**Chipotle Location Analysis**

**Dan Goetz and Joey Sabel**

**IST 707**

**Overview:**

Chipotle is one of the biggest fast casual brands in the country with locations all around the United States. We wanted to look at demographic data from 30,687 cities and predict the best places for Chipotle to open their next location as this company is ever-growing. We also wanted to see not only if Chipotle should open one location in a city, but how many they should open. For example, we should not expect New York city to only have one location since it is home to so many more potential Chipotle fans. The city/demographic data was collected from the US community survey in 2018 and the Chipotle locations were collected in August 2020 from Kaggle. They were joined together with a VLOOKUP using city and state and doing a count of locations per city and per state after doing a Pivot Table to get a count of locations per city. The Chipotle location dataset had addresses and latitude and longitude, which were not helpful, but considered in this approach. We decided that looking at various factors ranging from the median income to racial demographics would be able to help paint a picture of what an “ideal” target city would be for a new Chipotle location. Exploring characteristics of cities that have locations might make it easier to see if other cities fit the target demographic better than others.

**Methods Used:**

To explore the Chipotle locations, we utilized many different types of data analysis, both supervised and unsupervised. Initially, we ran a linear regression to predict the number of Chipotle locations in a given city, followed by a few logit regressions to try to predict whether we expect a city to have a Chipotle location or not. Next, we ran three K-Nearest Neighbors models in order to try to classify cities with and without locations of the restaurant. Our next method of analysis was to look at K-Means clustering to see whether there were certain characteristics of cities which are appealing to Chipotle, based on many demographic figures, where three models were run to compare the data just after initial preprocessing and after additional winsorization, as well as after normalizing the data. Lastly, a decision tree was run to try to simulate a decision-making process to determine whether a city is a good fit or not.

**Variable Selection:**

The variables used for all models (unless noted otherwise) consisted of city and state ID (used for classification only), median age of a city, median household income, population proper, percent of population who is male, average family size, percent of population who is college educated or higher, unemployment rate, and racial classifications which are defined as white, black, asian, native, pacific, multiple, and other. These variables were chosen as we believed they represent a city best out of all of the available demographic figures, rather than something like median rent in a city or labor force participation percent.

**Data Preprocessing:**

The initial methods of preprocessing embodied removing NA values and filtering based on population. There were roughly 12,000 cities removed in this step as some data was not available or cities were deemed too small to be home to a Chipotle location. Next, we adjusted population proper and median income to be at the 99th percentile and 90th percentiles, respectively, to adjust for some outliers, like New York City for population proper; some models would be thrown off if over 8 million people were included in a model that goes down to 601 people.

Other data preprocessing was used for certain models, specifically for our experimented K-Means model and our experimented logit regressions. For the experimented K-Means model, we winsorized all of the variables besides median income and population proper to be between the 5th and 99th percentiles. For the logit regressions, we filtered our data based on the logit results to explore the incorrectly marked cities. For the normalized K-Means model, we normalized every column of the dataset.

**Base Modeling:**

**Linear Regression:**

After preprocessing techniques, we were ready to run our first model with the dependent variable as the number of locations in each city and the independent variables as our chosen variables listed above. The attributes that were statistically significant according to the initial model were median age, median income, population, family size, education, race\_white, and race\_asian. Age, median income, family size, and race\_asian all had negative coefficients, meaning that they have an inverse relationship to the number of Chipotles in that city. For example, younger people with lower median income, smaller families are more likely to have a chipotle in their area. This makes sense because the fast-casual dining experience of a chipotle caters to a young college student who does not have much disposable income. Race\_white also had a negative coefficient. Population had a positive coefficient. This meant the more populated a city, the more likely it is to have more chipotles. In our model, the city with the highest residual was New York City This is because of its massive population. Therefore, our model predicted seventy locations but it actually has fifty-two. Our Adjusted R-squared of this model was .6798.

**Logistic Regression:**

Our next task was to run a logit regression to predict whether or not the city will or will not have any number of Chipotle restaurants. The same variables were significant in this model as in the linear regression, with percent\_male being added. Male had a negative coefficient meaning that more female population are a more target demographic for whether or not to add a Chipotle to the town. According to our model 766 cities have more than a 50% chance of having a Chipotle, which means that we predict them to have one. There are currently 1066 cities with a Chipotle so our model did a relatively accurate job of predicting. However, we realized the population skewing the data because big cities like NYC and Los Angeles had so many more people. The goal of this research was to predict whether or not we should have Chipotles in an average American city. In the final model, we got rid of the population and ran the model with the same other attributes.

