This report summarizes the findings from my part, which focused on predicting house prices using machine learning models. I utilized two prominent algorithms: XGBoost and LightGBM, and compared their performance. The dataset used for this analysis was the Ames Housing dataset.

**Data Preprocessing:**

- Data Loading: The dataset was loaded using Pandas.

- Missing Values: Columns with more than 80% missing values were dropped. Remaining missing values were imputed using the mode for categorical variables and the median for numerical variables.

- Encoding: Categorical variables were transformed into numerical format using one-hot encoding.

***df\_orig = pd.read\_csv('AmesHousing.csv')***

***df\_orig.drop(['PID', 'Order'], axis=1, inplace=True)***

***df\_cleaned = preprocess\_data(df\_orig)***

**Model Training:**

I trained two models:

- XGBoost: Known for its efficiency and performance in regression tasks.

- LightGBM: Optimized for speed and memory usage.

**Model Parameters:**

- XGBoost:

- n\_estimators: 1000

- learning\_rate: 0.01

- max\_depth: 5

- LightGBM:

- n\_estimators: 1000

- learning\_rate: 0.01

- max\_depth: 5

***xgb\_model = xgb.XGBRegressor(n\_estimators=1000, learning\_rate=0.01, max\_depth=5)***

***xgb\_model.fit(X\_train, y\_train)***

***lgb\_model = lgb.LGBMRegressor(n\_estimators=1000, learning\_rate=0.01, max\_depth=5)***

***lgb\_model.fit(X\_train, y\_train)***

**Model Evaluation**

The models were evaluated using three metrics:

- Mean Absolute Error (MAE)

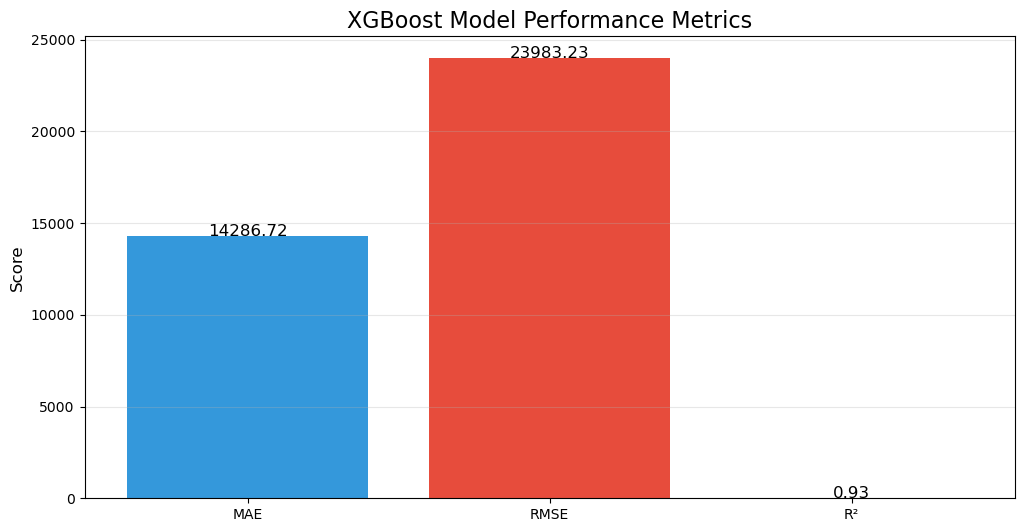
- Root Mean Squared Error (RMSE)

