45-952 ETE Analytics
Final Presentation

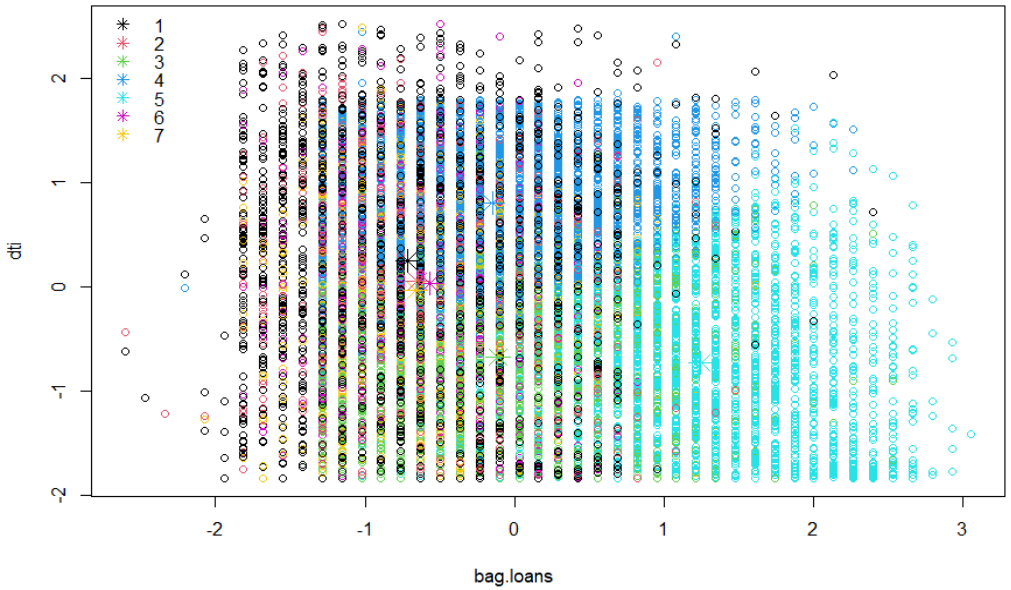
PT Team 2

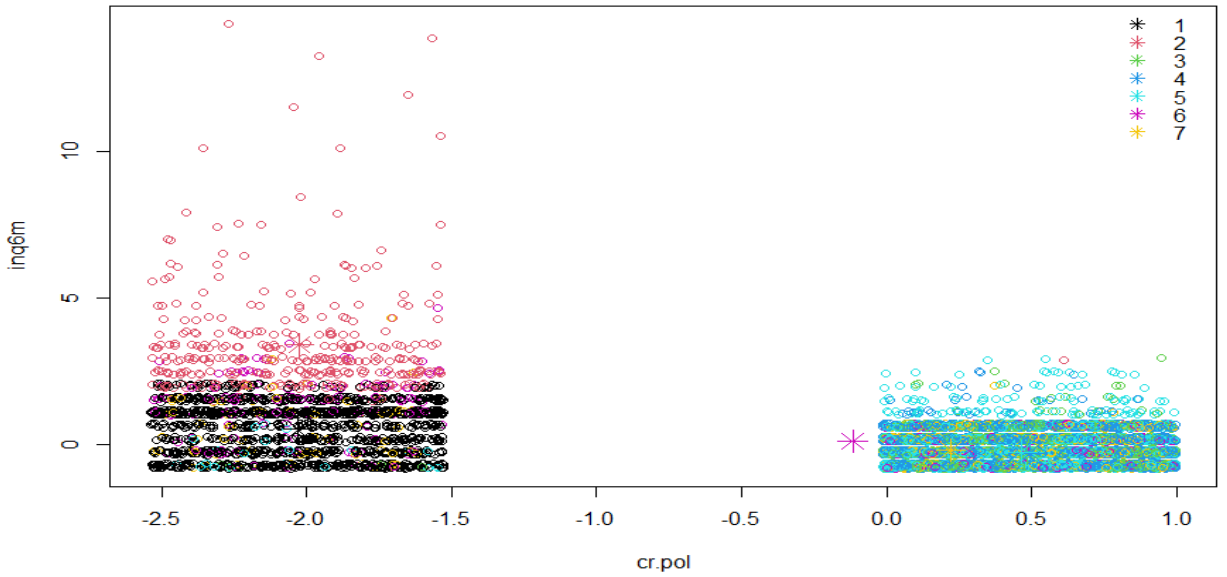
Amira Roux and Luke Profio

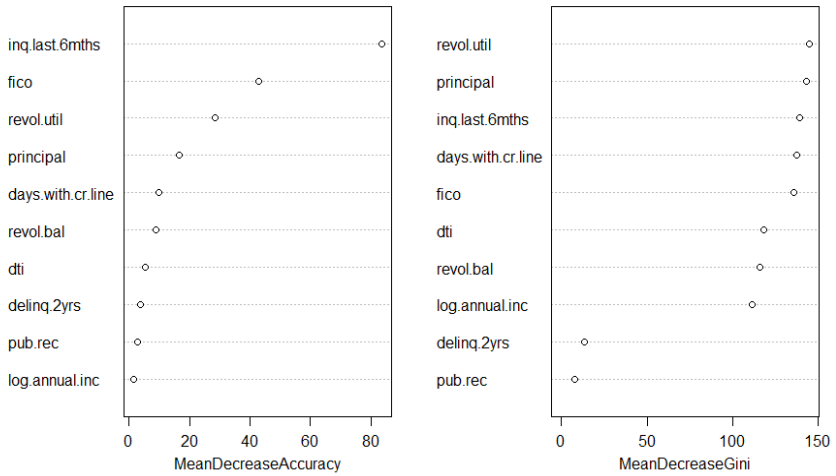
Summary of our method & Final Graphs:

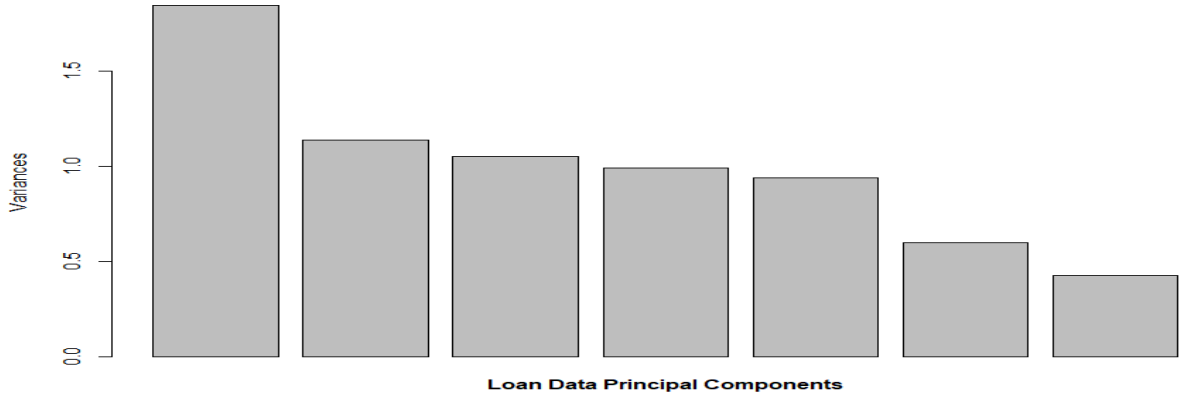
- 1. Summary Graphs:
- - Generated summary graphs, including barplots and correlation heatmaps, to provide a concise visual overview of the dataset.
-
- 2. Model Performance Graphs:
- - Presented ROC curves with AUC values to showcase the performance of both logistic regression and decision tree models.

