UNIVERSITY OF EDINBURGH
Business School

2019-2020

Predictive Analytics and Modelling of Data - CMSE11428

Final Report / Group 3

Antonis Photiou, Dimitrios Petkidis, Yury Arutyunov, Vangelis Kappos, Izgi Arda Ozsubasi and Rachad Saab

21 November 2019

Table of Contents

List of Figures

List of Tables

1. INTRODUCTION

Airline scheduling is known to have many complexities due to the multiple factors that go into taking a flight from origin to destination. This includes passenger check-in, security checks, aircraft check-up, taxiing and many more variables that come into play. Considering these factors, it is easy to see that delays are a common problem and sometimes a flight cancellation can occur. In fact, delays and cancellations cost the industry vast amounts of money every year. A study on the US domestic flight delays show that an annual cost of *$31-40* billion in *2010* (Ball et al., 2010). Prominent given reasons for delays and cancellations include factors such as weather conditions, air traffic congestion, late reaching aircraft, security issues and maintenance (Chakrabarty, 2019). Hence, it is important to study and identify the most notable causes of delays and cancellations in order to improve and act upon them. In this paper we would like to address the issue of modelling flight delays and cancellations in order to predict their occurrence ahead of time-based on forecast data. We will focus on examining the direct causes of flight delays and cancellations, such as process delays. In addition to that, we will examine indirect causes, such as weather conditions. The motivation of this attempt is to assist in scheduling flights with more accuracy once the probability of a delay or cancellation is known. We aim to help airlines, specifically F9 and MQ, to optimize their schedules to improve customer satisfaction, reduce churn and in turn limit unexpected expenses resulting from altered schedules.

2. RESEARCH QUESTIONS

Within the scope of this analysis, we had to focus on two US based airlines, Frontier Airline (F9) and Envoy Air (MQ). We used our dataset in order to perform a thorough exploratory analysis, using departure delay and cancellation as our dependent variables. Our task was to find data-motivated research questions that would drive the rest of the analysis. Through our research questions we would like to find the potential solutions, that F9 and MQ should consider to improve their performance.

A. FLIGHT DELAYS

Our exploratory analysis showed that several weather attributes are positively correlated with departure delays [Appendix A1]. Evaluating whether those attributes are important and the magnitude of their influence in predicting delays will be useful for both airlines and airports. Furthermore, we would like to see whether the airlines F9 and MQ are more prone to weather caused delays than other airlines and whether this is due to airport-specific factors. Finally, in the studied literature we found conflicting results on how the competitiveness of the airport affects flight cancellations. We would like to use the idea of competitiveness and market share within an airport to find out whether it affects departure delays. Competitiveness is included in two levels in our study. These are the percentage of total flights an airline company executes in a specific airport per day and per year respectively. Thus, we would like to examine whether competition levels are affecting delays for the selected airports. Therefore, the selected research questions for predicting delays are:

1. Which weather factors are most influential in predicting delays?
2. Our airlines are more affected by weather than other airlines?
3. Does competition affect flight delays?

B. FLIGHT CANCELLATIONS

Through our exploratory analysis we have found that more cancellations happen during a certain period within a day [Appendix A2]. We would like to study whether moving the departure time, if possible, will reduce the probability of the flight being cancelled. However, in some cases this might not be applicable, since there are a lot of factors that determine the departure time of a flight, such as the parking cost at the airport, connected flights, etc. Moreover, we are interested to study how the various factors affect different types of cancellations, Carrier, Weather and NAS. The reason behind that is because we found that MQ has high proportion of NAS cancellations while F9 has many weather cancellations. Finally, we would like to examine which airports at which time periods are more related to flight cancellations. Therefore, our research questions for this part are:

1. Should the airlines consider changing the time of departure to avoid airport congestion?
2. How do the various factors affect different types of cancellations?
3. Which airports and time periods are more prone to flight cancellations?

3. TECHNIQUES

A. DATA PRE-PROCESSING

i. DATA SELECTION

The data used in our research are sourced from the Bureau of Transportation Statistics, which is part of the U.S. Department of Transportation. However, more data regarding weather and airports are used to enrich this dataset, originating from IEM's AWOS network (publishing 10-minute interval data real-time, aggregated to one-hour interval data for this research) and OpenFlights respectively. The dataset is thoroughly presented in Appendix A3. The weather dataset was merged, not at the time of the flight, but 4 hours prior to it. We decided to use data 4 hours before the scheduled departure time throughout the project because it gives enough time to the carriers to try to prevent any unwanted situations and makes the models more useful for the carriers.

Moreover, after conducting Exploratory Data Analysis, we decided to focus on the ten busiest airports for the specific airlines we aimed to examine. Through the exploratory analysis we found that airports have different characteristics and differ a lot in terms of delays and cancellations, thus, to include that information in the modelling we decided to reduce the number of airports. There were more than *150* airports in the dataset, and some of them were not used by the airlines of interest. Using the whole dataset would allow us to discover more general trends but it wouldn't be suitable for making recommendations to the studied airlines. It was calculated that the ten busiest airports account for *20%* of the total flights in 2008 and *54,7%* of the flights operated by F9 and MQ airlines.

ii. MISSING VALUES

Missing values need to be processed before proceeding to implement the models since they have a strong impact on the model's performance. Chakrabarty (2019) completely removed missing values in order to predict flight delays. On the other hand, Sternberg et al. (2017) attempted to replace missing values with means or medians, and where the replacement didn't provide a good result, the values were dropped.

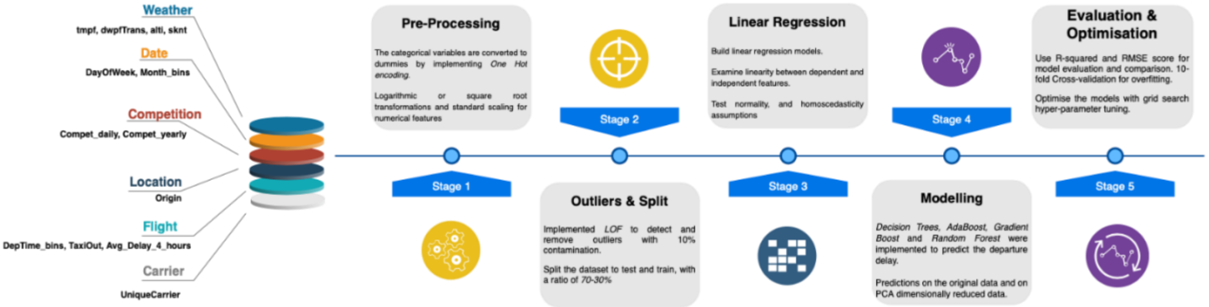
Observations with missing weather or/and arrival delays for non-cancelled flights were dropped (*0.23%* of the dataset). *CRSElapsedTime* values were filled with deterministic regression imputation. This type of imputation replaces the missing values with the exact result of the linear regression between explanatory variables and missing cases in a single variable.

iii. FEATURE ENGINEERING

We then employed feature engineering techniques to enrich our dataset by creating new features using the existing data. The feature engineering implemented, added five new features in the dataset. Competitiveness, calculated as the total percentage of flights an airline executes in each airport, both in a daily and yearly basis. The taxi-out time, the Average Departure Delay and the Average Departure Delay 4 hours prior to a flight were also added in the features. All 3 features are measured for each airport on a daily basis.

The *CRSDepTime* and *CRSArrTime* were converted into categorical values ranging from *1* to *24*. Since we wanted to study the differences between different times of a day and different periods of a year, we decided to reduce the number of categories of the variables *CRSDepTime* and *Month*, aiming to get better results. We then proceeded into binning the first one into Early Morning (*4-7am*), Morning (*8-11am*), Afternoon (*12-17pm*), Evening (*18-21pm*) and Night (*22pm-3am*). The *DayofMonth* was also binned into five-day categories, while the *Months* was binned into four seasons. The rest of the pre-processing procedures were implemented differently for each task and will be explained in detail in the corresponding report section.

B. PREDICTING DEPARTURE DELAY



Pre-Processing

The categorical variables are converted to dummies by implementing *One Hot encoding.*

Logarithmic or square root transformations and standard scaling for numerical features

Outliers & Split

Implemented *LOF* to detect and remove outliers with 10% contamination.

Split the dataset to test and train, with a ratio of *70-30%*

Linear Regression

Build linear regression models.

Examine linearity between dependent and independent features.

Test normality, and homoscedasticity assumptions

Modelling

Decision Trees, AdaBoost, Gradient Boost and *Random Forest* were implemented to predict the departure delay.

Predictions on the original data and on PCA dimensionally reduced data.

Evaluation & Optimisation

Use R-squared and RMSE score for model evaluation and comparison. 10-fold Cross-validation for overfitting.

Optimise the models with grid search hyper-parameter tuning.

Figure 1: Machine Learning pipeline for regression.

Flight delays have been studied extensively in the literature. Many factors have been considered to build models based on different techniques. Klein et al. (2010) considered weather conditions as a basis for their model. They constructed a *12*-component weather-impacted traffic index (WITI) to assess the weather at airports located in the United States, from *2007* to *2010*. The index included information on wind, snow, convective weather, visibility and more. Delay was quantified as the sum of departure and arrival delay in minutes. Using a multiple linear regression model resulted in an R_2 greater than *70%* so the researchers moved on to use their model on future flights using forecast weather.

On the other hand, Manna et al. (2017) decided to focus only on factors directly related to the flights, considering the *70* busiest airports in the United States, from April to October *2013*. The factors included the day of the week, carrier, origin and destination airport, CRS departure and arrival time, departure and arrival delay. Decision tree regression with gradient boosting was used and the final prediction was a weighted average of the sequence of predictions produced. To ensure efficiency, an additional classifier was added at every stage. The features were normalized to the same type on a scale from *0* to *1*. Furthermore, the outliers were removed from the dataset. Feature analysis was then performed and revealed that there was a high correlation between arrival and departure delays. Additionally, the average departure and arrival delays were found to be the highest on Wednesdays and Thursdays. Furthermore, airlines with IATA codes of WN and F9 had the highest average arrival and departure delays.

The model was trained on the two delay types separately. The performance of the model was evaluated using mean absolute error, coefficient of determination *(R_2)* and mean square error. The results showed a high R_2 and low error values indicating that selected features were good predictors for both arrival and departure delays.

