IDS 561 Analytics for Big Data

Fall 2022

FINAL REPORT

"Flight On-Time Prediction and Sentiment Analysis using Pyspark"

Submitted By:

Chris Lazarus: 673773993

Pratik Sharma: 667724536

Sayali Bonawale: 656488690

PROBLEM STATEMENT:

By examining delays and cancellations, we want to develop powerful machine-learning models to forecast flight performance on-time. Even if they are unavoidable, flight delays are one of the most serious issues facing the aviation sector. Unavoidable flight delays have a huge negative impact on the airlines' profits and losses. Since the information can be used to increase customer satisfaction and income for airline agencies, airlines must accurately forecast flight delays. In an additional effort to widen the scope of this project, we are analyzing tweets from US airline passengers to uncover the reasons for their dissatisfaction with a particular airline using sentiment analysis. To directly highlight how the sentiment corresponds to an airline and how we can use this concept into play to understand things at the broader spectrum.

DATASET DESCRIPTION:

To predict flight delay we are using the following dataset which is available to us from Kaggle.

This dataset contains departure data for flights by date, day of the week, and month, by the carrier. It also includes the departure time, how long the flight will take to reach the destination, and the flight distance.

Number of records: 275,000

Features: Month, Date, Day of Week, Origin, Distance, Airline ID, Flight ID, Depart, Flight Duration, Delay

Data source Link: **[1]**

To perform Twitter US Airline sentiment analysis, we are using the following source dataset. We are going to perform evaluations for determining whether the sentiment of the tweets is positive, neutral, or negative. From there we are going to segment the reasons for the dissatisfaction of passengers.

Number of records: 11,541

Features: Tweet ID, Comments, Airline, Name, Retweet count, Text, Created, Location, User Time zone

Data source Link **[2]**

TECHNIQUES:

Workflow for On-Time Prediction

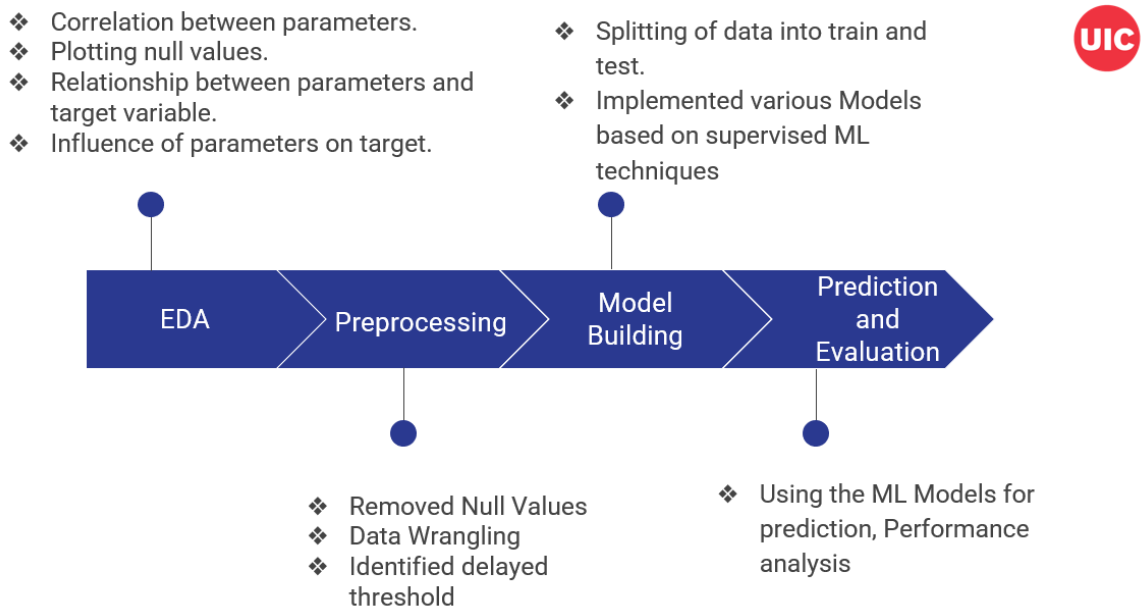


Fig.1: Process flow for building the predictive model

Our workflow for on-time prediction is shown in the diagram above. As previously said, we got our datasets from Kaggle.com. After that, Google Collaboratory was utilized for modeling and data preprocessing. We have employed pandas libraries for EDA. Preprocessing included removing Null values from our dataset, performing data wrangling using the String Indexer to transform a string column into an ML column of label indices, and combining all feature data and separate 'label' data in a dataset, we use VectorAssembler to train ML models.**[5]** We have determined a delayed threshold and trained our models using it. We have utilized Pyspark to implement multiple ML models for model building.

