ISEN 613, Project 1: Boston Housing Price Prediction

Mohammadmahdi Ghasemloo (934006883)

1 Dataset Description and Goal

The goal of this project is to develop a predictive model for housing prices using the Boston dataset. By leveraging machine learning techniques, specifically polynomial regression, the study aims to identify key features that influence housing prices and improve prediction accuracy. Feature selection is further implemented to reduce the model complexity and enhance the model accuracy. The overall accuracy of the model using R^2 score was derived about 86% at the end. This suggests that the analysis in this report is supported by the learning model. More complex models such as Artificial Neural Networks (ANNs) can be trained but the company's computational resources are limited to train such models.

2 Primary Numerial Analysis

The boxplots reveal the distribution of various housing attributes in the dataset. The per capita crime rate (CRIM) is highly skewed, with most values near zero but some extreme outliers above 80. The proportion of residential land zoned for large lots (ZN) shows a concentration around zero, with a few higher values reaching 100. The proportion of non-retail business acres per town (INDUS) is distributed between 0 and 27, with a wider spread. The nitrogen oxide concentration (NOX) ranges approximately from 0.3 to 0.8. The average number of rooms per dwelling (RM) varies between 3 and 9, with a median around 6. The proportion of owner-occupied units built before 1940 (AGE) is mostly above 50, extending to 100. The weighted distance to employment centers (DIS) spans from about 1 to 12. The property tax rate (TAX) is widely spread, with values ranging from around 180 to over 700. The pupil-teacher ratio (PTRATIO) falls between 12 and 22. The proportion of Black residents (B) is concentrated near its upper limit, with some lower outliers. The percentage of lower status population (LSTAT) has a range from close to 2 up to about 40. Lastly, the median value of owner-occupied homes (MEDV) varies between 5 and 50, with most values clustering in the lower range. These distributions highlight skewness and the presence of outliers in several features.

3 Feature Analysis

Housing prices are influenced by several key factors, including crime rates, educational quality, industrialization, accessibility, pollution levels, and property taxes. Notably, the interaction between **crime rate (CRIM) and pupil-teacher ratio (PTRATIO)** suggests that neighborhoods with higher crime rates and lower educational quality tend to experience depreciation in property values. Additionally, crime rates independently play a crucial role, as safer neighborhoods are generally

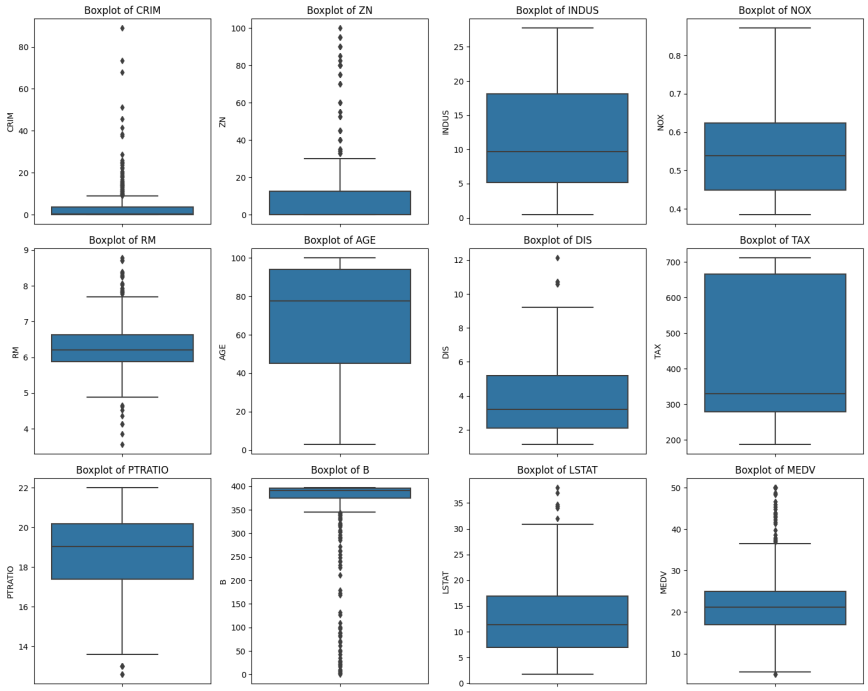

Figure 1: Boxplots of the features. The range of each feature can be seen here.

