Predicting House Price Ranges

Rajendra Kumar Durgesh Tiwari

Project-kumar13-wcutchin-dutiwar

Abstract

The" House Price Range Prediction (UK)" project utilizes data from the UK housing market spanning 1995 to 2023, as provided by HM Land Registry. This project aims to predict future home price ranges (basically categorized into four categories), offering valuable insights to potential home buyers based on property attributes such as locality, postcode, district, and county. To make this possible, we employ exploratory data analysis (EDA) to understand and refine the dataset, addressing the challenges of millions of records through data pruning and condensation. Furthermore, we utilize the random forest technique under the classification model to create an accurate predictive model for future housing price ranges. This project directly addresses the common problem faced by individuals planning to buy homes, providing them with a data-driven solution for budgeting and improved planning. The" House Price Range Prediction in the UK" project empowers potential home buyers to make informed decisions and work towards their housing goals.

Keywords

Property Sales, Investment Opportunities, Random Forest Classifier for price prediction, XG-Boost for house price, Logistic Regression on house price, Real estate market, ML in Real estate, future house price range, Model Evaluation, Classification, UK Real State.

1 Introduction

Navigating the intricacies of the United Kingdom's housing market, shaped by economic fluctuations and geopolitical events, is a daunting task for potential homebuyers. To address this challenge, our project utilizes extensive data from HM Land Registry spanning 1995 to 2023. Through exploratory data analysis (EDA) and predictive models, we aim to empower individuals with precise insights, enabling informed decision-making and effective financial planning. Our primary objective is to develop a user-friendly predictive tool categorizing house prices into four groups: Base, Low, Moderate, and High. Focusing on property details like location, postcodes, and districts, the tool guides potential homebuyers in predicting future house price ranges and planning budgets accurately for their desired location. The project adopts a systematic approach, starting with EDA to comprehend the 1.2 million-record dataset.

We address challenges in managing extensive data and implementing random forest and XG-Boost classification models for precise predictions. This methodical approach ensures a nuanced understanding of the housing market, facilitating the creation of a dependable predictive tool. Beyond traditional data analysis, this project aspires to impact users by providing a transformative tool. It delivers insights into future house price ranges, instills confidence in navigating

the housing market, and facilitates effective financial planning. The scalable nature of our approach allows for global application, offering reliable and accurate insights regardless of location.

Our ultimate goal is to demystify the housing market by providing potential homebuyers with a tool that ensures accurate budgeting for their dream homes. We aim to deliver the best price insights that remain robust and unchanging, empowering users to make informed decisions, plan effectively, and secure their ideal homes with confidence.

Previous work

The landscape of real estate price prediction has been shaped by several notable studies. The HM Land Registry Open Data has been instrumental in providing comprehensive datasets that serve as a basis for predictive analysis and provide a rich historical perspective on property valuations in the UK ref [a]. Similarly, the data curated on Kaggle has facilitated numerous machine learning projects, with UK property price data from 1995 to 2023 standing out as a central resource for modern models ref [b] but it has just 3 attemts and those also seems incomplete, so we tried to do our own work and see if we could be able to get a better perspective of solving this problem.

In the field of real estate price forecasting, the study at Science Direct ref [c]. represents a remarkable investigation into the effectiveness of machine learning methods. Recognizing the multiple determinants of property prices such as location, area and population, this research goes beyond the traditional house price index (HPI) to encompass a broader range of predictive factors. In a landscape increasingly influenced by machine learning, Rawool et alref [d]. have embarked on a journey to revolutionize the real estate market with their cutting-edge housing cost prediction model. The study conducted by Atharva et al and teamref [e]. offers an innovative approach to estimating the prices of real estate transactions by applying various machine learning algorithms. Taking into account the multiple factors that influence the selling prices of houses, such as plot size, location, building materials, age of the property, number of bedrooms and garages, the study develops a prediction model tailored to the complexity of the real estate market.

In a pioneering analysis of the Spanish real estate marketref [f], Mora-Garcia et al. delve into the capabilities of machine learning algorithms to predict housing prices with a keen eye on the socioeconomic upheavals brought by the COVID-19 pandemic. Their comprehensive methodology traverses the entire spectrum of predictive modeling, from meticulous data preparation and feature engineering to hyperparameter tuning and model interpretation

2 Methods

The data collection process, as outlined in Section 2.1, involves obtaining data from Kaggle, sourced from the HM Land Ministry of the UK. Figure 1 visually represents this process, incorporating a line graph, scatter plot, and box plot for comprehensive data analysis. Following this, Section 2.2 delves into data preprocessing, including outlier detection, duplicate removal, and handling null values, ensuring the integrity of the dataset. In Section 2.3, encoding techniques are applied (Figure 3) to enhance model accuracy and reduce training time. Feature engineering, explored in Section 2.4 and illustrated in Figure 2, aims to identify crucial parameters for model training while balancing the data. The subsequent section, 2.5, introduces machine learning algorithms and highlights feature identification. Finally, Section 2.6 provides an evaluation of the models, presenting recall, precision, and F1-score metrics for various algorithms. This comprehensive approach ensures a thorough understanding of the data analysis process, incorporating visualization, preprocessing, and model evaluation.

