Retail & Marketing Analytics

Marketing Analytics Report

Kyriakos Theodorides

Imperial College London

April 2020

1 Introduction

Consistently accurate sales forecasts are gold. They deliver the revenue predictability that is essential for companies to accelerate their growth and success. It also allows marketing teams to see if a future dip in sales could benefit from promotional offers. As a matter of fact, 49.8% of all food products and 58.6% of all non-food products in the UK are sold on promotion. Since inventory is expensive to keep on shelves and stockouts are detrimental to both short-term revenue and long-term customer engagement, marketing executives are interested in implementing more accurate forecasting methods; especially when data sources are insufficient to provide accurate or actionable predictions for example due to small sample sizes or coarseness of historical sales i.e new products, niche products. Thus, this paper attempts to implement a statistical and a machine learning model to make more accurate Profit forecasts. It compares a neural network and a Multiple Linear regression model on a dataset from a public platform. It is organised as follows: Section 2: Exploratory Data Analysis, Section 3: Neural Network, Section 4: Multiple Linear Regression, Section 5: Results & Comparison, Section 6: Discussion.

2 Exploratory Data Analysis

2.1 Dataset

This dataset has been extracted from the online platform Kaggle. It is consisted of 50 startups from New York, California, and Florida. It contains information regarding: Profit, R&D Spending, Administration Spending, and Marketing Spending for each startup.

2.2 Visualisation

Initial investigations have been performed on data by using Tableau Desktop software to: discover patterns (trends or seasonalities), spot any anomalies (outliers), test hypotheses and check any assumptions with the help of graphical representations.

The two dashboards (Figure 1 and 2) illustrate an overview of the data. In the first dashboard (Figure 1), there are 3 bar charts and a geomap, demonstrating the Profits and Expenses per state. In depth, it seems that the 3 states spend roughly the same amount on Research and Development. In respect of Profits generated per state, New York is leading with some $30k more than the second Florida and some $150k more than California. All states spend between $400-470k on Marketing Expenditures. Lastly, among the three states, Florida spends the highest amount on Administration.

