Credit Lending Decisions

2021 University of Georgia Data Science Competition

Cody Dailey[[1]](#footnote-20)

Ishaan Dave1

Yang Ge1

Vipul Shinde[[2]](#footnote-21)

15 April 2021

Table of Contents

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

## [[1]]  
## [1] "dplyr" "stats" "graphics" "grDevices" "utils" "datasets"   
## [7] "methods" "base"   
##   
## [[2]]  
## [1] "flextable" "dplyr" "stats" "graphics" "grDevices" "utils"   
## [7] "datasets" "methods" "base"   
##   
## [[3]]  
## [1] "knitr" "flextable" "dplyr" "stats" "graphics" "grDevices"  
## [7] "utils" "datasets" "methods" "base"

dir()

## [1] "DRAFT\_UGA\_DSC\_report\_20210414.docx" "DRAFT\_UGA\_DSC\_report\_20210414.log"   
## [3] "DRAFT\_UGA\_DSC\_report\_20210414.tex" "DRAFT\_UGA\_DSC\_report\_20210415.docx"  
## [5] "methods DSC .docx" "README.md"   
## [7] "UGA\_DSC\_report.Rmd"

getwd()

## [1] "C:/Users/daile/Documents/GitHub Projects/UGA-Data-Science-Competition/06-Report/01-Written"

For more details, please see appendix (add exactly where – figure XXX).

# Background

# 1 Competition Prompt

We are expected to work as a team to develop ***credit decision models*** using the simulated datasets for Bank XYZ, to interpret the results, document findings, and present to a panel of judges.

Specifically, we are asked to  
1. Conduct an exploratory data analysis (Sections 2 & 3)  
2. Develop and fit a logistic regression model (Section @(logistic))  
3. Develop an additional machine learning (ML) model: random forest, gradient boosting, or feedforward neural network (Section 6)  
4. Compare the logistic regression model with ML model, select one for credit approval, and describe the reason for our selection (Section 7)  
5. Answer the following questions: (Section 8)  
- Do customers who already have an account with the financial institution receive any favorable treatment in our model?  
- Suppose a credit card application is rejected using our model and the applicant asks us to provide an explanation as to why it was rejected. How would we explain the results to the customer?

# Bank XYZ and Credit Bureaus’ Credit Data

# 2 Univariate Descriptions

Table 2.1

Table 2.1: Descriptive Statistics of Training Data

| Variable | Levels | Defaulted | Did Not Default | Statistic | Parameter | p | Test |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | n = 1586 (7.9) | n = 18414 (92.1) |  |  |  |  |
| tot\_credit\_debt |  | 91915.57 (29920.36) | 94791.79 (22901.39) | -3.73522708815311 | 1748.60975862171 | 0.000193560594634868 | Welch Two Sample t-test |
|  | Missing | 0 (0) | 0 (0) |  |  |  |  |
| avg\_card\_debt |  | 17880.43 (21195.06) | 13761.61 (7363.26) | 7.69916624800648 | 1618.10844391613 | 2.36867032982657e-14 | Welch Two Sample t-test |
|  | Missing | 0 (0) | 0 (0) |  |  |  |  |
| credit\_age |  | 275.16 (61.99) | 298.55 (61.34) | -14.4324429070927 | 1862.47774919973 | 7.72558297814542e-45 | Welch Two Sample t-test |
|  | Missing | 0 (0) | 0 (0) |  |  |  |  |
| credit\_good\_age |  | 140.54 (33.56) | 150.57 (33.94) | -11.4132170587326 | 1875.31942243245 | 3.22845824673055e-29 | Welch Two Sample t-test |
|  | Missing | 0 (0) | 0 (0) |  |  |  |  |
| card\_age |  | 247.71 (58.6) | 269.76 (59.11) | -14.3719935016147 | 1873.67889535775 | 1.64707943912853e-44 | Welch Two Sample t-test |
|  | Missing | 0 (0) | 0 (0) |  |  |  |  |
| non\_mtg\_acc\_past\_due\_12\_months\_num |  | 0 [0, 2] | 0 [0, 0] | 19325868 |  | 0 | Wilcoxon rank sum test with continuity correction |
|  | Missing | 0 (0) | 0 (0) |  |  |  |  |
| non\_mtg\_acc\_past\_due\_6\_months\_num |  | 0 [0, 0] | 0 [0, 0] | 17862939 |  | 0 | Wilcoxon rank sum test with continuity correction |
|  | Missing | 0 (0) | 0 (0) |  |  |  |  |
| mortgages\_past\_due\_6\_months\_num |  | 0 [0, 1] | 0 [0, 0] | 18383330 |  | 0 | Wilcoxon rank sum test with continuity correction |
|  | Missing | 0 (0) | 0 (0) |  |  |  |  |
| credit\_past\_due\_amount |  | 3060.92 (5777.48) | 94.01 (1048.88) | 20.4221795705161 | 1594.01051969845 | 1.56191097097667e-82 | Welch Two Sample t-test |
|  | Missing | 0 (0) | 0 (0) |  |  |  |  |
| inq\_12\_month\_num |  | 2 [1, 4] | 1 [0, 3] | 17234540 |  | 1.58825932027483e-34 | Wilcoxon rank sum test with continuity correction |
|  | Missing | 0 (0) | 0 (0) |  |  |  |  |
| card\_inq\_24\_month\_num |  | 4 [2, 6] | 3 [1, 5] | 16837121 |  | 1.75437761438812e-24 | Wilcoxon rank sum test with continuity correction |
|  | Missing | 0 (0) | 0 (0) |  |  |  |  |
| card\_open\_36\_month\_num |  | 0 [0, 0] | 0 [0, 0] | 15023057 |  | 0.00247170924553078 | Wilcoxon rank sum test with continuity correction |
|  | Missing | 0 (0) | 0 (0) |  |  |  |  |
| auto\_open\_.36\_month\_num |  | 0 [0, 0] | 0 [0, 0] | 14743461.5 |  | 0.287739876179222 | Wilcoxon rank sum test with continuity correction |
|  | Missing | 0 (0) | 0 (0) |  |  |  |  |
| uti\_card |  | 0.56 (0.11) | 0.5 (0.11) | 21.4902150021776 | 1853.56884866499 | 1.18553508793446e-91 | Welch Two Sample t-test |
|  | Missing | 0 (0) | 0 (0) |  |  |  |  |
| uti\_50plus\_pct |  | 0.56 (0.11) | 0.51 (0.11) | 16.9887708483998 | 1855.95850699333 | 2.81352034855187e-60 | Welch Two Sample t-test |
|  | Missing | 0 (0) | 0 (0) |  |  |  |  |
| uti\_max\_credit\_line |  | 0.55 (0.11) | 0.5 (0.11) | 16.5067492322018 | 1858.45601121345 | 3.16764188025015e-57 | Welch Two Sample t-test |
|  | Missing | 0 (0) | 0 (0) |  |  |  |  |
| uti\_card\_50plus\_pct |  | 0.54 (0.12) | 0.48 (0.12) | 17.4766494080874 | 1641.42052530212 | 7.40236731448236e-63 | Welch Two Sample t-test |
|  | Missing | 174 (11) | 1881 (10.2) |  |  |  |  |
| ind\_acc\_XYZ | Account w/ XYZ | 328 (20.7) | 4843 (26.3) | 23.7634718425882 | 1 | 1.08930246049978e-06 | Pearson's Chi-squared test with Yates' continuity correction |
|  | No Account | 1258 (79.3) | 13571 (73.7) |  |  |  |  |
|  | Missing | 0 (0) | 0 (0) |  |  |  |  |
| rep\_income |  | 74522.99 (16775.13) | 75583.34 (16323.78) | -2.32028527342509 | 1701.35804644551 | 0.02044321906901 | Welch Two Sample t-test |
|  | Missing | 129 (8.1) | 1441 (7.8) |  |  |  |  |
| States | AL | 229 (14.4) | 2664 (14.5) | 4.05737720234524 | 6 | 0.668911785016524 | Pearson's Chi-squared test |
|  | FL | 219 (13.8) | 2638 (14.3) |  |  |  |  |
|  | GA | 209 (13.2) | 2648 (14.4) |  |  |  |  |
|  | LA | 228 (14.4) | 2621 (14.2) |  |  |  |  |
|  | MS | 241 (15.2) | 2586 (14) |  |  |  |  |
|  | NC | 224 (14.1) | 2674 (14.5) |  |  |  |  |
|  | SC | 236 (14.9) | 2583 (14) |  |  |  |  |
|  | Missing | 0 (0) | 0 (0) |  |  |  |  |

