Flight On-Time Prediction and Sentiment Analysis using Pyspark

-Group 7-

Chris Lazarus : 677377993
Pratik Sharma : 667724536
Sayali Bonawale : 656488690

IDS 561 Analytics for Big Data (38536) 2022 Fall

Overview

- Problem Statement
- Dataset Description
- Prediction Process Flow
- Exploratory Data Analysis
- Methodologies Used / Model Building
- Results and Performance Analysis
- Sentiment Analysis

Problem Statement

On Time Prediction

- On Time estimation is critical for airlines to get the degree of Customer satisfaction.
- Using different ML Classification and Regression Algorithms to propose the best method.

Sentiment Analysis

- Classification of Tweets so as to understand the Quality of Customer Experience.
- Categorize the tweets: Positive or negative sentiment

Dataset

Prediction Dataset :

Number of records: 275,000

Dataset: https://www.kaggle.com/

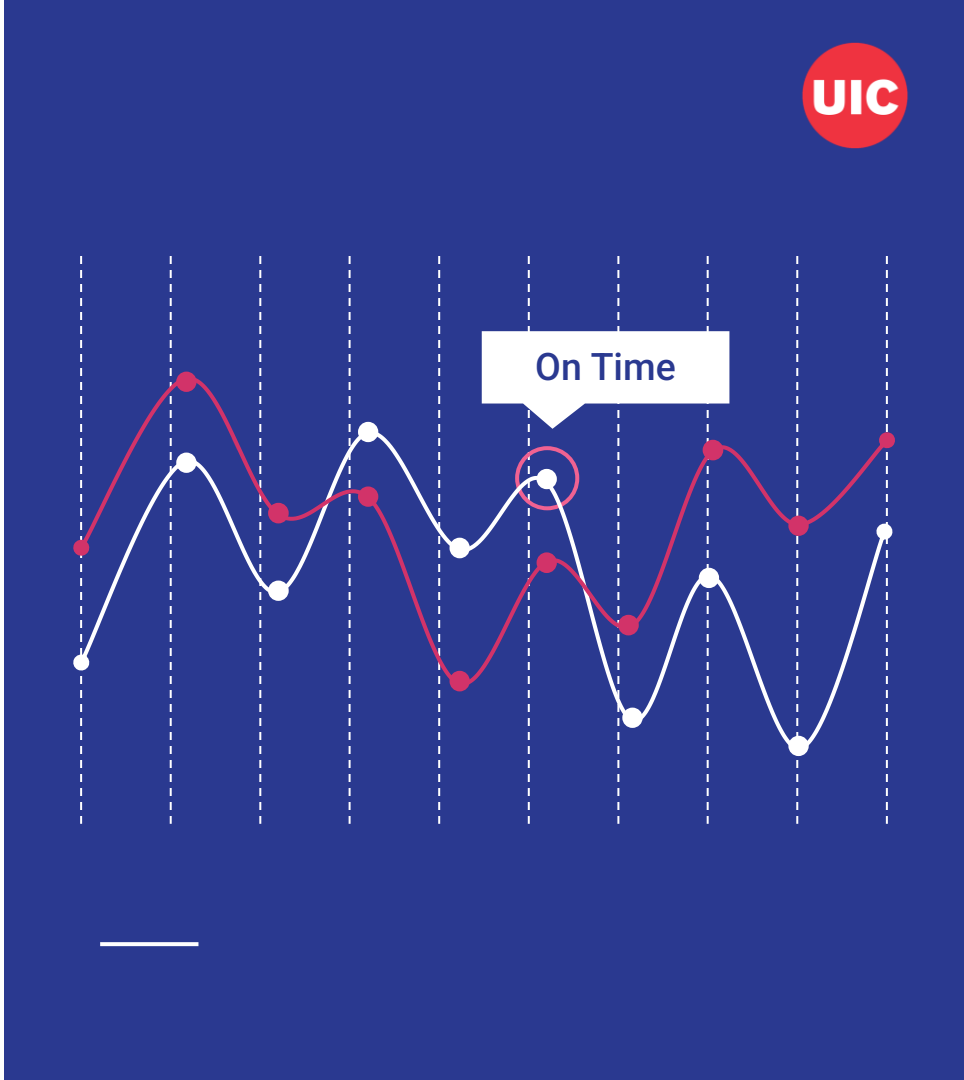
Features: Month, Date, Day of Week, Origin, Distance, Airline ID, Flight ID, Depart, Flight Duration, Delay

Sentiment Analysis Dataset :

