Jasper Evans, Krehl Kasayan, Paul Kiefer, Pablo Suarez
Milestone 3
INST 737
5/6/2024

Introduction: Additional Data Preparation

After unsuccessful attempts to contact the US Census Bureau for guidance on the Census variable code numbers for the 2020 Census, our team opted to move forward with the ACS data used in previous reports.

While ACS data is collected over five-year periods rather than in a single year, in this case, it may offer some advantages – namely because 20202 Census figures are notoriously unreliable.

Our attempts correct errors in the Census tract number field of the HMDA mortgage data — which could have provided information on loan denial rates and refinancing rates by tract — were also unsuccessful; the HMDA dataset's Census tract column includes numerous partial tract numbers, and we can't extrapolate the missing digits to correctly identify the tract without additional information.

We did, however, create a version of our core dataset with an additional column called foreclosure_hi_med_low — an ordinal variable that divides tracts into three groups based on the dependent variable foreclosure_pc_2020 — to allow us to test a classification SVM.

As a reminder, foreclosure_pc_2020 represents the number of reported foreclosures in a census tract between 2011-2023 divided by the estimated population of the tract in the 2016-2020 ACS. That may seem like an imperfect point of comparison, and in some ways, it is: ideally, we would work with foreclosures per capita in the most recent year for which we have both foreclosure and population data available.

Unfortunately, the Prince George's County dataset that records foreclosures by address contains very few records dated after 2020 – which may mean some foreclosures have been omitted – so we would have hardly any samples if we were to limit ourselves to recent years.

Instead, we consider all foreclosures between the end of the Great Recession – defined in our case as 2011, though that is up for debate – and the end of our dataset in 2011. This gives us a larger sample of tracts to assess, and we include variables measuring characteristics of each tract at various points in that decade to ensure we are considering changes in a tract's risks as part of our analysis.

Question 1A: SVM

We tested two SVM models with linear kernels: One using a regression approach and the continuous dependent variable foreclosure_pc_2020, and another using a classification approach and the ordinal (i.e. multiclass) dependent variable foreclosure_hi_med_low.

The output of the regression SVM is as follows:

```
[1] "Mean Squared Error (MSE): 384.843777031783"
[1] "Root Mean Squared Error (RMSE): 19.6174355365777"
[1] "R-squared (R²): 0.664195090677985"
```

In other words, only approximately 66.42% of the variance in foreclosure_pc_2020 is explained by the 40 independent variables in our model. This is far from ideal, but it is a start.

To test the statistical significance of the RSME, we run a hypothesis test comparing it to the standard deviation of foreclosure_pc_2020 at an alpha of 0.05. At that significance level, the RSME is not statistically significant; the p-value is 0.76, meaning there is not enough evidence to reject the null hypothesis that there is no difference between the RMSE and the standard deviation of foreclosure_pc_2020. In other words, the RSME and standard deviation of foreclosure_pc_2020 are similar enough that any observed differences between them are likely random fluctuations.

The output of the classification SVM – using foreclosure_hi_med_low as the output variable — is as follows:

```
Confusion Matrix and Statistics

          Reference
Prediction 1 2 3
         1 9 5 0
         2 2 4 4
         3 0 2 7

Overall Statistics

               Accuracy : 0.6061
                 95% CI : (0.4214, 0.7709)
    No Information Rate : 0.3333
    P-Value [Acc > NIR] : 0.001199

                  Kappa : 0.4091

 Mcnemar's Test P-Value : NA

Statistics by Class:
```

```
                     Class: 1 Class: 2 Class: 3
Sensitivity            0.8182   0.3636   0.6364
Specificity            0.7727   0.7273   0.9091
Pos Pred Value         0.6429   0.4000   0.7778
Neg Pred Value         0.8947   0.6957   0.8333
Prevalence             0.3333   0.3333   0.3333
Detection Rate         0.2727   0.1212   0.2121
Detection Prevalence   0.4242   0.3030   0.2727
Balanced Accuracy      0.7955   0.5455   0.7727

Precision:
 0.8181818 0.3636364 0.6363636
Recall:
 0.6428571 0.4 0.7777778
Specificity:
 0.6666667 0.6363636
F1 Measure:
 0.72 0.3809524 0.7
```

In other words, the model accurately classifies the foreclosure quantile of a tract 60.61% of the time.

The model classifies tracts in the first (low foreclosure rate per capita) and third (high foreclosure rate per capita) quantiles of tracts, but it is significantly less accurate when classifying tracts in the second quantile. Its negative predictions for the second quantile are more accurate than its positive predictions.

Question 1B: Non-linear Kernels

We applied two non-linear kernels to each of our SVMs:

Radial Basis Function (RBF) Kernel:

The **regression SVM with the RBF kernel** returned an RMSE of 21.2206779506785 – slightly larger than the RMSE of the model when computed with a linear kernel.

As above, we compared the RMSE of this model to the standard deviation of foreclosure_pc_2020 using a hypothesis test; the resulting p-value, 0.7701436, indicates that there is not enough evidence to reject the null hypothesis that there is no difference between the RMSE and the standard deviation of foreclosure_pc_2020

That said, the model still performed relatively worse than the regression model with the linear kernel.

The **classification SVM with the RBF kernel** returns the following output:

```
Confusion Matrix and Statistics

          Reference
Prediction 1 2 3
         1 7 5 0
         2 3 4 5
         3 1 2 6

Overall Statistics

               Accuracy : 0.5152
                 95% CI : (0.3354, 0.692)
    No Information Rate : 0.3333
    P-Value [Acc > NIR] : 0.02348

                  Kappa : 0.2727

 Mcnemar's Test P-Value : 0.42586

Statistics by Class:

                     Class: 1 Class: 2 Class: 3
Sensitivity            0.6364   0.3636   0.5455
Specificity            0.7727   0.6364   0.8636
Pos Pred Value         0.5833   0.3333   0.6667
Neg Pred Value         0.8095   0.6667   0.7917
Prevalence             0.3333   0.3333   0.3333
Detection Rate         0.2121   0.1212   0.1818
Detection Prevalence   0.3636   0.3636   0.2727
Balanced Accuracy      0.7045   0.5000   0.7045

Precision:
 0.6363636 0.3636364 0.5454545
Recall:
 0.5833333 0.3333333 0.6666667
Specificity:
 0.6666667 0.5454545
F1 Measure:
 0.6086957 0.3478261 0.6
```

This model is less accurate overall and for each individual quantile.

Polynomial Kernel:

The **regression SVM model with the polynomial kernel** returned an RMSE of 27.0167112718583 — larger than the RMSE of both the model computed with a linear kernel and the model computed with an RBF kernel.

As above, we compared the RMSE of this model to the standard deviation of foreclosure_pc_2020 using a hypothesis test; the resulting p-value, 0.7091711, indicates that there is not enough evidence to reject the null hypothesis that there is no difference between the RMSE and the standard deviation of foreclosure_pc_2020

That said, the model still performed relatively worse than both the regression model with the linear kernel and the regression model with the RBF kernel.

