Feature Importance Analysis using SHAP

Name: (Your Name)

Hall Ticket: (Your ID)

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.

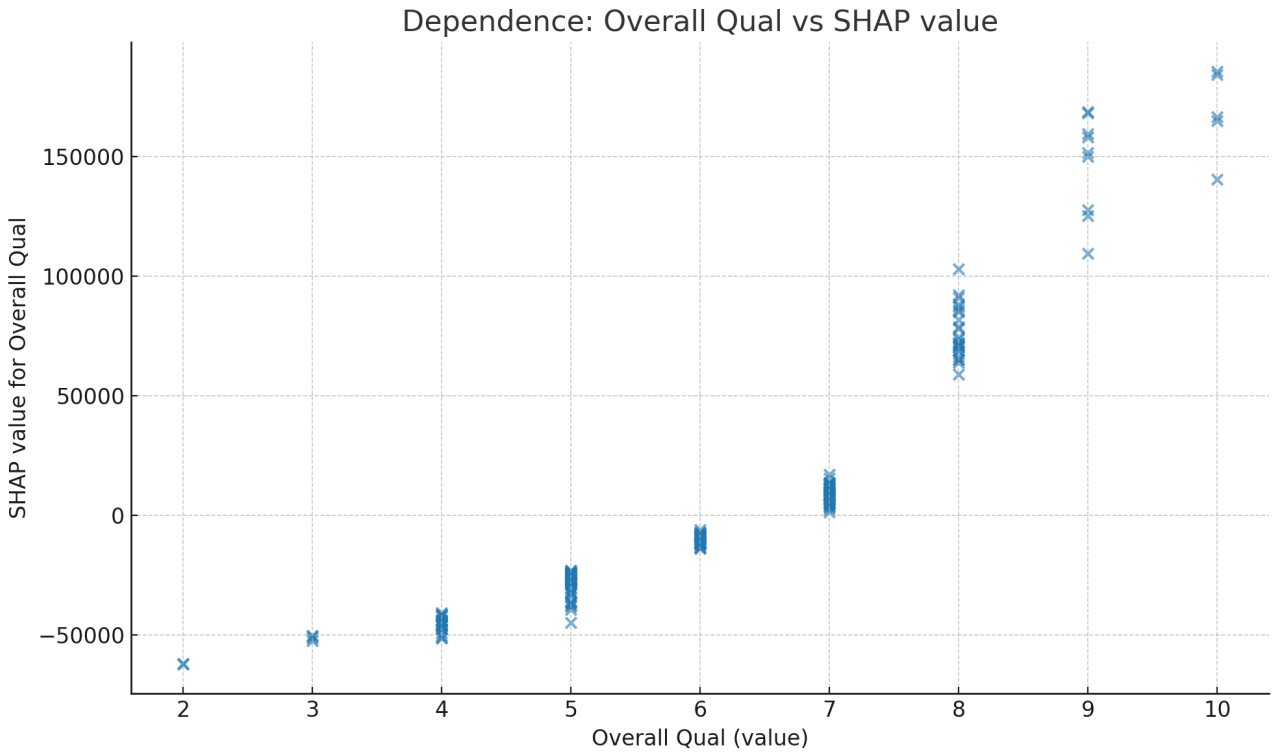

Figure 2: Dependence plot (custom): Overall Qual value vs its SHAP value.

