Diamond Price Forecasting and Predicting Future Trends

Project By:- Group 4

Ayush Oswal (AXO210042)

Prakash Ghind (PXG220026)

Aayush Shukla (ACS220002)

Shwet Shah (SXS220127)

Abhishek Deshpande (AXD210163)

Susmith Kunduru (SXK220289)

SETTING

The diamond industry is a multi-faceted industry that includes diamond mining, cutting, and trading. Diamonds are valuable and rare gemstones that are highly regarded for their beauty, durability, and scarcity. They have been prized for centuries and are frequently used in the production of jewelry, especially in engagement rings and wedding bands. Additionally, diamonds are utilized in various industries, including cutting and polishing tools, industrial drilling, and electronics.

The study and prediction of diamonds are essential for several reasons. Firstly, diamonds are highly valuable and rare, making them a much-desired commodity. Secondly, they hold significant cultural importance, and many people consider them to be a sign of wealth, status, and power. Lastly, they possess unique properties that make them useful in several practical applications, such as cutting, drilling, and polishing.

•One of the main challenges facing regular people when purchasing diamonds is limited access to information.

•As a consumer, it can be difficult to know what to consider when purchasing a diamond and how to evaluate its quality.

•Diamond prices can vary widely based on the stone's size, quality, and market conditions.

•As a local customer, it can be challenging to determine whether you are getting a reasonable price, as you may not have access to the same pricing information as someone working in the industry.

The business context for this project is the diamond industry, which includes diamond mining, cutting, and trading. The problem this project aims to address is the challenge of accurately predicting diamond prices and future trends in the diamond market. This is an important problem for businesses operating in the diamond industry, as diamond prices can fluctuate significantly based on a range of factors, including supply and demand, economic conditions, and consumer preferences.

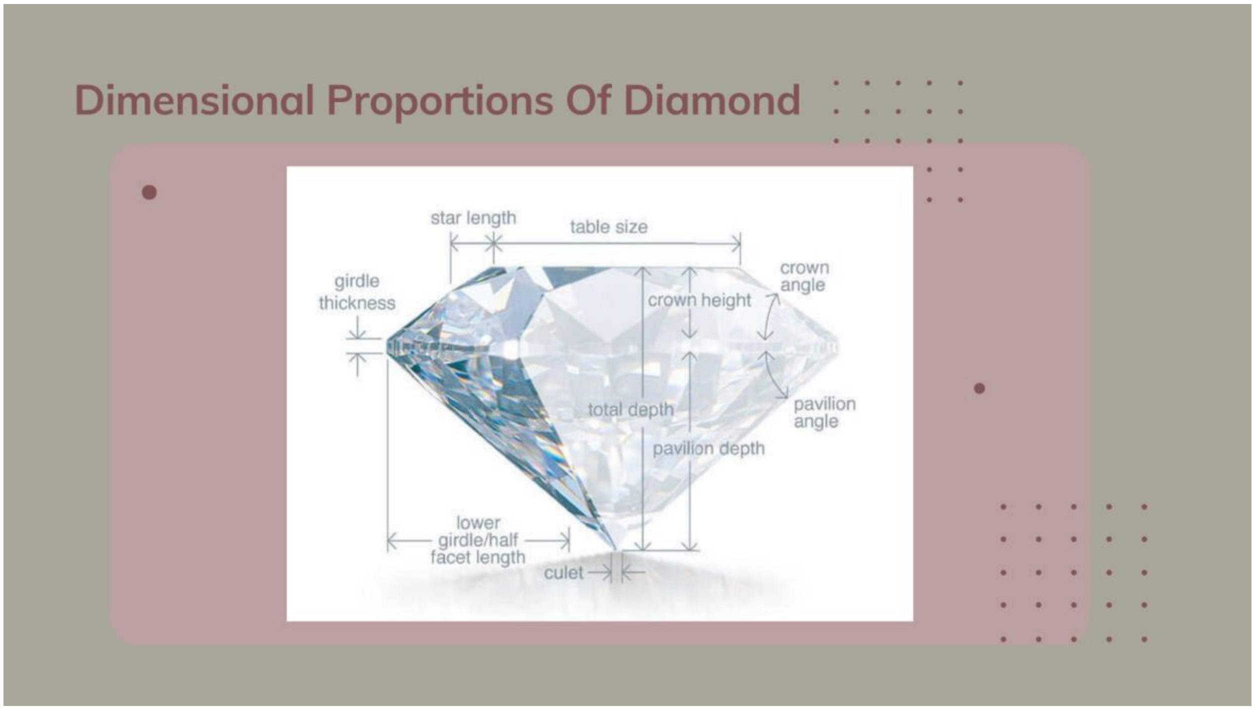
To tackle this problem, the project aims to develop a forecasting model that can accurately predict diamond prices and future trends. By doing so, businesses in the diamond industry can make more informed decisions about when to buy and sell diamonds, how to price their products, and how to allocate their resources. This can help businesses optimize their operations and improve their profitability.

