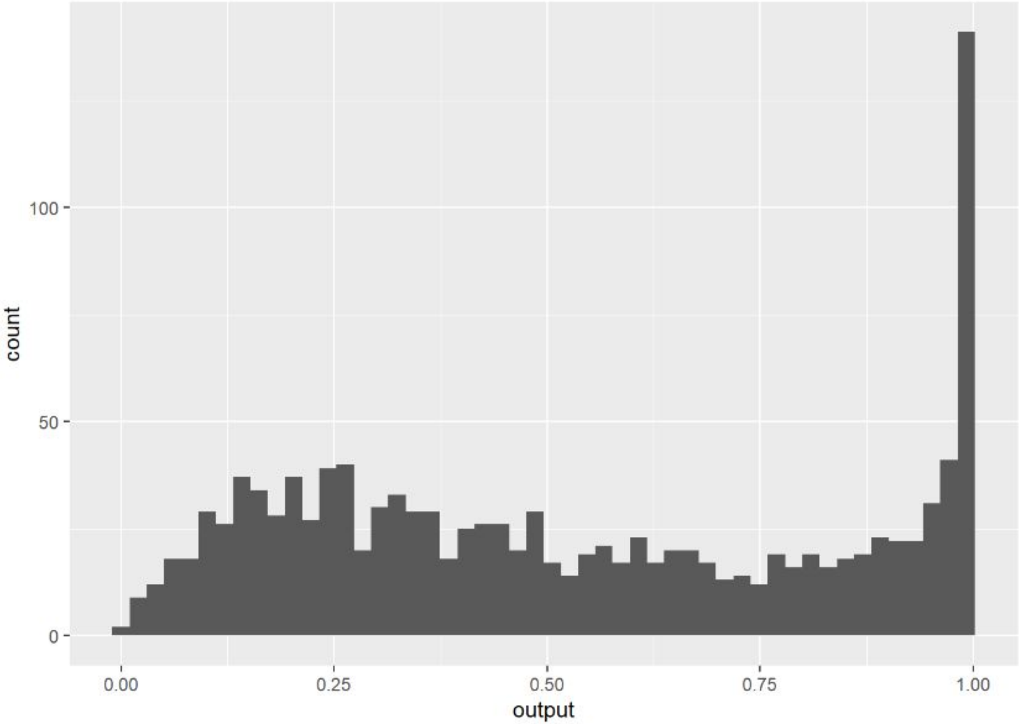
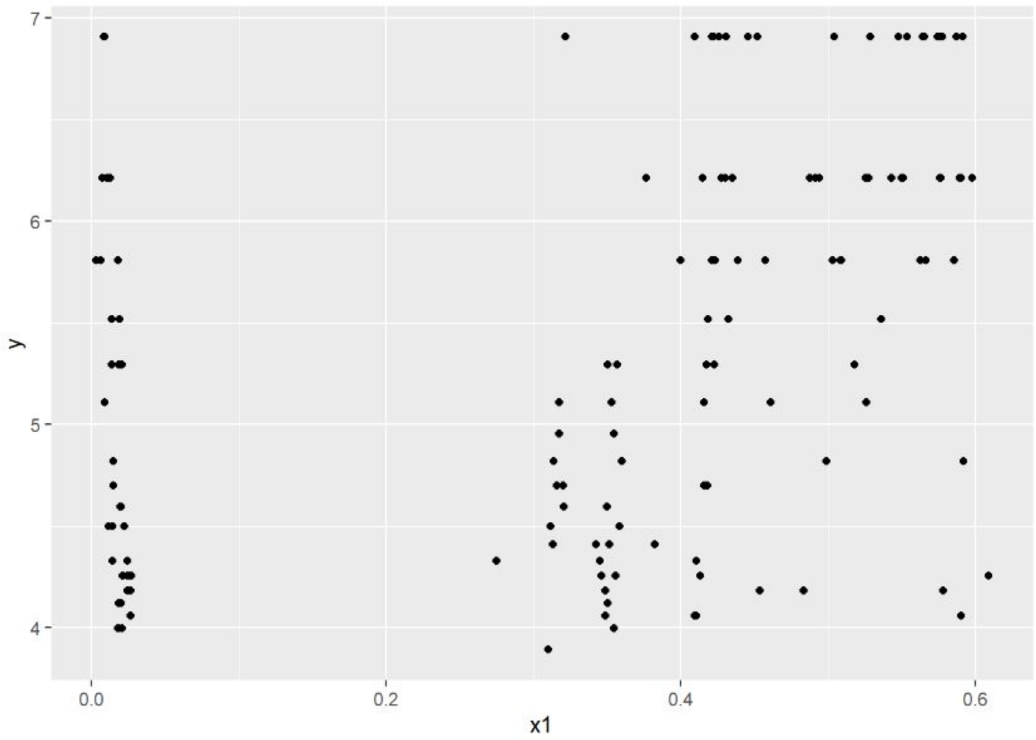
Minimal Corrosive Surface Coatings

