information

MDPI

*Article*

# Advanced Machine Learning Techniques for Predictive Modeling of Property Prices

**Kanchana Vishwanadee Mathotaarachchi [1], Raza Hasan [1,*] and Salman Mahmood [2]**

[1] Department of Computer Science, Solent University, Southampton SO14 0YN, UK; 6mathk26@solent.ac.uk
[2] Department of Information Technology, School of Science and Engineering, Malaysia University of Science and Technology, Petaling Jaya 47810, Selangor, Malaysia; salman.mahmood@phd.must.edu.my
* Correspondence: raza.hasan@solent.ac.uk

**Abstract:** Real estate price prediction is crucial for informed decision making in the dynamic real estate sector. In recent years, machine learning (ML) techniques have emerged as powerful tools for enhancing prediction accuracy and data-driven decision making. However, the existing literature lacks a cohesive synthesis of methodologies, findings, and research gaps in ML-based real estate price prediction. This study addresses this gap through a comprehensive literature review, examining various ML approaches, including neural networks, ensemble methods, and advanced regression techniques. We identify key research gaps, such as the limited exploration of hybrid ML-econometric models and the interpretability of ML predictions. To validate the robustness of regression models, we conduct generalization testing on an independent dataset. Results demonstrate the applicability of regression models in predicting real estate prices across diverse markets. Our findings underscore the importance of addressing research gaps to advance the field and enhance the practical applicability of ML techniques in real estate price prediction. This study contributes to a deeper understanding of ML's role in real estate forecasting and provides insights for future research and practical implementation in the real estate industry.

**Keywords:** real estate market dynamics; property price forecasting; machine learning techniques; predictive modeling applications; UK housing market

**Citation:** Mathotaarachchi, K.V.; Hasan, R.; Mahmood, S. Advanced Machine Learning Techniques for Predictive Modeling of Property Prices. *Information* **2024**, *15*, 295. https://doi.org/10.3390/info15060295

Academic Editor: Gabriele Gianini

Received: 5 May 2024
Revised: 19 May 2024
Accepted: 20 May 2024
Published: 22 May 2024



## 1. Introduction

The real estate market is a cornerstone of the economy, influenced by various factors such as interest rates, building costs, and disposable income [1]. However, this sector faces uncertainties exacerbated by events like the COVID-19 pandemic, which can impact investor sentiments and market performance [2]. Despite these challenges, the real estate industry remains a significant contributor to the global economy, particularly within the residential property sector [3].

In the dynamic landscape of the United Kingdom (UK) real estate market, precise predictions are essential for effective decision making. While this study focuses on the unique challenges and dynamics of the UK market, the methodologies and insights presented have broader applicability across various real estate markets globally [4].

This study aims to explore the potential of advanced machine learning methods to address the challenges of predicting property prices in the UK real estate market, with the overarching goal of developing robust predictive models that enhance decision making for stakeholders in the industry. It seeks to investigate these factors comprehensively, develop predictive models capable of accurately forecasting property prices based on relevant market variables, and evaluate the performance of various regression techniques, including Linear Regression, tree-based ensemble methods such as Random Forest, XGBoost, and LightGBM, as well as regularized regression models. Additionally, this study aims to assess the generalization capability of the developed models on unseen data

and evaluate their robustness in capturing underlying market trends. Furthermore, it seeks to provide insights into the potential applications of predictive modeling in real estate decision making and to highlight avenues for future research and development. By delineating these objectives, the research aims to address existing gaps in property price prediction methodologies, contribute to the advancement of predictive modeling techniques in real estate, and facilitate evidence-based decision making in the real estate industry.

This paper is structured as follows: Section 2 conducts a literature review on real estate challenges and machine learning models. Section 3 outlines the research methodology, covering design, data sources, and model application. In Section 4, results are presented and discussed, evaluating model performance. In Section 5, discussions about the implications and interpretations of the results are elaborated upon. Finally, Section 6 concludes this paper by addressing limitations, offering recommendations, and emphasizing practical implications for real estate stakeholders.

## 2. Literature Review

The application of ML techniques for real estate price prediction has gained significant traction in recent years, driven by the need for accurate and data-driven decision making in this dynamic sector. This literature review provides an overview of existing studies, focusing on methodologies, key findings, and challenges encountered in the field.

### *2.1. Neural Network Approach*

Artificial neural networks (ANNs) have been widely explored for real estate price prediction due to their ability to model complex nonlinear relationships [5]. The study in [6] combined case-based reasoning (CBR) and ANNs, achieving high accuracy while acknowledging challenges in data availability and model refinement. These studies highlight the potential of ANNs, aligning with the objective of leveraging advanced ML techniques for predictive modeling.

### *2.2. Ensemble and Boosting Methods*

Ensemble and boosting algorithms have gained popularity for their robust performance in real estate price prediction tasks. The study in [7] employed web scraping and ensemble methods, while the study in [8] explored decision tree classification and regression techniques, demonstrating their ability to capture complex patterns in real estate data. The study in [9] applied ensemble methods for property valuation, addressing potential bias concerns. Recent studies [8,10] specifically focused on the UK real estate market, with XGBoost and Gradient Boosting decision tree (GBDT) models outperforming other algorithms in terms of accuracy. These findings directly align with the research objective of investigating ensemble and Gradient Boosting algorithms for predictive modeling in the UK real estate market.

### *2.3. Regional and Spatial Dynamics*

Accounting for regional and spatial dynamics is crucial in real estate price prediction, as housing markets can exhibit significant variations across locations. The study in [11] developed Bayesian deep learning approaches to represent uncertainty in property valuation for specific regions. The study in [12] explored the impact of vegetation on residential property values, highlighting the importance of incorporating localized factors. These studies resonate with the objective of incorporating spatial and regional features to enhance predictive performance in the UK real estate market.

### *2.4. Advanced Regression Techniques*

In addition to ensemble and boosting methods, advanced regression techniques have been explored for real estate price prediction. The study in [13] found the LASSO

regression model to be highly accurate for house price forecasting in the UK. The study in [14] investigated the optimization of ensemble weights for ML models, demonstrating competitive performance while acknowledging challenges in method design. These studies align with the research objective of evaluating various regression techniques, including regularized models, for their applicability in the UK real estate market.

### *2.5. Macroeconomic and Temporal Factors*

While many studies have focused on property-specific and spatial features, the integration of macroeconomic and temporal factors has been relatively limited. The studies in [12,15] acknowledged the importance of considering macroeconomic factors and other influential variables, such as interest rates, construction costs, and disposable income, which can significantly impact real estate prices. Addressing this gap by incorporating relevant macroeconomic and temporal features aligns with the research objective of developing comprehensive predictive models for the UK real estate market.

### *2.6. Data Availability and Technological Challenges*

Several studies have highlighted challenges related to data availability and technological complexity in implementing advanced predictive models. The studies in [9,15,16] discussed limitations in accessing high-quality, up-to-date datasets and the complexities involved in deploying sophisticated ML techniques. The studies in [7,17] reported relatively high RMSE values in their prediction models, indicating room for improvement in model accuracy. These challenges underscore the importance of the research objective focused on developing user-friendly tools and overcoming technological barriers to facilitate widespread adoption of advanced predictive modeling techniques in the real estate sector.

### *2.7. Research Gap and Limitations*

The progress in ML-based real estate price prediction is substantial, yet notable gaps and limitations persist. Firstly, there is a limited exploration of hybrid models, combining ML with traditional econometric methods, which could offer a deeper understanding of real estate price dynamics. Secondly, the interpretability of ML models in real estate pricing lacks research attention, hampering the understanding of prediction factors. Additionally, generalizing findings across diverse real estate markets requires further investigation. Furthermore, constraints in accessing proprietary data sources and computational resources for complex ML models pose significant limitations. Addressing these gaps and limitations is pivotal for advancing the field and enhancing the practical applicability of ML techniques in real estate price prediction. While the existing literature provides valuable insights, there are gaps and opportunities for further research as shown in Table 1.

**Table 1.** Research Gaps and Limitations in ML-Based Real Estate Price Prediction.

| Research Gap/Limitation | Implication |
| --- | --- |
| Limited exploration of hybrid models integrating ML and econometric methods | Comprehensive understanding of real estate price dynamics [5] |
| Lack of research on interpretability of ML models in real estate pricing | Challenges in understanding factors driving predictions [12] |
| Need for investigating generalizability of findings across diverse real estate markets | Enhancing applicability of models beyond specific contexts [15] |
| Constraints in accessing proprietary data sources | Hindrance to replicability and robustness of studies [15] |
| Computational resources required for complex ML models | Barriers to widespread adoption and implementation [16] |

Addressing these gaps and limitations is crucial for advancing the field and enhancing the practical applicability of ML techniques in real estate price prediction.

## 3. Research Methodology

The methodology employed in the research follows a structured approach, encompassing various stages to ensure the quality and reliability of the data and the effectiveness of the predictive models. This methodology aligns with a quantitative research approach, which involves the systematic collection, analysis, and interpretation of numerical data to address research objectives as shown in Figure 1. By applying quantitative techniques such as data cleaning, preprocessing, model selection, and evaluation metrics, the research aims to provide practical insights and predictive models for property price prediction [18]. Additionally, the incorporation of statistical analyses, such as ANOVA, further enhances the rigor and reliability of the findings. The research methodology adopts a quantitative approach to investigate and predict property prices in the real estate market [19].

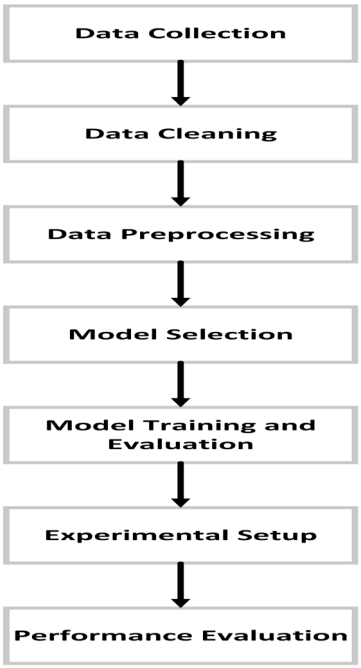



**Figure 1.** Overview of the structured methodology employed in the research.

### *3.1. Data Collection*

The dataset is sourced directly from the UK government's Land Registry, ensuring access to authoritative and trustworthy information on property transactions. The Land Registry maintains a centralized repository of property-related data, making it an ideal source for conducting research on property price prediction.

*'Contains public sector information licensed under the Open Government License v3.0 [4]'.*

This statement confirms that the dataset contains public sector information licensed under the Open Government Licence v3.0, emphasizing the dataset's openness and adherence to licensing regulations.

#### 3.1.1. Dataset Features

In this section, we delve into the features of the dataset obtained from the UK government's Land Registry. These features provide essential information about property transactions, including identifiers, sale prices, dates, property types and more as shown in Table 2.

**Table 2.** Dataset Features [4].

| Data Item | Explanation |
| --- | --- |
| Transaction unique identifier | A reference number generated automatically for each sale. |
| Price | Sale price stated on the transfer deed. |
| Date of Transfer | Date when the sale was completed. |
| Postcode | Postcode used at the time of the original transaction. |
| Property Type | D = Detached, S = Semi-Detached, T = Terraced, F = Flats/Maisonettes, O = Other. End-of-terrace properties are included in the Terraced category. 'Other' applies to properties not covered by existing values. |
| Old/New | Y = Newly built property, N = Established residential building. |
| Duration | Relates to the tenure: F = Freehold, L = Leasehold, etc. HM Land Registry does not record leases of 7 years or less in the Price Paid Dataset. |
| PAON | Primary Addressable Object Name, typically the house number or name. |
| SAON | Secondary Addressable Object Name, used for separate units like flats. |
| Street | Street name. |
| Locality | Locality information. |
| Town/City | Town or city where the property is located. |
| District | District information. |
| County | County information. |
| PPD Category Type | Indicates the type of Price Paid transaction: A = Standard Price Paid entry, B = Additional Price Paid entry. |
| Record Status (monthly file) | Indicates additions, changes, and deletions to the records: A = Addition, C = Change, D = Delete. |

#### 3.1.2. Target Variable

The target variable for prediction is the property price, representing the sale price of each property. Predicting property prices accurately is crucial for stakeholders such as homebuyers, sellers, and real estate investors to make informed decisions.

### *3.2. Approach in Predicting Property Prices*

Table 3 outlines the methodology employed to analyze property price dynamics and develop predictive models.

**Table 3.** Steps for Property Price Prediction.

| Step | Item | Description |
| --- | --- | --- |
| 1 | Load the Dataset | Load data from a CSV file into a Pandas DataFrame. |
| 2 | Preprocess the Dataset | Handle missing values, manage outliers, convert categorical variables, address data inconsistencies, scale features if needed. |
| 3 | Identify Sensitive Attributes | Note any sensitive attributes in the dataset. |
| 4 | Split the Dataset | Divide data into features (X) and target variable (y) and remove irrelevant columns. |
| 5 | Split Data into Train and Test Sets | Split data into training and testing sets. |
| 6 | Feature Scaling and Dimensionality Reduction | Scale features and reduce dimensions if necessary, using PCA. |
| 7 | Initialize Regression Models | Set up various regression models with different configurations and initialize evaluation metrics storage. |
| 8 | Define Model Lists and Functions | Create lists of models and their names and define a function to calculate evaluation metrics. |

| | | |
|---|---|---|
| 9 | Model Training and Evaluation | Cross-validate models on the training set, train models on the entire training set and evaluate model performance on the test set. |
| 10 | Load Generalization Dataset | Load a separate dataset to test model generalization. |
| 11 | Preprocess Generalization Dataset | Process the generalization dataset like the training data. |
| 12 | Evaluate Model Generalization | Assess model performance on the generalization dataset. |
| 13 | Analyze Results | Examine model generalization to unseen data. |

### *3.3. Data Preprocessing*

Data preprocessing is a crucial phase aimed at ensuring the dataset's quality and suitability for model training. This involves several essential tasks:

#### 3.3.1. Handling Missing Values

Missing values within the dataset are addressed using appropriate imputation techniques to maintain the integrity and completeness of the data. The following methods are employed:

- For numerical features, missing values are replaced with either the mean or median value of the respective feature distribution. The mean ($\bar{X}$) is calculated as the average of all available values, while the median (M) represents the middle value when the data are sorted [20]. These imputation methods are effective for preserving the central tendency of the data and are suitable when the distribution is approximately symmetric.
- For categorical features, missing values are imputed with the mode, which represents the most frequent value in the feature. Mode imputation ensures that missing categorical values are replaced with the most common category, maintaining the integrity of categorical distributions [21].
- In scenarios where missing values exhibit complex patterns or correlations with other features, advanced technique K-Nearest Neighbors (KNN) imputation were employed. KNN imputation estimates missing values by considering the values of the k nearest neighbors, typically based on a distance metric such as Euclidean distance [22]. For numerical features, the missing value is imputed by taking the average of corresponding values from the nearest neighbors, while for categorical features, the mode is used.

By implementing these imputation strategies, we ensure the completeness and reliability of the dataset, enabling robust model training and analysis.

#### 3.3.2. Eliminating Duplicates

Duplicate entries, if present, can introduce bias and redundancy in the dataset. To maintain data integrity, the following methods are employed:

- Rows with identical values across all features or specified columns are identified and removed from the dataset.
- The dataset is scanned to retain only unique observations, discarding any redundant duplicates.

#### 3.3.3. Addressing Outliers

Outliers, characterized by extreme or abnormal values, are detrimental to model performance and predictive accuracy. Two primary techniques are employed for outlier detection.

