Inflation Prediction model using Machine Learning

Omprakash Yadav[#1] Cynara Gomes[#2] Abhishek Kanojiya[#3] Abhishek Yadav[#4]
Department of Computer Engineering Xavier Institute of Engineering, Mumbai University
[1]omprakash.y@xavierengg.com
[2]c98gomes@gmail.com
[3]abkanojiya8998@gmail.com
[4]yadavabhishek3008@gmail.com

***Abstract* — Inflation can be defined as the loss of purchasing power of a fiat currency over a period of time. It plays a key role in designing the macro economic policy of central banks across the world. During periods of high growth rate, inflation rises and during periods of low or negative growth rates, deflation takes place. We undertook a study to find the correlation between inflation and CPI(Consumer Price Index). We have implemented various machine learning algorithms to find out pattern of Inflation over the time. This has helped us predicting the Inflation. Finally by using Karl Pearson's coefficient we have een successful in finding out top 5 factors which has affected inflation. Inflation is also a state when purchasing power of a currency i s falling,our model also predicts the value of money.To derive the best model for our dataset we have compared various algoritms like Linear Regression, Ridge Regression, Lasso Regression, XGboost Regression, Random Forest Regression.**

Keywords - Consumer Price Index, Linear Regression, Ridge Regression, Lasso Regression, XGboost Regression, Random Forest Regression.

I. INTRODUCTION

Severe inflation can cause a country's economic downturn. Therefore, inflation needs to be controlled. One of inflation control conducted by the government is to calculate and predict inflation every month using CPI indicators. The Consumer Price Index (CPI) is one of the most commonly used indicators to measure the inflation rate. Prediction with monthly frequency, could be too late, because inflation has been a few days and it is not known quickly. With the development of internet technology today, various data sources related to inflation are easily obtained in real-time. This data can be used for daily CPI prediction variables. Daily predictions will allow policy-makers to make better policies.

The daily inflation data can be obtained online or real time by using web scrapping techniques, data crawling techniques (twitter) and data API features. By using web scrapping techniques, we do not have to wait for web providers to generate API data. Scrapping the web will take the primary commodity price of many e-commerce and put it into CPI calculations and predictions. For example daily data that can be associated with inflation are commodities data, many hypermarkets publishing their prices on their sites. Social media data such Twitter and facebook data, many people discuss several topics including economics in social media. Twitter data and commodity prices in the e-commerce market can be used as predictive variables.

Many central banks use forecasting models based on machine learning methodologies for estimating various macroeconomic indicators, like inflation, GDP Growth and currency in circulation etc. Particularly those central banks that run an inflation targeting regime urgently require high quality inflation forecasts. A sound base for monetary policy decisions requires deep insights into the process that generates future inflation in general and the transmission mechanism from short-term interest rates to long term interest rates, exchange rates, real economic activity and inflation in particular. Inflation forecasting plays a major role in monetary policy and daily life. Based on different diagnostic and evaluation criteria, the best forecasting model for predicting inflation is identified. The results will enable policy makers and businesses to track the performance and stability of key macroeconomic indicators using the forecasted inflation.

In this report we have attempted monthly inflation of consumer price index (CPI) for India by using conventional time series forecasting based machine learning algorithms on the basis of monthly data between January 2013 to April 2018. The focus of

this study is to employ three methods: Linear Regression, Ridge Regression, Lasso Regression, XGboost Regression, Random Forest Regression for forecasting of inflation of CPI. We have carried out inflation forecast for the months till March 2018.

I. LITERATURE SURVEY

This review presents the works that influenced the choice of methods used in this paper. Works by Stock and Watson (2008) and Faust and Wright (2013) are frequently cited in the field of inflation forecasting. The basic models used in this study (random walk, autoregressions of orders 1 and p) were chosen based on the results obtained in these works. In the next two works by Chakraborty and Joseph (2017) and Garcia et al. (2017), ML methods are used to predict inflation. The work by Chakraborty and Joseph (2017) is a review of the best ML methods in terms of their practical application for solving several important problems faced by central banks.
In our study, we used the RF and boosting methods, which belong to the class of so-called 'ensemble' methods. The methods of this class are based on the general idea of using multiple training algorithms and subsequently combining them so that the final forecast is more accurate than any individual forecast. We expect that results of our study may be of practical value to the Bank of India, by expanding the range of potential methods for inflation forecasting.

