**Prediction the rate of success of the Starbucks stores in particular area**

**Capstone project**

Business Understanding and Data: January, 2019 – Vitaliy Pustovit

**Introduction**

The success criterion for the particular business can be converted to a predictive modeling criterion so the modeler can use it for selecting models. One of the key parameters was used in this project is a customer rating issued by the readers of the Foursquare.com service. Combined with other parameters like “cost of living” and “livability” it helps to carefully estimate the level of success of the particular restorant business for different geographical location.

## **Objective**

The objective of this project is to explore the possibilities to expand Starbucks cafe network in the Ohio state neighborhood and predict the place with greater chance of success due to the characteristics of the place (cost of living, livability). To do that we will analyze variety of Starbucks locations with different level of success and suggest places to open a new one following these characteristics.

### **Stakeholders**

Future business managers who want wisely locate restaurant network.

### **Selection criteria**

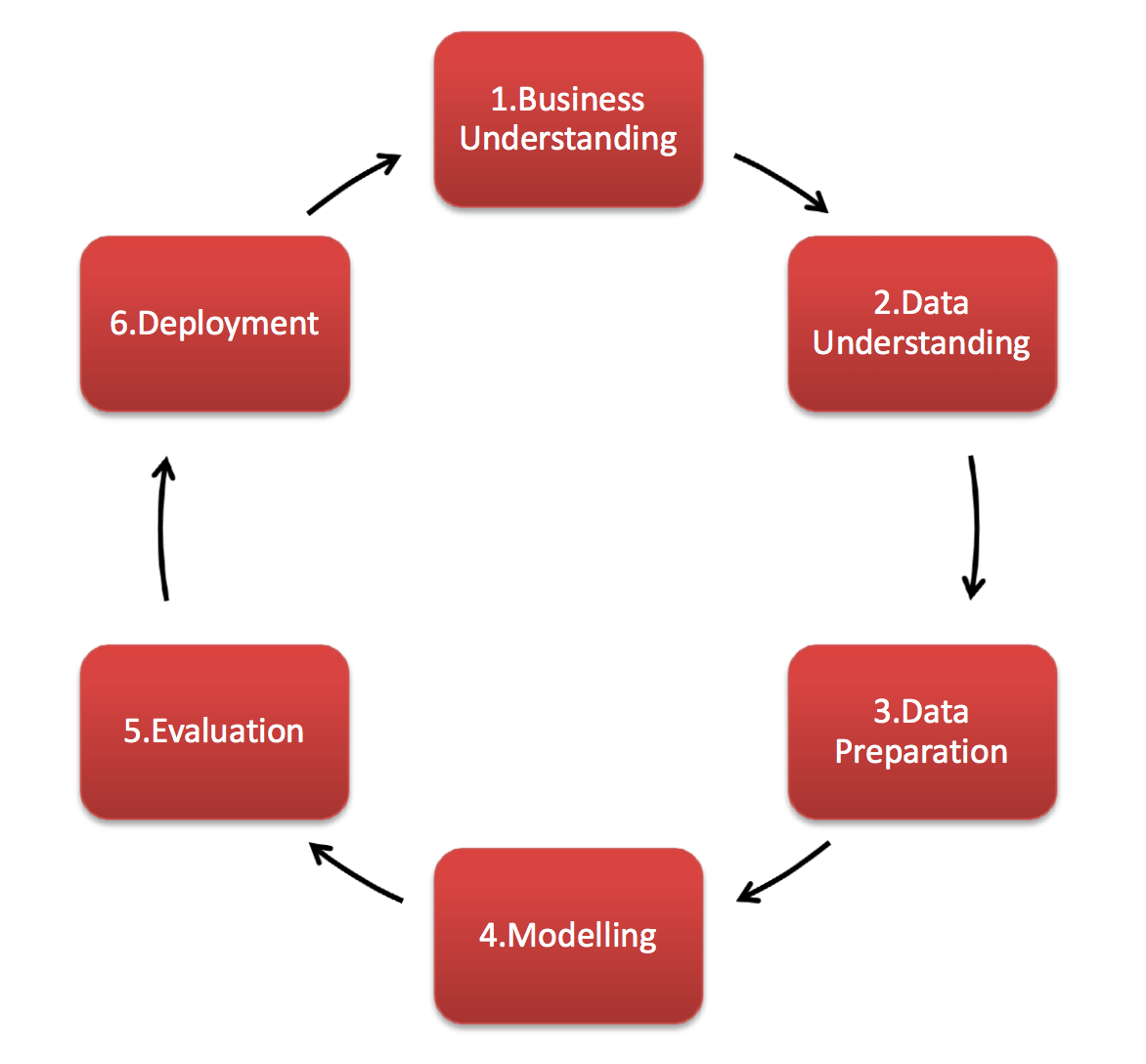
The location of the establishment can guarantee the success of sales and, for this, following some behaviors identified in Ohio we have, as points indicated for the opening of a cafe of this type some factors that increase the chance of consumer frequency and correspondently level of success:  
- Must be located within area with high cost of living  
- Must be located within area with high rating of livability

That information was taken from the realtors database: <https://www.areavibes.com>.

The classification and clustering were performed to classify data into two sets, one with high customer rating (1) and one with lower response (0). The corresponding Map is provided.

## **Methodology**

For this project the following methodological steps must be performed:



**1. Business Understanding**

Focuses on understanding the project objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem definition and a preliminary plan.

**2. Data Understanding**

Starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information.

**3. Data Preparation**

The data preparation phase covers all activities to construct the final dataset from the initial raw data.