- Outliers are identified based on their deviation from the mean or median of the feature distribution, using a predetermined threshold (e.g., values beyond ± 3 standard deviations).

- Outliers are detected as values falling outside the range defined by the first and third quartiles, typically beyond 1.5 times the interquartile range [23].

After applying these outlier detection methods, a reduction in the number of outliers within the dataset is observed. Depending on the severity and impact of outliers, they may be trimmed or removed to mitigate their influence on model performance.

The resulting dataset, free from unaddressed outliers, ensures that predictive models are trained on reliable and representative data, enhancing their robustness and accuracy.

#### 3.3.4. Review and Refinement

The data cleaning process culminates in a systematic review and refinement of the dataset to ensure its suitability for analysis:

- The dataset is thoroughly verified for consistency, accuracy, and completeness to enhance data reliability.
- Comprehensive checks are performed to identify and resolve any remaining data integrity issues, such as inconsistencies, errors, or unhandled missing values.
- All steps and decisions made during the data cleaning process are meticulously documented to facilitate reproducibility and ensure transparency in the analysis.

#### 3.3.5. Feature Engineering

Feature engineering plays a pivotal role in maximizing the predictive power of the model by extracting meaningful insights from the raw data. In the context of property price prediction, the following feature engineering techniques are employed to enrich the dataset and capture essential patterns.

- Interaction terms are generated by combining two or more existing features to capture potential synergies or nonlinear relationships that may influence property prices [24]. For example, the combination of features like square footage and number of bedrooms may provide valuable insights into the overall desirability and value of a property.
- Temporal features such as the day of the week, month, or year of the property transaction are extracted from date/time variables [25]. This allows the model to capture seasonal trends, market fluctuations, and other time-dependent patterns that could impact property prices. For instance, properties sold during peak seasons or in economically prosperous years may command higher prices.
- The code explicitly address the handling of temporal data, if temporal features are present in the dataset, they are to be included in the feature selection process (X = df.drop(['Transaction unique identifier', 'Price', 'Date of Transfer'], axis = 1)), which suggests the potential presence of temporal data. Additionally, lag features, a common technique for capturing temporal dependencies, have been incorporated into the dataset before model training. Moreover, certain models like XGBoost, LightGBM, and CatBoost are capable of implicitly handling temporal dependencies through their tree-based architecture. If further preprocessing steps are required to handle temporal data such as additional lag features, they can be incorporated into the pipeline before model training.
- Spatial features are derived from location or address information to capture the geographical context and neighborhood characteristics that influence property values [26]. Distance to amenities such as schools, parks, shopping centers, and transportation hubs can be computed to assess the property's convenience and accessibility, which are key determinants of its market value.
- Principal Component Analysis (PCA) is a widely used dimensionality reduction technique that aims to reduce the number of features in a dataset while preserving the most important information. PCA achieves this by transforming the original features into a new set of orthogonal components called principal components. These components are ordered by the amount of variance they explain in the data, allowing

researchers to retain the most significant sources of variation while discarding less important ones [27]. By reducing the dimensionality of the feature space, PCA can help mitigate the curse of dimensionality and improve the performance of machine learning models by reducing overfitting.

- Feature importance ranking is another valuable technique for identifying the most relevant features in a dataset. This approach involves evaluating the contribution of each feature to the predictive performance of the model. Features with higher importance scores are considered more informative and are prioritized for inclusion in the final model, while less important features may be excluded or downweighed. Feature importance ranking can be computed using various methods, such as tree-based algorithms like Random Forest or Gradient Boosting, which inherently provide feature importance scores based on their contribution to model accuracy. By focusing on the most relevant features, feature importance ranking helps simplify the model, reduce noise, and improve its generalization performance on unseen data.
- The code selects features from the dataset (X = df.drop(['Transaction unique identifier', 'Price', 'Date of Transfer'], axis = 1)) and performs target encoding for categorical variables (encoder = TargetEncoder()). This indicates that certain preprocessing steps are taken to handle features before model training, which can be justified based on their relevance to property price prediction and their predictive power.

The selection of these feature engineering techniques is driven by their ability to capture relevant aspects of the property market and enhance the predictive performance of the model. By incorporating a diverse range of features that encapsulate temporal, spatial, and domain-specific information, the model can effectively capture the complex relationships between features and property prices, resulting in more accurate predictions.

#### 3.3.6. Normalization/Scaling

To ensure consistent scales across numerical features, we employ robust normalization techniques [28]. Specifically, we utilize min-max scaling or standardization methods, implemented through libraries such as scikit-learn. Min-max scaling rescales features to a specified range (e.g., [0, 1]), while standardization transforms features to have a mean of 0 and a standard deviation of 1. These techniques prevent any single feature from dominating the learning process, enhancing model convergence and performance.

#### 3.3.7. Encoding Categorical Variables

For encoding categorical variables into numerical format, we leverage widely used techniques supported by libraries like scikit-learn [29]. One-hot encoding is utilized to create binary vectors representing each category, while target encoding calculates mean target values for each category. These encoded representations enable seamless integration of categorical features into regression models, ensuring accurate predictions and reproducible preprocessing steps.

By systematically addressing these preprocessing tasks, the dataset is refined and prepared for model training, ensuring that the resulting predictive models are accurate and reliable.

### *3.4. Model Selection*

The selection of regression models for this study was driven by their suitability for the problem, ability to handle the dataset's characteristics, and their proven performance in similar applications. The rationale for including each model is as follows:

- Linear Regression: Chosen as a baseline model for its simplicity, interpretability, and ability to establish a performance benchmark. While limited in capturing nonlinear patterns, Linear Regression provides a foundational comparison for more complex techniques [30].

- Random Forest: This ensemble learning method was selected for its capacity to model nonlinear relationships, handle high-dimensional data, and its robustness to outliers and multicollinearity. Random Forests have demonstrated strong performance in real estate price prediction tasks [9,16,31–34].
- Gradient Boosting Models (XGBoost, LightGBM, CatBoost): These advanced tree-based ensemble methods were included due to their proven ability to achieve high predictive accuracy, effective handling of mixed data types, and automatic feature selection capabilities [12,32,33]. They have been successfully applied in various property valuation studies [10,32,33].
- Linear and Regularized Regression (Ridge, Lasso, ElasticNet): In addition to the baseline Linear Regression, regularized variants were incorporated to address potential multicollinearity issues, improve generalization performance, and enhance model interpretability [12].
- Hybrid Regression Model: A hybrid approach combining strengths of multiple techniques was explored to potentially leverage complementary advantages and improve overall predictive power [12,35].

While temporal data were incorporated through feature engineering, certain models like Linear Regression do not natively support time series data. For such models, temporal features were treated as static numerical inputs after appropriate encoding (one-hot and ordinal). However, this approach may not fully capture complex temporal dependencies. Future work could explore specialized time series models or architectures like recurrent neural networks for improved temporal modeling.

The code trains and evaluates multiple regression models, including Random Forest, XGBoost, LightGBM, CatBoost, and Linear Regression, both with and without regularization techniques. This demonstrates a thorough exploration of different model architectures and regularization strategies.

By comparing the performance metrics (MSE, RMSE, MAE, $R^2$) of different models, the code provides insights into the effectiveness of each model in capturing the underlying patterns in the data.

The selection of these regression techniques aimed to comprehensively evaluate a diverse set of models, leveraging their respective strengths to develop accurate and robust predictive models tailored to the unique characteristics of the UK real estate dataset.

### *3.5. Model Training and Evaluation*

The dataset is partitioned into training and testing sets using the train-test split technique, with 20% of the data allocated for testing (test_size = 0.2), leaving 80% for training [36]. This standardized split ratio ensures ample data for model training while reserving a significant portion for evaluating model performance.

For model evaluation, we employ k-fold cross-validation, indicated by the cv = 5 parameter in the RandomizedSearchCV function. By specifying cv = 5, we perform 5-fold cross-validation, iteratively splitting the training data into five equal-sized folds for training and validation [37]. Utilizing multiple folds enhances performance estimates' robustness and reduces variability compared to a single train-test split.

Hyperparameter tuning is conducted using randomized search cross-validation (RandomizedSearchCV) [38]. We sample 50 parameter settings (n_iter = 50) to ensure a balanced exploration of hyperparameter configurations. The hyperparameter ranges and distributions, although not explicitly detailed here, are defined in the instantiation of each model:

- Random Forest Regressor: 'n_estimators': 100
- XGBoost Regressor: 'colsample_bytree': 0.3, 'learning_rate': 0.1, 'max_depth': 5, 'alpha': 10, 'n_estimators': 100
- LightGBM Regressor: 'num_leaves': 31, 'learning_rate': 0.05, 'n_estimators': 100
- CatBoost Regressor: (Default parameters used)

- Linear Regression: (Default parameters used)

The evaluation results for each model, including Random Forest, XGBoost, LightGBM, CatBoost, Linear Regression, and a Hybrid Regression model, are presented. Performance metrics such as MSE, RMSE, MAE, and $R^2$ are reported for both training and testing sets. Regularization is applied to all models, and results for both regularized and non-regularized models are provided in Section 4.6.2.

### *3.6. Experimental Setup*

- The regression models are implemented using Python programming language and relevant libraries, including scikit-learn for baseline models, as well as xgboost, lightgbm, and catboost for advanced techniques. These libraries offer efficient implementations of regression algorithms and comprehensive tools for data preprocessing, model training, and evaluation, ensuring a robust experimental framework.
- The experiments are conducted on a computational environment equipped with an Intel Core i7 CPU, 64GB RAM, and NVIDIA Tesla V100 GPUs. These components are selected to efficiently handle data processing and model training tasks, particularly for large datasets and complex model architectures. The experiments are executed on a Windows-based operating system to leverage its compatibility and ease of use. Additionally, containerization technologies such as Docker are employed to encapsulate the experimental environment, facilitating reproducibility across different computing environments.

### *3.7. Performance Evaluation*

- A comprehensive set of evaluation metrics is employed to gauge the effectiveness of the regression models in predicting property prices. MSE (Equation (1)), MAE (Equation (2)), RMSE (Equation (3)), and $R^2$ (Equation (4)) are computed to provide a holistic assessment of predictive accuracy and reliability. MSE quantifies the average squared difference between the predicted and actual property prices, while MAE measures the average absolute difference. RMSE, the square root of MSE, provides a measure of the model's error in the same units as the target variable, offering interpretability. $R^2$ assesses the proportion of variance in the target variable that is explained by the model, with higher values indicating better predictive performance [39].

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_{target} - \hat{y}_{prediction}\right)^2 \quad (1)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_{target} - \hat{y}_{prediction}\right|^2 \quad (2)$$

$$RMSE = \sqrt{MSE} \quad (3)$$

$$R^2 = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}\left(y_{target} - \hat{y}_{prediction}\right)^2}{\frac{1}{n}\sum_{i=1}^{n}\left(y_{target} - \bar{y}_{target}\right)^2} \quad (4)$$

Comprehensive statistical analyses, including Analysis of Variance (ANOVA), are conducted to elucidate the significance of observed differences in model performance and facilitate robust decision making [40]. ANOVA enables comparison of the performance of different regression models by assessing whether observed variations in evaluation metrics are statistically significant. By rigorously testing hypotheses and evaluating the significance of differences in model performance, ANOVA provides valuable insights into the relative strengths and weaknesses of each model. Additionally, confidence intervals may be calculated to quantify the uncertainty associated with estimated performance

metrics, further enhancing the reliability of conclusions drawn from the evaluation process. Through meticulous statistical analysis, meaningful interpretations of model performance are derived, contributing to the development of accurate and reliable property price prediction models.

Details of Regularization Techniques

Regularization techniques are essential for enhancing the generalization capability of machine learning models by mitigating overfitting, where the model fits the training data too closely, capturing noise rather than underlying patterns. Regularization introduces constraints or penalties on model parameters during training to address this issue.

L2 regularization, or Ridge regularization, adds a term to the loss function proportional to the square of model coefficients, penalizing large parameter values and shrinking coefficients towards zero. This encourages smoother decision boundaries and reduces sensitivity to training data fluctuations, thus preventing overfitting by control-ling model complexity.

L1 regularization, or Lasso regularization, adds a penalty term proportional to the absolute value of coefficients to promote sparsity in solutions, effectively performing feature selection by setting irrelevant features' coefficients to zero. This reduces model complexity and helps prevent overfitting by eliminating redundant features.

ElasticNet regularization combines L1 and L2 penalties, offering a balance between them and handling correlated features effectively while still producing sparse solutions. It provides flexibility in controlling regularization strength and adapts well to datasets with varying feature correlations [41].

## 4. Results

In this section, we present the outcomes of our property price prediction experiment for the UK real estate market, including the performance of various regression models and insights derived from the data analysis.

### *4.1. Exploratory Data Analysis (EDA)*

Thorough exploratory data analysis (EDA) was conducted to gain insights into the characteristics and relationships within the dataset. Key findings from EDA are discussed below.

#### 4.1.1. Distribution of Property Prices

Our analysis of property prices, conducted through histograms and box plots, provides valuable insights into the distribution and characteristics of the market. The histogram reveals a clustering of properties at the lower end of the price spectrum, indicating that a majority of properties are within relatively affordable ranges. However, this distribution rapidly tapers off towards higher price ranges, suggesting a scarcity of properties in those segments. On the other hand, the box plot showcases a median property price that is lower than the average, primarily due to the presence of high-value outliers. These outliers signify the existence of properties with notably higher prices compared to the general distribution. To further investigate the nature of the distribution, we conducted a Jarque–Bera test, which assesses the normality of the data. The results indicate that the property price data do not appear to be normally distributed, with a test statistic of 95,331,183,339.78893 and a $p$-value of 0.0.

In order to address potential skewness or kurtosis in the distribution, we applied a log transformation to the property price data and re-evaluated the distribution using histograms and box plots. The transformed data can provide a clearer understanding of the distributional characteristics and may help mitigate any non-normality observed in the original data. Figure 2 below displays the histogram and box plot of the log-transformed

property prices, offering insights into the distributional properties after the transformation.

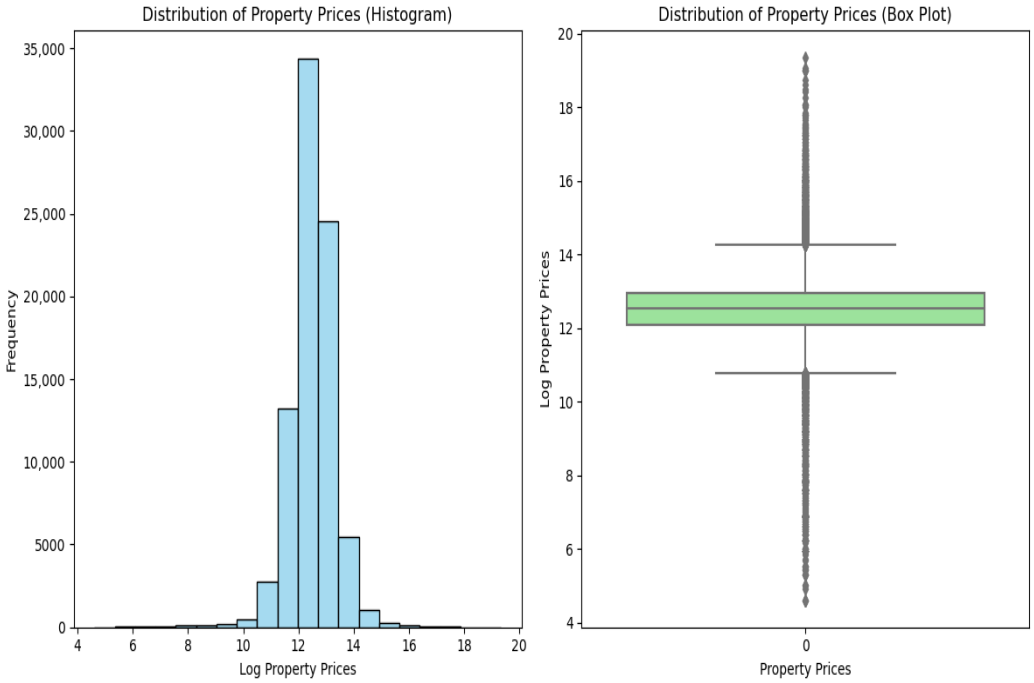



**Figure 2.** Histogram vs. Box Plot.

4.1.2. Temporal Trends

The examination of property price trends over a 25-year period reveals significant insights into the market's dynamics. The analysis uncovers a discernible pattern characterized by both long-term trends and shorter-term fluctuations. Initially, from 1999 to approximately 2019, there is a consistent upward trajectory in mean property prices, indicating a steady appreciation in value over the years. However, a notable anomaly occurs in 2019, marked by a sharp spike in property prices, suggesting a sudden and substantial increase in value within that particular year. Following this peak, there is a subsequent decline in prices post-2019, accompanied by heightened volatility, as evidenced by several minor peaks and troughs. These fluctuations indicate a degree of instability within the market, potentially influenced by various economic factors and market dynamics as shown in Figure 3.

