if, working together, are more competitively valid than operating without cooperation. Therefore, cross-business synergies represent the value added (improving quality, growth, financial benefit, …) by the corporate system. In formula:

$$TSP = value\left(\sum_{i=1}^{n} BU_i\right) - \sum_{i=1}^{n} value(BU_i)$$

where:

TSP = Total Synergy Potential of a firm (why "Potential" is explained in Par.2.1.1)

BU_i = i-th Business Unit

n = Total number of businesses owned

The concept of *value* is very similar to the NPV (Net Present Value) of a company, that is, it reflects the benefit deriving from the future cash flows due to its object (and it includes both cost reduction and increase in revenue of business involved).

Ansoff (1965) was the first to talk about synergies and how important they are in company strategies. And for some time, managers of many important multi-business companies have been talking about synergies and the efforts they are making to implement them. Indeed, Larsson & Finkelstein [11] stated that the establishment of effective synergies is the first reason for mergers and acquisition. Researches show that mergers and acquisitions, or in general diversification strategies, generate shareholder's benefit (increase in stock prices) if the new entity has significant advantage in distributions, bargaining power, demand, etc…, all things that are due to cross-business synergies. Stressing its importance is also the opposite: most diversification failures owe their cause to the lack or incorrect exploitation of synergies.

But synergy is not simply a fit between businesses, but also a relation between skills, management effort, resources, activities, and more others. More specifically, the synergistic effects are the sum of two distinct effects, even if sometimes not easily distinguishable: the first effect is the vertical one and is due to the presence of corporate management (result of the relationship between the top management and the business units), the second effect is instead horizontal, and considers the direct influence that different businesses have between them. This last effect is precisely that which arises spontaneously from related diversification, focusing on sharing resources and capabilities, while vertical effect is closely linked to the corporate level of the company and therefore to the management skills, and this is independent of the initial correlation between businesses.

The absence of synergies, or their poor management (lack of corporate vision and objectives, lack of good information network, dysfunctional leadership, etc…), is a heavy loss for companies and for this reason over time more and more large groups have been created, in order to reach competitive advantage over focused companies.

2.1.1 Synergies cost of realization and corporate advantage

In the formula of the previous paragraph we talked about TSP (Total Synergy Potential). The concept of "potential" is soon explained: since there are synergy

realization costs, the obtained benefit is the net effect between the synergic potential and the cost to erect it (*Figure 2.1*).

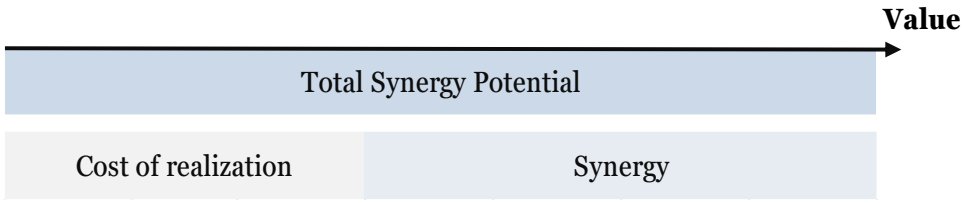


Figure 2.1. Relation between TSP, Cost of realization and Synergy. Source: author

Among costs we find:

- *Coordination costs* caused by linkages between businesses. Managers who work in roles to exploit synergies, in fact, spend about 70% of their time to coordinate with managers of other business unit or to organize the interaction.
- *Controlling costs* due to the existence of business executives who have internal decision-making power: the corporate level managers need to interact and control them in order to maintain the strongest decision-making power, and this need a lot of time spent.
- *Costs of compromise* arises when, to favour the performance of a business, the effectiveness of another is reduced.
- *Costs of inflexibility* arise when the interdependencies between the businesses or more generally the corporate strategies leave less decision-making power space to the businesses themselves, which no longer move to achieve the individual best.

These costs are very important to manage as they can also be far superior to the benefit of the synergies obtained. In fact, if you normally look for a super-additive business results (2+2=5), sometimes the reverse could happen (2+2=3).

Companies goal is to obtain a long-term sustainable competitive advantage on owned businesses, and Porter (1985) described the synergies between businesses as a powerful means of achieving it. Some classic theories of strategy, above all resource-based view theory and transaction cost theory, suggest how competitive advantage can be achieved through cross-business synergy exploitation: shared resources, that are assets or human resources, tangible or intangible, need to be valuable (they increase opportunities and reduce threats), rare (they are few in the market) and difficult to imitate (competitors struggle to get similar ones). Moreover, competitive advantage is favoured by obtaining a corporate advantage, that is the advantage of having more than one business. In fact, internal (between businesses) transaction costs can avoid hold-

up risks or higher transaction costs of the external market, and coordination advantage can be obtained when two or more business are so connected that they require constant mutual adjustment (with consequents re-contracting costs) which would be very complex to manage if the reference business is owned by another company.

2.2 Classifying cross-business synergies

Although Ansoff introduced a synergy types classification based on ROI components (sales, investment, operation, management), the author thinks a more recent classification, made by Sebastian Knoll [5] , a German researcher that has combined many historical fonts on synergies in an in-depth analysis of the phenomenon, is more appropriate for understanding the difference between classes. The classification is based on four type of resources (operative resources, market power resources, financial resources and corporate management resources) which give rise to different types of synergies: respectively operative synergies, market power synergies, financial synergies and corporate management synergies. The four types are described below.

2.2.1 Operative synergies

Operative Synergies are related to operative tangible and intangible resources, that refers to all those concerning production, facilities and operative assets. Operative synergies are divided into two sub-classes: efficiency synergies (based on cost advantage) and growth synergies (based on revenue advantage).

- **Efficiency synergy**

Efficiency synergies are present when multi-business companies achieve efficiency advantages by sharing operative resources across businesses, keeping high their utilization level. This is practically the classic concept of scope economies: scope economies occur when the Total Production Cost (TPC) of producing the output of the businesses together is more convenient than producing each output with separate resources (in or out company). In formula:

$$TPC(BU_1, BU_2, \dots, BU_n) < TPC(BU_1) + TPC(BU_2) + \cdots + TPC(BU_n)$$

But it is important to emphasize that is wrong to speak only of productive factors, because scope economies include efficiency even in other areas such as R&D or reputation. Efficiency synergies are favoured by some specific conditions and strategies, such as products similarity, excess capacity on a resource, reusing inputs

for more products, or sharing of intangible assets (brands, best practices, …) between more businesses.

- **Growth synergies**

Growth synergies are present when a multi-business company combine operative resources across businesses in order to achieve revenues super-additivities: Total Revenues (TR) obtained from utilizing the same resources in more businesses are greater than revenues obtained from utilizing different resources for different businesses (in or out company). In formula:

$$TR(BU_1, BU_2, \dots, BU_n) < TR(BU_1) + TR(BU_2) + \cdots + TR(BU_n)$$

Growth synergies emerge when managers recombine complementary resources (that is, they mutually reinforce each other) in more businesses for new market opportunities or to enhance value perceived by customers. More specifically, two resources R_1 and R_2 are complementary when, stopped the level of R_2, an increment in R_1 give an output value greater than the output value that would be obtained if R_1 was in isolation. In formula:

$$V(R_1, R_2) - V(0, R_2) > V(R_1, 0) - V(0, 0)$$

The differences between efficiency and growth synergies are shown in *Table 2.2*.

	EFFICIENCY SYNERGIES	GROWTH SYNERGIES
First Effect	Increased efficiency	Increased revenues
Strategic Implication	Cost subadditivities	Revenue Superadditivities
Value Driver	Operations efficiency	Customer Utility
Focus	Similarities in value chain functions	Customer and markets
Time of exploitation	Long period	Temporary

Table 2.2. Difference between Efficiency Synergies and Growth Synergies on some points. Source: author, based on document [5]

When efficiency and growth synergies are based on resources that are valuable, rare, and with a combination difficult to imitate, they are source of sustainable competitive advantage. Moreover, competitive advantage could be achieved even just sharing resources between businesses, thanks to benefits due to the lack of market transaction costs.

2.2.2 Market power synergies

Market power synergies refer to advantages that derived from reduction of market competition and increasing in prices. So, market power synergies are obtained by spreading market power across businesses, for example by undertaking anticompetitive behaviour, such as predatory pricing, bundling, reciprocal dealing, mutual forbearance and cross-subsidization, already widely explained in previous chapter (Par. 1.1.3).

2.2.3 Financial synergies

Financial resources are all those resources connected to the financial and risk management of the company. Financial synergies are present when a multi-business organization leveraging financial resources across more businesses. This type of synergies offers numerous advantages, and the most important ones are listed below.

First, a multi-business firm can reduce corporate risk because the variance of the trend in costs and revenues is less when, compared to a single account, more businesses contribute to the corporate performance. So, imperfect correlated cash flows can certainly reassure managers about sudden downs in single businesses. The main purpose of the search for financial synergies is just to mitigate the risk, and managers who want to achieve these synergies tend to prefer unrelated diversification, which helps in this regard.

Second, an internal capital market is established, that reduce financial costs (included cost of capital that every business should support by looking for capital on the external market), increase financial flexibility and capital allocation effectiveness.

Third, tax advantages can be obtained by exploiting corporate accounting: a system of compensation between profits and losses of businesses that minimizes the total fiscal commitment could be used (law permitting). Moreover, as multi-business companies have lower risk levels, they have higher debt capacity. Interest on this debt would serve as a tax shield on gross profit.

Lastly, multi-business firms may exploit financial synergies by increasing financial economies of scale: corporate financial management significantly reduces transaction costs in debt and equity management.

Financial synergies are not seen as potential bidders of competitive advantage, in fact financial resources are neither rare nor difficult to imitate, and investors are not willing to reward multi-business companies more, as they can diversify their portfolio by themselves.

2.2.4 Corporate management synergies

Corporate Management resources refer to managerial capabilities of top executives (at the corporate level). Corporate management synergies arise from the presence of an organizational system managed by a higher level, the corporate level. Particularly, corporate management synergies are advantages of multi-business companies from leveraging management capabilities across more businesses, so they capture the value added by the activities of corporate managers (high-level strategic actions).

Corporate management synergies contribute to competitive advantage if managers' capabilities are valuable, rare and difficult to imitate. So, companies are able to achieve high levels of performance thanks to best practices, successful organizational structures, effective strategies, and innovative processes. Moreover, complex interpersonal relationships and organizational routines help to make corporate manager teams difficult to imitate.

2.2.5 Classification summary

In *Figure 2.3* the classification of synergies is summarized.

OPERATIVE SYNERGIES
Related to leveraging operative tangible and intangible resources (all those concerning production and facilities) ◆ Efficiency Synergies: cost subadditivities by sharing (similar) operative resources across businesses ◆ Growth Synergies: revenue superadditivities by sharing (complementary) operative resources across businesses
MARKET POWER SYNERGIES
Related to spreading market power across businesses ◆ Predatory pricing: cutting prices to wipe away competitors from a business, thanks to cross-subsidization of another one ◆ Bundling: bundle two or more products in order to force the purchase of all, extending monopoly power of one business ◆ Reciprocal-dealing: simoultaneous agreement (also tacit) between two companies that operates in same businesses ◆ Mutual forbearance: in multimarket competition two or more companies tend to collude
FINANCIAL SYNERGIES
Related to leveraging financial resources across businesses ◆ Corporate risk reduction: the variance of trend in costs and revenues is less when there are more businesses ◆ Internal capital market: reduces cost of capital, increase flexibility and capital allocation effectiveness ◆ Tax advatages: a system of compensation between profits and losses of the businesses can decrease fiscal impact ◆ Financial economies of scale: corporate financial management reduces transaction costs in debt and equity management
CORPORATE MANAGEMENT SYNERGIES
Related to leveraging managerial capabilities across businesses ◆ Strategic and organizational corporate-level action improve more individual business performance

Figure 2.3. Main characteristics of the types of synergies.
Source: author, based on document [5]

2.3 Deepening: realizing growth synergies

The author intends to elaborate growth synergies more in depth, in particular he wants to explicate strategies to increase performance through growth synergies. This paragraph lays the foundations for the case study in chapter four, which requires the use of some strategies for the exploitation of synergies to increase revenues of a company.

It has been said, growth synergies are present when a multi-business company combine operative resources across businesses in order to achieve revenues super-additivities. Starting from the Ansoff matrix, that allow to determine four growth strategies to increase the business through existing or newly developed products in existing or new markets, we can define the respective four types of strategies for achieving growth synergies: joint market penetration, joint product development, joint market development and joint diversification (*Figure 2.4*). The idea of these concepts is to exploit strategies shown by Ansoff and apply them for the exploration of synergies of multiple businesses, therefore transporting the same strategies in a multi-business context in which businesses interact with each other, so as to increase performance (growth in revenues) both at individual business unit level and at corporate level.

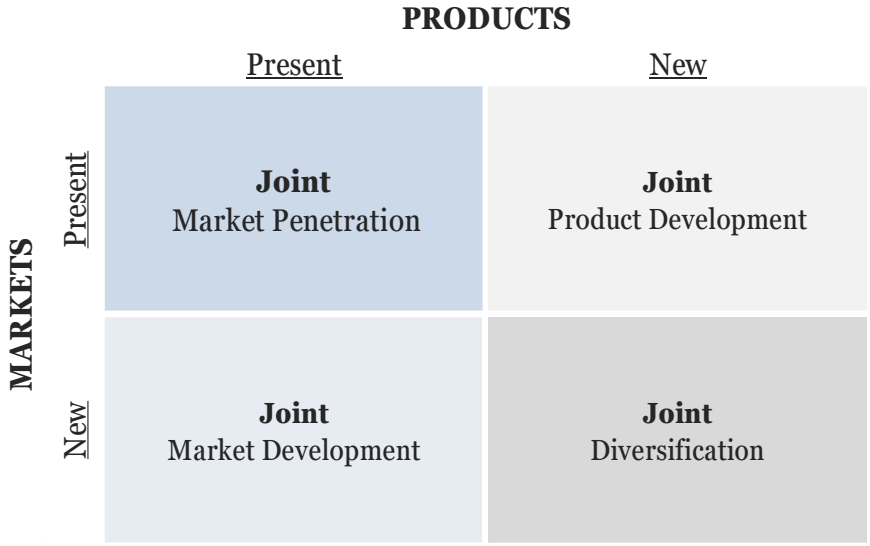


Figure 2.4. Strategies for achieving cross-business growth synergies.
Source: author

2.3.1 Joint market penetration

Joint market penetration includes all those strategies and techniques that aim to increase synergistically market share in existing markets with existing offerings. In other words, businesses interact together to achieve a better performance about their individual market shares. The main techniques implemented for this purpose are described below.

- **Cross-selling**

Cross selling strategies allow to leverage customers across businesses through collaboration in terms of product offerings. In other words, this technique tries to exploit the knowledge about already acquired customers to increase sales and their loyalty across businesses involved. In details, cross-selling is a sales strategy consisting in offering, to customers who has already purchased a particular product or service, the purchase of other (preferably complementary) products or services, often with a discount. An example of cross-selling is a company that, to the customer who buys its smartphone, offers a 10% discount on the headsets, or, in a multi-business context, a firm that offers discounts in his bar department, if they are bought for a certain amount household products (it is the same case of IKEA). In this way customers strongly loyal to the first business can be brought to the second, hoping that they will appreciate and then continue to purchase in the future. The added value given to customers, in addition to discounts, is the "one-stop shop", that is one or more places (or platforms) in which they can buy more products, even very different, with just one effort for choice. For this reason, cross-selling techniques allows the company to develop corporate advantage over competitors, so are very useful for obtaining competitive advantages in individual businesses due to the fact of being a multi-business firm.

- **Cross-business bundling**

Bundling strategy gives to customers the opportunity to buy a package (bundle) of multiple products at a unique combined price, usually lower than the sum of the individual prices (sometimes single products cannot be purchased individually). While in cross-selling products are suggested to the customer, in bundling the customer is obliged to buy the whole package, even if he only needs one of the products included. Bundling is exploitable for products of the same business, but it is very useful especially when the products in the bundle belong to different businesses (cross-business bundling), because in this way customers are forced to buy from different businesses, even if they are loyal to only one of these. An example of bundling is a multi-business company that offers the purchase of the call, internet and music streaming at a unique price. Amazon uses a lot bundling to incentive customers to

subscribe to Amazon Prime by offering them many other services such as Amazon Prime Video, Amazon Prime Music, Twitch, etc.

As already mentioned in the previous chapters, bundling is under control of antitrust authority, since, if abused, it leads to anti-competitive behaviour that therefore affects the entire society welfare. This happens especially when companies bundle essential products or super-used standard products together with other low importance products, forcing customers to buy the last ones against will. A case of cross-business bundling subject to provisions by antitrust authorities is that of Microsoft, which forced PC-makers to pre-install the office package if they wanted to install the Windows operating system.

Cross-business bundling generates higher revenues at lower selling costs, and allows customers, such as cross-selling, to reduce their transaction costs by buying a single package instead of the individual elements contained in it. However, it is a double-edged sword because, if it is true that it allows a very sold product to drag other less-sold ones, the opposite can also happen: the sale of all the products in bundle falls because the combination is not in the interest of consumers.

- **Umbrella brand**

Umbrella is a marketing strategy involving the use of the same (or very similar) brand name for two or more products. This technique becomes a powerful marketing weapon when the company has a very strong brand: new products are advertised under its brand in order to be immediately recognized as "relatives" of the strong product and therefore perceived as strong in the same way. Successful example is Apple Company who built his famous *Apple* brand over the years and applies it to a whole series of products, such as *Apple Watch* or *iPhone,* while the original brand referred to *Mac* computers.

This common brand across businesses can be a significant source of value, both for the company, which increases its sales, and for customers, who feel reassured by the name of the brand, even if the product is recently on the market. Brands are so important that they are considered real valuable assets: that of Apple, for example, is estimated to be worth as much as 214 billion dollars (2018, Interbrand, Best Global Brand Report).

- **Joint amrketing initiatives**

Joint marketing initiatives are cross-business marketing campaigns aimed at increasing customer loyalty and strengthening the company's reputation. In particular, we are talking about joint image campaign when there is a coordinated

business effort to create a strong corporate image, while we are talking about joint customer loyalty program when the objective is to retain customers as much as possible on the company's various businesses. The company can exploit the existence of a corporate organization by creation of “total company fidelity program”, and, moreover, obtain the advantage to sharing marketing costs across businesses.

2.3.2 Joint product development

Joint product development refers to the synergistic development (offer) of new products across existing businesses. The main techniques implemented for this purpose are described below.

- **Integrated cross-business solutions**

Integrated cross-business solutions are products created by the collaboration of more businesses as a solution to common customer problems. An example is an automotive and plants company that provides integrated multi-function solutions. This technique leads to a more complete market coverage, increasing sales, and allows the organization to create services with long-term contracts, because integrated solutions often require skill that only manufacturer can provide. Integrated solutions are very efficient in dynamic markets that require complex products.

- **Innovative cross-business fusions**

Innovative cross-business fusions have the function of creating a completely new product based on skills, resources or know-how of existing businesses. For example, Ferrero makes *Nutella & Go*, a single product that contains *Nutella* and *Estathè,* with the addition of breadsticks. Companies adopt this technique to create growth synergies by increasing revenues, creating a unique customer value that single-business companies are not able to create. Moreover, companies can bring new products to the market much faster than their competitors thanks to capabilities already present in the existing businesses involved.

- **Joint development platforms**

Platforms are technical solutions which are used to produce intermediary components or products. So, joint development platforms refer to the creation of intermediary solutions useful for more businesses. For example, Radio Frequency Identification (RFID) technology could be used for several finished products, or same car engine could be used for different end models. However, this strategy seems to be part of efficiency rather than growth. However, it helps a lot in creating different variations

for products, both in innovation, and in fast time to market of new products, favouring sales growth.

2.3.3 Joint market development

Joint market development refers to the synergistic development of new businesses with existing products. In other words, the company seeks to enter a new business through the collaboration of existing businesses that offer their products. Two techniques implemented for this purpose are described below.

- **Joint development of new geographical markets**

Multi-business firms grow by developing new geographical markets with existing offerings, that is a territorial expansion by synergistic contribute of more businesses with existing products. It is a very common strategy of growth, used when managers believe that the target territory can appreciate products already offered in other countries. In this way the company only needs to know the culture and possibilities of new countries, because its offer, and therefore the required capabilities, are always the same.

- **Joint development of new market segments with common customer**

The expansion into new markets with existing offering can also be done creating a new business that satisfies potential common customers to two or more company businesses, that is creating a cross-business offer composed by existing products for new not already targeted segments. The main objective of this strategy is to grow by increasing market coverage.

2.3.4 Joint diversification

Joint diversification is not to be confused with the simple concept of diversification. In fact, the first refers to the synergistic expansion of many businesses together to create an innovative new one that offers new products, the second refers to the simple entry in a new business, regardless of the others already existing. Joint diversification must, by definition, give rise to a business connected to the existing ones, therefore it belongs to related diversification strategies. A multi-business company expanded into new business with new products by leveraging capabilities and know-how of existing businesses. So, growth synergies are created by increasing sales of the participating businesses, that collaborate for a common goal, the search for blue ocean markets, that promise unexplored markets full of opportunities.

2.3.5 Other tips for growth synergies exploitation

Certainly, possible strategies for the exploitation of cross-business synergies are not reduced to the sub-points of the classification just presented, but they are many and depend a lot on the business case to which we are facing. So, some tips, which are not necessarily real strategies, are presented for the development of synergies that the author wants to present, emphasizing its importance.

- **Store layout and location**

When a company owns sales points, the exploitation of cross-business synergies between business also passes from their positions. The best way to lay the foundations for exploiting synergies is when products of two or more businesses are sold in the same location (that is a multi-business store). This allows managers to implement strategies to exploit synergies much more easily: just think of the ease with which cross-selling and cross-business bundling promotions can be made. Furthermore, customers would encounter a "shopping centre" effect, i.e. once they have purchased the products for which they came, they could go on to buy others from other businesses because they are close, so business reinforcing each other, in fact the greater level of collective offer means that a higher number of customers comes to the place (these are the reasons why premises in shopping centres have very high rent costs). It is not enough, however, that the products of the two businesses are sold in the same store to make the most of the synergies: often there is need a design of the internal layout of the store, that allows customers to participate in sales of the products of both businesses. This layout should be studied on a case-by-case basis. On the contrary, when the two businesses of the company are partly competitors on some product lines, selling them in the same location could be a mistake as there would be a high risk of cannibalization among businesses (so, the competition between the two businesses precipitates the demand for the "loser" one). In any case, the opening of a sales point requires a careful strategic analysis, a study to customers who live near it and the attention to potential competition present in the surroundings.

