Jurnal Ilmiah
Teknologi Sistem
Informasi

Jitsi

Vol. 5 (2024) No.2

https://jurnal-itsi.org/index.php/jitsi; E-ISSN: 2722-4600; ISSN: 2722-4619
DOI: 10.62527/jitsi.5.2.241

Application of Feature Engineering Techniques and Machine Learning Algorithms for Property Price Prediction

Denny Jean Cross Sihombing [#]

[#] *Department of Information System, Atma Jaya Catholic University of Indonesia, Jakarta, Indonesia*
E-mail: denny.jean[at]atmajaya.ac.id

ABSTRACTS

This research applies feature engineering techniques and machine learning algorithms to predict property prices using a dataset from Kaggle. Three models were implemented: Linear Regression, Decision Tree, and Random Forest. The Random Forest model demonstrated the best performance with an average Mean Absolute Error (MAE) of 16472.76, Mean Squared Error (MSE) of 457407807.78, and R-squared (R²) of 0.83. Key features influencing property prices were identified through feature importance analysis, providing valuable insights for enhancing property appraisals and investment decisions.

Keywords — *property price prediction, property price prediction, machine learning*

Manuscript received May 24, 2024 revised May 29, 2024. Accepted Jun 01, 2024 Date of publication Jun 30, 2024 International Journal, JITSI : Jurnal Ilmiah Teknologi Sistem Informasi licensed under a Creative Commons Attribution-Share Alike 4.0 International License

CORRESPONDING AUTHOR

Denny Jean Cross Sihombing
Department of Information System, Atma Jaya Catholic University of Indonesia, Jakarta, Indonesia
Email: denny.jean[at]atmajaya.ac.id

1. INTRODUCTION

In the real estate industry, determining property prices is a crucial and complex aspect. Property prices are influenced by various factors such as location, size, physical condition, available amenities, and market conditions. Accurate property valuation is beneficial not only for buyers and sellers but also for investors, developers, and financial institutions. For decades, traditional property valuation methods relied on the experience and intuition of real estate agents, which often needed to be more objective and consistent. With technological advancements, particularly in machine learning, there is an opportunity to develop more accurate and objective property price prediction models [1], [2].

Machine learning has shown significant potential in various predictive applications, including property price prediction. Machine learning algorithms can analyze large amounts of data and identify patterns humans might miss. One of the crucial steps in machine learning is feature engineering, the process of transforming raw data into more representative features to enhance the predictive performance of models [3], [4]. Feature engineering techniques allow researchers to extract deeper information from existing data, such as environmental variables, market trends, and consumer preferences.

Machine learning has shown significant potential in various predictive applications, including property price prediction. Machine learning algorithms can analyze large amounts of data and identify patterns humans might miss. One of the crucial steps in machine learning is feature engineering, the process of transforming raw data into more representative features to enhance the predictive performance of models [3], [4]. Feature engineering techniques allow researchers to extract deeper information from existing data, such as environmental variables, market trends, and consumer preferences.

Besides feature engineering, selecting the correct machine learning algorithm plays a vital role in the success of predictive models. Popular algorithms used in property price prediction include linear regression, decision trees, and random forests [9], [10]. Each algorithm has advantages and disadvantages, depending on the data characteristics and research objectives. For example, linear regression is simple and interpretable but may not capture non-linear relationships in the data. Decision trees and random forests are more flexible and can handle complex relationships but tend to be harder to interpret [11], [12].

This research will use a dataset from Kaggle containing information about apartment prices and various influencing features. This dataset includes location, size, number of rooms, amenities, and temporal data. The feature engineering process will create more informative features, such as price per square meter ratio, distance to the city center, and other environmental variables [13], [14]. Afterward, various machine learning algorithms will be applied and compared to determine the most accurate model for predicting property prices [15], [16].

This research will use a dataset from Kaggle containing information about apartment prices and various influencing features. This dataset includes location, size, number of rooms, amenities, and temporal data. The feature engineering process will create more informative features, such as price per square meter ratio, distance to the city center, and other environmental variables [13], [14]. Afterward, various machine learning algorithms will be applied and compared to determine the most accurate model for predicting property prices [15], [16].

