# Machine Learning Task 1 Report: Housing Prices Prediction

*Faculty of Computer Science*
*Misr International University, Cairo, Egypt*
Yahia Tamer (2202264)[1], Nouran Hassan Ahmed (2200062)[2],
Malak Mohamed (2207005)[3], Roaa khaled salah (2205885)[4]
{@miuegypt.edu.eg}

*Instructor:*
Dr. Manal Tantawy
{manal.csc0277@miuegypt.edu.eg }

***Abstract*—This paper presents a comprehensive machine learning approach to predicting housing prices based on multiple property features. We preprocess the dataset extensively by handling missing values, encoding categorical data, scaling numerical features, and performing feature selection. We implement and compare multiple regression models such as Linear Regression, Ridge Regression, and Lasso Regression, evaluating their performance using RMSE, MAE, and R-squared. Through extensive experimentation, we find that Lasso Regression outperforms other models due to its ability to perform feature selection while reducing overfitting. Additionally, we apply hyperparameter optimization techniques like Grid Search and Random Search to enhance model accuracy. Our findings demonstrate the impact of data preprocessing, feature engineering, and model tuning on prediction performance, providing key insights into the strengths and weaknesses of different machine learning approaches.**

## I. Introduction

Housing price prediction is a crucial area in real estate that benefits buyers, sellers, and investors by providing data-driven price estimations. Traditional valuation methods rely heavily on subjective expert opinions, leading to inconsistencies and biases. Machine learning offers a more objective, data-driven approach by analyzing historical data and identifying patterns that influence housing prices. The primary objective of this study is to develop a robust predictive model that minimizes error and enhances price estimation accuracy. This paper investigates different regression models, their effectiveness in predicting housing prices, and the influence of various preprocessing techniques on model performance, ultimately concluding that Lasso Regression is the most effective approach for this problem.

## II. Methodology

### *A. Data Collection and Exploration*

The dataset used in this study is sourced from a real estate database containing features such as location, number of rooms, square footage, neighborhood rating, and available amenities. The dataset undergoes an initial exploratory data analysis (EDA), which includes:

- Descriptive statistics to summarize numerical and categorical features.
- Correlation analysis to identify relationships between different attributes.
- Visualization techniques such as histograms, box plots, and scatter plots to observe distributions and outliers.

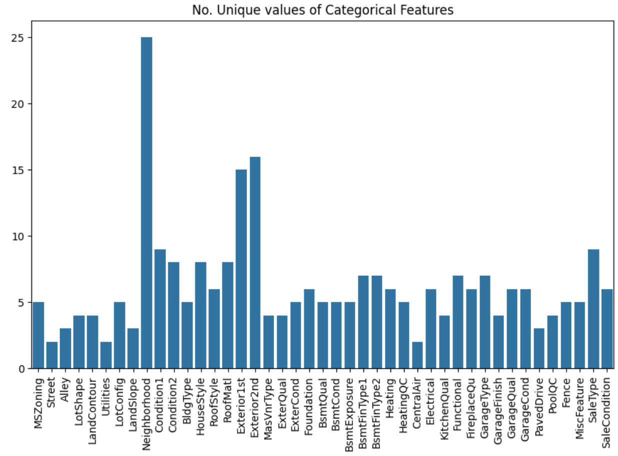

Fig. 1. Number of Unique Values of Categorical Features

### *B. Data Preprocessing*

Preprocessing is a crucial step in machine learning as it ensures that the dataset is clean, well-structured, and suitable for model training.

*1) Handling Missing Data:* Handling missing data is essential as it directly impacts model accuracy. We analyze missing data patterns and apply different strategies:

- Columns with excessive missing values (e.g., more than 50%) are dropped to maintain data quality.
- Numerical attributes are imputed using the mean or median values to prevent bias.
- Categorical attributes are imputed using the most frequent category to preserve data integrity.

*2) Feature Engineering:* Feature engineering enhances model performance by transforming categorical and numerical features appropriately:

- One-hot encoding is applied to categorical variables to convert them into numerical representations without introducing ordinal relationships.
- Numerical features are standardized using Min-Max scaling or z-score normalization to bring all features into a comparable range.
- Feature selection is conducted using techniques like Recursive Feature Elimination (RFE) and correlation analysis to eliminate redundant information.

