University of Essex

Department of Mathematical Sciences

Part 2: Comparative Study

Subject:

CE802 Machine Learning and Data Mining

Registration number:2003303

Supervisor: Dr Luca Citi

Date of submission (January 20 2021)

Word count: 850

Classification Study

Problem: A travel insurance company needs to anticipate if their clients will file a case or not later on. This forecast is upheld by the verifiable information of the customers given by the administrator of the organization. It is classification problem.

Data preparation: Two CSV files are provided. first file is for training and testing. Second file is for predicting classification. For understanding of data we generate descriptive statistic summary. Dataset has 15 features.

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11
count	1500.000000	1500.00000	1500.000000	1500.000000	1500.000000	1500.000000	1500.000000	1500.000000	1500.000000	1500.000000	1500.000000
mean	218.049333	14.90628	-0.829733	41.063440	133.485333	-1.592947	5.390220	3.854280	-7.308000	5.403333	-307.992000
std	234.669160	12.97563	5.090345	9.500727	71.416874	2.513850	7.662813	3.465276	7.159315	4.504907	120.565344
min	0.000000	0.00000	-17.560000	30.000000	54.040000	-6.070000	-19.800000	0.000000	-24.670000	0.000000	-609.750000
25%	30.000000	3.72000	-4.600000	32.160000	78.040000	-4.000000	-0.300000	0.600000	-14.670000	1.000000	-354.750000
50%	100.000000	6.21000	-2.810000	33.240000	98.040000	-0.610000	2.400000	1.020000	-3.670000	1.000000	-291.750000
75%	390.000000	24.60000	4.080000	49.350000	194.040000	0.350000	12.877500	7.050000	-1.670000	10.000000	-252.750000
max	780.000000	47.10000	8.300000	60.450000	299.040000	6.070000	18.810000	10.350000	1.330000	10.000000	35.250000

Figure 1(descriptive statistic table)

Generating correlation heatmap to examine connection between the features(as shown in Figure 2). All the data column are consistent excluding column F15.

