Global Scientific JOURNALS

GSJ: Volume 10, Issue 9, September 2022, Online: ISSN 2320-9186

www.globalscientificjournal.com

Data Science and Machine learning: A survey on retail sales forecasting and prediction in Rwanda Supermarkets

Dr.Musoni Wilson[1] PhD, Mr.Ukwizabigira Jean Baptiste[2] Master's, Mr.Cyemezo clement Pierre[2] Master's.

wilson.m@keplercollege.ac.rw,clement.c@keplercollege.ac.rw,jeanbaptiste.u@keplercollege.ac.rw

[1] Information communication and Technology (ICT), Kepler College, Kigali, Rwanda.

[2] Information communication and Technology (ICT), Kepler College, Kigali, Rwanda.

Subject area: Machine Learning.

Abstract

This article focuses on how supermarkets choose to stock products at different times of the year, in most supermarkets there is no real formal system in place it is just a case of the person in charge of stock making their own decisions based on their past experiences and what they know and have observed from their clientele. So, most of the estimations are done by people. This is where the proposed research comes in. Instead of solely depending on people making their own estimations we can now utilize trusted calculations and analysis of the previous sales in a supermarket using python libraries and built in functions to make accurate predictions of which products will sell more than others and by what margin.The article aims to find and define the spending patterns of consumers to help shop owners stock products more efficiently and reduce on stock waste. This article also aims to see if the future purchases of consumers can really be predicted. The recommendation is that this prediction tool can be helpful to local supermarkets to possibly increase sales and cut down on products expiring on shelves. This will also help local supermarkets to better plan their financial years.

1. Introduction

The rise of new technologies has positively affected our everyday lives especially in terms of business where a lot of business solutions have been provided by information technology. Now it is almost impossible to run a business without any information technology solutions. Even our day today lives are dependent Information Technology. One of Rwanda's Vision pillars is private-sector led development, and for this to happen successfully it is important to integrate ICT into the private sector and help small to medium business owners run their businesses more efficiently and this is where the sales forecasting system could come in handy, to help people who might not have the accounting or business background to understand the importance of predicting future expenses and sales to help better stock products, better plan the direction of the business in terms of advertising as well, as well as general planning for the financial year. If supermarket owners know which products will likely sell more during the course of a month or the year than they can plan marketing campaigns and materials efficiently? There are countless advantages to the integration of this sales forecasting system that could possibly be used in other business outside of just local Supermarkets

2. Methodology

Data Analysis

Analysis of data is a process of inspecting, cleaning, transforming, and modeling data with the goal of discovering useful information, suggesting conclusions, and supporting decision making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, in different business, science, and social science domains. Data mining is a particular data analysis technique that focuses on modeling and knowledge discovery for predictive rather than purely descriptive purposes.

Sales Forecasting is considered Supervised Machine Learning. Supervised learning is where you have input variables (x) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output. Y = f(X), The goal is to approximate the mapping function so well that when you have new input data (x) that you can predict the output variables (Y) for that data. It is called supervised learning because the process of an algorithm learning from the training dataset can be thought of as a teacher supervising the learning process. We know the correct answers, the algorithm iteratively makes predictions on the training data and is corrected by the teacher. Learning stops when the algorithm achieves an acceptable level of performance

Cleaning the Data

Cleaning data should be the first step in any Data Science (DS) or Machine Learning (ML) workflow. Without clean data it'll be much harder time seeing the actual important parts in the exploration. Once the training of ML models begins, they'll be unnecessarily more challenging to train. The main point is that to get the most out of a dataset, it should be clean. In the context of data science and machine learning, data cleaning means filtering and modifying the data such that it is easier to explore, understand, and model. *Filtering out* the parts you don't want or need so that you don't need to look at or process them. *Modifying* the parts, you *do need* but aren't in the format you need them to be in so that you can properly use them. The dataset used needed the following changes to be considered clean:

- Dropping of Rows with Null Values
- Removal of all negative numbers and replacing them with absolute values
- Breaking down the column Sales_Date into the following columns time, day, week, and month.
- Adding sales as a column which results from multiplying the Unit Price with Quantity of each Product.

