Machine Learning Forecasting of Cost-of-Living Trends in Cities across UK

Student Name: Zul Absar Ali
Student Number: 221056932
Supervisor Name: Hafiz Sherazi
Programme of study: MSc Computing and Information Systems

Abstract— ***The United Kingdom is currently witnessing dynamic shifts in housing, food, and transportation costs, exerting significant implications on both businesses and individuals. In response to these intricate dynamics, the integration of machine learning stands as a robust tool for forecasting the diverse facets of the cost of living, encompassing areas such as energy, housing, and sustenance. This study embarks on the task of predicting the ascending trajectories of energy, housing, food, and transportation expenditures across various UK cities through the adept utilization of advanced machine learning methodologies. By harnessing the predictive capabilities of machine learning algorithms, specifically employing the Random Forest Classifier and Decision Tree Regressor, this research aims to provide invaluable insights for strategic decision-making, optimal resource allocation, and prudent financial planning. Such insights enable stakeholders to adeptly navigate the evolving terrain of cost-of-living dynamics within the United Kingdom. The adopted methodology comprises comprehensive data collection, meticulous preprocessing, feature selection, application of machine learning algorithms, and rigorous model evaluation. While this study furnishes substantial predictive insights, it acknowledges potential constraints such as data scope and model intricacies, thus paving the way for future advancements in this domain.***

Keywords: Machine learning, Cost of living, random forest classifier, Decision Tree Regressor, feature selection

I. Introduction

The UK, in particular, has been enduring these changes with the citizens experiencing significant expenditure in various aspects such as housing costs, food prices, transport costs and others. Monitoring these changes is essential as it helps businesses, individuals, and policymakers formulate the most relevant decisions. The same concept of forecasting future trends and expectations based on the cost of living is applied in predicting the general cost of living. Consequently, the majority of the cities in the UK have been experiencing changes in various components that define the cost of living, such as inflation rates on different commodities, food prices, housing costs and others. These predictions based on these speculations can be widely utilized in guiding and advising any plan in finance decision-making and investments (Francis Devine et al. 2022). Machine learning is one of the critical tools that can be applied in this case for making extensive forecasting of the future of the cost of living. The predictions are based on its highly speculative algorithms in the sense of how a machine learning certain aspects under prediction can be in the future. Based on these predictions, the central theme under analysis in this context is the analysis of projections of the cost of living concerning energy costs, housing costs, food prices and transport costs of different cities in the UK with the help of machine learning techniques

II. Literature Review

A. *Machine learning*

Machine learning is a rapidly advancing field of computational techniques that aims to replicate human intelligence by acquiring knowledge from the surrounding environment. This field serves as a cornerstone in the era of extensive data, often referred to as "big data." El Naqa et al (2015)

B. *Supervisor learning*

Supervised learning has proven highly effective across various domains, including text and the web. It's often termed classification or inductive learning in the machine learning field. Comparable to human learning from past experiences, it uses historical data to enhance task performance. Yet, unlike humans, computers lack personal experiences, relying on historical data for learning. This data represents past occurrences in specific real-world contexts. Liu, B. (2011).

C. *Transport cost*

The paper introduces a methodology called DA4PT to analyze factors influencing bus ticket booking and purchases using a dataset of 3.23 million logs from a ticketing platform. Machine learning models accurately predict ticket purchases with 95% accuracy. Correlation rules between booking factors and purchases inform dynamic pricing strategies to increase ticket sales by 6% and overall revenue by 9%. The methodology provides insights into traveler behavior and optimization of bus ticketing systems to augment revenues. Branda et al (2022)

A potential drawback of this study is its narrow focus on the European long-distance coach market, which might restrict the applicability of results to diverse transportation settings or geographical areas. Future investigations could assess the approach's adaptability in various transportation markets and incorporate a wider array of variables to enhance the holistic comprehension of the efficacy of dynamic pricing strategies. Branda et al (2022)

