# **Predicting Country FINC using machine learning (Regression)**

Contents

[**Predicting Country FINC using machine learning (Regression)** 1](#_Toc94029063)

[**Introduction:** 1](#_Toc94029064)

[**Handling Missing Values:** 2](#_Toc94029065)

[**Data Separation:** 2](#_Toc94029066)

[Linear Regression: 3](#_Toc94029067)

[Pros and Cons: 4](#_Toc94029068)

[**Ridge Regression:** 4](#_Toc94029069)

[Pros and Cons: 4](#_Toc94029070)

[**Lasso Regression:** 5](#_Toc94029071)

[Pros and Cons: 5](#_Toc94029072)

[**Polynomial Regression:** 5](#_Toc94029073)

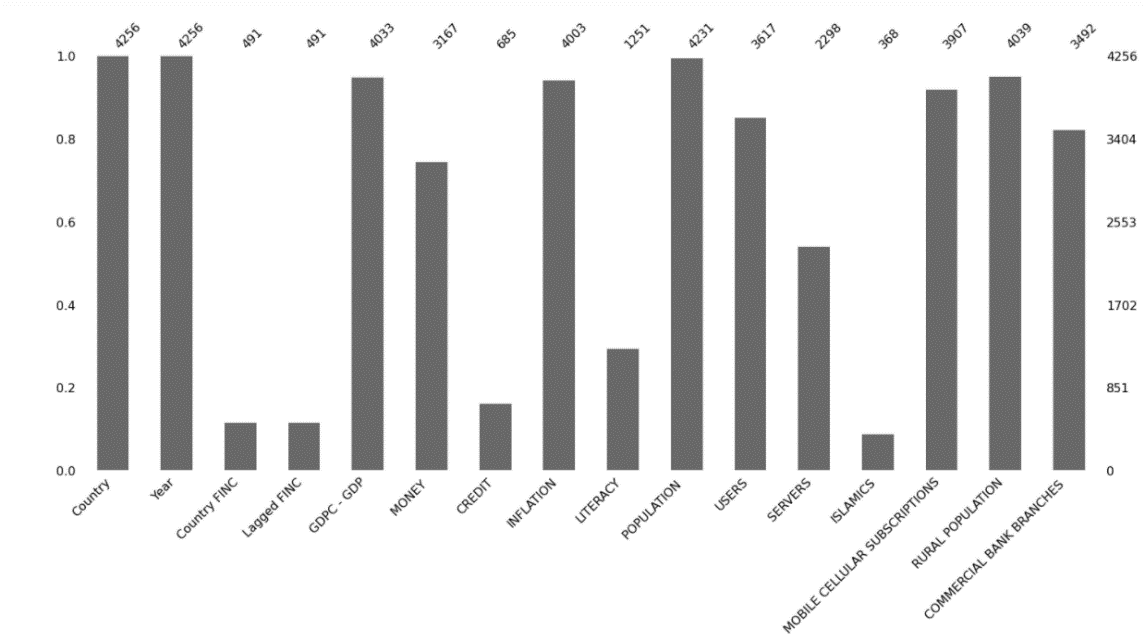
[Future Recommendation: 6](#_Toc94029074)

## **Introduction:**

“Prediction” refers to the output of an algorithm after it has been trained on a historical dataset and applied to new data when forecasting the likelihood of a particular outcome, such as whether or not a customer will churn in 30 days.

In our case, we have a dataset that contains many features like GDP, Money, Credit, Servers, etc. Based on these features, we have to predict the Country FINC. As mentioned above, predicting anything depends on the available historical data. After analyzing our dataset, we found that there is a huge missing in our dataset. For example, if we look at the country FINC column, out of 4275 rows, only 491 rows are given, the rest are missing. Now, this is the biggest problem in any machine learning prediction. Because we do predictions on the given data and there is no big historical data present of Country FINC.

Similarly, all other features’ data is also missing.



By Looking at the above graph, we can see the missing data from our dataset. For instance, let's look at the Country FINC and Lagged FINC features, there are only 491 rows where this feature’s data is given, the rest is missing. Similarly, the Money feature has 3167 entities and the rest are missing. The problem is the missing values but no worries, using machine learning we can deal with missing values. How?

## **Handling Missing Values:**

In machine learning, missing data is a headache. Machine learning has been the cornerstone in analyzing and extracting information from data and often a problem of missing values is encountered. Missing values occur because of various factors like missing completely at random, missing at random, or missing not at random.

In real-world data, missing values are a normal thing we face on regular basis, the reason is that we don’t care about our data or we enter data in hurry, in machine learning, missing values is also called data corruption, and there are many reasons due to which we miss the data at some point.

