5-2018

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

Stanislav Mamonov

Tamilla Triantoro

**Journal of Theoretical and Applied Electronic Commerce Research**
ISSN 0718–1876 Electronic Version
VOL 13 / ISSUE 2 / MAY 2018 / 15-28


This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762018000200103

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

**Stanislav Mamonov[1] and Tamilla Triantoro[2]**

[1] Montclair State University, Feliciano School of Business, Montclair, NJ, USA, stanislav.mamonov@montclair.edu.
[2] Quinnipiac University, School of Business, Hamden, CT, USA, tamilla.triantoro@quinnipiac.edu.

Received 21 March 2017; received in revised form 3 October 2017; accepted 17 October 2017

**Abstract**

Diamonds belong to a unique product category whose perceived value is largely dependent on socially constructed beliefs. To explore the degree to which the physical properties of a diamond can be used to predict the diamond price, we perform data mining on a large dataset of loose diamonds scraped from an online diamond retailer. We find that diamond weight, color and clarity are the key characteristics that influence diamond prices. The data mining results also suggest a high degree of subjectivity in diamond pricing that may reflect price obfuscation strategies employed by diamond retailers.

**Keywords:** Search costs, Price obfuscation, Diamond retail, Data mining, Pricing, Revenue management

**Journal of Theoretical and Applied Electronic Commerce Research**
ISSN 0718–1876 Electronic Version
VOL 13 / ISSUE 2 / MAY 2018 / 15-28


This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762018000200103

# 1 Introduction

Diamonds are a very unique consumer product. At the first glance, there is no apparent utility of owning a diamond. Although it is considered the hardest gem, that is 58 times harder than anything else on Earth [19], few people use diamonds for this utility. One of the most apparent reasons for owning or wearing diamonds lies in their perceived value as being rare and expensive objects. Although diamonds have been considered valuable objects for millennia, the consumption of diamonds significantly changed in the 20th century when it became customary to give diamond rings as engagement gifts. De Beers, the largest player in the diamond industry, has been using a slogan 'a diamond is forever' to encourage such gifts since 1948 [42]. Global sales of diamond jewelry in 2015 reached $79 billion, with the United States being the largest market, contributing $39 billion to total sales [5]. The demand for polished loose diamonds, globally was $25 billion in 2014 [4]. The top five markets for diamonds are the United States, China, India, Japan and the Persian Gulf region.

Traditionally diamonds were sold in jewelry stores. The Diamond District in New York City and the Diamond Quarter in Antwerp have been global centers for diamond trade. With the proliferation of electronic commerce, more diamonds are sold on the Internet, allowing for a broader consumer reach. While the perceived value of a diamond may be affected by the price paid [32], diamonds also have well-defined physical properties: weight, cut and color among them. In addition to expanding the potential markets for diamonds, the growing popularity of online commerce is expected to reduce the consumer search costs and make it easier for consumers to compare diamond prices in relation to diamond physical properties and do so across different retailers [1]. The reduction in the online consumer search costs and the ease of comparing physical diamond properties through online search would be expected to make diamonds more like a commodity product, whose value is determined by the physical characteristics.

The goal of our research 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. To answer this question, we perform data mining on a large dataset of diamonds available for sale at one of the largest online diamond retailers. The emergent insights contribute to our understanding of the relationship between the consumer search costs and variation in product prices for diamond goods. The results also have broader implications for the effects of online commerce on luxury goods pricing as well as on the potential strategies for retailers to overcome the pricing pressure created by e-commerce.

The paper is structured as follows. First, we provide an overview of prior research on consumer search and product price variation. Next, we review the key diamond physical properties that are known to affect diamond prices. We then discuss the literature on price obfuscation or retailer strategies to increase consumer search costs. After that, we describe the dataset, the methodology and present the data mining insights. We conclude with the discussion of the results, contributions and future research directions.

# 2 Search Costs and Price Dispersion

The extant research on consumer product information search and price dispersion provides the theoretical foundation for our work. Price dispersion - the variation in price for the same product across different retail channels - has been studied in economics [2], [38]. Price dispersion arises due to the information asymmetry and imperfect consumer information [2]. When consumers are differentially informed, firms can charge consumers different prices [38]. In order to find the price, consumers have to spend resources on search, and the search for product prices can be costly. The higher the consumer search costs, the greater the expected price dispersion in a given product category [2].

The Internet adds transparency and efficiency to the markets by providing valuable information about goods in one place [23]. It offers temporal and spatial shopping convenience and adds value through price comparison opportunities [16], [27]. It has long been proposed that the Internet reduces search costs and information asymmetry by putting all information at the consumer's fingertips [1]. The Internet is also known to increase the product selection and product availability via the Long Tail phenomenon [8]. The Long Tail refers to the fact that Internet-based retail can effectively service the demand for niche products, which are generally not offered by physical retailers due to the relatively low demand for each individual niche product. The early theoretical work on the effects of the Internet on online retailers suggested that given the low cost of online information search for consumers, online retailers would be expected to experience price competition that would eventually lead to price convergence for almost all goods offered on the Internet [1]. This is known as *the law of one price*. Subsequent research has shown that information intermediaries can make it easier for consumers to find the best price by aggregating pricing information across multiple retail channels [12], [22].

Although some studies found that price dispersion may be decreasing due to the growing popularity of e-retail [33], the empirical evidence has not always supported the theoretical predictions of price convergence driven by e-commerce [18], [28], [44]. For example, in the early stages of e-commerce, the prices for books and CDs varied as much as 50% across online retailers [1]. Similar observations have been made in the digital cameras market [3]. These findings were initially attributed to the immaturity of online markets [3], but a literature review spanning four decades,

This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762018000200103

suggests that price dispersion is the rule in markets with homogenous products in both offline and online markets and price dispersion commonly reaches 30 percent [2]. Further, theoretical arguments have been made that price dispersion may actually benefit consumers as well as vendors in some industries [34]. Recent studies across different industries show that even with the growing popularity of e-commerce significant price dispersion persists in many industries [25], [40].

# 3 Online Diamond Retail and Diamond Characteristics

One of the benefits of online diamond markets for consumers is the availability of all diamond related information in one place. Online diamond retailers also commonly enable the consumers to compare diamonds by creating dashboards with search attributes to assist in the diamond selection process. Reputable online diamond retailers provide images, detailed descriptions and certifications, and they offer money-back guarantees. The online diamond collections are vast and may include information on more than 100,000 diamonds of various shapes and attributes [45].

