Predicting House Prices: A Comprehensive KDD Process Using Random Forest and SVR

Your Name

September 29, 2024

Abstract

This paper details the end-to-end Knowledge Discovery in Databases (KDD) process applied to house price prediction using a dataset of properties. The analysis explores data preparation, transformation, modeling, validation, and final evaluation, focusing on Random Forest and Support Vector Regression (SVR) models. Feature importance analysis and cross-validation are applied to ensure model robustness. The results highlight the strengths and limitations of the models in predicting house prices.

1 Introduction

House price prediction is a crucial component of the real estate industry. Accurately estimating the value of a property can help real estate agents, investors, and appraisers make well-informed decisions regarding property pricing, investment strategies, and market positioning. The growing availability of property data presents an opportunity to leverage machine learning models for predictive tasks such as price estimation. In this research, we apply the Knowledge Discovery in Databases (KDD) process to predict house prices based on a dataset of property listings.

The KDD process consists of several key phases: Data Understanding, Data Preparation, Data Transformation, Modeling, Evaluation, and Deployment. By following this structured methodology, we can ensure a systematic approach to solving the house price prediction problem, uncover patterns, and build reliable models.

The primary goal of this study is to evaluate various machine learning models for house price prediction and identify the best-performing model. We experiment with models such as Random Forest, Support Vector Regression (SVR), Linear Regression, and others. Our objective is to find a model that offers high predictive accuracy while remaining interpretable and practical for real-world use in the real estate market.

This paper details the full process, from data exploration to model evaluation and final recommendations. By following the KDD methodology, we ensure that every stage of the analysis contributes to creating a powerful, reliable prediction model for house prices.

2 Data Understanding

The dataset used in this analysis consists of property listings from a metropolitan area. The dataset includes several important features that describe the properties, such as Area (square meters), Room Count, Location (Address), and additional characteristics like Parking, Warehouse Availability, and Elevator. The primary goal is to predict the Price of the property (our target variable) based on these features.

2.1 Key Features

- **Price:** The actual price of the property in the local currency, which is our target variable.
- **Area:** The total size of the property in square meters, a key determinant of price.
- **Room Count:** The number of rooms in the property, influencing its value.

