Predicting the Success of Starbucks Locations in United States Cities

Alexis Raymond

May 20, 2020

1. Introduction

1.1 Background

Founded in Seattle, Washington in 1971, Starbucks is one of the most important and successful coffee companies in the world today. Known for its taste, quality and, above everything else, customer experience, the brand has expanded to over 25,000 locations worldwide in the past 50 years. Some of these sites have known extreme successes, whereas others have had to close their doors following some kind of failure.

1.2 Problem

Multiple factors can be the cause of these successes and failures such as a difference in product offerings or a less capable workforce. However, this paper will focus on the most important determining factor in a Starbuck's success: it's location. In fact, since all Starbucks across the world offer the same quality and taste, and that the customer experience is similar across the board, the most variable factor is the location. This includes the competition in the area (*supply*), as well as the demographics of the targeted customers (*demand*). Hence, this project aims to predict whether or not a Starbucks will be successful based on its location.

1.3 Interest

The primary audience of this project is Starbucks itself. In fact, being able to predict whether a location will be successful or not would constitute an enormous competitive advantage for the Seattle coffee giant, both financially and for marketing purposes. The secondary audience are companies similar to Starbucks (e.g. Tim Hortons, Second Cup, Dunkin Donut) who could use the study for comparable purposes.

2. Data acquisition and cleaning

2.1 Data sources

In order to achieve the results presented above, two datasets and the use of a location data API (in this case, Foursquare) are needed. First, Kaggle offers a free dataset containing a record for every Starbucks store in operation as of February 2017 (Starbucks Locations Worldwide). Second, a list of all US cities with a population of at least 100,000 people as of July 1, 2017 can be scraped from the following Wikipedia page: List of United States cities by population. This dataset also contains each city's area and population density. Finally, the Foursquare Places API will provide a rating for every location of interest as well as a detailed list of their nearby venues.

2.2 Limitations of the data

2.2.1 Foursquare Places API

As powerful as it is, the Foursquare Places API has certain limitations. First, being primarily crowdsourced content, the platform does not have a profile for each Starbucks. Paired with some unexpected errors in the API calls (2), this resulted in only 3,807 Starbucks found in the Foursquare database out of the 5,866 that were looked for (success rate of 64.90%).

Second, Foursquare imposes a daily call quota of 99,500 regular calls and 500 premium calls to users in the *Personal Tier*. Since requesting a venue's rating falls in the premium calls category, it would be difficult and time consuming to obtain the rating for every 5,869 Starbucks in the United States with a Foursquare *Personal Tier* profile. Thus, 500 locations were chosen randomly, and API calls were made to obtain their ratings. From those 500, 2 did not have a rating, and 114 had less than 10 ratings, making them ineligible. Hence, only 384 out of the 500 locations obtained a valid rating. *(See below for more details on the data cleaning process)*

2.2.2 Accuracy of target variable

Being an external research project on Starbucks' success, only open-source data is available. Therefore, the only measure of success that can be used is online reviews. While ratings are a good indication of customer satisfaction, they do not show a complete picture. In fact, a venue can be financially successful without being liked by its customers, even though this is extremely rare. Hence, readers must remain sceptical of the results stated below as they only cover one component of success.

2.3 Data Cleaning

2.3.1 List of US cities

Being scraped from the Internet, the list of US cities with a population greater than 100,000 had a few problems. First, multiple entries in the City column of the dataset had footnotes in brackets next to their name. These were easily removed by writing a script to clean the city names.

Second, 23 had duplicate names; meaning 2 or more cities had the exact same name. This caused problems because though it would have been possible, it would have been difficult and time consuming to distinguish the cities based on the state of the Starbucks location. Thus, it was decided to remove all cities with duplicate names from the list. Therefore, the Starbucks locations in those cities were also removed from the dataset.

Table 1. Cities removed from the list because of duplication

City	States
Aurora	Illinois, Texas
Columbia	Missouri, South Carolina
Columbus	Georgia, Ohio
Glendale	Arizona, California
Kansas City	Kansas, Missouri
Lakewood	Colorado, New Jersey
Pasadena	California, Texas
Peoria	Arizona, Illinois
Richmond	California, Virginia
Rochester	Minnesota, New York
Springfield	Illinois, Massachusetts, Missouri

Third, the area and density columns had units at the end of their values and used commas to separate every third number. Once again, writing a script allowed the removal of the unwanted characters and allowed for the conversion of the column to numbers.