The most well-known attributes pertaining to diamonds are the 4Cs introduced by the Gemological Institute of America (GIA) in the 1950s - Cut, Carat, Color and Clarity. The 4Cs describe the unique qualities of each diamond and greatly influence diamond prices. Three of the 4Cs have a long history: carat weight, color, and clarity were used in the first diamond grading system created in India over 2,000 years ago [20].

The cut refers to a diamond's proportions and determines how well it reflects light. The cut scale ranges from poor to excellent. The cut of a diamond has additional three attributes: brilliance, or the amount of light reflected from a diamond; fire, or the dispersion of light into the colors of the spectrum; and scintillation, the flashes of sparkle when a diamond is moved around [21].

Carat is a standard unit of weight and corresponds to a diamond's size. One carat equals 0.2 grams. The name carat comes from the carob seed. Back in the day, traders started using carob seeds because of their fairly uniform weight to counterweight their balance scales. Only one in 1,000 diamonds weighs more than one carat. [20].

The color ranges from D for colorless to Z for a diamond with a hint of yellow or brown. Colorless diamonds have more sparkle and brilliance, thus diamonds graded D through F are considered superior and more expensive. Most color distinctions are subtle and almost unnoticeable to a human eye, but can greatly affect the price of a diamond.
The clarity corresponds to the lack of inclusions or natural flaws that a diamond has. Highly praised diamonds are flawless, and contain no inclusions or blemishes. The GIA Clarity Scale contains 11 grades from Flawless to Included. The majority of diamonds fall into categories of very slightly included (VS) or slightly included (SI) [21].

Another important attribute of a diamond is its shape. There are about ten popular shapes of the diamonds. A round shape is the most prevalent shape as it is considered to be the shape that reflects light exceptionally well. Other popular shapes, sometimes called fancy shapes, are princess, cushion, pear, radiant, marquis, asscher, oval, heart, and emerald.

In addition to 4Cs and the shape, there are many other attributes of diamonds such as polish, depth, table, symmetry, fluorescence and, of course, price that ranges from a few hundred to tens of millions of dollars. Prior research on the effects of diamond characteristics on price suggests that the diamond weight is the key factor affecting price [43] and that the degree of price dispersion increases with the diamond weight [45].

# 4 Price Obfuscation

In addition to the research focusing on the consumer search costs and overall market efficiency, there is also a parallel stream of studies that focuses on the retailer strategies for achieving price premiums and consequently increasing price dispersion. Price obfuscation refers to a number of different strategies that can be employed by retailers to make consumer search more costly and thus potentially increase price premiums. The research on price obfuscation suggests that product versioning and bundling can be used to make consumer price comparison more difficult [33]. Ellison and Wolitzky [14] suggest that online price obfuscation can take the form of providing the number of screens that a consumer must click through before the final price is known, including upgrades, shipping costs and service fees as well as the time that it takes each screen to load. Researchers have also advocated experimentation with price discrimination - offering the same product at different prices to different customers based on inferred willingness to pay [1].

While product information disclosures potentially make consumers more informed, too many details and information attributes, and the lack of understanding of price formation may complicate consumer decision-making. In the scenario when prices are available, but obfuscated by the difficulty of search, the consumers may search less, the prices of goods will be higher on average and there will be more price dispersion.

This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762018000200103

In online retail, sellers may opt for a tactic of increasing consumer search costs by presenting too much information. When consumers feel overwhelmed by search results, they may end up purchasing goods at suboptimal prices. In online diamond markets, considering the number of information attributes that a buyer has to select from, the obfuscation can result from the overwhelming effects of selecting the right diamond among the pool of available gems. For example, at the time of writing, an online diamond retailer JamesAllen.com offered 90,000+ loose diamonds for sale [31], Brilliance.com had 154,000+ loose diamonds on their website [31], and BlueNile.com's collection consisted of 165,500+ loose diamonds [31].

Considering that there are multiple attributes pertaining to each diamond, including 4Cs, price, shape, polish, fluorescence, symmetry, table, depth, pavilion depth, crown height, culet, girdle, certifying agency, etc. the total number of possible combinations of diamonds and their attributes can easily go over a million. Even after filtering diamond data by attributes, the number of possible combinations can be overwhelming. In addition, the majority of consumers purchase diamonds once or several times in their lifetime, and are at a disadvantage of building the experience and familiarity with purchasing diamonds. The lack of experience coupled with the information overload may force consumers to rely on shortcuts - eliminating attributes that they are not familiar with and assigning more value to attributes that make sense to them.

To assist consumers, almost all online diamond retailers include educational information and guidelines on how to choose diamonds starting with the diamond shape, cut, carat, color, clarity and certification being the most essential, and polish, fluorescence, symmetry, table, depth, pavilion depth, crown height, culet, girdle being offered as well. For example, according to the International Gem Society, out of 4Cs, the cut is the most important attribute of a diamond, followed by color, clarity, and the least important carat [10]. According to Blue Nile, the cut has the biggest effect on the sparkle, and even with perfect color and clarity, a poorly cut diamond will look dull [6]. However, the explanation of which particular attribute affects the price the most is generally missing. For example, the GIA explains that in addition to other attributes, prices of diamonds are affected by the fact that "some weights are considered magic sizes: half carat, three-quarter carat, one carat, etc." [20]. Diamonds whose weight is slightly above the magic one-carat size can increase the price as much as 20 percent with only a 6-point difference in weight [20].

Price obfuscation can potentially reduce a consumer's ability to fully understand the prices. While we can safely assume that larger and better quality diamonds fetch higher prices on the market, it may be difficult for consumers to ascertain the fair value of a diamond based on its physical attributes. This fact coupled with the lack of experience with diamonds for the majority of consumers, as diamond purchases are infrequent purchases, makes the understanding of prices even more challenging.

In summary, prior research examining the interplay between the consumer search costs and price dispersion revealed that the theoretical expectations of the reduction of price dispersion due the growing popularity of e-commerce and associated lower consumer search costs were not supported by the empirical evidence from a number of different product categories [9], [11], [33]. Prior research on price dispersion has been generally done comparing prices across different retail channels [9], [11]. Diamonds are a unique product category. Consumers generally possess limited knowledge about the relationship between the physical characteristics of a diamond and diamond prices. The relatively high number of diamond physical characteristics: weight, shape, cut, clarity, color, fluorescence, length, width, height, table, etc., make it difficult for consumers to evaluate the effect of each characteristic on price. In addition, diamond retailers commonly offer tens of thousands of loose diamonds for sale, further complicating consumer search. To explore the degree of diamond price dispersion in the context of a single retailer, we perform predictive data mining focusing on the predictive value of the physical characteristics in relation to diamond price in the context of a single diamond retailer. In the next section, we discuss the dataset in our study and the analytical methodology that progresses through several stages to gain insight on the predictability of diamond prices.