Exploratory Data Analysis and Visualization

We have done EDA on the dataset to understand the importance and influence of parameters on each other. Using the correlation matrix, we identified the relationship between the features. After that, we determined the number of Null values in the primary dataset. We plotted a bar graph between the target variable and a few features to identify the relationship between them and also the variation between them. For this, we analyzed that the more the variance between the target variable and a parameter, the more influence it has on the target variable.

-Below given is a correlation matrix plotted between all the parameters which tells us if the parameters are negatively or positively correlated.

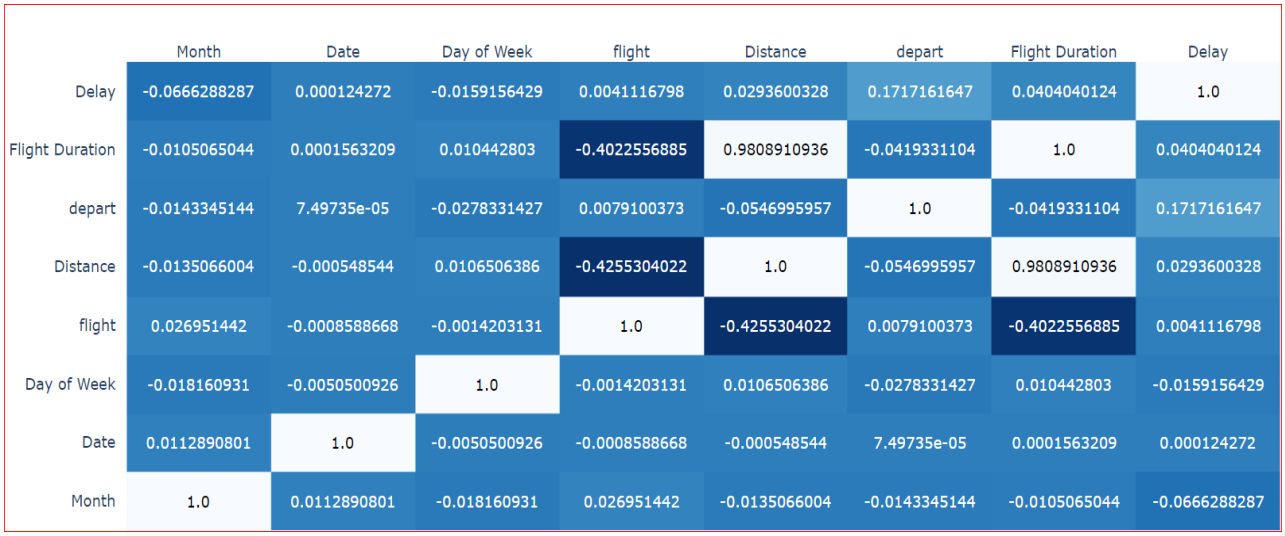


Fig.2: Correlation Matrix for flight parameters

-Next, we plotted a histogram to observe the Null value composition. We can see that only delay values have Null values present in the column which needs to be removed so that the records can be further used for manipulation.

