Understanding and Predicting Airline operation delay for Atlanta(ATL) International Airport.

Nidhi Mariam Jolly(20196502),Tressy Thomas(20194526)

Cape Breton University

Abstract. Flight delay is a problem that can adversely affect the efficiency of airline operations and degrade customer experience. With the help of data mining techniques we have explored the patterns of arrival and departure delays at the Hartsfield-Jackson International airport, Atlanta in Georgia, USA. Our study on the 2019 data collected from Bureau of Transportation Statistics suggests that flight delay is intense during the month of June and Saturdays have less flights and delays compared to other days of the week. It is also observed that the most flights are getting delayed during morning time. While Delta Airlines Inc accounts over 60 percentage of operations in and out of Atlanta airport, Frontier Airlines recorded highest percentage of delay in operations. In order to increase the efficiency of operations and convenience of customers, it is better to have the knowledge about the occurrence of delay before hand based on the temporal and location factors. We were able to predict whether a flight would be delayed or not in the future by employing three different classifiers: Logistic Regression, k-Nearest Neighbour(kNN) and Extreme Gradient Boost Classifier(XGBoost). We achieved an overall accuracy of 83.45 percentage and 60.67 percentage balanced accuracy with XGBoost but recommends using data imbalance treatments to have a useful model . It is observed that the day of the month and month are very significant factors that influences the delay.

Keywords: Delay Patterns · Classification · Flight delay · Atlanta airport· XGBoost · Imbalanced data.

1 Introduction

Airline industry is very much customer-centred and has been employing several steps to optimize its operations and reliability. Travelers in the US are increasingly using air transport, for it being the fastest mode of transport, mainly for time saving. Providing faster, cheaper and reliable services are crucial factors that affect the efficiency of airline operations as well as customer satisfaction. For this very reason flight departure and arrival delays are significant factors that impacts the reputation and operation of airline services. With the availability of the reliable data, it is possible to provide insights into the patterns and characteristics of airline delays.

Data mining techniques can be leveraged to discover interesting relationships from the data that impact flight delays. Identifying the patterns of delay, predicting the delay and important factors that influences the delay are of interest in this project. Doing so can optimize the operation of airlines, dependent services as well plan schedules and save time for the passengers.

In this study we are focusing on the flight delays pertaining to the Atlanta airport in Georgia, USA. Hartsfield-Jackson is a global gateway, offering nonstop service to more than 150 domestic and 70 international destinations [1]. It considered as one of the the busiest and most efficient airports in the world. We are interested in understanding and predicting the flight delays for Atlanta airport and considering the following research objectives. The main objectives of this research are to

1) Conduct exploratory data analysis to identify contribution of temporal factors such as time of the day, day of the week and week of the month towards flight delays.

2) Understand the severely affected airlines and routes originating or departing from Atlanta International airport.

3) Predict for the future, if a flight would be delayed or not.

The rest of this paper is structured as follows. Literature review section presents the related works of researchers done in the airline delay prediction. It is followed by the Materials and Methodology section where the datasets used in this study are explained. In addition to that the data pre-processing steps, different data models employed for prediction of airline delay and the evaluation metrics are also discussed. The Results section details the exploratory data analysis and airline delay prediction results from the data models experimented in this study. The literature is concluded with the conclusion section where the key observation from this study is presented along with the limitations and future scope for this subject.

2 Literature Review

Due to the availability of data and importance of the subject, there are several studies that addressed the airline delays and root causes using data mining and machine learning techniques.

Assent et. al [3] conducted a study for the classification of flights into delay categories based on the archived data at major airports in current flight information systems.They have proposed an efficient and effective algorithm for subspace classification for airline delay. Since the large number of attributes might occlude the dominant patterns of flight delays, and globally dimensionality reduction seem to be inappropriate, locally relevant attributes selection method is employed [3]. Another study by N. Etarni [4] developed a predictive model for on-time arrival flight of airliner by discovering correlation between flight and weather data. In this study, supervised machine techniques like Random Forest and Gradient Boosting Machines were applied with and without weather data to predict flight delays.It is concluded that Random Forest Classifier with weather data performed the best with 77.00 percentage accuracy to predict flight delay [4]. Another research by A. Aljubairy et. al [5] studied the flight delays from a new angle by utilising data generated from the emerging Internet of Things (IoT). A framework that aims at improving the flight delay problem is presented. Feature selection by employing correlation and modeling with Multiple Logistic Regression algorithm resulted in 85.74 percentage accuracy [5].

