Capstone Project 1:

Project Proposal

I) What is the problem you want to solve?

How long will a flight be delayed at the destination Airport?

Who is my client and why do they care about this problem? In other words, what will my client do or decide based on my analysis that they wouldn’t have done otherwise?

My clients are Airlines passengers, vacationers, the Department of Transportation/National Aviation System.

No one enjoys spending extra time at the airport because flight delays have kept their plane grounded. The analysis will help my clients to be mindful of the airlines that have the longest delays as well as the airlines that have the lowest delays and cancellations. The analysis will help The Department of Transportation(DOT)/National Aviation System to implement regulations that will reduce delays and provide better protection for consumers.

What data are you using? How will you acquire the data? I am using this free dataset from Kaggle: 2015 Flight Delays and Cancellations: https://www.kaggle.com/usdot/flight-delays This data has 3 sources: airlines.csv, airports.csv, flights.csv.

Briefly outline how you’ll solve this problem.

For the prediction I am using the basic models. I created a data set that can be used for both visualization and Model Building. The purpose of visualization is to get a better inference of the data.

For the Split test, the dataset has been separated into training and Testing data so that prediction can be done on the testing data. I also import Tableau that can make the visualization more feasible.

The technical aspects used for visualization are Matplolib and seaborn. Including data manipulation: pandas, NumPy, modeling -sklearn: Linear Regression, Random Forest & Decision Tree.

Here are the steps that I will follow to solve this problem.

1)**Uploading the csv files into pandas**

2) Cleaning The Data

We used isnull() to find the missing values in the dataset and dropped the rows which had missing values using dropna ()

We also converted categorical variables using one hot encoding but we did not use any other new variables.

3) Exploratory data/ Visualization

Exploratory Data Analysis on the data to discover patterns, to spot anomalies, to test hypotheses, and to check assumptions with the help of summary statistics.

Data visualization to understand the data by placing it in a visual context so that the patterns, trends and correlation that might not be detected be exposed.

4)Create predictive models/In depth Analysis

Using basic predictive model like Linear regression, Random Forest & Decision Tree to

Predict how long a flight will be delayed at the Destination airport.

5) Communicate results- comparison of the models in terms of performances.

6)Conclusion

Feedback from my mentor.

Hi Vertuile,

Yes, we can keep working with modifying the features and see if we can get better results. I am not sure it will improve the model a lot.

Keep in mind that the important factors are that you perform the visualization of the statistical tests and the model and you know that the data is not sufficient. Add a discussion that will describe which other data might improve the results. or what analysis you think might improve the results.

I think that we should continue the next project. If we will have time after we finish the next project we can have a look at this analysis again.

Please let me know what you think.

Best

Shmuel