Tiang wang in 2019 include in her research that Utilizing Artificial Intelligence (AI) and Machine Learning (ML), the research aims to uncover airlines' underlying pricing strategies and develop predictive models for price fluctuations. The models employed in this study encompass Random Forest and Support Vector Machine (SVM) algorithms, further enhancing its predictive capabilities. Leveraging datasets such as the Airline Origin and Destination Survey (DB1B), Air Carrier Statistics database (T-100), and macroeconomic data, the framework employs tailored machine learning algorithms. These algorithms generate predictive models, offering forecasts for average quarterly ticket prices across specific market segments defined by origin and destination pairs. Notably, the study demonstrates promising results, boasting an impressive, adjusted R-squared score of 86% on the testing dataset. Wang et al (2019)

However, it heavily hinges on the data quality and availability of chosen datasets, potentially affecting the models' robustness. Furthermore, the study lacks a deep exploration of machine learning model complexity, raising concerns about interpretability and overfitting. Wang et al (2019)

D. *Food prices*

The escalating global food price trend has garnered considerable attention among both practitioners and researchers. In response, this recent study delves into the empirical assessment of the influence of global factors on food price prediction, employing a comparative analysis between machine learning algorithms and time series econometric models. Encompassing eight international explanatory variables and spanning from January 1991 to May 2021, the research reveals that machine learning algorithms exhibit superior predictive performance compared to traditional time series econometric models. Ulussever et al (2023) The study's potential limitation lies in its focus on a limited set of explanatory variables, potentially overlooking the complexity of global food price dynamics. Additionally, its specific time frame and global variables might hinder broader applicability. Future research could expand variable inclusion and validate findings across varied contexts for a more comprehensive understanding of global food price trends. Ulussever et al (2023)

In the consumer market sector, essential daily commodities play a substantial role, encompassing a significant share of everyday expenditures. However, the volatility in their pricing poses challenges that reverberate through the calculus of living costs. Anticipating such fluctuations could offer proactive strategies to stabilize markets and optimize supply chains. This research leverages advanced Machine Learning algorithms, including Support Vector Machine, Random Forest, Bagging, AdaBoost, Gradient Boost, XGBoost, and LightGBM, to forecast the prices of six distinct daily commodities, spanning items like wheat, avocados, and dairy goods. Amin M.N(2020)

Elements such as variations in climate and shifts in demand contribute to price instability. Achieving precision in predictions demands a meticulous selection of variables, their transformations, and fine-tuning of parameters, thereby illuminating the intricacies inherent in this undertaking. Amin M.N(2020)

E. housing costs

understanding housing costs relies on predictive insights from machine learning algorithms and leveraging multi-sourced big geo data. This study addresses the appreciation of house prices, particularly the rate of cost appreciation, which has received limited attention despite the focus on housing costs. The research introduces a data integration framework to predict house price appreciation potentials by merging diverse data sources Kang et al. (2021),

Although the study offers a thorough approach to predicting house price appreciation through data fusion, a limitation lies in assuming equal contribution of all data sources to predictive accuracy. Variability in the importance of data sources across contexts might not be adequately addressed by the framework. Moreover, focusing solely on the Greater Boston Area restricts broader applicability to regions with distinct real estate trends. Future research could validate the framework's efficacy in various geographical settings and analyze how outcomes are influenced by variations in data source quality and quantity. Kang et al. (2021)

F. Energy costs

energy analysts have started consolidating ML (machine learning) strategies to speed up these advances. In this viewpoint, we feature ongoing advances in machine learning-driven energy research, frame flow and future difficulties and portray what is expected to utilize machine learning methods. Yao et al (2023), present a bunch of crucial execution pointers with which to look at the advantages of various machine learning sped-up work processes for energy research While machine learning (ML) offers advancements in energy research, drawbacks include potential opaqueness in decision-making, resource-intensive requirements, reduced focus on fundamental understanding, and ethical concerns related to biases and automated decision-making. Yao et al (2023),

This review examines cutting-edge approaches to electricity price forecasting (PF) and energy management, encompassing machine learning algorithms such as feedforward ANFIS, space vector regression (SVR), Binary Genetic Algorithm (BGA)Principal Component Analysis (PCA), and Firefly algorithm (FA). These methods enhance PF accuracy, optimize energy utilization, and reduce costs in an evolving energy landscape Yousaf et al. (2021) However, a potential drawback could be the complexity of integrating multiple algorithms and techniques, which might require a thorough understanding and expertise to implement effectively. Additionally, the paper's focus on specific buildings' cost reduction may limit the generalizability of its findings to broader energy management scenarios. Yousaf et al (2021)

