# ASC61013 Data Modelling and Machine Intelligence

# Course Work – Diamond

Ankur Singh Gulia

Registration Number: - 220181437

# *Content*

# 1. Introduction to Coursework

The data set provided consists of 53,940 instances with 7 meta-attributes and 4 features in the raw CSV data file named diamonds_coursework.

The 4 features presenting the data in the raw CSV datasets are Id, cut, color, and clarity.

The 7 meta-attributes that have presented themselves are carat, table, depth, x(length), y(width), z(depth), and price.

This coursework aims to present a domain analysis of the data and further associate the analysis with data cleaning and pre-processing of the datasets. With the understanding of the domain, define a correlation between the features with the target feature. After the correlations have been confirmed, with the help of a regression machine learning method to predict the price of the diamond. Also, the creation of a decision tree methodology and feature engineering to predict the diamond price based on three classes: Low, Medium, and High.

Once the predictions are confirmed, cross-validation of the machine learning pipeline is supposed to be the next step. The application of learning curves and classification evaluation metrics, prove that the pipeline is effective at preventing underfitting as well as overfitting.

# 2. Domain Analysis – Diamonds

It is said that like snowflakes, no two diamonds are alike. Each diamond has its characteristics that draw the potential of the diamond. Prominently, certain standards in features define the quality and in turn the price of the diamond.

These standards are upheld by a certain organization in the diamond industry which goes by the name of the Gemological Institute of America (GIA). According to the dataset provided, it can be classified into different features.

1. Id

This dataset allocates an instance of a clear address that helps in the verification of that particular instance. It does not affect the diamond in any way but it helps understand the characteristics of a particular piece in the examination.

2. Cut

The cut is never referred to as the diamond's shape. Instead, it provides a reference to the diamond's proportions, symmetry, and polish. The better the diamond is cut, the greater it can reflect or refract light. According to the dataset and GIA, the feature has been divided as:

2.1 Ideal 2.2 Premium 2.3 Very good 2.4 Good 2.5 Fair

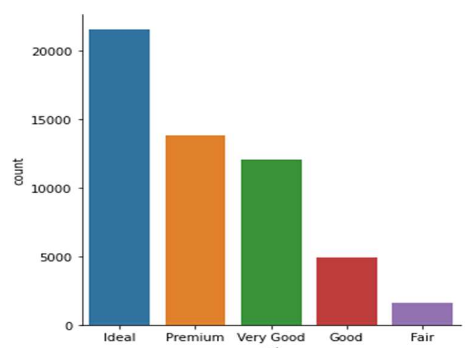



*Figure (1). Cut vs Count*

The graph shown in Figure (1) shows the frequency of the occurrence of the cut in the dataset provided.

3. Color

This feature, as the name suggests, defines the color of the diamond. Against the presumption that diamonds are clear, the fact remains that they can have subtle colors. According to the GIA standards, completely colorless diamonds are considered a rarity which makes them valuable. This feature has also been further categorized into 7 colors.

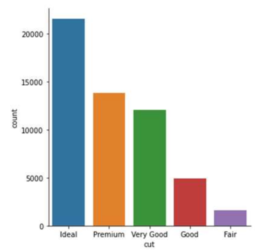



*Figure (2). Color vs Count*

4. Clarity

These diamonds are rocks formed under the ground with somewhat changes they went through which creates another feature that provides data to make up the instances. This feature entails assessing the quantity, size, relief, type, and location of the microscopic characteristics as well as their effect on the overall look of the stone.

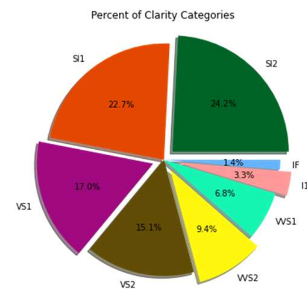



*Figure (3). Clarity Category*

Figure 3 clearly states the variation of clarity in the dataset.

The meta-attributes also play an essential role in defining a diamond. According to the dataset provided above, it is clear that there are 7 attributes:

- The first three attributes which describe the dimensions of the diamond are given as X(length), Y(width), and Z(depth).

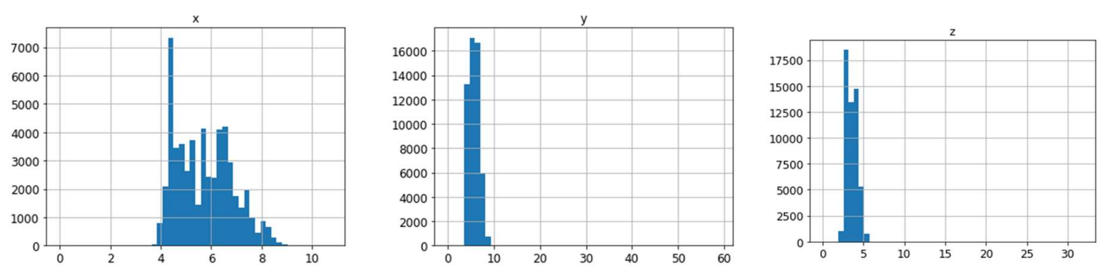



*Figure (4). Dimensions Vs Count*

The variation of the dimensions with count can be seen in Figure (4). For the above attributes, the range of variation, mean, and standard variation is given as:

| Description | X(length) | Y(Width) | Z(Depth) |
|---|---|---|---|
| Max Variation | 10.740000 | 58.900000 | 31.800000 |
| Mean | 5.732158 | 5.735530 | 3.539362 |
| Standard deviation | 1.122585 | 1.143023 | 0.706241 |

*Table 1*

- Carat

This attribute is the most important in defining the diamond as it is the unit of diamond measurement. Carat weight is often confused with the actual weight of the system but on the contrary, it depends on various factors such as density, shape, and formulation of the jewel. Carat weight can never be observed by the naked eye. Figure (5) displays the frequency of occurrence of a variety of carats.

