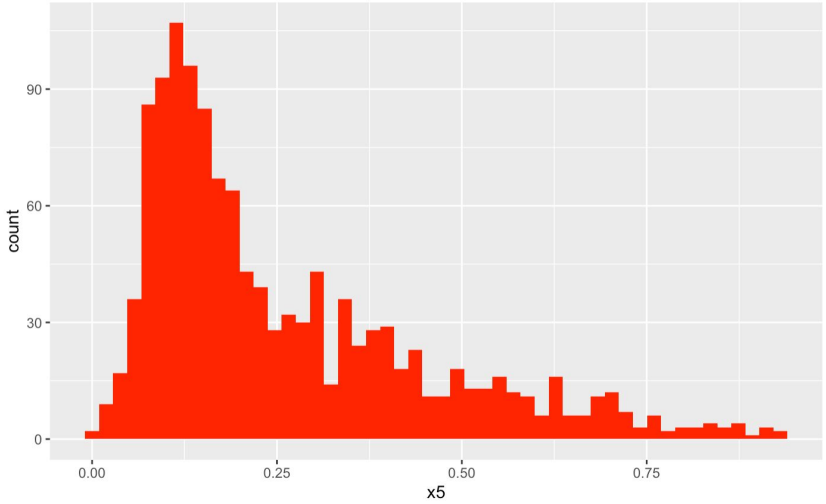
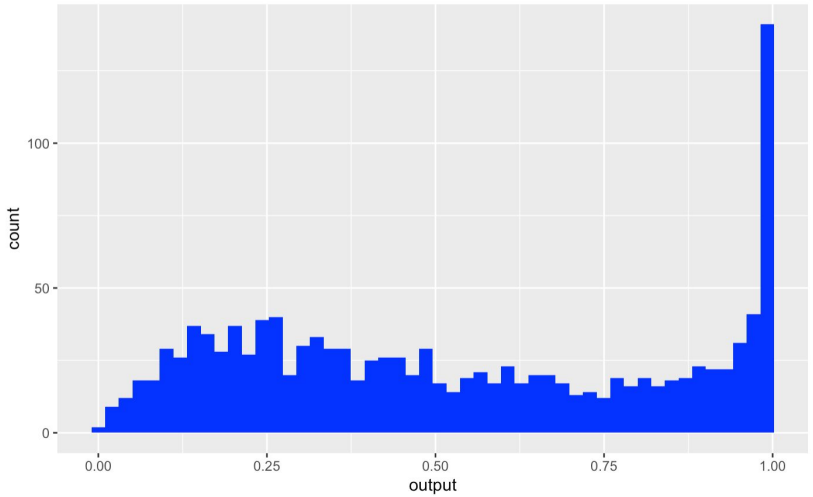
Modeling Solutions To Limit Corrosion

