Improving efficiency of HCD tax credit program

Palak Agarwal

10-30-2020

Introduction

The Department of Housing and Community Development (HCD) in Emil City seeks to launch a targeted campaign to encourage homeowners to take advantage of a $5,000 tax credit for home repairs. Typically, only 11% of the eligible homeowners they reach out to take the credit. This analysis attempts to improve the efficiency of HCD's outreach efforts, minimizing outreach to homeowners who are unlikely to take the credit while maximizing outreach to homeowners who are likely to take the credit.
The method used to maximize this efficiency is by generating a matrix that tells us how likely is the user to accept the credit or not. To make these prediction many variables are used which are related to the likely of taking the credit or not.

Our analysis uses a binary logistic regression to estimate whether eligible homeowners are likely to take the home repair tax credit based on a number of features. our goal is to create a model that can accurately predict instances of when a homeowner will and will not take the credit. After engineering features to make our predictive model as accurate as possible, we then use a cost-benefit analysis to search for an optimal threshold to limit 'costly' errors, or those create the greatest cost to HCD while producing the least benefit to homeowners. Based on our understanding of the credit program, we constructed some stylized facts to inform the cost-benefit analysis:

- For each homeowner predicted to take the credit, HCD will allocate $2,850 for outreach (this figure includes staff and resources to facilitate mailers, phone calls, and information/counseling sessions at the HCD offices).
- Given our new targeting algorithm, we assume 25% of contacted homeowners take the credit.
- The credit costs $5,000 per homeowner which can be used toward home improvement.
- Houses that transacted after taking the credit sold with a $10,000 premium, on average.
- An additional benefit of the credit is that homes surrounding the repaired home see an aggregate premium of $56,000 on average, which HCD would like to consider as a benefit in the cost-benefit analysis.

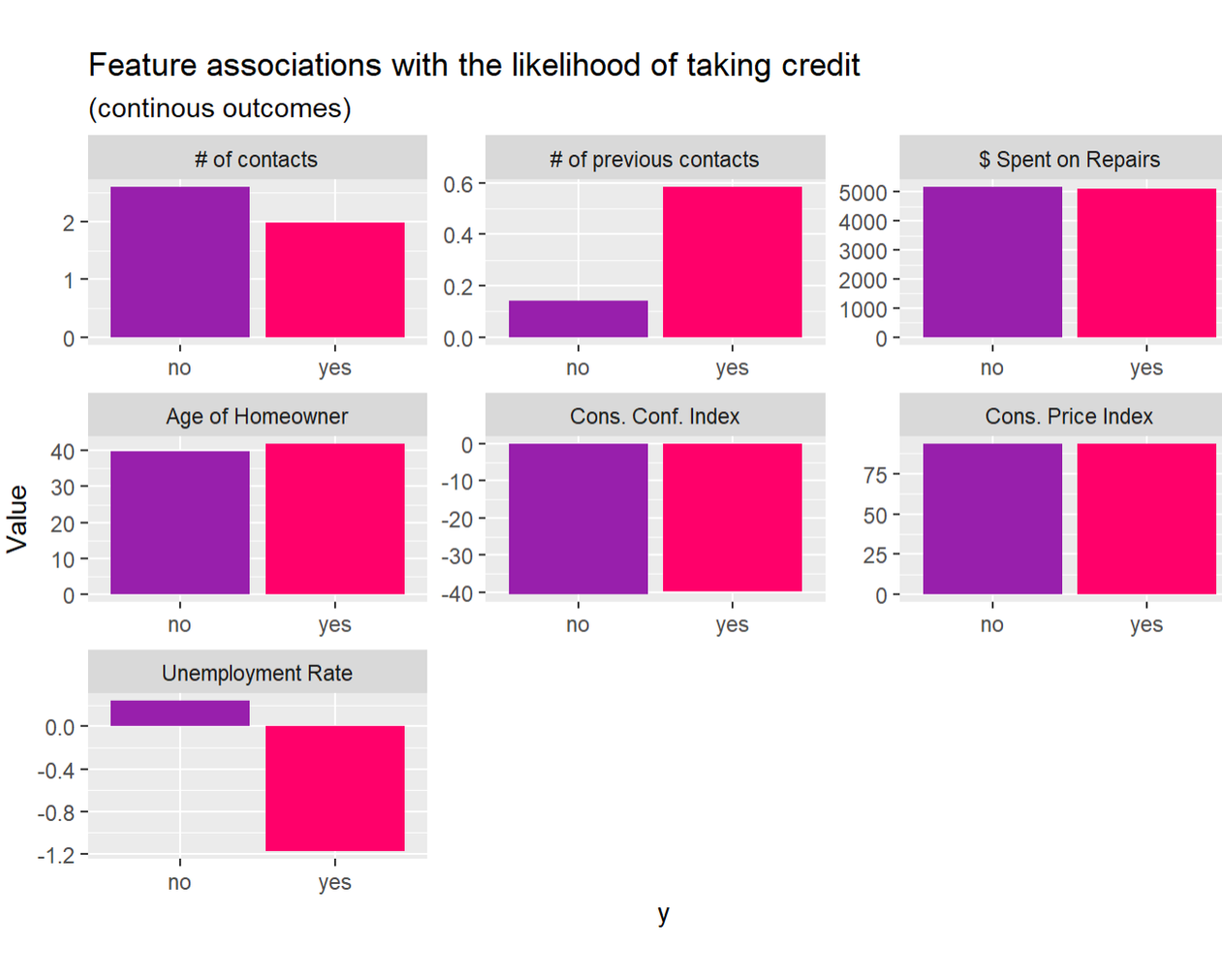
Initially the file is set up by loading the necessary libraries and the base file.

Set Up

Data Visualizations

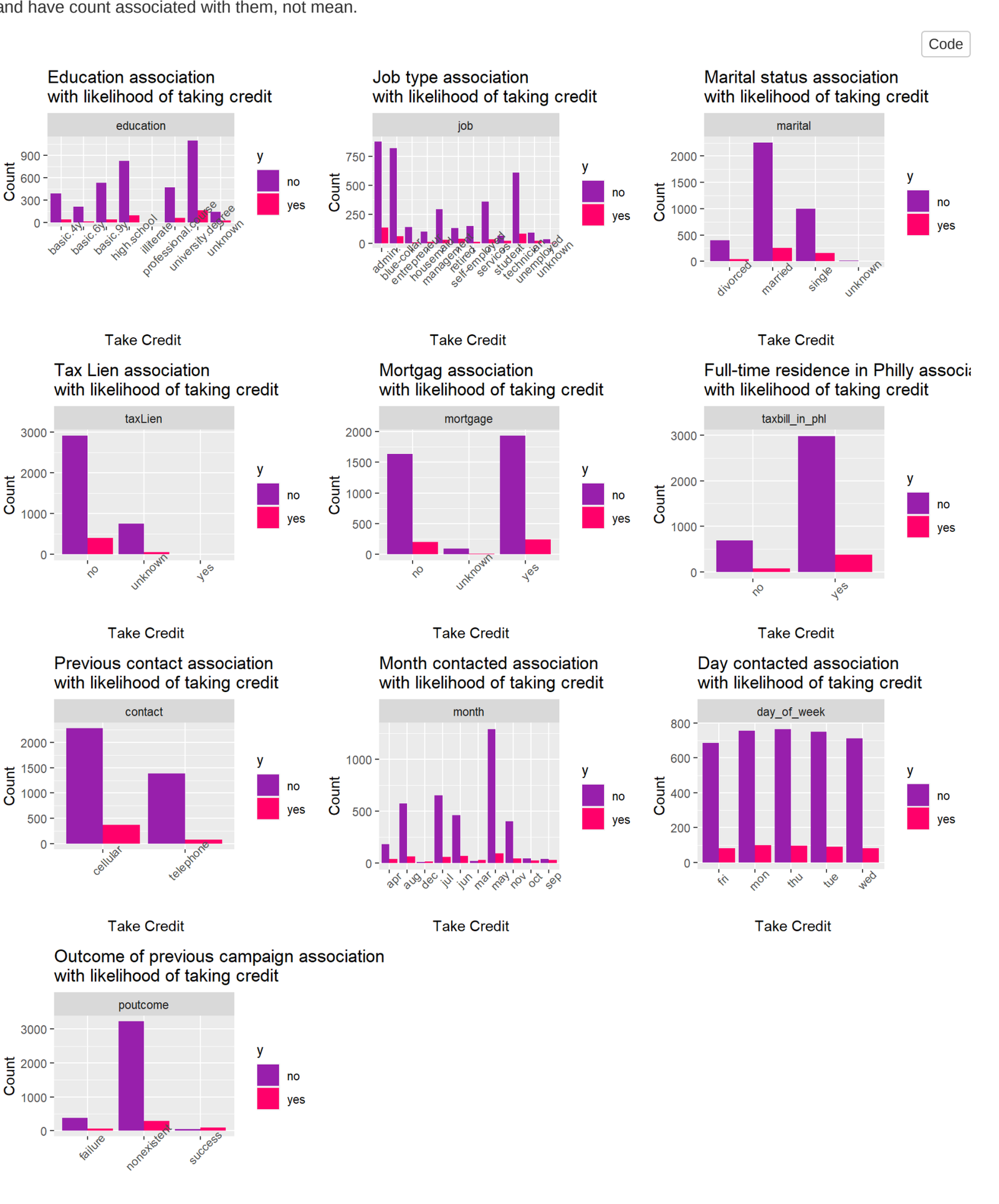
Below are some data visualizations - both of numeric and categorical variables.