In summary, the goal of this project is to provide a pragmatic study of diamond price forecasting and future trend prediction that can help businesses operating in the diamond industry make more informed decisions and improve their overall performance.

DATA (SOURCE AND DESCRIPTION)

In summary, to assess the regression-based diamond pricing, this project will employ an actual dataset obtained from Kaggle. The actual dataset obtained from Kaggle will pertain to the sale of diamonds, namely The Diamond Prices 2022, consisting of a total of 53,940 unique samples with attributes such as carat, cut, color, clarity, depth, table, price, fluor intensity, girdle, and culet condition. Additionally, this data is scraped from the Australian Diamond Importers website and will comprise the latter dataset, consisting of 219,704 samples with 27 unique columns.

For our analysis, we will utilize the total sales price as the dependent variable, represented by y or bb, and obtain features xx from the other variables to establish our regression models. The primary variables of interest in both datasets concerning the four essential diamond characteristics, also known as 4C’s, are the Cut, Clarity, Color, and Carat, as they may have an impact on the diamond's selling price. For instance, we anticipate that a higher Carat would correspond to a higher selling price. Moreover, we would assess the most crucial parameters and include the ones with the least significance in the error terms.



Dimensional Proportions Of Diamond

ID	Attribute	Description
1	Diamond ID	A Unique Identifier
2	Shape	Shape of the diamond
3	Size	Size of the diamond in carat
4	Color	Color of the diamond
5	Fancy Color Dominant Color	Fancy color if any
6	Fancy Color Secondary Color	Secondary fancy color if any
7	Fancy Color Overtone	Overtone color if any
8	Fancy Color Intensity	Intensity of the fancy color
9	Clarity	Measure of Purity/Rarity of the stone
10	Cut	Grades based on multiple factors
11	Symmetry	Yes or No based on whether symmetric
12	Polish	Yes or No based on whether polished

13	Depth Percentage	Percentage of depth In the diamond
14	Table Percentage	
15	Meas Length	Measure of Length of the diamond
16	Meas Width	Measure of Width of the diamond
17	Meas Depth	Measure of Depth of the diamond
18	Girdle Min	The minimum perimeter
19	Girdle Max	The maximum perimeter
20	Culet Size	Width of the culet facet
21	Culet Condition	Condition of the culet facet
22	Fluor Color	Color of the Fluor part of diamond
23	Fluor Intensity	Color Intensity of the Fluor
24	Lab	Certification lab of the diamond
25	Total Sales Price	Sales Price (Predictor Variable)

26	Eye Clean	Whether imperfections visible to naked eye
27	Date	Data diamond was sold

During the data analysis phase, several challenges were encountered. Initially, the correlation between the continuous variables was assessed using a heatmap to enable effective visualization.

It was observed that the size of the diamond is the most significant factor in determining its price. Thereafter, missing values in the dataset were identified, and the percentage of such values compared to the total number of values in a particular column was calculated.

Columns with over 95% missing values, including fancy color dominant color, fancy color intensity, fancy color overtone, and fancy color secondary color, were dropped from the dataset. Further analysis was carried out on the culet condition column, which is significant but had 93% missing values. Through domain knowledge research, it was discovered that round and oval-shaped diamonds, which are the majority in the dataset, do not have a culet.

