

machine learning(732A99) lab1

Anubhav Dikshit(anudi287)

26 November 2018

Assignment 1

Loading The Libraries

Loading Input files

```
spam_data <- read.xlsx("spambase.xlsx", sheetName = "spambase_data")
spam_data$Spam <- as.factor(spam_data$Spam)

tecator_data <- read.xlsx("tecator.xlsx", sheetName = "data")
tecator_data <- tecator_data[,2:NCOL(tecator_data)] # removing sample column
```

1.1 Import the data into R and divide it into training and test sets (50%/50%) by using the following code

```
set.seed(12345)

n = NROW(spam_data)
id = sample(1:n, floor(n*0.5))
train = spam_data[id,]
test = spam_data[-id,]
```

1.2 Use logistic regression (functions glm(), predict()) to classify the training and test data by the classification principles

Manual Feature Selection

```
best_model <- glm(formula = Spam ~ ., family = binomial, data = train)
summary(best_model)
```

```
##
## Call:
## glm(formula = Spam ~ ., family = binomial, data = train)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.5205  -0.4402  -0.0005   0.6584   3.6196
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.5081289   0.2011152   7.499 0.00000000000000644 ***
## Word1       -0.7191881   0.5014624  -1.434  0.151520
```

```

## Word2      0.0399446      0.3014438      0.133      0.894580
## Word3     -0.3528561      0.1839013     -1.919      0.055019 .
## Word4      0.1369914      0.1117461      1.226      0.220230
## Word5      0.1221434      0.1413138      0.864      0.387400
## Word6      0.2887448      0.4153010      0.695      0.486888
## Word7     -0.2947580      0.3348237     -0.880      0.378676
## Word8     -0.1033859      0.3509944     -0.295      0.768337
## Word9     -0.1084832      0.4053447     -0.268      0.788983
## Word10    -0.0309116      0.1668155     -0.185      0.852991
## Word11    -0.6088010      0.6789682     -0.897      0.369902
## Word12      0.1614068      0.1072619      1.505      0.132378
## Word13      0.7810953      0.3572174      2.187      0.028771 *
## Word14    -0.4818689      0.3003186     -1.605      0.108598
## Word15    -0.1305229      0.3884464     -0.336      0.736861
## Word16      0.3170597      0.2331654      1.360      0.173891
## Word17    -0.0920067      0.2822914     -0.326      0.744479
## Word18      0.0232959      0.2321682      0.100      0.920074
## Word19      0.0003694      0.0574594      0.006      0.994871
## Word20      0.0168793      0.3181387      0.053      0.957687
## Word21    -0.0282058      0.1072320     -0.263      0.792524
## Word22    -0.4767281      0.3200436     -1.490      0.136337
## Word23      0.2541081      0.3412766      0.745      0.456525
## Word24    -0.2483026      0.6254580     -0.397      0.691372
## Word25    -0.0782762      0.0585199     -1.338      0.181027
## Word26      0.0037333      0.1358210      0.027      0.978071
## Word27    -0.2238407      0.1077417     -2.078      0.037749 *
## Word28      0.1319556      0.1880126      0.702      0.482776
## Word29    -0.0813140      0.0911918     -0.892      0.372564
## Word30    -1.8150700      0.6195133     -2.930      0.003391 **
## Word31    -4.6944446      1.8530163     -2.533      0.011296 *
## Word32   -119.4495067 15134.1847623 -0.008      0.993703
## Word33    -2.8994424      0.6794146     -4.268 0.0000197622951954 ***
## Word34    -3.7098525      4.3523571     -0.852      0.394004
## Word35    -7.0325274      1.9964683     -3.522      0.000428 ***
## Word36    -1.6779354      0.3810251     -4.404 0.0000106400723118 ***
## Word37    -0.8583374      0.2174799     -3.947 0.0000792212684576 ***
## Word38    -0.6043466      1.2790588     -0.472      0.636575
## Word39    -1.8771543      0.4281730     -4.384 0.0000116464815552 ***
## Word40      0.0739289      0.3400368      0.217      0.827885
## Word41   -332.5541421 16559.5233247 -0.020      0.983978
## Word42    -5.3516109      1.3024490     -4.109 0.0000397576929183 ***
## Word43    -2.5919298      0.7352904     -3.525      0.000423 ***
## Word44    -2.9313923      0.6600870     -4.441 0.0000089575885095 ***
## Word45    -1.1408036      0.1681342     -6.785 0.0000000000116025 ***
## Word46    -3.2880431      0.5178052     -6.350 0.0000000002153691 ***
## Word47    -3.7410844      2.0304190     -1.843      0.065399 .
## Word48    -4.3898995      1.4729045     -2.980      0.002878 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 1696.82 on 1369 degrees of freedom
## Residual deviance: 928.54 on 1321 degrees of freedom

```

```
## AIC: 1026.5
##
## Number of Fisher Scoring iterations: 23
```

Prediction for probability greater than 50% and 90%

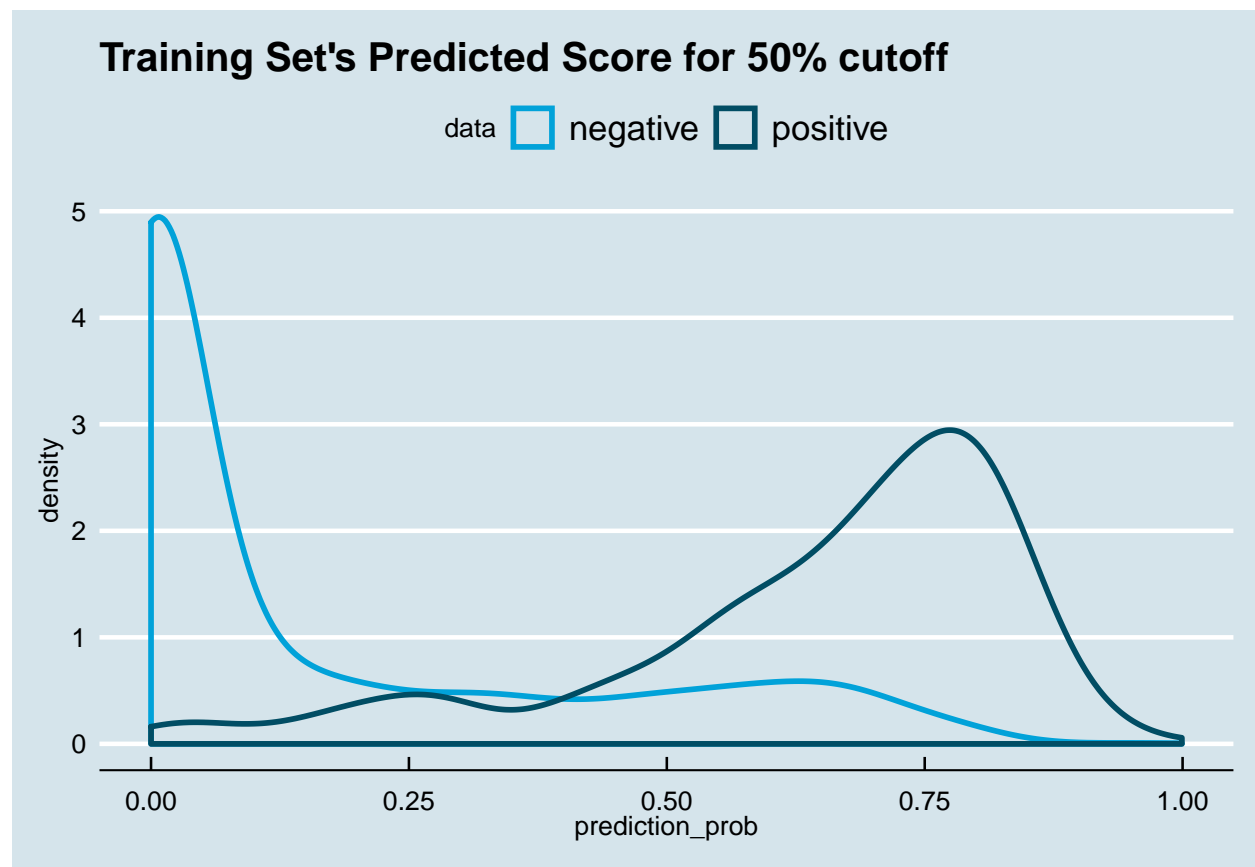
```
# prediction
train$prediction_prob <- predict(best_model, newdata = train, type = "response")
test$prediction_prob <- predict(best_model, newdata = test, type = "response")

train$prediction_class_50 <- ifelse(train$prediction_prob > 0.50, 1, 0)
test$prediction_class_50 <- ifelse(test$prediction_prob > 0.50, 1, 0)

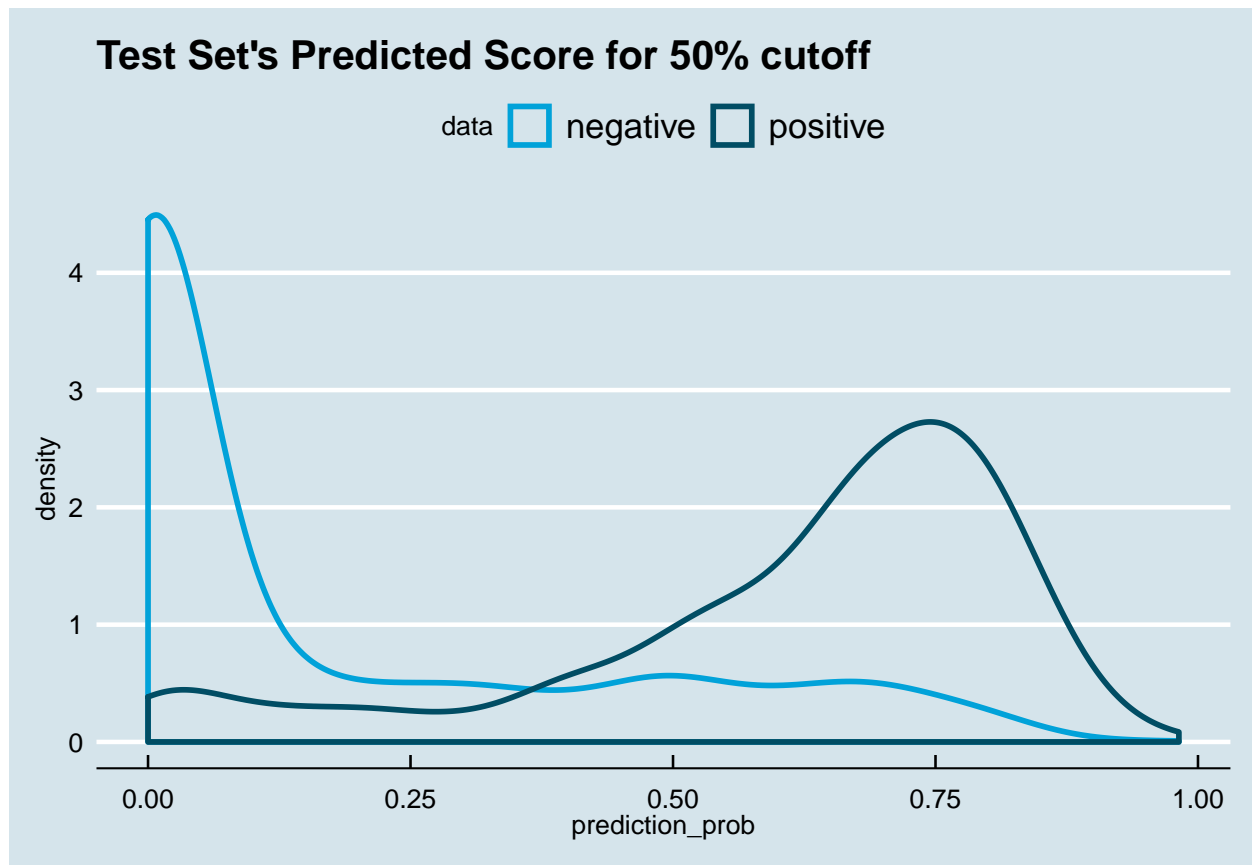
train$prediction_class_90 <- ifelse(train$prediction_prob > 0.90, 1, 0)
test$prediction_class_90 <- ifelse(test$prediction_prob > 0.90, 1, 0)
```

