Frank Sharpe – l39222594

Project Dissertation

AI tools were used for code debugging and conceptual planning. All outputs were reviewed, adapted, and validated

# Predicting House Prices Using Mathematical and ML Models

# Figures

[Figure 1 - (KEEP and Keep, 2025) - house building 7](#_Toc216688700)

[Figure 2 - (KEEP and Keep, 2025) - house price index 7](#_Toc216688701)

[Figure 3 - (Bekhruz Tuychiev, 2019) - shows the usage of models shows that linear regression is most popular 11](#_Toc216688702)

[Figure 4 - XGBoost - (Kavlakoglu and Russi, 2024) 12](#_Toc216688703)

[Figure 5 - (Kavlakoglu, 2021) - decision tree example! 13](#_Toc216688704)

[Figure 6 (Yao, 2024) - linear regression example. 14](#_Toc216688705)

[Figure 7 - (Abdul-Rahman et al., 2021) - actual price against predicted. 16](#_Toc216688706)

[Figure 8 - 2nd study reuslts - (Cui, Liu and Ma, 2024) 18](#_Toc216688707)

[Figure 9 - (Ons.gov.uk, 2025) - Uk house price index. 21](#_Toc216688708)

[Figure 10 - prop.csv data - (Chi et al., 2019) 22](#_Toc216688709)

[*Figure 11 - pipeline diagram explain process* 22](#_Toc216688710)

[Figure 12 - key fields used in the model to make the predictions. 23](#_Toc216688711)

[Figure 13 - cleaning steps and purpose. 24](#_Toc216688712)

[Figure 14 - data cleaning sql used by the dataset - (Chi et al., 2019) 25](#_Toc216688713)

[Figure 15 - clean hpi data - commentsi in code to explain steps used. 26](#_Toc216688714)

[Figure 16 - clean hpi data - part 2. 27](#_Toc216688715)

[Figure 17 - clean property data - comments in code to explain steps used. 27](#_Toc216688716)

[Figure 18 - linear regression model graph. 28](#_Toc216688717)

[Figure 19 – Random forest graph model. 28](#_Toc216688718)

[Figure 20 - Decision tree model. 29](#_Toc216688719)

[Figure 21 - XGBoost results visualised. 30](#_Toc216688720)

[Figure 22 - data used for train test split. 31](#_Toc216688721)

[Figure 23 - reuslts from all models used. 32](#_Toc216688722)

[Figure 24 - Top 10 coefficients. 33](#_Toc216688723)

[Figure 25 - size of dataset used. 33](#_Toc216688724)

[Figure 26 - how house prices changed over time - (Lloyds Banking Group, 2023) 36](#_Toc216688725)

[Figure 27 - average house price per region - (HM Land Registry, 2023) 41](#_Toc216688726)

[*Figure 28 - shap model graph.* 42](#_Toc216688727)

[*Figure 29 - shap model for MAE.* 43](#_Toc216688728)

[Figure 30 - demand for house types - (Zoopla.co.uk, 2023) 46](#_Toc216688729)

[Figure 31 - 500.xlsx - house price sale list 53](#_Toc216688730)

[Figure 32 - met.xlsx - property stat avg 53](#_Toc216688731)

[Figure 33 - libarys used 54](#_Toc216688732)

[Figure 34 - load both files merge deed date and time perdod 54](#_Toc216688733)

[Figure 35 - sort the data and create means for the data such as price paid and great Britain average 54](#_Toc216688734)

[Figure 36 - plot the data using the data as axis 55](#_Toc216688735)

[Figure 37 - graph to show the results 55](#_Toc216688736)

[Figure 38 - prop.csv - list of property data 56](#_Toc216688737)

[Figure 39 - prop.csv - number of rooms and size 57](#_Toc216688738)

[Figure 40 - load the excel files 58](#_Toc216688739)

[Figure 41 - clean the data removing £ signs and empty rows 58](#_Toc216688740)

[Figure 42 - clean prop.csv data 59](#_Toc216688741)

[Figure 43 - put property types into categories 59](#_Toc216688742)

[Figure 44 - combine the dates for hpi and sale date 59](#_Toc216688743)

[Figure 45 - display property sale date and price against hpi at sale 60](#_Toc216688744)

[Figure 46 - results of first 10 rows 60](#_Toc216688745)

[Figure 47 - library needed 61](#_Toc216688746)

[Figure 48 - clean the data from hpi 62](#_Toc216688747)

[Figure 49 - clean property data 62](#_Toc216688748)

[Figure 50 - put property types into catorgories 63](#_Toc216688749)

[Figure 51 - merge sale date against hpi at sale 63](#_Toc216688750)

[Figure 52 - create a final dataset to use for training 64](#_Toc216688751)

[Figure 53 - get the features needed to train x and y 64](#_Toc216688752)

[Figure 54 - features being used 65](#_Toc216688753)

[Figure 55 - train test split of 20/80 and linear regression implementation 65](#_Toc216688754)

[Figure 56 - visualise the linear regression 66](#_Toc216688755)

[Figure 57 - decision tree model libary 67](#_Toc216688756)

[Figure 58 - decision tree used to test the data 67](#_Toc216688757)

[Figure 59 - visualise the data 67](#_Toc216688758)

[Figure 60 - comparisons of the models 68](#_Toc216688759)

[Figure 61 - linear regression graph 69](#_Toc216688760)

[Figure 62 - decision tree graph 69](#_Toc216688761)

[Figure 63 - linear regression model 71](#_Toc216688762)

[Figure 64 - decision tree model 71](#_Toc216688763)

[Figure 65 - linear regression graph 73](#_Toc216688764)

[Figure 66 - decision tree graph 73](#_Toc216688765)

[Figure 67 - random forest 75](#_Toc216688766)

[Figure 68 - random forest graph code 75](#_Toc216688767)

[Figure 69 - xgboost code 76](#_Toc216688768)

[Figure 70 - shap model 77](#_Toc216688769)

# Table of Contents

[Predicting House Prices Using Mathematical and ML Models 0](#_Toc216688634)

[Figures 1](#_Toc216688635)

[Table of Contents 3](#_Toc216688636)

[Abstract 6](#_Toc216688637)

[1.0 Introduction 7](#_Toc216688638)

[1.1 Research questions 9](#_Toc216688639)

[1.1.1 RQ1. How accurately can machine learning models predict residential property prices using structured housing and market data? 9](#_Toc216688640)

[1.1.2 RQ2. How can model transparency and ethical considerations be improved when using “black-box” machine learning methods for property price prediction? 9](#_Toc216688641)

[2.0 Literature Review 10](#_Toc216688642)

[2.1 Hedonic pricing models 10](#_Toc216688643)

[2.2 Machine learning models in real estate 10](#_Toc216688644)

[2.3 XGBoost 12](#_Toc216688645)

[2.4 Random forest 13](#_Toc216688646)

[2.5 Linear regression 14](#_Toc216688647)

[2.6 Challenges in data 15](#_Toc216688648)

[2.7 Other studies to compare against 16](#_Toc216688649)

[2.7.1 Kuala Lampur 16](#_Toc216688650)

[2.7.2 2nd study - AI in Real Estate: Forecasting House Prices with Advanced Machine Learning Models 18](#_Toc216688651)

[3.0 Data and Methodology 19](#_Toc216688652)

[3.1 overview 19](#_Toc216688653)

[3.2 Data sets 21](#_Toc216688654)

[3.3 Pipeline Diagram 22](#_Toc216688655)

[3.4 fields 23](#_Toc216688656)

[3.5 Clean data 24](#_Toc216688657)

[4.0 Model outputs 28](#_Toc216688658)

[4.1 Mathematical model 1: Linear regression 28](#_Toc216688659)

[4.2 Model 2 : Random forest 28](#_Toc216688660)

[4.3 Model 3 : decision tree model 29](#_Toc216688661)

[4.4 Model 4 : XGBoost 30](#_Toc216688662)

[5.0 Evaluation 31](#_Toc216688663)

[5.1 Train Test Split 31](#_Toc216688664)

[5.2 Results 32](#_Toc216688665)

[5.3 Top 10 featured coefficients 33](#_Toc216688666)

[5.4 Size of dataset 33](#_Toc216688667)

[6.0 Discussion 34](#_Toc216688668)

[6.1 introduction 34](#_Toc216688669)

[6.2 evaluation metrics 34](#_Toc216688670)

[6.3 model selection reasoning 35](#_Toc216688671)

[6.4 model limitations 35](#_Toc216688672)

[6.5 external factors 36](#_Toc216688673)

[6.6 results overview 37](#_Toc216688674)

[6.7 real world impact 37](#_Toc216688675)

[6.8 XGBoost performance detail 38](#_Toc216688676)

[6.9 regional inequality 39](#_Toc216688677)

[6.10 feature considerations 39](#_Toc216688678)

[6.11 comparison with other studies 40](#_Toc216688679)

[6.12 coefficient interpretation 41](#_Toc216688680)

[7.0 Shap – Interpretation of model results 42](#_Toc216688681)

[8.0 Ethics and Limitations 45](#_Toc216688682)

[9.0 Conclusion 49](#_Toc216688683)

[10.0 Reflection 52](#_Toc216688684)

[Appendices 53](#_Toc216688685)

[Coding editions 53](#_Toc216688686)

[First edition just graph of results from excel file 53](#_Toc216688687)

[2nd edition 56](#_Toc216688688)

[3rd edition 61](#_Toc216688689)

[4th edition 67](#_Toc216688690)

[Graphs with 2000 rows of data 69](#_Toc216688691)

[Graphs with 10,000 rows of data 71](#_Toc216688692)

[graphs with 20,000 rows of data 73](#_Toc216688693)

[Random forest 75](#_Toc216688694)

[XGboost 76](#_Toc216688695)

[SHAP 77](#_Toc216688696)

[Reflection 78](#_Toc216688697)

[References 79](#_Toc216688698)

[Bibliography 92](#_Toc216688699)

# Abstract

This project evaluates machine learning models for predicting UK house prices by merging Price Paid Data (Land Registry) with the House Price Index (UK government index tracking price trends) to capture market trends. Four models used were Linear Regression, Decision Tree, Random Forest, and XGBoost were trained on 40,000 property records. Results show XGBoost achieved the highest accuracy (R²: 0.816, MAE: £59,324), yet this error margin represents over 20% of the average UK house price, highlighting significant practical limitations. The study critically finds that model performance is intrinsically linked to socio-economic features like postcode, raising substantial ethical concerns. Using SHAP analysis, the project addresses the 'black-box' dilemma, concluding that while ensemble methods offer superior predictive power, their real world application requires rigorous ethical safeguards to mitigate the risk of automating spatial inequality.

# 1.0 Introduction

As of August 2025, the index is at 104.6 and the average property price in the UK is £272,995. The cost of real estate has increased by 3.0% over the last year and by 0.8% over the preceding month (Data.gov.uk, 2025). These two graphs below show how houses being built have slowed down, this in hand has caused house prices to rise.

