UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

"Digital Automation Engineering" Master Degree
Course "Artificial Intelligence *and* Data Science"

Comparison between 3 different Machine Learning models for a Forecast regression based on US flights

Data Science Project

Simone Giovanardi 275785@studenti.unimore.it

Introduction

The aim of this project is to compare three different models used in supervised machine learning and to evaluate which is the better in similar cases to that studied in this paper. In particular the study regards a forecast regression applied on a dataset of U.S. monthly airline traffic from 2003 to 2023 [1].

Dataset choice

From the above mentioned dataset, the data that was considered is the "Total Air Travel Passengers". This can be one of the first steps in an analysis conducted by an airline company in order to predict routes and number of flights to be scheduled. A line graph of the data is reported below (Figure 1):

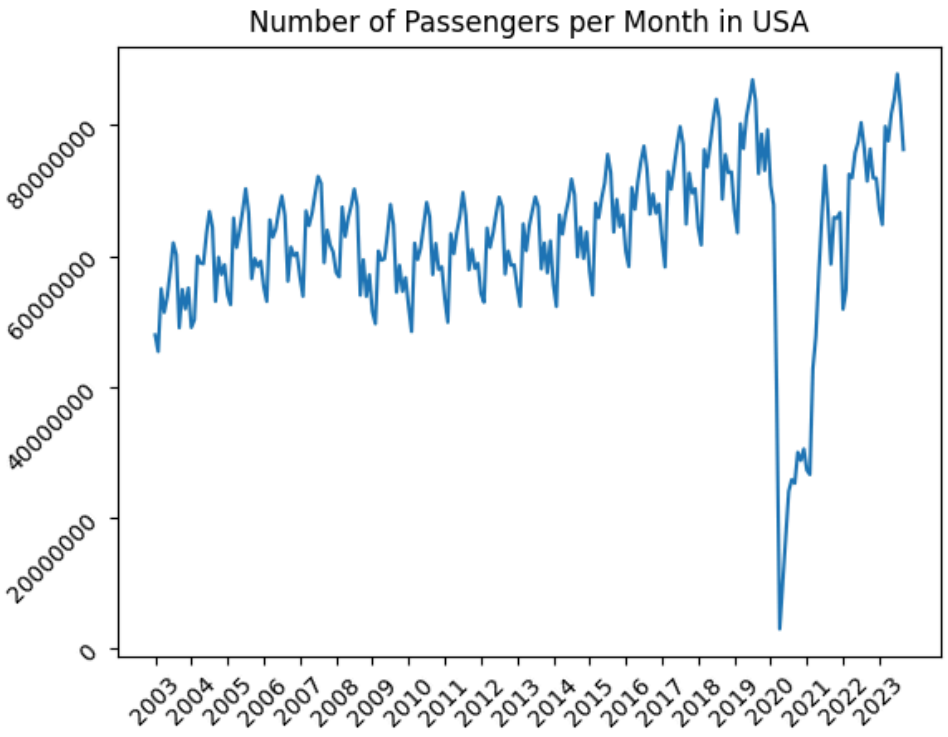

Figure 1. Number of passengers (Y-axis) per month in the USA.

It is possible to see that in 2020 there was a drastic reduction in the number of passengers and more generally in flights, and this is due to the Covid pandemic and associated restrictions. This is a highly unpredictable event and this external variable may affect the accuracy of the models and increase the error. From the graph it can be seen that there are frequent trends in base of the months of the year.

Models comparison

The dataset was splitted in the training and testing phases, considering the partition into windows (data needed to predict the output value) and the predictions. The validation phase was not performed due to the simplicity of the project and also because the number of data was not too large.
The three different model implemented are:

- Random Forest;
- Linear Regression;
- KNN K-Nearest Neighbors.

The metric which is used to evaluate and compare the different algorithms is the mean absolute error; also the "Mean squared error" was calculated, but the error was much greater in all three cases and it was preferred to omit it. With regard to the hyperparameters, window length and the number of neighbors for the KNN algorithm was taken into account, then some attempts were made to evaluate the impact of them on the goodness of the result. One could also try varying the prediction, but it was considered only the following month related to the window, instead of 2 or more months in the future.
The following values were considered:

- window = [6, 12, 24, 36]

#note the multiplicity of 12 (except for the 6, just to try a low number) in base of the type of data used and for the differences among a month and another in a year

- number of neighbors = [5, 10, 15]

Below, 4 different tables are reported. Each table collects the same value for the window hyperparameter and the 3 possibilities for the other hyperparameter. Cells are filled with the correspondent error.

