Master of Science in Computer Engineering

Internship report

Optimization of a mathematical model for predicting churn and improving cross sell policies for a B2B operator

Author: Ndjekoua Sandjo Jean Thibaut*

Advisors: Raybaut Daniel † and Daniele Apiletti‡

September 2020 - March 2021

Contents

*jean.ndjekouasandjo@telecom-paris.fr

†daniel.raybaut@egis.fr

‡daniele.apiletti@polito.it

Abstract

It is now widely accepted that firms should direct more effort into retaining existing customers than to attracting new ones. To achieve this, customers likely to defect need to be identified so that they can be approached with tailored incentives or other bespoke retention offers. Such strategies call for predictive models, capable of identifying customers with higher probabilities of defecting in the relatively near future. In addition, not all users have the same added value to the business. That's why it's just as interesting to set up customer segmentation strategies to better understand customers' needs and

provide them with offers that suit them. The aim of this study is to present the steps and main challenges faced in order to build a predictive algorithm and the different segmentation techniques proposed, in order to reduce the churn rate and improve the commercial campaigns for Easytrip Transport Services which is an entity of the Egis group.After testing different data analysis and processing techniques, and then different predictive models, gradient boosting produced the best performance on the churn prediction task. The unsupervised machine learning techniques did not produce the best performance. On the other hand, a better understanding of the business need allowed the definition of a data processing pipeline allowing the sales teams to derive different business rules. At the end of the POC phase, the cross sell project was launched in pilot phase, while the churn project remains to be finalised.

1 Introduction

As described in Figure 1, Egis is an international player in construction engineering and mobility services, Egis creates and operates intelligent infrastructures and buildings capable of responding to the climate emergency and the major challenges of our time, enabling more balanced, sustainable and resilient land use planning. A company 75% owned by Caisse des Dépôts and 25% by partner managers and employees, Egis puts all its expertise at the service of the community and cutting-edge innovation within reach of all projects, at every stage of their life cycle: consulting, engineering, operation. Through the diversity of its areas of intervention, Egis is a key player in the collective organisation of society and the living environment of its inhabitants throughout the world. 1.22 billion in revenues managed in 2019 with 15,800 employees.

The group is made up of different business units including the BU MENS (Project structuring, Operation and Services) where the internship was carried out. The MENS Business Unit is one of the world leaders in the financial structuring of PPP (Public-Private Partnership) projects and the operation of infrastructure and services. In the road sector, it operates and maintains 44 roads or motorways (including major engineering structures such as bridges, viaducts and tunnels) through 28 operating companies employing more than 9,300 people in 20 different countries, covering almost 4,400 km. In the airport sector, the BU operates, maintains and develops a network of 17 airport hubs around the world, handling more than 30 million passengers and 330,000 tonnes of freight per year. The MENS BU also offers an extensive range of services for its customers, such as electronic toll collection, inter operable device for Heavy goods vehicles and parking control for on-street and off-street car park.

The context of the project was mainly addressed to the Sales department of Easytrip Transport Services (ETS) entity which provides a range of services to HGV drivers across Europe as described in Table 1. Customers can be companies with many employees and trucks or single persons with one single truck, eventhoug most of the customers are actually large companies and that's why we will mainly talk about a B2B context.

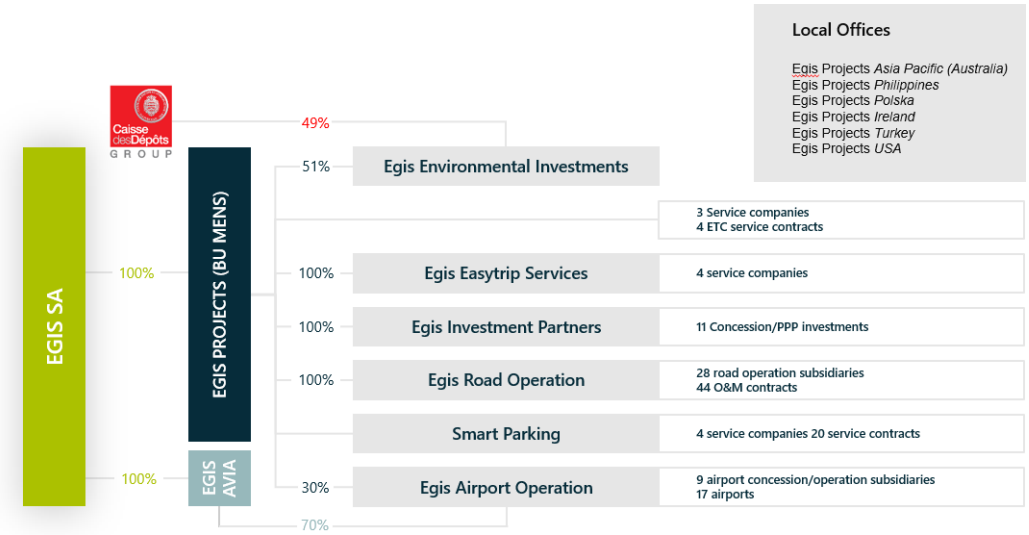

Figure 1: Egis group description with focus on BU MENS

In general, customer churn remains a particularly salient concept in contemporary marketing and should not be ignored by B2B companies. Nowadays, due to improved access to information, customers are more transient and it is easier and less costly for them to switch between competitors [1]. Firms recognize this and are interested in identifying potential churners in order to attempt to prevent defection by targeting such customers with incentives. As part of this general trend,the main reasons that motivated the actual study can be summarized in the following 4 points:

- ETS market is highly competitive and currently undergoing a price war.
- Competitors are reducing their commissions as a means to gain additional business and protect their customer base.
- This trend affects Easytrips' capacity to maintain its margins, and protect it's customer base.
- As a consequence, some customers tend to move to competitors as they receive a more affordable offer.

Based on the previous observations and considering the amount of user data stored in the various systems that had remained unexploited until now, it seemed natural to make the most of this data and to set up machine learning algorithms to predict and prevent customer churn. In combination with this anti churn campaign, different segmentation techniques were also explored to create user groups with different profiles and needs with the goal of improving cross sell policies and increase customers stickiness. The purpose of this study is therefore to present the steps taken to implement a predictive algorithm by detailing all the phases that have been followed, and the different challenges encountered with the proposed solutions.

Product Name	Description
TOLL	Inter operable device allowing drivers to travel in different countries and pay toll in EU.
VAT	VAT recovery service in different EU countries
EXCISE	Recovery of VAT on fuel invoices in different european union countries
TRAINS	Train reservation service for heavy goods vehicles in the EU
FERRIES	Ferries reservation service for heavy goods vehicles in the EU
CONNECTED SERVICES	Web portal enabling managers to geolocate their fleet of vehicles in real time.
ADVISORY & REPRESENTATION	Legal representation service for drivers in FRANCE and ITALY.

Table 1: Description of the different products or services of ETS

2 Description of EGIS

2.1 General overview of the group

Created in October 1997 and specialising in transport and development infrastructure engineering, the Egis holding company holds all the subsidiaries of the former Scetauroute, which is refocusing on its operational activities. There are 35 of these subsidiaries. They include: Scetauroute, Semaly, BCEOM, Isis, Beture Infrastructure, Jean Muller International, Axial, Dorch Consult, CMPS&F, Italconsult, Transinfra, Transroute International, Sofremer and UK Highways Services. Present in more than 80 countries, they have a consolidated turnover of 2.646 billion francs and 5,000 employees.

In 2007, EGIS decided, in agreement with its chairman Philippe Segretain and its managing director Nicolas Jachiet, to merge these companies - their highly diversified backgrounds were complementary. The integration into one company of several companies such as Scetauroute (which later became Egis Route), BCEOM, Semaly (now Egis Rail), Isis, which later became Egis Mobilité, quickly grew. Egis was joined in 2010 by the Guigues group, specialised in environmental engineering, and in 2011 by the Iosis group, the French leader in building and nuclear civil engineering; this integration provided the opportunity to open up 25% of the capital to managers and staff.

For its infrastructure development activity, the Egis group was organised into seven regional companies before 2007 :

- Beture Infrastructure (ex-Beture Setame, Beture, or Scet-Beture, created in 1960 as part of SCET) in Île-de-France, the east of the Centre and Haute ;Normandie;
- Seralp Infrastructure in Rhône Alpes, Burgundy and Auvergne;

- Beterem Infrastructure in the Mediterranean region;
- Ouest Infra in the Greater West (Pays de Loire, Brittany, Basse Normandie, Poitou-Charentes and the West of the Centre);
- ACI in the North;
- Est Ingénierie in the East;
- Sud-Ouest Infra in the greater South-West (Aquitaine, Midi-Pyrénées, Limousin).

These regional companies merged on 1 June 2007 to form Egis Aménagement, a national company, which itself became Egis France on 1 January 2011. On 1 July 2011, as part of the reorganisation of the Egis group, Egis Mobilité and the French management of Egis Route joined Egis France. BCEOM International became Egis International, integrating part of Egis Route, the other part of this company being integrated into Egis France (now Egis Villes et Transports).

In 2014 and 2015, two PSE (plan de sauvegarde de l'emploi) were launched. They concerned the subsidiaries AVP (Atelier Ville et Paysage): 37 employees were fired; and Egis Eau: 58 employees were fired. In 2016, Egis moved to London to develop its activities, particularly in the building sector, and to work more closely with British architects. In July 2016, the La Poste Group (a French company), Caisse des Dépôts and Egis joined forces to create Sobre, a joint company dedicated to the energy transition. In July 2015, Projacs, a company specialised in project management and construction supervision in the Middle East, joined Egis. At the end of 2015, the three Brazilian engineering companies Egis Lenc (roads, environment and geotechnics), Egis Engenharia e Consultoria (railways and urban transport) and Egis Aeroservice (airport consulting and engineering) joined forces to form a single entity, Egis Engenharia e Consultoria. In March 2016, Egis Easytrip Services, a subsidiary of Egis, acquired the Dutch company Versluis in order to strengthen its leading position in the field of services for heavy vehicle fleets, and In 2017, Egis acquires OCACSA, Mexico's leading independent motorway operator. The group is strengthening its presence in South East Asia with the acquisition of two companies based in Hong Kong: the architectural firm 10 DESIGN in 2017 and Inhabit, an envelope and environmental engineering consultancy in 2019. In France, the Plantier engineering office in Annecy, the Michel Frustié firm of economists in Montpellier and EXYZT, a company specialised in geomatics in Castres, joined Egis in 2018.

2.2 Activities of the group

Egis is a group with international activities in different sectors as shown in Figure 2

Egis is notably involved in the following projects :

Aviation

- Egis manages and operates sixteen airports worldwide - São Paulo-Viracopos (Brazil), Tahiti Faa'a, Bora-Bora, Raiatea, Rangiroa (French Polynesia), Brazzaville, Pointe-Noire, Ollombo (Congo), Abidjan (Ivory Coast), Pau, Bergerac,

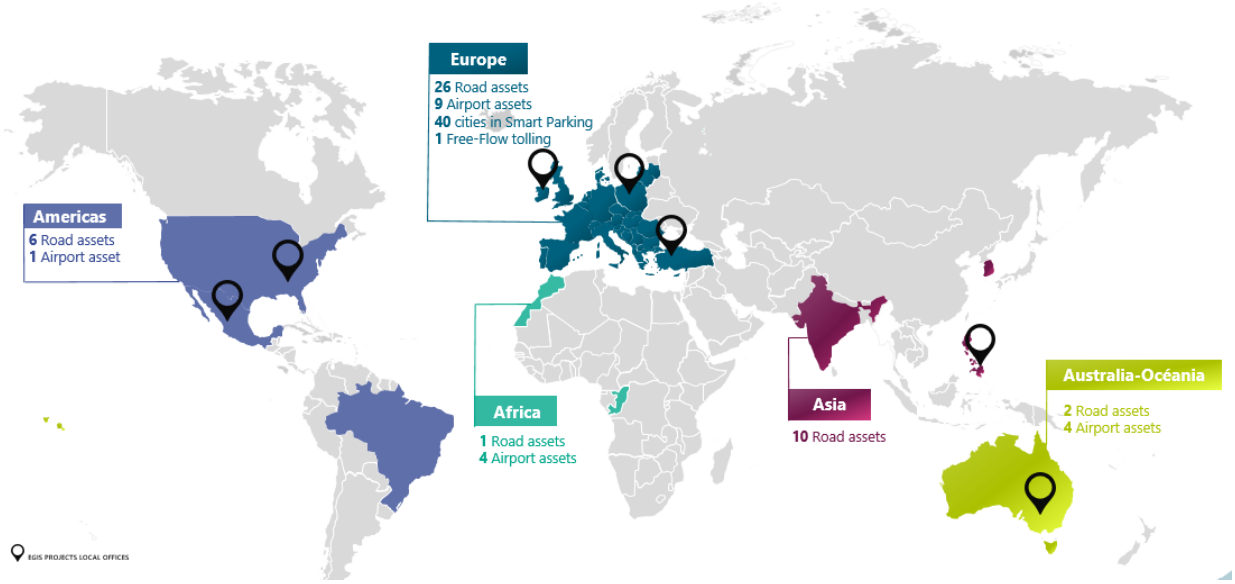

Figure 2: Egis World presence

Brest and Quimper (France), Larnaca, Paphos (Cyprus), Antwerp and Ostend-Bruges (Belgium) - totalling more than 28 million passengers and 366,000 tonnes of cargo in 2018.

- Support for the development of four regional airports in Saudi Arabia.
- Development of a patented system in partnership with Airborne Concept that monitors mini-UAVs with an ADS-B16 transmitter.
- Safety study and human factors support for the modernisation of the air traffic management and communication, navigation and surveillance systems for Aerothai.

Building

- Olembe sports complex in Cameroon: project management assistance and supervision of works.
- Alexis de Tocqueville Library in Caen: project management.
- Duo Towers in Paris: structural work.
- The group is paricularly involved on the topic of BIM (Building informatio modelling) and SIG(Systeme d'informations geographiques)

Major Works - Water / Environment / Energy

- Citadelle Bridge in Strasbourg, Grand Prix National de l'Ingénierie 2016: design and general contracting.
- Heating network for the ZAC de l'Arsenal in Rueil-Malmaison: turnkey contract in a consortium with Engie-Cofely and project management assignments.

- Waste treatment centre in Romainville: project management
- Offshore extension to the Portier cove in Monaco: project management.

