POLITECNICO DI TORINO

Master's Degree in ICT for Smart Societies

Master's Degree Thesis

Forecasting Public Transport Demand using Smart Cards Data

Supervisors
Prof. Silvia CHIUSANO
Dr. Elena DARAIO
Dr. Brunella CAROLEO

Candidate
Eleonora GASTALDI

October 2021

Abstract

The collection of mobility data through the validation of electronic tickets and smart cards allows to obtain personal information about users and their mobility patterns.
Having this knowledge available, it is possible to forecast the passengers' demand which is fundamental to optimize the allocation of resources (personnel and vehicles), the network planning, the frequency setting and therefore to reduce operating costs.

Granda Bus consortium [1] provides about 10 million smart card validations referring to the whole year 2019 in the Piedmont area, in the North-West of Italy. The study, conducted at the Links Foundation [2], exploits these data to answer the following research question: *"What is the estimated public transport demand at one bus stop for a selected route, given a specific day and time slot?"*.
To address this unknown, a methodology has been designed, developing the whole KDD process. It opens with a preliminary data analysis useful to understand the quality and the integrity of the data and to identify the best way to process them. The couple *bus stop-route* is the core element of the analysis: this choice is justified by the fact that the bus stops can have several routes, each one characterized by its own target of users and therefore trend of validations which differs significantly one from the other.
A clustering process has been applied to all the couples *bus stop-route*, based on the number of validations and importance of the offer to detect a set of representative couples. The study focuses on them and compares the performance of the different selected machine learning techniques.
In particular, the predictive models selected to conduct the analysis are: Average and Median Response, Random Forest Regressor, Gradient Boosted Decision Tree, Support Vector Regression and SARIMA.

The obtained results show that a temporal segmentation is needed, since the validations trend changes according to the period of the year, in correlation with the schools opening or closing. For each segment and cluster, the best machine learning model has been identified.

Keywords: mobility data analysis, smart card data, public transport demand, forecasting mobility demand, machine learning, Random Forest, GBDT, SVR, SARIMA.

Table of Contents

List of Tables

List of Figures

Acronyms

AI
: Artificial Intelligence

RF
: Random Forest

GBDT
: Gradient Boosted Decision Tree

SVR
: Support Vector Regression

MAE
: Mean Absolute Error

MASE
: Mean Absolute Scaled Error

Chapter 1

Introduction

1.1 Context

The forecasting of the public transport demand is a crucial point for the service provider: it allows to improve the offer and, consequently, to increase the profit but it represents a complex task.

The difficulty lies in having to take into account a huge number of factors that influence the prediction and its variabilities over time, such as the weather conditions, the zone in which the stop is located, the day of the week, if it is a working day or a holiday day.

The higher the variability of these factors is, the more useful is to estimate the number of passengers that will board in a specific place and time.

In particular, forecasting the demand during the peak hours when it quickly increases would give to the service provider a helpful tool to understand how to manage the fleet of vehicles, avoiding overcrowding and delays, which are the main reasons for users disappointment. The dissatisfaction of the users, who will search for alternative solutions, translates into loss of income for the company. At the same time, it is not worth to provide too many trips as the money requested could not be rewarded.

The ideal solution would be an on-demand service, where trips are provided according to users' needs. To achieve this goal, the ability to predict the demand becomes an important and challenging point on which to focus attention. Its difficulty is the reason why nowadays the on-demand services involve only one or few people who want to reach the same place. They look for the nearest car or bike or scooter, take it and leave it close to their destination.

In the public transport domain, it is hard to offer the same service since the capacity of the buses should be exploited to meet costs and this implies that the

interests of a group of people should match. Their trips should be similar in terms of space and time to define some classes of users with common characteristics, helping the service providers to plan and schedule the trips, satisfying the demand. This information can be taken from the spatial-temporal coordinates of the journey or from its reason, which can be inferred from the regularity of the validations: same or near boarding and alighting stops within the same time slot repeated in time. For example, in a residential zone, it is reasonable to expect to collect a peak of check-in validations in the morning and one of check-out in the evening, following the traditional office hours.

In isolated areas, the forecast can be more challenging: the stops are located far from each other and the number of passengers taking the bus is definitely lower due to the distance from their destination and perhaps to the weak offer, which does not meet the needs of the inhabitants. The trips must be carefully scheduled to make the service appealing and to entice users to use it.

To forecast the demand, the usage of electronic tickets and, in particular, the smart cards, is very useful.

The first ones give information about the validation and the typology of the ticket but they can not be associated with the user. This is possible using the smart cards and it allows to retrieve some additional information such as the anagraphic data of the passenger.

For the aforementioned reason, nowadays the usage of smart cards is strongly encouraged and new typologies of fee and tickets are being developed, as an example, the possibility to recharge subscriptions and transport credit documents over the same electronic card. The first allows to travel within a specific area in a time slot of validity, which usually start at the validation moment while the latter gives the possibility to charge some money on the card and depending on the departure and the arrival place it is calculated and subtracted the price of the trip.

To avoid annoying users, they validate the ticket only at the boarding stop but this represents a limitation for the analysis since the alighting stop is unknown. However, investigating the following validations of the user - like in a travel diary - it is possible to infer the destination of each trip. Thanks to the timestamp and the location of the validation, it is easy to understand if the user needs to change bus to reach his/her destination or not.

Moreover, through the history of the validations of each user, a periodicity can be discovered. This can be explained by regular activities that users do - such as go to school or to office - and they are correlated to the category at which the user belongs.

The information retrieved from the smart cards is valuable for forecasting purposes. The validations should be collected for a long period of time in order to take into account all the values of the variables which have an impact on their trend.

Some stops represent important interchange points and the demand at those is quite stable due to the daily routines of the users. In these cases, the forecast can be done through some methods which care about the number of validations and nothing else, returning the moving average of the validations in the previous time slot at the same stop or involving also the seasonality.
When the demand has an unpredictable trend over the time, the prediction is more complex: the relationship between the target value and each one of the predictors should be carefully studied.

1.2 Thesis outline

The structure of this thesis is the following:

- Chapter 2 refers to the state of the art,
- Chapter 3 is dedicated to the description of the data, the used libraries and packages and to the pre-processing of the data,
- Chapter 4 describes the methodology, the predictive models selected and the performance metrics,
- Chapter 5 shows the results and the related comments,
- Chapter 6 summarizes the analysis and leaves suggestions for future works.

Chapter 2

State of the Art

Planning the offer of the public transport presents lot of critical points and the estimation of the demand is one of the strategies that the service providers adopt to overcome them.
The following list summarizes the most relevant open issues, each one with references to some contributions:

1. **Segmenting customers**: users can be grouped according to their anagraphical information or to the regularity (both spatial and temporal) of their trips: in [3], the partitioning is based on the travelers age or economic conditions and it provides a first step to solve also problem 4, while in [4] a two-level generative model is used to regroup passengers basing on their temporal habits in their public transportation usage.

2. **Forecasting mobility at a certain stop, station or zone**: predict how many people will need a particular mobility service, at a certain place, within a specific time slot. References are reported in **Tab.2.1**.

3. **Estimating the most likely destination**: forecast the alighting stop of a passenger of the public transport, taking information about the boarding stop and about the boarding stop of his following trip during the same day, if present. [5] and [6] exploit the virtual-checkout algorithm, while in [7] and [8] alternative solutions have been implemented for cases in which the following validation is not available.

4. **Forecasting individual mobility**: in a certain day, determines whether the user will move and, if so, when the user will leave and where (origin and destination). The origin-destination matrices (OD) is a useful tool to represent the aggregated result of this analysis. OD rows represent the origins and OD columns represents the destinations. The value in cell *(i,j)* corresponds to

the number of people moving from i to j within the selected time interval. [9] adopts a neural network with Long-Short Term Memory (LSTM), while the study in [10] aims to recover the chain of journeys by applying Markov models.

5. **Estimating the time instant at which a certain vehicle of a certain route will get at a certain stop**: this topic is correlated with (2); in [11] has been implemented a variant of the SVM algorithm while in [12] a clustering over the historical data has been done.

6. **Predicting travel and dwell time**: the first refers to the time needed for a public vehicle to travel between two consecutive stops while the second refers to the time needed to let passengers boarding and alighting at a certain stop. To determine them, in [13] a neural network takes as inputs the estimation of the traffic and of the incoming demand, while in [14] the travel and dwell times are clustered based on their distribution along the routes.

7. **Predicting bunching and preventing it**: bunching is when two vehicles following the same shape simultaneously arrive at the same stop: this may cause the overcrowding of the first vehicle and an increased headway (and therefore an unexpectedly great waiting time for people at that stop), a possible solution has been proposed in [15], where a Least Squares Support Vector Machines regression has been conducted.

8. **Reorganising the routes of the public transport**: basing on travel and dwell times and on demand at each stop, as done in [16] for Baghdad (Iraq), where origin-destination matrices are again used to represent the journeys.

Points from 1 to 4 deal with the demand of mobility in general terms, which does not involve only public transportation. While the following points, from 5 to 8, are related to the offer.

This study focuses on the problem number 2: forecasting the public transport mobility at a certain stop for a specific route.

A review of the literature regarding this issue can be summarize in **Table2.1**:

Table 2.1: Models used to predict public transport demand

Model	Flow type	Input variables	Time Horizon	**Reference**	Python package
GBDT	variable	- Number of alighting passengers from adiacent bus stops at t, t-1, t-2, t-3; - 3 most relevant subway passengers demand at t, t-1, t-2, t-3; - Time of the day, weekday, month	15 and 30 min	[17]	sklearn.ensemble
LSTM	abnormal	- Temporal dimension: day type, time of the day; - Exogenous factors: grade of precipitation; - Temporal dependency: 3 LSTM layers, one for each temporal interval (hourly, daily, weekly)	10 min	[18]	keras
SVR-LSTM	abnormal	- SVR1: periodic features, to compute the steady passenger flow volume (steady series); - SVR2: temporal series (recently observed real volume) and steady series (abnormal features); - LSTM: same as SVR2	15 and 30 min	[19]	keras
ARIMA / SARIMA	periodic	- Temporal dimension: day of the week, if holiday or not; - Exogenous: weather conditions	7 days ahead	[20]	statsmodels.tsa.arima
RF LT	variable	- Temporal dimension: day, four-day weekend, public holiday, school holiday, time step	1 year	[21]	sklearn.ensemble
RF ST	variable	// // + past samples of all other stations	15 min	[21]	sklearn.ensemble
KALMAN	peak hour	- Temporal dimension: past passengers flows at each station series based on the periodic features	next time step	[22]	/

Chapter 3

Data Description

3.1 Public Transport Data of Piedmont Area

The mobility analysis at the core of this thesis investigates real data referring to the public transport demand and offer provided by Granda Bus.
The **Granda Bus Consortium**, founded in 2004, nowadays counts 16 of the main companies operating in the local public transport sector in the Province of Cuneo, in North-Western Italy. Since 2010 Granda Bus has been managing Local Public Transport services in the area of the Province of Cuneo, which includes the suburban service of the Province of Cuneo, the service of the conurbations of Bra and Alba and the urban service of Mondovì, Saluzzo, Savigliano and Fossano.

The analysed data exploit the GTFS (General Transit Feed Specification) format, which is common for the public transport planning and it also provides information about the associated spatial coordinates. The dataset refers to the whole year 2019. The analysis investigates the number of validations, the typology of the tickets and the category of users for some couple stop-route for all the 52 weeks of 2019.

The available data retrieved from each validation are:

- **Temporal information**: timestamp (date and time),
- **Spatial information**:
  - stop id, identifies the bus stop and where it is located,
  - trip id, it allows to understand the real terminals of the route - since sometimes the end of the route does not coincide with the terminal but it is a point between the terminals - and it is useful to retrieve the path that the particular route follows,

  - route id, useful to detect the shape.

- **Ticket information**:
  - ticket typology, which can be single ticket, carnet, weekly, monthly, three months, year, year for students, year for over65, year for people with disabilities,
  - if it is a check in or a check out,

Information about the ticket typology allow to better understand the demand. Depending on the typology, it is possible to retrieve information about the user category: if the ticket is only for retired people, for sure the user is over65. This insight method has been used to insert the user's category if it was not present.

In case of smart card, also

- **Anagraphic information**: user id, which is useful to retrieve anagraphic data such as age and sex, place of birth, user category (student, retired, over 65, etc).

If the *user id* is available, it is possible to obtain personal data. They are useful also to verify the coherence between the information inferred and retrieved, for example the age with the type of ticket and the user category.

In case of transport credit formula, the users have to validate both at the boarding and the alighting stops, providing additional information related to their destinations.

3.2 Tools

For the development of this thesis the selected programming language is Python. It offers several packages such as:

- *pandas* for managing tables and csv file,
- *numpy* for arrays and matrices,
- *gtfs kit* for analysing the GTFS tables,
- *time* and *datetime* for working with temporal data and convert from one format to another,
- *geopandas*, *geopy* and *leaflet* for retrieving some geographical information,

- *sklearn* and *pmdarima* for applying and evaluating the machine learning algorithms,
- *matplotlib* for plotting the graphs.

Chapter 4

Methodology

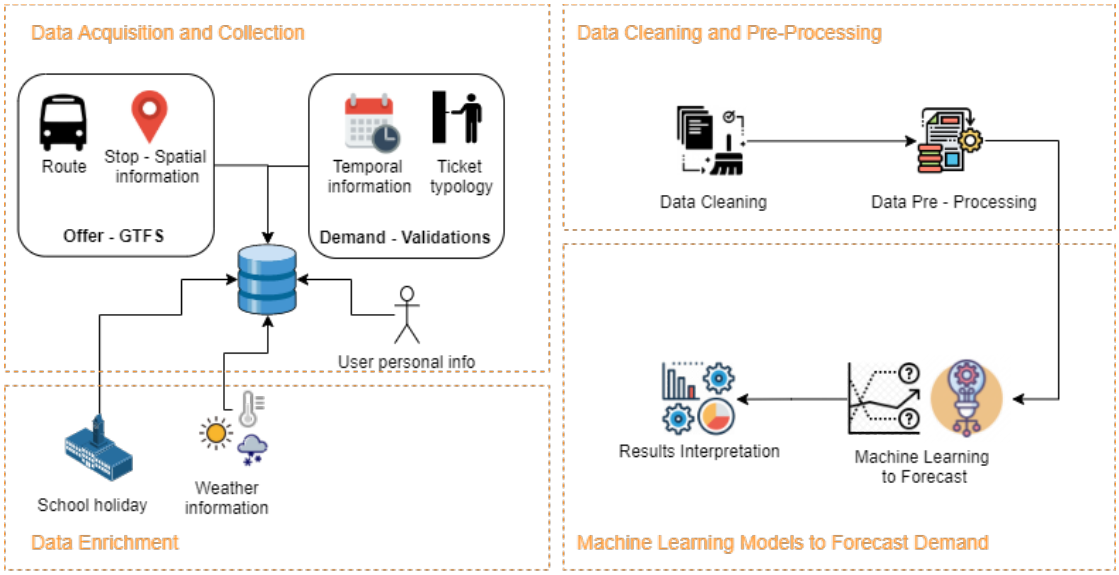


Figure 4.1: Proposed Framework

Fig. 4.1 shows the framework of the study, it is composed of three main blocks:

- **Data Acquisition, Collection and Enrichment**: collection of data provided thanks to the validation of smart cards and enrichment of such data with the addition of further details on exogenous and temporal contextual information,
- **Data Cleaning and Pre-Processing**: check integrity and consistency of the data applying specific filters and data preparation to run the models,

- **Machine Learning Models to Forecast the Demand**: selected machine learning predictive models have been used to forecast the demand in different temporal segments and then the results are evaluated.

4.1 Data Acquisition, Collection and Enrichment

The data used in the study are described in **Chapter 3**.
To enrich the basic information retrieved from the validations of the smart cards, some additional information have been added:

- **Temporal information**:
  - *Weekday*, explicit day of the week (Monday - Sunday),
  - *day type*, explicit type of the day (working day, holiday day or pre-holiday day),
  - *school holiday* (boolean value), it is equal to 1 (True) for the days in which the schools are closed: for example, during the Carnival's holiday, when students stay at home but workers do not and it is set to 1 as well as on Sunday when both schools and offices are closed. While it is equal to 0 (False) when schools are open and students have lessons. To assign the right value at this variable, the school calendar 2019 has been considered.

    The first two have been easily retrieved from the timestamp thanks to the *datetime* library in Python.

- **Spatial information**:
  - type of zone in which the stop is located (residential, working or mix). The socio-demographic data published by Istat [see 23] provide the number of the total buildings used and the number of residential and commercial/office buildings located in each census zone. The percentage of residential buildings is compared with the percentage of buildings used for production purposes, if one of these two is greater than 60% this category will define the area, otherwise the category will be mix.

- **Statistical information**:
  - average number of validations for the specific hour, weekday and day type,
  - median number of validations for the specific hour, weekday and day type,

These variables are computed over the dataset, taking into account hour, weekday and type of the day.

- **Weather information**: hourly minimum and maximum temperature, quantity of rain precipitation and categorical description of the situation (sunny, partly-cloudy, cloudy, variable, rainy, snowy, rainy-snowy, stormy, foggy).

Request HTTP library for Python allows to query web pages and download data. This function accesses the 3bmeteo portal and retrieves the weather conditions of each day of the dataset.

4.2 Data Cleaning and Pre-Processing

Before applying the specific techniques, data given in input to the models should be properly cleaned, filled and splitted in training dataset and test dataset.