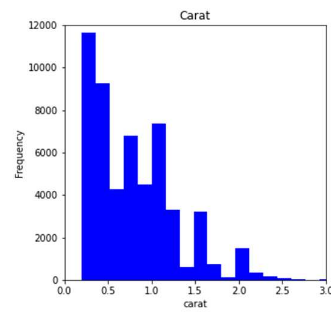



*Figure (5). Carat vs Count*

| Description | Range | Mean | Std Deviation |
|---|---|---|---|
| Carat | 5.01 – 0.20 | 0.798454 | 0.474411 |

*Table 2*

- Depth

The depth of a diamond is defined as the height or the distance from the table to the culet of the diamond. The information at the basic glance on the dataset shows:

| Description | Range | Mean | Std Deviation |
|---|---|---|---|
| Depth | 95.00 – 44.00 | 57.457981 | 2.233967 |

*Table 3*

- Table

A diamond's crown that extends from the top of the stone down to the girdle is known as the Table of the diamond. The dataset released the following information:

| Description | Range | Mean | Std Deviation |
|---|---|---|---|
| Depth | 95.00 – 44.00 | 57.457981 | 2.233967 |

*Table 4*

- Price

The last attribute but certainly not the least affecting the diamond. It is the market value that defines and differentiates between two diamonds. The instances show the variation amongst them for this attribute which can be tabulated and graphed as:

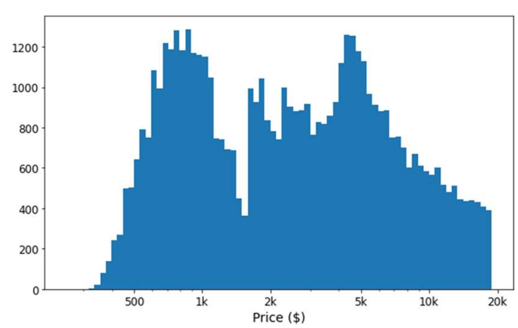



*Figure (6). Price vs Count*

| Description | Range | Mean | Std |
|---|---|---|---|
| Depth | 18823-326 | 3937.77 | 3993.28 |

## 3. Data Pre-processing

The dataset generally has certain unwanted or non-interactive data which does not allow the pipeline to work in accordance with the model which is required to work the dataset and features. By the book definition would be, a technique with which a raw data is converted to clean set of data is known as data Pre- processing.

For the dataset provided for this course work, it would be required to go through few steps to achieve a clean dataset. Before proceeding further, it is essential to consider the position the dataset

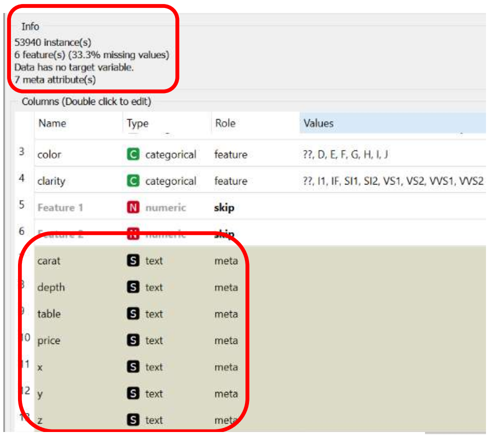



Figure (7). Meta-attributes and Features

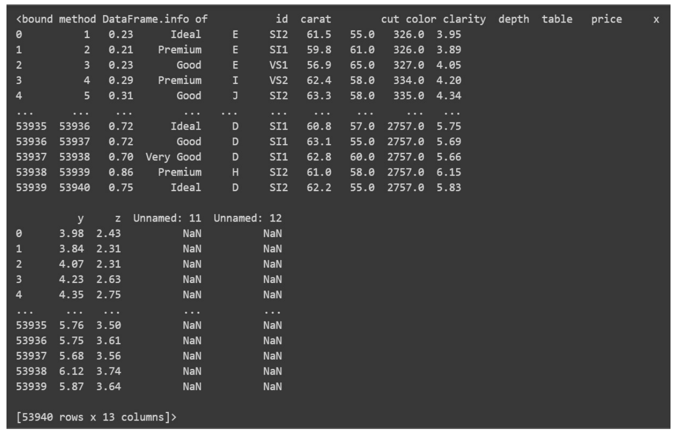



Figure (8). Data Description

It is clear from a quick analysis of the dataset that there is a certain missing data which needs to be taken care of as well as unwanted data such as: '??' to be removed. Also, the 7 meta-attributes affect the diamond characteristics, hence are required to be converted to features to be included in the pipeline. For this to work in a pipeline the data set must be pre-processed first.

