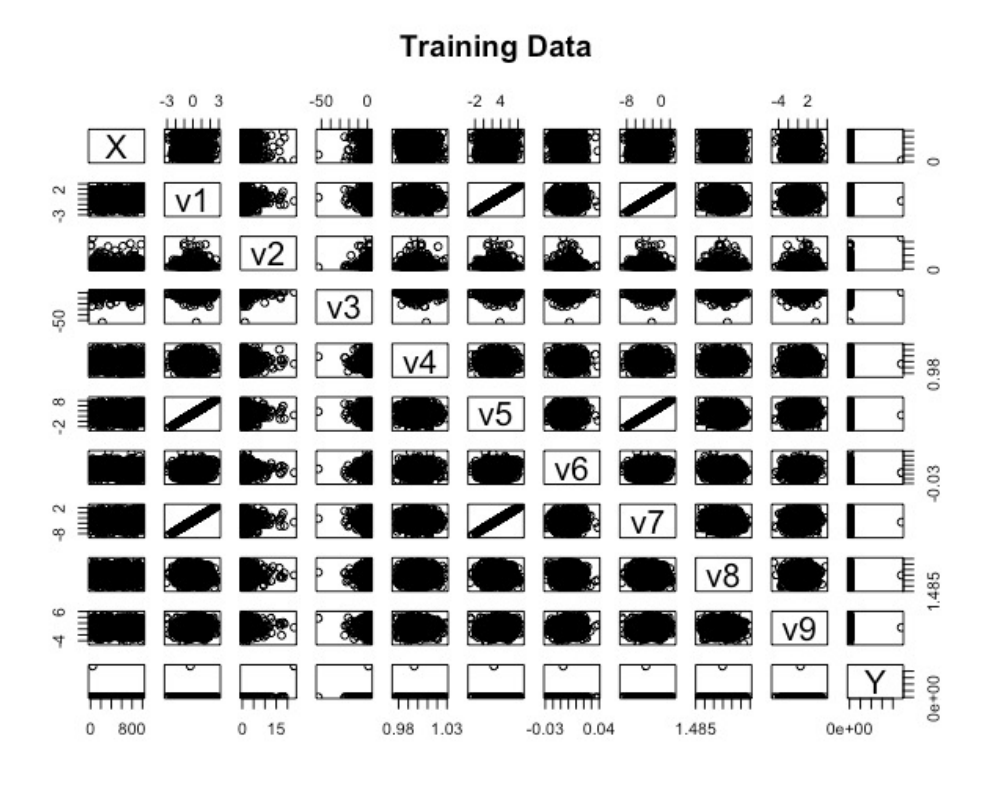
MID TERM ASSIGNMENT

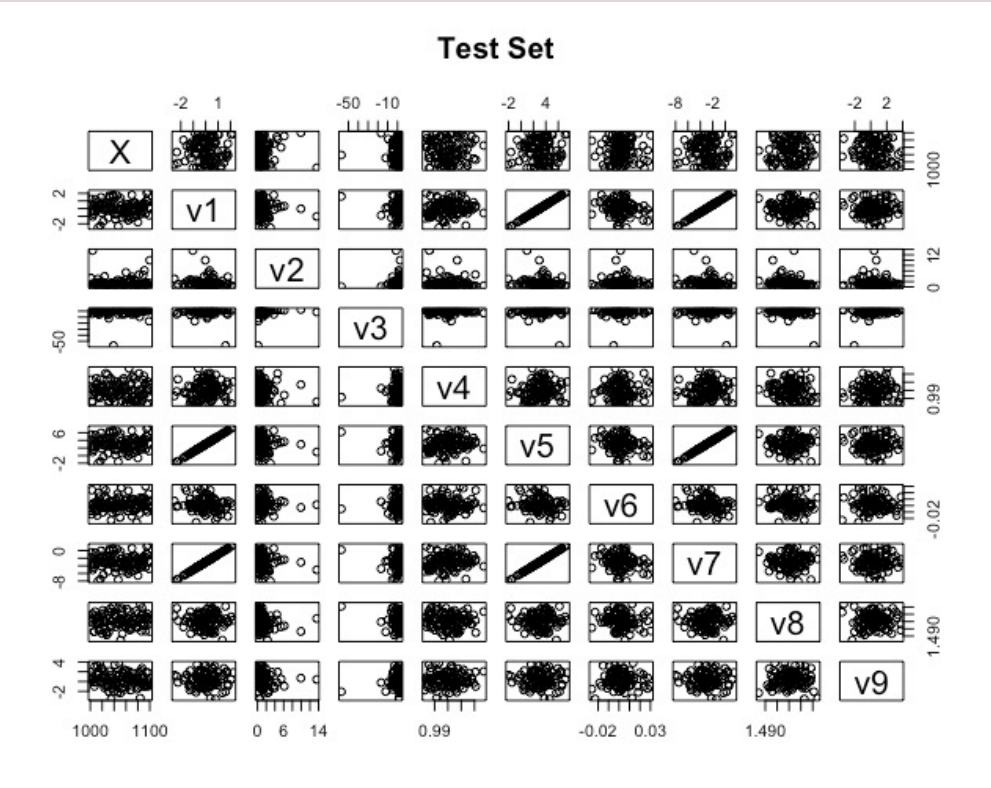
Statistical Learning and Data Mining (91940)

Submitted By:
Sheikh Adilina

Importing the Dataset

- The “read.csv” function was used to import both the csv files.
- The train dataset contained 9 columns for features and 1 column as the target
- The test dataset contained 9 columns for features

