ABSTRACT

Title of dissertation:	Mining of Business Data
	Shu Zhang, Doctor of Philosophy, 2009
Dissertation directed by:	Associate Professor Wolfgang Jank AMSC and R.H.Smith School of Business

Applying statistical tools to help understand business processes and make informed business decisions has attracted enormous amount of research interests in recent years. In this dissertation, we develop and apply data mining techniques to two sources of data, online bidding data for eBay and offline sales transaction data from a grocery product distributor. We mine online auction data to develop forecasting models and bidding strategies and mine offline sales transaction data to investigate sales people's price formation process.

We start with discussing bidders' bidding strategies in online auctions. Conventional bidding strategies do not help bidders select an auction to bid on. We propose a automated and data-driven strategy which consists of a dynamic forecasting model for auction closing price and a bidding framework built around this model to determine the best auction to bid on and the best bid amount.

One important component of our bidding strategy is a good forecasting model. We investigate forecasting alternatives in three ways. Firstly, we develop model selection strategies for online auctions (Chapter 3). Secondly, we propose a novel functional K-nearest neighbor (KNN) forecaster for real time forecasting of online auctions (Chapter 4). The forecaster uses information from other auctions and weighs their contribution by their relevance in terms of auction features. It improves the predictive performance compared to several competing models across various levels of data heterogeneity. Thirdly, we develop a Beta model (Chapter 5) for capturing auction price paths and find this model has advantageous forecasting capability.

Apart from online transactions, we also employ data mining techniques to understand offline transactions where sales representatives (salesreps) serve as media to interact with customers and quote prices. We investigate the *mental* models for salesreps' decision making, and find that price recommendation makes salesreps concentrate on cost related information.

In summary, the dissertation develops various data mining techniques for business data. Our study is of great importance for understanding auction price formation processes, forecasting auction outcomes, optimizing bidding strategies, and identifying key factors in sales people's decision making. Those techniques not only advance our understanding of business processes, but also help design business infrastructure.

Mining of Business Data

by

Shu Zhang

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2009

Advisory Committee:

Professor Wolfgang Jank, Chair/Advisor
Professor Wedad J. Elmaghraby
Professor Abram Kagan
Professor Itir Z. Karaesmen Aydin
Professor Mahesh Kumar
Professor Paul J. Smith

To Lijun, My Parents and My brother

Acknowledgments

I owe my gratitude to all people who have brought laughters and joy into my graduate life. Without you being on my side, this thesis would never be possible, and I would have never had so many cherishable moments in the past four years that I will remember forever.

First and foremost, I would like to thank my advisor, Dr. Wolfgang Jank, for his advice, support, and encouragement on research. I started as a rookie in the field of eCommerce research, and his mentoring has leaded me all the way to get here. It has been my honor and a great pleasure to work with and learn from such an extraordinary individual.

I'm grateful to Dr. Paul Smith, for his support and help in all aspects in my graduate study. He has brought up a lot of brilliant ideas that help improving this work significantly. Not only helping me in research with his incredible knowledge in statistics, he has also always been there to guide me through the depressed days when I got stuck in research.

The chapters about pricing strategies would have not been part of this dissertation without the help from Dr. Karaesmen and Dr. Elmaghraby who have always opened the door to me when I had questions. I have benefited a lot from their great passion towards research and extraordinary understanding and creative thoughts about pricing.

I would like to thank Dr.Kagan and Dr.Kumar for their kind support and guidance, and reading of my dissertation. I have learnt a lot from them. It is a great pleasure to have them in my committee.

Thanks are also due to Dr.Shmueli, who has given me tremendously amount of help and encouragement. It has been a wonderful experience to have such a talented and exceptional mentor in my graduate life.

I owe my deepest thanks to my family - my brother and parents who have always been on my side and support me from faraway China; and my beloved husband, Lijun, who takes great care of me every day. Their love has been my unlimited source of energy and confidence.

It is impossible to list everyone who has helped me in the few pages. I just want to thank you all. I am so lucky to have you in my life!

Table of Contents

List of Tables

List of Figures

List of Abbreviations

MAPE	Mean Absolute Percentage Error
WTP	Willingness To Pay
FDA	Functional Data Analysis
KNN	K Nearest Neighbors
fKNN	Functional K Nearest Neighbors
KL Distance	Kullback-Leibler Distance
B2C	Business to Customer
B2B	Business to Business
H2H	Human to Human
DST	Decision Support Tool
Salesrep	Sales Representative
Triplet	Salesrep-Customer-Product Triplet

Chapter 1

Introduction

Due to the availability of rich, high-quality data, employing data mining tools to solve business problem and related research have gained great popularity in recent years. Business data typically comes from two sources - online transactions, such as the bidding history for eBay online auction, and offline transactions, such as sales transaction data from a grocery product distributor. In this dissertation, we first develop several forecasting models and bidding strategies for online auction data. Then we investigate sales representatives' price formation process, in particular, the impact of price recommendation on such process, for sales transaction data.

1.1 Introduction to Online Auctions

1.1.1 Online Auctions

Online auction is a significant marketplace, which allows consumers and businesses to sell, buy, and bid on a variety of goods. People shop for consumer electronics on uBid (ubid.com), for consumer loans on Prosper.Com, and for almost everything on eBay (ebay.com). eBay is one of the major online marketplaces and currently the biggest consumer-to-consumer online auction site. Founded in 1995, eBay Inc. has attracted over 200 million registered users and touts net revenue of $8.37 billion for the year 2008 despite the ecomony recession. Dispersed across twenty thousands of categories, several millions of items are listed on any given day. In fact, [6] refers to online auctions as “one of the most successful forms of electronic commerce”.

Typically in an online auction, the opening price is set by the seller or the auction house, and bidders submit bids online. There are various auction formats: Online auctions can be ascending (e.g., in eBay auctions) or descending (e.g., in Dutch flower auctions where the price is bid down); first price or second price (i.e. whether the final price is equal to the highest bid or second highest bid); with fixed or soft closing time (i.e. where the auction duration extends with the arrival of new bids); for single items or bundles. On eBay, most auctions are second-price ascending auctions for single items, with a fixed duration. The seller sets the opening price, and bidders place ascending bids until the auction end time is reached. At that time, the winner is the highest bidder, and s/he pays the second highest bid (plus an increment).

Online auctions differ from their offline counterparts in their longer duration (typically several days), the anonymity of participants, the low barriers of entry, their global reach, and around-the-clock availability. These conditions lead to a highly dynamic environment, where bidders engage in competitive behavior that is motivated by both psychological effects and economic reasoning. Auctions allow bidders to adjust their behavior based on the previous progress of the auction of interest and competing auctions, which in turn contributes to the dynamic changes in auction progression and price.

1.1.2 eBay Data Structure

Empirical research of online auctions has been flourishing in recent years due to the important role that these auctions play in the marketplace and the availability of large amounts of high-quality bid data from eBay (as well as Yahoo!, OnSale, uBid, etc). eBay makes public a vast amount of rich bidding data that include all the bidding information as well as information about the bidders, the seller, and the product being auctioned. A typical example of the bid data for a single auction is shown in Figure 1.1. From the bid data, we can determine the price as shown on eBay at any time during the ongoing auction[1].

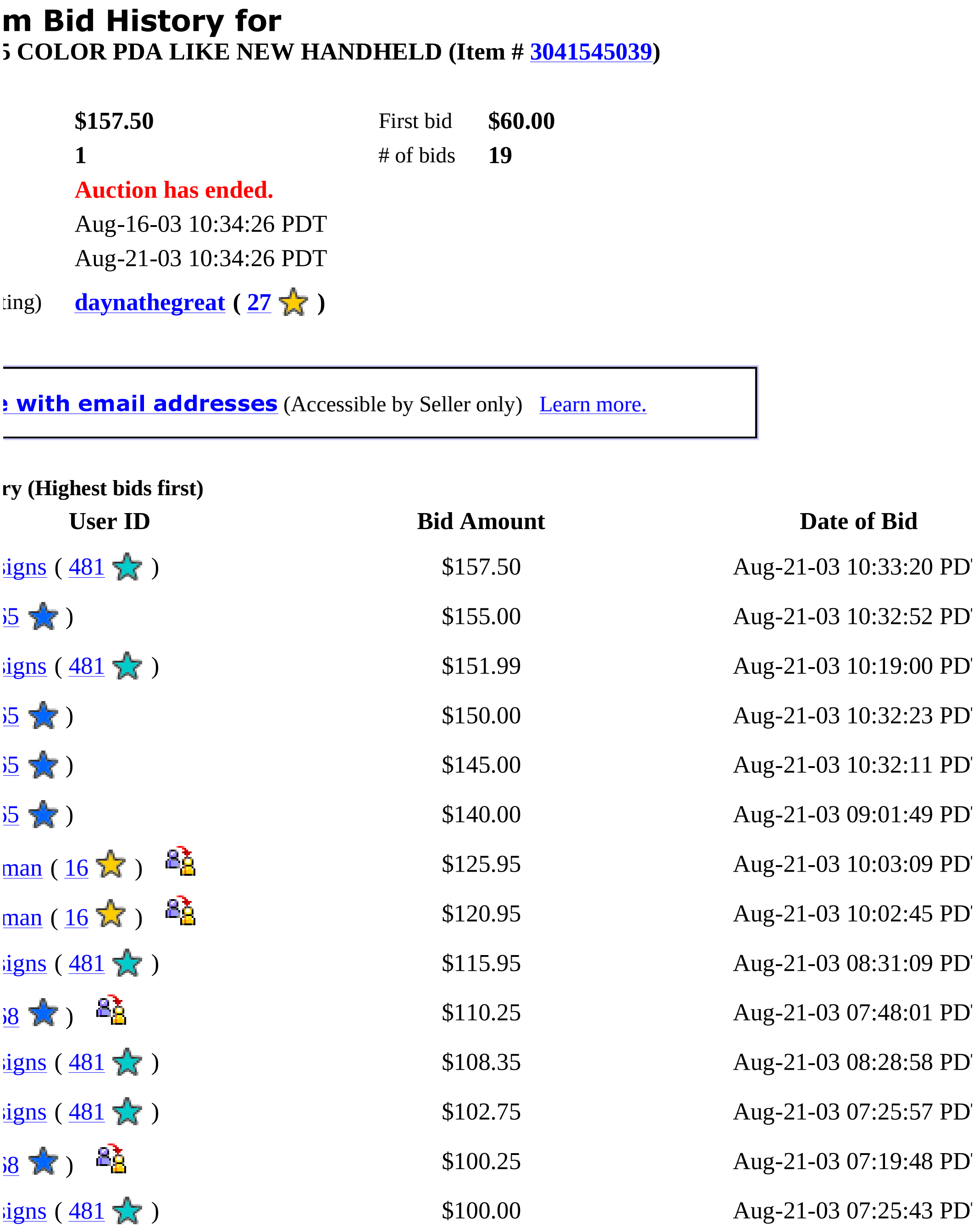
We use two eBay data sets about auction bidding history throughout the research. One includes the complete bidding records for 380 auctions for new Palm Pilot M515 handheld PDA's that took place on eBay between March and May, 2003; the other data set contains information on 4,965 laptop auctions that took place on eBay between May and June, 2004. For details about the two data sets, please see Appendix A.

1.1.3 Online Auction Literature

Statistical and data mining techniques have been extremely instrumental in gaining insights into auction processes, and we describe some of the major contributions to the online auction literature to date.

One important stream of research has focused on various auction features and

[1] On eBay, the price shown at any point in time is the second highest bid at that point rather than the highest bid. Thus, the bid data might include bids that are lower than the highest bid.



m Bid History for
5 COLOR PDA LIKE NEW HANDHELD (Item # 3041545039)

	$157.50	First bid	$60.00
	1	# of bids	19
	Auction has ended.		
	Aug-16-03 10:34:26 PDT		
	Aug-21-03 10:34:26 PDT		
ting)	daynathegreat (27 ★)		

with email addresses (Accessible by Seller only) Learn more.

ry (Highest bids first)

User ID	Bid Amount	Date of Bid
igns (481 ★)	$157.50	Aug-21-03 10:33:20 PD
55 ★)	$155.00	Aug-21-03 10:32:52 PD
igns (481 ★)	$151.99	Aug-21-03 10:19:00 PD
55 ★)	$150.00	Aug-21-03 10:32:23 PD
55 ★)	$145.00	Aug-21-03 10:32:11 PD
55 ★)	$140.00	Aug-21-03 09:01:49 PD
man (16 ★)	$125.95	Aug-21-03 10:03:09 PD
man (16 ★)	$120.95	Aug-21-03 10:02:45 PD
igns (481 ★)	$115.95	Aug-21-03 08:31:09 PD
58 ★)	$110.25	Aug-21-03 07:48:01 PD
igns (481 ★)	$108.35	Aug-21-03 08:28:58 PD
igns (481 ★)	$102.75	Aug-21-03 07:25:57 PD
58 ★)	$100.25	Aug-21-03 07:19:48 PD
igns (481 ★)	$100.00	Aug-21-03 07:25:43 PD

Figure 1.1: Bidding data for an eBay Palm PDA auction.

their impact on the closing prices. In fact, a seller's reputation [6], an auction's duration [37], opening and reserve prices [74], or an item's shipping costs [38], all have been shown to affect the final price (see also [89]). Statistical tools used for this type of research are mainly classical regression models, and the results from such analyses help answer sellers' questions about which auction setting or listing enhancements are worth the extra fee and improve the design of the online market.

Besides auction features, there has been much interest in understanding the dynamics of the price formation process recently, in an attempt to better capture, understand, and forecast price in online auctions. Novel statistical approaches have been developed in gaining deeper understanding about price dynamics. [52; 10] have shown that price dynamics can be very heterogeneous, even for auctions of the same product, using descriptive statistics. [51] have shown this for auction of new Palm PDA handheld devices sold on eBay; [21] found similar behavior in auctions for contemporary Indian art. [51] further segmented auctions based on price dynamics: "steady auctions" are those with constant dynamics, "low-energy auctions" are those with late dynamics, and "bazaar auctions" see mostly early activity. [86] developed a three-stage non-homogeneous Poisson process for capturing bid timing, and showed its flexibility in capturing the bid timing for various items and auction durations, etc. [97] introduced a single class of functional differential equation models that captures a wide range of auction price paths and dynamics. Finally, [96] developed real-time forecasting models for ongoing auctions that use as input the price path and its dynamics until the time of prediction. They show that the inclusion of the dynamic information significantly improves predictive accuracy compared to models that exclude such information.

Researchers also study the interplay between auction features and dynamics. [84] illustrated the effect of auction features (such as the opening bid) on auction dynamics, and found that higher opening bids result in lower price dynamics. [55] developed model-based regression trees to relate differential equation models for different auction dynamics to auction features.

Auction dynamics reflect unobservable dynamic behavior such as competition between bidders within the auction and across auctions. The fact that millions of auctions are taking place simultaneously and many of these auctions sell the same or similar items introduces competition both to the sellers and the bidders of the products, which results in competition thus cross-dependencies among auctions and their outcomes. Consequently, adequately capturing and modeling the price path can be used for studying the effects of competition. [44] developed visualizations for the price formation process and its dynamics to study the effect of concurrency among online auctions. [21] have investigated the relationship between within-auction and between-auction competition on price dynamics and have shown that price dynamics are good proxies for the harder-to-measure competition.

Although the presence and importance of competition are broadly noticed by many scholars, quantified study of its effect is rather limited. This is mainly due to lack of measures for competition in the dynamic environment. In this dissertation, we set out to quantify the competition between simultaneous auctions and use such information in our forecasting of auction outcomes and designing bidding strategies in this competition environment.

There exist two very well documented (and researched) bidding strategies for online auctions, early bidding and last-minute bidding. [8] have shown that early bidders often discourage potential competitors from entering the same auction by signaling their commitment early in the auction. In contrast, last-minute bidders [83; 86] wait until the very last moment to avoid being out-bid. However, neither strategy takes into account the effect of competition, thus provides no guide for bidders to select the right auction from many simultaneous auctions to bid on. We build a forecasting model which accounts for competition among simultaneous auctions, and develop a bidding strategy around the model that can determine not only the best auction to bid on but also the right bidding amount.

An alternative way to capture competition is to assign heavier weights to auctions with high level of competition when estimating the model and making forecasts. This is in contrast to conventional methods where information from each auction is weighted equally in the process of model estimation. Examples for conventional methods include [96] which used regression-based models to forecast an auction's final price in a dynamic fashion (see also [32; 59]) and [16] which employed a classification and regression tree method for forecasting. In this dissertation, we develop a method for computing weights for each auction based on auction similarity (therefore competition level) and making weighted forecasts.

1.2 Introduction to Pricing in B2B and B2C and Literature Review

Business transactions are normally divided into four groups: business-to-consumer (B2C), business-to-business (B2B), business-to-public administration, and consumer-to-public administration. We only discuss the first two types of transactions here because the latter two depend heavily on government policies which is basically a different research area.

A business can ensure profitability and longevity by utilizing appropriate pricing strategy. [69] stated that improved pricing can yield 20%-35% reduction in waste or unused inventory, 2%-4% increase in corporate revenues, and 1%-3% increase in profit. However, with the increasing production size and customer population, setting the right price has become an non-trivial task.

Pricing in B2C settings usually involves setting prices for hundreds even thousands of products/services over hundreds of stores nationally and/or internationally (e.g. retail stores, airlines, and hotels). Such complicated task is typically done by decision support tool (DST). DST collects vast amounts of data and employ data mining and optimization routines to uncover the holy grail of pricing - customer's willingness-to-pay (WTP) - based on which optimized price is computed. DST has proven itself to be extremely helpful in enduring profitability in B2C business. For instance, by the help of DSTs, Marriott's annual profit increase for individual hotels totaled $86 million after the rollout of their in-house developed pricing and revenue management system in 2004 [76].

