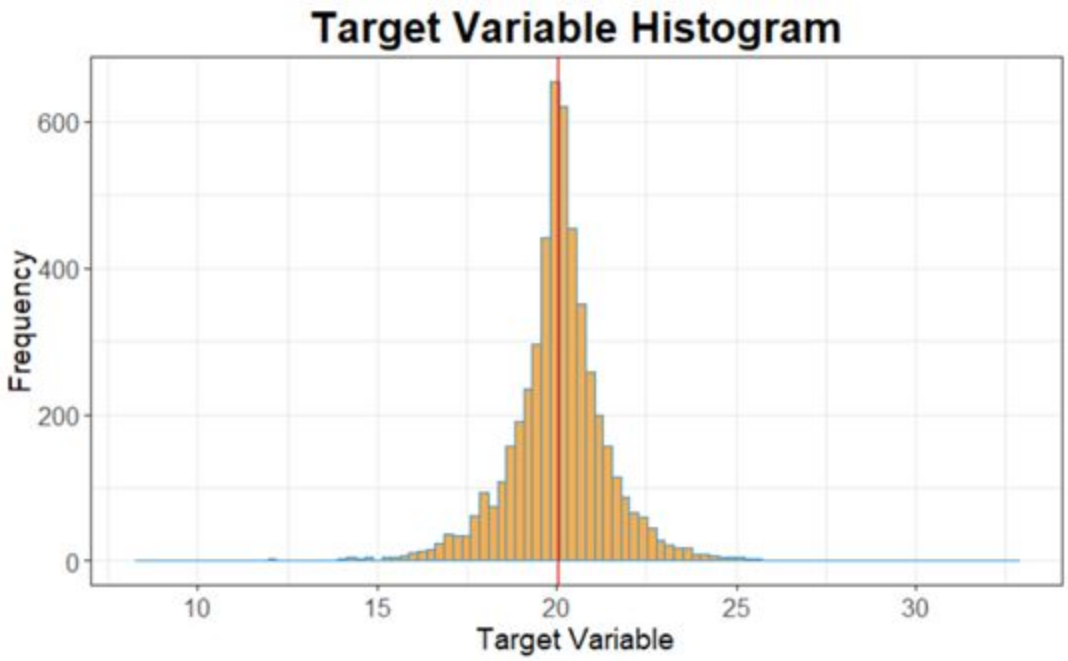
Blue Team 4: Angelica Spratley, Amy Hand Tolan Nguyen, Abhinav Pandey, Tommy Lang

Data Exploration

The data available for training has 6,350 observations, 86 variables and no missing values. There are 59 numeric variables (not including target), 24 logical variables and 2 categorical variables with 5 and 12 levels each. All the logical variables have at least 95.5% of their values as false.

The data was split into a train and test set with 5,080 and 1,270 observations each respectively (80-20 split). The plot below shows the distribution of the target variable for the train set. It is tightly centered around the middle with a mean of 20.04.



Ridge and Boruta algorithms were used to get a sense of which variables were good predictors. A lasso model was created first but all the coefficients for the predictors were zero or nearly zero. The Ridge regression model had an intercept of 20.036 which is very close to the target mean. All the other coefficients were small, and it suggested there is very little linear relationships. The Boruta algorithm suggested that 59 of the variables were more relevant as predictors than just random probes (shadow features).

Model Stacking

The top 9 most important numerical variables as suggested by the Ridge regression model and Boruta algorithm were selected and 45 new variables were created from them. We created 36 interaction terms and 9 squared terms from the 9 original variables to give linear algorithms a

chance. We also created a new logical variable that had a value of 1 if any of the top 5 logical variables had a value of True and 0 otherwise. In addition, we created another variable that contained the row sum of the logical variables. The new variables were added to original train set and the Boruta algorithm was used again for variable importance. The top 40 variables as suggested by the Boruta algorithm plus 3 that we thought would be useful were selected for training models.