**4. Modeling**

Modeling techniques are selected and applied.  Since some techniques like neural nets have specific requirements regarding the form of the data, there can be a loop back here to data prep.

**5. Evaluation**

Once one or more models have been built that appear to have high quality based on whichever loss functions have been selected, these need to be tested to ensure they generalize against unseen data and that all key business issues have been sufficiently considered.  The end result is the selection of the champion model(s).

**6. Deployment**

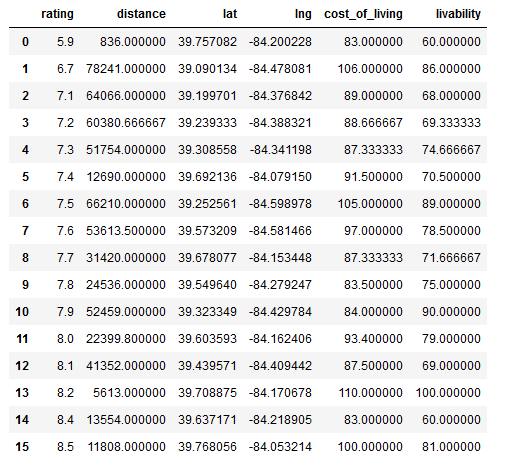
Generally, this will mean deploying a code representation of the model into an operating system to score or categorize new unseen data as it arises and to create a mechanism for the use of that new information in the solution of the original business problem.  Importantly, the code representation must also include all the data prep steps leading up to modeling so that the model will treat new raw data in the same manner as during model development.

## **Data Preparation**

The following data cleaning, normalization and aggregation were needed to allow correct use of the data mentioned above: The Pandas code retrieves the attributes for all starbucks location.

This code uses the data from foursquare.com to get the store number, latitude and longitude and customer rating. Also, an additional parameter the “cost of living” and “livability” rates were introduced to characterize particular neighborhood and taken from the realtors public database: <https://www.areavibes.com>.

Collected data for the set of Starbucks shops located on distance 100km from center of Dayton OH and grouped by customer raiting:



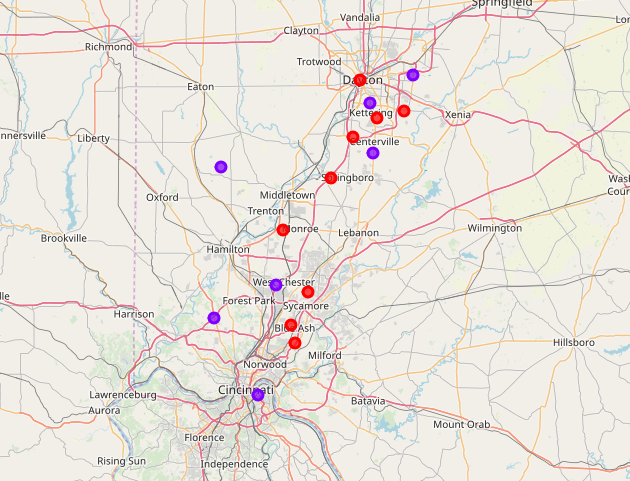
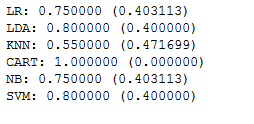


Figure1. A set of the Starbucks shop clustered into two groups by kmeans method with high and low customer raiting.

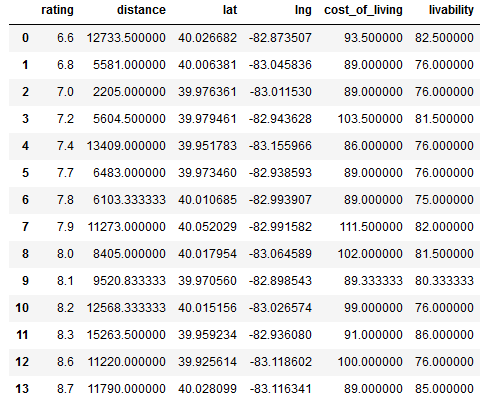
**Results and Discussion**

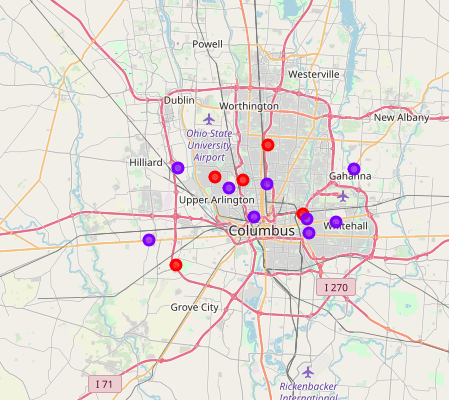
A spot check of the different ML algorithms were provided to select optimal ML training procedure:



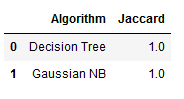
Then Decision Tree (CART) and Gaussian NB (NB) algorithms were applied to train model on the given data set mentioned above.

Next step was to pick (randomly) other set of stores around Columbus OH within distance 100km. Again, using this lat and long values, the Pandas code retrieves the attributes for all starbucks location. The classification and clustering again were performed to classify data into two sets: one with high customer rating (1) and one with lower response (0). Collected data for the set of Starbucks shops located on distance 100km from center of Columbus OH and grouped by customer raiting:





The corresponding Map is provided. The obtained “features” data was used on initially trained Decision Tree and Gaussian NB models to predict customer rating for this new set of stores. It seems prediction works well (100%).



Conclusions

It seems it was possible to classify set of Starbucks café in two groups based on consumer rating which we directly relate to the rate of their success. The model trained (ML methods) on that one set located in one geographical area seems working well on other set of cafes in the completely different area.