Diamond Price Prediction Using Machine Learning

Süleyman Buğra Gülsoy
Department of Statistics
Middle East Technical University
Ankara,Turkey
bugra.gulsoy@metu.edu.tr

***Abstract*—Gemstone price prediction is an essential part of trading. Gemstone market is volatile and unstable. Since the prices of gemstone are not fixed and trades are done with auction and bargaining. Individuals are usually faced with unfair trades. The market is inconvenient for new traders. Even selling a basic piece of gemstone is guarantee way to lose money for uniformed individuals. I aim to develop a model using different machine learning algorithms to create an efficient model to predict diamond prices. The diamond prices data is available at Kaggle.**

Keywords—MLR, ANN, SVM, XGB, RF, IMPUTATION, REGRESSION

I. Introduction

Gemstone trading is a profitable business. The trade of gemstones is done by bargaining and auction. The price of each individual gemstone is distinct and flexible since its attributes are affecting price. Since gemstones do not have a fixed price like gold correctly pricing them is essential. Individual gemstone holders, collectionists, inheritors or jewelers constantly trade their gemstones with hopes of profiting. The aim of the project is to propose a method to estimate the price of a diamond which is one of the most common, popular and valuable gemstones. The motivation is with the proposed method individuals can use this method to estimate the price of their diamonds to not sell it underpriced similarly buyers can use the proposed method to prevent any overpriced buy. Traders can use this method to make profitable trade, improving their negotiation. Normally individuals estimate the price by their own knowledge and personal experience. The project can stabilize fluctuations in prices of each trade.

II. Data Preprocess and Preparation

A. Data Description

The provenance of data is Australian diamond importers website(australiandiamondimporters.com.au) and scrapped on 24th Feb 2022. Scrapped data shared publicly in Kaggle by EMAD NASHED. (https://www.kaggle.com/datasets/enashed/diamond-prices)

The data has a total of 27 variables. Out of 27 variables, 7 of them are numerical, 18 of them are categorical variables. 1 variable is unique indicator of diamond's name, and 1 variable is date that data is scrapped which is equal for all observations. The target is continuous variable "total_sales_price" which indicates sales price of the diamond. Seven of the explanatory variables are NA inflated. Thus, they are beyond the level of usability, over 90 percent of the observations are NA for this seven variables. The variables are the features of each diamond's properties. The variable names are self-explanatory. The diamonds are graded by their properties and classified by institutes. The variables lab represents which institute classified the diamond. Because in order to tell the quality of certain property such as cut, symmetry, polish which are variables of my data the diamond needs to be expertise by someone who has certificate of this institutions. The most common institutions are GIA and IGI stands for Gemological Institute of America and The International Gemological Institute. 98% of the observations are graded by either one of them. After removing the NA inflated and unnecessary date and ID variables. 18 variables remain. 7 of them are numerical and 11 of them are categorical.

The frequency tables of categorical variables are examined. No miss written observations are detected. However, some categorical variables have frequency problems. To exemplify culet size as can be seen below has categories that is extremely rare (less than 10 observations out of 220000) Thus, while making feature engineering for categorical variables the categories that has almost 0 frequency won't be included in one hot encoding. Also, the type of variables has adjusted to correct factor forms.

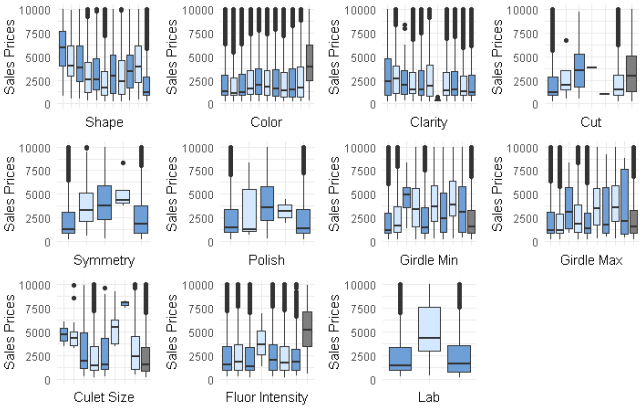

Figure 1 Association Between Categorical Variables and Price

The association between categorical variables and diamond prices is inspected. It is observed that The diamond price significantly varies between categories. Kruskal Wallis tests are conducted for all the categorical variables. The test also confirms ($p < 0.05$) That diamond prices change with different categories. To further have an analysis about significant differences between subcategories in same variable Dunn Test was conducted. Out of the total 302 subcategories 120 subcategories found statistically similar to each other. The subcategories that found unsignificant by Dunn Test, will be removed from study with dimensional reduction.

