Assignment 2

Problem 1) (25 points) In assignment 1, you created a logistic regression model for the Caravan Insurance dataset.

1. Visualize the ROC curve by changing the probability threshold for that model (20 points)
2. Report the Area Under Curve (AUC) (5 points)

Problem 2) (25 points) In assignment 1, you created a linear regression model for the cereal dataset. Create another linear regression after removing two variables (sodium and sugars). Compare the AIC, BIC, and Adjusted R-squared variables between the full model and the model with two less variables. Which one is a better model based on those model goodness measures? (split into 20/80% for test/train)

Problem 3) (25 points) add a new variable to the cereal dataset called sodium2 which has double the sodium value plus a random noise from a normal distribution mean zero and standard deviation of 5.

1. Create three linear regression models to predict rating using lasso, ridge, elastic net regularization techniques. Also create an ordinary linear regression model without regularization.
2. For the ridge model, find a proper value for the lambda (alpha) parameter by drawing the R-squared value against the lambda in a diagram (Use k-fold cross validation).
3. Calculate the variance and bias for each model of these four models and compare them.

problem4) KNN(25 points) : Use different Ks(1,2,3,5,10,15,20,25,30,35,40) and build a KNN model to predict the type of Iris in the iris dataset. Use K-fold cross validation and draw the misclassification rate(error rate) against the k in a diagram. What is the best K in this case?

Please submit a python ipynb notebook.