Sponsored By PPG Industries

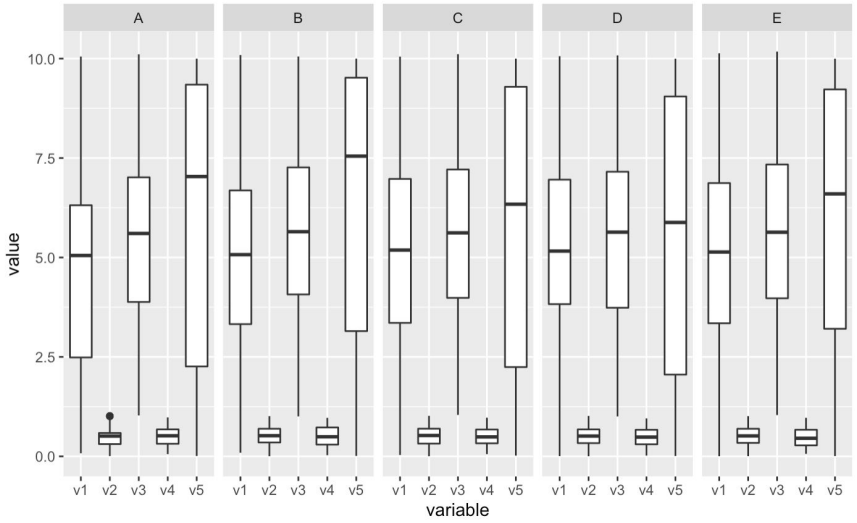
EDA: Examining Output Distributions

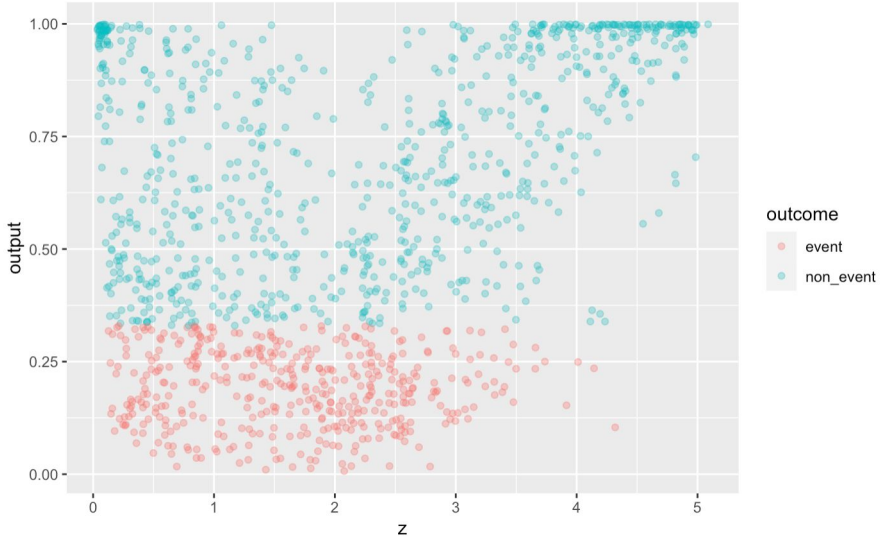




Inspecting the output showcased that many samples in the sample data had corroded completely. X5's distribution was also noticeably left skewed.

