PROJECT REPORT

CMPE 255 -Data Mining

Building a Flight Delay prediction model using ML algorithms
Team 3

TEAM INFORMATION

Sakshi Ahuja	- 015266823
Monalisha Parida	- 014637844
Yamini Aalla	- 014331018
Lingxiang Hu	- 015230631

TABLE OF CONTENTS

1. Introduction

1.1 Motivation:

Flight delay causes major exasperation in airports. It not only causes financial losses to airline companies but also brings dissatisfaction among passengers. Generally, flight delays represent the period by which the flight is late or cancelled [1]. Flight delay can be considered as a regression problem and as well as a classification problem. In our project we have implemented various machine learning algorithms to predict whether the flight is getting delayed.

1.2 Objective:

We aim to build a regression model for predicting the flight arrival delay considering various features which show a strong relation with arrival delay. We have also tried to frame the arrival delay and departure delay as a classification problem and used different classification algorithms that are taught in the class.

2. System Design and Implementation

2.1 Algorithms:

For Predicting the Flight Delay we have used the following algorithms:
Regression Algorithms:

- Linear Regression
- Ridge Regression
- Lasso Regression
- Multi Layer Perceptron
- Random Forest Regression
- Decision Tree Regression

Classification Algorithms:

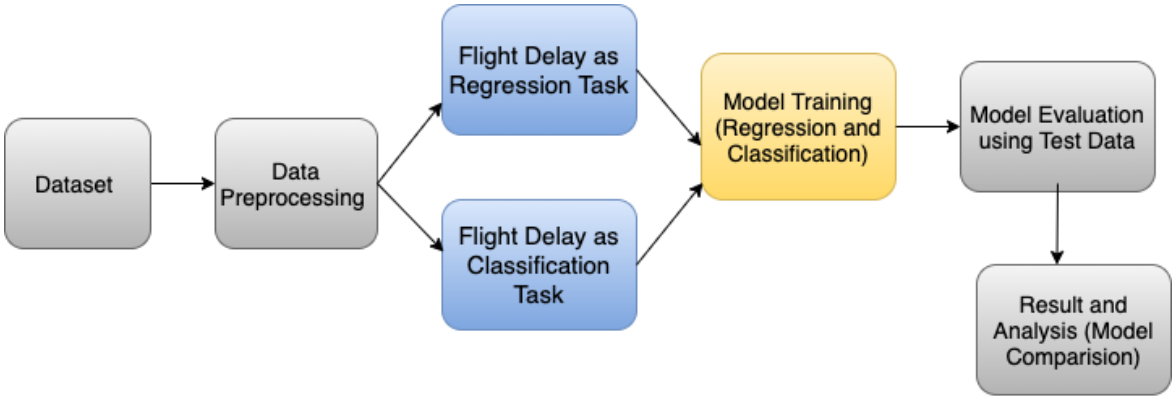
- Decision Tree Classification
- Naive Bayes Classifier
- Multinomial Logistic Regression
- Random Forest Classification
- SGD classifier
- Logistic Regression
- Gradient Boosting Classifier

2.2 Technology and tools used:

Programming Language	:	Python
Frameworks	:	Pandas, Numpy, Sklearn, Seaborn, Matplot Library
IDE	:	Jupyter Notebook, Google Collab
Version Control System	:	GIT

2.3 Architecture and System Design:

Below architecture diagram shows the flow of our project. We have considered the flight delay problem both as a regression and classification task and tried to evaluate the performance of each model.



Architecture Diagram

3. Experiments/ Proof of concept evaluation

3.1 Dataset Details:

Dataset used for this project is obtained from Kaggle (https://www.kaggle.com/usdot/flight-delays). Flight delay dataset was obtained from the U.S Department of Transportation which tracked the performance of domestic flights operated by large airline carriers.

Dataset contains three separate files for flight details, airline details and airport details. These three files are flight.csv (565MB), airlines.csv(360B) and airports.csv(24KB). The Flight.csv dataset contains 5819079 rows and 31 columns.

Below image shows the features from the datasets which are used for predictions and classification.

