Predicting Diamond Prices: A Statistical Modeling Approach

Chehak Arora[1], Jessica Ezemba[2], Sanika Gokakkar[1], Kirpa Kaur[1], Hyunseok Lee[1], Sahana Krishna Murthy[1]
Department of MCS Interdisciplinary [1], Mechanical Engineering [2]

Introduction

Diamond prices pose a unique forecasting challenge in the luxury goods market. While diamonds are prized for their unmatched hardness and beauty, their value is driven by complex, non-linear factors including carat, cut, clarity, table, and depth.[1] Current pricing methods rely heavily on the Rapaport price list and subjective expert assessments, creating inconsistencies between listed and actual transaction prices.[2] This research explores statistical modeling approaches to improve diamond price prediction accuracy, aiming to reduce uncertainty for investors and buyers in this high-stakes market.

Figure 1: Diamonds price by size

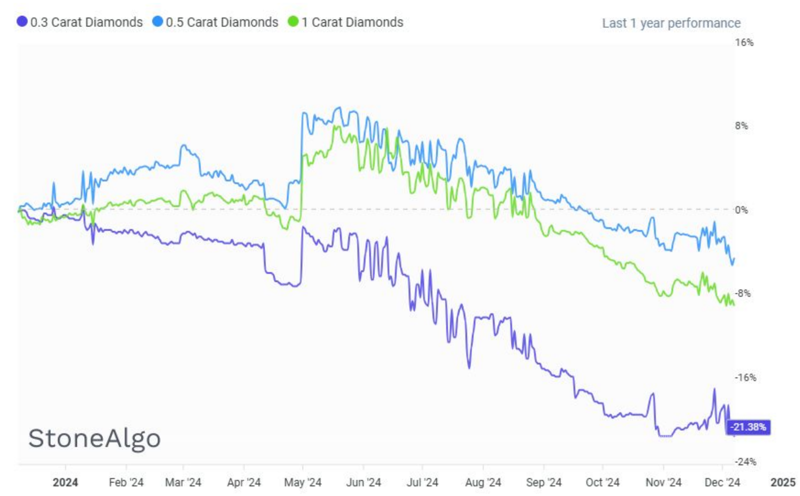

Figure 2: 2024 Diamonds price distribution

Data

The dataset consists of 53,940 rows and 11 variables. We observe that the response variable (price) is highly positively skewed, so we apply a logarithmic transformation.

Name	Description
Carat	Diamond weight (1 carat ~ 200 milligrams)
Cut	Graded quality (Fair, Good, Very Good, Premium, Ideal)
Color	Graded color (J is worst, D is best)
Clarity	Graded clarity (I1, SI2, VS2, ... in ascending order of quality)
X	Length of diamond (millimeters)
Y	Width of diamond (millimeters)
Z	Depth/height of diamond (millimeters)
table	Width of top part of diamond relative to widest point (percentage)
depth	Depth of top part of diamond relative to total depth (percentage)
Price	The price of the diamond (dollars)

Table 1: Overview of diamond characteristics and their definitions.

