Machine Learning Nanodegree Engineer

Capstone Project

Bharat Singh
Jan-19

Flight Delay Prediction

Table Contents

Definition

Project Overview

This project has been initiated from airline domain. The main objective of Flight Delay Prediction Machine learning project is to predict aircraft delay. This will help in resource management. It will give prior information of flight journey, that it will be delay or reach on time on destination airport. So by using prior knowledge, the can manage resource. Like ground staff, taxi and baggage allocation etc.
As I am going to classify that the given flight will be delay or will reach at scheduled time. It problem comes under the supervised Classification Problem.

Dataset:

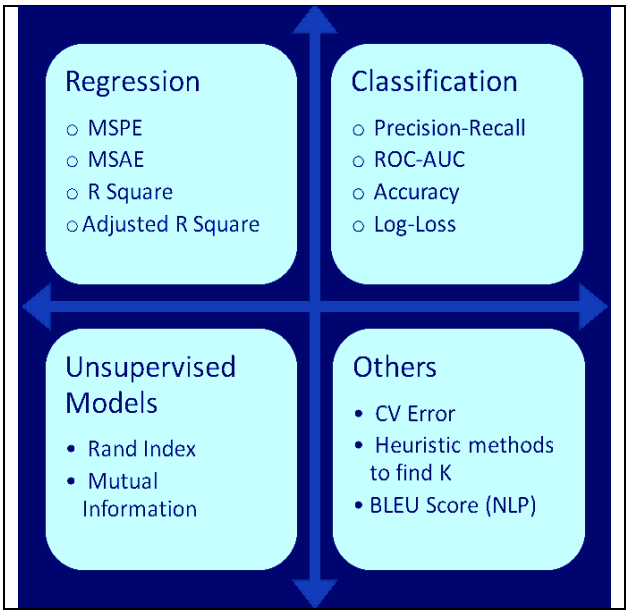
https://www.kaggle.com/fabiendaniel/predicting-flight-delays-tutorial/data

Problem Statement

This is the supervised learning problem, so we will use classification algorithm. It will classify, that given flight will be delay or not.

Matrices

For validation of machine learning model preformation on unseen data or for verifying, system have generalized well for unseen data or not. We need some evaluation matrices. Machine Learning have different matrices for different type of problem. Like Classification, regression or clustering etc..

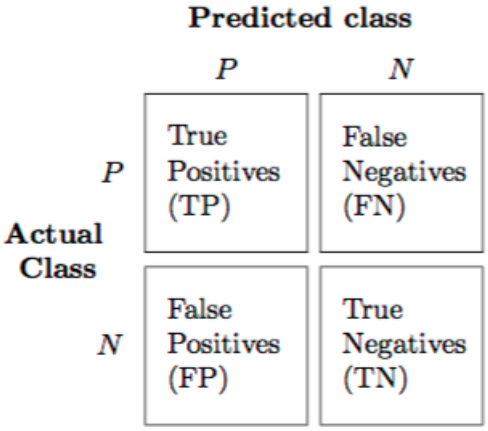


As our problem is related to Supervised Classification Machine learning, we will use all the classification related to matrices.

Confusion Matrix

Confusion matrix is a table representation of model output, which is used to validate the classification model preformation on set of testing data for

which resultant values are known. It also required to calculate Precision, Recall, Accuracy and AUC-ROC Curve.



https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

Accuracy

Accuracy is a metric, which is required to check the model accuracy on the unseen data.
https://developers.google.com/machine-learning/crash-course/classification/accuracy

Recall

Recall is an evaluation matric for Machine Learning Classification Model. It will show, what's the probability of correctly classifying for the given positive sample?
"Out of all the positive classes, how much model predicating correctly. It should be high as possible. It called as Sensitivity or Recall."

