Nitte Meenakshi Institute of Technology

Title: Diamond Price Prediction

Submitted in partial fulfilment of LA2

Bachelor of Engineering

In

Computer Science and Engineering

Submitted by:

Ayushman Shivam	**Veeresh**
1NT19CS223	1NT19CS213
Uday Kiran Chari	**Vaibhav Jamwal**
1NT19CS204	1NT19CS209

Under the Guidance of:

Dr. Vani V

Department of Computer Science Engineering

Table of contents

1. Introduction

Natural diamonds are one of the precious stones bought to wear as jewellery or as investment as well. Diamonds are not that glittery and beautiful in their raw form. The rough diamond stone is normal looking stone as others are. The miners filters the mined soil to find the rough diamonds and sell them to the manufactures. The manufactures do the creative work on those rough stones. There are many shapes of polished diamonds available in the market. Manufactures plan and polish the rough diamonds based on the maximum financial gain from the polished product. Here, the diamond's price depends upon hundreds of parameters but mainly on 4 C's (Carat, Cut, Clarity, Color).

1.1 Motivation

Natural diamonds come in a variety of shapes and forms. To the inexperienced these stones might not appear precious and are often sold at inappropriate prices. As such we decided to create a model that is able to put a proper price to these diamonds.

1.2 Problem Domain

Diamonds before undergoing proper processing are very hard to identify as they appear rather dull. Often such raw diamonds get sold cheaply in bulks mostly at inappropriate prices. The price of these stones is decided by the vendors. In order to stabilise the price of these stones it is important to find an automated way free from human factors to make the decision for the quality of the stone.

The method designed application should provide a simple uniform and standardised means to estimating the price of the stones.

Since, the prediction of price depends entirely on the measured parameters any intentional or technical error will bring forth wrong predictions.

1.3 Aim and Objective

To train a machine learning model that is capable of predicting the accurate price of a natural diamonds based on certain selected features to standardise and simplify the sales of these stones.

2. Data Source and Data Quality

2.1 Dataset Used

The data set consists of 6 CSV files based on their shape (cushion, emerald, heart, oval, radiant, round) with the following attributes: -

Id (Object): unique identification number of diamond

Shape (String): shape of the diamond

Weight (float64): weight of the diamond in Carats (the bigger the weight the expensive it is)

Clarity (Object): clarity of the diamonds (FL, IF, VVS1, VVS2, VS1, VS2, SI1, SI2, SI3, I1, I2, I3)

Colour (Object): colour shade of the diamond (D, E, F, … Z)

Cut (Object): cutting level of the polished stone (Poor, Fair, Good, very Good, Excellent)

Polish (Object): polish level of the stone

Symmetry (Object): over all symmetry of the stone's shape

Fluorescence (Object): Fluorescence is the ability of certain chemicals to give off visible light after absorbing radiation which is not normally visible, such as ultraviolet light

Messurement (String): The measurement of the diamonds (format: length-breadth x depth).

Price(float64/String): The price of the diamond.

2.2 Data Pre-processing

Normalisation

- Min Max normalisation used

- **Converting the attribute data to required datatype**
  - Converting Price attribute to float64
  - Converting Messurement attribute to string removing unnecessary symbols (- x)

- **Deriving secondary attributes from original**
  - Length, Breadth and depth attribute were created from the original Messurements attribute using ReGex.

- **Dropping records with null values**
  - All records with Nan value were removed for simplicity

- **Dropping redundant columns**
  - Columns irrelevant to decision making (Messurements, Shape, Id) were dropped

- **Reorganizing the data**
  - The six data frames were concatenated into one to increase the size of sample data

- **Removing outliers**
  - The outlier was calculated for price using quantile method and outlier records were removed.

3. Methods & Models

3.1 Data Mining Questions

- Are there outliers in the data?
- How does the feature affect the outcome?
- Are the results usable?

3.2 Data Mining Algorithms

- Random Forest Regression

It is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.

To get a better understanding of the Random Forest algorithm, let's walk through the steps:

1. Pick at random *k* data points from the training set.
2. Build a decision tree associated to these *k* data points.
3. Choose the number *N* of trees you want to build and repeat steps 1 and 2.
4. For a new data point, make each one of your *N*-tree trees predict the value of *y* for the data point in question and assign the new data point to the average across all of the predicted *y* values.

