Data Science Project:
Predicting Business Survival Rate within London Boroughs

Dewi Pori

August 2021

1. Introduction

1.1 Background:

Every year in London new businesses emerge in various industries, e.g., food & beverages, travel & hospitality, real estate, entertainment, health & wellness and many more. Recommending and approving new businesses for financing can be difficult, especially when the decision relies on determining potential business success according to the wider geographical space and neighbourhood conditions & safety.

1.2 Problem:

The project aims to provide a prediction on likelihood of business survival rate based on presence of other businesses within the borough and area security. For the scope of this project area security refers to the volume & outcome of crimes committed within the area. Data that can help to tackle this include popular venue types in London, enterprise survival rate according to different London Boroughs, crimes committed in different areas, outcome of crimes and population by borough.

1.3 Interest:

Main audience of this report include Banks, other Financial Institutions and Government bodies, as prediction outcome can assist in either approving new finances or initiate a review and amendment of existing policies & regulation relating to public safety.

2. Data Acquisition, Scope & Cleaning

2.1 Overview

The final datasets utilised for this project are based on the following:

- **London Business Rate**: Containing volume, birth and death rate of enterprises according to relevant London Borough (period: 2012-2019; source: **Wikipedia**, UK GOV Data).
- **London Crime Data**: Containing information relating to reported crime and outcome of the crime according to the relevant London Borough (period: 2018-2020; source: **Kaggle**).
- **London Venue**: Containing latitude, longitude, popular venue types and venue names according to the relevant London Borough (period: 2021; source: **Foursquare**).

The data above will be used to determine business survival rate, i.e., inverse of death rate, by looking into any correlation between the rate and a) available venues, b) types of crime committed within the area and c) outcomes of the crime committed. Thus, the better the survival rate, the chances of getting approval for financing will increase.

2.2 Data Acquisition & Cleaning:

The project contains three main datasets. These were obtained and cleaned as follows:

London Business Rate

The first data set has been webscraped from Wikipedia to obtain London Borough information. Further manual appending was applied to include the population of each borough as webscraping could not be done on the Statista website.

The second set of data was obtained from gov.uk site, which contains business birth/death rate according to London Borough. Original data contains information from 2004 to 2019, for the purpose of the modelling, I have filtered this to 2017-2019 to have a more up-to-date trend on birth/death rate. These two datasets are then joined, resulting in 99 records (3 years for each of the 33 London Boroughs).

Raw data source:

- **https://www.statista.com/statistics/381055/london-population-by-borough/**
- **https://en.wikipedia.org/wiki/London_boroughs**
- **https://data.london.gov.uk/dataset/business-demographics-and-survival-rates-borough?resource=1cff3cf1-b4a8-4194-b0ba-42839e94b432**

London Crime Data

Datafiles are uploaded on a monthly basis to Kaggle. Python was utilised to append all CSV files into a consolidated file, which contains 1,038,740 records. Further clean-up of the data consists of:

- "Year" column introduced to simplify data filtering to year instead of using month.
- "Borough" column introduced as original column "LSOA name" contains Borough name AND 4-letter code at end of value. This will be used to join with other datasets.
- "Resolved" column introduced to identify whether the crime has been resolved. This is by identifying whether the column "Last Outcome Category" has the string value "Offender" or not. Note: From original data, having value "Offender" refers to the following outcome

Last outcome Category
Offender deprived of property
Offender fined
Offender given a caution
Offender given a drugs possession warning
Offender given absolute discharge
Offender given community sentence
Offender given conditional discharge
Offender given penalty notice
Offender given suspended prison sentence
Offender ordered to pay compensation
Offender otherwise dealt with
Offender sent to prison

Raw data source: **https://www.kaggle.com/hark99/london-crime-data**

London Venue

As part of the project requirement, Foursquare was leveraged to obtain the top 10 popular venues and categories within 500m of each London Borough. This results in a total of 267 records.

2.3 Feature Selection:

Most of the data fields within the 3 datasets do not overlap one another. However, there were a few fields that were redundant from the London Crime Data. These redundant data includes, “LSOA Name” which is essentially the name of the Borough, “Context”, “Crime ID” and “LSOA Code”. All of these were dropped given that there are no other datasets that can be utilised to join the information.

