HarvardX Data Science Capstone
Project 2: Italian Airports Flights Prevision

Pietro D'Ambrosio

November 5, 2019

1. Executive Summary

This project aims to examine the public data of the main Italian airports with the emphasis on analyzing the volume of "movements" (departures and landings) starting from available historical data.

The input data are acquired from the web portals by the control bodies, some specialized sites and the trade associations.

Obviously, the forecast of flight "movements" is only a first demonstrative elaboration of data extracted from different sources in order to obtain useful information to predict evolutionary trends and market spaces for service companies operating in the aviation sector and cargo transporters.

The choice to address an original theme without being able to count on data sets ready for processing involved a considerable effortin the initial collection of data and in the data cleaning activity.

Furthermore, at this stage much more information was collected than was necessary for this specific project, so that it can be made available for other uses.

From the point of view of modeling, the particularity of the problem required the evaluation of various regression models and the use of techniques to found and manage recurrent elements in time series.

In the end the chosen model is composed of a set of two models chosen among the various experimental models and led to a Rsquared result of about 0.99 and an overall error, on the verification data, less than 5%.

2. Introduction

This section describes the objectives of the project, the input dataset and the key steps that were performed to achieve the results presented in this report.

2.1. Goal of the project

The main objective of the project is to create a system to predict the number of flights departing or arriving from a given airport starting from a series of historical data and context information. The objective is to identify the areas of development and guide the organizational and investment choices made by the operators in the sector.

For the purposes of the project we have limited the processing to the prediction of a single value (the total number of flight movements), but it is easy to assume an extension of the predictive process to other characteristic information, such as the number of passengers and the volume of the cargo handled.

2.2. Input data

As already mentioned, there is no starting dataset available. The dataset was built as part of the project by acquiring data from various web sources and doing all the necessary normalization and integration phases.

In particular the public data available on the following sites were used:

www.assaeroporti.com

Assaeroporti, established on 31 May 1967, is the Association of Italian Airport Management Companies.

Assaeroporti is a member of Confindustria, Federtrasporto and IFSC, and is present in Europe as a member of ACI EUROPE, the association of European airport managers.

Assaeroporti represents 33 airport management company from 42 italian airports (185.7 million passengers, 1,6 million flights, 1.1 million tons of goods).

www.flightradar24.com

Flightradar24 is a global flight monitoring service that provides real-time information on thousands of aircraft worldwide. Flightradar24 tracks over 180,000 flights in real time, from over 1,200 airlines, flying to or from over 4,000 airports worldwide.

For each of these airports Flightradar24 also provides detailed information as well as some interesting indexes including an airport "rating" that includes various assessments on the quality of the structure.

We believe this information can be a decisive predictor of establishing an airport's growth trend.

https://openflights.org/

OpenFlights is an open source project aimed at offering public tools and data to map flights all over the world, search for them and filter them in all ways and automatically calculate relative statistics.

On this portal a series of data is available including the list of all airport codes, all aircraft, all airlines and all routes (flight to and from any airport in the world).

We will use the "routes" data to calculate a new qualitative index based on the number of destinations of a given airport (the criterion is that the more routes are available the more important the airports are).

2.3 Key steps performed

After downloading the data, the first major difficulty encountered was the use of different airport codes in the various datasets.

The data of Assaeroporti carry the names in Italian (and in some cases the same airport is displayed with different names, if over the years it has changed its name).

The other datasets instead report the names of the airports in English or use the international 3-digit codes (eg "FCO" to indicate the "Leonardo da Vinci" airport in Rome Fiumicino).

After this first normalization phase it was possible to integrate the various data of interest, checking and removing the error cases and the non-significant data (eg very small airports without passenger or cargo traffic).

At this point a descriptive analysis of the data was carried out, through specific research, visualizations and graphics, in order to identify useful functions to define the best possible approach to processing.

Subsequently we proceeded to the definition of various modeling hypotheses starting from a very simple model and gradually increasing the level of complexity.