Assessing the Model

```
# plots
ggplot(train, aes(prediction_prob, color = Spam)) +
  geom_density(size = 1) + ggtitle("Training Set's Predicted Score for 50% cutoff") +
  scale_color_economist(name = "data", labels = c("negative", "positive")) +
  theme_economist()
```



```
ggplot(test, aes(prediction_prob, color = Spam)) +
  geom_density(size = 1) + ggtitle("Test Set's Predicted Score for 50% cutoff") +
  scale_color_economist(name = "data", labels = c("negative", "positive")) +
  theme_economist()
```



1.2 Assessing the Fit on train dataset for 50%

```
#confusion table
conf_train <- table(train$Spam, train$prediction_class_50)
names(dimnames(conf_train)) <- c("Actual Train", "Predicted Train")
confusionMatrix(conf_train)
```

```
## Confusion Matrix and Statistics
##
##               Predicted Train
## Actual Train  0    1
##               0 803 142
##               1  81 344
##
##               Accuracy : 0.8372
##               95% CI : (0.8166, 0.8564)
##               No Information Rate : 0.6453
##               P-Value [Acc > NIR] : < 0.00000000000000022
##
##               Kappa : 0.6341
```

```

## McNemar's Test P-Value : 0.00005872
##
##          Sensitivity : 0.9084
##          Specificity : 0.7078
##          Pos Pred Value : 0.8497
##          Neg Pred Value : 0.8094
##          Prevalence : 0.6453
##          Detection Rate : 0.5861
##          Detection Prevalence : 0.6898
##          Balanced Accuracy : 0.8081
##
##          'Positive' Class : 0
##

conf_test <- table(test$Spam, test$prediction_class_50)
names(dimnames(conf_test)) <- c("Actual Test", "Predicted Test")
confusionMatrix(conf_test)

## Confusion Matrix and Statistics
##
##          Predicted Test
## Actual Test  0    1
##          0 791 146
##          1  97 336
##
##          Accuracy : 0.8226
##          95% CI : (0.8014, 0.8425)
##          No Information Rate : 0.6482
##          P-Value [Acc > NIR] : < 0.00000000000000022
##
##          Kappa : 0.6018
## McNemar's Test P-Value : 0.002076
##
##          Sensitivity : 0.8908
##          Specificity : 0.6971
##          Pos Pred Value : 0.8442
##          Neg Pred Value : 0.7760
##          Prevalence : 0.6482
##          Detection Rate : 0.5774
##          Detection Prevalence : 0.6839
##          Balanced Accuracy : 0.7939
##
##          'Positive' Class : 0
##

```

Analysis: Distribution of the prediction score grouped by known outcome given that our model's final objective is to classify new instances into one of two categories (spam vs. non-spam). We will want the model to give high scores to positive instances (1: spam) and low scores (0 : not spam) otherwise. Ideally you want the distribution of scores to be separated, with the score of the negative instances to be on the left and the score of the positive instance to be on the right.

From the confusion matrix it is apparent that Accuracy on train and test dataset when cutoff=50% is about ~84% for train and ~82% for test, thus the misclassification rate is ~16 and ~18% for the train and test dataset.

1.3 Assessing the Fit on train dataset for 90%

```
#confusion table
conf_train1 <- table(train$Spam, train$prediction_class_90)
names(dimnames(conf_train1)) <- c("Actual Train", "Predicted Train")
conf_train1
```

```
##           Predicted Train
## Actual Train  0    1
##           0 944    1
##           1 419    6
```

```
conf_test1 <- table(test$Spam, test$prediction_class_90)
names(dimnames(conf_test1)) <- c("Actual Test", "Predicted Test")
conf_test1
```

```
##           Predicted Test
## Actual Test  0    1
##           0 936    1
##           1 427    6
```

Analysis: Strange, the model almost only predicts one class!! We know that the prediction of a logistic regression model is a probability, thus in order to use it as a classifier, we'll have to choose a cutoff value, or threshold (cutoff). Where scores above this value will be classified as positive, those below as negative. Let's find this optimum value.

Choosing the best cutoff for test

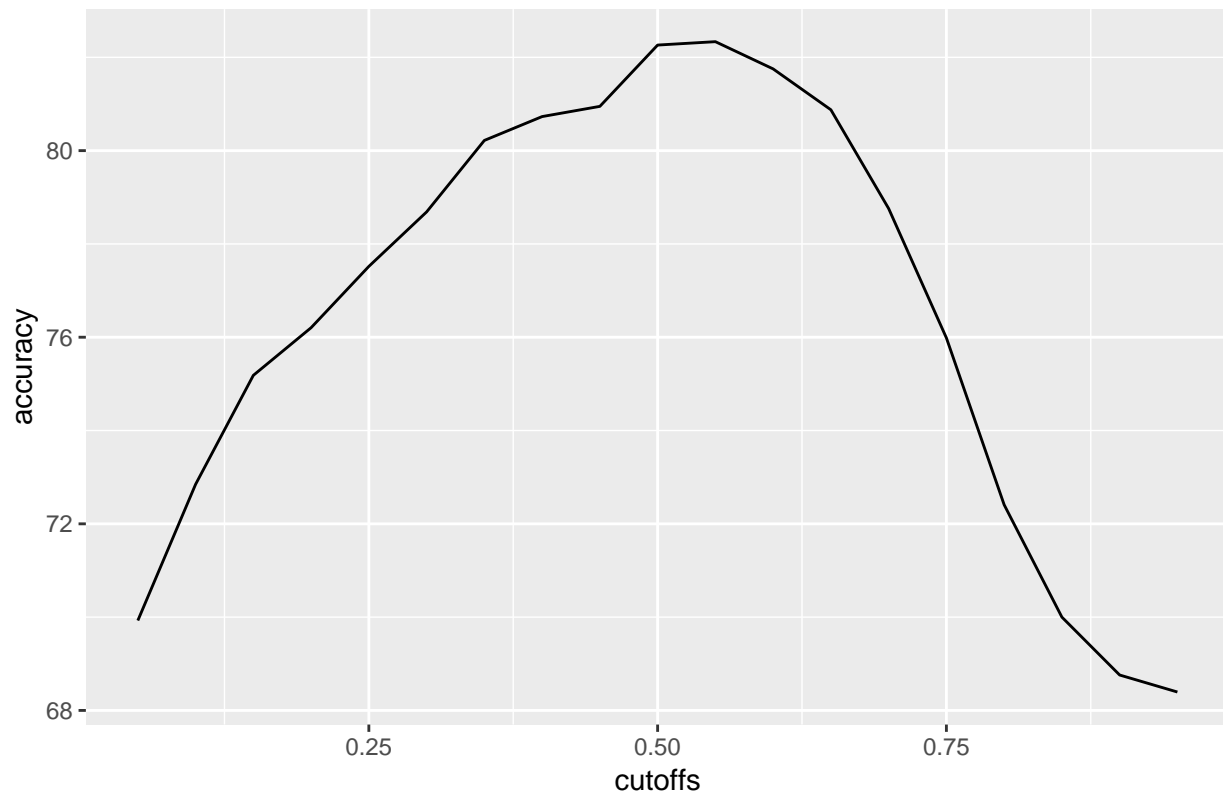
```
cutoffs <- seq(from = 0.05, to = 0.95, by = 0.05)
accuracy <- NULL

for (i in seq_along(cutoffs)){
  prediction <- ifelse(test$prediction_prob >= cutoffs[i], 1, 0) #Predicting for cut-off
  accuracy <- c(accuracy, length(which(test$Spam == prediction))/length(prediction)*100)}

cutoff_data <- as.data.frame(cbind(cutoffs, accuracy))

ggplot(data = cutoff_data, aes(x = cutoffs, y = accuracy)) +
  geom_line() +
  ggtitle("Cutoff vs. Accuracy for Test Dataset")
```

Cutoff vs. Accuracy for Test Dataset



Analysis: Our small detour suggests that the cutoff value of ~60% was the best for our purpose and going higher than this leads to worse results, at 70% and above the accuracy drastically reduces which is what we see when we make cutoff as 90%.

From the confusion matrix it is evident that the model becomes a trivial model(predicts all cases as one class) and thus the prediction is worse than just tossing a coin. This should be the absolutely the worst case that we should avoid.

The missclassification rate is about 31% for both the training dataset and test dataset.

1.4 Use standard classifier `knn()` with $K=30$ from package `knn`, report the the misclassification rates for the training and test data and compare the results with step 1.2.

```
knn_model30 <- train.kknn(Spam ~., data = train, kmax = 30)

train$knn_prediction_class <- predict(knn_model30, train)
test$knn_prediction_class <- predict(knn_model30, test)

conf_train2 <- table(train$Spam, train$knn_prediction_class)
names(dimnames(conf_train2)) <- c("Actual Train", "Predicted Train")
confusionMatrix(conf_train2)

## Confusion Matrix and Statistics
##
##               Predicted Train
```

```
## Actual Train   0   1
##               0 845 100
##               1  91 334
##
##               Accuracy : 0.8606
##               95% CI : (0.8411, 0.8785)
##               No Information Rate : 0.6832
##               P-Value [Acc > NIR] : <0.0000000000000002
##
##               Kappa : 0.6761
## Mcnemar's Test P-Value : 0.5627
##
##               Sensitivity : 0.9028
##               Specificity : 0.7696
##               Pos Pred Value : 0.8942
##               Neg Pred Value : 0.7859
##               Prevalence : 0.6832
##               Detection Rate : 0.6168
##               Detection Prevalence : 0.6898
##               Balanced Accuracy : 0.8362
##
##               'Positive' Class : 0
##
```

```
conf_test2 <- table(test$Spam, test$knn_prediction_class)
names(dimnames(conf_test2)) <- c("Actual Test", "Predicted Test")
confusionMatrix(conf_test2)
```

```
## Confusion Matrix and Statistics
##
##               Predicted Test
## Actual Test   0   1
##               0 801 136
##               1 139 294
##
##               Accuracy : 0.7993
##               95% CI : (0.7771, 0.8202)
##               No Information Rate : 0.6861
##               P-Value [Acc > NIR] : <0.0000000000000002
##
##               Kappa : 0.5348
## Mcnemar's Test P-Value : 0.904
##
##               Sensitivity : 0.8521
##               Specificity : 0.6837
##               Pos Pred Value : 0.8549
##               Neg Pred Value : 0.6790
##               Prevalence : 0.6861
##               Detection Rate : 0.5847
##               Detection Prevalence : 0.6839
##               Balanced Accuracy : 0.7679
##
##               'Positive' Class : 0
##
```


Analysis: Using KNN with $K = 30$, we increased our training accuracy to 86%, thus misclassification is 14%, however for the test dataset misclassification rate is about ~20%.

