Should I Stay or Should I Joe: Predicting Trader Joe's Presence with Economic and Demographic US Census Bureau Data

Lauren Sun

Fall 2024

Background

Trader Joe's is a much beloved California-native grocery store chain with over 500 locations in the United States. Its quirky store brand and unique marketing strategy has always catered to educated, middle-class consumers; Joe Coulombe, the founder, would often use the phrase "overeducated and underpaid" [source]. He envisioned a store that catered to university graduates who had "adventurous, health-oriented palates" but couldn't afford gourmet meals on a middle-class budget.

However, it has been noted that locations tend to pop up in increasingly gentrified neighborhoods. The effect has been observed with Starbucks. A study found that "every additional ten Starbucks reviews for a given ZIP code is associated with a 1.4 percent increase in home prices increases" [source]. Though there isn't an official definition or metric for gentrification, researchers tend to measure it with increases in home price and educated populations, and shifting demographics. From personal observation, when people talk about a neighborhood being gentrified, they mean an influx of wealthier white people.

Given that Trader Joe's seems to cater to a particular demographic and economic group, the goal is to predict whether a ZIP code has a Trader Joe's location using US Census Bureau about income, race, and education in that ZIP code. I also incorporate rent price increase data to see if there is a "gentrification" effect similar to that of Starbucks's. For the neighborhoods predicted by the model to have a Trader Joe's where there isn't one, it may be profitable to open one there.

Data

This Kaggle dataset from 2021 contains Trader Joe's 521 locations and their zip codes. Some zip codes have more than one location, so there are 505 unique zip codes with a Trader Joe's location.

Zillow published a dataset with its Rent Zestimates, which uses rental price data to estimate the price on all homes, including those not for rent, using their proprietary models. It contains price and price per square foot for May and December between 2010 and 2017 in 8,971 cities.

The demographic and economic data by ZIP code is accessible in the American Community Survey data collected by the US Census Bureau and is available at data.census.gov. I downloaded tables ACSDT5Y2017. B15002, which contains educational attainment data by gender, ACSST5Y2017. S1903, which contains income data by race, and ACSST5Y2017. S2406, which contains data about occupation industry for employed individuals. Each table has data for 33,122 ZIP codes in the US. Because there is only Zillow price data up to 2016, I use the American Community Survey data from 2017.

Preprocessing

The Zillow dataset is missing some values for 2010 and 2011, so I only consider the period from May 2012 to December 2016. Prices can be skewed depending on the typical square footage of homes in each city, so I only consider price per square foot. I calculate the rent increase by subtracting the May 2012 price per square foot from that of December 2016 for every city and call it `overall_diff` or just `diff` throughout the code.

However, though both the US Census Bureau and the Zillow datasets have cities, there are many duplicate city names like Monroe and Springfield, and not all ZIP codes are located near a city in the Zillow dataset.

To map each of the 33,122 ZIP codes in the American Community Survey to one of the 8971 cities in the Zillow data, I use this world cities dataset to obtain each city's latitude and longitude coordinates, then use this ZIP code dataset from simplemaps.com to obtain each ZIP code's coordinates, and assign each ZIP code its geographically closest city. Every ZIP code now has an attribute `overall_diff`, the increase in home price per square foot of the nearest city in the Zillow dataset.

As for the other predictors provided in the American Community Survey tables, the following are provided directly:

1. % White households
2. % Black households
3. % Indigenous households
4. % Asian households
5. % Hispanic / Latino households
6. % Male (mpct)
7. median age
8. % aged < 18 (au18)
9. % aged 18-24
10. % aged 25-29
11. % aged 30-34 (midage)
12. % aged > 65
13. median income
14. % aged over 16 employed (eo16)
15. % service jobs
16. % sales/office jobs
17. % labor jobs

The education table has educational attainment split by gender, so following are calculated as a weighted average using the % male column.