- **Customer segmentation and customized cross-business promotions**

Customer segmentation is a marketing activity, in which customers are divided into groups with same certainly characteristics. Segmentation variables could be: shared needs, common interests and buying behaviour, similar lifestyles, or even similar demographic profiles. Today segmentation techniques are many, operated by CRM software or implemented ad-hoc through controlled clustering or classification algorithms. Customer segmentation allows for customized cross-business promotions, that are bundling and cross-selling discounts (or in general joint market penetration

techniques) targeted to customers segment who are believed to respond more effectively to the strategy undertaken. Customized promotions are increasingly used in e-commerce business for ease of customer profiling.

- **Timing for growth synergies strategies**

The moment in which to implement strategies for the exploitation of synergies must be studied carefully, and demand and environment variables forecasting models become fundamental in this sense. In fact, think for example of a very sold product that presents a demand seasonality: for example, if it served a dragging of another product in a cross-business bundling strategy, it would certainly be convenient to propose the bundle during the season in which there are the peak. Moreover, all strategies listed above must be considered in a certain time range, as the environment variables change and some strategies, such as joint diversification, can be incredible opportunities at a certain time and dangerous errors if implemented too much early or too much late.

Chapter 3

DATA ANALYTICS FOR BUSINESS DECISIONS

In this chapter, disconnected from the previous ones, an overview is given on data analytics and on the choice of companies to base their decisions on data. In detail, in the first sub-chapter we deal with the importance of implementing data-driven decision-making processes in companies, with an overview of concepts such as Data Mining and Machine Learning, while in the second sub-chapter we deal with a system that easily allows this implementation: data warehouse systems. Finally, in the third sub-chapter, the main Data Mining and Machine Learning techniques are presented, supplemented by the explanation of some of the most important algorithms known in these fields.

3.1 Data driven decision making

Today, disproportionate amounts of data are always generated by the incessant internet traffic and by every kind of sensor installed on every type of modern device. Our way of living is changing and therefore the business world is changing with it. Companies without adequate capacity to manage and process large amounts of data to their advantage are increasingly outclassed by the competition that is capable of it. In this scenario, a new type of science (Data Science) develops as quickly as technological change is disruptive; is a science that offers methodologies for the interpretation and processing of large amounts of data that statistics with traditional methodologies is no longer able to manage. The goal of data science is to transform data into knowledge, that drive decisions.

In computer science, a datum is a simple element, consisting mostly of symbols. Information is the result of the elaboration of more data that organized, contextualized and interpreted, take on a real meaning. Therefore, information attributes give

meaning to data that, in isolation, would not have it. For example, by querying a relational database to obtain the identity of a person, a datum may consist of a name, a surname, or a date of birth, which aggregates represent the information of the identity. Through the analysis of one or more information it is possible to produce a knowledge on a specific subject, object or event. Knowledge of things is the means by which actions, behaviours and decisions are established. Structured and contextualised information allows us to understand the reality that surrounds us and therefore to be able to predict the evolution of the future to some extent. For this reason, people "who know" can establish a correct perception of reality and have the power to make decisions to change future events to their own advantage. Thus, the process of transforming information into knowledge represents a strategically important step for every company.

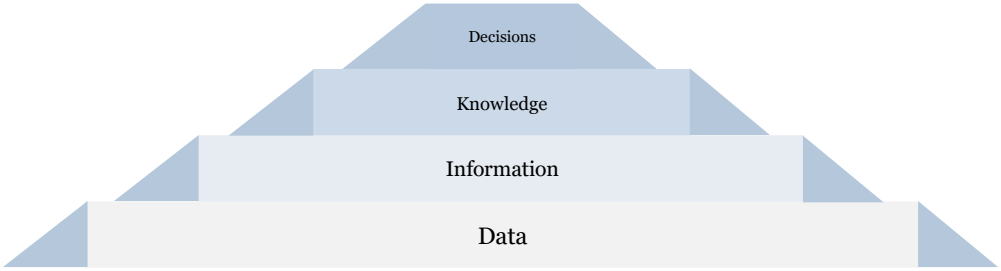


Figure 3.1. Pyramid of data-driven decision making.
Source: author

"Data science is a set of fundamental principles that guide the extraction of knowledge from data" [12] . The main concept of Data Science is the so-called Data-Driven Decision Making (DDD), that is the objective of making its own decisions based on precise and rigorous data analysis, rather than on intuition. "Data-driven decision-making (DDD) refers precisely to the practice of basing decisions on the analysis of data, rather than purely on intuition" [12]. In fact, companies in every sector, all over the world, are becoming more and more data-driven, focused on exploiting data for competitive advantage. Understanding the main concepts of business analytics or more generally become a data-driven organization helps assess opportunities and threats, and address to right decisions to increase profits. In fact, every company can benefit from the data available to it: IT company for intrusion detection, hospital for patient record anomalies, pharmaceutical company for disease penetration, etc... It is also enough to consider that many large companies, like Facebook, Twitter, Google and Amazon, base their business on data, having high valuations due primarily for this reason. The benefit in terms of performance that companies obtain through a data-driven decision-making process has been demonstrated by the economists Brynjolfsson, Hitt and Kim [13] : the more decisions are based on data, the more the company is productive.

This tendency to use more and more data is mainly since the technology offers the tools suitable for storing and processing data that did not exist before: computers have become more powerful and new efficient algorithm have been implemented. Moreover, the volume and the variety of data have far exceeded the possibility for manual analysis or for conventional databases.

New words in Data Science fields like Machine Learning, Data Mining and Big Data have increasingly taken place among the innovative challenge of companies.

Machine Learning is a sub-field of artificial intelligence that deals with developing methods for improving knowledge of an intelligent agent over time based on the experience that the agent himself develops in his experience. Usually, Machine Learning processes involve data analysis and prediction.

Data Mining is the process of extraction of information and patterns from large amounts of data. Many consider the field of Data Mining as a derivative of Machine Learning, or at least something that is closely related to it, in fact both areas deal with data analysis and share some algorithms. In detail, however, Machine Learning includes fields such as robotics, computer vision, and others, unrelated to Data Mining process. Data Mining is usually used for customer relationship and customer behaviour in order to obtain appropriate strategic solutions (and this is precisely what we need to exploit synergies). Moreover, effective Data Mining requires a lot of data, from databases, data warehouses, web or from many other sources. This will be discussed more in detail in Par. 3.2.

Big Data literally means "lots of data" and refers to the fact that data available to companies are many and in many forms, such that traditional methods and technologies are not able to analyse them. Big data technologies are born precisely for this purpose and are occasionally used to implement very sophisticated Data Mining techniques. For more details on these, please refer to more accurate studies.

3.1.1 From data to strategy: Data Mining process

So far it has been said that it is important for a company to use data for decision-making and strategic purposes. But how to analyse data to identify useful information? Every decision-making process is unique, it depends on the business and its objectives, on people, on the external environment and on its constraints. However, we can outline a basic structure that guides the various steps for the extraction of effective strategy actions from data. Point 2, 3 and 4 are known as *Data Mining process* or *knowledge discovery process*. All the steps are as follow:

1. Business understanding

There is no sensible analysis if you do not understand the business first: the market and competitors, the problems to be solved, the constraints, the opportunities and the people involved. Question like “What exactly do we want to do? How exactly would we do? What problems are there? What margins for improvement?” are necessary before going into advanced analysis, in fact these analyses would have no sense of existence if a clear objective is not defined first, indeed they would be confusing and anti-productive. This phase includes all the basic statistics of the business (known as “summary statistics”), such as information on customers and what they buy, revenues by point of sale, profits, etc..., that are all that is needed, precisely, to fully understand the problems and opportunities of the business.

2. Data understanding

Once you understand the business, you need to understand what data you need for advanced analysis, which ones are available and if they are enough to reach the set goals. In companies, historical data is often collected in traditional databases, spreadsheets or data warehouses. But often companies don't even know what they have available, and also finding data they need is not trivial if processes are messy. For this reason, data warehouses (Par. 3.2) are taking more and more shape for its simplicity of interrogation. Understanding what you have available for analysis is a fundamental step to avoid doing most of the work and only in a second moment realizing that you don't have the tools to do it totally. Moreover, understanding data can also direct the analysis towards objectives that are a little different from those previously set.

3. Data preparation

It is one of the most complex phases of the process. Tendentially, the data must be manipulated before being input into an algorithm. Some sub-phases of this step are:
a) *Data Cleaning* and *Data Quality*, that refers to remove noise and inconsistent data;
b) *Data Integration*, that means to combine multiple data source, usually with some conversions;
c) *Data Selection*, in which only the data useful for the analysis are chosen;
d) D*ata Transformation*, where data are transformed and consolidated into appropriate forms for mining by processing operations. For example, the algorithm requires categorical data, normalized data, in percentage form, etc.

These sub-phases are sometimes cyclical.

4. Modelling & evaluation

The modelling step is where Data Mining techniques are applied to data and includes a multitude of algorithms and intelligent methods that are continuously discovered or improved. Output of modelling is a sort of pattern that captures information or regularity in data.

The evaluation stage, through intelligent measures, is in charge of judging how the created model or the techniques used are effective. In fact, it is necessary to understand if the patterns extracted from data are true or if they are only particular anomalies, for example due to sampling. Another very important purpose of this phase is to understand if the set objectives have been satisfied.

5. Presentation & deployment

In this phase analysts present the results obtained to managers and interact on the concrete strategies to be performed to improve the business. It seems a trivial phase, but it is very difficult for technical roles to make himself understood by non-technical people. Good practice is to present the results not in the form of black-boxes, because managers, who have the real responsibility for strategic actions, will hardly trust blindly results that they cannot understand.

3.2 Data Warehousing: an overview

It has been said that data can be taken from different sources. However, some sources are decisively more "comfortable" than others, and not only in terms of extraction, but also of storage. Informational applications, such as warehouse management, sales management, order management and invoicing, analytical and financial accounting, produce data and perform a large number of elementary operations. These transactions are essentially static and photograph routine and repetitive operations. Therefore, the data thus produced are of poor use to decision-makers, who need summary and aggregated information to perform fast and effective analysis.

Data warehousing systems serve to produce essential and concise business information in a timely manner. This information is usually obtained with few and complex queries and data scientists often become quite skilful at writing queries to extract data they need.

A data warehouse, system that took place in the early 2000s, is a repository of information collected from multiple sources and stored under a unified schema, are

constructed from a process of data cleaning, data integration, data transformation, data loading, and periodic data refreshing. Data in a data warehouse system are reshaped into new subject-oriented structures (products, customers, suppliers, …), integrated (data is homogeneous), reusable (historical data is stored), and non-volatile (original data do not change over time). Therefore, given that they favour queries and maintain clean and reusable historical data, data warehouses are the best base for performing Data Mining.

Regarding the basic characteristics of data warehouses, the first step is to identify the type of data that will have to enter and become part of it, and then understand the most appropriate use for decision-making support. In doing so, however, it must be borne in mind that the primary objective of data warehouse is to allow users to manage the business of the company more efficiently and effectively, so it is not necessary to historicize all data, but only those that will serve the business. Furthermore, data is entered in data warehouse in aggregated form, if necessary even at multiple aggregation levels, and this is to favour the speed of queries, normally performed in SQL language, and saves memory ,which otherwise would not be enough to store all the non-aggregated data produced. Queries are, to be precise, the ones that actually play a key role in the choice of data to be stored: the most frequent and important queries for the business (for example monthly revenues per customer, revenues by point of sale, etc., depending of company needs and strategy) define which data to storage and at what level of aggregation.

Many find it easier to think of Data Warehouse as a multidimensional cube (*Figure 3.2*) , in which each dimension reproduces a variable of interest. Each point within the cube is the intersection of the coordinates defined by the sides and can be thought as the measure of the business for that particular combination of variables. For example, if the variables of interest are time, customer and store, and the measure of the business to be considered is revenue, each cell represents revenues that a certain

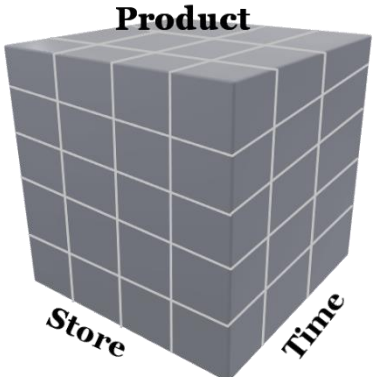


Figure 3.2. Data warehouse multidimensional cube.
Source: author

customer (ex. John Smith) has made in a certain store (ex. point sale X) in a certain period of time (ex. January 2019).

A model of this type is obtained by adopting a *Star Schema* (*Figure 3.3*): the name of the diagram derives from its resemblance to the image of a star, with a large central table and a set of smaller tables on the sides. Unlike traditional entity-relation schemes (which set relational normalized tables with primary keys and foreign keys without optimizing speed for queries), the star schema is strongly asymmetric: in the centre there is a dominant table, the *fact table*, of huge dimension, which connects with all the other tables (*dimensional tables*) through key fields, and contains measures that "explain the fact". Returning to the previous example, the fact table represents the business branch (sales), measures are attributes of the fact (revenues, discount, receipt line,...) and dimensional tables represent dimensions in which measures must be calculated (customer, store, time), and contain the attributes that describe and quantify them: each cell of the cube is represented by a record of the fact table and the sides of the cube are defined by the primary keys of the dimensional tables.

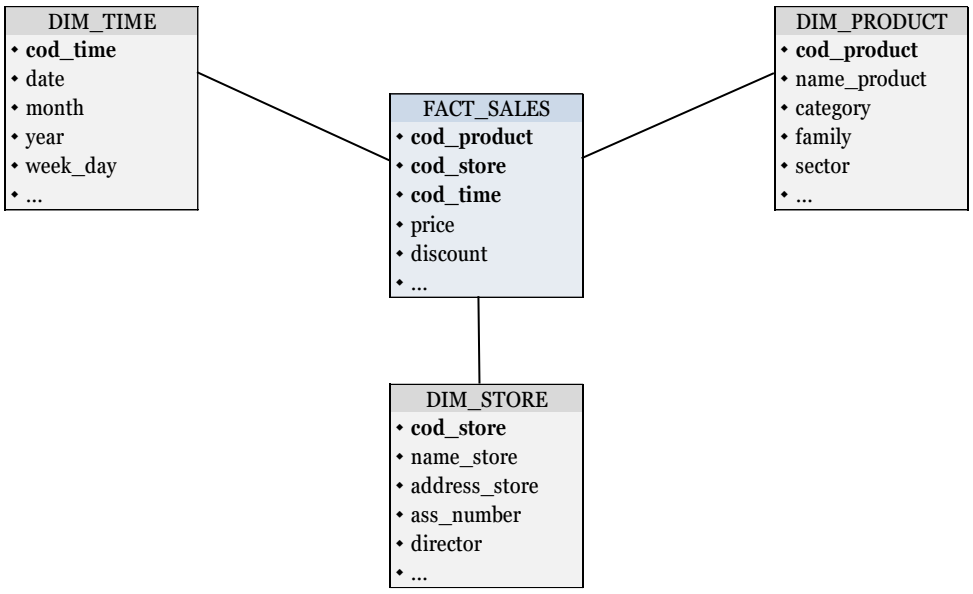


Figure 3.3. Example of a Star Schema.
Source: author

Furthermore, the fact table is the only normalized table (i.e. it does not contain data redundancy or aggregated), while dimensions contain redundant and descriptive attributes (for example, the store table can contain *cod_store*, *name*, *complete_name*, *tel_number*, *director*,... , the time table can contain *cod_time*, *day*, *month*, *year*, ...), and this favours the speed of interrogation at the expense of the memory used for archiving. In this way, in fact, during SQL queries, joins between tables (very expensive

computationally) and group by functions are drastically reduced. Moreover, the fact that dimensional tables are redundant is not a big problem since they contain few data (many orders of magnitude lower than the fact table) and they are statics, that is they do not change or change very slow over time (think for example of data personal details of a customer, or time schedule).

According to La Noce & D'Ercole [14] , the steps for creating an efficient data warehouse are:

1. *Choose a business to model* (in other words, choose a "fact", for example "sales");
2. *Identify the minimum level of granularity required for the fact* (for example sales per day per customer per store);
3. *Identify the minimum level of granularity required for each dimension* (for example "day" for calendar table);
4. *Choose the measures that populate the fact table* (measures are the data that "explain the fact", for example "revenues", "discount", ..., that will come into the SQL SELECT);

The technical and infrastructural level with which the data warehouse is installed, alimented (ETL) and managed is outside the scope of the document, so please refer to other more detailed studies.

Data warehouse systems allow OLAP (On-Line Analytical Processing) work sessions, that is a set of software techniques for interactive and fast analysis of large amounts of data, which can be examined in rather complex ways. In fact, the cube mentioned earlier is also known as an OLAP cube. OLAP analysis allows more specific search operations than simple relational databases. The most used functions are: roll-up, the reduction of the level of detail of one of the present dimensions (for example *group by store, months -> group by city, months*); drill down, the increase in the level of detail of one of the dimensions present (for example *group by city, months -> group by store, months*); slice and dice, the selection of a subset using predicates or combinations of predicates (for example *where 'months = January'*); pivot of tables, that is reorganization of the orientation structure without varying the level of detail. This type of "on-line" data reprocessing is allowed by the redundancy of the attributes in dimensions and in general by the star schema model.

It is important to repeat that the whole system is so made to favour the speed of data interrogation, storing only the necessary ones in one or more aggregation levels.

3.3 Main Data Mining, analytics and Machine Learning techniques

After having made an excursus on how important it is to base decisions on data, let's see some of the main techniques that Data Mining field offers. These techniques are chosen because they are applied in Chapter four for the purpose of exploiting synergies between businesses. The author is aware of the fact that the kind of patterns available are many, but here we analyse classification, clustering, association rules, forecasting and an introduction to neural networks.

For each technique we analyse an algorithm that belongs to it, sometimes the best known, sometimes the most useful for our applications.

3.3.1 Classification

"Classification and class probability estimation attempt to predict, for each individual in a population, which of a (small) set of classes this individual belongs to. Usually the classes are mutually exclusive" [12] .

Classification is the most commonly applied Data Mining technique. Objective of the classification is to find a descriptive profile for each class, which allows to assign objects of unknown class to the appropriate class. Basically, the process consists first in the construction of the model with a portion of data (training set) and then an evaluation of the model with the remaining portion (test set). Of these data, the class label is already known but for test set it is removed, foreseen by the model, and compared with the real label to understand how good the model is in classifying. Once the accuracy of the model has been evaluated, if it is satisfactory, the model can be used for new data, i.e. data of which the class label is not known. Classification offers endless applications, from customer segmentation to forecasting failures in all areas.

The best way to evaluate a classificatory is to consider the following metrics:
- Accuracy: how accurate is the classification;
- Efficiency: how burdensome the creation of the model is in terms of time;
- Scalability: how the model is usable for an indefinite number of records;
- Robustness: as the model is robust even if there are dirty data;
- Interpretability: how much the model can be explained (no black box).

Common algorithms include Decision Tree, K-Nearest Neighbour, Support Vector Machines, Naïve Bayes classification, and Neural Networks applications. Let's see CART and KNN models.

- **CART: Classification And Regression Tree**

CART model is a binary tree in which each node represents a single input variable (attribute) and a split point on that variable. The leaf nodes of the tree contain an output variable, that is the attribute to predict. To explain the detailed functioning of the algorithm, a simple example is used.

Suppose that, based on historical data (that are for example the attributes shown in *Table 3.4*: married, annual income, remaining years of mortgage), you want to predict whether the customer of a bank will pay his mortgage or not in future. Historical data are divided into two, in different colour in figure: training set, in grey, usually about 70%, and test set, in blue, about 30% of the total record available.

STATUS	ANNUAL INCOME	REMAINING YEARS	PAY
married	30K	3	YES
married	10K	8	NO
single	23K	4	YES
single	25K	12	YES
single	15K	1	YES
married	8K	1	YES
single	55K	11	YES
married	17K	4	NO
single	16K	3	NO
married	22K	4	YES
married	19K	4	NO
married	7K	1	YES
single	20K	5	YES
single	30K	4	YES
married	9K	3	NO
single	21K	4	NO

STATUS	ANNUAL INCOME	REMAINING YEARS	PAY
single	10k	12	?
married	22K	5	?
single	23K	2	?

Table 3.4. Example of data. "Pay" in blue table, is the attribute to predict.
Source: author

The algorithm creates, starting from the data of the training set, a binary decision tree, like the one in *Figure 3.5*. The algorithm creates this tree considering three factors, always modifiable by the user when creating the tree:

1. *Split formulas*: the algorithm decides on which attribute to split at each step (tree level). Particularly, the attribute that "best divides" data according to the split formula used is chosen. Formulas widely used, for example, are the Gini Index for continuous attributes and the Information Gain for categorical attributes. For example, $\mathrm{GINI(t)} = 1 - \sum_{j=1}^{n}[p(j|t)]^2$ where t is the node, and $p(j|t)$ is the relative frequency of class j at node t. The algorithm will chose the attribute for which GINI is minimum.

2. *Stop criterion*: are rules that determine the closure of a path with a leaf node. The most trivial: there are no more partitioning attributes, there are no more training data, the node contains only examples of the same class. Among the most complex, instead, there are the criteria that do not require all the examples to belong to the same class, but only high percentage (ex. 99%) to close the path.
3. *Pruning*: these are features that the algorithm offers so that the tree does not become excessively deep and complex. This would in fact lead to *overfitting*: overfitting is the situation for which the tree is so perfect for the training set, that the test set and future data will certainly not match well.

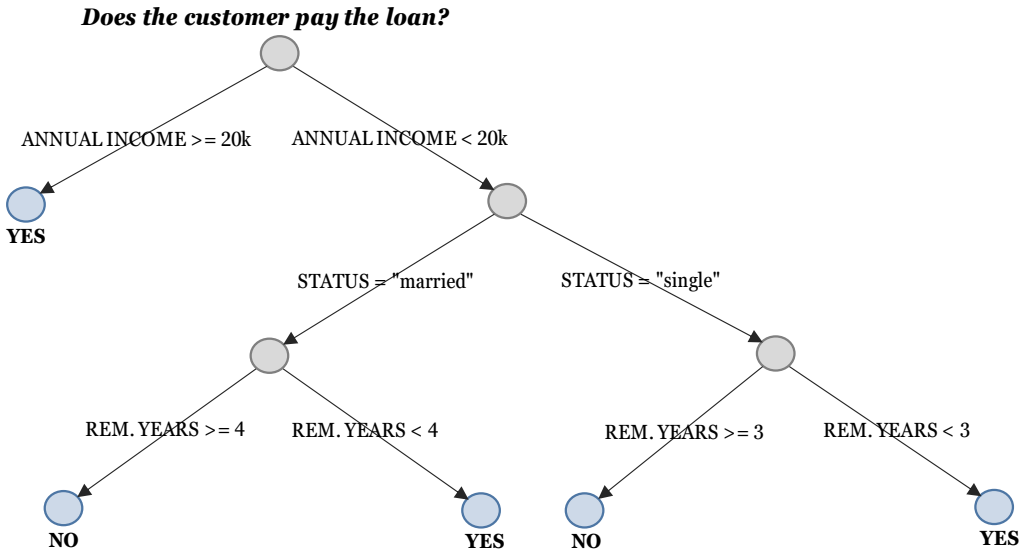


Figure 3.5. Example of a CART tree model.
Source: author

Software often offers many possibilities for customizing the tree, but describing all them would only cause confusion.

