GDP Growth Prediction of Countries Using Machine Learning Algorithm

M. Priya[1], Subhashini Al[2]

[1]Assistant professor, [2]Master of Engineering
[1, 2]CSE Tagore Institute of Engineering and Technology, Salem, India

Abstract

The main objective is to predict GDP Growth by help of other parameters like GDP Per Capita, Inflation Rate, Government Debt, Total Investment, Remittance, Unemployed Rate. The complex relations are obtained by machine learning algorithm among GDP Growth Rate and other parameters to predict GDP Growth Rate that may help everyone to get connected to the field of economy and also to the economist to demonstrate their prediction about the economy. With help of this it is easy to find out the possible way to improve the desire growth of GDP. This project can help to demonstrate our eco- social scenario of future. This project can help to set economic goals for our country and can find out which parameters are most directly related to our GDP Growth and which are less related to our GDP Growth and which are accountable for reducing our GDP Growth. For any country GDP growth is a very important think to follow up. This project will give the analyzed data and we will get proper information to take certain action to keep the growth in higher rate. Through this system it is possible to achieve accurate information about the GDP Growth.

Keywords: GDP Growth, ML Algorithms, Relation, Prediction, Complex, Parameters.

I. INTRODUCTION

A countries development is highly dependent on the growth of GDP of the country. Predicting the GDP can help the government or the authority to take right action and decision to continue the economy development of the country. Policy makers of government and economists need to realize the phase of economy to make best possible policy decisions. To make those decision they need to take count previous and present economic conditions. Gross domestic product (GDP) is the main meter for knowing the performance of our economy. It helps our economist and policy makers to determine whether our economy is expanding or contracting and enables them to make policy decisions. In this paper, we t ry to find out more accurate GDP predict according to Machine learning algorithm. We explore which model of Machine learning algorithm best fit to predict our GDP growth rate. We use last 40 years data to find out which Machine learning model is best fit for our prediction and which independent data/features/parameter are most related to our GDP growth.

In this paper, we work on the use of machine learning algorithm to predict the situation of the GDP growth considering other variables more the to find out best independent variable and best ML algorithm which helps us predict GDP growth of Bangladesh and try to understand how other variable impact out GDP growth. There is always a lake of communication and research about the economic growth. Economist cannot give useful instruction due to lake of data analysis. So here we are trying to help the economist and the government about the growth of the GDP. Our goal is to make economists comfortable with machine learning, show how machine learning can help develop economic theory. Test the limits of machine learning predictive ability and make economics more policy conscious by using machine learning. This project will help the economist and the economical authority to analyze the economic growth by predicting the GDP. In a country like Bangladesh economic state has depended on many factors due to the politica l conflicts, natural disasters hampering agricultural growth, imbalance in the national budget and many more. As a developing country Bangladesh needs a stable economic growth to achieve its goal to become a developed country. To ensure the economic stability economist must help the government and authorities about the current state of economy and the growth that leads to the future. But there is always a lake of communication and research about the economic growth.

II. LITERATURE REVIEW

There are no similar work or research was done which can predict GDP growth using machine learning in Bangladesh. So the background is the past and current situation of economy and the use of machine learning to predict the GDP growth of Bangladesh.

Adam Richardson et al. [1] check something out that acquiring accurate predict of real GDP growth for New Zealand, using common Machine Learning algorithms on real time dataset. The predictions obtained from these models are then compared with the predictive accuracy of a naive autoregressive benchmark and other data-rich approaches such as a factor model, a large Bayesian VA R and the combined GDP now casts obtained from the RBNZ's series of statistical models. They evaluate that maximum of machine learning models are capable of to e xhibit more reliable forecast than AR and other statically standard. The results also suggested that in combining individual ML forecasts, there are some benefits. Our results therefore suggest the use of ML algorithms as an alternative to GDP nowcasting models from a forecaster's suite.

Karim Barhoumi et al. [2] study sought to nowcast GDP growth in Sub-Saharan Africa using machine learning methods. The findings from this study indicate that machine learning methods, including elastic net, support vector machine, and random forest, are able to produce superior nowcasts compared to traditional regression methods. In context of the COVID-19 pandemic, it is hoped that the models used in this study can provide policymakers with a means to track the current state of economic activity as the situation rapidly evolves. A ne xt step in the research agenda is to assess the informational content of more f inancial indicators as well as employ data on google search trends and from satellite images in tracking economic activity in Africa.