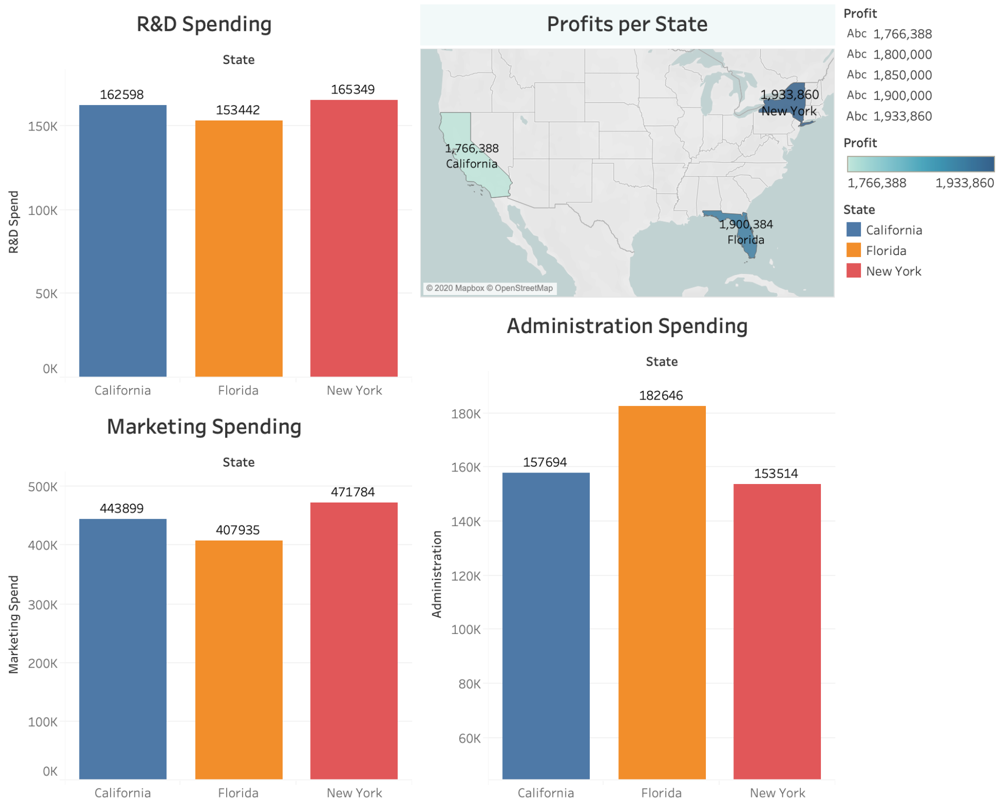


Figure 1: Bar Charts Per Variable

In the second dashboard (Figure 2), the correlations between Profit and the independent variables are shown. It is obvious that there is a strong relationship between Marketing Spend and Profit. In simple terms, more money being spent on Marketing, higher amount of Profits being generated. There is no apparent pattern between Administration expenses and Profits earns and Research and Development seems to have a linear relationship with Profits too.

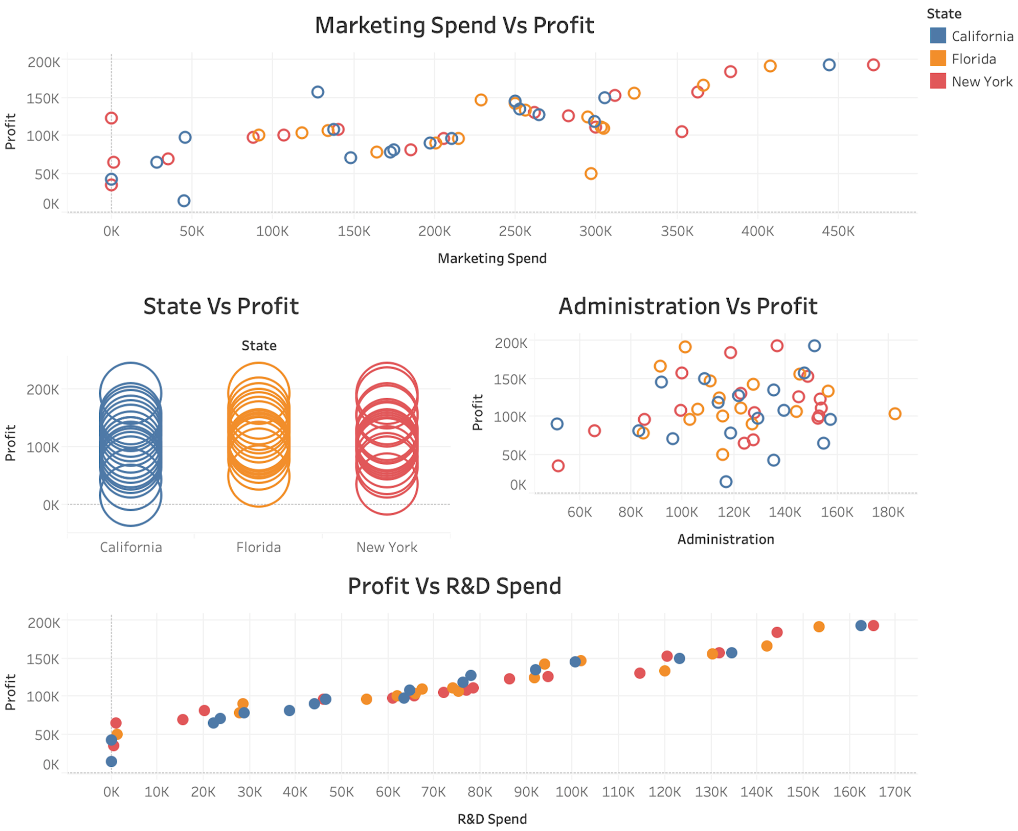


Figure 2: Correlations between Output and Inputs

3 Neural Network

Neural networks are a form of artificial intelligence that can capture complex relationships between outcome and predicting variables. Neural networks have been arguably labelled as a “black box” because they “provide little explanatory insight into the relative influence of the independent variables in the prediction process”.(Olden et al., 2002). They essentially search many models (including nonlinear models) to find a relationship involving the independent variables that best predict the dependent variable. As with regression, neural networks have a certain number of observations. Each observation contains a value for each independent variable and dependent variable. The main goal of neural networks is to make accurate predictions for the output cell (or cells)(Yildirim,2020).