Chakrabarty et al. proposed a model using Gradient Boosting Classifier for predicting flight arrival delay of American Airlines covering 5 most busiest airports in US [6]. In this study, 'Randomized Synthetic Minority Over-Sampling Technique', known as SMOTE, is applied for data balancing and hyper-parameter tuning with grid search is used. This approach achieved a flight delay prediction accuracy of 85.73 percentage [6]. Another study by B. Ye et. al researched the airline delay interactions between time, flight plan and previous delay [7]. It investigated four popular supervised learning methods: multiple linear regression, a support vector machine, extremely randomized trees and LightGBM and recommends LightGBM model that provided the best result, giving a 86.55 accuracy with their experiments [7].

It is observed from the review of related works that Gradient Boosting, Logistic Regression and Random Forest models are performing better to predict the flight delays. Our study is different from these previous studies that we are only utilising temporal and spatial factors to determine the flight delay.

3 Materials and Methodology

3.1 Dataset

For this project, data pertaining to ATL International Airport for nonstop domestic flights is collected from Bureau of Transportation Statistics [2]. It has the information about the domestic flights departed and arrived in ATL airport during January 2019 to December 2019. Cancellation and diversion of flights are not taken into account for this study.

3.2 Data Pre-processing

The attributes present in the original dataset and its description is given in Appendix A. Since our concern is only the flight delays from operation schedules, the instances that pertains to cancellation or diversion are removed for that matter. Two delay- indicating attributes are used in the creation of the target variable. The target variable is a binary variable to indicate the occurrence of delay. The attributes used in deriving the target variable, DEP DELAY GROUP and ARR DELAY GROUP, are removed from the dataset for the modeling purposes. Also we have removed the YEAR field from the dataset as we are only considering the year 2019. Our prepared dataset has 9 independent variables and one target variable which has 2 classes.

Imbalanced Data The target variable we consider in this classification indicates whether the flight is delayed or not. If there is a delay recorded at departure or at arrival, the flight is considered as a delayed flight. This dataset has 81.39 percentage on-time arrival instances(No delay, Class = 0) and 18.61 percentage late departure and arrivals (delay, Class = 1) to and from the Atlanta airport. The target variable classes are considered imbalanced as about 82 percentage of flights are not delayed and only 18 percentage being delayed. In addition to training the models with original dataset with all data instances available, we also experimented with down sampled data for model training. To balance out the skewness in the target class, we have considered various balancing techniques for data modeling. Due to the size of the dataset and limited computing resources, we used under-sampling In the case of under-sampling, the majority class is randomly under-sampled to match the minority class. Once the dataset is split into test and train datasets in 70:30 ratio, under sampling is employed on the train dataset so that the number of records in each of the classes match. After under-sampling the train dataset has a total of 204120 records with 102060 records in each class.

Since there is no missing values or outliers present in the dataset no data imputation or outlier-handling is performed.

3.3 Methods

For the flight delay prediction problem, we have used three data modeling techniques. A brief description of the data modelling algorithm used in this project is described below.

Logistic Regression Logistic Regression examines the relationship between discrete and continuous independent variables and those which have binary result dependent variable while also considering the risk factors as the probability of occurrence of each class [9]. The core of a logistic regression model is the odds ratio: the ratio of the outcome probabilities. In case of binary classification, the odds ratio is

$$oddsratio = \frac{P(1)}{1 - P(1)} \tag{1}$$

The non-linear relationship between the inputs and the odds ratio can be flattened out to a linear relationship by computing the log of the odds ratio [13].

