**ML Project**

**Flight Delay Prediction**

**1. Problem statement:**

Finding the delay in the arrival of the Flight based on the given datasets of flight and weather details. The prediction of the arrival delay is predicted once the flight has taken off.

**2. Introduction:**

The departure and the arrival timings of a scheduled flight always get varied depending on the weather at the time of flight departure or arrival.

Such a delay of flight always affects the economy of the airport as well as the passengers travel plans. The prediction of such a delay in the flight timings is achieved through this ML project which could get better customer satisfaction. In this project the data is pre-processed followed by which a two-stage model is used to find the flight delay.

**3. Data Pre-processing:**

the datasets consists of :

**weather data-**

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

**Flight data -**

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

The datasets of weather and flight is taken only for the 15 airports ATL, CLT, DEN, DFW, EWR , IAH, JFK, LAS, LAX, MCO, MIA, ORD, PHX, SEA and SFO.

An additional feature of arrival and departure absolute hour is added which consists of the hour closests to which the flight departs and arrives.

The flight data and the weather data is merged based on the features of place, date and the hour of arrival and departure.

The Flight data and the weather data is merged such that each record of the flight has the corresponding weather data at that hour.

The combined dataset consists of 33 features.

This combined dataset is used for the next two modules of classification and regression.

**4. Classification:**

The main purpose of this module is to predict whether a flight arrival will be delayed or not. ArrDel15 stores the values of 1 or 0 emphasizing whether the flight is delayed or not. Class 0 represents flight is not delayed whereas class 1 represents flight is delayed.

Totally 33 features are taken for training the classifier. The flight delay is predicted using the classifiers.

The records are split into train and test data. 80% of data is used for training and 20% for testing. The train data is used to train the model. And the accuracy of the model is tested using the test data.

**Results with no sampling:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classifier** |  | **Precision** | **Recall** | **F1 score** |
| **Extra trees**  **Classifier** | **Class 0**  Class 1  Weighted average | **0.83**  0.48  0.76 | **0.92**  0.30  0.79 | **0.87**  0.37  0.77 |
| **Decision tree**  **classifier** | **Class 0**  Class 1  Weighted average | **0.83**  0.36  0.74 | **0.82**  0.39  0.73 | **0.83**  0.37  0.73 |
| **XGB**  **Classifier** | **Class 0**  Class 1  Weighted average | **0.94**  0.73  0.90 | **0.92**  0.78  0.89 | **0.93**  0.75  0.89 |

Only 1/4th of the train records are class 1 whereas the rest are class 0. Because of such a difference the values of classifier are this poor.

Hence to make the two classes have the same number of records to get a better result the classes are sampled. There are two types of sampling which are oversampling and undersampling. Undersampling reduces the greater class to match with the lower class and in oversampling the smaller class is increased to meet the number of greater class.

These sampled values are added to the classifier models and the value is results are compared to find out which type of sampling should be followed and which is the best classifier model.

**Oversampling Results:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classifier** |  | **Precision** | **Recall** | **F1 score** |
| **Extra trees**  **Classifier** | **Class 0**  Class 1  Weighted average | **0.90**  0.83  0.88 | **0.97**  0.58  0.89 | **0.93**  0.68  0.88 |
| **Decision tree**  **classifier** | **Class 0**  Class 1  Weighted average | **0.92**  0.68  0.87 | **0.91**  0.71  0.87 | **0.92**  0.70  0.87 |
| **XGB**  **Classifier** | **Class 0**  Class 1  Weighted average | **0.94**  0.73  0.90 | **0.92**  0.78  0.89 | **0.93**  0.76  0.89 |

**Undersampling Results:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classifier** |  | **Precision** | **Recall** | **F1 score** |
| **Extra trees**  **Classifier** | **Class 0**  Class 1  Weighted average | **0.93**  0.59  0.86 | **0.86**  0.76  0.84 | **0.89**  0.66  0.85 |
| **Decision tree**  **classifier** | **Class 0**  Class 1  Weighted average | **0.94**  0.51  0.85 | **0.80**  0.80  0.80 | **0.86**  0.63  0.81 |
| **XGB**  **Classifier** | **Class 0**  Class 1  Weighted average | **0.80**  0.68  0.77 | **0.88**  0.05  0.70 | **0.88**  0.09  0.72 |

F1-score is calculated for different classifiers and the one with the best score is considered as the best model.

The delay predicted by the classifier is added as a new column to the existing merged csv file and this record is passed on to the next regression module.

**5. Regression:**

The records obtained from the module 2 prediction is used here to find the minutes of delay of a flight.

Only those records in which flight is delayed is used to train the regressor model. Similar to module 2 the data is split into train and test data. 80% train data and 20% test data. Additional feature of flight delay is added in this module so hence totally 34 features are used here.

The efficiency of different regressor models is compared using the r2, mean absolute error and the root mean square error scores.

**Regressor results:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Regressor** | **R2**  **score** | **Mean Absolute Error** | **Root Mean Square Error** |
| **Linear Regressor** | 0.945 | 11.80 | 16.83 |
| **Extra Trees Regressor** | 0.944 | 12.42 | 17.33 |
| **Random Forest Regressor** | 0.945 | 12.39 | 17.28 |
| **XGB Regressor** | 0.949 | 11.68 | 16.56 |

**Results and Conclusion:**

The two stage model for flight delay prediction was built. In the first stage the classifier is used to predict whether the flight is delayed or not. The dataset initially had a class imbalance giving poor results and after trying out the sampling methods, the oversampled data gave the best result of 0.93(class 0) and 0.76(class 1) with **XG boost classifier**. Thus this result was used for the second stage. In the regressor model only those data were taken which consisted of flights which were delayed. The best model obtained here was from **XGB regressor** which gave the scores of 0.95 (r2), 11.68 (mean absolute error) and 16.56(root mean square error). The mean errors where less compared to actual delay minutes and hence this is the best model for our prediction.

Thus the flight prediction model was successfully built and good results were obtained.