4.2.1 Data Cleaning

Before starting the analysis, it is a good rule to check the quality of the data and clean them.
For this reason, the following filters have been applied:

- **Without stop**: to remove all the validations without *stop id*,
- **Synchronisation**: to remove all the validations which have the same *user id* and *timestamp* - with a minute of granularity - but different *stop id* since a user can not be in two different place at the same time. However, before removing the validation, two checks are done:
  1. compare the distance that a human can cover in one minute (using the average walking speed) with the distance between the two stops since if these are close one to the other it is possible that the user was able to reach the following stop in less than a minute,
  2. in case of transport credit, verify that the previous validation is a check out and the following is a check in.
- **Users**: to remove all validations referring to a user whose age is incoherent with the ticket typology (too young or too old),
- **Frequency**: to remove all the validations whose *user id* has a daily frequency validations greater than 10,
- **Average Speed**: to remove all the validations whose user's average speed is greater than the average speed for the specific trip.

4.2.2 Data Filling

Once the data has been cleaned, it is necessary to check that there is no missing data and, eventually, fill it in. Since the objective is to forecast the demand at each hour of the day, the data has been resampled in hourly time slot: each day of the year counts 24 records in the dataset.
The study takes as unit of measure one week (168 records): this choice reflects the periodicity of the data and the typical interval time used by the service provider to plan the offer.
This means that the size of the training dataset is a multiple of one week and the size of the test dataset is fixed to one week so that each week can be separately forecasted.
According to this, each day of the year has to be composed by 24 samples, if it does not, some records are added. The gaps due to the lack of validations are filled with 0 in the validation column while temporal and spatial features are retrieved from the previous or following samples.
The data filling is typical for the hours in which the service is stopped - generally from 10 pm to 6 am - when the number of validations (demand) is for sure 0.

4.2.3 Data Preparation

The dataset used for the analysis is composed of several categorical variables: the weekday, the type of the day and the weather conditions.
The models need numerical features to correctly run, so the one-hot encoding is used to convert these categorical values into numerical indicators. This means that a new binary variable is added for each categorical value.

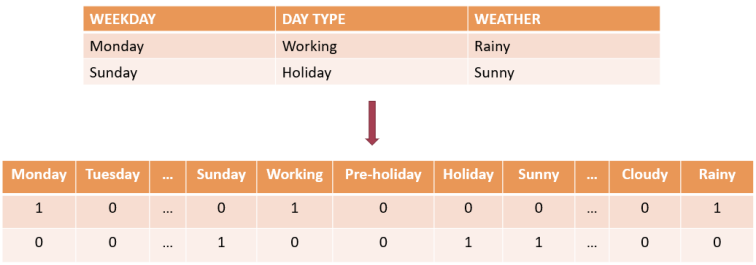


Figure 4.2: One-hot encoding example.

4.2.4 Data Splitting

Machine Learning algorithms use a dataset to train and fit the model, investigating the relationship between the variables given in input and the target one (*training*

set) and a dataset to test the performance of the model, evaluating the accuracy of the predictions (*test set*).
In this study, a grid search over the number of samples to take as training set (**N**) defines which is the best size of the training window for each model.
Instead, the number of test samples to predict is fixed for all the models and it is equal to 168. This choice reflects the trade-off between the need of the service provider to plan the offer in advance and the accuracy of the predictions.

4.3 Clustering

Following the study [24], a clusterization of all the couples *bus stop-route* has been performed to detect the representative ones, which can be used as models for the others.

The algorithm chosen for the creation of the clusters is the K-means [25]. It receives as input the data collected in October 2019 and outputs six clusters of couples. To create these grouping, it evaluates the number of validations and the importance of each couple in terms of offer (number of trips, terminal stop, number of interchanges within the route and frequency provided) and the volume of the demand.

4.4 Data Segmentation

At the beginning of the analysis, the dataset composed by all the data of the year has been used. It has been investigated a possible segmentation to avoid bad forecasting when temporal discontinuities occur: for example, in the first days of holidays in Summer and at the beginning of the academic year in Autumn.
The whole dataset has been into two segments basing on the following criterion:

- **Open Schools segment**, when schools are open: from the second week of January to the first week of June and from the second week of September until the third week of December,

- **Closed Schools segment**, when schools are closed: from the second week of June to the first week of September plus the Christmas holidays,

- **Hybrid segment**, composed by the weeks in which there are days in which schools are open and days in which schools are closed, such as the Carnival's week, the Easter's week and the 25^{th} April's week.

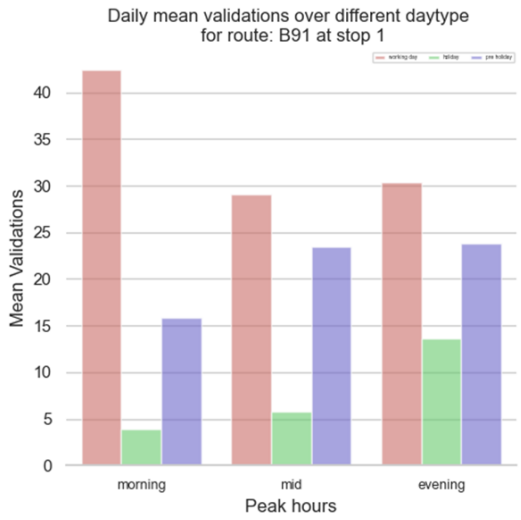


Figure 4.3: Holiday segment

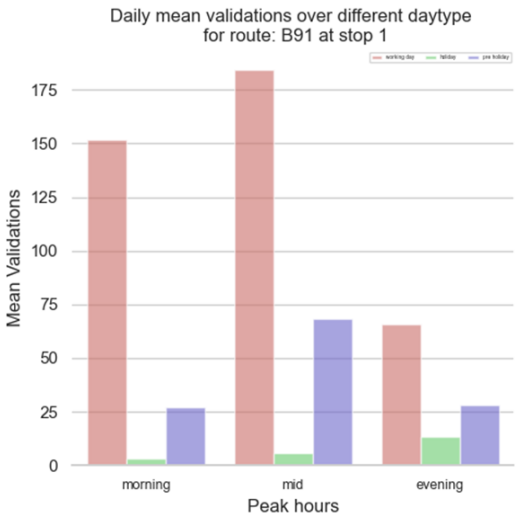


Figure 4.4: Working segment

The histograms in **Fig.4.3** and **Fig.4.4** represent the number of validations during the peak hours, which are the time slots in which the demand increases, for the different day types: working day, pre-holiday day and holiday day.

In particular, for working days:

- **Morning**: from 6 am to 8 am,
- **Mid day**: from 12 am to 2 pm,
- **Evening**: from 4 pm to 6 pm.

For holiday days:

- **Morning**: from 9 am to 11 am,
- **Mid day**: from 1 pm to 3 pm,
- **Evening**: from 5 pm to 7 pm.

These histograms confirm that the trend and the volume of the demand differ in the two scenarios.
It is clear that at least two segments are needed to have an accurate analysis, since the one that refers to the period of school holidays has a validation peak that is more than four times lower than the one which refers to the period in which the schools are open.

Hybrid Weeks

However, this is not a perfect separation because there are some weeks which are composed by both working and holiday days.

These weeks are called *"hybrid weeks"* and they are:

- **Carnival** week, from 4/03/2019 to 10/03/2019,
- **Easter** week - from 15/04/2019 to 21/04/2019,
- **25° April** week - from 22/04/2019 to 28/04/2019.

These weeks are evaluated independently, in a specific segment and the following strategy has been adopted: two prediction vectors are generated, one is the output of the model trained over a working week and the other is the output of the model trained over a holiday week. Then, each day of the test week is checked to understand if it is a working or a holiday day and, according to this, the prediction vector from which to take the output has been selected, (see **Figure 4.5**).

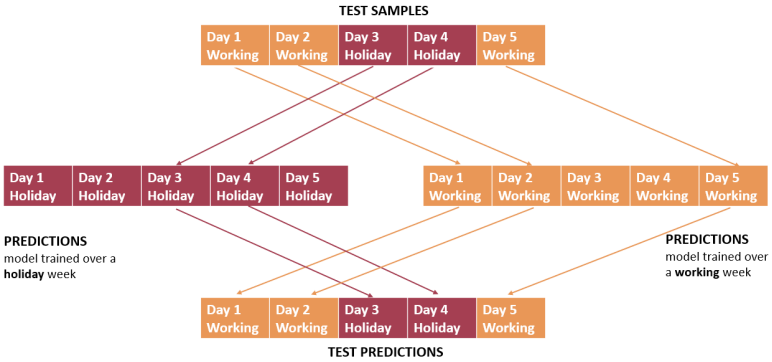


Figure 4.5: Association between test samples coming from hybrid weeks and corresponding predictions.

4.5 Machine Learning Models to Forecast the Demand

Machine learning is often used to build predictive models by extracting patterns from large datasets. Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy. To predict the demand of the public transport, the use of machine learning is strongly recommended to the forecasting.

According to **Tab.2.1** in **Chapter 2**, this section describes the predictive models

selected for this study.
In particular, these are: Average Response, Median Response, Random Forest (RF), Random Forest using the most important features (RF), Gradient Boosted Decision Tree (GBDT), Support Vector Regressor (SVR), SARIMA.
These algorithms will be described in details here after.

4.5.1 Average and Median Response

The Average and Median Response move from the concept of baselines, they are simple algorithms which do not involve artificial intelligence. They are useful to evaluate the benefits of the adoption of machine learning techniques.
These two techniques are very simple and there is no need to set any parameters.

They take as input the validations of the training dataset, the samples are grouped by the tuple weekday, daytype, hour and for each of these subsets the average and median values of validations are computed and saved. divide them grouping the same weekdays, day types and hours, compute the statistical value (average or median) over the data with the same properties and save it.
When they have to predict the number of validations for the test dataset, they evaluate the weekday, the day type and the hour of each sample and output the value stored with the same characteristics.

Proceeding in this way the predictions given in output change according to the training dataset. This is useful to take into account the differences that can occur between the trends of weeks within different periods, such as when schools are open or closed.

4.5.2 Ensamble Methods

Ensemble methods combine weak learners into a strong one and they are divided into two types: *Bagging* and *Boosting*.
Bagging, which stands for Bootstrap AGGregatING, consists of building successive learners on random - Random Forest is an example of this category.
Boosting consists of building a sequence of learners, where each learner focuses on the weaknesses of the previous one, and then building a strong classifier based on a weighted majority vote of the learners - Gradient Boosted Decision Tree is an example of this group.

Algorithm 1 Average and Median Response

```
1: procedure AVERAGE AND MEDIAN(w, d, h)
2:     ▷ w is the is the day of the week (from Monday to Sunday)
3:     ▷ d is the day type (working day, pre-holiday day, holiday day)
4:     ▷ h is the hour at which the validation occurs
5:     ▷ Execution
6:     ▷ Split the whole dataset into seven dataframe: one for each day of the
   week
7:     ▷ Compute the average or the median for each tuple (weekday, daytype,
   hour)
8:     for each sample of the test dataset do
9:         retrieve w, d, h
10:        ▷ Output is the value stored in the cell whose row correspond to the
   hour and column to the weekday
11:    end for
12: end procedure
```

Random Forest

Random Forest belongs to the family of *Bagging - Ensemble methods*, which is composed by some of the most useful machine learning techniques used nowadays as their performances reach good levels with relatively low cost.

Random Forest (RF), [26], is an algorithm consisting of many decision trees which can be used for both classification and regression.
It uses *bagging* and *feature randomness* when building each individual tree to try to create an uncorrelated forest of trees whose prediction is more accurate than that of any individual tree.

Bagging means that instead of generating one single binary tree, M different trees are built and each one of them is created and trained using different datasets, all originated by the starting one but with some random modification.
When a new input occurs, in the case of classification, M labels are retrieved and the majority rule gives the output while for the regression the output will be the mean of the M results.

Feature randomness means that each tree selects randomly a subset of features and this strategy helps in avoiding the overfitting problem.
RF model can be used for both short and long terms predictions.

Random Forest using the most important features

Random Forest model allows to identify the most relevant features computing the Gini impurity.
It measures the probability of incorrectly classifying a sample of the dataset if it is randomly labeled according to the class distribution in the dataset.
The importance of a feature is normalized therefore the values of the resulting array sum to 1. The higher, the more important the feature.

The selection of the features should improve the performance of the model - avoiding overfitting - and for this reason also a Random Forest model which takes into account only the most important features has been implemented.

Gradient Boosted Decision Tree

Gradient Boosted Decision Tree belongs to the *Boosting - Ensamble methods*, it is an approach where each tree learns from its predecessors and updates the residual errors. Hence, the tree that grows next in the sequence will learn from an updated version of the residuals.

The base learners in boosting are weak learners in which the bias is high, and the predictive power is just a tad better than random guessing.

The term *"gradient"* underlines that the algorithm uses a gradient descent algorithm to minimize the loss when adding new models.

It is proposed to handle different types of predictor variables, fit complex nonlinear relationships, and identify the interaction effects between influential factors

4.5.3 Support Vector Regression

Support Vector Regression (SVR) algorithm, [26], is a generalization of the binary classification problem solved by the Support Vector Machine (SVM) but, in this case, the output is a continuous value.

For this reason, a margin of tolerance ϵ is set and it represents the error that is accepted. However, the concept is the same as SVM: minimize the error taking into account ϵ determining the hyperplane which maximizes the margin.

The error function is composed by two parts: one penalises points that are misclassified, and the other penalises lines that are too close to each other:

$$Error = C * (ClassificationError) + DistanceError \tag{4.1}$$

The main hyperparameters to configure for the Support Vector Regression model are:

- the *kernel*, which can be non linear if the dataset is not linearly separable. In this case, the polynomial (poly) and radial basis function (rbf) kernel methods are the most popular choices.

- the coefficient *gamma*, which determines how far the impact of a single training sample arrives,

- the coefficient C which is a trade-off between keeping the errors as low as possible and keeping the lines used as support as far apart as possible.

Low values of gamma mean far and high values meaning close. C multiplies the first term of the error function **Eq.4.1**: large values of C to let the model focus more on classifying all training points correctly and low values of C to deal with a larger margin, therefore a simpler decision function, at the cost of training accuracy.

4.5.4 SARIMA

SARIMA is a time-series model, commonly used when an event is likely to happen at a regular interval of time, and to impact our target variable in a similar fashion each time it occurs.

SARIMA is an acronym of **S**easonal **A**uto **R**egressive **I**ntegrated **M**oving **A**verage.

This model works under the assumption of stationarity of the time series and it requires four parameters which together account for seasonality, trend, and noise.

- **p** represents the *Auto-Regressive* part, it is the number of past values to incorporate in the model as predictors,
- **d** represents the *Integrated part*, it is the number of past time points to subtract from the current value to obtain stationarity,
- **q** represents the *Moving Average* part, it allows to set the lagged forecast error of the model as a linear combination of the error values observed at previous time points,
- **s** specifies the length of time which defines the *seasonality* period.

4.6 Hyperparameter Tuning

For each predictive model it has been conducted a grid search to determine the best size of the training dataset window (**N**).
In particular, the interval of possible values for **N** ranges from 1 week to 7 weeks. For each one of these values a model has been trained and the prediction errors - in terms of MAE - have been evaluated.
The value of **N** used by the model with the lowest MAE becomes the selected one.

In addition to **N**, each model has its own *hyperparameters* to tune in order to perform better.
Hyperparameters are the settings of an algorithm that can be adjusted to optimize performance. While the model parameters are learned during training, the hyperparameters must be set by the data scientist before the training.
Scikit-Learn, the Python library exploited in this analysis, implements a sensitive set of predefined hyperparameters for all models, but they are not guaranteed to be optimal for the specific problem.
The best hyperparameters are difficult to determine in advance and optimizing a model is a trial-and-error-based process.

4.7 Performance Metrics

To have a more clear idea about the performances, a temporal sliding window is applied.

A temporal window composed by **N** training samples is fixed, as well as the size of the test samples to predict. The shift of this window generates several models.

Then, the mean of all the errors of the models is computed (see **Fig.4.6**).

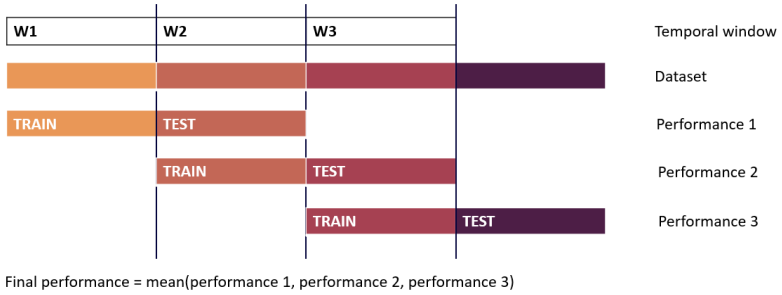


Figure 4.6: Performance evaluation of the model.

The performance metrics chosen are:

- **MAE**: Mean Absolute Error, it coincides with the difference between the predicted value and the real one: $e_t = (\hat{y}_t - y_t)$. The MAE is popular as it is easy to both understand and compute but its limit is that it can not be used to compare series characterized by different units because it is scale-dependent,
- **MASE**: Mean Absolute Scaled Error, it is a scale-free error metric and it is defined as follow:
  $MASE = |q_t|$, where $q_t = \dfrac{e_t}{\frac{1}{n-1}\sum_{i=2}^{n}|y_i - y_{i-1}|}$
  with $t = 1, \ldots, n$ which is the set of forecasting sample periods.

  The mean absolute error over all the series is used a scale factor, since both the numerator and denominator involve values on the scale of the original data, q_t is independent of the scale of the data. $MASE < 1$ if it arises from a better forecast than the average naïve forecast computed on the training data, conversely, $MASE > 1$ if the prediction is worse than the average naïve forecast computed on the training data.

  The main advantage of using MASE is that it never deals with undefined or infinite values, representing a good choice for intermittent-demand series

(which, as in this case study, arise when there are slots of zero demand in the forecast). It can be used on a single series, or as a tool to compare multiple series.