For each modeling hypothesis, a verification of the RMSE (and, as we'll see, of another LOSS functions) was performed and all the values were summarized in a table to allow a more immediate comparison of the results.

3. Methods and analysis

This section explains the process and techniques used, such as data cleaning, data exploration and visualization, the various insights gained, and the modeling approach chosen.

3.1. Process and techniques

The work begins with a quantitative analysis of the acquired data in order to understand the property of the available features and evaluate the need for any pre-processing activity.

Then a data preparation and data cleaning phase is carried out to proceed with the subsequent modeling phases. Finally, the data visualization phase allows the analysis of the single available features and the identification of the most suitable processing strategies to solve the problem.

3.1.1. Descriptive analysis

The "pd_italian_airports" dataset, obtained by processing the various input sources (using the script associated with this project) contains **5,661** rows and **35** variables.

The following table shows the first lines of the dataset.

	airport	year	month	tm_mov_naz	p_tm_mov_naz	tm_mov_internaz	p_tm_mov_internaz
594	Alghero	2003	1	604	12.26766	76	0.00000
595	Ancona	2003	1	488	14.82353	402	13.23944
596	Bari	2003	1	1226	-4.66563	284	226.43678
597	Bergamo	2003	1	366	-31.46067	2310	33.29486

	tm_mov_UE	p_tm_mov_UE	tm_mov_tcomm	p_tm_mov_tcomm	tm_mov_tgenav
594	62	0.00000	680	10.74919	46
595	263	42.93478	890	14.10256	746
596	178	0.00000	1510	9.97815	142
597	1828	27.83217	2676	18.04146	141

	p_tm_mov_tgenav	tm_mov_tot	p_tm_mov_tot	tp_mov_naz	p_tp_mov_naz
594	64.28571	726	13.084112	39735	49.166604
595	34.90054	1636	22.730683	15851	8.904157
596	-23.24324	1652	6.033376	87910	16.915588
597	-39.74359	2817	12.634946	5218	-65.861956

	tp_mov_internaz	p_tp_mov_internaz	tp_mov_UE	p_tp_mov_UE	tp_mov_tcomm
594	9259	24.73394	8296	19.62509	139
595	18970	52.30831	16175	58.78080	12
596	11208	870.38961	4430	0.00000	1203
597	98609	119.81498	80295	135.31049	1937

	p_tp_mov_tcomm	tp_mov_tgenav	p_tp_mov_tgenav	tp_mov_tot	p_tp_mov_tot	cod3
594	-26.84211	49133	43.44983	44	144.44444	AHO
595	100.00000	34833	28.93471	989	46.73591	AOI
596	778.10219	100321	31.16771	218	-19.55720	BRI
597	178.30460	105764	73.83672	229	-34.00576	BGY

	cod4	rating	nroutes	nro	tm_mov_tot_nm	tm_mov_tot_pycm	tm_mov_tot_pynm
594	LIEA	70	1922	4	670	642	796
595	LIPY	77	200	2	1496	1333	1243
596	LIBD	72	5000	5	1511	1771	3148
597	LIME	73	13122	6	3162	2501	2223

The value to be expected, instead, is contained in the **"tm_mov_tot_nm"** field corresponding to the value of the month following the current month. This value has a very high variability, as shown in the following table:

Measures	Values
Min value	2
Max value	32,931
Mean value	3,989.3
SD value	5,156.18