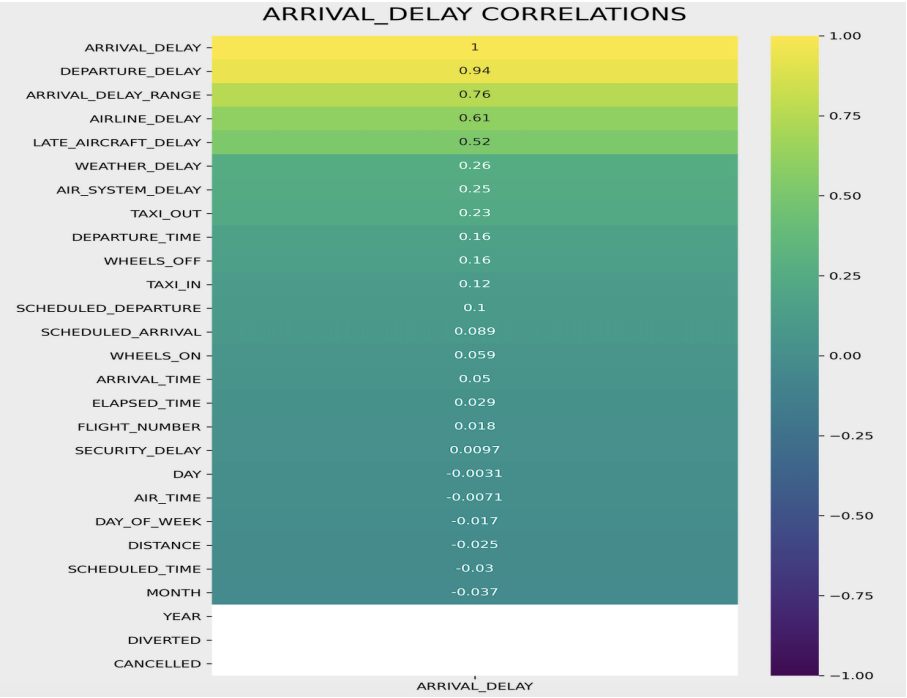
Data Fields	Description
Airline	Airline Identifier
Origin Airport	Starting Airport Name
Destination Airport	Destination Airport Name
Scheduled Departure	Planned Departure Time
Departure Time	Flights Information
Departure Delay	Flights Information
Arrival Delay	Arrival Time – Scheduled Arrival
Scheduled Time	Amount Of Time Planned for trip
Elapsed Time	Air Time+ Taxi Out + Taxi In

Data Exploration:

For Regression:

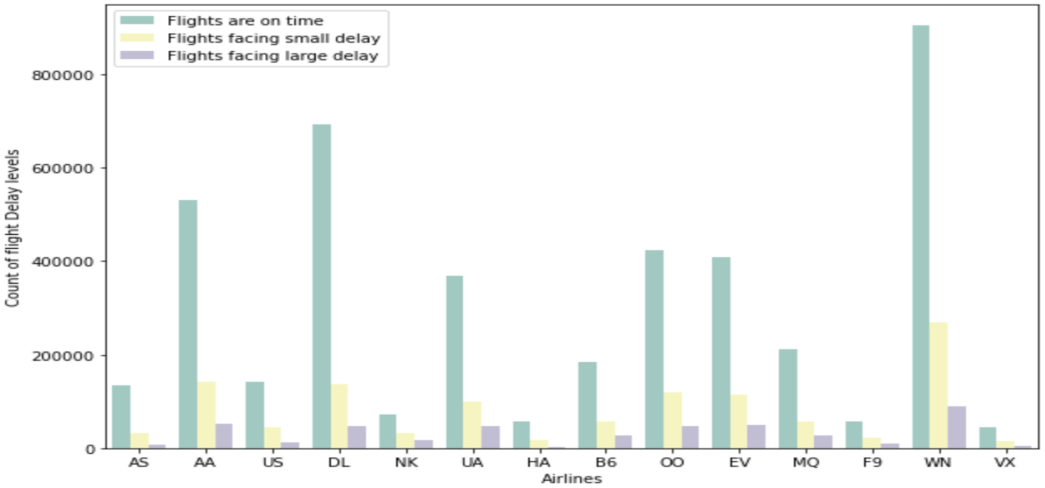
There are a total of 31 features in our flights.csv file. But not all columns are necessary to be used while predicting flight delay. We have found the correlation between the features in our dataset. Correlation generally evaluates the strength of relationship between two features. In our case, we have tried to find out the correlation between ARRIVAL_DELAY and rest of the features. Then we have dropped the unwanted features which are not highly correlated to ARRIVAL DELAY.

- The below graph shows correlation between Arrival Delay and each of the features in the flight dataset.