18. % highest education 4-year degree (bach)
19. % highest education HS diploma

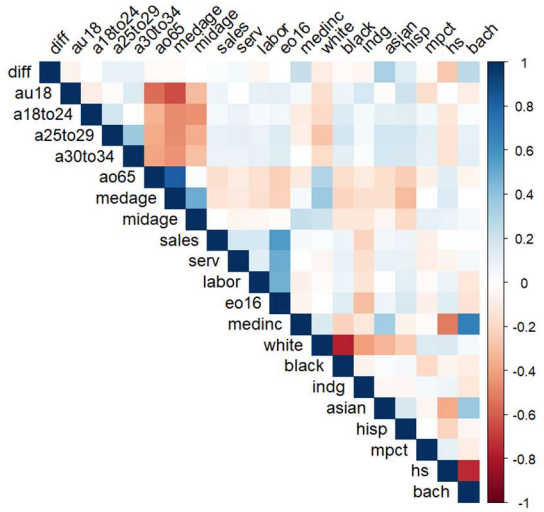
Lastly, the middle-aged population is grouped into one:

20. % aged 35-64

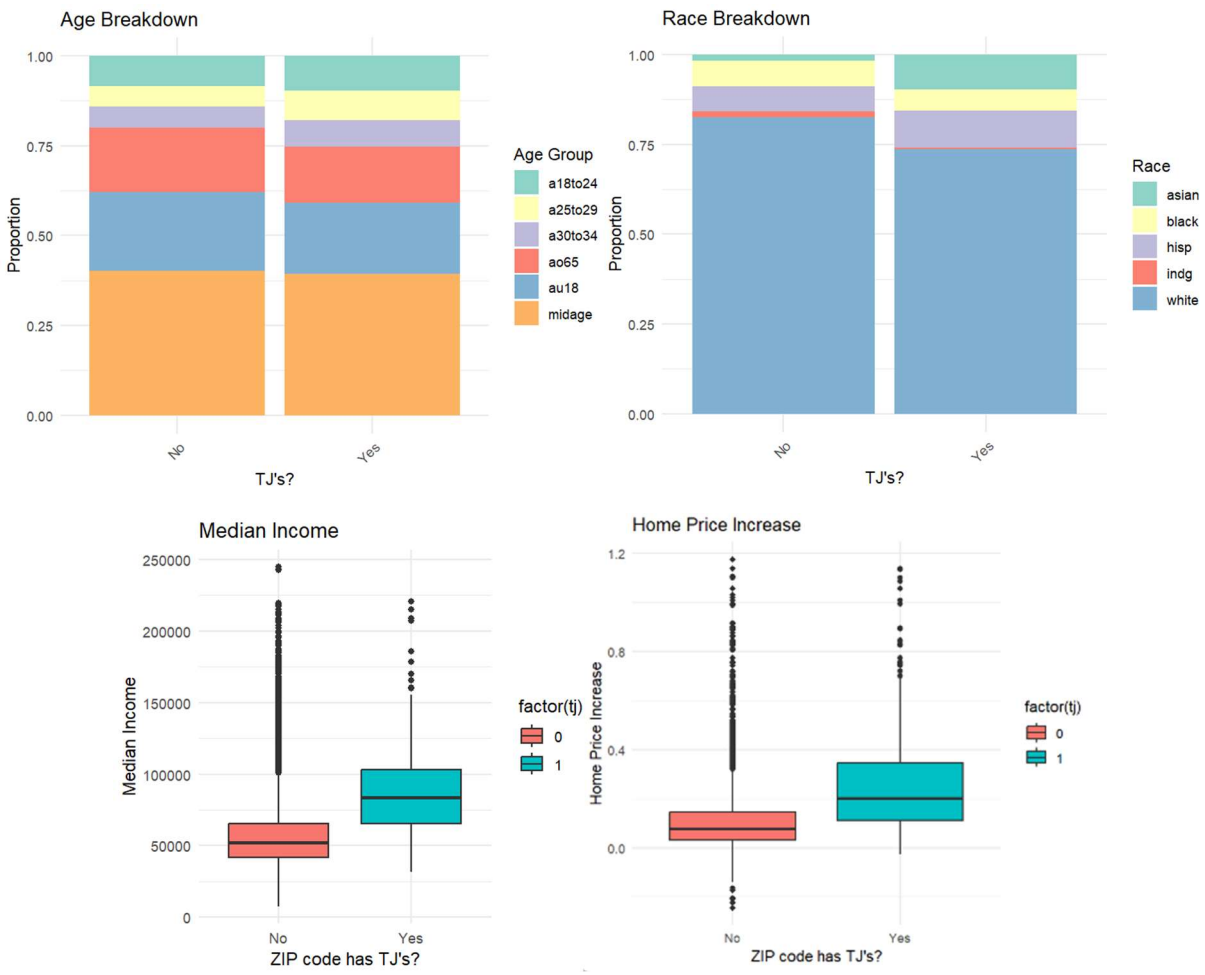
All percentages are listed between 0 to 100. The 21st predictor, is diff, as obtained above. The response variable, `tj`, is encoded as 1 if there is at least one Trader Joe's location in the ZIP code and 0 if there are none. After preprocessing, there are 499 positive instances and 28,817 negative instances.