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

Precision

ROC-AUC

ROC-AUC (Receiver Operating Characteristic – Area under the Curve). It's a

Analysis

Data Exploration

In this project we are going to use Kaggle competition dataset. Which is publically available on Kaggle. Data for Flight Delay Prediction has been taken from DOT's Bureau of Transpiration statistics. It's related to flight journey. It has data in CSV format. For data exploration, I will use

- airline.csv
- airports.csv
- flights.csv

For prediction of flight arrival delay, we will use mainly flights.csv data. It has 31 features including target feature (Arrival delay). We will drive another feature FLIGHT_DELAY from ARRIVAL_DELAY. I will contain 'YES' OR 'NO'.

Feature	Description
YEAR	Year of flight departure date
MONTH	Month of flight departure date
DAY	Day of flight departure date
DAY_OF_WEEK	Day of week of flight departure date
AIRLINE	Airline Name (Like Virgin, emirates etc)
FLIGHT_NUMBER	Flight unique identifier
TAIL_NUMBER	Flight Registration number
ORIGIN_AIRPORT	Flight Departure airport
DESTINATION_AIRPORT	Flight Arrival airport
SCHEDULED_DEPARTURE	Planned flight departure time.
DEPARTURE_TIME	Actual Departure time
DEPARTURE_DELAY	Actual departure delay in flight time.
TAXI_OUT	Flight left the gate
WHEELS_OFF	Flight wheels take-off from runway.
SCHEDULED_TIME	Flight planned time for journey.
ELAPSED_TIME	
AIR_TIME	Total time of traveling
DISTANCE	Distance from origin Airport to destination airport.
WHEELS_ON	Wheels touch the runway on arrival airport.
TAXI_IN	Arrival time at gate.
SCHEDULED_ARRIVAL	Planned arrival time on arrival airport
ARRIVAL_TIME	Actual arrival time on airport
ARRIVAL_DELAY	Arrival delay in journey. We will drive **Flight_DELAY** feature from it. Like Yes Or NO
DIVERTED	Flight diverted to another airport in between journey or not
CANCELLED	Flight got cancelled or not
CANCELLATION_REASON	Reason for flight cancellation
AIR_SYSTEM_DELAY	Flight delay because of air system. Air traffic or air control system.
SECURITY_DELAY	Is flight got delay, because of security checks?
AIRLINE_DELAY	Is flight got delay, because of airline?
LATE_AIRCRAFT_DELAY	Is flight got delay, because of aircraft was late?

Instead of directly using these feature, will try to drive some feature.
Total samples: 5819079
Total Features: 31
Drive Features: 9
Target Feature: FLIGHT_DELAY (YES, NO)

Exploratory Visualization

In this section we will try to explore the dataset by visualization and statistical. In our problem, we have counted sample per class. This will help in understanding in data is balanced or imbalanced and distribution

