Forecasting the Evolution of Vancouver's Business Landscape

Aakanksha Dimri, Keanna Knebel, Jasmine Qin, Xinwen Wang

2020/05/15

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

With government expenditures totalling 40% of the Gross Domestic Product and public organizations accounting for 20% of the country's employment, the Canadian public sector has established itself as a pillar in the national economy. Due to its importance and potential for impact, it is critical that the public sector operates in an efficient and transparent manner. Accessible public data portals (e.g., Vancouver Open Data), in conjunction with data science techniques, represent one approach to address these demands.

In particular, the current socio-economic environment in Vancouver presents the challenge of balancing significant budgetary constraints with the increasing demand for medical and social services. Developing an understanding of how Vancouver's business landscape has evolved over time can help mitigate this challenge and gain insight on how to most how to most efficiently allocate the city's resources and services.

All businesses operating in the city of Vancouver must have a valid business licence which is required to be renewed every calendar year. As such, this dataset represents a yearly snapshot of the entire Vancouver business landscape. The renewal of business licences provides information regarding the spatial distribution, the temporal trends, and the volatility of businesses across Vancouver's neighbourhoods. While business licence renewal takes place on the individual scale, it is influenced by broader regional factors, such as the proximity to public transport, the demographics of the neighbourhood, and the national economic health.

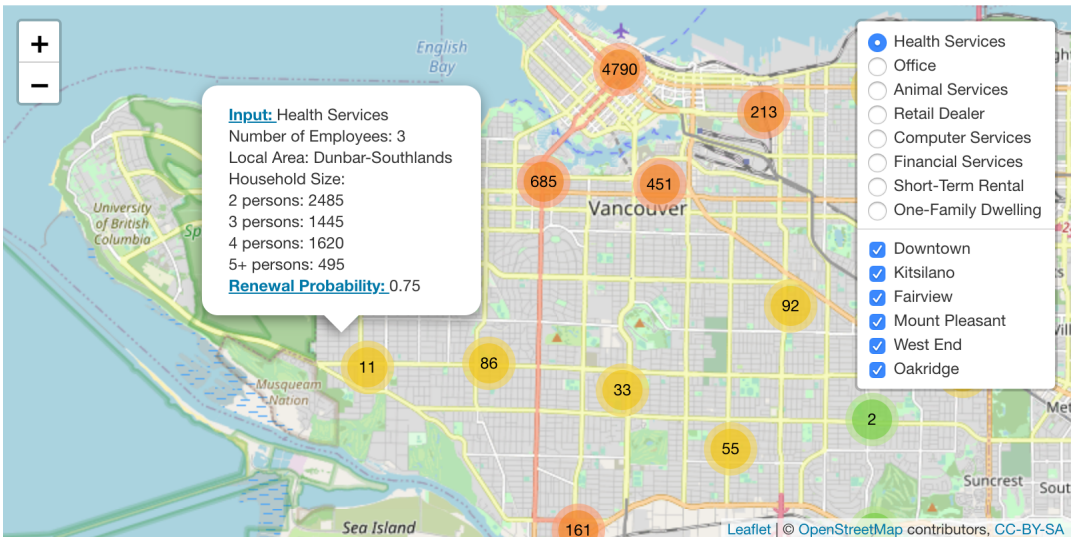

Figure 1: Descriptive and Simulated Information on Map

This proposal focuses on developing insight into Vancouver's business landscape which can be leveraged by policy-makers, planners, business owners, and others to improve efficiency. To achieve this we have established two main research objectives. (1) We will predict whether a business will renew its license, given a set of underlying factors. Given a

reasonable set of factors drawn from public data, we can also begin to interpret the model output to provide (2) a broader geospatial summary of the evolution of Vancouver's business landscape.

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.

Possible Data Sources

Data	Explanation	Spatial Scale	Temporal Scale
Business License 1997-2012	Business information and its license status for business in Vancouver area	Point location	Annual
Business License 2013-current	Business information and its license status for business in Vancouver area	Point location	Annual
Census Local Area Profile 2016	Census information in Vancouver	Local District	Annual
Census Local Area Profile 2011	Census information in Vancouver	Local District	Annual
Census Local Area Profile 2006	Census information in Vancouver	Local District	Annual
Census Local Area Profile 2001	Census information in Vancouver	Local District	Annual
Local Area Boundary	Geospatial boundary for Vancouver 22 local area	Local District	Not Applicable
Disablibity Parking	Location for disability parking space	Point Location	Not Applicable
Parking Meters	Parking meters location and rate in Vancouver	Point Location	Not Applicable
Business Improvement Area	Area where business join together to promote commercial viability of their district	Local District	Not Applicable

Methodologies

Synthesizing information across different data sources is a crucial initial step shared by both proposed research questions. It can provide an integrated summary of the business landscape by adding demographics and other social components and allows us to build a more valuable predictive model for renewal probability. The main difficulty in this stage is the different temporal and spatial scales among the collected datasets. For example, licences are applied annually and locations can be pinpointed on map, whereas census profile is collected every 5 years and aggregated to neighbourhood level. The best solution is to interpolate census data for a given year and build spatial configuration around an individual

business. 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. 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 information using Python and deploy the result.

The datasets pose some potential difficulties for us. First, we need to identify and address temporal and geospatial correlation between the variables. Moreover, some variables may be proxies for underlying factors that are not included in the model. For example, number of employees in the licence dataset might be a significant predictor. However, capital invested might actually be the lurking factor that is affecting both number of employees and success probability.

Timeline