Preliminary Data Analysis

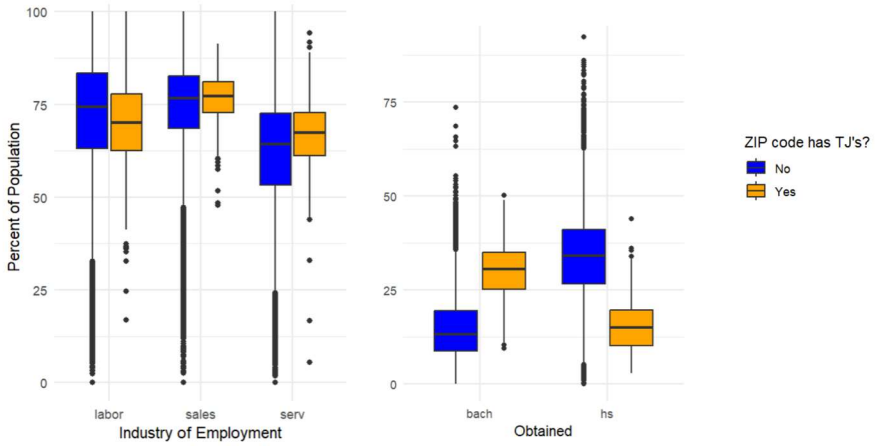
Most variables aren't that strongly correlated to the others, but the more salient ones include a strong positive relation between income and bachelor's degrees and a strong negative correlation between income and high school diploma. The most dramatic increases in home prices occurred in neighborhoods with more 4-year degree holders and Asians. White populations tend to be older than other groups. A higher white population was negatively associated with increase in home prices, which is surprisingly given that gentrification is used almost synonymously with an influx of white residents in some conversations. Other relationships can be seen in the correlation matrix.



I compare ZIP codes with and without Trader Joe's locations. The neighborhoods with a store location tend to have a slightly higher proportion of young adults between ages 18 and 34. Neighborhoods with Trader Joe's have a noticeably higher proportion of Asians and a lower proportion of white people.



Neighborhoods with a Trader Joe's location have higher median incomes and have seen greater increase in home price between 2012 and 2016. They also have a higher proportion of college-educated people and fewer residents with only a high school diploma. There's no noticeable difference in the breakdown of industry of employment.