Rail

- Grand Paris Express: infrastructure and systems project management for line 15 East and line 1623
- Extension of the Birmingham Tramway network in England24
- Midland Metro train at Wolverhampton St George's station
- Extension of line B of the Lyon metro to the Lyon Sud hospitals: complete project management25
- Rennes Metro Line B: complete project management

Cities, Roads and Mobilities

- Widening of the A10 motorway in Orléans: project management for the development of a 2 x 4 lane section of the motorway.

consulting

- Support for the Paris 2024 Olympic and Paralympic Games bid.
- The QUAI #3 consortium, of which Egis is the leader, is responsible for the design, production and deployment of all communications relating to the Grand Paris Express

Project structuring and new services This service is one of the historical one for Egis, and it generate more than 30% of the yearly turnover. On Figure 3 we can see the map of the worldwide road assets

- Operation of the Osman Gazi Bridge over the Bosphorus in Turkey
- The Osman Gazi Bridge
- Concession for self-service bicycles in Krakow, Poland
- Pau, Brest and Quimper airport concessions won in 2016

We can have a quick look at some roads projects exploited by Egis around the world on Figure 4

Middle East

- Operation of the new Jeddah airport terminal and development of Abha regional airport in Saudi Arabia
- Zamil Tower in Riyadh, Saudi Arabia: project management33
- Rehabilitation of the Al Karaana lagoons in Qatar: project management

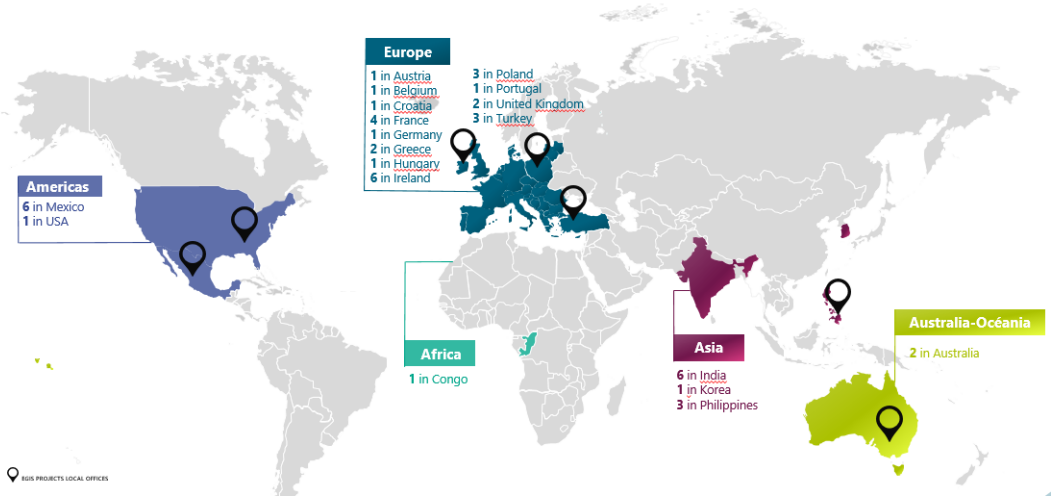

Figure 3: Egis Raod assets in the world

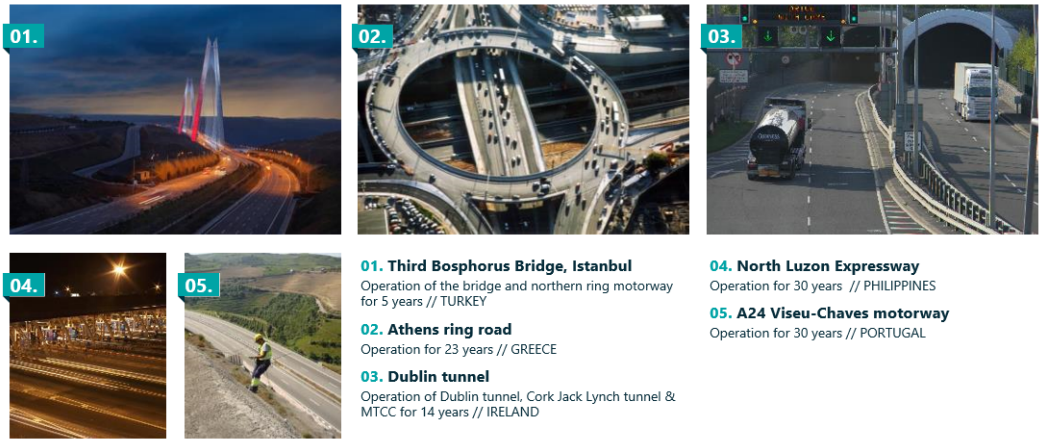

Figure 4: Some of the roads exploited by Egis around the world

India

- Management of the first Smart City programme in Bhubaneswar, India
- Mumbai Metro Line 3: project management
- Development of fishing ports, Gujarat State: project management

Brazil

- Railway lines for the Carajás mining complex in Brazil: project management studies and works
- Drinking water distribution network in the Federal District of Brazil: technical assistance
- Line 13 of the São Paulo metro: technical assistance to the project owner

International regions

- Foumban-Koupa Matapit and Ngaoundere-Paro roads in Cameroon: project management
- Improvement of road safety in Ukraine: technical assistance
- Line 3 of the Guadalajara metro in Mexico: supervision of the work on all systems and rolling stock

3 Motivations and context of the internship

The internship took place in the context of the development of the roadmap data within BU MENS(Project structuring and new services), and the motivations behind the subject can be summarised around the following points:

- As described in the group description section, Egis operates in different business areas and its activities generate a huge amount of data
- As described by the magazine "The Economist" in 2018 data is the new oil. The availability of data and the maturity of the technologies and skills to exploit it are driving companies to place it at the heart of their strategic activities to create value.
- Several attempts to exploit data for data science projects have been made in the past in collaboration with startups. Some had been successful, others not, but these lacked internal follow-up to animate and update the models: hence the need to have in-house skills capable of processing and exploiting large volumes of data.
- It was in this context that the MENS BU decided to set up its dataLab, the aim of which is to equip itself with the necessary skills in the field of data processing and to address all the data needs of the group's various entities, while ensuring the monitoring of projects.

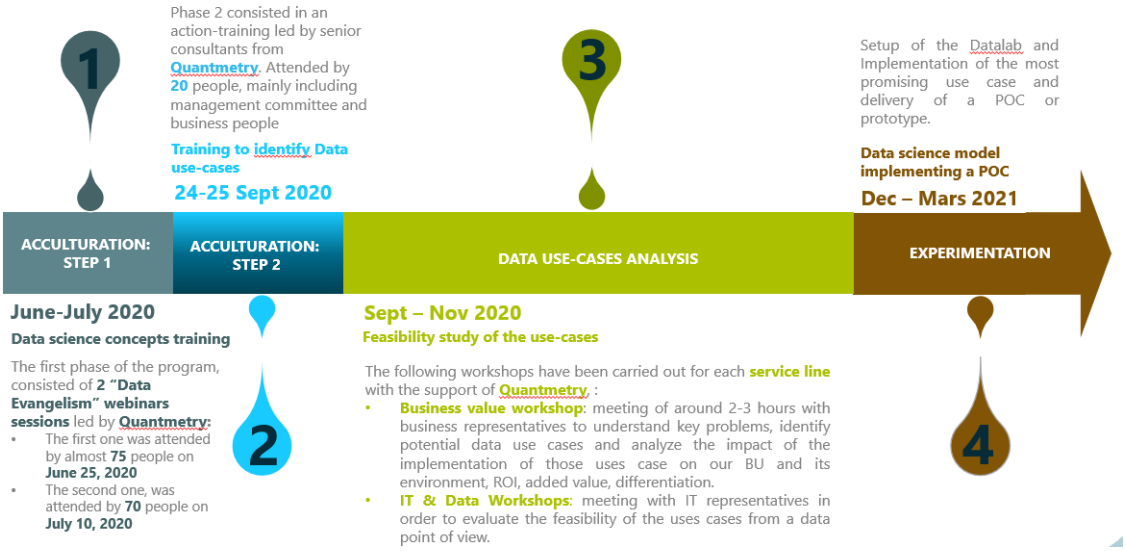

Figure 5: Steps to elaborate the data roadmap of the BU MENS

- The creation of this new entity began with meetings to acculturate the business staff with the aim of presenting them with the importance and the emerging issues related to data.
- This was followed by a macro-mapping of the various data initiatives taken within the group in the past in order to draw conclusions.
- On the basis of this macro-mapping, different use case identification workshops were conducted to build a roadmap of high potential topics.
- Once the use cases were identified, they were prioritised and the most promising one was implemented in the framework of my internship.

The detail about the different steps involved in the elaboration of the roadmap can be found on Figure 5

From the above description, it follows that the first 3 months of my internship consisted of taking stock of the different data science topics that have been conducted in the past and drawing conclusions. The last 3 months instead consisted in implementing a use case foe a specific entity of the group called Easytrip transport services.

4 Related work

Over the past few decades the availability of various marketing-related data such as scanner data and internet data along with organizations' demand for new analytical methods, has spurned an increasing interest in Artificial Intelligence (AI)-based marketing problem solving [2]. Of all the roles that AI-based systems can play in marketing, predictive modeling and more specifically, churn modeling is one of the most promising [3].

Customer churn prediction is the process of calculating the probability of future

churning behavior for each customer in the database, using a predictive model, based on past information/prior behavior [4]. Thus, with the aim of developing an effective customer retention program, the utilized models should be as accurate as possible [5], otherwise these systems would be very wasteful when spending incentive money on customers who will not churn. In this regard, data mining techniques, with their roots in AI, have been widely favored to model churn [6, 3]. The tendency to employ data mining techniques in customer churn prediction stems from the fact that churn is a rare event in a dataset and making an accurate forecast calls for techniques that emphasize predictive ability.

However, studies reveal that, in general, research directions calling for applications of strategic intelligence (i.e. business intelligence, competitive intelligence, and knowledge management) in industrial marketing have received insufficient attention from both academics and practitioners [9, 1]. Accordingly, [9] argue that the application of intelligent systems in handling industrial marketing problems has been limited, which means this research theme has been underdeveloped in business and management journals. Digging deeper to uncover the underlying reasons for this gap, [1] points out that as opposed to the B2C field, the availability of B2B 'big data' is more limited. Thus, mining large datasets to extract knowledge about customers is not as common as in the B2C field. Furthermore, in cases where the data is available, the practices to exploit this data and transform it into information are still underdeveloped in B2B companies. Similarly, by delving more deeply into the existing literature on churn modeling, one also notes that among all studies concentrating on predicting churn across different sectors such as telecommunications [8, 10], online retail [11], finance [6, 12], and retail [13], the majority are within B2C contexts and the application of data mining techniques in B2B churn prediction is still an underdeveloped area. Fortunately, most of the supervized learning techniques used in the context of B2C market can also be applied to B2B market.

Churn prediction belongs to the family of supervised learning problems and as such, several classifiers including decision trees, support vector machines (SVM), artificial neural networks (ANN) ,logistic regression have been identified by machine learning modelers as good candidate in solving this problem [14].

Decistion trees A decision tree is a tree-shaped structure that represents sets of decisions and is able to generate rules for the classification of a data set [16] as described in Figure 6. This technique is suitable for describing sequences of interrelated decisions or predicting future data trends, and is capable of classifying specific entities into specific classes based on feature of entities. Among all existing classification techniques, decision trees are one of the most popular in business since their underlying logic is typically more understandable to managers, and one of the main reasons behind the popularity of decision trees is their transparency and interpretability [15, 7]. This type of classifier is also characterized by a low bias but high variance, which prevents them from generalizing correctly. Although aware of this limitation, this alternative has been explored in this study because it gives results that are easily explained.

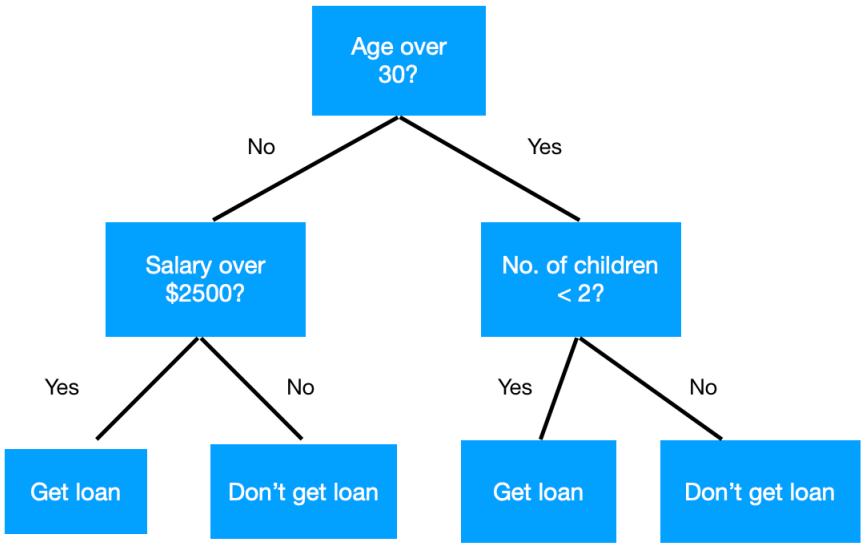

Figure 6: Example of Decision tree output

Logistic regression In cases such as churn prediction, where the dependent variable is binary (e.g. churner as '1' vs. non-churner as '0'), the ordinary linear regression is not applicable as it allows the dependent variable to fall outside the range of 0–1 as shown in Figure 7. Thus, as a special case of general linear models – logistic regression – is favored. Logistic regression models the probabilities for classification problems with two possible outcomes and it's an extension of the linear regression model for classification problems. Ease of use and robustness of results [24] have made logistic regression a popular binary classifier among marketing academics as well as one of the first choice for customer churn modeling [8, 10]. Therefore, in this study, logistic regression has been trained and the obtained performance compared with others classifiers.

Artificial Neural Networks (ANN) Artificial Neural Networks (ANN) are multilayer fully-connected neural nets that look like the Figure 8. They consist of an input layer, multiple hidden layers, and an output layer. Every node in one layer is connected to every other node in the next layer. We make the network deeper by increasing the number of hidden layers. ANNs are very flexible yet powerful deep learning models. They are universal function approximators, meaning they can model any complex function. There has been an incredible surge on their popularity recently due to a couple of reasons: clever tricks which made training these models possible, huge increase in computational power especially GPUs and distributed training, and vast amount of training data. All these combined enabled deep learning to gain significant traction. Although powerful and capable of modeling very complex functions, these models require large volumes of data in order to be efficient and are generally not interpretable. Given the fact that in this study we do not have a large volume of data (about 5000 samples), this approach has not been explored.

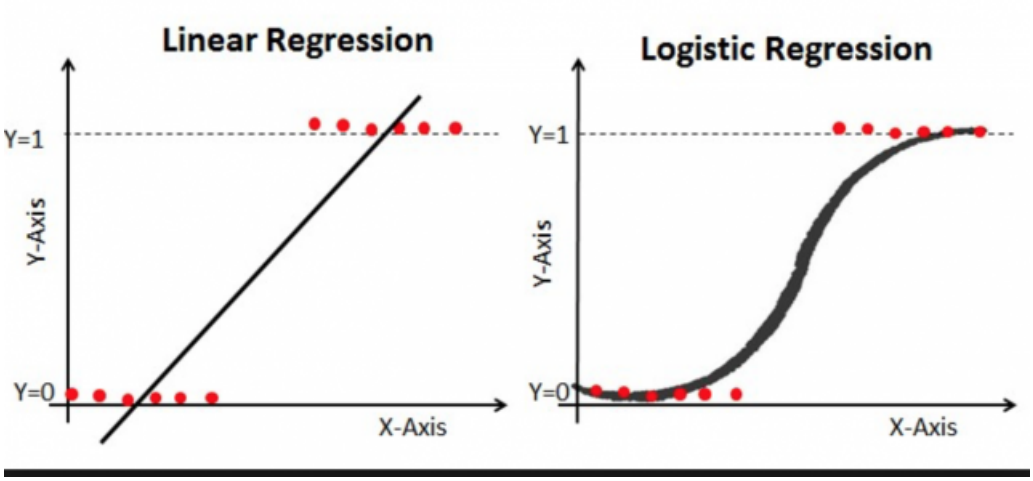

Figure 7: Output of Logistic regression Vs Linear regression

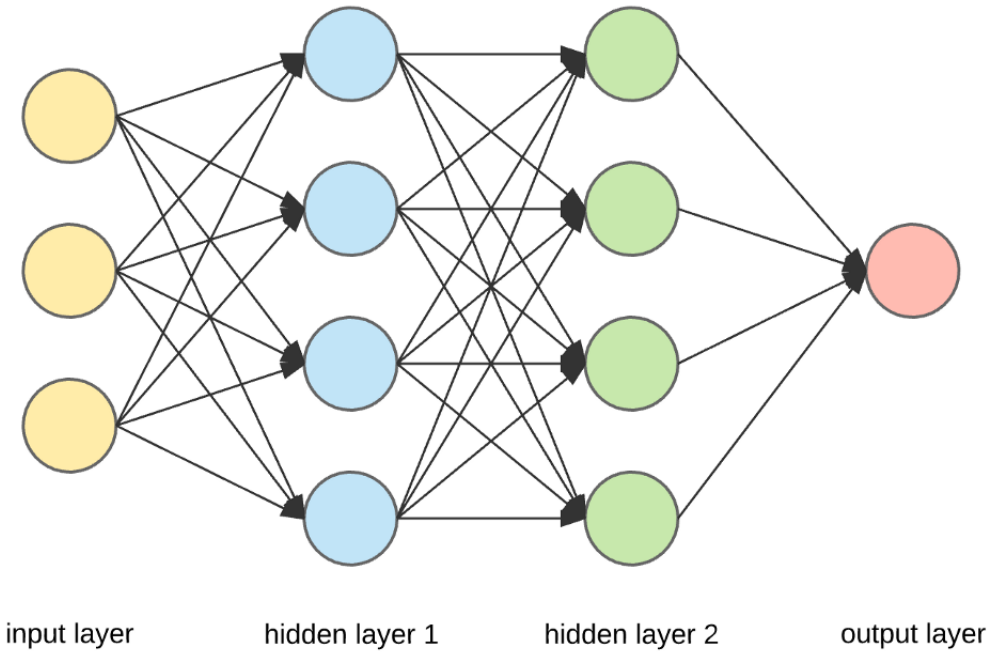

Figure 8: Artificail neural network architechture

Ensemble learning methods Since decision trees are classifiers with low bias but high variance, more sophisticated and complex techniques to reduce variance and improve classifier performance have been introduced, including boosting and bagging methods. Among exiting ensemble learners, the boosting technique is popular due to its outstanding churn prediction capabilities [17, 8]. Basically, the boosting technique manipulates the weight of misclassified instances by attributing more importance to them over multiple training iterations to help the classifier in the classification of instances which are difficult to classify correctly. Several versions of boosting exist, such as logitboost [20], adaptive boosting [21], and brownboost [22]. In this study we used adaptive boosting as it is one of the most well-known and capable boosting techniques. A recent and more efficient implementation of boosting has been proposed by Tianqi chen et al as described in[23], and this implementation is nkown in the data science community, to achieve state-of-the-art results on many machine learning challenges. The idea of aggregating classifiers instead was initially proposed by [19] Breiman (1996) who believed that the combination of several base classifiers can increase the overall accuracy of the aggregated model. In this regard, a class of ensemble learners such as random forests and bagging have been introduced within the data mining stream of churn modeling [19]. One of the disadvantages of boosting methods over bagging methods is that they are not easily parallelizable because the learning of each weak learner is done sequentially, unlike bagging methods where it can be parallelized. However, in the context of this study, given that the data available was small, this was not a limitation for which both alternatives were explored.

