Real Estate Price Prediction

Aurora Leone
Politecnico di Torino
Student ID: 334258
Data Science Lab: Process and Methods
Summer Call, A.Y. 2024/2025

***Abstract*—This study presents a comprehensive machine learning approach for real estate rental price prediction using LightGBM with extensive feature engineering. The methodology incorporates TF-IDF vectorization for textual analysis, robust categorical encoding strategies, geographic clustering techniques, and systematic feature selection. Through careful preprocessing and hyperparameter optimization, the final model achieves a Mean Absolute Error of 174.812 on test data, demonstrating strong predictive performance across diverse property types and geographic locations.**

I. Problem Overview

Real estate rental price prediction represents a challenging machine learning problem due to the complex interplay of heterogeneous data sources and dynamic market conditions. The dataset includes diverse information types such as numerical measurements (e.g., square footage, number of rooms), categorical attributes (e.g., property type, amenities), unstructured text descriptions, and geographic coordinates—each influencing prices in different ways.

Several intrinsic factors complicate the development of accurate pricing models. Rental prices exhibit extreme right-skewness (skewness = 9.13), primarily due to a small number of luxury listings that generate long-tailed distributions. Moreover, there is strong spatial heterogeneity across urban, suburban, and coastal regions, where comparable properties may command vastly different prices based solely on their location. The relationship between property size and rental price is also nonlinear and context-dependent, shaped by factors such as architectural efficiency, neighborhood desirability, and local market saturation.

Categorical features—such as pet policies, parking availability, and listing fees—introduce further variability by reflecting service-level differences and shifting tenant preferences. Data quality challenges, including missing values, inconsistent formatting, and potential outliers, add another layer of complexity and require careful preprocessing.

These characteristics demand a robust modeling pipeline capable of handling mixed data types, nonlinear patterns, spatial dependencies, and skewed target distributions, while also ensuring interpretability for practical decision-making in the real estate domain.

II. Proposed Approach

Our methodology addresses the multifaceted challenges of real estate price prediction through a systematic three-stage approach: comprehensive preprocessing to handle data heterogeneity, strategic model selection optimized for tabular data with mixed types, and rigorous hyperparameter tuning focused on generalization performance.

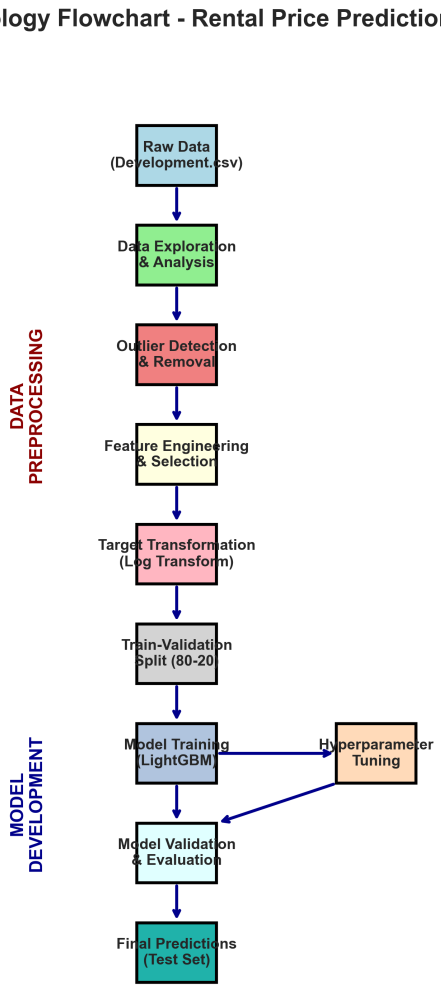

Fig. 1. Complete methodology flowchart showing the systematic pipeline from raw data preprocessing through feature engineering to final model training and evaluation.

A. Preprocessing

The preprocessing pipeline tackles four critical objectives to transform raw real estate data into a machine learning-ready

format while preserving meaningful relationships and reducing statistical anomalies.

1) Outlier Detection Strategy: Traditional outlier detection methods risk removing legitimate high-value properties that represent valid market segments. We implemented a conservative approach using extended percentile ranges to preserve data integrity while removing genuine anomalies.

