Literature Survey

Problem statement: Diamond price Prediction and analysis

Diamonds are the precious stone consisting of a clear crystalline form of pure carbon, they are the hardest gemstones known to man. They are formed deep within the Earth about 100 miles or so below the surface in the upper mantle, they are Rare because of the Incredibly powerful forces needed to create them.

Looking into the pricing history of diamonds : In a span of 60 years i.e. 1960-2020, the price has risen from 2700 USD to 28400 USD per carat.

As the popular saying goes:

“Diamonds are a girl’s best friend and an investor’s best investment!”

They are an excellent source of investment, most popular and costliest jewelry item and have variety of usage in medical, automobile and other industries.

These exceptional qualities and the great impact of diamond in economic market definitely demands for their price prediction and exploration of various other factors. It would not only help investors and companies but also help in the understanding of rise and fall of economic market.

Thus, it is a great driving force to perform in depth Data analysis to get required results for proper availability and pricing standards of this extraordinary stone.

However, the prediction process is difficult due to the wide variation in the diamond stones sizes and characteristics

Paper 1:

Title : Implementing Data Mining Methods to Predict Diamond Prices

Source : By – Jose M. Pena Marmolejos

Year of Publish : 2018

In this Paper, JMP applies 3 techniques namely linear regression, neural networks and M5P algorithm to predict the price of diamonds. Further in the report, the results of these 3 techniques are compared to find which one performs better. The analysis of the data has been done by utilizing the Python3 programming language. Weka toolkit (a collection of ML algorithms for data mining tasks, such as classification, regression, clustering, and visualization) has been used to experiment with these three data mining algorithms. Extensive work is done for accurate prediction of diamond prices. The technique we

plan to use as of now is multiple linear regression and neural networks and programming in Python language.

Assumption made here is that there is a relationship between price and carat, nonetheless it can be observed that the trend appears to fade away, this might signify higher price volatility for heavier diamonds, especially those above 2.5 carats.

The M5P model produced better overall results than linear regression and neural . There were small number of outliers that were affecting negatively on the performance of the model.

Dimensionality reduction by high correlation has proved to be a useful technique for better performance. This research paper is very productive and informative. It helps in understanding the key procedure of analysis & prediction. Further, it also brings into light some new methods and techniques which could be very useful for our work. The importance of cleaning and removal of outliers can be understood as it created a bad impact on the model.

Paper 2:

Title : Subjectivity of Diamond Prices in Online Retail: Insights from a Data Mining Study

Source : By – Stanislav Mamonov and Tamilla Triantoro

Year of Publish : Oct 2017

The goal of this research paper is to explore the relationship between the physical properties of diamonds and diamond prices to understand the degree to which diamond prices are determined by the physical characteristics. The dataset is obtained from one of the largest retailers that offers one of the largest collections of diamonds for sale. To assess the quality of data using linear regression we build a linear model which however results in an mean percentage error of 30.4% which is quite high from our perspective of building a predictive model. Hence, we resort to non-linear models. Diamond price is a continuous interval target variable that is based on a ratio scale, hence we employed the following prediction data mining techniques: decision forest, boosted decision tree, and artificial neural network.

On building and analyzing the three models we observe that the mean percentage error comes out to be 15.8%,23% and 24.3% respectively which is an improvement over the linear model but is still not good enough. We analyze the diamonds that were sold over a span of next 5 weeks and found that 96% of the diamonds that were sold are smaller than 2.5 carats in size. Thus narrowing our scope of analysis to diamonds weighing between 0.2 and 2.5

carat we observe the following percentage error with respect to multi-regression, decision tree, boosted decision tree and artificial neural networks respectively -->9.9%,13%,22,6% and 8.2%.We can thus see that artificial neural networks appears to the most optimal model for diamond price prediction. The results of the analysis suggest that the size of a diamond, as indicated by weight, width, length and height, along with color, clarity and shape influence the model the most.

