SCHOOL OF COMPUTING
Faculty of Engineering

Project Proposal Form MCST1043
Sem: 2 Session: 2024/25

SECTION A: Project Information.

Program Name: **Masters of Science (Data Science)**

Subject Name: **Project 1 (MCST1043)**

Student Name: Yang Mu

Metric Number: MCS241045

Student Email & Phone: yangmu@graduate.utm.my number:1162302346

Project Title: Big data driven: Forecast of global real estate market ups and downs in some regions

Supervisor 1:

Supervisor 2 / Industry Advisor(if any):

SECTION B: Project Proposal

Introduction:

The global real estate market is a complex system that is affected by many factors, including economic growth, interest rate changes, demographic changes, and government policies. (Mariia, 2021) In recent years, with the rapid development of big data and artificial intelligence technologies, it has become possible to use big data analysis technology to predict the trend of the global real estate market, and it has gradually become a research hotspot. However, most existing studies focus on forecasts for a single region, making it difficult to gain a comprehensive view of the overall trends and influencing factors of the global real estate market.

Problem Background:

Global real estate market data is rich, but there are data source dispersion, rigid

conditions inconsistent problems such as uneven, analysis Angle, lead to predict and analyze the global real estate market faces many challenges. For instance, there is a lack of a unified global real estate market forecasting and analysis method. (Lee Pei Wen, 2017) For example, the lack of a unified standard for describing and managing property data leads to difficulties in integrating and analyzing data from different

sources. In addition, most of the existing prediction models based on a single region or a single index, is difficult to accurately reflect the overall trend of the global real Estate market. (Ruoying Tan & Tze-Haw Chan 2021)

Problem Statement:

How to build a unified framework for the global real estate market based on big data?. The framework needs to formulate unified standards for describing and managing global real estate market data, and develop effective forecasting models on this basis, so as to improve the accuracy of forecasting changes in the global real estate market.

Aim of the Project:

The project aims to use advanced big data analytics techniques to exploit local areas and build global real estate market forecasting model data. The aim is to build global real estate market forecasting model data, enhance the accuracy and reliability of the prediction results, serve as the theoretical support for this project, and conduct in-depth research on the significance of its experimental results. (Petros et al , 2024).

Objectives of the Project:

Data Collection and Organization: To use advanced big data analytics techniques to exploit local areas and build global real estate market forecasting model data. To support for this project, and conduct in-depth research on the significance of its experimental results.

Data Framework Construction: To formulate a unified and standardized data framework and describe and manage the collected data. This framework ensures the usability and analyzability of data, while providing a solid foundation for its analysis and model establishment (Petros et al, 2024).

Data framework construction: To build a unified data framework to describe and manage collected housing data, ensuring its comparability and analysis, while providing a standardization basis for data integration and analysis (Petros et al, 2024).

Forecasting model development: To develop effective forecasting models, such as time sequence models and machine learning models, to predict the top, downs and trend of housing market in some areas, and to evaluate and verify the model, analyzing the accuracy and truth of the model.

Analysis and visualization of results: To visualize forecast results, analyze and interpret forecast results, and to put forward recommendations, such as forecasting and risk analysis of future real estate market trends, as well as investment strategies.

Scopes of the Project:

Data source: Mainly uses public global real estate transaction data, such as data from government agencies, real estate websites, research institutions, etc.

Prediction indicators: Mainly focus on the rise and fall trend of housing prices, and may include other indicators, such as transaction volume, rental level, etc.

Prediction method: Mainly uses time series analysis, machine learning models and other methods for prediction.

Result analysis: Compare and analyze the prediction results with the actual situation, explain the reasons behind the prediction results, and put forward some suggestions.

Expected Contribution of the Project:

Promote standardization of global real estate market data management: By establishing a unified metadata framework, this project will provide standardized specifications for global real estate market data management, promote the integration analysis of data from different sources, improve data quality and credibility, and provide the basis for more accurate market forecasting. Improve the accuracy and reliability of global real estate market forecasting: This project will develop a forecasting model based on big data analytics and combine multiple data sources and analysis methods to improve the accuracy and reliability of forecasting, provide investors and decision makers with more effective reference basis to help them better understand market trends and potential risks. (Ruoying Tan & Tze-Haw Chan, 2021)

Provide new perspectives and tools for business analysis: The research results of this project can provide new perspectives tools for business analysis. For example,

enterprises can use the metadata framework and prediction model built by this project to analyze real estate market trends, formulate more effective investment strategies, optimize business processes, and improve profitability. Urban planning and development: It can provide data support for urban planning and development, help urban planners better understand real estate market trends, and formulate more reasonable urban planning and development strategies. (Bo et al, 2022)

Project Requirements:

Software: Python, Pandas

Hardware: Computer, network, data

Technology/Technique/ Methodology/Algorithm: Data collection and preprocessing

- Collect and organize metadata from COVID-19 scientific datasets.
- Construct a metadata framework for COVID-19 scientific datasets,encompassing external features, content features, and sharing features.
- Utilize Protégé software to build an ontology for scientific datasets, defining core concepts and attribute relationships.
- Store the constructed knowledge graph using the Neo4j graph database, enabling query retrieval and reasoning for entities and their relationships.
- Data analysis
- Model Evaluation

Type of Project (Focusing on Data Science):

[ ] Data Preparation and Modeling

[√] Data Analysis and Visualization

[√] Business Intelligence and Analytics

[ ] Machine Learning and Prediction

[ ] Data Science Application in Business Domain

Status of Project:

[√] New

[ ] Continued

If continued, what is the previous title?

SECTION C: Declaration

I declare that this project is proposed by:

[√] Myself

[ ] Supervisor/Industry Advisor ( )

Student Name: Yang Mu

Yang Mu — **Signature**

07/04/2025 — **Date**

SECTION D: Supervisor Acknowledgement

The Supervisor(s) shall complete this section.

I/We agree to become the supervisor(s) for this student under aforesaid proposed title.

Name of Supervisor 1:

	Signature	**Date**
Name of Supervisor 2 (if any):		
	Signature	**Date**

SECTION E: Evaluation Panel Approval

The Evaluator(s) shall complete this section.

Result:

[ ] FULL APPROVAL
[ ] CONDITIONAL APPROVAL (Minor)
[ ] CONDITIONAL APPROVAL (Major)*
[ ] FAIL*

***** Student has to submit new proposal form considering the evaluators' comments.

Comments:

Name of Evaluator 1: .....

Signature **Date**

Name of Evaluator 2: .....

Signature **Date**

Chapter 1 Introduction

Abstract

In the current situation where everything is running at a fast pace, the real estate market is a complex system that is highly affected and closely related to people. Not only under the regulation of the economy, but also including the influence of policies, geographical factors, and even wars and natural disasters. Despite the progress made in big data and artificial intelligence, the existing research mainly focuses on local predictions and lacks an overall framework for global market analysis. And mostly analyze the impact of a single condition on the real estate market. This study aims to address this gap by constructing a unified data framework through the utilization of big data analysis and developing a predictive model for predicting regional real estate market trends. This project emphasizes the application and prediction of business analysis, the preprocessing and integration of diverse data, and the application and comparison of models. The expected contributions include improving the accuracy of predictions, standardizing data management practices, and providing actionable insights for investors, policymakers, and urban planners.

1.1 Introduction

The real estate market plays a very important role in economic stability and the development of people's livelihood. However, its volatility has brought significant challenges to almost all people, so accurate prediction tools are extremely necessary. However, the current research is still fragmented, with limited integration of cross-regional data and standardized methods under a single condition. However, with the development of society and technology, the progress of big data is also advancing rapidly. Therefore, accurately predicting the global real estate market is not out of reach. This study proposes a new method for unifying global real estate data management and comparing the optimal prediction models, with the aim of providing comprehensive insights into market trends and informing strategic decisions.

1.2 Background of the Problem

The global real estate market has generated a large number of datasets, but their different conditions, scattered sources and inconsistent standards have hindered effective analysis. Existing models often rely on data from a single region or isolated indices, and are unable to capture interrelated global trends. For instance, China's building height limit regulations almost affect different Southeast Asian countries, and the high housing management fees in the United States do not exist in many countries either. Furthermore, the traditional

single-condition prediction method is also difficult to accurately deal with and adapt to the real estate market that is dynamically influenced by multiple conditions. These restrictions emphasize the need for standardized data infrastructure and advanced analytical models to address the complexity of global market forecasting.

1.3 Statement of the Problem

The core challenge lies in the lack of a unified framework to integrate and analyze heterogeneous global real estate data. The current papers and studies have inconsistent conditions and data, regional biases and insufficient prediction accuracy. This study aims to answer: How to utilize big data technology to construct a standardized data framework, compare and obtain reliable prediction models, and predict fluctuations in the global real estate market.

1.4 Research Questions

How can the scattered and inconsistent global real estate data be standardized and integrated into a unified framework?

How to determine representative and effective market data for research under the vast amount of global data?

Which prediction models (such as Random Forest, Regression

Is machine learning the most effective for predicting the trends of the real estate market in multiple regions?

How to unify the influencing factors and make comparisons among the data?

How can the accuracy and reliability of these models be verified based on real market data?