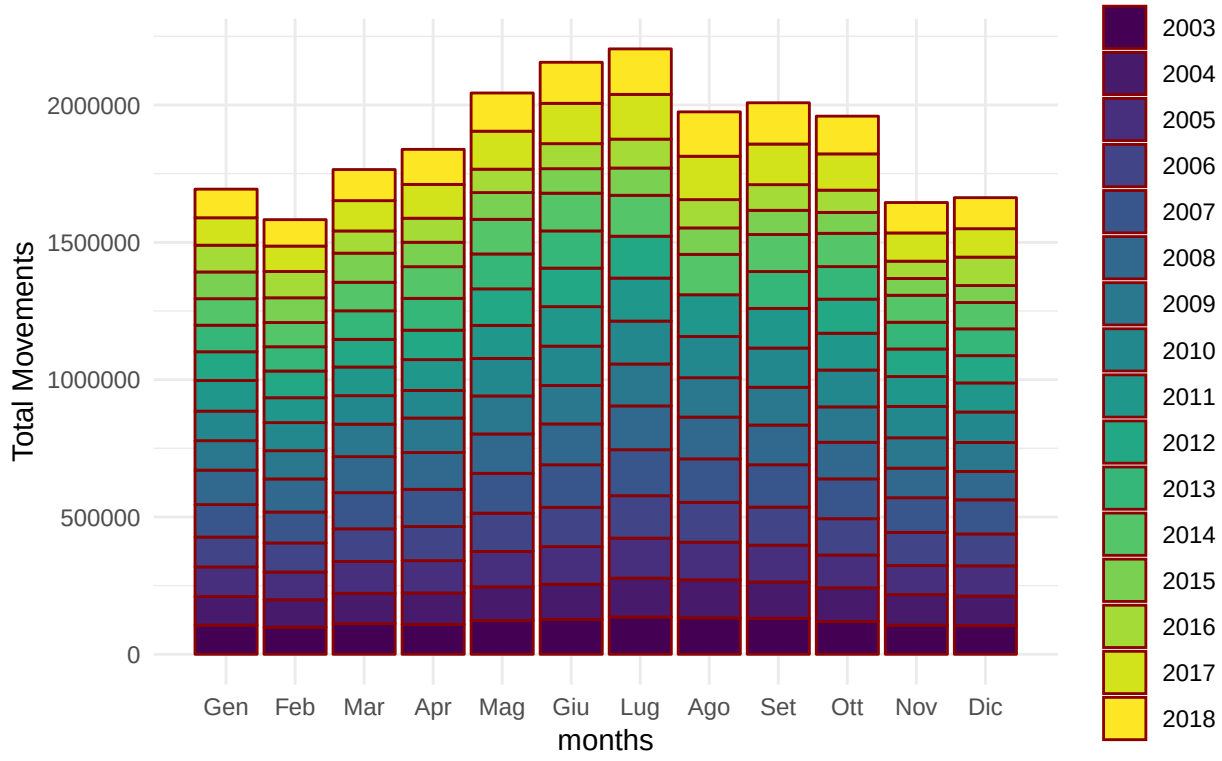
The dataset (in raw and clean version) ,for convenience and for any possible future use, have been made available on thefollowing GitHub repository: **https://github.com/pietrodambrosio/HarvardX_DataScience** together with a **data dictionary** which contains a description of all the variables contained in the dataset.

Most of the information in the dataset is numerical and refers to the data collected by the **"www.assaeroporti.it"** site, others have been determined based on the historical series (eg the values of the current month and the next month of the previous year), others were acquired from other sources such as the "rating" of the airport and the number of routes (from which the "nro" index was obtained).

In this phase we will try to understand how the values of some characteristic variables are distributed.

Month distribution of mov_tot values

As a first observation we would like to understand if there are substantial differences on the average values of the various months of the year. In fact, we expect that in particular months such as summer or Christmas holidays, there may be more travel and therefore more flights at all airports.