Our method defines outlier boundaries using the 15th and 85th percentiles with a 2.5× IQR multiplier:

$$[Q_{0.15} - 2.5 \times \text{IQR},\ Q_{0.85} + 2.5 \times \text{IQR}]$$

This conservative threshold removed only 344 entries (0.43%) from 79,589 total observations, preserving 99.57% of the dataset. The approach was applied only when outliers represented less than 5% of total data to prevent excessive information loss.

2) Missing Value Imputation: Rather than applying global imputation strategies that ignore structural relationships, we implemented hierarchical group-based imputation leveraging architectural logic. Missing bathroom counts were imputed using median values within groups defined by bedroom count and square footage ranges, preserving realistic room proportions. Similarly, missing square footage values were filled using grouped medians based on bedroom count and property type, maintaining architectural coherence across similar property categories.

This approach ensures that imputed values reflect realistic property configurations rather than dataset-wide averages that may not represent meaningful real estate relationships.

3) Target Variable Transformation: The extreme right-skewness of rental prices (skewness = 9.13) posed significant challenges for gradient-based learning algorithms. We applied a logarithmic transformation $\log(1 + \text{price})$ to stabilize the distribution, reducing skewness to 0.436 and creating a near-symmetric target variable suitable for regression modeling.

This transformation enables the model to better capture relative price differences across the entire price spectrum rather than being dominated by extreme high-value properties. Predictions are subsequently back-transformed using $\exp(\cdot)-1$ to restore interpretable price values.

4) Feature Engineering: We systematically expanded the original feature set to 669 engineered variables across five key dimensions to capture the multifaceted nature of real estate pricing.

Spatial Intelligence: Geographic location was encoded through two complementary approaches. K-means clustering with k=20 on latitude-longitude coordinates created a `geo_cluster` variable capturing sub-regional pricing patterns [1]. Additionally, we computed `estimated_price_by_city` using leave-one-out cross-validation to estimate median price per square foot by city, preventing target leakage while incorporating local market information [2].

Structural Features: We created architectural ratios including `rooms_ratio` and `rooms_per_sqft` to capture space utilization efficiency. Interaction terms such as `sqft_bedrooms_interaction` model the joint effect of size and room count. Polynomial and logarithmic transformations capture diminishing returns in property size and room count relationships.

Temporal Components: Extracted temporal features including `listing_month`, `day_of_week`, and `is_weekend` to capture potential seasonal effects and market timing patterns in rental listings.

Text Mining: Applied TF-IDF vectorization to property titles (1-3 grams, max 200 features) and descriptions (1-2 grams, max 400 features) to extract semantic information from unstructured text [3]. This captures property characteristics and marketing language that influence rental desirability.

Amenity Processing: Standardized amenity descriptions through rule-based mapping (e.g., "garage," "carport" $\rightarrow$ `parking`) and retained the 26 most frequent amenities as binary features, supplemented by a `total_amenities` count variable.

5) Categorical Encoding and Feature Selection: High-cardinality categorical variables were encoded using smoothed target encoding with regularization parameter $\alpha = 25$ balancing category-specific means with global averages to prevent overfitting. Low-cardinality variables received standard one-hot encoding to preserve all category information.

We implemented a two-stage feature selection process to manage dimensionality. First, features with variance below 0.005 were removed to eliminate uninformative variables. Second, we selected the 111 features with highest Pearson correlation coefficients with the log-transformed target, reducing dimensionality by 83% while retaining predictive diversity across all feature categories.

B. Model Selection

We selected LightGBM [4] as our primary modeling framework after systematic evaluation of multiple approaches, based on its superior performance characteristics for tabular data with mixed types and high dimensionality.

LightGBM's gradient boosting architecture provides several key advantages for real estate applications. Its native support for categorical variables eliminates the need for extensive preprocessing while maintaining computational efficiency. The leaf-wise tree growth strategy optimizes memory usage and training speed, enabling effective handling of our high-dimensional feature space.

The algorithm's built-in regularization mechanisms align with our emphasis on generalization over training set performance, crucial for real estate models that must perform reliably across diverse market conditions [5]. LightGBM's feature importance analysis also provides interpretable insights into pricing factor relationships, supporting business understanding and model validation.

