Predicting World GDP Based On Countries Social, Economic And Cultural Data

Project Report for CPSC-6300

Eshaa Deepak Sood, Vikas Garg

Abstract—**In recent times, decision making has been shifted to data driven outcomes. Analysing the past data, getting insights about how various parameters influence the decision are crucial steps. Data science is a domain where past available data is analysed using statistics, data visualizations, scientific methods and algorithms. We aim to predict the world gross domestic products (GDP) based on GDPs of various countries. The GDP of countries is impacted by various social, economical, cultural parameters. We are analysing those parameters from 1960 to 2017 and will predict future GDP of the world. We are using supervised learning methods to build our models. Our procedure includes data mining, data clustering, regressions, feature analysis and reduction. We are building our models using Python and its libraries.**

Index Terms—**GDP-prediction, Regression, Machine-Learning, Supervised Learning, PCA.**

I. Introduction

A large amount of data is created every single day or every single second. Exploring the data can give the industries insights about their product's position in the markets. The governments can also leverage the big data to identify and plan service delivery among various regions of the country. Manual analysis of such a large data may not be possible or even if possible may require a substantial amount of skilled human resources. However, the advent of technologies, computer, statistical analysis using software etc. has completely revolutionised how we shall deal with the data. Nowadays, everything is available at the snap of fingers. Like every e-commerce companies can analyse customers' purchasing behaviour and thus using computerised algorithms can recommend products the customer may like. This helps both companies to boost their sales and customers to get what they want. Similarly, the governments also leverage the statistical computing to prepare the public polices or predicting any emergency situations and preparing the road-map to avoid disasters.

Data Science is a field which helps in data mining, data cleaning, data clustering, and making prediction models using statistical algorithms available on internet. In our project we are working on the similar attributes of the data science. We are trying to predict the world gross domestic products (GDP) using past data available. GDP is a measure of the economic growth of any given country. The economic growth of the country depends on many factors such as Social, Economic and Cultural Environment. We are thus building a prediction model by including all such factors as independent variable and world GDP as dependent variable. We are including the data from 1960 to 2018 for world's 264 countries which is collected from World Bank [3], Kaggle [1]. We would be using data cleaning, data clustering techniques to get refined data and then build linear, polynomial regression model to predict the world GDP. We will be using test-train method to check the Minimum Squared Error (MSE) and if our model is over-fitting, we will be using regularisation such as Lasso, Ridge etc. We will also try to reduce the features using Principal Component Analysis (PCA). By using PCA we will try to include the components which result in maximum variance of the data. The organisation of the report is as follows: Section II discuss the previous related work available, Section III describes the numerical analysis we have performed, Section IV describes the results obtained from the numerical analysis, Section V includes the conclusion we have made from this project. The last part of the report includes the references we have used in preparing the report.

II. Previous Related Work

GDP is an important parameter to know the health and condition of a country compared to the other countries. Therefore, knowing beforehand about GDP helps in knowing whether a country is progressing or it's economic health is declining [6]. There has been considerable work available related to world GDP. The paper titled "Tracking world trade and GDP in real time" [7] has proposed a mathematical model to forecast world GDP on real time basis. This paper has prepared the World Bridge Model (WBM) to forecast GDP based on real time 7000 time series data sets. However this work is more focused on trade and economic parameters of the countries. The project titled "Predicting GDP: world countries" [5] has included some of the economic and geographical parameters to build a prediction model using Python Libraries. However, [5] has included approximately 20 features leaving out some of the features such as Number of ATM machines, Gender distribution in employments etc. These parameters may also impact the world GDP. Therefore, we are trying to look into the correlation of all such parameters in our model. Another paper titled "Prediction of GDP growth rate based on carbon dioxide (CO2) emissions" [8] predicts the GDP growth rate considering environmental impacts such as carbon dioxide (CO2) emissions. The paper [8] uses the Extreme Learning Machine (ELM) to predict the GDP growth based on the input data related to % of emissions from types of fuel used. The paper compares the statistical results such

Root-mean-square error (RMSE), Coefficient of determination (R^2) obtained using ELM with those obtained using genetic programming (GP) and artificial neural network (ANN). This paper also didn't account for the other parameters such as trade, economic, geographical to predict the GDP growth. The GDP can also be predicted using time series models as done in [7]. Similar study was done for regional GDP prediction in paper titled "Modeling and forecasting regional GDP in Sweden using autoregressive models" [4]. Here the author used Autoregressive Integrated Moving Average (ARIMA) model, the Vector Autoregression (VAR) model and the First-order Autoregression (AR(1)) model [4]. This article also considered limited factors affecting the GDP for the prediction models. Another work in predicting GDP using regression model is done in a thesis titled "Analysis of GDP using Linear Regression" [9]. In [9] the author has used the S&P 500 and various sectors are correlated with GDP. The author is trying to predict what parameters actually impact the GDP. The author has limited the model to only S&P 500 and certain industries so as to aid in decision making related to finance and economics. However, this model will not be a holistic overview of a country's GDP prediction model.