Summary of our Graphs



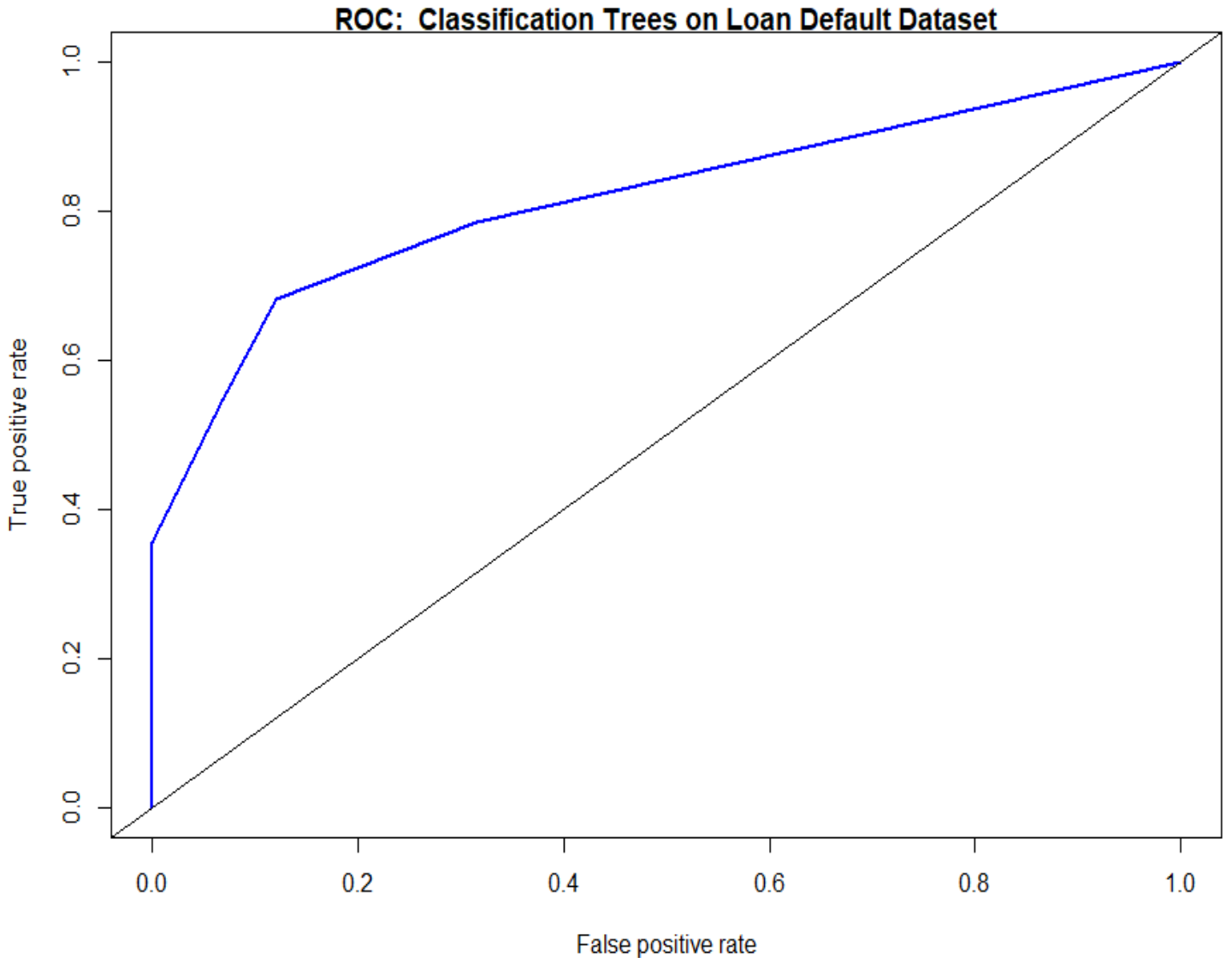






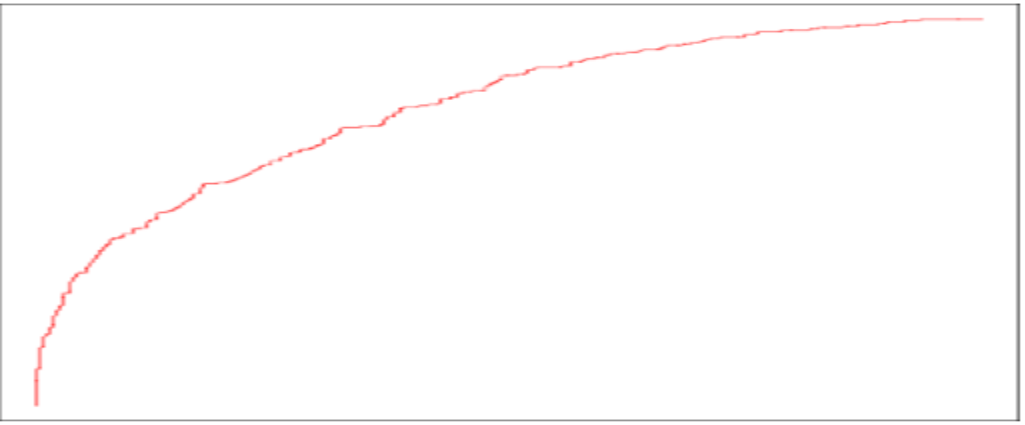
Logistic Regression Modeling:

- 1. Variable Selection:
- - Logistic regression models were built to predict loan defaults.
- - Significant predictors were identified based on p-values.
- - Variable selection was performed by excluding correlated variables and using stepwise regression.
-
- 2. Model Evaluation:
- - The performance of the logistic regression model was evaluated using ROC curves and the Area Under Curve (AUC).
- 3. Visualization of Model Response:
- - The impact of variables on the probability of default was visualized using the plotmo package.