Anna Heltz

Visualizations from EDA

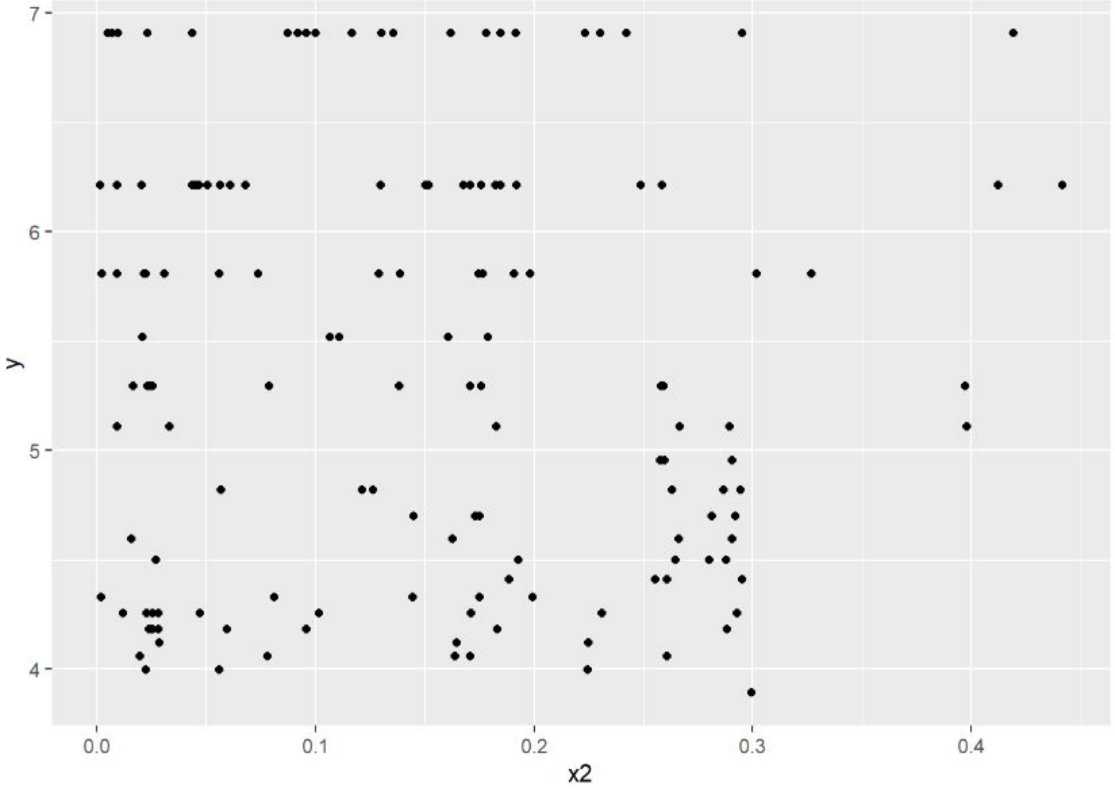


As you can see, a high proportion of the data is highly corroded. So we will filter this data so we can see the values of the inputs that account for the highly corroded surface coatings.



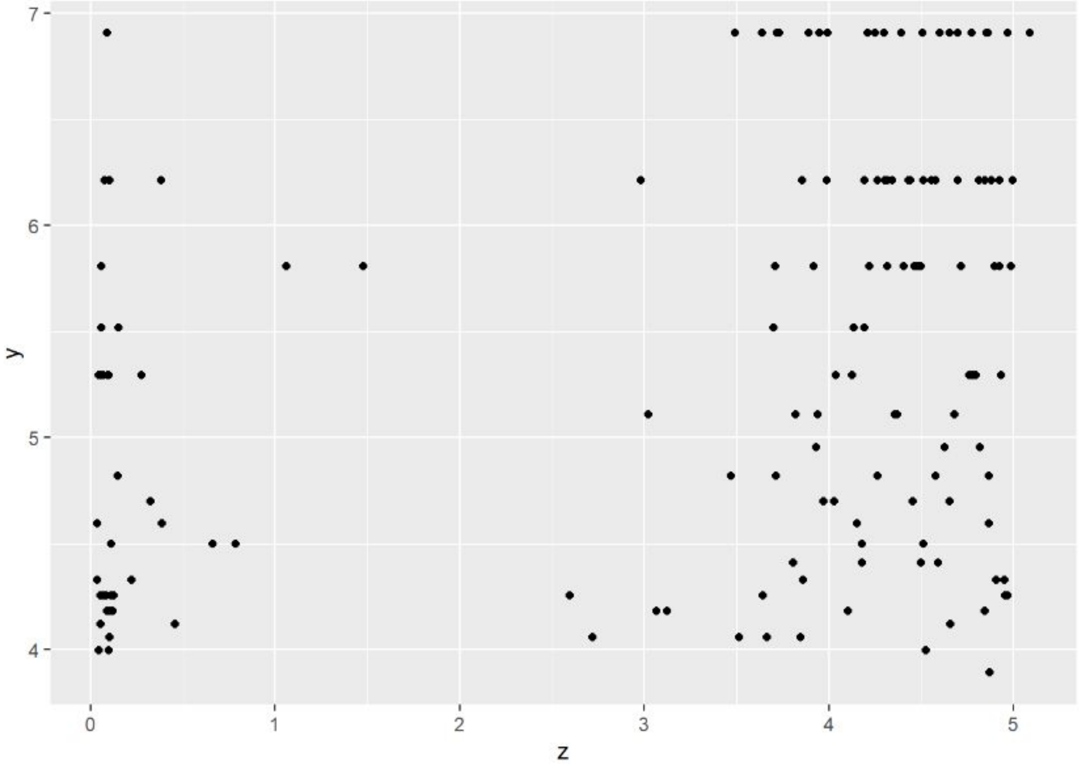
This is a graph of the highly corroded samples.

