Machine Learning Engineer Nanodegree 

Capstone Project

Bharat Singh

Dec, 2018

Flight Delay Prediction

Abstract:

The project proposal is designed for creating a machine learning model for predicting flight delay. So that ground staff and network team can plan for aircraft scheduling and ground handling staff and even passenger also can plan their journey accordingly. It will help flight operation and ground staff for ground handling and network operation. For achieving this goal, we are going to use Supervised Machine Learning.

Data for the flight delay and cancellation problem was collected and published by the DOT’s Bureau of Transpiration Statistics. This project will be implemented with the help of Scikit-learn, Tensorflow and Python.

Table Contents

[Domain Background: 2](#_Toc534642703)

[Problem Statement: 3](#_Toc534642704)

[Dataset and Inputs: 3](#_Toc534642705)

[Flights.csv 3](#_Toc534642706)

[Airline.csv 4](#_Toc534642707)

[Airport.csv 5](#_Toc534642708)

[Solution Statement: 5](#_Toc534642709)

[Benchmark Model: 6](#_Toc534642710)

[Evaluation Metrics: 6](#_Toc534642711)

[Confusion Matrix 6](#_Toc534642712)

[Accuracy 6](#_Toc534642713)

[Precision 7](#_Toc534642714)

[Recall or Sensitivity 7](#_Toc534642715)

[ROC-AUC 7](#_Toc534642716)

[Project Design: 8](#_Toc534642717)

[Language and Libraries 8](#_Toc534642718)

[Data Collection 8](#_Toc534642719)

[Data Visualization 8](#_Toc534642720)

[Feature Engineering 8](#_Toc534642721)

[Train & Test Dataset 8](#_Toc534642722)

[Model Training 8](#_Toc534642723)

[Model Testing 8](#_Toc534642724)

[Model Tuning 8](#_Toc534642725)

[Finalizing Model 9](#_Toc534642726)

[Production Deployment 9](#_Toc534642727)

[References: 9](#_Toc534642728)

Domain Background:

This project has been inherited from The Airline Domain. In Airline, if you want to travel or anyone who wants to travel, he has to book the flight from one place to another.

There are number of factors, which can impact the flight journey like Weather, flight departure time, boarding gate time and actually departure time etc. Keeping these factor in mind, we can decide that particular aircraft can be landed or arrive on time or not or how much it will be delay.

Every airline has their flight history past journey, which can help them in predicting future flight delay. We can implement a machine learning model, which will help us in the prediction of a flight delay.

The Motivation behind this project is to optimization of network operation, ground staff management and passenger.

Problem Statement:

This problem is related to the flight operation department of Airline Industry. Flight is flying from source to destination, sometimes it’s reaching on time, sometimes it reaching with some delay. This Delay in the journey can impact many things like it can block airline’s resource and in Respect to the passenger, they can miss their meetings etc.

Flight delay prediction means, how much time was estimated for journey and how much time actually aircraft took to reach from Origin to Destination.

It’s a binary classification problem. We will classify that the upcoming flight will reach on time or will it be the delay to reach destination airport. For overcoming this problem, we will drive a Supervised Classification Machine Learning.

Dataset and Inputs:

Data for this ML problem provided by DOT’s Bureau of Transpiration Statistics. It’s in CSV format. These CSV contains all the information related to the airport, flight and airline, which is necessary to for it.

This corpus has 3 csv files.

* flights.csv
* airports.csv
* airline.csv

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

For model prediction, we will use flights.csv. But for data insight, we will use other two also.

## Flights.csv

It contains all the past data related to flight schedule from source to destination. Which we will use in our model to test and train. It has 5819079 samples with 31 feature. As we are going to implement it as classification problem. We need to consider a sample per class.

In case data is imbalanced, Accuracy matrices will not be enough to validate model performance.

