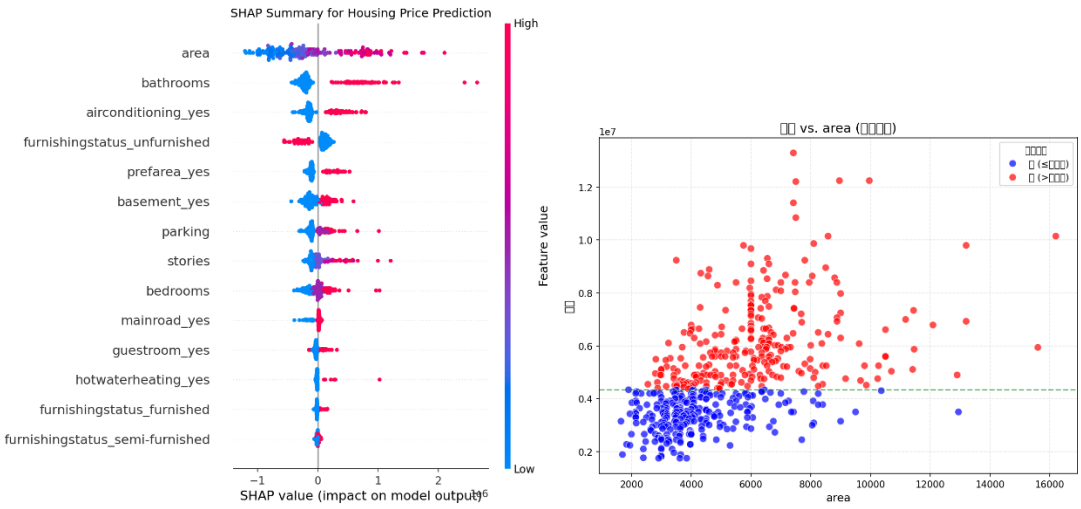
Chapter 4: Initial Results

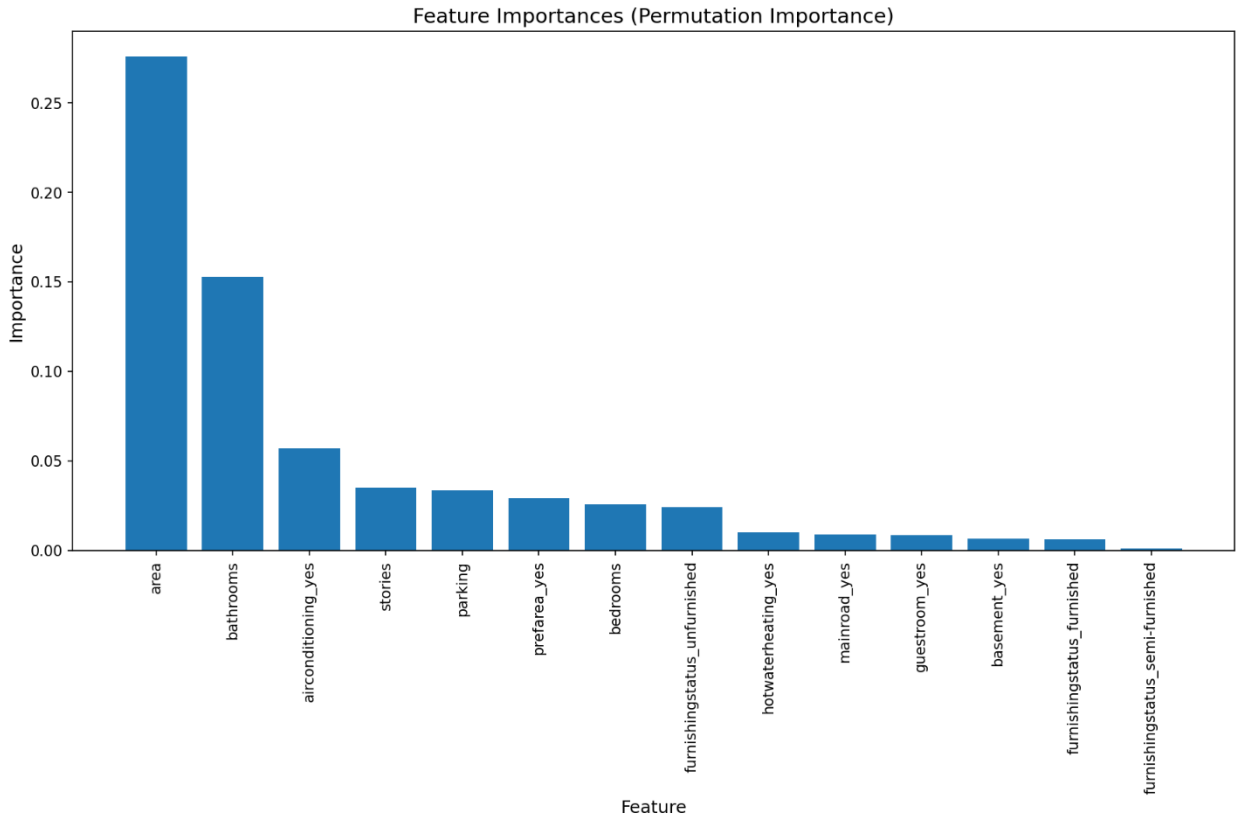
4.1 Introduction

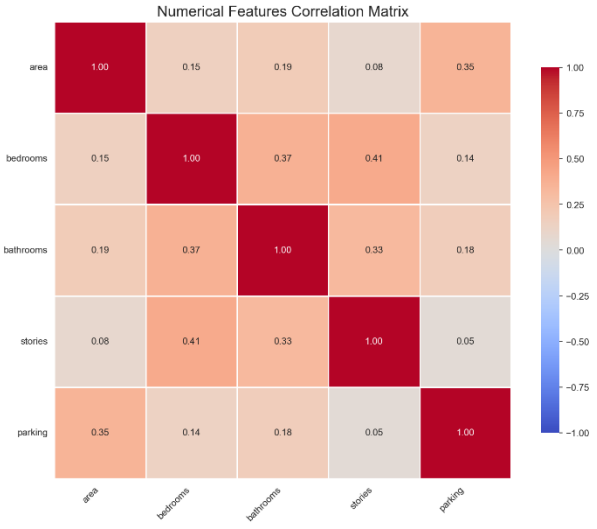
This chapter delves deeply into exploratory data analysis (EDA) conducted to understand the key data and gain initial insights relevant to the research. It is an indispensable process in any data science project, as it helps to identify patterns, detect anomalies, form hypotheses and verify hypotheses 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.

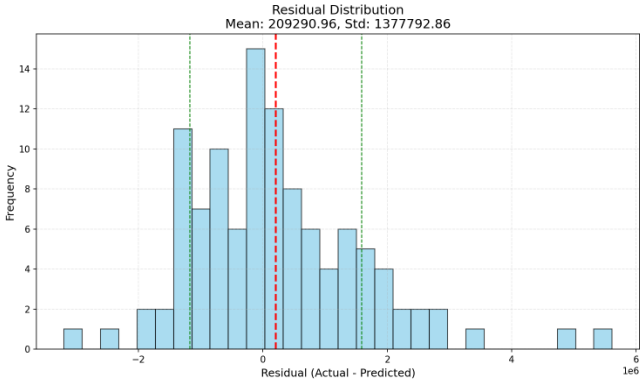
4.2 Exploratory Data Analysis (EDA)

