Predicting Starbucks Locations

Brett Waugh

University of South Florida

Author Note:

Brett Waugh, School of Information, University of South Florida

Waughb@mail.usf.edu

Predicting Starbucks Locations

Predicting the number of Starbucks locations has many practical uses and interested groups. Starbucks is interested in predicting where they should put a new franchise and competing companies may want this information in order to prepare for the new competition. This study walks through six different methods for predicting Starbucks locations. The first three methods use data Starbucks provided and show methods that everyday individuals may use to predict Starbucks locations. The next three methods will use data from several sources and more advanced modeling techniques to predict Starbucks locations. The benefits and costs of each method are discussed; exploring if using more complex modeling is warranted over other methods.

The original dataset

The original dataset was found on Kaggle (Starbucks, 2017). There were some datasets online that required a membership or cost to use, but the Kaggle dataset was from of cost. There were many fields in this dataset that were not needed for this project and were stripped away early on to expedite processing.

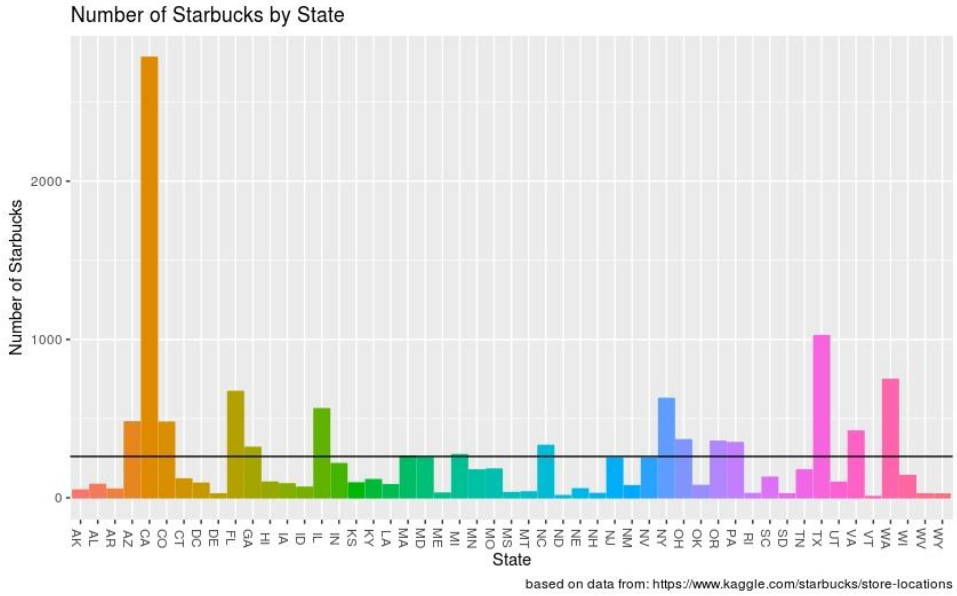
The scope of the project included all Starbucks located in the United States, regardless or ownership type. These distinctions were important to validate because the original dataset provided both Starbucks and Teavana because Starbucks now owns Teavana. The original dataset also provided the locations for all Starbucks in the world, so this needed to be trimmed down to only United States locations. With these altercations made to the dataset, the size of the file shrunk from about 10 MB to around 1.5 MB.

From the original dataset, an additional file was created for the number of Starbucks locations by state. This file is very small and allowed some calculations to be done much quicker.

Method One: Mean

There are many techniques commonly used for predicting values. Many of the most commonly used ones are used not because they achieve more accurate results, but because they provide a quick way to get a feel for the data. Averaging is one of these techniques that may not produce the best representation of the dataset, but it is popular and quick to implement.

Using the file with the associated states and number of Starbucks in that state, the average is quickly worked out to 261 Starbucks locations per state.

