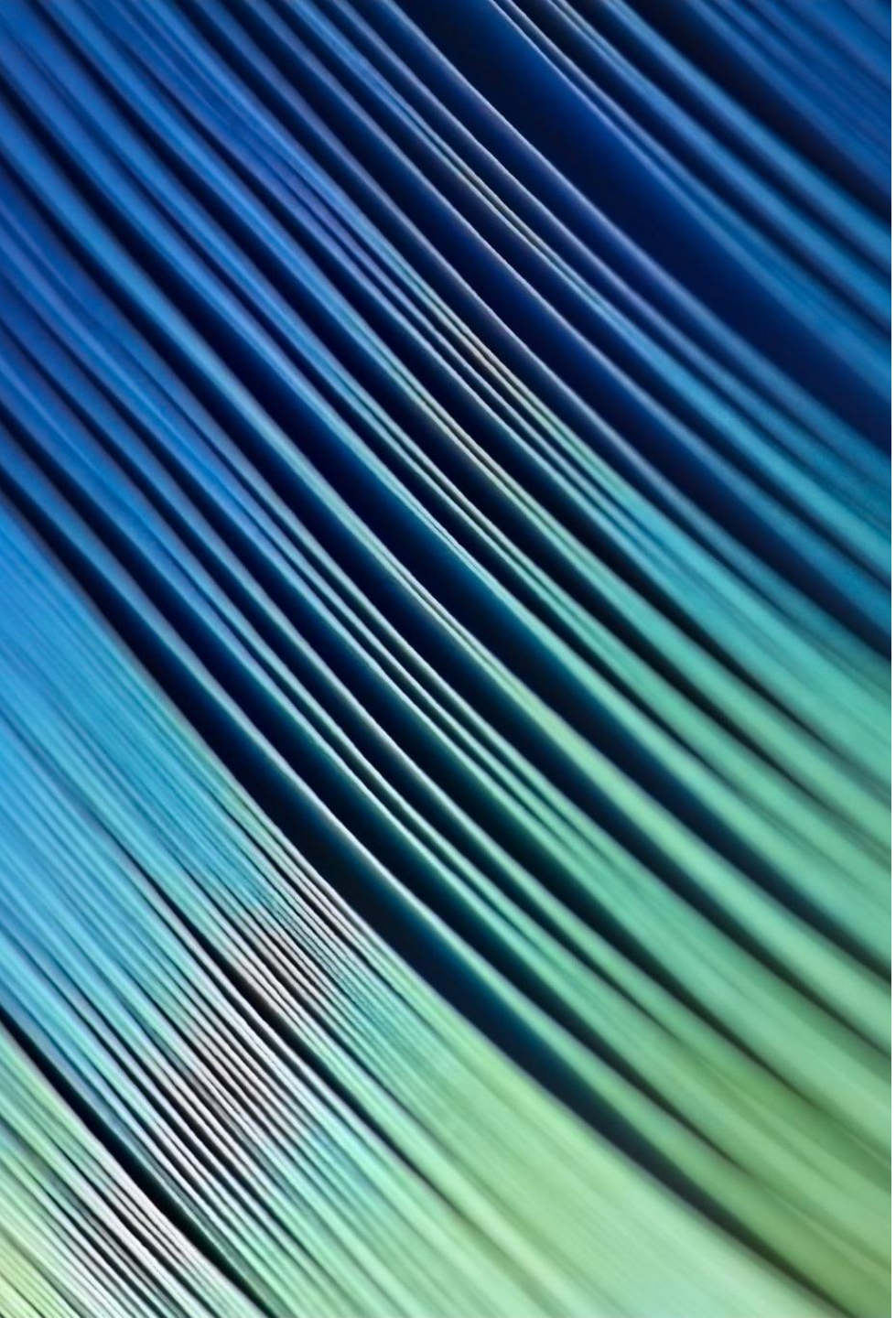
Project Presentation

Max Chis



Methodology

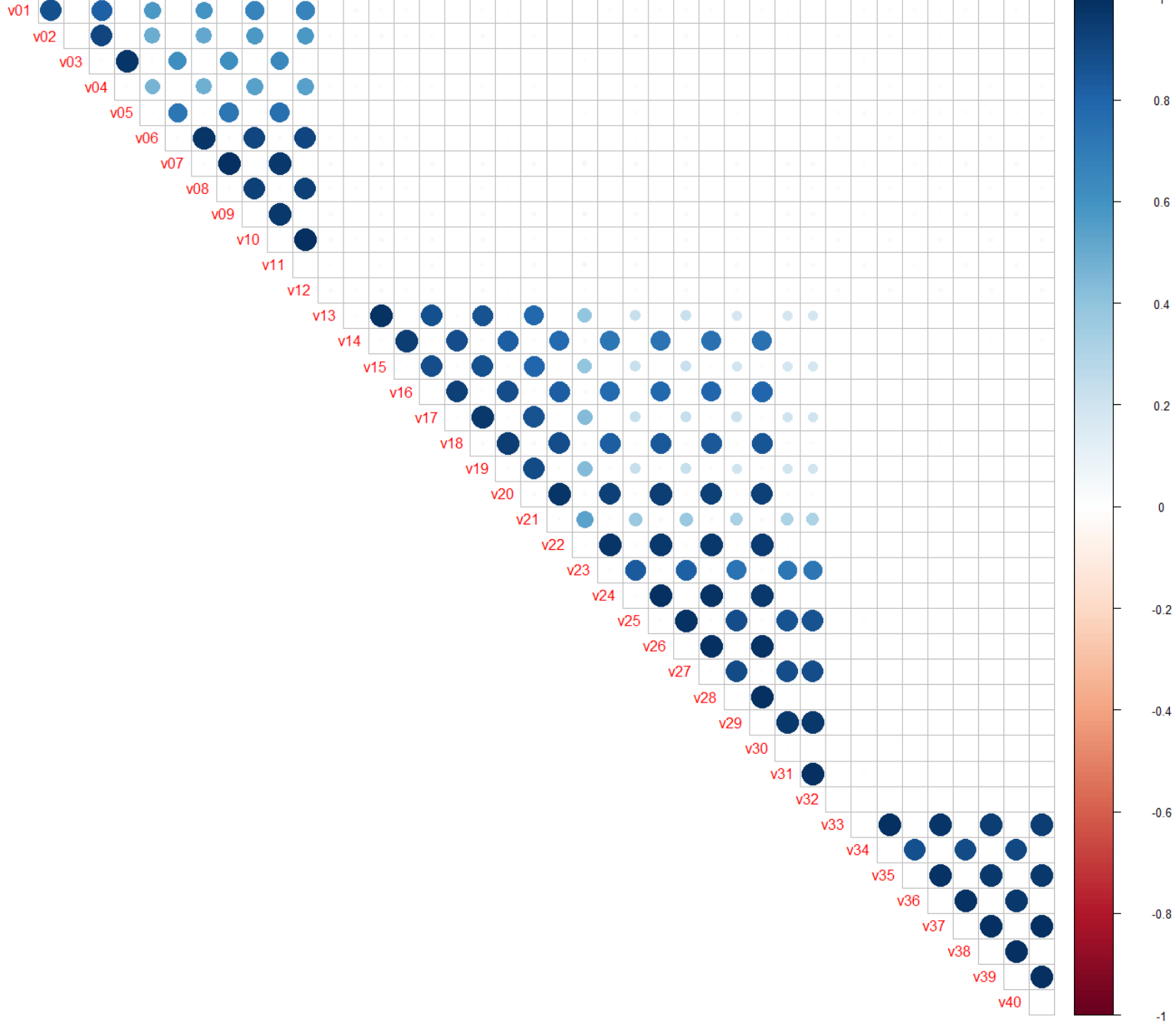
Data exploration with emphasis on variable correlation, relationship of variable values with event probability and continuous output

Regression and Classification Models trained, including:

- Linear and Quadratic Additive
- All pairwise interactions
- Interactions of splines of selected variables
- Elastic Net
- K Nearest Network
- Support Vector Machine
- Gradient Boosted Tree
- Random Forest
- Neural Networks

