COUB01
F014621

Predicting Inflation With Machine Learning

by

Ben M. Taylor

A Project Report

Submitted in partial fulfilment
of the requirements for the award of

B.Sc.
in
Computer Science
of
Loughborough University

2nd December 2024

Acknowledgements

I would like to thank my supervisor Mohamad Saada for his help and guidance throughout this project.

Abstract

In October 2022, the UK hit an inflation rate of 11.1%, the country's highest in over 40 years. Now more than ever, the ability to accurately predict inflation and other financial indicators is a crucial skill required by the government and the individual to prepare themselves for the future financially. In a time where Artificial Intelligence and Machine Learning are ever flourishing, it is only natural to attempt to use these tools at our disposal to predict and combat the issues we face.

In this paper I will attempt to predict inflation through the use of machine learning eventually presenting my findings and evaluations.

Keywords: Inflation, Artificial Intelligence, Machine Learning

Contents

List of Figures

List of Tables

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] Steven B Achelis. Technical analysis from a to z, 2001.

[3] Oguz Akbilgic, Hamparsum Bozdogan, and Mehmet Erdal Balaban. A novel hybrid rbf neural networks model as a forecaster. *Statistics and Computing*, 24:365–375, 2014.

[4] Mariette Awad and Rahul Khanna. *Machine Learning*, pages 1–18. Apress, Berkeley, CA, 2015.

[5] World Bank. World devlopment indicators, 2023.

[6] David Beckett. Consumer price inflation, uk: October 2023, Nov 2023.

[7] Leo Breiman. Bagging predictors. *Machine learning*, 24:123–140, 1996.

[8] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[9] Lijuan Cao and Francis E.H. Tay. Financial forecasting using support vector machines. *Neural Computing and Applications*, 10(2):192, 2001. Cited by: 272.

[10] François Chollet et al. Keras. `https://keras.io`, 2015.

[11] G. Ciaburro, V.K. Ayyadevara, and A. Perrier. *Hands-On Machine Learning on Google Cloud Platform: Implementing smart and efficient analytics using Cloud ML Engine*. Packt Publishing, 2018.

[12] John Harold Clapham. *The bank of England*. Number 7. CUP Archive, 1939.

[13] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20:273–297, 1995.

[14] Jeff Dean, David Patterson, and Cliff Young. A new golden age in computer architecture: Empowering the machine-learning revolution. *IEEE Micro*, 38(2):21–29, 2018.

[15] Jeffrey Dean. The deep learning revolution and its implications for computer architecture and chip design. *ArXiv*, abs/1911.05289, 2019.

[16] The Economist. Our big mac index shows how burger prices are changing.

[17] Brigid Francis-Devine. Rising cost of living in the uk. `https://commonslibrary.parliament.uk/research-briefings/cbp-9428/`, 2023. [Accessed 17-12-2023].

[18] Google. Using the python client library — ai platform training — google cloud.

[19] C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969.

[20] Nicholas Geoffrey Lemprière Hammond. *Alexander the Great*. Chatto & Windus London, 1981.

[21] Bruce E. Hansen. Threshold effects in non-dynamic panels: Estimation, testing, and inference. *Journal of Econometrics*, 93(2):345–368, 1999.

[22] Cynthia Harrington. Fundamental vs. technical analysis, 2003.

[23] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, sep 2020.

[24] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995.

[25] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

[26] Chip Huyen. Design a machine learning system, 2022.

[27] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167, 2015.

[28] Markus "Löning, Franz Király, Tony Bagnall, Matthew Middlehurst, Sajaysurya Ganesh, George Oastler, Jason Lines, Martin Walter, ViktorKaz, Lukasz Mentel, chrisholder, Leonidas Tsaprounis, RNKuhns, Mirae Parker, Taiwo Owoseni, Patrick Rockenschaub, danbartl, jesellier, eenticott-shell, Ciaran Gilbert, Guzal Bulatova, Lovkush, Patrick Schäfer, Stanislav Khrapov, Katie Buchhorn, Kejsi Take, Shivansh Subramanian, Svea Marie Meyer, AidenRushbrooke, and Beth" Rice. "sktime/sktime: v0.13.4", 2022.

[29] Ryszard Stanislaw Michalski, Jaime Guillermo Carbonell, and Tom M Mitchell. *Machine learning: An artificial intelligence approach*. Springer Science & Business Media, 2013.

[30] John J Murphy. *Technical analysis of the financial markets: A comprehensive guide to trading methods and applications*. Penguin, 1999.

[31] Eir Nolsøe. Bank of England's Covid money-printing spree 'drove up inflation' — telegraph.co.uk. `https://www.telegraph.co.uk/business/2023/04/18/bank-of-england-covid-money-printing-drove-up-inflation/#:~:text=During%20the%20pandemic%2C%20the%20Bank,to%20a%20record%20%C2%A3895bn.`, 2023. [Accessed 17-12-2023].

[32] OECD. Data warehouse. 2014.

[33] Ceyda Oner. Inflation: Prices on the rise, Jul 2019.

[34] The pandas development team. pandas-dev/pandas: Pandas, feb 2020.

[35] Michael Parkin. *Inflation*, pages 1–10. Palgrave Macmillan UK, London, 2016.

[36] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[37] Barbara Rockefeller. *Technical analysis for dummies*. John Wiley & Sons, 2019.

[38] F. Rosenblatt. The perceptron - a perceiving and recognizing automaton. Technical Report 85-460-1, Cornell Aeronautical Laboratory, Ithaca, New York, January 1957.

[39] Ehsan Sadrfaridpour, Korey Palmer, and Ilya Safro. Aml-svm: Adaptive multilevel learning with support vector machines, 2020.

[40] Tom Seegmiller. 10 metrics to measure the financial efficiency of your organization, Sep 2023.

[41] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[42] Yajiao Tang, Zhenyu Song, Yulin Zhu, Huaiyu Yuan, Maozhang Hou, Junkai Ji, Cheng Tang, and Jianqiang Li. A survey on machine learning models for financial time series forecasting. *Neurocomputing*, 512:363–380, 2022.

[43] Michael C Thomsett. *Getting started in fundamental analysis*. John Wiley & Sons, 2006.

[44] Ruey S Tsay. *Analysis of financial time series*. John wiley & sons, 2005.

[45] Ahmed. S. Wafi, Hassan Hassan, and Adel Mabrouk. Fundamental analysis models in financial markets – review study. *Procedia Economics and Finance*, 30:939–947, 2015. IISES 3rd and 4th Economics and Finance Conference.