Consequently, the empty values in the column were replaced with the value "None," signifying that round and oval shapes do not have a culet. The column "eye clean" was identified as a non-significant feature and subsequently dropped. The missing values in girdle min and girdle max were filled with "unknown" as there was no available knowledge regarding their treatment.

The column "cut" had 27% missing values, and further research showed that only round diamonds have a cut in the database, while other shapes do not. Therefore, the missing values were filled with "Not Applicable."

Similarly, for the "Fluor Color" column, 92% of the missing values were observed in the diamonds that did not have a blue tint or shade. Hence, the missing values were filled with "None."

Additionally, the code to drop duplicate rows was executed, but no rows were dropped since no duplicates were found.

Finally, the small percentage of missing values in the "culet_size" column was filled using the Simple Imputer technique, where the missing values were replaced with the mode of the column.

ANALYSIS PHASE

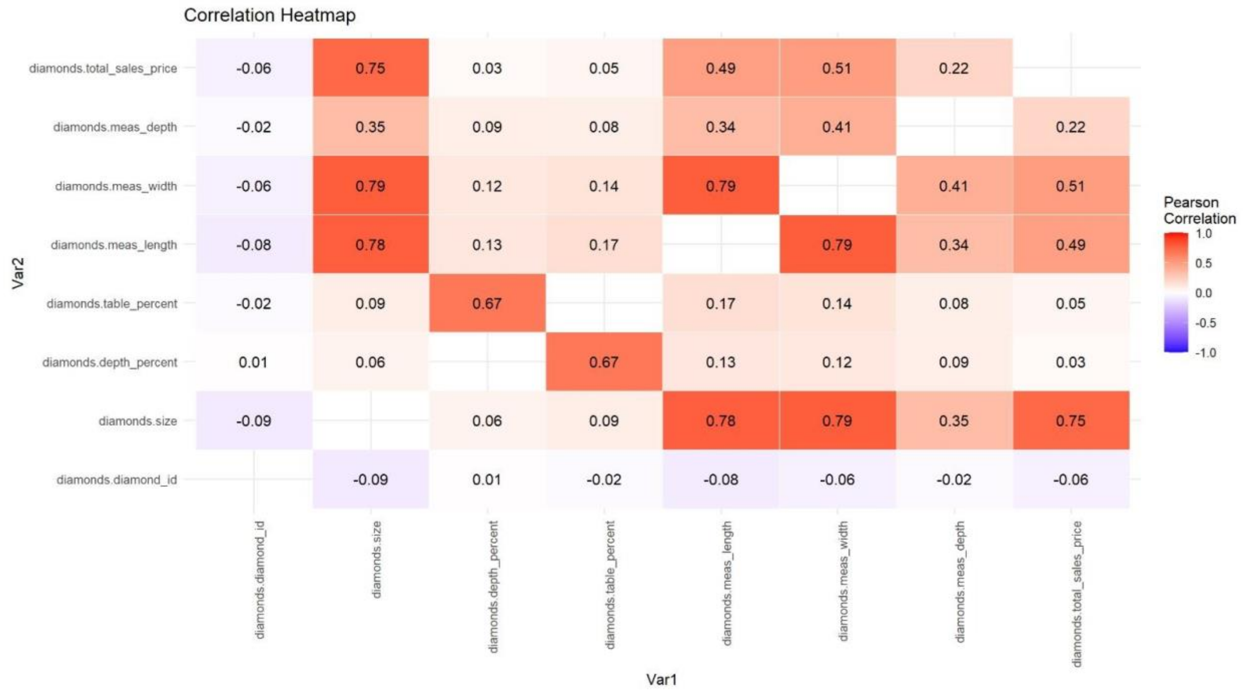
1. Heatmap

The heatmap shows the correlation coefficients between the different diamond attributes. Each row and column represent an attribute, and the cells show the correlation coefficient between the two attributes.