Prior research has suggested that consumers prefer diamonds of specific sizes. The half-carat and full carat size diamonds are particularly popular. On analyzing the box plot of prices of diamonds of various sizes we observe that diamonds having weights 0.5 carats and 1 carats tend to fetch the most prices. This paper besides showing how diamond prices tend to vary with different physical stats provides proof that linear regression is not optimal for diamond price prediction, rather artificial neural networks turns out to be the most efficient model for this task.

Limitation of linear regression for this dataset is clearly visible.

Paper 3:

Title : State of the Art Diamond Price Predictions using Neural Networks

Source : By – Charley Yejia Zhang

The goal in this paper was to predict the price of diamonds using features such as carat, clarity, color, cut, depth, table length, x, y, and z axis lengths in millimeters, there is a very strong positive correlation between {x, y, z, carat} and {price}. Other features showed weak to almost no correlation with price, although there is no direct correlation that can be seen in data, other features could have information embedded in them that will help predict price with higher accuracy so they were included.

The following feature vectors were taken and their performances on a 2 hidden layer neural network with fully connected layers of 100 and 100 nodes were evaluated, and the set of features with highest r2 scores were selected

{cu(1h),co(1h),cl(1h),d,t}->0.5102,{ca,x,y,z} ->0.7946 , {ca,d,t,x,y,z}->0.8112

{cu(num),co(num),cl(num),ca,d,t,x,y,z}->0.9775

With the feature vector decided, the next part is to choose the best NN architecture by training different networks and evaluating their performances, the r2 scores of the different set of layers are as follows,

{100, 100, 50}->-0.897, {100} ->,{50,100}->0.9762 ,{100,100}

->0.9762,{200,100}->0.9764. Deeper networks did not improve the performance, so the number of hidden layers were 2, so the chosen hidden layers have a {200,100} network.

With feature transformation (log prices) a r2 score of 0.99241 was observed, without feature transformation (raw prices) a r2 score of 0.98111 was observed. Neural networks achieved much better r2 score compared to other models such as Linear SVR Baseline and Linear Ridge Regression Baseline which achieved r2 score of -0.13001 and -0.14371 respectively.

Specific Conclusions:

The refined problem statement:

Diamond price prediction through various data models and EDA.

Here, we would be using different models to predict the prices and look out for the best fit model. Apart from price prediction, we would also take out informative insights .It will tell us more about the peoples choice of diamonds, like most popular colour, most expensive designs, most economic diamonds etc.

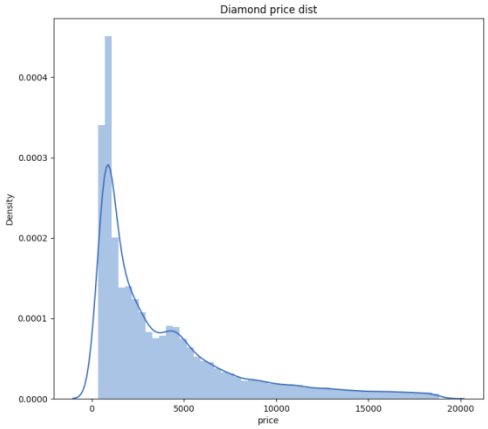
EDA:

Number of records in dataset are 53940 and Features are 10. All the columns are completely filled and no missing values are found.

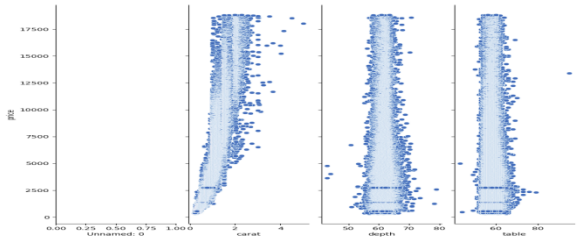
Although we didn’t find any missing values present in the dataset, some wrong data entries were discovered.