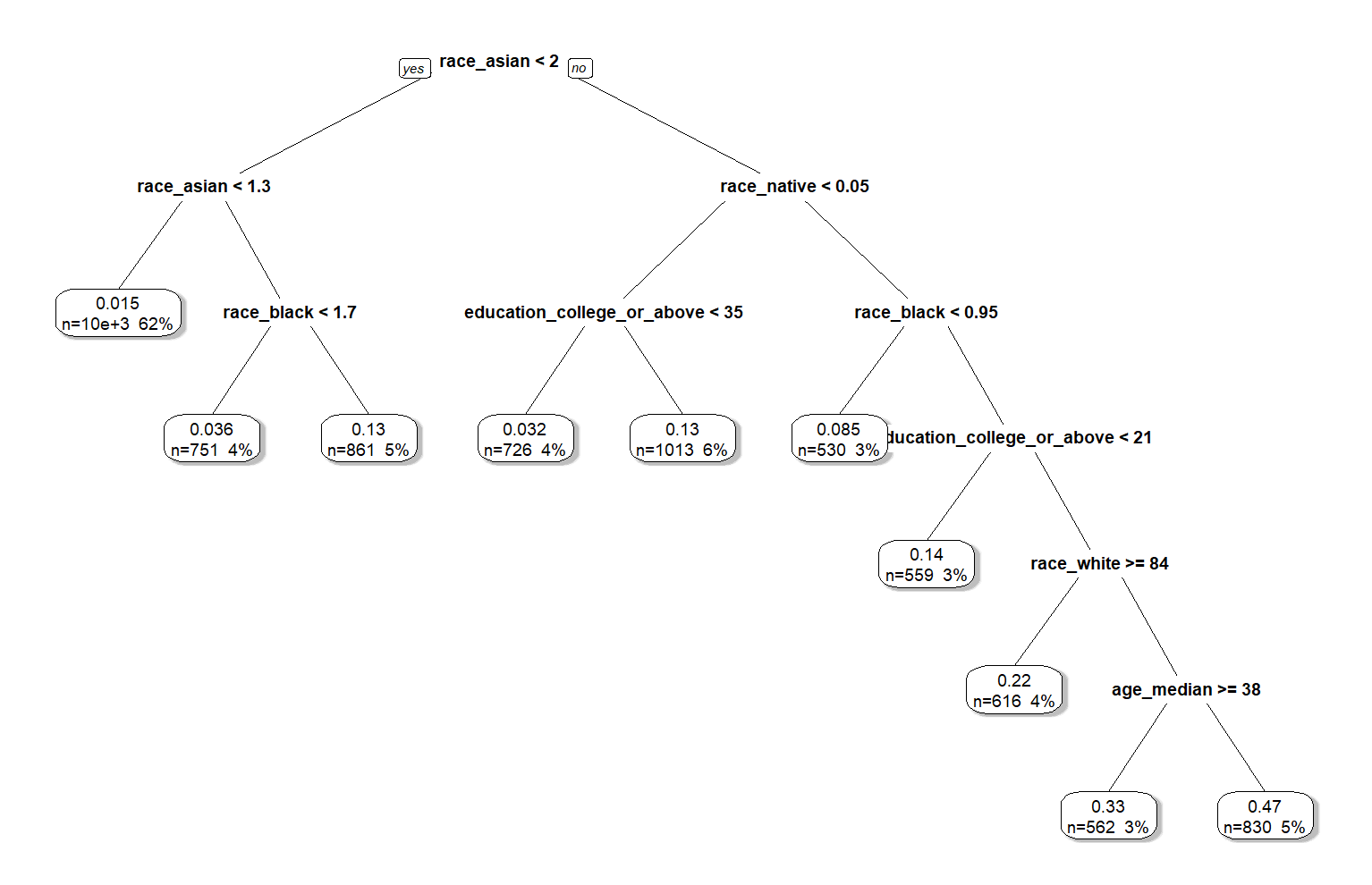
**K-Nearest Neighbor:**

In total, three K-Nearest Neighbor (KNN) models were run with different specifications for trainControl method and tuneLength. Our base model was run based on a training set of 80% of the data, repeatedcv, and tuneLength = ten, which happened to be our best model of the three. We changed repeatedcv to cv for the second model and used a tuneLength of twenty-five with regular cross-validation for the third model. These two had lower accuracies based on the KNN plot as well as the confusion matrices for each one. The first model was best with twenty-three nearest neighbors and that had a training accuracy of .94492, which is very high as well as a test confusion matrix accuracy of .9447.