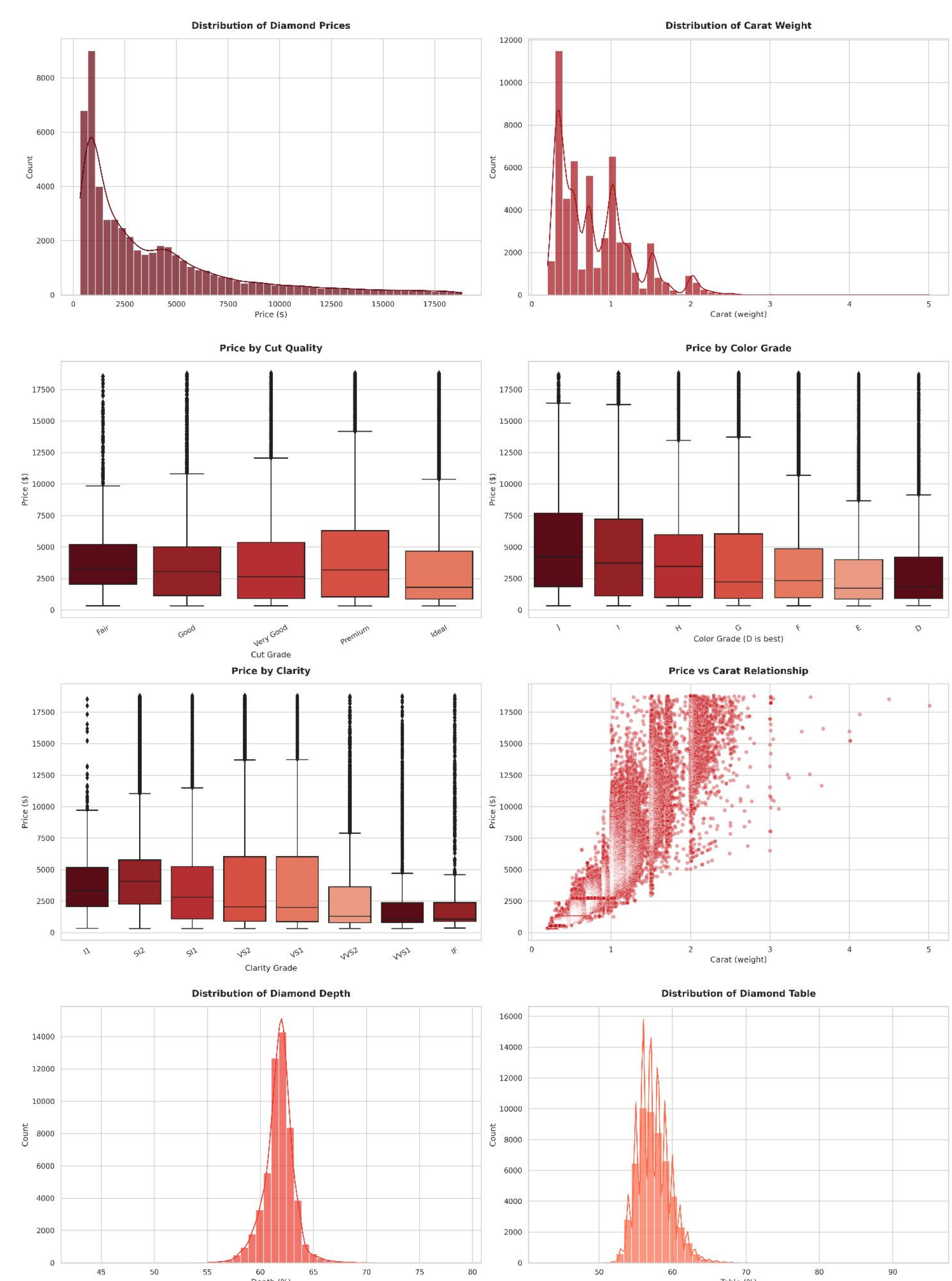

Figure 3: Exploratory Data Analysis of the Diamond Dataset

Analysis

Dataset Split

The dataset was split into 80% training and 20% testing sets, with a random seed ensuring reproducibility. The training set was used for model development, and the test set for performance evaluation.

Machine Learning Models

In this study, we explored and applied a range of machine learning models to predict diamond prices based on key features. These models included linear regression, multiple linear regression with variable subset selection, decision tree, random forest, and extreme gradient boosting (XGBoost).