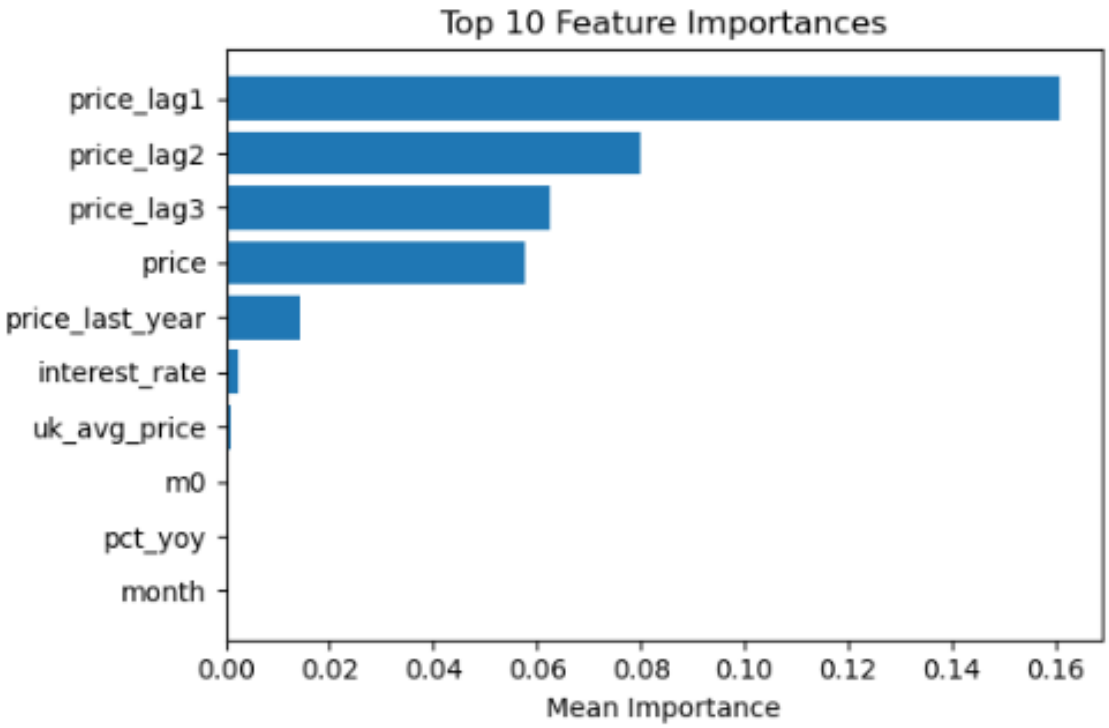
In order to explore the data, multiple visual and descriptive statistics were generated. The main visualizations include:

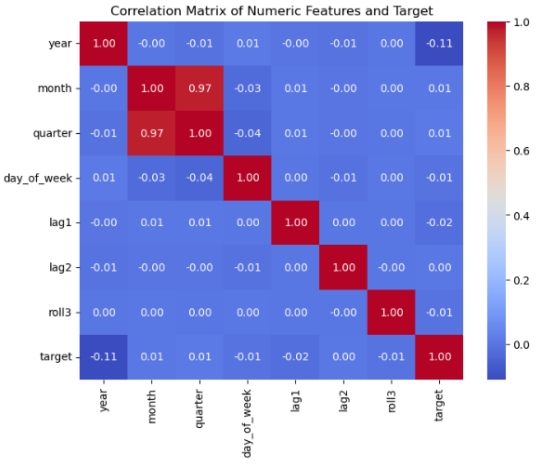


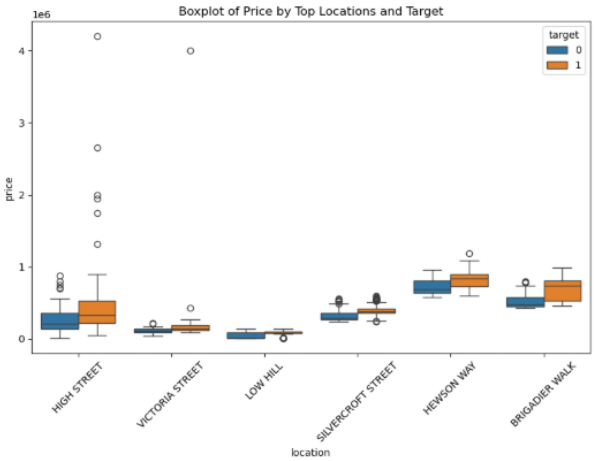

Feature Importances (Permutation Importance)
Importance
0.25
0.20
0.15
0.10
0.05
0.00
area
bathrooms
airconditioning_yes
stories
parking
prefarea_yes
bedrooms
furnishingstatus_unfurnished
hotwaterheating_yes
mainroad_yes
guestroom_yes
basement_yes
furnishingstatus_furnished
furnishingstatus_semi-furnished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