II. PROBLEM STATEMENT

We undertook a study to find out if a correlation between inflation and interest rates exist. Companies employ constructive use of credit for their growth and increasing shareholder value. A rise or fall in interest rates depends on the corresponding inflation in a prior period. A thorough analysis of the inflation dataset could help in predicting inflation. A machine learning based model has been developed as a solution to predict interest rates using inflation. This can help companies to time their credit application and save few percentage points on the interest rate. Our aim is to predict inflation with the help of Consumer Price Index (CPI) for India by using conventional time series forecasting based machine learning algorithms on the basis of monthly data between January 2013 to April 2018. The focus of this study is to employ three methods: Linear Regression, Ridge Regression, Lasso Regression, XGboost Regression, Random Forest Regression for forecasting of inflation of CPI.

III. PROPOSED SYSTEM

A machine learning based model has been developed as a solution to predict interest rates using inflation. This can help companies to time their credit application and save few percentage points on the interest rate. Though, the savings in interest rate percentage may only translate to a few percentage points but it would correspond to huge savings for the company and boost its Return on capital employed(RoCE), Earnings per share (EPS) and Earnings before interest, tax, depreciation and amortization (EBITDA).

Data flow Diagram is given as follows:

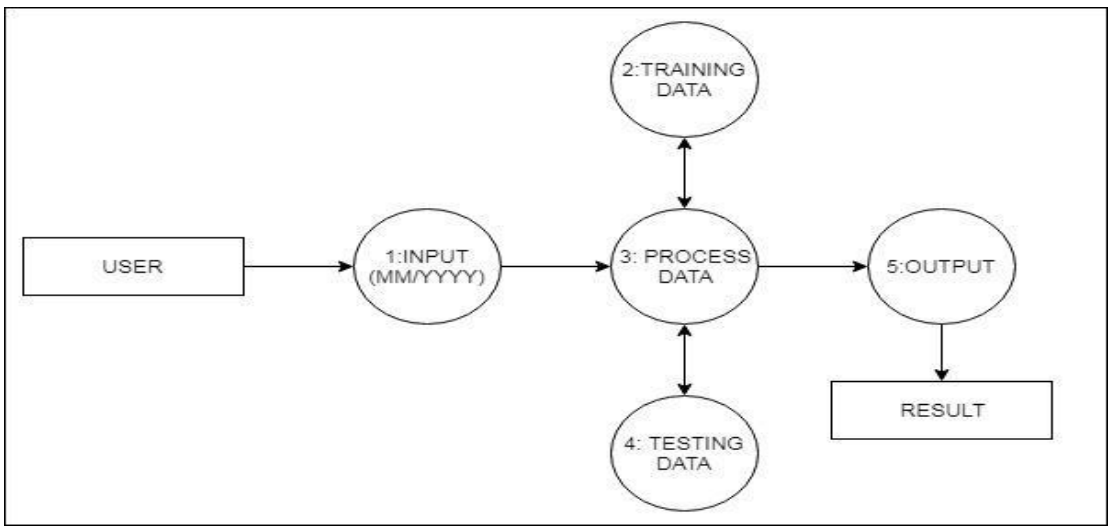

Fig. 1. 1-Level Data Flow Diagram

Block Diagram is given as follows:

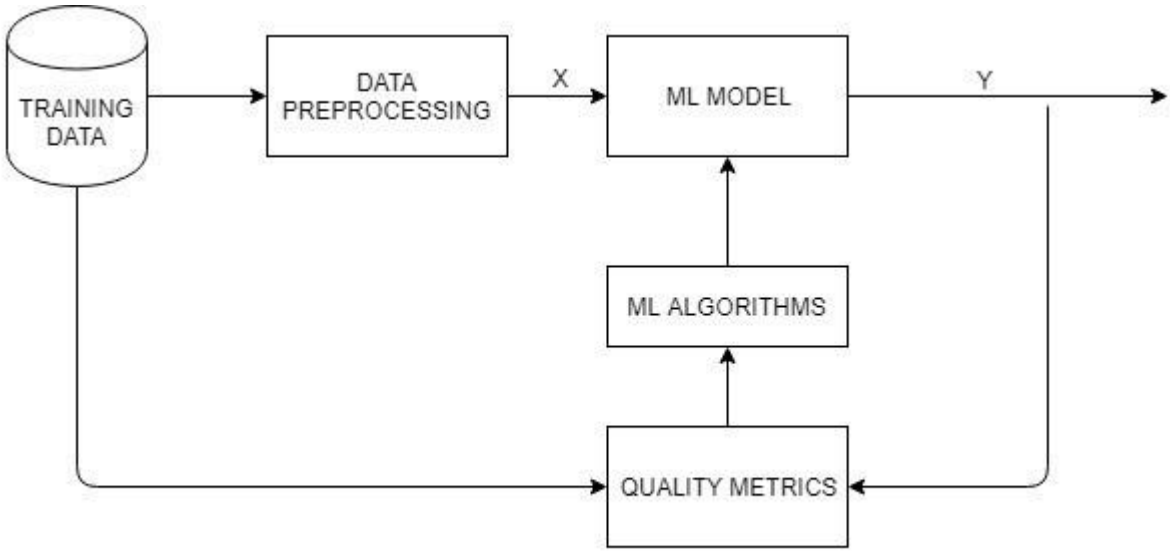

Fig. 2. Block Diagram

IV. METHODOLOGY

A .Data preprocessing

This research are used a daily commodity prices and exchange rates, can be obtained with a web scrapping techniques. This data contains Consumer Price data from www.data.gov.in .The data is segregated to rural and urban dataset of Consumer Price Index against time series.

Our challenge is to estimate the monthly CPI values. CPI is determined by using monthly models, it would be explained in Building model stages. Twenty-eight commodity prices were chosen as input variables, i.e. Cereals_and_products, Meat_and_fish,etc

B..Machine Learning Models

1. Lasso regression

Lasso regression uses the L1 penalty term and stands for Least Absolute Shrinkage and Selection Operator. The penalty applied for L2 is equal to the absolute value of the magnitude of the coefficients:

$$+\lambda \sum_{j=0}^{p} |w_j|$$

L1 regularization penalty term

Similar to ridge regression, a lambda value of zero spits out the basic OLS equation, however given a suitable lambda value lasso regression can drive some coefficients to zero. The larger the value of lambda the more features are shrunk to zero. This can eliminate some features entirely and give us a subset of predictors that helps mitigate multi-collinearity and model complexity. Predictors not shrunk towards zero signify that they are important and thus L1 regularization allows for feature selection (sparse selection).

2.Ridge Regression

Ridge regression uses L2 regularization which adds the following penalty term to the OLS equation.

$$+\lambda \sum_{j=0}^{p} w_j^0$$

L2 regularization penalty term

The L2 term is equal to the square of the magnitude of the coefficients. In this case if lambda(λ) is zero then the equation is the basic OLS but if it is greater than zero then we add a constraint to the coefficients. This constraint results in minimized coefficients (aka shrinkage) that trend towards zero the larger the value of lambda. Shrinking the coefficients leads to a lower variance and in turn a lower error value. Therefore Ridge regression decreases the complexity of a model but does not reduce the number of variables, it rather just shrinks their effect.

3.Linear Regression

The representation is a linear equation that combines a specific set of input values (x) the solution to which is the predicted output for that set of input values (y). As such, both the input values (x) and the output value are numeric.

The linear equation assigns one scale factor to each input value or column, called a coefficient and represented by the capital Greek letter Beta (B). One additional coefficient is also added, giving the line an additional degree of freedom (e.g. moving up and down on a two-dimensional plot) and is often called the intercept or the bias coefficient.

