International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
Volume 12– No.24, October 2019– www.ijais.org

Price Premiums Prediction using Classification and Regression Trees (CART) Algorithm in eBay Auctions

Mofareah Bin Mohamed
Master Student King Abdul Aziz University
Jeddah, Saudi Arabia

Mahmoud Kamel, PhD
Assistant Professor, King Abdul Aziz University
Jeddah, Saudi Arabia

ABSTRACT
Price premiums in internet auction are the percentage that the sale price exceeds or decreases the average price of this product, so, if the sale price exceeds the average price then the internet auction is price premiums otherwise it is non-price premiums.

The objective of the study is to analyze eBay auctions data using Classification and Regression Trees (CART), which is a type of decision trees induction. The information about previous auctions of a specific product was collected from the eBay site to the extent of its users and comprehensive, and the formulation of the previous information in the form of variables can be statistical operations on the processing of decision trees algorithms.

This study identifies the critical variables and ranks them according to their importance using the decision-making tree algorithms.

General Terms
Pattern Recognition, Algorithms, Artificial intelligence, Internet Auction.

Keywords
CART, Price premiums, eBay, K-fold-cross.

1. INTRODUCTION
eBay is leadership of e-commerce and internet auction that used to sell and buy goods and services over worldwide., with more than 168 million active buyers, and more than 20 million active sellers with billions$ transactions [1].

Price premiums in the Internet auction are defined as "the monetary amount above the average price received by multiple sellers for a certain matching product" [2].

CART is a type of a decision tree, it is classified as supervised machine learning, which means the category of each sample in data is known, in this study, there are two categories of data, reach price premiums or do not reach price premiums. This technique is for classification, so it is suitable for categorical variables like yes, no, always little, rarely and others, and also CART for regression which is suitable for continuous variables [3].

Predicting the final price of an internet auction is a critical problem, a seller who starts auction need to get a reasonable price and achieve a positive price premium. Many websites precede online auction services like Amazon, Yahoo, and eBay [4].

2. RELATED WORK
Stochastic Differential Equation (SDE) is used to modelling eBay price [5], in this modern research authors using SDE to represent price velocity and accelerator; their dataset is 93 7-days auctions of Microsoft X-box gaming system, 63 samples for training and 30 samples for test. Variables estimated as continues data; in fact, some variables can't be real like deliver option or available count and others. They use principal differential analysis, which is suitable for continues data; they approved that using SDE is better than ordinary differential equation ODE approach [6].

Price Prediction and Insurance for Online Auctions [7], is a service guarantee minimum end-price for the sellers in auctions like that in eBay, according to this study, statistics show that auction with reserve price option get worse price at the end, so it suppose that price insurance service is a right choice and has benefit for sellers rather than using reserve price premium from eBay that offers for some fee. The proposed service is trained by seller characteristics, product or functional characteristics, service providers used crawler software to collect data from eBay website, and crawler collected listings for many categories, over 2 months. Feature extraction techniques are applied to extract seller, items, and auction features, after that, training samples and using multiple classification regressions artificial intelligent algorithms.

Dynamic Price Forecasting in Simultaneous Online Art Auctions [8], it deals with multiple on-line auctions for the same product, computations conditions and information are fundamental to this model. Also, it depends on the price dynamics, not on the static data like initial price and seller reputation and others. But in fact also static data is essential, since initial price can be variable and depends on seller opinion, not on eBay policy, also the quality of product's images can change the characteristics of the product and enhance them From the point of view of the seller and thus increase the bidding price especially on the early stage of auction. Authors discussed the source of price dynamics in the term of buyer's competitions, Simultaneous Online Auctions (SOA) differs than eBay that sells multiple products at the same time, and all auctions start and end with the same time.

Kernel principal component analysis and support vector machines for stock price prediction [17], is a method for market complex data, using a neural network to solve non-parametric price prediction models, the power of this study is using a neural network to mapping non-linear data and approximate the end-price regardless of any assumptions about the data. The study targets stocks rather than auctions, but the objective is same, it predicts the price, authors generated over than hundred indicators which are investors for seller and buyers, the study tries to find the relation between these indicators and stock prices.

3. METHODOLOGY

The main contribution of this research is extracting auctions' variables, applying a CART algorithm analyzer on the data to construct the decision tree of training samples. After building a decision tree for a specific auction product, a final price of a new incoming auction of this product can be predicted to help the seller in making the decision about selling his product by giving him hints and information about the initial auction parameters like opening price, auction period, auction ending time, etc.