Compared to alternatives, LightGBM demonstrated superior efficiency characteristics. While XGBoost offers similar gradient boosting capabilities [6], LightGBM's faster training times enabled more extensive hyperparameter experimentation within computational constraints. Neural network approaches

would require significantly larger datasets and computational resources while sacrificing the interpretability essential for real estate applications.

C. Hyperparameter Tuning

Our hyperparameter optimization strategy focused on maximizing generalization and robustness across varying property types and market conditions, rather than merely optimizing performance on the training set.

We adopted dual regularization to control model complexity. L1 regularization ($\alpha = 0.2$) penalizes the absolute values of the model parameters, encouraging sparsity and implicitly selecting relevant features. L2 regularization ($\lambda = 1.5$) penalizes the squared magnitudes of the parameters, promoting smoother decision boundaries and reducing sensitivity to noise in the training data.

To ensure stable and gradual learning, we set a conservative learning rate of 0.04 and allowed up to 2500 boosting rounds. This configuration enables the model to capture subtle patterns while avoiding abrupt convergence. An early stopping mechanism was employed to automatically halt training once performance on a validation set ceased to improve, thereby preventing overfitting.

Tree structure parameters were tuned to balance expressiveness and control. We limited the number of leaves to 180 and the maximum tree depth to 16, ensuring the model could capture complex interactions without growing overly deep or fragmented trees.

Finally, we introduced stochasticity to enhance generalization. By setting the bagging fraction to 0.7, each tree was trained on a randomly sampled 70% of the dataset. Likewise, the feature fraction was set to 0.8, limiting each tree's view to 80% of the available features. These randomization strategies increase model robustness by promoting ensemble diversity and reducing over-reliance on specific subsets of the data.

III. Results

The final LightGBM model achieved strong predictive performance with a Mean Absolute Error of 174.812 on unseen test data, corresponding to approximately 11.7% of the mean rental price (1,494.64). This performance demonstrates robust generalization across diverse property types and geographic locations.

Model predictions exhibited excellent calibration with training data distributions. The test set prediction mean (1,492.43) and median (1,352) closely aligned with training statistics (mean: 1,494.64, median: 1,350), indicating minimal prediction drift and strong model stability.

Feature importance analysis revealed that spatial features dominated model decisions, with `estimated_price_by_city`, `latitude`, and `longitude` ranking among the top predictors. Engineered features including `log_square_feet`, `sqft_rooms_interaction`, and TF-IDF text components also contributed significantly, confirming the value of comprehensive feature engineering.

Fig. 2. Final prediction distribution versus training distribution, demonstrating excellent alignment between predicted and actual value distributions.

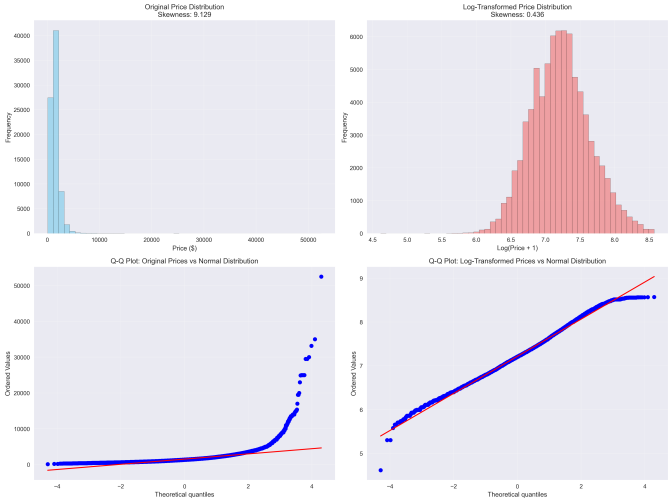

Fig. 3. Target variable distribution before (blue) and after (orange) logarithmic transformation, showing successful skewness reduction from 9.13 to 0.436.

The extensive preprocessing pipeline proved essential for model performance. The logarithmic transformation successfully normalized the highly skewed price distribution, enabling more stable gradient-based learning. Conservative outlier removal preserved data integrity while eliminating genuine anomalies that could mislead model training.

