**Srinivas Bhootam**

**DeVos Graduate School, Northwood University**

**115537-MGT-665-NW Solv Probs W/ Machine Learning!!**

**Dr. Itauma Itauma**

**June 2nd, 2025**

**House Price Prediction using Regression Models**

**Abstract**

In this work, I propose a housing value prediction model based on the multi-variable regression model for the housing value using a huge set of housing property data. The objective is to determine how well these regression methods will be able to predict the prices of the houses. The dataset was cleaned and preprocessed and each model was trained, tested and evaluated with MAE, MEAN, RMSE. It turned out that: Ridge Regression gave an optimum compromise and better generalization than the others.

**Introduction**

Accurate prediction of housing prices is indispensable in real estate domain in providing potential buyers, seller organizations and investors with necessary information for decision making. As data on housing has become more widely available, predictive modeling is an effective tool for valuing properties. This paper presents house price prediction model based on various features like carpet area, location, number of bathrooms and furnishing status using multiple regression frameworks and demonstrates their performance.

**Literature Review**

Several machine learning techniques have been used in the past to predict the value of houses. Kumar and Garg (2020) also investigated multiple regression techniques and demonstrated that the regularization methods such as Ridge and Lasso enhanced the models' stability in the presence of multicollinearity. Similarly, Wang et al. (2021) emphasized the importance of categorical encoding and feature engineering for model accuracy. Furthermore, recent work with ensemble approaches and neural networks, for example, has proven to be more predictive, more computationally and conceptually complex. In this investigation we emphasize representative regression models that are designed for simplicity, interpretability and performance.

**Methodology**

The data set had over 93,000 rows and with various features such as

* Carpet Area
* Property Status
* Floor
* Transaction Type
* Furnishing
* Facing
* Overlooking
* Society Name
* Bathroom Count
* Balcony Count
* Car Parking Availability
* Ownership Type
* Super Area
* Dimensions
* Plot Area
* Property Title
* Total Amount
* Price Per Square Foot
* Location Description

Preprocessing steps included filling missing values using median imputation, dropping non-informative text columns such as the property title, and converting categorical variables with one-hot encoding. Numeric features were further normalized where applicable.

EDA was used for relationship and distribution visualization. Primary visualizations were the house price distribution, carpet area vs price, bathroom count vs price and a correlation heatmap between numerical features. The EDA established that Carpet Area and Bathroom Count had a weak positive correlation with Price while most other features exhibited low Pearson correlation.

After pre-processing and some EDA there were three models : Linear Regression, Ridge Regression with L2 regularization, and Polynomial Regression (degree 2). I Had to train the Polynomial Regression on a smaller dataset of 10000 rows due to memory constraints on the Collab environment. MAE, MSE, and RMSE values were used to assess the model to explain predictability accuracy and error variance.

**Results**

A graph with numbers and a number of houses

AI-generated content may be incorrect. **Figure 1**

A screenshot of a graph

AI-generated content may be incorrect. **Figure 2**

**A graph with numbers and lines

AI-generated content may be incorrect.**

**Figure 3**

**A graph of a number of bathrooms

AI-generated content may be incorrect.**

**Figure 4**

**Model Evaluation Summary**

**Linear Regression:**

* MAE: 2,699.39
* MSE: 886,907,481.35
* RMSE: 29,780.99
* Serves as the baseline model with decent performance but no regularization.

**Ridge Regression:**

* MAE: 2,681.37
* MSE: 885,889,993.86
* RMSE: 29,763.90
* Slightly outperforms Linear Regression due to L2 regularization, offering better generalization.

**Polynomial Regression (Degree 2, Sampled):**

* MAE: 2,235.81 (Lowest among all models)
* MSE: 3,918,571,158.96 (Highest)
* RMSE: 62,598.49 (Highest)
* Although it has the lowest average error (MAE), the high RMSE and MSE suggest poor generalization and severe overfitting to the sample data.

**Discussion**

Based on model comparison, Ridge was the most compromising in both prediction performances (accuracy) and error consistency. While the lowest MAE was achieved by Polynomial Regression the high RMSE showed the existence of large outlier errors, most likely caused by overfitting. And the performance of Linear Regression was like Ridge, but Linear Regression didn’t contain the regularization term, so it was relatively more likely to be influenced by variance. The EDA validated that although features such as Carpet Area and Bathroom Count had a visible influence on price, no single feature was individually strongly predictive. This highlights the need to use features in combination and to regularize the model to prevent it from overfitting.

**Conclusion**

Three regression models were successfully tested to predict housing prices with a large real state database. Ridge Regression presented itself as the best model as it performed the best overall and was the most robust. The work demonstrates the necessity of elaborate data preprocessing, suitable features selection and model evaluation steps for developing trustable machine learning applications. Future works could place interest in ensembles, more sophisticated feature engineering and geographical clustering to improve accuracy.

All the codes and analysis are available at the link below:

**GitHub Link:**

**References**

Kumar, R., & Garg, A. (2020). Comparative study of regression techniques for house price prediction. *International Journal of Computer Applications*, 175(7), 1-5.

Wang, J., Liu, Y., & Zhao, H. (2021). Improving house price prediction accuracy through feature engineering and machine learning. *Procedia Computer Science*, 183, 271-278.

Scikit-learning Developers. (2023). *Scikit-learn: Machine Learning in Python*. <https://scikit-learn.org/stable/index.html>