Customer WTP is often endogenously determined by many factors, some observable and some not. The observable traits, such as customer's purchase history or price of the same products from competing companies [67], can be captured and incorporated into DST. The non-observable part, however, cannot be quantified. Such non-observable factors speak to how a customer perceives/ internalizes a price quote and reacts to it. For instance, [65] introduce the concepts of fairness, and they find it is unfair to exploit shifts in demand by raising prices. [64] discuss the notion of anchoring and how customers make adjustments under uncertain market conditions. [94] study the framing of the price quote, and find that customers respond differently to price quotes framed different by salesreps. While both observable and unobservable factors may exist and hence be useful in determining customer WTP in B2C markets, the relatively small dollar spend of each customer coupled with the large number of customers present in the market generally imply that DSTs can make reasonable pricing decisions while ignoring the unobservable traits.

The situation changes as we turn to B2B markets where the characteristics of each customers matter to pricing. In such settings, sale representatives (hereafter referred to as "salesreps") are entrusted with determining the impact of the unobservable customer traits on each customer's WTP, and managing the (relatively) large accounts of and relations with several business customers. For example, salesreps must assess if a customer will find a price to be fair (whether or not it is a price that is justified by current market conditions), how and on what the customer anchors his willingness to pay (e.g., the past price paid or possibly a competitor's current price), the strength of the relationship between salesrep and customer and hence whether a customer will trust a quoted price as being reasonable, how customer reacts to price increases, etc. To emphasize the human involvement, we hereafter refer to B2B settings by H2H (Human-to-Human).

With such intangible pieces of information about customers, salesreps are often considered experts for quoting appropriate prices. However, studies show that being an "expert" does not always imply better decisions [93]. No matter how experienced salesreps are, they are all human beings who are subject to their own decision biases and judgment heuristics (e.g., memory bias, [93], satisficing behavior, [77], status quo bias, [66]), which leaves space to improve pricing. For instance, salesresps decision is generally affected by irrelevant information [40], thus providing them with only most relevant info may lead to appropriate price quotes.

DST, on the other hand, can gather information across hundreds of salesreps, and is able to make better aggregate predictions about WTP and demand. Hence, DST price recommendation may provide a valuable reference point on which salesreps can anchor their price decisions. However, it is not very clear whether price recommendations to salesreps in H2H markets as they have in B2C business.

While there is a large amount of literature on pricing in economics, marketing, or operations management for B2C markets (e.g. [94; 20; 101]), surprisingly little research has been done on H2H pricing, and even less so on behavior of salesreps in this context. With limited understanding of what salesreps anchor on when making price quotes, it is difficult to improve pricing in B2B setting. We set out to study how salesreps form prices and respond to price recommendations in H2H markets in this lack of study. The results will aid designing of DSTs to counter salesrep biases and improving profitability for companies.

1.3 Contributions of the Dissertation

Applying statistical tools to help make informed business decisions has attracted enormous amounts of research interest in recent years. Because of the huge amounts of information available, distinguishing useful from noisy information and drawing informed conclusions from data becomes a non-trivial task and requires employment of novel statistical tools. In this dissertation, we develop/apply data mining techniques to two sources of business data - online auction data and H2H transaction data. We develop prediction models and bidding strategies in online auction setting and investigate the impact of DST price recommendation on sales representatives' pricing decision. This dissertation has resulted in several papers under review at Statistics and Business journals; and another paper is coming out at the end of summer.

1.3.1 Data Driven Bidding Strategy

Bidders participating in online auction often face many complicated bidding decisions. They have to decide whether to bid early or late, whether to place a single bid or multiple updates, whether to bid high or low. Bidding is further complicated by the existence of many auctions that offer the same, or similar item simultaneously. All in all, a complete bidding strategy has to include decisions on which auction to bid on and how much.

Many bidders rely on two conventional strategies, *early bidding* and *last-minute bidding*. Although proven to effectively yield a high winning probability for a careful selected auction, neither strategy answers bidders' question about which auction to bid on given thousands of simultaneous auction.

The first contribution of this dissertation is to propose a novel automated and data-driven bidding strategy which provides bidders with complete decision guides. Our strategy consists of two main components. First, we develop a dynamic, forward-looking forecasting model for price in competing auctions. Then, using the idea of maximizing consumer surplus, we build a bidding framework around this model that determines the best auction to bid on and the best bid-amount. We also conduct a simulation study which shows that our strategy results in a much higher surplus than two conventional bidding rules. This research, discussed in Chapter 2, is currently under the second round review at the INFORMS Journal of Computing [57].

1.3.2 Model Selection for Improved Forecasting

One important component of our bidding strategy is a good forecasting model for auction closing prices. Knowing the auction's closing price has several advantages for auction participants. Bidders can use this information to make more informed bidding decisions [57]. Sellers can use predictions to identify times when the market is more favorable to sell their products and to better evaluate the value of their inventory.

In this chapter, we investigate forecasting alternatives by developing model-selection strategies for online auctions. Model selection in this setting is different compared to classical time series analysis. In classical forecasting, one typically wants to forecast a particular time point; while in the context of online auction, one needs to forecast an entire time interval to satisfy bidders' need of bidding on any auction that is expected to close in that time window.

Our second contribution of this dissertation is to extend the classical model selection criteria which are applicable only to a time point to the setting where forecasting a time interval is required. we do so by computing an entire *distribution* of a model selection criterion over the prediction interval. In this Chapter, we investigate different ways to summarize the distribution and the impact of different summaries on the prediction task. We find that the models selected by the volatility of classical AIC or BIC's distribution over the prediction window have extremely poor prediction performance, while the models selected by minimum or maximum predict very well. This research is discussed in Chapter 3 and has been submitted to the Journal of Business and Economic Statistics for review [58].

1.3.3 Weighted Forecasting of Closing Prices

Besides studying model selection criteria for regression models, we also investigate forecasting alternatives by developing novel weighted forecasting methods. For the closing price of an ongoing auction, the natural reference points are the final prices of past auctions. Previous forecasting methods, including ones developed in Chapter 2 and 3, put equal weight on the information from all training auctions when estimating model coefficients and making forecasts. Nevertheless, assuming that more similar auctions contain more relevant information for forecasting, a forecasting method that weighing the information from each auction differently, depending on how similar that auction is to the auction of interest, is more appropriate. For this purpose, we apply the popular weighted prediction method - K-Nearest Neighbors (KNN) - for forecasting closing prices of an ongoing auction.

One key aspect to the success of KNN is the choice of distance metric based on which the distances between samples (i.e. the reciprocal of sample weights) are measured. This is especially challenging in the context of online auctions because auctions vary on many conceptually different dimensions, such as static (e.g. auction starting prices), time-varying (e.g. number of bids) and functional dynamics information (the dynamics/shapes of the auction price paths). Although there exist standard measures for static or time-varying information, measuring the distance between functional dynamic information (e.g., between two curves) is more involved because of infinite dimensionality.

An important contribution of this research is to point out a new research area - developing weighted forecasting models for better forecasts. In the study, we introduce a parametric Beta model to capture auction price paths, which allows measuring the distance between auctions' dynamics in a very parsimonious way via the Kullback-Leibler distance (KL distance). Furthermore, we define distance metrics that integrates information of various types, including dynamics. Using the reciprocal of the distance as weights, we find that weighing information unequally yields better forecasts compared to classical methods such as regression models or trees and this result holds in auctions of varying levels of heterogeneity. This research, discussed in Chapter 4, has been recommended for publishing in International Journal of Forecasting [102].

1.3.4 A Flexible Model for Price Dynamics in Online Auctions

Besides allowing measuring distance between auctions' dynamics via KL distance, the Beta model developed in Chapter 4 has many other useful properties in online auction context. We explore those properties in details and compare it with existing models for capturing auction price paths.

The fourth contribution of this dissertation is to study the characteristics of the parsimonious parametric Beta model and show its advantages as a representation for auction price paths over existing methods. We show that the model can accurately capture price paths and price dynamics of various types, summarize the bid timing distribution, measure pairwise distances between price paths or price dynamics curves, and is computational efficient. This work is discussed in Chapter 5 and currently under review at the Journal of the Royal Statistical Society (Series C) [56].

1.3.5 Decision Making in H2H Transactions

Different from B2C settings where decision support tools (DST) have been adopted and proven to be extremely valuable in aiding firms and improving their profits, sales representatives have significant responsibility in pricing decisions in B2B (H2H) transactions. Salesreps may rely on many observable and non-observable information, such as their personal expertise, knowledge of individual customers, and price recommendations from DST, to make price quotes. Given those many pieces of related information, especially DST price recommendation, it is not very clear which are the important factors that take effect in salesreps' *mental model*, by which we refer to their price formation process.

One important contribution of this dissertation is to identify important factors that determine a salesrep's mental decision model in a H2H setting. We study how sales people adjust price quotes for different products and different customers over time with special attention to the impact of DST price recommendation. We use various model selection criteria to identify most influential factors, and we find that salesreps anchor most on cost related information, including cost, sign and size of cost change, and types of products (perish commodities or non-commodities). Furthermore, we find that price recommendation from DST, whenever provided, influence salesreps' decisions in a positive way. It serves as a price focal point, without which, salesreps are influenced more by unobservable factors and thus make price decisions difficult to explain and predict. This work is anticipated to be submitted to a top business journal (such as Management Science) by the end of summer.

1.3.6 Summary of Dissertation Contributions

To summarize, the contributions of this dissertation are to:

• Propose a novel automated and data-driven bidding strategy which helps bidders make bidding decision (Chapter 2).

- Investigate various model selection criteria for forecasting over a time interval in online auction setting (Chapter 3).

- Propose a K-Nearest Neighbor forecaster for forecasting closing price of online auctions; introduce a parsimonious model to capture auction price paths that allows measuring distances between auctions' dynamics; and propose a novel distance metric for online auctions that takes into account both static and time-varying features as well as the auction's price dynamics information (Chapter 4).

- Study characteristics of the beta model and illustrate its advantages over existing models as representations of auction price paths (Chapter 5).

- Identify the key factors influence saleresps' pricing decisions and investigate the impact of decision support tool on salesreps (Chapter 6).

Chapter 2

An Automated and Data-Driven Bidding Strategy for Online Auctions

2.1 Introduction

The flexibility of time and location as well as the availability of many different products make online auctions an important marketplace. However, bidders participating in this marketplace often face many complicated bidding decisions. They have to decide whether to bid early or late, whether to place a single bid or multiple updates, whether to bid high or low. Bidding is further complicated by the existence of many auctions that offer the same, or similar item simultaneously. In that case, one's bidding strategy has to be expanded to include decisions on which auction to bid on, when to bid on that auction, and how much.

There exist two very well documented bidding strategies, *early bidding* and *last-minute bidding*. By signaling their commitment early, early bidders [8] discourage competitors from entering the same auction. In contrast, last-minute bidders [83; 86] wait until the very last moment as the chances of being out-bid decrease with the time left in the auction. However, both bidding strategies suffer from limitations since neither takes into account the effect of competition [39]. In other words, neither strategy considers the information from simultaneous auctions offering similar products. While there is emerging literature [100] that suggests that bidders should shade their bids in the presence of sequential auctions, the precise amount of the optimal shade on an auction-by-auction basis is not quite clear.

In this chapter, we propose a novel automated and data-driven bidding strategy. Our strategy consists of two main components. First, we develop a dynamic, forward-looking forecasting model for price in competing auctions. Then, using the idea of maximizing consumer surplus, we build a bidding framework around this model that determines several decisions: the best auction to bid on and the best bid-amount. This work is currently under the second round review at *Informs Journal of Computing*.

The first component of our automated bidding strategy is a dynamic forecasting model for the price in competing auctions. There has been considerable amount of work on predicting an auction's closing price using *static* (or pre-auction) information (see e.g. [6; 37; 38; 74]). One drawback to these approaches is that they only consider information available *before* the start of the auction and thus ignore the dynamic nature of the auction process.

Dynamics have only recently been found to affect the outcome of an online auction [10]. [51] find that auctions selling identical products fall into one of three segments of price dynamics, namely "steady auctions" which experience a constant flow of dynamics, "low-energy auctions" with late dynamics and "bazaar auctions" which see the largest jump of dynamics. [84] illustrate the effect of auction parameters (such as the opening bid) on an auction's dynamics and find that higher opening bids result in lower price dynamics. [97] show that an auction's price dynamics can be characterized well using a single class of functional differential equation models and [55] extend upon this idea and develop model based regression trees to relate differential equation models to auction covariates. Moreover, [96] show that the inclusion of price dynamics into forecasting models significantly improves the predictive capability of an auction (see also [7]). In this study, we build upon the ideas developed in previous studies. However, one key difference is that, in contrast to previous studies, we study dynamics in the context of competing auctions. That is, we study the effect of dynamics generated in simultaneous auctions, selling the same or similar product as the auction of interest. We incorporate the dynamic nature of the auction process by employing a modern statistical approach called *functional data analysis* (FDA). See [79] for a general introduction to FDA or [52] for an illustration of FDA in the context of electronic commerce.

Besides incorporating dynamics, our model also explicitly accounts for the information from competing auctions. Competition between auctions has come to the spotlight only recently (e.g. [39; 37; 3; 17]). One problem with competition is the precise quantification of its effect. [53] propose a spatio-temporal model to measure similarity among concurrent auctions. [44] take a functional approach to visualize concurrent auctions and their dynamics. (See also [50] for additional visualizations of concurrent auctions.) In this chapter, we propose several innovative measures for auction competition using the concept of functional data analysis. Our measures can be grouped into three conceptually different classes: measures that capture competition from static (or pre-auction) information; from time-varying information; and from dynamics. We perform variable selection to identify the most predictive set of competition measures.

Our proposed forecasting model incorporates measures of dynamics and competition as predictors. In contrast to [96], our model also takes into account competition; in contrast to [53], we measure competition in ways that are easily scalable and do not rely on spatial methodology. We compare our model's predictive capability to several alternate approaches, and find that our model can predict an auction's price with higher accuracy.

In the second component, we build a comprehensive bidding strategy around our forecasting model. The idea is based on maximizing consumer surplus, which refers to the difference between the bidders' *willingness to pay* (WTP) and the price *actually paid*. We formulate an automated algorithm for selecting the best auction to bid on, and for determining the best bid-amount. The best auction provides bidders with the highest surplus, and the best bid-amount equals the predicted closing price.

We conduct a simulation study where we compare our automated, data-driven bidding strategy with early bidding and last-minute bidding. We compare all bidding strategies on two different dimensions: the probability of winning an auction, and the surplus extracted. We find that although snipers have the highest probability of winning, our strategy results in a much higher surplus. We also investigate the impact of the *prediction window* on the resulting surplus. The prediction window is equivalent to the given time frame within which a bidders wants to win an auction. Shorter time frames correspond to bidders that want to win more quickly; longer time frames correspond to bidders that allow more time for search and selection. We find that, as the width of the prediction window is increasing, surplus goes up while the probability of winning goes down.

The chapter unfolds as follows: In Section 2.2, we describe the data; in Section 2.3, we derive our forecasting model. Results of model estimation are discussed in Section 2.4. In Section 2.5, we present the framework for our automated bidding strategy and the results of our simulation study. We conclude with further research directions in Section 2.6.

2.2 Data Description

The data used in this study are the complete bidding records for new Palm M515 handheld devices, auctioned between March 14, 2003 and May 25, 2003. The market price at the time of data collecting was $230 (based on `Amazon.com`). Each bidding record includes the auction number, starting- and closing-time and -prices, bids with associated time stamps, and other information, such as auction duration, shipping fee, seller's feedback score, whether the seller is a power seller, whether the product is from an eBay store, and whether auction's description includes pictures. A summary of this information can be found in Table A.1.

Figure 2.1 illustrates the information-overload that bidders face. In particular, we see, for each individual auction, the *live price curve*, that is, the price that bidders see at any given time during the ongoing auction. We can see that the information can be quite overwhelming: the amount of concurrent auctions, the variation in prices and the fact that some auctions are only in the early stages, while others are about to end, all cause challenges for properly processing the given information.

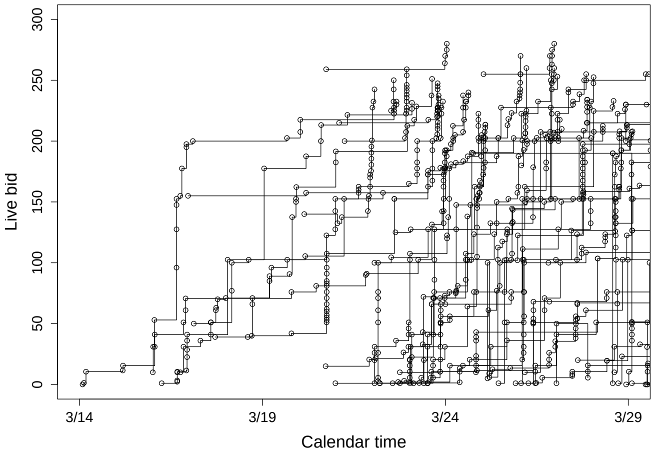


Figure 2.1: Snapshot of the live price curves during eBay auctions.

Moreover, we see that prices increase unevenly throughout most auctions. They increase fast in some auctions, but much slower in others. We refer to this as *price dynamics*, which will be an important factor in our modeling approach.

2.3 A Model for Forecasting Competing Auctions

Our forecasting model has several features: We model the real-time price process of ongoing auctions using functional data analysis (FDA), which allows us to incorporate information about the dynamics of price. We also propose several innovative ways of incorporating competition across concurrent auctions, and then we suggest an innovative way to perform model selection[1] and model updating. We describe these features in detail next.

[1] A more complete study about the model selection can be found in next chapter.

