**Milestone final**

**Predicting Housing Prices Based on Real Estate Features**

Bobga-Herman Gwanvoma

Bellevue University

DSC 630 – Predictive Analytics

Professor Andrew Hua

February 28, 2025

**Introduction**

Accurately estimating housing prices is an essential aspect of the real estate industry, affecting homebuyers, sellers, and investors. The ability to predict housing prices enables stakeholders to make informed decisions regarding property valuation, investment potential, and market trends. This project aims to use machine learning techniques to predict housing prices based on real estate features from the Ames Housing dataset. The dataset contains over 80 attributes that describe various aspects of residential properties, including size, location, quality, and condition. The study seeks to answer three fundamental questions: (1) What are the most influential factors in determining housing prices? (2) How accurately can machine learning models predict housing prices? (3) Which machine learning model performs best in terms of predictive accuracy? Through extensive data preprocessing, feature engineering, and model evaluation, this research aims to develop an optimal predictive model that can provide valuable insights into housing price trends.

**Data Selection**

The Ames Housing dataset was chosen for this study due to its comprehensive and structured nature. It provides detailed information on residential properties, including numerical and categorical features that significantly influence home prices. This dataset was sourced from Kaggle and is widely used in predictive modeling research due to its high-quality and well-documented variables. Key features of the dataset include GrLivArea (above-ground living space), OverallQual (quality of construction and materials), Neighborhood (geographical location), GarageCars (number of garage spaces), and YearBuilt (year of construction). These features, among others, play a crucial role in determining a home's market value.

To prepare the data for modeling, several preprocessing steps were performed. Missing values in critical features such as LotFrontage and GarageYrBlt were imputed using domain-specific strategies. LotFrontage values were filled using the median within the same neighborhood, ensuring location-based consistency. For homes without garages, GarageYrBlt was set to zero to maintain uniformity in the dataset. The target variable, SalePrice, exhibited right skewness, which could impact model performance. To correct this, a log transformation was applied to stabilize variance and improve predictive accuracy. Additional transformations were performed to handle multicollinearity, particularly among features such as 1stFlrSF and TotalBsmtSF, which were highly correlated. One-hot encoding was applied to categorical variables like Neighborhood, and numerical features were standardized for consistency across models.

**Modeling & Methods**

Multiple machine learning models were implemented to assess their effectiveness in predicting housing prices. The models tested included Linear Regression, Ridge and Lasso Regression, Decision Tree Regression, Random Forest Regression, and Gradient Boosting (XGBoost). Each model was evaluated based on key performance metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R² (coefficient of determination). RMSE measures predictive accuracy by quantifying the standard deviation of residuals, while MAE indicates the average magnitude of errors. The R² metric evaluates how well the model explains the variance in SalePrice.

The Gradient Boosting (XGBoost) model emerged as the best-performing model, achieving an R² score of 0.91, meaning it explained 91% of the variance in housing prices. Random Forest followed closely, while Ridge Regression slightly outperformed standard Linear Regression by addressing multicollinearity. Decision Tree Regression had the weakest performance, primarily due to overfitting.A feature importance analysis was conducted using the XGBoost model. The most influential factors in predicting home prices were GrLivArea, OverallQual, and Neighborhood. Larger homes, as indicated by GrLivArea, commanded higher prices, while higher construction quality (OverallQual) strongly correlated with increased value. Neighborhood also played a significant role, with properties in high-income areas such as NoRidge, StoneBr, and Timber selling at premium prices. Surprisingly, lot size had a weaker-than-expected effect, suggesting that buyers prioritize interior space over large outdoor areas.

**Results Interpretation**

The analysis confirmed that certain features significantly impact housing prices. The results indicated that home size, quality, and location are the strongest determinants of property value. The scatterplot comparing actual versus predicted prices demonstrated strong alignment between the two, confirming model reliability. A boxplot analysis of Neighborhood versus SalePrice reinforced the importance of location in determining house prices. The visualization revealed that properties in affluent neighborhoods consistently had higher median prices, validating real estate market trends. To further evaluate model performance, predictions from the XGBoost model were compared against actual sale prices. The model exhibited a strong ability to capture price variations, with minimal deviation from actual values. While minor prediction errors were observed, they were within an acceptable range, indicating robust predictive capability.

**Conclusion**

This study demonstrates that machine learning techniques can effectively predict housing prices based on real estate features. The results indicate that Gradient Boosting (XGBoost) is the most reliable model, achieving an R² score of 0.91. The most influential predictors of home prices include above-ground living space, overall home quality, and neighborhood. From a practical perspective, these findings offer valuable insights for various stakeholders. Homebuyers should prioritize factors such as property size and construction quality over lot size when making purchasing decisions. Real estate agents can leverage these insights to develop more accurate pricing strategies, emphasizing home quality and location. For investors, targeting high-income neighborhoods and focusing on property upgrades may yield the highest returns.

Future research could expand on this study by incorporating additional economic indicators, such as mortgage rates, employment levels, and broader market trends. Additionally, exploring deep learning techniques could enhance predictive accuracy by capturing complex interactions among features. A longitudinal analysis of housing price trends over time could also provide deeper insights into market dynamics and investment opportunities. Overall, this study highlights the power of predictive analytics in the real estate sector and underscores the importance of data-driven decision-making in property valuation.

**References**

Prevek18. (n.d.). Ames Housing Dataset [Data set]. Kaggle. Retrieved December 08, 2024, from https://www.kaggle.com/datasets/prevek18/ames-housing-dataset

XGBoost. (n.d.). XGBoost documentation. Retrieved December 08, 2024, from https://xgboost.ai