From the below plots the following interpretations can be made :
- The mean number of contacts prior to this campaign is higher for those who take the credit than those who do not. - The mean amount of money spent annually on repairs is about the same. - The mean age of those who take the credit is slightly higher than the mean age of eligible homeowners that do not take the credit.
- The mean number of times homeowners were contacted within one campaign is higher for those who do not take the credit. - The mean consumer confidence index and mean consumer price index at the time of the campaign is about the same for those who take the credit and those who do not. - The mean unemployment rate is higher during more successful campaigns.



Categorical Variables

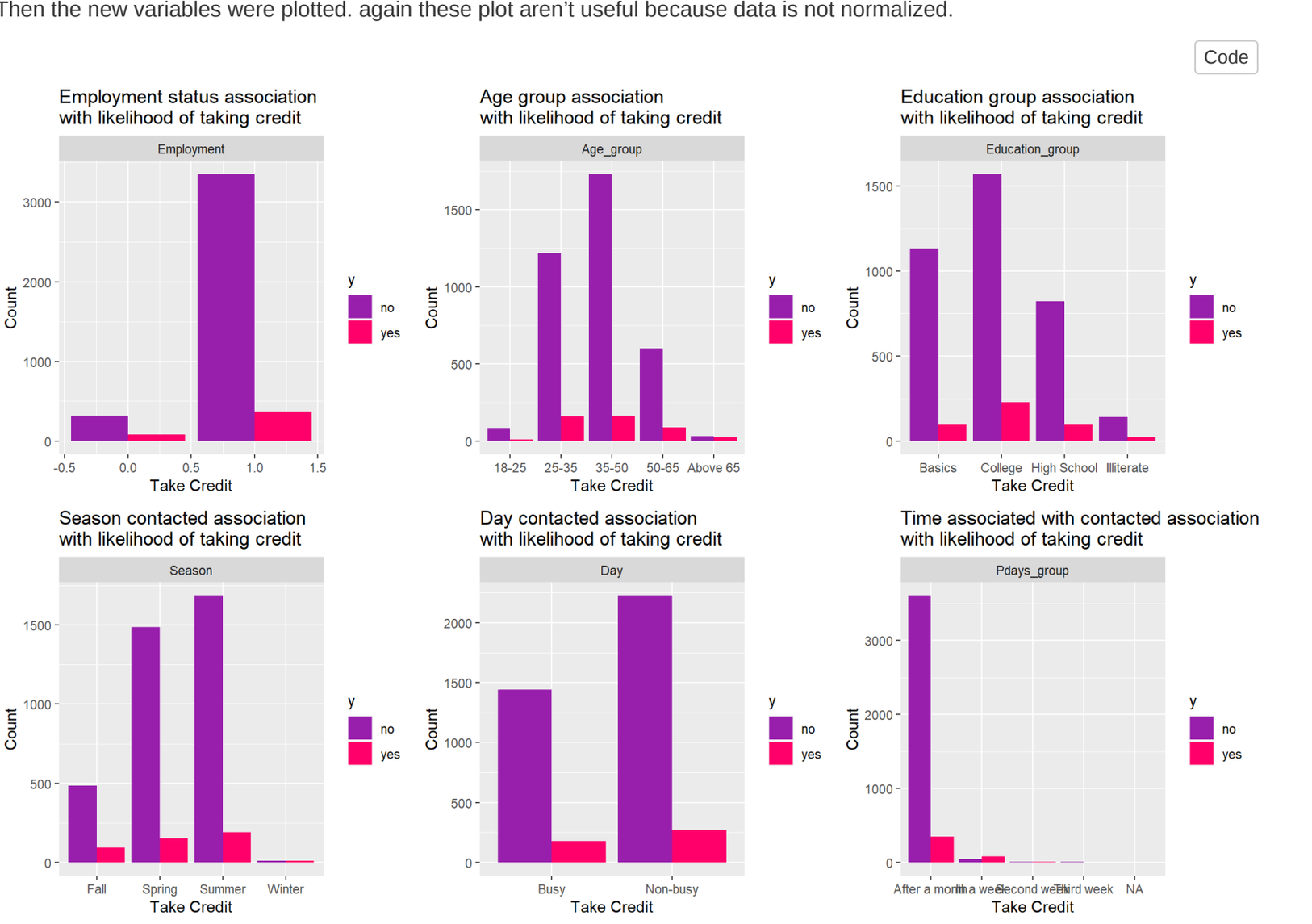
Counts of different levels of each categorical variable shown below. These plots are not useful because the data aren't normalized and have count associated with them, not mean.



Feature Engineering

For this part of the model, we used the above plots to identify a few variables that did not make sense to have as they were presented. Below are the modified features:
- Employment status - As the status of employment matters more than type of job, we reduced the job category to only 2 categories - employment and unemployed.
- Seasons - Checking when the user was contacted and identifying the season during that call improved the model. Seasons may impact when people are thinking about house repairs and are more likely to accept the credit during certain seasons.
- Days of week - Beginning and end of week calls to users might reduce their receptiveness to the marketing campaign when compared to calls during the mid week calls.
- Education & Age - The specific values of these columns are not as important as the larger brackets they fall under hence we feature engineered them to create smaller categories.
- Pday - When the consumer was contacted has been made categorical.