As you can see, there are no samples that are highly corroded when x1 is between .05-.25.



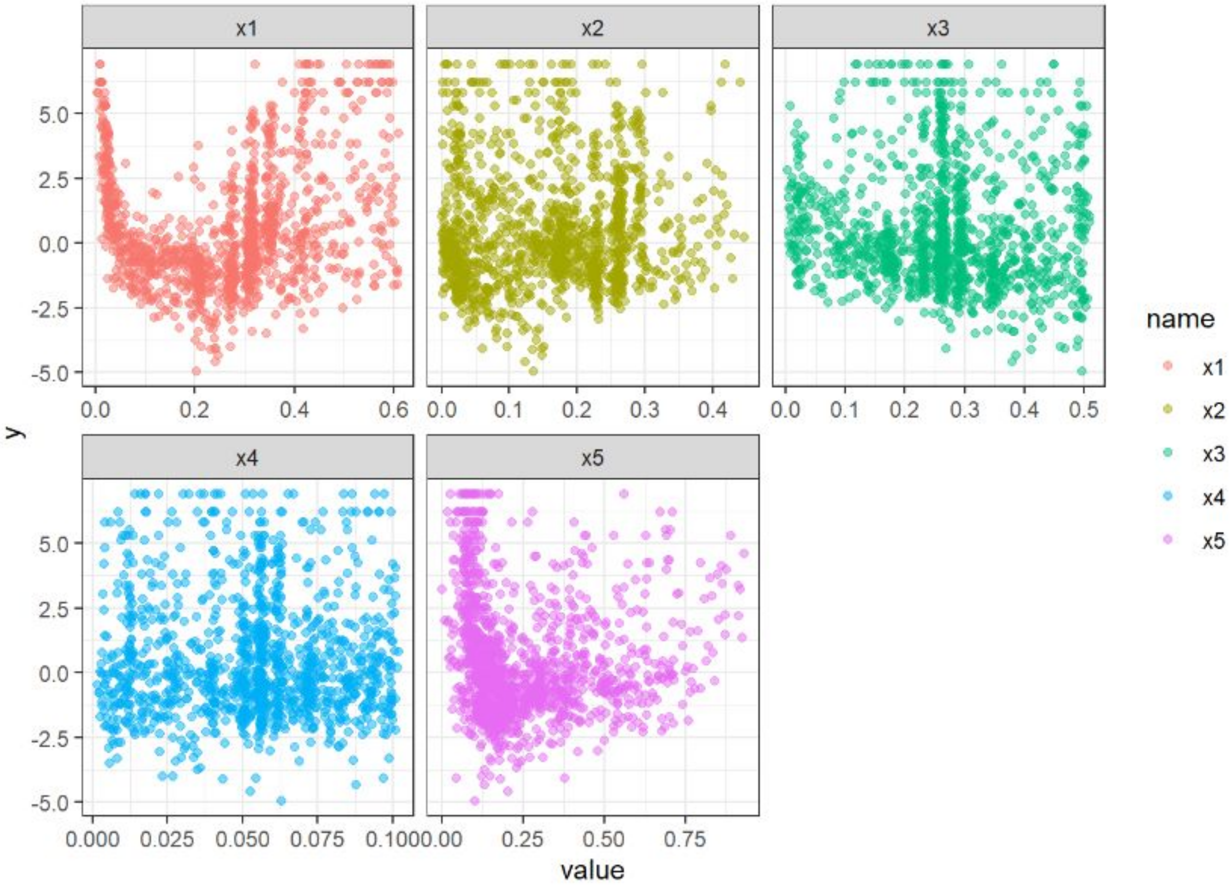
This is a graph of the highly corroded samples.

As you can see, there are minimal samples that are highly corroded when x2 is greater than .3.