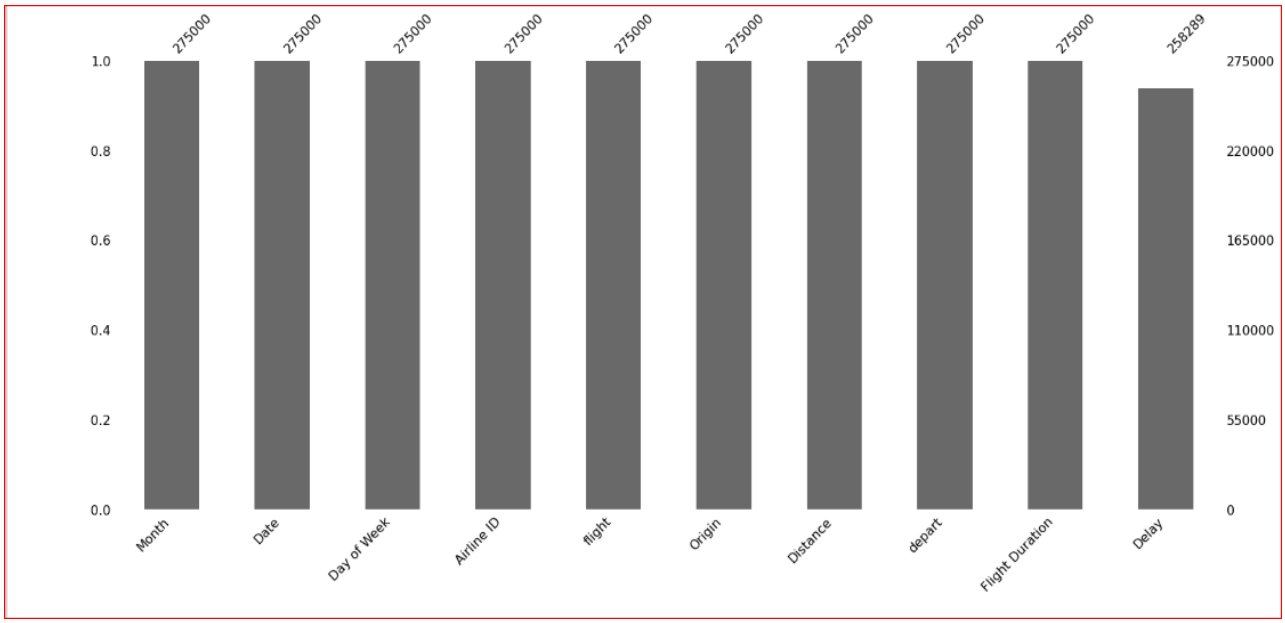
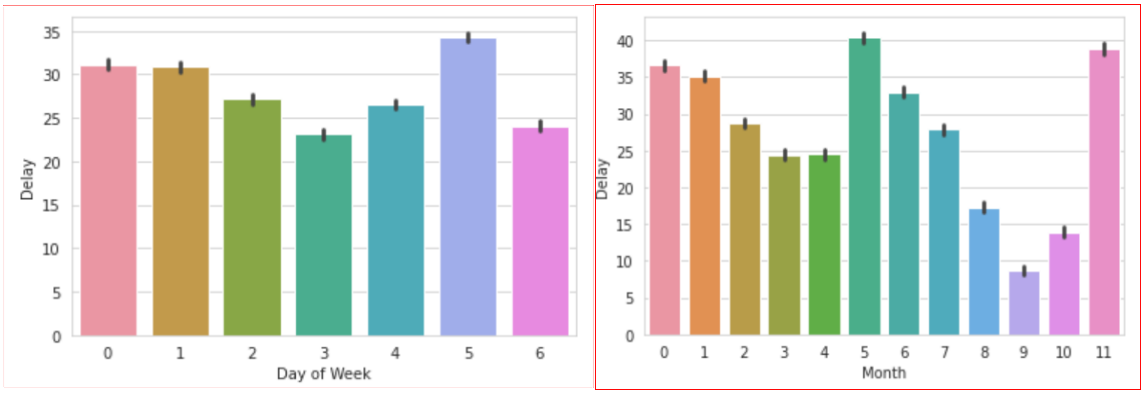
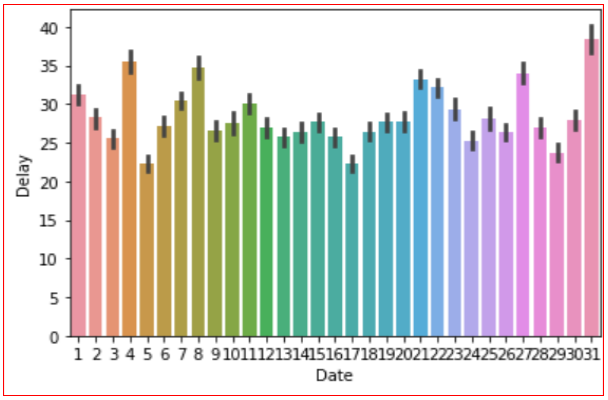


Fig.3: EDA graph 1: Null value identification

-Next, we plotted relationship graphs between all the parameters and Delays to understand the relationship between them and to derive any insights that could be used to prove our assumptions for the prediction.