2.3.2 Nearby Venues

Nearby venues (venues from the Foursquare Places API in a radius of 500 meters of the location) were found for most Starbucks location. However, out of the 5,869 locations that went through the script, 3 did not return any nearby venues. This can be explained by the fact that no nearby venues are listed on Foursquare in a radius of 500 meters of that specific location. For the locations that did return at least one nearby venue, one hot encoding was used to convert the categorical values to quantitative values that can be read by the model later by using the process of binarization. Finally, the venues were grouped by the Starbucks location they were associated with and the mean of their frequencies was taken in order to show the proportion of presence of each venue category close to a Starbucks location.

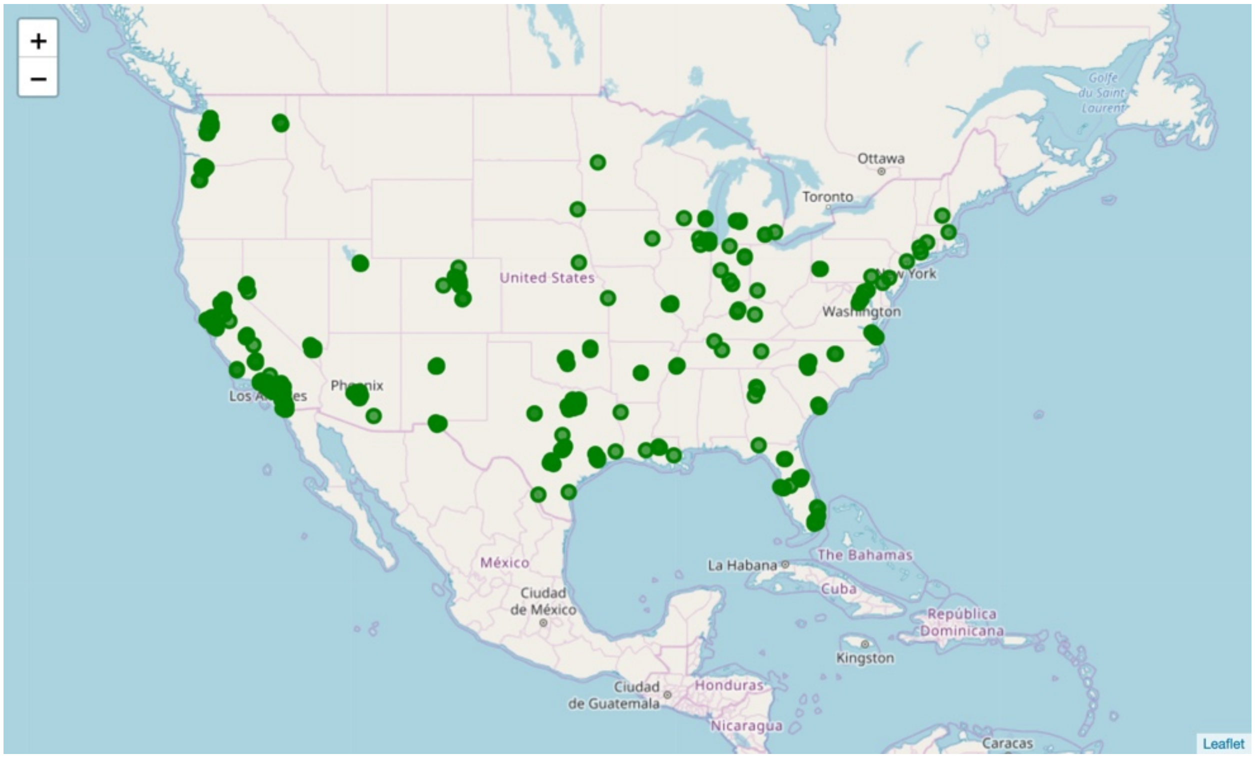
2.3.3 Ratings

In order to measure the success of each Starbucks location, its rating was pulled from its Foursquare profile using the API. The first step in doing so is to find the Foursquare ID for that specific Starbucks. The process used to achieve this was to search for all venues in a radius of 500 meters of the alleged coordinates of the store, filter for all Starbucks found, and keep the ID of the closest one. During this step, 2 locations generated request errors and 2,019 locations were not found because the store does not have a Foursquare profile. Thus, 3,807 were ready to go to the next step, which was to retrieve their rating. However, as mentioned above, the Foursquare Places API only allows 500 premium calls per day. Therefore, 500 locations were randomly selected out of the 3,807. From those 500, 2 did not have a Foursquare rating, and 114 did not have more than 10 ratings (arbitrary threshold decided to avoid small sample size problems). Hence, the final dataset contains 384 Starbucks locations.