# 5 Data and Methodology

We obtained the dataset for our study by scraping data from an online retailer that offers one of the largest collections of diamonds available for sale. The retailer operates several online storefronts targeting customers internationally. We limited the data scraping to the web site that focuses on the consumers in the United States. We were able to collect data about 138,654 diamonds. Although the retailer offers consumers an opportunity to incorporate the purchased diamonds into rings, earrings and other types of jewelry, we focused our analysis on the diamonds themselves, because while the loose diamond prices are relatively transparent on the retailer's website, the jewelry prices were more difficult to determine due to different types of discounts available through the retailer and third parties.

Our analysis has been done in several stages. First, we present the exploratory analysis of our data. In the second stage we develop a multiple linear regression model to predict diamond prices and assess its accuracy. In the third stage we expand our predictive modeling to include several data mining techniques that are able to capture non-linear relationships among the predictor variables and price. Next, drawing on the evidence of actual diamond sales, we narrow the range of diamonds in our analysis and reassess the accuracy of data mining models and key predictors of diamond prices. In the final stage of our empirical analysis we examine specific price discontinuities in the dataset.

**Journal of Theoretical and Applied Electronic Commerce Research**
ISSN 0718–1876 Electronic Version
VOL 13 / ISSUE 2 / MAY 2018 / 15-28


This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762018000200103

# 6 Analysis and Results

We obtained a set of attributes for each diamond. These attributes included diamond weight (measured in carats), cut, color, and clarity rating as well as the shape and physical dimensions of each diamond. Table 1 summarizes key descriptive statistics for the dataset (N = 138,654).

Table 1: Dataset descriptive statistics

| Feature | Descriptive statistics / Distribution | | |
|---|---|---|---|
| Carat (weight) | Mean = 0.91, SD = 0.8, Min = 0.01, Max = 22.74 | | |
| Length/width ratio | Mean = 1.08, SD = 0.18, Min = 0.75, Max = 2.95 | | |
| Depth | Mean = 0.6, SD = 0.04, Min = 0.01, Max = 0.82 | | |
| Table | Mean = 0.60, SD = 0.05, Min = 0.17, Max = 0.85 | | |
| Cut | Good | 17510 | 12.63% |
| | Very Good | 57852 | 41.72% |
| | Ideal | 59195 | 42.69% |
| | Sig. Ideal | 4096 | 2.95% |
| Color | D | 23686 | 17.1% |
| | E | 25675 | 18.5% |
| | F | 25081 | 18.1% |
| | G | 22506 | 16.2% |
| | H | 17523 | 12.6% |
| | I | 14371 | 10.4% |
| | J | 9809 | 7.1% |
| Shape | Round | 93384 | 67.4% |
| | Princess | 13202 | 9.5% |
| | Emerald | 6647 | 4.8% |
| | Pear | 6493 | 4.7% |
| | Cushion | 5558 | 4.0% |
| | Oval | 5160 | 3.7% |
| | Asscher | 2931 | 2.1% |
| | Marquise | 1812 | 1.3% |
| | Heart | 1740 | 1.3% |
| | Radiant | 1727 | 1.2% |
| Fluorescence | None | 107840 | 77.78% |
| | Faint | 15873 | 11.45% |
| | Medium | 7632 | 5.50% |
| | Strong | 4348 | 3.14% |
| | Medium Blue | 1661 | 1.20% |
| | Strong Blue | 796 | 0.57% |
| | Very Strong | 216 | 0.16% |
| | Faint Blue | 195 | 0.14% |
| | Very Strong Blue | 55 | 0.04% |
| | Medium Yellow | 9 | 0.01% |
| | Other | 19 | 0.01% |
| Price | Mean = $7,932.94, Median = 2,345.47, SD = $30,346.77, Min = $223.00, Max = $2,818,242.16 | | |

**Journal of Theoretical and Applied Electronic Commerce Research**
ISSN 0718–1876 Electronic Version
VOL 13 / ISSUE 2 / MAY 2018 / 15-28

This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762018000200103

As it is evident from the descriptive statistics, the dataset includes a very broad range of diamonds varying in weight from 0.01 to 22.74 carats and consequently varying in price from $223.00 to over $2.8 million. A weight/price distribution plot (Figure 1) shows that relatively few diamonds are over 10 carats in size and that price variance increases with the diamond weight.

To reduce the heteroscedasticity in our dataset we limited further analysis to diamonds larger than 0.2 carats and smaller than 10 carats by removing 71 records containing diamonds outside of this range. Following the recommendations in [30] to further reduce the heteroscedasticity in the dataset, we also transformed the weight and the price of diamonds by taking a natural logarithm of these variables. Figure 2 illustrates the relationship following the transformation.

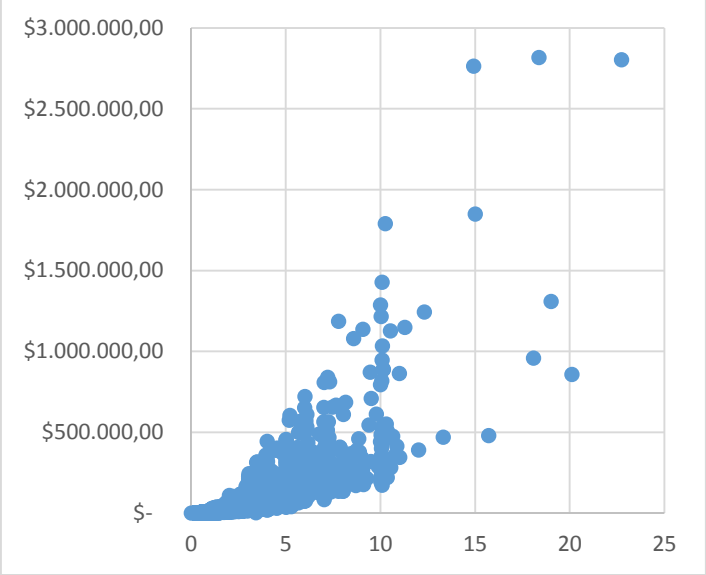


Figure 1: Price versus diamond weight, carats

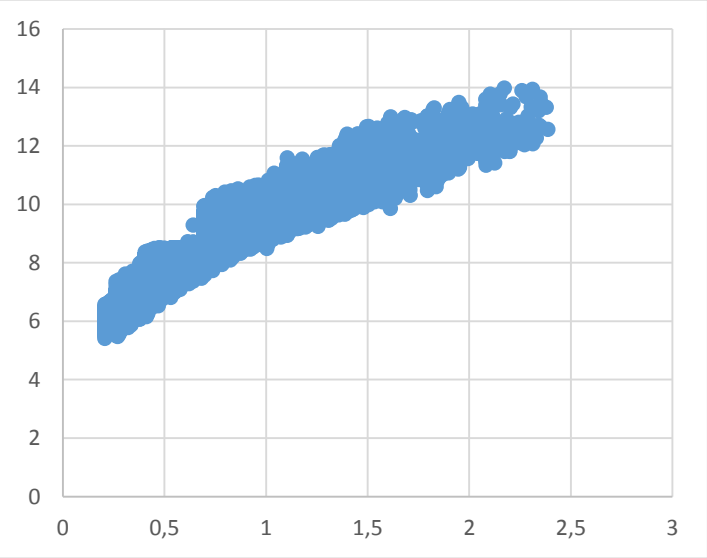


Figure 2: ln(Price) versus diamond weight, ln(carats)