**Delayed Flight:** 2086896

**Arrival Onetime Flight:** 3732183

We will use **arrival\_delay** to drive final target feature like **flight\_delay**.

|  |  |  |
| --- | --- | --- |
| Feature Name | Data Type | Description |
| YEAR | Integer | Year of travel |
| MONTH | Integer | Month of journey |
| DAY | Integer | Day of Journey |
| DAY\_OF\_WEEK | Integer | Day of week for the given journey |
| AIRLINE | String | Name of Airline |
| FLIGHT\_NUMBER | String | Unique identifier of Flight |
| TAIL\_NUMBER | String | It Aircraft registration number. |
| ORIGIN\_AIRPORT | String | Source airport of journey |
| DESTINATION\_AIRPORT | String | Destination airport of journey |
| SCHEDULED\_DEPARTURE | Float | Schedule departure time of aircraft |
| DEPARTURE\_TIME | Float | Actual departure time of aircraft |
| DEPARTURE\_DELAY | Float | Actual Delay in departure of aircraft. |
| TAXI\_OUT | Float | Time to leave the gate |
| WHEELS\_OFF | Float | Wheels take off from runway. |
| SCHEDULED\_TIME | Float | Schedule take of time of wheels |
| ELAPSED\_TIME | Float |  |
| AIR\_TIME | Float | Arrival time at airport |
| DISTANCE | Integer | The distance between ORIGIN\_AIRPORT and DESTINATION\_AIRPORT |
| WHEELS\_ON | Float | Landing time on the runway |
| TAXI\_IN | Float | Reached on the gate |
| SCHEDULED\_ARRIVAL | Float | Schedule time to reach on gate |
| ARRIVAL\_TIME | Float | Actual arrival time |
| ARRIVAL\_DELAY | **Float** | **Arrival Delay of flight to reach the destination.** |
| DIVERTED | Boolean | Flight diverted in between journey to any other airport. |
| CANCELLED | Boolean | Particular flight got cancelled or not. |
| CANCELLATION\_REASON | String | What was the reason of cancellation of flight? |
| AIR\_SYSTEM\_DELAY | Boolean | Air system issue. |
| SECURITY\_DELAY | Boolean | Security issue |
| AIRLINE\_DELAY | Boolean | Airline started delay |
| LATE\_AIRCRAFT\_DELAY | Boolean | Connecting flight delay |
| WEATHER\_DELAY | Boolean | Weather issue |

## Airline.csv

This csv contains information related to airline. It has 2 feature with 15 sample.

|  |  |  |
| --- | --- | --- |
| Feature Name | Data Type | Description |
| IATA\_CODE | string | IATA Code for airline. It’s unique identifier of airline |
| Airline | string | Airline Name |

## Airport.csv

This csv contains information related to airport. It has 7 feature with 322 sample.

|  |  |  |
| --- | --- | --- |
| Feature Name | Data Type | Description |
| IATA\_CODE | String | IATA code for airport |
| AIRPORT | String | Airport name |
| CITY | String | Airport city |
| STATE | String | Airport belongs to which state. |
| COUNTRY | Sting | Country of airport |
| LATITUDE | Float | Geographical location of airport |
| LONGITUDE | Float | Geographical location of airport |

Solution Statement:

It’s a binary classification problem. So for solving it, we will use Supervised Classification algorithms. But before applying algorithms we have to do the data pre-Processing.

1. Visualization (for getting data insight )
2. Missing value handling
3. Feature selection
4. Categorical and continues feature processing
5. Driving new features
6. Normalization

After Data Pre-Processing, as per thumb rule, we need to divide data into train and test for ML Model training and validation. There are several algorithms for classification, we will apply some of them for predicting aircraft delay.

1. Linear Classifiers: Logistic Regression, Naive Bayes Classifier
2. Support Vector Machines
3. Decision Trees
4. Boosted Trees
5. Random Forest
6. Neural Networks
7. Nearest Neighbour

I am planning to train my model on Logistic Regression, Decision Tree and Neural Networks. At the end on the bases of the evaluation matrix, I will select one model which will perform best for our problem.

Benchmark Model:

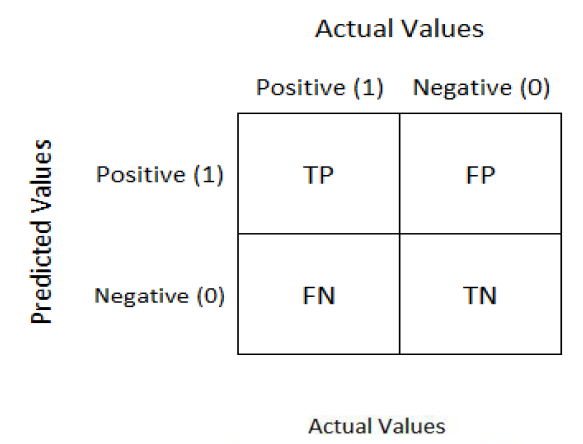
Planning to use Logistic Regression to train and test flight delay prediction. And the output of Logistic Regression, I will use as my benchmark model.