- R^2: R squared, it is the coefficient of determination and it is defined as follow: $R^2 = 1 - \frac{\sum_{i=1}^{N}[\hat{y}(n) - y(n)]}{\sum_{i=1}^{N}[y(n) - \bar{y}(n)]^2}$.

  It evaluates the ratio between the variance of the error - how much the prediction differs from the real value - and the variance of the measured data - how much the real value differs from the mean value. For this reason, desirable values of R^2 are as close to 1 as possible.

Chapter 5

Results and comments

In this section are reported the results of the study.

5.1 Data Cleaning

Table 5.1 shows that the most impactful filter is the first one, which is related to the validations without stop id, while the filters related to the synchronization and the frequency of the trips are not significative.
The average percentage over the whole year is around 88.70%, which means that the raw data do not present particular problems. However, the percentages of left data referring to the winter months are higher than the ones related to the summer months.

Table 5.1: Data Cleaning. Percentage of cleaned data after each applied filter.

MONTH	NO STOP	SYNC	USERS	FREQ	AV SPEED	Tot CLEANED
JANUARY	-8.04%	-0.01%	-1.12%	-0.25%	-2.49%	**88.42%**
FEBRUARY	-7.35%	-0.01%	-1.12%	-0.25%	-2.78%	**88.81%**
MARCH	-7.17%	-0.01%	-1.14%	-0.78%	-3.27%	**98.05%**
APRIL	-7.09%	-0.01%	-1.21%	-0.4%	-3.59%	**88.10%**
MAY	-6.05%	-0.01%	-1.13%	-0.47%	-3.21%	**89.45%**
JUNE	-6.46%	-0.02%	-1.61%	-0.57%	-4.53%	**87.33%**
JULY	-8.0%	-0.01%	-2.08%	-0.88%	-5.86%	**84.02%**
AUGUST	-12,65%	-0.02%	-1.39%	-0.96%	-5.95%	**80.19%**
SEPTEMBER	-6.0%	-0.02%	1.07%	-0.52%	-3.24%	**89.48%**
OCTOBER	-4.31%	-0.01%	-0.99%	-0.48%	-2.97%	**91.46%**
NOVEMBER	-5.98%	-0.01%	-1.05%	-0.54%	2.94%	**89.77%**
DECEMBER	-6.26%	-0.1%	-1.13%	-0.60%	-3.2%	**89.15%**

5.2 Data Filling

Following the criteria described in **Section 4.2.2**, the records related to the night hours - typically from 8 p.m. to 6 a.m. - and to the holiday days in which there was no service have been created.

5.3 Data Preparation

As explained in **Sections 4.2.3 and 4.2.4**, the data has been elaborated and organised properly. One hot encoding spreads the features from 31 to 46, the size of the training dataset depends on the model and the cluster chosen, while the size of test dataset is fixed to 168 which means 1 week (24 hours * 7 days).

5.4 Data segmentation

The data referring to 2019 has been organized into three temporal segments, as explained in **Section 4.4**.

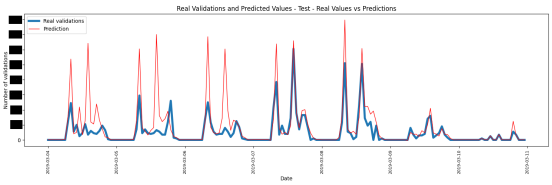
To confirm the importance of the segmentation, in particular of the hybrid weeks described in **Section 4.4**, the validations referring to the representative couple of group 1 (Stop 1 and Route B91) which occurred during the Carnival week have been evaluated in three different approaches:

1. **Working weeks**: working weeks are used as training dataset to predict the hybrid week,
2. **Holiday weeks**: holiday weeks are used as training dataset to predict the hybrid week,
3. **Working and Holiday weeks**: both working and holiday weeks are used as training dataset to predict the hybrid week.

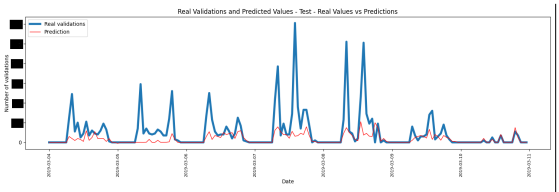
Fig.5.1a, **Fig.5.1b** and **Fig.5.1c** show the predicted number of validations (red lines) and the real number of validations (blu lines) during the Carnival week. It is clear that the red lines do not follow well the blu lines in both the first and the second case. **Fig.5.1a** represents the obtained results when working weeks are used as training dataset while **Fig.5.1b** shows the results when holiday weeks are used as training dataset.

This discrepancy between actual and predicted values can be explained by the fact that the students represent the largest component of users for the couple under analysis.
For this reason, the days within hybrid weeks in which schools are closed show a different trend of validations with respect to the ones referring to both working,

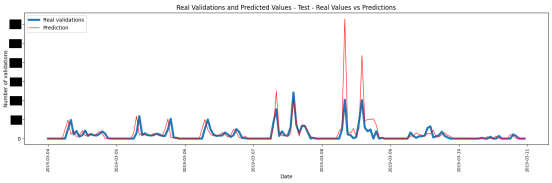
when schools and offices open, and holiday days, when schools and offices closed. It can be noticed a greater discrepancy when the model uses working days as training dataset but the number of validations does not meet the high expectations due to school closures or when the model uses holiday days as training dataset but schools are open and therefore the number of validations is much higher than expected. As can be seen in **Fig.5.1c**, the usage of both working and holiday weeks improves predictions because the training dataset is made up of days belonging to the same category as the days belonging to the test dataset.



(a) Working weeks as training dataset



(b) Holiday weeks as training dataset



(c) Working and Holiday weeks as training dataset

Figure 5.1: SVR predictions for Carnival week, using working weeks, holiday weeks or both for training: predictions and MASE.

5.5 Clustering

The clusterization described in **Section 4.3** identifies six clusters and their centroids couples have been analysed. In particular, they are:

- **Group 0**: low importance from the offer side (in terms of number of trips, terminal stop, number of interchanges) and few validations,
- **Group 1**: high importance from the offer side (in terms of number of trips, terminal stop, number of interchanges) and consistent volume of demand,
- **Group 2**: high importance from the offer side (in terms of number of trips, terminal stop, number of interchanges) and consistent number of validations,
- **Group 3**: low importance from the offer side (in terms of number of trips, terminal stop, number of interchanges) and low demand,

- **Group 4**: high importance from the offer side (in terms of number of trips, terminal stop, number of interchanges) but lower number of validations with respect to group 2,
- **Group 5**: similar characteristics of group 3 but with more validations.

Table 5.2: Clusterization of all the couples *bus stop-route* available

Cluster n°	Total Number of Couples	Representative Stop, Route
0	7113	26029, 299
1	5	1, B91
2	14	26041, 299
3	53	53, B91
4	325	4225, B42
5	77	541, B176

Tab. 5.2 reports for each group the number of elements which belong to the cluster and the centroid (*bus stop* and *route*).

In **Tab. 5.3** are specified the information of each representative route: the percentage of validations occurred in the given route within 2019 and its two terminals. From the data reported in the table, it is possible to notice that route B91 is the most popular one between the routes given as output by the clustering.

Table 5.3: Number of validations occurred in the representative Routes

Route	Percentage N° of Validations	Terminal A	Terminal B
299	2,43%	Torino	Saluzzo (CN)
B91	5.78%	Cuneo	Saluzzo (CN)
B42	1.04%	Gallo/Bivio Castiglione (CN)	Alba (CN)
B176	2.45%	Mondovì (CN)	Cuneo

In **Tab. 5.4** are specified the information of each representative stop: the percentage of validations occurred in the given stop within 2019 and its location. From the table it is possible to understand that stop 1 is the more popular since it represents an important interchange point, in particular to Torino and Cuneo which are the main cities in Piedmont. Moreover, according to the characteristics of each cluster, the representative stop of group 0 shows a very low percentage of validations, it may be a stop where the route not always stops.

Fig. 5.2 shows on a map the geographical locations of the explored stops. They are in Torino, Saluzzo (CN), Cuneo, Alba (CN) and Mondovì (CN).

Table 5.4: Percentage of validations occured in the representative Stops

Stop	Percentage N° of Validations	Location
26029	0.05%	Torino - Corso Unione Sovietica
1	4.79%	Saluzzo (CN) - Bus Station
26041	0.22%	Torino - Piazza Caio Mario
53	0.91%	Cuneo - Corso Giolitti
4225	0.85%	Alba (CN) - Bus Station
541	0.90%	Mondovì (CN) - Railway Station

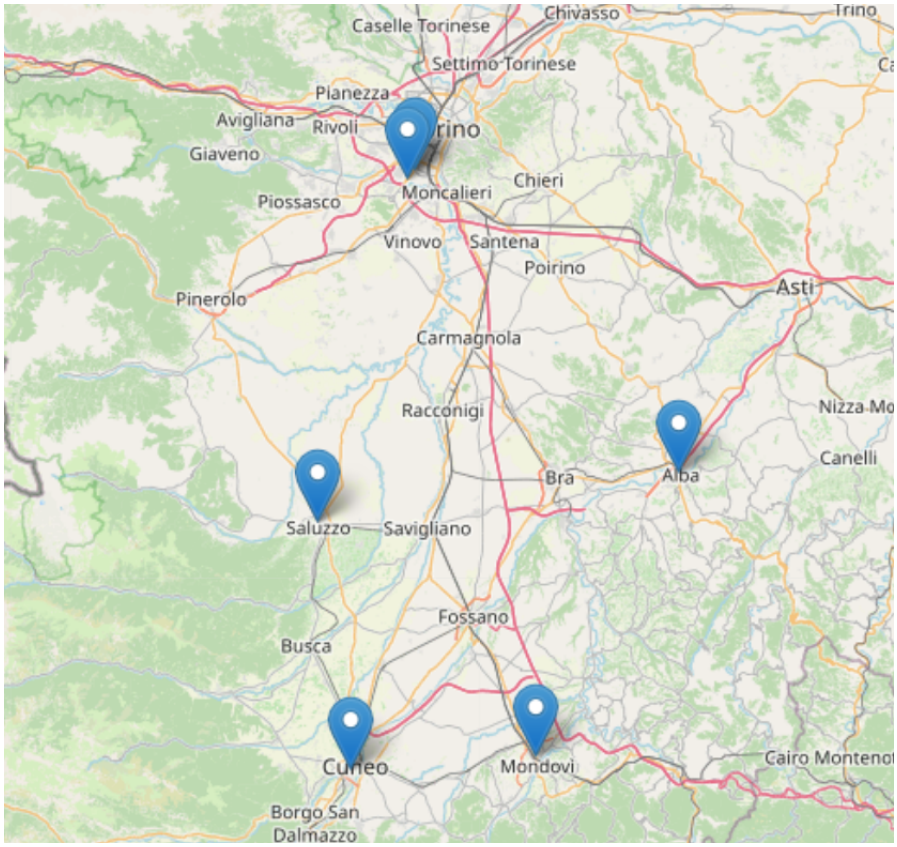


Figure 5.2: Location of the stops selected as representative for the clusters

5.6 Hyperparameters Tuning

According to what explained in **Section 4.6**, for each model it has been conducted a grid search to fix the best size of the training dataset window (**N**) and the hyperparameters .

The obtained results are reported in the following.

5.6.1 Random Forest

In the case of Random Forest, the most relevant hyperparameters include the number of decision trees (number of estimators) considered and the depth of the tree.
To set them grid searches have been conducted, the chosen ranges of values are:

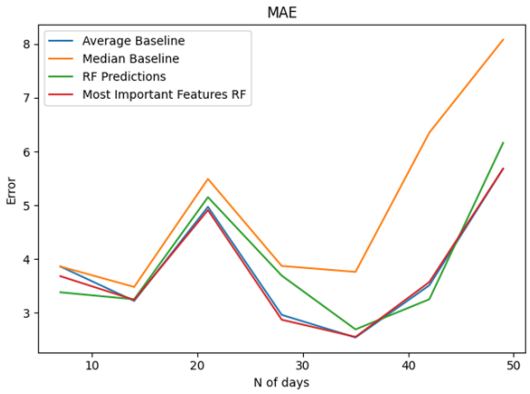
- **Number of Estimators**: from 10 to 200 with step 10,

- **Depth of the Trees**: [3, 5, 7, None], None is the default value which means that nodes are expanded until all leaves are pure

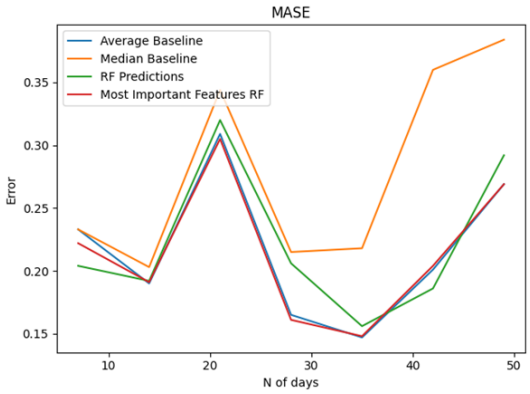
In this analysis, a Random Forest model which considers only the features whose importance is greater than 0.01 has been implemented.

Working Segment

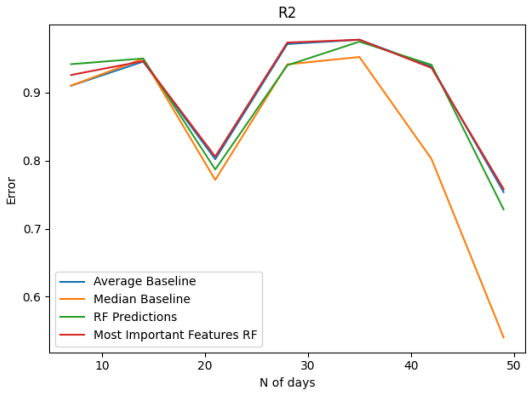
The first parameter to tune is **N**, the size of the training window size. Plots in **Fig. 5.3** represent MAE and MASE errors and coefficient $R\hat{2}$ for each value of N (number of days to take as training window). As shown in **Fig.5.3**, when N=14 there is a drop of the error in both **Fig.5.6a** and **Fig.5.3b** and an increase of the value of R^2. For this reason, for this scenario the optimum value of **N** is 14, which means 2 weeks of observations.



(a) MAE vs training window size



(b) MASE vs training window size

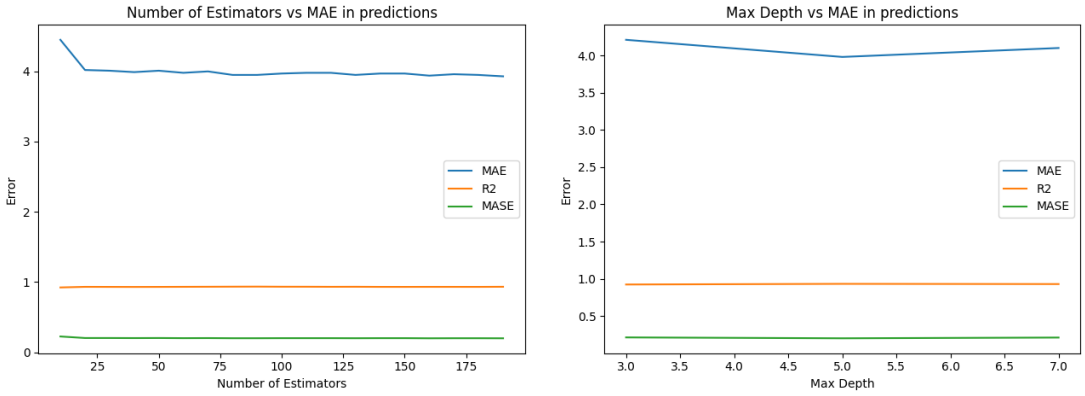


(c) R^2 vs training window size

Figure 5.3: Grid search for N - Working segment

Once N has been fixed, a grid search for the hyperparameters **Number of Estimators** and **Depth of Trees** has been performed.

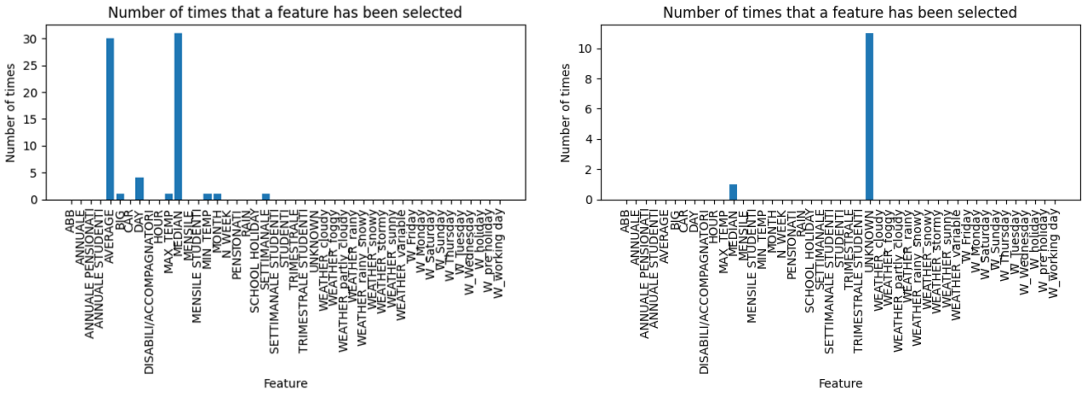
As can be seen in **Fig.5.4a**, the MAE related to the number of decision trees is quite stable. For this reason it has been set equal to 25, which corresponds to a down peak of the error.
The depth of the tree does not impact significantly the performance of the model in terms of MAE. However, in **Fig.5.4b**, there is a down peak when the maximum depth is set to 5.



(a) MAE vs Number of Estimators (b) MAE vs Max Depth

Figure 5.4: Grid search for hyperparameters

Histogram in **Fig.5.5a** refers to the Random Forest models which use only the most important features (see **Sec.4.5.2**). It shows on the x-axis all the features that the model can take and on the y-axis the number of times that the corresponding feature has been selected in all the models trained within the working segment.

