Flight Prediction Using Respective Flight and Weather Data

Shreyas Srinivasan | Machine Learning | 24/09/2021

ABSTRACT

A flight delay occurs if the flight arrives earlier or later than the scheduled time. This delay will be affected by factors such as weather conditions, air traffic, thunderstorm and other weather glitches. The aim of this project will be to build a two-stage predictive machine learning model to train a dataset and predicting if there will be a delay in arrival time and by how much time will the delay be.

INTRODUCTION

Flight delay affects airlines, airports and passengers. Prediction of flight delay helps passengers and airlines reduce financial losses caused by delays and also be wary of possible accidents that can be caused due to bad weather conditions. Hence it is essential to predict flight delays.

Flight data across 15 airports across the years 2016 and 2017 and weather data across these airports for the same years.

The flight and weather data are first preprocessed and merged appropriately to obtain the required data frame.

The data frame is then classified using appropriate classifier models, to predict if there is a delay in arrival time of the flight, and based on the results of metrics obtained, the best classifier is obtained.

The data frame is then regressed using different regressor models and is then used to check the delay period for delayed flights. Based on the metrics obtained, the best regressor is obtained.

DATASET

Airport codes:

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

Weather Data Columns:

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

Flight Data Columns:

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

DATA PREPROCESSING

The Flight and Weather data with the above columns are preprocessed by first merging the flight and weather data separately. Then the 2 are merged using left join and right join method for departure and arrival data of Flight Dataset respectively. After Flight and Weather data are merged, the rows with NULL and Non-Finite values are dropped.

CLASSIFICATION

It is essential to know if a flight will be delayed or not first before finding out by how much time will there be a delay. Hence the first stage of pipeline is classification. In classification, we classify whether the flight is delayed or not by 15 minutes or more. The ground truth used for classification is ArrDel15 which tells us if Arrival Delay of flight is more than 15 minutes or not. If ArrDel15 = 1.0, the flight is delayed by 15 minutes or more. ArrDel15 = 0.0 for other cases.

METRICS FOR CLASSIFICATION

Following notations for

TN: True Negative (the model has predicted correctly that a flight will not be delayed)

TP: True Positive (the model has predicted correctly that a flight will be delayed)

FN: False Negative (the model has predicted wrongly that a flight will not be delayed)

FP: False Positive (the model has predicted wrongly that a flight will be delayed)

Accuracy: Ratio of true results to the total number of results examined.

$$\text{Accuracy} = \frac{(\mathrm{TN} + \mathrm{TP})}{(\mathrm{TN} + \mathrm{TP} + \mathrm{FN} + \mathrm{FP})}$$

Precision: Ratio of true positives to total positives. Gives a proportion of predictive positives which are truly positive.

$$\text{Precision} = \frac{TP}{(TP+FP)}$$

Recall: Proportion of actual positives which are correctly classified.

$$\text{Recall} = \frac{TP}{(TP+FN)}$$

F1 Score: Harmonic mean of Precision and Recall.

$$\text{F1 Score} = 2 \times \frac{Precision \times Recall}{Precision+Recall}$$

Confusion Matrix: Gives the number of TP, TN, FP and FN in a matrix form.

$$Confusion\ Matrix = \begin{matrix} TP & FN \\ FP & TN \end{matrix}$$

CLASSIFIER MODELS

We use different Classifier models and determine the best classifier for our project based on F1 score and accuracy. The different classifier models we use are Logistic Regressor, Decision Tree Classifier, Extra Tree Classifier, Gradient Boost Classifier and Random Forest Classifier.