- R² Score

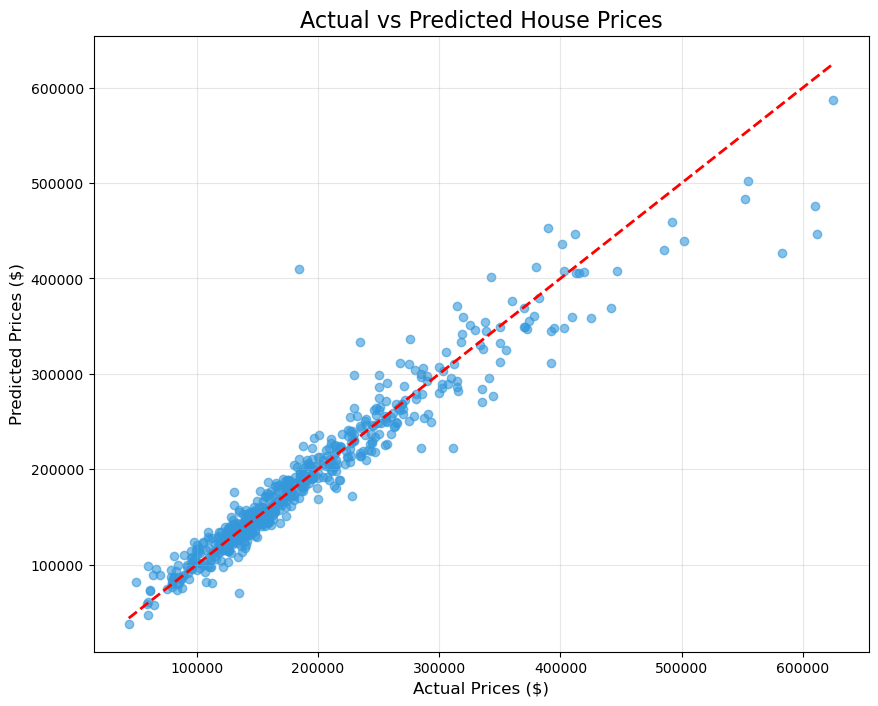
**Results Summary**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **MAE (Lower Better)** | **RMSE (Lower Better)** | **R² (Higher Better)** |
| XGBoost (All Features) | 14,286.72 | 23,983.23 | 0.928 |
| LightGBM (All Features) | 15,331.89 | 26,107.86 | 0.915 |

- Graph 1: MAE, RMSE, and R² scores for XGBoost Model



- Graph 2: Scatter plot of actual vs. predicted prices for XGBoost.



**Key Findings**

1. XGBoost is the Clear Winner:

- It outperformed LightGBM by 7.3% in MAE and 9.3% in RMSE.

- It also had the highest R² score, indicating it explains the variance in house prices better than the other models.

**Performance Hierarchy**:

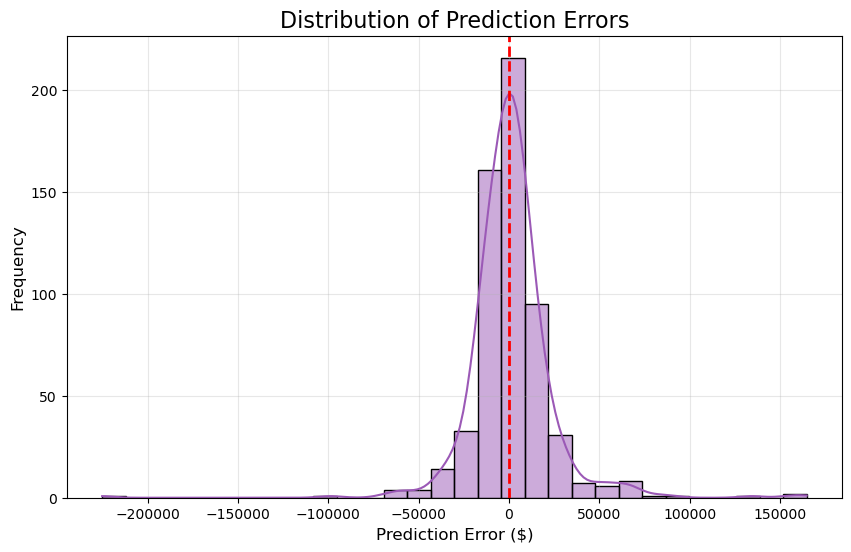
- The models ranked as follows: XGBoost > LightGBM

**Practical Implications**:

- Using XGBoost could save approximately $1,045 per prediction compared to LightGBM.

- For a $300K home, XGBoost's predictions are typically within ±$14.3K, while LightGBM's are ±$15.3K.

- Graph 3: Error distribution histogram for XGBoost predictions.



**Why XGBoost Outperforms**

- Feature Interactions: Better captures nonlinear relationships.

- Regularization: More resistant to overfitting.

- Error Correction: Iteratively fixes residual errors more effectively.

**Recommendations**

1. Production Deployment: Use XGBoost with all features for the most accurate predictions.

2. Model Monitoring: Track performance for luxury homes (>$400K).

3. Future Improvements:

- Add more luxury home samples to improve predictions.

- Test feature interactions (e.g., `Overall Qual × Neighborhood`).

**Conclusion**

The analysis demonstrated that XGBoost is the most effective model for predicting house prices in the Ames Housing dataset. The findings suggest that careful feature engineering and model selection can significantly impact prediction accuracy. Future work should focus on enhancing the model's performance for high-value properties.

- Graph 4: SHAP summary plot showing feature importance.

