L06 Ensemble Models

Data Science III (STAT 301-3)

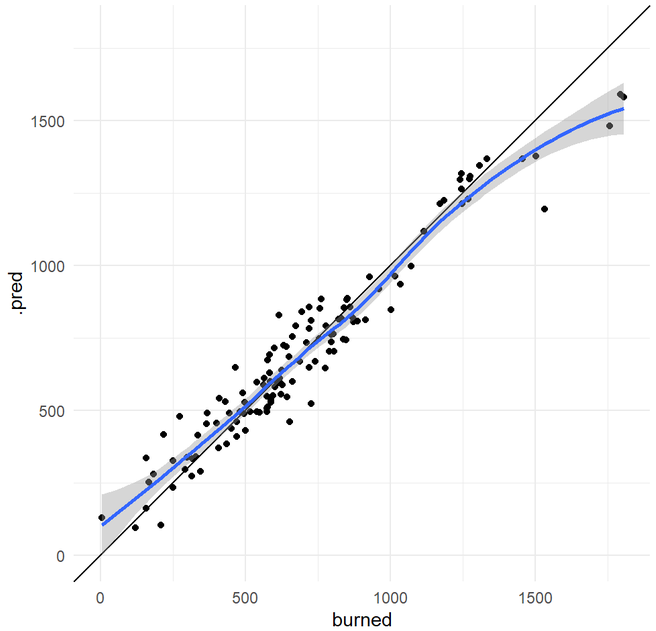
Austin Shinn

The stacked ensemble model considered 41 candidate models, with 6 retained in the final model. The models used were KNN, radial kernel support vector machines, and linear regression. The highest weighted type was a linear regression model, with a weight of .68154, and the 2nd highest weight was a support vector machine with a weight of .2847. These 2 make up the majority of the weight, with the next 4 being weighted .03 or less (SVM and KNN models).

```
## # A tibble: 6 x 3
##   member          type              weight
##   <chr>           <chr>              <dbl>
## 1 lin_reg_res_1_1 linear_reg       0.682
## 2 svm_res_1_19    svm_rbf          0.285
## 3 svm_res_1_23    svm_rbf          0.0369
## 4 knn_res_1_10    nearest_neighbor 0.0348
## 5 knn_res_1_11    nearest_neighbor 0.0119
## 6 knn_res_1_09    nearest_neighbor 0.00398
```

```
## # A tibble: 2 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard      93.3
## 2 rsq     standard       0.934
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



The ensemble model had an RMSE value of 93.297, and an R^2 of .934. The fit is generally quite accurate, but falls off from a simple linear fit near the higher values of burned.

```
## # A tibble: 8 x 4
##   .metric .estimator .estimate member
##   <chr>   <chr>          <dbl> <chr>
## 1 rmse    standard         0   burned
## 2 rmse    standard        93.3 .pred
## 3 rmse    standard       157.  knn_res_1_09
## 4 rmse    standard       158.  knn_res_1_10
## 5 rmse    standard       159.  knn_res_1_11
## 6 rmse    standard       360.  svm_res_1_23
## 7 rmse    standard        97.9 svm_res_1_19
## 8 rmse    standard        92.3 lin_reg_res_1_1
```

Looking at the final RMSE values for the models, the linear regression model actually had the lowest RMSE, with an RMSE value of 92.258 compared the 93.297 for the ensemble model. The next closest model was a support vector machine model with an RMSE of 97.882. After that the RMSEs increase drastically. So it appears that a simple linear regression model is more effective and should also save more time computationally.

Github Repo Link

https://github.com/STAT301III/L06-ensemble-models-shinnsplints