III. METHODOLOGY

Predicting the cost of living involves a complex engagement with a myriad of socioeconomic influences. In the context of diverse UK cities, comprehending and foretelling cost of living patterns holds substantial importance for individuals, enterprises, and policymakers. This segment offers an intricate exploration of the comprehensive methodology adopted in this study, utilizing machine learning techniques to prognosticate and scrutinize the multifaceted cost of living aspects. The methodology is intricately organized to furnish a methodical approach that captures the subtleties of cost-of-living fluctuations.

The progression unfurls through an interlinked series of stages, each meticulously crafted to unearth invaluable insights and furnish precise forecasts. By embracing this systematic framework, the study strives to deepen comprehension of the fundamental dynamics steering alterations in the cost-of-living landscape.

A. Data collection

Numerous datasets pertinent to specific aspects, such as inflation, CPI, or CPIH, are readily accessible online through official sources like the National Statistical Office (ONS). Nonetheless, the requisites of machine learning necessitate substantial datasets. Thus, for the purpose of this report, data has been sourced from Numero to ensure the availability of a sufficiently large dataset for analysis.

B. Pre-Processing

During the pre-processing phase, null values within the dataset were addressed by employing an averaging approach in Microsoft Excel. This step was taken with the aim of enhancing the overall accuracy of the dataset. Subsequently, in the Python programming environment, outlier detection and removal procedures were applied to further refine the dataset. These measures were implemented to ensure the robustness and reliability of the data for subsequent analytical and modelling tasks.

C. Feature selection

The process involves initial feature selection, a crucial step in preparing machine learning models. Here, "features" represent independent variables, and "targets" are dependent variables. Excluding less impactful columns like "Year" and "City" from features enhances model efficiency. Using Scikit-Learn's `train_test_split`, data is smartly divided into subsets for training and testing. This empowers creating and rigorously evaluating a model, gauging its performance, generalizability, and real-world applicability.

D. Machine Learning

Two distinct algorithms, namely Random Forest Classifier and Decision Tree Regressor, were employed to analyze the dataset, each tailored to predict specific outcomes based on different independent variables. The Random Forest Classifier is an ensemble technique that combines multiple decision trees. Each tree is built from a bootstrapped data sample and uses a random subset of features at each decision point. This dual randomness enhances accuracy and robustness. One standout feature is its automatic probability prediction for class membership. This makes Random Forests efficient and effective for predictive tasks. The Random Forest Classifier algorithm was applied to predict outcomes related to the independent variable "cities," Breiman (2001)

Decision Tree Regression is a method for predicting numerical values by creating a tree-like structure that captures relationships between features and targets. It's effective for modelling complex interactions in data. Abdurohman et al(2022). The Decision Tree Regressor algorithm was utilized to forecast outcomes based on the independent variable "years." This segmentation of algorithms was guided by the goal of achieving granular predictions for individual years in the case of DecisionTreeRegressor and for specific cities in the case of Random Forest Classifier.

This strategic methodology provided a dual focus on specific cities and individual years, concurrently generating predictions that catered to both dimensions. Particularly, it facilitated the ability to forecast outcomes for designated cities within precise years, allowing for the extraction of targeted insights for distinct city-year pairings. To rigorously assess the algorithms' performance and bolster their dependability, a prudent dataset division strategy was adopted. This division involved segregating the dataset into training and testing subsets adhering to the widely recognized 80-20 ratio. In this partitioning, a substantial 80% of the data was allocated to train the models, while the remaining 20% served the crucial purpose of evaluating the models' predictive prowess. By adhering to such a partitioning framework, an essential foundation for measuring the algorithms' precision was established, along with insights into their generalization capabilities across diverse subsets of data.

