Project 3

CPSC 483

Professor: Kenytt Avery

Submitted By: Vidhi Patel

CWID: 893375105

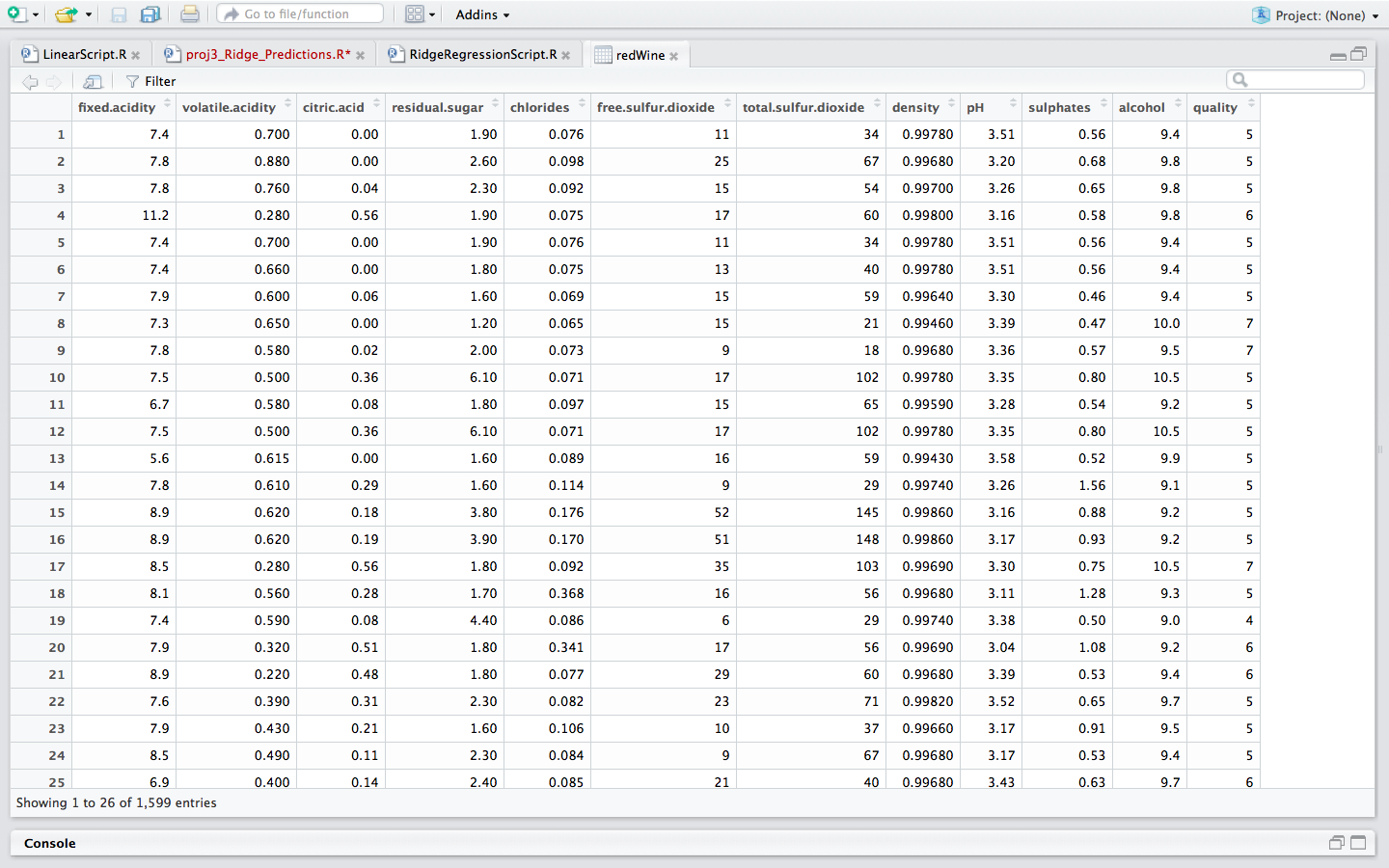


**Introduction**

There are large number of machine learning algorithms which perform different types of analysis on data sets. Here, I will use two different machine learning algorithms to perform analysis on a chosen dataset.

**A. The data set, features, tasks, and algorithms chosen: -**

**1.** I have chosen to analyze red wines data from Wine Quality Data Set which is available in the UCI Machine Learning Repository. The Red Wine Data consist of 12 different set of features. Using this feature, we will analyze the data.



