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 across Industries and by individuals to tackle everyday problems like; finding the quickest route home, tracking one's food deliveries etc. Decisions and future planning regarding management of one's assets and services be it on individual scale, industrial or by urban planners could also be approached in the same vein.

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 say 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. 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

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 the business type, the location of the business, and the 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 shown on the map]. As the project progresses, we may choose to further refine our model by obtaining census data aggregated at the postal code level.

To answer our research question, the first step to take is data synthesis by combining different datasets. The difficulty of this step is that different datasets were collected to answer different questions so they are not on the same time scale. We need to think about which scale is meaningful to answer our research question. Another part of data synthesis is to wrangle the data so they are suitable for the machine learning model. Due to the size of our datasets, we will use Postgres to build a database.

After the data is ready, we will build a baseline model using logistic regression. We choose logistic regression because it can give us a probability for the prediction result which can be useful for potential clients. One technique we will use after building the baseline model is regularization. This is because some of the covariates might not be significant so we

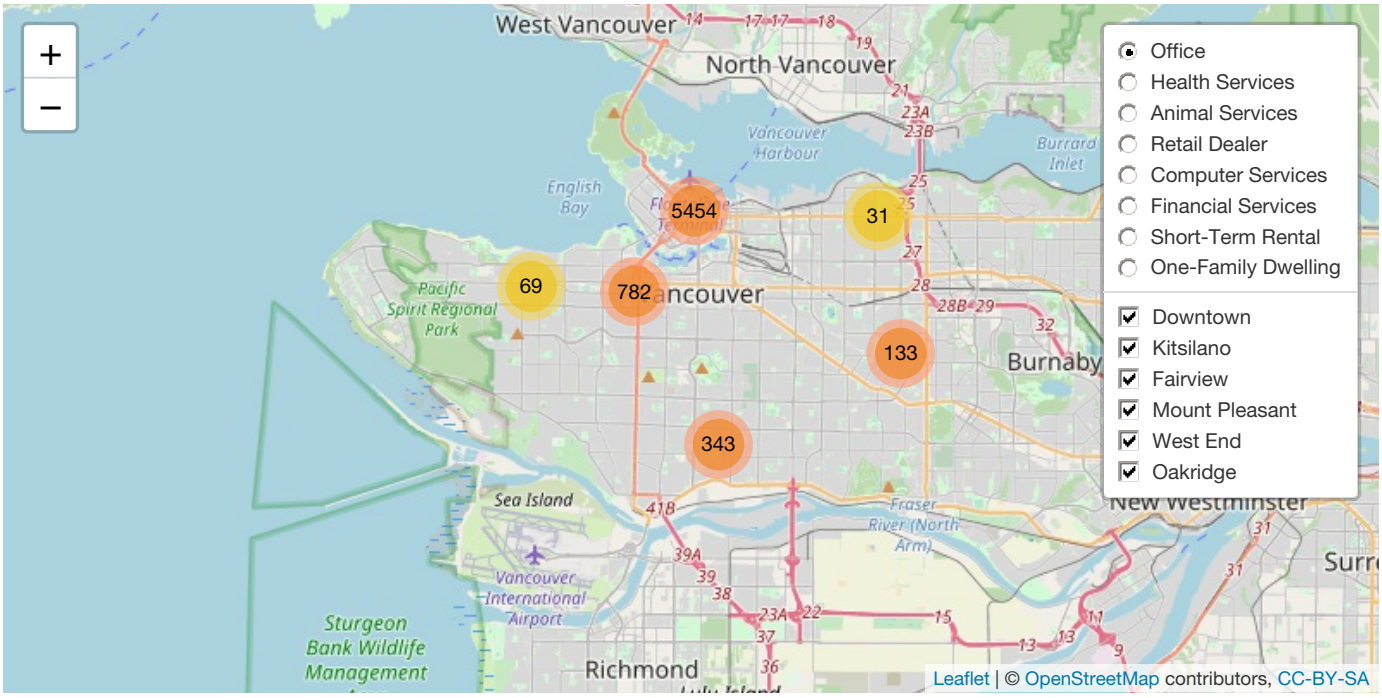
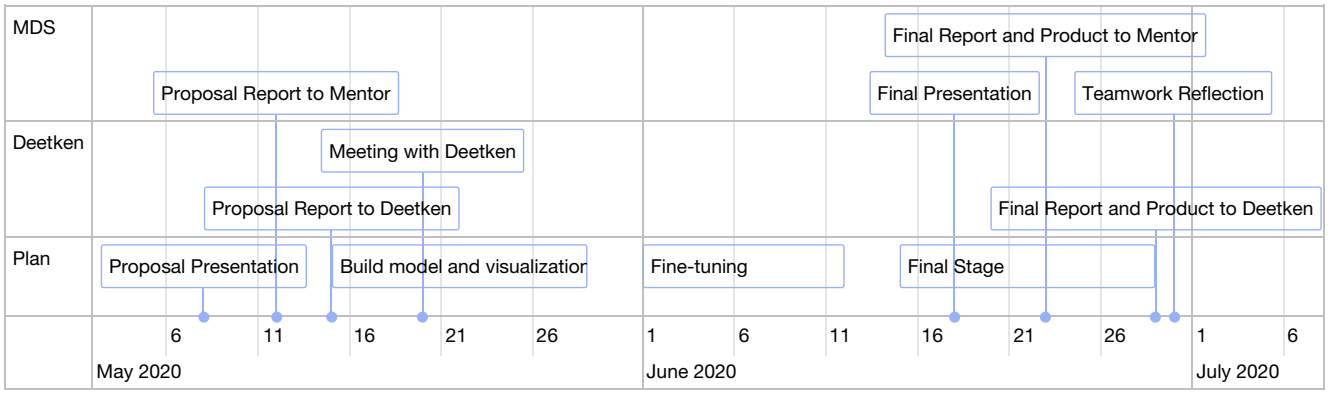


Figure 1: Descriptive and Simulated Information on Map

want to find out significant factors for our final model. Another potential technique we might use is survival analysis. At the same time of building the predictive models, we will also create a visualization to display the 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


MDS
Deetken
Plan
Proposal Report to Mentor
Final Report and Product to Mentor
Final Presentation
Teamwork Reflection
Meeting with Deetken
Proposal Report to Deetken
Final Report and Product to Deetken
Proposal Presentation
Build model and visualization
Fine-tuning
Final Stage
6
11
16
21
26
1
6
11
16
21
26
1
6
May 2020
June 2020
July 2020