Project Description: Diamond Price Prediction Model

Objective

The primary objective of this project is to develop a predictive analytics model that accurately estimates the price of diamonds based on their attributes. The model will be trained on a provided dataset containing various characteristics of diamonds and their corresponding prices.

Dataset Overview

The dataset includes the following attributes for each diamond:

carat: Weight of the diamond *(range: 0.2 - 5.01)*

cut: Quality of the cut (categories: Fair, Good, Very Good, Premium, Ideal)

color: Diamond color, graded from J (worst) to D (best)

clarity: Measurement of how clear the diamond is (grades: I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))

depth: Total depth percentage, calculated as 2 * z / (x + y) *(range: 43 - 79)*

table: Width of the top of the diamond relative to its widest point *(range: 43 - 95)*

price: Price in US dollars *(range: $326 - $18,823)*

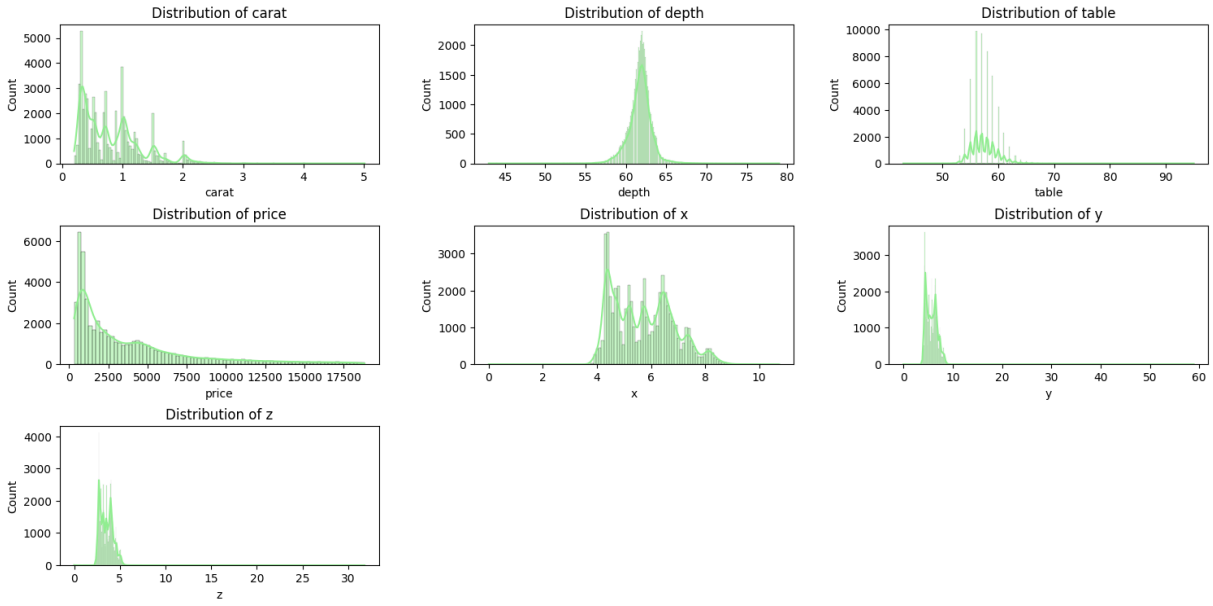
x: Length in mm *(range: 0 - 10.74)*

y: Width in mm *(range: 0 - 58.9)*

z: Depth in mm *(range: 0 - 31.8)*

Analysis based on distributions and descriptive statistics:

1.Distributions



- The distribution plot basically shows the distribution of data over the dataset.
- It gives the range of the each column and number of items on a specific value(frequency of values).
- Any human error or outliers, skewness can be observed with the help of this distribution.
- According to the diamond dataset each column range and number of items in each value within range is shown and there seems to be no massive outliers which will impact the model prediction.

Some key observations:

1)Carat - Carat value distribution seems to be *right skewed* meaning that smaller carat values or dominating in the dataset.

2)Price - Following that the price is a perfect *right skewed* distribution which makes even clear that high costs diamonds are low in count.

3)Dimensions – The dimensions also exhibits a typical *right skewed* distribution where some mid range values are hyped, which explains that mid ranged diamonds also prevails in the data set.

