Machine Learning Algorithms for predicting the diamond price

Sai Manjith Matta1 Soujanya Myla2 Anshul Reddy Anandula3 Sriharsha Katta4

700730744 700731894 700734890 700728960

**Abstract –**

Diamonds are one of the most valuable naturally occurring carbon compounds. These are in high demand due to their monetary rewards. However, unlike other precious metals like gold and silver, determining the price of a diamond is not straight forward as it depends on many factors including carat, cut, color, dimensions, depth percentage and Table width. To put in simple words, a diamond that is cut with high precision has more monetary value than one with less precision. Having a system that could help in deciding the intrinsic value of a diamond gives an immense opportunity to both business and customers in transacting diamonds. Through this project, we have built machine learning models to predict the price of the diamond. The dataset has been extracted from Kaggle. Dataset contains price and attributes of 50,000 diamonds. Attributes or features provided in the dataset are Price, carat, cut, Color, Clarity, Dimensions, Depth and Width. More details about the dataset are explained in the following sections in detail. Linear Regression, Decision Trees and K-Nearest Neighbours algorithms are trained using the dataset. Apart from that, hyper parameter tuning is also performed, and best model is reported. More details of the experiments performed are explained in the proposed framework section. With hyper parameter tuning of K Nearest Neighbours algorithm, we were able to produce 97% of R2 Score. Before training of the algorithms, Exploratory data analysis has also been performed along with feature engineering techniques. EDA helps us to build a robust model as we can know more about the dataset. In EDA, several plots were visualized in-order to better understand the data. As a part of feature engineering, techniques like label encoding, Standard Scaling were also applied to data. All the code and instructions on how to train are documented and well and uploaded to GitHub repository which can be accessed from [1].

Keywords—

Diamond price prediction, Machine Learning, Regression, Supervised Machine learning, Decision Trees, K Nearest Neighbours, Linear Regression.

1. **Introduction**

Diamonds are one of the most valuable metals in the world. Diamond is a pure form of carbon which has high monetary value in the market. Diamonds are primarily used as ornaments. Apart from ornaments, diamonds are also used for other purposes like cutting metals, glass. It is one of the toughest naturally occurring metal in the world. Chemical resistance and thermal conductivity are also very high for diamonds. Shining is also one another feature of diamond. As diamonds are highly durable, they last for very long without having any scratches. Diamonds are also of various types based on how they are made. Some of the factors influencing them are quality, cut, dimensions. Despite having these many advantages with diamonds, one of the biggest problems with diamonds are how they are priced. Unlike other precious metals like gold and silver whose price is merely determined by a linear relationship with weight and quality, these are dependent on several factors. Apart from weight and quality, price of a diamond also depends on factors like cut, colour, depth, width, and dimensions.

The motive of this project is to predict the price of a diamond using machine learning algorithms. Price forecasting is the process of using historical data on a given product to predict the future price. In recent times, price forecasting has been widely used in all domains to predict prices of various commodities. As the price is a continuous variable, unlike classification algorithms which can be applied on discrete variables, regression algorithms must be applied to predict the price of a diamond. Regression is a statistical procedure which is used to find out the relationship between dependant variable and independent variable. Here, in our problem, price of the diamond is dependant variable as it is heavily dependant on various other factors. In contrast, features like carat, colour, depth, dimensions and width are independent features as they don’t depend on each other. Regression analysis is typically a supervised machine learning technique. Supervised learning is a technique in machine learning in which algorithms are trained on labelled data. While training of the models, both inputs and outputs are provided to the model. As our dataset contains historic prices of nearly 50,000 diamonds along with the features, we can use both features and prices to train the machine learning models. In this project, we have chosen three machine learning algorithms to perform the predictions. They are linear regression, Decision Trees and K Nearest Neighbours. Python was used a programming language to build and train the algorithms. External modules like scikit-learn, pandas and matplotlib were made use of easier and quick implementations. In our literature survey, we could see many research works have tried implementing the same machine learning algorithms but some of the works haven’t properly performed exploratory data analysis and feature encoding while other related works haven’t focussed on tuning the algorithms which is also very important in building a robust machine learning predictor. Through this project, we will be exhaustively implementing all the machine learning techniques and build a robust model. Following are the b brief steps followed in developing the machine learning models. They are as follows:

1. Loading the dataset.

2. Exploratory Data Analysis

3. Data pre-processing

4. Splitting the data into train and test splits

5. Model training

6. Hyper parameter tuning

7. Model Evaluation

8. Real time prediction

Section 2 focusses on motivation behind this project. Following section briefs about main contributions and objectives of this project. Next, a brief overview of the literature survey is presented on what all has been achieved till now in this domain. Section 5 details about the proposed framework in depth. Description of the dataset follows in section 6. Thereafter, experiments done, and results obtained are displayed. Along with these, a comparison of models is also shown in this section. In the last sections, references followed while developing this project are listed.

1. **Motivation**

Diamonds are one of the most valuable naturally occurring carbon compounds. These are in high demand due to their monetary rewards. However, unlike other precious metals like gold and silver, determining the price of a diamond is not straight forward as it depends on many factors including carat, cut, colour, dimensions, depth percentage and Table width. To put in simple words, a diamond that is cut with high precision has more monetary value than one with less precision. As many factors play in determining the price of the diamond, it is burdensome to tag a diamond with a particular price. This problem gives an opportunity to leverage historic data and machine learning algorithms to predict the price of a diamond. Having a system that could help in deciding the intrinsic value of a diamond gives an immense opportunity to both business and customers in transacting diamonds. From a customer point of view, he/she can use the trained systems to assess before making a purchase. At the same time, it also helps business to understand trends/demand in the market for a particular diamond.

1. **Main Contribution & Objectives:**

* Our project aims to implement algorithms for predicting diamond prices. Algorithms we decided to implement are KNN, Decision Tree, Linear Regression algorithms.
* Using trained algorithms, predict the price of the diamond given its features
* Target on optimizing each model using various hyperparameter tuning mechanisms like Grid Search
* Our contribution through this project is we were able to build above mentioned models using python as programming language
* We were also able to hyper tune and produce a best model that has produced 97% score on the test dataset.

1. **Related Work**

As a part of our literature survey, we went through related works done in the same domain. There are several research articles proposed to predict diamond price prediction. A brief finding from our literature survey is presented here in this section. A research paper proposed by Harshvardhan et.al, [7] uses regression algorithms like XGBOOST, KNN, Decision trees, Support Vector machines, Random Forest regressors and other 4-5 machine learning algorithms to train on the dataset. They were able to achieve 98% of accuracy through their approaches. They have also used the same dataset as our dataset. They found out that cat boost algorithm is proven to be the best optimal algorithm among all the others they have experimented with. They have used R2 score to evaluate their models.

Another study by Graima etal., [6] on diamond price prediction is also reviewed by us as a part of literature survey. They have provided a comparative analysis of supervised models for diamond price prediction. In their study, they have evaluated eight different regression algorithms to train the predictor. Algorithms they have implemented in their study are Efficient Net, AdaBoost Regression, Random Forest, Gradient Boosting Regressor, Linear Regression, Lasso Regression and Decision trees and Random Forest. They claimed that out of the eight algorithms they have experimented with, Random Forest Regression model has given out the best accuracy. In the results section of the paper, they have compared scores of different algorithms. For comparison, they have used Root mean squared error as a metric to evaluate all models. But the research paper doesn’t mention about the details of data analysis they performed and pre-processing steps they have performed. Another downside of the paper is they haven’t performed any parameter tuning mechanism to improve the accuracy. Exploratory data analysis and pre-processing are vital prerequisites in a building a machine learning algorithm. Their best model Random Forest regression has resulted root mean square error of 580. They also mentioned the accuracy of random forest model as 97%. Another research paper by Avinash Chandra Pandey and other [5] has used ensembled learning to predict gold as well as diamond prices. Ensemble learning is a technique in machine learning where we use two or more different algorithms outputs to derive final output. It is basically a mixture of different algorithms. According to their study, they have performed dimensionality reduction techniques like Principal component Analysis (PCA) to reduce the features in the dataset. To ensemble different algorithms, they have chosen Random Forest Regressor, Adaboost, Lightbgm and Xgboost. They have presented in detail about the techniques they have used to pre-process the data. But this research also hasn’t performed any hyper parameter tuning mechanism which had the potential to improve model performance significantly. They also stated techniques like PCA, Recursive feature elimination has improved the model performance significantly but haven’t mentioned about by what factor they were improved.

