Delayed Flights: Study of Possible Factors Caused Flight Delays at Destination Airport

Janith Perera

Introduction

Purpose of the Project

Nowadays, the aviation industry plays a crucial role in the world’s transportation sector, and lot of businesses rely on various airlines to connect them with other parts of the world. But several factors may directly affect the airline services by means of flight delays. Those flight delays are inevitable and it plays an important role in both profits and loss of the air lines. An accurate estimation of flight delays allows passengers to be well prepared for the deterrent caused to their journey and enables airlines to respond to the potential causes of the flight delays in advance to diminish the negative impact and to increase customer satisfaction and incomes of the airline agencies.

The purpose of this project is to build a model for predicting flight delays that occurs in month of January.

Significance of the Project

Quote

Flight cancellations and delays have been surging, especially during heavy travel periods such as weekends and holidays. Since the start of June, nearly 26,000 flights, or 2.2% of all flights by U.S. carriers, have been canceled and 260,000, or 22%, have been delayed, according to FlightAware, a flight tracking company. (Los Angeles Times, 2022)

Unquote

As the air travels have a significant role in economy of agencies and airports, it is necessary for them to increase quality of their services. One of the important modern life challenges of airports and airline agencies is flight delay. In addition, delay in flight makes passengers concerned and this causes extra expenses for the agency and also the airports.

Research Question

Using exploratory analysis and popular machine learning classification models, can we predict the possible flight delays at the destination airport specifically for the month of January in upcoming years, based on factors like day of the week, date of the month, carrier, origin, destination and distance of the flight etc.

Description of Dataset

Dataset downloaded from www.kaggle.com which contains one binary target variable and 21 feature columns such as origin airport, destination airport, carrier, flight information and etc. with nearly 1.2 million instances.

This data is originally collected from the Bureau of Transportation Statistics, Government of United States of America. Dataset contains all the flights throughout the United States in the month of January 2019 and January 2020.

Data Preprocessing

Data Preparation

Initially two features named cancelled and diverted, integrated with arrival delayed feature since the total number of instances of cancelled and diverted were 2.24% from the total dataset, further same instances had the null values in target variable.

Next, all other instances with missing values were dropped because, null values to total dataset are less than 3% and dataset is quite large. Therefore, dropping few datapoints will not make significant difference in outcome.

After, hourly time block feature has created from both departure and arrival time features (continuous), because existing departure blocks are inaccurately categorized and there was no categorical feature for arrival time. Then, departure time and arrival time features has been dropped to reduce correlation between features.

Finally, dropped some other subset/same features which creates an inter correlation in between features. All categorical features one-hot-encoded using sklearn's LabelEncoder() method and target variable encoded with sklearn's LableBinarizer() method.

Exploratory Data Analysis

When scrubbing data, it is important to note some special characteristics of the dataset. It seems to be forth day of the week and twenty forth day of the month have the greatest number of delayed flights at destination airport. Further, ORD (Chicago O'Hare International Airport), DFW (Dallas/Fort Worth International Airport), ATL (Hartsfield-Jackson Atlanta International

Airport) are the top three airports having the greatest number of delayed departing flights and arriving flights. Further, it can clearly see WN (Southwest Airline), AA (American Airline), OO (Skywest Airline) are top three carriers who has greatest number of delayed flights at destination airports. It is unable to see significant relation in between delayed flights and distance flew.

Visualization

The heatmap shows there is correlation in between some of the features in the dataset which can badly affects the outcome of machine learning models. To avoid that negative effect, some features has been changed or dropped when doing the preprocessing step.

Data Splitting

Since the whole dataset is too large, it took unreasonable time to train a model on local computer. Therefore, I used pandas sample() method to downsize/ get random sample of 10% of total dataset.

Then, for modeling purpose downsized dataset split in to 70:30 ratio of train and test sets by using sklearn's train_test_split() function. Since this dataset is unbalanced and it can create negative influence when training the model, I have used the imblearn's RandomOverSampler() method to create balanced training dataset.

Model Building and Evaluation

Model Building

In this project, primarily used the classification models to train the dataset because the target variable is categorical. Initially seven of well-known classification models in sklearn package used. Those are Logistic Regression, Stochastic Gradient Descent (SGD) Classifier, Decision Tree Classifier, Random Forest Classifier, Support Vector Machine Classifier (SVC), Gradient Boosting Classifier and XG Boost Classifier. Models compared with overall accuracy, precision score, recall score, and F1 score. Some of the models seems to be overfit with the training dataset.