In terms of the research associated with predicting flight delays, Klein et al. (2010) have used a multiple regression to predict flight delays, based on different factors of weather. Even though, the resulting R_2 was found to be very significant, the paper lacks the variety of other features, which can potentially explain flight delay. We have intended to fix this problem by including other potentially important features, such as level of competition, taxi out time and more. Mann et al. (2017) in their study have used decision tree regression to predict flight delays. Their regression has included a more diverse set of features, such as day of a week, carrier type. However, the regression does not include one of the most important determinants of the flight delays, that is the weather, which we are using in our regression analysis. Furthermore, we are employing many different algorithms, tree-based models and linear regression to tackle the problem which gives allows us to compare models and determine the most suitable algorithm for the task.

i. PRE-PROCESSING

Beyond, the initial pre-processing, we needed to do additional steps, pivotal for regression models. In order to predict the flight delays, *Dest, CRSArrTime, FlightNum, CancellationCode and ArrDelay* features were removed. Since some of the features are categorical, they must be encoded before being utilized to machine learning models. One Hot Encoding is a method of categorizing data using binary variables. Chakrabarty (2019) used both Label and One Hot Encoding. However, One Hot is preferable to Label Encoding since it transforms the data without the risk of unintentionally assigning higher weights to higher numbers and hence treating every category equally. The categorical variables are thereby converted to dummies by implementing One Hot encoding. In order to implement regression models on the data, normalization is required. *TaxiOut, CRSElapsedTime, Distance, sknt* and *dwpf* are transformed logarithmically. *MinMaxScaler* was used in order to scale the values of *DepDelay* and *WeatherDelay* variables between *0* and *1*. The square root of the *MinMaxScaler* scaled values is then calculated, in order to normalize the distribution. All the numerical variables of the data are then normalized by using *StandardScaler* with a mean value of *0* and standard deviation of *1*. Finally, the dataset is split into train and test data, with a ratio of *70-30%*.

Extreme values can also have an impact on the model's performance. An extreme value is an observation with a particularly high or low value compared to the rest when performing univariate analysis, or a rare combination of variables in multivariate analysis. The existence of such values can be considered as noise and it is important to identify whether they affect the model's quality by implementing outlier detection techniques. Traditional methods tend to exclude the tails of the distribution, which are important since they contain extreme values. (Breunig et. al., 2000) suggested that *Local Outlier Factor* (LOF*)* includes most of the desirable properties and therefore it consists of a more accurate detection method for outliers. *LOF* is an algorithm that examines the density of a single observation, by analyzing its neighbors. In our research, a multivariate *LOF* algorithm is implemented in order to detect the outliers. The number of nearest neighbors examined was set to *30*, since *30* is the default setting of *Sklearn* library for a medium-sized dataset. Hence the focus of the algorithm is not on a local scale, resulting in a good tradeoff between local and dataset-wise level. The contamination parameter was set to *0.05*, assuring that outliers consisting *5%* of the total dataset are detected. That was done to ensure that delays were not caused by special unpredictable circumstances such as a major security emergency. We then proceed on dropping these outliers.

ii. MODELLING

In order to predict the delay time, different regression models should be implemented. However, before proceeding, the validity of linear regression assumptions needs to be checked. The linearity of the data is examined via scatterplots, heatmaps and pair plots. Only some pairs of variables show a linear relationship both before and, after the transformations. The normality of errors is then examined with the use of *QQ* plots and statistical tests *Shapiro-Wilk, Jarque-Bera* and *Kolmogorov-Smirnov,* where under the null hypothesis the errors are normally distributed. It appears that normality is violated, even after square root and logarithmic transformation. Finally, the homoscedasticity assumption is examined, by implementing *Breusch-Pagan* and *Goldfeld-Quandt* tests, where under the null hypothesis the error variances are equal. In conclusion, due to assumptions violation, linear regression cannot be implemented for the prediction of flight delays. [Appendix A5]

Therefore, *Decision Trees, AdaBoost, Gradient Boost* and *Random Forest* regression models are used in order to predict the delay times. For the evaluation of the models, we have used Residual squared (R_2) and *Root Mean Square Error (*RMSE*). RMSE* is an error measure, while R_2 compares the model to a benchmark giving us the idea of how good a model is compared to a random model. Both the evaluation metrics are used to determine the best algorithm, as each one has its own advantages and disadvantages. Furthermore, we used 10-fold cross-validation to verify that the models are not overfitting and ensure that our results are robust. Finally, to build our feature set, we used *Recursive Feature Elimination* (RFE) to get the *10* best features for each model. *RFE* removes the weakest feature each time until it reaches the predefined threshold which we set at *10*. *RFE* helps us to reduce the complexity of the models but also provides us with the most influential features.

Furthermore, *Principal Component Analysis* (PCA) was used to reduce the dimensionality of the dataset. Primarily, to evaluate whether running the models on lower dimensions would give us better results. To do that, we ran the models on *10* components which explain about *80%* of the dataset's variance. The effectiveness of the algorithms did not improve training on the components instead of the actual features. Nevertheless, using the components would not be suitable to make recommendations based on the findings.

C. PREDICTING FLIGHT CANCELLATIONS

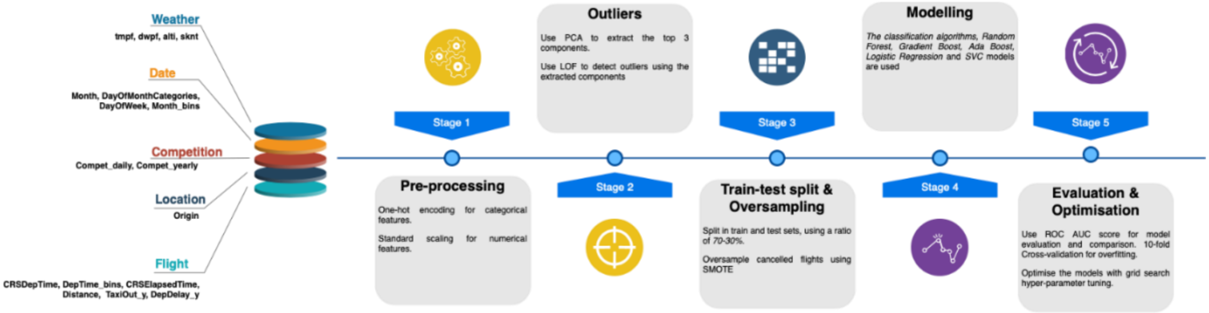

Weather
tmpf, dwpf, alti, sknt
Date
Month, DayOfMonthCategories, DayOfWeek, Month_bins
Competition
Compet_daily, Compet_yearly
Location
Origin
Flight
CRSDepTime, DepTime_bins, CRSElapsedTime, Distance, TaxiOut_y, DepDelay_y
Stage 1
Pre-processing
One-hot encoding for categorical features.
Standard scaling for numerical features.
Outliers
Use PCA to extract the top 3 components.
Use LOF to detect outliers using the extracted components
Stage 2
Stage 3
Train-test split & Oversampling
Split in train and test sets, using a ratio of 70-30%.
Oversample cancelled flights using SMOTE
Modelling
The classification algorithms, Random Forest, Gradient Boost, Ada Boost, Logistic Regression and SVC models are used
Stage 4
Stage 5
Evaluation & Optimisation
Use ROC AUC score for model evaluation and comparison. 10-fold Cross-validation for overfitting.
Optimise the models with grid search hyper-parameter tuning.

Figure 2: Machine Learning pipeline for classification.

In addition to weather data, Rupp and Holmes (2006) considered data on airport congestion, for United States domestic flights between *1995* and *2001*. They have also addressed the issue of correlated flights by removing two-way flights from their data set. A probit binary regression model was used with independent variables including the distance of the flight, weather conditions, aircraft age, competition levels on certain routes and more. Interestingly, competition was found to be a significant factor in determining flight cancellation, with the probability of flight being cancelled on competitive routes being much lower, compared to, for example, monopolistic routes. This however did not hold when researchers controlled for airport-specific factors, such as the size of an airport. Estimates on the link between occupancy rates and flight cancellations were found to be significant, where the full planes were less likely to be cancelled, compared to non-full ones.

Xiong and Hansen (2013) in their research have focused on modelling flight cancellations. The study has used data on US domestic flights in 2006. The model used was a logistic regression, where the dependent variable has recorded whether the flight was cancelled or not. The dependent variables in the model have included distance, average fare, aircraft size, and more. The results have revealed that all features were found to be statistically significant. When looking at the size of an aircraft, it was found that smaller aircrafts had a higher probability of cancellation compared to a larger aircraft. Furthermore, the results have revealed that flights of the major airlines were more likely to be cancelled, compared to regional airlines. The paper gives a great insight into the variety of factors, which can potentially contribute to flight cancellation. On the downside, researchers do not include any weather features, which were previously found to have a strong impact on both delays and cancellations (Rupp and Holmes 2006).

The set of features used in our classification differs significantly, when comparing to this paper. However, the main difference comes from the method implemented, where we used logistic regression, tree-based classifiers and support vector classifier, rather than probit.

i. PRE-PROCESSING

To predict the cancellations, features *mslp, vsby, ArrDelay, DepDelay, Dest, UniqueCarrier, CRSArrTime, FlightNum, Avg_Delay_4hours, TaxiOut* are removed from the original dataset. One Hot Encoding is implemented in order to convert the categorical variables into dummies. The numerical data are then normalized, using the *StandardScaler*, following a normal distribution with a mean value of *0* and a standard deviation of *1*.

To remove the outliers, *PCA* was implemented to extract the top three components. *LOF* was then used on the top *3* components, with *30* nearest neighbors and *5%* contamination. Based on the results of *LOF*, the outliers are then removed from the initial dataset.

Train and test split with a ratio of *70-30%* was then conducted. An imbalance between the classes can lead to a biased model which only focuses on the majority class. In order to balance the classes, under-sampling or oversampling techniques can be implemented. Chakrabarty (2019) used Randomized *SMOTE* in order to balance the classes in the dataset and predict flight delays for American Airlines. *SMOTE* is an oversampling technique based on nearest neighbors, which creates synthetic samples that are not a replication of current ones. Since there is a severe class imbalance in the training dataset, *SMOTENC* is employed to oversample the cancelled flights, as they contain both numerical and categorical features. Five nearest neighbors are used to create the artificial samples, thus ensuring that the artificial samples are representative and realistic for this dataset.

ii. MODELLING

The algorithms used are *Random Forest*, *Gradient Boost*, *Ada boost*, *Logistic Regression* and *Support Vector Machine*. Regarding the evaluation of the results, *Receiver Operating Characteristic Area Under the Curve* (ROC AUC score) was used instead of accuracy. The *ROC* curve is plotted as True Positive Rate against False Positive Rate at different thresholds. *ROC AUC* score ranges from *0* to *1*, where *0.5* means that the model cannot separate between the classes. It was selected over other metrics due to its' robustness to class imbalance and usage of the whole range of thresholds to compute the score.

As with regression, we use *RFE* to get the *10* top features for each algorithm, so we can then build the rest of our analysis based on them. Grid search is used to tune the model parameters and the 10-fold cross validation *ROC AUC* score is calculated in order to find out whether a model is overfitting on the training dataset.