From our literature survey, we have concluded that most of the related works despite using wide range of algorithms haven’t focussed much on data pre-processing and parameter tuning mechanism. As per our knowledge based on literature survey, this is the first research work to perform hyper parameter tuning on selected algorithms to improve accuracy.

1. **Proposed Framework**

This section is divided into four sections. They are:

1. Data loading and analysis

2. Data pre-processing

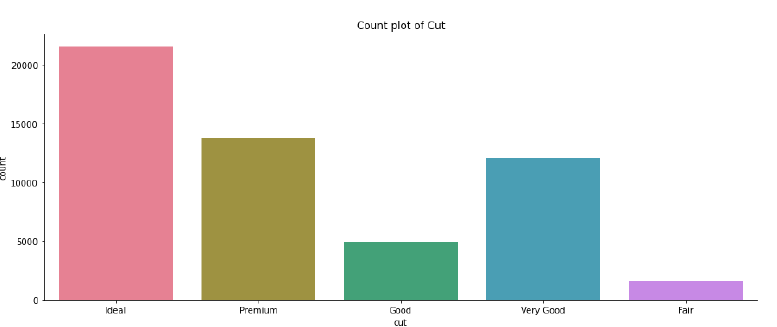
3. Model training

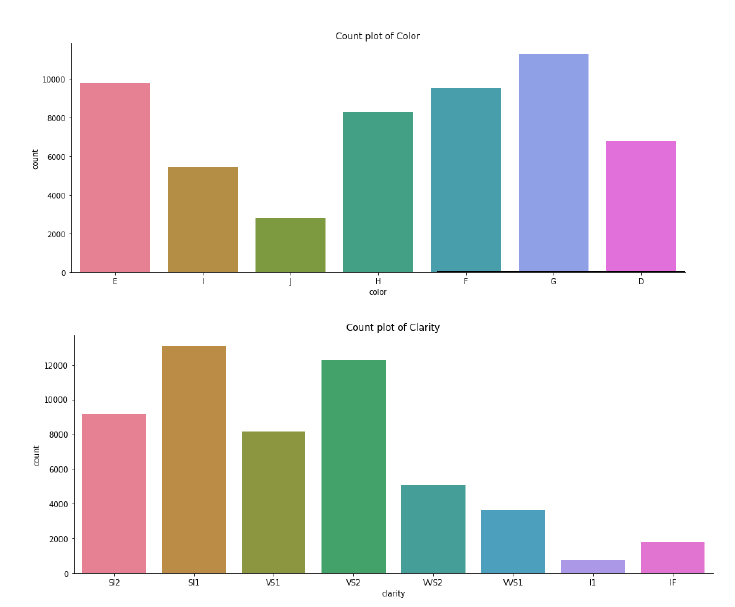
4. Parameter Tuning.

Let us go into each of the sub sections in detail.