Dealing with class imbalance Finally, in modeling customer churn, class imbalance is a common challenge for model developers. In such cases, it is usually the rare class that is of primary interest[17]. In most churn data sets, the number of **non-churners** is greater than the **churners**. Thus, despite the fact that misclassifying the real churners might not have a great impact on model's accuracy, it can cause a costly loss for the companies. Therefore, cost-sensitive learning methods have been utilized by academics to solve the problem of class imbalance in churn prediction [17, 18]. Basically, cost-sensitive learning methods consider the fact that the correct classification of churners has more value than correct classification of non-churners and this is done for a binary classification problem via assigning more cost to false negatives than the false positives [17]. In this study, after workshops with business managers and CEO of the company, it was concluded that the cost of misclassifying a churner was more important than the cost of misclassifying a non churner. Therefore, some learning methods with sensitivity on the cost function have been explored to improve the performance of the algorithm in detecting churners. Another interesting yet powefull technique for dealing with imbalance learning problems is sample handling. The key idea is to pre-process the training set to minimize any differences between the classes. In other words, sampling methods alter the priors distribution of minority and majority class in the training set to obtain a more balanced number of instances in each class [37].

Figure 9: Tools used in the project

5 Environment and libraries

A brief summary of the tools and environement library used in the project are described in Figure 9

5.1 Development environment

Within the framework of this study, the project was carried out following the classic method of the life cycle of a data project as shown in the Figure 10. When working on a data science project, a good part of the time will be dedicated to data analysis and exploration (stages 1, 2, and 3 in Figure 10), so it is important to have the possibility to create graphs but also to add textual comments. The usual development environments such as Pycharm or Spider are not necessarily the most suitable for this stage of the project, hence the importance of jupyter notebooks. The Jupyter Notebook application allows you to create and edit documents that display the input and output of a Python or R language script which are included by default, but with customization, Notebook can run several other kernel environments. Jupyter notebook can be installed and run on local computers as described in [38], and once saved, we can share the files with others. If we do not want to install it locally, we can still use Google Colaboratory which is an interactive computational environment provided by Google. In practice, it is a hosted Jupyter notebook service allowing to combine code execution, rich text, mathematics, plots and rich media (*Google Colab* or just *Colab* for short).

After an evaluation of the quantity and diversity of the different data sources that could be used in this project, the decision was made to carry out all the experimental work locally on a machine with a core i7 processor and a 32 GB of RAM. If the data to be analyzed would have been voluminous, then we would have needed to have access to more powerful and sophisticated tools and services such as a HADOOP clusters or a server with good computing power to facilitate and speed up the calculations.

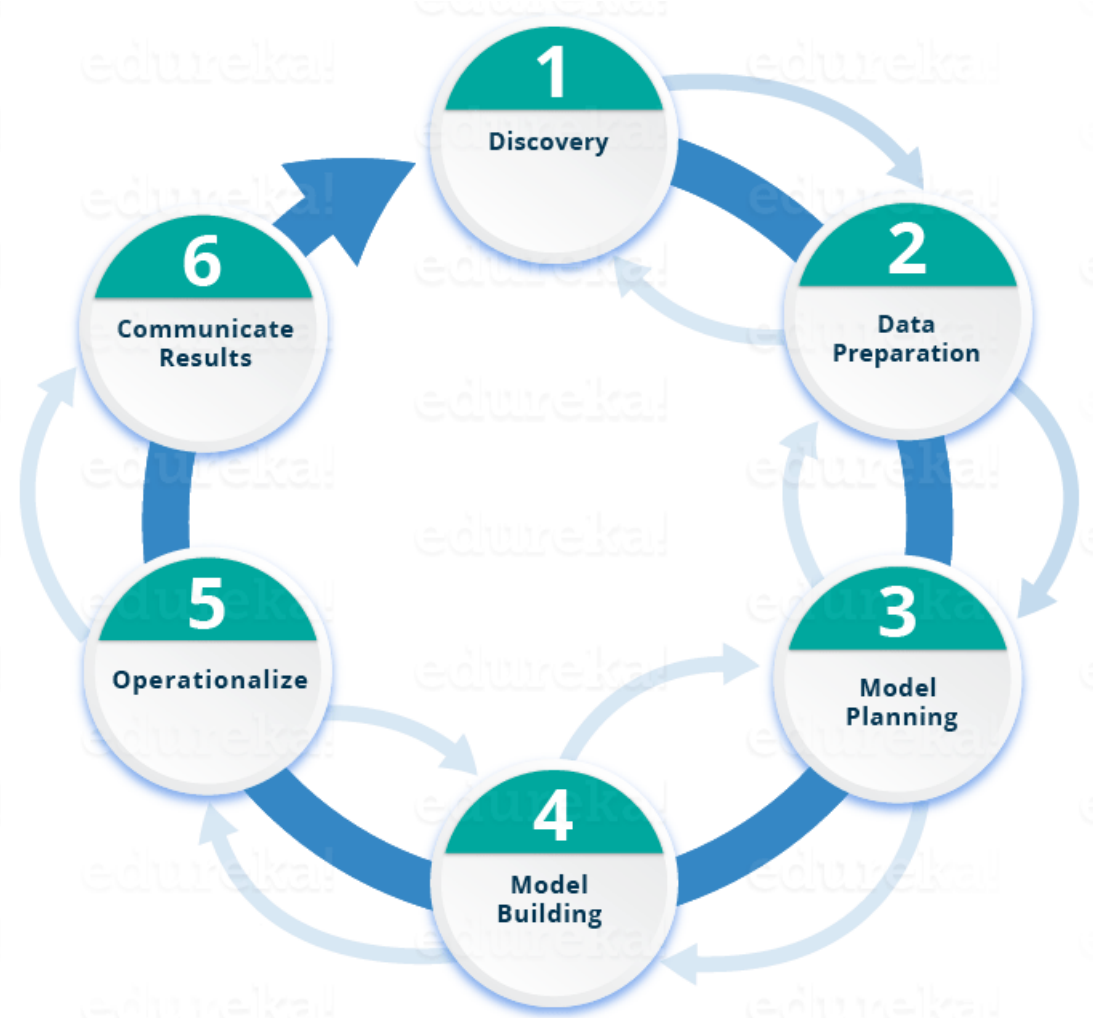

Figure 10: Stages of a data science project

5.2 Tools used to access and query data bases

Since we are talking about a data project, the essence of the project is the data. Contrary to school projects or competitions where the data is usually made available in csv format, in real projects it is usually necessary to connect to different databases to export the data. The data can be made available through APIs, non-relational MongoDB databases or relational databases such as SQL server. In this project, it was only a question of accessing 3 relational databases SQL server and MySQL, that's why we used SQLYog[39] tools to connect to the MYSQL database and SSMS (SQL Server Mangement Studio)[40] to connect to the SQL Server database.

5.3 Programming language and libraries

When it comes to programming in data science, the 2 most popular languages that are usually used are Python and R. Data scientists and programmers like Python because it is a general-purpose and dynamic programming language. Python seems to be preferred for data science over R because it ends up being faster than R with iterations less than 1000. It is also said to be better than R for data manipulation. This language also contains good packages for natural language processing and data learning and is inherently object-oriented. R is better for ad hoc analysis and exploring datasets than Python. It is an open-source language and software for statistical computing and graphics.

In the context of this project, python was choosen due it's flexibility and availability of differents libraries or packages such sklearn [43] to perform data analysis, preprocessing and build machine learning models. As far as graphs are concerned, they were realized using Matplotlib which is a library for plotting and visualizing data in graphical form[42]. The choice of this library is motivated by its ease of use and its full compatibility with Python. To conclude, we inevitably used Numpy library to facilitate the implementation of certain mathematical operations[44]. Since this was a supervised imbalanced learning problem, it was interesting to use over and under sampling techniques provided by the library imbalance learn as described in [35], to improve the performance of the algorithm.

6 Proposed approach

6.1 Modelling

6.1.1 Churn prediction process

The development of a churn model can be done by following a number of steps as described in Figure 11. This process is the one that has been followed and the activities carried out during each step will be described in the following sections.

6.1.2 Churn definition

To define churn, one must first determine whether or not we are in the context of contractual or non contractual services. In the context of contractual services, when a contract expires and is not renewed, then the end date of the contract is considered to

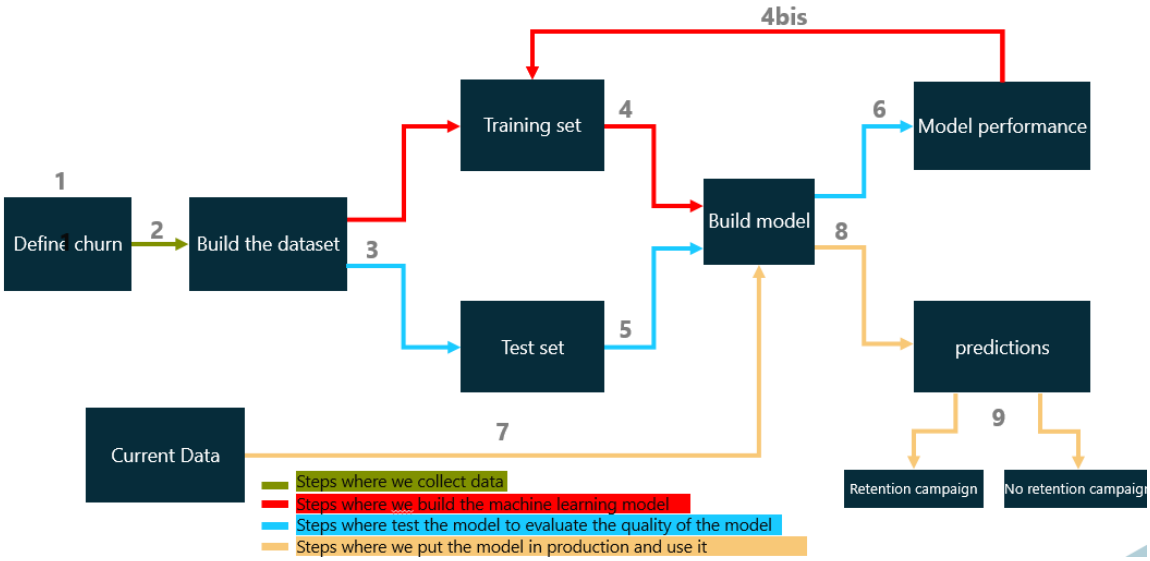

Figure 11: Stages of churn management framework [25]

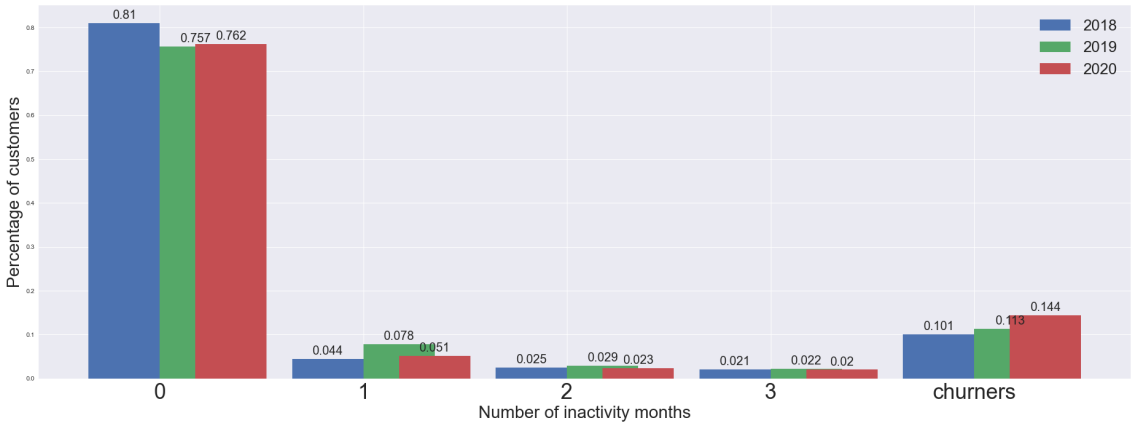

Figure 12: Inactivity graph of customers for years 2018, 2019 and 2020

be the date of churn. On the other hand, defining customer churn in non-contractual settings is complex. In the absence of a contract(s) between the focal company and its customers to be renewed or terminated, it is difficult to estimate the exact time of defection or churn. Because of the inherent difficulty of specifying churn in non-contractual settings, most of the extant churn literature has focused on contractual settings.

Analysts interested in non-contractual settings need to take into account exactly what is meant by "churn". In such settings, since it is probable that a customer returns after a period of inactivity (i.e. 'always-a-share' scenario) non-contractual churn has a different meaning [26, 31]. Due to the difficulty inherent to the definition of churn in a non-contractual context, within the framework of this project and in agreement with business managers, a churner has been defined as a user who has not been invoiced for a period of 3 months. The duration of observation is empirical and strongly depends on the type and field of activity. There are numerous other ways to define churn, but they all have different drawbacks in the context of this sudy. For instance, [27] defined churn based on changes in customers spending from one period to another. We employ a 'change in monetary' variable as a predictor of a churn rather than the definition of the churn itself. Based on the previous agreed definition, we ploted the so called inactivity graph to understand the percentage of customers not being invoiced for more than 3 months on a yearly basis for 3 different years 2018,2019 and 2020 Figure 12. This graph actually allowed to have an estimation of the yearly churn rate in the department and also an estimation of the yearly loss due to churn.

After deciding which definition of churn to use, it's also important to distinguish between voluntary and involuntary churn. The exact definition of these 2 concepts actually depends on the activity domain but in our case, involuntary churn has been defined as customer who churns because they are facing economical difficulties. Whereas voluntary churn has been defined as customers who churns because they are switching to competitor. Based on those 2 definitions, it was important to distinguish between these 2 categories of churners since it was not possible to perform any retention strategies on involuntary churners. That's why the focus have mainly been given to predicting voluntary churners. However, it is necessary to point out that the Figure 12 just gives an estimate of the churn rate and that some users with more than three months of unpaid invoices were found not to be churners. These cases were processed and corrected manually.

6.1.3 Data quality issues and prepossessing techniques

Data preprocessing transforms the raw data into a format that will be more easily and effectively processed for the purpose of the user.The data used in this study came mainly from 3 independent relational databases. These 3 systems contained data from promotional campaigns, sales and personal customer information. Due to the diversity of the systems, the first work in order to build the data set was to analyze these different data sources and identify possible data quality problems such as missing values, incomplete data and duplicate values. Table 2 gives a summary of the different tables than has been analysed in this study.

During the analysis of the data contained in these tables, various data quality problems were encountered. Some were just impossible to process and the use of the data source in question was abandoned. For example, after an analysis of the *Claim* table, it turned out that more than 30% of the historical transactions in this database had zero amounts. After verification with the IT managers, it turned out that some transactions were carried out in another system external to the one taken into account in the analysis. In addition, the customer table containing user information had more than 20% of the users with a creation date not filled in. This problem was solved by approximating the creation date with the date of first transaction for some users, for others it just wasn't possible to solve the problem. Considering that the link between the different tables had to be made on the basis of a unique identifier, it turned out that 10% of the identifiers was inconsistent between the different systems. This problem was therefore solved by defining ad hoc processing functions to normalize the identifiers.