2.3.1 Functional Data Analysis and Price Dynamics

The price process of online auctions is characterized by an extremely dynamic environment. One aspect of this environment is the changing bid density, where the number of bids per time unit changes constantly. The resulting unequally-spaced time-series of bids deem traditional models (which assume evenly spaced measurements) inadequate. Furthermore, the changing bidding patterns also result in varying price dynamics. By price dynamics we mean the change in price and the rate at which this change occurs. Traditional forecasting models, which do not account for such instantaneous change, fail to accurately predict auction prices [96]. To incorporate this dynamic environment, we take a functional data modeling approach.

Functional data analysis [79] uses smoothing methods[2] to recover (or estimate) the underlying price curve from observed bidding data. From the price curve, we then obtain estimates of the price dynamics via its first and second derivatives. Figure 2.2 illustrates the process of generating smooth price curves from observed data (left panel) and estimating the corresponding price velocities (right panel). We see that the smooth curves capture the trend of the price increase due to the discrete bids; the velocity captures the instant change of price increase. For more details on FDA in the context of online auctions, refer to [84] or [52]; and for more details on the smoothing process, see Section 5.2.1 in Chapter 5.

[2] In particular, we use polynomial smoothing methods (p-splines) in this study.

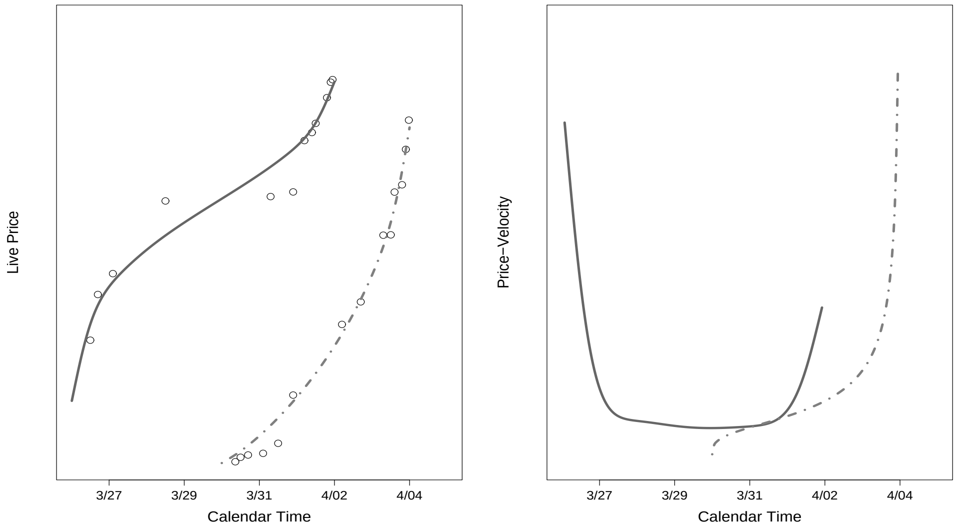


Figure 2.2: Smooth price curves (left panel) and corresponding velocities (right panel) for two sample auctions. The dots in the left panel denote the observed bids.

2.3.2 Capturing Competition

One major component of our model is competition. That is, we want to capture the effect of what happens in other, simultaneous auctions. To that end, we must first define meaningful measures for competition. There are many different ways of defining competition measures and we explore several alternatives below. All measures are driven by the same general principle which is illustrated in Figure 2.3. We define a focal auction (indicated by the solid line in Figure 2.3) as the auction for which a bidder wants to decide whether or not to bid on. At time T of decision-making, there are several other auctions that take place simultaneously (indicated by the dotted lines). One meaningful measure of competition is the level of price in other auctions. In our example, there are four different prices levels at time T, varying from high (p1) to low (p4). The price level in the focal auction at that time is p3. Thus, a possible measure for the price competition is given by the *average price* in concurrent auctions (which we denote by *c.avg.price*), that is, by the average of p1, p2 and p4. In similar fashion, the *average price velocity* (*c.avg.vel*) in concurrent auctions would be defined as the average of the corresponding price velocities, and so on.

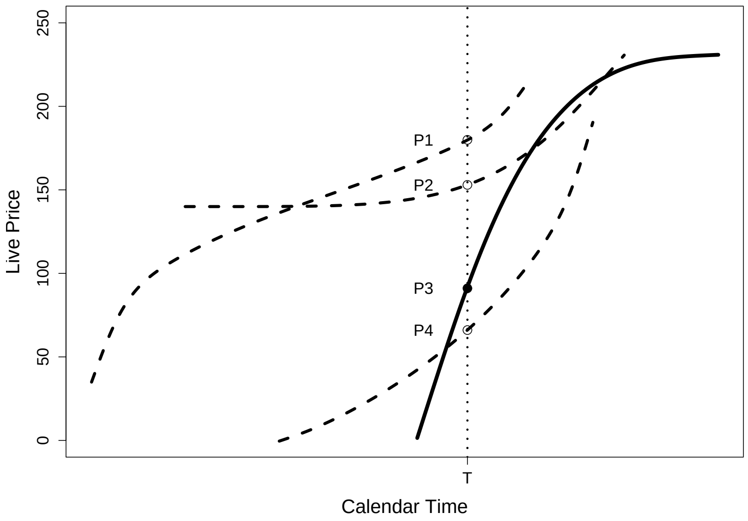


Figure 2.3: Illustrating competition: The sold black line denotes the focal auction; the dashed lines denote competing auctions; T denotes the time of decision-making.

In this chapter, we investigate several different competition features and their impact on the price of the focal auction. Table 2.1 categorizes these features by the information that they carry: *Static competition features* are known at the outset of the auction and do not change during the auction process; examples include the opening price of concurrent auctions (a high opening price in other auctions could discourage bidders and make them participate in the focal auction) or the duration of concurrent auctions (if competing auctions have a shorter duration, then bidders with an immediate desire may be attracted to those auctions); *Time-varying competition features* change during the auction process, such as the current price of concurrent auctions (if the price is low in other auctions, bidders may leave the focal auction) or the number of bidders of concurrent auctions (bidders may feel that their chances of winning are higher in auctions with lower competition); and *price dynamic competition features* capture the effect of changing dynamics in competing auctions (if the price dynamics increase in competing auctions, e.g., due increased bidding activity in those auctions, then the price speed in the focal auction is likely to slow down).

In Figure 2.4, we explore the relationships between some of the competition features from Table 2.1 and the future price in the focal auction. We can see that some features (e.g., the average price and its velocity in competing auctions) have a strong relationship with price, while others (e.g., the average opening bids or the shipping fee in simultaneous auctions) have a rather weak relationship. Pairwise correlation analysis (not reproduced here) also shows that, unsurprisingly, many of the features in Table 2.1 are multicollinear. Thus, a good modeling strategy will start with a suitable variable selection procedure. We will use the initial observations from Figure 2.4 for guidance when selecting the most relevant set of competition features in the next section.

2.3.3 Variable Selection

Many different pieces of information can affect price in online auctions. We differentiate between two main components, information from *within* the focal auction vs. information from other, *competing* auctions that take place simultaneously.

Table 2.1: Candidate Competition Features

Name	Description
	Static Features
c.openbid.avg	Average opening price of concurrent auctions
c.dura.avg	Average duration of concurrent auctions
c.ship.avg	Average shipping fee of concurrent auctions
c.feedback.avg	Average sellers' feedback of concurrent auctions
c.power.avg	Average number of power seller in concurrent auctions
c.store.avg	Average number of eBay stores in concurrent auctions
c.pic.avg	Average number of pictures in concurrent auctions
	Time-varying Features
c.price.avg	Current average price in concurrent auctions
c.price.vol	Current price volatility (stdev) in concurrent auctions
c.price.disc	Price discount (difference) between focal auction and highest concurrent price
c.t.left.avg	Average time left in concurrent auctions
c.t.left.vol	Volatility (stdev) of time left in concurrent auctions
c.nbids.avg	Average number of bids in concurrent auctions
c.nbids.vol	Volatility (stdev) of number of bids in concurrent auctions
c.nbidders.avg	Average number of bidders common to focal and concurrent auctions
c.nbidders.vol	Volatility (stdev) of number of bidders common to focal and concurrent auctions
	Price dynamic Features
c.vel.avg	Average price velocity in concurrent auctions
c.vel.vol	Volatility (stdev) of price velocity in concurrent auctions
c.acc.avg	Average price acceleration in concurrent auctions
c.acc.vol	Volatility (stdev) of price acceleration in concurrent auctions

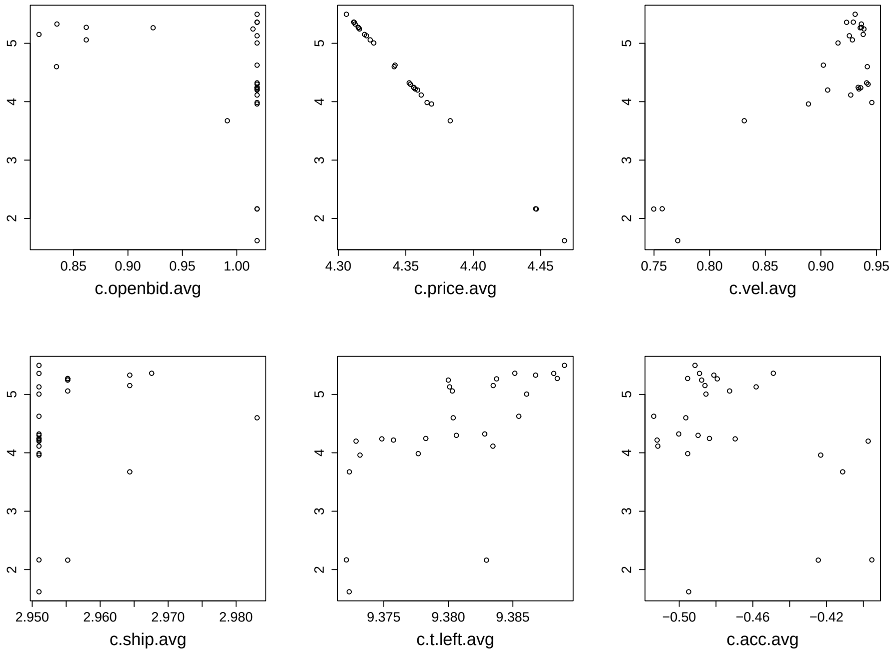


Figure 2.4: Pairwise relationships between some of the competition features from Table 2.1 (measured at time T) and the price (measured at $T+1$) in the focal auction.

Within each component, information can be further segmented into static, time-varying and price dynamic information, similar to Table 2.1. Table 2.2 lists all the different pieces of information that are candidates for our forecasting model.

Table 2.2 shows that there are over 30 different variables that are candidates for our forecasting model. Thus, an important first step in our modeling efforts is the selection of a parsimonious subset of relevant predictors. Variable selection has been researched in the statistics literature for a while [13] and it is receiving increasing attention today with the availability of more and more data sets featuring larger and larger number of variables [30]. A complicating factor in our situation is the time-varying nature of our model. Our goal is to find a model that predicts well at time $T+1$, *universally* across all time periods $T = 1, 2, 3, \ldots, N_T$. Classical variable selection procedures focus on cross-sectional data, that is, on data corresponding to a *single* time period only. Since our data varies over time, it is quite plausible that there exists *one* model that best predicts at time $T + 1$, while *another* (different) model best predicts at a different time $T' + 1$. Our goal is to find a model that is not geared to a single time period only, but applies rather universally to the eBay market over a longer time window. To that end, we choose a model which has good *average* performance[3], averaged over all time periods T of interest. We describe this approach next.

Table 2.2: Candidate information for the forecasting model

	Information from within the focal auction
Static information	opening bid, auction duration, shipping fee, seller's feedback, power seller, eBay store, picture
Time-varying information	current price, time left, current number of bids, current number of bidders
Price dynamic information	price velocity, price acceleration
	Information from competing auctions
Static information	c.openbid.avg, c.dura.avg, c.ship.avg, c.feedback.avg, c.power.avg, c.store.avg, c.pic.avg
Time-varying information	c.price.avg, c.price.vol, c.price.disc, c.t.left.avg, c.t.left.vol, c.nbids.avg, c.nbids.vol, c.nbidders.avg, c.nbidders.vol
Price dynamic information	c.vel.avg, c.vel.vol, c.acc.avg, c.acc.vol

[3] Note that our decision to use the criterion that has the best "average" performance is rather intuitive. We conduct a more complete study on different criteria in the next chapter.

Our model has the general form

$$\mathbf{y}_{T+1} = \boldsymbol{\beta}'_T \mathbf{x}_T \tag{2.1}$$

where $\mathbf{y}_{T+1}$ denotes the auction prices at $T+1$, $\mathbf{x}_T = (x_{T1}, \ldots, x_{Tp})'$ is a vector of predictors, and $\boldsymbol{\beta}_T = (\beta_{T1}, \ldots, \beta_{Tp})'$ is a vector of coefficients to be estimated from the data. The goal is to select only those predictors that are important for predicting the price $\mathbf{y}_{T+1}$, across all time periods $T = 1, 2, 3, \ldots, N_T$.

We accomplish this in several steps. In the first step, we run simple regressions (i.e. $p = 1$) between each individual predictor from Table 2.2 and the response $\mathbf{y}_{T+1}$ at *each* time period $T, T = 1, 2, 3, \ldots, N_T$. We then calculate the percentage of time points a predictor is significant (at the 5% significance level). That is, for each predictor $x_k = (x_{1k}, \ldots, x_{N_T k})'$, $k = 1, \ldots, p$, we compute the average[4]

$$p.sig_k := \frac{1}{N_T} \sum_T \mathbf{1}_{\{x_{Tk} \text{ significant at 5\% level}\}}. \tag{2.2}$$

Table 2.3 shows the results for a fine grid of hourly forecasts (i.e. $(T+1) - T = 1$ hour) which results in $N_T = 1,754$ different time periods. We can see that the predictors that *individually* have a strong effect on $\mathbf{y}_{T+1}$ (consistently across all time periods T) are the current price, price velocity and acceleration, time left and the number of bids (from within the focal auctions) and c.price.avg, c.price.vol, c.price.disc, c.t.left.avg, c.t.left.vol, c.nbids.avg, c.nbids.vol, c.vel.avg, c.vel.vol, c.acc.avg and

[4] While we use an un-weighted average, a possible alternative would be to weight each time point according to its distance from the close of the auction.

Table 2.3: Percentage of significant time points. The two leftmost columns refer to predictors from within the focal auction; the two rightmost columns refer to predictors from competing auctions.

Focal auction	*p.sig*	Competing auctions	*p.sig*
openbid	.199	c.openbid.avg	.193
duration	.032	c.dura.avg	.032
shipping	.039	c.ship.avg	.046
sellerfeed	.055	c.feedback.avg	.044
powerseller	.061	c.power.avg	.076
store	.092	c.store.avg	.104
picture	.028	c.pic.avg	.028
currenprice	1.00	c.price.avg	1.00
		c.price.vol	.886
		c.price.disc	1.00
timeleft	.775	c.t.left.avg	.771
		c.t.left.vol	.758
numbids	.780	c.nbids.avg	.777
		c.nbids.vol	.509
numbidders	.197	c.nbidders.avg	.188
		c.nbidders.vol	.086
price velocity	.762	c.vel.avg	.762
		c.vel.vol	.624
price acceleration	.308	c.acc.avg	.309
		c.acc.vol	.306

c.acc.vol (from competing auctions). It is interesting that most of these variables relate to price (or price movement) from the focal auction relative to competing auctions. This suggests that information about price and its dynamics effectively captures much of the relevant auction information such as information about the product, the auction format, the seller and competition between bidders. However, also note that the results so far are based only on simple regressions ($p = 1$) and thus may not fully reflect the *joint* effect of a predictor in the presence of other predictor variables. To that end, we investigate pairwise correlations (again, averaged across all time periods, $T = 1, \ldots, N_T$; correlation-table not reported here) and find high collinearity between ten pairs: the current price and c.price.avg, the current price and c.price.vol, the current price and c.price.disc, the current price and time left, the current price and c.t.left.avg, the current price and number of bids, the current price and c.nbids.avg, price velocity and c.vel.avg, price velocity and c.vel.vol, and price acceleration and c.acc.avg. This high collinearity is not surprising since many of these predictors carry similar information, only coded in a slightly different way. We eliminate all highly collinear predictors; next we derive our final model using the Bayesian Information Criterion (BIC).

In similar fashion to equ.(2.2), one can compute the *average* BIC across all time periods. That is, let avg.BIC := $\text{mean}_T(\text{BIC}(T))$, where $\text{BIC}(T), T = 1, 2, 3, \ldots, N_T$, denotes the Bayesian Information Criterion (e.g. [30]) of a model computed at time period T [5]. By comparing all possible subsets of non-collinear predictors, we arrive

[5] Note that we calculate the average of BIC across only the time points where BIC is applicable. That is, time points where BIC is not available due to non-sufficient data for modeling is ignored in this definition.

at our final forecasting model as

$$\mathbf{y}_{T+1} = \alpha_T + \beta_{1T}\text{current price}_T + \beta_{2T}\text{velocity}_T + \beta_{3T}\text{acceleration}_T - \beta_{4T}\text{c.acc.vol}_T. \tag{2.3}$$

Table 2.4 shows the avg.BIC of our final forecasting model (2.3) compared to several competitor models. We can see that our model results in the lowest avg.BIC. It is also interesting to see that models with *only* information from competing auctions perform almost as well as models with the corresponding information only from within the focal auction. This is yet another piece of evidence for the tight connectivity of the auction marketplace.

Table 2.4: Average BIC computed across all time periods T. The first row shows the value of $avg.BIC$ for our model in (2.3); the remaining rows show the corresponding values of several competing models.