In this paper [3], work on to achieved a machine learning framework. Apart from creating this framework, we also envisage this book as a sort of primer for using Machine Learning to answer economic questions . While Machine Learning itself is not a new idea, advances in computing technology combined with a dawning realization of its applicability to economic questions makes it a new tool for economists (Va rian, 2014)[10]. They identify the research questions and issues in section 2 and state our proposed methodologies in section 3. Section 4 describes our data as well as some of the issues in our dataset. Sections 5 and 6 present our results and some concluding thoughts about policy implications and avenues for future research.

In this paper [4], design to indicate that XGBoost has the potential to become a useful tool for macroeconomic prediction.

We also conclude that Google t rends data can be a suitable alternative to official data for predicting the monthly Canadian real GDP growth. We argue that while GT cannot replace official data for predicting GDP growth, it permits us to predict the monthly and quarterly GDP growth ahead of the release of the official figures with a substantial degree of accuracy.

Machine learning models are sometimes criticized as black-box models as they do not lend themselves to interpretation for policy-making purposes. We try to address these issues by including tools that can make machine learning more easily interpretable. To this end, we provide partial dependence plots, variable importance plots, and SHAP values. We show that these interpretable machine learning methods all point to the same predictors as being the most important ones for forecasting the Canadian RGDP growth rate. Thus, in the case of official data, they all select CA employment, CA e xpo rt, CA retail, trade, the shipment of all manufacturing, US shipment manufacturer and US industrial production as the most important predictors of GDP growth, while in the case of Google trends data they all select finance, roleplaying games, car electronics, and property development as the crucial features. Interestingly, we also find that principal component analysis can track the target variables quite well when using only “good” features. We will address the issue of constructing forecast intervals for XGBoost in our future research.

In this paper [5], by the help of machine learning algorithms and traditional time series regression models Pirasant Premraj design a catholic comparison of forecasting Gross Domestic Product (GDP) on the following economies: Australia, Japan,USA, Ge rmany, Euro Area, France, Sweden, Spain, Great Britain and Canada. The ML algorithms they employ some algorithms and these are Bayesian Additive Trees Regression Trees(BART), Stochastic Gradient Boosting (GBM) , Elastic-Net Regularized Genera lized Linear Models (GLMNET) and eXtreme Gradient Boosting (XGBoost) For present traditional time series regression methods they use Autoregressive (AR) models, Autoregressive Integrated Moving Average (ARIMA) models and Vector Autoregressive (VA R) the result shows that the multivariate VA R model are superior that indicates the picked variables and the model suitability of predicting GDP Growth.

In this paper [6], work on XGBoost algorithm where they describe a scalable end to end tree boosting system called XGBoost. XGBoost can scales beyond billions of e xa mples using far fewer resources than existing systems. In this paper[7], Spyros Makridakis, Evangelos Spiliotis try to improving forecasting accuracy by minimizing some loss function(like: sum of square error). In this paper [8] Shafiullah Qureshi, Ba M Chu and Fanny S. Demers showed the use state-of-the-art machine-learning (ML) algorithms to predict using both Google trends (GT) and official trends (GT), the monthly and quarterly actual GDP growth of Canada. Data that is available prior to the release by Statistics Canada of GDP data. In this paper [9], Rickard Nyman and Poul Ormerod train algorithm to predict potential to give early warning of recessions.

In this paper [10] showed major machine learning models for time series forecasting, a large scale comparison analysis. Specifically, on the monthly M3 time series competition data, they applied the models (around a thousand time series). In this paper [11], used a broad range of monthly indicators to perform a comprehensive comparison of the short-term forecasting skills of twelve statistical models and skilled analysts in a pseudo-real-time environment. In this paper [12], Leo Beriman try to monitor internal error, correlation and strength to show the response to the increasing number of features used to in the splitting.

