Diamond Exploratory Analysis and Price Prediction

Utkarsh Verma
B.Tech CSE
PES University
Bangalore, Karnataka
utkarshofficio@gmail.com

Anvesh Reddy
B.Tech CSE
PES University
Bangalore, Karnataka
anveshreddy595@gmail.com

Aman Prasad
B.Tech CSE
PES University
Bangalore, Karnataka
amanprsd0599@gmail.com

Shashank Barole
B.Tech CSE
PES University
Bangalore, Karnataka
shashankbarole410@gmail.com

***Abstract*—We consider a dataset containing 53000 rows and 10 columns describing different features of a diamond. We do a exploratory analysis on the dataset and plot our observations.Consequently we extend our analysis further to predicting diamond prices using the attributes so analysed.**

I. Introduction

Diamonds play and interesting and dominating role in a countries economy.Since old times we have had people professionally analysing diamonds.These diamond analysts go through hundreds of diamond daily analysing attributes and predicting their prices.However the common folk does not have the expertise nor the experience to analyse the diamonds so efficiently often leading to people poorly negotiating the prices and often end up paying much more than what the diamonds may be worth.However,the computer technology today has given us the resources to mimic human intelligence through various machine learning model and implement it to the common folks giving them a proper resource to a fair negotiation chance.One such model is the artificial neural networks(ANN).The artificial neural networks mimic a pattern similar to that of human neurons.We analyse the attributes provides through the dataset https://www.kaggle.com/shivam2503/diamonds and do an exploratory analysis whose results we further extend to our neural networks model to build an efficient price predictor.

II. Previous Work

A. Implementing Data Mining Methods to Predict Diamond Prices by Jose M. Pena Marmolejos (year 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 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, however it can be observed that this trend seems to diminish, this signifies higher price variation for heavier diamonds. The M5P model produced better accuracy than lr(linear regression) and nn(neural networks) .The presence of outliers were affecting the performance of the model. Dimensionality reduction 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.

B. Subjectivity of Diamond Prices in Online Retail: Insights from a Data Mining Study by Stanislav Mamonov and Tamilla Triantoro (year 2017)

The goal of this research paper is to explore the relationship between the different attributes of diamonds and prices of diamonds to understand by how much prices of the diamond are influenced by the each attribute of the diamond and which influences it the most. The dataset is obtained from one of the largest retailers which offers a healthy collection of different attributes of the iamond. To assess the quality of data. we use lr(linear regression) to build a model which however results in an mean percentage error of 30.4which is quite high from our perspective of building a predictive model. Hence, we resort to non-linear models. we employed the following algorithms for prediction: decision trees, boosted decision trees, and Neural networks. On building and analyzing the three models we observe that the mean percentage error comes out to be 15.8respectively 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 96were sold are smaller than 2.5 carats in size. Thus narrowing our scope of analysis to diamonds weighing between 0.2 and 2.5carat we observe the following percentage error with respect to multi-regression, decision tree, boosted decision tree and artificial neural networks respectively – 9.9and 8.2artificial neural networks appears to the most optimal model for diamond price prediction. The results of the analysis suggest that these features influence the diamond most size of a diamond,weight, width, length,height,color, clarity,shape helps in predicting the price compared to other features. 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.

C. State of the Art Diamond Price Predictions using Neural Networks by Charley Yejia Zhang

The goal in this paper was to predict the price of diamonds using various features like carat, clarity, color, cut, depth, table, length, x, y,z.there is a positive correlation between x,y, z, carat and price. Other features showed no correlation with price,other features could also help us to predict price with higher accuracy so they were included.The following feature vectors were taken and their performances on a 4 layer neural network with densely connected layers which has 2 hidden layers,1 input and 1 output 2 hidden layers have 100 neurons each were evaluated, and the set of features with highest r2 scores were selected (cut,color,clarity,depth,table) these four features have an r2 score of 0.5102 and (carat,x,y,z) has r2 score of 0.7946 , (carat,depth,table,x,y,z) has r2 score of 0.8112 (cut,color,clarity,carat,depth,table,x,y,z) in which categorical features were encoded with a label encoder unlike one hot encoding before has an r2 score of 0.9775 With the feature vector decided, the next part is to choose the best set of neural network architecture by training various set of networks with different number of hidden layers and various learning rates and evaluating their performances based on r2 scores of the different set of layers are as follows, (100, 100, 50)=-0.897,(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.

