Implementation of a Neural Network Approach for Predicting Sales Profit

Abstract

An artificial neural network (ANN) primarily relates to a computer device that is influenced by the process, function, and cognitive development similar to that of a human brain. It imitates the functionality of neurons operating in the human brain. By understanding, observing and recognizing the data patterns, it is able to solve complex and dynamic problems in real time with some reasonable probability constraint. This paper implements the Microsoft neural network in Microsoft SQL Server Management using visual studio on a publicly available and accessible dataset to prove the effectiveness of neural network in transforming raw information to an in-depth insight of the trends in the dataset which is not easily visualized. This paper also provides the error margin or standard deviation in the value prediction of the algorithm performed on the database. The dataset selected constitutes high volume of bike selling records across Europe and can therefore be considered as a "big data", that cannot be resolved using normal paper record approach or traditional data mining technique that have no learning capability. The paper also investigates the optimizers and modifiers that effects the prediction value processed by the SQL database. In brief, neural network can resolve big data information to useful knowledge in the form of predictions as compared to traditional data analysis methods that cannot accurately access or interpret data so complex and dynamic in order to provide quality forecasting.

Keywords: *Artificial Neural Network, Prediction, Forecasting*

1. Introduction

The word 'big data' has recently been emphasized a lot, but few individuals understand what big data is. Corporations, state agencies, pharmaceuticals, monetary and educational institutions use Big Data's potentials to increase organizational opportunities and better service offering [1]. Big Data is a terminology used to classify a set of information that is massive in quantity and is increasing dramatically with time. Then why is it necessary to extract data? Anyone can observe the impressive figures in different fields annually as the amount of data generated doubles. 90% of the global world contains unstructured data itself and is increased in just 2 years [2]. More data is not in any way means more awareness. It is just a set of information and without any knowledge describing the trends and predictions which make it useless to take any advantage of for any institution. In short, such data is so diverse and intricate that conventional data processing techniques are unable to access or analyse it effectively to have any proper forecasting. This paper study implementation of one of the modern data mining structures for comprehending an available open source database. Data Mining is characterized as the method of retrieving relevant data from large database collections. The applied data mining technique on the database is the neural network which is implemented using SQL server and visual studio. The dataset consisted of bike sales in Europe which is segregated in terms of gender, age and country with other information [3]. Using this dataset, the predictions are made for targeting and focusing, an optimized environment for the sales of bikes.

2. Literature Review

The data digging processes that identify hidden ties and forecast potential patterns have been around for several years. The word 'data mining' was not introduced till the 1990s, also known as information exploration in databases [4]. The framework encompasses these three overlapping fields: mathematics, artificial intelligence as well as machine-learning algorithms. The data mining tools continues to grow in conjunction with Big Data's infinite capacity and competitive processing resources. In the last decade improvement in capacity and rate of computation has contributed to a fast, simple and automatic statistical evaluation removing manual setups, complicated and repetitive activities. As further the diversity of the data sets gets, the related observations and accuracy are to be discovered become more promising. The distributors, financial institutions, vendors, service provider and insurance agency use data mining among other things to uncover how advertising incentives and analytics impact marketing strategies, profits, risk, competition, company operations and consumer perceptions.

So why does data extraction need to be done? Everybody can notice the amazing numbers of entries from numerous fields globally whose sum of data produced doubles and where 90% of the data globally itself involves unorganized material [2]. More data does not mean more insight in any way. It is just a collection of details that makes it worthless to use without expertise to explain patterns and forecasts. The data mining tools accomplish data analyses through scrolling the records and removing the noisy entries at a high speed. It recognizes what is apparent and then utilize it to determine the possible effects. Thus, speeding up rational decision-making.

A Neural Artificial Network (ANN) may be viewed as a classifier and a prediction tool. In SQL Server, it is named as Microsoft Neural Network and is predominantly more complex and advanced than Decision trees and Naïve Bayes data mining techniques. It attempts to replicate the functional intricacies of the biological system of the brain. Electric impulses are transmitted by neurons to provide input to the brain and relay them across the network when then triggers human brain output behaviour. The brain decides on action depending on the signal's directions, despite signals being similar. The brain analyses the features of the signals getting collected and can identify the sort of instruction obtained from that processing. The ANN, to mimic this behaviour, has multiple sections: the node function as the neuron, while the weightage given to different values and the relation between various nodes act similar to the biological network of the brain [5].

