TransVoyant Challenge

Peyman Ghahremani

January 24, 2019

Question: If you were tasked with generating an insight showing the risk (probability) of a departure delay at an airport for the next seven days, how would you do it? Please outline your methodology in around one page, provide the seven-day forecast results, and the code used to produce the results.

Answer: In this project, I was given two sets of data. One was the scheduled flights for the airport KDCA and the other contains the departure data of all flights from all airports around the globe. Both data range from the 1st day of January until 31st of January of 2018. I did the followings in this project:

1. By excessive data cleaning using spark, I took the necessary features from the departure data and schedule data and joined the schedule and departure flights on the flights at KDCA and obtained the delayed flights which are flights whose actualflighttimestamp - filedflighttimestamp is positive which we call delayed flights and then grouped them by day and took the average which would stand for the average delay of the airport for the January days.

2. Transformed the resulting spark data frame into pandas in order to ease our statistical classification. The resulting data frame is just day and delay data frame. To be more precise, it is just 31 rows and each row represents the number of hours that the airport KDCA experieced delays (which is the average delay for the day from each flight),

3. Defined the threshold to be the median of the average delays for January and then defined a function which maps numbers bigger than 56 (56 is the median of the average delays in January) to 1 and numbers less than 56 to zero. Therefore, in our new data frame, the average delays would be substituted by 0 or 1. 1 stands for the delay and 0 stands for no delays.

4. Trained 3 classification models such as Logistic Regression, Random Forest Classifier and Stochastic Gradient Descent Classifier and attained the result that Logistic Regression and Random Forest Classifier do way better than Stochastic Gradient Descent Classifier based on their corresponding confusion matrices.

5. Since Logistic Regression and Random Forest Classifier had the same accuracy, appealed to their corresponding ROC curves and attained the result that Random Forest Classifier does better as its ROC curves possessed more area.

6. Based on our Random Forest Classifier model, there would be no delays on the last 7 days of January.

7. Using sparl.sql commands, I attained that The top 10 airports by most departing

flights for the first 30 days of January 2018 are: KATL, KORD, KLAX, KDFW,ZBAA, KDEN, CYYZ, KCLT, ZSPD, KSFO.

8. Our model predicted that the airport would experience no delays on the last 7 days of January which is .71 accurate.

9. My code is available in my github which is:
https://github.com/PeymanGuitar/Flight-Delays/blob/master/LastVersion.ipynb

Clearly if the flight is being delayed does have connections with Arrival data (if the airplane was available on time for the next flight), if an airplane needed maintenance, if the airport itself was suffering technical issues and last but not the least if the weather of the origin airport and the destination airport is good enough for a flight. Such data was not available in this project but I believe with this data in hand a better classification could be made.