associated with higher home prices. Industrialization also impacts housing costs, as indicated by the relationship between **industrial land proportion (INDUS) and crime rate**, where highly industrialized areas with elevated crime levels correlate with lower housing demand. Furthermore, **accessibility** is a significant factor, with the interaction between **highway access (RAD) and property tax (TAX)** demonstrating that while well-connected areas are desirable, excessive taxation can diminish affordability. Moreover, **pollution levels (NOX) in conjunction with educational quality (PTRATIO)** contribute to property valuations, as households often prefer cleaner environments with access to high-quality education. In conclusion, the interplay of these variables highlights the complexity of real estate valuation, emphasizing that housing prices are shaped by both individual attributes and their interdependencies.

4 Common Scenarios

We will go over three common scenrios that we have faced recently.

A high-end suburban home is located in a low-crime, family-friendly neighborhood with access to high-quality schools and a favorable student-to-teacher ratio. Its proximity to major highways and business districts enhances commuting convenience. As a newly constructed property, it features modern architecture and energy-efficient systems, ensuring both aesthetic appeal and sustainability. Minimal industrial pollution further contributes to a clean and healthy living environment. Given these attributes, such homes are valued between $450,000 and $600,000. A mid-range urban apartment offers affordability within a city setting. The neighborhood is generally safe, though some security concerns may exist. Schools provide adequate learning opportunities but have a higher student-to-teacher ratio than suburban areas. The apartment is well connected to public transportation, making commuting convenient, though car accessibility may be limited. Constructed within the last two to three decades, it remains a viable option despite some aging, though higher

pollution levels are a drawback. Homes in this category typically range from $250,000 to $400,000. An affordable rural home is situated in a quiet, low-crime area, appealing to those seeking tranquility. However, educational opportunities are more limited, and access to major employment centers requires personal transportation. The home is often an older structure but offers potential for renovation. Its primary advantages include pristine air quality and abundant green space, contributing to a high quality of life. The estimated price range for such homes is between $100,000 and $250,000.

5 Features Affecting Expensive Houses

The most significant features influencing the top 5% of houses in terms of value include the proportion of Black residents (B), the percentage of the lower-status population (LSTAT), crime rate (CRIM), average number of rooms per dwelling (RM), and pupil-teacher ratio (PTRATIO). These factors play a crucial role in determining housing prices, with B and LSTAT showing the highest importance scores. The importance of each feature was calculated by first normalizing the dataset and then determining the absolute mean difference between high-value homes and the overall dataset. This highlights how much a feature deviates in high-priced homes compared to the entire dataset. To ensure reliability, we also considered the variance of each feature in the top 5% of houses. Features with high mean differences but low variance were ranked higher since they consistently differ in expensive properties without fluctuating unpredictably. This measure makes sense because it balances both the magnitude of deviation and the stability of each factor, ensuring that features identified as important are not merely outliers but consistently relevant indicators of high-value properties. The results suggest that socioeconomic and educational factors, along with crime rates and housing characteristics, significantly shape property values.

6 Trade-Offs

In our context, buyers seeking larger homes within the mean house value budget can make strategic trade-offs. One approach is sacrificing proximity to top schools by purchasing a home in an area with average school quality instead of a highly ranked district. This decision can lower home prices by 10-20%, making it possible to afford a more spacious property with additional rooms and a larger yard. Another option is choosing a longer commute to major employment centers, as homes located 30-40 minutes outside the city are typically 25-30% cheaper. This cost reduction allows buyers to purchase a home that is 30-50% larger while still maintaining access to job opportunities. Lastly, buyers can opt for an older home instead of a newly built property, as houses constructed more than 20 years ago often sell for 15-25% less. While such homes may require minor renovations, they provide significantly more square footage within the same budget. These trade-offs illustrate how adjusting location and property age can maximize home size while remaining within financial constraints in the Boston housing market.