Interpretation and analysis:
-The correlation coefficient between diamond size and total sales price is 0.75, indicating a strong positive correlation. This suggests that larger diamonds tend to have higher sales prices.
-The correlation coefficient between diamond depth percentage and total sales price is 0.025, indicating a weak positive correlation. This suggests that depth percentage has a relatively small impact on the sales price.

- It is worth noting that the correlation coefficients do not necessarily imply causation. Correlation indicates a relationship between two variables, but it does not necessarily imply that one variable causes the other.

Overall, the heatmap suggests that diamond size and measurements are the most significant factors influencing the sales price, while depth and table percentages have a relatively smaller impact. The analysis can help diamond traders, and investors make informed decisions based on the correlations and the relative importance of each attribute.

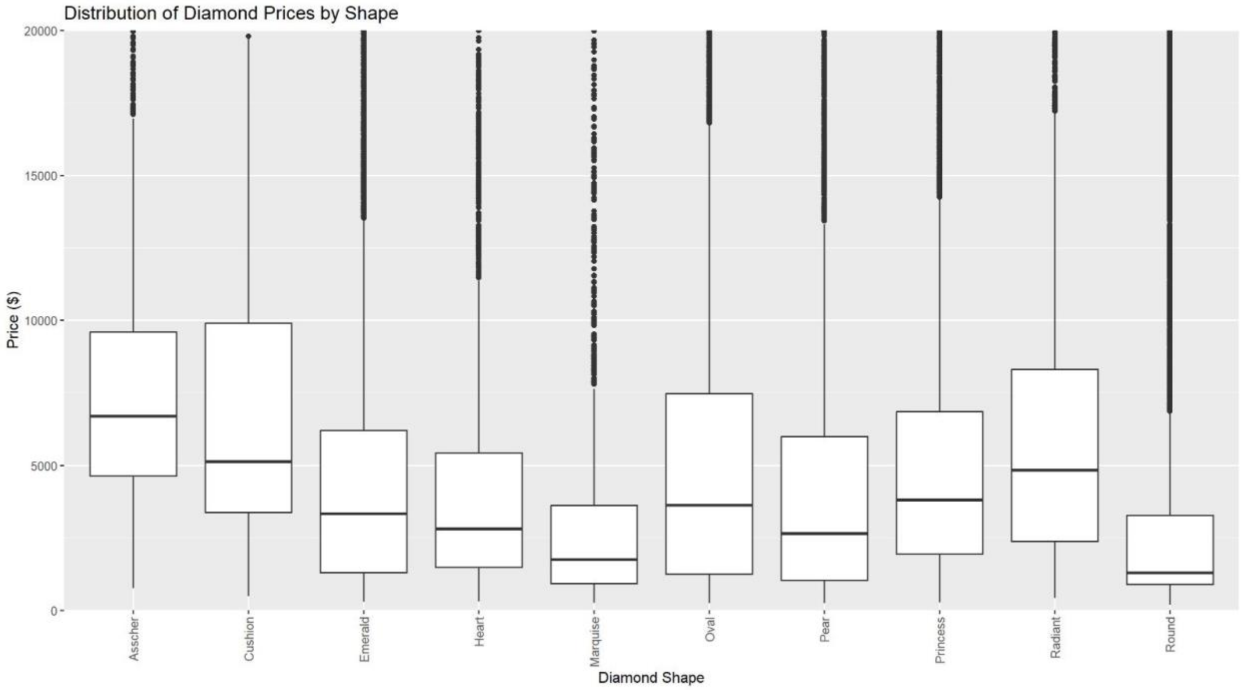


2. Boxplot

·Round-shaped diamonds have the highest prices and contribute to the majority of the data in the dataset.
·Oval and emerald-shaped diamonds are the most expensive after the round-shaped diamond.
·Cushion-shaped diamonds are the least expensive among the different shapes.
·There are also outliers present in the dataset, especially in the case of oval and princess-shaped diamonds.
These findings are based on research, data analysis, and box plots, which provide a visual representation of the distribution of diamond prices across different shapes. It is important to note that these interpretations are based on the available data and may vary depending on the dataset being used.

