Implementing Data Mining Methods to Predict Diamond Prices

José M. Peña Marmolejos
Graduate School of Arts and Sciences, Fordham University 113 W 60th Street
New York, NY 10023, USA

Abstract— In this research paper, an analysis of diamond price prediction is presented by implementing data mining algorithms. The dataset used for this research is the diamonds dataset which is publicly available from the Kaggle repository; the data contains 53,940 records, with 10 unique features. The proposed methodology learns the patterns of every diamond combination by applying three models using the following algorithms: neural networks, linear regression and M5P regression tree. Moreover, an assessment of the most relevant features contributing to diamond pricing is done by contrasting their correlation among each other. Consequently, an analysis of the effects of multicollinearity on the performance of the data mining models is conducted.

Keywords— Data Mining; Diamonds; MLR; M5P.

1 Introduction

Diamonds are valued for reasons beyond aesthetics including their hardness (they are the hardest naturally sourced mineral), abrasive nature and ability to disperse light, about one-fifth of diamonds are used in industry, for lasers, drill bits and surgical equipment, largely in the auto and aerospace sectors [1]. They also serve as insulation and have high heat conductivity, suggesting that diamonds may have applications in the semiconductor industry.

Nowadays precious metals around the globe are valued based on their weight, whereas diamonds involve many other factors that significantly affect their price; by the time a diamond reaches a retail store, it has undergone through a series of steps in the supply chain, adding more cost to the precious stone in every single iteration, this leads towards comparatively elevated prices for most available diamonds in the marketplace.

The rough gemstones must first be mined and cut. Once cut, diamonds are appraised to determine their value. Some of them undergo treatments to augment their appearance. Diamonds progress through these processes to be transformed into beautiful valuable pieces of personal adornment.

Some industry leaders like Rapaport, a major trade source for diamond prices, are working toward a standardization that would bring transparency to the pricing of diamonds. The diamond jewelry trade is fairly unified in claiming that this cannot be accomplished. Taking an adamant stance that diamonds cannot be considered a commodity, while by most definitions they can, the trade insists that each diamond is too unique and that standardization cannot account for the diamonds symbolic value of “enduring love and commitment” (which they believe and spend millions in advertising to convince consumers of this notion). The implied thought is that diamonds are nearly priceless and that their value would not hold should they be traded as a commodity [2].

Although the process of mining, cutting and polishing may set the baseline price for a diamond, the following features are the major defining factors to consider in order to obtain an accurate figure: carat weight, cut, color, clarity, length, width, depth, depth percentage and table width. The 4Cs describe the individual qualities of a diamond, and the value of an individual diamond is based on these qualities. The terms that people use to discuss the 4Cs have become part of an international language that jewelry professionals can use to describe and evaluate individual diamonds. Today, the descriptions of each of the 4Cs are more precise than those applied to almost any other consumer product. And they have a long history. Three of them—color, clarity, and carat weight—were the basis for the first diamond grading system established in India over 2,000 years ago [3].

The characteristics of the data can be observed to establish the factors associated with an outcome. Observation studies such as data mining, can reveal the association of the features to the target outcome. Data driven statistical research is becoming a common component to a plethora of areas such as stock market, product and business development. The discovery of diamond value prediction is possible by extracting the insights in the data that are directly related to the mineral.

The main purpose of this research paper is to present the data mining techniques to predict diamond prices in US dollars by employing the Kaggle diamond dataset and regression approaches to determine an accurate outcome. Moreover, a comparison of the results of neural networks, linear regression and M5P algorithms is performed to determine which one performs better for our task.

This paper is organized as follows. The next section reviews the related work. Section 3. presents the experiments conducted. Experimental results are presented in section 4. The conclusion is given in the last section.
2 Related work

There were many studies in the attempt to predict diamond prices using different methods. Chu [4] implemented multiple linear regression (MLR) to explore the relationship between the four C’s of diamonds and its pricing. The analysis was to focus on data pertaining to \( n = 308 \) round diamond stones. Given the information in the dataset, an MLR model is a natural path to explore. Generally speaking, one would expect the price (denoted in Singapore dollars) of a stone to move in tandem with the carat weight. However, the relationship may not be linear as heavier stones are more prized than the lighter ones. To illustrate this, an examination of the Kaggle diamond dataset (denoted in U.S. dollars) is performed through a scatter plot visualization. Figure 1.1 shows the trend of price against carats.

![Figure 1.1. Scatterplot of Price vs Carat](Image)

The obvious assumption is that there is an unquestionable 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. In order to satisfy the homoscedasticity assumption, the logarithm of price is the most common transformation. This is illustrated below in Figure 1.2.

![Figure 1.2. Scatterplot of logarithm(Price) vs Carat](Image)

Nevertheless, in this current study more features and a wider range of data analysis methods are taken into account; and unlike [4] the dataset in this study contains information of 53,940 diamond stones, adding therefore even more robustness and accuracy to the research.