In the literature discussed above, none of the work considered the holistic view of GDP dependence on Social, Economical, Geographical, Environmental impacts to predict world GDP. Most the work focused on building mathematical or time series models except works like [5]. Therefore, in our proposed model we are trying to include all such comprehensive parameters. We will look what parameters impact the GDP, what parameters are correlated, how we can reduce the parameters using Principal of Component Analysis. We will then use linear or polynomial regression models along with some regularisation techniques if our model is over-fitting the data. We will test our models using test data and estimate the performance of our model using significant coefficients.

III. Numerical Analysis

We have collected the 58 years' (1960-2017) data of 264 countries. We choose the data related to social, economic, and cultural parameters. Our data included the Year, Women_Informed_Choices, RuralPopulation_PerCent, LegalRights_Strength, CreditTo_PrivateSector, BirthsAttendedby_SkilledStaff, ATMMachines_Ratio, Agricultural_Machines. Our aim here is to predict the "value" parameter which corresponds to the GDP value of each country. Following are steps that we will be using in our numerical analysis. Each of the step is explained in the relevant section.

The figure 2 shows the statistics summary all of individual features. As we can see most of the data has large variation for example the minimum value of GDP is approximately 34 and the maximum value is approximately 200,000. So we tried to standardise the data using sickit-learn's StandardScaler preprocessing [2].

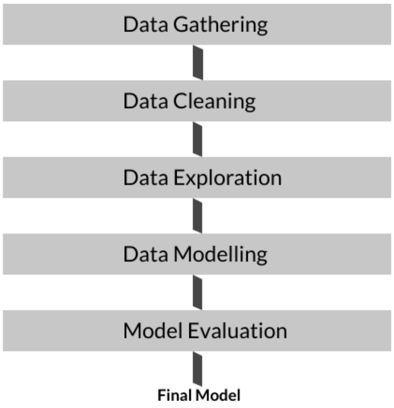
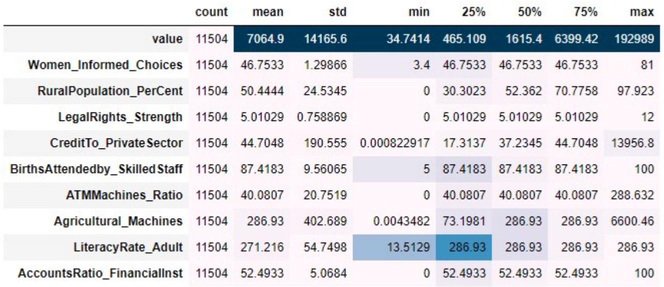


Fig. 1: Tasks Workflow



	count	mean	std	min	25%	50%	75%	max
value	11504	7064.9	14165.6	34.7414	465.109	1615.4	6399.42	192989
Women_Informed_Choices	11504	46.7533	1.29866	3.4	46.7533	46.7533	46.7533	81
RuralPopulation_PerCent	11504	50.4444	24.5345	0	30.3023	52.362	70.7758	97.923
LegalRights_Strength	11504	5.01029	0.758869	0	5.01029	5.01029	5.01029	12
CreditTo_PrivateSector	11504	44.7048	190.555	0.000822917	17.3137	37.2345	44.7048	13956.8
BirthsAttendedby_SkilledStaff	11504	87.4183	9.56065	5	87.4183	87.4183	87.4183	100
ATMMachines_Ratio	11504	40.0807	20.7519	0	40.0807	40.0807	40.0807	288.632
Agricultural_Machines	11504	286.93	402.689	0.0043482	73.1981	286.93	286.93	6600.46
LiteracyRate_Adult	11504	271.216	54.7498	13.5129	286.93	286.93	286.93	286.93
AccountsRatio_FinancialInst	11504	52.4933	5.0684	0	52.4933	52.4933	52.4933	100

Fig. 2: Statistical Summary of the input data

A. Data Processing and Cleaning:

We have collected common data based on country and country code. There were few datasets which could be merged together using 'inner join' for example the PopulationPerCountry and GDP by country could be merged together. Similarly we combined all of the features into common DataFrame. Then we analysed the data for any missing, NaN values. We had the option either to remove missing/null entries or impute those values. Removing the data may seem an easy option however it would lead to loss of information. Performance of any prediction model improves as more and more data is used to train-test the data. Therefore we imputed the missing value for each of the feature with the respective mean values.

In figure 3 we plotted the pairwise relation of all the features. As we can see none of the feature is strongly correlated to the "value" parameter. Also, most of the features follow linear trend with the "value" parameter, so we except that linear regression will give better results as compare to the polynomial however we will compare the performance of multiple linear regression and polynomial regression in our next subsections. We also tried feature combining to see if we get any strong correlation with the "value" parameters.

