Regression Challenge

By

MUHAMMAD ZAIN AMIN

UNIVERSITY OF CASSINO AND SOUTHERN LAZIO

DATA ANALYSIS

The initial stage of the challenge involves examining the datasets. Here's a brief overview of the summarized outcomes:

- **Missing Values**:

◦ We first checked the missing values in both the train and test sets and didn’t find any missing values in either of the sets.

- **Predictors with Outliers: -**

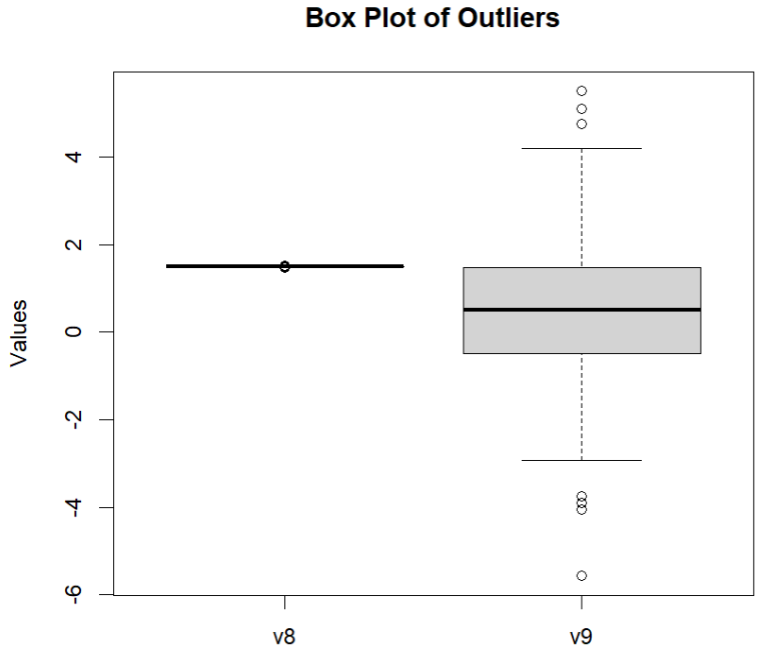
◦ We found outliers in the v8 and v9 predictors on specific rows in the training set.

◦ Rows with outliers: 1, 182, 277, 517, 568, 810, 898, 940.

- **Train and Test Sets Observations: -**

◦ Total number of observations in the training set -> 1000.

◦ Total number of observations in the testing set -> 100



Regression Framework

- **Data Preprocessing**

◦ Firstly, we removed the observation serial column from both the train and test datasets.
◦ After that we have separated the predictors from the labels in the training dataset.

- **Data Findings**

◦ **Correlation between Predictors: -**

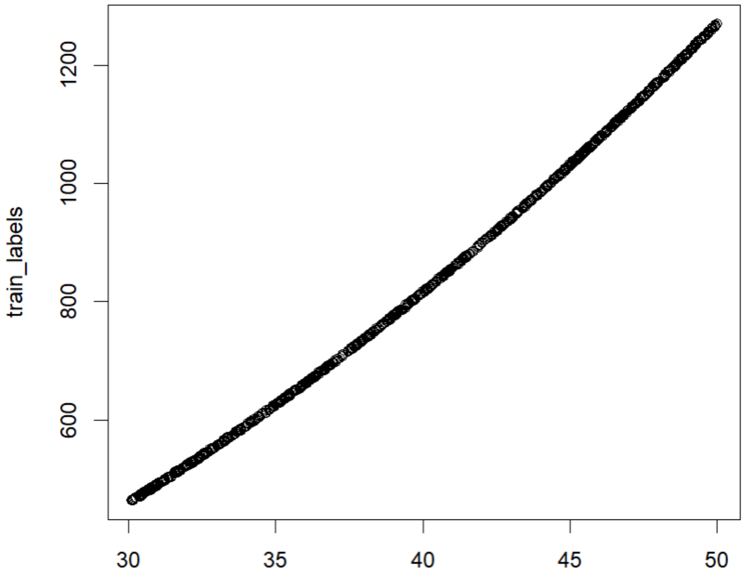
◦ The best single predictor found is v3. You can clearly see the scatter plot of v3 predictor.
◦ I also found out that the highly correlated predictor pairs are v5 and v7.

