Problem Set 6

Prof. Conlon

Due: 5/5/20

Packages to Install

The packages used this week are

- ggplot2
- xtable (build tables quickly)
- data.table (data tables are computationally efficient and IMHO easier to work with)

```
ERROR: DATA DIRECTORY NOT CORRECT
data_dir variable set to:  /Users/christopherconlon/Documents/applied_metrics/HW~+~6-Adv~+~ML/
```

Problem 1 (Coding Exercise: Next Year)

Discrete Choice Question Here

Problem 2 (Coding Exercise)

This exercise will walk you through a prediction task. I have downloaded data from a peer-to-peer lending platform, Lending Club. The dataset you will work with is: **lending_club_07_to_11_cleaned.csv**. Lending Club provides detailed characteristic information regarding loans, both information on the borrower, as well as, the loan itself. Your goal will be to build a model to predict the outcome of a loan, i.e. whether an individual paid off a loan or did not pay off a loan. In our case, a good outcome is if the loan is fully paid off, a bad outcome is if the loan is charged off.

The target variable for the analysis is **loan_status**, where:

$$loan\_status = \begin{cases} 1 \text{ if loan is paid off} \\ 0 \text{ if loan is not paid off} \end{cases}$$

a. This is going to be a more DIY style exercise, provide a list of the variables you plan to use for the analysis. Give a short discussion for why you excluded other variables.

b. Regularization is an important step when using an machine learning algorithm, regularzie the variables that you have included. Briefly, why is regularization important?

c. Provide a simple correlational table to give you a sense of the relationship between your covariates. Do you notice any interesting patterns?

d. Split the dataset into a single test and training set, a simple rule of thumb is an 40/60 split. How did you build these two sets?

e. Using your training set, run a logistic regression, a random forest, and a gradient boosted random forest. To show your results, present both a measure of miscalssification error, accuracy and a confusion matrix.

f. One easy way to improve model performance is cross-validation. Do a k-fold cross validation, where k=5, using the best performing model from part (e.). Re-report the misclassification error, accuracy and a confusion matrix.