Classifier performance

Class 0: non-delayed flights

Class 1: delayed flights

Algorithm	Precision		Recall		F1 Score		Accuracy
	0	1	0	1	0	1	
Logistic Regressor	0.80	0.89	1.00	0.07	0.89	0.12	0.80
Decision Tree Classifier	0.92	0.67	0.91	0.70	0.91	0.68	0.86
Extra Tree Classifier	0.85	0.45	0.85	0.46	0.85	0.45	0.77
Gradient Boost Classifier	0.92	0.90	0.98	0.68	0.95	0.78	0.92
Random Forest Classifier	0.87	0.85	0.98	0.45	0.92	0.58	0.87

Table 1: Performance of different classifiers

DATA IMBALANCE PROBLEM

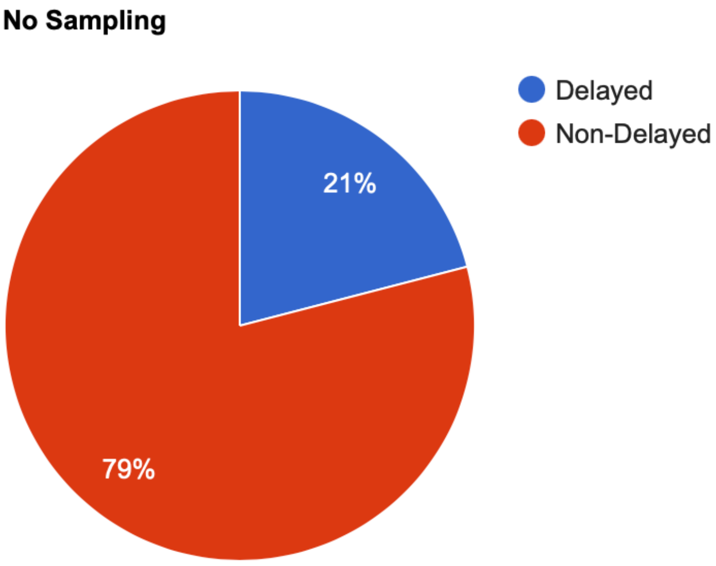


Figure 1: Representation of delayed and non-delayed flights with No Sampling

From table 1, it is shown that performance of Class 1 (non-delayed flights) is weaker than Class 0 (delayed flights). This is caused due to the significantly higher number of non-delayed flights compared to delayed flights.

This bias can be overcome by using Under sampling and Over sampling. Under sampling involves removal of delayed flights data from dataset to make it equal to the number of non-delayed flights. Over sampling involves addition of non-delayed flights data to dataset to make it equal to number of delayed flights.

Under Sampling

Under sampling of data is done using **RandomUnderSampler** function.

