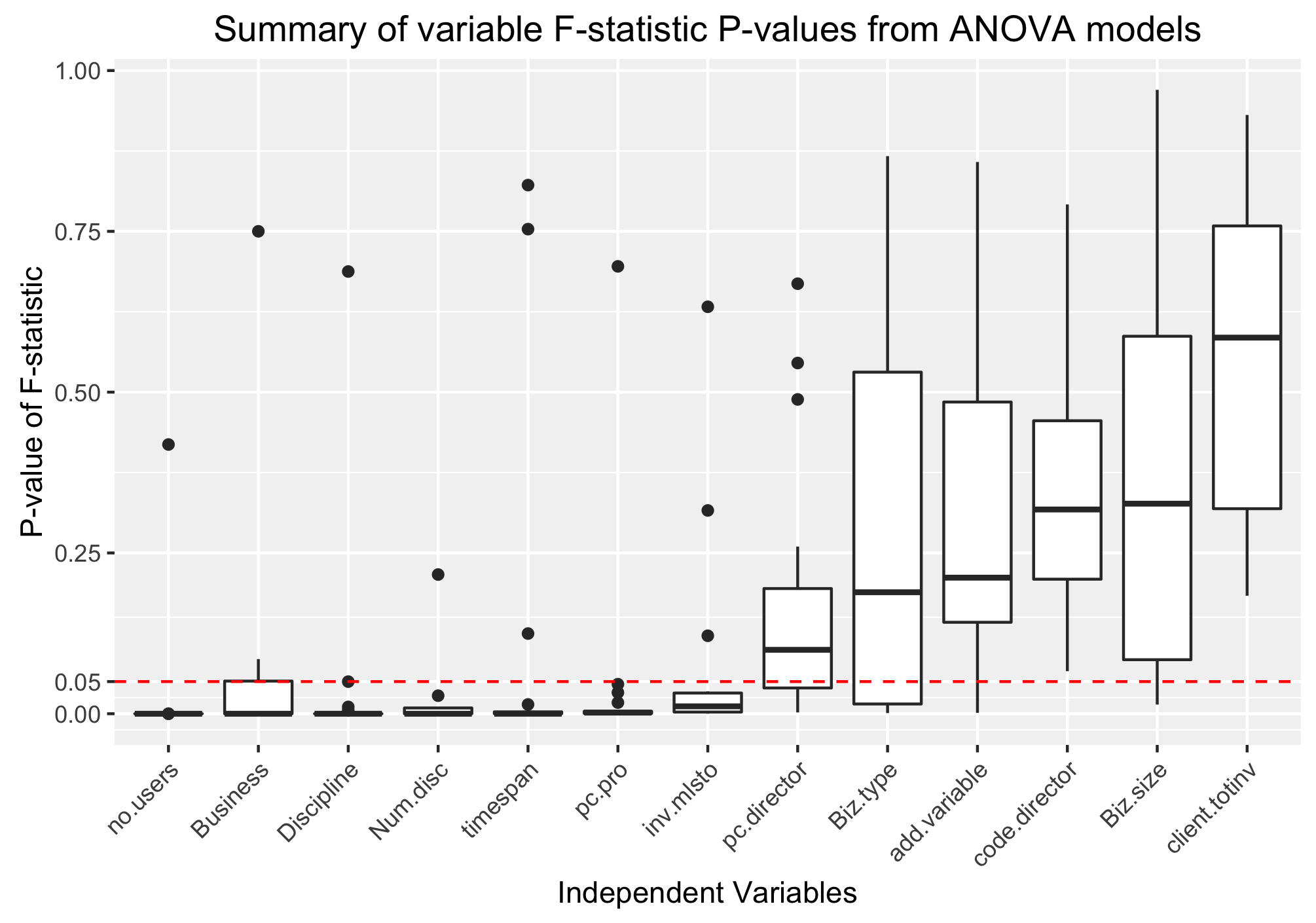
trials

Amy Cook

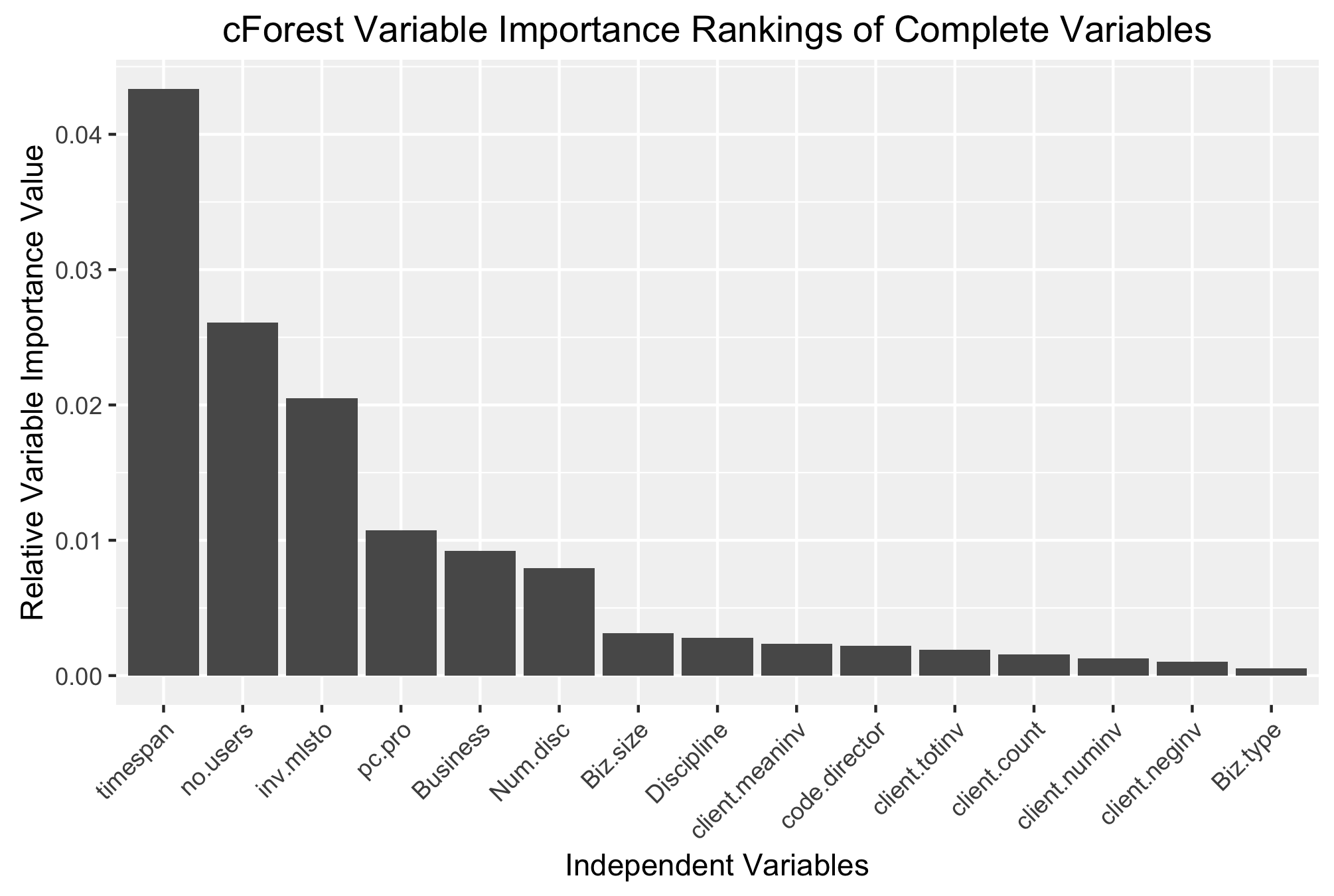
June 16, 2016

1



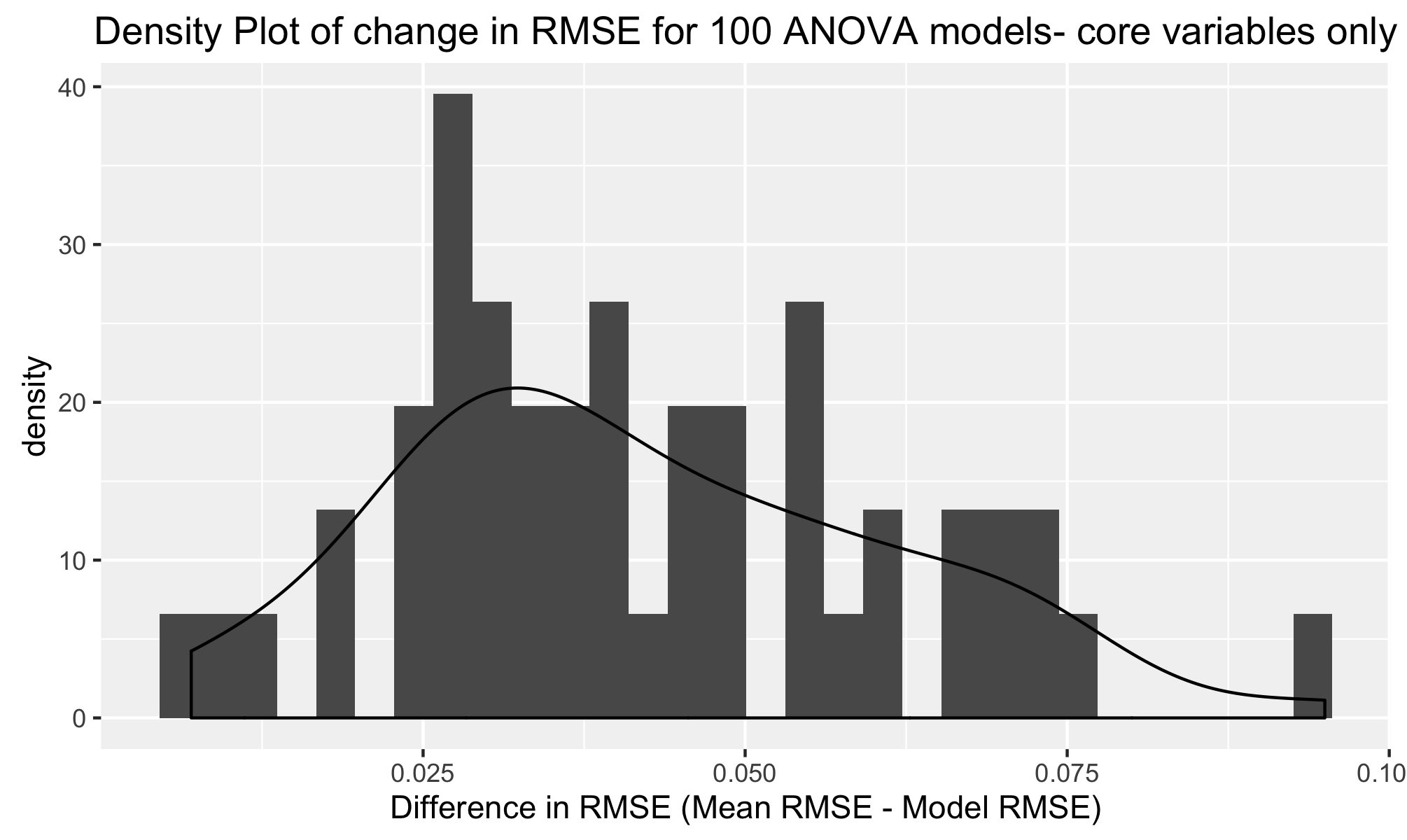
P-values of an ANOVA regression's F-statistics used to interpret variable importance