3.1 Data Normalisation

Data preparation includes techniques such as normalization to scale the input and output variables. It is a recommended pre-processing step when working with deep learning neural networks as the variables may have different units (e.g. thousands, kilometers, etc.) that means the variables have different scales. Differences in the scales across variables may increase the difficulty of the problem being modelled. For instance, large input values (thousands of units) could result with a model that learns large weight values. Large weight values make a model unstable, which eventually may suffer from poor performance during learning and may result in higher generalisation error. On this account, this dataset has been normalised and all the values were in the range of [0,1] with $\mu = 0, \sigma = 1$.

3.2 Data Splitting

In order to learn from and make predictions on data, input data will build the neural network model. Machine learning literature suggests initially to fit the model on the training set, which is a set of examples used to fit the weights of connections between neurons of the model. Then, the test (holdout) set will be used to provide a more general and unbiased evaluation of the final model fit on the training set. As per the forecasting literature suggestions, the dataset has been splitted in according with the rule of 70:30. So a randomly selection of 70% has been used to train the model, while the 30% will be used to validate the model.

3.3 Neural Interpretation Diagram

The far-left nodes on Figure 3, also known as the Input nodes, are consisted of: R&D Spend, Administration, Marketing Spend and State. They are the raw data variables which feed the model. The black arrows and their associated numbers are the weights which indicate the extent which each variable contributes to next node. The blue lines are the bias weights. The middle nodes, which include all the nodes between the input and output nodes, are the hidden nodes (which consist the hidden layer). Each of these nodes constitute a component that the network is learning to recognize. The far-right (output node) node is the final output of the neural network. All of this is omitting the activation function(as shown on the figure 4) that will be applied at each layer of the network as well.

According to Olden and Jackson(2002), "the relative contributions of the independent variables to the predictive output of the neural network depend primarily on the magnitude and direction of the connection weights". Hence, the input variables with larger connection weights, as R.D.Spend(4.098) and Marketing Spend(0.393), represent greater intensities of signal transfer and, therefore, are more important in the prediction process compared to Administration(-0.081) and State(-0.078) variables. Additionally, as mentioned on Olden and Jackson(2002) paper, these negative connection weights of Administration and State variables represent “inhibitory effects” on neurons which reduce the intensity of the incoming signal and decrease the value of the predicted response. This can be clearly seen on Figure 6, where the thickness of the connections is proportional to their connection weights.

