Osmow's new establishment for Toronto

Coursera Capstone Final Project

Compiled by Mohsin U. Syed

Version 1 Dated: 21 September 2019

Contents

1. Introduction

Osmow's, a modern authentic Mediterranean restaurant, received the 2018 Platinum Readers' Choice Awards. On June 13, 2019, Raptors won their 1st National Basketball Association (NBA) championship title. Osmow's hired Norman Powell and Fred Vanvleet, players from Raptors team, to create an advertisement for their fast food chain. Since then Osmow's restaurant has been growing exponentially. Currently there are 70+ locations in Ontario out of which, two are located in downtown Toronto, one at Queen & Bathurst (Fashion District) and another at College & Augusta (Kensington Market). Being a Torontonian and seeing the craving for this cuisine, as per the Osmow's official web site, I decided to do my capstone project on recommending a new location for establishing the Osmow's restaurant in one of the following locations: East Toronto, Central Toronto, Downtown Toronto or West Toronto.

2. Business Synopsis

The city of Toronto is known for its multiculturalism. Toronto is the fourth-most-populous North American city (2,826,498) as per the wikipedia.org. In Toronto city the following locations: East Toronto, Central Toronto, Downtown Toronto and West Toronto, have wide varieties of fast food restaurants, cuisines and many venues. As we are aware that currently Osmow's is located at Queen and Bathurst (Fashion District) and at College and Augusta (Kensington Market). One of the greatest challenges is to come up with the best location for a new restaurant for this brand.

In order to meet this business challenge, we will apply analytical approach of data science methodology, such as machine learning technique and also leverage the Foursquare API to provide solutions for business issue, which is also the requirement for the capstone project.

3. Stake holders

If someone is looking to open a restaurant or any investors, who are interested in opening up a Mediterranean cuisine or fast food restaurant, can become stake holders for this project. This article can be useful for Osmow's franchisees, franchisers and their fans because this will allow them to discover more opportunities for this business. Food lovers and healthy food trend followers may find this article to be interesting too. Throughout the year Toronto is explored by many tourists and new immigrants; this project will allow them to have a better insight regarding the different venues that exist at East Toronto, Central Toronto, Downtown Toronto and West Toronto.

4. Data Requirements

In order to eradicate the obstacle of finding the best location for the Osmow's modern Mediterranean cuisine, data requirements are enumerated below:

1. Data related to East Toronto, Central Toronto, Downtown Toronto and West Toronto is required.
2. Data for different types of venues and specially related to fast food restaurants and Mediterranean cuisine for the above locations needs to be collected.

3. In order to utilize the Foursquare location data, the latitude and the longitude coordinates for the above locations and venues are required. The Geocoder Python package will be utilised to collect the latitude and longitude. Recently it was noticed that when a request was made to get the latitude and longitude coordinates of a given postal code, often the return result is NaN value. In order to mitigate this issue, Geocoder Python package will not be used. Instead the following csv file will be used for this purpose: http://cocl.us/Geospatial_data
4. Folium libraries will be used for plotting point of interest on the map for the data visualization purpose.
5. Various Python packages will be required for data collection, data understanding and data preparation.

5. Data Collection

1. Web scraping library, BeautifulSoup4 package will be utilised for collecting and cleaning the data for the above mentioned locations for the city of Toronto from the following link: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M,
2. Toronto's table extracted from the Wikipedia portal mentioned above will be converted into Pandas DataFrame.
3. The following link: http://cocl.us/Geospatial_data will be used to download the csv file that has the geographical coordinates of each postal code.
4. The latitude and longitude columns of Toronto Pandas DataFrame will be populated with geographical coordinates based on each postal code column from the csv files.
5. Foursquare API will be used to collect the different venues for the above mentioned locations.
6. Upon retrieving the JSON flat file successfully using Foursquare API, Data Preparation stage will kick in and finally Data Wrangling will be performed to clean the JSON flat file and structure it into a Pandas DataFrame. This process will allow us to better customise and analyze the data to meet the above mentioned business objective.
7. Upon successfully completing the data preparation phase, the data modeling phase will be implemented. In this phase the data will be moduled to apply the K-Means clustering machine learning technique to segment the data into different clusters.
8. Based on the above result, the information will be plotted into the map and conclusion will be derived based on machine learning for the different clusters and finally the best location for the Osmow's modern Mediterranean cuisine will be predicted.