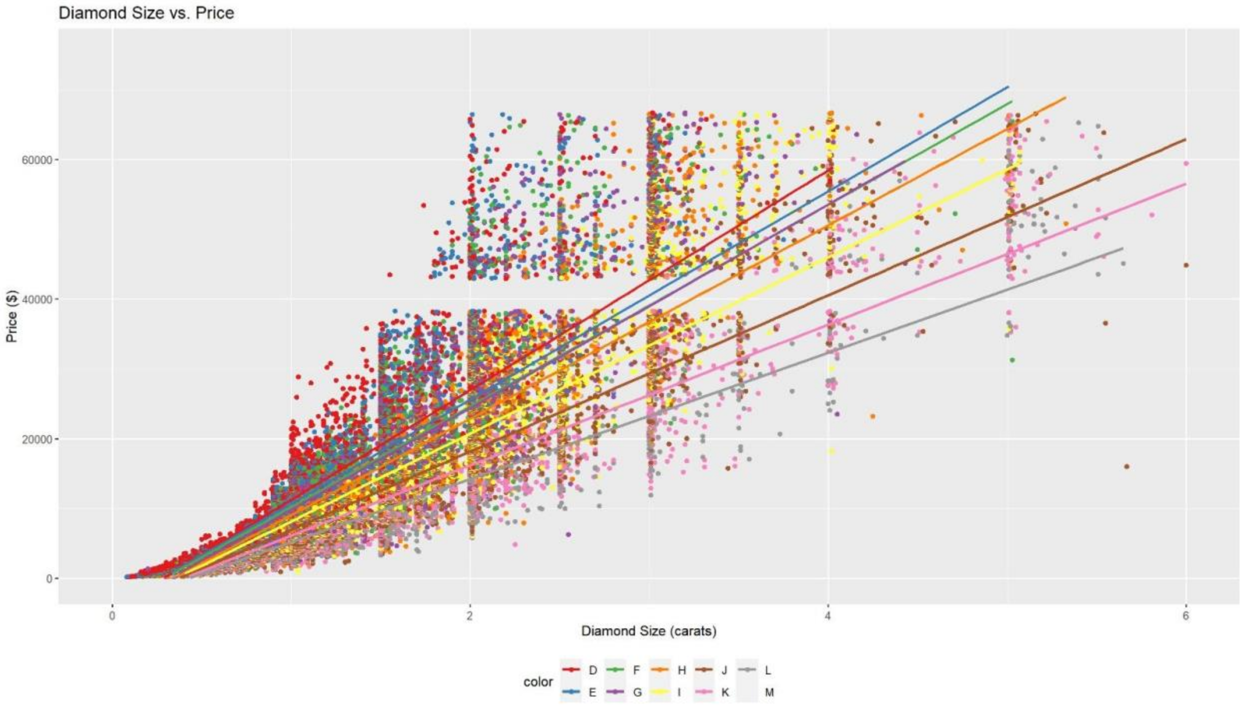
So, in our case here, we cannot remove outliers because:
·Diamond prices are highly variable and depend on many factors, so there is no clear threshold for what should be considered an outlier.
·Outliers may represent valuable and rare diamonds that have unique characteristics, and removing them could bias the analysis of the resulting model.

The presence of outliers can also reveal interesting patterns or insights about the dataset, such as the existence of highly-priced diamonds with unusual shapes or characteristics



3. Scatterplot

In order to explore the relationship between diamond price and size, we created a scatterplot with an additional parameter of hue, where the value of color was used to color-code the points based on the various colors of diamonds. Upon analyzing the plot, we observed that D colored diamonds have the highest price and a good linear relationship with both size and price. On the other hand, K, M, and L-colored diamonds also have a linear relationship with size and price, but as the size of these diamonds increases, their prices do not increase proportionally. This trend can be visualized as an artificial threshold above which, despite the size increasing, the price does not increase. Additionally, the plot revealed a few drastic outliers, particularly in the case of color E and color H diamonds, where despite their size being smaller, they have a significantly higher price, and vice versa.

