Flight Delay Prediction

Shanthakumar 3rd year IT

# Sri Venkateswara College of Engineering

Chennai



**Introduction**

The most important criteria that a flight should follow is the arriving and departing in time. If the flight does not follow the specified schedule then it may lead to may problem and chaos for both the passenger and the airline company. There will be days where the delay is unavoidable as it is not in our control , but the one that is in our control is that we can predict the delay using a machine learning model. Hence we can make plans accordingly.

The aim of this project is to predict the delay in arrival

of a particular flight. This project aims on training a machine learning module with weather and flight data of the departing and arriving airports.

**Specification**

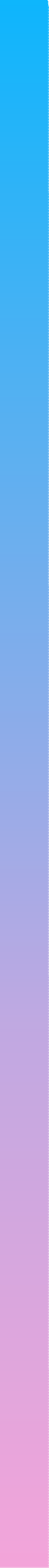
**The datasets used for this problem are:**

Weather data of 15 airports from 2013 to 2017 ( TABLE: 1) Flight data of all flights travelling in USA during 2016 and 2017 ( TABLE: 3)

DATA PREPROCESSING

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

**TABLE: 1**



The first step of the project is to segregate the given data and take only the required data of the given weather data. We will be taking the above mentioned fields alone .The data is in Json format , we will merge the data and store it in the form of csv.

The secod step is to merge all the flight data. As we will be dealing only with 15 airports, let us only consider the flights that are travelling to or from the below mentioned airports.

**TABLE: 2**

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

The features that will be considered for the flight are:

**TABLE: 3**

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

The next step is to merge the weather data and the flight data.

Each flight data is merged with its corresponding departure time weather and arrival time weather data. The features considered are:

* Date
* Time
* Airport





**Classification**

We need to classify whether the flight will be delayed or not. The features used to fit the classifier are ORIGIN AIRPORT, DESTINATION AIRPORT, DEPARTURE DELAY, and then weather features of both destination and origin airports. Since there is a huge imbalance between the delayed and not delayed flight. We should manipulate the training data. We will be using oversampling and resampling to manipulate the data. The Classifiers considered are: Extra Trees Classifier, XGBoost, Decision Tree, Logistic Regression.

**CLASS 0 represents the Flights which are not delayed**

**CLASS 1 represents the Flights which are delayed**

# 

# **True vs. False and Positive vs. Negative**

|  |  |
| --- | --- |
| True Positive (TP):   * Reality: Flight Dealyed. * Predection: “Delayed." * Outcome: Correct Prediction. | False Positive (FP):   * Reality: Flight NOT- Dealyed. * Predection: “Delayed." * Outcome: In-Correct Prediction. |
| False Negative (FN):   * Reality: Flight Dealyed. * Predection: “Delayed." * Outcome: In-Correct Prediction. | True Negative (TN):   * Reality: Flight NOT- Dealyed. * Predection: “NOT-Delayed." * Outcome: Correct Prediction. |

A **true positive** is an outcome where the model correctlypredicts the positive class. Similarly, a **true negative** is an outcome where the model correctly predicts the negativeclass.

A **false positive** is an outcome where the model incorrectlypredicts the positiveclass. And a **false negative** is an outcome where the model incorrectly predicts the negativeclass.

NOSAMPLING:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | PRECISION | RECALL | F1 SCORE |
| **XGBOOST** | Class 0 | 0.92 | 0.98 | 0.95 |
|  | Class 1 | 0.90 | 0.68 | 0.77 |
|  |  |  |  |  |
| **EXTRA TREES CLASSIFIER** | Class 0 | 0.93 | 0.95 | 0.94 |
|  | Class 1 | 0.80 | 0.74 | 0.77 |
|  |  |  |  |  |
| **DECISION TREE** | Class 0 | 0.93 | 0.95 | 0.94 |
|  | Class 1 | 0.80 | 0.74 | 0.77 |
|  |  |  |  |  |
| **LOGISTIC REGRESSION** | Class 0 | 0.92 | 0.98 | 0.95 |
|  | Class 1 | 0.88 | 0.68 | 0.77 |