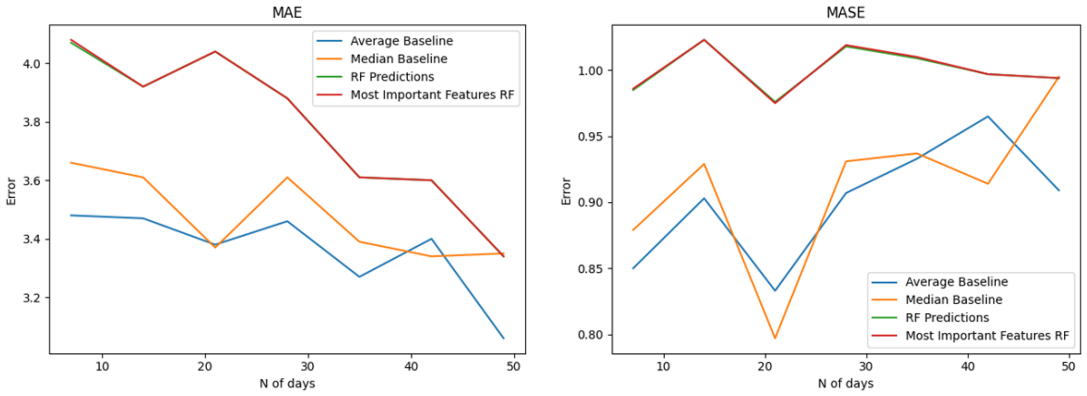


(a) Most important features - Working Segment (b) Most important features - Holiday Segment

Figure 5.5: Most important features - Random Forest

Holiday Segment

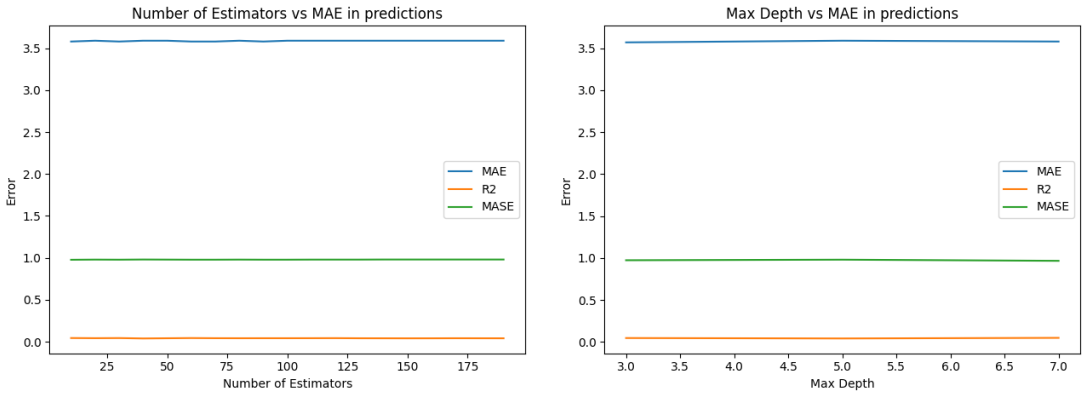
Also in the holiday segment the optimum value for **N** is 21, as can be seen in **Fig.5.6b**, since the corresponding MASE error is the lowest one while MAE error shows a negligible increase.



(a) MAE vs training window size (b) MASE vs training window size

Figure 5.6: Grid search for N - Holiday segment

As shown in **Fig.5.7a** and **Fig.5.7b**, the MAE related to the number of decision trees and maximum depth are stable. For this reason, the same values used to configure the working segment have been chosen.



(a) MAE vs Number of Estimators (b) MAE vs Max Depth

Figure 5.7: Grid search for hyperparameters - Holiday Segment

From histogram in **Fig.5.5b** it is evident that in this scenario the most important features are only two: unknown and median.

Hybrid Segment

Also in this case the grid search for **N** suggests to take a training window size of 2 weeks.

But in this case the most important features change, as shown in **Fig.5.8**.

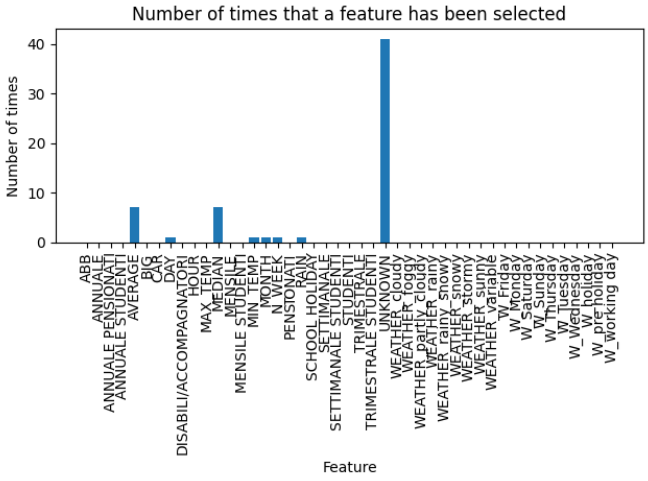


Figure 5.8: Most important features - Random Forest, Hybrid Segment

5.6.2 Gradient Boosted Decision Tree

There are three types of enhancements to basic gradient boosting that can improve performance:

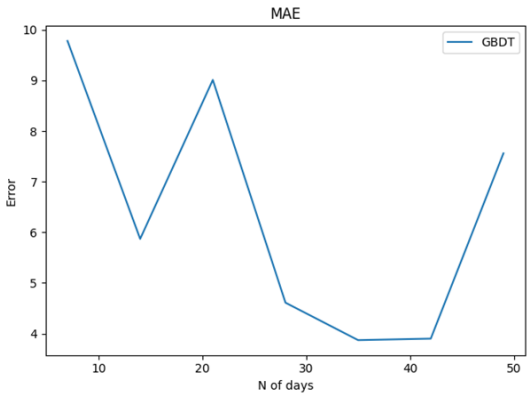
- *Tree Constraints*: such as the depth of the trees and the number of trees used in the ensemble,
- *Weighted Updates*: such as a learning rate used to limit how much each tree contributes to the ensemble,
- *Random sampling*: such as fitting trees on random subsets of features and samples.

In this study a grid search over the number of estimators, the maximum depth of each tree and the learning rate has been done. The used ranges are:

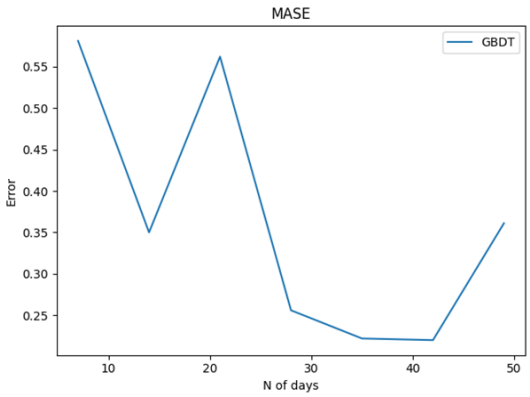
- **Number of Estimators**: [10, 20, 50, 100, 200, 500],
- **Maximum Depth**: [2, 3, 4, 5, 6, 7],
- **Learning Rate**: [0.0001, 0.001, 0.1, 1.0].

Working Segment

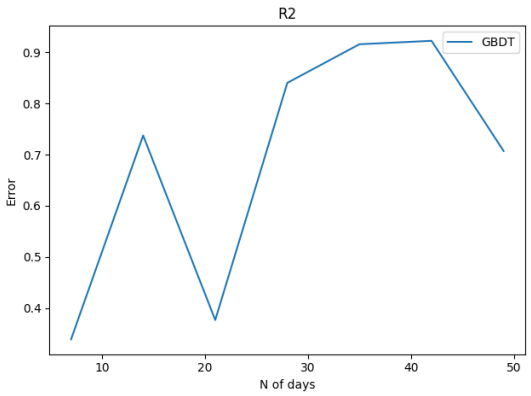
The grid search over **N**, the size of the training window, are reported in **Fig. 5.9**. It is clear that the best value for N is 28, since the corresponding MAE and MASE errors take the lowest values and R$\hat{2}$ presents a maximum.



(a) MAE vs training window size



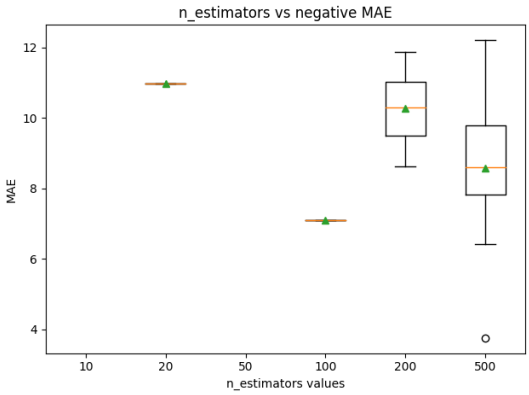
(b) MASE vs training window size



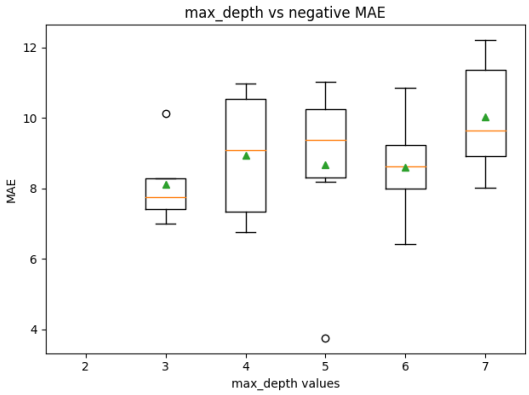
(c) R^2 vs training window size

Figure 5.9: Grid search for **N** - Working Segment

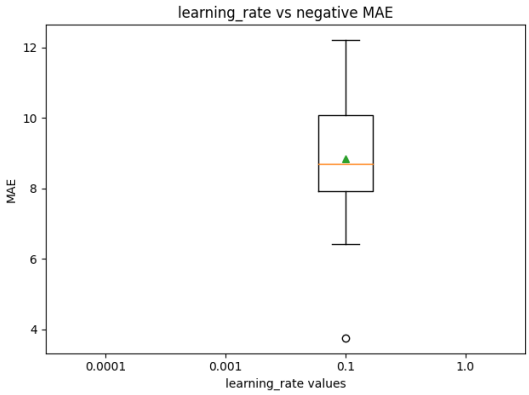
Fig.5.10 reports the results of the grid searches over the working segment. It is possible to see that the number of estimators which gives the lowest error is 100, the possible values for the maximum depth of the tree have errors which are close one to the others but the error related to max depth = 3 gives the lowest one. For the learning rate, the only value taken by the models is 0.1 so this represents the best value for this parameter.



(a) MAE vs Number of Estimators



(b) MAE vs Max Depth

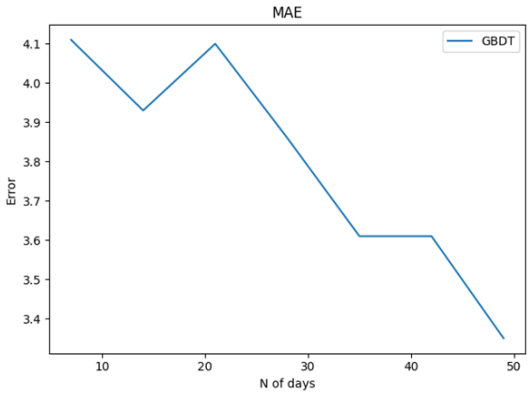


(c) MAE vs Learning Rate

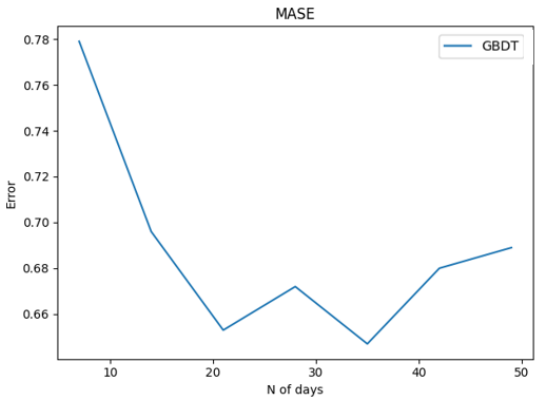
Figure 5.10: Grid search for hyperparameters - Working Segment

Holiday Segment

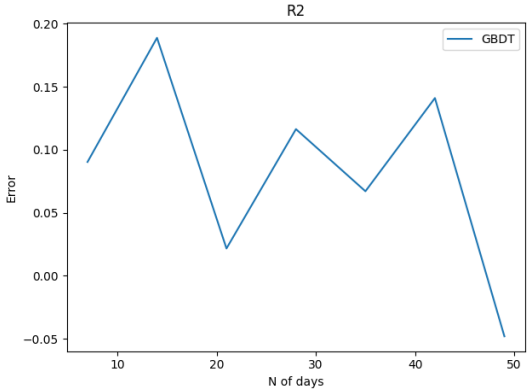
Fig. 5.11 represents the grid search over **N**, the size of the training window.



(a) MAE vs training window size



(b) MASE vs training window size

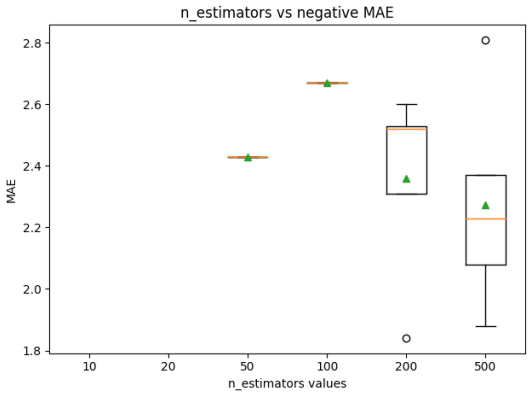


(c) R^2 vs training window size

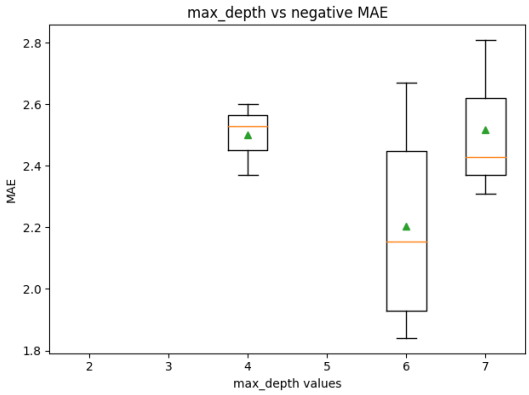
Figure 5.11: Grid search for **N** - Holiday Segment

In **Fig.5.11** can be seen that good indeces are reached for N=14, even if for MAE and MASE it does not correspond to the lowest value (N=49 and N=35, respectevely). However, since the differences in terms of MAE are not significant, N=14 has been chosen, in order to reduce the number of samples that are needed to the forecast.

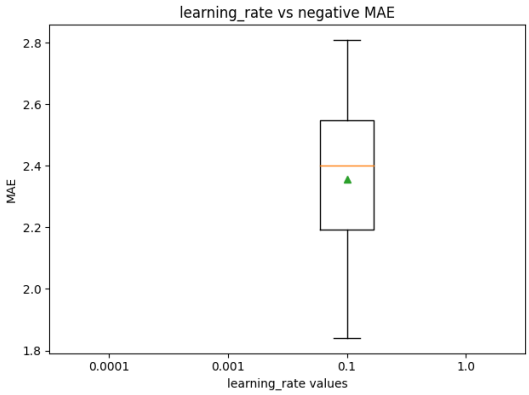
Fig.5.12 reports the results of the grid searches over the holiday segment. In this case, the best value for the number of estimators is 500, for the maximum depth is 6 and for the learning rate again 0.1.



(a) MAE vs Number of Estimators



(b) MAE vs Max Depth

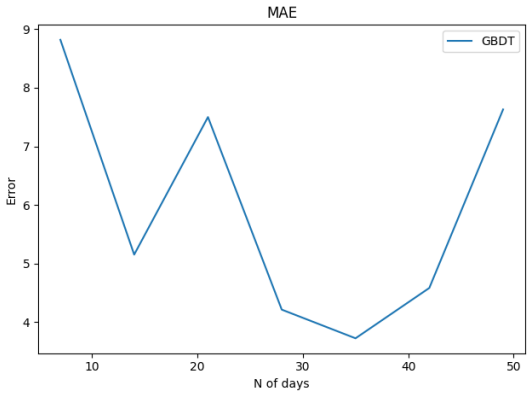


(c) MAE vs Learning Rate

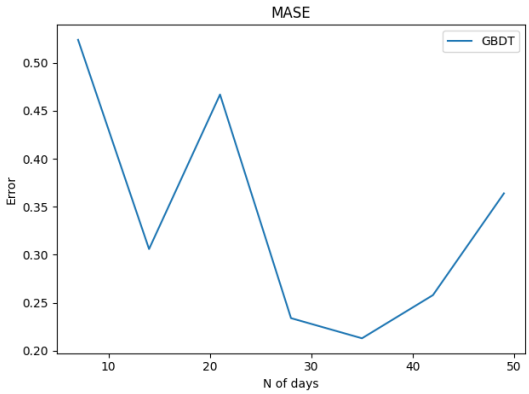
Figure 5.12: Grid search for hyperparameters - Holiday Segment

Hybrid Segment

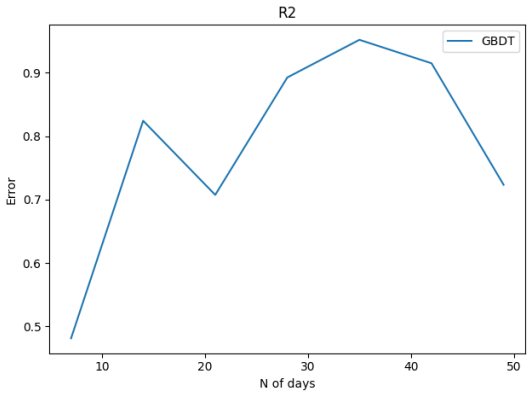
The results of the grid search over **N** are reported in **Fig. 5.13**.