- We have observed that Arrival Delay is highly correlated with Departure Delay,Arrival Delay,Late_Aircraft_Delay,AirSystem_Delay,Weather_Delay and Taxi_out.
- Format Time and convert to Timestamp format for these columns: SCHEDULED_DEPARTURE,DEPARTURE_TIME,SCHEDULED_ARRIVAL,ARRIVAL_TIME
- After merging airports and airlines file to flights dataset renamed the IATA code for airports and airlines to avoid confusion and check their unique ids.
- Encoded categorical features Airline, Origin Airport,Destination Airport to 0 to classes-1 using label encoder.

- The below graph shows relation between Airlines and Arrival Delays:

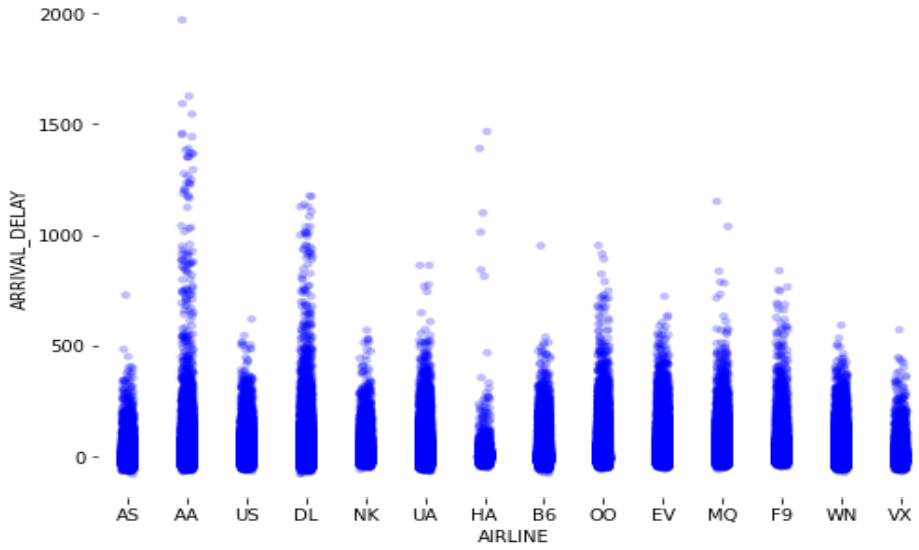


- We have divided arrival delays into three ranges less than five minutes, five to forty five minutes and greater than forty five minutes for each airline and found an interesting relation: a high percentage of airlines have less than five minutes and very few percent have large delays.

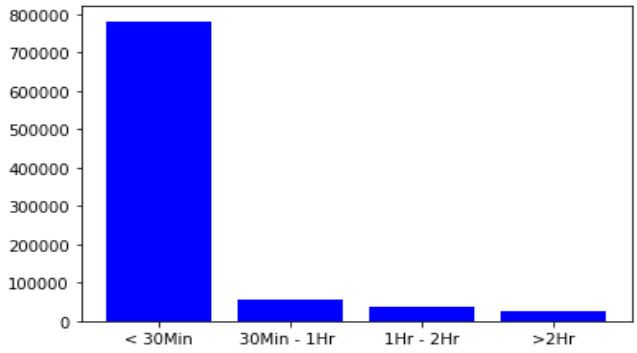
Data Preprocessing for Classification Algorithms:

As we don't have the infrastructure to process such a large dataset, we have considered a subset of dataset which contains January & February of 2015. It has 899159 flight information.
The input data has 31 attributes which are not relevant always. As a first step we wanted to know which are the important variables that 'ARRIVAL_DELAY' is mostly correlated to. The correlation with the rest of the attributes gave us a clear Idea about the major attributes that impact the Arrival delay.

- The below graph shows the Arrival delay for all the air lines. As we can see the average delay for all the airlines is close to 7 minutes.



- The attributes that are poorly correlated have less influence on Arrival delay. Based on correlation score we dropped those columns from our data set.
- The data matrix that we have is sparse as for many flights we don't have information for all the selected attributes. We replaced those unavailable data points by the mean of that attribute over the entire data set.
- As we wanted to frame the flight delay as a classification problem we binned the Arrival delay into 4 classes.
    1. Flights with Arrival delay less than 30 min
    2. Flights with Arrival delay between 30 Min to 1Hr
    3. Flights with Arrival delay between 1Hr to 2Hr
    4. Flights with Arrival delay more than 2Hr



Data Preprocessing for Classification Using the Departure Delay :