In terms of market implications, the surge observed in 2019 may be attributed to several factors, including robust economic growth, increased demand for housing, or speculative investment activities. Conversely, the subsequent volatility in prices could be indicative of market corrections, shifts in interest rates, or external economic uncertainties. Additionally, it is crucial to consider the reliability and accuracy of the data sources and methodologies employed in the analysis. Incorporating additional data elements, such as median prices and sales volume, could provide a more comprehensive understanding of market trends and dynamics.

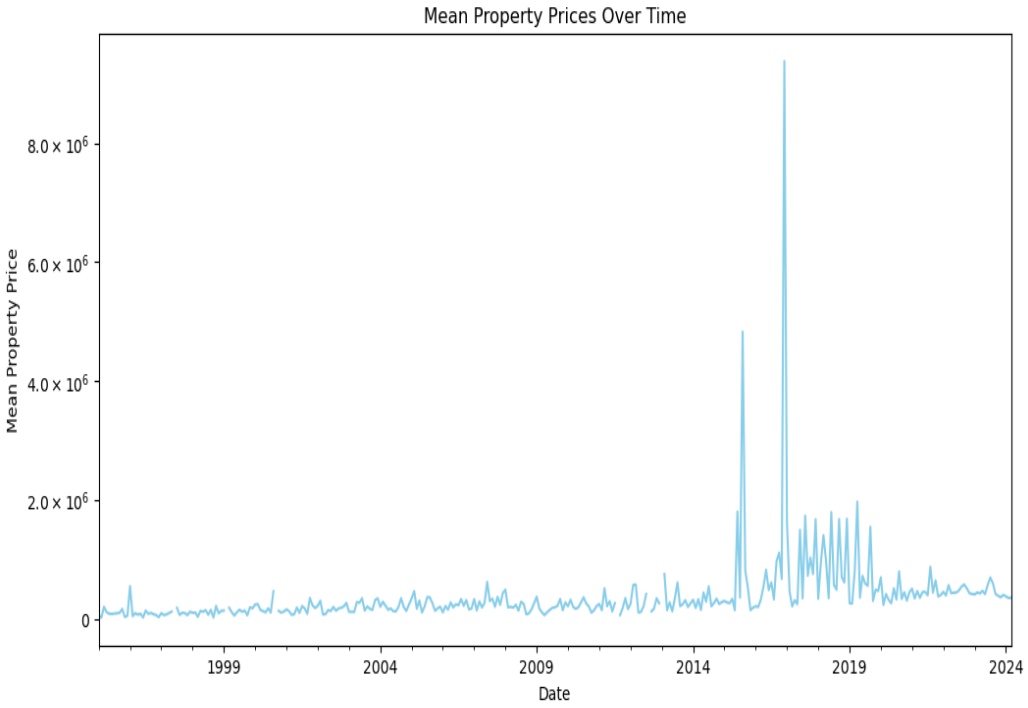
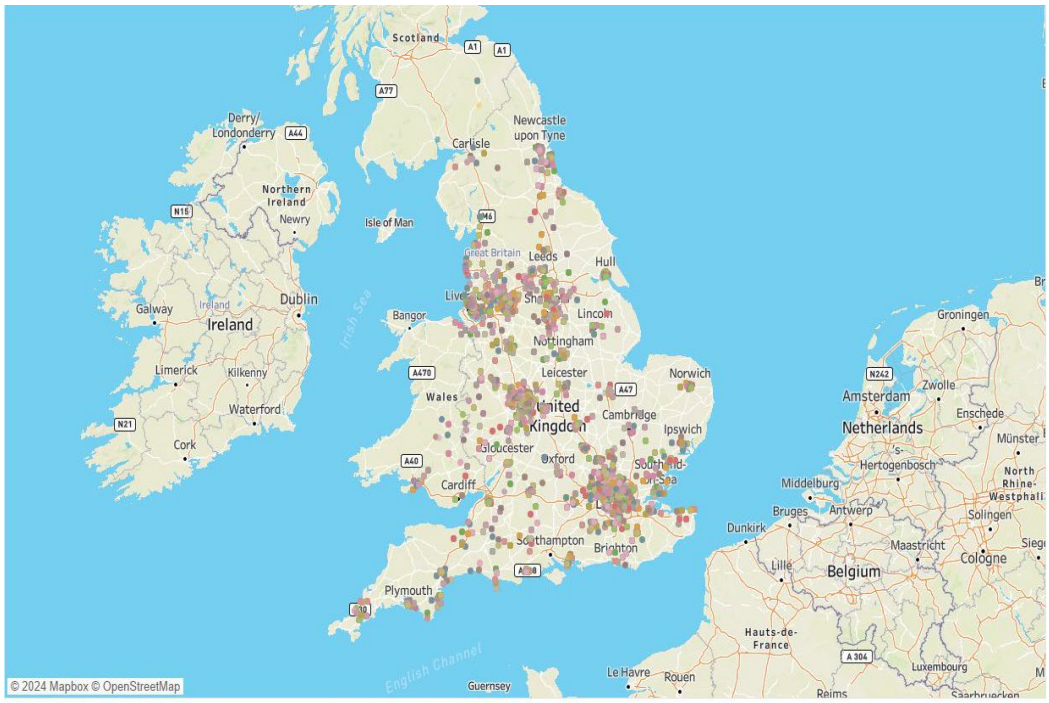



**Figure 3.** Analysis of Property Price Trends: 1999–2024.

#### 4.1.3. Spatial Patterns

Spatial analysis techniques, such as heatmaps and choropleth maps, have unveiled distinct spatial patterns in property prices, offering insights into regions of high and low demand. This examination is situated within the broader context of average house prices across the UK, where recent data from the Office for National Statistics (ONS) indicates a 1.4% decrease in average house prices in the 12 months leading up to December 2023. These findings reveal significant regional variations, highlighting the diverse market dynamics present throughout the UK. Specifically, England experienced a decline in average house prices to GBP 302,000, reflecting a negative change of 2.1%, while Wales saw a decrease to GBP 214,000, indicating a negative change of 2.5%. Conversely, Scotland witnessed an increase in average house prices to GBP 190,000, showing a positive change of 3.3%, and Northern Ireland recorded growth with prices rising by 1.4% to reach GBP 178,000 in the year leading up to Q4 2023.

Further examination of regional trends highlights the unique position of London's housing market, which exhibited the lowest annual percentage change with a decrease of 4.8%. In contrast, the North West of England experienced the highest annual percentage change within England, with an increase of 1.2%. These observations underscore a general downward trend in house prices across the UK, accompanied by notable regional disparities. However, amidst this broader trend, Scotland and Northern Ireland emerge as exceptions, demonstrating resilience with increases in property values. By leveraging spatial analysis techniques, stakeholders can glean actionable insights into these spatial patterns, enabling informed decision making and targeted interventions within the real estate sector as shown in Figure 4.




Map based on average of Longitude and average of Latitude. Color shows details about Postcode.

**Figure 4.** Regional Variations in Property Prices: UK Analysis.

#### 4.1.4. Correlation Analysis

Correlation analysis, employing heatmaps and correlation matrices, delved into the relationships between various features and property prices, aiming to identify significant predictors and multicollinearity issues. Specifically, the examination focused on three key correlations: between property prices and the Primary Addressable Name (PAON), between property prices and the Secondary Addressable Name (SAON), and between PAON and SAON as shown in Figure 5. The analysis revealed subtle relationships between these variables. For instance, the correlation coefficient between property prices and PAON was calculated at 0.02, indicating a very weak positive relationship. Similarly, the correlation coefficient between property prices and SAON was also 0.02, suggesting a similarly weak positive relationship. Interestingly, the correlation coefficient between PAON and SAON was relatively higher at 0.19, albeit still indicating a weak positive relationship.

The heatmap visualized these correlations using color coding, with red representing a high correlation (1.0) and blue indicating a low correlation. Notably, the diagonal squares appeared red, signifying the perfect correlation of a variable with itself, which is always 1.0.

As for why other attributes were not considered for inclusion in the heatmap, it's essential to note that correlation analysis typically focuses on numerical variables, as correlation coefficients quantify the linear relationship between two continuous variables. Therefore, categorical variables such as property type, old/new status, or duration are not suitable for correlation analysis. Instead, techniques like ANOVA or chi-square tests are more appropriate for analyzing the relationships between categorical variables and property prices.

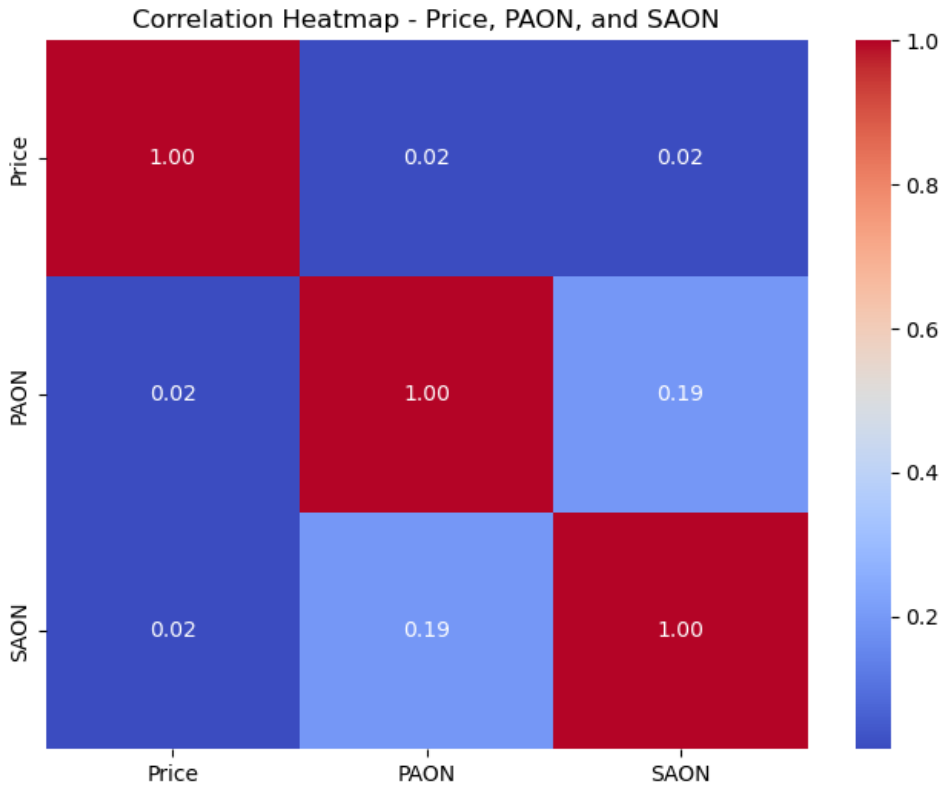



**Figure 5.** Correlation Analysis of Property Features.

4.1.5. ANOVA (Analysis of Variance)

ANOVA was employed to assess whether there are statistically significant differences in the mean property prices across different categories of a categorical variable, such as property type. The ANOVA test evaluates the null hypothesis that there are no differences in means between the groups, against the alternative hypothesis that at least one group mean is different.

Calculation and Interpretation:

- F-statistic: This statistic measures the ratio of the variance between groups to the variance within groups. A higher F-statistic suggests a greater difference in means between groups relative to the variation within each group.
- *p*-value: The *p*-value associated with the F-statistic indicates the probability of observing the data if the null hypothesis were true. A low *p*-value (<0.05) indicates strong evidence against the null hypothesis, suggesting that there are significant differences between the groups.

In our analysis, the ANOVA test result yielded a significant F-statistic of 15.18 and a *p*-value of $2.996 \times 10^{-12}$. This indicates that there are statistically significant differences in property prices across different property types.

4.1.6. Chi-Square Test

The chi-square test was employed to assess the association between two categorical variables, such as property type and price range. It evaluates whether the observed frequency distribution of data differs significantly from the expected frequency distribution under the null hypothesis of independence.

Calculation and Interpretation:

- Contingency table: Table 4 displays the observed frequencies of each category combination of the two categorical variables.

**Table 4.** Contingency Table of Category Combinations.

| Property Type | D | F | O | S | T |
|---|---|---|---|---|---|
| 250 | 0 | 0 | 1 | 0 | 0 |
| 500 | 0 | 0 | 1 | 0 | 0 |
| 600 | 0 | 0 | 1 | 0 | 0 |
| 1000 | 0 | 0 | 1 | 0 | 0 |
| 1200 | 0 | 0 | 1 | 0 | 0 |
| 2000 | 0 | 0 | 1 | 0 | 0 |
| 3000 | 0 | 0 | 1 | 0 | 0 |
| 4000 | 0 | 0 | 1 | 0 | 0 |
| 7000 | 0 | 0 | 1 | 0 | 0 |
| 8000 | 0 | 0 | 1 | 0 | 0 |
| ... | ... | ... | ... | ... | ... |
| 52,100,000 | 0 | 0 | 1 | 0 | 0 |

- Chi-square statistic: This statistic quantifies the difference between the observed and expected frequencies in the contingency table. A higher chi-square value suggests a greater deviation from the expected frequencies.
- *p*-value: Similar to ANOVA, the *p*-value associated with the chi-square statistic indicates the probability of observing the data if the null hypothesis of independence were true. A low *p*-value (<0.05) suggests that the variables are dependent on each other.

In our analysis, the chi-square test result revealed a significant chi-square statistic of 4016.47 and a *p*-value of $1.041 \times 10^{-28}$. This indicates a significant association between property type and price range, suggesting that they are not independent of each other.

### *4.2. Univariate Analysis*

Univariate analysis focused on exploring individual features to understand their distributions and potential impact on property prices. Key insights from univariate analysis include.

