Assignment 6 - Classification

Lacey Hartigan

Due 3/30/21

This problem set uses a kaggle.com dataset called: training_car.csv. Go to kaggle.com, search in the data section for "caravana" and download the training_car.csv dataset.

1. Calculate the proportion of lemons in the training dataset using the IsBadBuy variable.
2. Calculate the proportion of lemons by Make.
3. Now, predict the probability of being a lemon using a linear model (`lm(y~x)`, with covariates of your choosing from the training dataset (must have at least two covariates).
4. Make predictions from the linear model (in other words, add the predictions from your model to your dataset). Use head() to display the first 5-10 rows of data so I can see your variables.
5. Now, predict the probability of being a lemon using a logistic regression.
6. Make predictions from the logit model. Make sure these are probabilities. Use head() to display the first 5-10 rows of data so I can see your variables.
7. Create a confusion matrix from your linear model and one from your logit model (and compare the two).
8. Plot the probability of a car being a bad buy by car make.
9. Create a table that shows the probability of a car being a bad buy by make.

Optional Bonus: Create a heatmap of the probability of a car being a bad buy by make and size.