ISSN No:-2456-2165

Data Science and Machine Learning: Usage of Machine Learning Models for Forecasting to Improve Performance of Data Analytics in Non- Governmental Organization

Uwibambe Josiane [1]
Dr. Musoni Wilson[2] PhD,
[1] Master of Science in Information Technology (MICT), University of Kigali, Kigali, Rwanda

Abstract:- Estimating performance in relation to the expectation is a key component of many machine learning algorithms for decision-making. Measuring performance in accordance with expectations may not be very useful in many real-world situations. In this article, with deployment to a public dataset, we examine the viability and comparative analysis of Deep Learning techniques to anticipating the demand problem. We compare Deep Learning performance to that of various model approaches, such as Random Forest, Gradient Boosted Trees, and Support Vector Machine, using RMSE performance criteria. The forecasting issue is crucial for organizational decision-making. When making strategic decisions on valuable resources, risk-averse goals should be taken into account. This article aims to demonstrate the usage of ML models for forecasting and decision making to improve performance of data analysis of an organization. And to demonstrate that, especially when decision-makers are dealing with complicated limitations data, a Deep Learning algorithm can be a dominant answer to Machine Learning challenges for forecasting and decision-making.

I. INTRODUCTION

The forecasting problem is not novel. Science, engineering, and even commercial decision-making are all significantly impacted by forecasting. Demand forecasting and stock market forecasting are two specialized applications of forecasting methodologies. The better equipped the business will be, the more information there is about demand from product creation to production, logistics to sales. On the other side, if the predictions were wrong, there would be a risk of over or underproduction, poor service, or the simple sale of subpar goods. To at least maximize revenue, market prospects with the highest potential demand should be pursued. Obtaining the right amount of the commodities at the right time and place (production), calculating the right production volume and inventory (inventory), and optimizing the supply chain process for delivery are all important (supply chain). Additionally, it's crucial to uncover valuable prospects by leveraging long-term customer profiles to recognize and acquire prospects with similar traits.

The social sector's next big thing is now firmly established. International conferences, $25 million funding competitions, fellowships at esteemed universities, and initiatives introduced at Davos all revolve on machine learning today. In spite of this, it can be difficult to understand which social issues machine learning is best suited to solve, how businesses can use it practically to increase their impact, and what kinds of sector-wide investments are required to make ambitious uses of it for social good in the future possible. This is the point at which a non-government organization working on several projects with a social impact should start using machine learning models for forecasting.

II. METHODOLOGY

- *Data Analysis:*

Data analysis is the process that involve examining data, clean, transform, and model data with the goal of discovering useful information, drawing conclusions, and supporting decision-making. In several fields of business, science, and social science, data analysis has many features and methodologies, incorporating various techniques under a variety of titles. A specific type of data analysis approach called "data mining" concentrates on modeling and knowledge acquisition for predictive as opposed to just descriptive purposes.

A prediction study method was adopted to study the link between the independent variables and dependent variable. Each member of the sample provided at least two scores, one for each variable. This research approach will be suitable and fitting to our study because the researcher had to collect data based on current situation of to improve performance of data analytics in non-governmental organization.

- *Cleaning the Data:*

Data cleaning should be the initial step in any Data Science (DS) or Machine Learning (ML) methodology. Without clean data, it will be much more difficult to see the crucial components ofthe exploration. When training of ML models begins, they will become unnecessarily more difficult to train. The main point is that a dataset should be clean in order to maximize its potential.

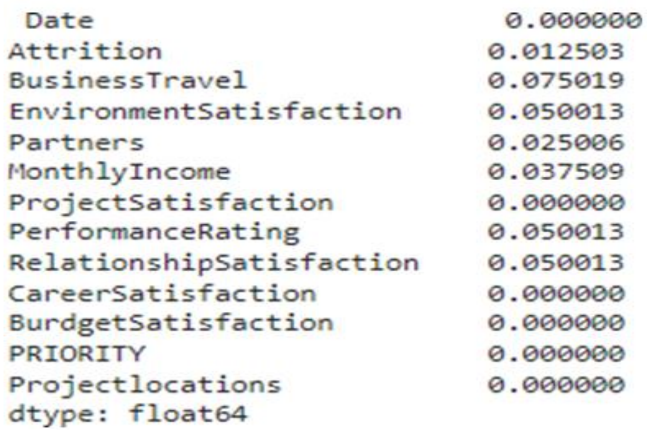
In data science and ML, data cleaning is the process of filtering and modifying data to make it easier to explore, understand, and model. Removing the parts, you do not want or need so you do not have to look at or process them modifying the parts that you do require but are not in the format that you require in order to use them properly. The dataset used needed the following changes to be considered clean:

Dropping of Rows with Null Values including NaN and null rows Removal of all negative numbers and replacing them with absolute valuesOne Hot Encoding

- *Null Values:*

The majority of Algorithms in data science do not accept null values (missing values). As a result, something must be done to eliminate them first or during data analysis. There are numerous methods for dealing with nulls. Which methods are suitable for a certain variable can be heavily influenced by the algorithms you intend to use, as well as statistical raw data patterns, particularly missing values and the randomness of their locations. Furthermore, Different methods could be suitable for various variables in a particular dataset. It is sometimes advantageous to apply multiple techniques to a single variable. Lastly, corrupt values are typically treated as nulls.

The figure1 below shows the rows in the dataset used in this project and the number of missing values in each row.