- Editing Domain

`This is done to convert the meta-attributes into features so that they can be included in the pipeline to confirm their effect the on the target feature. In orange tool, it can be done with the help of Edit domain widget. As in Figure (9), it will now be reinterpreted as numeric value instead of text.

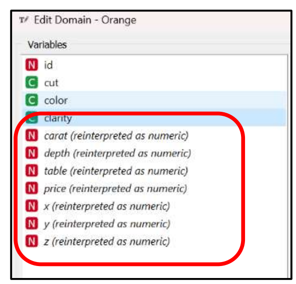



Figure (9). Edit Domain

This step can be skipped while working for python but the next applies for both orange and python.

- Imputing

This step is considered the base for the pipeline because it consists of cleaning of data to move ahead. If slight mistake or wanted data is removed, it would necessarily affect the decision making further down the pipeline.

It is achieved by choosing the impute widget and selecting all the column to dropping the unwanted or missing data. In this case, it was necessary to remove the unknown values.

For the imputing to work, the concerned features are selected to first drop. After confirming the selection of the features, action that is needed to be taken is selected. For instance, in this case, it is required to remove instances with unknow values and that is what is selected in the impute widget in the orange tool.

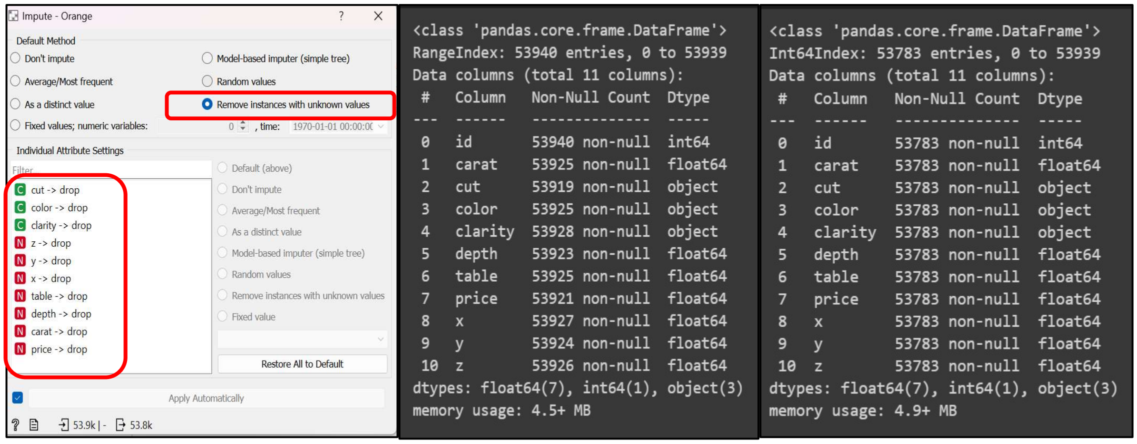



*Figure (10). Imputing data, and Effect of Imputing on data*

With respect to Figure (10), it clearly states that the non-null count has drop from different values to all the features having equal count of 53783.

## 4. Correlation

Machine learning models can be classified as good or based on the data that has been provided. The selection features contribute most to the quality of resulting model. The selection of such features that helps in predicting the variable more accurately is known as Feature selection.

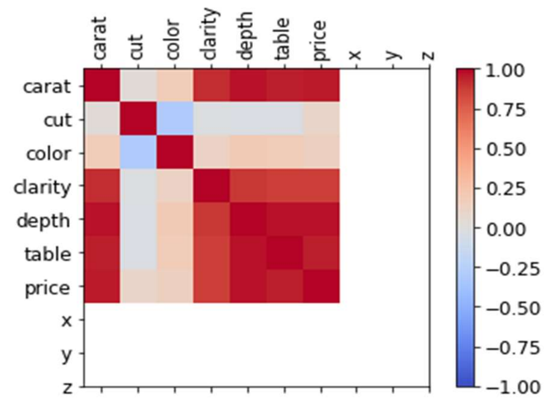



*Figure (11) Correlation Plot*

Feature selection is crucial task which is based on configuring the relation between the available features. With help of such relationships, one can identify the features important based on the target model.

| | carat | depth | table | price | x | y | z |
|---|---|---|---|---|---|---|---|
| carat | 1.000000 | 0.028246 | 0.181732 | 0.921578 | 0.975098 | 0.951694 | 0.953359 |
| depth | 0.028246 | 1.000000 | -0.295991 | -0.010723 | -0.025231 | -0.029280 | 0.094952 |
| table | 0.181732 | -0.295991 | 1.000000 | 0.127152 | 0.195527 | 0.183920 | 0.151075 |
| price | 0.921578 | -0.010723 | 0.127152 | 1.000000 | 0.884465 | 0.865419 | 0.861240 |
| x | 0.975098 | -0.025231 | 0.195527 | 0.884465 | 1.000000 | 0.974669 | 0.970746 |
| y | 0.951694 | -0.029280 | 0.183920 | 0.865419 | 0.974669 | 1.000000 | 0.951956 |
| z | 0.953359 | 0.094952 | 0.151075 | 0.861240 | 0.970746 | 0.951956 | 1.000000 |

*Figure (12). Correlation Table*

In case of Diamonds dataset, the correlations between features are given out in Fig (12) and Figure (11).