For example, in a simple regression problem (a single x and a single y), the form of the model would be:

y = B0 + B1*x

In higher dimensions when we have more than one input (x), the line is called a plane or a hyper-plane. The representation therefore is the form of the equation and the specific values used for the coefficients (e.g. B0 and B1 in the above example).

It is common to talk about the complexity of a regression model like linear regression. This refers to the number of coefficients used in the model.

When a coefficient becomes zero, it effectively removes the influence of the input variable on the model and therefore from the prediction made from the model (0 * x = 0). This becomes relevant if you look at regularization methods that change the learning algorithm to reduce the complexity of regression models by putting pressure on the absolute size of the coefficients, driving some to zero.

Now that we understand the representation used for a linear regression model, let's review some ways that we can learn this representation from data.

4.Random forest (RF)

The Random Forest model is based on bootstrap union of so-called decision trees and was first proposed by Breiman (2001). A binary decision tree is used as the basic algorithm for the Random Forest. A binary tree is a graph consisting of 'parent' or 'root' nodes (interim nodes) and 'leaf' terminal nodes. A decision tree is constructed in stages. The first stage is the optimal division of the entire sample X into two sub-samples: X1(i,p) = {x|xi ≤ p} and X2(i,p) = {x|xi > p} according to the specified quality function Q(X, i, p). Next, each of the sub-samples is iteratively broken down using the same principle. The breakdown stops when a stopping criterion is fulfilled.

After that, n leaf nodes are created, each of which corresponds to a certain preserved sub-sample (which could contain only a single point). If the regression analysis problem is solved (as in the case of modelling and forecasting of inflation), each leaf node is assigned the average value of the explanatory variable across the points (observations) in the corresponding sub-sample. The resulting tree is a connected graph of root nodes, each of which contains a threshold predicate that breaks down the sub-sample into two parts, and leaf nodes, each of which contains the predicted values of the explanatory variable. Usually, the quality function is specified in the following form:

$$-\frac{|X_2|}{|X|} H(X_2),$$

$$Q(\frown) = H(X) - \frac{|X_1|}{|X|}H(X_1)$$

where H(X) is an informativeness criterion.

The informativeness criterion shows how homogeneous the objects(observations) in the sub-sample are in terms of the explanatory variable. The main idea here is to maximise this homogeneity, and to break the sample down into two parts, in each of which the spread of values of the explanatory variable is minimal. Therefore, for the regression analysis problem, the quadratic deviation is used as a loss function and the following informativeness criterion is minimised:

$$H(X) \frown \frac{\hat{}}{|X|} \frown - c)\hat{}.$$

As we know, the minimum value of this kind of function is achieved when c is equal to the average value of the target (explained) variable.

$$H(X) \frown \frac{\hat{}}{|X|} \frown - \frac{\hat{}}{|X|} \frown)\hat{}.$$

In other words, the main goal of the algorithm at each stage is to minimise the sum of the weighted average variance within each of the two sub-samples resulting from the breakdown. Using the constructed tree, we can get predictions for values of the target variable with the new values of the explanatory variables.
The main advantage of the decision tree model is that the trees allow us to simply create effective (in terms of minimising the variance of the target variable) nonlinear dependencies. However, there is a serious drawback: the overfitting problem.

5. Gradient Boosting

The gradient Boosting model was first proposed by Friedman (2000). The idea behind the gradient boosting algorithm is similar to the idea behind the RF model: both algorithms are ensemble methods. The base Boosting model can represent any collection of models, but often, as with the RF model, a decision tree is selected. The main difference between the Boosting and the RF is that the base models are not trained independently, but rather taking into account the results of operation of the model on the previous iterations. The algorithm's operation can be described as follows:

1.) The first base model is trained on the whole sample:

$$b_1(x) = arg\min_b \sum_{i=1}^{l} (b(x_i) - y_i)^2.$$

2.) After the first step, the ensemble Boosting algorithm results in the first trained base model:

$$B_1(x) = b_1(x).$$

3.) Next, residuals are calculated that are equal to the difference between a true value and a predicted value based on the first Boosting model:

$$e_i^1 = y_i - B_1(x_i).$$

4.) The following model is trained on these residuals:

$$b_2(x) = arg\min_b \sum_{i=1}^{l} (b(x_i) - e_i^1)^2.$$

5.) We add a new model to the algorithm obtained in the previous step with a certain coefficient γ (0; 1]. This technique is called 'step reduction'. In this paper, the coefficient γ is equal to 0.2. This technique helps to improve the model's operation and avoid overfitting. A new model is obtained: ∈

$$B_2(x) = B_1(x) + \gamma b_2(x).$$

6.) Then, the algorithm is built up iterativelyuntil the end. As a result of operation of the algorithm, the final model is:

$$B_N(x)$$

$$= \sum_{i=1}^{N} \gamma^{i-1} b_i(x).$$

The algorithm terminates once all training cycles are completed. The number of training cycles is an important hyperparameter. Usually, in the training sample, errors tend to vanish as the number of iterations is increased. However, out of sample, too many iterations can actually increase errors as that the model begins to adjust for noise. In this paper, the number of iterations is 100. As already mentioned above, both untransformed (non-stationary) and transformed (stationary) data are used for the RF and Boosting models.

V. FUTURE SCOPE

Using inflation to predict interest rates

Companies employ constructive use of credit for their growth and increasing shareholder value. A rise or fall in interest rates depends on the corresponding inflation in a prior period. A thorough analysis of the inflation dataset could help in predicting inflation. A machine learning based model has been developed as a solution to predict interest rates using inflation. This can help companies to time their credit application and save few percentage points on the interest rate. Though, the savings in interest rate percentage may only translate to a few percentage points but it would correspond to huge savings for the company and boost its RoCE, EPS and EBITDA.

VI. RESULT AND DISCUSSION

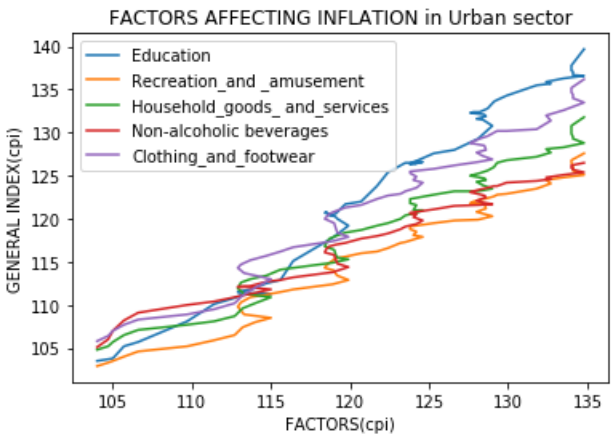

Fig 3:Fcators affecting Inflation(Urban)

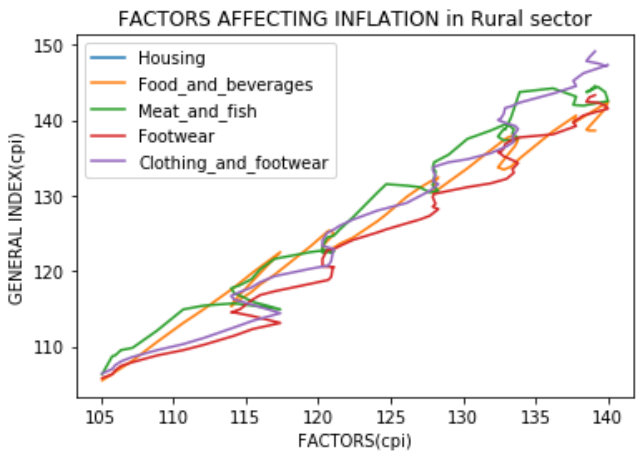

Fig 4:Fcators affecting Inflation(Rural)

The main results of this study are presented in Table 1. The Tables show the relative RMSE values of all specifications of each model for all forecast horizons Also, Table 1 gives the relative RMSE values can draw the following main conclusions from the results:

1)The use of ML methods can improve the quality of forecasting of Indian inflation compared to reference models (benchmarks) that use only lags of inflation as predictors. However, significant disadvantage of these models in comparison with classical econometric models is the loss of interpretability in the classical sense.