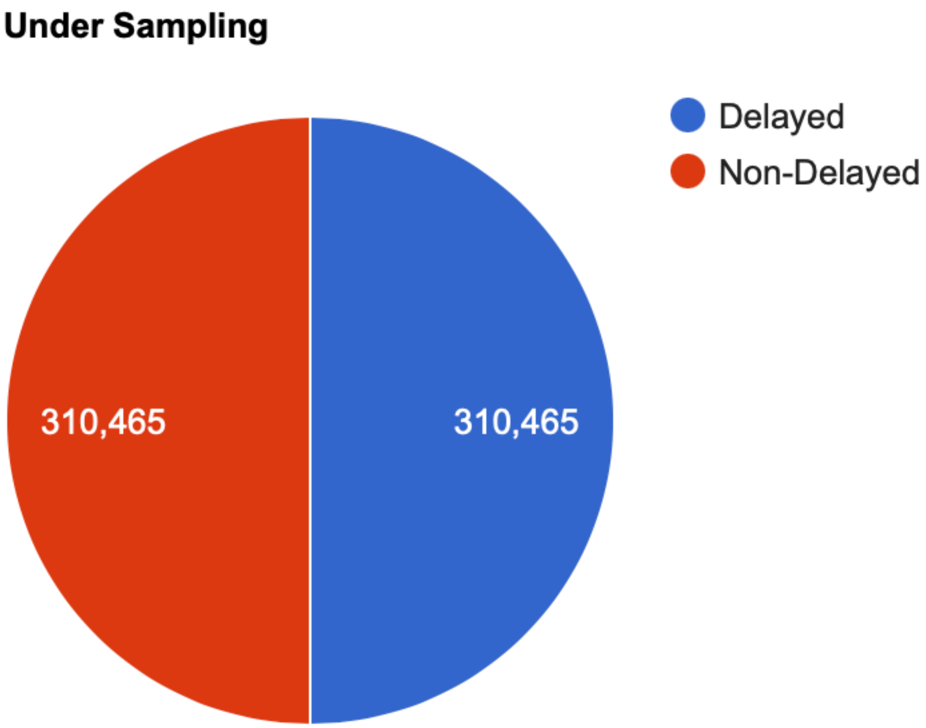


Figure 2: Representation of delayed and non-delayed flights after Under Sampling

Algorithm	Precision		Recall		F1 Score		Accuracy
	0	1	0	1	0	1	
Logistic Regressor	0.86	0.29	0.61	0.61	0.71	0.40	0.61
Decision Tree Classifier	0.93	0.49	0.78	0.79	0.85	0.61	0.78

Extra Tree Classifier	0.86	0.30	0.62	0.62	0.72	0.41	0.62
Gradient Boost Classifier	0.94	0.74	0.93	0.79	0.94	0.77	0.90
Random Forest Classifier	0.93	0.53	0.83	0.75	0.87	0.62	0.81

Table 2: Classifier Performance after Under Sampling using RandomUnderSampler

Over Sampling

Oversampling is done using **RandomOverSampler** function.

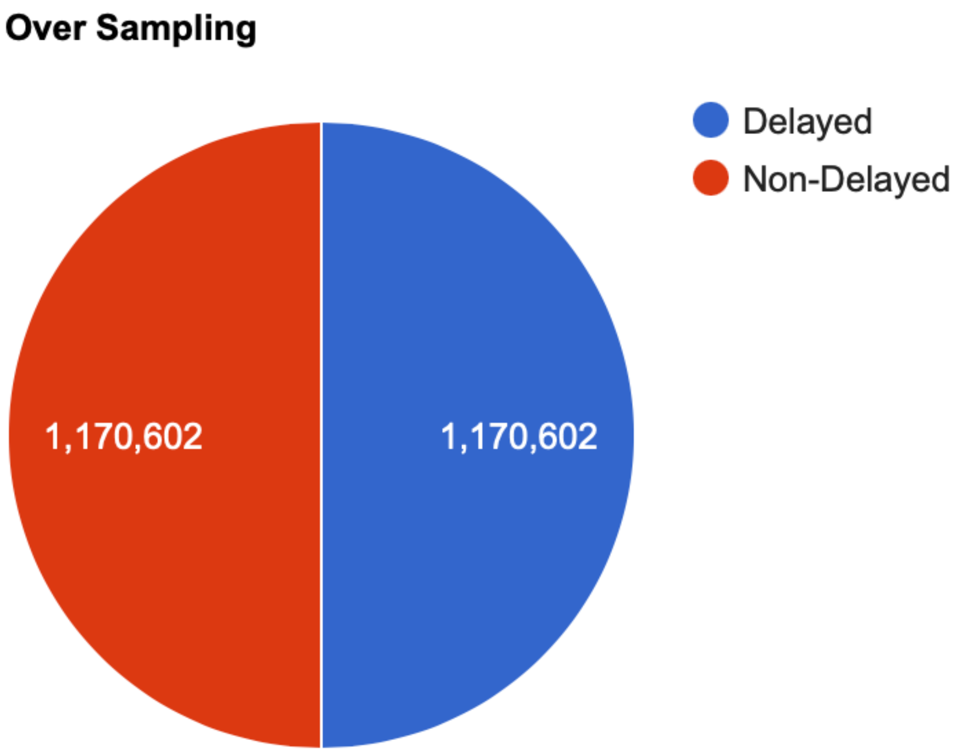


Figure 3: Representation of delayed and non-delayed flights after Over Sampling

Algorithm	Precision		Recall		F1 Score		Accuracy
	0	1	0	1	0	1	
Logistic Regressor	0.85	0.29	0.61	0.61	0.71	0.40	0.61
Decision Tree Classifier	0.92	0.68	0.91	0.69	0.91	0.68	0.87

Extra Tree Classifier	0.84	0.42	0.85	0.41	0.85	0.41	0.76
Gradient Boost Classifier	0.94	0.74	0.93	0.79	0.94	0.77	0.92
Random Forest Classifier	0.88	0.75	0.95	0.52	0.92	0.62	0.86

Table 3: Classifier Performance after Over Sampling using RandomOverSampler

Choosing Best Classifier

After analyzing the results in tables 1,2 and 3, we can infer that the best classifier is the **Gradient Boost Classifier after Over sampling of data** as it has highest F1 Score.

