MID TERM ASSIGNMENT

Submitted by: **Anwai Archit**

Subject: Statistical Learning and Data Mining (91940)

Challenge: Regression Problem using Parametric and Non Parametric Approach

Importing the Dataset

1. Installing the libraries and loading them in the challenge.
2. Importing Training *(train_ch.csv)* & Testing Dataset *(test_ch.csv)*.
3. Studying the multivariate datasets of 1000 observations.

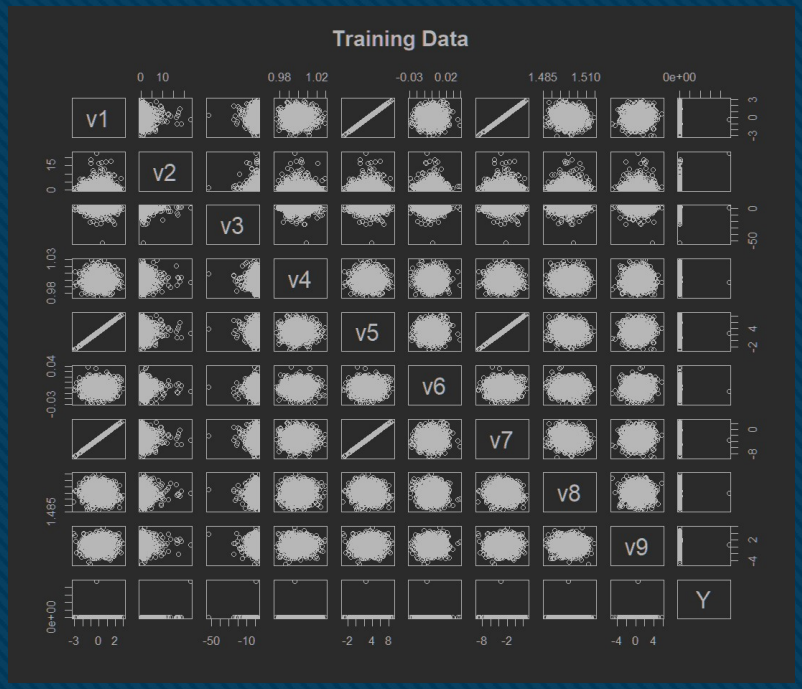

Fig.1

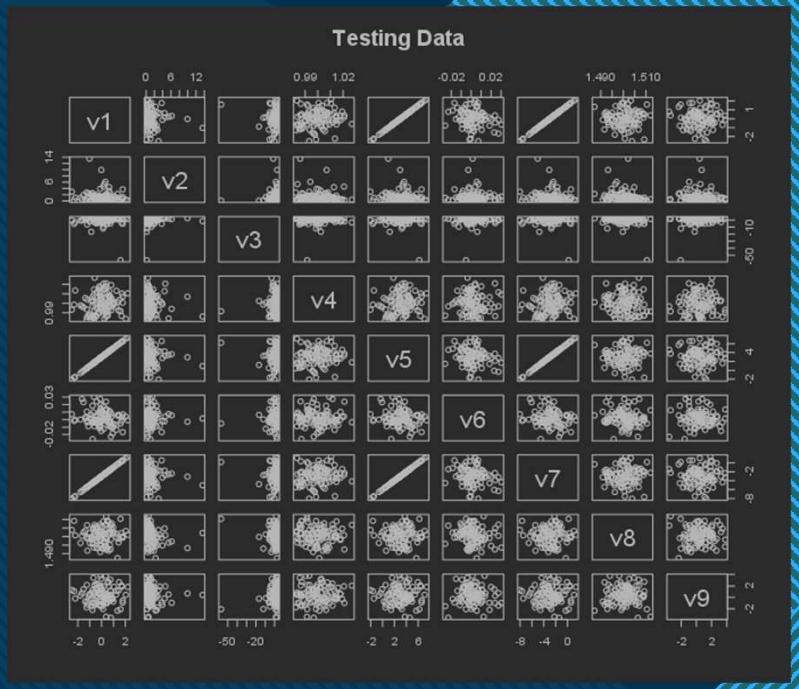

Fig. 2

Preprocessing the Dataset