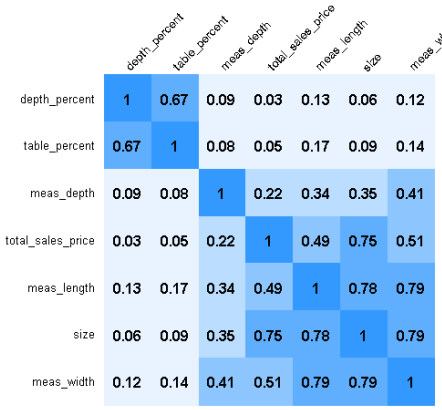

Figure 2 Correlation Matrix of Numeric Variables

The total sales price is the target variable. Thus, the most correlated variables with it are "meas length", "size" and "meas width" Hence, It is safe to assume that these 3 variables will be the most important ones in the analysis and modelling. On the other hand, the magnitude related variables are found correlated with each other. Which might cause multicollinearity problems. So, those variables can be combined into a single Principal Component "magnitude" to avoid multicollinearity.

Distribution of numerical variables are inspected. Found that numerical variables have 0 values which is not logical for the nature of data. To exemplify size of a diamond or depth of a diamond can is not 0. Hence, such observations changed into NA values. Also, Anderson Darling test has conducted for numerical variables. Found that none of the numerical variables were distributed normally. Yet, Anderson Darling test is susceptible to large sample size. It has a tendency to reject.

B. *Outlier Analysis*

Data is found to have a lot of extreme values to analyze any possible problematic outliers, percentile-based detection method is used. The method converts the numerical values into z- score or chi square scores(depending on the distribution of the variable). And observations that exceed chosen percentile, is selected as outliers. The fifth and 95th percentile is selected as outliers threshold. The analyses repeated for all the numerical variables and found that in average 8000 observation for each feature was outliers. However, after inspecting the observations none of them found problematic such as miss typed or not logical. Thus, observations were kept in the study.

C. *Missing Value Analysis*

The data suffers from missing value by a great margin. The majority of the categorical variables have missingness up to %30. Also, ten thousand numerical observations are missing after adjusting the 0 values. Little's test of detecting MCAR has been conducted. The p value was extremely small and indicated that missingness is not MCAR. For deciding whether the missingness in not random (MNAR) or random (MAR), the missing value from a feature is removed and summary statistics with and without missing values are compared. The method is applied for each feature and found that summary statistics and frequencies were similar with and without NA values. Thus, it is claimed that the missingness were random (MAR). To aid the missingness in data "Multiple Imputation Method" with "Predictive Mean Matching" feature is used. PMM is a imputation technique that replaces missing data with estimated likely values from observed data. This guarantees the imputed values to be similar to real values and preserves the original distribution. And It can also be used with multiple imputation and for both categorical and numerical variables. Which makes PMM computationally easy and effective. The PMM applied with number of imputations(m) 5 and with 5 iterations for each imputation. Thus, the algorithm imputed 5 different numbers to a missing value and for finding each imputation number it uses 5 iterations. Then combining the results to overall have more accurate imputation. After imputation the missing values, changes in the variables are examined. There does not appear to be significant change in all of the variable's characteristics. Their density or frequency remained similar compared to not imputed data.

D. *Normalization*

Three different normalization techniques are used for the data. Min-Max scaling, Z-score scaling and Decimal scaling. The normalization technique can vary in terms of effectiveness from data set to data set. Thus, It Is not possible to choose an overall advantages method. Moreover, each normalization method can work better for different algorithms. Thus, selecting a normalization method prior to model analyses can cause loss in prediction power. Hence, all tree differences normalization techniques are applied at the same time and obtained three different datasets. Each model will be trained with all three different normalized data. The Min-Max method fits data into specific ranges by subtracting each data point from its minimum value and dividing it with its range. Z-score method is the most common of all three. It converts each data point to a mean zero variance one Z-score. Decimal scaling technique moves the decimal point of each observation according to the maximum value of the feature. After applying the methods to each feature of data set excluding the target, three distinguished data have obtained.