$$\ln \frac{P(1)}{1 - P(1)} = w_0 + w1 * x_1 + w_2 * x_2 + ... + w_n * x_n \tag{2}$$

the values w_0, w_1, and so forth are the model coefficients or weights. The coefficient w0 is the constant term in the model, sometimes called the bias term. The variables $x_1, x_2, ...x_n$ are the inputs to the model.

$$\frac{P(1)}{1-P(1)} = e^{(w_0+w1*x_1+w_2*x_2+....+w_n*x_n)} \tag{3}$$

the calculation of the probability that the outcome is equal to 1 is

$$P(1) = \frac{1}{1+e^{-(w_0+w1*x_1+w_2*x_2+....+w_n*x_n)}} \tag{4}$$

This function known as the logistic curve. In our flight delay logistic model we consider the threshold boundary as 0.5.

p $<$ 0.5 = No Delay(class 0)

p $\geq$ 0.5 = Delay(class 1)

The logistic prediction with more than 0.5 threshold probability is considered as a class of flight with delay. Otherwise, the record is marked as a flight with no delay.

k-Nearest Neighbour(kNN) Classifier kNN Classification is one of the most fundamental and simple classification methods and is commonly used when there is little or no prior knowledge about the distribution of the data. kNN is an instance based "supervised" classification method where the class labels are identified by proximity of known data points with the data point for prediction [10].

Algorithm 1 kNN Classifier

```
Input: k, traindata, testdata
Output: testoutcome
procedure kNN Classifier(k ,traindata, testdata)
    Begin: For each test tuple
        Compute the distance(Euclidean distance in this case) between test tuple and the training tuples
        Find the 'k' nearest neighbors with respect to the distance
        Use mode of k-nearest neighbors' model to estimate the class of test tuple
    End
```

The basic algorithm for kNN is given in Algorithm 1. For each test tuple, the classifier finds the distance between the training data point and test tuple. The target class mode of nearest k neighbours are used as the test outcome class label.

We have scaled the the input data for the kNN classifier for better performance. We used the Nearest neighbour k value as 5, which appear to give better results from our experiments, in our final model .

Extreme Gradient Boosting(XGBoost) It is an efficient and scalable implementation of gradient boosting framework. The main idea of boosting is to combine a series of weak classifiers with low accuracy to build a strong classifier with better classification performance [11]. XGBoost is optimized under the Gradient Boosting framework and developed by Chen and Guestrin [12]. XGBoost can automatically do parallel computation generally over 10 times faster than Gradient Boosting Machine(GBM). The basic algorithm for boosted regression trees can be generalized to a stage-wise additive model of B individual regression trees

$$f(x) = \sum_{b=1}^{B} f^b(x) \tag{5}$$

We had used RStudio for the data processing, visualization, modelling and evaluation. The Processor we used is IntelCore™ i5 7200U CPU@2.50GHz on Windows 64-bit OS with 8GB RAM. Source code is available in Supplementary materials.

3.4 Evaluation Metrics

Confusion matrix will be generated for the test data and will be utilized to evaluate the metrics for model performance. Accuracy, Sensitivity, Specificity and Balanced Accuracy are be used.

$$Accuracy = \frac{Numberofcorrectpredictions}{Numberofobservations} \tag{6}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{7}$$

$$Specificity = \frac{TN}{TN + FP} \tag{8}$$

$$BalancedAccuracy = \frac{Sensitivity + Specificity}{2} \tag{9}$$

where TP is the number of true positive predictions,
TN is the number of true negative predictions,
FP is the number of false positive predictions,
FN number of false negative predictions.

4 Results

In this section we present the results of our analysis and data modelling experiments.

4.1 Exploratory Data Analysis

We studied the data to understand the different aspects of flight operations at the ATL airport during the year 2019.