In general when dealing with data issues, there is no general rule or pattern on how to solve them. Other techniques that could have been used in this study included replacing missing actual values with the mean or median where possible, or estimating them using a logistic regression model.

Table Name	**Description**	**Number of attributes**	**Number of rows**
Sales	Table containing the historical sales of 5 products	38	8 Millions
Customer	Table containing all personal customers data	15	44 thousands
Claims	Table containing sales of one specific service	25	500 thousand
Quotes	Table containing historical quotes activity of the sales team	25	5 thousand
Calls	Table containing all call activity of sales with customers	15	100 thousand

Table 2: Description of the different data sources used in this study

6.1.4 Target variable

The target variable in the current study is 'churn' which is defined based on business customers' transnational history in a certain period of time. In order to label each user as chunter or not, it is first necessary to define the length of the history that

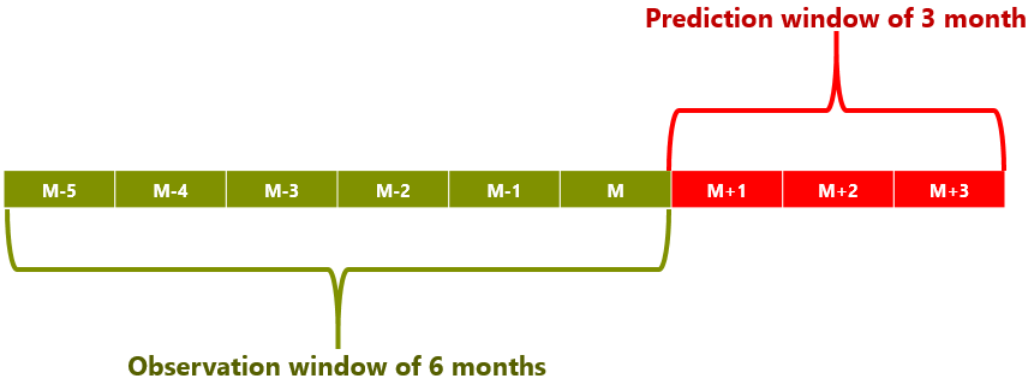

Figure 13: Churn setup to label customers

will be used to train the algorithm. In the context of this study, constraints related to data quality have limited the length of the transaction history to 3 years which are 2018, 2019 and 2020. Given the health situation that prevailed during 2020, the 2020 data was not representative of the behaviour of the company's customers. For this reason, the history used for the construction of the dataset was 2018-2019. On this basis, we defined for each user an observation window and a prediction window as shown in the Figure 13.

A user is labelled as chunter and coded with 1 when he was active during the observation period and inactive during the prediction period. Similarly, a user is defined as not chunter when it was active during the observation period and newly active during the observation period. The choice of the calibration period should not be done randomly because if it's too large, then you may miss to correctly increase the churn probability of a customer who is about to churn because you are looking at a long interval of time, whereas if it's too short, you may end up with too much false positive since the random decreases in revenues will lead a customer to be predicted as churner. In general, this value will be defined empirically making different experiments, observing how the model actually behaves and checking the meaning from a business point of view with the marketing team.

6.1.5 Predictors variables

The 3 main categories of data available for this study were user sales or consumption data, personal data and commercial interactions such as calls and promotional campaigns. The objective of this section is to present how these data were used to create interesting features that would be used to drive the predictive model.

Personal data in terms of personal information about the users, during this analysis we had at our disposal the user's VAT number, residence address, annual income, field of activity and date of start of activity. As mentioned above in the section 6.1.3, some users who did not have a date of registration in the database saw their inscription date estimated by the date of the first transaction. In general,these variables were relatively simple and did not require advanced processing techniques to be used

by the model. Therefore, they were just normalized and inserted as a feature for our model.

Management data The second source of data that could be exploited was information from commercial interactions between sales teams and users. This mainly consisted of follow-up phone calls from users or complaints from users, and the history of promotional campaigns and offers that were made to users. Unfortunately, no features from this data source could be created. Indeed, a detailed analysis of the history over a period of 3 years revealed that there were problems with the sales team filling the database. Therefore, the decision was made to focus on other data sources and use this data source later once the quality issues were resolved.

Behavioural data The last available data source for churn modeling was therefore user transactional data. In this case, it was a question of defining indicators to capture users' monthly consumption behavior. During the definition of these indicators, it is generally important and advisable to have workshops with the marketing department managers who know the business in order to describe what could characterize the churn of a user, and it is on the basis of these inputs that the indicators will be created. In this study, where the users were B2B customers who used a bundle of services and had to be billed on a monthly basis, one would expect a normal user to have a rather stable consumption with few variations. On the other hand, a sudden change in consumption was usually a sign of the beginning of a churn. In addition, the monthly consumption of some users could have considerable variations without being churners, that's why it was also important to check the seasonality of customers from a year to the other one. Having taken into account all these indications from sales representatives, it was therefore a question of creating what are known as behavioural variables. On this basis, recency and frequency of purchases, as well as the magnitude of changes in total spending of customers during a given paeriod, have been chosen as predictor variables to construct the models. Recency, frequency, and monetary variables have been proven to play an undeniable role in predicting customer churn [24, 4] . As noted in [28], the more recent a customer's purchase is, the more likely that the customer is active. In addition, according to [29] frequency of purchases made by a customer can be a measure of defection likelihood in future. Previous studies also suggest that the monetary value of past purchases of a given customer can be an indicator to predict the future behavior [30].

Table 3 and Table 4 gives a description of **Demogrpahic** and **Behavioural** variables that have been created to train the model.In particular,variables in Table 4 are computed considering the following quantities:

1. N : The current year taken into consideration
2. $m(N)_i$: The monthly commissions in the month i and year N for a given customer
3. $\Delta(N)_{i,i+1} = \frac{m(N)_{i+1} - m(N)_i}{m(N)_i}$: The monthly variation in revenues of a customer from month i to month $i+1$.

6.2 Model training

In machine learning, we can distinguish 2 main families of learning problems which are supervised and unsupervised learning. In Supervised learning, you train an algorithm using data which is well "labeled." It means some data is already tagged with the correct answer. It can be compared to learning which takes place in the presence of a supervisor or a teacher. Unsupervised learning on the other hand is a machine learning technique, where you do not need to supervise the model. Instead, you need to allow the model to work on its own to discover information, which means it mainly deals with the unlabelled data.

The churn prediction problem clearly belongs to the family of supervised learning problems where the history of past users is used to train the model to make predictions in the future as discussed in the section 6.1.2. To achieve this goal, there are different types of learning algorithms each with its own advantages and disadvantages as discussed in the section 4. In the context of this study, although characterized by a low bias and high variance as outlined in [20], the decision tree has been used as a benchmark for comparison with other models because it is easily explained and understood. The idea is that we will only move on to more complex and less explainable models if there is a real gain in performance. The results of this model were compared with those of other models such as logistic regression, gradient boosting and extreme gradient boosting. Given the fact that this is an unbalanced learning problem, some of these algorithms as discussed in section 4 have cost sensitive learning techniques to improve the performance of the classifier on the under represented class. In order to improve the performance of the classifier, some of these techniques have been explored to improve the prediction score. In addition, there are other techniques in the literature such as under sampling and oversampling to improve the performance of algorithms in supervised learning as described in [35].

6.3 Model evaluation

6.3.1 Evaluation metric

In general, the evaluation metric can be described as the measurement tool that measures the performance of a classifier. It plays a critical role in achieving the optimal classifier during the classification training. Thus, the selection of a suitable evaluation metric is an important key for discriminating and obtaining the optimal classifier. For binary classification problems, the discrimination evaluation of the best (optimal) solution during the classification training can be defined based on confusion matrix as shown in Figure 14. The row of the table represents the predicted class, while the column represents the actual class.

From this confusion matrix, **tp** and **tn** denote the number of positive and negative instances that are correctly classified. Meanwhile, **fp** and **fn** denote the number of misclassified negative and positive instances, respectively. From Figure 14, several commonly used metrics can be generated as shown in Table 5 to evaluate the performance of a classifier with different focuses of evaluations. As shown in the previous studies [32, 33], the accuracy is the most used evaluation metric in practice either for binary or multi-class classification problems. Through accuracy, the trained classifier

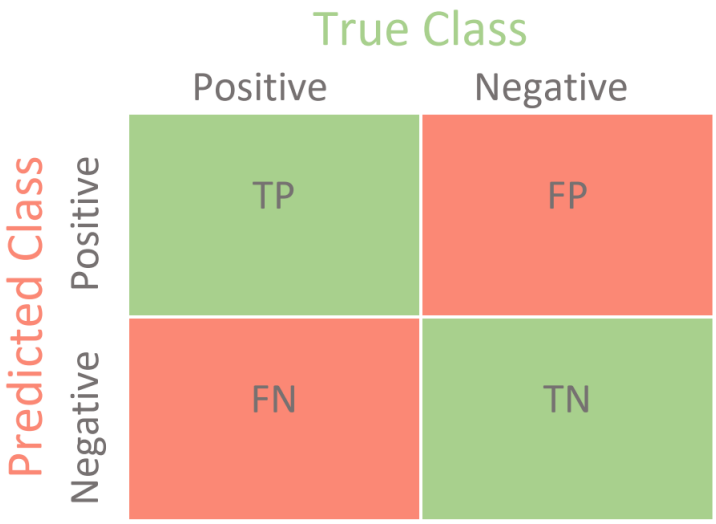

Figure 14: Confusion matrix for binary classification

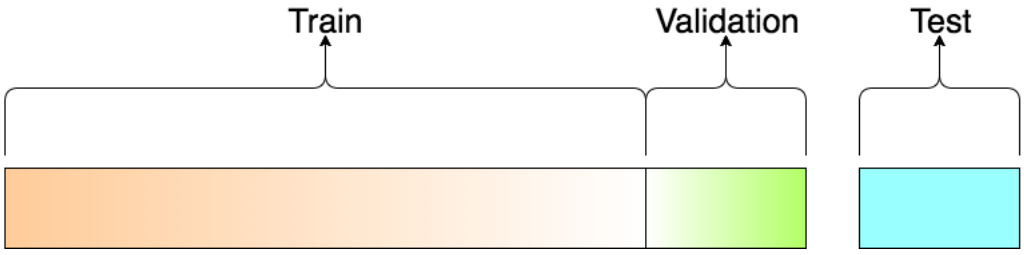

Figure 15: Visualization of the dataset splits

is measured based on total correctness which refers to the total of instances that are correctly predicted by the trained classifier when tested with the unseen data. However, this is still not always reliable, especially when it is a problem of unbalanced learning. For example, imagine a learning problem with a class distribution of 90:10 for negative and positive classes respectively. A classifier that predicts all instances as negative class would have an accuracy of 90% which might seem interesting. However, this model is useless because it does not solve the basic problem of distinguishing between positive and negative class elements.

AUC instead is one of the popular ranking type metrics and unlike the threshold and probability metrics, the AUC value reflects the overall ranking performance of a classifier. The AUC was proven theoretically and empirically better than the accuracy metric [34] for evaluating the classifier performance and discriminating an optimal solution during the classification training. Another interesting metric taken into account when comparing different models in this study was the F measure. As described in Table 5, it's computed considering the precision and recall of the algorithm and allows to distinguish between positive classes during training.

6.3.2 Validation method

In machine learning, model validation is referred to as the process where a trained model is evaluated with a validation data set. In order to validate the quality of the learning, the dataset is split in 3 main parts as described in Figure 15, and in the following we have a description of each part of that dataset:

Train dataset The actual dataset that we use to train the model (weights and biases in the case of a Neural Network). The model sees and learns from this data

Validation dataset The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyperparameters. The evaluation becomes more biased as skill on the validation dataset is incorporated into the model configuration.

Test dataset The sample of data used to provide an unbiased evaluation of a final model fit on the training dataset. The Test dataset provides the gold standard used to evaluate the model. It is only used once a model is completely trained(using the train and validation sets).

In the context of this study, the data has been splitted into 2 — Train and **Test**. After this, the Test set has been kept a part, and we've randomly choosen X% of the Train dataset to be the actual **Train** set and the remaining (100-X)% to be the Validation set, where X is a fixed number(say 80%), the model has then been iteratively trained and validated on these different sets. There are multiple ways to do this, and is commonly known as Cross Validation. Basically we use the training set to generate multiple splits of the Train and Validation sets. Cross validation avoids over fitting and is getting more and more popular, with K-fold Cross Validation being the most popular method of cross validation.

7 Experiments and results

7.1 EDA (Exploratory Data Analysis)

7.1.1 Analysis of the historical length of the sales data and feature engineering

As mentioned in the section 6.1.3, after analyzing the different data sources, the conclusion was that the most reliable and usable data source was the users' billing history. Therefore, for reasons of synthesis and coherence, we will only report some of the analysis carried out on this data and not on the other data sources. The analysis of the transaction history revealed that the history available for analysis spanned 7 years between 2013 and 2020 as shown in Figure 16. However, after checking with the IT services, the gap between the last three years and the previous ones was mainly due to the fact that this first part was not complete. Therefore, all the analyses were performed using the transactions of the last three years.In addition, in view of the year 2020, which has been affected by COVID-19, a comparative analysis was conducted to evaluate the impact of COVID on user consumption. The analysis was conducted by comparing monthly turnover for the three years 2018, 2019 and 2020.

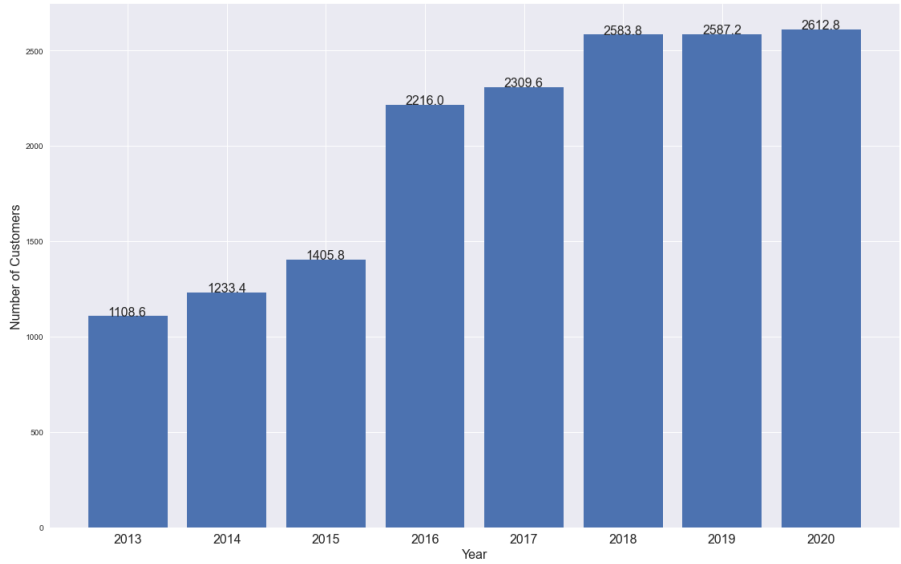

Figure 16: Historical yearly invoiced customers

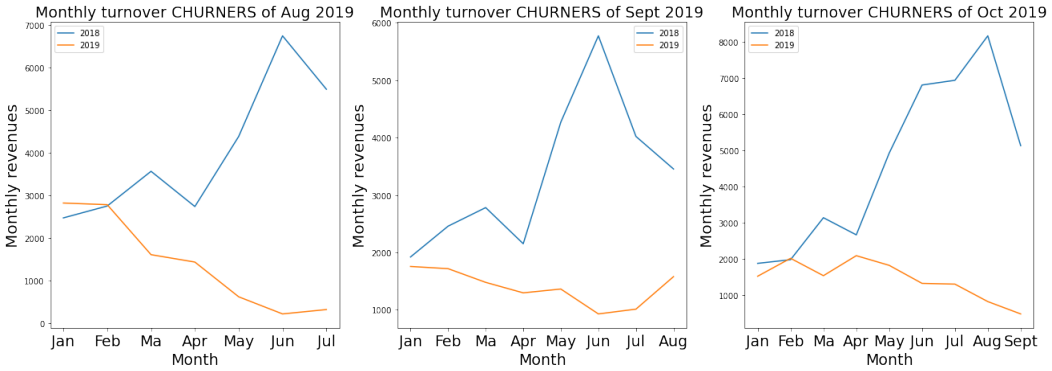

Figure 17: Monthly turnover of customers who churned in different months for 2019

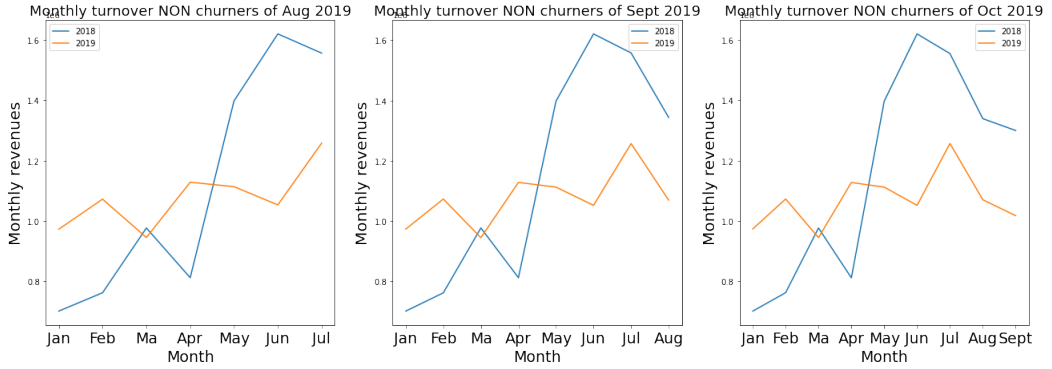

Figure 18: Monthly turnover of customers who did not churned in different months for 2019

The conclusion was that the COVID-19 has mainly affected the users' commissions during the first confinement between February and April 2020. The validity of this observation has been confirmed by the business teams, and transaction invoices from early 2021 confirms a return to normality. If this would not have been the case, then it would have been necessary to use domain adaptation techniques so that the data on which the model is trained is representative of the production data to avoid a learning bias. Again, for reasons of sensitivity, the numerical results of this analysis cannot be disclosed. As discussed in the section 4, from the transactional data, it is possible to extract variables such as Recency, frequency and monetary value which alone hold more than 80% of the predictive power of the machine learning model. However, the creation of these variables remains critical and highly dependent on the type of business and user behavior. Therefore, a detailed analysis of past churners was conducted to identify specific churn behaviors and create indicators accordingly.

