# Taiwan Credit Defaults

Henry Chang | Avani Sharma
Atindra Bandi | Abraham Khan
Group 14

# Situation

Taiwan economy grew 95% from 1990-2000

Banks loosened credit requirements to continue growth

People started borrowing more than they could pay

# Problem Statement

TEXAS McCombs

Decision: Identify high risk customers based their credit history

Key Questions:

How to identify potential defaulters?

What are the factors leading to potential default?

# Dataset

9 Categorical variables

Sex, Education, Marriage, and Payment status for 6 months

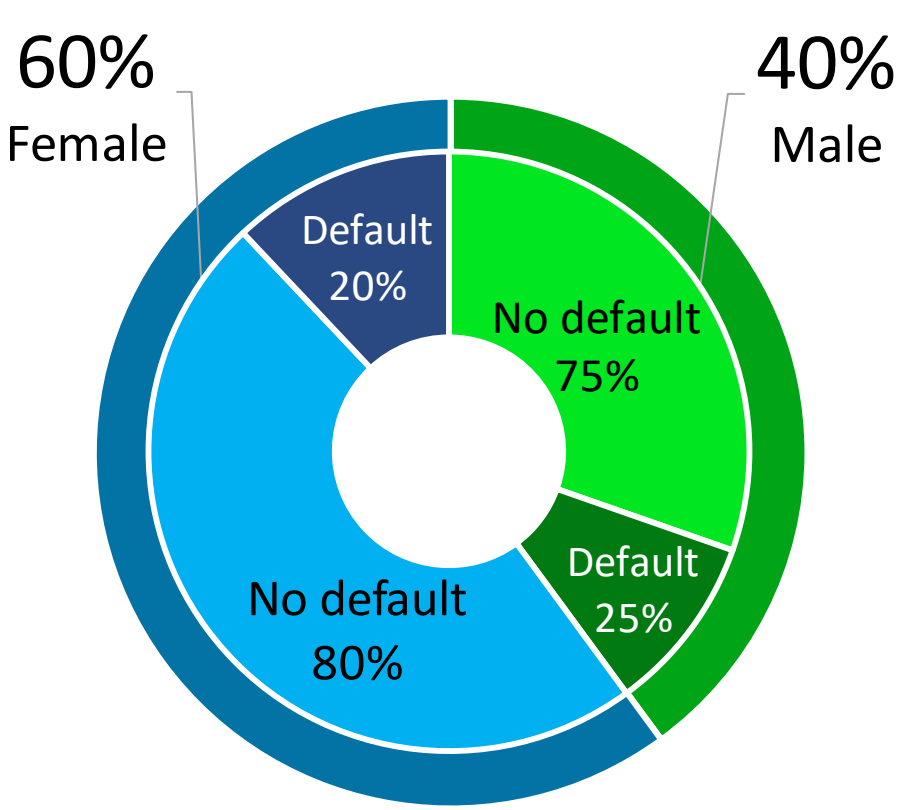
14 Numerical variables

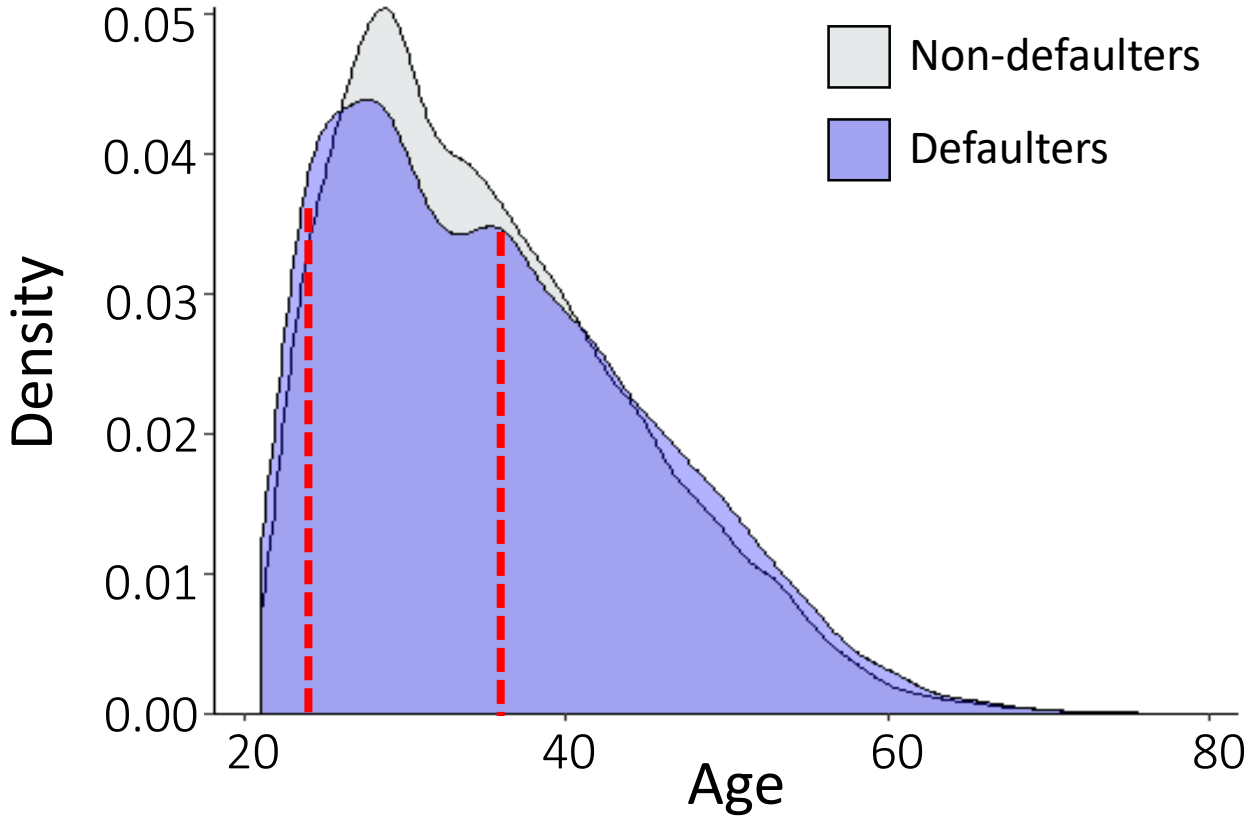
Age, Credit limit, Balance and Payment amounts for 6 months

> Predict default on credit card payment next month

Data Source: https://archive.ics.uci.edu/ml/datasets/default%20of%20credit%20card%20clients