The **classification SVM with the polynomial kernel** returns the following output:

```
Confusion Matrix and Statistics

          Reference
Prediction 1 2 3
         1 6 3 0
         2 3 7 6
         3 2 1 5

Overall Statistics
                                          
               Accuracy : 0.5455          
                 95% CI : (0.3635, 0.7189)
    No Information Rate : 0.3333          
    P-Value [Acc > NIR] : 0.009726        
                                          
                  Kappa : 0.3182          
                                          
 Mcnemar's Test P-Value : 0.134428        

Statistics by Class:

                     Class: 1 Class: 2 Class: 3
```

```
Sensitivity            0.5455   0.6364   0.4545
Specificity            0.8636   0.5909   0.8636
Pos Pred Value         0.6667   0.4375   0.6250
Neg Pred Value         0.7917   0.7647   0.7600
Prevalence             0.3333   0.3333   0.3333
Detection Rate         0.1818   0.2121   0.1515
Detection Prevalence   0.2727   0.4848   0.2424
Balanced Accuracy      0.7045   0.6136   0.6591

Precision:
 0.5454545 0.6363636 0.4545455
Recall:
 0.6666667 0.4375 0.625
Specificity:
 0.875 0.4545455
F1 Measure:
 0.6 0.5185185 0.5263158
```

While this model is slightly more accurate than the model with the RBF kernel, it is less accurate in all quantiles than the model with the linear kernel.

Question 2: Neural Networks

After scaling our 40 independent variables within the 0-1 range, we defined the following Neural Network configurations to test:

```
# Define configurations to test
configurations <- list(
  list(hidden_layers = 1, neurons_per_layer = c(10), activation_function =
"logistic"),
  list(hidden_layers = 2, neurons_per_layer = c(10, 5), activation_function
= "tanh"),
  list(hidden_layers = 1, neurons_per_layer = c(5), activation_function =
"logistic"),
  list(hidden_layers = 2, neurons_per_layer = c(5, 3), activation_function
= "tanh"),
  list(hidden_layers = 3, neurons_per_layer = c(8, 6, 4),
activation_function = "logistic"),
  list(hidden_layers = 3, neurons_per_layer = c(12, 8, 4),
activation_function = "tanh"),
  list(hidden_layers = 4, neurons_per_layer = c(10, 8, 6, 4),
activation_function = "logistic"),
  list(hidden_layers = 4, neurons_per_layer = c(12, 10, 6, 3),
activation_function = "tanh"),
  list(hidden_layers = 5, neurons_per_layer = c(15, 12, 10, 8, 6),
activation_function = "logistic"),
  list(hidden_layers = 5, neurons_per_layer = c(20, 15, 12, 8, 5),
activation_function = "tanh")
)
```

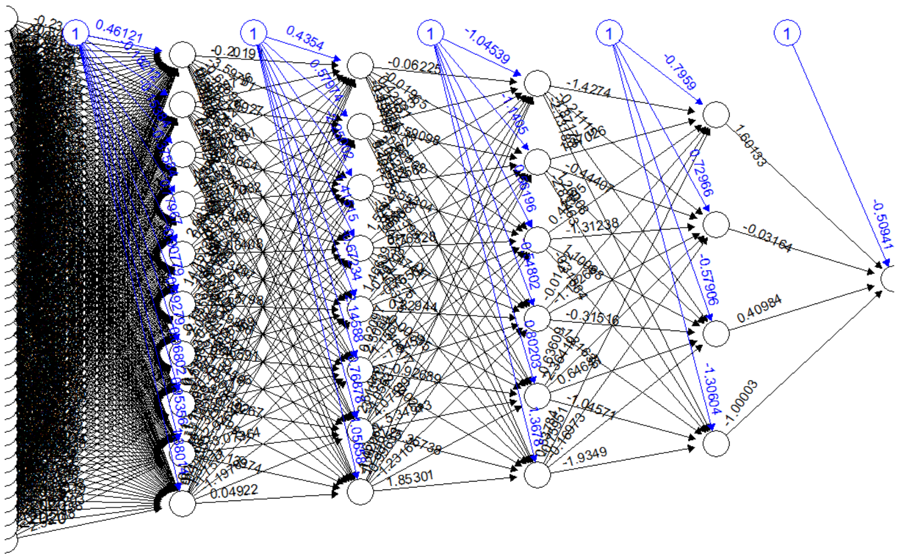
Those configurations return the following results:

```
Configuration: 1
RMSE: 0.2076627
Activation Function: logistic
Hidden Layers: 1
Neurons per Layer: 10

Configuration: 2
RMSE: 0.3605905
Activation Function: tanh
Hidden Layers: 2
Neurons per Layer: 10 5

Configuration: 3
RMSE: 0.3651585
Activation Function: logistic
Hidden Layers: 1
Neurons per Layer: 5

Configuration: 4
RMSE: 0.2237902
Activation Function: tanh
Hidden Layers: 2
Neurons per Layer: 5 3

Configuration: 5
RMSE: 0.1936369
Activation Function: logistic
Hidden Layers: 3
Neurons per Layer: 8 6 4

Configuration: 6
```

```
RMSE: 0.1968819
Activation Function: tanh
Hidden Layers: 3
Neurons per Layer: 12 8 4

Configuration: 7
RMSE: 0.1530879
Activation Function: logistic
Hidden Layers: 4
Neurons per Layer: 10 8 6 4

Configuration: 8
RMSE: 0.2548988
Activation Function: tanh
Hidden Layers: 4
Neurons per Layer: 12 10 6 3

Configuration: 9
RMSE: 0.174261
Activation Function: logistic
Hidden Layers: 5
Neurons per Layer: 15 12 10 8 6

Configuration: 10
RMSE: 0.4945652
Activation Function: tanh
Hidden Layers: 5
Neurons per Layer: 20 15 12 8 5
```

Based on those results, the seventh configuration — with four hidden layers — performed the best; in fact, it produced an RMSE lower than any of our SVMs.

While far from beautiful (given the number of independent variables), that model looks like this when plotted:



Question 3: Clustering

Hierarchical Clustering

After plotting our dendrogram and drawing our line, we determined that eight clusters (shown below) were the optimal number for our dataset. Using the cutree function, we saw the following tract distribution by cluster: 85, 27, 21, 4, 16, 14, 3, and 3. Among the variables with observable and distinguishable volatile variance were median income by tract for 2020 and 2010, post-2010 foreclosures, owner-occupied units in 2020, percentage of units built pre-1960, and the mortgaged_2010, 2015 and 2020 columns. Another variable of note was the pct_0_bed column, in which we observed zero values for the seventh cluster.

