DSCI 632
Caleb Miller and Hashim Afzal
February 19, 2025

Final Project Proposal: Regression Model for King County Housing Prices

Team Members

- Caleb Miller
    - Second to last quarter MSDS student currently working as a Data Engineer. Skilled in Python, SQL, web-scraping, and machine learning.
- Hashim Afzal
    - MSDS Student with a background in Biology. Currently applying machine learning concepts to environmental management problems.

Project Objective

The goal of this project is to develop a regression model that accurately predicts house prices in King County, Washington. By analyzing historical housing data, this model aims to provide insights into the key factors that influence property values and assist homebuyers, real estate professionals, and investors in making informed decisions. This predictive tool will help estimate the fair market value of a property based on various characteristics described below.

Data Source

The analysis will utilize a dataset acquired from Kaggle titled “House Sales in King County.” This dataset contains 21,614 records of house sales in King County, Washington, providing a rich source of information for modeling and analysis. The dataset includes the following variables:

- **Transaction Details:** Date of sale, price
- **Property Characteristics:** Number of bedrooms, bathrooms, home size (square footage), lot size, number of floors, square footage above ground, basement size
- **Additional Features:** Waterfront status, presence of a view, home condition, grade (quality of build and interior)
- **Geospatial Data:** Latitude, longitude, zip code, average square footage of the 15 closest neighbors

The dataset can be found here:
https://www.kaggle.com/datasets/harlfoxem/housesalesprediction

Methodology

This project will follow a structured approach consisting of the following phases:

1. **Data Preprocessing:**
    a. Ensure no missing values
    b. Drop unnecessary variables:
        i. ID (unique identifier - does not influence price)
        ii. Latitude (may be redundant when using zip codes)
        iii. Longitude (may be redundant when using zip codes)
        iv. Date (will be transformed into meaningful features)
        v. Year Built & Year Renovated (will be incorporated in feature engineering)
    c. Feature Engineering
        i. Calculate Month of Sale (from date)
        ii. Calculate Year of Sale (from date)
        iii. Calculate Age of House (from year built)
        iv. Calculate Years Since Renovation (from year renovated)
        v. Encode Waterfront
2. **Exploratory Data Analysis:**
    a. Understand the distribution of housing prices and feature variables
        i. Balance classes if necessary
    b. Identify correlations between features and price
    c. Check for multicollinearity
    d. Identify Outliers
    e. Visualize findings to a less technical audience
3. **Model Selection and Development:**
    a. Compare performance of multiple regression models, including:
        i. Linear Regression (Baseline model)
        ii. Decision Tree
        iii. XGBoost (Extreme Gradient Boosting)
        iv. MLP Regressor (Multi-Layer Perceptron Neural Network)
    b. Perform hyperparameter tuning to optimize model performance
4. **Model Interpretation and Insights:**
    a. Identify the most important features influencing housing prices

b. Understand how different variables impact predictions
c. Quantify accuracy through:
    i. Mean-Squared-Error
    ii. Root-Mean-Squared-Error
    iii. Mean-Absolute-Error
    iv. R Squared

Other Considerations

Several external and internal factors may impact the effectiveness of the prediction model:

1. **Market Trends and Inflation:** External economic conditions can impact pricing such as employment rates, mortgage rates, and inflation. The lack of this data may have adverse effects on our modeling.
2. **Geospatial Effects:** Some neighborhoods may have unique factors affecting housing prices (e.g. school districts, public transport availability, crime, etc.). These are not accounted for in this dataset.
3. **Data Limitations:** The dataset may have missing or outdated records, which could impact modeling accuracy.

All coding will be performed in PySpark. The final deliverables will be:

1. Research paper highlighting our findings.
2. Code necessary to replicate this project.