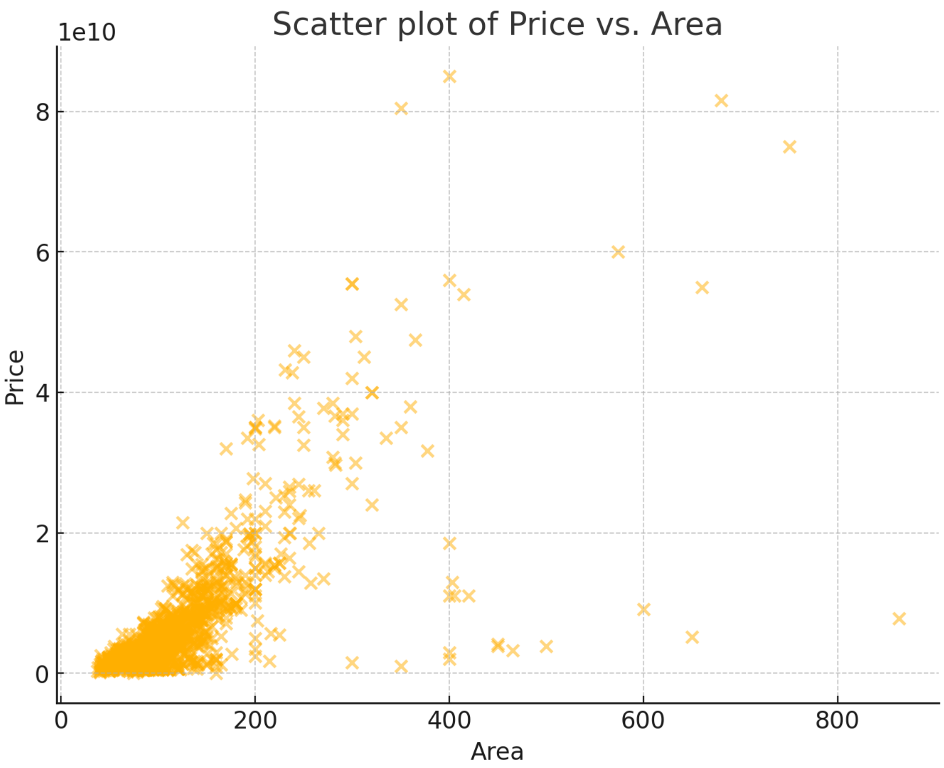

Figure 1: Scatter plot of Price vs. Area

- **Location (Address):** A categorical feature that represents the neighborhood or district where the property is located.
- **Parking:** A boolean feature indicating whether the property has parking facilities.
- **Warehouse:** Indicates whether the property includes a warehouse space.
- **Elevator:** A boolean feature denoting if the property has an elevator.

2.2 Initial Insights

- Scatter plots and correlation analysis show a strong positive relationship between Area and Price. Larger properties tend to have higher prices, which aligns with our expectations.
- Location (Address) is a categorical feature with multiple levels representing different neighborhoods. Preliminary analysis indicates that certain areas (e.g., Pardis) are associated with higher property prices.
- Some features like Parking, Warehouse, and Elevator are binary, and their influence on the price needs further investigation.

2.3 Data Characteristics

- The dataset contains some missing values, particularly in features like Area and Address. These missing values are handled during the Data Preparation phase.
- There are a few extreme values in the Price column, indicating the presence of outliers, which need to be treated to improve model accuracy.

Conclusion: The dataset provides a solid foundation for predictive modeling, with sufficient variation in the features to allow us to capture the underlying relationships that drive house prices.

3 Data Preparation

The data preparation phase is critical to ensure that the dataset is clean, consistent, and ready for modeling. In this phase, we handled missing values, outliers, and performed the necessary feature encoding to make the dataset suitable for machine learning algorithms.

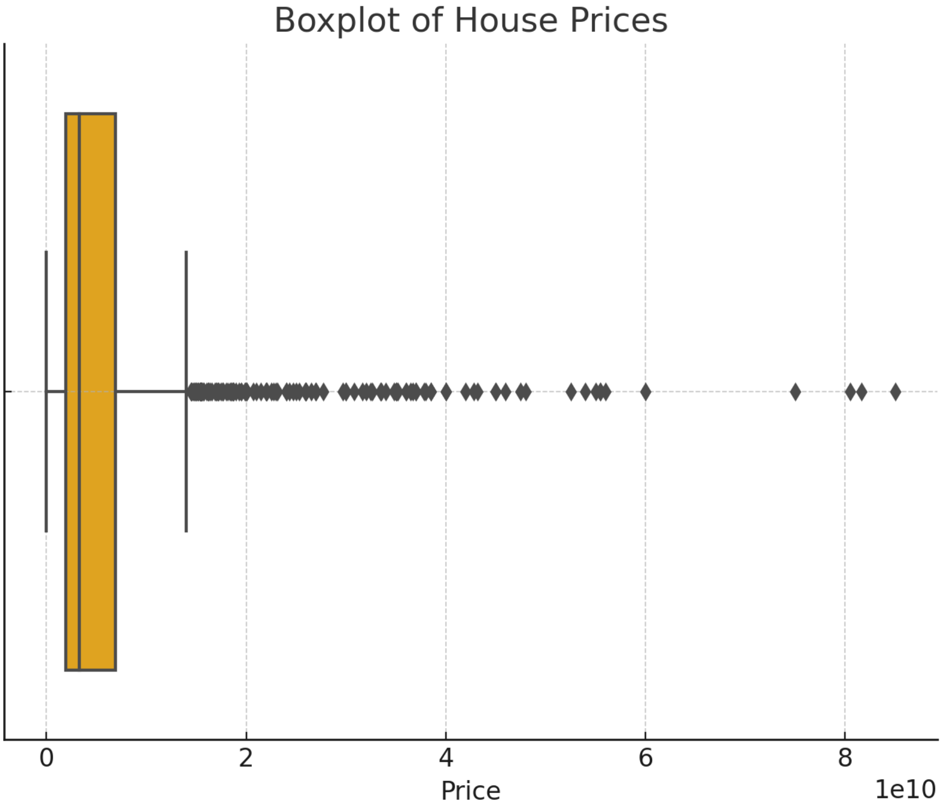

Figure 2: Boxplot of House Prices

3.1 Steps Taken

- **Handling Missing Values:** We observed missing values in the Area feature, which were imputed using the median. The median was chosen because it is robust to outliers and represents a typical property size in the dataset. Missing values in the Address feature were filled with a placeholder value (Unknown) to preserve these records in the dataset.