A comparative study was conducted by considering for each month **July, August, and September** of the year 2019 the monthly consumption of users who have churned and those who have not churned, and the result is presented on Figure 17 and Figure 18. From those graphs we can draw 3 conclusions:

1. The consumption curve of the churners persists in general a decrease before the churn occurs Figure 18.

2. There is generally a decrease when comparing the user's revenue over the previous year during the same period 17. Based on these observations, indicators have been created to monitor the monthly and periodic consumption variations of users as described in the section 6.1.5.

3. If we take a look at the monthly turnover curve of the non churners Figure 18, we can see that it is not only different from that of the churners but also does not show any particular trend.

7.1.2 Analysis of the products proposed by Easytrip

As mentioned in the section 1, propose a set of 6 products to HGV(Heavy Good Vehicles) as described in Table 1. Since the goals were to predict churn a promote cross sell, after understanding the historical length of data available, the next step was to understand the distribution of customers for each product line. It's important to notice that Easutrip deals with 2 types of customers : DIRECT and INDIRECT customers. In the case of DIRECT customers, thay pruchase the service directly from Easytrip commercial team hence the perceived commission is high, where as in the case of INDIRECT customers, they purchase the service through a channel partner - the perceived commission is therefore shared between Easytrip and the channel parther ans is lower than the one of DIRECT customers. In order to understand the distribution of DIRECT and INDIRECT customers,we need to refer to Figure 19. That figure particularly reveals that even if the number of INDIRECT customers is higher than the one of DIRECT customers, the yearly revenues generated from them is actually lower that the one of DIRECT customers. The main reason is because INDIRECT customers are manager through CHANNEL partners and the commissions need to be shared with them. Moreover, Easytrip's commercial team

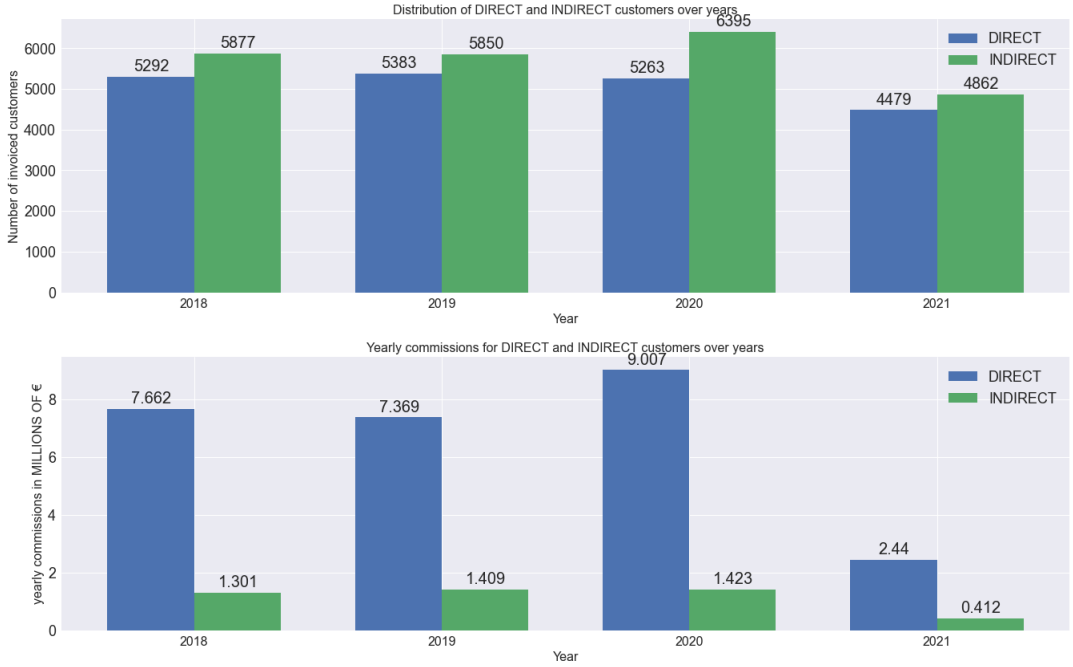

Figure 19: Comparison of DIRECT and INDIRECT customers for easytrip

has no power on INDIRECT customers which means even if we identify a cross sell or churn action to perform on them, we are not legally allowed to do so. That's the reason why all the following analysis have been performed using only DIRECT customers.

The data used to perform this analysis can be summarized in Figure 20

Churn For the churn analysis, the first thing that i had to check was how many customers we have for each product. And as shown in Figure 21, the 2 most important products or services where we have more than 80% of the customer base are TOLL (Inter operable pass) and VAT refund.

I have therefore tried to deepen the analysis by plotting the revenue distribution for each product in order to compare this trend with that of the number of users. The Figure 22 confirm that there is a strong correlation between the number of customers and the yearly turnover generated by the product. Indeed, as the number of customers grows , the yearly turnover will also grow and the most important services are still the same as before. Based on the previous analysis and observation, the decision was taken to build the churn algorithm only for TOLL and VAT services since they are the ones with most of the customers generate more than 90% of the yearly turnover of the company. Hence, all the following analysis only focus on the TOLL service which is at the end of the day the most important among all product lines.

Cross sell All the results of the analysis for the churn part have been used to draw some preliminary conclusions about the cross sell part. In particular, since TOLL is the most important service, the main object of the segmentation part will be to identify potential candidate that may have a high probability of buying other services. A futher analyssi that have been done was to plot the number of customers purchaising 1, 2 or more services. Figure 23 reveals most of the customers(more

1 - CUSTOMER PROFILE & FINANCIAL

What really matters are information about the customer(**Direct and Indirect**) such as:

- **PROFILE**
  - Company name
  - Addresses
  - Contacts
  - Subscription date
  - Number of trucks
  - Activity domains
  - Guaranty's type
  - Contract's information
  - Etc…
- **FINANCIAL INFORMATIONS**
  - Turnover
  - Risk level
  - Payment Terms
  - Account type
  - Etc..
- **SUBSCRIBED SERVICES**
  - For each customer, we to know the list of subscribed services.
- **GUARANTY**
  - Since some customers may prefinance their activity and need to provide some guaranty, it could be a good idea to consider that in the analysis.

2 - BILLING HISTORY

Since customers are invoiced each month, here we want to have for each customer the history of invoices since the date of subscription for each service:

- **INVOICES**
  - Company name
  - Date
  - Due date
  - Payment method
  - Currency
  - Amount
  - Related account
  - Supplier
  - Etc…

NOTE: we are interested here in the invoices sent to the customer with respect to a service consumed and not in the transaction invoices coming from a provider in the case of TOLL for example.

3 - QUOTES & CALLS ACTIVITY HISTORY

- the activity of sales team is recorded and stored, the goal here is to analyse the past quotes sent to each customer in order to have compute some statistics, segment and categorize customers based on the response rate, calls rate, etc..

Figure 20: Summary of data used to perform the analysis

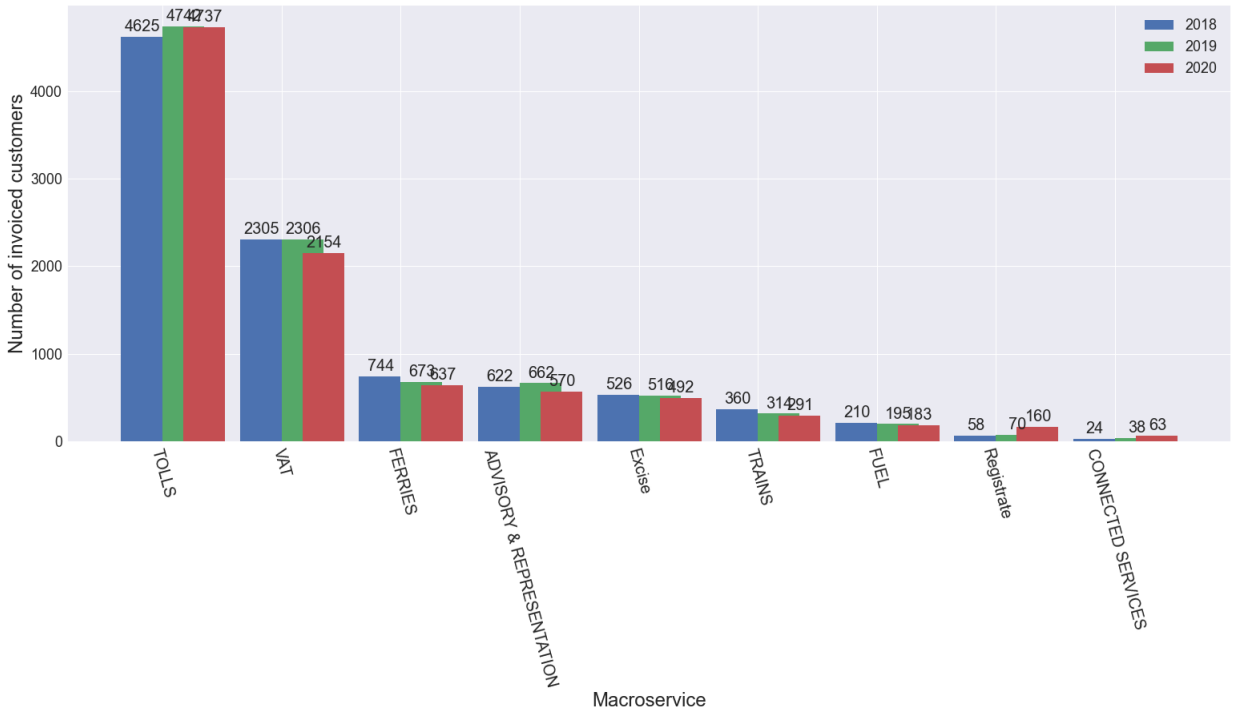

Figure 21: distribution of customers by product

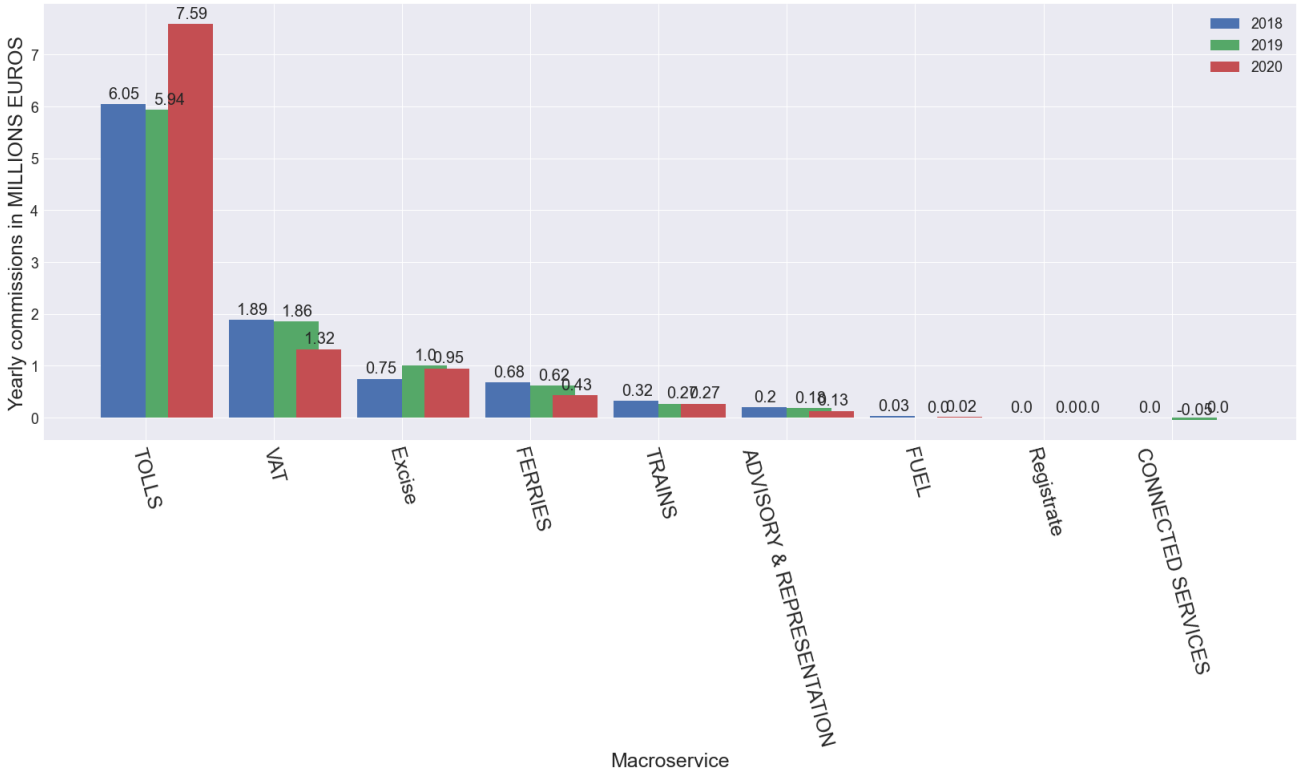

Figure 22: Distribution of revenues by product

than 60%) purchase only one service (TOLL) and less than 22% pruchase 2 services. The goals of the cross sell is there to increase the purchased product by customer, increasing the average basket size of each customer and also the customers stickiness.

7.1.3 Pareto's law for easytrip

The Pareto principle states that for many outcomes, roughly 80% of consequences come from 20% of the causes (the "vital few"). Other names for this principle are the 80/20 rule, the law of the vital few, or the principle of factor sparsity.