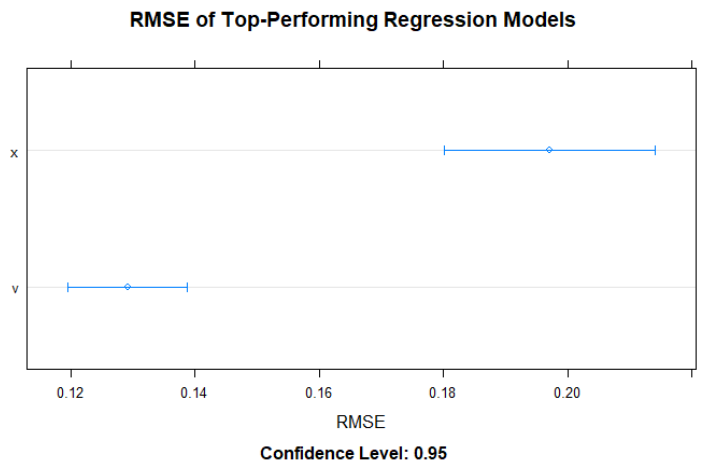
Variable Correlation

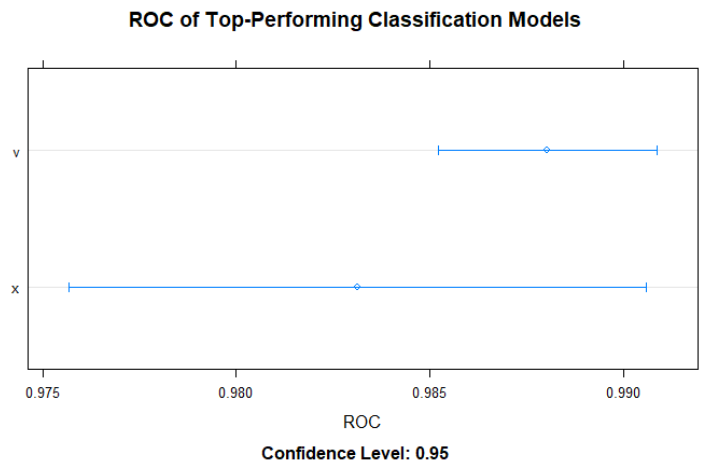
- V variables more likely to be strongly correlated, compared to X variables
- Highest X variable correlation was 0.15ρ
- Highest V variable correlation was 1ρ

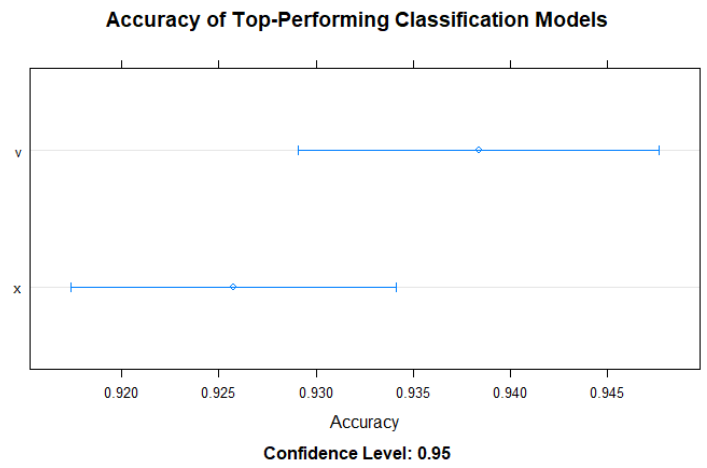


Comparing Performance Between V and X Models

- Top-performing V model outperformed top-performing X model in Regression
- Reverse is true for classification in terms of Accuracy
- Comparable performance for classification in terms of ROC

