Critical Analysis on Flight Cancellations and Predictive Analysis on Flight Delays using Automated Machine Learning

MSc Research Project
Data Analytics

Nikhil Salvi
Student ID: X20179529

School of Computing
National College of Ireland

Supervisor: Dr. Christian Horn

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Nikhil Salvi

Student ID: X20179529

Programme: MSc. Data Analytics **Year:** 2022-23

Module: Research Project

Supervisor **:** Dr. Christian Horn

Submission Due Date: 15th December 2022

Project Title: Critical analysis on flight cancellations and predictive analysis on flight delays using automated machine learning

Word Count: 4873 **Page Count:** 19

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Nikhil Salvi

Date: 15th December 2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Critical Analysis on Flight Cancellations and Predictive Analysis on Flight Delays using Automated Machine Learning

Nikhil Salvi

X20179529

Abstract

In the 21st century, every nation needs to grow in the global network and get connected to other countries, and for this purpose, the aviation industry plays a vital role. The aviation or airline industry is responsible for rapid transportation worldwide. This industry generates economic growth, cre- ates employment, and provides facilities. As the number of passengers and air traffic is increasing, the airline industries are adopting advanced technologies to make the processes to make the process quick and easy. But, with this, a tremendous amount of data gets generated. Industries are using this data to understand the business more and find out areas of improvement to provide the best services to their customers. This helps the industry to sustain itself in a competitive environment and grow its business. In this research project, a critical analysis of flight delay is done. Also, a predictive model is built using automated machine learning. This model has provided the best suitable algorithm to predict flight delays accurately so that the air- line industries strategize and prepare the airport management system.

1 Introduction

1.1 Background

Nowadays using various sources, big data is collected to make better analyses for businesses in every industry. Notably, as the number of passengers and travelers is proliferating, and as the airline industries are adopting advanced technologies such as self-check-in and intelligent check-in, a tremendous amount of data gets generated. Also, with new technologies such as ADS-D (Adaptive Dependent Surveillance-Broadcast), real-time data of the operating flights gets consistently collected. With the help of big data, in this 21st century, industries are becoming competitive rapidly. To grow the business industries, provide the best services to their customers, and identify the challenges and areas of improvement, big data is used on a large scale. This project uses data from domestic flights operated in the United States during 2015 to build analytical and predictive models. The data source is the U.S. Department of Transportation's (D.O.T.) Bureau of Transportation Statistics. But, for this study, the data set is collected from Kaggle.com, an open-source platform. Using this dataset, further steps are taken to critically analyze the data to build interpretations to understand the flight delays and cancellations patterns. Flight delays and cancellations cause substantial economic losses to airline industries. One flight delay causes a significant impact on the number of stakeholders. It also causes the delay of flights scheduled after that flight, which eventually causes more fuel consumption than expected. Hence, this project creates a predictive model using automated machine learning to determine which machine learning algorithm is best depending on the evaluation matrices and business requirements.

1.2 Research objectives

- To critically analyse the flight cancellations.
- To create predictive model to predict whether a flight will be delayed.

1.3 Research questions

Depending upon the business requirement and objective, which supervised leering machine algorithm provides the best results to predict delayed flights accurately? The business objective may focus on correctly predicted delayed flights out of actually delayed flights, or the overall accuracy of both classes, flights that are delayed, or flights that are not delayed.

2 Related Work

2.1 Data gathering

As air traffic rapidly increases, big data is collected on a large scale. With this big data, many machine learning and deep learning models are built on understand- ing the data precisely and building predictive analysis to improve airline services. As the air traffic gets dense and as a result, it creates more and more extensive data, and it gets challenging to build a machine learning or deep learning models by using traditional technologies such as primary and secondary surveillance radar. Therefore, in the research paper (Nijsure et al. 2015), new technologies such as automatic dependent surveillance-broadcast (ADS-B) are discussed. In this technology, a hardware setup is installed. This technology allows us to gather real-time flight data, such as the angle of arrival, latitude, longitude, flight speed, etc. To build machine learning predictive models, the quality of the data to be trained is essential. Therefore, as the ADS-B technology is accurate, it will help develop more accurate machine learning and deep learning models.

Between 2009 and 2015, 22% of domestic flights were delayed by more than 15 minutes in Brazil. In the research paper proposed by (Moreira et al. 2018), the researchers have used data transformation techniques such as normalization, smoothing, and data balancing techniques such as SMOTE. With these techniques, the researchers have tackled the challenge of data imbalance. As a result, the model built with balanced data got a 60% hit.

2.2 Supervised machine learning techniques to predict flight delays

In the research paper proposed by (Gui et al. 2020), the researchers have used the ADS-B system, discussed before in the previous section of related work in this study, to gather the data. The researchers have used random forest and long short-term memory algorithms to build predictive models. The random forest algorithm has been proven to improve by overcoming the challenge of overfitting and achieving an accuracy of 90%.

In the research paper proposed by (Murca & Hansman 2019), a machine learning model experiments on National Airspace System. A random forest is used to build the model. The accuracy for a short-term forecast of 1 hour is 83%, that for a 3-hour forecast is 63%, whereas the accuracy for more than 3 hours is 52%. The predictive model can be improved by gathering more accurate weather data.

In the research paper proposed by (Manna et al. 2017), a gradient-boosting algorithm is used to build a model to predict flight delays. The gradient boosting is built on the top of the decision tree algorithm. The model is run differently for arrival and departure cases. For arrival, the accuracy achieved is 92.31%, whereas, for departure, it is 94.85%. But the training data had only 70 airports. The model can be improved with more historical data for training.

Weather data plays a crucial role while building machine learning models. In the research paper proposed by (Choi et al. 2016), the researchers have used weather data to build predictive models. Four models are created using the decision tree, random forest, adaptive boosting, and k-nearest neighbour algorithms. Among these four models, the random forest has outperformed all the other algorithms by achieving an accuracy of 83.40%. Since the model uses weather data, uncertainty in weather forecasts would hamper the model's accuracy.

Another machine learning model is built in the research proposed by (Hu et al. 2021). A random forest algorithm was used to create the model. The model experimented on the data of Guangzhou Baiyun International Airport. The dataset contained the records of punctual flights, delayed flights, and the flights that arrived before the arrival time. The accuracy of this model was 90%. To improve this model, more variables such as airspace capacity data, airport capacity data, and airport taxi data can be considered.

2.3 Deep learning techniques to predict flight delays

In the research paper proposed by (Lv et al. 2014), the sparse autoencoder deep learning method is used to build and predict air traffic patterns. In the prediction layer, the logistic regression layer is used. The accuracy of this model ranged from 85% to 90%. However, as only logistic regression is used to build the model, the accuracy can be improved using more robust layers to create a deep learning model.

A deep learning method is used in the research paper proposed by (Pamplona et al. 2018). Using an artificial neural network, a predictive model is built to achieve an accuracy of 90%. As per this research, the three crucial components in this experiment are the day of the week, block hour, and the airline. The model can be improved if meteorological data is used to train the model.

Deep learning has achieved significant success. The research paper proposed by (Kim et al. 2016) creates the long-short-term RNN architecture with multiple models combined. The accuracy of the model ranged from 85% to 87%.