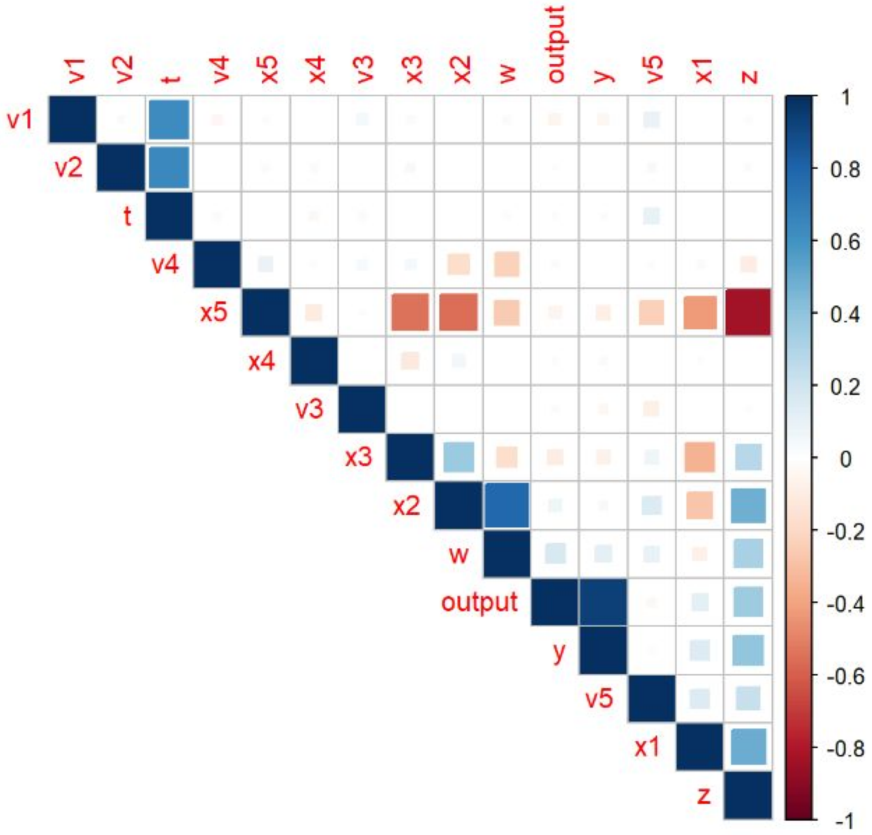


This is a graph of the highly corroded samples.

As you can see, there are minimal samples that are highly corroded when z is between 1 and 2.5.



Here is a graph of the ranges of the x inputs, as well as x5.

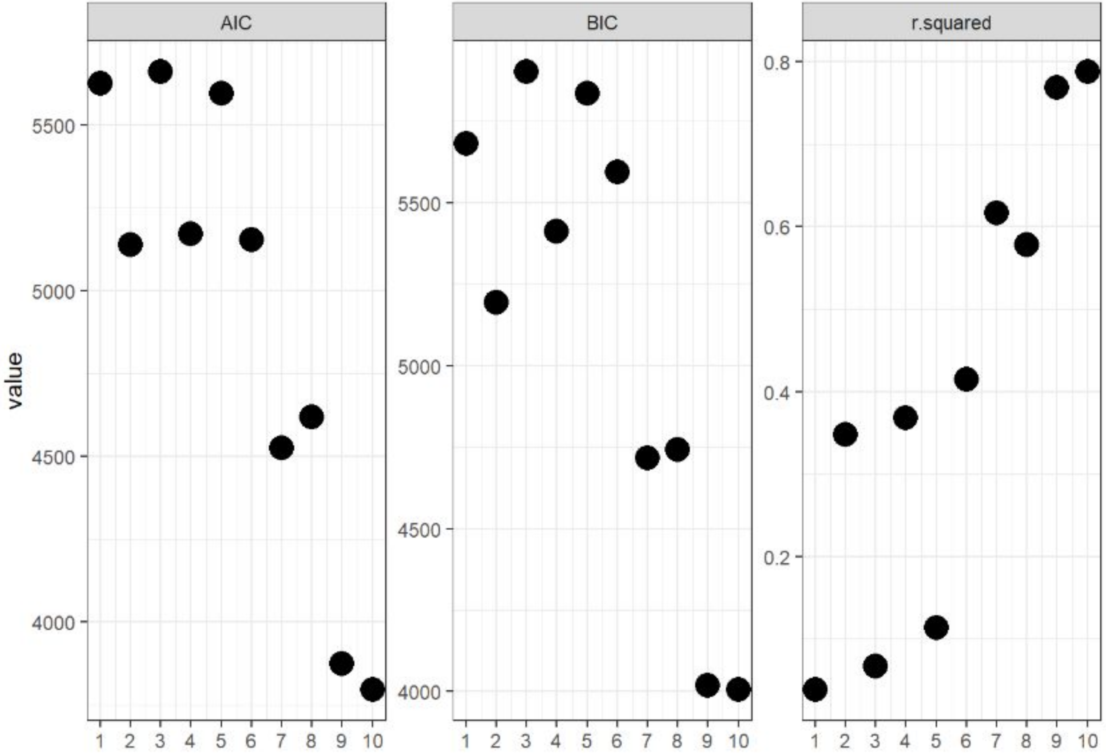


Here is a graph that shows the relationship between all of the variables.