In the next step of the analysis, we examined the predictability of diamond prices using a multiple linear regression model. We estimated the parameters in the following function:

$$\ln(P_i) = \alpha + \beta * X_i + \delta * \ln(W_i) + \varepsilon_i \quad (1)$$

where

- ln(Pi) is the natural log of a diamond price

**Journal of Theoretical and Applied Electronic Commerce Research**
ISSN 0718–1876 Electronic Version
VOL 13 / ISSUE 2 / MAY 2018 / 15-28

This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762018000200103

- ln(Wi) is the natural log of the diamond weight
- Xi is the vector of the diamond's features
- α, β, δ are estimated parameters
- εi is a random error term.

To assess the quality of the linear regression model, we randomly split the dataset 70/30 into training and validation subsets. We estimated the model parameters using the training dataset and we assessed the model performance by scoring the validation dataset using the model and comparing the model predictions to the actual prices. The average price, mean absolute error and mean absolute percent errors for the regression model using the validation data are provided in Table 2. The definitions of the model performance metrics: Mean absolute error and Mean absolute percent error, are provided in the Appendix A.

Table 2: Multiple regression model performance summary

| | |
|---|---|
| Average price | $ 14,987.07 |
| Mean absolute error | $ 4,556.07 |
| Mean absolute percent error | 30.4% |

The results of the multiple regression model suggest that a linear model may not be the best way to capture the effects of diamond characteristics on price. The mean absolute percent error is 30.4% implying a very high dollar price variation. In the next step of our empirical analysis we examined the ability of non-linear data mining techniques to accurately predict diamond prices.

Predictive data mining is the practice of building predictive models that can accurately forecast the target variable of interest [17]. Predictive data mining techniques can be generally separated into two general families of models. Classification models can be used to forecast a nominal target variable. Prediction models can be leveraged to forecast an interval numeric target variable. 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. We also included the multiple linear regression model to provide the baseline to evaluate the improvement in accuracy offered by the other techniques. The detailed discussion of the specific model types is beyond the scope of the current manuscript. Here we will only provide a brief overview of the individual data mining techniques.

Decision forest, also known as random forest, is an ensemble modeling technique which aggregates predictions across multiple individual decision trees [7]. Decision tree algorithms are one of the fundamental data mining techniques [36]. Several decision tree algorithms have been developed [36], but all share the general approach to building the tree. The decision tree algorithms iteratively partition the data in the training dataset and attempt to construct a set of sequential hierarchical splitting rules that can partition the dataset in such a way that reduces variance within each bucket of cases following the traversal through the decision tree. The decision rules are developed iteratively by considering potential binary data partitioning rules, for example weight <= 2 carats. Decision tree algorithms are greedy (focus on local optima) and therefore they tend to be globally suboptimal [36]. Decision trees also tend to over-fit the training data. Several ensemble data mining techniques have been developed that leverage the decision tree ability to capture non-linear relationships in the data. The decision forest technique builds multiple decision trees by subsampling data from the training dataset and also restricting the number of variables that are available for modeling within each tree [29]. The decision forest algorithm then estimates the value of the target variable by averaging the predictions of the individual tree models.

The boosted decision tree algorithm is another example of an ensemble model that is built on the foundation of the decision tree algorithm [13]. The boosted tree algorithm builds a series of decision trees, but it takes a unique approach to improving the accuracy of the ensemble model by increasing the weights assigned to the records with the largest error with each tree that is built (in our case, the error is the difference between the predicted and the actual diamond price). In other words, after building the initial tree, errors are assessed and the next tree is built to minimize the errors for the records that the first tree had the largest errors for. The process is repeated iteratively increasing the weights of the cases with the largest error in each round. Random forest and boosted decision trees afford the advantage of capturing non-linear relationships in data while safeguarding against overfitting the training dataset. This is in part accomplished by putting aside an out-of-bag subsample while building the models and assessing the improvement in model accuracy after each tree is added the model by using the out-of-bag subsample. The modeling stops when the models begin to over-fit the data as is indicated by an increasing error on the out-of-bag subsample.

Artificial neural networks (ANNs) are an entirely different approach to modeling non-linear relationships in data. Artificial neural networks evolved from the attempts to model human brain functioning [46]. The artificial neural networks are typically composed of nodes organized in input, inner and output layers. Each inner and output layer node functions

**Journal of Theoretical and Applied Electronic Commerce Research**
ISSN 0718–1876 Electronic Version
VOL 13 / ISSUE 2 / MAY 2018 / 15-28

This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762018000200103

as a processing unit receiving multiple inputs and sending a single output. The input nodes receive inputs corresponding to the predictor variables in the model. The inner layer nodes receive inputs from the input layer and transform the inputs using a mathematical function, e.g. a logistic regression, the outputs of the inner layer nodes then received as inputs by the output layer nodes. In case of modeling a single continuous numeric target variable there is a single output node. ANNs are *trained* by sending each record through the network, assessing the error on the output node and adjusting parameters affecting the output from each inner layer node iteratively with the goal of minimizing the errors. This variant of the ANNs is referred to as feedforward, error backpropagation models. ANNs can capture complex non-linear relationships in the data, but they are typically seen as *black box* models that provide little opportunity to understand how the values of input variables translate into the target prediction.

We used Microsoft Azure Machine Learning platform [47] to assess the accuracy of the data mining algorithms in relation to their ability to predict diamond prices. We split the dataset 70/30 into training and validation subsets. To compare the performance of the models we assess the average dollar and percent errors for each model using the validation dataset. Table 3 summarizes the performance statistics across the models that we examined.