Thus compared to using logistic model the misclassification error for the training dataset decreased by 2% to 14%, while for the test dataset the misclassification error increased by 2% to 20%.

1.5 Repeat step 4 for $K=1$ and compare the results with step 4. What effect does the decrease of K lead to and why?

```
knn_model1 <- train.kknn(Spam ~., data = train, kmax = 1)

train$knn_prediction_class <- predict(knn_model1, train)
test$knn_prediction_class <- predict(knn_model1, test)

conf_train2 <- table(train$Spam, train$knn_prediction_class)
names(dimnames(conf_train2)) <- c("Actual Train", "Predicted Train")
confusionMatrix(conf_train2)
```

```
## Confusion Matrix and Statistics
##
##               Predicted Train
## Actual Train  0    1
##               0 945    0
##               1    0 425
##
##               Accuracy : 1
##               95% CI : (0.9973, 1)
##      No Information Rate : 0.6898
##      P-Value [Acc > NIR] : < 0.00000000000000022
##
##               Kappa : 1
##  Mcnemar's Test P-Value : NA
##
##               Sensitivity : 1.0000
##               Specificity : 1.0000
##               Pos Pred Value : 1.0000
##               Neg Pred Value : 1.0000
##               Prevalence : 0.6898
##               Detection Rate : 0.6898
##      Detection Prevalence : 0.6898
##               Balanced Accuracy : 1.0000
##
##      'Positive' Class : 0
##
```

```
conf_test2 <- table(test$Spam, test$knn_prediction_class)
names(dimnames(conf_test2)) <- c("Actual Test", "Predicted Test")
confusionMatrix(conf_test2)
```

```
## Confusion Matrix and Statistics
##
##               Predicted Test
## Actual Test  0    1
##               0 750 187
```

```
##          1 150 283
##
##          Accuracy : 0.754
##          95% CI : (0.7303, 0.7766)
##    No Information Rate : 0.6569
##    P-Value [Acc > NIR] : 0.000000000000004691
##
##          Kappa : 0.4438
## Mcnemar's Test P-Value : 0.04987
##
##          Sensitivity : 0.8333
##          Specificity : 0.6021
##    Pos Pred Value : 0.8004
##    Neg Pred Value : 0.6536
##          Prevalence : 0.6569
##    Detection Rate : 0.5474
##    Detection Prevalence : 0.6839
##    Balanced Accuracy : 0.7177
##
##    'Positive' Class : 0
##
```

Analysis: Using KNN with $K = 1$, we increased our training accuracy to 100%, thus misclassification is 0%, however for the test dataset accuracy is ~75% thus misclassification rate is about ~25%, thus we improved on the training accuracy but did bad on the test case, thus this is an example of overfitting leading to more variance.

Explanation: The KNN works in the following way, An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors. If $k = 1$, then the object is simply assigned to the class of that single nearest neighbor. Thus $K=1$, makes the separation boundary to be very complex and locally optimised (lots of local clusters), while as K goes higher, the decision boundary becomes more linear/simple.

Assignment 2 Feature selection by cross-validation in a linear model

2.1 Implement an R function that performs feature selection (best subset selection) in linear regression by using k-fold cross-validation without using any specialized function like `lm()` (use only basic R functions)

```
subset_function <- function(X,Y,N){

  # X = swiss[,2:6]
  # Y = swiss[,1:1]
  # N = 5

  df <- cbind(X,Y)
  temp <- NULL
  final <- NULL

  for(i in 1:NCOL(X)){
    combs <- as.data.frame(gtools::combinations(NCOL(X), r=i, v=colnames(X), repeats.allowed=FALSE))
    combs <- tidyr::unite(combs, "formula", sep = ",")
  }
}
```

```

temp <- rbind(combs, temp)
}

set.seed(12345)
df2 <- df[sample(nrow(df)),]
df2$k_fold <- sample(N, size = nrow(df), replace = TRUE)

result <- NULL

for (j in 1:NROW(temp))
{
  for(i in 1:N){

train = df2[df2$k_fold != i,]
test = df2[df2$k_fold == i,]

vec <- temp[j,]
train_trimmed = lapply(strsplit(as.character(vec), ","), function(x) train[x])[[1]]
test_trimmed = lapply(strsplit(as.character(vec), ","), function(x) test[x])[[1]]

y_train = train[,c("Y"), drop = FALSE]
y_test = test[,c("Y"), drop = FALSE]

train_trimmed = as.matrix(train_trimmed)
test_trimmed = as.matrix(test_trimmed)
y_test = as.matrix(y_test)
y_train = as.matrix(y_train)

t_train = as.matrix(t(train_trimmed))
t_test = as.matrix(t(test_trimmed))

betas = solve(t_train %*% train_trimmed) %*% (t_train %*% y_train)

train_trimmed = as.data.frame(train_trimmed)
test_trimmed = as.data.frame(test_trimmed)

train_trimmed$type = "train"
test_trimmed$type = "test"
final <- rbind(train_trimmed, test_trimmed)

y_hat_val = as.matrix(final[,1:(ncol(final)-1)]) %*% betas
mse = (Y - y_hat_val)^ 2

data <- cbind(i, vec, mse, type = final$type)
result <- rbind(data, result)

}
}

result <- as.data.frame(result)

colnames(result) <- c("kfold", "variables", "mse", "type")

```

```

result$mse <- as.numeric(result$mse)
result$no_variables <- nchar(as.character(result$variables)) - nchar(gsub('\\\\', '\\', result$variables))

variable_performance <- result %>% group_by(kfold, no_variables) %>%
  summarise(MSE = mean(mse, na.rm = TRUE))

myplot <- ggplot(data = variable_performance, aes(x = no_variables, y = MSE, color=kfold)) +
  geom_line() + ggtitle("Plot of MSE vs. Number of variables by folds")

myplot2 <- ggplot(data = result, aes(x = variables, y = mse, color=kfold)) +
  geom_bar(stat="identity") + ggtitle("Plot of MSE vs. Features by folds") + coord_flip()

best_variables <- result %>% group_by(type, variables) %>% filter(type == "test") %>%
  summarise(MSE = mean(mse, na.rm = TRUE)) %>% arrange(MSE) %>% select(variables) %>% slice(1) %>% as.v

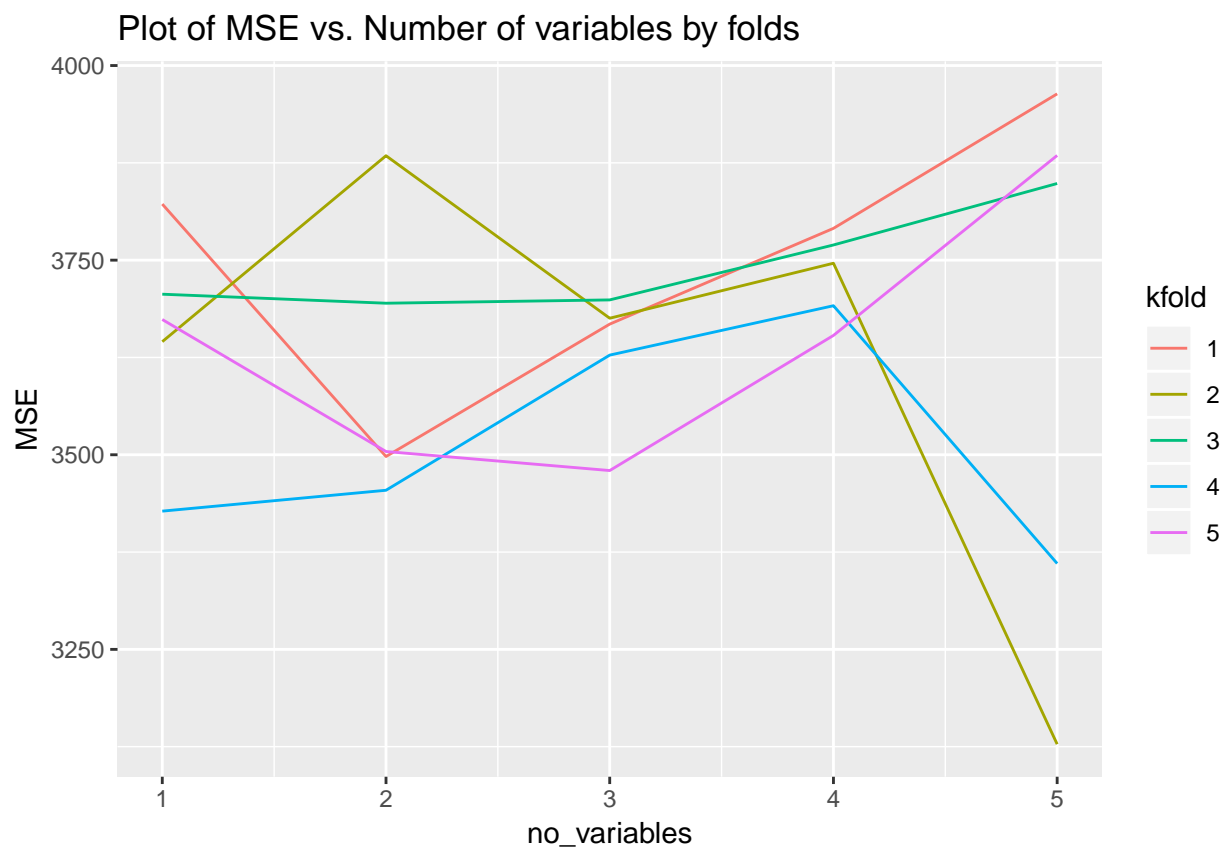
return(list(myplot, myplot2, best_variables))
}