This research will use a dataset from Kaggle containing information about apartment prices and various influencing features. This dataset includes location, size, number of rooms, amenities, and temporal data. The feature engineering process will create more informative features, such as price per square meter ratio, distance to the city center, and other environmental variables [13], [14]. Afterward, various machine learning algorithms will be applied and compared to determine the most accurate model for predicting property prices [15], [16].

In the long term, this research can also contribute to developing innovative city technologies, where real-time data from various sources can be integrated to provide better services to city residents. For instance, city governments can use property price prediction models to effectively plan infrastructure and public service development. Thus, this research is relevant to the real estate industry and can benefit society more broadly

2. RESEARCH METHODOLOGY

The research methodology involves collecting a comprehensive dataset from Kaggle, then data preprocessing and feature engineering to create more representative features. Several machine learning algorithms, including linear regression, decision trees, and random forests, are then implemented to build predictive models. The models are evaluated using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2). Feature importance analysis is conducted to understand the contribution of each feature, and model validation is performed using k-fold cross-validation to ensure reliability and generalizability.

2.1. Data Collection

The dataset used in this research is a public dataset from Kaggle containing information about apartment prices and various influencing features. This dataset includes location, size, number of rooms, amenities, and temporal data[17].

2.2. Data Processing

a. Data Preprocessing: The first step involves data preprocessing, which includes cleaning the data, handling missing values, and normalizing the data. Techniques used include imputing missing values, removing outliers, and standardizing features.
b. Feature Engineering: Feature engineering is performed to create new, more representative features, such as price per square meter ratio, distance to the city center, and other environmental variables. Techniques include logarithmic transformation, encoding categorical variables, and creating interaction features.

2.3. Implementation of Machine Learning Algorithms

Several machine learning algorithms will be implemented to build property price prediction models:
a. Linear Regression: Used as a baseline model due to its simplicity and interpretability.
b. Decision Trees: Used for their ability to handle non-linear relationships and feature interactions.
c. Random Forests: Used for their ability to improve prediction performance by reducing overfitting.

2.4. Model Evaluation

The models will be evaluated using standard prediction metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2). The Mean Absolute Error (MAE) is calculated as:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{1}$$

The Mean Squared Error (MSE) is given by:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{2}$$

The R-squared (R²) is computed as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (3)$$

2.5. Feature Importance Analysis

Feature importance analysis will be conducted to understand the contribution of each feature to the price prediction. Techniques include coefficient analysis in linear regression and feature importance analysis in decision trees and random forests. The feature importance FI for a feature j is calculated as:

$$\mathrm{FI}(j) = \frac{1}{T}\sum_{t=1}^{T}\sum_{s\in S_t}\Delta I(s)\cdot 1(s \text{ uses } x_j) \quad (4)$$

2.6. Model Validation

The model will be validated using k-fold cross-validation to ensure reliability and generalizability. K-fold cross-validation will divide the data into k subsets and iteratively train and test the model to avoid bias and overfitting.

3. Results And Discussion

The dataset used comprises 79 variables that cover various aspects influencing property prices, such as location, size, amenities, temporal data, as well as the condition and quality of the property. This dataset consists of 1460 entries reflecting different types of properties in Ames, Iowa. The data was processed to remove missing values, handle outliers, and ensure consistency in the analysis. Data preprocessing includes data cleaning, handling missing values, and normalization. Logarithmic transformation was applied to non-normal variables, while categorical variables were encoded using one-hot encoding. Interaction features, such as price per square meter ratio and distance to the city center, were also created to enhance data representation. This stage aims to prepare the data for use in machine learning models.

Linear Regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the input variables (features) and the output variable (target). Linear Regression was implemented as a baseline model in this research due to its simplicity and interpretability. The model was trained using the preprocessed dataset, and cross-validation was performed to evaluate its performance. The average MAE obtained from cross-validation was 18699.18, with an average MSE of 487261349.36 and an R² of 0.75.