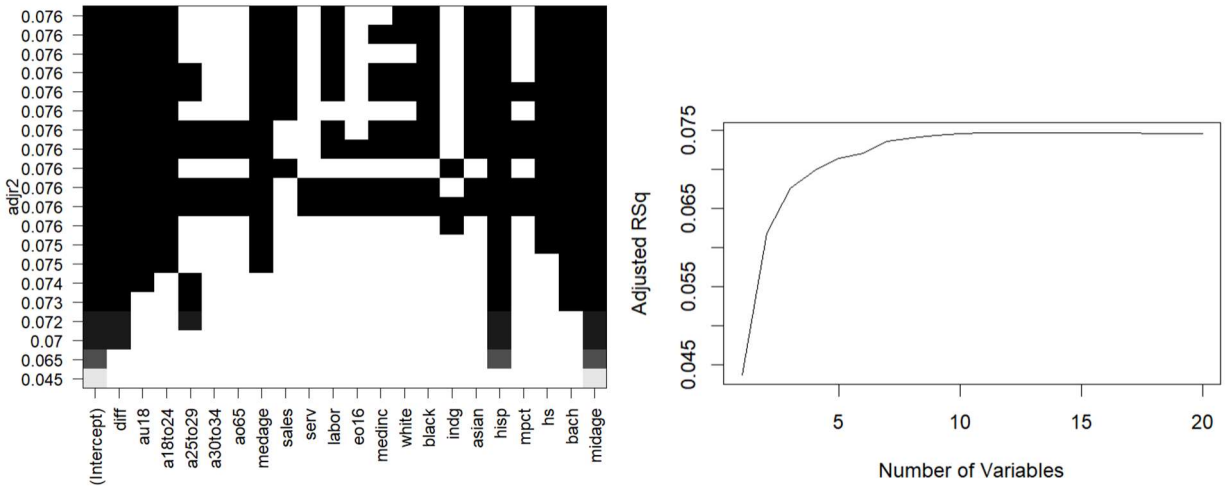
Models

As outlined in my proposal, I set out to try three different types of models for the binary classification task: a logistic regression model, a decision tree, and a neural network. I thought the logistic regression model should use 4 to 5 predictors, the decision tree about 12, and the neural network all 21 predictors. I wanted to see if more complex models really did better on this task. I chose logistic regression and decision trees because they are interpretable, and I want to know which predictors are most important. All models use the same training set with a 75/25 train/test split. There are 21988 training instances,

Because there aren't that many predictors, I conduct exhaustive best subset selection on the training set. The most important predictors, roughly, are bachelor's degree attainment, proportion of Asians, home price increase, and proportion in the 25-29 age group. The least useful predictors are those about industry of employment.

```
1 subsets of each size up to 20
Selection Algorithm: exhaustive
         diff au18 a18to24 a25to29 a30to34 ao65 medage midage sales serv labor eo16 medinc white black indg asian hisp mpct hs  bach
1  ( 1 ) " "  " "  " "     " "     " "     " "  " "    " "    " "   " "  " "   " "  " "    " "   " "   " "  " "   " "  " "  " "  " "
2  ( 1 ) " "  " "  " "     " "     " "     " "  " "    " "    " "   " "  " "   " "  " "    " "   " "   " "  "*"   " "  " "  " "  "*"
3  ( 1 ) "*"  " "  " "     " "     " "     " "  " "    " "    " "   " "  " "   " "  " "    " "   " "   " "  "*"   " "  " "  " "  "*"
4  ( 1 ) "*"  " "  " "     "*"     " "     " "  " "    " "    " "   " "  " "   " "  " "    " "   " "   " "  "*"   " "  " "  " "  "*"
5  ( 1 ) "*"  " "  " "     "*"     " "     " "  " "    " "    " "   " "  " "   " "  " "    " "   " "   " "  "*"   " "  " "  "*"  "*"
6  ( 1 ) "*"  "*"  " "     "*"     " "     " "  " "    " "    " "   " "  " "   " "  " "    " "   " "   " "  "*"   " "  " "  "*"  "*"
7  ( 1 ) "*"  "*"  "*"     " "     " "     " "  "*"    " "    " "   " "  " "   " "  " "    " "   " "   " "  "*"   " "  " "  "*"  "*"
8  ( 1 ) "*"  "*"  "*"     " "     " "     " "  "*"    " "    " "   " "  " "   " "  " "    " "   " "   " "  "*"   " "  "*"  "*"  "*"
9  ( 1 ) "*"  "*"  "*"     " "     " "     " "  "*"    " "    " "   " "  " "   " "  " "    " "   " "   " "  "*"   " "  "*"  "*"  "*"
10  ( 1 ) "*"  "*"  "*"     " "     " "     " "  "*"    "*"    " "   " "  " "   " "  " "    " "   " "   " "  "*"   " "  "*"  "*"  "*"
11  ( 1 ) "*"  "*"  "*"     " "     " "     " "  "*"    "*"    " "   " "  " "   " "  "*"    " "   " "   " "  "*"   " "  "*"  "*"  "*"
12  ( 1 ) "*"  "*"  "*"     " "     " "     " "  "*"    "*"    " "   " "  " "   " "  "*"    " "   " "   " "  "*"   " "  "*"  "*"  "*"
13  ( 1 ) "*"  "*"  "*"     " "     " "     " "  "*"    "*"    " "   " "  " "   "*"  "*"    " "   " "   " "  "*"   " "  "*"  "*"  "*"
14  ( 1 ) "*"  "*"  "*"     " "     " "     " "  "*"    "*"    " "   "*"  " "   "*"  "*"    "*"   " "   " "  "*"   " "  "*"  "*"  "*"
15  ( 1 ) "*"  "*"  "*"     " "     " "     " "  "*"    "*"    " "   "*"  " "   "*"  "*"    " "   " "   "*"  "*"   " "  "*"  "*"  "*"
16  ( 1 ) "*"  "*"  "*"     " "     " "     " "  "*"    "*"    " "   "*"  " "   "*"  "*"    "*"   " "   "*"  "*"   " "  "*"  "*"  "*"
17  ( 1 ) "*"  "*"  "*"     "*"     " "     " "  "*"    "*"    " "   "*"  " "   "*"  "*"    "*"   "*"   "*"  "*"   " "  "*"  "*"  "*"
18  ( 1 ) "*"  "*"  "*"     "*"     " "     " "  "*"    "*"    " "   "*"  "*"   "*"  "*"    "*"   "*"   "*"  "*"   " "  "*"  "*"  "*"
19  ( 1 ) "*"  "*"  "*"     "*"     " "     " "  "*"    "*"    " "   "*"  "*"   "*"  "*"    "*"   "*"   "*"  "*"   "*"  "*"  "*"  "*"
20  ( 1 ) "*"  "*"  "*"     "*"     "*"     "*"  "*"    " "    "*"   "*"  "*"   "*"  "*"    "*"   "*"   "*"  "*"   "*"  "*"  "*"  "*"
```