Table 2. Evolution of dataset

Step	Initial	Final	Removed	Reason
US cities	25,599	5,869	19,730	These locations were not in the identified US cities
Nearby venues	5,869	5,866	3	These locations did not have nearby venues in Foursquare
Foursquare ID	5,866	3,807	2 2,019	Request errors Starbucks not in Foursquare
Daily call quota	3,807	500	3,307	API daily call quota of 500 premium calls
Ratings	500	384	2 114	Rating not found Less than 10 ratings

Figure 1. Locations of the retained Starbucks



2.4 Feature Selection

2.4.1 City Data

The demographics of the city in which a restaurant or other kind of venue is located can have a great impact on the success (or failure) of this commerce. This is why two characteristics were captured for each Starbucks location's city: the area (in km^2) and the population density (in population per km^2).

2.4.2 Nearby Venues

The venues surrounding a Starbucks location give an idea as to whether or not there is a lot of competition in the surroundings. It is safe to assume that a Starbucks surrounded by many other coffeehouses will not be as successful as one alone. Therefore, by using one hot encoding, a list of all venues registered on Foursquare in a radius of 500 meters of a Starbucks location was added to the model.

2.4.3 Distance to headquarters

"Research has shown that distance constrains the flows of goods, capital, and information across and within countries, and both between and within firms."[1] Therefore it was important to capture, and include in the model, the distance between the store and the Starbucks headquarters in Seattle. The hypothesis here is that the probability of a location being successful decreases as it gets further from the head office.

2.4.4 Rating (target variable)

As mentioned above, being an external research project, the only accessible data to measure the success of every location was its customer satisfaction. The metric chosen was its rating (out of 10) on Foursquare, a popular virtual city guide with over 50 million users. In order to avoid small sampling issues, only locations with more than 10 reviews were retained.

[1] https://scholarship.sha.cornell.edu/cgi/viewcontent.cgi?article=1296&context=articles

Table 3. Retained features

Feature	Description
Area	Area of the city in which the Starbucks is located (in km^2)
Density	Population density of the city in which the Starbucks is located (in population per km^2)
Nearby Venues	Categories of venues located in a radius of 500 meters of the Starbucks
Distance to HQ	Distance between the Starbucks and the Starbucks head office in Seattle
Rating	Target variable: Measure of success of the Starbucks

3. Exploratory Data Analysis

3.1 Target feature analysis

As stated above, the target feature chosen for this research is the location's average rating on Foursquare. In order to better understand the meaning of this value, we must first of all understand how it is distributed. As we can tell from the following histogram (figure 2) and its associated boxplot (figure 3), the ratings seem to be normally distributed around the value of 7.75 with only 5 outliers between the value of 5.5 and 6.

Figure 2. Histogram of the ratings distribution

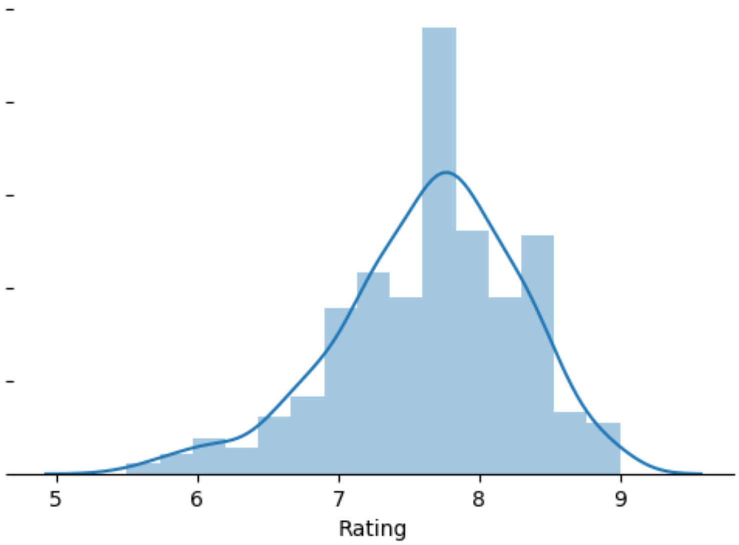
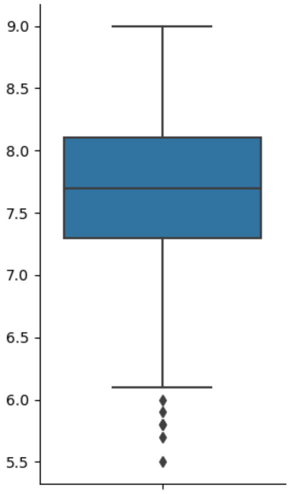