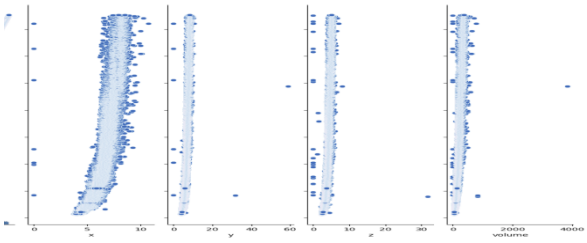
8,7,20 rows where x,y,z dimensions respectively were mentioned as 0. These data rows were dropped for accurate analysis.

Upon plotting density plot for diamond prices we find that most of the diamonds have prices less 5000 USD.



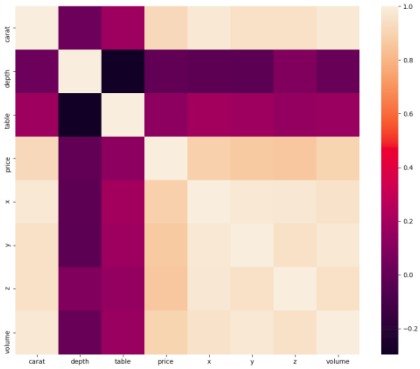
On plotting prices vs various attributes interesting facts were established.





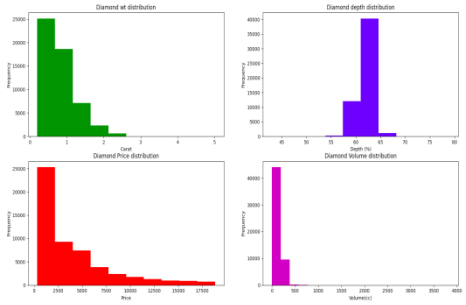
- We find that the diamond prices increase steeply as the carat weight increases uptil 2 carats after which increase is slower.
- We find that most of diamonds have a mean depth of 60 and diamond prices is generally found to be independent of the depth.
- Diamond prices also tend to show slow increase with the volume.

HeatMap shows various correlation trends.



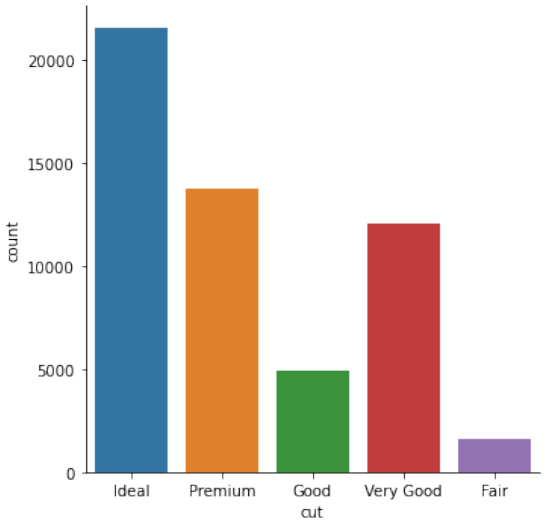
- Diamond prices show high correlation with carat weight.
- Diamond prices also show a decent correlation with the volume.
- Diamond prices also show a slightly negative correlation with table width(width of top of diamond relative to widest point).

Frequency distribution of diamonds reveal that



- Most of the diamonds have weight less than 1 carat.
- Most of the diamonds have a depth of 60%
- Most of the diamonds have prices less than 2500 USD.

Cut Count:



Most diamonds present are of Ideal cut, whereas Fair cut is the least.

Approach & Future Work:

After meaningful insights and conclusions we would be finally working upon data model for prediction.

The papers we referred compared various machine learning algorithm and then came on the conclusion that neural networks performs best for price prediction. Henceforth,we will be implementing neural networks to predict the prices of diamonds based on various attributes and also establish the which attributes contribute to diamonds either positively or negatively.