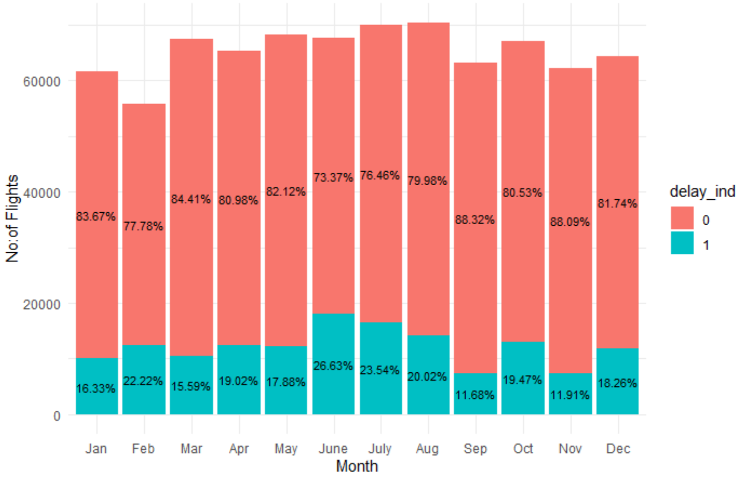

Fig. 1. Flight Operation based on Month

It is observed from the analysis and from Figure.1 that the highest percentage of flights are delayed during the month of June with

27.6 percentage delayed flight where the lowest being during September with 11.7 percentage.

More flights are getting delayed around the third week of the month relative to the other days. The analysis from Figure. 2 shows that the number of flights and the number of flights getting delayed are much less during Saturdays compared to other days of the week.

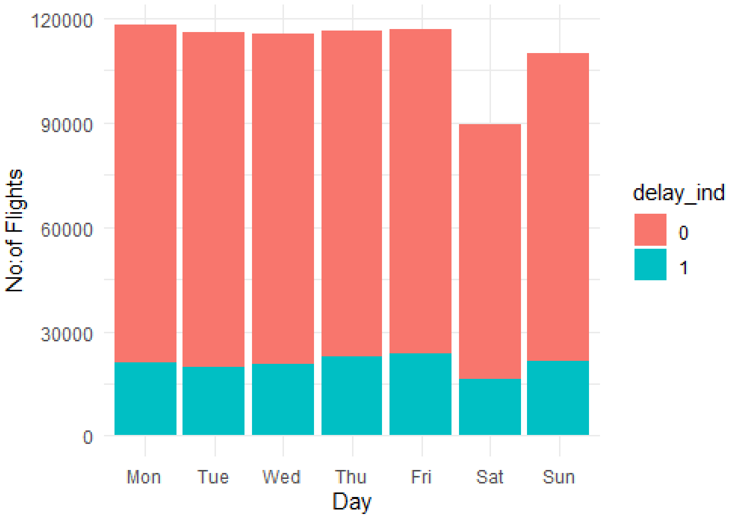

Fig. 2. Flight operation based on Day of week

During the day, it is between 5 AM - 8 AM that more flights are scheduled as well as delayed. Delta Airlines Inc recorded the most number flights, accounting over 60 percentage of flights operating in and out of Atlanta airport while Frontier Airlines Inc recorded Highest percentage of delayed flights with its 32 percentage of flights getting delayed.

The top 10 airport locations to and from which most flights are operating in and out of Atlanta airport are given in Table 1. We could observe that most flights arriving in and departing from Atlanta International airport is from New York International airport, totalling 9365 arrived and 9336 departed flights. To and from Orlando Airport to Atlanta airport is the second most busiest route and to and from to Fort Lauder-dale airport, being the third most busiest in the routes of flight operations.

Table 1. Top 10 Routes in and out of Atlanta airport

Rank	Out of ATLANTA	Towards ATLANTA
1	ATLANTA − > NEW YORK	NEW YORK− >ATLANTA
2	ATLANTA − > ORLANDO	ORLANDO− > ATLANTA
3	ATLANTA − > FORT LAUDERDALE	FORT LAUDERDALE− > ATLANTA
4	ATLANTA − > CHICAGO	CHICAGO− > ATLANTA
5	ATLANTA − > TAMPA	TAMPA− > ATLANTA
6	ATLANTA − > WASHINGTON DC	WASHINGTON DC− > ATLANTA
7	ATLANTA − > MIAMI	MIAMI− > ATLANTA
8	ATLANTA − > BALTIMORE	DALLAS− > ATLANTA
9	ATLANTA − > DALLAS	BALTIMORE− > ATLANTA
10	ATLANTA − > NEWARK	NEWARK− > ATLANTA

Table 2 below shows the routes of flights operations in and out of Atlanta airport that experienced departure or arrival delays. It can be seen that about 38.80% of flights departing from ATL airport to Islip airport are being delayed, which contributes to the highest rate. In addition, even though Newark airport finds the tenth position in the number of flights operating in and out of ATL airport, 34.60% of flights arriving to Newark airport are delayed.