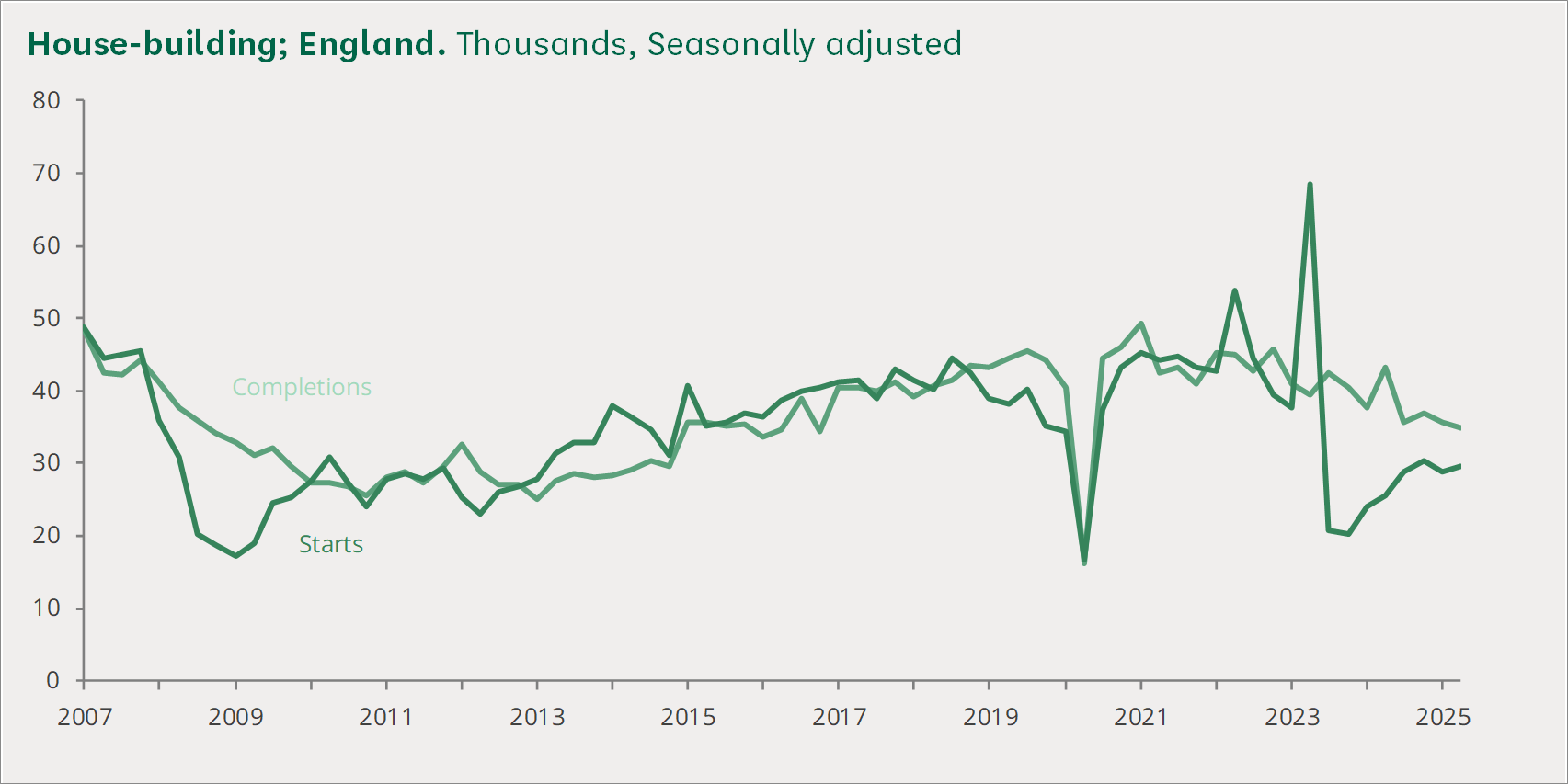


Figure - (KEEP and Keep, 2025) - house building

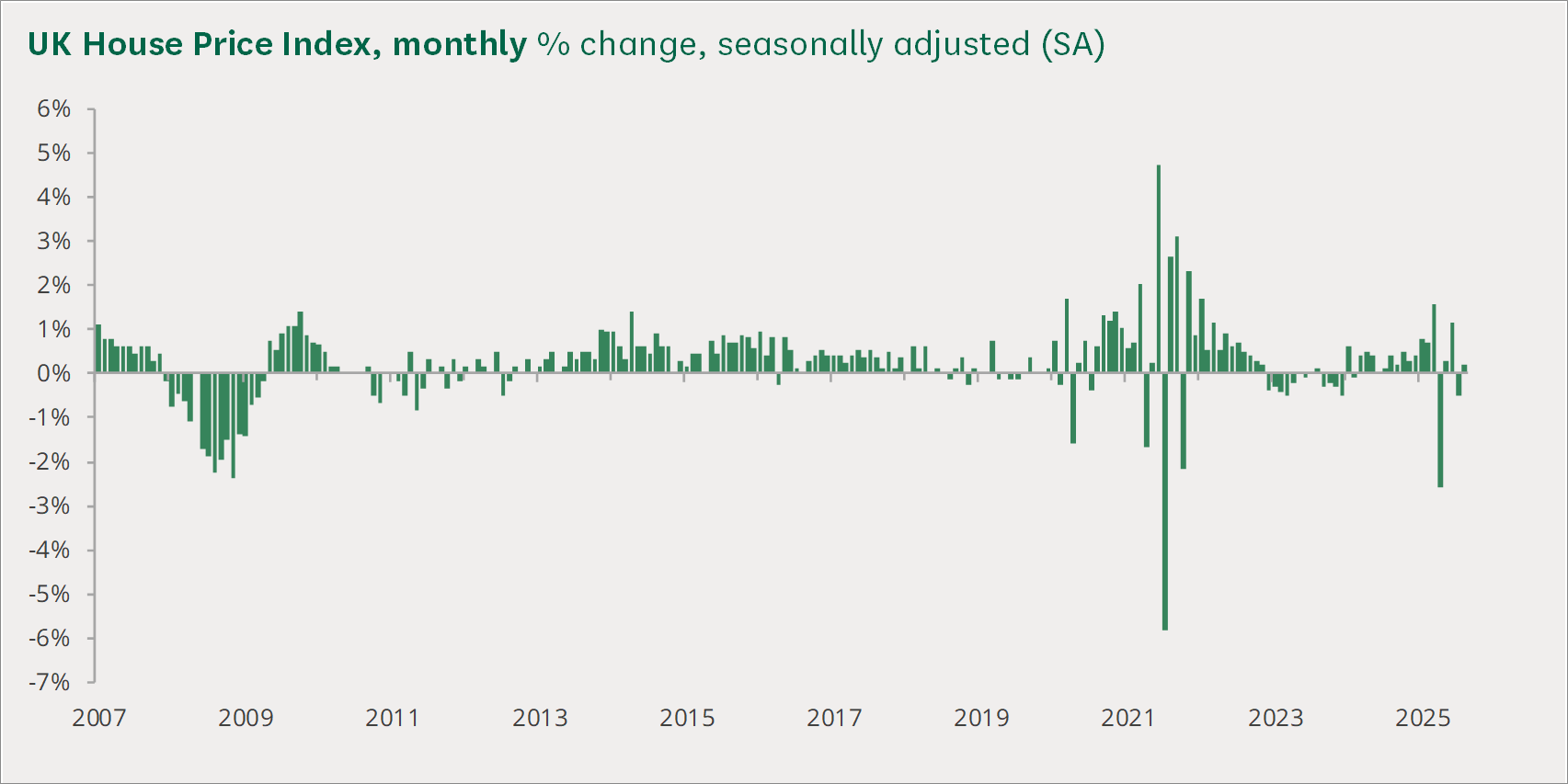


Figure - (KEEP and Keep, 2025) - house price index

House price models are essential because they reveal patterns that cannot be identified through traditional valuation alone, it can drive transparency for various stakeholders and increased efficiency and accuracy (Sharma, Harsora and Ogunleye, 2024). The traditional method of valuation for houses was comparing to recently sold prices, the cost of build the property or residual method which compares land prices this is used for development projects (FRICS, 2020). These traditional method can be useful for picking out the features and finding underlying problems such as loft conversion, as for valuation tools extra square footage usually isn’t added (YallaValue, 2025). However the time consuming process can take weeks or days also requiring an employee to visit the property meaning businesses hiring a employee just for valuation purposes (Nieberg, 2016).

The aim of this project is to evaluate how different valuation approaches perform and determine which method produces the most realistic and reliable outcomes. My aim is to find out how accurate valuation tools are and the effect it can have on the housing market and people looking to enter, it also aims to analysis the ethical problems that can caused by these tools. And compare results against existing research.

Predicting house prices is a difficult task in today's world, with rapid market changes it can effects predicting accuracy, inflation and interest rates set by the government can also have a large impact on the real estate market, however ML models can handle large datasets to increase the accuracy of the predictions, and can handle the changing interest rates and prices even hourly (Segal, 2021).

Accuracy in models is a challenging problem due to the non linear relationship between buyer behaviour and economic conditions, many studies rely on a single source for a dataset. This limits the realism and reproducibility of their models (White and Papastamos, 2022).

ML can improve valuation with its ability to handle large datasets with consistency and speed (Khan, 2024). Human bias is reduced as the large dataset can mitigate input risk, however is the dataset is corrupted or unclean with bias include against individuals or monitories it can cause ethical issues.

# 1.1 Research questions

## 1.1.1 RQ1. How accurately can machine learning models predict residential property prices using structured housing and market data?

This question aims to evaluate model performance across multiple algorithms, comparing traditional approaches such as Linear Regression with tree based methods including Decision Tree, Random Forest, and XGBoost. Metrics such as RMSE, MAE and R² are used to determine how well each model generalises to unseen property sales.

## 1.1.2 RQ2. How can model transparency and ethical considerations be improved when using “black-box” machine learning methods for property price prediction?

Tree ensembles and boosted models often lack inherent interpretability, which can raise ethical issues when these models are used in housing valuation, lending decisions or policy analysis. This question examines the limitations of black box models and evaluates how interpretability tools like SHAP can increase accountability in the prediction process.

# 2.0 Literature Review

The three primary issues of this section's analysis of the literature on home price prediction include classic valuation methods, real estate machine learning techniques, and the particular difficulties posed by UK housing date. The suitability of modern machine learning techniques for property is then assessed, including XGBoost, Random Forest, and Linear Regression. In order to view this study within the broader area, compare studies from various global markets are examined.

A large gap into other studies as they compare house prices and areas without considering the HPI index which shows a true overview of house prices of that period. few reports compare simple models against tree models which is where the justification of this research project is needed.

## 2.1 Hedonic pricing models

Hedonic pricing models are statistical models that assume a products price based features such as size and number of bedrooms (Hargrave, 2025). Traditional statistical models rely on fixed assumptions which limits their ability to represent the complexity of modern housing data and this reduces their reliability. (Bzdok, Altman and Krzywinski, 2018).

## 2.2 Machine learning models in real estate

Many studies achieve an accurate model with predictions , however they fail to see the trends with regional prices, this causes an overlook of long term market trends. The studies show that linear models are still effective for predicting prices, however ensemble methods outperform them by capturing the complexity of modern property markets (Fu, 2024).

House price prediction models use machine learning to help improve the accuracy they focus on location , square footage and number of rooms, an example of machine learning model that can be used is random forest, it combines the prediction of multiple decision tree to create a prediction (Truong et al., 2020).

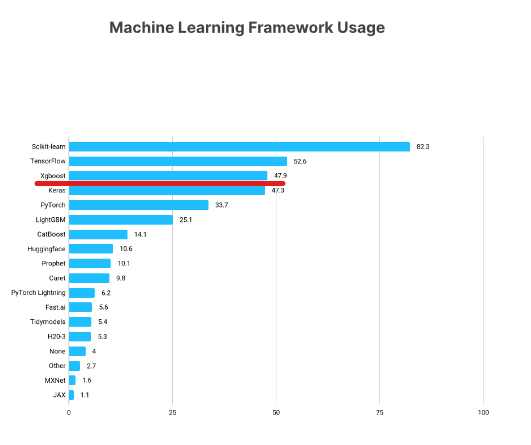


Figure - (Bekhruz Tuychiev, 2019) - shows the usage of models shows that linear regression is most popular

## 2.3 XGBoost



Figure - XGBoost - (Kavlakoglu and Russi, 2024)

XGBoost stands for extreme gradient boosting , it uses parallel tree boosting it is the leading machine learning library for regression, classification and ranking problems (NVIDIA Data Science Glossary, 2019).

The main advantage of this includes the high performance, speed and flexibility it also includes build in features for handling missing data and preventing overfitting data, however it is a high memory consumption model, it has many hyperparameters and finding the right combination can be time consuming (Krayonnz.com, 2025).