2



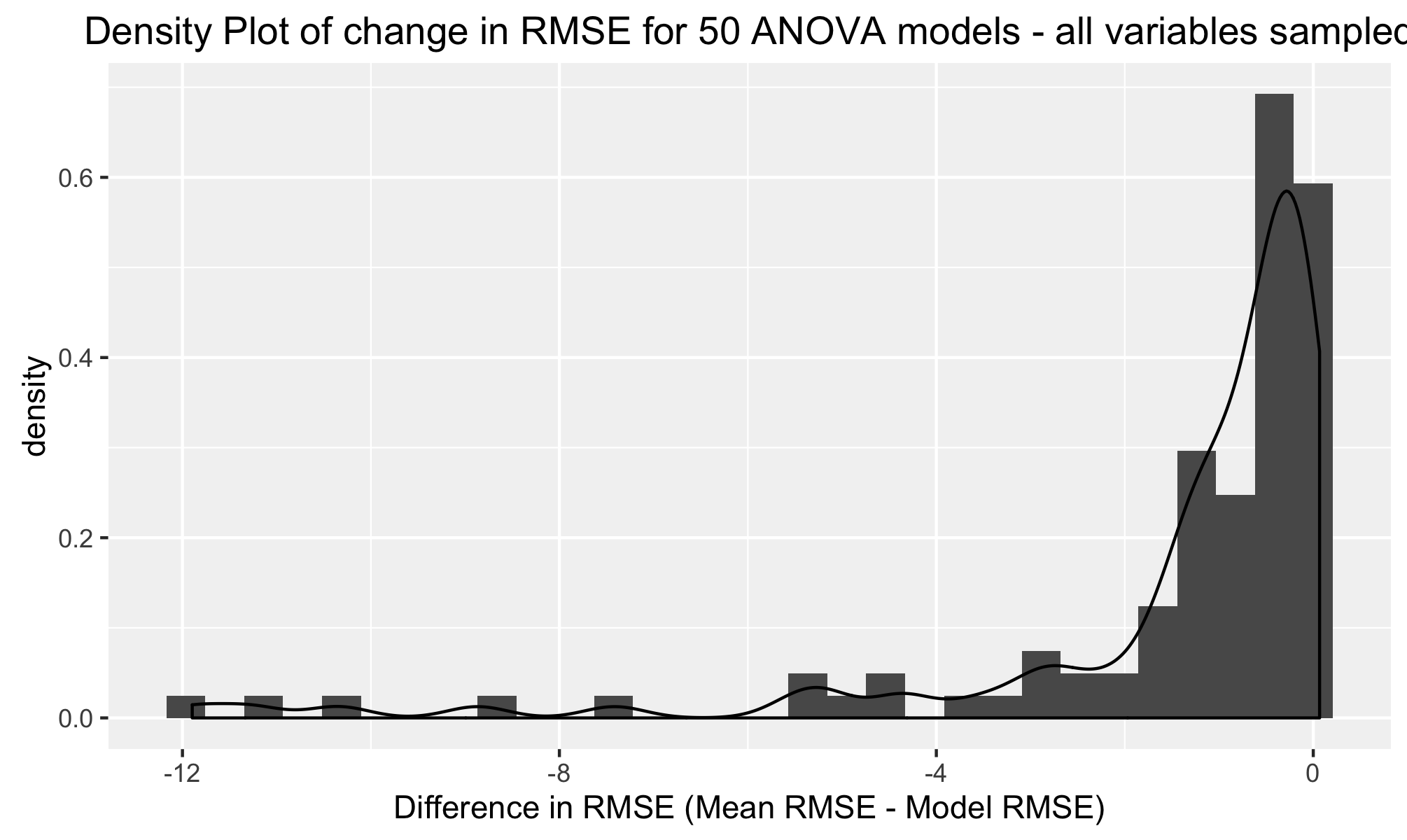
Variable importance output from a cForest built from 15 core variables

3



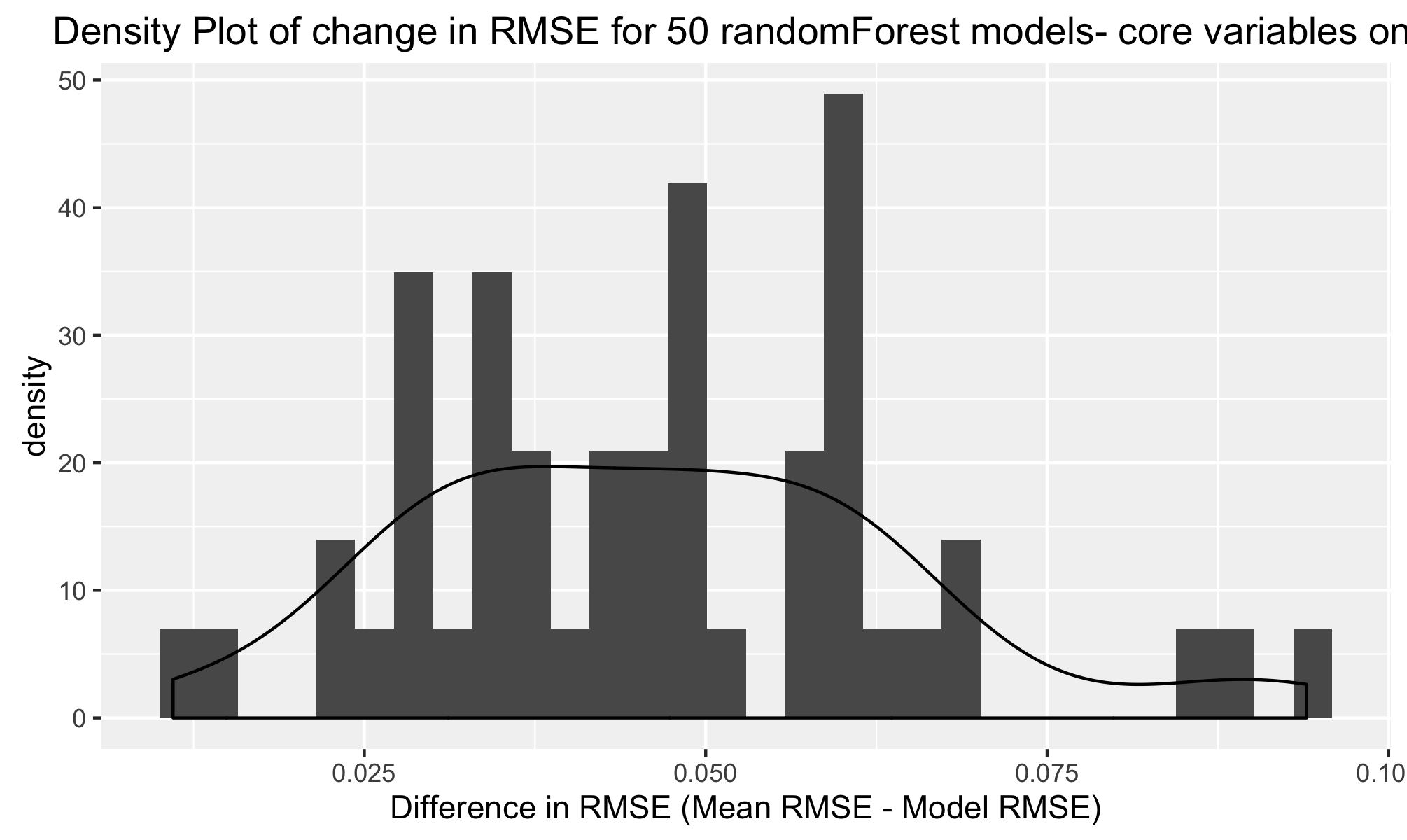
Distribution of the difference in RMSE between ANOVA model predictions (built on 6 core variables) and a base line predictor that uses the mean return per dollar as its prediction. If the ANOVA model is effective, the difference should be greater than 0.

4



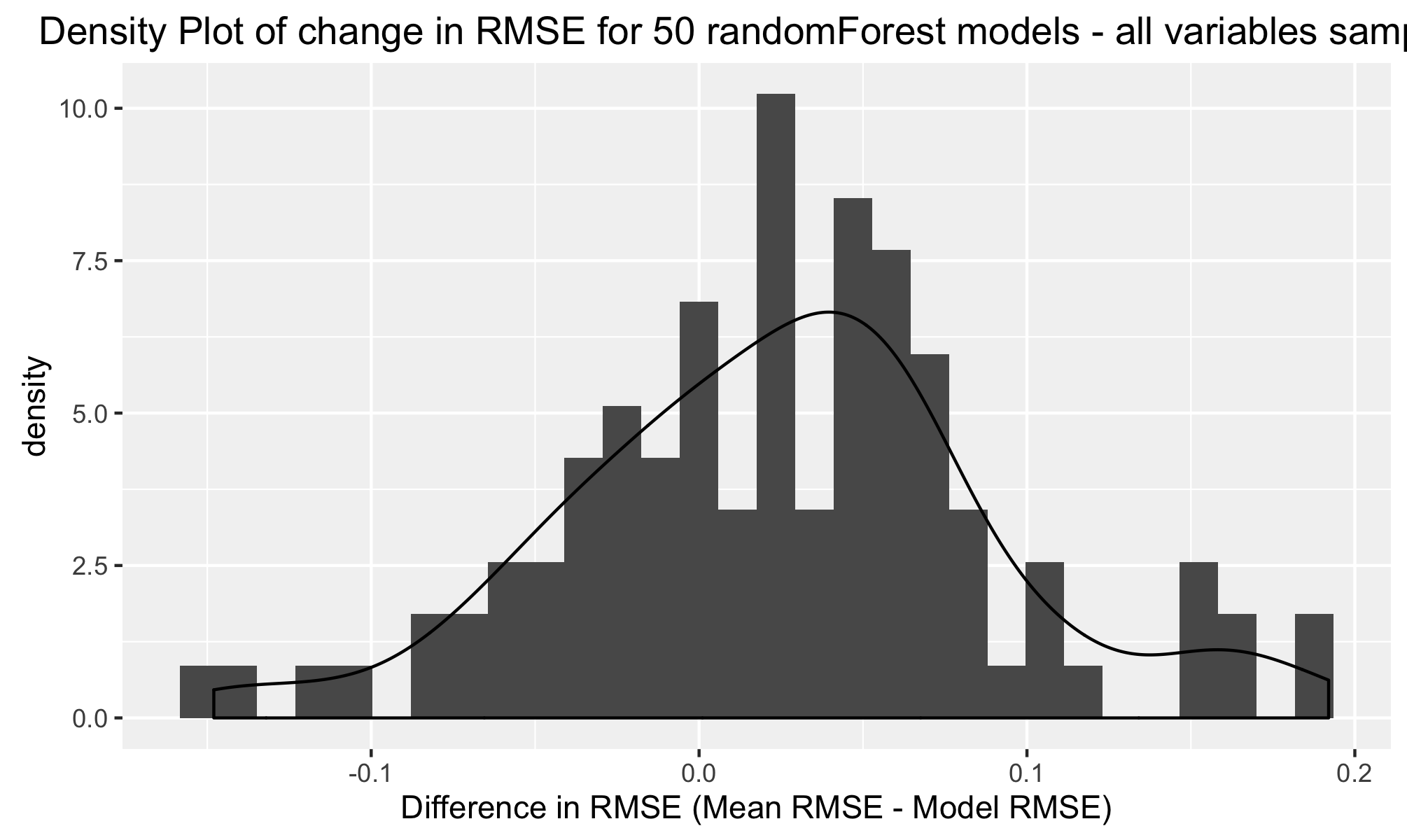
Distribution of the difference in RMSE between ANOVA model predictions (built on all variables) and a base line predictor that uses the mean return per dollar as its prediction. If the ANOVA model is effective, the difference should be greater than 0.

5



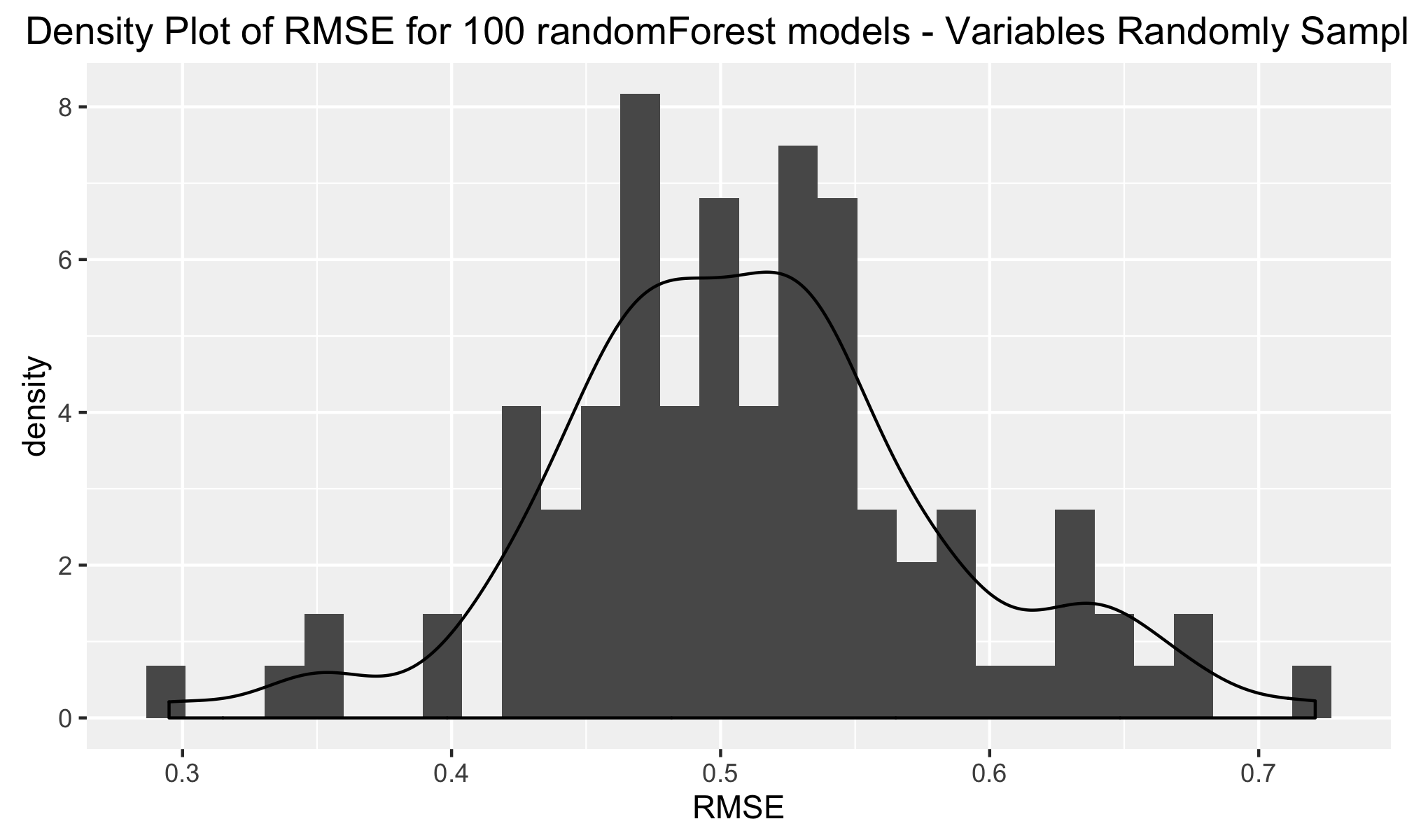
Distribution of the difference in RMSE between randomForest predictions (built on 6 core variables) and a base line predictor that uses the mean return per dollar as its prediction. If the ANOVA model is effective, the difference should be greater than 0.

