DSA Final Project

Diamond Price Prediction

Shivanshu Himanshu Kumar Gupta
AI21BTECH11027 AI21BTECH11012

May 2024

1 Introduction

Diamonds are one of the most valuable and sought-after gemstones in the world. The price of a diamond is determined by various factors such as carat, cut quality, color grade, clarity grade, depth, etc. Predicting diamond prices accurately is crucial for both buyers and sellers in the diamond industry. In this analysis, we aim to predict diamond prices using the properties of diamonds. For this we first applied data preprocessing techniques to clean and format the data. After that we did EDA and feature modelling to understand data followed by training different prediction model and comparing results.

2 Data description

Dataset used for this analysis is from kaggle which can be downloaded from here. It has total 53940 rows and 10 columns among which 9 are predictor variables and price is target variable. Predictor variables include diamond characteristics like carat, grade, etc. You can see the data columns and its description in fig 1.

Column Name	Description
carat	Weight of the diamond
cut	Quality of the cut (Fair, Good, Very Good, Premium, Ideal)
color	Diamond colour, from J (worst) to D (best)
clarity	How clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))
x	Length in mm
y	Width in mm
z	Depth in mm
depth	Total depth percentage = z / mean(x, y) = 2 * z / (x + y) (43--79)
table	Width of top of diamond relative to widest point (43--95)
price (target)	Price in US dollars (326--18,823)

Figure 1: Data columns description

3 Exploratory data analysis (EDA)

3.1 Feature Visualization

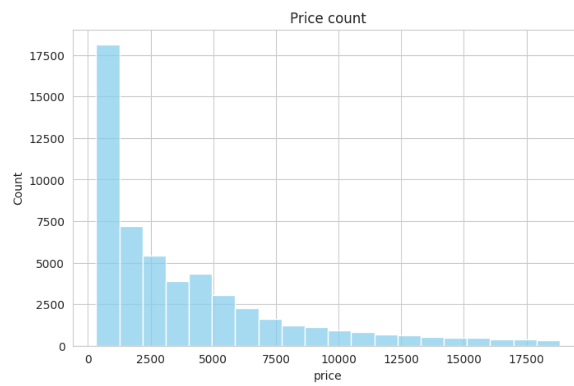

Figure 2: Price Vs Diamond Count

In fig 2 we can see that quantity of diamonds with less price are lot more compared to diamonds with more price.