## 2.4 Random forest

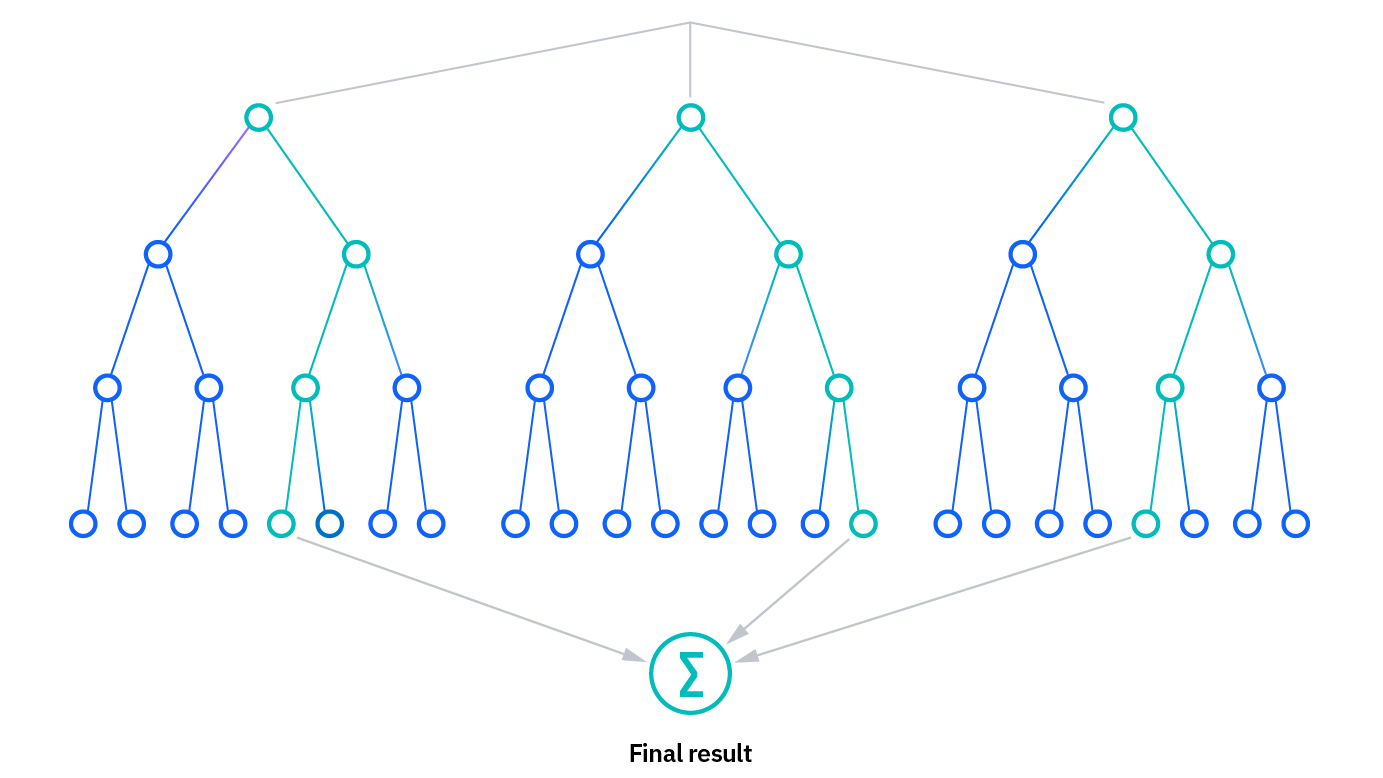


Figure - (Kavlakoglu, 2021) - decision tree example!

The benefit of using this method is the reduce risk of overfitting , it will lower the variance and prediction error. It is made to handle regression tasks with a high degree of accuracy so perfect for predicting house prices, however the accuracy depends on the size of the data set which can become time consuming, it can require more resources to process that data and the overall complexity (Kavlakoglu, 2021). An example of this would be postcodes unless they are encoded appropriately.

## 2.5 Linear regression

Another method many other papers use is linear regression it is supervised learning-based machine learning algorithm. It carries out a regression job. Regression uses independent variables to model a goal prediction value. It is mostly utilised for predicting and determining the link between variables (Maloku, 2024).

An example of running linear regression is.

from sklearn.linear\_model import LinearRegression

regressor = LinearRegression()

regressor.fit(X\_train, y\_train)

(Yao, 2024)

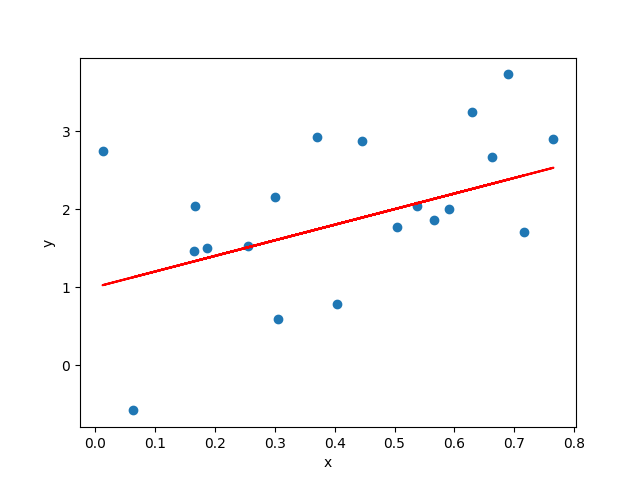


Figure (Yao, 2024) - linear regression example.

The simplicity to implement and interpret the outcomes makes it efficient, however outliers can have a huge impact on results , linear regression also assumes a linear relationship between dependent and independent variables (GeeksforGeeks, 2020).

2.6 Challenges in data

UK market is unusual with extreme regional variation with the rural Scotland to the built-up cities, house prices can vary in price largely with northern prices being close to £300,000 more in value. This creates a challenge for prediction models because postcode variation introduces structural inequality that the model must learn to handle (rick, 2023).

The largest issue is with finding datasets that have uk market house data, the uk price paid database doesn’t include number of rooms , size of the house or many other attributes. This missing data causes it hard for ML to have enough data to predict (Harel and Harpaz, 2024).

Other issues include outliers such as luxury homes and apartments which is included also in the datasets, cleaning addresses and encoding postcodes is a large part of increasing the accuracy of predicting model, postcodes can show an income area level, demographics and local infrastructure and how they effect prices (Code B website, 2024).

## 2.7 Other studies to compare against

### 2.7.1 Kuala Lampur

An example of a study to review is “Advanced Machine Learning Algorithms for House Price Prediction: Case Study in Kuala Lumpur”, this study focuses on a Malaysia housing dataset from Kaggle with around 21,995 rows of data. It provided locations, number of bedrooms , bathrooms, car parks , size , furnishing status , property type and distance to shop , school , hospital and stations(Abdul-Rahman et al., 2021).

A graph of a graph

AI-generated content may be incorrect.

Figure - (Abdul-Rahman et al., 2021) - actual price against predicted.

The models the study compared was multiple linear regression , ridge regression and light GBM, they evaluated the mean absolute error, root mean square error and adjusted r squared(Abdul-Rahman et al., 2021). Result of the tests was XGBoost performed best MAE = 0.148, RMSE = 0.197, adjusted R² = 0.911. The conclusion for the overall report was that the model was effective for Kuala Lumpur, but however for future work should include more attributes and expand beyond KL.

This study is focused on modern ML methods which help the accuracy, the paper has a clear structure and uses two different data sets, google maps was also used to help expand the attributes rather than basic features. However, the main disadvantage of this study is that it uses a simple train and test split, this can cause learning patterns from future data sets without the author realising. The geographical concern of the report being limited to Kuala lumpur, so the results don’t really show how well the model would work in other regions with different market behaviour. It is hard to understand the reliability as the authors don’t include any uncertainty estimates or interoperability, it is hard to understand why the model makes certain decisions (Abdul-Rahman et al., 2021).

Many studies use data sets from other countries, but the focus of this report is on the UK housing market, the data set will include many attributes to help with predictions and HPI index, making the accuracy increase. I want to focus on a simple model that will focus on accuracy and transparency to address the ethical concern regarding prediction models.

### 2.7.2 2nd study - AI in Real Estate: Forecasting House Prices with Advanced Machine Learning Models

Another study to compare against is ‘AI in Real Estate: Forecasting House Prices with Advanced Machine Learning Models’ , the authors aim to use 7 machine learning algorithms with linear regression, decision tree and XGBoost. The goal is to predict house prices based on a dataset from Kaggle.  
The dataset covered 1,460 houses and 81 features such as area, street and rooms. However before modelling, they performed exploratory data analysis, this showed a trend that quality is associated with higher prices.

Their results.

A table with numbers and letters

AI-generated content may be incorrect.

Figure - 2nd study reuslts - (Cui, Liu and Ma, 2024)

The results show that xgboost performed the best this correlation between kuala lapur shows that XGBoost is a more accurate and simpler model which can fit all prediction models, however the low dataset and the dataset from kaggle will have a large impact on bias cause them to become larger.

(Cui, Liu and Ma, 2024)

# 3.0 Data and Methodology

This section goes through the research design , data sources , data cleaning used and the techniques employed. This follows a description of the datasets used, followed by the data cleaning and feature engineering process. The implementation of the four predictive models is then shown, the evaluation of the results is then explained, the methodology ensures a reproducibility and aligns with the best machine learning models.

## 3.1 overview

The dataset comprises 79% of residential transactions from the Land Registry Price Paid Data, merged with the UK House Price Index to incorporate temporal market trends. Data cleaning involved SQL-based preprocessing to handle missing values and outliers, ensuring robustness for model training. This new linked dataset details 5,732,838 transactions in England and Wales between 2011 and 2019, along with each property's total floor area and the number of habitable rooms (Chi et al., 2019).

A large issue with UK data is that certain regions include more data, this can be due to housing supply and demand with urban areas having access to a larger number of supply, income inequality and land availability all contribute to data access for regions (Drayton, Levell and Sturrock, 2024). This raises a risk of domination in training data for urban data, another large issue is that the large change in patterns from old data as issues from recession and interest rates cause differences in house prices.

it will cover the whole of the Uk to help with complexity I will put a capacity of 40,000 rows which exceeds the other study complete for Kuala lumpar. I will also combine the Uk house price index this is used to calculate the overall market trends, if I used one model for training a jump in house prices may be unexplainable but the index can show the market trends for that year aligned with price (Bedford, 2020).

Dataset was capped at 40,000 rows this cause a large decrease in accuracy, however laptop limits on memory availability, this meant a trade off from accuracy however the feasibility of the project depended on it, in future having a larger memory usage would mean training limits could be longer helping increase the accuracy of the predictions.

This project addresses the gap in current research , by merging the datasets with the HPI index the project will be able to capture both transactional and market rend data, this reflects the real-world conditions prediction models will be able to handle such as covid and recessions.

## 3.2 Data sets

<https://reshare.ukdataservice.ac.uk/854942>

<https://www.gov.uk/government/statistical-data-sets/uk-house-price-index-data-downloads-may-2025>

A screenshot of a web page

AI-generated content may be incorrect.

Figure - (Ons.gov.uk, 2025) - Uk house price index.

A screenshot of a website

AI-generated content may be incorrect.

Figure 10 - prop.csv data - (Chi et al., 2019)

### 3.3 Pipeline Diagram

*Figure 11 - pipeline diagram explain process*

### 3.4 fields

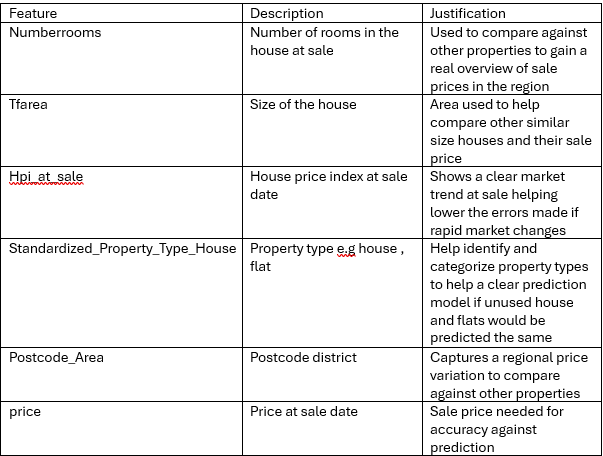


Figure - key fields used in the model to make the predictions.

### 3.5 Clean data

A white sheet with black text

AI-generated content may be incorrect.

Figure - cleaning steps and purpose.

(Thealliance.ai, 2025).

A screenshot of a computer program

AI-generated content may be incorrect.

Figure - data cleaning sql used by the dataset - (Chi et al., 2019)

A screen shot of a computer program

AI-generated content may be incorrect.

Figure - clean hpi data - commentsi in code to explain steps used.

A screen shot of a computer code

AI-generated content may be incorrect.

Figure - clean hpi data - part 2.

A screen shot of a computer program

AI-generated content may be incorrect.

Figure - clean property data - comments in code to explain steps used.

## 4.0 Model outputs

### 4.1 Mathematical model 1: Linear regression

A diagram of a graph

AI-generated content may be incorrect.

Figure - linear regression model graph.

### 4.2 Model 2 : Random forest

A graph showing a line of orange dots

AI-generated content may be incorrect.

Figure – Random forest graph model.

## 4.3 Model 3 : decision tree model

A green dotted line with a red line

AI-generated content may be incorrect.

Figure - Decision tree model.

## 4.4 Model 4 : XGBoost

A purple line with a red line

AI-generated content may be incorrect.

Figure - XGBoost results visualised.