The proposed algorithm of predicting price premium or the final bid price in e-bay electronic business-to-consumer (B2C) auction consists of two phases:

- Training phase for constructing a decision tree.
- Testing phase for determining initial auction parameters.

3.1 Training phase

In the training phase, a product is targeted if there are enough auctions at e-bay website for a specific period of days (30 days). Then, the researcher can gather the information about this product's auctions and select it as appropriate target product to construct a decision tree for it, of course, more auctions add more training data which leads to more accurate results.

Auction information contains variables that can be extracted as data samples. After that, statistics operation were performed on data to calculate the mean and the standard deviation for numerical variables, and frequencies for categorical variables.

Now the algorithm uses Classification and Regression tree analysis CART algorithm analyzer to constructing the decision tree [9], by selecting the most critical variables and ignoring the less important variables. The previous process is a classification-tree analysis for numerical data and regression-tree analysis for categorical data.

CART algorithm requires two dependent variables as a criterion variable to split data, and construct decision rules, the dependent variable of classification-tree analysis is the final bid price, and the dependent variable of regression-tree analysis is the number of bids since it is another metric of reach the price premium.

After that, algorithm stores the decision rules list in format that can be visualizes quickly by the computer application later, other information required to be stores like the product name or id, period of auctions (date of initial auction-date of final auction), number of price premium (PP) auctions and the number of non-price premium (NPP) auctions.

The main advantage of the result tree is that the human visualized and accessible to understood by ordinary persons without knowledge in computer science or economic fields.

Figure 1 illustrates the training process or phase, where data is collected about the target product, and the variables are extracted and preprocessing by applying statistics operations on it. After that, it is the turn to analyze data by CART algorithm to construct the classification tree and regression tree.

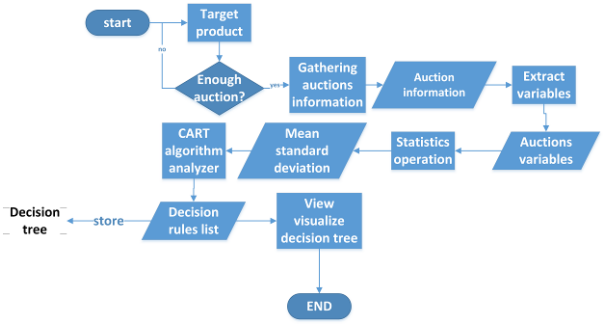

Fig 1: Proposed algorithm train phase data flowchart diagram

3.2 Testing phase

As any data mining artificial intelligent algorithm, the proposed algorithm requires training phase to construct the model of data samples, CART algorithm is a non-parametric procedure which is suitable for dealing with complex and non-linear relationships between variables as in this case of e-bay auctions. The process of testing is straightforward; the user selects the product that he wants to sell. Moreover, the system fetches decision rules that are stored during the training phase and draw a visualizing decision tree that can be scanned by a human.

In addition to decision rules and decision tree, another useful information will be shown too like the period of auctions, price premium and non-price premium percentages. It is not important how long the training process goes on, but what matters is that the testing phase needs to be short, fast, reliable and accurate; the proposed algorithm shows decision rules or tree for each product separated from others according to user (seller) selection.

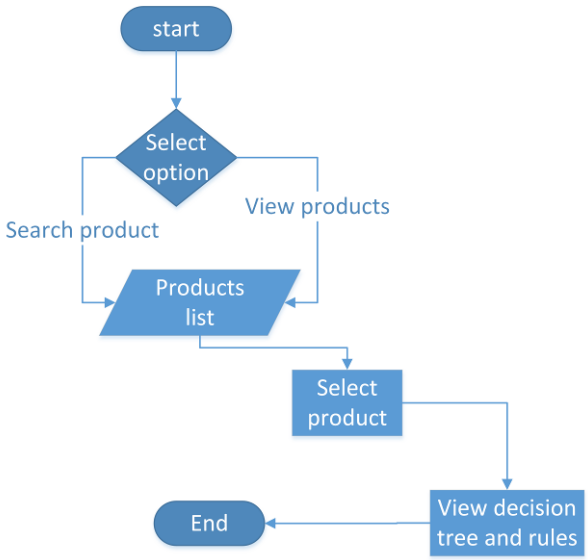

Fig 2: Proposed algorithm test phase data flowchart diagram.