6



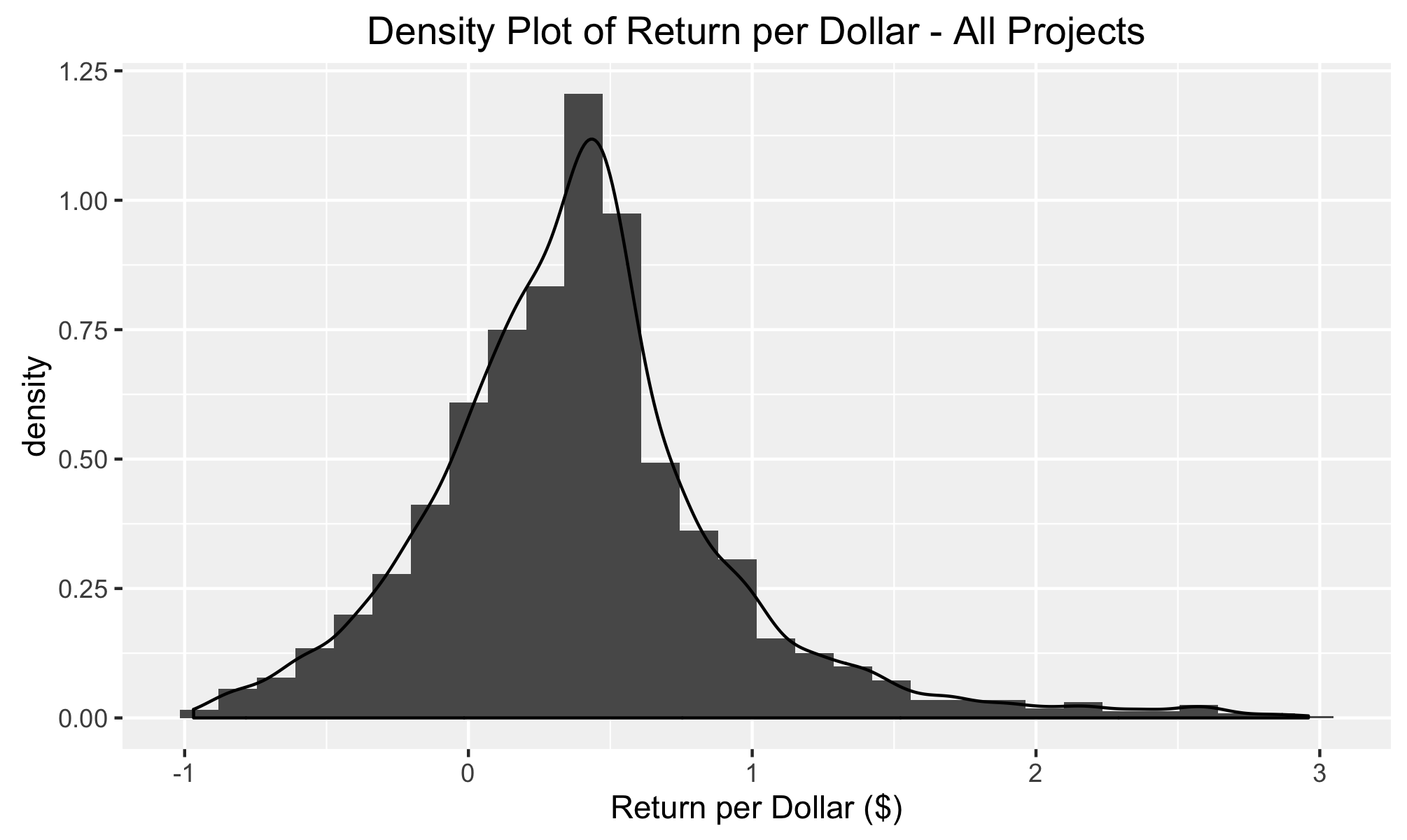
Distribution of the difference in RMSE between randomForest predictions (built on all variables) and a base line predictor that uses the mean return per dollar as its prediction. If the ANOVA model is effective, the difference should be greater than 0.

7



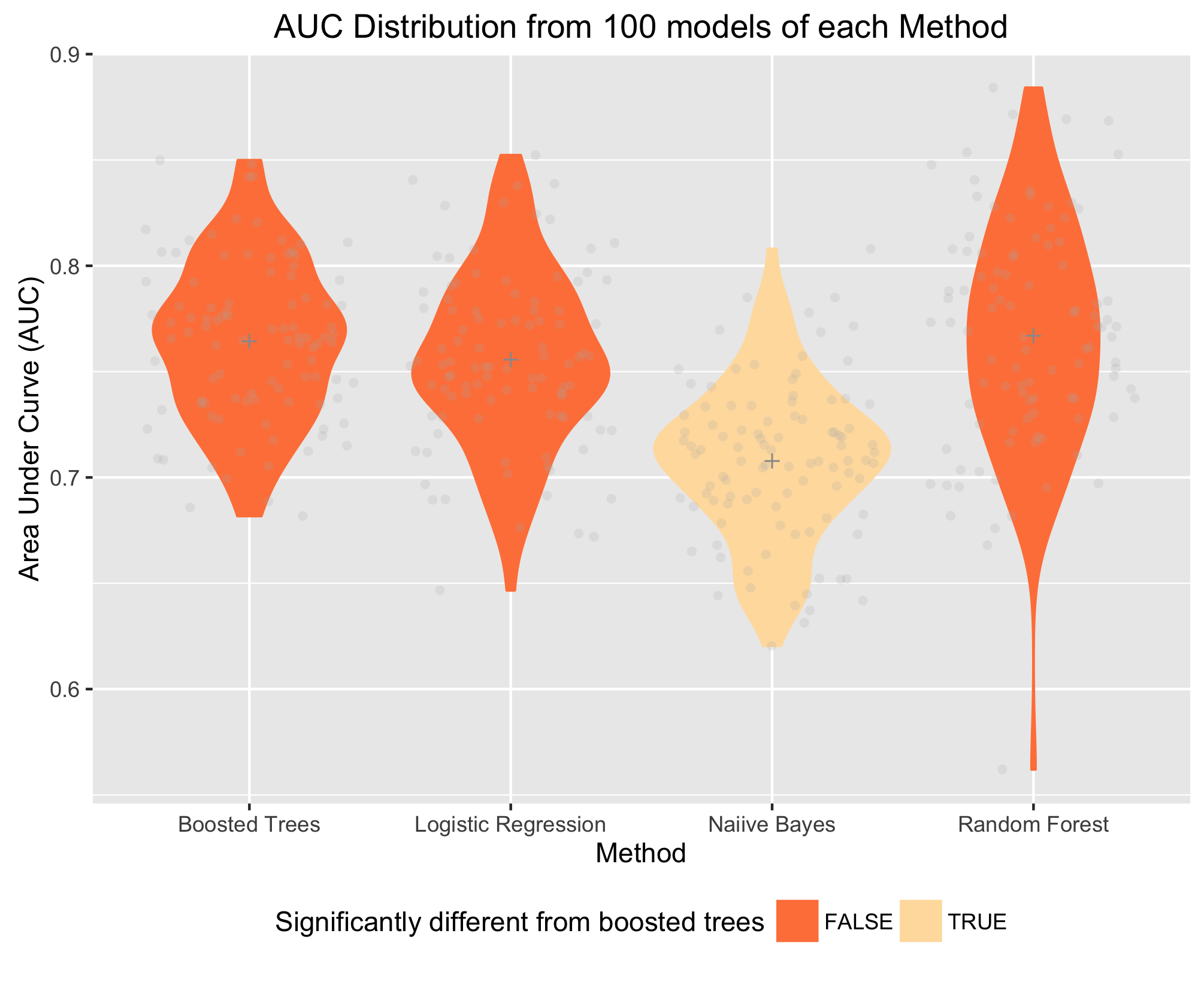
Distribution of RMSE's of 'return per dollar' from 100 randomForest models built on random subsamples of the variables

