Real Time Flight Delay Prediction

Abstract. A flight delay is when an airline flight takes off and/or lands later than its scheduled time. There are a multitude of causes that may cause a flight to deviate from its scheduled timing, but this project focuses on the weather and flight data.

1 Introduction

Air transport, which represents the next most substantial energy-consuming transport sector(First being Road transport), includes passenger and freight airplanes, that is, aircraft configured for transporting passengers, freight, or mail. Thus a higher demand for air transport results in more air traffic thus reducing reliability .

Air transport enables highly perishable and valuable products to be moved fast over long distances, but it lacks the environment control that is possible for other modes. Flight delays are economic, social, and environmental problems that cause inconvenience for both airline companies and passengers . Flight delays not only irritate air passengers and disrupt their schedules but also cause a decrease in efficiency, an increase in capital costs, reallocation of flight crews and aircraft, and additional crew expenses.

This project aims to predict whether a flight may delay or not depending upon environmental conditions at the time of arrival of that particular flight towards its airport and subsequently predict the total number of minutes the flight may be delayed in **minutes** using Machine Learning Algorithms, i.e., Classifiers and Regressors.

2 Data-set

The data-sets used to train the Machine Learning Models are Flight and Weather dataset. The flight dataset contains the schedule, details, arrival, departure etc.., of all the flights that flew inside the USA between 2016 and 2017. On the other hand the weather dataset contains the weather report recorded every 1 hour across the USA .

Table 1: Features selected from flight dataset

FlightDate	Quarter	Year	Month
DayofMonth	DepTime	DepDel15	CRSDepTime
DepDelayMinutes	OriginAirportID	DestAirportID	ArrTime
CRSArrTime	ArrDel15	ArrDelayMinutes	

The flight dataset is preprocessed and fifteen of the best features are selected as referenced as in Table 1.

Table 2: Features selected from weather dataset

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

The weather dataset is also preprocessed and fifteen of the best features are selected as referenced as in Table 2.

The flights are selected from 15 specific airports as referenced in Table 3 and segregated by that basis.After segregating the features from their core datasets, each individual flight is combined with its corresponding weather data at its time of arrival or departure.

Table 3: Airports selected

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

3 Classification

The Classification algorithm is a Supervised Learning technique that is used to identify the category of new observations on the basis of training data. In Classification, a program learns from the given dataset or observations and then classifies new observation into a number of classes or groups. Unlike regression, the output variable of Classification is a category, not a value. Since the Classification algorithm is a Supervised learning technique, hence it takes labeled input data, which means it contains input with the corresponding output.

The Classifiers are trained to classify whether a flight is delayed or not. Arrdel15 is used as the target variable(y-axis),which holds the value 1 if the flight produces a delay of more than 15 minutes and 0 otherwise.The features used in x-axis are selected using feature selection from the features referenced above in table 1 and table 2. The dataset is then split on a 7:3 train over test ratio to train the Classifiers. The **Classifiers** explored in this project to classify are **XGBooost**, **RandomForest**, **DecisionTree**, **LogisticRegression** and **ExtraTrees** Classifiers.

3.1 Evaluation Metrics

There are a lot of metrics to evaluate a Machine Learning Classification Model, but the metrics used to evaluate the Classifiers in this project are **precision**, **recall**, **f1-score**, **accuracy**. A confusion matrix is also used to figure out where the model is confused and therefore making incorrect predictions.

Precision Precision is a metric that quantifies the number of correct positive predictions made.Precision, therefore, calculates the accuracy for the minority class.It is calculated as the ratio of correctly predicted positive examples divided by the total number of positive examples that were predicted.

$$Precision = TruePositives/(TruePositives + FalsePositives) \tag{1}$$

Recall Recall is a metric that quantifies the number of correct positive predictions made out of all positive predictions that could have been made.Unlike precision that only comments on the correct positive predictions out of all positive predictions, recall provides an indication of missed positive predictions.

$$Recall = TruePositives/(TruePositives + FalseNegatives) \tag{2}$$

F1-Score F1-Score provides a way to combine both precision and recall into a single measure that captures both properties.Alone, neither precision or recall tells the whole story. We can have excellent precision with terrible recall, or alternately, terrible precision with excellent recall. F1-Score provides a way to express both concerns with a single score.

$$F1 - Score = (2 * Precision * Recall)/(Precision + Recall) \tag{3}$$

Accuracy Accuracy is one metric for evaluating classification models. Informally, accuracy is the fraction of predictions our model got right. Formally, accuracy has the following definition:

$$Accuracy = NumberofcorrectPredictions/TotalnoofPredictions \tag{4}$$

Confusion Matrix A confusion matrix is a performance measurement technique for Machine learning classification. It is a kind of table which helps you to the know the performance of the classification model on a set of test data for that the true values are known.