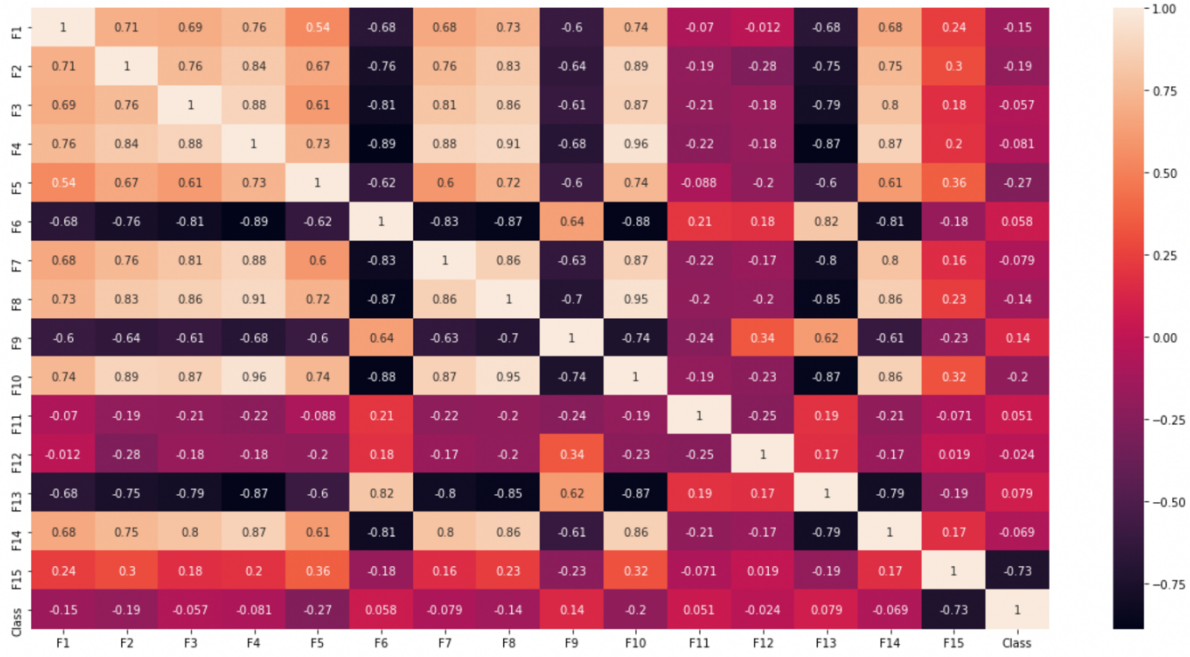

Figure 2(Corelation Heatmap)

Column F15 had 50% missing value but the value distribution was normal so replace missing values with mean of column F15 *(as shown in Figure 3)*.

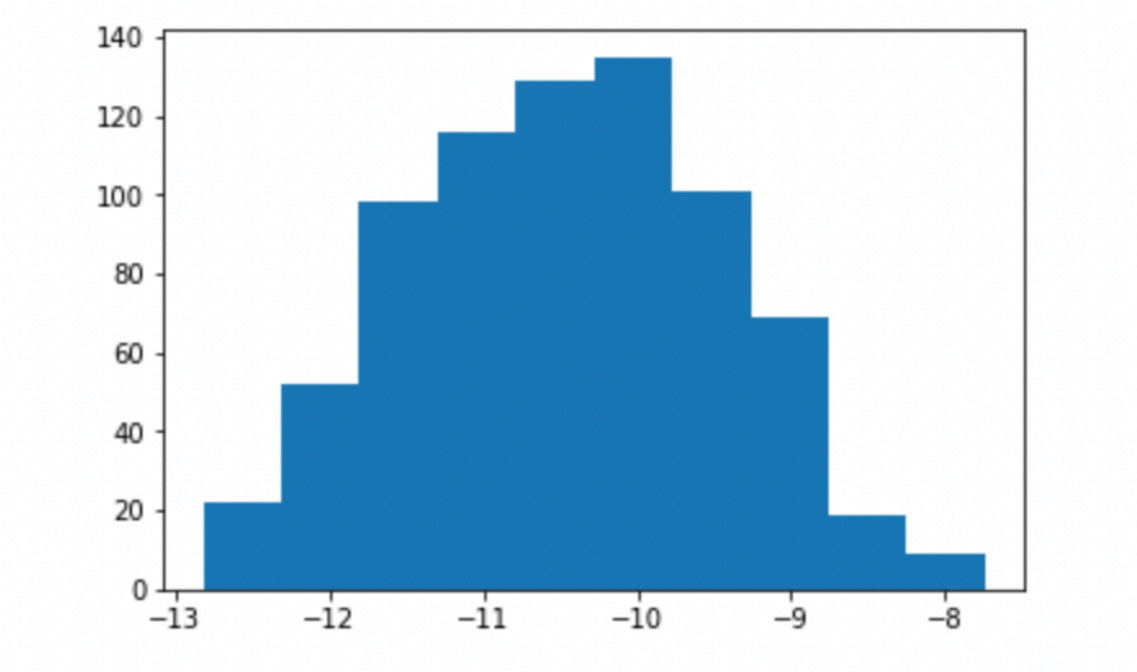

Figure 3(Normal Distribution)

Data normalization is part of data preparation. Data brought in range between 0 to 1 range while data has different ranges using min max scaler and split into training and testing dataset 75% and 25% respectively.

Building various ML models:

1. **Decision tree classifier**: Decision tree classifier is simplest and effective model to binary classification problem. For DTC criterion had set to 'gini' as we all know tree will partitioned the data recursively using greedy algorithm. Decision tree also help to handle missing values and useful for feature engineering. Confusion matrix for Decision tree classifier (*as Shown in Figure 4*).