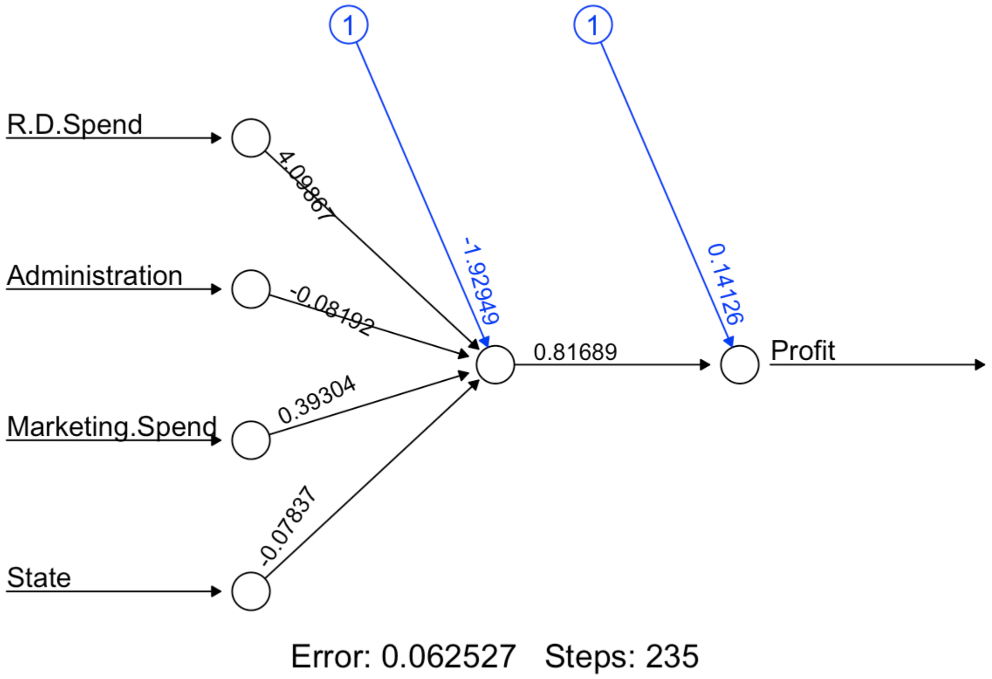


Figure 3: First Neural Network Model

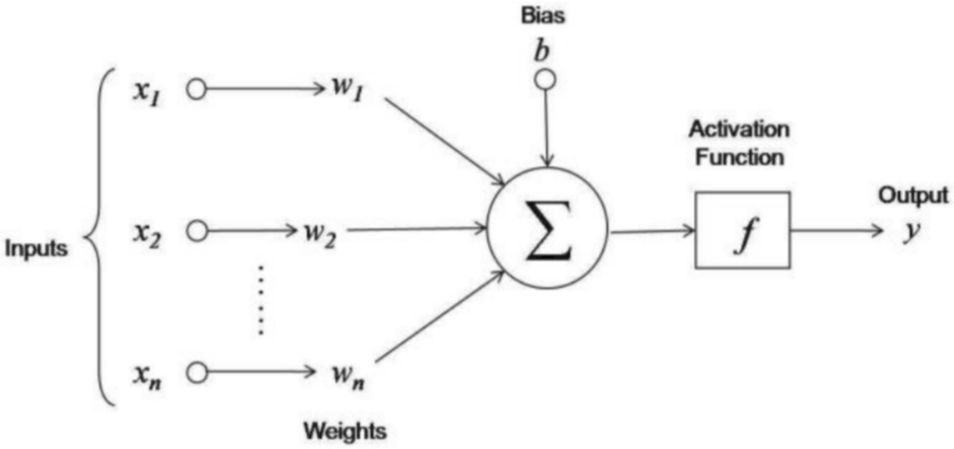


Image Source

$$Y = \sum(weight * input) + bias$$

Figure 4: Neural Network mapping

Worth to mention that the Overall Error is 0.06 and the number of steps needed to converge is 235.

Subsequently, the model has been redesigned (Figure 5) in order to compare the performance of neural network with 1 and 2 hidden neurons in the hidden layer. It is important to note that "the number of neurons in the hidden layer is the mean(μ) of the number of features in the independent and dependent variables"(Yildirim,2020).

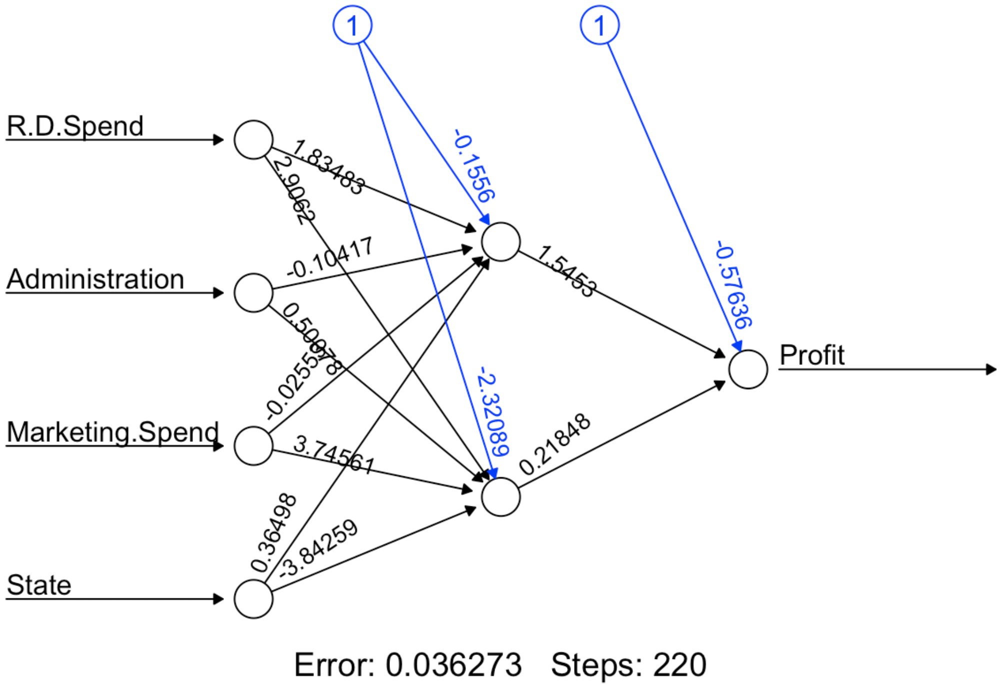


Figure 5: Second Neural Network Model

The SSE(Error) has been dropped to 0.04 and training steps had been reduced to 220 as the number of neurons under the hidden layer have been increased from 1 to 2.

