Nastassia Pukelik
Prof. Jairam
CIS 9660
August 4, 2024

Business Problem

This project focused on predicting whether a home is considered expensive based on its physical characteristics and location. The goal was to create a machine learning classifier capable of labeling a house as 'expensive' or 'not expensive' using real-world housing data from King County, Washington.

Data and Preprocessing

The dataset was sourced from Kaggle and includes residential property sales between 2000 and 2015. Features used include bedrooms, bathrooms, square footage, waterfront presence, view score, construction grade, latitude, and more. The dataset was cleaned, filtered, and split into training and testing sets (70/30). Numerical features were standardized using StandardScaler. A binary target variable was created to indicate whether a home was 'expensive' (above a certain price threshold).

Modeling and Results

Seven classification models were trained and compared: Logistic Regression, Naive Bayes, Decision Tree, Random Forest, Support Vector Machine, K-Nearest Neighbors and K-Means.

The Random Forest model achieved the highest test accuracy at 91.78%, with a weighted F1-score of 0.92. Key predictive features were 'sqft_living', 'grade', and 'lat'. Cross-validation accuracy for Random Forest was 91.85%, making it the most stable and effective model.

Insights and Takeaways

Homes with higher square footage, better construction grades, and northern locations within the county tended to be classified as expensive. Nearly all new homes were classified as expensive. Grade 9 and above was a significant threshold for predicting expensive homes. Random Forest's feature importance chart confirmed these insights. Visual analysis showed strong class separation by sqft and grade.

Limitations and Next Steps

The model does not account for zip code, school district, or recent renovations, which may also influence home value. Also, external economic factors were not included.

Kaggle Dataset Link: https://www.kaggle.com/datasets/harlfoxem/housesalesprediction

Appendix: Visualizations and Interpretations

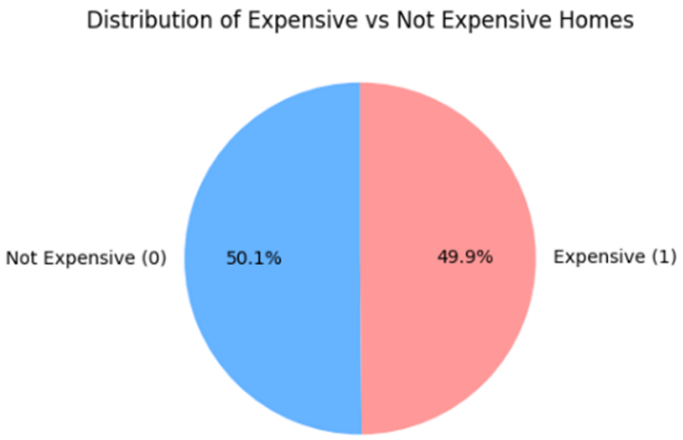

Figure 1: Class distribution (balanced dataset). Nearly 50/50 split between expensive and non-expensive homes.

