The Deetken Group

THE UNIVERSITY OF BRITISH COLUMBIA

Understanding the Evolution of Vancouver's Business Landscape

UBC MDS Capstone Project Final Report

Aakanksha Dimri, Keanna Knebel, Jasmine Qin, Xinwen Wang

Mentor: Simon Goring

Date: 2020/06/23

Executive Summary

With a rising demand for future social services planning in Vancouver, the goal of this project is to provide data-driven strategies for individual businesses, as well as, a broader understanding of resource allocations across neighbourhoods. To achieve this, we used a wide range of machine learning methodologies to (1) develop a model to predict whether a business will renew its licence and (2) provide a geospatial summary of the business licence history and the constructed model. The predictive Light-GBM model constructed had a final accuracy of 0.58 and a recall of 0.57 (non-renewed licences). Our data pipeline has been designed with efficiency and flexibility for synthesizing diverse datasets. Thus, despite its limitations in regards to predictive capabilities, the final data product serves as a solid foundation for understanding Vancouver's evolving business landscape.

Introduction

Strategic urban planning provides a framework for achieving socio-economic objectives driven by actionable and sustainable development. Centralized city plans serve to reach these objectives through the coordination of efforts from the government, the private sector, and the community. Importantly, by planning and anticipating a community's future needs, city leaders are better able to allocate municipal spending, mitigate potential risks, and capitalize on opportunities.

Developing an effective strategic plan requires a critical assessment of a diverse set of factors, including budgetary constraints, local demand for social services, accessibility, employment, and housing. Accessible public data portals (e.g., Vancouver Open Data), in conjunction with data science techniques, represent one approach to address these demands. In particular, developing an understanding of how Vancouver's business landscape has evolved over time can provide insight into how to 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, the collective set of business licences 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 licence renewal takes place on the individual business 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.

This project 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 generate a machine learning model to predict whether a business will renew its licence, 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.

Data Science Methods

Descriptions and justications of methods used

Given that the business licence dataset is chronological data records of licence applications in each year, one crucial initial step was to wrangle the dataset so that it fits into machine learning context. We identified unique businesses using business names along with locations and split them into train, validation, and test sets. For businesses that have multiple entries in one year, duplicates are sorted by extract date and only the latest record is kept for better data accuracy. In order to answer the proposed research question of whether a business is going to renew its licence next year, we lagged each unique business one year ahead so that next year's status can be used as the target variable in our model.

Synthesizing information across various datasets is another important step shared by both research questions. To extract valuable information from the census datasets, we manually went over all variables in each of the census years. The sub-categories in some census variables might be called in different names from year to year. For example, in industry distribution of labour forces, the sub-category 'industry - not applicable' is named ‘Industry - NAICS2012 - Not applicable’ in 2016. We used various techniques such as regex to fix entry errors and wrote functions for data cleaning to make sure the pipeline is fully reproducible.

After the tidy data was ready and before we could fit it into any model, we noticed that there is class imbalance issue. There are more renewed licences than not renewed ones. Therefore, instead of focusing just on getting high accuracy, we were more concerned about recall score. This is because the recall score of the minority class measures number of predicted non-renewed licences compared to actual non-renewed licences. By avoiding situations of high accuracy but low recall, we tried to be conservative about making a false sense of success for individual businesses.

In order to enhance model performance, we conitnued to extract more information from the licence dataset and added features that are tailored to individual businesses. The parking meters and disability parking datasets were originally appended to licence dataset as counts in each neighbourhood. However, in reality, an entire neighbourhood is too broad of a range for people looking for a parking lot. Instead, we counted the number of nearby parking spaces for each unique business by calculating real distances between them and applying an adjustable threshold of 150 meters. Another geo-spatial feature we managed to extract from the licence dataset is the count of nearby similar businesses. This variable is meaningful because businesses are not random points on the map, but are clustered or separated for a reason. By adding geo-spatial configuration, we are able to feed spatial correlations to the model. However, the current method only counts the same geometries, for example businesses in the same building, as a nearby location and takes a very long time to run so results have been saved to csv files.

