ST 4035 – DATA SCIENCE

Group Project Report

Group Number - 02

S. Raweesha Nethmini s14499
S.Wickramathilake s14518
U.A.B.M. Gunasekara s14330
M.R.N. Fernando s14429

1. Abstract

There are over 5,000 airlines around the world and The Southwest Airlines, is the largest low-cost airline in the world and one of the main airlines in the United States. This study has focused on developing a model that could accurately predict whether or not a specific Southwest Airlines aircraft would be delayed for 15 minutes and also discover the variables that significantly influence flight delays and the unique traits associated with them based on Southwest airline data. This research was conducted on the '2019 Airline Delays w/Weather and Airport Detail' dataset from Kaggle (www.kaggle.com, n.d.). Since the dataset contained a large amount of data, the study was conducted using only one airline category for the convenience of the study. For the purpose of deducting the size of the dataset, the records related to Southwest Airlines were drawn out since it is the most popular and customer-satisfied airline in the United States. Throughout this study, visualizations and model fitting were done using Python. There were no null values in the initial dataset, but some duplicates were there, and all of them were removed. The variables 'MONTH', 'DAY_OF_WEEK', 'DISTANCE_GROUP', and 'SEGMENT_NUMBER', which were initially integer types, were transformed to object types since they are categorical variables. And also, some of the unnecessary columns were removed from the dataset. After applying data preprocessing techniques, the exploratory data analysis was carried out for the prepared dataset. Prior to the further analysis, the dummy variables were created for categorical variables. The dataset was split into train and test data, and the train set had been balanced. Further, the train set was split again into two parts: train and valid for the purpose of hyper parameter optimization. The best random forest model was chosen through several steps of removing the least significant variables. Then the parameters of the chosen random forest model were optimized using 'GridSearchCV' and updated the model with optimized parameter values. Finally, XGBoost was applied to the final random forest model and further parameter tuning was done. The optimized 'XGBoost' model was selected as the best model which resulted in 0.794 accuracy.

2. Introduction

The demand for air transportation services has grown since it provides only a rapid global transportation network, making it significant for global business. People use airlines to travel for personal, business, and official purposes to a great extent. However, the aviation industry is now dealing with a number of issues, one of which is flight delays. Passengers experience several inconveniences due to flight delays, which often incur additional costs. Therefore, it is important to have some approaches for predicting flight delays, which assist in gaining a general understanding of whether a specific flight will be delayed and taking the necessary action to overcome such circumstances.

The Southwest Airlines Co., also known as Southwest, is the largest low-cost airline in the world and one of the main airlines in the United States. This study has mainly focused on developing a model that could accurately predict whether or not a specific Southwest Airlines aircraft would be delayed for 15 minutes based on the 2019 Airline Delays w/Weather and Airport Detail dataset that is accessible on Kaggle. When fitting the classification model, the information pertaining to aircraft, weather, airports, and employment was taken into account.

The results of this study would help the aviation industry to identify important risk factors related to flight delays and predict whether a flight would be delayed in advance. Hence, the aviation sector can pay more attention to those factors and take suitable actions to reduce flight delays, which helps to maintain the customer confidence and reputation of the organization.

3. Objectives

- The primary objective of this study is to identify a suitable model to forecast whether or not a specific Southwest Airlines aircraft would be delayed for 15 minutes based on factors relating to aircraft, weather, airports, and employment.
- Along with constructing the model, the study also aims to discover the variables that significantly influence flight delays and the unique traits associated with them, allowing the aviation industry to take the necessary steps to reduce delays.

4. Methodology

4.1 Data

The dataset which will be used in our study is a classification dataset with detailed airline, weather, airport, and employment information in the USA. The dataset contains 26 categorical and quantitative variables and nearly 01 million pieces of data. Here the TARGET variable is the binary of a departure delay over 15 minutes.

Since the dataset contains a large amount of data, we will be conducting the study using only one airline category for the convenience of the study. For the purpose of deducting the size of the dataset, the records related to Southwest Airlines will be drawn out since it is the most popular and customer-satisfied airline in the United States.