Once the training model is created, the test set records are passed in the tree. Then it is possible to evaluate the accuracy of the model by comparing the responses given by the model with respect to the actual data labels. If the correct response rate is considered valid, the model can be used to provide class labels that are not known.

It is important to underline that the example given is really too simplified. Trees can reach depths of even more than 100, created with hundreds of thousands of records and tested with as many before defining them as "good models".

The advantages of this algorithm are its ease of construction, its speed of classification and interpretation of results. However, it is not very robust as far as missing value is

concerned and, even if it does not seem to be, it is very difficult to use because setting all parameters correctly is complicated as they are many and influence each other.

- **KNN: K-Nearest-Neighbour**

Many believe it is the simplest Machine Learning algorithm. The concept behind the classification is in fact very simple: an object is classified according to the majority of the votes of its nearest *k* objects. But what does "near" mean? For the calculation of the distance, the objects are represented by vectors in a multidimensional space, after which a distance formula is used, usually the Euclidean or the Manhattan.

k is chosen by the user as a parameter, and the decision on which *k* to choose for a good classification can be made through many techniques, among which cross-validation for many *k* stands out: data is divided into groups of equal numbers, it is iteratively excluded one group at a time and try to predict with non-excluded groups. This is to verify the quality of the prediction model used for every *k,* avoiding as much as possible the randomness factor in the choice between training and testing.

The advantages of this algorithm are its simplicity of understanding and interpretation. Problems related to this algorithm arise from its simplicity, from the strong dependence on the data used and from the computational burden of calculating distances when the input attributes are many and not all numerical. Moreover, when the data is strongly unbalanced towards a class, this class will always be assigned to all test set data. KNN is also used for choosing parameters of other algorithms, as happens in DBSCAN case (see next Par.).

3.3.2 Clustering

“Clustering attempts to group individuals in a population together by their similarity, but not driven by any specific purpose” [12] . In other words, clustering is a process of grouping a set of objects into classes of similar objects. In fact, clustering algorithms consider data in tuples (records) and each record is assigned a cluster based on how “similar” it is to objects in the cluster and how “dissimilar” to objects in the other clusters. As in classification, also in this case the concept of "similar" refers to the distance between multidimensional vectors.

Although classification is an effective means of distinguishing groups or classes of objects, it requires a construction set and labelling of the training set, that are often expensive. It may often be desirable to proceed in reverse by first partitioning the data into groups based on their similarity (that means, using clustering) and, subsequently, assigning the labels to the relatively small number of groups thus obtained.

As a Data Mining feature, clustering can be used to examine distributions of data, to observe the characteristics of each distribution and to focus on those of major interest. Alternatively, it can be used as a pre-processing step for other algorithms, such as association rules, which operate on the identified clusters. In economics, clustering can help operators to discover distinct groups of typical customers based on their purchases, and that's what we'll do in the next chapter. Clustering can also be used to search for outliers (i.e. very distant values from each cluster), that in some applications are even more important than common values.

The quality of clustering can be measured with different indices, whose effectiveness depends on the particular algorithm used.

Clustering algorithms are divided into families based on the operation methods. Basically, every algorithm belongs to one of the following families:

- Partitioning methods
- Hierarchical Methods
- Density-based methods
- Grid-based methods
- Model-based methods

For reasons of time, space and objectives of the document, we cannot see them all. The two most famous and used algorithms, k-means (partitioning method) and DBSCAN (density-based method), are presented below.

- **K-means**

k-means is a partitional algorithm, in which a centroid is assigned to each class (that is its mean for k-means, its medoid for k-medoids, and others) and each point is assigned to the cluster with the nearest centroid. k is a parameter to be choose as input and represents the number of clusters to be obtained.

More in detail, the algorithm works as follows:

1. The algorithm randomly chooses k centroids in space;
2. Each object (record) is assigned to the cluster whose centroid is closest according to the chosen distance function (ex. Euclidean);
3. Cluster's centroids are recalculated. If centroids don't change, stop. Else, return to step 2.

It is clear that the final result depends on the position of the starting centroids.

One of the most used measure for k-means evaluation is the SSE (Sum of Squared Error): it is the sum of the squares of the distances from each object to the centre of its cluster. The smaller the SSE, the better the clustering result. Be careful, however, that

the SEE usually decreases with the increase in the number of clusters (groups) to be obtained, so a good compromise must be found. This compromise is found at the knee point of the SEE-k curve. So, the SEE is mainly used to choose the best k.

To get good results from k-means, in general what you need to do is:

a) try the algorithm for several k;
b) for each k, start it several times and choose the minor SSE solution;
c) put the best solution for each k on a SEE-k graph and choose the k of knee, that is the k that has the best compromise between cluster number and SSE quality (*Figure 3.6*).

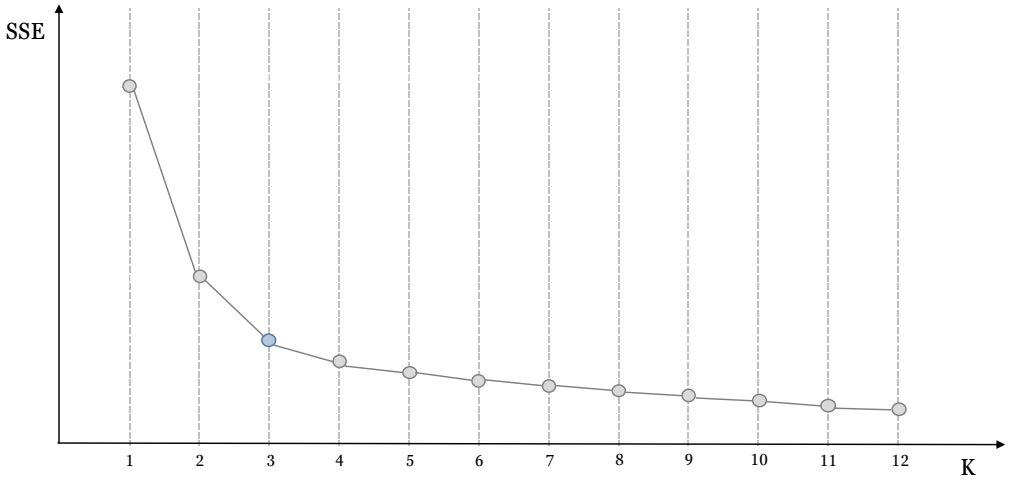


Figure 3.6. Example of SEE-K graph. In blue the knee of the curve.
Source: author

K-means offers the advantage of being a low computation load algorithm, so it converges very quickly. However, it has many non-negligible defects, including: the clusters are also very different in size, the result depends on the initial centroids (sometimes other algorithms are used to define good starting centroids), does not consider the density of objects in space, and it is very influenced by the outliers (because they move the centroids considerably).

In *Figure 3.7*, an example of successful k-means clustering is shown.

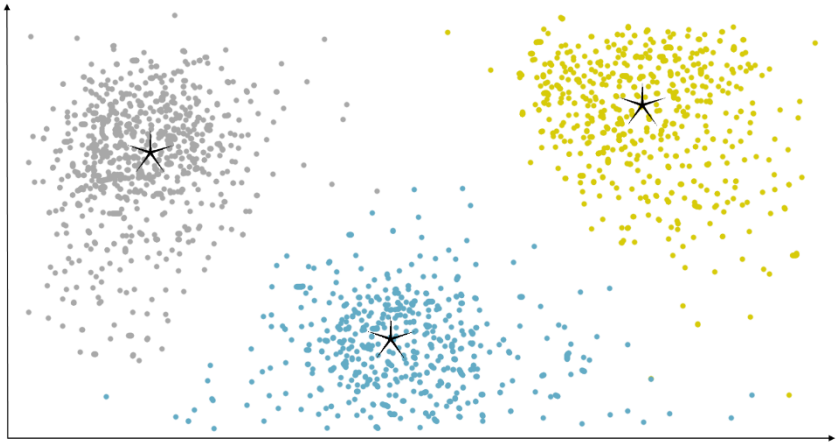


Figure 3.7. Example of k-means clustering (k=3) on two-dimensional approximation. Source: author

- **DBSCAN**

DBSCAN, more recent developed compared to k-means, is a density-based algorithm. This means that groups in clusters are based also on the density of points in space. In particular, it groups points of the high-density areas into clusters and considers points in low-density areas as outliers. The algorithm uses two key parameters chosen by user: *MinPnt* (min points) and *Eps*.

More in detail, the algorithm considers, after putting all records in a space in the form of points, that:

- An object, which becomes a point in space, is a *core point* when it contains at least MinPnt in its Eps radius;
- An object is a *border point* when it is not a core point, but is within the Eps range of a core point;
- An object is a *noise point* (outlier) when it is neither core point nor border point.

Once you understand this, the algorithm works simply:

1. The algorithm places all point in space considering them as multidimensional vectors;
2. Calculate, for all points, which type they belong to (core, border, noise);
3. It builds each cluster so that core points and border points of the same cluster are connected by eps of core points of the same cluster. All noise points are instead enclosed in a dedicated cluster (usually called cluster0).

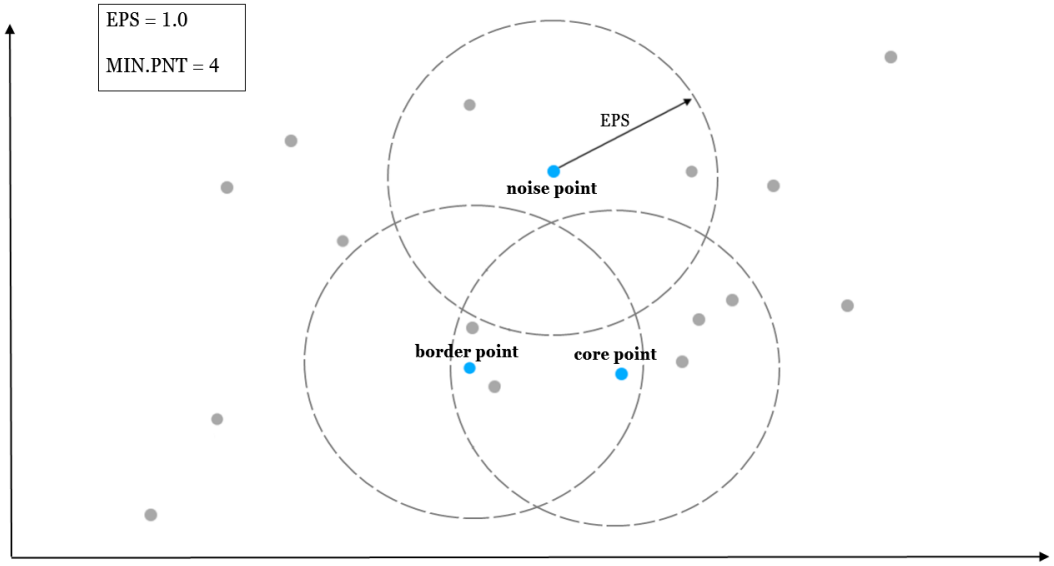


Figure 3.8. Two-dimensional representation of core points, border points, and noise points. Source: author, based on image [15]

Therefore, DBSCAN does not allow you to choose the number of clusters to be obtained a priori, but allows you to choose only MinPnt and Eps parameters. The difficulty of DBSCAN lies precisely in the choice of these parameters. The KNN is functional for this purpose: chosen a value of MinPnt, the KNN offers good advice for the value of Eps to choose. In particular, using the KNN with k=MinPnt, we construct the [X=point, Y=distance of the k^{th} point] graph ascending sorted: the knee point is chosen and consequently Eps is read on Y-axis. The knee point is in fact the best compromise between cluster density and number of outliers that will be obtained from DBSCAN (all points to the right of the knee point will be border or noise points).

To assess the quality of clusters found with DBSCAN, there are several measures, including SSE and entropy.

DBSCAN, if repeated several times, always gives the same result. However, it is often useful to use more DBSCAN cycles, especially when data has different densities that you want to intercept. With this method, for each result, one or more clusters obtained are given in input to a new DBSCAN algorithm.

DBSCAN has the advantage of being robust to outliers. Moreover, its power to detect them is also used before other cluster algorithms that are not very robust to outliers in order to remove them from data before processing. DBSCAN remains very useful even when you have data with really different densities that you want to intercept. Defects of this algorithm, however, are the high computational cost, and the impossibility of choosing the number of clusters to be obtained, if not through various attempts by changing parameters.

In *Figure 3.9*, an example of successful DBSCAN clustering is shown.

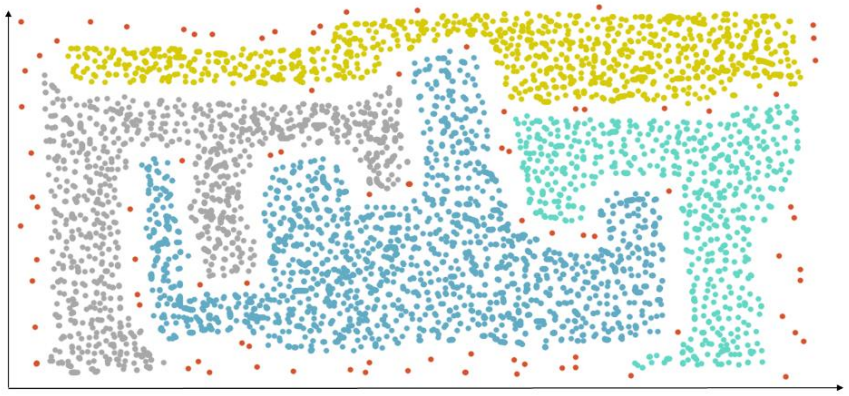


Figure 3.9. Example of DBSCAN clustering on two-dimensional approximation. Source: author

3.3.3 Association Rules

Association rules are used to find relationships between objects which are frequently found together. Applications of association rules are basket data analysis, classification, cross-marketing.

The objective of association rules is to extract frequent patterns from a transactional database: a transactional database is in the form "<*id*><*list of items*>" (for example "<id_receipt><tomatoes, peaches, water, ...>" in case of basket analysis), where lists of items are called *itemsets*.

An association rule occurs in the form:

$$A \Rightarrow B \text{ (reading: "A then B")}$$

Where A and B are itemsets (therefore they can include one or more objects). The meaning of the expression is that, in the transactional database, when A is found then B is also found. A represent the *Rule Body*, B represent the *Rule Head*.

The search for rules is done by setting two thresholds:

- *Support*: is the fraction of transaction containing both A and B. In formula:

$$support = \frac{\#\{A, B\}}{\#id}$$

- *Confidence*: is the frequency of B in transactions containing A. In formula:

$$confidence = \frac{\text{support}\{A, B\}}{\text{support}\{A\}}$$

To better understand, see this example below.

ID_RECEIPT	LIST OF ITEMS
10097	tomatoes, potatoes, water
10098	bread, tomatoes, potatoes
10099	bread, potatoes
10100	water
10101	tomatoes,water

Figure 3.10. Example of data in transactional form.
Source: author

Starting from the data in *Figure 3.10*:

The rule tomatoes ⇒ potatoes has $support = \frac{2}{5}$ and $confidence = \frac{2}{3}$;

The rule bread ⇒ potatoes has $support = \frac{2}{5}$ and $confidence = 1$;

The rule water ⇒ potatoes has $support = \frac{1}{5}$ and $confidence = \frac{1}{3}$;

… … … …

In general, given n itemsets, there are 2^n possible candidates itemsets. So, it would be decidedly senseless to produce all the possible rules for huge data. The search for patterns in data is therefore done by setting a support threshold, which ensures that only the frequent items are considered, and a confidence threshold, which determines the strength of the co-occurrence. In this example, placing $minsup = \frac{2}{5}$ and $minconf = \frac{3}{4}$, only the second of the rules brought for example will be given in output.

The confidence threshold, sometimes, is not very indicative of the strength of a rule. This happens when the Rule Head is in itself very frequent. For this reason, it is important to also consider lift. Lift is defined as:

$$lift = \frac{\text{support}\{A, B\}}{\text{support}\{A\}\text{support}\{B\}}$$

In general, the procedure for obtaining good association rules from a database is the following:

1. Manipulate the data to have them in transactional form;
2. Apply an association rules algorithm by setting support and confidence thresholds so that the desired number of rules is obtained (the higher the thresholds, the less the rules);
3. Order the rules obtained by decreasing lift;
4. Check the rules one by one and eliminate those that don't make sense or are too trivial.

There are many algorithms for association rules, which work differently but give the same result. The two best known algorithms are Apriori and FP-Growth. We see only the first.

- **Apriori**

Apriori looks for the rules first of all considering only the support threshold. Once he has found all the rules that exceed the support threshold, he filters them considering the confidence threshold.

Apriori seeks to improve efficiency in the search for rules by reducing the number of candidates itemsets to be valuable rules (pruning). In particular it follows the principle by which "If an itemset is frequent (that means, it exceeds the support threshold), then all of its subsets must also be frequent". In fact, the support of an itemset can never be higher than the support in its sub-itemset.

So, the algorithm looks for combinations of items that exceed the threshold starting from combinations of 2 objects. If itemsets of size 2 do not exceed the threshold, then all the potential items constructed adding items to these itemsets are not explored, that is, it is not checked how many times they appear in the dataset.

Even with this type of tricks, however, the algorithm remains a bit slow in searching for frequent itemsets. Some solutions may be to increase the support threshold, use statistical probability techniques to predict frequencies rather than calculate them precisely, or, if the dataset is really too large, work only on samples.

FP-Growth algorithm, using tree structures, is faster and improves several aspects of Apriori when faced with huge data.

3.3.4 Time series forecasting

Time series forecasting field tries to solve business problems that involve time component. A time series is a sequence of observations taken sequentially in time. The goal is to predict future observations based on historical data, also considering time dimension. The field of time series is not considered as much Data Mining as it is related to statistic field.

Having more historical data available always results in a more accurate prediction (in terms of probability). Furthermore, accuracy is greater when values are predicted closer, that is in near future, rather than values that are very distant in time. The role of the frequency, that is the interval in which the historical data are detected, does not have a strong impact on the quality of prediction, but in general it must be evaluated

case by case. The impact of outliers and missing values depends very much on the algorithm to be used.

At a theoretical level, the components of a time series are:

- *Trend*: exists when a series increase, decrease or remain in a constant level with respect to time;
- *Seasonality*: property that show repeated periodical patterns at a constant frequency;
- *Randomness (*or *residuals)*: everything that cannot be explained by trends or seasonality (random does not mean that it is completely due to the case, but it can also be due to causes external to seasonality or trend detected over time).

Figure 3.11 shows the decomposition of a time series in the three components just described.

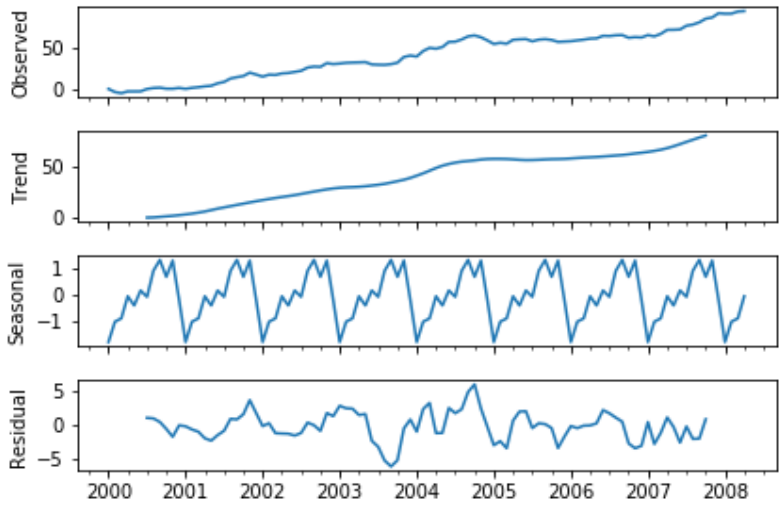


Figure 3.11. Decomposition of time series into trend, seasonality and randomness. Source: not subject to copyright

There are many algorithms, among which the most important ones are: linear regression (the simplest and most intuitive), exponential smoothing, ARMA and ARIMA, Dynamic Linear Model, and Neural Network applications. Each algorithm has its pros and cons, such as ease of use and programming, interpretability, implementation of confidence intervals for results, sensitivity to outliers, and so on. In this document we treat the ARIMA method.

- **ARIMA**

ARIMA, acronym of Auto Regressive Integrated Moving Average, is a more particular form of linear regression, very complex mathematically, so only its broad outline is explained.

The model can be understood by looking at the parts of its name individually:

- *Autoregression (AR)*: refers to a model that shows a changing variable that regresses on its own lagged values.
- *Integrated (I)*: it means that it uses the difference between data values and previous values instead of real values to predict future values. In this way the model tries to transform non-stationary data into desired stationary data (that means, data values do not depend on time, so seasonality and trend components are thus removed, or better, bypassed).
- *Moving Average (MA)*: the model incorporates in its predictive calculations the dependence of an observation from the previous values according to the formula of the moving average (the average of the q values preceding the one to predict, with q basically free to be chosen).

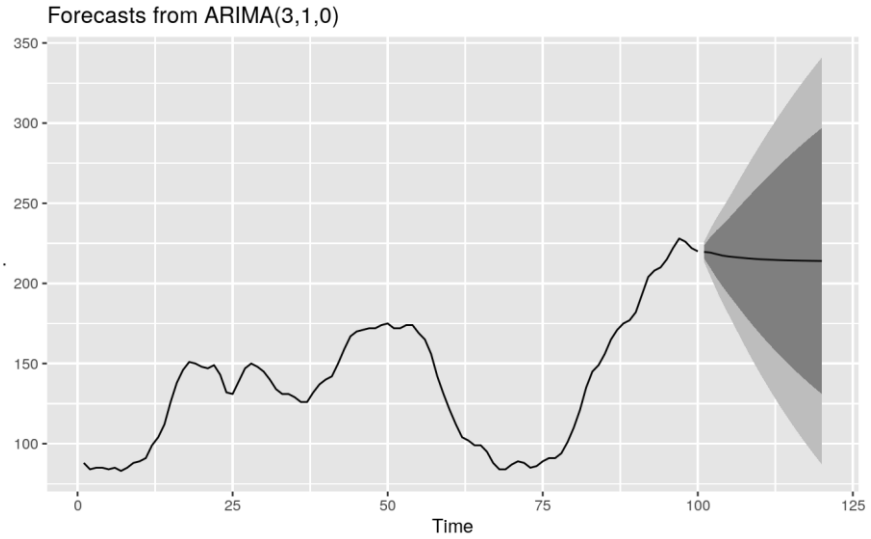


Figure 3.12. Example of ARIMA prediction with confidence intervals.
Source: not subject to copyright

The standard notation of ARIMA is*: ARIMA (p, d, q)*, where p, d and q are the settable parameters. In detail, p is the number of lag observations on which to do autoregression, d is the degree of finite differences in which data values and previous values are manipulated, and q is the size of the moving average window.