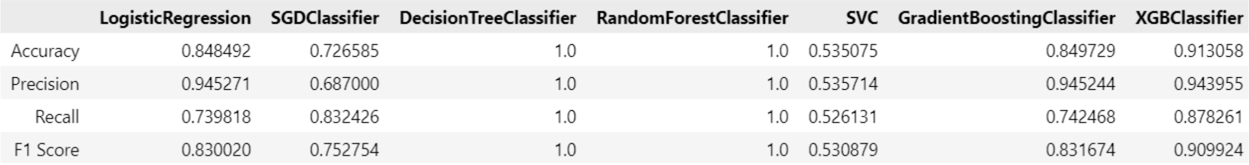
Therefore, same models trained with 5-fold cross validation and compared with the same test scores as before.

Model Selection and Model Optimization

The best fit model, with best test results which is Random Forest Classifier has selected out of seven models and retrained with grid search cross validation to obtained hyperparameter tuned optimal model.

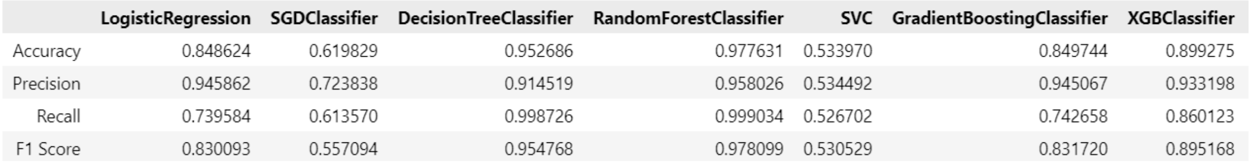
Model Comparison

At the initial step decision tree and random forest classifiers overfitted and SVC, SGD classifier, logistic regression, gradient boosting had poor performance than XG boost classifier. Below table shows the accuracy scores obtained at this step which trained with training dataset with default parameters on models.



	LogisticRegression	SGDClassifier	DecisionTreeClassifier	RandomForestClassifier	SVC	GradientBoostingClassifier	XGBClassifier
Accuracy	0.848492	0.726585	1.0	1.0	0.535075	0.849729	0.913058
Precision	0.945271	0.687000	1.0	1.0	0.535714	0.945244	0.943955
Recall	0.739818	0.832426	1.0	1.0	0.526131	0.742468	0.878261
F1 Score	0.830020	0.752754	1.0	1.0	0.530879	0.831674	0.909924

5-fold cross validation training with default parameters seems to be done better job on training the model.



	LogisticRegression	SGDClassifier	DecisionTreeClassifier	RandomForestClassifier	SVC	GradientBoostingClassifier	XGBClassifier
Accuracy	0.848624	0.619829	0.952686	0.977631	0.533970	0.849744	0.899275
Precision	0.945862	0.723838	0.914519	0.958026	0.534492	0.945067	0.933198
Recall	0.739584	0.613570	0.998726	0.999034	0.526702	0.742658	0.860123
F1 Score	0.830093	0.557094	0.954768	0.978099	0.530529	0.831720	0.895168

Among those models, random forest classifier has the best performance and finally hyperparameter tuned random forest model gave 1.00 and 0.92 overall accuracy on training set and testing set respectively.

Conclusion

Conclusion

In this project, it has been used flight data throughout United States of America in month of January 2019 and 2020 to predict flight arriving delays at destination airport. Project result shows that the Random Forest Classifier yields the best performance compared to all other models. Therefore, random forest classifier can use to assist in predicting flight delays at destination airport for the month of January in upcoming years based on selected factors.

Further, as per sklearn’s feature_importance() function, delays at departing airport is the most important feature of predicting delays at destination airport.

Recommendations

For further studies of delayed flights, I would recommend to have year-round flight delay data instead of one month, because we can study seasonal variations on flight delays.

Further, weather condition is well-known influencing factor, for flight delays. Therefore, I would recommend to populate this dataset with weather condition data features of the area such as visibility, snow, precipitation, wind speed and etc.