Model	*avg.BIC*
Our model from eq. (2.3)	**-381.59**
Full model (all 33 predictors from Table 2.2)	-147.08
All 13 predictors from the focal auction (Table 2.2)	-319.82
Only 2 focal auction dynamics	37.77
Only 4 focal auction time-varying predictors	-83.87
All 20 predictors from competing auctions (Table 2.2)	-313.52
Only 4 competing auction dynamics	37.58
Only 9 competing auction time-varying predictors	-84.29

A few comments about our final model in (2.3) are in order. It is interesting to see that the model relies only price related information. In particular, it is interesting to see that many variables that have been found significant in previous studies have dropped out of our model. For instance, [74] find, among other things, a significant effect of the seller's rating. One key difference between previous studies and our study is that while they take a static look at online auctions, our model captures the dynamic nature of the auction process. In other words, previous studies typically only look at the static, pre-auction information that is available before the start of the auction (such as the auction length, the opening bid or a seller's rating). In such a static view, the effect of the seller's rating is highly significant (since the seller's reputation and trustworthiness will impact the final price). However, our model is dynamic in the sense that all previous price considerations and bidding decisions have already been factored into the current price and its current dynamics $(\text{current price}_T, \text{velocity}_T, \text{acceleration}_T)$. In that sense, price dynamics reflect the expectations of all bidders about the product, the seller and the bidding competition up to this time point. It is thus not too surprising that all static variables drop from our final model. The effect that captures concurrency is more intriguing. Note that the information from concurrent auctions is captured in a single variable, the volatility of dynamics from competing auctions (c.acc.vol_T). As there has not been much prior research on the effect of concurrent auctions, it is hard to formulate an expectation about c.acc.vol_T. However, the negative sign indicates that higher variance in the price movements of competing auctions will result in smaller price advances of the focal auction. In other words, more price activity in different parts of the market will lead to price stalling of the focal auction. We will conduct a more complete study regarding the model selection in Chapter 3.

2.3.4 Model Updating

The goal of our model is to predict price at a future time $T+1$ using only information from the present (i.e. time T) and the past ($T-1$, $T-2$, etc.). We accomplish this by estimating the functional relationship between $T-1$ and T and then applying this relationship to predict $T+1$ from T. Figure 2.5 illustrates this updating scheme.

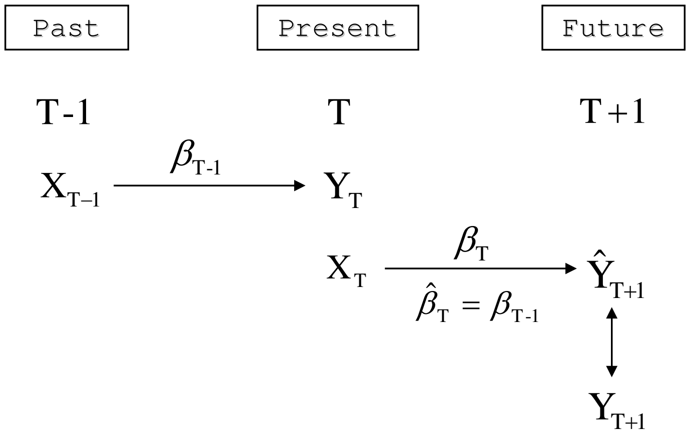


Figure 2.5: The illustration of the update scheme in the forecasting model

At time T (present), we wish to make a prediction about the future price at time $T+1$. Per our model, $\mathbf{y}_{T+1}$ is given by $\boldsymbol{\beta}'_T\mathbf{x}_T$, where $\mathbf{x}_T$ contains information observed in the present (or past). Note that we cannot estimate $\boldsymbol{\beta}_T$ directly since the response ($\mathbf{y}_{T+1}$) is yet unobserved. We therefore estimate the relationship from the past: We estimate $\boldsymbol{\beta}_{T-1}$ for the price at T ($\mathbf{y}_T$) and then estimate $\boldsymbol{\beta}_T$ via $\hat{\boldsymbol{\beta}}_T := \boldsymbol{\beta}_{T-1}$. In that sense, we "roll" the relationship from the past one time period forward.

We also investigated alternate updating approaches (such as estimating $\boldsymbol{\beta}_T$ via a moving average (MA) of prior relationships, $\hat{\boldsymbol{\beta}}_T := \text{MA}\{\boldsymbol{\beta}_{T-1}, \boldsymbol{\beta}_{T-2}, \boldsymbol{\beta}_{T-3}, \ldots\}$) but did not find significant improvements in model performance.

2.4 Estimation and Prediction Results

In this section we discuss estimation and prediction of our forecasting model in (2.3). We also compare its predictive capabilities to alternate forecasting approaches.

To that end, we divide our data set into a training set (80% of the data), and a validation set (remaining 20% of the data). Since our data varies over time (and since we are primarily interested in making accurate predictions of the future), our training set consists of all auctions that complete during the first 80% of our data's time span (i.e. between March 14 and May 10); the validation set contains all remaining auctions (i.e. between May 11 and May 25). In that sense, we first estimate our model on the training set; results of model estimation and -fit are discussed below. We then apply the estimated model to the validation set to gauge its predictive capabilities; this is discussed in the second half of this section.

2.4.1 Model Estimation

Figure 2.6 shows the estimated coefficients for the parameters of our forecasting model (2.3). Recall that we estimate the model at every time point T, $T = 1, 2, 3, \ldots, N_T$ in the training set. In our application, we consider time intervals of one hour over the time period between March 14 and May 10, hence the coefficients also vary over that time period. Figure 2.6 shows the resulting trend of the coefficients together with 95% confidence bounds.

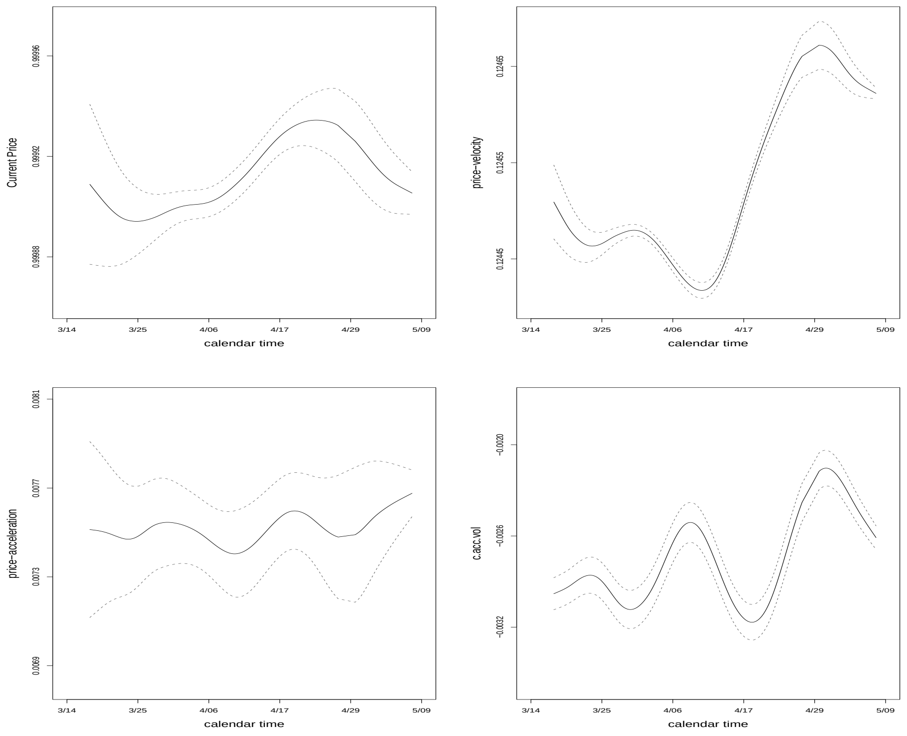


Figure 2.6: Estimated coefficients of model parameters, together with 95% confidence bounds. The x-axis denotes calendar time; the y-axis denotes the magnitude of the coefficient. The panels show (from top left to bottom right) current price, price velocity, price acceleration and the acceleration volatility of competing auctions (c.acc.vol).

We can see that information from within the focal auction (current price, price velocity and acceleration) has a positive relationship with the future price $\mathbf{y}_{T+1}$; in contrast, information from competing auctions (c.acc.vol) has a negative relationship. In other words, both the current level of price and its dynamics are positive indicators of future price. On the other hand, the volatility of price acceleration in competing auctions is a negative indicator. Price acceleration in competing auctions will be high if many bidders bid in auctions *different* from the focal auction.

A high volatility in price acceleration may suggest high uncertainty in the marketplace, with some auctions experiencing large price jumps and others experiencing no price movements at all. This high uncertainty results in depressed prices of the focal auction.

2.4.2 Model Fit and Varying Time Intervals

Figure 2.6 shows the estimated model coefficients for one hour time intervals; that is, for $(T + 1) - T = 1$ hour. Alternatively, one could also consider models with a larger time intervals; that is, models that forecast further into the future. Intuitively, since forecasting further into the future is harder, such models should not perform as well. Figure 2.7 (left panel) shows the model fit for the time intervals $(T + 1) - T = 1, 2, 3, \ldots, 14$ hours. We measure model fit by the average R^2 value, avg.$R^2 := (1/N_T) \sum_T R^2(T)$, where $R^2(T), T = 1, 2, 3, \ldots, N_T$, denotes the R^2 of a model computed at time period T. We can see that, as expected, the model fit decreases as the time intervals get larger. Notice though that even for the largest time interval (14 hours), the value of avg.R^2 is still larger than 99%.

2.4.3 Prediction Performance

As pointed out above, we estimate the model on the training set; then we gauge its predictive performance on the validation set. We measure predictive performance of a model in terms of its *mean absolute percentage error* (MAPE). For each time

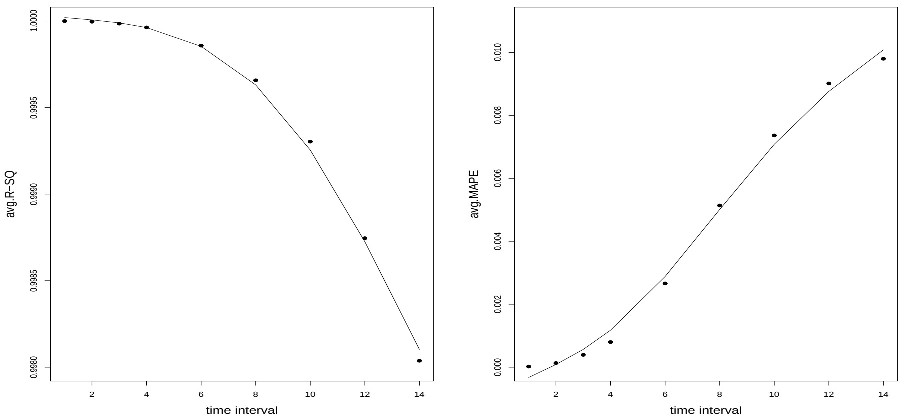


Figure 2.7: Model fit and prediction accuracy for different time intervals. The x axis represents the time interval (in hours; ranging from 1 hr to 14 hrs); the y axis represents the value of avg.R^2 (left panel) and the value of avg.MAPE (right panel).

period T, $T = 1, 2, 3, \ldots, N_T$, in the validation set, we compute

$$\text{MAPE}(T) = \frac{1}{m_{T+1}} \sum_{i=1} \frac{|y_{T+1,i} - \hat{y}_{T+1,i}|}{|y_{T+1,i}|} \tag{2.4}$$

where $y_{T+1,i}$ and $\hat{y}_{T+1,i}$ denote the true and predicted values of auction i at time $T + 1$, respectively, and m_{T+1} denotes the number of auctions available at time $T + 1$. We compute the *average MAPE* across all time periods as avg.MAPE := $(1/N_T) \sum_T \text{MAPE}(T)$. In similar fashion to Section 2.4.2, we investigate avg.MAPE for different time intervals, $(T + 1) - T = 1, 2, 3, \ldots, 14$ hours. The right panel in Figure 2.7 shows the results.

Unsurprisingly, we see that as we predict further into the future (i.e. as time interval gets larger), the predictive performance decreases (i.e. avg.MAPE increases). It is interesting to see that for predictions up to 4 hours into the future, the prediction error is less than 0.1%. For time intervals larger than 4 hours, the prediction error increases at a faster rate. However, even for predictions as far as 14 hours into the future, the error is still less than 1%. This predictive accuracy is quite remarkable as we will see in the next subsection where we benchmark our approach against several competing approaches. We also want to note that while we cannot claim generalizability to all eBay auctions, there has been prior evidence that real-time forecasting models can provide superior predictive accuracy, especially for books and electronics (see [96]).

2.4.4 Comparison with Alternative Models

We benchmark our model against five alternative models, the generalized additive model (GAM), classification and regression trees (CART), Neural Networks and two simper linear models: a purely static and an time-varying linear model.

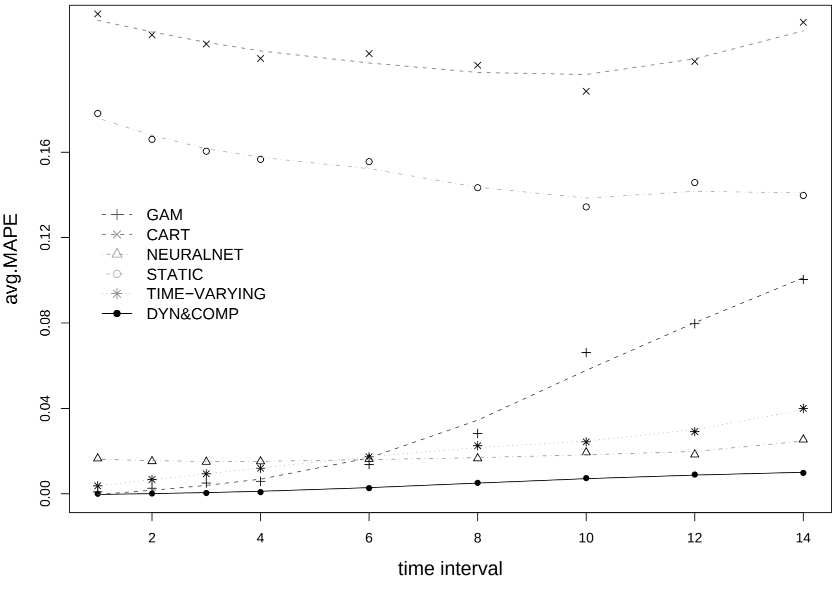


Figure 2.8: Prediction accuracy for competing models. The x axis represents the time interval (in hours); the y axis represents the value of avg.MAPE.

GAMs relax the restrictive linear model assumption between the response and predictors by a more flexible nonparametric form [41]. CARTs [15] provide a data-driven way to partition the variable-space and are thus often viewed as alternatives to formal variable selection. Neural Networks also provide a technique that can approximate non-linear functional relationships. In addition, we consider two linear models that use a subset of the variables from Table 2.2: one model that uses only static information from the focal auction and another one that uses static and time-varying information from the focal auction; we refer to these two models as "STATIC" and "TIME-VARYING," respectively. The static model corresponds to the information of many prior eBay studies (e.g.[74]) in that it only considers pre-auction information. The time-varying model accounts for changes due to the process of bidding, but it does not account for price dynamics or competition.

Figure 2.8 shows avg.MAPE (similar to Figure 2.7) for time intervals $(T+1)-T = 1, 2, 3, \ldots, 14$ hours, for all 6 different models. We refer to our model (2.3) as "DYN&COMP," since it incudes dynamics and competition features. We can see that STATIC and CART have the worst prediction performance, with an error uniformly larger than 10%. While our model performs the best, GAM, TIME-VARYING and Neural Nets are competitive, at least for smaller time intervals. In other words, for predicting less than 4 hours into the future, both GAM and TIME-VARYING pose alternatives with prediction errors not too much larger compared to DYN&COMP. However, their predictive performance breaks down for larger time intervals. In fact, the error of GAM is as large as 10% for predicting 14 hours into the future, which is 10 times larger than the corresponding prediction error of DYN&COMP. While the performance of TIME-VARYING is somewhat better, its prediction error is 4 time as large as DYN&COMP for 14 hours time intervals (similar for the Neural Network). In the next section, we use the excellent forecasting performance of our model and build an automated bidding strategy around it.

2.5 An Automated Data-Driven Bidding Decision Rule

We now discuss the second component of our bidding strategy, building an automated and data-driven decision rule around our forecasting model. The decision rule provides answers to three basic bidding questions: which auction to bid on, when to bid on it and how much.

2.5.1 Decision Framework

Our decision framework is built upon the principles of maximizing consumer surplus (e.g.[11]). Consumer surplus is the difference between the actual price paid and the consumer's willingness to pay for an item, CS = WTP - Price, where CS denotes consumer surplus, and WTP denotes willingness to pay. Therefore, the lower the price, the higher is a bidder's surplus.

For each individual auction, our forecasting model (2.3) provides bidders with that auction's estimated future price; combining this with a bidder's WTP leads to an auction's estimated surplus. For a set of competing auctions, a plausible decision rule is to bid on that auction with the highest estimated surplus. Moreover, in order to avoid a negative surplus, a bidder should only bid on an auction if the predicted price is lower than his WTP.

Note that our forecasting model depends on the length of the time interval $(T+1)-T$ (which we also refer to as the *prediction window*). Our model can only predict the final price of an auction that ends at or before time $(T+1)$. Therefore, longer prediction windows will result in a larger number of candidate auctions, that is, in a larger supply of *potential* auctions to bid on. On the other hand, we have also seen in Section 2.4.3 that a larger time interval leads to an increased prediction error. Therefore, our decision rule faces a trade-off between supply of candidate auctions and prediction accuracy for each individual auction. We will investigate this trade-off in detail below.

Our decision rule picks that auction with the highest estimated surplus, as long as the surplus is positive. After picking an auction, the next two questions are with respect to the *time* and *amount* of the bid. Since our forecasting model is based on a fixed time interval, nothing is gained by waiting. So we suggest placing the bid as soon as an auction is picked. Moreover, since our model predicts an auction's closing price at $\hat{y}_{T+1}$, we would expect to lose for bids lower than $\hat{y}_{T+1}$. Similarly, bids higher than $\hat{y}_{T+1}$ are expected to overpay. Therefore, we suggest to bid *exactly* the expected (or predicted) closing price $\hat{y}_{T+1}$. In summary, our decision rule picks the auction with the highest predicted surplus, it bids the predicted price, and it places the bid immediately.

