Evaluating Model Generalization Across Regions in Melbourne Housing Market

Hongtao Yao 1302148, Zeyu Yang 1555413, Zuqiang Yu 1358964

1. Introduction

Housing prices vary across Melbourne's regions due to various factors. If a machine learning model is trained only on data from certain regions, it often only learns the characteristics or patterns of that region and fails to reflect the overall trends across all of Melbourne. This leads to poor performance when applied to new, unseen regions. This phenomenon reveals a lack of generalization when the model is applied across regions.

This study aims to compare the cross-regional generalization capabilities of models of varying complexity, including ridge regression, random forest, and LightGBM, and to assess their predictive abilities in unseen regions. This study uses the Melbourne Housing Dataset (https://www.kaggle.com/datasets/ronikmalhotra/melbourne-housing-dataset) from Kaagle. After preprocessing and feature construction, the dataset contains 27,244 samples and 109 features, covering property prices and various characteristics across Melbourne's four major metropolitan areas.

2. Literature review

According to Sellam et al. (2024) note that early house price prediction relied on traditional regression analysis, primarily considering basic features such as building area and year of construction. However, with the advancement of technology and data resources, researchers have begun to introduce more accurate machine learning algorithms.

One such model is the LightGBM model proposed by Ke et al. (2017). In their paper, the authors improved the traditional gradient boosted decision tree (GBDT) algorithm by introducing two new techniques: gradient-based one-sided sampling (GOSS) and mutually exclusive feature bundling (EFB). GOSS reduces sample size by excluding samples with small gradients, while EFB reduces the number of features by bundling mutually exclusive features. By combining existing techniques such as histogram-based algorithm and leaf-wise tree growth strategy, the authors achieved higher efficiency than other GBDT algorithms while maintaining similar accuracy.

While the performance of modern regression models has improved, spatial heterogeneity remains a major challenge in house price prediction. This characteristic limits the model's transferability, resulting in reduced performance when predicting across regions (Yao & Fotheringham, 2014). This study uses the Melbourne Housing Dataset from Kaggle (Malhotra, 2022), which contains multiple metropolitan areas and detailed real estate information, and is used to evaluate the cross-regional generalization ability of different models.

3. Method

3.1 Data Preprocessing

To ensure data quality and consistency, we first examined missing values and outliers across all features.For features with fewer missing values (such as CouncilArea, PropertyCount, Postcode, and Distance), the corresponding samples were directly deleted. The target variable, Price, is the core prediction object of this study. Given the sufficient sample size, we chose to delete the missing samples instead of using the mean or model imputation to avoid introducing noise.

In the processing of numerical features, Latitude and Longitude are grouped by Suburb and RegionName, and the group means are used to fill in the missing values; Bedroom and Rooms are highly correlated, so the values of Rooms are used to replace the missing items. Similarly, for Bathroom, the missing values are filled in with the group means after grouping by Bedroom.

The distribution of the Car feature is significantly skewed (approximately 80% of the samples are concentrated in the range [1,2]). The initial linear and logistic regression model predict results were limited ($R^2 \approx 0.22$), so the Car was binned and one-hot encoded into three categories: low, medium, and high. After logistic regression prediction, the accuracy rate reached 59.14%.

Further analysis revealed that YearBuilt, Landsize, and BuildingArea also exhibited strong skewed distributions, direct use of KNN prediction ineffective for filling in the data. We adopted a similar processing approach as for the Car category: first, we divided the data into bins to reduce the impact of extreme values and then used KNN for filling. The results showed a significant improvement in model performance: the R^2 of YearBuilt increased to approximately 0.60, while Landsize and BuildingArea reached 0.6153 and 0.7420 respectively.

3.2 Feature Construction

During the data construction stage, the date field "Date" is first split into "Year", "Month", and "Day" to capture the temporal variation characteristics of house prices. Since the number of Suburb categories is large, direct one-hot encoding would lead to dimension explosion and overfitting. Therefore, it is mapped to a unique numerical ID (Suburb_ID) to control the number of features while retaining the geographical information. For the address field, only the house number is retained because latitude, longitude, Suburb, and RegionName can already fully represent the geographical location.

Based on our research question, the city was divided into four directions - east, south, west, and north - centered around the CBD of Melbourne to evaluate the generalization ability of the model in unexplored regions. The orientation of these data was classified based on the direction information of RegionName, which was used to compare the prediction performance in different orientations. The model performance was measured by the coefficient of determination R^2, RMSE and MAE, which are the main evaluation indicators to assess the degree of fitting of the model to the housing price trend and the degree of variance deviation. Additionally, one-hot encoding was performed on categorical features such as CouncilArea, Method, RegionName, ParkingArea and Direction; while SellerG was deleted due to its weak influence and excessive values to reduce noise and dimensionality.

