Predicting House Price Ranges

Durgesh Tiwari dutiwar@iu.edu
Rajendra Kumar kumar13@iu.edu

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, Data Analysis, Investment Opportunities, Random Forest Classifier, XGBoost Algorithm, Logistic Regression, Feature Selection, Model Training, Validation and Testing, 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 implement random forest and XGBoost 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.

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.

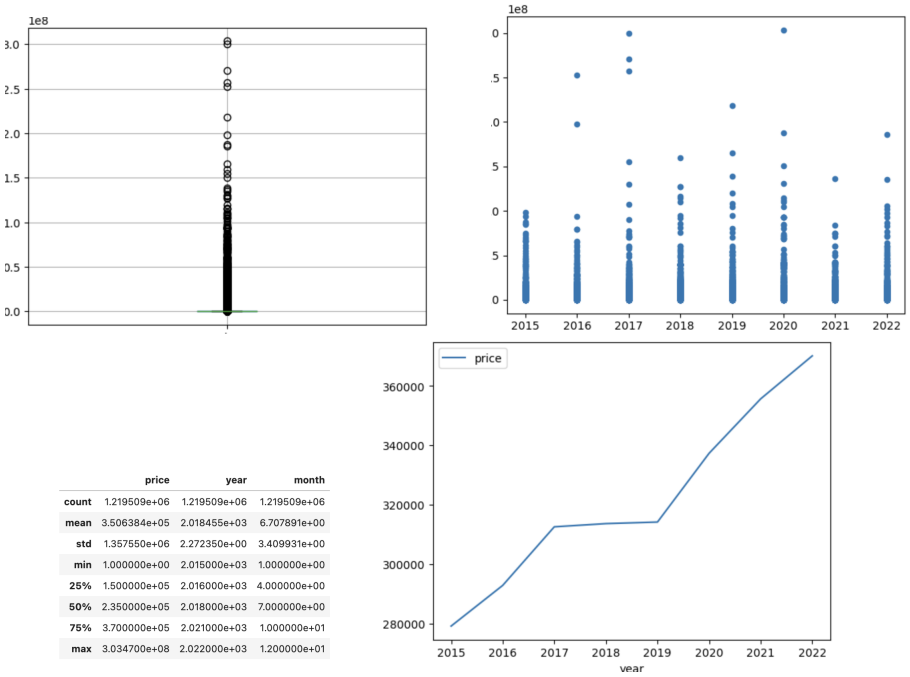
2.1 Data collection

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.

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.



	price	year	month
count	1.219509e+06	1.219509e+06	1.219509e+06
mean	3.506384e+05	2.018455e+03	6.707891e+00
std	1.357550e+06	2.272350e+00	3.409931e+00
min	1.000000e+00	2.015000e+03	1.000000e+00
25%	1.500000e+05	2.016000e+03	4.000000e+00
50%	2.350000e+05	2.018000e+03	7.000000e+00
75%	3.700000e+05	2.021000e+03	1.000000e+01
max	3.034700e+08	2.022000e+03	1.200000e+01

Figure 1: EDA Analysis Report to extract meaningful insight

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.

2.4 Feature engineering

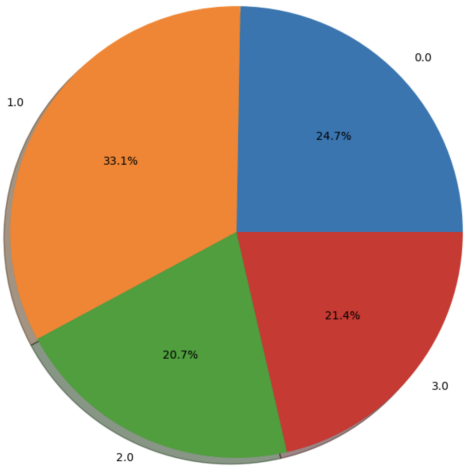


Figure 2: Show a balance data after Feature Engineering

Employing feature engineering is a crucial aspect of data pre-processing. We implemented this technique to significantly reduce training time for our model, especially when dealing with millions of records. In this process, we utilized encoding techniques to capture cyclic patterns, transforming the data to create new features that are more relevant for the algorithm. The goal is to effectively handle the data, making it more balanced to enhance the efficiency of model training, as illustrated in Figure 2.

2.5 Algorithms

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 XGBoost 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 are 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.

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" contributes the least to the prediction.