(a) MAE vs training window size



(b) MASE vs training window size



(c) R^2 vs training window size

Figure 5.13: Grid search for **N** - Hybrid Segment

The trends in **Fig.5.13** show that the optimum value for this scenario is N=28, since it corresponds to low MAE and MASE errors and to a high value of R^2.

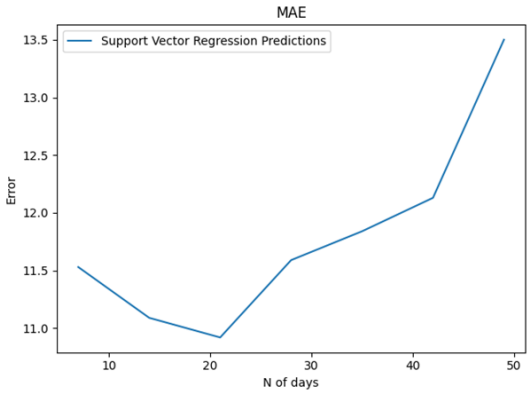
5.6.3 Support Vector Regression

As explained in **Section 4.5.3**, the main hyperparameters to configure for the Support Vector Regression model are:the *kernel*, the coefficient **C** and the coefficient **gamma**. The selected intervals of possible values are:

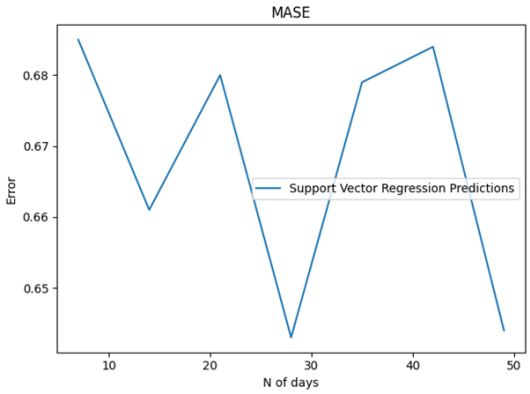
- **Kernel**: [linear, polynomial, radial basis function],
- **C**: [1, 10, 100, 1000, 10000],
- **gamma**: [0.001, 0.01, 0.2, 0.5, 0.6, 0.9].

Working Segment

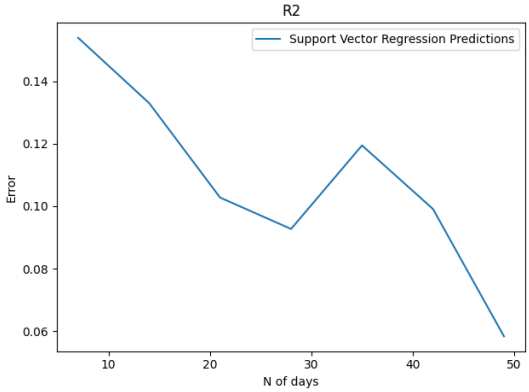
The results of the grid search over **N** are presented in **Fig. 5.14** in terms of MAE, MASE and R^2.



(a) MAE vs training window size



(b) MASE vs training window size



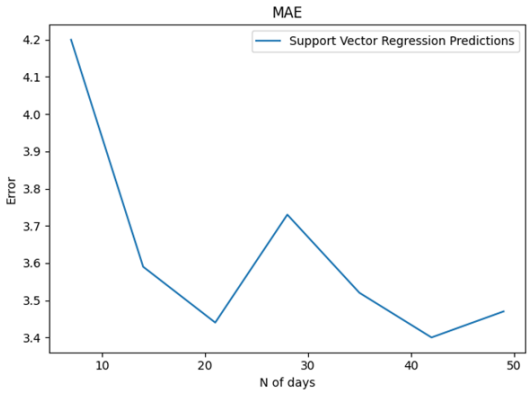
(c) R^2 vs training window size

Figure 5.14: Grid search for **N** - Working Segment

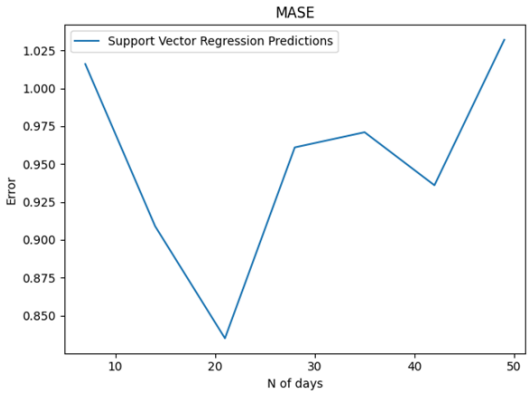
In this case, N has been set equal to 21. Even if the MASE corresponding at this value shows an increment, it is not far from the lowest peak of the line (0.678 and 0.645), therefore can be neglected, see **Fig. 5.14b**.

Holiday Segment

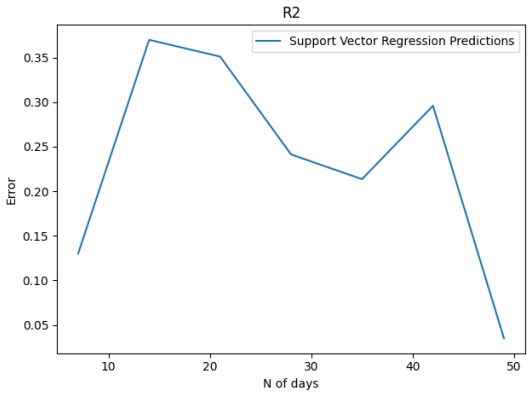
Also in this case, the optimum value for the size of the training window is 21. As can be seen in **Fig. 5.15**, it corresponds to the lowest MAE and MASE and to the maximum value of R^2.



(a) MAE vs training window size



(b) MASE vs training window size

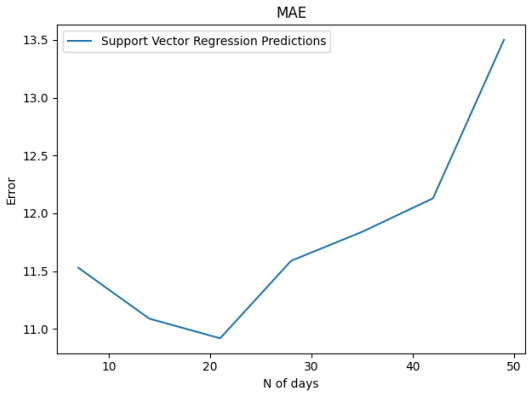


(c) R^2 vs training window size

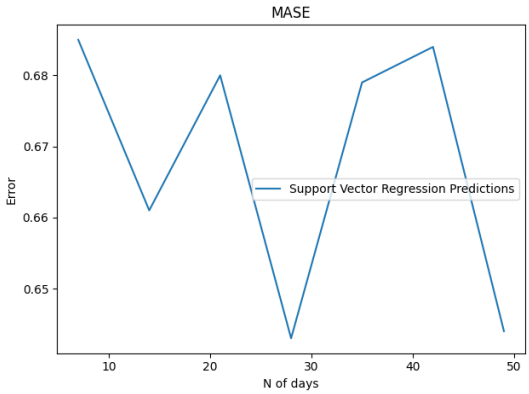
Figure 5.15: Grid search for **N** - Holiday Segment

Hybrid Segment

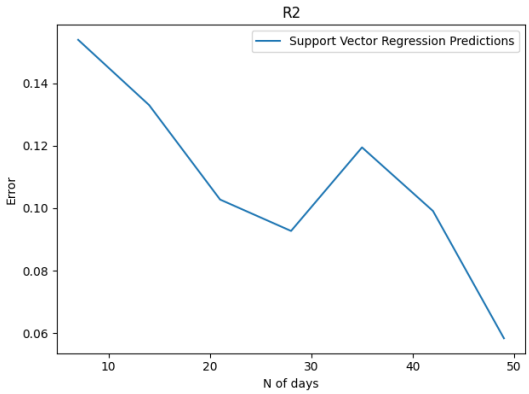
The hybrid segment reflects the choices of the other two segments: the training window takes 21 days of samples. **Fig. 5.16** justifies this choice and reports the grid search for **N** in terms of MAE, MASE and R^2.



(a) MAE vs training window size



(b) MASE vs training window size



(c) R^2 vs training window size

Figure 5.16: Grid search for **N** - Hybrid Segment

5.6.4 SARIMA

For SARIMA models **N**, the size of the training window has been fixed to 7 since this value represents the seasonality of the time series, *s*. However, before applying SARIMA models, it is necessary to verify the stationarity of the time series.
The most popular method is the Augmented Dickey-Fuller test (ADF) [27].
Once the stationarity has been checked, it is important to tune the seasonal order. This is a difficult and delicate task, but it has a great impact on performance. It aims to find the best values for the fundamental parameters, which are: *(p, q, d)*, as explained in **Sect.4.5.4**.

In this study, the function present in the package *pdmarima* [28] has been used to identify the set of best values for the aforementioned parameters. It is called *auto_arima*, it takes as input the training dataset and the extremes of the intervals

to perform the grid search for the parameters p and q.
auto_arima executes differencing tests to determine the order of differencing, d, and then fits models performing a grid search over the defined ranges for p and q. If the seasonal optional is enabled, *auto_arima* also seeks to identify the optimal P, Q and D hyper-parameters.
From the resulting models tuned after the internal grid search, it can be seen that d is equal to 0 most of the times, in this case, it is not necessary to differentiate the time series to achieve stationarity. Sometimes the parameters p, q, d take values (0,0,0), this means that the resulting model is a time series only containing a constant and white noise and the errors are uncorrelated across time. However, this does not give information about the size of the errors, so it is not an indication of good or bad fit.

5.7 Application of the Predictive Models

This section presents the results of the application of the models to the given dataset.

The investigated couples *bus stop-route* are the centroids of the detected clusters (see **Section 4.3** and **Table 5.2**). For each one of them, it has been identified the best predictive model for the three temporal segments (working, holiday and hybrid) and the proper size of the training window (N).

In **Table 5.5** for each temporal segment and cluster are reported: (i) the best model to use and its MASE error, (ii) the second best model, (iii) the difference in terms of MASE with respect to the first one, (iv) the number of samples to look at to make an accurate forecasting (N).

Moreover, the column *Gain** shows the MASE error gain of using the best model over using the Average Baseline. The last row of the table represents the average gain in the whole segment: as if the model identified as the best one was chosen for all clusters instead of using the simple Average Baseline.

Table 5.5: Best predictive techniques for each representative couple in the three different temporal segments.

Group	Working	N	MASE	Gain*	Holiday	N	MASE	Gain*	Hybrid	N	MASE	Gain*
0 - C.so Un. Sov. (TO)	**1. SVR**	21	0.57	**+38%**	**1. SARIMA**	7	0.97	+28%	**1. SVR**	21	0.85	+7%
	2. RF	21	+0.13		2. RF	14	+0.02		2. RF	21	+0.02	
1 - Saluzzo (CN)	**1. SVR**	21	0.23	+8%	**1. RF**	14	0.77	+2%	**1. SARIMA**	7	0.37	**+50%**
	2. RF	14	+0.00		2. Median	14	+0.01		2. SVR	21	+0.04	
2 - P.zza C. Mario (TO)	**1. RF**	28	0.56	+3%	**1. SVR**	35	0.96	**+30%**	**1. SARIMA**	7	0.76	+23%
	2. Average	28	+0.02		2. GBDT	21	+0.02		2. RF	28	+0.01	
3 - Cuneo	**1. RF**	28	0.47	+2%	**1. SVR**	35	0.95	+2%	**1. RF**	28	0.68	+7%
	2. Average	28	+0.01		2. Most Impo Features (RF)	21	+0.01		2. SVR	35	+0.05	
4 - Alba (CN)	**1. SVR**	28	0.71	+7%	**1. SVR**	21	0.95	+6%	**1. RF**	28	0.73	**+31%**
	2. RF	21	+0.01		2. RF	28	+0.03		2. SVR	21	+0.03	
5 - Mondovì (CN)	**1. RF**	14	0.43	+4%	**1. SVR**	28	0.97	+5%	**1. SVR**	28	0.67	**+38%**
	2. Most Impo Features (RF)	14	+0.01		2. GBDT	14	+0.01		2. GBDT	21	+0.03	
Avg Gain*				+10%				+12%				+26%

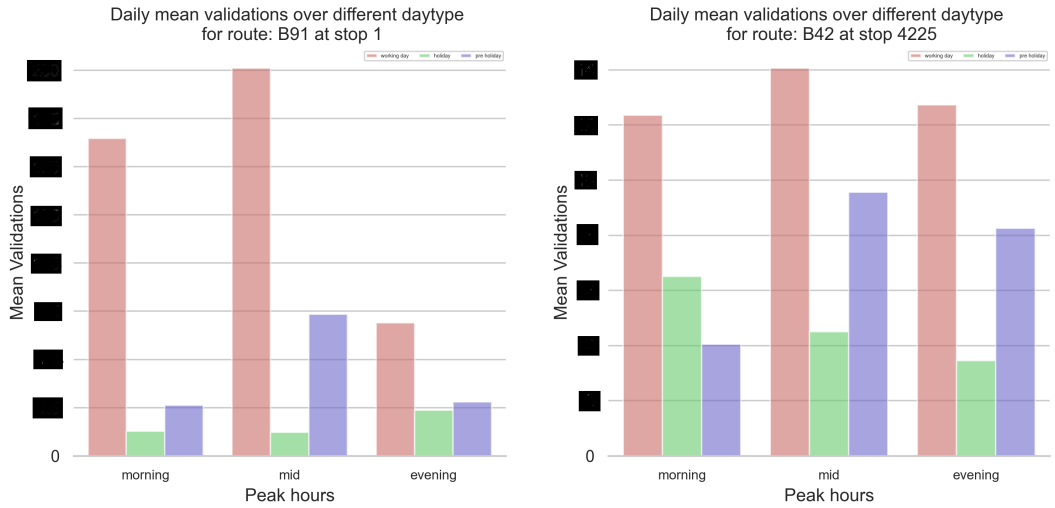
The table can be read in two different ways: vertically and horizontally.
The vertical point of view allows to identify which is the more appropriate model for each temporal segment, while the horizontal one detects the best model for each cluster and segment.
The vertical interpretation identifies for the working segment RF and SVR as the best performing models, for the holiday segment SVR but also RF and GBDT show good results and for the hybrid segment again SVR, RF and SARIMA are the ones whose results stand out.
If the model that is the best for each cluster is chosen within the working segment, the average gain of the MASE error is 10%, within the holiday segment it is 12% and within the hybrid it raises until 26%.
This means that machine learning increases the performance in all cases by at least 10% and the higher is the uncertainty of the validations – variability of the demand –, the higher are the ML improvements over the baseline. Therefore, it helps most in forecasting demand when it is not regular and easy to estimate, such as in hybrid weeks.

The horizontal reading of the reported values show that for some groups it is possible to use the same model for different segments. In particular, for clusters 0 and 3, the best model to predict weeks belonging to working and hybrid segments is the same: respectively, for the first SVR and RF for the second. For group 0 RF results as the second best model for all the three segments and investigating the distribution of the users, this could be influenced by the fact that the students component represents about the half of all the users and this percentage does not change through the different temporal periods (see **Fig. 5.19**).
As expected from the clustering results (see **Section 5.5**), the results of cluster 3 are similar to the ones obtained by cluster 5. Both the representative routes of those groups have one of the two terminals located nearby schools and the distribution of the users across the temporal segments is similar. Students are the main component

of users within the working segment, while in the others their percentage drastically drops down. The best model for the hybrid segment is the same of the holiday one: if the students do not take the bus, the number of validations decreases and this results in a significative drop of validations overall week. Consequently, the number of weekly validations is more similar to the holiday segment than to the working one.
Group 1 shows results very similar to the ones of groups 3 and 5 but in this case also in the holiday segment there is a high presence of students. The number of validations within this segment is lower with respect to the one typical of the working period, but the composition of the users do not change. Since the 5.78% of the total validations refers to route B91, as reported in **Table 5.3**, it is possible to infer that this route is important in terms of both offer and demand and that it is not only frequented by students who have to go to school but also by workers and other users.

Group 4 differs from the others: the same two models result to be the best ones in all the segments. The analysis of the users categories (see **Fig. 5.33**) and the location of the stop, nearby the Ferrero factory, suggest a reason for this: the most of the users are workers who take the bus to go to the establishment and since they have to go to work even during holiday weeks, the trend of the validations does not change significantly over the periods.

Some statistical consideration have been performed over the clusters to better understand the nature of the validations and the target of users, therefore the forecasting results.



(a) Group 1 (b) Group 4

Figure 5.17: Validations during peak hours

In **Fig. 5.17** are represented the number of validations which occur, on average, during the peak hours in working days (red), holiday days (green) and pre-holiday days (violet) for two different groups: 1, which needs to change model according to the temporal segment and 4, for which only one model can be used.
For group 1 the number of validations during working days is significative higher than the one referring to holiday or pre-holiday days, while for group 4 this difference is not so evident.
This discrepancy may be one of the reasons for which group 4 requires only one model for the whole year, while group 1 needs two models, one for the working and hybrid segments and one for the holiday one.
This consideration is valid also for clusters 0, 2, 3 and 5 which require more than one model. Moreover, an important characteristic of group 4 is that the routes which belong to it do not work on Sunday and during National holiday days.
At this point it should be recalled that the data predicted in the holiday segment refers to the weeks in which the schools are closed, even if it is not holiday for all workers.