Management consultant Joseph M. Juran developed the concept in the context of quality control, and improvement, naming it after Italian economist Vilfredo Pareto, who noted the 80/20 connection while at the University of Lausanne in 1896. In his first work, Cours d'économie politique, Pareto showed that approximately 80% of the land in Italy was owned by 20% of the population. The Pareto principle is only tangentially related to Pareto efficiency. Mathematically, the 80/20 rule is roughly described by a power law distribution (also known as a Pareto distribution) for a particular set of parameters, and many natural phenomena have been shown to exhibit such a distribution. It is an adage of business management that "80% of sales come from 20% of clients". This law can basically be applied in different domains as described in [45] such as:

Computing In computer science the Pareto principle can be applied to optimization efforts. For example, Microsoft noted that by fixing the top 20% of the most-reported bugs, 80% of the related errors and crashes in a given system would be eliminated. Lowell Arthur expressed that "20 percent of the code has 80 percent of

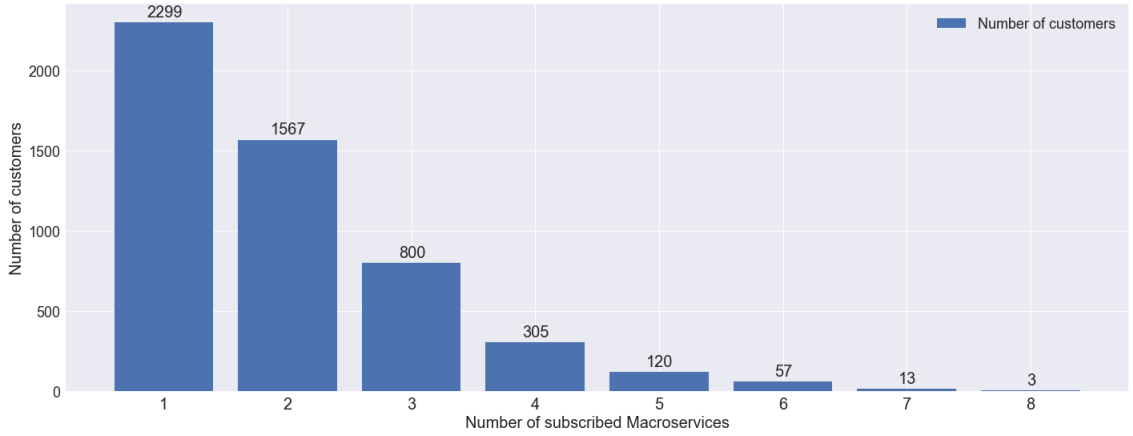

Figure 23: Customers breakdown by number of purchased services

the errors. Find them, fix them!" It was also discovered that in general the 80% of a certain piece of software can be written in 20% of the total allocated time. Conversely, the hardest 20% of the code takes 80% of the time. This factor is usually a part of COCOMO estimating for software coding. WordPerfect and other software developers identify what customers want most of the time and how they want to do it: the 80/20 rule (people use 20 percent of a program's functions 80 percent of the time). Software developers work to make high-use functions as simple and automatic and inevitable as possible.

Sports It has been argued that the Pareto principle applies to sport, where leading players often take the majority of wins. For instance in baseball, the Pareto principle is reflected in Wins Above Replacement (an attempt to combine multiple statistics to determine a player's overall importance to a team). "15% of all the players last year produced 85% of the total wins with the other 85% of the players creating 15% of the wins. The Pareto principle holds up pretty soundly when it is applied to baseball." It has been suggested (but not tested) that the principle applies to training, with 20% of exercises and habits having 80% of the impact, suggesting trainees should reduce the variety of training exercises to focus on this effective set.

Occupational health and safety Occupational health and safety professionals use the Pareto principle to underline the importance of hazard prioritization. Assuming 20% of the hazards account for 80% of the injuries, and by categorizing hazards, safety professionals can target those 20% of the hazards that cause 80% of the injuries or accidents. Alternatively, if hazards are addressed in random order, a safety professional is more likely to fix one of the 80% of hazards that account only for some fraction of the remaining 20% of injuries. Aside from ensuring efficient accident prevention practices, the Pareto principle also ensures hazards are addressed in an economical order, because the technique ensures the utilized resources are best used to prevent the most accidents.

Engineering and quality control The Pareto principle has many applications in quality control where it was first created. It is the basis for the Pareto chart, one of the key tools used in total quality control and Six Sigma techniques. The Pareto principle

serves as a baseline for ABC-analysis and XYZ-analysis, widely used in logistics and procurement for the purpose of optimizing stock of goods, as well as costs of keeping and replenishing that stock.In engineering control theory, such as for electromechanical energy converters, the 80/20 principle applies to optimization efforts. In the systems science discipline, Joshua M. Epstein and Robert Axtell created an agent-based simulation model called Sugarscape, from a decentralized modeling approach, based on individual behavior rules defined for each agent in the economy. Wealth distribution and Pareto's 80/20 principle emerged in their results, which suggests the principle is a collective consequence of these individual rules

Software testing The Pareto principle in the context of software testing is commonly interpreted as "80% of all bugs can be found in 20% of program modules. In other words, a half of the modules may contain no bugs at all. Applying Pareto Principle to quality control activities of a software can help reduce the testing time and increase the efficiency of the system, but the application of the principle itself will require good analytical and logical skills.

Health and social outcomes In health care in the United States, in one instance 20% of patients have been found to use 80% of health care resources. The Dunedin Study has found 80% of crimes are committed by 20% of criminals.[31] This statistic has been used to support both stop-and-frisk policies and broken windows policing, as catching those criminals committing minor crimes will supposedly net many criminals wanted for (or who would normally commit) larger ones. However, this principle has proven false in practice, as over 90% of citizens victimized by stop and frisk policies were found not to have committed any crime[citation needed]. The principle was erroneously applied, and instead residents were targeted by race, having little impact on crime[citation needed]. Improved economies overall have had a far greater correlation with lowering crime rates[citation needed]. Some cases of super-spreading conform to the 20/80 rule, where approximately 20% of infected individuals are responsible for 80% of transmissions, although super-spreading can still be said to occur when super-spreaders account for a higher or lower percentage of transmissions. In epidemics with super-spreading, the majority of individuals infect relatively few secondary contacts. The 80/20 rule has been suggested to account for a large proportion of transmission events during the ongoing COVID-19 pandemic.

General distribution operations The Pareto principle is often referred to in distribution operations, normally called the 80-20 rule. In distribution operations it is common to observe that 80 percent of the production volume constitute 20 percent of the SKUs (Stock Keeping Units). During facility design, this rule often governs the storage area and processing area configurations.

Product lines Many video rental shops reported in 1988 that 80% of revenue came from 20% of videotapes. A video-chain executive discussed the "Gone with the Wind syndrome", however, in which every store had to offer classics like Gone with the Wind, Casablanca, or The African Queen to appear to have a large inventory, even if customers very rarely rented them.

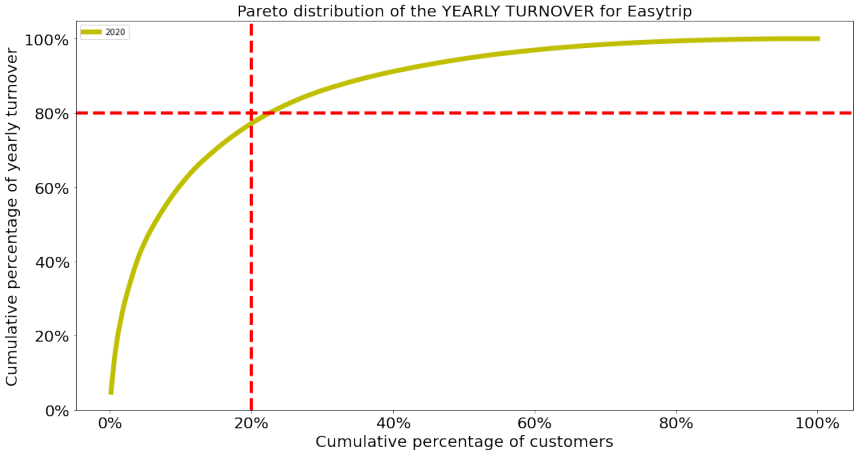

Figure 24: Pareto's Law for Easytrip

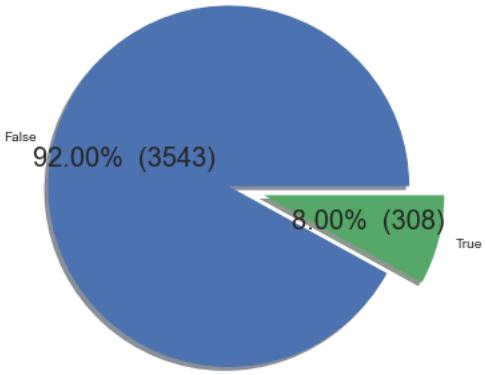

Figure 25: Churn rate distribution in the dataset

In Study In school, students usually have been asked to do a lot of work to achieve high marks, but depend on the personal experiment which illustrate that the 80% of the results could be achieved with 20% of the work, some suggestions are explained about the way of take notes, and how profesors' questions focus on the specific knowledge in the course, to make this practical, and to know how that might be suitable for you, try to select two ways of studying, and notice which one is more productive, because the way of study math is different of the way of study biology.

Having that idea in mind, i tried to check if this observation was also true in the case of easytrip especially for the TOLL product. The Figure 24 shows that the distribution of revenues for th TOLL service follows a pareto law. Indeed, we can notice that 80% of the yearly turnover is generated by only 21% of the customers. From a churn perspective, this basically means that this is the part of customers we do not want to loose, and this business will be taken into account when optimizing the machine learning model.

7.1.4 Churn rate distribution in the dataset

After defining the key variables, preprocessing the data and labelling each customer as churner and non churner, the next step was then to perform an analysis to understand the distribution of churn in our dataset. As shown in Figure 25, the churn rate in the

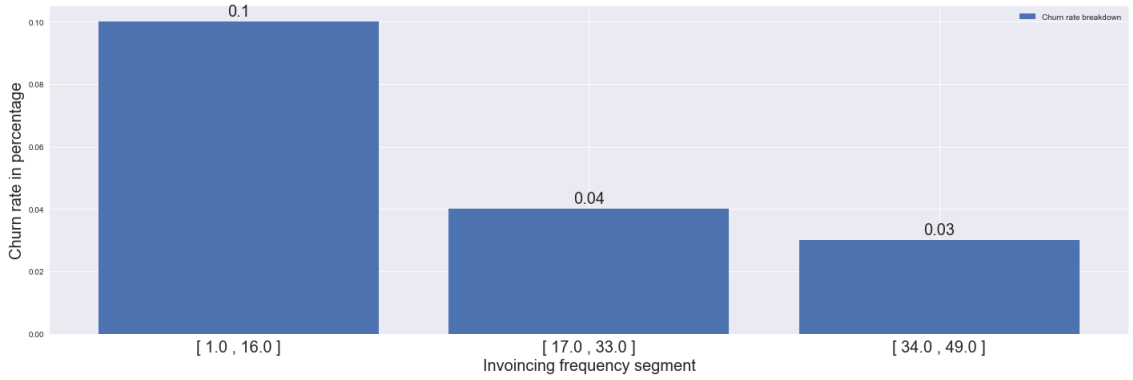

Figure 26: Churn rate breakdown by invoicing frequency

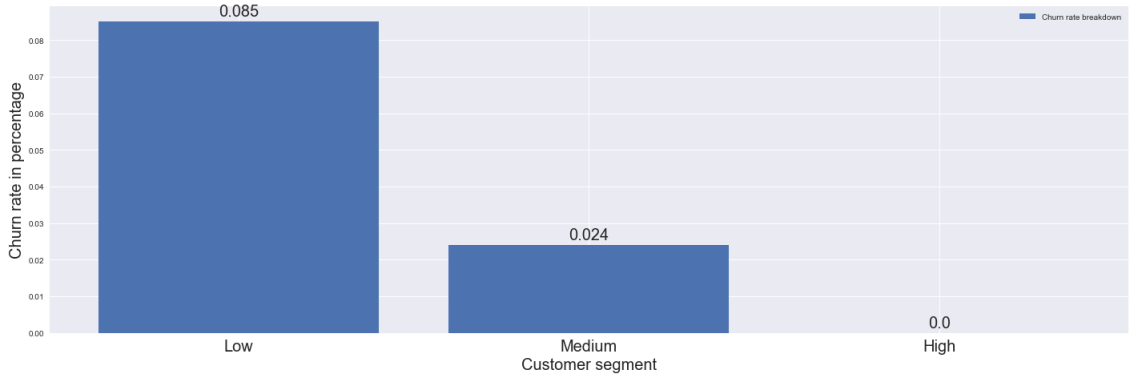

Figure 27: Churn rate breakdown by customer segment

dataset was only **8%**, hence we were dealing with an imbalance classification problem as described in [17]. In the section 7.5, we will present and discuss the results obtained using cost sensitive tehchniques designed to tacle this type of issue.

7.1.5 Bivariate Analysis

After creating the different variables, i also tried to perform different type of bivariate analysis to understand the impact of each variable on the target output. When analyzing the transaction history, I found that users were billed differently. Some were billed on a monthly basis, others twice a month or once every 1 or 2 months. And as I expected, the churn rate is different depending on the billing frequency of the user, indeed the more the annual billing frequency increases, the more the churn rate decreases as described in Figure 26

Churn rate by invoicing frequency

Churn rate by revenues segment After different workshops with business managers, rules were established to decide how to group users into different segments. After doing so, the Figure 27 shows that the churn rate is higher within the user population that takes with a low income as opposed to those with a high income where the churn rate is practically null in 2019. This same analysis has also been performed for other years and the result was basically the same.

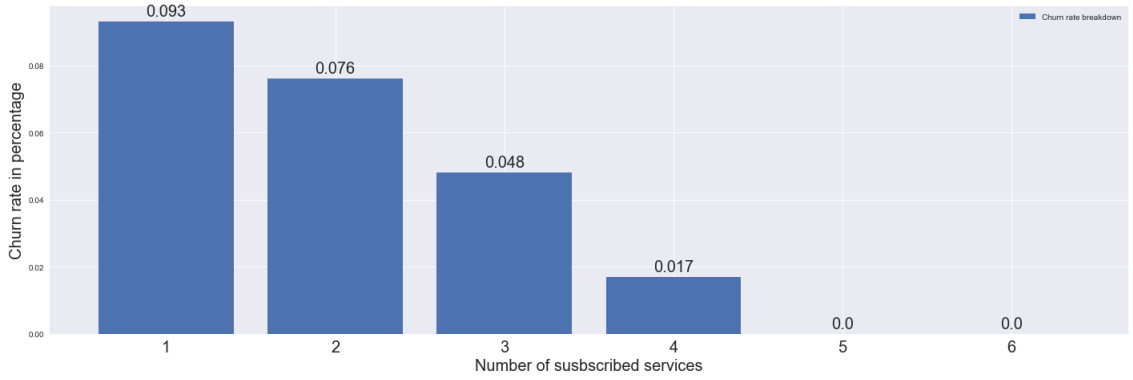

Figure 28: Churn rate breakdown by number of services

Churn rate by number of services Since users can subscribe to different services as described in the section 1, it was also interesting to analyze the churn distribution in relation to the number of services subscribed to. The rest is rather interesting because as shown in Figure 28, as the number of services increases, the churn rate decreases. This translates from a business point of view by the fact that when a user takes several, it increases his loyalty. This is also a justification of the incentive of cross-selling policies to retain customers. Several other analyses were conducted in order to evaluate the correlation between independent and dependent variables, but to keep the report synthetic and understandable, here we decided to report some that seemed relevant.

7.2 Results obtained using Demographics features

After collecting and analyzing the different data sources and dealing with any data quality issues, we first created the demographic variables as described in Table 3 in order to evaluate the performance of the model obtained by using these variables. As shown in Table 6, the accuracy obtained is around 92% for all the models. This may look quite interesting,but as discussed in the section 6.3 this score is misleading.Indeed, remembering the churn rate in the dataset shown in Figure 25 which is around 8%, we may conclude that even a naive classifier predicting all samples to be in the negative class will have an accuracy of 92%. This is mainly the reason why wen looking at the other metrics in Table 6, we can see that they are particularly low, especially the precision in the positive class which is equal to zero, meaning that none of the classifiers was able to predict a churner. In order to better understand what is going on, we can have a look at the confusion matrix of the different classifiers obtained making predictions on the testset as explained in the Figure 15.