Number of records: 11541

Dataset: https://data.world/crowdflower

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



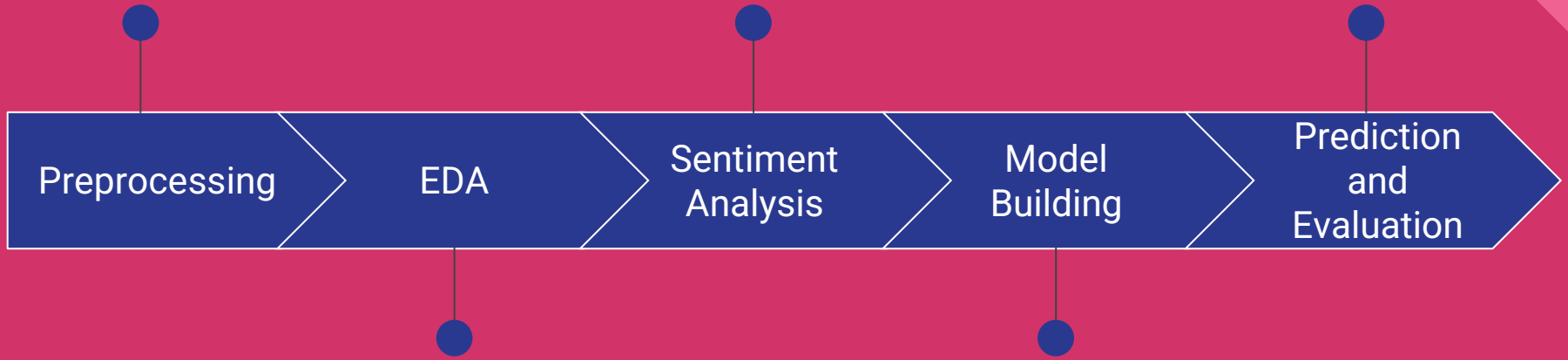
Airline Sentiment Analysis

- Tagged our sentiment feature as 0 for negative, 1 for positive and -1 as neutral.
- We then generated tokens from our text tweet dataset
- Processed data for removing Stop words.
- Transformed data using Hashing
- Trained Model using Logistic Regression and used this model for Prediction.

Removed Null Values, Special Characters, Tokenizing, Stopwords

Conducted sentiment Analysis on Twitter data

Used the LR Model to perform prediction, Performance analysis

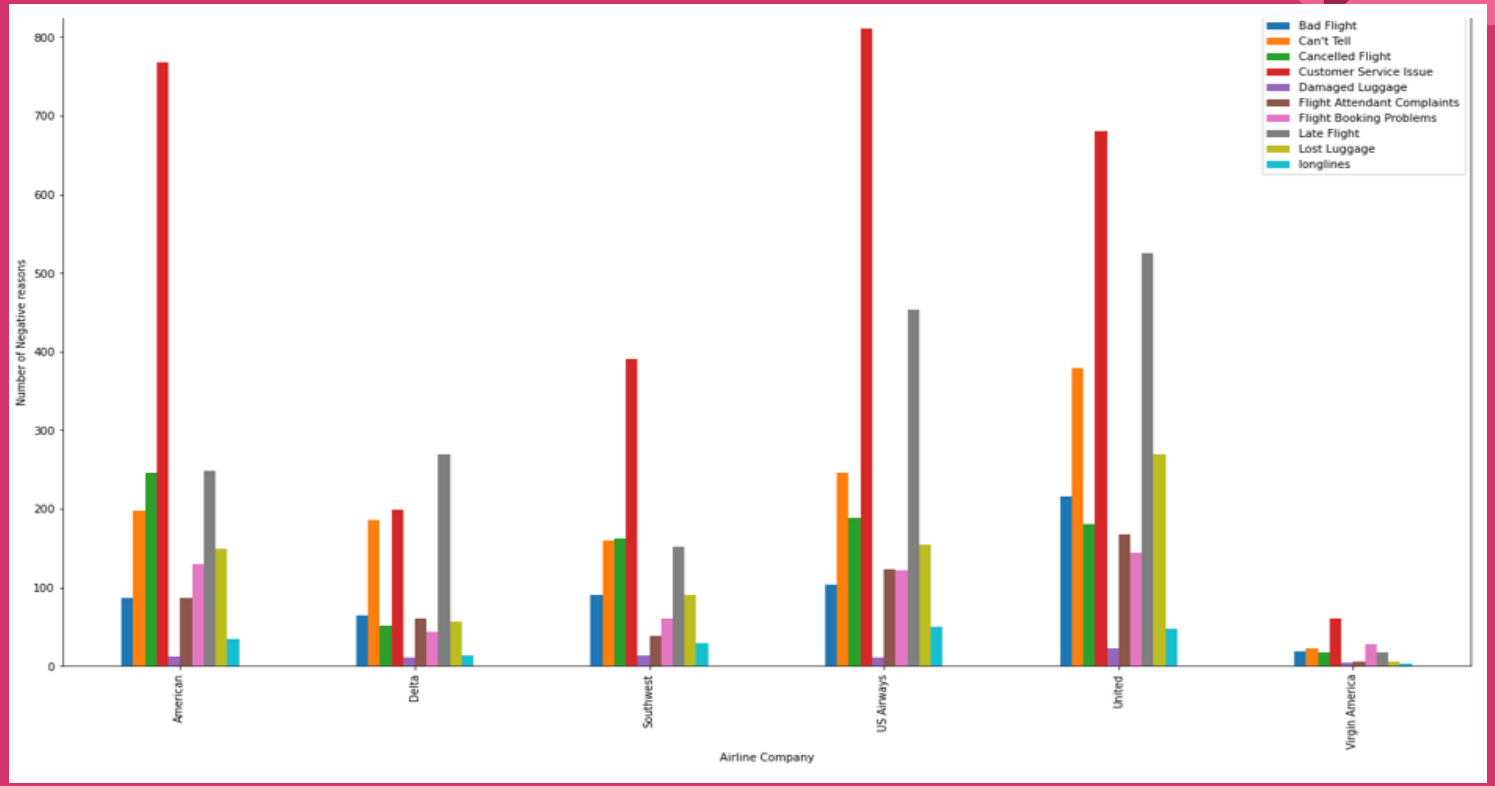


Categorized the tweets into positive and negative. Analysis of airlines using reasons