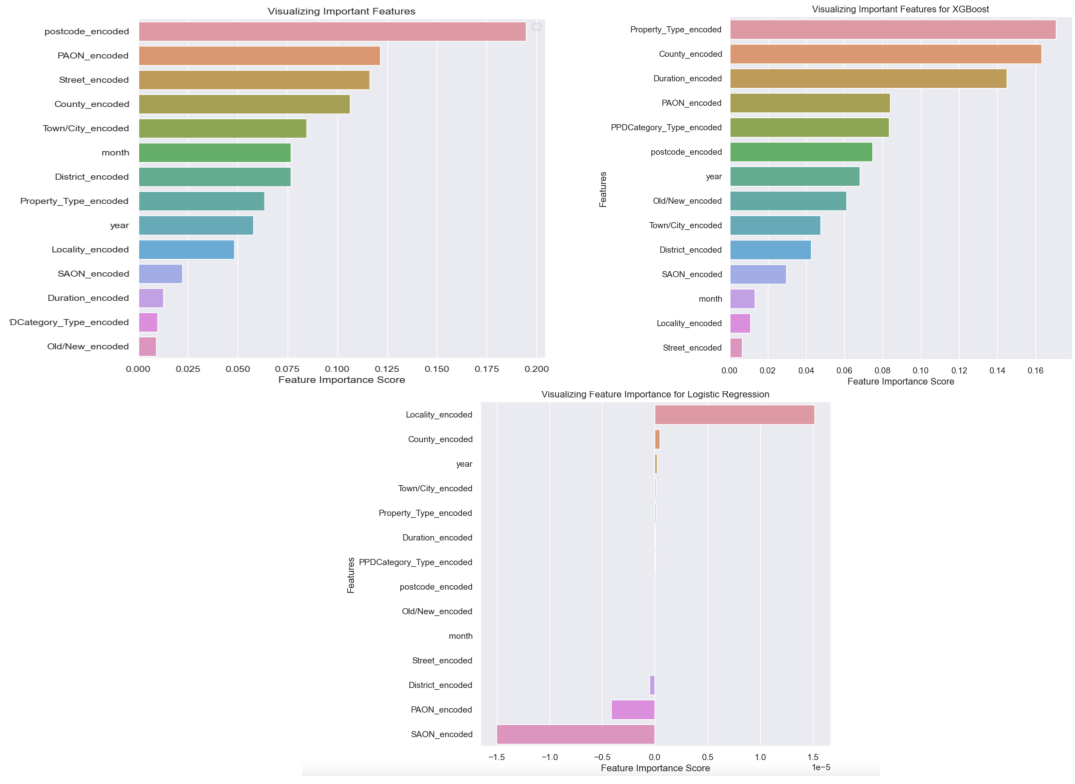


Figure 3: Visualizing Feature Importance for the data

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 analysis of recall, precision, and F1-scores highlights Random Forest's stability. In conclusion, the Random Forest model appears as the best option for the challenging task of predicting housing market changes, providing customers with a fair and accurate answer.

Classification Report:

	precision	recall	f1-score	support
0.0	0.28	0.02	0.04	75083
1.0	0.34	0.90	0.49	100794
2.0	0.00	0.00	0.00	62960
3.0	0.33	0.17	0.22	65000
accuracy			0.34	303837
macro avg	0.24	0.27	0.19	303837
weighted avg	0.25	0.34	0.22	303837

Figure 4: Logistic Regression classification report

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

Figure 5: Xgboost classification report

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

Figure 6: random forest classification report

Figure 7: classification report for different algorithm

3 Results

The given data were analyzed using different methods, involving cleaning, preprocessing, encoding, and algorithmic approaches.
Figure 1: Illustrates the distribution of data for 1.4 million records, encompassing outlier and noise analysis using boxplots, scatter plots, and line graphs to better understand how inflation affects house prices.
Figure 2: Represents a pie chart showcasing the balance achieved in different categorical data after implementing Exploratory Data Analysis (EDA) and other preprocessing steps.
Figure 3: Specifically displays the results of feature selection using various algorithms chosen to predict our target parameter.
Figure 4: Presents the logistic classification report, highlighting the most important feature for accurate predictions.
Figure 5: depicts the XGBoost classification report, indicating the features selected by XGBoost to predict the parameter.
Figure 6: exhibits the random forest classification report, showcasing the features chosen by the random forest to predict the parameter.

4 Discussion

The house price prediction range model is a predictive tool that assists users in estimating the future price of a house at a specific location based on historical data spanning from 1995 to 2023. Initially, data analysis revealed the extensive and challenging nature of the dataset, as depicted in Figure 1, which highlighted the presence of extreme values requiring normalization. Feature engineering was employed to address this, making the data more balanced and eliminating noise, ultimately enhancing model accuracy and training efficiency.

Additionally, feature importance visualization, as shown in Figure 3, played a crucial role in understanding the significance of different features in training the model. Random Forest identified 'postcode-encoded' as the most valuable feature, while XGBoost prioritized 'property-type-encoded.' Consequently, Random Forest achieved an accuracy of 68%, surpassing XGBoost's 63%.

The promising results indicate the model's potential impact on the real estate market. Future enhancements could extend the model's applicability globally, aiding potential buyers in making informed decisions and optimizing property investments for leasing or other purposes. This model serves as a valuable tool for individuals seeking their dream home or making strategic real estate investments worldwide.

References

1. HM Land Registry Open Data. *https://landregistry.data.gov.uk/.* Website Name. URL.
2. Kaggle Link . *https://www.kaggle.com/datasets/willianoliveiragibin/uk-property-price-data-1995-2023-04/code.*Website Name. URL.
3. Science Direct. *https://www.sciencedirect.com/science/article/pii/S1877050920316318* Website Name. URL.
4. Anand G. Rawool1 , Dattatray V. Rogye2 , Sainath G. Rane3 , Dr. Vinayk a. Bharadi. *https://www.irejournals.com/formatedpaper/1702692.pdf* Website Name. URL.

5. Alisha Kuvalekar, Shivani Manchewar, Sidhika Mahadik, Shila Jawale BHARADI *https://papers.ssrn.com/sol3/papers.cfm?abstract*$_id = 3565512.WebsiteNameURL$.

6. Atharva Chouthai1, Mohammed Athar Rangila, Sanved Amate, Prayag Adhikari, Vijay Kukre *https://www.irjet.net/archives/V6/i3/IRJET-V6I31151.pdf.* Website Name. URL.

7. Raul-Tomas Mora-Garcia, Maria-Francisca Cespedes-Lopez,Raul Perez-Sanchez . *https://www.mdpi.com/2073-445X/11/11/2100* Website Name. URL.