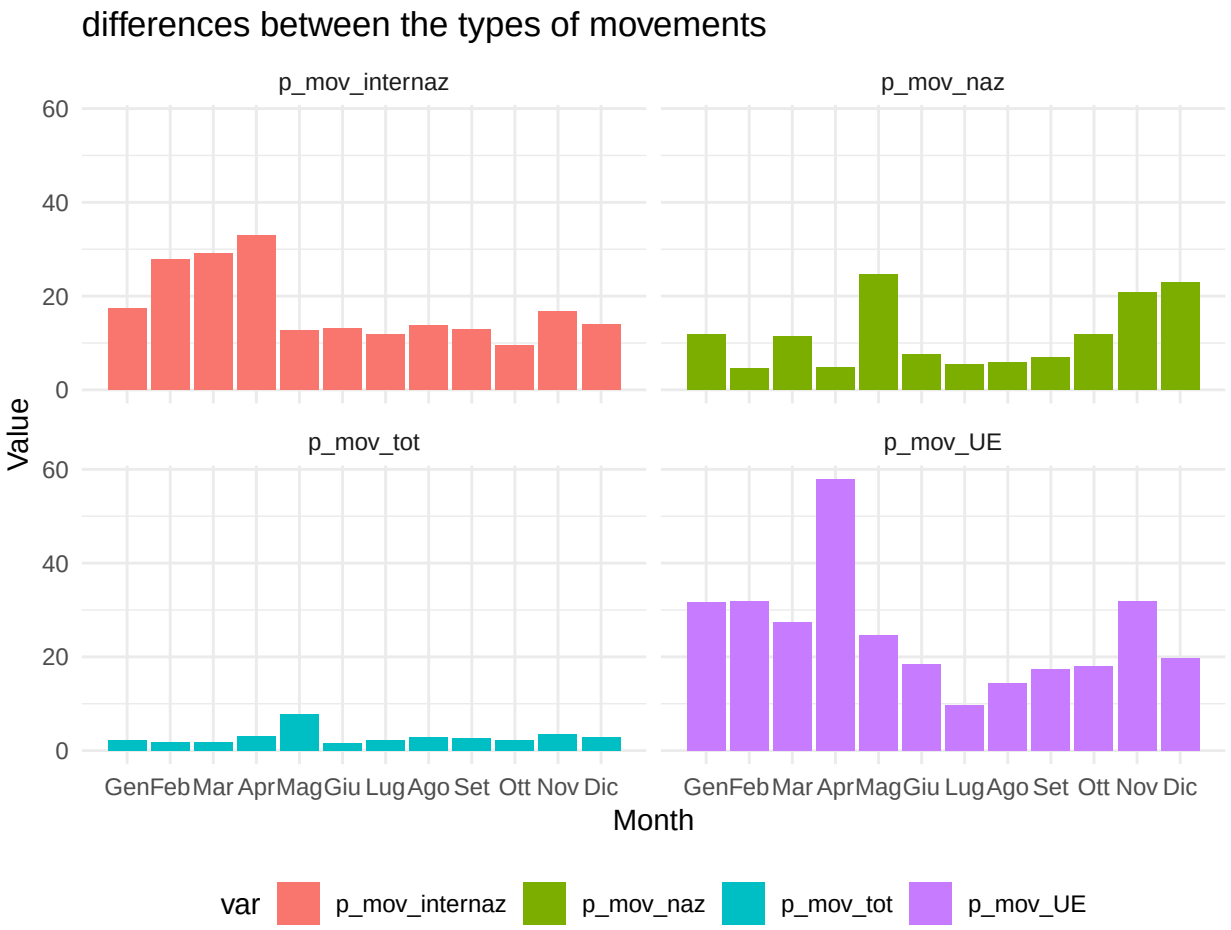
The previous graph clearly shows that, strangely, there is no increase in the number of flights in the last months of the year, while the growth expected for the summer period is confirmed. In any case, on average

(also taking into account a wide range of years), the traffic volume data vary within the year. This applies to all years.

From the graph, of course, we have excluded the years 2001 and 2009 because they are not complete. The graph also shows that substantially the monthly averages are substantially unchanged over the years.

