**Ramp Competition**

**Project Report**

**By Paul-Antoine Girard & Adrien Toulouse**

**Username: toulouse\_girard**

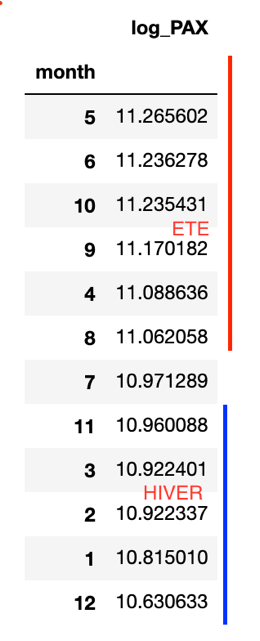
**Final submission name: ???**

***The aim of this report is to narrate our workflow, to detail everything we tried and explain the choices we made to build our best model and predict the number of air passengers.***

## Starting point, descriptive statistics and data preprocessing

Understanding the task and how to proceed was our first challenge as it was for both of us our first data science competition.

The first thing we did was look at the data we were given to try to look for patterns, correlations, missing values and to see which variables could help predict the number of air passengers.



Une image contenant capture d’écran

Description générée automatiquementFor example, we looked closely at the date after the one hot encoding.

Our analysis showed us that there was a difference in the number of air passengers for the different weekdays and the different months. Therefore, we built a variable ‘*weekend’* and a variable ‘*season’* to better capture this information. These variables improved our score when we used specific regressors such as SVM. However, they didn’t improve our score in our final gradient boosting model, so we didn’t keep them in our final submission.

Another thing we did was to make the date continuous. As dates are cyclical … etc

* Airport of departure and arrival mean encoding
* Regrouping WeeksToDeparture and their mean std
* Ajouter des jolis graphes

Looking at the external weather dataset, we didn’t find many correlations between the variables and the number of air passengers. After numerous tries (regrouping variables, doing operations between them like multiplying humidity and temperature), we ended up only keeping the variable ‘*Events’* which we thought could have an influence on the number of air passengers*.* As this variable was categorical, we encoded it into three numerical values:

* 0 when there were no particular events (we replaced missing values by 0)
* 1 for common events such as rain or fog
* 2 for more important events such as thunderstorms.

Finally, we thought that the weather data was not the necessarily the best thing to really improve our model. Since the variables we initially had didn’t allow us to explain enough the number of air passengers, we decided to add new external data.

## Adding relevant external data

Overall, we added the following four variables:

1. Distance in km between the airports: ‘Distance’
2. Number of inhabitants per city: ‘Population’
3. Revenue per state: ‘Revenue’
4. Crude oil price: ‘Oil\_Price’
   1. **‘Distance’**

Our first idea was to add the distance between the airports. We had the intuition that longer flights would carry more passenger and we thought that this information could help our algorithm to make a difference between long flights with more passengers and shorter flights.

We therefore found the geographical longitude and latitude of each airport and used a function to calculate the distance between each of the twenty airports.

This information consequently improved our score, so we decided to keep this variable.

We tried at some point to regroup the distance between airports into three categories of flights that we could encode: long flights, medium flights and short flights. However, this encoding did not allow us to improve or score and we lost a bit of precision, so we decided to stick with the distance variable.

* 1. **‘Population’**

Our second idea was that the number of passengers on a flight could be explained by the population of the city of arrival and that flights going to big cities would carry more passengers.

Therefore, we created a column with the number of passengers.

* 1. **‘Revenue’’**

**.**

**.**

* 1. **‘Oil\_price’**

.

.

## Choosing a machine learning algorithm: “Honey badger don’t give a crap “

After comparing the results from several regressors among which Random forest, Svm and Logistic regression, we decide to go for a gradient boosting algorithm as it fitted the best to our data and gave us the best score. We finally opted for the XGBRegressor.

Advantage of gradient boosting:

* F
* F

To try to improve the speed our model, we also tried to implement the LGBMRegressor. This regressor was quite performant too and a lot faster when we tried to test it with RAMP\_submission\_test. However, when we submitted with this regressor on the test dataset of the RAMP competition, we didn’t get results as good as with the XGBRegressor. Therefore, we dropped LGBM and kept XGB.

## Parameter tuning

Everytime we tried a new model or tried an existing model on new data, we tuned and adapted the parameters of our regressor to fit the data as well as possible. To do so, we used a cross validation GridSearch based on minimizing the RMSE.

For our best model, we used the following parameters:

* Max\_leaf.. Describe most important parameters and their effects
* …
* ..

## Score obtained, lessons learnt and conclusion

**In the end, our model obtained a RMSE of ??? and ranked ??? in the RAMP competition.**

Our good score was a combination of all the things we tried during these several weeks of competition: adding new useful variables to our dataset, preprocessing all our variables, choosing the best machine learning algorithm and its best parameters to fit our data.

Our workflow was very circular: every time we added new data or did relevant feature preprocessing, we would analyze how our score was impacted and try to choose the best regressor and adapt it so that it would fit the best the data.

We tried to improve our score at every step of our workflow and would continuously go back and forth all the steps.

Overall, this project was particularly interesting to understand the importance of data preparation and the particularities of the different regressors. It was also fun to look for new external data and see how they could improve our model. Undoubtedly our model could still be improved by adding more data and going deeper into parameter tuning.