T03 Diamond Dataset Analysis

Overview

This repository documents our analysis of a comprehensive diamond dataset, where we performed exploratory data analysis (EDA), data pre-processing, and regression analysis. Following is an overview of the key outcomes of our dataset.

Exploratory Data Analysis (EDA) & Preprocessing:

T1. Explore the dataset assigned to your team and provide:

a. A summary of the dataset (should include information columns present, attribute types, null values, and a summary of each attribute).

Summary of the dataset and features:

"The Largest Diamond Dataset Currently on Kaggle" is a comprehensive and extensive dataset that contains a wealth of information about diamonds. It reflects real-world information about diamonds, making it relevant for practical applications.

This dataset contains comprehensive information about diamonds, including their various attributes such as carat weight, cut quality, color, clarity, depth percentage, table percentage, price, and more. Due to its diverse attributes, this dataset is versatile and can be used for a myriad of purposes.

The data is ideal for projects related to diamond analysis, pricing prediction, and exploring factors affecting diamond quality and value. It can aid in understanding consumer preferences, assisting jewelers in inventory management, and guiding potential investors in the diamond market.

Researchers can leverage this dataset to gain valuable insights into the diamond industry. They can explore how different factors, such as carat weight, cut quality, and color, affect the desirability and value of diamonds.

Summary of the data:

The various features of the Diamond Dataset are discussed below:

1. Cut:

This feature describes the cut of a diamond. The cut of a diamond refers to how well it has been shaped and faceted from its rough form. The dataset mentions that there are around 10 common diamond cuts, with an additional one called the 'Cushion Modified.' Diamond shapes are an essential aspect of a diamond's appearance. Common diamond shapes include round, princess, emerald, pear, and more. Different shapes can affect how light interacts with the diamond and can influence its visual appeal. The cut of a diamond significantly affects its brilliance and overall visual appeal.

2. Color:

Diamonds are graded on a scale from D to Z, with D being completely colorless and Z having noticeable yellow or brown tints. Higher-grade diamonds (D, E, F) are considered more valuable because they are more colorless and allow more light to pass through, enhancing their sparkle.

3. Clarity:

Clarity refers to the presence of inclusions, which are internal flaws or imperfections within a diamond. Fewer and smaller inclusions are considered better because they can impact the diamond's brilliance. Clarity is typically assessed using a jeweler's loupe or microscope.

4. Carat Weight:

Carat weight refers to the mass of the diamond. Larger diamonds are generally more valuable, but other factors like cut and cut_quality also play a significant role in determining a diamond's worth.

5. Cut Quality:

The dataset mentions the GIA (Gemological Institute of America) Cut Grading System, which was developed in 2005 and has become the industry standard for evaluating a diamond's cut quality. The cut quality greatly influences how well a diamond reflects light and sparkles.

6. Lab:

This feature indicates the grading lab that assessed the diamond. The three major grading labs mentioned are GIA (Gemological Institute of America), IGI (International Gemological Institute), and HRD (Hoge Raad voor Diamant). These labs provide certificates that include important information about a diamond's characteristics.

7. Polish:

Polish refers to the quality of the surface finish of a diamond. It is one of the key factors that contribute to a diamond's overall appearance and visual appeal. When a diamond is cut and faceted, its facets or surfaces are created, and these surfaces need to be finely polished to achieve a smooth and reflective finish.

8. Symmetry:

Symmetry in a diamond refers to how precisely its facets are aligned and balanced in relation to one another. It evaluates the regularity and evenness of the diamond's shape and the alignment of its facets. Symmetry contributes to the diamond's aesthetic beauty and can also impact its optical properties.

9. Eye-Clean:

This feature likely describes the visibility of blemishes or inclusions to the naked eye. The dataset appears to have 10 grades to assess how noticeable these imperfections are without magnification.

10. Culet Size:

The culet size of a diamond is a characteristic that pertains to the very bottom facet of the diamond, which is typically shaped like a small point. This facet is known as the culet, and its size is an important aspect of a diamond's cut. In a well-cut diamond, the culet is often cut to a very small point, and ideally, it is either not visible or barely noticeable when viewed from the top (table) of the diamond.

11. Culet Condition:

This feature indicates if the culet has any chipping or damage. A damaged culet can affect the diamond's overall quality.

12. Fancy Color:

Fancy color diamonds are rare and highly valued gemstones that exhibit colors other than the traditional colorless or near-colorless range found in most diamonds. These diamonds are known for their vivid and intense hues, which can include shades of blue, pink, yellow, green, orange, and many more.

The features in our dataset related to fancy color are as follows:

- **Dominant Color:** The dominant color in a fancy color diamond refers to the primary or most prominent color visible when observing the diamond. It's the color that immediately catches the eye. For example, a fancy color diamond with a dominant color of "vivid blue" means that the most striking and noticeable color in the diamond is blue.
- **Secondary Color:** In some fancy color diamonds, there may be additional colors present besides the dominant color. These are known as secondary colors. Secondary colors can either complement or contrast with the dominant color, adding complexity and character to the diamond's appearance. For instance, a fancy color diamond with a dominant color of "vivid green" might have secondary colors of "yellow" and "blue." These secondary colors can appear as subtle accents within the diamond's overall coloration.
- **Overtone:** The overtone in a fancy color diamond refers to a subtle, additional hue that is visible when the diamond is examined closely. Overtone is distinct from the dominant color and secondary color. It is often described as a translucent or shimmering color that appears as a veil or sheen over the main color.
- **Fancy color intensity:** It refers to the strength, vividness, or saturation of the color in a fancy color diamond. It measures how strongly and vibrantly the diamond displays its dominant color. Intensity is a critical factor in assessing the overall beauty and value of a fancy color diamond.