3.3 E-bay auctions data variables

E-bay website is rich with economic data resource that is targeted by economist, traders, and scientific. There are many datasets which are gathering during years [10], researchers can target e-bay website by their self to collect data and needed information.

This research has to do with helping the seller to sell their goods and products, so the variables that are corresponding to the buyers or bidders are not in the list of this research, the variables used in this research are categorized as:

- Selling information variables.
- Seller information variables.
- Product information variables.
- Delivery information variables.

The following figure illustrates the eBay auction page.

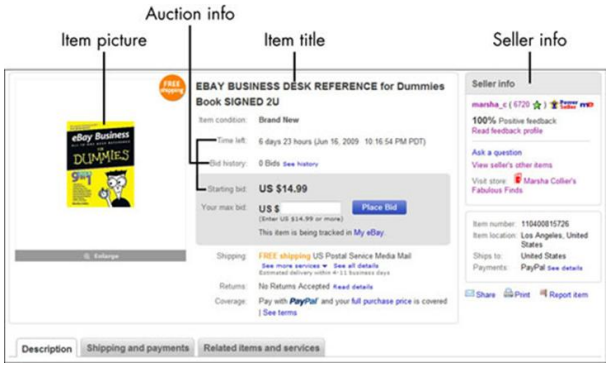

Fig 3: e-Bay auction page

3.4 CART algorithm

Many articles refer to CART algorithm with simple decision tree, CART algorithm is supervised machine learning algorithm, where the label of class for a given data sample is known, classification using supervised machine learning algorithm is powerful technique where there are available data samples with attributes which are called features or variables, usually these variables are independent, so any variable has its effect on building the model of the class. In some cases, the class is parametric, which means using the statistics and the probability theory we can build a mathematical model that represents the class.

Building class model uses training data with a known class label, after that any new sample data without a class label can be classified with this model using the mathematical model, sometimes when the number of features is vast, features reduction is needed to reduce the number of features.

The data of eBay is nonparametric, and it contains categorical, continues real, and integer variables data types. For eBay auctions data, there are two categories: price premium (PP) and non-price premium (PPP). The price premium is: "The monetary amount above the average price received by multiple sellers for a certain matching product".

CART is a nonparametric procedure, which is used to conduct the optimal decision tree, the main advantage of CART is supporting certain data types (Classification) and real or continues data type (regression). CART is simple to understand and visualize, and it is very used for features and variables selection, in this study, the variables that have the most effect on getting price premium to need to be selected. CART needs little data preparation before using it. The nonlinear relationships between parameters do not affect CART performance.

CART process consists of:

1. Features or variable to choose.
2. Conditions for splitting.
3. Stopping criteria.
4. Pruning.

3.5 Proposed algorithm evaluation

K-fold cross-validation is statistics method for classifier accuracy validation; the original sample is randomly partitioned into k equal sized subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining k − 1 subsample are used as training data. The cross-validation process is then repeated k times, with each of the k subsamples used exactly once as the validation data. The k results can then be averaged to produce a single estimation. The advantage of this method over repeated random sub-sampling is that all observations are used for both training and validation, and each observation is used for validation exactly once. 10-fold cross-validation is commonly used [11], but in general, k remains an unfixed parameter.

For example, setting k = 2 results in 2-fold cross-validation. In 2-fold cross-validation, we randomly shuffle the dataset into two sets d0 and d1, so that both sets are the equal size (this is usually implemented by shuffling the data array and then splitting it in two). Then, we train on the d0 as well as validate on the d1, followed by the training on d1 and validating on d0.

if k = n (observations' number), the k-fold cross-validation is exactly the leave-one-out cross-validation.

In stratified k-fold cross-validation, the folds are chosen so that the value of mean response is nearly equal in the whole folds. In the binary classification's case, this means that every fold does roughly contain the same proportions of the two kinds of labels of class.

In this study, the 10-fold cross-validation was used, so nine subsets are used as a training sample, and one subset is used as a test sample. In the10-fold cross-validation process, the data are divided into approximately 10 equal subsets, where subsets are determined by random sampling on the criterion variable, and the tree-growing process is repeated 10 times.

4. DATASET

4.1 Data collection

The dataset which is used in this research, was collected from the eBay U.S. website [12], eBay is a leadership of electronic auctions industry, researchers prefer real dataset rather than artificial data, using real data increase the realism of the study, and the data field is already generalized to the marketplace, and by using this real data field need not extra optimization methods [13].