6. Data Understanding and Data Wrangling

Our dataset **toronto_borough** consists of 38 rows and 5 columns. Figure 6-1 illustrates the breakdown of this dataframe:

Borough	Postcode	Neighbourhood	Latitude	Longitude
Central Toronto	9	9	9	9
Downtown Toronto	18	18	18	18
East Toronto	5	5	5	5
West Toronto	6	6	6	6

Figure 6-1

Using the FourSquare API along with the help of Toronto neighbourhood's coordinates provided from the toronto_borough dataframe, we received a JSON file with 1711 records related to different types of venues for the city of Toronto within 500 meters of each neighbourhood. These 1711 records were converted, cleaned and collected into new **venues_df** dataset.

For this project we are interested in data related to Mediterranean Restaurant venues or related to Osmow's Mediterranean Restaurants. So once again we have to analysis the venues_df data set for data related to restaurants, which was collected in new **restaurants_df.** This new restaurants_df holds 409 records, which are related to all sort of restaurants venues that are available in the vicinity of Toronto. We had to go over data collection process again to collect data related to Mediterranean restaurants venues or related to Osmow's restaurant from the restaurants_df. Finally we were able to collect **83** records for this project. Breakdown of **mediterranean_df** is illustrated below:

We found 4 records only related to Mediterranean Restaurants from the 410 records that exist in the restaurants_df. So we decided to collect records related to similar Mediterranean restaurants too. Based on this decision we were able to collect 83 records in total and their breakdown are given below in figure 6-2. Also note we have to add two new records related of the Osmow's Mediterranean Restaurant to the Mediterranean Restaurant categories as this is mentioned in the Business requirement section above.

Venue Categories	Record Counts
Restaurant	50
Fast Food Restaurant	11
Middle Eastern Restaurant	9
Mediterranean Restaurant	4
Falafel Restaurant	2
Afghan Restaurant	1
Doner Restaurant	1
Persian Restaurant	1
Total Records for Mediterranean_df	79

Figure 6-2

7. Data Visualization

A picture is worth a thousand words and we can prove this in our data science world easily. With the help of Folium API and **toronto_borough** dataset we were able to plot the 38 neighbourhoods, with a special effect of heat map displaying the neighbourhood area and the neighbourhood in different colours too for the city of Toronto as illustrated in figure 7-1 below:

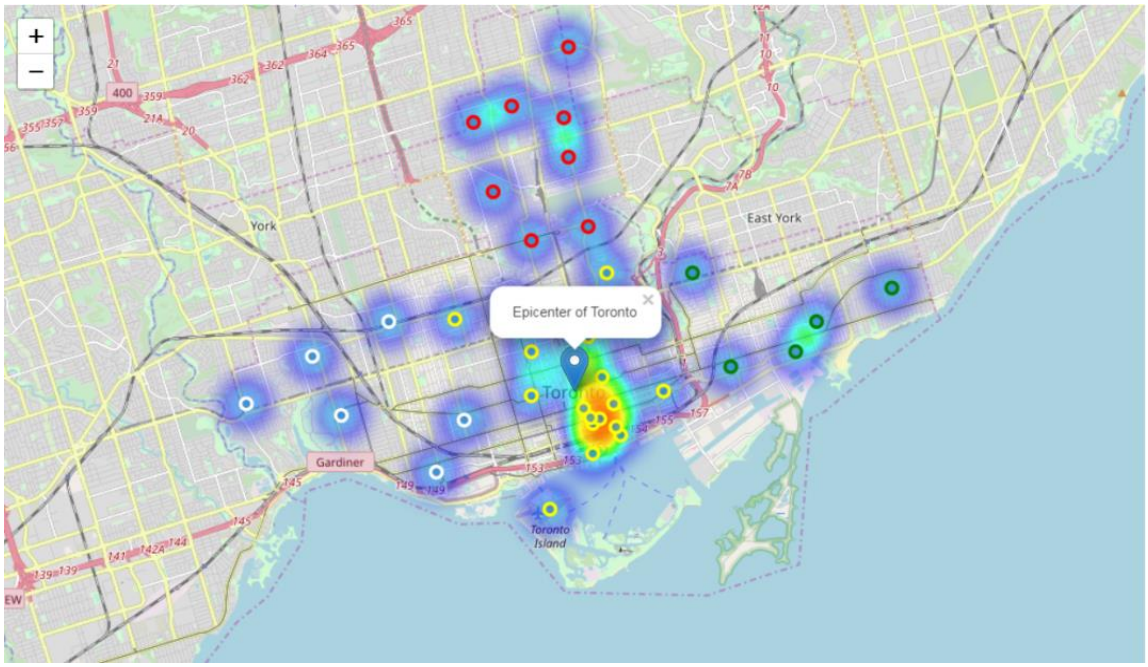

Figure 7-1

In total we found 50 unique restaurants categories for the city of Toronto. Figure 7-2 illustrates all the restaurants categories and there details. Mediterranean restaurants category are coloured in white. This allows us to visualize the current location for these restaurants and will help us in verifying if our model had predicted correctly on where to establish the new Osmow's restaurant, which is the objective of this project.

Figure 7-2

Figure 7-3 illustrates the different grouping of Mediterranean restaurants. Mediterranean restaurant grouping is displayed in white:

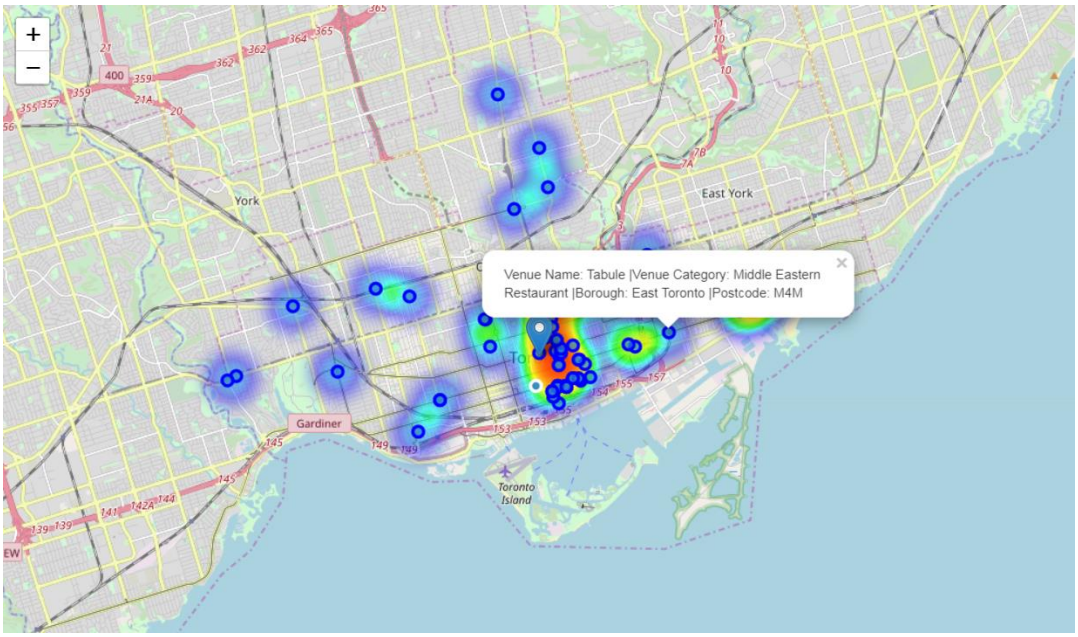

Figure 7-3

Figure 7-4 illustrates the location of two Osmow's restaurants in comparison to the other 7 Mediterranean restaurants category (Restaurant, Fast Food Restaurant, Middle Eastern Restaurant, Falafel Restaurant, Doner Restaurant, Afghan Restaurant and Persian Restaurant).