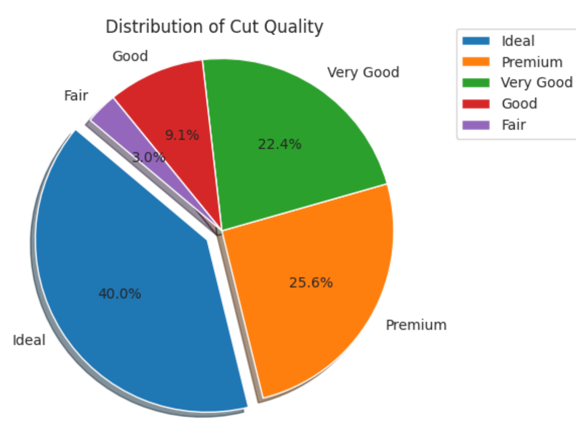

Figure 3: Distribution on the basis of cut Quality

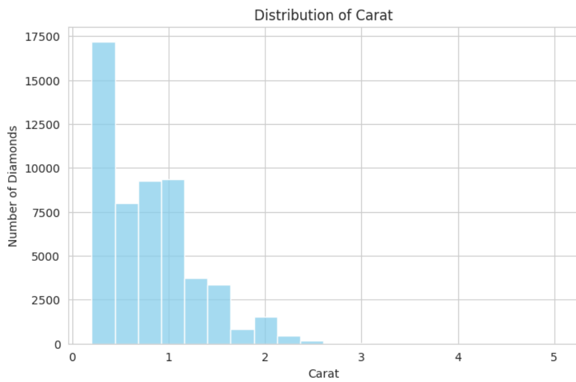

Figure 4: Carat Vs Diamond Count

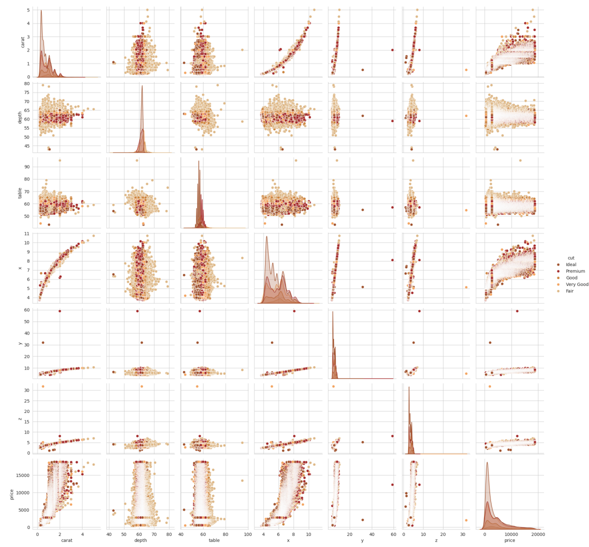

Figure 7: Pairplot

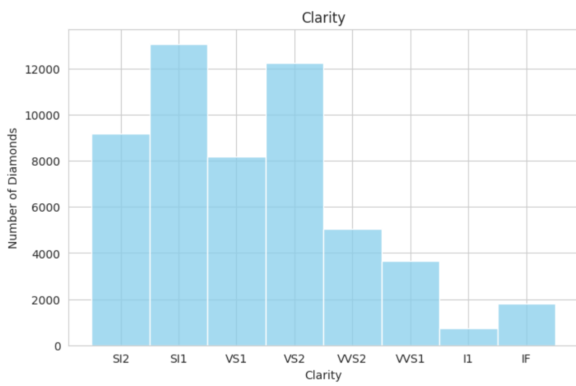

Figure 5: Clarity Vs Diamond Count

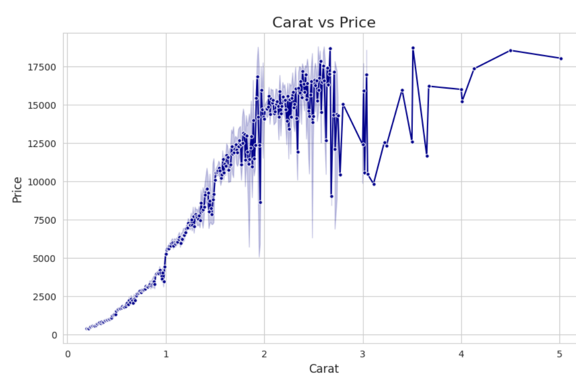

Figure 8: Plot showing with increase in carat Price is increasing

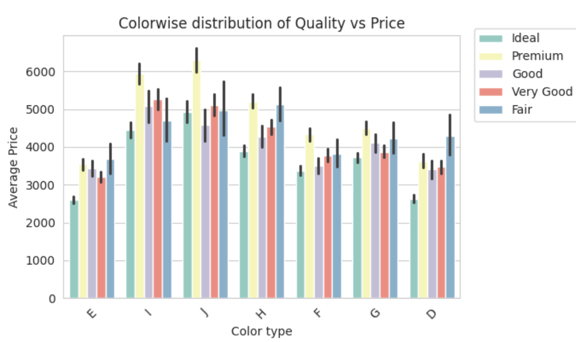

Figure 6: Color-wise distribution of Quality vs Price

3.2 Correlation analysis

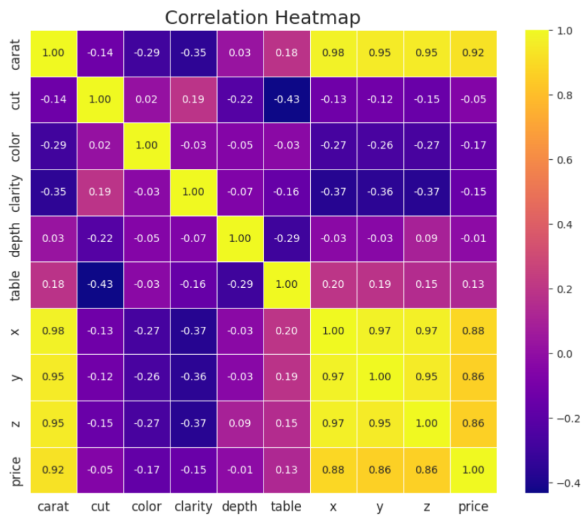

Figure 9: Correlation heatmap

In fig 9 we can see that x, y and z are too much correlated. This maybe because diamond shapes is generally fixed which makes ratio of x, y and z approximately remains constant.

