Statistical Data Mining I

Homework 2

Student Number:71

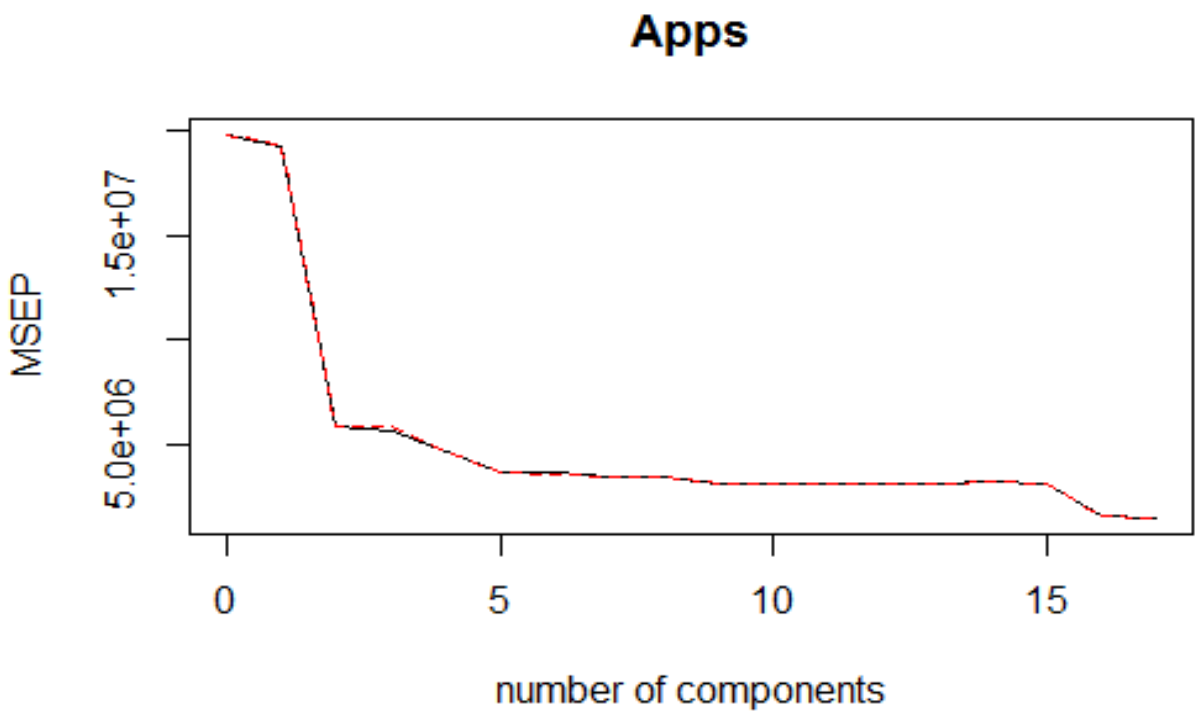
Problem 1:

a) The Test Error obtained after fitting a linear model using least squares method on the training data set is: 1104450

b) The Test Error obtained after fitting a ridge regression model on the training data set with λ (Lambda) is 1120098. The Test error for ridge regression is higher compared to the least squares method.

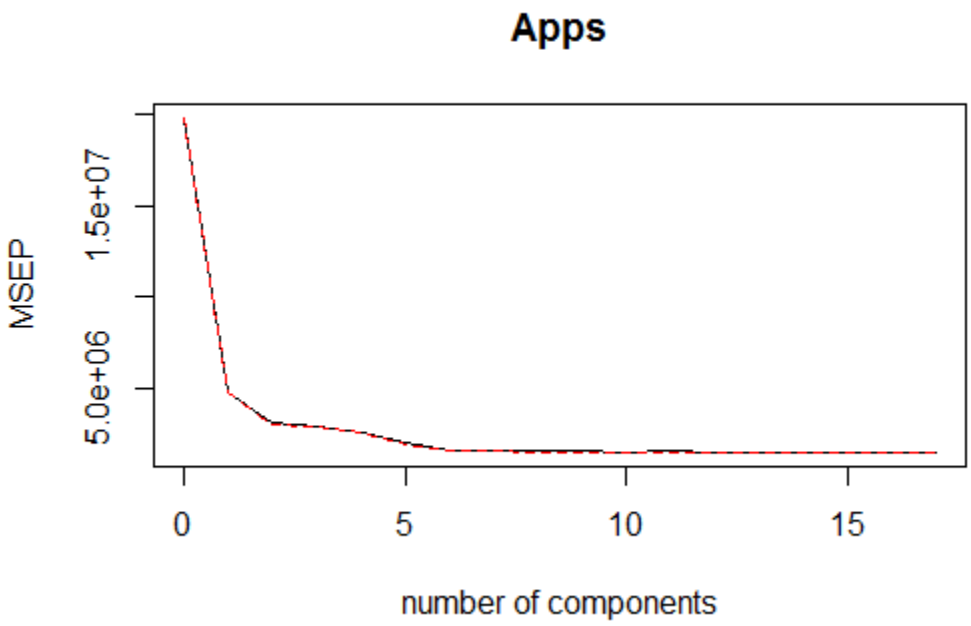
d) The Test Error obtained after fitting a lasso model on the training data set with λ chosen by cross-validation is 1089086. The number of non-zero coefficients estimates is 13.