Removing the Outliers to Improve the Result

Normalizing the Datasets and Establishing the Target Column

Plotting the Normalized Datasets

Checking Correlations of Variables

Feeding the Processed Datasets to the Models

Preprocessed Data

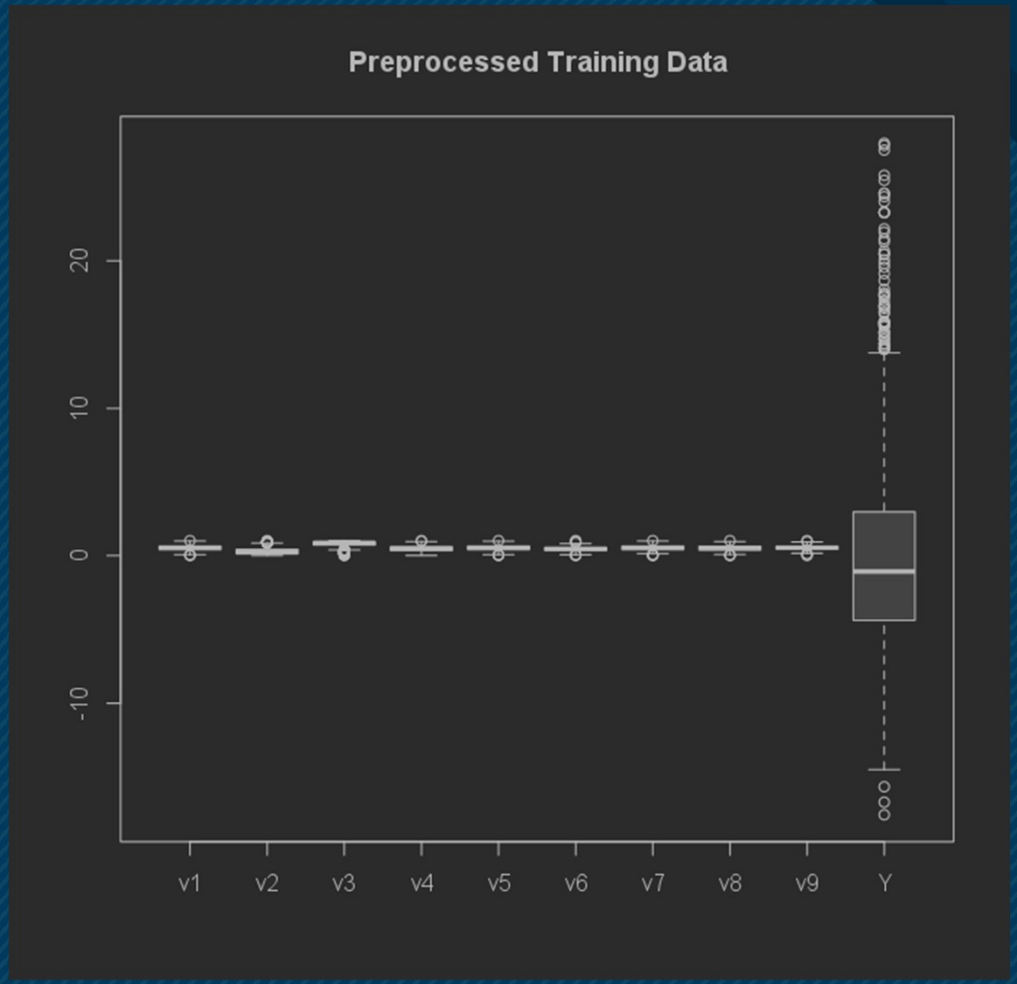

Fig. 3

```r
# Building our own `normalize()` function
normalize <- function(x) {
  num <- x - min(x)
  denom <- max(x) - min(x)
  return (num/denom) ^normalize
}
```

Fig. 4

- Removing the Outliers and Normalizing the Dataset reduced the observations to “843”.