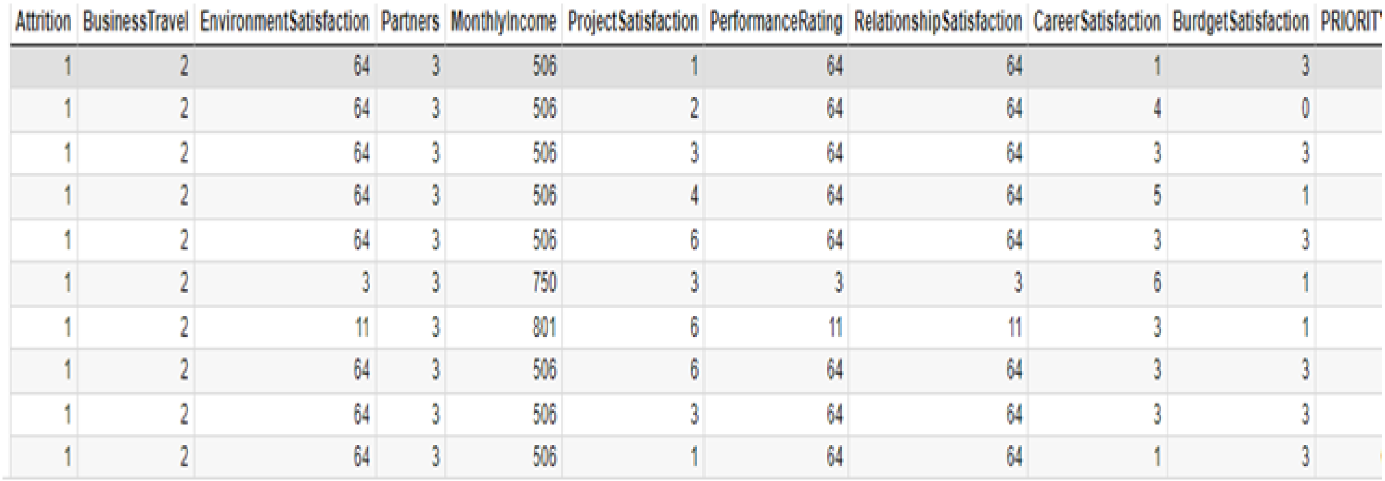
```
 Date                             0.000000
Attrition                        0.012503
BusinessTravel                   0.075019
EnvironmentSatisfaction          0.050013
Partners                         0.025006
MonthlyIncome                    0.037509
ProjectSatisfaction              0.000000
PerformanceRating                0.050013
RelationshipSatisfaction         0.050013
CareerSatisfaction               0.000000
BurdgetSatisfaction              0.000000
PRIORITY                         0.000000
Projectlocations                 0.000000
dtype: float64
```

Fig 1 Null Rows Dataset

It is clear that the value 0 (all bits at zero) is a typical value used in memory to denote null. It means that there is no absence of data or simply in these an impact on the algorithm implemented. So, the best option is to remove all the rows with Null Values as this is a big dataset and removing these rows will not have too much of an effect on the algorithm implemented.

- *One Hot Encoding:*

A one hot encoding is a categorical variable represented as binary vectors. In order to do this, the categorical values must first be converted to integer numbers. After that, each integer value is represented as a binary vector with all of its elements being zero except for the integer's index, which is denoted by a 1.



Attrition	BusinessTravel	EnvironmentSatisfaction	Partners	MonthlyIncome	ProjectSatisfaction	PerformanceRating	RelationshipSatisfaction	CareerSatisfaction	BurdgetSatisfaction	PRIORIT'
1	2	64	3	506	1	64	64	1	3	
1	2	64	3	506	2	64	64	4	0	
1	2	64	3	506	3	64	64	3	3	
1	2	64	3	506	4	64	64	5	1	
1	2	64	3	506	6	64	64	3	3	
1	2	3	3	750	3	3	3	6	1	
1	2	11	3	801	6	11	11	3	1	
1	2	64	3	506	6	64	64	3	3	
1	2	64	3	506	3	64	64	3	3	
1	2	64	3	506	1	64	64	1	3	

Fig 2 After Applying One Hot Encoding

- *Feature Selection:*

The effectiveness of a machine-learning model for a given job is influenced by a variety of factors. The choice of features is one of the fundamental ideas that significantly affects how well a model performs. Performance of a machine learning model is significantly influenced by the characteristics of the data used to train it. This is due to the negative impact irrelevant characteristics have on model performance. By eliminating redundant data, a feature selection approach decreases overfitting, enhances predictor performance, speeds up training, and increases model correctness. The method below was employed in this work for feature selection.

- *Data Visualization:*

Data visualization is the graphical display of data and information. Data visualization tools offer an easy approach to observe and analyze trends, outliers, and patterns in data by utilizing visual elements like charts, graphs, and maps. To analyze vast volumes of data and make data-driven decisions, data visualization tools and technologies are crucial in the world of big data.

➢ *Multivariate Plots:*

Designed to simultaneously reveal the link between several variables. There are a few fundamental properties of the relationship among sets of variables that are of interest, just as there were when looking at correlations among pairs of variables.