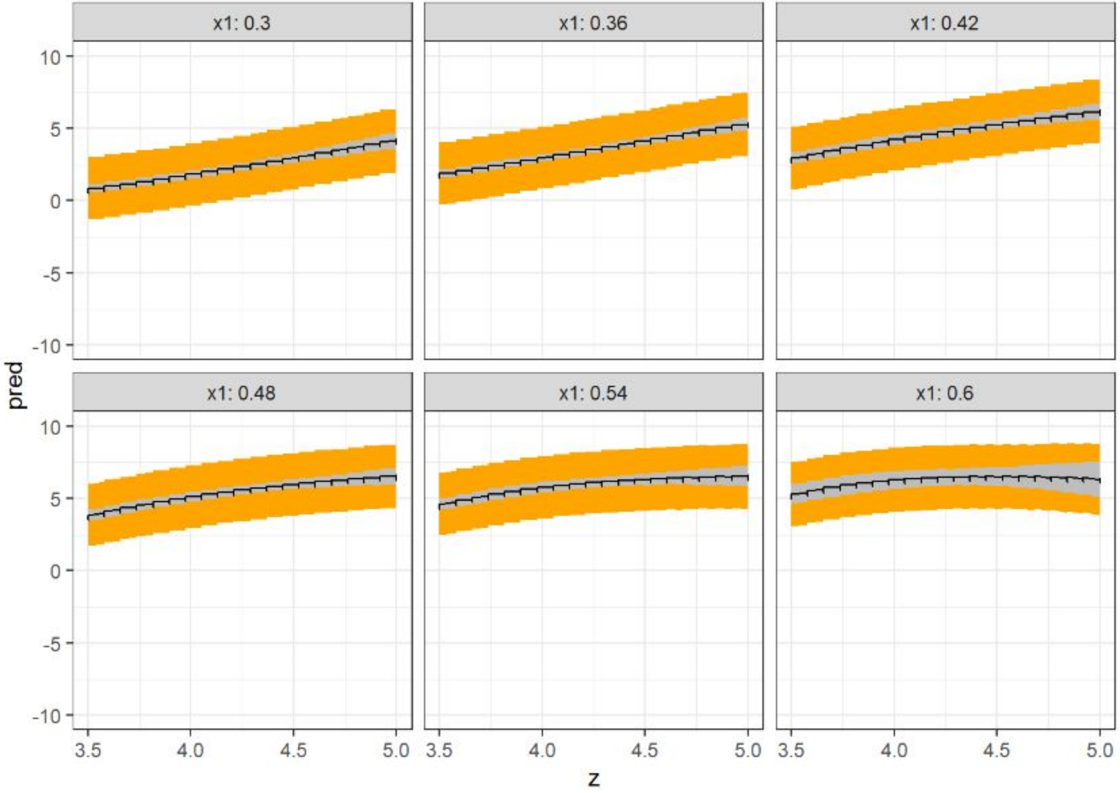
Takeaways from EDA

- X1 should be between .05-.25
- X2 should be greater than .3.
- Z should be between 1 and .25.
- Slight polynomial relationship in the x1 values.
- Based on these visualizations and takeaways, we will make models to test this data, to find the best ratio of inputs to make the most minimally corrosive surface coating.

Model Performances and Best Model



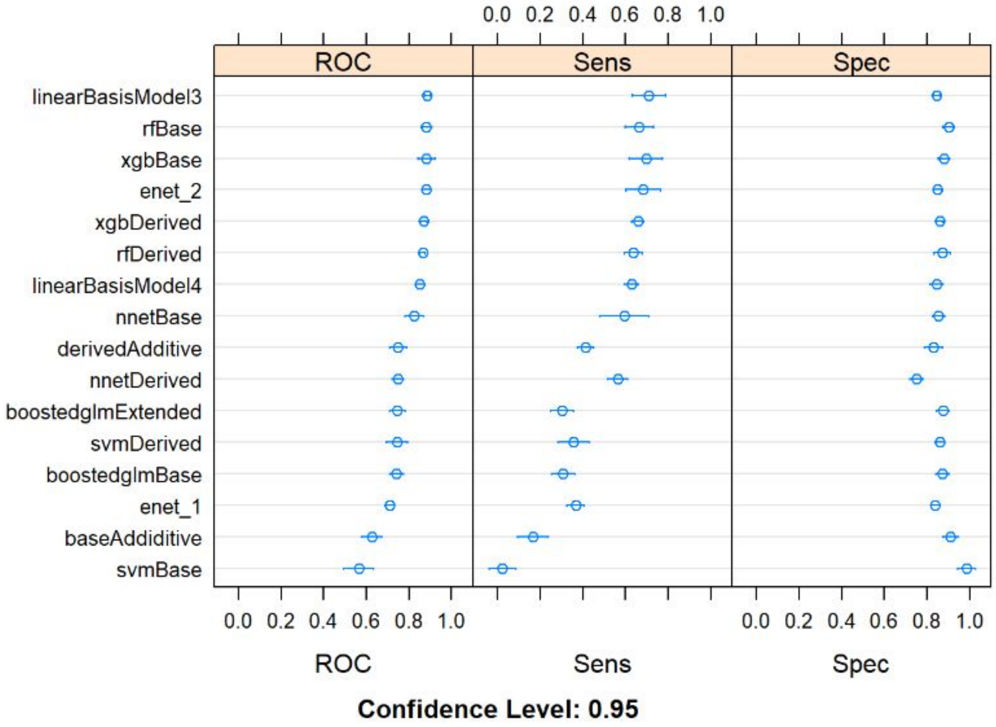
We fit many regression models, and tested then to find out which was the best based on three performance metrics; AIC, BIC, and R-Squared.



We made predictions on the better models, to find our confidence and error intervals.

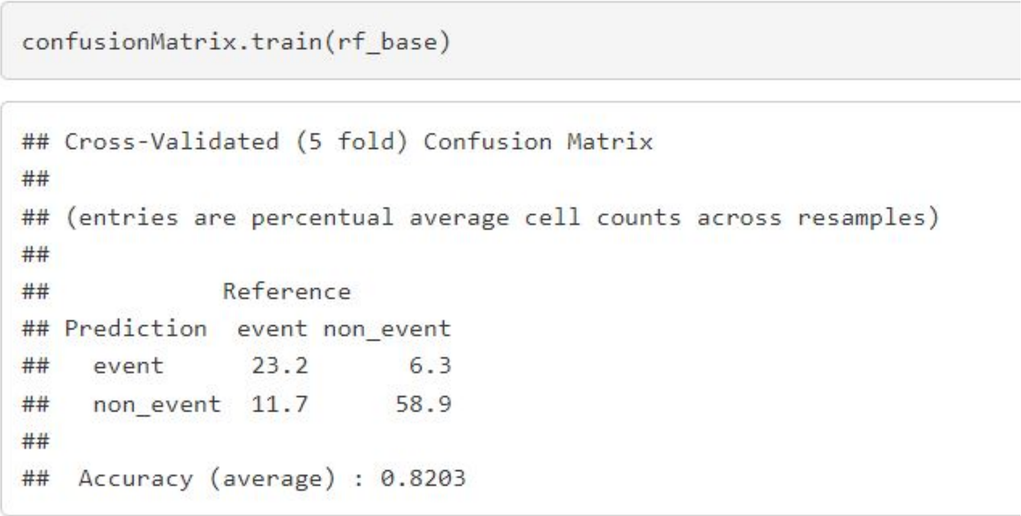
Best Model

- After we found our best models, we trained this model as well as other models based on RMSE. This in the end, helped us find that a neural network model was the best fit for regression



We also fit classification models and trained them based on the ROC Metric.