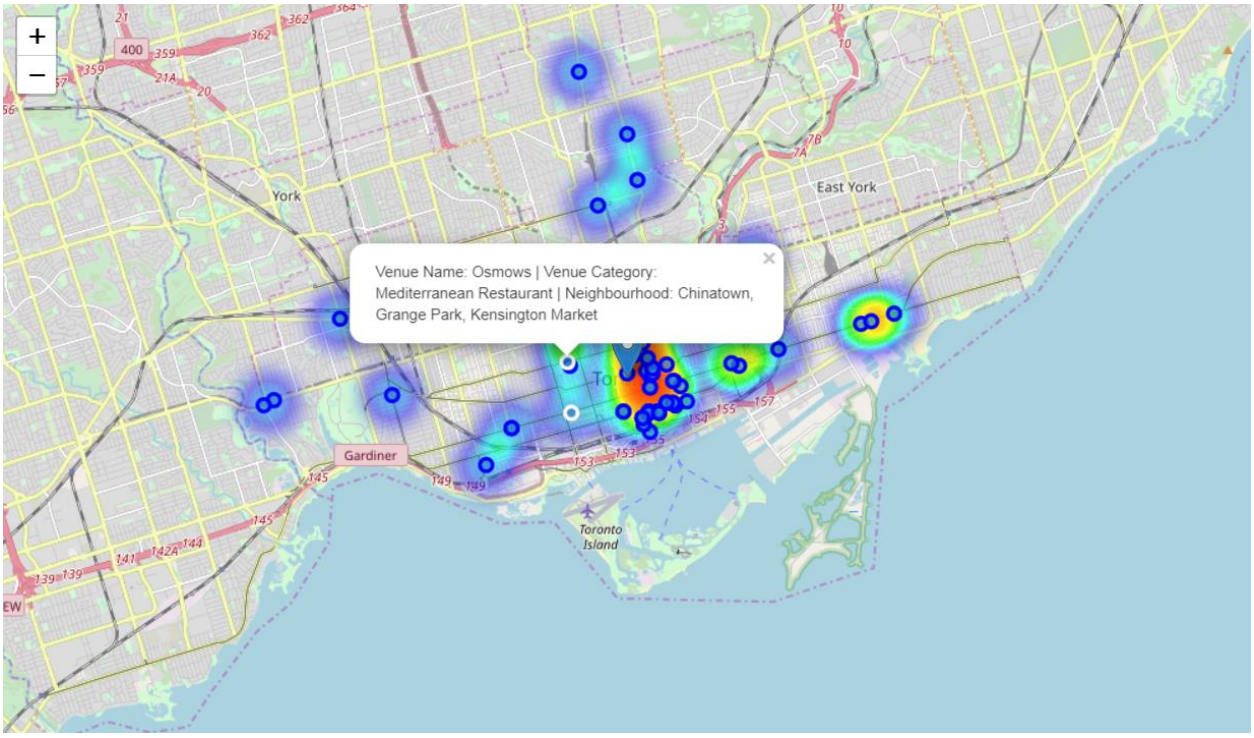

Figure 7-4

a. Analyze Each Borough of the Mediterranean restaurants

A boxplot is a standardized way of displaying the distribution of data based on a five number summary ("minimum", first quartile (Q1), median, third quartile (Q3), and "maximum"). It can tell you about your outliers and what their values are. Figure a-1 shows the distance of restaurants in Mediterranean_df in relation to the different boroughs (Central, Downtown, East and West Toronto).