Model	Test RMSE
Linear Regression	0.1670
Decision Tree	0.3080
Random Forest	0.0923
XGBoost	**0.0893**

Table 2: Test RMSE Across Machine Learning Models

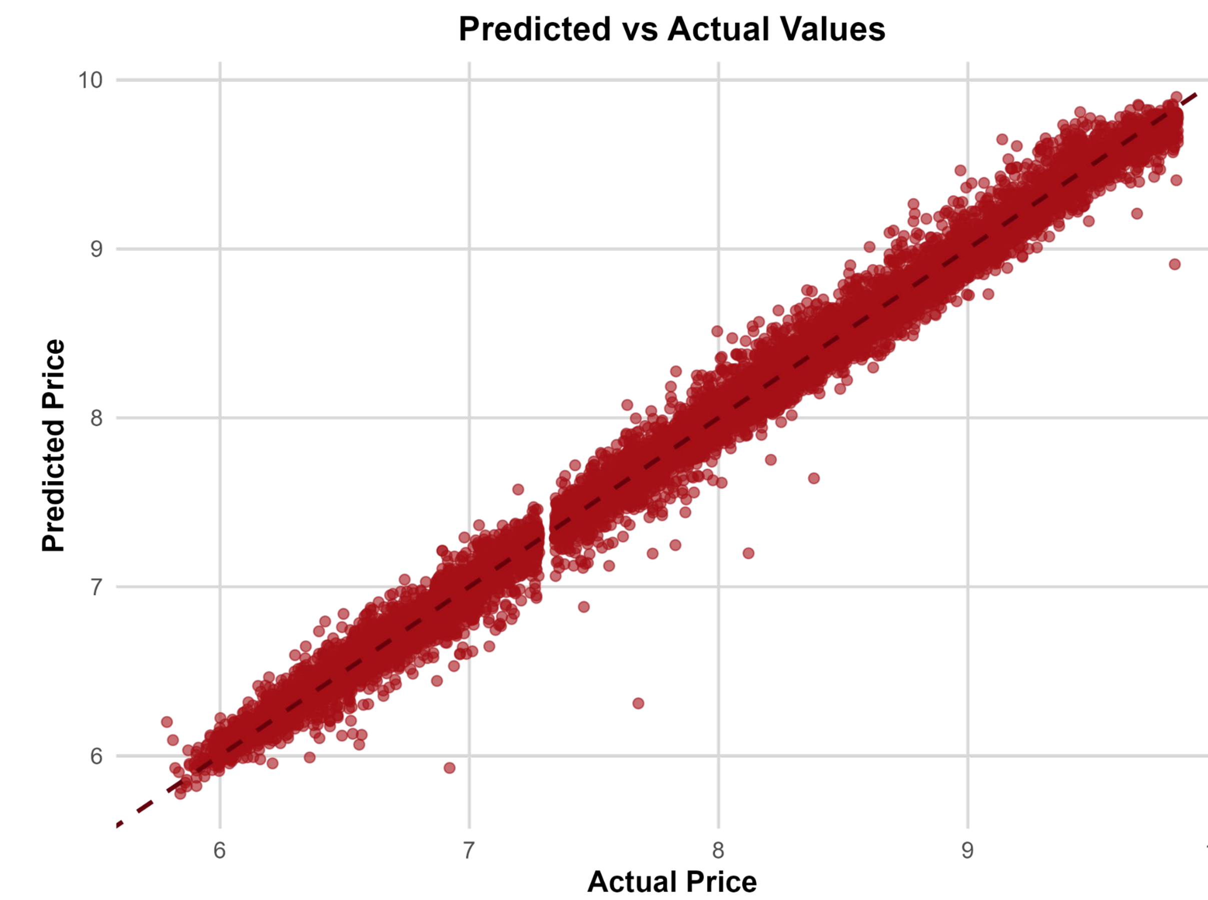

Figure 4: Predicted vs Actual Prices for the XGBoost Model

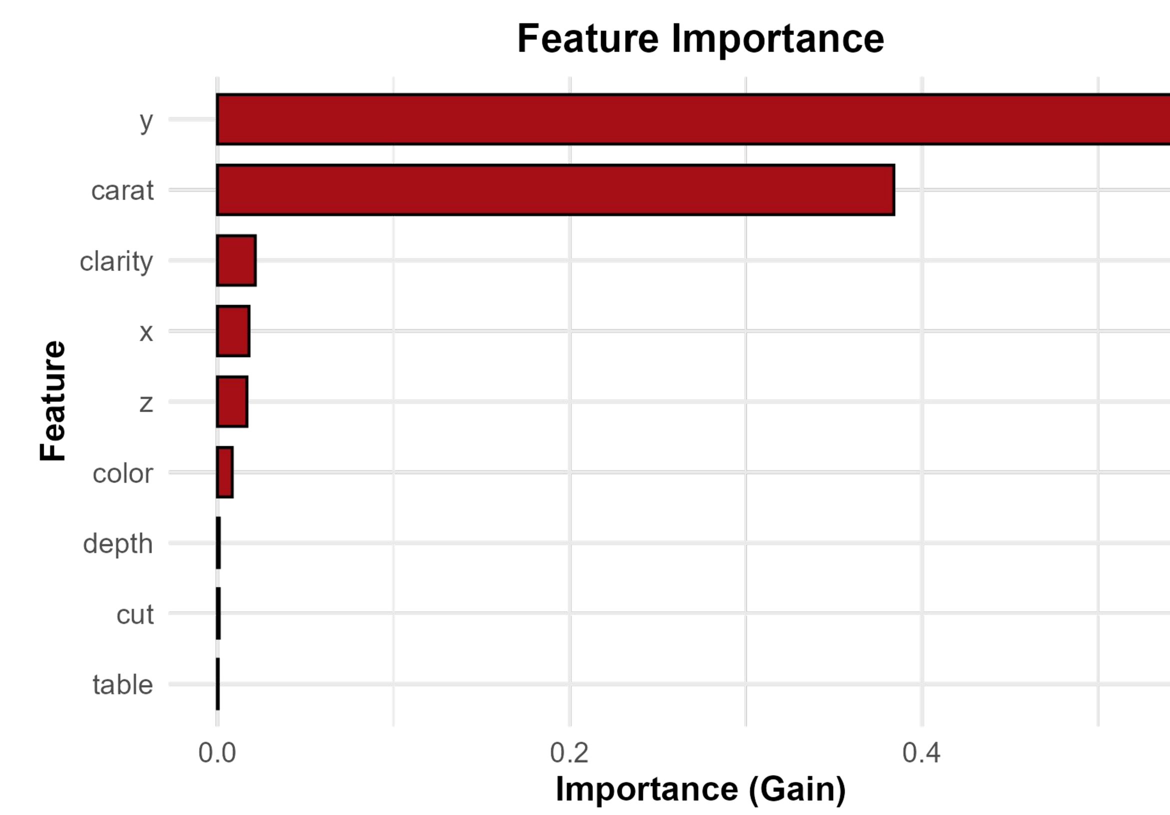

Figure 5: Feature Importance for the XGBoost Model

Conclusion

The XGBoost model outperformed linear regression, decision tree, and random forest in predicting diamond prices, achieving the highest R-squared value (0.992) and lowest RMSE (0.0893). Feature importance analysis highlights the key predictors, with y, carat, clarity, and x being the most influential variables, as shown in the accompanying plot. This performance demonstrates the model's capability to capture complex relationships, making it ideal for applications in diamond pricing, such as online marketplaces and valuation tools.

References

[1] Kigo, S. N., Omondi, E. O., & Omolo, B. O. (2023). Assessing predictive performance of supervised machine learning algorithms for a diamond pricing model. *Scientific Reports*, *13*(1), 17315.

[2] StoneAlgo. (n.d.). *0.3 carat diamond prices*. Retrieved December 9, 2024, from https://www.stonealgo.com/diamond-prices/0.3-carat-diamond-prices/