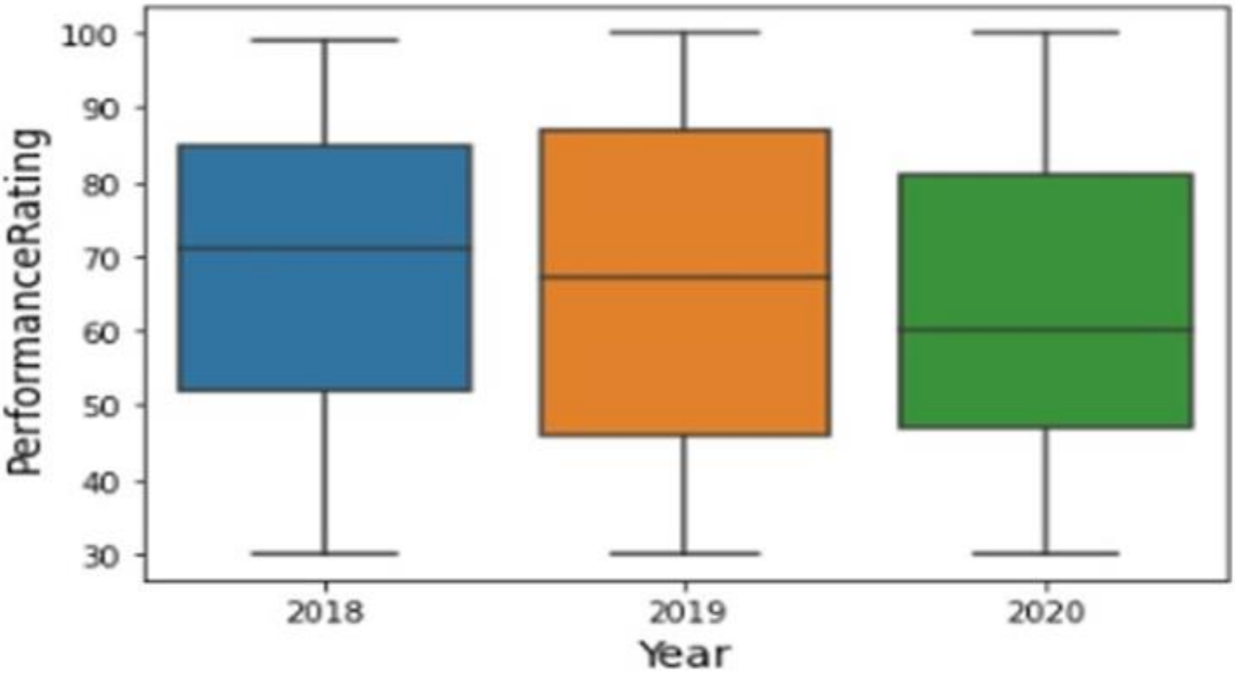


Fig 3 Graph of Year Vs Perfogrmance Rating in 3 years

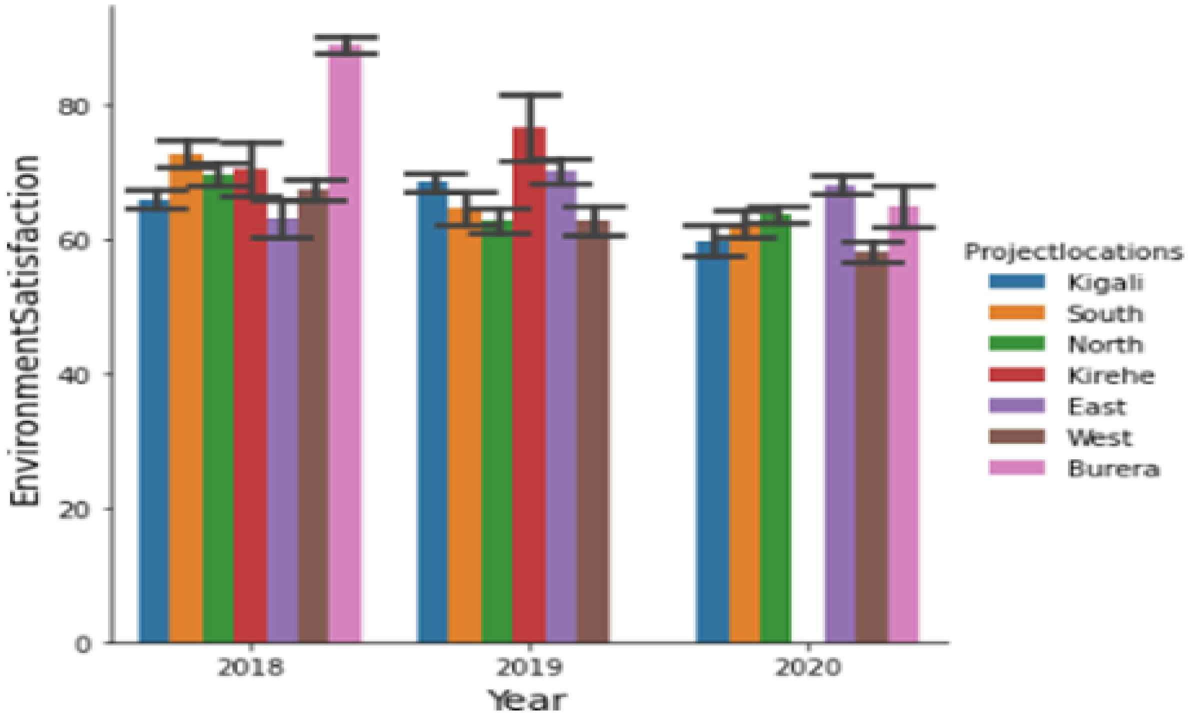


Fig 4 Graph year vs Environment satisfaction with project location

III. RESULT

➢ *Machine Learning Models*

Dataset was firstly divided into two variables X as the features we defined earlier and y as the consumer the target value for prediction.This is a regression problem so we will use Regression methods.Train test split will be a 8:2 ratio respectively.

- *Machine Learning Model Used:*
  Linear Regression, Random Forest Regressor, Lasso Regressor, Gradient Boosting Regressor, Decision Tree Regressor and Ridge Regressor.