## 5.0 Evaluation

### 5.1 Train Test Split

The 80/20 split was used as it allows the training data to have enough knowledge to perform predictions accurately, it can mitigate bias and variance (ifttt-user, 2023). This split means that with 40,000 rows that 32,000 rows of data will be used for training and 8,000 rows for testing. However the larger the dataset can help enhance the accuracy of the predictions made by the data. As it can help add more data for training and testing (or, 2020).

A computer screen shot of a black background

AI-generated content may be incorrect.

Figure - data used for train test split.

However the model stability may vary across different slices of data , this can create optimistic results depending on how the data is arranged, housing markets depend heavily on region and price band so one split may not represent the full distribution, the way to tackle this issue could use cross validation or multiple random splits this would greatly improve the reliability of the models performance (GeeksforGeeks, 2017).

### 5.2 Results

R2 (R squared), this is a way to review a models accuracy, can be misleading because it reflects how well a model fits past data rather than its real predictive strength, rather than true predictive powers, it also cannot show weather the predictions are bias, it also doesn’t show a reliability as you can get a high R squared for a poorly fitted model (Fernando, 2025). Areas with high variability can also misshape the predictive powers of the models, MAE and RMSE help give a clear view of the predictive powers and make it easier for buyers to be able to read.

A white sheet with numbers and black text

AI-generated content may be incorrect.

Figure 23 - reuslts from all models used.

While XGBoost's R² of 0.82 aligns with expectations from the literature (e.g., Cui et al., 2024), its £59k MAE starkly contrasts with their sub-£25k error, exposing the unique volatility and structural inequality of the UK market as the primary driver of predictive uncertainty.

### 5.3 Top 10 featured coefficients

A table with numbers and lines

AI-generated content may be incorrect.

Figure - Top 10 coefficients.

### 5.4 Size of dataset

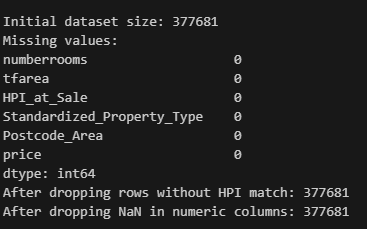
****

Figure - size of dataset used.

# 6.0 Discussion

## 6.1 introduction

Linear regression was chosen for a baseline result and interpretation, decision tree was selected to show the opposite of a nonlinear learning model, whereas random forest was later selected to reduce the overfitting that decision tree was experiencing (Sharma, 2020). XGBoost was used it can help remove outliers and other data challenges (GeeksforGeeks, 2021); however it also raises the black box problem of transparency (Lazyprogrammer.me, 2023).

## 6.2 evaluation metrics

RMSE ( root mean squared error) measures that average magnitude of the errors. The mean absolute error (MAE) it measures the average of the errors without considering direction , it is robust to outliers. R^2 score is a representation of the variance between prediction and real , a 0 means that you could predict that result yourself and a 1 shows a prefect prediction.

## 6.3 model selection reasoning

XGBoost was selected to handle the non linear relationship, comparing against decision tree model it tackles the overfitting through regularisation, it also performs efficiently on larger datasets (he /him, 2023).  
Linear regression performed well as it is a simple tool that can fit many uses, it efficiency to find relationships in data allows it to be easy to implement (KAVITA, 2021). It is a versatile tool that can handle the postcode areas, decision trees overfit training data and can capture noise in the data rather than underlying patterns (Kara, 2024).

## 6.4 model limitations

Linear regression limitations reflect the theoretical expectations that uk prices are ruled and shaped by interacting regional and temporal pressures, this limitation requires non linear models to capture the true market structure (Wan et al., 2025).  
random forest performed reasonably well but struggled with variance, likely due to its reliance on bootstrapped sampling, which can limit the model’s ability to capture fine grained patterns in highly heterogeneous datasets like UK housing (Iwaniuk et al., 2025).  
whilst XGBoost performed significantly well , its reliance on postcode as shown by SHAP risk automating systematic inequality, this becomes a social issue from ML models as it can increase bias from past historical data with regional disparities, accuracy must be balanced with a fairness , this trade off isn’t explored in many other studies.

## 6.5 external factors

External problems such as covid, which increased house prices by 20% in 2020 have changed the way prices traditional valuation models predict prices, whereas a ML can handle market changes and adapt to new situation and predict outcomes more accurately, giving a risk free knowledge to potential sellers and buyers (The Impact of Large Language Models in Finance: Towards Trustworthy Adoption, n.d.).

A graph of different colored lines

AI-generated content may be incorrect.

Figure - how house prices changed over time - (Lloyds Banking Group, 2023)

## 6.6 results overview

The results show that xgboost regression is the most accurate at predicting prices as shown by the R2 score being at 8 which is the closest to 1, decision tree shows a very inaccurate result being nearly 0.6 away from 1. Going through the RMSE and MAE results show that xgboost regression also has a lowest MAE meaning that its average prediction error is the smallest.

## 6.7 real world impact

A MAE error being off can mean that sellers could be out by £68k if they used linear regression, this mean that they have lost out on a large value from the prediction being wrong, this is a real world impact of prediction models and it can have devastating effects. The model safest is random forest for MAE however still being off by £65k which is a large sum of money to a large proportion of the population. This is shown by only 12% of the UK earning over £60,000 a year so if the prediction was wrong it can be wrong by over a persons yearly wage (Institute for Fiscal Studies, 2025).

## 6.8 XGBoost performance detail

Xgboost outperformed all other models however the persistence of a £59,000 rmse demonstrates that model superiority does not translate into practical valuation reliability, this increase is inline with the two other case studies, this accuracy increase can save buyers and sellers thousands of pounds. This comes from the merging the predictive models such as decision trees, this increases the overall accuracy it works really well with numerical data such as prices of properties. It comes with built in functions able to manage outliers and imbalanced data (Hossein Ashtari, 2024).  
However whilst XGBoost performed the greatest, It’s MAE showed a lack of 20% of the current UK house price ( data.gov.uk , 2025). For an induvial this questions the models predictiveness quality. Kuala lumpur suggested that a portion of predictive error is not a model failure, but a large regional inequality within the data itself this is a complex problem that advance algorithms still struggle with (Kaustubh Chakradeo et al., 2025).

## 6.9 regional inequality

The model showed a regional inequality and shown by the postcode impact, this is caused by income differences in areas such and London and up north where wage differences can reach nearly £8,400 more for London wages (Grace, 2022). The model strongly weights towards expensive regions this can cause predictions to be higher for northern regions and can be lowered for southern regions closer to London, the way to increase the accuracy and reliability of the predictions would need to increase the data used, however this is restricted to availability of data available which includes rooms and size of the house.  
Postcode the largest predictive quality of the models reflects the socio economic pattern rather than pure market behaviour, this shows the model learns structural inequality itself which is an issues not followed by previous studies however it is essential for interpretation of results and accuracy (Cribb, Wernham and Xu, 2023).

## 6.10 feature considerations

Rooms size and postcode are key features needed to make predictions as they are standard features, however increasing the reliability of the predictions adding in the rest of the data such as epc rating, bathrooms and school distance, however the trade of comes from the slower computing time and larger need for cleaning causing the complexity of the project to increase significantly. With the time constraints of 6 week the scope of the project was brought down to account for this.

## 6.11 comparison with other studies

Kuala Lumpur showed a difference with eXtreme Gradient Boosting being the most accurate which is a gradient boosted decision tree (Kavlakoglu and Russi, 2024). This combines decision trees and boosting to create and accurate predicting model.  
my results differ slightly from the other studies because uk housing data has far higher regional variation and postcode inequality than the kuala lumpur and kaggle datasets they used, so the models behave differently when the data is more complex.  
The result difference compared to the existing studies arise from a deeper force from within the uk market, This comes from regional inequality, volatility and postcode driven value patterns (Hilton, 2024). The merged data set along with shap reveals a clear view of these forces compared to kuala lampur and other studies.

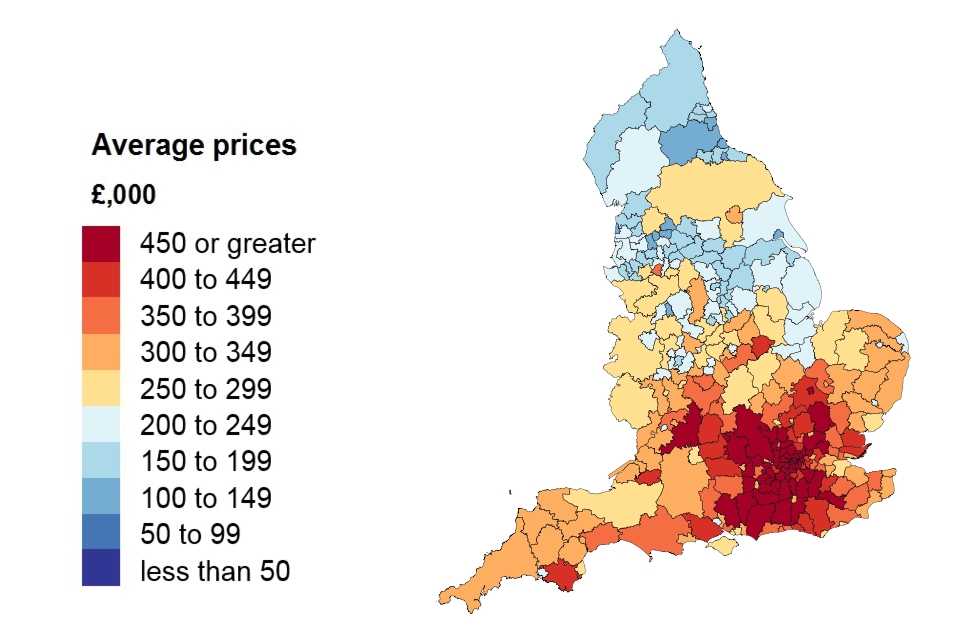


Figure - average house price per region - (HM Land Registry, 2023)

## 6.12 coefficient interpretation

Coefficients indicate predicted value such as a positive coefficient can mean that postcode is associated with a higher predicted value;  
Results such as Guildford region with a coefficient of £438,466.89 shows that having this postcode will increase predicted value by over £438k compared to baseline category, the largest negative postcode was Sheffield having a - £313,113.70 coefficient causing values to be predicted less in that area compared to the baseline. These can be affected by real world things such as GU being a London commuter belt and has high property prices compared to S1 postcode which covers more affordable properties and lower prices compared to London regions.

# 7.0 Shap – Interpretation of model results

A condition called the black box problem considers the transparency with ML models, the blur comes from the complex network architecture, massive and diverse training data and difficulty to understand outputs (Vera and Vaughan, 2023).

A screen shot of a graph

AI-generated content may be incorrect.

*Figure 28 - shap model graph.*

A graph of a graph with blue and white text

AI-generated content may be incorrect.

*Figure 29 - shap model for MAE.*

SHAP ( shapley additive explanations) was used to interpret XGBoost model, it can offer explanations to help build trust and validating the model behaviour, it can show how each feature impacts the predictions. It is useful for stakeholders and end users to easy-to-read graphs. It can help solve part of the black box problem associated with many ML driven financial decision-making tools (Coralogix, 2025). The results show that postcode area and HPI index were strongest drivers of predicted prices which aligns with other studies.

Whilst SHAP provides crucial interpretability, it unveils an ethical dilemma; accuracy depends on a learning pattern from features such as postcode. This creates a tension between predictive performance and fairness. The most accurate model may have to be problematic ethically. However a true ethical project deploys a fairness aware modelling techniques that constrain discriminatory outcomes (Lundberg and Lee, 2017).

# 8.0 Ethics and Limitations

In this research, the main limitations of the dataset is that the data is 5 years old, which lowers the prediction accuracy for any house up for sale after those dates, another main limitation is the technology as was not able to handle the whole dataset as it contained over 5 million sales.

The pre 2020 dataset ignores the post pandemic market shock, the models failure to capture recent volatility questions the predictive quality of applying such models to todays market without using newer data.

A major limitation is that the model treats postcode as a purely predictive feature, but postcode also reflects social and economic factors, so the model risks learning structural inequalities rather than neutral price patterns.

The dataset was so large that bias should have been mitigated by the large and different number of areas , this was shown by the prices being accurate to areas and showing London areas with larger prices compared to up north where the prices were less.