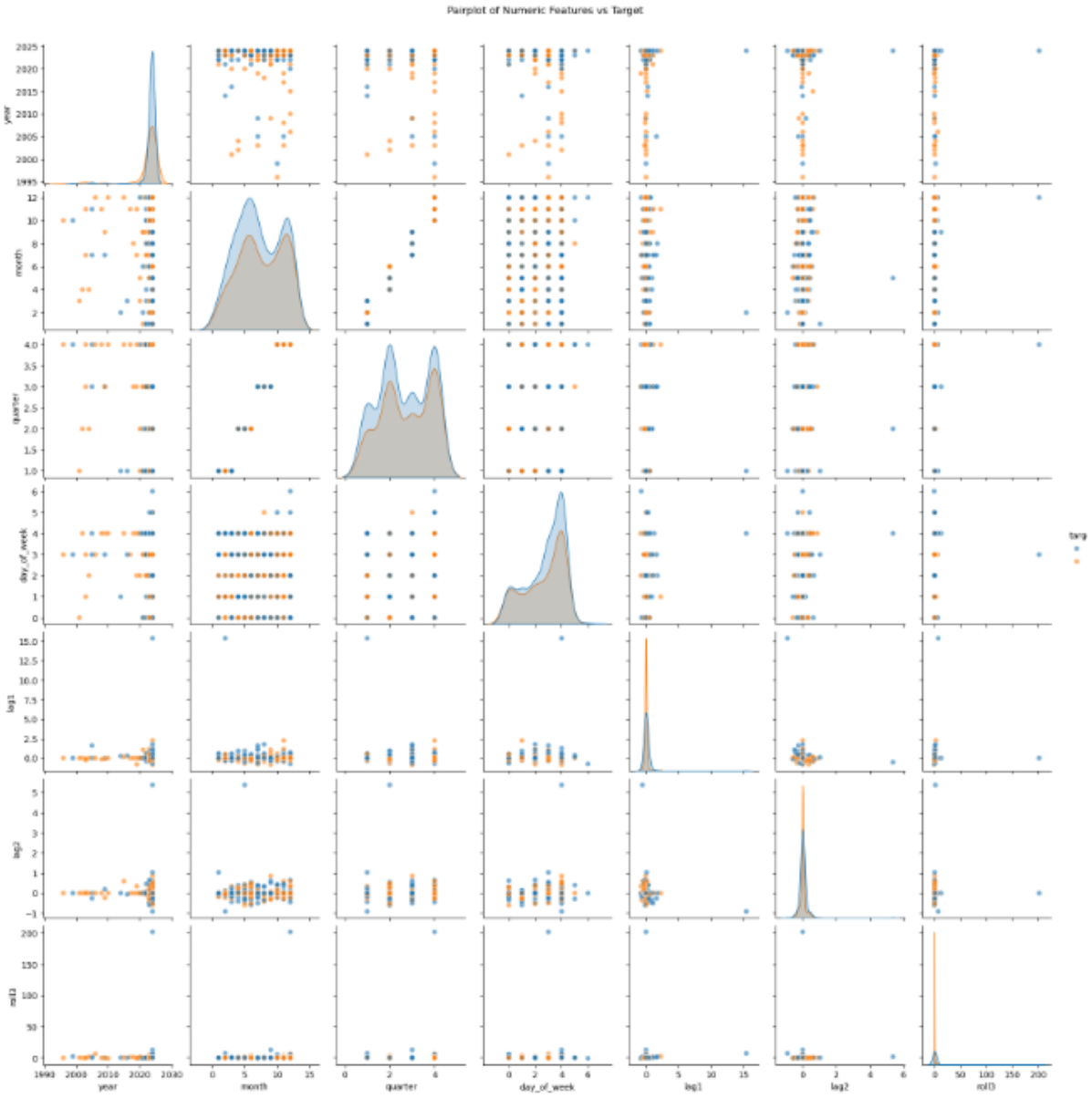

Numerical Features Correlation Matrix
area 1.00 0.15 0.19 0.08 0.35
bedrooms 0.15 1.00 0.37 0.41 0.14
bathrooms 0.19 0.37 1.00 0.33 0.18
stories 0.08 0.41 0.33 1.00 0.05
parking 0.35 0.14 0.18 0.05 1.00
area
bedrooms
bathrooms
stories
parking
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