1.5 Objectives of the Research

Data collection and standardization: Aggregate and preprocess global real estate data (such as prices and transaction volumes) to ensure consistency and reliability.

Framework development: Design a standardized metadata framework to support the integration and analysis of diverse and multi-regional data.

Model construction: Verify and compare the prediction model using Random Forest, Regression and machine learning algorithms.

Result interpretation: Visualize predictions, analyze model performance, and provide actionable suggestions for stakeholders.

1.6 Scope of the Study

Data source: Public datasets from government agencies, real estate platforms and research institutions.

Forecast indicators: Focus on the trend of housing prices, supplemented by transaction volume and rental levels.

Methods: the Random Forest, Regression, and machine learning techniques (such as LSTM, HistGradientBoostingClassifier).

Geographical focus: Select regions with market dynamics to test the universality of the model.

1.7 Significance of the study

The contributions of this research to the academic and industrial fields are as follows:

Promote data standardization: Establish a unified framework to enhance data comparability and quality.

Enhance predictive capabilities: Provide scalable models for accurate global market forecasting and reduce investment risks.

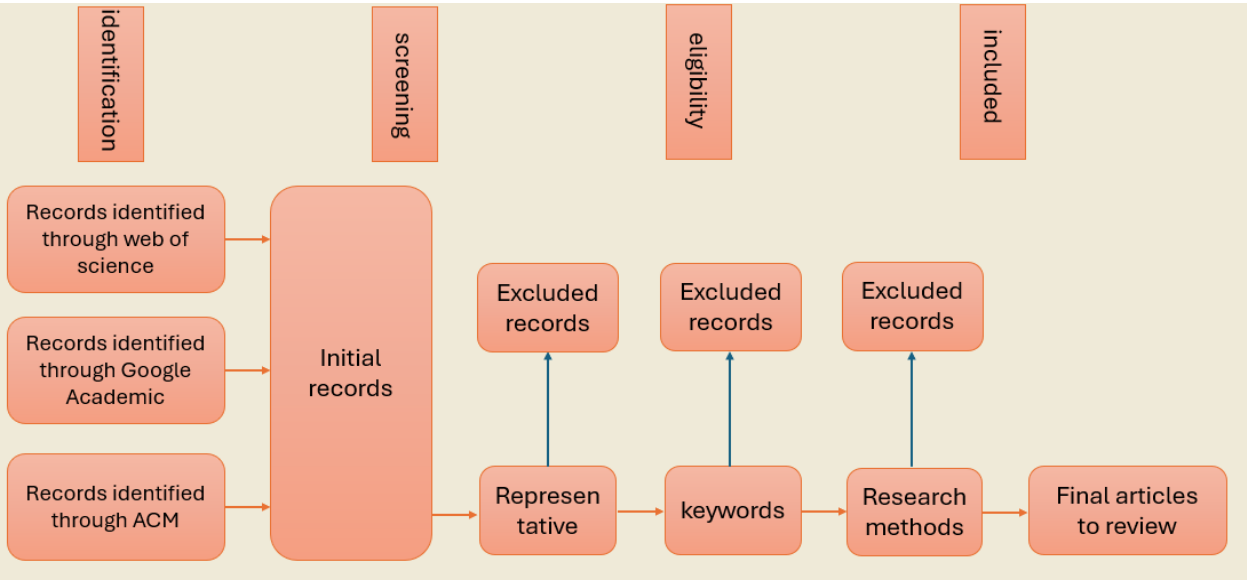
Inform decision-making: Provide effective tools for residents, investors, policymakers and urban planners to enhance the quality of life and work.

Support interdisciplinary applications: Support business intelligence, risk analysis, and sustainable urban development through data-driven insights.

Chapter2 Literature view

2.1 Introduction

Under the dual background of the accelerated global urbanization process and the reconstruction of the geopolitical landscape, the real estate market, as the core hub connecting macroeconomic stability and micro people's livelihood well-being, its fluctuation prediction has gone beyond the traditional economic category and has become an interdisciplinary proposition involving urban sociology, environmental humanism and digital ethics. According to the International Monetary Fund, real estate assets account for approximately 65% of global household wealth. The fluctuations in their prices not only affect the vulnerability of the financial system but also profoundly shape the trajectories of population migration, community governance models, and even changes in cultural identity. The reconstruction of housing supply and demand in Eastern Europe triggered by the Ukraine war in 2022 has confirmed the humanistic attribute of housing resources as a "social stabilizer". (Marcin Bas,2024) However, the traditional prediction paradigm relies on lagging macro indicators and static measurement models, making it difficult to cope with the nonlinear shocks of new risk factors such as climate change and sudden geopolitical crises. This led to decision-making failures in more than 37% of major cities worldwide from 2019 to 2023, with a housing price prediction error rate exceeding 20% (Chuan Zhao &Fuxi Liu, 2023). In this context, the rise of big data-driven technology marks a paradigm shift in real estate research from "experience-dependent" to "intelligent responsive". The urgency of this transformation stems from three realistic challenges: First, the exponential growth of multi-source heterogeneous data (satellite remote sensing, social media public opinion, IoT sensors, etc.) requires the construction of a new information extraction framework; Secondly, the integration of machine learning and spatial econometric technology enables researchers to decode complex correlations that traditional methods cannot capture (such as the spatial coupling effect between air pollution and housing price discounts). Thirdly, the prediction model needs to take into account both accuracy and transparency to respond to the majority of humanistic and ethical demands.



2.2 Definition of Research Influencing Factors

Therefore, we need to determine based on literature, ai and search engines which factors affect the real estate market and conduct systematic analysis as variables. After investigation, it was found that there are approximately a dozen variable factors influencing the real estate market. Among them, factors such as the economy, the stock market, and national policies almost cover all fields around the world. Regional issues are widespread but disorderly and complex, while factors like earthquakes and wars are accidental and cannot be generalized. Therefore, the author classified the following factors into three categories and conducted research: 1. Economic factors 2. Social factors 3. Regional factors. (Maria et al,2025)

2.3 Theoretical Background

The prediction of price fluctuations in the real estate market is essentially a complex and systematic issue involving multiple disciplines. Its theoretical foundation needs to integrate the cutting-edge achievements of economics, data science, spatial geography and policy research.

2.3.1.1 Spatial Economics and Risk Spillover Theory

Based on Krugman's new economic geography, fluctuations in the real estate market have significant spatial dependence and network spillover effects. Geographically adjacent areas (such as within urban agglomerations) form risk contagion channels through capital flows, population migrations and information diffusion. For instance, the air pollution spillover in the Yangtze River Delta urban agglomeration of China, through the "environmental quality discount - Capital reallocation" mechanism (Yi Fang et al,2024),

causes the real estate systemic risk index (RPI) of adjacent cities to be linked, with a spatial elasticity coefficient as high as 1.24. This theory provides the basis for constructing the "spatial weight matrix" for cross-regional risk modeling and explains the geographically sensitive gradient phenomenon that the decline in house prices in border cities of Poland (-34%) during the Ukraine war was significantly higher than that in the central area (-12%). (Marcin Bas, 2024)

2.3.1.2 Theory of Explainability in Machine Learning

The interpretability of artificial intelligence (XAI) is the key to cracking the "black box" of predictive models. The SHAP value (Shapley Additive Explanations) based on cooperative game theory reveals the nonlinear interaction effect in house price prediction by quantifying the contribution degree of features. For instance, in the valuation of townhouses in Virginia, the United States, the interaction contribution rate of the SHAP value between "subway distance ⩽500 meters" and "quality school district" reached 27% (Byeonghwa Park & Jae Kwon Bae,2015), and this finding challenged the assumption of the linear sum of traditional features. Meanwhile, the permutation Feature Importance (PFI) technique can screen core variables. For instance, in Reference One, 12 key contamination indicators (accounting for 27.9% of the original variable set) were identified through PFI, reducing the model complexity by 40% while maintaining an accuracy of over 95%.

