FINAL ASSIGNMENT

INTRODUCTION

This is the final course assignment. This assignment will be worth 50% of your 'final exam' grade.

The big goal is to use the provided dataset on health insurance charges to create a model that predicts charges as accurately as possible, based on the patient traits of age, sex, bmi, children, smoker, and region. As you generate this model, you should perform and document initial data quality checks, exploratory data analysis, and all of the models you try to fit.

You must submit a final report as a word document, as well as the code you wrote to generate the results.

QUESTIONS/REPORT COMPONENTS

General components:

- Brief data summary for EDA summarizing the model input variables
- Univariate summary of the model output (cost)
- Pick a loss function and explain why you think it is a reasonable choice.
- Implement a cross validation scheme. Explain how you did this in your report.
- For the machine learning sections, implement at least the following models:
  - A few different models using H2O (random forest, gbm, regularized regression, Auto-ML)
  - At least 2 different architectures of neural networks using keras and tensorflow, implementing some form of regularization
- A summary of the training error versus generalization error to ensure you didn't overfit the data.
- Estimate the generalization error in your final model. Clearly state how you chose to estimate the generalization error, as well as what final expected value is. This is based on your choice of loss function

Report Section headers:

- Introduction
- Methods summary
- Results
- Conclusions