Machine Learning Nanodegree Engineer

|  |  |
| --- | --- |
| Capstone Project  Bharat Singh  Jan-19 |  |

Flight Delay Prediction

Table Contents

[Definition 2](#_Toc534441385)

[Project Overview 2](#_Toc534441386)

[Problem Statement 2](#_Toc534441387)

[Matrices 2](#_Toc534441388)

[Confusion Matrix 3](#_Toc534441389)

[Accuracy 3](#_Toc534441390)

[Recall 3](#_Toc534441391)

[Precision 3](#_Toc534441392)

[ROC-AUC 3](#_Toc534441393)

[Analysis 3](#_Toc534441394)

[Data Exploration 3](#_Toc534441395)

[Exploratory Visualization 4](#_Toc534441396)

[Analysis 5](#_Toc534441397)

[Data Exploration 5](#_Toc534441398)

[Exploratory Visualization 5](#_Toc534441399)

[Algorithms and Techniques 5](#_Toc534441400)

[Benchmark 5](#_Toc534441401)

[Methodology 5](#_Toc534441402)

[Data Pre-processing 5](#_Toc534441403)

[Implementation 5](#_Toc534441404)

[Refinement 5](#_Toc534441405)

[Results 5](#_Toc534441406)

[Model Evaluation and Validation 5](#_Toc534441407)

[Justification 5](#_Toc534441408)

[Conclusion 5](#_Toc534441409)

[Free-Form Visualization 5](#_Toc534441410)

[Reflection 5](#_Toc534441411)

[Improvement 5](#_Toc534441412)

[Quality 5](#_Toc534441413)

[Presentation 5](#_Toc534441414)

[Functionality 5](#_Toc534441415)

[References 5](#_Toc534441416)

# 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:**

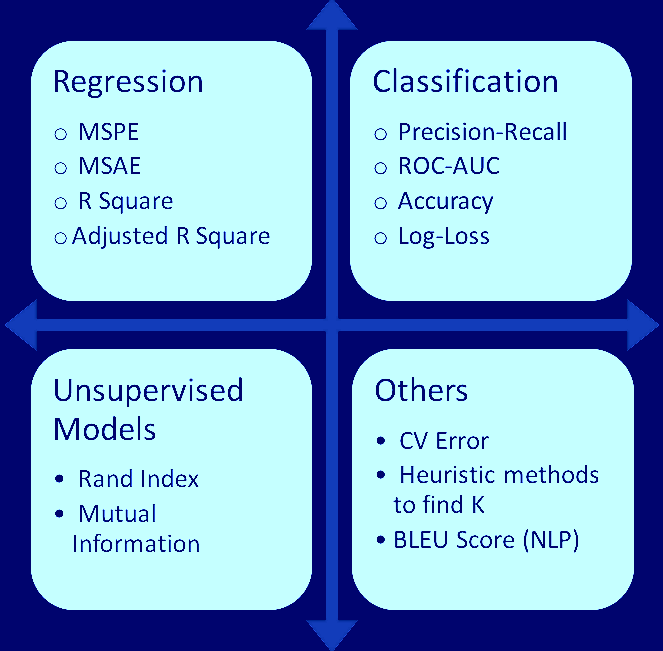
<https://www.kaggle.com/fabiendaniel/predicting-flight-delays-tutorial/data>

## 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

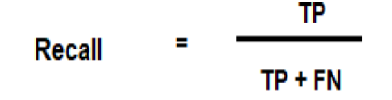
Accuracy is a metric, which is required to check the model accuracy on the unseen data.

<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.”



### Precision

### ROC-AUC

# 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

## Algorithms and Techniques.

# Analysis

|  |  |
| --- | --- |
| Data Exploration Data for flight delay prediction has been taken from Exploratory VisualizationAlgorithms and TechniquesBenchmark |  |

# Methodology

|  |  |
| --- | --- |
| Data Pre-processingImplementationRefinement |  |

# Results

|  |  |
| --- | --- |
| Model Evaluation and ValidationJustification |  |

# Conclusion

|  |  |
| --- | --- |
| Free-Form VisualizationReflectionImprovement |  |

# Quality

## Presentation

## Functionality

# References

<https://developers.google.com/machine-learning/crash-course/classification/accuracy>

* <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://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>