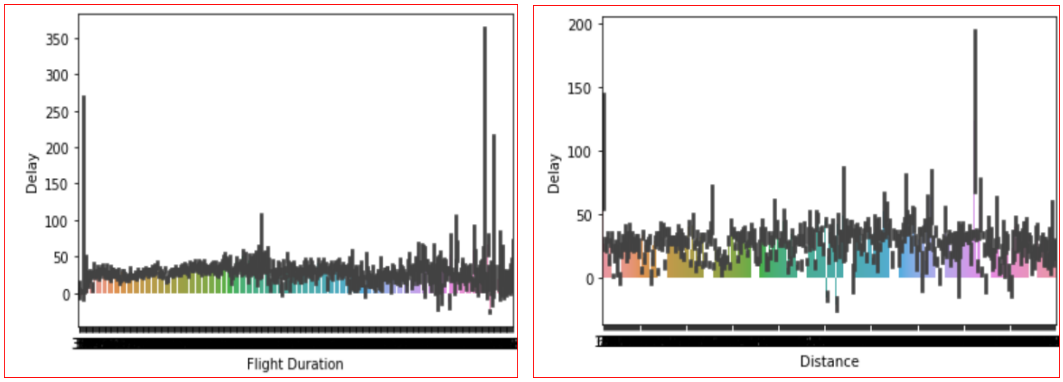


Fig.4: EDA graphs with week, month, flight duration, and distance parameters

- We plotted the graph to understand the distribution of delay counts more and less than the threshold value (which in our case is **10 minutes)**.

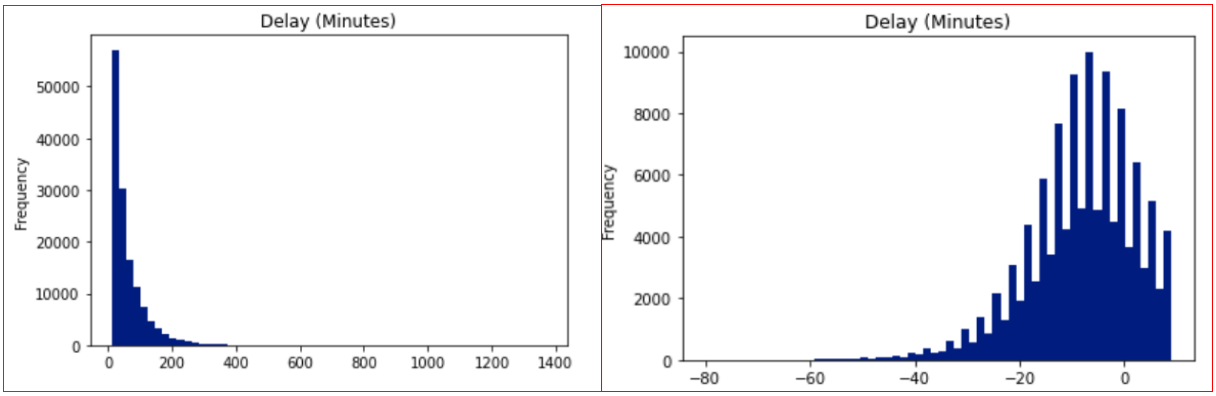


Fig.5: EDA graphs frequency vs Delay

-Lastly for EDA we plotted a graph against the Airlines and the Delay and no Delay count using Tableau.