Results of the model are provided by some libraries or software as an average value with attached confidence interval, for example at 95%.

3.3.5 Neural Networks

Neural networks have a fundamental role in Machine Learning field. For this reason, although they are neither deepened, nor used in chapter four, they still deserve to be mentioned and broadly described.

The idea of artificial neural networks is to emulate the functioning of the brain, or at least some learning functions. They are a mathematical model that is inspired, therefore, by the network of human neurons, in which small single elements connected to each other, the neurons, form an interconnected network. That of neural networks is a field in strong expansion and research because has a remarkable ability to derive meaning, patterns and trends from complicated data.

A neural network is a kind of black box (and for this reason, it is still far from being applied in some fields that require the interpretability of the results achieved), consisting of simple processing units, the neurons, and directed weighted connections between those neurons, which defines link strength.

More in detail, the algorithms of neural networks are structured in different levels, each consisting of a certain number of neurons (we speak of a "multilevel system"), which represent the nodes of the network. For every level, the algorithms involve the reception of external signals on a first layer nodes (which constitute the input layer), connected with innumerable nodes organized in multiple steps (hidden layers), which will give the processed data to the last nodes (output layer) for the final level result (*Figure 3.13*). Therefore, several internal input-hidden-output levels are hierarchically placed side by side in creating a large network with disproportionate amounts of neurons.

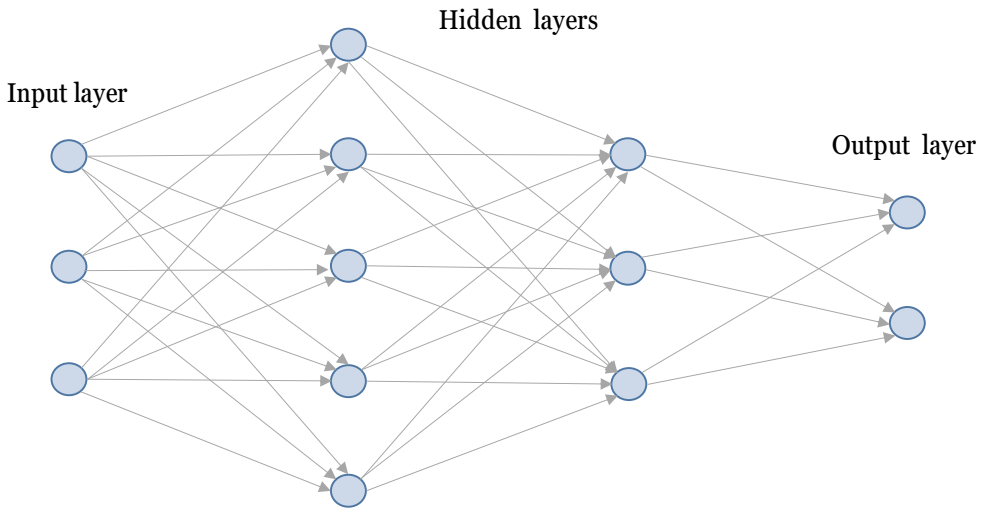


Figure 3.13. Simplified representation of a level of a neural network.
Source: author

Figure 3.14 shows a simple representation of the functioning of a neuron.

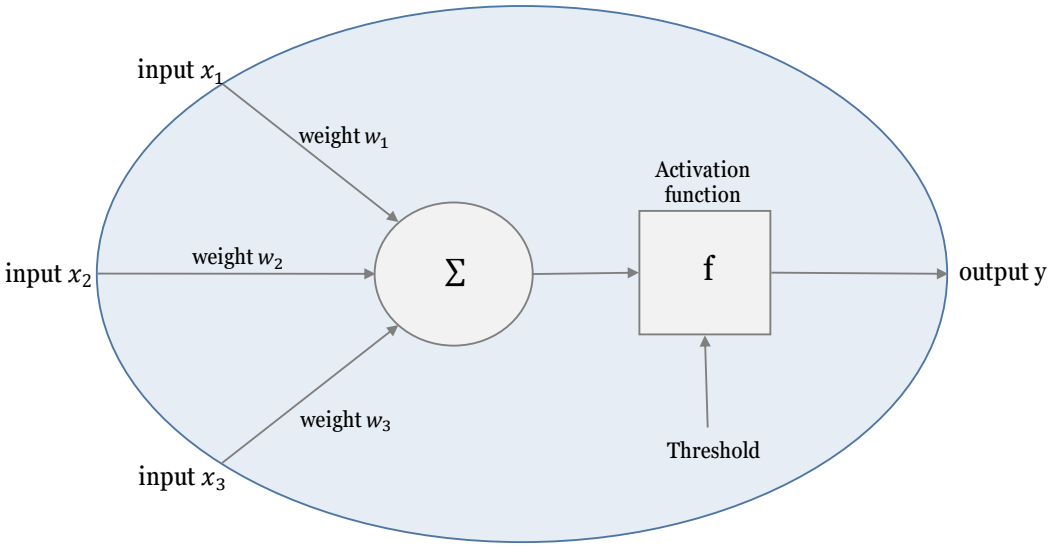


Figure 3.14. Simplified representation of a neuron.
Source: author

Each neuron receives *n inputs* (stimuli) x_1, x_2, ..., x_n on n links characterized by a certain value w_i, called *weight*, which represents the importance of a certain connection. Processing takes place in the unit Σ: it can be very complex but for example it can be considered that every stimulus is multiplied by its weight; then the sum of the results obtained is carried out and an *activation function* is applied and, if a certain *threshold* is exceeded, the neuron produces an *output*, which will be the stimulus of one or more following neurons.

Thanks to the weights of the connections, in this system preferential paths are delineated, that is those where the weights are greater, and the neurons have more possibilities to produce output and continue the chain. In particular, the function of the neurons of the input layer is to make "usable", through a transformation of the input data, the stimuli sent to the internal nodes. The inner layers then proceed to the actual processing of the data and provide the final layer the results of the processing that are here adapted and manipulated to be presentable and usable from the next level.

There are many types of neural networks, and the one just described is the *feedforward* type: neurons are grouped in layers (input, hidden, output), and each neuron in one layer has only directed connections to neurons of the next layer. Another widely used type is *recurrent*: a neuron can also influence itself (connected to itself). The choice of which type of neural network to use depends on the problem to be solved.

The most interesting characteristic of neural networks is their capability to familiarize with data by training, and then to be able to solve unknown data problems. The learning process through training, and then the application of what has been learned on data of which the result is not known, is called generalization.

Neural network learning models are very numerous and they depend a lot on why it was built. The learning models are grouped into three families: supervised, unsupervised and reinforced learning. In *supervised learning* training sets consists of input patterns with correct result, so the network can know his precise error once his result is given out and proceed with its improvement. In *unsupervised learning* models only input patterns are given and the network try to identify and classify them in similar groups, so the networks tries by itself to detect right results. In *reinforcement learning*, lastly, after a sequence of input, a value that defines when the answer is wrong is returned to the network, so that nodes adapt their characteristics according to the value they receive, in a reward/punishment system.

But what does it mean that the network "learns"? A neural network learns by changing itself, i.e. by developing new connections, deleting existing connections, changing weight, neurons thresholds or activation functions, creating new neurons or deleting existing ones. For example, in backpropagation algorithms (supervised), the network, once known its result errors, modify connections weights in a retroactive way so that the outputs increasingly converge towards the right values.

There are multiple applications of neural networks: speech recognition, calligraphy or artificial vision (reproduction of 3D environments from 2D images) are now a concrete reality thanks to neural networks. Moreover, a neural network can recognize the contents of a photo with extreme precision and to catalogue it directly without human aid, thousands of neurons analyse elements such as contrasts, pixels, or brightness. Another very important area is the control of product quality in industrial processes: with supervised learning the program observes various examples whose quality standard is excellent and learns which products are not suitable and which are defective. In this case, discard them, report them or even give the machine instructions to make them fit the standards. Many fields such as biology, physics, electronics, and many others, carry out advanced research on neural networks, both human and artificial; the same medicine uses this synergy to study cancer causes, the most effective treatments for each patient, and, more generally, to make more precise diagnoses. As far as Data Mining is concerned, neural networks are implemented for classification, optimization and forecasting models.

Chapter 4

CASE STUDY: EXPLOIT GROWTH SYNERGIES BY DATA ANALYTICS

The objective of this chapter is to apply Data Mining and Machine Learning methods on a real multi-business company, to identify business problems and focus on developing growth synergies. The author does not intend to show the results of the strategies proposed after the analysis, but to indicate effective methods for the exploitation of synergies, which serve as suggestions to decision-makers, subject to many more constraints than those considered in the chapter.

The first sub-chapter attempts to give an overview of the company, from the study of businesses, to the understanding of the data available for analysis. Furthermore, a first classification of customers is made, which allows us to identify business problems more precisely. Among the problems, we focus in particular on those concerning the failure to exploit growth synergies. The second sub-chapter, on the other hand, proposes some data-driven strategies and methods to improve the company's performance regarding the growth synergies between its businesses, through the implementation of some Machine Learning and Data Mining techniques discussed in the previous chapter.

4.1 Business understanding

Before starting any advanced analysis, you need to understand the business model you are facing and the data available to analyse it. Therefore, it is possible to proceed with initial data analysis to estimate performance and to detect any problems.

4.1.1 Business sectors and competition analysis

The company, which for rights reasons we will call "Alpha", is an Italian multi-business company, more precisely double-business, as it operates both in the supermarket business and in the plant nursery business. The company started as a garden centre and plant nursery, but in recent years it has tried to diversify into supermarket products, because they offer higher margins.

The company is therefore a multi-business that has carried out a related diversification, not so much for the affinity between the products, but because the supermarket business and the plant nursery business share the sales channel, the employees and the management. Furthermore, the distribution channel for incoming (suppliers) and outgoing (shipments of goods to customers) is common across the two businesses. According to another classification point of view, given the sharing of the stages of the value chain (sales channel), a horizontal diversification was implemented. It has been said in the previous chapters that related diversification allows a better exploitation of operative synergies, which are divided into efficiency synergies, characterized by cost subadditivities, and growth synergies, characterized by revenues super-additivities. As we will see in Par. 4.2.1, however, we will only have the availability of sales data, so we will concentrate on growth synergies.

In the theoretical chapters, we have also seen that a well-studied strategy of related diversification has a good chance of increasing sales of the two businesses thanks to the sharing of resources and management capabilities. The fact of owning two businesses, however, is more complex for management, which must integrate and manage two businesses with different products and needs, both temporal and space. Plant nursery in fact requires a certain flexibility given by the seasonality of the products, a more careful management of warehouse rotation and good management of spring peaks in demand. It also requires constant product care even during sale, and a difficult space management that depends a lot on the products coming from suppliers. The market survey, however, sees the plant nursery business very strong and recognized among customers. The goal of the company is to grow the supermarket business, which is newer and less known. But we will talk more about this in the following paragraphs.

Today the company has one sales point, in which it sells both supermarket products and plant nursery products (including plants, garden products and agricultural machinery). The internal organization of the products is, however, particular: in an entrance “A” Alpha sells only plant nursery products, in an entrance “B” sells supermarket products mixed with some plant nursery products. The choice of which

nursery products are placed on the supermarket side does not follow any precise logic. The rough layout of the store structure is shown in *Figure 4.1*.

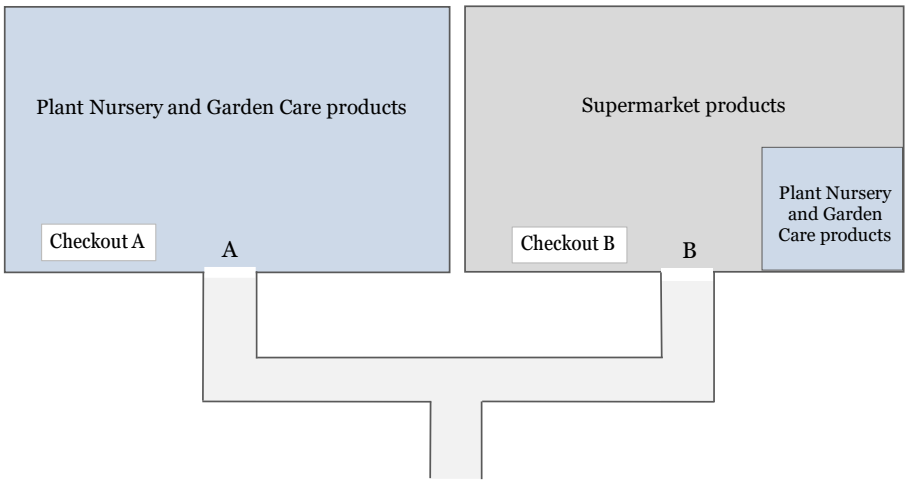


Figure 4.1. Simplified map of the business layout in the Alpha store.
Source: author

The performance of the two businesses, especially the supermarket, depend a lot on the proximity of competitors who sell the same products. As for the level of competitiveness of the territory regarding the businesses involved, refer to *Figure 4.2*.

Figure 4.2. Map indicating the distribution of competitors in the territory.
Source: Google Maps

As can be seen from the image, within a radius of 1km there are two supermarkets, which is fairly standard for an urban zone. Within 3 km, there is a shopping centre, 5 other supermarkets and two plant nurseries. Only further away there are 4 more nurseries and an agricultural machinery shop. So, at the supermarket level the

competition is very high, while as regards the plant nursery Alpha having to face a lower competitiveness, also because it is considered by the inhabitants as a reference point on this business.

Another important factor that characterizes the company is that it operates both in Business to Business and in Business to Customer, in other words it acts both as a supplier to other companies (especially for the plant nursery business), and as a sales point for final customers.

The goal of the company is to exploit growth synergies between the businesses on B2C market. In fact, it has diversified into the supermarket business precisely because the existence of the distribution and sales channels of the plant nursery business, seeing in the first a new growth opportunity. The company also has a data warehouse, which is really useful for undertaking data-driven strategies.

4.1.2 The available data

The company has a data warehouse. However, let's see only data available for the analysis: one-year sales data (2018) for B2C customers, customers data and products data. The star schema of the available data is shown in *Figure 4.3*.

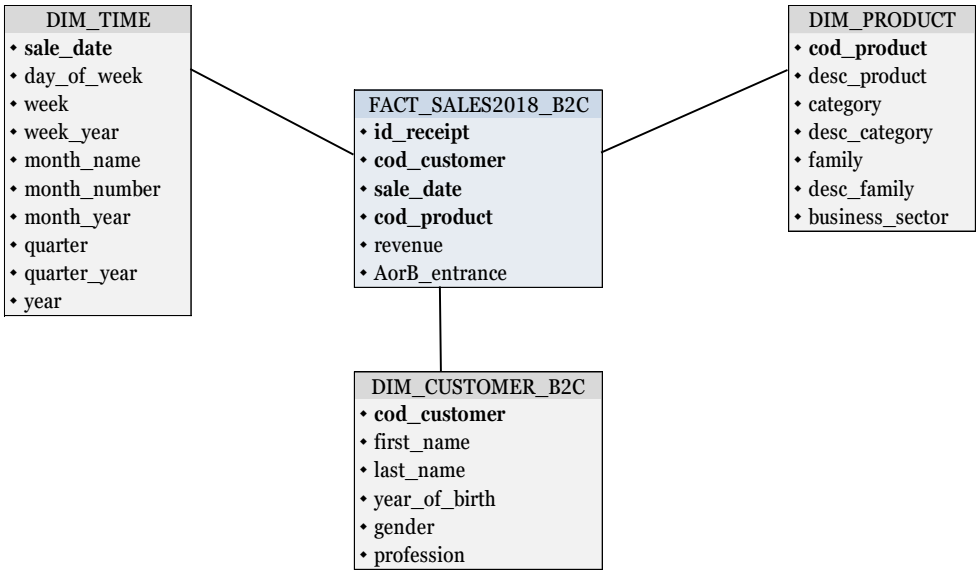


Figure 4.3. Star Schema of the available sales data of Alpha.
Source: author

Let's take a closer look at the attributes of each table:

- DIM_TIME

- **sale_date** [Date]: primary key in YYYY/MM/DD format. Ex. 2018/06/27
- day_of_week [String]: Ex. 'WED'
- week [Integer]: week of the year. Ex. 37
- week_year [String]: week of the year and year. Ex. '37_2016'
- month_name [String]: Ex. 'FEB'
- month_number [Integer]: Ex. 2
- month_year [String]: Ex. 'FEB_2017'
- quarter [String]: Ex. 'II'
- quarter_year [String]: Ex. 'II_2018'
- year [Integer]: Ex. 2018

- DIM_PRODUCT
  - **cod_product** [String]: primary surrogate key. Ex. '00023421'
  - desc_product [String]: Ex. 'Piantine ortive 27' (vegetable seedlings 27)
  - category [String]: Ex. 'A21'
  - desc_category [String]: Ex. 'Ortive primaverili' (vegetable seedlings)
  - family [String]:Ex. 'R43'
  - desc_family [String]: Ex. Piante da giardino (garden plants)
  - business_sector [String]: two values allowed: 'garden/plantnursery' or 'supermarket'.
- DIM_CUSTOMER_B2C
  - **cod_customer** [String]: primary surrogate key. Ex. 'c0034223567'
  - name [String]: Ex. 'John'
  - surname [String]: Ex. 'Smith'
  - year_of_birth [Integer]: Ex. 1965
  - gender [String]: two values allowed: 'M' (male) or 'F' (female)
  - profession [String]: Ex. 'artigiano' (artisan)

- FACT_SALES2018_B2C (2.693.061 records)
  - **id_receipt** [String]: primary surrogate key. Because one customer can buy the same product more than ones at different times in the same date. Ex. '0401019331749162018-06-1609.01.28'
  - **cod_customer** [String]: surrogate key. Ex. 'c0034223567'
  - **sale_date** [Date]: surrogate key in DD/MM/YYYY format. Ex. 27/06/2018
  - **cod_product** [String]: surrogate key. Ex.' 00023421'
  - revenue [float]: Ex. 12.50
  - AorB_entrance [String]: two values allowed: 'A' or 'B'.

Each line of the fact represents a line of a customer receipt for a certain amount of a product on a certain date.

It should also be underlined that if the OP value followed by a number appears in the item cod_customer in FACT_SALES2018_B2C, it means that the purchase was made by a customer without a loyalty card, and therefore not traceable. Purchases made in this way are obviously useless for analyses on customer behaviour.

AorB_entrance represents the entrance to which the purchase was made, in fact each section of the store has a separate cash point. Products purchased on one side must pass through the cash point on the same side. business_sector, on the other hand, indicates which business the product refers to. Since the entrance B offers both plant nursery and supermarket products, the two attributes do not have the same meaning. This last attribute has been inserted by the author based on product families because it is fundamental in subsequent analyses.

Since the available data is only the sales data, the best thing to do is to understand how to increase only growth synergies.

4.1.3 Classifying customer for historical analyses

Before looking in detail at the performance it is very useful to define the firsts customer classifications. To be honest the techniques used for this purpose are not part of the advanced analytics because they are not replicable on the future data but they fit well only on the current historical data. This is because in this sub-chapter we are only interested in analysing the historical performance of the business and not creating models for the future. Dividing customers on the basis of their 2018 buying behaviour opens the way for us to better understand the customers we are dealing with and the opportunities to grow sales. We also need to label customers and use different strategies depending on class membership, as well as division into groups to perform intra-class analysis.

Before moving on to the practical division of customers into groups, a brief excursion is made to discuss some of the techniques analysed for this purpose, so as to choose the most suitable technique starting from the distribution of the owned data. Three simple methods of dividing a multitude of objects into groups are presented: division by intervals, quantile division, and division by cumulative intervals. Before proceeding to the division, it is necessary to choose the classification value (ex. annual sale per customer, number of monthly receipts per customer, ...), extract the data of interest, sort them in ascending order and display them in a graph (in the following figures, for

example, a distribution of data that will be useful in practice is taken). Starting from the distribution of these data in *Figure 4.4*, we then choose the most intelligent or most

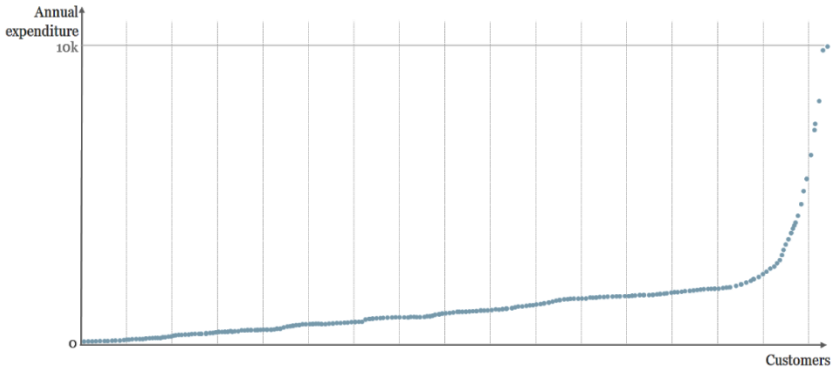


Figure 4.4. Ordered distribution of the sample data.
Source: author

useful class division for the objectives we want to achieve. For example, you are supposed to want to get 5 classes, based on annual expenditure values per customer.

- **Division by intervals (same range of class amplitude)**

Take the maximum and minimum between the values of the classification and create the class limits dividing the range by the number of classes.

This technique is appropriate when the distribution of values tends to a straight line. It can be seen from the *Figure 4.5* that the technique is not good for distributing sample data, because almost all customers are assigned class 5.

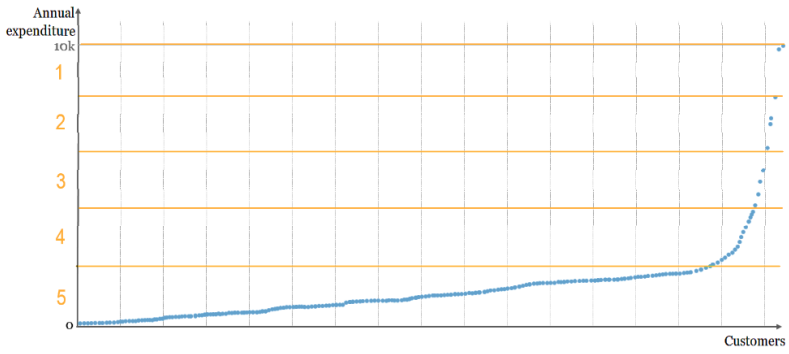


Figure 4.5. Division by intervals of sample data.
Source: author

❖ **Quantile division (same number of items per class)**

The total number of customers is divided into 5 parts and the corresponding class limits are calculated. In this way classes with the same number are obtained.