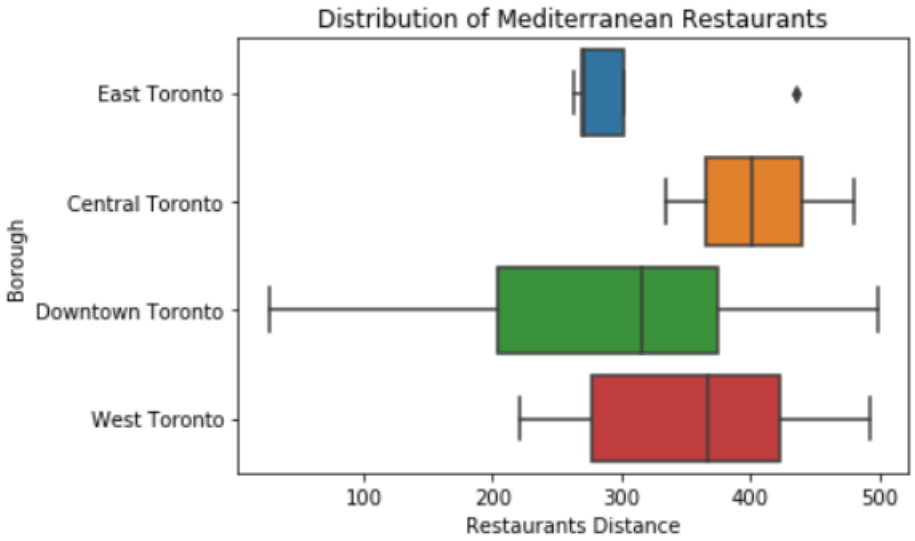

Figure a-1

Figure a-2 is a bar graph displaying distance, number of different types of Mediterranean restaurants located in each of the 4 boroughs.

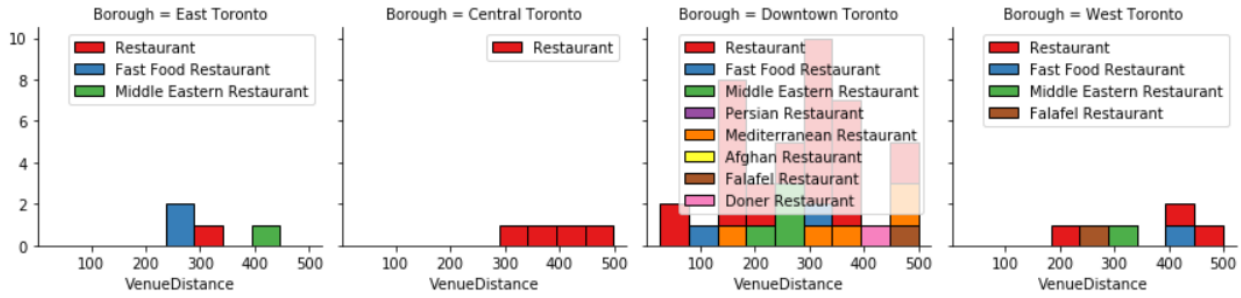

Figure a-2

Currently two Osmow's Mediterranean restaurants are located in Downtown Toronto borough. Please see figure a-3 for more detail. Currently this restaurant can be established in other 3 locations based on the data visualization. But we will wait for the ***KNN supervised machine learning*** result…