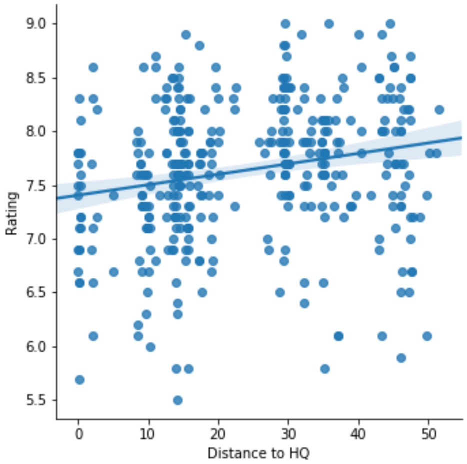
Figure 3. Boxplot of the ratings distribution



3.2 Relationship between distance to HQ and rating

A hypothesis made at the beginning of this research project was that the further away a Starbucks store was from the headquarter in Seattle, the harder it would be for the location to produce a positive customer experience, generating a lower rating. However, when performing a correlation analysis, we found that this is not necessarily the case. In fact, the correlation coefficient between the rating of a store and its distance to the HQ is 0.21. This value is too low to conclude that there is a relationship between the two elements but, the slope would imply the opposite of our hypothesis. As you can see in the regression plot below, the data suggests that the further away from the headquarter a store is, the higher his rating. Again, this is not conclusive because of the low coefficient.

Figure 4. Regression plot between the rating and the distance to HQ



3.3 City demographics

For a predictive model to be efficient and effective, it cannot have two correlated features. Therefore, we quickly checked that the *area* and *population density* features were not correlated. This was proven by their extremely low correlation coefficient of -0.06. We can use the features in our model.

Another hypothesis made at the beginning of this project was that the area and density of a city would have an impact on its performance. Once again, this was disproven by their respective correlation coefficients of -0.03 and -0.17. The two regression plots below visualize the absence of a relationship.

Figure 5. Regression plot between the rating and the area of the city

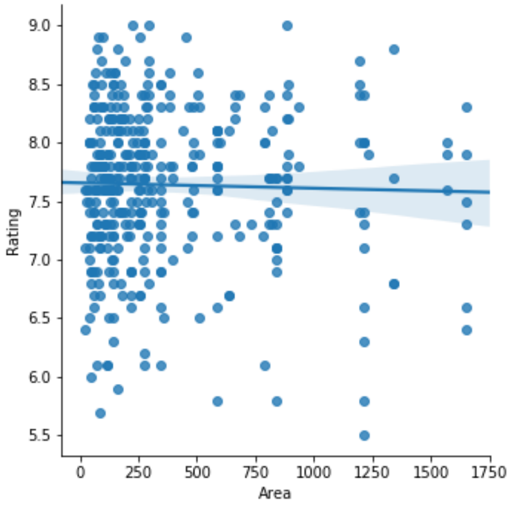
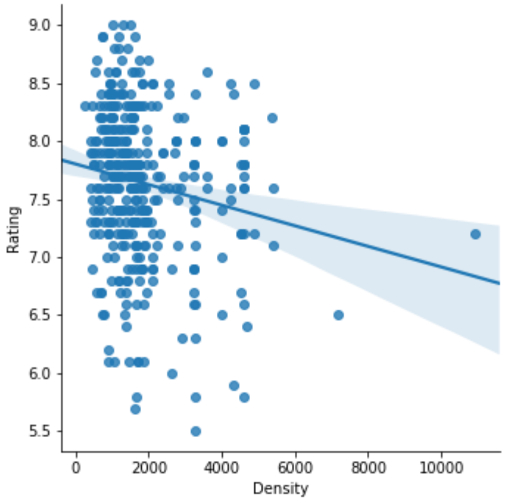


Figure 6. Regression plot between the rating and the density of the city



4. Predictive modeling

Now that we have our dataset and that some basic data analysis has been performed, we are ready to create the predictive model. Most of our hypothesis have been wrong so far but that's alright. In fact, they only constitute small factors in the Starbucks potential for success. The real goal is to use all the features collected to create a more accurate model.

4.1 Regression model

A linear regression model was used in this research as it is the best for predicting quantitative values with a limited dataset. Another option would have been to use a classification model to determine which Starbucks are likely to be rated high and which ones would have a low ranking. This option was not considered as I believed it would not offer significant information to the reader and determining an arbitrary value would have been subjective.

4.2 Results

The linear regression model created was trained using 617 features to predict a target variable, which in our case was the store's ranking on Foursquare. These features were composed of the area of the city in which it was located, its population density, the distance between the store and the Starbucks headquarter in Seattle and which buildings were in proximity to the store.