As you can see here, a linear basis model and a random forest model were the best fit for classification based on the ROC metric.



```
confusionMatrix.train(rf_base)
```

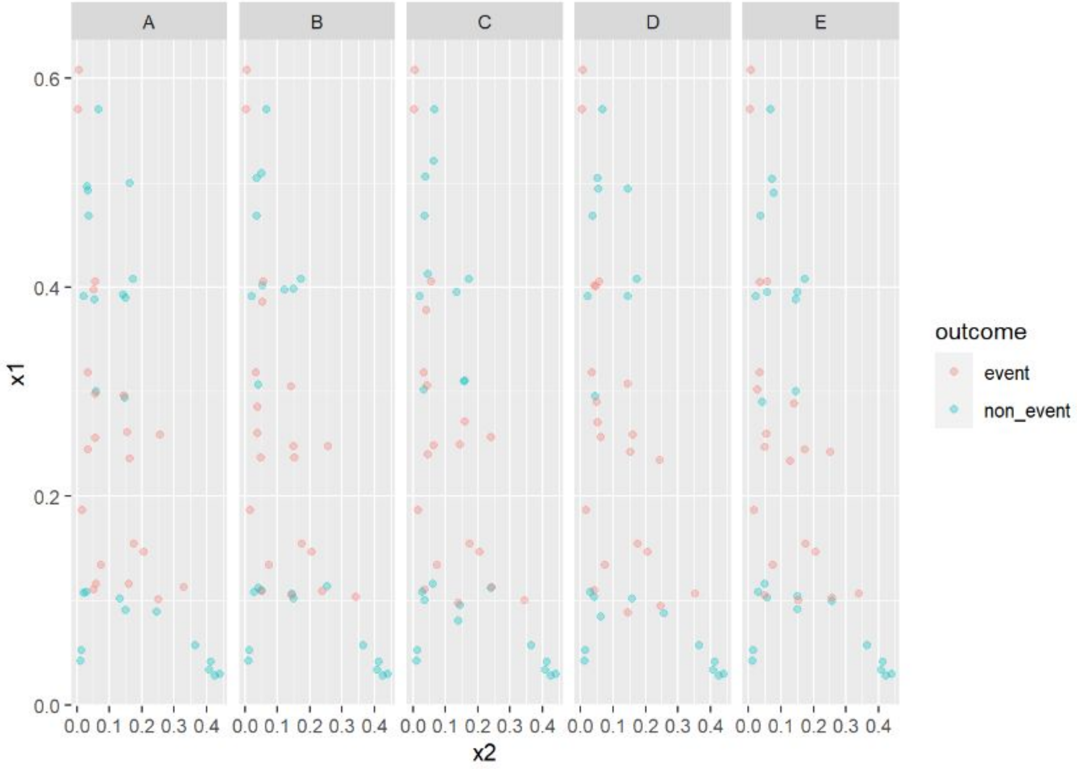
```
## Cross-Validated (5 fold) Confusion Matrix
##
## (entries are percentual average cell counts across resamples)
##
##            Reference
## Prediction  event non_event
##   event      23.2       6.3
##   non_event  11.7      58.9
##
##  Accuracy (average) : 0.8203
```

When we train these same models based on Accuracy, the random forest model comes out on top. This random forest model is based on the base features, not the derived features.

Variable Importances

We can find variable importances from our best models

- As I stated previously, we found best models for regression and classification models. The variables that these models had in common were x1,x2,x3, and z. This leads me to believe that these variables are important. We will also take the categorical variable m to be important as well because the models had this variable as well.