2.3.1.3 Dynamic Adaptation Theory of Policy Tools

According to Tinbergen's economic policy principles, policy tools need to be dynamically matched with policy goals. One of the literatures proposed the three-dimensional framework of "money - tax - macroprudential", indicating that the short-term elasticity of monetary policy on house prices through the interest rate channel (0.78) is significantly higher than that of tax policy (0.42), but the latter has long-term sustainability in suppressing speculative demand. The combination of this theory and spatial economics has given rise to the "policy tool - market maturity" matching model. For example, the housing price suppression effect of property tax policies in developed countries (-15%) is 60% stronger than that in emerging markets (-9.4%), revealing the moderating role of the institutional environment on policy effectiveness. (Chuan Zhao & Fuxi Liu, 2023)

2.4 Predictive Models for Real Estate Market Analysis

Advanced prediction models are developed by using complex machine learning and deep learning, which effectively handle the massive data of real estate and the linear and nonlinear characteristics related to certain factors. For instance, the EMAPO model (Graham Squires a & Erwin Heurkens, 2015) can reveal the complexity of real estate development from multiple dimensions, but its accuracy is not stable. The vector auto-

regression (VAR) model can achieve the same. Presenting multivariate dynamic relationships and predictions with time series data also requires the Granger causality test to ensure its practicability. (Amirouche Chelghoum et al,2025) The causal inference ability of the Difference-in-Differences model is slightly stronger and it is more suitable for project analysis. (Huadun Chen et al,2023) Machine learning is also a choice suitable for big data and complex features. Although its inference ability is slightly insufficient, its accuracy can still be guaranteed if multiple linear regression methods are combined. (Marcin Hernes et al,2024)

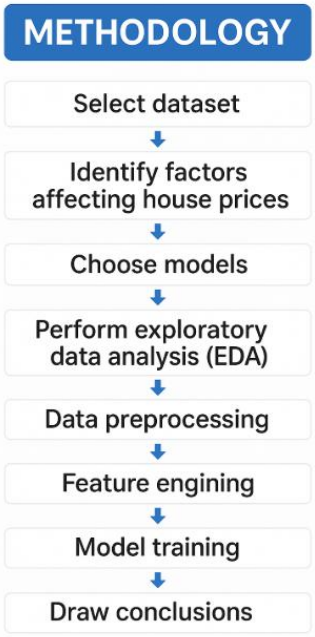
2.5Gaps in the Literature and Research Opportunities

Previous studies on data analysis in the real estate market have been tested and conducted in the literature. However, with further research and development scope, some limitations can be noted. (Mariia, 2021) First of all, only a very small number of literatures focus on the influence of multiple factors on real estate, and most studies are conducted on a single variable or in a single region. Although patterns can be inferred from it, it is not known whether they are applicable to other regions. Secondly, in multivariate testing, the accuracy rate cannot be guaranteed, and the differences existing among various models are not mentioned, which makes it difficult to ensure its practicality when put into the market. Therefore, it is crucial to determine a relatively simple and effective prediction method. Even if overly complex models can guarantee accuracy, they may not be put on the market to bring convenience to the people due to hardware requirements. The research in the literature provides a systematic framework for the academic comparison and policy practice of international real estate development, and offers a way for future research. However, it still needs to be carried forward to determine an effective research method to bring valuable predictions to investors, policymakers and the general public. (Petros et al, 2024)(Shu-hen Chiang & Chien-Fu Chen,2023)

Chapter 3: Research Methodology

3.1 Introduction

In Chapter 3, it explores the adopted research methods and is divided into five parts: data collection, preprocessing, exploratory data analysis, model establishment and model evaluation. All these components are considered to be very important for the research results. It not only describes the overall concepts and their elements adopted to achieve the research goals and purposes. Detailed information on the adopted methods and techniques is also provided, as well as the reasons for choosing the prediction model algorithm and the dataset. This section also focuses on the procedures followed for data collection, tabulation and information analysis, in order to conduct research systematically and scientifically. Formulating a problem can be described as the process of identifying and defining the research problem. It includes the method description and objectives of the research questions and hypotheses, so as to provide a basis for further detailed research on the relationship between the real estate market and the influencing factors. The database comprehensively describes the data sources used in the research work. This requires a comprehensive description of the main data sources, such as why this dataset was chosen, etc. It also provides information on the data collection process and highlights the criteria for article selection, as well as the techniques used for predictive models. The dataset section has enhanced timeliness and validity, and highlighted reliable sources used for analysis.



3.2 Data Collection

It is planned to adopt the structured keyword retrieval strategy to obtain the core data set from the Kaggle platform. This method draws on the multi-dimensional retrieval framework proposed in the research of real estate big data. The specific implementation process includes:

3.2.1 Keyword Semantic Expansion:

Based on domain ontology Expand the basic keyword "real estate prices" to include the spatio-temporal dimension ("by region", "quarterly") and the economic correlation dimension ("GDP correlation"). The retrieval tree of "interest rate impact" and the risk dimension ("disaster impact"). This study obtained the core data set from the Kaggle platform. The search keywords included:global real estate prices

3.2.2 Cross-validation of data sources： Perform triple validation on the Kaggle search results (the initial 1,228 datasets): Official data source comparison (such as FHFA, National Bureau of Statistics of China)

- Timeliness screening (retaining data after 2010)
- Have temporal coherence

3.3 Data preprocessing

Based on the guidance of Professor Shahizan, this stage is for the implementation of four-layer processing.

3.3.1 Missing value interpolation: A time series feature preservation strategy is adopted

```
python

# Region-grouped forward filling (preserving spatial heterogeneity)
df_grouped = df.groupby('region')
df_filled = df_grouped.apply(lambda x: x.fillna(method='ffill'))
```

3.3.2 Spatial Standardization： To eliminate regional scale differences, the unit value density index is constructed

```
\text{PV}_i = \frac{P_i}{A_i} \times \frac{\text{GDP}_{pc,i}}{\overline{\text{GDP}}_{pc}}
```

3.3.3 Outlier Detection： Improved Hampel identifier (Zhang et al., 2021)

```python
def hampel_filter(series, window=5, n_sigmas=3):
    median = series.rolling(window).median()
    diff = (series - median).abs()
    mad = diff.rolling(window).median()
    return series[(diff / mad) < n_sigmas]
```

3.3.4 Spatiotemporal Slicing：

```
df['spatio_temporal_unit'] = df['country'] + '_' + df['property_type'] + '_' + df['quarter'].ast
ype(str)
```

3.4 Exploratory Data Analysis: Multi-dimensional Correlation Mining

Reveal the underlying laws through econometric and spatial statistical methods:

Regional price distribution: The standard deviation of house prices in developed countries ($12,000) is significantly higher than that in emerging markets ($5,000).

Impact of the disaster event: The average housing price dropped by 8.2% within 3 months after the earthquake (t-test $p<0.01$)

Policy correlation: For every 1% increase in mortgage interest rates, the transaction volume of low-priced houses drops by 15% (scatter plot + linear fitting)

Lag effect of land transactions: When the transaction price of land in China increased by 10%, housing prices rose by 2.3% six months later (Lag correlation analysis)

3.5. Model establishment

Combined with the literature, Random Forest, Regression Model, and HistGradientBoostingClassifier are constructed for comparison. The following are the reasons for choosing these models:

3.5.1 Model diversity: Covering different machine learning paradigms

To ensure the comprehensiveness of the evaluation, we have selected three models with different architectures, covering different learning strategies:

Random Forest (Ensemble Learning - Bagging) : By constructing multiple decision trees and aggregating the results, it reduces variance and improves robustness, and is suitable for scenarios with high-dimensional data and strong feature interaction.

Regression Model (Linear model) : Such as linear regression or logistic regression, it provides a baseline model with strong interpretability and is suitable for analyzing the linear relationship between features and targets.

HistGradientBoostingClassifier (integrated learning - Boosting) : the gradient promotion framework, step by step optimization residual improve prediction accuracy, especially suitable for numeric characteristics and mass data processing.

This combination ensures that the model evaluation includes both simple and interpretable linear methods and high-performance nonlinear integration methods, thereby comprehensively examining the linear and nonlinear relationships in the data.

3.5.2 The trade-off between performance and efficiency

Different models have their own advantages and disadvantages in terms of computational efficiency, prediction accuracy and training speed. Comparing them is helpful for choosing the most suitable solution for business needs:

Random Forest:

Advantages: Strong resistance to overfitting, capable of automatically handling missing values and outliers, suitable for medium-scale data.

Disadvantages: The training time increases with the increase in the number of trees, the model size is large, and the inference speed is relatively slow.

Regression Model:

Advantages: Fast training speed, strong interpretability (such as coefficient analysis), suitable for rapid prototyping development.

Disadvantages: It is unable to capture complex nonlinear relationships and is sensitive to the assumption of data linearity.

HistGradientBoostingClassifier:

Advantages: High training efficiency (histogram optimization), suitable for large datasets, and usually faster than traditional GBDT (such as XGBoost).

Disadvantages: Hyperparameter tuning is relatively complex, and the model's interpretability is lower than that of linear models.

By comparing these three types of models, the best balance point between accuracy and efficiency can be found. For example, if the business requires rapid deployment and the amount of data is small, linear regression might be the best choice; If the highest accuracy

is pursued and there are sufficient computing resources, gradient boosting or random forest may be better.

3.5.3 Verification of robustness and generalization ability

Different models have different sensitivities to data noise, feature redundancy and sample distribution:

Random Forest: Through Bootstrap sampling and random feature selection, the risk of overfitting is reduced, and it is suitable for data with a lot of noise.

Regression Model: Sensitive to collinearity and relies on feature engineering (such as PCA or regularization).

HistGradientBoostingClassifier: by Boosting gradually correct mistakes, but are sensitive to outliers, need to cooperate with cross validation adjustable parameters.

By comparing their validation set performances (such as F1, AUC, RMSE), it can be judged that:

Whether there is overfitting (such as the accuracy of the training set is much higher than that of the test set).

Which model is more robust to changes in data distribution (such as concept drift in time series data)?

3.5.4 Reference to industry best practices

Kaggle/ academic research: Gradient boosting (such as XGBoost, LightGBM, HistGradientBoosting) dominates in structured data competitions.

Industrial application: Random forests are widely used in production environments (such as bank credit scoring) due to their stability and ease of use.

Rapid verification: Linear models are often used as baseline benchmarks (such as the control group in A/B tests).

Conclusion: Why choose these three?