5.8 Centroids Analysis

In the following performance metrics, pie charts and box plots related to each cluster have been reported. In particular, pie charts represent the categories of

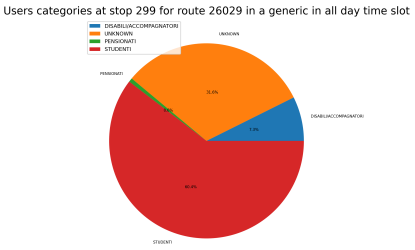
users and the ticket typologies while the box plots are useful to have information on the variability or dispersion of the prediction error. They show for each temporal segment and performed technique the collected MASE over all the predicted weeks. In the comparison between the temporal segments has to be underlined that the number of the considered weeks changes between the segments: in the working segments there are 33 weeks, in the holiday segment 15 and in the hybrid one only 3.

5.8.1 Group 0: Stop 26029, Route 299

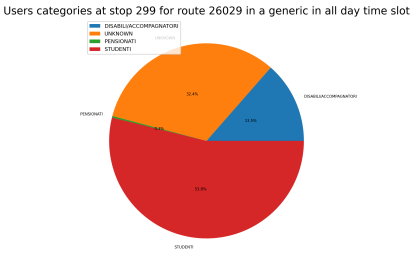
The representative couple for cluster 0 is Stop 26029, located in Turin - Corso Unione Sovietica, and Route 299, which links Torino to Saluzzo (CN), see **Figure 5.18b**.
As can be seen in **Fig. 5.18a**, the stop is in *Mirafiori*, a district of Turin built by FIAT for its employees in the 1960s, the years of its great success. It was also defined as *"dormitory neighborhood"* to underline the absence of services and daily life.
Nowadays it is a suburban and residential area, where mostly workers live.

(a) Stop 26029 (b) Route 299

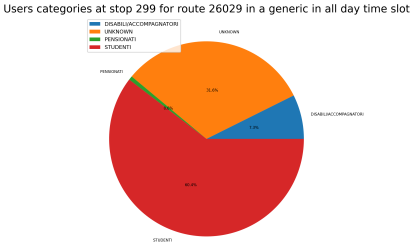
Figure 5.18: Group 0 - Stop and Route.



(a) Working Segment



(b) Holiday Segment



(c) Hybrid Segment

Figure 5.19: Group 0 - Users categories in different temporal segments

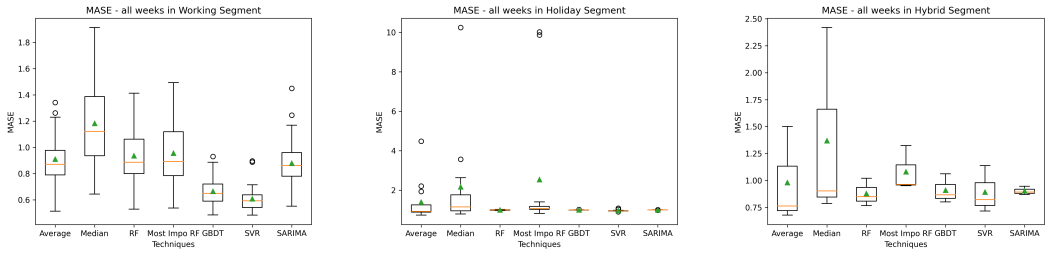
In cluster 0 the composition of the users does not change significantly between temporal segments, since the percentage of students in the working and hybrid ones is about 60% (see **Pie Chart 5.19a**) and in the holiday one it is about 54% (see **Pie Chart 5.19b**).

Table 5.6: Group 0 - Performance Metrics for all the techniques, for all temporal segments.

	WORKING				HOLIDAY				HYBRID			
Model	N	MASE	MAE	R2	N	MASE	MAE	R2	N	MASE	MAE	R2
Average	21	0.92	0.27	0.10	14	1.35	0.24	0.20	21	0.92	0.26	0.17
Median	21	1.18	0.29	0.07	14	1.48	0.26	0.23	21	1.16	0.27	0.21
RF	21	0.90	0.27	0.05	14	0.99	0.25	0.34	21	0.87	0.28	0.42
Most Impo RF	21	0.92	0.27	0.10	14	0.99	0.25	0.34	21	1.35	0.34	0.29
GBDT	14	0.70	0.31	0.20	14	1.01	0.24	0.37	14	0.93	0.30	0.32
SVR	**21**	**0.57**	**0.28**	**0.31**	14	1.11	0.23	0.29	**21**	**0.86**	**0.29**	**0.55**
SARIMA	7	0.87	0.25	0.29	**7**	**0.97**	**0.25**	**0.33**	7	0.90	0.31	0.61

In **Table 5.6** are reported the size of the training window and the performance metrics of the analysed predictive models in all the selected temporal segments. For each one of them the best model has been detected (bold values). The results are obtained averaging the values of all the predicted weeks.

For this cluster (see **Table 5.6**), the technique which gives the best results in terms of MASE is the same for both the working and the hybrid segments and it is SVR, while for the holiday segment it is SARIMA.



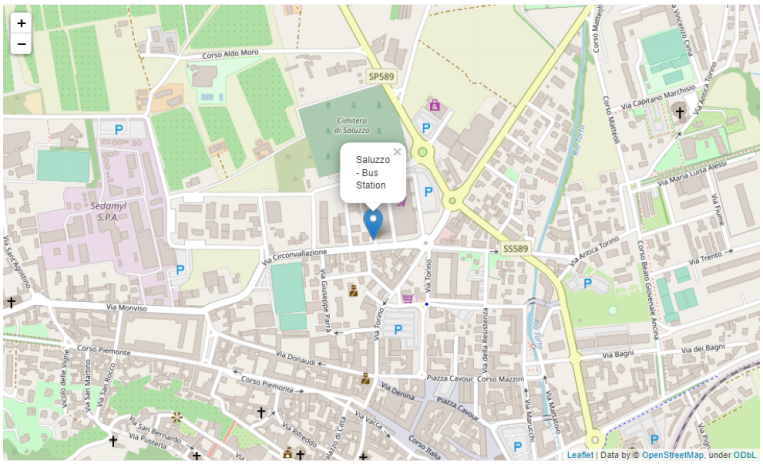
(a) Working Segment (b) Holiday segment (c) Hybrid segment

Figure 5.20: Group 0 - MASE box plots for all the temporal segments.

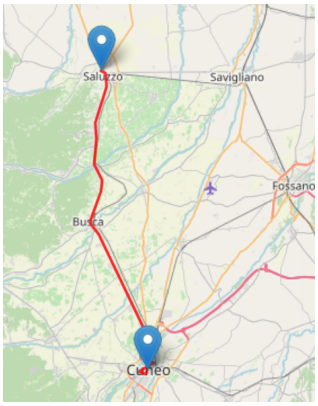
From **Fig. 5.20** can be seen that the techniques within the holiday segment have similar performances over the weeks in terms of average MASE but some outliers (white circles) are present. For the working segment the boxes spreads around a compact range of values and there are also some outliers. For the hybrid segment there are no outliers and the means (green triangles) of the different techniques are almost aligned.

5.8.2 Group 1: Stop 1, Route B91

For **group 1**, the study focuses on **stop 1**, which is the Saluzzo (CN) bus station, an important inter-exchange point for that area, and **route B91**, which links Saluzzo (CN) to Cuneo, the largest provincial capital of Piedmont, see **Fig. 5.21b**. From the map in **Fig. 5.21a** can be noticed that the stop is located near the cemetery, therefore it is not in the city centre, but on the edge of the it. This location confirms the role of inter-exchange of the stop: lots of routes arrive and leave there.

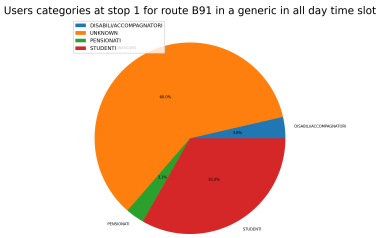


(a) Stop 1

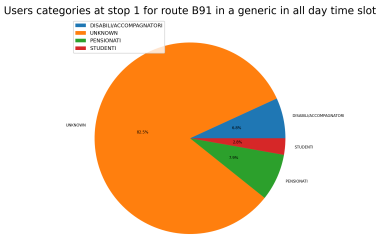


(b) Route B91

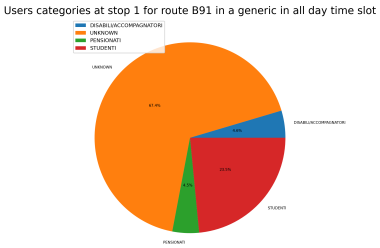
Figure 5.21: Group 1 - Stop and Route.



(a) Working Segment



(b) Holiday Segment



(c) Hybrid Segment

Figure 5.22: Group 1 - Users categories in different temporal segments

Within cluster 1 the students component (in red) decreases noticeably from working to holiday segment and this explains the drastic reduction of validations during the holiday segment (see **Fig. 5.17a**). While in the hybrid segment the students percentage is similar to the working one, as expected from the results reported in **Table 5.5**: for group 1 the same predictive model can be used for both

working and hybrid segments. So it is evident that the data changes significantly due to a known exogenous event: the opening or closing of schools.
This discrepancy justifies the segmentation of the year in two parts: working segment, when schools are open and holiday segment, when schools are closed.

An interpretation of these results may be that students live in a country village linked to Saluzzo (CN) - or in Saluzzo itself - and they take the bus to go to school to Cuneo, which is the main city in the zone.
As a consequence, it is clear that the distribution of the data in the working segment differs from that in the school holidays segment, in which students do not travel to school.

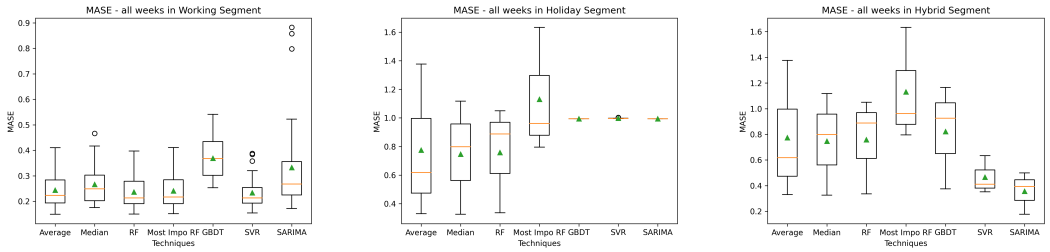
Table 5.7: Group 1 - Performance Metrics for all the techniques, for all the temporal segments.

	WORKING				HOLIDAY				HYBRID			
Model	N	MASE	MAE	R2	N	MASE	MAE	R2	N	MASE	MAE	R2
Average	14	0.25	4.28	0.88	14	0.79	3.25	0.18	14	0.78	4.61	0.03
Median	14	0.27	4.78	0.84	14	0.78	3.35	0.10	14	0.98	4.95	0.01
RF	14	0.23	4.16	0.89	**14**	**0.77**	**3.47**	**0.12**	14	0.78	4.9	0.11
Most Impo RF	14	0.24	4.17	0.89	14	1.02	3.57	0.10	14	0.99	7.10	0.10
GBDT	28	0.35	5.12	0.81	28	0.99	3.55	0.11	28	0.35	5.33	0.15
SVR	**21**	**0.23**	**4.13**	**0.90**	21	0.99	3.55	0.09	21	0.31	4.68	0.6
SARIMA	7	0.27	4.76	0.84	7	0.99	3.62	0.13	**7**	**0.27**	**5.03**	**0.84**

From **Table 5.7** it is clear that in the working segment the best models are SVR, with a window size of 3 weeks, and RF, with a window of 2 weeks. Considering also the other metrics, SVR has the lowest MAE and MASE and the highest R^2.

Into the holiday segment the best identified model is RF with 2 weeks and in the hybrid segment, the model with the lowest MAE is the SARIMA with a window size of 1 week.

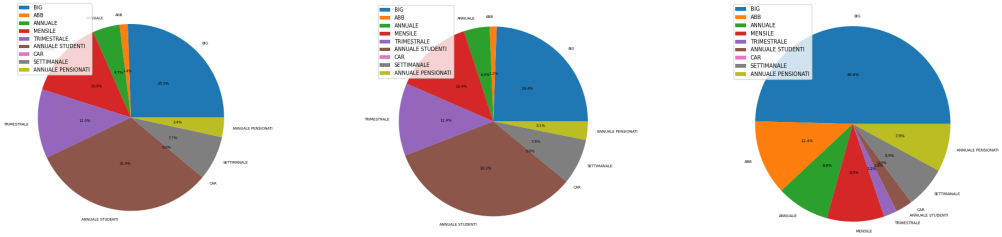
Fig. 5.23 shows the MASE box plots for each temporal segment.



(a) Working Segment (b) Holiday segment (c) Hybrid segment

Figure 5.23: Group 1 - MASE box plots for all the temporal segments.

Box plot 5.23a shows that the models under analysis have similar performance in terms of MASE and some outliers are present for SVR and SARIMA. In **Box plot 5.23b** can be seen that the MASE error over the weeks which belong to the holiday segment does not change for GBDT, SVR and SARIMA models, while it spreads between 0.3 and 1.6 for the other techniques. From **Box plot 5.23c** turns out that SARIMA and SVR differ considerably from the others, performing better.

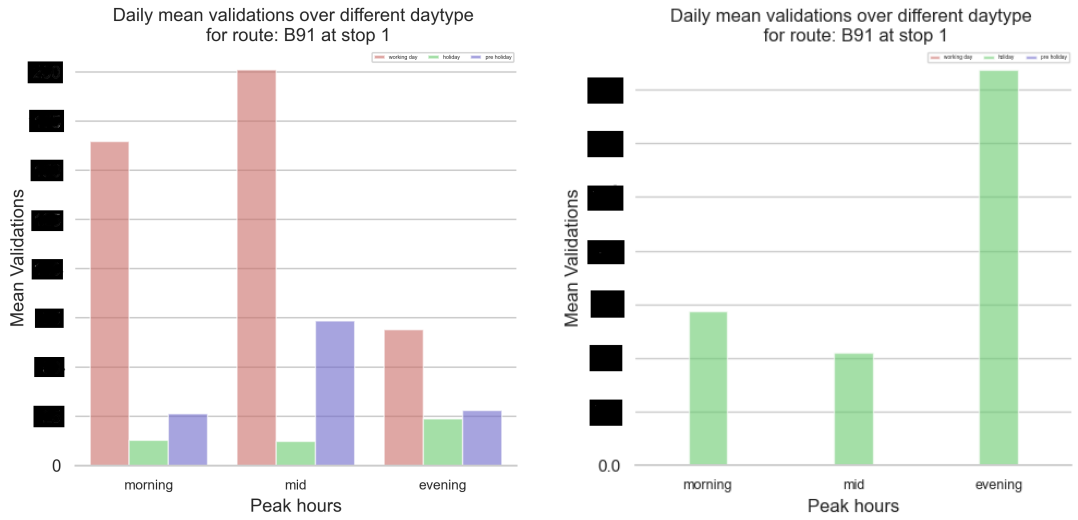


(a) All year Tickets (b) Working Segment Tickets (c) Holiday Segment Tickets

Figure 5.24: Group 1 - Tickets typologies statistics computed over different period of time: all the year 2019, school holidays segment and working segment.

Data reported in **Fig.5.24**, which refer to the ticket typologies, confirms that students use the public service mostly during the academic year.
It is also important to take into account that during the holiday segment some students may change ticket typology from "student annual pass" to others, such as single ticket or booklet, since some special tariffs dedicated to students only last 10 months, corresponding to the period in which schools are open.
This means that during the vacation period, the students who adopt the aforementioned ticket typology, have to buy a single ticket to take the bus.
However, since the single ticket does not allow to keep trace of the anagraphic data, the user, in this case a student, will not be classify as *student* but as *unknown*.

Moreover, the analysis of the hourly distribution of the validations helps in further investigating this discrepancy between the segments.



(a) Working segment

(b) Holiday segment

Figure 5.25: Group 1 - Distribution of validations over different peak hours (morning, mid day and evening) in working, holiday and pre-holiday days.

The trend of the holiday period is smoother, it is quite stable through the hours of the day. Within this segment the number of validated tickets drops significantly with respect to the working period - around five times lower. However, in both cases it is possible to identify some peaks which occur everyday in the same time slots. These are called *peak hours*, since they refer to the the couple of hours in which the demand grows.

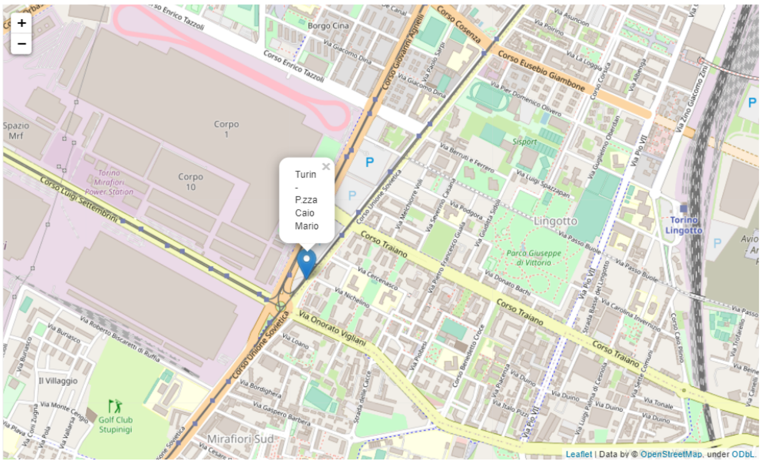
In **Fig.5.25** the validations are grouped in the peak hours.
Usually they reflect the common routine: in the early morning people go to work or to school, in the middle of the day students and part-time workers come back home and in the evening full-time workers leave their office and return to their home.

The highest peak of validations in the working segment is almost 4 times the one in the holiday segment.

