FLIGHT DELAY PREDICTION

A Project report on modelling a two stage predictive machine learning engine that forecasts the on-time performance of flights

Sri Gayathri Devi I

Solarillion Foundation

ABSTRACT:

In this fast paced world, time prevail as the most powerful and invaluable thing. To our concern here, Flight delays are considered to affect passengers, airlines, goods' transport resulting in increased block times on routes and higher carrier costs and indirect circular impact on the rest of the economy. Furthermore, in the domain of sustainability, it can even cause environmental harm by the rise in fuel consumption and gas emissions. Hence, these factors indicate how necessary it has become to predict the delays no matter the wide-range of airline meshes. As a result, air passenger travel decisions can be influenced by delay information. The insight of this project is to predict the delay and forecasts the on-time performance of flights in the USA airports.

Introduction:

Machine learning is a method of data analysis that automates analytical model building that learns from experience.The process involved are explained in this project with detailed analysis. The U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics tracks the status of the flights. They provide a summary of the arrival delays, departure delays, on-time arrival etc in their monthly report. This study aims at analyzing flight information of US domestic flights covering top 15 busiest airports of US and predicting possible arrival delay using Data Mining and Machine Learning Approaches.

Flight Delay and reasons:

- A flight delay is said to occur when an airline lands or takes off 15 minutes later than its scheduled arrival or departure time respectively.
- Notable reasons are adverse weather conditions, air traffic congestion, maintenance and security issues.

Supervised Learning:

In Supervised machine learning, we train the machine using data which is well "labeled." The techniques are,

- Regression: Regression technique predicts a single continuous output value using training data typically used in predicting, forecasting, and finding relationships between quantitative data. In our case,by how many minutes a particular flight is delayed is predicted.

- Classification: Classification means to group the output inside a class.Modelling a classifier that predicts whether a flight is delayed or not is projected here.

Workflow of the project:

- Gathering data
- Data pre-processing
- Researching the model that will be best for the type of data
- Training and testing the model
- Evaluation

The Dataset:

The data was contained in two folders: Weather and Flight. It comprises of weather information at 15 different airports in the USA between 2016 and 2017 and the flight details of all flights that flew inside the USA. Here we are interested in 15 airports in the years of 2016 and 2017.

- Weather data: It includes weather features such as WindSpeedKmph, Pressure, Cloudcover, tempF, etc for each hour in a day provided for each month.There were 25 hourly features and 6 daily features.Some of the significant features were,
- Flight data: Each file holds the On-Time Performance of flights.There are around 110 features. Some of the prominent features were,

Data Preprocessing:

The goal is to merge the flight details with its corresponding weather information. As a result, a flight flying at a particular time starting from a particular airport will have the corresponding hour's weather report.

- Weather- The .json files are in the form of nested dictionaries. Each file includes the weather report for a month containing hourly information for each day.All these files are processed into a single .csv file.

- Flight - They are found in csv format and combined accordingly for 2 years concerning only 15 significant airports and framed into a single file.

- Merge - The primary features for merge would be Date, Time, Airport.

FEATURE	DESCRIPTION
WindSpeedKmph	The rate at which air moves in a particular area
Visibilty	measure of horizontal opacity of the atmosphere at the point of observation and is
WindGustKmphs	a sudden, brief increase in speed of the wind
WindDirDegree	the degree of direction from which the wind originates
WeatherCode	it provide a precise location to the weather widget
Pressure	atmospheric or air pressure is the force per unit of area exerted on the Earth
Cloudcover	refers to the fraction of the sky obscured by clouds
tempF	refers to how warm or cold air is in Fahrenheit
WindChillF	is the lowering of temperature due to the passing-flow of lower-temperature air.
PrecipMM	any product of the condensation of atmospheric water vapour
DewPointF	the temperature the air needs to be cooled to achieve a relative humidity (RH) of 100%
Humidity	amount of water vapour in the air.
airport	The location of origin airport where above features are observed.
Time	an hour of above observation
Date	date at which above features are observed

Table 1: Weather Features

Features	Description
OriginAirportID	Origin Airport, Airport ID. An identification number assigned by US DOT
Origin	Origin Airport
DestAirportID	Destination Airport, Airport ID. An identification number assigned by US DOT
Dest	Destination Airport
CRSDepTime	CRS Departure Time (local time: hhmm)
DepTime	Actual Departure Time (local time: hhmm)
DepDelay	Difference in minutes between scheduled and actual departure time. Early departures show negative numbers.
DepDelayMinutes	Difference in minutes between scheduled and actual departure time. Early departures set to 0.
DepDel15	Departure Delay Indicator, 15 Minutes or More (1=Yes)
CRSArrTime	CRS Arrival Time (local time: hhmm)
ArrTime	Actual Arrival Time (local time: hhmm)
ArrDelay	Difference in minutes between scheduled and actual arrival time. Early arrivals show negative numbers.
ArrDelayMinutes	Difference in minutes between scheduled and actual arrival time. Early arrivals set to 0.
ArrDel15	Arrival Delay Indicator, 15 Minutes or More (1=Yes)

Table 2: Flight Features

- Handling Missing Values: Instances containing missing labels are removed from the dataset.