- *The Process of Modeling the Data:*
  Importing the model, Fitting the model, Regression metrics.

Regression score metrics: The mean of the absolute value mistakes is known as the mean absolute error (MAE) (absolute distance from true value). Mean Squared Error (MSE): The average of errors' squared values (squared distance from true value). R2 (coefficient of determination) is the score function for regression.

➢ *Linear Regression:*

For modeling the link between a scalar answer (or dependent variable) and one or more explanatory variables, statisticians use linear regression (or independent variables).

➢ *Future Warning:*

- MAE: 0.02
- MSE: 0.21
- R^2: 0.9995

Table 1 Predicted Linear Regression

Actuals Values	prediction
7	70
2.2	22
1.6	16
5.5	55
6.9	69
1.1	11
5.4	54
3.9	39
1.3	13
3.2	32

➢ *Random Forest Regressor:*

Random forest is an ensemble learning-based supervised learning technique for regression and classification. A 5-fold time series cross-validation strategy was used to train the dataset, with 80% of the data being used for training and 20% being used as the test set. Performances were assessed using the metrics MAE, R 2, and mean squared error MSE. The results obtained by the Random Forest Regressor are as follow:

- MAE: 0.47
- MSE: 0.37
- R^2 score: 0.9991

Table 2 Actual and Predicted Random Forest Regressor

Actual	Predicted
68.774915	70
22.257410	22
15.705080	16
55.306883	55
68.818260	69
10.502869	11
53.713084	54
1.760926	0
12.608953	13
32.595029	32

➢ *Graph Of Random Forest Regressor:*

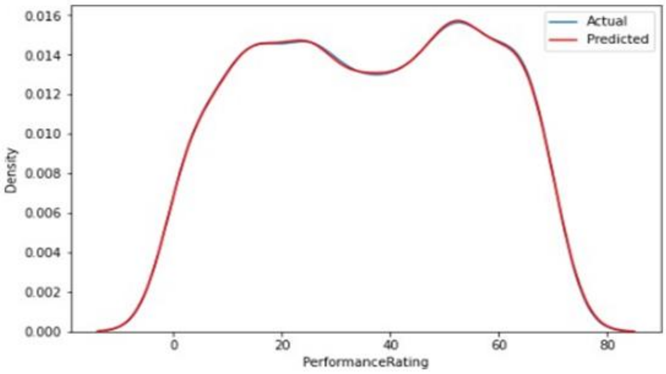


Fig 5 Graph of Random Forest Regressor

➢ *Lasso Regressor:*

Lasso (least absolute shrinkage and selection operator); is a technique for regression analysis that enhances the predictability and interpretability of the statistical model it produces through variable selection and regularization. 20% of the data was utilized as the test set, while the remaining 80% was used to train the model using a 5-fold time series cross-validation strategy. The performance was assessed using the metrics MAE and MSE.

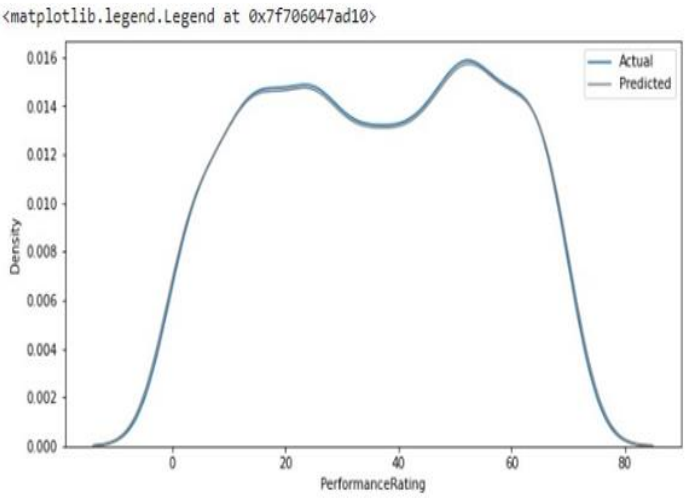
Results obtained by the Ridge Regressor are as follow:

- MAE: 0.15
- MSE: 0.03,
- R^2 score: 0.9999

Table 3 Actual and Predicted Lasso Regressor

Actuals	Predicted
69.709205	70
22.120236	22
16.171615	16
54.837652	55
68.717768	69
11.214431	11
53.846215	54
0.308625	0
13.197304	13
32.034605	32

- *Graph Of Lasso Regressor:*



```
<matplotlib.legend.Legend at 0x7f706047ad10>
```

Fig 6 Actual and Predicted Lasso Regressor graph

➢ *Decision Tree Regressor***:**

In Decision Tree Regressor, a 5-fold time series cross-validation strategy was used to train the dataset, with 80% of the data being used for training and 20% being used as the test set. Performances were assessed using the metrics MAE and MSE. The following are the results obtained by the Decision Tree Regressor:

- MAE: 0.54
- MSE: 0.36
- R^2 score: 0.9991

Table 4 Actual and Predicted Decision Tree Regressor

Actuals	Predicted
68.802974	70
22.003984	22
15.606272	16
55.530488	55
68.802974	69
10.673469	11
0.576642	0
12.734767	13
32.417910	32