4.2 Data Preprocessing

- All libraries and modules that are required for analysis were imported.
- Using the pandas library dataset was imported to the python notebook.
- Checked, Are there null/ missing values or duplicates? There were no null values, but some duplicates were there. All duplicates were removed. Because using duplicates we cannot get any additional information.
- Columns that have the same values for all records were dropped from the data frame.
- Next using the whole dataset Exploratory Data Analysis was done.
- All the outliers that appeared in the dataset were not removed because that outliers can be practically happened according to real-world scenarios.
- Data types of the categorical variables were changed into object data types except for the dependent variable.
- Using the get dummies function in the pandas library create dummy variables for categorical variables.
- Dataset was split into two parts train and test set. 80% of records were allocated for the train set.
- According to the results of the EDA dependent variable had not balanced. Using SMOTE technique train set was balanced. The train set was split into two parts train and valid (validation) set. (To use in hyperparameter optimization step)

4.3 Method

- Random forest model was fitted using all the variables (full model). And accuracy scores of the train, valid, and test sets were obtained.
- At the 4th step 'PREVIOUS_AIRPORT' and 'DEPARTING_AIRPORT' columns were dropped, and the same steps were done until the 11th step (2nd model).
- Values obtained at the 11th and 12th steps were compared.
- Variable importance plot was obtained according to the 2nd model. And some least important variables dropped (relative important less than 20%) and fitted the new 3rd model (less complex model than 1st and 2nd) was built.
- Accuracy scores of the above three models didn't deviate much. So, as the final random forest model 3rd model which has less complexity been selected.
- After that parameters of the random forest model were optimized using 'GridSearchCV'. (Tried to find the best value for each parameter of the 3rd model)
- Model was updated using optimized parameter values.
- Next, the valid test accuracy score was obtained. Accuracy scores of the train and valid sets were compared to check a significant difference. There was not any significant difference and concluded that from optimization no overfitting happened. (Before doing the parameter tunning there was a significant accuracy difference between the train and test set. It might be due to overfitting)
- Sensitivity of the test set in final random forest mode was low 'XGBoost' algorithm was also applied for the features that were used in the best Random Forest model.
- According to the classification report of the 'XGBoost' model there was little improvement in sensitivity.
- Next parameter tuning was done for the 'XGBoost' model also. After that sensitivity improved further.
- Optimized 'XGBoost' model was selected as the best model after considering the classification report. Finally, AUC values were calculated for both train and test sets of the best model using the ROC curve.