# Demographics

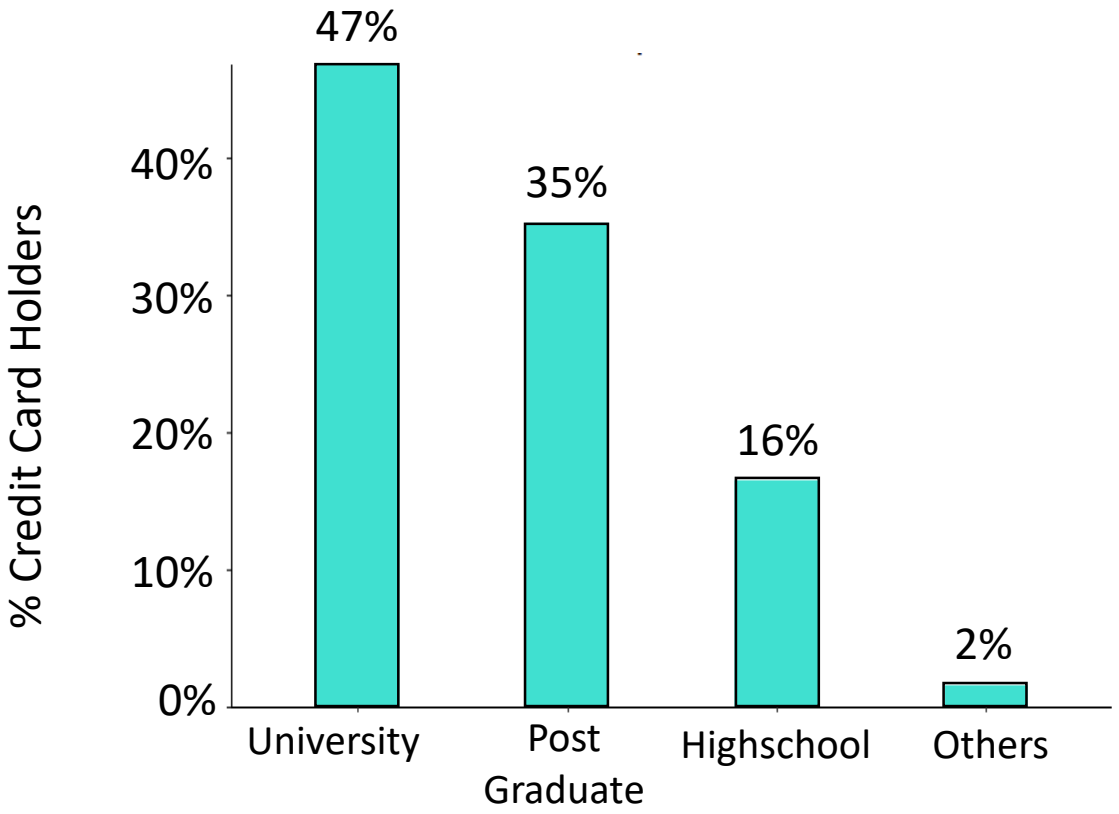


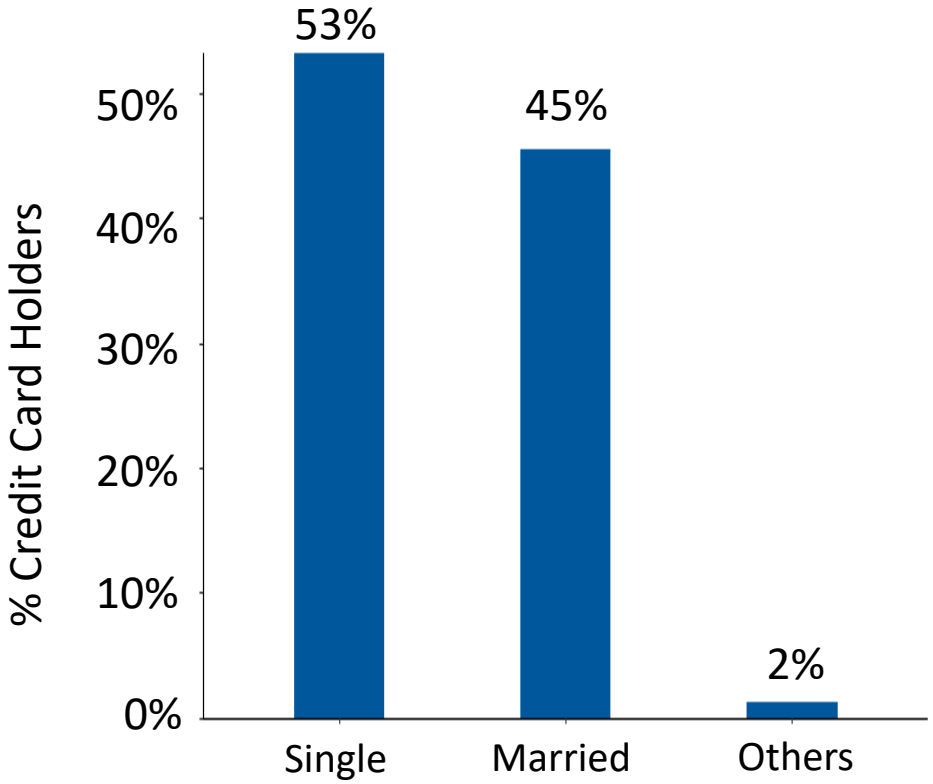
# Demographics

TEXAS McCombs

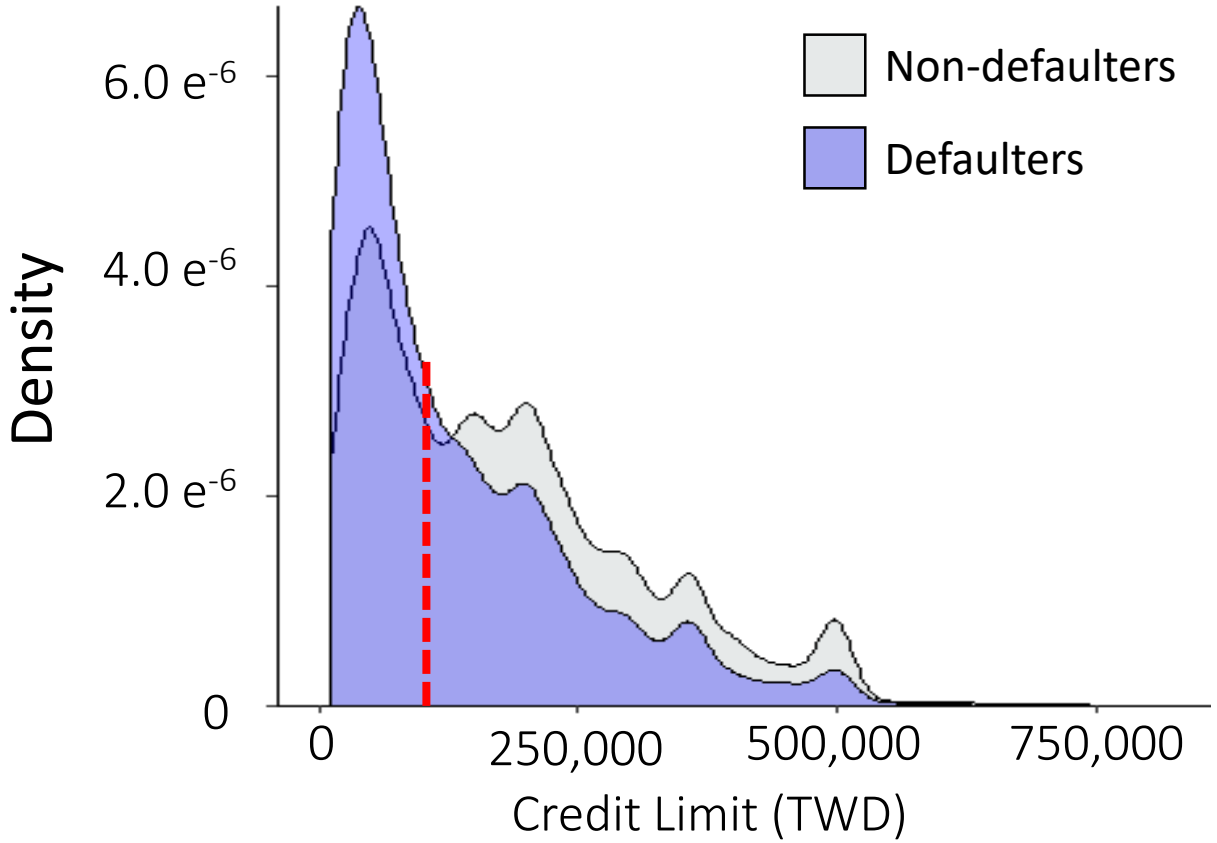
## Education




## Marital Status



# Credit Limits by Default Status




**17**% of customers with credit limit **above** TWD 100,000 defaulted

**29**% of customers with credit limit **below** TWD 100,000 defaulted

# Variable Creation

TEXAS McCombs

## Payment – Spending Ratio

$$\frac{\sum Payments}{\sum Spending}$$

## Weighted Payment Score

$(w_1 \cdot 1^{st}\ \text{Month Status}) + (w_2 \cdot 2^{nd}\ \text{Month Status}) + (w_3 \cdot 3^{rd}\ \text{Month Status})$
$+ (w_4 \cdot 4^{th}\ \text{Month Status}) + (w_5 \cdot 5^{th}\ \text{Month Status}) + (w_6 \cdot 6^{th}\ \text{Month Status})$

# Variable Selection

TEXAS McCombs

Logistic regression

- 10 fold cross validation, 20 times

Select variables that remained most often

- Credit limit
- Recent payment amounts
- Recent delayed payments
- Age of customer

