Capstone Project - The Battle of the Neighborhoods

Ying Li

May 16,2020

1. Introduction

1.1 Background

Choosing the right commercial location in local neighborhood can help businesses owner meet and exceed their sales and marketing goal. Location and Postcode analysis offer business the ability to combine all geographically relevant demographic, economic, and sociopolitical information with financial and other data that businesses generally possess.

Johnson County is a county with highest population and large number of employers in Kansas State. Restaurant owner is interested in exploring the population and business patterns in Johnson County neighborhood to discover the best location to open a pizza restaurant. Setting up pizza shop near an area with a lot of vehicle traffic and stable household income families will take the business to the next level.

1.2 Problem

Data contribute to making decision on pizza place location include venue category, venue address, distance to entrance, cross street, neighborhood postcode, neighborhood employee numbers, and neighborhood employee payrolls.

The aim of this project is to predict whether the given location in Johnson County is the best choice to start a pizza business.

2. Data

2.1 Data Source

Data to consider when choosing a pizza restaurant location involves:

- Demographics information extracted from Census API tool
- Business vendor information extracted from Foursquare API Tool
- Johnson County zip code data from Mongabay Zip Data

The census database I used in this project is 2017 Economic Annual Surveys. The basis of reporting ZIP Code Business Patterns is tabulated at the establishment level. An establishment is a single physical location at which business is conducted. Number of employees and annual payroll for each zip code area are extracted to add to this project model. Thus, the household wealthiness and the number target customers are added to the prediction modal as parameters. Johnson Count Business vendor data including vendor location, vendor cross street, distance to entrance, business categories is added to my analysis. I found the location by zip code level from Mongabay by use of Beautiful Soup to scrape the web.

2.2 Data Wrangling

2.21 Import Data and handle missing values

To limit my analysis in certain geographic areas, I create a latitude & longitude coordinates for centroids of Johnson county, KS neighborhoods. BeautifulSoup is used to capture zip codes of cities in Johnson County from Mongabay Zip Data. After adding the latitude and longitude keys to each city centers with pgeocode.Nominatim package, cleaning data, dropping NA and duplicate data, concatenating city areas with same postal code, the following (with first 10 rows visible) table is created. Now, I am able to use zip code to search for business venues with a radius

of 1500 meters of Johnson County neighborhood.

	city_name	Postal Code	county	state	latitude	longitude
0	Shawnee/ Shawnee Mission/ Sm	66218	Johnson County	Kansas - KS	39.0417	-94.7202
1	Overland/ Overland Park/ Shawnee Mission/ Sm/...	66223	Johnson County	Kansas - KS	38.8619	-94.6610
2	Overland/ Overland Park/ Shawnee Mission/ Sm	66282	Johnson County	Kansas - KS	38.8999	-94.8320
3	Lenexa/ Olathe/ Overland Park	66062	Johnson County	Kansas - KS	38.8733	-94.7752
4	Lenexa/ Overland/ Overland Park/ Shawnee Miss...	66215	Johnson County	Kansas - KS	38.9536	-94.7336
5	Lenexa/ Shawnee/ Shawnee Mission/ Sm	66227	Johnson County	Kansas - KS	38.9536	-94.7336
6	Lenexa/ Op/ Overland Park/ Shawnee Mission/ S...	66251	Johnson County	Kansas - KS	38.8999	-94.8320
7	Merriam/ Overland/ Overland Park/ Prairie Vil...	66204	Johnson County	Kansas - KS	38.9928	-94.6771
8	Edgerton	66021	Johnson County	Kansas - KS	38.7811	-95.0094
9	Spring Hill	66083	Johnson County	Kansas - KS	38.7631	-94.8246
10	Mission/ Overland Park/ Shawnee Mission/ Sm	66201	Johnson County	Kansas - KS	39.0278	-94.6558

- **Foursquare**

Use Foursquare API to get venue information in each city neighborhood. After taking a look at the data pattern for each venue in Olathe neighborhood, the following factors will be considered when choosing the business location:

- Access: One of the most important factors is how accessible your potential location is. Distance to entrance and crossroad information need to be added to the API request.
- Business type: Business type is what I need predict for. This item will be added to my test/train model.
- Popularity of the business: All the venues in Johnson County area have not been rated based on the data pulled from Foursquare. I will skip this part.
- Tax: The restaurant business income rates and sales tax rates imposed on customer are not changed too much within Johnson County. Combined tax rate for Johnson County is 9.48%.