# 3 Exploratory Data Analysis

## 3.1 Methods

Data were summarized using means and standard deviations or frequency and percentages as appropriate. For discrete count data, data were summarized using medians and interquartile range. Additional exploratory data analyses were conducted to examine normality of continuous variables and cross-tabulations of categorical variables between outcome levels. Bivariate relationships were examined using correlations (point-biserial between categorical and continuous, and Pearson correlation between continuous variables) and plotted in a correlation matrix (darker colors represent stronger correlations). To predict whether or not individuals will default, 2 models were fit: the first was an adaptive LASSO logistic regression model, and the second was a random forest machine learning algorithm for classification. Their predictive performances were compared by examining differences in area under the curve (AUC) and receiver operating curves (ROC), and the financial impact of correct (or incorrect) classification was assessed by assigning a cost to correctly or incorrectly classifying an individual.

## 3.2 Results

Figure Correlation Matrix

# 4 Modeling and Predicting Probabilities of Account Default

# 5 Logistic Regression

Our main outcome of interest is whether or not an individual defaults when asking for a loan. Traditionally, associations between predictors and outcome are tested individually through univariate regression models, then a subset of those covariates are selected for a multivariate regression analysis if univariate analyses yielded significant results. Given the large number of possible predictors, we suspected severe multicollinarity between some pairs of variables, and thus, overfitting of our model. To alleviate this, we “penalized” regression coefficients to select a parsimonious set of variables that leads to similar prediction results.

## 5.1 Methods

In a normal logistic regression with a binary outcome, our model is defined by logit(E(Y│X)=β^T (X), where Y is the binary outcome (whether or not an individual defaulted), X = (1,X\_1,…,X\_p) is the vector of covariate values, and β = (β\_0,β\_1,…,β\_p ) is the vector of regression parameters. In a penalized regression, estimation is generally in the form of an optimization problem that seeks to maximize the function in the form of L(β)-λ\*pen(β), where pen(β) is the “penalty” and λ is the “tuning parameter.” The penalty refers to a constraint surrounding the betas and the tuning parameter corresponds to the amount of “shrinkage” applied to the coefficient. In this analysis, an adaptive LASSO (least absolute shrinkage and selection operator) with logistic regression was employed to estimate coefficients and select the most important predictors. Adaptive LASSO is conducted in 2 parts – the first is a “ridge” regression is conducted that uses a penalty proportional to the sum of squares of the regression coefficients and efficiently handles multicollinearity (Equation 1).

(Equation 1)

The second step is the LASSO step that imposes a constraint on the sum of the absolute value of the regression coefficients and uses data-dependent weights (the inverse of the ridge regression coefficients) to penalize regression coefficients accordingly (Equation 2). Therefore strong predictors’ coefficients are shrunk less than coefficients of weak predictors.

(Equation 2).

Continuous predictors were standardized to have mean 0 and standard deviation 1 before penalized regression was conducted. The main focus of penalized regression is predictive performance rather than uncertainty around the coefficient itself. Confidence intervals for coefficients and predicted probabilities from penalized methods are problematic because coefficients are already biased toward the null because of the inherent penalty term and not presented with regression coefficients. In the table below, we show the standardized coefficient estimates for both ordinary logistic regression and adaptive lasso as well as the shrinkage associated with each coefficient.

The main purpose of LASSO is predictive performance rather than coefficient estimates. We used a training dataset to best inform the penalty and tuning parameters to be used in the LASSO. A validation test set was used to best determine the parameters in our model and our final model’s predictive performance was determined by running our model on a testing dataset and record the resulting AUC associated with classifying an individual as one who will or will not default. Moreover, a confusion matrix of actual vs. predicted counts of those who did and did not default was constructed to calculate predictive accuracy, classifications were performed at multiple predicted probability thresholds to determine that which resulted in highest predictive accuracy. To quantify predictive performance of our models, a monetary value was associated with both correctly and incorrectly classifying an individual.

