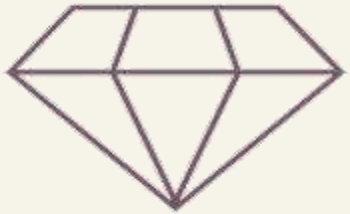
Project Proposal

Prediction How is Diamond Priced Based on its Attributes

SDAIA
الهيئة السعودية للبيانات والذكاء الاصطناعي
Saudi Data & AI Authority

Amal Nassier Almuteb



Introduction

Diamonds have become important indispensable resources for humans. Its versatility makes it a highly valued element in our society. Hence, people are willing to pay a lot of money for objects created using diamonds.

Due to the combination of different features for determining its worth, a diamond's price could range from a couple hundred dollars to millions of dollars.

SDAIA
الهيئة السعودية للبيانات والذكاء الاصطناعي
Saudi Data & AI Authority

Dataset Description

Diamonds dataset hosted on Kaggle.
and contains

53940
rows

11
variables

Columns represent:

'Unnamed: 0' - ID of the diamond
Carat - Measures weight of the diamond , 1 carat = 200 mg
Cut - Identifies quality of the cut (Ideal to Fair)
Color - Identifies color of the diamond (D - Best to J - Worst)
Clarity - Measures the clarity of the diamond (IF - Best, I1 - Worst)
Table - Width of top of diamond relative to its widest point
x, y, z - Length, width and height of the diamond respectively
Depth - Value derived from length, width and height of the diamond
Price - Price of the diamond

Loading Data

```
df = pd.read_csv('diamonds.csv')
```

```
df.head()
```

	Unnamed: 0	carat	cut	color	clarity	depth	table	price	x	y	z
0	1	0.23	Ideal	E	SI2	61.5	55.0	326	3.95	3.98	2.43
1	2	0.21	Premium	E	SI1	59.8	61.0	326	3.89	3.84	2.31
2	3	0.23	Good	E	VS1	56.9	65.0	327	4.05	4.07	2.31
3	4	0.29	Premium	I	VS2	62.4	58.0	334	4.20	4.23	2.63
4	5	0.31	Good	J	SI2	63.3	58.0	335	4.34	4.35	2.75

```
df.shape
```

```
(53940, 11)
```

SDAIA
الهيئة السعودية للبيانات والذكاء الاصطناعي
Saudi Data & AI Authority

Dataset Observations:

- There are no missing values in the dataset
- We can delete the IDs column as it provides no useful statistics.
- We can delete the Depth column as it is a derived column from length, width and height columns
- We need to rename the columns.
- In the minimum value of x ,y and z is zero and It doesn't make any sense to have length\width\depth of a diamond to be zero, so we can drop these rows
- The Price column is the dependent column and all others are independent columns

Identifying Types of Variables:

Categorical

- Cut
- Color
- Clarity

Numerical

- Carat
- Table
- x, y, z
- Depth
- Price

SDAIA
الهيئة السعودية للبيانات والذكاء الاصطناعي
Saudi Data & AI Authority

Project Target and Some potential problem that can be discussed in the dataset

Fully understand how the pricing system of diamonds worked.

Visualization the important qualities of a diamond.

Created a predictive model to predict the prices of diamonds.

SDAIA
الهيئة السعودية للبيانات والذكاء الاصطناعي
Saudi Data & AI Authority

Algorithm

Make the linear regression model based on the following steps:

01

Import Required Packages
Load the dataset
Data Preprocessing

02

Perform the exploratory data analysis (EDA)

03

Create a linear regression model

04

Train the model to fit the data

05

Make predictions using the trained model

SDAIA
الهيئة السعودية للبيانات والذكاء الاصطناعي
Saudi Data & AI Authority

Tools

Pandas:

a library offers data structures and operations for manipulating numerical tables and time series.

Numpy:

a library used for working with arrays. It also has functions for working in domain of linear algebra, fourier transform, and matrices.

Matplotlib:

a plotting library for the Python programming language and its numerical mathematics extension NumPy.

Seaborn:

a data visualization library built on top of matplotlib and closely integrated with pandas data structures in Python.