# Lasso Logistic Output

- Credit limit
- Recent payment amounts
- Recent delayed payments
- Age of customer

| | Estimate | Std. Error | z value | Pr(>\|z\|) | |
|---|---|---|---|---|---|
| (Intercept) | -1.499e+00 | 1.605e-01 | -9.340 | < 2e-16 | *** |
| LIMIT_BAL | -8.423e-07 | 1.783e-07 | -4.725 | 2.30e-06 | *** |
| PAY_AMT1 | -1.436e-05 | 2.787e-06 | -5.152 | 2.58e-07 | *** |
| PAY_AMT2 | -9.519e-06 | 2.204e-06 | -4.318 | 1.57e-05 | *** |
| PAY_AMT3 | -4.109e-06 | 1.812e-06 | -2.268 | 0.02332 | * |
| PAY_AMT4 | -2.830e-06 | 1.656e-06 | -1.709 | 0.08742 | . |
| PAY_AMT5 | -4.612e-06 | 1.810e-06 | -2.549 | 0.01082 | * |
| PAY_AMT6 | -2.999e-06 | 1.468e-06 | -2.043 | 0.04103 | * |
| PAY_1 | 5.642e-01 | 1.980e-02 | 28.488 | < 2e-16 | *** |
| PAY_2 | 6.601e-02 | 2.261e-02 | 2.919 | 0.00351 | ** |
| PAY_3 | 7.408e-02 | 2.523e-02 | 2.936 | 0.00332 | ** |
| PAY_4 | 4.856e-02 | 2.768e-02 | 1.754 | 0.07941 | . |
| PAY_5 | 3.180e-02 | 2.991e-02 | 1.063 | 0.28764 | |
| PAY_6 | 7.908e-03 | 2.431e-02 | 0.325 | 0.74491 | |
| pay_spend_ratio | -2.786e-04 | 7.217e-04 | -0.386 | 0.69945 | |
| BILL_AMT2 | -1.349e-06 | 3.225e-07 | -4.184 | 2.87e-05 | *** |
| AGE | 1.700e-02 | 4.192e-03 | 4.055 | 5.01e-05 | *** |
| SEXMale | 1.462e-01 | 5.060e-02 | 2.889 | 0.00387 | ** |
| EDUCATIONHigh School | 5.378e-01 | 1.987e-01 | 2.707 | 0.00680 | ** |
| EDUCATIONOthers | -1.951e+00 | 8.660e-01 | -2.253 | 0.02428 | * |
| EDUCATIONUniversity | 2.877e-01 | 1.533e-01 | 1.877 | 0.06057 | . |
| MARRIAGEOthers | 3.301e-01 | 6.799e-01 | 0.486 | 0.62729 | |
| MARRIAGESingle | 1.653e-02 | 1.497e-01 | 0.110 | 0.91205 | |
| SEXMale:MARRIAGEOthers | 3.122e-01 | 3.040e-01 | 1.027 | 0.30439 | |
| SEXMale:MARRIAGESingle | -1.139e-01 | 6.934e-02 | -1.643 | 0.10039 | |
| AGE:MARRIAGEOthers | -1.904e-02 | 1.580e-02 | -1.205 | 0.22814 | |
| AGE:MARRIAGESingle | -3.840e-03 | 4.085e-03 | -0.940 | 0.34716 | |
| AGE:EDUCATIONHigh School | -1.634e-02 | 5.112e-03 | -3.196 | 0.00139 | ** |
| AGE:EDUCATIONOthers | 1.978e-02 | 2.180e-02 | 0.907 | 0.36429 | |
| AGE:EDUCATIONUniversity | -1.111e-02 | 4.311e-03 | -2.576 | 0.00999 | ** |

