Group Members: Arjun Rajesh, Chris Breton,

Lena Weissman, Luke Hartfield, and Varsha Ramesh

Predicting Diamond Prices Using Physical and Quality Attributes

Description of Project Goals

Description:

We analyzed a dataset of 53,940 round-cut diamonds with 10 attributes: carat, cut, color, clarity, depth, table, price, and dimensions (length, width, depth). Our main research question is: "How do physical and quality factors influence a diamond's price, and can we accurately predict its price using these features?" To answer this, we will examine the relationships between each attribute and price. In addition, we will create some new features that will aid in the analysis of the dataset. Following this, we will develop a machine learning model to predict diamond prices for new data.

Importance of the Problem:

Diamond pricing impacts both consumers and the jewelry industry. A reliable prediction model can help buyers avoid overpaying and sellers price competitively. By building a model that can provide price estimates for given diamond specifications, we aim to help buyers and sellers make more informed decisions.

Summary Statistics:

Carat weights ranged from 0.20 to 5.01, with an average around 0.80. Depth percentage averaged about 61.75%, ranging from 43% to 79%. Table percentage (the size of the diamond's flat top surface compared to its overall width) averaged around 57.5%, ranging from 43% to 95%. Prices varied widely, from $326 up to over $18,000. The diamond dimensions (length, width, depth) showed some unusual extreme values, including zeros and very high numbers

Exploratory Analysis

Feature Engineering

We checked the dataset for any missing or null values and found none, so no further data handling was needed. To better analyze the set, we changed the cut, color, and clarity categories into rankings based on their quality. For example, cuts like Fair got a lower rank (1), and Ideal got a higher one (5). Similarly, colors from J (lowest quality) to D (highest quality) and clarity from I1 (lowest) to IF (highest) were ranked on a scale of 1-8. This way,

instead of just treating these as labels, we gave them numbers that represent how good they are, which helps the model understand their effect on price. In other words, we made some ordinal numerical features to attempt to have a better model.

We also created a new feature called volume by multiplying the diamond's length, width, and depth. This gave us a better idea of the diamond's overall size, rather than just using carat or individual measurements.

Initial EDA

Before building models to address our research question, we began with some initial data exploration to guide our approach. We first created visual plots of key variables against price to observe potential trends. This showed the following: Diamonds with a Premium or Fair cut tended to have higher average prices than those with an Ideal cut, as shown by Figure 1. For color, stones graded J and I had the highest average prices, while E and D were on the lower end, suggesting that other attributes may be influencing value beyond color alone. The results of the analysis on price by color can be seen in Figure 2.

For clarity, diamonds with an SI2 grade showed the highest average prices, with prices generally decreasing for higher-clarity grades like IF and VVS1, again indicating potential interactions with other features such as carat weight. The relationship with price based on clarity category can be seen in Figure 3. Next, we generated a correlation plot showing how each variable relates to price, giving us an early sense of which features might be most influential. Finally, we used a pairplot to examine relationships among all variables, which can sometimes reveal possible interaction effects. While pairplots alone make it difficult to clearly identify such effects, they still offer useful preliminary insight into the patterns within our data.

Solution and Insights

Feature Importance Analysis

Carat weight is the strongest predictor of diamond price, with volume (overall size) a close second. Clarity and color also significantly influence price, while depth, table, and cut have comparatively less impact according to Figure 4.

Modeling Approach, Evaluation & Results:

We evaluated multiple models, starting with a multiple linear regression and then moved to more complex approaches such as bagging and random forests, grow-prune, and XGBoost, in order to identify a model that can most

accurately predict diamond prices from physical and quality attributes. The dataset was split into training (80%) and testing (20%) sets, with standardization applied where necessary.

Models Evaluated:

As shown in Figure 5, our initial multiple linear regression model did not fit the data well. It failed to capture a clear pattern, consistently underestimating prices at the lower end and overestimating prices at the higher end. Overall, it performed only slightly better than the simplest possible baseline, predicting the mean price for all observations. Based on these results, we explored alternative models that could better capture the underlying relationships and reduce noise. The RMSE for this model was approximately $1,204.

Examining the coefficients provided additional insight into what drives the model. Carat, clarity, and color emerged as the variables with the largest coefficients, indicating they had the strongest influence on predicted price.