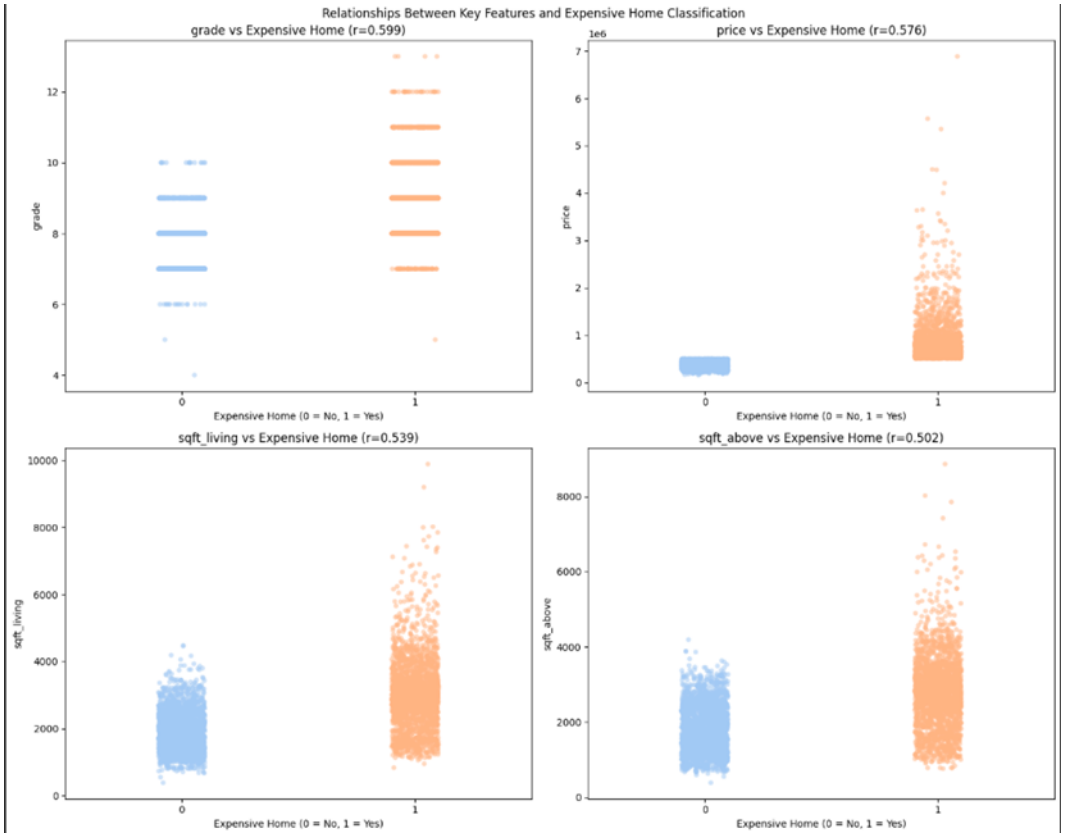

Figure 2: Top features by visual separation include grade, price, sqft_living, and sqft_above.

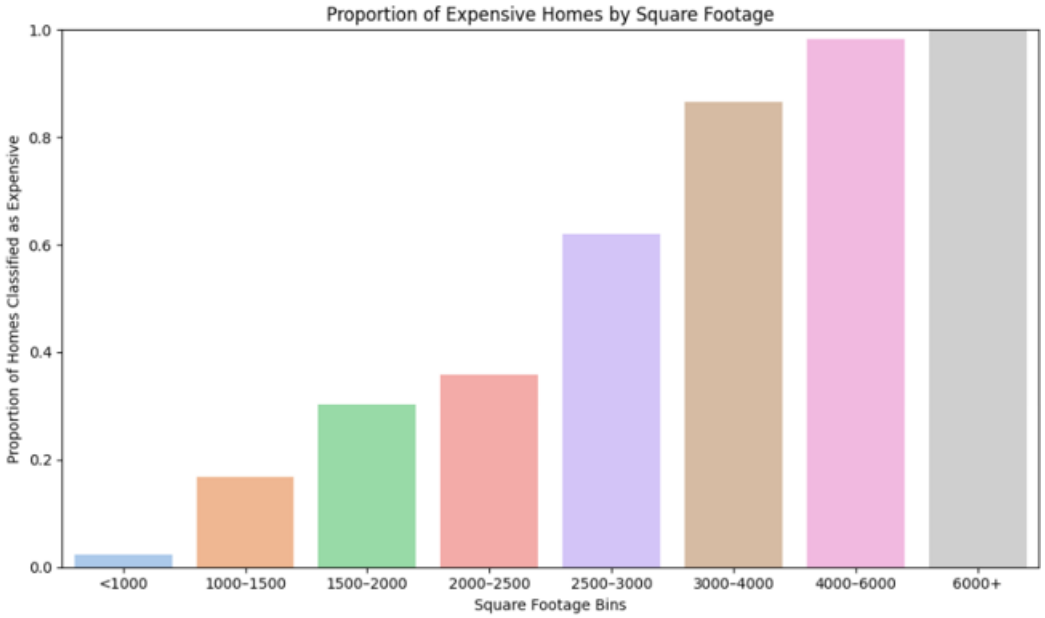

Figure 3: The proportion of homes classified as expensive increases sharply with square footage, especially beyond 2500 sq ft, with nearly all homes above 4000 sq ft labeled as expensive.

