Capstone Project

Bharat Singh

Dec, 2018

Flight Delay Prediction More than 15 Minutes Machine Learning

Abstract:

The project proposal is designed for creating a machine learning model for predicting flight delay more than 15 minutes for one lag. So that ground staff and network team can plan for aircraft scheduling and ground handling staff 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: 1](#_Toc533598573)

[Problem Statement: 2](#_Toc533598574)

[Dataset and Inputs: 2](#_Toc533598575)

[Solution Statement: 5](#_Toc533598576)

[Benchmark model: 6](#_Toc533598577)

[Project Design: 7](#_Toc533598578)

[References: 9](#_Toc533598579)

# 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, Flight reach on 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.

We can predict by airline delay by past history. With this history, 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 and ground staff management.

Problem Statement:

This problem is moreover related to airline industry. If you don’t have any prior knowledge about the delay of particular aircraft. It can block your number of resource.

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 total delay in aircraft to the reach at destination. For overcome this problem, we need to implement this project.

It’s a Supervised Regression Problem. In this problem, we need to predict flight delay.

Business benefit of this problem, resources will not go wastage and you can plan resources prior.

Dataset and Inputs:

For this problem, we have dataset in CSV format. These CSV contains all the information,

which is necessary to do the prediction.

Data Files:

Sr. No

File Name

Description

1

agent\_details.csv

Details for the agent.

2

ticket\_details.csv

Containing ticket details of group

3

group\_pnrs.csv

Group pnr details

4

group\_details.csv

Agent group information.

agent\_details.csv

Sr. No

Features Name

Feature details

1

AGENT\_TYPE

Details for the agent.

2

ticket\_details.csv

Containing ticket details of group

3

group\_pnrs.csv

Group pnr details

4

group\_details.csv

Agent group information.

ticket\_details.csv

Sr. No

Features Name

Feature details

1

agent\_details.csv

Details for the agent.

2

ticket\_details.csv

Containing ticket details of group

3

group\_pnrs.csv

Group pnr details

4

group\_details.csv

Agent group information.

group\_details.csv

Sr. No

Features Name

Feature details

1

agent\_details.csv

Details for the agent.

2

ticket\_details.csv

Containing ticket details of group

3

group\_pnrs.csv

Group pnr details

4

group\_details.csv

Agent group information.

group\_details.csv

Sr. No

Features Name

Feature details

1

agent\_details.csv

Details for the agent.

2

ticket\_details.csv

Containing ticket details of group

3

group\_pnrs.csv

Group pnr details

4

group\_details.csv

Agent group information.

Solution Statement:

As this is a classification problem, so we will use only supervised classification machine

Algorithms to tackle this problem. But before applying algorithms we will do the data pre-

Processing.

1. Missing value handling

2. Feature selection

3. Categorical and continues feature processing

4. Driving new features

5. Normalization

6. Visualization (for getting more intuition about the data)

After completing the data pre-processing, Data will be divide into train and test for our

Model training and validation. We will apply ML algorithms to predict the group booking

Cancelation rate. There are several algorithms for classification.

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:

We are going to use multiple model to train and test our problem. Like we are going to use

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.

6. 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.

1. Confusion Matrix

Confusion matrix is used for validating classification machine learning model. It’s table

representation of outcome.

C:\Users\S727953\Desktop\Project\_MLND\Capstone Project\ConfusionMatrix.png

C:\Users\S727953\Desktop\Project\_MLND\Capstone Project\AccuracyMesures.png

2. Accuracy

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

predicting correctly.

3. Precision

Precision is a measure that tells us what proportion of patients that we diagnosed as

having cancer, actually had cancer.

4.Recall or Sensitivity

Proportion of correctly classified form the given positive sample.

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

1. Programming Language and Libraries

Python 3.X \*

Tensorflow \*

Scikit-learn \*

2. Data Collection

For implementing machine learning model, we need data. In this group booking

problem we will collect data in CSV format.

3. 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.

4.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.

5. 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.

6. Model Training

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

7. Model testing

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

(Validating overfitting and under fitting)

8. Model tuning

In tuning, we will try to tune our algorithms hyper parameter to get high accuracy

and performance on test and train set.

9. Finalizing model

Selecting best final model for production promote.

10. 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>