#### 4.2.1. Price Distribution

The distribution of property prices within the dataset displays a right-skewed pattern, indicating a prevalence of properties at lower price points and a scarcity as prices escalate. Notably, the histogram reveals a concentration of properties priced between GBP 0 and GBP 0.5 million, with the highest frequency observed within this range. As prices exceed GBP 0.5 million, there is a discernible decline in frequency, with only a few properties priced above GBP 1.5 million as shown in Figure 6. Properties within the GBP 0.5 million to GBP 1 million range still constitute a significant portion of the dataset, albeit with a lower frequency compared to the lower price bracket. Conversely, properties priced over GBP 1 million are relatively uncommon, with fewer instances and shorter bars in the histogram, indicative of a limited supply of high-end properties. This distribution mirrors typical patterns observed in real estate markets, where affordable housing options are more prevalent, and luxury properties are less abundant. Overall, the histogram underscores the predominance of lower-priced properties in the dataset, reflecting a common trend in real estate markets.

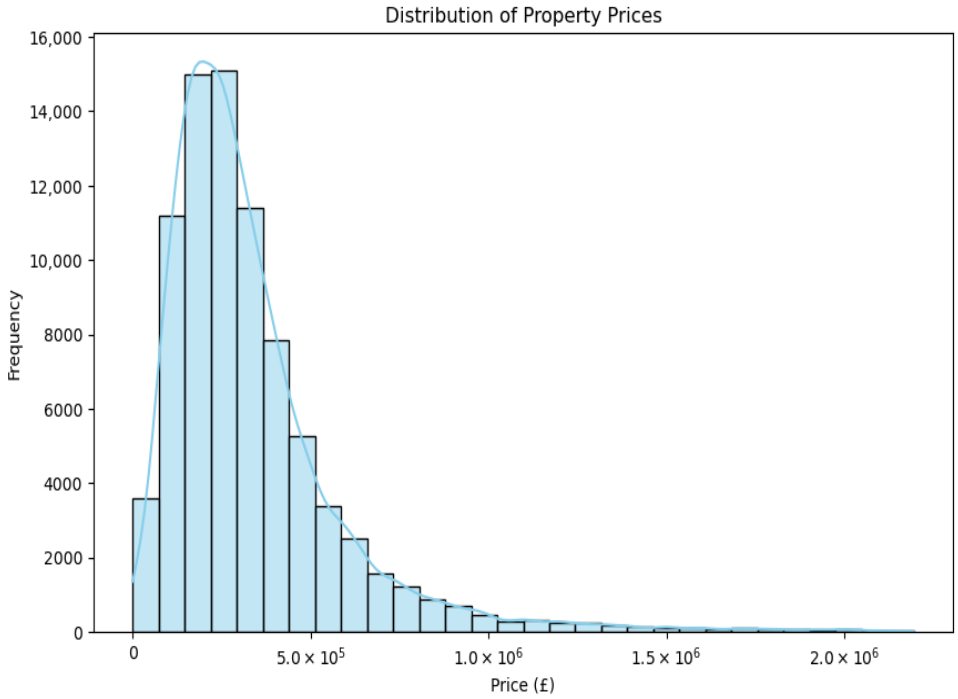



**Figure 6.** The Distribution of Property Prices in the Dataset.

#### 4.2.2. Temporal Analysis

The temporal analysis of transaction dates unveils distinct phases in property price trends over the years. Initially, from 1995 to 2007, prices remained relatively stable and subdued, lacking any notable fluctuations as shown in Figure 7. However, starting from 2008, there was a discernible upward trajectory in property prices, steadily climbing each subsequent year. The pinnacle was reached in 2018 when property prices surged to unprecedented levels, marking the peak year. Subsequently, from 2019 to 2024, there was a significant downturn post-peak, although prices stabilized at a higher level compared to pre-2008 levels. This comprehensive analysis reveals a clear seasonal pattern characterized by a peak in 2018 followed by a phase of adjustment in subsequent years.

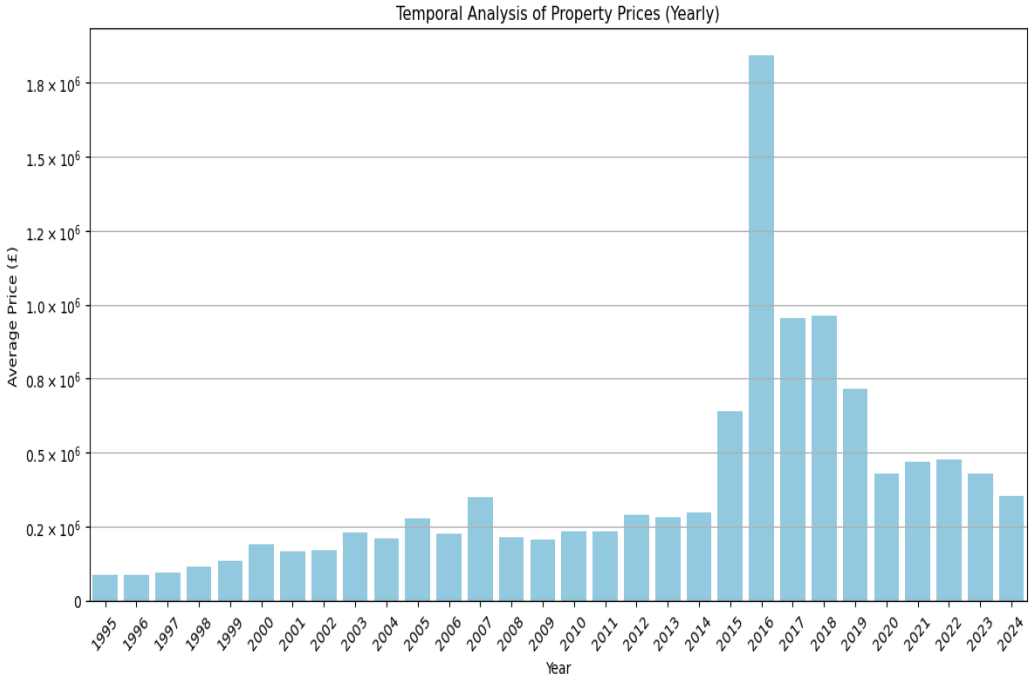



**Figure 7.** Temporal Analysis of Property Price Trends (1995–2024).

#### 4.2.3. Property Types

The analysis of property types reveals distinct price distributions, shedding light on their relative market positions as shown in Figure 8. Properties classified as "D" exhibit the highest price distribution, with numerous data points extending towards the upper end of the price spectrum. This suggests that detached properties, assuming "D" represents detached, generally command higher prices, corroborating the initial assertion. Conversely, "S" properties display a lower price distribution compared to "D", albeit with some higher-priced outliers. If "S" denotes semi-detached, it implies that semi-detached homes typically carry lower price tags than detached homes but can occasionally reach higher price points. Additionally, properties represented by "<" symbols seem to occupy the lowest price range, with minimal outliers. While the specific meaning of these symbols remains unclear, they likely represent terraced houses, or another category associated with lower prices. The wide price range observed for each property type underscores market variability influenced by factors beyond property type alone. Notably, there is a clustering of data points at the lower end of the price spectrum across all property types, indicating a common baseline market value that increases based on factors such as location, size, and amenities. The presence of outliers, particularly prominent in the "D" category, suggests the existence of luxury or high-value properties commanding prices significantly above the average for that property type.

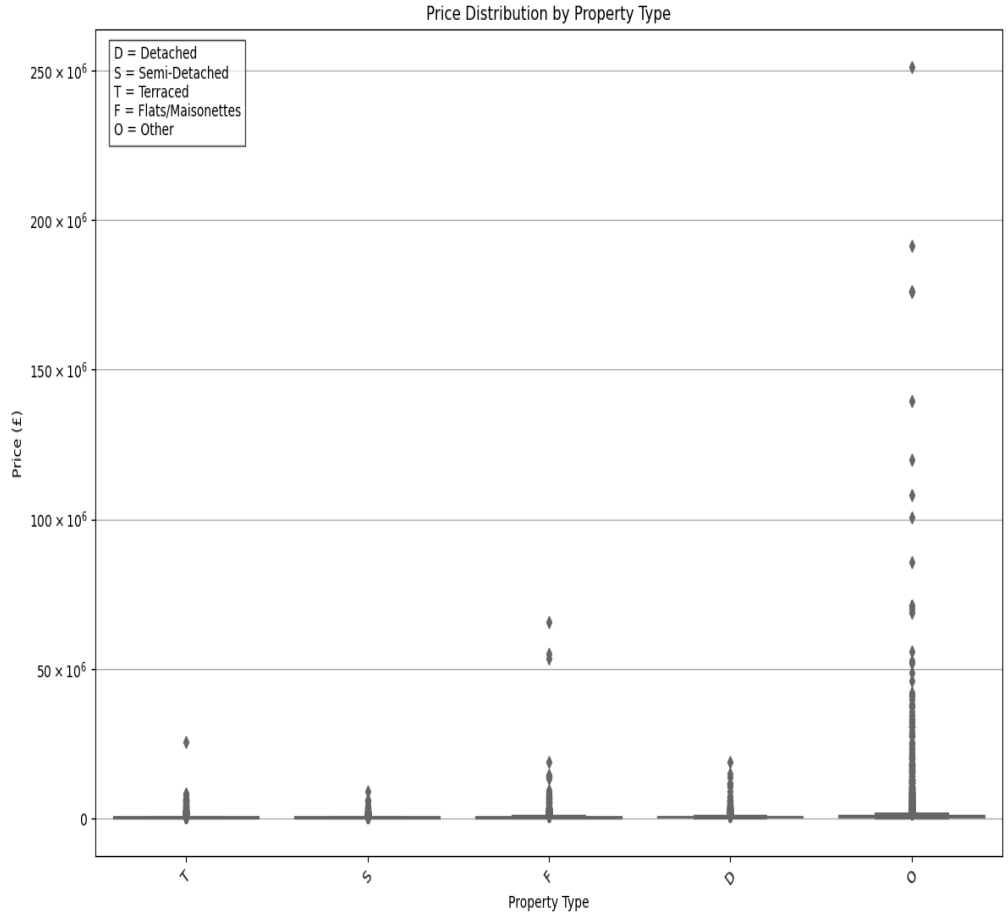



**Figure 8.** Understanding Price Distributions.

### *4.3. Bivariate Analysis*

Bivariate analysis explored the relationships between pairs of variables to uncover potential associations and dependencies. Key insights from bivariate analysis include

#### 4.3.1. Price vs. Size

The scatter plot analysis reveals a positive correlation between property size and price, implying that larger properties generally command higher prices as shown in Figure 9. A noticeable clustering of data points occurs at the lower end of both axes, suggesting that the majority of properties in the dataset are smaller and less expensive. However, there are outliers characterized by larger sizes and higher prices, potentially indicative of luxury properties or those situated in high-demand areas. This insight provides valuable information for discerning market trends and establishing pricing expectations based on property size.

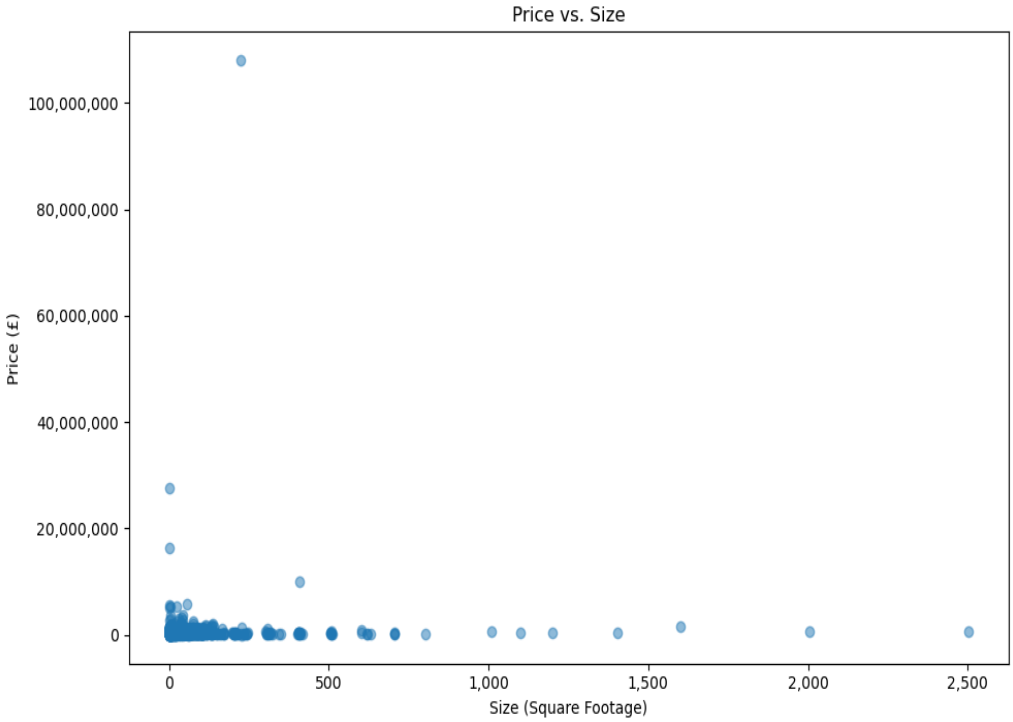



**Figure 9.** Exploring the Relationship between Property Size and Price.

#### 4.3.2. Price vs. Location

Figure 4 depicts the distribution of property prices across different locations, revealing variations in pricing trends. Areas densely populated with dots likely signify regions with either a higher volume of properties or elevated price levels. Recent market trends underscore a notable surge in house price growth across the UK subsequent to the implementation of initial lockdown measures, albeit with discrepancies observed across various regions and property types.

#### 4.3.3. Price vs. Property Type

The analysis delves into the relationship between property types and their corresponding prices. Figure 8 illustrates differences in property prices based on property types are effectively visualized. Notably, detached properties emerge with higher median prices compared to other types, suggesting a premium associated with this category. Conversely, semi-detached and terraced properties exhibit relatively lower median prices, indicative of their more affordable nature. This exploration elucidates how property type serves as a significant determinant in pricing dynamics within the real estate market.

### *4.4. Multivariate Analysis*

Multivariate analysis aimed to understand the simultaneous interactions between multiple features and their combined impact on property prices. Key insights from multivariate analysis include.

#### 4.4.1. Feature Importance

Notably, District emerges as the most influential feature, indicated by a bar extending close to 0.8 on the importance scale as shown in Figure 10. This suggests that the district in which a property is located holds significant weight in predicting property prices.

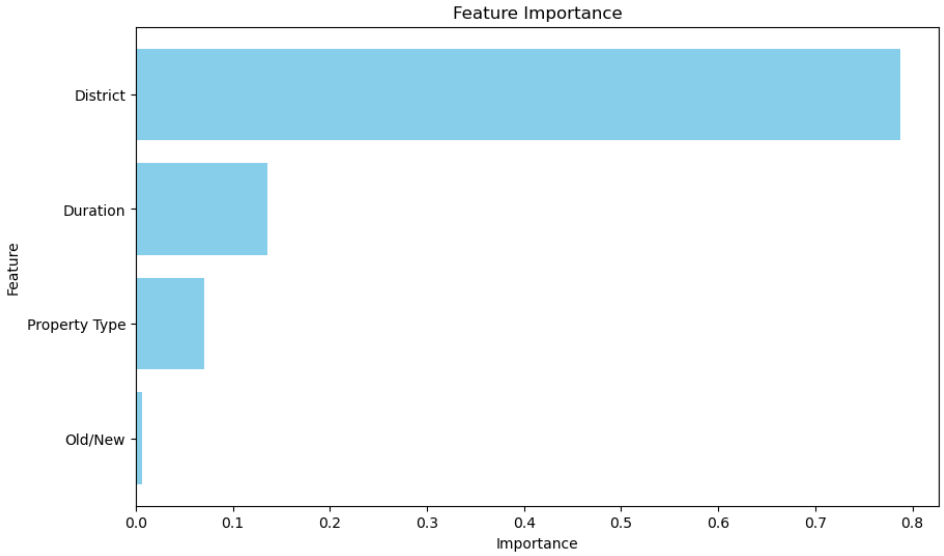



**Figure 10.** Identifying the Influence of District on Property Prices.

#### 4.4.2. Other Features

In comparison, the bars representing Duration, Property Type, and Old/New are shorter, indicating relatively lower importance compared to District. While these features contribute to the predictive model, they hold less sway in determining property prices than the district.