In addition to an adaptive LASSO logistic regression model, a random forest machine learning algorithm was run to compare predictive accuracy

## 5.2 Results

Figure Adaptive LASSO - Training –> ROC & Cost on training and validation datasets

Table - Coefficients / ORs / Probabilities - Confusion Matrices

Figure 5.1

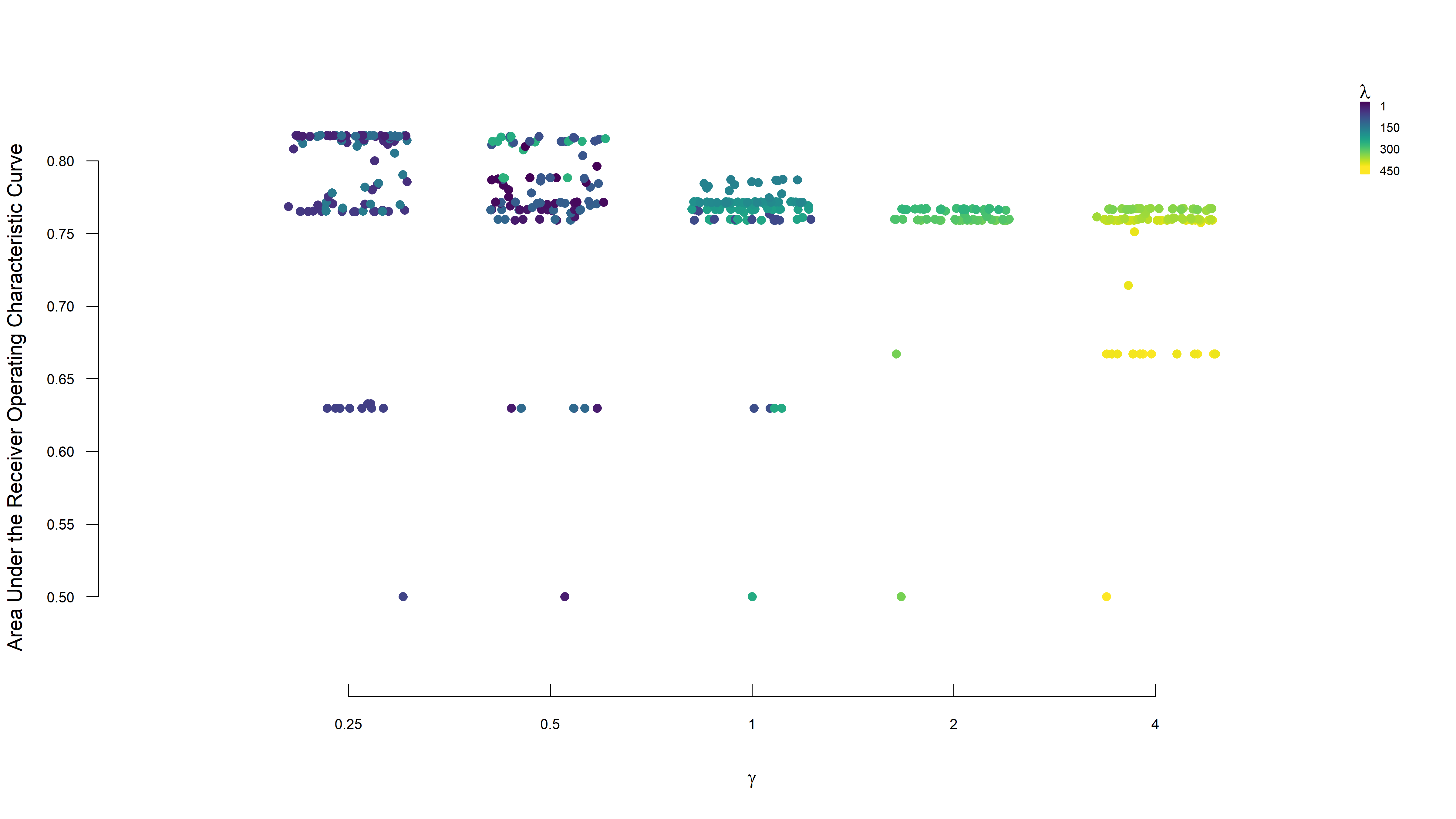


Figure 5.1: Adaptive LASSO Logistic Regression Hyperparameters and Areas Under the Receiver Operating Characteristic Curve