In this method, as seen in the following figure, it can be observed that at the simplest level there are three sections: input, hidden, and output.

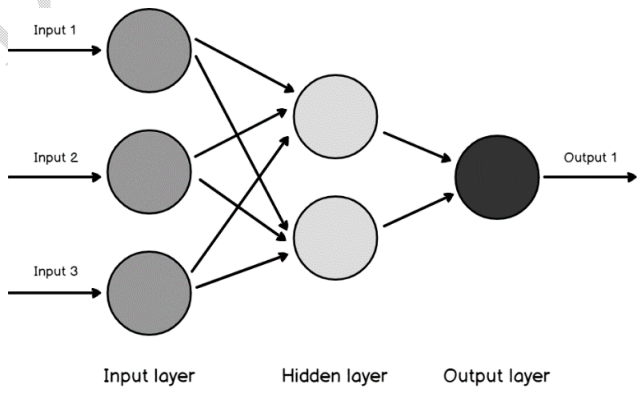

Figure 1. Artificial Neural Network hierarchy [6]

According to current dataset the output is set to profit while say the 3 inputs that is signifying the major change in the profit is the age, gender and country. The Hidden layer is a medium which provides all weighted inputs to every node within the hidden layer. The output is modelled according to the projected parameters. A neuron is a simple unit integrating many inputs for a single output. Input variations are created using various methods. The weighted sum technique is used by the Microsoft Neural Network. Once these input data have been determined, the evaluation function is used. In general, often small inputs do have a huge output and, while huge inputs could be negligible to output. As a consequence, usually non-linear equations are used for evaluation. In Microsoft Neural Network, tanh is used for the evaluation and the sigmoid operator for the output. Backpropagation is the central component of the Artificial Neural Network. Unlike most methods, this method has the potential to adapt as it learns. Learning feature is obtained by Backpropagation. In this procedure, the deviation is measured from actual value and the weighted values in hidden layer are adjusted.

In earlier phase, the Microsoft neutral network in SQL server allocates one on one arbitrary values as weightage. From the training set, the output and performance errors are determined by the neural network. The Backpropagation operation measures the inaccuracy of every output value and hidden layers of the system. After the error difference is calculated the weightages are updated. The iteration cycle repeats from the error determinations in the values from the training sets and updates the weights in the hidden layer of the network.

This dynamic and multi-layered framework for the modelling of machine learning provides with the concept of deep learning which is just the number of additional nodes added in the hidden layer in a given network contributing more accurately towards the forecasting of various characteristics of the dataset [7]. One drawback of SQL database algorithms such as is the execution of queries is performed through CPUs rather than the GPUs. The processing with GPU speeded up the SQL databases. Platforms such as blazingdb and mapd significantly provides high performance in database evaluation using distributed SQL engine [8]. The query and information exchange can be accompanied by the distributed SQL engine while local processing on the dataset can be performed using a GPU enhanced server.

3. Methodology

The data mining methodology used on the database is the neural network that is applied using the SQL server and the visual studio. Dataset is randomly divided into two groups, a testing set for validation of the model and a training set. Testing set is depended on the percentage selected from the maximum number of test data. It is used to verify the consistency of the model. The training set is a compilation of remaining data which is used to build a model for data mining. The dataset is comprised of 113036 entries from which 30% of the data is used for training the algorithm. Many trends can be visualized and predicted which effects the profit and thus an optimized marketing strategy can be executed. Through this technique aside from profit one can also include constraints which is in big data is very complicated to implement and visualize. These constraints can bound the profit forecasting to much narrower result i.e. suppose that the cost of the bike is not much flexible due to international market rates but at the same time you wanted to maximize the profit with keeping all the other variables of the data intact in the solution. By eliminating the cost variability, the output prediction would be much closer to the probability of getting higher revenue and thus profit. Similarly, gender and country are other aspects that can be taken as inputs for predicting the trend of customer buying bikes.