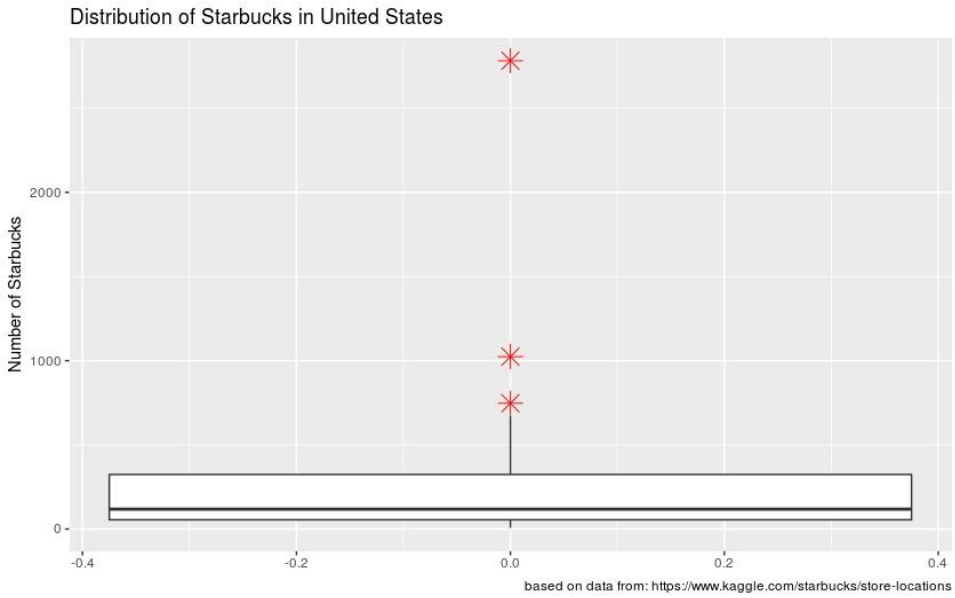


After viewing the histogram, we can see that this number does not encapsulate a single state's number of Starbucks locations. There are a few states that come close (MA has 262, MD has 252, MI has 272, NV has 249, and NJ has 249) but none that are this exact number.

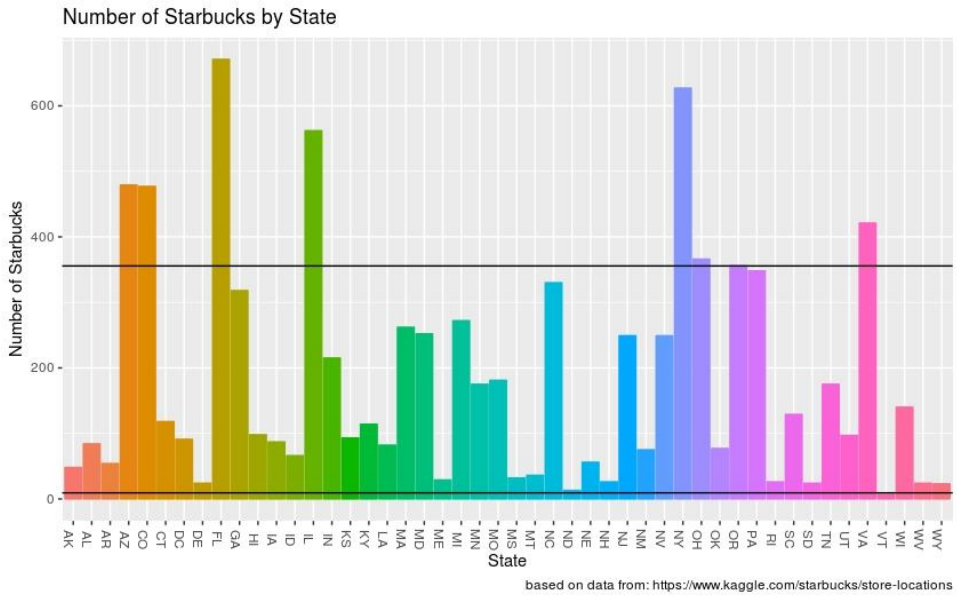
Averaging can be useful for getting a quick feel for a dataset, but it does not accurately represent the distribution of the dataset.

Method Two: Within a Standard Deviation, Removing Outliers

To improve upon the mistakes from Method 1, a deeper understanding into the distribution of the data is needed.