E. Model Evaluation

The assessment of your machine learning models encompasses a range of performance metrics, each shedding light on different facets of their predictive capabilities. Accuracy, a fundamental metric, quantifies the proportion of correctly predicted instances out of the total, reflecting the overall alignment between predictions and actual outcomes. F1 Score, an amalgamation of precision and recall, is particularly useful when dealing with imbalanced class distributions or varying consequences of false positives and false negatives. Moving to the realm of regression, Mean Absolute Error (MAE) gauges the average absolute difference between predicted and actual values, indicating the typical magnitude of errors. Meanwhile, Mean Squared Error (MSE) emphasizing larger errors, calculates the average of squared deviations between predictions and actual values. The R-squared (R2) Score, a measure of fit, elucidates the proportion of variance in the target variable explained by the model's features. Now, interpreting these metrics in the context of your models: The Random Forest Classifier attains an impressive accuracy of around 86.84%, signifying a solid match between predictions and actual outcomes. The F1 Score of approximately 85% underscores the balance between precision and recall. Transitioning to the Decision Tree Regression, the model exhibits a Mean Absolute Error (MAE) of about 65% and a Mean Squared Error (MSE) of around 123% These figures illustrate the average magnitude of deviations and squared deviations, respectively. Lastly, the Decision Tree Regression's R-squared (R2) Score signifies its ability to account for about 55.05% of the variance in the data, offering insights into the model's fit to the observed data. These comprehensive metrics provide valuable guidance to optimize and refine your models, ensuring their robustness and applicability.

Random Forest Classifier	Accuracy	0.868421052631579
	F1 score	0.8578947368421052
Decision tree regressor	Mean absolute Error	0.6578947368421053
	Mean square Error	1.236842105263158
	R square	0.5504656430908632

1.Confusion matrix

A confusion matrix serves as a foundational tool in the realm of machine learning, serving the purpose of evaluating the performance of a classification model. It presents a clear visual layout of how effectively the model's predictions align with the actual class labels. The structure of the matrix resembles a table, where rows correspond to the true data classes, and columns signify the predicted classes. The principal diagonal of the matrix represents accurate predictions for each class, whereas the off-diagonal elements expose instances where the model made incorrect predictions. By delving into the details of the confusion matrix, metrics such as accuracy, precision, recall, and the F1 score can be calculated, providing deeper insights into the model's competencies and limitations. (Géron 2019)