- After analysing the data for flights , we could see that the data contained numerical as well as categorical values . The most important column in this problem is 'DEPARTURE_DELAY' and the models will predict if there will be a delay in departure.
- Other than this column , there is a column 'CANCELLED 'which shows if the flight is cancelled or not . We tried to drop the rows where flights are cancelled because in the case if flights are cancelled, there will be no delay.
- After this, the labels were created out of 'DEPARTURE DELAY' column keeping it 1 if there is a delay and 0 if there was not.
- We checked for origin and destination airports and loaded the airports data and checked if the airports mentioned in the flights data are contained in the airports data.
- Those features were dropped that do not contribute to the departure delay prediction.
- The features which contributed most to the departure delay were extracted and it contained numerical as well as the categorica;l features , we separated those features and applied one-hot encoding using get_dummy function to convert them to Binary form.

3.2 Methodology

Regression Models:

Ridge Regression:
Ridge is a regularization technique that reduces overfitting. Ridge regression uses L2 regularization. We are predicting the arrival delay for the flights considering the features highly correlated to arrival delay.we have considered the Regularization value as 1.0.

Lasso Regression:
Ridge is a regularization technique that reduces overfitting. Ridge regression uses L1 regularization. We are predicting the arrival delay for the flights considering the features highly correlated to arrival delay.we have considered the Regularization value as 0.25.

Multi Layer Perceptron:
Multi-layer perceptron is a deep learning methodology. This deep neural network learns non- linear functions. It has hidden layers.The layers of input nodes produce different sets of outputs. We are using this algorithm to learn the non-linearity factor in the data. We have used rectified linear unit func-

tion(relu). The value of the regularization parameter alpha is 0.03. The maximum number of iterations taken is 600.

Decision Tree Regressor:
Decision tree regression is a supervised learning which splits the data into a format of tree based on test condition given.We have chosen maximum depth of tree as 8.

Gradient Boost Regressor:
Gradient boosting is also an ensemble method. It produces a strong learner through a combination of weak learners in an iterative fashion. We are using this model because the accuracy of the predictive results is higher as it minimizes the predictive error altogether.We considered maximum depth as 3, no of estimators as 100 and minimum sample split as 2.

Random Forest Regression:
Random forest is one of the ensemble methods which performs both Classification and Regression, where estimators are a large number of small independent decision trees.We have considered the no of estimators as 50 and maximum depth of tree as 7 for predicting the arrival delay.

Classification Models:

Decision Tree Classification:
Decision Tree classification is a supervised machine learning algorithm where we split the data based on certain test conditions that produces the highest information gain. We stop the splitting of data if all records belong to the same class or have identical attribute values. Depth of the decision was set to 8 for which we got the optimal result.

Naive Bayes Classifier:
It is a probabilistic algorithm used to solve classification problems. Naive Bayes algorithm considers independence among attributes.Here we have used GaussianNB classifier & that may be the reason having poor performance as most of the arrival process follows poisson distribution.

Multinomial Logistic Regression:
Multinomial Logistic Regression is an enhancement of logistic regression. It is best for considering greater than one categorical features as output parameters.Similar to logistic regression, multinomial logistic regression also tries to predict the probability for classification problems. For our case we have used L2 as the penalty term and used 100 iterations for the algo to converge.

Random Forest Classification:
Random forest Classifier fits multiple decision trees and aggregates the results which helps in improving the accuracy and prevents the conditions of overfitting . These classifiers work better than decision trees.In this , max_depth parameter was set to 6.

Logistic Regression:
Logistic regression is a predictive analysis method. It is a linear model used for classification and before applying it to this model we scaled our data .This model describes the data and tells the relationship between variables.The logistic regression model worked best with the default parameters.

Stochastic Gradient Descent Classifier:

Stochastic Gradient Descent is a classification model for fitting the linear models . Both logistic and SGD work similarly and are used to fit the linear models but SGD does this in small batches and hence it executes faster . For this the loss parameter is logarithmic and alpha value is 0.1.

Gradient boosting Classifier:
Gradient Boosting classifier is also an ensemble based classifier . Gradient boosting is similar to Random forest and used to improve the overfitting . At each stage, trees are created and fit to the model to improve the accuracy score . The best parameters for this models were n_estimators=100 and learning_rate=1.0

Flight Arrival Delay as Regression problem:
After data cleaning and feature selection we have 1063439 rows and 13 features which are highly correlated with arrival delay. We have split the dataset to train and test in the ratio of 70:30 respectively. After splitting the data we trained several models like random forest, decision tree, gradient boost, lasso, MLP and ridge. We evaluated the models considering RMSE,MAE and R2 scores.

