MACHINE LEARNING METHODS IN EVALUATING THE IMPACT OF ECONOMIC FACTORS ON THE CONSUMER PRICE INDEX IN ALBANIA

Lule Basha [1], *Llukan Puka* [2]

[1] Department of Applied Mathematics, Faculty of Natural Science, University of Tirana, Tirana, Albania, lule.hallaci@fshn.edu.al, ORCID: 0000-0003-3790-601X

[2] Department of Applied Mathematics, Faculty of Natural Science, University of Tirana, Tirana, Albania, llukan.puka@fshn.edu.al, ORCID: 0000-0003-4121-3232

Abstract:
The Consumer Price Index (CPI) in Albania is a measure of inflation that tracks changes in the prices of a basket of goods and services typically purchased by urban households in the country. It is a vital economic indicator used to assess changes in the cost of living and the overall price level in Albania. There are several factors that affect the levels and progress of the CPI, among them we have chosen: Euro/Lek and USD/Lek exchange rates, import levels, the monetary base, and salary data, from January 2007 to September 2023. In this paper, we investigate the efficiency of machine learning methods in determining the factors that have the greatest impact on the CPI.

In our analysis, we assess the effectiveness of decision-tree models, Random Forest and XGBoost algorithms, in predicting the CPI behavior in Albania. Based on our empirical findings, we conclude that the monetary base and wages play a crucial role in influencing the CPI, with imports and exchange rates following closely in significance. Additionally, our results indicate that the Random Forest model demonstrates superior accuracy and demands less parameter tuning time compared to the alternatives. This research underscores the critical role of model selection in achieving precision and dependability in CPI forecasting. It underscores the immense potential of machine learning models in enhancing forecasting accuracy. The implications of this study are significant, as they can foster the creation of more precise and dependable forecasting models, equipping policymakers with a deeper understanding of economic stability.

Keywords: 4 Consumer Price Index, Exchange rate, Imports, Machine learning, Decision-tree.

INTRODUCTION

The Consumer Price Index (CPI) is a widely used economic indicator that measures the average change in the prices paid by consumers for a basket of goods and services over time.

It provides valuable insights into inflation, allowing individuals, businesses, and policymakers to monitor the cost of living and assess the impact of price changes on households. The CPI typically includes a representative sample of items such as food, housing, transportation, and healthcare, and it plays a crucial role in determining adjustments in wages, social security benefits, and other financial instruments to maintain the purchasing power of consumers in an evolving economic landscape.

The field of using machine learning methods for macroeconomic forecasting is a relatively recent development (Carvalho et al., 2019, Linardatos et al., 2020). Medeiros et al., (2019) explore advances in machine learning (ML) methods and the availability of new datasets to forecast U.S. inflation. Garcia et al., (2017) explore the application of ML techniques with Brazilian data. In Kohlscheen's study from 2021, he investigates the factors influencing CPI inflation using a straightforward yet computationally demanding machine learning approach. To be more precise, the research involves forecasting inflation in 20 advanced nations from 2000 to 2021, employing 1,000 regression trees established on six essential macroeconomic factors. This impartial, data-centric method yields relatively strong predictive performance for outcomes. Additionally, Costa et al., (2021) focused on oil price point and density forecasting using ML methods in their research. In the age of vast data availability, they seek to determine if modern automated tools can enhance forecast accuracy compared to conventional methods. They generate oil price point and density predictions using a total of 23 techniques.

Araujo and Gaglianone, (2023), are delving into the realm of machine learning techniques to enhance inflation prediction in Brazil. They introduce two novel contributions in their research. The first one is the introduction of a fresh method for combining

*Corresponding author:

quantiles, termed the “quantile regression forest model.” The second contribution involves the utilization of a hybrid machine learning strategy to develop innovative machine learning techniques. Their collection of top-performing forecasts encompasses a variety of methods, including forecast combinations, tree-based algorithms like random forest and XGboost, breakeven inflation, and expectations gathered from surveys.