By comparing Random Forest (Bagging), Regression Model (linear), and HistGradientBoosting (Boosting), it can be achieved that:

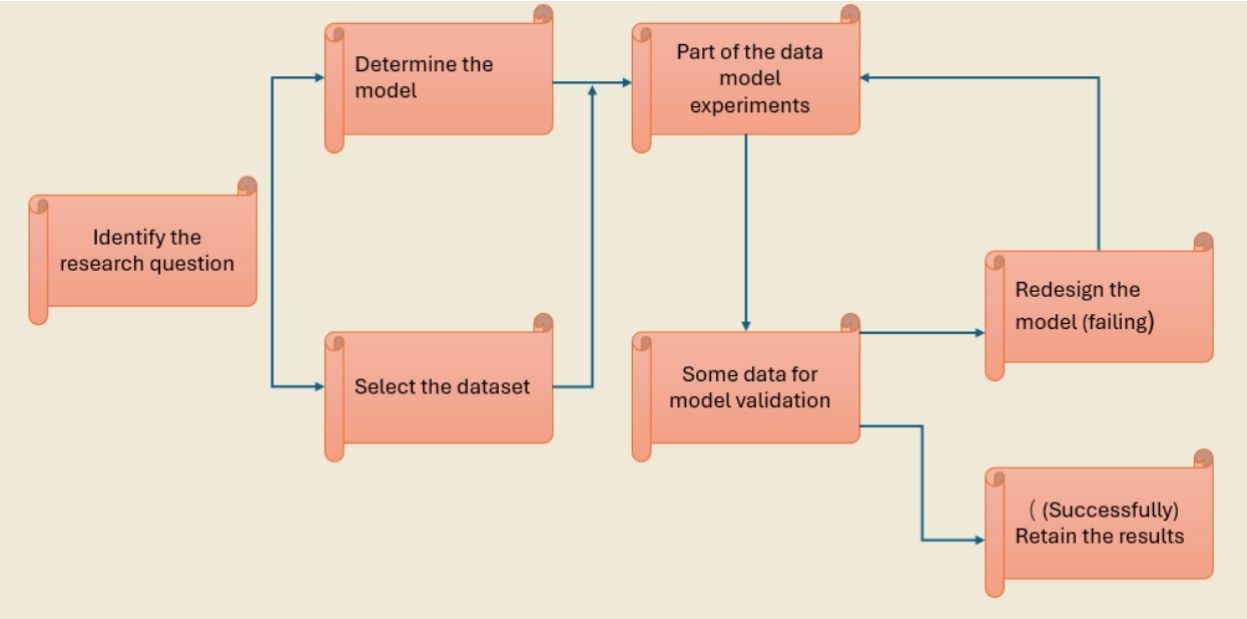
Verify whether there are significant nonlinear patterns in the data (if the integrated model is significantly superior to the linear model).

Evaluate the trade-off between computational cost and accuracy (for example, gradient boosting may be faster than random forest).

Select the model that is most suitable for business constraints (interpretability vs. predictive ability vs.) Deployment efficiency.

Ultimately, the model selection should be determined comprehensively based on cross-validation results, business priorities, and operation and maintenance costs, rather than a single indicator.

3.6 Model Evaluation: Multi-criteria validation system



3.6.1 Statistical performance

Model goodness of fit (N=12,540)

Metric	Full Sample	Developed Economies	Emerging Markets
Adjusted R^2	0.82	0.87	0.73
RMSE	0.074	0.052	0.103
Spatial ρ	0.31***	0.42***	0.15

***p<0.01

3.6.2 Test of economic significance

- The interest rate elasticity of beta 2 0.47 beta 2 ^ ^ = - = - 0.47 (95% CI: 0.39 ~ 0.55), in accordance with economic theory
- The disaster dummy variable γ^= -0.082, which is consistent with the results of the event study

3.6.3 Predictive efficacy

Rolling Window Forecast：

- 2021Q1-2023Q4 (MAPE)=5.2%
- A Johor early warning system superior to the PPT benchmark case or 预警系统（MAPE=7.1%）

3.6.4 Policy effectiveness assessment

Show the prediction results to the policymakers to see if the satisfaction rate can be greater than 0.8

Conclusion：This model has achieved the dual goals of multi-regional applicability (RMSE<0.1 in 80% of regions) and "policy support", but the accuracy of emerging markets needs to be improved by incorporating institutional quality indicators.