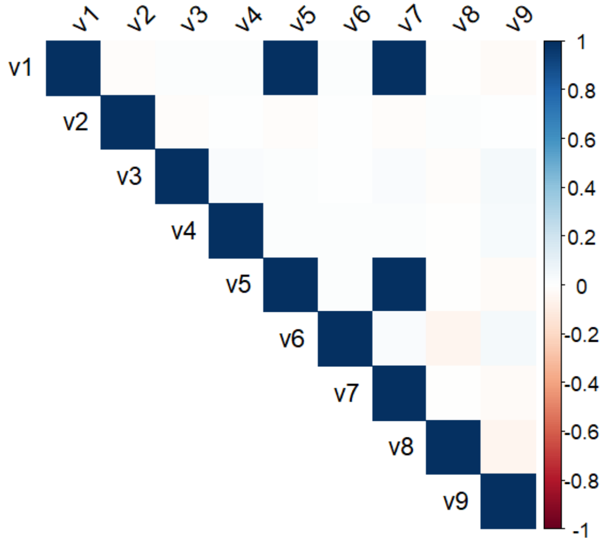
◦ **Assumption on Residuals**

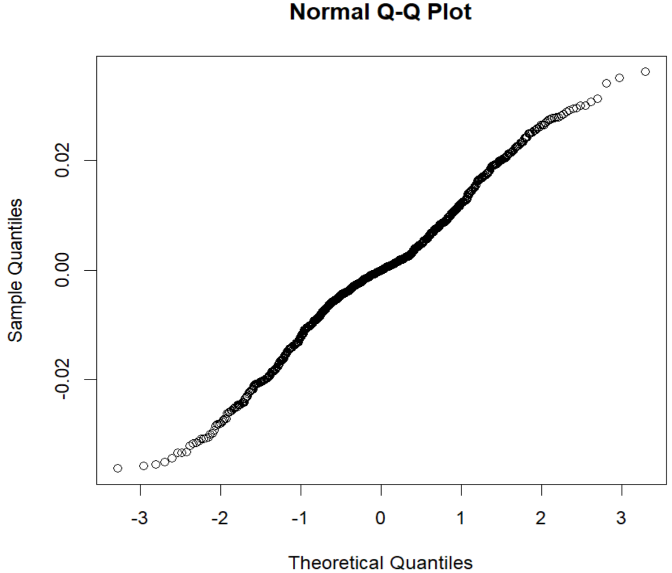
◦ Also, we checked the heteroscedasticity (varying spread of residuals) and normality of the residuals on the training dataset

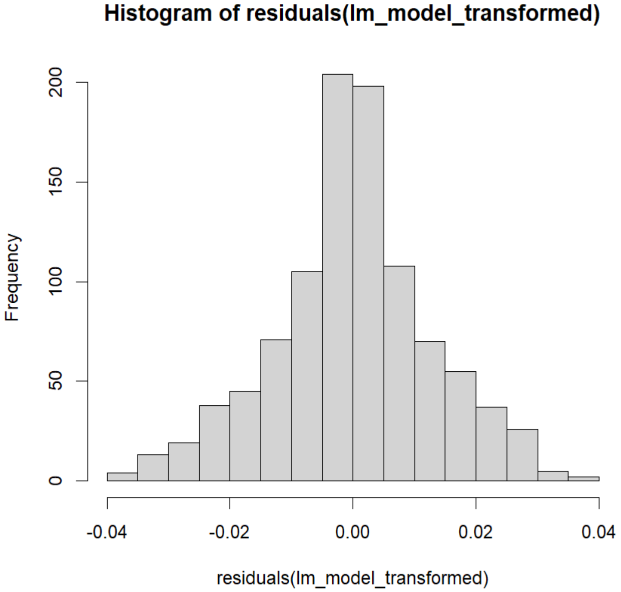
- **Linear Regression and KNN Regression Model**

◦ We will evaluate the performance of both the linear and KNN regression model based on multiple combinations of predictors and metrics.









TASK 1 Linear Regression

Let’s implement the linear regression model.