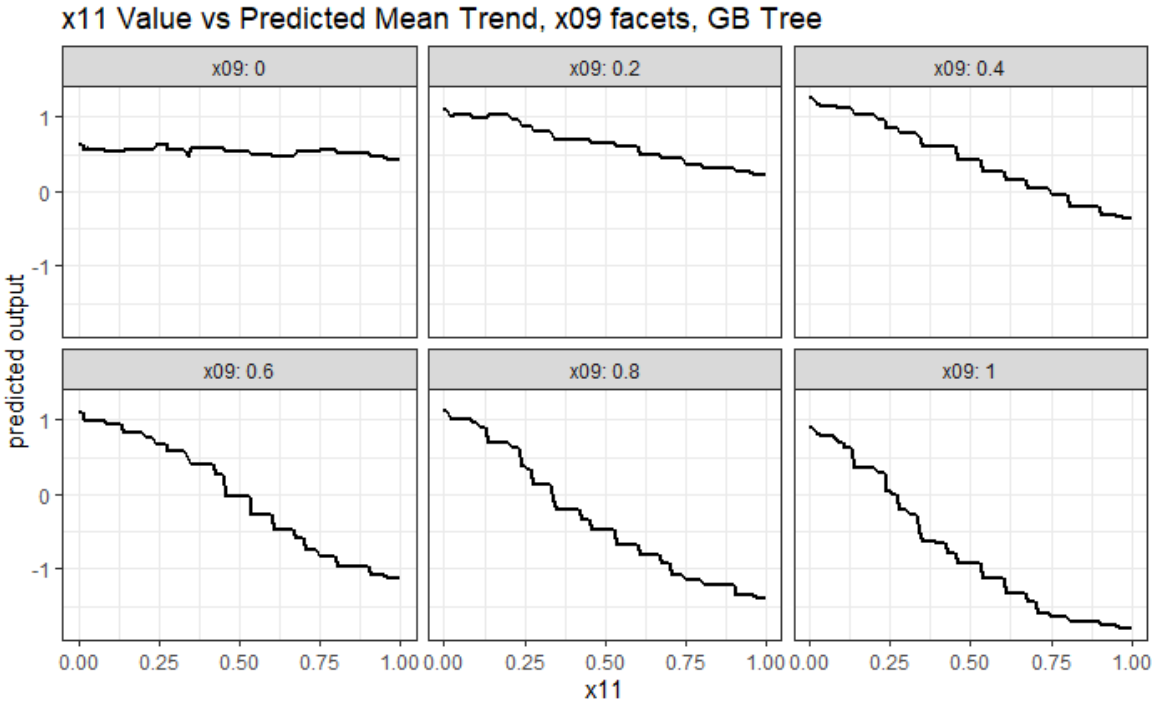




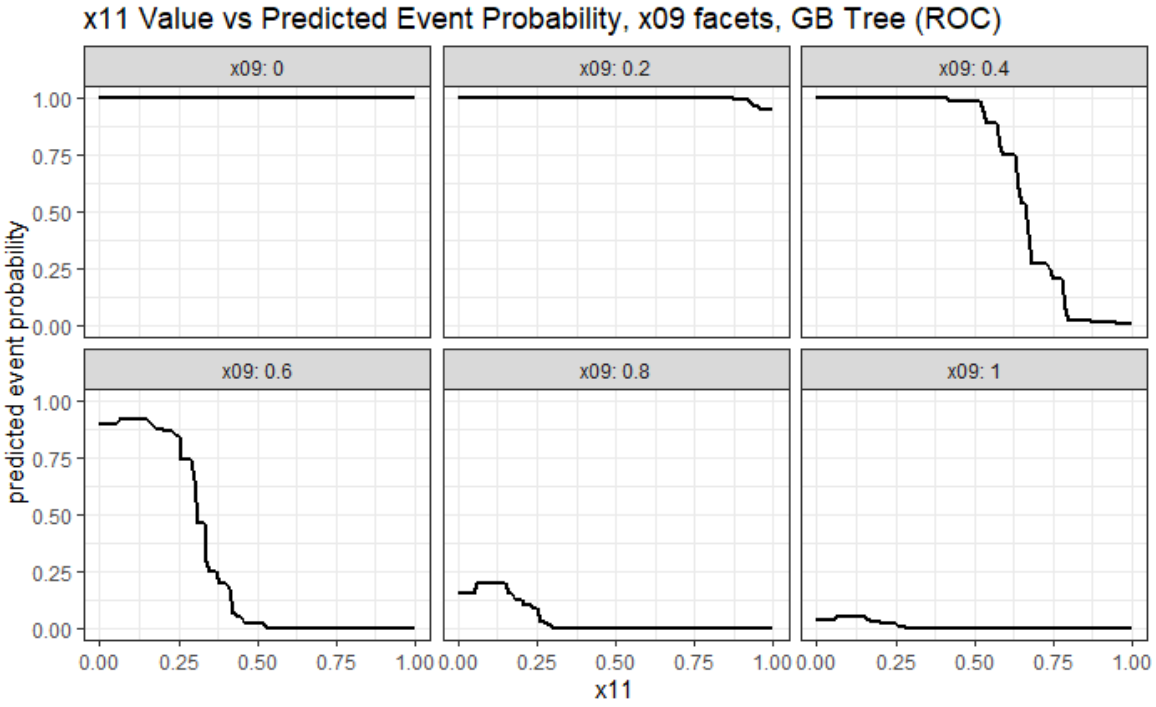


Best-Performing Model – X Variables

Continuous Output (Regression)



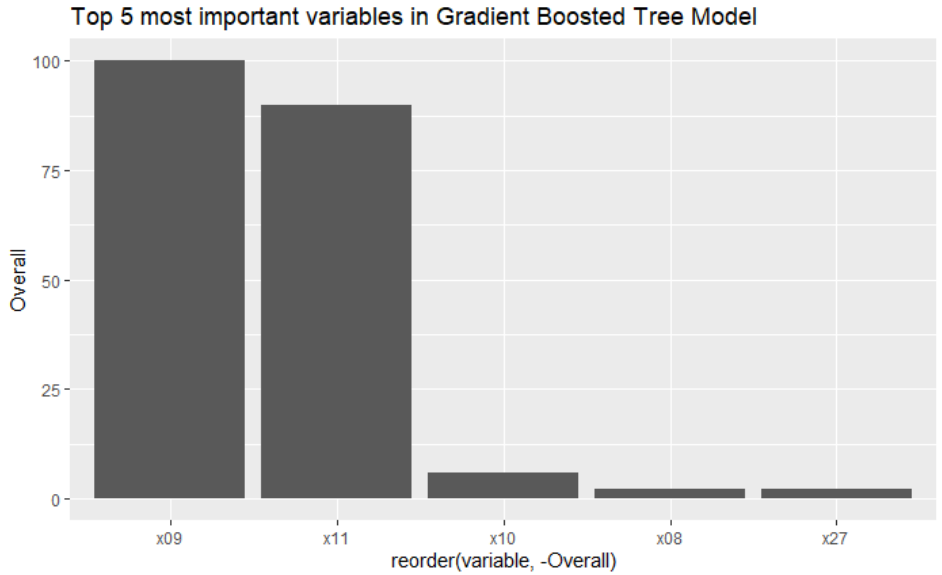
Event Probability (Classification)