- **Census**

I use census API tool to collect demographics data for the Johnson County neighborhood. The census database I used in this project is 2017 Economic Annual Surveys. Employee numbers and annual payroll ($1,000) are two variables which will be considered. So restaurants owners are able to make judgment on neighborhood size and neighborhood wealthiness of perspective customers.

The dataset with all relevant variables is ready.

	City	Postal Code	Latitude	Longitude	Venue	Venue Latitude	Venue Longitude	Venue Dis to entrance	Venue Category	Venue crosssSreet	Employee No.	Ann payroll
0	Shawnee/ Shawnee Mission/ Sm	66218.0	39.0417	-94.7202	West Flanders Park	39.030392	-94.713196	1396	0.0	0.0	1280.0	64208.0
1	Shawnee/ Shawnee Mission/ Sm	66218.0	39.0417	-94.7202	West Flanders Park & Walking Trail	39.029862	-94.713282	1447	0.0	0.0	1280.0	64208.0
2	Shawnee/ Shawnee Mission/ Sm	66218.0	39.0417	-94.7202	Service Source	39.041227	-94.723396	281	0.0	0.0	1280.0	64208.0
3	Shawnee/ Shawnee Mission/ Sm	66218.0	39.0417	-94.7202	Black Swan Lake	39.041802	-94.730938	928	0.0	1.0	1280.0	64208.0
4	Shawnee/ Shawnee Mission/ Sm	66218.0	39.0417	-94.7202	Bierman's Christmas Tree Farm	39.049750	-94.723537	941	0.0	1.0	1280.0	64208.0
...	...	...	...	...	...	...	...	...	...	...	...	...
1786	Leawood/ Overland/ Overland Park/ Shawnee Mis...	66211.0	38.9667	-94.6169	Ceramic Café	38.957999	-94.629457	1455	0.0	1.0	30534.0	2447778.0
1787	Leawood/ Overland/ Overland Park/ Shawnee Mis...	66211.0	38.9667	-94.6169	We B Nuts And Stuff	38.957999	-94.629457	1455	0.0	1.0	30534.0	2447778.0
1788	Leawood/ Overland/ Overland Park/ Shawnee Mis...	66211.0	38.9667	-94.6169	Hallmark Creations	38.958188	-94.629920	1472	0.0	1.0	30534.0	2447778.0
1789	Leawood/ Overland/ Overland Park/ Shawnee Mis...	66211.0	38.9667	-94.6169	Oz's MAQ Donut House	38.956166	-94.627651	1497	0.0	1.0	30534.0	2447778.0
1790	Leawood/ Overland/ Overland Park/ Shawnee Mis...	66211.0	38.9667	-94.6169	Hallmark	38.957614	-94.629727	1501	0.0	0.0	30534.0	2447778.0

