Credit Card Default Prediction Using Data Science & Machine Learning

Cody Dailey
Ishaan Dave
Yang Ge
Vipul Shinde

Summary

- Introduction
- Exploratory Data Analysis
- Classification
  - Logistic Regression
  - Random Forest Classifier
- Predictions
- Conclusions

Introduction

- Credit risk plays a major role in the banking industry business. Banks main activities involve granting loan, credit card, investment, mortgage, and others.
- Credit card has been one of the most booming financial services by banks over the past years. However, with the growing number of credit card users, banks have been facing an escalating credit card default rate.
- As such data science and machine learning can provide solutions to tackle the current phenomenon and management credit risks.
- Thus, Logistic Regression and Random Forest Classifier models are used to predict if the applicants for the credit card will default or not soon based on the given input data.

Exploratory Data Analysis

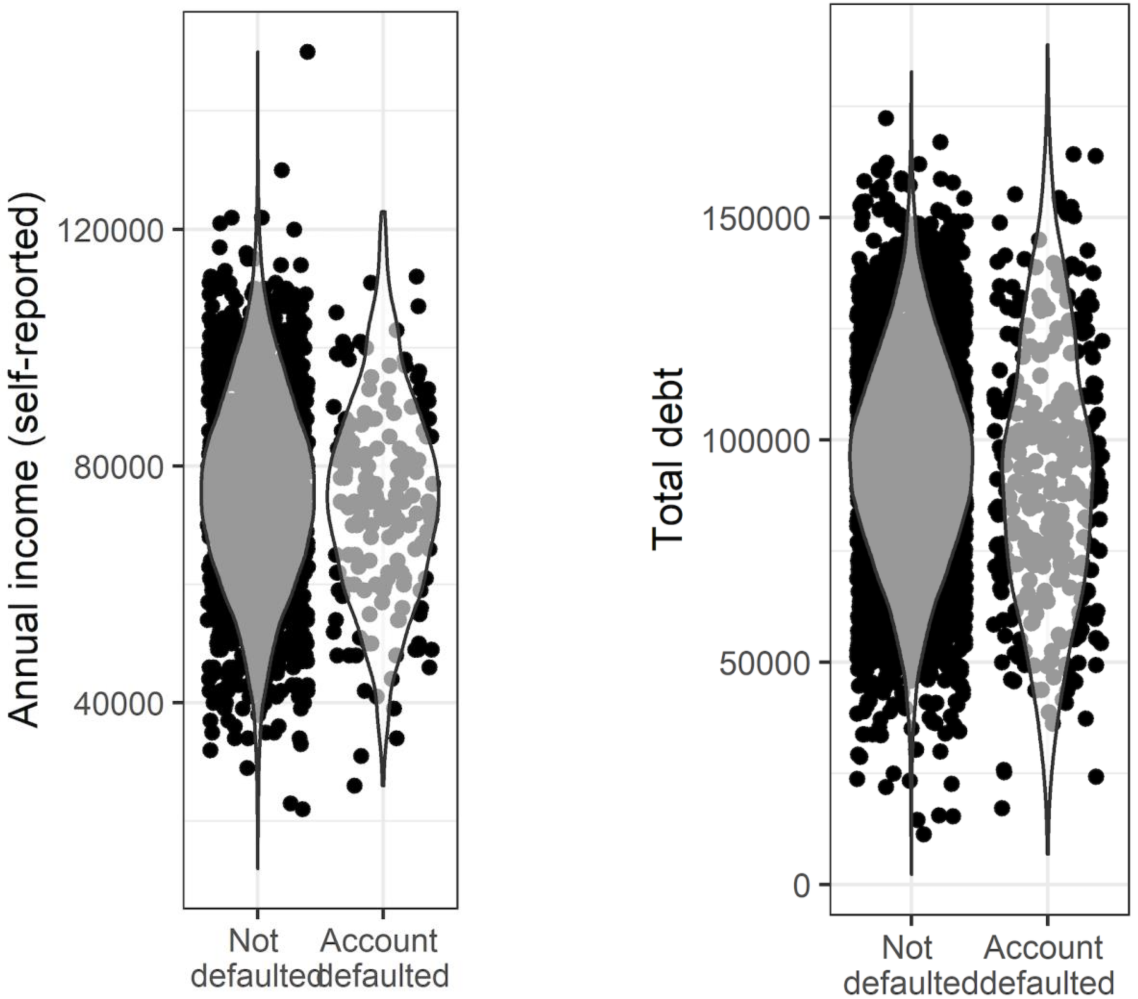
3 datasets

- Training – used to create model
- Validation – qualify performance
- Test – evaluate final model performance