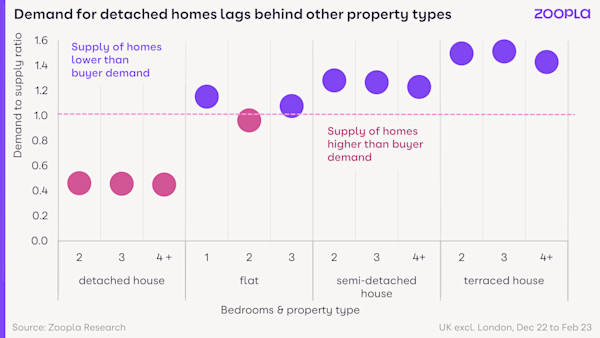


Figure - demand for house types - (Zoopla.co.uk, 2023)

Unusual factors which can effect the predictive quality of the models can come from factors such as leasehold vs freehold, this comes from freehold owning the building and the land compared to leasehold where you own the property for a set amount of time but not the land, this can effect prices as buyers will pay more for a freehold and demand for a leasehold is less (MaPS, 2025). Another factor to consider is that terraced, semi detached vs detached can also show an increase in demand and supply which can effect house prices greatly (Mason-Apps, 2022).

This project was a student built academic project, this is not for commercial use, however models being used in other markets can cause algorithmic pricing risk, this is caused when many algorithms will depict market prices based on a goal to cause price fixing, this is done when rival businesses agree to set prices rather than independently (Pricing algorithms Economic working paper on the use of algorithms to facilitate collusion and personalised pricing 8 October 2018 CMA94, n.d.).

An example of algorithmic pricing is Ubers surge pricing, this causes prices to fall, and rise based on factors such a supply and demand , location. This can happen hourly or even minutes apart (Uber, 2025). An example of price fixing would be The US Justice Department is presently suing RealPage on allegations that their rent-setting software, which suggests rent prices by combining public and private data, enabled price manipulation (Spira, 2025).

The project will have to be transparent and fair use with its data and predictions, if used incorrectly house prices could be fixed by many ML models having their parameters set this can cause inequality with social hierarchy and could even cause recessions (Arxiv.org, 2023).

AVM ( automated valuations models) are powered by machine learning and ai to revolution data driven valuations based on insights and market conditions, it increase the speed and efficiency and can even help reduced bias and be more accurate (Hayes, 2025). However, Training data must be cleaned carefully because uncorrected bias becomes embedded into the predictions, AVM struggle to account for subjective or specific property conditions and does not take into account features and specific upgrades such as a busy road nearby or a good view (Apex27 - Free Online Estate Agent CRM Software, 2022).

Another ethical risk involves the risk of algorithmic pricing this is involved when algorithms will collude with other algorithms to raise or lower prices as determined by companies for financial gain and benefit (Spann et al., 2025).

Although my project is for academic purposes, if I go to scale to deploy I would have to consider ethical laws and also the limitations on the datasets being used and the accuracy of the data. As the goal was on transparency of data I would have to be open about only using data of willing participant.

# 9.0 Conclusion

The aim of the project was Predicting House Prices Using ML Models, the project was completed successfully and the iterative development process strengthened the final approach. The most accurate model being xgboost, this could be due to the simplicity and efficient way of predicting (Yadav, 2024).

Future ways improve the project involves increasing the size of the dataset with brand new data, which is up to data, this will increase the accuracy and could involve predicting future trends into the future. This will be used to help with real estate and personal sellers know when is best to buy or sell their house (Zhang, 2025).

Compared to both studies thedecision tree shows a large difference with it being very inaccurate, comparing against the 2nd study linear regression MAE came out at 88,000 with this project coming out at 66,000 that shows a 22,000 improvement, however the r^2 score for their study came out higher showing a more accurate model.

Using SHAP allowed the model’s predictions to be interpreted and validated beyond pure accuracy metrics this also helped with the black box concern of models by improving the transparency and reducing the ethical concern many users and stakeholders have (Lomash Bhuva, 2025). This answered the 2nd research question with the 1st being answered by the performance of XGBoost whilst still being below the accuracy of the case studies it showed a significant increase in performance compared to linear regression.

Time constraints on the project caused scope of the project to be brought back , the goal of a further implementation would include user inputs of their home features to get a accurate prediction of a potential sale price.

Another future adaptation would be to develop an app that users could input their data into and get a predicted house price back this would help people be able to get a quicker and more efficient way of selling their home (Brown, 2025).

The potential for the usefulness of this project for real world buyers can include insights into prediction of house prices, they will be protected against market manipulation such as real estate agents or buyers under valuing properties. Business can use these tools to gain and insight onto how areas have grown overtime and they can use this to grow businesses. However this could cause postcode inequalities with areas on decline shown clearly causing business to move away.

A student-built systems lacks the technology and data constraints that a company is able to access, Rightmove and sites such as property API cost monthly to access databases and have access to new and updated data with more features. A laptop with 8 gb of a ram isn’t able to handle the million of rows of data, if access to technology with more memory the constraints could allow for larger dataset which is able to access all regions and map a clear predictive model. This advancement will mean mortgage companies will be able to use this tool reliability.

Reports on predictive models for the UK market are sparce and expensive with simple home buyers not able to access tools without costing a monthly subscription, this tool was built for a simple accurate tool home buyers could use to rely on a valuation.

To increase the quality of the project would require adding in all the attributes such as crime rate, bathrooms , garden size , distance to school, combine that with ML text extraction feature to extract text to property description can help increase the reliability of the project, also use other models to hopefully see the increase of the R^2 score to closer to 1 and lower the MAE to less than £20k would be a successful project. Lastly using hyper parameter tuning can optimise the variables set and maximize its performance, such as using grid search and random search would be used to complete this (Amazon Web Services, Inc., 2025). These would be added with new computing systems and power, and a new up to date dataset.

# 10.0 Reflection

A full reflective account of the project’s development, challenges, and personal learning is provided in Appendix under reflection..

# Appendices

## Coding editions

## First edition just graph of results from excel file

A screenshot of a computer

AI-generated content may be incorrect.

Figure - 500.xlsx - house price sale list

A screenshot of a computer

AI-generated content may be incorrect.

Figure - met.xlsx - property stat avg

A screen shot of a computer program

AI-generated content may be incorrect.

Figure - libarys used

A screen shot of a computer program

AI-generated content may be incorrect.

Figure - load both files merge deed date and time perdod

A computer screen shot of code

AI-generated content may be incorrect.

Figure - sort the data and create means for the data such as price paid and great Britain average

A screen shot of a computer code

AI-generated content may be incorrect.

Figure - plot the data using the data as axis

A screen shot of a graph

AI-generated content may be incorrect.

Figure - graph to show the results

## 2nd edition

A screen shot of a computer

AI-generated content may be incorrect.

Figure - prop.csv - list of property data

A screenshot of a table

AI-generated content may be incorrect.

Figure - prop.csv - number of rooms and size

A screen shot of a computer program

AI-generated content may be incorrect.

Figure - load the excel files

A computer screen shot of a program code

AI-generated content may be incorrect.

Figure - clean the data removing £ signs and empty rows

A screen shot of a computer code

AI-generated content may be incorrect.

Figure - clean prop.csv data

A screen shot of a computer

AI-generated content may be incorrect.

Figure - put property types into categories

A screen shot of a computer program

AI-generated content may be incorrect.

Figure - combine the dates for hpi and sale date

A screen shot of a computer program

AI-generated content may be incorrect.

Figure - display property sale date and price against hpi at sale

A black and white screen

AI-generated content may be incorrect.

