Data Mining Project

(Spring2020)

May 15, 2020

1 Project Description

The U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics tracks the on-time performance of domestic flights operated by large air carriers. Based on your success they have hired you as a machine learning and data science expert to solve some problems related to flights data; when a particular flight is delayed, canceled, and diverted.

The dataset that have provided has summarized information of on-time, delayed, canceled, and diverted flights. Which is published in their Consumer Report of 2015. You as a principle scientist are required to extract different patterns and meaning full information using all or any machine learning algorithm.

2 Dataset

The Dataset consists of 3 files with a compressed size of approx. 192 MB is uploaded on google classroom. It has 3 csv files; namely airlines.csv airports.csv and flights.csv. It is up to you as a lead scientist how you use these raw data files. Either you join them together or process them individually is up to you.

3 Evaluation

As there are no straight guidelines, you guys have an open field on how you use all what you have learned in this course. So, obviously every student will have a unique solution. This eliminates the possibility of coincidental or deliberate attempts to similar solutions. Furthermore, you can apply linear/ polynomial regression, classification, clustering etc. algorithms to find pattern and mine information.

You will be graded individually, on the amount of effort, extraction of meaningful information and predictions etc. Marking will be relative to the performance of the whole class. This means discussing your solutions with other is probably a bad idea and will make your solution not unique, which will result in your solution being ranked lower in class ranking thus losing marks.

Marks will be awarded based on ranking system. There are three brackets of solutions

- Best
- Medium
- Worse

So try to position yourself in the best bracket.

4 Guidelines

Dataset is attached in the project. This Ipython notebook (https://www.kaggle.com/adveros/flight-delay-eda-exploratory-data-analysis) consists of an Exploratory Data Analysis (EDA) of the data which can help you in understanding the data with visualizations. This file is not compulsory for you to follow but it can give you a jump start in starting your project quickly so do give it a look. You are required to submit two things on submission date. A solution file (Code) and a report which should show your results and analysis of YOUR solution. You can code the solution in any programming language in which you are comfortable in. The project can be done in groups of two.

Cheating/ plagiarism is a BIG NO and will lead to **F grade** in course if any part of the project is found plagiarized. Try to propose a solution yourself and remember no solution is a wrong solution for this project. Some solutions would be good others not.

Good luck