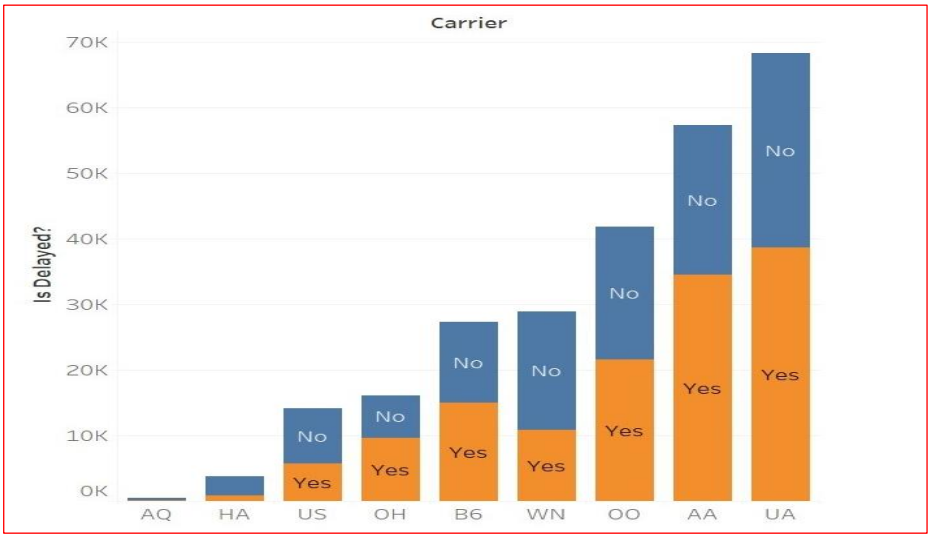


Fig.6: Delay Count vs Airline carrier

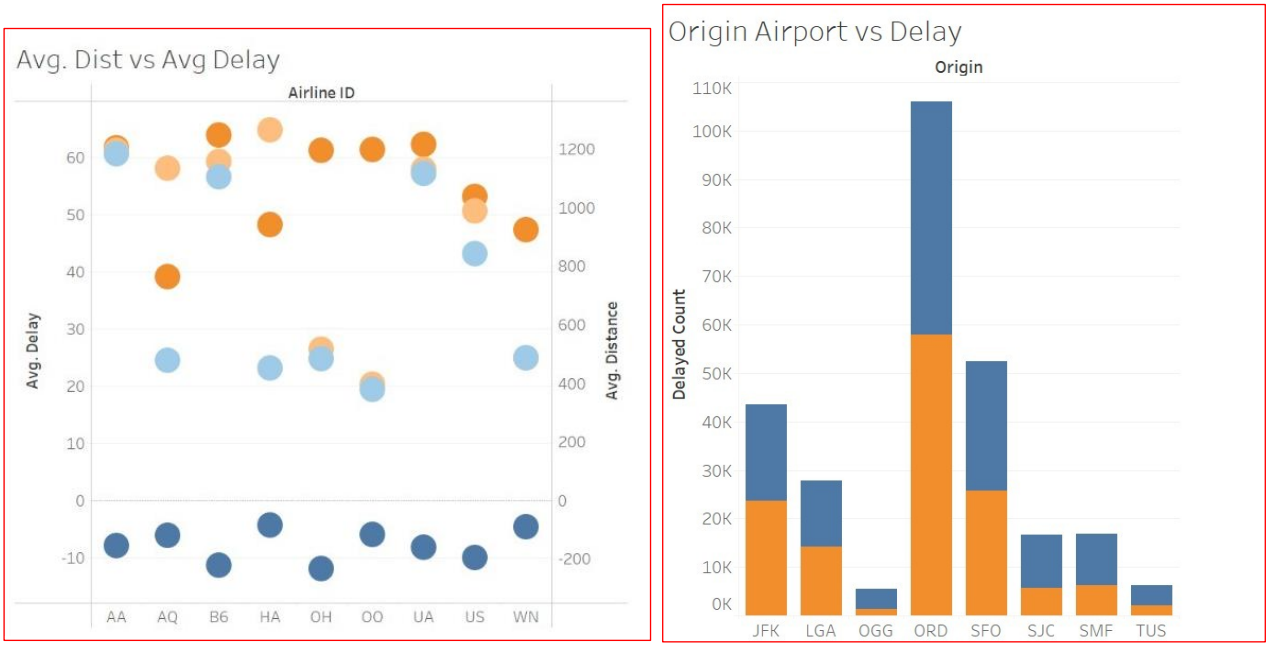


Fig.7: Delay vs Average distance and Origin Airport

Machine Learning Models

1. Decision Tree

Decision trees inherit simplicity and expandability and cover all the possible outcomes of a decision. It is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems.**[6]**

The threshold for the delay feature in this project has been set at 10. The target variable was then given the values 0 (Not Delay) or 1 (Delay) depending on the threshold. We divided the dataset into proportions of 70% train data and 30% test data. The findings will essentially be the prediction's accuracy and the likelihood that the model was accurate.