2.5.2 Experimental Set-Up

We conduct a simulation study to compare our automated bidding strategy to two alternate (and popular) bidding approaches: *early bidding* [8] and *last-minute bidding* [83]. Early bidding is often viewed as a bidder's strong commitment and intends to deter others from entering the auction. Last-minute bidding is popular because it does not allow much time for other bidders to react. In our simulation, we assume that a bidder's willingness-to-pay (WTP) is drawn from a uniform distribution [2] distributed symmetrically around the market value (\$230 at the time of data-collection). That is, we assume WTP $\sim$ Uniform(\$220, \$240). Our experiment then proceeds as follows. We randomly draw a WTP from that distribution. We also draw an auction from the validation set (i.e. we compare the bidding strategies on the same real-world data that we compare the forecasting models). The bidder then makes a bidding decision (whether or not to bid, and how much to bid) with each of the three bidding strategies outlined below. We repeat this experiment for all auctions in the validation set and for 20 different random draws from a bidder's WTP distribution.

2.5.2.1 Early Bidding Strategy

We assume that early bidders bid at the end of the first auction day [8]. In fact, we find that slightly earlier or later bid times barely affect the outcome of the experiment. The process of early bidding is illustrated in the left panel of Figure 2.9. A bidder compares his WTP with the auction's current price at the end of

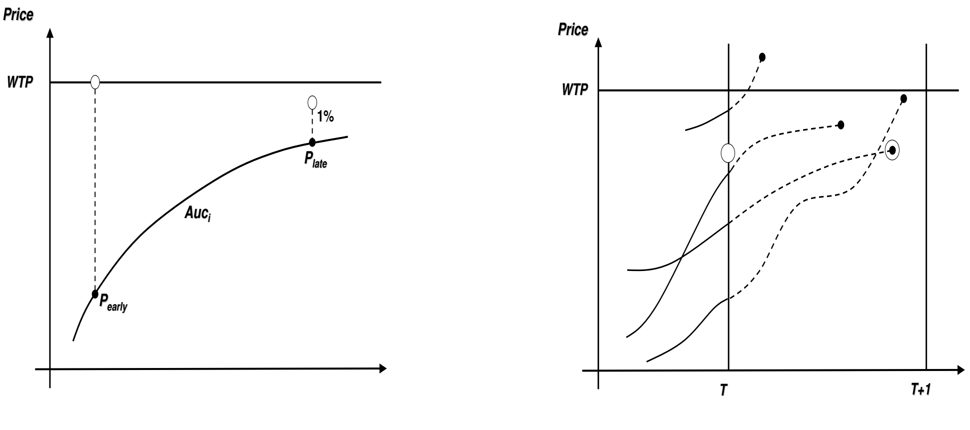


Figure 2.9: Illustration of bidding strategies. The left panel illustrates early and last-minute bidding; the right panel illustrates our automated bidding strategy.

the first day (p_{early}); if his WTP is higher, then he places a bid; otherwise, he does not place a bid and moves on to another auction. If he does place a bid, then the bid amount equals the WTP. Note though that due to eBay's proxy bidding system which incrementally bids up to the WTP on behalf of the bidder, the final price may be lower than the WTP. As a consequence, the bidder only pays the amount of the second-highest bid plus a pre-specified bid-increment (which ranges between $2.5 and $5 in our case). We also investigate alternate bidding heuristics in Appendix B. However, none of these heuristics beat last-minute bidding or our automated bidding strategy.

2.5.2.2 Last-Minute Bidding Strategy

We assume that last-minute bidders place their bid one minute before the auction closes [83]. Last-minute bidding carries the danger that the bid does not go through due to network congestion, but we will not explicitly consider this disadvantage in our simulations. The process of last-minute bidding is again illustrated in the left panel of Figure 2.9. A bidder compares his WTP with the auction's current price one minute before closing (p_{late}). Similar to early bidding, if his WTP is higher, then he places a bid; otherwise, he does not place a bid and moves on to another auction. If he does place a bid, then the bid-amount is only incrementally higher than the current price, since the chances of being outbid within the last 60 seconds are small. In our simulations, we bid an increment of 2% over the current price p_{late}. We also study the robustness to different increments in Appendix B and find that bid-increments of 1%, 2% or \$2 yield comparable results.

2.5.2.3 Our Automated Data-Driven Bidding Strategy

Our automated bidding strategy is conceptually different from early and last-minute bidding. Instead of making a bidding decision for each auction individually, our strategy requires a bidding decision for each time interval. Consider the right panel of Figure 2.9. At time T of decision making, there are four competing auctions, denoted by Auc_1-Auc_4, which all close before time $T+1$. The solid lines correspond to the observed part of the auction history; the dotted lines denote the future (and yet unobserved) price path. Since all auctions close before $T+1$, our model yields predictions of their final prices (denoted by the solid black circles). Note that Auc_4 has the highest predicted price; moreover, its predicted price is higher than the bidder's WTP; hence the bidder will never consider this auction. Auc_3

has the smallest predicted price; since the predicted price is also smaller than the bidder's WTP, he places a bid on this auction. He bids the predicted price and he bids immediately, i.e. at time T. If he wins, then the bidder's surplus will be the difference between his WTP and the *actual* closing price.

2.5.3 Simulation Results

Similar to [8], we compare all bidding strategies on two dimensions: the probability of winning the auction and the average surplus accrued. We compute the probability of winning (p.win) as the number of auctions won divided by the total number of auctions that the bidder placed a bid on. We compute the average surplus (avg.sur) as the corresponding difference between the WTP and actual price paid for an auctions. The results are shown in Table 2.5.

Table 2.5: Comparison of different bidding strategies. The first row corresponds to last-minute bidding; the second row corresponds to early bidding; and the last row corresponds to out automated bidding strategy. We report the mean estimates (with standard errors in parentheses).

	p.win	avg.sur
Last-moment bidding	95% (.5%)	\$17.97 (\$0.35)
Early bidding	53% (2%)	\$18.85 (\$0.57)
Automated bidding	61% (1%)	\$32.33 (\$1.95)

We see that last-minute bidders have the highest probability of winning (95%, compared to 61% for our automated bidding strategy). This is not surprising, since last-minute bidding is geared to out-witting the competition in the last moment. However, we also see that last-minute bidding accrue a significantly lower surplus compared to our automated bidding strategy (\$18 vs. \$32). Another way of comparing the two bidding strategies is via their *expected surplus*, i.e. the product ($p.win \times avg.sur$). We find that last-minute bidding yields an expected surplus of \$17.11 [6] while that of our automated bidding strategy is higher: \$19.72. Moreover while early bidders have a probability of 53% of winning the auction, their expected surplus is significantly lower: \$9.99 (=53% × \$18.85).

2.5.3.1 Effect of the Prediction Window

We have pointed out earlier that the length of the prediction window (i.e. the length of the time interval $(T + 1) - T$) has an effect on the outcome of our automated bidding strategy in that longer windows result in a larger supply of candidate auctions, but at the same time reduce the prediction accuracy of each individual auction. The results from the previous section (Table 2.5) are based on a prediction window of 12 hours and we have seen that it yields an expected surplus of \$19.72 for our automated bidding strategy. Longer prediction windows yield a larger number of candidate auctions and as such a larger probability of including an auction with a lower price (and hence a higher surplus). On the other hand, longer prediction windows also lead to less accurate predictions. Less accurate predictions can either lead to overpayment (if the predicted price, and hence our bid, are higher than the actual price); overpayment leads to a lower surplus. Less accurate predictions can also lead to a reduced probability of winning the auction (if the predicted price, and hence our bid, are lower than the actual price). Thus,

[6] For the expected surplus corresponding to other bid-increments, please refer to Appendix B.

a change in the prediction windows affects both the probability of winning as well as the average accrued surplus and it is not quite clear how it affects the overall *expected surplus*. To that end, we repeat the simulation study from Table 2.5 for prediction windows of different lengths. Table 2.6 shows the results.

Table 2.6: Tradeoff between the width of the prediction window and expected surplus. The first column denotes the width of the window; the second column denotes the probability of winning; the third column denotes the average accrued surplus; and the last column denotes the expected surplus, i.e. exp.sur = (p.win × avg.sur).

prediction window	p.win	avg.sur	exp.sur
14hrs	59.01%	\$35.29	\$20.82
12hrs	61.44%	\$32.33	\$19.80
9hrs	67.82%	\$30.99	\$21.02
6hrs	69.43%	\$29.60	\$20.55
3hrs	75.77%	\$27.21	\$20.62

We can see that larger prediction windows result in a larger average surplus which suggests that the effect of having a larger pool of candidate auctions outweighs the effect of overpayment. But we also see that larger windows result in a smaller probability of winning since the less accurate predictions more frequently yield bids below the auction's actual closing price and hence an unsuccessful auction. Interestingly, the *expected surplus* is maximized for a prediction window of 9 hours. While our results do not prove optimality, they suggest a very interesting global optimization problem for future research.

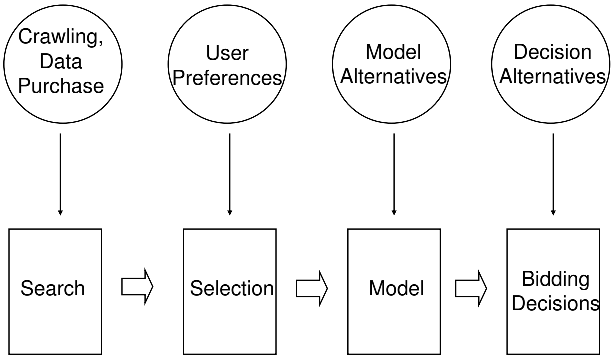


Figure 2.10: Process of automated bidding and alternatives.

2.5.4 Practical Considerations

It is important to understand that our automated bidding approach relies on a number of key ingredients. In this study, we assume that appropriate bidding records are available and we only focus on deriving a model from these bidding records and subsequently designing a bidding strategy around that model. Before deploying our automated bidding approach, one also needs to put in place methods for *searching* and *selecting* the right bidding records (see Figure 2.10). Finding suitable bidding records can be accomplished in several ways, e.g., using automated agents such as web crawlers (e.g. [9]) or by directly purchasing bidding data (from data vendors such as *Data Unison*). Having a pool of bidding records, the next challenge is to select, from this pool, the right set of *most relevant* bidding records. One could find this set via, e.g., a vector of desired product features (e.g., "iPod Nano, 8GB, yellow") and then selecting only those bidding records that are most similar to the feature vector. Deriving a suitable similarity metric can be done using, for example, the spatial feature model proposed in [53], or the comprehensive metric proposed in Chapter 4. Isolating product features from bidding records is made possible via eBay's effort of standardizing certain product descriptions (e.g., product descriptions for MP3 players require fields such as brand, product line, storage capacity, color or condition); additional information is often contained in the unstructured descriptive text which may take more effort to mine.

Related to the issue of search and selection is the issue of incorporating individual user preferences or risk tolerance into the bidding process. While some bidders may consider all relevant auctions as potential candidates, others may be more selective and wish to eliminate auctions based on certain constraints (e.g., eliminate auctions with seller ratings lower than a certain threshold, eliminate red iPods, etc.). This can again be accomplished in the selection step (see again Figure 2.10). In fact, when applying our method only to high-reputation sellers, the expected surplus increases to $20.05.

We also want to point out that in practice one would apply our method repeatedly until a consumer's demand is satisfied. We assume here that a consumer has demand for only a single unit (for more discussion on multiple units see the next section) and that s/he does not have any time constraints. Then, our automated strategy would place a bid while continuously monitoring the remaining market – which could be done at no extra cost for the bidder using automated agents. Once the outcome of the first bid is known, the strategy would then decide whether and (if the previous bid was unsuccessful) where to place the next bid, and so on. While a bidder could also decide to place more than one bid simultaneously, this runs the risk of winning two auctions which is undesirable in the case a single unit demand.

It is also important to note that the method proposed in this manuscript is modular in the sense that individual modules can be exchanged. For instance, one can replace the dynamic forecasting model by alternatives (such as regression models with different sets of variables, GAM, CART, or our K-Nearest Neighbor forecaster proposed in Chapter 4); similarly, one can replace the bidding decisions by an alternate set of rules. All in all, in order for the approach to be deployed, one will ultimately have to rely on agent-based technologies, similar to those currently in place for bid-sniping (e.g. `Cniper.com`). With such technology in place, our automated bidding strategy will not only yield real monetary benefits in terms of a higher expected surplus, but also less tangible benefits such as more convenience in terms of a truly automated bidding process.

2.6 Conclusions

The increasing popularity of online auctions puts more and more pressure on bidders to make informed bidding decisions in the face of competition. While classic bidding strategies such as early bidding or last-minute bidding are well-understood in the academic literature, they do not account for competition originating from simultaneous auctions selling same or similar items. Moreover, while it is unlikely that every bidder uses early or last-minute bidding in exactly the same way, to date they can only augment and adapt these strategies with gut-feeling, intuition or experience. We propose a novel automated and data-driven approach that provides bidders with valuable *objective* information about an auction's projected price in the face of competition.

Our approach consists of two main components. In the first component, we derive a novel dynamic forecasting model for price in competing auctions. We show that our model outperforms several competitor models. In the second component, we build a comprehensive bidding strategy around our forecasting model, using ideas from maximizing consumer surplus. We find that our strategy outperforms classical bidding strategies such as early bidding or last-minute bidding in terms of expected surplus accrued.

One important issue is the potential effect of a forecasting model on the market as a whole. If every bidder had access to the same model and bid on the same auction (with the lowest forecasted price), then forecasts, and as a consequence bidding decisions, would become unstable. This is very similar to the stock market where investment houses deploy complex math models to guide investment decisions. In such a scenario there is a risk that, if all investors base their decisions on the same model, the model – and not the investments' performance – could eventually drive the market. In this research, we are much less ambitious. While, at least in theory, one single model could eventually drive all bidding decisions on eBay, it is unlikely that it ever will. Rather, we view our automated bidding strategy, if ever deployed, as a decision tool that would be made available only to a few, select bidders and thus not destabilize the market.

There are several ways in which this research can be expanded. We have already pointed to the problem of selecting the optimal prediction window in Section 2.5.3.1. Another way to expand this research is via allowing for closing *and* continuing auctions. Recall that our current approach only consider auctions that close within the given prediction window. The reason is that our forecasting model is geared to the fixed time interval $(T + 1) - T$ so we can only predict the final price of auctions that end within that interval. Of course, one can roll the model one additional time period forward to make predictions at $T + 2$, based on the predicted values at $T+1$; however, predictions two time periods into the future (i.e. $T \rightarrow T+2$) are more uncertain than predictions only one step forward (i.e. $T \rightarrow T + 1$). It is not quite clear how to discount the additional prediction uncertainty in our decision framework. Another way to expand this research is via allowing for variable and adaptive WTP distributions. In our simulations, we assume that both early and last-minute bidders have the same WTP distribution. It may be possible that bidders with different strategies also have different product valuations. Moreover, we assume that the WTP distribution remains constant over our prediction window. While this may be realistic for short windows over only several hours, a bidder that wants an item immediately may have a different valuation compared to a bidder that is willing to wait several weeks. All-in-all, there are many opportunities for future research and we hope to inspire some of it with this study.

Chapter 3

Model Selection for Improved Forecasting

3.1 Introduction

People participate in online bidding day and night and from all over the world in a competitive fashion which sometimes results in price advantages for the consumer. However, given a choice between several hundred or thousand identical (or similar) options, all closing at different times, how can a consumer decide – in an efficient manner – which option results in the lowest possible price?

We accomplish this goal by developing a forecasting model for auction closing prices. Such models could alleviate the bidder's decision process by, e.g., ranking all available auctions by their lowest predicted price (see Chapter 2 for details). The bidder could then focus his or her bidding efforts only on those, say, K auctions with the lowest K predicted prices which greatly reduces the number of irrelevant choices and improves the efficiency of the search task.

One such model was developed in Chapter 2. To build that model, we first create many features, including static, time varying, and price-dynamics features for focal and competing auctions, to capture the many different pieces of information from the market that can affect the outcome of an auction, then conduct model selection to find our final model. However, the criterion based on which we selected the final model is rather intuitive.

Model selection is quite challenging in this setting because the forecasting goal is different compared to classical time series analysis. In classical time series analysis, one typically wants to forecast one particular *time point* of one particular series, such as sales at the end of the fourth quarter or the gasoline prices at the beginning of January 2009. This is different in the context of online auctions. In the auction context, one needs to forecast an entire *time window* of a stream of simultaneous auction processes. For example, a bidder discovers the need for a product, such as a Palm hand-held device on 4/14, and that s/he decides to purchase this item within the next 12 hours. There are many qualifying auctions available in this online market; some may close within the next hour, while others remain open for another 9 or 10 hours. Thus, the bidder needs to predict the outcome of each auction that closes within the next 12 hours. In other words, we need a forecasting model that not only predicts well at the beginning of the 12 hour time window or at its end, but *during its entire 12 hour duration.* Classical model selection criteria such as AIC or BIC optimize the model performance for only one time point, and are thus not directly applicable to our situation.

In this study, we investigate different approaches to overcome these challenges. We address the problem of model selection for a continuous time interval by computing an entire *distribution* of a model selection criterion rather than only a point value. In Chapter 2, we intuitively use the average BIC score over the time interval as the selection rule. We now investigate different ways to summarize this distribution for decision making and the impact of different distribution summaries on the prediction task. We find while the volatility of AIC or BIC's distribution over the prediction window results in extremely poor performance, their extremes work very well. We also find that both price dynamics and competition features play a crucial component in forecasting an online auction. This work is currently under review at the Journal of Business and Economic Statistics.