- I used the caret library “lm” function to implement the linear regression model.
- In linear regression, both the train and test datasets were used to train and test the “lm” model.
- I have tested the performance of the linear regression model with multiple series of predictors.
  - ◦ Results validated using the 100 observations from the test set.
  - ◦ Multiple metrices have been used to calculate the model performance.
  - ◦ The best model is used for predicting the test observations, provided in the test_ch.csv file.

Predictors Used	Residual Sum of Squares	Mean Squared Error	Root Mean Square Error	Mean Absolute Error	R-squared
v1,v2,v3,v4,v5,v6,v7,v8,v9	0.1631074	0.0001631074	0.01277135	0.009617715	1
v2,v3	1415.326	1.415326	1.189675	1.02709	0.9999745
v1, v2, v3, v4, v6, v8, v9	0.1637499	0.0001637499	0.01279648	0.009599386	1
v1, v2, v3, v4, v6, v8	0.1639203	0.0001639203	0.01280314	0.009593796	1
v1, v3	11.28107	0.01128107	0.1062124	0.08788478	0.9999998
v1, v2, v3, v4	4.394677	0.004394677	0.06629236	0.05624027	0.9999999
v1, v2, v3	4.398109	0.004398109	0.06631824	0.05625143	0.9999999
v3	1419.315	1.419315	1.19135	1.027492	0.9999745

Table 1. Linear Regression performance on different subsets of predictors

TASK 2 KNN Regression

Let’s implement the linear regression model.

- I used the FNN library “knn.reg” function to implement the kNN regression model.
- In KNN regression, the train and test datasets were used to train and test the “knn.reg” model.
- I have tested the performance of the KNN regression model with multiple series of predictors at different values of k.
  - ◦ Results validated using the 100 observations from the test set.
  - ◦ Multiple metrices have been used to calculate the model performance.
  - ◦ The best model is used for predicting the test observations, provided in the test_ch.csv file.

Predictors Used	K	Residual Sum of Squares	Mean Squared Error	Root Mean Square Error	Mean Absolute Error	R-squared
v1,v2,v3,v4,v5,v6,v7,v8,v9	9	46795.86	46.79586	6.84075	5.166068	0.9991718
v2,v3	4	3076.804	3.076804	1.754082	1.402978	0.9999447
v1, v2, v3, v4, v6, v8, v9	5	33677.3	33.6773	5.803214	4.305215	0.9993951
v1, v2, v3, v4, v6, v8	12	3979.905	3.979905	1.99497	1.479609	0.999929
v1, v2, v3, v4	11	3963.14	3.96314	1.990764	1.475916	0.9999294
v1, v2, v3	5	4518.19	4.51819	2.125603	1.672892	0.9999188
v3	5	2155.141	2.155141	1.46804	1.211451	0.9999612

Table 2. KNN Regression performance on different subsets of predictors

Best Performances of Linear and KNN Regression

Task 1 : Best Linear Regression Model

Predictors Used	Residual Sum of Squares	Mean Squared Error	Root Mean Square Error	Mean Absolute Error	R-squared
v1,v2,v3,v4,v5,v6,v7,v8,v9	0.1631074	0.0001631074	0.01277135	0.009617715	1

Task 2: Best KNN Regression Model

Predictors Used	K	Residual Sum of Squares	Mean Squared Error	Root Mean Square Error	Mean Absolute Error	R-squared
v3	5	2155.141	2.155141	1.46804	1.211451	0.9999612

Best Performances of Linear and KNN Regression

- It is clearly seen from the results that the best linear regression model achieved Residual Sum of Squares (RSS) of 0.1637499, Mean Squared Error (MAE) of 0.0001637499, Root Mean Square Error (RMSE) of 0.01279648, Mean Absolute Error (MAE) of 0.009599386, and R-squared score of 1.

- The best linear regression model, is also used to predict on the test_ch.csv observations.

- Also, the results that the best KNN regression model achieved showed good scores of Residual Sum of Squares (RSS) 2155.141, Mean Squared Error (MAE) of 2.155141, Root Mean Square Error (RMSE) of 1.46804, Mean Absolute Error (MAE) of 1.211451, and R-squared score of 0.9999612.

Thank
you