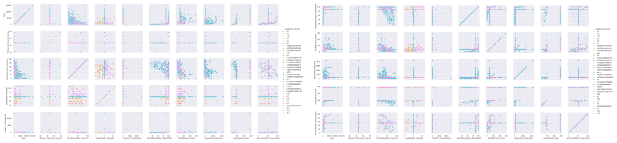


Fig. 3: Pairwise relation of the features

B. Feature Combining and Correlation:

We have tried to combine the social-cultural, cultural-economic, social-economic features. The features we combined are Literacy_creditToPriva, Literacy_RuralPop, Literacy_AgriMach, Literacy_AccountRa, Literacy_ATM, Literacy_BirthAT, Literacy_Legal, Literacy_Woman, Woman_Rural, Woman_CreditToPriv, Woman_AgriM, Woman_ATM, Woman_BirthAT. Then we compared the correlation of existing parameters and newly combined features with that of dependent feature. Fig 4 shows how the correlation heat map of original features and the combined features.As we can see that the combined features don't have a strong or changed correlation with the dependent value, so we expect that our model's performance will remain same for the original variables. We have tested the modified features with our best fit model.

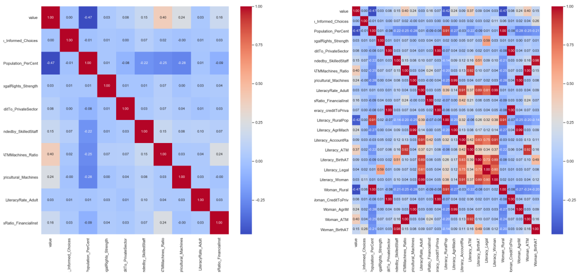


Fig. 4: Correlation Heatmap of features

C. Model Validation Methods:

Evaluating a model's performance is important to understand it's accuracy and reliability. Also model evaluation techniques can be used to compare the performance of different models and decide the best fit for the data. The various model evaluation techniques that we used are described below.

1) Mean Square Errors: In Data Science, the mean squared error (MSE) or mean squared deviation (MSD) of an estimator is used to measure the average of the squares of the errors—that is, the average squared difference between the estimated values and the actual value. The MSE measures the quality of an estimator—it is always non-negative, and values closer to zero are better. Therefore lower the MSE better the model is for the data.

2) R-squared: R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression. It gives the percentage of the response variable variation that is explained by a linear model denoted by the formula R-squared = Explained variation / Total variation R-squared always ranges between 0 and 100 percent:

- 0 percent indicates that the fitted model explains none of the variability of the response data around its mean.
- 100 percent indicates that the fitted model explains all the variability of the response data around its mean.

Therefore the higher the R-squared, the better the model fits your data.

3) Test-Train Split:: For any prediction model, it is necessary to test the models on various datasets. This is done to check whether the trained model is overfitting or underfitting the train dataset or it may be the case that the model is reducing the errors for those specific train datasets. If we pass multiple data sets and the model is giving similar results, then we can say the model is not biased towards dataset.
Therefore, we split our dataset into train & test. We used the ”sklearn.model_selection.train_test_split” method to divide the GDP dataset. We used test_size=0.20 & random_state=40 as the input parameters for the test_train split i.e. we have divided the data 20% into the test and rest 80% into the train dataset. Therefore our test set contained 2300 observations and train set contained 9204 observations. As can be seen from fig 5 the actual GDP values of 264 contries is identically distributed with that of the test dataset. So this will ensure that test dataset is actual representation of the actual dataset. This is very important part of model fitting and validations.Besides test-train split validation of the models we also performed the k-fold validations which will be discussed in relevant subsection below.

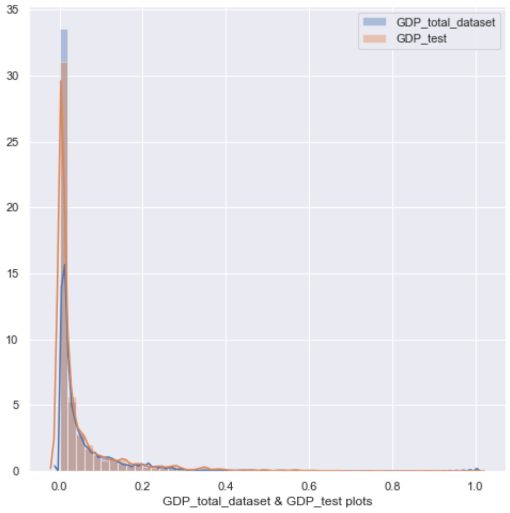


Fig. 5: Normal distribution of Actual & Test split of GDP

4) k-Fold Cross Validation:: Cross-validation is a statistical method which is used to estimate the skill of machine learning models. It is commonly used in applied machine learning to compare and select a model for a given predictive modeling problem. The algorithm of k - fold cross validation is as follows:

- The original training data set is partitioned into k equal subsets. Each subset is called a fold. Let the folds be named as f1, f2, . . . , fk .
- For i = 1 to i = k The fold fi is kept as as Validation set and all the remaining k-1 folds are in the Cross validation training set. The machine learning model is trained using the cross validation training set and accuracy of the model is calculated by validating the predicted results against the validation set. The accuracy of the machine learning model is calculated by averaging the accuracies derived in all the k cases of cross validation.

