Data Mining of Diamond Characteristics using the SEMMA Process

October 1, 2023

AVI AJMERA

SAN JOSE STATE UNIVERSITY

1 Abstract

This research paper delves into the application of the SEMMA process for data mining on a diamond dataset. The goal is to extract insights and potentially predict diamond prices based on various features. The findings emphasize the significance of methodical data mining and highlight the intricacies of diamond pricing.

2 Introduction

Data mining is the process of discovering patterns, correlations, and knowledge from large amounts of data stored in databases, data warehouses, or other information repositories. It encompasses a variety of techniques from statistics, machine learning, and database systems. The goal is to extract valuable information from data and transform it into an understandable structure for further use.

The SEMMA process is a methodological approach for data mining and predictive analytics projects. SEMMA stands for Sample, Explore, Modify, Model, and Assess. It offers a systematic approach to data mining projects, ensuring that the entire lifecycle, from data collection to model assessment, is addressed in an organized manner. Developed by the SAS Institute, it is widely recognized and adopted, especially by organizations using SAS software for data mining.

Diamonds are precious stones that have been traded and valued for centuries. Their value is determined by various characteristics, including carat weight, cut quality, color, and clarity. This study aims to apply the SEMMA process to a diamond dataset to extract insights and potentially build predictive models to understand the factors that influence diamond prices.

3 Methods

3.1 Sample

The dataset was loaded successfully, and the following attributes were identified: carat, cut, color, clarity, depth, table, price, x, y, z, and an index column named 'Unnamed: 0'.

The dataset comprises 53,940 rows and 11 columns. Given its size, it was determined that there's no immediate need for sampling, allowing us to work with the full dataset.
To validate the model, the data was partitioned into training and validation sets using an 80-20 split, yielding 43,152 entries for training and 10,788 entries for validation.

3.2 Explore

Summary statistics were computed for the training data. These statistics highlighted potential anomalies such as zero values for the diamond dimensions x, y, and z.
Histograms were utilized to illustrate the distributions of key continuous features such as carat, price, depth, and table.
Distributions of categorical variables like cut, color, and clarity were also examined. The majority of diamonds possess an 'Ideal' cut, with colors G, E, and F being the most prevalent. SI1 and VS2 were the most common clarity grades.

3.3 Modify

Anomalies were identified in the diamond dimensions: x had 7 zero values, y had 6 zero values, and z had 19 zero values. Zero values were replaced with the median of their respective columns.
Additionally, a new feature was engineered to represent the volume of the diamond, calculated using the formula: Volume = x $\times$ y $\times$ z. This feature captures the overall size of the diamond and could be related to its price.

3.4 Model

The analysis focused on a regression problem, predicting a diamond's price based on its features. Two models, Random Forest and Gradient Boosting Machines (GBM), were trained. The Random Forest model outperformed the GBM in terms of RMSE and MAE on the validation set. Scatter plots visualizing actual vs. predicted prices were created for both models to further visualize their performance.

4 Results

The Random Forest model achieved an RMSE of 630.97 and an MAE of 293.02 on the validation set. In contrast, the GBM model achieved an RMSE of 854.93 and an MAE of 437.54. The scatter plots further confirmed the superior performance of the Random Forest model, with its predictions clustering more closely around the line of identity compared to the GBM model.

5 Discussion

The study's findings underscore the importance of the SEMMA process in methodically approaching data mining projects. The analysis emphasized the intricacies of diamond pricing based on their features, such as carat, cut, color, clarity, and the newly engineered volume feature.
While both Random Forest and GBM are powerful algorithms, the former's superior performance in this study may be attributed to the nature of the dataset and the features used.

However, further tuning and experimentation with other algorithms could yield even better results.

6 Conclusion

This research applied the SEMMA process to a diamond dataset, yielding valuable insights into diamond pricing. The study highlighted the importance of methodical data mining, the significance of feature engineering, and the potential of machine learning models in predicting prices. Future work could delve deeper into model tuning, feature selection, and exploring other algorithms for improved performance.