III. PROPOSED METHODOLOGY

Throughout the proposed model, we are trying to find out which parameters of our dataset are closely related to our dependent parameter. To define this relation we follow several steps:

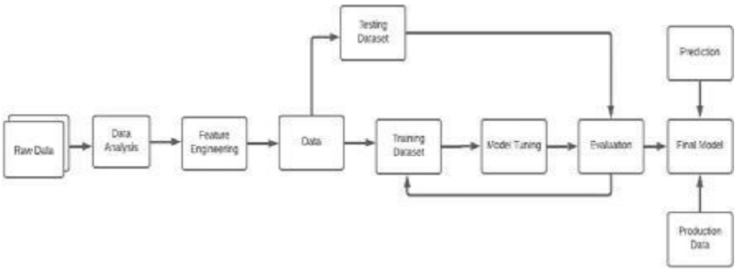
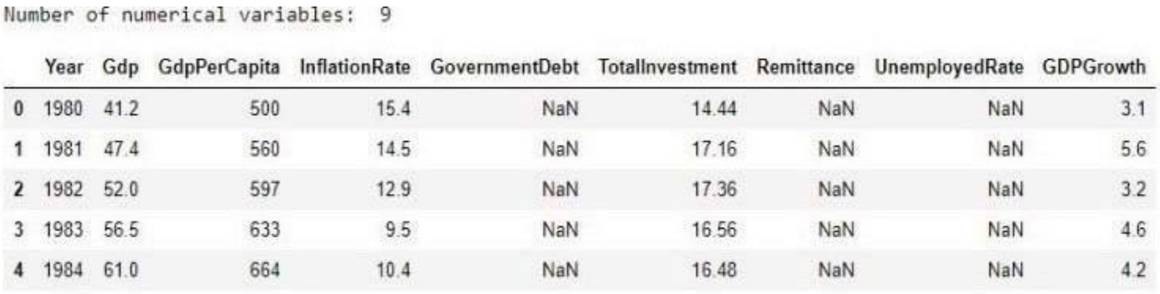


Figure 1.Methodology block diagram

A. Raw Data

Collection of our research dataset from ‘Kaggle.com’[9] where our dataset have 40x9 e xa mple vectors. In this dataset, there are last 40 years data of various parameters of bd economy which have 9 features and every feature have 40 columns. And there are 9 features in dataset and these are: ‘Yea r’, ‘GDP’, ‘GdpPerCapita’, ‘InflationRate’, ‘Govern mentDebt’, ‘TotalInvrstment’, ‘Remittance’, ‘UnemployedRate’, ‘GDPGrowth’. Since, we want to predict ‘GDPGrowth’ rate thus, ‘GDPGrowth’ is our dependent features and other are independent features.



```
Number of numerical variables:  9
```

	Year	Gdp	GdpPerCapita	InflationRate	GovernmentDebt	TotalInvestment	Remittance	UnemployedRate	GDPGrowth
0	1980	41.2	500	15.4	NaN	14.44	NaN	NaN	3.1
1	1981	47.4	560	14.5	NaN	17.16	NaN	NaN	5.6
2	1982	52.0	597	12.9	NaN	17.36	NaN	NaN	3.2
3	1983	56.5	633	9.5	NaN	16.56	NaN	NaN	4.6
4	1984	61.0	664	10.4	NaN	16.48	NaN	NaN	4.2

Figure 2. Dataset example

Table I: Dataset

B. Data Analysis:

i.Missing values

After collecting these data we find out the missing values of our dataset and then try to check the relation between missing data and target data. The result shows us relation with target data. That relation can define as proportionally related with target data that means our ‘GDPGrowth’ is less where null value are present.

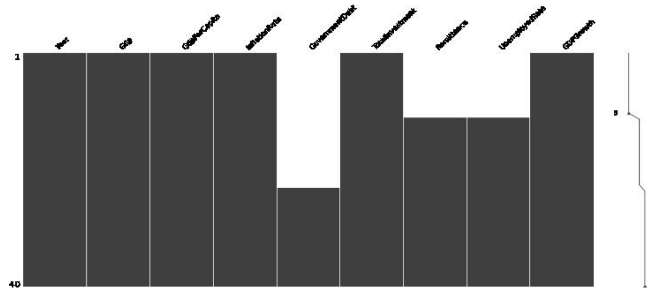


Figure 3. Graphical presentation of missing values

ii. Numerical Variables

After that, we check how many numerical values in our dataset. Then we divide all features into two categories one is, discrete and continuous.

We got all features are numerical and in these features 7 are continuous (‘GDP’, ‘Inflation Rate’, ‘Government Debt’, ‘Total Investment’, ‘Remittance’, ‘Unemployed Rate’, ‘GDP Growth’) and left features are discrete(‘Year’, ‘Gdp Per Capita’).

iii. Outliers

The next step is we check out outliers through box plotting thus, we can reduce distance of R^2.