1791 rows × 12 columns

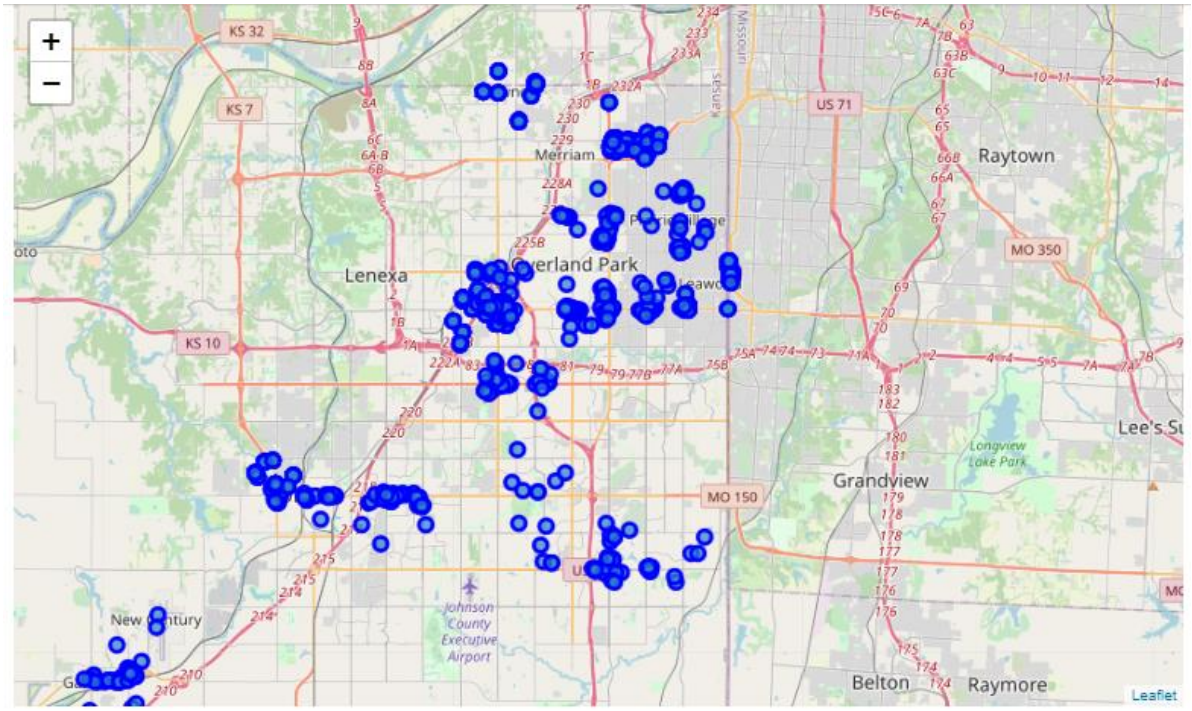
2.22 Data Standarization and Data Normalization

Putting the different variables on the same scale is my next step to analyze the statistics. Standardize data allows me to compare scores between different types of variables. After checking the datatypes of all the variables, I need to identify pizza place from venue category and transform this category item to numerical value. This is predicted value I will use to build models.

Venue cross Street also needs to be transformed to numerical value. All the NA values will be transformed to 0.

Here is the first 5 rows of my standardized dataset and the maps plotting all the Johnson county venues with folium.map package.

	City	Postal Code	Latitude	Longitude	Venue	Venue Latitude	Venue Longitude	Venue Dis to entrance	Venue Category	Venue crosssSreet	Employee No.	Ann payroll
0	Shawnee/ Shawnee Mission/ Sm	66218.0	39.0417	-94.7202	West Flanders Park	39.030392	-94.713196	1396	0.0	0.0	1280.0	64208.0
1	Shawnee/ Shawnee Mission/ Sm	66218.0	39.0417	-94.7202	West Flanders Park & Walking Trail	39.029862	-94.713282	1447	0.0	0.0	1280.0	64208.0
2	Shawnee/ Shawnee Mission/ Sm	66218.0	39.0417	-94.7202	Service Source	39.041227	-94.723396	281	0.0	0.0	1280.0	64208.0
3	Shawnee/ Shawnee Mission/ Sm	66218.0	39.0417	-94.7202	Black Swan Lake	39.041802	-94.730938	928	0.0	1.0	1280.0	64208.0
4	Shawnee/ Shawnee Mission/ Sm	66218.0	39.0417	-94.7202	Bierman's Christmas Tree Farm	39.049750	-94.723537	941	0.0	1.0	1280.0	64208.0
...	...	...	...	...	...	...	...	...	...	...	...	...



3. Methodology

I will focus on detecting the business patterns in the area of Johnson County, and trying to find the correlation of different variables for pizza shops in Johnson County. The analysis is limited to area 1.5 km around city centers.

Firstly, based on the folium library visualize geographic details, I will emphasize on Segmenting and Clustering model (k-means clustering) to check the business density, and business types in Johnson County. The plot map shows the density of existing various venues and especially pizza shops in Johnson County neighborhood.

Secondly, 77 pizza shops out of 1779 venues are located in Johnson County areas. I will use Sklearn and matplotlib to build different models to predict whether the chosen location is the best choice with comparable easy access to entrance and prospective customers with comparable stable household income.

Next, logistic regression, SVM and Decision tree models will be applied to my dataset. After train and test the dataset, Jaccard index and F1 Score and decision tree accuracy score will be used for accuracy evaluation. By choosing the best model and together with the consideration of the clustering model result, I can make the decision on whether to open a pizza business with the given location.