This technique is appropriate when the distribution of values tends to a Gaussian or a straight line.

The technique is not good for class division of sample data: the width of class 1 is enormously greater than the others (*Figure 4.6*).

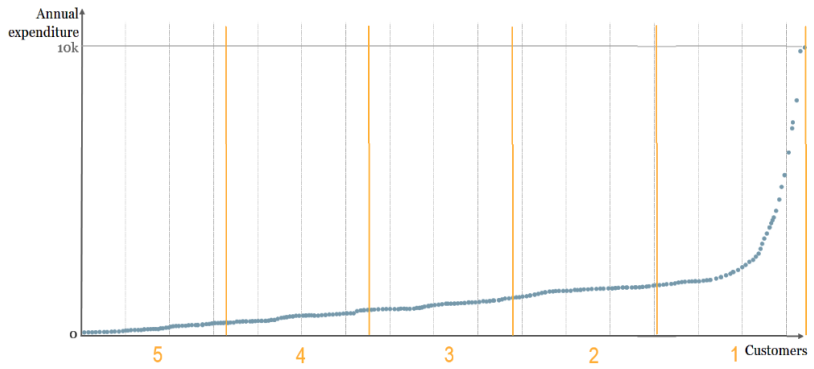


Figure 4.6. Quantile division of sample data.
Source: author

❖ **Division by cumulative intervals (equal participation of value per class)**

The classes are divided so that each gives the same contribution of value over the total. Therefore, the sum of the values of customers belonging to a class is constant for all classes. Both the class thresholds and the number of customers change from class to class. Operation is not so complex: after ordering the data, a cumulative function of values is created: each index is assigned the previous value added to its value. Then the cumulative function is divided according to the first method proposed. At this point the class numbers found are transposed in the starting function. The final result is shown in *Figure 4.7*.

This technique is appropriate for strongly unbalanced and asymmetrical distributions. The technique seems the most appropriate for distributing sample data, also to treat classes with a same importance.

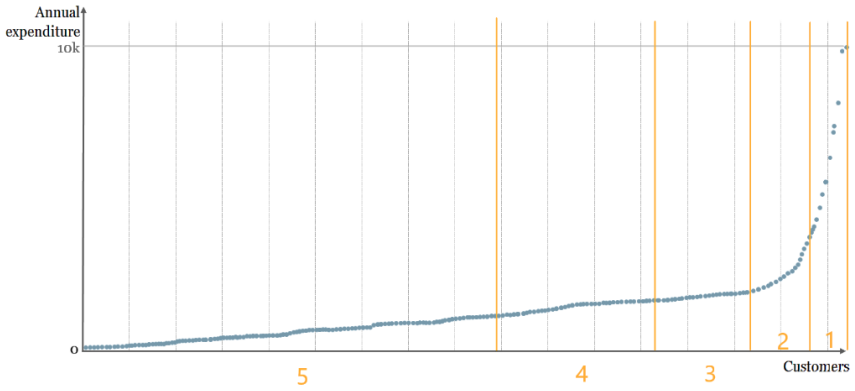


Figure 4.7. Division by cumulative intervals of sample data.
Source: author

For the practical purpose of the business case, customers are divided through two classifications, presented below.

a) Customer classification by revenue

Customers are divided into 5 classes based on the value of purchases made during the year (2018), from class 5 (the lowest) to class 1 (the highest). Since the distribution is of the type in *Figure 4.8*, it seems appropriate to use division by cumulative intervals method.

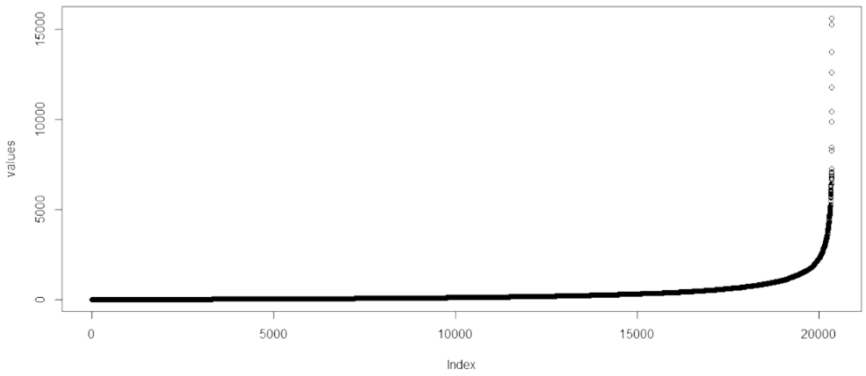


Figure 4.8. Ordered distribution of expenditure by all Alpha's customers.
Source: author (by the following code)

❖ **Code: customer classification by revenue**

Query for csv file:

```
SELECT f.cod_customer,
       sum(f.revenue)as Tot_revenue
FROM FACT_SALES2018_B2C f,
     DIM_CUSTOMER_B2C c
WHERE f.cod_customer = c.cod_customer
AND f.cod_customer not like '%OP%'
GROUP BY f.cod_customer
ORDER BY sum(f.revenue)
```

Script R language (Source: author):

```
#-------------input----------------------------------
#input format: <cod_customer><tot_revenue>
DATA <- read.csv (file="path", header=TRUE, sep=",")
values <- DATA$Tot_revenue
values <- sort(values)
plot(values)
perc_top <- 0.2
num_class <- 5
unit <- 0.01
#----------------------------------------------------
num_row <- nrow(DATA)
class <- c(1:num_class)
class <- rev(class)
global_value <- sum(values)
value_partition_inf <- global_value*perc_top
value_partition_sup <- global_value*(1-perc_top)
cum <- cumsum(values)
```

```
  plot(cum)
  min_value<-NA
  max_value<-NA
  min_value[1]<-values[1]
  max_value[num_class]<-999999999999

  index <- 1
  while ((cum[index]<value_partition_inf) || (values[index] == values[index+1])) {
        index=index+1}

  min_value[2]<-values[index-1] + unit
  max_value[1]<-values[index-1]

  index <- length(cum)
  while ((cum[index] > value_partition_sup) || (values[index] == values[index-1])) {
          index<-index-1}

  min_value[num_classi] <- values[index] + unit
  max_value[num_classi-1] <- values[index]

  values_rim <- NA
  j <- 1
  for(i in 1:length(values)){
    if(values[i] > max_value[1] && values[i] < min_value[num_class]){
       values_rim[j] <- values[i]
       j <- j+1}}

  global_value_rim <- sum(values_rim)
  values_rim <- sort(values_rim)
  cum_rim <- cumsum(values_rim)
  value_partition <- global_value_rim/(num_class-2)

  index_partition <- NA
  i <- 1
  inc_partition <- value_partition
  for (index in 1:(length(values_rim) - 1)){
    if((cum_rim[index] >= inc_partition) && (values_rim[index] !=
values_rim[index+1])){
        index_partition[i] <- index
        i <- i+1
        inc_partition <- inc_partition + value_partition}}
  i <- 1
  for(j in 3:(num_class - 1)){
      min_value[j] <- values_rim[index_partition[i]] + unit
      i <- i+1}
  i <- 1
  for(j in 2:(num_class - 2) ){
    max_value[j] <- values_rim[index_partition[i]]
    i <- i+1}
  #-------------output-------------------------------------------
  DIM_CLASS <- data.frame(class,min_value,max_value)
  write.csv(DIM_CLASS, file = "path", row.names = FALSE)
```

The result is shown in *Table 4.9*.

class_revenue	min_value (€)	max_value (€)
5	0	282.77
4	282.78	623.02
3	623.03	1163.76
2	1163.77	2259.31
1	2259.32	9999999999

Table 4.9. Class range of the classification by revenue.
Source: author (by code)

Figure 4.12 shows the result achieved more clearly.

❖ Customer classification by purchase frequency

Customers are divided into 5 classes based on the number of distinct times they purchase something during the year (2018), from class 5 (the lowest) to class 1 (the highest). Since the distribution is of the type in *Figure 4.10*, it seems appropriate to use division by cumulative intervals method.

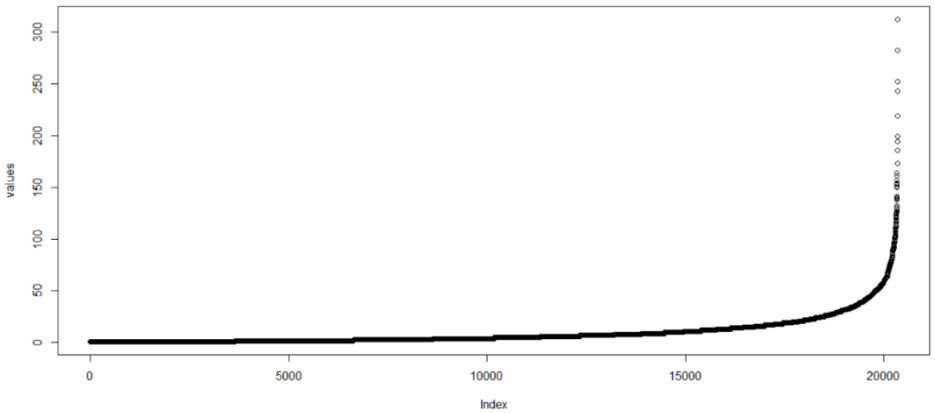


Figure 4.10. Sorted distribution of number of receipts by all Alpha's customer.
Source: author (by the following code)

❖ Code: customer classification by purchase frequency

Query for CSV file:

```
SELECT f.cod_customer,
       COUNT(DISTINCT id_receipt)as frequency
FROM FACT_SALES2018_B2C f
WHERE cod_customer not like '%OP%'
GROUP BY f.cod_customer
```

Script R language (Source: author):

```
#-------------input-----------------------------------------
#input format: >cod_customer><frequency>
DATA <- read.csv(file="path", header=TRUE, sep=",")
values <- DATA$frequency
values <- sort(values)
plot(values)
perc_top <- 0.2
num_class <- 5
unit <- 1
```

```
#-----------------------------------------------------------
num_row <- nrow(DATA)
class <- c(1:num_class)
class <- rev(class)
global_value <- sum(values)
value_partition_inf <- global_value*perc_top
value_partition_sup <- global_value*(1-perc_top)
cum <- cumsum(values)
plot(cum)
min_value<-NA
max_value<-NA
min_value[1] <- values[1]
max_value[num_class] <- 999999999999

index <- 1
while ((cum[index] < value_partition_inf) || (values[index] == values[index + 1])) {
        index = index + 1}
min_value[2] <- values[index - 1] + unit
max_value[1] <- values[index - 1]

index<-length(cum)
while ((cum[index] > value_partition_sup) || (values[index] == values[index - 1])) {
       index<-index - 1}

min_value[num_class] <- values[index] + unit
max_value[num_class-1] <- values[index]

values_rim <-NA
j <- 1
for(i in 1:length(values)){
   if(values[i] > max_value[1] && values[i] < min_value[num_class]){
      values_rim[j] <- values[i]
      j <- j + 1 }}

global_value_rim <- sum(values_rim)
values_rim <- sort(values_rim)
cum_rim <- cumsum(values_rim)
value_partition <- global_value_rim/(num_class-2)

index_partition <- NA
i <- 1
inc_partition <- value_partition
for (index in 1:(length(values_rim) - 1)){
    if((cum_rim[index]>=inc_partition) && (values_rim[index]!= values_rim[index+1])){
        index_partition[i] <- index
        i <- i + 1
    inc_partition <- inc_partition + value_partition}}

i <- 1
for(j in 3:(num_class - 1)){
    min_value[j] <- values_rim[index_partition[i]] + unit
    i <- i + 1}

i  <- 1
for(j in 2:(num_class - 2) ){
    max_value[j] <- values_rim[index_partition[i]]
    i <- i+1}
#-------------output-----------------------------------------
DIM_CLASS <- data.frame(class,min_value,max_value)
write.csv(DIM_CLASS, file = "path", row.names=FALSE)
```

The result is shown in *Table 4.10*:

class_frequency	min_value	max_value
5	1	8
4	9	17
3	18	29
2	30	51
1	52	999999999

Table 4.10. Class range of the classification by purchase frequency.
Source: author (by code)

Figure 4.12 shows the result achieved more clearly (the inaccuracies in the percentages are due to the fact that the frequency is a discrete and not continuous value. Thus, customers with the same frequency value cannot be separated into different classes).

It's comfortable now to join the two classifications in a table formed by cod_customer and the two classes to which it belongs. The result is therefore a table that integrates the star schema, as shown in *Figure 4.11*. Classes are not inserted in the table DIM_CUSTOMER_B2C because these classes refer only to the purchasing behaviour of the year 2018, while this dimensional table is time independent.

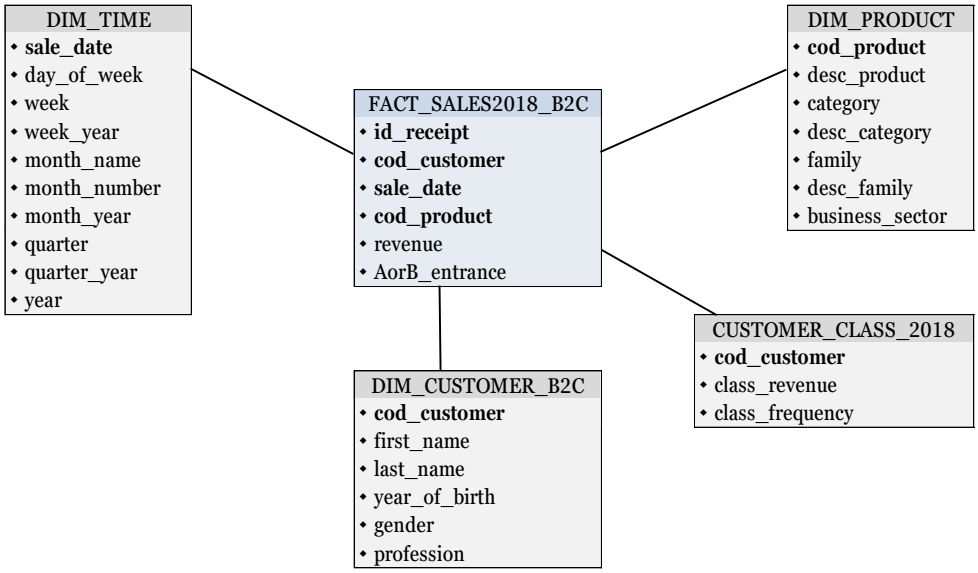


Figure 4.10. Star schema updated after classifications.

❖ Code: creating the classification table

Query:

```
SELECT TAB.cod_customer,
       CR.class_revenue,
       CF.class_frequency
FROM(
    SELECT cod_customer ,
```

```
            SUM(revenue) tot_revenue,
            COUNT(DISTINCT id_sale) frequency
    FROM FACT_SALES2018_B2C f
    WHERE cod_customer NOT LIKE '%OP%'
    GROUP BY cod_customer ) TAB,
    CLASS_CUSTOMER_REVENUE CR,
    CLASS_CUSTOMER_FREQUENCY CF
WHERE TAB.tot_revenue BETWEEN CR.min_value AND CR.max_value
AND TAB.frequency BETWEEN CF.min_value AND CF.max_value
```

4.1.4 Summary statistics: problems

With the data we have, we can now make an analysis of the business performance of Alpha. It would not make sense to use advanced analysis techniques without having first performed descriptive analyses of the business. After all, how do you run if you haven't learned to walk before? Furthermore, while studying performance, we identify the problems that Alpha has in carrying out its business.

As show in *Figure 4.11*, Alpha had a turnover of 12.19 € Million thanks to purchases of about 20350 people. Customers therefore spent an average of 599€ /year in the store.

Moreover, we see that *only 55% of revenues come from registered customers*, that is, those who bought using the loyalty card. This is certainly too low and means that the margin of error on subsequent analyses is quite high, for two reasons: the first is that almost 45% of revenues comes from a source that cannot be deepened and indeed introduces wide variability to the generalization on customers purchasing behaviour by purchasing behaviour of registered customers. The second is that the registered customers could have made purchases without the loyalty card, so that their buying behaviour could be "dirty".

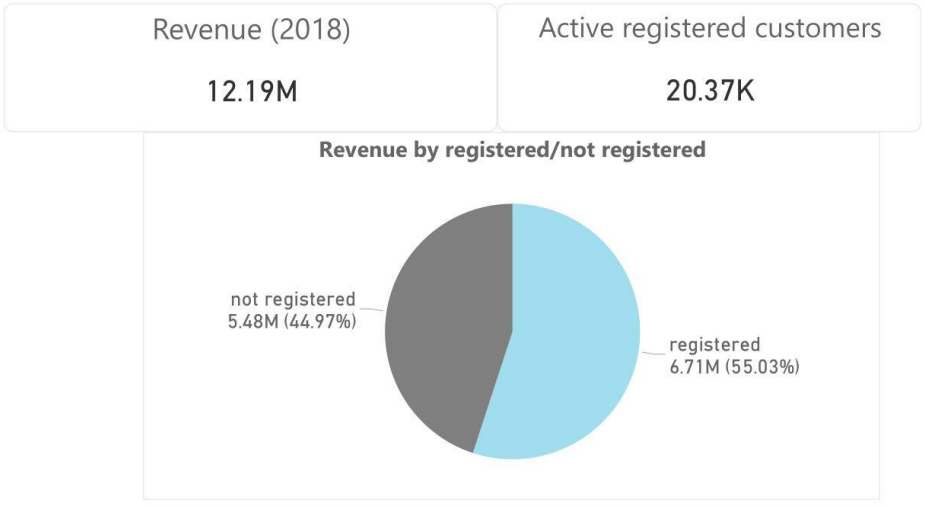


Figure 4.11. Revenue and number of customers.

Another big problem that the data show us (*Figure 4.12*) derive from classes 5 of the newly created classifications: 13.9K registered customers out of 20.35K active have bought for only 20% of the total receipts of registered customers (and these customers come to buy only 1 to 8 times a year). Furthermore, 14.3K registered customers on 20.35K contribute only 20% to the company's revenues deriving from loyalty cards.

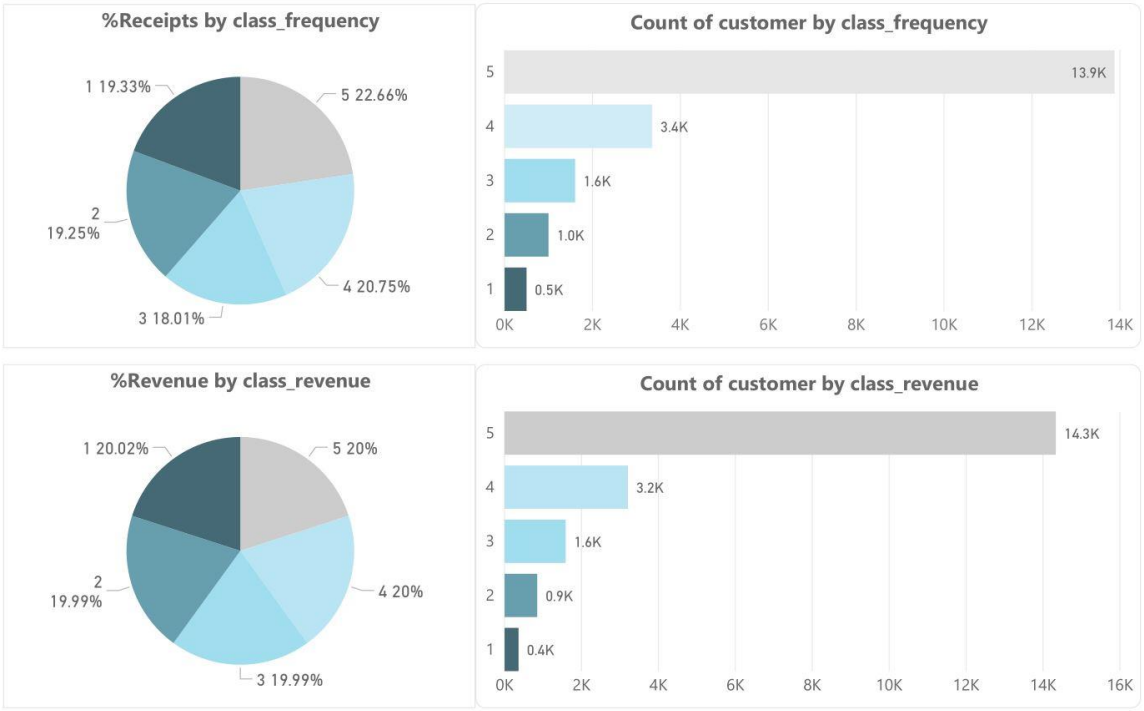


Figure 4.12. Revenues/Receipts and number of customers by classes.

This data is absurd: the problem is enormous. *Many customers come to buy and do not return much*. Why? Certainly, this is an aspect to be improved, that is to increase the customer loyalty through marketing strategies in order to make it come back more often, so also revenues benefit from it. In fact, what these data show us is that there is a huge growth potential, given that the number of customers reached is a lot, but they come few times to buy.

Figure 4.13 show the revenues by business sector. From the difference between the pie chart of registered customers and that of all customers, it is clear how customers who prefer the supermarket register more. Why? Probably the discounts with the loyalty card are better at the supermarket and almost non-existent in the plant nursery business. Then, looking at the revenue trend for business, we can see how the nursery grows a lot in the spring months, however, the increase in sales in the nursery does not translate into an increase in supermarket sales. The same happens to parts reversed in the last months of the year. So, *the two businesses seem to have a quite distinct trends, meaning they affect very little each other*. This is precisely a problem of failed

exploitation of synergies: the huge pool of customers who go to buy plants in April and May can be a huge opportunity for the supermarket business. In order for the synergies to be exploited to the full, we must try to grow on the supermarket side in the spring months (growing the nursery side in the winter months is difficult for botanical reasons, plants dry out). It is also recalled that the nursery is also subject to less competitiveness in the territory, and for this it must try to drag the supermarket, a business subject to stronger competition.

Figure 4.13. Revenues by business sector (the nursery in green, the supermarket in red).

We also see that, considering the two classifications made previously, the less the customers spend and the less the customers come frequently, the more their reference business is the plant nursery (*Figure 4.14*). *Alfa is therefore seen by most customers as a plant nursery and not a supermarket.* In addition, customers presumably go to the nursery in the spring months, buy, and occasionally return. From this graph we can also see, vice versa, that the supermarket business is populated by habitual customers and that they spend a lot (in fact, remember that in class 1, for example, there are only 500 clients that make up about 20% of Alpha's revenues), which means that a significant goal is *to make customers perceive Alpha also as a supermarket, make them habitual* in the purchase of its products, as the business supermarket favours the loyalty of the customer.