D. Oridnary Least Square

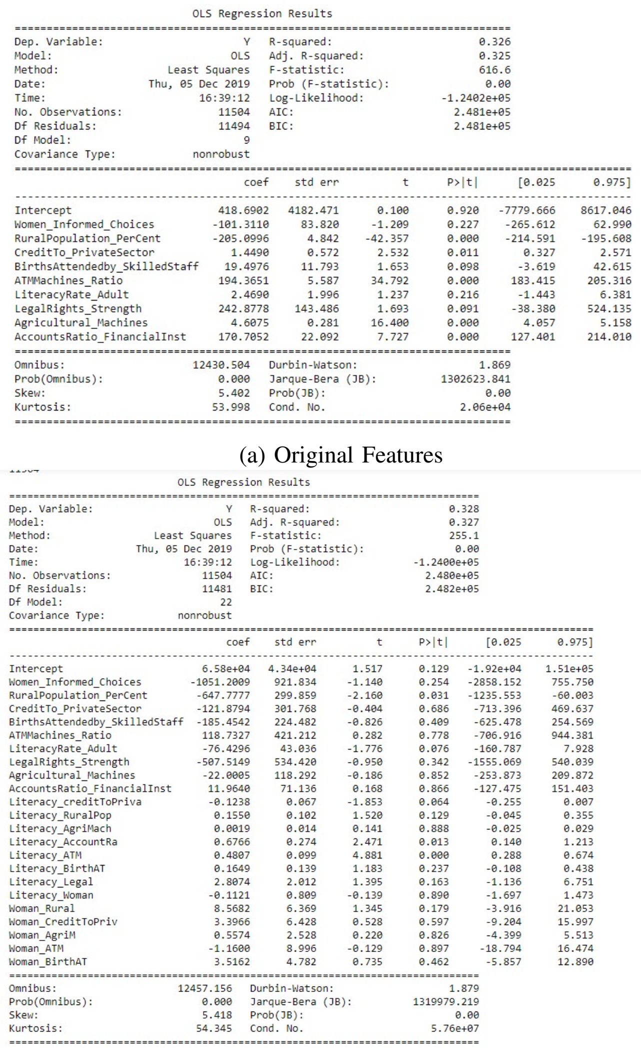
Ordinary least squares (OLS) regression is a statistical analysis of predicting the relationship of dependent and independent variables. The relationship is calculated by minimizing the sum of squares of difference between predicted and actual value. We have performed the OLS on both original and combined features. Fig 6 summarizes the OLS results. From the figure, it is evident that R-square, Adjusted R-square are same for both the models. The difference of residuals is also same.

E. Model Selections

In data science there are basically two type of statistical learning methods supervised learning and unsupervised learning. In the supervised learning we have the ground truth i.e. we have prior knowledge of the dependent variable and we want to fit the model for close prediction to the known ground truth. In the unsupervised learning we are not aware of the output's prior values. Unsupervised learning helps identifying the patterns among the independent variables. For our dataset we already know the GDP values which is the traget value. Therefore, we will be using supervised learning methods. There are numerous types of models available which can be used for supervised learning. For our task we have used following 7 types of prediction models to find out the best suited model for the dataset.

1) Multiple Linear Regression
2) Polynomial Regression
3) Decision Tree Regression
4) Random Forest Regression
5) Ridge Regression
6) Lasso Regression
7) Elastic Net Regression