OVERSAMPLING:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | PRECISION | RECALL | F1 SCORE |
| **GNB** | Class 0 | 0.94 | 0.93 | 0.93 |
|  | Class 1 | 0.74 | 0.76 | 0.75 |
|  |  |  |  |  |
| **EXTRA TREES CLASSIFIER** | Class 0 | 0.94 | 0.94 | 0.94 |
|  | Class 1 | 0.77 | 0.76 | 0.77 |
|  |  |  |  |  |
| **DECISION TREE** | Class 0 | 0.92 | 0.91 | 0.92 |
|  | Class 1 | 0.68 | 0.70 | 0.69 |
|  |  |  |  |  |
| **LOGISTIC REGRESSION** | Class 0 | 0.94 | 0.93 | 0.93 |
|  | Class 1 | 0.74 | 0.78 | 0.76 |

RESAMPLING:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | PRECISION | RECALL | F1 SCORE |
| **EXTRA TREES CLASSIFIER** | Class 0 | 0.93 | 0.95 | 0.94 |
|  | Class 1 | 0.80 | 0.74 | 0.77 |
|  |  |  |  |  |
| **DECISION TREE** | Class 0 | 0.92 | 0.92 | 0.92 |
|  | Class 1 | 0.70 | 0.70 | 0.70 |

**Precision** :

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations .

**Precision = TP/TP+FP**

**Recall**:

Recall is the ratio of correctly predicted positive observations to the all observations in actual class

**Recall = TP/TP+FN**

**F1 score** :

F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it’s better to look at both Precision and Recall

**F1 Score = 2\*(Recall \* Precision) / (Recall + Precision)**

Extra trees classifier after resampling was the best performing classifier because it had a good F1 score and recall for both class 0 as well as class 1, unlike other classifiers which was biased towards the class 0.

**Regression**

The number of delayed flights must be predicted . To do that, We will be taking all the flight records. The same values will be passed to regressor .

The following regressors are used:

* Extra Trees Regressor,
* Linear Regressor
* XGBoost Regressor

The RMSE and MAE are taken into account for finding the right regressor. MSE is MEAN SQUARE ERROR while MAE is MEAN ABSOLUTE ERROR. Both MSE and MAE gives the error in the predicted value when compared to the actual value. So lesser the error, the better is the model built.

R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression.

**R-squared = Explained variation / Total variation**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **RMSE** | **MAE** | **R-SQUARED** |
| **LINEAR REGRESSOR** | 17.6887 | 12.2594 | 0.9398 |
| **EXTRA TREE REGRESSOR** | 16.9530 | 11.8720 | 0.9436 |
| **XGB REGRESSOR** | 17.0413 | 11.7324 | 0.9441 |

From the following table, we see that Extratree Regressor has given a MAE of **11.87** and RMSE of **16.95**, which is considered to be less because the actual arrival delay is pretty high when compared to these values, and it is better than the other regressors. The R-square(**0.9436**) value is also high when compared to the other regressors.

**Pipeline**

We will now predict the number of flights delayed. We will be taking only the flight records which was predicted to arrive late by the extratrees classifier. The same values will be taken to regressor aswell.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **RMSE** | **MAE** | **R-SQUARED** |
| **LINEAR REGRESSOR** | 18.8519 | 14.0365 | 0.9397 |
| **EXTRA TREE REGRESSOR** | 19.3984 | 14.6161 | 0.9350 |
| **XGB REGRESSOR** | 19.2290 | 14.4818 | 0.9373 |

From the following table, we see that Linear Regressor has given a MAE of **14.03** and RMSE of **18.85**, which is considered to be less because the actual arrival delay is pretty high when compared to these values, and it is better than the other regressors. The R-square(**0.9397**) value is also high when compared to the other regressors.

**Conclusion**

The model was able to predict the flight delay with a good accuracy. Extra tree classifier was able to classify the data into delayed and not delayed. The extra tree classifier was able to give a f1 score of 0.94 for class0 and 0.77 for class1 after resampling. The predicted arrival delay by the Linear regressor had a Mean absolute error of 14 minutes and Mean square error of 18 minutes.