Table 3: Data mining model performance summary

| | Mean absolute error | Mean absolute percent error |
|---|---|---|
| Multiple linear regression | $4,556.07 | 30.4% |
| Decision forest | $1,205.10 | 15.8% |
| Boosted regression trees | $2,209.96 | 23.0% |
| Artificial neural network | $2,261.89 | 24.3% |

Although the decision forest, the boosted regression trees and artificial neural network improved on the performance of the multiple linear regression model, all models struggled to make accurate predictions of diamond prices in our dataset. In the next step of our analysis, we decided to examine which diamonds are actually being sold with the goal of narrowing the range of diamonds in the dataset. Five weeks after the initial data collection, we again scraped the data from the retailer's website to see which diamonds in our original dataset (each is identified by a unique id on the retailer's site) were still available for sale. 57,391 of 138,654 diamonds (41.4%) were no longer available. Figure 3 shows the distribution analysis of the sold diamonds by weight. The analysis reveals that 96% of the diamonds that were sold are smaller than 2.5 carats in size.

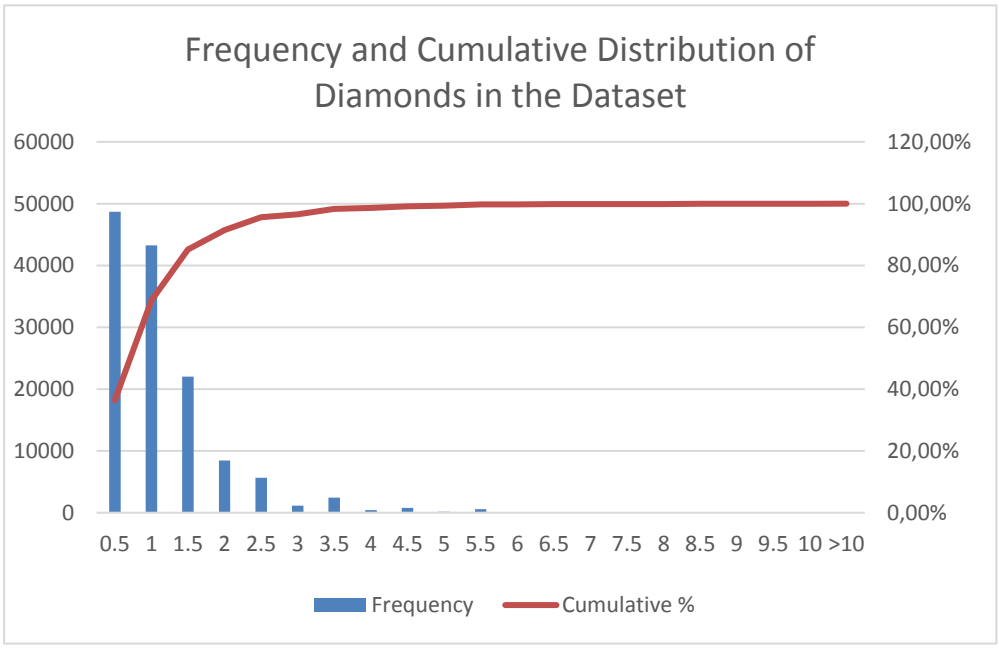


Figure 3: Number of diamonds and cumulative percentage by diamond weight, carats

Given that diamonds weighing between 0.2 and 2.5 carats constitute the bulk of the diamonds sold, we narrowed the range of diamonds in our dataset and reassessed the predictive accuracy of different data mining models. Diamonds

**Journal of Theoretical and Applied Electronic Commerce Research**
ISSN 0718–1876 Electronic Version
VOL 13 / ISSUE 2 / MAY 2018 / 15-28

This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762018000200103

weighing between 0.2 and 2.5 carats represent 95.8% of all records. Table 4 below summarizes model performance on this subset of diamonds.

Table 4: Model performance summary for 0.2 to 2.5 carat diamonds

| | Mean absolute error | Mean absolute percent error |
|---|---|---|
| Multiple linear regression | $527.07 | 9.9% |
| Decision forest | $787.86 | 13.0% |
| Boosted regression Trees | $1,163.31 | 22.6% |
| Artificial neural network | $388.35 | 8.2% |

The results in Table 4 indicate that artificial neural network model delivers the best performance using diamonds data ranging from 0.2 to 2.5 carats in size based on both metrics. The ANN model has the lowest mean absolute error and mean absolute percent error. In the next step, we examined the relative contribution of different diamond characteristics to the price by using the permutation based feature importance approach [41]. Briefly, the approach examines the effects of each variable on the model accuracy by withholding each of the variables in turn and assessing the effect on the average model error. Figure 4 below summarizes the relative importance scores for the individual diamond characteristics for the accuracy of the artificial neural network model which was the best performing model in the previous step of our analysis.

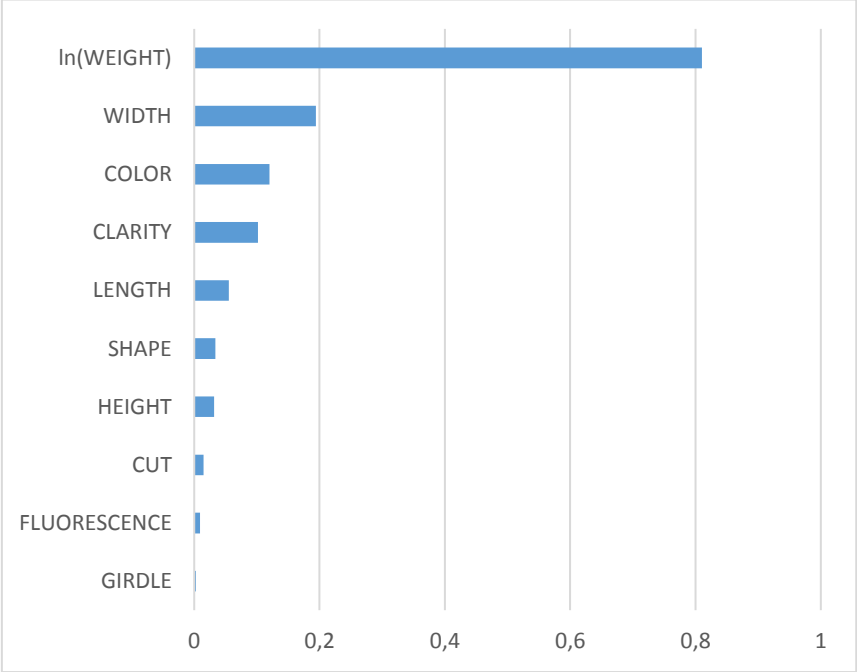


Figure 4: Diamond feature importance in influencing diamond prices

The results of the analysis (Figure 4) 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 [39]. The half-carat and full carat size diamonds are particularly popular. We decided to explore the price distributions of diamonds by weight specifically focusing on these weight ranges.