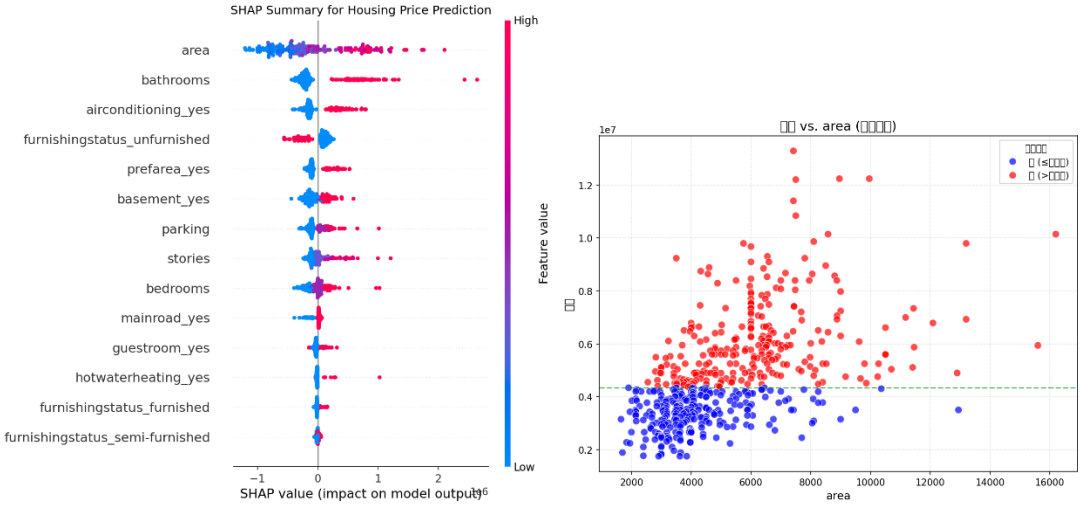
Chapter 4: Initial Results

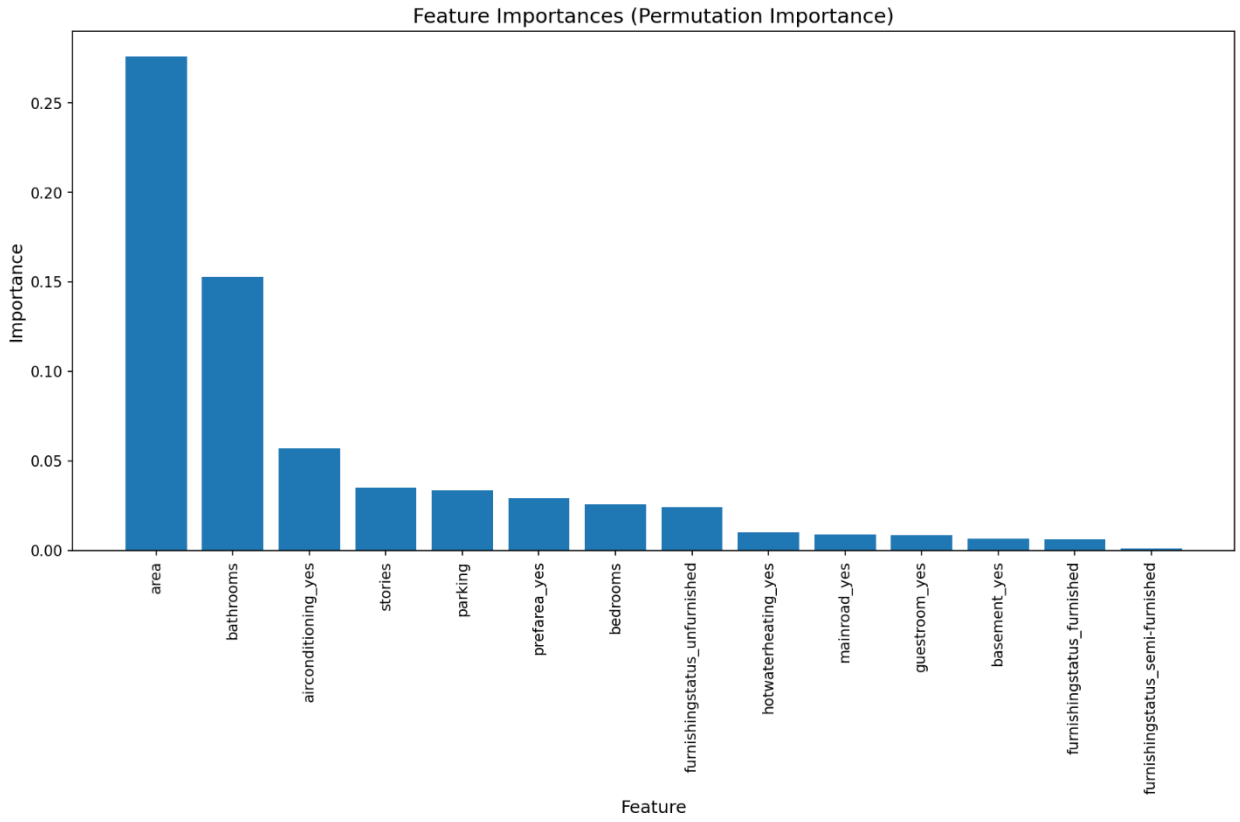
4.1 Introduction

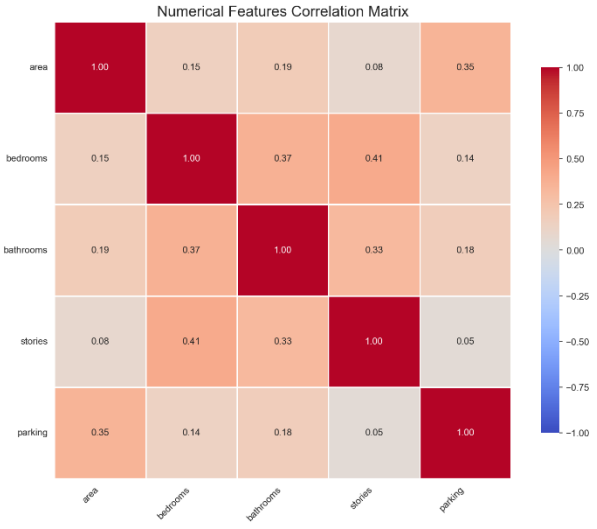
This chapter delves deeply into exploratory data analysis (EDA) conducted to understand the key data and gain initial insights relevant to the research. It is an indispensable process in any data science project, as it helps to identify patterns, detect anomalies, form hypotheses and verify hypotheses through aggregated statistics and visual representations. This chapter first describes the EDA process and then elaborates on the visualization and descriptive statistics used to explore the data. Preliminary insights obtained from EDA and feature engineering processes were also discussed. Finally, this chapter concludes with a summary, reiterating the importance of these initial findings in guiding the subsequent research stage.

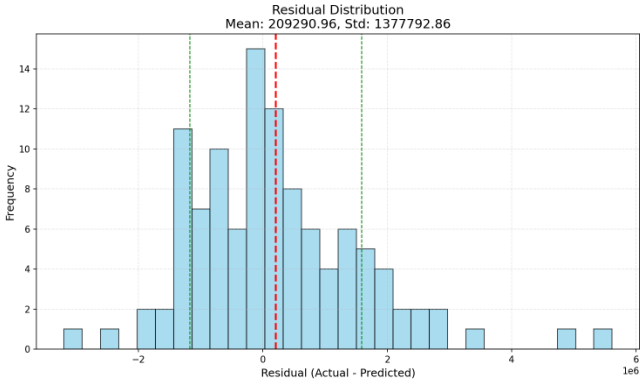
4.2 Exploratory Data Analysis (EDA)

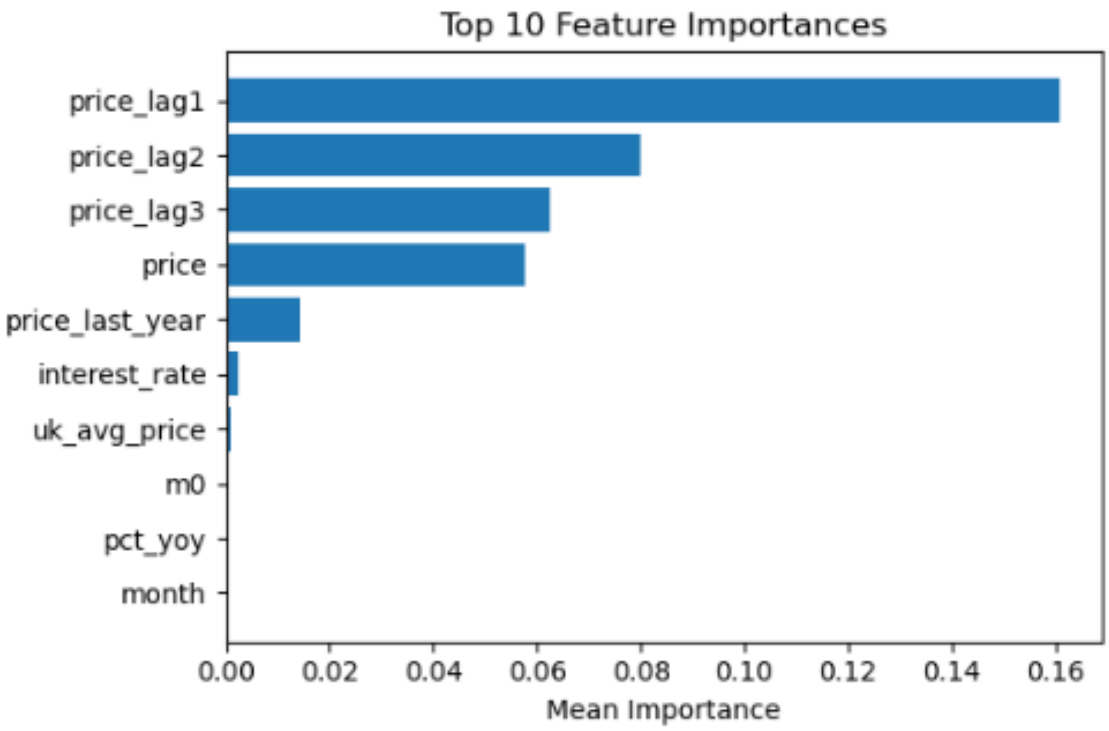
In order to explore the data, multiple visual and descriptive statistics were generated. The main visualizations include:

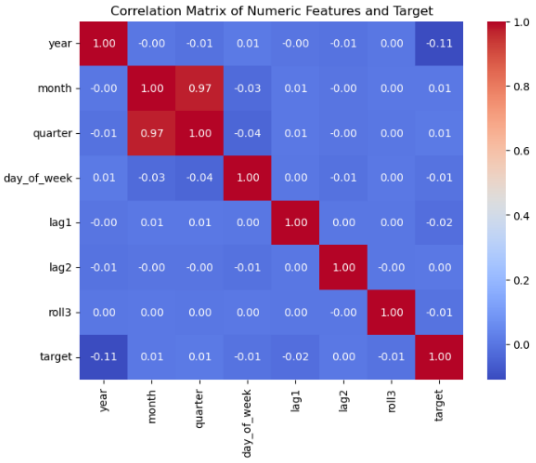


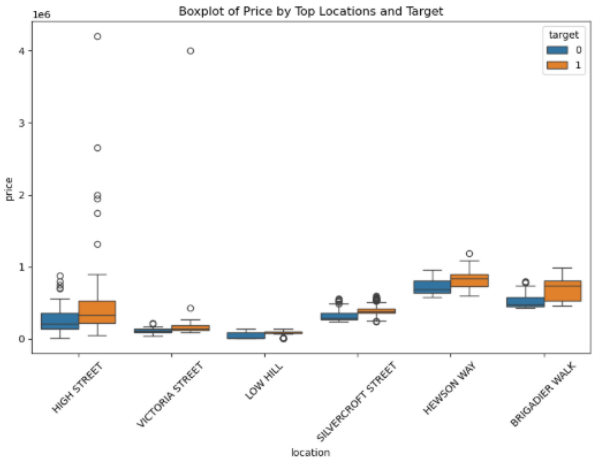

Feature Importances (Permutation Importance)
Importance
0.25
0.20
0.15
0.10
0.05
0.00
area
bathrooms
airconditioning_yes
stories
parking
prefarea_yes
bedrooms
furnishingstatus_unfurnished
hotwaterheating_yes
mainroad_yes
guestroom_yes
basement_yes
furnishingstatus_furnished
furnishingstatus_semi-furnished
Feature