8

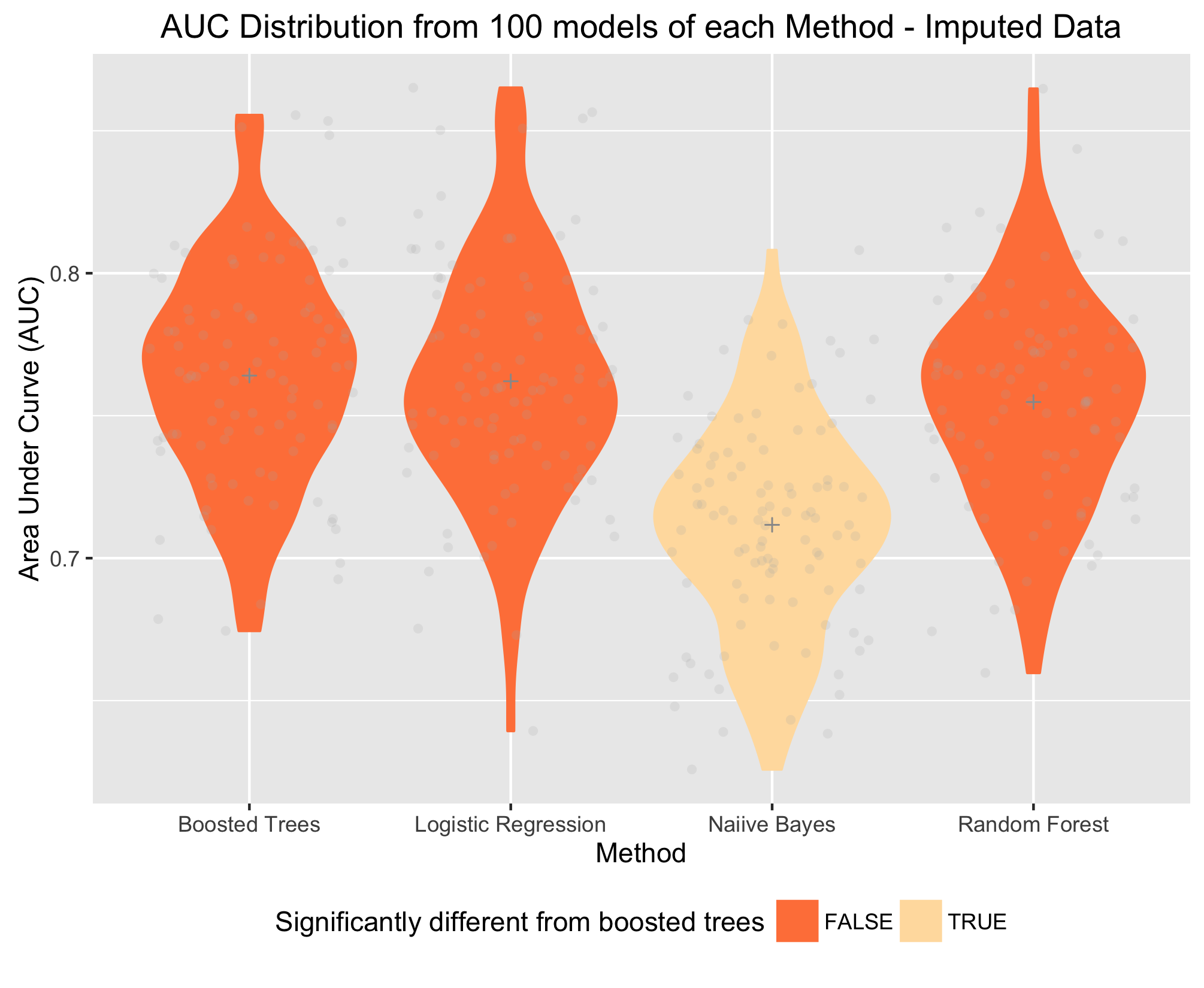


Distribution of 'return per dollar' for all projects

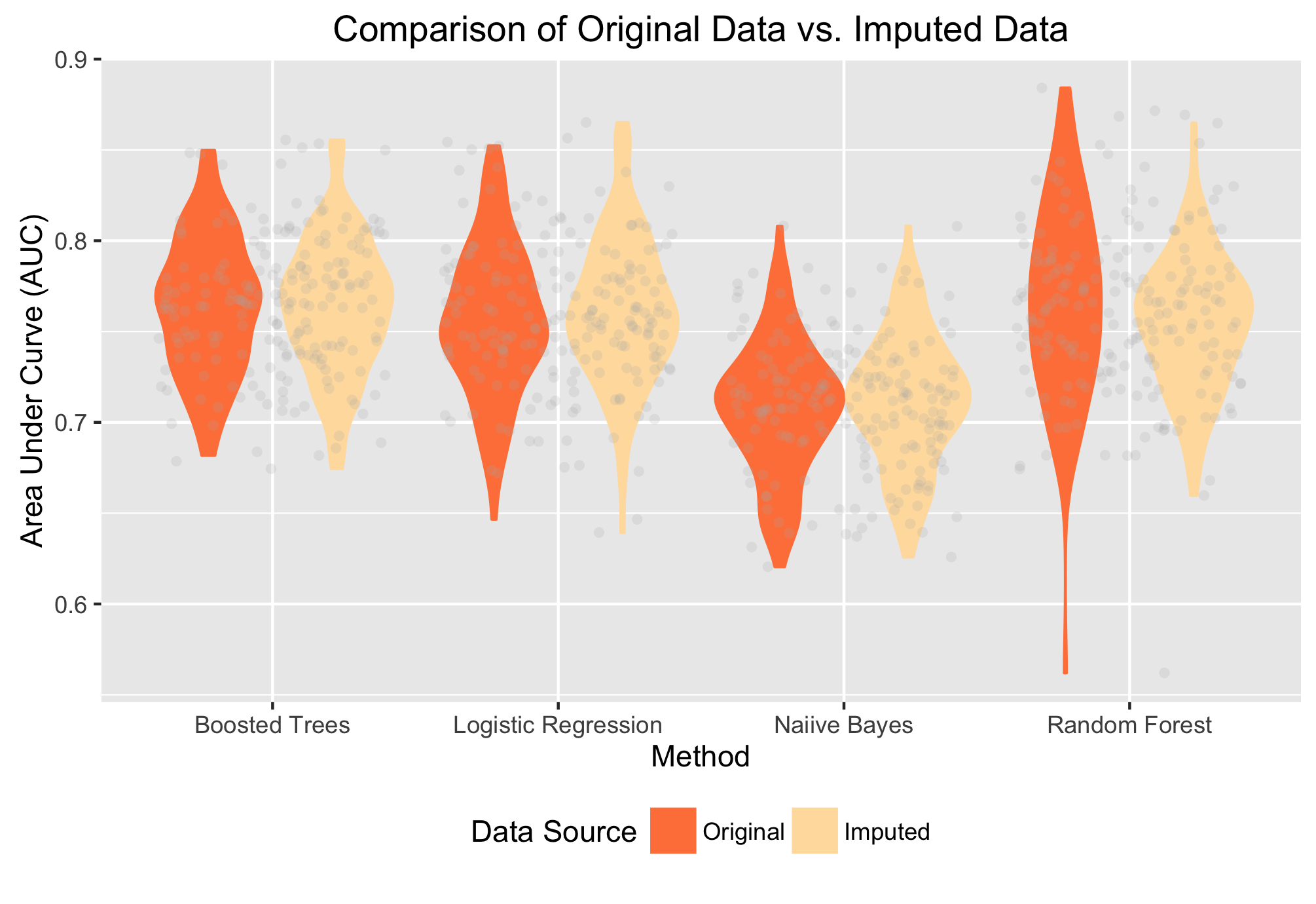
9



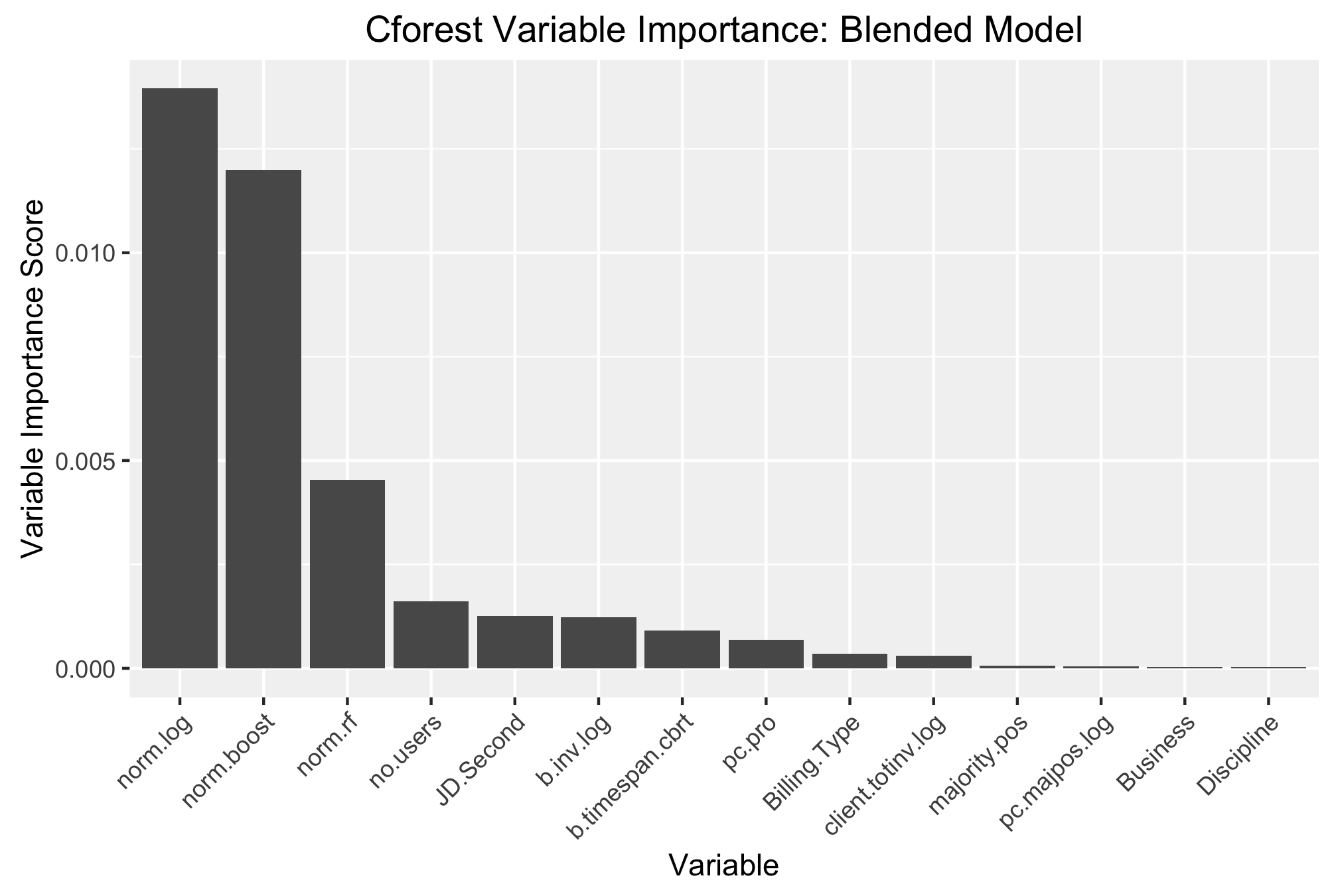
Violin plot vertically illustrating the distribution of AUC values from each of the methods when predicting profit/loss. Subsets of the data were used for Logistic Regression and randomForests in order to provide datasets without missing values.



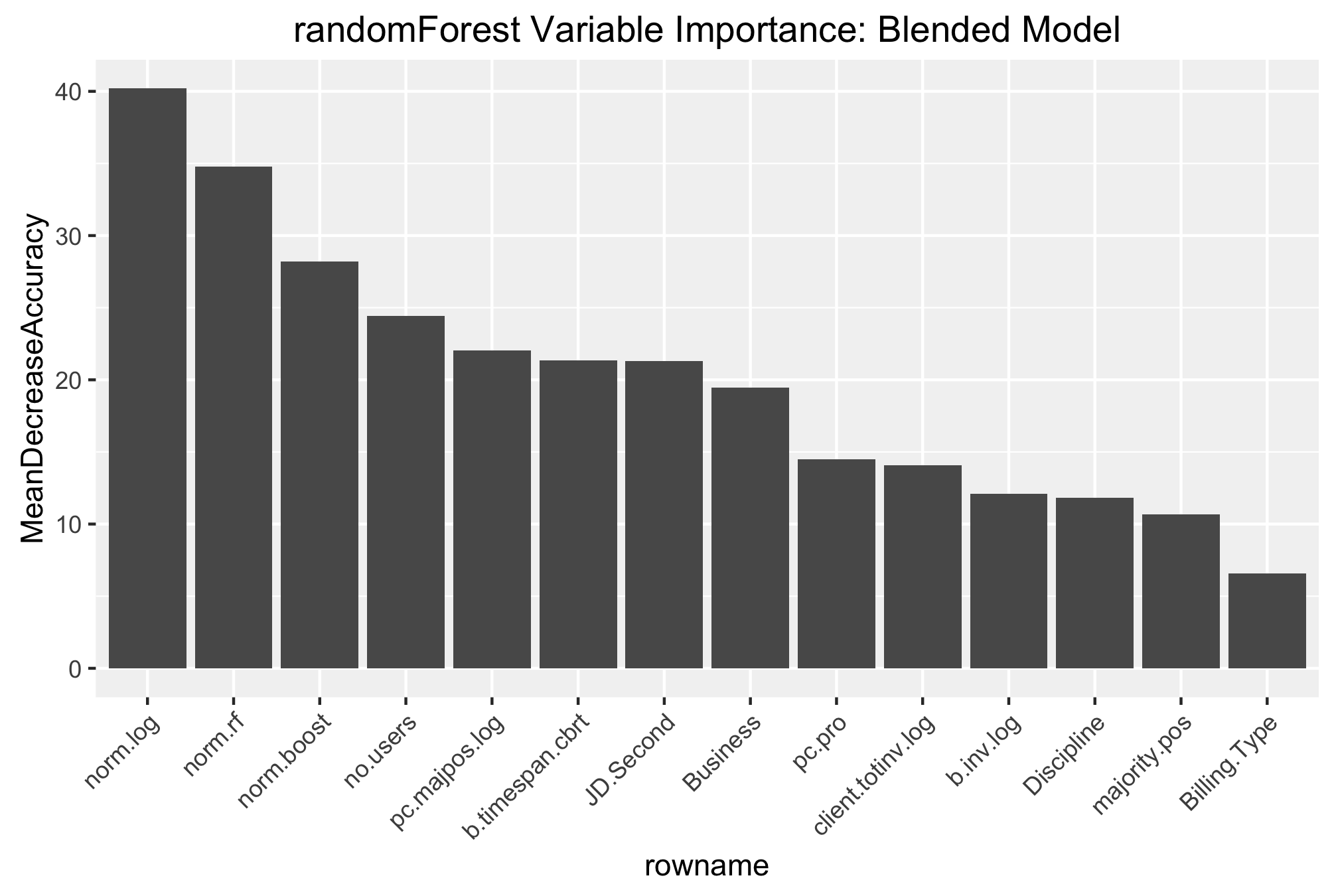
Violin plot vertically illustrating the distribution of AUC values from each of the methods when predicting profit/loss. Each method was fed the same imputed full dataset.