window = 6	Neighbors = 5	Neighbors = 10	Neighbors = 15
Random forest	12004858.24	12004858.24	12004858.24
Linear regression	**7378778.26**	**7378778.26**	**7378778.26**
KNN	13161310.30	13799651.42	13630323.41

Figure 2.1. Table of Mean absolute errors with window = 6.

window = 12	Neighbors = 5	Neighbors = 10	Neighbors = 15
Random forest	18966534.18	18966534.18	18966534.18
Linear regression	18559158.33	18559158.33	18559158.33
KNN	**14277986.47**	**14653979.94**	**14792068.48**

Figure 2.2. Table of Mean absolute errors with window = 12.

window = 24	Neighbors = 5	Neighbors = 10	Neighbors = 15
Random forest	20579496.98	20579496.98	20579496.98
Linear regression	**10566843.38**	**10566843.38**	**10566843.38**
KNN	17202216.43	17150948.22	17559827.58

Figure 2.3. Table of Mean absolute errors with window = 24.

window = 36	Neighbors = 5	Neighbors = 10	Neighbors = 15
Random forest	21570421.98	21570421.98	21570421.98
Linear regression	**11631044.35**	**11631044.35**	**11631044.35**
KNN	19790896.04	19384082.19	19719457.46

Figure 2.4. Table of Mean absolute errors with window = 36.

Obviously the error metric in Random Forest and Linear Regression methods do not vary with the number of neighbors.

Conclusions

Seeing the tables in the previous section (Figures 2) some considerations can be made:

- In 75% of the cases, the Linear Regression is the more efficient method;
- Varying the number of neighbors in KNN model, doesn't affect in a relevant way the performance of the algorithm;
- The big changing in the trend occurred in 2020 maybe may have affected on the global performances;
- KNN error increases, increasing the window dimension in a linear way (for each value of window, the mean KNN error is taken) (Figure 3) :

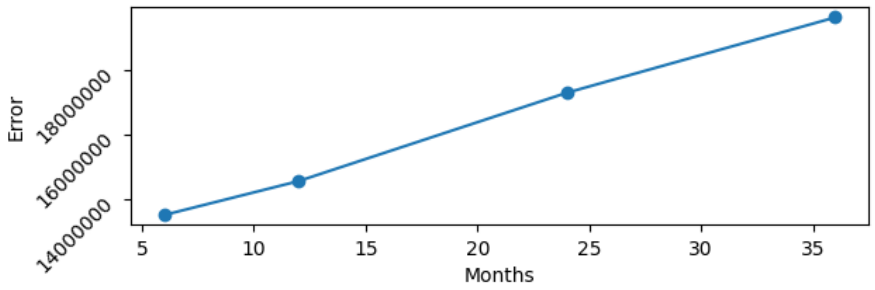

Figure 3. Linear relation between window choice and KNN error.

- Random Forest error increases, increasing the window dimension (Figure 4):

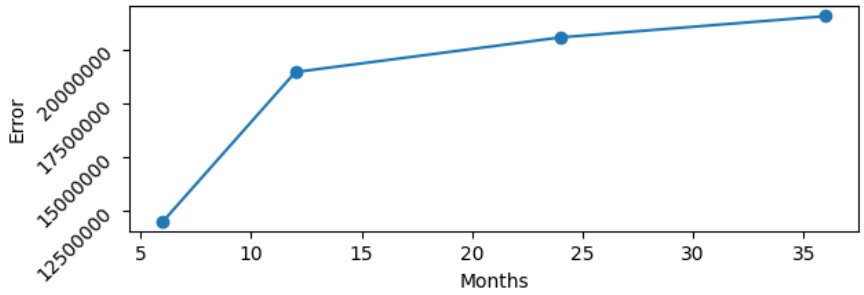

Figure 4. Relation between window choice and Random Forest error.

- There is not a clear correlation between Linear Regression error and window dimension (Figure 5):

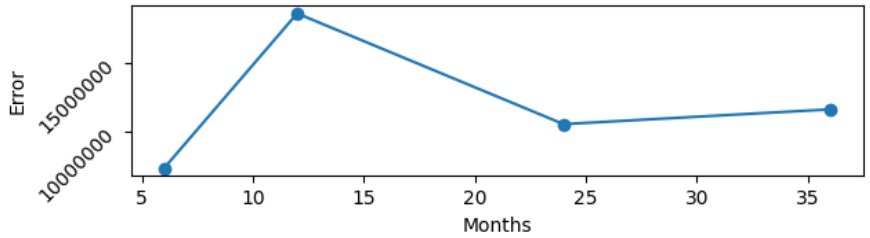

Figure 5. Relation between window choice and Linear Regression error.

Reference

[1] https://www.kaggle.com/datasets/yyxian/u-s-airline-traffic-data