About 61.50% of total flights operating from Missoula airport and Reno airport are being delayed when they arrive at ATL airport.

Table 2. Top 10 Routes with delay in and out of Atlanta airport

Rank	Route with Departure delay at ATL	% delayed	Route with Arrival delay at ATL	% delayed
1	ATLANTA − > ISLIP	38.8	MISSOULA − > ATLANTA	61.5
2	ATLANTA − > NEWARK	34.6	RENO− > ATLANTA	61.5
3	ATLANTA − > ANCHORAGE	33.9	PALP SPRINGS− > ATLANTA	54.6
4	ATLANTA − > TRENTON	33.4	FARGO− > ATLANTA	47.8
5	ATLANTA − > OAKLAND	32.8	ANCHORAGE− > ATLANTA	37.5
6	ATLANTA − > SAN FRANCISCO	31.9	BURBANK− > ATLANTA	35.8
7	ATLANTA − > RAPID CITY	30.8	ISLIP− > ATLANTA	32.3
8	ATLANTA − > COLORADO SPRINGS	29.0	ASPEN− > ATLANTA	32.3
9	ATLANTA − > NEW YORK	28.7	CHRISTIANSTE− > ATLANTA	31.5
10	ATLANTA − > LOS ANGELES	28.7	RAPID CITY− > ATLANTA	30.8

The below figures 3 and 4 illustrate the origin and destination locations which have most delays during year 2019 operating in and out of Atlanta airport.

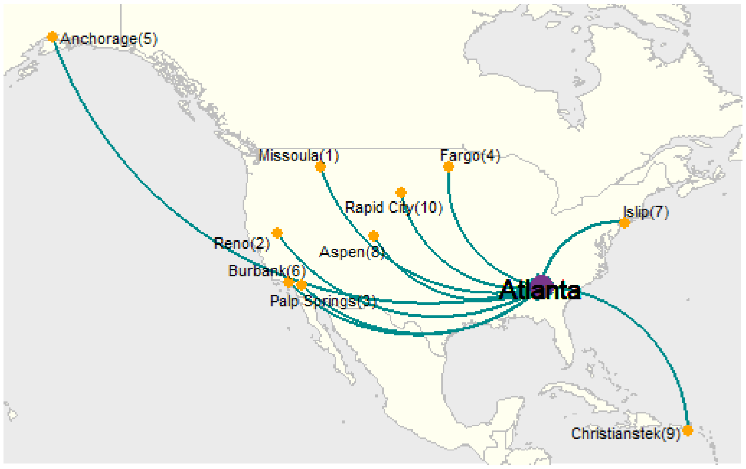

Fig. 3. Top ten routes origination with delay arriving at Atlanta

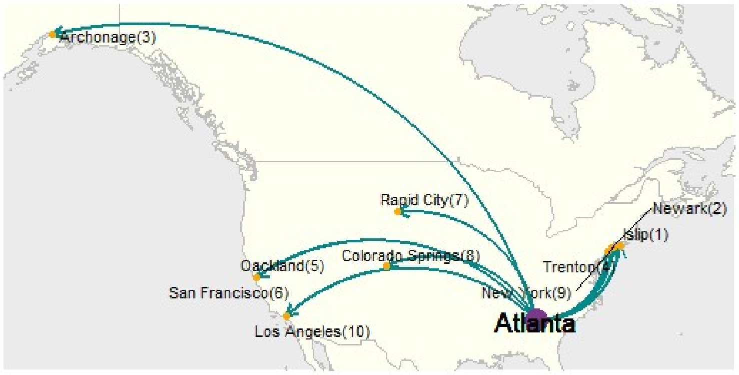

Fig. 4. Top ten routes destination with delay departing from Atlanta

Table 3 below shows the routes of flights operations in and out of Atlanta airport that consistently operated on-time as scheduled on departure and arrival. It can be inferred that about 91.30% of flights operating from Atlanta airport reached Fargo Airport without any delay.Also, about 89.50% of flights departed to Montrose airport arrived ontime.