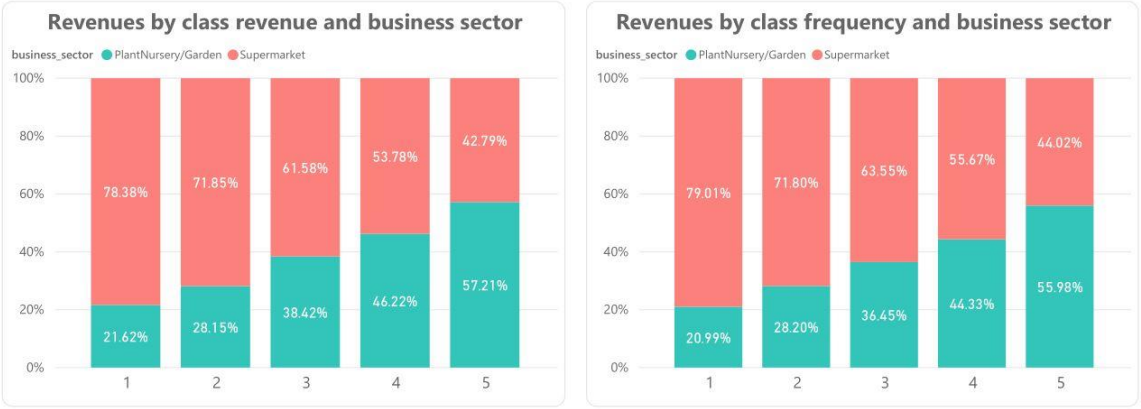


Figure 4.14. Revenues by classes divided by business sector.

For the next point, we want to understand how many customers buy both from A and B, how many only from "A" and how many only from "B". For this purpose, it is necessary to create a new table in which each customer has made receipts in "A", "B", or both on a given date.

❖ **Code: creating 'single or double entrance' table**

Query:

```
CREATE TABLE CUSTOMERS_PURCHASE_PLACES AS
  (SELECT DISTINCT cod_customer,
                   sale_date,
                   CASE WHEN purchase_places = 2 THEN 'A+B'
                      ELSE CASE WHEN purchase_places=1 AND AorB_entrance='B' THEN 'B'
                      ELSE  'A' END  END AS purchase_places
   FROM(
      SELECT cod_customer,
             sale_date,
             AorB_entrance,
             COUNT(DISTINCT AorB_entrance) OVER (PARTITION BY cod_customer,
                                                 sale_date) AS purchase_places
      FROM FACT_SALES2018_B2C
      WHERE cod_customer NOT LIKE '%OP%'
      GROUP BY cod_customer, sale_date, AorB_entrance )TAB
```

Analysing the data of the new table (*Figure 4.15*), we see that 11.18% of customers, when they buy from Alpha, buy only from "A", 75.82% buy only from "B" (which makes up 87.88% of revenues) and only "13%" buy from both. Attention: we remind you that receipts "A" and "B" have different cash registers and therefore if the customer wants to buy both "A" and "B", he must make two separate receipts.

From these data we can calculate the probability that a customer, once entered on one side, also enters the other. The calculation is the following:

Probability that the customer also goes to "B" since he wants to buy something from "A": $\frac{P[(A+B)]}{P[(A+B)]+P[A]} = \frac{0.13}{0.13+0.1118} = 0.5242 = 52.42\%$

Probability that the customer also goes to "A" since he wants to buy something from "B": $\frac{P[(A+B)]}{P[(A+B)]+P[B]} = \frac{0.13}{0.13+0.7582} = 0.1464 = 14.64\%$

Figure 4.15. Revenues by entrance and number of customers by entrance in the same date.

Therefore, at 52.42% a customer who buys from "A" will buy something from "B", while only 14.64% of customers who buy from "B" will go to buy something from "A". The first fact is significant that half of people, once they buy plants or garden products, take the opportunity to go to the supermarket. In this case the synergies between the businesses are positive. The second data is very low and the reason could be the following: since in "B" there are the products of the two businesses, customers do not need to go even to "A" because plant nursery products, even if to a limited extent, are also present in "B". This means that the nursery products in "A" are unfavourable to buy, even if they were of better quality. In this case *the location between the businesses products greatly disadvantages the plant nursery business*, which is cut off for most of its products, cannibalized from the nursery products at the entrance "B". It should be remembered for greater clarity that the two businesses contribute almost equal to the revenues of Alpha, but only about 12% derives from "A".

4.2 Improving performance through synergies

Now it's time to propose strategies based on advanced analysis to be able to increase the synergies between the two businesses, of which we have seen problems a little while ago. **Among the growth synergies that we are going to study, however, we will refer only to those that can be developed with Joint Market Penetration, that is through existing products in existing markets (Par. 2.3.1). The point of the store's layout will also be deepened** (i.e. the management of the two entrances), **and the moment in which the strategies must be implemented** (timing for growth synergies).

4.2.1 Customer clustering

Through clustering, we are going to divide customers into groups (clusters) based on their buying behaviour. For this purpose, reference is made to the product categories that appear at least once in FACT_SALES2018_B2C. From here on we only consider registered customers, which can be associated with a "buying behaviour".

In particular, in input to the clustering algorithm we give a table formed by a cod_customer column and a column for each product category. Inside the cells the revenue values that each customer has for the category. A draft of the table is shown in *Figure 4.16*.

cod_customer	category1	category2	category3	category4	category5	category6	category7	...
cod_customerA	A1 (€)	A2 (€)	A3(€)	A4(€)	A5(€)	A6(€)	A7(€)	...
cod_customerB	B1 (€)	B2 (€)	B3(€)	B4(€)	B5(€)	B6(€)	B7(€)	...
cod_customerC	C1 (€)	C2 (€)	C3 (€)	C4 (€)	C5 (€)	C6 (€)	C7 (€)	...
cod_customerD	D1 (€)	D2 (€)	D3 (€)	D4 (€)	D5 (€)	D6 (€)	D7 (€)	...
...	...	...	...	...	...	...	...	...

Figure 4.16. Clustering input table schema.

The real result is a table 20345x263.

❖ **Code: creation of the clustering input table**

Query:

```
SELECT cod_customer,
       desc_category,
       SUM(revenue) as expenditure
FROM FACT_SALES2018_B2C f,
     DIM_PRODUCT p
WHERE f.cod_product = p.cod_product
AND cod_customer not like '%OP%'
GROUP BY cod_customer, desc_category
```

The rows and columns of the table are then overturned using the excel pivot function.

Once the table is obtained, before giving it as input to the clustering algorithms it must be normalized. The normalization formula used is the min-max normalization per column. With normalization we make the categories that in general have higher sales do not weigh more than the others in the calculation of the distances between the points of the algorithm.

❖ **Code: normalization of the clustering input table**

Script R language (Source: author):

```
library(tidyverse)  # data manipulation
library(cluster)    # clustering algorithms
library(factoextra) # clustering algorithms & visualization
library(dbscan)
#--------------------input ---------------------------------
DATA <- read.csv(file="path", header=TRUE, sep=",")
```

```
DATA[is.na(DATA)] <- 0
DATA[DATA<0] <- 0         #replace negative values with 0

DATASET <- DATA [2:ncol(DATA)]      #delete customer column
#--------------------NORMALIZATION MIN-MAX ------------------
DATASET_NORM <- DATASET
for(i in 1:ncol(DATASET)){
    xmin <- min(DATASET[,i])
    xmax <- max(DATASET[,i])
    for(j in 1:nrow(DATASET)){
         DATASET_NORM[j,i] <- ((DATASET[j,i]-xmin)/(xmax-xmin))}}

DATASET_NORM[is.na(DATASET_NORM)] <- 0  #replace Nan with 0
DATASET_NORMALIZED <- data.frame(DATA$cod_customer, DATASET_NORM)
#--------------------output-----------------------------------
write.csv(DATASET_NORMALIZED, file = "path", row.names=FALSE)
```

The table is now ready to be fed to clustering algorithms. Given that expenditure values > 0 are very scattered in the table, because most customers rarely come and spend little, the points in the multidimensional space are very concentrated near 0. For this reason, they are applied three DBSCAN cycles in which in each step only two clusters are consciously created: a dense cluster and an outlier cluster. The points of the latter are input to the next DBSCAN. From these passages, 3 clusters of decreasing density are created and a cluster outlier from the last cycle. The latter cluster is given to k-means, which offers a further 3 clusters. The result is therefore 6 clusters, which, to avoid confusion with the classes, are called with letters of the alphabet.

❖ **Code: customer clustering by purchased categories**

Script R language: first DBSCAN cycle (Source: author from code [16]):

```
library(tidyverse)  # data manipulation
library(cluster)    # clustering algorithms
library(factoextra) # clustering algorithms & visualization
library(dbscan)

DATA <- read.csv(file="path", header=TRUE, sep=",")
DATA[is.na(DATA)] <- 0
DATA[DATA < 0] <- 0
DATASET <- DATA [2:ncol(DATA)]

#dbscan::kNNdistplot(DATASET, k =  30) #advice on eps - K nearest neighbour

db <- dbscan(DATASET, eps=0.008, minPts =30, weights = NULL, borderPoints = TRUE)
db
hullplot(DATASET, db)       ## plot clusters , usa le due colonne pilota del dbscan

DATA$cluster <-  NA
DATA$cluster <- db$cluster

write.csv(DATA, file = "path", row.names=FALSE)
```

The result (from R Studio editor):

```
> db <- dbscan(DATASET, eps=0.008, minPts =30, weights = NULL, borderPoints = TRUE)
> db
DBSCAN clustering for 20345 objects.
Parameters: eps = 0.008, minPts = 30
The clustering contains 1 cluster(s) and 19167 noise points.

    0     1
19167  1178
```

Cluster 1 is called ‘cluster A’
Cluster 0 is given as input to the second cycle

Script R language: second DBSCAN cycle (Source: author from code [16]):

```
library(tidyverse)
library(cluster)
library(factoextra)
library(dbscan)

DATA <- read.csv(file="path", header=TRUE, sep=",")
DATA[is.na(DATA)] <- 0
DATA[DATA < 0] <- 0
DATA2 <-  subset(DATA, cluster != 1)
DATASET <- DATA2 [2:(ncol(DATA) - 1)]

db <- dbscan(DATASET, eps=0.05, minPts =40, weights = NULL, borderPoints = TRUE)
db
hullplot(DATASET, db)

DATA2$cluster <-  NA
DATA2$cluster <- db$cluster

write.csv(DATA, file = "path", row.names=FALSE)
```

The result (from R Studio editor):

```
> db <- dbscan(DATASET, eps=0.05, minPts =40, weights = NULL, borderPoints = TRUE)
> db
DBSCAN clustering for 19167 objects.
Parameters: eps = 0.05, minPts = 40
The clustering contains 1 cluster(s) and 13105 noise points.

    0     1
13105  6062
```

Cluster 1 is called ‘cluster B’
Cluster 0 is given as input to the third cycle

Script R language: third DBSCAN cycle (Source: author from code [16]):

```
library(tidyverse)  # data manipulation
library(cluster)    # clustering algorithms
library(factoextra) # clustering algorithms & visualization
library(dbscan)

DATA <- read.csv(file="path", header=TRUE, sep=",")
DATA[is.na(DATA)] <- 0
DATA[DATA < 0] <- 0
DATA2 <-  subset(DATA, cluster != 1)
DATASET <- DATA2 [2:(ncol(DATA)-1)]
db <- dbscan(DATASET, eps=0.15, minPts =50, weights = NULL, borderPoints = TRUE)
db
hullplot(DATASET, db)

DATA2$cluster <-  NA
DATA2$cluster <- db$cluster

write.csv(DATA2, file = "path", row.names=FALSE)
```

The result (from R Studio editor):

```
> db <- dbscan(DATASET, eps=0.15, minPts =50, weights = NULL, borderPoints = TRUE)
> db
DBSCAN clustering for 13105 objects.
Parameters: eps = 0.15, minPts = 50
The clustering contains 1 cluster(s) and 6345 noise points.

   0    1
6345 6760
```

Cluster 1 is called 'cluster C'
Cluster 0 is given as input to k-means

Script R language: k-means (Source: author from code [17]) :

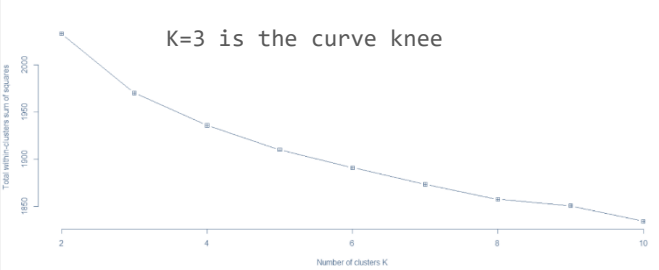
```
library(tidyverse)  # data manipulation
library(cluster)    # clustering algorithms
library(factoextra) # clustering algorithms & visualization
library(dplyr)

DATA <- read.csv(file="path", header=TRUE, sep=",")
DATA[is.na(DATA)] <- 0
DATA[DATA < 0] <- 0
DATA2 <-  subset(DATA, cluster != 1)
DATASET <- DATA2 [2:(ncol(DATA)-1)]

k2 <- kmeans(DATASET, centers = 2, nstart = 5)
k3 <- kmeans(DATASET, centers = 3, nstart = 5)
k4 <- kmeans(DATASET, centers = 4, nstart = 5)
k5 <- kmeans(DATASET, centers = 5, nstart = 5)
k6 <- kmeans(DATASET, centers = 6, nstart = 5)
k7 <- kmeans(DATASET, centers = 7, nstart = 5)
k8 <- kmeans(DATASET, centers = 8, nstart = 5)
k9 <- kmeans(DATASET, centers = 9, nstart = 5)
k10 <- kmeans(DATASET, centers = 10, nstart = 5)

set.seed(123)
# function to compute total within-cluster sum of square
wss <- function(k) {
kmeans(DATASET, k, nstart = 10 )$tot.withinss }

# Compute and plot wss for k = 1 to k = 10
k.values <- 2:10
# extract wss
wss_values <- map_dbl(k.values, wss)
plot(k.values, wss_values, type="b", pch = 12, frame = FALSE,
     xlab="Number of  clusters K", ylab="Total within-clusters sum of squares")
```




```
DATA2$cluster <- k3$cluster
write.csv(DATA2, file = "path", row.names=FALSE)
```

The result (from R Studio editor):

```
> k3$size
[1] 4871  787  587
```

respectively called 'cluster D', 'cluster E', cluster 'F'

From the cluster tables created it is easy to merge them into a single table: the new cluster table is inserted in the star scheme, in order to connect to the overall analysis (*Figure 4.17*).

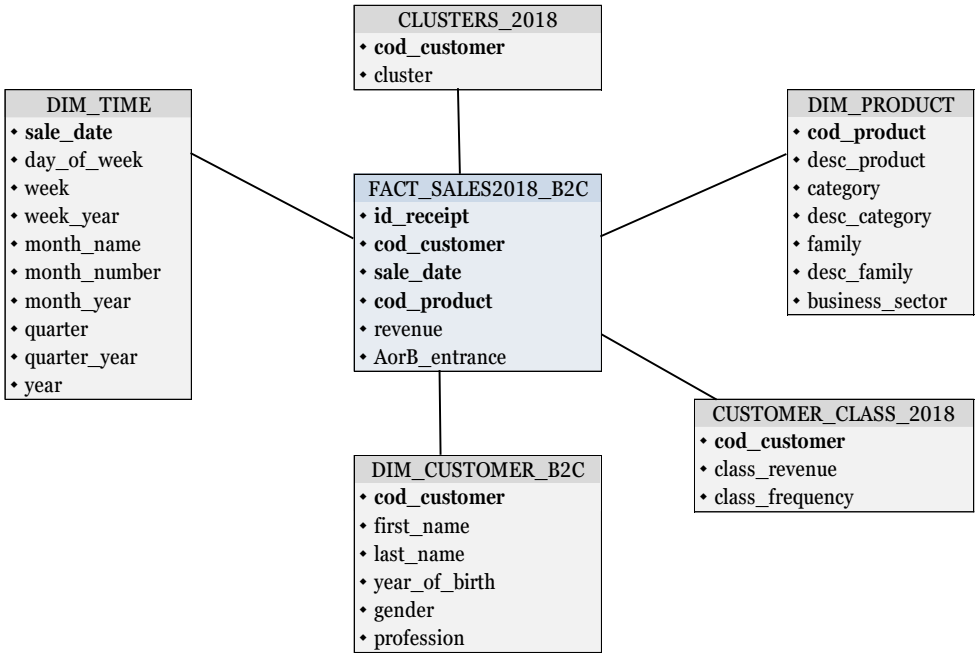


Figure 4.17. Star schema updated after clustering.

Numerically, each cluster contains the number of customers shown in *Figure 4.18*.

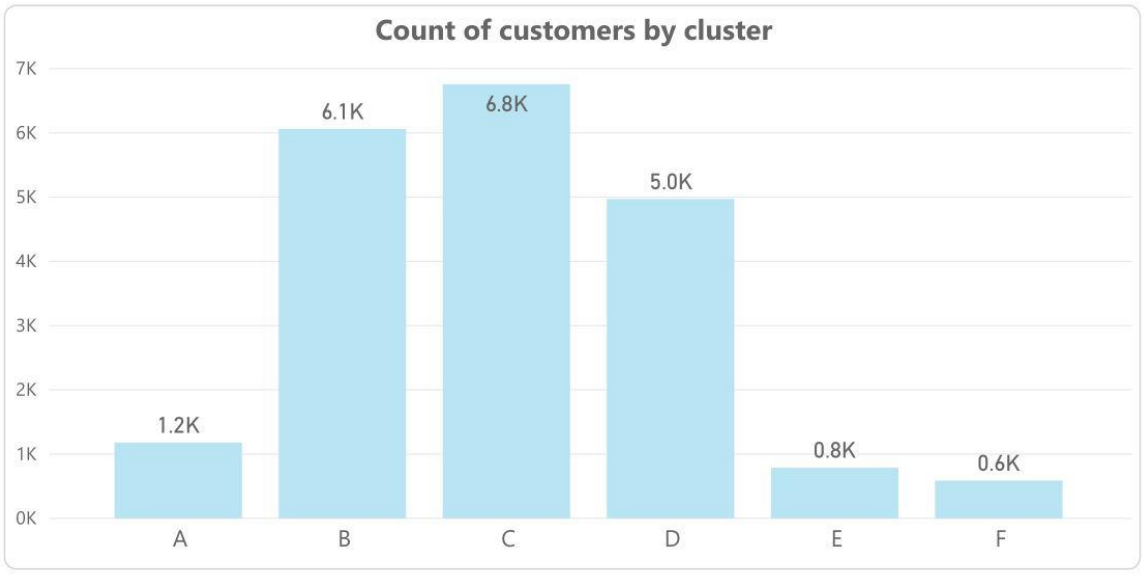


Figure 4.18. Number of customers per cluster.

Let's see in detail the characteristics of the found clusters: for each cluster there are three analyses: two on the number of customers belonging to the two types of class previously created, and one on the top 10 categories for revenues (*Figure 4.19*).

In cluster A, there are only customers who have come to buy very little from Alpha (1 to 8 times a year). As previously assumed, in light of the most purchased product categories, these customers buy plant from the nursery and few more.

Cluster B is very similar to cluster A, but much more numerous. In proportion the cluster B customers buy something extra from the supermarket.

In cluster C, which is also very numerous, some class 4 customers are starting to appear both in terms of revenues and in terms of frequency of purchase. However, the trend of the categories does not change much, except that, compared to cluster B, the supermarket business begins to grow.

Cluster D contains a wide variety of customer classes. It is therefore a very heterogeneous cluster in terms of class, and homogeneous in terms of business involved. Here, in fact, the supermarket is definitely preferred to the nursery business.

Cluster E is very similar to cluster D in terms of partitioning customers into classes, but slightly shifted to better classes. However, it is diametrically opposite in the purchase categories. In this cluster all those customers who strongly believe in the plant nursery of Alpha are present.

The F cluster is impressive. It is made up of customers who come more and spend more. To confirm what has been said so far, the data show that habitual customers rely heavily on the supermarket.

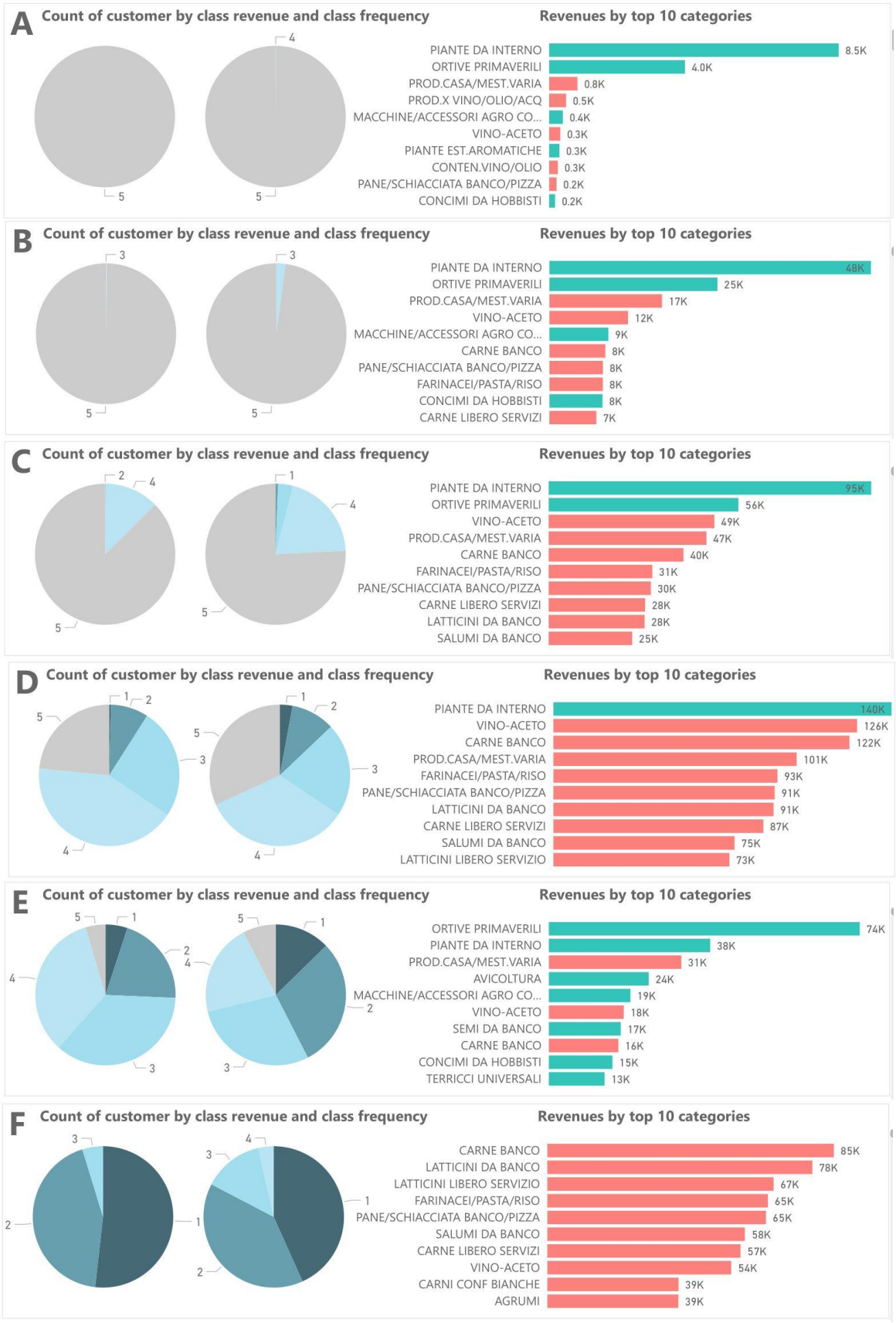


Figure 4.19. Cluster analysis by classes, and revenues by categories.