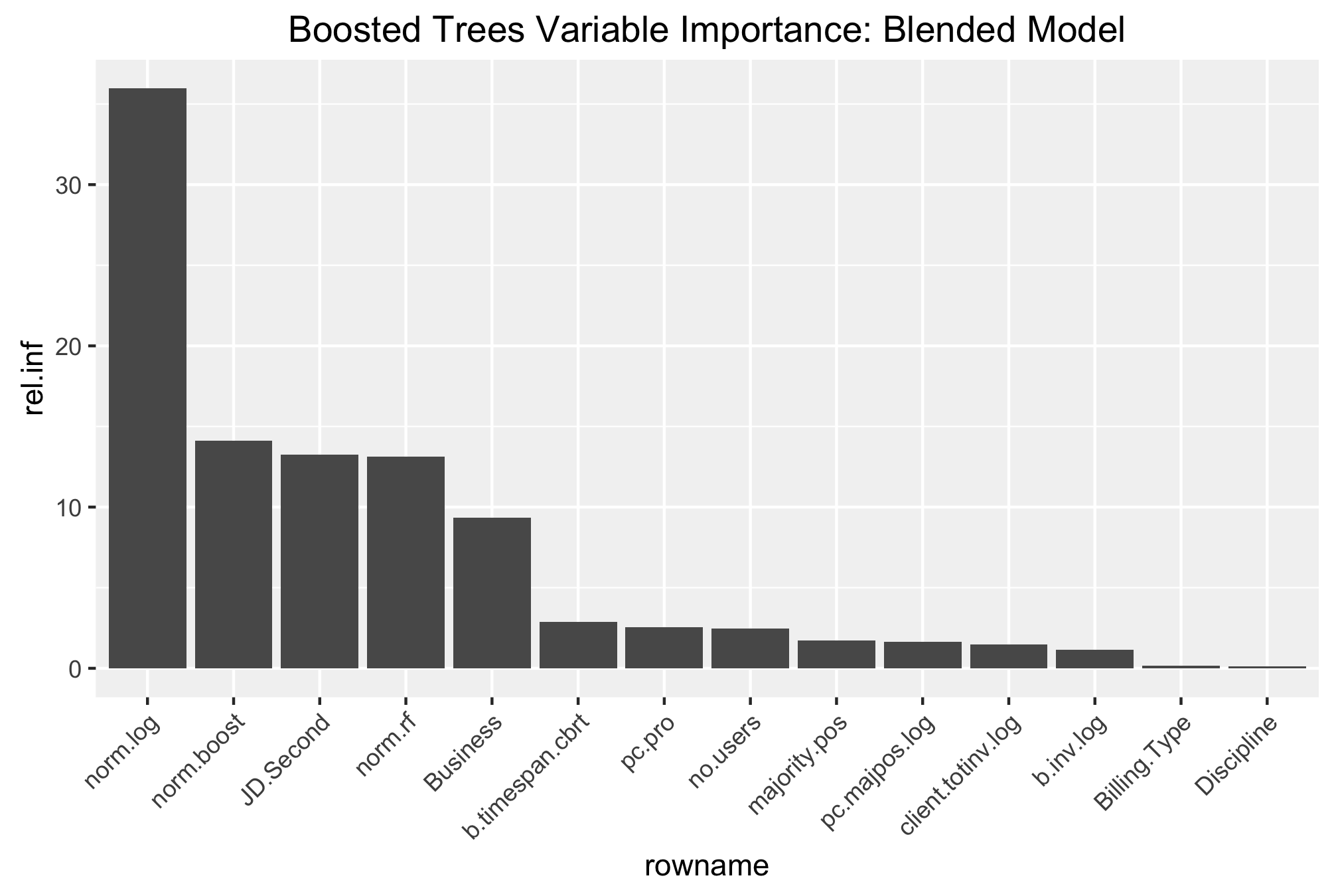
Combined violin plots comparing AUC results for each method predicting profit/loss using subsets of data vs. the full imputed data set.



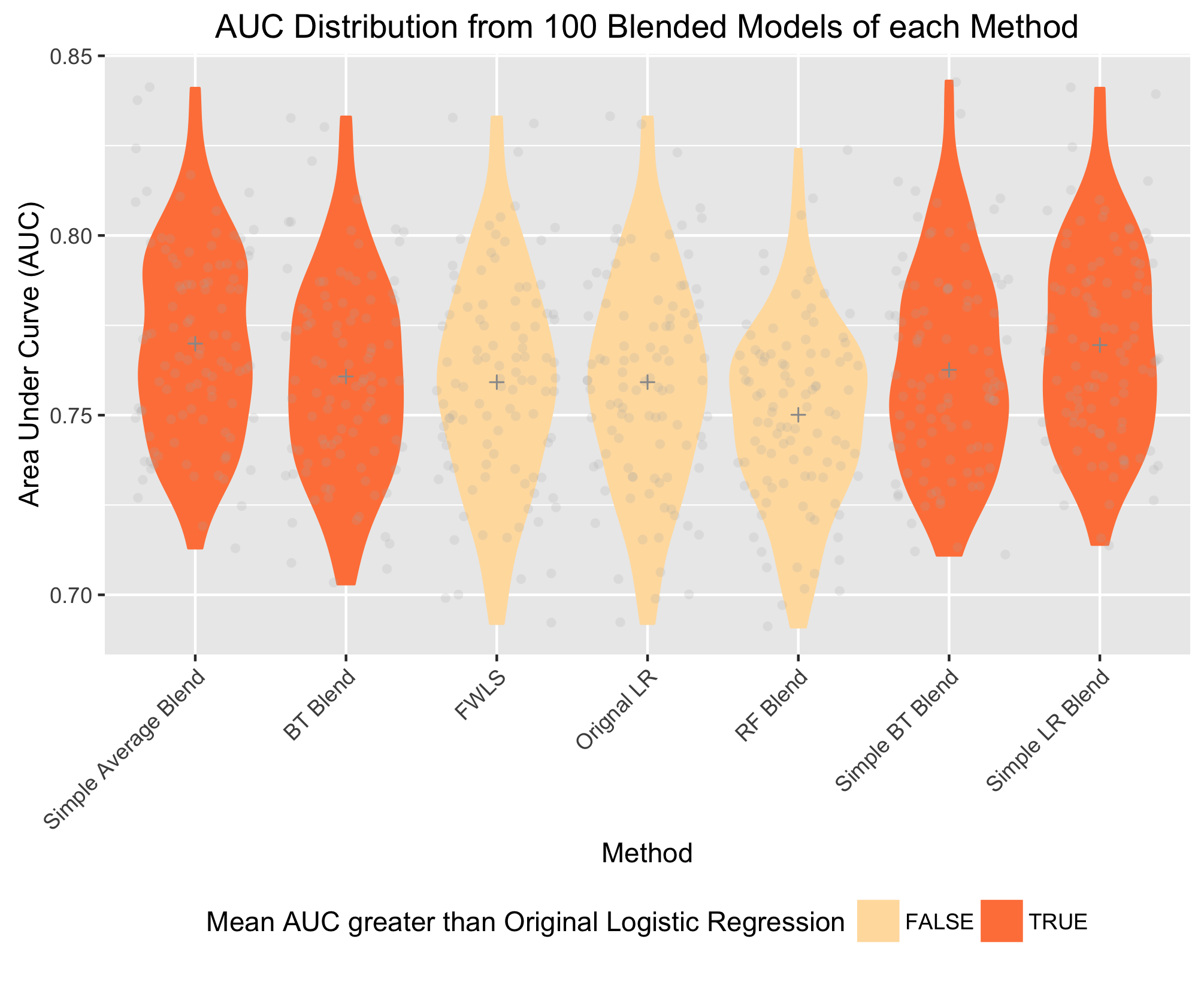
Variable importance output from a cForest blended model. The results from the three best performing models were added as explanatory variables.



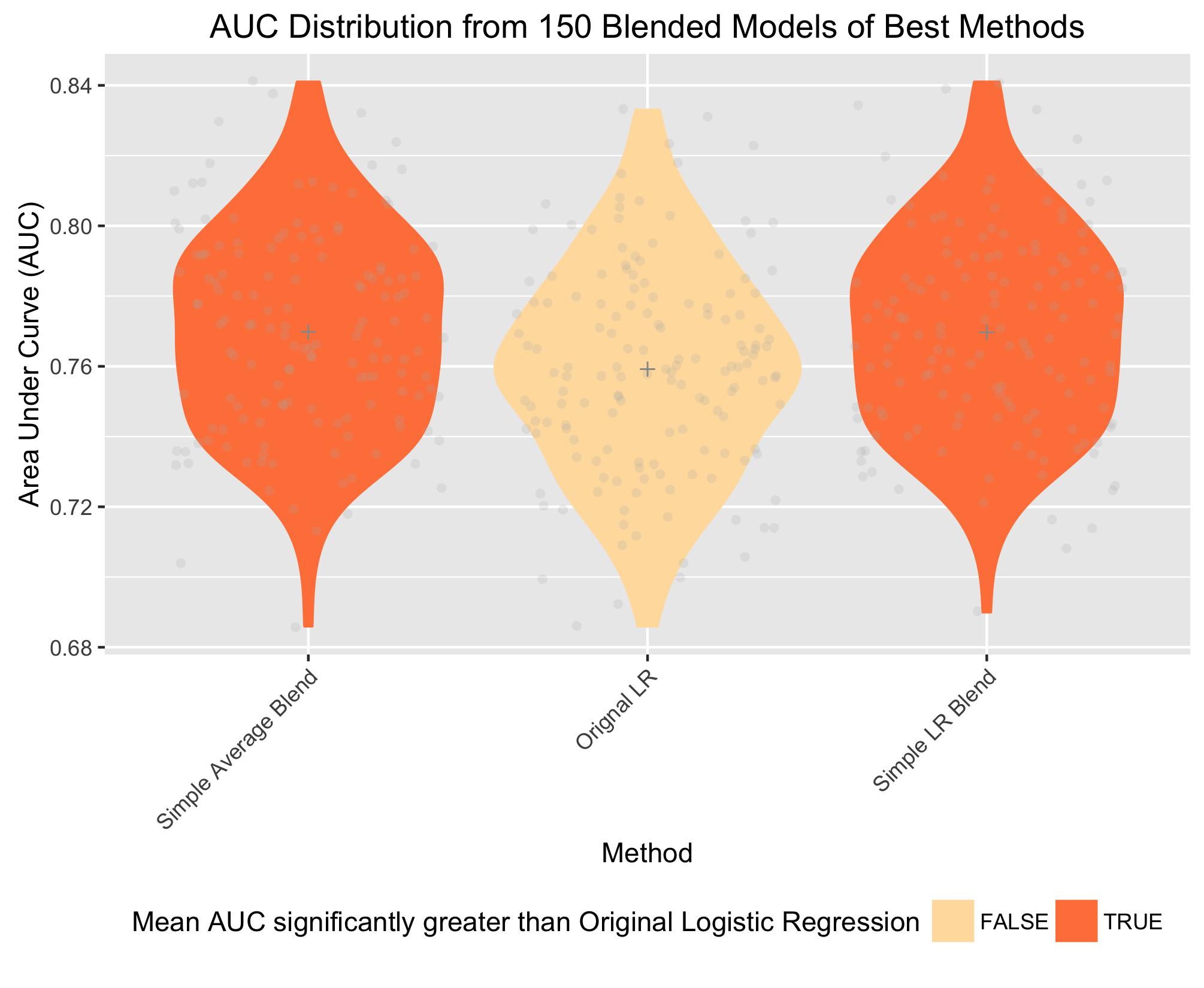
Variable importance output from a randomForest blended model. The results from the three best performing models were added as explanatory variables.



