A Study on Different Methods and Algorithms to Predict End Prices in Online Auctions

[1.]Ms. Selva Mary. G, [2.]Mr. Likhesh N. Kolhe, [3.]Ms. Rucha D. Pathari
[1,2, 3]Asst. Professor, Alamuri Ratnamala Institute of Engineering & Technology, Mumbai, India

***Abstract*— Online auctions have become one of the fastest growing modes of online commerce transactions. eBay has 94 million active members buying and selling goods at a staggering rate. These auctions are also producing large amounts of data that can be utilized to provide services to the buyers and sellers, market research, and product development.**

We collect historical auction data from eBay and use machine learning algorithms to predict end-prices of auction items. We describe the features used, and several formulations of the price prediction problem. Understanding and being able to better predict the outcome of these auctions is important to buyers and sellers alike for individual profit maximization.

In this paper, we explore different methods like the use of multinomial logistic regression, Naive Bayes (NB), and uniform prior Naive Bayes (UPNB) algorithms to predict both whether or not an item will sell and how much it will sell for. In predicting whether an item will sell or not, we find that the NB classifier performs with greater than 75% accuracy over a test data set. For final price prediction, we find that for multiclass binary prediction decision tree models, UPNB is able to outperform both the general NB and softmax algorithms. Unlike algorithms from previous works, our UPNB classifier uses the item title as the sole contributor to the feature vector. We provide a discussion on the results, as well as some insight about our particular data set and avenues for future exploration

Keywords*— Autoregressive model; Dynamics; Electronic commerce, Online Auctions, methods.

I. INTRODUCTION

Online auctions have become increasingly popular in recent years, and as a consequence there is a growing body of empirical research on this topic. Online auctions are generating a new class of fine-grained data about online transactions. This data lends itself to a variety of applications and services that can be provided to both buyers and sellers in online marketplaces. We collect data from online auctions and use several classification algorithms to predict the probable-end prices of online auction items.

One of the main drivers of this interest is eBay (*www.eBay.com*). On any given day, there are several million items, dispersed across thousands of categories, for sale on eBay. Most of that research treats data from online auctions as cross sectional, and consequently ignores the changing dynamics that occur during an auction. In this article we take a different look at online auctions and propose to study an auction's price evolution and associated price dynamics. Specifically, we develop a dynamic forecasting system to predict the price of an ongoing auction. By dynamic, we mean that the model can predict the price of an auction "in progress" and can update its prediction based on newly arriving information. Forecasting price in online auctions is challenging because traditional forecasting methods cannot adequately account for two features of online auction data:

(1)the unequal spacing of bids

(2)the changing dynamics of price and bidding throughout the auction.

Our dynamic forecasting model accounts for these special features by using modern functional data analysis techniques. Specifically, we estimate an auction's price velocity and acceleration and use these dynamics, together with other auction-related information, to develop a dynamic functional forecasting model. We also use the functional context to systematically describe the empirical regularities of auction dynamics.

Forecasting price in online auctions can have benefits to different auction parties. For instance, price forecasts can be used to dynamically score auctions for the same (or similar) item by their predicted price. On any given day, there are several hundred, or even thousands of open auctions available, especially for very popular items such as Apple iPods or Microsoft Xboxes. Dynamic price scoring can lead to a ranking of auctions with the lowest expected price. Such a ranking could help bidders focus their time and energy on only a few select auctions, that is, those that promise the lowest price. Auction forecasting can also be beneficial to the seller or the auction house.

II. ALGORITHMS AND METHODS

A. Functional Regression and Auction Dynamics

To understand the motivation for our forecasting model, it is useful to first take a closer look at eBay auction data. We have pointed out earlier that the data are characterized by rapidly changing price dynamics. We illustrate this phenomenon in this section by investigating the relationship between eBay's auction dynamics and other auction-related information. This will also lay the ground for the forecasting model that we describe in the next section.

We investigate the empirical regularities in eBay's auction dynamics using *functional regression analysis*. Functional regression analysis is similar to classical regression in that it relates a response variable to a set of predictors. However, in contrast to classical regression where the response and the predictors are vector valued, functional regression operates on *functional objects*, which can be a set of curves, shapes, or objects.