Those models are quite often used in applied data science. In the following section we are describing the relevant parameters we used and the respective results for each of the model. Then in the resultIV section discussed the performance of all model altogether.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.326
Model:                            OLS   Adj. R-squared:                  0.325
Method:                 Least Squares   F-statistic:                     616.6
Date:                Thu, 05 Dec 2019   Prob (F-statistic):               0.00
Time:                        16:39:12   Log-Likelihood:            -1.2402e+05
No. Observations:               11504   AIC:                         2.481e+05
Df Residuals:                   11494   BIC:                         2.481e+05
Df Model:                           9
Covariance Type:            nonrobust
=============================================================================================
                                coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------------
Intercept                   418.6902   4182.471      0.100      0.920   -7779.666    8617.046
Women_Informed_Choices     -101.3110     83.820     -1.209      0.227    -265.612      62.990
RuralPopulation_PerCent    -205.0996      4.842    -42.357      0.000    -214.591    -195.608
CreditTo_PrivateSector        1.4490      0.572      2.532      0.011       0.327       2.571
BirthsAttendedby_SkilledStaff 19.4976    11.793      1.653      0.098      -3.619      42.615
ATMMachines_Ratio           194.3651      5.587     34.792      0.000     183.415     205.316
LiteracyRate_Adult            2.4690      1.996      1.237      0.216      -1.443       6.381
LegalRights_Strength        242.8778    143.486      1.693      0.091     -38.380     524.135
Agricultural_Machines         4.6075      0.281     16.400      0.000       4.057       5.158
AccountsRatio_FinancialInst 170.7052     22.092      7.727      0.000     127.401     214.010
==============================================================================
Omnibus:                    12430.504   Durbin-Watson:                   1.869
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1302623.841
Skew:                           5.402   Prob(JB):                         0.00
Kurtosis:                      53.998   Cond. No.                     2.06e+04
==============================================================================
```

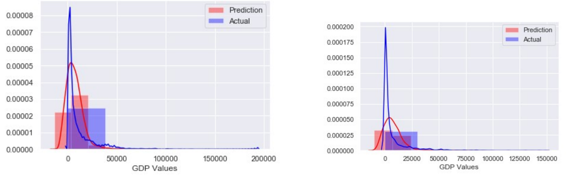
(a) Original Features

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.328
Model:                            OLS   Adj. R-squared:                  0.327
Method:                 Least Squares   F-statistic:                     255.1
Date:                Thu, 05 Dec 2019   Prob (F-statistic):               0.00
Time:                        16:39:12   Log-Likelihood:            -1.2400e+05
No. Observations:               11504   AIC:                         2.480e+05
Df Residuals:                   11481   BIC:                         2.482e+05
Df Model:                          22
Covariance Type:            nonrobust
=============================================================================================
                                coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------------
Intercept                   6.58e+04   4.34e+04      1.517      0.129   -1.92e+04    1.51e+05
Women_Informed_Choices    -1051.2009    921.834     -1.140      0.254   -2858.152     755.750
RuralPopulation_PerCent    -647.7777    299.859     -2.160      0.031   -1235.553     -60.003
CreditTo_PrivateSector     -121.8794    301.768     -0.404      0.686    -713.396     469.637
BirthsAttendedby_SkilledStaff -185.4542 224.482     -0.826      0.409    -625.478     254.569
ATMMachines_Ratio           118.7327    421.212      0.282      0.778    -706.916     944.381
LiteracyRate_Adult          -76.4296     43.036     -1.776      0.076    -160.787       7.928
LegalRights_Strength       -507.5149    534.420     -0.950      0.342   -1555.069     540.039
Agricultural_Machines       -22.0005    118.292     -0.186      0.852    -253.873     209.872
AccountsRatio_FinancialInst  11.9640     71.136      0.168      0.866    -127.475     151.403
Literacy_creditToPriva       -0.1238      0.067     -1.853      0.064      -0.255       0.007
Literacy_RuralPop             0.1550      0.102      1.520      0.129      -0.045       0.355
Literacy_AgriMach             0.0019      0.014      0.141      0.888      -0.025       0.029
Literacy_AccountRa            0.6766      0.274      2.471      0.013       0.140       1.213
Literacy_ATM                  0.4807      0.099      4.881      0.000       0.288       0.674
Literacy_BirthAT              0.1649      0.139      1.183      0.237      -0.108       0.438
Literacy_Legal                2.8074      2.012      1.395      0.163      -1.136       6.751
Literacy_Woman               -0.1121      0.809     -0.139      0.890      -1.697       1.473
Woman_Rural                   8.5682      6.369      1.345      0.179      -3.916      21.053
Woman_CreditToPriv            3.3966      6.428      0.528      0.597      -9.204      15.997
Woman_AgriM                   0.5574      2.528      0.220      0.826      -4.399       5.513
Woman_ATM                    -1.1600      8.996     -0.129      0.897     -18.794      16.474
Woman_BirthAT                 3.5162      4.782      0.735      0.462      -5.857      12.890
==============================================================================
Omnibus:                    12457.156   Durbin-Watson:                   1.879
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1319979.219
Skew:                           5.418   Prob(JB):                         0.00
Kurtosis:                      54.345   Cond. No.                     5.76e+07
==============================================================================
```

(b) Combined Features

Fig. 6: Ordinary Least Squares Analysis

1) Multiple Linear Regression:: Fig 7 represents the GDP value comparison between actual and predicted values. As we can see from the fig, the predicted values are not close to the actual values. This is evident from the RMSE values as well. The RMSE values are, train=11098.55, test=11098.55, & k-Fold RMSE=13701.24. Those RMSE are approximately equal to one standard deviation of the GDP value distribution.

