Springboard---DSC
Capstone Two
Predicting House Price

By Grace Riccadonna
May 25, 2024

1 Introduction

The problem we aim to address in this project is to predict housing prices by analyzing various housing features such as property size, location, condition, and more. House prices fluctuate frequently and can sometimes form significant bubbles, misleading both investors and buyers. This project is pertinent for real estate investors and buyers as it provides insights into the true value of properties and aids in making better decisions.

By building a predictive model, we can answer key questions such as: What are the key factors influencing housing prices? How accurately can we predict the price of a house based on its features? How can investors and buyers leverage this model to make informed decisions?

Our intended stakeholders include real estate investors, homebuyers, and financial institutions. These stakeholders will benefit from the insights provided by our model, enabling them to make more informed decisions regarding property investments and purchases.

Data science results will be presented through predictive models, visualizations, and actionable recommendations. Detailed implementation steps can be found in the notebooks developed throughout the project. For more information and access to all project deliverables, please visit the root folder of our GitHub repository here.

2 Data Acquisition and Wrangling

To investigate the problem of predicting housing prices, we utilized a dataset containing historical information on housing prices and various features. The dataset was sourced from a reputable real estate agency known for providing comprehensive data. The key features of the dataset include:

- **Housing Price Data:** Information about housing prices, such as sale prices, property types, and locations.
- **Property Features:** Data on property size, number of rooms, condition of the house, age of the property, and other relevant features.
- **Geospatial Information:** Location data, such as neighborhoods and regions, to explore geographical variations.

2.1 Data Acquisition

The dataset contains a total of 79 features, encompassing both numerical and categorical data. We followed these steps to acquire and prepare the data for further analysis:

- **Data Collection:** We downloaded the dataset from the Kaggle competition website. The data was provided in CSV format, which is convenient for processing and analysis.
- **Data Loading:** We loaded the dataset into a pandas DataFrame using Python. This allowed us to easily manipulate and analyze the data using various data science libraries.

2.2 Data Wrangling

Data wrangling was necessary to clean and prepare the data for analysis. This process included handling missing values, transforming data types, and ensuring the data was in a suitable format for modeling. The following steps were taken:

- Summary Statistics: We generated summary statistics for both numerical and categorical features. This helped us understand the distribution and central tendencies of the data.
- Identifying Missing Values: We identified missing values by calculating the percentage of missing data for each feature. This allowed us to determine which features had significant amounts of missing data.

- Removing High Null Value Features: We removed the top 5 features with the highest percentage of missing values to improve model accuracy. These features were: 'PoolQC' 、'MiscFeature'、'Alley' 、'Fence' 、'FireplaceQu'

Each of these features had over 47% missing values, which could potentially decrease the model's accuracy if included.

3 Exploratory Data Analysis

3.1 Exploring the Dependent Variable

The dependent variable in our study is the housing sale price. To ensure accurate modeling, it's crucial to analyze its distribution and decide whether normalization is necessary.

Normalizing the Dependent Variable

After examining the distribution of the sale prices, we observed a skewed distribution, indicating the need for normalization. Normalizing the sale prices helps in improving the performance of various machine learning models.

3.2 Key Questions for Deeper Insights

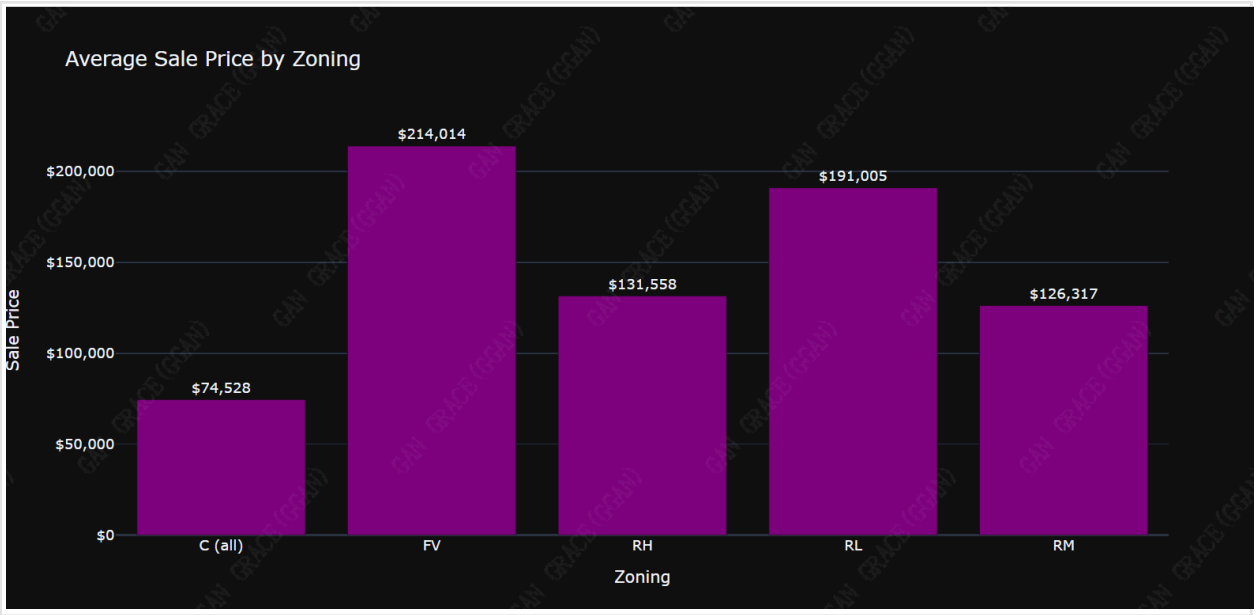
(1) Distribution of Dwelling Types and Their Relation to Sale Prices