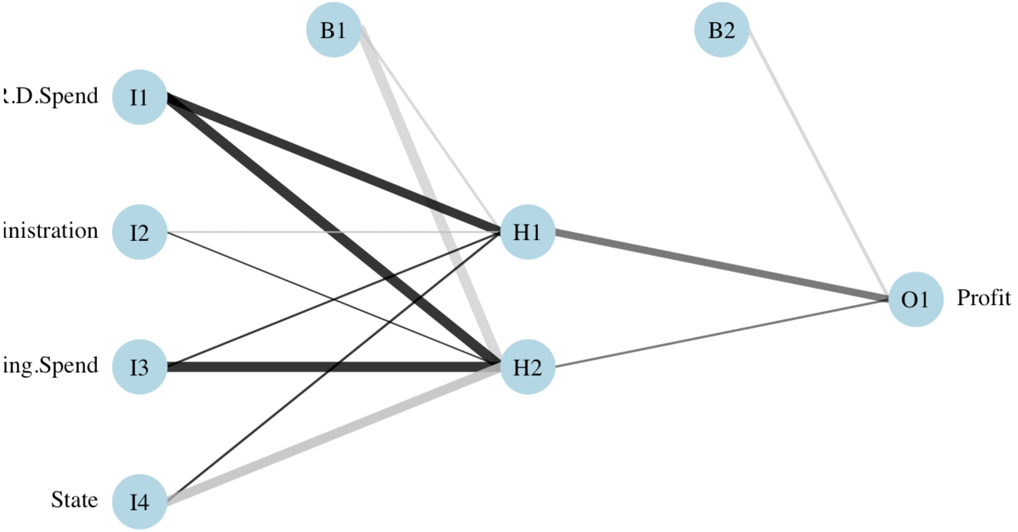


Figure 6: Second Neural Network Model(thickness proportional to magnitude

In the final step, the data had been de-normalised in order to get the actual prediction on Profits. Also, the RMSE (9277) has been computed to evaluate the model performance and to compare it with Linear Regression error.

4 Multiple Linear Regression

After running the neural network model, a Multiple Linear Regression was analysed to predict Profits from R.D Spend, Administration, Marketing Spend and State. Then, a comparison between the two model performances will determine the highest predictor power of the models.

An important limitation of Multiple Linear Regression is that it makes several key assumptions.

4.1 Assumptions

4.1.1 Linear Relationship between Outcome variable and Independent variables

Scatterplots can show whether there is a linear or curvilear relationship

4.1.2 Multivariate Normality

The residuals are normally distributed.

4.1.3 No multicollinearity

Independent variables are not highly correlated with each other. This assumption can be tested using Variance Inflation Factor values.

4.1.4 Homoscedasticity

The variance of error terms are similar across the values of the independent variables. A plot of standardized residuals versus predicted values can show whether points are equally distributed across all values of the independent variables.

A general rule of thumb for the sample size is that regression analysis requires at least 20 cases per independent variable. In this case we have 50 non-null records.

All tests of the assumptions are included in the R Markdown file.

4.2 Model Summary

The multiple linear regression equation is:

$$
\begin{aligned}
\hat{\text{Profit}} = 51870 &+ 0.792(\text{R.D.Spend}) \\
&- 0.043(\text{Administration}) \\
&+ 0.028(\text{Marketing.Spend}) \\
&+ 586.5(\text{State})
\end{aligned}
$$

4.2.1 Model Accuracy

Importantly, the model has a high Adjusted R^2 : 93.02%. That implies that the independent variables in the model explain by 93.02% the variation in the dependent variable.

The Residual Standard Error is 10690 . The RSE corresponds to the prediction error, which is the average difference between the observed outcome values and the predicted values by the model. The lower the RSE, the best the model fits to the data.

4.2.2 Coefficients significance

The β(beta) coefficients can be interpreted as the average effect on y of a one unit increase in predictor, holding all other variables constant.

For example, for a fixed amount of Administration, Marketing Spend and State, spending additional 1000 dollars on Research and Development leads to an increase in Profits by approximately (0.709)*1000 = 709 Profit units, on average.

4.2.3 Making predictions

In order to evaluate the performance of the regression model, predictions have been made using the test (holdout) dataset.