4 Data Preprocessing

In this section, we describe the steps taken to preprocess the dataset before analysis.

4.1 Missing data handling

```
RangeIndex: 53940 entries, 0 to 53939
Data columns (total 10 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   carat    53940 non-null  float64
 1   cut      53940 non-null  object
 2   color    53940 non-null  object
 3   clarity  53940 non-null  object
 4   depth    53940 non-null  float64
 5   table    53940 non-null  float64
 6   x        53940 non-null  float64
 7   y        53940 non-null  float64
 8   z        53940 non-null  float64
 9   price    53940 non-null  int64
dtypes: float64(6), int64(1), object(3)
```

Figure 10: Columns info

Above we can see that there are no null values in any columns so we don't need to handle null values.

4.2 Unwanted data removal

	carat	depth	table	x	y	z	price
count	53940.000000	53940.000000	53940.000000	53940.000000	53940.000000	53940.000000	53940.000000
mean	0.797940	61.749405	57.457184	5.731157	5.734526	3.538734	3932.799722
std	0.474011	1.432621	2.234491	1.121761	1.142135	0.705699	3989.439738
min	0.200000	43.000000	43.000000	0.000000	0.000000	0.000000	326.000000
25%	0.400000	61.000000	56.000000	4.710000	4.720000	2.910000	950.000000
50%	0.700000	61.800000	57.000000	5.700000	5.710000	3.530000	2401.000000
75%	1.040000	62.500000	59.000000	6.540000	6.540000	4.040000	5324.250000
max	5.010000	79.000000	95.000000	10.740000	58.900000	31.800000	18823.000000

Figure 11: Columns statistics

In fig 11 we can see that the minimum value of x, y and z is 0 which is not possible as diamonds can't be dimensionless. So we removed the rows which had any x, y, or z values as 0. Now we left with 53920 rows.

4.3 Categorical variables encoding

In fig 10 we can see that there are 7 float, 1 int and 3 object(string) data columns. So we need to handle object data type. Since the values are hierarchical in all 3 columns, we need to do ordinal encoding. Specifically, the 'cut' feature was encoded with values ranging from 1 to 5, representing the ordinal quality levels from 'Fair' to 'Ideal'. Likewise, the 'color' feature was transformed into numerical values from 1 to 7, denoting color grades from 'J' to 'D', with higher values indicating superior color quality. Similarly, the 'clarity' feature was encoded with values spanning 1 to 8, delineating clarity grades from 'I1' to 'IF'. Such encoding ensured that the categorical features were appropriately interpreted by machine learning algorithms.

5 Feature Engineering

5.1 PCA

Since in the correlation matrix there are many predictors which are highly correlated. So to remove this we applied PCA which decreased the number of features from 9 to 5 explaining 95% variance. We tried training the models on the transformed data also and got satisfactory results.

