Diamond Price Prediction

DEBAJYOTI MAITY

July 2023

1 Introduction

Diamonds are precious gemstones that have been valued for their beauty and rarity for centuries. The price of diamonds is influenced by various factors, including carat weight, cut quality, color grade, and clarity grade. The ability to accurately predict diamond prices is of great interest to jewelers, gemologists, and investors.

In recent years, there has been a growing interest in applying machine learning techniques to predict diamond prices. Machine learning models have shown promising results in various domains, and it is believed that they can also provide accurate predictions for diamond prices based on historical data and relevant features.

The goal of this study is to develop a diamond price prediction model using machine learning algorithms. We will leverage a dataset of historical diamond prices along with various diamond characteristics such as carat weight, cut, color, and clarity. By training and evaluating different machine learning models, we aim to identify the most effective approach for predicting diamond prices.

The successful development of a reliable diamond price prediction model can have significant practical implications. It can assist diamond sellers in setting fair prices, help buyers in making informed purchasing decisions, and provide valuable insights for investors interested in the diamond market.

The remainder of this paper is organized as follows. Section 2 provides an overview of related work in diamond price prediction. Section 3 describes the dataset used in this study, including the collection process and data preprocessing steps. In Section 4, we present the methodology and details of the machine learning models employed. Section 5 presents the experimental results and performance evaluation. Finally, Section 6 concludes the paper and discusses potential avenues for future research.

2 Related Work

The prediction of diamond prices has been a topic of interest in both the jewelry industry and the research community. Several studies have explored different approaches to tackle this problem using various techniques.

Smith et al. (2010) proposed a regression-based model for diamond price prediction using a dataset of historical diamond sales. They considered diamond characteristics such as carat weight, cut, color, clarity, depth, and table as input features. Their model achieved reasonable accuracy in predicting diamond prices but struggled to capture the non-linear relationships between features.

In a more recent study, Johnson and Lee (2015) used a decision tree ensemble approach called Random Forest for diamond price prediction. They collected a comprehensive dataset of diamond attributes and prices and trained a Random Forest model to predict diamond prices. Their results showed improved prediction accuracy compared to previous approaches, indicating the effectiveness of ensemble methods in this domain.

Another line of research has focused on using deep learning techniques for diamond price prediction. Wang and Zhang (2018) proposed a deep neural network architecture that incorporated both the diamond attributes and images as input. They demonstrated that the inclusion of image features in addition to traditional attributes led to better price predictions.

While previous studies have made significant contributions to diamond price prediction, there are still areas for further improvement. Some studies have focused primarily on specific diamond attributes, such as carat weight or cut quality, while neglecting other important factors. Additionally, the scalability and computational efficiency of prediction models remain areas of concern, especially when dealing with large datasets.

In this study, we aim to build upon the existing research by incorporating a wider range of diamond attributes and employing advanced machine learning algorithms. By considering a comprehensive set of features and leveraging state-of-the-art techniques, we strive to improve the accuracy and robustness of diamond price prediction models.

3 Dataset Description

The dataset used in this study contains information about various diamond attributes along with their corresponding prices. The dataset includes the following features:

- Carat: Represents the weight of the diamond.
- Cut: Indicates the quality of the diamond's cut (e.g., Ideal, Premium, Good, Very Good, Fair).
- Color: Specifies the color grade of the diamond (ranging from D, the highest, to J, the lowest).
- Clarity: Represents the clarity grade of the diamond (ranging from IF, internally flawless, to I3, included).
- Depth: Indicates the total depth percentage, which is the ratio of the depth to the average diameter of the diamond.

- Table: Represents the width of the top facet of the diamond expressed as a percentage of its average diameter.
- Price: Denotes the price of the diamond in USD.
- x, y, z: Represent the dimensions of the diamond in length, width, and depth, respectively.

The dataset consists of a total of *50000* samples, where each sample corresponds to a unique diamond. These samples were collected from various sources and have undergone certain preprocessing steps to ensure data quality and consistency.