Until now, the predictive model of supervised machine learning and deep learning has been discussed. But, in the research paper proposed by (Xing & Tang 2016), the researchers have used a genetic algorithm to optimize the air traf- fic flow. By doing this, the flight delays are distributed so that the total delay time is reduced for all the flights. This model was examined on the data collected from a western airport in China.

Some factors to predict flight delays are significant, while some are minor. In the research paper proposed by (Dhanawade et al. 2019), such factors are ana- lyzed. This experiment is done on domestic and international flights in India, and the data contains 14 different airlines. In India, a growth of 10.76% is observed in domestic passengers and 8.32% in international passengers during the period of 2007-08 to 2017-18. The critical analysis built interpretations such as, for some airlines, flights delay are less, while for others, it is more. Therefore, as per this research, the airline is essential in building a predictive model.

Another predictive model created using a long short-term memory network is discussed in the research paper (Jiang et al. 2020). Tensorflow and Keras are used to build this model. The error of this model is 6.76%. The model can be improved if more data is used to train the model.

Weather, traffic intensity, etc., are the significant factors affecting flight delays. In the research paper proposed by (Demir & Demir 2017), the data is gathered using some installed sensors at airports. Also, information about flight schedules is used along with the data collected from sensors. An artificial neural network is used to create a predictive model, and the accuracy of this model is 96%.

The delay in one flight causes delays in the flights scheduled after that flight. The researchers analyzed significant factors affecting flight delays in the research paper proposed by (Gao et al. 2015). For this research, data is collected from Beijing capital international airport, and a neural network model is trained. The error of this model is 25%, which is an error of 10 minutes. This model can experiment on data collected from other airports.

3 Research Methodology, Implementation and Evaluation

To achieve the objective of this study, the CRISP-DM method will be followed. There are six different stages in CRISP-DM, and those are explained below:

3.1 Data Understanding

The Bureau of Transportation Statistics is one of the principal federal statistical agencies. The U.S. Department of Transportation's (D.O.T.) Bureau of Transportation Statistics tracks the on-time performance of domestic flights operated by various airlines in the U.S. The data used for this study is acquired from Kaggle.com[1]. There are three datasets in the data, which are airlines.csv, air- ports.csv, and flights.csv. The airlines.csv contains the IATA (International Air Transport Association's Location Identifier) numbers of all the airlines operating during 2015. There is a total of 14 airlines in the data. The airports.csv contains the data of all the airports operating in 2015 in the U.S.A., and the data also includes the latitudes and longitudes of the respective airports. Using this data and the GeoPandas python library, actual locations can be plotted on the map of the United States. In the analysis part ahead, these operations are performed. The flights.csv file contains the data crucial for this study, including the data of all the flights operated or cancelled during 2015. This research paper will use this data with the airports.csv data to build analysis and predictive models.

[1] https://www.kaggle.com/datasets/usdot/flight-delays?select=flights.csv

3.2 Data Pre-processing and Preparation

The original data of flights contains more than 5 million records. This is population data. In this project, the sample of this data will be used to build an analysis and a predictive model. The sampling of the data has been done randomly. Below, figure 1 is the distribution of population data. The X-axis contains the month, and the Y-axis represents the total number of flights in the particular month.

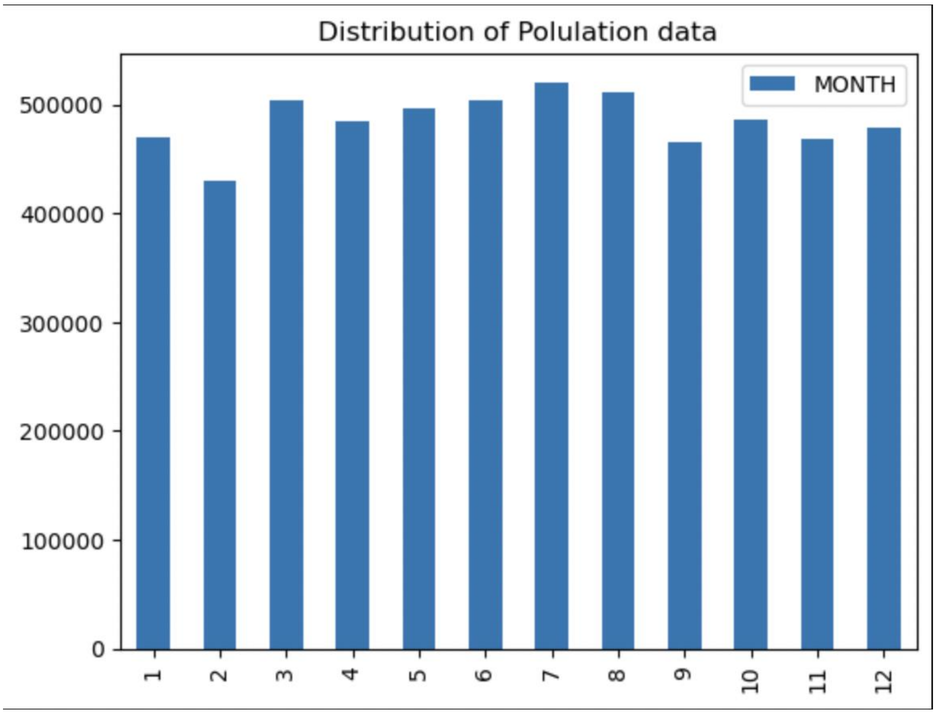


Figure 1: Month-wise distribution of population data

Figure 2 is the distribution of sample data. The distribution of the total number of flights in the months is similar to that of population data. Hence, it can be considered that the sample data represents the population data. Therefore, this sample data is used further in this study. The sample data is named flights sample.csv and contains more than 50 thousand records.

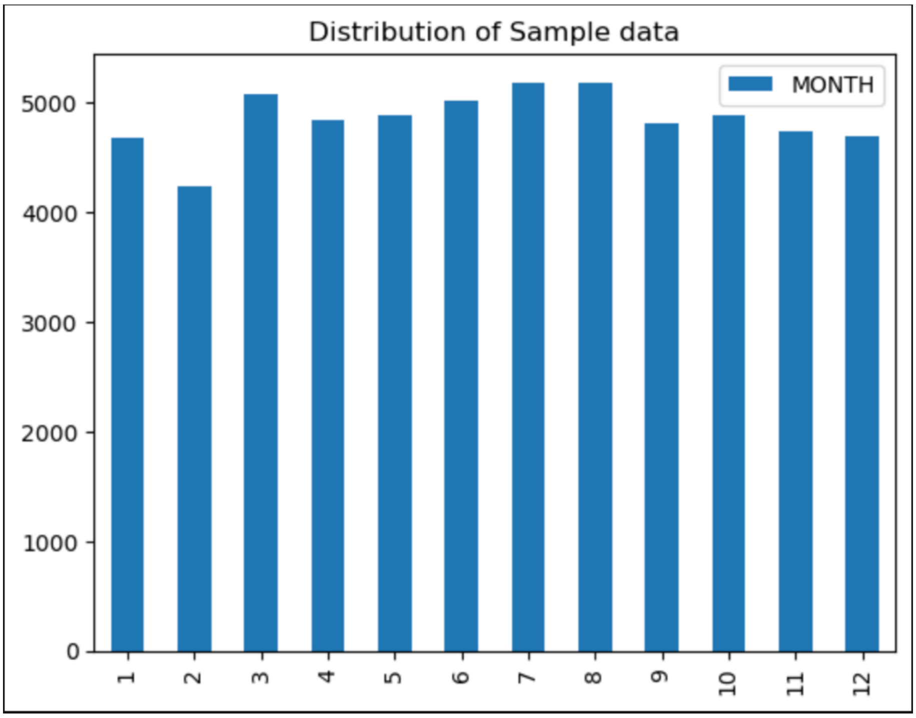


Figure 2: Month-wise distribution of sample data

The airports.csv and flight sample.csv files are imported into a data frame using panda's library. The airports.csv file contains seven variables: IATA code, Airport, City, State, Country and Latitude, and longitude of the location of the airports. The flight sample.csv file includes a column named ”Origin airport,” which contains the IATA codes of the airports from where the flights took off. With this column, the airport.csv data has been merged with the left join with flights sample.csv data to create a master data. The master data contains the record of all the flights and the latitude and longitude of the airports from which the flights were taken off.