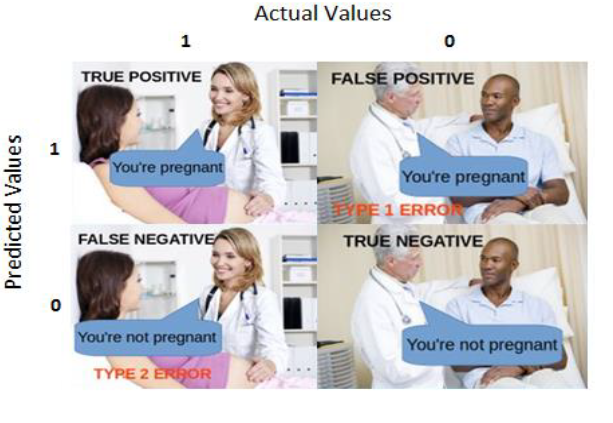
Evaluation Metrics:

Evaluation or performance matrices, after features selection, features engineer, and model training. We need to test the performance of our model, there are couple of matrices to test the model performance or evaluation.

## Confusion Matrix

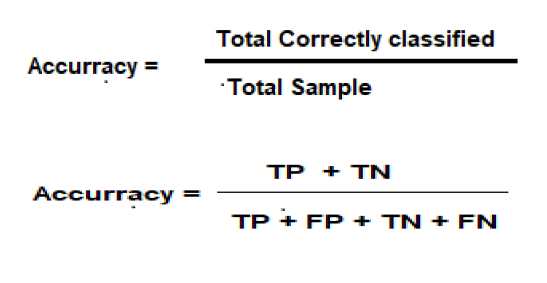
The Confusion matrix is used for validating classification machine learning model. Its table representation of outcome.





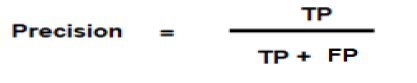
## Accuracy

Accuracy is the measure of calculating that how often Machine learning Model is predicting correctly.



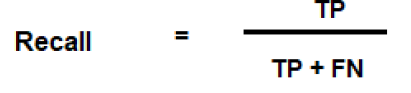
## Precision

Precision is a measure that tells us what proportion of patients that we diagnosed as having cancer, actually had cancer.



## Recall or Sensitivity

The Proportion of correctly classified form the given positive sample.



## ROC-AUC

ROC (Receiver Operating Characteristic Curves) it is used for checking the performance of binary classification model. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The true-positive rate is also known as sensitivity, recall or probability of detection in machine learning.

**“It tells how much model is capable of distinguishing between classes.”**



**TPR: TRUE Positive Rate**

**FPR: False Positive Rate**

Project Design:

Every Machine Learning Project have some steps to achieve the goal. Below the steps or action, we need to perform for any ML project. I will follow the same for this project

## Language and Libraries

* Python 3.X
* Tensorflow
* Scikit-learn

## Data Collection

For implementing a machine learning model, we need data. In this problem, we will collect data in CSV format.

## Data Visualization

With the help of data visualization, we will try to get insight of data. In visualization, we can see the correlation in between features of dataset like HeatMap, Scatter plot.

## Feature Engineering

Feature engineering is the main step in ML model designing. In this, we will do the feature analysis, which feature is more relevant and which are less impacting the outcome.

In feature Engineering, we will do feature normalization. So because of high magntiude one should not dominate another feature.

It very important step in the machine learning model. It can drastically impact model performance.

* <https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html>
* <https://machinelearningmastery.com/an-introduction-to-feature-selection/>
* <https://towardsdatascience.com/a-feature-selection-tool-for-machine-learning-in-python-b64dd23710f0>

## Train & Test Dataset

We will Split the dataset into train and test, Training set we will use for our training and testing set for model validation.

## Model Training

Training selected model on train dataset and validating on Training set.

## Model Testing

Testing is the process to test the model performance or accuracy on test data set.

(Validating overfitting and underfitting)

## Model Tuning

In tuning, we will try to tune our algorithms hyperparameter to get high accuracy and performance on test and train set with the help of GridSearch algorithm.

## Finalizing Model

Selecting best final model for production promote.

## Production Deployment

For production deployment we can use any Python framework, we will use Flask for your production deployment. We will create Rest Endpoint, So service will be available as rest API.

References:

* <https://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://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5>
* <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>