PREDICTING PROFITABILITY OF RESTAURANT

Afaf Kayani

1. Introduction

1.1 Background

Business investors examine the conditions and situations carefully before deciding the place where they want to put their resources in. The objective of business is to maximise profits through serving customers. Investors want to ensure they will have stable customer base for their choice of business. This is done with thorough research and analysis.

In order to open a restaurant, the factors contributing to profits are examined to decide the worth of investment, if any. One of the important factors is location of restaurant. The location demonstrates how well a restaurant will run. A populous location will attract more audience. The culture of social gathering is more favourable with more people in neighbourhood.

Neighbourhoods have different features that are employed in the decision process. Younger people are more likely to eat out. High income individual is more likely to spend money on eating out. Families go to restaurants more than individuals.

1.2 Problem

Running a business is challenging task as the efficiency of business is determined by profit percentage according to business investors' point of view. To ensure effective restaurants are opened are run thoroughly, analysis is done to for maximum profit.

Different neighbourhoods are examined to denote their favourability in opening a restaurant there. This leads to calculated prediction for business investors to make their decisions effectively.

1.3 Interest

The main stakeholders in opening the restaurants are investor and consumers. Investors analyse consumers to decide if they should proceed with opening the restaurant. The money that will be allocated for a project is also under consideration for investors. Budget is also decided by consumers as low-income households will be reluctant to spend a lot in restaurants.

2. Data Acquisition and Wrangling

2.1 Data Acquisition

The neighbourhood for Toronto was acquired from Wikipedia. It provided number of categories and sub-categories of features including overall population and population of several age groups, education qualifications, types of employment and ethnic groups. There was a feature set of 2384 features and 140 neighbourhoods that were examined to make a calculated decision.

2.2 Feature Selection

Population determines customer base for restaurant. Generally, more the population, more people will visit restaurants. However, if there belong to low-income households, they are unlikely to visit restaurants or only visit cheap restaurants. This means population alone is not enough to determine effectiveness of profit. Income was also taken as feature to determine whether restaurants should be open. So out of 2384 features, population and income was observed for 140 neighbourhoods.

2.3 Data Extraction

First neighbourhood data was filtered out and stored in a list. A new data frame with neighbourhoods and income and population headers was assigned. Then population and average income was filtered out from Toronto data frame and assigned to the new data frame.

2.4 Data Normalization

The values of population were quite large and difficult to comprehend, they were scaled down through normalisation. Similarly, income values were difficult to understand, hence they were normalised as well. Min-max normalisation was used. The result was values on scale from 0 to 1.

3. Methodology

3.1 Data Refinement

After neighbourhood's data is filtered out along with their respective income and population, it is normalised through Min-Max normalisation as shown by below equation:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

This normalises the values from 0 to 1 and increases our understanding.

Min-max linearly transformed the input features. The min and max are the minimum and maximum values in X, where X is the set of observed values of input feature. It can be easily seen that when x=min, then y=0, and. When x=max, then y=1

3.2 Model Selection

Model selection is the task of selecting a statistical model from a set of candidate models, given data. As target variable is whether there should be or should not be investment and whether it should it be low, moderate or high, hence this is classification problem as opposed to regression. Regression problem relates to target variable being a numerical quantity. Classification problem deals with classes of target variable, as it is in this case. As there are two features under observation, clustering is more appropriate technique to examine both variables.

3.3 Clustering

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is an example of classification technique. As the input data was numerical instead of categorical, clustering is better for this type of data than other types of classification. Decision trees are used if input features are of numeric type rather than categorical.

The scattering of data towards bottom left and top right and top left of graph shows that there should be three clusters. Income-population graph shows more neighbourhoods lie towards low population and low-income margin. Few neighbourhoods lie on low population and high-income margin and few lie on high population and low-income margin.

This gives rise to 3 clusters:

Cluster No	Population Bracket	Income Bracket
1	High	Low
2	Low	High
3	Low	Low

3.4 Model Improvement

Examining the graph closely shows that low population and low-income can be further divided into 2 clusters for better classification. This results in 4 clusters rather than 3 clusters that was decided earlier.