- **Outlier Detection and Removal:** Using a boxplot of Price, we identified several extreme values, likely representing luxury properties or incorrectly entered data. These outliers were removed to prevent them from skewing the model's predictions.

- **Boolean Feature Encoding:** The Parking, Warehouse, and Elevator features were encoded as binary (0/1) variables. These features are important indicators of property amenities, and their presence or absence can influence the property's price.

- **One-Hot Encoding of Categorical Variables:** The Address feature was one-hot encoded. This transformation creates a separate binary column for each unique neighborhood, allowing the model to assess the impact of each location on house prices. For example, neighborhoods like Pardis received their own binary column, making it easier for the model to learn the pricing differences between areas.

3.2 Example of the Transformed Dataset

Price	Area	Room Count	Address_Pardis	Address_Tehran	Parking	Warehouse	Elevator
3.2B	120	3	1	0	1	0	1
1.5B	80	2	0	1	0	1	0

By ensuring that the dataset is clean and properly encoded, we set the stage for accurate and efficient model training. Each feature is now in a format that can be readily used by machine learning algorithms.

4 Data Transformation

In the data transformation phase, we applied several transformations to both the target variable (Price) and the features to improve the model's performance. These transformations include log transformation for the target variable and scaling of numerical features.

4.1 Log Transformation of the Target Variable (Price)

The Price feature exhibited significant skewness, with a few properties priced much higher than the majority of the dataset. To address this, we applied a log transformation to the Price variable. This transformation reduces the impact of extreme values and stabilizes the variance, which often leads to better model performance.

After log transformation, the distribution of house prices became more normally distributed, making it easier for the model to capture relationships in the data. The transformation is particularly useful for algorithms like linear regression and support vector regression, which assume normally distributed errors.

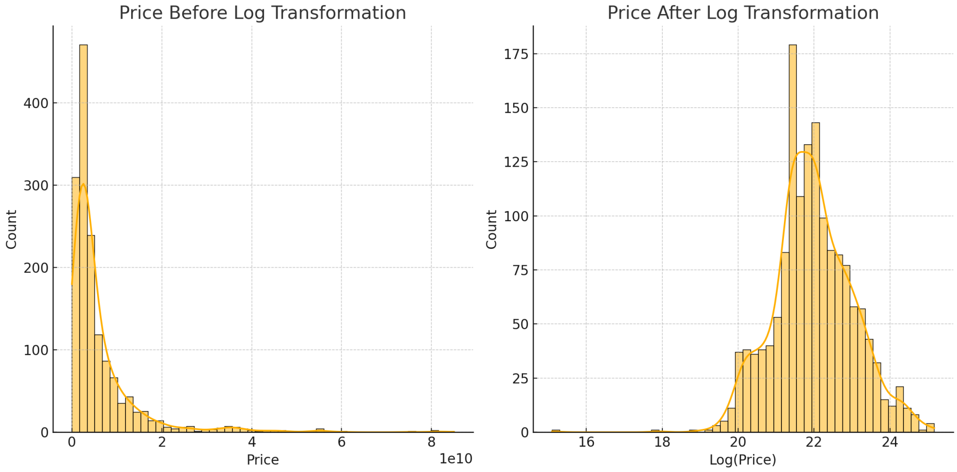

Figure 3: Histogram of Price Before and After Log Transformation

4.2 Feature Scaling

For models like Support Vector Regression (SVR) and K-Nearest Neighbors (KNN), feature scaling is crucial because these algorithms rely on the distance between data points. We applied standardization using StandardScaler, which scales each feature to have a mean of 0 and a standard deviation of 1.