4. RECOMMENDATIONS

After formulating our research questions based on interesting findings through exploratory analysis, we need to answer those questions in order to provide actionable insights to the studied airlines. To answer the questions, we used machine learning modeling with the task of predicting whether a flight will be cancelled or delayed based on the available information, 4 hours before the flight. We then evaluated the performance of the used algorithms and the impact of each available feature to determine the ones that make the difference.

A. FLIGHT DELAYS

MARKET SHARE / AIRPORT COMPETITION

The results showed that the features are very relevant and had a significant relationship with the probability of a flight being delayed. Based on these findings we recommend airlines to re-evaluate their flight management process focusing on the periods they have a relatively high number of flights per day at a specific airport. In addition to that, given that the average 4-hour delay was primarily the most important feature, the airlines should reconsider operating during peak times at airports having a large average delay as it might be the airport's inability to accommodate large number of people causing the delays.

We have noticed that for the airports where our airlines have a high market share, the average departure delays tend to be higher. To summarize, the airlines should look to re-evaluate their flight management at the airports where they have a high percentage of the airport flights during a day. More specifically, MQ should be more careful at operating flights departing from DFW, ORD and RDU, while F9 should do the same for DEN, since the high percentage of flights during a day, seems affect delay times. The table of market share and average departure delays can be found in Appendix A6.

OTHER FINDINGS

Weather is an important factor in the airline industry as it greatly affects the ability of an aircraft to commence its flight safely. This fact leads us to study the importance of weather factors on a flight’s delay by taking into account multiple weather measurements 4 hours prior to the flight’s scheduled departure time. Using tree-based models, we ran four different algorithms on all the available features on three datasets that were divided into the following three sub-datasets:

1. All airlines
2. F9 and MQ only
3. All Other airlines

Using a feature elimination technique and then observing the feature importance ranking according to each model and dataset, in addition to partial dependence plots, we can answer the first question about flight delays. We found that weather factors were indeed important in determining whether the flight will be delayed or not. The most notable features across all datasets and models were temperature (*tmpf*) and dew point temperature (*dwpf*), as presented in Figure 3.

Following up on these findings, we studied whether our airlines are more affected by weather factors than others. Using the three datasets mentioned above, we used the top two models with the best results (*GradientBoost* and *RandomForest*) to test the effect of weather features on the model performance. We ran the models with cross validation under three different input scenarios which consisted of weather features only, all selected features except the weather features and all the features. The results showed insignificant difference across all three datasets, indicating that our airlines are not unique when it comes to weather caused delays, hence being an industry wide issue. Moreover, the effect of weather features on the performance of the model was found to be minor.

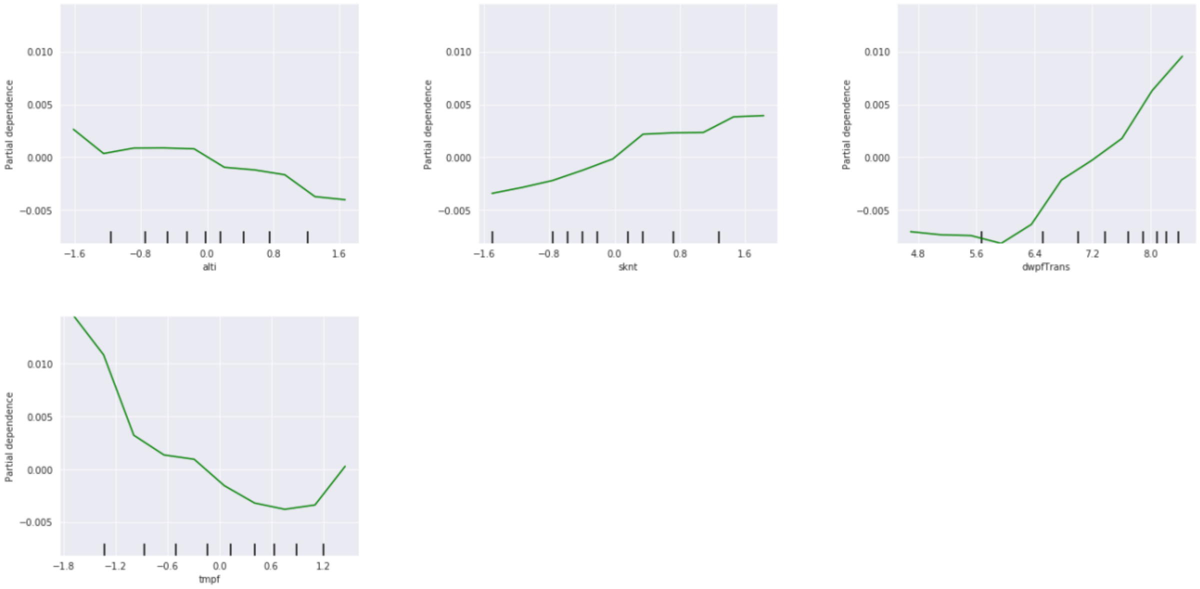


Figure 3: Partial dependence plots of Gradient Boost for the four weather features.

B. FLIGHT CANCELLATIONS

Flight cancellations are a huge burden to an airline’s revenue and reputation; thus, even minor causes should be investigated in order to reduce the probability of cancellations happening.

MOVING DEPARTURE TIME

Answering the first research question, we found that the probability of a flight being cancelled depends on the time of the day. As explained before, we divided the day into five parts, “Early Morning”, “Morning”, “Afternoon”, “Evening” and “Night”. All our algorithms indicate that flights operated early in the morning or during the night have fewer chances of getting cancelled [Appendix A7]. On the other side, afternoon flights are constantly ranked higher in the partial dependency plots, implying that flights during afternoon are more prone to cancellations. This finding might be explained by the airport traffic, as afternoon tends to be busier in terms of flight volume, while early morning and night flights are usually less congested. While in some cases it is not feasible, we would recommend that the studied airlines should consider changing the departure time for some of their flights in order to reduce the possibility of their flight being cancelled.

REDUCE THE USAGE OF ORD

Furthermore, we studied the airports and time periods which are more prone to flight cancellations. After using the *Origin* and *Season* separately in our models, we found that certain airports were positively correlated with cancellations. More specifically, from the coefficients of Logistic regression and the partial dependence plots of the ensemble methods, we found that the airports ORD and DCA showed a positive relationship with cancellations. Regarding the seasons, autumn was the only season with negative coefficients, implying that flights during that period have a lower probability of being cancelled. [Appendix A8]

In search of more specific results, we used the interaction between the two features. We observed that the interaction features, enhanced the results of Random Forest in relation to the results without the interaction. The following combinations were positively affecting the probability of a flight being cancelled for the algorithm: ORD-Spring, ORD-Winter and SAN-Winter. [Appendix A8]

We have repeatedly found that ORD airport is positively correlated with cancellations. MQ uses ORD for the majority of its flights and also it is the airport with the highest cancellation rates for the carrier. We would like to recommend to MQ, to reduce the usage of ORD, if possible, especially during Winter and Spring. ORD is located in the city of Chicago which is home to four other airports.

OTHER FINDINGS

Answering the second research question, it was found that models including the weather were able to make better predictions for NAS and Weather cancellations but functioned poorly in predicting Carrier cancellations. This might mean that carrier cancellations are not affected by weather, and other reasons may explain this. We are unable to find those reasons within the used dataset, because after training only on Carrier cancellations, the model still operates poorly. Determining, what exactly causes Carrier cancellation is something worthwhile looking into in the future. We assume that Carrier cancellations are more related to staffing and other airport operations which are not included in our data. Furthermore, it seems that NAS cancellations are mainly affected by the weather. In Appendix A10 we can see the results of the predictions, made on the separate datasets.

C. DENVER AIRPORT

After carrying out separate analyses for delays and cancellations, we discovered something interesting in relation with Denver International Airport (DEN). As found in the classification analysis, DEN is the airport with the highest negative effect on the assigned probability of a flight being cancelled as seen on Figure 12 (Appendix A8). This finding implies that flights from DEN are less likely to get cancelled compared to the other airports studied. Furthermore, as presented in Table 5 (Appendix A6) DEN has the highest average departure delay time between the airports. That leads us to the conclusion that flights departing from DEN airport are more likely to be delayed instead of cancelled.

From the studied airlines, only F9 uses DEN. Furthermore, DEN is the largest airport in the United States, thus, this finding might be explained by the magnitude of a flight being cancelled in that airport. We would advice F9 to ensure that departure delays from DEN are not related to their operations and flight management. If the delays are the consequence of the airport's administration, F9 should take into account that departing from DEN implies longer departure delays. We would therefore recommend F9 to take this into consideration, when scheduling the flights.

5. CONCLUSIONS

After exploring the dataset and formulating the research questions, the analysis was carried out under two separate paths. Using classification methods for determining flight cancellations and regression methods for flight delays. The study answered questions relative to weather factors affecting flight delays as well as other prominent factors affecting flight cancellations. Both classification and regression models have performed substantially well, with regression model scoring an R_2 of 0.2 and classification model scoring an AUC of 0.72. Based on our findings we were able to recommend managerial insights to our airlines, Frontier Airline (F9) and Envoy Air (MQ), that would help them improve their operations. Recommendations were based on the machine learning results and were tailored to our airlines based on the most relevant features. In conclusion, the airlines were advised to revisit their fleet management during times of high competition as well as avoid highly congested airports, especially during peak times.

6. REFERENCES

[1] Ball, M., Barnhart, C., Dresner, M., Hansen, M., Neels, K., Odoni, A., Peterson, E., Sherry, L., Trani, A., Zou, B. and Britto, R., 2010, October. Total delay impact study. In *NEXTOR Research Symposium, Washington DC. http://www. nextor. org*.

[2] Breunig, M.M., Kriegel, H.P., Ng, R.T. and Sander, J., 2000, May. LOF: identifying density-based local outliers. In *ACM sigmod record* (Vol. 29, No. 2, pp. 93-104). ACM.

[3] Chakrabarty, N., 2019. A Data Mining Approach to Flight Arrival Delay Prediction for American Airlines. *arXiv preprint arXiv:1903.06740*.

[4] Klein, A., Craun, C. and Lee, R.S., 2010, October. Airport delay prediction using weather-impacted traffic index (WITI) model. In *29th Digital Avionics Systems Conference* (pp. 2-B). IEEE.

[5] Manna, S., Biswas, S., Kundu, R., Rakshit, S., Gupta, P. and Barman, S., 2017, June. A statistical approach to predict flight delay using gradient boosted decision tree. In *2017 International Conference on Computational Intelligence in Data Science (ICCIDS)* (pp. 1-5). IEEE.

[6] Rupp, N.G. and Holmes, G.M., 2006. An investigation into the determinants of flight cancellations. *Economica*, *73*(292), pp.749-783.

[7] Sternberg, A., Soares, J., Carvalho, D. and Ogasawara, E., 2017. A review on flight delay prediction. *arXiv preprint arXiv:1703.06118*.