- *Graph of Decision Tree Regressor:*

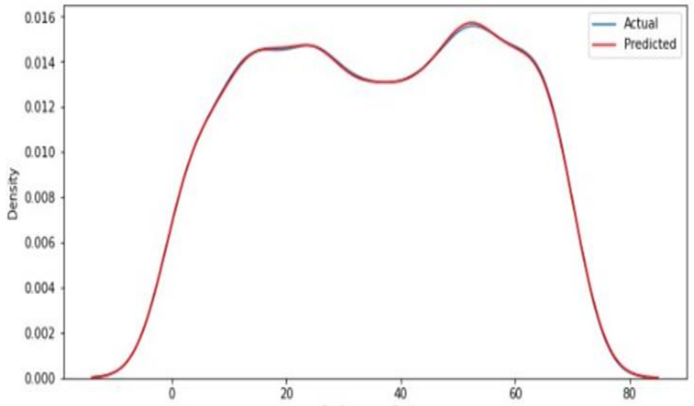


Fig 7 Decision Tree Regressor

➢ Ridge Regressor:

Ridge regression is a method of model tuning applied to multicollinear data analysis. With this approach, L2 regularization is accomplished. Least squares are unbiased and variances are high when there is a multicollinearity issue, leading to projected values that are greatly off from the actual values.

- MAE: 0.68
- MSE: 0.7
- R^2 score: 0.9983

Table 5 Actual and Predicted Ridge Regressor

Actuals	Predicted
70	68.080802
22	22.454124
16	16.672250
55	54.158046
69	67.744628
11	11.828008
54	53.396039
0	2.002357
13	13.764932
32	32.116512

- *Graph Of Ridge Regressor:*

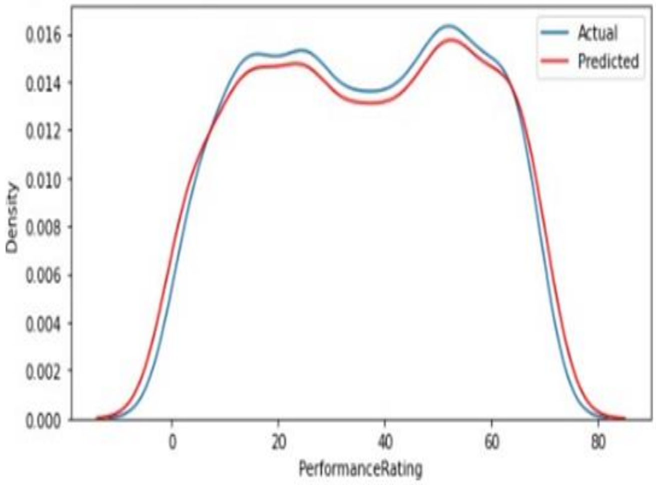


Fig 8 Graph of Ridge Regressor

➢ Gradient Boosting Regressor:

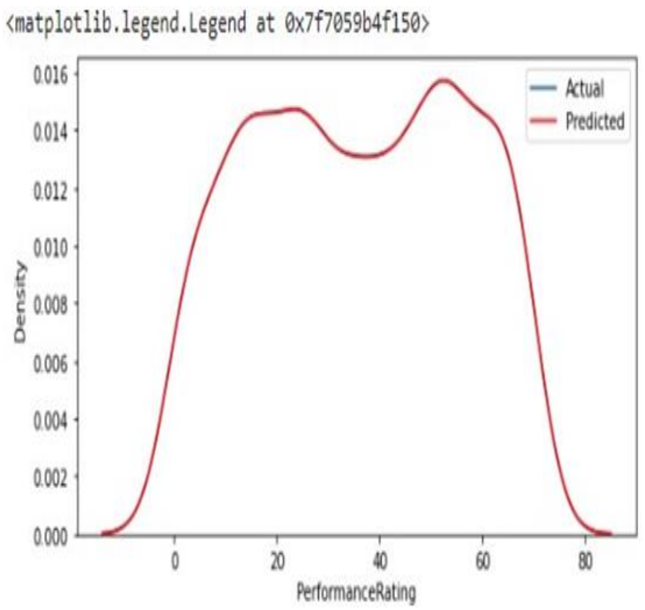
Gradient boosting is a machine learning method that is used to do tasks like classification and regression; among other things, it returns a prediction model in the form of a group of weak prediction models, typically decision trees. Gradient Boosting Regressor is trained using a 5-fold time series cross-validation approach, where 80% of the data is utilized for training and 20% is used as the test set. Performances are assessed using the metrics MAE, R2, and MSE. The following are the results obtained.

- MAE: 0.0
- MSE: 0.0
- R^2 score: 1.0

Table 6 Actual and Predicted Gradient Boosting Regressor

Actual	Predicted
69.989911	70
21.999743	22
16.000673	16
55.000637	55
68.994912	69
11.001352	11
53.999160	54
0.010568	0
13.000602	13
32.010449	32

- *Graph Of Gradient Boosting Regressor:*



```
<matplotlib.legend.Legend at 0x7f7059b4f150>
```

Fig 9 Graph of Gradient Boosting Regressor

- *A Multilayer Perceptron (MLP):*

The MLP model stands for Multilayer Perceptron is feed-forward artificial neural network (ANN) where the information runs from the input layer towards the output layer through the hidden layer. The activation function for the Multilayer Perceptron algorithm is the Rectified Linear Unit. Backpropagation is a supervised learning method used by MLP to train the network. The fault spreads throughout the network backward during backpropagation. The error is determined by subtracting the network output from the actual output. Based on this methodology, the network's weights parameters are changed to reduce this inaccuracy. A stopping condition is reached after multiple iterations of this operation.

