PES University, Bangalore

(Established under Karnataka Act No. 16 of 2013

UE16CS322 – DATA ANALYTICS

Assignment 3- Regression Analysis

Date of Submission: 02/10/2019 **Max. Marks: 20**

In this assignment you will explore the use of Linear Regression, as well as some regularisation techniques (Ridge and Logistic Regression) to predict housing prices - the quintessential regression problem.

For questions 1-3, you will use the Ames Housing Dataset. Question 4 is based on the Shenzhen Housing Prices Dataset. Ensure that you use the correct dataset for each question.

Question 1 (5 points)

One of the assumptions of multiple linear regression is that there must be no multicollinearity between pairs of dependent variables -

a) Why is it necessary to do away with multicollinearity? **(1 point)**
b) Select only the **quantitative** variables and plot a correlogram to visualise the degree of correlation between pairs of variables. If you were to drop variables based on this plot, which one(s) would you drop and why? Go ahead and drop them from the train and test set. (Consider a threshold of 0.8) **(2 points)**
c) Pearson's correlation coefficient cannot be used on categorical variables. Provide a suitable visualisation for the relationship between a categorical variable and the target variable. Plot this graph for the variable "OverallQuality" and indicate its relationship with the target "SalePrice". Why is this variable categorical? **(2 points)**

Question 2 (5 points)

a) Fit a linear regression model to your data, using all of the variables you decided to keep from the previous question **(1 point)**

b) Plot the standardised residuals versus fitted values. What does an ideal plot of this kind look like? Which prerequisite for the application of linear regression does the plot violate? **(1 point)**
c) Use the histogram of the target variable to decide what kind of transformation you can apply to correct the problem you identified. Compare the plots of the original and corrected model **(2 points)**
d) Analyse the residuals to decide whether or not they are normally distributed **(1 point)**

Hint: Plot the summary of your model for these plots (plot(summary(model)))

Question 3 (4 points)

a) What does the measure R^2 indicate? Is it a reliable measure of the goodness of fit of a model? Compare the R^2 values for both models. **(2 points)**
b) Compute the mean absolute squared error for the two models you fit in the previous question (Make sure you account for all the transformations you carried out). Which model gives you a better MAE? **(2 points)**

Question 4 (6 points)

This section explores Regularisation with Ridge and Lasso Regression. You are provided with another dataset - Shenzhen Housing Prices.

a) When would you typically use Regularisation? **(1 point)**
b) Build a model using the training set to predict the attribute SalePrice using **Ridge Regression**. Use the test set to evaluate the model by computing **R^2 and RMSE (2 points)**
c) Build another model using the training set to predict the attribute SalePrice using **Lasso Regression.** Use the test set to evaluate the model by computing **R^2 and RMSE (2 points)**
d) How do Ridge and Lasso Regression differ from each other? **(1 point)**

(Hint : Use the glmnet function in the glmnet package to build the ridge and lasso regression models. Think about how to figure out the regularisation parameter - lambda -> the value that minimises cross validation loss)