**To upgrade or not to Third Wave Coffee**

Virat G Singh

January 26th, 2020

1. Introduction
   1. Background

The third wave coffee movement is a movement led by both consumers and manufacturers to consume, enjoy, and appreciate high-quality coffee. This movement considers coffee an artisanal food, like wine, whose consumption experience can be enhanced with greater education, connoisseurship, and sensory exploration beyond just a cup of coffee. While all coffee comes through a similar value stream, third wave coffee seeks to highlight the unique characteristics that result from the diversity of coffee bean cultivars, growing and cultivation methods, processing methods, roasting methods, and the variables in beverage preparation.

* 1. Problem

There are many coffee shops in Bengaluru and most of them haven’t embraced the third-wave of coffee. As third-wave of coffee is gathering steam all around the world and new ventures in Bengaluru are starting to offer it, which coffee shops in Kengeri (an area in Bengaluru) should or shouldn’t upgrade to offer third wave coffee?

Investors, knowing the Third Wave Coffee revolution in US and its potential in Bengaluru would be curious to know which ones to invest in.

1. Data Acquisition
   1. Data Source
      1. Third-wave and Non Third-wave coffee shops

These coffee shops are scraped from the yelp website as the training set would consist of ones from California. This can be done using BeautifulSoup python library.

* + 1. Coordinates of those areas

Using geocoder package, coordinates of these areas can be obtained which will then be used to get its venue id in Foursquare API. The venue id is important as it acts as a key in FourSquare API to open up details of that particular venue.

* + 1. Details of the Coffee Shops

Further information of these coffee shops can be obtained by calling them using their venue IDs in Foursquare API. These will be details like Likes, Rating, Top five Venues visited after visiting the coffee shop, Total number of shops present in top five venues details.

* + 1. Preparing Test Dataset

Using coordinates of Kengeri area, coffee shops near it and its venue id are collected and constructed as a table. After this, venue id column of the dataset is used to collect all features collected for the training dataset using FourSquare API.

* 1. Data Cleaning

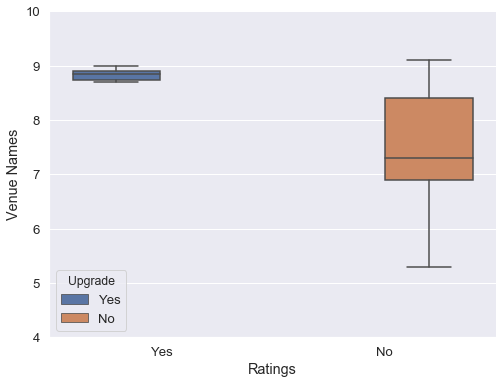
Once the dataset is obtained with all the details required, some of the instances were incomplete. These were replaced by “NaN” and dropped from obtained table. The sum of five NextVenue columns will be taken as “TotalShops” column.

* 1. Feature Explanation
     + - Upgrade
* "Yes" means the venue should and has upgraded to serve third-wave coffee.
* "No" means the venue shouldn't and hasn't upgraded to serve third-wave coffee.
  + - * The five NextVenues
* these are the top five venues visited after visiting the Coffee Shop.
* The places types are primarily categorised as either a "Shopping Activity" or a "Non-Shopping Activity"
  + - * Likes
* Total number of people who have liked the venue in FourSquare API
  + - * Rating
* An out-of-ten score of the venue in FourSquare API
  + - * TotalShops
* Total number of shops visited after visiting the coffee shop, ranges from 0 to 5