   Precision and Recall table:

	Precision	Recall
False	77	76
True	71	72

   Confusion Matrix:

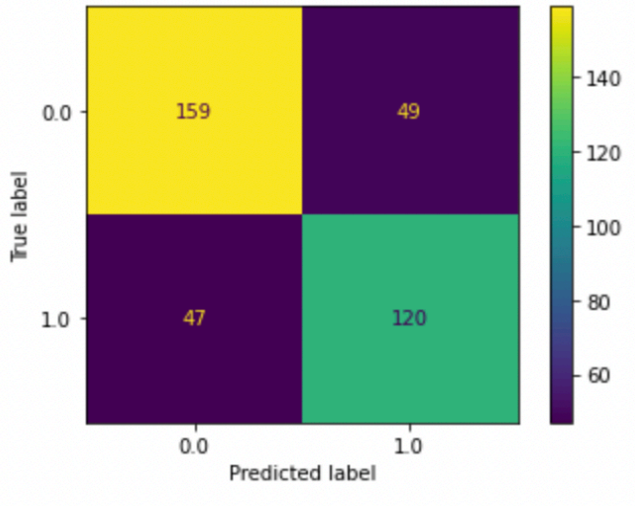

*Figure 4 (**Decision Tree Classifier**)*

2. **Random forest classifier:** Random forest is an intuitive model.it performed well because of its stochastic approach. It makes saveral trees and combine them to examine. It has better result than Decision tree. Confusion matrix for Random forest classifier (*as Shown in Figure 5*).

Precision and Recall table:

	Precision	Recall
False	75	88
True	80	64

Confusion Matrix:

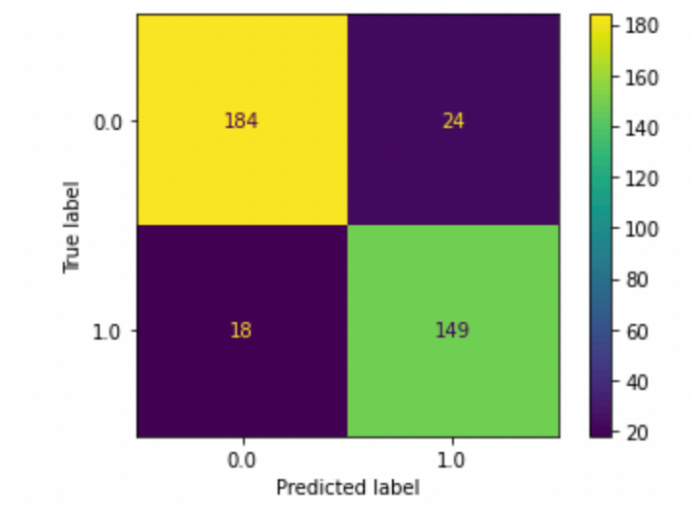

*Figure 5(**Random forest Classifier**)*

3. **Support Vector Machine (SVMs):** SVMs are majorly used for classification problem. In SVMs each data get plot on n-dimensional space with value of each feature then hyper plane classifies the data. Confusion matrix for Support Vector Machine (*as Shown in Figure 6*).

Precision and Recall table:

	Precision	Recall
False	75	88
True	80	64

Confusion Matrix:

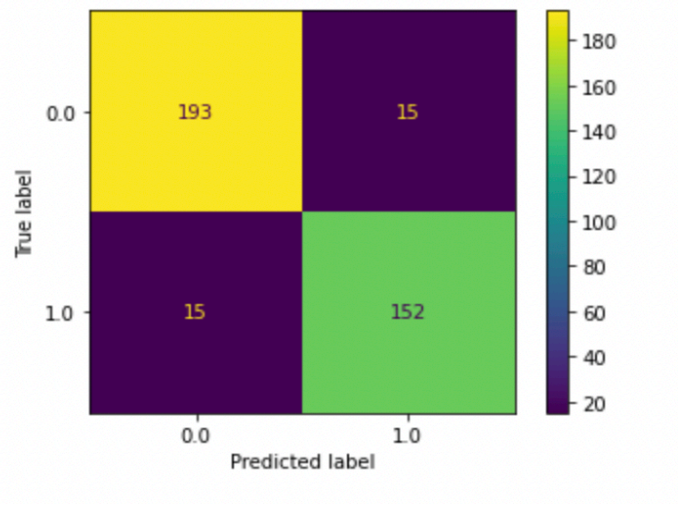

*Figure 6(**Support Vector Machine**)*

Evaluation: We had built three different classifier Machine learning model and same training and testing data used in all three classification models in which SVM performed better than random forest classifier and decision tree(*as shown in Figure 7*).SVM model was the robust and stable model for this classification problem.

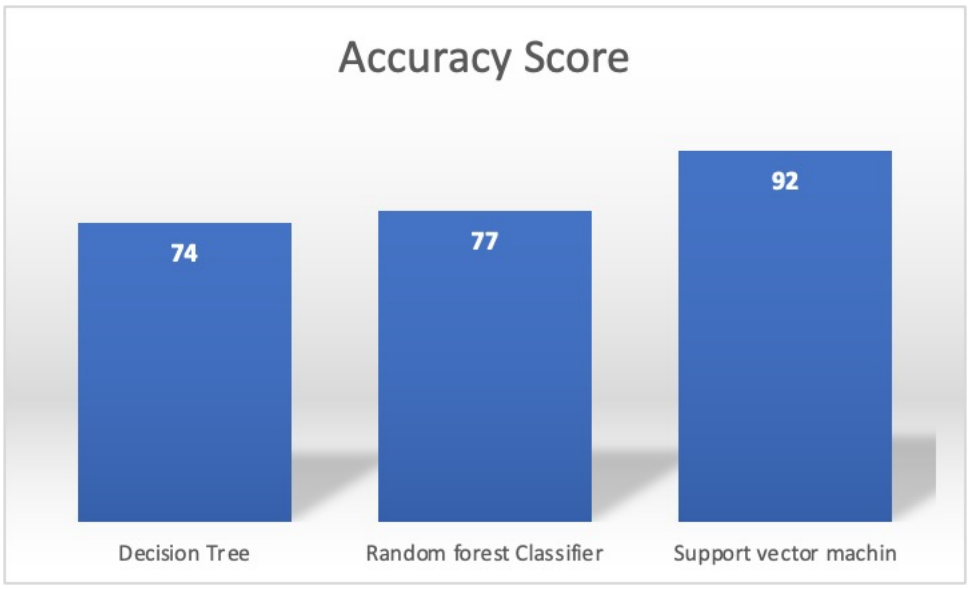

*Figure 7(**Accuracy score**)*

Regression Study

Problem: As we predicted whether customer will claim the insurance or now. The extension of previous problem we need to predict the continuous value of the claim. It is a regression problem.

Data preparation: Two CSV files are provided. first file is for training and testing. Second file is for predicting value of the claim. For understanding of data we generate descriptive statistic summary. Dataset has 16 features.

	F2	F3	F4	F5	F6	F7	F8	F9	F11	F12	F13
count	1500.00000	1500.000000	1500.000000	1500.000000	1500.000000	1500.000000	1500.000000	1500.000000	1500.000000	1500.000000	1500.000000
mean	-630.78434	-92.061967	-2311.234440	24115.645020	4.041660	60.448080	-6.962873	11.828000	127.205547	8.109187	-3.634847
std	904.26205	29.857614	889.404292	13329.106617	2.893909	58.613624	3.024250	5.430108	1548.308592	5.812853	3.040026
min	-3708.93000	-210.780000	-5503.920000	-35169.510000	0.040000	-144.460000	-16.390000	0.000000	0.000000	0.080000	-13.700000
25%	-1233.33000	-112.110000	-2926.252500	17982.390000	1.945000	21.720000	-9.080000	9.000000	0.260000	3.815000	-5.590000
50%	-641.53500	-92.015000	-2302.755000	24118.815000	3.415000	62.060000	-6.960000	12.000000	2.020000	6.740000	-3.530000
75%	-3.56250	-71.195000	-1703.347500	30824.220000	5.492500	99.740000	-4.840000	15.000000	13.610000	10.945000	-1.655000
max	2521.86000	13.160000	671.100000	85176.180000	23.390000	239.820000	2.060000	30.000000	54949.060000	45.140000	6.070000