- MAE: 0.05
- MSE: 0.01
- R^2 score: 1.0

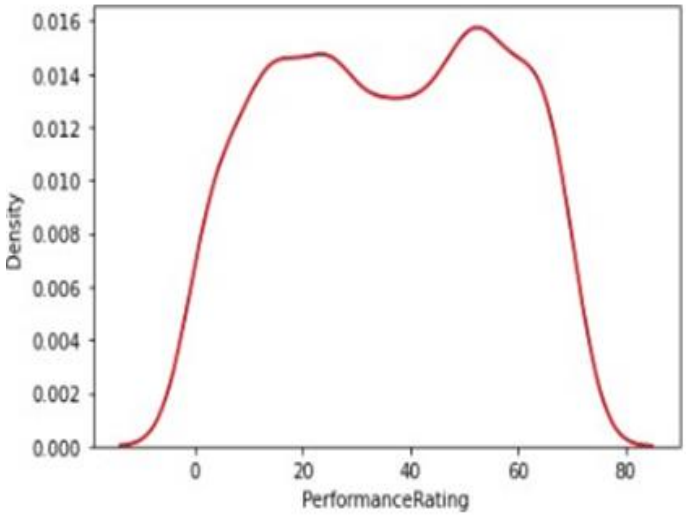
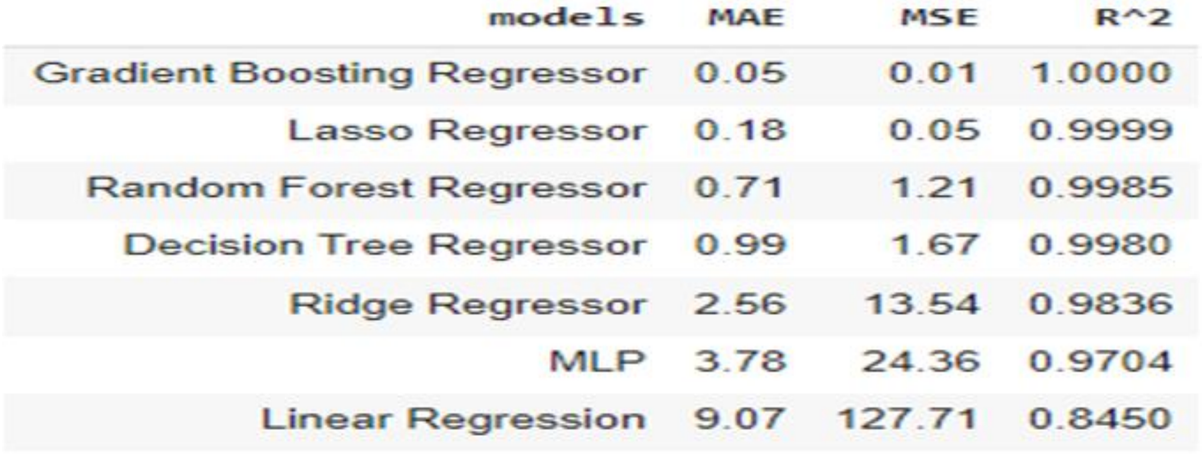


Fig 10 Graph of MLP

IV. COMPARATIVE STUDY OF THE FITTING MODELS BASING ON METRICS

- *Table Metrics (MAE, MSE and R^2):*

Table 7 Table Metrics (MAE, MSE and R^2)



models	MAE	MSE	R^2
Gradient Boosting Regressor	0.05	0.01	1.0000
Lasso Regressor	0.18	0.05	0.9999
Random Forest Regressor	0.71	1.21	0.9985
Decision Tree Regressor	0.99	1.67	0.9980
Ridge Regressor	2.56	13.54	0.9836
MLP	3.78	24.36	0.9704
Linear Regression	9.07	127.71	0.8450

- *Mean Absolute Error:*
  5-fold time series validation tests using MAE from regression models

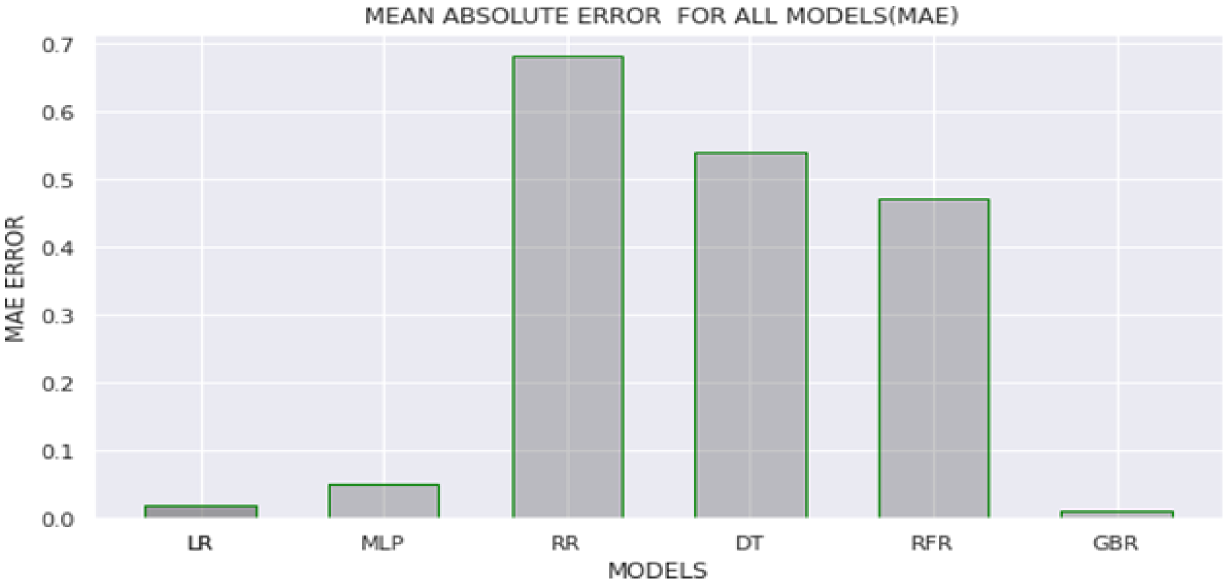


Fig 11 MAE Obtained from Regression Models on 5-Fold Time Series Validation Tests