From the Correlation plot and the table, it is clearly understandable which features are having an impact where and how are they affecting the model and target

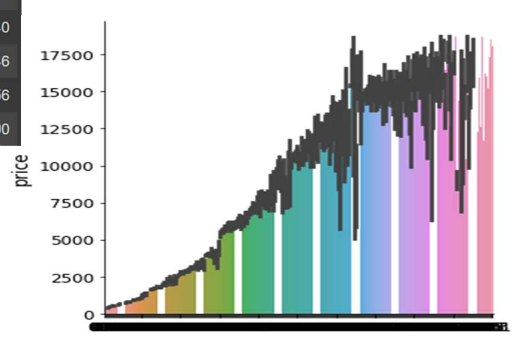



*Figure (13). Price vs Carat*

feature. As per the coursework, the designated target feature is given as Price.

Upon further studying the correlation table and plot, it justifies the variables affecting the target feature and further help in predicting the variable accurately. From the correlation plot, it can be deduced that the major feature affecting the price of diamond is carat.

The relation between both features can be easily interpreted by the graph. It suggests that as the carat weight increases the price of the diamond also increases. As we have already studied in the domain analysis that for a carat weight of 0.20 the price of the diamond can be considered as 326 as they create a minimum limit of the range. Whereas for the maximum range, it will be a diamond of 5.01 carat weight and 18823 price range.

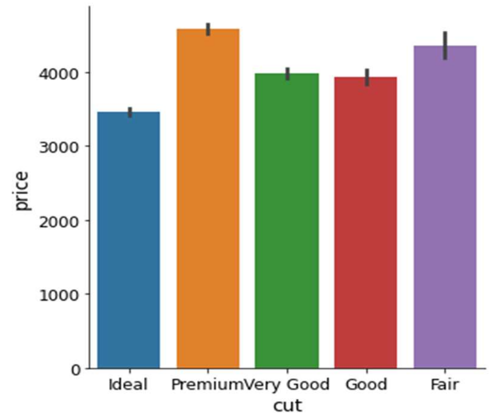



*Figure (14) Price vs Cut*

Other features that seem to be affecting the price is cut, which is divided into 5 categories namely ideal, premium, very good, good and fair. With each category, the decision of the price varies. Hence considering this feature to be a correlational feature to target price.

As per the domain analysis and study of the factors affecting the price of a diamond, clarity has been another factor among these.

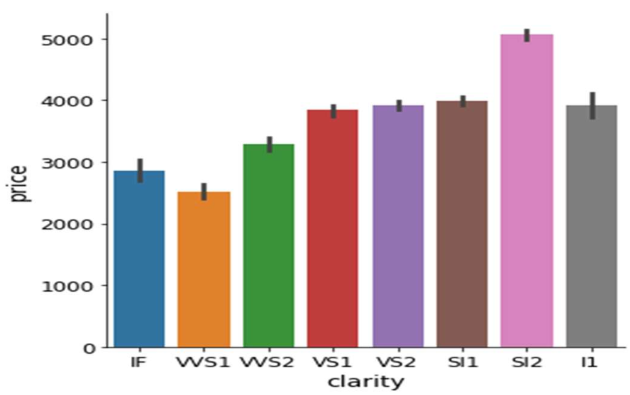


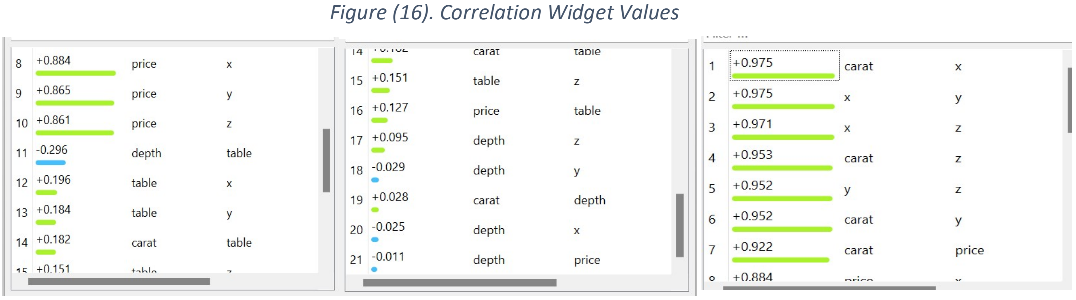

*Figure (15). Price vs Clarity*

The Figure (15) explain the price with respect to clarity in clear details. It is assumed that the for lower grades of clarity the increase in price is linear but as the grades of clarity increase the price also increases but the increase in non-linear.

Last categorical feature that affects the price of the diamond is color. Diamond is a at its core a rock which is formed due to high pressure and intense heat being provided to it. Almost always the diamond has some impurity or mineral mixed with it that makes up for the color of the diamond. Color is supposed to be the rarity of a diamond, and as always with increase in rarity, the price increases.