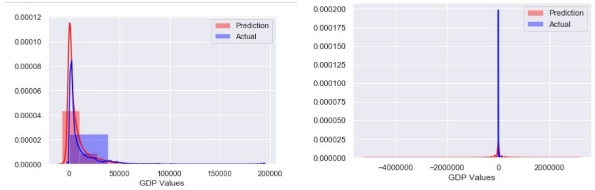


(a) Train

(b) Test

Fig. 7: Linear Regression-Comparison,Actual vs Predicted

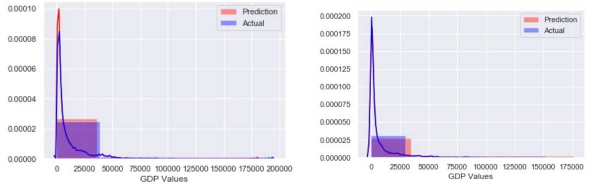
2) Polynomial Regression:: We used the polynomial feature of ”sklearn.preprocessing” to make the features into polynomial of degree=3. Then we fitted the linear model for those polynomial features. Fig 8 represents the GDP value comparison between actual and predicted values. The predicted values are more closer as compare to the multiple linear regression. However, the RMSE doesn’t improve much. The RMSE values are, train=9165.46, test=161698, & k-Fold RMSE=13701.24. As we can there is a huge difference between test and train RMSE. So based on this we can say multiple linear regression is better compared to polynomial regression.



(a) Train

(b) Test

Fig. 8: Polynomial Regression-Comparison,Actual vs Predicted



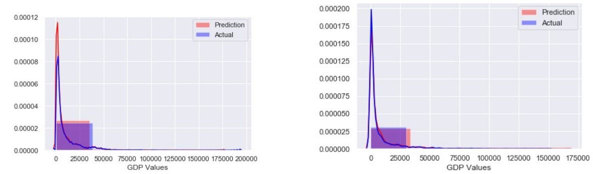
(a) Train

(b) Test

Fig. 9: Decision Tree-Comparison,Actual vs Predicted

3) Decision Tree Regression:: Decision tree regression builds a tree structure based on subset of features. Decision tree is both used for classification and regression. We have defined the maximum depth of the tree as ”30”. Fig 9 compares the GDP values for both test and train. As it can be seen that for both test and train the predicted values are close to actual values and show less variation among test and train. This can seen from RMSE values as well. The RMSE values are, train=2255.18, test=6584.06, & k-Fold RMSE=15318.41. The RMSE for test and train as comparitavly low as compared to the linear and polynomial regression. The cross validation RMSE doesn’t improve much. However the R-squared value for decision tree is 0.97 for train and those values for linear and polynomial regression are 0.37 & 0.57 respectively.

4) Random Forest Regression:: Similar to Decision Tree regression, Random Forest regression is also an ensemble learning method. In this method, trees are build in parallel and Random Forest calculates the mean for each of the node built based on the selected features. Random Forest randomly selects the features and prevents overfitting. From fig 10 it can be seen that similar to Decision Tree, predicted values and actual values are close for Random Forest as well. The RMSE values are, train=2873.09, test=5582.56, & k-Fold RMSE=11649.48. Apart from improved RMSE values, the R-squared value for Random Forest is 0.95 which 0.97 for Decision Tree. As R-squared value increases the model is considered to be a better fit.



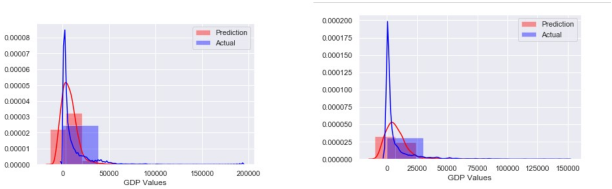
(a) Train

(b) Test

Fig. 10: Random Forest-Comparison,Actual vs Predicted

We also tried regularization methods to see if the model’s performance increases. However, looking at the RMSE and R-squared values of linear and polynomial regression we anticipate that those models are not overfitting. We tried Ridge, Lasso, & Elastic Net regularization regression methods.

5) Regularizations:: In the regularization we reduced the overfitting at the expense of adding bias to the model. In python, ”sklearn.linear_model.Ridge” library is used to implement ridge regression. Here we need to define the regularisation parameter (alpha) which puts constraint on the model to reduces the overfitting of the model. In our model we used alpha=0.0000001. Fig 11 clearly indicates that as we expected regularisation doesn’t improve the model performance. The RMSE values are train=11098.55,test=11357.50, & k-Fold RMSE=12424.70. The R-squared value is 0.37. Those values are similar to the linear regression. Therefore, we can say our models don’t overfit.