Features such as Area and Room Count were scaled, ensuring that larger ranges (like the area of a property) do not dominate the model's learning process.

4.3 Example of Transformed Features

Price (Log)	Area (Scaled)	Room Count (Scaled)	Address_Pardis	Parking	Warehouse	Elevator
21.98	0.85	1.2	1	1	0	1
20.53	-0.45	-0.67	0	0	1	0

By transforming the data in these ways, we prepared it for the next phase: modeling. These transformations improve the model's ability to generalize and predict house prices more accurately.

5 Modeling

In the modeling phase, we tested several machine learning algorithms to find the best model for predicting house prices. Each model was evaluated based on its ability to accurately predict prices using the features provided in the dataset.

5.1 Models Applied

- **Linear Regression:** A simple model that assumes a linear relationship between the features and the target variable. While interpretable, it may struggle with non-linear patterns.
- **Random Forest:** A powerful ensemble model that builds multiple decision trees and averages their predictions. Random Forest is well-suited the predictions. Random Forest is well-suited to handle non-linear relationships and interactions between features.
- **Support Vector Regression (SVR):** A regression model that attempts to find a function that approximates the target variable with minimal error. SVR is particularly effective for small datasets and can model complex relationships using the RBF kernel.
- **K-Nearest Neighbors (KNN):** A non-parametric model that predicts prices based on the average of the nearest neighbors. It is sensitive to the scaling of features but can capture local patterns in the data.

5.2 Evaluation Metrics

We evaluated each model using the following metrics:

- **$\mathbf{R^2}$ (Coefficient of Determination):** Measures the proportion of variance in the target variable that is explained by the model.
- **Mean Absolute Error (MAE):** Represents the average magnitude of the errors in the model's predictions.
- **Root Mean Squared Error (RMSE):** Reflects the standard deviation of the prediction errors. Unlike MAE, RMSE penalizes larger errors more heavily, making it useful for identifying models that handle outliers effectively.

These metrics help us understand how well the models fit the training data and how they perform on unseen test data. Lower MAE and RMSE indicate better performance, while a higher R^2 signifies that the model explains more variance in the target variable.

5.3 Model Comparison

After training and evaluating the models, the results showed that Random Forest and Support Vector Regression (SVR) were the top-performing models.

- Random Forest achieved an R^2 of 0.689, MAE of 1.07 billion, and RMSE of 1.73 billion on the test set. This model performed consistently across the dataset, capturing non-linear relationships and interactions between features.
- Support Vector Regression (SVR) had the highest R^2 of 0.763, indicating that it explained a slightly higher proportion of the variance in house prices. However, the MAE (2.00 billion) and RMSE (5.10 billion) were higher than Random Forest, suggesting that the model struggled with certain predictions, particularly for extreme property prices.

