Gaby Lio
Reed Dalton
Reece Wooten
Ashwin Param
Bhavana Vijay

INTRO TO PREDICTIVE MODELING: GROUP 8

DATA ANALYSIS

To build a predictive model to price cars based on available features, we first needed to understand the data set before we could make predictions. Our first objective once we had access to the *Cars* dataset was to clean and visualize the data. After viewing scatter plots of the response variable *Price* and the predictor variables, we noticed there was limited distinct variability between the *state* and *region* factors based off of comparisons of the *State* vs. *Price* scatter plot and the *Region* vs. *Price* scatter plot. Therefore, we decided to drop the *State* variable to minimize the amount of variability error.

Upon further analysis, we decided to remove the variable *Sub Trim* from our predictive model, as it did not capture significantly more information than what was captured in the associate variable *Trim.* The variables *Mileage*, *Condition*, *Displacement*, *Trim*, and *Year* seem to have the most distinct information from the data visualization, which was later corroborated by the output of our models.

LINEAR MODELS

Least Squares

Although KNN is a simpler model, the presence of multiple categorical variables moved us to start our exploration with linear models. Initially we tried a regression model using backward stepwise selection, resulting in a RMSE of $62,889.88, which was almost as much as the mean of the price data ($67,001). We calculated RMSE on an out of sample test set, and created the models on a training set, splitting the data 30/70 respectively. After studying the residual plot, we observed a large variance in the residuals, leading us to apply the logarithmic function to the predictor variables. While the variance of the residuals decreased when we fed in the new set of predictors, the RMSE ended up increasing by $1,778.85. Since we saw no obvious interactions between the predictor variables, we felt that adding interaction terms to our linear model would not significantly decrease our RMSE, and we instead chose to allocate our efforts to exploring other methods.

LASSO/Ridge

Hoping to decrease the predictive errors found in the least squares model, we utilized ridge regression and LASSO selection to tease out the car features most correlated with price in the context of our linear model. Unfortunately, LASSO only brought 8 of our 69 variable coefficients to zero, and therefore did not provide further significant insights.

RANDOM FOREST, BAGGING, BOOSTING

Next we explored fitting a random forest, bagging, and boosting model to our data set to see if we could yield a lower RMSE than the RMSE we obtained with our linear models. Since bagging, boosting, and random forest all use samples from the population to build their regression trees, the computational cost of also applying k-fold cross validation did not outweigh the marginal improvements of the model. Instead we decided to split our data into a training and test set, and compute the RMSE on an out-of-sample set.

In order to control for overfitting, we adjusted the parameters in each model. For random forest, we controlled the number of *n* trees, the number of *m* variables randomly selected, and the node size. For bagging, we controlled the number of *n* trees and the node size. For boosting, we tuned the the number of *n* trees, the shrinkage parameter, and the interaction depth. This allowed us to reduce the variance error.

To evaluate our tree-based models, we plotted the prices we predicted versus the actual prices of the cars. We then used the variable importance graphs to manipulate the parameters in our model.

We found that using more than 200 trees had a negligible impact on the RMSE. Despite some variables seeming insignificant (i.e. *Color*, *Wheel Size*, *Featurecount*, etc.), removing these variables actually led to a higher RMSE. The variables *Mileage*, *Year*, *Trim*, *Condition*, and *Displacement* seemed to have the largest effects on *Price*. Even in our linear models, these variables were significant at a 0.05 level, and in our tree models, these variables were the dominant nodes when making the splits in the trees.

After repeating the process of building the model, adjusting the parameters, and calculating the RMSE for multiple models, we were satisfied with our random forest model as it yielded a low RMSE and controlled for overfitting. Random forest continually out performed bagging and boosting, with our final model consisting of 200 trees, 20 nodes, and 6 total variables resulting in an RMSE of $6,688.96.

INSIGHTS

As a whole, our tree models outperformed our linear models. Compared to the linear models, using random forest and bagging models reduced the RMSE by roughly $4,500. Since our data had many categorical variables, each composed of multiple factor levels, it was hard to find a linear relationship between the variables and the price. Due to this challenge, trees were a better choice. Because trees are flexible fitters, they intrinsically capture non-linearity and interactions between variables in ways that are not native to linear models.

We chose to evaluate each method by their RMSE as RMSE and price are both measured in dollars and the RMSE can tell us on average the magnitude of our model's mistakes. Since our RMSE is relatively small compared to car prices, we can be confident that our predictions will be accurate, give or take around $6,700.

LIMITATIONS

While we were able to develop and tune a model to predict car prices based on available features, it is important to acknowledge that we were forced to relax some criteria for proper model building along the way.

From the plot of *Predicted Prices* vs. *Residuals* for our random forest model, the residuals appear to fall in a predictable pattern and do not appear to be randomly distributed. The plot also shows that our model overpredicts the price of cars that are valued over $120,000. More expensive cars seem to have higher variance in their residuals, leading us to believe that there is some element of heteroskedasticity in our analysis. Despite the fact that tree models do not require standardization of variables, given more time, we would have liked to tried and find a way to account for these errors in our model.