Exploratory Data Analysis

Univariate Statistics

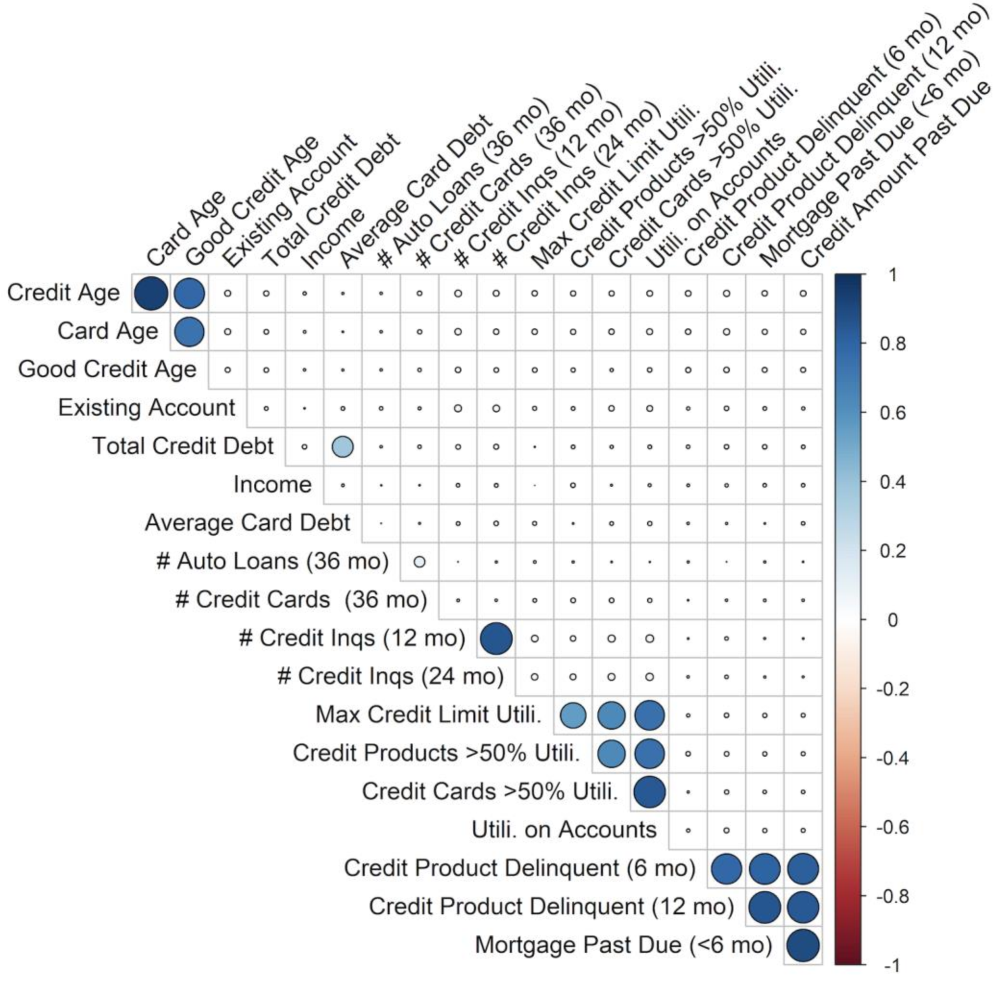
- Mean +/- SD and normality for continuous
- N (%) for categorical
- Median [IQR] for discrete counts
- N missing for each variable
- Compared between outcome groups (Did / did not default)