Sample of dataset

	Product_ID	Mass	Product_Class	Product_Type	Unit_Price	Quantity sold	Branch_Id	Sales_Date	Branch_Size	Branch_Location	Branch_Type	Total sales in RWF
0	KLMS17	12.090	Less fat	Cheese	3922.00444	11.0	S009	1/3/2019	Level2	Gasabo	Simba_Branch1	43142.04884
1	CHKS22	7.696	Normal	Sodas	757.82644	12.0	S008	1/3/2019	Level2	Nyarugenge	Simba_Branch2	9093.91728
2	BSLS12	22.750	Less fat	Meat	2223.40260	13.0	S009	1/3/2019	Level2	Gasabo	Simba_Branch1	28904.23380
3	DHDS18	24.960	Normal	Fruits line	2858.89150	14.0	S001	1/3/2019	NaN	Nyarugenge	Simba_Branch4	40024.48100
4	YXLS01	11.609	Less fat	kitchen items	845.62398	15.0	S003	1/3/2019	Level1	Nyarugenge	Simba_Branch1	12684.35970

Figure 1.0 Sample of dataset

Null Values

Most data science algorithms do not tolerate nulls (missing values). So, one must do something to eliminate them, before or while analyzing a data set. There are many techniques for handling nulls. Which techniques are appropriate for a given variable can depend strongly on the algorithms you intend to use, as well as statistical patterns in the raw data, in particular, the missing values, and the randomness of the locations of the missing values. Moreover, different techniques may be appropriate for different variables, in a given data set. Sometimes it is useful to apply several techniques to a single variable. Finally, note that corrupt values are generally treated as nulls.

The figure below shows the rows in the dataset used in this article and the number of missing values in each row.

Null Rows dataset

```
Branch_Size          2261
Total sales in RWF      0
Branch_Type             0
Branch_Location         0
Sales_Date              0
Branch_Id               0
Quantity sold           0
Unit_Price              0
Product_Type            0
Product_Class           0
Mass                    0
Product_ID              0
dtype: int64

 Product_ID           0.000000
Mass                 0.000000
Product_Class        0.000000
Product_Type         0.000000
Unit_Price           0.000000
Quantity sold        0.000000
Branch_Id            0.000000
Sales_Date           0.000000
Branch_Size         28.269567
Branch_Location      0.000000
Branch_Type          0.000000
Total sales in RWF   0.000000
dtype: float64
```


Figure.1.1 Null Rows dataset

It is clear that Branch_size has the highest amount of Null Values, followed by Description, these are both important rows and should be removed, So the best option is to remove all the rows with Null Values as this is a large dataset and removing these rows will not have too much of an impact on the algorithm implemented.

Removal of Negative Numbers

After attempting to plot a few graphs, it was clear that within the data set there were negative values in the Unit Price and Quantity column which makes no sense because a customer cannot buy a negative amount of Stock in terms of quantity or pay for a product with a negative Unit

Price. So the natural assumption is that this must have been errors which is expected when handling real world data.

The best option would be to assume that these are indeed the correct Quantity and Unit Prices but incorrectly entered as negative values, therefore it is necessary to return the absolute numbers in each row to get rid of all the negative values.

Breaking down Sales_Date

It is important to make sure all columns are in the correct format to be able to be processed by the chosen algorithm. So the first step was to transform Sales_Date into the right date format.

Sales_Date includes the day, month, year, and time the purchase was made by a customer. To properly see how each of these factors affect sales, it is necessary to turn each of these factors into individual columns. Below is a picture of the final result after the mentioned operations.