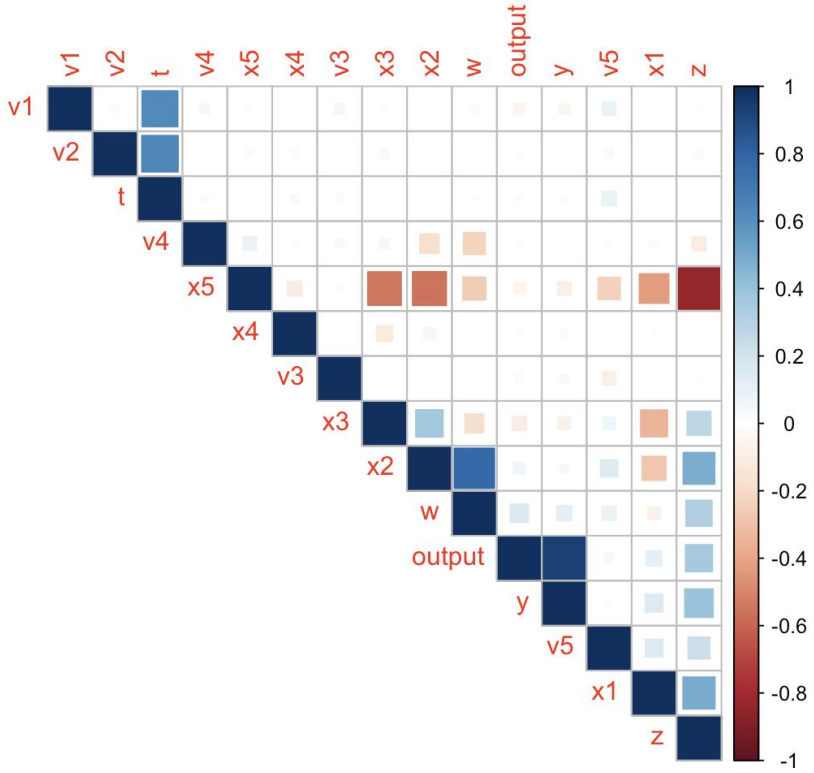
EDA: Data Insights





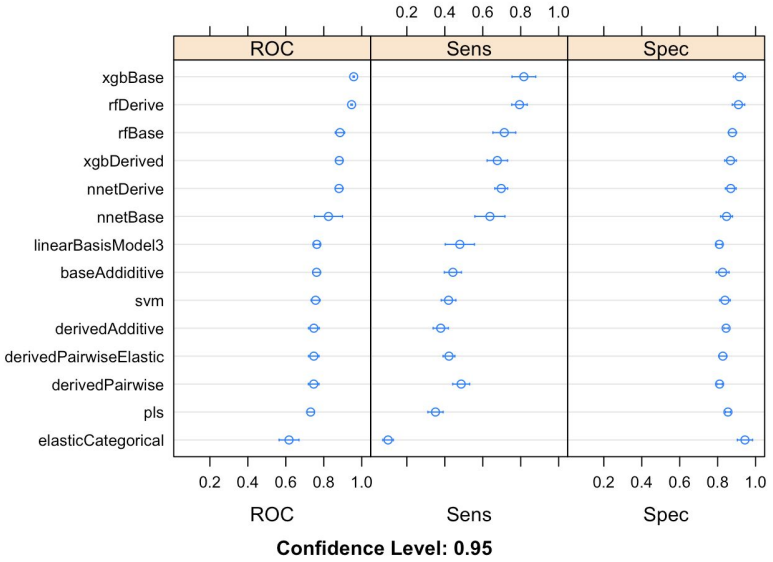
Boxplots of the manufacturing and chemistry input values at each machine revealed that every machine was given an approximately equivalent sample set. Another insight was that high values of z produced many non-events (corroded samples).

EDA: Feature Correlation



Z and X1 had a fairly strong positive correlation, while x5 and z had a very strong negative correlation.

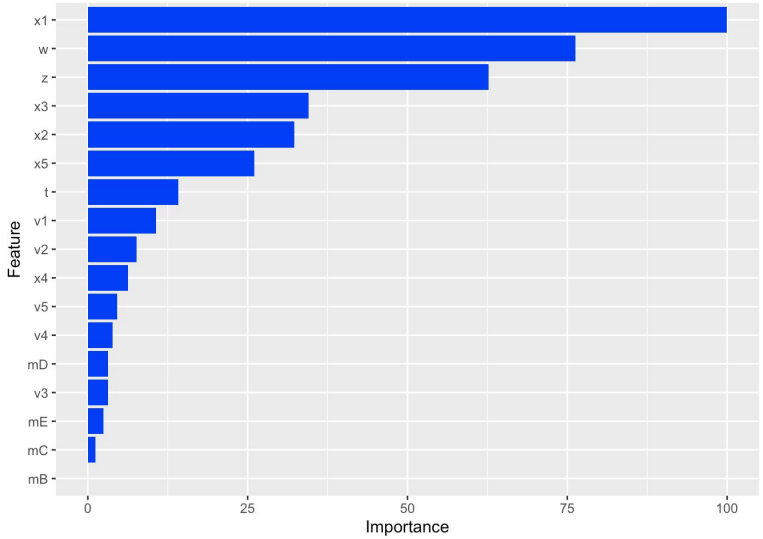
Model Selection: Comparing Model Performance

