Data Mining Project Report

BDS-6A

Group members:
Misbah Munir
Mehak Fatima
Minahil Mobin

Introduction:

In the field of real estate analysis, it is critical to comprehend the complex dynamics affecting house prices. We undertook a thorough investigation by utilizing a dataset that included a wide range of property variables, from size and amenities to location details. By using painstaking preprocessing methods and four different regression approaches, we attempted to reveal the underlying trends that drive property values. Furthermore, by adding a new characteristic to our dataset called "price category," we were able to explore the field of classification and improve our comprehension of housing market segmentation. Come along with us as we explore this fascinating environment and uncover knowledge that is essential for making wise decisions in the fast-paced real estate market.

Motivation:

Our main goal was to employ "price category" as the target variable in classification models and "price" in regression models. Our goal with regression analysis was to precisely estimate residential properties' market value by considering a variety of characteristics such size, location, amenities, and condition. Regression analysis was used to determine the intricate correlations between these characteristics and house prices, which would help stakeholders make well-informed decisions about sales, property investments, and market trends. Our goal using categorization models, on the other hand, was to divide the housing market into several groups according to price ranges. This methodology made it easier to classify properties into discrete groups or sectors, which in turn allowed for more focused marketing campaigns, well-informed investment choices, and thorough market research.

Data set Description:

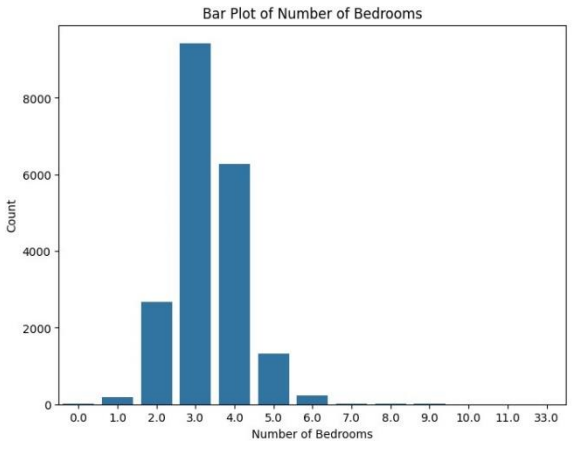
A comprehensive source of property data, the home price dataset provides a wide range of attributes essential for in-depth analysis and machine learning applications. It offers a comprehensive view of the homes that are offered, from fundamental elements like bedrooms and baths to specific metrics like living area size and exact geographical information. Additionally, the dataset makes it easier to spot trends and patterns in the housing market.

Preprocessing Graphs:

Scatter Plot:

This scatter plot shows that living space square footage and property prices are positively correlated, meaning that larger homes often fetch higher prices. Fewer homes at higher price points and square footages are scattered across the data, which is densely packed at lower values. Variability in costs points to additional important elements, such amenities and location. Outliers might be rare or opulent houses, especially in high-priced and square footage locations. In general, the graph helps investors, purchasers, and real estate experts comprehend the connection between living space and property value within a certain market or dataset.

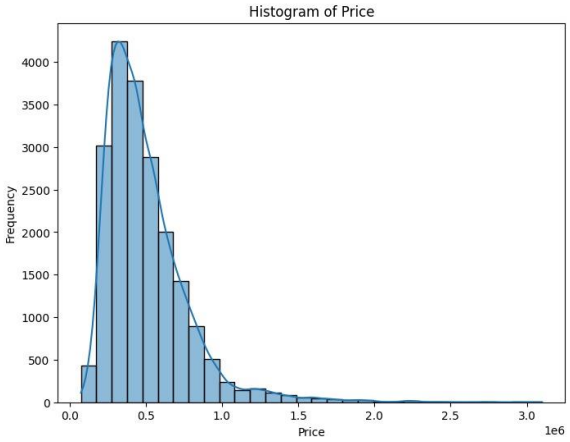
Bar Plot:



The distribution of bedrooms among homes is shown by the bar plot. Three bedrooms is the most typical arrangement, followed by two and four

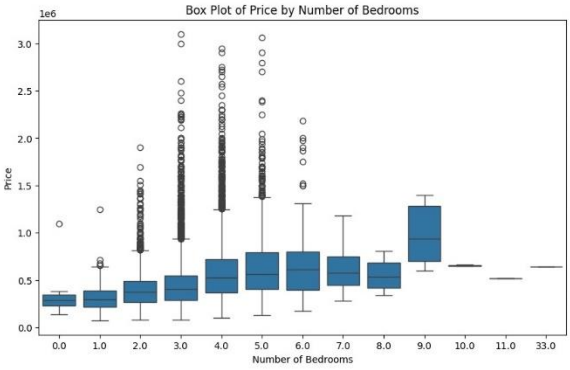
bedrooms. Fewer than five-bedroom homes exist, and the number significantly decreases after four bedrooms. The "33" bedroom bar probably indicates data aggregation issues or outliers.

Histogram:



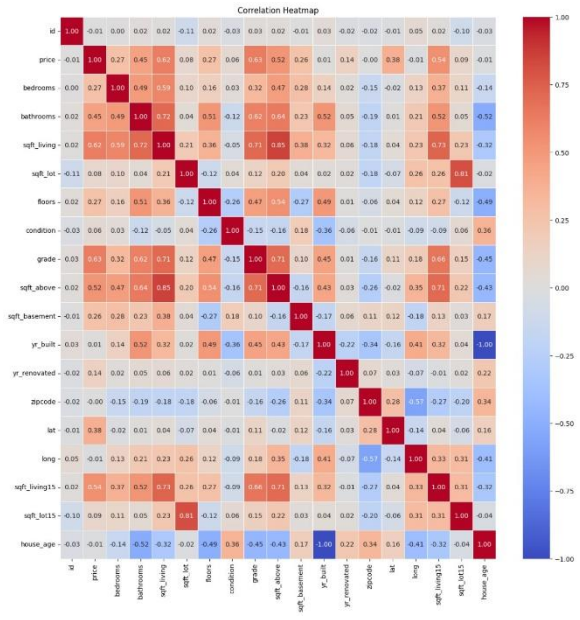
The distribution of prices is depicted by this histogram, which demonstrates that the majority of the data is centered around lower price values and taper off as prices rise. The price range that is most frequently seen is shown by the peak frequency, which falls between 0 and 0.5 million.

Box Plot:



This picture shows a box plot graph that shows how real estate values vary according to how many bedrooms a property has. Plotting the median, quartiles, and outliers for various bedroom counts reveals a general pattern of rising costs as the number of bedrooms increases.

Correlation Heatmap:



This picture shows a correlation heatmap, which uses varied color intensities based C correlation values to visually portray the relationships between various variables in a dataset. Stronger, weaker, and no linear connection between the variables are shown, accordingly, by red tones for positive correlations, blue tones for negative correlations, and white tones for values around zero.

Regression analysis:

Linear regression:

The goal of linear regression is to build a linear relationship between independent and dependent variables to provide a basic and understandable model for predicting numerical outcomes. Its simplicity and interpretability make it extensively applicable in a variety of sectors.

Insights:

Although the model performs quite well, the lack of an intercept component in the optimal model configuration raises the possibility of underfitting. This implies that biased predictions could result from the model's incomplete representation of the connection between the features and the target variable. This problem may require more research and testing to be resolved and possibly increase the predicted accuracy of the model.

Ridge Regression:

The goal of slope regression is to reduce the issues of multicollinearity and overfitting that are frequently present in linear regression models. This is accomplished by adding a regularization term, commonly referred to as the L2 penalty, which incentivizes the use of simpler models with smaller coefficients by penalizing big coefficients.

The distribution of prices is depicted by this histogram, which demonstrates that the majority of the data is centered around lower price values and taper off as prices rise. The price range that is most frequently seen is shown by the peak frequency, which falls between 0 and 0.5 million.