3.3 Exploratory Data Analysis

Using the GeoPandas module, the inbuilt map of North America has been im- ported. Using the data of latitude and longitudes of airports in airports.csv, the actual locations of the airports are plotted. Figure 3 shows the areas of the airports operational during 2015.

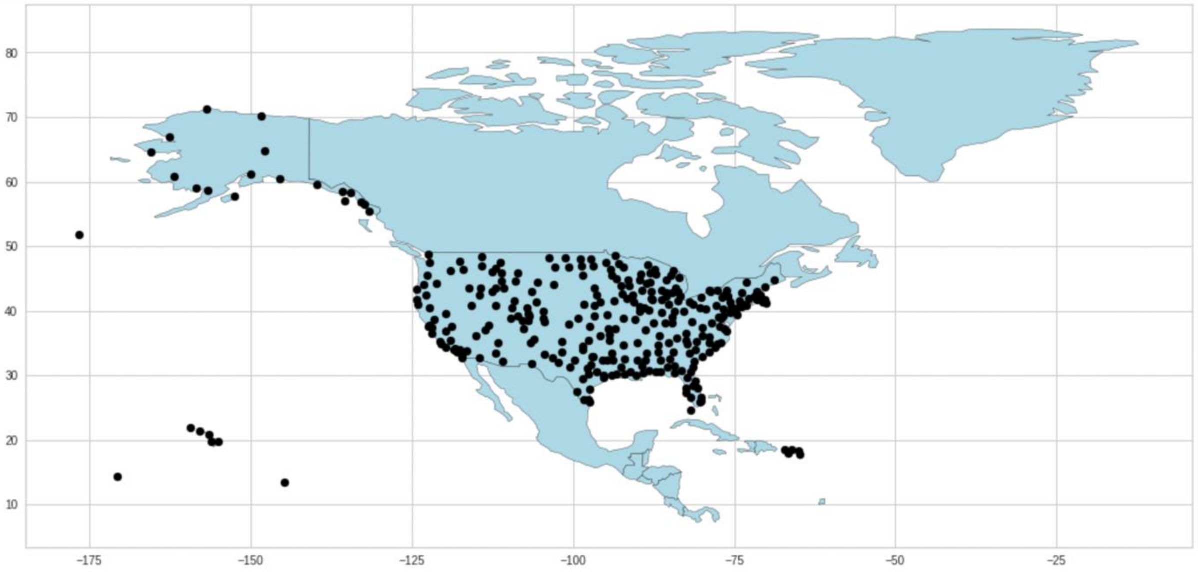


Figure 3: Plots of airport locations using Geopandas

When the data was analysed, it was found that 871 flights were cancelled. Figure 4 below is the heatmap of the null values. The blank spaces in the heatmap represent the missing data.

The data on arrival and departure times are missing from the records for which the flights were cancelled. Also, if the missing records are dropped, vital information about canceled flights will be lost. Therefore, in this study, a separate analysis of cancelled flights is done further.

3.4 Analysis of cancelled flights

A separate data frame of canceled flights has been formed to analyze the canceled flights. This data frame is named ”month cancelled” and contains 882 records. Figure 5 below represents the bar plot of the total number of flights cancelled in 2015.

```
<matplotlib.axes._subplots.AxesSubplot at 0x7efdaf11e1c0>
```

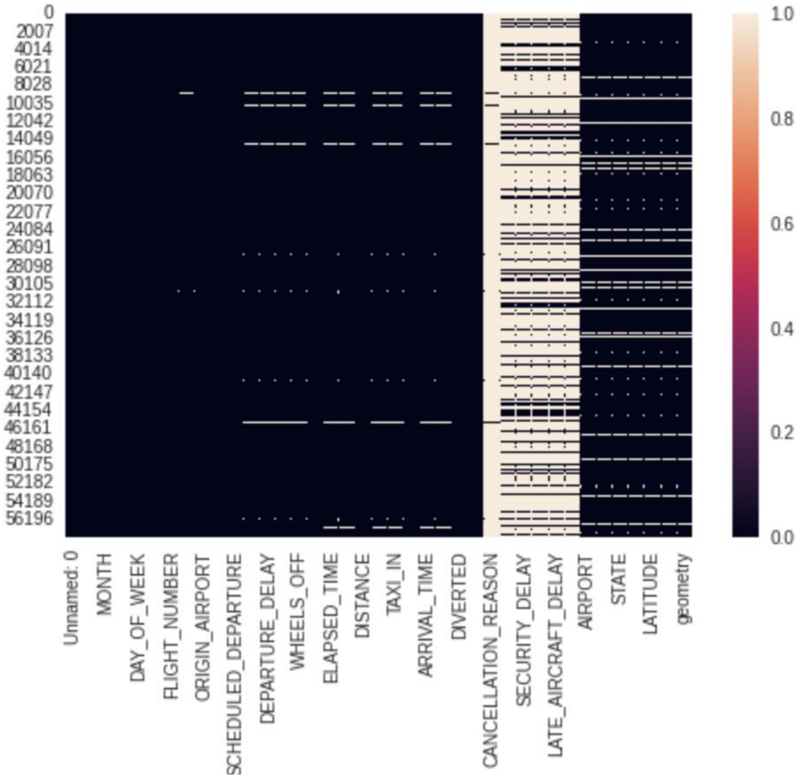
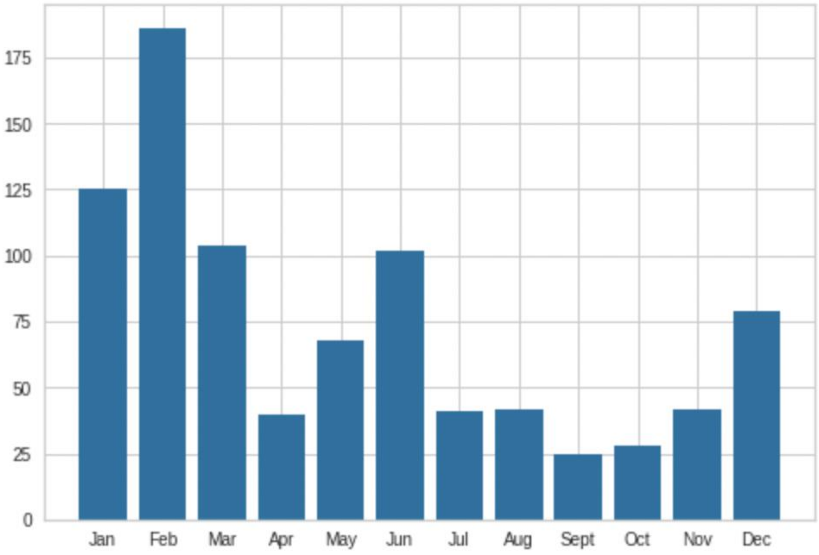


Figure 4: Heatmap of null values



```
<Figure size 7200x1440 with 0 Axes>
```

Figure 5: Month-wise distribution of number of cancelled flights

As can be seen, most flights were cancelled during January, February, and March. This is winter in the United States of America. Figure 6 below shows the locations of the airports where the flights were cancelled in January 2015.