```

2.2 Test your function on data set swiss available in the standard R repository:

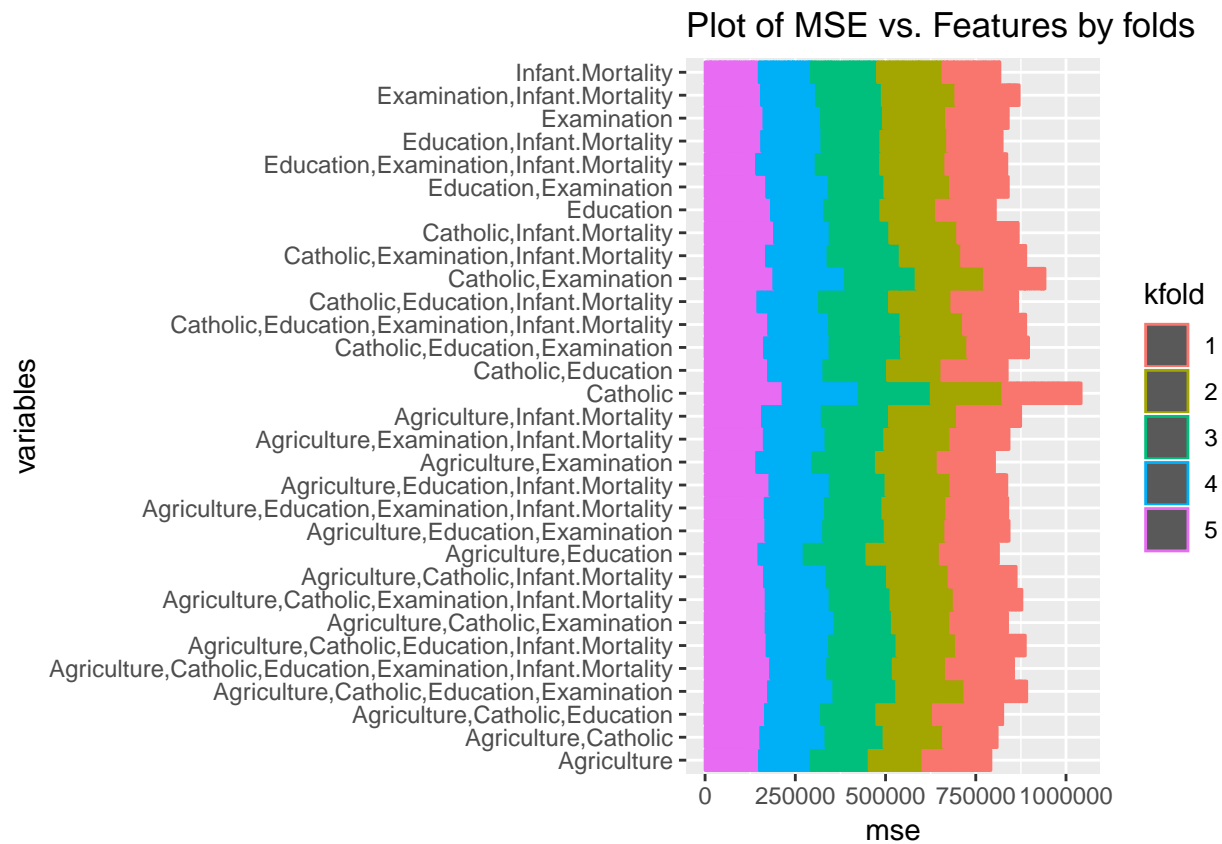
```
subset_function(X = swiss[,2:6], Y = swiss[,1:1], N = 5)
```

```
## [[1]]
```



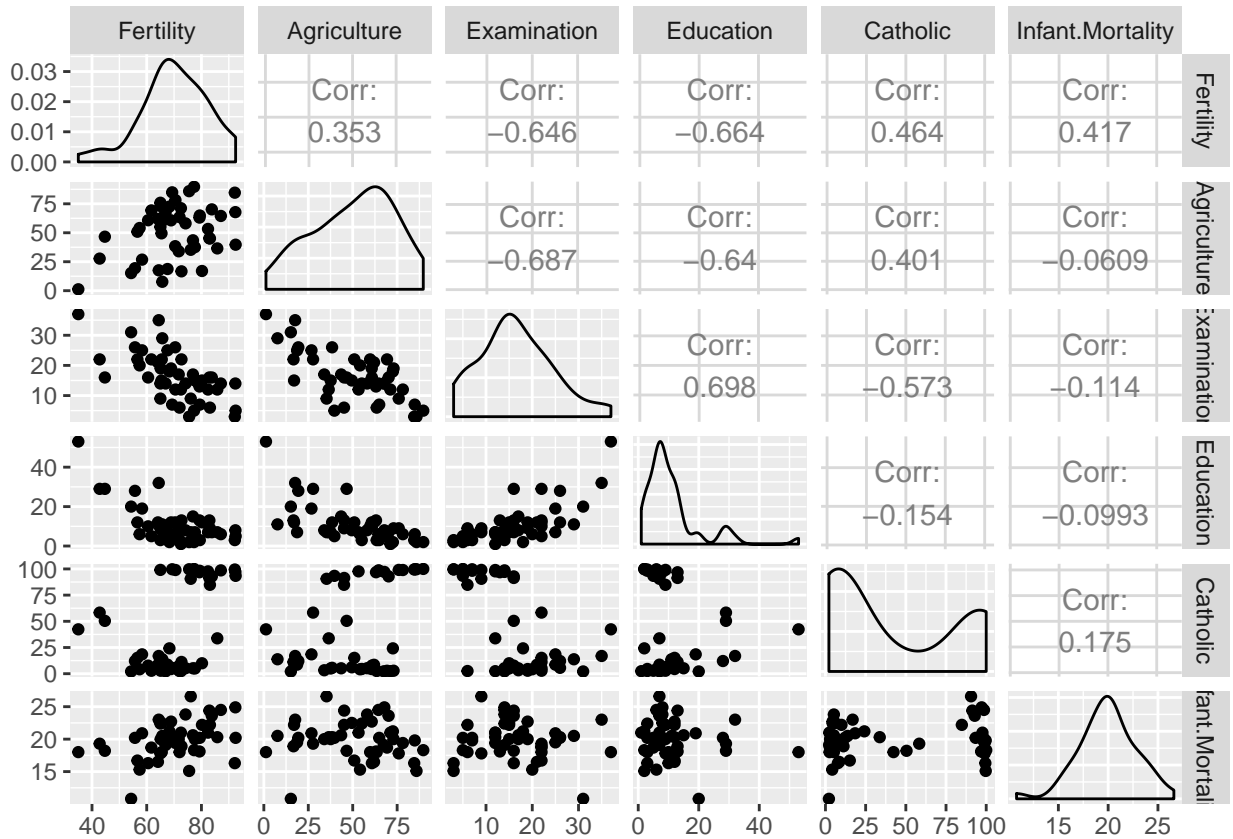
```
##
```

```
## [[2]]
```



```
##
## [[3]]
## # A tibble: 1 x 2
## # Groups:   type [1]
##   type variables
##   <fct> <fct>
## 1 test  Agriculture,Examination
```

```
# Using pairs on swiss dataset
GGally::ggpairs(swiss)
```



Analysis: Swiss dataset is a dataset which represents fertility measure and socio-economic indicators for each of 47 French speaking provinces of Switzerland.

From the dataset we have the following information:

Fertility -> 'common standardized fertility measure' Agriculture -> % of males involved in agriculture as occupation Examination -> % draftees receiving highest mark on army examination Education -> % education beyond primary school for draftees. Catholic -> % 'catholic' (as opposed to 'protestant'). Infant.Mortality -> live births who live less than 1 year.

From the graph we can see that there is correlation between Fertility and other variables, for variable 'Catholic' and 'Education', the relationship is not strong, while other variable do exhibit a reasonable correlation.

From our model we can see that the models with 2 features gives us the least mean squared error across most folds. Implies our best model is a function of 2 variables.

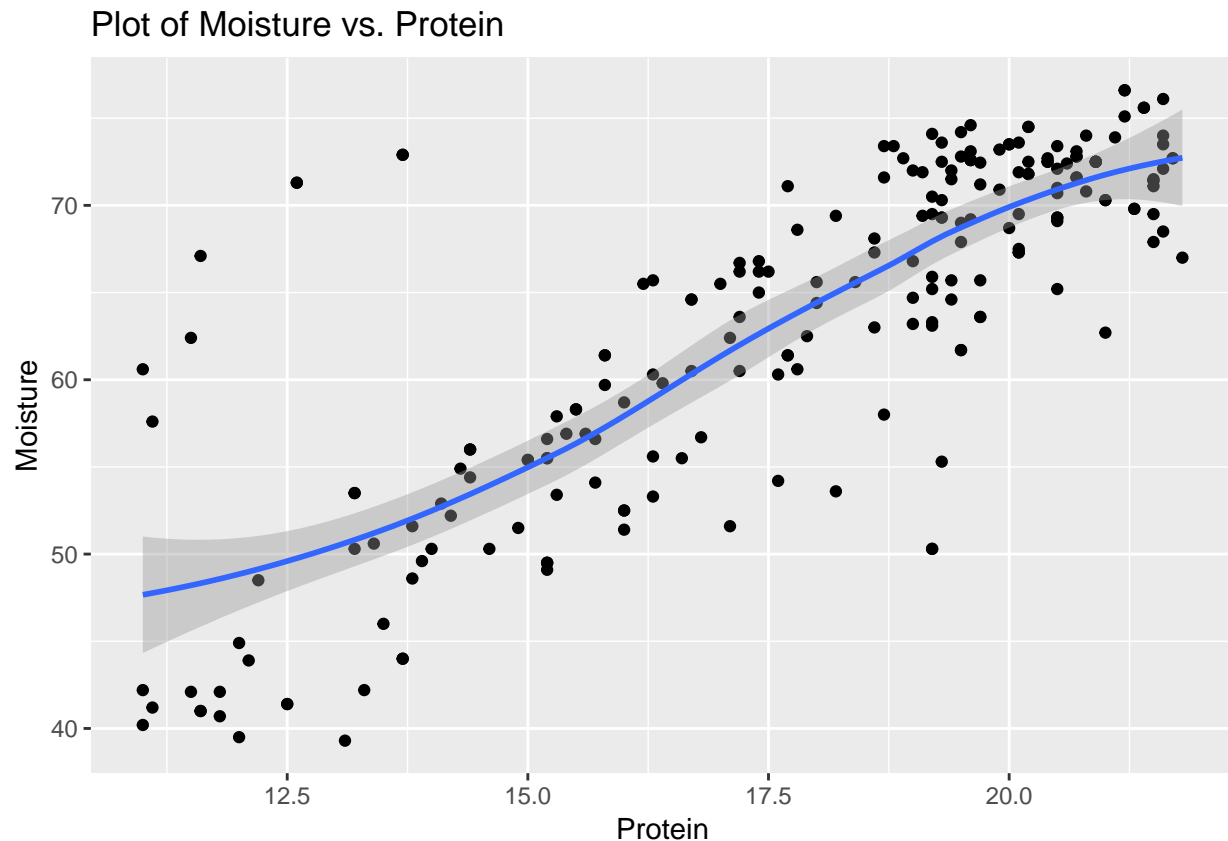
Our model predicts that variable 'Examination' and 'Agriculture' are the strongest predictors of 'Fertility', these results makes sense but from the graph it could be seen that even 'Infant.Mortality' might have a strong predictor.

Logically speaking infant mortality cannot be direct estimator of fertility, higher infant mortality could be responsible in population decrease but not in reduction in fertility.

Assignment 3 Linear regression and regularization

3.1 Import data to R and create a plot of Moisture versus Protein. Do you think that these data are described well by linear model.

```
ggplot(data = tecator_data, aes(x = Protein, y = Moisture)) +  
  geom_point() +  
  geom_smooth( method = 'loess') +  
  ggtitle("Plot of Moisture vs. Protein")
```



Analysis: The data seems fairly linear in nature however there are many outliers. As we can see that data is fairly distributed around the line drawn (above and below) thus there is little bias.