2. Logistic Regression

Logistic Regression helps in determining the relationship between dependent categorical variables and multiple independent variables based on the Probability score. It is a popular Machine Learning algorithm, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.

We wanted to understand which ML model performs better, hence we have compared the precision, recall, and F1-score of the models. Using logistic regression, we determined the probability score.

3. Naive Bayes

A Naive Bayes classifier is a probabilistic machine learning model that's used for classification tasks. Using Bayes theorem, we can find the probability of A happening, given that B has occurred. Here, B is the evidence and A is the hypothesis. Here, it is assumed that the predictors and features are independent. In other words, the presence of one specific feature has no impact on the other. The term "naive" is a result. By applying the Naïve Bayes class and fitting the model to train data, we have defined the decision tree classifier model.**[4]**

In our project, we used the transform() method to predict the test data. After the model had been trained, we predicted test data and looked at the metrics for accuracy. Using the function "evaluate" from ***pyspark.ml.evaluation***, we have utilized *MulticlassClassificationEvaluator* to examine the accuracy in this case.

4. Random Forest

Random Forest produces outcomes based on predictions of decision trees. Helpful in generating close to accurate predictions. It belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and improve the performance of the model.**[6]**

SENTIMENT ANALYSIS:

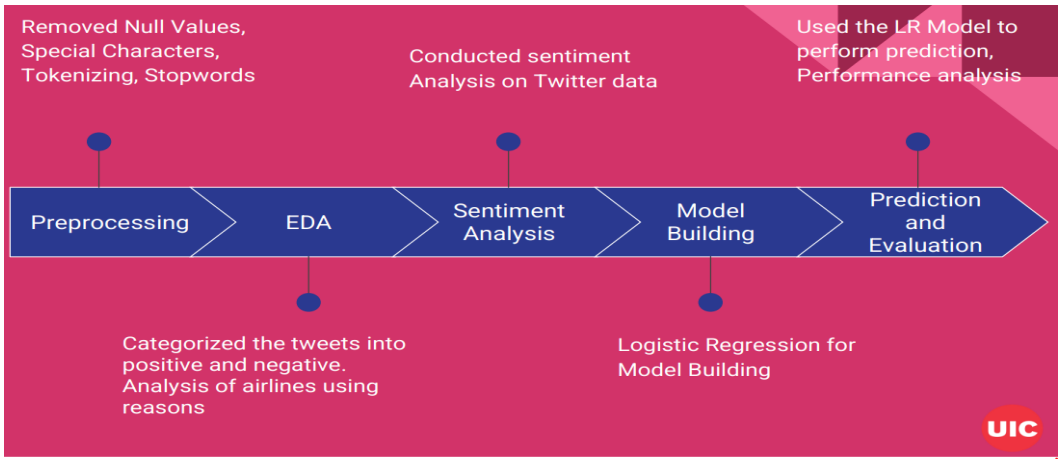


Fig.8: Process Flow Sentiment Analysis

Our workflow for sentiment analysis is shown above. We will use the Twitter sentiment airline data file for preprocessing checking if the information is complete and by the analysis requirements.

Next, we used the processed data for deriving sentiment insights related to the airline and the information stated about the airline by which we got the following graph.

- The graph below shows us the negative and positive sentiment composition for each airline which gives us an overall sentiment observation for all the Airlines.