This chapter is organized as follows. In the next section, we briefly restate how to capture all potentially relevant information, both from within an auction as well as from simultaneously competing auctions as we did in Chapter 2. Section 3.3 proposes an idea to perform model selection with the goal of making good forecasts across a continuous time interval rather than only a single time point. We conduct empirical studies using the Palm data set (see Appendix A for data description) to compare the different approaches, both in terms of the different models they select as well as in terms of their actual predictive capabilities in Section 3.4. We conclude with further remarks in Section 3.5.

3.2 Create Features to Capture Important Information

Many different pieces of information potentially matter for the outcome of an ongoing auction. For example, we have discussed ways to capture dynamics and competition information for forecasting in Chapter 2. We now summarize features that are created to capture related information in order to forecast the outcome of an auction.

The outcome of an auction may be affected by what happens *within* that auction. We therefore create a set of features to capture the information from within the focal auction, including static information (such as condition of the product or the rating of the seller), time varying information (such as the number of bidders and time left), and price-dynamics (e.g. price-velocity and acceleration).

Besides what happens within an auction, the outcome of an auction may also be affected by what happens *outside*, that is, in simultaneous auctions that all sell the same (or similar) product and thus compete for the same bidders. For instance, the seller ratings, the current prices or the number of bidders in those simultaneous auctions could all affect the outcome of the focal auction. Therefore, we create another set of features to capture the information from competing auctions. To the end, we use the average of the features in concurrent auctions to capture the average market condition and use the standard deviation to capture the volatility of the market. For example, the price competition is given by the average and standard deviation of prices in concurrent auctions, and the average price-velocity in concurrent auctions would be defined as the average of the corresponding price-velocities, and so on.

A complete list of created features can be found in Table 2.2 in Chapter 2, and a more detailed description of creation of all features is described in section 2.3.1 - 2.3.2 in Chapter 2. It is clearly seen that the competition features can be categorized into *static competition features*, *time varying competition features*, and *price-dynamic competition features* by the information that they carry. Moreover, our created features incorporate both price dynamics and competition information which is necessary for accurately forecasting the extremely dynamic and competitive environment.

Table 2.2 in Chapter 2 shows that over 30 different variables are candidates for our forecasting model. Thus, an important first step in our modeling efforts will be the selection of a parsimonious subset of relevant variables.

3.3 Model Selection for Auctions Markets

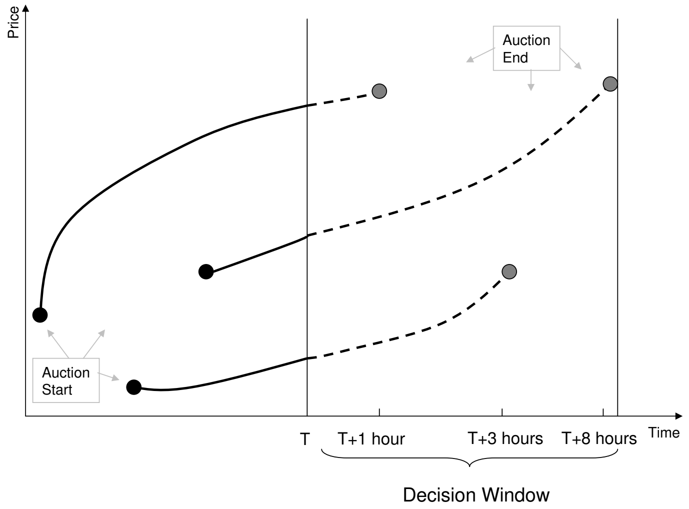


Figure 3.1: Illustration of the modeling task.

Our task is to find a model for the auction market. As pointed out earlier, the market consists of all auctions that sell the same (or similar) item (during a certain period of time). Take Figure 3.1 for illustration. In that market, we have 3 auctions selling the same item. What complicates the modeling task is that all auctions start (and hence end) at different times: one auction ends in the next hour, another one ends in 3 hours and the third auction ends in a little more than 8 hours. At time T, the bidder wants to make a decision on which of the three auctions to bid on. Since the bidder's decision window extends 8 hours into the future, the market model must consider all time points inside that window. The implication for model selection is that we need a model that works well not only at a single time point, but across the entire time window. Thus, our goal is to find a model that best describes price in this market, taking into account the effect of competition between auctions as well as differences in price dynamics across auctions.

Forecasting an entire time window is challenging since all statistical models are geared towards a single time point only. Take for illustration the (intentionally simple) time series model

$$y_{T+1} = y_T + \epsilon_T. \tag{3.1}$$

(The same argument would apply for more complex models also.) Model (3.1) implies that, given information up to time point T, we can forecast the response at time (T+1). However, model (3.1) also implies that we can *only* forecast the response at (T+1), and not at $(T + \frac{1}{2})$ or at $(T + \frac{2}{3})$. Thus, the best model selected (e.g. using model selection criteria such as AIC or BIC) is optimal only for time steps of length $\delta := (T + 1) - T$, and not for any time steps that are shorter (or longer). As we've argued above, such a model is not very meaningful for the eBay bidder!

We propose to investigate new model selection criteria that can overcome this challenge. Model selection has been researched in the statistics literature for a while [13] and it is receiving increasing attention today with the availability of more and more data sets featuring larger and larger number of variables [30]. Our goal is to find a model that, given a desired time T of decision making, predicts well *universally* across an entire time window, say, $T + \Delta$, where Δ could be as small as a few hours or as long as a few days, depending on the time frame within which the bidder wants to place a bid. We describe our idea next.

3.3.1 Model Selection for Time Windows

Classical model selection criteria, such as AIC or BIC, are geared towards models such as in equation (3.1) and thus only produce *pointwise* optimal results. We propose to generalize this idea to apply to entire time windows and thus to produce a *distribution* of model selection results.

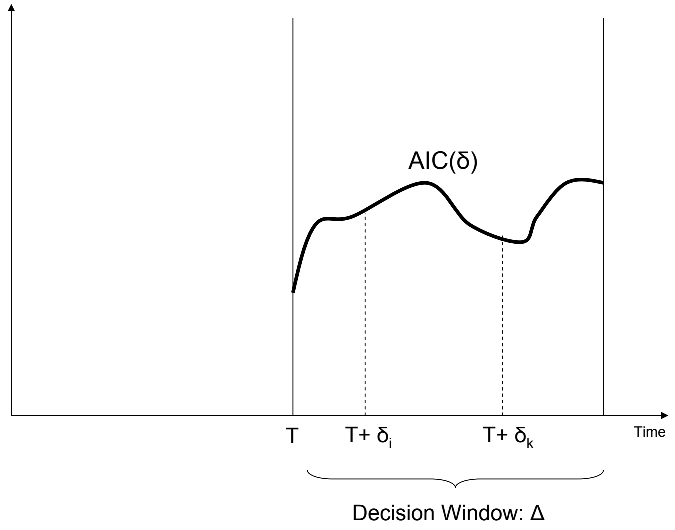


Figure 3.2: Distribution of model selection criterion.

The basic idea is illustrated in Figure 3.2. Suppose we have a model of the form

$$y_{T+\delta} = f(y_T, X_T) + \epsilon_T,\ \ \delta \in [0, \Delta], \tag{3.2}$$

where T denotes the time of decision making, and δ denotes the time increment which we would like to predict. (While $f()$ could denote any functional relationship between response and predictors, we consider linear models in our application.) Note that we let this time increment vary in the interval $[0, \Delta]$, where Δ corresponds to the length of our decision window.

For each time increment δ, we can compute a corresponding model selection criterion. Let us assume (for the moment) that we choose Akaike's information criterion (AIC) for decision making. (Later, we will also consider alternate model selection criteria such as the Bayesian Information criterion, BIC.) Then, for each $\delta \in [0, \Delta]$, we compute AIC(δ) and thus obtain a *distribution* of model selection values. This distribution is indicated by the solid black line in Figure 3.2. Note that we can only compute AIC(δ) for a training set, that is, for a set of data for which we know all values at T as well as at $T + \delta$. Later, we will apply the model to a holdout set, that is, a set of data for which we only know values at T and we wish to predict future values $T + \delta$.

Our objective is to select, among a set of candidate models, a model that performs well across all time increments $\delta \in [0, \Delta]$. In practice, there are two challenges associated with this objective. First, AIC(δ) is measured over the continuous interval $[0, \Delta]$. Since we cannot evaluate AIC(δ) over a continuous interval, we first select a fine grid $0 \leq \delta_1 < \delta_2 < \cdots < \delta_n \leq \Delta$ and then compute AIC(δ_i) for all

$i \in \{1, 2, \ldots, n\}$. The second challenge is that requiring a model to perform *uniformly* better than all other models does not lead to any results in our application. In other words, let $\mathrm{M}_k, k = 1, \ldots, K$, denote a set of candidate models and let $\mathrm{AIC}^{\mathrm{M}_k}(\delta_i)$ denote the k^{th} model's AIC value at time increment δ_i. Then, in our application, we may not find any model, say M_*, for which

$$\mathrm{AIC}^{\mathrm{M}_*}(\delta_i) \leq \mathrm{AIC}^{\mathrm{M}_k}(\delta_i), \ \forall i = 1, \ldots, n.$$

Thus, we resort to an approach where we do not require uniformly better performance, but rather performance that is better as measured by an appropriate summary statistic of $\mathrm{AIC}(\delta)$.

We consider several different summary statistics, such as the average, the extremes as well as the variance, to elicit a model that performs well across the entire interval $[0, \Delta]$. More specifically, we compute, for $i \in \{1, 2, \ldots, n\}$,

$$\begin{aligned} \mathrm{AIC}_{avg} &:= (1/N_i)\sum_{N_i}\{\mathrm{AIC}(\delta_i)\} \qquad (3.3)\\ \mathrm{AIC}_{med} &:= \mathrm{Median}_i\{\mathrm{AIC}(\delta_i)\}\\ \mathrm{AIC}_{min} &:= \mathrm{Min}_i\{\mathrm{AIC}(\delta_i)\}\\ \mathrm{AIC}_{max} &:= \mathrm{Max}_i\{\mathrm{AIC}(\delta_i)\}\\ \mathrm{AIC}_{sd} &:= \text{Standard Deviation}_i\{\mathrm{AIC}(\delta_i)\}\\ \mathrm{AIC}_{mean+sd} &:= \mathrm{AIC}_{avg} + \mathrm{AIC}_{sd} \end{aligned}$$

The rationale for investigating these 6 different summary model selection criteria is as follows. Both AIC$_{avg}$ [1]and AIC$_{med}$ consider a model's central performance and are as such natural candidate selection criteria. However, a model that is performing well on average may not be the best model for decision making. To that end, AIC$_{min}$ and AIC$_{max}$ consider a model's extreme performance. While AIC$_{min}$ points out a model's best performance, AIC$_{max}$ gauges the worst performance. In that sense, choosing AIC$_{max}$ for model selection finds the model that minimizes the worst loss – very similar in spirit to a minimax criterion. On the other hand, AIC$_{min}$ is the most optimistic selection criterion which identifies the time increment δ and associated model with the best performance. The next model selection criterion, AIC$_{sd}$, gauges the volatility of a model's predictive performance over the interval $[0, \Delta]$. The rationale is that models with less volatility will, in general, lead to more stable decision making. The last criterion, AIC$_{mean+sd}$, combines the effect of good average performance (i.e. small AIC$_{avg}$) and little volatility (i.e. small AIC$_{sd}$). Since we have no a-priori knowledge on which of these 6 model selection criteria will perform best, we will let the data speak.

3.3.2 Variable Pre-Selection and Multicollinearity

As we saw in Table 2.2, we have 33 different candidate variables for our forecasting model. Our goal is to select the subset of these 33 variables that results in the best-possible model. There are 8,589,934,591 different ways of selecting a subset of $m, 1 \leq m \leq 33$, variables, too many to enumerate manually. Moreover, in order

[1] Notice that compare with the definition for AIC$_{avg}$ in Chapter 2, all time points, including those where AIC (or BIC) is not available due to non-sufficient data for modeling, is now counted in this definition of AIC$_{avg}$.

to gauge a model selection criterion's entire distribution as in Figure 3.2, we need to evaluate each candidate model on a fine grid $0 \leq \delta_1 < \delta_2 < \cdots < \delta_n \leq \Delta$ which becomes computationally infeasible as n becomes larger. Our goal here is not to derive a computationally efficient algorithm for this task (which would indeed be an interesting challenge for future research), but rather to illustrate the performance of the different model selection summary criteria in Equation 3.3. Therefore, we first preselect a subset of *more realistic* candidate variables from the total of 33, and then perform exhaustive search on the remaining variables. We select this subset of more realistic candidate variables considering potential multicollinearity among predictors.

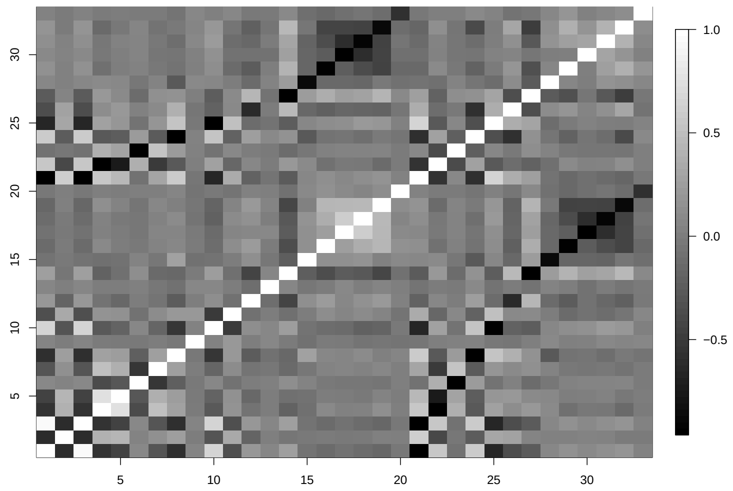


Figure 3.3: Pairwise mean correlation among the 33 variables. The x- and y-axes denote the variable index (between 1 and 33); the color corresponds to the strength (between -1 and 1) of the pairwise correlation.

To that end, we investigate the pairwise correlations of all 33 variables. Note that this correlation may change, depending on the density at which a variable is measured. Therefore, in similar fashion to above, we first evaluate each variable at different densities, $T + \delta_i$, $0 \leq \delta_i \leq \Delta$, and then compute pairwise correlations between pairs of variables at each density level. After that, we summarize the correlations in the same way as in Equation 3.3. Figures 3.3 and 3.4 show *image plots* of the corresponding pairwise *mean* correlations, *minimum* correlations and *maximum* correlations, respectively.

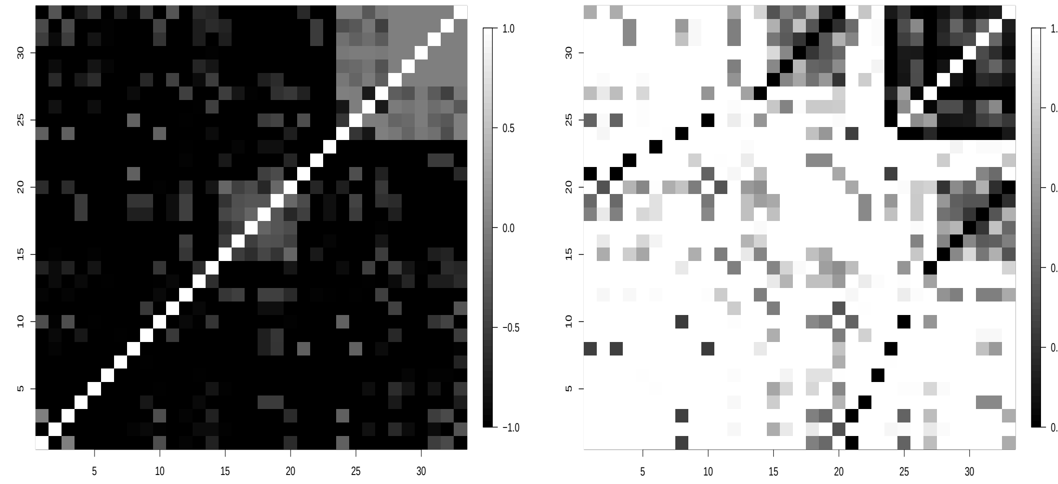


Figure 3.4: Pairwise minimum (left panel) and maximum (right panel) correlation among the 33 variables. The x- and y-axes denote the variable index (between 1 and 33); the color corresponds to the strength (between -1 and 1) of the pairwise correlation.

We can see that many of the variables are highly correlated. For instance, we find high collinearity between the current price and c.price.avg, the current price and c.price.vol, the current price and c.price.disc, the current price and time left, the current price and c.t.left.avg, the current price and number of bids, the current price and c.nbids.avg, price-velocity and c.vel.avg, price-velocity and c.vel.vol, and price-acceleration and c.acc.avg. This high collinearity is not surprising since many of these predictors carry similar information, only coded in a slightly different way. We eliminate all highly collinear predictors. In particular, for each pair of collinear predictors, we keep the variable that measures price or price-dynamics of the focal auction (i.e. current price, price-velocity or price-acceleration). While we could also keep alternate variables, this approach results in the most parsimonious model. In that fashion, we retain 6 variables for further analysis; these 6 variables are listed in Table 3.1.

3.4 Results

3.4.1 Model Selection

Table 3.1 lists the 6 candidate variables that we retain for further analysis including both information from within (price, price-velocity and -acceleration) and outside (volatility of acceleration, time left and number of bids of competing auctions). For ease of notation, we refer to each variable by a number. For instance, variable #1 refers to the price, variable #2 refers to the price velocity, and so on.