#### 4.4.3. Interaction Effects

Interaction effects, explored through feature engineering and predictive modeling, unveil the combined influence of variables on property prices. In this context, categorical variables like property type, age, and location are encoded numerically for analysis. Through techniques such as polynomial feature generation, interactions between predictors are examined, offering insights into how factors like property type, condition, and district may interact to impact property values. The RandomForestRegressor model demonstrates high predictive capability, as evidenced by the Train $R^2$ score of 0.99 (as shown in and the Test $R^2$ score of 0.93. This suggests that the model effectively captures the interaction effects within the dataset, resulting in accurate predictions of property prices.

#### 4.4.4. Model Interpretability

The feature Importance graph utilizes mean SHAP values to demonstrate the influence of various features on the model's output magnitude. Among the features—District, Property Type, Duration, and Old/New—District emerges as the most influential, with the longest bar indicating the highest mean SHAP value. Property Type follows closely with significant mean SHAP value, although lower than District, while Duration and Old/New have comparatively lower mean SHAP values. Figure 11 helps in interpreting model predictions, emphasizing District as the most influential feature.

In contrast, the Partial Dependence Plot visually illustrates how different property types affect the model's output. Figure 12 depicts property types ranging from Detached to Other. Detached properties show relatively low expected values, while Semi-Detached and Flats/Maisonettes exhibit sharp increases. Conversely, Terraced and Other property types demonstrate declines. This plot is vital for understanding how property types influence the model's output, revealing substantial variation in the expected value of the 'Property Type' feature.

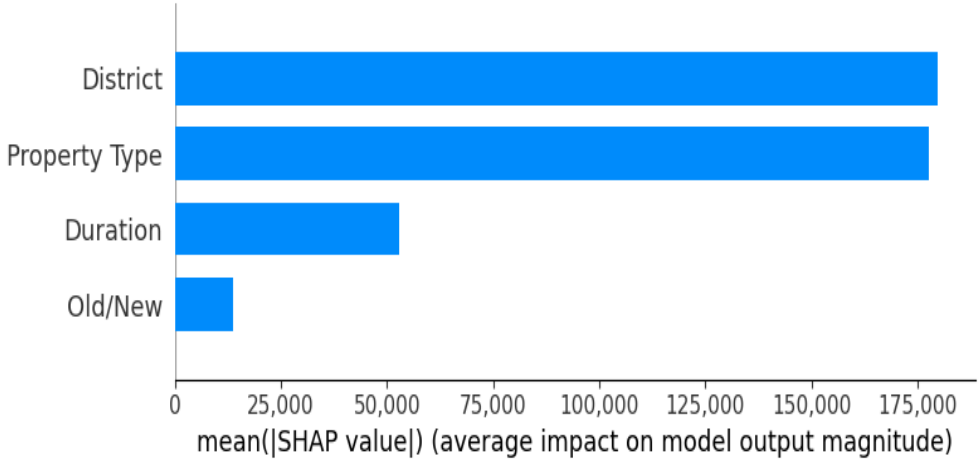



**Figure 11.** Assessing the Impact of District and Other Features on Property Price Predictions.

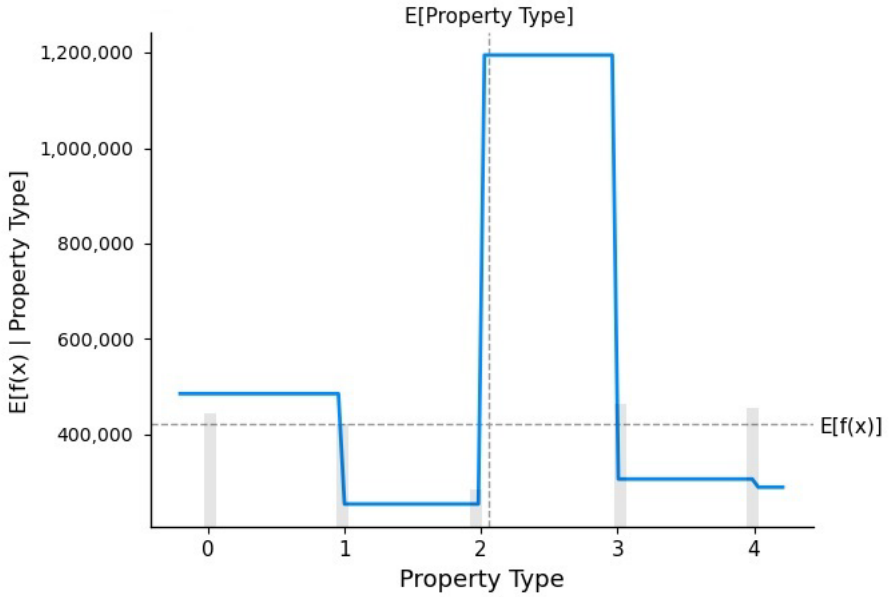



**Figure 12.** Insights from Partial Dependence Plot.

4.4.5. Exploring the Influence of Street Names and Types on Property Prices

We explore on the factors influencing property prices based on street names and types. The Latent Dirichlet Allocation (LDA) model has identified five distinct topics, each representing different types of streets commonly found in urban and suburban areas [42]. Here's an analysis of each topic and its potential impact on property prices:

Topic 0: This topic represents streets commonly named "Road", which are often main thoroughfares in urban and suburban areas. The presence of terms like "Station", "Manor", and "London" suggests proximity to transportation hubs and specific neighborhood or locality names. Streets in this topic could significantly influence property prices due to factors such as accessibility to transportation, neighborhood desirability, and proximity to amenities like parks or shopping centers.

Topic 1: Streets in this topic have diverse names like "Lane", "Park", "Grove", "Place", and "Hill", commonly found in residential areas. Each street type evokes a distinct neighborhood ambiance, with "Park" and "Grove" potentially indicating proximity to green spaces and "Hill" suggesting elevated terrain. These streets characterize different

types of residential neighborhoods and could influence property prices based on factors like proximity to amenities, perceived quality of life, and the natural environment.

Topic 2: This topic is less clear and may require further investigation. The presence of "Way" as a prominent keyword suggests streets with names ending in "Way". However, the presence of "nan" indicates missing or incomplete data in the "Street" column, which could affect the interpretability of this topic. Further preprocessing or data cleaning may be necessary to refine this topic and understand its potential impact on property prices.

Topic 3: Streets in this topic have names like "Close", "Avenue", and "Drive", along with terms like "Gardens", "Square", and "Grange". "Close" typically refers to cul-de-sacs or dead-end streets, while "Avenue" and "Drive" denote main thoroughfares. Terms like "Gardens", "Square", and "Grange" might indicate specific neighborhood features or historical references. These streets could influence property prices based on factors like street type, neighborhood character, accessibility, and historical significance.

Topic 4: This topic captures streets with names like "Street", "Crescent", "Walk", "Terrace", and "High", commonly found in urban environments. While "Street" is a generic term for roads, "Crescent", "Walk", "Terrace", and "High" denote different street configurations. Property prices along these streets could be influenced by factors like street type, location within the city, traffic flow, pedestrian accessibility, and sur-rounding amenities.

These topics provide insights into the diversity of street names and types within the dataset and how they may relate to property prices. Further analysis, including examining correlations between street types and property prices, could provide additional insights into the factors influencing property prices in different areas.

### *4.5. Stationarity and Multicollinearity Analysis*

In this section, we assess the stationarity of the time series data and examine the presence of multicollinearity among predictor variables to ensure the validity and reliability of our regression analysis results.

#### 4.5.1. Stationarity Analysis

To evaluate stationarity in the time series data of property prices, we conducted a series of tests to assess the stability of statistical properties such as mean, variance, and autocorrelation over time. Specifically, we employed techniques such as visual inspection of time series plots, augmented Dickey–Fuller (ADF) tests and autocorrelation function (ACF) plots.

Figure 13 depicts the property prices over time, providing a visual representation of the trend and fluctuations in property prices. The $x$-axis represents time, while the $y$-axis represents property prices. This plot enables us to observe any long-term trends, seasonality, or irregular patterns in the property price data.

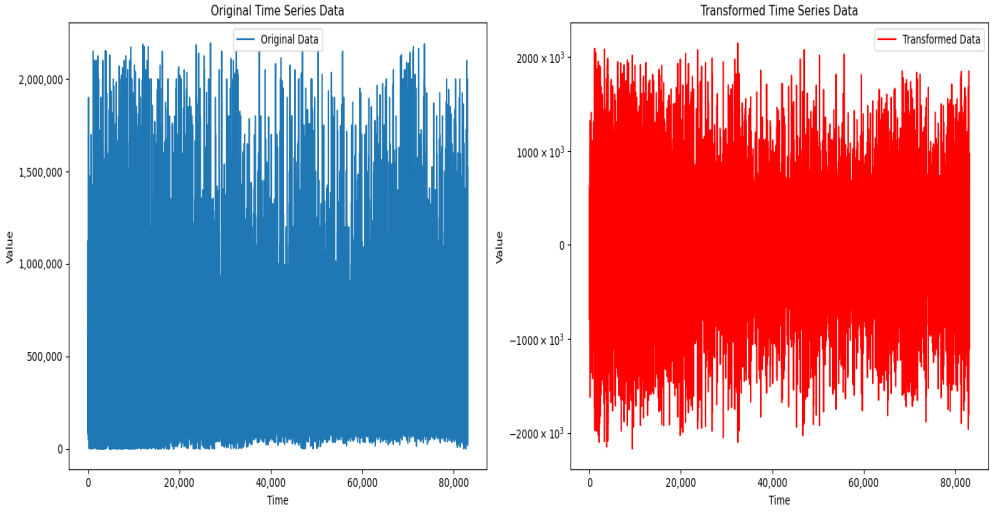



**Figure 13.** Trend and Fluctuations in Property Prices Over Time.

Figure 14 illustrates the frequency distribution of property prices, presenting the distribution of property prices across different value ranges. The $x$-axis displays the property price intervals, and the $y$-axis represents the frequency or count of properties falling within each price interval. This plot allows us to visualize the central tendency, dispersion, and skewness of the property price distribution.

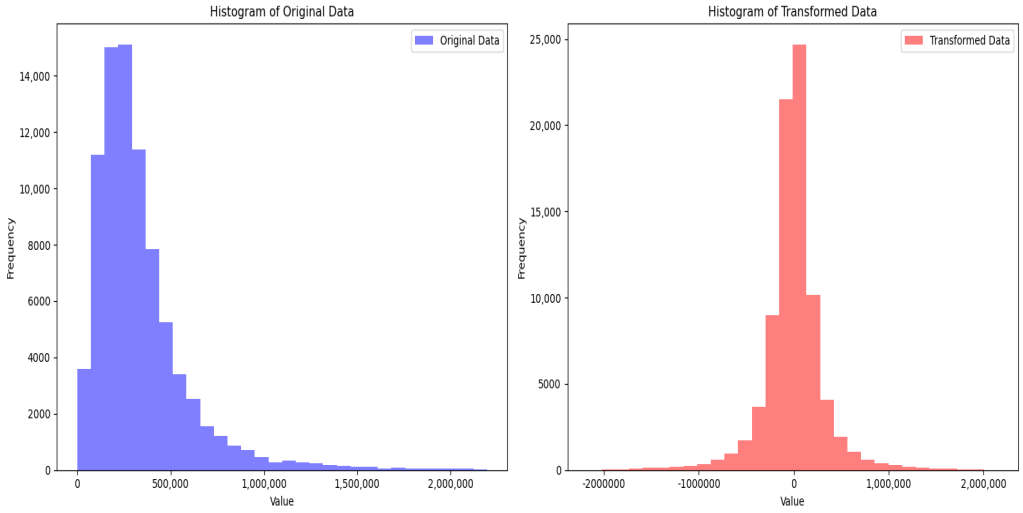



**Figure 14.** Frequency Distribution of Property Prices across Value Ranges.

Figure 15 provides a graphical summary of the distribution of property prices, highlighting key statistical measures such as the median, quartiles, and outliers. The box represents the interquartile range (IQR), with the median indicated by the line inside the box. The whiskers extend to the minimum and maximum non-outlier values, while any data points beyond the whiskers are considered outliers. This plot facilitates the identification of central tendency, spread, and variability in property prices.

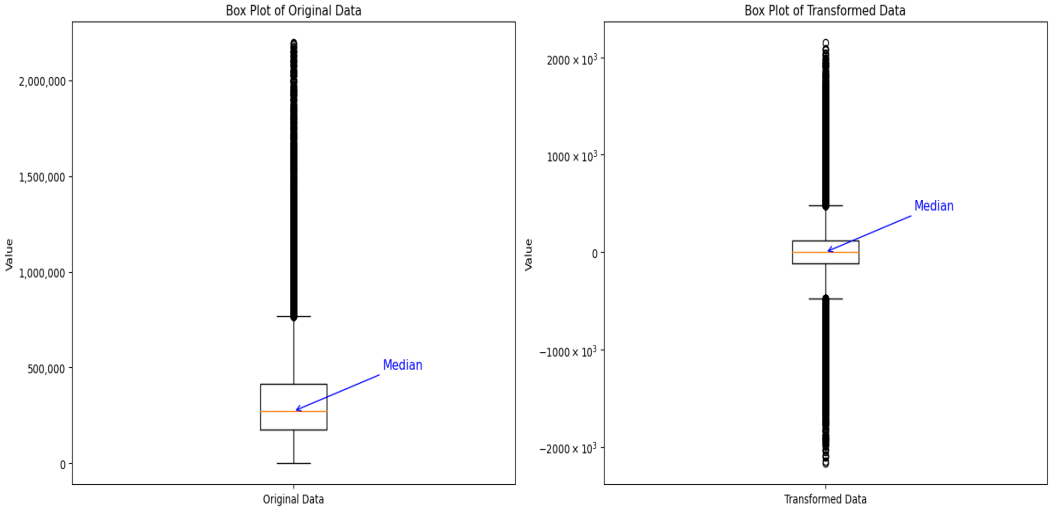



**Figure 15.** Summary of Property Price Distribution with Statistical Measures.

Figure 16 displays the correlation of property prices with lagged versions of itself, revealing any temporal dependencies or autocorrelation within the time series data. The $x$-axis represents the lag or time interval, while the $y$-axis indicates the autocorrelation coefficient, which measures the strength and direction of the relationship between property prices at different time lags. This plot assists in identifying any significant autocorrelation patterns, which can inform time series modeling and forecasting.