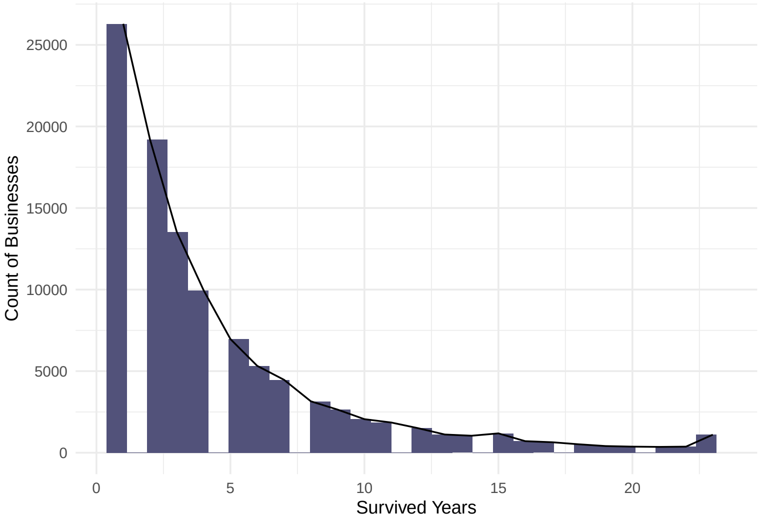


Figure 1: Number of Years Businesses have Survived

It seems like businesses are more likely to close down when they are new compared to businesses that survived a long time. Therefore, we decided to add in a new history variable, which is an indicator of whether a business has survived for more than 5 years or not. One last feature engineering we did using the licence dataset was to uncover the chain business information for an issued licence. To do so, we counted the number of times the same name has occurred in the dataset. There are limitations in the current method used for this feature simply because it is time consuming to go through all chain businesses and make sure their names are spelled the same. However, this feature does capture the cases when a business had opened up at a different locations many years ago.

Possible future improvements

There are several things that can be improved for the data science techniques. The first one is the way to deal with missing values. The current method uses the median of existing values to fill in. Since there are a lot of businesses with missing data, using the same value reduces the predictive power of that predictor. An improvement could be using year as a predictor to fit a regression model to predict the missing value. This method makes sense because the data is chronological. However, this requires fitting and tuning another model so we didn't implement it due to the time limit.

The second improvement is how to convert characters like business type into input acceptable by machine learning models. The method implemented is one-hot encoding which outputs the same numerical representation to the same business name. However, we are losing some information using this method. For example, we know that restaurants are more similar to cafes than gasoline stations. That information is lost after one hot encoding. Thus, one improvement is to use natural language processing techniques that will keep such relationships while converting those business names into numerical representation. We didn't implement this because this requires training another machine learning model just for one variable which might not be an important predictor.

The third improvement is to explore more predictive models. There are many factors that will affect business survival like global economic factors, local policies, the district demographic and they are likely to interact with each other in subtle ways that we might not have anticipated. We used a tree-based model LightGBM to deal with interactions but that might not be enough. Neural networks are another powerful predictive tool to model complex relationships. The number of training examples is not that large so the result of a complex neural network might not be very good. Moreover, For the purpose of our project it is important to see the important features which are harder to achieve for neural network models.

Data Product and Results

The final data product consists of a fully-reproducible machine learning model pipeline and a visualization dashboard with the model embedded. The entire pipeline along with usage instructions and commands required to host the Dash dashboard are documented in the project's private GitHub repository. Currently the dashboard is only available to be viewed locally, but it can be easily deployed to Heroku or other cloud servers if public sharing is needed in the future.