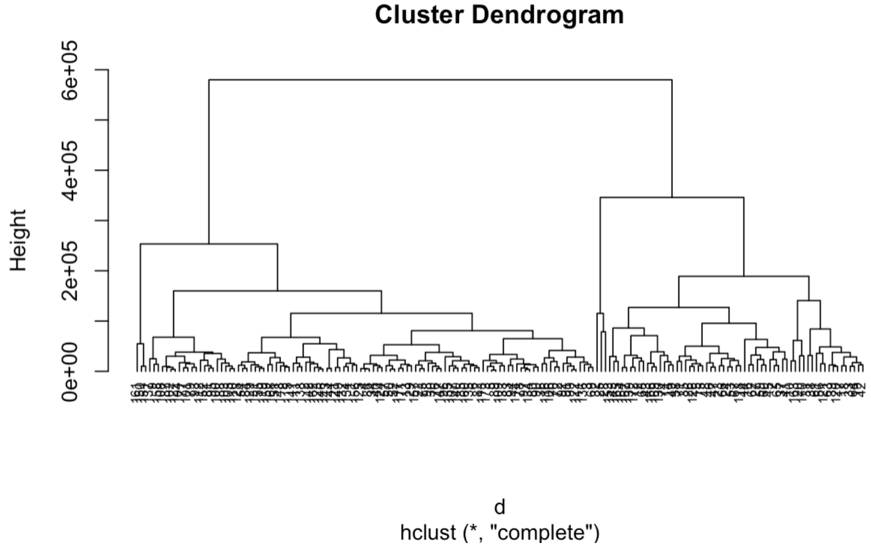
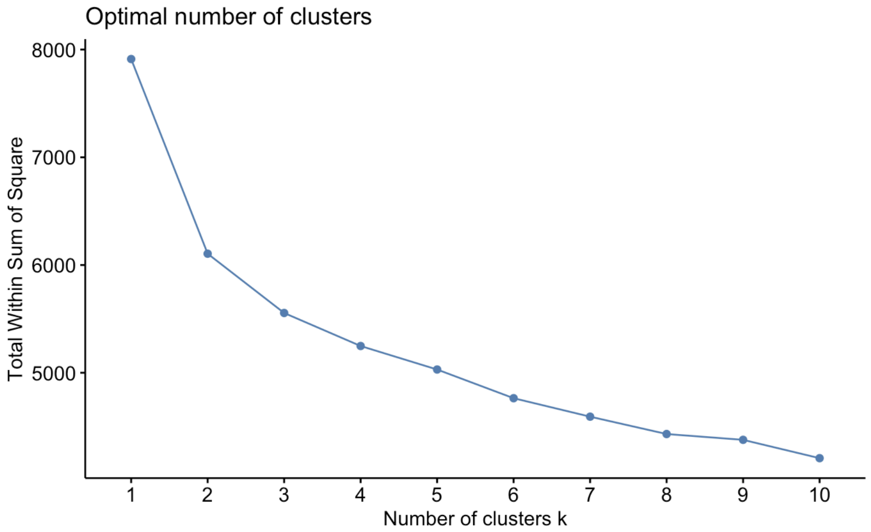
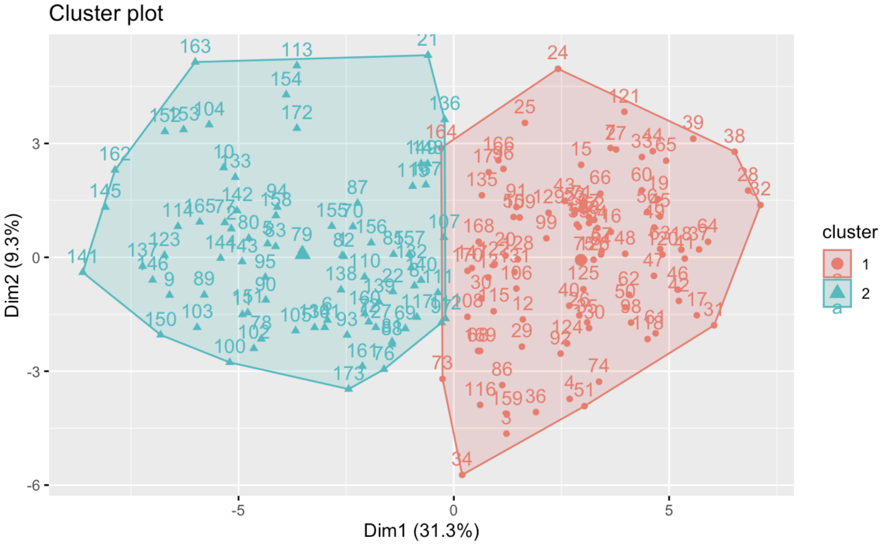


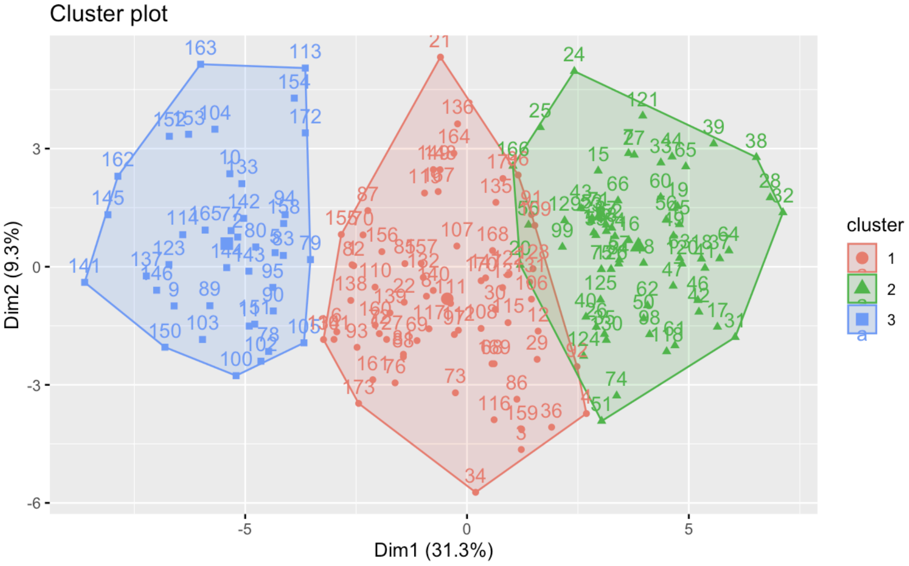
Figure: Hierarchical cluster dendrogram.

Partitional Clustering

For this step, we conducted two k-means clustering analyses—one for two clusters and another for three clusters. Our elbow graph indicated that two clusters, in this case, were the optimal number of clusters. A comparison of our plotted two and three-cluster graphs showed that there was almost no observable overlap in the two-cluster version, while the three-cluster version displayed a noticeable overlap between clusters one and two. In the two-cluster version, we observed a distribution of 94 to 79, while the three-cluster version saw a 69-54-40 distribution. The within-cluster sum of squares for the two-cluster version was 3,420 for cluster one and 2,685 for cluster two, indicating that cluster two was considered “tighter.” The three-cluster version saw within-cluster sums of squares with 2,473 for cluster one, 1,907 for cluster two, and 1,175 for cluster three. For both versions, the total sum of squares was 7,912, indicating that the variance in the data was consistent across both analyses.