3.4.1 January 2015

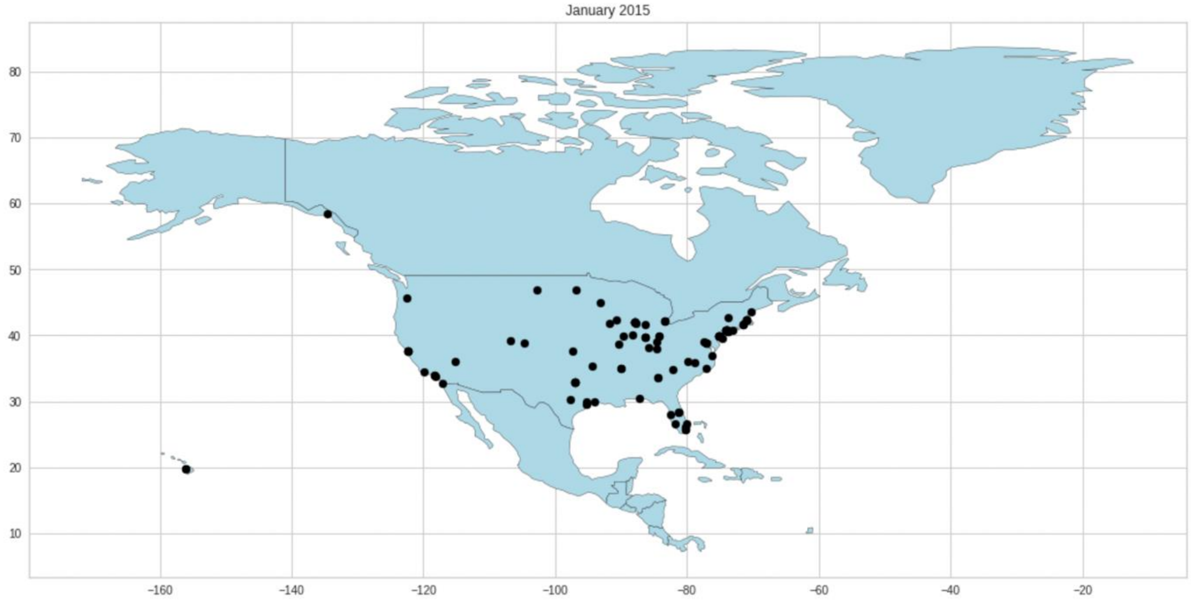


Figure 6: GeoPandas plot of January 2015

As can be seen in the figure 6, the maximum number of flights were cancelled on the United States' east coast. Also, if the below figure 7 is analysed, many flights were cancelled during the last week in January 2015.

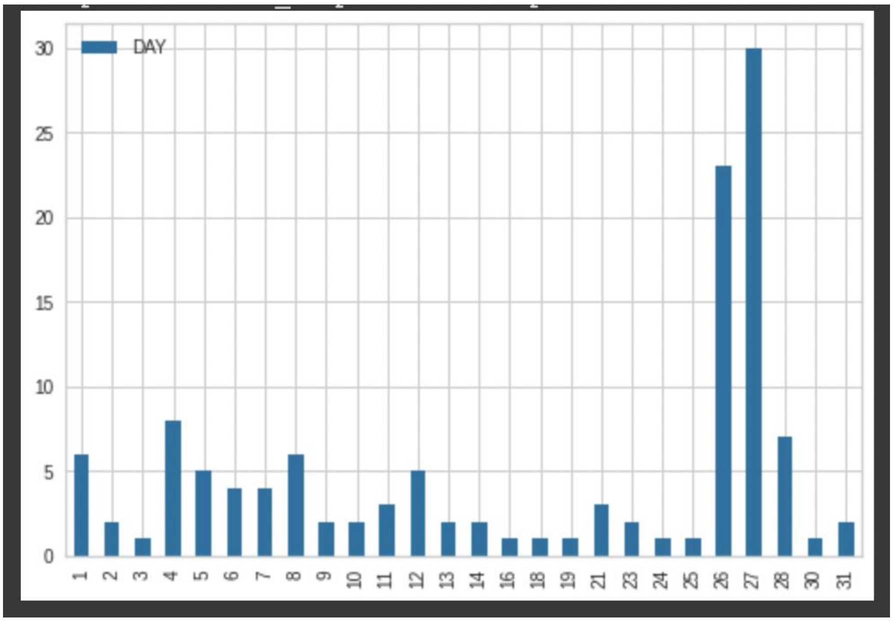


Figure 7: Day-wise distribution of number of cancelled flights in January 2015

3.4.2 February 2015

Figure 8 below shows the locations of the airports at which flights were cancelled during February. As can be seen, again, the maximum number of flights were cancelled on the east coast of the U.S.A.; in fact, the number has increased, and with this, the cancellation of the flights had been raised in the middle part of the U.S.A. The bar plot below shows the number of flights cancelled on each day of the month.

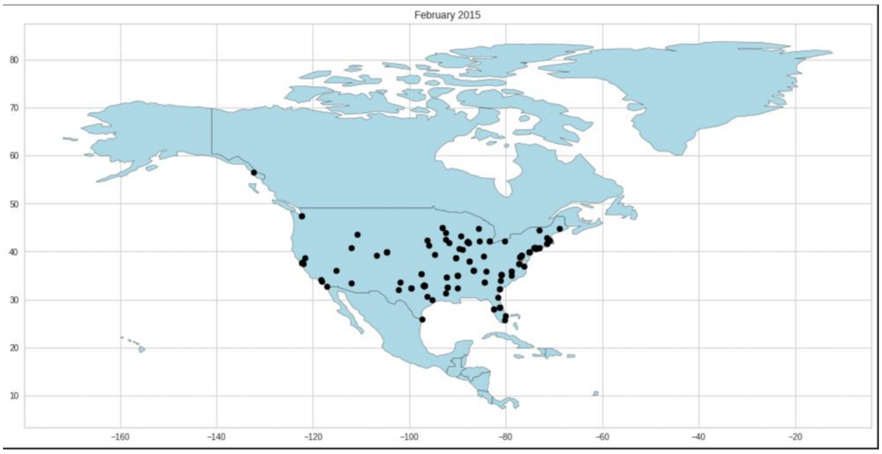


Figure 8: GeoPandas plot of February 2015

The maximum number of flights were cancelled on 2nd February. As per the article in ”BW Businessworld,” thousands of flights were cancelled in New York City during the first week of February. This was caused by the Winter Juno Storm, which was moved in during those days in February. New York, New Jersey, and Connecticut were severely affected by this storm.