Exploratory Data Analysis

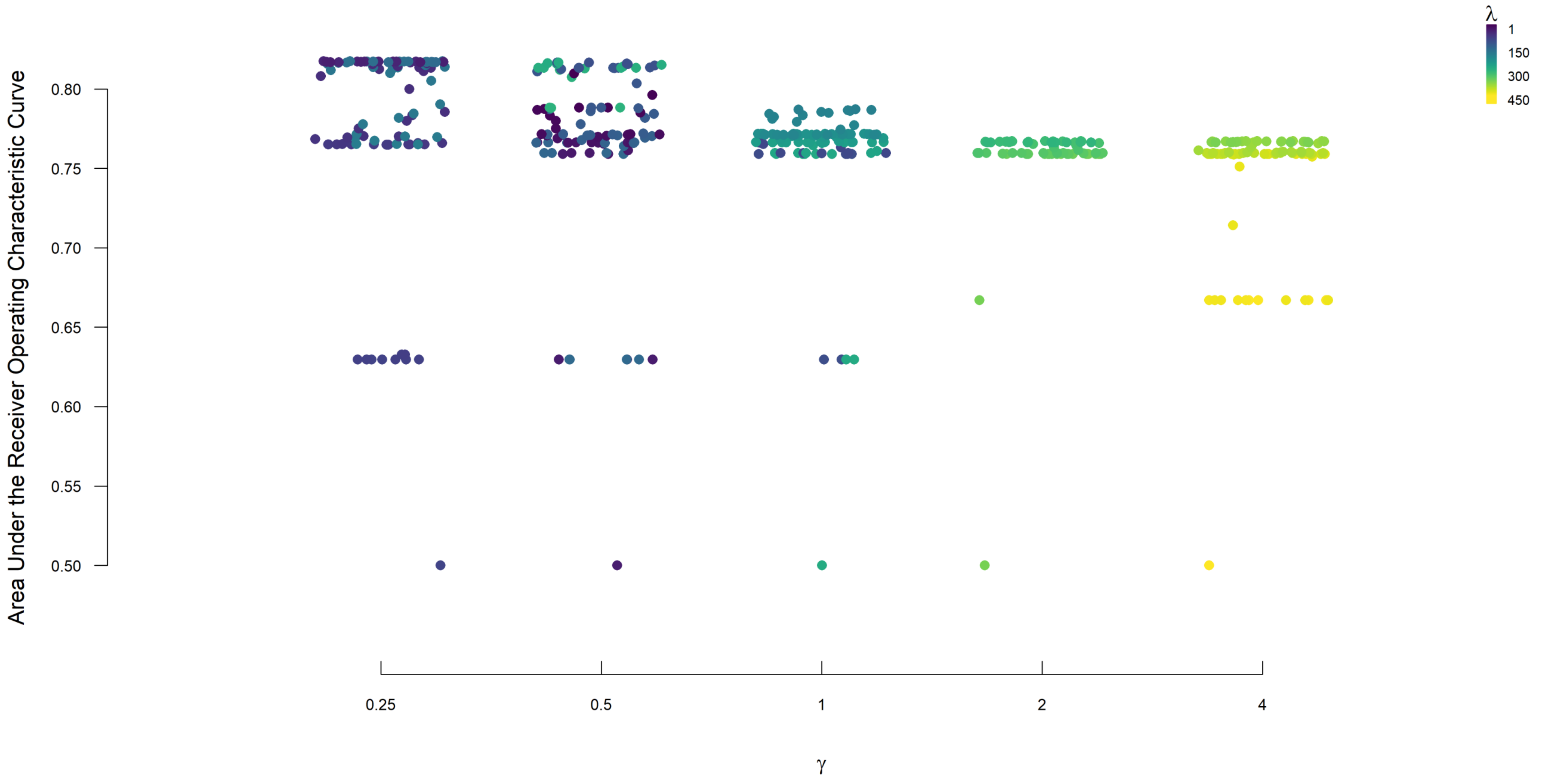
Correlations between variables

- 20 potential predictors
- Suspected multicollinearity
- Correlation matrix to examine what variables may vary together
-

