Flight Delay Detection

Abstract

This project focuses on leveraging Machine Learning techniques to predict flight arrival delays, by utilizing a comprehensive dataset comprising weather and flight-related information across 15 airports (refer to Table 1). A pipeline containing a classifier to predict if there is an arrival delay or not, and a regressor to predict the period of delay is implemented and analysed. Different classification and regression models are trained on the data and the model that performs the best is considered. The F1 score of the classifier and the R2 score of the regressor were observed to be 0.85 and 0.81 respectively.

1 Introduction

Flight delays can have significant impacts on both passengers and airlines, and the weather condition is a potential factor causing them. Predicting flight delays accurately can help airlines manage their schedules more effectively and assist passengers in making informed travel decisions. By combining flight and weather data and feeding it to a Machine Learning model, the efficiency in predicting and analysing flight delay patterns can be improved. The data pertaining to the below listed airports (Table 1) are considered for implementation.

ATL	CLT	DEN	DFW	EWR
IAH	JKF	LAS	LAX	MCO
MIA	ORD	PHX	SEA	SFO

Table 1: Airport codes

2 Dataset Description

The weather data contains the hourly weather condition specifics on every day in each month, at every airport. The data is available for the years 2013-2017. Details like *windspeed, dew point, temperature*, etc. among the others shown in table 2 below, are present.

WindSpeedKmph	WindDirDegree	WeatherCode	precipMM
Visibilty	Pressure	Cloudcover	DewPointF
WindGustKmph	tempF	WindChillF	Humidity
date	time	airport	

Table 2: Weather data

The flight data contains details regarding a flight's arrival, departure, airport code, origin airport, destination airport among others, for the above mentioned 15 airports. The dataset contains various features like *Origin, Destination, OriginAirportID, DestAirportID, CRSArrTime,CRSDepTime* etc. among the others, as mentioned in table 3 below.

FlightDate	Quarter	Year	Month
DayofMonth	DepTime	DepDel15	CRSDepTime
DepDelayMinutes	OriginAirportID	DestAirportID	ArrTime
CRSArrTime	ArrDel15 (label)	ArrDelayMinutes (target)	

Table 3: Flight data

3 Pre-processing

Data related to flights and weather conditions are acquired from different sources, corresponding to the years 2016 and 2017 and the 15 airports mentioned, and the features are identified upon parsing and analysing.The flight dataset contains about 110 features, out of which only the above mentioned 15 features are considered, as these features have more meaning and information to explain the target variable considered, than the others. Values of the required features are stored in a data frame, one each for flight and weather data.These individual datasets are then merged based on the 'time' feature in both the datasets.

The resultant data frame is checked for null values which are to be dropped, following which the categorical variables present are subjected to one-hot encoding. The data is analysed for presence of correlation among the feature variables using correlation matrix (represented as a heatmap in Figure 1). Of the highly correlated columns, one among those is retained, while others are dropped. This pre-processed data is then used for the classification and regression tasks, as explained later.

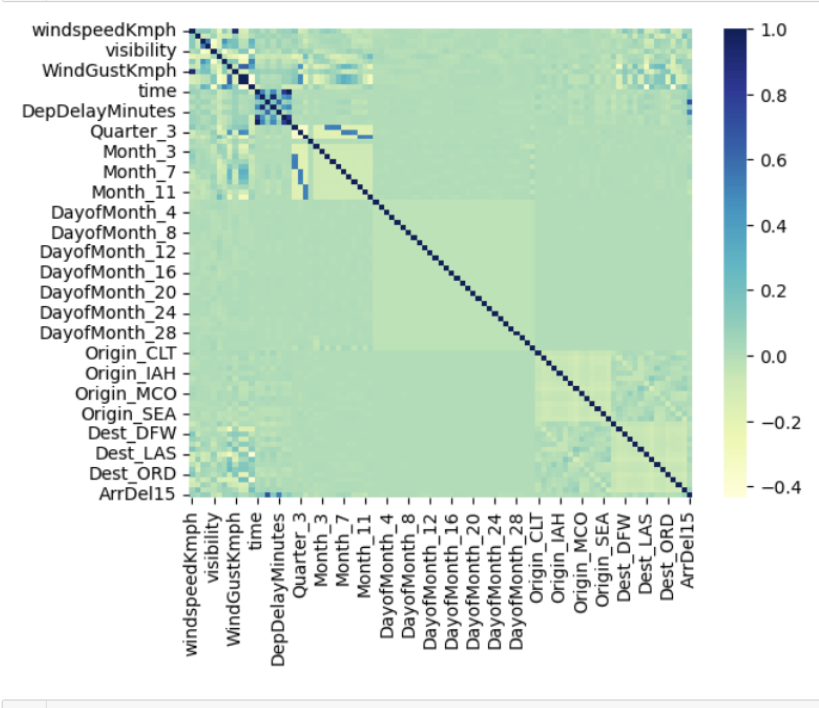

Figure 1: Correlation Heatmap

4 Classification

4.1 Definition

Classification is a supervised machine learning method, where the model tries to predict the correct label (or class) of a given input data. In classification, the model is fully trained using the training data, and then it is evaluated on test data before being used to perform prediction on new unseen data. It gives a discrete value as output. Various machine learning models used for this task are logistic regression, Naïve Bayes, Support Vector Machine, Decision Tree, Gradient Boosting, etc.