The models with the best r^2 value are shown below.



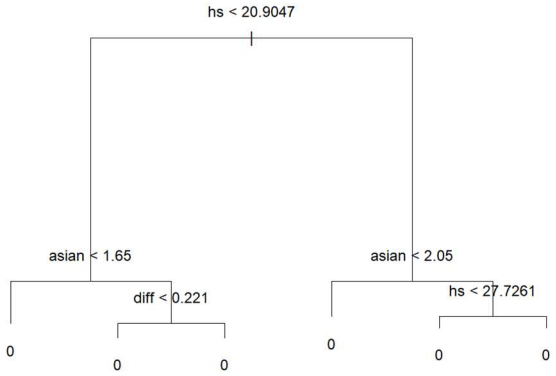
Most of the best r^2 can be obtained with 5 predictors and adding more than about 7 predictors doesn't do much to increase the model's accuracy, so I stick with my original decision to keep the 5 best predictors. The model's coefficients are listed below.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.266e-02  6.217e-03  -8.471  < 2e-16 ***
diff         9.210e-02  6.070e-03  15.172  < 2e-16 ***
a25to29      2.404e-03  2.677e-04   8.979  < 2e-16 ***
hisp         2.040e-04  5.912e-05   3.451 0.000559 ***
bach         2.509e-03  9.232e-05  27.175  < 2e-16 ***
midage       1.039e-04  1.340e-04   0.775 0.438193
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
          Reference
Prediction    0    1
         0 7210  106
         1    8    5

               Accuracy : 0.9844
```

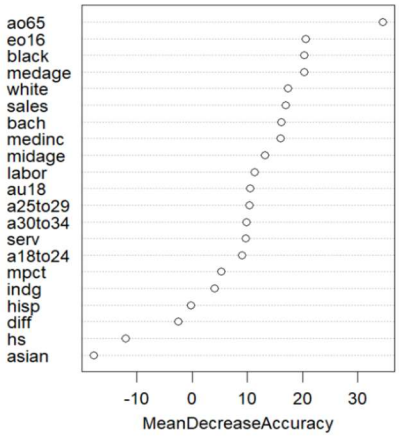
Because the age, Hispanic, and bachelor's degree variables are all percentages, it can be said that the age and bachelor's degree variables have 10x more weight than the Hispanic variable in the model. The model correctly predicted 98.44%, though it's only marginally better than the baseline 98.30% if it had guessed all 0's.

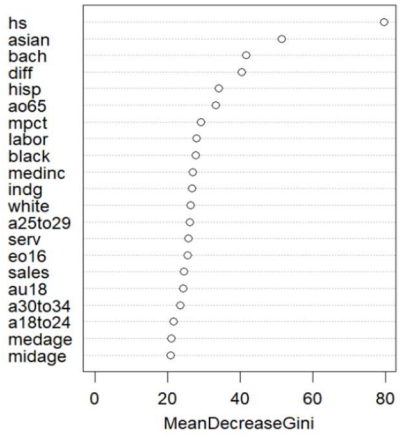
I then try a decision tree model. Unfortunately, the tree simply learns to predict all instances as 0 because there are so many more of them.



I instead try a random forest of 500 trees, limiting each tree to 12 predictors. Its accuracy of 98.51% is slightly better than the baseline. At least it does not learn to predict everything as a 0.

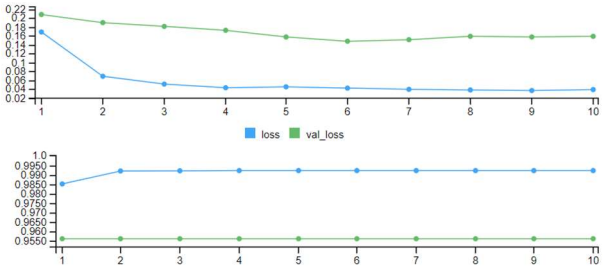
```
          Reference
Prediction    0    1
         0 7217  108
         1    1    3