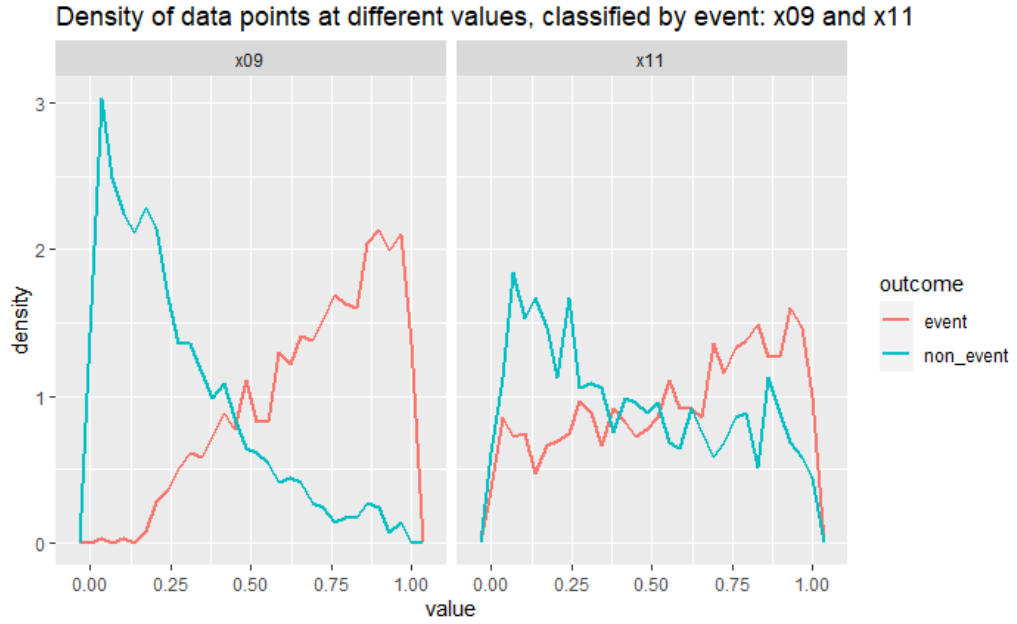


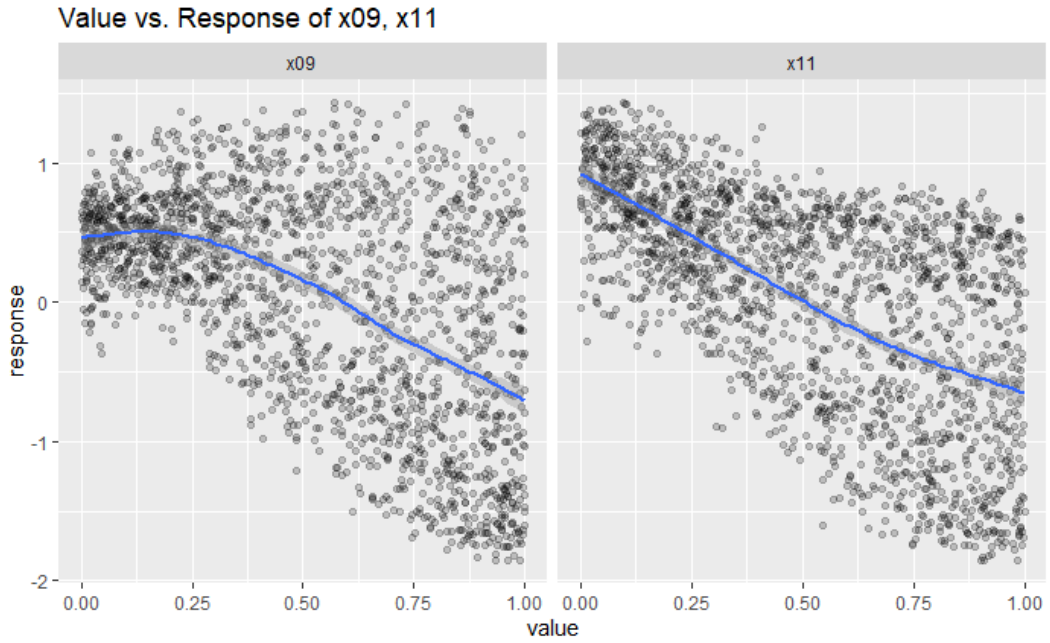
- Model fitted using gradient-boosted tree.
- Maximizing x09 and x11 minimizes continuous output and error probability

Variable Importance – X Variables

- X09 and x11 consistently found to be most important variables
- All other variables have marginal importance to model.
- Importance of X09 and X11 consistently found across different models.

