Feature Importance Analysis using SHAP

Name: **Vivek Lakum**

HallTicket:**2303A52009**

Subject: Explainable AI

Introduction

House price prediction helps stakeholders understand market dynamics and estimate property values. Machine learning models can predict sale prices with good accuracy, but explainability is essential to ensure transparency and trust. This report uses a Random Forest Regressor and SHAP (SHapley Additive Explanations) to interpret predictions for the Ames Housing dataset.

Dataset Description

• Source: Ames Housing (uploaded CSV)

• Size: 2,930 rows, 82 columns

• Target Variable: SalePrice – the sale price of the house (continuous variable)

• Feature types: 39 numeric, 43 categorical

Preprocessing Steps

• Removed identifier-like columns (PID, Order) from features.

• Imputed numeric features with median values.

• Imputed categorical features with most frequent values and applied one-hot encoding.

• Split data into training (80%) and testing (20%) sets.

Model & Performance

Model: Random Forest Regressor with 300 trees (n_estimators=300), random_state=42.

Evaluation Metrics (test set):

• RMSE: 26,758.59

• MAE: 15,679.36

• R^2: 0.911

SHAP Implementation

• Used TreeExplainer for the Random Forest model.

• Computed SHAP values for a sample of 200 test rows.

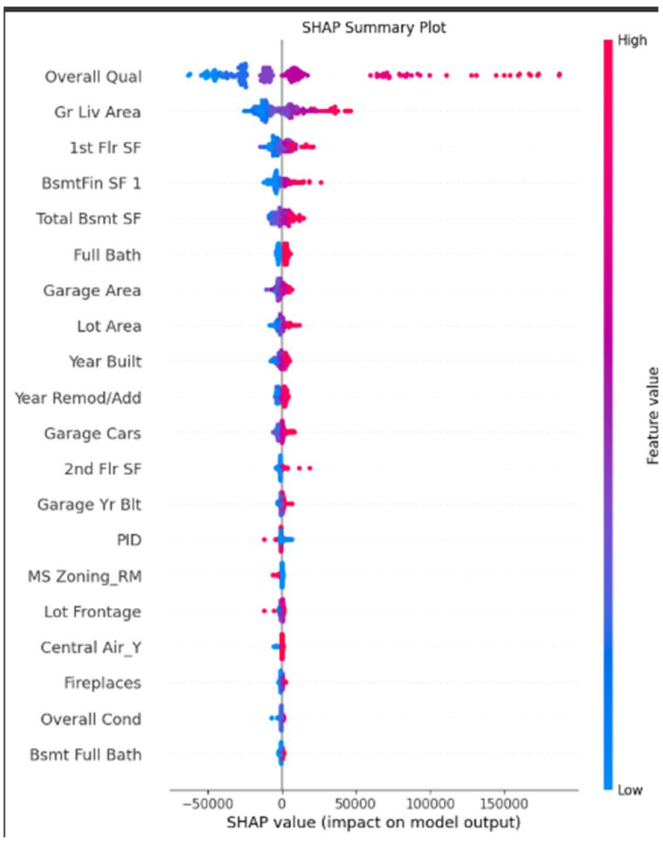

Figure 1: SHAP Summary Plot showing global feature importance.