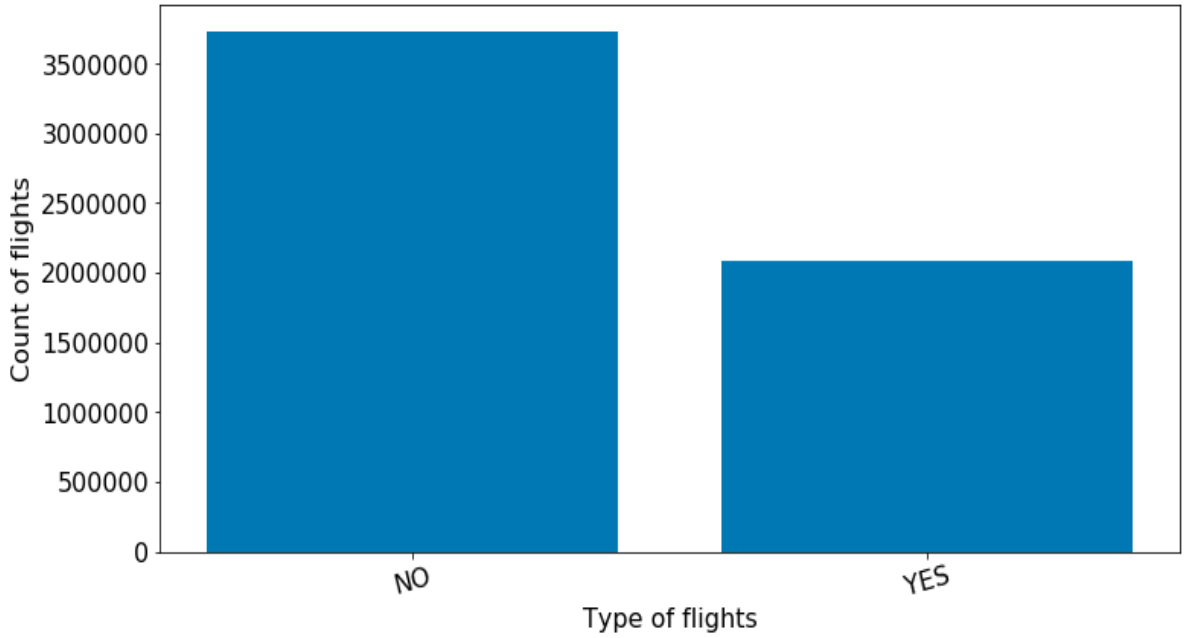


Figure-1

As per above figure 1, our data distribution is highly imbalance. But data has significate sample for each class.

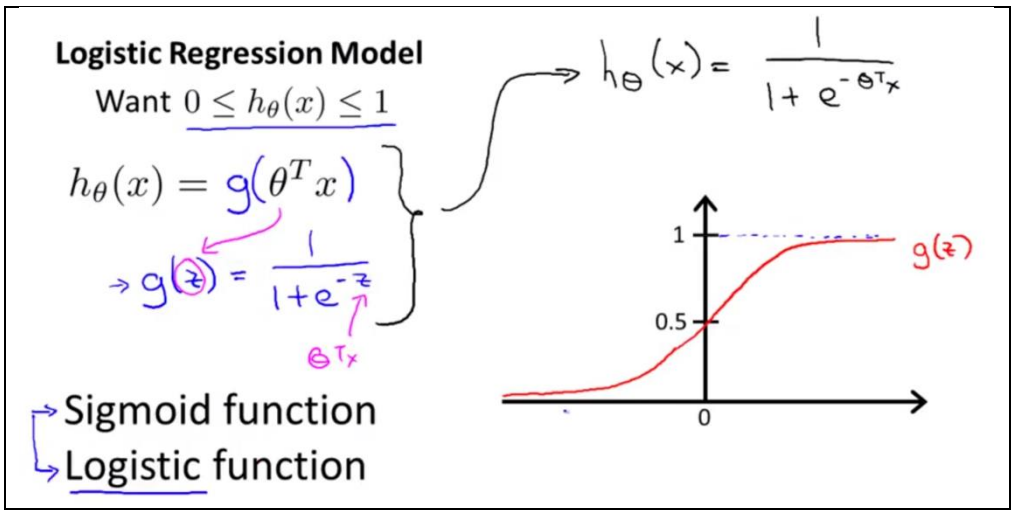
Observation:
Data is imbalance

Algorithms and Techniques.

This problem is related to Supervised Machine Learning. Because it has target feature ('FLIGHT_DELAY'). So for handing it, we need Supervised Machine Learning Algorithm. But the target feature is categorical even its binary category. So finally we need Supervised Classification algorithm.

For some instance dataset is imbalance for flight delay. In respect to imbalance dataset better to use tree base algorithm. Like Decision tree, Random Forest. But at this stage we are not much sure, which application will perform better in compression to other. So, we will train our model on multiple classification algorithms.