For data analysis and visualisation purposes, these three datasets are kept separate, with only survival rate being appended accordingly to each dataset.

Upon further investigation, the period of data between the three datasets ranges greatly, with some having data backdating to 2004 (business survival rate), whilst another only having 2021 (Foursquare data). Given the data and time constraint, for the purpose of this project, we will focus on the annual information 2018-2019 for business rate and crime data and 2021 for venue category from Foursquare.

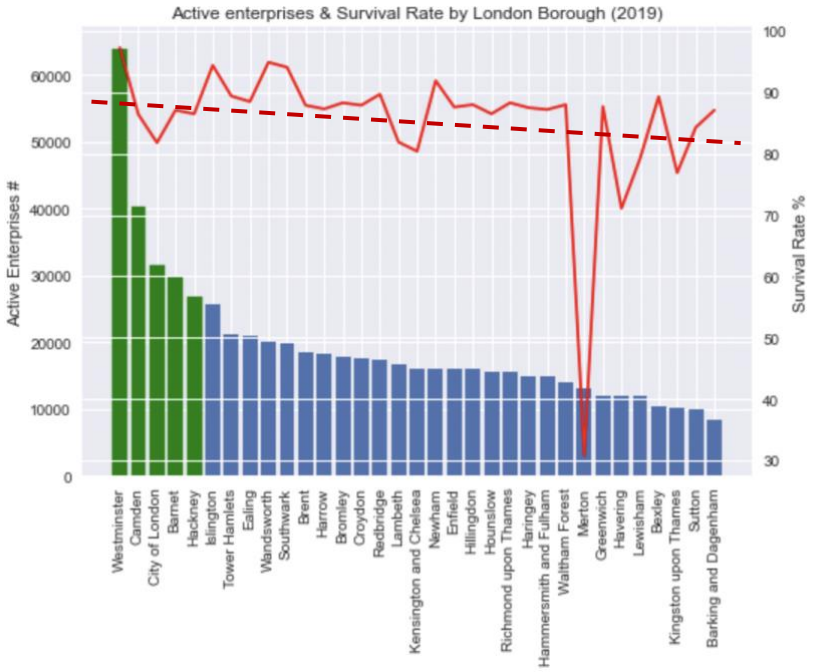
For modelling, I have joined all three datasets and **applied one-hot encoding** to “Crime Type” and “Venue Category” and summarising up to London Borough level. This will optimise the results as much as possible as unfortunately, business rate only exists on an annual basis and by London Borough only. Thus, in the following analysis, to obtain the business survival rate, I have mainly averaged the value by the specific sub-segments.

3. Methodology & Results:

3.1 Exploratory Data Analysis:

Top London Borough for Business Activities

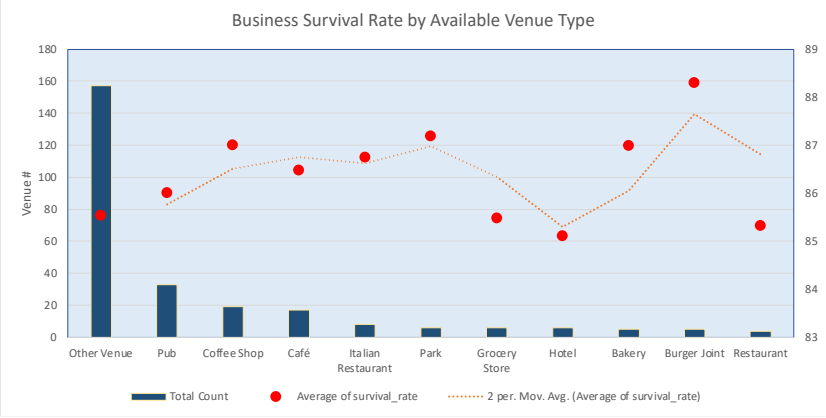
By volume of active enterprises, Westminster, Camden, City of London, Barnet and Hackney are the most popular borough with survival rates all above 80%. City of London has the lowest rate.