5. Analysis

5.1 Explanatory Data Analysis

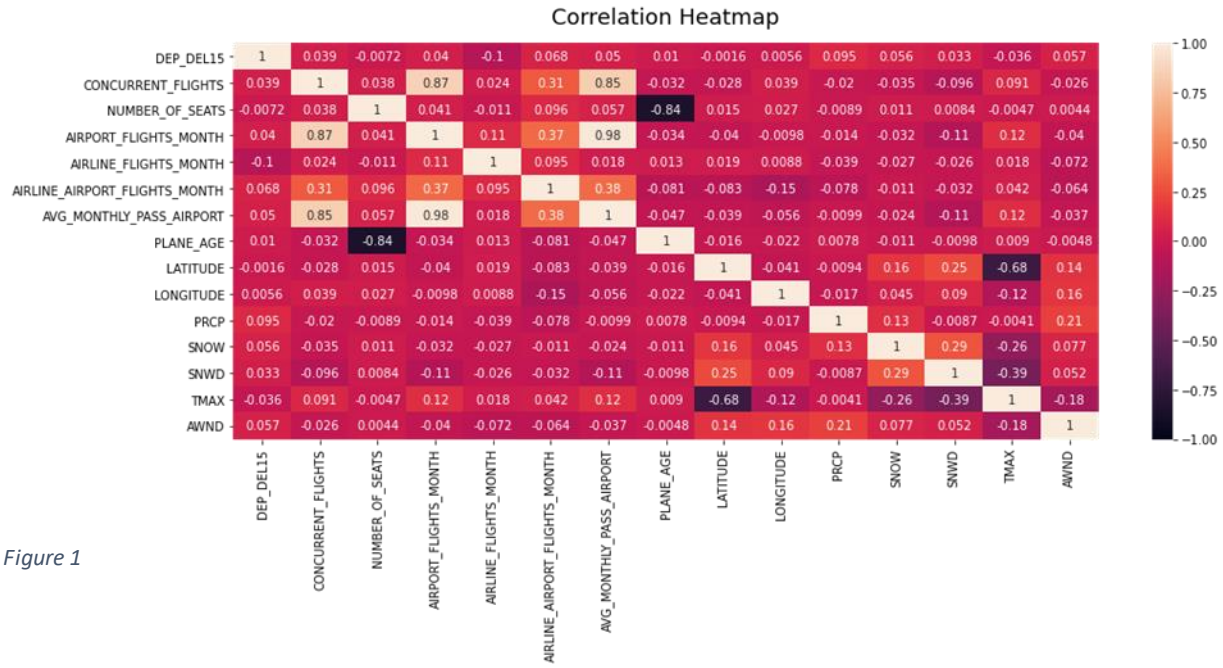


Figure 1

Figure 1 shows the correlation between the numerical variables in the dataset. According to that there is no strong relationship between departure delay and other variables. But this shows high multicollinearity among variables. There is a moderate negative correlation between max temperature and latitude And between plane age and the number of seats. According to the matrix airport flights month, concurrent flights and average passengers for airline for month appear to be highly positively correlated with each other. There is a strong correlation between average passengers for airline and airport flights per month.

According to the figure 3 scatter plot when the Average Passengers for the departing airport for the month increasing, the Airport flights per month also increasing. Some outliers can be observed from the plot.

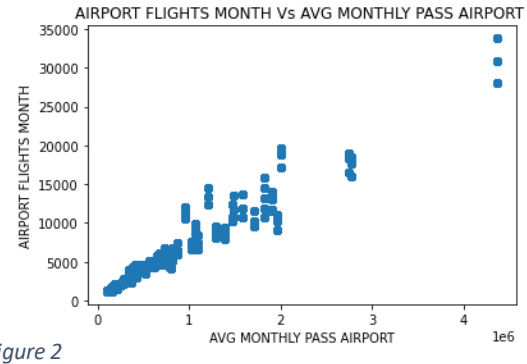
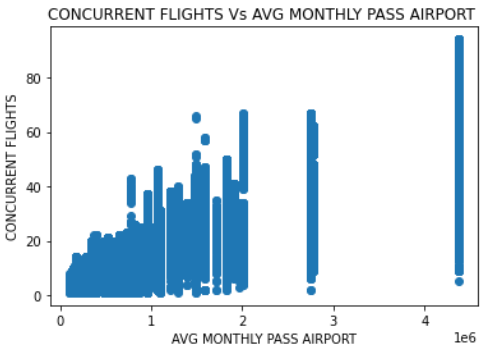


Figure 2



The scatter plot of Concurrent flights Vs Average Passengers for the departing airport for the month leaving from the airport shows an increasing trend.

Figure 3

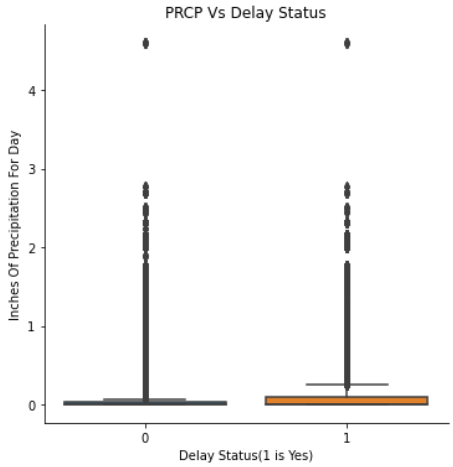


Figure 4

Figure 5 shows the PRCP vs delay status. PRCP stands for the inches of precipitation for a day which means the amount of water that is falling out of the sky as rain or drizzle. The PRCP is high in delayed flights. Outliers can be seen in both cases.