Logistic Regression: As per the name it looks like regression algorithm, but it belongs to classification family. And it binary classifier like 0 or 1. So it has the hypothesis

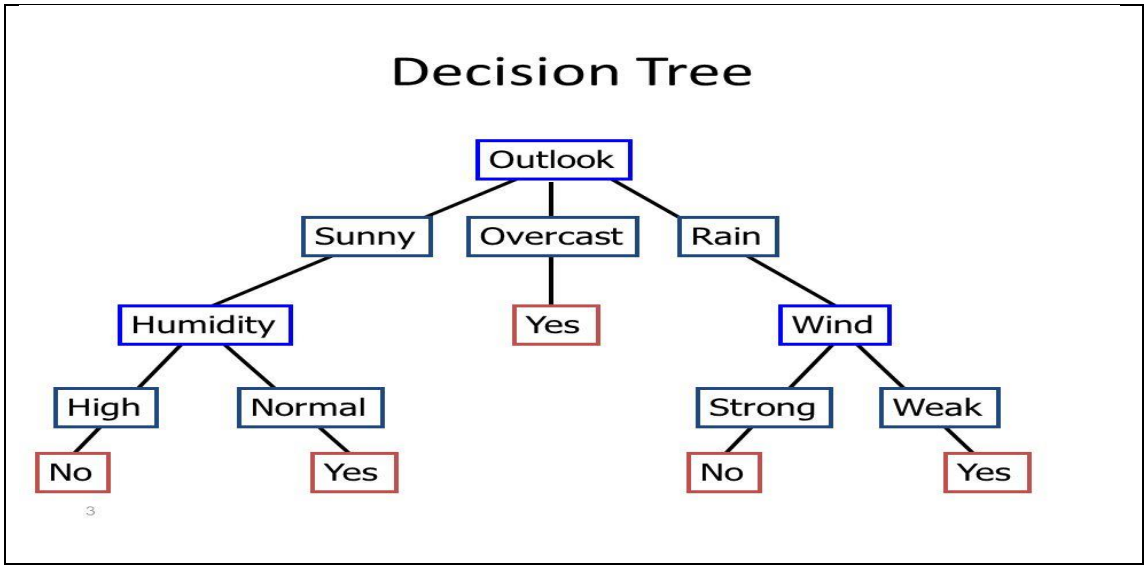


Here $g(z)$ is a sigmoid activation function.

Using the logit function, we will get the simplicity of the methodology of linear without disadvantage. Which means independent variables don't have to be normally distributed or have equal variance in each group.

Decision Tree:

This is tree base algorithm. It works on tree theory. DT can be used for both regression and classification problem. It has one root node, intermediate nodes and leaf node. Leaf nodes are our model outcome. It can handle both numerical and categorical features. Non-linear relationship in between variable don't affect the tree performance.

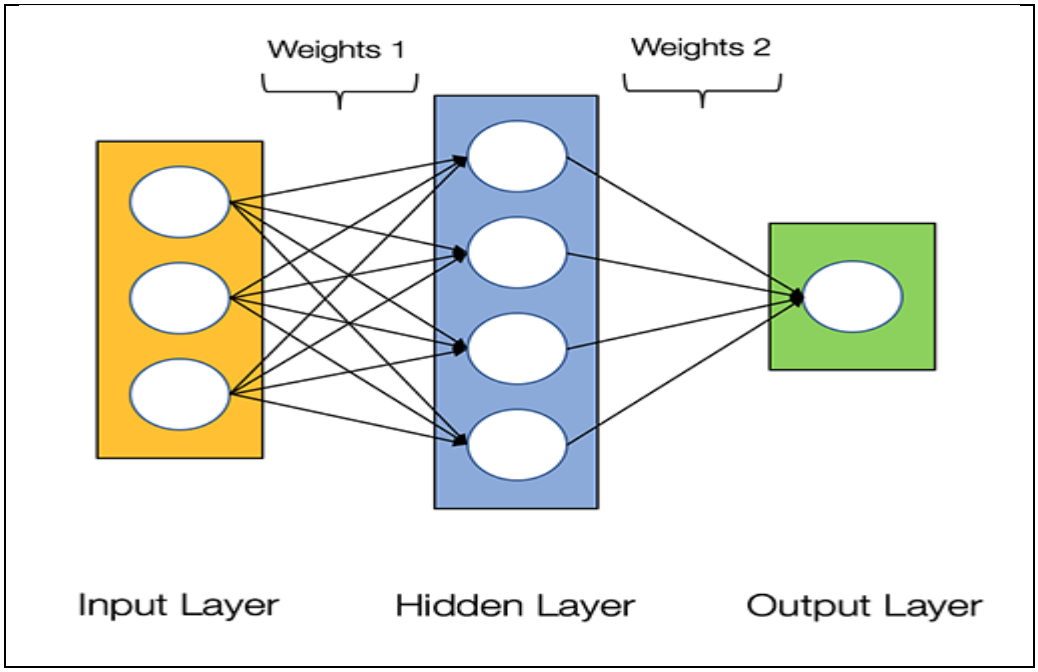


Neural Network:

Neural network refers to interconnected populations of neurons or neuron simulations that form the structure and architecture of nervous system. The theory behind the machine learning neural network, has been taken from brain system.

Neural Networks consist of the following components: It's taking input as sample data, performing some calculation and returning output.

1- An **Input Layer** X
2- An arbitrary amount of **hidden layers**
3- An output function **y_hat**
4- A set of weights and biases between each layers. **W and B**
5- A **Activation function** for each layer



Benchmark

I will consider logistic regression model accuracy as Benchmark of our model and we will try to beat it with other models.

Matrices	Value
Precision	
Recall	
F_1 Score	

Methodology

Data Pre-processing

In this Data Pre-processing, I will performed several step for get clean and good data

- Filling data missing values
- Driving feature from existing features
- Removing less and high correlated features.
- Convert categorical feature into dummy feature
- Normalization of features
- Splitting data in feature and target variable.
- At the end, I will split data in 2 part train and test. Its' required for model training and validation.

Implementation

Process flow of machine learning model shown below figure.

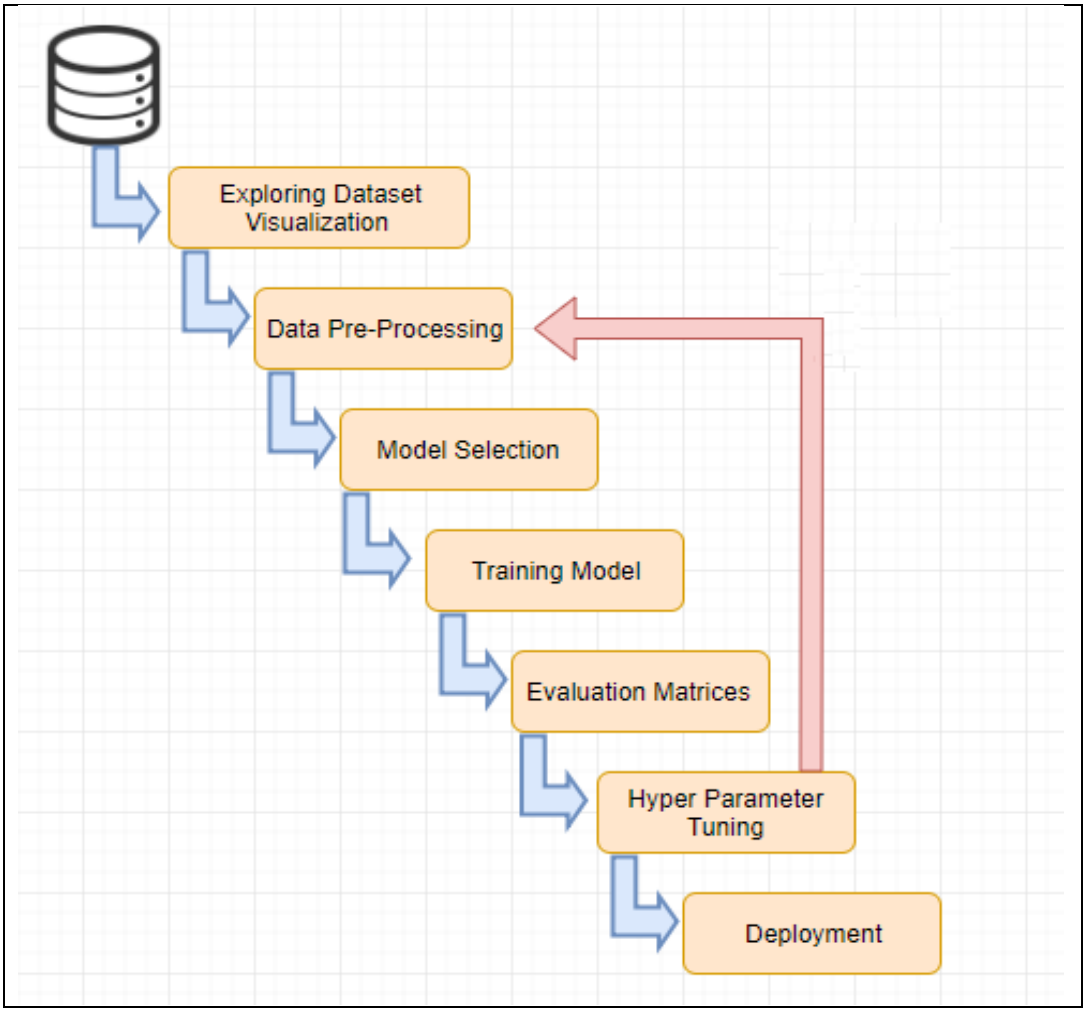


Figure_2

Most of the machine learning model follow the same workflow in implementation of machine learning model.

1- We will load data form CSV files.
2- After loading data in data frame, we will do data visualization and expletory analysis on dataset. With the help of visualization, I will try to understating distribution of data. We will check the correlation in between features, try to remove or drive another feature form highly correlated features. We will drive scatterplot to understand the correlation.
3- After visualization and exploratory analysis of dataset. We will data pre-processing.
    a. Split data in feature and target.
    b. We will check the missing values of features. And based on understanding we will fill them or remove them fully form dataset.
    c. Convert categorical future into dummy features.
    d. And normalization of numeric features.
    e. Splitting data into train and test set.

4- Chose model for training our data
5- Training model on train set
6- Testing the model accuracy by several evaluation matrices
7- Tuning hyper parameter for improving model preformation.
