Conclusion

This chapter briefly summarizes the main findings of the project, conducts a final observation of its advantages and disadvantages, and puts forward suggestions for further research or improvement.

5.1 Summary of Main Findings

1. The effect of feature engineering is remarkable: After introducing temporal features (year, month, quarter, day of the week), spatial features (postcode area), as well as historical growth lags (lag1, lag2) and rolling mean (roll3), the model can better capture the temporal and spatial patterns of housing price growth.

2. HistGradientBoostingClassifier is superior to other models: Under strict temporal segmentation verification (training in the first four months and testing in the fifth month), HistGradientBoostingClassifier in Accuracy, Recall and F1, ROC AUC on indicators leading RandomForest and LogisticRegression. It indicates that its modeling of category characteristics and nonlinear patterns is stronger.

3. Preventing data leakage is of vital importance: The roll3 feature. shift(1) and the target encoding and hyperparameter search are completed within the Pipeline + TimeSeriesSplit, effectively eliminating the inflated metrics caused by historical data leakage and making the model performance evaluation more reliable.

5.2 Project Success and Limitations

• 5.2.1 Success Points

o has built an end-to-end Pipeline: from reading multi-month data, strict time series segmentation, to feature encoding, model parameter tuning, and visual evaluation, a complete and reusable set of processes has been established.

Multi-model comparison and visualization: Multiple charts such as ROC/PR curves, confusion matrices, calibration curves, and hyperparameter tuning curves were simultaneously generated, visually presenting the strengths and weaknesses of each model.

o proposed the optimization threshold and class imbalance processing scheme (class_weight, target encoding), which significantly improved the prediction ability of minority classes.

• 5.2.2 Limitations

Lack of macroeconomic and external factors: The model only relies on transaction prices, time and location, and does not incorporate macro information such as interest rates and supply and demand indicators, thus limiting its predictive ability.

Rough spatial granularity: postcode_area still belongs to a large geographical unit, and the heterogeneity within the region has not been fully captured. Local hotspots or cold areas may be ignored.

The generalization of the o model remains to be verified: Although it was tested on the data of the fifth month, it did not cover a longer time span. The robustness of the model under different cycles (such as the epidemic and policy fluctuations) is still unclear.

5.3 Future Research or Improvement Suggestions

1. Introduce macro and meso indicators: Integrate external features such as interest rates, mortgage policies, regional construction data, and population mobility into the model to enhance the comprehensive modeling of price drivers.

2. Refine spatial resolution: Attempt to capture small-scale price fluctuations based on more refined geographical units (such as streets and communities), or by using latitude and longitude grids and spatial embeddings (Graph Embedding/Geohash).

Model: 3. The integration and stack on HistGradientBoostingClassifier and LightGBM, random forest model to vote or stacked, with their respective advantages, improve stability prediction.

4. Cross-cycle robustness testing: Extend the data period to one year or longer, and use sliding window CV or rolling validation to evaluate the model's applicability under different market environments (policies, economic cycles).

5. Online learning and dynamic update: Consider deploying an incremental learning mechanism to update model parameters in real time, respond promptly to market emergencies and structural changes, and maintain prediction accuracy.