Sample of dataset after changes

	Product_ID	Mass	Product_Class	Product_Type	Unit_Price	Quantity sold	Branch_Id	Sales_Date	Branch_Size	Branch_Location	Branch_Type	Total sales in RWF
0	KLMS17	12.090	Less fat	Cheese	3922.00444	11.0	S009	1/3/2019	Level2	Gasabo	Simba_Branch1	43142.04884
1	CHKS22	7.696	Normal	Sodas	757.82644	12.0	S008	1/3/2019	Level2	Nyarugenge	Simba_Branch2	9093.91728
2	BSLS12	22.750	Less fat	Meat	2223.40260	13.0	S009	1/3/2019	Level2	Gasabo	Simba_Branch1	28904.23380
3	DHDS18	24.960	Normal	Fruits line	2858.89150	14.0	S001	1/3/2019	Level2	Nyarugenge	Simba_Branch4	40024.48100
4	YXLS01	11.609	Less fat	kitchen items	845.62398	15.0	S003	1/3/2019	Level1	Nyarugenge	Simba_Branch1	12684.35970

Figure 1.2 Sample of dataset after changes

Data Visualization

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. In the world of Big Data, data visualization tools and technologies are essential to analyze massive amounts of information and make data-driven decisions.

3. Result

Machine learning models

we divided our dataset into two variables X as the features we defined earlier and y as the Product_Branch_Sales the target value we want to predict.

This is a regression problem so we used Regression methods.

Train test split will be ratio respectively.

Machine Learning Models used:

- Linear Regression
- Random Forest Regressor
- Lasso Regressor
- Gradient Boosting Regressor
- Decision Tree Regressor
- Ridge Regressor

Deep Learning Model used:

Artificial Neural Network

The Process of Modeling the Data:

Importing the model

Fitting the model

Predicting Product Sales

Regression metrics

Score Metrics for Regression: Mean Absolute Error (MAE) - Mean of the absolute value of errors (absolute distance from true value):

Mean Squared Error (MSE) - Mean of the squared value of errors (squared distance from true value):

R^2 (coefficient of determination) - Regression score function.

Linear Regression

In statistics, linear regression is a linear approach to modelling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). Linear regression was the first type of regression analysis to be studied rigorously, and to be used extensively in practical applications. This is because models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters and because the statistical properties of the resulting estimators are easier to determine

Mean Absolute Error: 1.2588781740944013e+17

Mean Squared Error: 1.5455189915440857e+37

R^2 Score: -2.963116350531758e+25

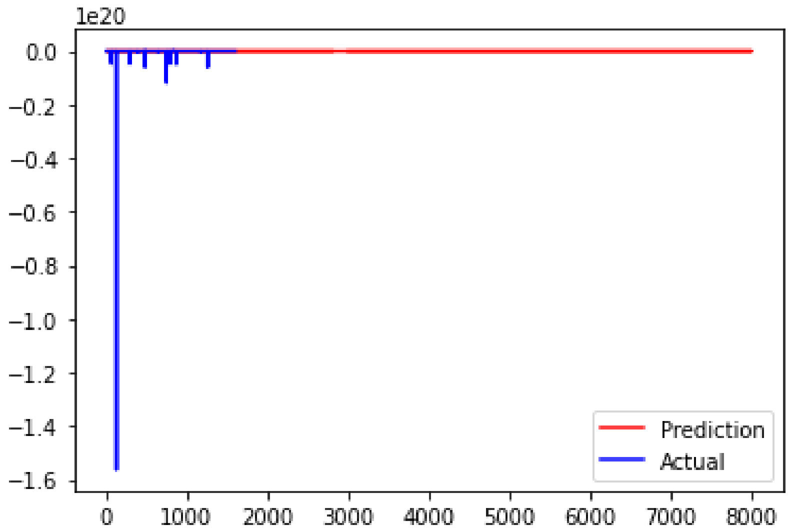

Figure 2.1 Linear Regression graph

Random Forest Regressor

Random forest is a Supervised Learning algorithm which uses ensemble learning method for classification and regression. It operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees

Mean Absolute Error: 81005.29
Mean Squared Error: 18076378196.53
R^2 Score: 0.9653