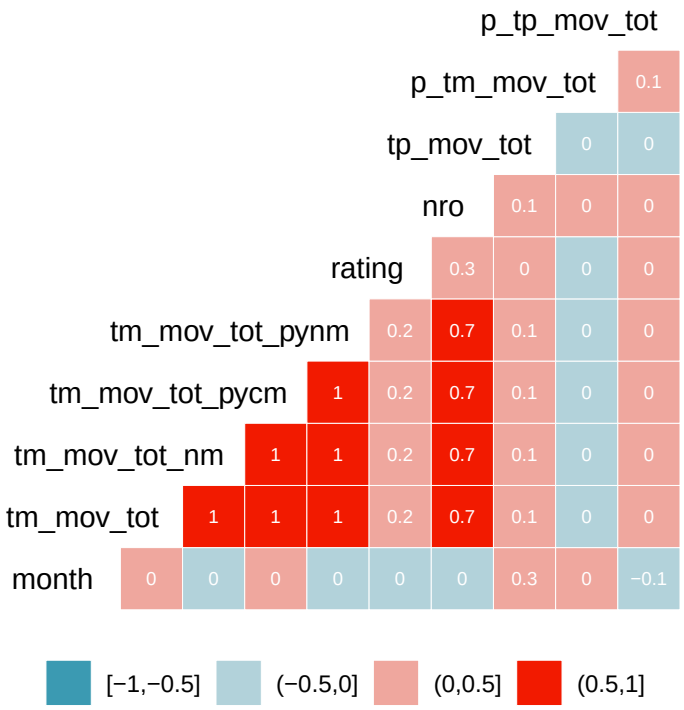
Month distribution of others values

Let's try to understand now if, instead, there are differences between the types of movements (national, international, UE) compared with the total movements.



The graph above shows some variability in the monthly average values of the percentage changes.

Correlation matrix



The correlation matrix, which was created on the fields of greatest potential interest of analysis, shows a strong correlation of the number of total movements with some variables as, for examples, the corresponding values of previous year.

It can be noted, however, that other variables, including the "rating", are not very correlated with all other variables and therefore could be useful in the model.

3.1.3. Data Pre-processing

The value that we have to forecast is the total of the movements of the month following the current one. For the purposes of the forecast, considering the characteristic periodicity of these values and the obvious relationship with the airport size, we think it may be useful to integrate the input dataset with other information obtainable from historical data and in particular the values of previous years (current and next month).

It is also necessary to add to the data the value of the following month (the one we want to predict) to be able to train the algorithm.

So, first of all we insert in the database the field that must be predicted tm_mov_tot_nm, taking it from the data of the following month. Then create previous year's values for the current month and the following month: * tm_mov_tot_pycm = total month movements of previous year - current month * tm_mov_tot_pynm = total month movements of previous year - next month The choice to use the values of the previous years obliges us to no longer be able to take into consideration the values of the year 2002 (as they will be used to enrich the data of the following year).

The **"clean"** dataset now contains **5,661** rows and **35** variables.

3.2. Modeling approach

For the choice of the best approach we will proceed step by step, trying out various models (starting from the simplest) and then we will try to create a "set" of the best performing methods to see if we get better results.

The models will be tested on a subset of the randomly chosen dataset with a size equal to 20% of the complete dataset.

Below we will present the individual models evaluated, starting from the simplest possible (the simple replication of the previous year's value) and continuing with the addition of other factors. In some cases we used "cross-validation" for model optimization, while for model evaluation we used two different metrics:

- the typical RMSE error loss function (mean residual square error), as defined by the following formula:

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}$$

- a MAPE function (Mean Absolute Percentage Error) that can be more intuitive and help us better understand how our forecast approaches reality. The our MAPE function is defined as follow:

$$MAPE = \sum_{i=1}^{N}(\hat{y}_i - y_i) / \sum_{i=1}^{N}(y_i)$$

3.2.1. Model 1 - First basic model

This model is as simple as possible and obviously serves as a reference point for the analysis of the performance of the various models we will build and to understand whether we are proceeding in the right way or not.

In this model we use the next month data of previous year to predict the next month of this year. In this case we are not interested in training the model so we define this measure on the entire available dataset.

From the evaluation carried out on the test dataset this model presents an RMSE of **561.1441**.