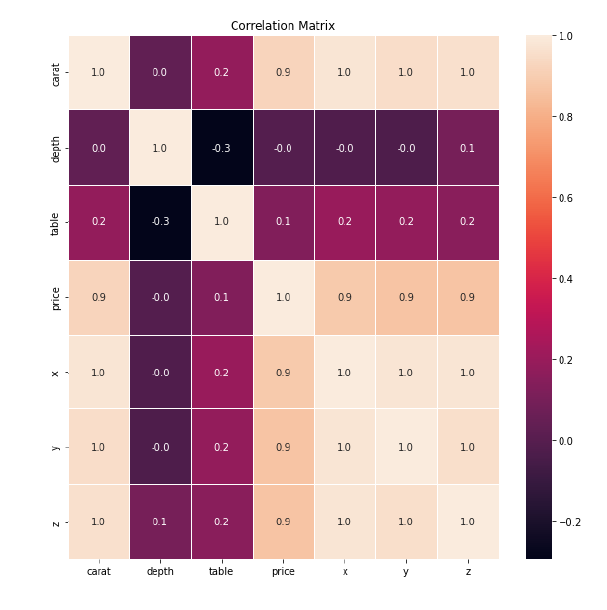
1. **Data loading and analysis:**

As an initial step, all required dependencies were installed in the environment. Thereafter, the dataset has been loaded into the python environment using panda’s library. Once data is loaded, it is verified whether data loading is done correctly. Data is checked for null values. There are no null values in the dataset but upon describing the dataset columns, we could find the dimensions in some columns are zero. Such rows are discarded. Now, variables are visualized using various kinds of plots. A count plot is drawn for three attributes colour, cut and clarity. Visualizations are shown below.





Thereafter, a correlation heatmap is generated for all of the features available in the dataset. Correlation is a technique that helps us in understanding the relationship between pairs of features. Correlation heatmap is visualized using seaborn library as shown in below figure.



**Fig**: Correlation Matrix

Followed by histogram plot and pair plot of all numerical features are also visualized to understand the relationship between attributes.

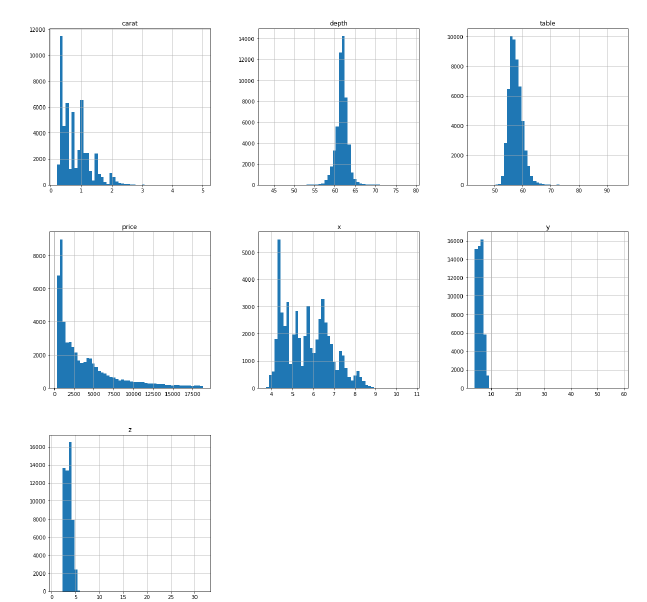


Fig 3: Histogram representation

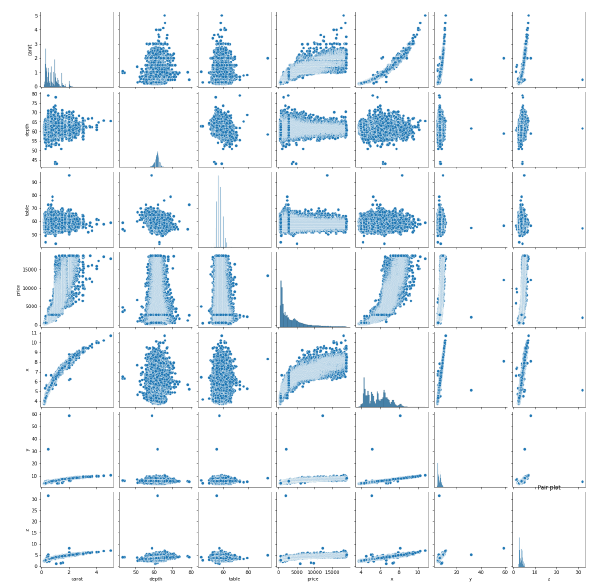


Fig 4: Pair plot representation

2. **Data pre-processing:**

In this section, we will go through different pre-processing techniques implemented in this project. As x, y and z are width, height and depth of the diamond, we will multiple all of these features to create a more meaningful feature ‘volume’ and append that to existing data. Then, Label encoding is performed on the categorical columns cut, colour and clarity. Label encoding is the process of converting categorical variables to numerical variables. Machine learning algorithms can only interpret numerical data. So, we need to transform them using label encoders to feed them to the model. Once, label encoding is completed, scaling is performed on the whole data to reduce variance on the data.