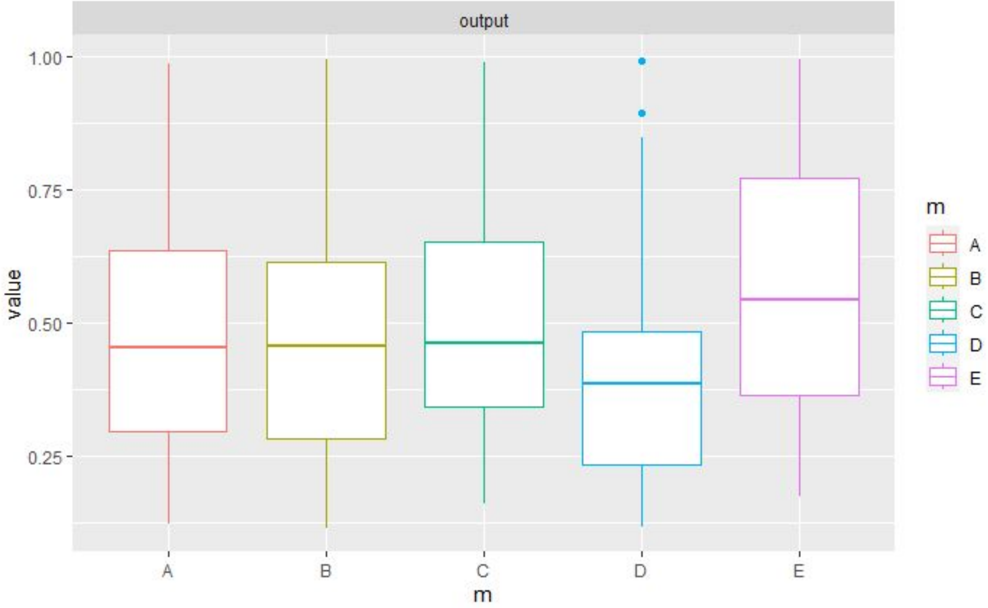


As you can see here, we test our best models on the test data set. When we graph the predictions we can clearly see where the events will be based on the two inputs x1 and x2.

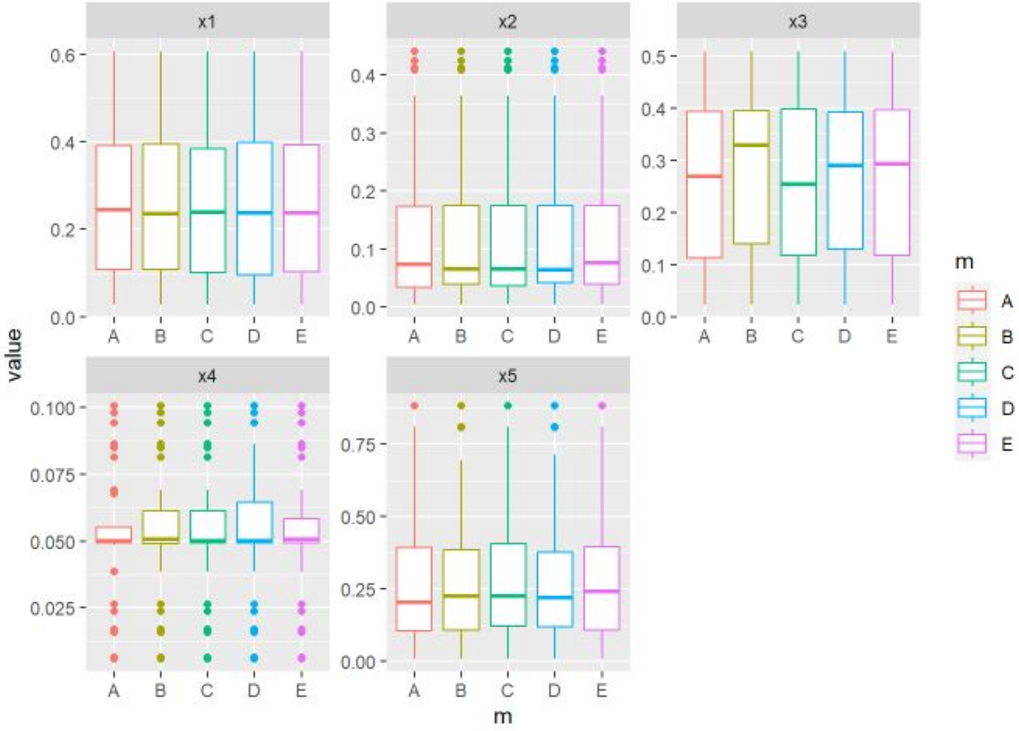
Last Slide Continued

- With the visualization on the last slide, we made similar visualizations.
- This gave us the conclusions that x1 should not be between .4-.6 when x2 is between 0-0.2, because there are not a lot of events in this range
- X3 should be between .3-.5 when z is between 0-.2 because there are a lot of events.
- For the same reason, x2 should be between .2-.25 when z is between 1-3, and x1 should be between .1-.3 when z is between 0-3, and x1 should be between .1-.3 when z is between 0-5.