13. Fluorescence (Fluor):

Fluorescence refers to how a diamond reacts to long-wave UV light. Some diamonds exhibit fluorescence, and it can be noticeable to an expert. The dataset may include information about whether a diamond exhibits fluorescence and its intensity.

The dataset has two features related to fluorescence as shown below:

- **Flour Color:** Fluorescence color refers to the color of the visible light emitted by a diamond when it fluoresces under UV light.
- **Flour Intensity:** Fluorescence intensity measures the strength or degree of fluorescence exhibited by a diamond when exposed to UV light. It quantifies how prominently the diamond fluoresces.

14. Depth Percent:

Depth percent, in the context of a diamond, is a measurement that indicates the relative depth of the diamond's pavilion (the lower part of the diamond) compared to its diameter (width when measured across the girdle). It is expressed as a percentage.

15. Table Percent:

Table percent, also known as table size, refers to the size of the flat, top facet of the diamond, known as the table facet, expressed as a percentage of the average diameter (width) of the diamond.

16. Measurements (Length, Width, Depth):

These features provide the absolute measurements of the diamond, including its length, width, and depth in some units of measurement (e.g., millimeters).

- **Meas_Length:** The length of a diamond refers to the measurement from one end of the diamond to the other along one of its longest axes.
- **Meas_Width:** The width of a diamond refers to the measurement from one side of the diamond to the opposite side, perpendicular to the length.
- **Meas_Depth:** The depth of a diamond refers to the measurement from the table (the flat, top facet) to the culet (the pointed or flat bottom). It represents the diamond's height or how deep it extends into the pavilion (the lower part of the diamond).

17. Girdle Min/Max:

The girdle is the outer edge of the diamond, and these values likely describe the minimum and maximum thickness of the girdle. The girdle thickness can affect how light is reflected within the diamond.

18. Total Sales Price:

This feature indicates the price of the diamond in dollars, representing its market value or cost.

Understanding these features is essential for assessing the quality and value of a diamond when buying or evaluating its worth in the jewelry market.

Note:

Attribute types and null values have been mentioned in the code itself.

b. Data Visualization, summarizing insights about the dataset through EDA.

All the insights have been mentioned in the ML_total_sales_price_prediction.ipynb and ML_carat_weight_prediction.ipynb file along with the code.

The reasoning for each of the preprocessing steps has also been provided wherever needed.

Regression Analysis:

T2. Identify and list regression problems on your assigned dataset. Which one does seem the most interesting to you and why?

We have identified two regression problems on our Diamond Dataset:

1. Predicting the sales price of the diamond.
2. Predicting the carat weight of the diamond.

Among these two problems, the best one is the prediction of the sale price of the diamond due to following reasons:

- **Business Impact**: Predicting the sales price directly relates to the financial aspect of the diamond industry. It can help businesses make informed decisions about pricing, marketing, and inventory management.
- **Customer Engagement**: Knowing the expected sales price can help jewelers communicate more effectively with customers. They can provide accurate price estimates and potentially increase customer satisfaction.
- **Market Insights**: Understanding the factors that influence diamond prices can provide valuable insights into market trends and consumer preferences.
- **Practical Application**: Predicting sales prices is a common problem in the diamond industry. Solving this problem can directly benefit businesses and professionals in the field.

T3. Build an end-to-end Machine Learning pipeline for your assigned dataset for the aforementioned most interesting regression problems found in T2. Your pipeline should include components for dataset preprocessing, transformation, regression model building hyperparameter tuning, grid search or optimization, and evaluation. Report results on the regression models with hyperparameter tuning, and report the best hyperparameter values. Report results using at least two relevant evaluation metrics like RMSE, MAE.

Compare results for different models and give the reasoning for that.

- End-to-End Machine Learning Pipeline

We developed an end-to-end machine learning pipeline for the selected regression problem. Our pipeline includes the following components:

Data preprocessing
Feature transformation
Regression model building
Hyperparameter tuning
Model evaluation

Evaluation Metrics: We assessed the performance of our regression models using at least two relevant evaluation metrics, such as Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE).

Model Comparison: We compared the results for different regression models like Linear Regression, Polynomial Regression, XGBoost Regressor and Random Forest Regressor, and provided the reasoning behind our choices.

Results:

1. Total Sales Price Prediction:
   The best model for this prediction is ‘Polynomial Regression’ with hyperparameter degree 2.

2. Carat Weight Prediction:
   The best model for this prediction is ‘XGB Regressor’ with hyperparameter learning_rate of 0.1 and n_estimators as 710.

Conclusion
This project showcases our comprehensive analysis of the diamond dataset, highlighting the importance of EDA, data preprocessing, and regression analysis.

All the analysis has been done in the code file along with the best model prediction for the problem.

CONTRIBUTIONS:

Each member conducted an individual Exploratory Data Analysis (EDA), which can be accessed in their respective branches, named after the team member. These individual analyses served as a foundation for our collective understanding of the dataset's characteristics.

Following the EDA phase, we convened for joint efforts in preprocessing the dataset, identifying regression problems, and implementing various regression models. Several meetings were held and the code was written during those meets.

Team members :

1. Deven Patel (202101264)
2. Takshay Makadia (202101414)
3. Sakshi Patadiya (202101469)
4. Nancy Patel (202101491) (Leader)
5. Ishita Rathod (202101516)