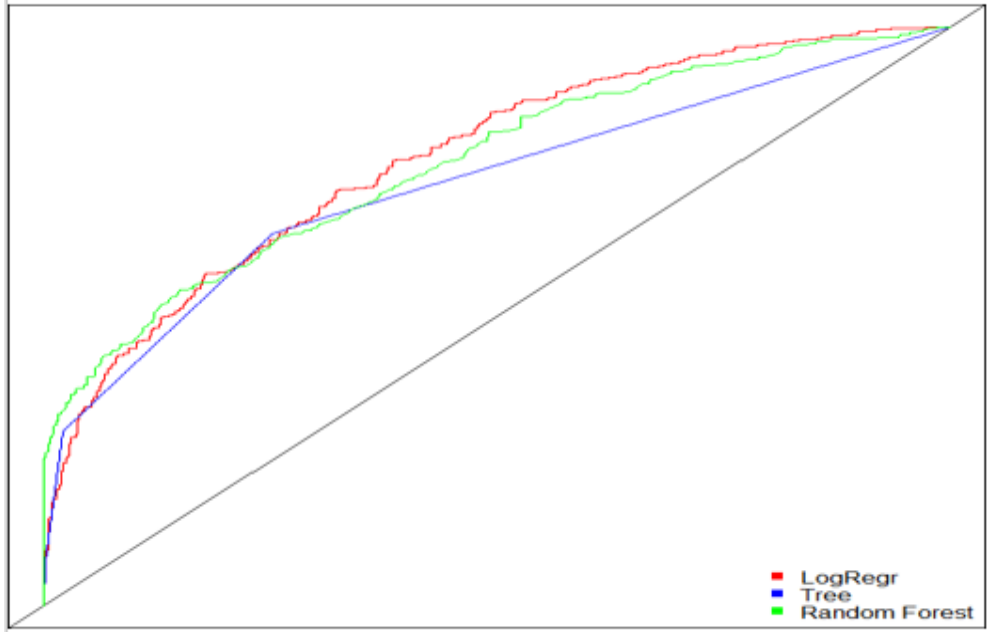
Continuation...

Graphical representantion of defaulting probability by the lr model



1. The logistic regression (LR) model is effective, with statistically significant coefficients for key features such as interest rate, FICO score, and revolving utilization.
2. LR model's robust fit is demonstrated by deviance metrics like lower residual deviance compared to null deviance.
3. The random forest (RF) model assessment uses MeanDecreaseGini values for variable importance.
4. Comparing performance metrics, such as accuracy, precision, recall, and F1-score along with confusion matrices, can give insights into both models' performances.
5. Preferring the LR model over the RF model is rational due to its strengths in interpretability and computational efficiency.

Visuals or descriptions to best compare your two most effective models (the tree model has been included and it has a discrete curve)