4.2 The Class Imbalance Problem

A class imbalance problem in a classification problem occurs when there is an imbalance in the number of labels belonging to each class in the data. This results in the model being biased, (i.e.) misclassification of the minority class as the majority class, which can lead to inaccurate predictions. This problem is crucial, especially in an anomaly detection task, and needs to be handled before the data is given to the model for training. Resampling techniques like ran-

dom under-sampling, random over-sampling, Synthetic Minority Over-Sampling Technique (SMOTE), etc. are used to handle this condition.

In this data, there are two target classes: class 0 which denotes *NoDelay* and class 1 which denotes *Delay*. As about 80% of the samples belong to class 0 (majority class), a class imbalance problem is observed (80:20) as shown in Figure 2 below and thus, techniques like under sampling, over sampling, SMOTE, etc. are applied on the training data. As the results of the resampling techniques are almost similar, the under-sampled data is used for training.

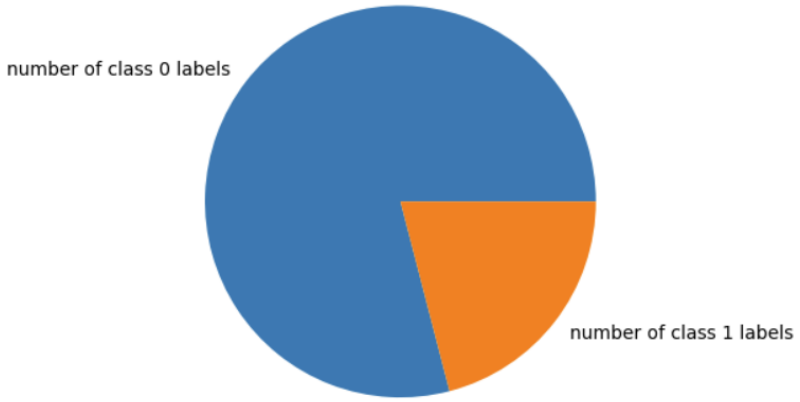

Figure 2: Class Imbalance Visualization

4.3 Model training and Performance Analysis

The data is fit with different classification models like logistic regression, Extra Tree Classifier, Random Forest Classifier, Decision Trees and XGBoost Classifier. The under-sampled data is given to the models for training, and the corresponding accuracy, precision, recall, and F1 score, for each model is as tabulated below (Table 4):

Model	Accuracy	Precision	Recall	F1 score
Random Forest Classifier	0.90	0.86	0.85	0.85
XGBoost Classifier	0.90	0.86	0.85	0.85
Logistic Regression	0.88	0.81	0.86	0.83
Extra Trees Classifier	0.87	0.80	0.86	0.82
Decision Tree Classifier	0.78	0.71	0.79	0.73

Table 4: Performance of various Classification Models

Accuracy considers only the ratio of samples correctly classified over the entire data. But considering the accuracy as a metric can be misleading for data with class imbalance problem. Precision focuses on minimising false positive errors, but doesn't consider false negatives. On the other hand, recall penalises false negatives heavily but does not account false positives. Thus, the **F1 score**, which considers both precision and recall, is used as the performance evaluation metric. Upon analysis, it is observed that the Random Forest Classifier gives the maximum score and the Decision Tree Classifier gives the least score for the given data. The Random Forest Classifier has the best Accuracy, Precision, Recall and F1 scores.

5 Regression

5.1 Definition

Regression is a technique used to estimate the relationships between a dependent(target) variable and one or more independent(feature) variables, using which the value of the target variable for a given set of feature variables is predicted. Here the period of delay is the target variable to be predicted from the available features.