2)Descriptive Statistics

Descriptive statistics gives the basic structure of the dataset which helps in better understanding of each columns. It provides each columns mean, IQR, max, min values etc.

Data Transformations

1) Encoding categorical features

- Model require numerical values to perform well so the categorical values in the data set is transformed into numerical values with the help of techniques one hot encoding, label encoding etc.
- Categorical features ‘cut’ , ‘color’, ‘clarity’ are encoded into numerics using label encoder as they are recognized as ordinal data.

Standardization

When there are larger differences between the values are there in the features of the dataset, Standardization comes into play where it provides a scale with standardized input ranges and thus increasing the efficiency of the model training.

In our diamond data set *Standarscalar* method was used to standardize the features like carat , dimensions , table, depth etc.

Correlation Matrix

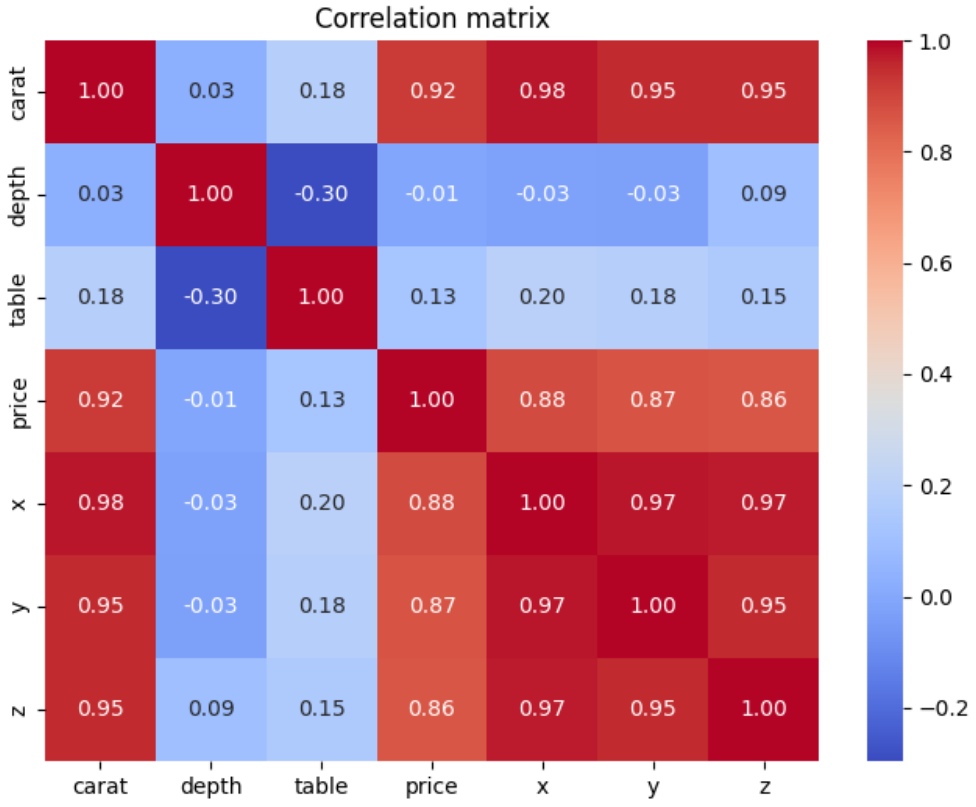
The Correlation matrix is used to find relationships among the features of the dataset. It ranges from -1 to 1 where,

1 → Positive Relationship

0 → No Relationship

-1 → Negative Relationship

Correlation matrix of diamond dataset



Strong Relationships

1.Carat

From the diagram its clear that carat is strongly positively correlated with **'price'(0.92) , 'x'(0.98) , 'y'(0.95)** and **'z'(0.95) .** This indicates that increase in carat will result in increase of the price.

2.Price

It is also seen that price is strongly connected with **'price'(0.92)** and **'dimensions'(0.87)** which indicates that price increases with increase in carat and dimensions.

3.Dimensions

Dimensions are also strongly connected with the ‘price’ and ‘carat’.

Weak Relationships

1.depth

Depth is maintaining negative relationships with price and some other features. This shows that depth values are not much associated with price.

2.table

Similarly table also maintaining weak or no relationship with other features and indicates it is not associated with the price.