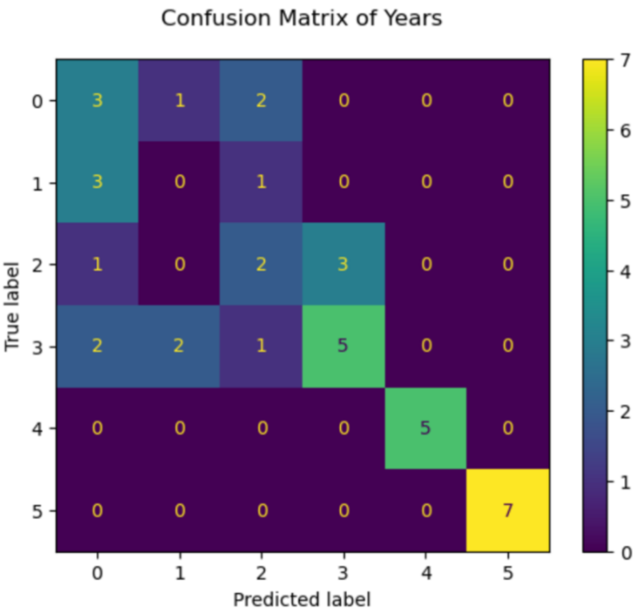


Fig (1) Confusion matrix for years

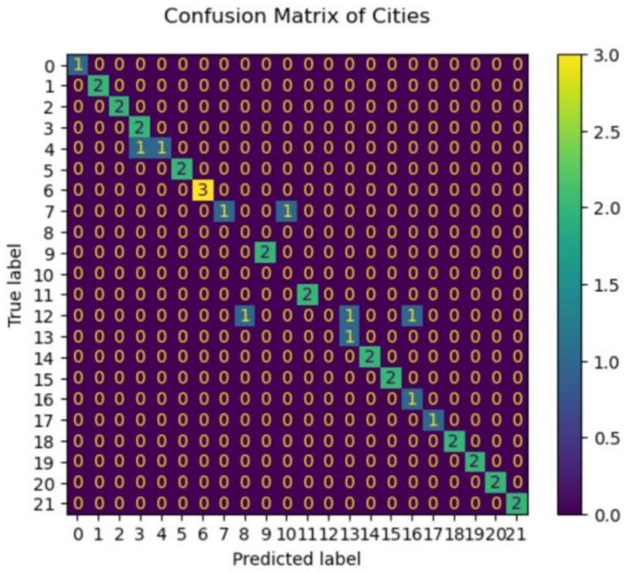


Fig (2) Confusion matrix for cities

2. *Correlation coefficient*

A correlation coefficient is a numerical metric used in statistics to gauge the extent of connection or link between two variables. It gauges the extent to which changes in one variable match changes in another. This coefficient generally falls between -1 and 1: -1 signifies a strong negative correlation, 1 indicates a strong positive correlation, and 0 implies no discernible connection between the variables. This metric finds broad application in statistical analyses, aiding in comprehending the intensity and direction of the relationship between variables. (Géron 2019).

The Fig (3) we can see in the graph white area is denoted as 1 and the peach colour is denoted as -2 Each variable is compared to itself, resulting in a diagonal line that runs from the top-left to the bottom-right corner of the heatmap. This diagonal line is shaded in white, representing a perfect correlation of 1. This makes sense because a variable is perfectly correlated with itself. The other cells in the heatmap show the correlation values between pairs of variables, using different colours to indicate the strength and direction of their relationships.

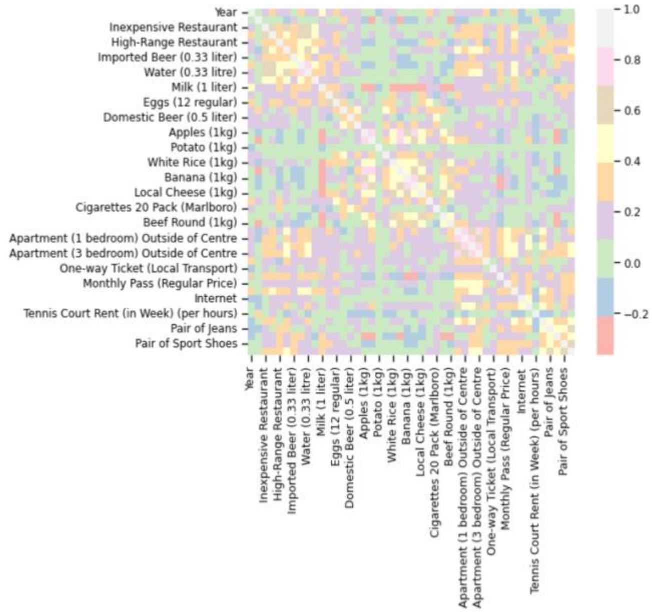
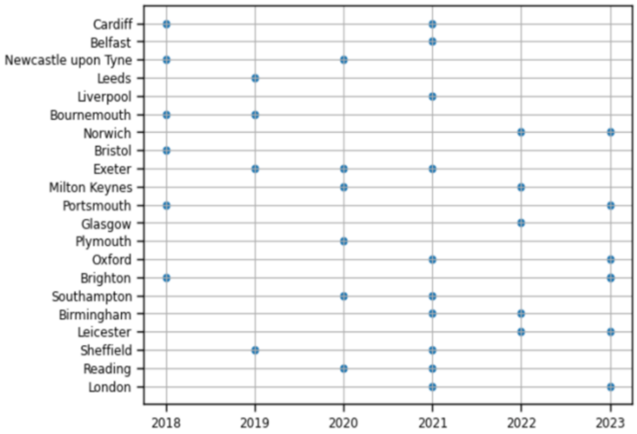


Fig (3)