**Journal of Theoretical and Applied Electronic Commerce Research**
ISSN 0718–1876 Electronic Version
VOL 13 / ISSUE 2 / MAY 2018 / 15-28


This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762018000200103

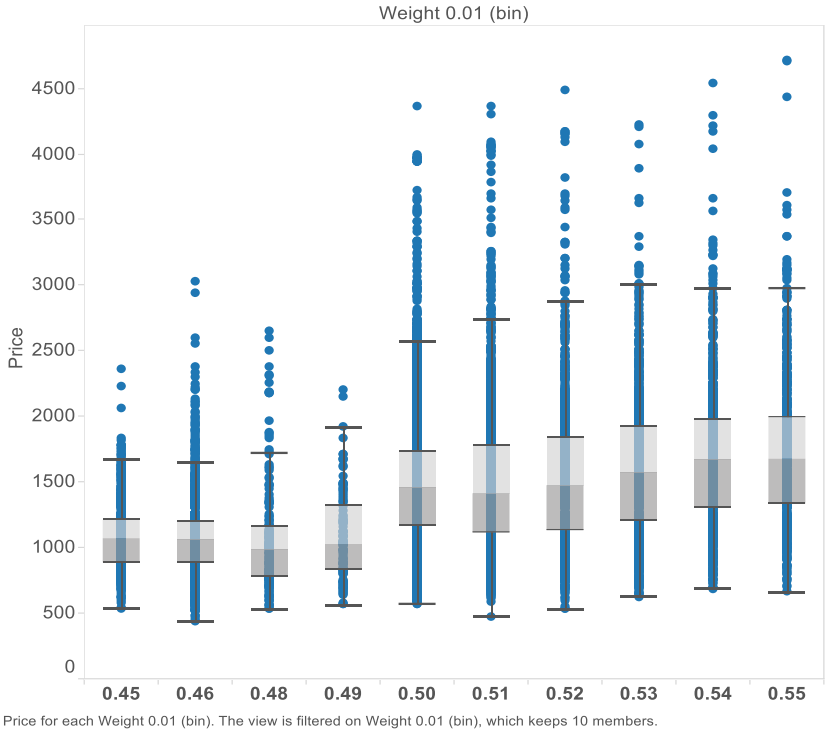


Figure 5: Price distribution by diamond weight, 0.45-0.55 carat

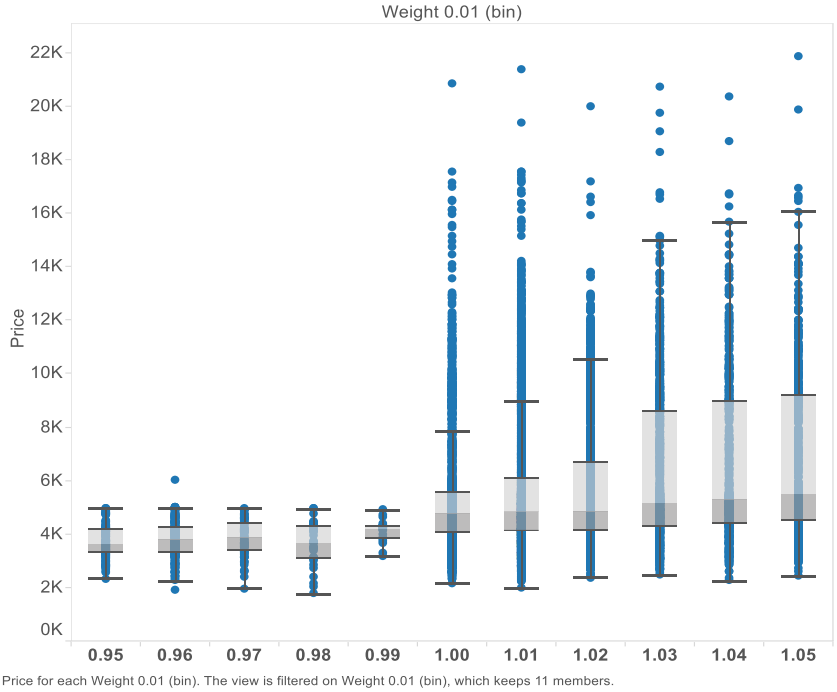


Figure 6: Price distribution by diamond weight, 0.95-1.05 carat

Figures 5 and 6 show the price dispersion using box-and-whisker plots for diamonds binned by 0.01 carat weight. The visualizations show dramatic increases in diamond price dispersion at the half and full-carat weights. To further evaluate the price dispersion, we calculated a measure of dispersion as a function of the diamond weight by dividing the standard deviation for the diamond prices by the average value for each weight range. Table 5 summarizes these results. The dispersion metric increases drastically from 0.11-0.23 range for diamonds 0.95-0.99 carats to 0.38-0.47 for diamonds 1-1.05 carats in weight.

**Journal of Theoretical and Applied Electronic Commerce Research**
ISSN 0718–1876 Electronic Version
VOL 13 / ISSUE 2 / MAY 2018 / 15-28

This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762018000200103

Table 5: Average price, standard deviation and dispersion as a function of diamond weight

| Weight in carats | Average Price | Standard Deviation | Coefficient of Variation |
|---|---|---|---|
| 0.95 | $3,693.01 | $615.60 | 0.17 |
| 0.96 | $3,759.68 | $660.85 | 0.18 |
| 0.97 | $3,847.52 | $671.52 | 0.17 |
| 0.98 | $3,621.83 | $814.99 | 0.23 |
| 0.99 | $4,091.72 | $447.43 | 0.11 |
| 1 | $5,222.84 | $2,008.89 | 0.38 |
| 1.01 | $5,602.56 | $2,394.06 | 0.43 |
| 1.02 | $5,718.56 | $2,451.88 | 0.43 |
| 1.03 | $6,376.02 | $2,951.17 | 0.46 |
| 1.04 | $6,639.26 | $3,092.23 | 0.47 |
| 1.05 | $6,912.87 | $3,119.90 | 0.45 |

# 7 Discussion

Drawing on the theoretical and empirical research on the relationship between consumer search costs and price dispersion, we investigated the predictability of diamond prices as a function of diamond physical properties by applying predictive data mining techniques on a dataset of 138,654 diamonds scraped from a leading online diamond retailer. The analysis of the effects of individual diamond physical properties on price shows that diamond weight and associated dimensions (length, width, height) play the largest role, while color, clarity and shape are also important. The initial data mining results produced relatively high errors, which were reduced by narrowing the range of diamonds between 0.2 and 2.5 carats in weight. This range selection is based on the empirical evidence that these types of diamonds are most frequently bought by consumers online. The artificial neural network model delivered the best results with the mean absolute percent error of 8.8%. In other words, even a sophisticated modeling algorithm would struggle to make an accurate price forecast based on diamond physical characteristics.