2.1 Data collection

The dataset employed in our project is formally acquired from Kaggle, originating from the HM Land Ministry of the UK. This data is meticulously collected from various locations across the UK, offering insights into the inflation dynamics of housing prices influenced by the prevailing conditions. Sourced from a reputable platform like Kaggle and endorsed by the HM Land Ministry, the dataset ensures credibility and reliability. Focusing on the geographical variations, the dataset provides a nuanced understanding of how housing prices are affected by ongoing situations. Our analysis aims to unravel these patterns, contributing valuable perspectives to navigate the complex terrain of the UK housing market.

2.2 Data Preprocessing

The data we collected has challenges like outliers and imbalanced data. To address this, we're actively cleaning the data—removing duplicates, handling null values, and identifying outliers. Given the data's size, splitting it is crucial. Employing exploratory data analysis (EDA), we gain insights into the data's nuances, guiding decisions on what segments to use for training and testing. These steps ensure a refined and reliable dataset, laying a strong foundation for subsequent analyses and modeling processes.

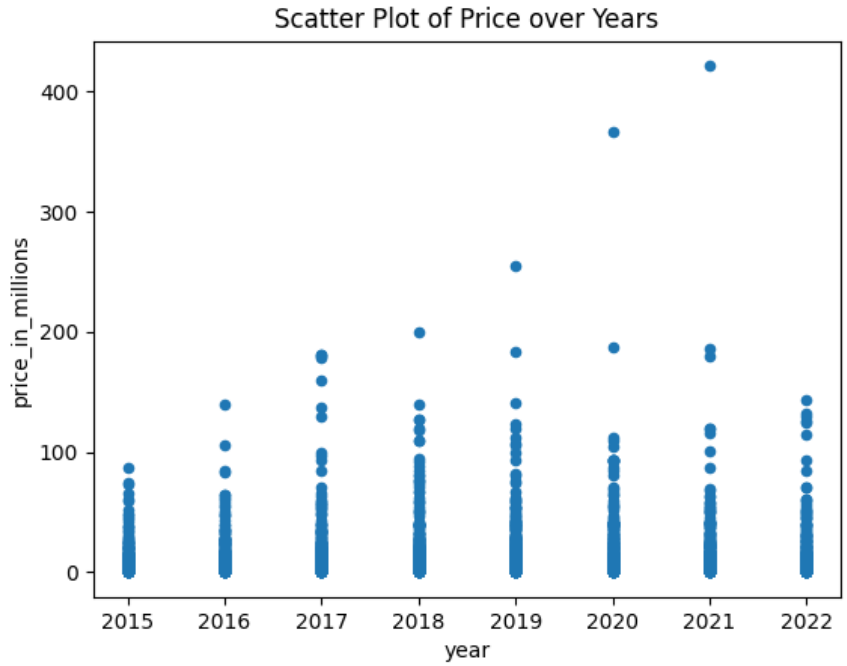
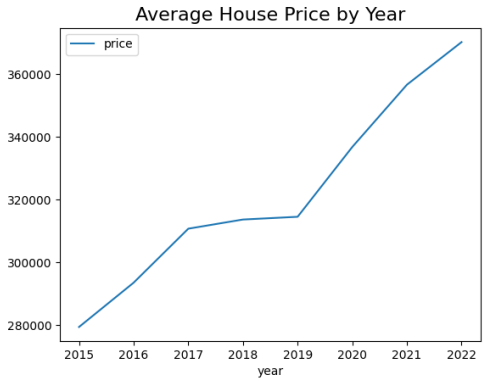


Figure 1: Scatter plot showcasing the distribution in years

The figure below shows that we are using a boxplot to observe how the data is distributed, especially the prices of houses in different locations. Additionally, a scatter plot is used to observe how prices change over time in various locations in the UK. Lastly, the line plot demonstrates how economic recessions and COVID-19 have impacted the overall inflation of house prices over time.