3.2.2. Model 2 - add a percent variation of current month

Now let's see how we can add some elements to the model to consider "the percentage change of the current month" compared to the previous year. This information is already available in the dataset and is contained in the field: **"p_tm_mov_tot"**.

the RMSE calculated with this model is equal to **430.476**. both the RMSE and the MAPE value are better than those obtained with the previous model, as shown in the following table.

method	RMSE	MAPE
Model 1 - simply next month of previous year	561.1441	0.0756
Model 2 - add percent variation of current month	430.476	0.0502

3.2.3. Pre-processing necessary to use real machine learning models

Before starting to do a real machine learning process, we have to do two things: 1) prepare the train set and the test set starting from the complete dataset, 2) eliminate the columns that do not add information to the model (like cod3 and cod4 which are only other encodings of the airport field). After these elaborations we will have a "train" dataset, containing 80% of the data of the input dataset (cleaned) and a "test" dataset containing the remaining 20% of the data. The choice of rows to include in test and train dataset was made randomly.

3.2.4. Model 3 - 'glmboost' (gradient boosting)

We start using a gradient boosting method. Gradient boosting is a machine learning technique for regression (and classification) problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It is a very powerful technique for building predictive models. Gradient boosting is applicable to many different risk functions and optimizes prediction accuracy of those functions, which is an advantage to conventional fitting methods.Gradient boosting then is a useful tool for predictive tasks, and provides higher accuracy results compared to conventional single strong machine learning models.

the RMSE calculated with this model is equal to **476.2927** and it is worse than that obtained with the two simple previous models, as shown in the following table:

method	RMSE	MAPE
Model 1 - simply next month of previous year	561.1441	0.0756
Model 2 - add percent variation of current month	430.476	0.0502
Model 3 - 'glmboost' model (gradient boosting)	476.2927	0.0644

3.2.5. Model 4 - 'saic' (glm with Stepwise Feature Selection)

The next method we use is the "glmStepAIC" (called breafly "saic"). This model starts from the best one-variable model, and each time add the variable that brings the largest improvement (in terms of AIC). AIC stands for "Akaike's Information Criteria", a metric developped Hirotugu Akaike in 1970. The basic idea of AIC is to penalize the inclusion of additional variables to a model adding a penalty that increases the error when including additional terms. Then the lower the AIC, the better the model.

This model worked better than the previous one (with RMSE of **439.1131**) but it is still worse than one of the two very simple initial models.

method	RMSE	MAPE
Model 1 - simply next month of previous year	561.1441	0.0756
Model 2 - add percent variation of current month	430.476	0.0502
Model 3 - 'glmboost' model (gradient boosting)	476.2927	0.0644
Model 4 - 'saic' model (glm with Stepwise Feature Selection)	439.1131	0.0598

3.2.6. Model 5 - classical 'random forest' model

Now let's use the classical "rf" model. Random forests are an overall learning method for classification, regression and other machine learning activities. It works by building a multitude of decision trees at the time of training and producing the class (classification) or the average forecast (regression) of the individual trees. This type of process may take a long time to optimize and execute, so in this case we used parallel processing (12 thread).

This model achieved an RMSE of **382.6312** and has been the best of all the models examined so far, as shown in the following summary table.

method	RMSE	MAPE
Model 1 - simply next month of previous year	561.1441	0.0756
Model 2 - add percent variation of current month	430.476	0.0502
Model 3 - 'glmboost' model (gradient boosting)	476.2927	0.0644
Model 4 - 'saic' model (glm with Stepwise Feature Selection)	439.1131	0.0598
Model 5 - classical 'random forest' model	382.6312	0.0499

3.2.7. Model 6 - 'ranger' model (fast implementation of random forests)

The "ranger" method is a fast implementation of random forests (Breiman 2001) or recursive partitioning, particularly suited for high dimensional data. It is quite recent (2017) and is used very often for some types of processing that require efficiency (for example Data from genome-wide association studies).