4.2.2 Change store layout

Let us now make an argument, starting from what we have said so far about the division of the store into two entrances, now that we have clusters (*Figure 4.20*). In clusters A and B, purchases in sector “A” increase dramatically. This confirms that customers who spend little and come little see Alpha much more as a plant nursery than a supermarket. As customers tend to spend more and come more often, they keep going to "B", except for a slight increase in “A” in cluster E. This is a sign that the "B" entry is definitely stronger and more profitable. We also note that clusters C and E are the two most synergistic clusters, that is, the customers within them tend to enter more into the two entrances, and consequently also to buy from all the two businesses. For this reason, these will be the fundamental clusters in the strategies for growth with joint market penetration. We have seen, therefore, that the clusters A, B and C substantially go in "A" more than the average and come little and spend little. And over 14,000 customers are in this condition (members of cluster A only with a probability of 3% go in "B" after having bought from "A"). This huge pool of customers must be brought to “B”, to buy more from supermarket. Conversely, the customers of the clusters in which much is spent, D and F in particular, buy practically only from “B”: in this case the nursery products of "A" are cannibalized from those of "B".

To these customers, who are almost 6,000, we must make it clear that Alpha also has a large plant nursery sector and not just a supermarket mixed with some garden products.

However, the main objective remains: the nursery must bring people into the supermarket after all during the peak months (April and May), to retain those who see alpha only as a plant nursery.

To achieve all objectives, we need to unite the sections into one. The synergies between the two businesses, as anticipated in theoretical chapter, are exploited more when customers use the "shopping center" effect ("as long as I am here, I also buy this"). Keeping the cash boxes separate is also a huge disincentive to go and make another queue and another receipt. Let's look at one of the many possible layout redefinition strategies.

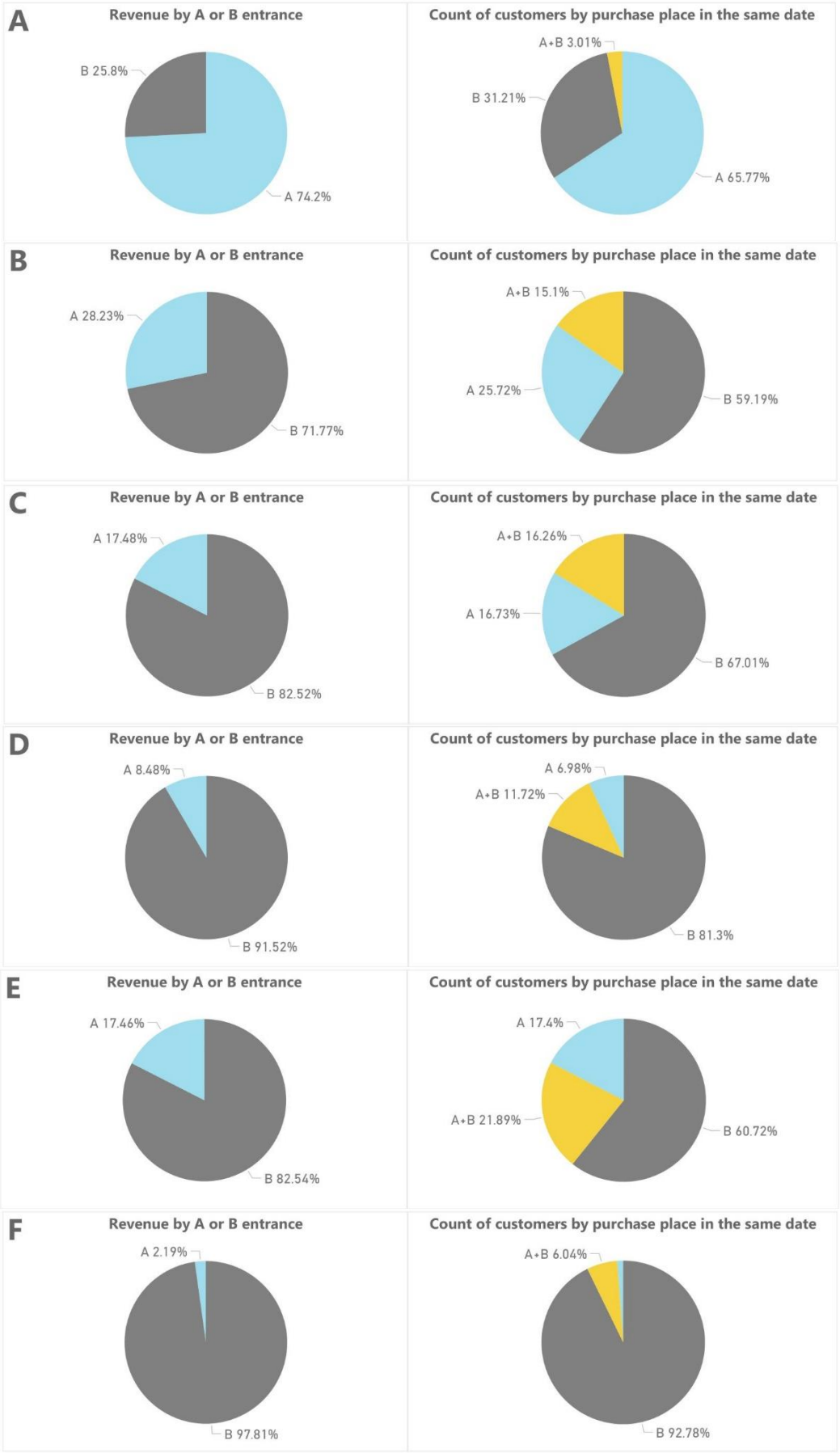


Figure 4.20. Cluster analysis by entrance.

A good business strategy is to create a unique store where people must pass through both sectors in order to pay. The path to do, could be to show first the products of the plant nursery business and then that of the supermarket business. It's a bit of IKEA strategy (the idea is: "after offering what you came for, I'll show you that you can also buy your dinner"). However, in this way it would be to the detriment of customers who come only to the supermarket and spend a lot, as they may not want to do all the way, and we know that there are many supermarket competitors in the area. For this reason, it would be advisable to leave the two entrances, in which on the one hand it is sold only plant nursery, from the other only supermarket. However, checkouts are only from the supermarket side (*Figure 4.21*). Of course, a lot of space is needed to implement this strategy, as garden products, often very large, must find space for a dedicated corridor that passes through the supermarket. However, companies that deal with gardening and plants do not usually have space problems.

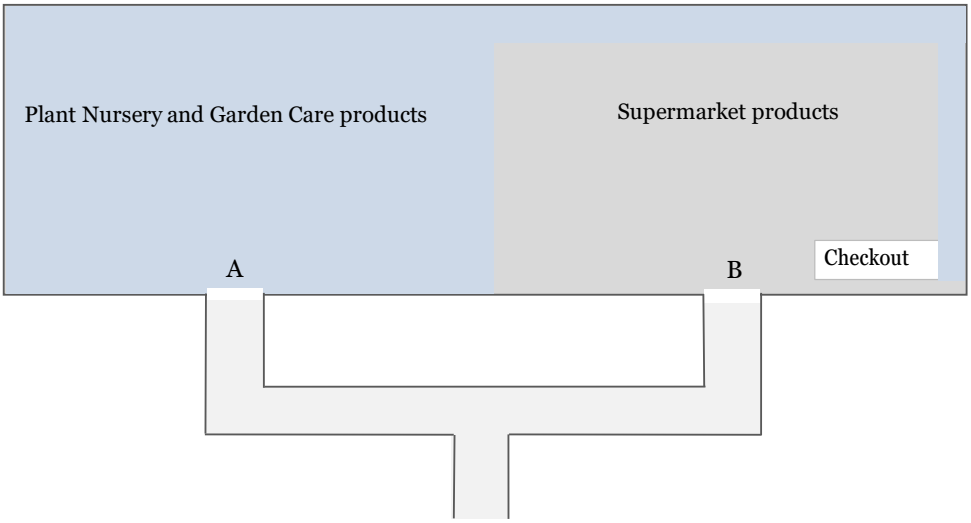


Figure 4.21. Proposal to change store layout.

4.2.3 Association rules for joint-market penetration strategies

Association rules are a useful tool to understand the combinations of products that customers buy.

Instead of using the products, which are really many and constantly changing in the two businesses, it was preferred to use categories. In this way, moreover, it is consistent with the analysis by cluster carried out just above. Reasoning by categories, moreover, price makers have more freedom on product pricing: if it is convenient to promote a certain product category, managers can choose which products cover by discounts, so as to remedy problems with deadlines or price impositions by suppliers.

To use association rules, you need to have a transactional dataset, where for each line there is an ID and a field containing the list of categories of products associated with that ID. In this case, however, the ID is not the id_receipt field, because we need to trace the purchase of products from both entrances on the same date as a single transaction, which instead have separate cash registers (and receipts). For this reason, the ID used is a concatenation between cod_customer and date_sale.

Below is the code to carry out association rules, through the apriori algorithm.

❖ Code: Association rules extraction

Query:

```
SELECT CONCAT(f.cod_customer, f.sale_date)as ID,
       p.desc_category,
       c.cluster_2018
FROM FACT_SALES2018_B2C f,
     DIM_PRODUCT p,
     CLUSTERS_2018 c
WHERE f.cod_product = p.cod_product
AND f.cod_customer = c.cod_customer
AND f.cod_customer not like '%OP%'
AND f.cod_product not like '%BUSTA%'
```

Script R Language (Source: author and from 'arules' package guide) :

```
library(arules)
library(dplyr)

DATA <- read.csv(file="path", header=TRUE, sep=",")
#DATA<- filter(DATA, CLUSTER_2018=='A')
DATA <- DATA [1:(ncol(DATA)-1)]
DATA$ID <- as.numeric(DATA$ID)
dir.create(path = "tmp", showWarnings = FALSE)
write.csv(DATA, "./tmp/tall_transactions.csv")

order_trans <- read.transactions(
file = "./tmp/tall_transactions.csv",
format = "single",
sep = ",",
cols=c("ID","DESC_CATEGORY"),
rm.duplicates = T,
header = TRUE)

rules <- apriori(order_trans, parameter = list(minlen=2, maxlen=3, support=0.009,
confidence=0.5))

rulesDataFrame <- as(rules, "data.frame")
rulesDataFrame
```

Playing with the thresholds, as is usual with the association rules algorithms, the rules of the entire dataset were obtained. They therefore reflect the buying behaviour of all customers. *Figure 4.22* shows few of the most interesting rules found (in terms of support, confidence and lift), ordered by lift (it is clear that the rules like "oil -> oil container", even if they have very high parameters, do not have business value, for this reason they are omitted).

Rules (849 found)	Support (0.009)	Confidence (0.5)	Lift
{ALBICOCCHE} => {PESCHE}	0.01955332	0.6055464	8.548363
{PELARGONIUM} => {PIANTE DA INTERNO}	0.01004472	0.6117689	6.079103
{CAVOLI,ZUCCHE} => {VERDURE FOGLIA}	0.01269581	0.5093359	6.029388
{ERBACEE PERENNI} => {PIANTE DA INTERNO}	0.001148809	0.5963303	5.950927
{FIORITURE PRIMAVERILI} => {PIANTE DA INTERNO}	0.01433949	0.5336549	5.430576
{CETRIOLI} => {POMODORI}	0.02741824	0.658647	4.961006
{FRUTTA ESOTICA,PERE} => {MELE}	0.01363253	0.5731979	4.874913
{BISCOTTI,LATTE E DERIVATI} => {LATTICINI LIBERO SERVIZIO}	0.02616928	0.6249297	4.047203
{INFUSI/POLV/LIEVITI/AMIDI} => {FARINACEI/PASTA/RISO}	0.02505582	0.5996898	3.718294
{SURGELATI/GELATI} => {LATTICINI LIBERO SERVIZIO}	0.01855768	0.5264037	3.566783
{SALUMI LIBERO SERVIZIO} => {LATTICINI LIBERO SERVIZIO}	0.02437832	0.531262	3.590473
{PRODOTTI BIOLOGICI} => {FARINACEI/PASTA/RISO}	0.02246364	0.5538931	3.510701
DOLCI DA BANCO} => {PANE/SCHIACCIATA BANCO/PIZZA}	0.03391049	0.7522216	3.167901
{BIBITE-SUCCHI FRUTTA} => {PANE/SCHIACCIATA BANCO/PIZZA}	0.02490854	0.5722019	2.649084
{ACQUE} => {PANE/SCHIACCIATA BANCO/PIZZA}	0.03929516	0.5598926	2.613609
{ALCOOLICI-SPUMANTI} => {PANE/SCHIACCIATA BANCO/PIZZA}	0.023771511	0.5400161	2.556325
{ARGILLA} => {TERRICCI UNIVERSALI}	0.002916208	0.5113636	2.445665
...	...	...	...

Figure 4.22. Relevant entire dataset rules.

What interests us, beyond the found associations, is that none of the 849 rules found are of a cross-business type, that is, it presents categories of two different businesses in the head and in the body. This shows once again how synergies are not present to a great extent.

An idea, therefore, for the company to exploit growth synergies is the following:

1) Find a synergistic cluster in which customers buy from both business in fairly high quantities and often (a “good cluster”).
2) Then identify another cluster, numerous but with low revenues and purchasing frequencies (a “bad cluster”), which has a purchasing behaviour similar to the first cluster.
3) Find cross-business association rules in the “good cluster”, whose rule body is a category bought a lot from both clusters. The business sector of the body category will bring the business sector of the category into the head (we want the plant nursery to drag the supermarket).
4) Check that the same rules are not present in the "bad cluster".
5) Propose cross-selling, cross-business bundling and cross-marketing strategies on the categories that appear among the cross-business association rules of the

"good cluster" on customer in the "bad cluster": since the buying behaviours are similar, the body of the rule (in common between the clusters) would be able to drag the category in the head even in the "bad cluster".

It is sometimes wrong to think of growing by offering cross promotions to customers who already have rules that include the association of the categories to put in cross: customers already buy those categories in an associated way, making a promotion on that could only be a loss of revenues. It can be useful, however, in cases where there is a surplus of offers compared to the demand of these categories, but it does not have a real purpose of growth.

Below is an example, among the multitude of possible ones.

1)

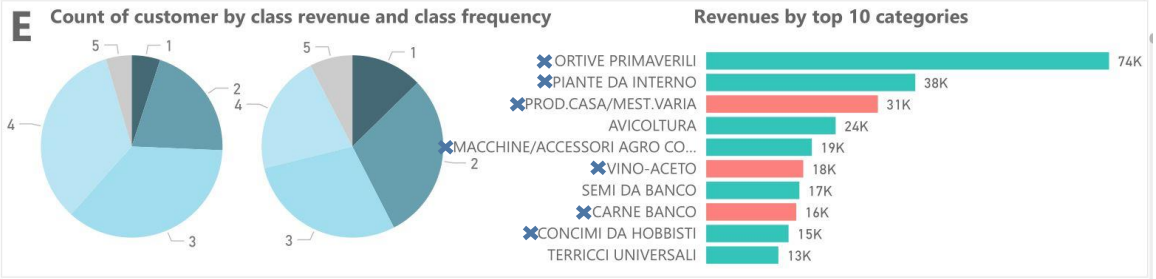


Figure 4.23. Cluster E analysis by classes and categories.

Cluster E is a cluster where customers buy both nursery and supermarket in good quantities with good frequency. It is also very synergistic on entrances.

2)

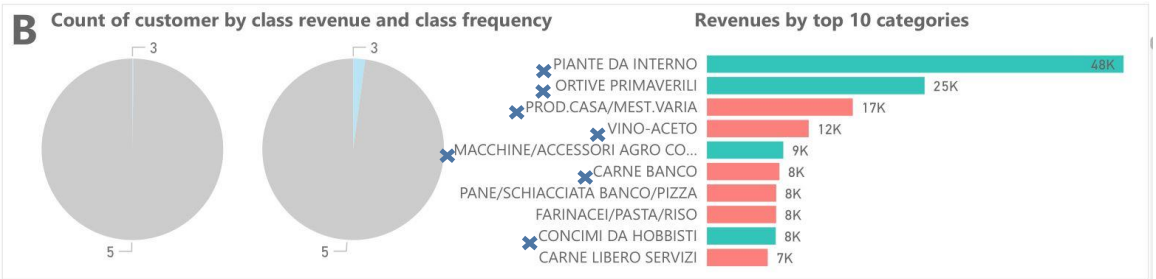


Figure 4.24. Cluster B analysis by classes and categories.

Cluster B is a cluster where many customers spend little and rarely come to buy. Furthermore, it shares with E 7/10 its top 10 categories. Also, products of both businesses are bought in B.

3)

The following cross-business association rules are found in cluster E, where in the body of the rule there is a plant nursery category among top 10 categories of both clusters and in the head of the rule there is a supermarket category:

Rules (993 found)	Support (0.02)	Confidence (0.3)	Lift
{CONCIMI DA HOBBISTI} => {LATTICINI LIBERO SERVIZIO}	0.025461403	0.3134622	3.0823568
{PIANTE DA INTERNO} => {CARNE BANCO}	0.021463245	0.4235784	2.2896048

Figure 4.25. Cross-business rules in cluster E.

4)

The same rules, in cluster B are not present, not even with very low support threshold.

Rules (643 found)	Support (0.005)	Confidence (0.3)	Lift
{CONCIMI DA HOBBISTI} => NULL			
{PIANTE DA INTERNO} => OTHERS			

Figure 4.26. Cross-business rules in cluster B.

4)

Going to offer cross-selling promotions or cross-business bundling between the categories of point 3) to customers of cluster B, there is a good chance that the customers of this cluster will begin to buy more supermarket categories in the head of the rules. In this way, approximately 6,100 customers of cluster B would develop synergies between the two businesses, becoming more aware of Alpha as a double business company. That is, we repeat, it is due to the fact that in cluster E, which has very similar buying behaviours but is populated by more habitual customers, these rules are quite strong. Given the similarity, among all the possible cross-business combinations of products, these categories of cluster E rules are by liking probability, the best for cluster B. Which to choose between a cross-selling, a cross-business bundling or a marketing strategy associated with the products of the examined categories, is a decision maker's choice based on his constraints and his assessments in the field. The objective of the three strategies is however almost the same, if not some smallness seen in chapter two. All methods, however, aim to increase synergies through joint market penetration, that is through existing products in existing markets. If we were to advise the means by which to implement the offers, they could be discount vouchers that cashiers offer, discounted bundling, or targeted advertisements by mail or phone.

This is just one of the many examples that could be found. The data has enormous potential, but time is needed to process it to the fullest. The biggest challenge is to process them in such a time that the environment does not change much in the meantime.

4.2.4 Predictive Analytics: timing of growth synergies

So far we have talked about "how" strategies could be adopted to exploit growth synergies. Now let's talk about "when" the above strategies are more effective, that is when it is convenient to implement them.

Predictive analysis, or those that try to know the future, allow decision makers to implement strategies that are currently most appropriate for them to have positive results. Predictive analytics are a powerful tool, and even in this business context they could be used for many different purposes.

In this document, applications are proposed with the aim of bringing more effective strategic solutions to the problem that up to now we have defined "primary": to bring clients who spend little and come little (clusters A and B, and in a lesser way C) to recognize the company also as a supermarket, so as to make them "habitual" through a synergic transport from the plant nursery business, preferred by these customers, to the supermarket business. Take for example the low performance of the numerous cluster B, already covered in association rules.

Without going into analysis by category, we see from the *Figure 4.27*, that customers buy a lot in the months of April and May.

Figure 4.27. Revenues by business sector by month in cluster B.

The huge customer base in these two months is absolutely not used in the supermarket business. For this reason, the strategies of joint market penetration, seen in the previous paragraph, would be optimal in these months: with the sale of plant nursery products, promotions could be made for the supermarket, so as to make the products of the latter business known also to who ignores its existence or quality. The same type of considerations could also be made for weekly analyses. For example, if over the weekend there is plenty of revenue for the nursery, the strategies could be implemented more in these days.

Entering the categories in detail, considering the rule in *Figure 4.25*, we have, for 2018, revenues in *Figure 4.28*.

Figure 4.28. Revenues by categories in cross-business rules by month in cluster B.

The trend of the two categories seems to reflect roughly that of the two sectors. So, it would seem appropriate to implement joint market penetration strategies seen in the chapter of association rules in the spring months, where indoor plants are abundantly sold and could drag the meat.

However, this reasoning presents a fairly consistent imprecision, as we are projecting historical data, exactly as they are, on the future, when they will certainly change. To predict future data based on historical data, such as product demand, forecast algorithms can be used. the ARIMA method is very useful and precise, but in this case it cannot be applied because there is only one annuity of data. In fact, ARIMA needs more years of data to break down data by detecting seasonality. Assuming you also have sales data for 2016 and 2017, referring to the topic just discussed, we could estimate the revenues for the two categories of the rules in 2019. Below are the steps, and the code, which can be used if the data is available.