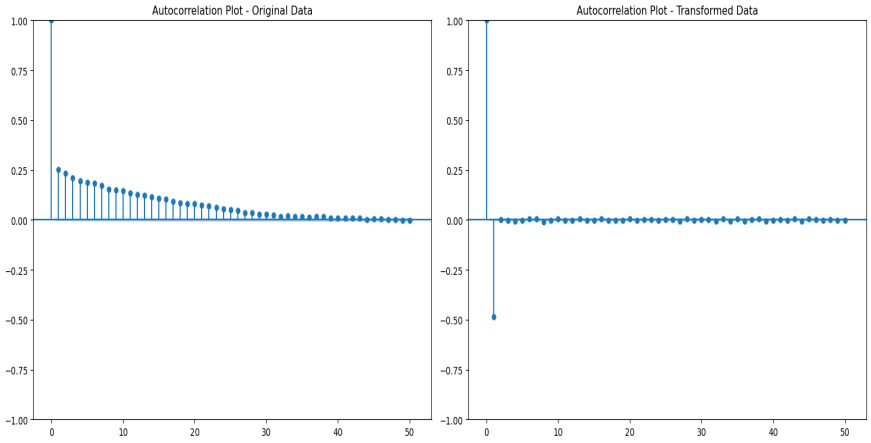



**Figure 16.** Autocorrelation of Property Prices across Time Lags.

The visual inspection of time series plots revealed potential trends and seasonality in the data, prompting further analysis. Subsequently, ADF tests were conducted to formally test for stationarity, with the null hypothesis being non-stationarity as shown in Table 5.

**Table 5.** Critical Values for Hypothesis Testing.

| Confidence Level | Critical Value |
|---|---|
| 1% | −3.430429535692449 |
| 5% | −2.8615751534462155 |
| 10% | −2.5667887109639147 |

The ADF test statistic of −40.69 indicates a significant deviation from the null hypothesis of non-stationarity, with a $p$-value of 0.0, suggesting strong evidence against the null hypothesis. Additionally, comparing the ADF statistic to critical values at the 1%, 5%, and 10% significance levels further confirms the rejection of the null hypothesis, indicating

that the time series data are stationary. This implies that the statistical properties of the data, including mean, variance, and autocorrelation structure, remain relatively constant over time.

The stationarity of the data is a crucial assumption for many time series models, including regression analysis. By establishing stationarity, we ensure the validity of regression analysis results, as violating stationarity assumptions can lead to biased estimates and unreliable forecasts. Therefore, the confirmation of stationarity through the ADF test instills confidence in the subsequent regression analysis conducted in this study.

#### 4.5.2. Multicollinearity Assessment

Multicollinearity among predictor variables in our regression models was assessed to ensure the stability and interpretability of regression coefficients. We employed variance inflation factor (VIF) analysis to quantify the degree of multicollinearity among predictor variables as shown in Table 6.

**Table 6.** VIF for Predictor Variables.

| Predictor Variable | VIF |
|---|---|
| Property Type | 1.141504 |
| Old/New | 1.061866 |
| Duration | 1.576009 |
| PAON | 1.461037 |
| SAON | 1.832881 |
| Street | 1.008614 |
| Locality | 1.059770 |
| Town/City | 1.220228 |
| District | 1.423253 |
| County | 1.246195 |
| PPD Category Type | 1.030142 |
| Record Status | 1.007994 |
| Intercept | 63.021976 |

The VIF values ranged from 1.0086 to 63.022, with most variables exhibiting low to moderate levels of multicollinearity. The high VIF value for the intercept variable indicates potential multicollinearity issues, which may warrant further investigation and remedial actions.

#### 4.5.3. Ridge and Lasso Coefficients

The examination of coefficients obtained from Ridge and Lasso regression models provided valuable insights into the relative importance and impact of predictor variables on property prices. The Ridge and Lasso techniques are regularization methods that introduce penalties on the regression coefficients, effectively shrinking or driving some coefficients to zero. This process can mitigate issues such as multicollinearity and overfitting, while also facilitating feature selection.

The Ridge coefficients, ranging from 126,286.12 to −27,175.84, revealed the magnitude and direction of the relationships between predictor variables and property prices. Positive coefficients indicated a direct relationship, where an increase in the predictor variable corresponded to an increase in property prices, while negative coefficients suggested an inverse relationship.

Similarly, the Lasso coefficients, ranging from 151,507.63 to −36,508.07, provided insights into the relative importance of each predictor variable. The Lasso method's ability to drive less important coefficients to zero effectively performed feature selection, highlighting the most relevant predictors for property price estimation.

By examining the Ridge and Lasso coefficients, stakeholders can gain a better understanding of the factors that significantly influence property prices. This knowledge can guide decision-making processes, such as feature selection for model development, identification of key drivers for property valuation, and targeted interventions or policies to address specific market dynamics as shown in Figures 17 and 18.

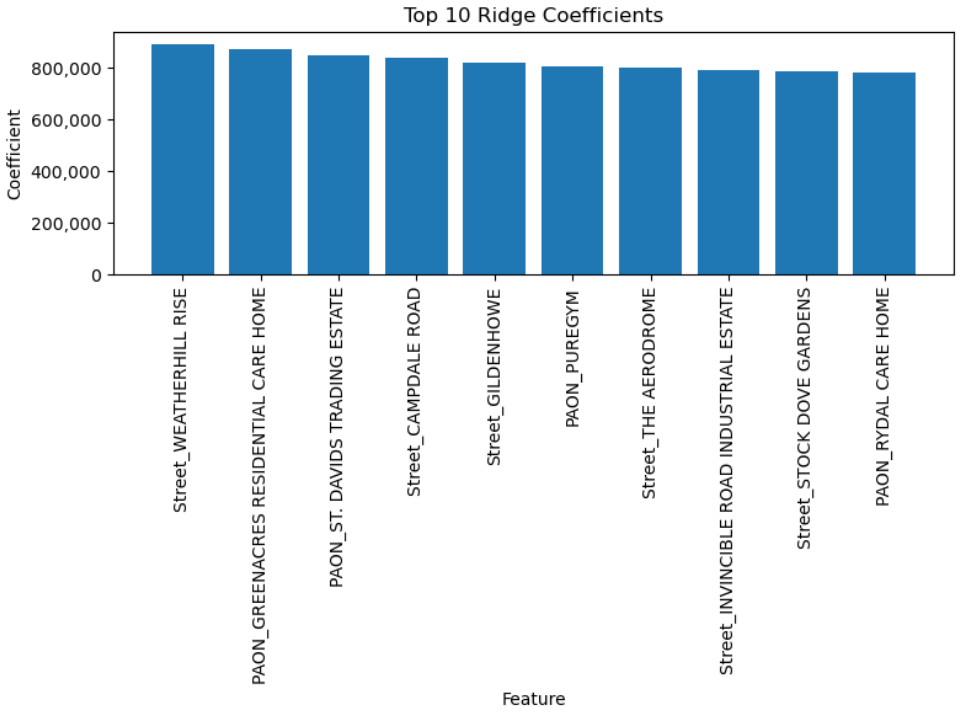



**Figure 17.** Top 10 Ridge Coefficients Graph.

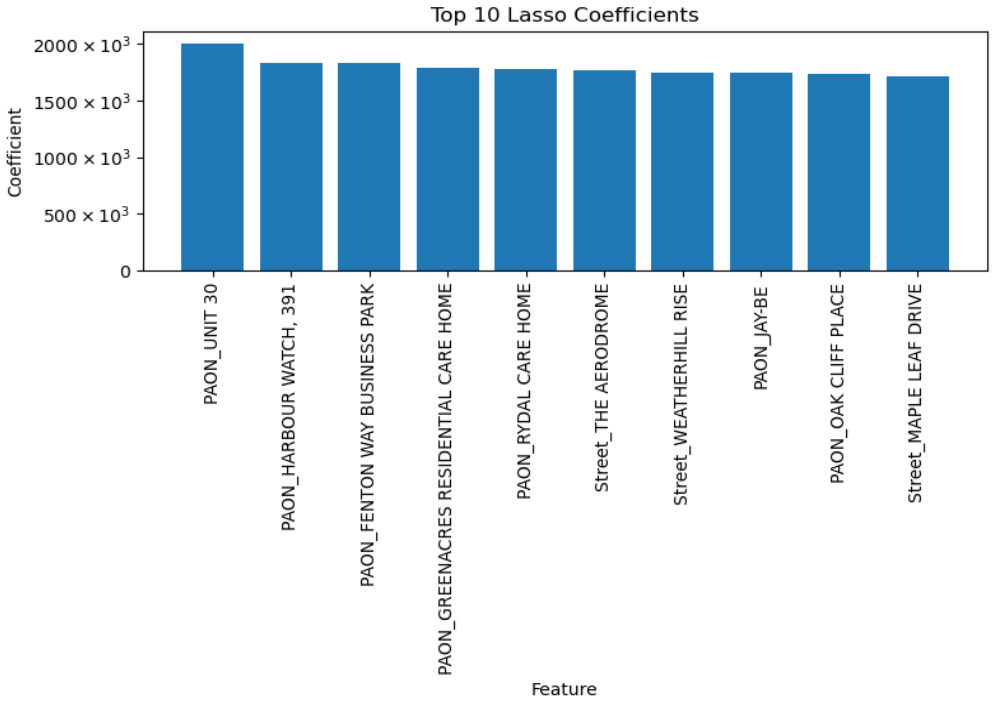



**Figure 18.** Top 10 Lasso Coefficients Graph.

However, it is important to note that the interpretation of these coefficients should be performed in conjunction with other diagnostic measures and domain knowledge. While Ridge and Lasso coefficients offer insights into variable importance, they may not fully capture complex interactions or nonlinear relationships present in the data. Additionally, the coefficients should be evaluated within the context of the specific modelling approach and the underlying assumptions made during the analysis.

Incorporating regularization techniques like Ridge and Lasso can enhance the robustness and interpretability of regression models, providing stakeholders with valuable tools for informed decision making and strategic planning within the real estate sector.

#### 4.5.4. OLS Regression

In regression analysis, the coefficients estimate the relationship between the predictor variables and the dependent variable (property prices). Each coefficient represents the change in the dependent variable associated with a one-unit change in the corresponding predictor variable, holding other variables constant as shown in Table 7.

**Table 7.** Regression Results for Predictor Variables.

| Predictor Variable | Coefficient Estimate | Standard Error | T-Value | *p*-Value |
|---|---|---|---|---|
| Property Type | −46.3 k | 592.646 | −78.117 | <0.001 |
| Old/New | 2.76 k | 2264.071 | 1.217 | 0.224 |
| Duration | −132.2 k | 2356.411 | −56.102 | <0.001 |
| PAON | 11.06 | 0.356 | 31.113 | <0.001 |
| SAON | 8.51 | 2.659 | 3.201 | <0.001 |

The regression analysis uncovered key factors influencing property prices. Notably, Property Type wielded significant influence, with certain types precipitating a considerable decrease in price (~GBP 46,300). Conversely, Old/New status, despite a positive coefficient (~GBP 2760), failed to exhibit a significant effect on prices. Duration emerged as a pivotal factor, showcasing a substantial negative impact (~GBP 132,200 per unit increase), suggesting that longer ownership durations correlate with lower prices. Moreover, PAON and SAON displayed notable positive effects on prices, with increases of approximately GBP 11.06 and GBP 8.51 per unit, respectively. These findings underscore the multifaceted nature of property price determination and highlight the significance of addressing diverse predictors in analyses.

### *4.6. Model Performance*

The performance of various regression models was evaluated using a comprehensive suite of metrics, including MSE, MAE, RMSE and $R^2$. The models were assessed both with and without regularization.

#### 4.6.1. Without Regularization

The performance of each model without regularization is shown in Tables 8 and 9. The models evaluated include Random Forest, XGBoost, LightGBM, CatBoost, Linear Regression, and a Hybrid Regression model.

**Table 8.** Regression model performance training set metrics without regularization.

| Model | MSE (Millions) | RMSE (Millions) | MAE (Thousands) | $R^2$ |
|---|---|---|---|---|
| Radom Forest | 3421 | 184,970 | 14,400 | 0.99 |
| XGBoost | 676,250 | 822,340 | 100,500 | 0.83 |
| LightGBM | 1,168,250 | 1,080,850 | 63,500 | 0.71 |
| CatBoost | 50,510 | 224,740 | 41,250 | 0.99 |
| Linear Regression | 218,270 | 467,190 | 105,820 | 0.94 |

| | | | | |
|---|---|---|---|---|
| Hybrid Regression | 356,730 | 597,270 | 61,180 | 0.91 |

**Table 9.** Regression model performance testing set metrics without regularization.

| Model | MSE (Millions) | RMSE (Millions) | MAE (Thousands) | $R^2$ |
|---|---|---|---|---|
| Radom Forest | 610.76 | 781.51 | 38.58 | 0.93 |
| XGBoost | 1909.24 | 1381.75 | 114.88 | 0.77 |
| LightGBM | 3888.00 | 1971.80 | 83.67 | 0.53 |
| CatBoost | 4220.00 | 2054.26 | 77.46 | 0.49 |
| Linear Regression | 251.48 | 501.48 | 108.84 | 0.97 |
| Hybrid Regression | 2748.89 | 1657.98 | 81.87 | 0.67 |

#### 4.6.2. With Regularization

The performance of each model with regularization is shown in Tables 10 and 11. The models evaluated include Ridge, Lasso, ElasticNet, Random Forest, XGBoost, LightGBM, CatBoost, Linear Regression, and a Hybrid Regression model.

**Table 10.** Regression model performance training set metrics with regularization.

| Model | MSE (Millions) | RMSE (Millions) | MAE (Thousands) | $R^2$ |
|---|---|---|---|---|
| Ridge | 2140 | 1.46 | 413 | 0.88 |
| Lasso | 2140 | 1.46 | 413 | 0.88 |
| ElasticNet | 2.18 | 1.48 | 404 | 0.88 |
| Radom Forest | 3640 | 1.91 | 233 | 0.86 |
| XGBoost | 4470 | 2.11 | 186 | 0.97 |
| LightGBM | 4410 | 2.10 | 200 | 0.97 |
| CatBoost | 4580 | 2.14 | 104 | 0.98 |
| Linear Regression | 2140 | 1.46 | 413 | 0.88 |
| Hybrid Regression | 4340 | 2.08 | 154 | 0.98 |

**Table 11.** Regression model performance testing set metrics with regularization.

| Model | MSE (Millions) | RMSE (Millions) | MAE (Thousands) | $R^2$ |
|---|---|---|---|---|
| Ridge | 1070 | 1.04 | 427 | 0.91 |
| Lasso | 1070 | 1.04 | 427 | 0.91 |
| ElasticNet | 1010 | 1.01 | 415 | 0.92 |
| Radom Forest | 2170 | 1.47 | 282 | 0.82 |
| XGBoost | 5820 | 2.41 | 279 | 0.52 |
| LightGBM | 1960 | 1.40 | 272 | 0.84 |
| CatBoost | 4680 | 2.16 | 251 | 0.61 |
| Linear Regression | 1070 | 1.04 | 427 | 0.91 |
| Hybrid Regression | 3860 | 1.96 | 252 | 0.68 |

### *4.7. Generalization*

In addition to evaluating the performance of predictive models on the testing set, we further validated the models on an independent dataset to assess their generalizability. The independent dataset was sourced separately from the Ames Housing dataset, available at https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data (accessed on 13 May 2024), ensuring that it represents a distinct set of observations from the original dataset used for training and testing. By validating the models on this independent dataset, we aimed to corroborate their performance and ensure their applicability beyond the specific data used in model training. The results of this validation

process are summarized in Table 12, providing insights into the models' robustness and generalizability.

The generalization performance of each model is shown in table below. The models evaluated include Ridge, Lasso, ElasticNet, Random Forest, XGBoost, LightGBM, CatBoost, Linear Regression, and a Hybrid Regression model.