Figure represents the Mean Absolute error from the results of the predictions produced by the Linear Regression, Random Forest Regressor, Regressor Lasso Regressor, Gradient Boosting, Decision Tree , Ridge Regressor, A multilayer perceptron (MLP) on 5-fold time series cross- validation tests. From the figure, Linear Regression has the highest; Gradient Boosting Regressor, has least MAE for the folds and thus can be said as a best-performed algorithm. RidgeRegressor has the highest MAE and thus it can be said as worst performer.

- *Comparison of R^2 Obtained By Regression Models:*

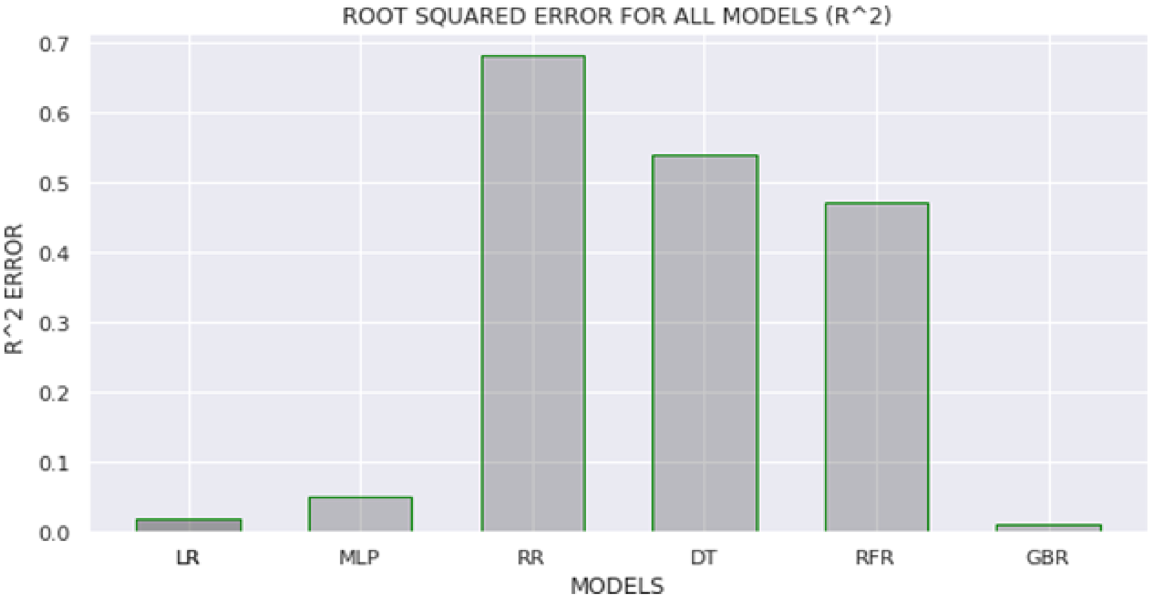


Fig 12 Comparison Of R^2 Obtained By Regression Models

From figure, it can say that Gradient Boosting and MLP are the best performer with the lowest error and Ridge Regression is the worst performer with the highest error. From figure 22, it can be noticed that Still Gradient boosting has shown spectacular overall performance with the least R^2 with the least value and Rigde Regressor has shown worst performance with the highest R^2value 0.9983.

- *Comparison of MSE obtained by regression models:*

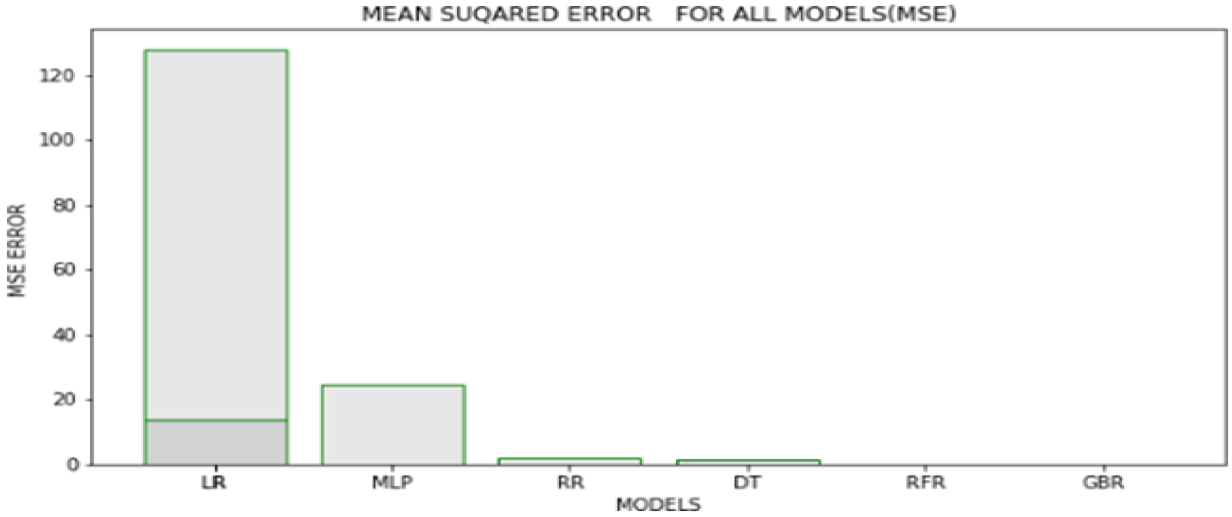


Fig 13 Comparison of MSE Obtained by Regression Models

From figure, it can be noticed that Still Gradient boosting has shown spectacular overall performance with the least *MSE* with the least value (0.00) and Ridge Regressor has shown worstperformance with the highest *MSE* with the value (0.70)