Optimal number of clusters
Total Within Sum of Square
8000
7000
6000
5000
1
2
3
4
5
6
7
8
9
10
Number of clusters k


Cluster plot
Dim2 (9.3%)
3
0
-3
-6
-5
0
5
Dim1 (31.3%)
cluster
1
2

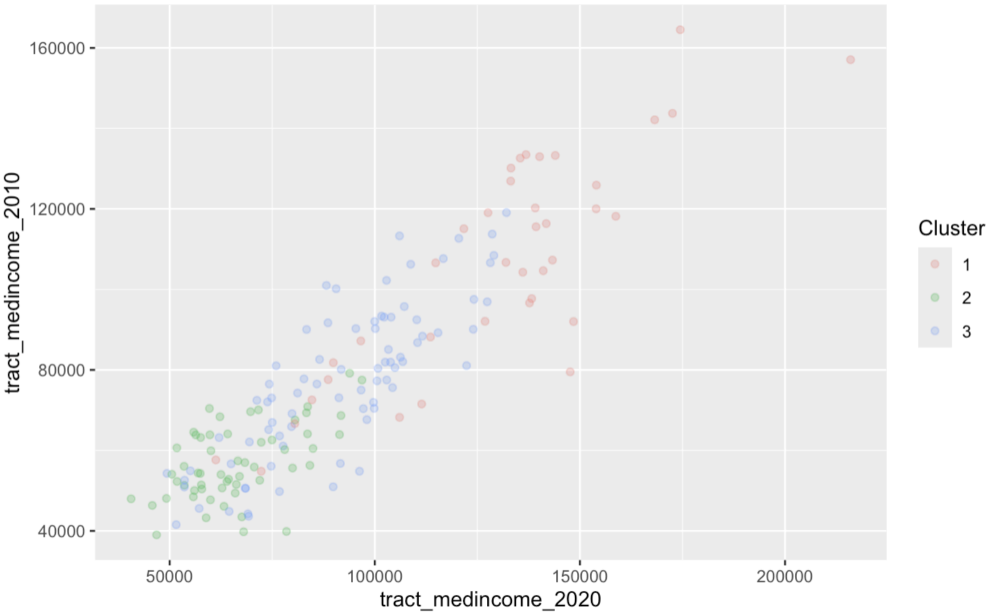

Cluster plot
Dim2 (9.3%)
3
0
-3
-6
-5
0
5
Dim1 (31.3%)
cluster
1
2
3

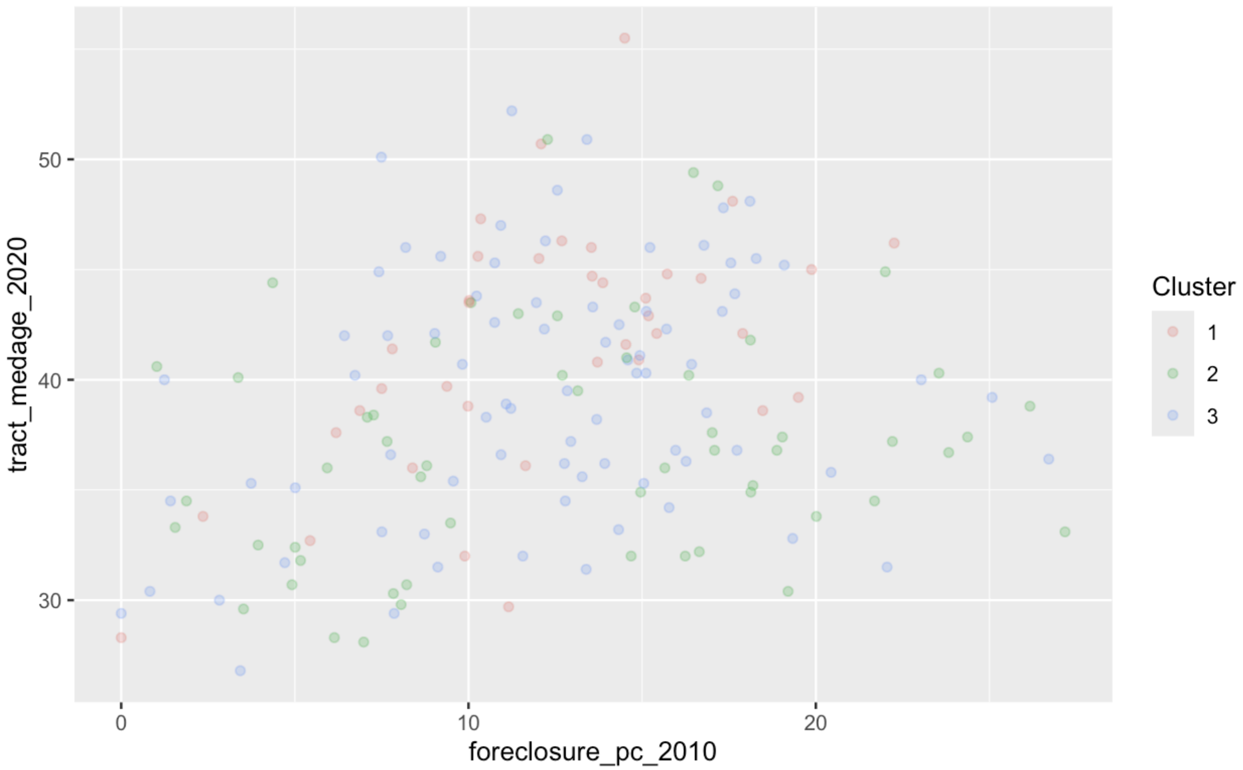
Spectral Clustering

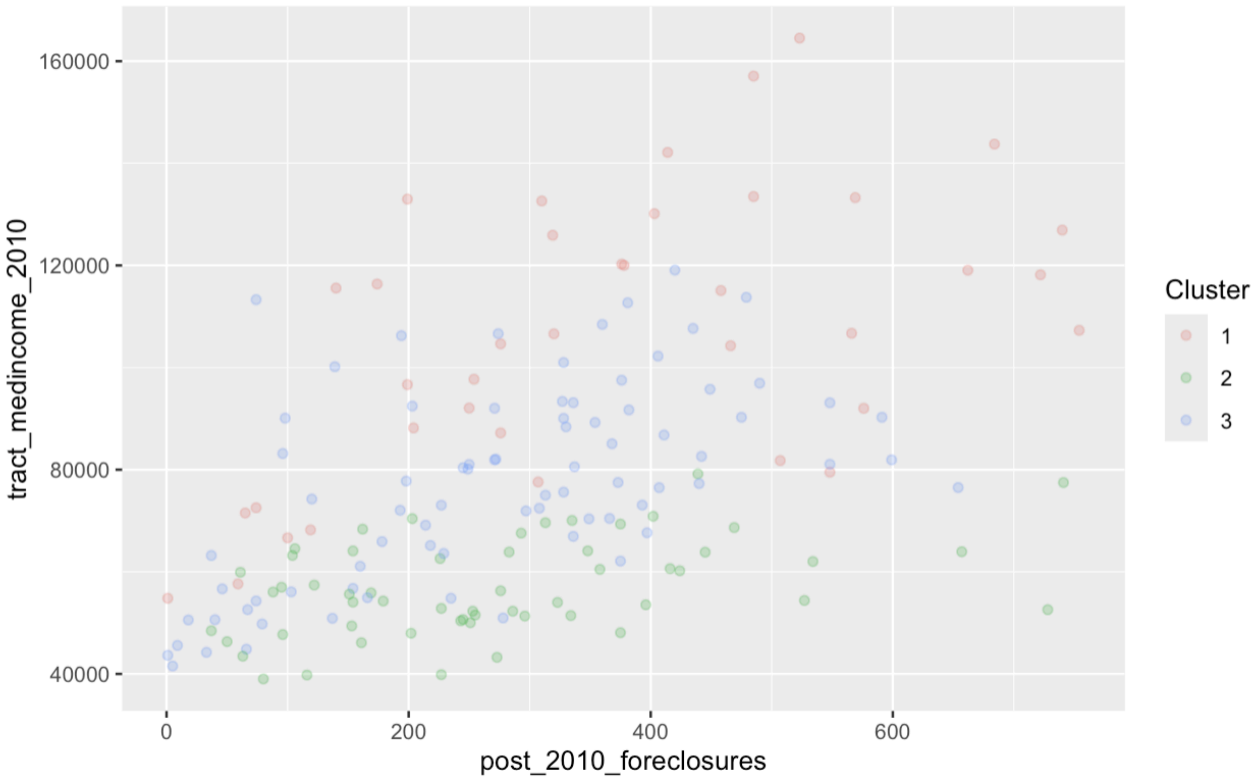
We opted for spectral clustering as our third technique. We began by using the k-means method. After testing various values, we chose to set k as 3 because the groups seemed the most distinctive with that value. We pulled the tested variables for this method from our previous clustering techniques and our linear regression model in milestone 2. We then plotted these variables to determine which factors had the most impact on identifying which cluster group each tract belonged to.

We first plotted two variables from our dataset—median income from 2010 and median income from 2020. There's a visible relationship between the incomes. However, because they are the same variable, it doesn't necessarily mean that these are the variables that had a large impact on the clustering groups. It's difficult to determine which factors had the most impact because there is a significant number of variables and a 2D plot does not enable us to plot all of them at once. We also saw a correlation between these variables, which is to be expected because they are essentially the same at different points in time.

In the following two plots, we looked at these variables—tract median age from 2020, foreclosure per capita in 2010, tract median income from 2010, and post-2010 foreclosures. Based on the variables from milestone 2, we wanted to see how they'd look with the clustering groups. The second plot does not yield any valuable information for us to draw any conclusions from. The third plot shows a semblance of a relationship, but drawing conclusions is difficult. It's also difficult to recommend spectral clustering as an optimal technique for our dataset, given the sheer number of variables we possess.