               Accuracy : 0.9851
                 95% CI : (0.9821, 0.9878)
    No Information Rate : 0.9849
    P-Value [Acc > NIR] : 0.449

                  Kappa : 0.0512

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.99986
            Specificity : 0.02703
         Pos Pred Value : 0.98526
         Neg Pred Value : 0.75000
             Prevalence : 0.98485
         Detection Rate : 0.98472
   Detection Prevalence : 0.99945
      Balanced Accuracy : 0.51344

       'Positive' Class : 0
```





Finally, I try a neural network. It's a multilayer network whose structure is pictured below. It has three hidden layers with ReLU activation functions. Its accuracy seems to converge after 10 epochs.

```
input <- layer_input(shape = c(21))
output <- input %>%
  layer_dense(units = 16, activation = "relu") %>%
  layer_dropout(rate = 0.4) %>%
  layer_dense(units = 8, activation = "relu") %>%
  layer_dropout(rate = 0.3) %>%
  layer_dense(units = 8, activation = "relu") %>%
  layer_dropout(rate = 0.3) %>%
  layer_dense(units = 2, activation = "softmax")
```



However, it only learns to predict all 0's! I was surprised that even a relatively complex model such as a neural network would learn that.

Balanced Dataset

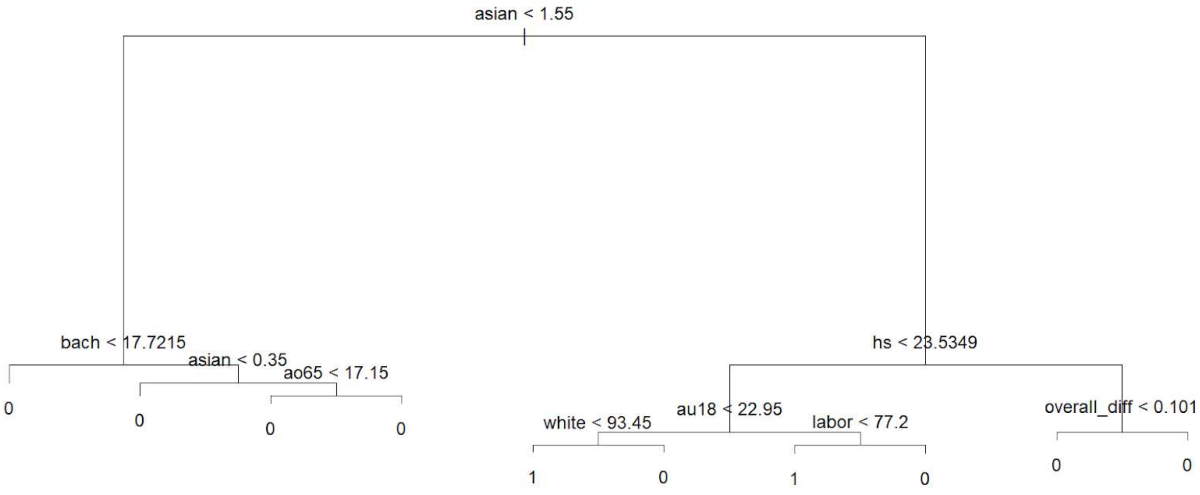
To deal with the problem of severe class imbalance, I only sampled 1,000 of the 28,817 negative instances to create a more balanced dataset. The new dataset now has 1,499 data points, 499 of which are positive observations and 1,000 of which are negative. The baseline is now 66.7%.

Best subset selection yielded an output similar to when it was trained on every single one of the negative instances. Its best model with 5 predictors is also quite similar. The performance of the logistic regression using this balanced dataset is 89.3%, which is much better than the baseline. Note that, when I train the logistic regression model using all 21 predictors, it only predicts one more instance correctly than the 5-predictor model.