Model overview

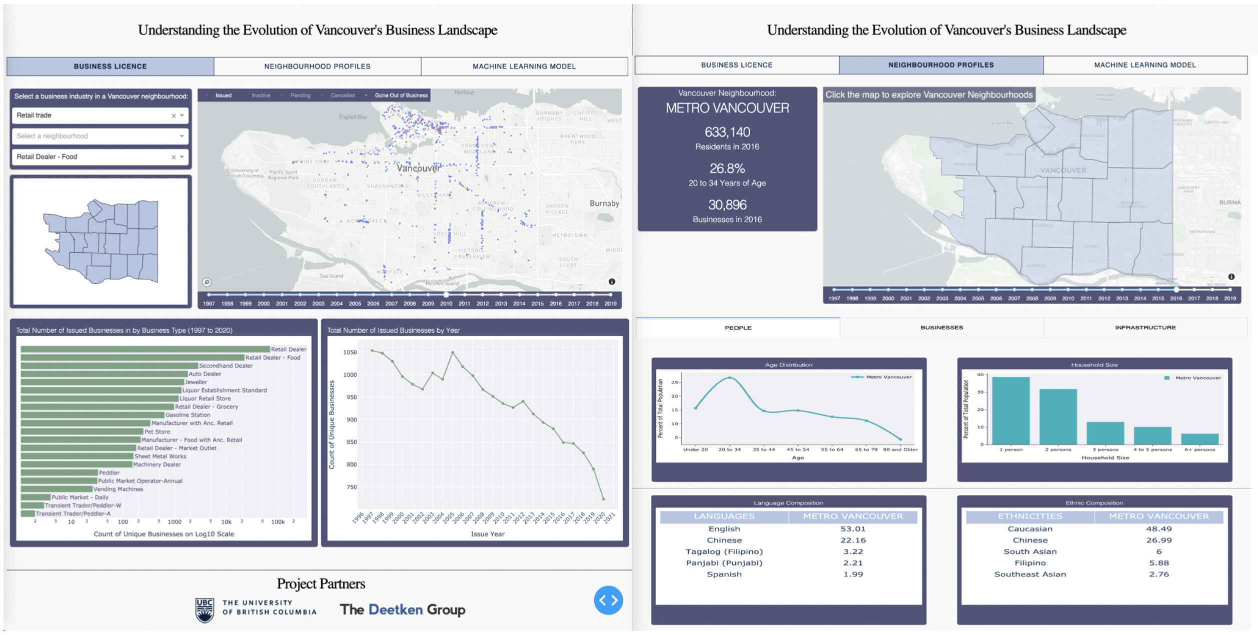
For modelling, we built a data pipeline that could automatically pre-process, train and perform hyperparameter tuning. We used the LightGBM model, a tree-based model to make predictions. Due to the class imbalance problem, it is important to look at the recall rate for the not renewed class as well. The recall rate for not renewed class is for all the businesses that have not renewed their licences, how many of them the algorithm correctly predicts. The accuracy for the current model is and the recall rate of not renewed cases is 0.55. Most of the hyperparameters for the LightGBM model deal with overfitting problems. Since the current model does not have such problems, the only hyperparameter that is different from default is class_weight. The class_weight parameter is set to balanced to deal with the class imbalance problem.

Dashboard overview

An integrated data product is desired to reflect the reality that predictive modelling of landscape evolution is a sophisticated task that cannot be well explained alone by a model. Cities are dynamic living organisms and associated datasets often contain both temporal and spatial dimensions. Entrepreneurial activities and demographics of neighbourhoods that are being fed into our pipeline are actual tangible components of Vancouver that are constantly evolving at different paces. In order to aid the collaboration and communication among different stakeholders in city planning or business decision making, it is intuitive to visualize these factors on top of Vancouver's physical structure.

The first and second tabs of the dashboard focus on providing a holistic view of Vancouver's business landscape and demographics with descriptive charts and map of the city. The third machine learning model tab allows users to select levels of available variables for a specific business. Instead of just computing the renewal probability in the background, modelling performance and outcomes are visualized on a map to incorporate the geospatial dimension. The design of this visualization solution is intended to draw conclusions from the city's past and present data for better forecast into the future. With this in mind, we have also added a slider bar to each of the maps to present the temporal dimension of the data.

Suppose June is a city planner working for the City of Vancouver. She noticed that quite a few auto repair stores failed to renew their license this year in Marpole and Mount Pleasant. In order to devise an intervention



Select a business industry in a Vancouver neighbourhood:

Retail trade

Select a neighbourhood

Retail Dealer - Food

Vancouver Neighbourhood:
METRO VANCOUVER
633,140
Residents in 2016
26.8%
20 to 34 Years of Age
30,896
Businesses in 2016

Language Composition

LANGUAGES	METRO VANCOUVER
English	53.01
Chinese	22.16
Tagalog (Filipino)	3.22
Panjabi (Punjabi)	2.21
Spanish	1.99

Ethnic Composition

ETHNICITIES	METRO VANCOUVER
Caucasian	48.49
Chinese	26.99
South Asian	6
Filipino	5.88
Southeast Asian	2.76

Figure 2: Dashboard Overview

for future zoning designation planning, she wants to be able to explore the business licence dataset alongside some social factors.

