Jennifer Guo, Jintong Hong, Sihan Chen, Wenjie Han
BIOSTAT 625 - Fall 2024

Final Project Proposal

Introduction:

Accurately determining the market value of properties is a critical step in the home-selling process. The price of a property is influenced by a variety of factors, including its location, the floor area, the number of rooms, and additional features such as balconies and amenities. Our project aims to develop a predictive model that estimates house prices based on property attributes, thus, providing a practical price referencing tool for buyers and sellers. The dataset used in this analysis is derived from real estate listings on agent websites in India (Bhojani, 2023). Specifically, we will employ and evaluate the performance of two modeling approaches—Elastic Net and Random Forest—to identify the most effective model for price estimation.

Methods:

Data Cleaning

The dataset consists of 21 variables and 187,531 observations. The data are the real estate property details and prices scraped from the Magic Bricks website, which is available on Kaggle by Bhojani Juhi. An initial exploratory analysis was conducted to assess the data quality, identify missing values, and ensure that the appropriate covariate type was assigned. Some numerical variables, such as "balconies" and "bathrooms", were transformed into categorical variables to mitigate the influence of extreme outliers. Bivariate encoding was applied for categorical covariates with only two categories, while one-hot encoding was employed for variables with multiple categories. In particular, the "location" variable, which contains 81 unique categories, will be recategorized with clustering against the socioeconomic level of each city (planning to scrape from the official Indian website) to reduce dimensionality. Numerical variables, including "carpet area," "current floor," and "total floors," were standardized to improve model performance. Normalization was excluded due to the presence of significant outliers suggested by the boxplot analysis.

Modeling Approaches

Elastic Net: Elastic Net is a regression method that uses two types of regularization, L1 (Lasso) and L2 (Ridge). Our dataset contains some variables with high dimensionality such as “location”, as well as variables that are highly correlated to each other such as “carpet area” with “super area” or “current floor” with “total floors”. Elastic Net is effective at handling high dimensionality and multicollinearity. It is also possible to balance the effects of the L1 and L2 regularization with a mixing parameter, making this a flexible model.

Random Forest: To address the considerable missing values, outliers, and categorical covariates present in the housing price dataset, we have chosen to employ random forest classifiers for model construction. This machine-learning approach can handle both continuous and categorical data well while being resilient to the impact of outliers. In addition, hyperparameter optimization will also be performed in order to further enhance the model’s predictive performance.

Model Performance Assessment:

Cross-validation is a robust statistical method designed to evaluate the predictive performance of models, particularly in the fields of data science and machine learning. It assesses how the results of a statistical analysis will generalize to an independent data set, particularly useful in scenarios where model generalizability is paramount. The technique is fundamental to avoiding biases associated with the random sampling of training and test datasets. Therefore, we are going to use cross-validation to check the accuracy of our model, both for random forest and elastic net, then to see which model performs better for our dataset.

References

Berrar, Daniel. "Cross-validation." (2019): 542-54

Bhojani, J. (2023). House price dataset. Kaggle. https://www.kaggle.com/datasets/juhibhojani/house-price

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x