A Random Forest Regression model is powerful and accurate. It usually performs great on many problems, including features with non-linear relationships.

3.3 Data Mining Models

- Random Forest Regressor

A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. It makes use of the random forest regression algorithm. It is the most accurate model available when performing regression tasks and hence was selected for this project.

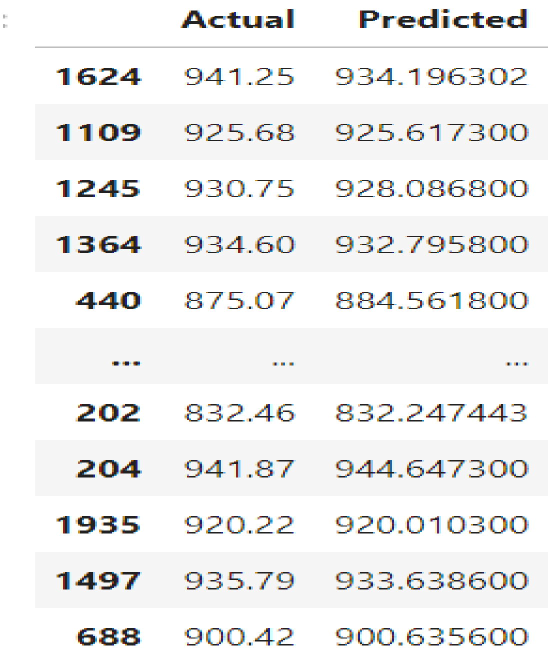
4. Model Evaluation & Discussion

The following metrics to evaluate the performance of our model :

- Accuracy for test set (0.718)

- Accuracy for the training set (0.9602)
- Mean absolute error (14.15)
- Mean squared error (805.6)
- Accuracy of the model (98.33%)

The disparity between test set and training set accuracy indicates that the data has been overfitted to the training set. The values of the mean square and mean absolute error indicate that the model is not able to predict all values with accuracy. Future attempts must be made to minimise the error values.



	Actual	Predicted
1624	941.25	934.196302
1109	925.68	925.617300
1245	930.75	928.086800
1364	934.60	932.795800
440	875.07	884.561800
...	...	...
202	832.46	832.247443
204	941.87	944.647300
1935	920.22	920.010300
1497	935.79	933.638600
688	900.42	900.635600

Finally, the accuracy of the model leaves very little difference between the predicted and actual values as such it can be used in the field. However, more efforts must be paid to make the model better.

5. Conclusion & Future Direction

We were able to train a model using the Random Forest Regressor to perform the prediction task with a 98.33% accuracy. This number being exemplary considering the beginning stages of development.

However, since the data has been overfitted there is a possibility that the model might not perform as well to the new data.
As such more data must be gathered in order to train the model to a better degree.

Keeping the future development in mind, we may extend our application using the following new features: -

- In the future we may make the price prediction unique to each shape making use of some extra features like regional availability etc.

- There is also room for adding a computer vision-based model that will be able to recognise the shape of the stone from its image and correctly classify it into the right category (heart, oval etc).

- The final goal would be to extract features of the stone using high resolution images taken from 360deg without the need of any tools. However, this is likely to require much longer a time.

6. Reflection Portfolio

Going through this project we were able learn the following things:

- Use of python libraries namely pandas, numpy, sklearn for data manipulation and evaluation.
- Use of plotting libraries like matplotlib, plotly, seaborn to visually represent various forms of data.
- Performing Exploratory data analysis.
- Using functions to minimise redundant code.
- Applying normalisation and other pre-processing techniques to data.
- Calculating outlier using quantile method.
- Evaluating the importance of each feature to the decision making.
- Splitting the data into training and test set
- Identifying overfitting and underfitting.
- Selecting the best machine learning problem according to the problem by evaluating accuracies.

7. References

- Kaggle
- Tutorials Point
- One Stop Data Analysis
- Ask Python

Appendices

a. Link to the dataset chosen

https://www.kaggle.com/harshitlakhani/natural-diamonds-prices-images

b. Python Codes Implementation

https://github.com/Despicabug/Diamonds-Classification