4. Exploratory Data Analysis

4.1 K means cluster

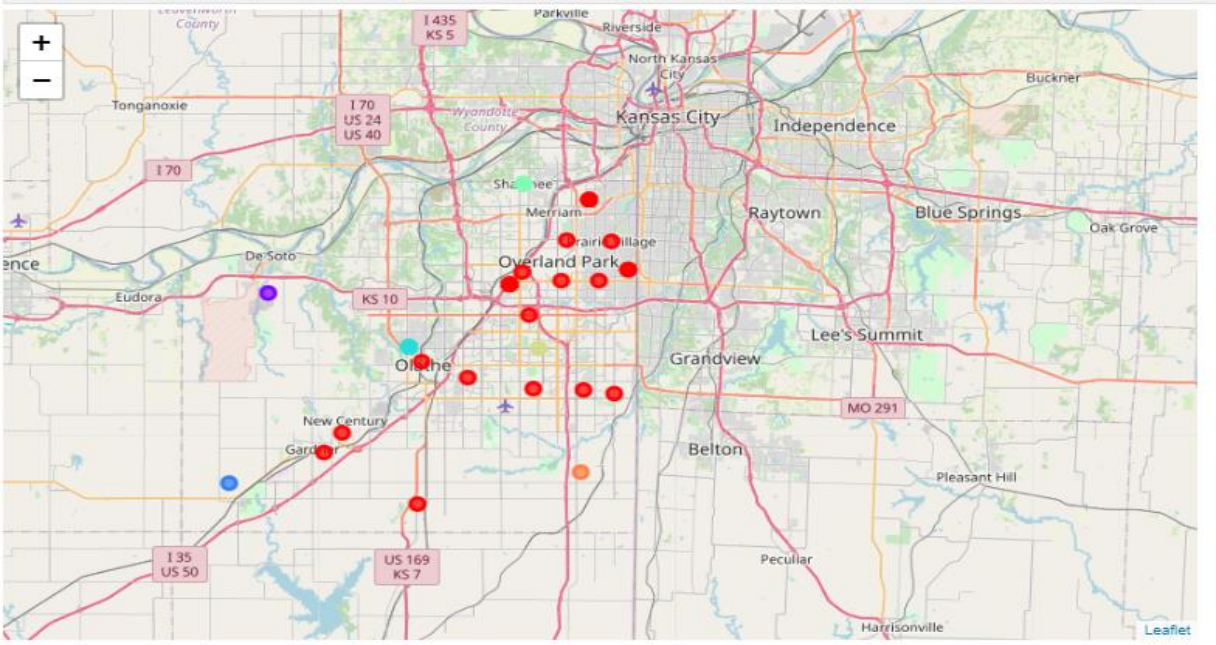
I will do research on the business category within Johnson County and dig in further from the venue dataset by using k means cluster method and create a map showing the pizza shop density in neighborhood.

The primary reason is that clustering creates groups from continuous variables, so if you're looking to create groups, clustering does a really nice job of finding the boundaries between groups for you. Venue segmentation is the process of dividing venues into groups based upon certain boundaries; clustering is one way to generate these boundaries. Lastly, I'm using the k-means clustering technique because it can efficiently handle large datasets and iterates quickly to good solutions.

The Johnson County venues are segmented based upon the business categories. Within the 169 different business categories in Johnson County neighborhood, the business category which appears most often is labeled as 1st most common venue. I choose the first 10 most common venues to establish 7 groups. Each group is clustered upon the frequency of same business

category. I can conclude from below category map and cluster 0 output: most of the recreation and food business are centralized in first cluster which is located Overland park and Olathe areas.

