Machine Learning Nanodegree Engineer

Capstone Project
Bharat Singh
Jan-19

UDACITY

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:

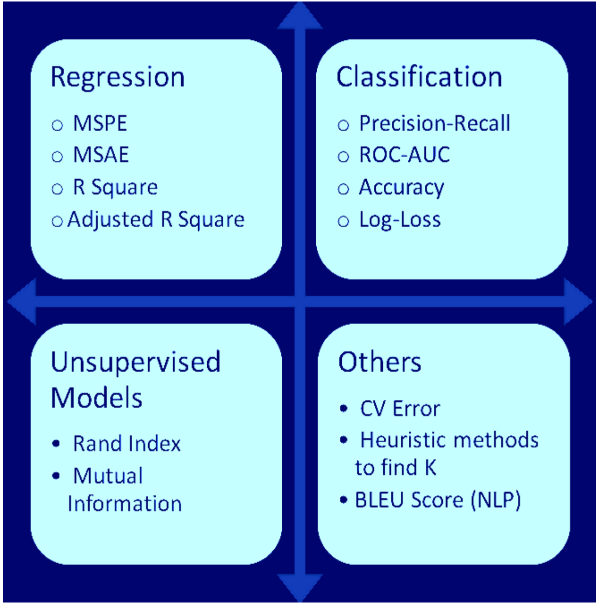
https://www.kaggle.com/fabiendaniel/predicting-flight-delays-tutorial/data

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

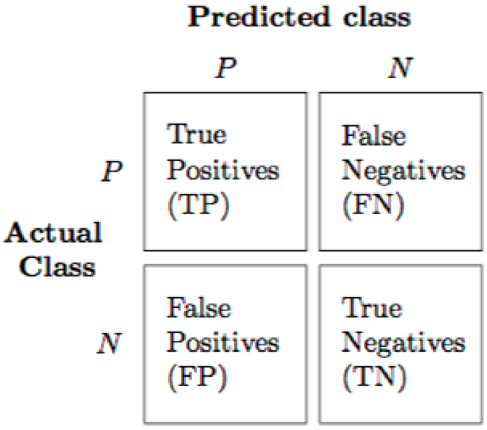
For validation of machine learning model preformation on unseen data or for verifying, system have generalized well for unseen data or not. We need some evaluation matrices. Machine Learning have different matrices for different type of problem. Like Classification, regression or clustering etc..



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.



		Predicted class	
		P	N
Actual Class	P	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

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."

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

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 Visualization

Algorithms and Techniques

Benchmark

Methodology

Data Pre-processing

Implementation

Refinement

Results

Model Evaluation and Validation

Justification

Conclusion

Free-Form Visualization

Reflection

Improvement

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