1. Exploratory Data Analysis
   1. Target variable

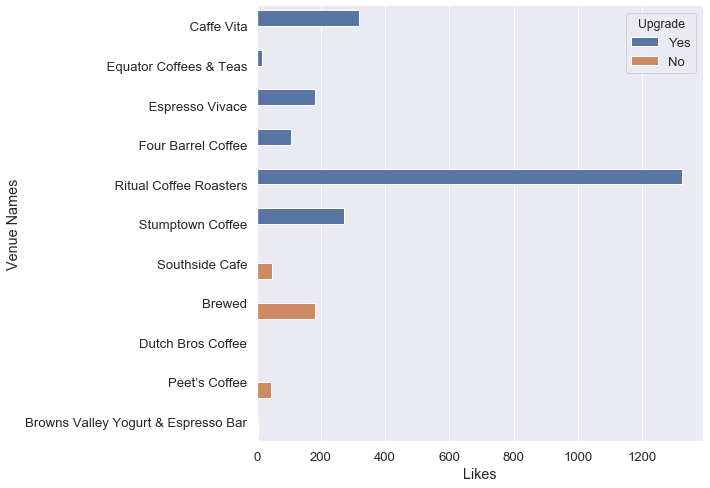
The target variable is “Upgrade” as it shows whether the Coffee Shops should upgrade or not, which is needed to solve the problem stated above.

* 1. Relationship between Ratings and Upgrade feature



The above figure shows that there is a relationship between Rating and Upgrade feature of a Venue and it turns out to be a significant one too. It shows that Coffee Shops with high ratings are most likely to be serving third-wave coffee when compared to ones which do not serve third-wave coffee.

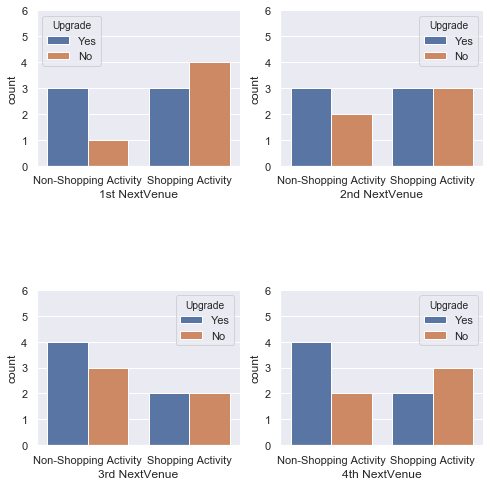
* 1. Relationship between Likes and Upgrade feature



The above figure shows that there is a relationship between Likes and Upgrade feature of a Venue and it turns out to be a significant one too. It shows that Coffee Shops with huge number of Likes are most likely to be serving third-wave coffee when compared to ones which do not serve third-wave coffee.

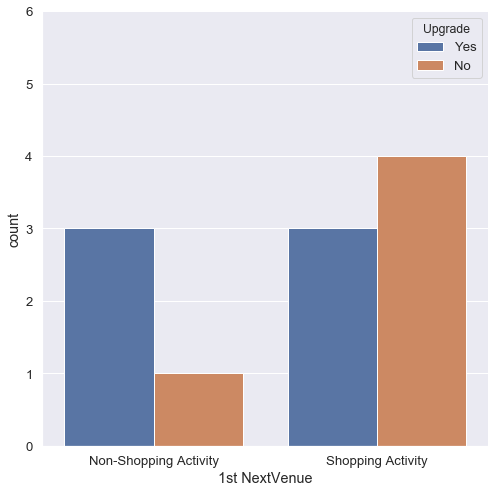
This shows that third-wave coffee shops are mostly the ones able to attract the masses, which could be due to its higher quality of coffee and places surrounding it.

* 1. Relationship between Top four NextVenues visited and Upgrade feature



Now, users go other places after having coffee at the coffee shop which can be used to paint a picture of places around it. After getting the above visual, it looks like only the most visited venue after having coffee has a relationship with upgrade feature. The other three NextVenues has no relationship with Upgrade feature as there is no relevant change observed regardless of it being a third-wave coffee shop or not.

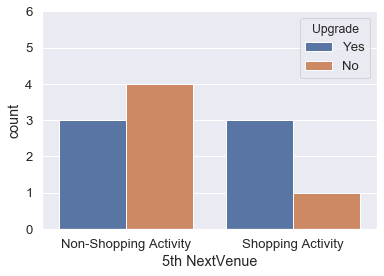
* 1. Relationship between 1st NextVenue and Upgrade feature



After getting the above visual, it looks like the most visited venue after having coffee has a relationship with upgrade feature. The non-Third wave coffee Shops most visited venue after having coffee is a Non-Shopping activity, this means that if the most visited venue after coffee is a Shopping activity then it’s mostly likely to be done after having coffee from a Non Third-wave coffee shop.