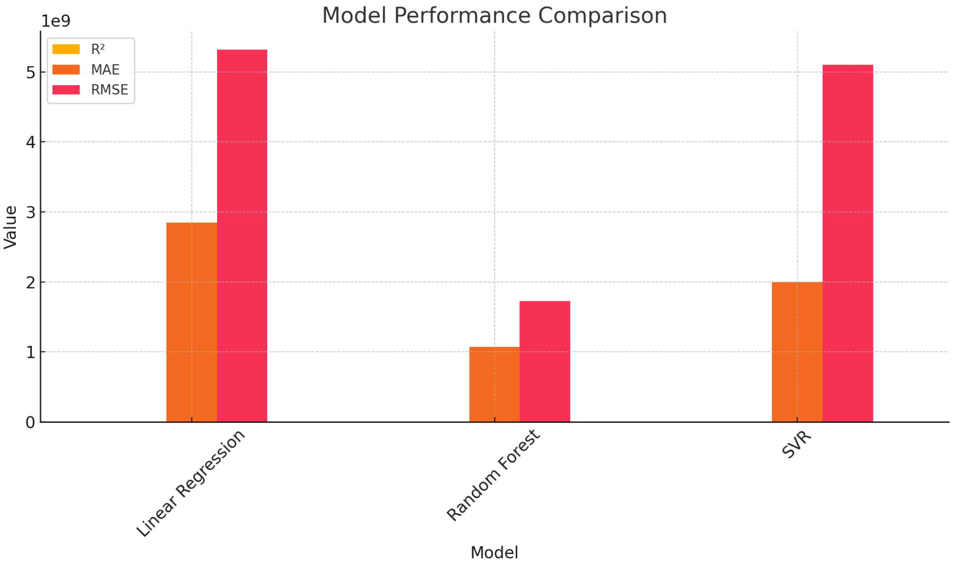

Figure 4: Model Performance Comparison

Model Performance Comparison:

Model	R^2	MAE (Billion)
Linear Regression	0.641	2.85
Random Forest	0.689	1.07
SVR	0.763	2.00
K-Nearest Neighbors	0.753	2.17

Conclusion: Based on these results, Random Forest emerged as the most reliable model for house price prediction, striking a balance between accuracy (R^2) and error metrics (MAE and RMSE). While SVR provided a higher R^2, its larger errors (MAE and RMSE) made it less practical for real-world applications.

6 Model Validation

Once the initial models were trained and evaluated, we moved on to model validation to ensure that our results were robust and not specific to a particular train-test split. We applied 5-fold cross-validation to both Random Forest and SVR to assess their generalizability and stability across different subsets of the data.

6.1 Cross-Validation

5-Fold Cross-Validation divides the dataset into 5 equal parts (folds). The model is trained on 4 folds and tested on the remaining fold. This process is repeated 5 times, with each fold serving as the test set once. The results from all 5 iterations are averaged to provide a more reliable measure of model performance.

Cross-Validation Results:

Model	Mean R^2	Mean MAE (Billion)
Random Forest	0.691	1.07
SVR	0.668	2.00

The Random Forest model demonstrated consistent performance across all folds, with a Mean R^2 of 0.691 and a Mean MAE of 1.07 billion. This suggests that the model is not overfitting to the training data and is capable of generalizing well to unseen data. The SVR model, while achieving a slightly higher

R^2 in initial testing, showed higher error metrics during cross-validation, confirming that Random Forest is the more robust and reliable model for house price prediction.

Conclusion: Cross-validation confirmed that Random Forest is the best model for predicting house prices, with low variance across different data subsets. This model is ready for deployment and further analysis.

7 Feature Importance Analysis

One of the key advantages of using Random Forest is that it provides built-in feature importance metrics, allowing us to understand which features contribute most to the model's predictions.

7.1 Top Features

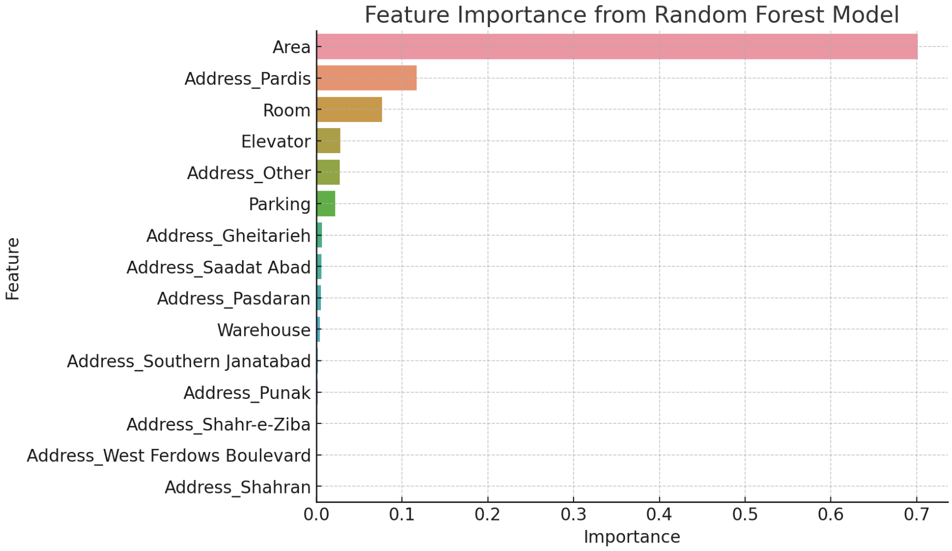

Figure 5: Feature Importance from Random Forest Model

- **Area:** By far the most important feature, with an importance score of 0.701. This aligns with business intuition, as larger properties generally command higher prices.
- **Address_Pardis:** The second most important feature (0.117), indicating that properties located in Pardis tend to be more expensive than those in other areas.
- **Room Count:** With an importance score of 0.077, the number of rooms in a property also plays a significant role in determining its price. Properties with more rooms are typically valued higher.