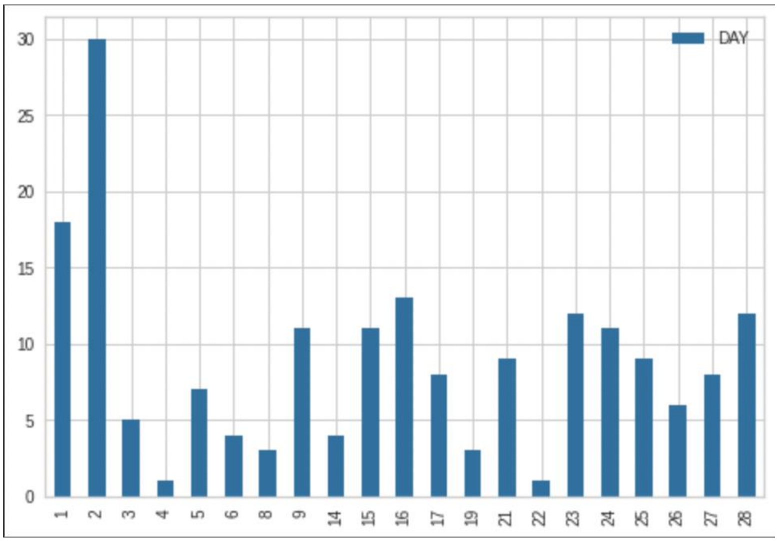


Figure 9: Day-wise distribution of number of cancelled flights

3.4.3 March 2015

The figure below shows the airports where flights were cancelled during March. It can be interpreted that the number of flights cancelled has significantly dropped. Lesser number of flights were cancelled on the eastern coast. Figure 12 is the bar plot of several flights canceled each day in March. On 5th March, this number was at its peak. This was again the effect of storms and bad weather in the continent. As per the article published by the U.S.A. today, 21% of all flights were cancelled during the first week of March.

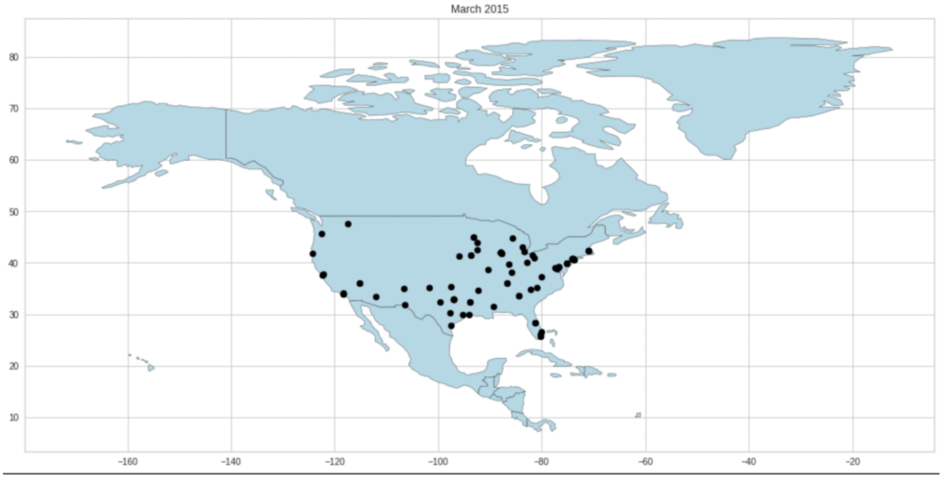


Figure 10: GeoPandas plot of March 2015

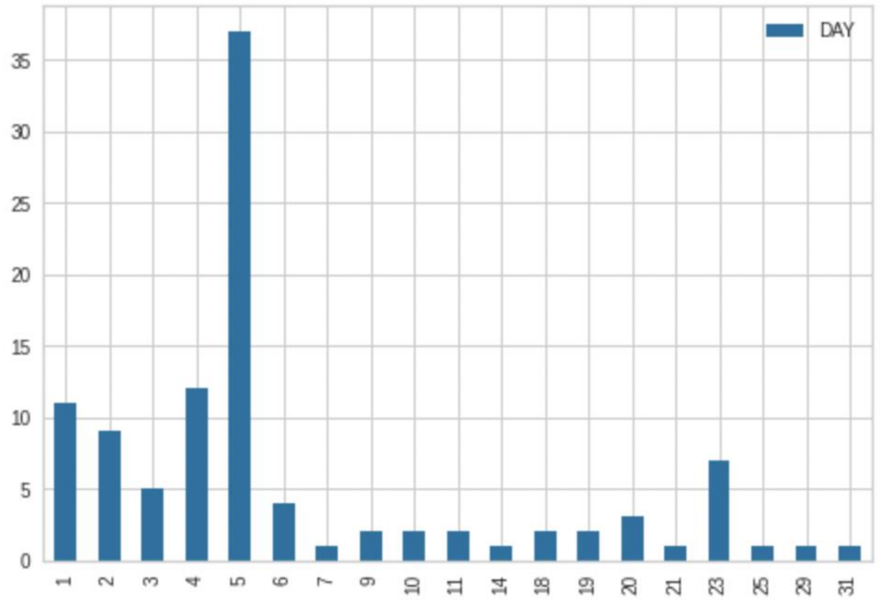


Figure 11: Day-wise distribution of cancelled flights in March 2015

Figure 12 below shows the geographical locations of the airports at which the flights were cancelled in all the other months. As it can be interpreted, the chances of flights getting canceled at the beginning of the year are higher, and on the eastern coast of the U.S., the chances are even higher. As the middle of the year approaches, the events of flight cancellations get lesser. However, During December, the number of flight cancellations increased. Hence, again, as the end of the year approaches the chances of flight cancellations increase.

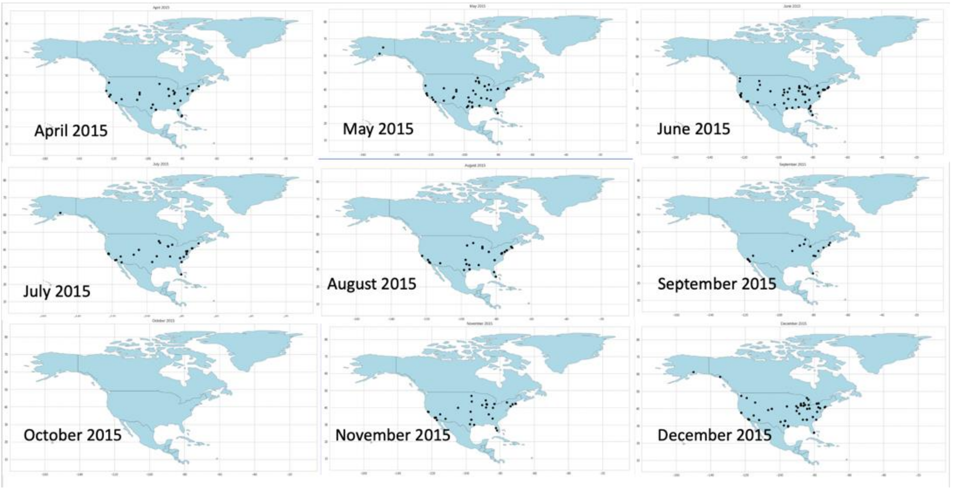


Figure 12: GeoPandas plot of the months from April 2015 to December 2015

With this critical analysis of canceled flights, many factors could be worked out. If flight cancellations are predicted, many better services could be planned and provided to the passengers to improve the benefit of the airline industry. To predict flight delays depending on certain variables, an automated predictive model has been built, and further is the documentation of the model.