D. Problem Statement We address

We would be analysing different models to predict the prices and look out for the best fit model. Using this best fit model we will predict the prices based on diamond characteristics.Further the EDA will further help us to extract interesting facts about diamonds in market.

III. Proposed solution

We begin with an exploratory data analysis to gain insights into various attributes contributing to the diamond characteristics.We do a density plot for diamond prices and check for the most common price for the diamonds in market this helps us to separate diamonds based on quality.Secondly we do prices vs diamond attribute plot which gives us interesting insights into the data.Further we can segregate diamonds into various categories using histograms because histogram will separate diamonds into various classes based how frequently that diamond is found.Finally heat map provides us insights into various correlation that exists between attribute.Once we have enough insights we proceed with building the model.On analysing the different models used in previous works we find that neural networks happens to yield the best results.Thus we decide to implement neural networks to generate our model.We begin by preprocessing the data.We first label the categorical columns(cut,clarity and color) using a labelEncoder()(LabelEncoder transforms the categories into numbers.).We observe that x,y,z can be replaced by volume(xyz),thus we consolidate x,y,z to volume(xyz).After this we normalize depth,xyz,table columns.Once we are done with preprocessing we move over to model building and training the data.First we split the columns into feature of diamond (X) and price of diamond (Y).Then we proceed by splitting the dataset into train and and test set using splitting ratio of 0.3.We then move on to building the neural network model.We use keras library to build our neural network.We use different sets of layers and neurons to define our architecture and based on the best result obtained we layout of architecture.On observing different architecture we finally decide to use 3 hidden layers which has 512,256,128 neurons each to achieve the best r2 score.We use mean absolute percentage error as our loss function.

$$M = 1/n * \sum((Y_{true} - Y_{pred})/Y_{true})$$

We use r2 score to evaluate our model.

$$R2 = 1 - SS_{RES}/SS_{tot}$$

where:

$$SS_{res} = \sum(y_{true} - y_{pred})^2$$

$$SS_{tot} = \sum(y_{true} - y_{mean})^2$$

We initially attempted to use MSE but due to the difference in prices absolute mean square percentage error was used.We used adam optimizer with learning rate of 0.006 and 100 epochs with a batch size of 100 and relu activation function for hidden layers to acheive the optimal result.

