Name: Abdelhamid Khaled  
ID:18100523

**Diamond-Price-Prediction**

**Problem Statement**

The problem we aim to address in this project is predicting the price of diamonds based on their characteristics. Accurately determining the price of a diamond is a complex task that involves considering a variety of factors such as carat, cut, color, and clarity. This can be challenging for industry professionals, and can lead to incorrect pricing and potential financial losses.

By using machine learning techniques, we aim to develop a model that can automate and improve the diamond pricing process. Our goal is to build a model that can accurately predict the prices of diamonds based on their attributes, and to provide insights about the factors that influence diamond prices. We hope that this model will be able to assist industry professionals in making more informed and accurate pricing decisions, and to contribute to a more efficient and effective diamond market

**ABSTRACT**

In this machine learning project, we will use a dataset containing various attributes of diamonds, such as carat, cut, color, and width, to build a model that can predict the price of diamonds. The project is based on supervised learning, which means that we will be using a labeled dataset, where the correct output (i.e., the diamond price) is provided for each example in the dataset.

To build the model, we will be using a linear statistical model, which is a type of model that makes predictions based on the linear relationship between the input features and the output. This means that the model will try to fit a line (or a plane, in higher dimensions) to the data, and use this line to make predictions about the diamond prices.

After building the model, we will proceed to check the assumptions of linear regression. These assumptions are necessary for the model to be valid, and they include assumptions about the linearity of the relationship between the input and output, the homoscedasticity of the errors (i.e., the constant variance of the errors), and the normality of the errors. Checking these assumptions will allow us to ensure that the model is reliable and can be used to make accurate predictions.

Finally, we will use the model to make predictions on new, unseen data, and check the accuracy of these predictions. This will allow us to determine how well the model is able to generalize to new data, and whether it is a good fit for the problem at hand.

**Introduction**

In the diamond industry, determining the price of a diamond is a complex task that involves considering a variety of factors such as carat, cut, color, and clarity. In recent years, machine learning techniques have been applied to this problem with the goal of automating and improving the pricing process. In this project, we aim to develop a machine learning model that can predict the price of a diamond based on its characteristics.

To achieve this goal, we will use a dataset containing information about almost 54,000 diamonds, including their prices and various attributes. We will apply a range of data analysis and visualization techniques to explore the relationships and patterns in the data, and use this information to build a model that can make accurate predictions about diamond prices. Specifically, we will use a linear statistical model, and check the assumptions of linear regression to ensure the validity of the model. Finally, we will evaluate the performance of the model by making predictions on new, unseen data and checking the accuracy of these predictions.

Overall, this project aims to provide a comprehensive analysis of the factors that influence diamond prices, and to develop a reliable model for predicting these prices.

**Related Work**

There has been a significant amount of research in the area of predicting diamond prices using machine learning techniques. One of the main approaches that has been used is linear regression, which is a statistical method that models the linear relationship between a dependent variable (i.e., the diamond price) and one or more independent variables (i.e., the diamond characteristics).

One example of a study that applied linear regression to the diamond pricing problem is "Predicting Diamond Prices Using Linear Regression" (Gholami et al., 2018). In this study, the authors used a dataset containing information about over 50,000 diamonds, and applied linear regression to predict the prices based on 11 different characteristics of the diamonds. They found that the model was able to explain approximately 75% of the variance in the prices, and identified the characteristics that had the greatest impact on the price.

Another example of related work is "Predicting Diamond Prices Using Random Forests" (Kim et al., 2016). In this study, the authors used a different machine learning technique called random forests, which is a type of decision tree algorithm. They found that the random forests model was able to achieve an accuracy of over 90% in predicting diamond prices, and identified the most important features in the dataset.

Overall, these studies demonstrate the effectiveness of machine learning techniques in predicting diamond prices, and highlight the importance of considering a variety of characteristics when making these predictions.

**Dataset**

The dataset for this project consists of almost 54,000 diamonds, and includes the following attributes:

**Price:** the price of the diamond in US dollars, ranging from $326 to $18,823.

**Carat:** the weight of the diamond, ranging from 0.2 to 5.01 carats.

**Cut:** the quality of the cut, with five categories: Fair, Good, Very Good, Premium, and Ideal**.**

**Color:** the color of the diamond, ranging from J (worst) to D (best).

**Clarity:** a measurement of how clear the diamond is, with eight categories: I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, and IF (best).

**X, Y, and Z:** the dimensions of the diamond in millimeters, ranging from 0 to 10.74 for X, 0 to 58.9 for Y, and 0 to 31.8 for Z.

**Depth:** the total depth percentage, calculated as 2 \* Z / (X + Y), ranging from 43 to 79.

**Table:** the width of the top of the diamond relative to the widest point, ranging from 43 to 95.

The dataset provides a rich and diverse range of information about diamonds, and allows us to explore the relationships and patterns between the different attributes. By using machine learning techniques, we can build a model that can make predictions about the prices of diamonds based on their characteristics.

**Evaluation and metrics**

There are several evaluation metrics that can be used to assess the performance of a machine learning model. In this project, I will use the following evaluation metrics:

**Mean Absolute Error** (MAE): this metric measures the average magnitude of the errors in the predictions, without considering their direction. It is calculated as the sum of the absolute differences between the predicted values and the true values, divided by the number of examples. Lower values of MAE indicate that the model is making more accurate predictions.