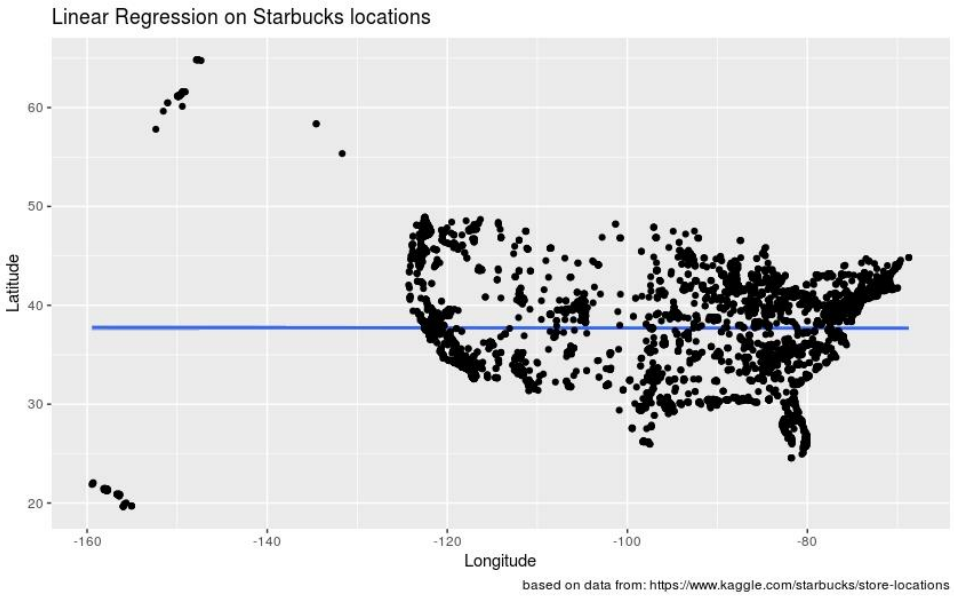
As shown in the boxplot, there are three upper bound outliers. The largest one (CA) being so much more than the rest that it heavily distorts the graph. Removing outliers usually increases the understanding of a dataset because those events are not representative of the overall dataset (Downey, 2015). The standard deviation before removing any outliers is 421.41, but by removing the three upper bound outliers the standard deviation drops drastically to 173.17.



Notice that by removing the three upper bound outliers and staying within one standard deviation, much more of the state's Starbucks locations are encapsulated. This method produced an upper and lower bound of [9.29, 355.62] which included 76.47% of the states included. This is a great method if all of the data is readily available but does not discriminate with any features of a particular state. A part of the success to this method is the drastic range in the bounds, a range of 346.33.

Method Three: Linear Regression

Linear regression is a useful technique in finding patterns in data. The underlying principle is attempting to create a line that includes the greatest number of points (Zhao, 2013). In the current dataset, the latitude and longitude are the only numerical values that linear regression can be used on.



The plot above shows every United States Starbucks location. Linear regression does not fit the data well in this case, the correlation was –0.003 with an adjust R-squared value of –6.82 x 10^-5. The failure in using linear regression to find a pattern in the placement of Starbucks shows that Starbucks locations depend on more features than just latitude and longitude for placement.

The need for more data

Working only with the data provided in the original dataset, it is clear that more information is needed to determine how Starbucks places its locations. Additional data was collected from several locations and combined to provide additional information on state's: population (Enchanted Learning, & U.S. Census Bureau, 2017), median income (U.S. Census Bureau, & Wikipedia,2018), number of universities (National Center for Education Statistics, & U.S. Department of Education, 2013), and crime rate per 100,000 people (Johnson, 2016). With these additional features, better predictions may be made regarding locating Starbucks.

Method Four: Multivariate Linear Regression

Similar to the fourth method, multivariate linear regression is trying to match variables to an outcome variable (Zhao, 2013). With multivariate linear regression, many more features are used. This method is particularly useful in determining useful features for other models, because it is so easy to setup and processes quickly.

For the first linear model, the outcome variable is the number of Starbucks and the other variables are: population, median income, number of Universities, and crime rate. After running the model, the results indicate that only population, median income, and number of Universities are significant features (P Value < 0.05) to the model. The first model has an RMSE of 164.62 and an Adjusted R sqaured value of 0.83. For the second model, the crime variable was removed and the RMSE increased slightly to 165.00 while the Ajusted R squared stayed at 0.83.

The crime variable made little difference in the performance of the models. The three features with the most significance are population, median income, and number of Universities. The RMSE of around 165 is a reasonable value considering the dataset. The models were able to be setup quickly and easily, with little hassle involved.

Method Five: Logistic Regression

The model often used for logistic regression is the Generalized Linear Model (GLM) (Dietrich et al., 2015). This model shares the fourth method's ease of use and can be quickly setup.

The first logistic model is setup with all the features: population, median income, number of Universities, and crime. This model received an RMSE of 164.62. This model performed about the same as the models from the fourth method. For the second logistic model, the crime feature is removed and the RMSE increases to 165.00, similar to what happened in the fifth method.