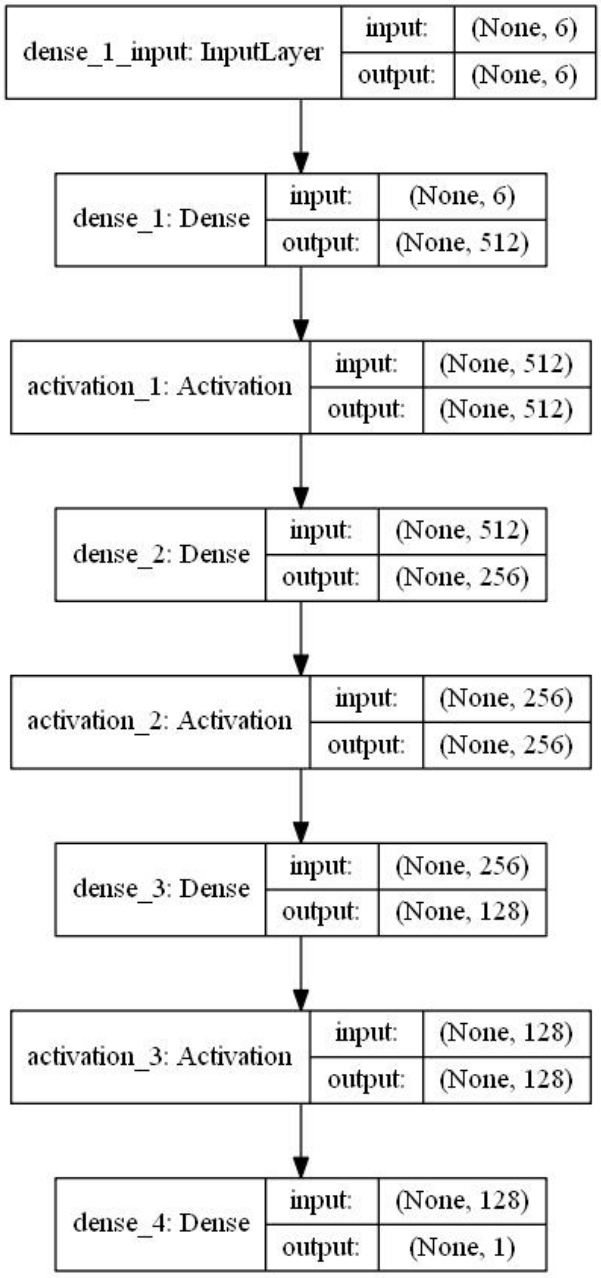

Fig. 1. Neural network architecture

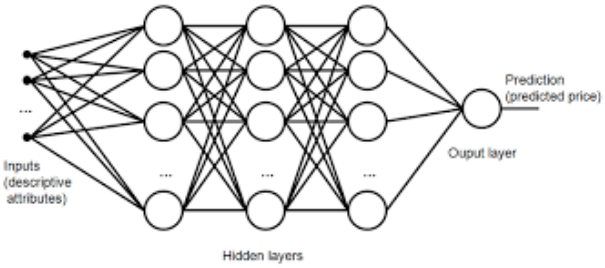

Fig. 2. Neural network

IV. Experimental results

A. EDA

On plotting the graph for the price distribution we find that most of the diamonds have prices less 5000 USD.

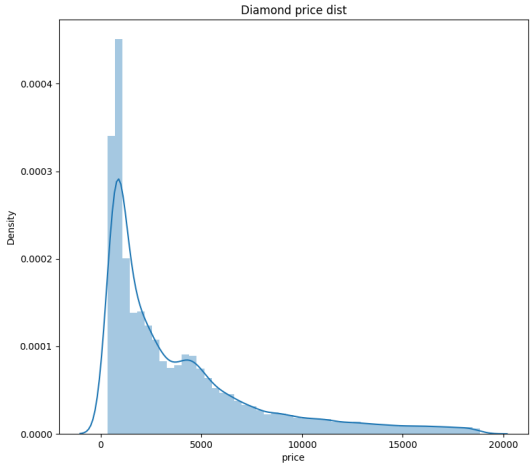

Fig. 3. Price distribution

On plotting prices vs various attributes interesting facts were established.

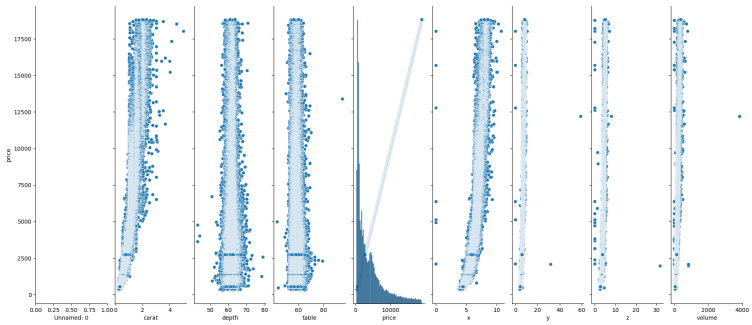

Fig. 4. Price vs attributes

- 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.

Frequency distribution of diamonds reveal that

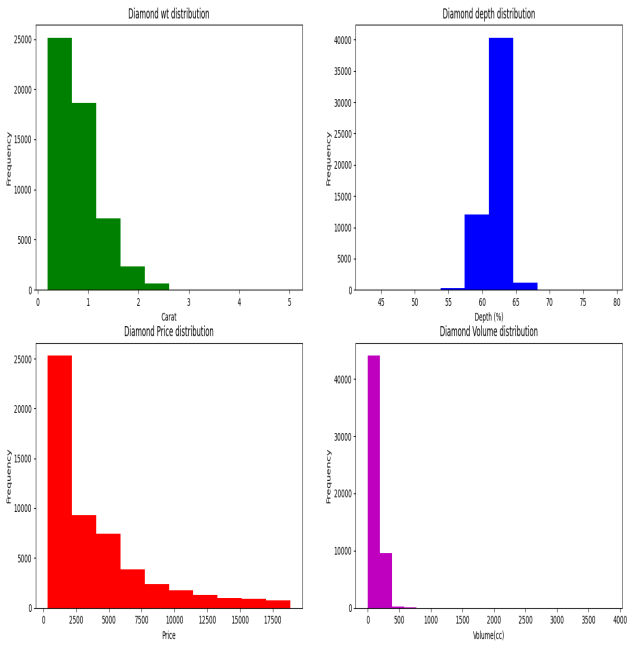

Fig. 5. frequency distribution

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

HeatMap shows various correlation trends.