As shown in Figures 29,30,31,32, the tp is always equal to zero 0, which means the classifier is not learning anything. The main reason behind this is because the real predictive power of the churn algorithm will mainly come from **Behavioural features** as explained in [4].Since i tried different optimization techniques but did not manage to improve the accuracy of the models, the conclusion after this first iteration was to start thinking of how to create some features allowing to capture the consumption trend of the customer on a monthly basis. In the following, we will

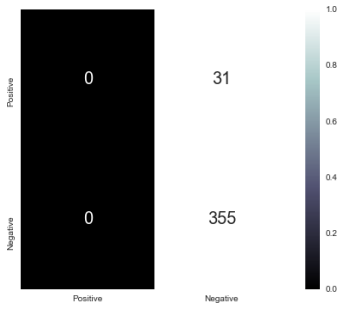

Figure 29: confusion matrix for Decision Tree on test set

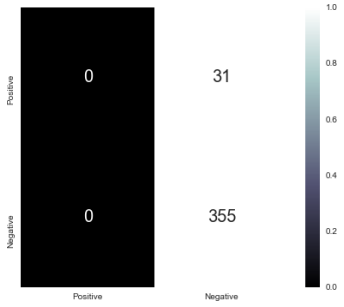

Figure 30: confusion matrix for Logistic Regression on test set

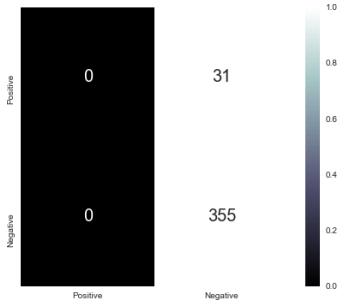

Figure 31: confusion matrix for Random Forest on test set

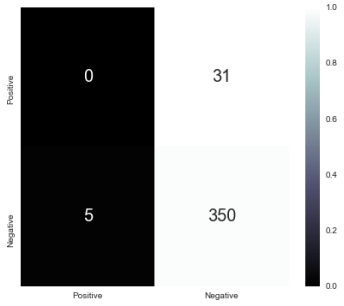

Figure 32: confusion matrix for gradient boosting on test set

see how the performance of the classifier actually changes when adding some basic behavioural variables.

7.3 Results obtained using basic feature engineering features

After the first iteration of testing the power of demographic variables for predicting churn, the conclusion was that they were not informative enough to allow the classifiers to learn how to distinguish true positives from false positives. Therefore, in this section, we will present and discuss the performance obtained by creating fairly simple and intuitive behavioral variables.The definition of behavioral variables for churn prediction is a rather delicate process and strongly depends on the type of business we are dealing with. For example, in the e-retail industry, given the fact that users' purchases are seasonal and do not necessarily follow a specific frequency, certain variables such as clumpiness and periodicity of purchases have been created for the churn modeling by [11].

In the telecom industry on the other hand, [10] has created a kind of graph that allows to monitor the frequency of calls from a user to other users of the same network in order to monitor its average consumption. However, these indicators must be defined according to business intuitions and must have a meaning for them, because after all the algorithm will be used as a tool to improve their efficiency. In the context of this study, given that users are billed on a monthly basis, it was therefore a question of defining indicators to monitor the monthly consumption of users. Referring to Figure 13 where we distinguish between an observation period and a prediction period, the indicators are defined for all users in the observation window. The first challenge was to define the length of this observation window. Indeed, a window that was too long (e.g. 1 year) was not very informative and gave too many false negatives because the focus was not put on the critical period of revenue decrease preceding the churn as shown in Figure 17.

On the other hand, a much too short window gave rise to too many false positives because users with small cascading declines were predicted to be churned by the algorithm. Since this is a variable that is generally estimated experimentally, different tests were conducted and the 6 month window proved to be adequate, which is why all behavioral indicators were defined for all users over a 6 month observation period.Once the observation window has been chosen, the first indicators that have been created are **consecutiveIncrease, and consecutiveDecrease** as described in the table 4. The performance of the classifier obtained by using the indicators is represented in the Table 7

As illustrated in the table Table 7, we notice a slight improvement of the accuracy of cross validation compared to the previous iteration and in particular of the AUC and F measure metrics. This means that the different models this time have learned something, and succeed in distinguishing some churners from non churners. This result can be confirmed by examining the confusion matrices of the different models obtained by making predictions on the test set as shown on the figures. The main observation that can be made is that the true positive number is no longer zeros as

before, which means that the model starts to learn to distinguish churners from non churners. This observation proves that we are moving in the right direction and in the next section we will try to further improve these performances by adding more intelligent behavioral variables.

7.4 Results obtained improving the feature engineering

So far we have managed to verify in the Section 7.2 that the demographic variables alone did not have a great predictive power on the model, which is why we have moved on in the next Section 7.3 to the definition of basic behavioral variables that have slightly improved the performance of the model. In this section, the goal is to continue the previous initiative and define even more intelligent and discriminative variables. Indeed, the previous variables were quite general because they just captured the general consumption trends of the user without quantifying this variation. For example, a user who goes from 100€ to 10€ monthly and another who goes from 100€ to 98€ monthly would both have the **consecutiveDecrease** variable increased by 1. However, the decrement is more important in the first case than in the second. This is why the behavioral indicators such as **minVariation and averageVariation** have been defined as shown in the Table 4. After defining all these variables, the modes have been trained again and the performances are shown in Table 8. We can see again a slight improvement of the performance of the different models, and in particular of the gradient boosting and random forest models. This proves the importance and the added value of these new variables in the discrimination of churners and non churners. So far, there are still some problems that have not been properly addressed, and will therefore be the subject of the next section.

7.5 Results obtained using cost sensitive methods

As mentioned at the end of the previous section, there is still one aspect of our problem that has not been explicitly taken into account so far. Indeed, if we recall the Figure25, we will realize that it is an unbalanced learning problem where the most important class (the churners) is under represented with only 8%. However, when training the different models obtained in the previous sections, they would consider the two classes as having the same importance, which is obviously not true. In this section, we propose to overcome this problem by providing additional information to the different models during training in order to allow them to give more importance when they make an error on the positive class which is more important. This technique is generally called learning with cost as described in the Paragrpah 4, and considering that the library sklearn [43] has been used for the training of the different models, the API of the different models contains attributes such as **classWeight** for Logistic Regression and Random Forest or **scalePosWeigth** for Gradient Boosting allowing to achieve this goal. The results obtained by considering this additive information are summarized in the Table 9.

As shown in the Table 9, except for the decision tree which has difficulty in achieving good results due to the complexity of the problem, all the other computational models have performances that have improved significantly compared to those observed in the Table 8. In particular, gradient boosting was the best performer of all the different models, followed by random forest. This is not so surprising knowing

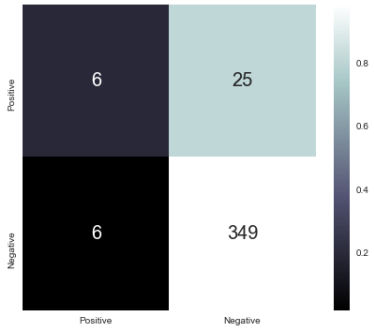

Figure 33: confusion matrix for Decision Tree on test set with basic feature Engineering

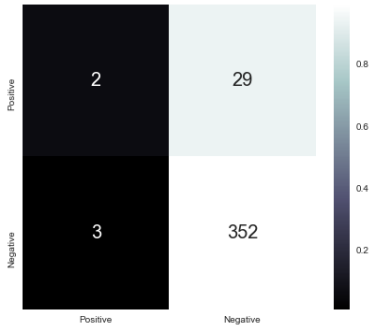

Figure 34: confusion matrix for Logistic Regression on test set with basic feature Engineering

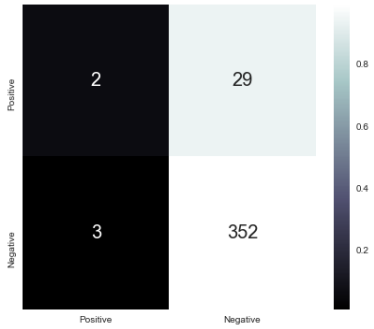

Figure 35: confusion matrix for Random Forest on test set with basic feature Engineering

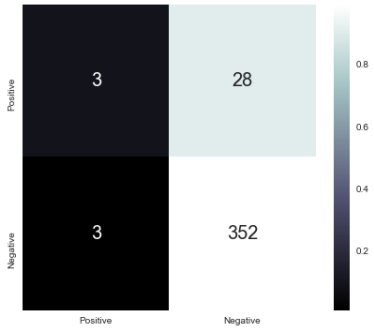

Figure 36: confusion matrix for gradient boosting on test set with basic feature Engineering

that the ensemble methods in general besides being robust, allow to obtain good performances because they are based on weak learners with low bias especially decision trees and thus minimize the global variance by creating several trees. The results obtained during the cross validation can be further confirmed looking at the confusion matrices of the different models obtained making predictions on the test set as shown in Figures 37, 38, 39, 40

Although the random forest and the extreme gradient boosting have similar performances and slightly oriented towards the extreme gradient boosting, the real reason why the extreme gradient boosting has been definitively preferred to the random forest is that it has a much better performance on the test as presented on the Figures 39, 40.One question that might come to mind right now is to understand what the model learns, and why it classifies a user as churner or non churner. This is generally important and especially in the context of churn because for a sales manager to decide to implement a customer retention action, he should at least know why. This is the question we will answer in the next section.

8 Model explanability

Even if in the previous section decision trees did nor provide good results, one of it's advantages as discussed in the section 4, is that we can plot the tree and have a look at it in order to understand which variables have been used to make predictions and how the model actually works as shown in Figure 41. At the contrary of this simple model, gradient boosting is more complex and robust model, but which is not explainable. In the context of reliable AI, one of the pillars of which is based on the interpretability of decisions made by a model, different works have been done to produce tools to understand what machine learning algorithms learn. Thus, the gradient boosting module in sklearn has a function allowing to plot the importance attributed to each variable by the model as illustrated on the Figure 42. This graph shows that the most important variables for the prediction of churn are the behavioral variables, while the demographics variables have a rather marginal contribution. However, one piece of information that is missing from this study is how the probability of churn varies when the values of these different variables changes. Another useful technique to understand the impact of different variables on the predictions of a machine learning model is the Shap graph [36]. The goal of SHAP is to explain the prediction of an instance x by computing the contribution of each feature to the prediction as illustrated in Figure 43. We can see again that the most important features are behavioural ones and the impact on the model output is rather coherent with what we were expecting.In fact we can say that the more the first attribute **min6monthsVariations** decreases, the more the probability of churn increases, on the contrary the more it increases, then the probability of churn decreases. This means from a business point of view that the more the user has a good performance during the last 6 months, the less we have to worry, but the more he will have a very negative performance, the more we will have to worry because the customer have may be about to leave.

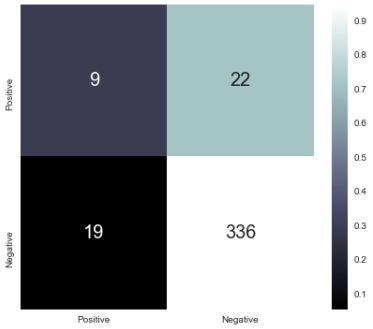

Figure 37: confusion matrix for Decision Tree on test set with **Cost sensitive learning**

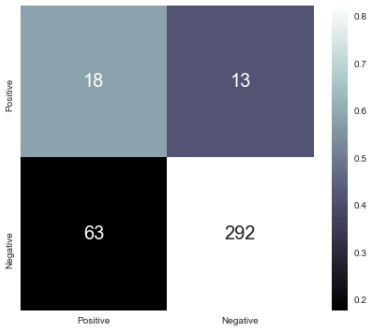

Figure 38: confusion matrix for Logistic Regression on test set with **Cost sensitive learning**

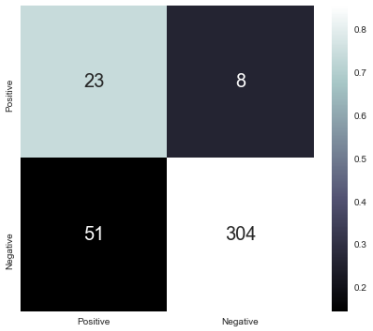

Figure 39: confusion matrix for Random Forest on test set with **Cost sensitive learning**

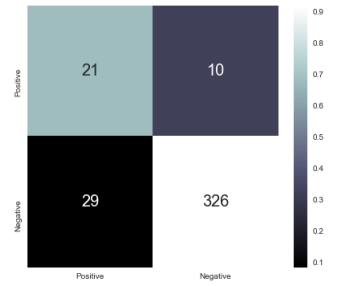

Figure 40: confusion matrix for gradient boosting on test set with **Cost sensitive learning**

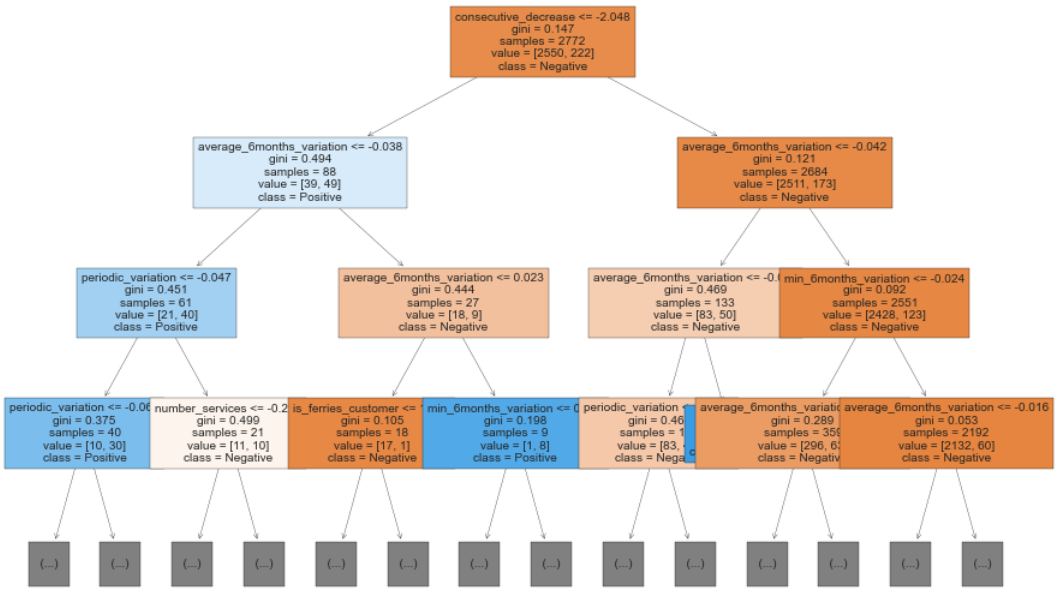

Figure 41: Graph of the decision tree built during training

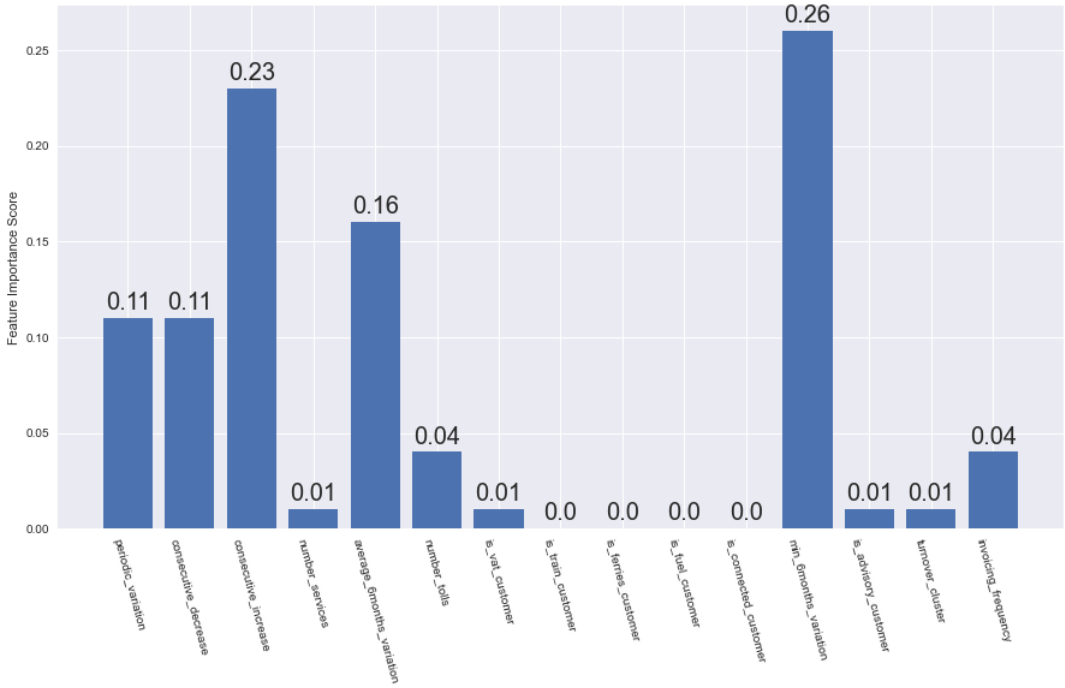

Figure 42: Feature importance graph of the extreme gradient boosting model

Figure 43: Shap graph to explain feature importance of Extreme gradient boosting algorithm

9 Cross sell strategies

As mentioned in the introduction, in parallel to the churn subject, we also explored unsupervised learning techniques to segment customers and improve the efficiency of marketing campaigns.However,after different workshops with business managers to understand what kind of segmentations they needed to create, and after an analysis of the feasibility, it was concluded that the best technique to meet this need was mainly the definition of user groups based on business rules. So unlike the churn prediction topic where data science and machine learning techniques were extensively explored and used, in this project of segmentation, the bulk of the work consisted of understanding the business need, identifying the data sources needed to meet it, collecting and processing them, and being a little creative in defining the format for the results.Furthermore, another reason that prevented the use of unsupervised learning techniques was the scarcity of personal user information. However, given that this is a B2B context and that for each user we have the **VAT number**, there are external databases(open data) that allow us to collect user information such as the number of employees, size, field of activity, etc. This is an alternative that, although not explored in this study for reasons of time, could later be considered.