Apart from the above-mentioned correlation, the possible correlations can understand with the help of orange tool by adding a correlation widget to the pipeline. In this case, the Pearson correlation for all combinations are shown as:

*Figure (16). Correlation Widget Values*



| | | | |
|---|---|---|---|
| 8 | +0.884 | price | x |
| 9 | +0.865 | price | y |
| 10 | +0.861 | price | z |
| 11 | -0.296 | depth | table |
| 12 | +0.196 | table | x |
| 13 | +0.184 | table | y |
| 14 | +0.182 | carat | table |

| | | | |
|---|---|---|---|
| 15 | +0.151 | table | z |
| 16 | +0.127 | price | table |
| 17 | +0.095 | depth | z |
| 18 | -0.029 | depth | y |
| 19 | +0.028 | carat | depth |
| 20 | -0.025 | depth | x |
| 21 | -0.011 | depth | price |

| | | | |
|---|---|---|---|
| 1 | +0.975 | carat | x |
| 2 | +0.975 | x | y |
| 3 | +0.971 | x | z |
| 4 | +0.953 | carat | z |
| 5 | +0.952 | y | z |
| 6 | +0.952 | carat | y |
| 7 | +0.922 | carat | price |

As per the domain analysis and the study of the correlation from both python and orange. It can stated that the features that have strong relation with each other are:

1. **X(length)-Y(Width)-Z(Depth)**

As per Figure (12) and (16), the correlational values from the widget and the table suggest strong influence over each other, and also the domain analysis states that the dimension of diamond always have an impact over each other.

2. **Price – Carat**

According to the domain analysis and standard guidelines from the GIA, they have a strong co-dependency upon each other.

In accordance with the above discussion and the domain analysis, the three variables that closely correlate with the target price column are: **Carat, Cut & Clarity.**

## 5. Regression Model and Decision Tree

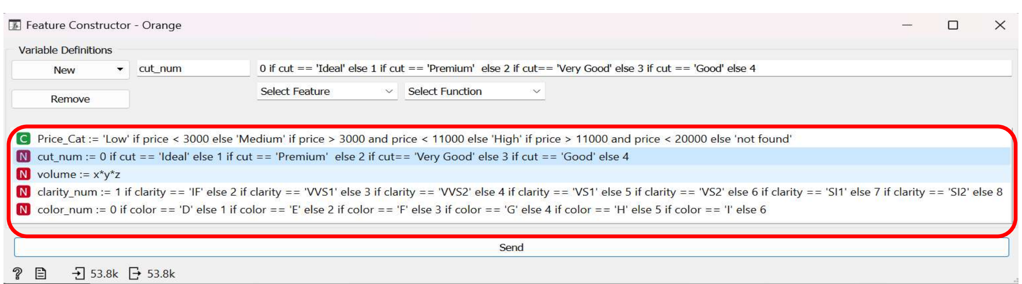
- Decision Tree Classifier

A Decision tree methodology is a data mining technique for establishing classification systems based on multiple covariates or for developing a prediction algorithm for a target variable. For a decision tree classifier to act on the target variable, it is first required to make the necessary changes to the data by the help of feature engineering.

➢ Feature Engineering

A process of creating a feature consisting the necessary requirements depending on the target to be applicable for a classifier. In this case, the target feature must be converted to classification of 3: Low, Medium and High.

It is also observed that converting the categorical features to numeric features will also affect the accuracy of the classifier as the data refines itself. Also, from the correlational analysis, it was observed that features X, Y, Z can be combined as well. Hence, the new creations are done with the help of feature constructor widget.



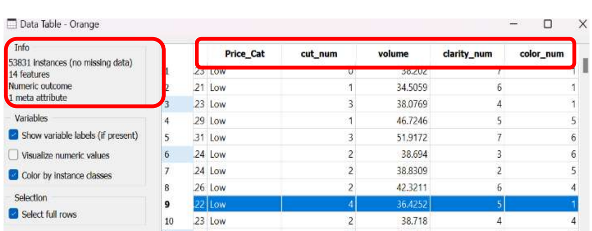




Figure (18). Data For Decision tree

*Figure (17). Feature Constructer*

As presented in the Figure (18), new features have been created. All that is left to do is to use a sampler on the required data and take up the Decision Tree Classifier.

Once the necessary columns are selected, a sample data and test data is created to configure the system for classifier. In this case, a ratio of 70:30 is taken for the current pipeline.

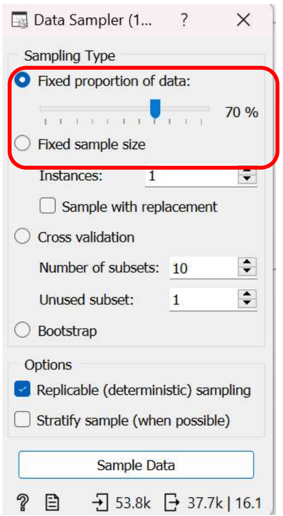



Figure (20). Data Sampler

Once the data sampler is ready for the pipeline, next step is to use this sample data for Tree Classifier and the rest remaining Data is used for testing the prediction. The pipeline for the classifier is given as:

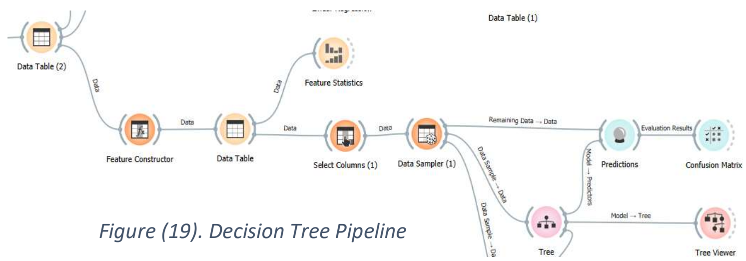



Figure (19). Decision Tree Pipeline

From the above pipeline, the predictions are passed through the confusion matrix to make the sure of underfitting and over fitting for the classification.