F. *Data visualization*

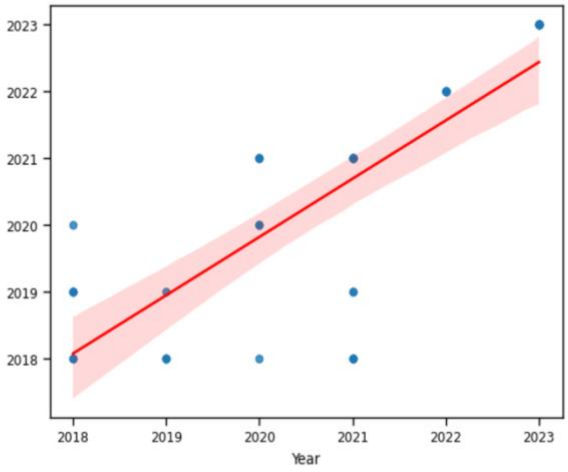
Fig (1) gives an analysis that helps us see how things change over time and which cities might be good places to live in. The X-axis denotes the years that started from 2018 to 2023. The grid in the graph assists us to have a precise allocation of the points this was the result of using a decision tree regressor. And the Y-axis of the graph shows us different cities. For the cities, we used a Random Forest Classifier as it was a string.

By combining the outcomes of these two models (Random Forest Classifier and decision tree regressor, we yield a graph that concurrently provides dual insights. Specifically, it elucidates the potential temporal context, indicating plausible timeframes. Simultaneously, it accentuates cities that hold promise as commendable habitats during those specific periods. Thus, this graph assumes the role of sequential and spatial mapping, facilitating a nuanced comprehension of the congruent instances and locales wherein diverse cities might manifest as suitable abodes.



(Fig 4)

Underfitting arises when the model struggles to achieve a sufficiently low error value within the training dataset. Zhang et al (2019). When we applied a decision tree regressor to predict 'years' within the training dataset, the resultant output exhibited signs of underfitting. This could be attributed to the model's oversimplified nature, failing to capture the intricate patterns and variations present in the data. (Fig 2)



(Fig 5)

IV. Limitation

This study significantly contributes to the field by utilizing machine learning techniques for predictive modelling of the cost of living. Nevertheless, a comprehensive acknowledgement of its limitations is imperative. An inherent constraint lies in the limited scope and accessibility of the employed datasets, which predominantly encompass inflation, CPI, and CPIH. The exclusivity of these sources might inadvertently exclude vital socioeconomic variables that intricately shape the cost-of-living dynamics. Furthermore, the study's temporal framework, while pertinent, might inhibit the incorporation of broader economic trends that potentially wield influence over predictive accuracy.

The study's focal point is on collective predictions encompassing energy costs, housing expenditures, food prices, and transport outlays. A promising avenue for enhancement lies in segmenting predictions into discrete facets of the cost-of-living spectrum. This nuanced stratification holds the potential to unravel the intricate interplay among diverse determinants that collectively mold the broader cost of the living landscape, particularly across heterogeneous urban centers.

To bolster the analytical rigour and fidelity, embracing diverse indices and datasets warrants attention. Enriching the predictive model with an expanded repertoire of economic indicators, labor market dynamics, and demographic insights could engender a more holistic comprehension of the intricate cost-of-living milieu. The incorporation of extrinsic factors like environmental sustainability and quality of life may further amplify the evaluative framework, furnishing a comprehensive assessment of the multifaceted influences steering individuals' day-to-day fiscal commitments.

In summation, while the existing study offers valuable insights via machine learning applications, untapped potential for advancement remains. This potential resides in the amplification of data sources, the subdivision of predictions into discrete cost-of-living dimensions, and the assimilation of a wider array of indices and variables. Such an encompassing approach has the potential to yield a more nuanced and precise portrayal of the evolving cost of living dynamics within diverse UK municipalities.

V Conclusion

In summary, this study has explored the dynamic field of predictive modelling concerning the cost of living in various UK cities, utilizing the powerful capabilities of machine learning. The ever-changing landscape of living expenses, covering housing, energy, food, and transportation costs, carries substantial importance for individuals, businesses, and policymakers. Through a comprehensive methodology, the study successfully anticipated the trajectories of these critical factors, shedding light on the continuously evolving financial dynamics.