10 Conclusion

Given the context of this use case, which was part of the first use case of a data organization that is under construction, I had the opportunity at the very beginning to participate in different training seminars led by senior consultants of the consulting firm specialized in DataScience Quantmetry, in different workshops of identification and framing of data use cases within the business unit. This allowed me to familiarize with certain methodologies of acculturation of business people to data science

subject, and understand how to lead use case identification workshops. Moreover, I benefited during the internship of the punctual support of a senior data scientist of Quantmetry to support me from a methodological point of view, and to unblock certain technical subjects.

The work allowed to build a machine algorithm to predict the users who are going to churn with 1 month in advance, and also to define customer segmentations to identify groups of users with specific needs. As part of this study, i had the opportunity to work on the lifecyle of a data science project, managing from start to finish the different steps related to data collection, analysis and processing, including the construction and quality evaluation of the algorithm. It is interesting to note that a large part of the time (around 75%) was spent on data collection, analysis and processing, which is completely in line with the distribution of time spent by a data scientist on different phases of a data science project. The reasons are mainly due to the fact that unlike school projects where data are usually available in CSV format, in real use cases they have to be queried in generically heterogeneous databases, and usually present huge data quality problems that have to be analyzed and processed. Based on the principle of *Garbage in, Garbage out*, it is important to take care of the data that are given into the model if we want to obtain good results.

After conducting different analysis and testing different algorithms, the gradient boosting algorithm was selected as the best candidate because of the best cross-validation result and also because of its robustness. Although performing incredibly well on the test set, there are other methods to improve the performance of the classifier such as under and oversampling as described in [35]. However, due to time constraints these techniques have not been explored and would be an alternative to test in order to improve the performance of the algorithm. On the other hand, the choice was made to test this algorithm in production on billing data from 2021 in order to evaluate the relevance and consistency of the predictions. The first test results were quite conclusive and positive, as some of the users who had been predicted as churners have been confirmed as having already churned by the business managers.

To finish, one of the main challenges we are currently facing is to solve, in collaboration with the IT teams, the frequency of availability of monthly invoices. Indeed, one of the assumptions of the model in order to make forecasts for a specific month M, is to have these invoices available at the latest at the end of this month in order to make the forecasts for the following month (M+1). If the IT team cannot find a solution, an alternative is to improve the algorithm's predictions and move from M+1 to M+2. Once this is done, the next step will be to move to the industrialization of the project, which unlike IT projects requires answering more difficult and delicate questions. Indeed, within the framework of a web application for example, once the front end, back end and data base are developed, these components can simply be embedded in containers, deployed on servers after the configuration of characteristics such as RAM, disk space, directory etc., then the application exposed through an API to make it usable. On the other hand, in the framework of Data projects, once the experimentation phase is over, the first questions that need to be answered mainly concern the necessary IT infrastruce, in particular the data infrastructure (DataLake, datawareHouse), deployment on premise or in the cloud depending on the costs, se-

curity and privacy constraints, and the model's update frequency. In addition to this, we must also think about implementing a data governance strategy to ensure the quality of data that are given into the model. These are therefore critical points that will have to be addressed later as part of an eventual industrialization of this data use case.

11 References and Figures

References

[1] *F. Wiersema. The B2B Agenda: The current state of B2B marketing and a look ahead, Industrial Marketing Management, 42 (4) (2013), pp. 470-488.*

[2] *A. Orriols-Puig & F.J. Martínez-López & J. Casillas & N. Lee : Unsupervised KDD to creatively support managers' decision making with fuzzy association rules: A distribution channel application, Industrial Marketing Management, 42 (4) (2013), pp. 532-543.*

[3] *B. Wierenga & J. Casillas & F. Martínez-López (Eds.) : Marketing and artificial intelligence: Great opportunities, reluctant partners & Marketing intelligent systems using soft computing, Vol. 258, Springer, Berlin Heidelberg (2010), pp. 1-8.*

[4] *K. Coussement & D. Van den Poel: Customer churn prediction in the online gambling industry: The beneficial effect of ensemble learning, Journal of Business Research, 66 (9) (2013), pp. 1629-1636.*

[5] *K. Coussement & K.W. De Bock:Integrating the voice of customers through call center emails into a decision support system for churn prediction Information & Management, 45 (3) (2008), pp. 164-174.*

[6] *H. Risselada & P.C. Verhoef & T.H.A. Bijmolt: Staying power of churn prediction models, Journal of Interactive Marketing, 24 (3) (2010), pp. 198-208.*

[7] *C.-P. Wei & I.-T. Chiu: Turning telecommunications call details to churn prediction: A data mining approach Expert Systems with Applications, 23 (2) (2002), pp. 103-112.*

[8] *A. Lemmens & C. Croux: Bagging and boosting classification trees to predict churn Journal of Marketing Research, 43 (2) (2006), pp. 276-286.*

[9] *F.J. Martínez-López & J. Casillas :Artificial intelligence-based systems applied in industrial marketing: An historical overview, current and future insights Industrial Marketing Management, 42 (2013), pp. 489-495.*

[10] *S.A. Neslin & S. Gupta & W. Kamakura & L. Junxiang& C.H. Mason:Defection detection: Measuring and understanding the predictive accuracy of customer churn models Journal of Marketing Research, 43 (2) (2006), pp. 204-211.*

[11] *X. Yu & S. Guo & J. Guo & X. Huang:An extended support vector machine forecasting framework for customer churn in e-commerce Expert Systems with Applications, 38 (3) (2011), pp. 1425-1430.*

[12] *D. Van den Poel & B. Larivière:Customer attrition analysis for financial services using proportional hazard models European Journal of Operational Research, 157 (1) (2004), pp. 196-217.*

[13] *W. Buckinx & D. Van den Poel:Customer base analysis: Partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting European Journal of Operational Research, 164 (1) (2005), pp. 252-268.*

[14] *J. Han & M. Kamber & J. Pei:Data mining: Concepts and techniques (3rd ed.), Morgan Kaufmann (2011).*

[15] *S. Olafsson & X. Li & S. Wu:Operations research and data mining European Journal of Operational Research, 187 (3) (2008), pp. 1429-1448.*

[16] *S.J. Lee & K. Siau:A review of data mining techniques Industrial Management and Data Systems, 101 (1) (2001), pp. 41-46.*

[17] *Burez and Van den Poel, 2009: Handling class imbalance in customer churn prediction Expert Systems with Applications, 36 (3) (2009), pp. 4626-4636.*

[18] *G.M. Weiss: Mining with rarity: A unifying framework Sigkdd Explorations, 6 (1) (2004), pp. 7-19.*

[19] *L. Breiman: Bagging predictors Machine Learning, 24 (2) (1996), pp. 123-140.*

[20] *J.H. Friedman & T. Hastie & R. Tibshirani: Additive logistic regression: A statistical view of boosting The Annals of Statistics, 28 (2) (2000), pp. 337-374.*

[21] *Y. Freund & R.E. Schapire: Experiments with a new boosting algorithm, Paper presented at the 13th International Conference on Machine Learning, Bari, Italy (1996).*

[22] *Y. Freund : An adaptive version of the boost by majority algorithm Machine Learning, 43 (3) (2001), pp. 293-318.*

[23] *T.Chen & C.Guestrin : A Scalable Tree boosting, KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*

[24] *W. Buckinx & D. Van den Poel :Customer base analysis: Partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting, European Journal of Operational Research, 164 (1) (2005), pp. 252-268*

[25] *Datta P & Masand B &Mani DR & Li B : Automated cellular modeling and prediction on a large scale. Issues on the Application of Data Mining 2001; 485–502.*

[26] *R.T. Rust & K.N. Lemon & V.A. Zeithaml: Return on marketing: Using customer equity to focus marketing strategy Journal of Marketing, 68 (1) (2004), pp. 109-127.*

[27] *V.L. Miguéis & D. Van den Poel & A.S. Camanho & J. Falcão e Cunha : Modeling partial customer churn: On the value of first product-category purchase sequences Expert Systems with Applications, 39 (12) (2012), pp. 11250-11256*

[28] *C. Wu & H.-L. Chen : Counting your customers: Compounding customer's in-store decisions, interpurchase time and repurchasing behavior European Journal of Operational Research, 127 (1) (2000), pp. 109-119*

[29] *Reinartz & Kumar, 2000:On the profitability of long-life customers in a non-contractual setting: An empirical investigation and implications for marketing Journal of Marketing, 64 (4) (2000), pp. 17-35*

[30] *D.C. Schmittlein & R.A. Peterson:Customer base analysis: An industrial purchase process application Marketing Science, 13 (1) (1994), pp. 41-67*

[31] *R. Venkatesan & V. Kumar: A customer lifetime value framework for customer selection and resource allocation strategy Journal of Marketing, 68 (4) (2004), pp. 106-125.*

[32] *N.V. Chawla, N. Japkowicz and A. Kolcz : "Editorial: Special issue on learning from imbalanced data sets", SIGKDD Explorations, 6 (2004) 1-6.*

[33] *Q. Gu, L. Zhu and Z. Cai: "Evaluation Measures of the Classification Performance of Imbalanced Datasets", in Z. Cai et al. (Eds.) ISICA 2009, CCIS 51. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 461-471.*

[34] *J. Huang and C. X. Ling,: "Using AUC and accuracy in evaluating learning algorithms", IEEE Transactions on Knowledge Data Engineering, 17 (2005) 299-310.*

[35] *Guillaume.L, Fernando.N, Christos K. Aridas : Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning.*

[36] *Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in Neural Information Processing Systems. 2017.*

[37] *Mohamed Bekkar, Taklit Akrouf Alitouche : "Imbalance data learning approches review." International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.4, July 2013 DOI : 10.5121/ijdkp.2013.3402 15.*

[38] *https://docs.anaconda.com/ae-notebooks/user-guide/basic-tasks/apps/jupyter/*

[39] *https://webyog.com/product/sqlyog*

[40] *https://docs.microsoft.com/en-us/sql/ssms/download-sql-server-management-studio-ssms?view=sql-server-ver15*

[41] *www.research.google.com/colaboratory/faq.html*

[42] *www.matplotlib.org*

[43] *www.scikit-learn.fondation-inria.fr*

[44] *www.numpy.org*

[45] *https://en.wikipedia.org/wiki/Pareto_principle*

Variable Name	Type	Description	Formula
Country	Demographic	The country of the customer	None
Local office	Demographic	The office of the customer's accountmanager	None
Number of services	Demographic	Number of subscribed services by the customer	sum of all services
isFerriesCustomer	Demographic	boolean flag indicating if the customer take this service	None
isTrainsCustomer	Demographic	boolean flag indicating if the customer take this service	None
isTollCustomer	Demographic	boolean flag indicating if the customer take this service	None
isVatCustomer	Demographic	boolean flag indicating if the customer take this service	None
isExciseCustomer	Demographic	boolean flag indicating if the customer take this service	None
isConnectedCustomer	Demographic	boolean flag indicating if the customer take this service	None
isAdvisoryCustomer	Demographic	boolean flag indicating if the customer take this service	None

Table 3: Description of **Demographic** variables created to train the model

Variable Name	Type	Description	Formula
consecutiveIncrease	Behavioural	Variable counting the number of times customers monthly invoices have been increasing in the past **6** months	$count_{i=1,6}\Delta(N)_{i,i+1}$ where $\Delta(N)_{i,i+1} > 0$
consecutiveDecrease	Behavioural	Variable counting the number of times customers monthly invoices have been decreasing in the past **6** months	$count_{i=1,6}\Delta(N)_{i,i+1}$ where $\Delta(N)_{i,i+1} <= 0$
averageVariation	Behavioural	Variable describing the average variation in monthly commissions of the customer in the past 6 months **6** months	$\frac{\sum_{i=1}^{6}\Delta(N)_{i,i+1}}{5}$
minVariation	Behavioural	Variable describing the minimum variation in monthly commissions of the customer in the past 6 months **6** months	$min_{i=1,6}(\Delta(N)_{i,i+1})$
PeriodicVariation	Behavioural	Variable describing for a given period in a year, the precentage of variation with respect to the previous year and same period	$\frac{\Delta(N)_{i,i+1}-\Delta(N-1)_{i,i+1}}{\Delta(N-1)_{i,i+1}}$
InvoicingFrequency	Behavioural	Variable describing the number of time the customer is invoiced on yearly basis	None

Table 4: Description of **Behavioural** variables created to train the model

Metric	Formula	Evaluation Focus
Accuracy (acc)	$\frac{tp+tn}{tp+tn+fp+fn}$	in general, the accuracy metric measures the ratio of correct predictions over the total number of instances evaluated.
Sensitivity (sn)	$\frac{tp}{tp+fn}$	This metric is used to measure the fraction of positive patterns that are correctly classified
Specificity (sp)	$\frac{tn}{tn+fp}$	This metric is used to measure the fraction of negative patterns that are correctly classified.
Precision(p)	$\frac{tp}{tp+fp}$	Precision is used to measure the positive patterns that are correctly predicted from the total predicted patterns in a positive class
Recall (r)	$\frac{tp}{tp+fn}$	Recall is used to measure the fraction of positive patterns that are correctly classified
F-Measure(FM)	$\frac{2*p*r}{p+r}$	This metric represents the harmonic mean between recall and precision values
AUC	$\frac{S_p-\frac{n_p*(n_n+1)}{2}}{n_p*n_n}$	the AUC value reflects the overall ranking performance of a classifier : S_p is the sum of the all positive examples ranked,n_p and n_n denote the number of positive and negative examples respectively

Table 5: Different methods for Classification evaluations

Model	Accuracy	AUC	F-measure
Decision tree	0.919+- 0.007	0.490+-0.005	0
Logistic regression	0.919+-0.008	0.551+-0.006	0
Random forest	0.919+-0.008	0.527+-0.006	0
Gradient boosting	0.906+-0.005	0.526+-005	0

Table 6: **5 fold cross validation score** of the model trained using **ONLY demographic** features

Model	Accuracy	AUC	F-measure
Decision tree	0.9186 +- 0.006	0.7507 +- 0.0509	0.2942 +- 0.0458
Logistic regression	0.931 +- 0.0056	0.8438 +- 0.0311	0.3648 +- 0.1
Random forest	0.929 +- 0.003	0.8474 +- 0.0277	0.2745 +- 0.0532
Gradient boosting	0.9307 +- 0.0038	0.861 +- 0.0286	0.3578 +- 0.0702

Table 7: **5 fold cross validation score** of the models trained using **Basic behavioral** features

Model	Accuracy	AUC	F-measure
Decision tree	0.9027 +- 0.0058	0.6923 +- 0.0105	0.4053 +- 0.0227
Logistic regression	0.9293 +- 0.0055	0.8391 +- 0.0402	0.3592 +- 0.0986
Random forest	0.9267 +- 0.0021	0.912 +- 0.0098	0.4098 +- 0.0334
Gradient boosting	0.9273 +- 0.0061	0.9123 +- 0.0119	0.4309 +- 0.0618

Table 8: **5 fold cross validation score** of the models trained using **Improved behavioral** features

Model	Accuracy	AUC	F-measure
Decision tree	0.9027 +- 0.0058	0.6923 +- 0.0105	0.4053 +- 0.0227
Logistic regression	0.824 +- 0.012	0.8445 +- 0.0375	0.3972 +- 0.0283
Random forest	0.8563 +- 0.0088	0.9194 +- 0.0094	0.4693 +- 0.0192
Gradient boosting	0.8975 +- 0.0089	0.9197 +- 0.0101	0.5226 +- 0.0277

Table 9: **5 fold cross validation score** of the models trained using **ALL** features with **cost sensitive** techniques