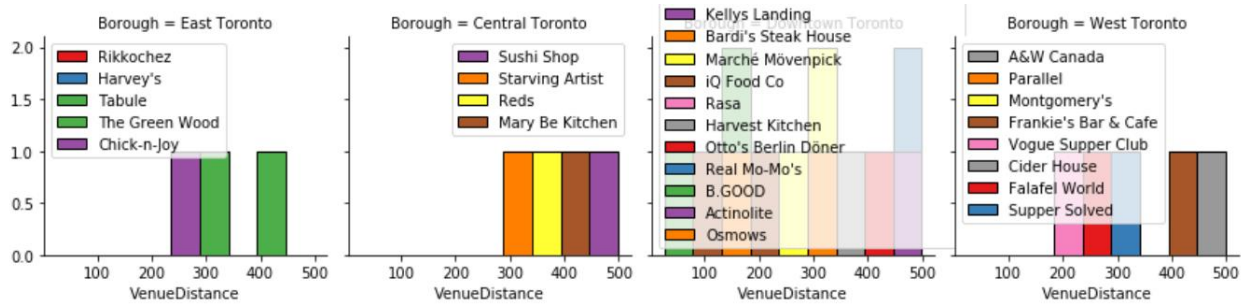

Figure a-3

8. LabelEncoder & OneHotEncoder and DictVectorizer

Machine learning is an important aspect in the data science stream. One important part is feature engineering. It is very common and easy to mess up because there are many categorical features in a dataset. Machine learning or algorithm is all about reading numerical values. Special precaution and attention is required to ensure that encoding is performed correctly. Yan Liu, a data scientist, found a way to automate this process in python with few lines of code. I used his code to perform the data encoding process for this project.

a. mediterranean _df following categorical values were converted into numerical values:
    i. Borough column value before applying encoding:

```
mediterranean_df['Borough'].value_counts()
```

```
Downtown Toronto    66
West Toronto         8
East Toronto         5
Central Toronto      4
Name: Borough, dtype: int64
```

ii. After applying encoding values are: Downtown Toronto is 1, West Toronto is 3, East Toronto is 2 and Central Toronto is 0

```
#LabelEncoder & OneHotEncoder and DictVectorizer
mediterranean_df['Borough'].value_counts()
```

```
1    66
3     8
2     5
0     4
Name: Borough, dtype: int64
```

iii. Before and after values of VenueCategory column:

```
#mediterranean_df.groupby(['Borough'])['VenueCategory'].value_counts(normalize=True)
mediterranean_df['VenueCategory'].value_counts()
```

```
Restaurant                  52
Fast Food Restaurant        11
Middle Eastern Restaurant    9
Mediterranean Restaurant     6
Falafel Restaurant           2
Doner Restaurant             1
Afghan Restaurant            1
Persian Restaurant           1
Name: VenueCategory, dtype: int64
```

```
#LabelEncoder & OneHotEncoder and DictVectorizer
mediterranean_df['VenueCategory'].value_counts()
```

```
7    52
3    11
5     9
4     6
2     2
6     1
1     1
0     1
Name: VenueCategory, dtype: int64
```

- mediterranean_df before and after DictVectorizer…

Before labelEncoder, OneHotEncoder and DictVectorizer

```
mediterranean_df.head()
```

	Postcode	Borough	Neighbourhood	BoroughLatitude	BoroughLongitude	VenueName	VenueLatitude	VenueLongitude	VenueCategory	VenueDistance
0	M4K	East Toronto	The Danforth West, Riverdale	43.679557	-79.352188	Rikkochez	43.677267	-79.353274	Restaurant	269
1	M4L	East Toronto	The Beaches West, India Bazaar	43.668999	-79.315572	Harvey's	43.666663	-79.315081	Fast Food Restaurant	263
2	M4M	East Toronto	Studio District	43.659526	-79.340923	Tabule	43.659731	-79.346341	Middle Eastern Restaurant	436
3	M4R	Central Toronto	North Toronto West	43.715383	-79.405678	Sushi Shop	43.713861	-79.400093	Restaurant	480
4	M4S	Central Toronto	Davisville	43.704324	-79.388790	Starving Artist	43.701538	-79.387240	Restaurant	334

After DictVectorizer run:

DictVectorizer

```
print(mediterranean_df.shape)
mediterranean_df.head()
```

```
(83, 10)
```

	Postcode	Borough	Neighbourhood	BoroughLatitude	BoroughLongitude	VenueName	VenueLatitude	VenueLongitude	VenueCategory	VenueDistance
0	0	2	30	43.679557	-79.352188	42	43.677267	-79.353274	7	269
1	1	2	29	43.668999	-79.315572	20	43.666663	-79.315081	3	263
2	2	2	28	43.659526	-79.340923	48	43.659731	-79.346341	5	436
3	3	0	22	43.715383	-79.405678	46	43.713861	-79.400093	7	480
4	4	0	11	43.704324	-79.388790	44	43.701538	-79.387240	7	334