Accurate inflation forecasting in a data-rich environment is challenging, with unanswered questions on extracting predictive information from correlated predictors. Traditional factor models have been used, but recent studies suggest machine learning models like random forests can help. Aras and Lisboa, (2022), promotes using machine learning models alongside or independently from factor models, incorporating new tree-based models, and combining feature selection techniques with Shapley values for concise inflation predictions. Experiments in volatile Turkey show that tree-based ensemble models offer both accuracy and explainable predictions.

Numerous prior investigations have centered on the variables influencing the Consumer Price Index (CPI) (Beckmann and Czudaj, (2013); Gao et al., (2014); Binner et al., (2010)). Nguyen et al., (2023), used various models, including multivariate linear regression (MLR), support vector regression (SVR), autoregressive distributed lag (ARDL), and multivariate adaptive regression splines (MARS), to forecast the US CPI from January 2017 to February 2022. The models considered factors like crude oil prices, world gold prices, and the federal fund effective rate. Evaluation metrics indicated that the MARS model outperformed the others in forecasting US CPI, which could aid the US government in shaping economic policies and promoting economic development. Several research studies have utilized regression models to examine the relationship between economic growth and environmental quality (Riofrío et la., (2020), Wang et al., (2023)).

In Gjika et al., (2020), the authors employ multivariate methods along with time series forecasting models to model the CPI indices in Albania. In a separate study by Gjika et al., (2016), they explore Albania’s economic growth and its links with the consumer price index (CPI), unemployment rate, inflation, and life expectancy. Basha and Gjika, (2023), assess the performance of machine learning and traditional models for forecasting time-series data of the Consumer Price Index (CPI), both for the total CPI and its 12 component groups. They evaluate the effectiveness of ARIMA models, Prophet models, and combinations of ARIMA and Prophet models with XGBOOST algorithms.

In Albania, similar to numerous other countries, the Consumer Price Index (CPI) fluctuates annually owing to shifts in the costs of products and services, variations in consumption habits, and various economic determinants. The National Institute of Statistics in Albania (INSTAT) frequently releases CPI data to monitor these fluctuations. The CPI serves multiple purposes, including measuring the inflation rate in Albania, acting as a deflator in National Accounts, Short-Term Statistics, and for adjusting the cost of living for households. It also plays a role in the monetary policy of the Central Bank of Albania. The CPI can fluctuate due to various economic and political factors, so it’s essential to use the most recent and accurate data when analyzing inflation trends and making financial decisions.

For this purpose, the objective of this paper is to examine various factors and assess their influence on the Consumer Price Index (CPI). Specifically, the factors under investigation include the Euro/Lek and USD/Lek exchange rates, import levels, the monetary base, and salary data, all within the timeframe from January 2007 to September 2023. The analysis was performed using machine learning: decision-tree, random forest and XGboost methods.

The rest of the paper proceeds as follows. Section 2 provides an explanation of the model methodologies, introducing the three mentioned models briefly. In Section 3, we present our data and findings. Section 4 offers conclusions, discussing potential implications of the current research and outlining future directions.

Materials and methods

Decision tree models: are a popular machine learning and data analysis technique used for classification and regression tasks. They are one of the earliest and simplest forms of predictive modelling, and they have been widely used in various fields, including data mining, artificial intelligence, and statistics. The concept of decision trees can be traced back to the early 1960s, with the development of the ID3 (Iterative Dichotomiser 3) algorithm by Ross Quinlan in 1986. The ID3 algorithm was one of the first practical decision tree algorithms and laid the foundation for subsequent variations like C4.5, C5.0, and CART (Classification and Regression Trees).

Decision trees are essentially a flowchart-like structure that helps make decisions based on the input data. They recursively split the dataset into subsets based on the most significant attribute or feature, resulting in a tree-like structure of decision nodes and leaf nodes. At each node, a decision is made based on a feature, leading to different branches (child nodes) and, eventually, a prediction or classification at the leaf nodes. Decision trees have several advantages, including their simplicity, interpretability, and the ability to handle both numerical and categorical data. However, they are prone to overfitting, which can be mitigated with techniques like pruning Han et al., (2012).