Logistic Regression for Model Building

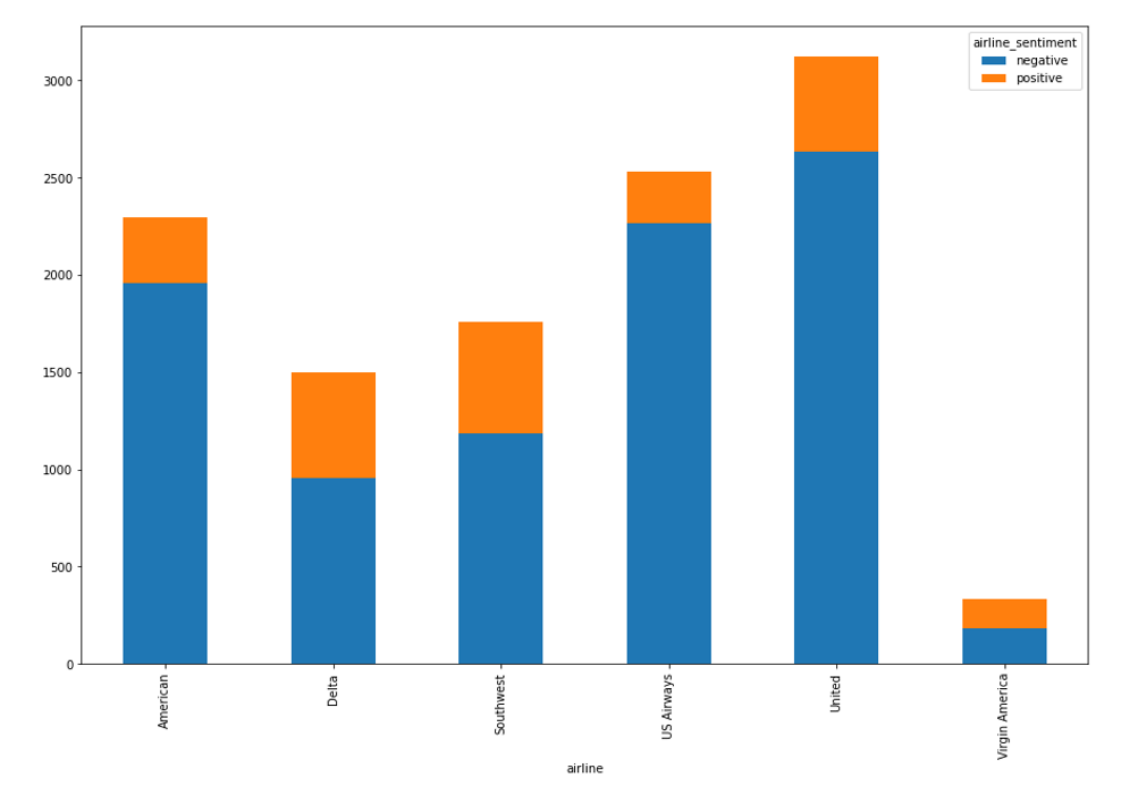
UIC

Airline Sentiment Analysis



Analysis of Airlines using additional comments and reasons

Airline Sentiment Analysis



Polarity scores of Major Airlines

Results and Performance Analysis

Model	Accuracy
Logistic Regression	89.8%

Polarity scores of Major Airlines

On Time Prediction

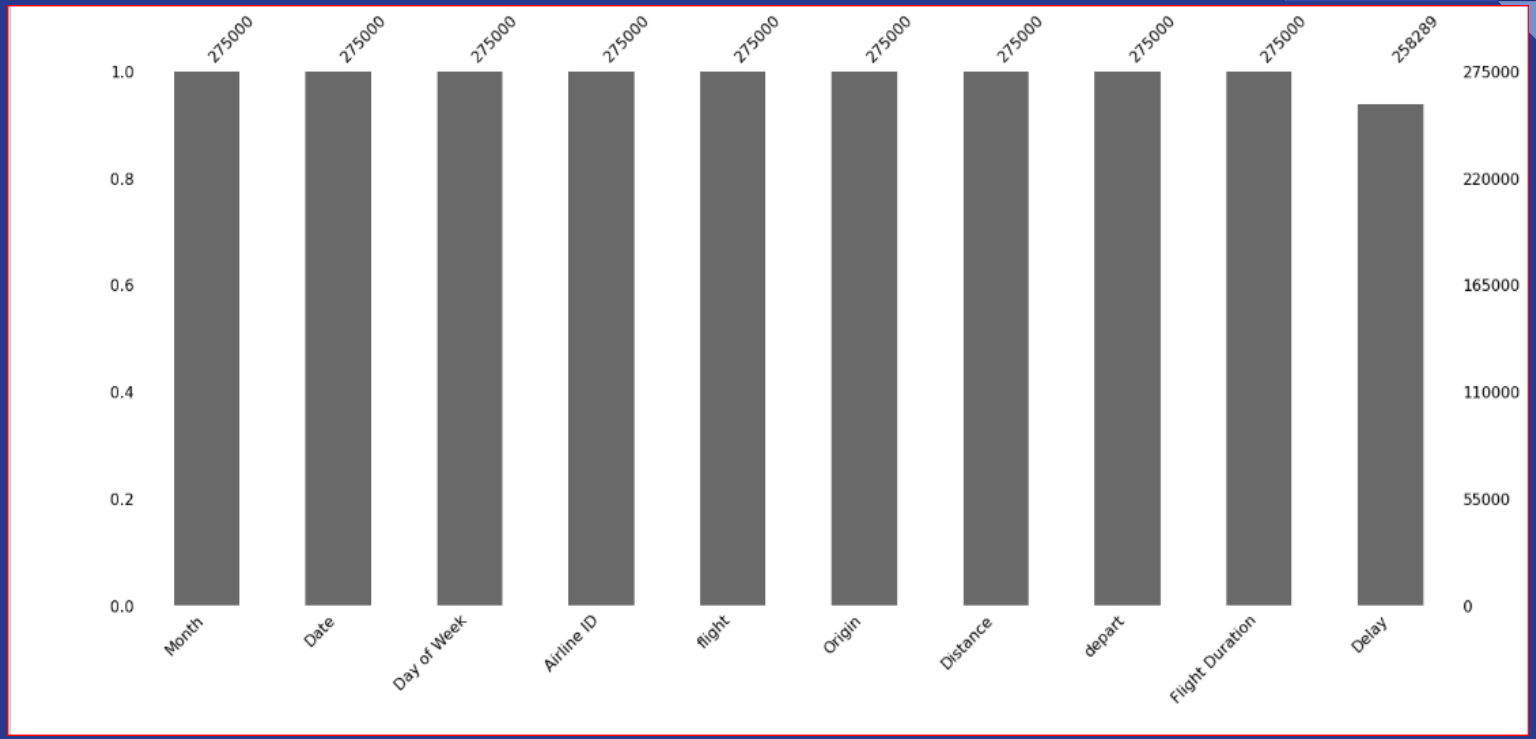
Correlation Matrix

Exploratory Data Analysis

	Month	Date	Day of Week	flight	Distance	depart	Flight Duration	Delay
Delay	-0.0666288287	0.000124272	-0.0159156429	0.0041116798	0.0293600328	0.1717161647	0.0404040124	1.0
Flight Duration	-0.0105065044	0.0001563209	0.010442803	-0.4022556885	0.9808910936	-0.0419331104	1.0	0.0404040124
depart	-0.0143345144	7.49735e-05	-0.0278331427	0.0079100373	-0.0546995957	1.0	-0.0419331104	0.1717161647
Distance	-0.0135066004	-0.000548544	0.0106506386	-0.4255304022	1.0	-0.0546995957	0.9808910936	0.0293600328
flight	0.026951442	-0.0008588668	-0.0014203131	1.0	-0.4255304022	0.0079100373	-0.4022556885	0.0041116798
Day of Week	-0.018160931	-0.0050500926	1.0	-0.0014203131	0.0106506386	-0.0278331427	0.010442803	-0.0159156429
Date	0.0112890801	1.0	-0.0050500926	-0.0008588668	-0.000548544	7.49735e-05	0.0001563209	0.000124272
Month	1.0	0.0112890801	-0.018160931	0.026951442	-0.0135066004	-0.0143345144	-0.0105065044	-0.0666288287