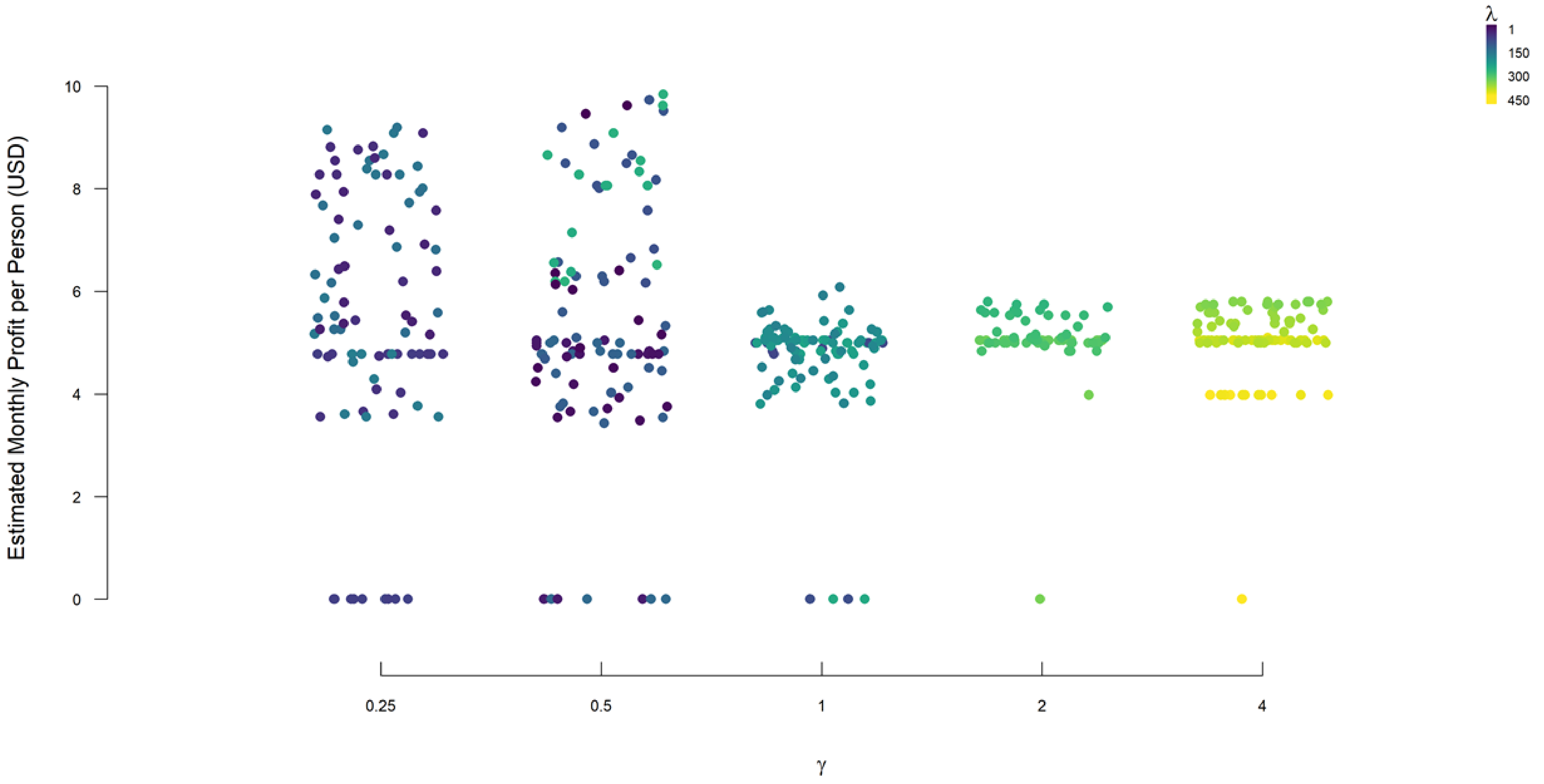


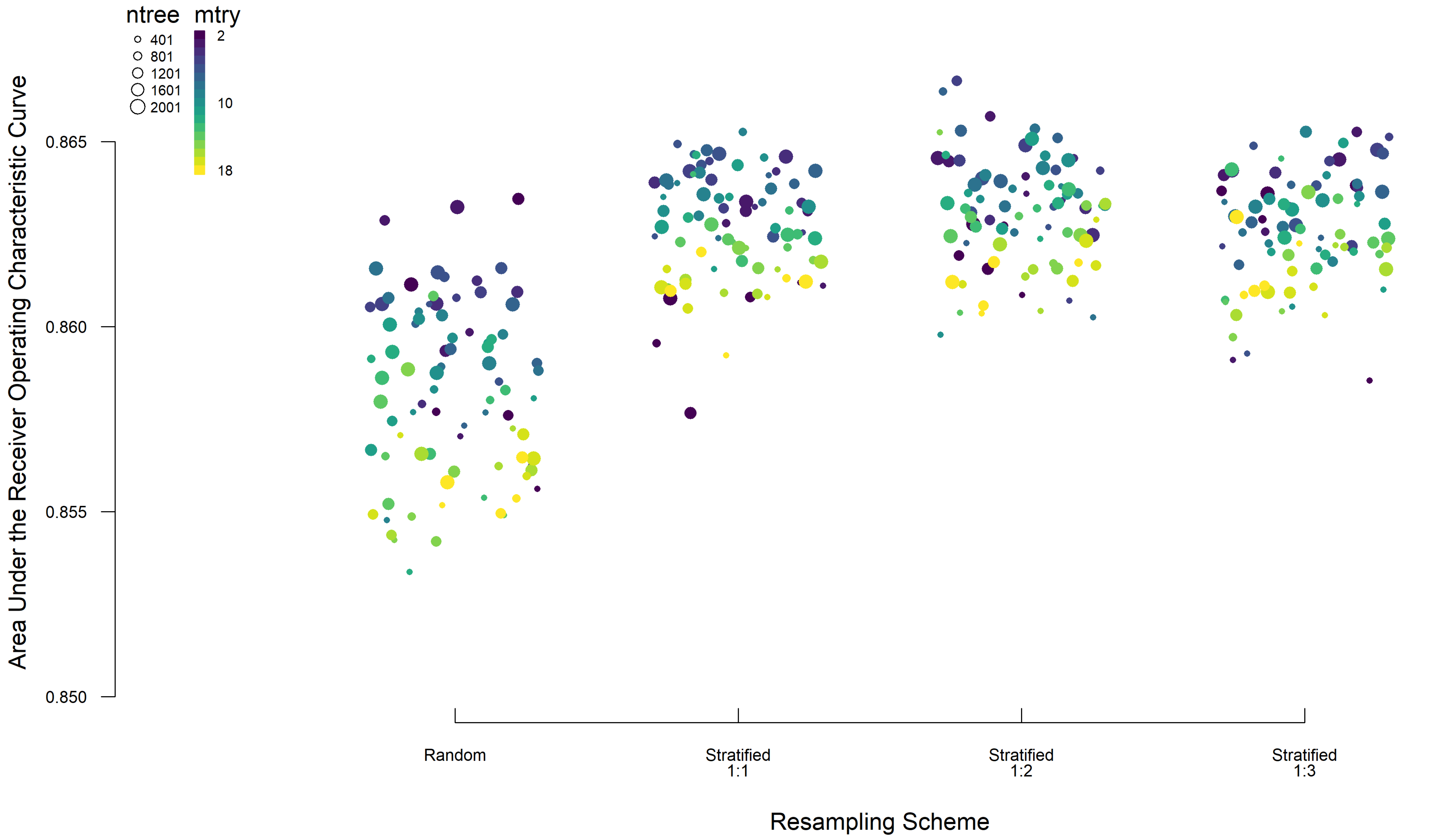
Feature Selection

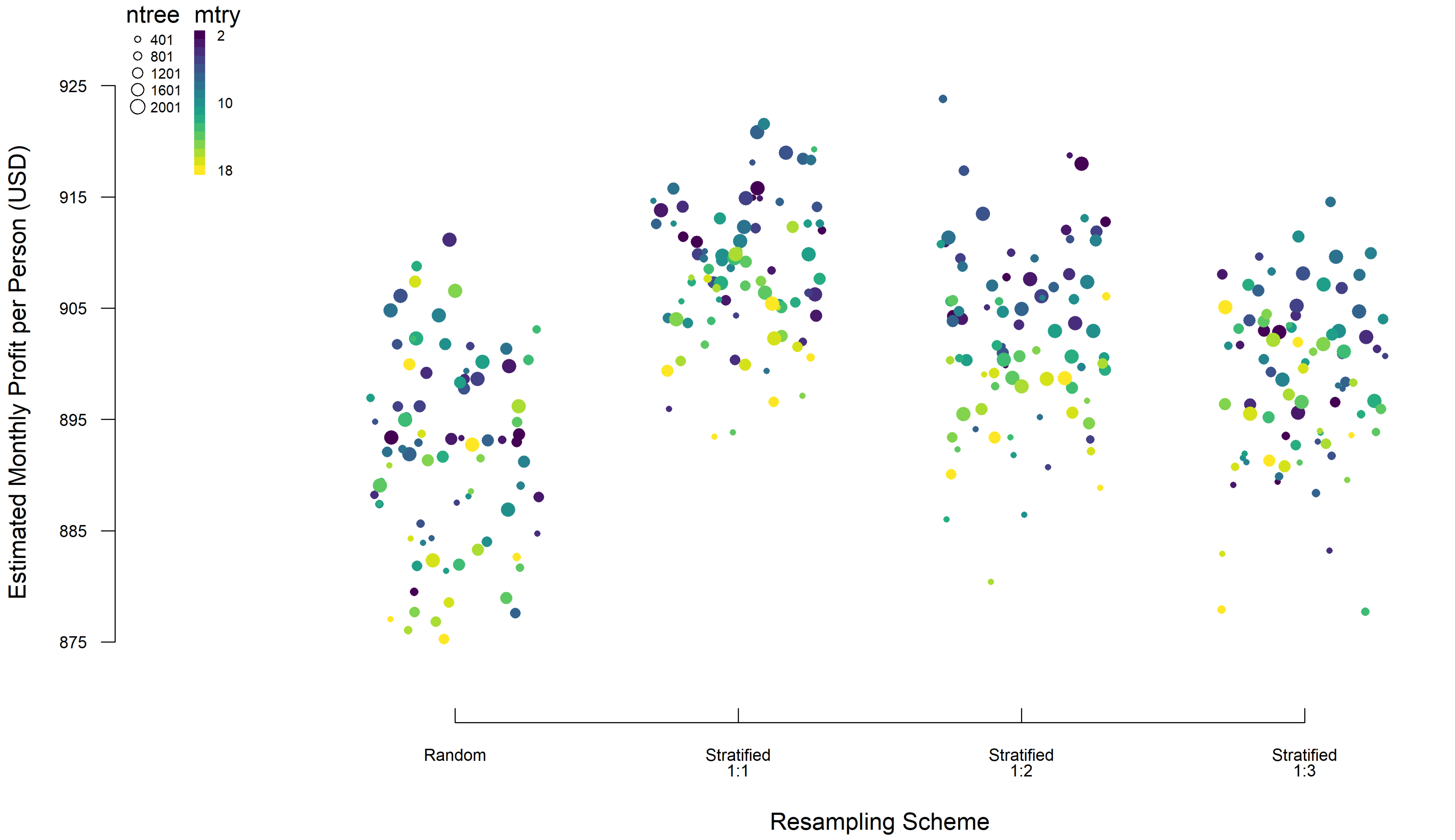
- We want to select covariates that are important in predicting whether or an individual will default or not and "penalize" (shrink) those that are not important
- Adaptive Least Absolute Shrinkage and Selection Operator (LASSO) Logistic Regression was performed to select covariates and predict binary outcome
- $logit(E(Y|\boldsymbol{X}) = \boldsymbol{\beta}^{\boldsymbol{T}}(\boldsymbol{X}),$
  - Y is the binary outcome (whether or not an individual defaulted)
    - **X** = $(1, X_1, \ldots, X_p)$ is the vector of covariate values
    - $\boldsymbol{\beta}$ = $(\beta_0, \beta_1, \ldots, \beta_p)$ vector of regression parameters we want to estimate
- Estimation is an optimization problem in the form L$(\boldsymbol{\beta}) - \lambda * pen(\boldsymbol{\beta})$, where $pen(\boldsymbol{\beta})$ is the "penalty" and λ is the "tuning parameter" (amount of shrinkage)