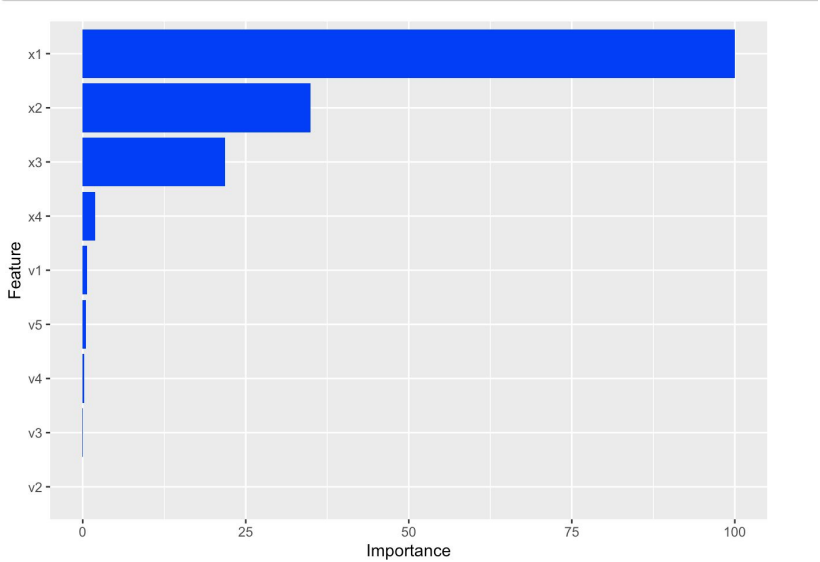


Model <chr>	RMSE <dbl>
baseModelAdditive	1.7605565
derivedModelAdditive	1.7721405
linearCreatedMod1	1.3507039
linearCreatedMod2	1.9805634
ElasticRegression	2.2037546
NeuralNetBase	1.1129444
NeuralNetExpand	1.2828495
RForestBase	1.1753743
RForestExpand	0.7292399
GradientBoostBase	0.6260892

After training and tuning many different types of models, I discovered that base feature set gradient boosted trees and the expanded feature set random forest models performed the best for both predicting binary classification and the logit output.