3 Experiments

In this paper three data mining algorithms for regression tasks were taken into account, such as: Multilayer Perceptron Neural Network, Multiple Linear Regression and Regression Tree M5P. A results comparison of the three former algorithms is conducted in order to predict the price of Kaggle diamond dataset.

The Neural Networks technique represents a brain metaphor for information processing. These models are biologically inspired rather than an exact replica of how the brain actually functions. Neural networks have been shown to be very promising systems in many forecasting applications and business classification applications due to their ability to “learn” from the data, their nonparametric nature (i.e., no rigid assumptions), and their ability to generalize [5]. An artificial neural network has a more complex structure than that of a perceptron model. The additional complexities may arise in a number of ways:

The network may contain several intermediary layers between its input and output layers. Such intermediary layers are called hidden layers and the nodes embedded in these layers are called hidden nodes [6]. The resulting structure is known as a multilayer neural network (see Figure 2).

![Figure 2. Multilayer Perceptron Neural Network](Image)

The goal of the ANN learning algorithm is to determine a set of weights \( w \) that minimize the total sum of squared errors:

\[
E(w) = \frac{1}{2} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2
\]  

In Multiple Linear Regression, there are \( p \) explanatory variables, and the relationship between the dependent variable and the explanatory variables is represented by the following equation:

\[
y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + e_i
\]  

Where \( \beta_0 \) is the constant term and \( \beta_1 \) to \( \beta_p \) are the coefficients relation to the explanatory variables of interest [7].

M5P, a binary regression tree models in its final nodes (leaves) in order to produce linear regression functions that can produce continuous numerical attributes. The model is based on tree splitting method. The first step involves the use of a divergence metric to produce a decision tree. Branching criterion for M5P model tree algorithm is the behaved class values that reach a node as the quantification of the error and the expected reduction in error as a result of testing each attribute at that node is calculated.

The formula for calculating the standard deviation reduction (SDR) is as follows:

\[
SDR = sd(T) - \frac{\sum_{i=1}^{n}|T_i|}{|T|} \times sd(T)
\]  

SD = standard deviation is indicated. Because of the split process, data in child nodes have lesser than standard deviation from the parent node and hence purer. After maximizing all possible ramifications, M5P selects which attribute will reduce the expected maximum. This division often creates a tree-like structure that is over-fitting. To overcome the problem of over-fitting, the tree should be pruned back, for example, by replacing a sub-tree with a leaf. Therefore, the second step of the design model including tree pruning, tree removal and replacement of trees, plants grown with linear regression functions. This technique generates a tree model, the parameter space into regions (subside) and split each of them makes a linear regression model [8].

3.1 Multicollinearity

The term multicollinearity refers to a situation in which there is an exact (or nearly exact) linear relation among two or more of the input variables, (Hawking, 1983). If there is no linear relationship between the regressors, they are said to be orthogonal. Multicollinearity is a case of multiple regression in which the predictor variables are themselves highly correlated. If the goal is to understand how the various \( X \) variables impact \( Y \), then multicollinearity is a big problem. Multicollinearity is a matter of degree, not a matter of presence or absence. In presence of multicollinearity the ordinary least squares estimators are imprecisely estimated. There are several methods available in literature for detection of multicollinearity. By observing correlation matrix, variance influence factor (VIF), eigenvalues of the correlation matrix, one can detect the presence of multicollinearity [9]. In this study, this issue is addressed by removing the high correlating variables that can potentially be removed, this is shown later in Figure 3 within the “Dimensionality Reduction” section.

The Weka toolkit has been used to experiment with these three data mining algorithms. Weka is a collection of machine learning algorithms for data mining tasks, such as classification, regression, clustering, association rules and visualization [11]. The software is developed in Java and is open source under the GNU General Public License.

The data source for this purpose is the Kaggle repository website, which is a robust, reliable and ample resource for investigating a variety of topics related to data mining. The Kaggle diamond database provides the main characteristics of the precious stone and the average selling price for different setups [12]. The features of the diamonds are presented in Table 1.

### Table 1. Data Features

<table>
<thead>
<tr>
<th>Name</th>
<th>Values</th>
<th>Data Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Price (USD)</td>
<td>326…18823</td>
<td>Numerical</td>
</tr>
<tr>
<td>Carat Weight</td>
<td>0.2…5.01</td>
<td>Numerical</td>
</tr>
<tr>
<td>Cut quality</td>
<td>Fair, Good, Very Good, Premium, Ideal</td>
<td>Categorical</td>
</tr>
<tr>
<td>Color</td>
<td>J, I, H, G, F, E, D</td>
<td>Categorical</td>
</tr>
<tr>
<td>Clarity</td>
<td>I1, SI2, SI1, VS2, VS1, VVS2, VVS1, IF</td>
<td>Categorical</td>
</tr>
<tr>
<td>X (length in mm)</td>
<td>0…10.74</td>
<td>Numerical</td>
</tr>
<tr>
<td>Y (width in mm)</td>
<td>0…58.9</td>
<td>Numerical</td>
</tr>
<tr>
<td>Z (depth in mm)</td>
<td>0…31.8</td>
<td>Numerical</td>
</tr>
<tr>
<td>Depth percentage</td>
<td>( z / \text{mean} (x, y) = 2 * z / (x + y) ) (43–79)</td>
<td>Numerical</td>
</tr>
<tr>
<td>Table width</td>
<td>43…95</td>
<td>Numerical</td>
</tr>
</tbody>
</table>