```
Coefficients:
(Intercept)      a25to29         hisp
 -0.6811283    0.1681260    0.0257066

         hs        bach      midage
 -0.1288011    0.0717491    0.0002936
```

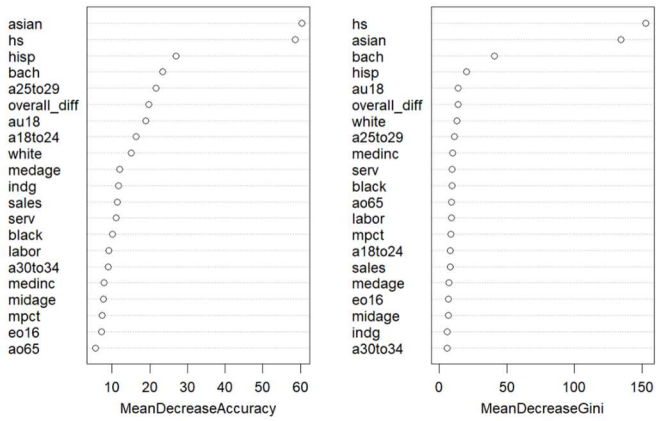
```
          Reference
Prediction   0   1
         0 227  17
         1  23 107

               Accuracy : 0.893
```

The decision tree also performed better. It actually learned to classify some instances with 1's. These are neighborhoods with a higher proportion of Asians and a lower proportion of white people, people whose highest degree is a high-school diploma, and blue-collar workers.

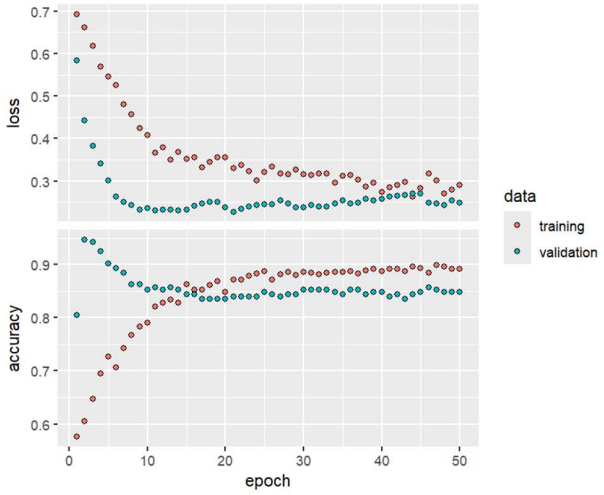


Curiously, the random forest of 500 trees, each with at most 12 predictors, did not perform that much better with an accuracy of 90.64%. The predictors it found most important by far are the proportion of Asians and the proportion with only high school diplomas.

```
          Reference
Prediction   0   1
         0 233  18
         1  17 106

               Accuracy : 0.9064
                 95% CI : (0.8723, 0.9339)
    No Information Rate : 0.6684
    P-Value [Acc > NIR] : <2e-16
```



Lastly, I train the neural network on the balanced dataset. Its structure is the same as the one trained the full training data. Its accuracy did not converge after 10 epochs as it did when trained on the full training data, so I increased it to 50.