All categories of venues clustering map in Johnson County, KS

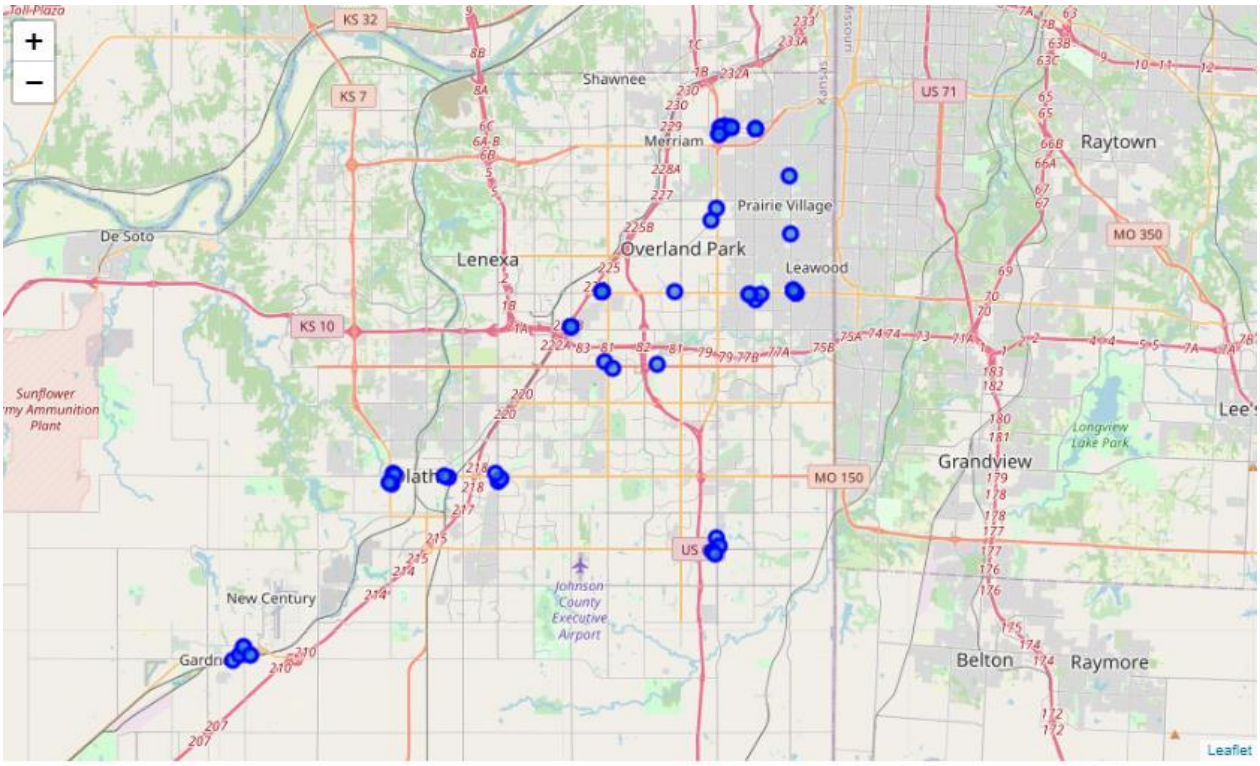


Recreation cluster details:

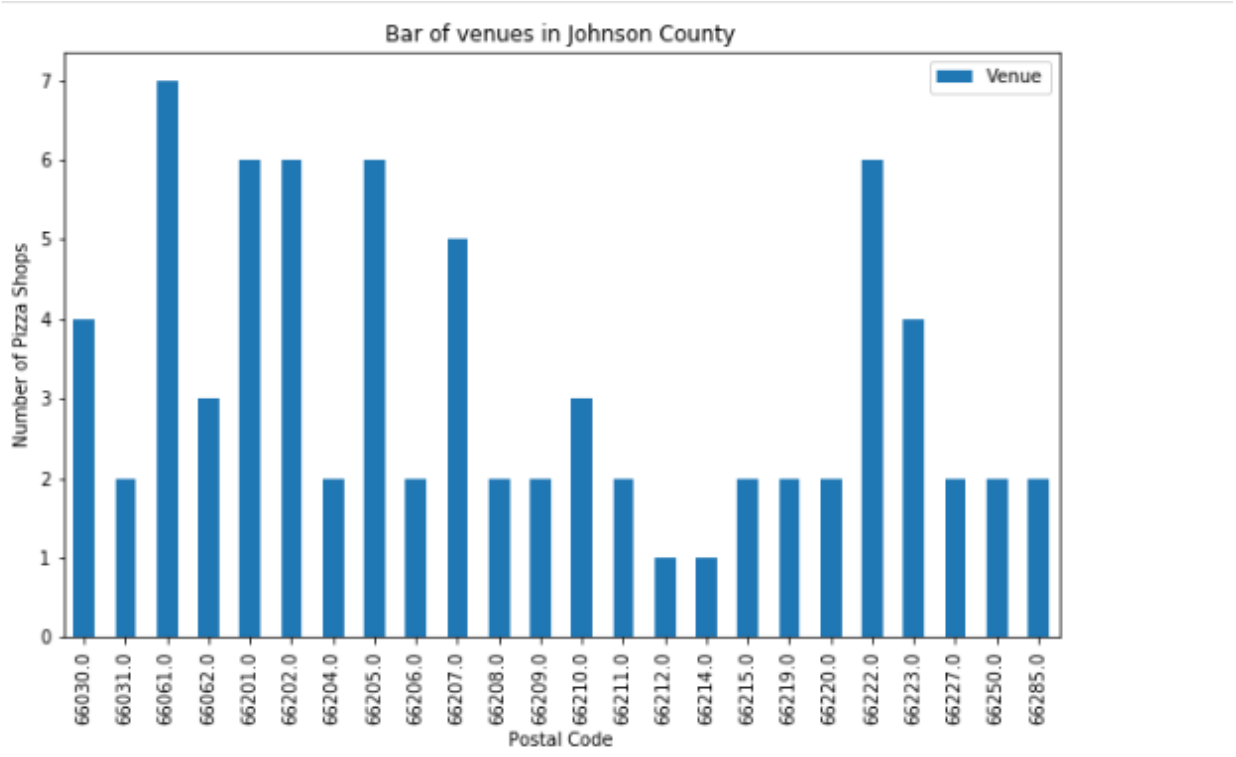
	city_name	Postal Code	county	state	latitude	longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Mos Commor Venu
1	Overland/ Overland Park/ Shawnee Mission/ Sm/...	66223	Johnson County	Kansas - KS	38.8619	-94.6610	0	Pizza Place	Pharmacy	Yoga Studio	Mexican Restaurant	Liquor Store	Grocery Store	Rental Ca Locatio
3	Lenexa/ Olathe/ Overland Park	66062	Johnson County	Kansas - KS	38.8733	-94.7752	0	Fast Food Restaurant	Pizza Place	Cosmetics Shop	Sandwich Place	Coffee Shop	Mexican Restaurant	Discoun Stor
4	Lenexa/ Overland/ Overland Park/ Shawnee Miss...	66215	Johnson County	Kansas - KS	38.9536	-94.7336	0	Clothing Store	Department Store	Cosmetics Shop	Shoe Store	Ice Cream Shop	American Restaurant	Mexicar Restauran
5	Lenexa/ Shawnee/ Shawnee Mission/ Sm	66227	Johnson County	Kansas - KS	38.9536	-94.7336	0	Clothing Store	Department Store	Cosmetics Shop	Shoe Store	Ice Cream Shop	American Restaurant	Mexicar Restauran
7	Merriam/ Overland/ Overland Park/ Prairie Vil...	66204	Johnson County	Kansas - KS	38.9928	-94.6771	0	Mexican Restaurant	Asian Restaurant	Bakery	Sports Bar	American Restaurant	Salon / Barbershop	Pizz Plac
9	Spring Hill	66083	Johnson County	Kansas - KS	38.7631	-94.8246	0	Golf Course	Brewery	Baseball Stadium	Baseball Field	Shop & Service	Fast Food Restaurant	Yog Studi
10	Mission/ Overland Park/ Shawnee Mission/ Sm	66201	Johnson County	Kansas - KS	39.0278	-94.6558	0	Pizza Place	Sandwich Place	Fast Food Restaurant	Mexican Restaurant	Discount Store	Coffee Shop	Clothing Stor

More specifically, pizza places are the first common venue in overland park, Shawnee Mission, Leawood and Olathe cities within radius of 1500m. Thus, I will emphasis on the demographics information for the pizza shops in these four areas.

Johnson County Pizza shop map:



The following bar chart shows the different numbers of pizza shops in Johnson County area.



Leawood/overland park (zip code 66211) and lenexa/Olathe (zip code 66062) neighborhoods are the two communities with high population density and household income, but low pizza shop density (compared with 66030 and 66061 neighborhoods).