The database for implementing data mining structure requires 3 types of data and at least require one entry in each column of these data types. These 3 categories of data are key, input and output. Input and output as the name suggest are the entries and the solution of the system on which algorithm are operated on. As in normal SQL query, the key is the most important data category as this uniquely identify each entry of the database. SQL keys guarantee the redundant information is not present in rows. Not only does it support, it also helps create a links with various database tables [9]. However in this database it is kept simple and all the data is present in single database table. The table although do not have a unique key to identify each row. To resolve this problem each entry of the database is manually given an ID number, but not through going each entry one by one but using Microsoft excel as a tool. A new column is added at the leftmost side of the table. The first few cells are selected of the ID column with number added to them for the series function to follow and is autofill till the end [10].

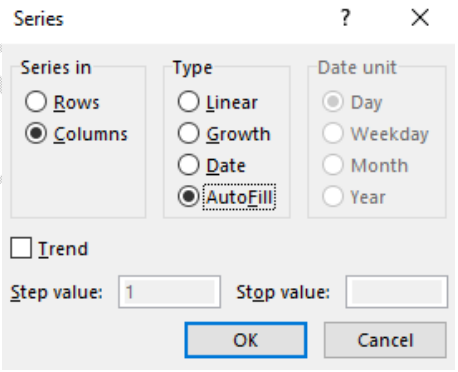

Figure 2. Unique ID provision for each entry

This new data table is uploaded to the Microsoft SQL server management and stored in Database subcategory of an instance on local server. Furthermore, the database is auto protected and is only accessed by the person who have server key. This also does not let anyone even the owner to transfer or link it to any other application without special permissions configured by the owner. These settings are configured in the security section under users. A new user is added and given the permissions to read the dataset so when it is used by visual studio it can implement the data mining algorithm and not prevented from reading the data.

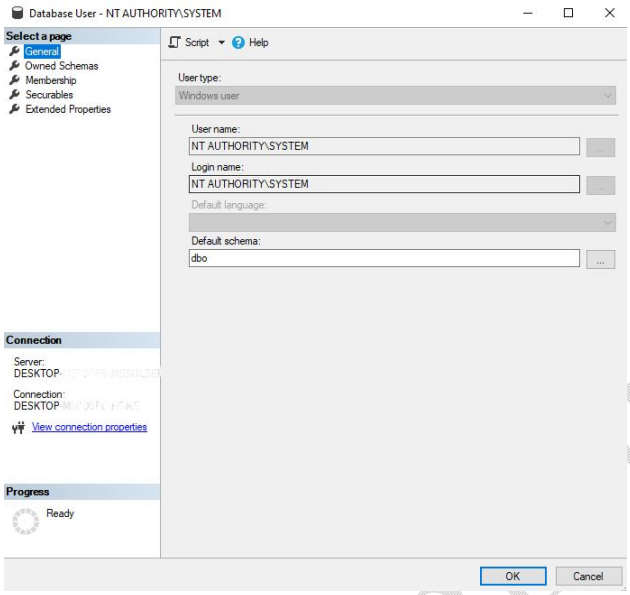

Figure 3. Database permission

Two further data files are added/created for generating mining structure for the data mining algorithm.

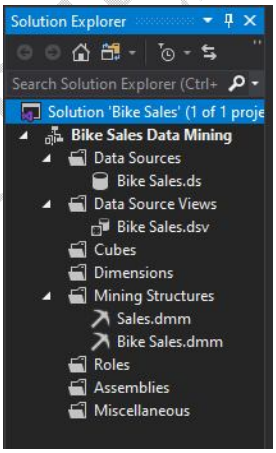

Figure 4. Visual Studio solution explorer

First one is the selection of data source where the actual table consisting of all the entries is located. This table is selected in the second data file which is data source view. This file enables to create relations between different tables of the same data source and view them in a much more conceptualized manner and easily visualized. Although the database created here does not have extra tables and therefore no relationship configurations are required.