8- Deployment of model as service

Refinement

In this section we will try to improve the performance of our model.

- Reinvestigation of features of model.
- Change in train test split.
- Hyper parameter tuning, in this section, I will try to tune Decision Tree Classifier hyper parameter.

Hyper Parameter	Infor	Options	Best Values

Results

Model Evaluation and Validation

Initially, I have splatted our dataset into train and test. And we have evaluated our model with the test set. In the final model training we have use tuned hyper parameter. For beating the untoned model benchmark model.

As our dataset is imbalance, we can't relay only accuracy metrics for evaluation. So we will use Recall, Precision and F1_Score metrics.
As this problem is binary classification, we will ROC-AUC to test the final model performance.

Matrices	Infor	Options	Best Values
Recall			
Precision			
F1_Score			

Code spnit

Roc-AUC

The code used to evaluate this model is pushed on GitHub repository.

Justification

Although the final model result are good in respect to performance. But still there is room for improvement in the model. Someone can improve the model accuracy, by changing the feature selection, drive new feature etc.

Conclusion

Free-Form Visualization

Reflection

Improvement

Quality

Presentation

Functionality

References

- https://developers.google.com/machine-learning/crash-course/classification/accuracyhttps://www.kaggle.com/c/flight-delays-spring-2018/data
- https://medium.com/greyatom/performance-metrics-for-classification-problems-in-machine-learning-part-i-b085d432082b
- https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/
- https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html
- https://www.quora.com/What-is-the-difference-between-the-project-background-problem-definition-aims-of-project-project-justification-and-scope-of-project
- https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62
- https://en.wikipedia.org/wiki/Precision_and_recall
- https://www.kaggle.com/c/flight-delays-fall-2018/kernels
- https://www.kaggle.com/c/flight-delays-prediction
- https://machinelearningmastery.com/data-leakage-machine-learning/
- https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html
- https://medium.com/greyatom/lets-learn-about-auc-roc-curve-4a94b4d88152
- https://towardsdatascience.com/how-to-build-your-own-neural-network-from-scratch-in-python-68998a08e4f6