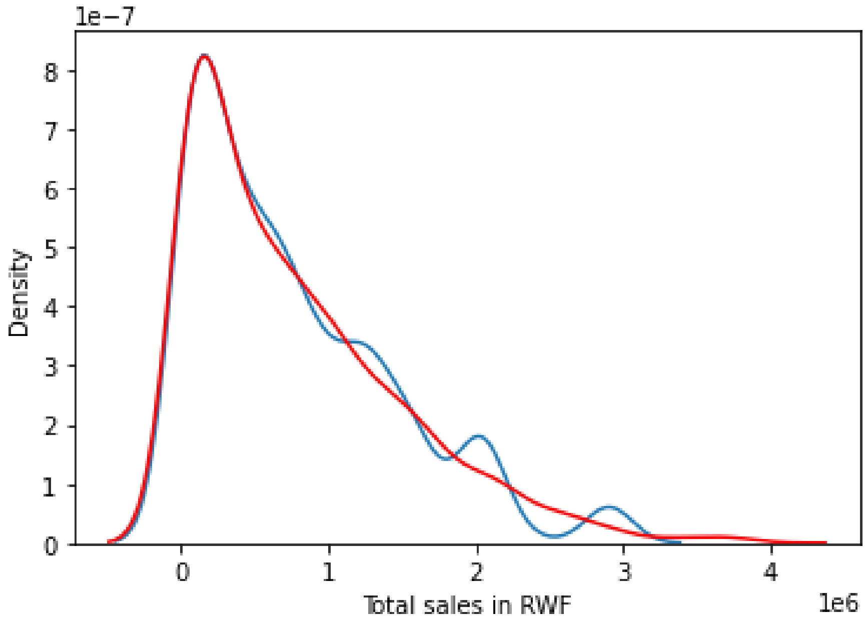

Figure 2.2 Random Forest Regressor graph

Lasso Regressor

In statistics and machine learning, lasso (least absolute shrinkage and selection operator; also Lasso or LASSO) is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces

Mean Absolute Error: 204450.06
Mean Squared Error: 86313641129.96
R^2 Score: 0.8345