All flights departed from Elmira airport reached ATL airport on-time. About 91.60% of flights operated from Tucson airport reached ATL airport, experiencing no delay.

Table 3. Top 10 Routes with On-Time in and out of Atlanta airport

Rank	Routes with Ontime Departure from Atlanta	% Ontime	Routes with Ontime Arrival at Atlanta	% Ontime
1	ATLANTA − > FARGO	91.3	ELMIRA/CORNING− > ATLANTA	100
2	ATLANTA − > MONTROSE	89.5	TUCSON− > ATLANTA	91.6
3	ATLANTA − > EAGLE COUNTY	88.9	SCRANTON/WILKES-BARR− > ATLANTAE	90.9
4	ATLANTA − > DAYTON	88.1	GREENSBORO− > ATLANTA	90.6
5	ATLANTA − > RENO	88.0	CHARLESTON− > ATLANTA	90.3
6	ATLANTA − > HARRISBURG	87.7	FLINT− > ATLANTA	90.3
7	ATLANTA − > GREENSBORO	87.6	JACKSON− > ATLANTA	90.1
8	ATLANTA − > JACKSON	87.6	MYRTLE BEACH− > ATLANTA	90
9	ATLANTA − > VALPARAISO	87.6	MANCHESTER− > ATLANTA	89.9
10	ATLANTA − > NEWPORT NEWS	87.6	NORFOLK− > ATLANTA	89.6

4.2 Flight delay Prediction

We developed three models for classifying whether a particular flight operation would get delayed or not. We employed Logistic Regression, kNN Classifier and XGBoost with undersampling and the evaluation metrics are shown in 4. Also, the models are trained without any under-sampling and the results shows that except for Logistic Regression the overall accuracy of the classifier is better when no under-sampling is employed.

Classifier performance with under-sampled training Table.4 below summarises the performance of the classifiers evaluated on test data after training the model with under-sampled training data.

Table 4. Evaluation metrics for Classifiers with under-sampling

Evaluation metric	Logistic Regression	kNN Classifier	XGBoost
Accuracy	0.5818	0.5840	0.6920
95 perc CI	(0.5798, 0.5838)	(0.5820, 0.5860)	(0.6901, 0.6939)
Sensitivity	0.5732	0.5793	0.6990
Specificity	0.6192	0.6047	0.6612
Pos Pred Value	0.8681	0.8650	0.9002
Neg Pred Value	0.2491	0.2474	0.3344
Balanced Accuracy	0.5962	0.5920	0.6801

With Logistic Regression model we had achieved a model accuracy of 58.18% .As can be seen from the metrics, the minority class prediction percentage is low at 26.28 percentage. With the balanced accuracy at 59.62 the model is not much promising. With kNN classifier, using k=5, an accuracy of 58.40 % is achieved. With respective to sensitivity, specificity and balanced accuracy metrics, not much improvement is observed compared to the logistic model in this case. XGBoost outperformed both other classifiers with 69.20% accuracy rate. Specificity, sensitivity and negative predictive power is also much better compared to other classifiers.

Classifier performance without under-sampled training We also trained our classifier with all of the training data i.e. with 70% from original dataset, and we observed that both kNN and XGBoost gave better results than previous down-sampled version.