The classifiers are evaluated with these metrics and the model which produces the best result for the given dataset is chosen.

3.2 Classification Report

The scores of the classifiers used are tabulated below:

Table 4: Classifier scores

Classifiers		Precision	Recall	f1-score
DecisionTree	Class 0	0.90	0.89	0.90
Classifier	Class 1	0.60	0.63	0.62
XGBoost	Class 0	0.85	1.00	0.92
Classifier	Class 1	0.96	0.33	0.49
ExtraTrees	Class 0	0.87	0.89	0.88
Classifier	Class 1	0.82	0.72	0.78
RandomForest	Class 0	0.90	0.97	0.94
Classifier	Class 1	0.85	0.73	0.79
Logistic	Class 0	0.81	1.0	0.90
Regressor	Class 1	0.99	0.15	0.25

The best classifier is chosen in terms of its f1-score. It can be observed that some classifiers produce high f1-score for class 0 but poor score for class 1. Thus, it can be inferred that the classifier predicted majority to be no delay even if it was delayed, which defeats the purpose of the project, hence the best classifier is chosen in consideration of the f1-score of the minority class(1). Thus, from the above classifiers **RandomForest Classifier** is seen to have the highest **f1-score** for the minority class and a reasonable score for the majority class, namely 0 and 1, than the other four classifiers used.

3.3 Data Imbalance

It can be observed that the dataset is largely imbalanced(see Fig. 1), Class 0 takes upto 79% of the data . This may result the classifier to get used to majority class more than minority class , which can be seen in table 4 , making the classifier show bias and being not very accurate.This data imbalance might have a bigger impact on the efficiency of the classifier . Thus , methods such as **Undersampling** and **Oversampling** is used.

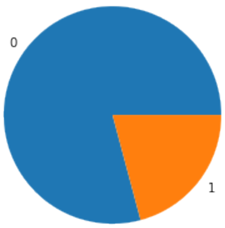


Fig. 1: 1-Delayed flights
0-No Delay

3.3.1 UnderSampling

Undersampling is a technique to balance uneven datasets by keeping all of the data in the minority class and decreasing the size of the majority class. In our dataset, class 1 is the minority class, which is kept as it is, and class 0 is the majority, which is decreased to the size of class 1. As mentioned above, RandomUnderSampler is used for undersampling the dataset. The scores that the classifiers produced are tabulated below:

Table 5: DecisionTree Classifier

		Precision	Recall	f1-score
DecisionTree	Class 0	0.94	0.79	0.86
Classifier	Class 1	0.51	0.80	0.62
XGBoost	Class 0	0.94	0.92	0.93
Classifier	Class 1	0.73	0.79	0.76
ExtraTrees	Class 0	0.95	0.88	0.92
Classifier	Class 1	0.65	0.81	0.72
RandomForest	Class 0	0.88	0.92	0.90
Classifier	Class 1	0.80	0.78	0.79
Logistic	Class 0	0.92	0.93	0.93
Regressor	Class 1	0.74	0.76	0.75

Even after Undersampling the dataset, **RandomForest** Classifier still produced the best results.Overall, the scores haven't improved much to emphasize.

3.3.2 OverSampling

Oversampling involves duplicating examples in the minority class, although these examples don't add any new information to the model. Instead, new examples can be synthesized from the existing examples. This is a type of data augmentation for the minority class and is referred to as the Synthetic Minority Oversampling Technique, or SMOTE for short. In our case, Class 1 is the minority,so it is duplicated until it equalizes the majority in frequency. While, Class 0 is the Majority, so it is left as it is.

