Milestone 3: Audience Questions: Predicting House Prices

Author: Ramesh Talapaneni

Bellevue University – DSC 680-T302: Applied Data Science

Instructor: Amirfarrokh Iranitalab

Due Date (Week 8): February 02, 2025

1. What are the primary factors that influence house prices based on your analysis?

Based on our exploratory data analysis, the most significant factors influencing house prices include OverallQual (Overall Material and Finish Quality), which has a strong positive correlation with higher prices, as homes with better finishes and materials are more desirable. GrLivArea (Above-Ground Living Area in Square Feet) also plays a critical role, as larger living spaces typically command higher prices. GarageCars (Garage Size in Terms of Cars) and TotalBsmtSF (Total Basement Area in Square Feet) further contribute to property value, with larger garages and basements adding to the utility and appeal of a home. Additionally, YearBuilt (Original Construction Year) indicates that newer homes tend to be priced higher due to modern amenities and reduced maintenance needs.
Finally, Neighborhood is a key factor, as location significantly impacts demand andpricing, with some areas being more sought-after than others.

2. How did you handle missing data and outliers in the dataset, and why did you choose these specific methods?

To handle missing data, we imputed numerical values using the **median**, as it is less sensitive to extreme values compared to the mean, ensuring a more robust representation of the data. For categorical features, we filled missing values with **"None"** to maintain consistency and avoid introducing bias. For outliers, we used the **Interquartile Range (IQR)** method to identify and remove extreme values from key variables like **GrLivArea** and **SalePrice**. This approach helped reduce noise and improve model performance. While log transformations were considered to normalize skewed data, we opted not to apply them universally to preserve the interpretability of the results.

3. Why did you select the specific machine learning models (Linear Regression, Random Forest, XGBoost) for this project, and how do their performances compare?

We selected **Linear Regression** for its simplicity and interpretability, making it an excellent baseline model for comparison. **Random Forest** was chosen for its ability to handle nonlinear relationships and provide insights into feature importance, which is valuable for understanding the drivers of house prices. **XGBoost** was included due to its superior performance in regression tasks and its ability to model complex data patterns effectively. In terms of performance, **XGBoost** achieved the lowest RMSE (~18,000), indicating the highest predictive accuracy. **Random Forest** followed with an RMSE of ~19,500, offering a balance between interpretability and performance. **Linear Regression** had the highest RMSE (~22,000), reflecting its limitations in capturing nonlinear relationships.

4. What are the key insights from the correlation heatmap and feature importance analysis?

The correlation heatmap revealed that **OverallQual** and **GrLivArea** had the strongest positive correlations with **SalePrice**, highlighting their importance in determining house prices. The feature importance analysis from both **Random Forest** and **XGBoost** models confirmed these findings, with **OverallQual**, **GrLivArea**, **GarageCars**, and **TotalBsmtSF** emerging as the most influential variables. These insights underscore the significance of home quality, size, and additional features like garages and basements in driving property values.

5. How does neighborhood impact house prices, and which neighborhoods were identified as having the highest and lowest median sale prices?

Neighborhood plays a critical role in determining house prices due to factors like desirability, accessibility, and local amenities. Our analysis identified **NridgHt** and **StoneBr** as the highest-priced neighborhoods, likely due to their high-quality homes and prime locations. On

the other hand, **MeadowV** and **IDOTRR** had the lowest median sale prices, which can be attributed to smaller homes and less desirable locations. These findings emphasize the importance of location in real estate valuation.

6. What ethical considerations were considered when designing the predictive model, particularly with respect to bias and fairness?

To ensure ethical modeling, we took several steps to mitigate bias and promote fairness. We avoided using potentially discriminatory variables, such as demographic information, to prevent the model from perpetuating biases. Additionally, we ensured that the model did not disproportionately favor affluent neighborhoods by carefully selecting features and validating predictions across diverse areas. Transparency was also a priority, as we documented all preprocessing steps and assumptions to maintain accountability and allow for reproducibility.

7. What preprocessing steps were applied to the dataset, and how did these steps improve model performance?

Several preprocessing steps were applied to enhance the dataset and improve model performance. Missing numerical values were imputed using the median to reduce bias, while categorical features were filled with **"None"** to maintain consistency. Feature encoding was used to convert categorical variables into numerical formats, enabling their use in machine learning models. Scaling numerical features ensured that all variables were on a comparable scale, improving model convergence. Finally, outlier removal using the IQR method reduced noise and enhanced predictive accuracy.

8. How do the evaluation metrics (RMSE) differ across the models, and what does this tell us about their predictive power?

The evaluation metrics, particularly **Root Mean Squared Error (RMSE)**, provide insights into the predictive power of each model. **XGBoost** achieved the lowest RMSE (~18,000), indicating the highest accuracy in predicting house prices. **Random Forest** followed with an RMSE of ~19,500, offering a good balance between accuracy and interpretability. **Linear Regression** had the highest RMSE (~22,000), reflecting its limitations in capturing complex, nonlinear relationships. These results highlight the trade-offs between model complexity and performance.

9. What additional datasets or features could improve the accuracy and robustness of the model in future work?

To further enhance the model, additional datasets and features could be incorporated. **Macroeconomic indicators**, such as interest rates and inflation trends, could provide context for broader market conditions. **Crime rates** and **school ratings** would offer insights into neighborhood safety and education quality, both of which influence property values. **Climate data** could also be valuable, as weather conditions and natural disaster risks impact home desirability and pricing. Including these features would make the model more comprehensive and robust.

10. How can the insights and predictions from this project be practically applied by real estate stakeholders, such as agents, buyers, and investors?

The insights and predictions from this project have practical applications for various real estate stakeholders. **Real estate agents** can use the predictive pricing models to set competitive listing prices and advise clients effectively. **Home buyers** can leverage the model to identify underpriced properties and make informed purchase decisions. **Investors** can evaluate potential return on investment (ROI) by analyzing historical price trends and neighborhood

growth patterns. Overall, the model provides actionable insights that can enhance decision-making and strategy in the real estate market.