**2.** There are 12 different types of features present in the dataset. Some of the features play a significant role in performing the analysis of data while some of the features does not help in carrying out the analysis. They are just present in the data set giving some additional information. So in order to carry out the analysis by the algorithm, the first thing we will do is to choose appropriate set of features from the data set that play a significant role in carrying out the analysis. To select the Appropriate set of features, I have used a **Forward Direction Approach.**

To perform forward selection, we need to begin by specifying a starting model.

Eg- > step(null, scope=list(lower=null, upper=full), direction="forward")

This tells R to start with the null model and search through models lying in the range between the null and full model using the forward selection algorithm.

Using this approach, the significant features that I am able to select is,

**quality, alcohol, volatile.acidity, sulphates, total.sulfur.dioxide, chlorides, pH, free.sulfur.dioxide**

**3.** I have chosen the set of features that are significant in carrying out the analysis, the next step is to choose a task, that is an algorithm to perform analysis on the chosen features of the data set.

I will use Regression algorithms to predict the quality of the wines depending on the 7 significant attributes chosen in the forward selection. The two different regression algorithms used to predict the quality of wine is Multiple Linear regression and Ridge Regression. I will then compare the performance of the

**B. The procedures used to analyze the data and build models: -**

1. Once the features and algorithms to be used to perform the tasks are decided, the next step is to divide the data set into training and and testing data set,I have used caTools library to split the data into 80% as the training dataset and 20% as the testing data set.

* train\_rows = sample.split(y, SplitRatio=0.8)
* train = data[ train\_rows,]
* test=data[!train\_rows,]

**2.** The next step is to perform the multiple linear regression and Ridge Regression on the testing dataset by using the training dataset.

* **Multiple Linear Regression: -**

Multiple Linear Regression in R is performed using lm() model. So we will use the dependent variable as quality and we will give the chosen 7 attributes as the input to the model and the dataset will be training data.

> lm.fit = lm(quality~.,data=train)

Once the model is trained, we will now predict the quality of the test dataset by using the trained data model which is obtained by using lm() model.

> testdata = data.frame(test)

> pred\_data = predict(lm.fit, testdata)

* **Ridge Regression:-**

Ridge Regression in R is performed using lm.ridge() model. To use this model we will have to import the MASS library. To perform Ridge Regression, we will first find out the best lambda.

* select(lm.ridge(train$quality~.,data=train2,lambda=seq(0,10,0.001)))

modified HKB estimator is 10.67167

modified L-W estimator is 10.80418

smallest value of GCV at 10

Using the lamda 10, and dependent variable as the quality and data as the training set, we will perform Ridge Regression,

> ridgereg = lm.ridge(train$quality~.,data=train2,lambda=10)

> lm.ridge(train$quality~.,data=train2,lambda=10)

Once we train the model with the training data set, we will predict the quality of the wine of the test dataset using the this model

pred\_data <-scale(testdata[,1:7],center=F,scale=ridgereg$scales)%\*% ridgereg$coef

**C. The performance of each of the models on training and test sets.**

The performance of each algorithm can be measured by calculating the error of the training and the test sets.

1. The error observerd in the multiple linear regression is :-

> error <- sqrt((sum((test\_quality-pred\_data)^2))/length(test\_quality))

> error

**[1] 0.6609043**

**Proj3_Linear_Predictions.pdf**

**predicted quality vs observed quality of test data set**

2. The error observerd in the Ridge Regression with 8 significant features is :-

> sumh =0;

> #Add the intercept value obtained from lm.ridge function.

> tt <-nrow(testdata)

> for(i in 1:tt){

+ sumh = sumh+(pred\_data[i]+4.18-testdata[i,8])^2

+ }

> sumg = sumh/tt

> sumg

**[1] 0.4479991**

**part3Ridge.pdf**

**predicted quality vs observed quality of test data set**

3. The error observed in the Ridge Regression with all features is :-

> sumh<-0

> tt <- nrow(test)

> #Add the intercept value obtained from lm.ridge function.

> for(i in 1:tt){

+ sumh=sumh+(pred\_data[i]+6.14-test[i,12])^2

+ }

> error = sumh/tt

> error

**[1] 0.421975**

Part3_ridge1.pdf

**predicted quality vs observed quality of test data set**

On observing the error of all the three models, I conclude that Ridge Regression helps in better predictions than Multiple Linear Regression even if we don’t select the features and perform the Ridge Regression using all the features. Because Ridge Regression helps in Shrinking the features instead of removing it.

**4. Refernces:-**

**1.** Lichman, M. (2013). UCI Machine Learning Repository [<http://archive.ics.uci.edu/ml>]. Irvine, CA: University of California, School of Information and Computer Science.

**2**. P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by

data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.