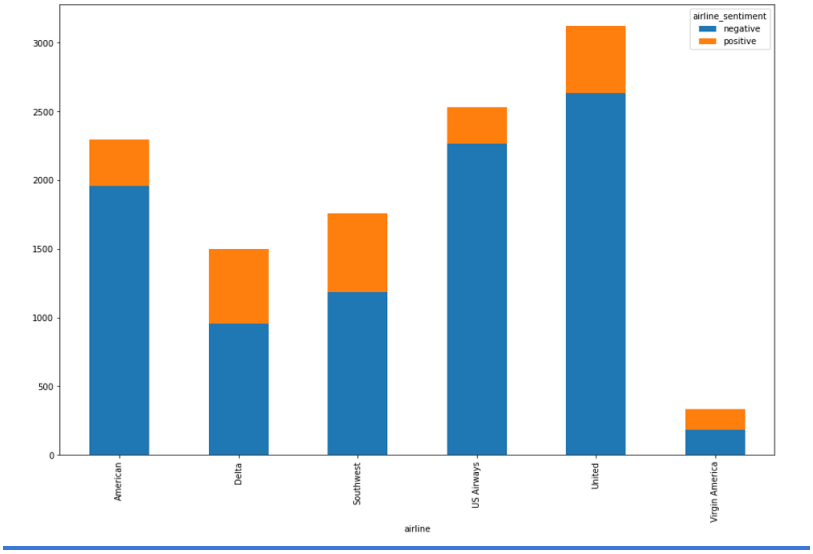


Fig.9: Sentiment type vs Airline Carrier

- Next, we use the previous information and plot a graph related to the negative reasons for the airlines which gives us a measure of how a particular reason results in the major negative sentiment.

-In our case, we can see that after customer service, late flight or delay is the major contributor to the negative sentiment.

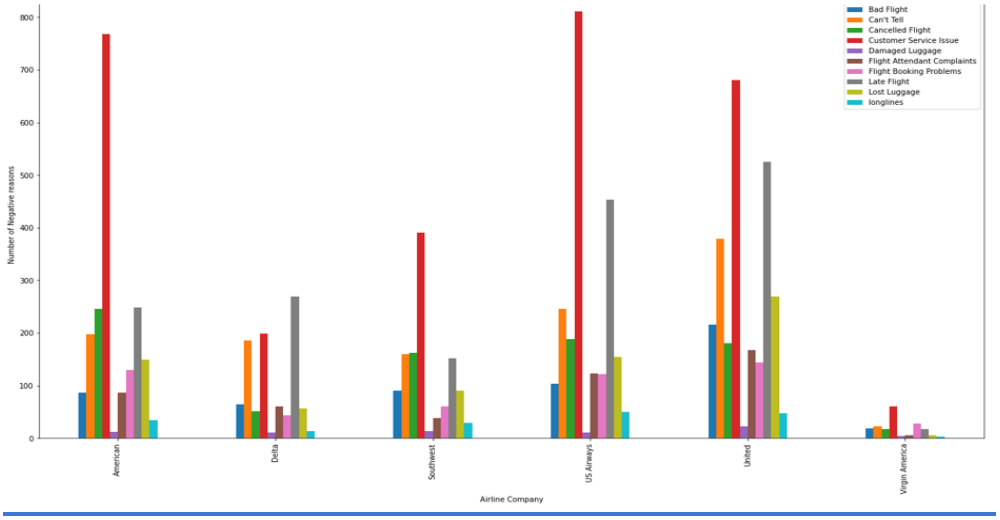


Fig.10: Reason for Delay vs Airline Carrier

We then split the data into training and testing and performed indexing as 0 and 1 respectively. For preparing our prediction model we tokenized the training data, used the tokenized data to remove stop words, and then performed hashing transformation to obtain the vector which is used further to train our logistic regression model.