References

[1] Los Angeles Times. (2022, July 22). *Retrieved from.* https://www.latimes.com/business/story/2022-07-22/flight-delays-cancelations-summer-2022

[2] Kaggle. (n.d.) *Retrieved from.* https://www.kaggle.com/datasets/divyansh22/flight-delay-prediction

Appendix

Figure 1: *Flight delays by year*

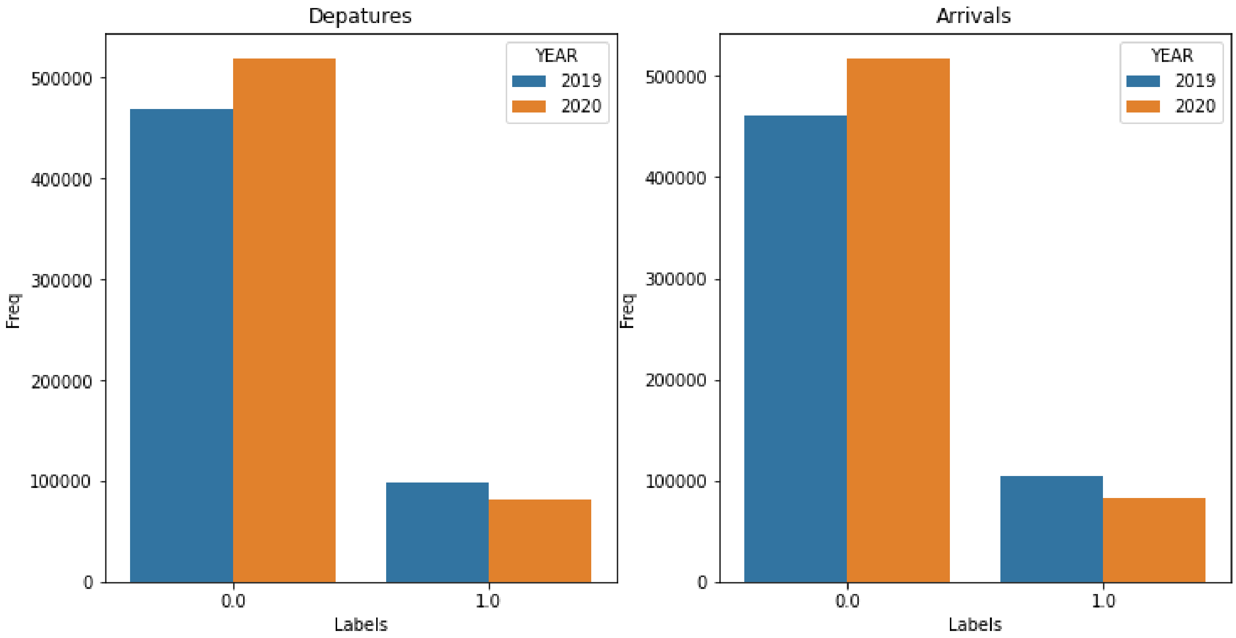


Figure 2: *Delayed flights by day of week*

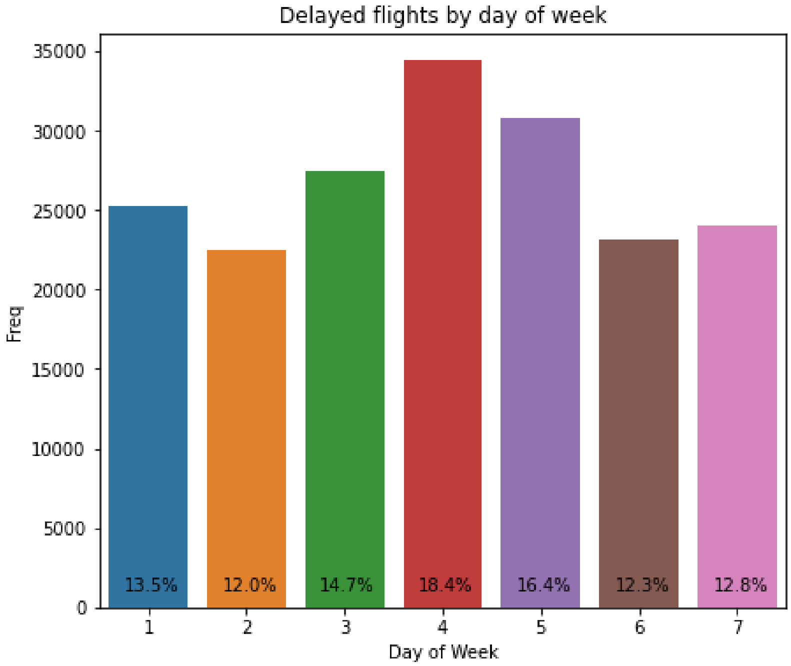


Figure 3: *Delayed flights by day of month*

