Zahid Akbar

The Battle of Neighborhoods

Capstone Project

Contents

[Introduction 2](#_Toc31289604)

[Background 2](#_Toc31289605)

[Scope 2](#_Toc31289606)

[Data 2](#_Toc31289607)

[Sources 2](#_Toc31289608)

[Data cleaning 4](#_Toc31289609)

[Exploratory Data Analysis 5](#_Toc31289610)

[Modelling 7](#_Toc31289611)

[Feature selection 7](#_Toc31289612)

[Data wrangling 7](#_Toc31289613)

[Model evaluation 7](#_Toc31289614)

[Descriptive analysis 8](#_Toc31289615)

[Conclusions 10](#_Toc31289616)

[Discussions 10](#_Toc31289617)

# Introduction

## Background

The success of a new business depends on a number of factors, one of them is to perform market research. Market research allows a prospective business owner to identify gaps in supply as well as determine if an area has demand for their goods. Doing this research will ensure a new business is set up for success and their hard work does not go to waste, however it is currently very time consuming.

This project aims to identify demand for different types of cuisines based on availability and population demand. The findings will allow new restaurant owners to decide where to open their venue as well as what type of food they will serve.

## Scope

For the purpose of this project, and scope of this project will only target Canada and be limited to food shops.

# Data

## Sources

This project will utilize data from several sources:

1. Geospatial data will be gathered from <http://www.geonames.org/>
2. Census data from Statistics Canada <https://www.statcan.gc.ca/>
3. Current business/venue data from <https://foursquare.com/>

The geospatial and census data will be combined to form a larger dataset, the data will be joined by post code. Finally, the FourSquare API will be used to identify restaurants within a 30KM radius. The responses from FourSquare will be cached due to rate limitations on the API.

Preview of the geospatial data, there is a total of 1656 records in the entire dataset:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| country code | postal code | place name | state name | state code | latitude | longitude | accuracy |
| CA | T0A | Eastern Alberta | Alberta | AB | 54.766 | -111.717 | 6 |
| CA | T0B | Wainwright Region | Alberta | AB | 53.0727 | -111.582 | 6 |
| CA | T0C | Central Alberta | Alberta | AB | 52.1431 | -111.694 | 5 |
| CA | T0E | Western Alberta | Alberta | AB | 53.6758 | -115.095 | 5 |
| CA | T0G | North Central Alberta | Alberta | AB | 55.6993 | -114.453 | 6 |

*Note: 4 columns were removed as they consist of nan values*

Preview of census data, there is a total of 1650 records in the full dataset:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| postal code | Geographic name | Province or territory | Incompletely enumerated Indian reserves and Indian settlements, 2016 | Population, 2016 | Total private dwellings, 2016 | Private dwellings occupied by usual residents, 2016 |
| 1 | Canada | NaN | T | 35151728 | 15412443 | 14072079 |
| A0A | A0A | Newfoundland and Labrador | NaN | 46587 | 26155 | 19426 |
| A0B | A0B | Newfoundland and Labrador | NaN | 19792 | 13658 | 8792 |
| A0C | A0C | Newfoundland and Labrador | NaN | 12587 | 8010 | 5606 |
| A0E | A0E | Newfoundland and Labrador | NaN | 22294 | 12293 | 9603 |

The data above was combined and cleaned to form a new table which contained the relevant information for analysis and calling the four square APIs:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| postal code | place name | state name | state code | latitude | longitude | Population |
| T0A | Eastern Alberta | Alberta | AB | 54.766 | -111.717 | 59234 |
| T0B | Wainwright Region | Alberta | AB | 53.0727 | -111.582 | 64072 |
| T0C | Central Alberta | Alberta | AB | 52.1431 | -111.694 | 62701 |
| T0E | Western Alberta | Alberta | AB | 53.6758 | -115.095 | 43729 |
| T0G | North Central Alberta | Alberta | AB | 55.6993 | -114.453 | 42905 |

Using the geospatial data, the foursquare API was used to get all food venues within a 30 KM radius. A total of 112, 732 records retrieved. The records are limited as a maximum of 99 venues can be retrieved for each area.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Neighborhood | Population | Venue | Venue Latitude | Venue Longitude | Venue Category |
| Eastern Alberta | 59234 | a&w | 54.76694 | -111.981 | Fast Food Restaurant |
| Eastern Alberta | 59234 | subway lac la biche | 54.76208 | -111.964 | Sandwich Place |
| Eastern Alberta | 59234 | taco bell | 54.77161 | -111.973 | Taco Place |
| Eastern Alberta | 59234 | subway® restaurants | 54.77097 | -111.976 | Sandwich Place |
| Eastern Alberta | 59234 | tarra's pizza | 54.76901 | -111.979 | Pizza Place |