[46] Michael L. Waskom. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021.

[47] Zhongyuan Wei. *A SVM approach in forecasting the moving direction of Chinese stock indices*. Lehigh University, 2012.

[48] Daniela Hristova * Wojciech W. Charemza and Peter Burridge. Is inflation stationary? *Applied Economics*, 37(8):901–903, 2005.

[49] Jeffrey Marc Wooldridge. *Introductory Econometrics: A Modern Approach*. ISE - International Student Edition. South-Western, 2009.

[50] G. Udny Yule. Why do we sometimes get nonsense-correlations between time-series?–a study in sampling and the nature of time-series. *Journal of the Royal Statistical Society*, 89(1):1–63, 1926.

Chapter 1

Introduction

1.1 Motivation

Prices for goods and services are ever-changing, constantly affecting individuals, organizations, and governments. These clients would all benefit from the ability to predict the rise and fall of prices as it can impact their choices to spend, invest, or save. Ultimately allowing them to make the most out of the resources they have.

However, the value of inflation is an incredibly difficult indicator to predict with even establishments like the Bank of England, who have direct control over interest rates, failing to correctly forecast it. With the ability to accurately predict inflation having so many stakeholders and huge organizations working on the problem, I am under no illusion that an undergraduate like myself will be able to find a solution better than all those before me. Instead, I aim to understand which machine learning (ML) methods produce the best results when faced with this task and potentially find out why this is the case. This project's results will ultimately contribute to the ongoing research into inflation forecasting and help clarify which ML methods may be most suitable.

1.2 Aims and Objectives

1.2.1 Aims

This project aims to create multiple machine learning models and compare their predictive abilities when it comes to UK inflation. The findings will then be presented in this report, outlining the advantages and disadvantages of each tested model. A conclusion will then be made about which models performed best and in which scenarios certain models should be used over others.

1.2.2 Objectives

This project can be broken down into a list of objectives that not only provide a strong path to follow to complete the project but also a way to evaluate the project's success post-completion. I have placed these objectives into phases corresponding to the project's work plan.

1. Phase 1: Research
   (a) Conduct a literature review to understand the current landscape of inflation forecasting.
   (b) Research prominent models in the literature.
2. Phase 2: Source Data
   (a) Source the data to be used in the models.
   (b) Evaluate the usefulness and appropriateness of the data.
   (c) Clean the data: dropping and retaining specific variables.
3. Phase 3: Creating Models
   (a) Choose at least 3 appropriate models to use.
   (b) Develop and tune the models.
   (c) Train the models on the dataset.
4. Phase 4: Evaluation and Report
   (a) Evaluate the models using statistical tests.
   (b) Conclude the findings.
   (c) Present the findings in this final report.

Chapter 2

Literature Review

2.1 Motivation

Embarking on a literature review before developing our project offers numerous benefits. Understanding existing knowledge in Machine Learning, specifically when used to predict financial indicators, helps to contextualise our research, positioning it within the existing field. Reviewing previous literature also provides the benefits of identifying gaps in current research and finding supporting arguments that can guide our work and help us to avoid, as much as possible, redundancy in our and others' works. Having completed the literature review, we should have a strong foundation to start and guide our project.

2.2 Available Literature and Context

There is certainly a strong monetary incentive to produce research on how best to predict financial indicators. The correct predictions can not only allow organizations and individuals to profit greatly but also to avoid loss. This results in a myriad of papers being written, experimenting with a variety of techniques to predict future values, most of which we can learn from to help structure our models.

This report's topic focuses on the prediction of inflation through the use of machine learning. To accomplish this we can view papers predominantly addressing two types of topics. The first type is papers that focus on the topic of predicting inflation or other economic indicators and time series. The second type of papers we can research are ones that deal with different machine learning techniques. Additionally, it is pertinent to survey the current literature on inflation: its causing factors, effects, and significance.

2.2.1 Financial Indicator Prediction Papers

According to "Analysis of Financial Time Series" by Ruey S. Tsay "Financial time series analysis is concerned with the theory and practice of asset valuation over time."[44] There are many financial time series (FTS for short) prediction methods both theoretical and practical that have attracted attention, the taxonomy of which is shown in figure 2.1. The predominant analysis strategies for predicting financial market behaviour are fundamental analysis and technical analysis[22].

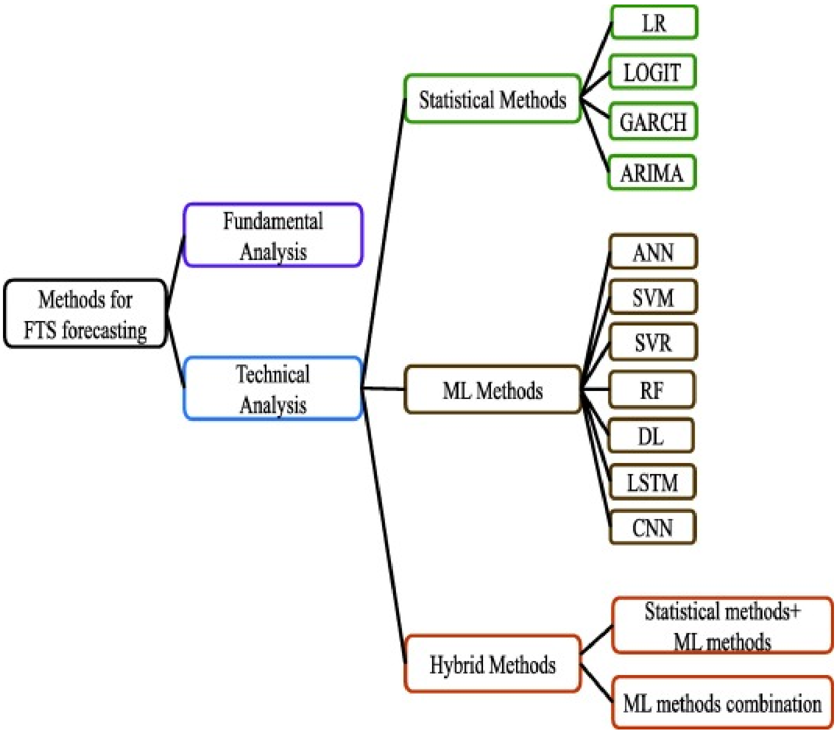


Figure 2.1: FTS Forecasting Methods.
Figure from page 3 of 'A survey on machine learning models for financial time series forecasting by Yajiao Tang et al.'[42]

Fundamental Analysis

Fundamental analysis[43] attempts to measure the intrinsic value of an asset by looking at current market and economic conditions. Additionally, fundamental analysis frequently makes use of techniques - such as sentiment analysis - that often deal with unstructured data. The success of fundamental analysis often relies on the financial efficiency of the target[45], which according to Tom Seegmiller is defined as "... how successful your organization is at turning expenses into revenue"[40].

Technical Analysis

Technical analysis[2] attempts to identify opportunities and predict investments by viewing movements and trends in market data alongside using a variety of technical indicators. Unlike fundamental analysis, technical analysis does not take into account many of the same fundamentals that can help indicate an asset's current value such as quarterly revenue. This is partially because it is often argued that technical indicators such as inflation or a stock's value are already priced according to the fundamentals that cause or contribute to them[30]. From this, we can come to the understanding that while fundamental analysis is the idea of looking at the current factors affecting an asset and using them to evaluate to asset's true value; Technical analysis is built upon the idea that past performance can predict future performance. Traditionally, technical analysis has relied heavily on statistical models to forecast the future performance of assets[37]. Furthermore, the application of using past values to predict future values has been widely implemented for years, with one of the earliest uses of autoregressive models being used to predict time series created by U.G.Yule in the 1920s[50]. However, with the increase of big data and the internet, ever-larger amounts of financial predictive data are continually being produced. Nowadays, simple statistical models may struggle to produce accurate future predictions when faced with big data sets containing complex characteristics[3].

2.2.2 Machine Learning Papers

This brings us to machine learning algorithms[29]. According to Mariette Awad et al. machine learning "is a branch of artificial intelligence that systematically applies algorithms to synthesize the underlying relationships among data and information"[4]. Currently, there is a massive abundance of fresh machine learning papers constantly being produced in the field. [14]

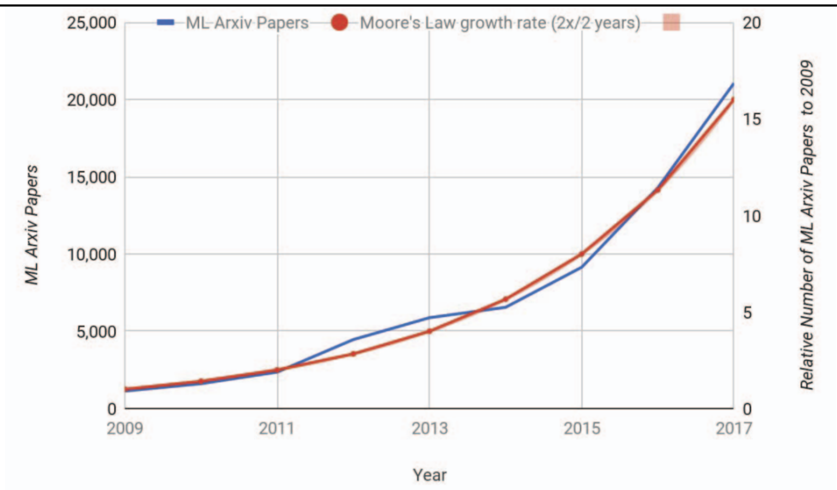


Figure 2.2: ML arXiv articles per year.
Figure from page 4 of 'A New Golden Age in Computer Architecture: Empowering the MachineLearning Revolution' by Jeff Dean, David Patterson, and Cliff Young[14]

As you can see in figure 2.2 articles on machine learning posted to arXiV (an archive for scholarly articles) have more than doubled every two years. Additionally, in 2018 the number of articles released reached 100 per day, summing to more than 33,000 by the end of the year. The number of articles released has steadily continued to increase in the years since[15]. Naturally, to read this many articles is impossible, however, the sheer quantity bodes well for this project as it means there will be plenty of guidance on how best to select and develop our predictive models.

2.2.3 Inflation Papers

According to Ceyda Oner at the International Monetary Fund "Inflation is the rate of increase in prices over a given period of time."[33] Often inflation is used to broadly indicate a countries of the global state of price fluctuation, however, inflation can also be used for certain services, goods, or food. Inflation affects everyone leading to an abundance of different papers linking to the topic. The papers focusing on inflation cover a variety of topics such as inflation's various effects, forecasting inflation, the causes and trends of inflation, and more. [35]

2.3 Problem Domain

This section of the literature review will go more in-depth into the area that this report will tackle as well as which methods could potentially be used.

2.3.1 UK Inflation

Inflation has been around since money has been used with one of the earliest recordings of inflation being caused by the death of Alexander the Great in around 300 BC[20]. Because of inflation's pervasiveness and its intrinsic link to both macro and microeconomics, there is a surplus of literature surrounding the subject. This report will focus on inflation in the UK as opposed to global inflation. Several indicators can be used to measure inflation, the most common of which are Consumer Price Index (CPI), Consumer Price Index with Housing (CPIH), and Retail Price Index (RPI). There are also more novel measurements of inflation as well such as viewing the Big Mac index[16] and recording its change over time.

In the UK, inflation is the responsibility of the Bank of England who set monetary policy eight times a year with the goal of controlling and stablising inflation. The Bank of England (BoE) calculates inflation using their own in-house models[12]. The BoE is not transparent with the specifics of the models or exactly how they calculate inflation, however, they state their models do factor in market expectations. In October 2022 inflation rates reached a peak of 11.1% the highest in over 40 years. Yet by October 2023, the annual rate was the lowest since October 2021 at 4.7%[6]. Granted this is still higher than the targeted 2% imposed on the BoE by the government. According to a research briefing published by the House of Commons Library, the main causes for the extremely high current inflation were: Covid-19 lockdowns causing supply chain disruptions and Russia's invasion of Ukraine causing increased energy prices due to the UK's prior reliance on Russian fuel[17]. Additionally, several pundits have suggested the BoE's slow reaction to combating high inflation rates and their increased money-printing during the pandemic[31]. The main way that the BoE combats inflation is through the manipulation of interest rates. The general premise is that as inflation rates grow, the growth of inflation should slow and inflation should eventually decrease. This is because increased interest rates mean the overall spending in the economy lessens as money is instead being spent on the increased interest.

Through this project we hope to be able to forecast future inflation through the use of machine learning models.

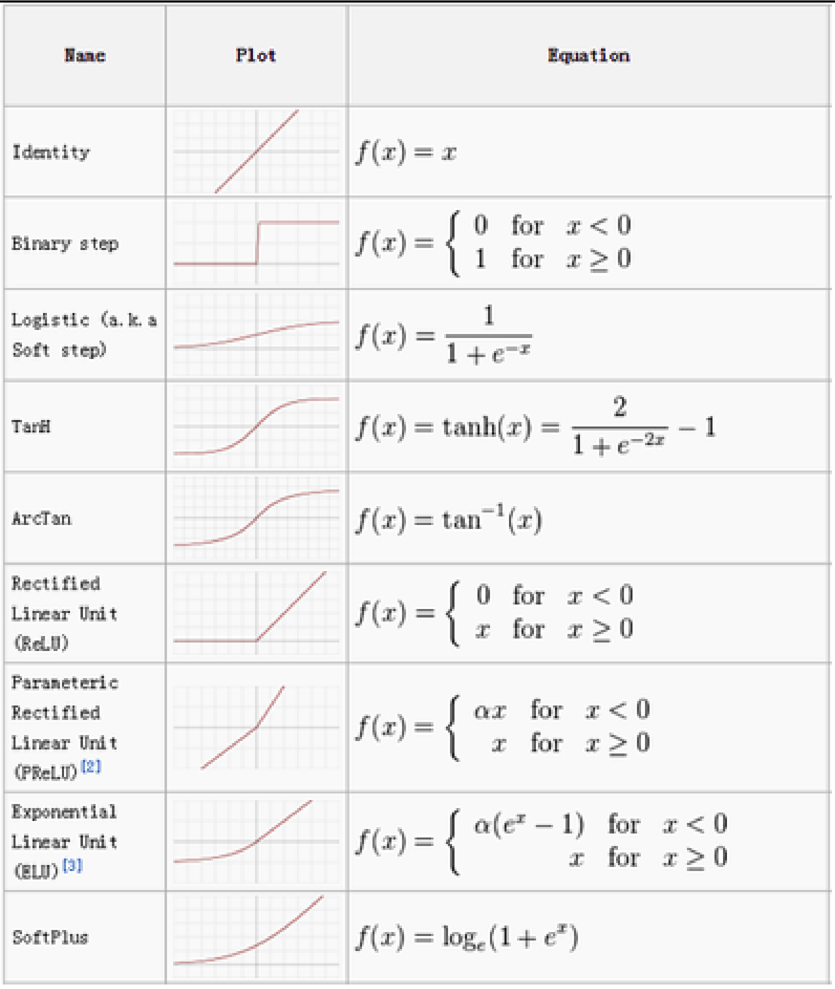
2.3.2 Machine Learning Models

By utilising machine learning techniques in financial forecasting we can endeavor to improve upon the performance of traditional statistical models. Generally, the goal of FTS forecasting can be placed into two main categories: 1. Price prediction 2. Price movement prediction (this includes volatility predictions) These two goals also reflect two types of machine-learning problems: 1. Regression 2. Classification. This paper will focus mainly on the price prediction/regression categories regarding inflation. This means that we will aim to predict future values of inflation as opposed to whether inflation will increase or decrease. Naturally, the areas we study will be with this regression problem in mind. We shall now cover some of the potential ML forecasting models that can aid us in predicting inflation.

Artificial Neural Networks

An artificial neural network is a machine learning model that is made up of an interconnected group of nodes (also known as neurons) organised into layers. The model's inspiration stems from how neurons in the human brain interact with one another. In 1957 Frank Rosenblatt invented the perceptron, one of if not the first implementations of an artificial neural network [38].

ANNs are composed of 3 types of layers: the input, hidden, and output layers. The input layer receives input data and passes it through to the first hidden layer. Hidden layers receive weighted inputs, perform an activation function on said input, and then pass the new data to the next layer. The output layer receives data from the final hidden layer and then produces the resulting prediction. The neurons within an ANN can either be excited or inhibited. Neurons are connected between layers and the strength of these connections (the weight) is decided by how excited or inhibited a neuron is. Each neuron in the hidden and output layers contains biases and activation functions (with the activation function in the output layer typically being different from the one in the hidden layer). Activation values are passed from node to node through the connections in the network. When a neuron receives the activation value, it sums and modifies it based on the neuron's activation function and bias. The predicted results can then be compared to the true values and the weights and biases of the network are updated accordingly. Artificial neural networks can be modified with a variety of techniques to alter their accuracy. One of these alterations is changing the activation function of the neurons. Figure 2.3 shows some common activation functions. Other commonly used activation functions include sigmoid, Guassian, and leaky ReLU.



Name	Plot	Equation
Identity		$f(x) = x$
Binary step		$f(x) = \begin{cases} 0 & \text{for} \quad x < 0 \\ 1 & \text{for} \quad x \geq 0 \end{cases}$
Logistic (a.k.a Soft step)		$f(x) = \frac{1}{1+e^{-x}}$
TanH		$f(x) = \tanh(x) = \frac{2}{1+e^{-2x}} - 1$
ArcTan		$f(x) = \tan^{-1}(x)$
Rectified Linear Unit (ReLU)		$f(x) = \begin{cases} 0 & \text{for} \quad x < 0 \\ x & \text{for} \quad x \geq 0 \end{cases}$
Parameteric Rectified Linear Unit (PReLU) [2]		$f(x) = \begin{cases} \alpha x & \text{for} \quad x < 0 \\ x & \text{for} \quad x \geq 0 \end{cases}$
Exponential Linear Unit (ELU) [3]		$f(x) = \begin{cases} \alpha(e^x - 1) & \text{for} \quad x < 0 \\ x & \text{for} \quad x \geq 0 \end{cases}$
SoftPlus		$f(x) = \log_e(1+e^x)$

Figure 2.3: Common Activation Functions [11]

ANNs have many advantages:

- Thanks to the many interconnected neurons, ANNs have strong learning capabilities[21].
- ANNs do not have a fixed structural equation making them very adaptable.
- ANNs can be finely tuned with many changes to the network in order to find the best-fitting model. For example, changing the number of hidden layers, the number of nodes in a layer, the activation functions, and so on.

ANNs also come with some disadvantages:

- The higher complexity of ANNs means that they require more resources than traditional statistical models.
- Due to the nature of the hidden layers, ANNs can be hard to interpret.
- ANNs can be overfitted[41].

Support Vector Regression

Created by Vladimir Vapnik and Alexey Chervonenkis in the 1960s and later built upon further by Vapnik et al. with the addition of the kernel trick and soft margin [13], Support Vector Regression and Support Vector Machines are supervised learning methods used for regression and classification respectively. Both models use non-linear mapping to transform the dimension of the input data. Then utilise a hyperplane in order to either best fit or categorize the data. The hyperplane for these models is found by using an ε-insensitive tube, meaning that any errors within the range of the tube are ignored. This is unlike a standard line of best fit that takes into account the ε (distance) of all points to the line, instead, only errors outside of the tube are considered pertinent. The hyperplane is then placed in a way in which the sum of all points outside of the tube is minimised. The points outside of the tube are called support vectors hence the name support vector regressions.

Advantages of Support Vector Regression:

- SVRs are simple and easy to implement as well as producing easily interpretable results.
- SVRs require less computational resources than other models.
- SVRs can maintain stability despite noisy input data thanks to the ε-incentive tube[47].

Disadvantages of Support Vector Regression:

- Deciding a suitable kernel function can cause difficulty [9].
- SVRs may struggle with big data[39].

Random Forest

The first random forest (RF) algorithm was created by Tin Kam Ho in 1995[24] which was later developed upon by Leo Breiman[8]. The random forest model makes predictions by consulting multiple decision trees. Each tree is trained on a random subset of data taken from the training set, this is called bagging or bootstrap aggregation. The final prediction is an average taken from all of the trees' predictions.

Advantages of random forest:

- Random forest can prevent overfitting by combining the results of several weak learners instead of using one powerful learner[7].

- Rf models have good prediction accuracy as the result is an average making it unlikely to be an outlier.
- Rfs are stable as changes to the data set may affect one tree but are unlikely to affect many trees.

Disadvantages of random forest:

- Rfs suffer from increased training time. This is due to the fact that to make a prediction you need predictions from all of the trees to get an average.

2.3.3 Regression Metrics and Model Evaluation

Evaluating a model is extremely important not only to know the quality of your model's predictions but also in order to make improvements to the model to achieve a more desirable result. Evaluation metrics can be used to evaluate the performance of the model. Some useful metrics that can be applied to measure a model's performance are: Mean absolute error (MAE), Mean squared error (MSE), Mean absolute percentage error (MAPE), Root mean absolute error (RMAE), Normalised mean square error (NMSE), Root mean squared error (RMSE), Relative root mean squared error (RRMSE), Correlation coefficient of prediction (R) The most commonly used metrics are MSE, MAE, MAPE, and R.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Y_i - X_i|$$

n is the number of data points.

Y_i is the ith true value.

X_i is the ith predicted value.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i)^2$$

n is the number of data points.

Y_i is the ith true value.

X_i is the ith predicted value.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_i - X_i}{Y_i}\right|$$

n is the number of data points

Y_i is the true value.

X_i is the predicted value.

$$R = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

X_i and Y_i are the data points.

$\bar{X}$ is the mean of the x-value and $\bar{Y}$ the mean of the y values.

FOR MSE, MAE, and MAPE a lower result indicates a more accurate model, and when a model has no error the value will be zero. R is always between -1 and 1, when R=0 it indicates that there is no linear relationship between the values. If R is -1 then there is a perfect negative linear relationship and if R is 1 then there is a perfect positive linear relationship. These formulas can be used to understand the predictive skills of a model.

2.4 Summary and Conclusion

Through this literature review, we have covered a sample of the literature on machine learning methods, financial time series forecasting, and inflation. There is a multitude of literature available in these areas due to the potential monetary gain as well as the new and emerging technologies being explored in the fields. Therefore, it would be nearly impossible to extensively cover all of the relevant articles to this project. Instead, the main focus was on understanding the most common techniques used to predict FTS and which machine learning models are often applied. We also covered the UK's relationship with inflation, who is tasked with controlling it, and how they attempt to do so as well as some of the factors contributing to it.

Having completed the literature review we are now equipped with general knowledge of the problem of inflation, its causes, and how it is controlled. This will help us to select appropriate data to train and build our models on. Furthermore, due to reviewing current literature in the field of FTS forecasting and machine learning, we have the knowledge of which available models may be best suited to forecasting inflation as well as an understanding of how to evaluate the models we choose to use.

Chapter 3

Design

The design process for machine learning algorithms varies depending on the size, complexity, and use cases of the algorithm. According to Chip Huyen, author of the book Designing Machine Learning Systems, the machine learning project flow has four distinct steps. Project setup, data pipeline, modeling and training, and finally: serving. These four steps can be iterated through as seen in figure 3.1.

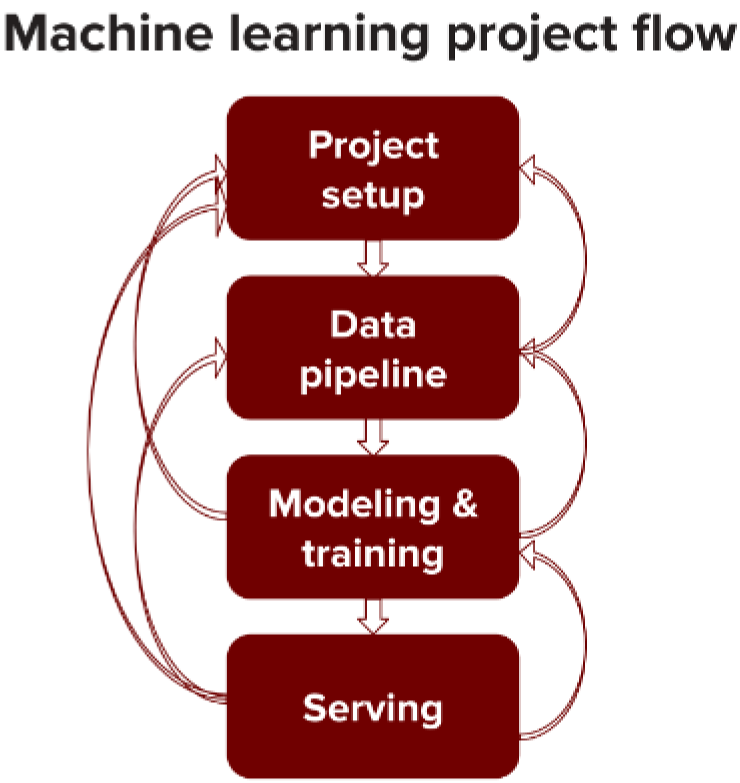


Figure 3.1: Machine Learning Design Pipeline by Chip Huyen.
Figure from https://huyenchip.com/machine-learning-systems-design/design-a-machine-learning-system.html[26]

This is the project flow that will be followed throughout this design chapter.

3.1 Project Setup

The project setup step of Heung's design flow mainly focuses on outlining the details of the project. Examples of these details are the projects: goals, user experience, evaluation, personalization, and constraints. This report's project setup shall focus mainly on the goals, evaluation, and constraints sections as user experience and personalization are not key aspects of the project.

3.1.1 Goals

As the title of the project outlines, the goal of the project is to predict inflation using machine learning. This project will evaluate the use of various machine learning methods and assess their ability to predict inflation. Further aims and objectives are outlined in Chapter 1. Introduction.

3.1.2 Evaluation

Evaluation outlines how the system's performance shall be evaluated: what metrics shall be used, what visualisations shall be produced, and so on. The models created in this project will be regression models attempting to predict an exact value of inflation. Thus the methods selected to evaluate the models should be suited for regression and continuous values. So methods like confusion matrixes that are better suited for classification would not be a good choice. As stated in the literature review, the most common regression metrics were MSE, MAE, MAPE, and R/ R^2. As each of these metrics are straightforward to implement, they will all be used for evaluating the models created. Furthermore, visualisations displaying each model's predictions alongside the values they were attempting to predict will be produced. This will aid the view in understanding the model's effectiveness beyond numerical metrics. Where appropriate, certain models should produce loss functions as these will be able to guide hyperparameter optimisation, such as selecting the correct number of epochs to train a model.

3.1.3 Constraints

The Constraints of a project can follow two veins: performance constraints and project constraints. Performance constraints address how a project must produce its results. Examples of performance constraints are how fast predictions should be made and how precise predictions should be. Of these two constraints, this project's main concern is with precision. The predictions of each model created will be evaluated and compared to one another in order to conclude the type of machine learning model best suited to predicting inflation.

On the other hand, project constraints are real-world constraints that may limit the project such as the time or manpower available. A typical project constraint may be the software available. However, since it is common practice to use Python for artificial intelligence research (even for billion-dollar companies such as Google[18]). Additionally, Python is a free-to-use, open-source language. The constraints of this project are mainly generic constraints that will apply to most final-year projects. One such generic constraint is the fact that only one student will be working on the project and the student has several other modules to focus on throughout the year. This means that a limited amount of time and resources will be dedicated to the project. In the context of predicting inflation with machine learning, this means that it is extremely unlikely that any of the models produced by the end of the project will be more capable or effective than current models in use from large organisations such as the Bank of England. As these organisations will have dedicated more time, manpower, and resources toward producing models with greater sophistication and accuracy than those that could be created by a lone student. Even if models of similar complexity were to be created from this project, there would not be the hardware available to consistently and reasonably train and test such models. These factors mean that given the task of predicting inflation with machine learning, it is more suitable to create numerous models and compare their effectiveness rather than creating a single extremely sophisticated model that would require more resources to produce.

3.2 Data pipeline

The data pipeline section of designing machine learning algorithms deals with the selection and preprocessing of data.

3.2.1 The Systems Input and Output

One of the first questions that needs to be addressed is: "What is the system producing and from what information are these predictions being produced?" The input is real-world historical economic data relating to UK inflation. The output is an array of values predicting future inflation. As all of this data is publicly available the project should have very few privacy and bias concerns.

3.2.2 Data Selection

The saying "garbage in garbage out" succinctly illustrates the importance of selecting good data for machine learning models. It does not matter how powerful a model is if the data selected is poor or inappropriate. Selecting data is a key step to building effective machine-learning models. Thankfully, there are plenty of large open-source data sets available online, despite how time-consuming and expensive gathering data may be. Yet picking an appropriate set may still present a challenge. The criteria for a dataset for this project are as follows:

- A large number of data points (preferably in the thousands) as inflation is a complex feature to predict.
- Several related economic indicators to use as features for predicting inflation and to test their ability to predict inflation.
- Minimal missing or erroneous data as this data can lower a model's accuracy.
- Reliable data - the data should be taken from a reliable source in order to accurately predict real inflation.

During the project, several datasets were tested, for example, various Kaggle datasets and the "World Development Indicators" by the World Bank [5]. The dataset that was finally settled on was the UK financial statistics from OECD (The Organisation for Economic Co-operation and Development)[32]. The reason for selecting this set is due to its large variety of indicators, the reliable provider, and the fact that it is open-source (the dataset is classified as public under the access to information classification policy).

Initially, datasets from Kaggle were tested but they lacked a sufficient number of data points for multivariate machine learning analysis of inflation. Data from the World Bank was also preprocessed and tested with basic models. The World Bank is known for reliable and accurate data which should avoid issues of bias, or poor quality. However, the World Bank sourced a majority of its indicators annually dating back to the 1960s meaning that there would only be around 60 rows of data up to the present day. Thus, although the set had a large number of interesting features, the number of data points was insufficient for a machine-learning model. Additionally, due to the large number of features (over 1400), it would need a sufficient amount of cleaning/preprocessing in order to remove irrelevant features.

3.2.3 The Final Dataset

Due to the issues encountered with the other datasets, the OECD dataset was selected for use alongside monthly CPI data sourced from the World Bank. The first main benefit of the OECD dataset is that the data is taken monthly so there are a lot more data points than datasets that collect data annually. The features from the OECD dataset were taken from 1972 to 2022. Unfortunately, a few of the features had less data as they were taken from later than 1972 with the latest being taken in 1987. However, as this is only a small fraction of the overall dataset, replacing the missing values with zeros is unlikely to have a profound effect on the final results. The second benefit is that the dataset contains a moderate number of features to choose from (not too many that the features need to be drastically reduced (as was the case with the World Bank dataset)). There were twenty different features in total (not including CPI), these features were: composite business confidence, composite consumer confidence, composite leading indicator (CLI), consumer prices, hourly earnings, immediate interest rates call money interbank rate, long-term interest rates, M1, M3, merchandise exports, merchandise imports, nominal exchange rates, passenger car registrations, production volume, GDP, retail trade volume, share prices, short-term interest rates, and unemployment Histograms were plotted for each feature in order to assess their distributions. Understanding the distribution of a feature can give some insight as to whether to use normalisation (rescaling the data between two set values) or standardisation (setting the mean of the data to 0 and the standard deviation to 1) before giving the data to the machine learning models. In addition, a correlation matrix was formed for all of the features to determine if any features were overlapping. A correlation matrix gives us an indication of how the features relate to each other one on one.

Figure 3.2: A histogram of each feature in the dataset

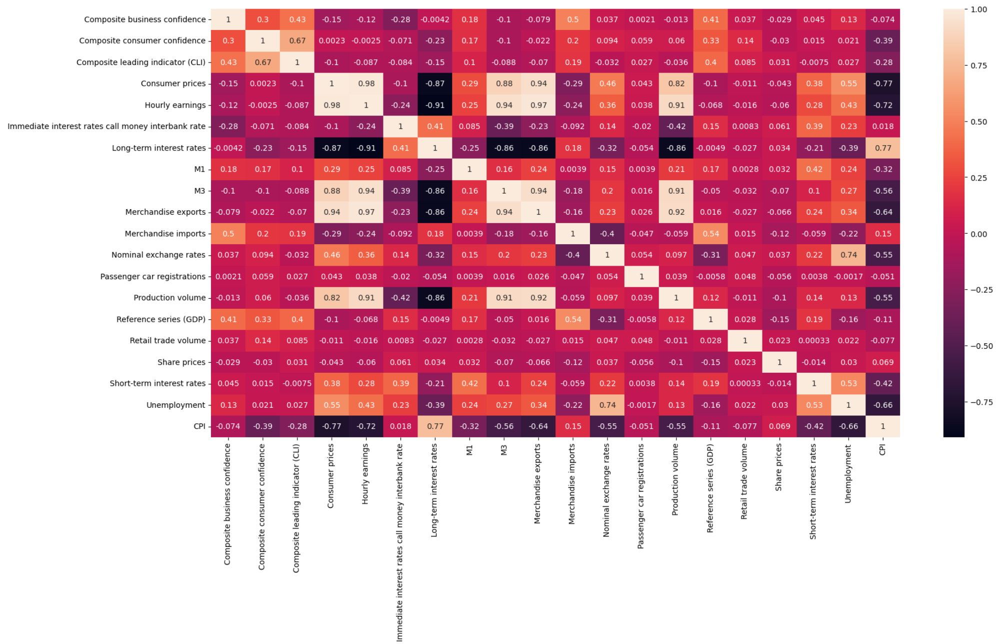


Figure 3.3: Correlation Matrix for the Features in the Dataset

The histograms display a variety of different distributions, this is understandable as many time series do not follow a normal distribution. The distributions indicate that normalisation may be preferable to standardisation as normalisation is preferable when several features do not have a Gaussian (normal) distribution. This is because normalisation will preserve the distribution of the data by rescaling the data where as standardisation makes the asssumption that the date has a Gaussian distribution and may modify the data's distribution if it is not already Gaussian. As the distribution of the dataset will remain non-normal, some precautions may need to be observed when using models that assume a normal distribution (such as linear regression).

3.2.4 Data Pre-processing and Transformation

Often, datasets have several outstanding issues or properties that make them imperfect for use in a machine-learning model. Data pre-processing often includes steps such as data cleaning (handling outliers, noise, or missing values) and data integration (combining data from multiple sources). Data transformation is the task of molding the data to an appropriate size and dimensionality. Both of these processes are to create a dataset that is formatted in a manner that suits our models and objectives. Poorly processed data can be difficult for both machines and humans to use and can lead to poor results.

Issues With The OECD Dataset

There were some outstanding issues with the original dataset taken from OECD and thus several steps were taken to clean the data. These issues included:

- Features containing missing data.
- Incorrect format of the dataset (required transformation).
- Duplicate features.
- Unnecessary/erroneous data.
- A large amount of metadata that needed to be removed.

Any data points that fell under these issues were removed/cleaned and all null values were replaced with zeros to improve readability.

Finally, a Granger causality test was carried out on the data to gather a better understanding of our data and its correlation to the target feature (CPI).

Granger Causality

Granger causality is a statistical concept in economics used to show if time series A is useful at forecasting time series B. Clive Granger originally proposed the test in 1969 in the article "Investigating Causal Relations by Econometric Models and Cross-spectral Methods[19]. The test only shows predictive causality and not true causality. Additionally, the test only provides information about forecasting ability and not the actual causal relationship. Another issue with the use of the Granger test in the context of inflation is that the test works best on stationary data. Whether inflation is best treated as stationary or non-stationary data is currently inconclusive[48]. But for our purposes, these limitations are fine as we only want to understand an indication of how useful the data will be for inflation forecasting.

Granger Testing The Data

With two indicators X and Y, X causes Y if a series of tests on lagged values of X produce a p-value of less than 0.05. The closer the p-value is to zero the more likely it is for X to granger cause Y. Lagged values are values from a time series shifted forwards or backward in time. In the case of the Granger test, Jeffery Woolriddge proposes that fewer lags should be used for annual data compared to quarterly or monthly data in order to not lose degrees of freedom[49]. The Granger test was run on the preprocessed indicators comparing each of them to CPI. The function ran 12 lags to see if data has a potential causal relationship with inflation within the past year. This means that if at some point during the last 12 lags a p-value¡0.05 was produced the data will be labeled as having the potential to be useful in forecasting CPI. Of the 20 features tested, 9 produced results that indicate Granger Causality. These 9 were: 'Composite business confidence', 'Composite consumer confidence', 'Consumer prices', 'Long-term interest rates', 'Merchandise imports', 'Reference series (GDP)', 'Retail trade volume', 'Share prices', 'Unemployment' This indicates that almost half of the features have the potential to predict CPI. This, however, does not mean that the 9 features found to Granger Cause CPI cause inflation as the Granger Causality test is only a measure of the potential of one value to cause another. Furthermore, just because a value does not Granger Cause inflation does not mean that the feature is not useful in predicting inflation. We now have a better understanding of the potential our dataset has to predict inflation.

3.3 Modelling and Training

3.3.1 Model Selection

The models that will be used to predict inflation are a Random forest model, a support vector regression model, a linear regression model, a long-short-term memory model, and a standard dense neural network model These models were selected through the literature review due to their common use in similar regression problems. Several of these models can be implemented through the use of libraries such as Scikit-learn[36]. Implementing a model through a high-level library lets the user create an already well-documented and tested model with only a few lines of code. This saves time while also providing additional support, if necessary, from the community and documentation.

3.3.2 Model Training

The models will be created using Python and several of its supported libraries and written in a Jupyter Notebook. The reason for using a Jupyter Notebook is that the cell structure allows for the steps to be run separately and the results to be displayed in real time without the entire file being run. This makes experimenting with hyperparameters and training models much more convenient. As the results of each cell are displayed separately this also improves the ability to quickly create and display visualisations through tools such as Matplotlib [25]. Jupyter also supports multiple languages and libraries enabling users to work in the preferred environment.

3.4 Serving

Serving is the process in which the models and their results are presented to the end user. As this is an academic project this report will act as the presentation method. This means that both the model and their results need to be suitably presented within a report. This will be done through the use of graph visualisations created in Matplotlib. These visualisations should display and compare the model's predictive ability along with their learning rates. This will be done by creating a graph for each model that displays their predictions compared to the actual values of inflation, indicating how well they can predict both training and testing data. Loss graphs will be used to show each model's learning rate over time.

Chapter 4

Implementation and Testing

The goal of the project is to implement a machine-learning model that has the ability to predict inflation. In order to find the most suitable model, several different models were implemented and their results were compared. This chapter of the report will cover how the models were implemented, which models were implemented, how the models were tuned and tested, as well as some of the issues encountered in the implementation process. This chapter will not cover the results of the models, their comparisons, or their evaluations.

4.1 Tools

4.1.1 Coding Environemnt

The IDE used for this project was Jupyter Notebook provided through the use of Anaconda. Anaconda is a platform that includes a packet manager, this makes it easy to control packages and their dependencies that are installed in an environment. Each environment in Anaconda is isolated which allows version control and preventing conflicts between the installed versions and packages of different projects. This is valuable when working on multiple projects simultaneously. Furthermore, Anaconda provides Jupyter Notebook integration which is beneficial as it is a common IDE for data science and has an active community making it easy to learn.

4.1.2 Visualisation

Jupyter Notebook provides visualisation as output from individual cells. This was useful for debugging throughout the project. Both Matplotlib and Seaborn were used for further graphical visualisation.

4.1.3 libraries

The project's code was written in Python and several libraries were used to assist the coding process.

NumPy[23] and Pandas[34] - These libraries were used for data manipulation and cleaning. NumPy provides several tools for dealing with multidimensional arrays. Pandas is built upon NumPy and is used for dealing with tabular data (data that is organised into a table with rows and columns). Pandas also has built-in tools for dealing with time-series data which is useful as inflation is time-series data.

TensorFlow[1] and Keras[10] - TensorFlow is a platform developed by Google for training machine learning models. It makes defining a model straightforward while still providing enough control and flexibility of the model's structure, inputs, and outputs. Using pre-written functions and nodes that were written by a reliable source both saves time and resources during development. Keras is built on top of TensorFlow and aims to provide a simplified and more user-friendly interface for creating and training machine learning models.

SciKit-Learn[36] - Sklearn is a machine learning library that provides a wide range of easy-to-implement machine learning models and metrics. Its extensive documentation makes it easy to learn and implement.

Matplotlib[25] and Seaborn[46] - Matplotlib and Seaborn were used for all graphical data visualisation. These libraries provide a plethora of graphing options along with plenty of community support.

Sktime[28] - Sktime is a library for machine learning with time series data. This was used to implement the univariate time series forecasting of inflation.

4.2 Data Preperation

Features relating to inflation were sourced from the OECD Databank and Monthly CPI indicators were taken from the World Bank's online statistics. All of this data is relating to the United Kingdom's general economic indicators. The data was compiled into a tabular format to facilitate the use of the Pandas library and to streamline further data manipulation tasks. By using a pandas dataframe several functions can be used to quickly assess the state of the data such as the head or shape functions. This is useful in the context of machine learning as model layers require inputs of certain types or shapes.

4.2.1 Normalising the Data

Normalisation, also known as feature scaling, is used to rescale data between a given range. The importance of normalisation is due to the fact that many of the features may have varying ranges, this could lead to features with a large range having a greater impact on the result. By adjusting all of the features to range between the same values, no single feature will dominate and all features should contribute equally to the final result. Additionally, normalisation increases the speed at which gradient descent converges, which will decrease the runtime of the models.[27] Normalising the data is often done either in the range [0,1] or [-1,1]. The dataset for this project was normalised between the range of [0,1] with the use of the MinMaxScaler from the sklearn library. There are several benefits to scaling to a small range such as reducing the impact of outliers and increasing the speed of the algorithms.

4.2.2 Splitting the Data into Training and Testing Splits

It is common practice to split a dataset into training and testing splits. The training split is then used for training the machine learning model and the testing split is used for testing the model's performance on unseen data. If the model was tested on data it had already been trained on it would become difficult to evaluate the model's generalization (its ability to perform on new/unseen data). The difficulty in deciding the size of the training and testing splits lies in the fact that too small a training split and our model may not perform well but too small a testing split and our performance metrics will have a greater variance. The project dataset was split into 80% training data and 20% testing data following the advice from the article: A scaling law for the validation-set training-set size ratio by Isabelle Guyon[**?**].

As the dataset is time series data it is likely to contain some temporal dependency. Temporal dependency is when data contains stronger associations between events that happened within the same time period. To preserve the temporal dependency of the data The dataset was then split using the "train_test_split" function from sklearn.

4.3 The Machine Learning Models

Five different models were implemented and tested for the predictive ability on the same data set. The different machine learning models were linear regression, random forest, support vector regression, long short-term memory network, and a custom artificial neural network. The reason for these five models was because they are some of the most common models used for regression analysis and, as found in the literature, many of the models are used for economic indicator analysis (such as stock predictions). Three of the five models were implemented with the Scikit-learn library. These were linear regression, random forest, and support vector regression. The remaining two models were implemented through the use of TensorFlow and Keras.

4.3.1 Univariate Implementation

Before Implementing each model with the full dataset that contained twenty different features, some of the simple models were implemented. The models that were implemented were linear regression, random forest with 20 trees (as this produced the best results), and support vector regression. This was a quick implementation to test if predicting inflation solely based on its past values would suffice. As expected, the results of this experiment were lackluster likely due to inflation's complexity and the fact that many outside factors affect it. Of the three models, random forest regression produced the best results although the quality of the predictions was still not satisfactory.

4.3.2 Multivariate Analysis

On the topic of outside factors affecting inflation, the curated dataset that was used for this project contained 20 different features, each with monthly data dating back to 1972 at the earliest and 1987. The OECD features are macroeconomic indicators that focus on broad trends from the UK with global implications. Macroeconomic indicators are better predictors of inflation as inflation happens on a macro scale.

Chapter 5

Results and Evaluation

Chapter 6

Conclusion

Chapter 7

References