tract_medage_2020
50
40
30
0
10
20
foreclosure_pc_2010
Cluster
1
2
3


tract_medincome_2010
160000
120000
80000
40000
0
200
400
600
post_2010_foreclosures
Cluster
1
2
3

Model Comparison

The next step in our process was to take the models we have developed and trained, the SVM, Random forest,, Neural Network, and Decision tree and first see the accuracies of each model when compared to one another and then see what would be considered as our “best” model.

Unset

```
#Crossvalidation method for resampling
ctrl <- trainControl(method = "cv", number = 10)
# Train the model with trControl

train_data$foreclosure_pc_2020_binary <- ifelse(train_data$foreclosure_pc_2020
< 0.5, 0, 1)
train_data$foreclosure_pc_2020_binary <-
as.factor(train_data$foreclosure_pc_2020_binary)

train_data <- subset(train_data, select = -foreclosure_pc_2020)

# Train the SVM model
svm_model <- train(foreclosure_pc_2020_binary ~ ., data = train_data, kernel =
"radial",trControl = ctrl)
random_forest <- train(foreclosure_pc_2020_binary ~ ., data = train_data,method
= "rf", trControl = ctrl)
neural_network <- train( foreclosure_pc_2020_binary ~ ., data =
train_data,linear.output = TRUE, act.fct = "logistic", trControl = ctrl)
decision_tree <- train(foreclosure_pc_2020_binary ~ ., data = train_data,
trControl = ctrl)
# Pass the control object to resamples

results <- resamples(list(RF = random_forest, SVM = svm_model, NN =
neural_network, DT = decision_tree))
```

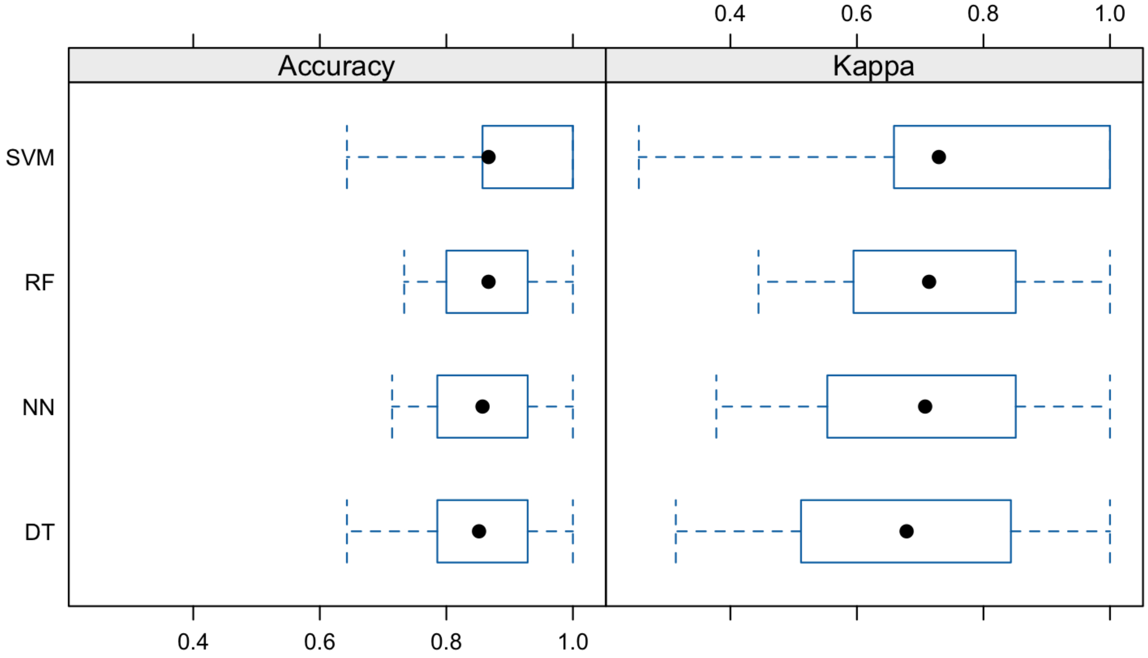
Using the Caret Package learned in lecture, we imported our models into the environment running the code above. We then set a trainControl object to train our models for 10 folds. During development we found that the model comparisons would not run due to the foreclosure_pc_2020 variable not being in classifying terms, so the solution to that was to turn it into a binary variable by classifying any value below 0.5 as 0 and above as 1. This then generates the following output when the resample() function was made.