9. Model Evaluation and Refinement:

K-Nearest Neighbors is an algorithm for supervised machine learning. Data is 'trained' with data points, in our case it was K=5, then for the best K with Ks=13 and the result was:
array([1. , 0.81927711, 0.81927711, 0.81927711, 0.80722892, 0.81927711, 0.79518072, 0.79518072, 0.79518072, 0.79518072, 0.79518072, 0.79518072])

10. Results and Discussion:

Finally ran for the best accuracy and the returned value is 1, which is illustrated in figure 10-1.

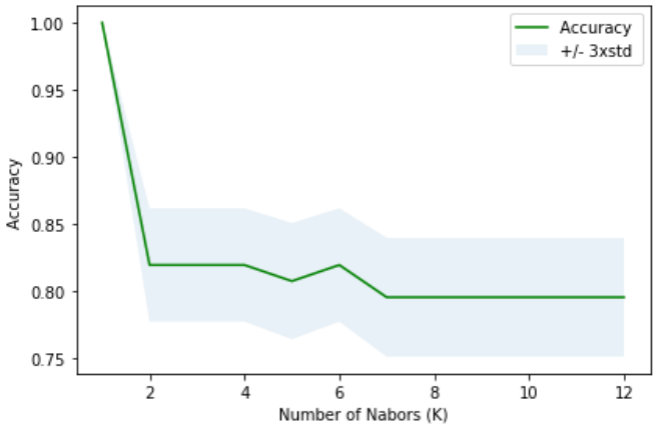

Figure 10-1

In pattern recognition, the k-nearest neighbour algorithm (k-NN) is a non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. If k = 1, then the object is simply assigned to the class of that single nearest neighbor.

Algorithm	Jaccard	F1-score	LogLoss
KNN	0.80	0.70	NA
Decision Tree	1.00	1.00	NA
SVM	1.00	1.00	NA
LogisticRegression	0.82	0.75	0.49

Figure 10-2

Jaccard similarity coefficient is a statistic used in understanding the similarities between sample sets. The measurement emphasizes similarity between finite sample sets,
and is formally defined as the size of the intersection divided by the size of the union of the sample sets. In our case the values are very promising.

F1 score (also F-score or F-measure) is a measure of a test's accuracy. Used for statistical analysis of binary classification, The F1 score is the harmonic mean of the precision and recall, where an F1 score reaches its best value at 1 (perfect precision and recall) and worst at 0. In our case the value is ranging between 0.7 and 1.

Decision tree learning is a method commonly used in data mining. The goal is to create a model that predicts the value of a target variable based on several input variables. In our case we have

four different choices to select from and make a decision. For this purpose I used it and the result is 1 which means the outcome at the end of the leaf is correct.

Support Vector Machine is a linear model for classification and regression problems. It can solve linear and non-linear problems and work well for many practical problems and in our case it worked well.

Logistic regression is a statistical analysis method used to predict a data value based on prior observations of a data set. It then scores new cases on their probability of falling into a particular outcome category.

Log Loss, logarithmic loss (related to cross-entropy) measures the performance of a classification model where the prediction input is a probability value between 0 and 1. The goal of our machine learning models is to minimize this value and in our case it is 0.49, which is fairly good. Log loss increases as the predicted probability diverges from the actual label but in our case the predication is close to 0.82 which is pretty good.

I used the above machine learning algorithms, just to ensure that we get the best result and in our case we did a fairly good job.

11. Conclusion:

Based on the data visualization results we discovered that currently both Osmow's restaurants are located in Downtown Toronto neighbourhood. ***KNN supervised machine learning*** also predicted the same result. The best accuracy and the returned value is 1, which is illustrated in figure 10-1. I will leave it to best interest of my stakeholder to decide on the new establishment, as there is a huge opportunity in the other 3 boroughs, which is East, Central and West Toronto. As per the Osmow's official website they are growing significantly in the East of Toronto.