2)The ensemble methods (RF and Boosting) predict average inflation better than the base model from the second month onwards.

3)Among all three specifications of the RF model, the specification with untransformed data gave the best result when forecasting both inflation in individual months and average inflation over entire forecast horizon. Comparing the results of the two specifications of the Boosting model, with stationary and non-stationary data, leads to the same conclusion.

4)Relatively speaking, the regularised models provide less accurate forecasts over all forecasting horizons. The AR(1) model combined with LASSO gave results worse than the base AR(1) model, except for forecasts one month ahead.

5)The AR(1) model combined with LASSO gave the highest quality results when forecasting inflation over the horizon of one month. The models that use only lags of inflation as predictors (RW and AR) gave the same quality of forecasts over this horizon. The remaining methods were less accurate with respect to forecasts than the base model.

```
                     MSE       MAE  R_SQUARE
Ridge           6.880963  2.263657  0.796656
Linear          7.109186  2.301780  0.789912
XGBoost         4.608483  1.924700  0.863812
Random forest   4.908897  1.701424  0.854934
LASSO           6.838353  2.260129  0.797915

for rural:
                     MSE       MAE  R_SQUARE
Ridge           5.740988  1.870448  0.906644
Linear          5.688829  1.876631  0.907492
XGBoost         7.911959  2.546240  0.871341
Random forest   5.570834  1.624224  0.909411
LASSO           5.738247  1.875894  0.906688

for rural/urban:
                     MSE       MAE  R_SQUARE
Ridge           8.164872  2.271230  0.910248
Linear          7.997177  2.248933  0.912091
XGBoost        11.203740  2.987607  0.876843
Random forest   3.597780  1.621509  0.960452
LASSO           8.242583  2.294532  0.909394
```

Fig.5: Error Calculation

ALGORITHMS	ACCURACY
XGBoost	91.2363%
LASSO	86.7%
RIDGE	86.686%
LINEAR	86.683%
RANDOM FOREST	82.54%

Fig 6: Accuracy

VII. CONCLUSION

This paper aimed to prove the viability of ML methods for forecasting Indian inflation, compared to traditional methods. As the results demonstrate, this conjecture has been confirmed. Not all methods performed equally well in solving this problem: the regularisation models showed lower forecasting quality compared to the base model.

Both ensemble methods (RF and Boosting) showed results comparable to the basic AR(1) model in predicting monthly inflation. At the same time, they showed significantly better results when forecasting average inflation over a horizon of more than two months. We can therefore conclude that the RF and Boosting models show promise when applied to the task of forecasting Indian inflation.

This paper also addressed the issue of data transformation. According to our results, the RF and Boosting models perform better with untransformed rather than transformed data. This conclusion could be of use in further research using ML methods, since data transformation is a standard preparatory element of almost any empirical macroeconomic research using time series.

In addition to the models used in this article, there exist a number of other nonlinear ML algorithms, such as neural networks. Researchers would be well advised to also test these algorithms in future work on forecasting Indian inflation.

VIII. REFERENCES

[1] Doran Pandapotan Manik and Albarda, "Using Big Data in Statistics Indonesia," in International Conference on Information Technology System and Innovations (ICITSI), Badung Bali, 2015.

[2] Rumana Hossain and Shaukat Ahmed, "Forecast of Inflation in Bangladesh using ANN model," IJASCSE, vol. 2 No. 1, July 2013.

[3] Gour Sundar Mitra Thakur, Rupak Bhattacharyya, and Seema Sarkar Mondal, "Artificial Neural Network Based Model for Forecasting of Inflation in India," in Fuzzy Information and Engineering, 2015, pp. 87-100.

[4] Linyun Zhang and Jinchang Li, "Inflation Forecasting Using Support Vector Regression" in Fourth International Symposium on Information Science and Engineering, 2012.