Flight Arrival Delay as Classification problem:
After data preprocessing, the dataset contains 899159 rows and 12 features. As we wanted to frame the flight delay as a classification problem we binned the Arrival delay into 4 classes. To verify that all our models are being generalised, we divided the input data into Training and Testing data with 70:30 split. We used 5 machine learning models like Decision Tree, Naive Bayes, Multinomial logistic regression, Random Forest classifier and XGBoost Classifier. We have evaluated these models by calculating precision, recall, f1 score and support and found that XGBoost gave better accuracy compared to others.

Flight Departure Delay as Classification problem:
After data preprocessing, total data contained 457986 rows and 324 features.
For fitting to the model, the data set was separated into train-validation-test data as 70%-15%-15%. After splitting, the models were fit with the training data and predictions were made on the validation data and the model was then evaluated with the testing data and metrics were calculated .

3.3 Analysis of Results:

Evaluation Metrics:

For evaluating the performance imported root mean square error and mean absolute error,r2 score from sklearn.
Mean Absolute Error(MAE): MAE averages the absolute difference between actual predicted values.
Root Mean Square Error(RMSE): RMSE is the square root of mean square error. Where mean square error is the difference between actual and predicted values.It ranges from 0 to infinity and the rmse score is less for small error.
R2 Score: It is a regression score function which takes actual and predicted values and calculates the closeness of data to a fitted regression model.it ranges from 0 to 100 percent.

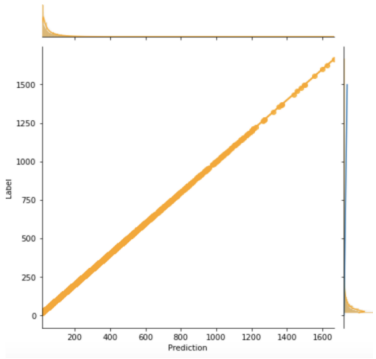
For classification models:
Accuracy Score: It predicts whether the number of true positives and true negatives to the total data . Accuracy Score helps us in predicting how often our model predicts correctly .

AUC score: AUC score represents the probability that a model provides a higher rank to positive samples than negative samples.
Precision Score : Precision helps in predicting when the model provides the true positives, how often these are correct.
Recall Score : It is very helpful when the cost of false negatives is very high.
F1 Score: It is a combined measure of precision and recall and optimise the evaluation metric.

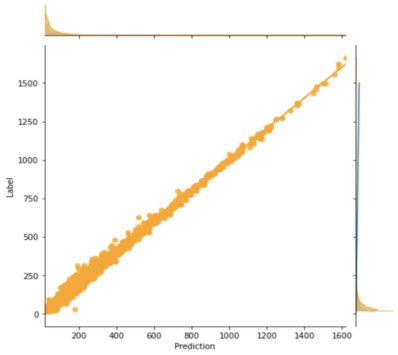
Analysis for Regression Models:

The Graphs Represent Prediction of models vs actual values for test data.