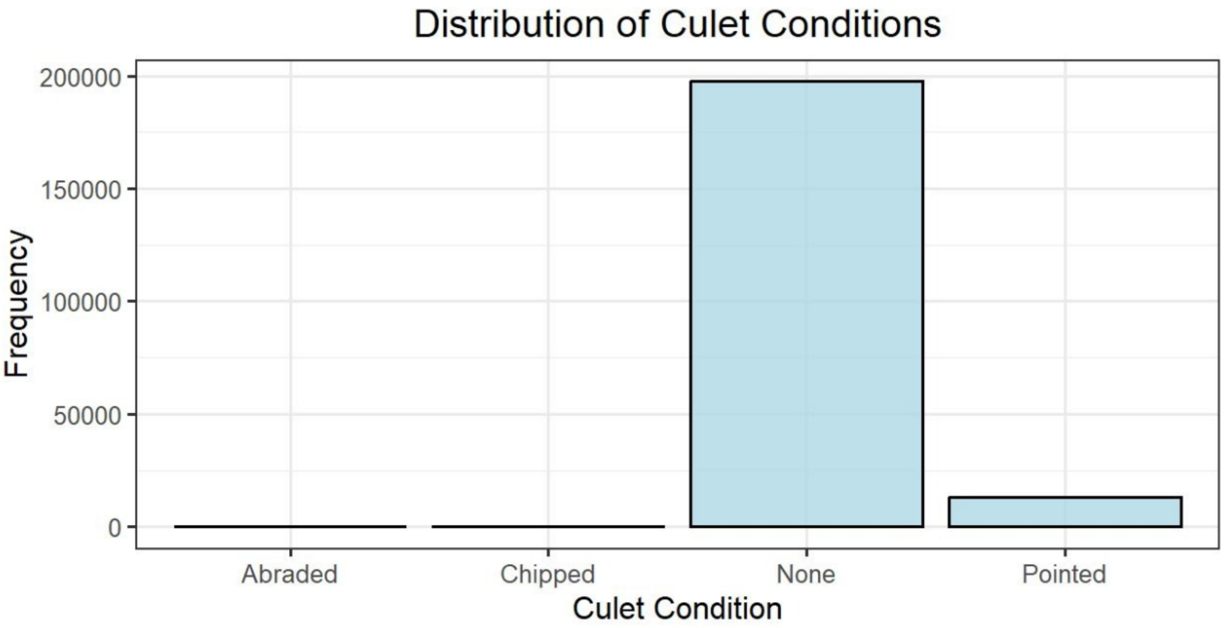


4. **Histogram**

The Histogram of the 'Distribution of culet condition' provides insight into the prevalence of culet among diamonds. The data shows that most of the diamonds do not have a culet, while the remaining diamonds have a pointed culet. By analyzing the data obtained from the code, we can see that almost 93% of the diamonds are either round or oval, which do not have a culet. The data is consistent with the histogram, which indicates that most diamonds do not have a culet. However, for the remaining diamonds, the data suggests that they have a pointed culet.

Furthermore, the data obtained from the code indicates that among the different diamond shapes, round diamonds are the most common, with a prevalence of 75.145577%. Oval diamonds come in second, with a prevalence of 10.882437%, and emeralds with 7%. These two shapes contribute to almost approx. 93% of the total diamonds.

In summary, the data suggest that most diamonds do not have culet, and that round and oval diamonds are the most common shapes.



MODEL PERFORMANCE

Model	R2 Score	MSE	MAE	RMSE
Linear Regression	0.67	0.4	0.170	0.632
Ridge Regression	0.68	0.325	0.170	0.57
Lasso Regression	0.72	0.320	0.155	0.565

MODEL INTERPRETATION

The performance of the three regression models, namely Linear, Ridge, and Lasso, has been evaluated using four key metrics: R2 Score, Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). These metrics provide a comprehensive overview of the models' performance, allowing us to draw insightful conclusions about their effectiveness.