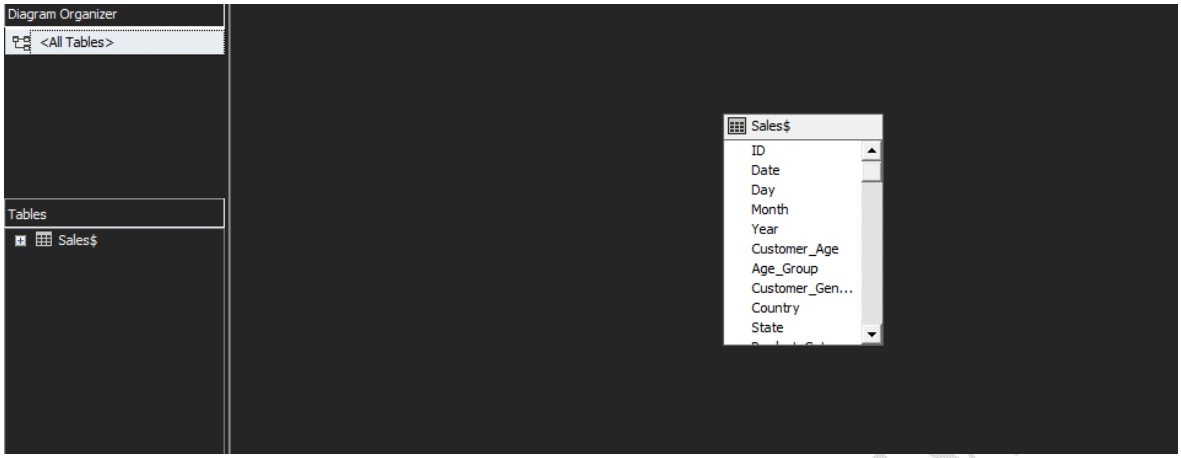

Figure 5. Relationship organiser

For the implementation of data structure, the existing relational database or data warehouse can be selected which is already included in the project solution. In mining structure, the data mining technique is set to Microsoft neural network and here the inputs and outputs, including the key are identified and selected.

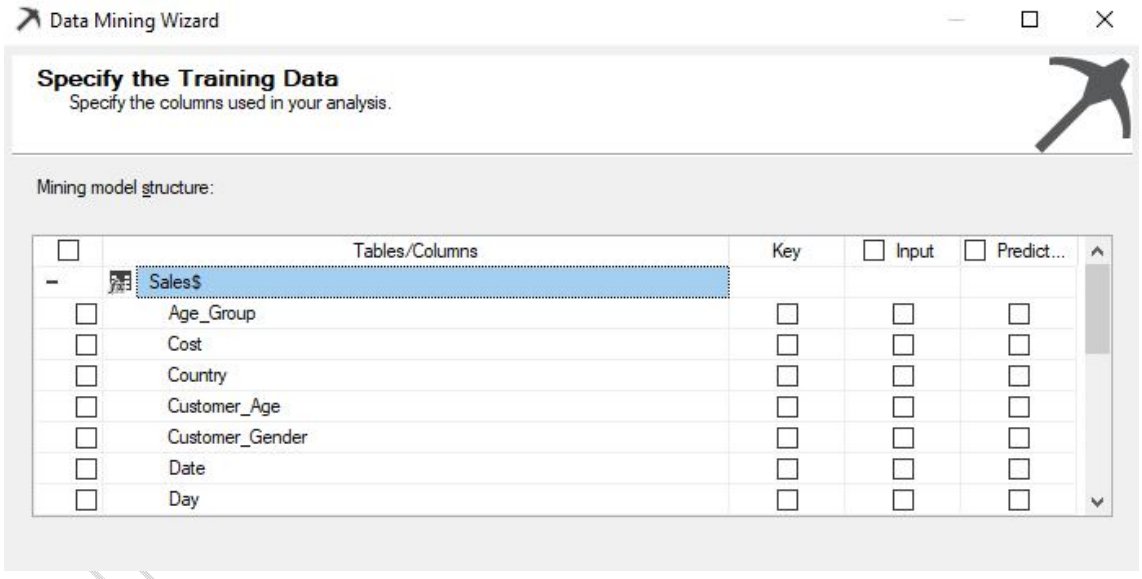

Figure 6. Neural Network mining structure

The key as already discussed is the ID provided manually through excel. The inputs represent the mining model and as a criterion from which the output can be drawn. Output designates the predicting or predictable column or more specifically the variable that is needed to be optimized.