In addition to the data above we will use more granular census data to get details about each state individually, this dataset includes a lot of demographic details such age, language spoken, and family characteristics. It is not possible to display a sample of all the data as there are nearly 800 features, here is an example of the first 5:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| State | Age characteristics 15 years | Age characteristics 16 years | Age characteristics 17 years | Age characteristics 18 years |
| NB | 8740 | 9030 | 9285 | 9455 |
| NL | 5805 | 5840 | 5900 | 6020 |
| NS | 11040 | 11085 | 11420 | 11755 |
| ON | 21235 | 22140 | 21715 | 22995 |
| PE | 1965 | 1895 | 1830 | 1915 |
| QC | 93910 | 95980 | 97700 | 100035 |

## Data cleaning

Several steps were taken to clean the data throughout the process to improve the accuracy of the analysis.

|  |  |
| --- | --- |
| Where | What |
| Census data: place name | Removal of suburb names usually presented in brackets. This is unnecessary as the foursquare API does not use it. |
| Census/Geospatial data | Removal of any rows with missing latitude & longitude |
| Geospatial data | Remove any whitespace from place names |
| Foursquare data | Removal of any areas where foursquare did not have any records. |
| Foursquare data | Remove any areas that are not in Canada (Foursquare does area searches that retrieve venues from the USA in the border states) |

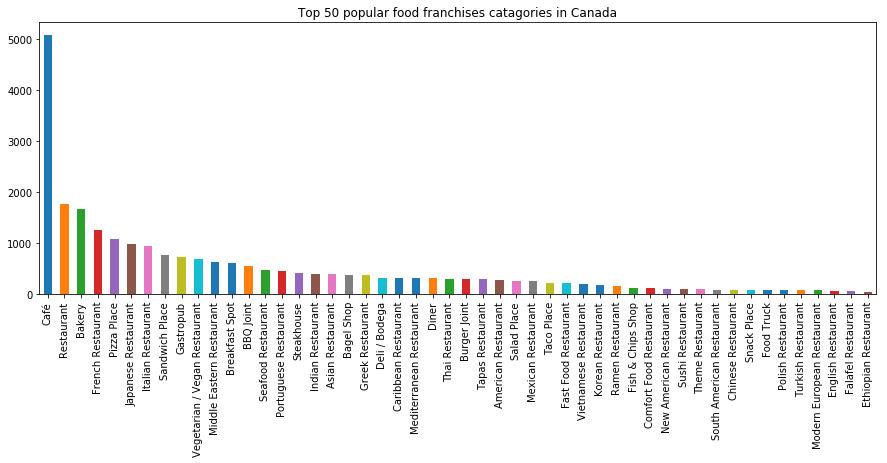
After data clean-up there is a total of 25400 records and 807 attributes.

# Exploratory Data Analysis

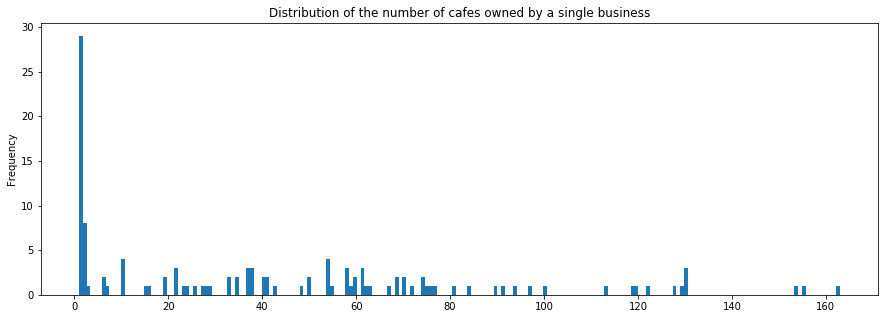
Once the data has been cleaned a range of insights can be gathered. Firstly we need to identify what type of venues are likely to succeed in Canada. Firstly we can observe that Canada has a range of franchises that are popular. According to the data, the Maddy's Pub & Restaurant franchises is the most popular food venue in Canada.



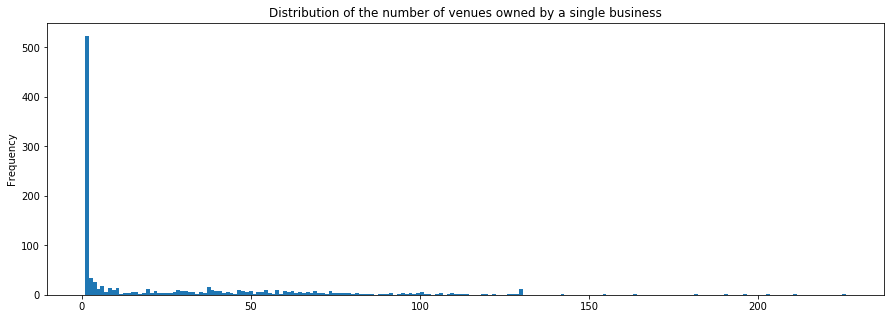
Observing the types of venues in Canada, it can be said that the most popular is the café with a density more than double of the next popular type of venue. This is rather counter intuitive as the top franchises in Canada don’t have many cafés.