e)



1. The test error obtained after fitting the PCR model on the training dataset with k chosen by cross validation is 1134126.
2. The value of k selected by cross-validation is 16 as the minimum CV is at k=16.
3. The Test error for PCR model is also higher compared to the least squares method.

f)

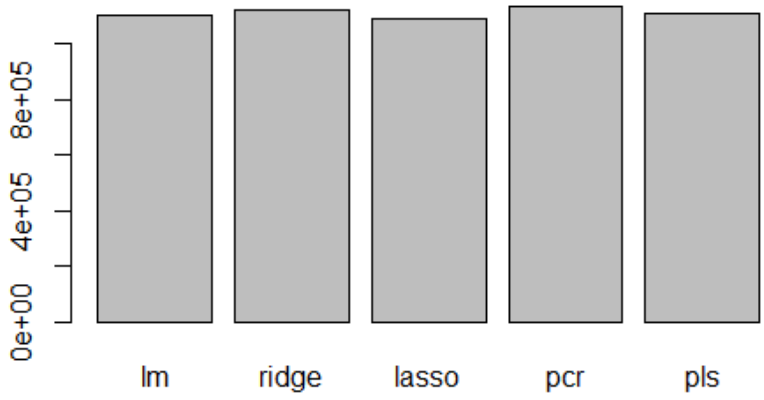


1. The test error obtained after fitting the PLS model on the training dataset with k chosen by cross validation is 1108956.

2. The value of k selected by cross-validation is 10 as the minimum CV is at k=10.

3. The Test error for PLS model is also higher compared to the least squares method.

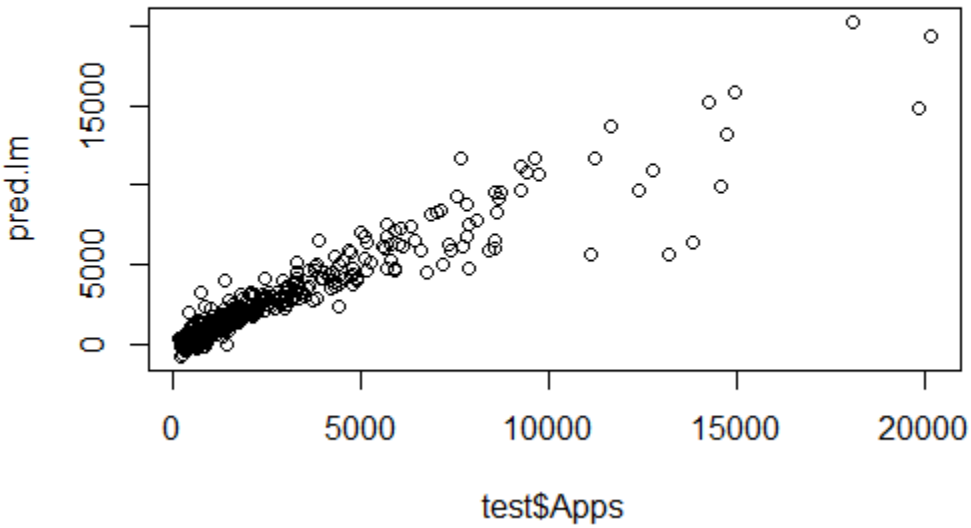
g)



1. There's not much difference between the test errors.

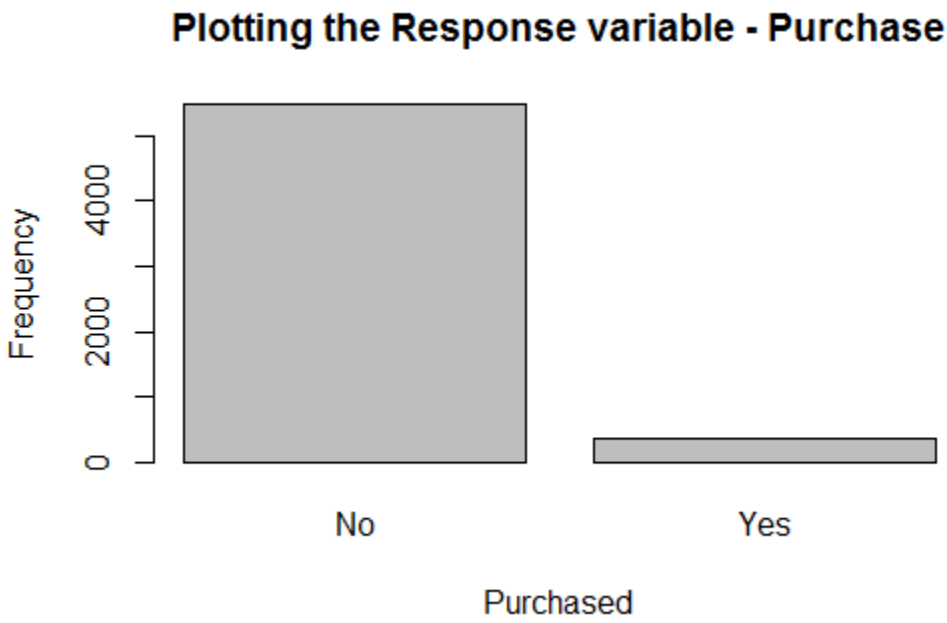
2. The ridge and lasso perform somewhat better than the other models.

