DS 432 Predictive Modeling for Data Science Lab Assignments (Last Updated: 25/9/2018)

A. CRISP-DM

Reference - *Larose & Larose, Data Mining and Predictive Analytics, Wiley, Second Edition. (Part I - Data Preparation; Chapters 1, 2, 3, and 4.)*

Perform the CRISP-DM analysis as per instructions (Data Preprocessing, Exploratory Data Analysis, and if required, Dimension-Reduction Methods) of the four data sets given below.

1. Bank Loan Default
2. Campaign Offers
3. Retail Forecasting
4. Telco Churn

B. Regression - I

Reference - *Julian Faraway, Linear Models with R.*

1. Consider three data sets given below. Remove every tenth observation from the each data set for use as a test sample. Use the remaining data as a training sample and build (Diagnostics -> Transformation -> Variable Selection -> Diagnostics) an appropriate linear regression model. Use the models you find to predict the response in the test sample and compare the results.
   (a) RegD1.txt
   (b) RegD2.txt
   (c) RegD3.txt
2. Use the data set RegD14.txt, to fit a model. Perform regression diagnostics on this model to answer the following questions. Display any plots that are relevant.
   (a) Check and comment on the constant variance assumption for the errors.
   (b) Check and comment on the normality assumption.
   (c) Check and comment on the large leverage points.
   (d) Check and comment on the outliers.
   (e) Check and comment on the influential points.
   (f) Check and comment on the structure of the relationship between the predictors and the response.
   (g) Compute and comment on the condition numbers.
   (h) Compute and comment on the correlations between the predictors.
   (i) Compute and comment on the VIF.

3. Use the data RegD4.txt to fit a model using the following methods.

   (a) Least squares.

   (b) Least absolute deviations.

   (c) Huber method.

   (d) Least trimmed squares.

   Compare the results. Use diagnostic methods to detect any outliers or influential points. Remove these points and then use least squares. Compare the results.

4. Use the data RegD7.txt to fit a model with Y as the response and only $X3$, $X4$, and $X5$ as predictors. Use the Box-Cox method to determine the best transformation on the response.

5. Use the data RegD8.txt to fit a linear model. Implement the following variable selection methods to determine the "best" model.

   (a) Backward Elimination.

   (b) AIC, AICC, BIC.

   (c) R^2, R_a^2.

   (d) Mallows C_p.

6. Consider the data RegD9.txt. Remove every tenth observation from the data for use as a test sample. Use the remaining data as a training sample building the following models.

   (a) Linear regression with all predictors.

   (b) Linear regression with variables selected using AIC.

   (c) Principle component regression.

   (d) Partial least squares.

   (e) Ridge regression.

   Use the models you find to predict the response in the test sample. Make a report on the performance of the models.