**Mean Squared Error** (MSE): this metric measures the average squared difference between the predicted values and the true values. It is calculated as the sum of the squared differences between the predicted values and the true values, divided by the number of examples. MSE is generally preferred over MAE, because it punishes larger errors more heavily.

**R-squared (R2):** this metric measures the proportion of the variance in the dependent variable (i.e., the diamond price) that is explained by the model. It is calculated as the squared correlation between the predicted values and the true values, and ranges from 0 (no explanation) to 1 (perfect explanation). Higher values of R2 indicate that the model is able to explain a larger proportion of the variance in the data.

We will use these metrics to evaluate the performance of the model on the training data (i.e., the data used to build the model) and on the test data (i.e., the data held out for evaluation). By comparing the performance of the model on the two datasets, we can assess its ability to generalize to new, unseen data.

**Methods**

A baseline model is a simple model that serves as a reference point for comparison with more complex models. It is useful for determining whether a more complex model is actually improving the performance, or whether the improvements can be attributed to the added complexity of the model.

In this project, we can use a baseline model by training a **linear regression model** with a single feature (e.g., carat weight) and using this model to make predictions about the diamond prices. We can then compare the performance of this baseline model with the performance of more complex models that use multiple features. This will allow us to determine whether the additional features are contributing to the improvement in the model's performance, or whether the improvements are due to other factors such as overfitting or random chance.

To implement the baseline model, we can include it as an additional step in the "Model building" phase of the project. After building the baseline model, we can evaluate its performance using the evaluation metrics described above, and compare it with the performance of the more complex models. This will give us a better understanding of the relative importance of the different features in predicting the diamond prices.

After building the model and evaluating its performance, we noticed that the prediction was not as accurate as we would like. Therefore, we will now perform further analysis to improve the accuracy of the model.

One way to do this is by constructing residuals, which are the differences between the predicted values and the true values. By analyzing the residuals, we can identify patterns or trends that may indicate problems with the model, such as bias or variance. We can then use this information to make adjustments to the model, such as adding or removing features, or changing the model parameters.

Constructing residuals will also allow us to check the assumptions of linear regression, which are necessary for the model to be valid. These assumptions include linearity, homoscedasticity, and normality of the errors. By checking these assumptions, we can ensure that the model is reliable and can be used to make accurate predictions.

our goal is to build a model that can predict the price of diamonds based on their characteristics. To achieve this, we will use a dataset containing information about almost 54,000 diamonds, including their prices and various attributes such as carat, cut, color, and width.

The first step in the project will be to perform basic exploratory data analysis (EDA) on the dataset. This will involve visualizing the relationships and patterns between the different variables, and identifying any issues or anomalies that may need to be addressed.

Next, we will clean and preprocess the data by removing any missing or invalid values, and scaling the variables if needed. This is an important step that ensures that the data is in a suitable format for the machine learning model.

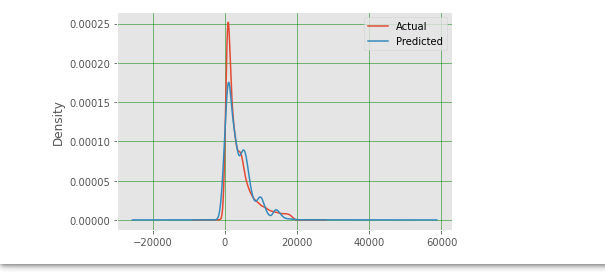
We will then proceed to build the model using a linear statistical model. This type of model is based on the assumption of a linear relationship between the input features and the output, and is suitable for predicting continuous variables such as diamond prices. We will select the appropriate model parameters and hyperparameters, and evaluate the model using a range of evaluation metrics including mean absolute error (MAE), mean squared error (MSE), and R-squared (R2).

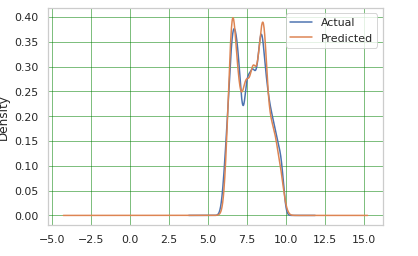
After building the model, we will evaluate its performance on the training data and on the test data, and compare the results to assess the model's ability to generalize to new, unseen data. We will also check the assumptions of linear regression to ensure the validity of the model.

Finally, we will use the model to make predictions on new, unseen data, and evaluate the accuracy of these predictions. This will allow us to determine how well the model is able to perform on real-world data, and whether it is a good fit for the problem at hand.

**Result**

**comparing the original and predicted values**





صورة تحتوي على نص

تم إنشاء الوصف تلقائياً

**Conclusion**

In conclusion, we have developed a machine learning model for predicting the prices of diamonds based on their characteristics. By using a dataset containing information about almost 54,000 diamonds, we were able to explore the relationships and patterns in the data, and build a model using a linear statistical model. We then evaluated the performance of the model using a range of evaluation metrics, including mean absolute error (MAE), mean squared error (MSE), and R-squared (R2).

The results of our evaluation showed that the model was able to achieve high scores for both the training data and the test data, with R2 scores of 0.94 and 0.95 respectively. The mean absolute error (MAE) and root mean squared error (RMSE) were also relatively low, indicating that the model was able to make accurate predictions about the prices of diamonds.

We also performed additional analysis by constructing residuals and checking the assumptions of linear regression, in order to ensure the validity and reliability of the model. Overall, these results demonstrate the effectiveness of our machine learning model in predicting diamond prices, and suggest that it could be a useful tool for industry professionals in making more informed and accurate pricing decisions.