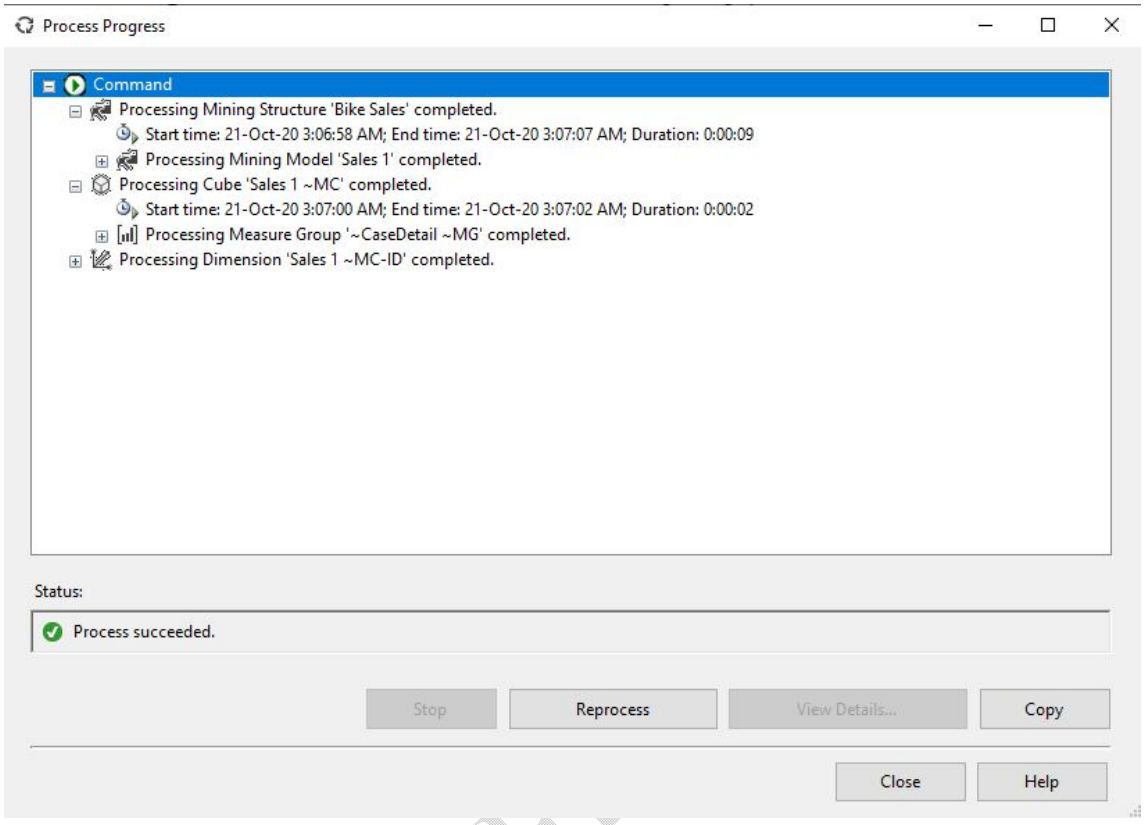

Figure 7. Execution of mining algorithm on the dataset

The mining model is processed and the system at the back end perform all the necessary operations to implement the neural network. After the deployment went successfully the software notify the user that it is successful and is operational.

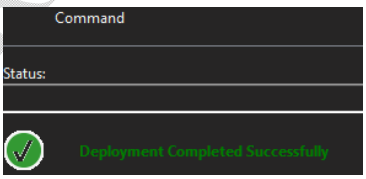

Figure 8. Final validations of every process

In the mining model viewer, the output data forecast can be viewed with different fixed attributes and various ranges of profit can be visualized limited to certain probability values.