Dynamic Auction Forecasting Via Functional Data Analysis

We now describe our dynamic forecasting model. We have shown in the previous section how unequally spaced data can be overcome by moving into the functional context and also

that online auctions are characterized by changing price dynamics. Our forecasting model consists of four basic components that capture price dynamics, price lags, and information related to sellers, bidders, and auction design. First, we describe the general forecasting model, which is based on the availability of price dynamics. Then, we describe how to obtain forecasts for the price dynamics themselves.

- *The General Forecasting Model*

Our model combines all information that is relevant to price. We group this information into four major components:
(1) static predictor variables,
(2) time-varying predictor variables,
(3) price dynamics
(4) price lags.
Static predictor variables are related to information that does not change over the course of the auction. This includes the opening bid, the presence of a secret reserve price, the seller rating, and item characteristics. Note that these variables are known at the start of the auction and remain unchanged over the duration of the auction. Time-varying predictor variables are different in nature. In contrast to static predictors, time varying predictors *do* change during the auction. Examples of time-varying predictors are the number of bids at time *t* or the number of bidders and their average bidder rating at time *t*. Price dynamics can be measured by the price velocity, the price acceleration, or both. Finally, price lags also carry important information about the price development. Price lags can reach back to price at times $t-1$, $t-2$, and so on. This corresponds to lags of order 1, 2, and so forth. We obtain the following dynamic forecasting model. Let $y(t|t-1)$ denote the price at time *t*, given all information observed until $t-1$. For ease of notation, we write $y(t) \equiv y(t|t-1)$.

The model has two practical challenges: (1) price dynamics appear as coincident indicators and must, therefore, be forecasted *before* forecasting $\tilde{}y(T+h|T)$; and (2) the static predictor variables among the *xi*'s do not change their value over the course of the auction and must, therefore, be adapted to represent time-varying information.

- *Forecasting Price Dynamics*

The price dynamics *D(j)y(t)* enter (6) as coincident indicators. This means that the forecasting model for price at time *t* uses the dynamics from the same time period! However, because we assume that the observed information extends only until $t-1$, we must obtain forecasts of the price dynamics before forecasting price. This process is described next.

We model *D(j)y(t)* as a polynomial in *t* with autoregressive (AR) residuals. We also allow for covariates *xi*. The rationale for these covariates is that dynamics are strongly influenced by certain auction-related variables such as the opening bid.

- *Integrating Static Auction Information*

The second structural challenge that we face is related to the incorporation of static predictors into the forecasting model. Take, for instance, the opening bid. The opening bid is static in the sense that its value is the same throughout the auction, that is, $x(t) \equiv x, \forall t$. Ignoring all other variables, we can rewrite model (5) as $y(t) = \alpha + \beta x$. (12)

Because the right-hand side of (12) does not depend on *t*, the least squares estimates of α and β are confounded! The problem outlined previously is relatively uncommon in traditional time series analysis because it is usually only meaningful to include a predictor variable in an econometric model if the predictor variable itself carries time-varying information.

As pointed out earlier, our dynamic forecasting model consists of two basic parts: one part forecasts the price dynamics, and the other part uses these forecasted dynamics as input into

the price forecaster.

B. *A Novel Method for Predicting the End-Price of eBay Auctions*

- *Naive Bayes*

In order to predict whether or not a given item would sell, we implemented a multinomial Naive Bayes classifier with the item title as the sole contributor to the feature vector. A complete word list V = f1; ::::; jV jg was created from all item titles in the complete set. For each new item on which to make a prediction, our feature vector x(i) is of length jV j, where the vth entry represents the number of times that the vth word appears in the title. The maximum likelihood estimates, _jjy=0;1, for each word are calculated from the training set data: Before performing this analysis, words that appeared only once amongst all titles were removed from the word list. Furthermore, words shorter than three characters, and the following common words, were removed from the list: and, for, from, that, there, these, this, and those. This analysis provided an error _S = 0:25 when predicting on all genres.