Finally, to capture the nonlinear relationships among the key variables, polynomial feature expansion was performed on Rooms, Bedroom, Bathroom, Distance, and PropertyCount. After the above processing, the final dataset contained 109 valid features.

3.3 Cross Validation

To systematically evaluate the generalization ability of the model across different geographical regions, this study employed a nested k-fold cross-validation based on regional rotation. The Melbourne urban area was divided into four regions: east, south, west, and north. In each experiment, three of these regions were selected as the training and validation sets, while the remaining one served as the independent "unseen region" test set.

During the training phase, the data from the three regions undergoes outer 10-fold cross-validation, with 9 folds used for training and 1 fold for validation. Within each outer training set, an inner cross-validation is conducted, and the hyperparameters of the model are optimized through grid search.

After obtaining the optimal parameters, the model is re-trained on the complete outer training set and the performance metrics are calculated on the validation set. This process is repeated in all 10 cross-validation fold combinations for all three regions to obtain the average validation performance (mean values of R^2, rmse and mae). Subsequently, the model is used with the optimal parameters to make predictions in the unseen region and the performance of the test set is calculated.

By comparing the performance of the validation set and the test set, we analyze the performance differences of the model in the geographical migration task, thereby evaluating the generalization ability of different models.

3.4 Algorithms

3.4.1 Ridge

Ridge regression adds an L2 regularization term to the ordinary least squares objective to suppress multicollinearity among features and stabilize coefficient estimation. It is chosen as the linear baseline model in this study for its simplicity and parameter transparency. The model uses the cleaned and feature-engineered dataset.csv, covering four metropolitan regions of Melbourne (Eastern, Western, Southern, and Northern). To prevent information leakage, one-hot encoded regional columns were removed. Hyperparameter tuning and performance evaluation follow a nested cross-validation combined with leave-one-region-out (LOCO) design: in each experiment, one region is held out as the test set while the remaining three are merged for training. Within the training data, an outer 10-fold cross-validation is performed, and for each outer training subset, an inner 3-fold validation iterates over α ∈ {1, 10, 100} to select the value minimizing the mean RMSE. The selected α is then used to retrain and evaluate the model on the outer validation set, and average performance and parameter selection frequencies are summarized. Finally, the most stable α from the inner validation (typically α = 10) is used to train the model on all training data and tested on the held-out region.

3.4.2 Random Forest

Random Forest aggregates multiple decision trees through feature-subsampled bagging to capture nonlinear relationships and feature interactions while effectively reducing variance. It is adopted as the medium-complexity model to complement the linear baseline. The number of trees is fixed at n_estimators = 200, and the parameter grid includes max_depth ∈ {None, 12, 24}, min_samples_leaf ∈ {1, 5, 10}, and max_features ∈ {1.0, "sqrt", 0.5}. Hyperparameter tuning and evaluation also follow the nested cross-validation + LOCO framework (outer 10-fold, inner 3-fold). The inner folds iterate over all parameter combinations to compute mean RMSE; when ties occur, preference is given to shallower trees, larger leaf sizes, and smaller feature proportions. The outer folds use the best inner parameters to retrain the model and record RMSE, MAE, and R^2. After completing all folds, the configuration with the smallest and most stable inner RMSE is selected to train the final model on all training data and tested on the excluded region. The optimal configurations are generally deep trees (max_depth = None/24/16), min_samples_leaf = 1, and smaller max_features (0.3–0.5).

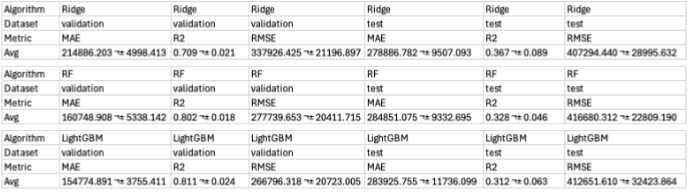
3.4.3 LightGBM

As forementioned, LightGBM is a GBDT optimized for efficiency. This algorithm incorporates two unique techniques: GOSS and EFB, to reduce sample size and feature space (Ke et al., 2017). As the dataset contains high-dimensional, mixed-type data, LightGBM's leaf-wise growth technique (although not unique) should allow for high accuracy when the model is trained on this dataset while achieving high-efficiency with GOSS and EFB. This unique combination of accuracy and efficiency makes LightGBM promising for our research question.

Due to computation resource constraints, the hyper-parameter tuning for LightGBM in our implementation focuses on high-leverage parameters such as number of leaves (num_leaves) and learning rate (learning_rate), with the tuning of these two parameters using 3 values. The rest of the hyper-parameters either have a reduced parameter search grid (in the case of minimum child samples and feature fraction) or have fixed parameter value (in the case of max depth, bagging fraction and the lamdas). The implementation of hyper-parameter is largely the same with previous algorithms, with the exception that the final hyper-parameter set is selected via taking the mode of the best hyper-parameter set from each of the 10 outer folds (as each outer fold would generate 1 best hyper-parameter).