Figure - results of first 10 rows

## 3rd edition

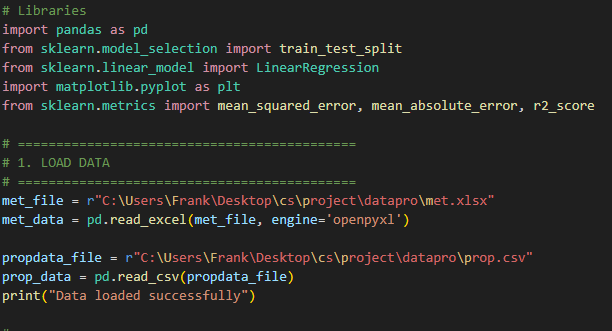


Figure - library needed



Figure - clean the data from hpi

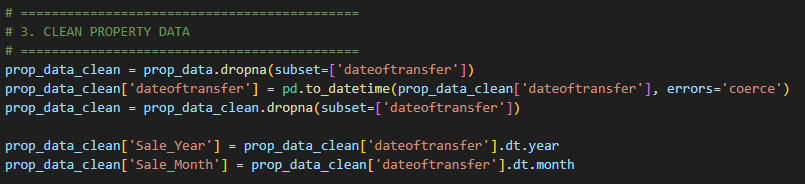


Figure - clean property data



Figure - put property types into catorgories

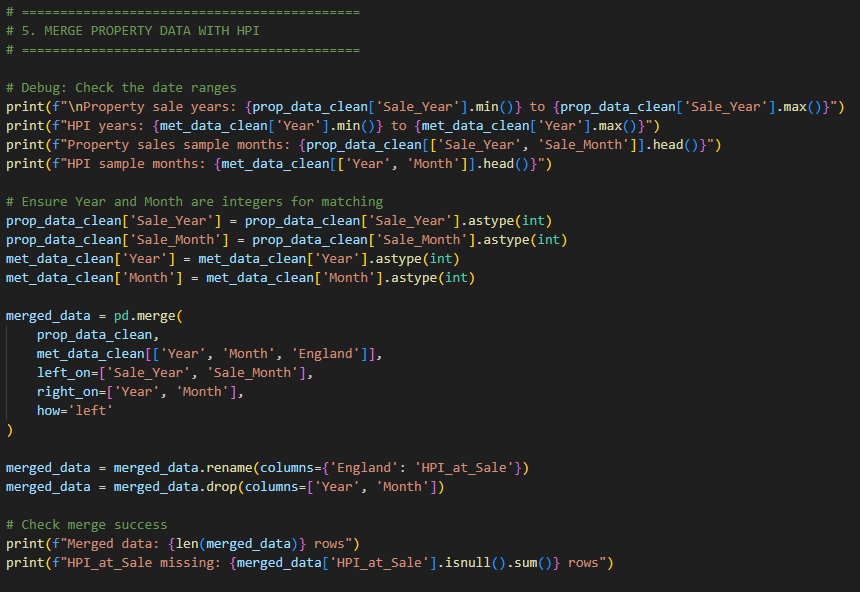


Figure - merge sale date against hpi at sale

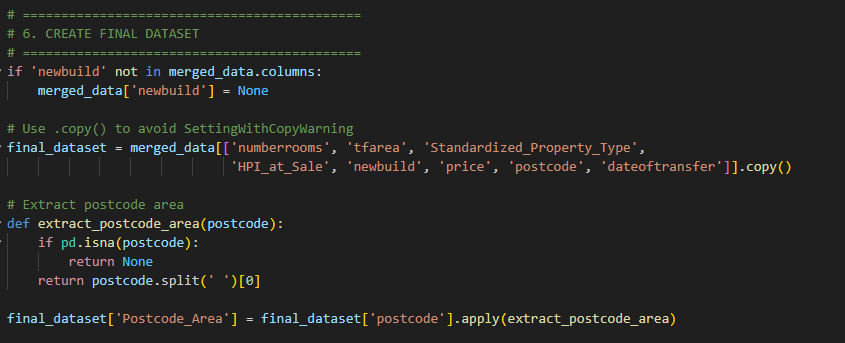


Figure - create a final dataset to use for training

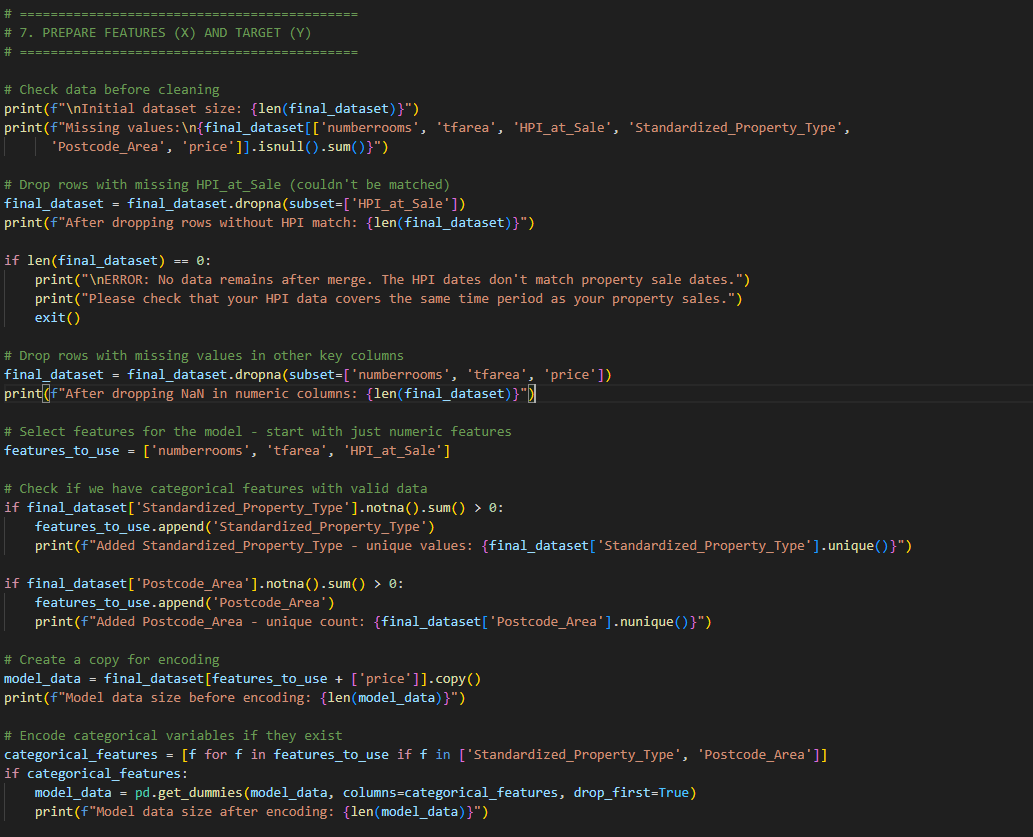


Figure - get the features needed to train x and y

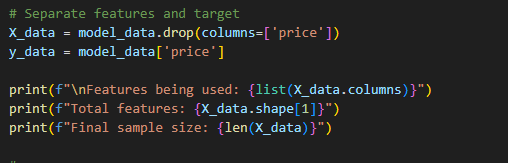


Figure - features being used

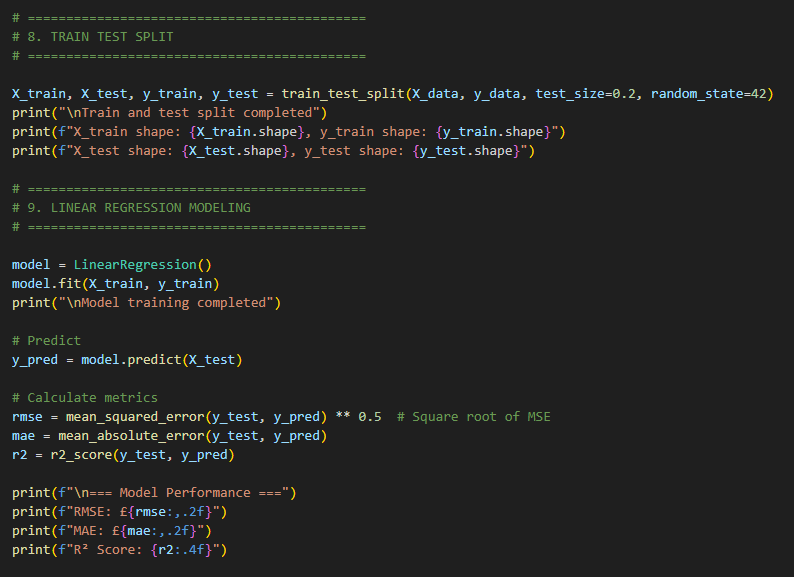


Figure - train test split of 20/80 and linear regression implementation

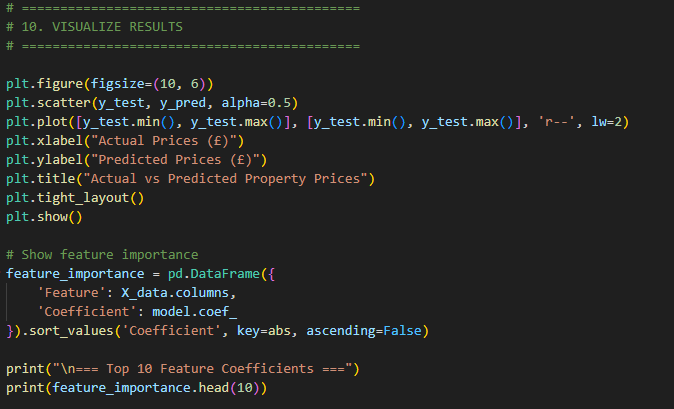


Figure - visualise the linear regression

## 4th edition



Figure - decision tree model libary

A computer screen shot of a program

AI-generated content may be incorrect.

Figure - decision tree used to test the data

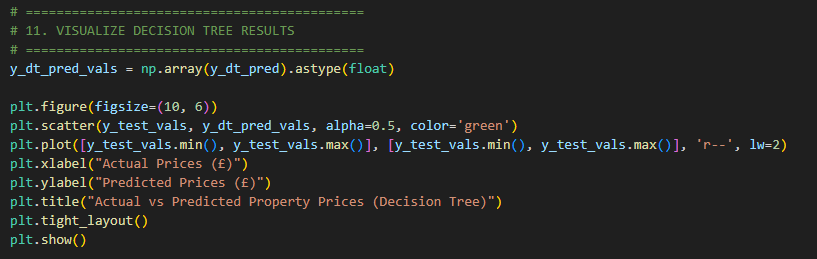


Figure - visualise the data

A computer screen with text and numbers

AI-generated content may be incorrect.

Figure - comparisons of the models

## Graphs with 2000 rows of data

A graph showing a line of blue dots

AI-generated content may be incorrect.

Figure - linear regression graph

A screen shot of a graph

AI-generated content may be incorrect.

Figure - decision tree graph

=== Top 10 Feature Coefficients ===

Feature Coefficient

25 Postcode\_Area\_TS4 -394942.205150

26 Postcode\_Area\_TS5 -309790.392464

23 Postcode\_Area\_TS1 -274600.436037

11 Postcode\_Area\_LE2 -227224.030656

12 Postcode\_Area\_LE3 -225727.921343

13 Postcode\_Area\_LE4 -209646.727763

14 Postcode\_Area\_LE5 -188378.671886

27 Postcode\_Area\_TS7 -187602.430453

24 Postcode\_Area\_TS3 -174763.752928

28 Postcode\_Area\_TS8 -124128.830915

=== Comparison of Models ===

Model RMSE MAE R� Score

0 Linear Regression 105663.442940 73394.814451 0.847304

1 Decision Tree Regressor 153050.295516 97875.714674 0.679635

## Graphs with 10,000 rows of data

A graph showing a line of blue dots

AI-generated content may be incorrect.

Figure - linear regression model

A green line with a red line

AI-generated content may be incorrect.

Figure - decision tree model

=== Top 10 Feature Coefficients ===

              Feature    Coefficient

25  Postcode\_Area\_TS4 -394942.205150

26  Postcode\_Area\_TS5 -309790.392464

23  Postcode\_Area\_TS1 -274600.436037

11  Postcode\_Area\_LE2 -227224.030656

12  Postcode\_Area\_LE3 -225727.921343

13  Postcode\_Area\_LE4 -209646.727763

14  Postcode\_Area\_LE5 -188378.671886

27  Postcode\_Area\_TS7 -187602.430453

24  Postcode\_Area\_TS3 -174763.752928

28  Postcode\_Area\_TS8 -124128.830915

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **Linear regression** | **Decision Tree Model** | **Interpretation** |
| **RMSE** (Root Mean Squared Error) | **£99,899.70** | £146,042.94 | The **average magnitude** of the prediction errors. Lower is better. |
| **MAE** (Mean Absolute Error) | **£71,659.06** | £88,324.62 | The **average absolute difference** between predicted and actual prices. Lower is better. |
| **$R^2$ Score** ($R^2$) | **0.8522** | 0.6840 | The **proportion of the variance** in house prices explained by the model. Closer to 1.0 is better. |

## graphs with 20,000 rows of data

A graph of a graph showing a line of blue dots

AI-generated content may be incorrect.

Figure - linear regression graph

A green dotted line with a red line

AI-generated content may be incorrect.

Figure - decision tree graph

=== Top 10 Feature Coefficients ===

               Feature    Coefficient

5   Postcode\_Area\_KT13  548139.347921

4   Postcode\_Area\_GU25  443143.705919

67   Postcode\_Area\_TS1 -396347.675423

56   Postcode\_Area\_SM2  361004.208824

71  Postcode\_Area\_TS13 -322948.615527

42   Postcode\_Area\_S13 -315152.705354

72  Postcode\_Area\_TS14 -301936.513461

86  Postcode\_Area\_WA13 -297111.869914

49   Postcode\_Area\_S63 -287478.449794

54   Postcode\_Area\_S80 -286214.317271

=== Comparison of Models ===

                     Model           RMSE           MAE  R� Score

0        Linear Regression  104029.960354  67267.240898  0.802110

1  Decision Tree Regressor  139277.126947  85053.389370  0.645296

## Random forest

A screenshot of a computer program

AI-generated content may be incorrect.

Figure - random forest

A computer screen shot of a program code

AI-generated content may be incorrect.

Figure - random forest graph code

## XGboost

A computer screen shot of a program code

AI-generated content may be incorrect.

Figure - xgboost code

## SHAP

A computer screen shot of text

AI-generated content may be incorrect.

Figure - shap model

## Reflection

The idea of machine learning is easy to put into practice, but there is a chance that you won't know the final result or how the model predicted it. Finding the dataset was the largest obstacle since the housing price paid database lacked parameters that were essential to the accuracy of the forecast model, such as rooms and home size. A larger technological and dataset reach would be part of another effort at this project. It was extremely difficult and required several tries to solve coding issues in order to properly combine the CSV for the time period so that the HPI index would merge; this was crucial to the project and accurate pricing. All of this strengthened coding knowledge and enhanced problem-solving abilities.

Professional growth came from time management skills increased having to stick to project deadlines to make sure the project completed on time and the quality is still increase, version control skills with keeping clear push and pull using github to keep track of version, this helped with debugging issues and helping solve problems that new code errors created. Having clear iterations of coding steps and combined with clear and well written documentation helped increase the speed and lowered the complexity of the workload.

The results by combining transactional and market trend data significantly improves predictive accuracy, shows the importance data can have on ML models, future work would include cross validation with model interpretation techniques, it would focus on social economic factors related to house prices increase the robustness and fairness of the prediction model.

# References

Abdul-Rahman, S., Zulkifley, N.H., Ibrahim, I. and Mutalib, S. (2021). Advanced Machine Learning Algorithms for House Price Prediction: Case Study in Kuala Lumpur. International Journal of Advanced Computer Science and Applications, 12(12). doi:<https://doi.org/10.14569/ijacsa.2021.0121291>.

Amazon Web Services, Inc. (2025). *What is Hyperparameter Tuning? - Hyperparameter Tuning Methods Explained - AWS*. [online] Available at: https://aws.amazon.com/what-is/hyperparameter-tuning/#:~:text=When%20you're%20training%20machine,not%20set%20by%20data%20scientists. [Accessed 25 Nov. 2025].

Apex27 - Free Online Estate Agent CRM Software. (2022). Instant House Valuations Are They Really Any Good. [online] Available at: <https://apex27.co.uk/estate-agency-blog/instant-house-valuations-are-they-really-any-good#:~:text=The%20major%20downside%20to%20automated,a%20likely%20selling%20price%20either>. [Accessed 24 Nov. 2025].

Arxiv.org. (2023). Towards Trustworthy AI: A Review of Ethical and Robust Large Language Models. [online] Available at: <https://arxiv.org/html/2407.13934v1> [Accessed 24 Nov. 2025].

Bedford, D. (2020). How Does A Recession Affect The Property Market? [online] Pettyson.co.uk. Available at: <https://www.pettyson.co.uk/about-us/our-blog/691-recession-affect-property-market#:~:text=In%20terms%20of%20the%20direct,wane%20over%20the%20coming%20months>? [Accessed 24 Nov. 2025].

Brown, P. (2025). The transformative potential of AI in real estate. [online] Lewissilkin.com. Available at: <https://www.lewissilkin.com/insights/2025/02/26/the-transformative-potential-of-ai-in-real-estate> [Accessed 21 Nov. 2025].

Bekhruz Tuychiev (2019). *Using XGBoost in Python Tutorial*. [online] Datacamp.com. Available at: https://www.datacamp.com/tutorial/xgboost-in-python?dc\_referrer=https%3A%2F%2Fwww.google.com%2F [Accessed 15 Dec. 2025].

‌

Bzdok, D., Altman, N. and Krzywinski, M. (2018). Statistics versus machine learning. Nature Methods, 15(4), pp.233–234. doi:<https://doi.org/10.1038/nmeth.4642>.