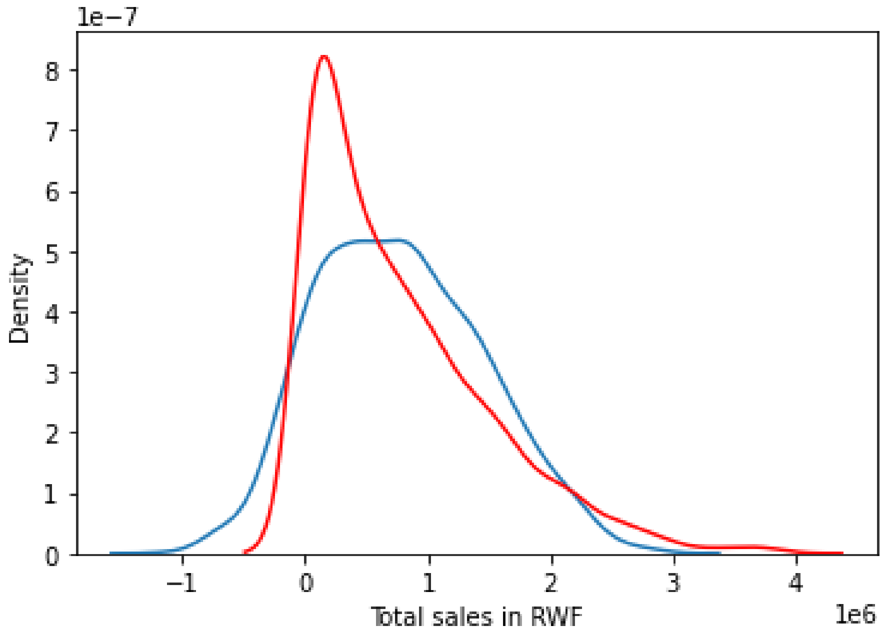

Figure 2.3 Lasso Regressor graph

Decision Tree Regressor

Mean Absolute Error: 97020.48

Mean Squared Error: 21931465357.44

R^2 Score: 0.958

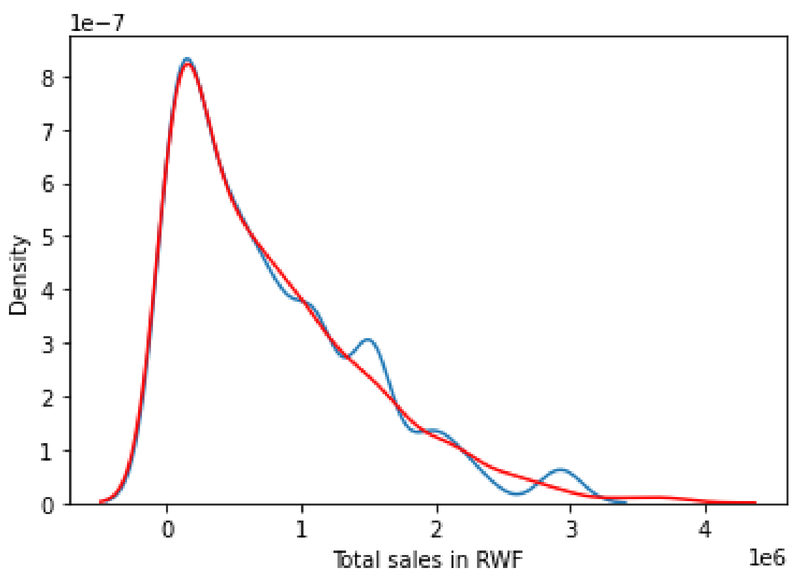

Figure 2.4 Decision Tree Regressor

Ridge Regressor

Mean Absolute Error: 189059.25
Mean Squared Error: 72037890447.05
R^2 Score: 0.8619