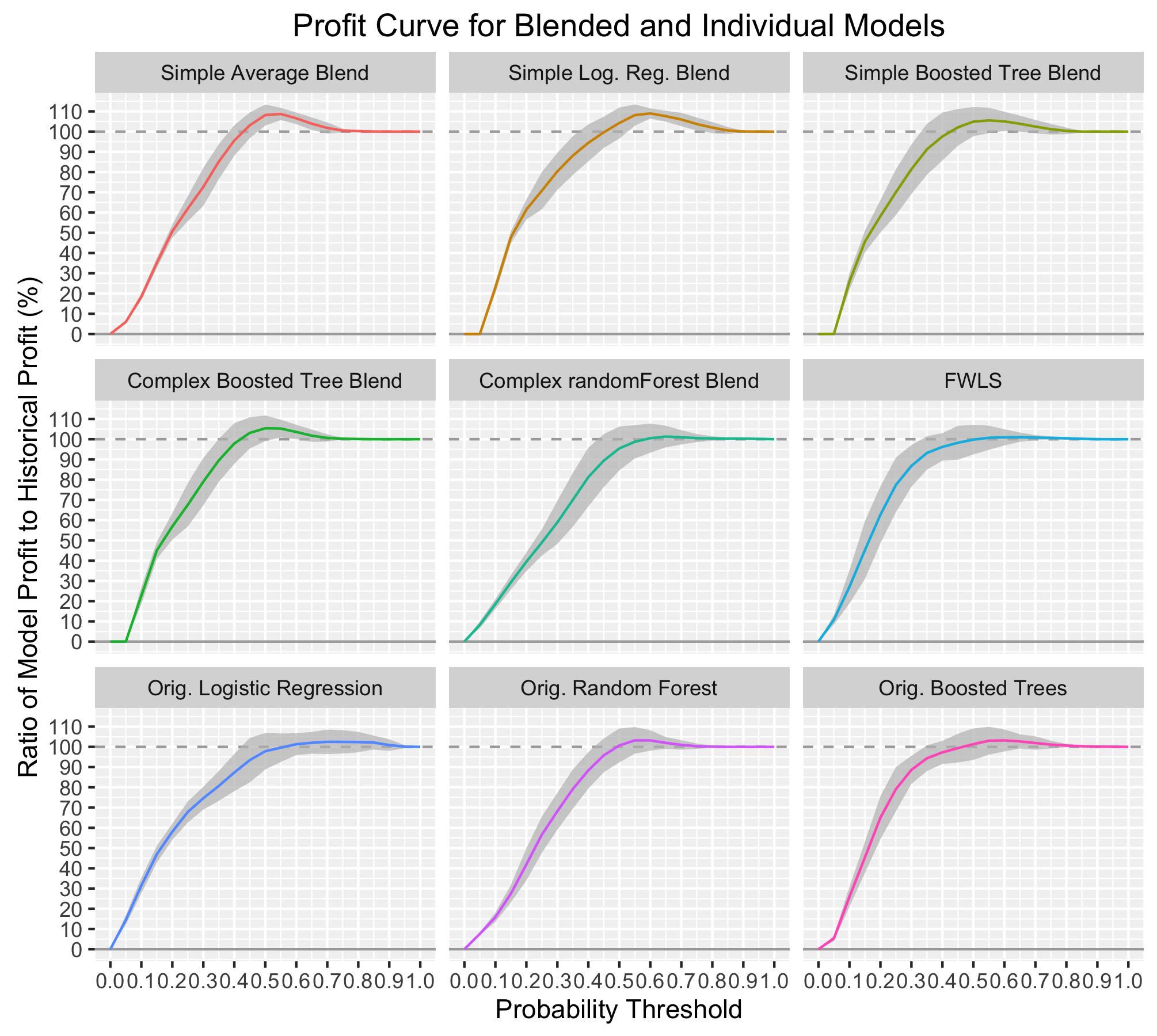
Variable importance output from a Boosted Tree blended model. The results from the three best performing models were added as explanatory variables



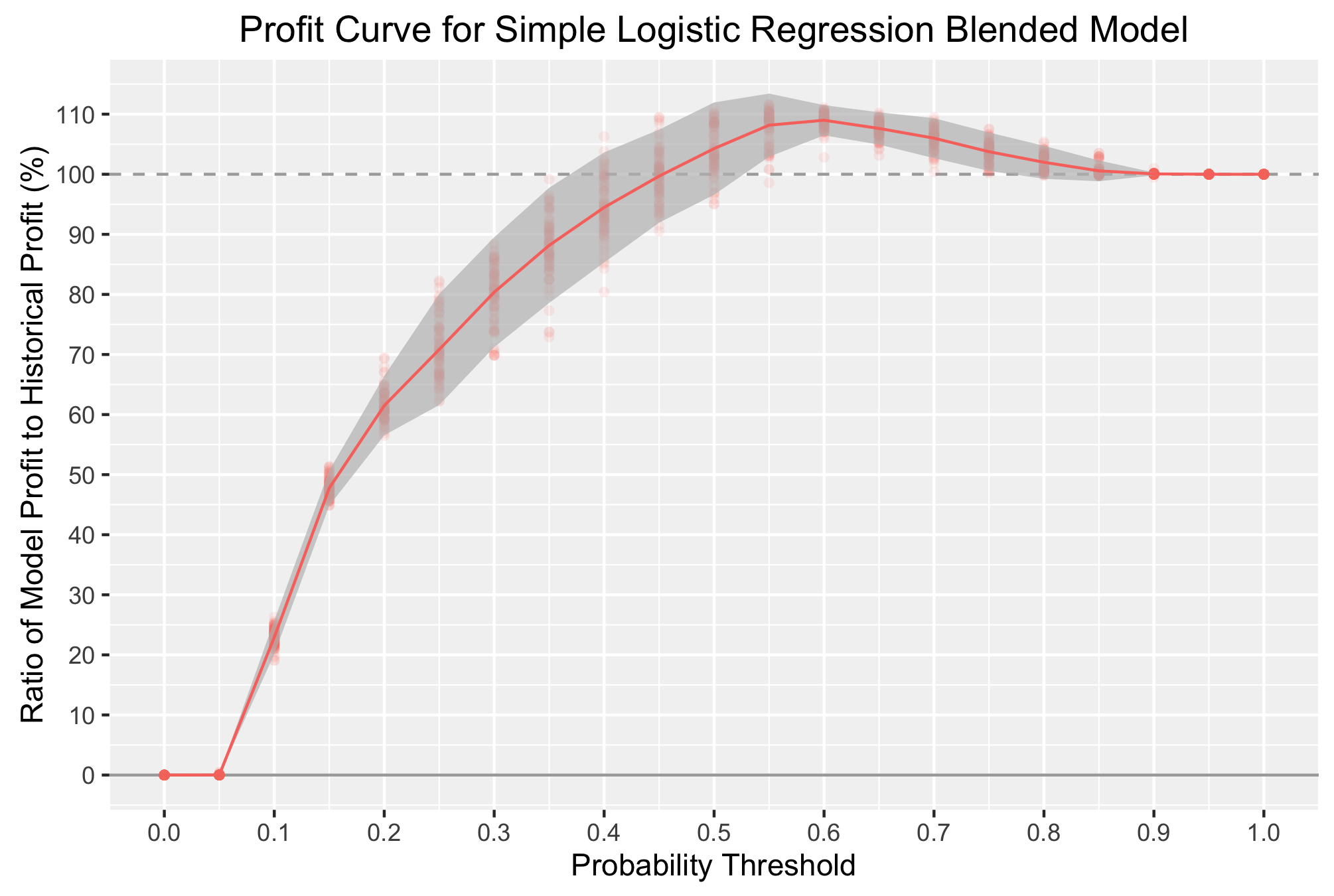
Violin plot vertically illustrating the distribution of AUC values from each of the blending methods when predicting profit/loss. 100 models were built for each method.



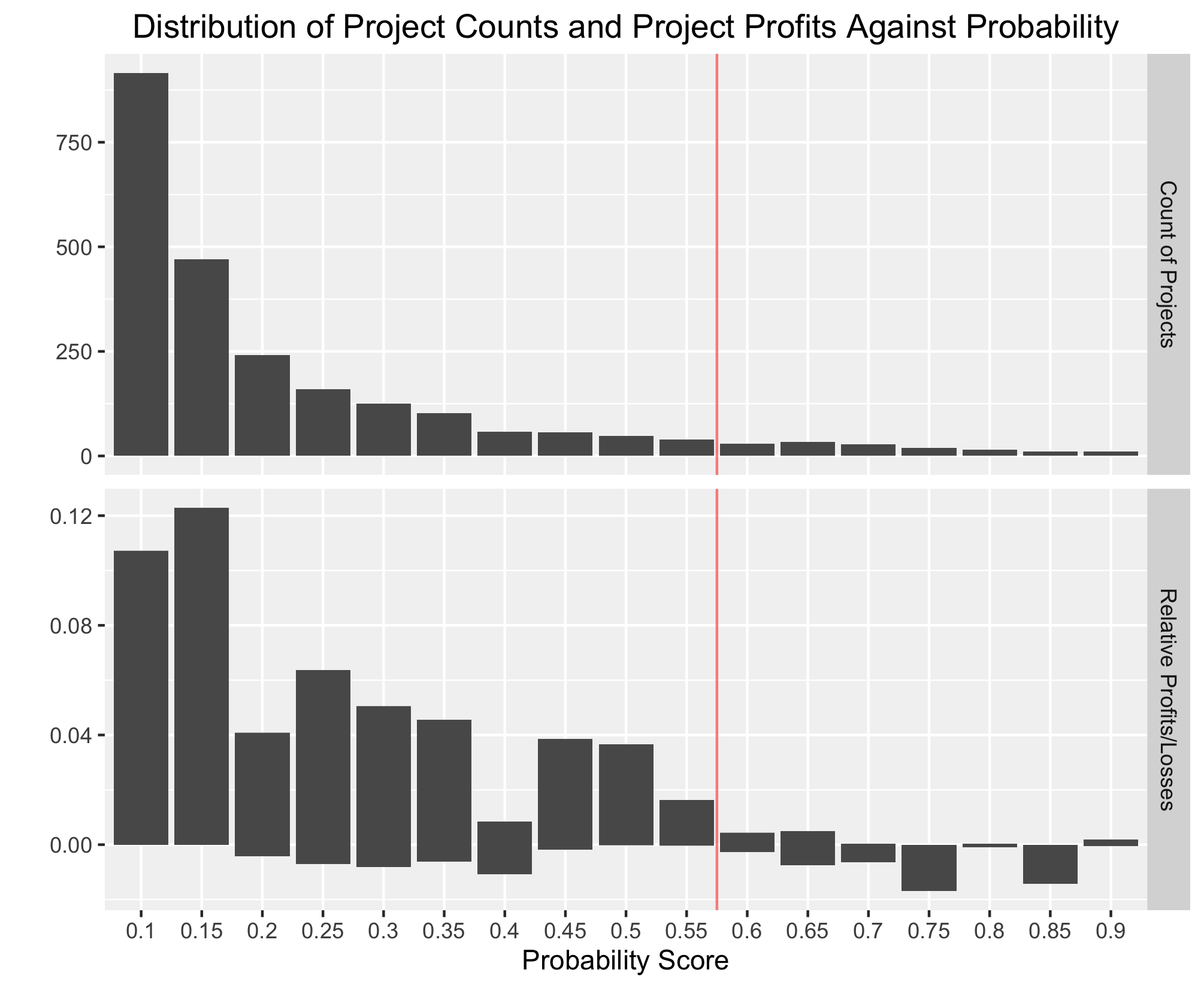
Violin plot illustrating the distribution of AUC values from the best three blending methods when predicting profit/loss. 150 models were built for each method to achieve a statistical power of 0.8.



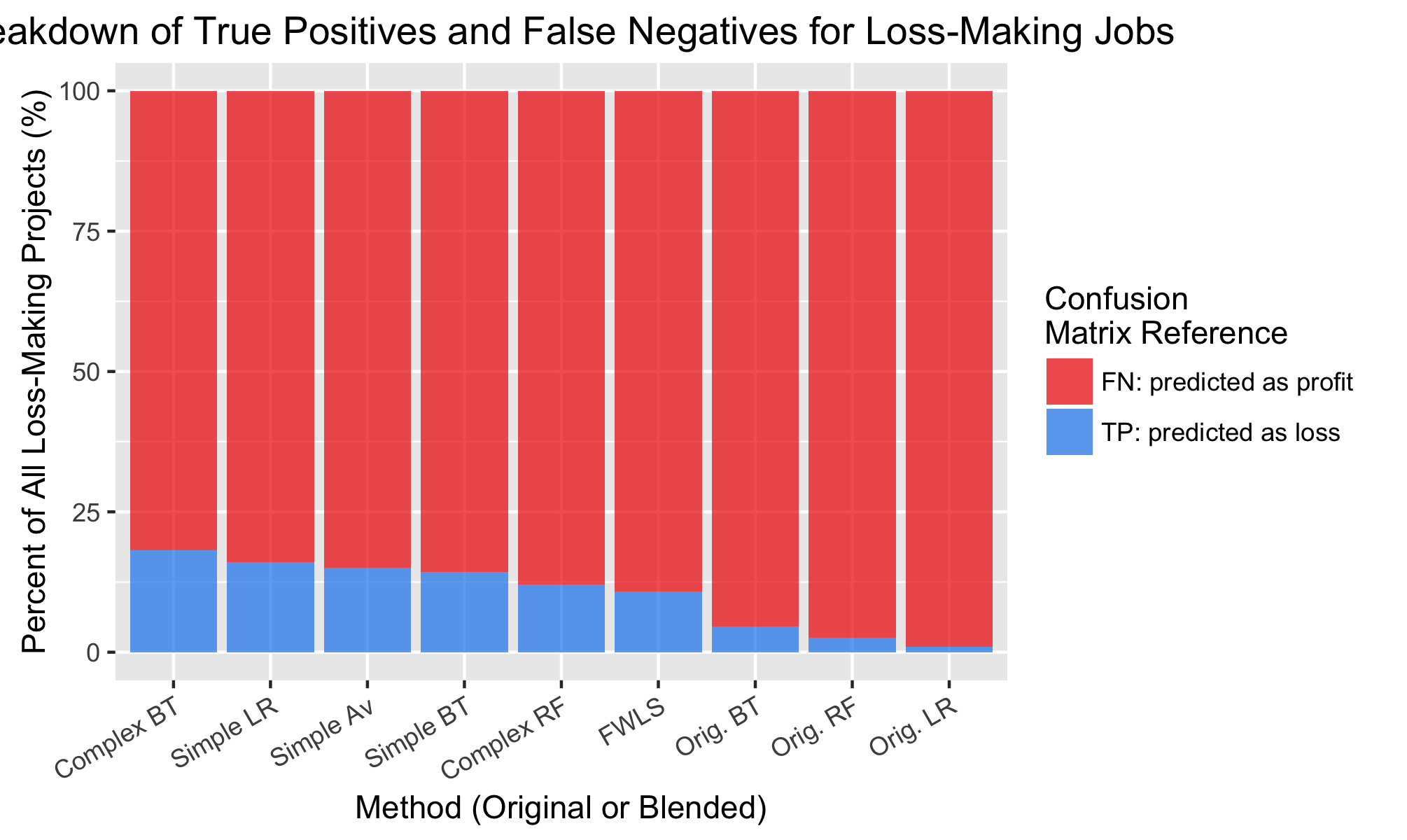
Profit Curves summarising results from 100 models of 9 methods: 3 simple blends, 3 complex blends, and the original 3 best methods.