- *Key Concepts Behind the Performance of Algorithms are Explored as Follows:*

- Gradient Boosting has performed significantly well compared to the other models. This may be because of the regularization approach where the variance is reduced at the cost of some bias initiation, which makes it robust to outliers and overfitting.
- Random Forest Regressor has performed surprisingly well compared to the Linear Regression, Decision Tree regressor and lasso regressor; this may be because of its generalization capability. Kernel function for the chosen parameters is set to ‘Rbf’, which means Radial basis function (RBF) is a function whose value changes depending on how far away a center is. Other kernel functions like linear, radial may produce better results.
- Random Forest Regressor has not shown good performance, this may be because of overfitting problem, which may be preventing from generalizing the model.
- Ridge Regression produced bad results compared to others because of overfitting problem and Ridge Regression is harder to tune compared to other models results.
- Ridge Regressor is the worst Model for this analysis, since for both the values of R^2 equals to 0.9983, MAE equals to 0.68 and MSE Equals to 0.70.

- *Comparison of Performance Evaluation Results:*

This project makes use of the GBR model stands for Gradient Boosting Regressor is a type of machine learning boosting. It is based on the assumption that the best next model, when combined with previous models, minimizes the overall prediction error. If a small change in a case's prediction results in no change in error, the case's next target outcome is zero.

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

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

V. CONCLUSION

From the study results, it was found that the ability of machine learning models to analyze and learn from real-time data and historic pact Rwanda dataset helps management of pact Rwanda to predict performance and enhanced productivity, the study results could help a non-government organization better prediction performance. 5. 0% of global Organization should use Artificial intelligent and machine learning, implying that performance efficiency must be greatly increased. The Non-governmental organization must embrace AI-driven performance as the solution. An efficient performance is critical to organization success in today's competitive world. Every day, disruptive technologies such as AI and ML play an important role in making it better.

REFERENCES

[1]. Boehmke, Bradley. (2019). "Gradient Boosting". . Hands-On Machine Learning with R. Chapman & Hall.

[2]. Brownlee. (2020). Machine Learning Algorithms. machine learning mastry.Burns, E. (2022). techtarget.

[3]. Hope, C. (2021). Overview of the Python 3 programming language. Computer Hope.

[4]. Jason, B. (2018). A Gentle Introduction to the Bootstrap Method. Machine learning master.

[5]. michael. (2019). Further Regression Algorithms.

[6]. Pedamkar, P. (2020). Machine Learning vs Neural Network. EDUCAB.

[7]. Rankish, K., & Ramsey, W. M. (2014). The Cambridge handbook of artificial intelligence. p.337.

[8]. Singhal, S. (2021). Defining, Analysing, and Implementing Imputation Techniques. analyticdyhya.

[9]. Twumasi. (2021). Machine learning algorithms for forecasting and backcasting. international journal, Pages 1258-1277.

[10]. Umar Ishfaq et al. (2022). Empirical Analysis of Machine Learning Algorithms for Multiclass Prediction.

[11]. Bradshaw, P., Murray, V., Wolpin, J. (2010), Do Nonprofit Boards Make a Difference? An Exploration of the Relationships Among Board Structure, Process, and Effectiveness, Nonprofit and Voluntary Sector Quarterly. 120 THE PERFORMANCE OF NON-GUVERNMENTAL ORGANIZATIONS

[12]. Brown, A. W. (2015), Exploring the Association Between Board and Organizational Performance in Nonprofit Organizations, Nonprofit Management & Leadership, vol. 15

[13]. P., Holland, T. P., Taylor, B. E. (2019), The Effective Board of Trustees New York: Macmillan.

[14]. Ciucescu, N. (2014), The Social Performance Of Non-Governmental, Proceedings of the Annual Session of Scientific Papers ”ÏMT Oradea”-2014

[15]. Green, J. C., Griesinger, D. W. (2015), Board Performance and Organizational Effectiveness in Nonprofit Social Service Organizations, Nonprofit Management and Leaderships.

[16]. Herman, R. D., Renz, D. O. (2014), Thesis on Nonprofit Organizational Effectiveness, Nonprofit and Voluntary Sector Quarterly.

[17]. Jackson, D. K., Holland, T. P. (2018), Measuring the Effectiveness of Nonprofit Boards, Nonprofit and Voluntary Sector Quarterly.

[18]. Medina-Borja, A.,Triantis, K. (2017), A conceptual framework to evaluate performance of nonprofit social service organizations, International Journal of Technology Management, vol. 37.

[19]. Străinescu, I., Ardelean B. (2017), Managementul ONG, Editura Didactică şi Pedagogică, Bucureşti. Verboncu, I. (2016), Management, eficiență și eficacitate, Management & Marketing, The official journal of the Society for Business Excellence.

[20]. Vlăsceanu, M. (2013), Organizaţii şi comportament organizaţional, Editura Polirom, Bucureşti.