REGRESSION

After finding out if a flight is delayed or not, it is important to know the delay time. In regression, delay period (in minutes) of delayed flights is obtained using different regression models. Delayed flights are those with ArrDel15 = 1.0. Our ground truth for regression will be ArrDelayMinutes, which is the time difference (in minutes) between the CRSArrTime (scheduled Arrival Time) and ArrTime (actual Arrival Time).

METRICS FOR REGRESSION

Following notations for

$\overline{Y}$: Mean Value of Y

$\hat{Y}$: Predicted Value of Y

N: Number of Data Points

R2 Score: Statistical measure that measures the proportion of variance for a dependent variable that's explained by an independent variable.

$$R2\ Score = 1 - \frac{\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{N}(Y_i - \overline{Y}_i)^2}$$

Mean Square Error (MSE): Statistical measure that gives us the square of average error in predicted value from the original value.

$$MSE = \frac{1}{N} \times \sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2$$

Root Mean Square Error (RMSE): Square root of MSE which gives average error in predicted value from original value.

$$RMSE = \sqrt{MSE}$$

Mean Absolute Error (MAE): Statistical measure that gives us the average of absolute value of the error between predicted and original value.

$$MAE = \frac{1}{N} \times \sum_{i=1}^{N} |Y_i - \hat{Y}_i|$$

REGRESSOR MODELS

Different Regressor models are used and the best regressor is chosen based on R2 score, RMSE and MAE. The different regressor models used are Linear Regressor, Extra Tree Regressor, Gradient Boost Regressor, Random Forest Regressor and Ridge CV Regressor.

REGRESSOR PERFOMANCE

Algorithm	R2 Score	RMSE	MAE
Linear Regression	**0.03**	**68.93**	**41.99**
Extra Tree Regressor	**0.93**	**18.77**	**12.16**
Gradient Boost Regressor	**0.82**	**29.56**	**19.37**
Random Forest Regressor	**0.93**	**19.12**	**12.21**
Ridge CV Regressor	**0.03**	**68.93**	**41.99**

Table 4: Regression metrics for different models

Choosing Best Regressor

Based on the results in Table 4, the **best regressor is the Extra Trees Regressor** as it has highest R2 Score and least Root Mean Square Error and almost the lowest Mean Absolute Error.

REGRESSION ANALYSIS

The arrival delay time of flights ranges from 0.0 to 2142.0 minutes. Regression analysis is done to check performance of the best regressor, obtained from Table 4, on smaller distributions of time periods and form an analysis based on the same.

Time Period	Number of Flights	R2 Score	RMSE	MAE
0-100	323550	0.64	13.27	10.11
100-200	48954	0.64	16.15	12.50
200-500	14232	0.88	22.73	16.72
500-1000	1128	0.94	33.07	25.19
1000-2142	175	-0.06	156.69	75.73