Figure 5.2

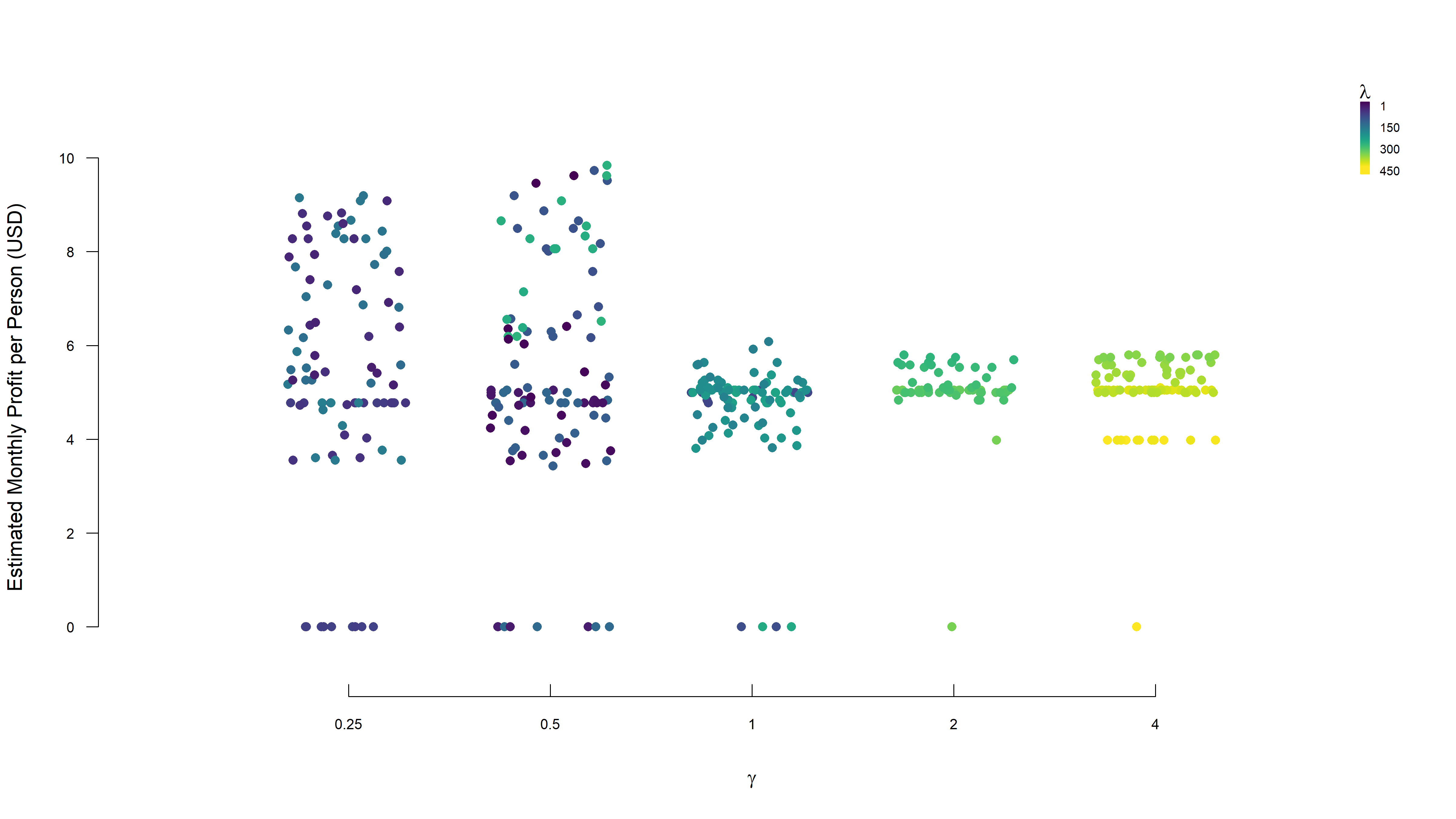


Figure 5.2: Adaptive LASSO Logistic Regression Hyperparameters and Estimated Monthly Profit per Applicant

# 6 Random Forest

## 6.1 Methods

## 6.2 Results

Figure - Training –> ROC & Cost on training and validation datasets - Variable importance - Proximity? - “Best” tree?

Table - Confusion matrices

Figure 6.1

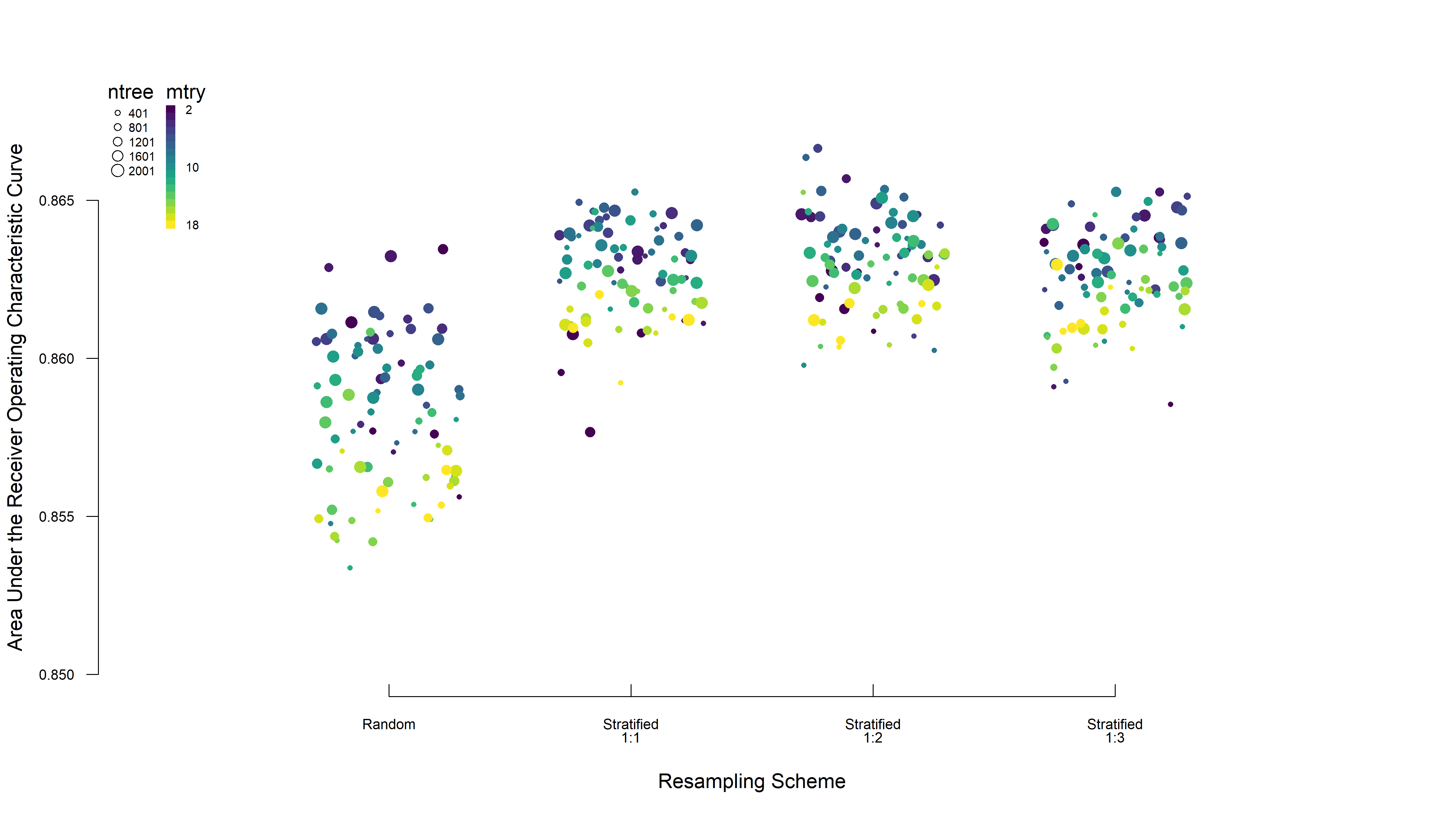


Figure 6.1: Random Forest Hyperparameters and Areas Under the Receiver Operating Characteristic Curve