ରl=Ƭ− σ(−)ରlσ(−

)ରl

Then we find out some outliers In Remittance GDP Growth and Inflation rate. For the sake of good prediction we remove outliers from Remittance and the others are negligible.

iv. Relation Between Independent And Dependent Feature

In this step, we try to capture the best parameters for our machine learning model. First we try to find out which are the parameters is highly related to our GDP Growth rate then, we see Total Investment, GDP, GDP Per Capita is positive and highly related to our GDP Growth. And Govern ment Debt and Inflation Rate are inversely related to GDP Growth.

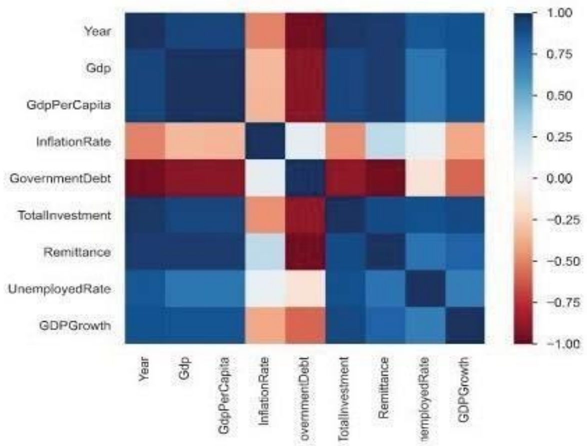


Figure 4. Presentation of correlation

After then, we can see Year, GDP, GDP Per Capita are internally highly correlated so, we don’t need to take all three as independent parameters for our machine learning model.

C. Feature Engineerning

i.Replaching Missing Value

End the all these process we calculate median value of that feature than null value were replaced by median.

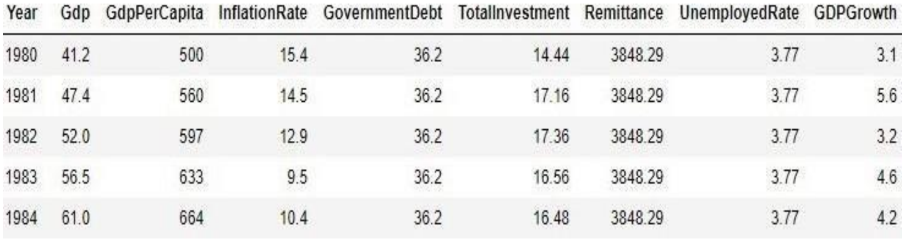
Ȁ ◌|◌| ф ꝗ 1 –

Median = (◌|◌|1)Ȁត or, ត

After replacing all missing value by median the output is:

```
GovernmentDebt      0
Remittance          0
UnemployedRate      0
dtype: int64
```

and our dataset look like:



Year	Gdp	GdpPerCapita	InflationRate	GovernmentDebt	TotalInvestment	Remittance	UnemployedRate	GDPGrowth
1980	41.2	500	15.4	36.2	14.44	3848.29	3.77	3.1
1981	47.4	560	14.5	36.2	17.16	3848.29	3.77	5.6
1982	52.0	597	12.9	36.2	17.36	3848.29	3.77	3.2
1983	56.5	633	9.5	36.2	16.56	3848.29	3.77	4.6
1984	61.0	664	10.4	36.2	16.48	3848.29	3.77	4.2

Figure 7. Dataset without missing values

ii. Features Scaling

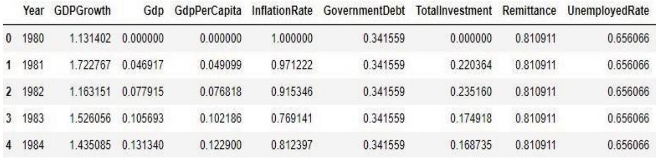
Since, our dataset is highly varying values or unit thus, we use this technique to standardize the independent parameters in dataset in a fixed range. We Min-Max Normalization (7) technique to scaling our data.

$$= \frac{-\min(\ )}{\max(X)-\min(X)} \quad \ldots\ldots\ (7)$$

w

D. Dataset

After Features Engineering, our dataset is ready for training. Since, we did scaling on our dataset now our dataset is scaled in 0 to 1.