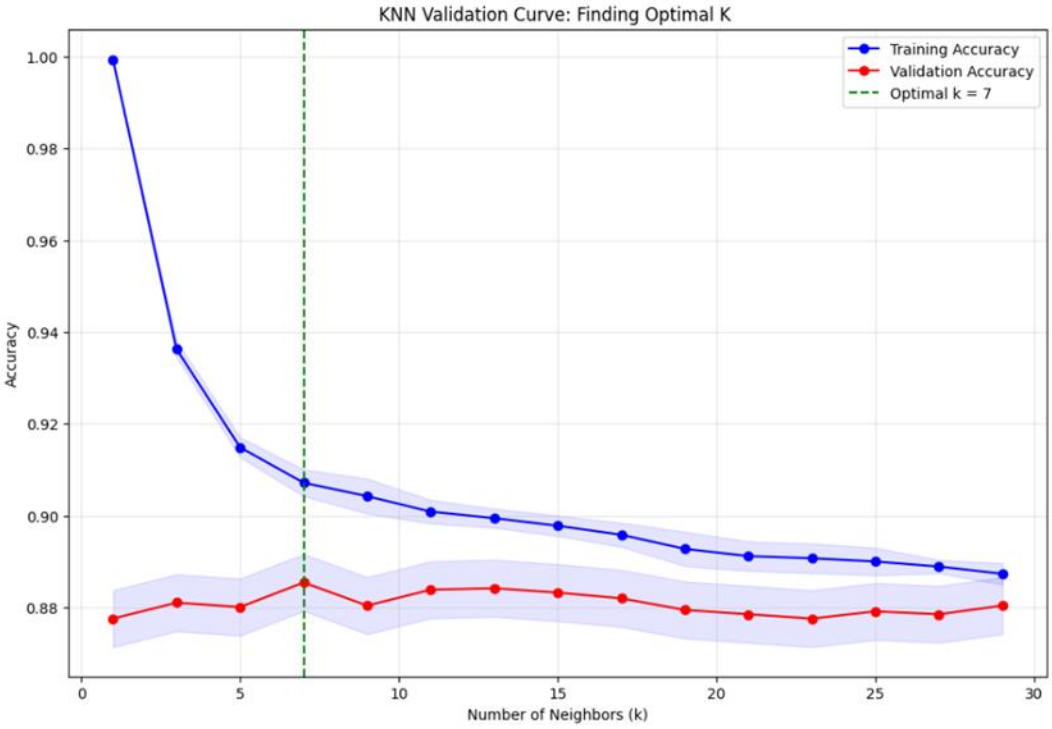

Figure 4: KNN validation curve showing optimal performance at k = 7, where validation accuracy peaks before declining with higher values.

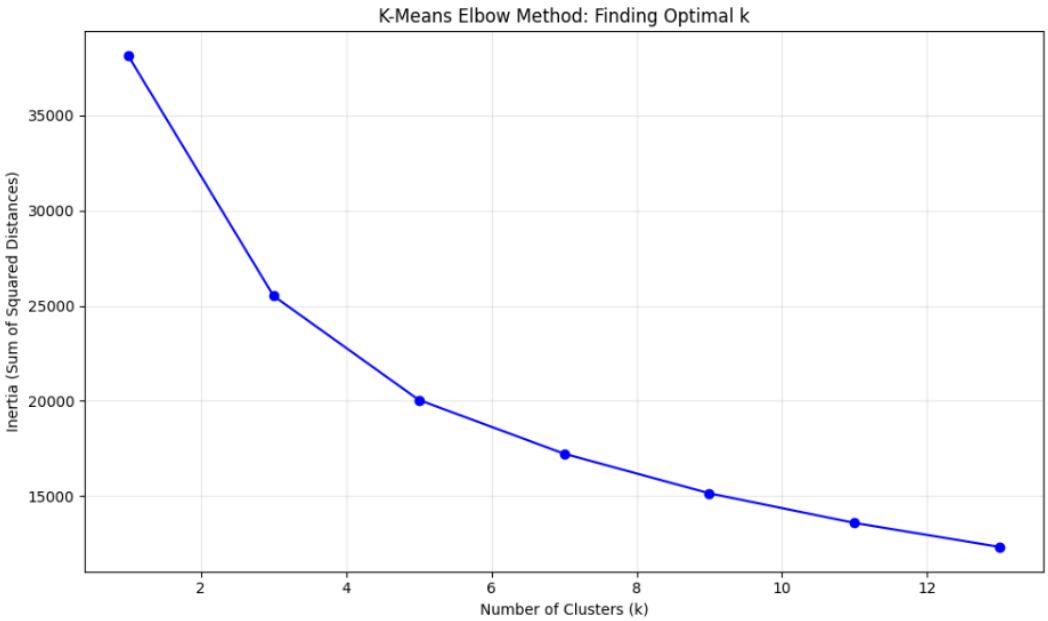

Figure 5: K-Means elbow plot indicating the optimal number of clusters is around k = 3 to 5, where the rate of inertia reduction begins to level off.