Graph 1 – Enterprises in London Borough

Popular Venue Category by Average Survival Rate

"Pub" turns out to be the most popular venue type when looking at all London Boroughs. However, based on the data, the availability of different venue type does not heavily impact the business survival rate as the rates do not have a clear pattern.

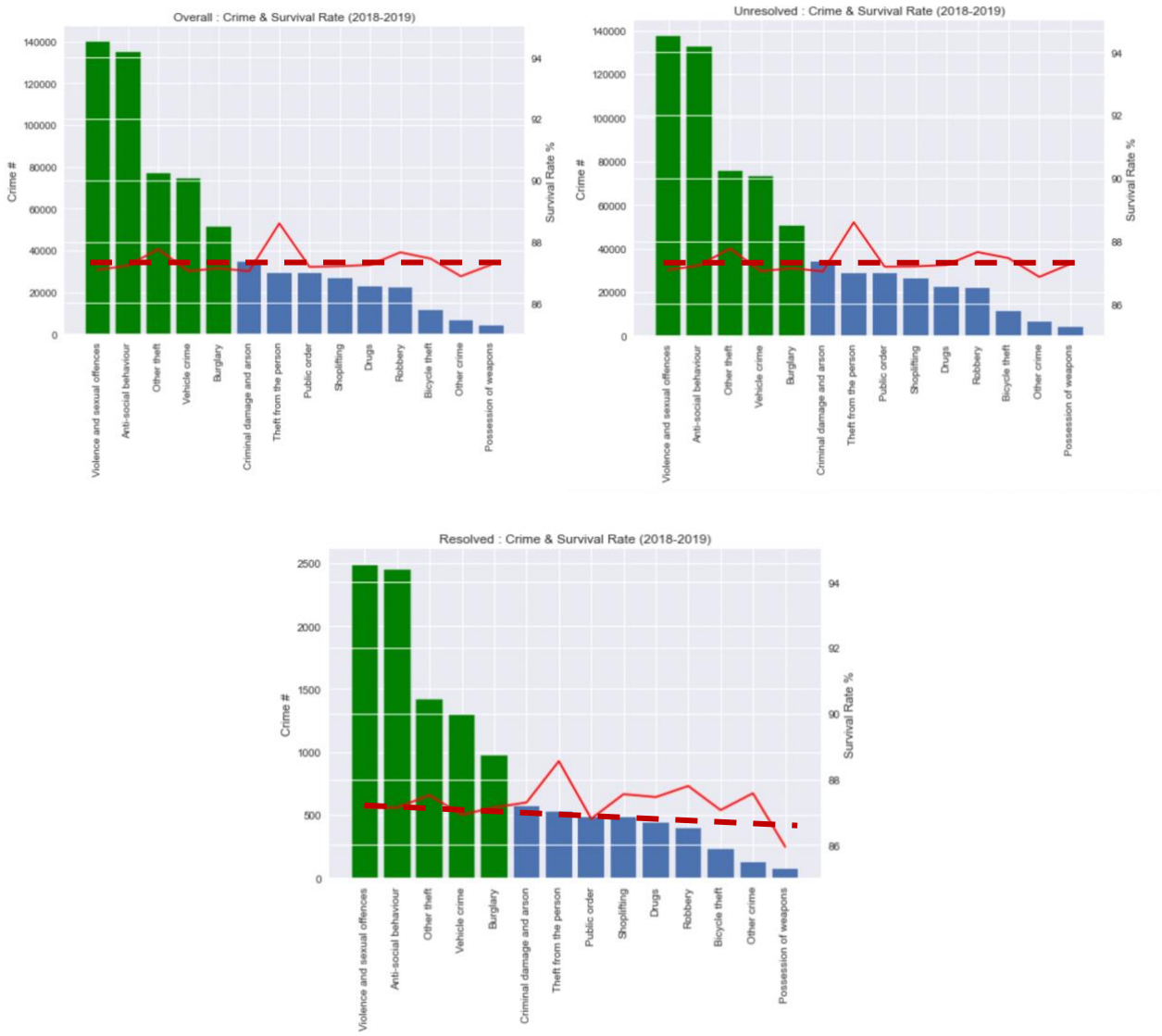


Graph 2 – Top 10 popular venue Type

Survival rate varies according to venue type; I want to highlight that the most popular venue types are related to food & beverage (7 out of 10) and the survival rate for these tends to be on the higher side. Thus, as part of further refinement recommendation, it could be worth grouping these venue categories into bigger segments.