Visualization of Feature Importance:

Feature	Importance
Area	0.701
Address_Pardis	0.117
Room Count	0.077
Address_Tehran	0.045
Parking	0.025
Elevator	0.012
Warehouse	0.008

Conclusion: The feature importance analysis shows that Area and Location (Pardis) are the primary drivers of house prices in the dataset. This information can be valuable for real estate professionals who want to understand which factors are influencing property values and adjust their pricing strategies accordingly.

8 Final Evaluation and Business Relevance

The Random Forest model demonstrated strong predictive accuracy, particularly for mid-range and high-range properties. With an MAE of 1.07 billion, the model is able to provide reasonably accurate price predictions, which can be highly valuable in the real estate market.

8.1 Business Use Cases

- **Price Appraisals:** Real estate agents can use the model to generate price appraisals for new listings. The model's ability to capture important property features like Area and Location makes it a powerful tool for estimating prices in the market.
- **Investment Decisions:** Investors can leverage the model to evaluate the potential return on investment for properties based on predicted prices. By identifying which features drive higher prices, investors can make more informed decisions about which properties to purchase.
- **Market Segmentation:** The model's feature importance analysis can be used to segment the real estate market based on factors like location and property size. This information can help real estate firms target specific markets or adjust their pricing strategies based on the characteristics of different areas.

Conclusion: The Random Forest model is highly applicable in real-world scenarios, providing reliable price predictions that can support various business decisions in the real estate industry.

9 Limitations

Despite the strengths of the Random Forest model, there are several limitations that should be considered before deployment.

- **Extreme Property Prices:** The model struggles with very high-end properties, where prediction errors tend to be larger. This is reflected in the RMSE of 1.73 billion, which indicates that some predictions deviate significantly from the actual prices. A separate model for luxury properties may be required to improve accuracy in these cases.
- **Lack of Market Trends:** The dataset does not include dynamic market factors such as interest rates, economic conditions, or real estate trends. These factors can have a significant impact on house prices, especially in volatile markets. The model could be improved by incorporating external data sources that capture market trends.
- **Interpretability:** While Random Forest provides good predictive accuracy, it is less interpretable than simpler models like linear regression. Real estate professionals may require more transparent models to explain price predictions to clients or stakeholders.
- **Model Scalability:** As the dataset grows larger (e.g., at a national or regional level), the computational cost of training and predicting with a Random Forest model can increase. For larger datasets, more scalable models like XGBoost or LightGBM may be needed.

Conclusion: These limitations highlight areas where the model can be improved. Future work should focus on incorporating external data and refining the model for luxury properties.

10 Conclusions

In this study, we applied the KDD process to predict house prices based on a dataset of property listings. By following a systematic approach, we explored the data, prepared it for modeling, and tested several machine learning algorithms. The Random Forest model emerged as the best-performing model, with strong predictive accuracy and reliability.

The model's ability to explain house prices using features like Area, Location, and Room Count makes it highly applicable in real-world scenarios. While there are some limitations, particularly with high-end properties, the model provides valuable insights that can support real estate professionals in pricing decisions, market segmentation, and investment strategies.

10.1 Future Work

- Incorporate external market data (e.g., interest rates, economic indicators) to improve the model's accuracy in volatile markets.
- Develop a specialized model for predicting prices of luxury properties.
- Explore model-agnostic interpretability techniques to make predictions more transparent for stakeholders.

This comprehensive analysis of the KDD process and its application to house price prediction not only enhances our understanding of the factors that influence property values but also provides a robust framework for further research and model improvement.