Our second approach used bagging. By bootstrapping the data and generating multiple shallow trees, we aimed to improve predictive accuracy. However, with the same set of features as the linear regression, the improvement was minimal: RMSE decreased only from $1,204 to $1,154. Given this small gain, the next logical step was to introduce randomness into feature selection as well as sampling, leading us to test random forests.

The random forest model produced considerably better results, as shown in Figure 6. RMSE dropped to $535 compared to $1,154 from bagging, representing a substantial improvement. The random feature selection in this model likely contributed to its performance by reducing overreliance on the most dominant variables. By incorporating a broader range of features across trees, the model captured more of the underlying patterns in the data and reduced noise.

We also tested a decision tree using the grow–prune algorithm. The idea was to grow a deep tree that could capture complex relationships, then prune it back to remove less important splits and reduce overfitting. Figure 7 shows this tree. However, the out-of-sample RMSE increased sharply to $1,333, indicating the model overfit the training data and performed worse than simpler models like linear regression.

Finally, we applied boosting, a method widely used in industry. In our case, it achieved an RMSE of $542, nearly identical to that of the random forest. Given these similar results, either model could be used with reasonable confidence to predict diamond prices using the selected features. As shown

in Figure 8, the boosting model fits the data well, generally avoiding both underfitting and overfitting. While a few outliers remained difficult to predict accurately, the overall performance was strong. The process of building shallow trees, iteratively correcting residual errors, and combining the results proved just as effective as the random forest's approach of random feature selection and bagging.

Results

Among all models tested, Random Forest and XGBoost produced the lowest RMSE values, our primary measure of predictive performance. While MSE measures the same error as RMSE, it is less interpretable for a non-technical audience. R^2 provides an indication of fit, but given that all models used the same predictors, with Random Forests drawing random subsets for each tree, the consistently high R^2 values across models were expected and therefore less useful for differentiating performance. Figure 9 displays all the models and their corresponding RMSE, MSE, and R^2 for comparison. Random Forest is the clear predictive winner with an R^2 of 98% and the lowest RMSE of 535.14.

The feature engineering described at the start of the paper appears to have contributed to better model performance, though quantifying its exact impact would be a useful next step. Our initial exploratory analysis offered hypotheses that guided the modeling process and were largely supported by the results. Feature importance analysis for XGBoost and Random Forest also showed that volume and carat were by far the most influential predictors, followed by clarity and color, while cut, depth, and table had smaller but still measurable effects on price.

It is important to note that these models are predictive rather than causal, they do not explain why a diamond has a certain price, only that given its characteristics, the model can provide an estimate. Based on our results, we can predict diamond prices with an average error of about $540, calculated as the midpoint between the RMSE values for XGBoost and Random Forest. This prediction model can be useful for consumers seeking fair pricing, jewelers aiming to price competitively, and marketplaces that want to increase transparency in an industry that is traditionally opaque.

Appendix

Figure 1:

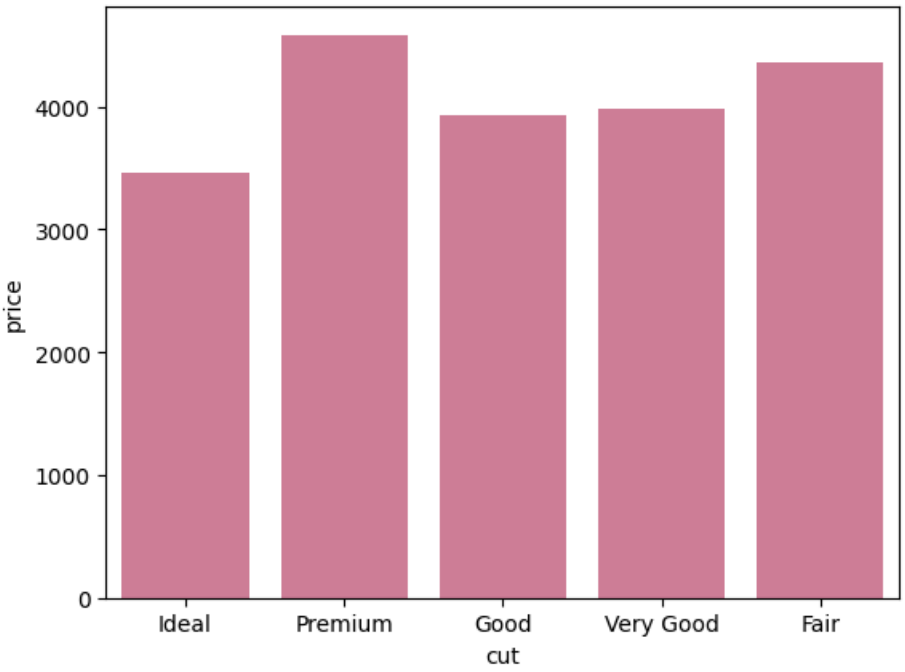

Figure 2:

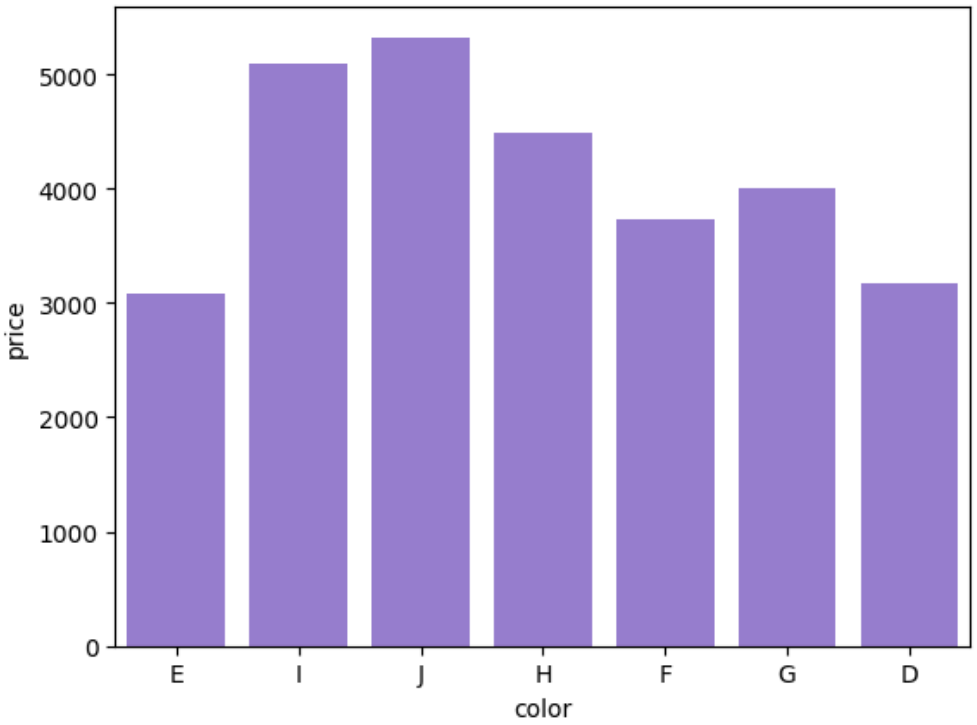

Figure 3:

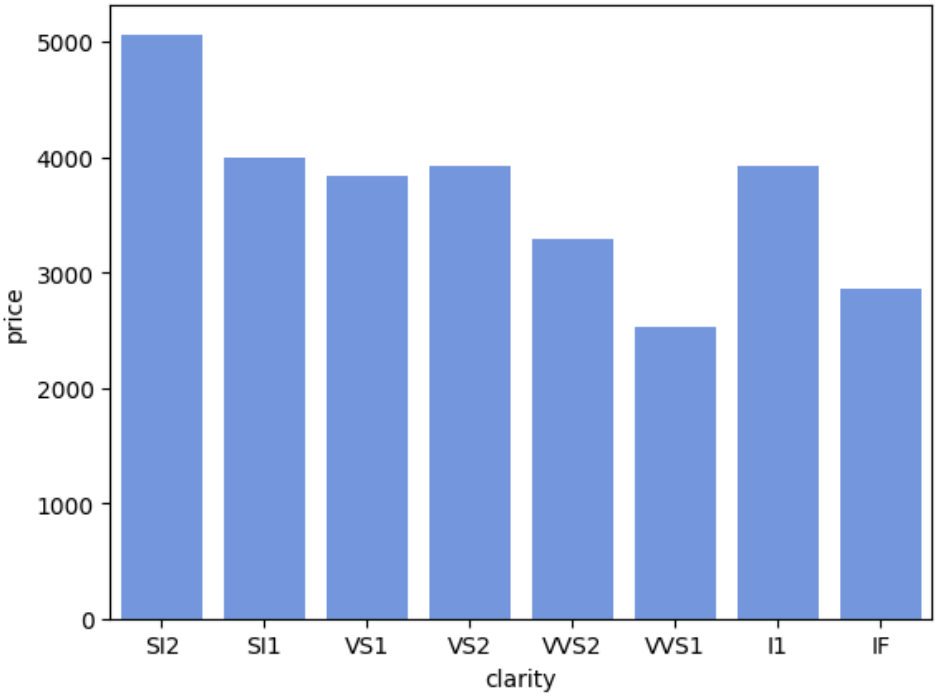

Figure 4:

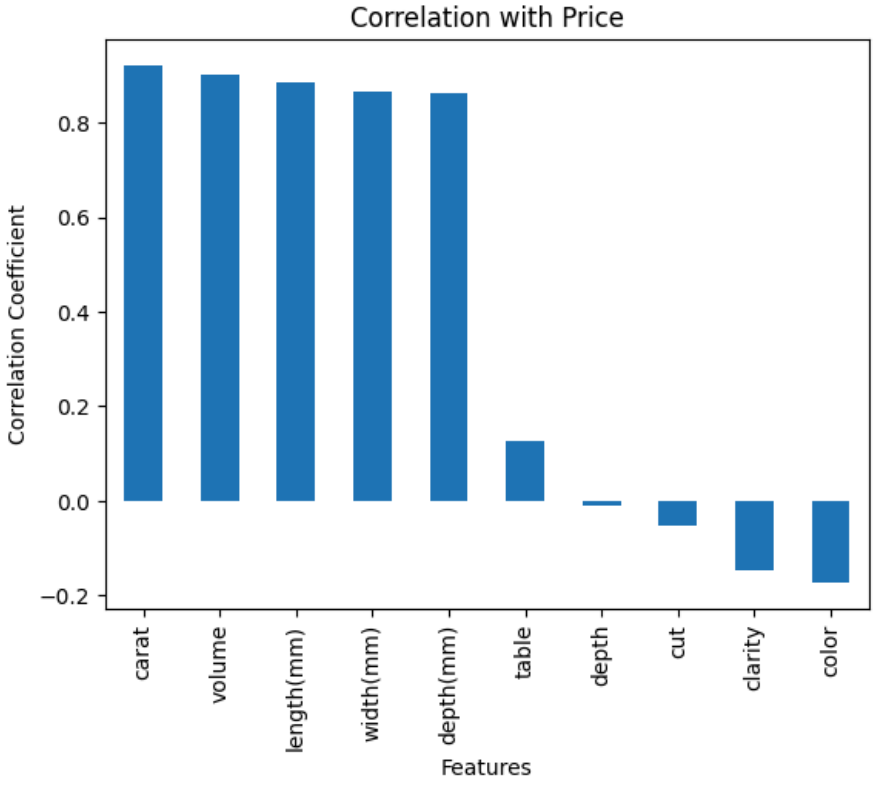

Figure 5:

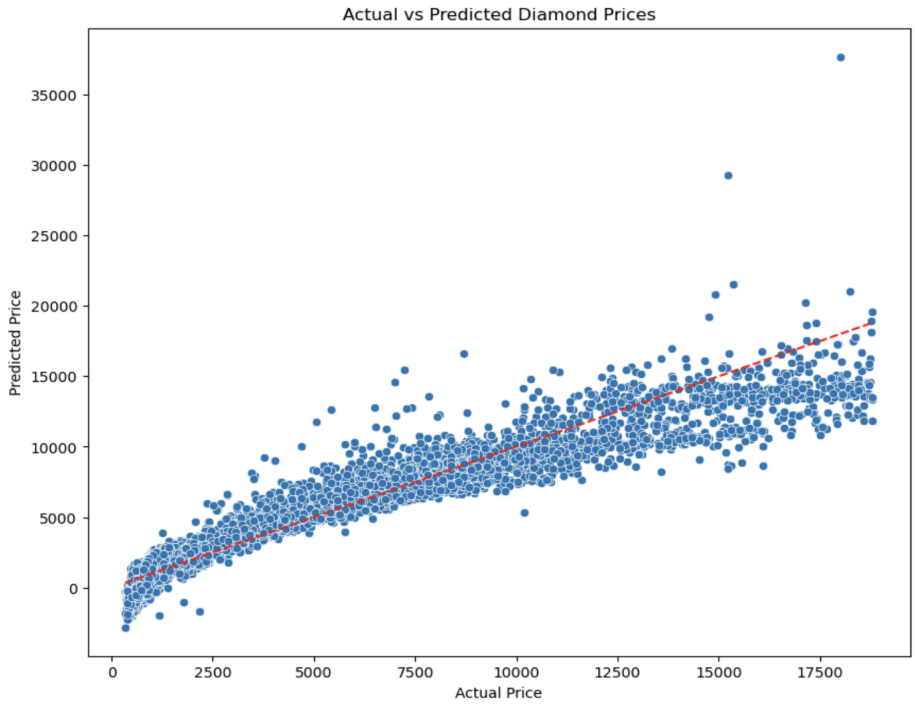

Figure 6:

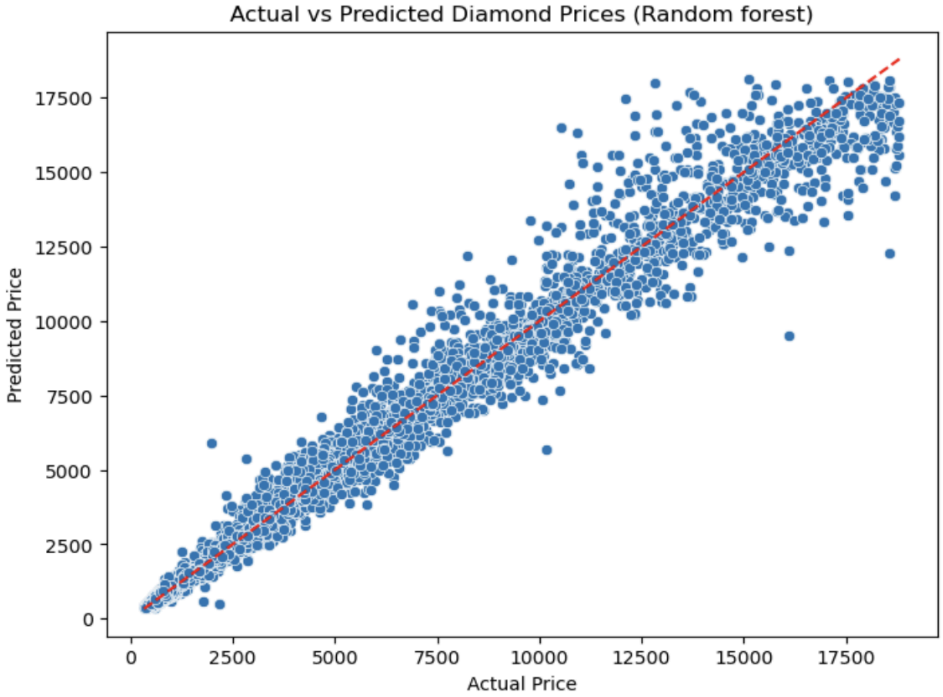

Figure 7:

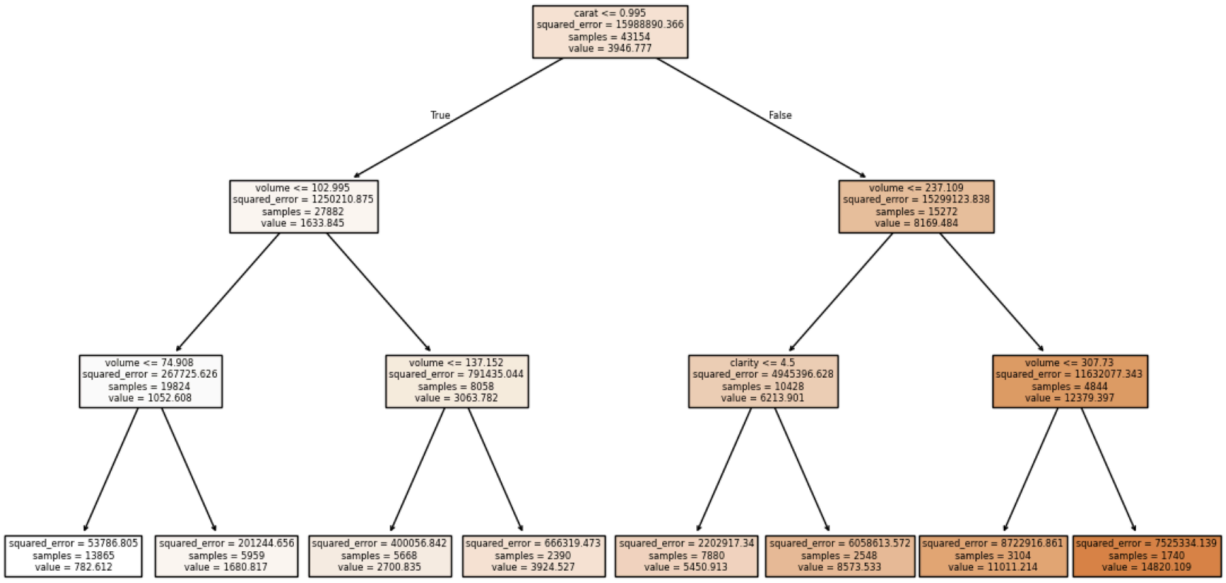

Figure 8:

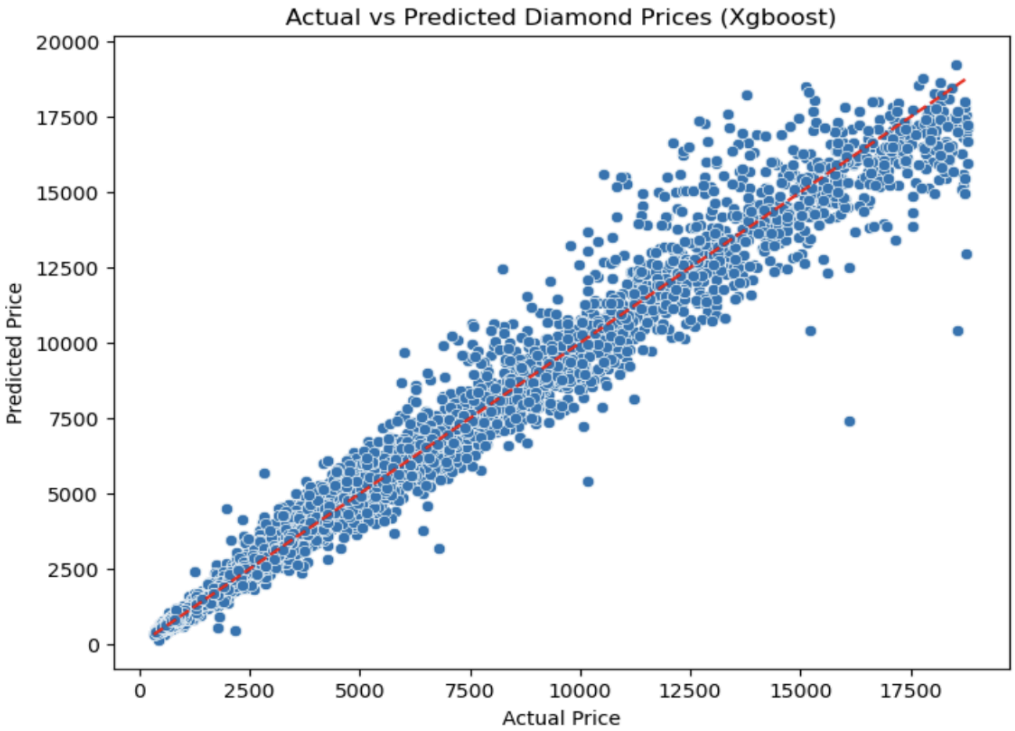

Figure 9:

Model	RMSE	MSE	R-squared
Linear Regression	1204.86	1,451,708	0.91
Decision Tree	711.50	511,684	0.97
Pruned Tree	1333.14	1,777,275	0.89
Random Forest	535.14	286,563	0.98
XGBoost	542.84	294,682	0.88
Bagging	1154.12	1332001	0.91