**Table 12.** Regression model performance testing set metrics with generalization.

| Model | MSE (Millions) | RMSE (Millions) | MAE (Thousands) | $R^2$ |
|---|---|---|---|---|
| Ridge | 1755 | 1.325 | 970.6 | 0.653 |
| Lasso | 1755 | 1.325 | 970.0 | 0.653 |
| ElasticNet | 1685 | 1.298 | 964.8 | 0.667 |
| Radom Forest | 1913 | 1.383 | 1019.4 | 0.622 |
| XGBoost | 2068 | 1.438 | 1057.5 | 0.591 |
| LightGBM | 1926 | 1.388 | 1023.0 | 0.619 |
| CatBoost | 1814 | 1.347 | 977.8 | 0.641 |
| Linear Regression | 1755 | 1.325 | 970.0 | 0.653 |
| Hybrid Regression | 1800 | 1.341 | 1000 | 0.650 |

These combined analyses provide comprehensive insights into the factors influencing property prices and the effectiveness of regression models in predicting them accurately.

## 5. Discussion

In this study, we conducted a comprehensive analysis of property price dynamics, leveraging various analytical techniques and regression models. Our findings shed light on the multifaceted nature of the real estate market, highlighting key trends, patterns, and factors influencing property prices. This discussion section aims to delve deeper into the implications of our findings, address the limitations of our analysis, and suggest avenues for future research and practical applications.

### *5.1. Evaluation of ML Models for House Price Prediction*

This analysis examines the effectiveness of various regression models in predicting property prices. It highlights the factors influencing model performance and the trade-off between accuracy and interpretability.

Random Forest emerges as a strong contender as it achieved a high $R^2$ value of 0.99 on the training set, indicating excellent predictive power. However, a slight decrease on the testing set 0.93 suggests some overfitting. This aligns with previous research where Random Forest demonstrates strong performance but can be sensitive to overfitting [16,32].

XGBoost and LightGBM show promise but struggle with overfitting, while these models perform well on the training set, their performance significantly drops on the testing set [9,33]. This highlights their potential overfitting tendencies in real-world applications [35].

Linear Regression offers interpretability and consistency, while not the most ac-curate, Linear Regression exhibits consistent performance across training and testing sets, making it a reliable baseline model. Additionally, it provides easy interpretation of coefficients, which is crucial for stakeholders to understand the impact of different factors on property prices.

CatBoost shows potential, like Random Forest, CatBoost achieves a high training set R20.99 but experiences a substantial drop on the testing set. While its performance varies across studies, it warrants further exploration for its potential in house price estimation [12].

Regularization techniques can help such as Ridge, Lasso, and ElasticNet can improve generalization performance by reducing overfitting. However, they may lead to a slight decrease in training set accuracy (Tables 10 and 11).

High accuracy models Random Forest, XGBoost, LightGBM, CatBoost as these models achieve impressive accuracy but can be complex and less interpretable. Their "black box" nature makes it difficult to understand how they arrive at predictions.

On the other hand, Linear Regression is the simpler model and offers clear interpretations of coefficients, allowing stakeholders to understand which features most significantly impact property prices. However, its accuracy might be lower compared to more complex models.

The optimal model selection depends on the specific needs. If interpretability is paramount (e.g., real estate decision making), a model like Linear Regression might be preferred. However, if maximizing accuracy is the primary goal, Random Forest could be a good choice, with the caveat of potential overfitting.

In terms of interpretability, Linear Regression is typically the preferred choice due to its simplicity and ease of interpretation. It provides clear insights into how each feature affects the predicted house prices through its coefficients. Regularized Regression Models like Ridge and Lasso also offer some level of interpretability while addressing multicollinearity and overfitting issues. They penalize large coefficients, making the model more interpretable while still maintaining reasonable accuracy.

In terms of accuracy, Gradient Boosting Models like XGBoost, LightGBM, and CatBoost often offer high predictive accuracy. These models are adept at capturing complex patterns and interactions in the data, resulting in accurate predictions. However, they may lose some interpretability due to their ensemble nature and black-box modelling approach.

In terms of complexity, Random Forest and Gradient Boosting Models (XGBoost, LightGBM, CatBoost) tend to be more complex than Linear Regression and Regularized Regression Models. They involve multiple decision trees or boosting iterations, making them computationally more intensive and potentially harder to interpret. Hybrid Regression Models, which combine the strengths of different techniques, may offer a balance between complexity and accuracy. However, they may require more computational resources and expertise to implement effectively.

Ultimately, the choice depends on the specific requirements of the problem and the priorities of the stakeholders. If interpretability is crucial and the relationships between features and target variable are relatively simple, Linear Regression or Regularized Regression Models may be preferred. If maximizing accuracy is paramount and interpretability is less critical, Gradient Boosting Models could be the right choice. Hybrid Regression Models may offer a compromise between accuracy and interpretability but may also introduce additional complexity.

### *5.2. Understanding Property Price Dynamics*

The EDA conducted on the dataset has unveiled valuable insights into various aspects of property prices and their underlying dynamics.

Firstly, the distribution of property prices was examined through histograms and box plots, revealing a clustering of properties at the lower end of the price spectrum, with fewer properties available in higher price ranges. Additionally, the presence of high-value outliers indicates the existence of properties with significantly higher prices. A Jarque–Bera test further confirmed that the property price data are not normally distributed. To address potential skewness or kurtosis, a log transformation was applied to the data, providing a clearer understanding of the distributional characteristics.

Temporal trends in property prices were analyzed over a 25-year period, showcasing both long-term trends and shorter-term fluctuations. A consistent upward trajectory in mean property prices was observed from 1999 to approximately 2019, with a notable anomaly in 2019 marked by a sharp spike in prices followed by subsequent volatility. The

surge in 2019 may be attributed to factors such as robust economic growth or in-creased demand for housing, while the subsequent volatility could stem from market corrections or economic uncertainties.

Spatial analysis techniques, including heatmaps and choropleth maps, revealed significant regional variations in property prices across the UK. While England and Wales experienced declines in average house prices, Scotland and Northern Ireland witnessed growth, highlighting diverse market dynamics. Notably, London's housing market exhibited the lowest annual percentage change, contrasting with the Northwest of England, which saw the highest increase.

Correlation analysis explored relationships between various features and property prices, indicating subtle relationships between property prices and address-related variables. ANOVA and chi-square tests further assessed differences in property prices across different categories of categorical variables, revealing statistically significant variations.

Univariate analysis delved into individual features, such as price distribution and temporal trends, uncovering insights into the prevalence of lower-priced properties and distinct phases in property price trends over time.

Bivariate analysis explored relationships between pairs of variables, revealing positive correlations between property size and price, as well as variations in pricing trends across different locations and property types.

Multivariate analysis aimed to understand simultaneous interactions between multiple features and their combined impact on property prices. Feature importance analysis highlighted the district as the most influential feature, while interaction effects were explored through predictive modelling, demonstrating high predictive capability.

Overall, the comprehensive EDA provides stakeholders with valuable insights into the complexities of property price dynamics, enabling informed decision making and targeted interventions within the real estate sector.

### *5.3. Implications for Stakeholders*

Property price predictions play a crucial role in empowering stakeholders across the real estate spectrum, offering valuable insights for investors, Policy makers, real estate developers, and homeowners alike. By leveraging predictive modelling techniques, stakeholders can make informed decisions, mitigate risks, and capitalize on opportunities in the dynamic real estate market landscape.

For investors seeking to navigate the complexities of real estate investments, understanding temporal and spatial trends in property prices is paramount. Predictive modelling techniques, particularly Linear Regression and Regularized Regression Models (such as Ridge, Lasso, and ElasticNet), offer interpretability and consistency, enabling investors to identify potential profit avenues and mitigate risks associated with market fluctuations. These models provide insights into property price trends, allowing investors to optimize their investment strategies and maximize returns.

Policy makers, tasked with addressing housing affordability issues and stimulating economic growth, can benefit greatly from property price predictions. Linear Regression and Regularized Regression Models offer valuable tools for Policy makers to formulate targeted interventions. By analyzing predictive models, Policy makers can identify regions experiencing rapid property price appreciation and implement measures such as subsidies, tax incentives, or zoning regulations to promote affordable housing options and drive economic growth.

Real estate developers rely on predictive modelling techniques to assess market demand and identify areas with high potential for development. Models like Linear Regression and Regularized Regression Models provide developers with data-driven insights into property price predictions, enabling them to make informed decisions about where to invest in new projects, optimize pricing strategies, and allocate re-sources effectively to maximize returns on investment.

For homeowners, property price predictions offer valuable insights into the current and future value of their properties. Linear Regression and Regularized Regression Models empower homeowners to make informed decisions about selling, renovating, or refinancing their homes. By understanding predicted market trends, homeowners can identify opportunities to increase the value of their properties through strategic upgrades or renovations, ultimately enhancing their investment.

### *5.4. Enhancing External Validity and Generalizability*

To improve the external validity of our findings, future research endeavors could undertake comparative analyses across multiple real estate markets. By examining similarities and differences in property price dynamics, market drivers, and regulatory environments, researchers can identify common patterns and unique characteristics across diverse contexts. This comparative approach not only validates the robustness of our predictive models but also provides valuable insights into global trends and regional variations in real estate markets.

The stationarity analysis conducted through techniques such as visual inspection of time series plots, ADF tests and ACF plots played a crucial role in validating the assumptions underlying our regression models. By confirming the stationarity of the property price data, we ensured the reliability of our regression analysis results and instilled confidence in the subsequent forecasting and decision-making processes. However, it is essential to acknowledge that stationarity is a temporal concept, and market dynamics can evolve over time, potentially leading to non-stationarity in the future. Therefore, it is imperative for stakeholders to continuously monitor the stationarity of property price data and adapt their models accordingly. This could involve incorporating time-varying coefficients or adopting dynamic modelling techniques that can capture evolving market trends and non-stationarities.

Furthermore, the assessment of multicollinearity through VIF analysis provided valuable insights into the stability and interpretability of our regression coefficients. While most predictor variables exhibited low to moderate levels of multicollinearity, the high VIF value for the intercept variable warrants further investigation and potential remedial actions. Addressing multicollinearity is crucial for ensuring the robustness and reliability of regression models, as it can lead to inflated standard errors, unstable coefficient estimates, and reduced predictive power. Stakeholders should remain vigilant about potential multicollinearity issues and explore techniques such as variable selection, principal component analysis, or Ridge regression to mitigate its effects.

The research methodology, characterized by its structured approach and quantitative techniques, holds potential for transferability to other geographical regions. By documenting our methodology in detail and providing guidelines for its adaptation, we enable researchers in different contexts to leverage our framework for analyzing their respective real estate markets. This methodological transferability enhances the reproducibility of our findings and facilitates cross-market comparisons, thereby contributing to the advancement of real estate research on a global scale.

Acknowledging the sensitivity of our models to contextual factors is essential for assessing their external validity. While our predictive models demonstrate efficacy within the UK real estate market, it is imperative to evaluate their performance across various socio-economic contexts, regulatory frameworks, and cultural landscapes. Sensitivity analyses can elucidate the extent to which our models generalize to different settings, thereby informing stakeholders about the potential applicability and limitations of our research findings.

### *5.5. Challenges and External Factors*

While the UK has emerged from the COVID-19 pandemic in a relatively strong position, supported by significant fiscal stimulus, the subsequent Russo–Ukrainian crisis has introduced new challenges [43,44]. Global oil prices have increased by 11%, and wholesale

gas prices in the UK have risen by 40% since the invasion [45]. These events have likely impacted the real estate market, albeit to an extent that warrants further investigation. A comparative projection of the evolution of the UK real estate market with the European one in the post-pandemic period of 2020–2024 could provide valuable insights into the sector's resilience and potential trajectories [46].

The UK real estate market faces several key challenges that make accurate analysis and prediction difficult. Specifically, the market is characterized by high regional variations in prices and trends, influenced by diverse socioeconomic factors [47]. The cyclical nature of the market, subject to economic upswings and downturns, also contributes to its complexity. These aspects create uncertainties that can influence investor perceptions and the performance of various property sectors, necessitating robust predictive models to navigate the intricate landscape.

We acknowledge the importance of external factors such as changes in government policies, economic conditions, and geopolitical events, which could significantly influence property prices. Future research endeavors should consider incorporating these external factors into predictive models to enhance their accuracy and robustness. By accounting for a broader range of influences, researchers can better understand the complexities of the real estate market and improve the reliability of their predictions.

### *5.6. Limitations and Future Research Directions*

While our analysis encompasses a broad range of factors, there are several limitations that signal areas for future exploration. Primarily, our study focused predominantly on quantitative variables, overlooking crucial qualitative factors such as neighborhood amenities, housing preferences, and cultural influences. Integrating qualitative data in future investigations could significantly enhance the predictive accuracy and robustness of our models by capturing the nuanced dynamics of the market.

Additionally, our reliance on historical data presents a limitation, potentially obscuring emerging trends and market disruptions. To address this constraint, future research endeavors could delve into dynamic modelling techniques adept at capturing real-time market dynamics and forecasting future trends, thereby offsetting the static nature of historical data analysis.

Furthermore, our analysis primarily focused on the UK real estate market, potentially limiting the generalizability of our findings beyond this geographic domain. To mitigate this limitation, future studies could expand our analysis to encompass global real estate markets. By broadening the scope to include diverse geographic regions, researchers can facilitate cross-country comparisons and glean deeper insights into the phenomena of market convergence and divergence.

Additionally, a thorough discussion of potential biases in the models or data, along with an analysis of how the models perform across different demographics and regions, would enhance the depth and rigor of our paper, providing a more comprehensive understanding of real estate market dynamics.

Acknowledging the limitations associated with assuming linear relationships between variables and property prices is crucial, particularly in the context of Linear Regression models. While Linear Regression provides a straightforward framework for analyzing relationships between variables, it may overlook complex nonlinear dynamics that could influence property price predictions.

Nonlinear relationships, such as exponential growth or diminishing returns, may exist between certain predictor variables and property prices, which cannot be adequately captured by linear models. To address this limitation, future research endeavors could explore more sophisticated modelling techniques capable of capturing nonlinear relationships, such as polynomial regression, spline regression, or machine learning algorithms like Random Forests or Gradient Boosting machines.

These models offer greater flexibility in modelling complex interactions and nonlinear patterns in the data, potentially improving the accuracy and robustness of property

price predictions. Additionally, sensitivity analyses could be conducted to assess the impact of nonlinear dynamics on model performance and compare the predictive capabilities of linear and nonlinear models. By considering both linear and nonlinear modelling approaches, researchers can gain a more comprehensive understanding of the underlying dynamics driving property prices and make more informed predictions.

To address the limitation of external validity and foster cross-market generalizability, future research initiatives could adopt a multidisciplinary approach. By integrating insights from economics, sociology, urban studies, and data science, researchers can develop comprehensive frameworks for analyzing real estate markets worldwide. Moreover, collaborative efforts involving international partnerships and data-sharing agreements can facilitate access to diverse datasets, enabling researchers to conduct cross-national studies and validate predictive models across multiple regions.

### *5.7. Practical Applications and Recommendations*

Based on our comprehensive analysis, we propose several actionable recommendations for stakeholders within the real estate sector. Firstly, stakeholders should embrace a holistic approach to property valuation, integrating both quantitative metrics and qualitative insights into their decision-making frameworks. This multifaceted approach ensures a more accurate assessment of property worth and market positioning.

Secondly, Policy makers ought to prioritize initiatives that promote housing affordability, facilitate sustainable urban development, and address disparities in housing accessibility. By implementing targeted policies and incentives, governments can foster inclusive housing markets and support equitable access to housing opportunities for all segments of society.

Thirdly, real estate developers can harness the power of predictive analytics and machine learning algorithms to optimize pricing strategies, enhance market competitiveness, and mitigate investment risks. By leveraging advanced technologies, developers can gain valuable insights into market trends, buyer preferences, and future demand dynamics, enabling them to make informed decisions and drive profitability.