(a) Average house price by year

(b) boxplot price

Figure 2: Exploratory Data Analysis

2.3 Data Encoding

Since we are using categorical data for our project, it is nonetheless important to use encoding techniques to transform the data in such a manner so the model can interpret the meaning of the data correctly. To implement this, we use one-hot encoding and label encoding. Encoding techniques help us achieve accuracy compared to before when it took more time to process and the algorithm's accuracy was low. Lastly, we label-encoded all the features and observed the accuracy increase. Earlier we tried frequency encoding and one-hot encoding which resulted in accuracy ranging between 40% to 55%.

2.4 Feature Engineering

As part of a price range forecasting project focused on the UK real estate market, we are in the process of developing an important new column entitled 'Flag". This newly developed feature will categorize properties into one of four different classifications based on their price range. The essence of this categorical feature is to capture the different price segments of UK properties and provide a structured and simplified view that aligns with the project's aim of predicting a property's price range.

We observed various other features and realized that we had fewer time to analyse all those 16 features and come out with new features and it would also cause us to exhaust more computation power that usual. So we dropped the idea of identifying more features from the available ones.

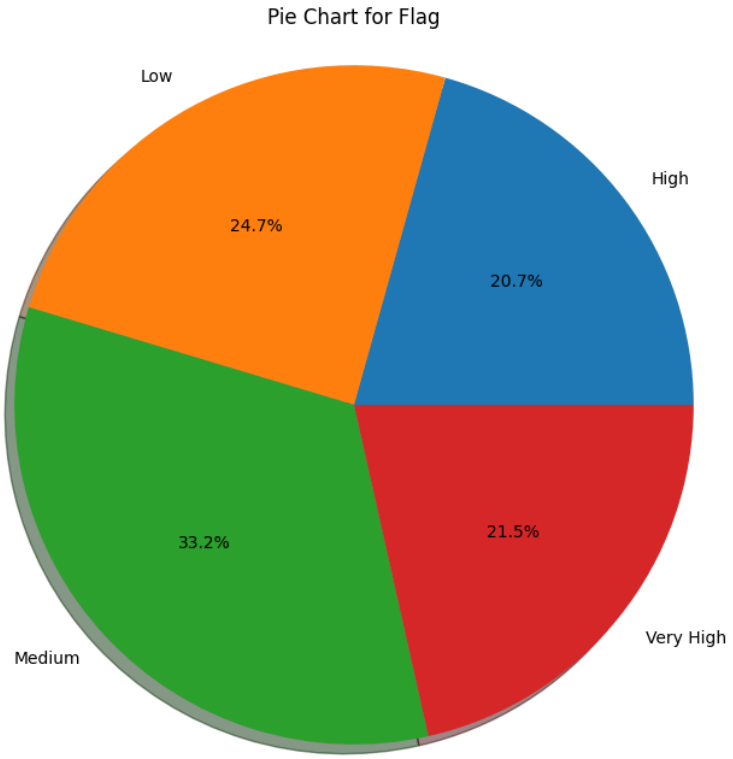


Figure 3: Pie-chart showcasing the data distribution for flag

To enable this, the 'flag' feature is carefully elaborated by dividing the continuous' price' data into quartiles denoting different price bands. Following derivation, this categorical attribute is transformed by coding labels. This is a technique that converts the category labels into a numerical format suitable for model training. This coding is a crucial step because it enables the subsequent prediction models to interpret and utilize the 'flag' feature effectively, thereby enabling them to infer the price range of the houses in a quantifiable and algorithm-friendly manner. Here the two pie-charts shows different distribution because we have eliminated some of the records where there are houses with more than 100 millions in prices and restricted by 10k as lowest price.

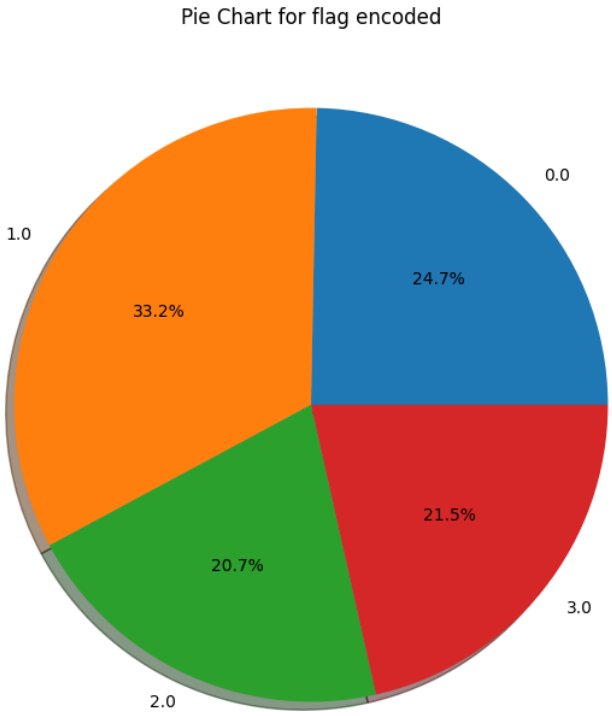


Figure 4: Pie-chart, distribution for flag post-encoding.

