**Machine Learning Model**

**Overview**

The purpose of the machine learning model is to predict the fare for domestic flights in the USA territory given only two inputs, origin airport code and destination airport code. To accomplish this purpose the dataset had to be fitted, trained and tested.

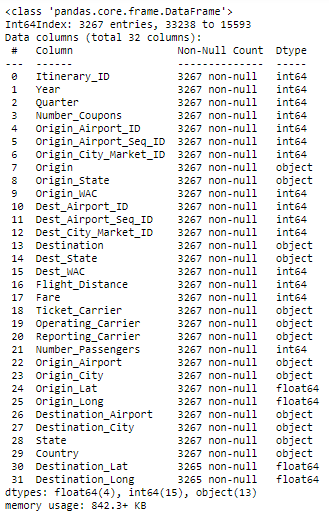
**Data**

Figure 1 Data

The dataset used for the model is from the [*Bureau of Transportation Statistics*](https://www.bts.gov/) which according to their website “is one of more than a dozen principal federal statistical agencies”. The sample used for the model is a 5% of all the available data from the first quarter of 2020 After cleaning the data, this ended up in 32 columns and 3,267 rows.

**Model**

To build this model, first we needed a correlation between our target, the Fare column, and the rest of the columns. With that result we decided to use the columns that had more than 5.6% correlation as futures in our model, the other columns were ignored because they were not truly relevant.

Then we fit the data and divide it into 80% for training purposes and 20% for testing purposes, these parameters will apply to all the tested models. Also, the data of all the tested model was fitted and unfitted to see the results in both ways.

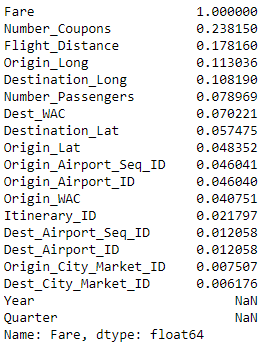
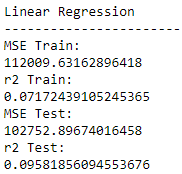
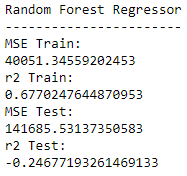


Figure 2 Model

First, the data was tested with linear models which included a Linear Regression, Logistic Regression model, K Neighbors Classifier, Support Vector Machine (SVM) and also Tree Models like: Random Forest Regressor, Decision Tree Classifier, Bagging Classifier, Random Forest Classifier, Ada Boost Classifier, Gradient Boosting Classifier and XTREME Gradient Boosting Classifier. That is a total of 22 models.

Out of these 22 models the two with best results were the Linear Regression and Random Forest Regressor. Here is a comparison for both models:

Although the Random Forest Regressor had a negative r2 score on the test, it had a better result on the Train and that is why it was tested on production with better results.

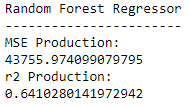
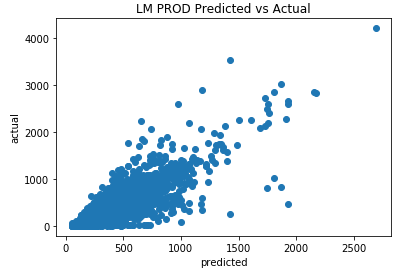


Figure 3 Selected Model



Because no one knows exactly how many miles their flight is going to take given that it is one of our main features and that the app is only asking for the origin airport code and destination airport code, it was necessary to pull the latitude and longitude from each one of the airports and then calculate the distance between both coordinates to be able to get the distance in miles and include it into our futures.

**Limitations**

This dataset was not divided by dates, it was only divided by quarters. This was a huge limitation because the date is especially important when it comes to determine the price of the tickets. When compering the results with actual tickets prices from different airlines it is important to select a roundtrip flight on the same day, the model cannot predict flight fares from future days.

There were also some outliers in the dataset that affected the model, the range of all the fares was from $4.00 dollars up to $17,000 dollars. This data probably included different classes or promotions, and this was not specified.

**Future Work**

Overall, it was a particularly good dataset, but it can be improving with the things mentioned on the limitations, like adding tickets types and dates. It took a lot of work and time pulling the dataset together and we only had two weeks to work on this project.

Also, other machine learning models could be included on the web page like Neural Network to categorize flights or Computer Vision to be able to analyze images.