Random Forest: is an ensemble learning technique developed by Leo Breiman and Adele Cutler in the early 2000s. It's an extension of decision tree models and is widely used in machine learning for classification and regression tasks. Random Forest works by constructing multiple decision trees during the training phase and combining their predictions to make more accurate and robust predictions. The method: randomly select a subset of the training data (with replacement). This creates multiple subsets of data, known as bootstrap samples. For each bootstrap sample, a decision tree is constructed.

However, Random Forest introduces randomness into the process. At each node of the tree, instead of considering all features, it only considers a random subset of features. This helps to decorrelate the trees. When making predictions, each tree in the forest makes its prediction. For classification tasks, the class that receives the majority of the votes among the trees is the final prediction. For regression tasks, the predictions are averaged. Random Forest is a powerful and versatile ensemble method that often outperforms individual decision trees. It's less prone to overfitting compared to a single decision tree and it can handle both categorical and numerical data (Breiman,L., (2001).

XGBoost, which stands for "Extreme Gradient Boosting," is a machine learning algorithm known for its speed and performance in supervised learning tasks, particularly in regression and classification. It was developed by Tianqi Chen and released in 2014. XGBoost is a gradient boosting algorithm that builds an ensemble of decision trees to make predictions. It works by iteratively adding decision trees to correct the errors made by the previous trees. The algorithm starts with a single decision tree, which can be a shallow tree with a single node, representing the average target value of the entire dataset. It calculates the gradient of the loss function with respect to the model's current predictions. This gradient represents the direction and magnitude of the error. Then constructs a new decision tree to minimize the loss function. This tree is added to the ensemble and weighted according to its contribution to reducing the error. XGBoost includes regularization terms in its objective function to control model complexity and prevent overfitting. Gradient Boosting, Additive Tree Construction and Regularization are repeated for a specified number of iterations or until a stopping criterion is met. XGBoost offers several advantages, including high predictive accuracy, handling missing values, feature selection, and excellent performance on structured datasets. It's widely used in machine learning competitions and real-world applications.

Results

For the study conducted in this paper, we decided to focus on the Consumer Price Index (CPI) in Albania and various factors such as: Euro/Lek and USD/Lek exchange rates, import levels, the monetary base, and salary data, for the period January 2007 to September 2023. The initial official Consumer Price Index (CPI) was computed in December 1991, with December 1990 as the base period. Monthly CPI calculations began in 1992. The CPI's consumer basket was revised in 1993 based on the results of the Household Budget Survey (HBS) carried out by INSTAT. The base period was updated to December 1993, featuring a basket with 221 items categorized into 8 main groups. In 2000, a new HBS was conducted, leading to further revisions in the CPI basket. December 2001 became the new base period, encompassing 262 items classified into 12 main groups by COICOP classification, including food and non-alcoholic beverages, clothing and footwear, housing, transport, education, health, recreation, and more. This allowed for the analysis of how different sectors contributed to inflation. Currently, the base period is December 2020 (December 2020=100), with potential future changes in base period revisions. The data were obtained from Institute of Statistics (INSTAT) and the Bank of Albania.

Following an initial data analysis, it was determined that the dataset did not contain any missing values. Moreover, various analytical steps were taken, including time series anomaly detection, the examination of series components, checking stationarity properties, and providing descriptive statistics for the data. Subsequently, the data was divided into two sets: the first 80% was allocated for model training, and the remaining 20% was designated for model prediction. This choice of an 80-20 split was made considering data limitations, and since the data pertained to monthly time series, this split was guided by best practices. Once the data was organized, decision-tree, random forest, and XGBoost models were constructed to evaluate the factors impact on the Consumer Price Index (CPI).