Crime Type and Volume by Survival Rate

By area security, the top 5 main crime type are “Violence and sexual offences”, “Anti-social behaviour”, “Other theft”, “Vehicle crime” and “Burglary”.



Graph 3 – Performance of survival rate by crime type and resolve status

Findings shows that survival rate do not vary greatly by crime types (unlike venue categories). But counter-intuitively, we find that the overall survival rate drops as more crimes have been resolved – noted by the downward **red** gradient line in the “***Resolved”*** graph.

Assumptions:

- COVID impact not taken into consideration. The available popular venues taken from Foursquare reflects business environment pre-covid i.e., no rapid closure.
- The available survival rate applies to the popular venue as a whole and have not been weighted according to categories – due to low volume of data.

3.2 Modelling:

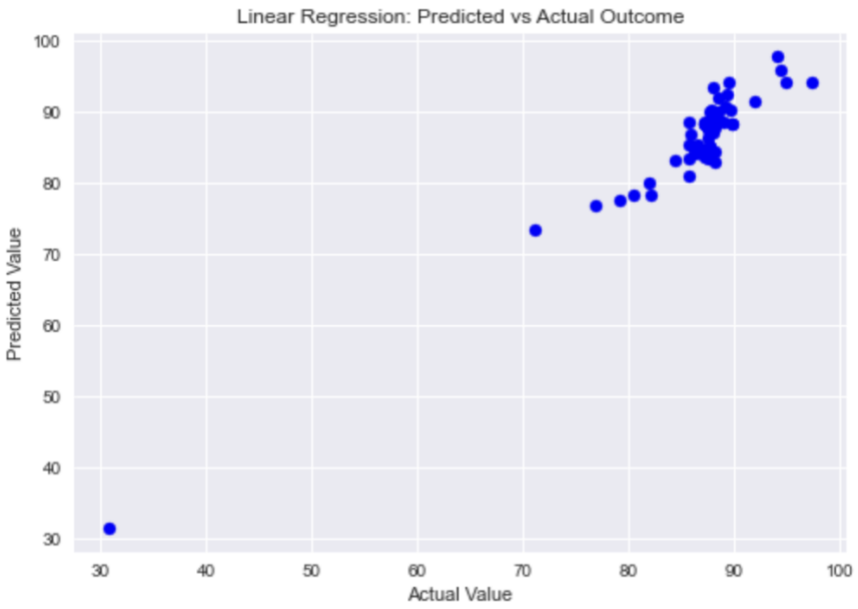
To demonstrate potential of the project, I have applied two common models for each modelling type: Linear Regression for regression models and Logistic Regression for Classification models. For evaluation, I have utilised Mean-squared-error (MSE), R-squared, confusion matrix, log loss and Jaccard score value respectively. Whilst there is various limitation to this project – which I will discuss in the next section (refer to Section 4 – Discussion), my main objective is proof of concept

Target Variable

As the aim of the project is to predict business survival rate, I have utilised the death rate of enterprises and calculated the inverse to get the survival rate.

Linear Regression

The scatterplot shows that predicted value are within the same range as actual value, with a 75.9% score on the R-square method and 138 on Mean-Square-Error. Despite low volume of data, the model has managed to obtain an acceptable range of accuracy.