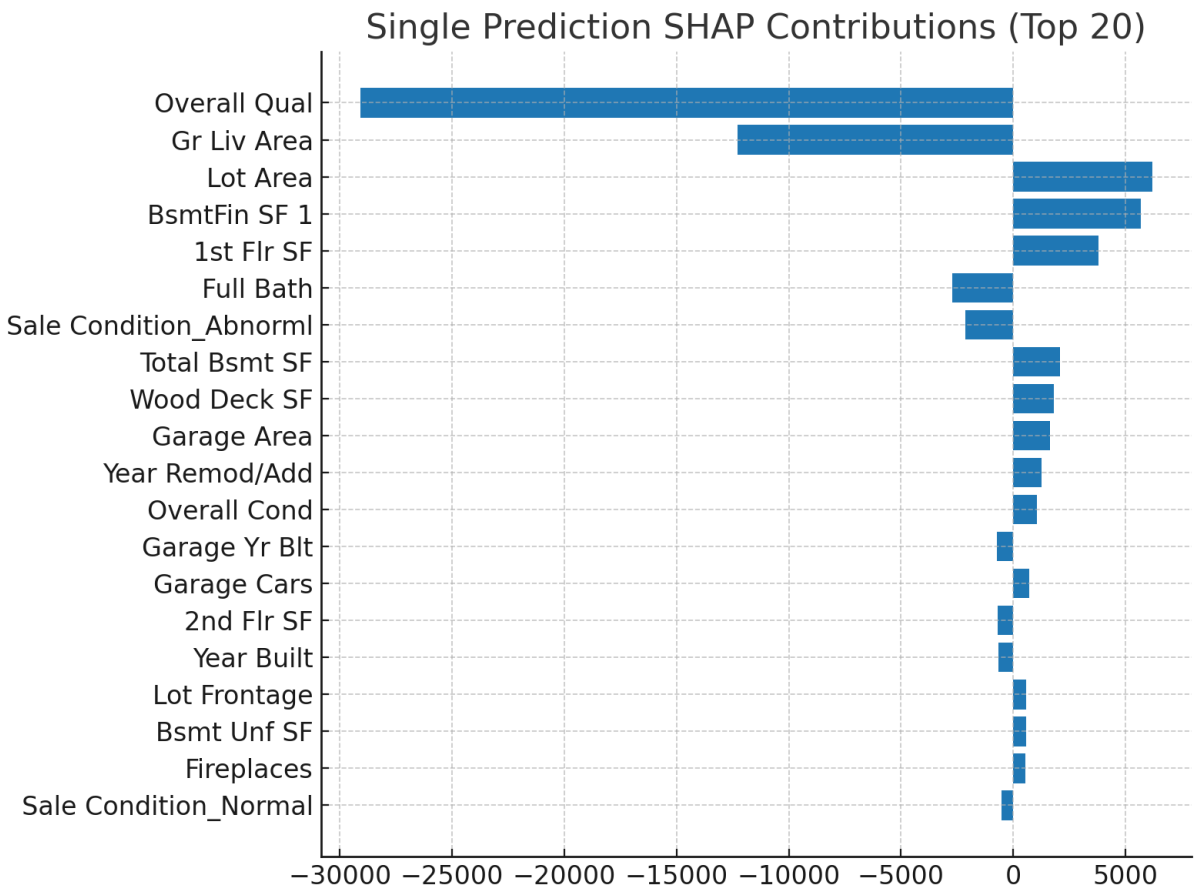

Figure 3: SHAP Waterfall (single prediction) or top-20 contributions.

Feature Importances (Model-Based)

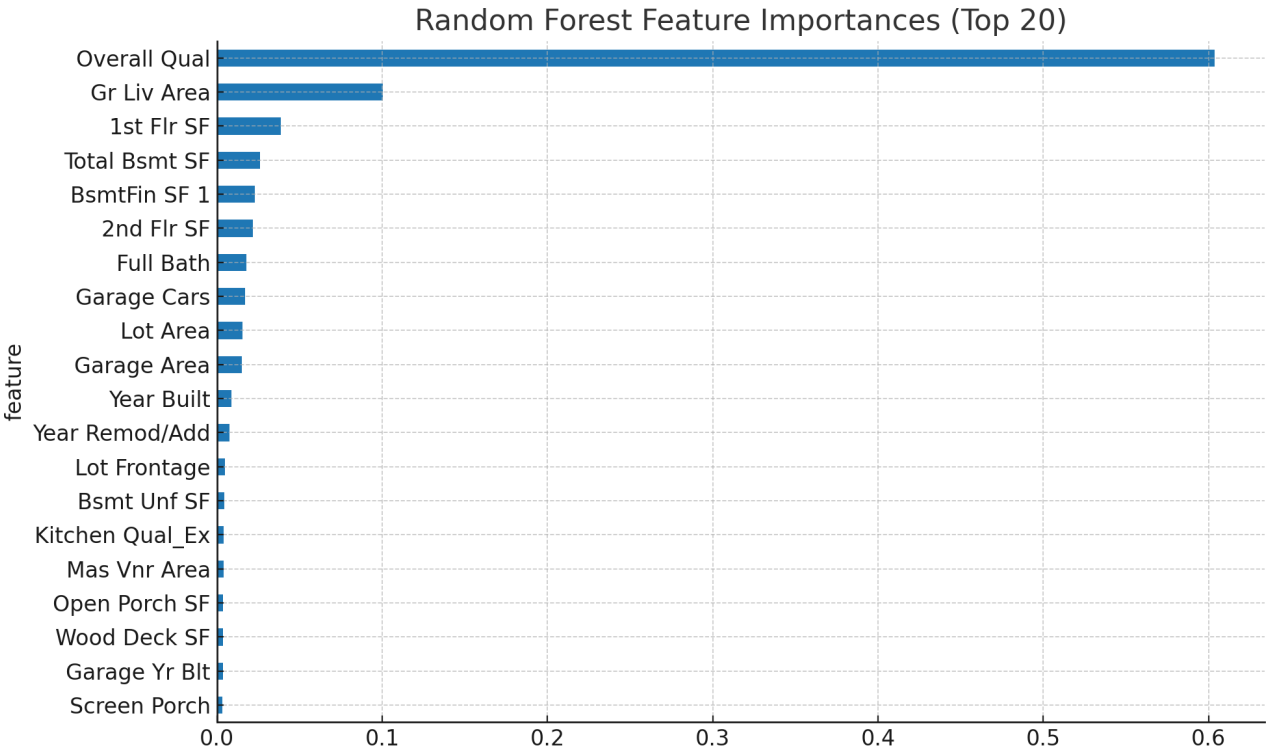

Figure 4: Random Forest Feature Importances (Top 20).

Result Interpretation

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

1. Overall Qual

2. Gr Liv Area

3. 1st Flr SF

consistent.

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).