Table 5: Frequency Table with Regression Metrics

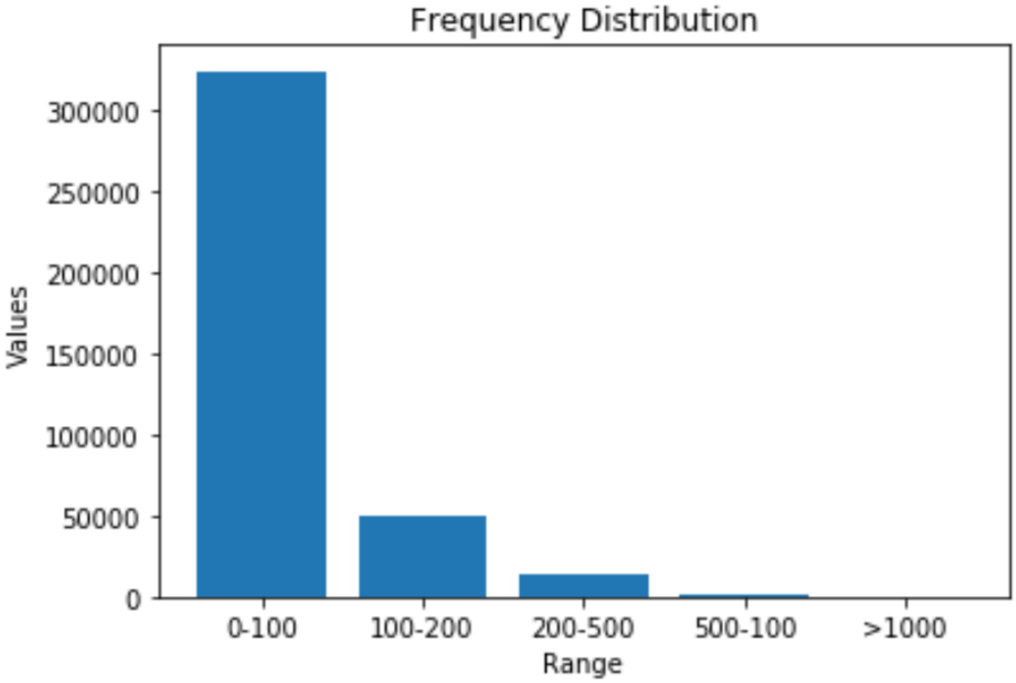
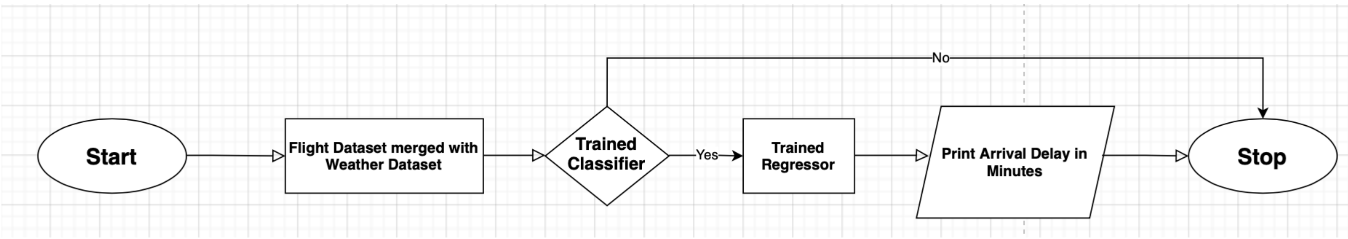


Figure 4: Frequency Distribution of Flights Delayed in Minutes

Low RMSE value indicates that predicted data is close to original data and low MAE value indicates the performance of the regressor. High R2 Score indicates how well the predicted data has been fit among original data.

Based on the same, it can be inferred from Table 5 that best performance has been shown in the 500-1000 range with a good R2 Score and an average value for RMSE and MAE.

PIPELINE


Start
Flight Dataset merged with Weather Dataset
Trained Classifier
Yes
No
Trained Regressor
Print Arrival Delay in Minutes
Stop

In pipelining we use our best classifier and regressor to classify and regress our data respectively. We first classify our data using our best classifier which is Gradient Boost Classifier using Oversampling to classify flights based on whether they're delayed or not by 15 minutes. We then check our accuracy, precision, recall and F1 Score.

For regression we use our best regressor, Extra Trees Regressor, for the delayed flights. After regression, we check our R2 Score, Root Mean Square Error and Mean Absolute Error.

Metrics	**Value**	
	0 (non-delayed)	1 (delayed)
Precision	0.94	0.74
Recall	0.93	0.79
F1 Score	0.94	0.77
Accuracy	0.90	
R2 Score	0.92	
RMSE	20.92	
MAE	12.91	

Table 6: Pipelining Metrics

CONCLUSION

A two-stage predictive machine learning model has been successfully built which predicted if a flight will be delayed or not and by how much time will there be a delay if it exists. With this information, passengers and airlines can significantly reduce losses caused by delays and also help in prevention of accidents caused due to bad weather conditions.