Insights:

In comparison to basic linear regression, ridge regression performs better thanks to the addition of regularization. The regularization term results in a more generalized model by penalizing big coefficients, which helps prevent overfitting. Ridge regression appears to efficiently address overfitting, resulting in more dependable predictions, as seen by the decrease in mean square error (MSE) on both test and cross-validation sets. More testing with various alpha values could be necessary to optimize the model's functionality.

Lasso Regression:

Like Ridge regression, Lasso regression uses regularization to deal with overfitting and multicollinearity. On the other hand, it makes use of an L1 penalty, which effectively does feature selection by encouraging the values of sparse coefficients by decreasing part of the coefficients to zero.

Insights:

The comparable MSE values suggest that the Lasso regression model performs similarly to simple linear regression. This result implies that the model's predictive ability on this specific dataset may not have been greatly impacted by the L1 penalty. To possibly improve the model's performance, more research or tweaking of hyperparameters, like the regularization strength (alpha), may be necessary. Furthermore, examining the possible causes of the L1 penalty's minimal influence on feature selection might reveal details about the properties of the dataset.

KNN Regressor:

Regression using K-nearest neighbors (KNN) is a non-parametric technique used to forecast numerical results. The method works by averaging, or weighing, the target values of the K data points that are closest to the query point.

Insights:

The higher MSE values suggest that the KNN regressor performs worse than the linear and ridge regression models. This implies that additional fine-tuning of the KNN model may be necessary to increase its forecast accuracy or that it may not be a good fit for this specific dataset. Changing the number of neighbors, investigating various distance measurements, or taking into account alternate weighting techniques are some possible research topics. Further examination of the features of the dataset and any possible outliers may further shed light on the model's performance constraints.

Classification Models:

Multinomial Logistic Regression:

One well-liked approach for binary classification problems is logistic regression. When the target variable has more than two classes, it can also be expanded to handle multiclass classification issues. Here, we employ Multinomial Logistic Regression, which makes use of the softmax function to expand binary logistic regression to several classes.

Insights:

With an accuracy of about 55%, the Multinomial Logistic Regression model demonstrates its capacity to categorize cases into several classes. The classification report, however, shows that the model's performance differs depending on the class, with the 'high' and 'low' categories showing better precision and recall than the 'medium' category. This points to the possibility of a class imbalance or the necessity of additional feature engineering to boost the performance of the model. Furthermore, there may be opportunity for improvement in the prediction of the continuous target variable, perhaps by feature selection or regularization adjustment, as indicated by the relatively high MSE for Lasso Regression. This needs to be corrected.

Random Forest:

Several decision trees are constructed using the Random Forest ensemble learning technique, which then combines the predictions of the trees to increase accuracy and decrease overfitting. It is frequently applied to classification tasks in which discrete class labels make up the target variable.

Insights:

The great performance of the Random Forest Classifier is further validated by the grid search with cross-validation, as seen by the high mean accuracy score attained during cross-validation. The chosen hyperparameters imply that the best model performance is achieved by letting trees grow without maximum depth restrictions, establishing minimum samples per leaf, and splitting the data to low values.

Gradient Boosting:

Gradient Boosting Classifier is an ensemble learning technique that creates a series of decision trees one after the other, correcting each other's mistakes. It works especially well for classification tasks where the goal is to combine the predictions of several weak learners to increase predictive accuracy.

Insights:

The grid search with cross-validation, with the chosen hyperparameters optimizing model accuracy, further demonstrates the Gradient Boosting Classifier's outstanding performance.
The resilience and dependability of the model are demonstrated by the high mean cross-validation accuracy score, which shows that the model generalizes effectively to new data.
Overall, the grid search-tuned model and the default Gradient Boosting Classifier both perform exceptionally well, demonstrating the usefulness of gradient boosting for this classification job.