The scores that the classifiers produced are tabulated below:

Table 6: DecisionTree Classifier

		Precision	Recall	f1-score
DecisionTree	Class 0	0.92	0.91	0.92
Classifier	Class 1	0.68	0.71	0.69
XGBoost	Class 0	0.93	0.95	0.94
Classifier	Class 1	0.81	0.73	0.77
ExtraTrees	Class 0	0.93	0.92	0.93
Classifier	Class 1	0.76	0.75	0.76
RandomForest	Class 0	0.90	0.93	0.92
Classifier	Class 1	0.81	0.79	0.80
Logistic	Class 0	0.94	0.93	0.94
Regressor	Class 1	0.75	0.77	0.76

From the scores produced, it can be inferred that the scores after oversampling have subsequently gotten better than the scores with No-sampling/Under-sampling. This implies that the data-imbalance did actually have an impact on the classifiers, affecting their performance.

4 Regression

Regression is a method that attempts to determine the strength and character of the relationship between one dependent variable (usually denoted by Y) and a series of other variables (known as independent variables-X).It is a supervised learning technique which helps in finding the correlation between variables and enables us to predict the continuous output variable based on one or more predictor variables. In this project **ArrDelMinutes** is the dependant variable(Y) and all the other columns mentioned in Table 1 and Table 2 are the independant variables(X).

The regressors are trained with two different datasets separately. The first set of data is taken from the core dataset of delayed flights only.While, the second set of data is taken from the core dataset which was predicted to be delayed by the best classifier(RandomForest). The regressors are trained to predict the delay of a flight in **minutes**.

4.1 Evaluation Metrics

To build and deploy a generalized model we require to Evaluate the model on different metrics which helps us to better optimize the performance, fine-tune it, and obtain a better result. The metrics used to evaluate the regressors are **Mean absolute error**, **Mean squared error** and **r2 score**.

Mean absolute error It is the difference between the actual value and predicted value that is an absolute error but we have to find the mean absolute of the complete dataset used.

$$MAE = (1/n)\sum_{i=1}^{n}\left|x_i - x_p\right| \tag{5}$$

Mean squared error It represents the squared distance between actual and predicted values. The squared is done to avoid the cancellation of negative terms and it is the benefit of MSE.

$$MSE = (1/n)\sum\left|x_i - x_p\right|^2 \tag{6}$$

R2 Score R2 score is also known as Coefficient of Determination or sometimes also known as Goodness of fit. A good R2 score is when it s closer to 1.

$$R2Squared = 1 - (SSr/SSm)$$

SSr = Squared sum error of regression line
SSm = Squared sum error of mean line
x_p = Predicted value
x_i = Mean value

4.2 Regression Scores

The scores produced by the regressors are tabulated below:

Table 7: Core Dataset

Regressors	MSE	MAE	R2score
Linear	19.81	14.45	0.9234
ExtraTrees	16.36	10.90	0.9477
XGB	17.02	11.68	0.9435
RandomForest	16.82	11.79	0.9448

The scores of all the regressors , namely Linear regressor, XGB, ExtraTrees, RandomForest , are tabulated above in Table 7. After observing the scores, **Extratrees Regressor** is chosen as the best regressor among the four regressors explored , because it has lower MAE and MSE, implying that the error is the lowest than any other regressors.

5 Pipeline Architecture

The regressors are trained using two different datasets separately to see which produces the best result. The first dataset comprises of data which belongs to flights which were predicted to be delayed by the best classifier(RandomForest). The second dataset comprises of data which consists of only delayed flights from the core dataset. The process is depicted below in fig 2.

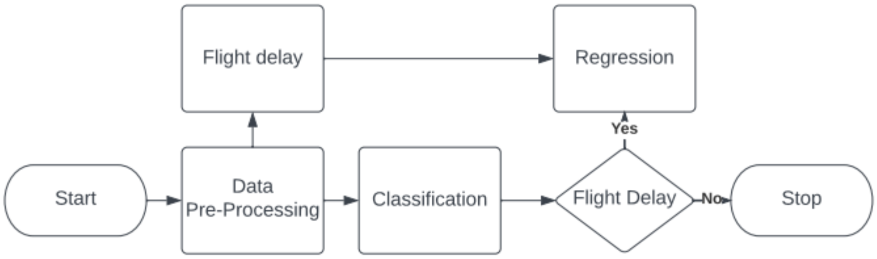


Fig. 2: Pipeline Architecture

It can be observed that both the goodness of fit as well as the error made by the regressors has increased noticeably which were trained using classifier predicted dataset. The scores are tabulated below:

Table 8: Classification Dataset

Regressors	MSE	MAE	R2score
Linear	17.95	12.77	0.9527
ExtraTrees	17.27	11.80	0.9563
XGB	17.10	11.52	0.9571
RandomForest	17.42	11.78	0.9555

6 Regression Analysis

Regression analysis is a statistical method to model the relationship between a dependent (target) and independent (predictor) variables with one or more independent variables. More specifically, Regression analysis helps us to understand how the value of the dependent variable is changing corresponding to an independent variable when other independent variables are held fixed.

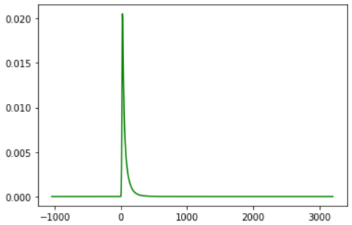
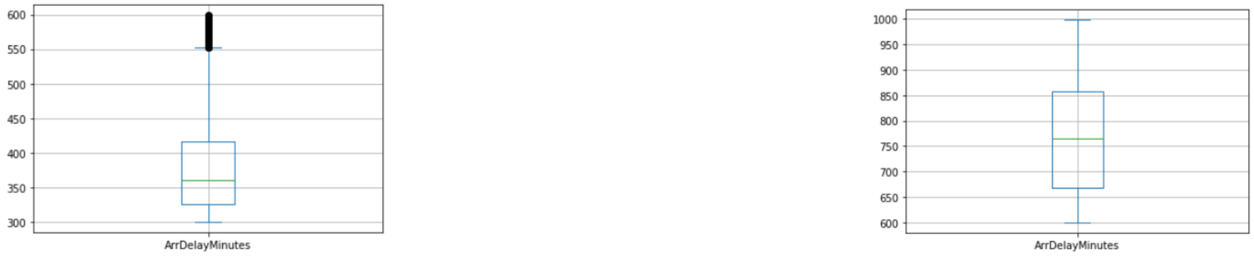


Fig. 3: Delay minutes density

The scores which were produced by the regressors with different intervals of the dataset are tabulated below:

Table 9: Regression Analysis

Intervals	MSE	MAE	R2score	Datapoints
15-300	18.35	13.84	0.8560	374349
300-600	30.03	20.74	0.8289	4144
600-1000	19.93	15.73	0.9685	699
1000-1500	23.92	18.75	0.9614	163
1500+	33.25	26.87	0.3661	11



(a) 600-1000

(b) 1000-1500

Fig. 4: Intervals

It can be inferred from the above table that the regressor performed best in the interval 600-1000 producing low errors and good fit. It can also be seen that the regressor produced poor result in the interval 1500+ due to lack of datapoints to train and learn from. The interval 15-300 had the highest datapoints to train from and predict, thus achieving low errors and a overall good fit. The interval had significantly low datapoints to learn from than its previous interval, but, still, it produced a good fit but a noticeable error increase.The interval 1000-1500 produced a better fit than the intervals having higher datapoints, and has mediocre error margin.

From the above scores it can be concluded that the model made the least error and performed well in the intervals 15-300 and 600-1000. However, it produced a lot of errors in the interval 1500+ due to lack of datapoints. Surprisingly, the model had a high error percentage in the interval 300-600 even though it had the second highest datapoints to train from.

6.1 Regression Analysis

- Regression estimates the relationship between the target and the independent variable.
- It is used to find the trends in data.
- It helps to predict real/continuous values. By performing the regression, we can confidently determine the most important factor, the least important factor, and how each factor is affecting the other factors.
- It shows which part of the dataset does the model make more/less errors.

7 Conclusion

The RandomForest classifier was able to predict whether a flight is delayed or not with satisfactory accuracy and gave a f1-score of 0.94 for the majority class and 0.79 for the minority class. Likewise, the ExtraTrees regressor produced a decent result in terms of prediction and gave an average error of 11 minutes to the actual arrival delay. The MSE on the other hand was close to 16 minutes. Overall, this project was successful in achieving the desired objectives. However, further development of this project can yield better scores and prove to be more accurate .