Chi, B., Dennett, A., Oléron-Evans, T. and Morphet, R. (2019). A new attribute-linked residential property price dataset for England and Wales 2011-2019 - ReShare. Ukdataservice.ac.uk. [online] doi:<https://reshare.ukdataservice.ac.uk/854942/1/tranall2011_19.csv>.

Code B website. (2024). Encoders in Machine Learning: Applications & Use Cases. [online] Available at: <https://code-b.dev/blog/encoders-machine-learning> [Accessed 24 Nov. 2025].

Cribb, J., Wernham, T. and Xu, X. (2023). Housing costs and income inequality in the UK. *Econstor (Econstor)*. [online] doi:https://doi.org/10.1920/re.ifs.2023.0288.

‌

Coralogix. (2025). *SHAP: Are Global Explanations Sufficient in Understanding Machine Learning Predictions?* [online] Available at: https://coralogix.com/ai-blog/shap-are-global-explanations-sufficient-in-understanding-machine-learning-predictions/#:~:text=Advantages%20of%20SHAP%20\*%20SHAP%20can%20be,on%20predictions%20i.e.%2C%20negative%20and%20positive%20contributions. [Accessed 4 Dec. 2025].

‌

Cui, J., Liu, Z. and Ma, Y. (2024). AI in Real Estate: Forecasting House Prices with Advanced Machine Learning Models. pp.144–148. doi:https://doi.org/10.5220/0013511000004619.

‌

Data.gov.uk. (2025). UK House Price Index. [online] Available at: <https://landregistry.data.gov.uk/app/ukhpi/> [Accessed 13 Nov. 2025].

Drayton, E., Levell, P. and Sturrock, D. (2024). Working paper The determinants of local housing supply in England. [online] Available at: <https://ifs.org.uk/sites/default/files/2024-07/WP202435-The-determinants-of-local-housing-supply-in-england.pdf>.

Fernando, J. (2025). R-Squared: Definition, Calculation, and Interpretation. [online] Investopedia. Available at: <https://www.investopedia.com/terms/r/r-squared.asp> [Accessed 24 Nov. 2025].

FRICS, J.L. (2020). APC: the five valuation methods. [online] Rics.org. Available at: <https://ww3.rics.org/uk/en/journals/property-journal/apc-5-valuation-methods.html> [Accessed 13 Nov. 2025].

Fu, Y. (2024). A Comparative Study of House Price Prediction Using Linear Regression and Random Forest Models. *Highlights in Science Engineering and Technology*, [online] 107, pp.96–103. doi:https://doi.org/10.54097/vcy5n584.

‌

GeeksforGeeks (2021). *XGBoost*. [online] GeeksforGeeks. Available at: https://www.geeksforgeeks.org/machine-learning/xgboost/ [Accessed 4 Dec. 2025].

GeeksforGeeks (2020). ML Advantages and Disadvantages of Linear Regression. [online] Available at: <https://www.geeksforgeeks.org/machine-learning/ml-advantages-and-disadvantages-of-linear-regression/> [Accessed 14 Nov. 2025].

Grace (2022). Persistent wage gap between North and South - Northern Powerhouse Partnership. [online] Available at: <https://www.northernpowerhousepartnership.co.uk/persistent-wage-gap-between-north-and-south/> [Accessed 24 Nov. 2025].

GeeksforGeeks (2017). *Cross Validation in Machine Learning*. [online] GeeksforGeeks. Available at: https://www.geeksforgeeks.org/machine-learning/cross-validation-machine-learning/ [Accessed 8 Dec. 2025].

Harel, A. and Harpaz, G. (2024). Why stock analysts may make wrong predictions? Economics Letters, 244, p.111956. doi:<https://doi.org/10.1016/j.econlet.2024.111956>.

Hargrave, M. (2025). Understanding Hedonic Pricing: Definition, Uses, and Real Estate Examples. [online] Investopedia. Available at: <https://www.investopedia.com/terms/h/hedonicpricing.asp> [Accessed 23 Nov. 2025].

Hossein Ashtari (2024). *XGBoost vs. Random Forest vs. Gradient Boosting: Differences | Spiceworks*. [online] Spiceworks Inc. Available at: https://www.spiceworks.com/tech/artificial-intelligence/articles/xgboost-vs-random-forest-vs-gradient-boosting/ [Accessed 4 Dec. 2025].

Hilton, D. (2024). *What’s happening with UK house prices? Latest property market moves and forecasts*. [online] MoneyWeek. Available at: https://moneyweek.com/investments/house-prices/house-prices [Accessed 9 Dec. 2025].

‌

Hayes, A. (2025). Automated Valuation Model (AVM): Definition and How It Works. [online] Investopedia. Available at: <https://www.investopedia.com/terms/a/automated-valuation-model.asp> [Accessed 24 Nov. 2025].

he /him, C.R. (2023). *The Notorious XGBoost | Towards Data Science*. [online] Towards Data Science. Available at: https://towardsdatascience.com/the-notorious-xgboost-c7f7adc4c183/ [Accessed 4 Dec. 2025].

‌

HM Land Registry (2023). UK House Price Index England: June 2023. [online] Available at: <https://www.gov.uk/government/statistics/uk-house-price-index-for-june-2023/uk-house-price-index-england-june-2023> [Accessed 25 Nov. 2025].

IBM (2021). Linear Regression. [online] Available at: <https://www.ibm.com/think/topics/linear-regression> [Accessed 13 Nov. 2025].

Institute for Fiscal Studies. (2025). *Your household’s income : Where do you fit in? | Institute for Fiscal Studies*. [online] Available at: https://ifs.org.uk/tools\_and\_resources/where\_do\_you\_fit\_in [Accessed 25 Nov. 2025].

Iwaniuk, M., Jarosz, M., Borycki, B., Jezierski, B., Cwalina, J., Kaźmierczak, S. and Mańdziuk, J. (2025). *The Impact of Bootstrap Sampling Rate on Random Forest Performance in Regression Tasks*. [online] arXiv.org. Available at: https://arxiv.org/abs/2511.13952 [Accessed 4 Dec. 2025].

‌

ifttt-user (2023). On Common Split for Training, Validation, and Test Sets in Machine Learning. [online] Available at: <https://towardsai.net/p/artificial-intelligence/on-common-split-for-training-validation-and-test-sets-in-machine-learning-2> [Accessed 23 Nov. 2025].

Kara, M. (2024). What are some pros and cons of decision trees? [online] Medium. Available at: <https://markmkara.medium.com/what-are-some-pros-and-cons-of-decision-trees-981926602359> [Accessed 24 Nov. 2025].

KAVITA (2021). Linear Regression in Machine Learning. [online] Analytics Vidhya. Available at: <https://www.analyticsvidhya.com/blog/2021/10/everything-you-need-to-know-about-linear-regression/> [Accessed 24 Nov. 2025].

Kaustubh Chakradeo, Huynh, I., Balaganeshan, S.B., Dollerup, O.L., Hjørdis Gade-Jørgensen, Laupstad, S.K., Malham, M., Nguyen, T.-L., Hulman, A. and Varga, T.V. (2025). Navigating fairness aspects of clinical prediction models. *BMC Medicine*, [online] 23(1), pp.567–567. doi:https://doi.org/10.1186/s12916-025-04340-3.

‌

Kavlakoglu, E. and Russi, E. (2024). XGBoost. [online] IBM.com. Available at: <https://www.ibm.com/think/topics/xgboost> [Accessed 23 Nov. 2025].

KEEP, M. and Keep, M. (2025). Housing market: Economic indicators. [online] House of Commons Library. Available at: <https://commonslibrary.parliament.uk/research-briefings/sn02820/> [Accessed 13 Nov. 2025].

Khan, F.T. (2024). Scaling Up Machine Learning: Efficient Strategies for Handling Large Datasets. [online] Medium. Available at: <https://medium.com/@ftech/scaling-up-machine-learning-efficient-strategies-for-handling-large-datasets-1d329c608470> [Accessed 23 Nov. 2025].

Krayonnz.com. (2025). Krayonnz : Social Learning Network. [online] Available at: <https://www.krayonnz.com/user/doubts/detail/623b2b7235e21e005f953106/what-are-the-advantages-and-disadvantages-of-XGBoost> [Accessed 23 Nov. 2025].

Liao, Q.V. and Vaughan, J.W. (2024). AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap. Harvard Data Science Review, (Special Issue 5). doi:<https://doi.org/10.1162/99608f92.8036d03b>.

Lundberg, S. and Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions*. [online] arXiv.org. doi:https://doi.org/10.48550/arXiv.1705.07874.

‌

Lomash Bhuva (2025). *Understanding Feature Importance in Machine Learning🌟🚀*. [online] Medium. Available at: https://medium.com/@lomashbhuva/understanding-feature-importance-in-machine-learning-d86ec50e0055 [Accessed 4 Dec. 2025].

‌

Lloyds Banking Group. (2023). How the pandemic reshaped the UK housing market. [online] Available at: <https://www.lloydsbankinggroup.com/media/press-releases/2023/halifax-2023/three-years-on-how-the-pandemic-reshaped-the-uk-housing-market.html> [Accessed 25 Nov. 2025].

Lazyprogrammer.me. (2023). *XGBoost — Machine Learning and Data Science Compendium*. [online] Available at: https://lazyprogrammer.me/mlcompendium/ensemble/xgboost.html#:~:text=and%20improve%20generalization.-,Disadvantages%20of%20XGBoost,be%20tuned%20for%20optimal%20performance. [Accessed 25 Nov. 2025].

Maloku, F. (2024). House Price Prediction Using Machine Learning and Artificial Intelligence. Journal of Artificial Intelligence & Cloud Computing, 3(4), pp.1–10. doi:<https://doi.org/10.47363/jaicc/2024(3)357>.

Mason-Apps, E. (2022). *The difference between a semi-detached and an end-terrace house | National Statistical*. [online] Ons.gov.uk. Available at: https://blog.ons.gov.uk/2022/03/10/the-difference-between-a-semi-detached-and-an-end-terrace-house/ [Accessed 25 Nov. 2025].

MaPS. (2025). *Leasehold vs freehold: What’s the difference? | MoneyHelper*. [online] Available at: https://www.moneyhelper.org.uk/en/homes/buying-a-home/leasehold-vs-freehold-whats-the-difference [Accessed 25 Nov. 2025].

Nieberg, D. (2016). The Pros and cons of using house price evaluation tools in the current market. [online] Property Rescue. Available at: <https://propertyrescue.co.uk/useful-guides-articles/the-pros-and-cons-of-using-house-price-evaluation-tools-in-the-current-market/> [Accessed 13 Nov. 2025].

NVIDIA Data Science Glossary. (2019). What is XGBoost? [online] Available at: <https://www.nvidia.com/en-gb/glossary/xgboost/> [Accessed 23 Nov. 2025].

or (2020). 80-20 or 80-10-10 for training machine learning models? [online] Data Science Stack Exchange. Available at: <https://datascience.stackexchange.com/questions/69907/80-20-or-80-10-10-for-training-machine-learning-models> [Accessed 23 Nov. 2025].

Pricing algorithms Economic working paper on the use of algorithms to facilitate collusion and personalised pricing 8 October 2018 CMA94. (n.d.). Available at: <https://assets.publishing.service.gov.uk/media/5bbb2384ed915d238f9cc2e7/Algorithms_econ_report.pdf#page=10&zoom=100> [Accessed 24 Nov. 2025].

rick (2023). How AI Will Revolutionise the Property Market in the UK. [online] Available at: <https://addpropertyfinance.co.uk/how-ai-will-revolutionise-the-property-market-in-the-uk/> [Accessed 24 Nov. 2025].

Segal, T. (2021). How Interest Rates Affect Property Values. [online] Investopedia. Available at: <https://www.investopedia.com/articles/mortgages-real-estate/08/interest-rates-affect-property-values.asp> [Accessed 23 Nov. 2025].

Sharma, A. (2020). *Decision Tree vs Random Forest | Which Is Right for You?* [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/#:~:text=A%20decision%20tree%20is%20simpler,%2C%20healthcare%2C%20and%20deep%20learning. [Accessed 25 Nov. 2025].

Sruthi (2021). *Random Forest Algorithm in Machine Learning*. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/ [Accessed 27 Nov. 2025].

‌

Sharma, H., Harsora, H. and Ogunleye, B. (2024). An Optimal House Price Prediction Algorithm: XGBoost. Analytics, 3(1), pp.30–45. doi:<https://doi.org/10.3390/analytics3010003>.