```
          Reference
Prediction   0   1
         0 219   5
         1  31 119

               Accuracy : 0.9037
```

However, the neural network's accuracy was 90.37%, which is lower than what I expected for a more complex model.

Discussion

The models' improved performance when trained on a much smaller but better-balanced dataset illustrates that it may not be better to include more observations if it will drastically skew the training data. Once the datasets are balanced, however, the simpler models like the logistic regression and decision tree perform just as well as more complex models like the random forest and neural network. The problem of predicting whether or not there is a Trader Joe's in a

particular neighborhood given some information about its residents just not that difficult of a problem, and it doesn't require that complex of a model. In accordance with Occam's razor and in the interest of interpretability, we can resort to a simple model to answer this question.

As for what we can learn about Trader Joe's choices of locations from this analysis, it seems that the best predictors are those we may have guessed just looking at the preliminary analysis. Neighborhoods with higher income, more educated residents, greater recent home price increases, and higher Asian populations are more likely to have a Trader Joe's location. However, I did not provide the models with a wide variety of features to work with. For example, one model might select the proportion of the Asian population as a predictor, while another might choose the Hispanic proportion. Similarly, one model might focus on the 18-24 age group, while another opts for the 25-29 age group. Since the models were given many age-related variables and many more race variables, it's almost trivial to observe that age or race emerge as a predictor—it reflects the limited scope of the input features rather than any deeper insight.

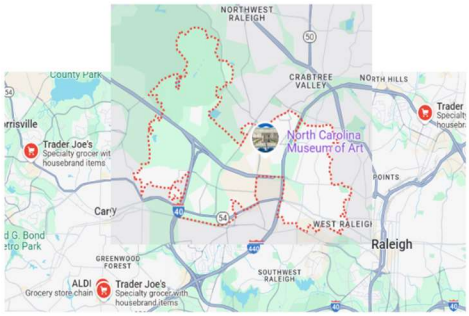
However, of the models used the industry of employment as a predictor, but we already have seen that in the preliminary analysis.

Interestingly though, as hinted in the preliminary analysis and shown in the decision tree, a higher proportion of white people is negatively associated with the presence of Trader Joe's. Given that both of these are sometimes spoken about in the context of gentrification, I'm surprised that the association is negative. If anything, the analysis shows that the indicators of gentrification, higher home prices and an influx of educated higher earners, can be more strongly associated with a higher proportion of Asians.

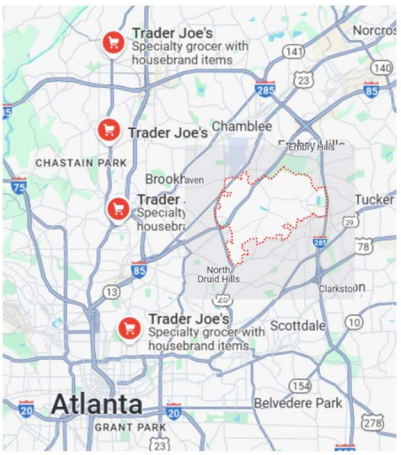
This analysis does offer another useful bonus—a list of potential new Trader Joe's locations. These ZIP codes were predicted by the logistic regression, decision tree, random forest, and neural network to have a Trader Joe's location when it in fact does not.

zip	city	overall_diff	medage	medinc	asian	white	black	tj
<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<fct>
37204	Nashville	0.256	30.1	76250	2	81.8	14	0
27607	Raleigh	0.1	24	74786	4.9	81.9	9.7	0
55405	Minneapolis	0.198	31.9	56365	3.8	78.4	11.8	0
79119	Amarillo	0.096	36.8	87565	2.2	93.4	1.1	0
19301	Paoli	0.148	44.4	93211	6	89.3	4.2	0
30345	Atlanta	0.416	34.1	71914	6.2	64.1	25.3	0

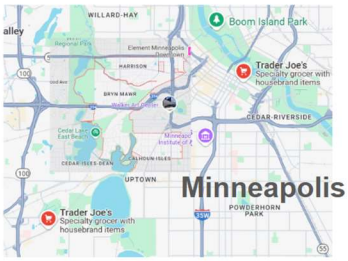
When the ZIP codes are plotted (in red) on a map with nearby Trader Joe's, these constitute pretty reasonable suggestions for new store locations.


NORTHWEST RALEIGH
CRABTREE VALLEY
NORTH HILLS
North Carolina Museum of Art
POINTS
Cary
WEST RALEIGH
Raleigh
GREENWOOD FOREST
SOUTHWEST RALEIGH
Trader Joe's
Specialty grocer with housebrand items
ALDI
Grocery store chain


Valley Forge National Historical Park
King of Prussia
Trader Joe's
Devault
Berwyn
Paoli
Malvern
Wayne
PA


Trader Joe's
Specialty grocer with housebrand items
Norcross
Trader Joe's
Chamblee
Embry Hills
CHASTAIN PARK
Brookhaven
Tucker
North Druid Hills
Clarkston
Scottdale
Atlanta
Belvedere Park
GRANT PARK


Shelby Park
Nashville
THE NATIONS
WEST NASHVILLE
THE GULCH
MIDTOWN
MUSIC ROW
SYLVAN PARK
Trader Joe's
Specialty grocer with housebrand items
12 SOUTH
SOUTH NASHVILLE
Berry Hill
GREEN HILLS
Belle Meade
Oak Hill


WILLARD-HAY
Boom Island Park
HARRISON
BRYN MAWR
Trader Joe's
Specialty grocer with housebrand items
CEDAR-RIVERSIDE
CEDAR-ISLES-DEAN
UPTOWN
Minneapolis
POWDERHORN PARK
Trader Joe's
Specialty grocer with housebrand items