A well-defined dataset is procured from the raw data.

Feature Selection:

Here we understand which columns/features are not helping ML pipeline and are removed. Techniques used to select the best set of features are,

1.Principal Component Analysis
2.Univariate Selection(SelectKBest with ANOVA)
3.Feature Importance
4.Correlation Matrix with Heatmap

Among the above methods, best features from Univariate and Feature Importance were filtered out.

DepDelayMinutes	tempF'
DepTime	pressure
CRSDepTime	winddirDegree
DestAirportID	WindGustKmph
OriginAirportID	windspeedKmph
time	WindChillF
'DayofMonth	DewPointF
'Quarter'	humidity
Month	weatherCode

Table 3: Filtered Features

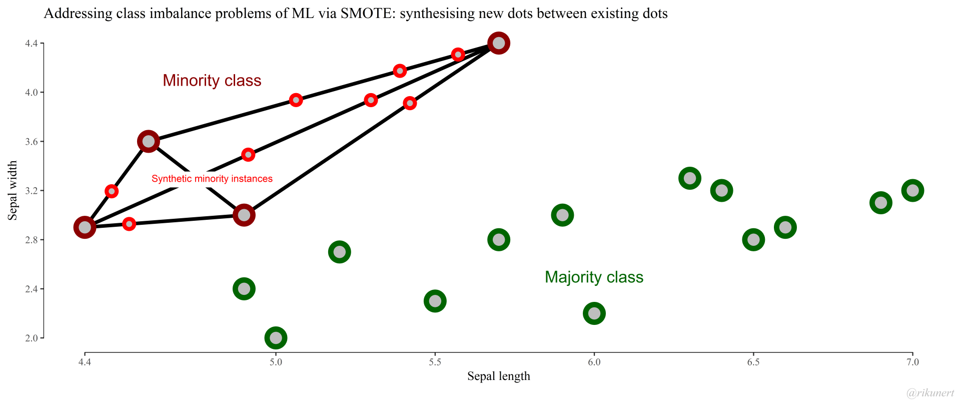
Class Imbalance and Sampling:

So far, in the implementation perspective, it is understandable that a binary classification needs to be performed on the label, Arr Del 15 which assumes binary values 0 and 1. But, the ratio of number of instances with label 0 to label 1 is found to be 4:1. Hence, the dataset is highly imbalanced. This might lead to incorrect accuracy.Techniques used to overcome class imbalance are,

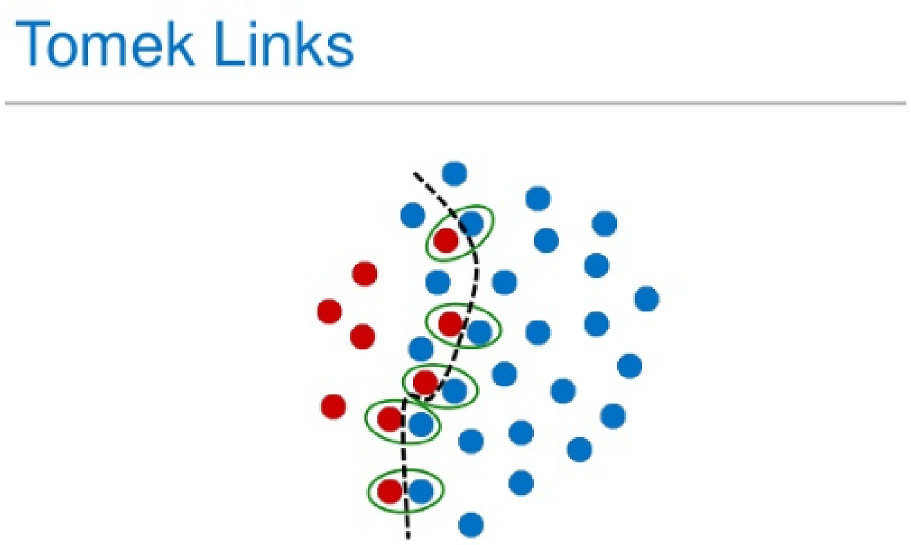
1.Under-Sampling(Random UnderSampling)
2.Over-Sampling(SMOTE, ADASYN)
3.Combined-Sampling(SMOTETomek)

- Random Under-Sampling: Balances the class distribution by randomly eliminating majority class.

