How Manhattan and Toronto neighborhoods are different?

Artem Ramus

October 2021

1. Introduction

1.1 Background

Neighborhoods in cities can be characterized in different ways. In capstone project of IBM Coursera Data Science Certification, New York Manhattan and Toronto neighborhoods were classified by venues using clustering model k-means.

1.2 Problem

It was found that there are distinguishable similarities and differences between neighborhoods in Manhattan and Toronto based on various venues that are located in the neighborhoods. This project will discuss a question “How Manhattan and Toronto neighborhoods are different?”. We will see if available classification models are able to distinguish between neighborhoods based on relatively small amount of neighborhoods, approximately 40 for each class.

2. Methodology

Data extraction:
The data set built from two parts, "Manhattan venues frequency" and "Toronto venues frequency". The data sets are combined and redundant values marked as zeroes.
Data cleaning:
A type of all the features is float, the target is integer. The data set is checked for duplicates, null values and homogeneous features.
Machine learning:
Preparation of the machine learning work includes train-test split, normalizing the model with standard scaler and training of KNN, logistic regression, decision tree and random fores models.
5-fold cross validation score is calculated for all the models.
Showing top 35 venues influencing the classification of random forest model.

3. Data acquisition and cleaning

3.1 Data sources

The data set built from two parts, "Manhattan venues frequency" and "Toronto venues frequency". These data sets were produced by following steps:

- listing of all the neighborhoods
- adding Geo-location
- retrieving venues related to the location from Forsquare API
- re-ordering the data frames in one-hot-encoding manner
- summarizing mean the venues appearance

The data sets are combined and redundant values marked as zeroes.

3.2 Data cleaning

The data set has above 350 features. A type of all the features was set to float. The target type was set to integer. The data set was checked for duplicates, null values and homogeneous features. The

3.3 Feature selection and feature engineering

All the venues were selected as features after thorough cleaning and verification of being appropriate data types. All the features were divided in train set and test set. Firstly, the train set was trained and normalized with standard scaler and then test set was normalized based on statistical distribution of the train set. This is in order to avoid over-fitting.