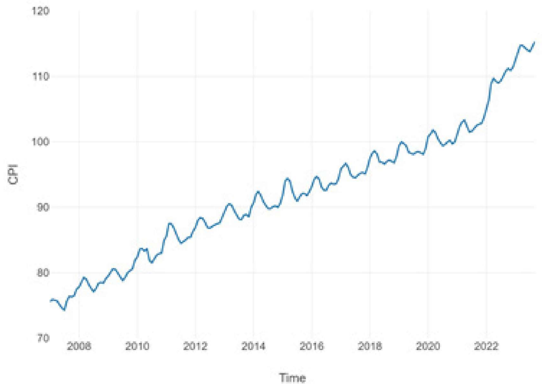


Figure 1. Consumer Price Index in Albania from January 2007 to September 2023. Source: Authors

In the year 2022, particularly in September and October, Albania experienced significant monthly fluctuations in the Consumer Price Index (CPI). The annual CPI change for September 2022 stood at 8.1%, marking a substantial increase compared to the 2.5% change a year prior. In September 2023, the CPI registered a value of 115.3, using December 2020 as the reference period. The annual CPI rate for September 2023 is 4.1%, showing a decline from the 8.1% rate in the previous year. The annual growth rate in September was primarily influenced by the prices of the "Food and non-alcoholic beverage" group, contributing a significant +2.80 percentage points, followed by the "Housing, water, electricity, and other fuel" group.

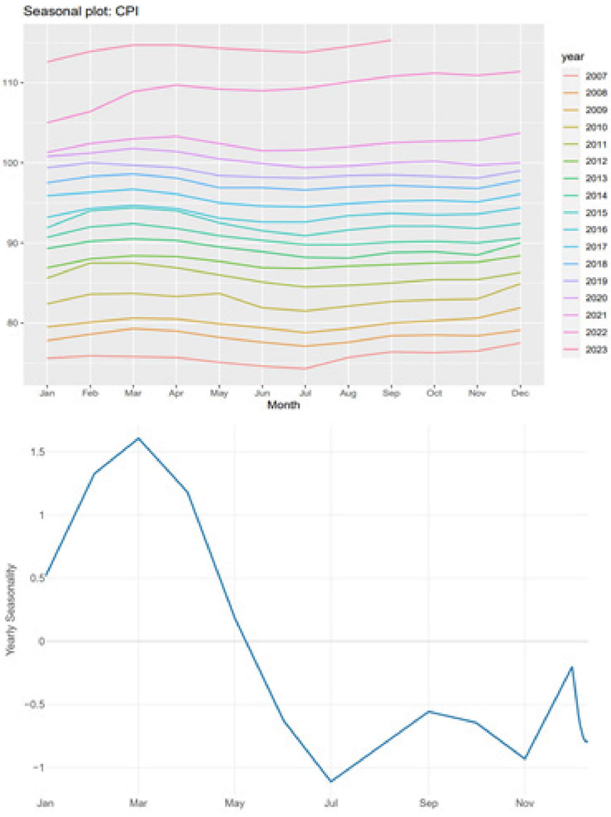


Figure 2. Seasonal plot. Source: Authors

The seasonal plot and time series graph of the Consumer Price Index (CPI) in Albania reveal a clear linear trend, for each year that passes, there is a higher value than the previous year for the index, and a seasonality pattern with a 12-month cycle. Figure 2 provides insights into assessing the annual cycle's shape consistency over time and identifying unique characteristics. The shapes of these annual cycles remain fairly similar, although the amplitude of the yearly fluctuations has decreased in recent years. Notably, March consistently exhibits a higher CPI compared to other months, while July consistently shows a lower CPI