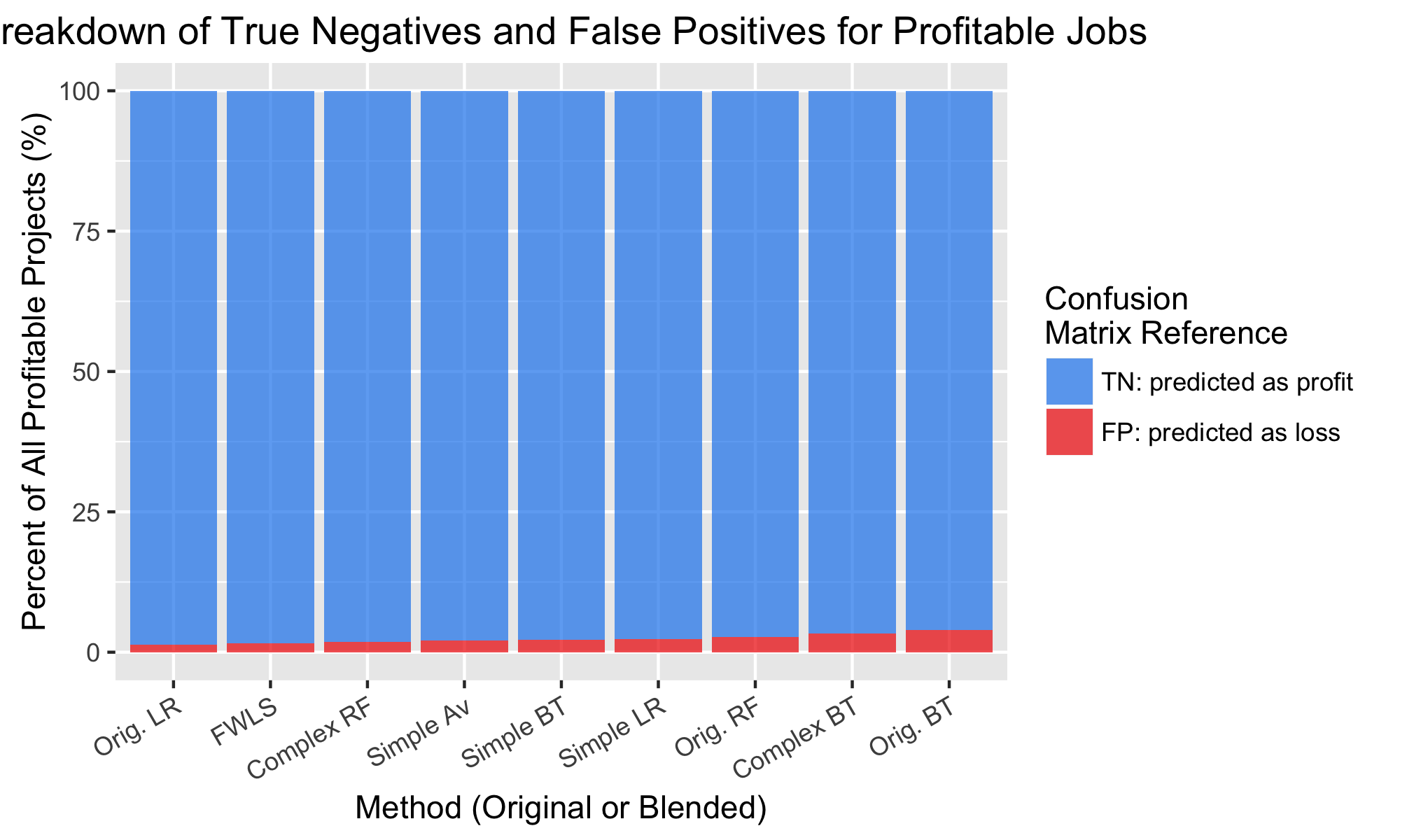
Profit Curve of the best performing method: the simple Logistic Regression Blended Method



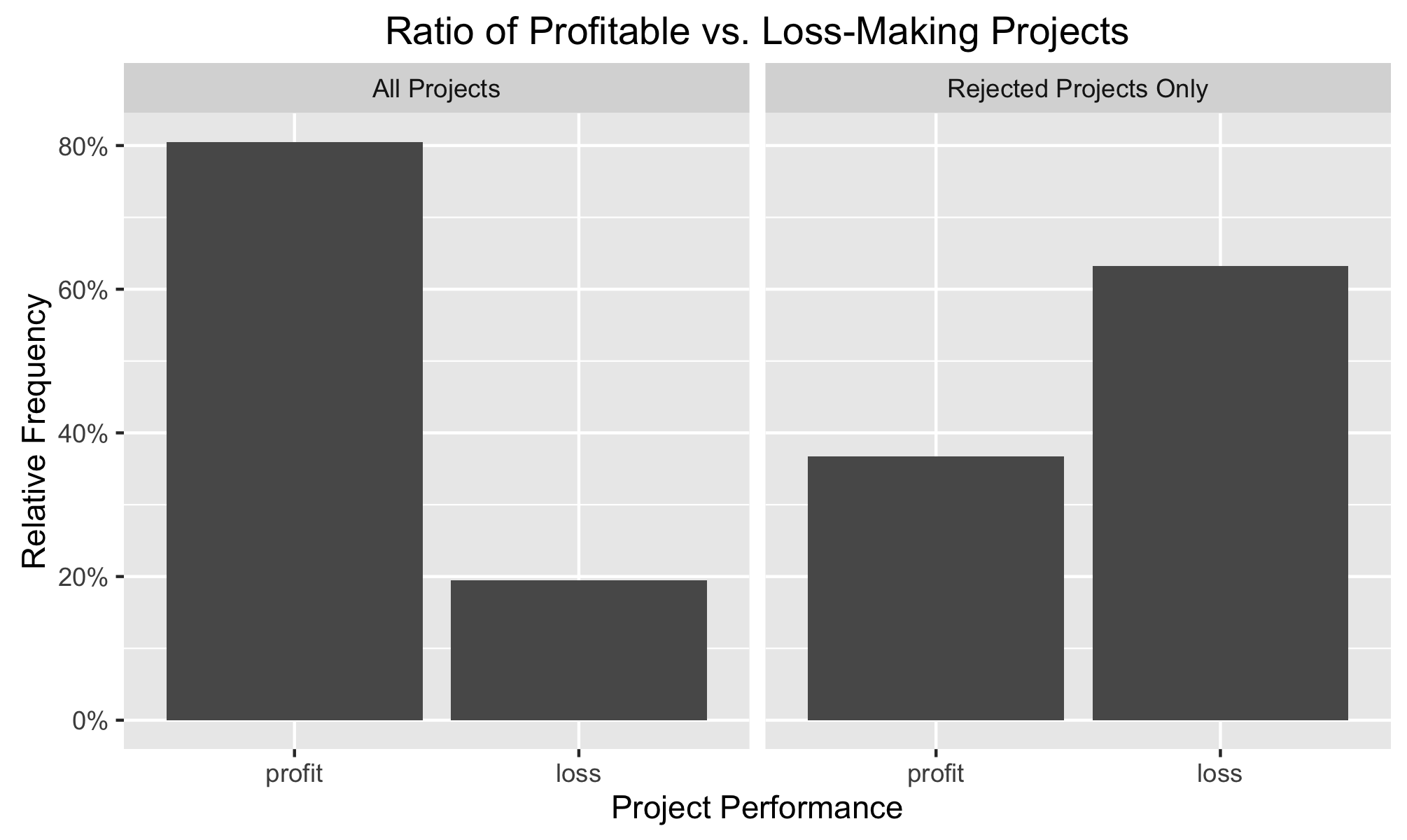
Distribution of projects according to the probability outputs from a typical simple Logistic Regression blended model. A bargraph of counts and relative profits are shown.



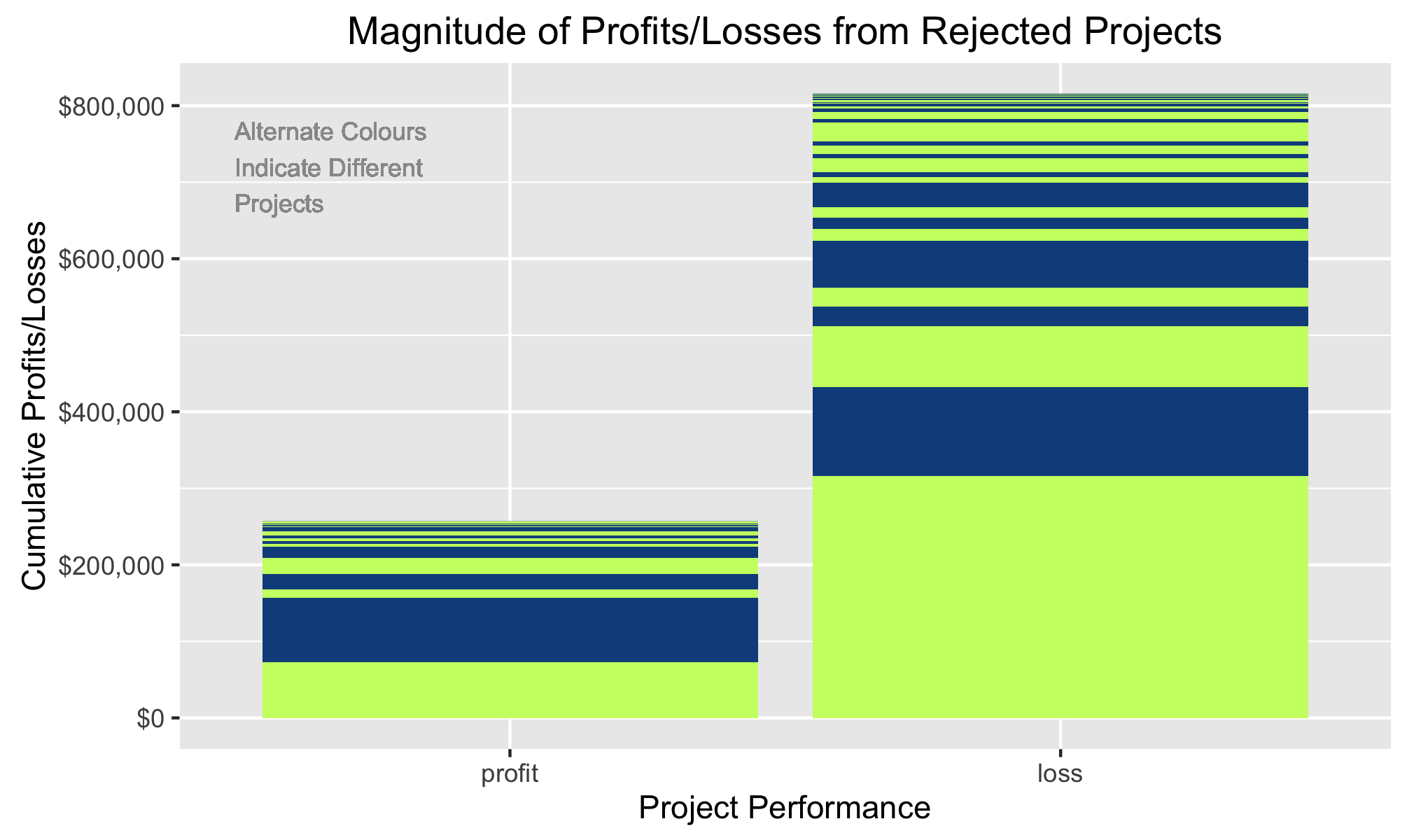
True positive and false negative rates for each blended method using a typical model from each - displayed as a bar graph.



