FINAL PROJECT REPORT

Team Name: Flight Delay Analysis

INTRODUCTION:

According to a 2010 report by the US Federal Aviation Administration, the economic price of domestic flight delays entails a yearly cost of 32.9 B dollars to passengers and airlines.

More than half of that amount comes from the pockets of passengers who not only lose time waiting for their planes to leave, but they also miss connecting flights, spend money on food and have to sleep on hotel rooms while they're stranded.

what are the causes for these delays?

In order to answer this question, we are going to analyze the provided dataset, containing up to 1.936.758 different internal flights in the US for 2008 and their causes for delay, diversion and cancellation.

The data comes from the U.S. Department of Transportation's (DOT).

The Dataset used has 29 columns which will help us to identify the possible delays in the flights thereby helping people to save their money and time, also helping the economy of the country by a surprising lot.

We have done some pre-processing and found out that there were many null values, we've dropped all null values as there are still some significant number of values to carry out our analysis.

Previous Work

A significant amount of research work has been done on the topic in the past years and by developing our research papers we have found out that that cancellation prediction is very difficult to predict and has a very low accuracy score

So we've tried to name out some questions which may help the passengers to decide whether they can take a flight or not on the basis of the time of the day/day of the week and other factors as machine learning algorithms didn't yield a good result

Further we've also used some machine learning algorithms such as KNN, Logistic regression and Deep Neural Networks(which didn't workout that well).

Proposed Solution:

We've done the prior EDA and cleaned the data and done the required pre-processing

The questions we tried to tackle were:

1) When is the best time of the day, day of the week, week of the month to travel to minimize delay?
2) Do older planes suffer more delay?
3) How does the number of people flying between different locations change over time?

We've only used python and no ML model to answer these questions.
Further we used some ML models

RESULTS:

Cancellation Classification Summary:

Unsurprisingly, all of the machine learning algorithms presented here did a poor job predicting cancelled flights. Only the support vector classifier was able to correctly predict any cancelled flights, but only about 2% of them. If airlines do employ machine learning models to predict cancellations, they are most likely more sophisticated and include all sorts of data that we don't have access to here. Namely, current weather and aircraft/airport maintenance data. It makes sense that if you didn't need that current data to predict cancellations, there would be a lot fewer last-minute cancellations.

CONCLUSION:

Even so, the models did not do a great job predicting flight delay times. The KNN regressor did the best job with and R-squared score of 0.139, while both the linear regressor (with Ridge regularization) and the neural network (with 3 hidden layers each of size 150) had R-squared scores lower that 0.1. Similarly to the cancellation prediction models, these could be augmented for real-time data to give better results.

The features used for predicting delay times did not use data that could not be know days/weeks in advance of the flights. However, the argument could be made that the averaged features include "future" data from the perspective of many of the flights. I don't think this introduces any significant concerns of "data leakage", as there is no reason to expect these averages to be changing significantly with time. Furthermore, any "production ready" models could be engineered to avoid this problem and only use averages of flights before the one being predicted.

Contribution of Each Member:

Vishnu Harith

(PES2UG20CS387):

Data Preprocessing and Cleaning.

Veeravalli Ruthvik

(PES2UG20CS571):

Exploratory Data Analysis and Cleaning.

Varun

(PES2UG20CS568):

Model Building and Evaluation.

INTERESTING INSIGHTS:

We really enjoyed doing the project and the pre-processing and Cleaning part took a lot of time as we had heard before getting to work on a real project was really fun and it will surely be helpful in the future.

APPENDIX:

Here's a beautiful correlogram from our EDA: