A dissertation submitted to the **University of Greenwich**
in partial fulfilment of the requirements for the Degree of

Master of Science
in
Data Science

A system for house price prediction in the UK

Name: Suganthan Balasuriyan

Student ID: 001232197

Supervisor: Mohammad Al-Antary
Submission Date: December, 2023
Word count: 10,120

A SYSTEM FOR HOUSE PRICE PREDICTION IN THE UNITED KINGDOM

XXXXX

Computing & Mathematical Sciences, University of Greenwich, 30 Park Row, Greenwich, UK.

(Submitted 24 December 2023)

ABSTRACT.

The assessment of economic stability heavily relies on house prices, and accurately forecasting their variations is of utmost importance for participants in the housing market. This thesis investigates sophisticated ensemble machine learning algorithms for forecasting house prices, with the goal of providing accurate predictions to empower decision-makers and improve market transparency. The study centres on the role of qualities in the process of prediction. The thesis highlights a gap in the existing literature, specifically examining the progression of machine learning models in predicting property prices. Current ensemble machine learning regressors provide superior performance, but specific models such as the 'Extra Tree regressor' and ‘ConvLSTM’ have not been thoroughly investigated for this application. The objective of the study is to fill this gap and provide an improved system for predicting housing prices in the UK. The study utilises a range of models, such as conventional regression and sophisticated ensemble models, to forecast house prices in the UK. The Extra Tree regressor demonstrates superior performance compared to other models, with a mean MAPE (Mean Absolute Percentage Error) of 0.81 and an R-squared value approaching 1. The thesis also highlights the marginal contribution of macroeconomic variables, such as the retail price index, CPI value, and household savings ratio, on housing price forecasts, using explainableAI technique, SHAP values. In summary, the suggested approach improves precision, hence facilitating better decision-making for stakeholders in the real estate market.

Keywords: House price prediction, machine learning, predictive modelling, UK houses, economics, explainableAI

Acknowledgements

I would especially like to thank my thesis supervisor, Mohammad Al-Antary for agreeing to be my supervisor and for his consistent advice, feedback, guidance and support throughout the lifecycle of this MSc project.

Many thanks go to my friends and family for their support.

Table of Contents

List of Figures

List of Tables

1 INTRODUCTION

House prices are a major gauge of the financial stability of the economy, and the housing market is essential to it. Comprehending the variables that impact home values and effectively forecasting their fluctuations is crucial for multiple stakeholders, such as prospective purchasers, brokers, financiers, investors, legislators, and real estate firms. A strong home price forecasting system can offer insightful information, facilitate decision-making, and promote a more transparent and effective housing market. Predictive modelling of housing prices has become a vital undertaking, driven by progress in machine learning, statistical methods, and the abundance of extensive datasets. Precise forecasts empower stakeholders to make well-informed choices, minimise risks, and optimise investment plans. This thesis explores the intricacies of predicting housing prices with advanced and latest ensemble machine learning algorithms. Further, explaining and interpreting the contribution of attributes will support decision-makers, stakeholders and economists.

1.1 Background

As people's living standards improve, there is a growing need for housing. House sales in the United States have had a 34% growth over the past decade, culminating in a record-breaking 5.51 million sales last year. In Australia, there has been a 36% surge in house sales since 2013. House price prediction has garnered significant attention due to its potential to assist real estate stakeholders in making well-informed decisions. Buyers utilise house price prediction to choose potential houses that align with their financial capacities. Likewise, homeowners would require it to continuously watch the market and actively pursue the most advantageous chance for selling their homes. In addition, real estate sales agents utilise home price prediction to assist customers in identifying market trends. The precision of these predictions has emerged as a crucial factor in evaluating the trustworthiness of house sales agents.

Machine learning has been applied in various fields, including business, computer engineering, industrial engineering, bioinformatics, medicine, pharmaceuticals, physics, and statistics, to acquire knowledge and make predictions about future events. Machine learning is an advanced method that may be employed to detect, understand, and analyse very complex data structures and patterns (Ngiam & Khor,2019). Consequential learning is facilitated, and model predictions are enhanced through the systematic incorporation of more recent data. Modern studies in machine learning are a branch of artificial intelligence (AI) that aims to teach computers new information by inputting data, such as words, images, and numerical values, and facilitating its communication with other computer networks. Given the current expansion in the real estate industry, the utilisation of machine learning can be crucial in accurately forecasting property prices. Nevertheless, there has been limited study conducted on utilising machine learning algorithms to analyse and predict the selling price of real estate properties. In the realm of real estate, real estate agents, buyers, and sellers all play crucial roles. Homeowners have the option to engage the services of a real estate agent when they want to sell their townhouse. Given this environment, the process of machine learning diverges substantially from conventional econometric approaches employed in social and economic assessments. In the subject of econometrics, economic models are created to explicitly analyse different features that indicate the effects of changes in demand and/or supply factors. Moreover, comprehending the spatial and temporal fluctuations in property prices can aid in implementing finance mechanisms like land value capture. The government can utilise these tools to expedite the construction of innovative urban infrastructure, such as new metro lines.

The Housing Price Index (HPI) is a weighted index that measures the average price changes of properties via repeated sales or refinancing transactions. The data is acquired through an analysis of recurring mortgage transactions on individual residential properties, the mortgages of which have been bought or securitised by Fannie Mae or Freddie Mac from January 1975. The data science community has shown significant interest in precise projections of housing prices, driven by the increasing popularity of machine learning algorithms and their strong predictive capabilities (Irizarry, 2019). The growing prevalence of Big Data has led to the widespread adoption of machine learning as a crucial predicting technique in recent years. It can precisely predict property prices based on their characteristics without depending on historical data. Tree-based algorithms, such as random forest and XGBoost, consistently get better results than standard prediction approaches when dealing with complex data (Park & Kwon Bae, 2015). Nevertheless, due to the absence of complete factors influencing property values in certain research, the accuracy of forecast outcomes is often insufficient.

1.2 Motivation and Goals

The existing algorithms for predicting housing prices have demonstrated potential; however, they frequently encounter constraints and difficulties. The drawbacks encompass imprecision, insufficient incorporation of pertinent variables, and restricted user engagement functionalities. Our project aims to rectify these deficiencies by implementing substantial enhancements in the precision, comprehensiveness, and user interface of house price estimates. Average UK house prices increased by 4.1% from 12 months to March 2023, down from 5.8% in February 2023) (*UK House Price Index - Office for National Statistics*, n.d.). Due to the uncertain variations and fluctuations in house prices in the UK, a more robust and accurate model is required. Over the past 10 years, the U.K. housing market has experienced frequent fluctuations and changes as a result of economic shifts, market dynamics and the pandemic.

The goal of this project is to create a house price prediction system that utilises machine learning methods to anticipate housing prices in the UK. The algorithm will generate precise and timely projections of house values by examining several aspects, including geography, property attributes, economic indicators, and market movements.

1.3 Objectives

The main objectives of this study are listed below.

- To survey and develop feature generation on the attributes that contribute to the house prices and the changes and collect those data reliable resources.
- To survey literature and gain a sound understanding of the different approaches for the project, researching the similar work carries out. Understanding the methodologies adopted and how it has contributed to the work. This can help to plan the methodology that can be adopted for this project.
- To study the different machine learning models and approaches that would suit the project. There is an advanced ensemble of machine learning models available for prediction analysis. Understanding their differences and uses becomes crucial to acquiring the best-suited model for the selected project.
- To experiment on the machine learning models to increase the performance and accuracy of the prediction. This experiment will be carried out using the evaluation criteria that will be used to evaluate the model performance.

- Further studies and experiments are needed to improve the system's usability for the user.

1.4 Thesis structure

This thesis contains seven chapters, including the introduction. The chapters are described as follows:

- Chapter 2 briefly covers the related works on house price prediction and reviews the literature in a chronologically evolving manner
- Chapter 3 provides a detailed explanation of the data utilised in this study, starting from the collection of the data, statistical and exploratory data analysis, and, ultimately, preprocessing of the data to be used for machine learning model development.
- Chapter 4 presents our selected proposed machine learning models, both ensemble and deep learning models, with a small theoretical background. Further, the experimental setup of our processed data and the criteria to evaluate the performance of the proposed models and explainable AI are provided.
- Chapter 5 presents the results obtained from the ensemble model and the comparison with other models and their analysis. In addition, the insights obtained from the feature important analysis of our proposed model are stated there.
- Chapter 7 serves as the final section of the research, providing a summary and presenting potential avenues for further research.

2 LITERATURE REVIEW

Several studies have been conducted to create models that predict house prices and property values. Hedonic regression, which was initially created in the 1960s, has become the predominant method due to its ability to break down overall housing spending into its separate components. The house price prediction procedures utilise a comprehensive analysis of the correlation between house prices and other property characteristics. These approaches generate an anticipated house price by inputting the relevant house parameters. Based on the fundamental concept of whether it depends on the global model or not, the current approaches for predicting housing prices can be classified into two types (Ja'afar et al., 2021).

The global model employs several characteristics of a property to calculate its price and is commonly used on the entire dataset of dwellings. Substantial advancements have been achieved in this domain. In their study, (Nguyen, N. and Cripps, 2001) compared the use of multiple regression analysis (MRA) and artificial neural networks (ANN) by examining three different training sets of single-family houses with different sizes. In 2006, (Liu et al. 2006) proposed a prediction model that utilises a fuzzy neural network grounded in hedonic price theory. The objective of this model is to predict the most favourable pricing range for newly developed properties. The experimental findings showcased the robust function approximation capabilities of the fuzzy neural network prediction model, making it highly suitable for real estate price prediction. (Selim, 2009) conducted a study to compare the predicting capacities of hedonic regression and artificial neural network models. This study emphasised the effectiveness of artificial neural network models as a superior option for forecasting house values in Turkey.

(Gu et al., 2011) introduced a hybrid methodology called G-SVM, which combines a genetic algorithm and support vector machine for predicting housing prices. The results from instances in China demonstrated the method's predictive capability. (Wang et al., 2014) introduced an innovative technique utilising Support Vector Machines (SVM) to forecast the mean residential property value over various years. Temur et al. developed an innovative approach to forecast housing prices in Turkey and other countries. They combined an autoregressive integrated moving average model with an LSTM network. Furthermore, they employed mean absolute percentage error (MAPE) and mean squared error as evaluation metrics. Their hybrid model demonstrated superior performance compared to previous models, exhibiting a decreased error rate and precise predicted outcomes.

The study conducted by (Shukry et al., 2012) utilised a Neural Network to forecast Property Price Indices in Malaysia. The model incorporated indicators such as the unemployment rate, population size, interest rate, and family income. A training dataset consisting of quarterly data from 2000 to 2009 was extracted and subsequently evaluated using out-of-sample testing in 2010 and 2011. The Neural Network achieved a Mean Absolute Percentage Error (MAPE) of 8%, thereby surpassing the performance of traditional multiple regression.

(Kok et al., 2017) investigated the efficacy of different machine learning techniques in the context of real estate appraisals. A comprehensive analysis was performed on a dataset containing 84,305 observations collected from the states of California, Florida, and Texas, covering the period from 2011 to 2016. A comparison was made between different learning techniques, namely ordinary least squares regression (OLS), random forest (RF), gradient boosting regression (GBR), and extreme gradient boosting (XGBM) methodologies. The results demonstrated that, in general, XGBM proved to be the most efficient method. (Čeh et al., 2018) performed a comparative investigation to ascertain the superior forecasting technique between the RF algorithm and the hedonic price model. The researchers employed a dataset comprising 7407 homes in Ljubljana, Slovenia, covering the period from 2008 to 2013. The findings demonstrated that the RF model displayed greater predictive efficacy. (Fan et al., 2018) employed a range of prediction algorithms, such as RF, SVM (with multiple kernels), XGBM, ridge, and LASSO linear regression, to propose a method for forecasting home values in a competition organised by Kaggle.com. The data was obtained from Ames Housing in Iowa, comprising records from 2006 to 2010. The results showed that ridge, LASSO, and XGBM had a lower prediction error.

Current research has concentrated on forecasting housing prices based on local perspectives, which is increasingly being considered as a viable alternative and expansion of traditional methods for modelling house prices. The study conducted by (Hu et al., 2019) investigated the prediction precision of supervised learning systems in estimating rental expenses for properties in Shenzhen, China. The researchers utilised RF, ETR, GBR, SVR, MLP-NN, and k-NN algorithms. The results indicated that the RF and ETR algorithms had greater predictive ability. In a study conducted by Hong (2020), a comparison analysis was performed to evaluate the predictive effectiveness of the Hedonic Price Model (HPM) and machine learning approaches. Specifically, three algorithms (XGBM, LGBM, CatBoost) were employed to anticipate the transaction price of apartments in Seoul. To accomplish this goal, the author employed a dataset that covered the time period from 2009 to 2019. The findings indicated that machine learning algorithms exhibited greater predictive capability compared to ordinary least squares (OLS) regression. Furthermore, it was shown that the CatBoost algorithm exhibited superiority in forecasting pricing, even in the presence of outliers. Moreover, the composite model, comprising the three algorithms, demonstrated superior accuracy compared to the individual algorithms. The time-series technique entails examining the correlation between current and past rates. The second approach entails employing hedonic pricing and linear regression. In his investigation, the researcher employed the Random Forest algorithm as the second technique. In recent years, there has been extensive exploration of machine learning approaches for the purpose of price prediction.

Recently (Ekberg, 2022) conducted a comparative analysis of various machine learning methods in terms of their predictive capacity for house prices in Stockholm. The study exclusively considers house-specific attributes, such as location and living area. The researchers discovered that the random forest model surpasses the K-nearest neighbours and neural network models in terms of performance. In contrast, (Revend, 2020) discovered that XGBoost outperforms random forest in predicting property prices by utilising house-specific characteristics. Nevertheless, the training of such models results in a significant drawback; the house-specific characteristics are unable to account for variations in the macroeconomic climate over a period of time. However, these two studies were done in Sweden and for countryside houses. Therefore, a house price prediction system for the UK is needed since various factors influence the variations of house prices across countries, such as their own economic policies and laws.

In summary, the evolution of machine learning models for house price prediction across various countries of the world is presented in chronological order. Moreover, the positives and drawbacks of different types of machine learning models are also indicated. From that, recent ensemble machine learning regressor models show better performance. However, it is noted that deep learning sequence-based models also have competitive performance with the time-evolving dataset, where they can learn hidden patterns and memories across trends and seasons. As presented, some studies of ridge regression proved to show better performance with minimal data for this kind of prediction problem. Moreover, in the evolvement of ensemble ML models, the recent ‘Extra Tree regressor’ is not utilised for this problem. Therefore, considering the above facts, this study tries to fill the gap in literature and compare the selected best-performing models for another kind of prediction problems towards achieving a better system for house price prediction in the UK.

3 EXPLORATORY DATA ANALYSIS

This chapter provides a detailed explanation of the data utilised in this study, starting from the collection of the data, statistical and exploratory data analysis, and, ultimately, preprocessing of the data to be used for machine learning model development.

3.1 Data

The primary dataset is collected from the Office for National Statistics (ONS. The UK House Price Index (HPI) quantifies the fluctuation in the value of residential houses that have been sold in the United Kingdom. It offers a thorough and precise measure of property price patterns on a national, regional, and local scale. The Home Price Index (HPI) is computed by the Office for National Statistics (ONS) and Land Registry and released on a monthly basis. The data is available from the 1970s until the present (April 2022).

3.2 Time series data analysis

Time series data is a collection of observations or measurements that are recorded and gathered in a sequential manner over a period of time. The data points are usually arranged in chronological order and might be collected at regular intervals, such as hourly, daily, monthly, etc. Time series analysis examines the patterns, trends, and behaviours present in the data in order to create predictions or get insight into the underlying mechanism. In our problem, the House Price Index (HPI) is the time series data that has monthly variation. Therefore, we have analysed the seasonal decomposition of the data from the perspectives of trend, season, cycle and residual. They are shown in the Figure 1 given below.

It is clearly observed that there is an increasing trend in the house price. However, there is a sudden drop observed during the 2009 first quarter, and further, it rises back to the increasing trend in the following years. In addition to that, we can observe some flat plateaus in the increasing trend in some periods. (e.g. around 2019). Nevertheless, a sudden rise can be noticed right after that plateau in the years 2020 and 21. This can be inferred as the post-covid period. This variation can be explained using other attributes collected. Moreover, even though we could observe a regular seasonality, there are fluctuations of residuals in the seasonal decomposition of the HPI. This analysis creates a need for an advanced machine learning model to learn the pattern of the HPI variation and to forecast any sudden drops or rises based on other multiple available attributes.

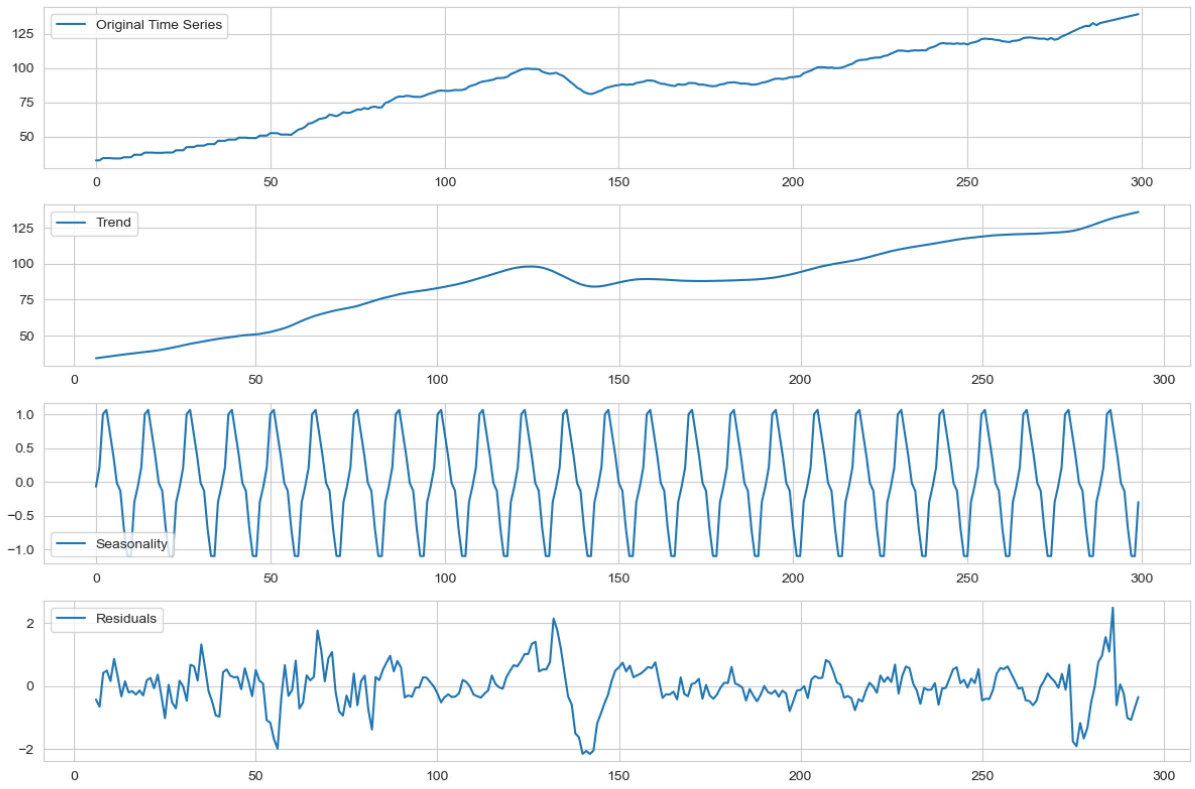


Figure 1: Exploratory time series data analysis

3.3 Data attributes or features available

The table below lists the total 16 features we used to train the machine learning model and more detailed simple explanations of them are provided next.

Table 1: Feature attributes for machine learning model and their detailed explanation

Feature	**Details**
DATE	The timestamp of each data record, that is month and year (e.g: 01-1970)
Household savings ratio	The proportion of household income that is saved rather than spent.
GDPHC	Gross Domestic Product of both households and non-profit institutions Gross Domestic Product (GDP): The aggregate worth of commodities and services generated inside a nation, including revenue that comes from households and non-profit organisations.
Retail Price Index	The Consumer Price Index (CPI) is a metric used for evaluating inflation by monitoring fluctuations in the prices of a selection of

	products and services commonly bought by households.
Whole Economy Index	An index representing the overall performance of the entire economy.
Claimant Count2rate	The unemployment rate is calculated by considering the number of individuals who are receiving unemployment-related benefits.
EngBank Rate	The Bank of England's official bank rate, which influences interest rates in the economy.
House Equity Withrawal	The amount of money withdrawn by homeowners from the equity in their homes through loans or mortgages.
GDP	Gross Domestic Product (GDP) refers to the aggregate value of all products and services that are generated within the geographic limits of a nation.
Durable Goods Total	The total value of long-lasting consumer goods, such as appliances and cars.
AEI Index	All Employee Jobs Index: An index measuring changes in the number of jobs in the economy.
UnEmpRate	The unemployment rate is a measure that indicates the proportion of the labour force that is currently without a job.
CPI VALUE	The Consumer Price Index (CPI) is a metric used to gauge the average variation in prices paid by consumers for goods and services over a certain period.
HPI Value	**Response variable.** House Price Index: An index measuring the changes in residential property prices over time

3.4 Variance Inflation Factor (VIF)

Collinearity refers to the condition in which two variables demonstrate a strong correlation and provide similar information regarding the variability within a specific dataset. To identify collinearity among variables, generate a correlation matrix and identify variables with significant absolute values. The Variance Inflation Factor (VIF) is a metric employed to evaluate the degree to which the variance of a calculated regression coefficient is amplified as a result of the existence of correlation among the predictors. Elevated VIF values signify a significant level of multicollinearity. In order to mitigate the issue of multicollinearity, one may opt to exclude or merge variables that exhibit strong correlations, acquire additional data, or employ regularisation methods. In addition to this, with the help of correlation analysis, we have dropped highly correlated features.

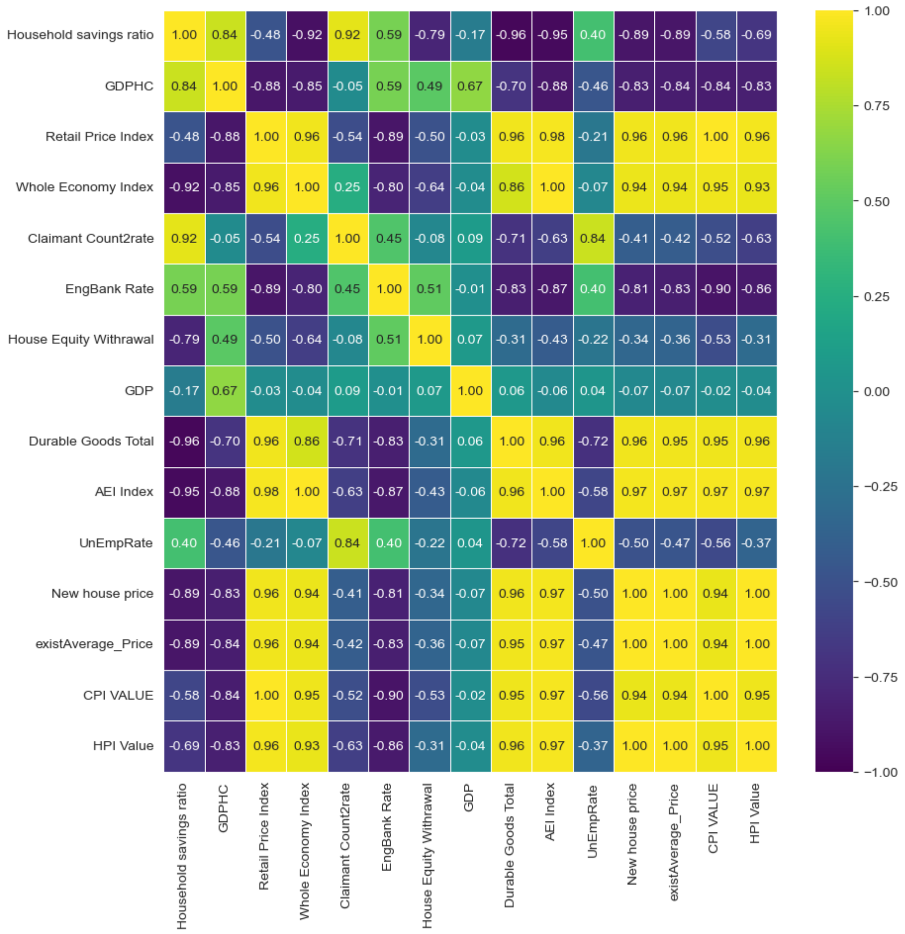
Table 2: Variable Inflation Factors of features

	Variable	**VIF**
1	Household savings ratio	28.9919
2	Retail Price Index	5706.7607
3	Whole Economy Index	699.0035
4	Claimant Count2rate	7.2931
5	EngBank Rate	32.2896
6	House Equity Withdrawal	5.8864
7	GDP	3.2880
8	Durable Goods Total	75.9584
9	AEI Index	785.0501
10	UnEmpRate	14.9272
11	New house price	347.8486
12	ExistAverage_Price	1339.0718
13	CPI VALUE	4436.2970
14	HPI Value	953.2961

15	Month	1.4550

3.5 Correlation analysis

Correlation analysis is a crucial data preprocessing step in the development of machine learning models. Correlation analysis is a method that allows for the identification of patterns, dependencies, and potential multicollinearity in a dataset by studying the correlations between distinct variables. Data scientists can enhance model efficiency and interpretability by making informed decisions regarding feature selection, which requires understanding the link between features. A strong correlation between two variables suggests that one of them may be redundant, and it is advisable to eliminate one of them to simplify the model and prevent overfitting. On the other hand, Variables that are correlated with the target variable can serve as valuable predictors. Furthermore, correlation analysis assists in discovering insignificant or poorly associated characteristics that could potentially be eliminated, thus optimising the model and enhancing computing efficiency. In summary, the careful implementation of correlation analysis is essential for improving the input feature set and maximising the effectiveness of machine learning models.



	Household savings ratio	GDPHC	Retail Price Index	Whole Economy Index	Claimant Count2rate	EngBank Rate	House Equity Withrawal	GDP	Durable Goods Total	AEI Index	UnEmpRate	New house price	existAverage_Price	CPI VALUE	HPI Value
Household savings ratio	1.00	0.84	-0.48	-0.92	0.92	0.59	-0.79	-0.17	-0.96	-0.95	0.40	-0.89	-0.89	-0.58	-0.69
GDPHC	0.84	1.00	-0.88	-0.85	-0.05	0.59	0.49	0.67	-0.70	-0.88	-0.46	-0.83	-0.84	-0.84	-0.83
Retail Price Index	-0.48	-0.88	1.00	0.96	-0.54	-0.89	-0.50	-0.03	0.96	0.98	-0.21	0.96	0.96	1.00	0.96
Whole Economy Index	-0.92	-0.85	0.96	1.00	0.25	-0.80	-0.64	-0.04	0.86	1.00	-0.07	0.94	0.94	0.95	0.93
Claimant Count2rate	0.92	-0.05	-0.54	0.25	1.00	0.45	-0.08	0.09	-0.71	-0.63	0.84	-0.41	-0.42	-0.52	-0.63
EngBank Rate	0.59	0.59	-0.89	-0.80	0.45	1.00	0.51	-0.01	-0.83	-0.87	0.40	-0.81	-0.83	-0.90	-0.86
House Equity Withrawal	-0.79	0.49	-0.50	-0.64	-0.08	0.51	1.00	0.07	-0.31	-0.43	-0.22	-0.34	-0.36	-0.53	-0.31
GDP	-0.17	0.67	-0.03	-0.04	0.09	-0.01	0.07	1.00	0.06	-0.06	0.04	-0.07	-0.07	-0.02	-0.04
Durable Goods Total	-0.96	-0.70	0.96	0.86	-0.71	-0.83	-0.31	0.06	1.00	0.96	-0.72	0.96	0.95	0.95	0.96
AEI Index	-0.95	-0.88	0.98	1.00	-0.63	-0.87	-0.43	-0.06	0.96	1.00	-0.58	0.97	0.97	0.97	0.97
UnEmpRate	0.40	-0.46	-0.21	-0.07	0.84	0.40	-0.22	0.04	-0.72	-0.58	1.00	-0.50	-0.47	-0.56	-0.37
New house price	-0.89	-0.83	0.96	0.94	-0.41	-0.81	-0.34	-0.07	0.96	0.97	-0.50	1.00	1.00	0.94	1.00
existAverage_Price	-0.89	-0.84	0.96	0.94	-0.42	-0.83	-0.36	-0.07	0.95	0.97	-0.47	1.00	1.00	0.94	1.00
CPI VALUE	-0.58	-0.84	1.00	0.95	-0.52	-0.90	-0.53	-0.02	0.95	0.97	-0.56	0.94	0.94	1.00	0.95
HPI Value	-0.69	-0.83	0.96	0.93	-0.63	-0.86	-0.31	-0.04	0.96	0.97	-0.37	1.00	1.00	0.95	1.00

Figure 2: Coefficient of correlation values of all features

As observed in the coefficient of correlation and VIF multi-collinearity, we have decided to opt out of the following two features since they have a very high correlation them.

1. New house price
2. Existing average house price

Another major reason for opting out of these features is the sparse data for these features since more than 35% of missing values were found for these attributes.

3.6 Data preprocessing

Data preprocessing is crucial for the effectiveness of predictive modelling as it transforms raw data into a format that is appropriate for machine learning algorithms. This essay explores essential data preprocessing techniques, such as imputing missing values, handling outliers, encoding categorical data, and performing feature engineering. Each technique tackles distinct issues in the data, guaranteeing that the input to machine learning models is dependable, enlightening, and free from irregularities. This section demonstrates the importance of these strategies in improving datasets to achieve optimal model performance through a thorough analysis. There are various data preprocessing techniques available for preparing our data to input machine learning models. Dealing with time series data entails additional factors and difficulties compared to static datasets. Time series data consists of observations gathered in a sequential manner, and the temporal aspect necessitates careful consideration when applying data preparation procedures. Here, we did the following data preprocessing methods.

i. Missing value imputing

ii. Outlier treatment

iii. Categorical data encoding

3.6.1 Missing value imputing

Missing values are a prevalent obstacle in real-world datasets and can greatly hinder the performance of models. Imputing missing values entails substituting or approximating these values by diverse techniques such as mean, median, or regression imputation. Imputation is a process that methodically deals with missing data, ensuring that predictive models are trained on datasets that are both complete and representative. Time series data frequently displays temporal dependencies, necessitating the inclusion of the time dimension when filling in missing values (Senan et al., 2021). To impute missing data of all considered attributes, we utilised the temporal context by employing two methods, such as the last observed value (ahead fill) and further explored more sophisticated techniques, such as regression imputation, while considering the time sequence and trends. The latter method has the higher accuracy while experimenting with the machine learning models in our case.

3.6.2 Outlier treatment

Detecting anomalies and determining appropriate methods for handling them are crucial stages in the data cleansing and processing procedure. Several techniques can be employed to identify outliers, including graphical methods like box plots and scatter plots, as well as statistical approaches like Z-score. Initially, the box plots were drawn to each feature. Nevertheless, there were some attribute values that deviated significantly from the norm. The process of identifying and addressing outliers is conducted utilising the incremental ‘three-sigma rule' since the data is time series and we have observed trends. Identify anomalies by taking into account the chronological sequence of data items. I applied exponentially weighted moving averages to identify anomalies in the data and impute them with the corresponding interpolated value.

3.6.3 Data encoding

Machine learning methods commonly necessitate numerical input, therefore requiring the encoding of categorical variables. Methods such as one-hot encoding and label encoding transform categorical data into a numerical representation without adding any biases. This guarantees that the algorithm can accurately comprehend and utilise categorical data, hence avoiding any misinterpretation of ordinal or nominal properties. In our case, the date column is converted to DateTime format and made as an index for deep learning models. At the same time, for regression models using the feature crossing technique for encoding, the ‘month’ column is created as a new feature by encoding the ‘DATE’ column.

4 METHODOLOGY

In this chapter, We introduce our chosen machine learning models, including both ensemble and deep learning models, accompanied by a concise theoretical foundation. Additionally, we provide a detailed explanation of the experimental setup used to process our data, as well as the criteria used to assess the effectiveness of the proposed models. Finally, we present an emerging sector of machine learning models that is explainable AI to interpret the features used for the proposed model and their contribution to better accurate results.

4.1 Developments of machine learning models

The gap in the literature on machine learning models for UK house price prediction is clearly depicted in the concluding paragraph of the literature review section. From that, we utilised machine learning models from linear regression, ensemble tree-based regressor and deep learning sequence models. The theoretical foundation for constructing such models is clarified in the subsequent sub-sections.

4.1.1 Linear Regression

Linear regression is a statistical technique that represents the connection between a dependent variable and one or more independent variables. Simple linear regression refers to a situation in which there is just one independent variable. When there are several explanatory factors, the technique is known as multiple linear regression. This statement contrasts with multivariate linear regression, which entails predicting multiple correlated dependent variables rather than a single scalar variable.

Linear regression is a statistical method that uses linear predictor functions to model relationships. The model parameters are determined based on the data that is available. These models are commonly known as linear models. Generally, it is presumed that the mean value of the response variable, given the values of the explanatory variables, conforms to a linear function. Alternatively, the conditional median or another quantile may be utilised, although this is less common. Linear regression, similar to other regression analyses, investigates the probability of the response variable depending on the predictor values. It fails to take into account the joint probability distribution of all variables, which is the central emphasis of multivariate analysis.

$$y_i = \beta_0 + \beta_1 x_{i,}1 + \beta_2 xi,2 + \cdots + \beta p xi,p + \varepsilon i, \qquad i = 1, 2, . . . , n,$$

While the linear regression method is a rather simplistic way to capture the complexity of housing predictions, there are fundamental concepts in linear regression that are used to develop other regression techniques. Many modern statistical learning approaches, such as splines and generalised additive models, can be considered generalisations or extensions of linear regression.

4.1.2 Ridge regression

Ridge regression, often referred to as Tikhonov regularisation, was introduced by Andrey Tikhonov (Tikhonov, 1966). It addresses the linear regression problem in a distinct manner from the least squares technique, with the aim of mitigating over-fitting.

This is achieved by incorporating an additional weight that is determined by a parameter commonly referred to as the l2-parameter or α. If the value of α is equal to zero, it results in an equivalent scenario to that of linear regression. The method is referred to as L2 regularisation, which involves the inclusion of a matrix in the equation to enhance smoothness. This matrix is constructed by multiplying the identity matrix I by the scalar α. The parameter α and the identity matrix I are involved in the equation.

4.1.3 Decision tree Regressor

A regression tree is created iteratively using a binary partitioning procedure, serving as a distinct type of non-linear regression. The regression tree method predicts the target variable by iteratively dividing the data into branches with leaves and nodes. Each branch corresponds to an attribute value that either leads to a decision node or divides into a separate branch. The Residual Sum of Squares (RSS) is calculated by taking the difference between the predicted value and the actual value at each split point, squaring this difference, and then totaling up all of these squared differences. The method chooses the branch that minimises the residual sum of squares (RSS).

4.1.4 Ensemble Regressor models

All models that will be constructed in this section for ensemble learning models are based on decision trees. They are a class of non-linear machine-learning methods. Ensemble methods are machine learning techniques that seek to generate multiple learners, referred to as weak learners, which are subsequently merged to form a unified prediction. This can significantly enhance the predicted accuracy and differs from conventional models where a single learner is optimised rather than merging multiple weak learners. Weak learners can be generated using several techniques, like as neural networks or decision trees. Ensemble learning is more efficient in reducing the generalisation error compared to weak learners, as a single decision tree commonly experiences overfitting. The majority of ensemble methods can be categorised into two types of ensemble learning approaches: sequential and parallel. Sequential ensemble approaches include the creation of individual models that learn from the previous model in a certain way. In parallel ensemble methods, the weak learners are constructed simultaneously rather than sequentially, meaning that each decision tree does not learn from the previous tree in the sequence.

1. Random forest regressor

Random forest is a machine learning methodology that falls under the classification of parallel combining methods, which will now be explained in detail. It elaborates on the notion of bagging, which is a technique for concurrently training many trees. Bagging involves constructing trees using data created using bootstrap aggregating, a method that selects several random samples with replacements from the original dataset. Every subsample B is utilised to train each tree f. Therefore, every tree acquires knowledge from numerous, somewhat distinct subsamples.

By employing bagging, it becomes feasible to generate numerous decision trees using the identical dataset, resulting in reduced variance and improved generalisation error. Nevertheless, bagging results in a substantial correlation among the regression trees due to the consistent utilisation of the same variables in all the trees. Consequently, this will augment the variability and exacerbate the overall generalisation mistake. In the context of random forest, this issue is addressed by employing a technique known as decorrelation, which aims to reduce the correlation between the individual trees. When considering only a specific number of variables at each split, the trees will vary not only in the data they are constructed on but also in the characteristics they are constructed with. During each split, a set of features j is produced randomly, and the feature with the splitting point t that minimises (1) is subsequently chosen. Due to the distinct characteristics of each tree, which are randomly determined, the primary factor influencing the outcome may not necessarily be considered first during the splitting process. As a result, the less significant variables have an opportunity to influence the final result. As a result, the model will exhibit improved performance when applied to data that was not used during its training. The random forest approach incorporates hyperparameters to mitigate overfitting to the training data. These parameters govern the number of trees used by the model, the maximum depth of each tree, and the number of features to be evaluated at each split. (Breiman, 1996)

2. Gradient Boosting

Boosting is a collection of algorithms that utilise sequential ensemble techniques to merge multiple weak models into a more powerful and dependable prediction model. The concept of boosting involves constructing a series of weak learners that, in each iteration, learn from the errors made by the preceding weak learners. Weak learners exhibit a slightly higher ability to predict the dependent variable compared to random guessing. However, when combined in an ensemble, they transform into strong learners with significantly improved predictive performance.

Gradient Boosting is a powerful combination machine-learning technique that belongs to the boosting genre of algorithms. The process involves training weak learners, usually decision trees, in a sequential manner and then aggregating their predictions to create a robust predictive model. The Gradient Boosting Machine (GBM) is a widely used implementation of Gradient Boosting. It aims to minimise the loss function by iteratively including weak learners that repair the errors made by prior learners. XGBoost, LightGBM, and CatBoost are improved iterations of the conventional Gradient Boosting method. These variations integrate optimisations and distinctive characteristics to improve overall performance. These algorithms are extensively utilised in diverse fields because of their capacity to manage intricate relationships in data, resilience against overfitting, and effective processing of sizable datasets.

3. XGBoost (Extreme Gradient Boosting)

XGBoost is a scalable algorithm for boosting trees that exhibits quicker training times, is susceptible to overfitting, and demonstrates robustness in the presence of outliers. It is known for its efficiency, scalability, high predictive performance, handling of large datasets, incorporation of regularisation techniques, and provision of feature importance analysis. XGBoost model used second-order Taylor expansion and added a regularisation term. The algorithm is comprised of a series of regression trees that repeatedly employ gradient descent to enhance the capabilities of weak learners (mostly decision trees) and elevate them to strong learners, hence enhancing the model's performance. The XGBoost algorithm expands a tree by continually adding new trees and executing feature splitting. A new function can be trained to fit the residual of an earlier prediction with each addition.

The original publication on XGBoost provides a comprehensive exposition of all the processes and computations, as presented by (Chen & Guestrin, 2016.). The XGBoost algorithm, short for Extreme Gradient Boosting, offers a highly optimised model that is renowned for its strong performance across a wide range of tasks and datasets. The approach is rooted in the gradient-boosting technique and offers significant benefits in terms of scalability and speed when compared to other gradient-boosting methods. When the technique was first introduced in 2016, it demonstrated a minimum speed advantage of tenfold over other widely used algorithms at the time.

4. CatBoost regressor

CatBoost is an advanced gradient boosting method specifically developed to handle categorical features in a dataset effectively. The system integrates a highly effective approach for handling categorical data by employing an oblivious decision tree algorithm. CatBoost reduces the danger of overfitting and improves predictive accuracy by using a customised approach to encode categorical features and a seamless mechanism to manage missing data. In their study, (Prokhorenkova et al., 2018) introduced the CatBoost algorithm in [2] and conducted a comparative analysis with XGBoost and LightGBM. The CatBoost learner is described, including the changes made to the GBDT method. CatBoost's primary improvement over Gradient Boosting is in its approach to handling categorical variables with large cardinality. CatBoost employs one-hot encoding for categorical data with low cardinality. The exact meaning of low cardinality varies based on the specific computing environment and the usage of CatBoost in specialised modes by the user. The present iteration of CatBoost, specifically version 0.23.2, assigns a default value of 255 in certain circumstances when executed on GPU's. Conversely, when executed on CPUs and certain other specified conditions are not satisfied, the default value is set to 2. This is a clear and significant example of CatBoost's responsiveness to hyperparameters. Modifying this hyper-parameter can lead to varying outcomes in terms of execution time and other performance measurements, as it affects both the selection of the processor utilised by CatBoost and the approach used to encode categorical information. The technique additionally incorporates a resilient kind of regularisation to mitigate the negative impact of model complexity on performance. CatBoost is a highly efficient tool for managing large datasets, automatically optimising hyperparameters, and providing built-in support for categorical features. This makes it an excellent choice for a wide range of applications, especially in domains where feature engineering and handling categorical variables are crucial factors to consider.

5. Light Gradient Boosting (LightGBM)

The LightGBM Regressor is a very efficient and powerful method specifically designed for regression problems in machine learning. It is a variation of the Light Gradient Boosting Machine (LightGBM). LightGBM utilises two novel methods, Gradient-based One-Side Sampling and Exclusive Feature Bundling, to efficiently manage a substantial amount of data instances and a significant number of features, respectively. LightGBM utilises a histogram-based learning approach, which discretises continuous features into distinct values to create histograms. This technique effectively reduces memory consumption and processing expenses. The tree's development approach, which focuses on optimising leaf-wise splits, enhances the training process by increasing its speed. However, when the sample size is large or the feature dimension is really high, the efficiency and accuracy of GBDT still do not yield satisfactory results. Gradient Boosting Decision Trees (GBDT) is an ensemble algorithm that utilises decision trees as its base classifier. The primary computational expense lies in identifying the optimal split points during the learning process of decision trees. (Ke et al., 2017)introduced a highly efficient gradient-boosting decision tree method called LightGBM, which incorporates gradient-based one-side sampling (GOSS) and exclusive feature bundling (EFB). Moreover, LightGBM integrates advanced methods, including gradient-based one-sided sampling and exclusive feature bundling, which improve its predictive accuracy. LightGBM Regressor is a popular choice for regression problems due to its seamless and effective parallel computing and flexibility in parameter adjustment. It is known for producing accurate predictions and preserving computational efficiency in various applications.

6. Extra trees regressor

The Extra Trees Regressor is a complex collaborative learning method employed for regression challenges. It is a variant of the Random Forest algorithm and is renowned for its exceptional capabilities. The procedure entails generating multiple decision trees during the training phase, where each tree is trained on a randomly selected portion of the features and training data. The Extra Tree Regression (ETR) technique is a variant derived from the Random Forest (RF) model and was initially proposed by (Geurts, P and Ernst, 2006). The Extra Tree Regression

(ETR) technique use the conventional top-down methodology to produce a collection of unpruned decision or regression trees. The Random Forest (RF) model utilises two steps, bootstrapping and bagging, to do regression. During the bootstrapping procedure, a set of decision trees is created by expanding each tree using a randomly chosen sample from the training dataset. The bagging step, which comprises of two stages, is utilised to divide the decision tree nodes after the ensemble is formed. This entails the selection of numerous random groups of training data during the initial bagging step. The decision-making process is carried out by selecting the most advantageous subset and its related value. The Extra Trees Regressor excels at managing noisy data, capturing complex correlations, and surpassing traditional regression algorithms in difficult scenarios. The ETR and RF systems reveal two significant distinctions. The ETR algorithm use all accessible cutting points to divide nodes by randomly selecting from these locations. Moreover, it use the entire collection of training data to cultivate the trees with the objective of minimising bias. The ETR approach utilises two parameters, namely k and nmin, to regulate the splitting process. The parameter k dictates the amount of features that are randomly chosen in each node, while the nmin parameter specifies the minimum sample size needed to separate nodes. The variables K and nmin control the level of attribute selection and the average strength of the output noise, respectively.Its versatility and effectiveness make it a potent tool in predictive modelling across several scientific and practical fields.The reference is from the study conducted by (Mastelini et al., 2023).

4.1.5 Deep learning models for time series forecasting

Here we incorporated the basic neural network (Multilinear perceptron: MLP) and recent advanced state-of-the-art Convolutional LSTM. The ConvLSTM model was introduced for weather forecasting (Shi et al., 2015), which is a type of RNN with a novel combination of convolution and LSTM layers. The convolution layer learns close neighbours in the input 4-D tensor. A ConvLSTM network is a chain of repeating modules called $cells$ (Ct) and contains the following gates and states:

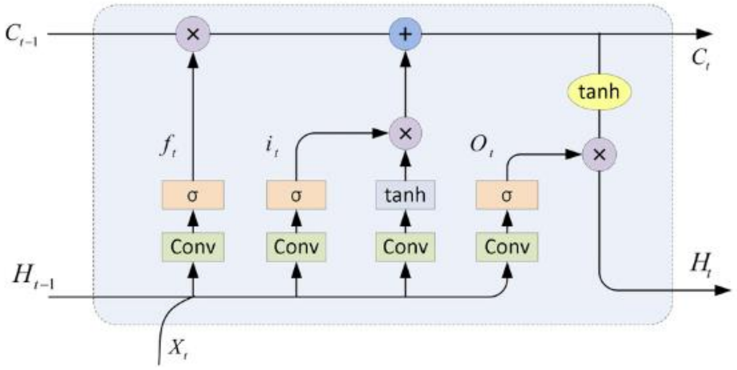


Figure 3: The structure of the ConvLSTM cell

i_t: The input gate updates the cell state using sigmoid activation of the previous hidden state and new input state

$$i_t = \sigma(W^i * x_t + R^i * h_{t-1} + U^i \circ c_{t-1} + b^i)$$

ft: Forget gate that selectively remembers or forgets information.

$$f_t = \sigma(W^f * x_t + R^f * h_{t-1} + U^f \circ c_t + b^f)$$

c_t: Cell state

$$c_t = f_t \circ c_{t-1} + i_t \circ tanh(W^c * x_t + R^c * h_{t-1} + b^c)$$

O_t: The output gate decides the value of the subsequent concealed state, which is transmitted to the next time step together with the updated cell state.

$$o_t = \sigma(W^o * x_t + R^o * h_{t-1} + U^o \circ c_t + b^o)$$

h$_t$: Hidden state or the cell output

$$h_t = o_t \circ tanh(c_t)$$

In the given equations, σ represents the sigmoid activation function, * symbolises the convolution operation, and ∘ represents the element-wise product.

Figure 3 demonstrates that ConvLSTM is an upgraded iteration of LSTM that integrates convolutional layers (depicted by green-colored boxes) in both the input-to-state and state-to-state transitions. Next, four ConvLSTM cells are arranged in a stacked configuration to create an encoder-decoder architecture. In order to simulate the multi-input time series of a 4-D tensor, the fully connected layer is replaced with the time-dispersed layer, resulting in the generation of multiple outputs. Furthermore, dropout layers are incorporated to mitigate overfitting, and batch normalisation is employed to enhance the performance of the neural network.

4.2 Experiments

The input data for the proposed model are made to time series data for deep learning ConvLSTM model and tabular data for regression models. Every model undergoes training using 70% of the observations and is then evaluated using the remaining 30%. Every model utilises identical training and test data. The specified split is a conventional proportion used for dividing the data (Brownlee, 2018). The hyperparameters of each model have been adjusted to the associated training dataset using 5-fold cross-validation to enhance their performance. The grids of values were generated by the utilisation of random search, followed by fine-tuning based on the best outcomes.

4.3 Criteria to measure the performance of models.

The performance of the proposed model with baseline models is evaluated by the following evaluation metrics.

1. **Mean absolute error (MAE) :**

Mean Absolute Error (MAE) quantifies the average absolute deviation between the expected and actual values. It exhibits a lower sensitivity to outliers. Smaller MAE values indicate superior model performance.

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i|$$

2. **Mean absolute percentage error (MAPE) :**

MAPE measures the average percentage difference between the predicted and true values, taking the absolute value. It is widely used in forecasting models. Lower MAPE values indicate better model performance.

$$MAPE = \frac{1}{N}\sum_{i=1}^{N} |\frac{y_i - \hat{y}_i}{y_i}$$

3. **Root mean square error (RMSE) :**

RMSE is the square root of the Mean squared error (MSE), providing an interpretable metric in the same units as the target variable. It **penalises** large errors more than other metrics.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

- yi is the true (actual) travel time of ith trip among N samples
- y^ is the predicted travel time of ith trip among N samples i

4. **R Squared - Coefficient of determination**

   The coefficient of determination, denoted as R2 in statistics, is the proportion of the dependent variable's variability that can be precisely predicted using the independent variables.

   When evaluating the appropriateness of the match between simulated (Y_pred) and measured (Y_obs) values, it is not appropriate to depend just on the R2 of the linear regression equation (i.e., Y_obs = m × Y_pred + b). The coefficient of determination, R2, quantifies the degree of a linear association between the observed Y and the predicted Y. However, while assessing the adequacy of the fit, only a specific linear correlation should be taken into account. R-squared is a statistical metric that quantifies the level of accuracy of a model's fit. Assuming that the coefficient of determination, R squared, has a value of 0.49. This signifies that 49% of the variance in the dependent variable has been accounted for, while the remaining 51% of the variance remains unaccounted for. The coefficient of determination, R2, in regression, is a statistical measure that quantifies the extent to which the regression predictions closely reflect the actual data points. A R2 score of 1 indicates a complete correspondence between the regression predictions..

4.4 SHAP values for feature importance

Examining explanatory variables helps elucidate the relationship between various factors and their impact on bus journey times. This text will elucidate the forecasts generated by the machine learning algorithm, which was previously regarded as an enigmatic entity. We utilised the SHapley Additive exPlanations (SHAP) method (Lundberg et al., n.d.). SHAP computes the contribution of each feature to the individual predictions of the model using cooperative game theory principles. The algorithm evaluates all potential coalitions (i.e., all possible combinations) of features and calculates the incremental impact of each feature on the forecast by comparing the prediction with and without the feature. The Shapely value is calculated as the mean of each feature's marginal contributions. These figures indicate that contributions can be either positive or negative. The equation below demonstrates the straightforward calculation steps for Shapely values.

$$\phi_i = \frac{1}{M} \sum_{k=1}^{M} \left[f(x_{\pi_k(i)}) - E(f(x_{\pi_k(i)}) \mid x_{\pi_{-k}(i)}) \right]$$

Figure 4: SHAPly calculation

Model interpretability using SHAP values can be achieved at both a global and local level. The global explanations provide a clear understanding of how the attributes contribute to the output of the model. The information is represented using a concise graphical representation. Simultaneously, the local explanations offer insight into the rationale behind the model's specific decision or feature variations for each unique prediction. The visualisation of this can be achieved by utilising waterfall plots. The insights derived from this interpretation of feature importance will ultimately assist authorities in making decisions, managing, and regulating house prices.

5 RESULTS AND ANALYSIS

This chapter elaborates on the results of the performance evaluation of the proposed models and compares their predictions. Moreover, feature explanatory analysis is also presented in the next sub-section,

5.1 Comparison of predictions of machine learning models

The table displays the Mean Absolute Percentage Error (MAPE) and its standard deviation for several regression algorithms, offering significant insights into their performance on a specific dataset. The Mean Absolute Percentage Error (MAPE) is a widely used metric for assessing the precision of regression models. It quantifies the percentage deviation between predicted and actual values.

Table 3: MAPE comparision of machine learning models

Algorithms	**Mean MAPE**	**Std MAPE**
Extra Trees Regressor	**0.814995**	**0.115721**
Random Forest	1.098841	0.16913
ConvLSTM	1.138123	0.099568
CatBoost Regressor	1.158862	0.221649
Extreme Gradient Boosting	1.16043	0.080856
Decision Tree Regressor	1.273823	0.162001
Light Gradient Boosting Machine	1.336081	0.186578
AdaBoost Regressor	2.439847	0.120401
Linear Regression	2.600792	0.268639
Bayesian Ridge	2.62009	0.266996
Ridge Regression	2.631679	0.26679

The "Extra Trees Regressor" exhibits the lowest mean MAPE at 0.814995, indicating high predictive accuracy. The standard deviation of MAPE provides insights into the variability of model performance. Lower standard deviations, such as those for "Extra Trees Regressor" and "Extreme Gradient Boosting (XGBoost)," suggest more consistent and stable predictions.

Table 4: Comparision of machine learning models using R2 value

Algorithms	**Mean R-squared**	**Std R-squared**
Extra Trees Regressor	**0.999089**	**0.000147**
Random Forest	0.998364	0.000478
ConvLSTM	0.998230	0.000209
CatBoost Regressor	0.997955	0.001072
Extreme Gradient Boosting	0.997839	0.000497
Light Gradient Boosting Machine	0.997669	0.000608
Decision Tree Regressor	0.995957	0.001653
AdaBoost Regressor	0.992885	0.001300
Linear Regression	0.992606	0.000636
Bayesian Ridge	0.992601	0.000662
Ridge Regression	0.992592	0.000678

- "Extra Trees Regressor" stands out with an exceptionally high mean R-squared of 0.999089, indicating an outstanding fit to the data. The low standard deviation (0.000147) suggests consistent and robust performance.
- "Random Forest" and "ConvLSTM" also demonstrate excellent fits with mean R-squared values exceeding 0.998.
- While high mean R-squared values are desirable, it's essential to consider the standard deviation. A model with both a high mean R-squared and a low standard deviation is preferable, as it indicates both accuracy and stability in performance.

Hence, the Extra Trees Regressor is the best predicting model.

5.2 Feature explanation using SHAP values

The computed mean SHAP values of each feature for the house price prediction model are plotted in the horizontal bar chart shown in Figure 6 below.

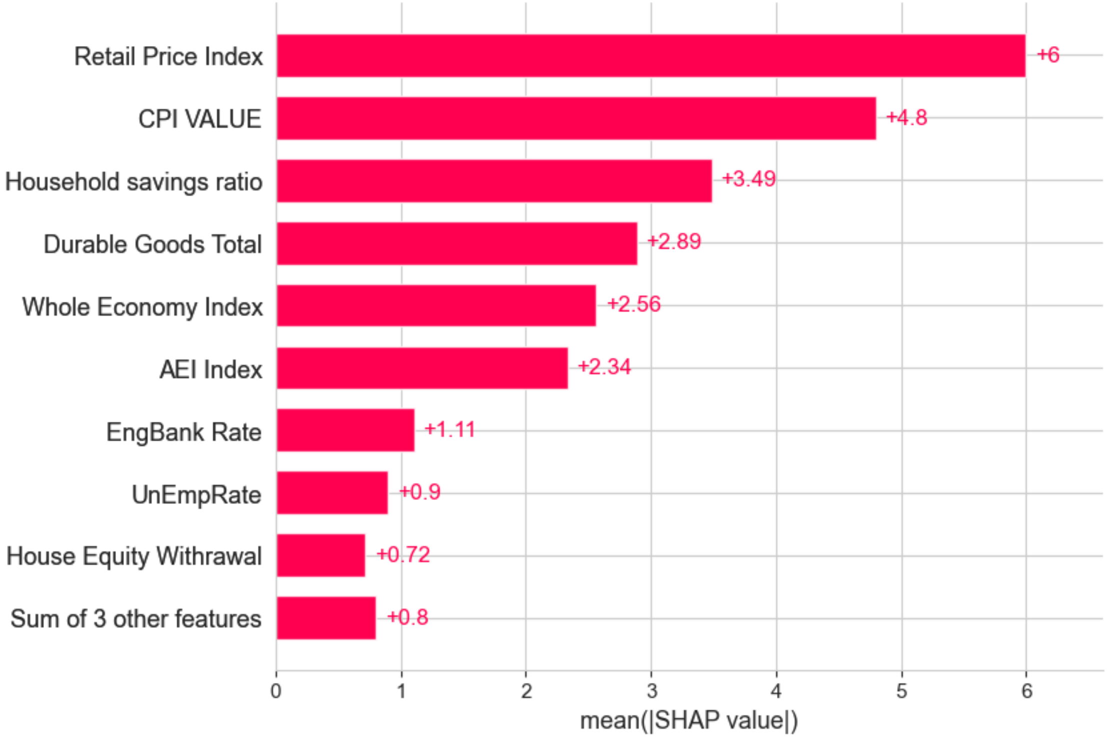


Figure 5: mean SHAP values of all features

It is concluded that the 'Retail Price Index' and CPI value have a significant contribution to the prediction of house prices. It is obviously true that other market prices heavily influence the price of houses.

The next part of this analysis is the local explanation. It is visualised by waterfall plots, and we can analyse the positive and negative contribution of each feature for every predicted value. When the characteristic has a beneficial effect, it augments the predictive value beyond it's initial value, and opposite. Such waterfall plots for a selected instance for the house price prediction model are shown in Figure 4 below. For the prediction of an HPI value of 97.656, the model starts from the estimated average value of 87.063, and then the way the other features contribute to a more accurate prediction is explained in the waterfall plot.

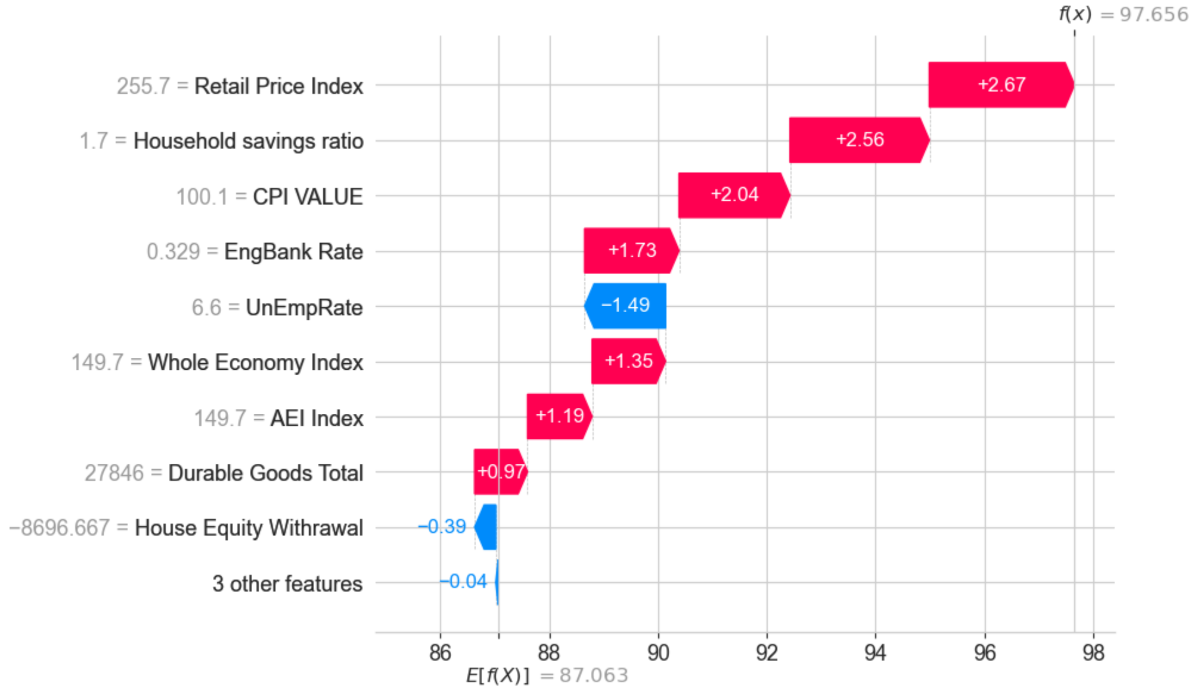


Figure 6: Waterfall plot showing local explanation

The next part of the feature importance analysis is the global explanation provided by SHAP values. The features' impact on the model output is visually represented in a violin plot, which provides an organised and comprehensive overview. If the feature has a positive impact (that is, this feature value increases the house price index), the Shapely value will be positive and vice versa. The violin plots for the prediction model with an Extra tree regressor are shown in Figure 5 below.

There, we can observe that the two features, ‘Unemployment rate’ and ‘House equity withdrawal’, have a negative impact on prediction as they reduce HPI while they increase. Among the other features, the retail price index, CPI value, and household savings ratio significantly impact the prediction, and these attributes highly explain the HPI response variable. It must be noted that we need to maintain the accuracy and correctness of the data collected for the above three attributes more minutely.

We did the SHAP value calculation with all the proposed regression tree-based models, and all showed similar plots. Therefore, the plots associated with the higher accurate model, ‘Extra tree regressor’, are illustrated here for more detailed depictions.

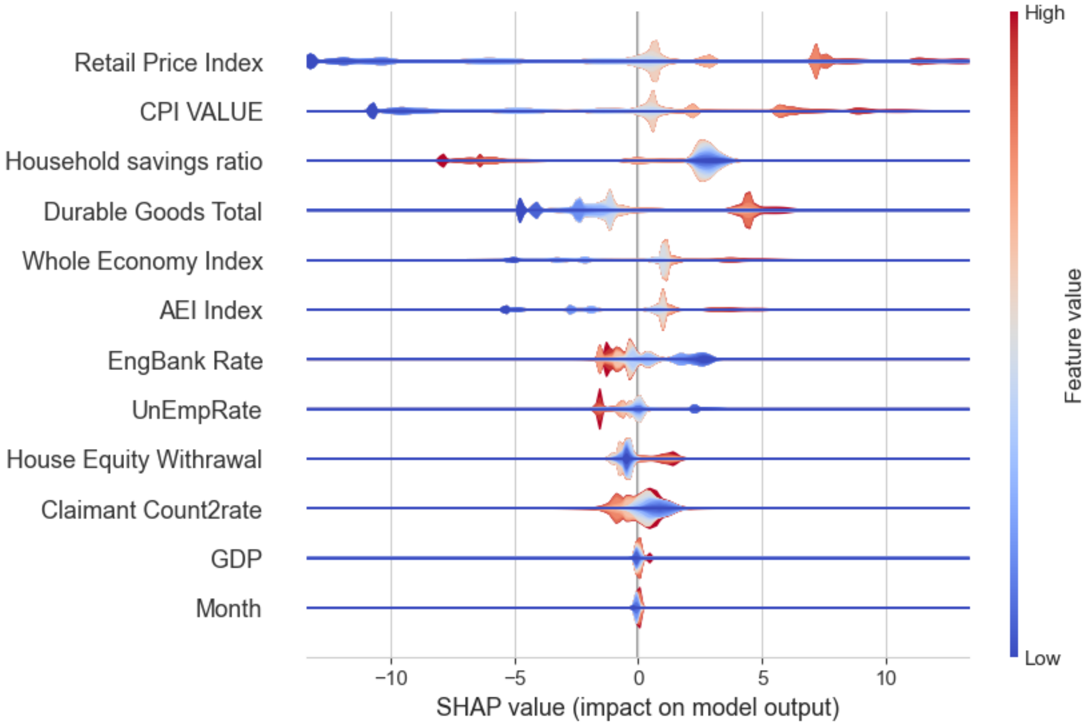


Figure 7: Violin plot showing a global explanation of all features.

At last, the overall summary of SHAP values for every feature is visualised through the heatmap, as shown in Figure 7 below.

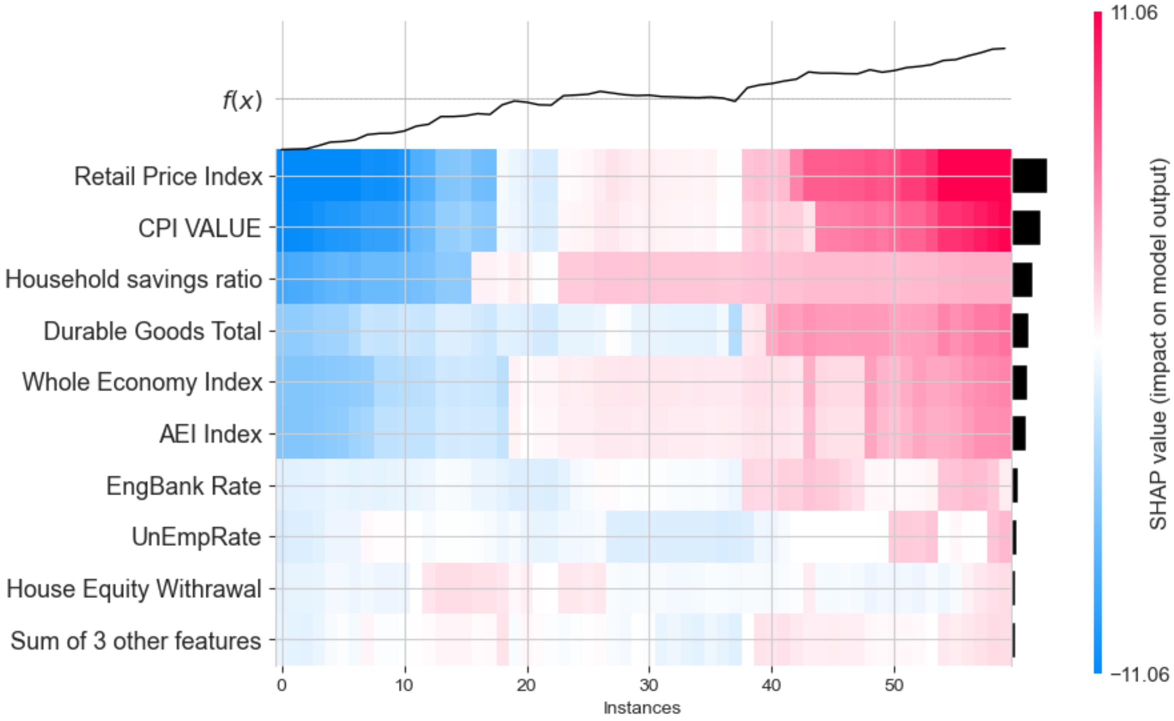


Figure 8: Heat map of SHAP values

The three mentioned forms of plots can be implemented to graphically highlight the individual impact of each attribute on the projected values of the machine learning model. This SHAP values ground breaks the ideology that the machine learning models are a black box. Further, the importance of features can be supportive in decision-making by the relevant authorities to manage efficient and reliable house price prediction systems.

6 CONCLUSIONS

By employing a range of analytical and graphical techniques, we assessed the accuracy of different housing pricing models when applied to actual data on single-family homes in UK. Furthermore, our models were instrumental in determining the home attributes that had the highest correlation with price and had the most potential to account for the majority of price fluctuations. In addition, we enhanced the accuracy of our models' predictions by including the influence of spatial position.

The study included many statistical techniques including simple and multiple linear regression, lasso regression, decision trees, KNN, SVM, and ensemble models such as Random Forest, XGBoost, and Extra tree regressor. Additionally, deep learning-based models like MLP and ConvLSTM were also utilised. The Extra tree regressor demonstrated superior performance compared to the other models by a substantial margin. It holds a mean MAPE of 0.81 and an R-squared value of almost 1 (one). This is the highest accuracy found for a machine learning model for house product prediction in the UK. This is possible with the very suitable and complementary feature/attribute selection.

The primary objective of this thesis was to assess the significance of each predictor in elucidating the fluctuations in price for a specific collection of home features. Overall, the retail price index, CPI value, and household savings ratio significantly impact the prediction, and these attributes highly explain the HPI response variable. The variations of these attributes hugely affect the marginal difference in the predictions of HPI. Hence, the stakeholders and people can utilise this system and make aware of the other macro-economic factors and can decide on HPI in prior for their profitable avenues.

REFERENCES

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140. https://doi.org/10.1007/BF00058655/METRICS

Brownlee, J. (2018). *Deep learning for time series forecasting: predict the future with MLPs, CNNs and LSTMs in Python*. https://books.google.com/books?hl=en&lr=&id=o5qnDwAAQBAJ&oi=fnd&pg=PP1&dq=J.+Brownlee.+Deep+learning+for+time+series+forecasting:+predict+the+future+with+MLPs,+CNNs+and+LSTMs+in+Python.+Machine+Learning+Mastery,+2018.&ots=yI-4yPsi58&sig=c-SwM6KrP4e4Y3Ayoss2WYSXqq4

Čeh, M., Kilibarda, M., Lisec, A., & Bajat, B. (2018). Estimating the Performance of Random Forest versus Multiple Regression for Predicting Prices of the Apartments. *ISPRS International Journal of Geo-Information 2018, Vol. 7, Page 168*, *7*(5), 168. https://doi.org/10.3390/IJGI7050168

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/2939672

Ekberg, J. J. L. (2022). *Comparison of different machine learning methods' capability to predict housing prices*. 15. https://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-319820

Fan, C., Cui, Z., & Zhong, X. (2018). House prices prediction with machine learning algorithms. *ACM International Conference Proceeding Series*, 6–10. https://doi.org/10.1145/3195106.3195133

Geurts, P and Ernst, D. (2006). Extremely randomized trees. *Machine Learning*.

Gu, J., Zhu, M., & Jiang, L. (2011). Housing price forecasting based on genetic algorithm and support vector machine. *Expert Systems with Applications*, *38*(4), 3383–3386. https://doi.org/10.1016/J.ESWA.2010.08.123

Hong, J. (2020). An Application of XGBoost, LightGBM, CatBoost Algorithms on House Price Appraisal System. *Housing Finance Research*, *4*, 33–64. https://doi.org/10.52344/HFR.2020.4.0.33

Hu, L., He, S., Han, Z., Xiao, H., Su, S., Weng, M., & Cai, Z. (2019). Monitoring housing rental prices based on social media:An integrated approach of machine-learning algorithms and hedonic modeling to inform equitable housing policies. *Land Use Policy*, *82*, 657–673. https://doi.org/10.1016/J.LANDUSEPOL.2018.12.030

Ja'afar, N. S., Mohamad, J., & Ismail, S. (2021). MACHINE LEARNING FOR PROPERTY PRICE PREDICTION AND PRICE VALUATION: A SYSTEMATIC LITERATURE REVIEW. *PLANNING MALAYSIA*, *19*(3), 411–422. https://doi.org/10.21837/PM.V19I17.1018

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems*, *30*. https://github.com/Microsoft/LightGBM.

Kok, N., Koponen, E. L., & Martínez-Barbosa, C. A. (2017). Big Data in Real Estate? From Manual Appraisal to Automated Valuation. *The Journal of Portfolio Management*, *43*(6), 202–211. https://doi.org/10.3905/JPM.2017.43.6.202

Liu, J. G., Zhang, X. L., & Wu, W. P. (2006). Application of fuzzy neural network for real estate prediction. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *3973 LNCS*, 1187–1191. https://doi.org/10.1007/11760191_173/COVER

Lundberg, S., information, S. L.-A. in neural, & 2017, undefined. (n.d.). A unified approach to interpreting model predictions. *Proceedings.Neurips.CcSM Lundberg, SI LeeAdvances in Neural Information Processing Systems, 2017•proceedings.Neurips.Cc*. Retrieved December 29, 2023, from https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

Mastelini, S. M., Nakano, F. K., Vens, C., & De Carvalho, A. C. P. D. L. F. (2023). Online Extra Trees Regressor. *IEEE Transactions on Neural Networks and Learning Systems*, *34*(10), 6755–6767. https://doi.org/10.1109/TNNLS.2022.3212859

Tikhonov, A. (1966). On the stability of the functional optimization problem. *Elsevier USSR Computational Undefined*. https://www.sciencedirect.com/science/article/pii/0041555366900036

Nguyen, N. and Cripps, A. (2001). Predicting housing value: A comparison of multiple regression analysis and artificial neural networks. *The Journal of Real Estate Research*, *22*(3), 313-336.

Park, B., & Kwon Bae, J. (2015). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. *Expert Systems with Applications, 42*(6), 2928–2934. https://doi.org/10.1016/J.ESWA.2014.11.040

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*, *31*. https://github.com/catboost/catboost

Revend, W. (2020). *Predicting House Prices on the Countryside using Boosted Decision Trees*. https://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-279849

Selim, H. (2009). Determinants of house prices in Turkey: Hedonic regression versus artificial neural network. *Expert Systems with Applications, 36*(2), 2843–2852. https://doi.org/10.1016/J.ESWA.2008.01.044

Senan, E. M., Abunadi, I., Jadhav, M. E., & Fati, S. M. (2021). Score and Correlation Coefficient-Based Feature Selection for Predicting Heart Failure Diagnosis by Using Machine Learning Algorithms. *Computational and Mathematical Methods in Medicine, 2021*. https://doi.org/10.1155/2021/8500314

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., Woo, W.-C., & Kong Observatory, H. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Proceedings.Neurips.CcX Shi, Z Chen, H Wang, DY Yeung, WK Wong, W WooAdvances in Neural Information Processing Systems, 2015•proceedings.Neurips.Cc*. https://proceedings.neurips.cc/paper/2015/hash/07563a3fe3bbe7e3ba84431ad9d055af-Abstract.html

Shukry, M., Radzi, M., & Muthuveerappan, C. (2012). *FORECASTING HOUSE PRICE INDEX USING ARTIFICIAL NEURAL NETWORK*.

UK House Price Index - Office for National Statistics. (n.d.). Retrieved December 29, 2023, from https://www.ons.gov.uk/economy/inflationandpriceindices/bulletins/housepriceindex/march2023

Wang, X., Wen, J., Zhang, Y., & Wang, Y. (2014). Real estate price forecasting based on SVM optimized by PSO. *Optik*, *125*(3), 1439–1443. https://doi.org/10.1016/J.IJLEO.2013.09.017