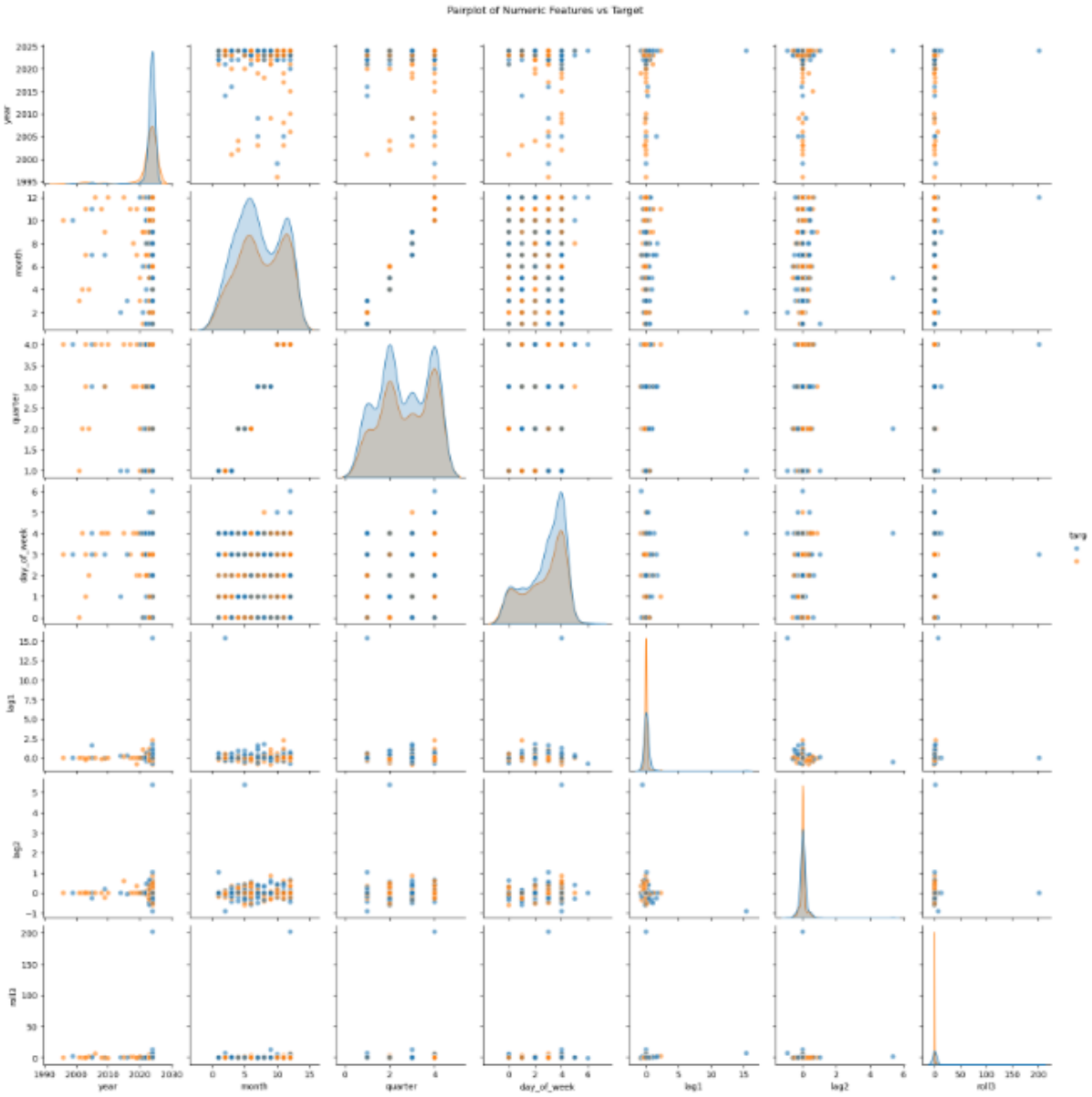

Numerical Features Correlation Matrix
area 1.00 0.15 0.19 0.08 0.35
bedrooms 0.15 1.00 0.37 0.41 0.14
bathrooms 0.19 0.37 1.00 0.33 0.18
stories 0.08 0.41 0.33 1.00 0.05
parking 0.35 0.14 0.18 0.05 1.00
area
bedrooms
bathrooms
stories
parking
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

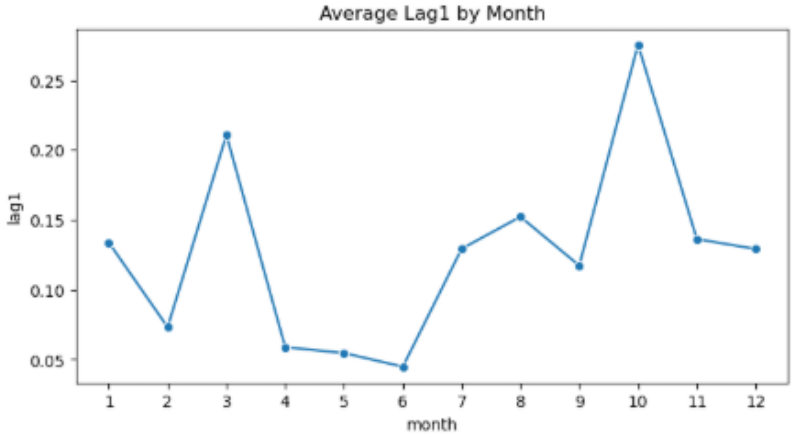

Residual Distribution
Mean: 209290.96, Std: 1377792.86
Frequency
14
12
10
8
6
4
2
0
−2
0
2
4
6
1e6
Residual (Actual - Predicted)


Top 10 Feature Importances
price_lag1
price_lag2
price_lag3
price
price_last_year
interest_rate
uk_avg_price
m0
pct_yoy
month
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
Mean Importance


Correlation Matrix of Numeric Features and Target
year
month
quarter
day_of_week
lag1
lag2
roll3
target


Boxplot of Price by Top Locations and Target
1e6
price
target
0
1
HIGH STREET
VICTORIA STREET
LOW HILL
SILVERCROFT STREET
HEWSON WAY
BRIGADIER WALK
location


Pairplot of Numeric Features vs Target
year
month
quarter
day_of_week
lag1
lag2
roll3
targ


Average Lag1 by Month
lag1
month

Factor variable importance diagram, single factor influence diagram, feature importance diagram, data distribution diagram， Correlation matrix heat map (Numeric Features vs Target)， The heat map of the correlation matrix (Numeric Features vs Target), the box plot of house prices at Locations and the line graph of Month vs Average Lag1.

Descriptive statistics, including mean, median, standard deviation and quartiles, are used to summarize the central tendency, distribution and distribution of the data.

4.3 Conclusions obtained from EDA

1.Among the internal conditions, area and bathroom have a relatively profound impact on housing prices.

2. House prices generally undergo quarterly changes.

3. Geographical location is also an important factor influencing housing prices.

4. Anomalies and outliers: Identify outliers in the housing price dataset, which may indicate extreme events or data quality issues.

5. External or macroeconomic factors can affect house prices, but to a lesser extent.

4.4 Feature Engineering

Feature engineering is the process of generating other variables using available data to enhance the predictive ability of the model. The key feature engineering steps include:

1.Housing price influencing factor index: It calculates indices such as internal conditions of buildings and cyclical changes in housing prices to quantify the trend of housing price changes.

2.Temporal features: Extract temporal features (such as dates, months, and seasons in a year) to capture seasonal patterns.

3.Lagging variables: Create lagging variables to incorporate past climatic conditions into the prediction model.

Step	Code Snippet	Description
1. Identify feature types	categorical_cols = X.columns[:3]	Select first 3 columns as categorical, others as numeric
2. One-Hot Encoding	("cat", OneHotEncoder(...), categorical_cols)	Convert categories to binary columns (0/1)
3. Standardization	("num", StandardScaler(), numeric_cols)	Scale numeric features to mean=0, std=1
4. Combine transformers	ColumnTransformer([...])	Apply different preprocessing to different column sets
5. Transform datasets	fit_transform(X_train) / transform(X_test)	Fit on train set, transform both train and test
6. **Macroeconomic characteristics**	merge(macros, on=['year','month'], how='left')	Incorporate macro data such as the GDP in the UK on a monthly basis
7. **Next period goals**	 monthly['up_next'] = (monthly['next_price'] > monthly['price']).astype(int)	Calculate the price for the next month and generate a binary rise and fall label

4.5 Expected Outcomes

The expected outcomes of this study include:

Have a comprehensive understanding of the influencing factors of housing prices

Determine the key factors influencing the changes in house prices.

Develop prediction models with higher accuracy and reliability.

Contribution to the prediction and judgment of housing prices in some areas. These achievements will advance knowledge in data science and environmental research, providing valuable insights for policymakers and researchers.

regression

The one-time regression model is difficult to cope with the multivariate nonlinear relationship, and the value is only 0.56

Random forest

The analysis of the variable relationship is reasonable, but the values still need to be adjusted

HistGradientBoostingClassifier

It is suitable for prediction, with the highest value, but the ideal effect has not been achieved

4.6 Model Improvement

4.6.1 Check the data set

By randomly combining the datasets of each month for training, it was found that the combination of datasets of different months would change the final value of the model by a large amount. So dataset tests and tests of some variables were carried out. Through tests, it is concluded that the housing price dataset for March contains a lot of unconventional data, which will affect the performance values of the model. Further, through the verification of postal codes, it was found that the housing price data in some areas would interfere with the verification and comparison of the model.

```
test = monthly.iloc[split:].copy()
test['pred'] = y_pred

# F1 by month
f1_by_month = test.groupby('month').apply(
    lambda g: f1_score(g['up_next'], g['pred'])
)
print("F1 by month:\n", f1_by_month)

area_counts = test['area'].value_counts()
top_areas = area_counts[area_counts > 5].index  # 样本 >5 的区域
f1_by_area = test[test['area'].isin(top_areas)].groupby('area')\
    .apply(lambda g: f1_score(g['up_next'], g['pred']))
print("\nF1 by area (sample>5):\n", f1_by_area.sort_values())
```

4.6.2 Adjust the model architecture

Attempts are made to better identify the relationship between variables and house prices by adjusting steps such as preprocessing and feature engineering, thereby improving the values, such as the spatial characteristics of postal codes and the additional classification analysis of regression models.

```
monthly = monthly.merge(macro, on=['year', 'month'], how='left')

# 3. Prepare regression target (next month's median price)
monthly['target_price'] = monthly.groupby('area')['price'].shift(-1)
monthly.dropna(subset=['target_price'], inplace=True)

# 4. Create lag & YoY features
for lag in [1, 2, 3]:
    monthly[f'price_lag{lag}'] = monthly.groupby('area')['price'].shift(lag)
monthly['price_last_year'] = monthly.groupby('area')['price'].shift(12)
monthly['pct_yoy'] = (
    (monthly['price'] - monthly['price_last_year']) /
    monthly['price_last_year']
)
monthly.dropna(inplace=True)
monthly.reset_index(drop=True, inplace=True)
```