Figure 6.2

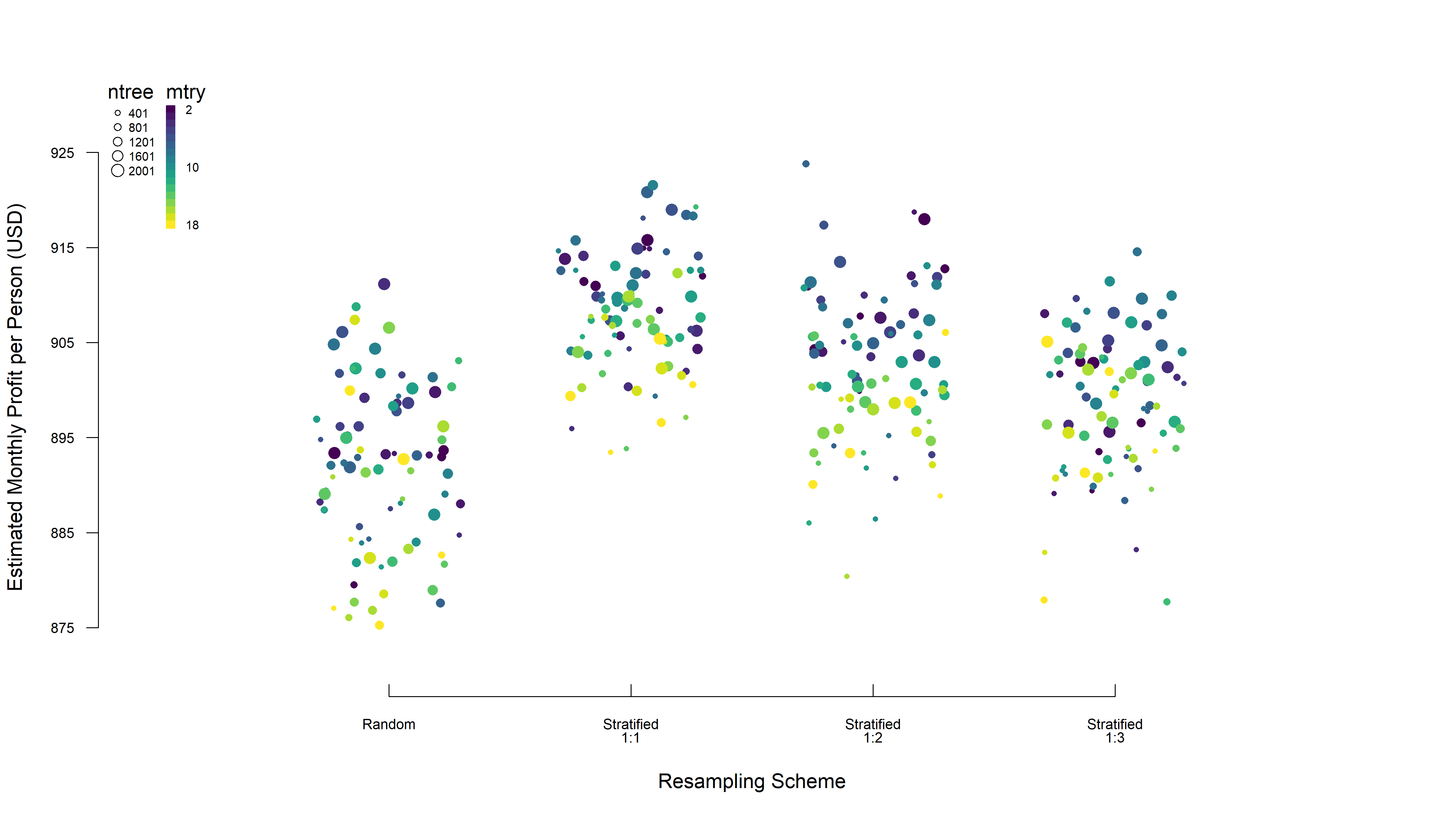


Figure 6.2: Random Forest Hyperparameters and Estimated Monthly Profit per Applicant

Figure 6.3

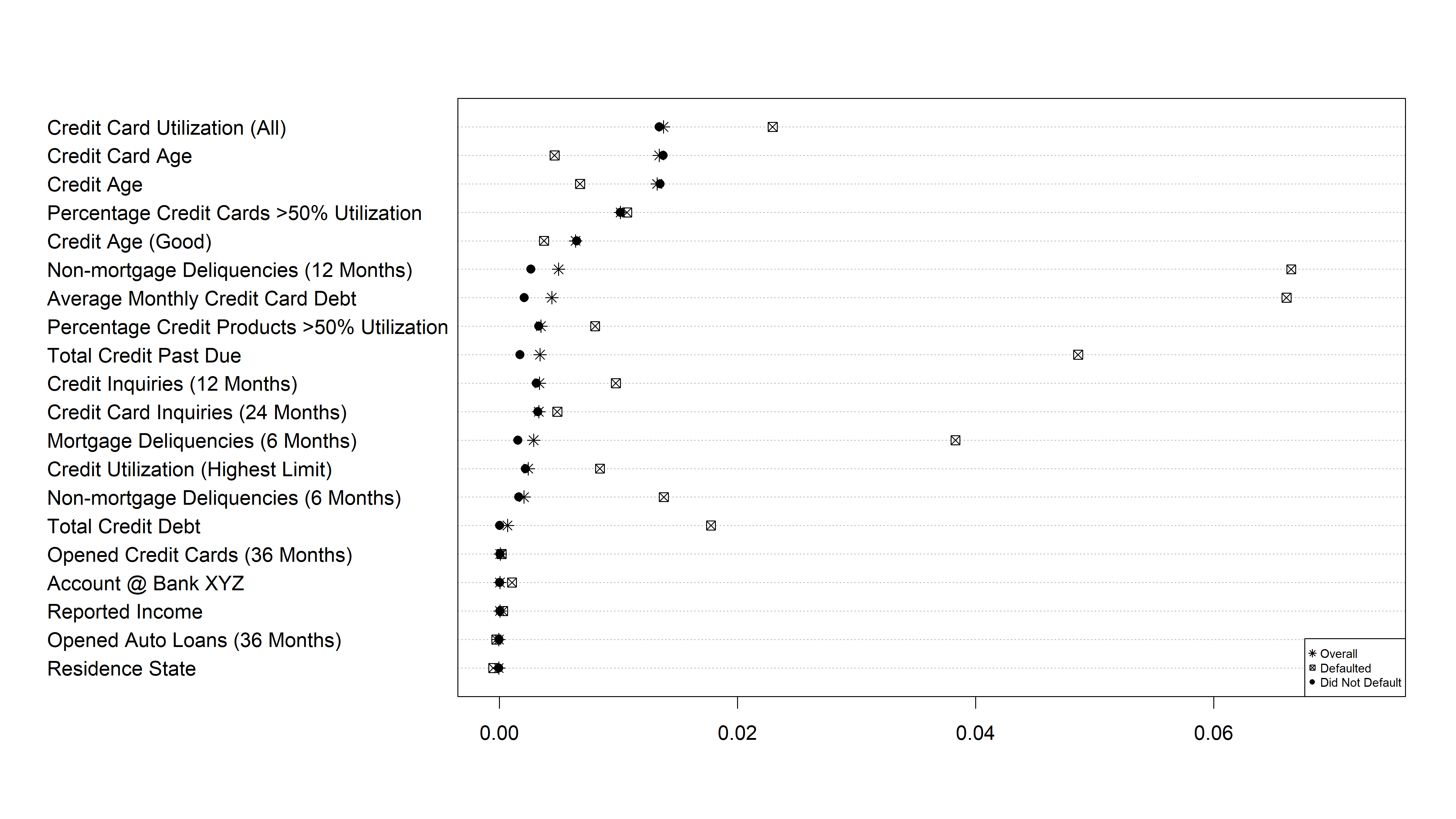


Figure 6.3: Variable Importance in Random Forest Model

# 7 Model Comparisons

Table 7.1

Table 7.1: Confusion Matrix for Adaptive LASSO Logistic Regression Model Predictions on the Test Data

|  | | **Truth** | |  |
| --- | --- | --- | --- | --- |
|  |  | *Defaulted* | *Did Not Default* |  |
| **Prediction** | *Defaulted* | 398 | 4,150 | 4,548 |
| *Did Not Default* | 3 | 449 | 452 |
|  |  | 401 | 4,599 | 5,000 |