Table 5. Evaluation Metrics for Classifiers without under-sampling

Evaluation Metric	Logistic Regression	kNN Classifier	XGBoost
Accuracy	0.8139	0.7921	0.8345
95 perc CI	(0.8123, 0.8155)	(0.7905, 0.7938)	(0.833, 0.836)
Sensitivity	1.0000	0.9361	0.9696
Specificity	0.0000	0.1625	0.2439
Pos Pred Value	0.8139	0.8302	0.8487
Neg Pred Value	0	0.3677	0.6472
Balanced Accuracy	0.5000	0.5493	0.6067

Logistic Regression model, as expected, resulted in a highly biased model and became no better than random guess with zero negative class predictive power. kNN classifier produced an accuracy of 79.21% which is much improved than the previous case. XGBoost with 1000 iterations was able to produce an accuracy of 83.45% which is the best of all models we evaluated in this study. The balanced accuracy value of 60.67% also show the superiority of this model in predicting airline delay.

But considering all the evaluation metrics to measure the performance it is understood that the classifiers performed better with the imbalanced data treatment with under-sampling. The specificity of the model is an important metric here due to the nature of data and the problem under study. The minority class here, indicating the presence of delay, needs to be predicted with better which is of significance in this model. So, we suggest the use of XGBoost Classifier with under-sampling to achieve this prediction.

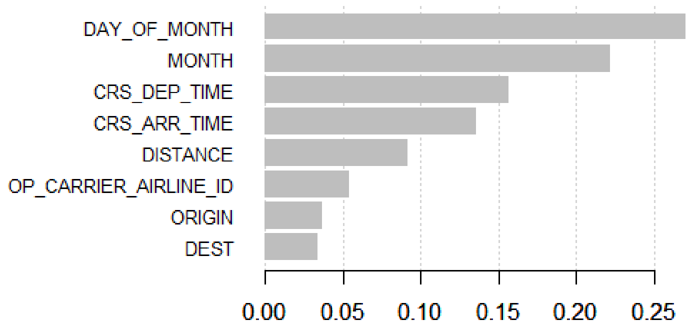

Fig. 5. Relative feature importance based on XGBoost Model

Upon analysing the attributes that can better classify the flight delay, we observed that the temporal factors such as the day of the month, month, departure and arrival time have the most influence on having a delay. The location and distance of the routes seem less relevant to the model in predicting the delay.

5 Conclusion

This study analysed the data from Atlanta airport for the year 2019 and identified the major temporal aspects that influenced the airline departure and arrival delay using data aggregation and visualization techniques. Frontier airlines recorded most delays recorded at Atlanta airport while most operations are towards New York and Orlando. Saturdays seem to have less flights and delys while its during 5:00 AM and 8:00 AM timeframe that most flights are scheduled and also delayed. Out of the three classifiers trained to predict the delay, XGBoost achieved a overall test accuracy of 83.45 percentage which is promising, but we suggest the model trained with under-sampled data because of its better specificity and balanced accuracy rate. Further improvement can be made to the model by optimizing the hyper-parameters of the XGBoost model by grid search. The limited resources posed a hindrance to achieve this further step in this study. Within the limitations the study was successful in understanding the temporal characteristics of flight delay and also predicting the delay Atlanta airport.

