Applied Datascience Capstone Project

The Battle of Neighbourhoods

Criteria: This capstone project is worth **70%** of your total grade. The project will be completed over the course of **2 weeks**. Week 1 submissions will be worth **30%** whereas week 2 submissions will be worth **40% of your total grade**.

For this week, you will be required to submit the following:

1. A description of the problem and a discussion of the background. (**15 marks**)
2. A description of the data and how it will be used to solve the problem. (**15 marks)**

Clearly define a problem or an idea of your choice, where you would need to leverage the Foursquare location data to solve or execute. Remember that data science problems always target an audience and are meant to help a group of stakeholders solve a problem, so make sure that you explicitly describe your audience and why they would care about your problem.
*This submission will eventually become your **Introduction/Business Problem** section in your final report. So I recommend that you push the report (having your Introduction/ Business Problem section only for now) to your Github repository and submit a link to it.*

1. Introduction

1.1 Background

We are a company that provides Software as a Service (SaaS) to small and medium sized enterprises (SMEs) in the Greater Toronto Area (GTA). One of our key client groups are restaurant and cafe owners (RCO) in the Food and Beverage (F&B) sector. They come to us for business consultancy services that leverage off our Big Data and Machine Learning capabilities.

We provide basic data on the different types of restaurants and cafes that operate in the neighbourhoods around the GTA, including but not limited to, addresses, contact details, types of cuisine and menus, which are provided to our RCO clients for free.

We provide a paid version of our services that, in addition to the basic data, monitors the different ratings and reviews from Key Opinion Leaders (KOL). We maintain contacts with KOLs by inviting them to promotional events for our RCO clients, like tastings and cocktails, when launching a restaurant or cafe. We also analyse the average spending power of the local residents in the neighbourhood, which are often the first customers to try out a new restaurant or cafe in their locale. We identify surrounding venues of interest where digital and traditional forms of advertising for our RCO clients can be arranged to maximise the impact for their advertising dollars.

1.2 The Problem

Set-Up: Traditionally, RCO clients relied on local knowledge and word of mouth to decide where to open a new restaurant or cafe. Many RCOs do not have the resources to do market research, customer segmentation and wallet analysis. For example, if a bubble tea

cafe opens at a specific locale and does well, many copy-cat bubble tea cafes quickly follow, driving up competition and driving down profits. With the power of crowd-sourced data and combining that with other compiled data, we can apply data analytics to help our RCO clients make better business decisions. If an upscale supermarket opens in a neighbourhood with above average home prices, opening a boutique cafe next door may better capture spill-over customers from the supermarket than opening a Tim Horton's fast-food outlet.

Promotion: Traditional advertising was centered on a local columnist visiting the restaurant or cafe and writing a good review in the local newspaper or magazine. Some restaurants would even pay popular actors/local celebrities to dine at their restaurant and take photographs with them for advertising. With the crowd-sourced platforms, good reviews by KOLs which have built up large followings of "foodies" can matter more than traditional advertising. Passing out fliers at the corner of the block can be replaced by email advertising sent straight to the Inbox of customers or by location-based messaging sent straight to smartphones of passer-byes.

1.3 The Solution

The preparation and serving of good food hasn't changed much, but the set-up and promotion of a good restaurant or cafe has changed greatly. RCOs have an enormous need for data analytics but few have the resources to dedicate to these functions. Our company can fulfil this need and provide RCOs with SaaS that improves their chances of launching a successful F&B business.

2. Data Acquisition and Cleaning

Describe the data that you will be using to solve the problem or execute your idea. Remember that you will need to use the Foursquare location data to solve the problem or execute your idea. You can absolutely use other datasets in combination with the Foursquare location data. So make sure that you provide adequate explanation and discussion, with examples, of the data that you will be using, even if it is only Foursquare location data.
*This submission will eventually become your **Data** section in your final report. So I recommend that you push the report (having your **Data** section) to your Github repository and submit a link to it.*

2.1 Data Sources

Data is sourced from multiple channels. Foursquare provides location based data on different sorts of venues obtained through crowd-sourcing. Foursquare indexes each location through the longitude and latitude of that location, and maintains data on the category of venue, contact details, US ZIP codes, hours of operation, link to websites (if any), menu and pricing (if F&B outlet), tips and reviews from customers and KOLs, and real-time trending data on popular venues.

For Toronto, Foursquare does not map locations to their postal codes. Postal codes for each neighbourhood has to be obtained from the following Wikipedia page, https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M, A separate Google Maps Geocoder API is needed to get the geographical coordinates of each postal code before Foursquare can return the location based data from these geographical coordinates.

Data on average spending power of GTA residents is difficult to obtain as income tax returns are private and confidential in Canada. Our company uses average housing prices in the GTA as a proxy for average spending power as economic studies have shown that higher housing prices have a strong correlation to higher spending power in any given neighbourhood. [**Quote source**]. Historical average housing prices are indexed by postal codes and can be obtained by scraping data from https://housepricehub.com.

2.2 Data Cleaning

Data from Wikipedia shows 103 postal codes starting with M, which is reserved for the GTA, whereas data on historical average housing prices from housepricehub.com shows only 99 postal codes starting with M. Further inspection revealed that certain postal codes have been assigned by Canada Post to high-traffic areas like an Amazon warehouse in Mississauga and a Gateway sorting facility, etc. These postal codes have no residential homes and therefore no average housing prices. Some housing prices have null entries for certain months due to lack of property transactions, so the historical average price needs to be inferred from months with valid entries and this average price is then used to backfill null entries.

Average housing prices then have to normalised across neighbourhoods, disregarding outliers at the very high end. In certain exclusive neighbourhoods, multi-million dollar mansions dominate and skew average housing prices to the high side, but there are no commercial spaces that can be rented by our RCO clients to conduct F&B business due to zoning restrictions. Failure to disregard these outliers could lead to misleading results.

Data scraped from different Data Sources are combined into one dataframe and postal codes with no historical average housing prices are dropped. Certain venues have identical names, which could mean they belong to the same chain store, like McDonald's or Tim Horton's. These were kept in the dataframe and distinguished using their geographical coordinates because different outlets of the same chain may have different tips and reviews from their customers, even though their menu and pricing may be the same.

Trending data often produced a null return. This could be because venues in Toronto were not popular enough with users of Foursquare to generate enough real-time reviews and tips to allow for the generation of trending data.

2.3 Feature Selection

After Data Cleaning, there are [99] neighbourhoods with [1,xxx] venues and [2x] features of which [1x] were selected for their relevance to solving our Problem.

The features relevant to Part 1- Set Up are: [Name, Neighbourhood, Postal Code, Geographical Coordinates, Average Housing Prices, Category of Venue, Type of F&B, Pricing of F&B, Opening Hours]. With the specifications of our RCO client in mind, we will use machine learning techniques to narrow down our search results to a handful of locations where we recommend our client to set-up.

We then add other features to solve Part 2 - Promotion. These features are: [Tips and Reviews, Agree and Disagree Counts, User ID and User Name of Reviewers, Venue ID and Contacts of Venue]. We will use machine learning techniques to identify KOLs that we recommend should be invited for promotional activities and other Venues that we can involve in digital marketing to help our client launch a successful F&B business.