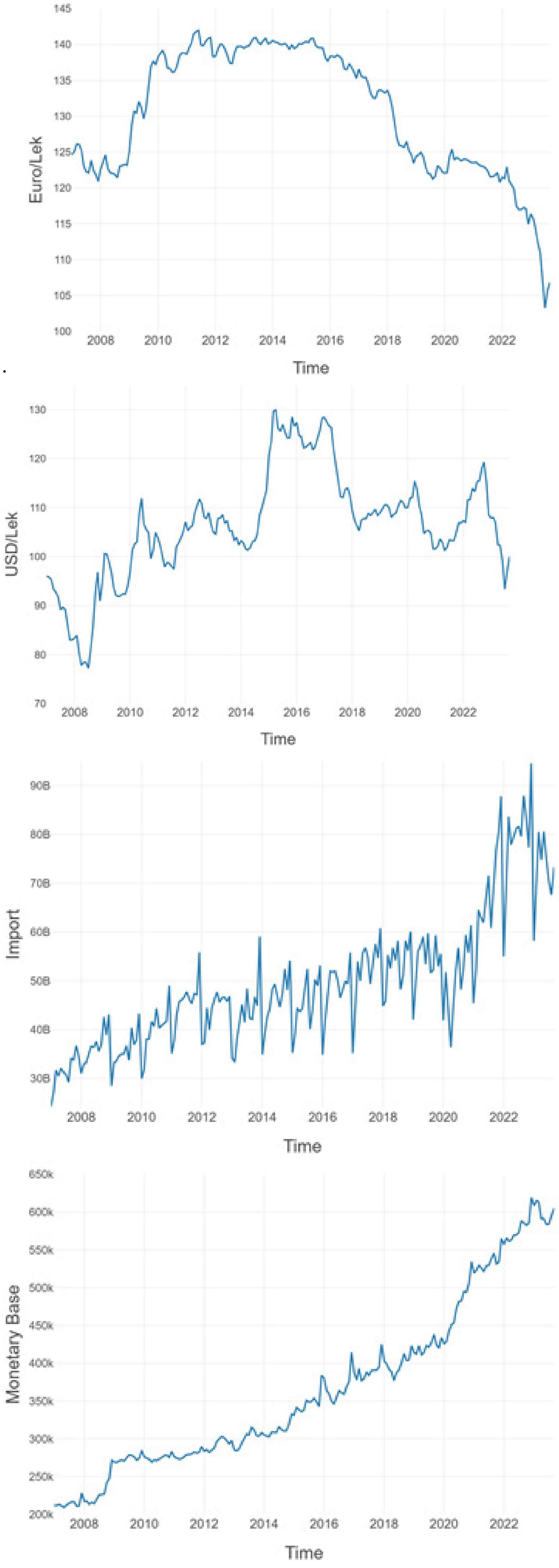
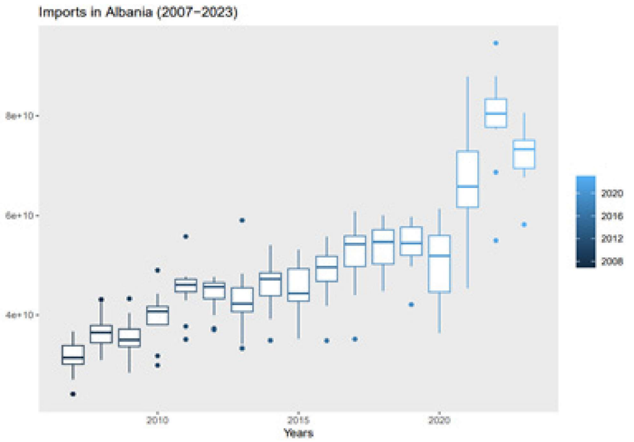


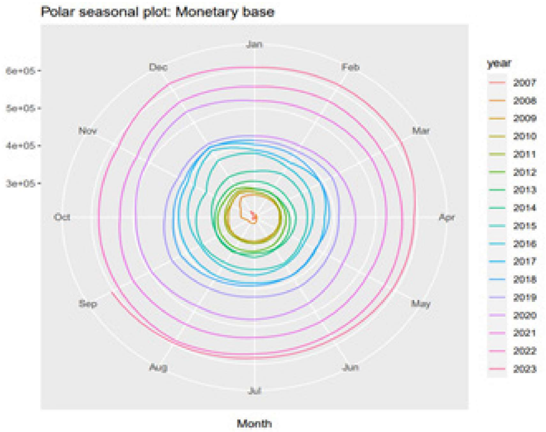
Figure 3. Currency Exchange Rate USD/Lek and Euro/Lek; Imports and Monetary Base in Albania from January 2007 to September 2023. Source: Authors

Exchange rates play a very important and critical role for most of the world's free market economies. The economic effects of the exchange rate changes are among the most controversial issues in the literature. The exchange rate's effect on the CPI is primarily through its impact on the prices of imported goods, inflation, interest rates, tourism, exports, and consumer sentiment. Exchange rate movements can contribute to both inflationary and deflationary pressures on the CPI, depending on the direction and magnitude of the currency's fluctuations. The Euro/Lek exchange rate hit its lowest point in July 2023, reaching 103.24, while its peak was recorded in June 2011 at 141.97. The average exchange rate over the years stands at 130.48. Conversely, the USD/Lek exchange rate reached its lowest point in July 2008, dipping to 77.24, while it reached its highest value in April 2015 at 129.97. The average exchange rate for the USD/Lek pairing over the years is 106.39.