3.cut, color and clarity

The encoded categorical values even though scores only decent relationship than table and depth, these positive and negative relations can be added for predicting the price as long as it is not closer to no relationship(0).

Observation

From the correlation matrix it is observed that strong relationship features like carat, x, y & z are significant predictors of the price. So the weak features like table and depth can be dropped.

Feature Selection

Feature selection is the process where important features that helps in predicting the target value are selected lowering the training time and process with

accurate results. It is used in cases when the dataset contains a large number of features which makes the model difficult to train.

1)Mutual Info Regression and KBest Selector

In this project one of the method implemented to select important features influencing the target price is KBest Selector which uses the Mutual Info Algorithm to find out the important features.

2)Feature importance method

Each model will come up with a built in method called feature_importances_ which will provide scores for each of the features based on its influence on calculating the target value.

Results of different models

Why Regression?

Regression is used to predict continuous output values as of our diamond price value. There are various types of regression and lets see in action which of the selected regression models performed well.

1)Linear Regression

- Linear regression is an easy to implement algorithm that uses in-dependent variables to predict the dependent variables.
- Linear regression is selected for the diamond price prediction as there are independent variables like carat, color, dimensions etc and the value to be predicted is a continuous value ‘price’.

- Even though linear regression is simple and easy to understand model it cannot handle higher variations among prices and thus provides an average predicted output which not satisfies the expected outcome.
- Mean Squared Error: 1881952.883447832
- R2 Score: 0.881614517139921

2)Decision Tree Regressor

- Decision Tree Regressor observes the features and trains a model using the tree structure.
- Decision Tree Regressor provides a good prediction values but not the best prediction.
- The tree structure looks pretty complex for this price prediction as different aspects involved like cut, carat , dimensions , clarity. It causes overlooked prediction as inspecting on various trees.
- Mean Squared Error: 515039.76967463846
- R2 Score: 0.9676010848298338

3) Gradient Boosting Regression

- Gradient Boosting Regression on the other hand is one of the best model in predicting prices like in our case diamond price prediction.
- It analyzes the relation between x and y values and factors ox x influencing y values there by makes the prediction.
- It provides almost 95% of accuracy in predicting the diamond prices.
- Mean Squared Error: 393682.0900619299

- R2 Score: 0.9752351694161643

4)Random Forest Regressor

- Random Forest Regressor is also a technique which uses tree based approach to train a model.
- Random forest have the specialty to play well with skewed values as our dataset contains skewed features that we observed in distribution.
- This reduces overfitting by averaging multiple tree decisions.
- Random Forest and Gradient Boosting Regression are top predictors of diamond prices with almost 95% of accuracy.
- Mean Squared Error: 315114.03666001407
- R2 Score: 0.9801775444464689

5)KNN Regressor

- KNeighbor Regressor is also an easy to understand algorithm which train the model with the help of similarity that is using the neighbors around it.
- Its accuracy is close to the Random Forest Regressor and Gradient Boosting Regressor and it handles variances well.
- But it takes time to train when compared to the other well performing algorithms.
- Mean Squared Error: 328977.80378290697
- R2 Score: 0.979305435065018

Best Model

According to the scores collected from each model and undergoing focus on how these models works and considering other evaluations like complexity, two of the models performed well,

1) **Gradient Boosting Regression**
2) **Random Forest Regressor**

Score Comparison

1)Random Forest

R2 Score: Random Forest - Mean R-squared: 0.28699734280923206 - Std: 0.4302028338672336

Neg mean squared error: Random Forest : 561.0090206903575

Explained Variance - Mean: Random Forest :0.980153097941707 - Std: 0.0012909868045886087

2)Gradient Boosting

R2 Score: Gradient Boosting - Mean R-squared: 0.39347947792717514 - Std: 0.33164721417849274

Neg mean squared error: Gradient Boosting : 678.3160562705917

Explained Variance - Mean: Gradient Boosting :0.9710769680484 - Std: 0.001435362521023079

Among these two I would go with RANDOM Forest Regressor which is a tree based algorithm and on my observation I see that Random Forest Regressor works well on Diamond prediction.

Limitations of Random Forest Regressor

1) Random Forest Algorithm struggles with categorical values and is sensitive to outliers.
2) Random forest is a bit time consuming algorithm.
3) Not easily interpretable meaning that decisions taken using this algorithm is not easily understandable.