[8] Xiong, J. and Hansen, M., 2013. Modelling airline flight cancellation decisions. Transportation Research Part E: Logistics and Transportation Review, 56, pp.64-80.

7. APPENDIX

A1. FLIGHT DELAYS EXPLORATORY ANALYSIS

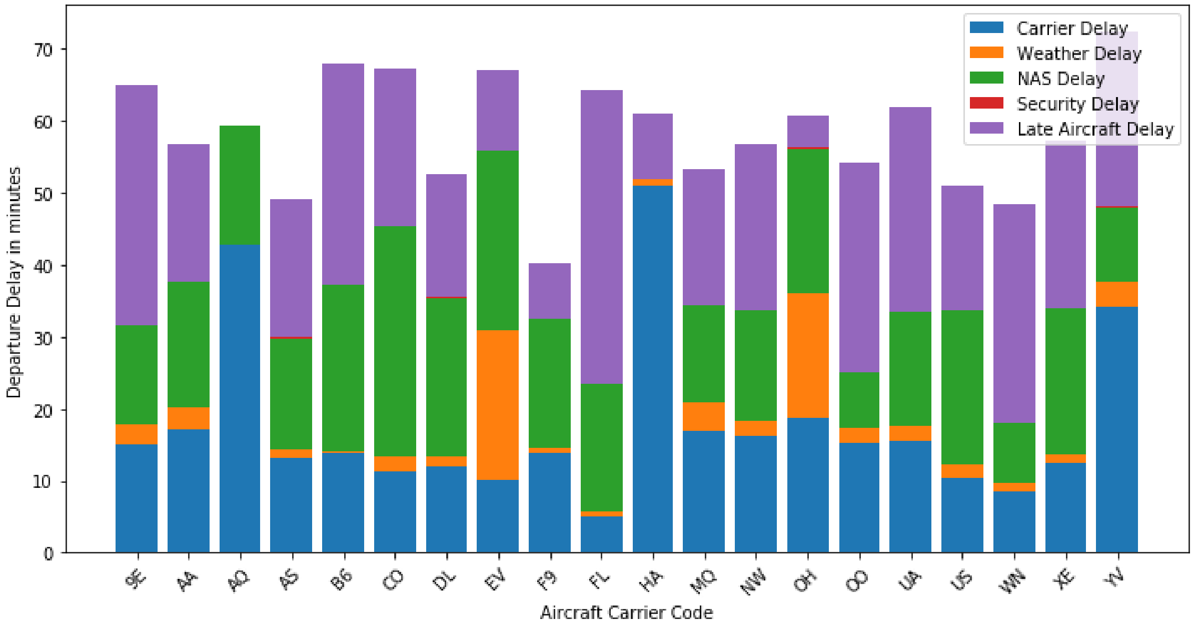


Figure 4:Average Departure Delay in minutes for each carrier as the total of different delay reasons.

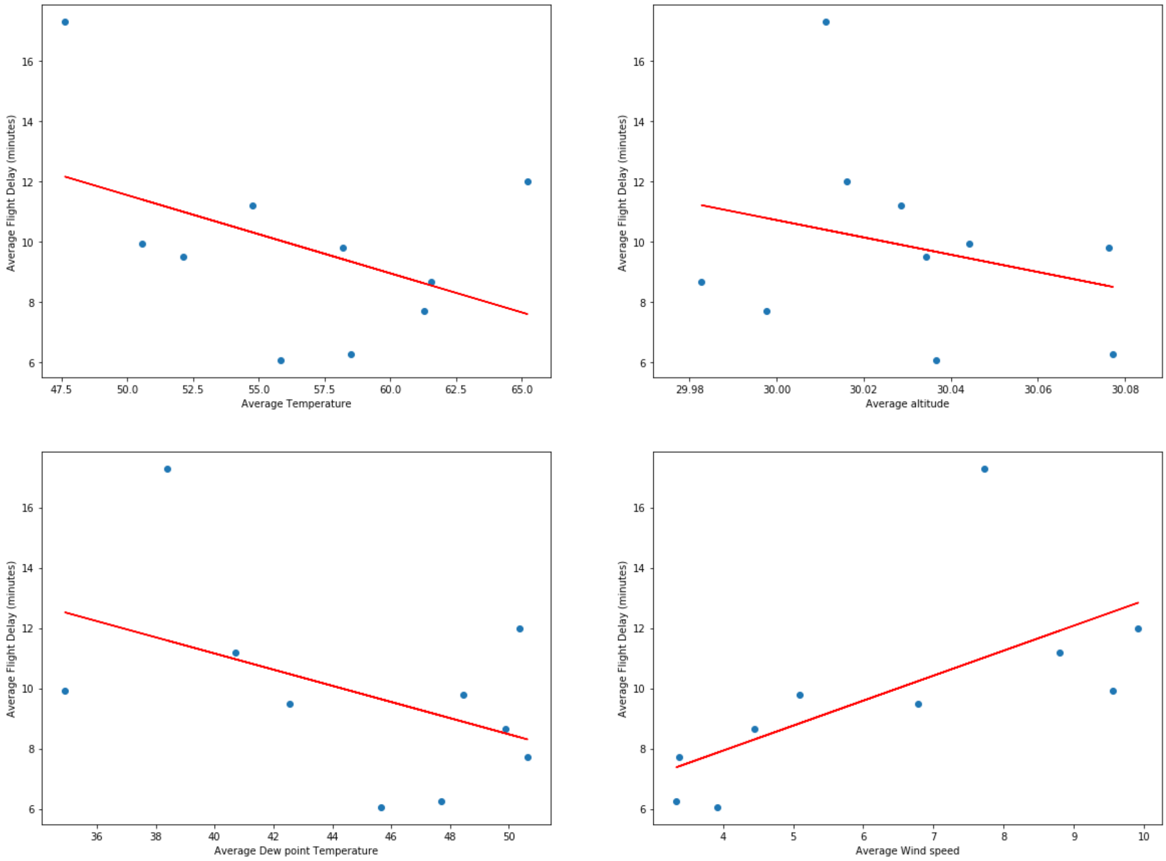


Figure 5: The averages of the four weather features plotted against the average Departure Delay for the 10 studied airports.

A2. FLIGHT CANCELLATIONS EXPLORATORY ANALYSIS

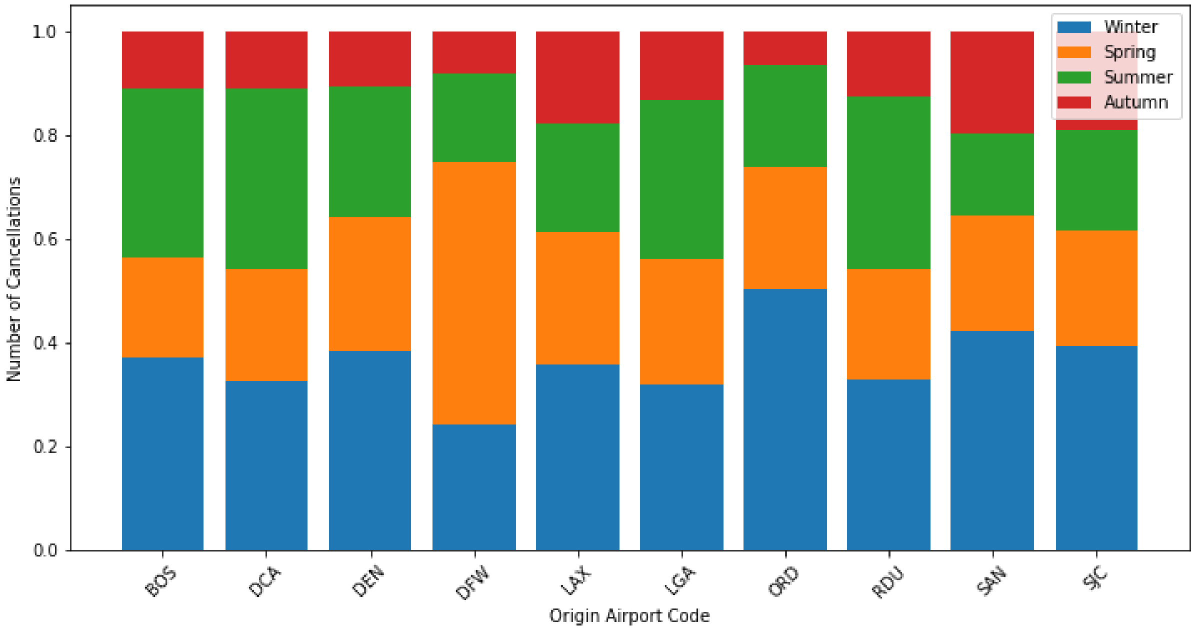


Figure 6: Number of cancellations for each airport colored by season.

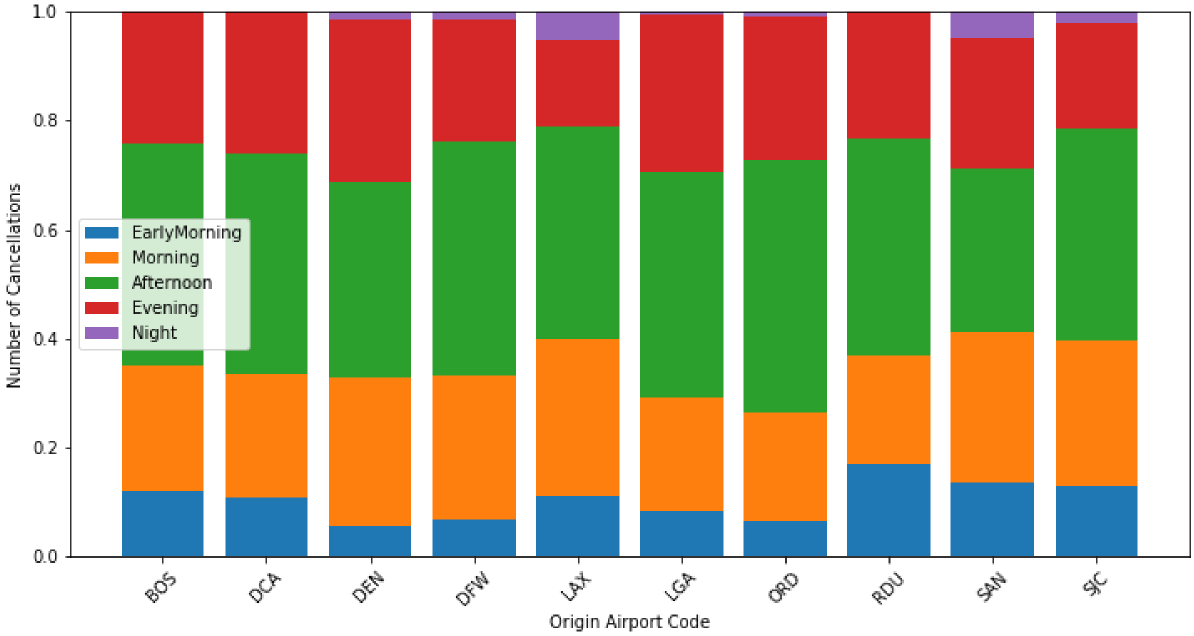


Figure 7: Number of cancellations for each airport colored by part of the day.