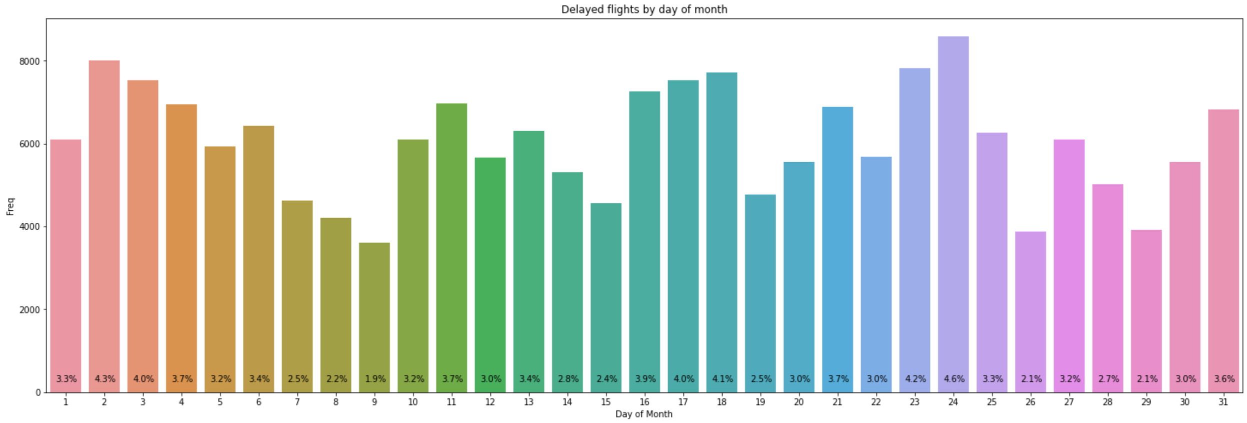
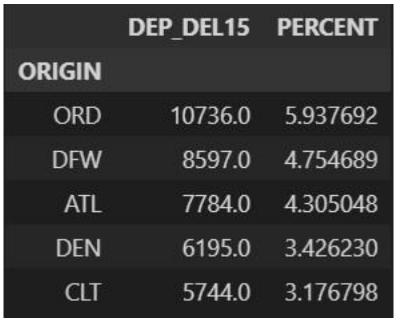
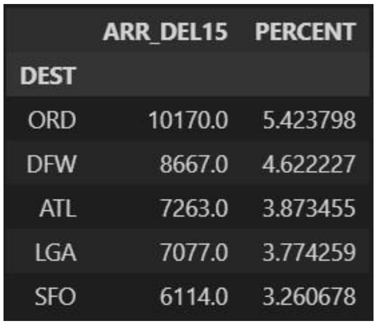


Figure 4: *Top five airports has delayed departures*



	DEP_DEL15	PERCENT
ORIGIN		
ORD	10736.0	5.937692
DFW	8597.0	4.754689
ATL	7784.0	4.305048
DEN	6195.0	3.426230
CLT	5744.0	3.176798

Figure 5: *Top five airports has delayed arrivals*



	ARR_DEL15	PERCENT
DEST		
ORD	10170.0	5.423798
DFW	8667.0	4.622227
ATL	7263.0	3.873455
LGA	7077.0	3.774259
SFO	6114.0	3.260678

