Drazen Zack

Paul Keglman

Ankur Patel

**Optimal Starbucks Locations**

Where to put the next Starbucks is a question that can make millions of dollars or cost millions of dollars. The goal of the project was to find optimal new Starbucks locations in three cities: Chicago, Philadelphia, and the borough of Manhattan within New York City. The search for optimal new locations in these cities was based on (1) data from the Census Bureau, (2) transit data from each city, and (3) existing store locations in each city. A logistic regression model was used to identify the optimal new locations. A separate model was developed for each city. Each model was trained on two of the cities and tested on the unused city.

The first step was to gather data that could be useful in the prediction of new locations for Starbucks. The first data source that was examined was the Census Bureau. It was a good source for demographic data for the three cities. The census data with the highest spatial detail that could be found was census tract data. Census tracts are areas roughly the size of neighborhoods established by the Census Bureau. The Census Bureau API was used to retrieve the data, which made the process quick and easy. Examples of demographic data extracted from the Census Bureau API were total population, median rent, time people go to work, how the get to work and the median house value. The statistics came from the American Community Survey (ACS) which is a survey conducted by the Census Bureau annually. The data was from the year 2016, the most recent data that could be accessed.

Next, ridership data for every station in each city’s subway or train system was obtained. Ridership is the total number of people that go in and out of the station turnstiles each day. The average ridership was calculated for the weekdays and weekend for each census tract. In some census tracts there is more than one station so each station’s ridership was added together. The data for the transit information was found on the MTA website for Manhattan, the Chicago transit website for Chicago, and the Southeastern Pennsylvania Transportation Authority website for Philadelphia.

The Starbucks location data was found on Kaggle from a user who scraped the locations from the Starbucks store locator. This data was as of February 2017. The census tract of each Starbucks was obtained by entering the address of each store location into the Census Bureau API, which then returned the corresponding census tract information.

Once all of the data was collected and organized into a useable format for modeling, the next step was to find optimal new store locations for each city. The method chosen was logistic regression, a generalized linear model with the sigmoid function. Logistic regression was chosen because the predicted variable was a binary variable (i.e. “No-Starbucks” or “Starbucks”) for each census tract. Logistic regression provides the probability of a Starbucks being in a census tract based on the training data from the other two cities. Once the probabilities were conducted for each census tract for the test city, the top five census tracts that didn’t have a Starbucks were seen as optimal locations for a new Starbucks. The top five census tracts for each city are in a table in the appendix at the end of the paper.

An examination of the results from all three cities revealed patterns in the selected locations. The rent and home values for the selected census tracts tend to be much higher than the city average for all three cities. Also, these optimal locations have a higher percentage of people with either a bachelor’s degree or master’s degree than other census tracts in the city. Another observation was that the optimal selected locations tend to be in census tracts that are right outside of the city’s downtown area. This could come from the fact that there are already a lot of Starbucks locations in downtown area’s or these are additional residential areas that Starbucks hasn’t yet taken advantage of.

During the planning and execution of the project several limitations and opportunities for improvement were identified. One issue was that when gathering data, some variables in the Census Bureau dataset had “NA” values. One could always find a way to estimate these values with the mean or median, but when it involved counting people this could lead to a count that is greater than the total population of the census tract. When this problem was encountered the variable was eliminated from use in the model. Also, a second issue was the consistency of the transportation data among the three cities. For example, the Manhattan data was down to the individual turnstile whereas the Chicago and Philadelphia data were only at the station level. A place for improvement in the modeling of human traffic volume in the vicinity of the locations would be the use of some type of street level traffic data. This has the potential to be a key factor in the analysis. Another place for improvement for gaining additional data could have been adding the number of competitors that are in a census tract.

In summary, a logistic regression model with Census Bureau and transportation data from each of the three cities was used to identify five optimal locations for each city. These locations tended to have higher rent, higher home values and a larger percentage of people with a bachelor’s degree or master’s degree. The optimal locations found in this project have the potential to provide a lucrative return on an investment for an entity who is interested in using this information.

**Appendix**

1. Chicago

|  |  |  |  |
| --- | --- | --- | --- |
| **Census Tract** | **Tract Name** | **Probability of No Starbucks** | **Probability of Starbucks** |
| 17031080202 | Census Tract 802.02 | 42% | 58% |
| 17031240500 | Census Tract 2405 | 46% | 54% |
| 17031070300 | Census Tract 703 | 48% | 52% |
| 17031071000 | Census Tract 710 | 49% | 51% |
| 17031842300 | Census Tract 8423 | 49% | 51% |

1. Philadelphia

|  |  |  |  |
| --- | --- | --- | --- |
| **Census Tract** | **Tract Name** | **Probability of No Starbucks** | **Probability of Starbucks** |
| 42101000801 | Census Tract 8.01 | 37% | 63% |
| 42101000300 | Census Tract 3 | 40% | 60% |
| 42101001202 | Census Tract 12.02 | 41% | 59% |
| 42101000401 | Census Tract 4.01 | 44% | 56% |
| 42101000901 | Census Tract 9.01 | 45% | 55% |

1. Manhattan

|  |  |  |  |
| --- | --- | --- | --- |
| **Census Tract** | **Tract Name** | **Probability of No Starbucks** | **Probability of Starbucks** |
| 36061007800 | Census Tract 78 | 13% | 87% |
| 36061012600 | Census Tract 126 | 18% | 82% |
| 36061008603 | Census Tract 86.03 | 21% | 79% |
| 36061005502 | Census Tract 55.02 | 23% | 77% |
| 36061010800 | Census Tract 108 | 23% | 77% |