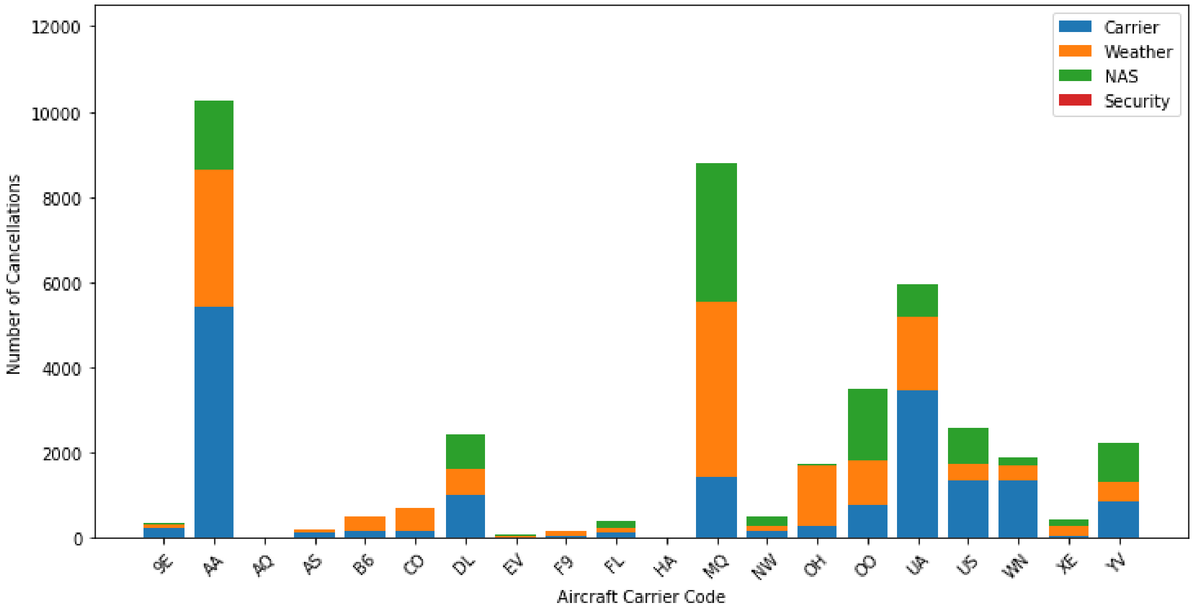


Figure 8: Total Cancellations by carrier colored by Cancellation Code.

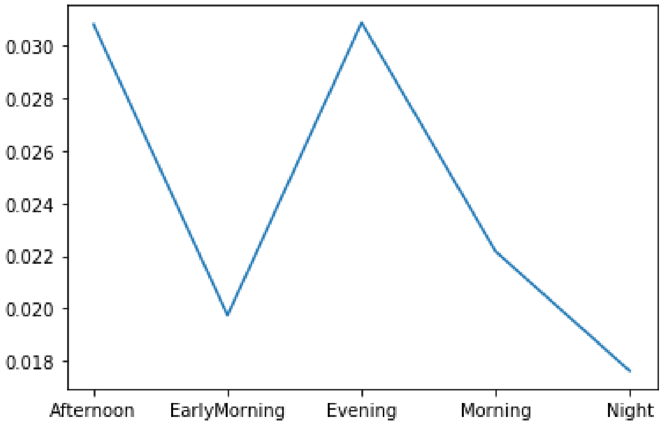
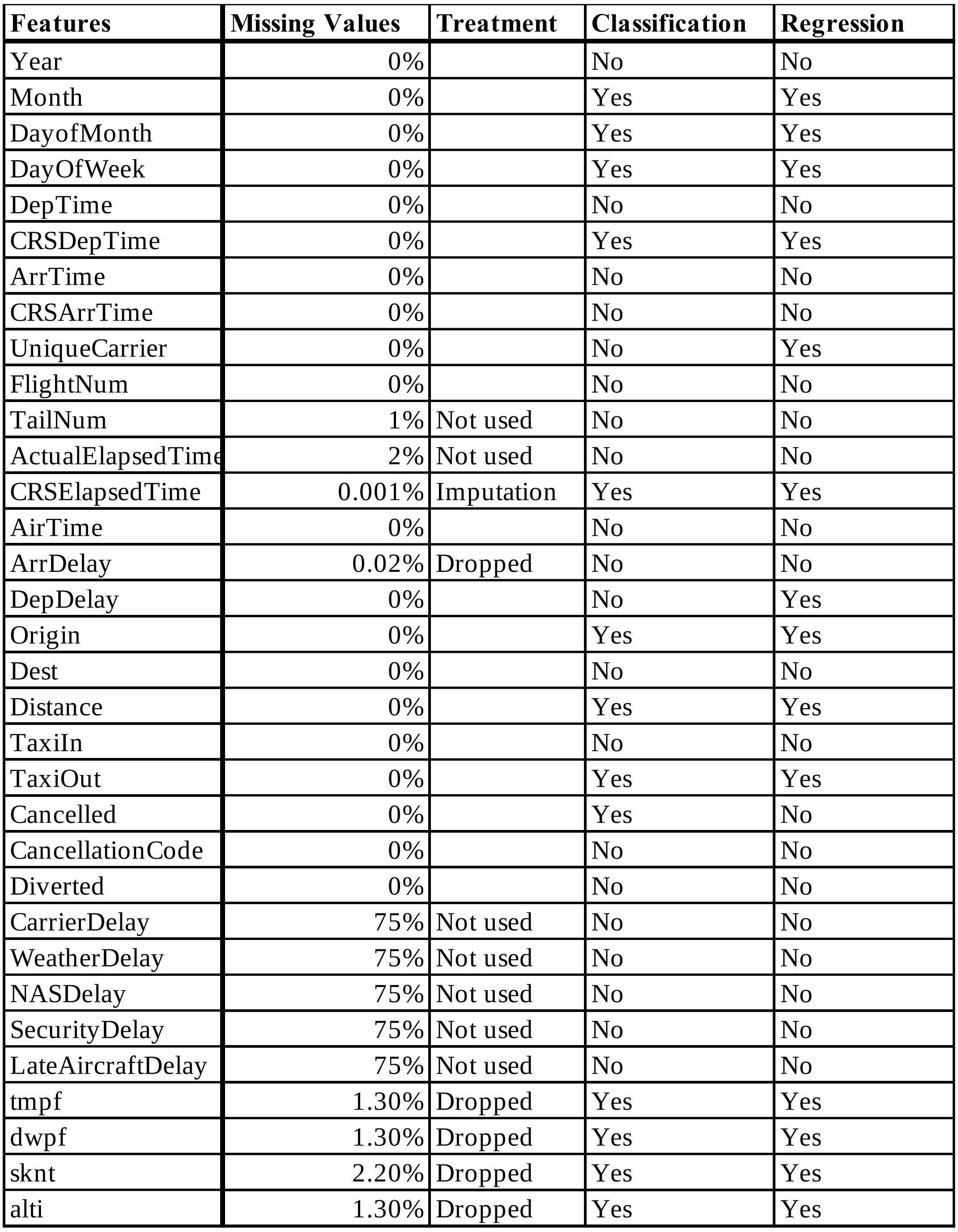


Figure 9: Percentage of Cancelled flights for different parts of the day.

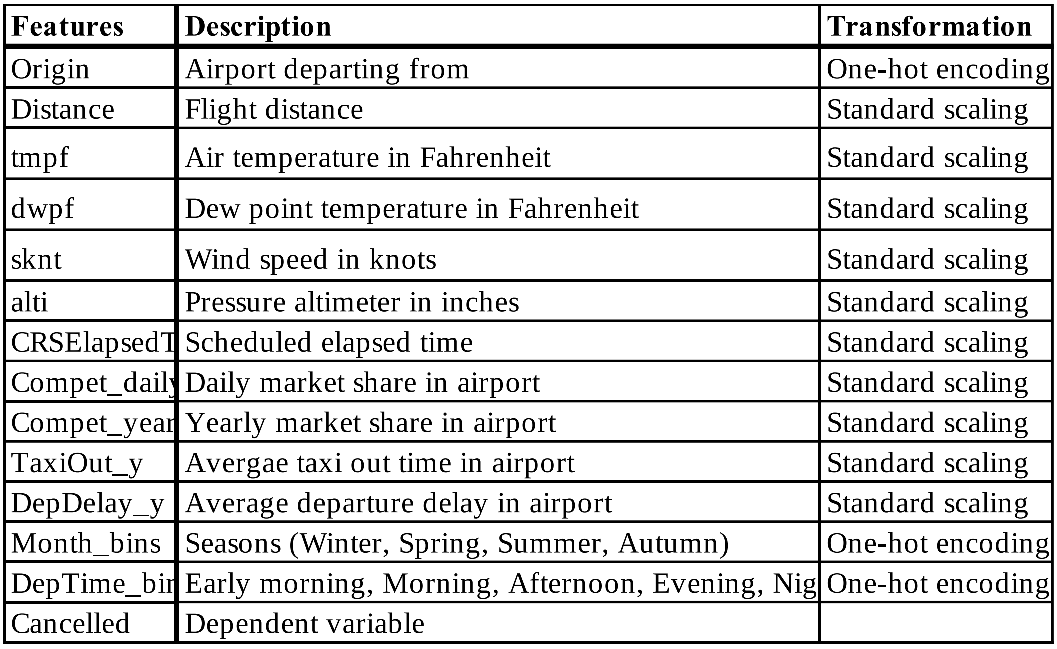
A3. DATA DESCRIPTION



Features	Missing Values	Treatment	Classification	Regression
Year	0%		No	No
Month	0%		Yes	Yes
DayofMonth	0%		Yes	Yes
DayOfWeek	0%		Yes	Yes
DepTime	0%		No	No
CRSDepTime	0%		Yes	Yes
ArrTime	0%		No	No
CRSArrTime	0%		No	No
UniqueCarrier	0%		No	Yes
FlightNum	0%		No	No
TailNum	1%	Not used	No	No
ActualElapsedTime	2%	Not used	No	No
CRSElapsedTime	0.001%	Imputation	Yes	Yes
AirTime	0%		No	No
ArrDelay	0.02%	Dropped	No	No
DepDelay	0%		No	Yes
Origin	0%		Yes	Yes
Dest	0%		No	No
Distance	0%		Yes	Yes
TaxiIn	0%		No	No
TaxiOut	0%		Yes	Yes
Cancelled	0%		Yes	No
CancellationCode	0%		No	No
Diverted	0%		No	No
CarrierDelay	75%	Not used	No	No
WeatherDelay	75%	Not used	No	No
NASDelay	75%	Not used	No	No
SecurityDelay	75%	Not used	No	No
LateAircraftDelay	75%	Not used	No	No
tmpf	1.30%	Dropped	Yes	Yes
dwpf	1.30%	Dropped	Yes	Yes
sknt	2.20%	Dropped	Yes	Yes
alti	1.30%	Dropped	Yes	Yes