When June logs on to the dashboard, on the first tab she can select the auto repair business type in the top left user control panel. The main scatter map plots all issued licences in a given year and distinguishes renewal statuses by different colour. June indeed sees that red dots, or gone out of business ones, concentrate in the two neighbourhoods. She then selects Marpole which triggers zoom in on the map, and filters data for the bar and line charts. When June navigates to the bar chart, she finds out that auto repair is a major business component of Marpole. However, the line chart shows that the increase in the number of issued unique businesses has been slowing down and the map shows that physical distribution of the stores has been shifting towards another major intersection.

June hypothesizes that this might be caused by some underlying changes in demographics and switches to the second tab which provides more detailed neighbourhood profiles. By clicking on the choropleth, she is able to select Mount Pleasant and view a summary statistics of census data. Among the three tabs in the bottom section: people, businesses, and infrastructure, June decides to investigate individual factors under people characteristics. Mount Pleasant has more people at age 45 or below than average Metro Vancouver and more small household sizes. By selecting a different census year, she might find out that the composition has changed significantly in recent years.

Pros and cons of the product

As discussed earlier, modelling of any dynamic component of a city requires aggregating data from socio-economic, environmental, and cultural perspectives. These datasets are collected to serve for various purposes and are thus often multi-dimensional on different scales. During the course of this project, we have encountered multiple types of data, such as chronological data entries, national surveys, and geojson shape files. Therefore, we have designed our pipeline to efficiently unify data from various sources and build configurations around specific time points or geo-locations, while allowing flexibility in adding new attributes of a city.

The fact that the dashboard is closely related to the model pipeline provides users an interface to explore available features, instead of reading off a plain feature importance ranking. Our visualizations are built to present the datasets and showcase the model in a clean and interactive way, allowing users to hypothesize scenarios and draw conclusions that suit their needs. However, since all visualizations are built around variables used in the model and given that our model lacks deterministic predictors, the amount of information users can learn from the data product is limited. For example, our city planner June might also be interested in the effect from the broader economy as well as other indicators such as crime rates and traffic information.

Possible future improvements

One feature that we want to implement is a SQL database. In the future, there will be more datasets added for predictions so the size of the datasets will grow. For example, instead of the census information for each local area, one could use census results at postal code level which will provide more precise demographic information. There are some datasets like the budget information for the business improvement area which we didn't include because it is in pdf format, transit data and data unique to individual businesses like their rent. We couldn't add all data due to time limit but our pipeline makes it convenient to add in data that is in similar format with the existing datasets. Thus, it will be more efficient to query and update data using a database. Moreover, datasets used in this project could also be used for other projects.

Another potential improvement for data products is to include a video demo that could show new users how they can interact with the dashboard. We didn't implement this because the dashboard will continue to evolve and features might change places so that it would be better to include this demo when the dashboard reaches the version to be shared with the public.

Conclusions and Recommendations

The dashboard along with the machine learning models provide a tool for people to explore the business survival in different local areas. It shows the change of distribution of different business types in each district so it can aid the decision on municipal spending. Moreover, the machine learning algorithm can predict the probability of a certain business surviving in an area that can help the business owner by letting them know the risk before they invest their time and money.

The major limitation faced by our model is the poor accuracy and recall scores, which makes it unsuitable for reliably predicting whether a business will renew its licence. Currently, our model lacks deterministic factors specific to individual businesses; however, we have designed our pipeline to efficiently unify data from a variety of sources, allowing flexibility in adding new attributes. As such, our model pipeline and associated geospatial visualization serve as a great basis for understanding the evolution of Vancouver's business landscape which Deetken can utilize and further enhancing through the incorporation of additional datasets.

References

This project involves the work of
de Jonge (2018); Stéfan van der Walt and Varoquaux (2011); McKinney (2010);

de Jonge, Edwin. 2018. *Docopt: Command-Line Interface Specification Language.* http://docopt.org/.

McKinney, Wes. 2010. *Data Structures for Statistical Computing in Python, Proceedings of the 9th Python in Science Conference.* http://conference.scipy.org/proceedings/scipy2010/mckinney.html.

Stéfan van der Walt, S. Chris Colbert, and Gaël Varoquaux. 2011. *The Numpy Array: A Structure for Efficient Numerical Computation, Computing in Science & Engineering.* https://ieeexplore.ieee.org/document/5725236.