We explored how different dwelling types relate to sale prices. The visualizations below show the distribution of various dwelling types and their corresponding sale prices.

- **Findings:**
  -Single-family detached homes are the most common, followed by townhouse end units.
  Two-family conversions are the least common.
  -Sale prices for single-family homes and townhouse end units are comparable, indicating a premium on townhouse end units.
  -From an investment perspective, townhouse end units appear to be a good option due to their relatively high sale prices and demand.

(2) Does Zoning Impact Sale Price?

We analyzed how different zoning classifications impact sale prices. The following visualizations provide insights into the relationship between zoning and sale prices.



- **Findings:**
  Floating village residential zones have the highest sale prices.
  For both investors and buyers, zoning is a significant factor affecting sale prices. The model will consider zoning information to make accurate predictions.

We also explored below questions to have deeper insights of dataset:

Does Street type affect the sale price?
What is the Average sale price by property shape?
Is there a Correlation between Property Age and Sale Price
Is there a Correlation between Living Area and Sale Price
Does the price change year to year?

These analyses provide a snapshot of how certain features influence housing prices. The use of inferential statistics and visualizations helps in understanding these relationships and informs better decision-making in real estate investments. For a more comprehensive analysis, refer to the full notebooks available in the GitHub repository.

3 Modeling

3.1 Data pre-processing

Before machine learning algorithms, we need to have some pre-processing to ensure the data in a suitable format for modeling and it can improve the performance and stability of the model. All key pre-processing steps used in this project listed below:

- Handling Missing Values: Missing values in numerical columns are replaced with the mean, and missing values in categorical columns are replaced with a constant. This ensures that the dataset has no missing values, which can otherwise cause issues during model training.
- Scaling Numerical Features: Standardizing numerical features ensures that they are on a similar scale, which is crucial for many machine learning algorithms that are sensitive to feature scaling.
- Encoding Categorical Variables: One-hot encoding transforms categorical variables into a binary format, allowing machine learning algorithms to process categorical data effectively.

3.2 Model Selection

We tested 4 different machine learning regression models: **Linear Regression, Random Forest Regressor, and XGBoost Regressor**. The output shows the RMSE for XGBoost is the lowest among three models indicating the best predictive performance. However, due to Linear Regression result is far off from the actual values, we used **Principal Component Analysis (PCA)** on a preprocessed dataset to reduce its dimensionality while retaining as much variance as possible, in this way, we can ensure that the dataset is transformed to a lower-dimensional space while retaining most of the important information, making it more suitable for further analysis or modeling.

After applying PCA on the dataset, linear regression changed to 0.164 and XGBoost was the best RMSE with value of 0.136.

We also create some new features such as ‘PropertyAge’, ‘TotalISF’ and others to improve the predictive performances of machine learning models.

Through hyperparameter tuning and cross-validation, the best model-Linear Regression is the best model with the value of 0.134.

4 Conclusions and Future Work

4.1 Conclusions

The linear regression model achieved an average error of 13.39% relative to actual house prices, demonstrating robust predictive performance. By incorporating various housing features, this model can effectively estimate property values, thereby mitigating market bubbles. This predictive capability supports real estate agents, homebuyers, and financial institutions in making data-driven decisions. The insights generated by our model provide valuable guidance for property investments and purchases, ultimately enhancing the decision-making process for our key stakeholders.

4.2 Future Work

- **Model Evaluation Metrics:** Instead of using RMSE, converting it into a percentage metric helps key stakeholders better understand the model's accuracy. This conversion provides a clearer interpretation of prediction errors relative to actual house prices, enhancing comprehensibility for non-technical audiences.
- **Feature Engineering:** Transforming the 'year' feature into a categorical variable rather than a continuous one acknowledges the independence of each year based on market conditions. This approach captures the unique market dynamics of each year, improving the model's ability to handle temporal variations effectively.
- **Data Volume:** Expanding the dataset beyond the current 1460 rows would likely enhance the model's accuracy. A larger dataset provides more comprehensive training data, enabling the model to learn more effectively and generalize better to unseen data.

5 Consulted Resources

- Anna Montoya, DataCanary. (2016). House Prices - Advanced Regression Techniques. Kaggle. https://kaggle.com/competitions/house-prices-advanced-regression-techniques