The following table shows the 5 coefficients that lead to a high ranking and the 5 coefficients that lead to a low ranking. These values can be interpreted the following way: *Holding all other features fixed, a 1 unit increase in proximity to a certain type of building results in an increase/decrease of y in the store's ranking*, where y is the value of the coefficient.

Table 4. Top positive and negative coefficients

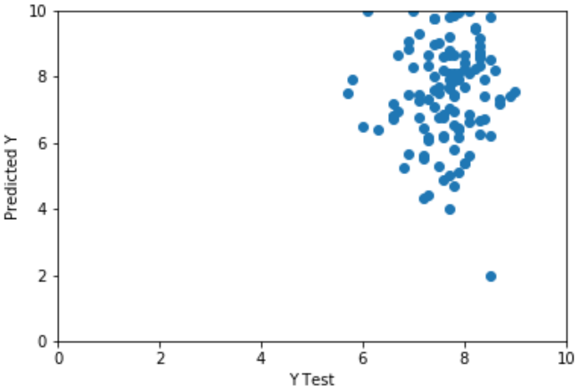
Feature	Coefficient	Feature	Coefficient
Pool Hall	42.08	Recreation Center	-43.06
Football Stadium	29.26	Beer Bar	-36.62
Garden Center	28.74	Auto Dealership	-34.84
Pier	26.60	French Restaurant	-31.34
Plaza	24.07	Shoe Repair	-29.35

4.3 Performance of Model

4.3.1 Predicted vs real ratings

The first evaluation method use was to create a quick and easy visualization of the relationship between the predicted values for the ratings and their real values in a scatter plot. This showed us that the model contained some outliers but that in general, there was a correlation between predicted and real values. An example of an outlier is the point located at approximately (8.25, 2). This shows that the model expected this Starbucks to be poorly rated at 2/10 but that in reality the store received an excellent rating of 8.25/10. This is a perfect example proving that the user should take the results of this study with a grain of salt and not consider them absolute.

Figure 7. Predicted ratings vs real ratings



4.3.2 Performance metrics

The predictive model achieved the following metrics when predicting the ratings of our testing dataset.

Table 5. Performance metrics

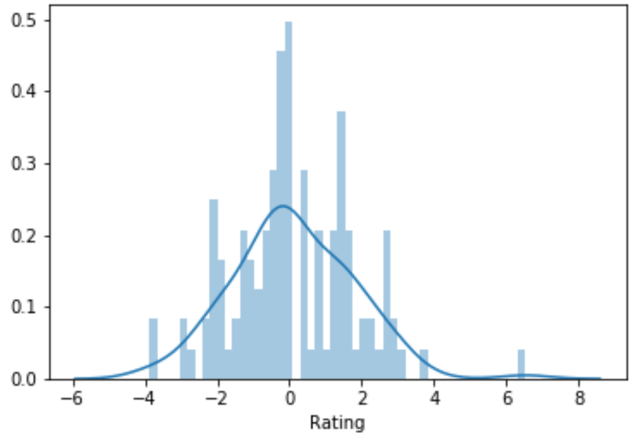
Metric name	Value
Mean Absolute Error	1.28
Mean Squared Error	2.80
Root Mean Squared Error	1.67

In a perfect model, these values would be 0 as there would be no differences between the real values and the predicted ones. However, when creating a predictive model, you want to achieve some error to avoid overfitting. Overfitting refers to a phenomenon where your model is only effective with the values passed to train it. It cannot be used to predict other values.

4.3.3 Residual distribution

Finally, we wanted to ensure that the residuals of our model followed a normal distribution. This would mean that it is not biased towards predicting a higher rating as opposed to a lower one. As you can see in the distribution plot below, the residuals are normally distributed around 0, meaning that it does not appear to be biased.

Figure 8. Residual Distribution



5. Conclusions

In this study, I analyzed the relationship between a Starbucks rating and its location by analyzing the impact of the city demographics, the distance to the headquarter and the buildings around it. The result of this research was that the only good indicators of whether a Starbucks will have a good customer satisfaction rate is the type of commerce around it. However, executives should use the results outlined above lightly as they do not show a complete picture of how successful the store really is.

6. Future directions

To improve the accuracy of the model and the usefulness of the results, the experience should be repeated with more data. This can include gathering data from all around the world (vs just in the US), receiving ratings from more than one source (e.g., google maps, yelp) and collecting more demographic data such as the average household income of the city, its average age, and most prominent ethnicity. Additionally, the results would become much more valuable if the research was able to include financial data about the store.