We then used this model on the test data to perform prediction using the concept of our already build Logistic Regression Model and got an accuracy of **89.8%**

RESULTS:

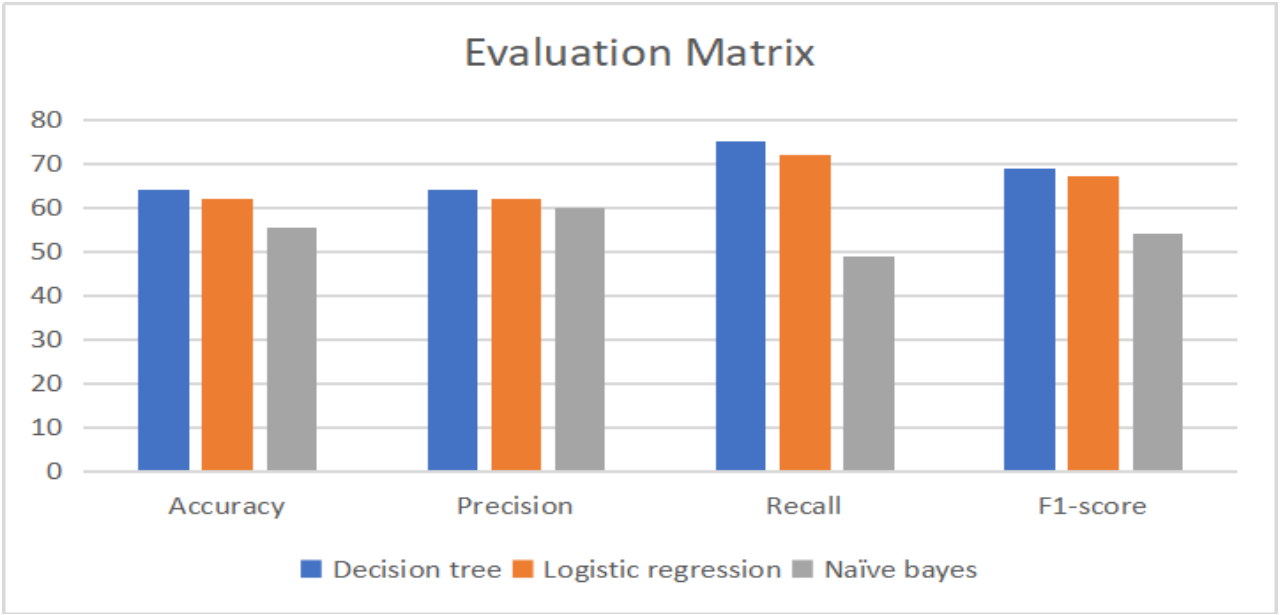


Fig.11: Evaluation Matrix Comparison for all the models

We decided to evaluate our ML models using a confusion matrix. From the diagram above, we can see that based on the label class False Negative i.e., Recall is more costly. Additionally, we found that logistic regression and decision trees perform more effectively.

For Sentiment Analysis, We used the logistic regression prediction model on the test dataset of

about 3501 records out of which using our model, we found that 3168 records were found to be predicted correctly. Our accuracy for the implemented model is approximately 90%.

Model	Accuracy
Logistic Regression	89.8%

Fig.12: Table for sentiment Analysis results

ROLE OF TEAM MEMBERS:

Our project primarily comprises two phases. The first phase consists of the sentiment analysis where we have analyzed how On-Time is a major factor resulting in negative sentiment for an airline. In the second phase, we used the ML models to use them for our prediction. We have worked together on each aspect of the project so that we can optimize our understanding of big data techniques.

Tasks	Team Members
Exploratory Data Analysis	Pratik, Sayali, Chris
Data Cleaning, Data Preprocessing	Sayali, Chris
ML Models	Pratik, Sayali
Performance and Sentiment Analysis	Chris, Pratik

References:

[1] https://www.kaggle.com/

[2] https://data.world/socialmediadata/twitter-us-airline-sentiment

[3] https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c

[4] https://medium.com/@nutanbhogendrasharma/feature-transformer-vectorassembler-in-pyspark-ml-feature-part-3-b3c2c3c93ee9

[5] https://python.plainenglish.io/decision-trees-random-forests-in-pyspark-d07546e4fa7d

[6] Kaur, Gurpreet & Malik, Kamal. (2021). A Sentiment Analysis of Airline System using Machine Learning Algorithms. International Journal of Advanced Research in Engineering. 12. 731-742. 10.34218/IJARET.12.1.2021.066.

[7] Ravi Kumar, G. & Kongara, Venkata Sheshanna & Babu G, Anjan. (2021). Sentiment Analysis for Airline Tweets Utilizing Machine Learning Techniques. 10.1007/978-3-030-49795-8_75.