The data was collected from eBay for a month, the target product is Palm Pilot M515 PDA coloured handled, all items in this study are brand news, never-used so in each category all items are same, so no bias for any auction and no item has extra value over similar products.

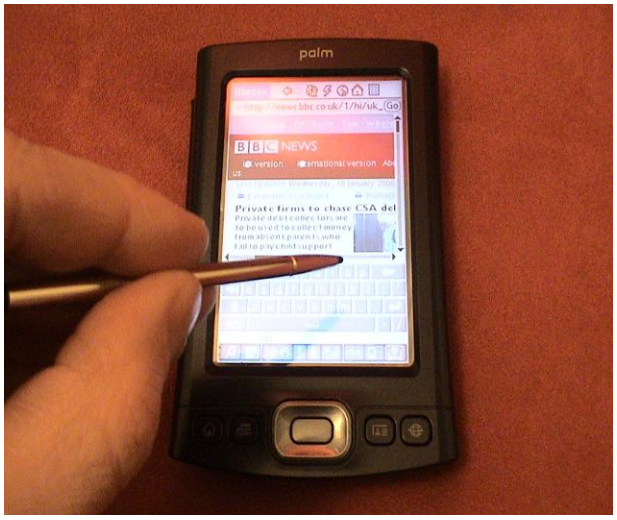

Fig 4: Palm Pilot M515 PDA coloured handled

4.2 Data statistics

In order to pre-process the data, statistics operations were applied to auctions' data to get the descriptive statistics which are the mean and standard deviation for continuous or real (float values) data variables, and the frequencies of the categorical (auction end time 4) data variables, the following table summarize the descriptive statistics of data field:

Table 1: Descriptive statistics

	Palm Pilot M515 PDA	
	Mean	**St.dev**
Criterion dependent variable		
Final Bid Price (FBP)	229.4409	21.9659
Independent numerical variables		
Initial Bid Price	78.1442	92.1557
Auction Duration	5.5882	1.74
Number of Bids	17.3706	11.2954
Independent categorical variables		
	Frequencies	
Auction Ending Time:		
1) Weekday Morning	105	
2) Weekday Afternoon	129	
3) Weekend Morning	60	
4) Weekend Afternoon	46	

5. EXPERIMENTS AND RESULTS

This section illustrates the experimental results of the proposed algorithm to construct the decision tree of eBay auctions for the seller in term of getting price premium.

5.1 Palm Pilot M515 PDA auctions decision tree

The CART algorithm which is implemented with classregtree MATLAB function, this dataset contains 340 auctions data samples. The number of Price Premium PP auctions is 178 auctions, the number of Non-Price Premium NPP auctions is 162 auctions. The following figure 5 shows the decision tree using CART algorithm, the performance of the decision tree is 94.12%.