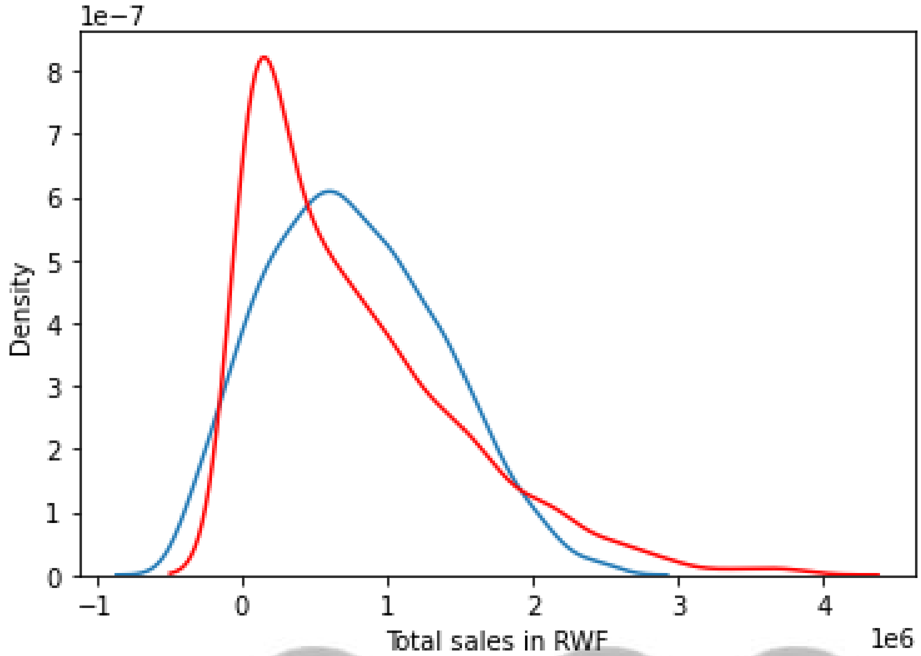

Figure 2.5 Ridge Regressor

Gradient Boosting Regressor

Mean Absolute Error: 21281.12
Mean Squared Error: 882118510.38
R^2 Score: 0.9983

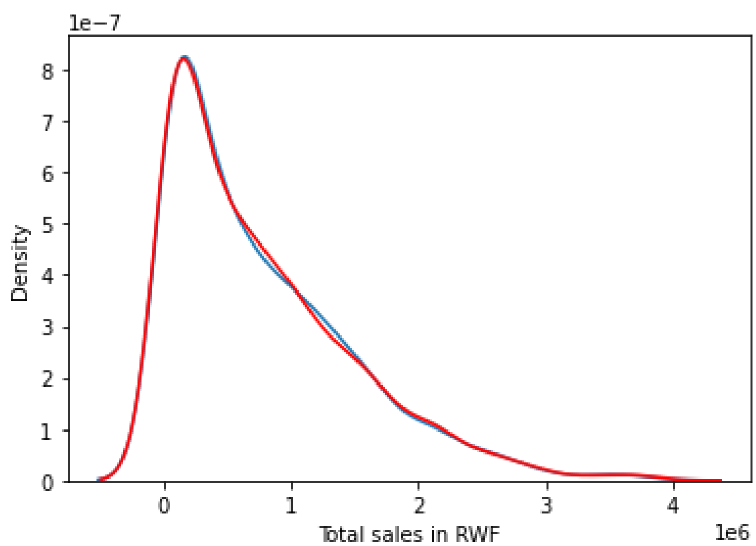

Figure 2.6 Gradient Boosting Regressor

Multilayer Perceptrons

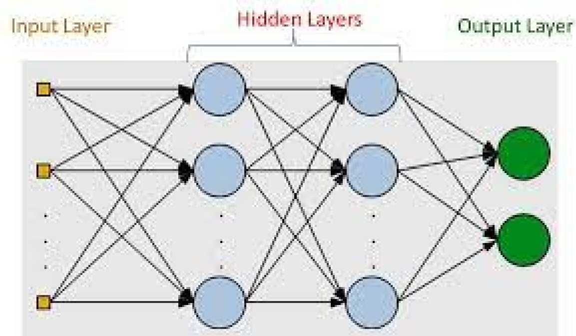

Figure 2.7 Archictecture graph of MLP

The MLP model stands for Multilayer Perceptron is a type of feed-forward artificial neural network (ANN) where the information flows from the input layer towards the output layer through the hidden layer. Rectified Linear Unit is used as activation function for Multilayer Perceptron algorithm. MLP makes use of a supervised learning algorithm called backpropagation for training the network. In backpropagation, the error is propagated backward throughout the network. The error is calculated by taking the difference between the network output and the actual output. The network parameters called weights are modified to minimize this error based on this method. This process is repeated several times until a stopping condition is reached.

Mean Absolute Error: 51421.09
Mean Squared Error: 11623311059.29
R^2 Score: 0.9777

