INFSCI 2160: Homework #1 (37 points)

This homework will cover:

- Linear regression and evaluation techniques
- Logistic regression and evaluation techniques
- KNN and evaluation techniques

Part I: Linear Regression

- Load the dataset 'HW_1_DATASET_DIRECT_MARKETING.csv' available on Piazza under "Homework Datasets"

Question 1: Create a histogram for 'AmountSpent' & 'Salary'. Then describe the shape of both columns as discussed in Week 1 of class:

- 2 points
  - 1 point for successful histograms
  - 1 points for correct shape description

Question #2: We are eventually going to predict 'AmountSpent'. The shape of 'Salary' and 'AmountSpent' go against one of the statistical assumptions of multiple linear regression. What is this statistical assumption? Transform the variables so it satisfies this assumption.

- 2 Points:
  - 1 point for description of the statistical assumption
  - 1 point for transformation

Question 3: Now that 'Salary' is transormed, perform LASSO and Ridge multiple linear regression to predict 'AmountSpent'. Evaluate each model in terms of fit. Which model performs best? How did you find your alpha value? What are your alpha values for the ridge and LASSO model?

- 10 Points

Question #4: Plot the fitted values of your best model and the residuals. Is there a pattern between these points?

- 2 points:
  - 1 point for successful residual vs. fitted values plot
  - 1 point for identifying if there seems to be a pattern or not

Part II: Classification:

- Load the dataset 'HW_1_flight_delays.txt' available on Piazza

Question #5: What is the balance of the "delay" column? Show a count of the on-time vs. late flights.

- 1 point

Question #6: Perform logistic regression to predict flight delays. Implement elastic-net for variable selection. Which L1 ratio did you choose?

- 6 points

Question #7: Create ROC curves for your train and test sets. Do you feel that the model is over or underfit? What are the AUC's?

- 4 points
  - 1 point for each ROC curve
  - 1 point for identifying fit
  - 1 point for AUC's for train and test

Question 8: Decide on a threshold to classify flights in your test set as "on time", or "delayed". Create a confusion matrix. What is your:

- Sensitivity
- Specificity
- PPV
- NPV
- F1 score
- Accuracy

7 points: 1 for each metric, 1 for a confusion matrix

Question #9: Perform KNN on the same dataset to predict flight delays. Plot the AUC for KNN on your test set and cite the AUC. How does it perform in comparison to your logistic regression model? Better or worse in terms of AUC?

HINT : Only include continuous variables are usable for KNN! It is also important to scale all variables together (at once)

- 3 points
  - 1 point for successful KNN on test set
  - 1 point for plotting ROC curve for KNN test set
  - 1 point for correctly comparing and citing which AUC is better between KNN model and logistic regression model on test set