IV. Discussion

Our methodology demonstrates several key strengths in addressing real estate price prediction challenges. The conservative outlier removal strategy preserved 99.6% of data while eliminating genuine anomalies, avoiding the common pitfall of removing legitimate high-value properties. Hierarchical missing value imputation based on architectural logic maintained realistic property relationships rather than applying uninformed global averages.

The comprehensive feature engineering approach successfully integrated diverse information sources. Spatial clustering and city-level price estimates captured geographic market effects, while TF-IDF processing extracted semantic information from property descriptions. The systematic feature selection process reduced dimensionality by 83% while preserving predictive diversity across all engineered categories.

LightGBM proved well-suited for this tabular prediction task, efficiently handling mixed data types while providing

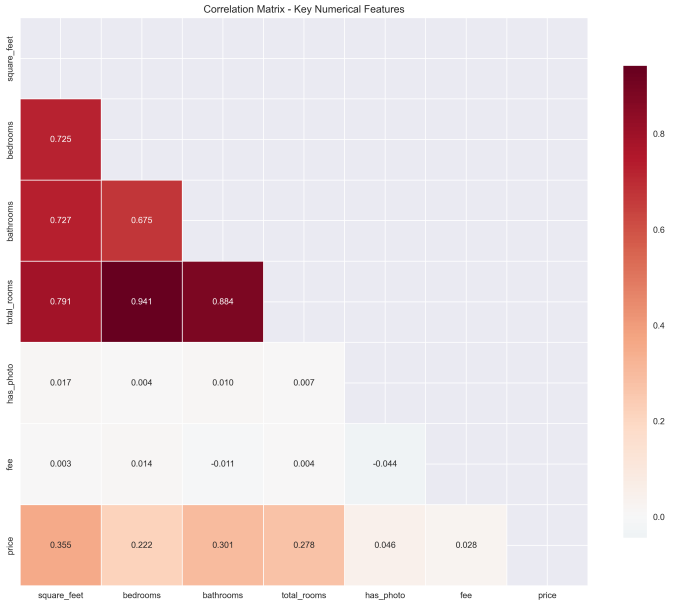

Fig. 4. Correlation matrix of selected features highlighting key relationships, including strong positive correlation between price and square footage measures.

interpretable feature importance insights. The regularization-focused hyperparameter strategy successfully balanced model complexity with generalization capability.

However, several limitations suggest directions for future improvement. First, TF-IDF text processing captures only surface-level linguistic patterns—advanced NLP techniques like BERT embeddings could extract richer semantic information from property descriptions [7]. Second, geographic modeling relies on basic clustering—sophisticated spatial indexing incorporating neighborhood characteristics and external geographic data could improve location-based predictions [8].

Third, the current approach assumes homogeneous pricing mechanisms across property types—segment-specific models for different real estate categories could capture distinct market dynamics. Fourth, temporal modeling is limited to basic seasonal features—time series approaches could capture market evolution and cyclical patterns [9].

Future research directions include ensemble methods combining diverse algorithms [10], specialized neural architectures for tabular data [11], dynamic modeling incorporating market temporal evolution, and advanced spatial analysis using external geographic databases. Production deployment would require robust preprocessing pipeline replication, instance-level model interpretability tools, and automated retraining mechanisms to adapt to evolving market conditions.

The modular pipeline design facilitates incremental improvements and adaptation to new markets or property types, establishing a solid foundation for continued development in this commercially significant domain.

REFERENCES

[1] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[2] M. Stone, "Cross-validatory choice and assessment of statistical predictions," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 36, no. 2, pp. 111–133, 1974.

[3] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing & Management*, vol. 24, no. 5, pp. 513–523, 1988.

[4] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, 2017, pp. 3146–3154.

[5] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of Statistics*, pp. 1189–1232, 2001.

[6] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, pp. 4171–4186.

[8] A. S. Fotheringham, C. Brunsdon, and M. Charlton, *Geographically weighted regression: the analysis of spatially varying relationships*. John Wiley & Sons, 2003.

[9] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[10] T. G. Dietterich, "Ensemble methods in machine learning," *International Workshop on Multiple Classifier Systems*, pp. 1–15, 2000.

[11] S. Ö. Arik and T. Pfister, "Tabnet: Attentive interpretable tabular learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 8, 2021, pp. 6679–6687.