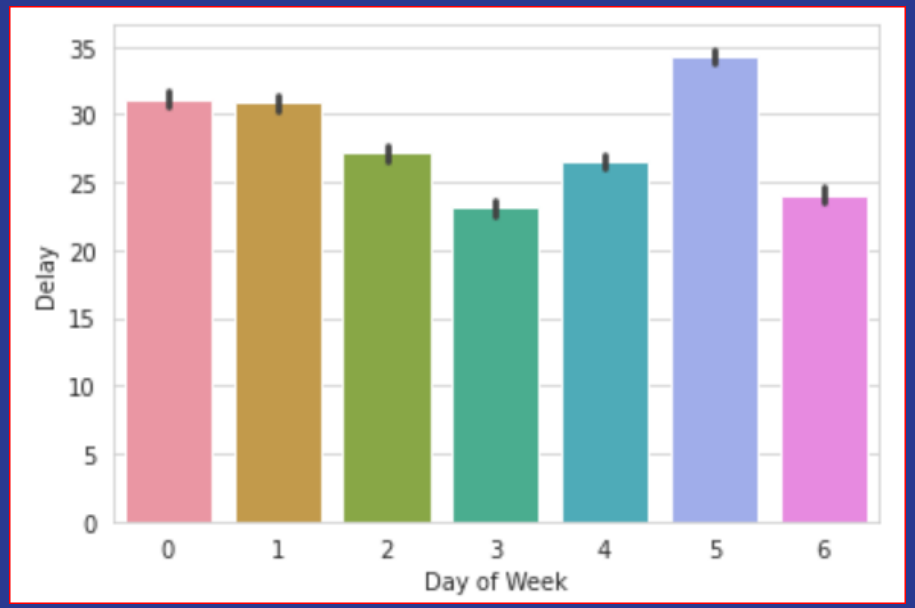
Correlation Matrix

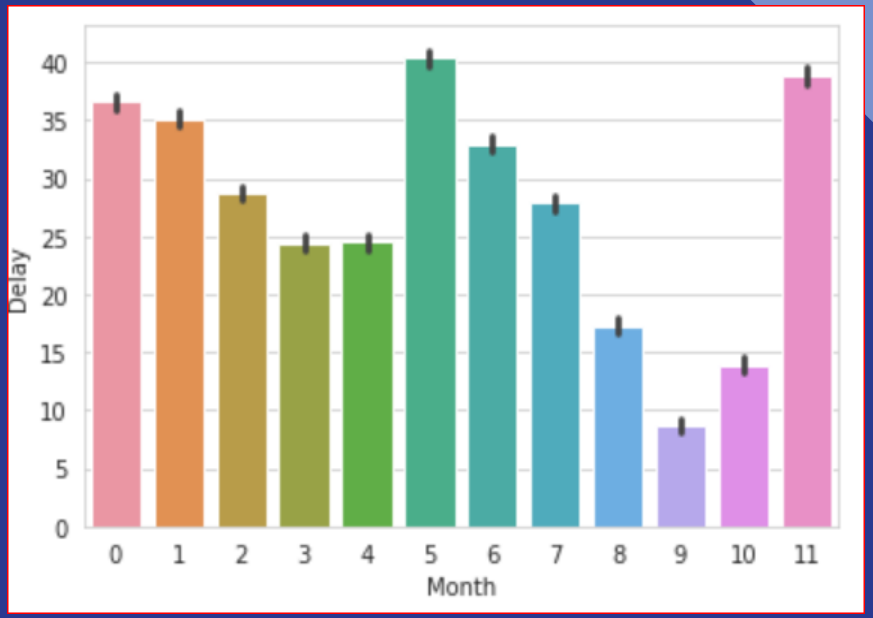
Exploratory Data Analysis



Null Value Composition

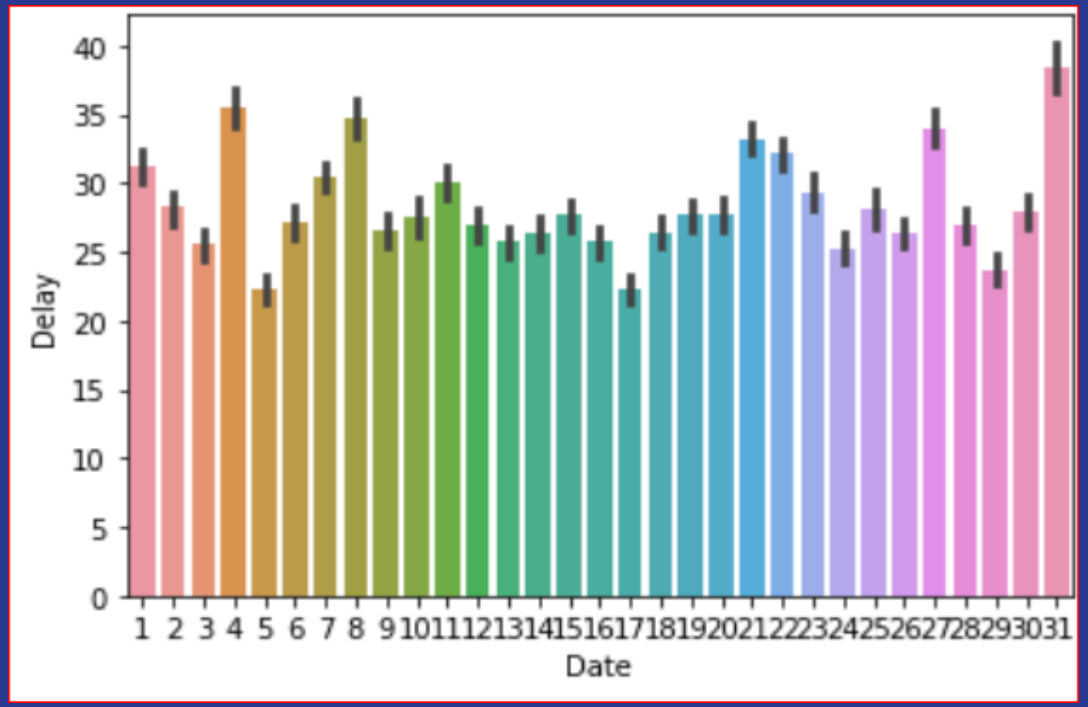
Exploratory Data Analysis





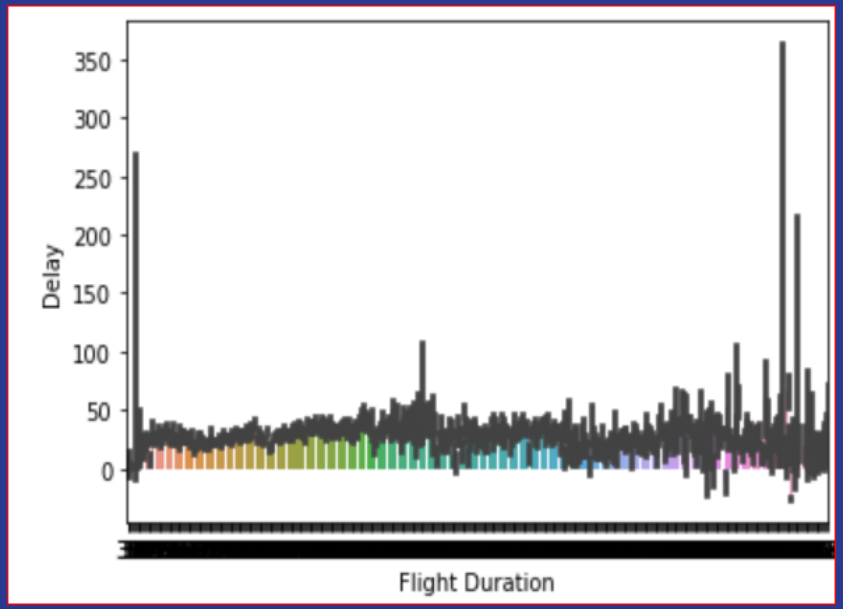
Relationship between Parameters and Target variable

