King County House Price Prediction Analysis Report

1. Data Overview

Sample Size: 30 property records (Note: The small sample size may limit the robustness of the analysis)

Structured Features: 20 attributes (e.g., house area, number of bedrooms, geographic location)

Text Features: 200 attributes (terms extracted from property descriptions using TF-IDF)

2. Model Training and Evaluation

2.1 Model Incorporating Text Features

Evaluation Metrics:

RMSE: 162,137.95

MAE: 151,258.19

R^2: 0.1722

Interpretation:

High Error: The Root Mean Square Error (RMSE) of $162,137.95 indicates significant deviation in predictions.

Low Explanatory Power: An R^2 of 0.1722 suggests that the model explains only a small portion of the variance in house prices.

2.2 Baseline Model (Using Only Structured Data)

Evaluation Metrics:

RMSE: 140,985.62

MAE: 119,773.66

R^2: 0.3741

Interpretation:

Lower Error: An RMSE of $140,985.62, which is lower than that of the model incorporating text features, indicating better prediction accuracy.

Higher Explanatory Power: An R² of 0.3741 suggests that this model explains a greater portion of the variance in house prices.

2.3 Possible Reasons

(1) Quality of Text Features: Property descriptions may contain limited information relevant to price prediction.

(2) Feature Processing**: The TF-IDF method might not capture deep semantic meanings, leading to less effective text features.

(3) Model Complexity**: Incorporating high-dimensional text features could increase model complexity, potentially causing overfitting or reduced performance.

3. Causal Inference Analysis (Impact of House Age on Price)

Insignificant Impact of Age: In both models, the effect of house age on price is statistically insignificant (P-values > 0.05).

Low Explanatory Power: The low R² values in both models indicate that house age alone does not substantially explain the variance in house prices.

4. Summary and Recommendations

1. Model Performance Summary:

Model Incorporating Text Features: Exhibits higher error and lower explanatory power, suggesting that text features did not enhance model performance effectively.

Baseline Model (Structured Data Only): Shows lower error and higher explanatory power, indicating that structured data is more effective for predicting house prices.

2. Causal Inference Summary:

The impact of house age on price is statistically insignificant in both models, implying that factors other than age play more substantial roles in determining house prices.