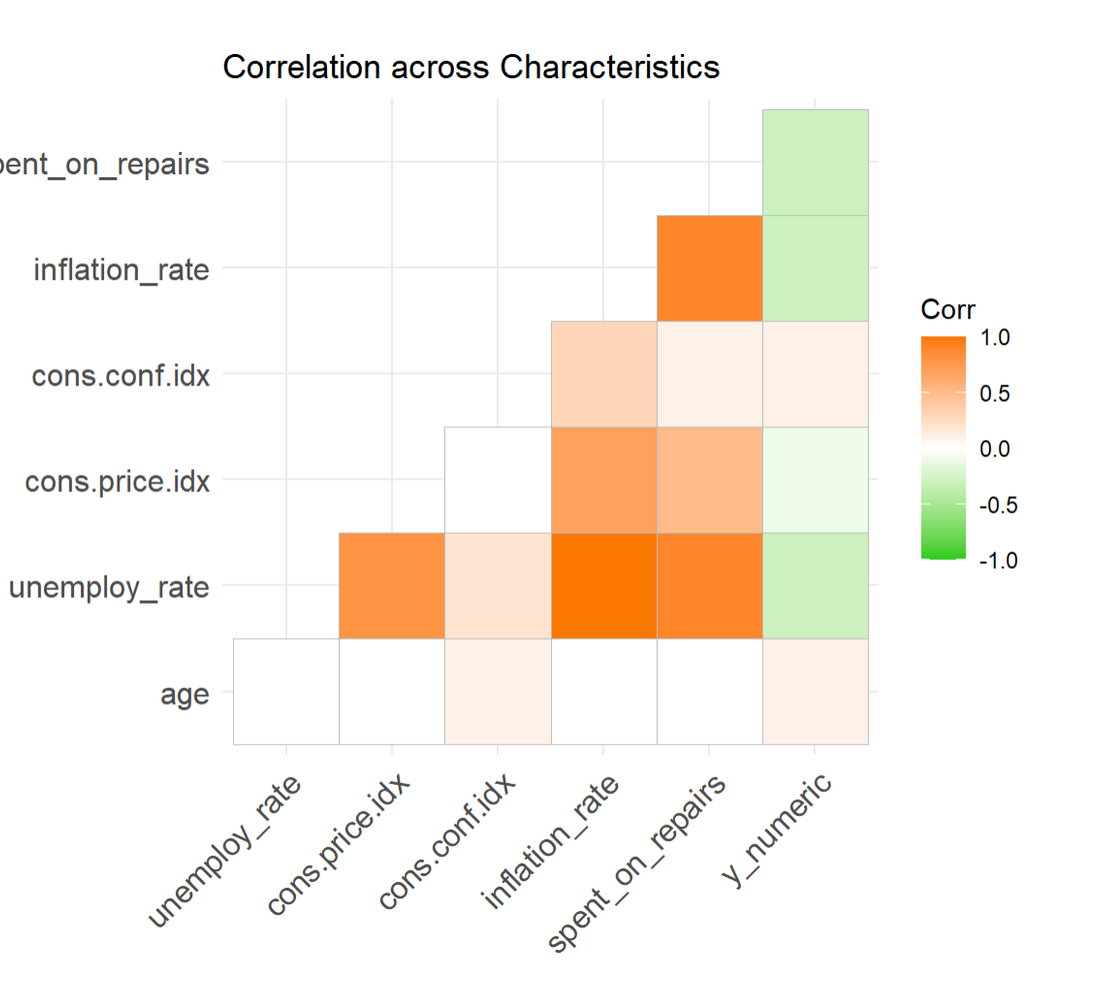
Then the new variables were plotted. again these plot aren't useful because data is not normalized.



Correlation

This plot shows that none of our numeric variables are strongly correlated with whether or not an eligible homeowner takes the credit.

The plot suggests that some of our numeric variables may be colinear. The unemployment rate at time of contact is strongly positively correlated with the inflation rate at time of contact. The inflation rate at time of contact is also strongly positively correlated with the amount of money spent annually on repairs.



Compare Models

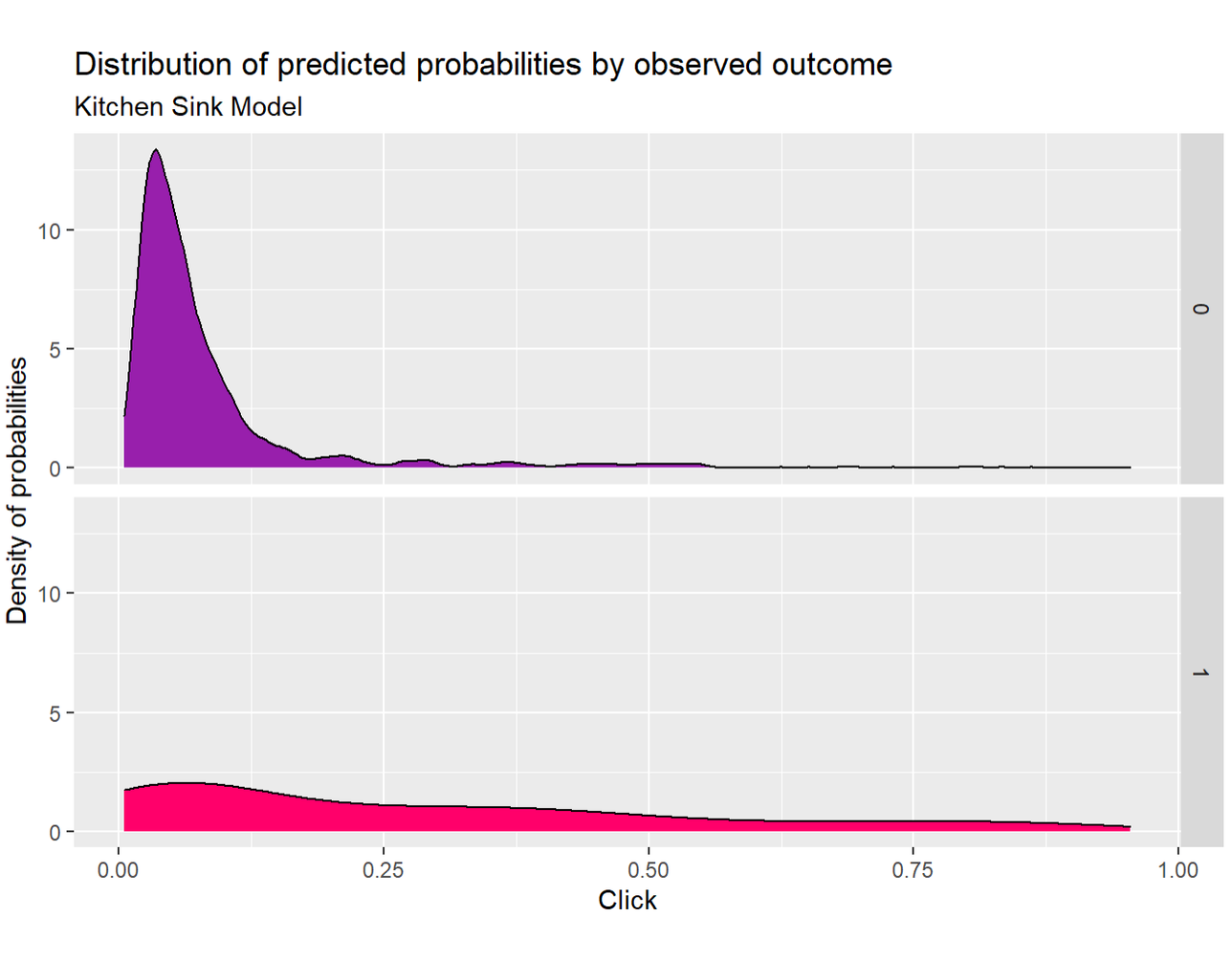
There are two models used in the regression below. One is the **Kitchen Sink Model** and the other is the **Feature Engineered Model.** Our model with feature engineering is not better than the "kitchen sink" model. Both have low sensitivities (true positive rate).