Imports have a multifaceted impact on the CPI, affecting price levels, inflationary pressures, supply chain dynamics, competition, and consumer behaviour. From Figure 4 (a), we can see that the value of imports in Albania throughout the years under study is increasing. Imports had their lowest value in January 2007, and their highest value is in December 2022, with 94 billion. Changes in the prices and availability of imported goods can influence the overall inflation rate, making them an important component of CPI calculations. Imported inflationary pressures can contribute to overall inflation. When the prices of imported goods increase, it can lead to cost-push inflation. This, in turn, affects the overall CPI, as inflation measures the average change in prices for a basket of goods and services, including both domestically produced and imported items.



(a)



(b)

Figure 4. (a) Box-plot of Imports and (b) Polar seasonal plot of Monetary Base in Albania through 2007-2023. Source: Authors

The monetary base represents the amount of currency in circulation and the reserves held by commercial banks. When the central bank increases the monetary base by injecting more money into the economy (often through open market operations or quantitative easing), it can lead to an increase in the broader money supply. From polar seasonal plot Figure 4 (b), we can see that the monetary base in Albania throughout the years under study has a linear trend. An expansion of the money supply can contribute to demand-pull inflation, which, in turn, affects the CPI. Increases in the monetary base can contribute to inflation by expanding the money supply, lowering interest rates, stimulating economic activity, influencing exchange rates, and shaping inflation expectations. Conversely, a decrease in the monetary base may have deflationary effects on the CPI. The average salary in Albania during this time is 49068.7 Lek, with a minimum of 33750 and the highest value of 70905 reached during the year 2023.

In the second phase of the work, the impact that the factors taken in the study have on CPI was evaluated, applying the decision-tree, random forest and XGboost methods.

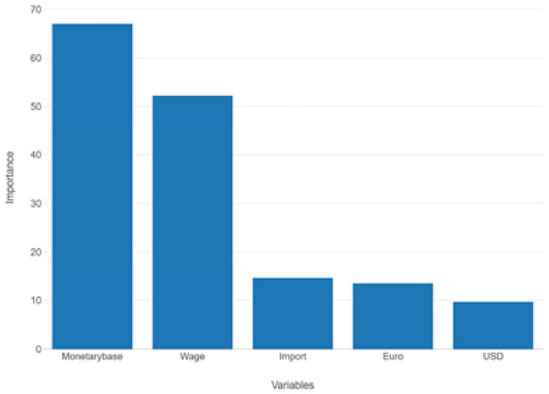


Figure 5. Variable importance plots. Source: Authors Variable

importance is a concept often used in statistics and machine learning to determine the significance or contribution of different variables (features or predictors) in a model, data analysis, or decision-making process Figure 5. It helps identify which variables have the most impact on the outcome of interest. We can clearly see that monetary base and wages have significant importance in the behaviour of the CPI in Albania, followed by imports and exchange rate. Variable importance is valuable for feature selection, model interpretation, and understanding the factors that contribute most to the model's predictive accuracy.

In the realm of predictive modelling, accuracy serves as a critical gauge for evaluating a model's effectiveness in making decisions. This accuracy can be assessed through in-sample data, which pertains to the data employed for model training, or out-of-sample data, which may encompass unseen data or a collection of observations utilized for testing. The performance of machine learning models can be evaluated using various metrics depending on the type of problem (classification, regression, clustering, etc.) and the specific goals of the analysis.