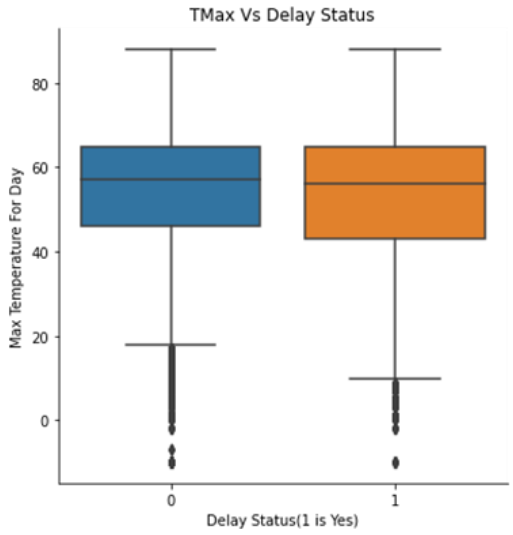


Figure 5

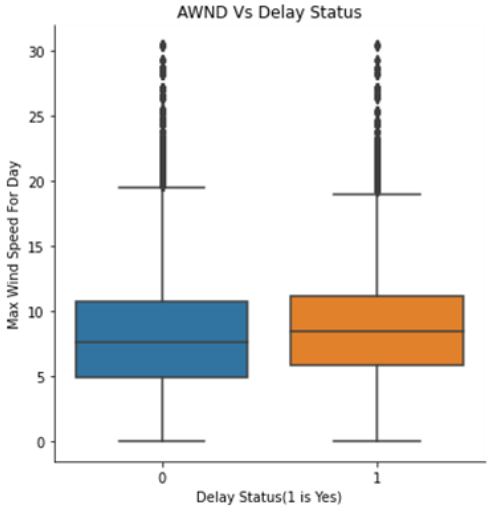


Figure 6

AWND stands for max wind speed for the day and TMAX stands for max temperature for the day. The mean of AWND for delayed flights is higher than the not delayed flights. The mean of TMAX is high in not delayed flights than the delayed flights.

5.2 Analysis of the Model

In this analysis, since the target variable in the dataset was a binary variable, we used classification techniques to fit a model. Accordingly, here first used the random forest for the classification model. In a random forest, first, get the bootstrapping samples from the original dataset, fit the classification tree for each bootstrap sample, and finally get the generalized prediction. But in a random forest, it doesn't consider all the variables in each split. In there get some random samples from an original set of variables and define the split. Then from the selected sample of variables, it tries to find out the best variable in each split.

First consider the random forest for the full model, which includes all the variables. After fitting the random forest model for the full train set, it was found that the model was overfitting. Model accuracy for the full model is 0.9977.

Since the model was overfitted with the train set, then reduced the model by only considering the most important variables. For that, we used feature importance using a random forest classifier. According to that,