The relatively high degree of error in our data mining models implies a significant degree of subjectivity and consequently price dispersion in diamond prices. We uncovered further evidence of price dispersion by examining the price distributions around half-carat and one carat diamond weights, which are known to be favored by consumers. Our results indicate that price dispersion increases markedly for diamonds equal to or larger than these critical values. The standard deviation of diamond prices reaches 47% of the average price for diamonds weighing 1.04 carats. This is rather remarkable given that our data is obtained from a single retailer where consumers do not even need to leave the retailer web site to compare different diamonds.

Our study makes a number of contributions to theory. First, while price dispersion has been explored in different markets, it was always investigated by comparing prices across different retailers [2], [11], [33], [40]. We find a unique context where significant price dispersion is apparent on the web site of single online retailer. The finding of significant price dispersion on a single retailer's web site is consistent with the early theoretical arguments suggesting that even a monopolist can achieve excess profits by introducing variation in prices for its product [37].

The observation of price variation in the context of a single retailer introduces an important nuance into the existing theoretical perspectives on consumer decision-making. Consumers generally want to see choices being available before making a purchase decision [35]. The online retailer that is the subject of our study certainly delivers choices to consumers. For example, there are 4,965 different 1-carat diamonds in our dataset. While satisfying the consumer desire for choice, the retailer also creates a significant search cost for consumers to identify the best available option. It would be very difficult for a consumer to evaluate all available 1-carat diamonds. Exposure to so much information may lead to information overload [15]. Prior studies on consumer information overload in e-retail contexts suggest that exposing consumers to too many choices may cause confusion, less confidence in the choices made, and lower overall satisfaction [24], [26]. These potential effects of so many choices in online retail require further investigation, as it is a growing phenomenon. Amazon.com frequently lists dozens of potential vendors that can supply a particular product at different prices on the same page. TripAdvisor also lists dozens of hotels and multiple potential booking partners with different prices on the same page. It would be important to examine the optimal presentation of such choices to avoid consumer dissatisfaction.

This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762018000200103

Our insights for practice emerge from data mining results. The results suggest that diamond retailers may be employing price obfuscation through high price dispersion to increase consumer search costs. Furthermore, the diamond cut is often lauded by diamond retailers as the most important product attribute [6]. Our data mining results reveal that in addition to diamond weight and dimensions, color and clarity have more pronounced effects on price than cut. By misdirecting the consumers towards the diamond property that is difficult to assess online and which actually plays a minor role in influencing diamond prices, the retailers likely further complicate the consumer search process.

Lastly, we should note that no research is without limitations. We conducted an exploratory study using a dataset from a single online diamond retailer. The insights may not necessarily generalize to other contexts. For example, it is likely that the consumer experience of shopping for diamonds offline would likely be very different. Anecdotal evidence suggests, offline diamond business is relatively opaque [31]. Our study is methodologically limited to the data that were able to obtain from the retailer's website. These data do not yield any information on the consumer motivation for purchasing diamonds. For example, consumers could be purchasing diamonds as an investment and such purchases may be affected by a different set of considerations beyond the diamond attributes that we considered in our study. Further research would be required to assess the generalizability of our observations.

# References

[1] Y. Bakos, Reducing search costs : Implications for electronic marketplaces, Management Science, vol. 43, no. 12, pp. 1676-1692, 1997.
[2] M. R, Baye, J. Morgan and P. Scholten, Information, search, and price dispersion, Economics and Information Systems, vol. 1, pp. 323-373, 2006.
[3] K. Baylis and J. M. Perloff, Price dispersion on the Internet: Good firms and bad firms, Review of Industrial Organization, vol. 21, no. 3, pp. 305-324, 2002.
[4] De Beers Group of Companies. (2015) Global consumer demand and diamond jewellery retail. De Beers Group. [Online] Available: https://www.debeersgroup.com/content/dam/de-beers/corporate/images/insight-report/pdf/DeBeers_Insight_Report_2015_GlobalConsumerDemand.pdf.downloadasset.pdf.
[5] De Beers Group of Companies. (2016) Diamond Jewellery Demand and Outlook 2016. De Beers. Group. [Online]. Available: http://www.debeersgroup.com/en/reports/insight/insight-reports/insight-report-2016/outlook.html.
[6] BlueNile. (2016) Choose your diamond. Blue Nile Web Site. [Online]. Available: http://www.bluenile.com/education/rings/engagement-ring-guide/choose-diamond.
[7] L. Breiman, Random forests, Machine Learning, vol. 45, no. 1, pp. 5-32, 2001
[8] E. Brynjolfsson, Y. Hu and D. Simester, Goodbye Pareto principle, hello long tail : The effect of search costs on the concentration of product sales, Journal Management Science, vol. 57, no. 8. Pp. 1373-1386, 2011.
[9] A. Chandra, Consumer search and dynamic price dispersion : An application to gasoline markets, The Rand Journal of Economics, vol. 42, no. 4, pp. 681-704, 2011.
[10] D. Clark. (2016) How to choose a diamond. International Gem Society. [Online]. Available: https://www.gemsociety.org/article/choosing-a-diamond/.
[11] E. K. Clemons, I.-H. Hann and L. M. Hitt, Price dispersion and differentiation in online travel: An empirical investigation, Management Science, vol. 48, no. 4, pp. 543-549, 2002.
[12] W. Duan, Analyzing the impact of intermediaries in electronic markets: An empirical investigation of online consumer-to-consumer (C2C) auctions, Electronic Markets, vol. 20, no. 2, pp. 85-93, 2010.
[13] J. Elith, J. R. Leathwick and T. Hastie, A working guide to boosted regression trees, Journal of Animal Ecology vol. 77, no. 4, pp. 802-813, 2008.
[14] G. Ellison and A. Wolitzky, A search cost model of obfuscation, RAND Journal of Economics, vol. 43, no. 3, pp. 417-441, 2012.
[15] M. J. Eppler and J. Mengis, The concept of information overload: A review of literature from organization science, accounting, marketing, MIS, and related disciplines, The Information Society, vol. 20, no. 5, pp. 325-344, 2004.
[16] S. A. Eroglu, K.A. Machleit and L. M. Davis, Atmospheric qualities of online retailing: A conceptual model and implications, Journal of Business Research, vol. 54, no. 2, pp. 177-184, 2001.
[17] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, From data mining to knowledge discovery in databases, AI Magazine, vol. 17, no. 3, p. 37, 2006.
[18] L. Gan, S. He, T. Huang, and J. A. Tan, Comparative analysis of online grocery pricing in Singapore, Electronic Commerce Research and Applications, vol. 6, no. 4, pp. 474-483, 2007.
[19] GIA. (2016) GIA gem encyclopedia. GIA Website. [Online]. Available: http://www.gia.edu/diamond.
[20] GIA. (2016) Diamond quality factors. GIA Website. [Online]. Available: http://www.gia.edu/diamond-quality-factor.
[21] GIA. (2016) GIA 4Cs. GIA Website. [Online]. Available: http://www.gia.edu/gia-about/4Cs-Cut.
[22] M. F. Gorman, W. D. Salisbury and I. Brannon, Who wins when price information is more ubiquitous? An experiment to assess how infomediaries influence price, Electronic Markets, vol. 19, no. 2-3, pp. 151-162, 2009.
[23] N. F. Granados, A. Gupta and R. J. Kauffman, Online and offline demand and price elasticities : Evidence from the air travel industry, Information Systems Research, vol. 23, pp. 164-181, 2012.
[24] T. T. Hills, T. Noguchi and M. Gibbert, Information overload or search-amplified risk? Set size and order effects on decisions from experience, Psychonomic Bulletin & Review, vol. 20, no. 5, pp. 1023-1031, 2013.
[25] G. Kaplan, G. Menzio, L. Rudanko, and N. Trachter, Relative price dispersion: Evidence and theory, The National Bureau of Economic Research, New York, Working Paper, 2016.