Figure 8(descriptive statistic table)

Generating correlation heatmap to examine connection between the features(*as shown in Figure 9*). All the data column are consistent excluding column F11.

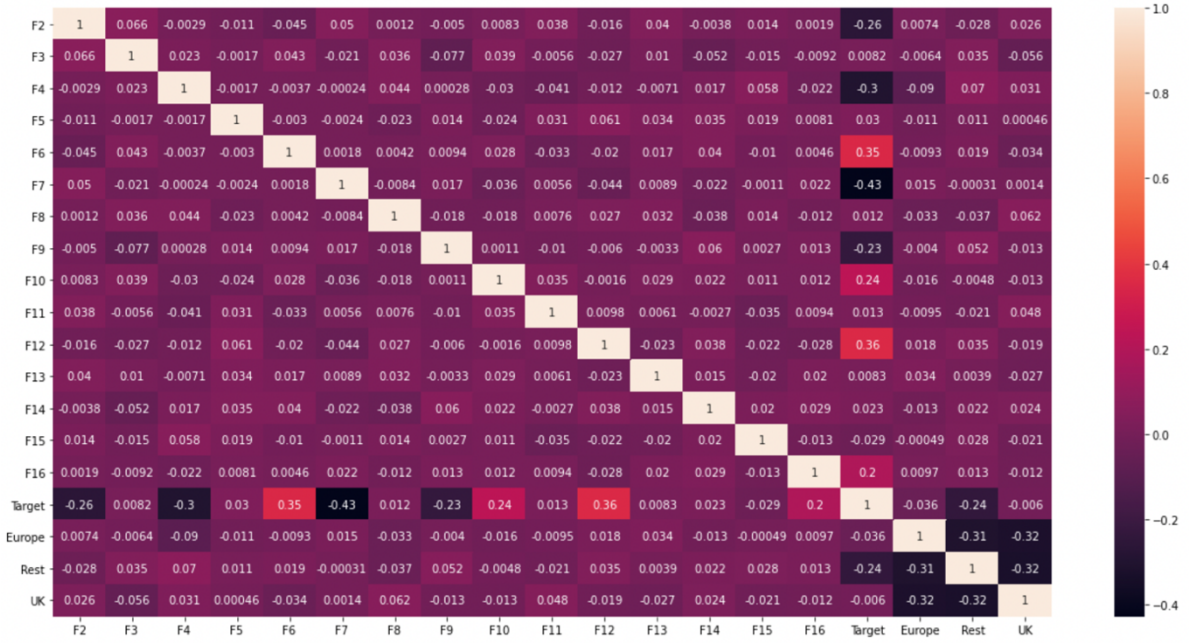

Figure 9(Corelation Heatmap)

Column F11 had outlier values. To handle outlier we replace outlier values by median value using 3 standard deviation*(as shown in Figure 10)*.