Despite its valuable insights, this study is not without limitations. The reliance on specific datasets, while pertinent, may inadvertently disregard essential socioeconomic variables contributing to the nuanced fluctuations in the cost of living. The focus on a specific time frame could potentially obscure broader economic shifts that influence the accuracy of predictions. Moreover, the collective predictions encompassing various dimensions of the cost of living might benefit from a more precise approach, acknowledging the distinct influences on each facet. Nevertheless, this study lays the groundwork for improvement and expansion.

Incorporating a broader array of indices, including economic indicators, labor market trends, and demographic insights, could lead to a more comprehensive understanding of the multifaceted cost-of-living landscape. The integration of external factors like environmental sustainability and quality of life further enriches the evaluative framework, capturing the diverse influences shaping individuals' financial commitments.

As the cost of living continues to evolve and machine learning techniques advance, there is significant potential for refining predictive accuracy. By enhancing data sources, segmenting predictions, and embracing a more comprehensive range of indices, predictive models could yield more nuanced, precise, and actionable insights. Such an approach empowers individuals, businesses, and policymakers to make informed decisions amid the ever-changing economic dynamics and shifting trends in the cost of living. In a world where economic decisions have far-reaching impacts, harnessing machine learning to predict the cost of living is invaluable. This study marks a substantial step in that direction, with the assurance of ongoing enhancement, increased accuracy, and enriched insights in the continued pursuit of understanding and navigating the intricates of the cost of living in the diverse urban landscape of the United Kingdom

VI References

1. Abdurohman, M., Putrada, A.G., & Deris, M.M. (2022). A Robust Internet of Things-based Aquarium Control System using Decision Tree Regression Algorithm.

2. Amin, M.N. (2020). Predicting Price of Daily Commodities using Machine Learning.

3. Branda, F., Marozzo, F., & Talia, D. (2020). Ticket Sales Prediction and Dynamic Pricing Strategies in Public Transport. Big Data Cogn. Comput.

4. Breiman, L. (2001). Random Forests. Machine Learning,

5. El Naqa, I., Murphy, M.J. (2015). What Is Machine Learning?. In: El Naqa, I., Li, R., Murphy, M. (eds) Machine Learning in Radiation Oncology. Springer, Cham.

6. Francis-Devine, B., Bolton, P., Keep, M., & Harari, D. (2022). The rising cost of living in the UK. Research Briefing, House of Commons Library, October.

7. Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media.

8. Harley, Q., & Leeds, M.H. (2023). The Cost of Living Crisis.

9. Kang, Y., Zhang, F., Peng, W., Gao, S., Rao, J., Duarte, F., & Ratti, C. (2021). Understanding house price appreciation using multi-source big geo-data and machine learning. Land Use Policy.

10. Liu, B. (2011). Supervised Learning. In: Web Data Mining. Data-Centric Systems and Applications. Springer, Berlin, Heidelberg.

11. Saharan, S., Bawa, S., & Kumar, N. (2020). Dynamic pricing techniques for Intelligent Transportation System in smart cities: A systematic review. Computer Communication.

12. Ulussever, T., Ertuğrul, H. M., Kılıç Depren, S., Kartal, M. T., & Depren, Ö. (2023). Estimation of Impacts of Global Factors on World Food Prices: A Comparison of Machine Learning Algorithms and Time Series Econometric Models. Foods, 12, 873.

13. Wang, T., Pouyanfar, S., Tian, H., & Tao, Y. (2019). A Framework for Airfare Price Prediction: A Machine Learning Approach. 2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science

14. Yao, Z., Lum, Y., Johnston, A., MejiaMendoza, L.M., Zhou, X., Wen, Y., Aspuru-Guzik, A., Sargent, E.H., & Seh, Z.W. (2023). Machine learning for a sustainable energy future. Nature Reviews Materials.

15. Yousaf, A., Asif, R.M., Shakir, M., Rehman, A.U., Alassery, F., Hamam, H., & Cheikhrouhou, O. (2021). A Novel Machine Learning-Based Price Forecasting for Energy Management Systems. Sustainability.

16. Zhang, H., Zhang, L., & Jiang, Y. (2019). Overfitting and Underfitting Analysis for Deep Learning Based End-to-end Communication Systems. 2019 11th International Conference on Wireless Communications and Signal Processing (WCSP).