The confusion matrix provides an accuracy to the predictions on the system so as to validate the pipeline.

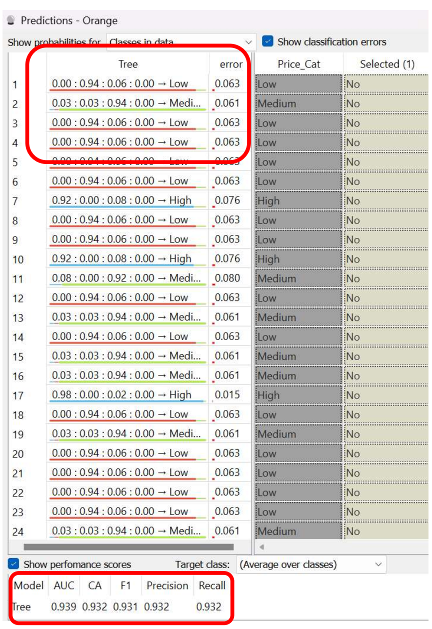



Figure (21). Predictions

The Prediction shows the individual instance's predicted classification based on the pipeline into the three categories created under the feature price_cat with presumptuous error.

Also, the precision for the classifier can also be seen in the same widget's window. As per the pipeline created, the system precision is given as 0.932.

The prediction only classifies but it is not nearly as capable to visualize the data. Hence a confusion matrix is used to give out the exact instances for the classification in terms of numbers as well as percentage.

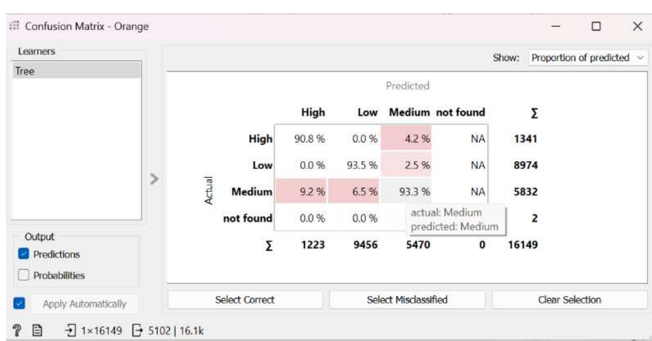



Figure (22). Confusion Matrix

The study of confusion matrix can be easily suggest out that accuracy of the pipeline and gives a better visualization of the dataset classification.

- Regression model

Looking at the domain analysis and the correlational study of the dataset, it would be better to use multiple regression and cross validate them to get an accurate result for the system/pipeline. But it is not required to use the feature engineering done on the pipeline.

For the regression model to bee applicable, the pipeline would look like:

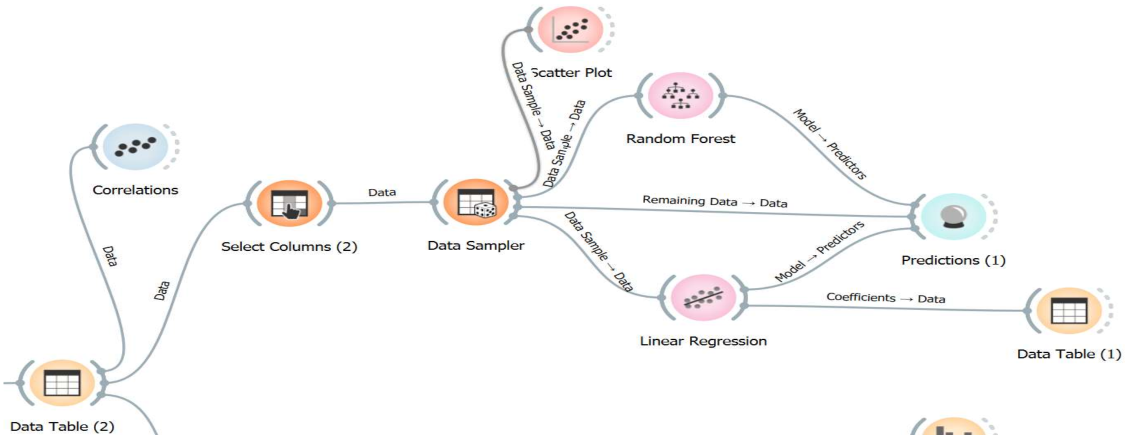



*Figure (23). Regression model pipeline*

As per the above pipeline, selecting the column that have an impact on the regression model based domain analysis and correlational study is the first step.Also, it could be assessed through the scatter plot between features. An example would be between price and carat.