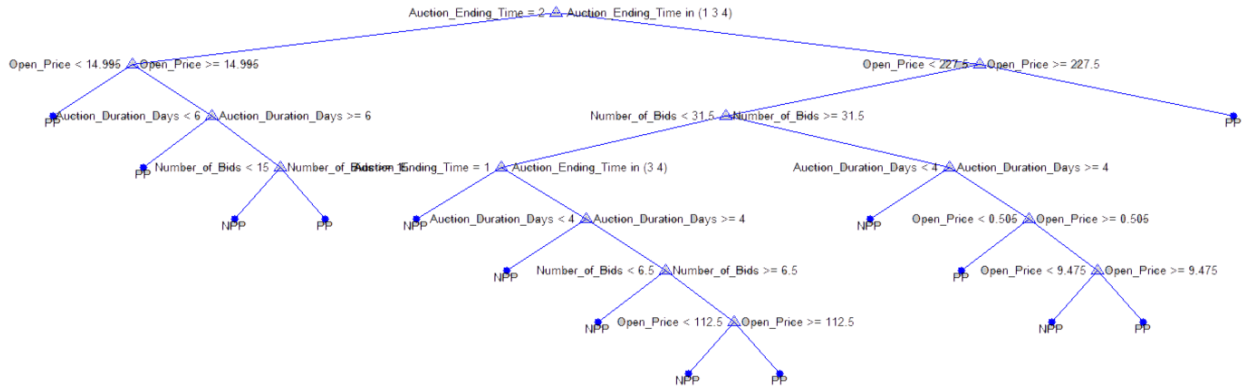

Fig 5: The decision tree result

The following code illustrates Decision tree for classification

```
1. if Auction_Ending_Time=2 then node 2 elseif Auction_Ending_Time in {1 3 4} then node 3 else PP
2. if Open_Price<14.995 then node 4 elseif Open_Price>=14.995 then node 5 else PP
3. if Open_Price<227.5 then node 6 elseif Open_Price>=227.5 then node 7 else NPP
4. class = PP
5. if Auction_Duration_Days<6 then node 8 elseif Auction_Duration_Days>=6 then node 9 else PP
6. if Number_of_Bids<31.5 then node 10 elseif Number_of_Bids>=31.5 then node 11 else NPP
7. class = PP
8. class = PP
9. if Number_of_Bids<15 then node 12 elseif Number_of_Bids>=15 then node 13 else PP
10. if Auction_Ending_Time=1 then node 14 elseif Auction_Ending_Time in {3 4} then node 15 else NPP
```

```
11. if Auction_Duration_Days<4 then node 16 elseif Au
    ction_Duration_Days>=4 then node 17 else NPP
12. class = NPP
13. class = PP
14. class = NPP
15. if Auction_Duration_Days<4 then node 18 elseif Au
    ction_Duration_Days>=4 then node 19 else NPP
16. class = NPP
17. if Open_Price<0.505 then node 20 elseif Open_Pric
    e>=0.505 then node 21 else PP
18. class = NPP
19. if Number_of_Bids<6.5 then node 22 elseif Number
    _of_Bids>=6.5 then node 23 else NPP
20. class = PP
21. if Open_Price<9.475 then node 24 elseif Open_Pric
    e>=9.475 then node 25 else NPP
22. class = NPP
23. if Open_Price<112.5 then node 26 elseif Open_Pric
    e>=112.5 then node 27 else PP
24. class = NPP
25. class = PP
26. class = NPP
27. class = PP
```

In this experiment, the algorithm results were validated using k-fold-cross-validation algorithm, k value is preferred to be 10 [14], so data is divided into ten subsets, 9 subsets for training and the remaining subset set for testing, and repeat the experiment 10 times by selecting a new subset as test subset and the remainder 9 subsets for training, by this way 10 pruned decision trees were generated.

The proposed algorithm using CART algorithm selected the order of the four variables according to the highest effect:

1. Auction Ending Time Period
2. Initial Bid Price (open price)
3. Number of Bids
4. Auction Duration Days

The following table summarizes the 10-fold-cross-validation; the maximum value is 1, which means the accuracy is 100%.

Table 2: 10-fold-cross validation

0.79	0.94	0.88	0.88	0.85	0.79	0.82	0.82	0.82	0.91

The average performance=0.85=85%.

5.2 Results discussion

In Palm Pilot M515 PDA coloured handled dataset the proposed algorithm selected 4 variables as the essential variables that listed at table 1, the result shows that if the seller wants to sell a product like Palm Pilot M515 PDA coloured handled which has average price about 230 $.

The decision tree has two main branches, right branch when the auction ending time is in [Weekday Morning, Weekend Morning, Weekend Afternoon], and left branch when the auction ending time is Weekday Afternoon.

The decision tree shows that at its left branch if the auction ending time is Weekday Afternoon period and the initial bid price or open price is less than 14.995 $, then the seller gets the price premium.

If the open price is equal or greater than 14.955 $, but the auction duration is less than 6 days then the seller gets the price premium, but if the auction duration is more than 6 days and the number of bids is more than 15 bids then the seller gets the price premium otherwise he gets the price premium.

At the right branch, where the auction ending time is Weekday Morning, Weekend Morning, or Weekend Afternoon, if the initial bid price (open price) is equal or greater than 227.5 $, then the seller gets the price premium.

At the same right branch, if the initial bid price is less than 227.5 $, a left child branch is created, and a condition checks the number of bids variable.

If the previous condition says that the number of bids is equal or greater than 32 bids, but auction duration is less than 4 days then the seller gets the price premium, otherwise if the auction duration is equal or greater than 4 days, the CART algorithm and its decision tree checks a new condition which is the initial bid price or open price.

In case of the initial bid price is less than 0.505 $ then the seller gets the price premium, but if the initial bid price is more than 0.505 $ and also more than 9.475 $ then the seller gets the price premium otherwise, he gets the price premium.