Results and Discussion

After training, validation, and testing, we have obtained the following metrics for each of the three models, averaged across all region combinations:

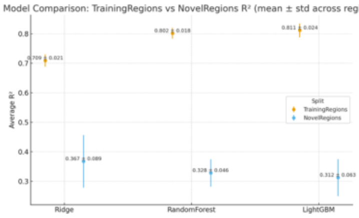


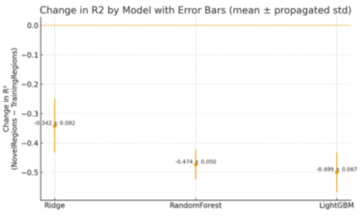
Algorithm	Ridge	Ridge	Ridge	Ridge	Ridge	Ridge
Dataset	validation	validation	validation	test	test	test
Metric	MAE	R2	RMSE	MAE	R2	RMSE
Avg	214886.203 ¬± 4998.413	0.709 ¬± 0.021	337926.425 ¬± 21196.897	278886.782 ¬± 9507.093	0.367 ¬± 0.089	407294.440 ¬± 28995.632

Algorithm	RF	RF	RF	RF	RF	RF
Dataset	validation	validation	validation	test	test	test
Metric	MAE	R2	RMSE	MAE	R2	RMSE
Avg	160748.908 ¬± 5338.142	0.802 ¬± 0.018	277739.653 ¬± 20411.715	284851.075 ¬± 9332.695	0.328 ¬± 0.046	416680.312 ¬± 22809.190

Algorithm	LightGBM	LightGBM	LightGBM	LightGBM	LightGBM	LightGBM
Dataset	validation	validation	validation	test	test	test
Metric	MAE	R2	RMSE	MAE	R2	RMSE
Avg	154774.891 ¬± 3755.411	0.811 ¬± 0.024	266796.318 ¬± 20723.005	283925.755 ¬± 11736.099	0.312 ¬± 0.063	412651.610 ¬± 32423.864

From these results, we can observe a degradation in accuracy in unseen regions. Specifically, the LightGBM algorithm showed an R^2 change of approximately -0.499, while Random Forest and Ridge only showed an R^2 change of approximately -0.474 and -0.342, respectively. These results shows that while our complex algorithm performs well in validation, it does not generalize well when met with unseen regions, this trend can also be observed in the error bar figures:





This indicates that there are subtle region-specific feature interactions that differ from region to region. Without the 4th region, complex models cannot learn its region-specific patterns and therefore overfits the other three regions.

To observe the generalization performance of model complexity in unseen regions, we compared the performance of three models in different unseen regions. The ridge regression model focuses on the linear weights of data features and is robust in controlling multicollinearity. The R^2 in all four unseen regions remains positive, but it is sensitive to complex nonlinear relationships in the features. Random forest is better at capturing nonlinear relationships and has high accuracy during training. However, it also over-relies on local patterns in the training area. When switching to unseen regions, the R^2 fluctuates significantly, and the r^2 in the west region is even close to 0. This may be because the housing prices or characteristics in the west region are very different from the three regions seen during training. LightGBM is the best in both fitting accuracy and operational efficiency, but its leaf-node-first growth strategy is very sensitive to regional differences, and the generalization ability drops most significantly when testing across regions.

To mitigate such degradation in performance, we could potentially add a variance penalty into future models for this type of task and give robust features higher weights. We could also add simple baseline models such as ridge into an ensemble with our complex model. These measures should all contribute to a better bias-variance balance.

Conclusion

The research compared the prediction results of the models in the known region and the unknown region. The results showed that all three models experienced a decline in performance when predicting "unseen regions". Among them, Ridge performed the most stably, while Random Forest and LightGBM had higher training accuracy but weaker cross-region generalization ability and were prone to overfitting local patterns.

This research reveals the spatial heterogeneity problem in housing price prediction. The weights of housing price features vary significantly in different regions, and these differences lead to a decline in the prediction performance of the models in the unseen regions. Moreover, the results indicate that a more complex model does not necessarily mean stronger generalization ability. For example, simple models like Ridge performed more stably in cross-region predictions, indicating that a balance should be achieved between model complexity and robustness in spatial prediction tasks.

Bibliography

Fotheringham, A. S., Crespo, R., & Yao, J. (2015). Exploring, modelling and predicting spatiotemporal variations in house prices. The Annals of Regional Science, 54(2), 417-436.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 30.

Malhotra, R. (2023, January 15). *Melbourne Housing Dataset*. Kaggle. https://www.kaggle.com/datasets/ronikmalhotra/melbourne-housing-dataset

Sellam, Z. A., Distante, C., Taleb-Ahmed, A., & Mazzeo, P. L. (2025). Boosting house price estimations with multi-head gated attention. Expert Systems with Applications, 259, 125276.