6 Model training and prediction

For predicting diamond price given its properties we used 3 models-

- Linear regression

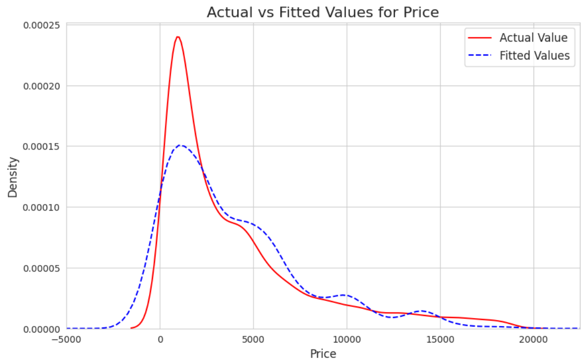

Figure 12: Linear Regressor

- Decision Tree Regression

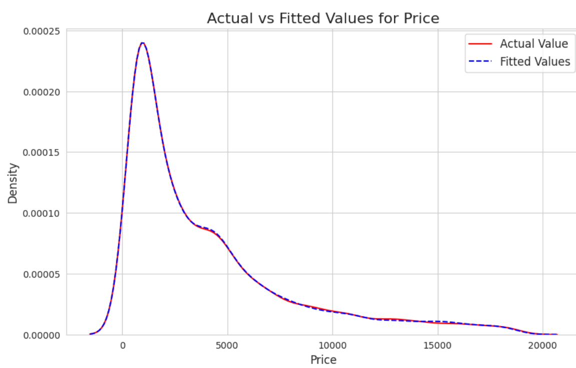

Figure 13: Decision Tree Regressor

- Random forest Regression

7 Model Evaluation

7.1 Evaluation Metrics

Linear regression, Decision Tree, and Random Forest are popular machine learning models used for regression tasks. In this evaluation, we assess the performance of these models on a regression problem using three key metrics: Root Mean Square Error (RMSE), Accuracy, and Mean Absolute Error (MAE).

RMSE measures the average deviation of predicted values from the actual values. A lower RMSE indicates better performance of the model. Accuracy measures the proportion of correct predictions made by the model. MAE represents the average absolute difference between the predicted and actual values.

The table below presents the evaluation metrics for each regression model:

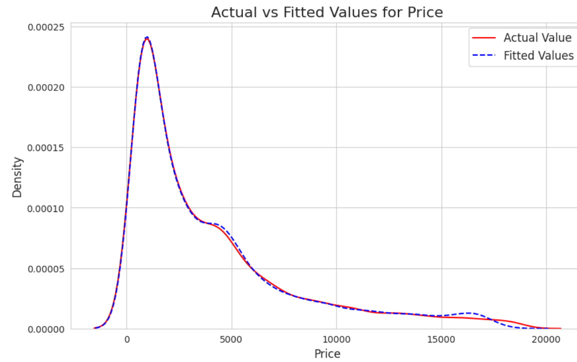

Figure 14: Random Forest Regressor

Table 1: Regression Model Evaluation Metrics

Model	**RMSE**	**Accuracy**	**MAE**
Linear Model	1201.39	0.910	259.12
Decision Tree	715.49	0.968	351.61
Random Forest	513.76	0.984	259.12

7.2 Results

Table 2: Accuracy Without PCA

	Linear Model	**Decision Tree**	**Random Forest**
Train	0.905	0.999	0.997
Test	0.909	0.968	0.983

Table 3: Accuracy With PCA

	Linear Model	**Decision Tree**	**Random Forest**
Train	0.846	0.999	0.996
Test	0.842	0.951	0.975

8 Conclusion

In this analysis, we investigated the factors influencing diamond prices and built predictive models to estimate diamond prices based on these factors.

8.1 Key Findings

Through exploratory data analysis (EDA), we identified several key features that strongly influence diamond prices. Carat and dimensions(x, y,

x) were found to be the most important factors affecting diamond prices.

We trained and evaluated several machine learning models, including Decision Tree, Random Forest, and Linear Regression among which decision tree and random forest are giving mostly same result with latter one being on high side.

After applying PCA we had got that 5 features are explaining 95% of variance. On using PCA transformed data we got nearly same results as before in decision tree and random forest. So we can say that PCA comes out to be helpful as it is making model simple(only 5 features in input instead of 9) without compromising performance.

On the other hand PCA didn't seem to be helpful in case of linear regression as it is decreaing accuracy abruptly. The reason can be that PCA features lose their linearity to capture more variance.

Moreover, the Random Forest Regressor outperformed the Decision Tree Regressor in terms of RMSE, achieving a lower value of 513.76 compared to 715.49. However, the Linear Regressor, while achieving a respectable R-squared value of 0.910, exhibited higher error metrics with an RMSE of 1201.39. These findings suggest that ensemble methods like Random Forest are well-suited for diamond price prediction tasks, providing more accurate and reliable results compared to traditional linear models.