Table 7.2

Table 7.2: Confusion Matrix for Random Forest Model Predictions on the Test Data

|  | | **Truth** | |  |
| --- | --- | --- | --- | --- |
|  |  | *Defaulted* | *Did Not Default* |  |
| **Prediction** | *Defaulted* | 390 | 3,060 | 3,450 |
| *Did Not Default* | 11 | 1,539 | 1,550 |
|  |  | 401 | 4,599 | 5,000 |

Table 7.3

Table 7.3: Various Performance Metrics Comparing Predictions on the Test Data from Adaptive LASSO Logistic Regression and Random Forest Models

|  |  |  |
| --- | --- | --- |
| *Metric* | **LASSO** | **RF** |
| Accuracy | 16.94% | 38.58% |
| Error | 83.06% | 61.42% |
| Profit | $31205.95 | $97091.97 |
| Profit.mp | $6.24 | $19.42 |
| PPV | 0.088 | 0.113 |
| NPV | 0.993 | 0.993 |
| Sensitivity | 0.993 | 0.973 |
| Specificity | 0.098 | 0.335 |
| Youden | 0.090 | 0.307 |
| d2c | 0.902 | 0.666 |
| LR.pos | 1.100 | 1.462 |
| LR.neg | 0.077 | 0.082 |
| DOR | 14.35 | 17.83 |

## 7.1 Accuracy

Figure - ROC & Cost on test data

## 7.2 Other Considerations

Transparency …

## 7.3 Selection

# Discussion

# 8 Scenario Application

***Do customers who already have an account with Bank XYZ receive any favorable treatment?***

Was it picked up in any of our models? What advantage might it have that wouldn’t be evident in data?

***Suppose a credit card application is rejected, how would you explain the results to the customer?***

For TP and FP, does it differ?

Figure - “Tipping point” plots (if we took any value from specific applicant and replaced with mean of someone who would have been approved, does it change probability enough to cross threshold? iterative combinations of strongest more variable predictors or z-score of particular customer?)

Figure 8.1

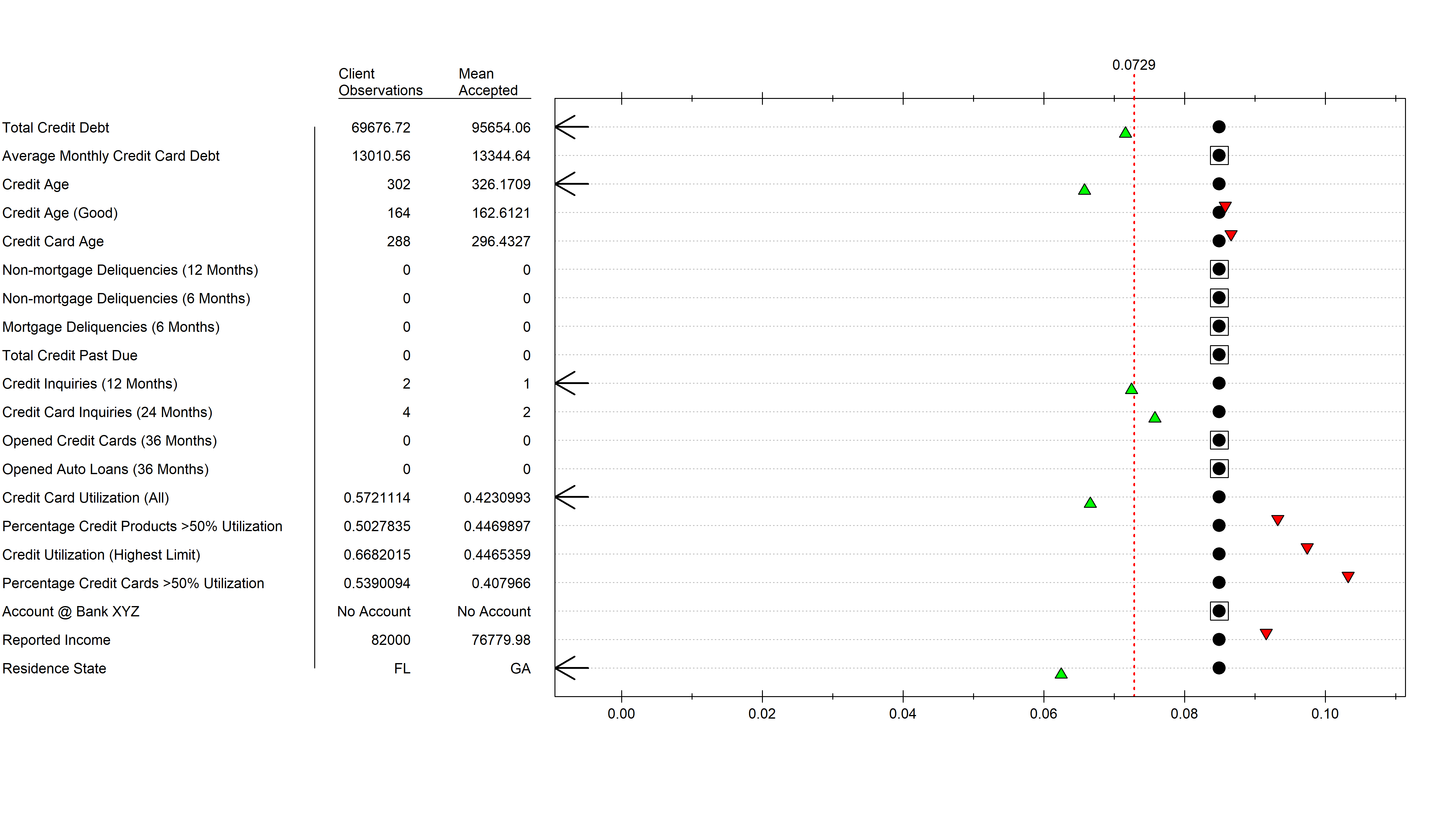


Figure 8.1: Explanation of A Rejected Applicant

# 9 Conclusions

Dilemma of basing credit approvals on default data (inherently missing approvals that should have been made, identifying only those that shouldn’t have been)