**K-Means Clustering:**

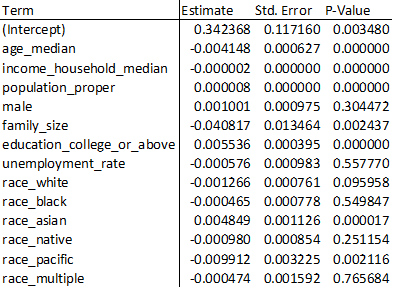
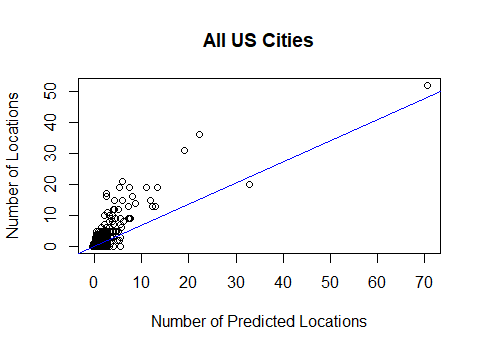
For K-Means clustering, the first approach was to figure out how many clusters we would need using the Weighted Sum of Squares (WSS) method, which left us with an elbow chart indicating five clusters is ideal. From here, we calculated the euclidean distance between the cities to find the center city in order to reduce the number of labels on our final plot, which was overrun with labels initially. The center cities were Stafford, OR; McCord, OK; Marlborough, MA; Camargo, KY; and Sunrise Manor, NV for each of their respective clusters. After this, we found a lot of outlier cities, so we winsorized them and clustered the cities. This did not end up doing much as we ended up with five clusters again and with only one changed center city, Camargo, KY became Oquaka, IL. For the first two sets of clustering, only two of the clusters had close to 50% of the cities with a Chipotle location with the other three clusters not having many in comparison by percent of the total. Only one of the cities switched clusters, otherwise there was no difference between winsorized and non-winsorized clusters. Finally, we wanted to normalize all of the variables used in the cluster analysis. We decided to use 4 clusters for our final K-Means cluster analysis because of the elbow plot. We again used euclidean distance between the cities to find the center city.

**Decision Tree Classifier:**

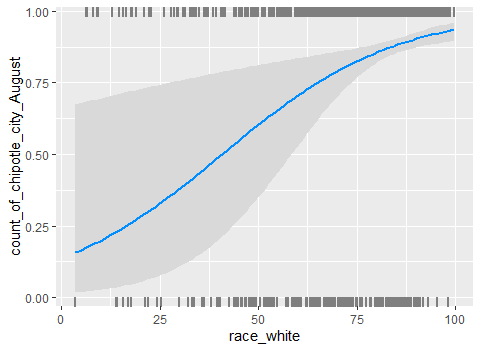
The decision tree classifier was one of the most important models that we ran because we wanted to see exactly what factors led to a chipotle opening in a certain city. Using the rpart package, we created a tree that is shown below:

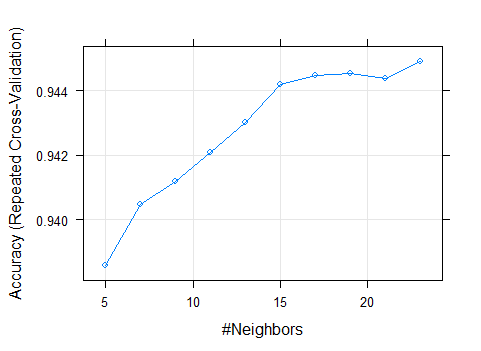
The tree shown had specifications of a complexity parameter of 0.003, a min bucket of 350, along with a minimum split of twenty-five. We felt that these parameters gave us the best visualization of what a chipotle executive goes through when deciding to open a new chipotle. Again, it shows that racial demographics, education, and age are important factors to consider when opening a Chipotle.

**Results:**

 Using the linear regression, we were able to see that all of the races are negative in relation to race\_other, which we would assume is mostly composed of Latinx people since they are not specified anywhere else in the data. Chipotle is Mexican food, so it would make some sort of sense that the restaurants are in cities with similar food cultures.

