Forecasting the Evolution of Vancouver's Business Landscape

Aakanksha Dimri, Keanna Knebel, Jasmine Qin, Xinwen Wang

2020/05/12

Executive summary

Our proposal introduces the context of the capstone project provided by Deetken. The problem statement is refined into quantifiable data science research questions. We then further deep dive into relevant data sets, elaborate on analysis methodologies, and propose final deliverable.

Introduction

In the continuing age of digital revolution, data is used ubiquitously among industries and individuals. Everyday problems, such as finding the quickest route home and tracking food deliveries, have been effectively tackled by data. The same approach could be implemented in the public sector, helping to inform policy decisions and future planning.

With COVID-19 governing our economy and lifestyle, it is critical to understand the evolution of the city's neighbourhoods and leverage that to predict potential future outcomes. One way to approach this problem, is to track the evolution of businesses in Vancouver's diverse neighbourhoods and develop/extract meaningful insights.

To address this we have defined the following research questions for the project:
- Will a business renew their license in the coming year?
- Geospatial summary of Vancouver's business landscape

Apart from observing changes in the business landscape, several other factors like access to public transportation, demographics of the region, public fund allocation etc. also affect this decision. To integrate these, we plan to use several datasets from the Vancouver city's open data catalogue.

The proposed final product consists of a data pipeline as well as a geospatial visualization of Vancouver's business landscape. Users will be able to locate a specific zone on the interactive map and view relevant descriptive information, such as business type distribution and census data [**figure 1**]. The data pipeline will pass processed input data of a specific business to a machine learning model and produce a predicted renewal probability.

Data Science Techniques

Data Sources

The primary dataset utilized in this project consists of all Vancouver Business Licence applications from 1997 to the current date. This data is made available as part of the city of Vancouver's Open Data Portal and regulated under the terms of the Open Government Licence – Vancouver. The most pertinent features present in this dataset are business type, location, and number of employees.

In addition to the business licence dataset, the Canadian census surveys provide another important source of data for this project. The census data is hosted on the Vancouver Open Data Portal and provides demographic information such as population density, average income, age distribution, and ethnicity. The current census dataset aggregates the demographic data by Vancouver neighbourhoods. As the project progresses, we may choose to further refine our model by obtaining census data aggregated at the postal code level.

Methodologies

To answer the proposed research question, we will commence with an initial data synthesis stage. This stage consists of combining the aforementioned datasets and wrangling the data into a suitable format. The main difficulty we will need to address during data synthesis is the different temporal and spatial scales among the collected datasets. Additionally, due to the size of the datasets, we will use a Postgres database to store the processed data.

After processing the data, we will proceed with building a baseline model using logistic regression. We choose logistic regression as it can produce a probability distribution of the predicted result, affording greater insight to the clients.

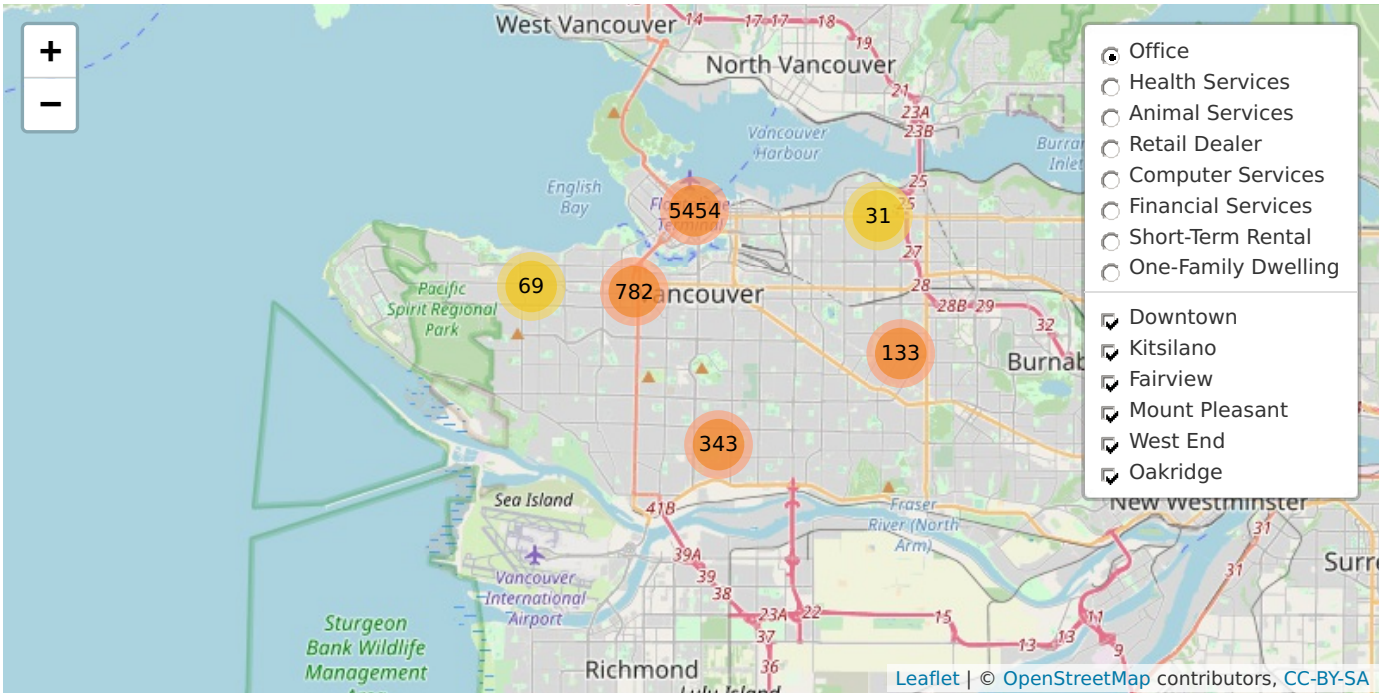

Figure 1: Descriptive and Simulated Information on Map

Once as the baseline model has been established, we will implement regularization techniques to mitigate overfitting and model complexity. We may also draw on techniques from survival analysis to examine business renewal over extended periods. In parallel with the development of predictive models, we will also create a visualization to display geospatial business information. We will use Python and deploy our result with Dash on Heroku.

The datasets pose some potential difficulties for us. First, we need to identify and address any existing temporal and geospatial correlation between the variables. Moreover, some variables in the model may be proxies for other factors that are not included in the model. For example, we have the number of employees in our dataset and we might find this factor to be significant in determining the renewal probability. However, it might be the capital invested that is actually causing this difference which is not included in the model but correlated with the number of employees. Thus, the number of employees is a proxy for the capital invested. As discussed before, how to combine features from different data sources will also be challenging.

Timeline