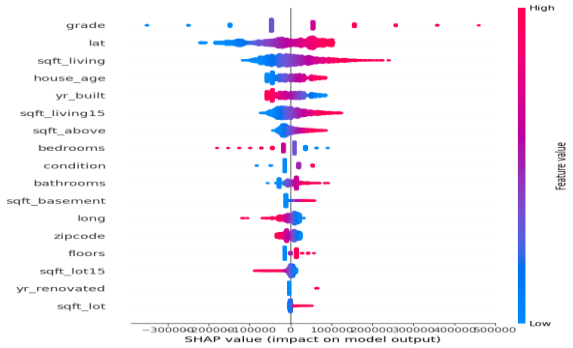
XGBoost Classifier:

The gradient boosting technique known as XGBoost (Extreme Gradient Boosting) has become well-known due to its quickness and effectiveness in machine learning contests. It is renowned for being scalable, effectively managing big datasets, and producing innovative outcomes for a variety of applications, including classification.

Insights:

Performance is enhanced by fine-tuning the model hyperparameters using the grid search with cross-validation.
The model's resilience and reliability are demonstrated by the high mean cross-validation accuracy score and test set accuracy, which show how effectively the model generalizes to new data.
Although there are some small differences between the confusion matrix and the classification report, overall performance is still very good, with XGBoost successfully categorizing examples into the appropriate classes.

Overall, exceptional performance is demonstrated by the XGBoost Classifier, highlighting the usefulness of XGBoost for this classification problem.



The feature's effect on the model's output is indicated by its position on the x-axis. Points on the right indicate a greater influence on raising the prediction, whilst points on the left indicate a greater influence on lowering the prediction.
The feature's value is represented by the color; red denotes high values and blue low values.
The y-axis lists the features, which are normally arranged according to the total of the SHAP value magnitudes for all samples. In this specific model, the property at the top (grade) is the most significant, and the feature at the bottom (sqft_lot) is the least significant.The distribution of the effects is shown by the width of the "cloud" of points for each feature. A larger dispersion indicates a more unpredictable influence of the characteristic on the model's predictions.
The most significant factor in this particular figure is "grade," with greater values (red) typically pushing the model's forecasts higher. The second most important factor is "latitude," with higher latitudes

significantly increasing predictions. However, the features'sqft_lot' and 'yr_renovated' appear to have a less significant and more inconsistent effect on the model's predictions.

Conclusion:

Regression and classification models were employed in our comprehensive real estate study to investigate the complex relationship between property qualities and prices. We carefully preprocessed and feature-engineered the dataset to make it ready for modeling, taking into account attributes like condition, size, location, and amenities. Although linear regression produced a basic model, multicollinearity and overfitting problems were addressed via ridge and Lasso regression, which produced predictions that were more accurate. The KNN regressor performed worse, indicating that more optimisation is required. Multinomial logistic regression performed well in classifying cases; however, feature engineering could yield better results. XGBoost, gradient boosting, and random forest classifiers all performed remarkably well, with XGBoost particularly excelling in accurately classifying attributes. Given the circumstances, our analysis provides real estate stakeholders with insightful information to help them make educated decisions and identify potential future directions.

Future Directions:

Advanced Modelling Techniques: To further enhance forecast performance and capture intricate linkages in the data, investigate more complicated modelling techniques like neural networks or ensemble methods.

Refine feature selection procedures to find the most important features and boost the effectiveness and interpretability of the model.

- ***Add More Data:***
  To improve the study and obtain a more thorough grasp of the housing market, add extra data sources such as demographic data, economic indicators, or neighbourhood features.
- ***Time-Series Analysis:***
  To improve forecasting and long-term planning, use time-series analysis to look at trends and patterns in real estate values over time.

Perform geographic analysis to pinpoint emerging markets or profitable investment prospects, as well as to comprehend spatial variances in real estate pricing.

We can improve our comprehension of the real estate market and give stakeholders useful information for making strategic decisions by going in these future directions.