4.2 Logistic Regression, Support-vector Machine and Decision Tree Model

4.21 Choosing Models

Next, I will turn to regression and classification models to train and test my data. Regression analysis mathematically describes the relationship between a set of independent variables and a dependent variable. Classification model attempts to draw some conclusion from observed values. Classification model will predict the outcome with given inputs. For these reasons, I will apply regression and classification models to predict whether to open the pizza shop (1) or not (0).

In my case, venue location close to city center, venue crossStreet, venue distance to entrance, neighborhood annual payroll, and neighborhood employee numbers are variables which will be used to predict the binary dependent variable---whether it is appropriate to open the pizza shop in the chosen area.

Then which model is best to choose? SVM works well with unstructured and semi-structured data like text and images while logistic regression works with already identified independent variables; SVM is based on geometrical properties of the data while logistic regression is based on statistical approaches.

I will first try to use logistic regression to see how the model does, and then try using SVM RBF kernel. I will also try decision tree model since this binary classifier share these traits: Easy to understand; able to find odd interactions; and make minimal assumptions.

4.22 Apply models

During this process, I will select some features for modeling. Also I change the target data type to be integer, as it is a requirement by the scikit-learn algorithm.

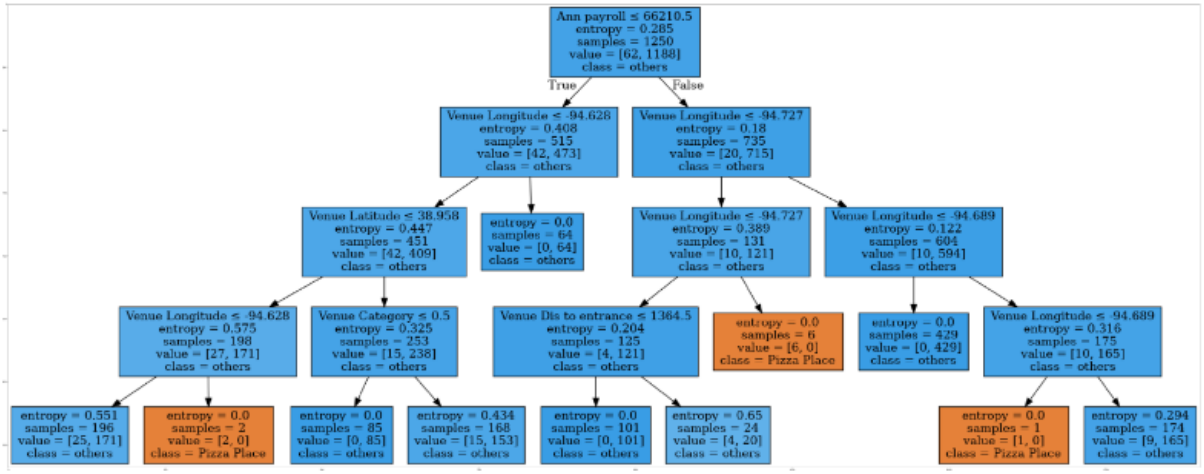
Here is the first 5 rows of the datasets which will be trained and tested:

	Venue Latitude	Venue Longitude	Venue Dis to entrance	Venue Category	Venue crosssSreet	Employee No.	Ann payroll	Postal Code
0	39.030392	-94.713196	1396	0	0.0	1280.0	64208.0	66218.0
1	39.029862	-94.713282	1447	0	0.0	1280.0	64208.0	66218.0
2	39.041227	-94.723396	281	0	0.0	1280.0	64208.0	66218.0
3	39.041802	-94.730938	928	0	1.0	1280.0	64208.0	66218.0
4	39.049750	-94.723537	941	0	1.0	1280.0	64208.0	66218.0

The independent variables which will be chosen to establish my models are venue distance to entrance, venue crossStreet, employee numbers of neighborhood, annual payroll of neighborhood. The dependent variable is whether to open the pizza shop (1) or not (0). I will create arrays for all the variables using numpy package, and then preprocessing the data with standardScaler. Standardization of a dataset is a common requirement for many machine learning estimators: they might behave badly if the individual features do not more or less look like standard normally distributed data.

A test size of 20% is selected to split the train-test set. Using sklearn.model_selection package, I got a train set of 1423 records and test set of 356 records. After importing the logisticRegression, SVM, and DecisionTree Classfier models from Scikit-Learn and training my data by applying models, I get these three models established. Then test set is used to do the evaluation by comparing the predicted y numbers with test y number.