This model is even better than the previous one but above all it should be noted that the processing required a considerably smaller time than the classic random forest implementation.

method	RMSE	MAPE
Model 1 - simply next month of previous year	561.1441	0.0756
Model 2 - add percent variation of current month	430.476	0.0502
Model 3 - 'glmboost' model (gradient boosting)	476.2927	0.0644
Model 4 - 'saic' model (glm with Stepwise Feature Selection)	439.1131	0.0598
Model 5 - classical 'random forest' model	382.6312	0.0499
Model 6 - 'ranger' model (fast implementation of random forests)	375.8211	0.0493

3.2.8. Model 7 - 'xgb' (extreme gradient boosting)

The "XGBoost"" is an algorithm that has recently been dominating applied machine learning and Kaggle competitions for structured or tabular data. It is an implementation of gradient boosted decision trees designed for speed and performance and was developed as a research project at the University of Washington. It is a great combination of software and hardware optimization techniques to yield superior results using less computing resources in the shortest amount of time.

The xgb model turned out to be the best performing model and recorded a RMSE of **360.7649** and a MAPA of **0.048**.

method	RMSE	MAPE
Model 1 - simply next month of previous year	561.1441	0.0756
Model 2 - add percent variation of current month	430.476	0.0502
Model 3 - 'glmboost' model (gradient boosting)	476.2927	0.0644
Model 4 - 'saic' model (glm with Stepwise Feature Selection)	439.1131	0.0598
Model 5 - classical 'random forest' model	382.6312	0.0499
Model 6 - 'ranger' model (fast implementation of random forests)	375.8211	0.0493
Model 7 - 'xgb' model (extreme gradient boosting)	360.7649	0.048

We will consider the last two models (those that have registered the best performances) to verify if, using an "ensable" technique, we can obtain a better results.

3.2.9. Model 8 - Ensamble of 'ranger' and 'xgb' models (best performed)

Our overall forecast will be realized considering an equal contribution of the two previous forecasts (ranger and xgb) to determine a new forecast value with this formula: **prev_ensamble = 50% prev_ranger + 50% prev_xgb**.

method	RMSE	MAPE
Model 1 - simply next month of previous year	561.1441	0.0756
Model 2 - add percent variation of current month	430.476	0.0502
Model 3 - 'glmboost' model (gradient boosting)	476.2927	0.0644
Model 4 - 'saic' model (glm with Stepwise Feature Selection)	439.1131	0.0598
Model 5 - classical 'random forest' model	382.6312	0.0499
Model 6 - 'ranger' model (fast implementation of random forests)	375.8211	0.0493
Model 7 - 'xgb' model (extreme gradient boosting)	360.7649	0.048
Model 8 - Ensamble of 'ranger' and 'xgb' models (best performed)	358.6764	0.0475

4.2. Modeling performance

The "ensamble" model obtained an RMSE of **358.6764** and a MAPE of **0.0475**, so it was the best of all the models developed. The RMSE result obtained is very good if we consider the high value of the standard deviation (5,156.18) of the "tm_mov_tot" field and its average value (3,989.3). In fact the MAPE obtained shows that the average error we record with this model is less than **5%** of the reference value.

5. Conclusion

The most complex aspect that has been addressed in the project has been the preparation of the input dataset starting from the information publicly available on the websites of some international organizations and of a trade association that gathers most of the Italian airports (Assaeroporti).

The subsequent phases of the work were aimed at carrying out a descriptive analysis of the information contained in the input datasets with the aim of obtaining useful insights to guide the subsequent modeling phases. As regards predictive models, models based on linear regression (with increasing complexity) were first tested. Then we proceeded to apply the random forest and gradient boosting family which showed the best results. The model chosen was an "ensamble" of two model and obtained a final RMSE equal to **358.6764** and a MAPE of **0.0475**.

It would be interesting to use the other information available in the input dataset to obtain other types of forecasts and experiment with further pre-processing to increase the accuracy of the chosen model. It could also be useful to try to extend the forecast to two or more months ahead, starting from the available historical data.