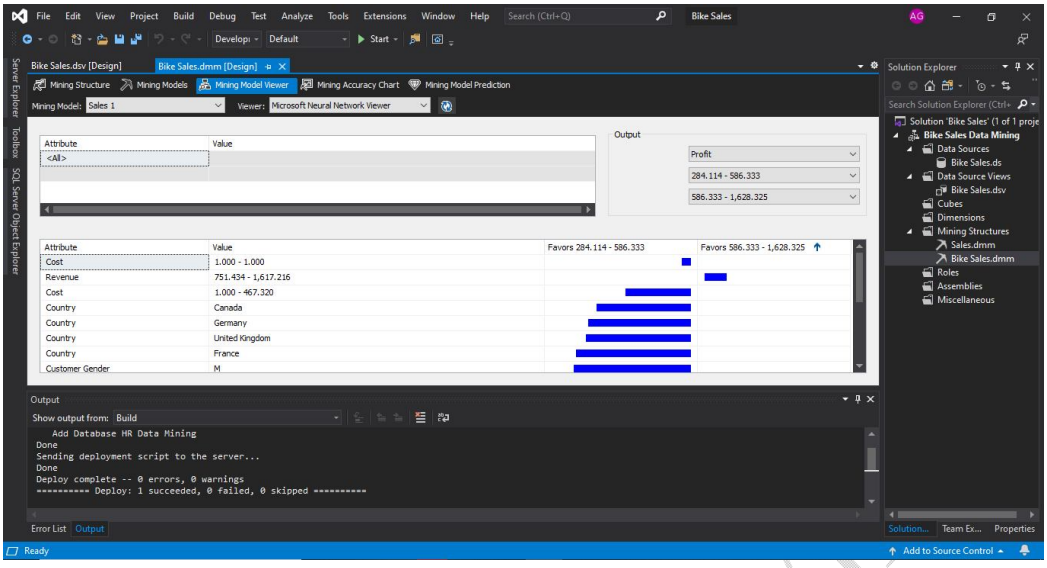

Figure 9. Implemented neural network viewer

This model can also be implemented on values other than the current provided table with its respective entries from which test sets and training sets are acquired, through the use of mining model prediction. One can provide at least 2 columns that will interact with each other and provide the values based on the predictions done by the neural network algorithm. This would provide a much more in-depth insight to the big data and from which the industries can improve their business strategy. The mining model prediction tab can also be used to perform specific SQL queries without any need to write code for it or if someone desires there is option to write code manually as well.

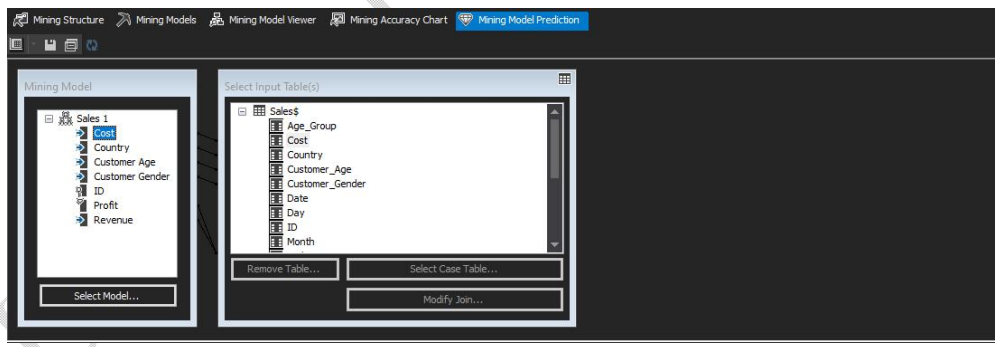

Figure 10. Mining model predictor

4. Results

The algorithm successfully predicted the output and below graph shows how much the ideal data predictions and real data differentiates after the algorithm is trained from 30% of the dataset. Through cross validation it is seen that log Standard Deviation between dataset and prediction are 0.0528.

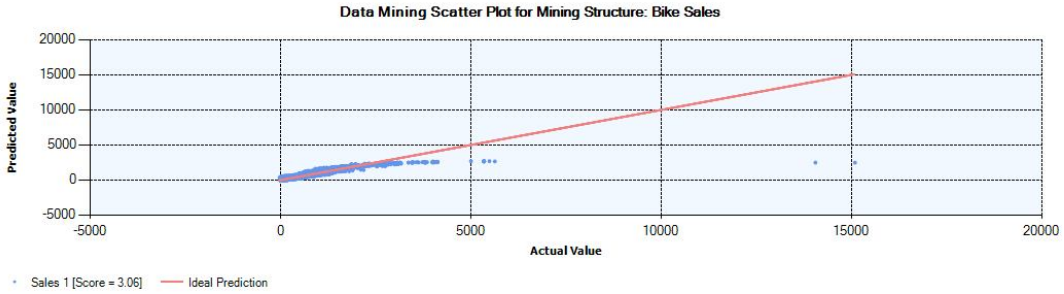

Figure 11. Scatter plot of the output comparison between predicted and actual values

This new data generated by predictions can provide insight to the business owners or vendors to check the average age of the population of certain area and see the trend which is best suited to attract the customers of that age group to buy their bikes.