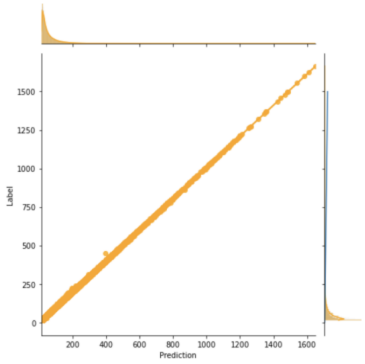
Ridge Regression



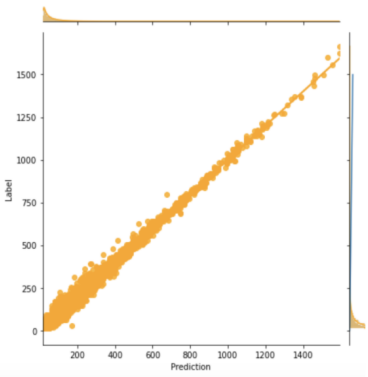
RandomForest Regression



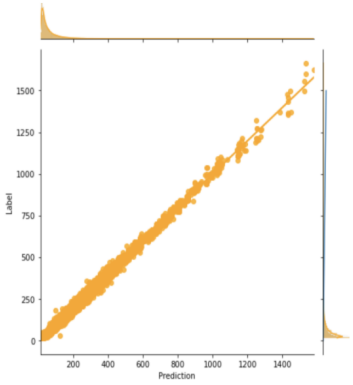
Lasso Regression



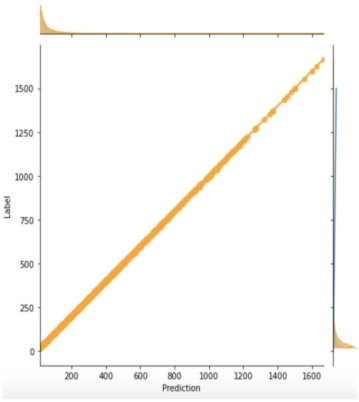
Decision Tree Regression



Gradient Boost Regression



MLP Regression



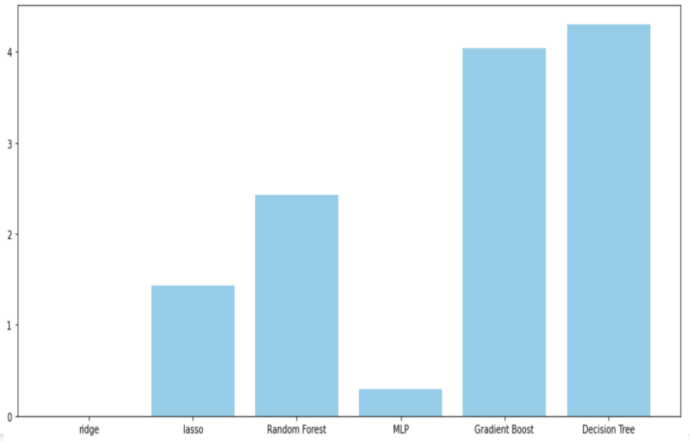
Comparison of Regression models:

- Root mean square error and R2 score were represented in the table below where it's evident that Ridge Regression with regularization value of 1.0 out performs other models by giving best RMSE and R2 scores.
- The graph represents models on x axis and RMSE scores on y axis where the Decision tree was giving the highest RMSE score and Ridge regression was giving very low RMSE score.

RMSE and R2 values for all the Models

ALGORITHMS	RMSE	R2 SCORE
Random Forest Regressor	1.428156	0.998560
Multi Layer Perceptron(MLP)	1.983941	0.999978
Decision Tree Regressor	4.043684	0.995484
Lasso Regression	0.298147	0.99950
Ridge Regression	0007591	0.99999
Gradient Boost Regression	2.425940	0.996000

