PHASE TWO PROJECT REPORT (GROUP 6)

PROJECT OVERVIEW

This project focuses on using multiple linear regression modeling to analyze house sales in a northwestern county.

BUSINESS UNDERSTANDING

The selling price of a home is a critical factor in the real estate market, as it directly influences the financial outcome for both buyers and sellers. For homeowners, the selling price represents the return on their investment and can significantly impact their financial well-being. Potential buyers, on the other hand, rely on the selling price to make informed decisions about purchasing a property within their budget and assessing its value relative to similar properties in the market.

House price is influenced by a multitude of factors, which can be broadly categorized into three main categories: property-specific factors, market factors, and external factors. Property-specific factors encompass attributes such as location, size, condition, amenities, architectural style, and age of the property. Market factors include supply and demand dynamics, interest rates, mortgage availability, and prevailing economic conditions. External factors can range from neighborhood characteristics, such as school quality and crime rates, to broader influences like government policies, infrastructure development, and demographic trends.

Understanding the factors that influence the selling price of residential properties is of paramount importance to various stakeholders involved in the real estate industry. Real estate agents need this knowledge to provide accurate pricing recommendations and effective marketing strategies for their clients. Homeowners can benefit from understanding these factors to make informed decisions when pricing their properties. Investors and developers can leverage this knowledge to identify promising investment opportunities and maximize their returns.

BUSINESS PROBLEM

The real estate market is highly volatile, influenced by economic conditions, housing demand, and external factors. Setting inappropriate prices and making uninformed decisions on when to sell a house can be counterproductive. Research is essential to understand market trends and identify the best time to sell a home to maximize its selling price. Analyzing property characteristics such as location, size, amenities, condition, and recent market trends through research aids in setting an appropriate selling price. By understanding how different home characteristics impact selling prices, the agency can help homeowners mitigate the risk of setting inappropriate prices or making poor investment decisions.

OBJECTIVES

1. To understand the top four factors that influence the prices of a house.
2. To develop a model that can predict housing prices based on various features.
3. To investigate how the important factors affecting price obtained in the first objective vary with the target variable.

DATA UNDERSTANDING

Through meticulous exploration and analysis of this dataset, we aim to glean critical insights into the factors that play a pivotal role in determining the sale prices of homes. By understanding the nuances and correlations within the data, we will be better equipped to provide tailored advice to real estate agencies enabling them to make informed decisions that align with their aspirations and investment goals.

Our dataset has 21,597 home sale records which include a record of the ID, the date a house was sold, the sale price, the number of bedrooms, the number of bathrooms, the square footage of living space in a home, square footage of the lot, number of floors in a house, whether a house is on a waterfront, quality of the view from a house, overall condition of the house, the overall grade of a house, square footage of house apart from the basement, square footage of the basement, the year a house was built, the year when a house was renovated, ZIP Code, latitude coordinate, longitude coordinate, square footage of interior housing living space for the nearest 15 neighbors, and the square footage of the land lots of the nearest 15 neighbors.

The dataset also has numerical variables that represent quantities that can be measured or counted example; the sale price, the number of bedrooms, the number of floors, etc., and categorical variables that represent discrete categories or labels example; the overall grade of a house, the quality of the view, etc.

DATA CLEANING

This is the process of identifying and correcting errors, inconsistencies, and inaccuracies in a dataset to improve its quality and reliability. It involves several key steps which include: handling missing values, removing duplicates, outlier detection and handling, standardizing data formats, dealing with inconsistent labels, etc.

Upon reviewing the dataset, it became evident that the 'sqft_basement' column had been mistakenly categorized as categorical data primarily due to the presence of "?" values, which accounted for approximately 454 instances. To rectify this, we decided to address this issue by removing these questionable entries, as there was no reasonable explanation for their inclusion in the dataset, suggesting that they were erroneous or invalid inputs.

We identify outliers by selecting rows where the 'bedrooms' value is greater than or equal to 11. This filter isolates instances where the number of bedrooms in a house is significantly higher than the typical range found in the dataset. Addressing outliers is essential for data quality and statistical analysis, as extreme values can distort the overall patterns and insights derived from the data. This data cleaning step aimed to ensure the accuracy and reliability of the data, ultimately leading to a more appropriate representation of the 'sqft_basement' column in the analysis or modeling process.
After dropping some missing values we lost about 1.2% of the data and concluded that the loss would not significantly impact the modeling process..