3. There's no improvement in PCR and PLS from the full linear regression model.



Problem 2:

Individuals who have purchased the caravan policy:



From "summary(Caravan$Purchase") we can say that

348/5822 = 0.05977 ≈ 6%

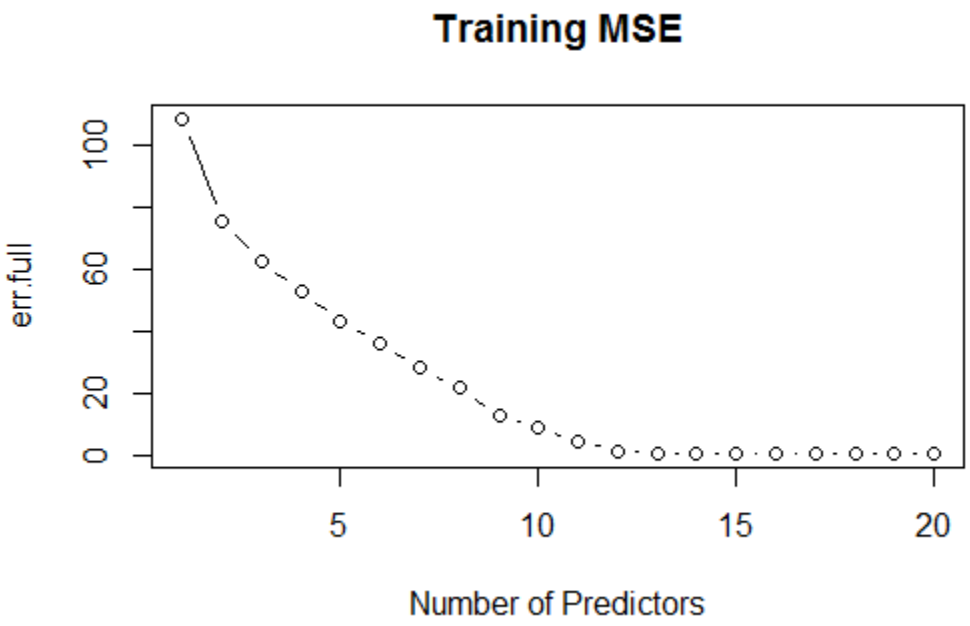
Therefore, the Caravan Policy is purchased by only 6% individuals.

Individuals Interested in buying Caravan Policy:

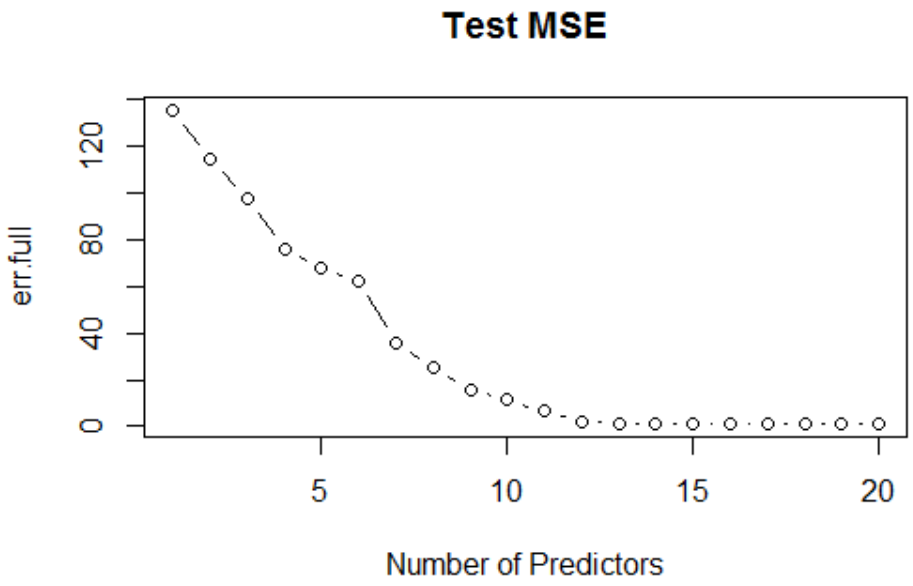
After performing KNN (Taking k=5) we get a prediction that 33% people will purchase the insurance.

Problem 3:

- The training set MSE associated with the best model of each size:



- The test set MSE associated with the best model of each size:



- The model size for which the test set MSE take on its minimum value is 13 variables models.
- The best subset model selected all the correct predictors.
- The best model also caught all zeroed out coefficients.