- *Multinomial Logistic Regression*

The first multi-class algorithm we attempted was a softmax regression. Our main assumption is that there exist k multivariate Gaussian distributions for each xjy, with equivalent covariance matrices after zeroing out the mean. The feature vector for each item included the starting price, starting time, condition, seller feedback score, whether returns were accepted, shipping price, duration of the auction, genre (e.g. rock, country, etc.), and a constructed "page views" parameter. The page views parameter roughly corresponds to the average page views per hour during the auction listing. Our reasoning was that this gives some insight to the popularity of the item and may help in estimating its end price.

This method worked decently, predicting with _s = 0:56 for the items in our test set with b = 5 and a maximum price range of $50. After implementing a feature selection process, a surprising result we found was that the starting price of the item dominated the regression more than we initially predicted. Excluding the starting price did not improve our results. Seeing this, we decided to look for new methods to make our predictions.

- *Multi-class Naive Bayes*

In an effort to extend the title-based Naive Bayes algorithm, we developed a multi-class Naive Bayes scheme. The final prices are discretized into bins of size b dollars, and a series of title-based Naive Bayes classifications are made. The first classifier asks, does the item sell for more or less than $0? The second asks, does the item sell for more or less than $(b)? The third, does the item sell for more or less than$(2b)?, and so on. Each decision is made using a standard Naive Bayes classifier and requires a retraining of the data set. The first classification is identically the sold/unsold problem and the remaining classifications determine which final price bin, ^y, that the algorithm should predict.The practical implementation is illustrated in figure (2). For the first price at which an item is deemed not to sell, the algorithm predicts the preceding bucket as the final sale price. In the case where ^yi < binOf(START_PRICEi), ^y is reassigned to ^y _ binOf(START_PRICEi). Those that were deemed to sell at greater than all bin values are assigned the maximum bin value.

- *Multi-class, Uniform Prior Naive Bayes*

Our final algorithm is derived from an analysis of the end price distributions. We were finding that weighting the probabilities, p(xjy), by their class priors, p(y = j), tended to skew the multi-class classifier predictions. We propose something new: given a multi-class binary decision tree process, we assume a uniform prior over all classes. This effectively states that the probability of a given item for selling above a price is equal to the probability of it selling below that price. This algorithm predicts better than softmax and NB in general, reducing our classification error to _S = 0:28 with $5 bins and _S = 0:3 with $2 bins. This result is presented in comparison with the pure Naive Bayes algorithm performance. We hypothesize that this improvement could be attributed to our decision making process, and we are still exploring the full repercussions of assuming a uniform prior.

C. Price Prediction As A Machine Learning Problem

Given the features described in the previous section, the task now is to predict the end-price of a new auction. There are several ways in which this problem can be tackled with machine learning algorithms. We defined the problem in three ways to compare the relative merits of each approach and judge their effectiveness in isolation as well as for offering Price insurance:

- *Regression:*

We treat the price prediction task as a regression task and use the training data to learn regression coefficients. The output of the model, when applied to new data is a specific (continuous) price. For the results reported in the following section, we used linear regression, polynomial regression with degrees 2 and 3, and CART (Classification & Regression Trees)

- *Multi-Class Classification***:**

We discretize the end-price (target variable) into $51 intervals and create discrete categories. Each instance now falls in one of these categories. The price prediction problem can then be treated as a multiclass classification problem with the output being a $5 range instead of the specific price (as in the case of regression). We use decision trees (C5.0) and neural networks to implement the multiclass classification in our experiments.

- *Multiple Binary Classification tasks:*

We create multiple binary classifiers, with each classifier learning a binary classification task: whether the end-price of the auction will be more than $X or not. For the experiments in this paper, we varied X in $5 intervals to be comparable to the multiclass approach. For example, one classifier for classifying whether price is more than $5, the next for $10, and so on, going up to the maximum price in the training set. This technique was motivated by the small amounts of training examples that are available for any item in online auctions. Although there are a large number of auctions going on, auctions for any single kind of item are limited in number. This creates the need to use the scarce training data in an efficient manner. The multiclass classification scheme (described earlier) is not very effective since the positive examples for each category ($5 interval) are limited to the ones in that category. In contrast, for the binary classification case with multiple classifiers, the positive examples for the classifier that is predicting whether the price is going to be greater than $45, consist of all the examples where the price is greater than 45 (and not just in the range $45-$50).