# 10 References

# Appendix

## 10.1 Extended Exploratory Data Analysis

## 10.2 Training Details

### 10.2.1 Class Imbalance

### 10.2.2 Cutoff Thresholds

## 10.3 Cost Rationale

# 11 Table 1s

Table 11.1: Descriptive Statistics of Validation Data

| Variable | Levels | Defaulted | Did Not Default | Statistic | Parameter | p | Test |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | n = 222 (7.4) | n = 2778 (92.6) |  |  |  |  |
| tot\_credit\_debt |  | 89556.78 (26786.54) | 89332.34 (20248.69) | 0.122083400819777 | 241.604573266439 | 0.902934450461725 | Welch Two Sample t-test |
|  | Missing | 0 (0) | 0 (0) |  |  |  |  |
| avg\_card\_debt |  | 19620 (24905.18) | 13044.33 (6037.51) | 3.92472468575384 | 223.08025290287 | 0.000115683060819151 | Welch Two Sample t-test |
|  | Missing | 0 (0) | 0 (0) |  |  |  |  |
| credit\_age |  | 262.61 (62.37) | 284.97 (62.53) | -5.13838307450266 | 257.791882405844 | 5.47629504321888e-07 | Welch Two Sample t-test |
|  | Missing | 0 (0) | 0 (0) |  |  |  |  |
| credit\_good\_age |  | 139.05 (30.96) | 148.16 (30.63) | -4.22086528840504 | 256.813923360661 | 3.37926816433808e-05 | Welch Two Sample t-test |
|  | Missing | 0 (0) | 0 (0) |  |  |  |  |
| card\_age |  | 232.65 (60.18) | 254.88 (62.49) | -5.27988615455778 | 260.571440407357 | 2.72679108457251e-07 | Welch Two Sample t-test |
|  | Missing | 0 (0) | 0 (0) |  |  |  |  |
| non\_mtg\_acc\_past\_due\_12\_months\_num |  | 0 [0, 2] | 0 [0, 0] | 407327 |  | 4.1492466957563e-62 | Wilcoxon rank sum test with continuity correction |
|  | Missing | 0 (0) | 0 (0) |  |  |  |  |
| non\_mtg\_acc\_past\_due\_6\_months\_num |  | 0 [0, 0] | 0 [0, 0] | 378854 |  | 3.71308270383415e-78 | Wilcoxon rank sum test with continuity correction |
|  | Missing | 0 (0) | 0 (0) |  |  |  |  |
| mortgages\_past\_due\_6\_months\_num |  | 0 [0, 1] | 0 [0, 0] | 388314 |  | 8.68654080752202e-92 | Wilcoxon rank sum test with continuity correction |
|  | Missing | 0 (0) | 0 (0) |  |  |  |  |
| credit\_past\_due\_amount |  | 2833.97 (5200.92) | 136.63 (1243.37) | 7.70978683181569 | 223.02299795649 | 4.13242752939403e-13 | Welch Two Sample t-test |
|  | Missing | 0 (0) | 0 (0) |  |  |  |  |
| inq\_12\_month\_num |  | 1 [0, 2] | 0 [0, 1] | 371269 |  | 4.91584737988872e-09 | Wilcoxon rank sum test with continuity correction |
|  | Missing | 0 (0) | 0 (0) |  |  |  |  |
| card\_inq\_24\_month\_num |  | 1 [0, 3.75] | 0 [0, 2] | 347265.5 |  | 0.000562231349256168 | Wilcoxon rank sum test with continuity correction |
|  | Missing | 0 (0) | 0 (0) |  |  |  |  |
| card\_open\_36\_month\_num |  | 0 [0, 0] | 0 [0, 0] | 318575 |  | 0.213333057664059 | Wilcoxon rank sum test with continuity correction |
|  | Missing | 0 (0) | 0 (0) |  |  |  |  |
| auto\_open\_.36\_month\_num |  | 0 [0, 0] | 0 [0, 0] | 302745 |  | 0.537114808680682 | Wilcoxon rank sum test with continuity correction |
|  | Missing | 0 (0) | 0 (0) |  |  |  |  |
| uti\_card |  | 0.55 (0.1) | 0.49 (0.1) | 8.83252861550425 | 258.107147001719 | 1.605430179006e-16 | Welch Two Sample t-test |
|  | Missing | 0 (0) | 0 (0) |  |  |  |  |
| uti\_50plus\_pct |  | 0.53 (0.11) | 0.48 (0.11) | 6.28386098573946 | 257.444998802966 | 1.40193993897306e-09 | Welch Two Sample t-test |
|  | Missing | 0 (0) | 0 (0) |  |  |  |  |
| uti\_max\_credit\_line |  | 0.5 (0.1) | 0.46 (0.1) | 6.71183073181734 | 257.927169712516 | 1.21500849127748e-10 | Welch Two Sample t-test |
|  | Missing | 0 (0) | 0 (0) |  |  |  |  |
| uti\_card\_50plus\_pct |  | 0.51 (0.1) | 0.45 (0.11) | 7.88449194816807 | 239.278427319455 | 1.11703888980296e-13 | Welch Two Sample t-test |
|  | Missing | 19 (8.6) | 278 (10) |  |  |  |  |
| ind\_acc\_XYZ | Account w/ XYZ | 53 (23.9) | 718 (25.8) | 0.321772517749248 | 1 | 0.570544368859077 | Pearson's Chi-squared test with Yates' continuity correction |
|  | No Account | 169 (76.1) | 2060 (74.2) |  |  |  |  |
|  | Missing | 0 (0) | 0 (0) |  |  |  |  |
| rep\_income |  | 82321.61 (15793.4) | 82939.95 (15704) | -0.53214474079498 | 229.652811556469 | 0.595140187660208 | Welch Two Sample t-test |
|  | Missing | 23 (10.4) | 230 (8.3) |  |  |  |  |
| States | AL | 34 (15.3) | 417 (15) | 3.12344543757152 | 6 | 0.793201318738328 | Pearson's Chi-squared test |
|  | FL | 28 (12.6) | 405 (14.6) |  |  |  |  |
|  | GA | 29 (13.1) | 371 (13.4) |  |  |  |  |
|  | LA | 25 (11.3) | 388 (14) |  |  |  |  |
|  | MS | 39 (17.6) | 422 (15.2) |  |  |  |  |
|  | NC | 32 (14.4) | 397 (14.3) |  |  |  |  |
|  | SC | 35 (15.8) | 378 (13.6) |  |  |  |  |
|  | Missing | 0 (0) | 0 (0) |  |  |  |  |