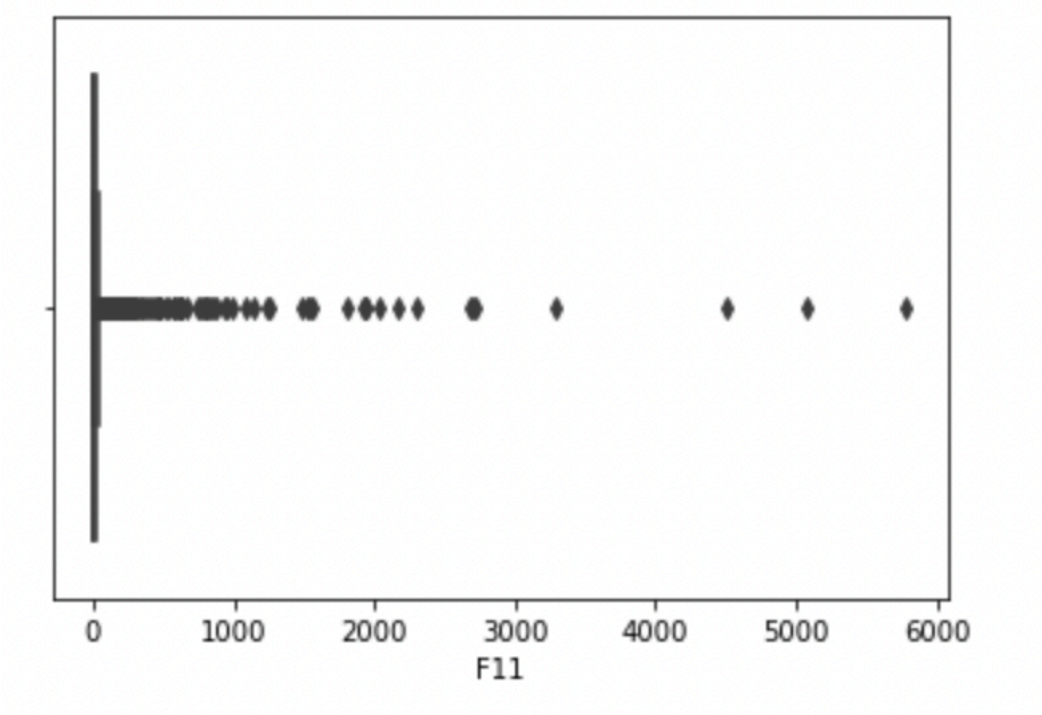

*Figure 10(**Box plot**)*

We had use may encoding techniques on dataset such as One-Hot encoding and Label Encoding on F1 and F10 respectively.

Building various ML models:

1. **Neural Network**: Artificial neural network is computational algorithm it consist of a large number of simple elements, called neurons or perceptron. We had use Sequential artificial neural network which has one input layer and 3 hidden layer. first and third hidden layer has each 50 neurons and second hidden layer has 25 hidden layer. Input layer has 18 neurons. For building ANN we had use keras library ,optimizer is 'Adamax' and 'relu' activation function. Trained data on ANN with 700 epochs.

Matrix Table:

Metrics	Score
Mean absolute error	78.24788
Mean squared error	23018.5321
Root mean square error	151.71

2. **Linear Regression**: Linear Regression is statistical method to find relationship between dependent and independent variable. Linear regression statistic model built using sklearn module without parameter.

Matrix Table:

Metrics	Score
Mean absolute error	393.1436
Mean squared error	249004.1134
Root mean square error	499.0031

3. **Random Forest Regressor**: Random Forrest is an ensemble machine learning technique of decision tree. Parameters of random forest are max_depth and n_estimators are set to 10 and 100 respectively. Refer Figure 11 of random forest cumulative graph.

Matrix Table:

Metrics	Score
Mean absolute error	449.1350
Mean squared error	371373.3450
Root mean square error	609.4040