Let’s get straight to our topic, we have missing values, now how to handle them? There are many ways to handle missing values, some of them are given below:

* Deleting Rows with missing values
* Statistical analysis
* Impute missing values for continuous variable
* Impute missing values for categorical variable
* Other Imputation Methods
* Using Algorithms that support missing values
* Prediction of missing values
* Imputation using Deep Learning Library — Datawig

In our case, we cannot remove rows with missing values, the reason is that our dataset has many missing values, removing them will make our dataset so small which is not a solution. We will instead use the statistical analysis to tackle missing data, statistical analysis is the process where we take the Mean, Median, and Mode of the data given, based on the result we fill the missing values. We have taken the Median of the data and then filled the missing values with the result of the Median.

NOTE: we took the median of all the data and filled missing values except the Country FINC column because we are going to predict it.

The following code will take the median and fill the missing places with it.



We have removed the Country FINC and Lagged FINC from the dataset then we wrote the above Python code to fill all the missing values present in our dataset with the Median of the given values. We will attach the Country FINC again to the dataset having no missing values.

## **Data Separation:**

Data separation is often called splitting data into test and train data. Data splitting is commonly used in machine learning to split data into a train, test, or validation set. Each algorithm divided the data into two subsets, training/validation. The training set was used to fit the model and validation for the evaluation.

Our problem is slightly different, how? As we are going to predict Country FINC, but the Country FINC have lots of missing values, now these values will result in false predictions, some machine learning algorithms don’t accept missing values data, we are using regression models to predict Country FINC and regression models don’t accept missing values features as an input.

The solution to this problem is that we will take all those data where the Country FINC is given, on this data, we will train our regression models to predict the rest of the missing values in Country FINC. In other words, we are predicting the missing values using machine learning regression techniques.

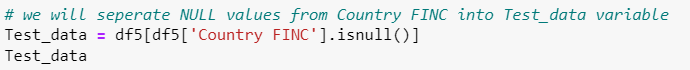
The following Python code will take all those rows in our dataset where the Country FINC is given.



We have separated all the data where the Country FINC is given, this is the training data, the amount of the data is 491 rows because we only 491 rows where Country FINC is given. Based on this data, we will train our regression model, and then we will do the prediction.

After that we will make another dataset called Test\_data, this data have those missing values of Country FINC, after training the model, we will then feed this Test\_data into the model to predict the missing Country FINC values.

Following is the Python code which will make another dataset called Test\_data:



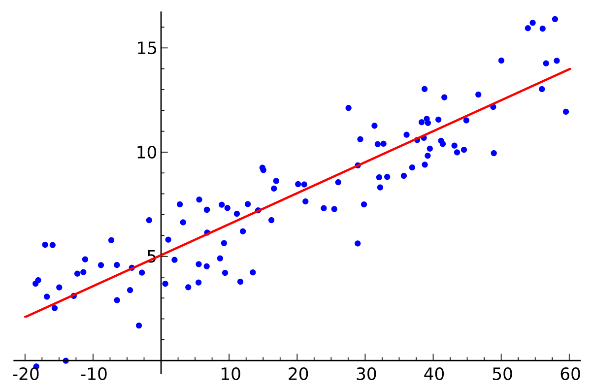
There are many regression models used in the market, but the most used regression models are:

* Linear Regression
* Ridge Regression
* Lasso Regression
* Polynomial Regression

We are going to train and test each of the above regression models and see what is the result?

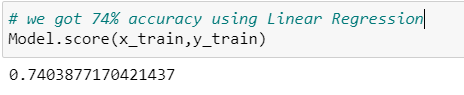
## Linear Regression:

In the simplest words, Linear Regression is the supervised Machine Learning model in which the model finds the best fit linear line between the independent and dependent variable i.e., it finds the linear relationship between the dependent and independent variable.



The first and the most famous regression model is Linear Regression is trained on the training data, we use a function called fit() to feed the data into the model, then the model will train itself after the training is completed, model is now ready for prediction. We are now checking the accuracy of each model. After getting the model with the highest accuracy score, we will use that regression model to do the predictions.

Linear Regression gave 74% accuracy on training data which is a good accuracy in machine learning.



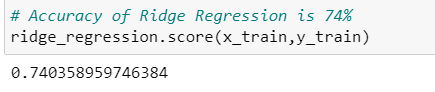
### Pros and Cons:

Linear Regression is easier to implement, interpret, and very efficient to train. 3. Linear Regression is prone to over-fitting but it can be easily avoided using some dimensionality reduction techniques, regularization (L1 and L2) techniques, and cross-validation.