The procedure was as follows:

1)Predict the Profit values based on the test data

2)Assess the model performance by computing the RMSE and R^2. The prediction error RMSE (Root Mean Squared Error) is 5964. It represents the average difference between the observed known outcome values in the test data and the predicted outcome values by the model. The lower the RMSE, the better the model. The R^2(95%) represents the correlation between the observed outcome values and the predicted outcome values. The higher the R^2, the better the model.

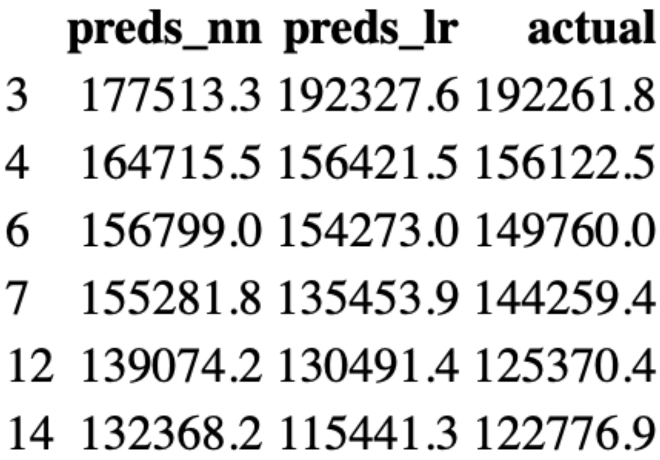
5 Results & Comparison

Recalling the predictions from the models on the test set:

Neural netowrk: RMSE(9277)

Linear Regression: RMSE (5964)

This shows that linear regression is superior to artificial neural network model in this case in terms of accuracy.



	preds_nn	preds_lr	actual
3	177513.3	192327.6	192261.8
4	164715.5	156421.5	156122.5
6	156799.0	154273.0	149760.0
7	155281.8	135453.9	144259.4
12	139074.2	130491.4	125370.4
14	132368.2	115441.3	122776.9

Figure 7: Predictions from NN&LR and the actual Profits

6 Discussion

In recent years, there has been much hype about the superior performance of neural network analysis. The results of the problems considered in this paper clearly shows that regression methods are much reliable once the data are adjusted such that the assumptions of the method are reasonably satisfied. It is difficult, though, to compare the efficacy of the techniques and determine the best one because their performance is data-dependent. This attempt is not to underestimate the potential of neural network but to point out that the performance of neural networks can be improved by knowledge of the statistical methods. We should also have in mind that accuracy alone is not the most important criterion. The forecaster may wish to forgo some accuracy in favor of other techniques. Also, the forecaster can improve the forecast by: combining forecasts, simulating a range of input assumptions or selectively applying judgement. Because the forecaster cannot predetermine which techniques will be superior in any situation, combining forecasts, simulating a range of input assumptions or selectively applying judgement, could offer an assured way of improving quality.

References

[1] Farhan. (2018) *50 startups.* Available from: https://www.kaggle.com/farhanmd29/50-startups

[2] Yildirim G. (2020) *Retail and Marketing Analytics.*

[3] Brownlee J. (2019) *How to use Data Scaling Improve Deep Learning Model Stability and Performance.* Available from: https://machinelearningmastery.com/how-to-improve-neural-network-stability-and-modeling-performance-with-data-scaling/

[4] Olden J., Jackson D.(2002) *Illuminating the "black box": A randomization approach for understanding variable contributions in artificial neural networks* Available from: https://www.sciencedirect.com/science/article/abs/pii/S0304380002000649

[5] Avlani A.,(2019) *Neural Network Models in R.* Available from: https://www.datacamp.com/community/tutorials/neural-network-models-r

[6] Kumar U. (2005) *Comparison of neural networks and*

regression analysis: A new insight. Available from: https://www.sciencedirect.com/science/article/abs/pii/S0957417405000904

[7] LaRiviere, J. McAfee, P. Rao, J. Narayanan, K. & Sun, W.(2016) *Where Predictive Analytics Is Having the Biggest Impact.* Available from: https://hbr.org/2016/05/where-predictive-analytics-is-having-the-biggest-impact

[8] *Rpubs* Available from: https://rpubs.com/thirus83/453911

[9] *Rpubs* Available from: https://rpubs.com/shyambv/linear_neural_network

[10] *Training, validation, and test sets* Available from: https://en.wikipedia.org/wiki/Training,_validation,_and_test_sets