The third model was constructed using a Bayesian GLM instead of the same method as the previous two. This model took more effort to setup and was performed with ten-fold cross validation. The RMSE for this model was 201.29 with an R squared of 0.80.

The logistic models performed about the same as the linear models, save for the Bayesian GLM that increased the RMSE by about 40. The first two models were simply constructed while the third model took more understanding about model creation to perform.

Method Six: SVM Models

The Support Vector Machine (SVM) models are considered one of the best out of the box classifiers (Witten et al., 2017). This method handles multiple variables well but takes more knowledge to setup and can take more computational power (Zaki & Meira , 2014; Welling, 2010).

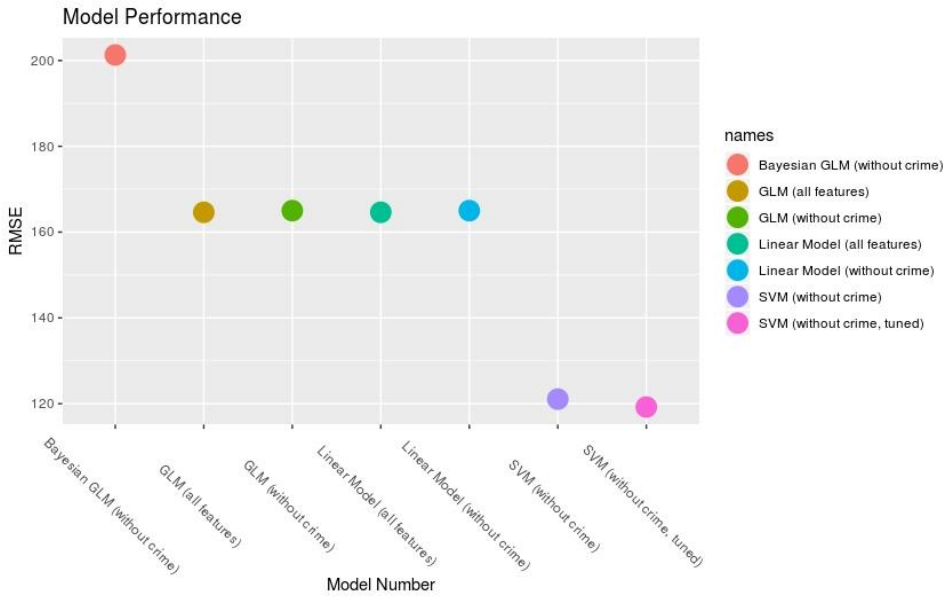
The first SVM was trained with the features: population, median income, and crime rate. It also used a Cost of 10, Epsilon of 1, and ten-fold cross validation. This model was able to produce RMSE of 121.06, significantly less than the RMSE of the previous methods.

Because SVMs have more variables to test than just the features put into it, grid search was used to select the best Cost and Epsilon variables for the second SVM model (Rich, 2017). Grid search saves time because the user does not have to manually test each value when creating the SVM. Using grid search is computationally expensive and can take a while to perform.

Using the Cost and Epsilon values from the grid search, along with the same features of the other model, and ten-fold cross validation produced an RMSE of 119.23. This is the lowest RMSE so far but costed the most computational power and required significant background knowledge to perform.

Results from the Models

A variety of models were used in methods four, five, and six. Each of these models took different amounts of background knowledge, computational resources, and time to setup correctly. The linear models in method four took the shortest amount of time to setup, while the SVMs in method six took the most amount of time to setup. The SVMs also took the most computational resources because of the grid search. The comparison of the RMSE results is shown below.



The best performing model ended up being the final SVM that was tuned using grid search. The other SVM had a comparable score using standard default values for the Cost and Epsilon. The logistic and linear models performed almost identically, save for the Bayesian GLM which produced a significantly higher RMSE than the other models.

Conclusion

There are numerous methods to help people predict the number of Starbucks in an area. The first three methods focused on more traditional methods that people who are not familiar

with more advanced techniques would use. These techniques gave people a sense of familiarity with the data but did not provide much depth to the technique. The first three methods also had a very wide range, with little distinction between states. The last three methods built off other features to give more context into the areas that Starbucks tend to appear in. Using these types of techniques, Starbucks location prediction is much more accurate.

If more granular data was available regarding population, median income, crime rates, and number of Universities by city instead of solely by state then the models may have performed better. The original intent was for the models to perform at the city-level, but a lack of data for all cities forced the models to perform at the state level.