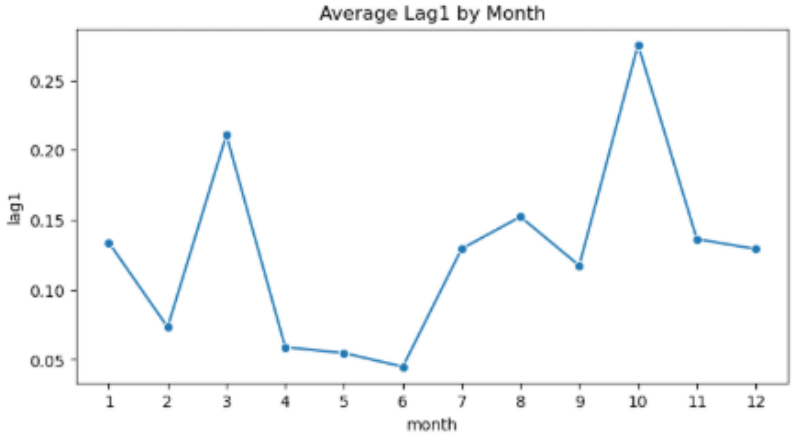

Residual Distribution
Mean: 209290.96, Std: 1377792.86
Frequency
14
12
10
8
6
4
2
0
−2
0
2
4
6
1e6
Residual (Actual - Predicted)


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 1.00 -0.00 -0.01 0.01 -0.00 -0.01 0.00 -0.11
month -0.00 1.00 0.97 -0.03 0.01 -0.00 0.00 0.01
quarter -0.01 0.97 1.00 -0.04 0.01 -0.00 0.00 0.01
day_of_week 0.01 -0.03 -0.04 1.00 0.00 -0.01 0.00 -0.01
lag1 -0.00 0.01 0.01 0.00 1.00 0.00 0.00 -0.02
lag2 -0.01 -0.00 -0.00 -0.01 0.00 1.00 -0.00 0.00
roll3 0.00 0.00 0.00 0.00 0.00 -0.00 1.00 -0.01
target -0.11 0.01 0.01 -0.01 -0.02 0.00 -0.01 1.00
year
month
quarter
day_of_week
lag1
lag2
roll3
target
1.0
0.8
0.6
0.4
0.2
0.0


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
tarq


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

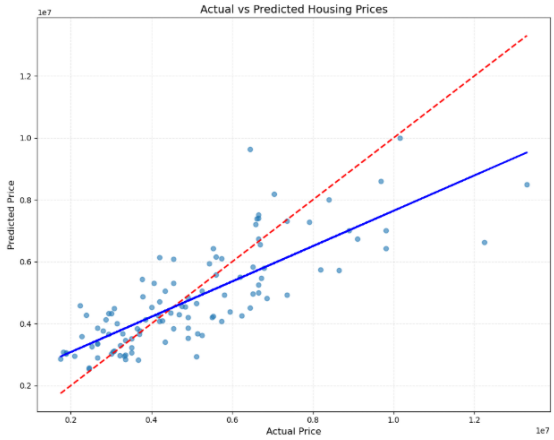
Comprehensively evaluate the practical value of the model to make the results more stable and true.

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 $\geqslant$ opt_threshold) accordingly.

For the regression $\rightarrow$ 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.