Figure 6: *Delayed flights by carrier*

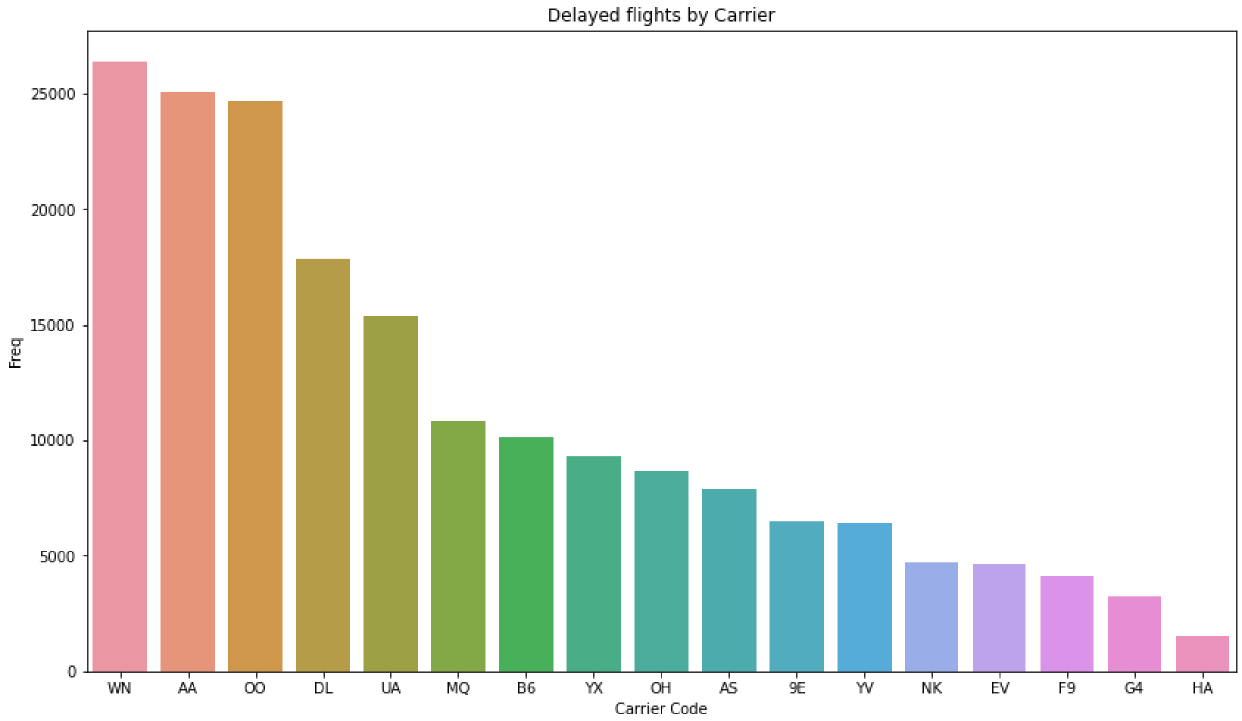


Figure 7: *Delayed flights by distance flew*

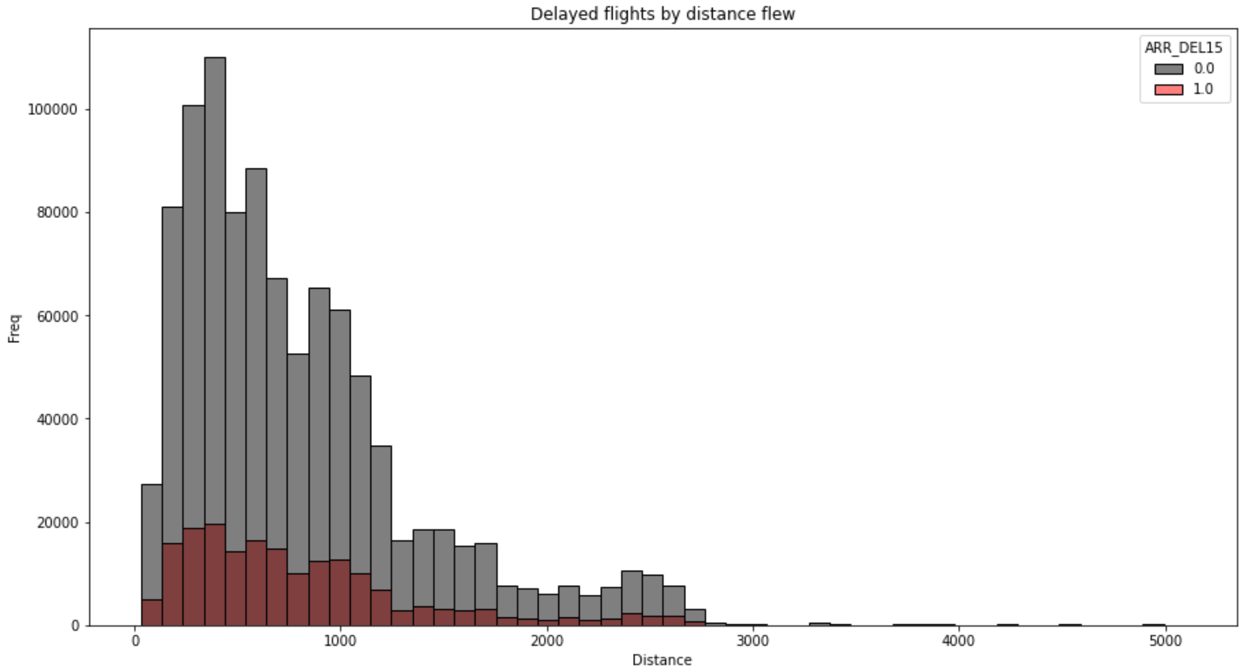


Figure 8: *Correlation heatmap before data preparation*

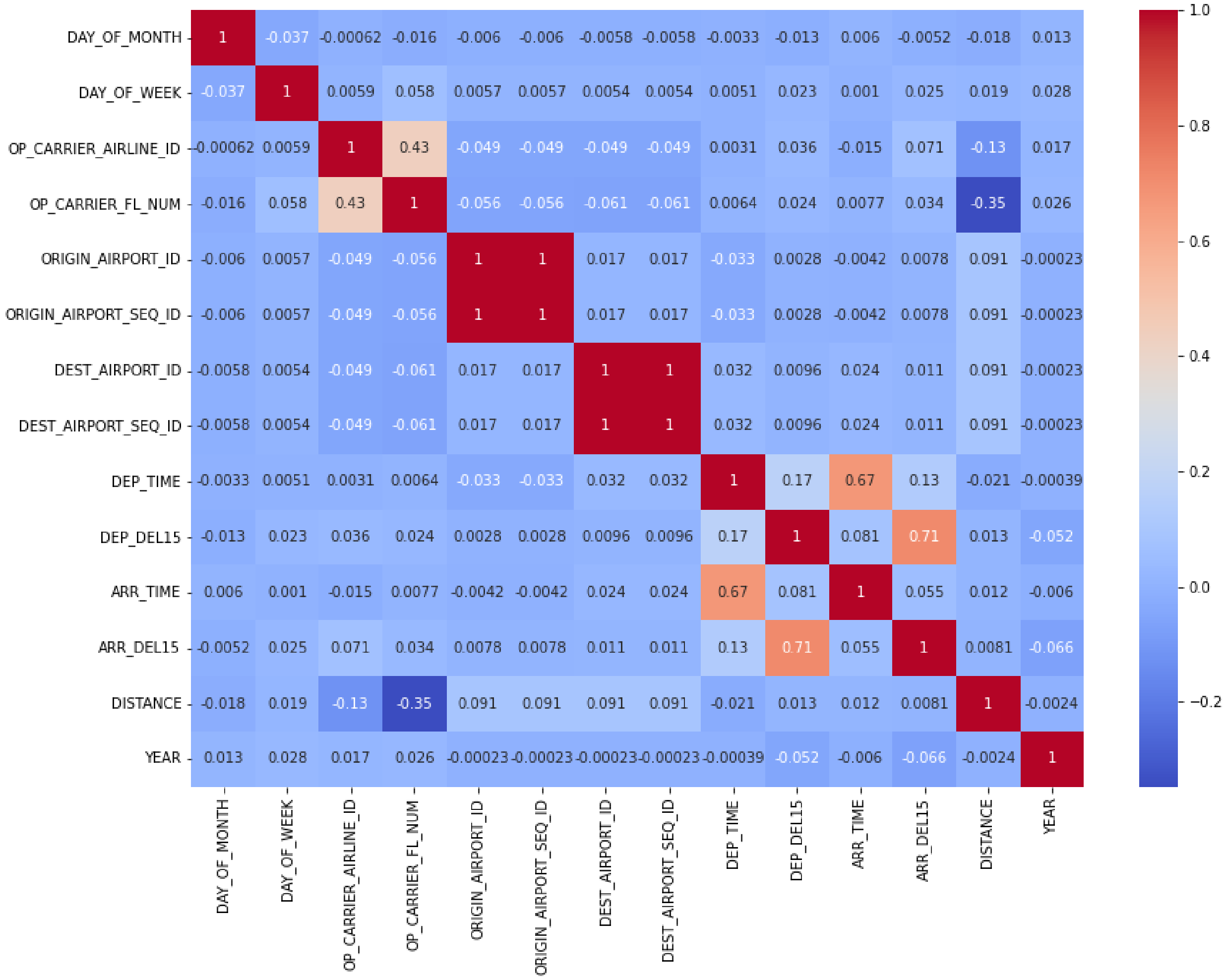


Figure 9: *Unbalanced and balanced training sets before model training*

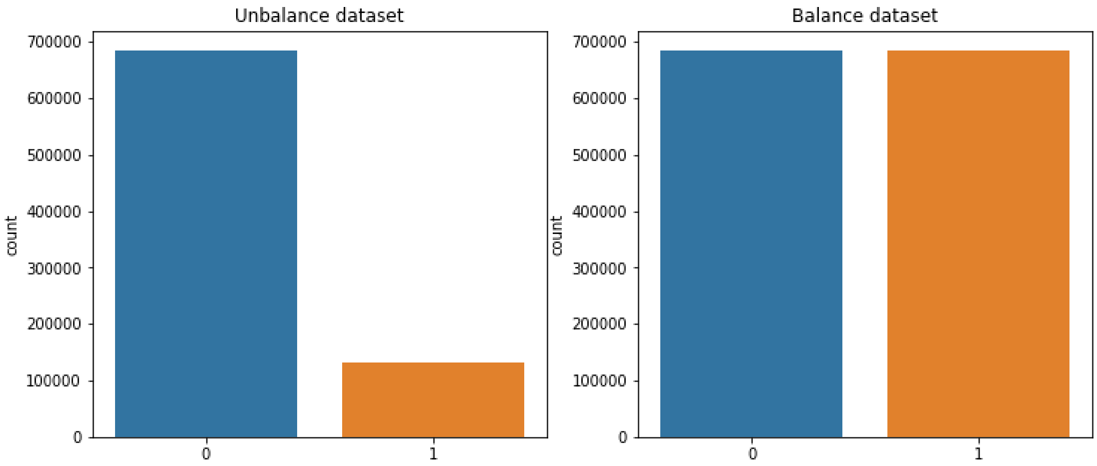
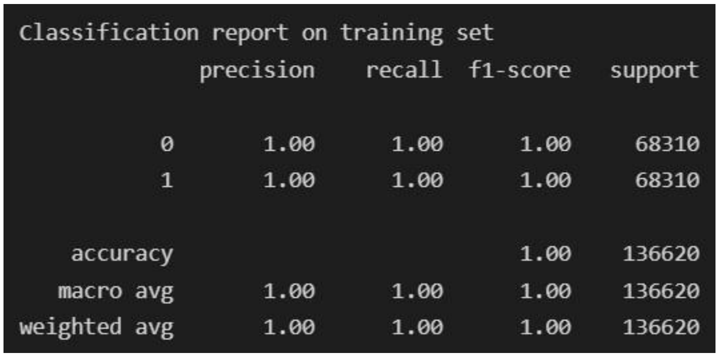


