NYC Open Data Final Report

Jack Wolfgramm

June 2022

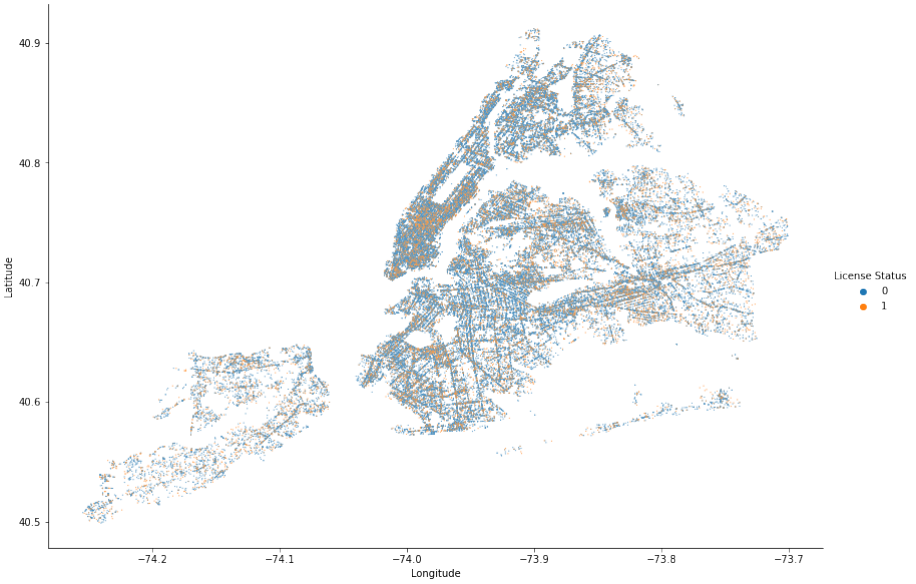

Figure 1: Scatter plot showing the location of business licenses within NYC, the color indicates whether the license is active or not. The outline of Central Park can be seen as there are few businesses inside.

1 Project Idea and Goal

The central plan of the project was to create a model that would predict duration of operation for a given business license by utilizing clustering. This would be useful to any business which was considering a license. For example, the profitability of any business venture depends on the duration of time the venture will be active. Having a more accurate regression model for the duration of license validity would provide value for businesses.
However, the models created without feature analysis often perform poorly,

which drove me to try to use clustering to improve model performance. The key idea is that the duration of business license duration in the nearby vicinity should provide good information, and hence be a good feature to add. Adding the clusters increased the R-squared of the model by .2, which is a rather large increase.

2 Data Sourcing

New York City keeps many datasets on `https://opendata.cityofnewyork.us/`. The dataset in question is located at `https://data.cityofnewyork.us/Business/Legally-Operating-Businesses/w7w3-xahh/data`, and contains around 150,000 entries. Features include the date the license was given, when the license will expire, whether the license is active, and the type of license.

3 Data Cleansing

A few things needed to be cleaned from the data. The main ones were, first, many licenses were located at a single location in Pennsylvania (The reason for this is unclear, but it may be the default location given). Second, some entries had license creation dates which were later than the license expiration date.

4 EDA

I discovered some interesting facts. First, licenses which were currently active had a longer average time between creation and expiration (likely a survivorship effect). Second, the average duration of license validity was fairly consistent across the most common types of business licenses, as shown by the box plots in figure 2. Finally, through exploring running k-means clustering on the businesses yielded the interesting insight that the duration of license validity between the clusters was statistically significant (even if the average cluster size contained 300-500 businesses).

5 Modelling

When it came time to actually build a regression model, it was important to tune not just the hyperparameters of the model, but also the number of clusters, and the way that the clusters were created (i.e. how heavily to weigh the distance vs the time since the license was created). Since creating clusters can take quite a while, I used Optuna to help find the ideal set of paramters. Optuna uses more advanced algorithms to find the maximum in the parameter space, and tends to greatly outperform not just grid search, but also random search.
Building a data-leakage free pipeline was an unforeseen but highly instructive difficulty of the modelling process. I had to create my own function that would

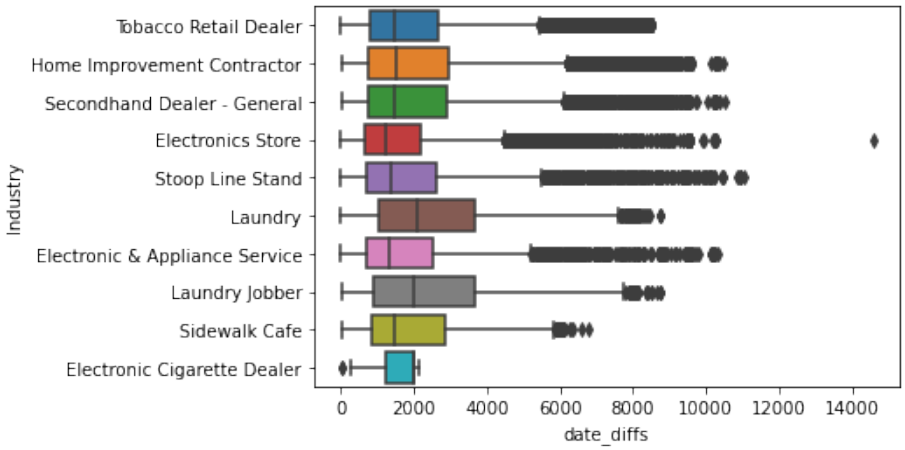

Figure 2: Box plots of the most common business licenses. The means are similar though statistically significant.

cluster only over the training data, and when running over the test data assign the new data to old clusters. Additionally data leakage due to one hot encoding was an interesting technical issue!

6 Results and Future Directions

The final model created had an R-Squared improved by about 0.2, though it was only about -0.051. This is still a significant improvement that could possibly help businesses make decisions.
However, there are certain limitations of the model. According to my mentor, the data is very "wishy-washy", which may account for the low R-squared. In addition, the model performs very well on some subsets of the data, and poorly on others. This is would be a good place to look in the future to see if either other features, or a different kind of model would improve the performance. Overall, I learned a great deal from this project and hope to do more in the future!