3. Model training:

Once data is pre-processed, it is time to train the algorithms. A summary of three algorithms implemented is as follows.

1. Linear Regression:

Linear regression is a supervised machine learning algorithm where the dependant variable is predicted using independent variable. In this algorithm, linear approach is used between dependant and independent variable. The goal of this algorithm is to find out the linear relationship between dependant and independent variable.

2. Decision Tree Regressor:

Decision tree is another supervised machine learning algorithm where questions are asked iteratively and based on that, classification is done. Below is a visual representation of the decision tree.

3. K Nearest Neighbours Regressor:

KNN Classifier is a simple and lazy machine learning algorithm. Given number of neighbours k, the algorithm tries to classify the new data point to one of the neighbours based on the distance metric like Euclidean Distance. The algorithm is implemented in Scikit-learn. Below is the code snippet to train and test the algorithm.

Once models are trained, hyper parameter tuning is done using GridSearchCV approach. In this approach, we will provide the values that we want to experiment with. The algorithm creates all possible combinations of parameters in a grid manner and returns the parameter combination with best accuracy. This is done with the help of scikit-learn.

1. **Data Description:**

The dataset has been extracted from Kaggle. The dataset contains around 50,000 records/samples of different diamond features and their prices. The dataset has 9 attributes. They are as follows:

1. Price/cost of the diamond : cost at which diamond was sold. The price of the diamonds is mentioned in US dollars. The prices in the dataset are in the range of $326 - $18,823.
2. Diamond Carat - Carat is the metric in which diamonds physical weight is measured. It is divided into 100 points.
3. Cut – Based on the quality of the cut, it is assigned one of the mentioned tags.

1. Ideal

2. Premium

3. Very Good

4. Good

5. Fair

This is determined the diamond grader which is done after evaluating the cutter’s skills.

1. Diamond Colour - Colour of the diamond varies in various hues. Depending on the colour, price is varied. Typically it occurs in many hues. It could be colourless as well. Here as well, like cut, each diamond was labelled into a category. Categories ranges from J to D where J is the worst and D is the best.
2. Clarity of the diamond - Based on the clarity of the diamond, a tag is assigned to that. The clarity of the diamond is carefully examined and assigned into one of the categories available.
3. X, Y and Z dimensions of the diamond- These are the physical dimensions of the diamond. These are measured using measuring equipment’s and given.
4. Depth of the diamond- Depth is same as the height of the diamond. It is calculated using formula.
5. Table Width of the diamond – Table width determines the refraction of the diamond. Table refers to the facet of the diamond when seen from face up. An ideal cut diamond has the most brilliance.



Fig: Description of diamond

1. **Results/Experimentation & Comparison:**

Following are the results of various experiments done.

1. **Linear regression:**

Linear regression is a supervised machine learning algorithm that predicts the dependent variable with the help of the independent variable. A linear approach is used between the dependent and independent variables in this algorithm. This algorithm's goal is to discover the linear relationship between the dependent and independent variables.



1. **Decision Tree Regression:**

Another supervised machine learning algorithm is decision tree, which asks questions iteratively and then classifies the answers. The decision tree is depicted graphically below.