MODELS VS RMSE SCORES

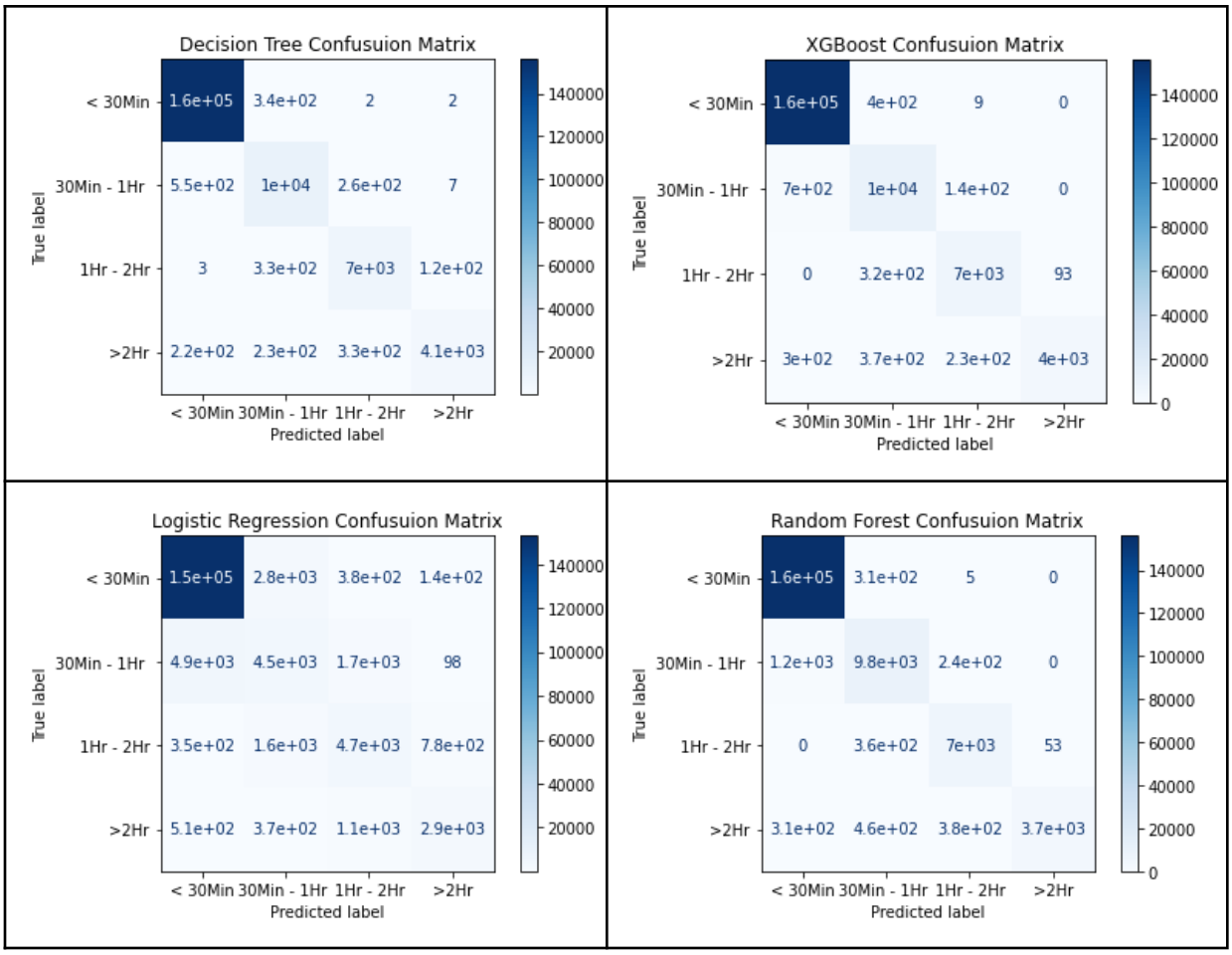


Analysis for Arrival Delay Classification:

Looking at the comparison table it seems all the models generalize well but XGBoost produced better accuracy compared to other models. The performance of Naive bayes is low because of the gaussian approximation (for prior) that we have taken. Most of the arrival/delay processes usually follow Poisson's distribution and as a next step we will explore the Naive Bayes with Poisson distribution being as a prior for the model.

	Model	Train_Accuracy	Test_Accuracy
4	XGBoost	0.986	0.986
0	Decision Tree	0.979	0.979
1	Random Forest	0.976	0.976
2	Logistic Regression	0.919	0.918
3	Naive Bayes	0.891	0.892

Below confusion matrix shows the Predicted Label and Actual Label for all 4 classes.