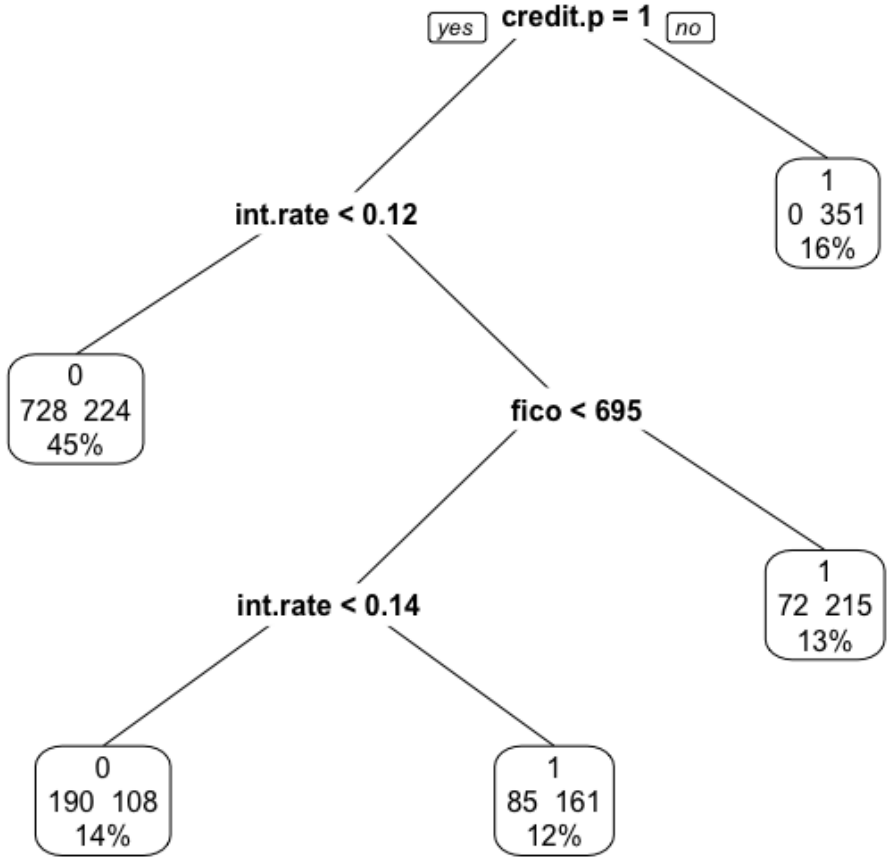
Objectives

- Model Selection
  - Minimize false negative rate
  - Maximize AUC

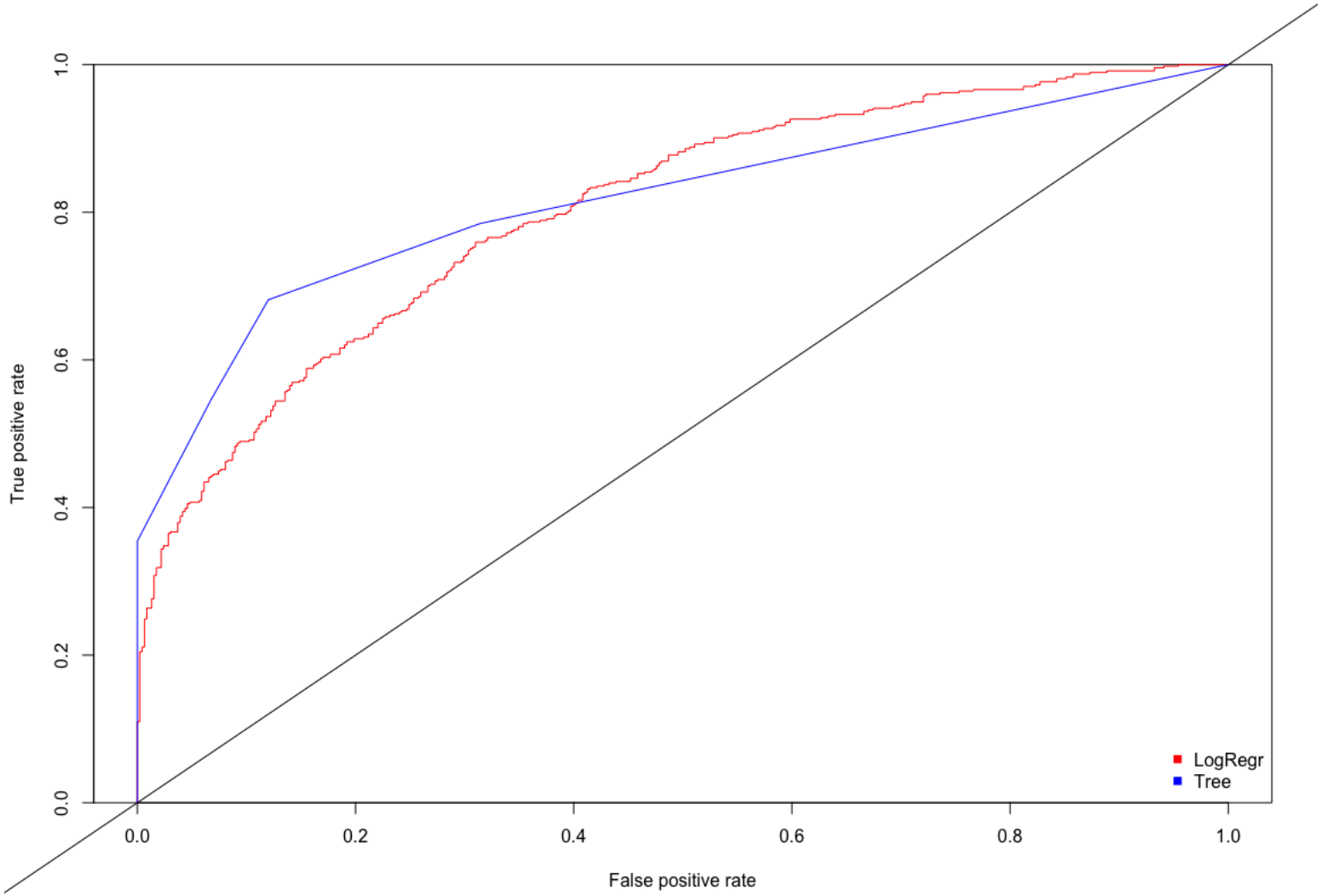
- Optimization
  - Balance the maximization of:
    - Return on Investment (ROI)
    - Market Share
    - Profit
      - Emphasis on profitability given we are dealing with a mature industry
      - Factors into consistent returns to stakeholders