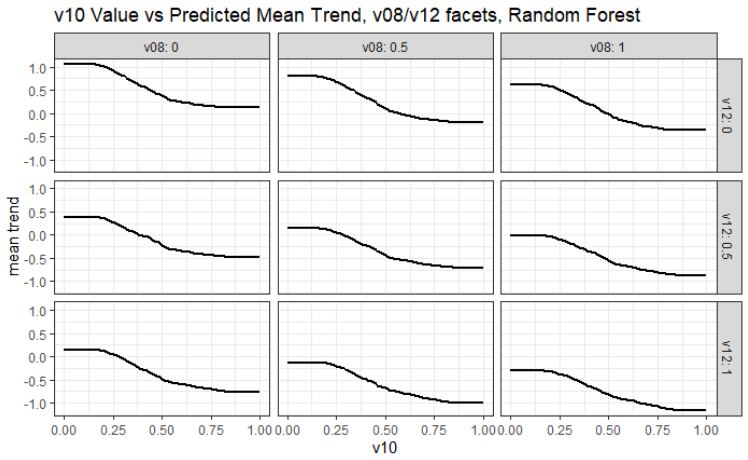


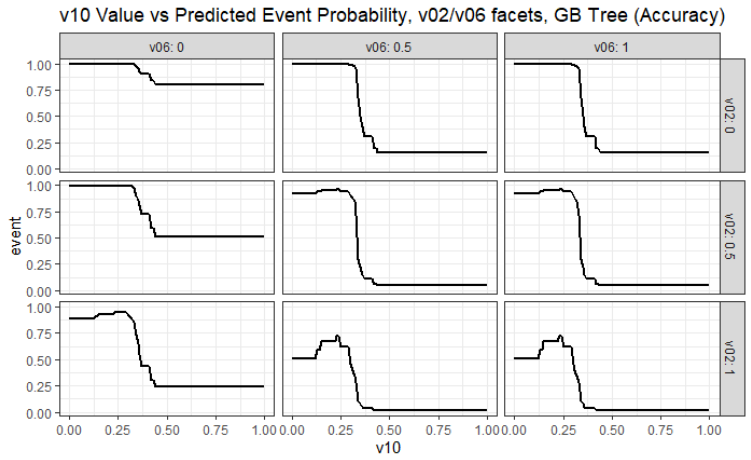


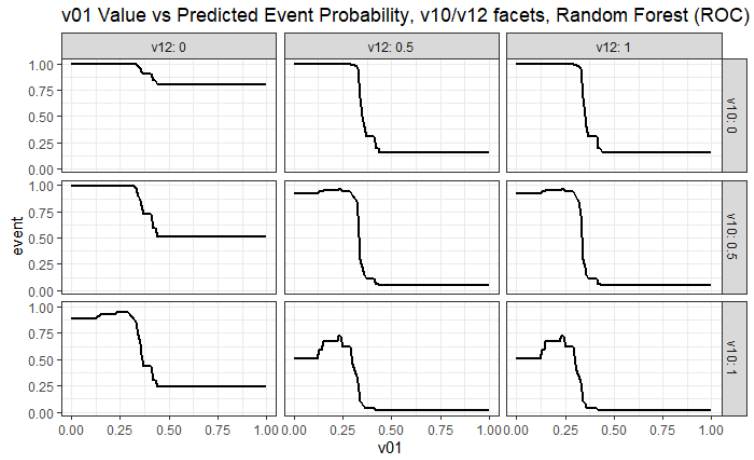


Best Performing Model – V Variables

- Gradient-Boosted tree maximized Accuracy (in Classification) and RMSE (in Regression)
- Random Forest maximized ROC scores in Classification
- Maximizing v01,v02, v06, v08, v10, v12 minimizes continuous output and event probability.