3.5 Predictive Analysis

A different data frame was formed to analyse cancelled flights in 2015 critically. To build a predictive model, exploratory data analysis is performed separately. The initially created master data, which contains the data of all the flights, including the longitudes and latitudes of the airports from where the flights took off, has been used to build the model.

3.5.1 Exploratory Data Analysis

The variables ”Unnamed,” and ”flight number” have unique values for each record. Such columns do not teach the model about the target variable. These columns will not affect the target variable, so they are not very useful for building the predictive model. Therefore, these columns are dropped in the initial stage. In addition to these columns, variable ”year” has the same value for all the records, which is 2015, and variable ”country” has the same value, which is ”U.S.A.,” as the data is of flights operated in the U.S.A. during the year 2015. Therefore these two columns are also dropped as they don’t add any value to build a predictive model.

There are three variables, ”scheduled departure,” ”departure time,” and ”de- parture delay.” The ”scheduled departure” variable is the scheduled time of the departure of flights, and the variable ”departure time” contains the data of the time at which flights departed. The variable ”departure delay” is nothing but the difference between variables ”scheduled departure” and ”departure time.” There- fore, as the three variables add the same value, two variables are dropped, and only the ”departure delay” variable is kept. This variable is the target variable of the model.

As the target variable is ”departure delay,” the origin of the airport is essential, and the airport’s destination is less important. Therefore, the variable ”destination airport” is dropped. In addition, the variable ”diverted” represents the value of flights diverted during the operation. In the dataset, only 128 flights were diverted from more than 50 thousand flights, meaning less than 0.5% were diverted. There- fore, this variable adds minimum value to the predictive model, as the values for all the records are the same, except for those 0.5% records. Hence, this variable is also dropped.

As previously seen in Figure 4 (heatmap) in exploratory data analysis the heatmap of the data ,the blank white spaces represent the missing records in the data. The variables such as ”security delay,” ”weather delay,” ”late aircraft delay,” ”Air system delay,” and ”Elapsed time” have null values of more than 80% of the records. Therefore, the best solution is to drop those variables. After dropping the variables mentioned above, the null values in the variables are treated by removing those records. After removing all the null records the data is clean and ready for further procedures.

In this study, it has been considered that if the departure delay of the flight is beyond 15 minutes, it is considered delayed. This assumption is purely on the experience basis and has been considered only for this research purpose, which is to build a predictive model. Therefore, a new variable called ”Flight delayed” has been generated. This variable is a binary-type variable, and the values are either 0 or 1. If the departure delay is more than 15, the flight is considered delayed, and the value of the record on the ”flight delayed” variable is 1. If the departure delay is less than 15 minutes, the value is zero, which means the flight is considered to be not delayed. This variable will be the target variable. As the variable is binary, a classification model has been built further.

3.5.2 Oversampling of the data

Figure 13 below is the bar plot of value counts in the ”flight delayed” variable, which is the target variable. As can be seen, the number of values of 1 is far lesser than those of 0.Less number of flights were delayed. Therefore this data is biased. The class distribution in the variable is imbalanced. And if the predictive model is trained with imbalanced data, it will be biased toward predicting the larger class. Therefore, a sampling technique is used in this study to overcome this problem. As the class of 0 is less than 50% of the class of 1, the oversampling technique is preferred over undersampling because undersampling will cause the loss of information. Figure 16 is the bar plot of the target variable after performing the oversampling. The classes in the variable are balanced.