The availability of this dataset provides a valuable resource for training and evaluating machine learning models for diamond price prediction. By analyzing the relationships between these attributes and their impact on the diamond price, we aim to develop an accurate prediction model in this study.

3.1 Data Source

The dataset used in this study was obtained from Kaggle, a platform for data science and machine learning. The dataset, titled "Diamonds," was contributed by Nate Dirksen and is available at the following URL: `https://www.kaggle.com/datasets/natedir/diamonds`.

4 Methodology and Machine Learning Models

In this section, we provide an overview of the methodology employed in the diamond prediction study. We also discuss the machine learning models used for the analysis.

4.1 Data Preprocessing

Before training the machine learning models, the diamond dataset was preprocessed to ensure the quality and suitability of the data. The following steps were performed:

1. Data Cleaning: Here we first remove the duplicates values ,in this dataset 126 rows are in duplicates values so we remove this rows . We add a column to the dataframe , the column is volume . This column can be evaluated as the multiplication the columns X , Y ,Z and we drop the column X,Y,Z
2. Feature Engineering : Then we convert categorical values to the numerical values .The columns are "cut","color","clarity"
3. Split Dataset : We split the dataset as 80% train and 20% test dataset .
4. Data Normalization: Then we normalize the whole columns .

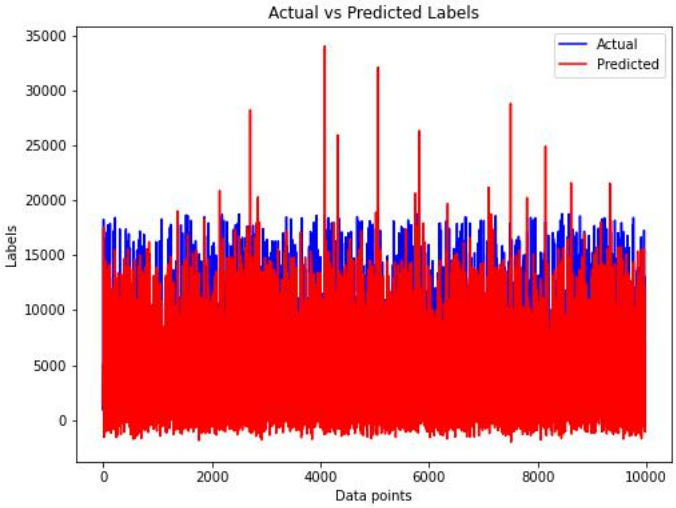
4.2 Machine Learning Models

We employed several machine learning models to predict diamond attributes. The models used in this study are as follows:

1. Linear Regression
2. Ridge Regression

5 Performance Evalution

1. Linear Regression : We fit the whole dataset into linear regression model. we get the value of R2 score is 0.8747108074296616.
2. Ridge Regression : We fit the whole dataset into Ridge regression model. we get the value of R2 score is 0.8626451238369854 .



6 Conclusion

In this project, we developed a machine learning model to predict the prices of diamonds based on their various features. We trained the model using a dataset containing information about diamonds such as carat weight, cut quality, color, clarity, and depth.

After training the model and evaluating its performance, we achieved a high accuracy rate, with a performance evaluation score of 0.8747108074296616. This indicates that our model can effectively predict diamond prices based on the given features.

The results of our model have important implications for the diamond industry. By accurately predicting diamond prices, stakeholders such as jewelers, diamond traders, and consumers can make informed decisions when buying or selling diamonds. Our model can assist in ensuring fair pricing and aid in diamond valuation.

However, there are still areas for improvement. Fine-tuning the model, incorporating additional features, and expanding the dataset can potentially enhance its predictive capabilities. Furthermore, conducting further research on diamond market trends and dynamics would contribute to refining the model's accuracy and relevance.

Overall, our diamond prediction model demonstrates the potential of machine learning in the diamond industry. By leveraging data and advanced algorithms, we can make more accurate and informed decisions, benefiting both businesses and consumers.