Lastly, homeowners should remain vigilant of evolving market trends, seek professional valuation services regularly, and formulate long-term investment strategies aligned with their financial objectives. By staying informed and proactive, homeowners can maximize the value of their properties and navigate changing market conditions effectively.

By implementing these recommendations, stakeholders can adapt to dynamic market environments, capitalize on emerging opportunities and contribute to the sustainable growth and development of the real estate sector.

## 6. Conclusions

In conclusion, this study has provided valuable insights into the complex dynamics of property prices and the efficacy of predictive modelling in understanding and forecasting these trends within the UK real estate market. Through a comprehensive EDA and a review of the existing literature, we have uncovered temporal, spatial, and multivariate patterns that significantly influence property prices.

Our findings carry significant implications for various stakeholders across the real estate spectrum. Investors can leverage our insights to make informed investment decisions and diversify their portfolios effectively. Policy makers can utilize the findings to devise targeted interventions aimed at addressing housing affordability, promoting sustainable urban development, and reducing regional disparities. Real estate developers stand to benefit from predictive modelling techniques to identify market opportunities, optimize pricing strategies, and mitigate investment risks. Additionally, homeowners can gain valuable insights into property valuation trends and make in-formed decisions regarding property investments and renovations.

Despite the progress made in understanding property price dynamics, there re-main several avenues for future research and improvement. Firstly, the integration of

qualitative factors alongside quantitative variables in predictive modelling could enhance predictive accuracy and robustness. Secondly, incorporating dynamic modelling techniques capable of capturing real-time market dynamics and forecasting future trends could provide more accurate predictions in rapidly changing real estate markets. Moreover, expanding the scope of analysis beyond the UK real estate market to encompass global markets would facilitate cross-country comparisons and yield deeper insights into market convergence and divergence.

This study underscores the importance of adopting a multidimensional approach to property valuation, incorporating both quantitative metrics and qualitative insights. By embracing advanced analytics and machine learning techniques, stakeholders can adapt to dynamic market environments, capitalize on emerging opportunities, and contribute to the sustainable growth and development of the real estate sector.

This study contributes to the growing body of knowledge surrounding property price dynamics and predictive modelling, providing stakeholders with actionable insights to make informed decisions and drive positive outcomes within the real estate sector.

**Author Contributions:** K.V.M. and R.H. conceptualized and designed the overall research framework. S.M. developed the experimental approach and contributed essential materials. K.V.M. created and implemented the computational tools. K.V.M., R.H. and S.M. verified the accuracy and reliability of the results. K.V.M. conducted statistical analyses. S.M. performed the experiments and gathered data. K.V.M. organized and managed the research data. S.M. wrote the initial draft of the manuscript. K.V.M. and R.H. critically revised and improved the manuscript. S.M. prepared figures and visual representations. R.H. oversaw and supervised the research process. K.V.M. coordinated project logistics. R.H. secured the financial support for the research. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not Applicable.

**Informed Consent Statement:** Not Applicable.

**Data Availability Statement:** The data presented in this study are openly available at: https://www.gov.uk/government/statistical-data-sets/price-paid-data-downloads (accessed on 14 April 2024).

**Acknowledgments:** The authors would like to acknowledge the use of ChatGPT 24 May 2023 version (OpenAI, San Francisco, CA, USA), specifically to assist in some content for improved clarity and effectiveness.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Brown, P.M. The changing United Kingdom residential real estate market. *Land Dev. Stud.* **1990**, *7*, 119–133. https://doi.org/10.1080/02640829008724007.
2. Lădaru, G.R.; Gombos, C.C.; Spiridon, C.; Troaca, V.A. Analysis of real estate market in United Kingdom. In Proceedings of the International Conference on Business Excellence, Bucharest, Romania, 24–26 March 2022; Volume 16, pp. 336–345. https://doi.org/10.2478/picbe-2022-0033.
3. Frodsham, M. The continuing uncertainty of property investment markets. *J. Prop. Invest. Finance* **2023**, *41*, 460–467. https://doi.org/10.1108/JPIF-01-2023-0002.
4. United Kingdom: November 2020 Price Paid Data Home Page. Available online: https://www.gov.uk/government/statistical-data-sets/price-paid-data-downloads (accessed on 2 April 2024).
5. Alzain, E.; Alshebami, A.S.; Aldhyani, T.H.H.; Alsubari, S.N. Application of Artificial Intelligence for Predicting Real Estate Prices: The Case of Saudi Arabia. *Electronics* **2022**, *11*, 3448. https://doi.org/10.3390/electronics11213448.
6. Adnan Diwan, S. Proposed study on evaluating and forecasting the resident property value based on specific determinants by case base reasoning and artificial neural network approach. *Indones. J. Electr. Eng. Comput. Sci.* **2020**, *17*, 1467. https://doi.org/10.11591/ijeecs.v17.i3.pp1467-1473.
7. Saiful, A. Prediksi Harga Rumah Menggunakan Web Scrapping dan Machine Learning Dengan Algoritma Linear Regression. *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)* **2021**, *8*, 41–50. https://doi.org/10.35957/jatisi.v8i1.701.

8. chougale, J.; Shinde, A.; Deshmukh, N.; Sawant, D.; Latke, V. House Price Prediction using Machine learning and Image Processing. *J. Univ. Shanghai Sci. Technol.* **2021**, *23*, 961–965. https://doi.org/10.51201/JUSST/21/05280.
9. Alfaro-Navarro, J.; Cano, E.L.; Alfaro-Cortés, E.; García, N.; Gámez, M.; Larraz, B. A Fully Automated Adjustment of Ensemble Methods in Machine Learning for Modeling Complex Real Estate Systems. *Complexity* **2020**, *2020*, 1–12. https://doi.org/10.1155/2020/5287263.
10. Weng, W. Research on the House Price Forecast Based on machine learning algorithm. *BCP Bus. Manag.* **2022**, *32*, 134–147. https://doi.org/10.54691/bcpbm.v32i.2881.
11. Lee, C.; Park, K.K. Representing Uncertainty in Property Valuation Through a Bayesian Deep Learning Approach. *Real Estate Manag. Valuat.* **2020**, *28*, 15–23. https://doi.org/10.1515/remav-2020-0028.
12. Saraswat, O.; Arunachalam, N. House Price Prediction Based on Machine Learning: A Case of King County. In Proceedings of the 2022 7th International Conference on Financial Innovation and Economic Development (ICFIED 2022), Zhuhai, China, 14–16 January 2022; Volume 757, pp. 1071–1087. https://doi.org/10.2991/aebmr.k.220307.253.
13. Liu, Z. Real Estate Price Prediction based on Supervised Machine Learning Scenarios. *Highlights Sci. Eng. Technol.* **2023**, *39*, 731–737. https://doi.org/10.54097/hset.v39i.6637.
14. Optimizing Ensemble Weights for Machine Learning Models: A Case Study for Housing Price Prediction. Available online: https://lib.dr.iastate.edu/imse_conf/185 (accessed on 14 April 2024).
15. Khrais, L.T.; Shidwan, O.S. The Role of Neural Network for Estimating Real Estate Prices Value in Post COVID-19: A Case of the Middle East Market. *Int. J. Electr. Comput. Eng. (IJECE)* **2023**, *13*, 4516. https://doi.org/10.11591/ijece.v13i4.pp4516-4525.
16. Sweta, R.K.; Swati, B.; Hasan, P. Predicting House Price with Deep Learning: A Comparative Study of Machine Learning Models. *Int. J. Multidiscip. Res.* **2023**, *5*, 1-10. https://doi.org/10.36948/ijfmr.2023.v05i02.1849.
17. Yücebaş, S.; Doğan, M.; Genç, L. A C4.5—Cart Decision Tree Model for Real Estate Price Prediction and the Analysis of The Underlying Features. *Konya J. Eng. Sci.* **2022**, *10*, 147–161. https://doi.org/10.36306/Konjes.1013833.
18. Gong, Y.; Liu, G.; Xue, Y.; Li, R.; Meng, L. A survey on dataset quality in machine learning. *Inf. Softw. Technol.* **2023**, *162*, 107268. https://doi.org/10.1016/j.infsof.2023.107268.
19. Liu, G. Research on Prediction and Analysis of Real Estate Market Based on the Multiple Linear Regression Model. *Sci. Program.* **2022**, *2022*, 1–8. https://doi.org/10.1155/2022/5750354.
20. Manikandan, S. Measures of central tendency: Median and mode. *J. Pharmacol. Pharmacother.* **2011**, *2*, 214–215. https://doi.org/10.4103/0976-500X.83300.
21. Studies from China Medical University Have Provided New Information about Health and Medicine (The ability of different imputation methods for missing values in mental measurement questionnaires). *Ment. Health Wkly. Dig.* **2020**, 789.
22. Lee, J.Y.; Styczynski, M.P. NS-kNN: a modified k-nearest neighbors approach for imputing metabolomics data. *Metabolomics* **2018**, *14*, 153. https://doi.org/10.1007/s11306-018-1451-8.
23. Kwak, S.K.; Kim, J.H. Statistical data preparation: management of missing values and outliers. *Korean J. Anesthesiol.* **2017**, *70*, 407–411. https://doi.org/10.4097/kjae.2017.70.4.407.
24. Oh, S. Feature Interaction in Terms of Prediction Performance. *Appl. Sci.* **2019**, *9*, 5191. https://doi.org/10.3390/app9235191.
25. Petropoulos, F.; Apiletti, D.; Assimakopoulos, V.; Babai, M.Z.; Barrow, D.K.; Ben Taieb, S.; Bergmeir, C.; Bessa, R.J.; Bijak, J.; Boylan, J.E.; et al. Forecasting: theory and practice. *Int. J. Forecast.* **2022**, *38*, 705–871. https://doi.org/10.1016/j.ijforecast.2021.11.001.
26. Visser, P.; VAN Dam, F.; Hooimeijer, P. Residential environment and spatial variation in house prices in the netherlands. *Tijdschr. voor Econ. en Soc. Geogr.* **2008**, *99*, 348–360. https://doi.org/10.1111/j.1467-9663.2008.00472.x.
27. Jolliffe, I.T.; Cadima, J. Principal component analysis: a review and recent developments. *Phil. Trans. R. Soc. A* **2016**, *374*, 20–20150202. https://doi.org/10.1098/rsta.2015.0202.
28. de Amorim, L.B.V.; Cavalcanti, G.D.C.; Cruz, R.M.O. The choice of scaling technique matters for classification performance. *Appl. Soft Comput.* **2023**, *133*, 109924. https://doi.org/10.1016/j.asoc.2022.109924.
29. Zhu, W.; Qiu, R.; Fu, Y. Comparative Study on the Performance of Categorical Variable Encoders in Classification and Regression Tasks. *arXiv* **2024**, arXiv:2401.09682. https://doi.org/10.48550/arxiv.2401.09682.
30. Ober, S.; Rasmussen, C. Benchmarking the Neural Linear Model for Regression. *arXiv* **2019**, arXiv:1912.08416. https://doi.org/10.48550/arxiv.1912.08416.
31. Wu, H.; Wang, C. A new machine learning approach to house price estimation. *New Trends Math. Sci.* **2018**, *4*, 165–171. https://doi.org/10.20852/ntmsci.2018.327.
32. Cajias, M. Can a machine understand real estate pricing?—Evaluating machine learning approaches with big data. *IDEAS Work. Pap. Ser. RePEc* **2019**. https://doi.org/10.15396/eres2019_232
33. Tekin, M.; Sari, I.U. Real Estate Market Price Prediction Model of Istanbul. *Real Estate Manag. Valuat.* **2022**, *30*, 1–16. https://doi.org/10.2478/remav-2022-0025.
34. Chhiller, M.; Shivam, S.; Kumar, R. Real Estate Price Prediction Using Machine Learning. *Int. J. Res. Appl. Sci. Eng. Technol.* **2023**, *7*, 6431–6435. https://doi.org/10.22214/ijraset.2023.53166.
35. Lee, J.; Kim, H.; Shim, G. Comparison of Real Estate Price Prediction Based on LSTM and LGBM. *Highlights Sci. Eng. Technol.* **2022**, *10*, 274–283. https://doi.org/10.54097/hset.v49i.8521.

36. Singh, V.; Pencina, M.; Einstein, A.J.; Liang, J.X.; Berman, D.S.; Slomka, P. Impact of train/test sample regimen on performance estimate stability of machine learning in cardiovascular imaging. *Sci. Rep.* **2021**, *11*, 14490. https://doi.org/10.1038/s41598-021-93651-5.
37. White, J.; Power, S.D. k-Fold Cross-Validation Can Significantly Over-Estimate True Classification Accuracy in Common EEG-Based Passive BCI Experimental Designs: An Empirical Investigation. *Sensors* **2023**, *23*, 6077. https://doi.org/10.3390/s23136077.
38. Malakouti, S.M.; Menhaj, M.B.; Suratgar, A.A. The Usage of 10-Fold Cross-Validation and Grid Search to Enhance ML Methods Performance in Solar Farm Power Generation Prediction. *Clean. Eng. Technol.* **2023**, *15*, 100664. https://doi.org/10.1016/j.clet.2023.100664.
39. Jadon, A.; Patil, A.; Jadon, S. A Comprehensive Survey of Regression Based Loss Functions for Time Series Forecasting. *arXiv* **2022**, arXiv:2211.02989. https://doi.org/10.48550/arxiv.2211.02989.
40. Kim, T.K. Understanding one-way ANOVA using conceptual figures. *Korean J. Anesthesiol.* **2017**, *70*, 22–26. https://doi.org/10.4097/kjae.2017.70.1.22.
41. Demir-Kavuk, O.; Kamada, M.; Akutsu, T.; Knapp, E.W. Prediction using step-wise L1, L2 regularization and feature selection for small data sets with large number of features. *BMC Bioinform.* **2011**, *12*, 412. https://doi.org/10.1186/1471-2105-12-412.
42. Gupta, R.K.; Agarwalla, R.; Naik, B.H.; Evuri, J.R.; Thapa, A.; Singh, T.D. Prediction of research trends using LDA based topic modeling. *Glob. Transit. Proc.* **2022**, *3*, 298–304. https://doi.org/10.1016/j.gltp.2022.03.015.
43. Kumar, R.; Mishra, R.S. COVID-19 Global Pandemic: Impact on Management of Supply Chain. *Int. J. Emerg. Technol. Adv. Eng.* **2020**, *10*, 132–139.https://doi.org/10.46338/IJETAE0416.
44. Nicola, M.; Alsafi, Z.; Sohrabi, C.; Kerwan, A.; Al-Jabir, A.; Iosifidis, C.; Agha, M.; Agha, R. The socio-economic implications of the coronavirus pandemic (COVID-19): A review. *Int. J. Surg.* **2020**, *78*, 185–193. https://doi.org/10.1016/j.ijsu.2020.04.018.
45. Fang, Y.; Shao, Z. The Russia-Ukraine conflict and volatility risk of commodity markets. *Finance Res. Lett.* **2022**, *50*, 103264. https://doi.org/10.1016/j.frl.2022.103264.
46. Hilber, C.A.L.; Vermeulen, W. The Impact of Supply Constraints on House Prices in England. *Econ. J.* **2016**, *126*, 358–405. https://doi.org/10.1111/ecoj.12213.
47. Bracke, P. House Prices and Rents: Microevidence from a Matched Data Set in Central London. *Real Estate Econ.* **2015**, *43*, 403–431. https://doi.org/10.1111/1540-6229.12062.