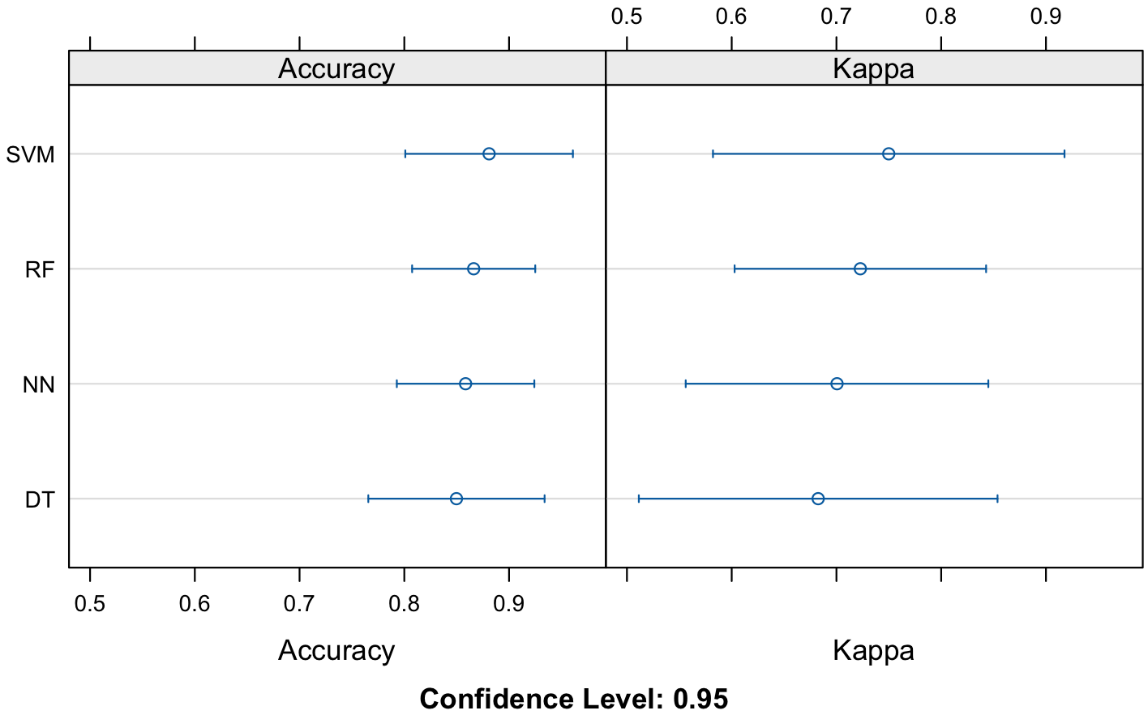
Unset

```
Call:
summary.resamples(object = results)

Models: RF, SVM, NN, DT
Number of resamples: 10

Accuracy
         Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
RF  0.7333333 0.8115385 0.8619048 0.8718681 0.9230769 1.0000000    0
SVM 0.7333333 0.8008242 0.8928571 0.8643956 0.9285714 0.9333333    0
NN  0.7142857 0.8571429 0.8571429 0.8585714 0.8642857 1.0000000    0
DT  0.7142857 0.8115385 0.8928571 0.8639194 0.9285714 0.9333333    0

Kappa
         Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
RF  0.4444444 0.5907336 0.7225877 0.7314360 0.8403223 1.0000000    0
SVM 0.4117647 0.5900983 0.7826748 0.7163659 0.8510638 0.8571429    0
NN  0.4166667 0.6661247 0.7083333 0.7042221 0.7312030 1.0000000    0
DT  0.3777778 0.6146959 0.7796986 0.7130490 0.8556231 0.8648649    0
```

Looking at the median column we can see that the SVM and Decision tree had the same median so they were tied in terms of accuracy with a 0.8928571 median. Next was the random forest with a 0.8619048 median. Followed by the Neural network with a 0.8571429 median.

Below is that same data visualized into a box and whisker plot (top) and a dot plot (bottom)





Feature Selection

To test feature selection methods, we chose two of our worse-performing models – our regression SVM with a linear kernel and our Naive Bayes model – and one of our best-performing models: our most predictive multivariate regression model.

If feature selection can improve the predictiveness of the less-predictive original models, we have more satisfactory options to choose from. If feature selection can bring an already successful model even closer to the ideal, we have improved on our most reliable tool.

Regression SVM with filter feature selection:

For our regression SVM with a linear kernel, we used a straightforward filter to identify the most predictive independent variables — similar to the process we used in Milestone 2 to choose the most predictive sets of variables.

After identifying the ten most predictive independent variables, we tested their VIF values to identify multicollinearity, filtered out all but one of the variables with a VIF value over 10, and ran the SVM with a new set of only 6 independent variables. This is the output:

```
[1] "Mean Squared Error (MSE): 110.529305958683"
[1] "Root Mean Squared Error (RMSE): 10.5132918707074"
[1] "R-squared (R²): 0.93073093386578"
```

In other words, the filter dramatically improved the predictiveness of our SVM. The R-squared value increased from 0.6642 to 0.9307, and the RMSE is the lowest of any we have seen thus far.

As expected, these are the most predictive variables:

```
"mortgaged_2010" "foreclosure_pc_2010" "tract_medage_2020" "pct_1_bd"
"poverty_2010"  "mortgaged_2020"
```

Naive Bayes with wrapper feature selection:

For our Naive Bayes model, we used the stepwise function to compute combinations of independent variables. In this case, our output variable is foreclosure_quantile, which divides the dataset into five groups based on foreclosure_pc_2020.