Original Tree Model Results

- We predicted the binary decision of default as a function of the interest rate, credit policy and fico score.
- Credit policy is the root node; having it accounts for 84% of the tree's predictive capabilities which are broken down by interest rate- and fico-derived nodes



AUC Analysis – Original Tree Model & Logistic Regression



- The initial KPI used in comparing our tree model to the logistic model is AUC.
- Based off our AUC scale, both models fall within the “good” category within a small margin (0.8-0.9), however the tree model performs slightly better.

Counterfactual Analysis – Original Tree Model and Logistic Regression

- Our counterfactual analysis reveals a more significant difference in performance between the models.
- Through our analysis, we observe an approximately 2% false negative rate of the tree model.
- Therefore, the logistic model significantly outperforms the tree model despite disproportionately lower SUC.

- Tree counterfactual analysis (~12% false negative rate):

	FALSE	TRUE
0	4206	3839
1	572	961

- Logistic counterfactual analysis (~10% false negative rate):

	FALSE	TRUE
0	4582	3463
1	519	1014

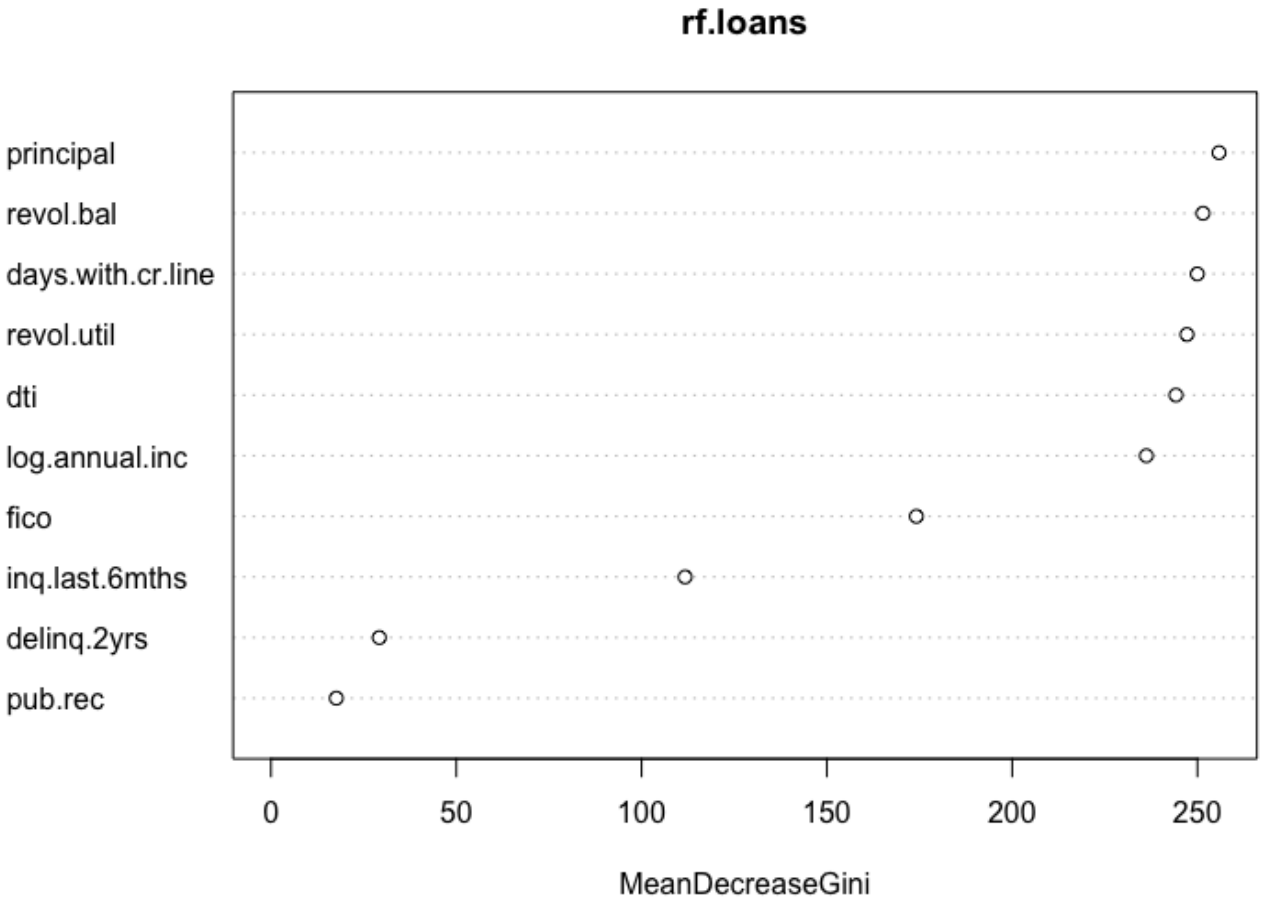
Model Comparison and Selection – Logistic (Rev.)