Linear Regression

```
fit <- lm(formula = Y ~ v1*v2*v3*v4*v5*v6*v7*v8*v9, data = data_norm_train)
```

1. Fitting the Linear Regression model using all the parameters on the normalized training dataset.
2. Predicting the fitted model on the training (*lmPredict*) and testing (*lm_pred*) dataset.
3. Finding the RMSE for the Linear Regression on the training dataset.
4. Checking the Homoscedasticity for the fitted model.

```
lmPredict <- predict(fit, newdata = data_norm_train)
lm_pred <- predict(fit, newdata = data_norm_test)
```

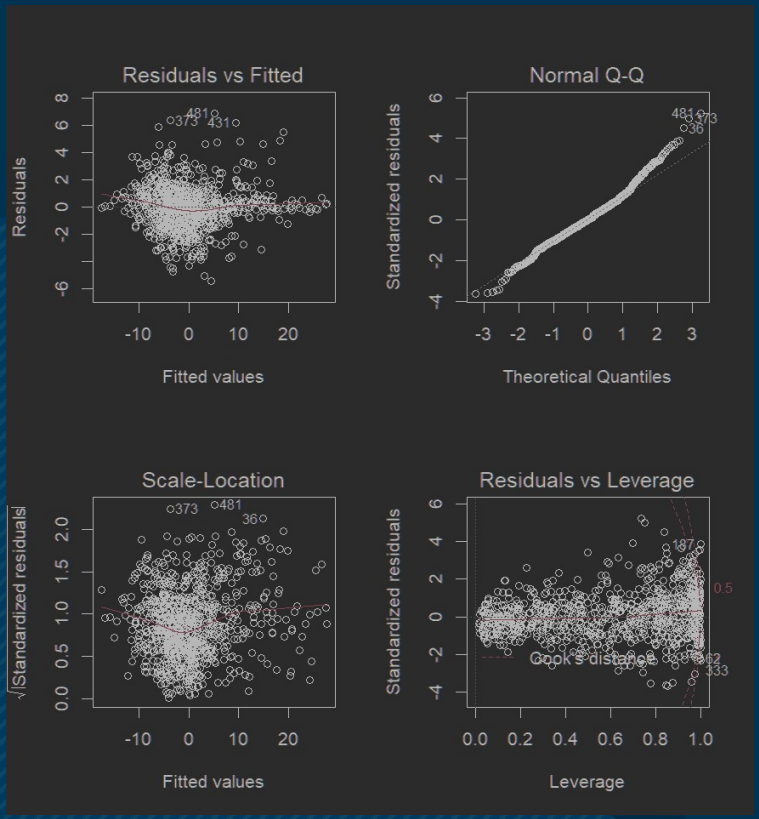

Fig. 5

```
Selection Algorithm: forward
         v1  v2  v3  v4  v5  v6  v7  v8  v9
1  ( 1 ) " " "*" " " " " " " " " " " " " " "
2  ( 1 ) "*" "*" " " " " " " " " " " " " " "
3  ( 1 ) "*" "*" "*" " " " " " " " " " " " "
```

Fig. 6

KNN

```
ctrl <- trainControl(method="repeatedcv", number = 10, repeats = 10)
knnFit <- train(Y ~ v1+v2+v3, data = data_norm_train, method = "knn", trControl = ctrl)
```

1. Applying the 10-fold Cross Validation Method.
2. Fitting the KNN model using all the parameters on the normalized training dataset.
3. Selecting the best features using Forward Feature Selection, procuring v1; v2 & v3 as the optimal combination.
4. Predicting the fitted model on the training (*knnPredict*) and testing (*knn_pred*) dataset.
5. Finding the RMSE for the KNN on the training dataset.

```
knnPredict <- predict(knnFit, newdata = data_norm_train)
knn_pred <- predict(knnFit, newdata = data_norm_test)
```


MAIA
THANK YOU!