Figure 10: *Classification report on training set with hyperparameter tuned model*

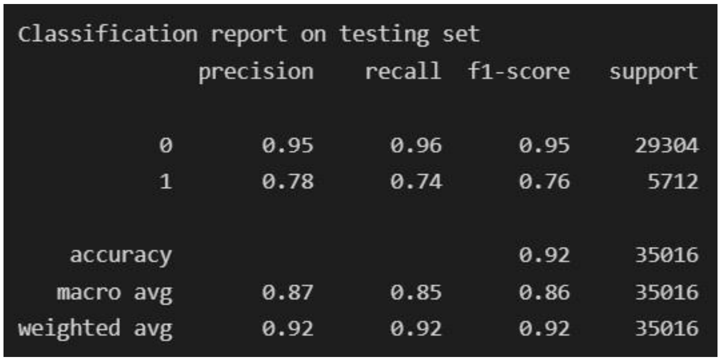


```
Classification report on training set
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     68310
           1       1.00      1.00      1.00     68310

    accuracy                           1.00    136620
   macro avg       1.00      1.00      1.00    136620
weighted avg       1.00      1.00      1.00    136620
```

Figure 11: *Classification report on testing set with hyperparameter tuned model*



```
Classification report on testing set
              precision    recall  f1-score   support

           0       0.95      0.96      0.95     29304
           1       0.78      0.74      0.76      5712

    accuracy                           0.92     35016
   macro avg       0.87      0.85      0.86     35016
weighted avg       0.92      0.92      0.92     35016
```