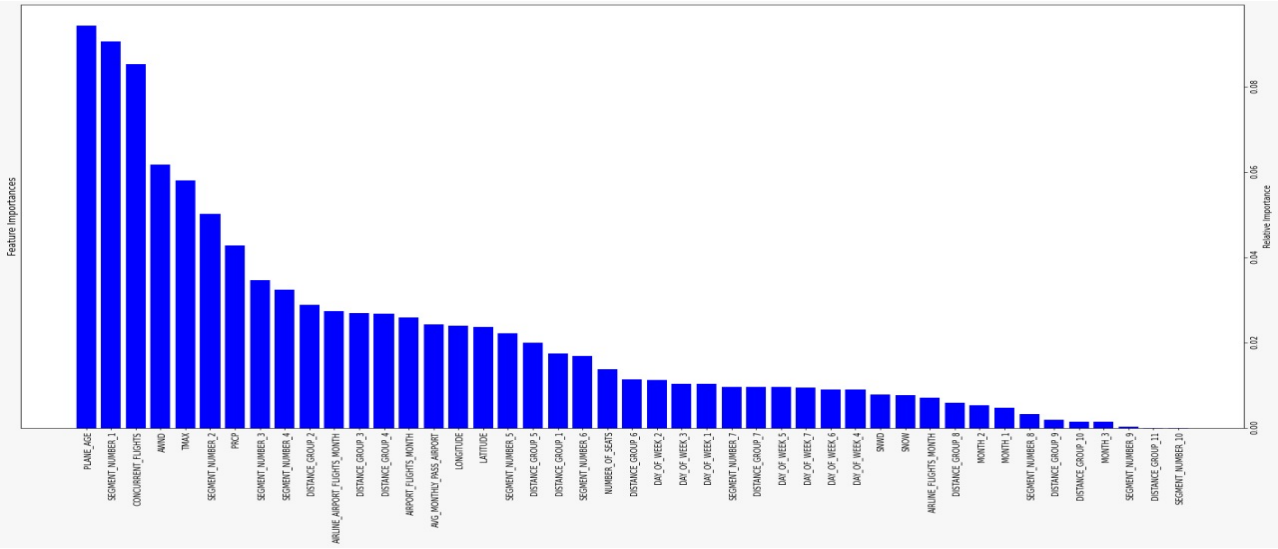


Figure 7: Feature Importance Plot

By only considering the most significant variables, fit the random forest for the reduced training set. Then noticed that the reduced model was also overfitted with the train set with the 0.9976 accuracy score.

Hence hyperparameter optimization has used the prune the random forest model. Pruning can be used for reducing the complexity of the trees. In there, it can be defined, as maximum depth, maximum features of considering for split, and the maximum number of terminal nodes likewise. There are several techniques that can be used to do the pruning. For classification, it can be used Gini impurity and information gain (entropy).

After running the hyperparameter tuning, it can be found that Gini was the best criterion, max_features were 25, max_depth was 30, n_estimators were 400, min_samples_leaf and min_samples_split was 25. Using these values, prune the reduced random forest model. Then found that the model was not overfitted with the pruning model. The accuracy score for the train test is 0.8564. Then the accuracy score for the test set is 0.7862.

Since the sensitivity value is less, which is 0.5, we used extremely randomized trees (extra trees) to fit the model.

Extremely randomized trees are another ensemble technique like a random forest. But the difference is in extra trees it is not used bootstrap samples. In there, consider the original train set and fit the model. Here take the randomly selected subset of features using and doing the splits.

First fit the XGBoost for the reduced model without considering the pruning. For a better fit, then fit the XGBoost model after pruning. After applying the hyperparameter optimization, it was observed that n_estimators were 200, and max_depth was 6. Then using those value prune the XGBoost and fit the model. The fitted final model accuracy was 0.7941.

After that obtained the confusion matrix. Using that we can understand the background behind the model's accuracy score.

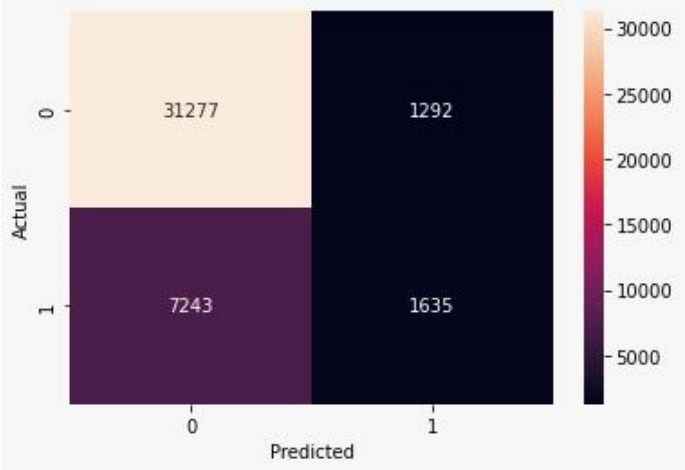


Figure 8: confusion Matrix

	precision	recall	f1-score	support
0	0.81	0.96	0.88	32569
1	0.56	0.18	0.28	8878
accuracy			0.79	41447
macro avg	0.69	0.57	0.58	41447
weighted avg	0.76	0.79	0.75	41447

Figure 9: Classification report

According to the confusion matrix,

- In fact, only 31,277 fights are predicted to have not been delayed out of the total number of fights that hadn't delay. The remaining 1292 are predicted as delayed.
- Of those who actually delayed, only 1635 are predicted as delayed. The remaining 7243 are predicted to as didn't delay. This is not a good situation.