2.5 Algorithms and Implementation

In our project, we used three machine learning algorithms to achieve our goal. We employed Random Forest and XGBoost for handling categorical data, while Logistic Regression played a role in understanding essential parameters for optimizing model efficiency. Upon examining the classification report, it is evident that Random Forest outperforms the XGB algorithm, training in less time in comparison. Random Forest excels, particularly in precision for classes 0 and 3, showcasing its ability to make positive predictions. XGBoost lags behind in precision and recall metrics, especially for class 2. The F1-scores for Random Forest is generally higher, indicating a more balanced trade-off between precision and recall. Furthermore, the overall accuracy for Random Forest is 0.68, outnumbering XGBoost at 0.63, demonstrating a higher ability to predict our target.

Classification Report:

	precision	recall	f1-score	support
0.0	0.76	0.77	0.77	75083
1.0	0.65	0.70	0.68	100794
2.0	0.57	0.48	0.52	62960
3.0	0.73	0.75	0.74	65000
accuracy			0.68	303837
macro avg	0.68	0.68	0.68	303837
weighted avg	0.68	0.68	0.68	303837

(a) Classification report for Random Forest

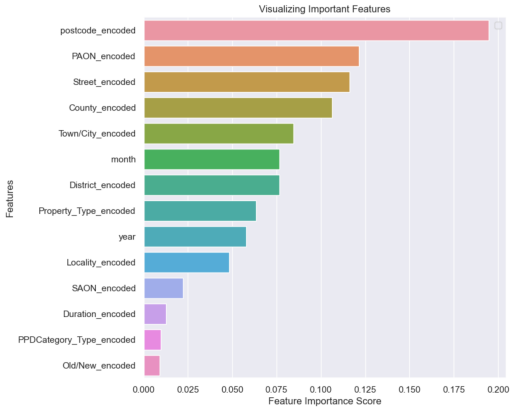
Classification Report:

	precision	recall	f1-score	support
0.0	0.71	0.70	0.70	75083
1.0	0.58	0.69	0.63	100794
2.0	0.53	0.36	0.42	62960
3.0	0.69	0.71	0.70	65000
accuracy			0.63	303837
macro avg	0.62	0.61	0.61	303837
weighted avg	0.62	0.63	0.62	303837

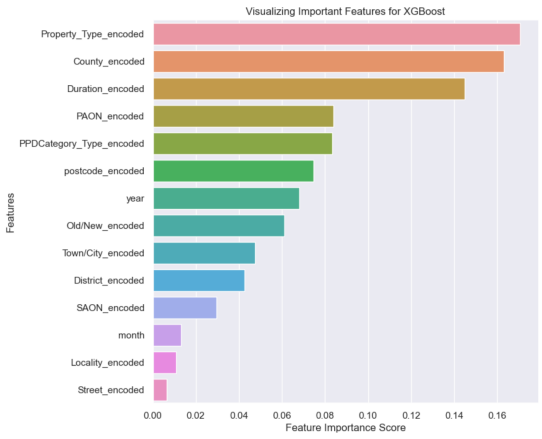
(b) Classification report for XGBoost

Figure 5: Classification report for RF and XGB

Moreover, in the visualization importance report, Random Forest selects parameters close to those of XGBoost, making it more reliable for predicting our parameter. Finally, Logistic Regression helps identify which particular features impact the overall prediction. The visualization report highlights that" locality-encoded" has the most significant impact, while" SAON-encoded" has the most significant impact. contributes the least to the prediction.



(a) Feature Importance for Random Forest



(b) Feature Importance for XGBoost

Figure 6: Feature Importance for RF and XGB

2.6 Model Evaluation

When evaluating machine learning models for predicting housing markets, the Random Forest model performs better than XGBoost, with an accuracy of 68%. surprisingly, Random Forest does exceptionally well in forecasting positive outcomes for classes 0 and 3, as seen by its higher recall as well as accuracy. But XGBoost has difficulties, especially with class 2, which shows in its lower accuracy and recall numbers. With a low accuracy of 34%, Logistic Regression trails much behind, suggesting its limits in accurately forecasting all classes. The detailed recall, precision, and F1-scores analysis highlights Random Forest's stability. In conclusion, the Random Forest model appears to be the best option for the challenging task of predicting housing market changes, providing customers with a fair and accurate answer. The we observed is working with Logistic regression. We tried to fine-tune the model, but the performance on this data was super low. Although we have not tried the hyperparameter tuning in any of these models, we observed that the ensemble models have the highest accuracy and F1 score.