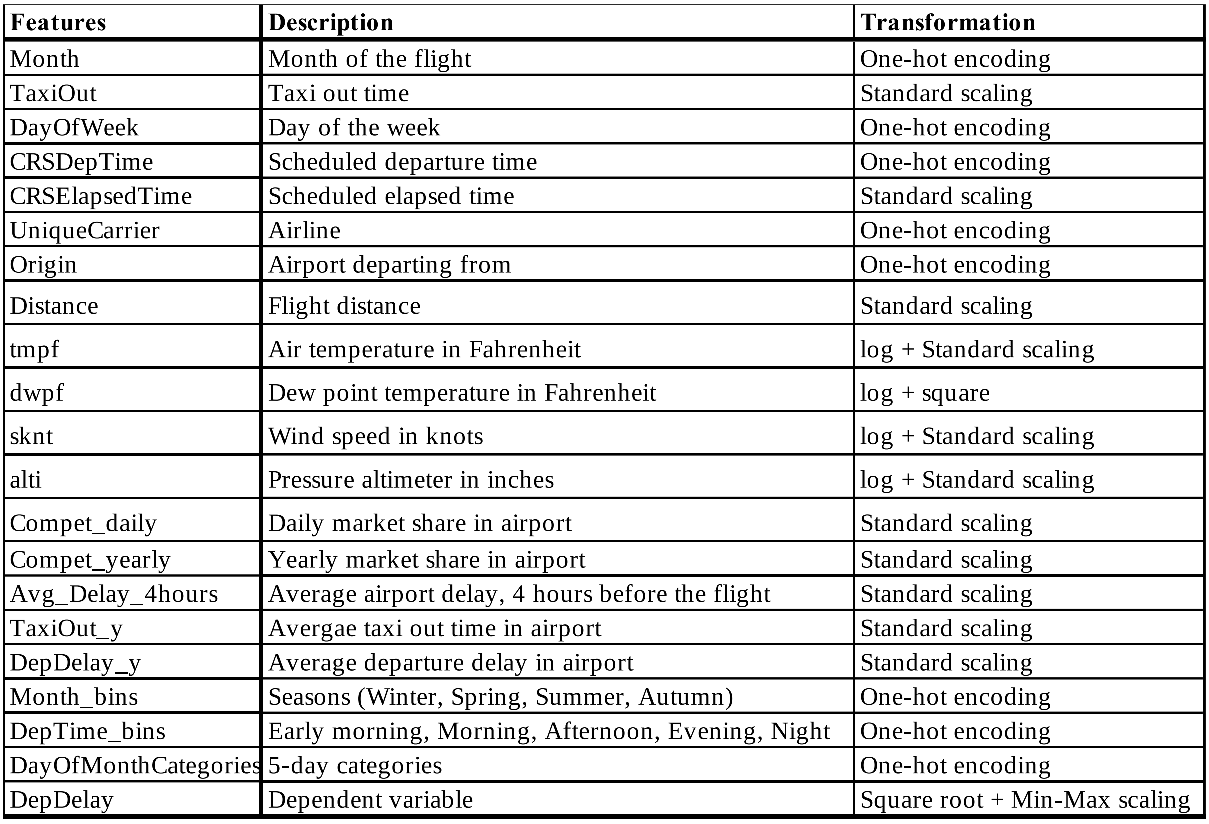
Table 1: Initial dataset plus the four weather features.

A4. FEATURES SELECTED



Features	Description	Transformation
Origin	Airport departing from	One-hot encoding
Distance	Flight distance	Standard scaling
tmpf	Air temperature in Fahrenheit	Standard scaling
dwpf	Dew point temperature in Fahrenheit	Standard scaling
sknt	Wind speed in knots	Standard scaling
alti	Pressure altimeter in inches	Standard scaling
CRSElapsedT	Scheduled elapsed time	Standard scaling
Compet_dail	Daily market share in airport	Standard scaling
Compet_year	Yearly market share in airport	Standard scaling
TaxiOut_y	Avergae taxi out time in airport	Standard scaling
DepDelay_y	Average departure delay in airport	Standard scaling
Month_bins	Seasons (Winter, Spring, Summer, Autumn)	One-hot encoding
DepTime_bir	Early morning, Morning, Afternoon, Evening, Nig	One-hot encoding
Cancelled	Dependent variable	

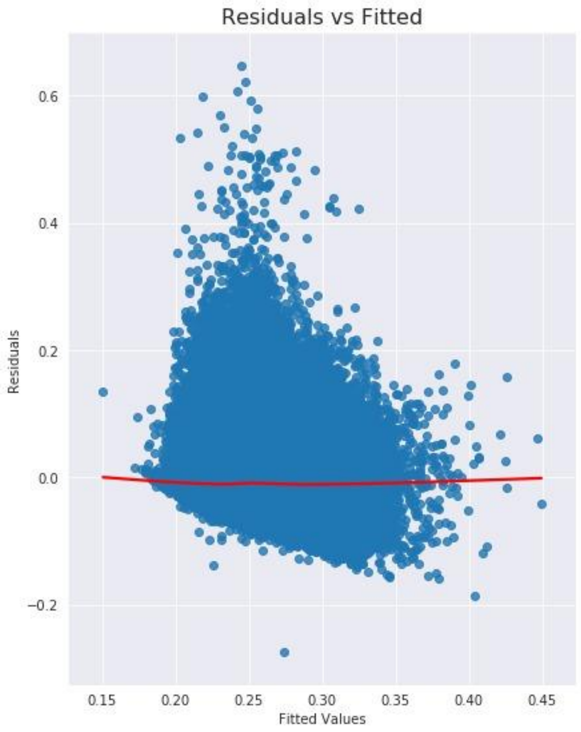
Table 2: Features selected for classification including description and transformation before modeling.

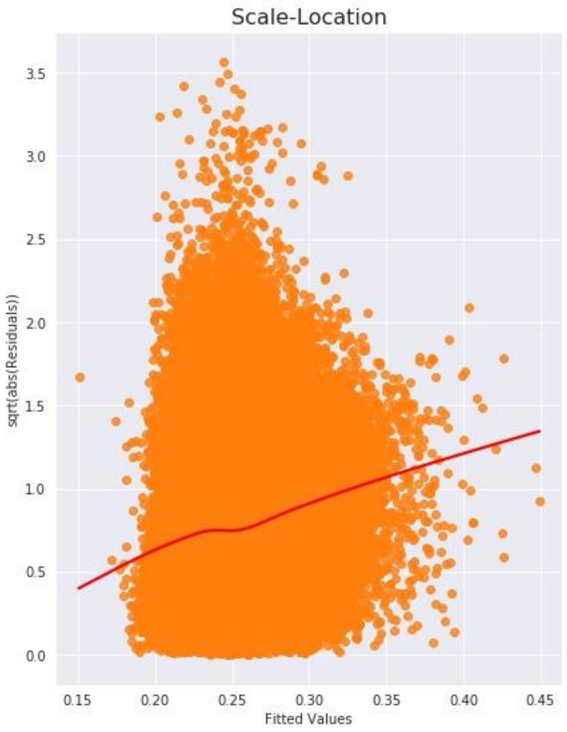


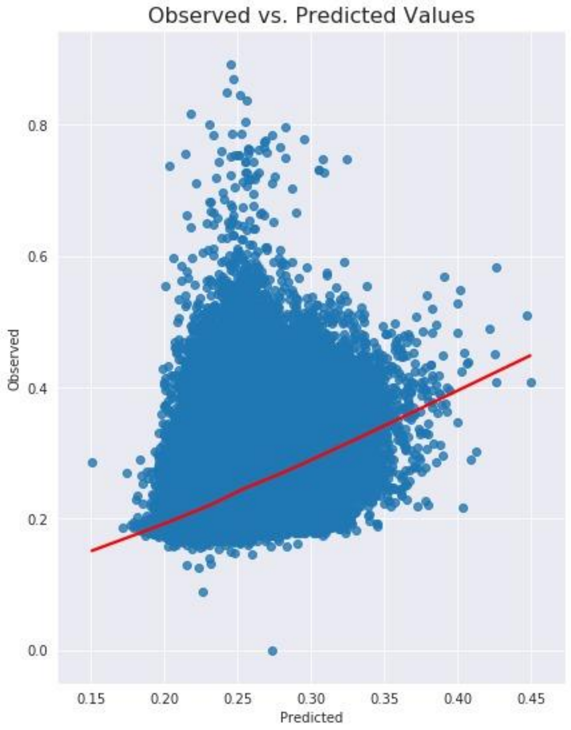
Features	Description	Transformation
Month	Month of the flight	One-hot encoding
TaxiOut	Taxi out time	Standard scaling
DayOfWeek	Day of the week	One-hot encoding
CRSDepTime	Scheduled departure time	One-hot encoding
CRSElapsedTime	Scheduled elapsed time	Standard scaling
UniqueCarrier	Airline	One-hot encoding
Origin	Airport departing from	One-hot encoding
Distance	Flight distance	Standard scaling
tmpf	Air temperature in Fahrenheit	log + Standard scaling
dwpf	Dew point temperature in Fahrenheit	log + square
sknt	Wind speed in knots	log + Standard scaling
alti	Pressure altimeter in inches	log + Standard scaling
Compet_daily	Daily market share in airport	Standard scaling
Compet_yearly	Yearly market share in airport	Standard scaling
Avg_Delay_4hours	Average airport delay, 4 hours before the flight	Standard scaling
TaxiOut_y	Avergae taxi out time in airport	Standard scaling
DepDelay_y	Average departure delay in airport	Standard scaling
Month_bins	Seasons (Winter, Spring, Summer, Autumn)	One-hot encoding
DepTime_bins	Early morning, Morning, Afternoon, Evening, Night	One-hot encoding
DayOfMonthCategories	5-day categories	One-hot encoding
DepDelay	Dependent variable	Square root + Min-Max scaling

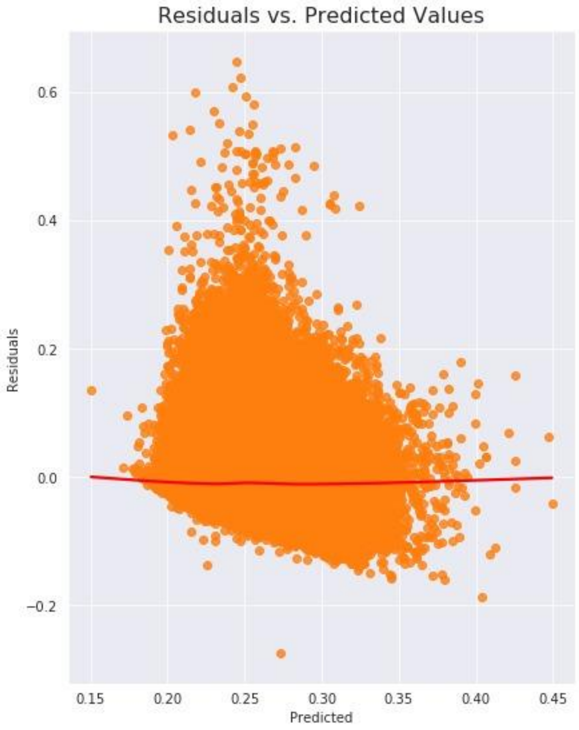
Table 3: Features selected for regression, including description and transformation before modeling.