# Model Comparison

Sensitivity = true positive rate

Specificity = true negative rate

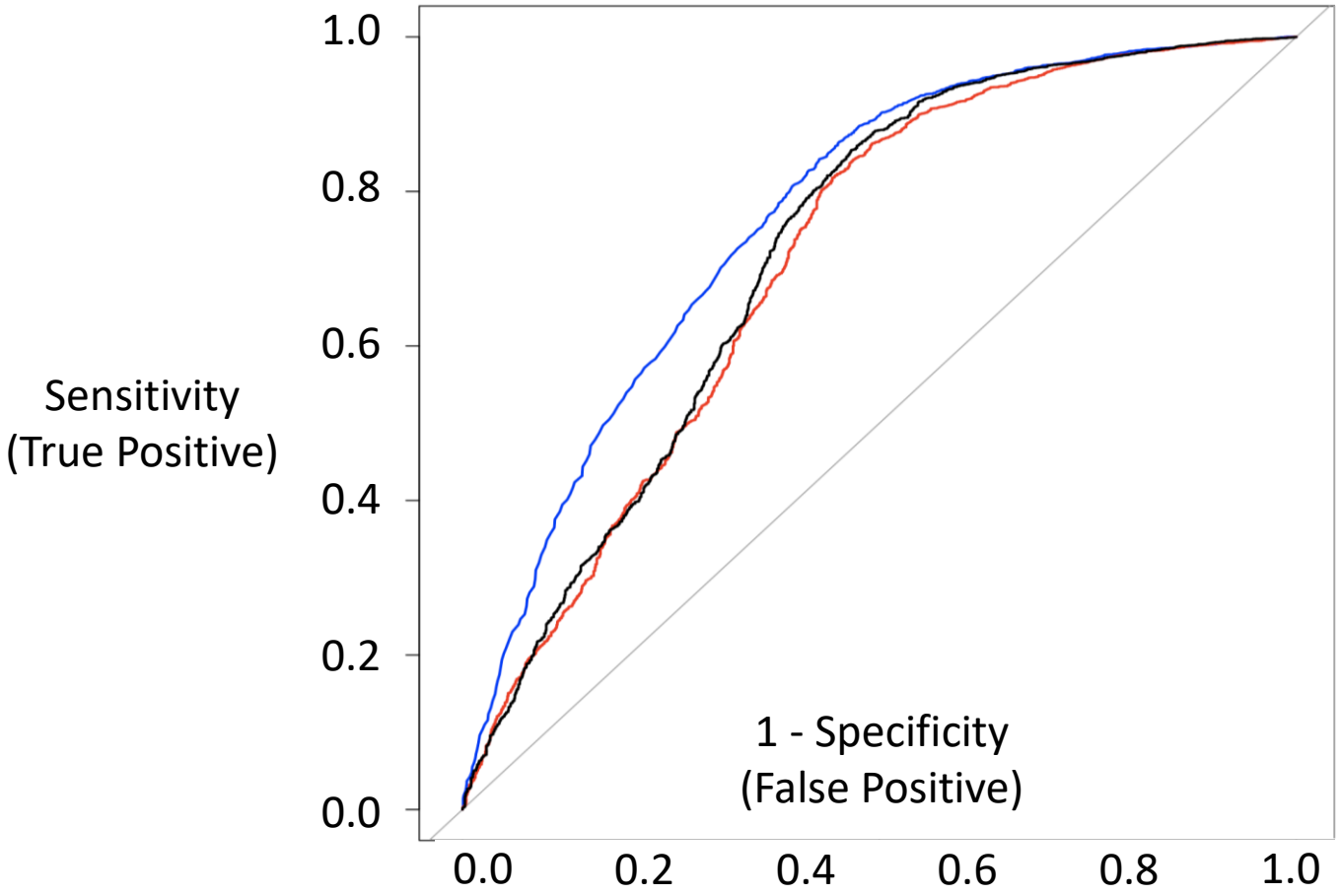
Accuracy = correct prediction rate

AUC = area under ROC curve

Cut-off = threshold for calculating default

| Models | Logistic Regression | Naïve Bayes | Random Forest |
|---|---|---|---|
| Accuracy | 80% | 80% | 80% |
| Sensitivity | 15% | 46% | 53% |
| Specificity | 98% | 90% | 88% |
| AUC | 73% | 73% | 76% |
| Cut-off | 0.57 | 0.97 | 0.32 |