3 Results

The given data were analyzed using different methods involving cleaning, preprocessing, encoding, and algorithmic approaches. Figure 1 and 2: Illustrates the distribution of data for 1.2 million records, encompassing outlier and noise analysis using boxplots, scatter plots, and line graphs to better understand how inflation affects house prices. Figure 3 presents a pie chart showcasing the balance achieved in different categorical data after implementing Exploratory Data Analysis (EDA) and other preprocessing steps. We identified that there was a huge amount of data and normal computation power would not work, so we reduced it to 1.2 million records. Figure 4: Specifically displays the results of encoding the flags and the result of data reduction by applying a few rules. Figure 5, Presents the classification report where we tried to compare the two models' performance. Random Forest has the highest performance without hyperparameter tuning. Figure 6 depicts the Random Forest and XGBoost feature importance reports where we can clearly observe that the importance given by the two different models is completely different; the postcode has the highest importance in Random Forest, whereas the property type has the highest importance in XGBoost model.

In conclusion, the best model for this type of problem would be the bagging model, like the Random Forest model. This resulted in better performance with label-encoded data and has the highest F1 score when compared to others.

4 Discussion

The development of our model to predict the price range for houses represents a significant advance in real estate analysis. This model uses historical data from 1995 to 2023 to forecast property prices in a given location, providing users with a valuable resource for planning future investments. Our initial foray into the dataset has revealed its complexity, which can be clearly seen in Figure 1. It shows that careful normalization is required due to the outliers.

In our efforts to refine the data, feature engineering proved to be a crucial step. It leveled the playing field by balancing the data and removing interfering noise, improving the model's prediction accuracy and speeding up its training process. An essential part of our analysis was visualizing the importance of the features, as shown in Figure 3. This visualization was key to deciphering the influence of the different variables on the learning of the model. The feature 'postcode-coded' proved to be the top performer in the Random Forest algorithm, while the feature 'property-type-coded' gained importance in the XGBoost approach.

In the evaluation, the Random Forest model showed superior performance, not only in terms of accuracy (68% versus 63% for XGBoost), but also in terms of the training and test datasets. Its robustness is based on its inherent ability to maintain an optimal balance between bias and variance, ensuring that the model performs reliably and generalizes well to new, unseen data.

The progress that the model has made shows that it is very promising for the real estate sector. Looking to future iterations, we want to expand the reach of the model on a global scale and provide a beacon of insight to potential buyers and investors looking to maximize their returns - whether through buying, leasing or other investment ventures. For those embarking on the search for their perfect home, or for those seeking a path through the complex waters of real estate investing, this model is a testament to the power of data-driven decision-making.

This project has the capability to expand to a world level and has the ideology that can be implemented on most real-world problems, whether it be a real state problem or it can be related to any industry-related problem like deciding the price range for any items in a market or materials required in the manufacturing industry.

5 Author Contribution

Rajendra Kumar: He did most of the coding and decided to build this model on house price data. Also corrected the final report that Durgesh Tiwari initially created. He corrected the figures and indentation and also fixed some of the sentences where they were required.
Durgesh Tiwari: Mainly worked on the project proposal and helped with the code for some of the Python coding requirements and understanding the model behavior. He also worked on the final report.
3rd member: Our 3rd member dropped out of the course, and he helped with nothing during the course of this project.

References

HM Land Registry Open Data. https://landregistry.data.gov.uk/, a. 2023-11.

Kaggle project Link for house price pridiction. https://www.kaggle.com/datasets/willianoliveiragibin/uk-property-price-data-1995-2023-04/, b. 2023-11.

Science Direct, Journals and Books. https://www.sciencedirect.com/science/article/pii/S1877050920316318, c. 2023-11.

Anand G. Rawool1 , Dattatray V. Rogye2 , Sainath G. Rane3 , Dr. Vinayk a. Bharadi. https://www.irejournals.com/formatedpaper/1702692.pdf, d. 2023-11.

Atharva Chouthai1, Mohammed Athar Rangila, Sanved Amate, Prayag Adhikari. https://www.irjet.net/archives/V6/i3/IRJET-V6I31151.pdf, e. 2023-11.

Raul-Tomas Mora-Garcia, Maria-Francisca Cespedes-Lopez,Raul Perez-Sanchez. https://www.mdpi.com/2073-445X/11/11/2100, f. 2023-11.