(a) Train

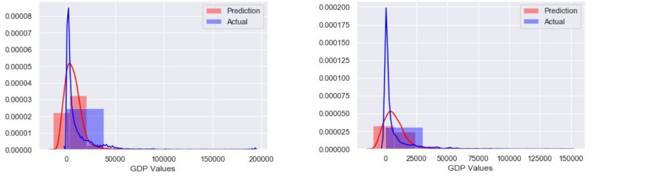
(b) Test

Fig. 11: Ridge Regression-Comparison,Actual vs Predicted

Similar results were obtained for other two types of regularization methods. Fig 12 is the representation of the Lasso Regression method. We used the same regularization parameter (alpha=0.0000001). Same observations were made for Elastic Net Regression as evident from Fig 13. For both type of regularisation the R-sqaured value is 0.37 which is equal to Ridge regression. Also the RMSE values are approximately.

From these regularization we observe that our models don’t overfit. Therefore, we don’t need to use the regularizations for our work.

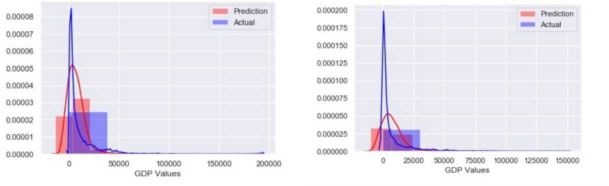
6) Principal Component Analysis:: Principal component analysis is used to reduce the number of features which



(a) Train

(b) Test

Fig. 12: Lasso Regression-Comparison,Actual vs Predicted

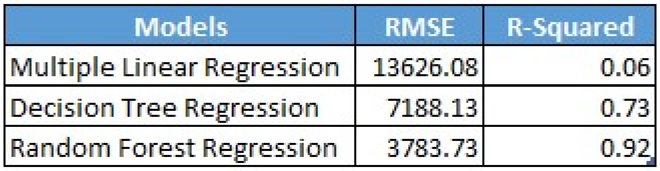


(a) Train

(b) Test

Fig. 13: Elastic Net-Comparison,Actual vs Predicted

explains the maximum variations so that model's performance improves. In our analysis the number of features used are 10 and feature combinations goes up to 23 features. In our analysis, we chose the features which will explain the 95% variances. We transformed the data and then used the Multiple Linear, Decision Tree, & Random Forest regression to check how the RMSE and R-sqaured values are affected. As evident from fig 14, those values don't improve. Hence, PCA is not necessary for our models.

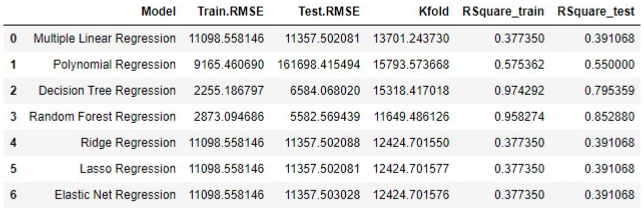


Models	RMSE	R-Squared
Multiple Linear Regression	13626.08	0.06
Decision Tree Regression	7188.13	0.73
Random Forest Regression	3783.73	0.92

Fig. 14: PCA-performance comparison

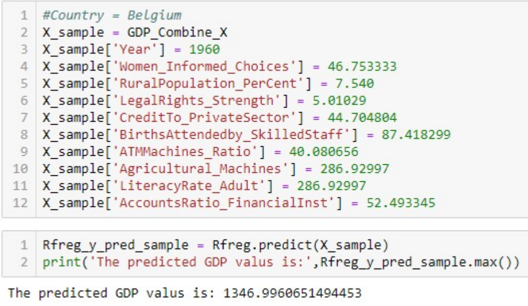
IV. Results

In fig 15 we have summarized the results of all the 7 models. The RMSE values for Multiple Linear, Polynomial, Ridge, Lasso, and Elastic Net regression are approx. same and quite high as compared to the Decision Tree and Random Forest regression models. The R-squared values for Decision Tree and Random forest regression are more than 95%. As we know higher the R-squared value and lower the RMSE, better the model is. Therefore, we can say the Decision Tree and Random Forest regression models are the best fit for our data. We also tested the accuracy of those best fit models by randomly selecting the input for any country for the year 1960 and 2016. For the year 1960, we checked accuracy of Belgium country whose actual value is 1273.69 and the predicted value is 1346.99 (16) and for year 2016, we tested the accuracy of the USA whose actual value is 57638.15 and the predicted value is 56823.51 (17). So the accuracy of our models is close to approx. 95%.