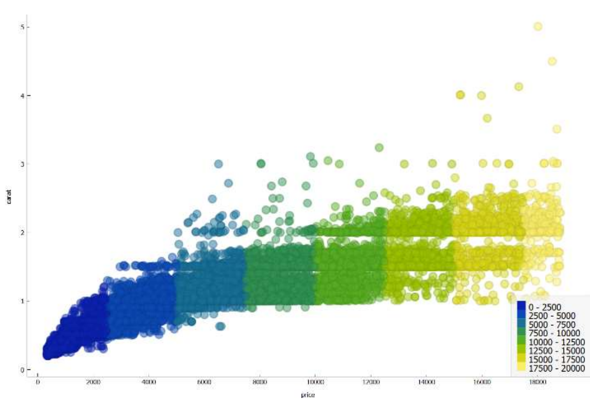



*Figure (24). Scatter plot: Price vs Carat*

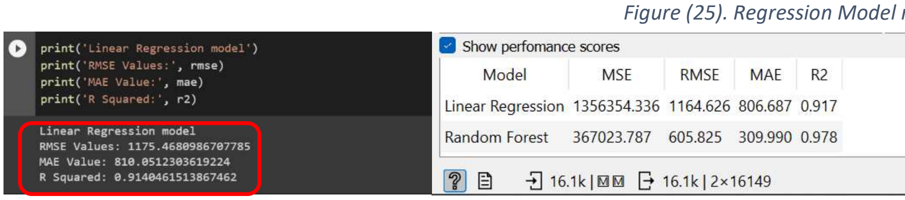
Through the scatter plot, it can be understood the variation with different carat and price.

Once that is accomplished, all that is left is to split the data into test and train dataset and to run the regression models through the prediction to figure out which one works better. The result of the predictions are visible in the Prediction widget.

The prediction stated that the random forest regression model is underfit based on the learning curve. The regression model for the linear regression seems to be a perfect fit as the validation curve and training curve work as parallel as plotted in graph. The values form orange tools are shown as:

Whereas the model for python shows the result as follows:

*Figure (25). Regression Model result*



```
print('Linear Regression model')
print('RMSE Values:', rmse)
print('MAE Value:', mae)
print('R Squared:', r2)

Linear Regression model
RMSE Values: 1175.4680986707785
MAE Value: 810.0512303619224
R Squared: 0.9140461513867462
```

Show perfomance scores

| Model | MSE | RMSE | MAE | R2 |
|---|---|---|---|---|
| Linear Regression | 1356354.336 | 1164.626 | 806.687 | 0.917 |
| Random Forest | 367023.787 | 605.825 | 309.990 | 0.978 |

16.1k | 16.1k | 2×16149

From the above regression model result it is clearly visible that the Linear regression model is much more effective for the dataset based on diamonds.

## 6. Cross Validation

For a Machine learning model, a pipeline is created based on the requirement. But to actually verify the accuracy and the legitimacy of the model, cross validation is done. Essentially, it is method that evaluates and compares the learning algorithm by dividing data in two segments: one to train the model and the other one is used to test the predictive capabilities if the model such as accuracy, error, etc.

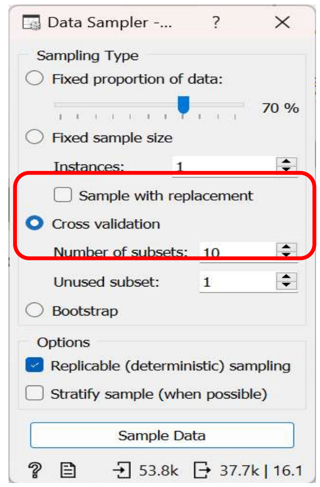



*Figure (26). Validation for data Sampler*

For this case, there are two methods, one is by using the orange tool's data sampler widget, wherein the subsets of the data changes to verify the pipeline.

And by changing the model, number of subset it, whether the accuracy changes or not.

The second method is by import the cross validation score from python. Which show the result for linear regression and random forest regression as:

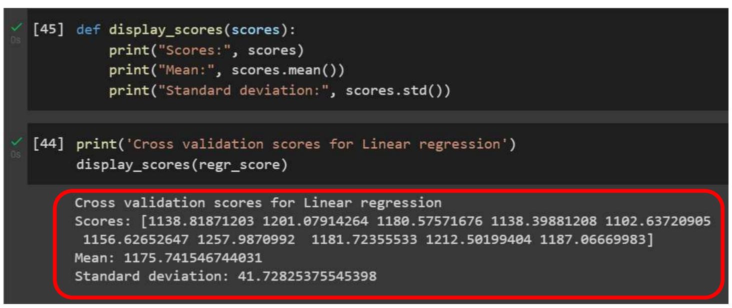



*Figure (27). Cross Validation scores for Linear regression*

The validation helps prevent bias and plays a crucial role in maintining the dataset required for the system to work. It does have an impact on the learning model as the test and train dataset must always be keep apart and have the exact ratio as mentioned during the time sampling.