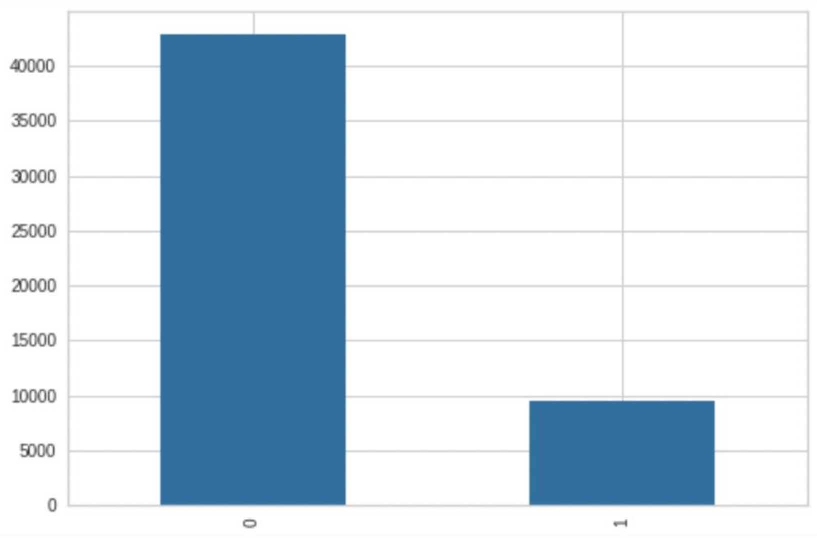


Figure 13: Visualization of imbalanced classes in target variable

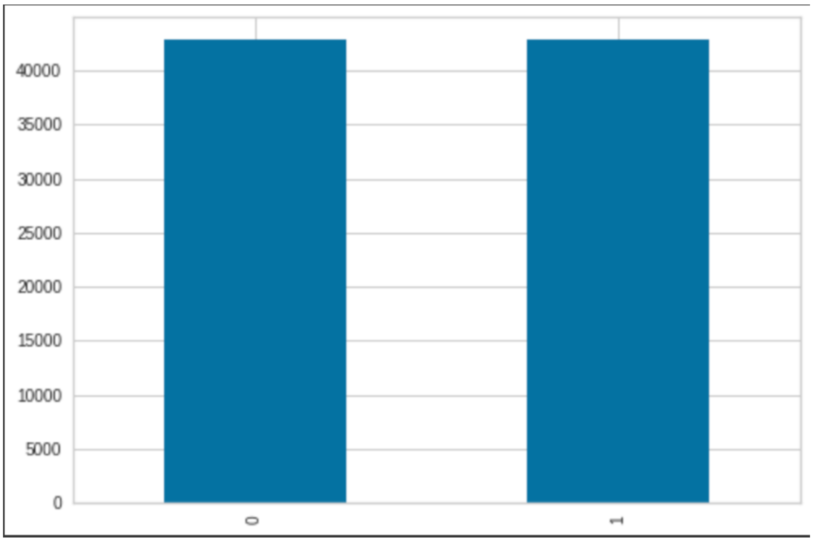
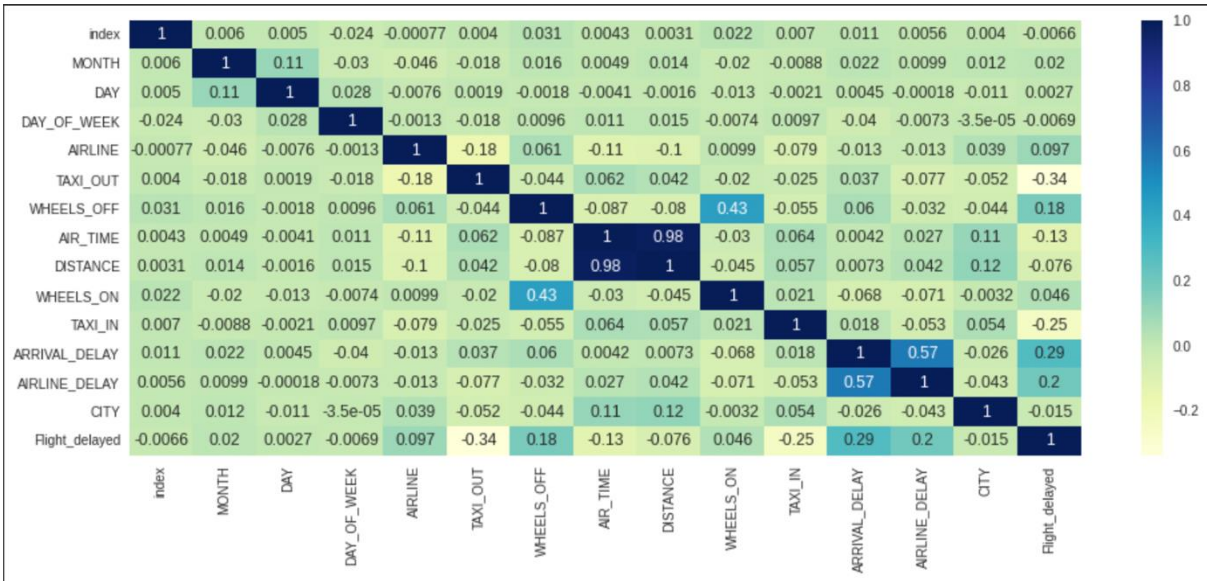


Figure 14: Visualisation of balanced data of classes in target variable

Figure 15 shows the heatmap of correlation matrices. It shows the correlation between each variable with other variables. This can be used to identify multi-collinearity. Multi-collinearity causes a problem because it undermines the statistical significance of other independent variables. As can be observed, some of the variables are highly correlated to each other, such as distance and airtime. To overcome this, principal component analysis has been used further.



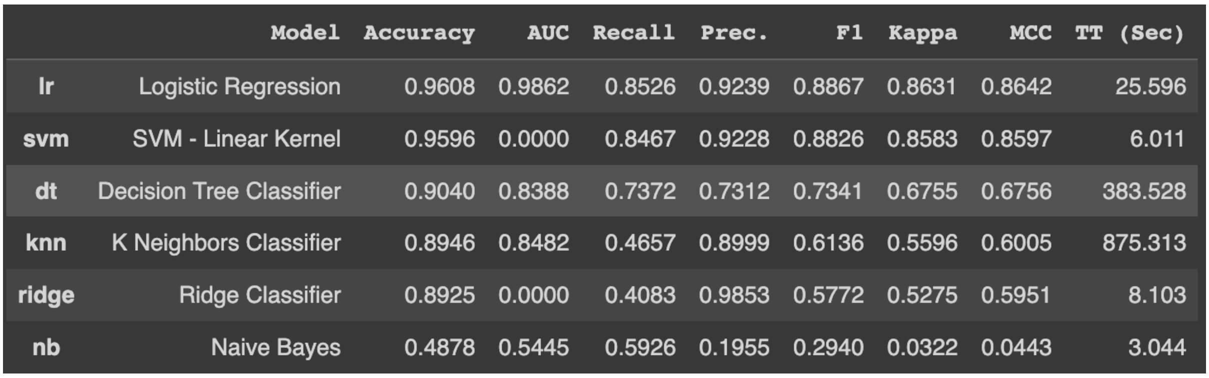
	index	MONTH	DAY	DAY_OF_WEEK	AIRLINE	TAXI_OUT	WHEELS_OFF	AIR_TIME	DISTANCE	WHEELS_ON	TAXI_IN	ARRIVAL_DELAY	AIRLINE_DELAY	CITY	Flight_delayed
index	1	0.006	0.005	-0.024	-0.00077	0.004	0.031	0.0043	0.0031	0.022	0.007	0.011	0.0056	0.004	-0.0066
MONTH	0.006	1	0.11	-0.03	-0.046	-0.018	0.016	0.0049	0.014	-0.02	-0.0088	0.022	0.0099	0.012	0.02
DAY	0.005	0.11	1	0.028	-0.0076	0.0019	-0.0018	-0.0041	-0.0016	-0.013	-0.0021	0.0045	-0.00018	-0.011	0.0027
DAY_OF_WEEK	-0.024	-0.03	0.028	1	-0.0013	-0.018	0.0096	0.011	0.015	-0.0074	0.0097	-0.04	-0.0073	-3.5e-05	-0.0069
AIRLINE	-0.00077	-0.046	-0.0076	-0.0013	1	-0.18	0.061	-0.11	-0.1	0.0099	-0.079	-0.013	-0.013	0.039	0.097
TAXI_OUT	0.004	-0.018	0.0019	-0.018	-0.18	1	-0.044	0.062	0.042	-0.02	-0.025	0.037	-0.077	-0.052	-0.34
WHEELS_OFF	0.031	0.016	-0.0018	0.0096	0.061	-0.044	1	-0.087	-0.08	0.43	-0.055	0.06	-0.032	-0.044	0.18
AIR_TIME	0.0043	0.0049	-0.0041	0.011	-0.11	0.062	-0.087	1	0.98	-0.03	0.064	0.0042	0.027	0.11	-0.13
DISTANCE	0.0031	0.014	-0.0016	0.015	-0.1	0.042	-0.08	0.98	1	-0.045	0.057	0.0073	0.042	0.12	-0.076
WHEELS_ON	0.022	-0.02	-0.013	-0.0074	0.0099	-0.02	0.43	-0.03	-0.045	1	0.021	-0.068	-0.071	-0.0032	0.046
TAXI_IN	0.007	-0.0088	-0.0021	0.0097	-0.079	-0.025	-0.055	0.064	0.057	0.021	1	0.018	-0.053	0.054	-0.25
ARRIVAL_DELAY	0.011	0.022	0.0045	-0.04	-0.013	0.037	0.06	0.0042	0.0073	-0.068	0.018	1	0.57	-0.026	0.29
AIRLINE_DELAY	0.0056	0.0099	-0.00018	-0.0073	-0.013	-0.077	-0.032	0.027	0.042	-0.071	-0.053	0.57	1	-0.043	0.2
CITY	0.004	0.012	-0.011	-3.5e-05	0.039	-0.052	-0.044	0.11	0.12	-0.0032	0.054	-0.026	-0.043	1	-0.015
Flight_delayed	-0.0066	0.02	0.0027	-0.0069	0.097	-0.34	0.18	-0.13	-0.076	0.046	-0.25	0.29	0.2	-0.015	1

Figure 15: Heatmap of correlation between variables

3.5.3 Model building and evaluation:

Many researchers have used many predictive machine learning algorithms. So, an automated machine learning algorithm is used instead of testing various algorithms. In this project, an automated machine learning module PyCaret is used to predict the flight delay. This module builds predictive models using different algorithms, and based on evaluation matrices such as accuracy, precision, recall, etc., it provides the ranking of the algorithms.