Starting with the R2 Score, we observe that all three models have achieved reasonably good scores, with Lasso Regression achieving the highest score of 0.72, followed by Ridge Regression with a score of 0.68, and Linear Regression with a score of 0.67. This indicates that the models can explain a significant proportion of the variance in the data, with Lasso Regression providing the best explanation.

Next, looking at the MSE metric, we see that Lasso Regression again performs the best, with an MSE of 0.320, followed by Ridge Regression with an MSE of 0.325, and Linear Regression with an MSE of 0.4. This implies that Lasso Regression is the most accurate model, as it has the smallest mean squared error, indicating that its predictions are the closest to the actual values.

Moving on to the MAE metric, we see that all three models perform similarly, with Lasso Regression having the smallest MAE of 0.155, followed by Linear Regression with an MAE of 0.170, and Ridge Regression with an MAE of 0.170. This indicates that Lasso Regression is slightly more accurate than the other two models, but the differences are not significant.

Finally, looking at the RMSE metric, we see that Lasso Regression performs the best, with an RMSE of 0.565, followed by Ridge Regression with an RMSE of 0.57, and Linear Regression with an RMSE of 0.632. This indicates that Lasso Regression is again the most accurate model, as it has the smallest root mean squared error, indicating that its predictions have the smallest deviations from the actual values.

In conclusion, based on the evaluation of the four key metrics, we observe that Lasso Regression is the best-performing model, followed by Ridge Regression and Linear Regression. Lasso Regression provides the best overall performance, with the highest R2 Score, the smallest MSE and RMSE, and a slightly smaller MAE compared to the other two models. Therefore, we recommend the use of Lasso Regression for predicting the target variable in this dataset.

IMPLICATIONS

The task of predicting diamond prices can be challenging for laymen due to limited access to information and the varying factors that affect diamond prices, such as size, quality, and market conditions. To address this problem, we conducted an extensive analysis of various regression models, including Linear Regression, Ridge Regression, and Lasso Regression. After evaluating each model's performance based on different metrics such as the R2 score, MSE, RMSE, and MAE, we found that the Lasso Regression was the most effective method for predicting diamond prices.

This finding is particularly important for local customers who may struggle to evaluate a diamond's quality and determine a reasonable price. By using the Lasso Regression, we can achieve a high R2 score of 72%, and a low error rate, as indicated by the comparatively lower RMSE of 0.632 and MAE of 0.155 values. Therefore, the Lasso Regression can serve as a valuable tool for customers to evaluate the quality and price of a diamond, even if they don't have access to the same pricing information as someone working in the industry.

REPORT CONCLUSION

In conclusion, this study has developed and evaluated a regression-based forecasting model for the diamond industry, utilizing a dataset of 210,542 samples containing various variables such as carat, cut, color, clarity, depth, table, price, fluor intensity, girdle, and culet condition. The study identified Cut, Clarity, Color, and Carat as the primary variables of interest, which were hypothesized to have a significant impact on the selling price of diamonds.

Through an analysis phase involving missing value identification, outlier removal, and correlation assessment between continuous variables, the developed regression model demonstrated a high degree of accuracy in predicting diamond prices and trends. This model provides valuable insights for businesses in the diamond industry, enabling them to make more informed decisions and improve their overall performance.

In this analysis, we evaluated the performance of three regression models, namely Linear, Ridge, and Lasso, using four key metrics: R2 Score, Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). Our findings indicate that Lasso Regression is the best-performing model, with the highest R2 Score, smallest MSE and RMSE, and slightly smaller MAE compared to the other two models. Therefore, we recommend the use of Lasso Regression for predicting the target variable in this dataset.

Overall, the evaluation of the three regression models provides valuable insights into their effectiveness and accuracy. The results demonstrate the importance of carefully selecting and evaluating different models using multiple metrics to determine the best-performing model for a given dataset. By doing so, we can make informed decisions and ensure the accuracy of our predictions, leading to better decision-making and improved outcomes in a variety of applications.

References

- https://www.kaggle.com/code/karnikakapoor/diamond-price-prediction
- https://beyond4cs.com/color/choosing-a-diamond-color/
- https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9689412&tag=1