3.2 Multiple Models of varying degree.

$$M_i = \sum_{i=0}^p X^i Protein * \beta_i + \epsilon$$

$$\epsilon \sim N(0, \sigma^2)$$

$$\epsilon = M_i - \sum_{i=0}^p X^i Protein * \beta_i$$

$$M_i \sim N \left(\sum_{i=0}^p X^i Protein * \beta_i, \sigma_M^2 \right)$$

or

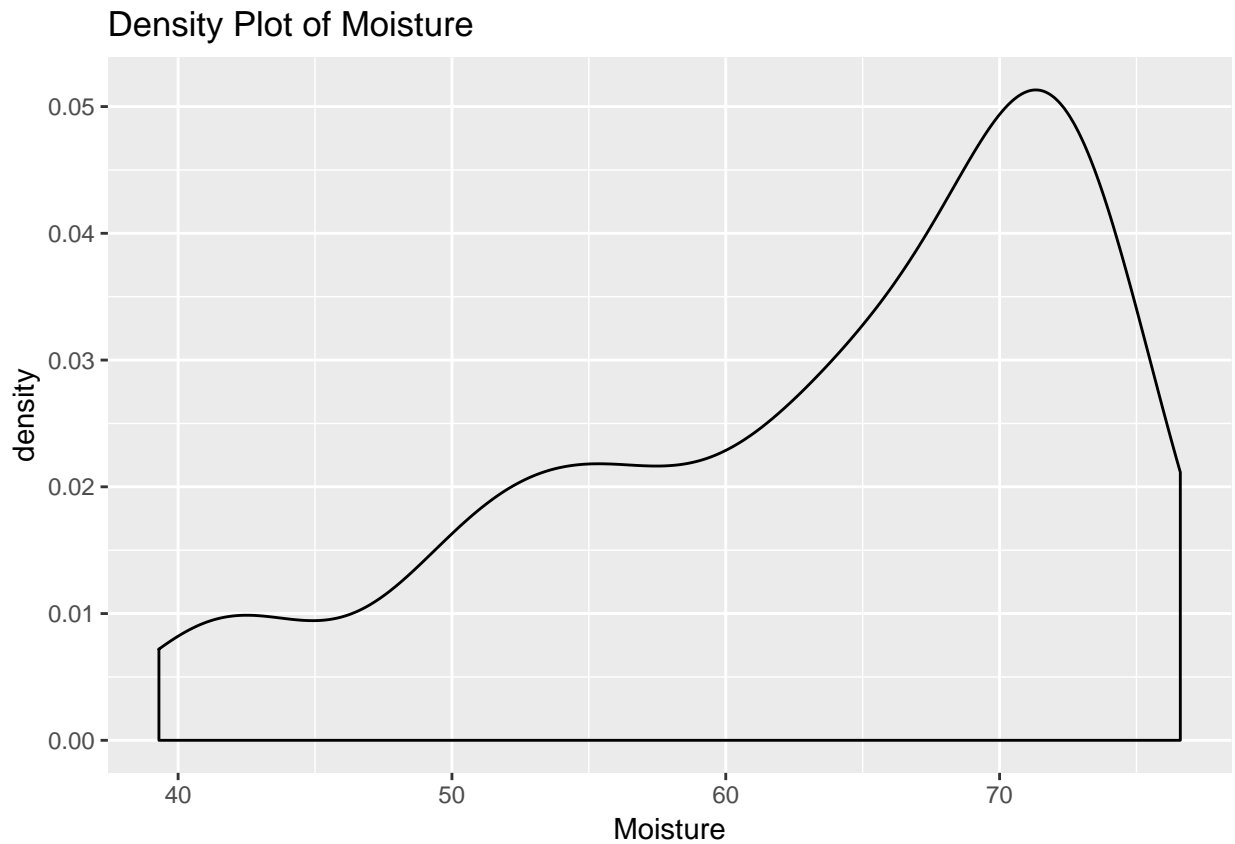
$$P \left(M_i | X_{Protein}, \vec{\beta} \right) = N \left(\sum_{i=0}^p X^i Protein * \beta_i, \sigma_M^2 \right)$$

Where,

σ_M^2 : variance of Moisture

p : degree of the polynomial

```
ggplot(data = tecator_data, aes(x = Moisture)) +  
  geom_density() +  
  ggtitle("Density Plot of Moisture")
```



Analysis: In this case we are given that moisture is normally distributed, thus the loss function to minimize had to be $(\text{actual} - \text{predicted})^2$, if we were to minimize the absolute value, then this would assume a Laplace distribution.

3.3 Validation of the Model

```
final_data <- tecator_data

magic_function <- function(df, N)
{
  df2 <- df
  for(i in 2:N)
  {
    df2[paste("Protein_",i,"_power", sep="")] <- (df2$Protein)^i
  }

  df2 <- df2[c("Protein_2_power", "Protein_3_power",
              "Protein_4_power", "Protein_5_power",
              "Protein_6_power")]

  df <- cbind(df,df2)
  return(df)
}

final_data <- magic_function(final_data, 6)

set.seed(12345)
n = NROW(final_data)
id = sample(1:n, floor(n*0.5))
train = final_data[id,]
test = final_data[-id,]

# model building
M_1 <- lm(data = train, Moisture~Protein)
M_2 <- lm(data = train, Moisture~Protein+Protein_2_power)
M_3 <- lm(data = train, Moisture~Protein+Protein_2_power+Protein_3_power)
M_4 <- lm(data = train, Moisture~Protein+Protein_2_power+Protein_3_power+
          Protein_4_power)
M_5 <- lm(data = train, Moisture~Protein+Protein_2_power+Protein_3_power+
          Protein_4_power+Protein_5_power)
M_6 <- lm(data = train, Moisture~Protein+Protein_2_power+Protein_3_power+
          Protein_4_power+Protein_5_power+Protein_6_power)

train$type <- "train"
test$type <- "test"

final_data <- rbind(test, train)

# predicting new values
M_1_predicted <- predict(M_1, newdata = final_data)
M_2_predicted <- predict(M_2, newdata = final_data)
M_3_predicted <- predict(M_3, newdata = final_data)
M_4_predicted <- predict(M_4, newdata = final_data)
M_5_predicted <- predict(M_5, newdata = final_data)
M_6_predicted <- predict(M_6, newdata = final_data)

# calculating the MSE
```

```

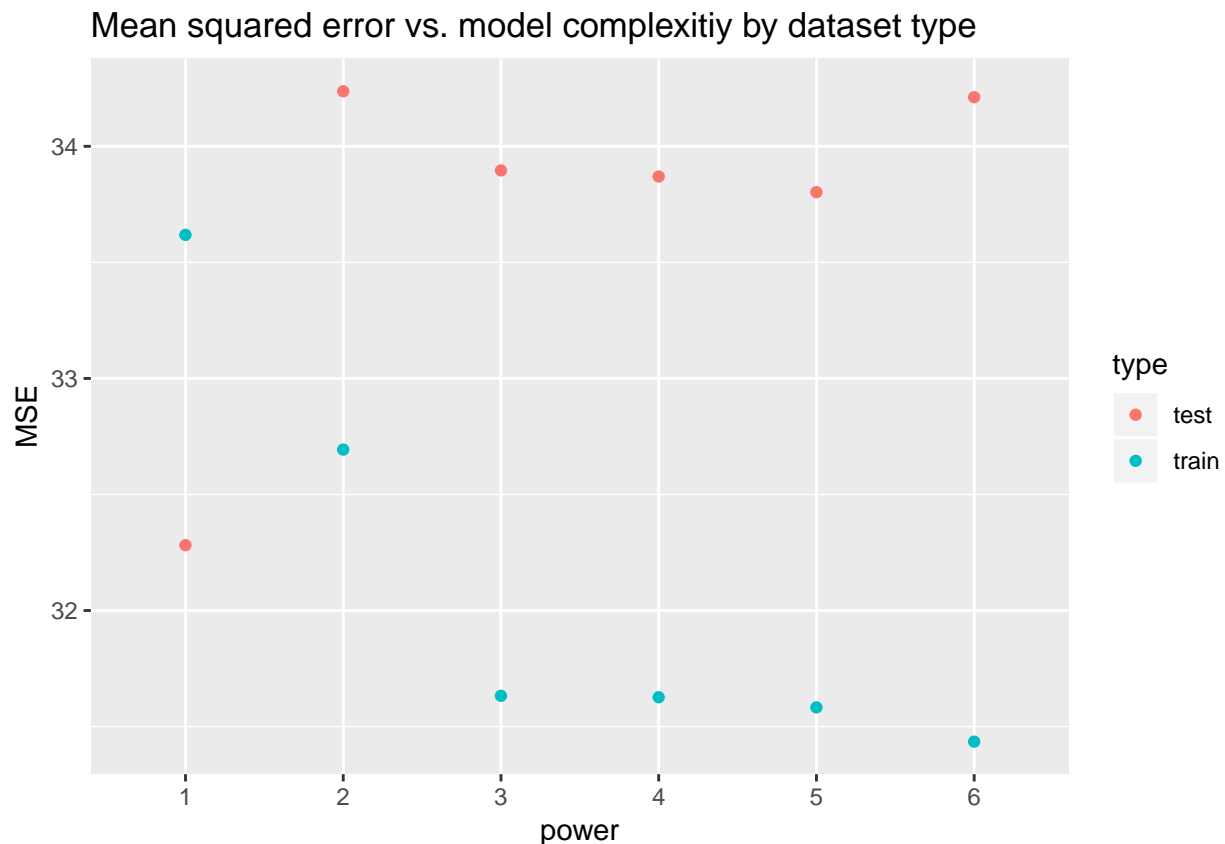
final_data$M_1_error <- (final_data$Moisture - M_1_predicted)^2
final_data$M_2_error <- (final_data$Moisture - M_2_predicted)^2
final_data$M_3_error <- (final_data$Moisture - M_3_predicted)^2
final_data$M_4_error <- (final_data$Moisture - M_4_predicted)^2
final_data$M_5_error <- (final_data$Moisture - M_5_predicted)^2
final_data$M_6_error <- (final_data$Moisture - M_6_predicted)^2

# Chaining like Chainsaw
final_error_data <- final_data %>% select(type, M_1_error, M_2_error, M_3_error,
                                          M_4_error, M_5_error, M_6_error) %>%

  gather(variable, value, -type) %>%
  separate(variable, c("model", "power", "error"), "_") %>%
  group_by(type, power) %>%
  summarise(MSE = mean(value, na.rm=TRUE))

ggplot(final_error_data, aes(x = power, y = MSE, color=type)) + geom_point() +
  ggtitle("Mean squared error vs. model complexitiy by dataset type")

```



Analysis: As evident from the plot above, we see that as we increase the model complexitiy (higher powers of the 'protein'), the training error reduces however the model becomes too biased towards the training set (overfits) and misses the test datasets prediction by larger margins in higher powers.

The best model is M1, that is Moisture~Protein as evident from the least test error (MSE).

The above is a classical case of bias-variance trade-off, which is as follows, as one makes the model fit the training dataset better the model becomes more biased and its ability to handle variation to new dataset decreases(variance), thus one should also maintain a good trade off between these two.