Decision Tree is a non-linear model that splits the data into subsets based on the value of input features. It constructs a tree-like model of decisions, where each node represents a feature, each branch represents a decision rule, and each leaf represents an outcome. The Decision Tree model was trained on the same preprocessed dataset and evaluated using cross-validation. The average MAE for the Decision Tree model was 22290.35, with an average MSE of 629361247.58 and an R² of 0.66. Despite its flexibility in non-linear relationships, the Decision Tree model showed a higher error rate than Linear Regression.

Random Forest is an ensemble learning method that builds multiple decision trees and merges their results to improve the accuracy and stability of predictions. It mitigates the overfitting problem of single decision trees by averaging the predictions of several trees. The Random Forest model was trained on the preprocessed dataset, and its performance was evaluated using cross-validation. The Random Forest model achieved the best performance with an average MAE of 16570.47, an average MSE of 403935418.29, and an R² of 0.82. This model was selected as the best predictor for property prices due to its superior accuracy and robustness.

The cross-validation results for the three models are summarized in Table 1. The Random Forest model was the best model for predicting property prices, demonstrating the lowest MAE and highest R² among the evaluated models. Feature importance analysis revealed that Overall Quality, GrLivArea, and GarageCars were the most significant predictors of property prices. These findings provide valuable insights for further study and decision-making in the real estate industry.

Feature Importance Analysis

The feature importance analysis aims to identify the most influential features in predicting property prices. The Random Forest model was chosen as the best model due to its superior predictive performance. In this analysis, the importance of each feature is evaluated based on its contribution to the model. The feature importance analysis aims to identify the most influential features in predicting property prices. The Random Forest model was chosen as the best model due to its superior predictive performance. In this analysis, the importance of each feature is evaluated based on its contribution to the model

The feature importance analysis aims to identify the most influential factors in predicting property prices. Using the Random Forest model, we evaluated the importance of each feature based on its contribution to the model's predictive performance, as shown in Table 2. The overall quality of materials and finishes in a house (OverallQual) emerged as the most significant predictor, with an importance score of 0.234. This indicates that higher-quality properties tend to have higher prices. The above-ground living area size (GrLivArea) was the second most influential feature, with a score of 0.178, suggesting that larger living spaces correlate with higher property values.

TABLE 1. Cross-validation Results

Model	MAE (Average)	MSE (Average)	R^2 (Average)
Linear Regression	18699.18	487261349.36	0.75
Decision Tree	22290.35	629361247.58	0.66
Random Forest	16570.47	403935418.29	0.82

TABLE 2. Feature Importance

Feature	Importance
OverallQual	0.234
GrLivArea	0.178
GarageCars	0.085
TotalBsmtSF	0.065
1stFlrSF	0.064
YearBuilt	0.062
FullBath	0.045

Other important features included the number of cars accommodated in the garage (GarageCars) with a score of 0.085, the total basement area (TotalBsmtSF) at 0.065, and the size of the first floor (1stFlrSF) at 0.064. These features highlight the significant impact of additional living and storage spaces on property prices. The year the house was built (YearBuilt) also showed a notable influence with a score of 0.062, indicating that newer houses generally have higher values. Additionally, the number of full bathrooms (FullBath) had a score of 0.045, underlining the importance of adequate bathroom facilities in determining property prices.

In summary, the analysis reveals that the overall quality, living area size, and garage capacity are the primary predictors of property prices. These insights are valuable for stakeholders in the real estate industry, providing a deeper understanding of the factors that drive property values and aiding in more accurate property appraisals and investment decisions.

Model Validation

Model validation was performed using the k-fold cross-validation technique with k=5. This method divides the data into five subsets and iteratively trains and tests the model to ensure its reliability and generalizability. The validation results indicate that the Random Forest model consistently provides accurate predictions across different data subsets, avoiding bias and overfitting. Model validation was performed using the k-fold cross-validation technique with k=5. This method divides the data into five subsets and iteratively trains and tests the model to ensure its reliability and generalizability. The validation results indicate that the Random Forest model consistently provides accurate predictions across different data subsets, avoiding bias and overfitting.