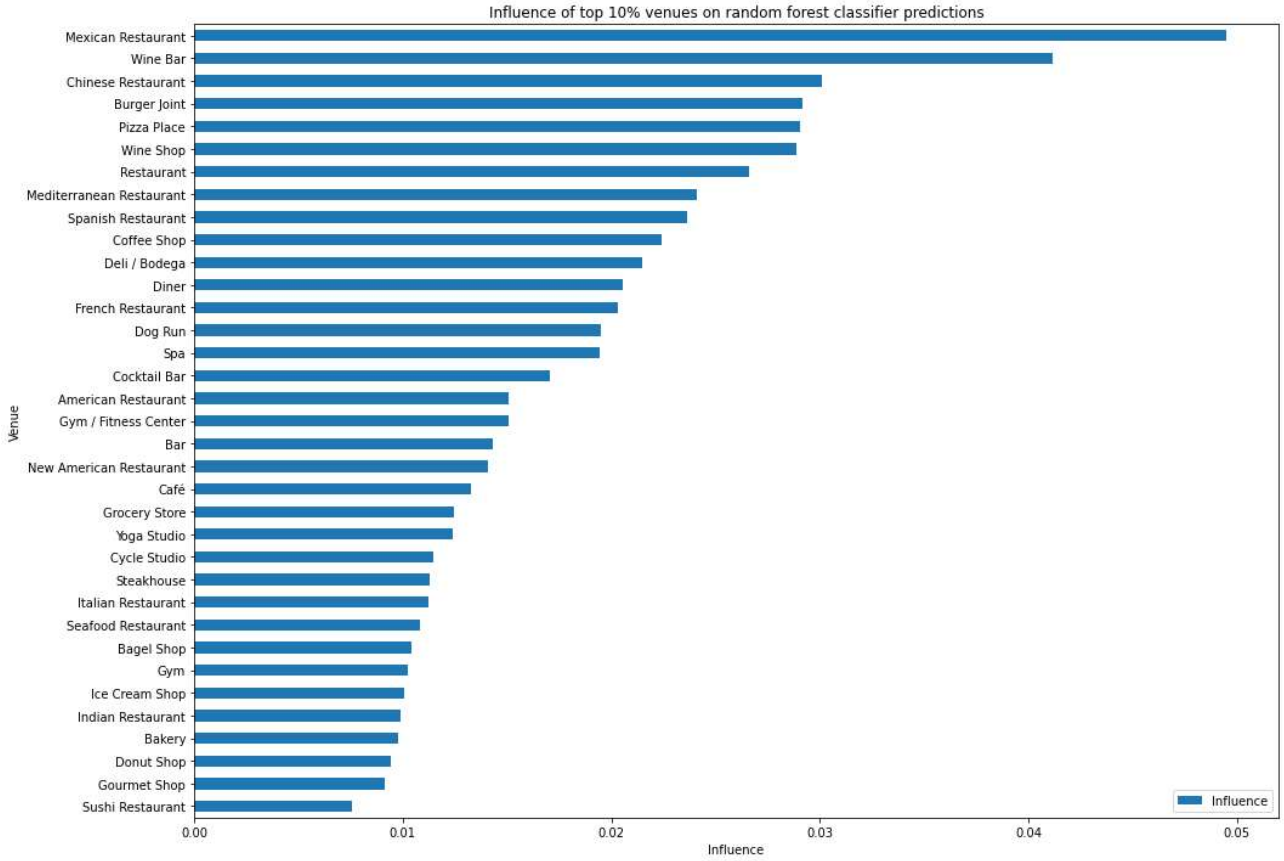
4. Exploratory data analysis

4.1 Definition of target variable

The target variable for the classification task was chosen as “Code”. 0 indicates "Manhattan" and 1 indicates "Toronto". Before the analysis, the target variable is separated from all the features. During the training, the model elaborates the features and learns if this is “0” or “1”. When predicting, the model elaborated the new, test, features and predicts if they are “0” or “1”. The goodness of this prediction then might be checked with classification metrics that are suitable for the problem.

4.2 Influence of venues on the classification

Various features have different degree of influence on the classification prediction. In order to explore this influence, special method of random forest classifier called “feature_importances_” was used. The method estimates a part of overall influence of the feature on classification prediction. It’s output is a list of float numbers in accordance with the features order in the data frame. In a picture below, the influence of top 10% of the features on random forest classifier predictions are listed. There are 35 features presented.



There is a gradual descend from top influencer with 5% strength to about 1% in the and of the list. The most influencing feature between the top 10% influencers is that 31 of them, this is 90%, are related to food. There is restaurants, bars, cafes, coffee shops, and other food-related stores. The rest 10% of the top 10% influences are related to sports, and they are mostly studios.

5. Predictive modeling

5.1 Classification models

Four classification models were chosen for the task: KNN, logistic regression, decision tree and random forest. KNN classifier implements the k-nearest neighbors vote. Logistic regression is a ‘statistical learning’ technique categorized in ‘supervised’ machine learning methods dedicated to ‘Classification’ tasks. Decision tree is a supervised machine learning algorithm that uses a set of rules to make decisions, similarly to how humans make decisions. Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees.

5.2 Solution of the problem

Preparation of the machine learning work includes train-test split, normalizing the model with standard scaler and training of KNN, logistic regression, decision tree and random fores models.

Hyper-parameters selection of KNN classifier was manually performed. 5 neighbors was found the best parameter. The rest of classifiers were employed with default parameters.

In order to be sure that train-test split does not influence the quality of the classification, 5-fold cross validation was employed for all the four models. Jaccard score was calculated for all the models to reflect a match between the prdictions.

5.3 Performance of different models

Performance of the four models was estimated based on the mean value of Jaccard score of 5 cross validation. Here are the results:

K Neighbors:

Mean Jaccart score is 0.73

Cross validation scores: 0.62, 0.88, 0.69, 0.60 and 0.87

Logistic Regression:

Mean Jaccart score is 0.60

Cross validation scores: 0.63, 0.56, 0.63, 0.53 and 0.67

Decision Tree:

Mean Jaccart score is 0.77

Cross validation scores: 0.75, 0.81, 0.69, 0.87 and 0.73

Random Forest:

Mean Jaccart score is 0.87

Cross validation scores: 0.69, 1.0, 0.81, 0.87 and 1.0

The random forest model made the highest score prediction of 87%.

6. Conclusions

The best classifier of those being probed for this problem is the random forest. It was found that 87% of the neighborhoods might be classified as being belong “Manhattan” of “Toronto” areas. This classification is based on venues types and quantities in the neighborhood. The conclusion is that most, 9 of 10, Manhattan neighborhoods are different from those in Toronto.

7. Discussion

It is interesting to mention that most prominent features that differentiate between Manhattan and Toronto neighborhoods are all about the food and sports-related activities. This might indicate that each city has it’s

own food habits and preferences and favorite sports activities. This is where the diversity expressed in a measurable way. It might be induced that there are some venues that equally present in the both cities and they are less helpful for answering the question “How Manhattan and Toronto neighborhoods are different?”.