After importing the Pycaret module, the names of categorical variables are stored in a separate list. This list is a parameter that is used while setting up the model. Also, while setting up the model, the parameter for principal component analysis has been set to true. This parameter will help reduce the dimension of the model, which will overcome the issue of multicollinearity. Also, the parameter of removing outliers has been set to true. When this parameter is true, it removes the outliers from the training data. The default percentage of outliers to be re- moved of 5%. The parameter of normalizing is also set to true. This feature will scale down the numeric variables to the given range. The default method for normalization is the 'z score.' The fold parameter, which determines the number of cross-validations, has been set to its default value, 10. The training size is kept at default, which is 70%.



	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
lr	Logistic Regression	0.9608	0.9862	0.8526	0.9239	0.8867	0.8631	0.8642	25.596
svm	SVM - Linear Kernel	0.9596	0.0000	0.8467	0.9228	0.8826	0.8583	0.8597	6.011
dt	Decision Tree Classifier	0.9040	0.8388	0.7372	0.7312	0.7341	0.6755	0.6756	383.528
knn	K Neighbors Classifier	0.8946	0.8482	0.4657	0.8999	0.6136	0.5596	0.6005	875.313
ridge	Ridge Classifier	0.8925	0.0000	0.4083	0.9853	0.5772	0.5275	0.5951	8.103
nb	Naive Bayes	0.4878	0.5445	0.5926	0.1955	0.2940	0.0322	0.0443	3.044

Figure 16: Score table of evaluation matrices

With the parameters mentioned above, PyCaret is trained on the data, and the results shown in figure 16 are acquired. As it can be seen, as per PyCaret, the logistic regression algorithm is proven to be the best, with the highest accuracy of 96.08%, if accuracy is considered. But, if the recall is regarded as an essential parameter, the Random Forest and Extra trees classifier performed better than the Logistic Regression classifier. And if the precision is considered a significant factor, SVM performed better than the Logistic regression.

4 Conclusion and Future Work

This research aims to gain a thorough understanding of flight data and build inter- pretations with the help of exploratory data analysis and predictive models. The exploratory data analysis provides valuable insights, which could be used to build strategies. The business strategies are aimed at providing the best services to the customers. Accurate business strategies will help the industry to stay competitive, and it will also help expand the business.

Flight delays and cancellations cause substantial economic losses to airline industries. In this research, using Geopandas and other visualization modules, the patterns in flight cancellations are understood. With the help of such insights, the airline industry could stay prepared for any unexpected challenges.

This research uses an automated machine learning module PyCaret, to build predictive models. Using automated machine learning, comparing the evaluation matrices made the model selection process more manageable. For example, sup- pose the business problem focuses on the correctly predicted number of delayed daily flights. In that case, the recall will be the best matrix to select the best algorithm to build a predictive model.

The results of this research are useful for future work. As the data size is extensive and we have used only the sample data for this project, big data stor- age and handling of big data, cloud storage, and cloud services like A.W.S. and Microsoft Azure can be used.

5 Acknowledgement

I thank my supervisors, Prof. Christian Horn, for their support during this research study. I have achieved these results due to their continuous feedback and guidance in the right direction. I am also thankful to respective organizations for providing open-source datasets and software for academic research work. Lastly, I thank my parents and friends for their patience, support, and guidance throughout the research and course.

References

Choi, S., Kim, Y.J., Briceno, S. and Mavris, D., 2016, September. Prediction of weather-induced airline delays based on machine learning algorithms. In *2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC)* (pp. 1-6). IEEE.

Demir, E. and Demir, V.B., 2017, May. Predicting flight delays with artificial neural networks: case study of an airport. In *2017 25th Signal Processing and Communications Applications Conference (SIU)* (pp. 1-4). IEEE.

Dhanawade, R., Deo, M., Khanna, N. and Deolekar, R.V., 2019, March. Analyzing factors influencing flight delay prediction. In *2019 6th International Conference on Computing for Sustainable Global Development (INDIACom)* (pp. 1003-1007). IEEE.

Gao, Y., Huyan, Z. and Ju, F., 2015, December. A prediction method based on neural network for flight turnaround time at airport. In *2015 8th International Symposium on Computational Intelligence and Design (ISCID)* (Vol. 2, pp. 219-222). IEEE.

Gui, G., Liu, F., Sun, J., Yang, J., Zhou, Z. and Zhao, D., 2019. Flight delay prediction based on aviation big data and machine learning. *IEEE Transactions on Vehicular Technology*, *69*(1), pp.140-150.

Hu, P., Zhang, J. and Li, N., 2021, October. Research on Flight Delay Prediction Based on Random Forest. In *2021 IEEE 3rd International Conference on Civil Aviation Safety and Information Technology (ICCASIT)* (pp. 506-509). IEEE.

Jiang, Y., Miao, J., Zhang, X. & Le, N. (2020), 'A multi-index prediction method for flight delay based on long short-term memory network model', *2020 IEEE 2nd International Conference on Civil Aviation Safety and Information Technology (ICCASIT* pp. 159–163.

Kim, Y. J., Choi, S., Briceno, S. & Mavris, D. (2016), 'A deep learning approach to flight delay prediction', *2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC)* pp. 1–6.

Lv, Y., Duan, Y., Kang, W., Li, Z. & Wang, F.-Y. (2014), 'Traffic flow prediction with big data: A deep learning approach', *IEEE Transactions on Intelligent Transportation Systems* pp. 1–9.

Manna, S., Biswas, S., Kundu, R., Rakshit, S., Gupta, P. & Barman, S. (2017), 'A statistical approach to predict flight delay using gradient boosted decision tree', *2017 International Conference on Computational Intelligence in Data Science(ICCIDS)* pp. 1–5.

Moreira, L., Dantas, C., Oliveira, L., Soares, J. & Ogasawara, E. (2018), 'On eval- uating data preprocessing methods for machine learning models for flight de- lays', *2018 International Joint Conference on Neural Networks (IJCNN)* pp. 1– 8.

Murca, M. C. R. & Hansman, R. J. (2019), 'Identification, characterization, and prediction of traffic flow patterns in multi-airport systems', *IEEE Transactions on Intelligent Transportation Systems* 20, 1683–1696.

Nijsure, Y. A., Kaddoum, G., Gagnon, G., Gagnon, F., Yuen, C. & Mahapatra, R. (2015), 'Adaptive air-to-ground secure communication system based on ads-b and wide-area multilateration', *Transactions on Vehicular Technology* 65(5), 3150–3165.

Pamplona, D. A., Weigang, L., de Barros, A. G., Shiguemori, E. H. & Alves, C. J. P. (2018), 'Supervised neural network with multilevel input layers for predicting of air traffic delays', *2018 International Joint Conference on Neural Networks (IJCNN)* pp. 1–6.

Xing, Z. & Tang, Y. (2016), 'The model for optimizing airport flight delays al- location', *2016 8th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)* pp. 188–191.