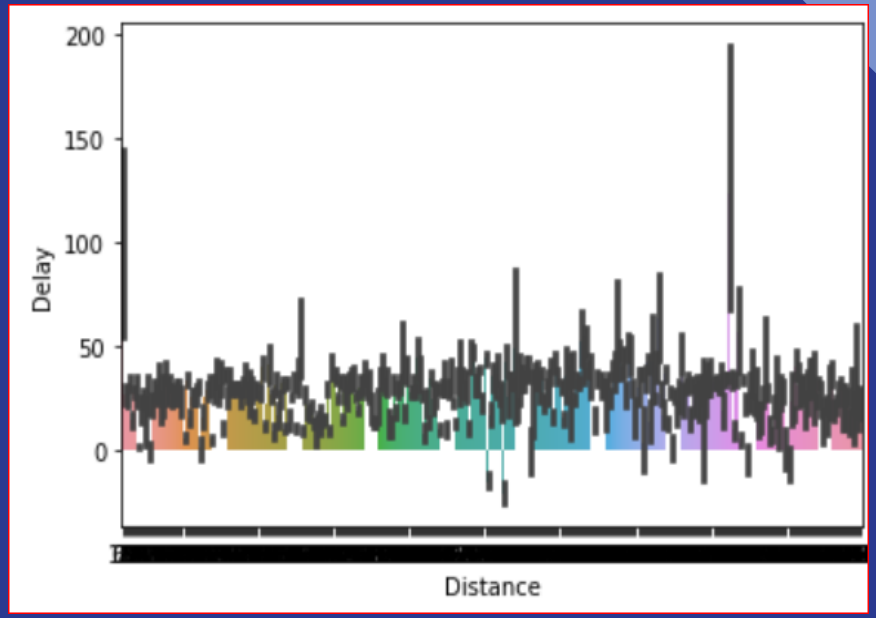
Exploratory Data Analysis



Relationship between Parameters and Target variable

Exploratory Data Analysis

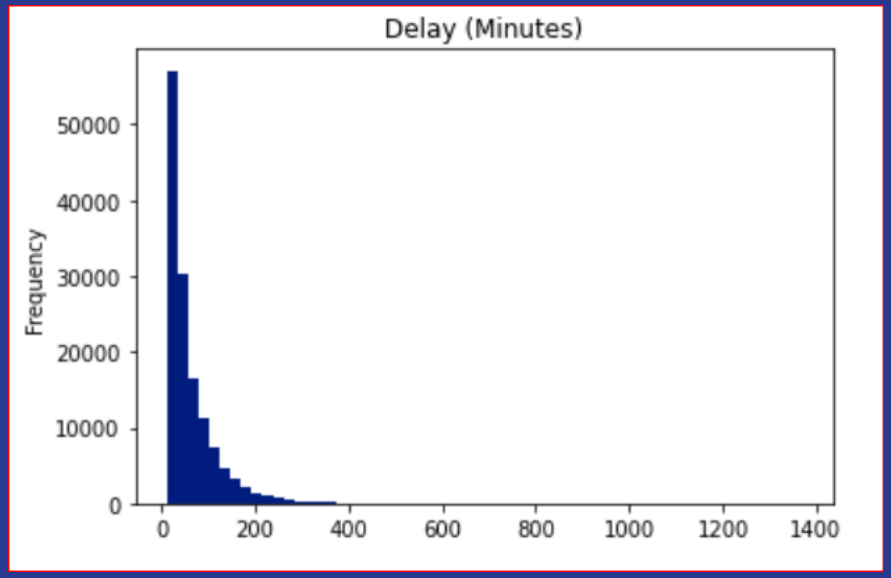




Influence of few features

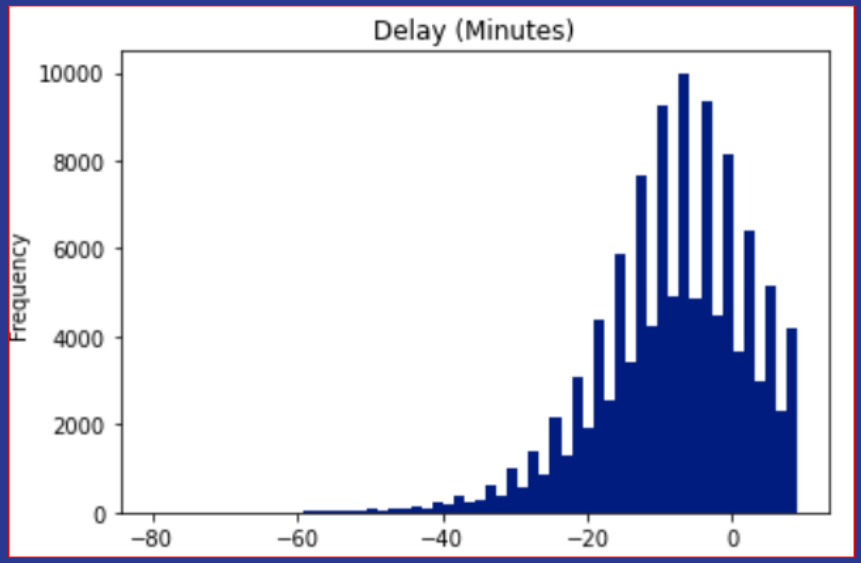
Exploratory Data Analysis

We have taken Delay threshold as 10 and analyzed the frequency of flights more and less than the threshold