3.3 Evaluation metrics

The performance of the models developed in this study have been assessed using various standard statistical performance evaluation criteria. The considered statistical measures have been correlation coefficient (R), mean absolute error (MAE), and root mean square error (RMSE).

\[
R = \frac{\sum_{i=1}^{n}(P_i - \bar{P})(A_i - \bar{A})}{\sqrt{\sum_{i=1}^{n}(P_i - \bar{P})^2 \sum_{i=1}^{n}(A_i - \bar{A})^2}}
\]

MAE = \( \frac{1}{n} \sum_{i=1}^{n} |P_i - A_i| \)

RMSE = \( \sqrt{\frac{1}{n} \sum_{i=1}^{n}(P_i - A_i)^2} \)

Where \( A_i \) is the price of diamonds, \( P_i \) is the predicted value, \( n \) is the total number of data points in validation, \( A \) is
the mean value of observations, and \( P \) is the mean value of predictions.

Additionally, all the three models are assessed using tenfold cross validation. This technique consists of dividing the overall data in 10 disjoint subsets, or folds. Each model is then trained using 9 of the subsets and evaluated using the tenth subset. The process is repeated 10 times and each time, a different subset is used for testing and the remaining 9 subsets are used to train the model. The algorithm is evaluated by averaging the prediction metrics from the 10 different models.

## 4 Results

The performance of linear regression, neural network and M5P algorithms evaluated on tenfold cross validation for the three models is given in Table 2.

<table>
<thead>
<tr>
<th>Method</th>
<th>Correlation</th>
<th>MAE (1017.5341)</th>
<th>RMSE (1112.2541)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear Regression</td>
<td>0.9621</td>
<td>713.4142</td>
<td>1071.5341</td>
</tr>
<tr>
<td>Neural Network</td>
<td>0.9859</td>
<td>439.8863</td>
<td>671.9511</td>
</tr>
<tr>
<td>M5P</td>
<td>0.987</td>
<td>293.6454</td>
<td>630.9923</td>
</tr>
</tbody>
</table>

As can be seen from Table 2, the M5P model produced better overall results than linear regression and neural network when evaluated on tenfold cross validation. In order to obtain these results, an in depth look at the data had to be taken, which led to the discovery of a small number of outliers that were affecting negatively on the performance of the models, by removing the outliers a significant decrease in the errors was produced. Additionally, the correlation coefficient improved after this procedure.

### 4.1 Dimensionality reduction

In this section we analyze the effects of dimensionality reduction by high correlation on the performance of the models. The current interest in this method raises several issues to be addressed. Although the previous results were good enough to draw some conclusions, one would expect to improve the performance of the models by removing features that carry very similar information. Generally, only one of them would suffice to execute the machine learning model, nevertheless in this particular scenario, three features were found to possess almost identical information and by leaving only one of them, the performance of the model suffered a slight increase in errors. However, when all the three correlating features were removed, there was a small improvement in the performance of the models. These insights prove the principle of multicollinearity and avoid misleading results by selecting only the features that carry the most meaningful information, hence improving the accuracy of the predictions. Figure 3. shows the correlation coefficient among the different features the diamond data contains. At a glance one can identify the high similarity in correlation among the \( x \), \( y \) and \( z \) features.

## 4.2 Performance comparison of the models

The M5P and neural network models demonstrated significantly better performance than linear regression in terms of mean absolute error and root mean squared error, however the M5P model had the best performance overall with the least amount of errors and the highest correlation coefficient when evaluated on tenfold cross validation with the original and the dimensionally reduced data. This is illustrated in Figure 4.
As shown in Figure 4. when evaluated on tenfold cross validation with the feature reduced data the M5P model had a great decrease in RMSE, hence demonstrating significant prediction power. Conversely the linear regression model had a small increase in errors whereas the neural network improved marginally, thus demonstrating that not every model benefit from this procedure. One hypothesis would be that the simplest algorithms such as linear regression suffer from prediction accuracy loss when working with large datasets which generally tend to be nonlinear, whereas the neural network can easily learn this type of dependencies.

Figure 5. shows the correlation coefficient of the three models. An interesting fact is that unlike the M5P and neural network models which benefited from the dimensionality reduction, the linear regression model suffered a small decrease in correlation coefficient.

5 Conclusion

From this study it can be concluded that data mining techniques such as linear regression, neural network and M5P model tree can be employed to estimate diamond prices. M5P model has shown to possess a high capacity for determining continuous numerical values. Dimensionality reduction by high correlation has proved to be a useful technique for performance improvement, although it might not apply for every scenario, as shown in Figures 4 and 5. Future work shall include more regression models and time series data, in order to extend the robustness and precision of the predictions.

6 References