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.

Machine Learning Engineer Nanodegree

Table Contents

Machine Learning Engineer Nanodegree

Domain Background:

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

There are number of factor, 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 prediction of flight delay.

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

Problem Statement:

This problem is moreover related to airline industry. If you don't have any prior knowledge about the delay of particular aircraft. In Respect to airline, it can block your number of resource and in Respect to passenger, you can miss his meetings etc.

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

In this problem we will try to predict delay in aircraft to the reach at destination.
It's a binary classification problem. We will classify that upcoming flight will reach on time or will it be delay to reach at destination airport. For overcome this problem, we need to implement this project.
It's a Supervised Classification Problem.

Dataset and Inputs:

For this problem, we have dataset in CSV format. These CSV contains all the information related to airport, flight and airline, which is necessary to for it.

Data Files:

File Name	Description
Flights.csv	Flight schedule information
Airports.csv	Airport information
Airline.csv	It's containing the airline details

Airline.csv

Feature Name	Data Type	Description
IATA_CODE	string	IATA Code for airline. It's unique identifier of airline
Airline	string	Airline Name

Machine Learning Engineer Nanodegree

Airport.csv

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

Flights.csv

Feature Name	Data Type	Description
YEAR	Integer	Year of travel
MONTH	Integer	Month of journey
DAY	Integer	Day of Journey
DAY_OF_WEEK	Integer	It day of week
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 on airport
DISTANCE	Integer	Distance between ORIGIN_AIRPORT and DESTINATION_AIRPORT
WHEELS_ON	Float	Landing time on runway
TAXI_IN	Float	Reached on gate
SCHEDULED_ARRIVAL	Float	Schedule time to reach on gate
ARRIVAL_TIME	Float	Actual arrival time
ARRIVAL_DELAY	Float	Delay in arrival
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

Machine Learning Engineer Nanodegree

Solution Statement:

As this is a regression problem, so we will use only supervised regression machine Algorithms to tackle this problem. But before applying algorithms we will 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 regression, 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

We are planning to train our model on Logistic Regression, SVM, and Decision Tree and neural Networks. At the end on the bases of evaluation matrix we will selected one model which will perform best for our problem.

Benchmark model:

I am planning to use multiple model to train and test flight delay prediction. We are going to use below given list of models

1. Linear Regression
2. SVM (Support Vector Machine)
3. Decision Tree
4. Neural Network

We will you one as benchmark model for other. Whichever model will perform best for our problem, we will chose that as final 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

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

Actual Values

	Positive (1)	Negative (0)
Positive (1)	TP	FP
Negative (0)	FN	TN

Predicted Values

Actual Values

Machine Learning Engineer Nanodegree

Accuracy

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

$$Accurracy = \frac{\text{Total Correctly classified}}{\text{Total Sample}}$$

$$Accurracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Precision

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

$$Precision = \frac{TP}{TP + FP}$$

Recall or Sensitivity

Proportion of correctly classified form the given positive sample.

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

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

Programming Language and Libraries

- Python 3.X
- Tensorflow
- Scikit-learn

Data Collection

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

Machine Learning Engineer Nanodegree

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.

Feature Engineering

Feature engineering is a 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 magnitude one should not dominate another feature.

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 under fitting)

Model tuning

In tuning, we will try to tune our algorithms hyper parameter to get high accuracy and performance on test and train set.

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://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