Author: Ifrah Zeb
Email: ifrah_zeb@outlook.com
Date: 22/08/2021

Project Report: Assessing the New York Fire Service Response Time

Goal

The goal of the project is to examine the fire service's travel response time to an incident in the city of New York, and how other variables affect it. By finding links between travel time and what affects it, I can recommend how the fire service could improve travel response time.

Hopefully the recommendation would improve the travel time for fire services and thus increase the chances of a better outcome in the case of fire emergencies. Having researched fire service travel time, there is not a great deal of existing analysis, however one paper I read based on the UK fire service did examine travel time. In that case, it has steadily been increasing over the years, rather than, as would be hoped, decreasing- or at least staying the same. It was found that in 2018/19 travel time for primary fires (fires that could potentially be serious) increased by 11 seconds from the previous year, and 33 seconds from 2013/14.

The Dataset

My data was acquired from the New York Open Data website. This is a very reliable source of information, run by *The Mayor's Office of Data Analytics (MODA)* and the *Department of Information Technology and Telecommunications (DoITT)* . It is the same data used by the New York City Council. The dataset is very recent, updated with incidents up to May 7 2021. It seems that the data is updated annually, resulting in a continually growing dataset that is available to the public. It is admirable that they make such data publicly available. I would consider training a model to update and run it for changes every year if I was to work for MODA.

The data provided me with the target variable (y); travel response time to an incident and a very large dataset starting at ~ 8, 500, 000 rows. However there are a couple of variables that I would want more from:

- The location of the incidents- it was not granular/ explicit enough for me to plot onto a map using Tableau unfortunately.
- The data description provided is a good source of information, but if they could provide a key to what the numerical information in the location columns represents it would have been ideal, e.g. the police precinct district for each incident was presented as a numerical value.

Further exploration would involve finding the data to complete the latter point. However, the granular/ explicit location would be too hard to fill in (very manual) so I would change the input going forward if I did work for the MODA.

Author: Ifrah Zeb
Email: ifrah_zeb@outlook.com
Date: 22/08/2021

Whilst the data wasn't perfect it was still a very great source of information given in a downloadable CSV dataframe.

Columns chosen

I was surprised to find that all incidents were included in the data set, including those where no travel was recorded. I believe this should be a separate dataset to be examined in its own right to compare these zero travel incidents with incidents where travel was made. I found that Manhattan made 34% of these zero travel incidents whilst only being 19% of the New York population. It would be great to explore the reasons for this in future to perhaps lower the amount of calls made or to support the fire service staff in Manhattan more.

The columns chosen for the final dataframe are:

1. Season (extracted from the column `Incident Datetime`)
2. Incident Borough
3. Alarm Source Description
4. Highest Alarm Level
5. Incident Classification Group
6. Dispatch Response Seconds Quantity
7. Valid Incident Response Time
8. Incident Response Seconds Quantity
9. Incident Travel Time Seconds Quantity

I conducted one hot encoding (this is *cracking open* a column with different categories stored within such as the five boroughs of New York, giving each category a column where 0 is false and 1 is true for each incident) on all the categorical columns (columns 1, 2, 3, 4, 5, and 7) to make them compatible with the continuous target variable `Incident Travel Time Seconds Quantity`.

After cracking open the columns with one hot encoding the shape of my dataframe changed to 41 columns (from 9) with ~734, 000 rows.

Modelling

I performed a number of models on the data including different types of linear regression (scaled, non-scaled, lasso, ridge, and forward stepwise regression). Though linear models were not the best model the forward stepwise regression model was informative (it showed me the top 10 variables that affected the target column most strongly):

- The summer season, closely followed by the autumn season correlated strongly with travel time.

Author: Ifrah Zeb
Email: ifrah_zeb@outlook.com
Date: 22/08/2021

- The Queens borough, closely followed by the Richmond/Staten Island borough, were the only boroughs in the top 10 shown, which is curious because the latter borough is only 6% of the New York population.
- The phone as the source of the alarm greatly affects travel time which is in line with it being the most used source of alarm
- Bars as the source of alarm came in as the tenth variable, which surprised me. The relationship between bars and the fire service could be improved to improve the overall average travel time (e.g. putting more safety measures in place)

The decision tree model ended up being my best model to predict travel time- not overfitted (as the linear regression models) or under-fitted (as the KNN model), with approximately 81% train accuracy and 77% test accuracy.

Conclusion

In conclusion the decision tree model is a good predictor of travel response time (multi_variable), while the forward stepwise linear regression model is a great model to view which variables most affect the travel response time.

Though I would have to gain more information before making any strong recommendations such as why Richmond/Staten Island holds significance to travel response time more so than Brooklyn, the largest borough, these two models taken together are instructive in determining where such investigations are needed.