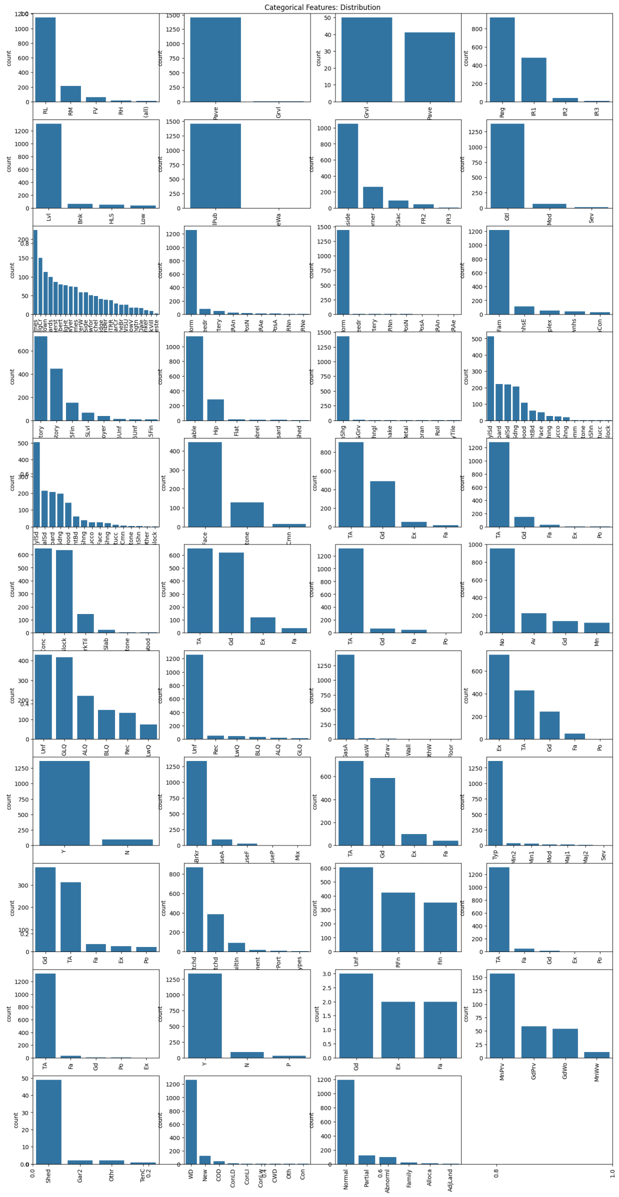

Fig. 2. Categorical Features: Distribution

*3) Data Splitting:* To ensure robust model evaluation, the dataset is split into:

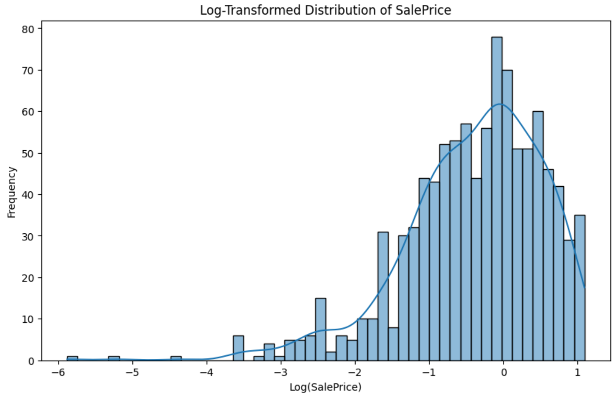

Fig. 3. Log-Transformed Distribution of SalePrice

- 70% training data
- 15% validation data
- 15% test data

This split ensures a proper balance between training and testing performance.

### C. Machine Learning Models

We implement and compare multiple regression models:

- **Linear Regression**: Assumes a linear relationship between independent features and the target variable. It serves as a baseline model.
- **Ridge Regression**: Uses L2 regularization to penalize large coefficients and reduce overfitting, particularly in high-dimensional datasets.
- **Lasso Regression**: Applies L1 regularization, which not only prevents overfitting but also performs feature selection by setting some coefficients to zero. This results in a simpler, more interpretable model with better generalization.

### D. Hyperparameter Tuning

To optimize model performance, we employ:

- **Grid Search**: A systematic approach that tests all possible combinations of hyperparameters to find the optimal configuration.
- **Random Search**: A more efficient approach that randomly samples different hyperparameter combinations within specified ranges, reducing computational cost.

## III. Results and Discussion

We evaluate each model using RMSE, MAE, and R-squared metrics. A comparative analysis before and after hyperparameter tuning reveals performance improvements. Our experiments indicate that Lasso Regression achieves the lowest RMSE and the highest R-squared score, making it the best-performing model. The feature selection capability of Lasso further enhances model interpretability and reduces noise in the dataset.

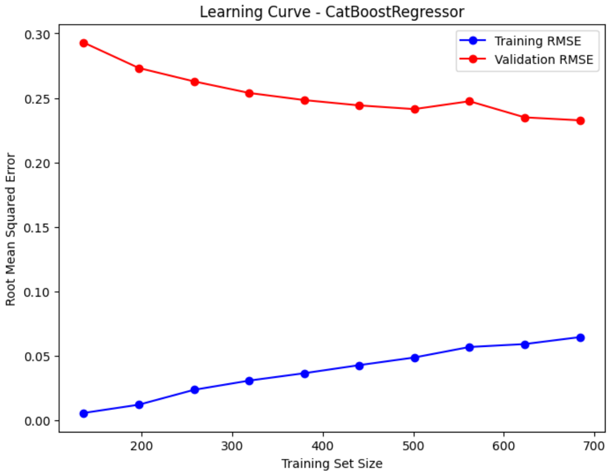

Fig. 4. Learning Curve - Cat Boost Regressor

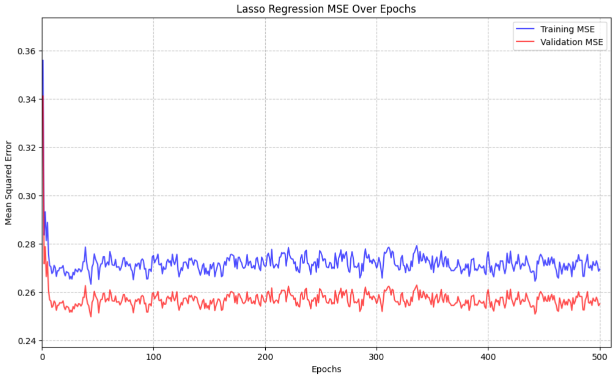

Fig. 5. Lasso Training and Learning Rate

## IV. Conclusion

This study demonstrates the effectiveness of machine learning models in housing price prediction. After extensive evaluation, we determine that Lasso Regression provides the best performance due to its ability to select relevant features while minimizing overfitting. Future work could explore deep learning techniques and integrate additional datasets to enhance model accuracy.