Table 1. *Performance of the models.* *Source: Authors*

MODEL	RMSE	R-squared
Decision-Tree	2.025191	0.959383
Random Forest	1.240062	0.984771
XGBOOST	3.12004	0.903596

Now that different models have been tried, we may compare the outcomes that have been found. Putting the outcomes in Table 1 together allows for a comparison to be made. As done for all the models, besides the RMSE parameter used in the fine tuning, we also report values for the R-squared for testing set. Root Mean Squared Error (RMSE), providing a more interpretable measure of error. R-squared (R2) measures the proportion of the variance in the CPI that is explained by the model. A higher R-squared indicates a better fit. Based on the results of Table 1, we can conclude that the best model for our data is the Random Forest model. Random Forest model shows the best performance both for training and testing data.

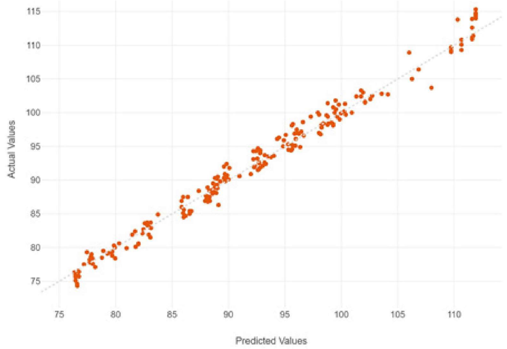


Figure 6. Actual versus predicted values, taken from Random-Forest model. Source: Authors

The graph of actual values versus predicted values is a visual tool that allows you to assess how well a model's predictions align with the actual data, Figure 6. Understanding the dispersion, pattern, and outliers in the graph helps in evaluating and improving the model's performance. Patterns in the data points form a linear pattern in our case indicates a strong, relationship between actual and predicted values. There is no significant evidence of outliers. So we can conclude that we have a highly accurate model.

Conclusions

This paper extensively analyzed the Consumer Price Index (CPI) in Albania, which serves as the official indicator of inflation, carrying immense importance for policymakers and economists. In the year 2022, particularly in September and October, Albania experienced significant monthly fluctuations in the Consumer Price Index (CPI). The study also investigated the influence of multiple factors, including exchange rates, imports, the monetary base, and salary data, over the period spanning from January 2007 to September 2023. To achieve our goals, we employed three distinct methodologies: decision-tree models, random forest models, and the XGBoost algorithm. Through variable importance analysis, we identified the key contributors to CPI fluctuations, with the monetary base and wages emerging as the most impactful, followed by imports and exchange rates.

Our model evaluation demonstrated that the Random Forest model outperformed the alternatives, exhibiting the lowest Root Mean Squared Error (RMSE), 1.240062 and the highest R-squared, 0.984771, values for both training and testing data. The graphical representation of actual versus predicted values from the Random Forest model confirmed its exceptional accuracy, displaying a clear linear pattern with no significant outliers.

In summary, this study provided valuable insights into the factors influencing the CPI in Albania and demonstrated the effectiveness of machine learning models, particularly the Random Forest model, in predicting CPI changes. These findings can be instrumental for policymakers and individuals in making informed economic decisions and understanding the dynamics of inflation in Albania.

References

Aras, S., & Lisboa, P. J. G. (2022). Explainable inflation forecasts by machine learning models. *Expert Systems with Applications, 207*(117982), 117982. https://doi.org/10.1016/j.eswa.2022.117982

Araujo, G. S., & Gaglianone, W. P. (2023). Machine learning methods for inflation forecasting in Brazil: New contenders versus classical models. Latin

American Journal of Central Banking, 4(2), 100087. https://doi.org/10.1016/j.latcb.2023.100087

Basha, L., Gjika, E. (2023) Forecasting Consumer Price Index With ARIMA, Prophet And Xgboost: A Comparative Analysis. IV. International Applied Statistics Congress (UYIK - 2023), September 26-29, 2023, Sarajevo / Bosnia and Herzegovina. ISBN: 978-975-7328-89-6

Beckmann, J., & Czudaj, R. (2013). Oil and gold price dynamics in a multivariate cointegration framework. International Economics and Economic Policy, 10(3), 453–468. doi:10.1007/s10368-013-0237-8

