Team: Data Vortex

Gemstone Price Prediction

Introduction

Gem Stones Co Pvt. Ltd is a leading manufacturer of cubic zirconia, an affordable diamond alternative with similar qualities. The company aims to optimize its profits by accurately predicting gemstone prices based on various attributes provided in the dataset. This project focuses on building predictive models to distinguish between higher and lower profitable stones, ultimately improving the company's profit margins.

Goals and Importance

The primary goals of this project are:

Price Optimization: Develop predictive models to accurately forecast gemstone prices, enabling the company to set competitive yet profitable prices in the market.

Profit Maximization: Identify key factors influencing gemstone prices to guide strategic decisions aimed at maximizing profits.

Quality Control: Gain insights into the relationship between gemstone attributes (e.g., carat weight, cut, color, clarity) and prices to improve quality control measures and enhance product value.

Dataset Description

The dataset contains records of almost 27,000 cubic zirconia gemstones, each with attributes such as carat weight, cut quality, color, clarity, dimensions (length, width, height), and price. These attributes are crucial determinants of a gemstone's value in the market.

Key Features:

Carat: Carat weight of the cubic zirconia.

Cut: Quality of the cut, ranging from Fair to Ideal.

Color: Color grade of the gemstone, with D being the best and J being the worst.

Clarity: Level of clarity, indicating the absence of inclusions and blemishes.

Depth: Height of the gemstone, measured from the culet to the table, divided by its average girdle diameter.

Table: Width of the gemstone's table expressed as a percentage of its average diameter.

Dimensions (X, Y, Z): Length, width, and height of the gemstone in millimeters.

Price: Price of the cubic zirconia.

Data Preprocessing

Handling Missing Values: Checked for missing values and imputed them using the median and mean for the 'depth' and 'price' columns, respectively.

Removing Duplicates: Eliminated duplicate records to ensure data integrity.

Outlier Detection and Removal: Identified and removed outliers using the Interquartile Range (IQR) method to enhance model robustness.

Feature Engineering: Performed one-hot encoding for categorical variables ('cut', 'color', 'clarity') and normalized numerical features using Min-Max scaling.

Exploratory Data Analysis (EDA)

Analysis of Categorical Columns: Visualized the distribution of cut, color, and clarity categories using pie and count plots.

Analysis of Numerical Columns: Examined the histograms of numerical features to understand their distributions.

Checking Correlation: Investigated correlations between features using a heatmap to identify relationships and potential multicollinearity.

Q-Q Plot: Conducted Q-Q plots to assess the normality of numerical features.

Model Development

Pipelines for Regression Models: Built pipelines for various regression models including Linear Regression, XGBoost, Random Forest, KNeighbors, and Decision Tree regressors.

Cross-validation: Evaluated model performance using negative root mean squared error (RMSE) during cross-validation.

Model Interpretability and Evaluation

Statistical Significance: Employed Ordinary Least Squares (OLS) regression to assess the significance of predictor variables based on p-values.

Model Evaluation: Calculated R-squared (R^2) values to measure the proportion of variance explained by the models on the test data.

Insights and Recommendations: Identified important features influencing gemstone prices and provided actionable insights for pricing strategies and quality control measures.

How to Run the Code

1. Download and extract the files from the compressed folder.
2. The folder contains the dataset as a csv file and the jupyter notebook.
3. Make sure the csv file and the ipynb file are within the same directory.
4. Change the path to your local directory path to ensure the dataset is imported.
5. Run the jupyter notebook to execute the project (ipynb file).

Limitations and Challenges

Data Limitations: Acknowledged potential biases and limitations in the dataset, such as missing external factors influencing gemstone prices.

Continuous Improvement: Emphasized the need for continuous refinement and validation of models to ensure reliability and accuracy in real-world applications.

Conclusion

Through thorough data preprocessing, exploratory analysis, and model development, this project provides valuable insights and predictive models to support Gem Stones Co Pvt. Ltd in optimizing their pricing strategy and maximizing profits in the cubic zirconia market.

Further enhancements and validations can lead to more accurate predictions and better decision-making processes.