Don’t Get Kicked

Tyler Smith

Problem Background

- ”Kicked cars” refer to cars purchased by auto dealerships at auto auctions that turn out to be bad purchases for various reasons.
- Searching for a new car is stressful enough; customers shouldn’t be worried about getting scammed into buying a piece of junk!
- Lucky for us, machine learning is making it easier for auto dealerships to keep their inventory up to their customers’ standards by identifying kicked cars.

Feature Engineering

- Response variable = IsBadBuy (0 or 1)
- My recipe has many steps that cater to specific complexities of the Don't Get Kicked data:
- **step_novel** and **step_unknown** are used to address novel factor levels in nominal predictors that might appear in new data but are not present in the training set. We then assign any unknown factor levels to their own group using step_unknown.
- **step_lencode_mixed** is where we add target encoding to the mix. Target encoding is when we convert a categorical variable to the mean of our response variable, which is IsBadBuy in our case.
- **step_impute_mean** helps us address missing values even further. This is going to assign the mean of respective columns if missing values are present.
- **step_corr** addresses the potential issue of multicollinearity. If features are highly correlated with each other, this step will remove one of them.
- **step_zv** removes any variables in the dataset that have zero variance.
- **step_normalize** normalizes numerical predictors. The normalized data will have a mean of 0 and standard deviation of 1.

Model Comparison

- BART = .237
- Penalized Logistic Regression = .234
- Boosting = .223
- Naive Bayes = .222
- KNN = .178

Bayesian Additive Regression Models (BART)

- Large number of regression trees are fitted to the data and each tree explains a small portion of the response variable.
- **Bayesian:** Use probabilities to express uncertainty in model parameters.
- **Additive:** Each tree of the BART model will explain a small portion of the overall prediction. We then combine what we find out from each tree to make our final prediction.
- **Regression Trees:** Data is split into subsets. Each leaf node makes a class prediction based off the class majority since the outcome is binary.

Final Score

BART gave me the highest score out of all the models.

Final score = .237