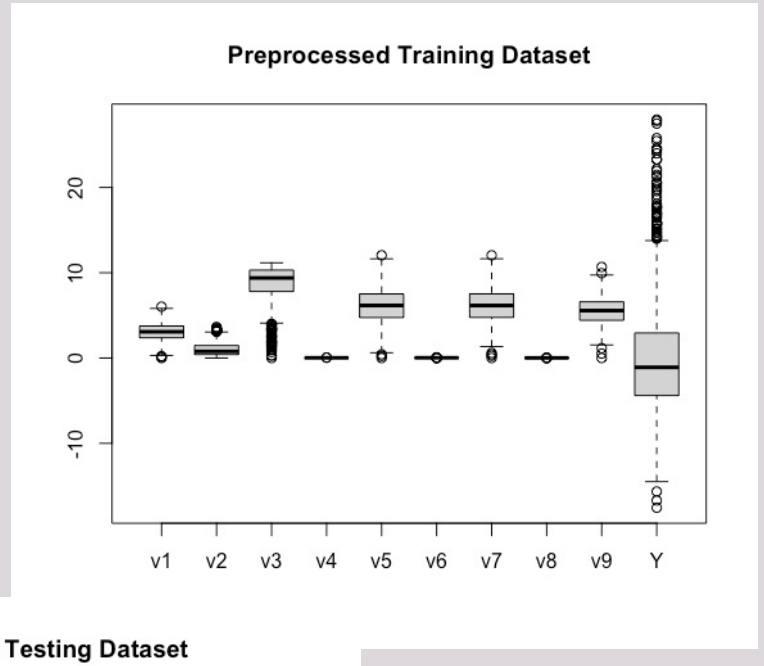


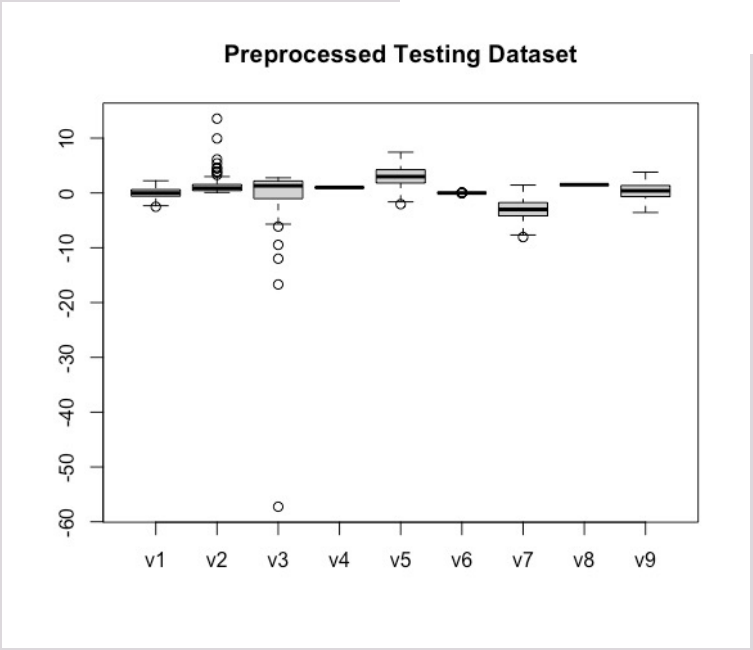


Preprocessing of Data

Some preprocessing were required to prepare the data before training the model.

- ✓ In the very beginning, the first column "X" was removed because it contained only the serial number of the instances.
- ✓ Secondly, the outliers were removed which reduced the number of instances from 1000 to 843.
- ✓ Thirdly, the whole data was centered and scaled using the Pre-processing transformation which is available in the caret library in R.





Linear Regression

Linear Regression creates a model by fitting a line to the training data points.

1. I gave "v1*v2*v3*v4*v5*v6*v7*v8*v8" as the parameter for feature set which means that the lm() function will evaluate performance on all possible combination of features and select the optimum feature set.

```
ctrl <- trainControl(method="cv", number = 10)
lm_fit <- lm(Y ~ v1*v2*v3*v4*v5*v6*v7*v8*v9, data = train_full,
             trControl = ctrl)
summary(lm_fit)
```

Fitting model using the optimum feature set

2. The model found out the best parameters using the 10 fold cross validation and fit the training data using those parameters.

3. The model's RMSE score and final predictions were obtained using the model obtained from the lm() function.

K Nearest Neighbors

K Nearest Neighbor is a non-parametric algorithm where it does not create a model to predict a new instance. It only looks at the nearest k data points to assign a label.

1. I applied the forward selection using the 10 fold cross validation to find out the best feature set with maximum performance on the training data set.

```
train.control <- trainControl(method = "cv", number = 10)
step.model <- train(Y ~., data = train_full,
                    method = "leapForward",
                    trControl = train.control)
```

2. The result showed that the best feature subset was "v1", "v2" and "v3".

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

3. I fitted the knn model using the optimum feature set. The model used 10 fold cross validation to obtain the best value for k and fitted the training dataset using that k value.

```
ctrl <- trainControl(method="cv", number = 10)
knn_fit <- train(Y ~ v1+v2+v3, data = train_full, method = "knn",
                 trControl = ctrl)
knn_fit
```

Fitting model using the optimum feature set

4. The model's RMSE score and final predictions were obtained using the model obtained from the train() function.

Thank you !