DSC 423: Data Analysis and Regression
Assignment 05: Variable Screening

Your submission must include your name and student ID. Your submission must include the honor statement: "I have completed this work independently. The solutions given are entirely my own work." Your submission must be submitted as a PDF.

1. Short Essay. The purpose of *k*-fold cross validation is often misunderstood.
    a. (10 points) How do you use cross validation to select a final (or production) model? Note: it is **not** the "best" of the *k* models you have built using cross validation.
2. PGA. The pgatour2006.csv dataset contains data for 196 players. The variables in the dataset are:
    - Player's name
    - PrizeMoney = average prize money per tournament
    - DrivingAccuracy = percent of times a player is able to hit the fairway with his tee shot
    - GIR = percent of time a player was able to hit the green within two or less than par (Greens in Regulation)
    - BirdieConversion = percentage of times a player makes a birdie or better after hitting the green in regulation
    - PuttingAverage = putting performance on those holes where the green was hit in regulation.
    - PuttsPerRound= average number of putts per round (shots played on the green)
    - Etc.

    a. (10 points) Build a complete first-order model. Evaluate the model using 5-fold cross validation. If necessary, remove a non-significant variable and repeat until you have your final first-order model. Present the model.
    b. (10 points) Evaluate scatterplots to determine which second-order terms should be tested. Test them using 5-fold cross validation and add them one-by-one until you arrive at a model you feel is appropriate. Present the model.
    c. (10 points) Beginning from scratch, engineer all possible second-order terms and add them to your dataset. From this dataset, produce a model using backward selection. Evaluate this model using 5-fold cross validation. Do you arrive at the same model as above? Explain.
    d. (10 points) You have used two procedures to build a second-order model. Compare these two procedures. Which do you think is "best"? Explain.