3.4 Perform variable selection of a linear model in which Fat is response and Channel1:Channel100 are predicted by using stepAIC.

```
min.model1 = lm(Fat ~ 1, data=tecator_data[,-1])
biggest1 <- formula(lm(Fat ~., data=tecator_data[,-1]))

step.model1 <- stepAIC(min.model1, direction = 'forward', scope=biggest1, trace = FALSE)
summ(step.model1)
```

```
## MODEL INFO:
## Observations: 215
## Dependent Variable: Fat
## Type: OLS linear regression
##
## MODEL FIT:
## F(29,185) = 4775.35, p = 0.00
## R2 = 1.00
## Adj. R2 = 1.00
##
## Standard errors: OLS
##
```

	Est.	S.E.	t val.	p
## (Intercept)	93.46	1.59	58.86	0.00 ***
## Moisture	-1.03	0.02	-54.25	0.00 ***
## Protein	-0.64	0.06	-10.91	0.00 ***
## Channel100	66.56	48.18	1.38	0.17
## Channel141	-3268.11	826.92	-3.95	0.00 ***
## Channel17	-64.03	20.80	-3.08	0.00 **
## Channel148	-2022.46	254.46	-7.95	0.00 ***
## Channel142	4934.22	1124.96	4.39	0.00 ***
## Channel150	1239.52	236.09	5.25	0.00 ***
## Channel145	4796.22	783.38	6.12	0.00 ***
## Channel166	2435.79	1169.85	2.08	0.04 *
## Channel156	2373.00	540.06	4.39	0.00 ***
## Channel190	-258.27	247.22	-1.04	0.30
## Channel160	-264.27	708.11	-0.37	0.71
## Channel170	14.25	327.12	0.04	0.97
## Channel167	-2015.92	543.74	-3.71	0.00 ***
## Channel159	635.71	996.31	0.64	0.52
## Channel165	-941.61	1009.23	-0.93	0.35
## Channel158	1054.24	927.95	1.14	0.26
## Channel144	-5733.84	1079.19	-5.31	0.00 ***
## Channel118	299.80	88.43	3.39	0.00 ***
## Channel178	2371.11	361.25	6.56	0.00 ***
## Channel184	-428.99	338.35	-1.27	0.21
## Channel162	3062.97	769.59	3.98	0.00 ***
## Channel153	-804.39	203.44	-3.95	0.00 ***
## Channel175	-1461.42	402.26	-3.63	0.00 ***
## Channel157	-3266.79	876.71	-3.73	0.00 ***
## Channel163	-2844.66	906.40	-3.14	0.00 **
## Channel124	-308.71	97.87	-3.15	0.00 **
## Channel137	401.64	151.76	2.65	0.01 **

Analysis: 29 variables were choose out of 107. Even among these there are many which have very low p values thus statistically it is a practice to remove variables which are above 0.0005 p values, thus the true

variables may not even include these many.

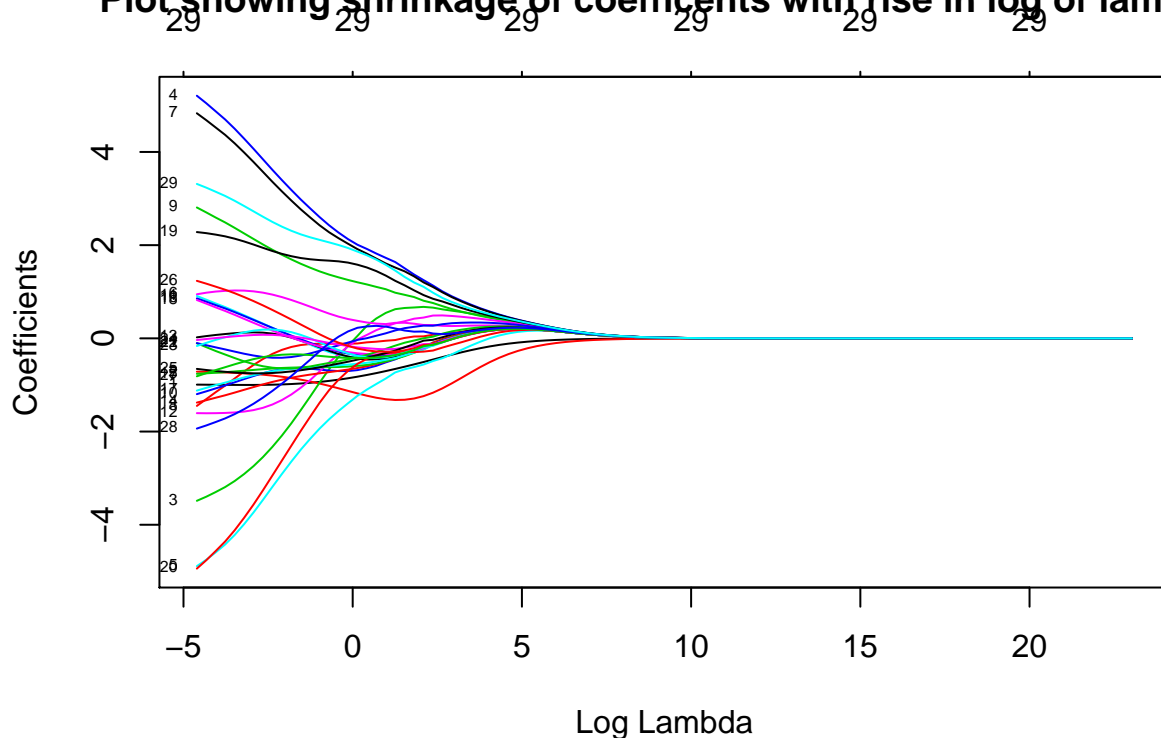
3.5 Fit a Ridge regression model with the same predictor and response

```
y <- tecator_data %>% select(Fat) %>% data.matrix()
x <- tecator_data %>% select(Moisture, Protein, Channel100, Channel141,
Channel17, Channel148, Channel142, Channel150, Channel145,
Channel166, Channel156, Channel190, Channel160, Channel170,
Channel167, Channel159, Channel165, Channel158, Channel144,
Channel118, Channel178, Channel184, Channel162, Channel153,
Channel175, Channel157, Channel163, Channel124, Channel137) %>% data.matrix()

lambda <- 10^seq(10, -2, length = 100)

ridge_fit <- glmnet(x, y, alpha = 0, family = "gaussian", lambda = lambda)
plot(ridge_fit, xvar = "lambda", label = TRUE,
     main = "Plot showing shrinkage of coefficients with rise in log of lambda")
```

Plot showing shrinkage of coefficients with rise in log of lambda



```
## Change of coefficient with respect to lambda
result <- NULL
for(i in lambda){
  temp <- t(coef(ridge_fit, i)) %>% as.matrix()
  temp <- cbind(temp, lambda = i)
  result <- rbind(temp, result)
}
```

```

result <- result %>% as.data.frame() %>% arrange(lambda)

table_cofe <- head(result, 10) %>% select(Channel18, Channel78, Channel84, Channel62,
                                          Channel53, Channel75, Channel57, Protein,
                                          lambda)

knitr::kable(table_cofe, caption = "Coefficient of Ridge Regression vs. Lambda")

```

Table 1: Coefficient of Ridge Regression vs. Lambda

Channel18	Channel78	Channel84	Channel62	Channel53	Channel75	Channel57	Protein	lambda
-4.941301	-0.1051478	-0.1017920	-0.1633868	-0.0352441	-0.6553389	1.2347381	-0.7149075	0.0100000
-4.761083	-0.1887872	-0.1473275	-0.0964082	-0.0142679	-0.6793011	1.1737952	-0.7194685	0.0132194
-4.563045	-0.2715621	-0.1935728	-0.0326085	0.0062755	-0.7006308	1.1089341	-0.7250500	0.0174753
-4.358319	-0.3472078	-0.2375663	0.0238972	0.0250101	-0.7180721	1.0436040	-0.7316380	0.0231013
-4.118947	-0.4230293	-0.2843872	0.0796662	0.0435796	-0.7335720	0.9689262	-0.7400267	0.0305386
-3.852036	-0.4931377	-0.3295873	0.1285556	0.0597302	-0.7449765	0.8871034	-0.7505390	0.0403702
-3.566312	-0.5524528	-0.3687106	0.1654546	0.0713550	-0.7508913	0.8003197	-0.7635225	0.0533670
-3.260550	-0.5987880	-0.3988579	0.1874355	0.0767648	-0.7506024	0.7073567	-0.7796364	0.0705480
-2.942882	-0.6299926	-0.4157692	0.1901664	0.0744131	-0.7433256	0.6093713	-0.7994381	0.0932603
-2.623836	-0.6453441	-0.4166940	0.1718246	0.0638187	-0.7293919	0.5084256	-0.8233650	0.1232847