	Model	Train.RMSE	Test.RMSE	Kfold	RSquare_train	RSquare_test
0	Multiple Linear Regression	11098.558146	11357.502081	13701.243730	0.377350	0.391068
1	Polynomial Regression	9165.460690	161698.415494	15793.573668	0.575362	0.550000
2	Decision Tree Regression	2255.186797	6584.068020	15318.417018	0.974292	0.795359
3	Random Forest Regression	2873.094686	5582.569439	11649.486126	0.958274	0.852880
4	Ridge Regression	11098.558146	11357.502088	12424.701550	0.377350	0.391068
5	Lasso Regression	11098.558146	11357.502081	12424.701577	0.377350	0.391068
6	Elastic Net Regression	11098.558146	11357.503028	12424.701576	0.377350	0.391068

Fig. 15: Model performance comparison

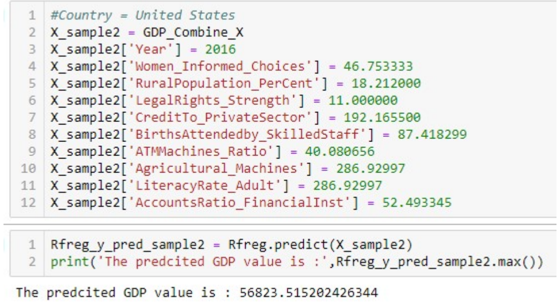


```
#Country = Belgium
X_sample = GDP_Combine_X
X_sample['Year'] = 1960
X_sample['Women_Informed_Choices'] = 46.753333
X_sample['RuralPopulation_PerCent'] = 7.540
X_sample['LegalRights_Strength'] = 5.01029
X_sample['CreditTo_PrivateSector'] = 44.704804
X_sample['BirthsAttendedby_SkilledStaff'] = 87.418299
X_sample['ATMMachines_Ratio'] = 40.080656
X_sample['Agricultural_Machines'] = 286.92997
X_sample['LiteracyRate_Adult'] = 286.92997
X_sample['AccountsRatio_FinancialInst'] = 52.493345
```

```
Rfreg_y_pred_sample = Rfreg.predict(X_sample)
print('The predicted GDP valus is:',Rfreg_y_pred_sample.max())
```

```
The predicted GDP valus is: 1346.9960651494453
```

Fig. 16: Belgium



```
#Country = United States
X_sample2 = GDP_Combine_X
X_sample2['Year'] = 2016
X_sample2['Women_Informed_Choices'] = 46.753333
X_sample2['RuralPopulation_PerCent'] = 18.212000
X_sample2['LegalRights_Strength'] = 11.000000
X_sample2['CreditTo_PrivateSector'] = 192.165500
X_sample2['BirthsAttendedby_SkilledStaff'] = 87.418299
X_sample2['ATMMachines_Ratio'] = 40.080656
X_sample2['Agricultural_Machines'] = 286.92997
X_sample2['LiteracyRate_Adult'] = 286.92997
X_sample2['AccountsRatio_FinancialInst'] = 52.493345
```

```
Rfreg_y_pred_sample2 = Rfreg.predict(X_sample2)
print('The predcited GDP value is :',Rfreg_y_pred_sample2.max())
```

```
The predcited GDP value is : 56823.515202426344
```

Fig. 17: USA

V. Conclusion

Through this project we realized the importance of each of the data science steps explained in our task workflow and realised that an accurate model cannot be devised unless the data is prepared and analyses appropriately. We explored all the supervised regression models in order to get the best fitting models. We also evaluated our model using various validation methods to get a clear overview of the performance of the models on the basis of various aspects. We also explored some efficient techniques like Ordinary Least Squares for efficient selection of features to be used in a model. We were able to create a model with a very high accuracy of about 95 percent which can be easily seen in our results table, graph visualization between predicted and actual data and evaluation on ground truth data.

References

[1] kaggle datasets. https://www.kaggle.com/datasets. last accessed-October 31st, 2019.

[2] Sickit learn. https://scikit-learn.org/stable/modules/preprocessing.html. last accessed-November 30,2019.

[3] The world bank data. https://data.worldbank.org/. last accessed-October 31st, 2019.

[4] A. George Assaf, Gang Li, Haiyan Song, and Mike G. Tsionas. Modeling and forecasting regional tourism demand using the bayesian global vector autoregressive (bgvar) model. *Journal of Travel Research*, 58(3):383–397, 2019.

[5] Marek (Tax Automation at Dell). kaggle datasets. https://www.kaggle.com/stieranka/predicting-gdp-world-countries. last accessed-November 6,2019.

[6] Tim Callen. International monetary fund. https://www.imf.org/external/pubs/ft/fandd/basics/gdp.htm#:~:text=GDP%20is%20important%20because%20it,the%20economy%20is%20doing%20well. last accessed-November 6,2019.

[7] Roberto Golinelli and Giuseppe Parigi. Tracking world trade and GDP in real time. Temi di discussione (Economic working papers) 920, Bank of Italy, Economic Research and International Relations Area, July 2013.

[8] Vladislav Marjanović, Miloš Milovančević, and Igor Mladenovic. Prediction of gdp growth rate based on carbon dioxide (co2) emissions. *Journal of CO2 Utilization*, 16:212–217, 12 2016.

[9] Rayan Mayolo. Analysis of gdp using linear regression. https://digitalshowcase.lynchburg.edu/cgi/viewcontent.cgi?article=1062&context=utcp.