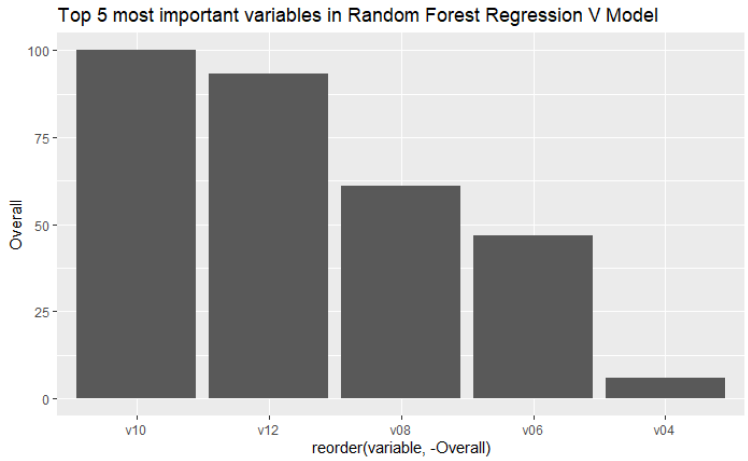


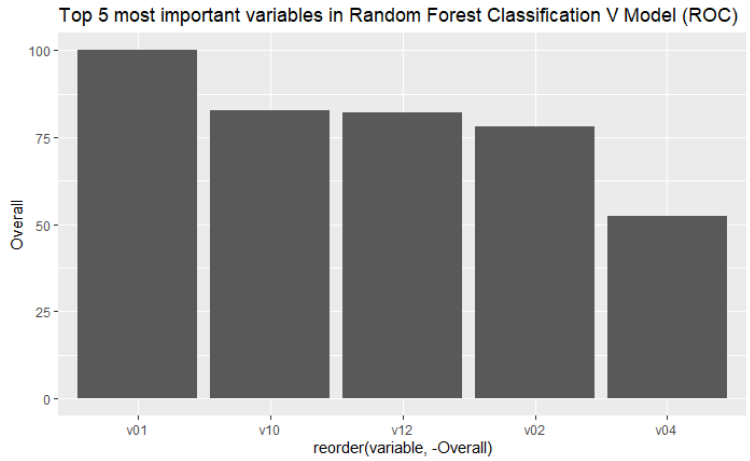


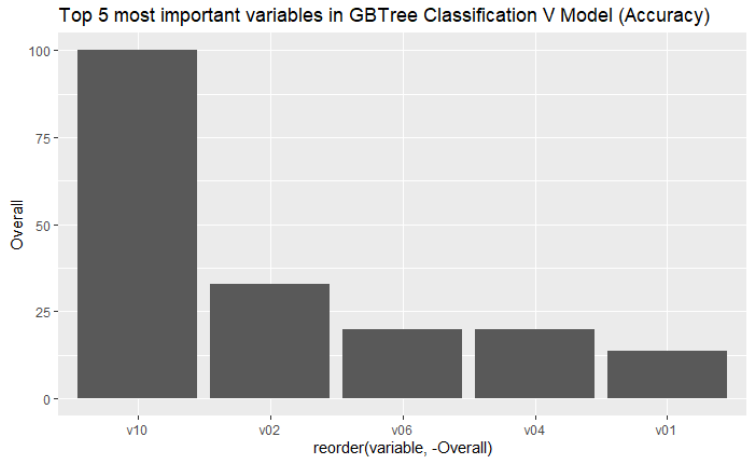


Variable Importance – V Variables

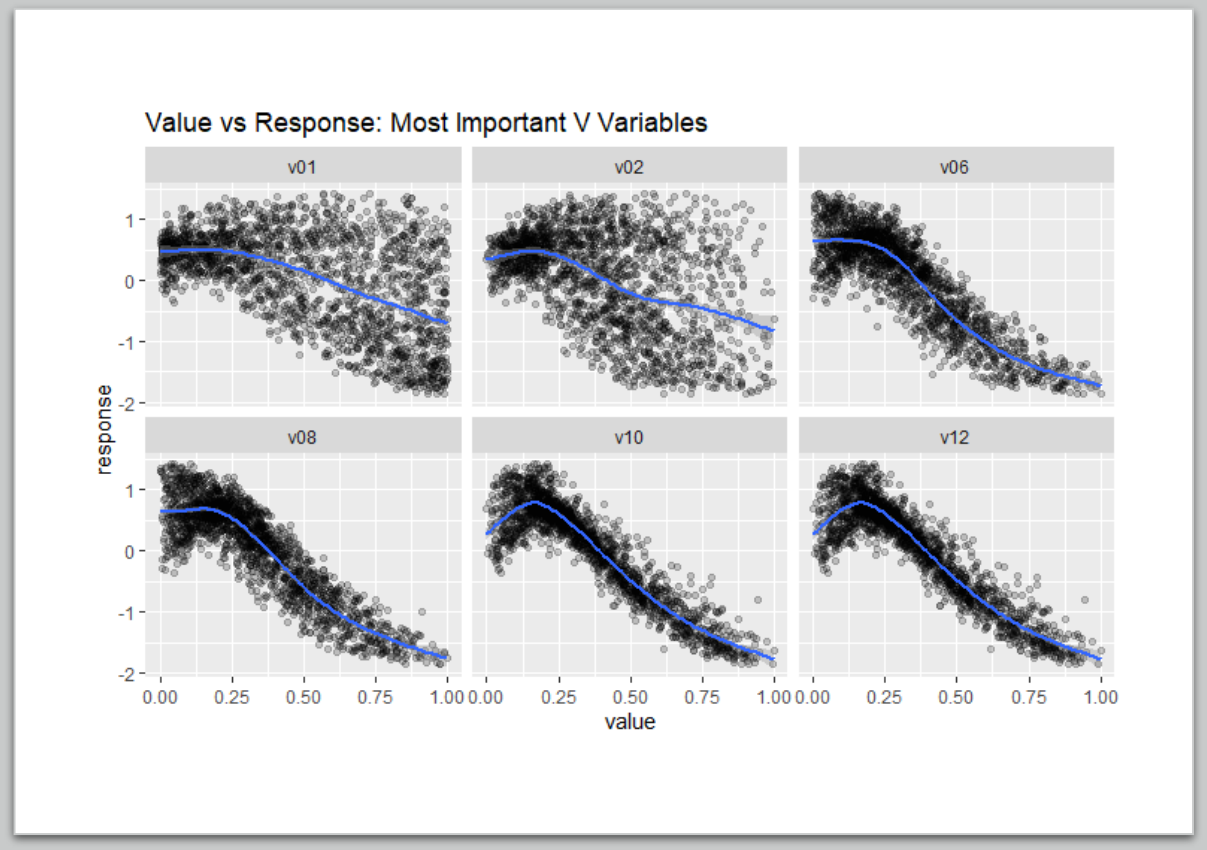
- V models disagreed more than X models on which variables were most important
- Consistent top-performers included v01, v02, v06, v08, v10, and v12