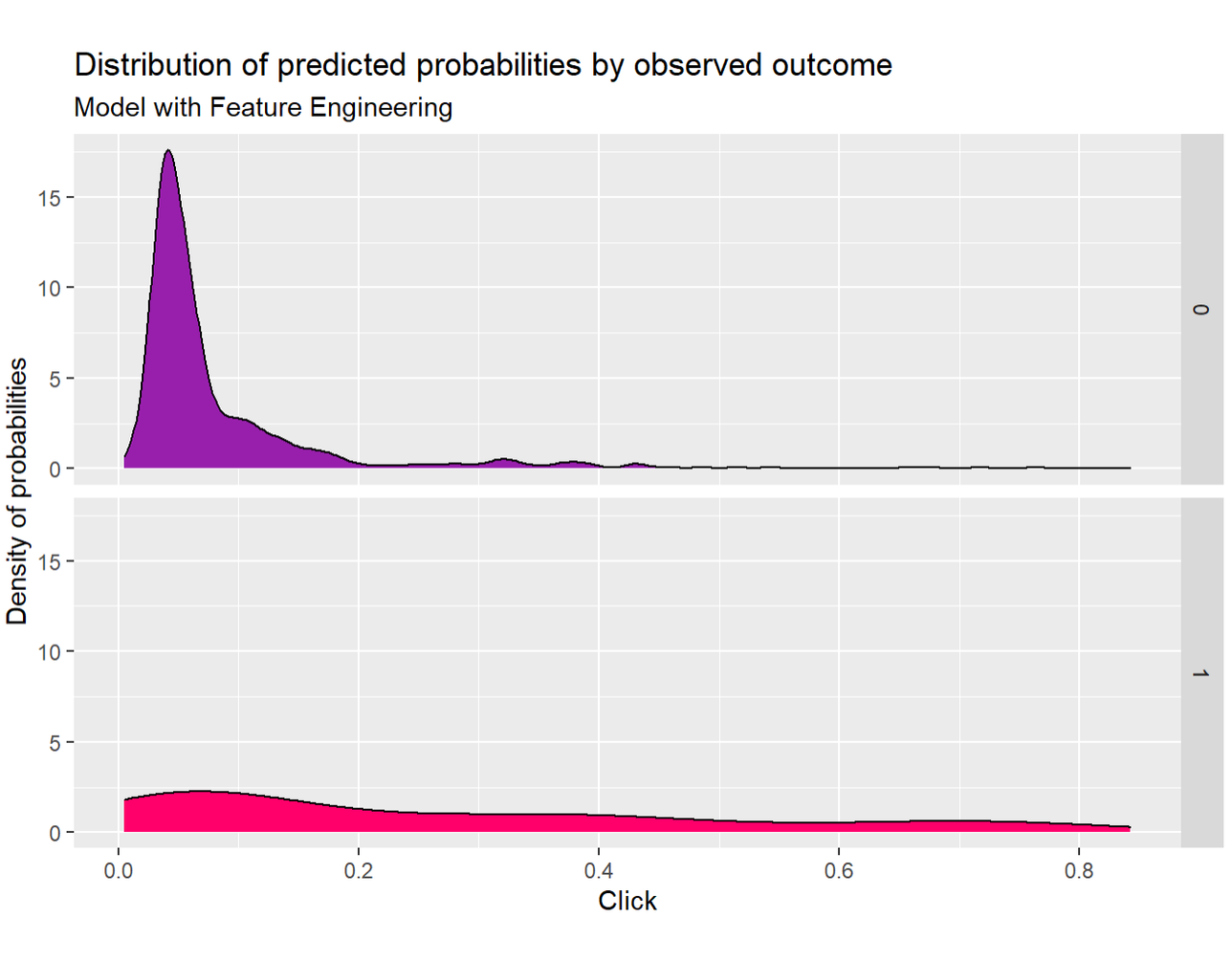
Confusion Matrix :
The confusion metrics as you can see has 4 possible outputs - combinations of 0 and 1. 1.True positive means we predicted they would take the credit and they took it. However, our research suggests that only 25% of our true positives will actually take the credit. (1 - observed, 1 - predicted)
2.True negative means we predicted they would not take the credit and they didn't. (0 - observed, 0 - predicted)
3.False positive means we predicted they would take the credit, but they did not. (0 - observed, 1 - predicted)
4.False negative means we predicted they would not take the credit, but they did. (0 - observed, 1 - predicted)

A very predictive regression would show clustering of 0 (don't take the credit) around 0 and clustering of 1 (do take the credit) around 1. Neither model shows clustering around 1, indicating that these models have poor sensitivity(true positive rate).

```
## fitting null model for pseudo-r2
```

```
##          llh      llhNull          G2     McFadden         r2ML         r2CU
## -779.5714022 -991.2194277  423.2960511    0.2135229    0.1425685    0.2776816
```






```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1216  102
##          1   21   27
##
##                Accuracy : 0.91
##                  95% CI : (0.8935, 0.9246)
##     No Information Rate : 0.9056
##     P-Value [Acc > NIR] : 0.3087
##
##                   Kappa : 0.2676
##
##  Mcnemar's Test P-Value : 0.0000000000005459
##
##             Sensitivity : 0.20930
##             Specificity : 0.98302
##          Pos Pred Value : 0.56250
##          Neg Pred Value : 0.92261
##              Prevalence : 0.09444
##          Detection Rate : 0.01977
##    Detection Prevalence : 0.03514
##       Balanced Accuracy : 0.59616
##
##        'Positive' Class : 1
##
```

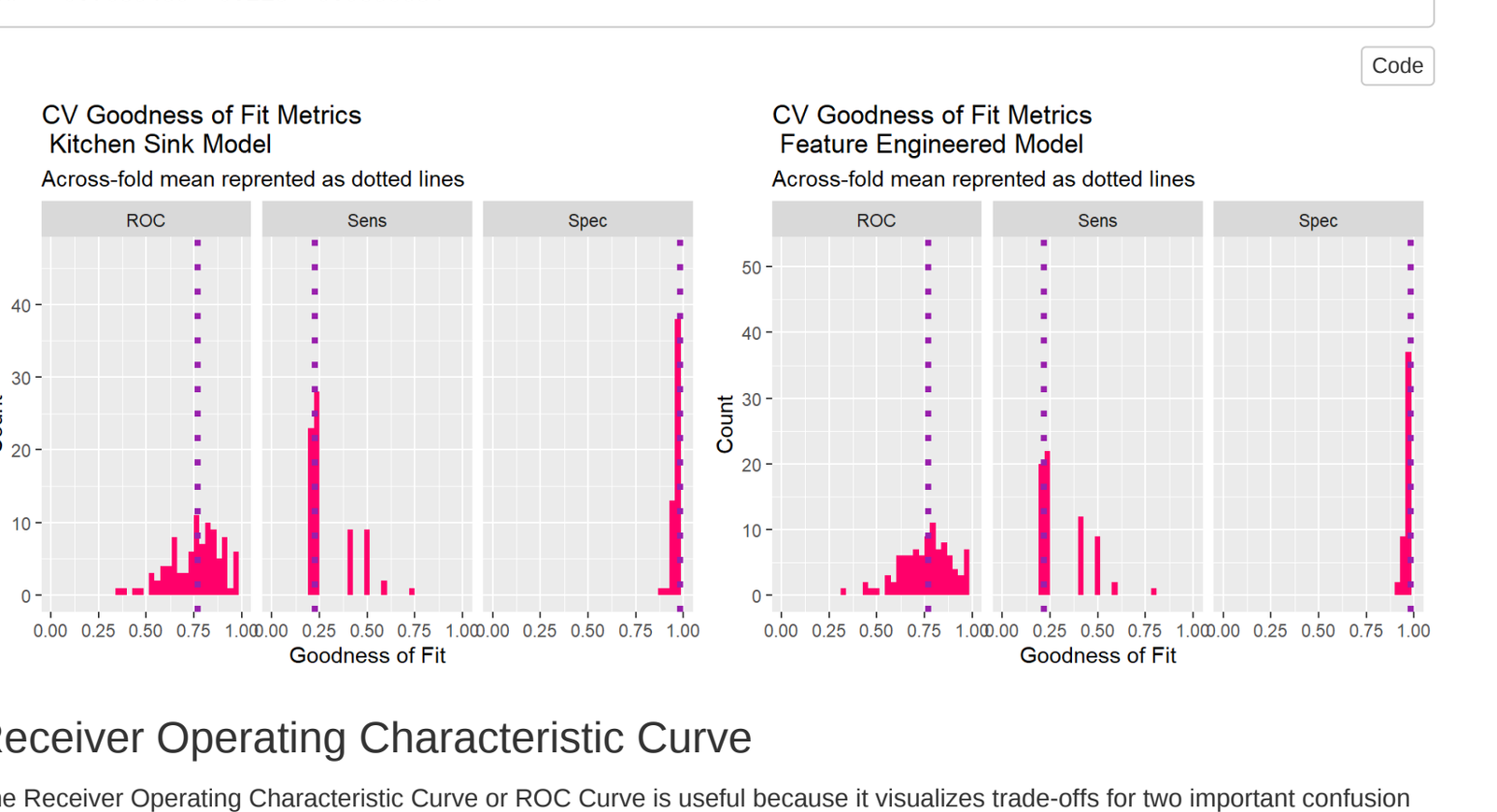
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1220  103
##          1   17   26
##
##                Accuracy : 0.9122
##                  95% CI : (0.8959, 0.9266)
##     No Information Rate : 0.9056
##     P-Value [Acc > NIR] : 0.2172
##
##                   Kappa : 0.2678
##
##  Mcnemar's Test P-Value : 0.000000000000008533
##
##             Sensitivity : 0.20155
##             Specificity : 0.98626
##          Pos Pred Value : 0.60465
##          Neg Pred Value : 0.92215
##              Prevalence : 0.09444
##          Detection Rate : 0.01903
##    Detection Prevalence : 0.03148
##       Balanced Accuracy : 0.59390
##
##        'Positive' Class : 1
##
```

Cross Validation

We tested both models using cross-validation. again, shows low sensitivity (true positive) and high specificity (true negative) rates for both models. The same can be seen on the plots that show the area under the ROC curve, sensitivity, and specificity. If model were generalizable and had a good goodness of fit, would expect all of these plots to be clustered around mean. As its visible, the specificity and the ROC of the models are good and can be a measure of goodness of fit but the variables need to be feature engineered better to get better sensitivity values.