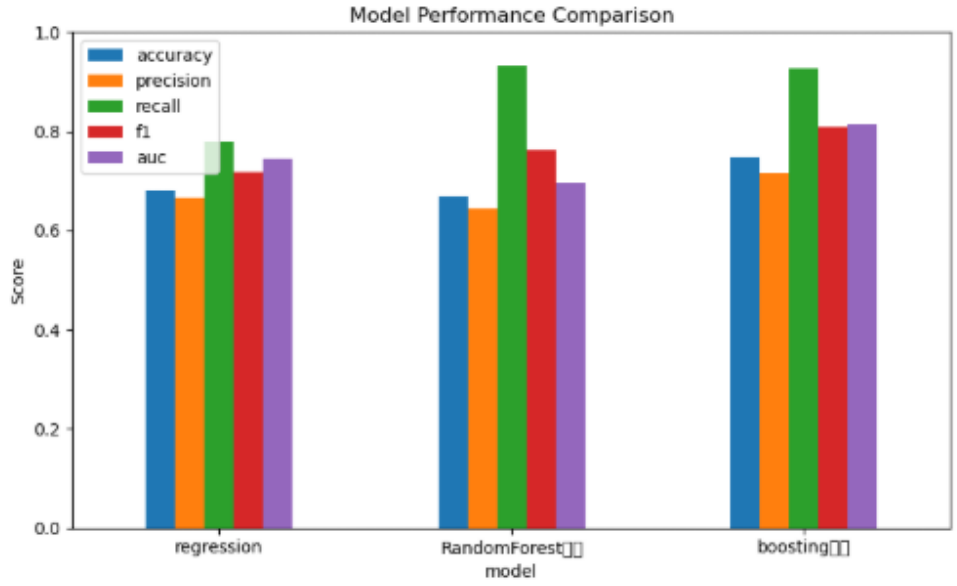
4.8 Presentation of Achievements

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

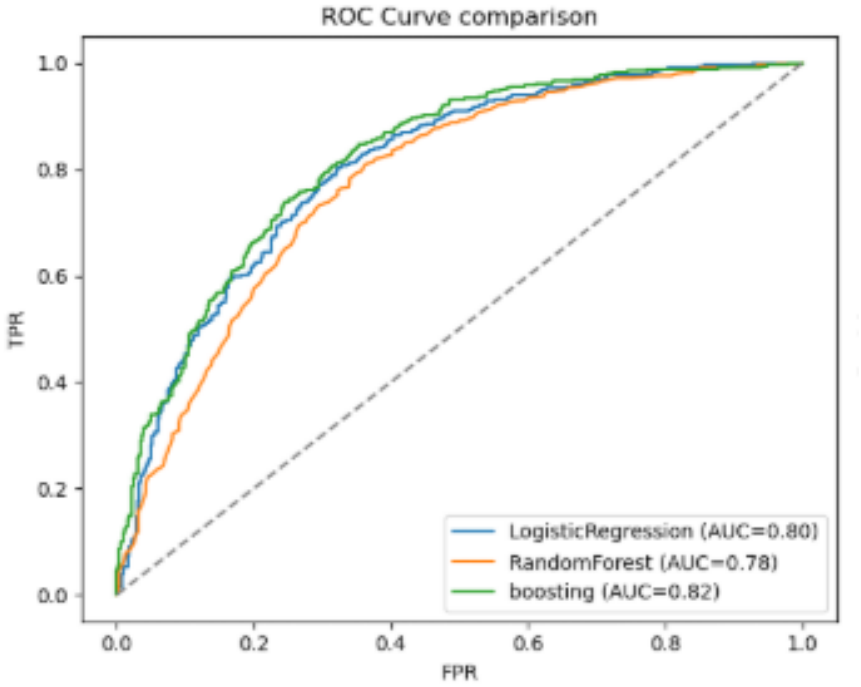


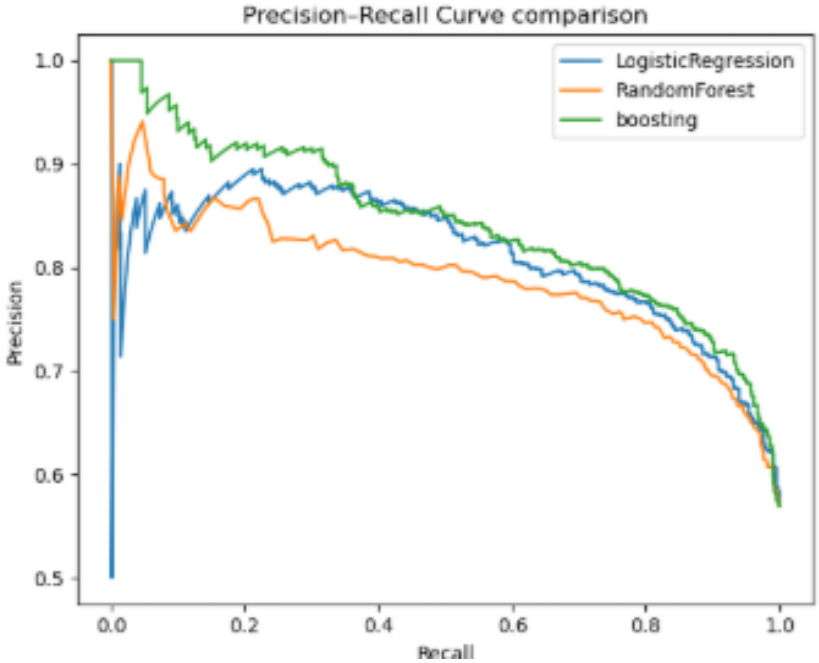
4.8.3Model Selection Guide:

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