III. APPLICATIONS

The ability to predict the ending price of online auction items lends itself to a variety of applications. We described Auction Price Insurance as one of those applications in the previous section. Here, we briefly describe some other possible applications of the same technique.

Listing Optimizer: The model of the end-price based on the input attributes of the auction can also be used to help sellers optimize the selling price of their items. When the seller enters their personal information and the item they want to sell in an auction, a service could give suggestions for the auction attributes (such as starting time, starting bid, use of photos, reserve price, words to describe the item, etc.) that would maximize the end price.

Auction Arbitrage: Instead of offering Price Insurance, Price Prediction can be used to do arbitrage in online auctions. If an item is being sold for much lower than the expected selling price due to some parameter of the auction that can be changed by the seller, the item can be acquired at the low price and then re-listed using optimal or better auction parameters (better ending time, duration, description, highlighting, etc.) to sell for a higher price. There are several other applications that can be enabled by the price prediction techniques described in this paper. While we have not provided an exhaustive list of applications, we believe that having access to the likely end-price of auction items opens up a large variety of services that can be offered to both buyers and sellers in online auctions.

CONCLUSION

There are several other applications that can be enabled by the price prediction techniques described. While we have not provided an exhaustive list of applications, we believe that having access to the likely end-price of auction items opens up a large variety of services that can be offered to both buyers and sellers in online auctions. In this paper Forecasting price in online auctions can have benefits to different auction parties. For instance, price forecasts can be used to dynamically score auctions for the same (or similar) item by their predicted price. On any given day, there are several hundred or even thousands of open auctions available, especially for very popular items such as Apple iPods or Microsoft Xboxes. Dynamic price scoring can lead to a ranking of auctions with the lowest expected price. Such a ranking could help bidders focus their time and energy on only a few select auctions, that is, those that promise the lowest price. Auction forecasting can also be beneficial to the seller or the auction house. The different methods are being studied with their advantages and process

REFERENCES

[1]. [W. Vickrey, “Counterspeculation, auctions, and competitive sealed tenders”, The Journal of Finance, 16(1), 8-37, 1961.

[2]. R. Ghani and H. Simmons, “Predicting the end-price of online auctions”, Decision Support Systems 44 (2008) 970-982, 2004.

[3]. Auction Software Review. http://www.auctionsoftwarereview.com/article-ebaystatistics.

[4]. Bajari, P. and A. Hortacsu, “Winner’s Curse, Reserve Prices, and Endogenous Entry:

[5]. Empirical Insights from Ebay Auctions,” (2002), The Rand Journal of Economics

[6]. Blum, A., & Mitchell, T. (1998). Combining labeled and unlabeled data with co-training.

[7]. Proceedings of the 11th Annual Conference on Computational Learning Theory (pp. 92-100).

[8]. Bryan, D., Lucking-Reily, D., Prasad, N., Reeves, D. Pennies from eBay: the Determinants of Price in Online Auctions., January 2000

[9]. Heijst, D. V.; Potharst, R.; and Wezel, M. V. 2008. A Support System for Predicting Ebay End Prices. Decision Support Systems 44(4):970–982.

[10]. Lucking-Reiley, D.; Bryan, D.; Prasad, N.; and Reeves, D. 2007. Pennies from eBay: The Determinants of Price in Online Auctions. The Journal of Industrial Economics 55(2):223–233.

[11]. Wellman, M.; Reeves, D.; Lochner, K.; and Vorobeychik, Y. 2004. Price Prediction in a Trading Agent Competition. Journal of Artificial Intelligence Research 21:19–36.

[12]. Chawla, N., Japkowicz, N. and Kolcz, A. (editors), SIGKDD Explorations, Special Issue on Class Imbalances , SIGKDD Explorations 6(1), June 2004.

[13]. Diettrich, T. & Bakiri, G. (1995). Solving Multiclass Learning Problems via Error-Correcting Output Codes. Journal of Artificial Intelligence Research,2;263--286, 199

[14]. Oren Etzioni, Rattapoom Tuchinda, Craig A. Knoblock, Alexander Yates: To buy or not to buy: mining airfare data to minimize ticket purchase price. KDD 2003: 119-128