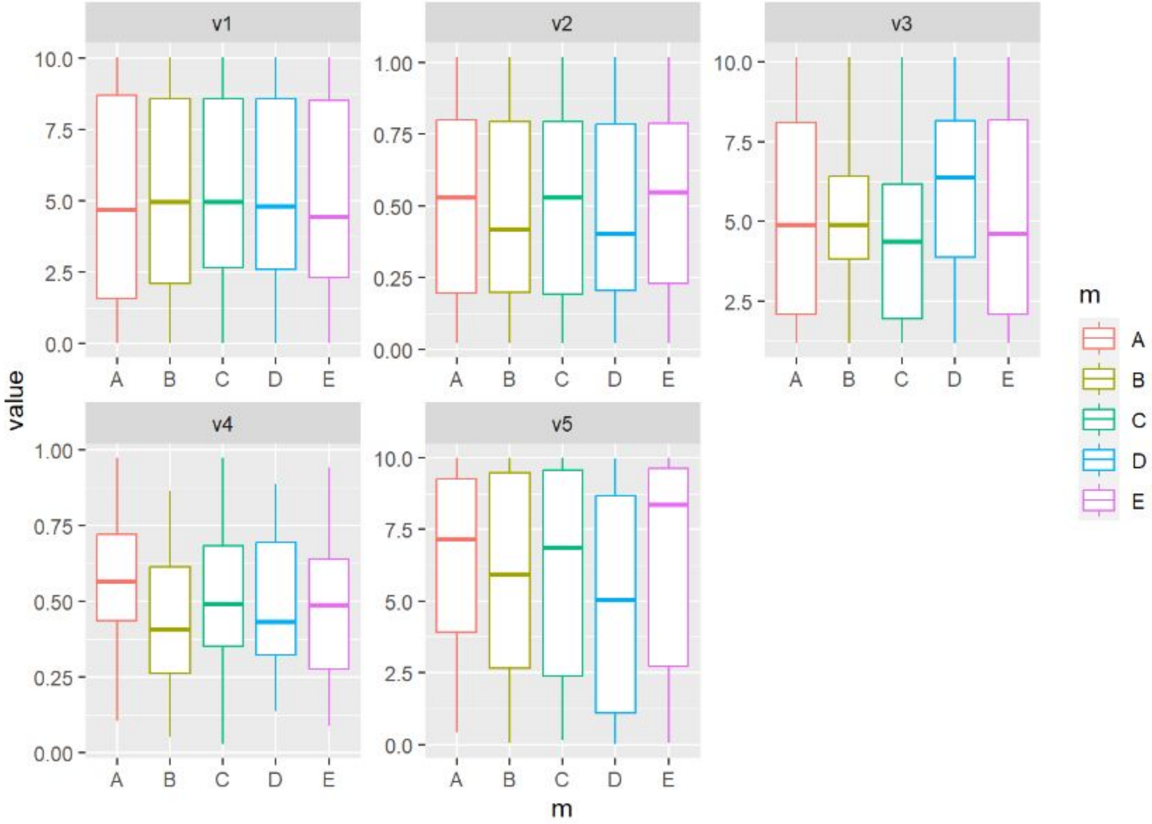
Model Predictions



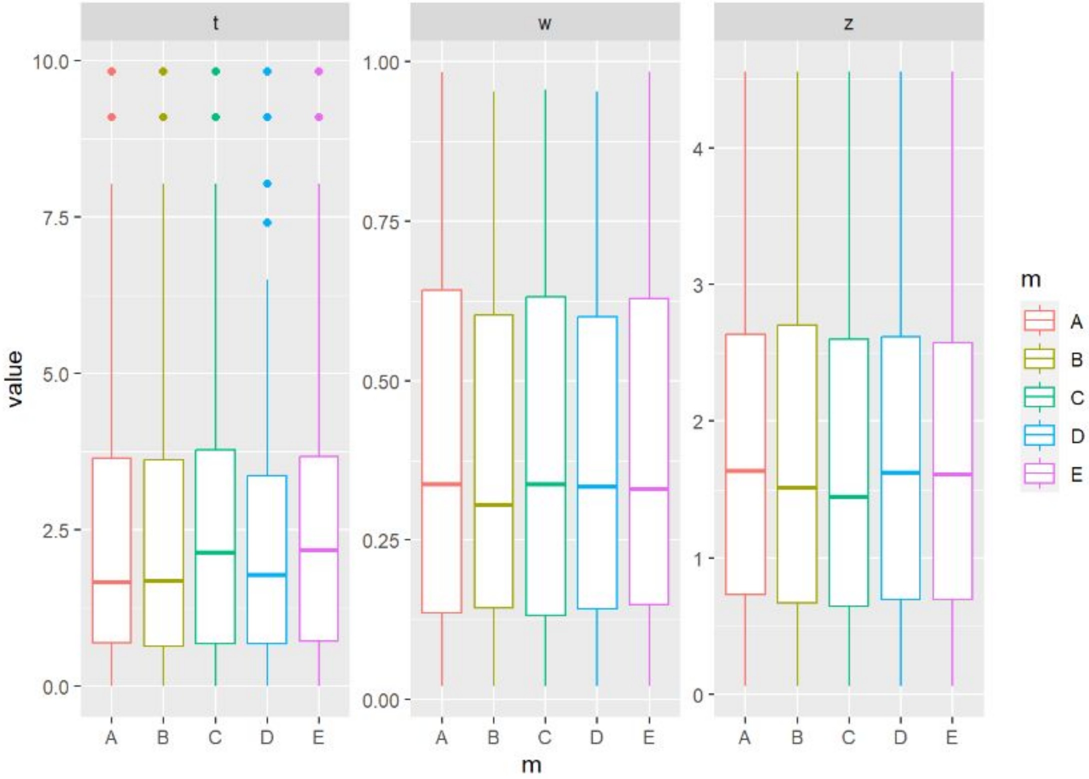
As you can see here, machine D produces the samples with the lowest output value, meaning that its samples will corrode the least.



Here is a graph of the predicted values of the x inputs that will minimize corrosion. We made similar visuals of the v inputs as well as the t,z, and w derived features to find the optimal values of those as well.



Here is a graph of the predicted values of the v inputs that will minimize corrosion.



Here is a graph of the predicted values of the derived inputs that will minimize corrosion.

Input Settings that Minimize Corrosion

Conclusion

- Based on our classification and regression models we can make some conclusions about the input settings that will minimize corrosion.
- Machine D is predicted to produce the samples with the lowest output value, meaning its samples will corrode the least.
- Optimal settings: x1=.25, x2=.07, x3=.25, x4=.05, x5=.2, v1=5, v2=.5, v3=5, v4=.5, v5=7, t=2, w=.35, z=1.6.