5.2 Model Training and Performance Evaluation

The period of delay for the data samples classified to have a delay in arrival is calculated and appended to the existing data. This data is fit with various regression models like Linear Regressor, Extra Trees Regressor, Random Forest Regressor and XGBoost Regressor, and the corresponding R2 scores and mean absolute errors are as tabulated below(Table 5):

Model	R2 score	MAE
Extra Trees Regressor	0.809	86.60
Random Forest Regressor	0.783	96.104
XGBoost Regerssor	0.701	140.54
Linear Regression	0.154	343.22

Table 5: Performance of various Regression Models

R2 score and Mean Absolute Error (MAE) are considered as performance evaluation metrics. R2 score is supposed to be as close to 1 as possible and MAE is to be as less as possible. Upon analysis, it is observed that the extra trees regressor gives the maximum R2 score (and least MAE) and Linear Regression gives the least R2 score (and greatest MAE) for the given data.

6 Pipeline

The best models of classification and regression, from tables 4 and 5, are applied on the entire dataset after training, treating it like the test data.
These models are used to build a pipeline whose structure and function is as follows (Figure 3):

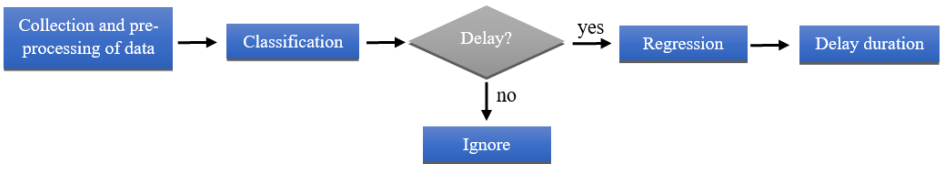

Figure 3: Pipeline

The pre-processed data is taken and fed to the classifier where the datapoints are classified into classes 0($NoDelay$) and 1($Delay$), and the datapoints that are classified as 1, are passed on to the regressor to predict the delay duration in minutes.

7 Regression analysis

As the number of datapoints with delay more than 2000 minutes are negligible, the following ranges of period of delay from 15 to 2000 are considered, and the performance of the regressor for values in each range is observed and analysed based on MAE values. The datapoint distribution in each range is given in Figure 4 and the Mean Absolute Error values are tabulated in Table 6 as shown below:

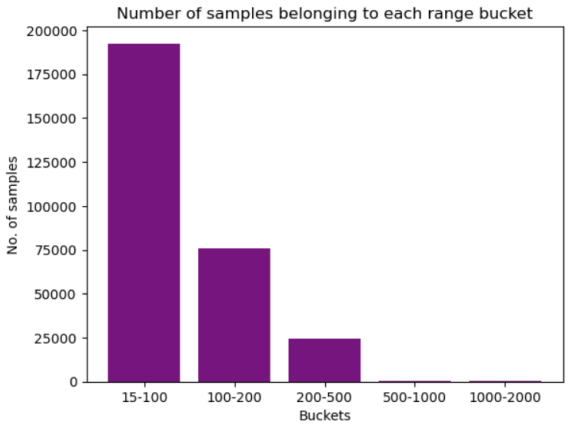

Figure 4: Number of datapoints in each range

Interval Range	MAE
15-100	18.859765
100-200	21.783713
200-500	29.931543
500-1000	63.126052
1000-2000	1426.684866

Table 6: MAE values

It can be observed that, the number of datapoints with the period of delay between 15 minutes to 100 minutes is the maximum, and the number of datapoints with the period of delay between 1000 minutes to 2000 minutes is the minimum. The performance of the regressor differs in each interval range, and it performs the best for the values in the range between 15 and 100. The performance declines as the value approaches the range between 1000 and 2000.

The least MAE score occurs at the range 15-100, and the greatest score occurs in the range 1000-2000. This implies that the predictions are the most accurate when the period of delay is between 15 to 100 minutes and the least accurate when the delay iis between 1000-2000 minutes. Thus, in the range 15-100, the delay can be accepted with high confidence, and mitigation measures can be accordingly framed and implemented. As the error in prediction increases for subsequent intervals, the risk increases and the confidence decreases and so, suitable risk management along with mitigation measures must be scrutinized.

8 Conclusion

Thus, the flight and weather data acquired were pre-processed using appropriate techniques, and the individual datasets were merged using features common to both datasets. The data was observed to have a Class Imbalance, which was handled by applying the under-sampling technique.

The data is given to various classifiers and regressors, out of which the **Random Forest Classifier** and the **Extra Trees Regressor** gave the best score for their respective tasks. These models are used to build a pipeline, in which the pre-processed data is fed to the Classifier to predict if there is a delay or not, and the data points classified to have a delay are fed to the Regressor to predict the duration of delay. The **Classifier** performed with a **F1 score of 0.85** and the **Regressor** performed with a **R2 score of 0.81** respectively. A Regression

Analysis on the performance of the regressor in various range of delay duration values was performed to evaluate the model's performance in each interval of values.

The performance of the models can further be enhanced by trying different sampling techniques, and trying feature selection techniques like PCA on the data.