❖ Code: ARIMA method

Queries:

```
SELECT SUM(revenue) as revenues_per_month
FROM FACT_SALES2018_B2C F,
     DIM_PRODUCT P,
     DIM_TIME T,
     CLUSTERS_2018 C
WHERE F.cod_product = P.cod_product
AND F.sale_date = T.sale_date
AND C.cod_customer = F.cod_customer
AND desc_category ='PIANTE DA INTERNO'
AND cluster_2018='B'
GROUP BY month_number
ORDER BY month_number

SELECT SUM(revenue) as revenues_per_month
FROM FACT_SALES2018_B2C F,
     DIM_PRODUCT P,
     DIM_TIME T,
```

```
     CLUSTERS_2018 C
WHERE F.cod_product = P.cod_product
AND F.sale_date = T.sale_date
AND C.cod_customer = F.cod_customer
AND desc_category ='CARNE BANCO'
AND cluster_2018='B'
GROUP BY month_number
ORDER BY month_number
```

R Script for ARIMA (Source: author and from 'tseries' package guide):

```
library('ggplot2')
library('forecast')
library('tseries')

data <- read.csv('path',header = TRUE, sep=';')

revenues_tot <- data$revenues
revenues <- data$revenues[1:36] #3 years available
plot(revenues)

ts_revenues <- ts(revenues,frequency = 12)
ts_revenues
plot(ts_revenues)

decomp = decompose(ts_revenues)
deseasonal_revenues <- ts_revenues - decomp$seasonal
seasonal <- decomp$seasonal
plot(decomp)

model <- auto.arima(deseasonal_revenues, seasonal=FALSE)
#tsdisplay(residuals(model), lag.max = 45)

fcast <- forecast(model, h = 12) #one year to forecast
plot(fcast)

values<- NA
forecast <- fcast$mean
for(i in 1: length(forecast)){
values[i] <- forecast[i] + seasonal[i] }

temp <- NA
for(j in 1: length(revenues)){
    temp[j] <- NA }

values <- c(temp,values)
x2<- c(1:(length(values)))
revenues_frame <- data.frame(revenues_tot)
values_frame <- data.frame(values)
p=ggplot()+ geom_line(data = revenues_frame,aes(x=x2, y=revenues_frame$revenues_tot),
  colour="blue") + geom_line(data = values_frame,aes(x=x2, y=values_frame$values),
  colour = "red") + xlab('date')+ ylab('revenues')
print(p)

values <- c(values, data$revenues[37:48])
values <- c(ts_revenues, values)
ts_values <- ts(values)
plot(ts_values)
values
```

Therefore, starting from the eventual estimated values for 2019, it is possible to implement strategies for the development of growth synergies where the body of the

rule is well sold and the head of the rule is little sold (and with an overabundance of offer) in order to trigger good drag effects.

It is however true that, if we want to adopt joint market penetration strategies aimed at individual customers, a priori we cannot find out whether they will or will not come to buy in a given month. In order to better estimate who and how many customers could take part in the promotions, predictive classification is helpful. In particular, we want to understand, or better to foresee with a good probability, starting from the buying behaviour of a customer in certain months, whether or not it will come in a certain subsequent month. To do this we use, to create the model, all the data of 2018, in order to apply it to the new data that will arrive in 2019. Below are all the steps to implement the algorithm. Then follow the considerations on the result obtained. It is assumed that you want to predict, for example, whether and which customers will come in May from their buying behaviour in Jan, Feb, Mar, Apr , from personal information like the most bought business sector, the total expenditure, the number of receipts, gender and profession.

❖ **Code: predictive classification for customer purchases**

Query to create customer information table:

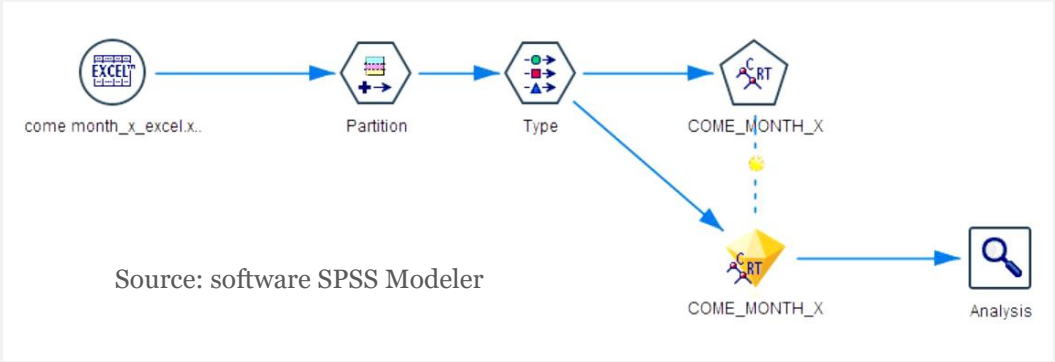
```
SELECT DISTINCT CC.cod_customer,
       CC.top_sector,
       CC.tot_expenditure,
       CC.num_receipts,
       EE.profession,
       EE.gender,
       CC.come_month_x
FROM (
       SELECT TAB.cod_customer,
              TAB.business_sector as top_sector,
              TAB.tot_expenditure,
              TAB.num_receipts,
              CASE when SUP.freq >= 1 THEN 'YES' ELSE 'NO' END AS come_month_x
       FROM (
              SELECT TABLE1.cod_customer,
                     TABLE1.business_sector,
                     TABLE2.tot_expenditure,
                     TABLE2.num_receipts
              FROM (
                     SELECT tmp.cod_customer,
                            tmp.business_sector,
                            tmp.ranking
                     FROM (
                            SELECT cod_customer,
                                   P.business_sector,
                                   RANK() OVER (PARTITION BY cod_customer ORDER BY
                                                SUM(F.revenue) DESC) ranking
                            FROM FACT_SALES2018_B2C F,
                                 DIM_PRODUCT P,
                                 DIM_TIME T
                            WHERE F.sale_date = T.sale_date
                        AND F.cod_product = P.cod_product
                        AND f.cod_customer not like '%OP'
                        AND T.month_name in ('JAN', 'FEB', 'MAR','APR')
```

```
                        GROUP BY F.cod_customer,
                                 P.business_sector ) tmp
               WHERE tmp.ranking = 1) TABLE1,
                     (SELECT F.cod_customer,
                             SUM(revenue) as tot_expenditure,
                             COUNT(DISTINCT id_receipt)as num_receipts
                      FROM FACT_SALES2018_B2C F,
                           DIM_TIME T,
                           DIM_CUSTOMER_B2C C
                      WHERE F.sale_date = T.sale_date
                      AND F.cod_customer = C.cod_customer
                      AND T.month_name in ('JAN', 'FEB', 'MAR','APR')
                      GROUP BY F.cod_customer) TABLE2
          WHERE TABLE1.cod_customer = TABLE2.cod_customer) TAB
     LEFT OUTER JOIN
     (SELECT cod_customer,
             COUNT (DISTINCT id_receipt) freq
      FROM FACT_SALES2018_B2C F,
           DIM_TIME T
           WHERE F.sale_date = T.sale_date
           AND T.month_name in ('MAY')
           GROUP BY cod_customer) SUP
           ON TAB.cod_customer = SUP.cod_customer)CC,
           DIM_CUSTOMER_B2C EE
           WHERE  CC.cod_customer= EE.cod_customer
```

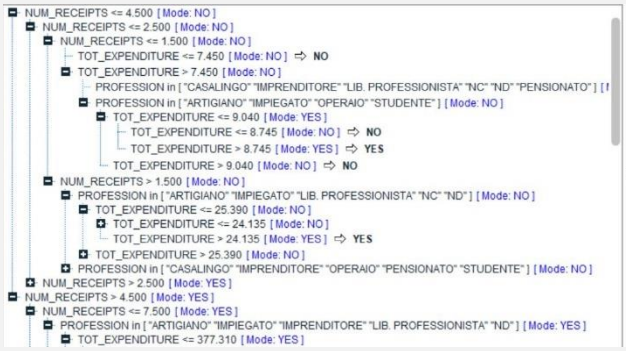
The result is a table composed by:

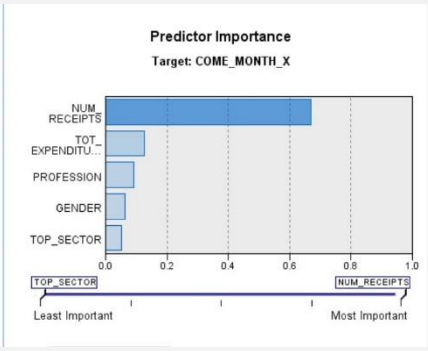
`cod_customer` `top_sector` `tot_expenditure` `num_receipts` `profession` `gender` `come_month_x`

CART tree by SPSS stream (Source: author)



Source: software SPSS Modeler





The first node is used to load the previously created table; the second (Partition) is used to divide data into training (70%) and test (30%) set; the third (type) serves to change the format of the data so as to be read well by the algorithm; the fourth (CART) is the CART algorithm node: in this node many parameters can be set, including tree depths, variables to predict and predictors; the fifth is the result produced by the previous node and shows the tree obtained by training set data and the importance of

```
predictors in the prediction of the labels (figure under the stream);  finally, the
last node (analysis) calculate the accuracy of the result obtained, that is how much
predictions on test_set are corrects.
```

The overall result is shown in *Figure 4.29*.

test_set results		
correct	3325	78,29%
wrong	922	21,71%
total	4247	100%

Figure 4.29. Result of predictive classification (accuracy).

The tree model created with the training set, therefore, offers 78.29% of probability of correct prediction on the test set. Therefore, it can be assumed that for the new data of 2019, with the same analysis and assuming constant boundary conditions, can be foreseen with a probability of success of 78.29% if a customer will come in the month of May when the data of the previous four months are available. So, in this way, the targeted strategies of joint market penetration can be perfected even considering customers purchase periods. It is obvious that this is only an example, the predictive classification can be applied for many other purposes.

4.2.5 Summary, results and limits

Starting from the data of just one year, 2018, we have succeeded in making good reasoning. The useful data in the data warehouse made it possible to highlight substantially two problems for Alpha:

a) *Poor customer loyalty;*

b) *Failure to exploit synergies between businesses*.

The first problem is highlighted both by the results of the ad-hoc classification created for the case (many customers spend little and rarely buy), and by some basic statistics including the distribution of sales and the percentage of customers registered on the total.

The second problem is highlighted by the unrelated sales trends of the two businesses and by problems due to the physical structure of the store, separated into two entrances. Furthermore, the second problem is linked to the first because customers who come more rarely contribute more to the plant nursery business particularly in spring months, suggesting that the huge pool of customers who behave in this way must be loyal, pushing them to buy from the supermarket, preferred by those who spend more and come more often. In this way the sales of the supermarket, a business

with higher competitiveness on the territory, would benefit from an advantage over the competition for the attraction of the customers by the plant nursery.

To increase growth synergies between businesses (only growth because of available data), increasing the sales of the supermarket, and therefore retaining customers, two approaches have been taken substantially:

1) *proposing the change of the store layout;*

2) *proposing strategies of joint market penetration starting from the purchase behaviour of customers.*

For the first point, a layout of the store was proposed that obliges customers who buy almost only nursery products to go to the supermarket sector to pay, while still keeping the two separate entrances so as not to discourage those who only want to buy from the supermarket.

For the second point, instead, starting from a clustering that allowed the grouping of customers by purchasing similar product categories, a method was offered to suggest which product categories to propose in joint market penetration strategies for growth synergies through association rules: proposing, to "bad clusters" similar to "good clusters" for purchased categories, synergistic rules of "good clusters", in which the body of the rule is common in all the clusters in analysis. Moreover, the strategies of joint market penetration (cross-selling, cross-business bundling or cross-marketing) on the products of selected categories, are better if applied in certain moment: that is when the category that drags is in a high point of its sales trend, and the category dragged is in a down point of its sales trend or has an overabundance of offer. To recognize these time points, however, it is not enough to know historical trends, but it is necessary to estimate future trends through forecasting techniques, such as ARIMA. Moreover, if it is considered appropriate to offer targeted promotions to individual customers, it is possible to estimate with good probability of success if a customer will come to buy from the company in a certain period, given his purchase behaviour of the previous periods, through predictive classification algorithms.

The results obtained are therefore not real strategies, but rather suggestions to be given to the decision maker to implement them, starting from the evidence of data.

This type of approach to the business case, however, presents some problems, or rather simplifies some conditions, which in reality have a sufficient influence on the results of the strategies. First of all, data quality is important: data warehouse seems to contain too much dirty data and often missing values, especially those on customer profiling. Having more data (more than one year) and of better quality undoubtedly increases

the quality of the analysis. In addition, only 55% of the company's revenues come from customers who paid with the loyalty card: the distortion of the assumptions and results, for this reason, could be great, just think of the distortion of the buying behaviour of customers that only sometimes use the card. Another problem for decision makers in this business is that sometimes prices are imposed by suppliers, and there is no freedom to do promotions. Finally, the most important is the fact that it is necessary to consider, in this type of analysis, also the company costs, so as to make an analysis of marginality (and therefore of profit) before implementing growth synergy strategies. For these reasons, the aim of this work, is not to impose strategies for exploiting synergies based on real and effective results, but to suggest methods based on data analysis to identify and deliver to decision makers problems and growth opportunities in the area of business synergies.

Conclusions

The exploitation of synergies is a fundamental objective for managers who want to undertake diversification strategies. The study of multi-business companies in terms of organization and diversification strategies has therefore served as a basis for better understanding the concept of synergy between businesses, and in particular that operational and managerial management has a huge impact on good and successful diversification with synergy target. By classifying the synergies we realized that different synergy objectives require different efforts and allow different results to be obtained: for example, if we focus on operational synergies, a correlated diversification is appropriate because the sharing of operative resources, while if you focus on financial synergies, unrelated diversification seems to be the best solution, as it allows you to mitigate financial fluctuations.

We then deepened growth synergies, that are operative synergies that aim to obtain revenues super-additivities: when there are two or more businesses, the revenues obtained from the business management of the joint businesses are greater than the sum of those obtained from separately managed businesses. The main strategies to unlock this type of synergy have been disclosed and in particular depend on the business objectives according to the Ansoff matrix: for example, if we want to exploit growth synergies to grow in existing markets with existing products, the strategies of joint market penetration seems to do for us, while for other combinations of new/existing markets and new/existing products we use joint market development, joint product development or joint diversification strategies.

We then went into the field of Data Analytics, emphasizing how today the use of data for business decisions is essential for being a competitive company. After giving an overview of data warehouse systems supporting data driven decision making, we analysed in detail the main data mining and machine learning techniques, along with some of the most popular algorithms.

Starting from all these notions, we went into a real case of a double-business company (plant nursery and supermarket) where only the sales data for a year, the catalogue of products sold and some customer information are available: the support of data analysis in the classification of customers based on expenditure and purchase

frequency, after having performed a first competitive and performance analysis, showed us that there are problems of customer loyalty and failure to exploit the synergies between the two businesses. Given the availability of only revenue data, we focused on growth synergies and in particular on joint market penetration strategies, as growth must take place through existing markets with existing products.

More in detail, the two problems are summarized in the fact that many of the registered customers come very rarely and spend little and tend to see the company as only a plant nursery. Moreover, the small of this business is in the spring months and does not in any way drag supermarket sales of the same period, underlining that revenue trends of the two businesses are completely disconnected. The clustering of customers by purchasing behaviour (categories purchased) confirmed the presence of the aforementioned problems, and showed how much the business supermarket is preferred by regular customers (who often come and spend a lot), setting the goal of expanding this business for occasional customers. Starting from this, some data-based solutions have been undertaken, or rather suggestions to be offered to decision-makers to solve problems and exploit the opportunities for synergies. First of all, there was a proposal of change the layout of the store that forced customers to switch from supermarket products before paying at the cash desk. Then, a method was proposed for the synergistic growth of the business starting from the association rules serving joint penetration strategies (cross-selling, cross-business bundling and cross-marketing). Moreover, we stressed the importance of applying this type of strategy at the right time with demand forecasting techniques, and, if the strategies are targeted to the individual customer, through predictive classification, predicting with a fair degree of accuracy the presence or absence of customers in the store in a certain period, so as to perfect the strategy. The objectives, namely, to propose data-driven methods and strategies for exploiting synergies with decision-makers, have therefore been met with a satisfactory result.

It is clear, however, that the analysis does not take into account numerous constraints, some of which are also quite relevant: decision-makers cannot always adjust prices at will; furthermore, the costs and margins of the products (whose data are not available) and all the financial and organizational constraints of the companies are never taken into consideration.

But, for the author, what was pressing was to provide data analysis tools to serve the decision-making processes by focusing on synergies, by linking more with *how* to increase performance than *how much*.

As for future developments, it is clear that it would be really good to test how precisely the techniques are effective in the case study, something that was not possible due to business and time constraints. Furthermore, it would be advisable to carry out this type of analysis with much more data (more years and more detailed), much cleaner and more consistent, pushing the customers who buy to the membership, requesting new information and checking those entered in the system. It could also be interesting to use external data, such as climate, social habits, information on the territory and the financial performance of the market at different times of the year, to improve algorithm precision.

Much more sophisticated techniques could therefore be implemented requiring much greater efforts both in substitution of the techniques used, and in integration. Consider, for example, the use of neural networks instead of the predictive classification used for deep learning predictions about the future buying behaviour of customers. In order to use all these techniques, however, both the ever-increasing quality of data in companies and increasingly cutting-edge information technologies ,that allow the convergence of algorithm results more quickly, are needed. For example, gigantic neural networks require today timing even beyond human life for mass consumption machines. For this reason, the largest and most powerful companies are increasingly investing in very expensive super-computers to obtain a guaranteed super competitive advantage over their competitors.

From the point of view of diversification strategies for the development of synergies, as we have widely said, they are today under the attention of many managers with a long-term vision. Data support, in this, objectively reduces the possibility of negative decisions and outcomes.

References

[1] Hussain A., Salma U., *A Comparative Study on Corporate Diversification and Firm Performance across South Asian Countries*, 2018, Faculty of Management Science, University of Lahore – Sargodha Campus, Pakistan

[2] Bryce D., Dyer J.H., Godfrey P., Jensen R., *Strategic Management: Concepts and Cases*, 2017, Wiley 2nd Edition

[3] Hayes A., *Mergers and Acquisitions – M&A*, 2019, https://www.investopedia.com/terms/m/mergersandacquisitions.asp, last access: 24/08/2019

[4] *Mergers & Acquisitions*, The Account Angle Management Service FZC LLC, https://theaccountangle.com/services/merger-acquisition/, last access: 24/08/2019

[5] Knoll S. revised by Günter Müller-Stewens and Probst G., *Cross-Business Synergies: A Typology of Cross-business Synergies and a Mid-range Theory of Continuous Growth Synergy Realization*, Wiesbaden 2008, University of St. Gallen

[6] Calori R., Harvatopoulos Y., *Diversification: les règles de conduit, 1988*, Harvard Expansion

[7] Dawid H., Reimann M., *Diversification: a road to inefficiency in product innovations?*, 2011

[8] Geiger S.W., Lamont B.T., Martin D., *Diversification strategy and top management team fit*, 2004, Journal of Managerial Issues, 16(3), 361-381

[9] Sekuli V., *Corporate strategy development and competitive advantage of enterprise*, 2010

[10] Balmer J., Gray E., *Corporate Brands: What are They? What of Them?*, 2003, European Journal of Marketing

[11] Finkelstein S., Larsson R., *Integrating Strategic, Organizational, and Human Resource Perspectives on Mergers and Acquisitions: A Case Survey of Synergy Realization*, 1999

[12] Fawcett T., Provost F., *Data Science for Business*, 2013, O'Reilly Media, Inc.

[13] E. Brynjolfsson, L. Hitt, H. Kim, *Strength in Numbers: How does Data-Driven Decision-making Affect Firm Performance?*, 2011

[14] D'Ercole L., La Noce F., *Data Warehousing - Dal dato all'informazione*, 1998, FrancoAngeli 5th edition

[15] Kumar V., Steinbach M., Tan P., *Introduction to Data Mining*, 2006, Mc Graw Hill

[16] Hahsler M., Piekenbrock M., *Package dbscan - Density Based Clustering of Applications with Noise (DBSCAN) and Related Algorithms*, 2019,University of Maryland

[17] *K-means Cluster Analysis*, AFIT Data Science Lab R Programming Guide, https://afit-r.github.io/kmeans_clustering, last access: 22/07/2019

Ansoff H.I., *Corporate Strategy: An Analytic Approach to Business Policy for Growth and Expansion*, 1965, McGraw-Hill, New York

Ansoff H.I., *The New Corporate Strategy*, 1988, Wiley, New York

Baralis E., Cerquitelli T., *Clustering Foundamentals*, Database and data mining group, Politecnico di Torino

Baralis E., Chiusano S., *Association Rules Foudamentals*, Database and data mining group, Politecnico di Torino

Baralis E., *Classificazione*, Database and data mining group, Politecnico di Torino

Brownlee J., *What is Time Series Forecasting?*, 2016, https://machinelearningmastery.com/time-series-forecasting/ last access: 21/07/2019

Chen J., *Autoregressive Integrated Moving Average (ARIMA)*, 2019, https://www.investopedia.com/terms/a/autoregressive-integrated-moving-average-arima.asp, last access: 21/07/2019

D'Ercole L., La Noce F., *Data Warehousing - Dal dato all'informazione*, 1998, FrancoAngeli 5th edition

Ensign P., *Interrelationships and horizontal strategy to achieve synergy and competitive advantage in the diversified firm*, 1998, Wilfrid Laurier University

Fawcett T., Provost F., *Data Science for Business*, 2013, O'REILLY 1st edition

Fernández A. & Montoya R., *Multi-business companies: The Leonisa case*, 2018, Cuadernos de Administration journal by Universidad De Valle

Gary S., *Capturing Synergy in the Growing Multi-Business Firm*, London Business School

Grant R.M. (2015), *Contemporary strategy analysis: Text and Cases*, John Wiley & Sons Inc 9th edition

Han J., Kamber M., Pei J., *Data Mining* ,2012, Morgan Kaufmann 3rd edition

Hargrave M., *Joint Venture*, www.investopedia.com/terms/j/jointventure.asp, last access: 26/6/2019

Horbas' I., Stepanova A., Zhylinska O., *Effective Synergic Interaction of Strategic Business Units of Diversified Company*, 2017, Problems and Perspectives in Management, Volume 15, Business Perspectives

Kabeyi M., *Organizational strategic diversification with case studies of successful and unsuccessful diversification*, 2018, International Journal of Scientific & Engineering Research Volume 9

Kannan P., Saravanan R., *Diversification - Strategies for managing a business*, 2012, International Journal of Multidisciplinary Management Studies, Vol.2

Kenny G., *Diversification Strategy – How to grow a business by diversifying successfully*, Kogan Page, 2009

Kenton W., *Consortium*, https://www.investopedia.com/terms/c/consortium.asp, last access: 26/6/2019

Kenton W., *Mergers and Acquisitions (M&A) Definition*, 2019-05-08, www.investopedia.com

Kenton W., *Strategic Alliances*, *www.investopedia.com/terms/s/strategicalliance.asp*, last access: 26/6/2019

Kopp C., *Partnership*, www.investopedia.com/terms/p/partnership.asp, last access: 26/6/2019

Kriesel D., *A Brief Introduction to Neural Networks*, 2005, ZETA2-EN

Neelamegam S., Ramaraj E., *Classification algorithm in Data mining: An Overview*, 2013, International Journal of P2P Network Trends and Technology (IJPTT) - Volume 3

Nwaeke L., *Business Synergy and Corporate Competitive Advantage: A Theoretical review*, 2010, The University Advanced Research Journal, Rivers State University of Science and Technology

Ramageri B., *Data Mining Techniques and Applications*, Indian Journal of Computer Science and Engineering

Shahri M., *The effectiveness of Corporate Branding Strategy in Multi-business Companies,* 2011, Australian Journal of Business and Management Research

Thomas J. G., revised by Mason W., *Diversification Strategy*, 2019-06-11, www.referenceforbusiness.com

Villasalero M., *Intra-network knowledge roles and division performance in multi-business firms*, 2014, Journal of Knowledge Management.