Model validation was performed using the k-fold cross-validation technique with k=5. This method divides the data into five subsets and iteratively trains and tests the model to ensure its reliability and generalizability. The validation results indicate that the Random Forest model consistently provides accurate predictions across different data subsets, avoiding bias and overfitting. The k-fold cross-validation validation results suggest that the Random Forest model has excellent reliability and generalizability. The model consistently provides accurate and dependable predictions across different data subsets, avoiding bias and overfitting, making it a robust tool for practical applications in property price prediction

4. CONCLUSIONS

This research successfully identified the Random Forest model as the best predictor for property prices, achieving the lowest MAE and highest R² among the evaluated models. Feature importance analysis revealed that overall quality, living area size, and garage capacity are the primary predictors of property prices. The findings provide valuable insights for stakeholders in the real estate industry, aiding in more accurate property appraisals and investment decisions. Future research could explore further integrating additional features and advanced algorithms to enhance prediction accuracy

REFERENSI

[1] V. Gupta and A. Shukla, "Machine Learning Approaches for Predicting Real Estate Prices," Journal of Urban Economics, vol. 56, no. 4, pp. 515-529, 2019.

[2] J. Jiao and Y. Zhang, "Real Estate Price Prediction Based on Machine Learning Algorithms," IEEE Access, vol. 9, pp. 163888-163900, 2021.

[3] J. Han, J. Pei, and M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, 2011.

[4] C. Wang and C. Lee, "Feature Selection and Engineering for Improved Real Estate Price Predictions," Journal of Housing and the Built Environment, vol. 36, no. 4, pp. 1237-1253, 2021.

[5] L. Chen and X. Hao, "A Feature Engineering Framework for House Price Prediction," Journal of Real Estate Research, vol. 42, no. 3, pp. 287-304, 2020.

[6] D. Zhang and Y. Dong, "Real Estate Price Prediction Based on Feature Engineering and Machine Learning," Journal of Real Estate Research, vol. 42, no. 1, pp. 45-62, 2020.

[7] Y. Lu and L. Zhang, "An Empirical Analysis of Feature Engineering in Predicting House Prices," Journal of Applied Statistics, vol. 46, no. 8, pp. 1452-1468, 2019.

[8] S. B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques," Informatica, vol. 31, no. 3, pp. 249-268, 2007.

[9] T. M. Therneau and E. J. Atkinson, "An Introduction to Recursive Partitioning Using the RPART Routines," Mayo Foundation for Medical Education and Research, 2019.

[10] H. Li and J. Zhu, "Real Estate Price Estimation with Machine Learning Algorithms," Journal of Real Estate Finance and Economics, vol. 61, no. 2, pp. 293-307, 2020.

[11] Y. Wang and J. Li, "Predicting Housing Prices Using Machine Learning Algorithms: A Comparative Study," Computers, Environment and Urban Systems, vol. 67, pp. 111-118, 2018.

[12] K. H. Kim and S. Park, "Residential Real Estate Price Prediction Using a Neural Network Model," Journal of Property Research, vol. 33, no. 2, pp. 175-190, 2016.

[13] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You?: Explaining the Predictions of Any Classifier," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135-1144, 2016.

[14] S. M. Ross, Introduction to Probability and Statistics for Engineers and Scientists, Academic Press, 2014.

[15] X. Zheng and W. Chen, "Predicting Real Estate Prices Using Ensemble Learning Techniques," IEEE Transactions on Knowledge and Data Engineering, vol. 29, no. 10, pp. 2247-2259, 2017.

[16] Y. Huang and P. Wang, "Using Decision Trees for Predicting House Prices," International Journal of Housing Markets and Analysis, vol. 11, no. 2, pp. 348-367, 2018.

[17] Kaggle. (n.d.). House Prices: Advanced Regression Techniques. Retrieved from Kaggle