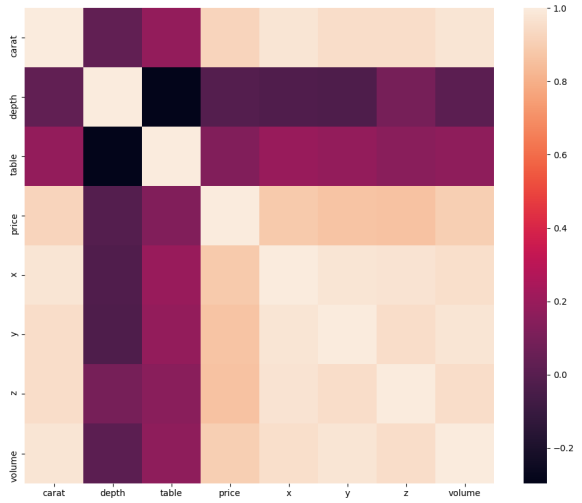

Fig. 6. Heat Map

- 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).

B. ANN model observations

We use different set of hidden layers and learning rate and get the following data

TABLE I
DIFFERENT NN ARCHITECTURE PERFORMACES

Layer Description	***Performace(R2 score)***
(512,256,128)	0.968
(512,256)	0.964
(512)	0.90

TABLE II
LEARNING RATE VS PERFORMANCE

Learning Rate	***Performace(R2 score)***
0.006	0.968
0.01	0.94
0.001	0.92

On building the model and testing it we get the following results:-

- r2 score(Test set)=0.968
- r2 score(Train set)=0.9723

We find the following relation between r2 score vs epochs

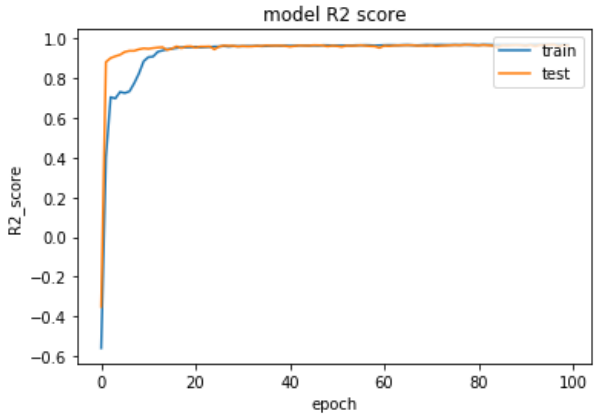

Fig. 7. R2 score vs Epochs

We find the following relation between mean absolute percentage error vs epochs

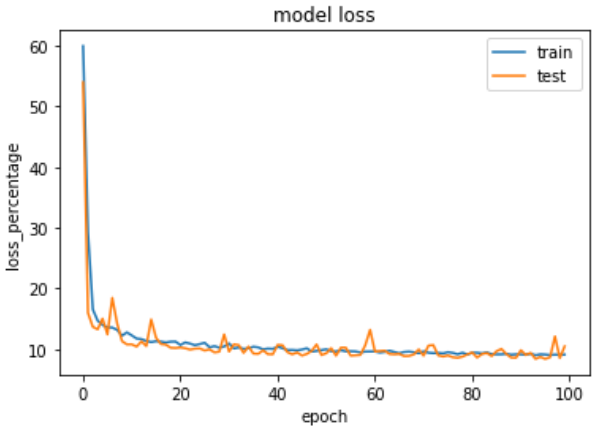

Fig. 8. mean absolute percentage error vs epochs

V. Conclusion

The EDA and the insights inferred helped us to build a model based on ANN networks which can very efficiently analyse a given diamond based on its attributes and predict the prices with a high r2 score and low percentage error.

Acknowledgment

We would first like to extend our at most gratitude to Dr. Gowri Srinivasa for her guidance.We would also like to thank Dr. Mamatha HR for her support as well as the entire Data Analytics faculty for their guidance.

References

- Implementing Data Mining Methods to Predict Diamond Prices by Jose M. Pena Marmolejos 2018
- Subjectivity of Diamond Prices in Online Retail: Insights from a Data Mining Study by Stanislav Mamonov and Tamilla Triantoro
- SOA Diamond Price Predictions using Neural Networks by Charley Yejia Zhang