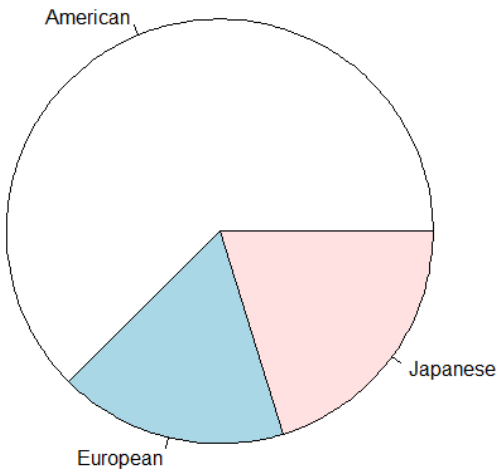
Evaluation of Classification Models in R
Eddie Figone

Introduction:
This analysis seeks to compare the performance of a random forest, boost, tree, LDA, and KNN model in classifying the country of origin for cars in the Auto dataset. This data was taken from the Statlib Library at Carnegie Mellon University. This cleaned dataset has 392 rows and 7 independent variables. In this analysis, origin is our response variable, and our predictor variables are mpg, displacement, weight, cylinders, horsepower, acceleration, and year. Displacement represents the volume of the cylinders in the engine. The models will be trained and tuned using the training data set. The models will then be compared using 500 iterations of Monte Carlo cross validation on the training and testing sets.

Exploratory Data Analysis:
Origin was chosen to be our response variable in this analysis. Origin represents the country of origin, with 1=American, 2=European, and 3=Japanese. Looking at our response variable, the majority of cars are of American origin (245) with the remaining fraction split between European (68) and Japanese (73) origin. This is significant, and the unevenness of our response will factor into our choice to use “information” for our splitting criteria for our tree model.

Breakdown of origin (response variable)



Next we calculate the covariance of each independent variable with our response. We see changes in displacement as the leading cause of change in origin and we will use that as part of our models.

Correlation of independent variables with origin

Variable	Covariance w/ origin
displacement	**-0.61**
weight	-0.59
cylinders	-0.57
mpg	0.57
horsepower	-0.46
year	0.18
acceleration	0.21

Before selecting other independent variables I mapped out the top 20 multicollinearity between independent variables, as seen below. All of the most predictive variables are correlated with displacement, so we opt not to use weight, cylinders, mpg, or horsepower. While acceleration and year do not have the highest correlation with our response variable, they have a tolerable level of multicollinearity with displacement so we opt to include them in our models.

Multicollinearity of independent variables (top 20)

Var1	Var2	Correlation
displacement	cylinders	0.95
weight	displacement	0.93
weight	cylinders	0.90
horsepower	displacement	0.90
weight	horsepower	0.86
horsepower	cylinders	0.84
weight	mpg	-0.83
displacement	mpg	-0.81
horsepower	mpg	-0.78
cylinders	mpg	-0.78
acceleration	horsepower	-0.69
year	mpg	0.58

acceleration	displacement	-0.54
acceleration	cylinders	-0.50
acceleration	mpg	0.42
acceleration	weight	-0.42
year	horsepower	-0.42
year	displacement	-0.37
year	cylinders	-0.35
year	weight	-0.31

Method:
In this section I will provide a brief explanation of each model, their optimal values chosen, and the comparison that was run with all the models. For each model we used the independent variables of displacement, acceleration, and year.

- **Random Forest:** This model is an ensemble method of a decision tree applying bagging. In addition to bagging, this approach uses random feature selection to avoid a dominant feature skewing the results. In my model I found the optimal number of trees to be 173 by testing values from 100 to 550. Looking at the feature importance of all features, I found displacement to be the most important, which is in line with our earlier findings.
- **Boosting (gbm):** This is another ensemble approach that combines the predictions of multiple weak learners to make an accurate prediction. Models iterate over the data with residuals being weighted, in this way the errors of a previous model are corrected by the future iterations. I found the optimal interaction depth to be 2 by testing values from 1 to 10 and shrinkage to be .04 by testing .01 to .1.
- **Decision Tree:** This model seeks to split the data to reduce entropy and increase purity of the dataset. Decision trees are a greedy algorithm that looks for the largest improvement per split. Trees are also trimmed back via pruning to avoid overfitting. I used the information splitting criteria rather than Gini as it is better suited to unbalanced responses.
- **LDA:** This classification model seeks a combination of features that split maximizes separation between classes and reduces dimensionality. Uses Fisher's criterion to find direction that optimally separates classes.
- **KNN:** Our last model uses k points near the new point based on distance to predict the class of the new point. We found the optimal k value to be 2, however 38 was very close, and less likely to be skewed by outliers, so we opted to use that.

Once our models were built, they were run on 500 iterations of Monte Carlo cross validation to measure their testing error. The results were averaged and their variance was also taken. The results are below.

Results:

Model	Test Error	Variance
RF	0.18	0.0010
Boost	0.21	0.0012
Tree	0.23	0.0016
KNN	0.28	0.0014
LDA	0.30	0.0011

Findings:
Based on our 500 iterations of Monte Carlo cross validation. We found the Random Forest model to perform the best in terms of test error as well as variance. Not only does this model make the most accurate predictions, but also does so most consistently. Our other ensemble method was the next best performance in terms of testing error and third in terms of variance. It seems that ensemble methods in general are more reliable than single models.

One thing to consider is the ability to fit the data so well in ensemble methods will likely lead to overfitting. In scenarios where less is known about the data or there is more variance in samples an individual model may be a better approach than an ensemble.