3.6 Fit a Lasso regression model with the same predictor and response

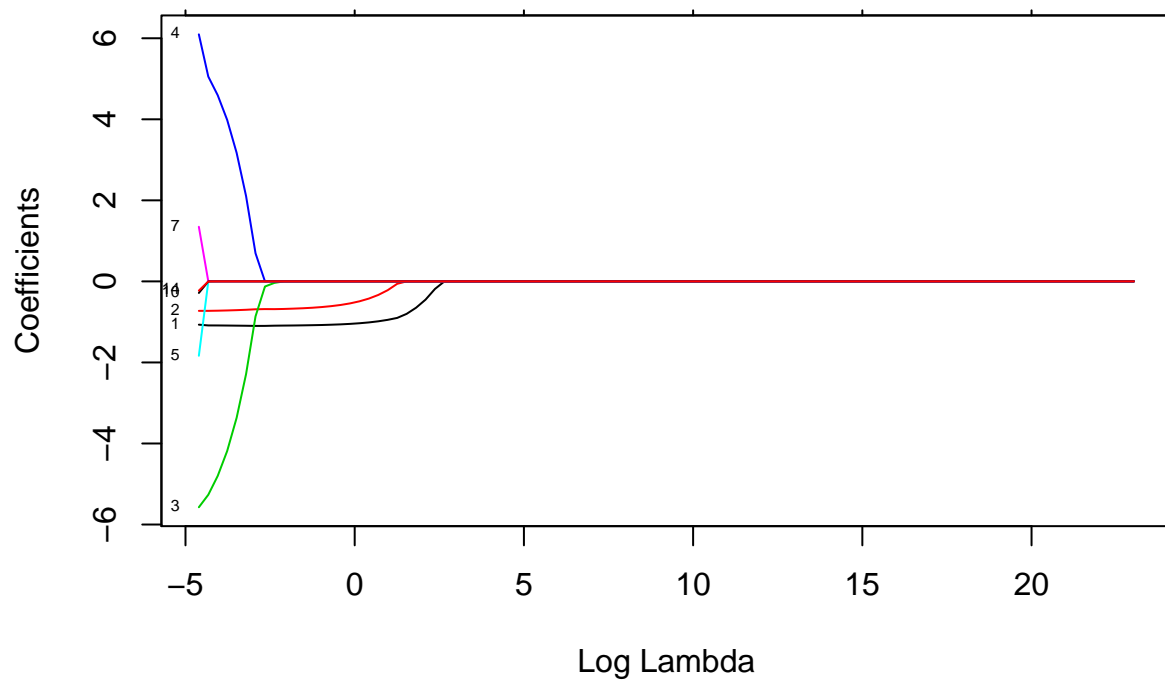
```

lambda <- 10^seq(10, -2, length = 100)

lasso_fit <- glmnet(x, y, alpha = 1, family = "gaussian", lambda = lambda)
plot(lasso_fit, xvar = "lambda", label = TRUE,
     main = "Plot showing shrinkage of coefficients with rise in log of lambda")

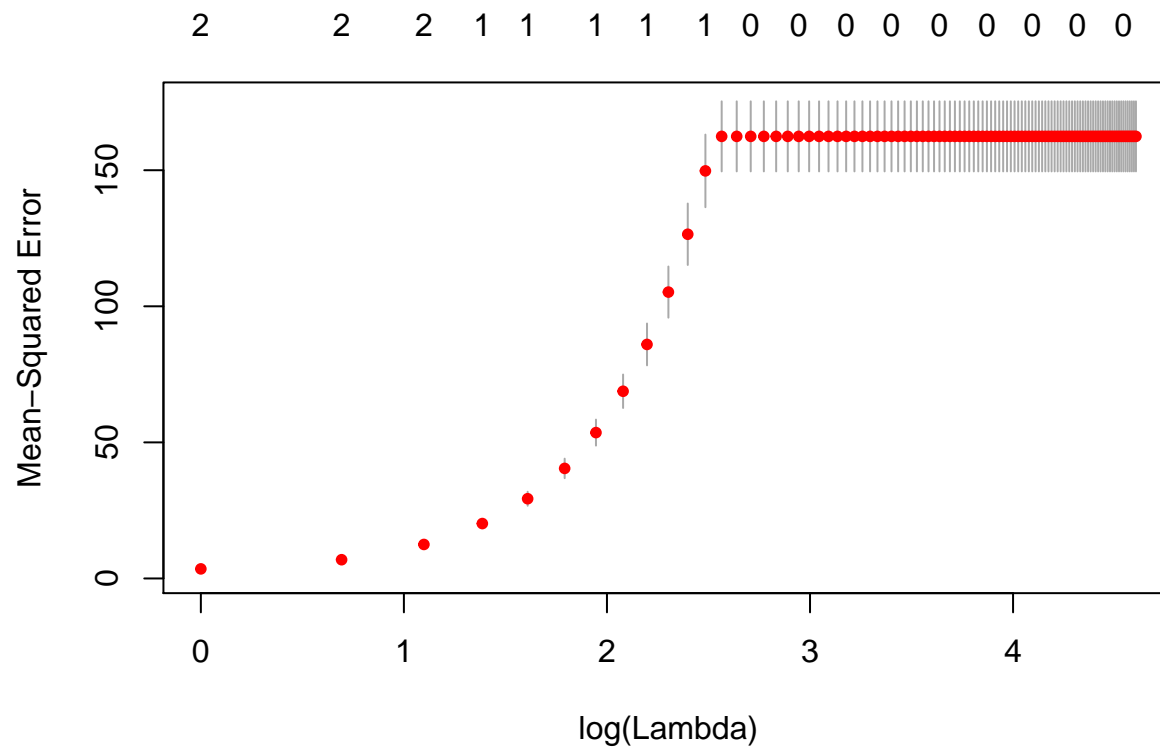
```

Plot showing shrinkage of coefficients with rise in log of lambda



3.7 Use cross-validation to find the optimal LASSO model, report the optimal and how many variables were chosen by the model and make conclusions

```
# #find the best lambda from our list via cross-validation
lambda_lasso <- seq(from = 0, to = 100, length.out = 101)
lasso_cv <- cv.glmnet(x,y, alpha=1, lambda = lambda_lasso, type.measure="mse")
plot(lasso_cv)
```



```
coef(lasso_cv, lambda = lasso_cv$lambda.min)
```

```
## 30 x 1 sparse Matrix of class "dgCMatrix"
##           1
## (Intercept) 95.33217672
## Moisture    -0.98182405
## Protein     -0.70566057
## Channel100  -4.63745993
## Channel41    6.80729763
## Channel17   -7.07697243
## Channel48    1.49163201
## Channel42    4.00513173
## Channel50   -4.89759364
## Channel45    3.36522225
## Channel66   -3.94883424
## Channel56    2.30364658
## Channel190  -0.96872375
## Channel60    0.05963838
## Channel70   -1.46513454
## Channel67    0.02934877
## Channel59    1.83426475
## Channel65   -1.65666952
## Channel58    2.02691211
## Channel44    2.51530613
## Channel118  -5.62061309
## Channel78    0.83812663
```

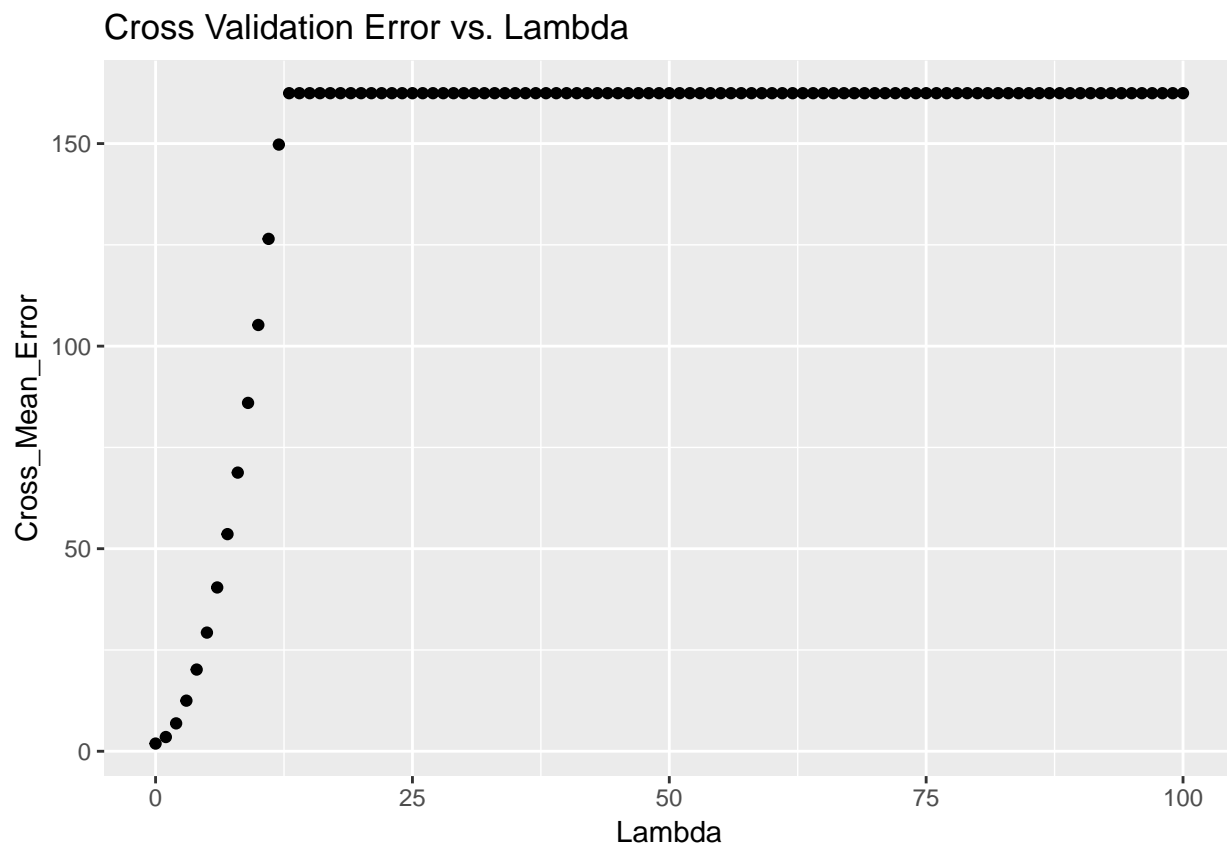
```
## Channel84      0.83643283
## Channel62     -0.96191077
## Channel53      0.52683349
## Channel75     -0.81286413
## Channel57      1.94836541
## Channel63     -1.88267554
## Channel24      0.49535686
## Channel37      3.18804687
```

```
lowest_lambda = lasso_cv$lambda.min
```

```
## Change of coefficient with respect to lambda
result_lasso <- NULL
for(i in 1:length(lambda_lasso)){
  temp <- lasso_cv$cvm[i] %>% as.matrix()
  temp <- cbind(CVM_error = temp, lambda = lasso_cv$lambda[i])
  result_lasso <- rbind(temp, result_lasso)
}
```

```
result_lasso <- result_lasso %>% as.data.frame() %>% arrange(lambda)
colnames(result_lasso) <- c("Cross_Mean_Error", "Lambda")
```

```
ggplot(result_lasso, aes(x = Lambda, y = Cross_Mean_Error)) + geom_point() + ggtitle("Cross Validation Error vs. Lambda")
```



Analysis: The minimum value of lambda was 0, implies zero penalisation. The variables selected are: Moisture, Protein, Channel 100, Channel41 along with intercept.

Appendix

```
knitr::opts_chunk$set(echo = TRUE)
if (!require("pacman")) install.packages("pacman")
pacman::p_load(xlsx, glmnet, MASS, jtools, huxtable, ggplot2,
               ggthemes, gridExtra, ROCR, broom, caret, e1071,
               kknn, tidyr, dplyr, reshape2, glmnet)

options("jtools-digits" = 2, scipen = 999)

spam_data <- read.xlsx("spambase.xlsx", sheetName = "spambase_data")
spam_data$Spam <- as.factor(spam_data$Spam)

tecator_data <- read.xlsx("tecator.xlsx", sheetName = "data")
tecator_data <- tecator_data[,2:NCOL(tecator_data)] # removing sample column
set.seed(12345)

n = NROW(spam_data)
id = sample(1:n, floor(n*0.5))
train = spam_data[id,]
test = spam_data[-id,]
best_model <- glm(formula = Spam ~., family = binomial, data = train)
summary(best_model)
# prediction
train$prediction_prob <- predict(best_model, newdata = train, type = "response")
test$prediction_prob <- predict(best_model, newdata = test, type = "response")

train$prediction_class_50 <- ifelse(train$prediction_prob > 0.50, 1, 0)
test$prediction_class_50 <- ifelse(test$prediction_prob > 0.50, 1, 0)

train$prediction_class_90 <- ifelse(train$prediction_prob > 0.90, 1, 0)
test$prediction_class_90 <- ifelse(test$prediction_prob > 0.90, 1, 0)
# plots
ggplot(train, aes(prediction_prob, color = Spam)) +
  geom_density(size = 1) + ggtitle("Training Set's Predicted Score for 50% cutoff") +
  scale_color_economist(name = "data", labels = c("negative", "positive")) +
  theme_economist()

ggplot(test, aes(prediction_prob, color = Spam)) +
  geom_density(size = 1) + ggtitle("Test Set's Predicted Score for 50% cutoff") +
  scale_color_economist(name = "data", labels = c("negative", "positive")) +
  theme_economist()

#confusion table
conf_train <- table(train$Spam, train$prediction_class_50)
names(dimnames(conf_train)) <- c("Actual Train", "Predicted Train")
confusionMatrix(conf_train)

conf_test <- table(test$Spam, test$prediction_class_50)
names(dimnames(conf_test)) <- c("Actual Test", "Predicted Test")
confusionMatrix(conf_test)
```

```

#confusion table
conf_train1 <- table(train$Spam, train$prediction_class_90)
names(dimnames(conf_train1)) <- c("Actual Train", "Predicted Train")
conf_train1

conf_test1 <- table(test$Spam, test$prediction_class_90)
names(dimnames(conf_test1)) <- c("Actual Test", "Predicted Test")
conf_test1

cutoffs <- seq(from = 0.05, to = 0.95, by = 0.05)
accuracy <- NULL

for (i in seq_along(cutoffs)){
  prediction <- ifelse(test$prediction_prob >= cutoffs[i], 1, 0) #Predicting for cut-off

  accuracy <- c(accuracy,length(which(test$Spam == prediction))/length(prediction)*100)}

cutoff_data <- as.data.frame(cbind(cutoffs, accuracy))

ggplot(data = cutoff_data, aes(x = cutoffs, y = accuracy)) +
  geom_line() +
  ggtitle("Cutoff vs. Accuracy for Test Dataset")

knn_model30 <- train.kknn(Spam ~., data = train, kmax = 30)

train$knn_prediction_class <- predict(knn_model30, train)
test$knn_prediction_class <- predict(knn_model30, test)

conf_train2 <- table(train$Spam, train$knn_prediction_class)
names(dimnames(conf_train2)) <- c("Actual Train", "Predicted Train")
confusionMatrix(conf_train2)

conf_test2 <- table(test$Spam, test$knn_prediction_class)
names(dimnames(conf_test2)) <- c("Actual Test", "Predicted Test")
confusionMatrix(conf_test2)
knn_model1 <- train.kknn(Spam ~., data = train, kmax = 1)

train$knn_prediction_class <- predict(knn_model1, train)
test$knn_prediction_class <- predict(knn_model1, test)

conf_train2 <- table(train$Spam, train$knn_prediction_class)
names(dimnames(conf_train2)) <- c("Actual Train", "Predicted Train")
confusionMatrix(conf_train2)

conf_test2 <- table(test$Spam, test$knn_prediction_class)
names(dimnames(conf_test2)) <- c("Actual Test", "Predicted Test")
confusionMatrix(conf_test2)

subset_function <- function(X,Y,N){

# X = swiss[,2:6]
# Y = swiss[,1:1]