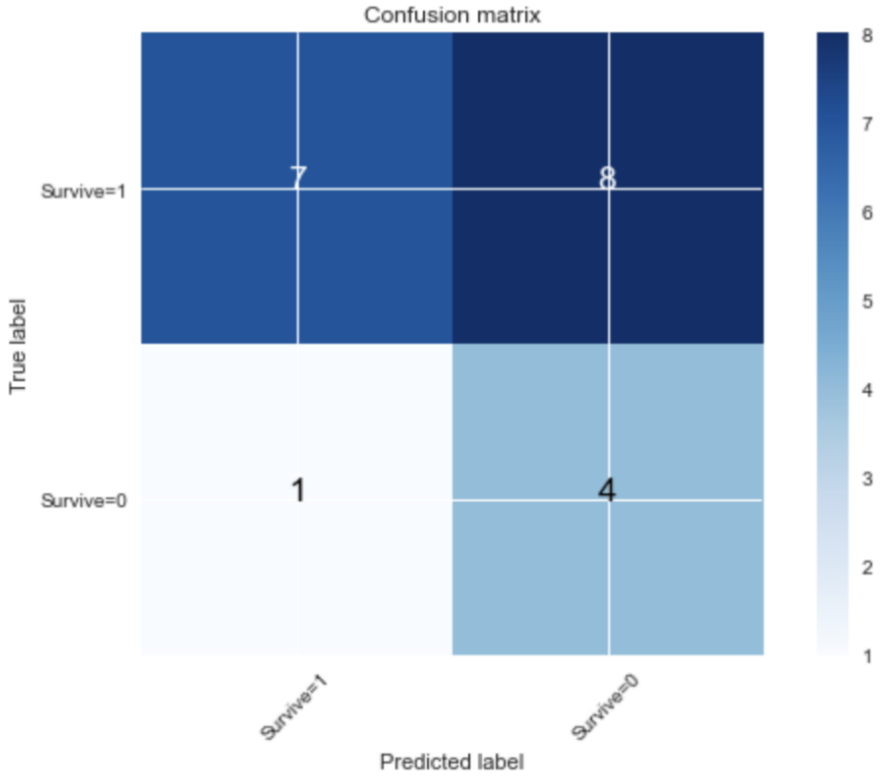
Graph 4 – Predicted value against actual value using Linear Regression

Evaluation Metric	Value
Residual Sum of Squares	138.3
Mean Squared Error	0.76
R-square	0.759

Logistic Regression

In contrast, the application of Logistic Regression – the classification model, proves to be inadequate. This mainly stems from low volume of data. There is also complexity in determining what constitutes as "survived" based on the survival rate, since most London Borough has a survival rate of greater than 80%.

The results show an accuracy of 30.8% (Jaccard), with log loss of 0.684. This implies the model was only able to accurately predict and classify 30% of the data and that the predicted probability diverges a lot from the actual value (log loss=1 indicates complete divergence from actual value).



Graph 5 – Confusion Matrix of the Logistic Regression outcome

Evaluation Metric	Value
Log Loss	0.684
Jaccard Score	0.308

4. Discussion

Due to lack of available data, the information is limited to yearly basis. Even with low volumes of data, the model shows promising results with the regression model providing a better predictor compared to classification - 76% vs 30% accuracy respectively.

As a short exercise with the intention of getting more data to improve the prediction, I created an alternate view where I did not split between training and testing. When just comparing the entire predicted value against actual value, results improved significantly as more data were available for training. Thus, I believe this model can be further strengthened with more data, especially having the business survival rate on a monthly basis instead of yearly basis. This was the main issue I faced when trying to split the survival rate into further sub-segments.

Since the success of the project hinges on the availability of data, assuming we can extract more, precise and recent data, the following are points to consider for follow-up or future projects for improvements:

1) Further deep-dive analysis on the survival rate by venue category greater segments. Clustering method (e.g., Hierarchical) can be used to group the categories. This may be easier to implement within the banking/financial sector, if there is least number of segments as possible.
2) Investigation into the counter-intuitive results of decreasing business survival rate with more resolved crime within the area. From here, we can determine whether it is coincidental, or if it is worth a review from the relevant government bodies.

5. Conclusion

In this study, I attempted to analyse the relationship among business survival rate, nearby popular venues and area security. Whilst the analysis can be further deep dived, the outlook of the regression model managed to obtain a good prediction accuracy. This implies that although the relation between the main variables are raw, there is potential in deciphering it further with more data.

I maintain my stance that this can be a useful model with primary usage within the banking/financial sector. For example, when a business financing application is received, the bank can run it through model to determine the survival rate and apply further policies or regulations.