References

Dietrich, D., Heller, B., & Yang, B. (2015). *Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data*. Indianapolis, IN: Wiley.

"Data Science & Big Data Analytics" has been used in multiple classes of mine at USF. This text provides a wide array of topics, while still providing depth. The book is in an easy to read format and includes many visualizations for context. Many of the techniques and methods used are common in the industry, making it a great reference.

Downey, A. B. (2015). *Think Stats* [2.0.38]. Retrieved January 20, 2019.

A more recent book, Downey explores probability and statistics using Python. This book mixes traditional teaching with case studies to show the importance of each concept. Downey covers a wide variety of topics, from probability to survival analysis. The book is free online, and she encourages people to explore the code and concepts, she even published all of the code from the book on GitHub.

Enchanted Learning, & U.S. Census Bureau. (2017, July 1). US States: Population and Ranking. Retrieved April 13, 2019, from https://www.enchantedlearning.com/usa/states/population.shtml

This source was used to create an additional dataset during the project. The dataset can be trusted since it is from the U.S. Census Bureau.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An introduction to Statistical Learning: With Applications in R*. Retrieved January 20, 2019.

James, Witten, Hastie & Tibshirani's "An Introduction to Statistical Learning: With Applications in R" covers a broad range of topics. This book cover statistical analysis in great detail and comes in at over 400 pages. The book combines traditional teaching with labs to reinforce concepts.

The authors of the book are all professors at the: University of Southern California, University of Washington, and Stanford University (two are here), respectively. All of them are published several times in many different places, but usually dealing with statistical analysis. The book is also recent as it was just completed two years ago.

Johnson, M., Jr. (2016, December 28). United States crime rates by county. Retrieved April 13, 2019, from https://www.kaggle.com/mikejohnsonjr/united-states-crime-rates-by-county

This source was used to create an additional dataset during the project. The dataset can be trusted since it provided additional resources and references about the origins of parts of the data.

Matloff, N. (2013). *The Art of R programming: A tour of statistical software design*. San Francisco, CA: No Starch Press.

This text was used to reference some commands and workarounds in R. Some of the syntax for commands can be difficult to remember because they are similar across multiple programming languages.

National Center for Education Statistics, & U.S. Department of Education. (2013, August 12). Colleges and Universities in the United States of America (USA) by State/ Possession. Retrieved April 13, 2019, from http://www.univsearch.com/state.php

This source was used to create an additional dataset during the project. The dataset can be trusted since it is from the U.S. Department of Education.

Rich, K. T. (2017). Machine learning intro in R: Support Vector Regression. Retrieved April 16, 2019, from https://rpubs.com/richkt/280840

This source was used to see how a grid search was constructed for an SVM. The code seems to be a standard method for constructing a grid search.

Starbucks. (2017, February 13). Starbucks Locations Worldwide. Retrieved February, 2019, from https://www.kaggle.com/starbucks/store-locations

U.S. Census Bureau, & Wikipedia. (2018, December 28). List of U.S. states and territories by income. Retrieved April 13, 2019, from https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_income

This source was used to create an additional dataset during the project. The dataset can be trusted since it is from the U.S. Census Bureau.

Welling, M. (2010). *A First Encounter with Machine Learning*. Retrieved January 20, 2019.

Welling's "A First Encounter with Machine Learning" will be used to start me into machine learning. Welling is an assistant professor at the University of California Irvine's School of Information and Computer Science. He is presently at the University of Amsterdam as a research

chair. While this resource is older, the methods I will be using from the book have not changed drastically during this time.

Zaki, M. J., & Meira, W. (2014). *Data mining and analysis: Fundamental concepts and algorithms*. Retrieved January 20, 2019.

Zaki & Meira's "Data mining and analysis: Fundamental concepts and algorithms" does a great job reinforcing fundamental concepts. Zaki is from Rensselaer Polytechnic Institute, Troy, New York, Meria is from Universidade Federal de Minas Gerais, Brazil.

Zhao, Y. (2013). *R and Data Mining: Examples and Case Studies*. Retrieved January 20, 2019, from http://www.RDataMining.com

This text was used to verify certain methods and provide greater understanding to why things could be done. The way the book was written gave a much quicker and easier understanding of the concepts than other texts I have reviewed. The free and open nature of the text makes it a common reference in data analytics.