```
## Generalized Linear Model
##
## 4119 samples
##   19 predictor
##    2 classes: 'c1.yes', 'c2.no'
##
## No pre-processing
## Resampling: Cross-Validated (100 fold)
## Summary of sample sizes: 4078, 4079, 4078, 4079, 4077, 4078, ...
## Resampling results:
##
##   ROC        Sens    Spec
##   0.7702121  0.2265  0.9806306
```

```
## Generalized Linear Model
##
## 4119 samples
##   17 predictor
##    2 classes: 'c1.yes', 'c2.no'
##
## No pre-processing
## Resampling: Cross-Validated (100 fold)
## Summary of sample sizes: 4077, 4077, 4077, 4077, 4078, 4079, ...
## Resampling results:
##
##   ROC        Sens   Spec
##   0.7685439  0.218  0.9833634
```

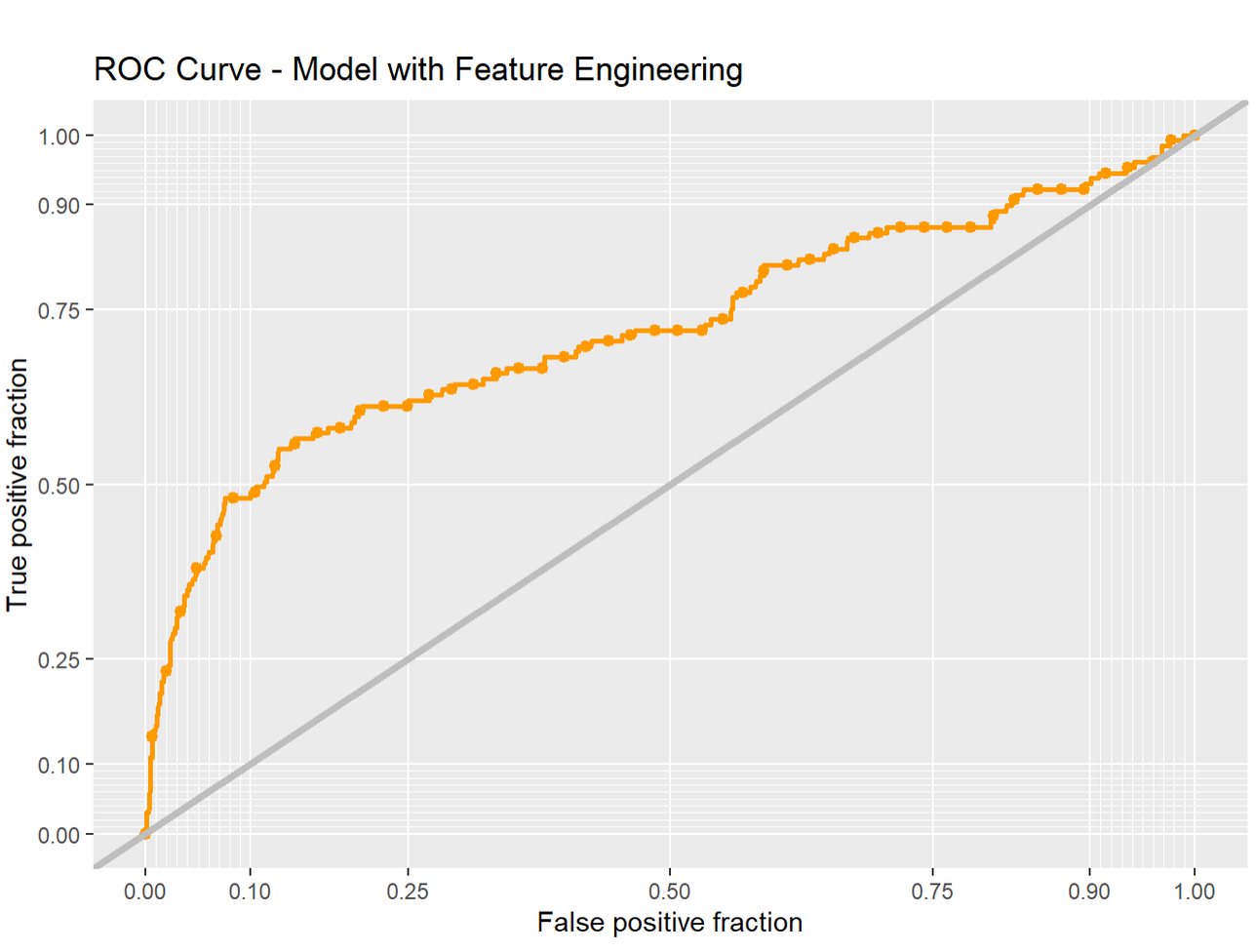


Receiver Operating Characteristic Curve

The Receiver Operating Characteristic Curve or ROC Curve is useful because it visualizes trade-offs for two important confusion metrics, while also providing a single goodness of fit indicator. When increase true positives, also increase false positives, that means HCD will waste money on marketing more. For example, according to the ROC Curve, a threshold that predicts taking the credit correctly 50% of the time, will predict taking the credit incorrectly 10% of the time.

The AUC is an indicator of goodness of fit. A 100% is overfit, 50% would be coin flip, and anything between the two is a useful fit. The AUC of our model is 72.24% which indicates that our model predicts reasonably well and it is a goodness of fit metric.