Decision Tree output:



4.23 Evaluate models

Algorithm	Jaccard	F1-score	Accuracy
SVM	0.9593	0.9421	NA
Logistic Regression	0.9579	0.9373	NA
Decision Tree			0.96816

SVM model has a higher Avg F1 score and Jaccard score. If A potential location in Overland park/Olathe/Shawnee mission/Leawood area has been chosen, I can add this address to my model and predict whether the predicted dependent variable equals to my expected outcome---opening a pizza place.

In my model, my pizza shop dataset is quite imbalanced. I have a 2 class (binary) classification of venue types (pizza shop and others) with 1778 instances. Only 77 instances are labeled with Pizza Shops and the remaining instances are labeled with others. Under this circumstance, I would choose the classifier that get high F1 scores and Jaccard. score. From the Decision tree model, I can tell from the train/test datasets: most of the pizza places are clustered in western Johnson County neighborhood with easy access to entrance.

5. Results

The results can be explained in the following areas:

Competition--- Being too close to established competition may help with marketing, but if too close to your competition, the shop owner will have a tough time gaining benefits in the community. The best location for a pizza business would be close to all other categories of business venues while comparably staying away from most pizza shops. The folium map of the business density and especially pizza shops density show Overland Park/Leawood neighborhood having a high density of business venues while low density of pizza shops.

Visibility---Though you want to be careful not to place yourself in the position of losing customers because traffic is too heavy near your restaurant, you also want to ensure that you are visible to drive-by and foot traffic. Actually, the busiest street in Overland park/Olathe/Shawnee mission/leawood area is street line between Kansas and Missouri State, 119th and 135th street. However, the crossroads on these streets are perfect for customer to get access to business. They are not as busy as New York Downtown which cannot even finding a parking lot.

Access---One of the most important factors is how accessible your potential location is. The accessibility of cars makes it as easy as possible for customers to visit the pizza place. Distance to entrance and cross Street accessibility have been added as parameters to establish the regression and classification models.

Demographics---Demographics are the statistical factor of marketing used to identify population segments by characteristics. Businesses must have a targeted approach to customers. Census business surveys provides employee numbers and employee payroll incomes data regarding how many potential customers and local economy at community level. Leawood/Overland Park (zip 66211) and Lenexa/Olathe (zip 66062) are the two top areas with high level in both elements.

6. Discussion

My machine learning models in this project are built based on imbalanced datasets. Out of the total 1778 business venues I selected in Johnson County areas, only 77 of them are pizza shops. Data imbalance reflects an unequal distribution of classes within a dataset. When training on imbalanced data set, this classifier will favor the majority classes and create a biased model. Further analysis and studies will be applied to fix the imbalanced data set. For example, BalancedBaggingClassifier could be used to resampling of each subset of the dataset before training each estimator of the ensemble. That way, I can train a classifier that will handle the imbalance without having to undersample or oversample manually before training.

In addition, the accuracy and precision of the model is restricted to parameters chosen. Some of the factors to consider the location have not been added to my analysis: the size of their parking lots, the location in relation to your equipment and food suppliers, real estate availability, labor costs, safety/crime rates, and other important factors. Business stakeholders have to take account the above elements to make location decision.

7. Conclusion

The purpose of this project is to find a perfect location to start a pizza business in Johnson County, KS local area. By collecting business venue location, venue category, distance to entrance, crossStreet, neighborhood employees, neighborhood wealthiness information from Foursquare & Census business database, I am able to analyze on pizza business competition from pizza shop density map using k-means cluster model. After taking account of the census demographics data, I conclude that the two neighborhood which would be the best location to open the pizza hut are Leawood/Overland Park (zip 66211) and Lenexa/Olathe (zip 66062). Next, the logistic regression model and decision tree model are built to narrow down the location chosen by making binary classification of venue types (pizza shop and others). The business stakeholder now is able to tell whether the location they choose in 66211 and 66062 neighborhoods is perfect to start the new business.

Other elements, like the location in relation to your equipment and food suppliers, real estate availability, labor costs, safety/crime rates and other important factors, will be considered to help the pizza shop owner to make the final decision.