A5. REGRESSION ASSUMPTIONS









```
-----regular fitted regression
Jarque-Bera test ---- statistic: 16047.6406, p-value: 0.0
Shapiro-Wilk test ---- statistic: 0.8915, p-value: 0.0000
Kolmogorov-Smirnov test ---- statistic: 0.4647, p-value: 0.0000
-----Polynomial fitted regression
Jarque-Bera test ---- statistic: 13266.6373, p-value: 0.0
Shapiro-Wilk test ---- statistic: 0.9075, p-value: 0.0000
Kolmogorov-Smirnov test ---- statistic: 0.4628, p-value: 0.0000
```

Table 4: Tests for normality of residuals

A6. RECOMMENDATION 1 – MARKET SHARE

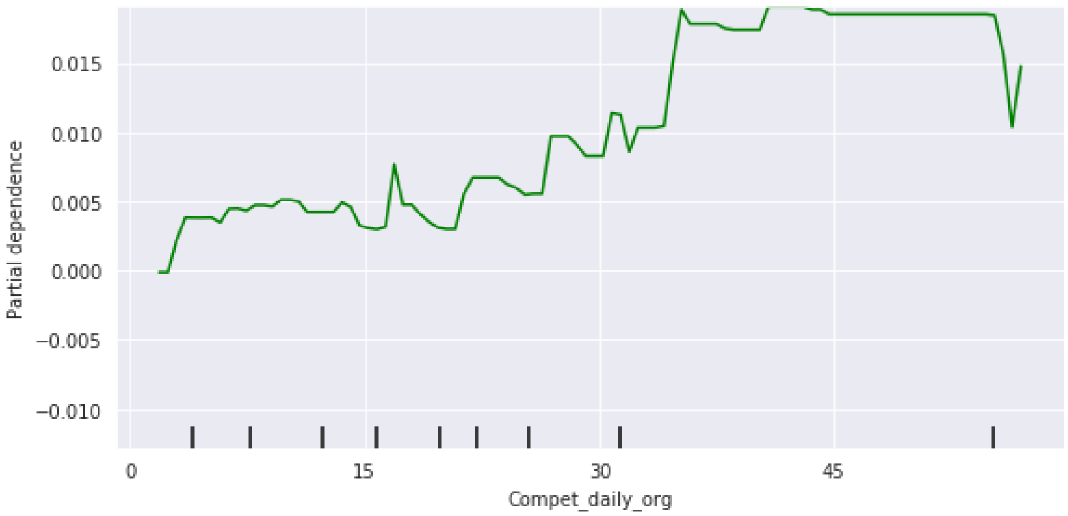


Figure 10: Partial dependence plot of daily market share. Shows the influence of the predictor on the final prediction.

Airport	Average Delay	% of flights
ORD	13.710104	25.016271
DFW	9.539688	31.759251
LGA	8.545529	14.641151
LAX	7.631385	7.536807
SAN	7.414322	8.048483
RDU	7.2422	23.133156
BOS	6.279447	8.535869
SJC	5.60239	10.908021
DCA	4.27522	11.969232

Airport	Average Delay	% of flights
LGA	0.820112	15.719757
DFW	0.7199	10.860125
DEN	19.79357	8.193517
SAN	2.272827	7.649847
LAX	1.05075	7.445946
SJC	2.288267	4.327108
DCA	1.296439	4.035921

Table 5: Average delay and average % of flights for MQ(left) and F9(right).

A7. RECOMMENDATION 2 – DEPARTURE TIME

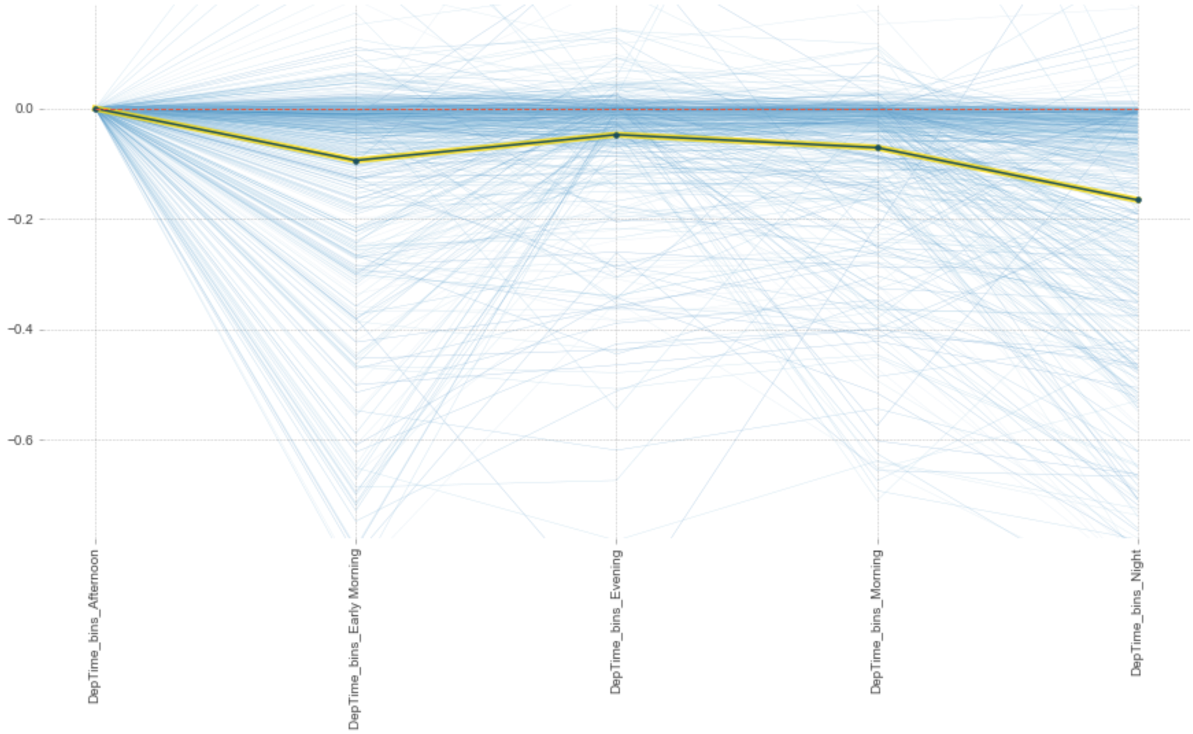


Figure 11: Partial dependence plot of the part of the day vs. the change in the assigned probability of a flight being cancelled for the Gradient Boost Classifier.

	Logistic Regression	Random Forest	Gradient Boost	AdaBoost
Early Morning	-0.9133	0.013	0.01	0.005
Morning	-0.86	0.025	0.01	0.005
Afternoon	Base class	0.059	0.039	0.01
Evening	-0.5401	0.009	0.008	0
Night	-1.6422	0.001	0.002	0.005

Table 6: Coefficients and Feature Importances for each part of the day.

A8. RECOMMENDATION 3 – REDUCE USAGE OF ORD

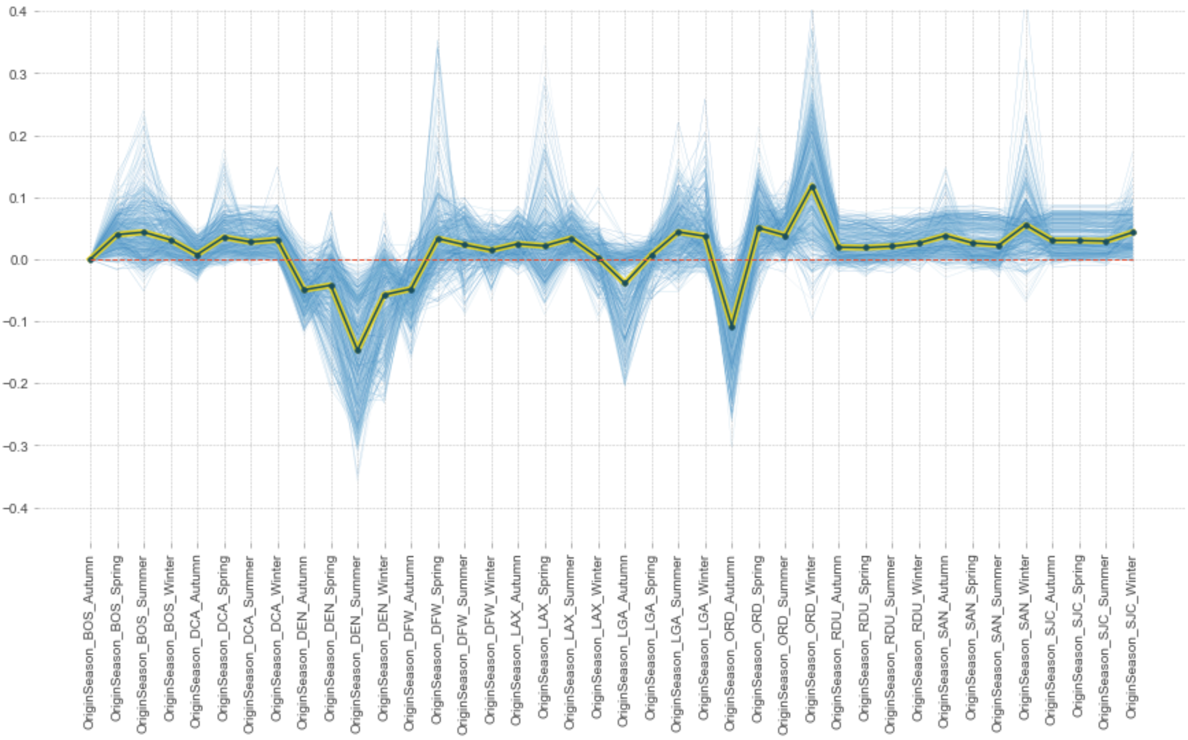


Figure 12: Partial dependence plot of the interaction feature, Origin-Season, vs. the change in the assigned probability of a flight being cancelled for Random Forest Classifier.