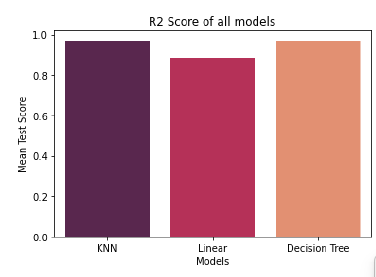


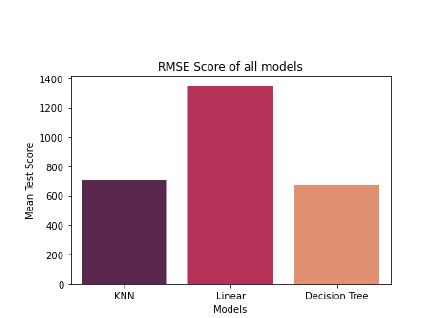
1. **K-Neighbours Regression:**

The KNN Classifier is a straightforward and lazy machine learning algorithm. Given the number of neighbors k, the algorithm attempts to assign the new data point to one of the neighbors using a distance metric such as Euclidean Distance. Scikit-learn is used to implement the algorithm. The code for training and testing the algorithm is provided



**Comparison of all models:**





To sum it up, on comparing the mean squared error, mean absolute error and mean squared error in all the models. The result we have got is the accuracy is greater when we used the Decision tree algorithm.

1. **References**

[1] GitHub repository containing code of the project

[2] https://www.kaggle.com/datasets/shivam2503/diamonds

[3] Wikipedia contributors. (2022, November 26). Diamond. Wikipedia. https://en.wikipedia.org/wiki/Diamond

[4] Editor. (2019, September 18). Price Forecasting Using Machine Learning Approaches in Electricity, Flights, Hotels, Real Estate, and Other Industries. AltexSoft. https://www.altexsoft.com/blog/business/price-forecasting-machine-learning-based-approaches-applied-to-electricity-flights-hotels-real-estate-and-stock-pricing/

[5] Alsuraihi, W., Al-hazmi, E., Bawazeer, K., & AlGhamdi, H. (2020, March). Machine learning algorithms for diamond price prediction. In Proceedings of the 2020 2nd IVSP (pp. 150-154).

[6] Pandey, A. C., Misra, S., & Saxena, M. (2019, August). Gold and diamond price prediction using enhanced ensemble learning, IEEE.

[7] Sharma, G., Tripathi, V., Mahajan, M., & Srivastava, A. K. (2021, January). Comparative analysis of supervised models for diamond price prediction.(pp. 1019-1022). IEEE.

[8] Marmolejos, José M. Peña. “Implementing Data Mining Methods to

Predict of the Diamond Prices.” in 2018

ICDATA’18.

[9] Mihir, H., Patel, M. I., Jani, S., & Gajjar, R. (2021, December). Diamond Price Prediction using Machine Learning. In 2021 2nd International Conference on Communication, Computing, and Industry 4.0 (C2I4) (pp. 1-5). IEEE.

[10] Cardoso, M. G., & Chambel, L. (2005). A valuation model for cut diamonds. ITOR - International Transactions in Operational Research, 12(4), 417-436.

[11] Kotsiantis, S. B. (2013). Decision trees: a recent overview. Artificial Intelligence Review, 39(4), 261-283.

[12] Weron, R. (2014). Electricity price forecasting: A state-of-the-art review with a look ahead. International journal of forecasting, 30(4), 1030-1081.

[13] Diamond-The most popular gemstone [online] https://geology.com/minerals/diamond.shtml

[14] Brandt, J. A., & Bessler, D. A. (1983). Price forecasting and evaluation: An application in agriculture. Journal of Forecasting, 2(3), 237-248.

[15] 3.2. Tuning the hyper-parameters of an estimator. (n.d.). Scikit-learn. https://scikit-learn.org/stable/modules/grid\_search.html

[16] 1.10. Decision Trees. (n.d.). Scikit-learn. https://scikit-learn.org/stable/modules/tree.html

[17] Supervised vs Unsupervised Learning - Javatpoint. (n.d.). www.javatpoint.com. https://www.javatpoint.com/difference-between-supervised-and-unsupervised-learning

[18] sklearn.linear\_model.LinearRegression. (n.d.). Scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.linear\_model.LinearRegression.html

[19]https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html

[20] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html