We now enumerate all possible models based on these 6 variables; that is, we evaluate a total of 63 different models. For each model, we evaluate the 6 different summary model selection criteria in equation (3.3), using $\Delta = 12$ hours and a grid of $0 \leq \delta_1 < \delta_2 < \cdots < \delta_n \leq \Delta$ of $n = 12$ different time increments. We use a grid-density of $\delta_i - \delta_{i-1} = 1$ hour since, for our data, on average 1 bid arrives every hour (during the last 12 hours of the auction). We do this for both AIC and BIC model selection criteria. Tables C.1-C.9 (Appendix C) list all the results.

Table 3.1: Variable Index

Var Number	Var Name
1	price
2	price-velocity
3	price-acceleration
4	c.acc.vol
5	c.tleft.vol
6	c.nbids.vol

Table 3.2: Best models according to BIC (top half) and AIC (bottom half). The number of model parameters is denoted by p. The bold cells correspond to the best model within each column.

p	BIC$_{avg}$	BIC$_{sd}$	BIC$_{med}$	BIC$_{min}$	BIC$_{max}$	BIC$_{mean+sd}$
1	1	2	1	1	1	1
2	1,2	2,3	1,2	1,2	1,2	1,2
3	1,2,3	2,5,6	1,2,3	***1,2,3***	1,2,3	1,2,3
4	1,2,3,5	2,3,5,6	1,2,3,5	1,2,3,6	***1,2,3,5***	1,2,3,5
5	***1-5***	***2-6***	***1-5***	1-4,6	1-3,5,6	***1-5***
6	1-6	1-6	1-6	1-6	1-6	1-6

p	AIC$_{avg}$	AIC$_{sd}$	AIC$_{med}$	AIC$_{min}$	AIC$_{max}$	AIC$_{mean+sd}$
1	1	2	1	1	1	1
2	1,2	2,3	1,2	1,2	1,2	1,2
3	1,2,3	2,5,6	1,2,3	1,2,3	1,2,3	1,2,3
4	1,2,3,5	2,3,5,6	1,2,3,5	***1,2,3,6***	***1,2,3,5***	1,2,3,5
5	1-5	***2-6***	1-5	1-4,6	1-3,5,6	1-5
6	***1-6***	1-6	***1-6***	1-6	1-6	***1-6***

Table 3.2 summarizes the results from Appendix C. In each cell, the table lists the best model for a corresponding combination of (selection criterion)× (number of model parameters (p)). For instance, the cell corresponding to $p = 1$ and BIC$_{avg}$ says that the best 1-parameter model picked by BIC$_{avg}$ is a model with only the price (see again Table 3.1 for the index of all variables). The highlighted cell within each column corresponds to the best model across all values of p (for a specific selection criterion). For instance, the highlighted cell ({1-5}) in the first column of the top half says that BIC$_{avg}$ picks a model with all variables (except c.nbids.vol) as the best model across all values of p.

We can make several observations in Table 3.2. First, we note that the cells in the top half are identical to the cells in the bottom half; this means that both AIC and BIC pick the same model (given a certain selection criterion and fixed value of p). However, we also note that the highlighted cells in the top and bottom tables are *not* identical; this implies that while for a given p, AIC and BIC result in the same model, they select different models across p. For instance, while BIC$_{avg}$ picks a 5-parameter model, AIC$_{avg}$ picks a 6-parameter model. (Similar for BIC$_{med}$ and BIC$_{mean+sd}$.) Only BIC$_{sd}$ and BIC$_{max}$ select the same models as their AIC counterparts.

Looking across rows, it is interesting to note that, for a given p, almost all selection criteria pick the same model – the main exception being BIC$_{sd}$ and AIC$_{sd}$. That is, while BIC$_{avg}$, BIC$_{med}$, BIC$_{min}$, BIC$_{max}$ and BIC$_{mean+sd}$ almost always agree on the same model (and similar for their AIC counterparts), BIC$_{sd}$ consistently disagrees. This seems to suggest that using the volatility of the model selection distribution measures a very different aspect of model performance compared to its center or its extremes.

In conclusion, we find that different selection criteria can result in different models, with the volatility of the model selection distribution showing the strongest deviations. However, we also want to point out that we have not yet determined the overall winner. The reason is that we have not yet determined which of the selected models actually results in the best forecasts. To that end, we investigate each model's predictive performance on a holdout set. We discuss this next.

3.4.2 Prediction Accuracy

We measure each model's predictive capabilities on a holdout set. To that end, we divide our data into a training set (80% of the data), and a validation set (remaining 20% of the data). Since our data varies over time (and since we are primarily interested in making accurate predictions of the future), our training set consists of all auctions that complete during the first 80% of our data's time span (i.e. between March 14 and May 10); the validation set contains all remaining auctions (i.e. between May 11 and May 25).

We measure predictive performance of a model in terms of its *mean absolute percentage error* (MAPE). For each $T + \delta_i$, $0 \leq \delta_i \leq \Delta$, we compute

$$\text{MAPE}_i = \frac{1}{m} \sum_{j=1} \frac{|y_{i,j} - \hat{y}_{i,j}|}{|y_{i,j}|} \tag{3.4}$$

where $y_{i,j}$ and $\hat{y}_{i,j}$ denote the true and predicted values of auction j at time increment

δ_i, respectively, and m denotes the number of auctions available. Figures 3.5 and 3.6 show the results.

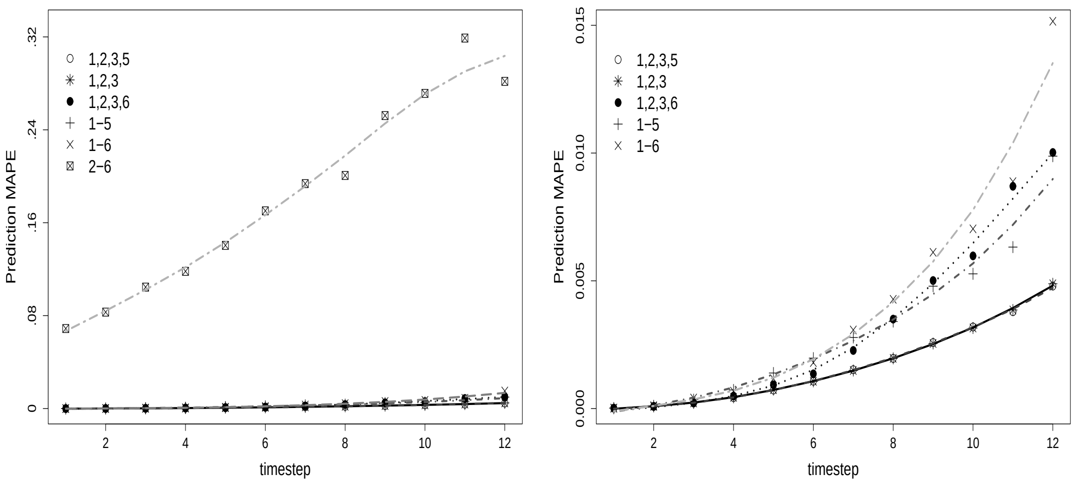


Figure 3.5: Prediction accuracy of the 6 top models. The left panel shows the prediction accuracy for all 6 models; the right panel shows a zoom-in on the 5 best models (leaving out the worst model of the right panel). The x-axis corresponds to the time increment δ_i, $0 \leq \delta_i \leq 12$, at which we make a prediction.

The left panel in Figure 3.5 shows the performance of the 6 best models. We can see that one model (with variables {2-6}) performs extremely poorly relative to the remaining 5 models. For that model, the prediction error is 8% for forecasting only one hour into the future; it increases to almost 32% for forecasting 12 hours into the future. Clearly, that model is not a candidate for the best predictive model; in order to get a better understanding of the remaining 5 models, we zoom-in (right panel of Figure 3.5). We can see that, of these 5 models, 2 (with variables {1,2,3,5} and {1,2,3}) have almost identical performance. We can further distinguish the performance of these 2 models in Figure 3.6. We can see that model {1,2,3,5} (slightly) outperforms model {1,2,3} for most time increments (especially for small,

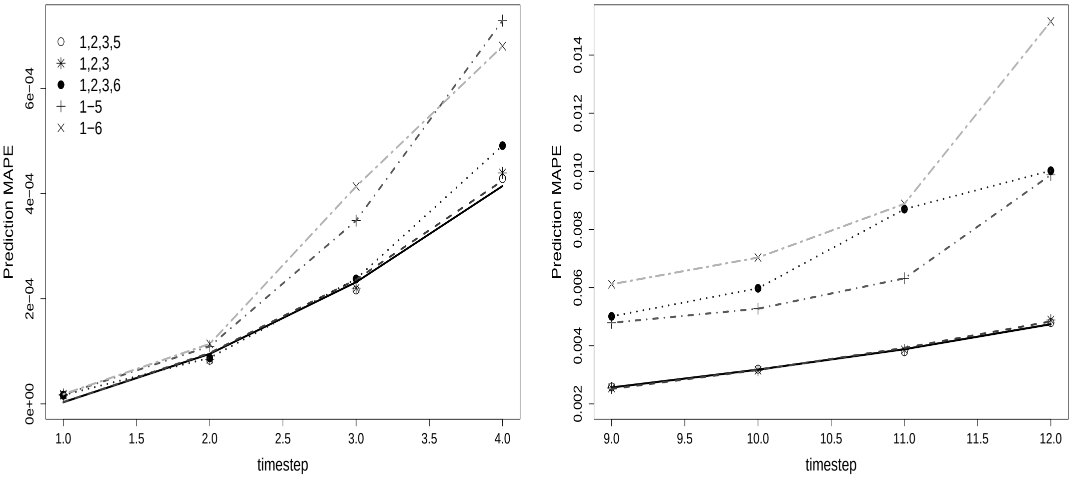


Figure 3.6: Prediction accuracy of the 5 best models from Figure 3.5. The left panel shows a zoom-in on the first 4 hours, the right panel shows a zoom-in on the last 3 hours.

e.g. 1 hour, and large, e.g. 12 hour, time increments). We comments further on the difference of these two models below.

3.4.3 The Winner

Let's recap: Section 3.4.1 illustrated the performance of each individual summary model selection criterion and we learned that while different model summarizes can point to different models, some select the same model. Section 3.4.2 investigated the performance of the top 6 models and we learned that while there is one clear loser, the two top models perform almost equally well. Now, we connect the analysis from Sections 3.4.1 and 3.4.2 and we reveal which summary model selection criterion leads to the best predictive model.

Table 3.3 shows the results. The fist column denotes the name of each model selection criterion and the second column denotes the best model it selects. (Note that since some selection criteria select the same model, we list more than one name in the first column). This information is taken from the model-selection analysis in Section 3.4.1. The remaining 3 columns refer to the prediction-based *ranking* of each model. Since the goal is to develop a model that predicts well in the short-term (i.e. for small time increments) as well as in for medium to longer time increments, we present 3 different rankings: one for predicting the next hour, one for predicting 6 hours into the future and one for predicting 12 hours into the future. The rankings are taken from the prediction-accuracy discussed in Section 3.4.2.

We can see that BIC_{max} and AIC_{max} dominate: both criteria find the (same) model (with variables $\{1,2,3,5\}$) that predicts best for short (1 hour) as well as long (12 hours) time increments. These two criteria are only outdone for the medium time increment (6 hours) for which BIC_{min} selects the best model (one which drops variable #5). It is interesting to see that while the maximum as a distribution summary performs equally well under both AIC and BIC, AIC_{min} performs significantly worse than BIC_{min} (and similar for the average, medium and medium + sd). Using the volatility as selection criterion (i.e. BIC_{sd} or AIC_{sd}) performs uniformly worst.

We can learn from Table 3.3 that the extremes as summaries of the model selection distribution result in the best performance. Both BIC_{max} and AIC_{max} select the best model; we pointed out earlier that choosing the maximum – just like a *minimax* criterion – protects against the worst loss, so this performance may not come too much as a surprise. It is more surprising though that BIC_{min} fares almost equally well; the minimum selects the most optimistic model across all time increments which seems to suggest that there is not too much heterogeneity across

Table 3.3: Selected Models and Prediction Accuracy

Selection Criteria	Selected Model	Prediction Rank 1 hour	6 hours	12 hours
BIC$_{max}$, AIC$_{max}$	1,2,3,5	1	2	1
BIC$_{min}$	1,2,3	2	1	2
AIC$_{min}$	1,2,3,6	3	3	4
BIC$_{avg}$,BIC$_{med}$,BIC$_{mean+sd}$	1-5	4	5	3
AIC$_{avg}$,AIC$_{med}$,AIC$_{mean+sd}$	1-6	5	4	5
BIC$_{sd}$,AIC$_{sd}$	2-6	6	6	6

different time increments (at least for our data). The performance of the central statistics (mean, median) is most surprising since, at least intuitively, one would expect good performance from a model that is selected according to average model quality. The poor performance of the volatility as a summary measure indicates that controlling the variability of a model's quality (around its mean) does not gain much in terms of predictive accuracy.

3.5 Conclusion

In this chapter we consider model selection when the goal is to find a forecasting model that works well across an entire range of time increments, producing an entire distribution of model selection criteria. We investigate different ways of decision-making based on that distribution find that the extremes lead to the most accurate forecasting models.

There are several avenues for future research. As pointed out earlier, efficient algorithms are necessary to perform model selection. While classical model selection can already be very computationally intensive, searching for models that work well over an entire distribution of time increments multiplies the computational burden. Moreover, it would be interesting to see if different ways for summarizing the distribution of a model selection criterion leads to better models, or if summaries can be combined in a more efficient way.

Chapter 4

Real-Time Forecasting of Online Auctions via Functional K-Nearest Neighbors

4.1 Introduction

Online auctions, such as those on eBay.com, have received a surge of popularity in recent years. This is in part due to their wide accessibility, their low participation barriers, and also due to the auction mechanism which engages its participants in stimulating competitive behavior. The popularity of online auctions has lead to a growth in related research and particularly in the desire to *predict* the outcome of an auction before its close. Knowing the auction's closing price has several advantages for auction participants. Bidders can use this information to make more informed (and perhaps even automated) bidding decisions [57]. Sellers can use predictions to identify times when the market is more favorable to sell their products and to better evaluate the value of their inventory.

Different approaches have been proposed to predict the price of an ongoing auction. We used regression-based models to forecast an auction's final price in a dynamic fashion in Chapter 2 and 3 (see also [32; 59; 96]). Common across these models is that they use information from a set of past auctions to predict an ongoing auction of interest. Moreover, for the purpose of model estimation, they weigh the information from each past auction equally. For instance, if the goal is to predict the price of a laptop auction based on a sample of historical auctions, then estimating a regression-type model will put equal weight on the information from a *Dell* laptop and from an *IBM* laptop – which may be inappropriate if the goal is to predict an auction for a *Sony* laptop. While some of the brand and product differences can be controlled using appropriate predictor variables, there might still be intrinsic differences that are hard to measure. An alternative to regression-based models which was proposed by [16] is a classification and regression tree. However, the authors point out that the prediction can be poor if prices in each final tree-node vary significantly. Moreover, while trees, unlike regression, manage to partition the data in a very flexible way, their predictions, like those of regression, are also based on the *un-weighted* information in each final node. In this chapter, we propose a novel and flexible approach for forecasting online auction prices based on the ideas of *K-Nearest Neighbors* (KNN). This work has been recommended for publication at *International Journal of Forecasting* with minor revision.

KNN is a forecasting approach that weighs the information from each record differently, depending on how similar that record is to the record of interest. For instance, if our goal is to predict the price of an auction for a *Sony* laptop, then it will put more weight on information from other *Sony* laptops and it will down-weight the information from, say, *Dell* or *IBM* laptops. More specifically, KNN predicts a record based on the weighted average of the K nearest neighbors of that record, where the weight is proportional to the proximity of the neighbor to the predicted record. KNN has been proven to converge to the true value for arbitrarily distributed samples [91; 23; 63], but studies show that its effectiveness is greatly affected by the choice of the number of neighbors (K) and the choice of distance metric [19; 34; 87; 62].

In the context of online auctions, the choice of the distance metric is challenging because auctions vary on many conceptually different dimensions. In particular, online auctions vary in terms of three types of information: *static, time-varying* and *dynamic* information. Static information comprises of information that does not change throughout the auction. This includes product characteristics (e.g., brand, product condition), or auction and seller characteristics (e.g., auction length, whether there is a secret reserve price, or whether the seller is a powerseller). Time-varying information updates itself during the auction (e.g., the number of bids or bidders). Both static and time-varying information have been shown to be important for forecasting the auction price because differences in product or bidding characteristics all influence bidders' decisions and hence the final price. Finally, auctions also vary in terms of their *dynamic* information. Dynamic information refers to the price path and its dynamics. These include the price-speed and the rate at which this speed changes throughout the auction. Auction dynamics are important for forecasting the final price because an auction that experiences fast price movements in the earlier stage will likely see a slow-down in price in later stages; conversely, auctions whose price travels very slowly at the beginning often see price-accelerations towards the end (e.g. [96; 54; 85]).

Auction price dynamics can be captured via functional objects such as curves. This means that bids are viewed as a discrete realization of an underlying smooth price path. Using smoothing methods (see [88]), this price path is recovered from the discrete observations and the smoothness of the resulting object allows gauging of dynamics via taking derivatives. In this chapter, we propose a novel *functional KNN* forecaster (*fKNN*), which combines functional and non-functional data, for forecasting price in online auctions.

One challenge with functional methods is the choice of smoother. Typical smoothers include penalized splines (p-splines) or monotone splines (see Section 5.2 in Chapter 5 for more details). However, while p-splines cannot guarantee the monotonic nature of the auction price growth, monotone splines can be computationally burdensome. An alternative is to use a flexible parametric approach that can capture different types of price growth patterns. [45] proposed a set of four parametric growth models for capturing price paths of online auctions (For details, see Section 5.2 in Chapter 5). In Section 4.3, we propose a parsimonious parametric form that generalizes these four growth models. Our parametric model has many appealing features such as monotonicity and computational efficiency. It is particularly important within the context of fKNN since it allows us to measure the distance between auctions' dynamics in a very parsimonious way via the Kullback-Leibler distance [12].