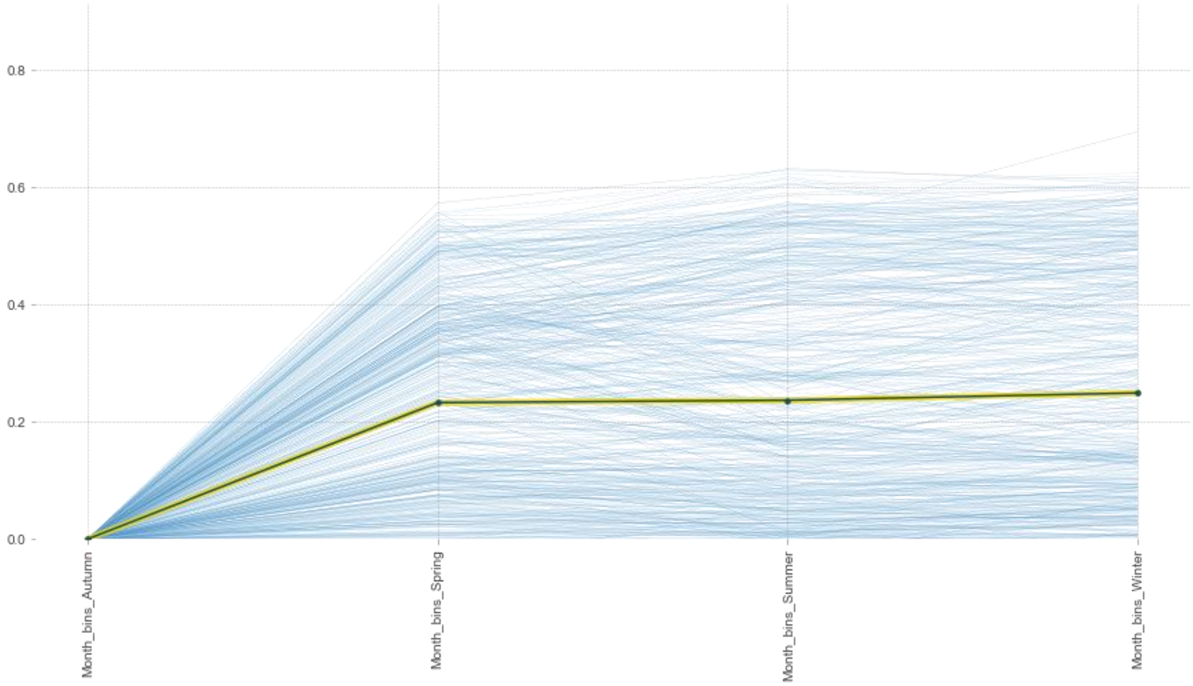


Figure 13: Partial dependence plot of season vs. the change in the assigned probability of a flight being cancelled for Random Forest Classifier.

A9. REGRESSION COLLECTIVE RESULTS

Random Forest			
R-Squared	0.14198	**RMSE**	**0.048954**
Selected Feature	**Importance**	**Selected Feature**	**Importance**
Avg_Delay_4hours	0.60031	*tmpf*	0.00020961
TaxiOut	0.39033	*Origin_DFW*	9.22E-05
Compet_daily	0.0036576	*Origin_LGA*	3.16E-05
Origin_ORD	0.0034706	*DepTime_bins_Afternoon*	2.96E-05
Compet_yearly	0.0018629	*Month_bins_Winter*	0

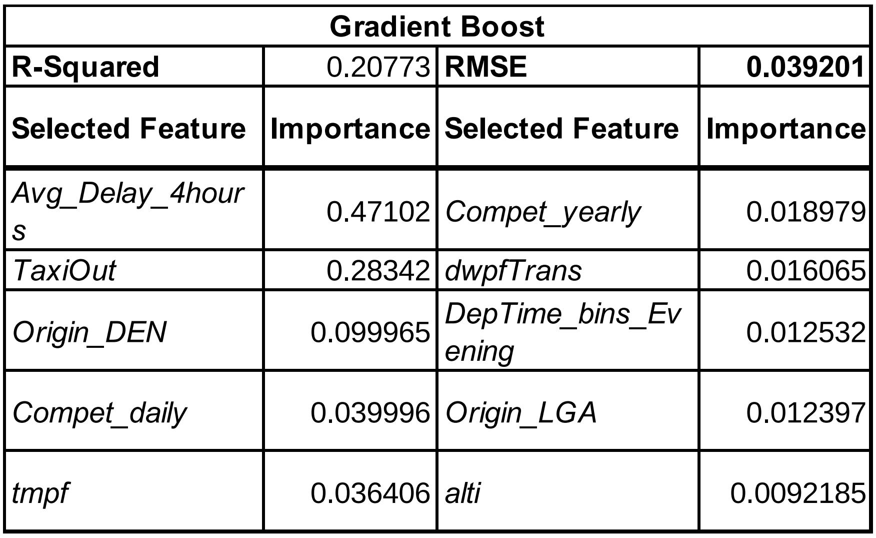
Table 7:Random Forest results on all the airlines, after applying RFE to get the top 10.

Gradient Boost			
R-Squared	0.17984	**RMSE**	**0.047862**
Selected Feature	**Importance**	**Selected Feature**	**Importance**
Avg_Delay_4hours	0.48724	*Origin_LGA*	0.022774
TaxiOut	0.32342	*dwpfTrans*	0. 010994
Compet_yearly	0.065583	*tmpf*	0.010885
Compet_daily	0.034944	*DepTime_bins_Afternoon*	0.010103
Origin_ORD	0.027557	*DepTime_bins_Morning*	0.0064964

Table 8:Gradient Boost results on all the airlines, after applying RFE to get the top 10.

Random Forest			
R-Squared	0.17797	**RMSE**	**0.03993**
Selected Feature	**Importance**	**Selected Feature**	**Importance**
Avg_Delay_4hours	0.54926	*Compet_daily*	0.0016798
TaxiOut	0.32485	*alti*	0.00069005
Origin_DEN	0.10506	*DayOfWeek_4*	0.00059103
tmpf	0.015	*sknt*	0.00036498
dwpfTrans	0.0021897	*Origin_BOS*	31013

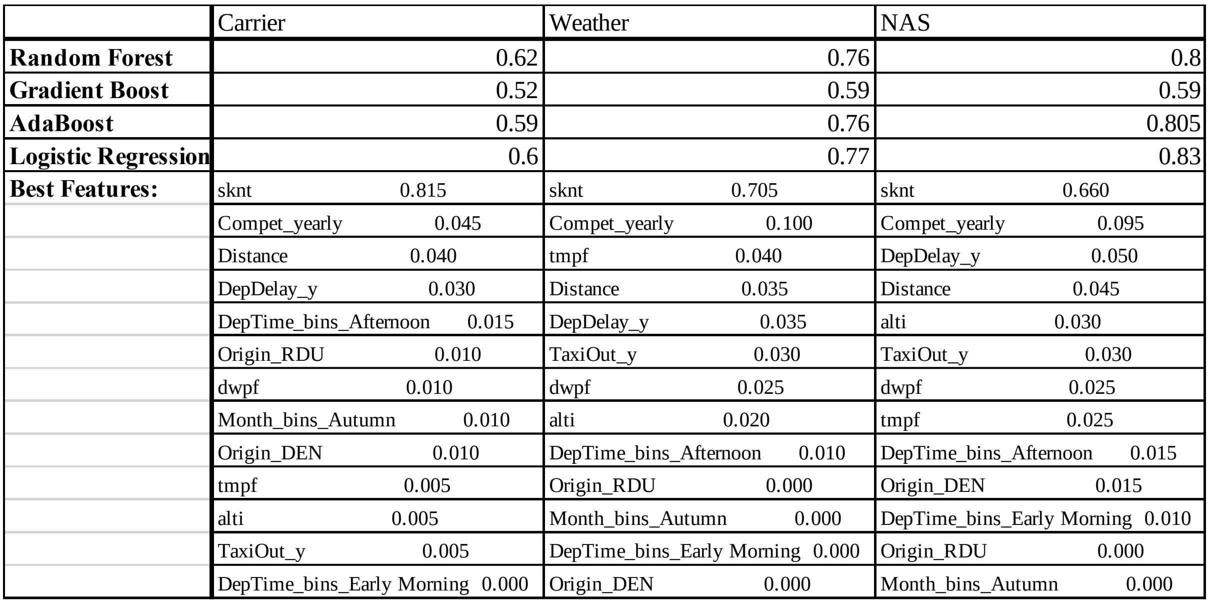
Table 9: Random Forest results on MQ and F9, after applying RFE to get the top 10.



Gradient Boost			
R-Squared	0.20773	**RMSE**	**0.039201**
Selected Feature	**Importance**	**Selected Feature**	**Importance**
Avg_Delay_4hours	0.47102	*Compet_yearly*	0.018979
TaxiOut	0.28342	*dwpfTrans*	0.016065
Origin_DEN	0.099965	*DepTime_bins_Evening*	0.012532
Compet_daily	0.039996	*Origin_LGA*	0.012397
tmpf	0.036406	*alti*	0.0092185

Table 10: Gradient Boost results on MQ and F9, after applying RFE to get the top 10.

A10. DIFFERENT CANCELLATION CODES



	Carrier	Weather	NAS
Random Forest	0.62	0.76	0.8
Gradient Boost	0.52	0.59	0.59
AdaBoost	0.59	0.76	0.805
Logistic Regression	0.6	0.77	0.83
Best Features:	sknt 0.815	sknt 0.705	sknt 0.660
	Compet_yearly 0.045	Compet_yearly 0.100	Compet_yearly 0.095
	Distance 0.040	tmpf 0.040	DepDelay_y 0.050
	DepDelay_y 0.030	Distance 0.035	Distance 0.045
	DepTime_bins_Afternoon 0.015	DepDelay_y 0.035	alti 0.030
	Origin_RDU 0.010	TaxiOut_y 0.030	TaxiOut_y 0.030
	dwpf 0.010	dwpf 0.025	dwpf 0.025
	Month_bins_Autumn 0.010	alti 0.020	tmpf 0.025
	Origin_DEN 0.010	DepTime_bins_Afternoon 0.010	DepTime_bins_Afternoon 0.015
	tmpf 0.005	Origin_RDU 0.000	Origin_DEN 0.015
	alti 0.005	Month_bins_Autumn 0.000	DepTime_bins_Early Morning 0.010
	TaxiOut_y 0.005	DepTime_bins_Early Morning 0.000	Origin_RDU 0.000
	DepTime_bins_Early Morning 0.000	Origin_DEN 0.000	Month_bins_Autumn 0.000

Table 11: Results of the algorithms on different cancellation codes.