Area Under the Receiver Operating Characteristic Curve
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0.25
0.5
1
2
4
γ
λ
1
150
300
450


Estimated Monthly Profit per Person (USD)
10
8
6
4
2
0
0.25
0.5
1
2
4
γ
λ
1
150
300
450


ntree
401
801
1201
1601
2001
mtry
2
10
18
Area Under the Receiver Operating Characteristic Curve
0.850
0.855
0.860
0.865
Random
Stratified 1:1
Stratified 1:2
Stratified 1:3
Resampling Scheme



ntree
mtry
401
801
1201
1601
2001
2
10
18
925
915
905
895
885
875
Estimated Monthly Profit per Person (USD)
Random
Stratified 1:1
Stratified 1:2
Stratified 1:3
Resampling Scheme

Predictions and Comparisons

- Logistic Regression

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

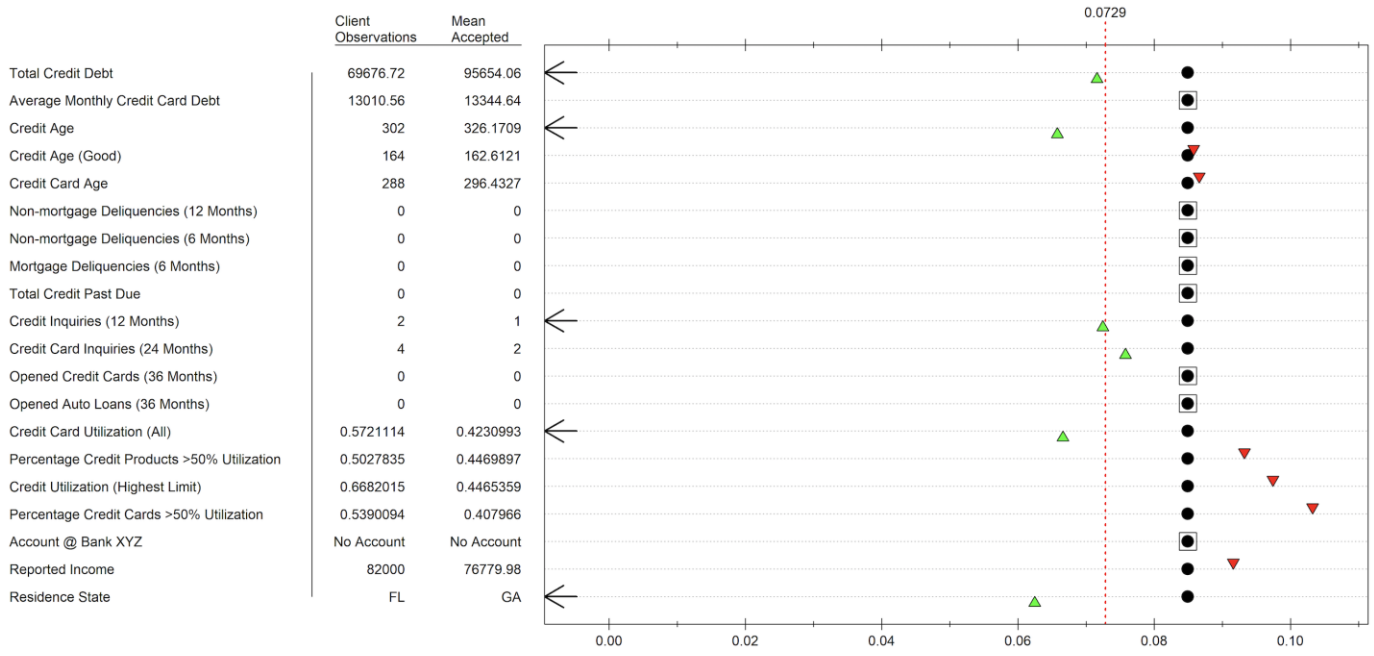
- Random Forest

		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

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

Conclusions

- Rejection example based on Random Forest that shows how a change in a variable impacts model performance.
- As seen in figure to the right, those who already have an account with bank XYZ do not receive favorable treatment



Thank you!