After seven steps, we reached an AIC of 68.97 with the following combination of variables:

```
nhwhite_2020 + foreclosure_pc_2010 + tract_medage_2020 +
    avg_bed + mortgaged_2010 + pct_built_pre_1960
```

While this group of variables includes some of our most predictive – namely foreclosure_pc_2010 and tract_medage_2020 – its confusion matrix and summary statistics indicate it is actually less predictive than our original Naive Bayes model.

```
Reference
Prediction 1 2 3 4 5
         1 6 1 1 0 0
         2 3 3 2 0 1
         3 0 1 3 1 1
         4 0 1 1 1 6
         5 0 0 0 1 2

Overall Statistics

               Accuracy : 0.4286
                 95% CI : (0.2632, 0.6065)
    No Information Rate : 0.2857
    P-Value [Acc > NIR] : 0.04981

                  Kappa : 0.3


Statistics by Class:
                     Class: 1 Class: 2 Class: 3 Class: 4 Class: 5
Sensitivity            0.6667  0.50000  0.42857  0.33333  0.20000
Specificity            0.9231  0.79310  0.89286  0.75000  0.96000
Pos Pred Value         0.7500  0.33333  0.50000  0.11111  0.66667
Neg Pred Value         0.8889  0.88462  0.86207  0.92308  0.75000
Prevalence             0.2571  0.17143  0.20000  0.08571  0.28571
Detection Rate         0.1714  0.08571  0.08571  0.02857  0.05714
Detection Prevalence   0.2286  0.25714  0.17143  0.25714  0.08571
Balanced Accuracy      0.7949  0.64655  0.66071  0.54167  0.58000
```

For comparison, these are the confusion matrix and summary statistics for our original Naive Bayes model:

```
Reference
Prediction 1 2 3 4 5
         1 6 2 1 0 0
         2 3 2 0 0 1
         3 0 1 4 1 2
         4 0 0 1 0 3
         5 0 1 1 2 4

Overall Statistics

               Accuracy : 0.4571
                 95% CI : (0.2883, 0.6335)
    No Information Rate : 0.2857
    P-Value [Acc > NIR] : 0.02302

                  Kappa : 0.3073

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4 Class: 5
Sensitivity            0.6667  0.33333   0.5714  0.00000   0.4000
Specificity            0.8846  0.86207   0.8571  0.87500   0.8400
Pos Pred Value         0.6667  0.33333   0.5000  0.00000   0.5000
Neg Pred Value         0.8846  0.86207   0.8889  0.90323   0.7778
Prevalence             0.2571  0.17143   0.2000  0.08571   0.2857
Detection Rate         0.1714  0.05714   0.1143  0.00000   0.1143
Detection Prevalence   0.2571  0.17143   0.2286  0.11429   0.2286
Balanced Accuracy      0.7756  0.59770   0.7143  0.43750   0.6200
```

In other words, the accuracy of the model slightly decreased after introducing feature selection. The original model was nominally better at predicting tracts in class 5 — meaning tracts with the highest rates of foreclosures since 2011.

Multivariate Linear Regression with embedded feature selection:

The third iteration of our multivariate linear regression model from Milestone 2 was among our most successful.

That model used a subset of 15 of our independent variables chosen to avoid multicollinearity. For this test, we used Ridge regression to embed feature selection and included all 40 of our independent variables.

To measure the predictiveness of our original model, we used two metrics:

```
Correlation between predicted and real values: 0.7934256
Mean Squared Error: 437.2731
```

After updating the model for feature selection and the full list of independent variables, its predictions became more accurate:

```
Correlation between predicted and real values: 0.8419894
Mean Squared Error: 271.8911
```

While the 0.05 change in the correlation between predicted and real values is notable, the substantial decrease in the MSE is even more pronounced. In short, the Ridge regression was more effective at feature selection than we were when using more rudimentary methods in Milestone 2.

Ethics

This semester, Team 2's has focused its efforts toward answering the research question, "Is it possible to forecast the per capita rate of foreclosure for a census tract in Prince George's County, Maryland, using identified independent variables?" This question aimed to uncover environmental variables, local to prince George's county, which may precipitate concentrations of foreclosures within specific Prince George's County census tracts.

The intent behind investigating a research question such as the one above was to determine whether it was possible to use predictive modeling to identify at-risk areas within Prince George's County. Ultimately, these at-risk areas could be reasonable candidates for state/federal government, non-profit, or private sector intervention. At the same time, a tandem goal of Team 2's efforts was to identify characteristics of tracts determined to be at risk for foreclosure. Successful and reliable identification of at-risk census tracts in Prince George's County would be a critical step for decision makers in formulating commensurate measures to combat foreclosure.

To this end, Team 2 has leveraged a combination of Federal and State government-level data and has used open-source tools/services to add additional fidelity to the original datasets Team 2 opted to use for its investigation. Ethical considerations accompanying Team 2's methods will be discussed in this section of the report through the lenses of Data provenance, created/enriched variables, research question implications, ideological argument avoidance, participant identification avoidance, and fair representation of the target population.

Provenance of Datasets Leveraged

1. *Prince George's County Foreclosure Dataset*

a. This dataset was downloaded by Team 2 from the Prince Georges County Open data portal. This dataset was devoid of foreclosed homeowner biographical information (names, birth dates, ages); however, included address-level information on foreclosed properties in Prince George's County, MD. Further, this dataset included property tax account numbers, property IDs, zip codes, as well as dates the properties listed were reported foreclosed.

b. The Prince Georges County open data portal did not note the source of the above dataset or accompanying collection methods. To this end, Team 2 contacted the Prince Georges County open data team via email to obtain clarity on the attribution of the dataset and what methods may have been used to gather the information included in its foreclosure dataset. Team 2 has not received contact from the Prince George'ss County open data team related to the above questions.

c. Based on public documentation of the foreclosure process in Prince George's County, MD it is most likely that information used by Team 2 (the county's public foreclosure dataset) was collected via a form required to be completed by lenders registering properties slated for foreclosure. On this form, lenders are required to enter address information, property tax account numbers, and zip codes, and biographical details about the owner(s) of the property. This information appears to match information present in the Prince Georges County open data portal, foreclosure dataset. It appeared the foreclosure dataset leveraged by Team 2 had been sanitized of property owner biographical details and contact numbers prior to publication. This assessment has been made based on the difference in variables present in the foreclosure dataset and fields present in the lender form noted above. It appears this sanitization was conducted to protect the privacy of the homeowners/borrower's undergoing foreclosure; however, based on addresses present in the dataset and dates the foreclosures were entered into the dataset, it may be possible to identify an individual owner after using open source tools/search methods.

2. *US Census Data, Tract Level: American Community Survey (ACS)*

a. Team 2 leveraged data from the US census Bureau's American Community Survey (ACS) specific to Prince George's County. The ACS is an annual survey conducted across the US that collects information on socio-economic and demographic characteristics related to the US population each year. The results of these surveys are reported by Census in five-year increments. From this data source, Team 2 queried variables related to housing characteristics in Prince George's County tracts such as, year of construction, number of bedrooms, percentage of owner occupied homes, etc.

b. The US Census Bureau collects ACS data via the mailing of surveys to a random sample of addresses in a defined geographic location. According to information on the Census Bureau's website, the Bureau ensures the sampling will result in representative coverage of a target population/area. For surveys that are responded to, the Census Bureau will conduct follow up via mail, phone and in-person visits to resolve any discrepancies; these follow-up efforts are additionally undertaken in the pursuit of obtaining responses to surveys that have not been returned. Participant responses to Census Bureau surveys are federally mandated. The US census Bureau offers extensive materials online about how ACS data is used after its collection by businesses, educational institutions, journalists, nongovernmental organizations as well as state and federal-level organizations.

c. The US Census bureau anonymizes it's datasets, generally, via stripping names, addresses, or other features which may be attributed to an individual. Information about individuals is aggregated and sorted into demographic variables and presented at the neighborhood level. The US Census bureau takes its disclosure avoidance responsibilities quite seriously and does not release information to the public that reveals the identities of survey respondents. In some cases, Census employs cryptographic techniques, to modify statistics to further obfuscate participant identities while retaining data accuracy.