The identification of irregularities is another advantage of using neural network approach in data mining. It is sometimes not enough to merely monitor trends or categorize information to comprehend given data set. For example, in a product which is although influenced by men, a company can see an odd spike of women's clients. In the scatter plot, it can be seen that due to these spikes, the actual value increases to nearly 15000 sales while the predicted value retains the value below 5000 sales where most of the data cluster is situated. These spikes are called outliers and are a huge problem in big data that prevents businesses from increasing their sales and profits.

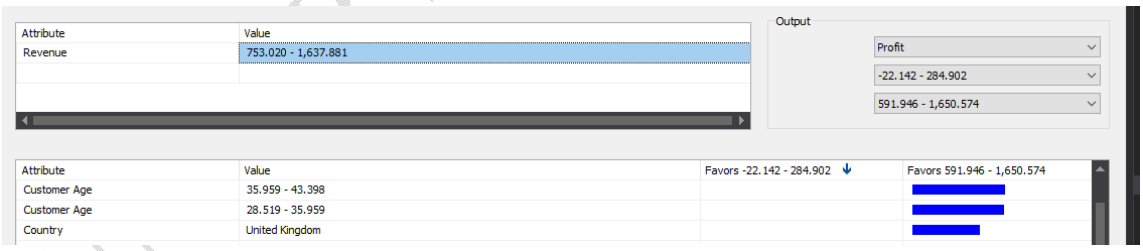

Figure 12. Data mining predictions

The predictions showed that the marketing focus should be on the customer with age of approximately between 28 to 43 and United Kingdom as a first choice of country to have significant profit range without any fear of loss. Similarly, other attributes can be compared with different conditions or find other options just because Sometimes it is not viable to go with the first option i.e. foreign relations.

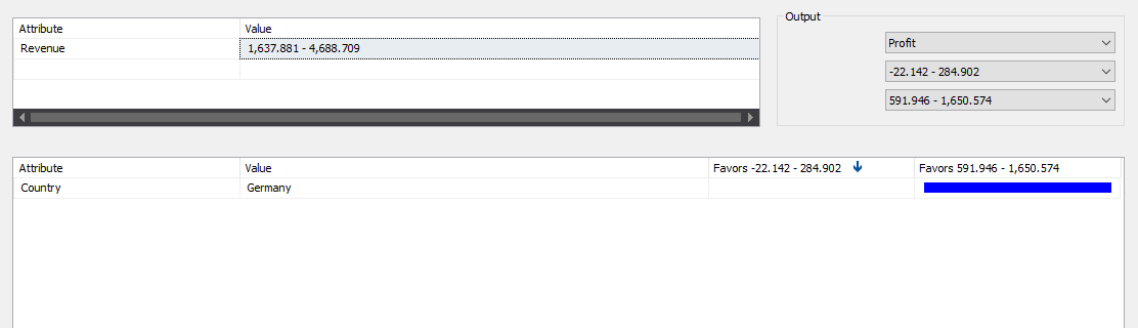

Figure 13. Optimisation of profit output

The attributes that form the prediction matrix is another important contributor in modifying the activating function in the hidden layer of the neural network. Each attribute is in some way putting weightage to the input values of the neural network. In the figure the mining structure is changed to constitute unit price and cost aside from the revenue and other modifiers it was generating. It can be observed that once the attribute is fixed to have maximum revenue on a fixed maximum profit in this new mining structure, the output shows that Germany is the best country and modifier in the specified outcome. Previously with old mining structure, the output showed that Canada being the best country for selling most bikes for maximum profit. This actually signify that the operator needs to provide the neural network some groundwork relations as the system does not exactly know aside from iterative procedures of assigning weightage to different values, that how much certain relation could have an impact on the predicted values. These input relations are the optimizers for the algorithm predictions while when these inputs are selected at some fixed range then they become modifiers for the predictions as desired or required by the user.