```
## Area under the curve: 0.7224
```



Cost/Benefit Analysis

For the cost and benefit table was split into HCD expenditures and increases to house values. As the HCD is a non profit and has no direct relation to the increase in the costing of the house, we felt was apt to split the two costing. Below is the explanation of how we split the costing among the possible outcomes.

- True negative: no HCD expenditure, no housing increases, **Count*0**
- True positive: HCD spends $2850 on marketing for all true positives and $5000 on tax credit for the 25% that actually take it, equation is **(Count2850)+(Count.255000)); benefit is $10000 increase in home value plus $56000 increase in value of surrounding homes, equation is (.25Count10000)+(.25Count*56000)**
- False negative: HCD did not spend money on marketing, but homeowner still took credit. Since we are analyzing impact of marketing campaign, we zero this out. Equation for both is **Count*0**
- False positive: HCD spends money on marketing, equation is **count2850. No increase in values to homes. Equation is Count0**

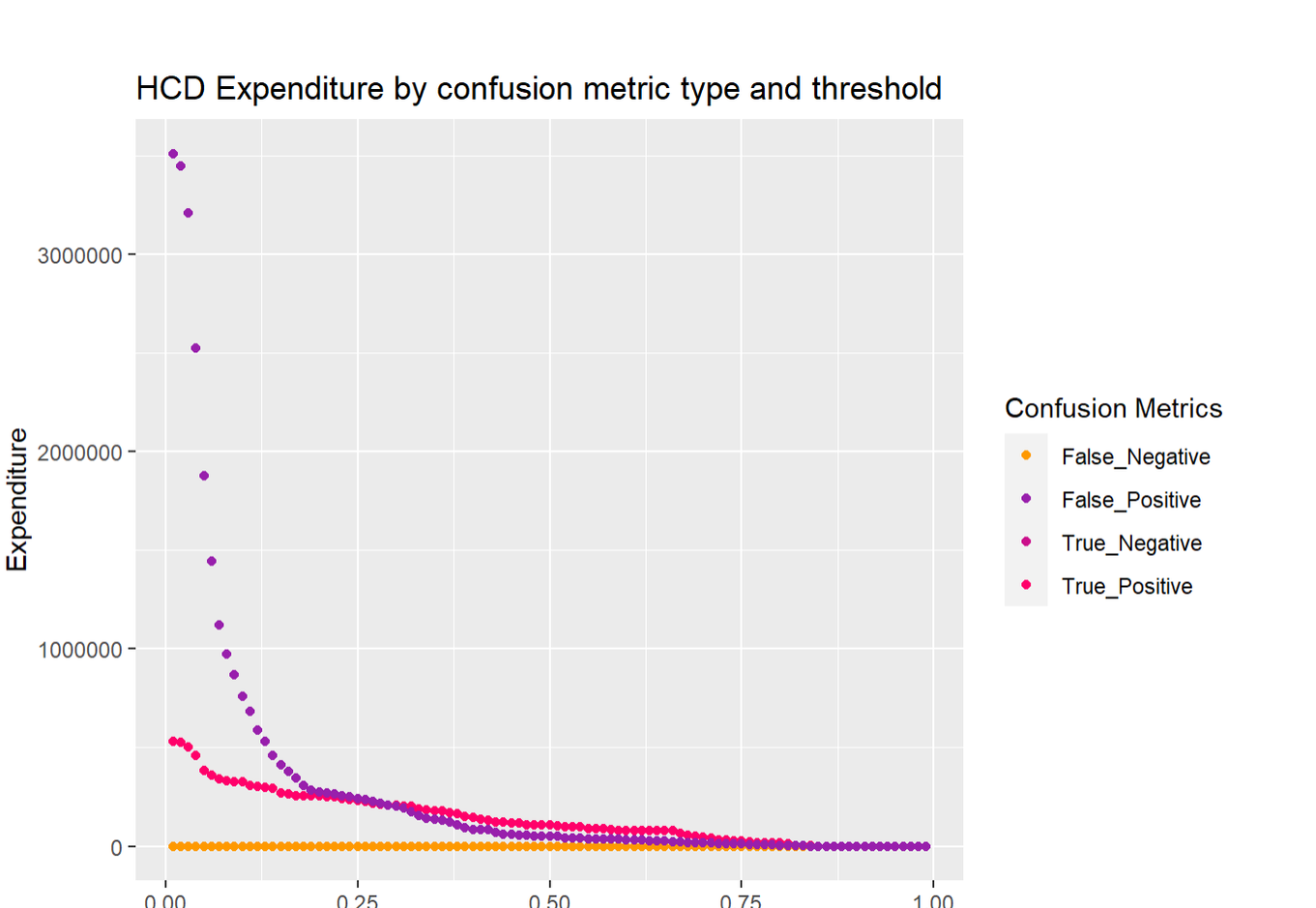
Cost/Benefit Table

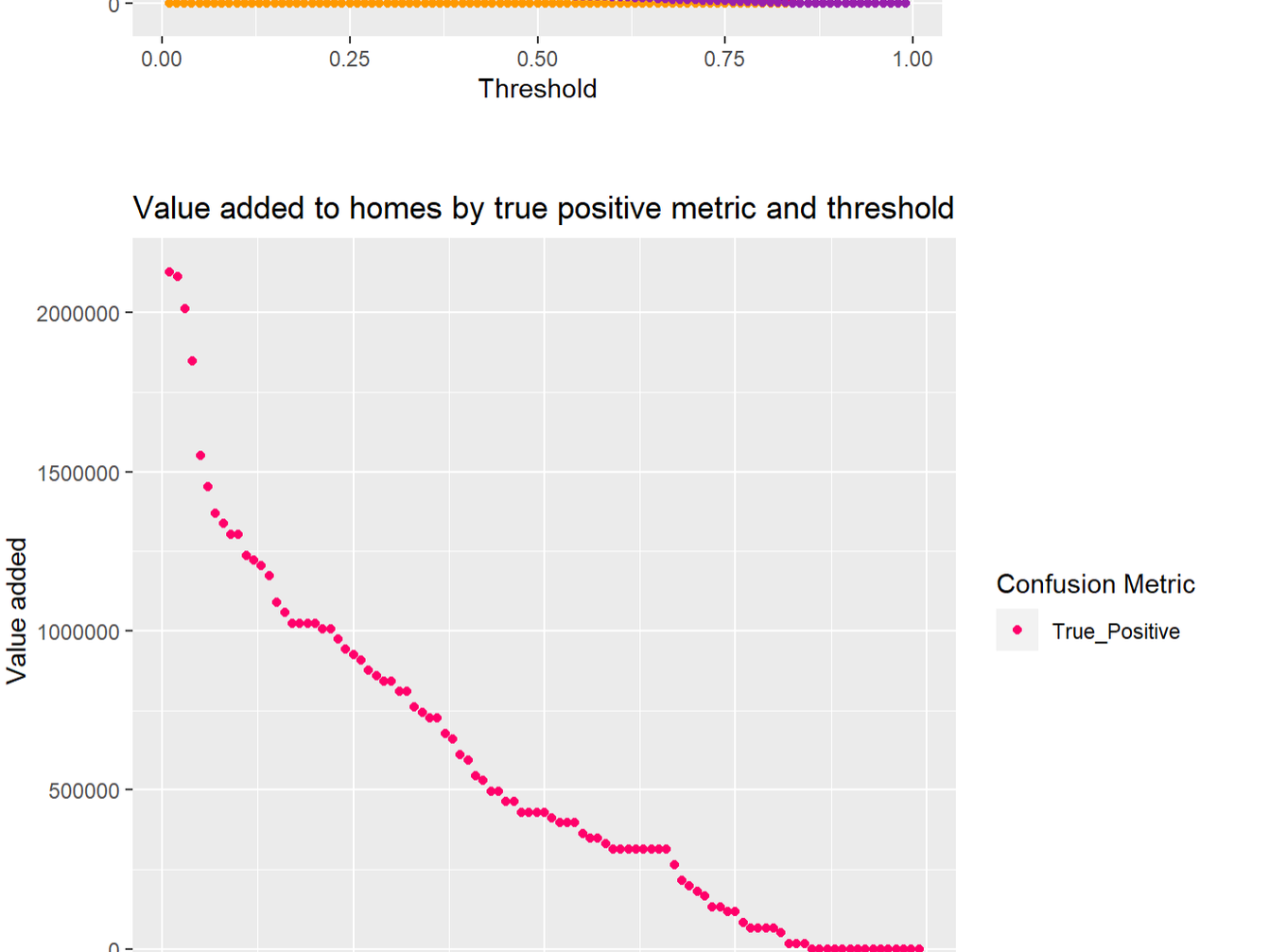
Variable	Count	HCD_Expenditure	Home_Value_Added	Number_Credits	Description
True_Negative	1220	0	0	0.0	We correctly predicted not taking credit
True_Positive	26	106600	429000	6.5	We correctly predicted taking credit
False_Negative	103	0	0	0.0	We predicted would not take credit and customer took credit
False_Positive	17	48450	0	0.0	We predicted customer would take credit and customer did not take credit

Comparing Thresholds

Confusion Metric Plots

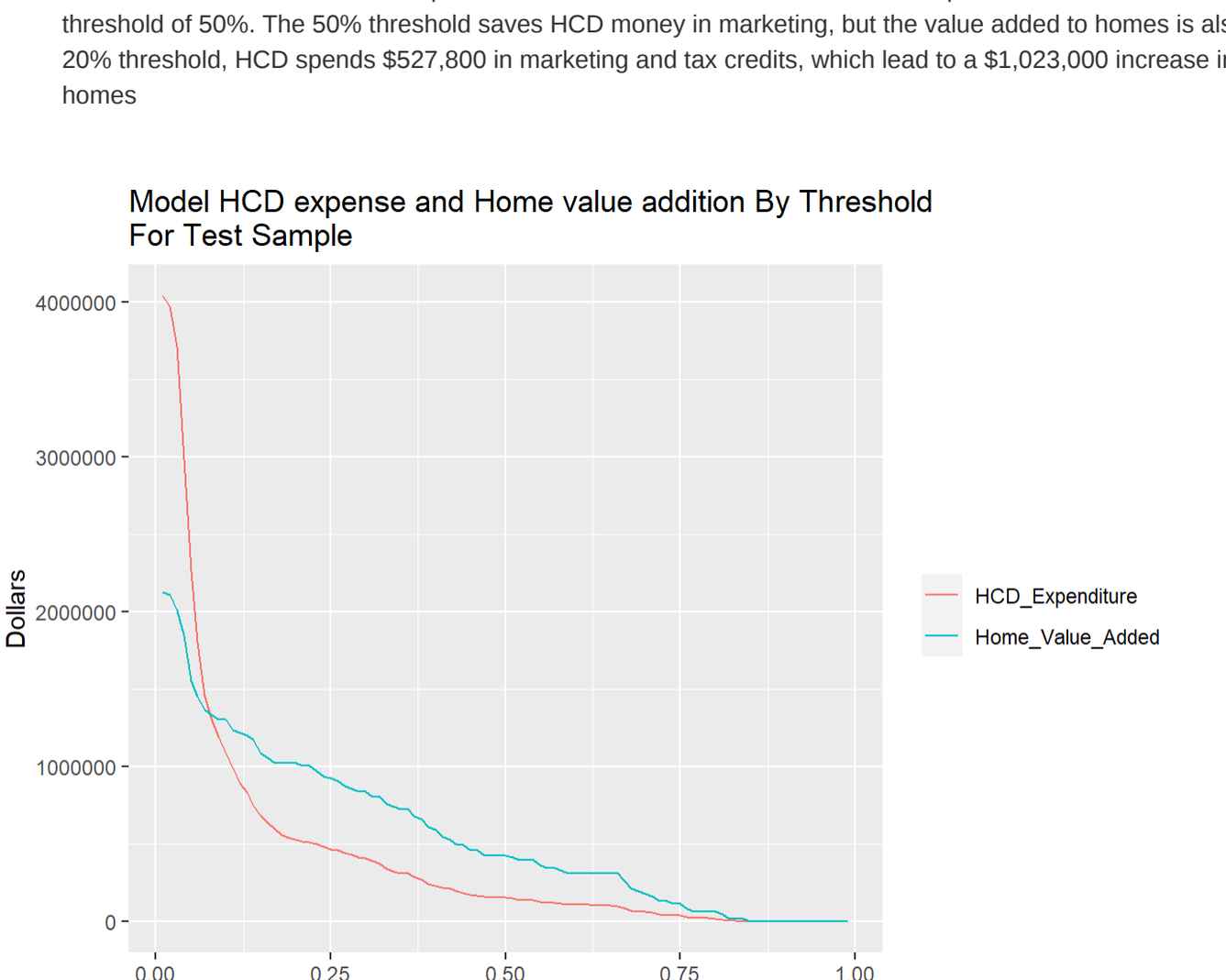