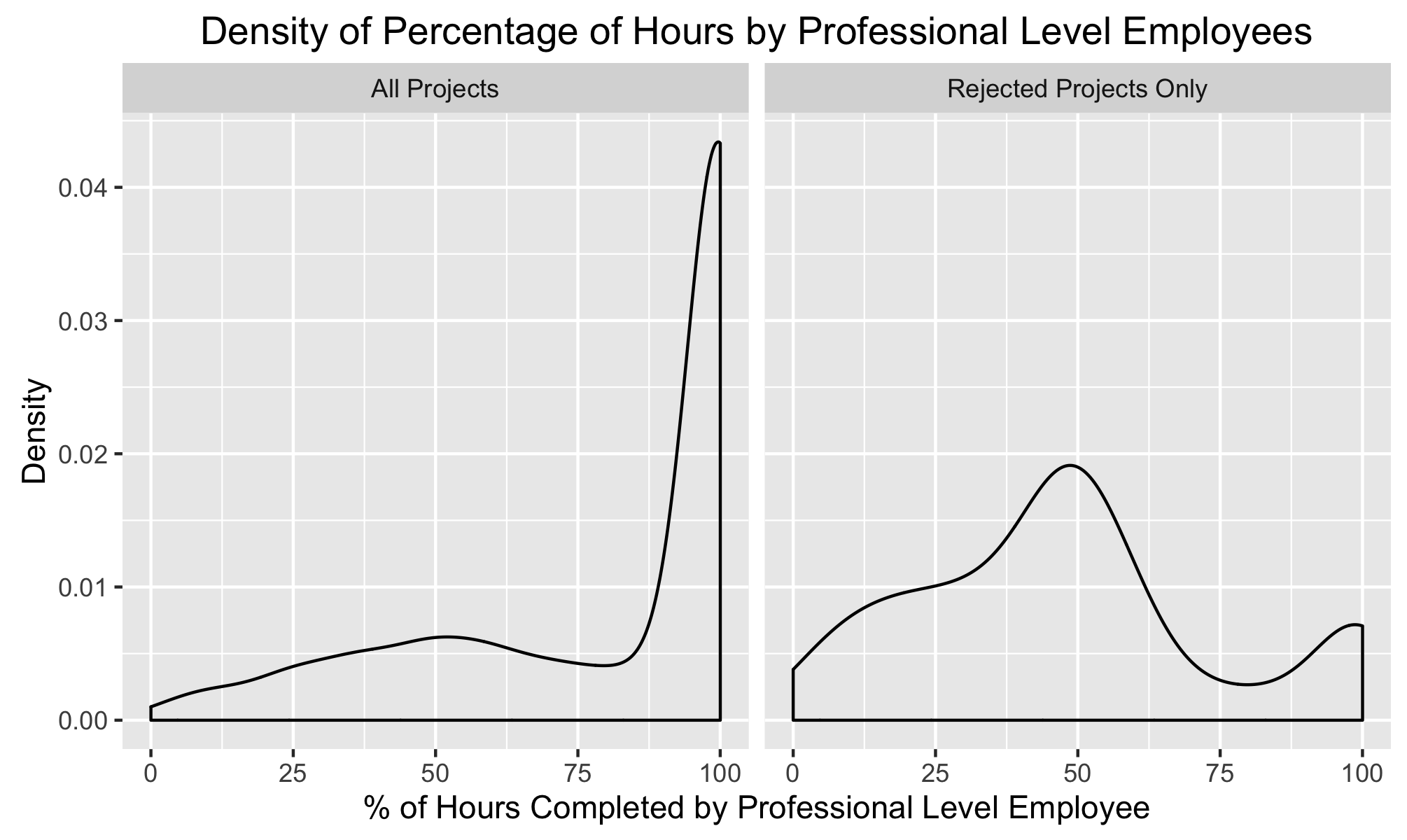
True negative and false positive rates for each blended method using a typical model from each - displayed as a bar graph.



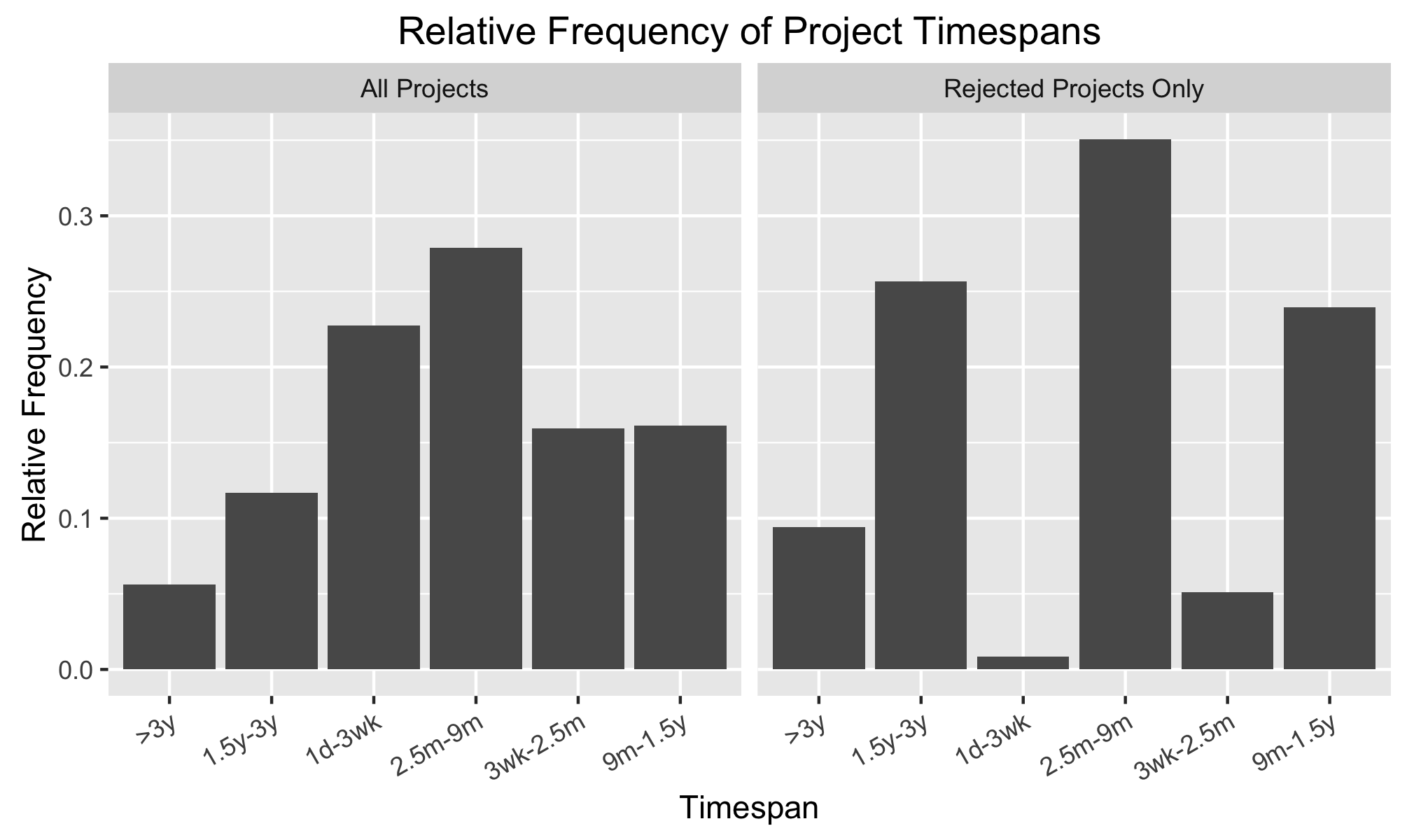
Proportion of profitable and loss-making jobs in the full data set of projects vs. the rejected projects only. Predictions from a typical simple Logistic Regression were used.



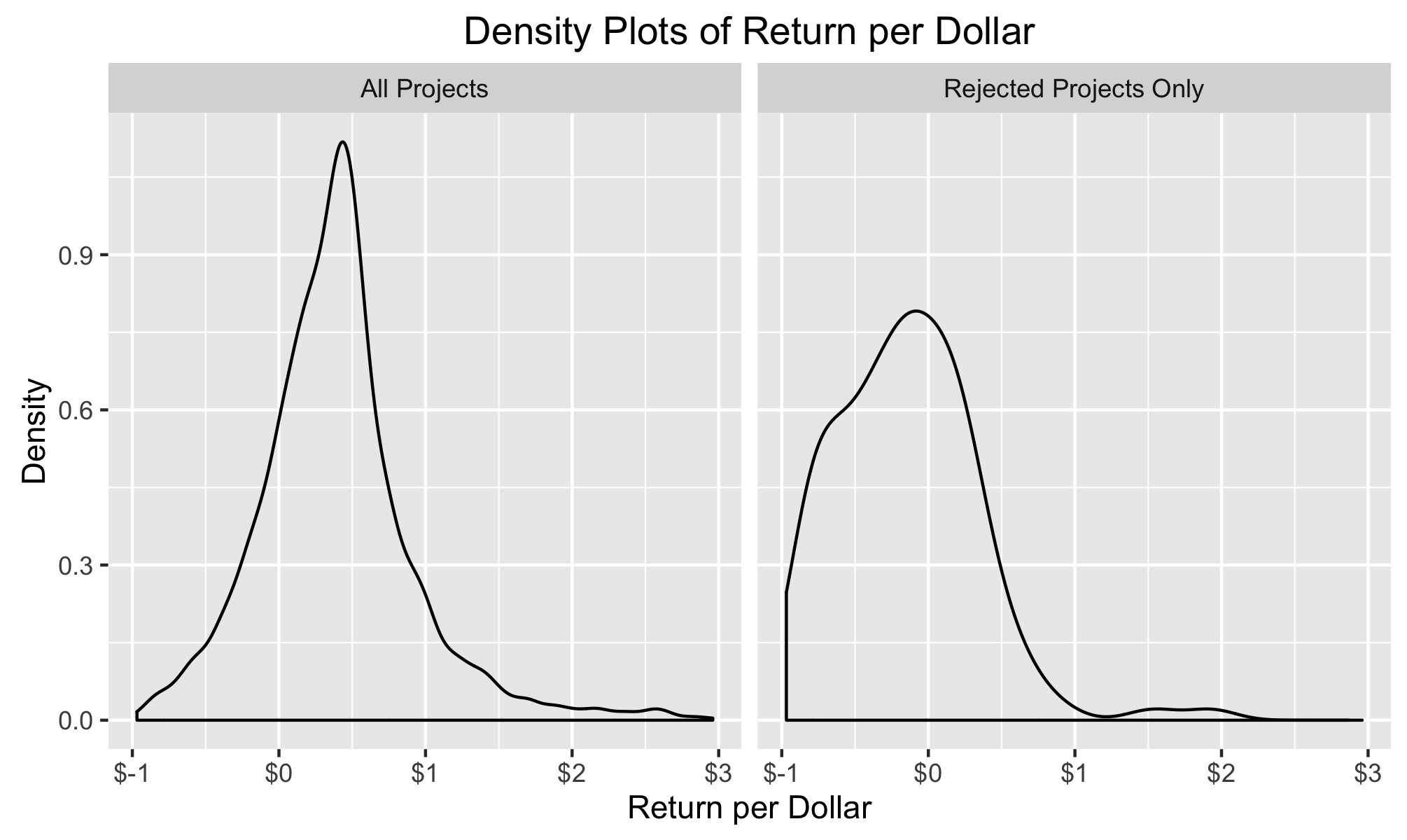
Proportion of the absolute value of profits and losses from profitable and loss-making jobs in the rejected projects only. Colours are alternated between different jobs to indicate profit/loss magnitude. Predictions were from a typical simple Logistic Regression.



Distribution of the percentage of hours completed by a professional level employee in the full data set of projects vs. the rejected projects only. Predictions from a typical simple Logistic Regression were used.



Proportion of each timespan category for projects in the full data set vs. the rejected projects only. Predictions from a typical simple Logistic Regression were used.



Distribution of 'return per dollar' in the full data set of projects vs. the rejected projects only. Predictions from a typical simple Logistic Regression were used.