	Year	GDPGrowth	Gdp	GdpPerCapita	InflationRate	GovernmentDebt	TotalInvestment	Remittance	UnemployedRate
0	1980	1.131402	0.000000	0.000000	1.000000	0.341559	0.000000	0.810911	0.656066
1	1981	1.722767	0.046917	0.049099	0.971222	0.341559	0.220364	0.810911	0.656066
2	1982	1.163151	0.077915	0.076818	0.915346	0.341559	0.235160	0.810911	0.656066
3	1983	1.526056	0.105693	0.102186	0.769141	0.341559	0.174918	0.810911	0.656066
4	1984	1.435085	0.131340	0.122900	0.812397	0.341559	0.168735	0.810911	0.656066

Figure 5. Scaled dataset

E. Features Selection:

We know feature selection is most core idea in machine learning that very impactful on the performance of machine learning model. Thus,

Irrelevant or partially relevant features can negatively impact model performance

Now, before select our model, we need to know which features are most related to our dependent feature. For feature selection we implement Lasso Regression (8).

Residual Sum of Squares + λ * (Sum of the absolute value of the magnitude of coefficients)

Lasso Regression reject these parameters which slopes are nearer to zero. Then we perform selectFromModel object from sklearn for selecting these nor-zero coefficient features. And these are Year and Total Investment that means our dependent features 'GDPGrowth' is highly dependent on 'Year' and 'Total Investment'.

F. Training Dataset

Since, we need to train our model thus, we separated our features into to group. One is y_train where we capture dependent feature 'GDP Growth' and the other is X_train where we captures all independent features such as 'Year', 'Government Debt', 'GDP', 'GDP per Capita', 'Inflation Rate', 'Total Investment', 'Remittance', 'Unemployed Rate'

After features scaling us find two independent features and we put this two features in X and in y we capture 'GDP Growth'.

Now we split our data set into training and testing using sklearn.model_selection.**train_test_split**
We use those training data for model tunning and use test data for observing performance of model.

G. Selecting Alpha value

We perform various alpha value for selecting best features for our dataset thus, we can ensure best fit machine learning model for prediction. When we select alpha value
0.005 than ML algorithm work best. [Here, you can read alpha as lamda]. When we select our alpha value *0.004* the best score provided by *Random Forest Regressor* and that is *0.761995* and when we select alpha value *0.005* the best score provided by *Random Forest Regressor and* that is *0.807730*. Here, we can see a tendency of reduce score and that means when we minimize the independent features we find the best value of alpha. Thus, the selected the alpha value is *0.005*.

H. Model Tunning I

When we select alpha value .004 for, sklearn.linear_model.**Lasso**……(15)
Than our selected features were 'Year', 'Total Investment', 'Unemployed Rate' and those are capture for X (use for train test split).

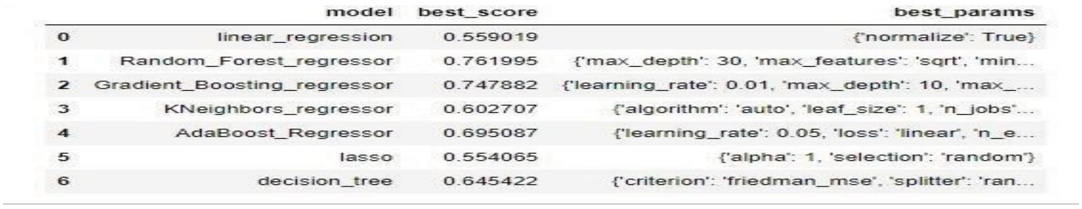
We perform Grid Search CV for find out best model. We do perform on 7 Regression algorithms

Table I: Model Parameters for tunning

Algorithms Name	**Parameters for algorithm**
Linear Regression	normalize: [True, False]

Random Forest Regressor	n-estimators: [1000] min-samples-split: [2] min-samples-lea f: [1, 2, 5] max-features: ['sqrt'] max-depth: [5, 10, 15, 20, 25, 30]
Gradient Boosting Regressor	n-estimators: [100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 1100, 1200] min-samples-split: [2,4,8] learning-rate: [0.01,0.1,1] max-features' ['auto', 'sqrt', 'log2'] max-depth': [5, 10, 15, 20, 25, 30]
K-Neighbors Regressor	n-neighbors:[4,5,6,7] leaf-size:[1,3,5] algorithm:['auto', 'kd-tree'] n-jobs: [-1]
AdaBoost Regressor	n-estimators: [50, 100] learning-rate[0.01,0.05,0.1,0.3,1] loss:['linear','square','exponential']
Lasso	alpha: [1,2] selection: ['random', 'cyclic']
Decision Tree Regressor	criterion:['mse','friedman_mse'] splitter: ['best','random']