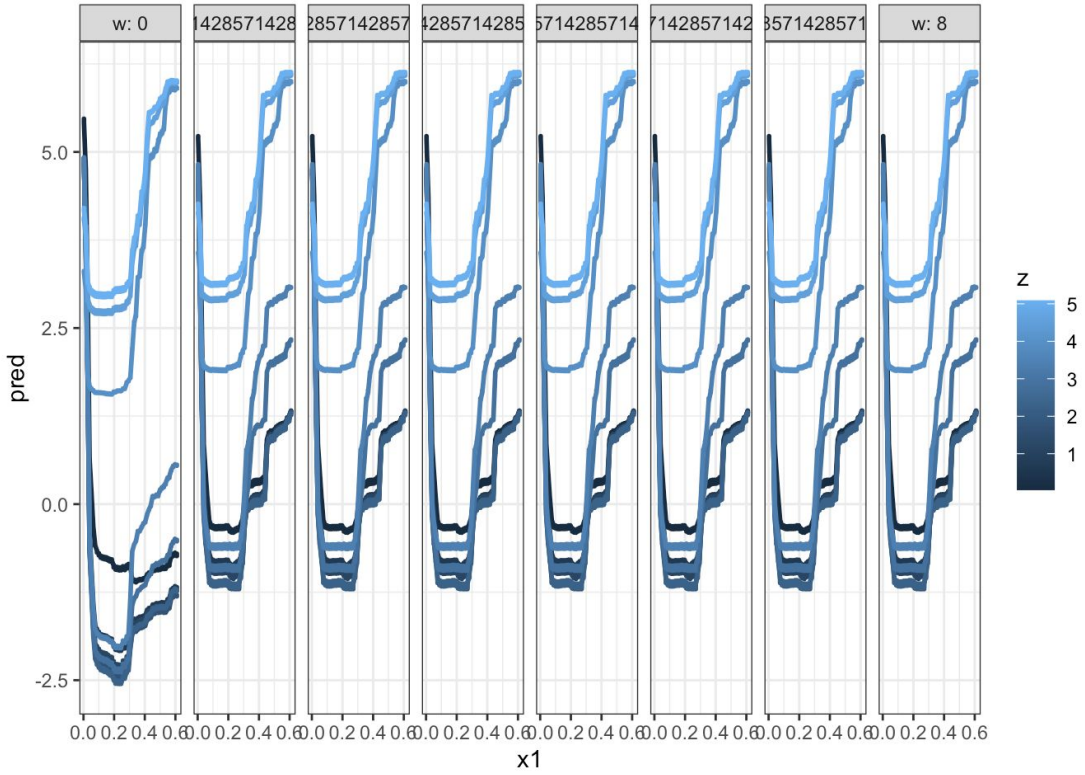
Analysis: Feature Importance





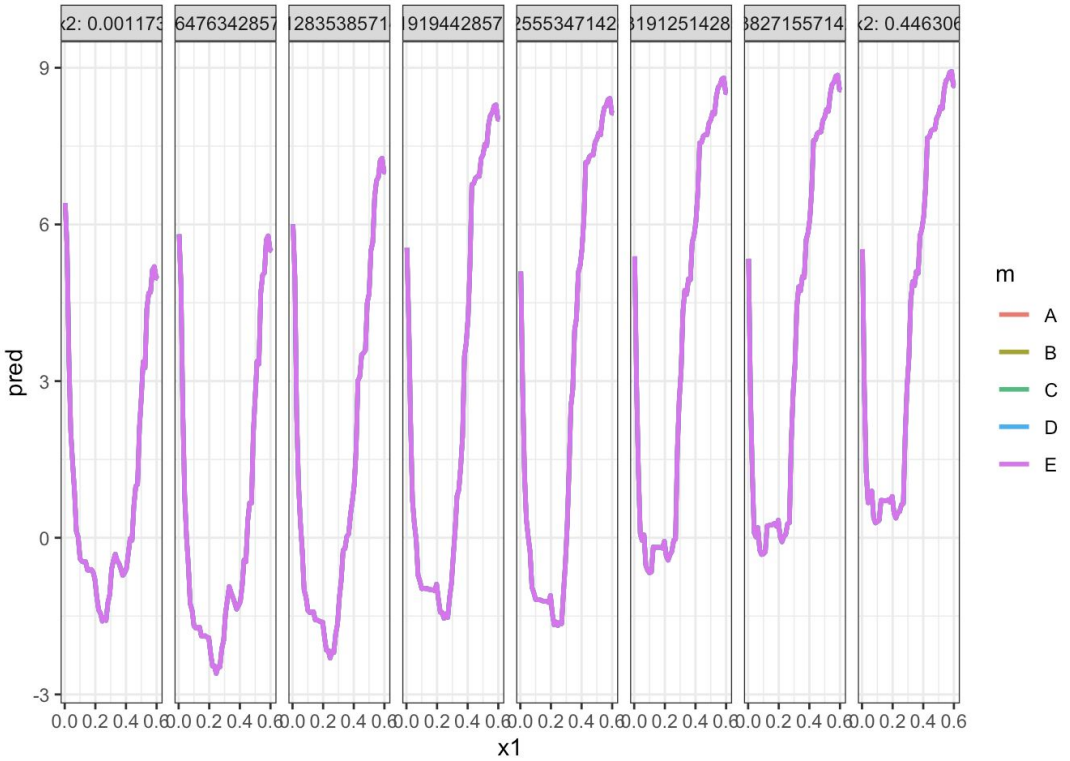
The derived features of z and were critical to my random forest expanded model and x1 was very important across both top models.

Predictions: Optimizing Input Settings



Graphing predictions using my random forest model showcased that a *w* value of 0 to .28, paired with a *z* value of 1 to 2 and a *x1* value of .2 to .3 drastically reduced corrosion.

Predictions: The Categorical Input



Predictions for the logit output with respect to each machine revealed that these optimal chemistry input settings do not vary much between different machines, so they can be used for all of them

Results: Closing Thoughts

predictionResults

.metric	.estimator	.estimate
rmse	standard	0.964862338155856
rsq	standard	0.595259690858915
mae	standard	0.764980576449625
accuracy	binary	0.894009216589862
mn_log_loss	binary	0.342130205593
roc_auc	binary	0.936224489795918

My top model test predictions yielded a high accuracy, high ROC AUC score, and low rmse score. This indicates its utility in making future predictions in this field.