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              9.395e+00  1.087e+00   8.642  < 2e-16 ***
fico                    -9.680e-03  1.191e-03  -8.124 4.51e-16 ***
purposecredit_card      -5.820e-01  1.282e-01  -4.539 5.64e-06 ***
purposedebt_consolidation -3.971e-01  9.052e-02  -4.387 1.15e-05 ***
purposeeducational      -5.413e-02  1.810e-01  -0.299 0.764887
purposehome_improvement  1.759e-02  1.504e-01   0.117 0.906876
purposemajor_purchase   -3.600e-01  1.976e-01  -1.822 0.068466 .
purposesmall_business    6.137e-01  1.338e-01   4.585 4.53e-06 ***
inq.last.6mths           1.073e-01  1.393e-02   7.700 1.36e-14 ***
tot.payment              3.467e-05  5.297e-06   6.545 5.94e-11 ***
log.annual.inc          -4.489e-01  6.813e-02  -6.589 4.42e-11 ***
revol.bal                3.611e-06  1.079e-06   3.345 0.000822 ***
revol.util               3.828e-03  1.444e-03   2.652 0.008011 **
pub.rec                  2.670e-01  1.104e-01   2.419 0.015585 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
    FALSE TRUE
  0  1359 1031
  1   143  305
> # Accuracy of the logistic regression model
> (Conf.lr[1,1]+Conf.lr[2,2])/sum(Conf.lr)
[1] 0.5863284
> # False negative rate of the lr model
> (Conf.lr[2,1]/(Conf.lr[1,1]+Conf.lr[2,1]))
[1] 0.09520639
>
```