# Receiver Operating Curve




Random Forest AUC = 76.8%

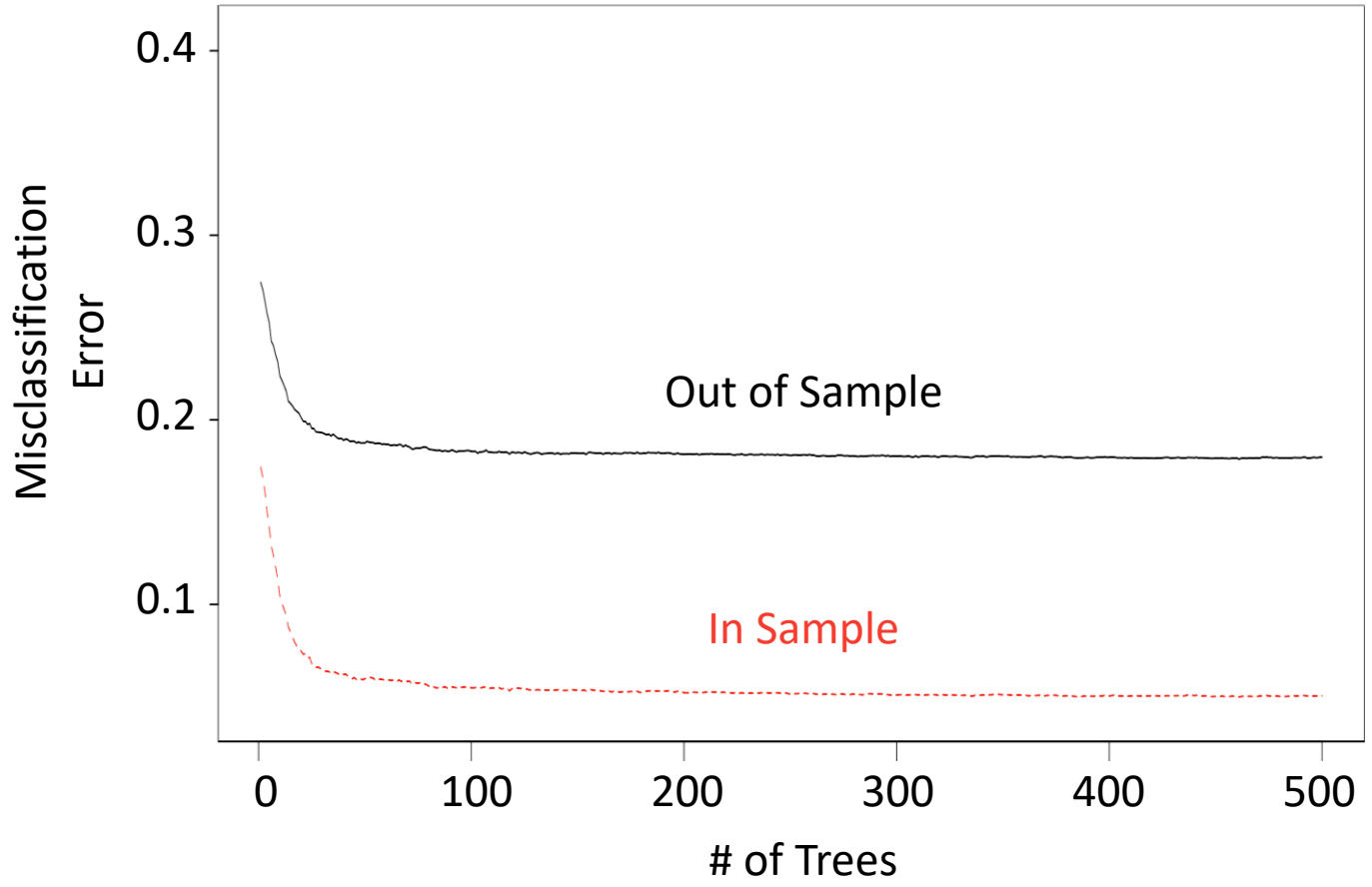
Logistic Regression AUC = 72.3%

Naïve Bayes AUC = 71.3%

# Random Forest

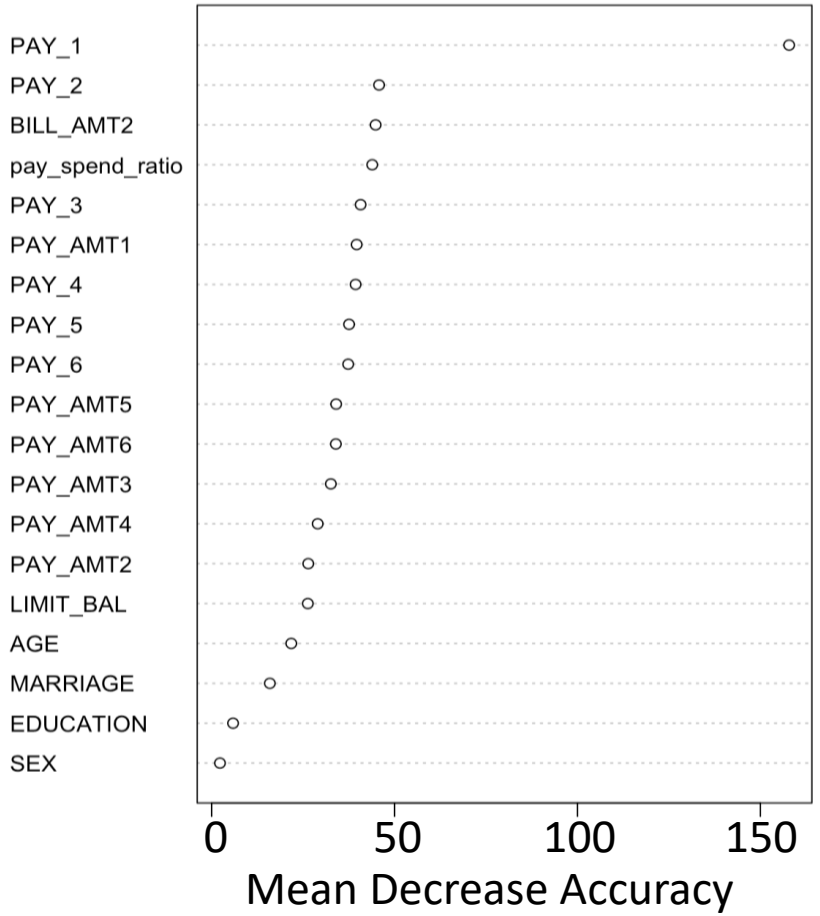
Run cross validation to optimize parameters

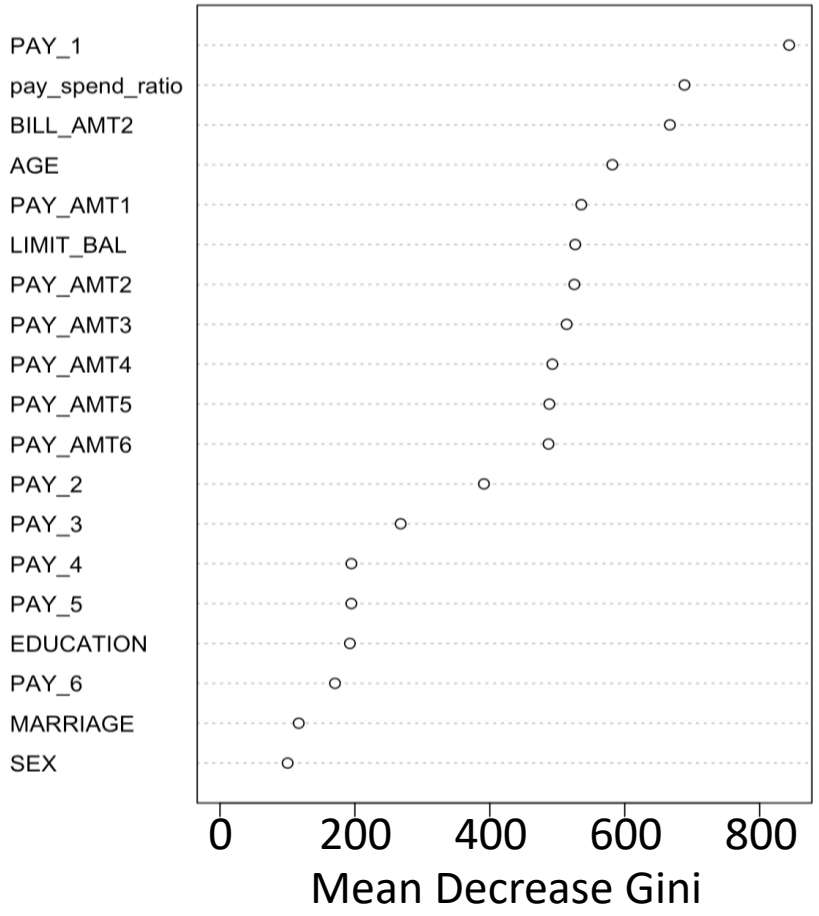
- Threshold
- Number of trees
- Number of variables



# Random Forest – Variable Importance

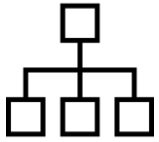
TEXAS McCombs



# Key Takeaways

TEXAS McCombs



Random forest has the best balance between TPR and FPR

Recent payments are the most important variables for prediction

Demographic variables are not important predictors for defaulting

Feature engineering can be tricky, but insightful


That's all Folks!