Usually, the model underpredicted large city locations as seen by the plot above. The average mean difference when we overpredicted was .224 locations. The average mean differences when we underpredicted the amount of Chipotles was .126 locations.

For our final logistic model, we decided to use all the variables except for population, since it skewed the data too much. This model performed much better than the other logistic regressions as there were only 183 incorrect predictions (fifteen false positives and 168 false negatives). For space limitations, we opted to not include that output. With this in mind, a graph of the logit prediction compared to the most statistically significant variable, race\_white, shows that as there are more white people in a city, the likelihood increases dramatically with a small error. On the other hand, there are some cities with a low race\_white percentage but a high likelihood, this likely comes from other factors that have positive coefficients, like race\_asian, race\_black, and age\_median.

 Our next method was K-Nearest Neighbor. Our best method was the result using repeated cross-validation with a tuneLength of ten. We used twenty-three neighbors in the final model, as represented by our plot below. This had a training accuracy of .9449. This means that for classification purposes, we looked at the twenty-three closest points. The test accuracy was very similar at .9447. The KNN method proved more efficient at predicting whether a Chipotle would open in a particular city than the logistic model.

The K-Means model shows certain groups of cities, plotted by the first and second dimensions of the dataset. Cluster 1 has relatively small populations (mean of 8000 people) but has a high median income of over $88,000. Because of this, we predict that these cities are not the best demographic to open a Chipotle. Cluster 2 also has small, very white populations with low educated people. Because of this we also predict these cities to not have Chipotles. Cluster 3 has the largest cities we have seen so far with household median income above the US average. They are still mostly white but have a higher population of minorities. Therefore, we predict that these cities are good places to open a Chipotle. Cluster 4 has very small populations with very low income. They are the most minority population and have a very high unemployment rate. These cities are not predicted to have a chipotle. Cluster 5 is the big city (NY and Los Angeles). They have a very high population and are predicted to have a chipotle (most likely more than one). Cluster 5 is however, our smaller cluster with under 400 cities.

**Chart

Description automatically generated** In our normalized K-Means model, we ended up with four clusters when we ran our clustering model again, which is different from before. Cluster 1 is by far the biggest with over 11,000 cities, Cluster 2 has 700, Cluster 3 has almost 6,000, and Cluster 4 has just under 2,000. Cluster 1 on average has a small population, where the mean is .032 after normalization and has the lowest ratio of Chipotles to cities, with only 2.59% of the cities having a location; these cities appear to be too small to have one. Cluster 2 has the most cities with locations and have a much higher population, seen by their mean of .6917 after normalization. These cities appear to be fairly diverse as well, with a lower proportion of white people compared to the US average (.6548 compared to .8284 for the US average after normalizing). Cluster 3 is set apart by the high incomes in their cities, with an average of .8688, which is very high; this cluster has 10.19% of its cities with a location, which is the second highest ratio of all four clusters. Cluster 4 is the first majority black cluster with its normalized mean as .4327 compared to the white population’s normalized mean being .3849.

For our last model we did a decision tree to try to model the thought process of Chipotle when selecting new locations. In the tree above, we see that the first branch looks at the Asian population of a city, followed by the Native population. This might not be the best model as these are split at such small percentages, but 81% of the cities are broken down in these first two steps, leaving only 19% of US cities to receive more consideration for a Chipotle location. This does not mean that there are no Chipotles in these cities, however, just a small number relative to the size of the branches.

**Conclusion**

In our analysis, we have come to the conclusion that the United States is a very diverse place, and it very difficult to come up with a perfect location for a Chipotle. Some cities might have a location for one reason or another. However, based on the results or our different models, we decided to look at one particular location where we think a Chipotle would thrive. Lowell, Massachusetts, a college town with a population of 111,346 would be a great fit for a Chipotle location. The city currently has none. We believe with their diverse, high educated population, we assume that there would be plenty of interest whether that the be from the student of UMass-Lowell or the rest of the community. Our model consistently shows that a more diverse population along with high education rates lead to a higher chance of there being a chipotle in that location. According to our final logistic regression, Lowell had a 96% chance of having a Chipotle. Also, it is in Cluster 2, where over 70% of cities in that cluster currently have a Chipotle. With all of this in mind, lots of other cities would be great fits for the next Chipotle location. Our preliminary research shows factors Chipotle corporate should consider when deciding on their next location.