```

```

# N = 5

df <- cbind(X,Y)
temp <- NULL
final <- NULL

for(i in 1:NCOL(X)){
  combs <- as.data.frame(gtools::combinations(NCOL(X), r=i, v=colnames(X), repeats.allowed=FALSE))
  combs <- tidyr::unite(combs, "formula", sep = ",")
  temp <- rbind(combs, temp)
}

set.seed(12345)
df2 <- df[sample(nrow(df)),]
df2$k_fold <- sample(N, size = nrow(df), replace = TRUE)

result <- NULL

for (j in 1:NROW(temp))
{
  for(i in 1:N){

train = df2[df2$k_fold != i,]
test = df2[df2$k_fold == i,]

vec <- temp[j,]
train_trimmed = lapply(strsplit(as.character(vec), ","), function(x) train[x])[[1]]
test_trimmed = lapply(strsplit(as.character(vec), ","), function(x) test[x])[[1]]

y_train = train[,c("Y"), drop = FALSE]
y_test = test[,c("Y"), drop = FALSE]

train_trimmed = as.matrix(train_trimmed)
test_trimmed = as.matrix(test_trimmed)
y_test = as.matrix(y_test)
y_train = as.matrix(y_train)

t_train = as.matrix(t(train_trimmed))
t_test = as.matrix(t(test_trimmed))

betas = solve(t_train %*% train_trimmed) %*% (t_train %*% y_train)

train_trimmed = as.data.frame(train_trimmed)
test_trimmed = as.data.frame(test_trimmed)

train_trimmed$type = "train"
test_trimmed$type = "test"
final <- rbind(train_trimmed, test_trimmed)

y_hat_val = as.matrix(final[,1:(ncol(final)-1)]) %*% betas
mse = (Y - y_hat_val)^ 2

```

```

data <- cbind(i, vec, mse, type = final$type)
result <- rbind(data, result)

}
}

result <- as.data.frame(result)

colnames(result) <- c("kfold", "variables", "mse", "type")

result$mse <- as.numeric(result$mse)
result$no_variables <- nchar(as.character(result$variables)) - nchar(gsub('\\\\', '\\', result$variables))

variable_performance <- result %>% group_by(kfold, no_variables) %>%
  summarise(MSE = mean(mse, na.rm = TRUE))

myplot <- ggplot(data = variable_performance, aes(x = no_variables, y = MSE, color=kfold)) +
  geom_line() + ggtitle("Plot of MSE vs. Number of variables by folds")

myplot2 <- ggplot(data = result, aes(x = variables, y = mse, color=kfold)) +
  geom_bar(stat="identity") + ggtitle("Plot of MSE vs. Features by folds") + coord_flip()

best_variables <- result %>% group_by(type, variables) %>% filter(type == "test") %>%
  summarise(MSE = mean(mse, na.rm = TRUE)) %>% arrange(MSE) %>% select(variables) %>% slice(1) %>% as.v

return(list(myplot, myplot2, best_variables))
}

subset_function(X = swiss[,2:6], Y = swiss[,1:1], N = 5)

# Using pairs on swiss dataset
GGally::ggpairs(swiss)
ggplot(data = tecator_data, aes(x = Protein, y = Moisture)) +
  geom_point() +
  geom_smooth( method = 'loess') +
  ggtitle("Plot of Moisture vs. Protein")
ggplot(data = tecator_data, aes(x = Moisture)) +
  geom_density() +
  ggtitle("Density Plot of Moisture")

final_data <- tecator_data

magic_function <- function(df, N)
{
  df2 <- df
  for(i in 2:N)
  {
    df2[paste("Protein_", i, "_power", sep="")] <- (df2$Protein)^i
  }

  df2 <- df2[c("Protein_2_power", "Protein_3_power",
              "Protein_4_power", "Protein_5_power",
              "Protein_6_power")]

```

```

df <- cbind(df,df2)
return(df)
}

final_data <- magic_function(final_data, 6)

set.seed(12345)
n = NROW(final_data)
id = sample(1:n, floor(n*0.5))
train = final_data[id,]
test = final_data[-id,]

# model building
M_1 <- lm(data = train, Moisture~Protein)
M_2 <- lm(data = train, Moisture~Protein+Protein_2_power)
M_3 <- lm(data = train, Moisture~Protein+Protein_2_power+Protein_3_power)
M_4 <- lm(data = train, Moisture~Protein+Protein_2_power+Protein_3_power+
  Protein_4_power)
M_5 <- lm(data = train, Moisture~Protein+Protein_2_power+Protein_3_power+
  Protein_4_power+Protein_5_power)
M_6 <- lm(data = train, Moisture~Protein+Protein_2_power+Protein_3_power+
  Protein_4_power+Protein_5_power+Protein_6_power)

train$type <- "train"
test$type <- "test"

final_data <- rbind(test, train)

# predicting new values
M_1_predicted <- predict(M_1, newdata = final_data)
M_2_predicted <- predict(M_2, newdata = final_data)
M_3_predicted <- predict(M_3, newdata = final_data)
M_4_predicted <- predict(M_4, newdata = final_data)
M_5_predicted <- predict(M_5, newdata = final_data)
M_6_predicted <- predict(M_6, newdata = final_data)

# calculating the MSE
final_data$M_1_error <- (final_data$Moisture - M_1_predicted)^2
final_data$M_2_error <- (final_data$Moisture - M_2_predicted)^2
final_data$M_3_error <- (final_data$Moisture - M_3_predicted)^2
final_data$M_4_error <- (final_data$Moisture - M_4_predicted)^2
final_data$M_5_error <- (final_data$Moisture - M_5_predicted)^2
final_data$M_6_error <- (final_data$Moisture - M_6_predicted)^2

# Chaining like Chainsaw
final_error_data <- final_data %>% select(type, M_1_error, M_2_error, M_3_error,
  M_4_error, M_5_error, M_6_error) %>%
  gather(variable, value, -type) %>%
  separate(variable, c("model", "power", "error"), "_") %>%
  group_by(type, power) %>%
  summarise(MSE = mean(value, na.rm=TRUE))

ggplot(final_error_data, aes(x = power, y = MSE, color=type)) + geom_point() +

```

```

ggtitle("Mean squared error vs. model complexitiy by dataset type")

min.model1 = lm(Fat ~ 1, data=tecator_data[,-1])
biggest1 <- formula(lm(Fat ~., data=tecator_data[,-1]))

step.model1 <- stepAIC(min.model1, direction = 'forward', scope=biggest1, trace = FALSE)
summ(step.model1)
y <- tecator_data %>% select(Fat) %>% data.matrix()
x <- tecator_data %>% select(Moisture, Protein, Channel100, Channel141,
Channel17, Channel48, Channel42, Channel50, Channel45,
Channel66, Channel56, Channel90, Channel60, Channel70,
Channel67, Channel59, Channel65, Channel58, Channel44,
Channel18, Channel78, Channel84, Channel62, Channel53,
Channel75, Channel57, Channel63, Channel24, Channel37) %>% data.matrix()

lambda <- 10^seq(10, -2, length = 100)

ridge_fit <- glmnet(x, y, alpha = 0, family = "gaussian", lambda = lambda)
plot(ridge_fit, xvar = "lambda", label = TRUE,
     main = "Plot showing shrinkage of coefficients with rise in log of lambda")

## Change of coefficient with respect to lambda
result <- NULL
for(i in lambda){
temp <- t(coef(ridge_fit, i)) %>% as.matrix()
temp <- cbind(temp, lambda = i)
result <- rbind(temp, result)
}
result <- result %>% as.data.frame() %>% arrange(lambda)
table_cofe <- head(result, 10) %>% select(Channel18, Channel78, Channel84, Channel62,
                                         Channel53, Channel75, Channel57, Protein,
                                         lambda)

knitr::kable(table_cofe, caption = "Coefficient of Ridge Regression vs. Lambda")
lambda <- 10^seq(10, -2, length = 100)

lasso_fit <- glmnet(x, y, alpha = 1, family = "gaussian", lambda = lambda)
plot(lasso_fit, xvar = "lambda", label = TRUE,
     main = "Plot showing shrinkage of coefficients with rise in log of lambda")

# #find the best lambda from our list via cross-validation
lambda_lasso <- seq(from = 0, to = 100, length.out = 101)
lasso_cv <- cv.glmnet(x,y, alpha=1, lambda = lambda_lasso, type.measure="mse")
plot(lasso_cv)
coef(lasso_cv, lambda = lasso_cv$lambda.min)

lowest_lambda = lasso_cv$lambda.min

## Change of coefficient with respect to lambda
result_lasso <- NULL
for(i in 1:length(lambda_lasso)){
temp <- lasso_cv$cvm[i] %>% as.matrix()

```

```

temp <- cbind(CVM_error = temp, lambda = lasso_cv$lambda[i])
result_lasso <- rbind(temp, result_lasso)
}
result_lasso <- result_lasso %>% as.data.frame() %>% arrange(lambda)
colnames(result_lasso) <- c("Cross_Mean_Error", "Lambda")

ggplot(result_lasso, aes(x = Lambda, y = Cross_Mean_Error)) + geom_point() + ggtitle("Cross Validation L")

```