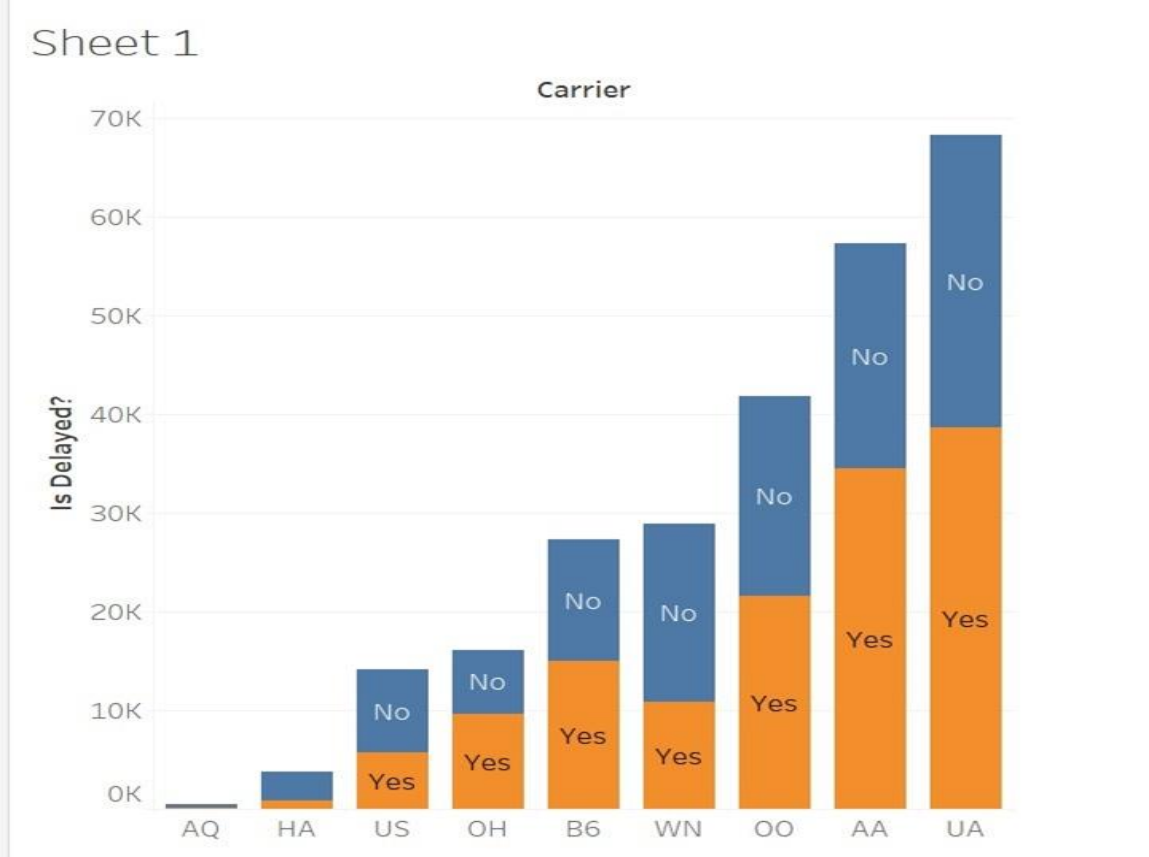


With Delay

Without Delay

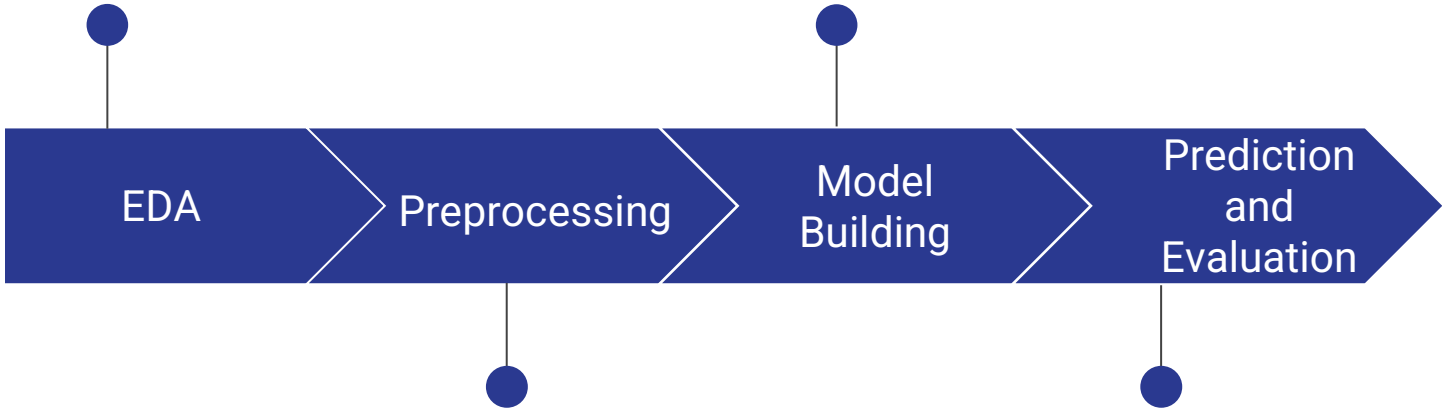


Exploratory Data Analysis



- Correlation between parameters.
- Plotting null values.
- Relationship between parameters and target variable.
- Influence of parameters on target.

- Splitting of data into train and test.
- Builded various Models based on supervised ML techniques



- Removed Null Values
- Data Wrangling
- Identified delayed threshold

- Using the ML Models for prediction, Performance analysis

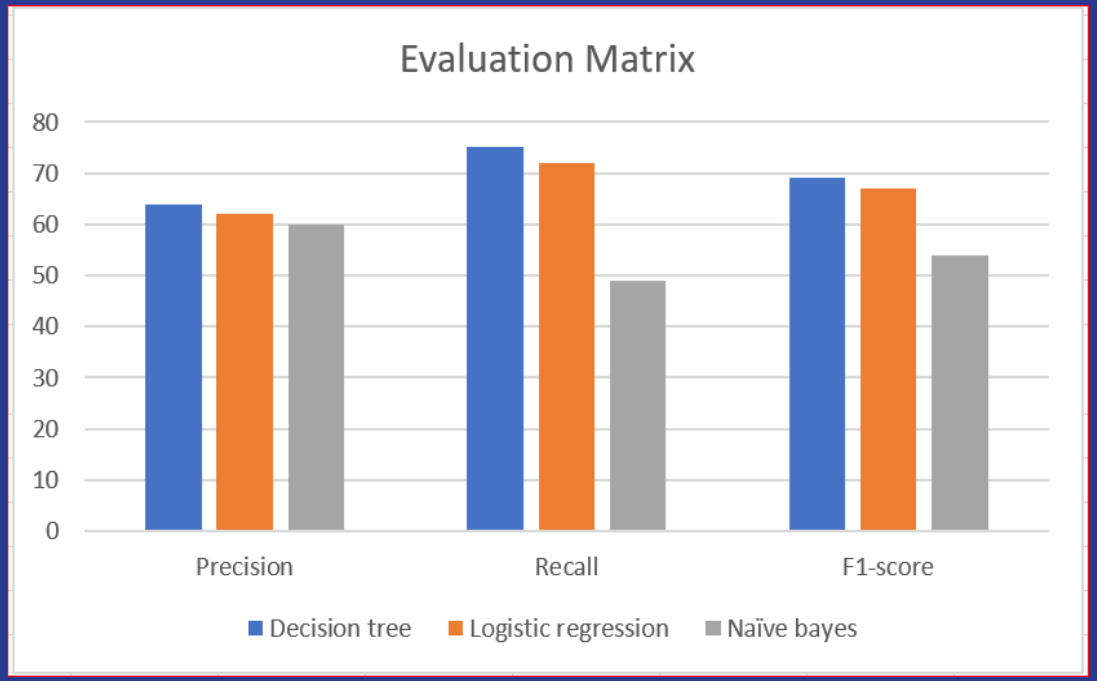
Model Building

Dataset	Algorithms Used
On Time Prediction	Decision Tree, Logistic Regression, Naive Bayes, Random Forest
Sentiment Analysis	Logistic Regression

Model Implementation

- Setting the threshold for delay feature
- Assigning the target label (0,1) based on threshold
- Positive class : 1 : Delay
- String Indexer : to convert a string column of labels to an ML column of label indices
- Vector Assembler
- Fitting Supervised learning models : Decision Tree, Logistic Regression, Naive Bayes, Random Forest

Results and Performance Analysis



- Confusion matrix
- Cost Analysis :Based on the label class False Negative is more costly : Recall
- Logistic regression and Decision tree performs well

Thank You!