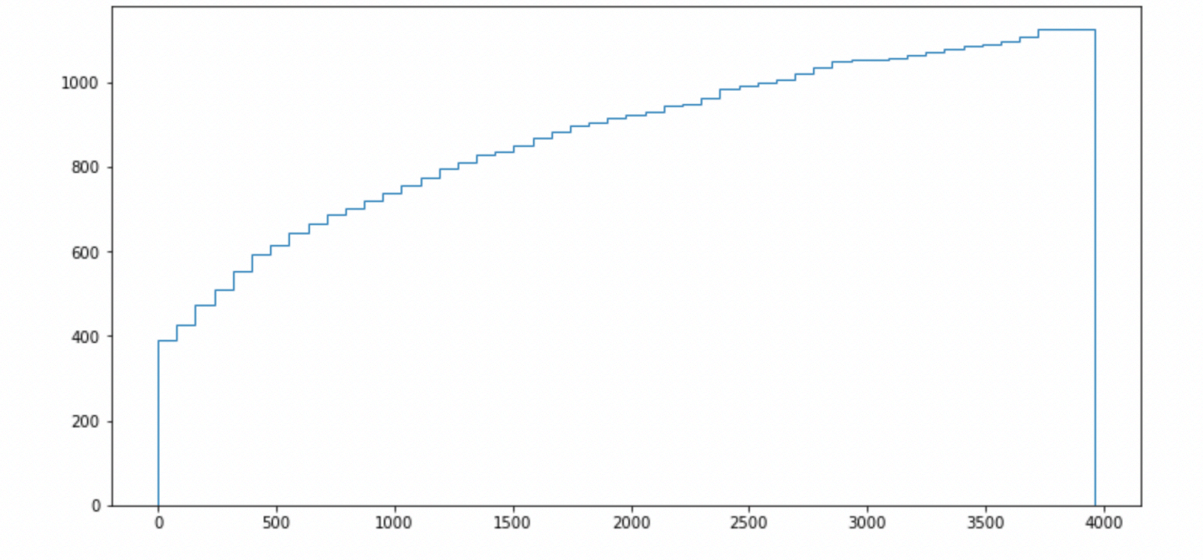

Figure 11(Random Forest Cumulative graph)

Evaluation: We had built three different regression Machine learning model and same training and testing data used in all three regression models in which Neural network performed better than random forest regressor and linear regression(*as shown in Figure 7*).Neural network model was the robust and has better R2 value among all models for this regression problem.

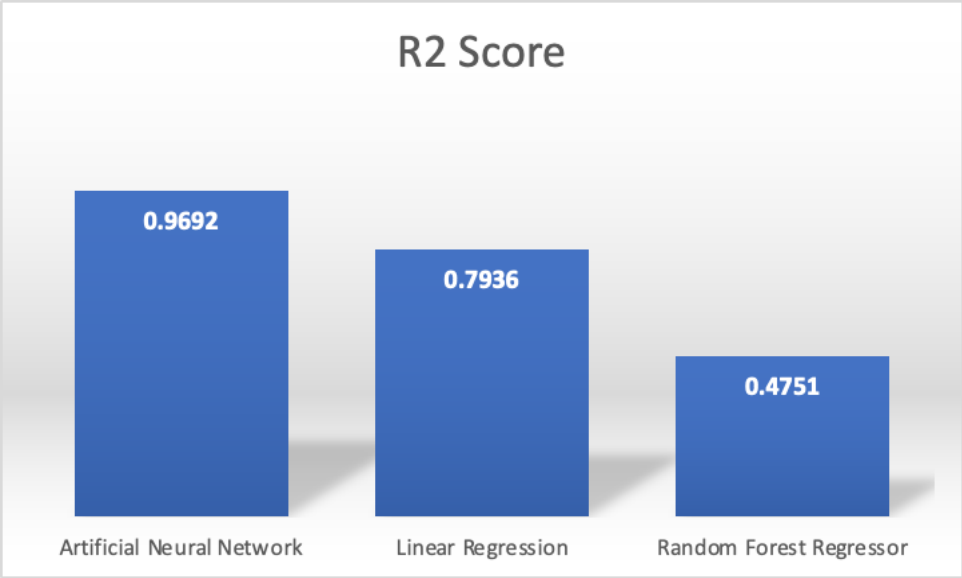

*Figure 12(**R2 Score** for each regression model)*