## **Ridge Regression:**

After training and testing Linear Regression, now let’s check Ridge Regression, Ridge regression is a model tuning method that is used to analyze any data that suffers from multicollinearity. This method performs L2 regularization. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values being far away from the actual values.

After training the model on the training data, we got the same accuracy as Linear Regression gave (74%).



### Pros and Cons:

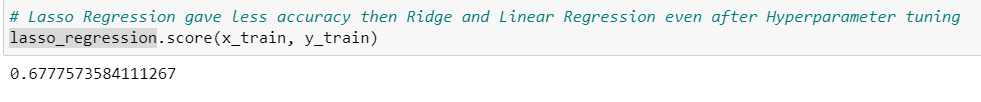
Avoids overfitting a model. They do not require unbiased estimators. They add just enough bias to make the estimates reasonably reliable approximations to true population values.

See, the accuracies of Linear and Ridge are the same, which means that this model will predict the same. Let’s move to the Lasso Regression and see what results it gives?

## **Lasso Regression:**

Lasso Regression is another regression model used to solve regression problems. Lasso regression is what is called the Penalized regression method, often used in machine learning to select the subset of variables. It is a supervised machine learning method. Specifically, LASSO is a Shrinkage and Variable Selection method for linear regression models. That is the variable selection process.

Oops, Lasso gave us less accuracy rate than Linear and Ridge regression models. We got the following accuracy rate using the Lasso Regression model.



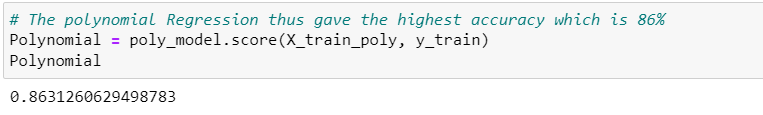
### Pros and Cons:

LASSO is a penalized regression method to improve OLS and Ridge regression. LASSO does shrinkage and variable selection simultaneously for better prediction and model interpretation. A disadvantage of LASSO: LASSO selects at most n variables before it saturates. LASSO can not do group selection.

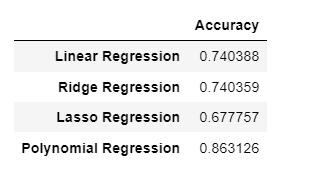
## **Polynomial Regression:**

Till now, we have trained three regression models, Linear Regression model gave us a 74% accuracy score, Ridge Regression gave the same accuracy as Linear Regression which is 74%, but the Lasso Regression gave is less accurate than the two above which is 67%. Now let’s check the 4th regression model called Polynomial Regression. Polynomial Regression is a form of linear regression in which the relationship between the independent variable x and dependent variable y is modeled as an nth degree polynomial. Polynomial regression fits a nonlinear relationship between the value of x and the corresponding conditional mean of y, denoted E(y |x).

Wow, after training this model, we got the highest accuracy score, we got an 86% accuracy rate.

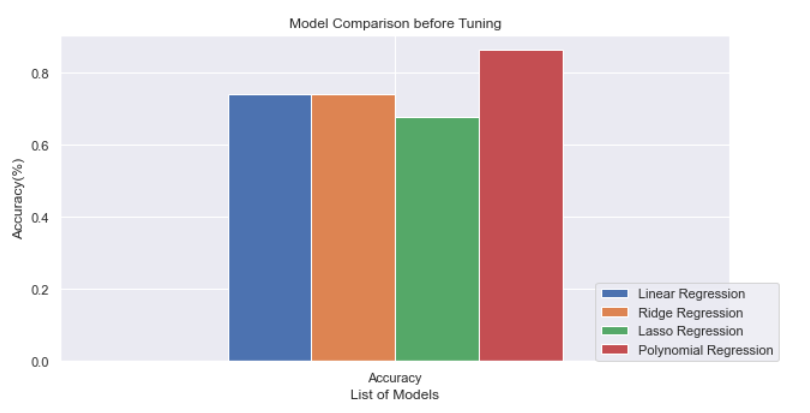


This is a good accuracy rate, based on the following comparison, we are going to predict the missing values of Country FINC in our dataset using the Polynomial Regression model.



We can see from the above comparison table that the Polynomial Regression has the highest accuracy rate than all the other regression models. We will use this model to predict our Test\_data.

The Following is the visual form of accuracy comparison.



## Future Recommendation:

As we have seen and analyzed our dataset has a huge number of missing values, these missing values are a big problem in machine learning and, the solution to avoid this type of problem from the future data is to collect the data carefully and completely. The data should be collected and handled in responsible hands. We have assumed missing values using statistical analysis which is not a good practice because the missing values are assumed by the machine, the results can be very different from the actual data.