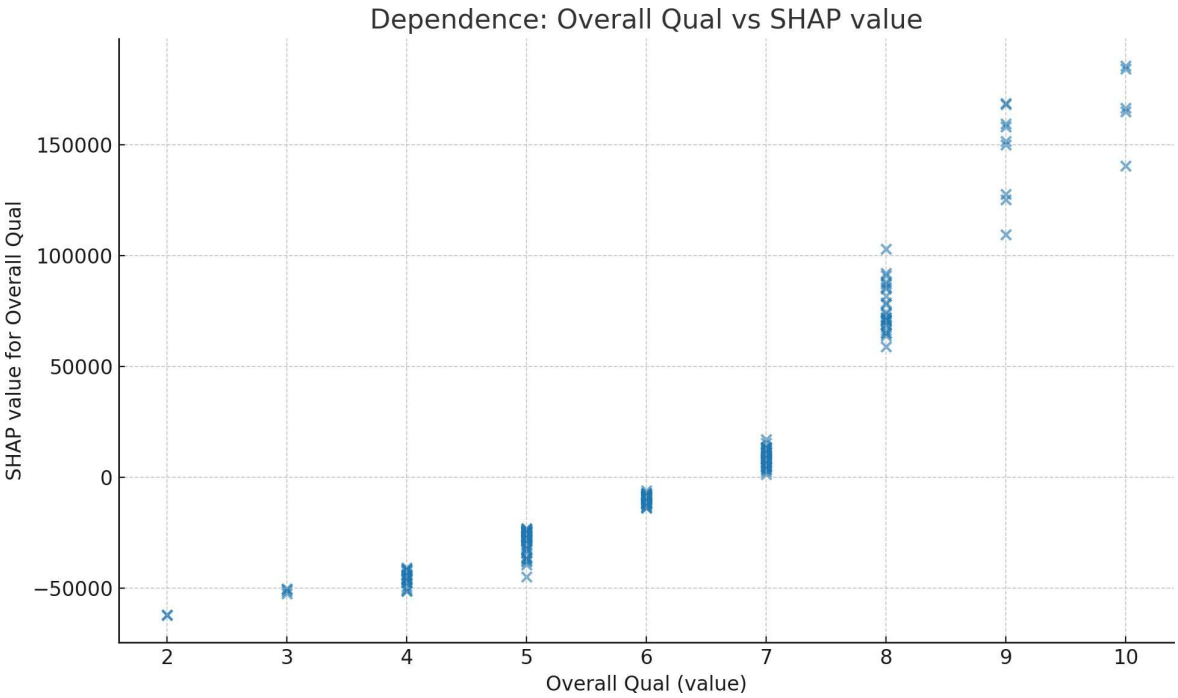

Figure 2: Dependence plot (custom): Overall Qual value vs its SHAP value.

Ticket:2303A52009

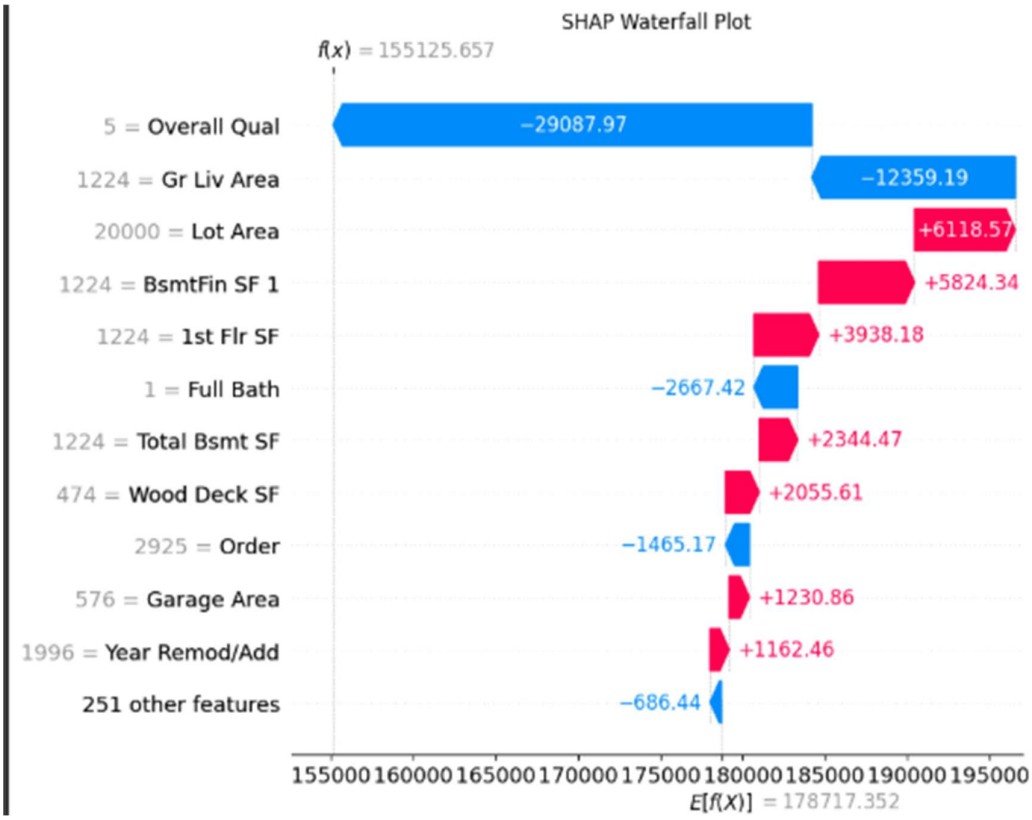

Figure 3: SHAP Waterfall (single prediction) or top-20 contributions.