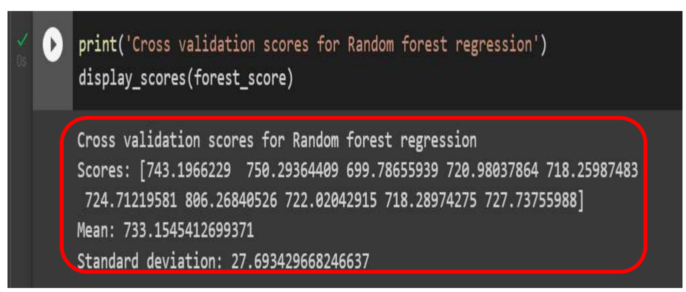



*Figure (28). Cross Validation for Random Forest regressor*

## 7. Learning Curves – Underfitting and Overfitting

Learning curves are the plots that represent the performance and the impact of the model at the training data set increases. The two major causes for error in the pipeline presented are:

Bias: it describes the model so that it makes simpler assumptions such that, for the model, it is easier to approximate.

Variance: It describes the variability of the model's prediction with the change in dataset used for training the model.

What the learning curves helps in representing the exact regressor to be used for prediction as the system requires a good fit.

For this case, 3 regressors were used and trained on the same data sets. The regressors used were Linear regression, Random Forest regression and Support Vector Machine regression. The plots are shown below:

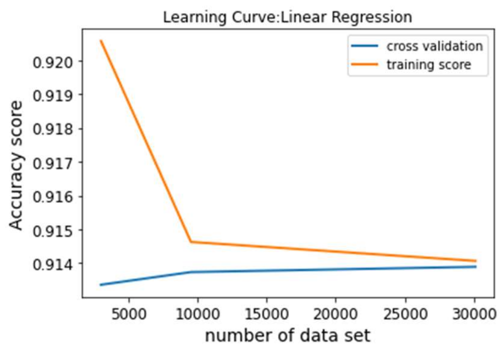



*Figure (29). Learning curve for Linear regression*

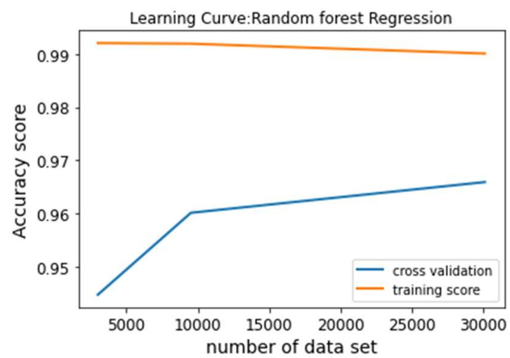



*Figure (30). Learning curve for Random Forest regression*

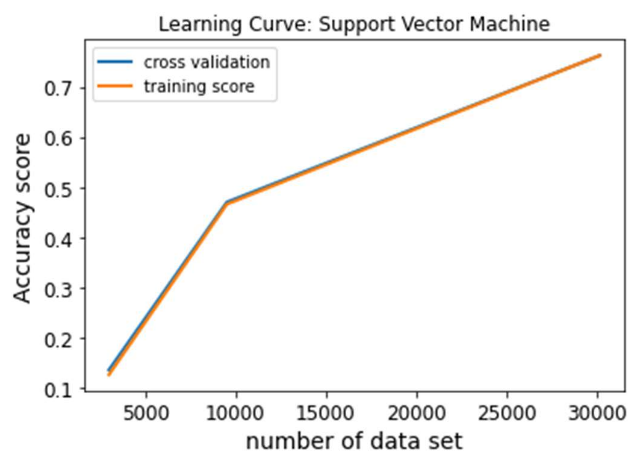



*Figure (31). Learning curve for Support Vector Machine Regression*

The reason for choosing the linear regression as fit for the machine learning model can be easily deduced from the above plots for all the regression models. If the training curve at any given point coincides with the validation, it is said to be an overfit.

If the training curve and the validation curve are running almost parallel to each other, it is considered as a good fit.

If the training curve and the validation curve are very far from each other it is considered to be an underfit case.

From the above plots and the knowledge based on the learning curve is can be justified that Linear regression is best suited for the Machine learning Model.

## 8. Conclusion

This course work consisted of creation of a complete pipeline for a data set based on study of 11 features of diamond. The process started with the domain analysis of the dataset and Diamond. Once the domain analysis was complete. It gave enough information to construct a pipeline based on the features for prediction of prices of diamond and classifying them in three categories: low, Medium and High.

This process started with editing the domain to make the meta-attribute into features. After the domain was complete. Next was data cleaning of unwanted or missing data instances. This was done to make the system completely based on complete datasets. With clean set of data, correlational study was done to achieve the particular features affecting the target feature price. It was noticed that X, Y, Z and depth was have been neglected as it made no impact on the target feature.

From here on, model was divided into two directions: First was creation a model that predict the price of a diamond. Second was the creation of a predictor that classified the prices of the diamonds into three categories.

For the first part of the pipeline, a sampler was taken and the data was divided into a ratio of 70:30. The bigger set was used to train the model and the smaller set to test the regression models in place. It was through cross validation and learning curve found that the Linear regression model was a good fit for the machine learning model.

For the second part, feature engineer was done to categorise the class of the price and then through the sampler, a decision tree classification was done to achieve the confusion matrix that explain the prediction with error as well. Also, the pipeline showed a ROC analysis to confirm that the tree classifier was the correct choice.

With the completion of a machine learning model, the price was predicted and the classification was done which helped in creating an effecting ML Model.