On the flipside, it is the opposite for Third-Wave Coffee Shop and couldn’t be represented here due to less number of instances present in the dataset.

* 1. Relationship between 5th NextVenue and Upgrade feature



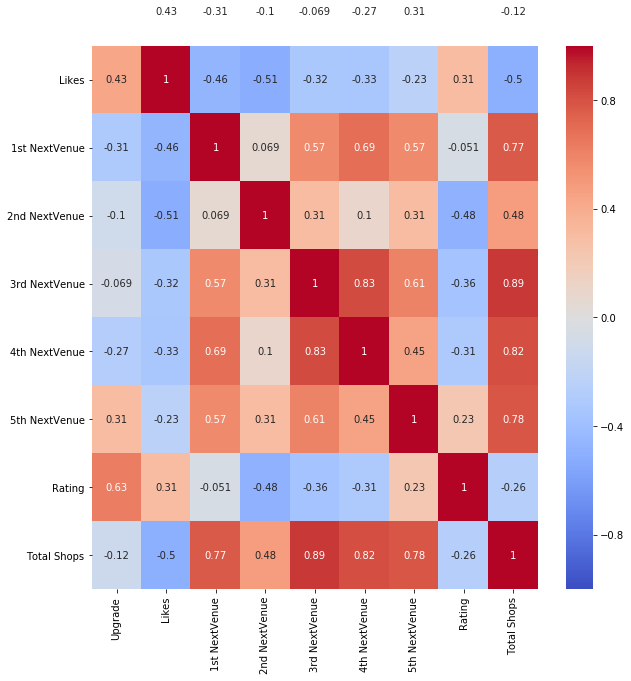
After getting the above visual, it looks like the fifth most visited venue after having coffee has a relationship with upgrade feature. The non-Third wave coffee Shops fifth most visited venue after having coffee is mostly a Shopping Activity; this means that if the fifth most visited venue after coffee is a Non-Shopping activity then it’s mostly likely to be done after having coffee from a Non Third-wave coffee shop.

On the flipside, it is the opposite for Third-Wave Coffee Shop and couldn’t be represented here due to less number of instances present in the dataset.

1. Predictive Modeling

To predict the target feature, classifications algorithms like KNN, SVM, Decision Tree and Logistic Regression will be used to as most of the variables are categorical variables and had to be converted into numbers using dummy values in order to fit into the model.

After doing Correlation analysis, a heatmap is obtained. It has all the features of the dataset and is given below.

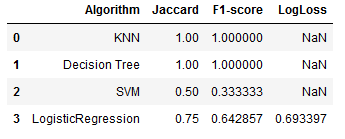


Based on above statistics, following independent variables are to be taken into consideration for predicting dependent variable "Upgrade".

* + - * Likes
      * Rating
      * 1st NextVenue
      * 5th NextVenue

1. Result

The following table, which shows the model evaluation scores of all classifications algorithms was obtained,



From above table we can conclude that Decision Tree is the best algorithm to predict as it not only has the most accuracy but also can return similar numbers when large testsets are given unlike KNN, which wouldn't be able to do if the instances aren't similar to each other.

After applying the test dataset to Decision Trees, it predicted that coffee shops near Kengeri haven’t and shouldn’t upgrade to serve Third-wave Coffee.

1. Discussion

At the end, we get the following observations.

* Coffee Shops with high ratings are most likely to be serving third-wave coffee when compared to ones which do not serve third-wave coffee.
* If the most visited venue after coffee is a Shopping activity then it’s mostly likely to be from a Non Third-wave coffee shop.
* If the fifth most visited venue after coffee is a Non-Shopping activity then it’s mostly likely to be done after having coffee from a Non Third-wave coffee shop.
* Third-wave coffee shops are the ones able to attract the masses, which could be due to its higher quality of coffee and places surrounding it.

1. Conclusion

Although certain models shows a 100% accuracy, the training dataset doesn’t have as many instances as required than required but the above observations are relevant and can be backed even better using larger datasets. It would also increase the relevance of the top five NextVenues relationship with target feature.

More features like user check-ins could help in making models more effective as it could present different scenarios which, at the end of the day is important to generate trustworthy results and in general understand the Third-wave coffee scene in USA.