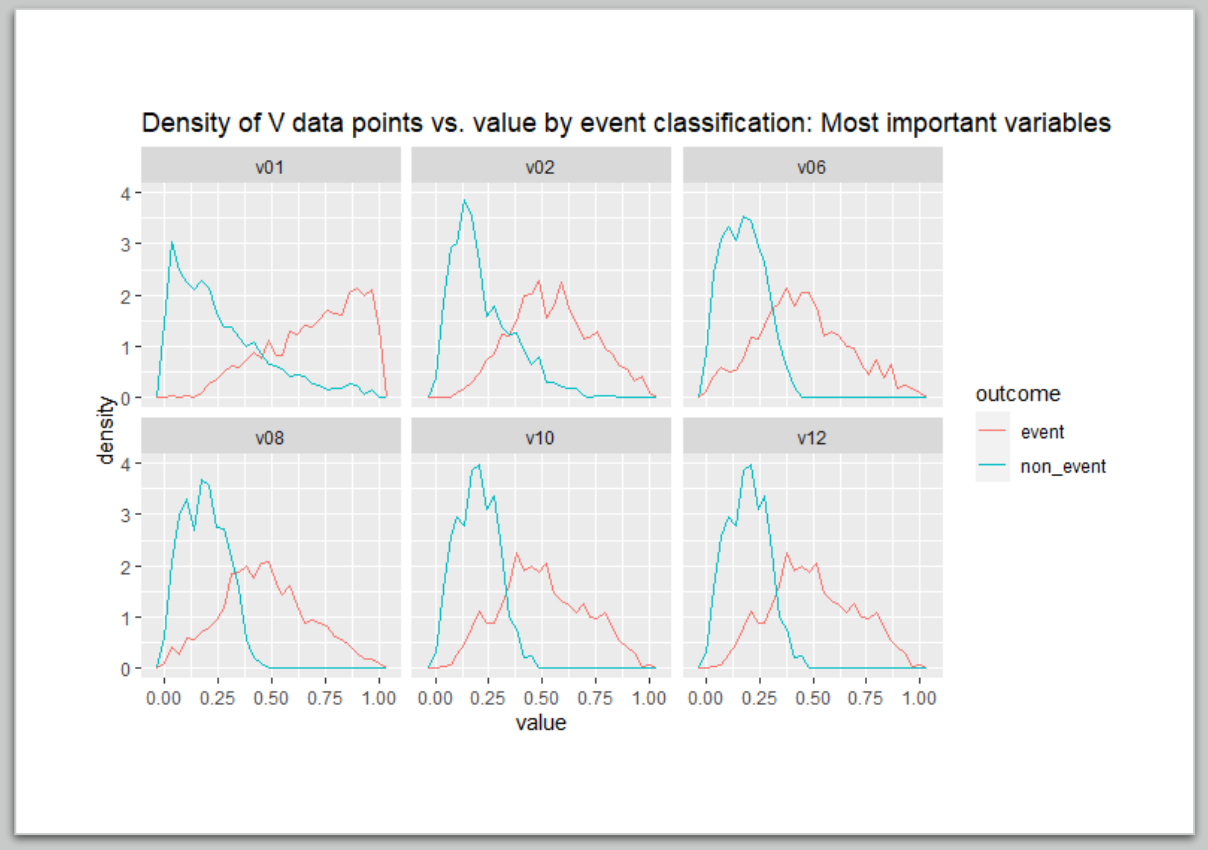






Plotting Most Important V Variables





Performance on Holdout Set

X Holdout Set (GBTree) utilized

The holdout test set performance metrics associated with your model are displayed below.

Note: nothing is shown until you upload the CSV file.

.metric	.estimator	.estimate
rmse	standard	0.21
rsq	standard	0.93
mae	standard	0.16
accuracy	binary	0.91
mn_log_loss	binary	0.21
roc_auc	binary	0.98