Binner, J.M., Tino, P., Tepper, J., Anderson, R., Jones, B., & Kendall, G. (2010) Does money matter in inflation forecasting? Phys. Stat. Mech. Appl., 389 (21), pp. 4793-4808 https://doi.org/10.1016/j.physa.2010.06.015

Breiman, L., & Cutler, A., (2000) Random forests – classification manual. https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#workings

Breiman, L., (2001). Random forests. Machine learning, 45(1):5–32 https://doi.org/10.1023/A:1010933404324

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. Electronics, 8(8), 832. https://doi.org/10.3390/electronics8080832

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. KDD '16, August 13-17, 2016, San Francisco, CA, USA https://doi.org/10.48550/arXiv.1603.02754

Costa, A. B. R., Ferreira, P. C. G., Gaglianone, W. P., Guillén, O. T. C., Issler, J. V., & Lin, Y. (2021). Machine learning and oil price point and density forecasting. Energy Economics, 102(105494), 105494. https://doi.org/10.1016/j.eneco.2021.105494

Gao, L., Kim, H., & Saba, R. (2014). How do oil price shocks affect consumer prices? Energy Economics, 45, 313–323. doi:10.1016/j.eneco.2014.08.001

Garcia, M. G. P., Medeiros, M. C., & Vasconcelos, G. F. R. (2017). Real-time inflation forecasting with high-dimensional models: The case of Brazil. International Journal of Forecasting, 33(3), 679–693. https://doi.org/10.1016/j.ijforecast.2017.02.002

Gjika (Dhamo), E., Basha, L., Allka, X., & Ferrja, A. (2020, June 9). Predicting the Albanian economic development using multivariate Markov chain model. 11th International Scientific Conference "Business and Management 2020". Presented at the 11th International Scientific Conference „Business and Management 2020", Vilnius Gediminas Technical University, Lithuania. https://doi.org/10.3846/bm.2020.581

Gjika, E., Zaçaj, O., & Gjecka, A. (2016). Projeksioni i indeksit të çmimeve të konsumit nëpërmjet metodave të serive kohore ((Rasti i Shqiperise). Buletini i Shkencave te Natyres, ISSN 2305-882X, Botimi Nr. 22, 138-147. http://buletini.fshn.edu.al/

Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques (3rd ed.). Elsevier, ISBN 978-0-12-381479-1 https://doi.org/10.1016/C2009-0-61819-5

Kohlscheen, E. (2022). What does machine learning say about the drivers of inflation? https://doi.org/10.48550/arXiv.2208.14653

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable AI: A Review of Machine Learning Interpretability Methods. Entropy (Basel, Switzerland), 23(1), 18. https://doi.org/10.3390/e23010018

Medeiros, M. C., Vasconcelos, G. F. R., Veiga, Á., & Zilberman, E. (2021). Forecasting inflation in a data-rich environment: The benefits of machine learning methods. Journal of Business & Economic Statistics: A Publication of the American Statistical Association, 39(1), 98–119. doi:10.1080/07350015.2019.1637745

Nguyen, T.-T., Nguyen, H.-G., Lee, J.-Y., Wang, Y.-L., & Tsai, C.-S. (2023). The consumer price index prediction using machine learning approaches: Evidence from the United States. Heliyon, 9(10), e20730. https://doi.org/10.1016/j.heliyon.2023.e20730

Quinlan, J. R. (1986). Induction of decision trees. Machine Learning, 1(1), 81–106. https://doi.org/10.1007/BF00116251

Riofrío, J., Chang, O., Revelo-Fuelagán, E.J., & Peluffo-Ordóñez, D.H., (2020) Forecasting the Consumer Price Index (CPI) of Ecuador: a comparative study of predictive models Int. J. Adv. Sci. Eng. Inf. Technol., 10 (3), pp. 1078-1084

Wang, Q., Zhang, F., & Li, R. (2023). Free trade and carbon emissions revisited: The asymmetric impacts of trade diversification and trade openness. Sustainable Development. doi:10.1002/sd.2703

Bank of Albania https://www.bankofalbania.org/home/

The National Institute of Statistics in Albania INSTAT https://www.instat.gov.al/