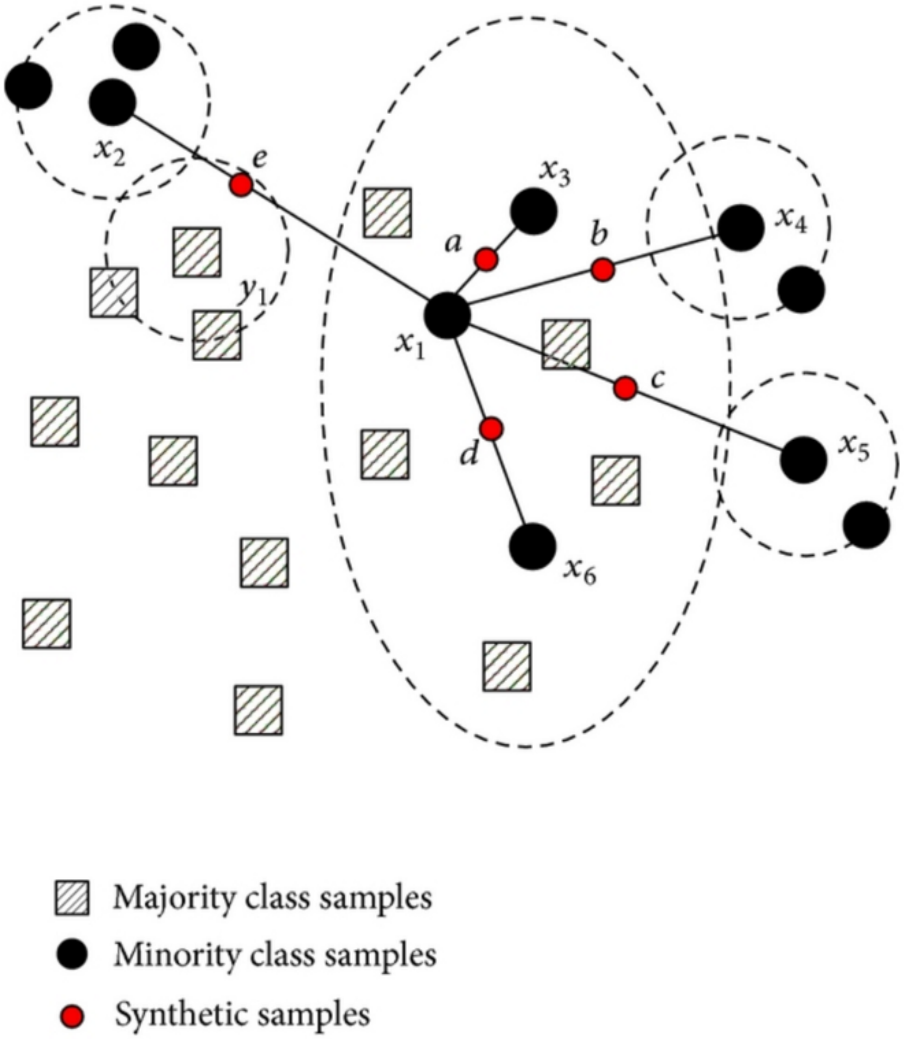
- SMOTE(Synthetic Minority Oversampling Technique): K-nearest neighbours are identified in feature space and drawing a line between the examples in the feature space where new samples are generated at a point along that line.



- SMOTETomek: Performs over-sampling using SMOTE and cleaning using Tomek links.


Tomek Links

- ADASYN (ADAptive SYNthetic):The algorithm uses Euclidean distance for KNN Algorithm. The key difference between ADASYN and SMOTE is that the former uses a density distribution, as a criterion to automatically decide the number of synthetic samples for each minority sample by adaptively changing the weights of the different minority samples to compensate for the skewed distributions.



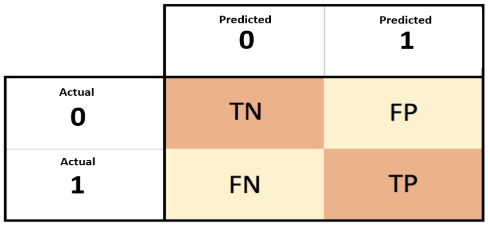
SMOTETomek is found to produce better sampling compared to other techniques.

Performance metrics:

Before modelling a Classifier or Regressor, we ought to know how to evaluate these models. To find how effective a model is, various performance metrics are used.

- **Classification metrics:**

  1.Confusion Matrix: It is a table with two dimensions (“Actual” and “Predicted”), and sets of “classes” in both dimensions.It contains the number of predicted samples that belong to TN, FP, FN and TP. Ideal case is when False Negatives and False Positives are zeros. Minimising FN and FP gives an effective model.