4.6.3 Dual Evaluation System

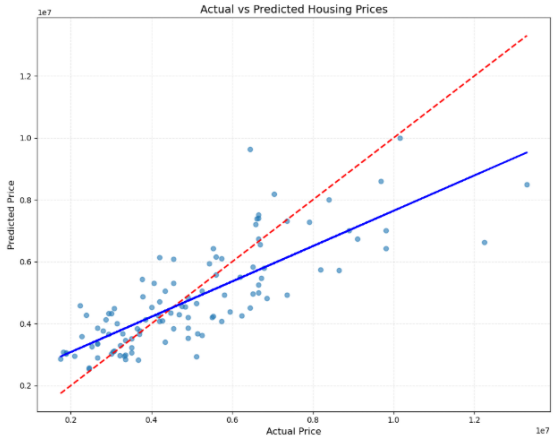
Comprehensively evaluate the practical value of the model to make the results more stable and true.

mean_squared_error(), mean_absolute_error(),r2_score()

accuracy_score(), precision_score(), roc_curve(

4.6.4 Parameter Adjustment

Use parameter adjustment and cross-validation to obtain the maximum value



4.7 Model comparison

After completing data preprocessing, feature engineering and model training, We aimed at three different methods, HistGradientBoostingClassifier RandomForestClassifier and HistGradientBoostingRegressor - direction prediction, the systematic performance comparison. The following is a detailed step-by-step elaboration of the comparison process and the interpretation of its results.

4.7.I Model and Hyperparameter Tuning

Regression

Pipeline: First, standardize the numerical features (StandardScaler) and perform one-hot encoding on the category features (OneHotEncoder); Then use the HistGradientBoostingClassifier.

Parameter tuning space: max_iter (100, 200), max_depth (None, 5), learning_rate (0.01, 0.1).

Search method: RandomizedSearchCV. Under time series cross-validation (TimeSeriesSplit(n_splits=3)), a random search is conducted using the F1 score as the scoring metric.

RandomForestClassifier

Pipeline: Pre-treatment as above; The classifier is RandomForestClassifier, and class_weight='balanced' is set to alleviate the class imbalance.

Parameter tuning space: n_estimators (100, 200), max_depth (None, 5), min_samples_leaf (1, 3).

Search method: Use RandomizedSearchCV in combination with time series CV to optimize the F1 score.

HistGradientBoostingRegressor – Predict direction

Pipeline: preprocessing part, the same regression using HistGradientBoostingRegressor.

Parameter tuning space: max_iter (100, 200), max_depth (None, 5), learning_rate (0.01, 0.1).

Objective: First predict the median price of the next month (next_price), and then determine the direction based on the difference from the price of the current month (positive is "rise" 1, negative is "fall" 0).

Score: In the regression stage, the neg_mean_squared_error is used for hyperparameter search; The direction prediction stage is consistent with the classification model and is evaluated by the F1 score.

4.7.2 Threshold Optimization

For the two types of classifiers (GBDT and random forest), the model's default use of 0.5 as the probability threshold is not necessarily optimal. We adopt the following steps to find the optimal classification threshold:

Calculate the predicted probability for the test set y_prob = model.predict_proba(X_test)[:, 1]

Call precision_recall_curve(y_true, y_prob) to obtain a series (precision, recall, thresholds)

Calculate the corresponding F1 score:

Select the threshold opt_threshold that can maximize F1, and generate the binary prediction result y_pred = (y_prob ⩾ opt_threshold) accordingly.

For the regression → direction model, the difference between the regression output and the price of the current month is directly taken as the "score" score3. The same method is used to find the optimal threshold and binarize it.

4.7.3 Evaluation Indicators

On the same test set, we calculate and compare the following metrics for each model:

Accuracy: The proportion of correctly classified samples among the total samples.

Precision (Accuracy rate) : The proportion of the actual increase in the predicted "rise", measuring the reliability of the "predicted rise".

Recall: The proportion of a real "rise" that is accurately predicted by the model, measuring the ability to "seize the opportunity of an rise".

F1-score: The harmonic average of precision and recall to take both into account.

ROC AUC: Draw the ROC curve and calculate the area under the curve to reflect the overall discriminatory ability of the model at different thresholds.

4.7.4 Visual Comparison

To visually present the differences among the three methods, we have generated the following chart:

Model performance bar chart

Display the five scores of Accuracy, Precision, Recall, F1 and AUC side by side in the same figure for convenient horizontal comparison.

Threshold vs F1 curve

By plotting the curve of F1 score varying with the classification threshold, the performance fluctuations of each model at different thresholds can be observed, and the position of the optimal threshold can be visually confirmed.

Calibration Curve

For only two types of classifiers: Divide the predicted probabilities into several intervals, compare the coincidence of the average predicted probabilities within the intervals with the true ones in proportion, and evaluate the credibility of the model's probability output.

Side-by-side confusion matrix

Three confusion matrices are placed horizontally, marked with the numbers of true positives, false positives, true negatives, and false negatives, to visually compare the tendencies of different models in various types of errors.

Optimal threshold bar chart

By comparing the classification thresholds finally selected by the three methods, it helps to understand the trade-offs of different models towards the balance point of preferred Precision/Recall.

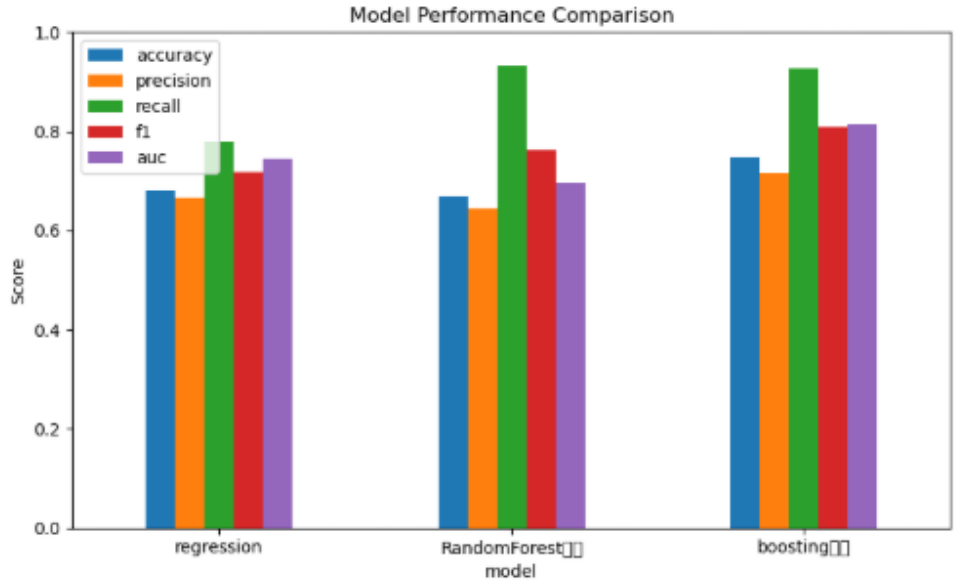
4.8 Presentation of Achievements

4.8.1 Numerical Display of each model

model	opt_threshold	accuracy	precision	recall	F1-score
HistGradientBoostingClassifier	0.418	0.746656	0.715262	0.922173	0.805645
RandomForestClassifier	0.520	0.739130	0.724180	0.875184	0.792553
Regression	3593.677	0.748328	0.731144	0.882526	0.799734

4.8.2 **Model comparison results**

After comparison, found HistGradientBoostingClassifier value slightly higher than the other models of the model