Next we move on to plotting the confusion metrics for all thresholds from 1% to a 100%. To do this we use a function called **iteratethreshold** which saves the value of HCD expenditure, value added to the homes and the number of credits for each threshold.
In the below plots you can see that for HCD expenditure by confusion metric, the true negative is not visible asits below the false negative because it is also 0 and are overlaid one top of other. The plot also shows that the HCD expenditure on false positives decreases steeply as threshold increases until a threshold of about 12.5%.
For added value to homes by confusion metric plot, value added would be highest at a threshold of 0. However, a lot of money would be wasted on marketing, as seen in the previous plot. No other confusion metric is shown as no other metric sees a change in home values.

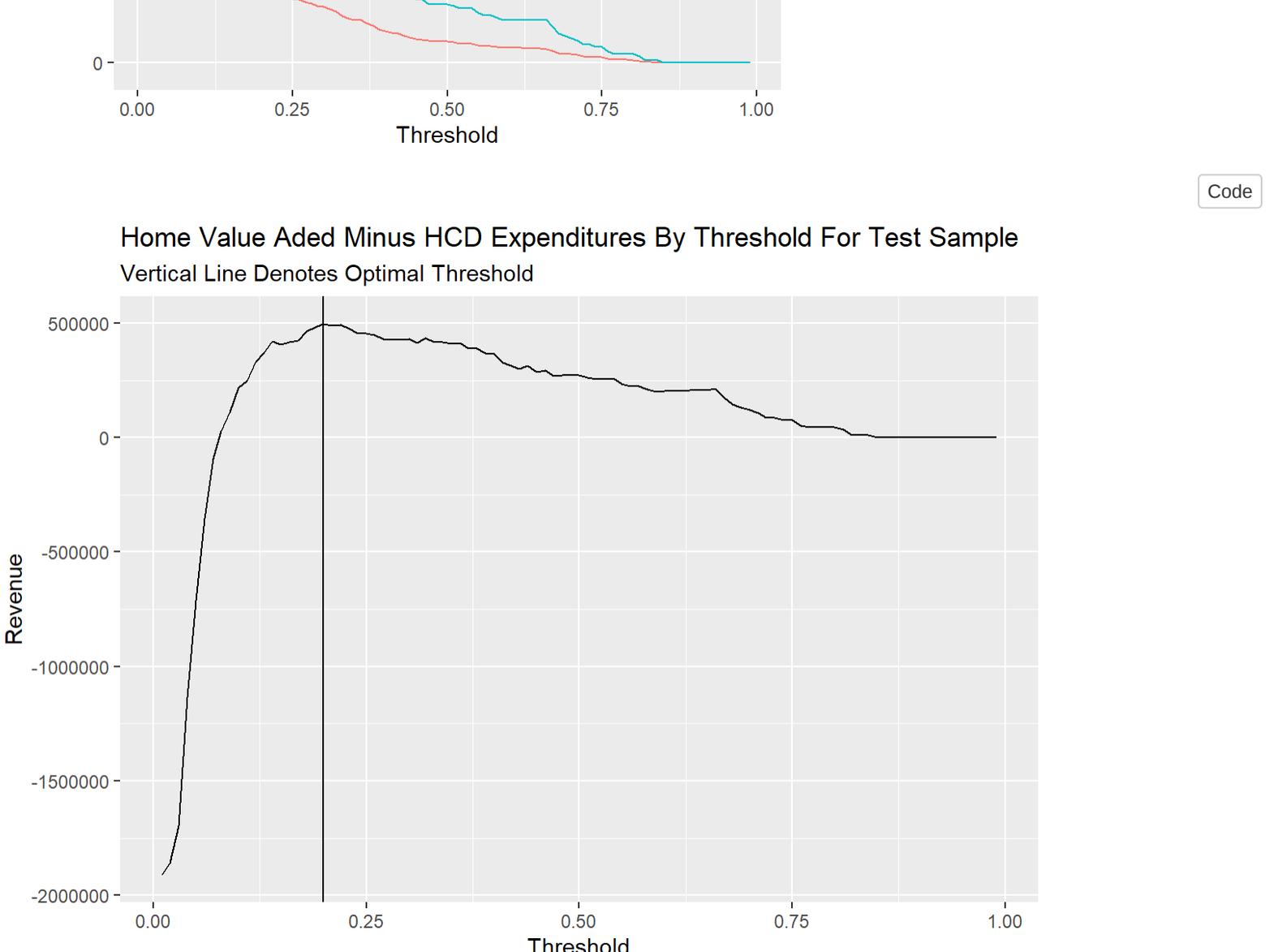


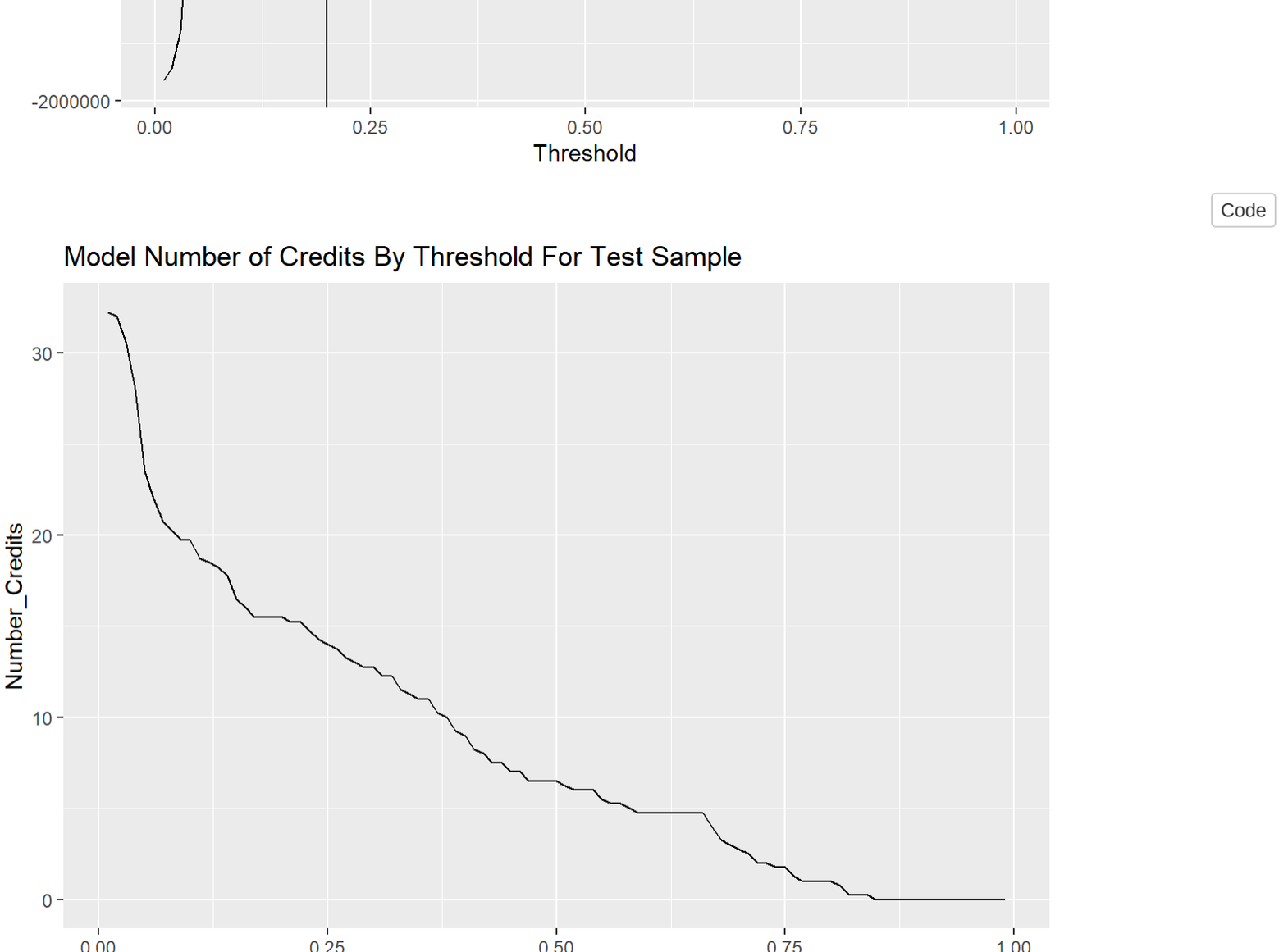


Threshold Plots

- First plot shows both HCD expenditures and value added to homes for each threshold. The goal is to minimize HCD expenditures while maximizing value added to homes. HCD being a government agency it has no profit gain calculation of its own and works for the betterment of the public.
- The second plot combines these terms, taking the value added to homes and subtracting HCD expenditures for each threshold. This shows that the optimal threshold is about .2, meaning any homeowner with a 20% probability of taking the credit is considered a "positive" case. This plot also shows that at thresholds below about 10%, HCD spends more money than is gained by adding value to houses
- The third plot shows the number of tax credits claimed at each threshold.This shows that the highest number of tax credits would be claimed with a very low threshold, meaning almost every eligible homeowner would receive marketing materials. However, the previous plots show that this would cost too much money compared to the financial benefits to the homeowners
- The final table shows the HCD expenditures and value added to homes for our optimal threshold of 20% and a default threshold of 50%. The 50% threshold saves HCD money in marketing, but the value added to homes is also quite low. At a 20% threshold, HCD spends $527,800 in marketing and tax credits, which lead to a $1,023,000 increase in the value of the homes







Optimum Threshold and 50% Threshold

Threshold	HCD_Expenditure	Home_Value_Added	Number_Credits
0.2	527800	1023000	15.5
0.5	155050	429000	6.5

Conclusion

In general, I would recommend putting this model into production as it will benefit the highest number of households and generate the highest amount of direct and indirect benefits to the Emil City community but it would need smarter features and more demographic data. The main issue is that the sensitivity is very low, meaning that the model does not do a good job of predicting actual positive credit acceptances. This is likely because there are so few "yes" outcomes in the underlying data.
In order to improve the model, I would recommend working with more data to improve the model, or engineer better features for predicting. To ensure that the marketing materials resulted in a better response rate, I would first test my improved method as a pilot program. This could serve as a test case to get a sense of whether the new method is working, or if it needs to be further improved before being implemented at a larger scale. In cases where there are limited resources available, it may be better to be cautious (thus using a pilot approach) before implementing an entirely new and untested method.