E. *Feature Engineering*

Machine Learning algorithms can require data to be numerical only. For make the data applicable for modelling and separate each subcategory from each other one hot encoding is applied. One hot encoding is separating a categorical variable that has "n" categories into "n" newly created variables such that each new variable can take only values 0 or 1. After applying one hot encoding data set had 80 features, the majority of them being dummy variables. For resolving the dimensionality problems and to get rid of unnecessary features as previously found. Some features had frequency less than optimal and Some dummy variables are not significant. The LASSO method is used for reducing the number of attributes. The LASSO is L1 regularization technique that when used in regression it shrinks the variable weight up to 0. Thus, is can be used for feature selection. LASSO model was created for each data. And nonzero coefficients of LASSO model is chosen to be significant features. All the parameters that their coefficients were shrinked to zero by LASSO removed from the dataset. After the process is completed. The number of features in the data sets decrease to 19. The variables "size" "depth", "meas_length" , "meas_width", "shape", "color", "clarity",

"lab" has found significant according to LASSO. Thus, the data sets were preprocessed for modelling.

III. Modelling

Multiple Linear Regression

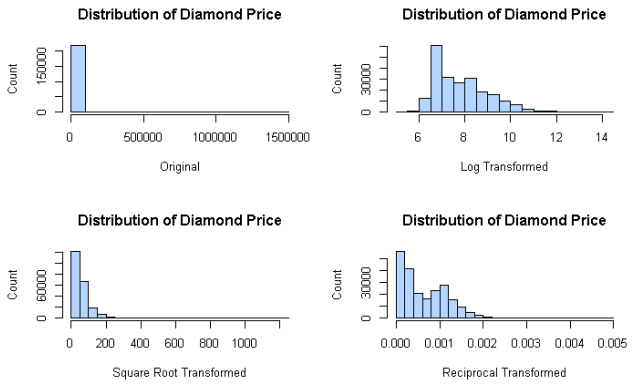

Figure 3 Distribution of Diamond Price

The multiple linear regression assumes the target variable to be normally distributed. The diamonds price variable on the other hand, has cumulation in lower values and extreme values are disturbing the distribution. The extreme right skewness is present. The table above shows the drastic level of skewness. The mean value and medians value is quite far and extreme values exist towards upper values of price. Which is logical for the nature of diamond prices.

Table 1 Summary of Diamond Price

Min	First Quartile	Median	Mean	Third Quartile	Max
200	958	1978	6908	5207	144988 1

To fix the issue. Log transformation, square root transformation and reciprocal transformation have been applied. And Anderson Darling normality test has conducted for each. The Anderson Darling test found original and all of the transformed variables not normally distributed. But the test is volatile to huge sample size and extreme values. Moreover, graph of log transformed diamond price shows similarity to normal distribution. Thus, log transform method has accepted as successful and MLR model is created after splitting the data to train and set with respect to 80% 20% rule. The adjusted R square for the model found as 0.8691. And the model found all of the variables significant. Dimension reduction successfully selected most related features. The RMSE for train data calculated as 27145.45 While, RMSE for test data calculated as 25677.17 Thus, It can be observed that model functioning better with unseen data. And the model can be assumed quite successful. But the model violated the normality error assumption. Residuals have heavy tails and are not normally distributed. Even though the target was log transformed already. The aid the issue, instead of OLS WLS was applied. The adjusted R square for WLS model calculated as 0.844. It decreased slight from OLS model. However, the RMSE for train data was obtained as 27145.38 While, RMSE for test was data 25677.09. Thus, in terms of predictive power both models are identical. On the other hand, WLS method was also unsuccessful to ensure the normally of residuals assumptions.

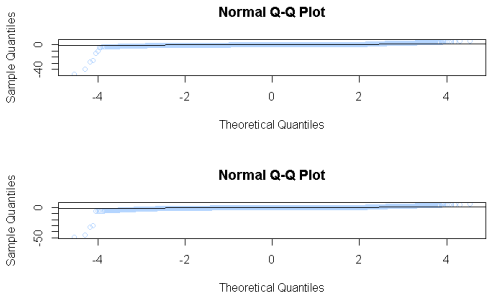

Figure 4 Normal Plot of Residuals

The qq plot above is in residuals from OLS and plot below is residual from WLS both. Also, Anderson Darling test confirms that both residuals are not normally distributed. Since, after transforming the target variable and using WLS was not ensured the MLR assumptions. It is concluded that the data was not application for MLR model. The model is not validated to formally used in future work. But it will be used as a benchmark value. To compare other models. For deeper analysis four more models are applied. ANN model, SVM model, RF model and XGB model. To increase the model performances repeated cross validation technique is used. With 5 validation sets and 5 repeats. In total 5 models were created for each three different normalized data sets. In total 15 models are compared with respect train and test errors. For error matrices RMSE is used. The goal is selecting the best performed combination to tune hyper parameters.