EXPLORATORY DATA ANALYSIS(EDA)

Exploratory Data Analysis (EDA) is a crucial initial step in data analysis. It involves examining and summarizing the main characteristics of a dataset to gain insights, discover patterns, and identify potential problems. It typically includes descriptive statistics, data visualization, and distribution analysis among others.

The analysis was conducted based on two broad categories: the categorical variables and the continuous data. We therefore created a heat map to visually represent the correlations between the numerical features to understand the relationships between the numerical features in our dataset. We also calculated and examined the correlation coefficients between the numerical features and the target variable price to understand which features have the strongest positive or negative correlation with the target variable which is essential for feature selection and model building by creating a heat map to visually represent the correlations between the numerical features.

Scatterplots were also created to explore the relationships between multiple variables. These scatterplots allow for the visual exploration of relationships between price, and other relevant variables in the dataset. We also used the Variance Inflation Factor(VIF) to check for multi-collinearity and found out that the square footage of the living space, the square footage of the house apart from the basement, and the square footage of the basement have extremely high VIF values (infinity), which indicates perfect multi-collinearity. This means that these variables are linear combinations of each other, and they provide the same information. thus we considered removing the square footage of the house apart from the basement and also the square footage of the basement variables to address multi-collinearity.

We also conducted an analysis to understand the relationships between the numerical features in our dataset, especially in relation to the target variable 'price.' we calculated and examined the correlation coefficients between the numerical features and the target variable 'price.' This

allows us to understand which features have the strongest positive or negative correlation with the target variable, which is essential for feature selection and model building.

MODELING AND REGRESSION ANALYSIS

Modeling in data analysis involves creating mathematical or computational representations of real-world phenomena or processes based on observed data. These models aim to capture relationships, patterns, or trends in the data, allowing for predictions, simulations, and insights. Regression analysis is a statistical technique used in data analysis to model the relationship from the data. The goal of regression is to understand how changes can be affected by the relationship between the categories.

After testing out different modeling techniques, we found the best model with a score of 86.5% accuracy as well as a low normalized RSME score of 0.0172 is the XGBoost model. A low RMSE is the best go-to approach when choosing the best model, especially considering where accuracy is of more significance. XGBoost offers better accuracies at fast training speeds and is therefore the ideal model for large data especially where scalability has to be considered. XGBoost is good at handling missing values and also offers feature-importance visuals which are very crucial to extracting deeper insights. When incorporated with the Partial Dependence plots, more meaningful insights are generated which provide very clear information to various stakeholders just like in our case, the King House county. On top of that, it comes with parallel computing which is a very important factor where computational power is of significance and consideration. This model also has the feature of hyper-parameter tuning in the event the accuracy score has to be improved. Tuning specifies the depth of learning trees as well as the learning rate.

FEATURE IMPORTANCE

Feature importance refers to the assessment of the impact or contribution of individual features (also known as variables or attributes) in a dataset towards achieving a specific outcome or prediction. It helps identify which features have the most influence on the target variable and provides insights into the underlying relationships within the data.

To answer our business objectives, the feature importance plots offer nice visuals with easy interpretations. After assessing the features in our dataset, we found out that the grades, waterfronts, the square foot of the living space, and ages of the houses are the important features driving the price of the houses.

We therefore plotted partial dependence plots to give us more insights on how feature values impact the prediction scores. From the partial dependence plots, we realized that Increasing the living space from 650 square feet to 700 square feet attracts higher prices but it should not exceed 700 because after that the prices start declining. We also found out that increasing the square foot basement from around 75-80 has better prices, but any increase after this fetches lower prices. Houses with more than 7 bedrooms fetch higher prices compared to those with 6 bedrooms. Increasing the number of bathrooms from 1 to 12 has a steady increase for the houses and houses with more than 20 bathrooms fetch very high prices and this may involve mansions and more luxurious houses. Finally, we also realized that houses with less than four floors fetch lower prices compared to those with more than four floors.

CONCLUSION

From the analysis results, we came to the conclusion that there are factors driving house prices in the King County region. These factors include the location of houses since the most low-priced houses seem to be in the southern direction of the King County region, the square feet of the living space, the house grades, the age of the house, and the number of bathrooms in a house. We also concluded that Increasing the living space from 650 square feet to 700 square feet attracts higher prices but it should not exceed 700 because after that the prices start declining.

RECOMMENDATIONS

Based on our comprehensive analysis of the factors influencing house prices in the King County region, we recommend improving house grades right from construction with better designs and investing in high-end neighborhoods to realize a good return on investment. For future proposals, the provision of more data is highly recommended to gain more accuracy and insights.