This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762018000200103

[26] B. K. Lee and W. N. Lee, The effect of information overload on consumer choice quality in an on-line environment, Psychology & Marketing, vol. 21, no. 3, pp. 159-183, 2004.
[27] H. G. Lee, S. C. Lee, H. Y. Kim, and R. H. Lee, Is the internet making retail transactions more efficient?: Comparison of online and offline CD retail markets, Electronic Commerce Research and Applications, vol. 2, no. 3, pp. 266-277, 2003.
[28] B. Li and F. F. Tang, Online pricing dynamics in Internet retailing: The case of the DVD market, Electronic Commerce Research and Applications, vol. 10, no. 2, pp. 227-236, 2011.
[29] A. Liaw and M. Wiener, Classification and regression by random forest, R News, vol. 2, no. 3, pp. 18-22, 2002.
[30] W. G. Manning, The logged dependent variable, heteroscedasticity, and the retransformation problem, Journal of Health Economics, vol. 17, no. 3, pp. 283-295, 1998.
[31] L. Mongelli, K. Fasick and E. Saul, Diamond dealer ruined the biggest purchase of my life, New York, New York Post, 2016.
[32] Y. -K. Ng, Diamonds are a government's best friend: burden-free taxes on goods valued for their values, The American Economic Review, vol. 77, no. 1, pp. 186-191, 1987.
[33] X. Pan, B. T. Ratchford and V. Shankar, Price dispersion on the internet: a review and directions for future research, Journal of Interactive Marketing (John Wiley & Sons), vol.18, no. 4, pp. 116-135, 2004.
[34] T. Rayna, J. Darlington and L. Striukova, Pricing music using personal data: mutually advantageous first-degree price discrimination, Electronic Markets, vol. 25, no. 2, pp. 139-154, 2015.
[35] E. Reutskaja and R. M. Hogarth, Satisfaction in choice as a function of the number of alternatives: When goods satiate, Psychology & Marketing, vol. 26, no. 3, pp. 197-203, 2009.
[36] S. R. Safavian and D. Landgrebe, A survey of decision tree classifier methodology, IEEE Transactions on Systems, Man, and Cybernetics, vol. 21, no. 3, pp. 660-674, 1991.
[37] S. Salop, The noisy price monopolist : Imperfect information, price dispersion and price discrimination, The Review of Economic Studies, vol. 44, no. 3, pp. 393-406, 1977.
[38] S. Salop and J. E. Stiglitz, Bargains and ripoffs: A model of monopolistically competitive price dispersion, The Review of Economic Studies, vol. 44, no. 3, pp. 493-510, 1977.
[39] F. Scott and A. Yelowitz, Pricing anomalies in the market for diamonds: Evidence of conformist behavior, Economic Inquiry, vol. 48, no. 2, pp. 353-368, 2010.
[40] A. Sengupta and S. N. Wiggins, Comparing price dispersion on and off the internet using airline transaction data, Review of Network Economics, vol. 11, no. 1, pp. 1-36, 2012.
[41] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, Conditional variable importance for random forests, BMC Bioinformatics, vol. 9, no.1, p. 1, 2008.
[42] C. Sullivan. (2013) How diamonds became forever. The New York Times. [Online]. Available: http://www.nytimes.com/2013/05/05/fashion/weddings/how-americans-learned-to-love-diamonds.html?_r=1.
[43] N. Vaillant and F. Wolff. (2013) Understanding Diamond Pricing Using Unconditional Quantile Regressions. HAL. [Online]. Available: https://halshs.archives-ouvertes.fr/halshs-00853384/document
[44] Z. Walter, A. Gupta and B.-C Su, The sources of on-line price dispersion across product types: An integrative view of on-line search costs and price premiums, International Journal of Electronic Commerce, vol. 11, no. 1, pp. 37-62, 2006.
[45] F.-C. Wolff, Does price dispersion increase with quality? Evidence from the online diamond market, Applied Economics, vol. 47, no. 55, pp. 5996-6009, 2015.
[46] B. Yegnanarayana, Artificial Neural Networks. New Delhi: Prentice-Hall of India, 2009.
[47] Microsoft Azure. (2016) Azure machine learning studio. Microsoft Azure. [Online]. Available: https://azure. microsoft. com/en-us/services/machine-learning/.

**Journal of Theoretical and Applied Electronic Commerce Research**
ISSN 0718–1876 Electronic Version
VOL 13 / ISSUE 2 / MAY 2018 / 15-28


This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762018000200103

# Appendix A

## Definitions of Model Accuracy Metrics

Mean absolute percent error (MAPE)

$$MAPE = \frac{100}{n} \sum_{k=1}^{n} \left| \frac{P_a - P_p}{P_a} \right| \quad (2)$$

Mean absolute error (MAE)

$$MAE = \frac{\sum_{k=1}^{n} \left| P_a - P_p \right|}{n} \quad (3)$$

Where $P_a$ is the actual price, $P_p$ is the predicted price and $n$ is the number of records.