Since the overall accuracy of this model is good, so looked at how the classes are predicted. Here considers the classification report.

According to the classification report, it is observed that the sensitivity of the model is 0.56 and the specificity is 0.81. Flights that were not delayed predicted correctly than the flights which were delayed.

Then consider the how ROC curve is distributed;

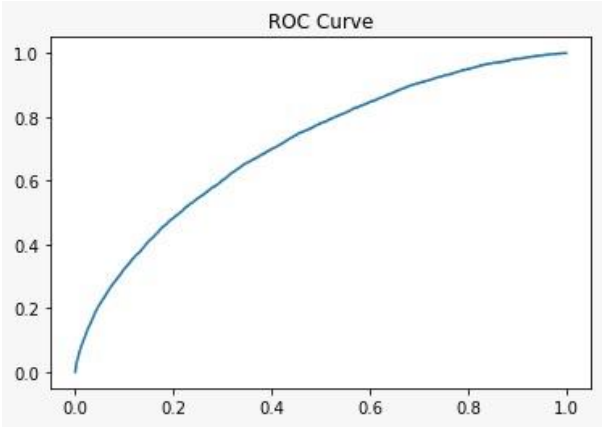


Figure 10: ROC Curve

According to the ROC curve, it was observed that the area under the curve (AUC value) is 0.7121. This final model fit is an averagely good fit.

Using the below table, it can be identified how model values were distributed with the several models.

Fitted Model	Accuracy Score of Train set	Accuracy Score of Test set	Specificity of test set	Sensitivity of test set
Random forest Full Model	0.9977	0.7688	0.83	0.44
Random forest Reduced Model	0.9976	0.7672	0.82	0.43
Random forest Reduced Optimize Model	0.8564	0.7862	0.81	0.50
XGBoost Reduced Model	0.8597	0.7915	0.81	0.54
XGBoost Reduced Optimize Model	0.8623	0.7940	0.81	0.56

6. Conclusions

- This study has focused on developing a model that could accurately predict whether or not a specific Southwest Airlines aircraft would be delayed for 15 minutes and also discover the variables that significantly influence flight delays and the unique traits associated with them based on Southwest airline data.
- As the initial step, data preprocessing techniques were applied to the data set. There were no null values in the initial dataset, however some duplicates were there, and all of them were removed. The variables 'MONTH', 'DAY_OF_WEEK', 'DISTANCE_GROUP', and 'SEGMENT_NUMBER', which were initially integer types, were transformed to object types. And also, some of the unnecessary columns were removed from the dataset.
- The exploratory data analysis was carried out to the prepared data set and there it was found that airport flights month, concurrent flights and average passengers for airline for month are highly positively correlated with each other and also there is a strong correlation between average passengers for airline and airport flights per month. In addition, it was observed that the PRCP is high in delayed flights. The mean of AWND for delayed flights is higher while the mean of TMAX is lower in delayed flights.
    - Finally, several classification models were fitted in order to find the best model for predicting the flight delays. The train set had been balanced prior to fitting the model. The train set was split into two parts: train and valid for the purpose of hyper parameter optimization. The best random forest model was chosen through several steps of removing the least significant variables. Then the parameters of the chosen random forest model were optimized using 'GridSearchCV' and updated the model with optimized parameter values. Finally, XGBoost was applied to the final random forest model and further parameter tuning was done. Optimized 'XGBoost' model was selected as the best model which resulted in 0.794 accuracy.