	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

  2.Accuracy score: number of correct prediction over all predictions

$$accuracy = \frac{TruePositives + TrueNegatives}{TotalSamples} \tag{1}$$

3.Precision: the probability that the decision is correct

$$precision = \frac{TruePositives}{TruePositives + FalsePositives} \tag{2}$$

4. Recall: the proportion of actual positives that was identified correctly

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives} \tag{3}$$

5.AUC curve(Area Under Curve): is equal to the probability that the classifier will rank a randomly chosen positive example higher than a randomly chosen negative example.

6. Classification report: This report consists of the scores of Precisions, Recall, F1 and Support.

- **Regression metrics:**

  1.Mean Absolute Error (MAE): It is the average of the difference between the Original Values and the Predicted Values

$$MAE = \frac{1}{n} \sum \mid Y - \hat{Y} \mid \tag{4}$$

2.Root Mean Squared Error(RMSE):

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\Big(\frac{d_i - f_i}{\sigma_i}\Big)^2} \tag{5}$$

3.R2 score:provides an indication of the goodness or fit of a set of predicted output values to the actual output values.

$$R^2 \ = \ 1 - \frac{\frac{1}{n}\Sigma_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\frac{1}{n}\Sigma_{i=1}^{n}(Y_i - \hat{Y}_i)^2} \tag{6}$$

4.Cross Validation Technique and K-Fold Technique: It shows how a model would generalize to an independent data set. The model dataset is divided into three sets: Training, test, and validation. The entire set is divided into K-folds. Then, the K-1 folds are sent for training and the learning is done on it, then the model's generalization is checked on the test set, which contains just the remaining one fold; and this process goes on till the last fold.

Training and testing sets:

- Training Dataset: The sample of data that we use to train the model. The model learns from this data.

- Test Dataset: The sample of data used to provide an unbiased evaluation of a final model fit on the training dataset.

The dataset is split into training and testing set.

Classification:

It is the process of assigning a 'class label' to a particular item. Our objective is to predict where a fight would be delayed or not on its Arrival. Thus, the target variable is "ArrDel15" denoting 1,0 for delayed and on time flights respectively.

The dataset is first split into training and testing data and training data is sampled and then fit into classifier models.
XGBoost, DecisionTree, GradientBoost, ExtraTrees, and Stochastic Gradient Descent were uesd.

Below is the performance of each classifier.

	Unsampled	**RandomUnderSampling**	**ADASYN**	**SMOTETomek**
Accuracy	0.868	0.791	0.866	0.916
Precision	0.678	0.501	0.672	0.913
Recall	0.705	0.802	0.704	0.919

Table 4: DecisionTreeClassifier

	Unsampled	**RandomUnderSampling**	**ADASYN**	**SMOTETomek**
Accuracy	0.919	0.881	0.918	0.918
Precision	0.899	0.922	0.891	0.893
Recall	0.692	0.834	0.693	0.693

Table 5: XGBoostClassifier

	Unsampled	**RandomUnderSampling**	**ADASYN**	**SMOTETomek**
Accuracy	0.868	0.791	0.866	0.916
Precision	0.678	0.501	0.672	0.913
Recall	0.705	0.802	0.704	0.919

Table 6: ExtraTreesClassifier

	Unsampled	**RandomUnderSampling**	**ADASYN**	**SMOTETomek**
Accuracy	0.917	0.895	0.900	0.910
Precision	0.895	0.735	0.760	0.818
Recall	0.684	0.786	0.767	0.737

Table 7: GradientBoostingClassifier

	Unsampled	**RandomUnderSampling)**	**ADASYN**	**SMOTETomek**
Accuracy	0.876	0.884	0.851	0.905
Precision	0.669	0.696	0.671	0.968
Recall	0.808	0.797	0.698	0.565

Table 8: Stochastic Gradient Descent

Regression:

Regression model predicts values of a desired target quantity. Of those flights that are classified as delayed, the arrival delay minutes are predicted with various regressors.
"ArrDelayMinutes" being the target variable.

Regressors such as Linear, ExtraTrees, RandomForest, GradientBoosting were used. Among which GradientBoosting regressor performs well. Results are shown below.

	MAE	RMSE	R2 Score
LinearRegression	14.64	20.01	0.9213
ExtraTreesRegressor	12.03	17.17	0.9422
GradientBoostingRegressor	11.74	17.09	0.9426
RandomForestRegressor	11.93	16.97	0.9433

Further, GradientBoosting regressor was used to predict the delay minutes at different intervals of time in order to assess the performance of regressor on different intervals of time.

Conclusion:

The dataset was explored and intended features were extracted according to our objective and Machine Learning models were implemented appropriately. Prediction of arrival delay and their range is successfully established with good amount of accuracy.The accuracy can still be improved with hyper parameter tuning using GridSearchCV or RandomSearchCV and further feature engineering.