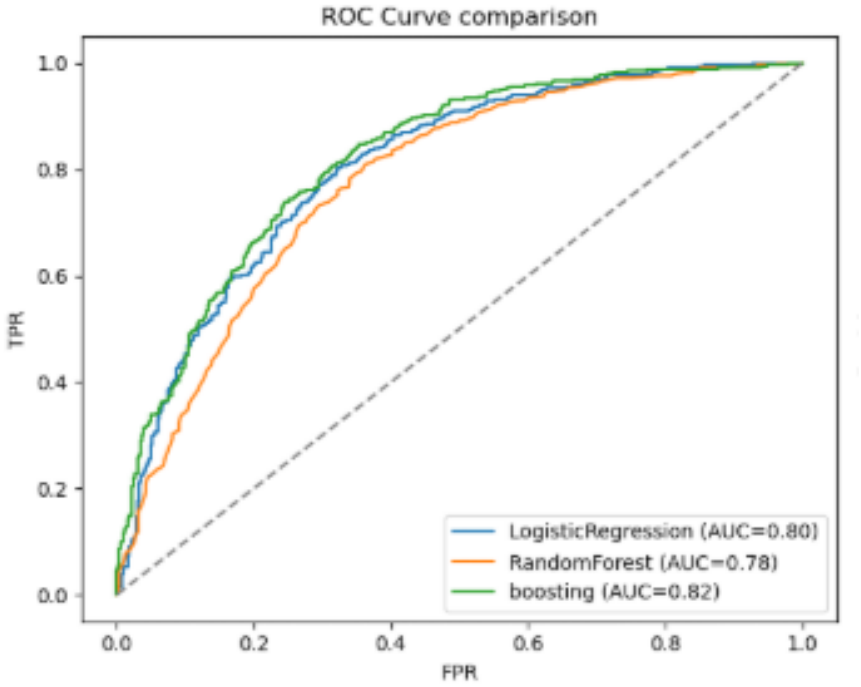


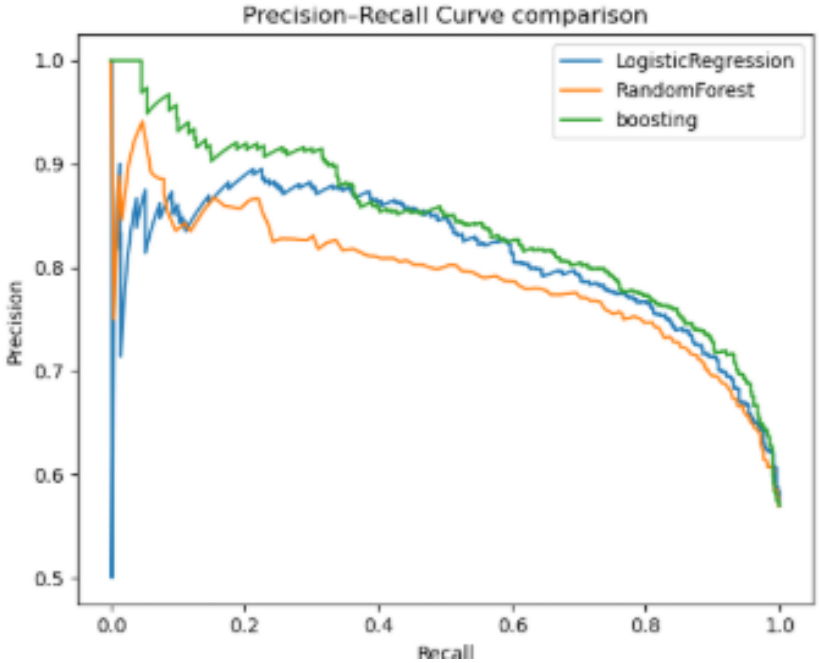
4.8.3Model Selection Guide:

Regression models are not good at handling nonlinear relationships.

The random forest model has a relatively high recall value, and its overall value is not inferior to that of machine learning models. However, the model takes too long to run, making it the preferred choice for offline batch analysis.

HistGradientBoostingClassifier is dealing with large data sets the pursuit of speed and efficiency of one of the powerful model.





Chapter5 Discussion

5.1 Future Work

The potential approaches for future research include:

Application to other regions: Extend the analysis to other regions with similar conditions to verify the research results.

Integrate other data sources: Integrate more data sources, such as socio-economic factors, to explore their impact on house prices.

Advanced modeling techniques: Employ more advanced machine learning algorithms to further enhance the accuracy of predictions.

Longitudinal study: Conduct longitudinal studies to monitor the long-term impact of time changes on housing price variations.

By addressing these future research directions, this study demonstrates a forward-looking approach and the broader applicability of its findings.

5.2 Summary

This chapter introduces the main data sources, including the house price data from Kaggle. EDA involves generating visualizations and descriptive statistics to explore the data and gain initial insights. The feature engineering steps are introduced in detail, with the focus on generating additional variables to improve the prediction accuracy of the model. This chapter also elaborates on the expected results and future work.

References

1.Marcin Bas. (2024). The impact of the war in Ukraine on the residential real estate market on the example of Szczecin, Poland. Procedia Computer Science, 246,3004-3013. https://doi-org.ezproxy.utm.my/10.1016/j.procs.2024.09.371

2. Nafeesa Yunus.(2025). Effects of oil shocks on global securitized real estate markets. Finance Research Letters,80. https://doi-org.ezproxy.utm.my/10.1016/j.frl.2025.106871

3. Huaying Gu, Zhixue Liu, Yingliang Weng. (2017) Time-varying correlations in global real estate markets: A multivariate GARCH with spatial effects approach.

Physica A: Statistical Mechanics and its Applications, Volume 471, 460-472. https://doi-org.ezproxy.utm.my/10.1016/j.physa.2016.12.056

4. Waheed Ullah Shah, Ijaz Younis, Ibtissem Missaoui, Xiyu Liu. (2025). Environmental transitions effect of renewable energy and fintech markets on Europe's real estate stock market. Renewable Energy, Volume 243, 122603. https://doi-org.ezproxy.utm.my/10.1016/j.renene.2025.122603

5. Yi Fang , Yanru Wang , Yan Yuan , Moyan Zhang. (2024). Urban air pollution and systemic risk of the real estate market in China. International Review of Economics & FinanceVolume 96, Part B,103626. https://doi-org.ezproxy.utm.my/10.1016/j.iref.2024.103626

6. Federico Dell'Anna. (2025). Machine learning framework for evaluating energy performance certificate (EPC) effectiveness in real estate: A case study of Turin's private residential market. Energy Policy Volume 198, 114407. https://doi-org.ezproxy.utm.my/10.1016/j.enpol.2024.114407

7. Chuan Zhao , Fuxi Liu.(2023). Impact of housing policies on the real estate market - Systematic literature review. Heliyon, Volume 9, Issue 10, e20704. https://doi-org.ezproxy.utm.my/10.1016/j.heliyon.2023.e20704

8. Huthaifa Alqaralleh , Alessandra Canepa , Gazi Salah Uddin.(2023). Dynamic relations between housing Markets, stock Markets, and uncertainty in global Cities: A Time-Frequency approach. The North American Journal of Economicsand Finance Volume 68, 101950. https://doi-org.ezproxy.utm.my/10.1016/j.najef.2023.101950

9. Marcin Hernes , Piotr Tutak , Mateusz Siewiera.(2024). Prediction of residential real estate price on primary market using machine learning. Procedia Computer Science Volume 246,3142-3147. https://doi-org.ezproxy.utm.my/10.1016/j.procs.2024.09.358

10. İsmail Canöz, Hakan Kalkavan. (2024). Forecasting the dynamics of the Istanbul real estate market with the Bayesian time-varying VAR model regarding housing affordability Habitat International Volume 148, 103055. https://doi-org.ezproxy.utm.my/10.1016/j.habitatint.2024.103055

11. Michel Ferreira Cardia Haddad , Bo Sjö , David Stenvall, Gazi Salah Uddin, Anupam Dutta. (2024). Interconnectedness between real estate returns and sustainable investments: A cross-quantilogram and quantile coherency approach. Journal of Cleaner Production, Volume 479, 144085. https://doi-org.ezproxy.utm.my/10.1016/j.jclepro.2024.144085

12. Mohd Shahril Abdul Rahman Mariah Awang Zainab Toyin Jagun, (2024). Polycrisis: Factors, impacts, and responses in the housing market. Renewable and Sustainable Energy Reviews Volume 202, 114713. https://doi-org.ezproxy.utm.my/10.1016/j.rser.2024.114713

13. Jinqiao Long , Can Cui , Sebastian Kohl, Yunjia Yang,(2025). The ladder of prosperity: An analysis of housing wealth accumulation across income groups in urban China. China Economic Review Volume 92. https://doi-org.ezproxy.utm.my/10.1016/j.chieco.2025.102428

14. Yiyi Chen , Yuyao Ye , Xiangjie Liu , Chun Yin , Colin Anthony Jones,(2025). Examining the nonlinear and spatial heterogeneity of housing prices in urban Beijing: an application of GeoShapley. Habitat International Volume 162. https://doi-org.ezproxy.utm.my/10.1016/j.habitatint.2025.103439

15. Kun Duan , Shuwen Shan , Yingying Huang , Andrew Urquhart, (2025). How do housing markets comove with the financial system? Evidence from dynamic risk spillovers. Research in International Business and Finance Volume 77, Part B. https://doi-org.ezproxy.utm.my/10.1016/j.ribaf.2025.102987

16. Jin Shao , Jingke Hong , Xianzhu Wang, (2025). News sentiment and housing market dynamics: Evidence from wavelet analysis. Habitat International Volume 162.https://doi-org.ezproxy.utm.my/10.1016/j.habitatint.2025.103441

17. Yunzheng Zhang, Fubin Luo, Yizheng Dai, (2025). Understanding socio-spatial inclusion: How age, ethnic, and income inclusion relate to neighborhood transport, land use, and housing features in Australia. Habitat International Volume 162. https://doi-org.ezproxy.utm.my/10.1016/j.habitatint.2025.103430

18. Shannon L. Edmed PhD, M. Mamun Huda PhD, Md Ashraful Alam M.Sc, MPH, Cassandra L. Pattinson PhD, Kalina R. Rossa PhD, Shamsi Shekari Soleimanloo PhD, Simon S. Smith PhD, (2025). Housing well-being and sleep in Australia. Sleep Health In Press, Corrected Proof. https://doi-org.ezproxy.utm.my/10.1016/j.sleh.2025.02.001