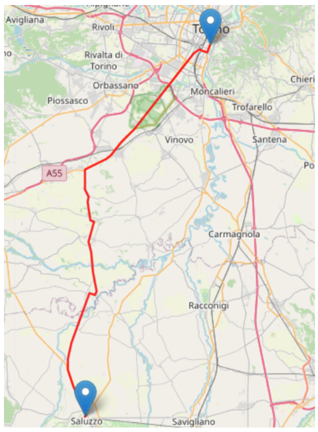
Another difference is that in the working segment, in each peak hours slot, the number of validations within a working day is at least 3 times the one referred to the holiday or pre-holiday day.
While in the holiday segment the number of validations within the same time slot for different day types does not change significantly.

5.8.3 Group 2: Stop 26041, Route 299

The representative couple for group 2 is the stop number 26041, located in Turin - Piazza Caio Mario and the route 299 which links Torino to Saluzzo (CN).
Fig. 5.26a shows that the stop is not far from the representative stop of cluster 0 (**Fig. 5.18a**). In this case, it is closest to *Lingotto*, a more commercial zone.

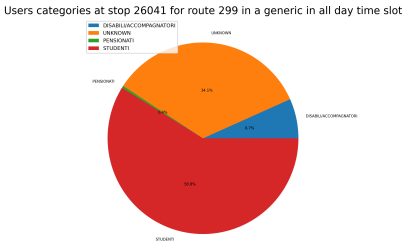


(a) Stop 26041

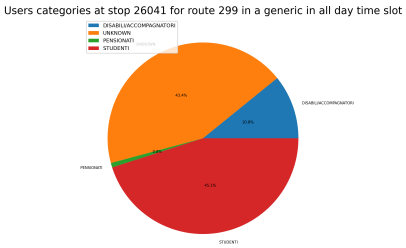


(b) Route 299

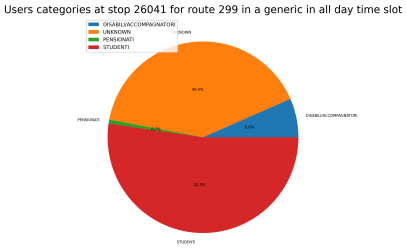
Figure 5.26: Group 2 - Stop and Route.



(a) Working Segment



(b) Holiday Segment



(c) Hybrid Segment

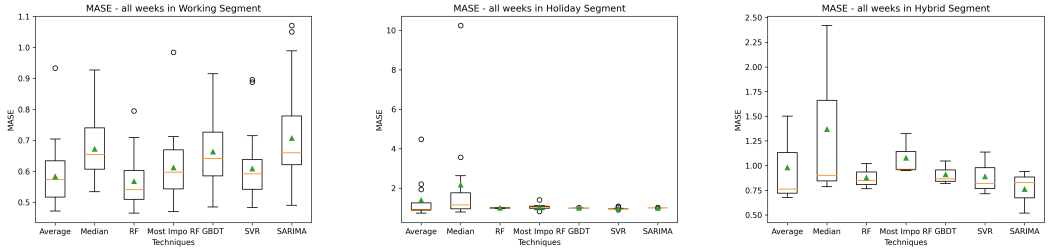
Figure 5.27: Group 2 - Users categories in different temporal segments

Pie Charts 5.27 represent the categories of the users belonging to cluster 2. It can be seen that the percentage of students changes by 15% between working and holiday segments. This reflects the discrepancy between the number of validations during working and holiday segments. The percentage remains quite stable from working to hybrid segment.

Table 5.8: Group 2 - Performance Metrics for all the techniques, for all the temporal segments.

	WORKING				HOLIDAY				HYBRID			
Model	N	MASE	MAE	R2	N	MASE	MAE	R2	N	MASE	MAE	R2
Average	28	0.48	1.50	0.65	21	0.97	1.08	0.21	28	0.73	1.34	0.61
Median	28	0.55	1.89	0.47	21	1.16	1.32	0.16	28	0.76	1.51	0.43
RF	**28**	**0.47**	**1.53**	**0.63**	21	0.98	1.17	0.14	**28**	**0.68**	**1.18**	**0.63**
Most Impo RF	28	0.49	1.59	0.61	21	0.96	1.17	0.18	28	1.01	1.83	0.18
GBDT	28	0.54	1.75	0.58	21	0.99	1.21	0.15	28	0.74	1.33	0.13
SVR	28	0.49	0.66	0.10	**35**	**0.95**	**1.13**	**0.19**	35	0.73	1.13	0.63
SARIMA	7	0.60	1.35	0.43	7	0.97	1.31	0.12	7	0.81	1.15	0.70

Table 5.8 detects the best models for each temporal segment for the cluster 2. RF results to be the more performing technique for both working and hybrid segments, while for the holiday one it is the SVR. This is in accordance with what has been observed in the previous graphs, relating to the composition of users (**Figure 5.27**).



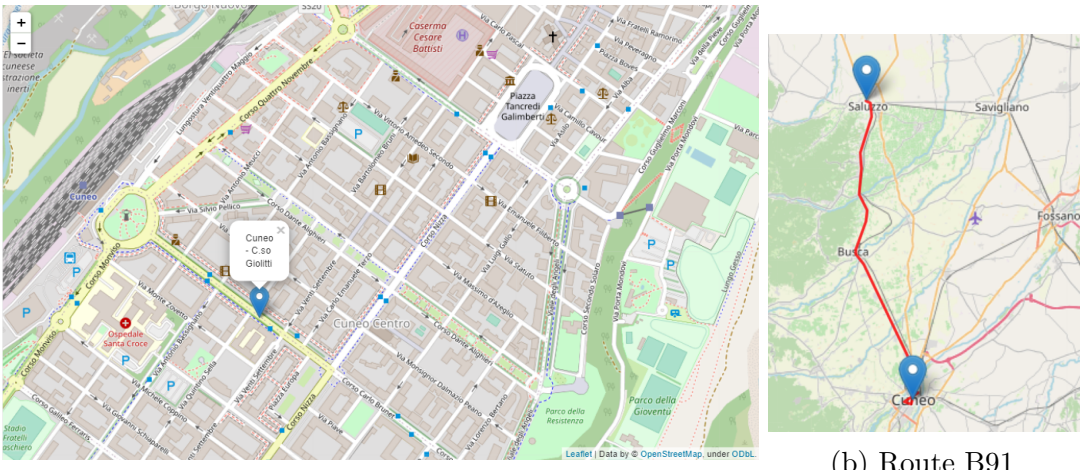
(a) Working Segment (b) Holiday segment (c) Hybrid segment

Figure 5.28: Group 2 - MASE box plots for all the temporal segments.

Figure 5.28 show the box plots for the temporal segments, in this case the MASE error is almost homogeneous between the techniques in all the three cases, with a visible drop for RF in the hybrid period.

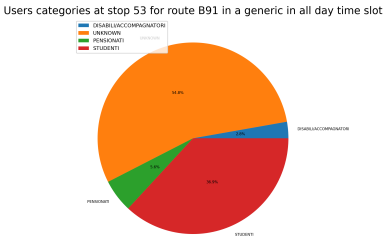
5.8.4 Group 3: Stop 53, Route B91

Stop 53, situated in Cuneo - Corso Giolitti, and Route B91, which connects Cuneo to Saluzzo (CN), are the elements used to model the group number 3.
Map in **Fig. 5.29a** shows that the stop is in the centre of Cuneo, near high schools. For this reason, it is expected that the prevalent users category is the *students* one.

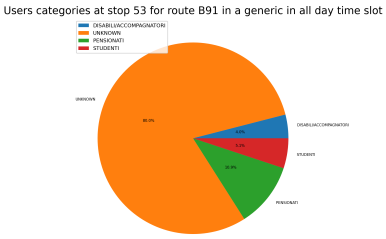


(a) Stop 53 (b) Route B91

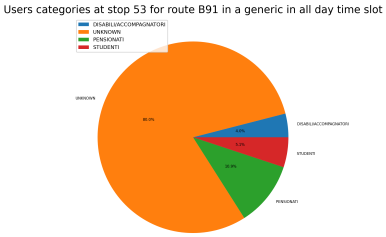
Figure 5.29: Group 3 - Stop and Route.



(a) Working Segment



(b) Holiday Segment



(c) Hybrid Segment

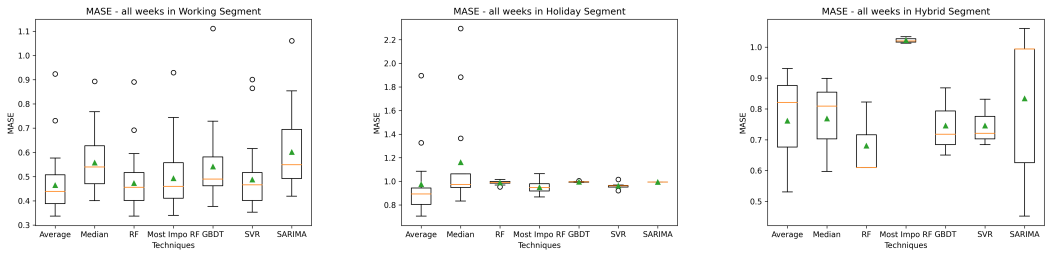
Figure 5.30: Group 3 - Users categories in working and holiday segments

Cluster 3 presents a drastic decrease of students percentage from working to holiday segment. **Pie Charts 5.30b** and **5.30c** present a similar division in user categories, this reflects the similar results obtained for the holiday and hybrid segments, reported in **Table 5.5**. It turns out that RF and SVR are the best predictive models for both. These results confirm the expectations linked to the position of the stop. The presence of high schools nearby explains why the hybrid segment is similar to the holiday one: as students represent the majority of users, even if some days within the hybrid week are working days and offices and commercial activities are open, they can be considered as holiday ones because students stay at home and the number of validations drastically decreases.

Table 5.9: Group 3 - Performance Metrics for all the techniques, for all the temporal segments.

	WORKING				HOLIDAY				HYBRID			
Model	N	MASE	MAE	R2	N	MASE	MAE	R2	N	MASE	MAE	R2
Average	28	0.48	1.50	0.65	21	0.97	1.08	0.21	28	0.75	1.34	0.61
Median	28	0.55	1.89	0.47	21	1.16	1.32	0.16	28	0.76	1.51	0.43
RF	**28**	**0.47**	**1.53**	**0.63**	21	0.98	1.17	0.14	**28**	**0.68**	**1.18**	**0.63**
Most Impo RF	28	0.49	1.59	0.61	21	0.96	1.17	0.18	28	1.01	1.83	0.18
GBDT	28	0.54	1.75	0.58	21	0.99	1.21	0.15	28	0.74	1.33	0.13
SVR	28	0.49	0.66	0.10	**35**	**0.95**	**1.13**	**0.19**	35	0.73	1.13	0.63
SARIMA	7	0.60	1.35	0.43	7	0.97	1.31	0.12	7	0.81	1.15	0.70

From **Table 5.9** can be identified the model which performs the best in each temporal segment. RF shows the lowest MASE values in both working and hybrid segments, while for the holiday one SVR is the best. It can be noticed that SVR performs well also in the hybrid segment, according to the similar distribution of users.



(a) Working Segment (b) Holiday segment (c) Hybrid segment

Figure 5.31: Group 3 - MASE box plots for all the temporal segments.

For this cluster the box plots referring to the working and holiday segments (**Fig.5.31a** and **Fig. 5.31b**) show some outliers but the mean values of all the applied models are comparable. The box related to the Most Important Features RF in **Fig. 5.31c** catches the attention since it has the highest mean value, while the others ranges between almost the same interval.

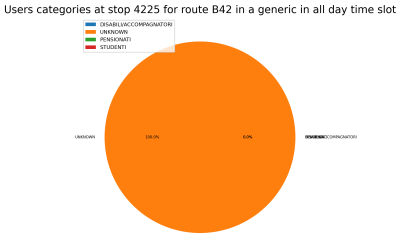
5.8.5 Group 4: Stop 4225, Route B42

Group 4 is represented by the stop 4225, the bus station of Alba (CN), and by the route B42, connecting Gallo - Bivio Castiglione (CN) to Alba (CN).
Fig. 5.32a shows that the stop is located between the Ferrero factory and one of the most popular squares of the city centre, where lots of buses stop. From the spatial characteristics of the stop, it is reasonable to expect that the major component of users is represented by workers.

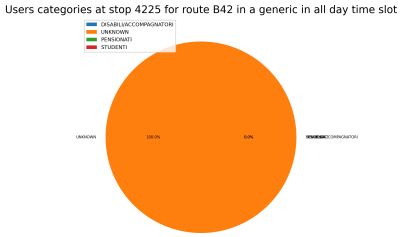
(a) Stop 4225

(b) Route B42

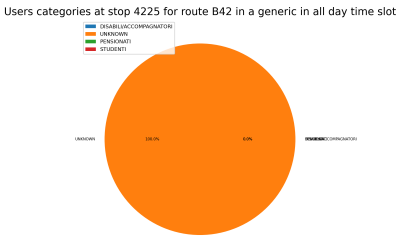
Figure 5.32: Group 4 - Stop and Route.



(a) Working Segment



(b) Holiday Segment



(c) Hybrid Segment

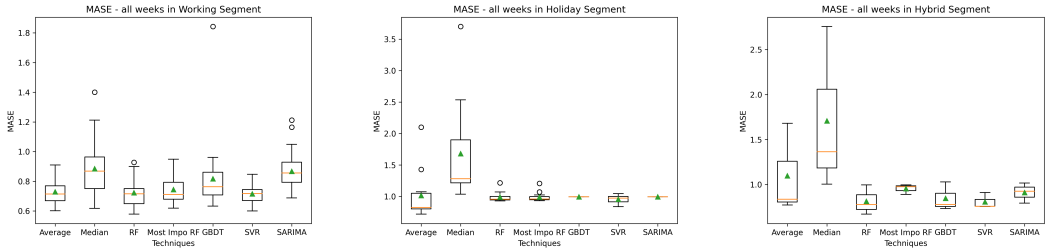
Figure 5.33: Group 4 - Users categories in different temporal segments

In group 4 the students percentage is null in all the temporal segments. This can be explained by the fact that this route has been design to link Ferrero factory to Alba (CN) and neighboring villages, so the most of the users are workers.

Table 5.10: Group 4 - Performance Metrics for all the techniques, for all the temporal segments.

	WORKING				HOLIDAY				HYBRID			
Model	N	MASE	MAE	R2	N	MASE	MAE	R2	N	MASE	MAE	R2
Average	21	0.77	1.15	0.39	28	1.01	1.66	0.10	28	1.07	1.15	0.32
Median	21	0.88	1.37	0.05	28	1.37	1.67	0.06	28	1.53	1.46	0.21
RF	21	0.72	1.15	0.41	28	0.98	1.49	0.12	**28**	**0.73**	**1.13**	**0.37**
Most Impo RF	21	0.74	1.18	0.38	28	0.99	2.33	0.18	28	0.82	1.35	0.29
GBDT	35	0.80	1.31	0.39	21	0.99	1.70	0.17	21	0.80	0.98	0.34
SVR	**28**	**0.71**	**1.15**	**0.43**	**21**	**0.96**	**1.61**	**0.24**	21	0.78	0.94	0.41
SARIMA	7	0.86	1.35	0.38	7	0.99	1.72	0.13	7	0.82	1.35	0.22

Table 5.10 shows that for working and holiday segments the best model to use is SVR, while for hybrid it is firstly RF and then SVR. The data homogeneity may be the reason for this: as can be seen in **Pie charts 5.33**, students component is not represented at all in the three temporal segments.



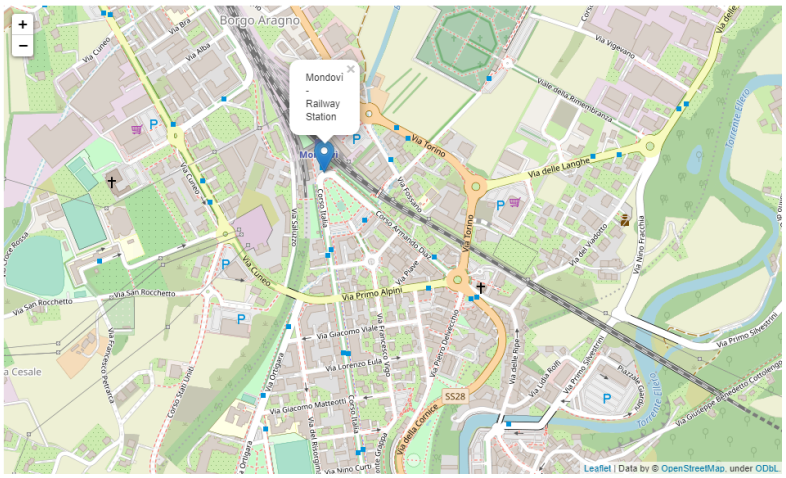
(a) Working Segment (b) Holiday segment (c) Hybrid segment

Figure 5.34: Group 4 - MASE box plots for all the temporal segments.

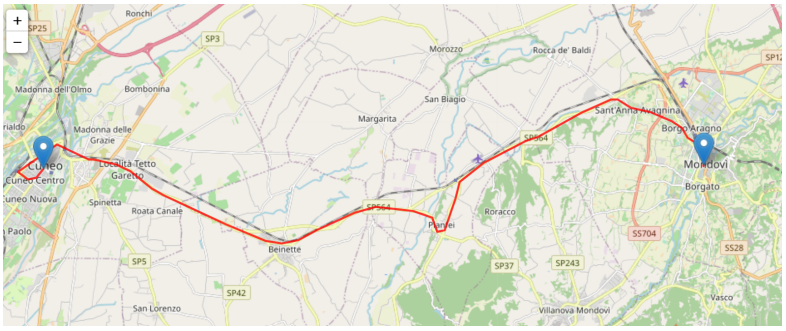
Fig. 5.34 show that the chosen models obtain almost the same performance with the exception of the Median technique. For the holiday and hybrid segments, it reports MASE mean values which are visibly higher than the others (see **Box plots 5.34b** and 5.34c).