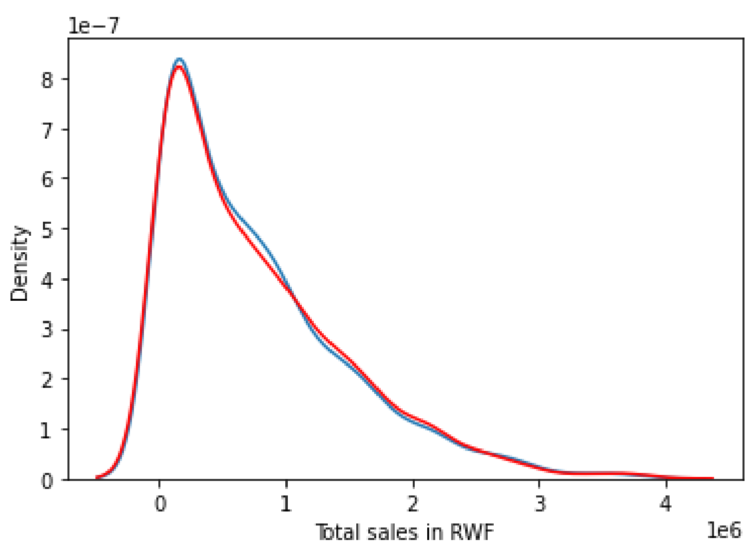

Figure 2.8 Multilayer Perceptrons

Final result of all models that has been tested

	models	MAE	MSE	R^2
5	Gradient Boosting Regressor	2.128112e+04	8.821185e+08	9.983000e-01
6	MLP	5.142109e+04	1.162331e+10	9.777000e-01
1	Random Forest Regressor	8.100529e+04	1.807638e+10	9.653000e-01
3	Decision Tree Regressor	9.702048e+04	2.193147e+10	9.580000e-01
4	Ridge Regressor	1.890592e+05	7.203789e+10	8.619000e-01
2	Lasso Regressor	2.044501e+05	8.631364e+10	8.345000e-01
0	Linear Regression	1.258878e+17	1.545519e+37	-2.963116e+25

Figure 2.9 Final result

This Article makes use of the GBR model stands for Gradient Boosting Regressor is a type of machine learning boosting. It relies on the intuition that the best possible next model, when combined with previous models, minimizes the overall prediction error. If a small change in the prediction for a case causes no change in error, then next target outcome of the case is zero

The GBR model is well suited for this article because of the following reasons:

- GBR is suitable for this article because it classified prediction problems where inputs are assigned a class or label
- GBR method is used to forecast the sales revenue of upcoming period. According to results there are high similarities between forecasted and actual data.
- GBR is suitable for regression prediction problems where a real-valued quantity is predicted given set of inputs

4. Conclusion

The main objective of this article was to create a management decision making tool, which would help supermarkets predict sales for future time periods in order to better plan their stock purchases, marketing campaigns and budgets for their financial year.Supermarket was used as the case study. Interviews and observation were the data collection techniques used which led to the following findings. There is a system in use which does not make any future predictions but instead works hand in hand with the stock management system to alert the management whenever the quantity of stock in store is too low, and therefore more stock needs to be purchased. After analyzing the existing system it was now time to design a prediction algorithm using python and tableau to implement the application. The Desired results were an online dashboard to analyze the sales dataset and prediction reports. This tool can be improved and even used in different businesses and not just supermarkets, as all businesses have to have decision making tools to help better plan for future financial periods. The accuracy of the prediction model is rather high and can be taken to market as is, several real life tests can be done on several supermarkets to test how well the predictions fair against real sales of the rime period the prediction model was meant to forecast. The cleaning process and modeling process require a data scientist to handle all these processes and therefore data scientists must always be at the help of this article to get things to run smoothly. In conclusion the desired results were achieved and the Sales Forecasting System was implemented and is running as expected.

5. References

Jie Hu, Li Shen, and Gang Sun (2017) Squeeze-and-excitation networks. arXiv
preprint arXiv

Andrew Ng (2016) Nuts and bolts of building ai applications using deep learning.

Hamilton, M. (1994). A Full Life Cycle Systems Engineering and Software Development Environment," cover story, Special Editorial Supplement,.

Hamilton, M. (2003). A Full Life Cycle Systems Engineering and Software Development Environment," cover story, Special Editorial Supplement,.

Hamlton. (2003). A Full Life Cycle Systems Engineering and Software Development Environment," cover story, Special Editorial Supplement,.

Daniel Kahneman and Patrick Egan(2011). Thinking, fast and slow, volume 1.
Farrar, Straus and Giroux New York.

Gerd Gigerenzer, Peter M Todd (1999) the ABC Research Group, et al. Simple
heuristics that make us smart.

https://www.forbes.com/sites/forbestechcouncil
three-ways-machine-learning-is-improving-the-hiring-process/
#4d9518c290e8.

Fake news challenge.

http://www.fakenewschallenge.org/.Arxiv sanity preserver.

http://arxiv-sanity.com. Wikimedia research: Detox.

https://meta.wikimedia.org/wiki/Research:Detox.

https://docs.aws.amazon.com/machine-learning/latest/dg/training-ml-models.html
https://mathworks.com/discovery/deep-learning.html

https://docs.aws.amazon.com/machine-learning/latest/dg/training-ml-models.html

https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/

https://en.wikipedia.org/wiki/Linear_regression