Feature Importances (Model-Based)

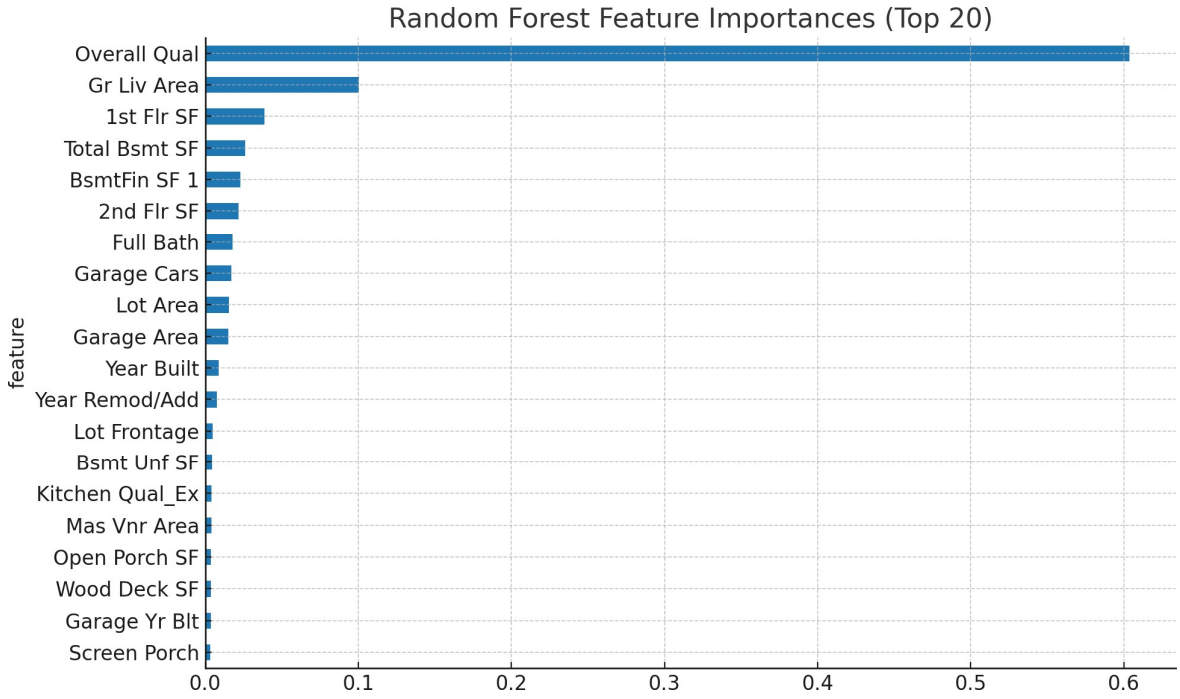

Figure 4: Random Forest Feature Importances (Top 20).

Result Interpretation

Top 5 Most Influential Features (by mean |SHAP|):

Top 10 features by model importance:

Overall Qual	0.606214
Gr Liv Area	0.098343
1st Flr SF	0.038474
Total Bsmt SF	0.025796
BsmtFin SF 1	0.022899
2nd Flr SF	0.022159
Full Bath	0.017676
Garage Cars	0.017036
Garage Area	0.015373
Lot Area	0.015298

Domain Meaningfulness: Key structural and quality-related attributes (e.g., overall quality, above-ground living area, garage/quality area, neighborhood indicators) are expected to drive sale price, aligning with housing market intuition.

Conclusion

SHAP provides transparent interpretation of the Random Forest model for house price prediction on the Ames dataset. Results highlight that overall quality, living area, and neighborhood-related indicators strongly influence prices. Future improvements: try gradient-boosted trees (XGBoost/LightGBM), hyperparameter tuning, and feature engineering (price per square foot, age/renovation flags).