Table 11.1: Descriptive Statistics of Test Data

| Variable | Levels | Defaulted | Did Not Default | Statistic | Parameter | p | Test |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | n = 401 (8) | n = 4599 (92) |  |  |  |  |
| tot\_credit\_debt |  | 92382.69 (30861.3) | 94791.52 (23043.76) | -1.52635947630644 | 439.745938385848 | 0.127639255912994 | Welch Two Sample t-test |
|  | Missing | 0 (0) | 0 (0) |  |  |  |  |
| avg\_card\_debt |  | 20132.24 (25007.03) | 13760.17 (6783.76) | 5.08629805500289 | 405.14820409324 | 5.59425374140873e-07 | Welch Two Sample t-test |
|  | Missing | 0 (0) | 0 (0) |  |  |  |  |
| credit\_age |  | 275.08 (60.68) | 299.41 (60.72) | -7.69899999283012 | 472.585306347121 | 8.0924808654733e-14 | Welch Two Sample t-test |
|  | Missing | 0 (0) | 0 (0) |  |  |  |  |
| credit\_good\_age |  | 140.78 (32.6) | 151.2 (33.29) | -6.13030917803221 | 475.728204579475 | 1.84357232459723e-09 | Welch Two Sample t-test |
|  | Missing | 0 (0) | 0 (0) |  |  |  |  |
| card\_age |  | 248.9 (60.94) | 270.22 (58.27) | -6.74219544155584 | 466.064196770262 | 4.62770568868965e-11 | Welch Two Sample t-test |
|  | Missing | 0 (0) | 0 (0) |  |  |  |  |
| non\_mtg\_acc\_past\_due\_12\_months\_num |  | 0 [0, 2] | 0 [0, 0] | 1229079 |  | 1.04424409163499e-125 | Wilcoxon rank sum test with continuity correction |
|  | Missing | 0 (0) | 0 (0) |  |  |  |  |
| non\_mtg\_acc\_past\_due\_6\_months\_num |  | 0 [0, 0] | 0 [0, 0] | 1141015.5 |  | 3.58081572133281e-179 | Wilcoxon rank sum test with continuity correction |
|  | Missing | 0 (0) | 0 (0) |  |  |  |  |
| mortgages\_past\_due\_6\_months\_num |  | 0 [0, 1] | 0 [0, 0] | 1159925.5 |  | 9.70317336530627e-188 | Wilcoxon rank sum test with continuity correction |
|  | Missing | 0 (0) | 0 (0) |  |  |  |  |
| credit\_past\_due\_amount |  | 2918.61 (5546.8) | 82.03 (941.01) | 10.2277497264292 | 402.009894907583 | 5.60495159214712e-22 | Welch Two Sample t-test |
|  | Missing | 0 (0) | 0 (0) |  |  |  |  |
| inq\_12\_month\_num |  | 2 [1, 3] | 1 [0, 3] | 1052994 |  | 1.23735039481109e-06 | Wilcoxon rank sum test with continuity correction |
|  | Missing | 0 (0) | 0 (0) |  |  |  |  |
| card\_inq\_24\_month\_num |  | 4 [1, 6] | 3 [1, 5] | 1026505.5 |  | 0.000146420742989445 | Wilcoxon rank sum test with continuity correction |
|  | Missing | 0 (0) | 0 (0) |  |  |  |  |
| card\_open\_36\_month\_num |  | 0 [0, 0] | 0 [0, 0] | 951673.5 |  | 0.0857469907792185 | Wilcoxon rank sum test with continuity correction |
|  | Missing | 0 (0) | 0 (0) |  |  |  |  |
| auto\_open\_.36\_month\_num |  | 0 [0, 0] | 0 [0, 0] | 926770 |  | 0.783828670614223 | Wilcoxon rank sum test with continuity correction |
|  | Missing | 0 (0) | 0 (0) |  |  |  |  |
| uti\_card |  | 0.56 (0.12) | 0.5 (0.11) | 10.8939093753159 | 463.754142500041 | 9.21668948746017e-25 | Welch Two Sample t-test |
|  | Missing | 0 (0) | 0 (0) |  |  |  |  |
| uti\_50plus\_pct |  | 0.56 (0.11) | 0.51 (0.11) | 9.38377124957334 | 471.818806230872 | 2.69212908555055e-19 | Welch Two Sample t-test |
|  | Missing | 0 (0) | 0 (0) |  |  |  |  |
| uti\_max\_credit\_line |  | 0.55 (0.11) | 0.51 (0.11) | 8.37420345619788 | 472.756415666828 | 6.36427101761314e-16 | Welch Two Sample t-test |
|  | Missing | 0 (0) | 0 (0) |  |  |  |  |
| uti\_card\_50plus\_pct |  | 0.55 (0.13) | 0.48 (0.12) | 9.58141384893526 | 427.980576652133 | 7.89445892054665e-20 | Welch Two Sample t-test |
|  | Missing | 29 (7.2) | 470 (10.2) |  |  |  |  |
| ind\_acc\_XYZ | Account w/ XYZ | 67 (16.7) | 1211 (26.3) | 17.4510633868205 | 1 | 2.94799334094036e-05 | Pearson's Chi-squared test with Yates' continuity correction |
|  | No Account | 334 (83.3) | 3388 (73.7) |  |  |  |  |
|  | Missing | 0 (0) | 0 (0) |  |  |  |  |
| rep\_income |  | 75896.74 (15265.15) | 75423.39 (16154.82) | 0.567934321586526 | 441.290880074828 | 0.570368408655156 | Welch Two Sample t-test |
|  | Missing | 33 (8.2) | 350 (7.6) |  |  |  |  |
| States | AL | 57 (14.2) | 681 (14.8) | 9.00387088365482 | 6 | 0.173360492152025 | Pearson's Chi-squared test |
|  | FL | 49 (12.2) | 639 (13.9) |  |  |  |  |
|  | GA | 64 (16) | 677 (14.7) |  |  |  |  |
|  | LA | 64 (16) | 619 (13.5) |  |  |  |  |
|  | MS | 52 (13) | 706 (15.4) |  |  |  |  |
|  | NC | 44 (11) | 620 (13.5) |  |  |  |  |
|  | SC | 71 (17.7) | 657 (14.3) |  |  |  |  |
|  | Missing | 0 (0) | 0 (0) |  |  |  |  |

1. Department of Epidemiology and Biostatistics [↑](#footnote-ref-20)
2. Department of Computer Science [↑](#footnote-ref-21)