Table 2 Comparison of Models

	RMSE for Train			RMSE for Test		
MLS	27145.45	26288.96	35701.28	25671.17	23626.57	27189.21
ANN	27147.78	29207.08	37469.17	25629.56	28328.13	34828.06
SVM	31502.36	30938.28	39251.36	33792.41	35826.74	42791.84
RF	21165.99	23721.36	28749.33	23008.14	24487.26	31305.38
XGB	7784.39	8163.18	9701.18	11215.42	15372.76	17949.76
	Min-Max	Z-Score	Decimal	Min-Max	Z-Score	Decimal

IV. Results

In order to find the best model five different models are compared in terms of RMSE metric in test data set. It has been found that XGB algorithm is extremely superior compared to other algorithms when it comes to prediction problems. The mean square error for XGB was significantly less than other algorithms. And the XGB gave the best performance once min-max scaling. The RMSE value calculated as 11215.42 for proposed min max scaled XGB model. Thus, in order to develop the model even better hyper parameter tuning has applied. To tune the parameters grid search method was used. Parameter grid was created with the values of hyper parameter

that wanted to be optimized. For computational limitations the grid was only limited with 3 different values for each hyperparameter only. The “nrounds”, “max_depth” , “gamma”, “colsample_bytree” , “subsample” hyper parameters are tuned to find optimal values. And resulted that the best optimal values for the stated values are respectively nrounds 200, max_depth 9, eta 0.001, gamma 0.01, colsample_bytree 0.6, subsample 0.8. Moreover, tuned parameters successfully increased the model performance and resulted with RMSE test value 9827.01 which is quite lower than the XGB without tuned.

V. Conclusion

The project aimed to provide a model to effectively predict the value of diamond, which is the most common gemstone. The project tried to provide an optimal method to preprocess the data. And instead of choosing a specific normalization method I proceeded with all to compare them at the end with all possible combinations of models. Dimension reduction proved to be successful in data. While MLR was inappropriate. The proposed best model is found to be XGB with Min-Max scaling and optimal parameters.

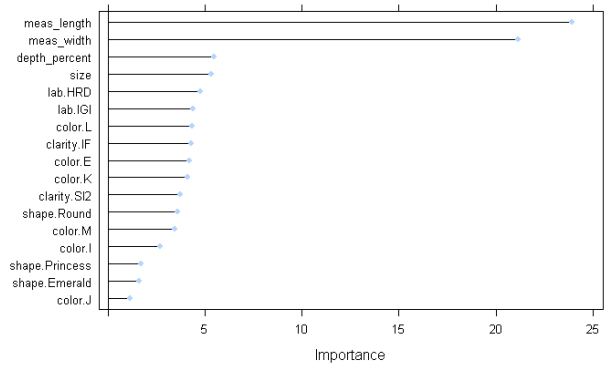

Figure 5 Variable Importance

The most valuable features are found as meas_length and meas_width. We can conclude that the magnitude of the diamond is the most influential property. Which is expected. However, an unorthodox result can be deducted as well. The lab that diamond has graded has also critical value on diamond’s price. With the proposed method in the future. Individuals can use the model to estimate the price of their diamonds. With the study diamond trading industry would be less fluctuated and less risky for new comers, fresh investors.

VI. References

S. Gopal Krishna Patro, Kishore Kumar sahu, “Normalization: A Preprocessing Stage”

V. Kumutha and S. Palaniammal, “An Enchanced Approach on Handling Missing Values using Bagging K-NN Imputation”

M. Schumacher, R. Robner and W. Vach, “Neural networks and logistic regression,”

S. Lu, Z. Li, Z. Qin, X. Yang and R. S. M. Goh, “A hybrid regression technique for house prices prediction,”