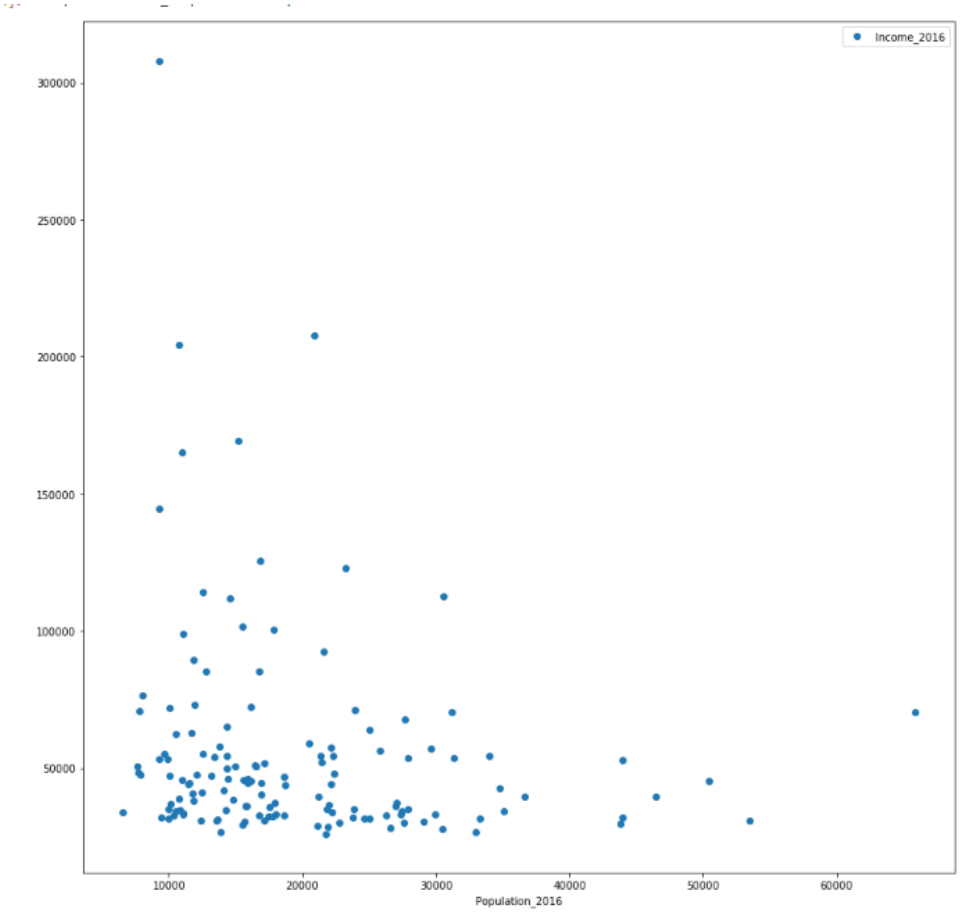
Cluster No	Population Bracket	Income Bracket
1	High	Low
2	Low	High
3	Low	Low
4	Moderate	Moderate

The centre points of clusters were adjusted accordingly for better classification as discussed in Results Section ahead.

4. Results

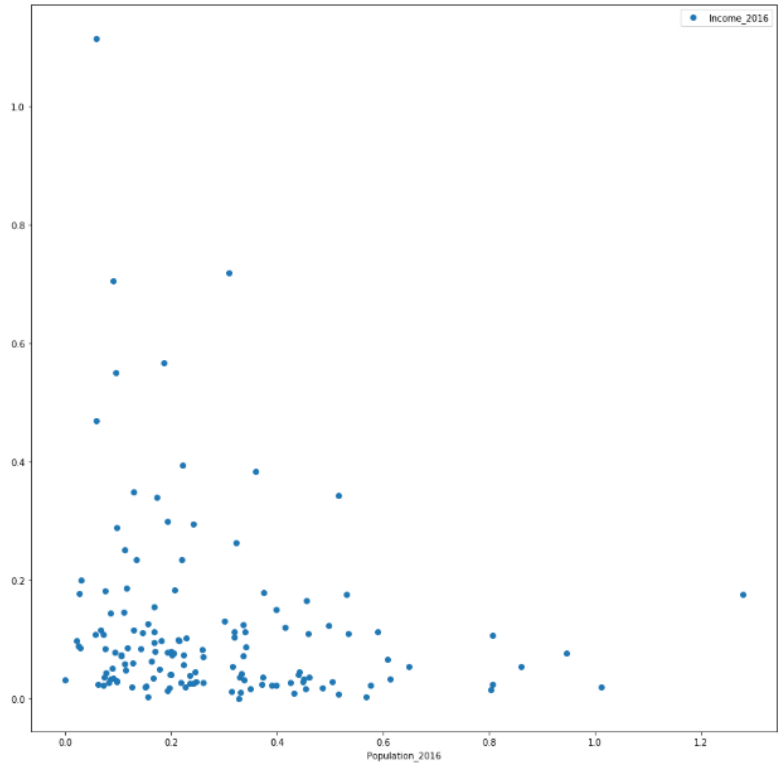
4.1 Raw Data Visualisation

The scatter graph for income-population was observed as follows:



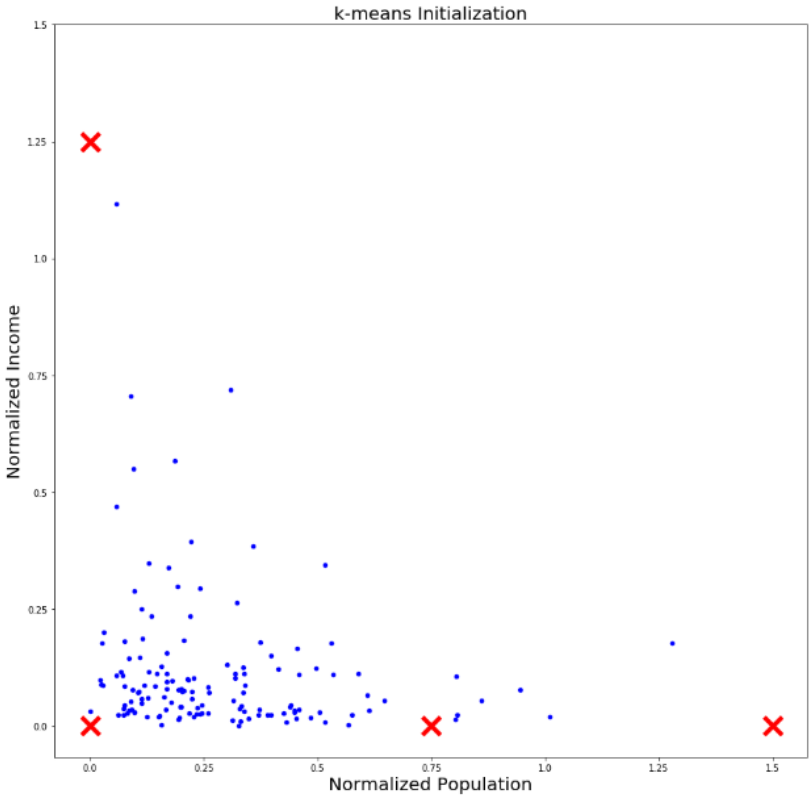
4.2 Normalised Data Visualisation

The scatter graph for income-population after normalisation was observed as follows:



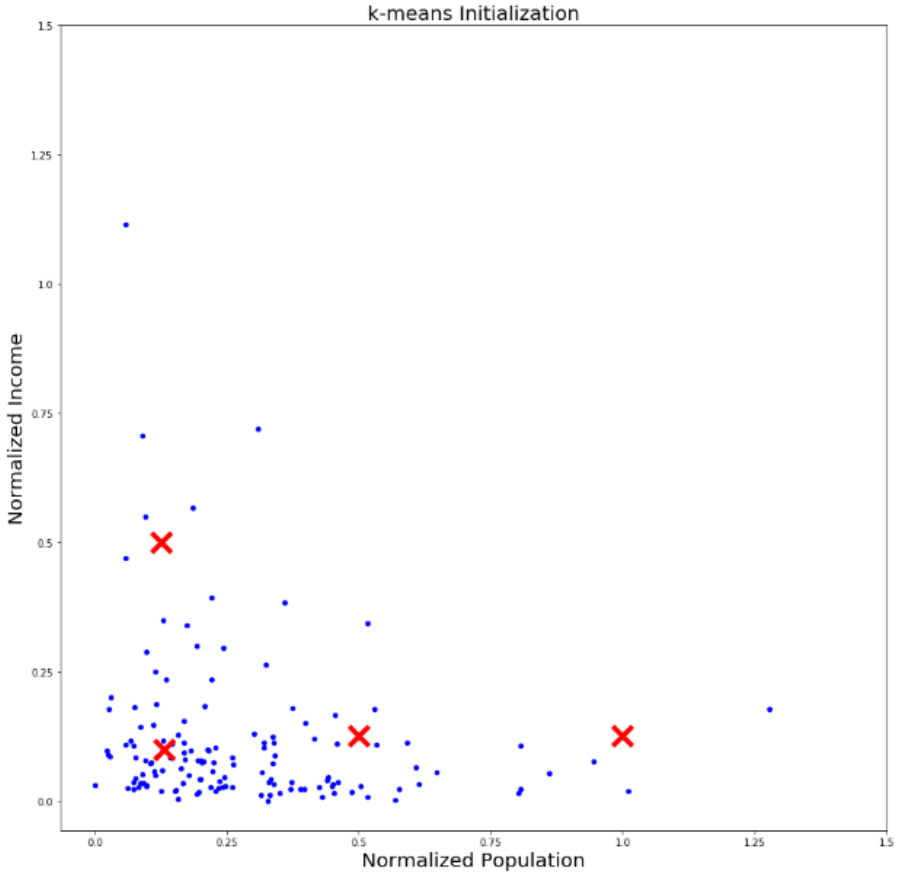
4.3 Initial Cluster Formation

Four clusters were observed as follow:



4.4 Final Cluster Formation

The centroids of clusters were observed as follows:



5. Discussion

5.1 Observations

Income-population graph shows more neighbourhoods lie towards low population and low-income margin. Few neighbourhoods lie on low population and high-income margin and few lie on high population and low-income margin. This gives rise to 4 clusters:

Cluster Coordinates	**Population Bracket**	**Income Bracket**
[1,0.125]	High	Low
[0.125,0.5]	Low	High
[0.130, 0.1]	Low	Low
[0.5, 0.125]	Moderate	Moderate

6. Conclusion

6.1 Conclusion

Neighbourhoods that lie close to cluster (0.5,0.125) will have moderate profit, while those on extremes are not likely to be much profitable. Low income, high population will have less customer or customers that will spend very less. High income, low population will have few customers eating in restaurants. Generally, Toronto does not have high population and high-income neighbourhoods so high investment is too risky and not favoured.

6.2 Improvements

Other factors can also be examined such as population so specific age brackets as younger people are more likely to eat out that older people. The population of families can also be examined as families are more likely to eat out than individual adults.