```
> (auc.lr = as.numeric(auc.tmp@y.values))
[1] 0.6812575
```

Model Comparison and Selection – Random Forest

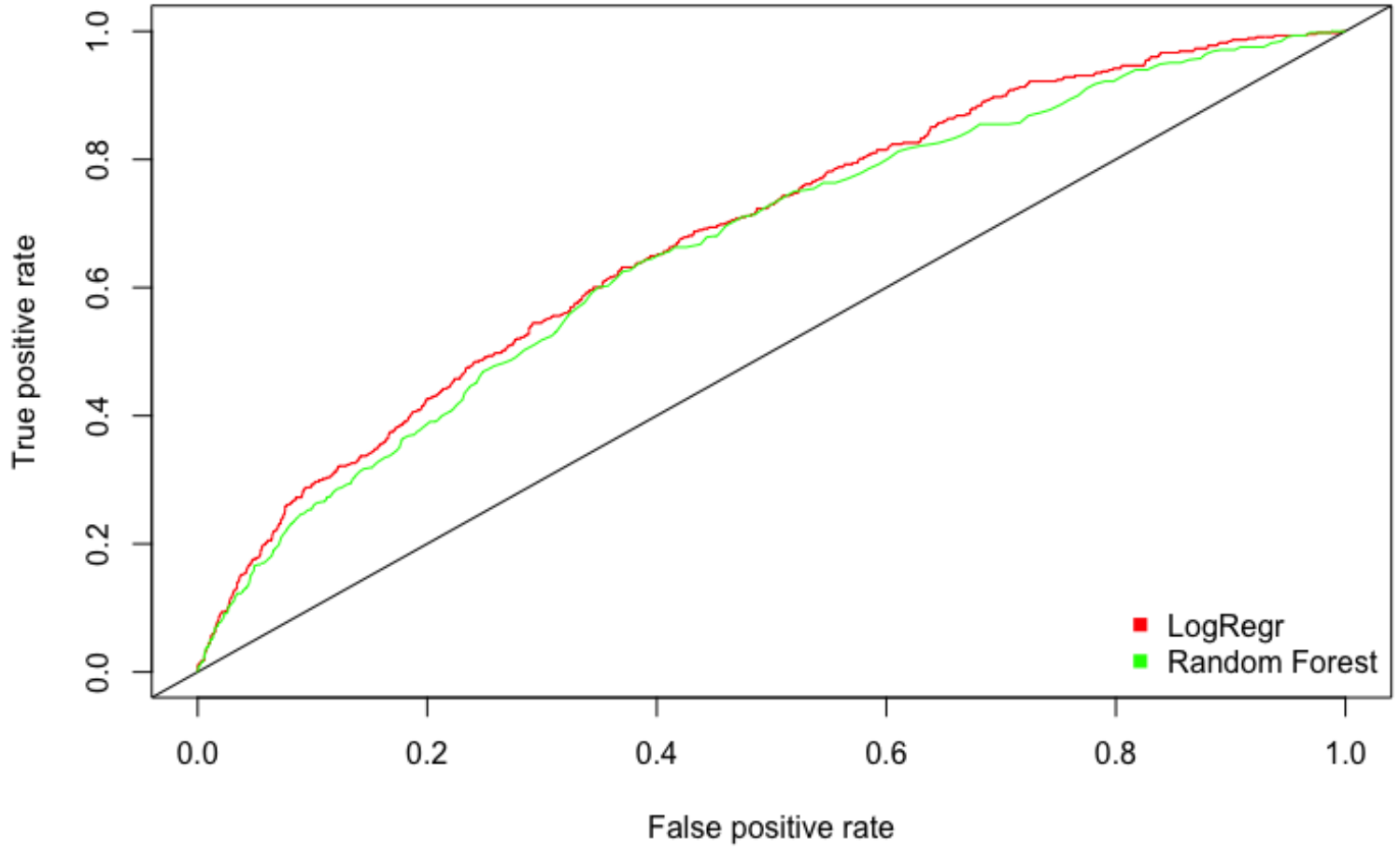



```
> Conf.rf = table(test$default, rf.default.prob[,2] > 0.15)
> Conf.rf

    FALSE TRUE
  0  1257 1133
  1   130  318
> # Accuracy of the random forest model
> (Conf.rf[1,1]+Conf.rf[2,2])/sum(Conf.rf)
[1] 0.5549683
> # False negative rate of the rf model
> (Conf.rf[2,1]/(Conf.rf[1,1]+Conf.rf[2,1]))
[1] 0.09372747
```

```
> (auc.rf = as.numeric(auc.tmp@y.values))
[1] 0.6624514
```

Model Comparison and Selection



- False Negative Rate:
  - 9.52% **(Logistic Regression)**
  - 9.37% (Random Forest)
- AUC:
  - 68.23% **(Logistic Regression)**
  - 66.25% (Random Forest)

Optimization Model Description

	LR	RF	LR	RF			
Threshold	1	1	Total loans out of 2838				
0.16	1	1	1676	1500			
Recovry rate	1	1	Total Principal Invested				
0.1	1	1	12459138	11473249			
	1	1	Total Expected Profit			Total Actual Profit	
	1	0	3256879	3302681		3326223.458	3272135
	1	1	Expected ROI Principal			Total Actual ROI Principal	
	1	1	0.261405	0.287859		0.266970595	0.285197
	0	1					
	0	0					
	0	0					

- Optimal Threshold: **16%**
- A 16% threshold is not only market-realistic – it also maximizes total expected profit from rates between 0 and 16%.

Conclusion

- Model Selection
  - Given comparatively slight difference in the false negative rates, move forward with the **revised logistic regression** model given significantly higher AUC scoring.

- Model Optimization
  - Move forward with a **16%** threshold to maximize profitability (with market share and ROI as ancillary to this optimization maximization).