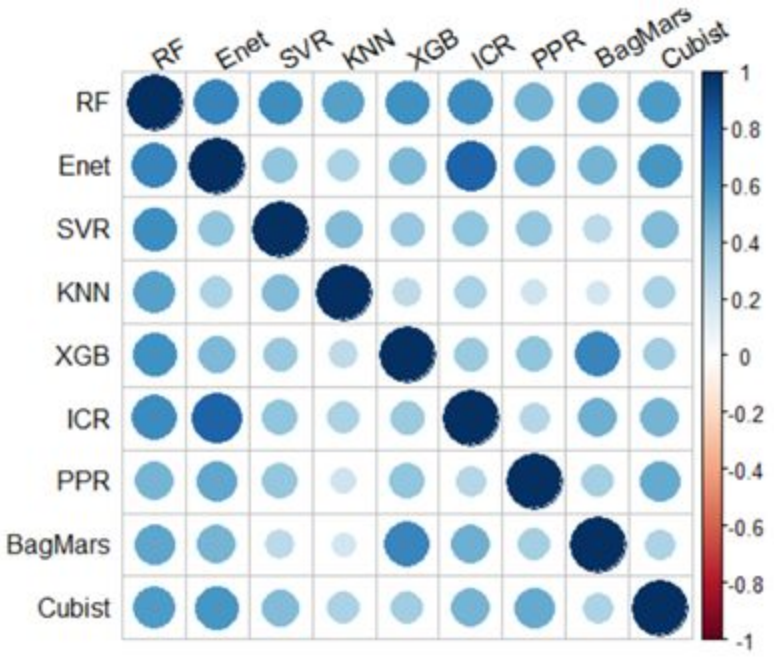
We created 9 models and a stacked model with the selected predictors. All the models were tuned with 5 fold cross validation and grid searches / “grad student” descent optimization. The same 5 folds for cross validation were used for all 9 models. The table below summarizes the 9 models created and the null / baseline model of just using the mean as the prediction.

Model	Hyperparameters Tuned	CV MAE
RandomForest (ranger)	Mtry, min node size, split rule	0.953
Elastic Net Regression	Alpha, lambda	0.954
Support Vector Regression (RBF)	Sigma, C	0.949
K – Nearest Neighbors	K	0.956
XGboost (tree base)	Nround, max_depth, eta, gamma, colsample_bytree, min_child_weight, subsample	0.949
Independent Component Regression	n.comp	0.953
Projection Pursuit Regression	nterms	0.959
Bagged MARS	Degree, nprune	0.952
Cubist	Committees, neighbors	0.961
Null (mean)		0.953

The SVR and xgboost models performed the best but they are all pretty poor considering the null model. The models have about the same cross validated MAE as the null model.

We used 8 out of the 9 models’ out of sample predictions from 5 fold cross validation as meta features for a level 2 ensemble learner. The predictions from the Random Forest model wasn’t

used because they were correlated with the other predictions. The figure below illustrates the correlation between the out of sample predictions.



A stacked model was created using xgboost and the MAE was 0.955 from cross validation. It's not very surprising that the stacked model is worse than some of the individual models because they all lack skill.

Support Vector Regression

A second support vector regression model was tuned using cross validation and 10 predictors. It used numerical variables 1, 3, 4, 5, 6, 18, 23, 32, 58 and a binary variable that had a value of 1 if any of the categorical variables 6, 9, 11, 19, 16, 21 had a value of true. The variables used were suggested to be the most important by the ridge regression model and Boruta algorithm. Also, the categorical variables used to derive the new binary variable all had negative coefficients in the Ridge regression model. The model has a cross validated MAE of 0.949. It appears to have similar skill to the models created for stacking.

Conclusion

The 3 candidate models we considered for submission are the xgboost and SVR model created for model stacking and the second SVR model. We choose the second SVR because it had the lowest MAE on the test set of 0.973. We combined the train and test set and refitted the SVR model with the hyperparameters found from cross validation and made predictions on the 77 observations for submission.