5.8.6 Group 5: Stop 541, Route B176

The couple which represents the group 5 is composed of stop 541, the railway station of Mondovì (CN), and route B176 which links Villanova (CN) to Mondovì (CN).
As can be seen in **Fig. 5.35a**, the stop is located in front of the railway station of Mondovì, so it is an important inter-exchange point.

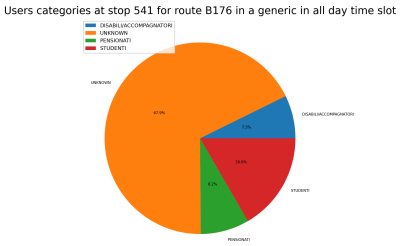


(a) Stop 541

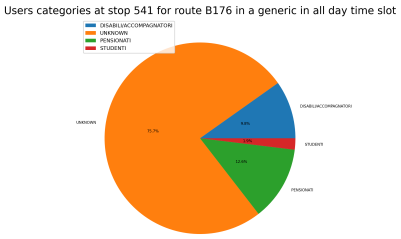


(b) Route B176

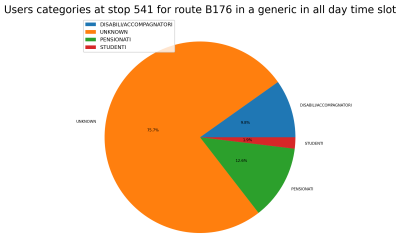
Figure 5.35: Group 5 - Stop and Route.



(a) Working Segment



(b) Holiday Segment



(c) Hybrid Segment

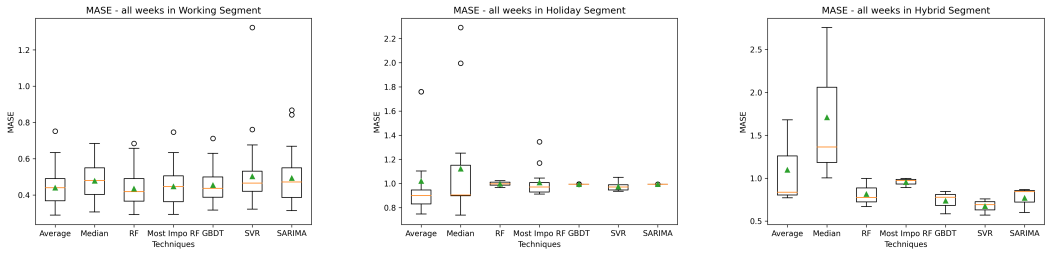
Figure 5.36: Group 5 - Users categories in different temporal segments

In cluster 5 the student component is visibly higher in the working segment (**Pie Chart 5.36a**), and its percentage is similar in holiday and hybrid segments. This justifies the results reported in **Table 5.11**: the same model (SVR) can be used to predict weeks belonging to both holiday and hybrid segments, while in the working segment it is prefereable to use RF.

Table 5.11: Group 5 - Performance Metrics for all the techniques, for all the temporal segments.

	WORKING				HOLIDAY				HYBRID			
Model	N	MASE	MAE	R2	N	MASE	MAE	R2	N	MASE	MAE	R2
Average	14	0.45	0.88	0.74	14	1.02	0.94	0.10	14	1.08	0.91	0.68
Median	14	0.47	1.01	0.76	14	1.12	1.06	0.10	14	1.54	1.02	0.56
RF	**14**	**0.43**	**0.88**	**0.75**	14	0.99	0.98	0.22	14	0.73	0.90	0.70
Most Impo RF	14	0.44	0.89	0.74	14	1.01	0.98	0.18	14	0.78	1.65	0.25
GBDT	14	0.45	0.98	0.75	14	0.98	0.99	0.25	21	0.70	1.01	0.26
SVR	14	0.49	1.09	0.26	**28**	**0.97**	**0.93**	**0.19**	**28**	**0.67**	**0.92**	**0.37**
SARIMA	7	0.48	1.01	0.70	7	0.99	1.02	0.14	7	0.72	1.41	0.21

For this cluster within the working and holiday segments (**Figg. 5.37a and 5.37b**) the techniques show MASE mean errors which are really close one to the others, while in the hybrid segment (**Fig. 5.37c**) Average and Median techniques differ from the others, with worse performance.



(a) Working Segment (b) Holiday segment (c) Hybrid segment

Figure 5.37: Group 5 - MASE box plots for all the temporal segments.

Chapter 6

Conclusions and Future Works

The use of electronic tickets allows to keep track of the travels of passengers in order to understand which are the most popular routes and at which hour they are more crowded in order to provide an efficient service that satisfies the needs of users.
To design a new network planning, schedule a customized frequency of trips and in general to optimize the allocation of resources, it is necessary to forecast the public transport demand.
To achieve this goal about 10 million validations collected in 2019 have been exploited and different machine learning models have been tested and compared. In particular, they are: RF, GBDT, SVR, SARIMA.

The first result of the study is the importance of the segmentation: it is necessary to take into account if schools are open or closed to reach good performance since the trend of validations changes a lot in these two different scenarios.
The analysis of the categories at which users belong to show that students represent a significative component and this confirms the great impact of the *school holiday* variable over the number of validations.

Another relevant point is the importance of clustering: it allows to group together couples *bus stop-route* with similar characteristics in terms of offer and demand (number of validations and users composition) so that all the elements of each cluster can be analysed using the same model within the same temporal segment.

Following this approach, the answer of the research question can be found in

the final table, where it is evident that machine learning leads to better results with respect to the other techniques which are not based on artificial intelligence. In particular, the advantage in the usage of more complex models which involve machine learning is more evident within the hybrid segment, when the trend of validations is not homogeneous over the weeks according to the different nature of the days within the same week.
This is reasonable since during the working segment the users tend to follow their own rigid daily routine, so a simpler model can be enough to forecast the demand. However, this study exploits data referring to 2019, before Covid pandemic. Today the results may be different due to the adoption of smart working and distance learning which lead to a reduction in the number of validations and to a smoother trend throughout the day as to avoid gatherings, offices and schools have introduced staggered entrances so peak hours should be no longer detected.

To improve the forecasting performance when special events occur (such as sports competitions, demonstrations, strikes), it is possible to insert some indicators that have to be taken into account in the case in which these events happen. For example, on 25 th November 2019 in the Cuneo area there was a serious flood which led to the closure of schools. Of course, the forecast results for this day are not good as this is an unpredictable event and the machine learning models have predicted the number of validations for this day as if this were a standard working day. To help in the management of these situations, some historical collected data referring to similar scenarios can be used as training.
In general, the data collected during the entire previous year could be used to train the predictive model in order to look at the same week instead of the N days immediately before the week to predict.
This strategy could improve the performance, avoiding the segmentation of the dataset into three temporal periods but since one year is a significative time interval, the demand may be changed and in this case more recent data would give better results.
Both approaches can be used to find a good compromise: the yearly one provides a forecast and the errors related to the predictions made over the previous weeks help to adjust this prediction.

Furthermore, another possible interesting future work could be to approach the problem from the service provider's point of view and interpret the difference between the actual number of validations and those expected, rather than as an error, as a starting point for understanding what did not work as expected by the model, on the offer side. This requires an analysis that focuses on the offer and finds out what are the critical factors influencing the results.

Acknowledgements

Ultima pagina, simbolica, perchè chiude non solo questo lavoro di tesi ma anche un percorso accademico, portandomi al raggiungimento di un grande traguardo.
Voglio quindi ringraziare tutte le persone per le quali nutro un grande affetto e condividere con loro questa vittoria, perchè è anche loro.
Ringrazio la Professoressa Chiusano e la Dottoressa Elena Daraio per la disponibilità ed il supporto tecnico, la Dottoressa Brunella Caroleo per il suo sorriso, entusiasmo e spiccata competenza e tutta l'area FCC della Links Foundation per avermi accolta affettuosamente.
Grazie ad i miei genitori ed alla mia famiglia, che hanno sempre creduto in me, spronandomi ad intraprendere questa strada e standomi accanto con immenso amore.
Grazie ai miei fedelissimi compagni di viaggio, Matteo e Brendan, con cui ho condiviso tanti bei momenti tra successi e scleri. In particolare, grazie a Matteo che mi ha accompagnata e sostenuta, da cui ho imparato tanto da un punto di vista tecnico e la cui presenza nella mia vita mi ha portato spensieratezza, felicità e forza, ingredienti fondamentali per affrontare ogni giorno con positività e spirito di sfida.
Grazie ai miei amici di sempre, Martina, Stefano (x2), Luca, Francesca e Deborah.
Grazie al fondamentale gruppo di amiche, Sabrina, Alessandra, Rebecca, Sara e Sara che con le loro risate hanno reso tutto più bello.
Grazie anche alle mie ex Professoresse che con la loro preziosa vicinanza ed il loro puntuale supporto mi fanno sentire speciale.
Infine, un grazie al Politecnico che mi ha permesso di imparare ad avere coraggio ed a sforzarmi costantemente per raggiungere gli obiettivi prefissati.

Bibliography

[1] «Granda Bus». In: (). URL: https://grandabus.it (cit. on p. i).

[2] «Links Foundation - About». In: (). URL: https://linksfoundation.com/en/about/ (cit. on p. i).

[3] Vera Costa, Tânia Fontes, Pedro Maurício Costa, and Teresa Galvão Dias. «Prediction of Journey Destination in Urban Public Transport». In: *Progress in Artificial Intelligence*. Ed. by Francisco Pereira, Penousal Machado, Ernesto Costa, and Amílcar Cardoso. Cham: Springer International Publishing, 2015, pp. 169–180. ISBN: 978-3-319-23485-4 (cit. on p. 4).

[4] Anne-Sarah Briand, Etienne Côme, Martin Trépanier, and Latifa Oukhellou. «Analyzing year-to-year changes in public transport passenger behaviour using smart card data». In: *Transportation Research Part C: Emerging Technologies* 79 (2017), pp. 274–289. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2017.03.021. URL: https://www.sciencedirect.com/science/article/pii/S0968090X17301055 (cit. on p. 4).

[5] Maurizio Arnone, Tiziana Delmastro, Giulia Giacosa, Mauro Paoletti, and Paolo Villata. «The Potential of E-ticketing for Public Transport Planning: The Piedmont Region Case Study». In: *Transportation Research Procedia* 18 (2016). Efficient, Safe and Intelligent Transport. Selected papers from the XII Conference on Transport Engineering, Valencia (Spain) 7-9 June., pp. 3–10. ISSN: 2352-1465. DOI: https://doi.org/10.1016/j.trpro.2016.12.001. URL: https://www.sciencedirect.com/science/article/pii/S2352146516307542 (cit. on p. 4).

[6] Arnone, Dmastro, Negrino, Arneodo, Botta, and Friuli. «Estimation of public transport user behaviour and trip chains through the Piedmont Region e-ticketing system». In: Proceedings of 14th ITS European Congress, Lisbon, Portugal. 2020 (cit. on p. 4).

[7] Martin Trépanier, Robert Chapleau, and Nicolas Tranchant. «Individual Trip Destination Estimation in a Transit Smart Card Automated Fare Collection System». In: *Journal of Intelligent Transportation Systems: Technology, Planning and Operations* 11(1) (Apr. 2007), pp. 1–14. DOI: 10.1080/15472450601122256 (cit. on p. 4).

[8] Li He and Martin Trépanier. «Estimating the Destination of Unlinked Trips in Transit Smart Card Fare Data». In: *Transportation Research Record: Journal of the Transportation Research Board* 2535 (Jan. 2015), pp. 97–104. DOI: 10.3141/2535-11 (cit. on p. 4).

[9] Florian Toqué, Mohamed El Mahrsi, Etienne Côme, and Latifa Oukhellou. «Forecasting Dynamic Public Transport Origin-Destination Matrices with Long-Short Term Memory Recurrent Neural Networks». In: Nov. 2016. DOI: 10.1109/ITSC.2016.7795689 (cit. on p. 5).

[10] Zhan Zhao, Haris Koutsopoulos, and Jinhua Zhao. «Individual mobility prediction using transit smart card data». In: *Transportation Research Part C Emerging Technologies* 89 (Apr. 2018), pp. 19–34. DOI: 10.1016/j.trc.2018.01.022 (cit. on p. 5).

[11] Haiyang Yu, Zhihai Wu, Dongwei Chen, and Xiaolei Ma. «Probabilistic Prediction of Bus Headway Using Relevance Vector Machine Regression». In: *IEEE Transactions on Intelligent Transportation Systems* 18 (Jan. 2016), pp. 1–10. DOI: 10.1109/TITS.2016.2620483 (cit. on p. 5).

[12] Fangzhou Sun, Yao Pan, Jules White, and Abhishek Dubey. «Real-Time and Predictive Analytics for Smart Public Transportation Decision Support System». In: May 2016. DOI: 10.1109/SMARTCOMP.2016.7501714 (cit. on p. 5).

[13] Md. Shalihin Othman and Gary Tan. «Predictive Simulation of Public Transportation Using Deep Learning: 18th Asia Simulation Conference, AsiaSim 2018, Kyoto, Japan, October 27–29, 2018, Proceedings». In: Oct. 2018, pp. 96–106. ISBN: 978-981-13-2852-7. DOI: 10.1007/978-981-13-2853-4_8 (cit. on p. 5).

[14] Teresa Cristóbal, Gabino Padrón, Alexis Quesada-Arencibia, Francisco Hernández, Gabriele de Blasio, and Carmelo García. «Using Data Mining to Analyze Dwell Time and Nonstop Running Time in Road-Based Mass Transit Systems». In: *Proceedings* 2 (Oct. 2018), p. 1217. DOI: 10.3390/proceedings2191217 (cit. on p. 5).

[15] Haiyang Yu, Dongwei Chen, Zhihai Wu, Xiaolei Ma, and Yunpeng Wang. «Headway-based bus bunching prediction using transit smart card data». In: *Transportation Research Part C: Emerging Technologies* 72 (Nov. 2016). DOI: 10.1016/j.trc.2016.09.007 (cit. on p. 5).

[16] Noor Asmael and Mohanned Waheed. «Demand estimation of bus as a public transport based on gravity model». In: *MATEC Web of Conferences* 162 (Jan. 2018), p. 01038. DOI: 10.1051/matecconf/201816201038 (cit. on p. 5).

[17] Chuan Ding, Donggen Wang, Xiaolei Ma, and Haiying Li. «Predicting Short-Term Subway Ridership and Prioritizing Its Influential Factors Using Gradient Boosting Decision Trees». In: *Sustainability* 8 (Oct. 2016), p. 1100. DOI: 10.3390/su8111100 (cit. on p. 6).

[18] Yang Liu, Zhiyuan Liu, and Ruo Jia. «DeepPF: A deep learning based architecture for metro passenger flow prediction». In: *Transportation Research Part C Emerging Technologies* 101 (Feb. 2019), pp. 18–34. DOI: 10.1016/j.trc.2019.01.027 (cit. on p. 6).

[19] Jianyuan, Zhen Guo, Ying Xie, Limin Jia Qin, Yaguan, and Wang. «Short-Term Abnormal Passenger Flow Prediction Based on the Fusion of SVR and LSTM». In: *IEEE* (Mar. 2019). DOI: 10.1109/ACCESS.2019.2907739 (cit. on p. 6).

[20] Milos Milenkovic, Libor Svadlenka, Vlastimil Melichar, Nebojsa Bojovic, and Zoran Avramović. «SARIMA modelling approach for railway passenger flow forecasting». In: *Transport* (Oct. 2015), pp. 1–8. DOI: 10.3846/16484142.2016.1139623 (cit. on p. 6).

[21] Florian Toqué, Mostepha Khouadjia, Etienne Côme, Martin Trépanier, and Latifa Oukhellou. «Short and long term forecasting of multimodal transport passenger flows with machine learning methods». In: Oct. 2017, pp. 560–566. DOI: 10.1109/ITSC.2017.8317939 (cit. on p. 6).

[22] Pengpeng Jiao, Ruimin li, Tuo Sun, Zenghao Hou, and Amir Ibrahim. «Three Revised Kalman Filtering Models for Short-Term Rail Transit Passenger Flow Prediction». In: *Mathematical Problems in Engineering* 2016 (Mar. 2016), pp. 1–10. DOI: 10.1155/2016/9717582 (cit. on p. 6).

[23] ISTAT: Istituto nazionale di STATistica. «Descrizione dei dati geografici e delle variabili censuarie delle Basi territoriali per i censimenti: anni 1991, 2001, 2011». In: (2016). URL: https://www.istat.it/it/files//2013/11/Descrizione-dati-Pubblicazione-2016.03.09.pdf (cit. on p. 11).

[24] Andrea Attili. «The demand for public transport: analysis of mobility patterns and bus stops.» In: *Rel. Silvia Anna Chiusano. Politecnico di Torino, Corso di laurea magistrale in Ingegneria Matematica, 2021* (Feb. 2021) (cit. on p. 14).

[25] Imad Dabbura. «K-Means». In: (). URL: https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a (cit. on p. 14).

[26] L. Serrano. *Grokking Machine Learning*. Manning Publications, 2021. ISBN: 9781617295911. URL: https://books.google.it/books?id=jJiDzQEACAAJ (cit. on pp. 19, 20).

[27] In: (). URL: https://www.machinelearningplus.com/time-series/augmented-dickey-fuller-test/ (cit. on p. 41).

[28] Taylor G. Smith et al. *pmdarima: ARIMA estimators for Python*. [Online; accessed <today>]. 2017–. URL: http://www.alkaline-ml.com/pmdarima (cit. on p. 41).