At the main right branch where the initial bid price is less than 227.5 $ but the number of bids is less than 32 bids then a new left child is created as shown at figure 2, the CART algorithm says that if the auction ending time is Weekday Morning then the seller will not get the price premium, but if the auction ending time is Weekend Morning, or Weekend After but the auction duration is less than 4 days, then seller also will not the price premium.

If the auction duration at the previous condition is more than 4 days and the number of bids is less than 7 bids, seller also will not get the price premium, but if the number of bids is equal or greater than 7 bids and the initial bid price is equal or more 112.5 $, seller will get the price premium, otherwise if the open price is less than 112.5 $, seller will not get the price premium.

6. CONCLUSION AND FUTURE WORK

In this research, eBay auctions are studied, to propose an algorithm to construct a decision tree to help the seller to predict if he can get the price premium, and what are the conditions that can guarantee the price premium, the researchers used real datasets collected from eBay website.

CART algorithm used a greedy algorithm which is a local optimum technique which assumes that the optimum local leads to global one [15], using another technique in the more complicated dataset increases the accuracy like multivariate classifier [16].

Another limitation of this research, and opens a new field of work, that the study only considered one product, introducing

more types of products is very usefully and testing the performance of CART algorithm and the proposed algorithm to new products is open for future work.

7. ACKNOWLEDGEMENTS

hanks to Dr Mahmoud Kamel who have contributed towards the development of the template.

8. REFERENCES

[1] Ueffing, N. 2018. Automatic Post-Editing and Machine Translation Quality Estimation at eBay. In Proceedings of the AMTA 2018 Workshop on Translation Quality Estimation and Automatic Post-Editing (pp. 1-34).

[2] Ba, S., & Pavlou, P. A. T2002. Evidence of the effect of trust building technology in electronic markets: Price premiums and buyer behavior. MIS quarterly, 247-248.

[3] Breiman, L., Friedman, J., Olshen, R., & Stone, C. 1984. Classification and regression trees. Belmont, CA: Wadsworth International Group.

[4] Cason, T. N., & Friedman, D. 2018. An empirical analysis of price formation in double auction markets. In The double auction market (pp. 253-284). Routledge.

[5] Liu, W. W., Liu, A., & Chan, G. H. 2018. Modeling eBay Price Using Stochastic Differential Equations. Journal of Forecasting.

[6] Hsu, S. B. 2013. Ordinary differential equations with applications (Vol. 21). World Scientific Publishing Company.

[7] Ghani, R. 2005. Price prediction and insurance for online auctions. In Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining (pp. 411-418). ACM.

[8] Dass, M., Jank, W., &Shmueli, G. 2010. Dynamic price forecasting in simultaneous online art auctions. In Marketing Intelligent Systems using Soft Computing (pp. 417-445). Springer, Berlin, Heidelberg.

[9] Mullender Breiman L., Friedman J.H., Olshen R.A. and Stone C. J. 1984. Classification and Regression Trees (2nd Ed.). Pacific Grove, CA; Wadsworth.

[10] "E-Bay auctions datasets", Internet: https://www.kaggle.com/onlineauctions/online-auctions-dataset/. Retrieved Feb 7, 2019.

[11] Kohavi, R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In Ijcai (Vol. 14, No. 2, pp. 1137-1145).

[12] Baker, J., & Song, J. 2008. Exploring decision rules for sellers in business-to-consumer (b2c) internet auctions. International Journal of E-Business Research (IJEBR), 4(1), 1-21.

[13]] Dholakia, U. M. 2005. The usefulness of bidders' reputation ratings to sellers in online auctions. Journal of Interactive Marketing, 19(1), 31-40.

[14] Rodriguez, J. D., Perez, A., & Lozano, J. A. 2010. Sensitivity analysis of k-fold cross validation in prediction error estimation. IEEE transactions on pattern analysis and machine intelligence, 32(3), 569-575.

[15]] Murthy, S. K., & Salzberg, S. 1995. Decision Tree Induction: How Effective Is the Greedy Heuristic?. In KDD (pp. 222-227).

[16] Brodley, C. E., & Utgoff, P. E. 1995. Multivariate decision trees. Machine learning, 19(1), 45-77.

[17] Ince, H., &Trafalis, T. B. 2007. Kernel principal component analysis and support vector machines for stock price prediction. IIE Transactions, 39(6), 629-637.