Final Report

Samuel Burge, Chad Austgen, Skylar Liu

April 19, 2022

Supervised Learning Task

For the supervised learning task we first attempted to approach the problem with as many probable solutions as possible. This included:

- Regularized Logistic Regresion (Ridge, Lasso and Elastic Net)
- Boosted Trees
- Naive Bayes
- Random Forest
- Support Vector Machines with Radial and Polynomial Kernals

We initially avoided some techniques such as support vector machines as our set of predictive variables was relatively large and this can cause the SVM method to run slowly. In this case however we did not find it so slow as to be prohibitive and so we pushed forward with it as a methodology.

K-Nearest neighbors was rejected outright as its limitations for high-dimensional data make it very unsuitable for this task.

Methodology

Logistic Regression: We importantly verified our assumptions which validates logisitic regression as a methodology. Namely, before the analysis we created large plot matrices and visually verified across all predictors that X~Y were not well separated.

```
##       Training Error Cross Validation Error
## Lasso      0.3201453              0.3509644
## Net        0.3201453              0.3862684
## Ridge      0.2016434              0.4178184
```

The results of our analysis are depicted in the table above. Of note is the degree to which Ridge under performs despite low training error. This is likely over fitting and we attribute this to high excess variance in the data set compared to bias. The Lasso and Net eliminate this variance by performing variable selection and eliminating many predictors.

Boosted Trees:

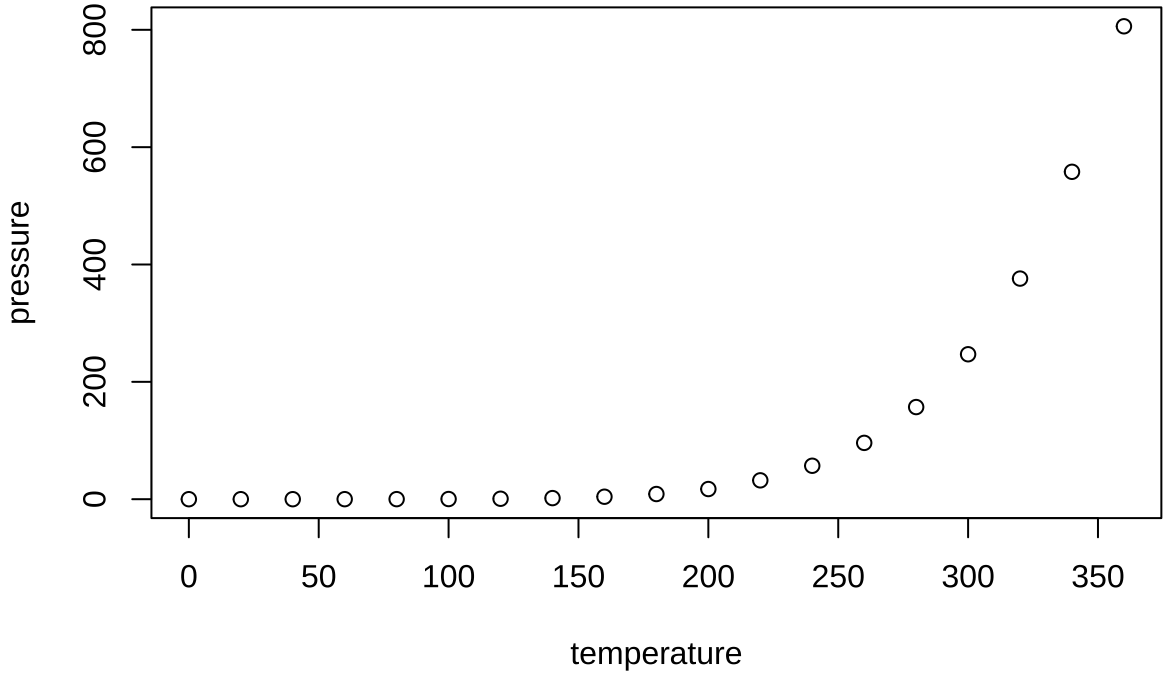
Boosting provided the best results for us of any procedure. In order to generate our results we used a cross validation method which iterated over multiple values for shrinkage rate, bag inclusion, and interaction depth. We then compared the models with a withheld validation set. We found that as is common with validation set approaches there was significant variance in the accuracy of the model with respect to the validation set. In order to identify higher performing models we resampled repeatedly and fed the data into the cross validation trees algorithm and measured against a corresponding validation set. We chose our model manually using a slight bias toward models which were simpler and had less variance.

Results

We can put a description of the results, like a table of the average training and cross-validation errors, in this section. I was also thinking we could include ROC curves.

Unsupervised Learning Task

Text here.



References