Variables created and/or enriched by Team 2.

1. Generation of Census tract Identification Numbers

a. In this operation, Team 2 used a service called Geocodio, which is an open-source geocoding service. Geocodio offers users the ability to upload datasets containing addresses and enrich them via the addition of corresponding Census county-level FIPS codes. Team 2 used this service specifically to attribute census county-level FIPS codes to addresses present in the Prince Georges County Foreclosure data set.

b. At the conclusion of Geocodio's enrichment, Team 2 conducted manual checks of the geocoding work to ensure its accuracy. Team 2 moved forward after satisfactory cleaning had been completed. This step later allowed Team 2 to conduct joins in R that resulted in the core dataset used to investigate Team 2's overarching research question.

c. Geocodio is a SOC 2 annually audited service. SOC 2 includes a five-point trust criteria used to assess a business's ability to manage customer data securely, with confidentiality, privacy, processing availability, and accessibility in mind. Geocodio's reputation was a factor in Team 2's decision to leverage the service. Team 2 remained aware that additional quality checks would be required when using a service such as Geocodio to fill missing data. Further, such quality checks were essential in ensuring fair representation of the population's data at the center of Team 2's analytic efforts.

2. Tract Bedroom Numbers

   a. Team 2 undertook efforts to create/transform a variable for the average number of bedrooms per home at the tract level. This creation entailed using arithmetic functions (multiplication and division) as well as the ROUND function in R studio to create a column of tract-level average bedroom numbers. This new column was derived from a variable queried by Team 2 from the US Census Bureau's ACS data noted under item 1 within this section of the report.

3. Average Construction Year

   a. Team 2 created/transformed a variable holding values for averages of year ranges. As the original Census dataset leveraged included year date ranges, Team 2 needed to assign values to these ranges to make them easily manipulable within the data model constructed by team 2. Team two leveraged an arithmetic formula similar in composition to the one noted in the *Tract Bedroom Numbers* section above.

Implications of the research question Team 2 has attempted to answer.

1. Team 2's research question aimed to assess the foreclosure susceptibility of certain census tracts in the locale of Prince George's County. Opposed to investigating foreclosure risk on a national scale (US domestic), Team 2 has opted to focus its efforts within one of Maryland's most populous counties. Scoping Team 2's activities to a localized area does carry some ethical implications in a general sense. Some of which could be:

   a. <u>Smaller Geography</u>: Smaller geographies tend to entail smaller populations. Within smaller populations it may be easier to identify individuals based on unique combinations of demographic characteristics specific to the area being examined.

   b. <u>Data Sparsity</u>: Smaller samples tend to lack diversity, and samples which are limited in this respect have the potential to limit the learning ability of the model.

   c. <u>Generalization Issues</u>: (stemming from Data sparsity) It can be difficult to generalize the results of models trained on localized data due to insufficient sample sizes. In these scenarios, overfitting stemming from the existence of noise/outliers within limited samples tends to have a greater effect.

   d. <u>Contextual Nuances</u>: Models may learn context specific patterns that do not apply to larger samples.

   e. <u>Cultural Bias</u>: Stemming from a lack of diversity in localized samples; models may reinforce existing stereotypes, cultural biases, or underrepresentation within a given community/environment. All these factors would influence a model's ability to make fair predictions.

2. Team 2 acknowledged for this classifier to be leveraged by a decisionmaker to identify communities/census tracts ideal for intervention, it would be necessary for the classifier's training data sample size to be larger to minimize the ethical and computational concerns noted above. A more robust training data set may yield results with higher reliability/accuracy, in turn, maximize the effectiveness of intervention efforts.

Avoiding Ideological Arguments.

1. Team 2's investigation of its research question involved the handling and transformation of demographic and socio-economic data which had been gathered from the public through collection efforts undertaken by state and federal government agencies. To this end, Team 2 has remained contentious of how its data is labeled, so as to avoid downstream consequences stemming from ideologically biased labeling and data entry. Much of the data Team 2 has used in its efforts was collected and compiled by the US Census Bureau. The Census Bureau is tasked with producing datasets that accurately depict a portrait of the United States populus. The Census bureau employs a high standard when it comes to labeling its variables so as not to introduce ideological bias. The US Census Bureau uses internal resources (Government employees) to label and construct its datasets.

2. Team 2 has retained most of the original data labels ascribed by the US Census Bureau across the transformed dataset used by Team 2. In the limited scarios Team 2 has been called to create labels of its own, the Team has aimed to apply labels that fairly and accurately represent the data they describe.

3. The research question Team 2 attempted to answer pertained to Prince George's County Maryland. By nature, this research question is localized, and the sample size is somewhat limited when compared to an inquiry conducted at the national level. The data Team 2 has used to train its model is specific to Prince George's County; the results generated by Team 2's model are only applicable to Prince George's County. Team 2 has been transparent in stating the localized limits of the classifier it has constructed for this project.

Avoiding The Identification of Individuals Residing in At Risk Census Tracts

1. Privacy is a foundational principle of data ethics, and users of information collected from the public must employ the requisite safeguards to protect participant data as breaches of privacy can result in an erosion of institutional trust, threats to personal autonomy, voided informed consent agreements, and legal exposure.

2. Using the datasets Team 2 has constructed, it is not possible to identify specific individuals within Census tracts Team 2's classifier may determine to be at risk for foreclosure. This is mainly due to the fact Team 2's dataset is primarily composed of US Census Bureau Data, and this data has been anonymized by Census via the stripping of individually identifying information. Additionally, the Census data leveraged by Team 2 is grouped by tract rather than

address. At the address level it may be easier to identify individuals associated with specific housing units.

3. Team 2 does believe if a classifier such as the one constructed by the Team were to be used by state/federal-level decision makers, successful intervention efforts could still be undertaken with insights derived from the classifier at the tract level. This information would be enough for a decision maker to compose follow-on efforts to investigate or design mitigation of the effects of foreclosure in an at-risk tract.

Fair Representation of Target Population

1. At the core of Team 2's dataset is US Census Bureau ACS data. The Census Bureau gathers its ACS data via the mailing of surveys to a random sample of addresses in a defined geographic location. In its public documentation, the Bureau states its methods ensure proper coverage/representation of an area they choose to survey.

2. ACS data is updated each year in the form of estimates. These estimates include information that is current and includes data for numerous socio-demographic, housing, and economic variables. ACS surveys are designed to answer useful questions about the US domain, and their results are credible; however, ACS survey results suffer from a larger margin of sampling and non-sampling error than decennial census results. The US census bureau is transparent about this reality and openly cautions users of ACS data streams to take this factor into account when leveraging and deriving experimental results using ACS data. Team 2 has been transparent about its use of ACS data in the completion of this project and has been proactive in caveating the Team's results accordingly to ensure fair representation of conditions affecting the population of specific tracts of Prince George's County.