Looking closer at cafés it quickly becomes clear that the results observed above is due to the fact that most cafés are small businesses with less than 5 stores:



This helps to confirm that small business can thrive in Canada.



# Modelling

It is important to understand that it is a sample and it is not completely representative of the population. One of the reasons for this is due to the fact the Foursquare API can only return 100 venues for a given place. This means that when we aggregate the dataset, every record where the venue count is 100 may actually mean that there are more than 100 venues; this issue can be solved with machine learning.

## Feature selection

All non-identifiable (i.e. Neighbourhood) data was selected for modelling. Population and Venue counts were scaled using Min max scaling, and the Venue Category was encoded using the label encoder.

Additionally, the data retrieved for areas where 100 records were retrieved were ignored.

## Data wrangling

Machine learning models are sensitive to class imbalance, the dataset used for modelling has a lot of class imbalance. To solve this issue the imblearn package was used to oversample minority classes in the data used for modelling.

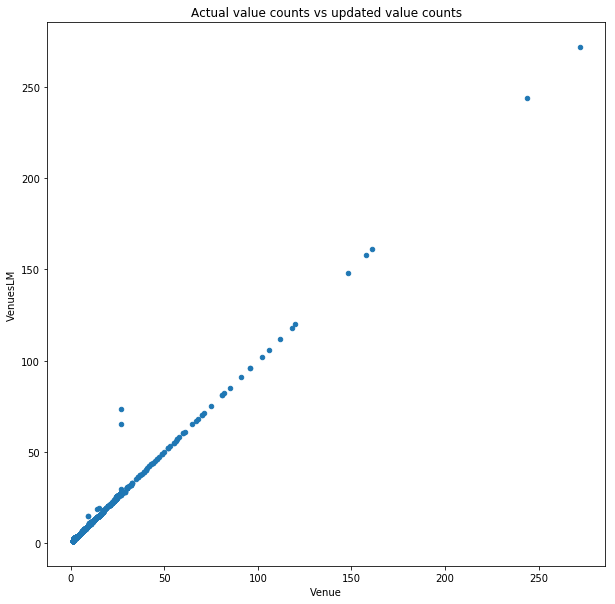
## Model evaluation

A variety of models were tested including:

* Descension tree
* Random Forest
* K Nearest Neighbour
* Multilayer Perceptron (Neural Network)

The Random Forest model preformed the best with an average cross validation score of 0.98

It can be observed that the model only affected the values in a small way, there were only a handful of values where the value the model predicted was more than 1 off the original value.



# Descriptive analysis

Going back to the original purpose of the project, three matrices will be used to determine if and what type of business will succeed in a given area:

1. Area population/venue density vs national statistics
2. Area population change vs national statistics
3. Area population/venue/category density vs national statistics

In the associated notebook, a user can select the area they wish to analyse using a dropdown and run the next cell. It will produce a report, for example, for the area of Downtown Toronto the report is:

|  |
| --- |
| Downtown Toronto has a total population of 15951 and 1700 venues. The standardised population/venue ratio is 0.423 which is more than the national median of 0.026, nationally the population/venue is an average is 22087 and 25450 per Neighbourhood.  Downtown Toronto has had a population change of 0.011% p/a which is less than the national average of 0.01%. Therefore, new businesses in Downtown Toronto has a greater chance of success.  Based on the data, there are 7 types of venues that are likely to do well in Downtown Toronto (see graph below). A Portuguese Restaurant is most likely to do well and a Café is likely to do the worst as it is oversaturated based on the national data.  Top 10 Venue Categories for new businesses in Downtown Toronto:  0 Portuguese Restaurant  1 Tibetan Restaurant  2 Bagel Shop  3 Fast Food Restaurant  4 Noodle House  5 Deli / Bodega  6 Turkish Restaurant  7 Afghan Restaurant  8 Hot Dog Joint  9 Falafel Restaurant  Name: Venue Category, dtype: object  Bottom 10 Venue Categories for new businesses in Downtown Toronto:  72 Vegetarian / Vegan Restaurant  73 Steakhouse  74 Sandwich Place  75 Diner  76 Bakery  77 Restaurant  78 Japanese Restaurant  79 Pizza Place  80 Italian Restaurant  81 Café  Name: Venue Category, dtype: object  C:\Users\theza\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\9F049988.tmp  C:\Users\theza\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\6F0FA0D6.tmp |

# Conclusions

In this study, census, geospatial, and Foursquare data was combined to determine if a new food business would survive in a given area and what type of food they should serve to increase the chance of success.

A range of data science methods were used to clean, modify, and combine data and then model it using various types of machine learning algorithms.

## Discussions

This project was completed as an assignment and as such it should not be used to make any real-world decisions. The data used in fairly old (census data from 2011) and the geospatial data isn’t the most accurate (crowdsourced).