Then the output of tuning



```
                        model  best_score                                        best_params
0           linear_regression    0.559019                                {'normalize': True}
1     Random_Forest_regressor    0.761995   {'max_depth': 30, 'max_features': 'sqrt', 'min...
2  Gradient_Boosting_regressor   0.747882  {'learning_rate': 0.01, 'max_depth': 10, 'max_...
3        KNeighbors_regressor    0.602707      {'algorithm': 'auto', 'leaf_size': 1, 'n_jobs'...
4          AdaBoost_Regressor    0.695087      {'learning_rate': 0.05, 'loss': 'linear', 'n_e...
5                       lasso    0.554065                  {'alpha': 1, 'selection': 'random'}
6               decision_tree    0.645422      {'criterion': 'friedman_mse', 'splitter': 'ran...
```

Figure7. Score of all applied algorithms

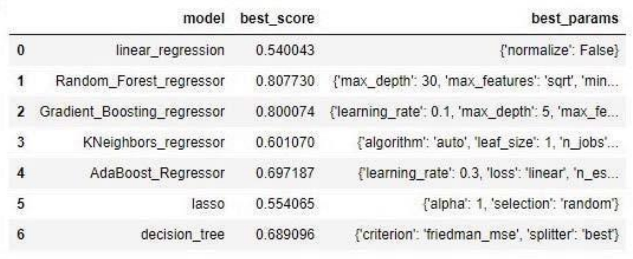
As the figure shows Random forest regressor works best as algorithm for our dataset and fit for serve our purpose.

A. Testing Dataset

Now, in this step we select alpha value as 0.005 for purpose of improving score for machine learning model. For this value our selected features were change and those are ‘Year’, ‘Total Investment’ which we capture as X (use for train test split).

J. Model Tunning II

When we select alpha value .005 for, sklearn.linear_model.**Lasso**……(15) than after tunning all model the output look like:



	model	best_score	best_params
0	linear_regression	0.540043	{'normalize': False}
1	Random_Forest_regressor	0.807730	{'max_depth': 30, 'max_features': 'sqrt', 'min...
2	Gradient_Boosting_regressor	0.800074	{'learning_rate': 0.1, 'max_depth': 5, 'max_fe...
3	KNeighbors_regressor	0.601070	{'algorithm': 'auto', 'leaf_size': 1, 'n_jobs'...
4	AdaBoost_Regressor	0.697187	{'learning_rate': 0.3, 'loss': 'linear', 'n_es...
5	lasso	0.554065	{'alpha': 1, 'selection': 'random'}
6	decision_tree	0.689096	{'criterion': 'friedman_mse', 'splitter': 'best'}

Figure 8. Score of all applied algorithms

IV. RESULT AND ANALYSIS

Now we evaluate which of the model is best fit for our dataset.

A. Training, Testing

We separated our data into training and testing. Since, our dataset was not large, we follow a technique called K- Fold Cross-Va lidation. We split our selected independent and dependent features into 5 folds and for each fold we allocated 20% of data testing.

B. Model Performance

We can see in our model outcome, there are two algorithm that score equal 80% and those two are:

Random Forest Regressor (81%)

…….(13)

Gradient Boosting Regressor (80%)

…….(14)

Table III: Comparison Between two Models

Model Name	Score on training data	Score on testing data	Accuracy
Random Forest Regressor	0.96068	0.74296	0.74296
Gradient Boosting Regressor	0.99966	0.65388	0.65388

Since, *Random Forest Regressor Algorithm* is more accurate than *Gradient Boosting Regressor* thus, we select *Random Forest Regressor* as best model for our project.