Figure 12: *Heatmap on confusion matrix*

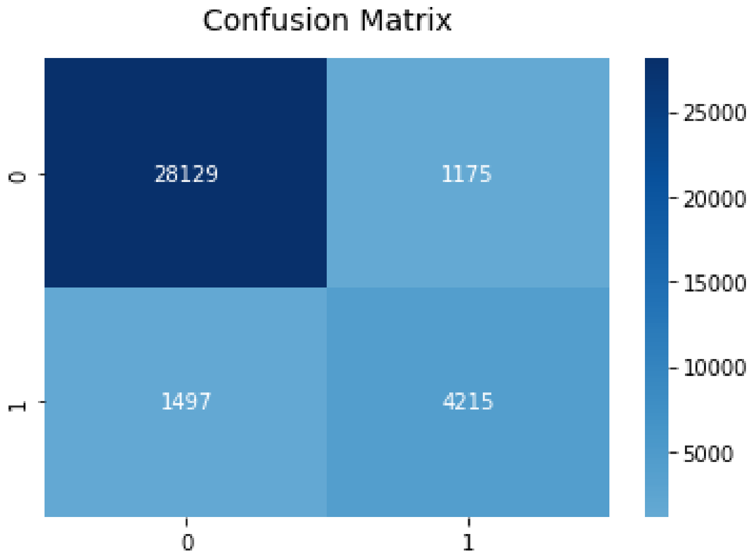
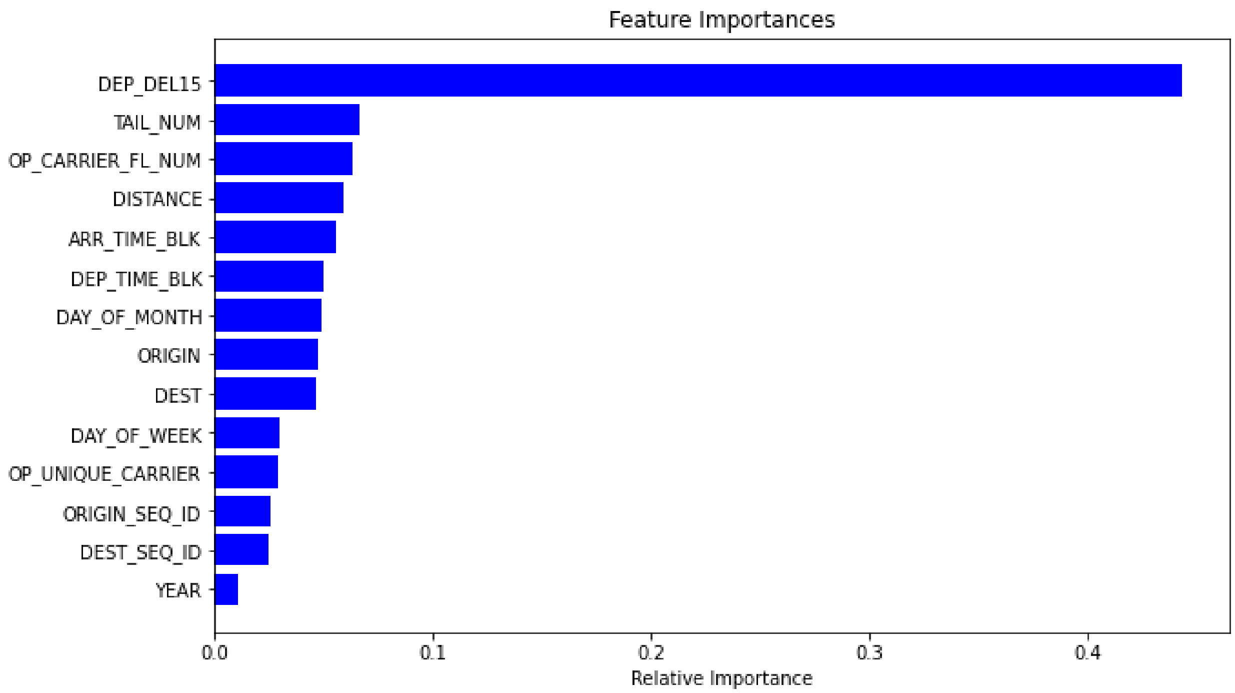


Figure 13: *Feature importance*


Feature Importances
DEP_DEL15
TAIL_NUM
OP_CARRIER_FL_NUM
DISTANCE
ARR_TIME_BLK
DEP_TIME_BLK
DAY_OF_MONTH
ORIGIN
DEST
DAY_OF_WEEK
OP_UNIQUE_CARRIER
ORIGIN_SEQ_ID
DEST_SEQ_ID
YEAR
Relative Importance