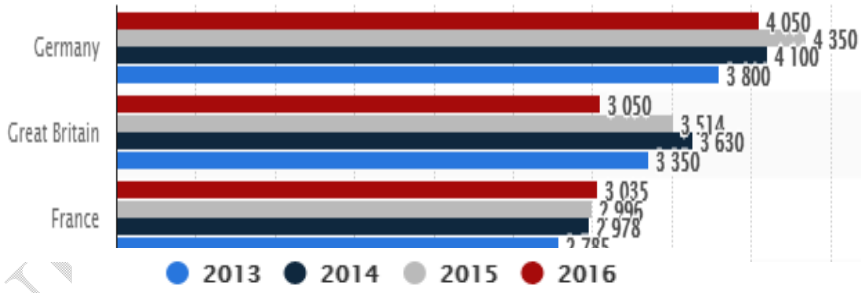

Figure 14. Statistics of the of top 3 countries in sales volume of bikes in thousand units between the years 2013 to 2016 [11]

To see the extent of this hypotheses, it is verified through internet statistics such that during the year 2013 to 2016 Germany, UK and France constitutes over 50% of all sales. The above graph shows sales volumes in thousand units [11]. Even in our mining prediction the second best country for most profitable outcome after Germany is United Kingdom aka Great Britain, thus proving that the neural network can identify the trends and provide forecast for selected variable for output prediction.

5. Conclusion

By just providing 30% of the database as training set the neural network accomplished an accurate estimation of country which would accommodate maximum profit with fix a maximum range of revenue. It detected the outliers that could have affected the output forecasts and the impact of these outliers is mitigated automatically with the back-end iterations performed by the neural network algorithm. It is also observed that the output predictions depend not significantly on the number of entries of the database for training but the number of input relations with the predicting output variable. The algorithm significantly reduced the effort and autonomously solved the complex relationship between the inputs and the prediction output. The big data constituting of 113036 entries with raw information were computed and resolved to provide useful insights into the data. This data can now be able to use by managers to manipulate the business strategy for maximum profit. The log Standard Deviation between dataset and prediction observed using cross validation is just 0.0528 while mean absolute error have a standard deviation of merely 2.5063. Thus, proving that predictions are accurate and is implementable in actual business strategy.

References

[1] R. H. Hariri, E. M. Fredericks and K. M. Bowers, "Uncertainty in big data analytics: survey, opportunities, and challenges," *Journal of Big Data*, vol. 6, no. 1, 2019.

[2] S. K. Pushpa, T. N. Manjunath, T. V. Mrunal, A. Singh and C. Suhas, "Class result prediction using machine learning," *Proceedings of the 2017 International Conference On Smart Technology for Smart Nation, SmartTechCon 2017*, 2018.

[3] s. shah, "Bike Sales in Europe," kaggle, November 2020. [Online]. Available: https://www.kaggle.com/sadiqshah/bike-sales-in-europe.

[4] S. Shanmuganathan, "From data mining and knowledge discovery to big data analytics and knowledge extraction for applications in science," *Journal of Computer Science*, vol. 10, no. 12, 2014.

[5] N. Zhang, S. L. Shen, A. Zhou and Y. S. Xu, "Investigation on Performance of Neural Networks Using Quadratic Relative Error Cost Function," *IEEE Access*, vol. 7, 2019.

[6] D. Asanka, "Implement Artificial Neural Networks (ANNs) in SQL Server," SQLShack, 14 April 2020. [Online]. Available: https://www.sqlshack.com/implement-artificial-neural-networks-anns-in-sql-server/.

[7] Z. Gao and X. Wang, "Deep learning," *EEG Signal Processing and Feature Extraction*, 2019.

[8] Y. Zhang, Y. Zhang, J. Lu, S. Wang, Z. Liu and R. Han, "One size does not fit all: accelerating OLAP workloads with GPUs," *Distributed and Parallel Databases*, vol. 38, no. 4, 2020.

[9] H. Köhler, U. Leck, S. Link and X. Zhou, "Possible and certain keys for SQL," *VLDB Journal*, vol. 25, no. 4, 2016.

[10] N. I. K, "How do you auto-number a large list of rows in Excel?," microsoft, 18 January 2010. [Online]. Available: https://answers.microsoft.com/en-us/msoffice/forum/msoffice_excel-mso_other-msoversion_other/how-do-you-auto-number-a-large-list-of-rows-in/b530980b-4e27-4fa2-b474-af70b27fc23f.

[11] Statista Research Department, “Number of bicycles sold in the European Union (EU) from 2013 to 2016, by country (in 1,000 units)*,” statista, 31 January 2020. [Online]. Available: https://www.statista.com/statistics/393948/bicycle-sales-volume-in-the-european-union-eu-by-country/#:~:text=This%20statistic%20shows%20the%20number,50%20percent%20of%20all%20sales..