Mean-Square-Error = ^{1}σ ◌|◌11(−)^{ଗ} ….(7)

The Result of our model MSE=0.004635

σ 1ä ◌|◌͡ ä

Mean-Absolute-Error = …(7)

The Result of our model MAE: 0.062673

Root-Mean-Square-Error= $\frac{1}{\,}$ σ ଶ …(7)

C. Result Compression

Before beginning our thesis based project, we read through a lot of projects, but, alas there are only few paper which are related to our projection work. At last, we found "Forecasting Canadian GDP growth using XGBoost" where we found three options to compare our model and result. First one is MSE and our value is less than "Forecasting Canadian GDP growth using XGBoost''. The significant value of MSE is closer to zero and our value is less than "Forecasting Canadian GDP growth using XGBoost". Second one and third one is MAE, RMSE and the significant value of them nearer to 0.5.In our project our MAE and RMSE closer to 0.5 than "Forecasting Canadian GDP growth using XGBoost". And our accuracy of training set and test set is respectively 96% and 75%.

Table II: Evaluation of some previous works

Work	Model Name	MSE	MAE	RMSE
Forecasting Canadian GDP growth using XGBoost	XGBoost	0.019763	0.016845	0.019763
Gdp Growth Prediction of Bangladesh using machine learning algorithm	Random Forest Regressor	0.004635	0. 06267 3	0.068084

V. CONCLUSION AND FUTURE WORK

A country's GDP is very important. It gives more details about A countries GDP is very important. It gives more details about the size of the country's economy and the sustainability of that country's economy. The general health of the economy is reflected by the growth rate of real GDP. In developing the idea, we will recognize that a rise in real GDP is seen as an indication that the economy is doing well. And the decrease is a sign that the economy is not doing well. Predicting the increase and decrease of the GDP will be a good indicator to take action in certain sectors to improve and help increase the GDP. In a country like Bangladesh economic state has depended on many factors due to the political conflicts, natural disasters hampering agricultural growth, imbalance in the national budget and many more. As a developing country Bangladesh needs a stable economic growth to achieve its goal to become a developed country. To ensure the economic stability economist must help the government and authorities about the current state of economy and the growth that leads to the future. By this research we will find out in which para meters we should focus more to ensure the better GDP growth testing the limits of machine learning predictive ability. We will show how machine learning can help develop economic theory. In our future works , we will include more para meters of the other countries in which our economic growth depends , such as, EU (e xport), China (foreign investment), Middle East (remittance). We will make economists comfortable with machine learning. We will make economics more policy conscious by using machine learning to focus on causal pathways that have a meaningful impact on dependent variables.

REFERENCES

[1] A Richardson, T Mulder - Nowcasting New Zealand GDP using machine learning algorithms - papers.ssrn.com

[2] K Barhoumi, SM Choi, T Iyer, J Li, F Ouattara, A Tiffin… - A Machine-Learning Approach to Nowcast the GDP in Sub-Saharan Africa - aip.uneca.org

[3] James Thomas Bang,Atin Basuchoudhary,T inni Sen-New Tools for Predicting Economic Growth Using Machine Learning: A Guide for Theory and Policy-researchgate

[4] Shafiullah Qureshi,Ba M Chu,Fanny S. Demers-Forecasting Canadian GDP growth using XGBoost - repec.org

[5] Pirasant Premraj-Forecasting GDP Growth-nhh brage

[6] Tianqi Chen , Carlos Guestrin-XGBoost: A Scalable Tree Boosting System.

[7] Spyros Makridakis, Evangelos Spiliotis. Statistical and Machine Learning forecasting methods: Concerns and ways forward

[8] Shafiullah Qureshi, Ba M Chu and Fanny S. Demers-Forecasting Canadian GDP growth using XGBoost

[9] Rickard Nyman , Poul Ormerod-Predicting Economic Recessions Using Machine Learning Algorithms

[10] N K Ahmed, A F Atiya, N E Gayar, H El-Shishiny-An Empirical Comparison of Machine Learning Models for Time Series Forecasting

[11] W J Jansen, X Jin, J M De Winter-Forecasting and nowcasting real GDP: Comparing statistical models and subjective forecasts

[12] Leo Beriman. Random Forests

[13] Nils Adriansson & Ingrid Mattsson- Forecasting GDP Growth, or How Can Random Forests Improve Predictions in Economics?

[14] Jerome H. Friedman-Stochastic Gradient Boosting.

[15] Matthias Feurer and Aaron Klein and Katharina Eggensperger and Jost Tobias Springenberg and Manuel Blum and Frank Hutter-Auto- sklearn: Efficient and Robust Automated Machine Learning.