Spann, M. et al. (2025). Algorithmic pricing: Implications for marketing strategy and regulation. International Journal of Research in Marketing. doi:<https://doi.org/10.1016/j.ijresmar.2025.05.001>.

Spira, L.S. (2025). Companies are using algorithms to charge you more. [online] TechEquity Collaborative. Available at: <https://techequity.us/2025/08/18/algorithmic-price-fixing/> [Accessed 24 Nov. 2025].

Thealliance.ai. (2025). Mastering Data Cleaning for Fine-Tuning LLMs and RAG Architectures. [online] Available at: <https://thealliance.ai/blog/mastering-data-cleaning-for-fine-tuning-llms-and-r> [Accessed 24 Nov. 2025].

The Impact of Large Language Models in Finance: Towards Trustworthy Adoption. (n.d.). [online] Available at: <https://www.turing.ac.uk/sites/default/files/2024-06/the_impact_of_large_language_models_in_finance_-_towards_trustworthy_adoption_1.pdf>.

Truong, Q., Nguyen, M., Dang, H. and Mei, B. (2020). Housing Price Prediction via Improved Machine Learning Techniques. Procedia Computer Science, 174, pp.433–442. doi:<https://doi.org/10.1016/j.procs.2020.06.111>.

Uber. (2025). Uber. [online] Available at: <https://www.uber.com/gb/en/drive/driver-app/how-surge-works/> [Accessed 24 Nov. 2025].

White, M. and Papastamos, D. (2022). Buyer behaviour and price expectations: a spatial analysis of the Athens residential market. *Journal of European Real Estate Research*, [online] 15(3), pp.463–481. doi:https://doi.org/10.1108/jerer-03-2020-0013.

‌

Vera, L.Q. and Vaughan, J.W. (2023). AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap. [online] arXiv.org. Available at: <https://arxiv.org/abs/2306.01941> [Accessed 24 Nov. 2025].

Wan, H., Chowdhury, P.K.R., Yoon, J., Bhaduri, P., Srikrishnan, V., Judi, D. and Daniel, B. (2025). Explaining drivers of housing prices with nonlinear hedonic regressions. *Machine Learning with Applications*, [online] 21, p.100707. doi:https://doi.org/10.1016/j.mlwa.2025.100707.

‌

Yadav, A. (2024). Linear Regression for Stock Market Prediction. [online] Medium. Available at: <https://medium.com/@amit25173/linear-regression-for-stock-market-prediction-6039f1ea5c1b> [Accessed 21 Nov. 2025].

YallaValue. (2025). Comparing Instant and Traditional Property Valuations. [online] Available at: <https://yallavalue.com/blog/instant-property-valuations-vs-traditional-property-valuations-pros-and-cons> [Accessed 13 Nov. 2025].

Yao, E. (2024). Predict Housing Price using Linear Regression in Python. [online] Towards Data Science. Available at: <https://towardsdatascience.com/predict-housing-price-using-linear-regression-in-python-bfc0fcfff640/> [Accessed 14 Nov. 2025].

Zhang, H. (2025). Residential real estate price prediction based on adaptive loss function and feature embedding optimization. Humanities and Social Sciences Communications, 12(1). doi:<https://doi.org/10.1057/s41599-025-05217-9>.

Zhao, Y., Zhao, J. and Lam, E.Y. (2024). House Price Prediction: A Multi-Source Data Fusion Perspective. Big Data Mining and Analytics, 7(3), pp.603–620. doi:<https://doi.org/10.26599/bdma.2024.9020019>.

Zoopla.co.uk. (2023). *Less is more - smaller properties favoured by buyers - Zoopla*. [online] Available at: https://www.zoopla.co.uk/discover/property-news/less-is-more-smaller-properties-favoured-by-buyers/ [Accessed 25 Nov. 2025].

‌

# Bibliography

Babitz, K. (2019) *Introduction to Plotting with Matplotlib in Python*. Available at: <https://www.datacamp.com/tutorial/matplotlib-tutorial-python> (Accessed: 20 November 2025).

Cbre.co.uk (2024) *Artificial Intelligence and real estate market forecasting*. Available at: <https://www.cbre.co.uk/insights/articles/artificial-intelligence-and-real-estate-market-forecasting> (Accessed: 21 November 2025).

Cbre.co.uk (2024) *Real estate firms should move faster on AI adoption, with new technology poised to transform sector*. Available at: <https://www.cbre.co.uk/press-releases/real-estate-firms-should-move-faster-on-ai-adoption-with-new-technology-poised-to-transform-sector> (Accessed: 21 November 2025).

Chi, B., Dennett, A., Oléron-Evans, T. and Morphet, R. (2019) *A new attribute-linked residential property price dataset for England and Wales 2011–2019*. Available at: <https://reshare.ukdataservice.ac.uk/854942/1/tranall2011_19.csv>.

EXPO REAL (2025) *EXPO REAL*. Available at: <https://exporeal.net/en/industry-topics/ai-real-estate/> (Accessed: 21 November 2025).

GeeksforGeeks (2018) *Linear Regression in Machine Learning*. Available at: <https://www.geeksforgeeks.org/machine-learning/ml-linear-regression/> (Accessed: 19 November 2025).

GeeksforGeeks (2018) *Reading an Excel File Using Python*. Available at: <https://www.geeksforgeeks.org/python/reading-excel-file-using-python/> (Accessed: 15 November 2025).

GeeksforGeeks (2019) *Random Forest Regression in Python*. Available at: <https://www.geeksforgeeks.org/machine-learning/random-forest-regression-in-python/> (Accessed: 20 November 2025).

GeeksforGeeks (2021) *How to Merge Multiple Excel Files into a Single File with Python*. Available at: <https://www.geeksforgeeks.org/python/how-to-merge-multiple-excel-files-into-a-single-files-with-python/> (Accessed: 15 November 2025).

Kavlakoglu, E. (2021) *Decision Trees*. Available at: <https://www.ibm.com/think/topics/decision-trees> (Accessed: 19 November 2025).

Matplotlib.org (2024) *Matplotlib — Visualization with Python*. Available at: <https://matplotlib.org/> (Accessed: 20 November 2025).

Matterport (2025) *AI in Real Estate: How To Use Technology To Elevate Your Marketing*. Available at: <https://matterport.com/en-gb/blog/ai-real-estate> (Accessed: 21 November 2025).

Stojiljković, M. (2024) *Linear Regression in Python*. Available at: <https://realpython.com/linear-regression-in-python/> (Accessed: 16 November 2025).

Perry, C. (2022) *SimpleBar*. Available at: <https://www.altusgroup.com/insights/the-real-estate-ai-debate-risk-or-revolution/> (Accessed: 21 November 2025).

Placester (2025) *How to Use AI for Real Estate? 12 Winning Strategies*. Available at: <https://placester.com/real-estate-marketing-academy/how-to-use-ai-in-real-estate> (Accessed: 21 November 2025).

scikit-learn (2025) *1.10. Decision Trees*. Available at: <https://scikit-learn.org/stable/modules/tree.html> (Accessed: 19 November 2025).

scikit-learn (2025) *train\_test\_split*. Available at: <https://www.scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html> (Accessed: 19 November 2025).

Szumilo, N. and Wiegelmann, T. (2024) ‘Real Estate Insights AI: real estate’s new roommate – the good, the bad and the algorithmic’, *Journal of Property Investment & Finance*, 42(2), pp. 211–217. doi:10.1108/JPIF-01-2024-0001.

Readthedocs.io. (2025). *Get Started with XGBoost — xgboost 3.1.1 documentation*. [online] Available at: https://xgboost.readthedocs.io/en/stable/get\_started.html [Accessed 4 Dec. 2025].

Tyson, L. (2025) *How AI in Real Estate Is Shaping a Smarter, More Efficient Market*. Available at: <https://medium.com/@lenaztyson/how-ai-in-real-estate-is-shaping-a-smarter-more-efficient-market-348b6db90148> (Accessed: 21 November 2025).

W3Schools.com (2025) *Matplotlib Pyplot*. Available at: <https://www.w3schools.com/python/matplotlib_pyplot.asp> (Accessed: 15 November 2025).

Yildiz, D.M. (2025) *Understanding the Differences and Similarities Between Feature Importance and Feature Coefficients*. Available at: <https://medium.com/@muslumyildiz17/understanding-the-differences-and-similarities-between-feature-importance-and-feature-coefficients-08ab29bdb5b8> (Accessed: 20 November 2025).

Zekri, M. (2025) ‘Data-driven literature analysis of artificial intelligence (AI) in real estate research: a bibliometric and machine learning-based approach’, *Property Management*, pp. 1–25. doi:10.1108/PM-04-2025-0059.

Houseprice.ai. (2025). *home*. [online] Available at: https://www.houseprice.ai/user/sign-in?next=/ [Accessed 8 Dec. 2025].

AccuVal. (2024). *AccuVal*. [online] Available at: https://accuval.co.uk/ [Accessed 8 Dec. 2025].

‌ Edri, E. (2025). *Will AI Replace Chartered Surveyors? A Deep Dive into Free AI Property Valuation Tools in the UK (2025)*. [online] Lendlord. Available at: https://lendlord.io/will-ai-replace-chartered-surveyors-a-deep-dive-into-free-ai-property-valuation-tools-in-the-uk-2025 [Accessed 8 Dec. 2025].

‌ Wheaton, W.C. and Xu, C. (2024). Using AI to Improve Price Transparency in Real Estate Valuation. [online] doi:https://doi.org/10.2139/ssrn.5056315.

‌ Alzain, E., Alshebami, A.S., Aldhyani, T.H.H. and Alsubari, S.N. (2022). Application of Artificial Intelligence for Predicting Real Estate Prices: The Case of Saudi Arabia. *Electronics*, [online] 11(21), p.3448. doi:https://doi.org/10.3390/electronics11213448.

‌ Asfi, M., Warsito, B. and Wibowo, A. (2024). Enhancing Explainable AI: Leveraging SHAP for Transparent Decision-Making in Machine Learning. *2024 Ninth International Conference on Informatics and Computing (ICIC)*, [online] pp.1–6. doi:https://doi.org/10.1109/icic64337.2024.10956244.

‌ Keita, Z. (2023). *Explainable AI - Understanding and Trusting Machine Learning Models*. [online] Datacamp.com. Available at: https://www.datacamp.com/tutorial/explainable-ai-understanding-and-trusting-machine-learning-models [Accessed 8 Dec. 2025].

‌ Harish K (2025). *EBM vs SHAP: Built-in Transparency vs Post-hoc Explanations — Which Path to Explainable AI?* [online] Medium. Available at: https://medium.com/@harishk3493/ebm-vs-shap-built-in-transparency-vs-post-hoc-explanations-which-path-to-explainable-ai-3cc875ae9b56 [Accessed 8 Dec. 2025].

‌ GeeksforGeeks (2017). *Decision Tree*. [online] GeeksforGeeks. Available at: https://www.geeksforgeeks.org/machine-learning/decision-tree/ [Accessed 8 Dec. 2025].

‌ Avinash Navlani (2018). *Decision Tree Classification in Python Tutorial*. [online] Datacamp.com. Available at: https://www.datacamp.com/tutorial/decision-tree-classification-python [Accessed 8 Dec. 2025].

‌ scikit-learn. (2025). *LinearRegression*. [online] Available at: https://scikit-learn.org/stable/modules/generated/sklearn.linear\_model.LinearRegression.html [Accessed 8 Dec. 2025].

‌ Waples, J. (2024). *Simple Linear Regression: Everything You Need to Know*. [online] Datacamp.com. Available at: https://www.datacamp.com/tutorial/simple-linear-regression [Accessed 8 Dec. 2025].

‌ Readthedocs.io. (2025). *XGBoost Documentation — xgboost 3.1.1 documentation*. [online] Available at: https://xgboost.readthedocs.io/en/stable/ [Accessed 8 Dec. 2025].

‌ Bekhruz Tuychiev (2019). *Using XGBoost in Python Tutorial*. [online] Datacamp.com. Available at: https://www.datacamp.com/tutorial/xgboost-in-python [Accessed 8 Dec. 2025].

‌ Bergmann, D. (2025). *Machine learning*. [online] Ibm.com. Available at: https://www.ibm.com/think/topics/machine-learning [Accessed 8 Dec. 2025].

‌ Amazon Web Services, Inc. (2025). *Machine Learning*. [online] Available at: https://aws.amazon.com/ai/machine-learning/ [Accessed 8 Dec. 2025].

‌