References

1. Statistics — ATL — Hartsfield-Jackson Atlanta International Airport. (2018). http://www.atl.com/business-information/statistics/
2. “OST R — BTS — Transtats.” https://www.transtats.bts.gov/TableInfo.asp (accessed Oct. 19, 2020).
3. Assent I., Krieger R., Welter P., Herbers J., Seidl T. (2009) Data Mining For Robust Flight Scheduling. In: Cao L., Yu P.S., Zhang C., Zhang H. (eds) Data Mining for Business Applications. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-79420-4 19
4. Etani, N. Development of a predictive model for on-time arrival flight of airliner by discovering correlation between flight and weather data. J Big Data 6, 85 (2019). https://doi.org/10.1186/s40537-019-0251-y
5. Aljubairy, A., Zhang, W.E., Shemshadi, A. et al. A system for effectively predicting flight delays based on IoT data. Computing 102, 2025–2048 (2020). https://doi.org/10.1007/s00607-020-00794-w
6. N. Chakrabarty, “A data mining approach to flight arrival delay prediction for American airlines,” 2019. doi: 10.1109/IEMECONX.2019.8876970.
7. B. Ye, B. Liu, Y. Tian, and L. Wan, “A methodology for predicting aggregate flight departure delays in airports based on supervised learning,” Sustainability (Switzerland), vol. 12, no. 7, 2020, doi: 10.3390/su12072749.
8. M. Robnik-ˇ Sikonja and I. Kononenko, “Theoretical and Empirical Analysis of ReliefF and RReliefF,” 2003.
9. Ranganathan, P., Pramesh, C., Aggarwal, R. (2017). Common pitfalls in statistical analysis: Logistic regression. Perspectives in Clinical Research, 8(3), 148151. https://doi.org/10.4103/picr.PICR8717
10. Peterson, L. (2009). K-nearest neighbor. Scholarpedia, 4(2), 1883. https://doi.org/10.4249/scholarpedia.1883
11. Chen, T., Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 13-17-August-2016, 785–794. https://doi.org/10.1145/2939672.2939785
12. Chen, M., Liu, Q., Chen, S., Liu, Y., Zhang, C. H., Liu, R. (2019). XGBoost-Based Algorithm Interpretation and Application on Post-Fault Transient Stability Status Prediction of Power System. IEEE Access, 7, 13149–13158. https://doi.org/10.1109/ACCESS.2019.2893448

13. Abbott, D. Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst. 2014

1 Appendix A: Dataset Field Description

Table 6: Attributes in the dataset.

Attribute	**Type**	**Description**	**Values**
Attribute	Type	Description	Values
YEAR	Categorical	Year of Flight Operation	2019
MONTH	Categorical	Month of Flight Operation	1-12
DAY OF MONTH	Categorical	Day of the month of Flight Operation	0-31
DAY OF WEEK	Categorical	Day of the month of Flight Operation	1: Monday 2: Tuesday 3: Wednesday 4: Thursday 5: Friday 6: Saturday 7: Sunday 9: Unknown

(To be continued)

Attribute	**Type**	**Description**	**Values**
OP CARRIER AIRLINE ID	Categorical	Unique Airline Carrier Code	19393: Southwest Airlines Co. 19690: Hawaiian Airlines Inc. 19790: Delta Air Lines Inc. 19805: American Airlines Inc. 19930: Alaska Airlines Inc. 19977: United Air Lines Inc. 20304: SkyWest Airlines Inc. 20363: Endeavor Air Inc. 20366: ExpressJet Airlines 20368: Allegiant Air 20378: Mesa Airlines Inc. 20397: PSA Airlines Inc. 20398: Envoy Air 20409: JetBlue Airways 20416: Spirit Air Lines 20436: Frontier Airlines Inc. 20452: Republic Airline
ORIGIN	Categorical	Origin Airport Code	169 Distinct Airport Codes. Ex: DFW, ATL
DEST	Categorical	Destination Airport Code	169 Distinct Airport Codes. Ex: LAX, ATL
CRS DEP TIME	Time	Scheduled Departure time	Time of day in 24 Hr Format Ex: 2210

(To be continued)

Attribute	**Type**	**Description**	**Values**
DEP DELAY GROUP	Categorical	Departure delay group	-2: Delay less than -15 minutes -1:Delay between -15 and -1min 0:Delay between 0 and 14min 1:Delay between 15 to 29min 2:Delay between 30 to 44min 3:Delay between 45 to 59min 4:Delay between 60 to 74min 5:Delay between 75 to 89min 6:Delay between 90 to 104min 7:Delay between 105 to 119min 8:Delay between 120 to 134min 9:Delay between 135 to 149min 10:Delay between 150 to 164min 11:Delay between 165 to 179min 12:Delay greater than 18minutes
CRS ARR TIME	Time	Scheduled Arrival time	Time of day in 24 Hr Format Ex: 2210
ARR DELAY GROUP	Categorical	Arrival delay group	Same as DEP DELAY GROUP

(To be continued)

Attribute	Type	Description	Values
DISTANCE	Numerical	Distance between airports in miles	Range of values: 83 to 4502 miles.

2 Appendix B: Source Code

https://github.com/tressythomas/DM-Group-Project