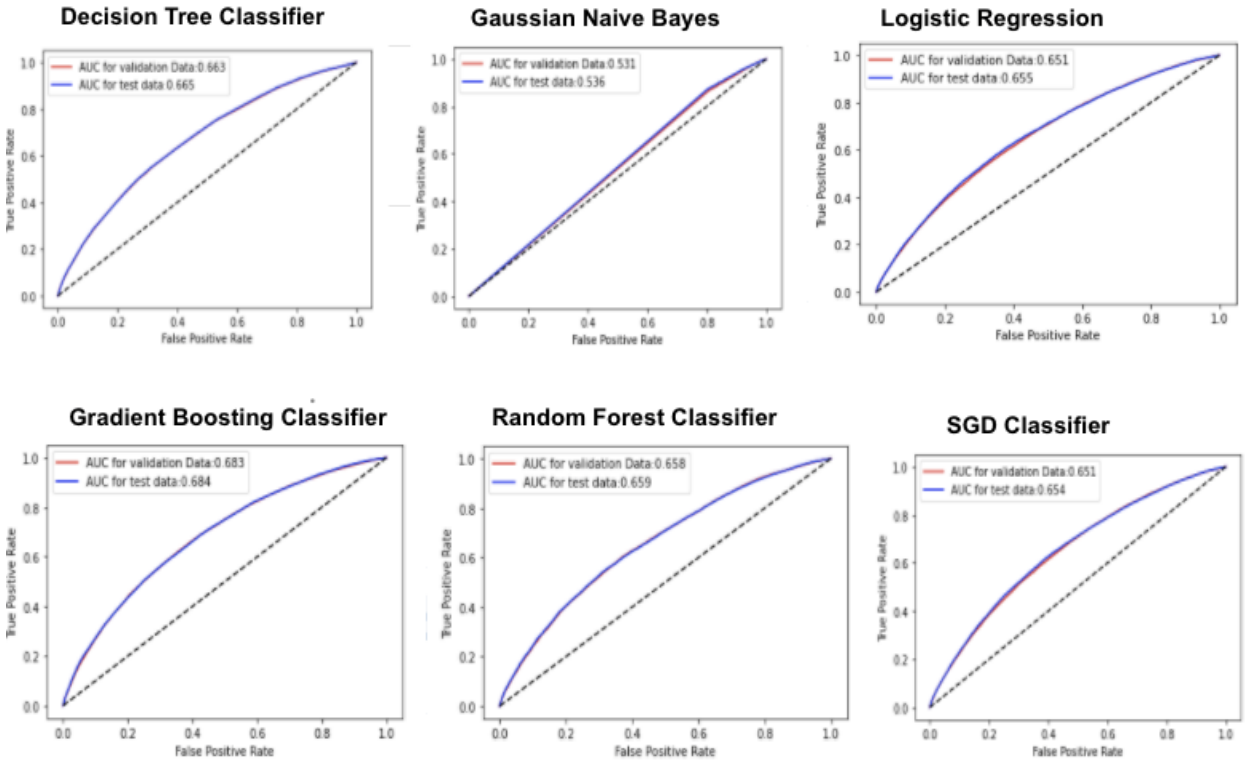


Analysis for Departure Delay classification:

Out of these models, we observed that the Gradient boosting Classifier outperformed the other models. We were able to accurately predict 66 % of the departure delays .

Algorithms	AUC Score	Accuracy	Precision	Recall	F1 Score
Logistic Regression model	0.655	0.644	0.575	0.295	0.390
Decision Tree Classifier	0.665	0.652	0.602	0.288	0.288
Gaussian Naive Bayes	0.536	0.437	0.399	0.914	0.556
SGDClassifier	0.654	0.641	0.597	0.214	0.315
Random Forest Classifier	0.659	0.620	0.757	0.021	0.042
Gradient Boosting Classifier	0.684	0.662	0.596	0.377	0.462

Following are the ROC curves for the above models.



4.Discussion and conclusion

4.1 Decisions made & Difficulties encountered

- The number of features for the flight dataset were 31 to improve the performance of the model we narrowed down to features which are positively correlated to delays.
- The order of the data preprocessing steps affects the data exploration.
- The classification models with all the features in the dataset were reducing the AUC score as well as accuracy. Selecting the required features was a difficult task.

4.2 Things that didn't work well

- For regression analysis we have selected Decision tree and Gradient Boost regression algorithms which are giving high RMSE values.
- Tried regression models for calculating the departure delay, but ended up having the results with huge error scores.
- Due to the huge dataset for some models, the notebook was crashing.

4.3 Conclusion

Through this project, we created classification and regression models for predicting the departure and Arrival Delays. For the Classification model for Departure delay, many models were evaluated and we found that the Gradient Boosting algorithm worked well. For the Classification model for Arrival delay,

XGBoost classifier is producing higher accuracy compared to other models. For regression models, ridge regression is performing best compared to other models.

5. Project Plan and Task Distribution

SJSU ID	NAME	TASK
015266823	Sakshi Ahuja	Data Cleaning, Data Analysis, Classification model for flight Departure delay(Model1), Project Report,Project Presentation
014331018	Yamini Aalla	Data Cleaning, Data Analysis, Regression model for flight Arrival delay(Model3), Project Report, Project Presentation
014637844	MonaLisha Parida	Data Cleaning, Data Analysis, Classification model for flight Arrival delay(Model2), Project Report, Project Presentation
015230631	Lingxiang Hu	Data Analysis

6.Reference :

[1] **V. Natarajan, S. Meenakshisundaram, G. Balasubramanian and S. Sinha, "A Novel Approach: Airline Delay Prediction Using Machine Learning,"** ***2018 International Conference on Computational Science and Computational Intelligence (CSCI)*****, 2018, pp. 1081-1086, doi: 10.1109/CSCI46756.2018.00210.**

[2]**https://pandas.pydata.org/docs/**

[3]**https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_curve.html**

[4]**https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html**

[5]**https://scikit-learn.org/stable/modules/generated/sklearn.metrics.recall_score.html**

[6]**https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_auc_score.html**

[7]**https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html**

[8]**https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html**

[9]**https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html**

[10]**https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html**

[11]**https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html**

[12]**https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html**

[13] SOURCE CODE: https://github.com/Sakshisjsu/CMPE-255Project