Our fKNN forecaster, which integrates information of various types, uses different distance metrics for each data-type. In Section 4.4 we discuss the different distance measures and how they are combined into a single distance metric.

We also discuss another important aspect of KNN forecasters, which is the choice of K. While choosing K too small eliminates important information, choosing

K too large results in noise that deteriorates forecast accuracy. The goal is to find a value that best balances signal to noise. [91] found that K can depend on the distribution of the data and that the optimal K often grows with the sample size. In this study, we investigate the optimal value of K as a function of different distance metrics as well as of data size and heterogeneity.

The chapter is organized as follows. In Section 4.2, we introduce the two sets of eBay data used in this study and discuss their level of heterogeneity. In Section 4.3, we discuss a flexible parametric model for capturing the price path in online auctions. Section 4.4 investigates the choice of the distance metric (combining distance metrics for static, time-varying and dynamic data) and the optimal choice of K. In Section 4.5, we describe the results of applying the *f-KNN* forecaster to the two datasets, and compare it to some competing approaches. We conclude and discuss possible extensions in Section 4.6.

4.2 Data

We use two datasets from the popular marketplace eBay. The datasets vary in terms of heterogeneity. The first dataset contains auctions that sell an identical product - Palm Pilot M515 PDA, while the second dataset contains auctions for various laptops. Each dataset is described briefly next and in further detail in Appendix A.

4.2.1 Palm PDA Auctions

Our first dataset includes the complete bidding records for 380 auctions that transacted on eBay between March and May, 2003. Each auction sold the same product, namely, a new Palm M515 handheld device. At the time of data collection, the market price of the product was about USD $230 (based on Amazon.com). Each bidding record includes the auction ID, the starting and closing times and prices, all bids with associated time stamps, and other information such as auction duration, shipping fee, seller's feedback score, whether the seller is a power seller, whether the product is from an eBay store, and whether the auction descriptions include a picture. All these variables contain information that can affect the final price of the auction. The complete summary statistics for these variables is presented in Table A.1 in Appendix A.

We now briefly describe what aspect of the auction process the individual variables measure and how they are related to the final price. The opening price is set by the seller and is known to influence the number of bidders the auction attracts. As for the final price, eBay uses second-price auctions where the winner is the highest bidder and s/he pays the second highest bid (plus an increment). Hence the final price is equal to the second highest bid plus an increment. Auctions can vary in their duration (between 3 and 10 days, in our data), with 7-day auctions being the default. In terms of auction competition, the average number of bids is 17.45 and the average number of bidders is almost 9. The average shipping fee, set by the seller, is $15.44. This fee is often perceived as a “hidden cost”. Another piece of relevant information is the seller's feedback score, which is approximately the number of transactions that the seller completed on eBay. A seller's feedback score often proxies for his/her credibility. In our data the highest seller rating is 27,652.

We can also learn from Table A.1 that over 87% of all auctions featured a picture. Pictures carry visual information about products, thus enhance bidders' confidence in the quality of the item. Power sellers are sellers with consistently high volumes of monthly sales, over 98% positive ratings, and PayPal accounts in good financial standing. We can see that 30% of sellers are power sellers. And lastly, sellers with feedback scores of 20 or higher, verified ID, and PayPal accounts in good financial standing are permitted to open "stores" on eBay. Stores provide easy management of accounts and improved brand boosting when the sellers have multiple items listed. In our data approximately 30% of all auctions are associated with an eBay store.

4.2.2 Laptop Auctions

While the Palm PDA dataset is very homogenous in terms of the product sold, the second dataset consists of auctions for a collection of laptops, featuring products of many different makes and models.

The data contain information on 4,965 laptop auctions that took place on eBay between May and June, 2004. Table A.2 in Appendix A summarizes the data. We can see that while some auction variables are similar to those of the Palm PDA data, others are different. For instance, Buy-It-Now auctions are listings that have the option of a fixed-price transaction and thus forego the auction mechanism. Over 20% of the laptop auctions included that feature. Moreover, a secret reserve price is a floor price below which the seller is not required to sell. This feature is particularly popular for high-value auctions. We can see that roughly 30% of all laptop auctions make use of the secret reserve price feature.

The main difference between the Palm PDA data and the laptop data is that the latter include products of a wide variety of makes and models. Table A.2 show that the data include over 7 different brands, and for each brand laptops differ further in terms of their memory size, screen size, processor speed, whether they are a new or used product, and whether or not they include an Intel chip or a DVD player. All-in-all, the products sold in these auctions are of a wide variety which is reflected in the wide range of closing prices (between $445 and $1,000).

4.3 A Functional Model for Capturing Price Growth Patterns

Our fKNN forecaster includes both functional and non-functional data. By functional data we mean a collection of continuous objects such as curves, shapes or images. Examples include measurements of individuals' behavior over time, digitized 2- or 3-dimensional images of the brain, or recordings of 3- or even 4-dimensional movements of objects traveling through space and time. In our context, we consider the price path of an online auction. Such data, although often recorded in discrete fashion, can be thought of as continuous objects represented by functional relationships. This gives rise to the field of functional data analysis [79].

Functional data consist of a collection of continuous objects. Despite their continuous nature, limitations to human perception and measurement capabilities allow us to observe these objects at discrete time points only. Thus, the first step in functional data analysis is to recover, from the observed data, the underlying continuous functional object. This is usually done with the help of data smoothing. Typical data smoothers include penalized splines or monotone splines [88]. In this chapter, we suggest a novel approach to recover the functional objects via a Beta model. The main advantage of the Beta model is that it allows us to measure distances between two functional objects via the Kullback-Leibler distance. In contrast to penalized splines, it guarantees monotonicity of the resulting functional object, which is important for modeling monotonic price growth behavior in auctions. Compared to monotone splines (which also result in monotonic representations), the Beta model is computationally much more efficient[1]. Recently, [45] proposed a family of four growth models for representing auction price paths. Our approach via the Beta model generalizes this idea and includes the four growth models as special cases.

4.3.1 The Beta Model

We model an auction's price path using the Beta cumulative distribution function (CDF). The Beta distribution is a continuous probability distribution defined on the interval $[0, 1]$ with two shape parameters, α and β, that fully determine the

[1] We use the popular R function smooth.monotone in the fda package. An alternative is to use the pcls function (and accompanying functions gam, smoothCon, and mono.con) in the mgcv package, which is computationally more efficient, but from our experience it produces inferior fits.

distribution. Its CDF can be written as

$$\mathrm{F}(x, \alpha, \beta) = \frac{\int_0^x u^{\alpha-1}(1-u)^{\beta-1}du}{B(\alpha, \beta)} \tag{4.1}$$

where $B(\alpha, \beta)$ is the *beta* function[2] ([1]), a normalization constant in the CDF to ensure that $F(1, \alpha, \beta)$ equals to unity.

We model auction price paths with the Beta CDF in the following way. Let p denote the sequence of observed prices with associated time-stamps t. Since auctions can be of varying durations, we normalize the time sequence by $t_n = t/Duration$, which yields time-stamps between 0 and 1. Similarly, auctions close at different prices, so we normalize the observed prices by $p_n = p/ClosingPrice$ which yields values of p_n between 0 and 1. The goal is then to find the values of α and β that satisfy $p_n = \int_0^{t_n} u^{\alpha-1}(1-u)^{\beta-1}du/B(\alpha, \beta)$ for every element of p_n and t_n.

In the context of real-time forecasting, we only observe price paths up to some time T (with associated price P). We therefore estimate α and β by normalizing the time and price scales to $[0, T]$ and $[0, P]$, respectively (i.e. $t_n = t/T$ and $p_n = p/P$). Estimation is done by error minimization (The algorithm for efficiently fitting the Beta model to auction data is described in detail in Section 5.3.1 in Chapter 5).

Figure 4.1 shows typical paths produced by the Beta model for different values of α and β. The solid black line represents the case of rapid price growth at the beginning and at the end, but only little growth during the middle; this case would be representative of auctions with intense early and last-moment bidding, but only

[2] $B(\alpha, \beta) = \int_0^1 u^{\alpha-1}(1-u)^{\beta-1}du$

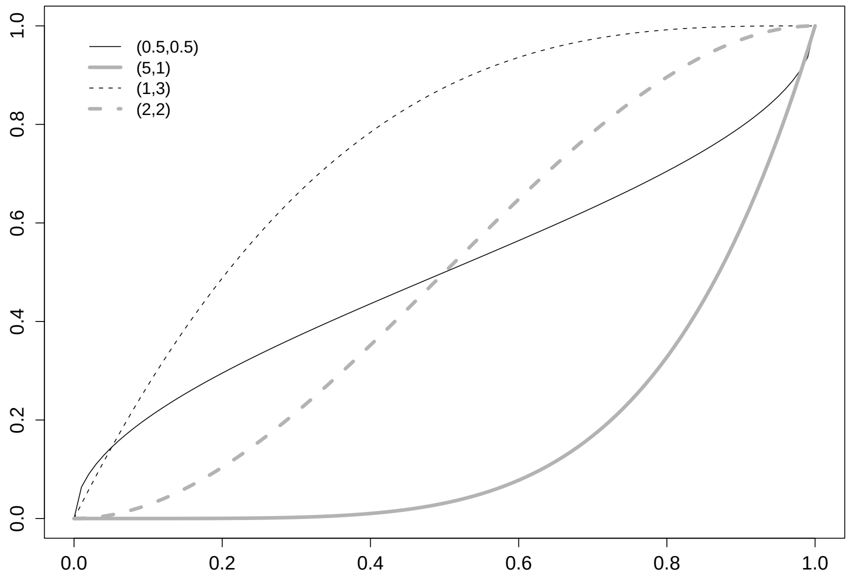


Figure 4.1: Typical price paths based on the Beta CDF with varying shape parameters (α, β)

little bidding activity in between – a case that is pretty common on eBay. The solid gray represents auctions that experience little bidding activity during most of the auction with bidding picking up only towards the end. In contrast, the dotted black line corresponds to auctions with high early activity which levels off as the auction progresses. And lastly, the dashed gray line corresponds to auctions where most of the bidding occurs during the middle part (and not at the beginning or the end), a case that, while rather uncommon, occurs from time to time on eBay.

One important consequence of the Beta model is its closed form of representation of price dynamics and acceleration by the first and second order derivatives of the *price*$\sim$*time* model. Besides this, there are other nice properties of the beta model which make it advantageous over existing models. We will explore those properties in Chapter 5.

4.3.2 Model Estimation

We estimate the Beta model for our auction data in a way that optimizes fit in both the x and y directions. In the auction context, the x direction corresponds to time and a good fit in that direction is necessary in order to accurately capture points of different bidding activity (e.g., early or last-minute bidding). We also require our model to fit well in the y-direction, which corresponds to price. A good fit in y-direction will guarantee accurate forecasts of an auction's final price which is the main goal of this study.

Since we fit the model in both x and y directions, we measure goodness of fit by examining the residual error in both directions. For the i^{th} auction with n bids, we define the residual as

$$\text{Resid}_i = \frac{1}{n}\sum_{k=1}^{n}\left[0.5(y_k - \hat{y}_k)^2 + 0.5(x_k - \hat{x}_k)^2\right],$$

which is the average of the sum of squared errors in both x and y directions. Note that the smaller the residual error, the better our model represents the auction price-path.

Figure 4.2 illustrates the model fit for the Palm PDA data. The left panel shows the distribution of residuals for the Beta model; the other two panels show the corresponding distributions for the growth models [45] and penalized splines, respectively. We can see that the Beta model results in the best model fit, i.e. in the smallest residual error[3]. The results are very similar for the laptop auctions.

[3] We conduct a more complete comparison of all smoothing models in Chapter 5.

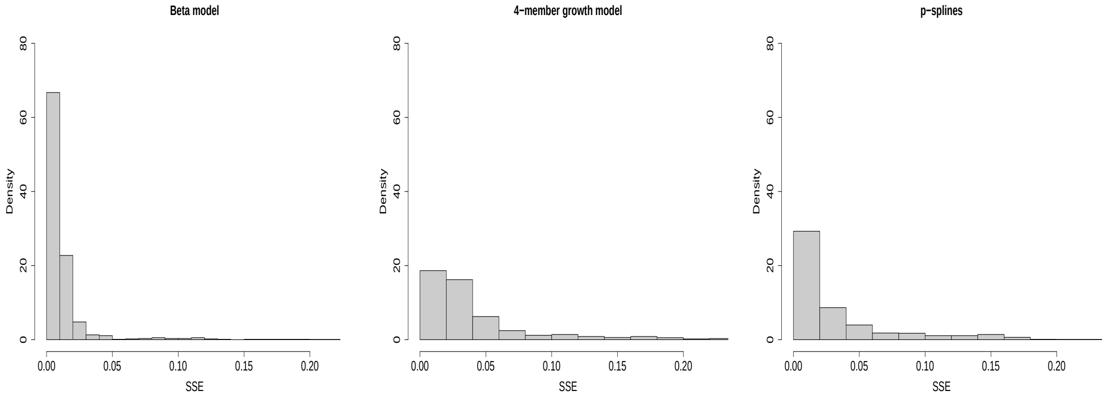


Figure 4.2: Residual comparison for fitting three models: Beta model (left), growth model (middle), and p-splines (right).

4.3.3 Kullback-Leibler Distance

Since the fKNN forecaster uses both functional and non-functional data, we must define distance measures for both data types. While there exist standard measures for the distance between non-functional data (e.g., Euclidian distance), measuring the distance between functional data (e.g., between two curves) is more involved because of infinite dimensionality. One of the main advantages of the Beta model is that it allows us to measure the distance between two auction price paths in a very parsimonious way via the Kullback-Leibler (KL) distance.

The KL distance [68] is a non-commutative measure of the difference between two probability distributions. For two distributions X and Y, it measures how Y differs from X. The KL distance is widely used in the field of pattern recognition for feature selection (e.g. [12]) or in physics for determining the states of atoms or other particles (e.g. [75]). In our case, X and Y both refer to the Beta distribution with parameters α, β, and α', β', respectively. The KL distance between X and Y

is then given by a very simple function of the Beta parameters [81]:

$$D_{\mathrm{KL}}(X,Y) = \ln \frac{B(\alpha',\beta')}{B(\alpha,\beta)} - (\alpha'-\alpha)\psi(\alpha) - (\beta'-\beta)\psi(\beta) + (\alpha'-\alpha+\beta'-\beta)\psi(\alpha+\beta), \tag{4.2}$$

where B and ψ denote the *Beta* and *Digamma* function, respectively ([1]).

Returning to the four auctions in Figure 4.1, consider the solid black line ($Beta(0.5, 0.5)$) as the focal auction that we want to forecast. Using equation (4.2), the KL distance to the focal auction is 9.69 from the solid gray line ($Beta(5, 1)$), 6.40 from the dotted black line ($Beta(1, 3)$), and 7.10 from the dashed gray line($Beta(2, 2)$). While the dashed gray line may, at least visually, not appear very distant from the focal auction, its distribution is in fact very different, as captured by the KL distance.

4.4 Functional K-Nearest Neighbors (fKNN)

In this section we discuss the components of our functional KNN forecaster. We start by explaining the basic forecasting idea and then discuss the two main elements of our fKNN implementation: the choice of a suitable distance metric and the choice of K.

4.4.1 Overview

Our goal is to predict the final price of an ongoing auction. Consider Figure 4.3. The solid line corresponds to the price-process of an auction that is observed

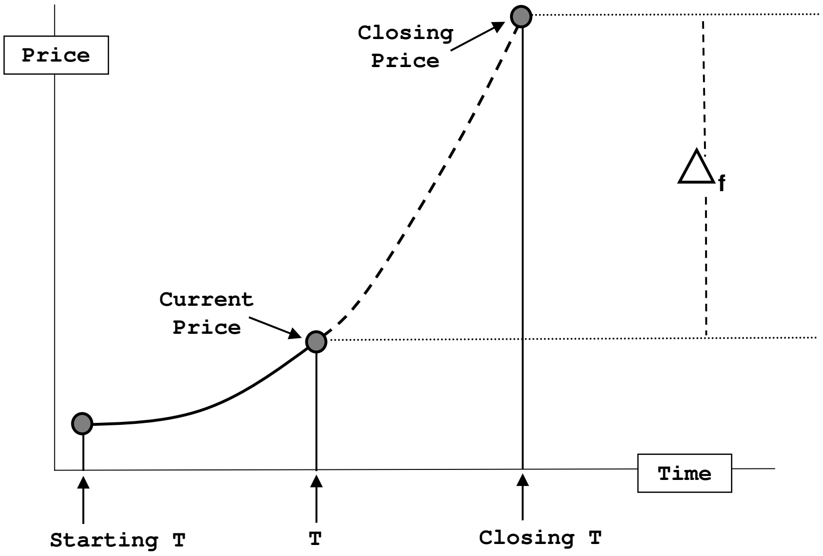


Figure 4.3: Illustration of the forecasting idea.

until time T. The dotted line corresponds to the (future) price path until the close of the auction. Our goal is to predict the closing price. As the closing price is determined by the current price plus the price-increment $\triangle_f$, our forecasting problem is equivalent to predicting $\triangle_f$. We will therefore use fKNN to estimate $\triangle_f$ based on a training set of completed auctions.

In order to estimate $\triangle_f$, we look for the K most similar auctions in the training set. Consider Figure 4.4 for illustration. In that scenario, we have a training set with 6 auctions, $\triangle_1$ – $\triangle_6$. We also have associated distances, D_1–D_6, between the focal auction and each of the auctions in the training set. If K equals 3, then we will estimate $\triangle_f$ by the weighted average of the 3 nearest auctions, in this case by

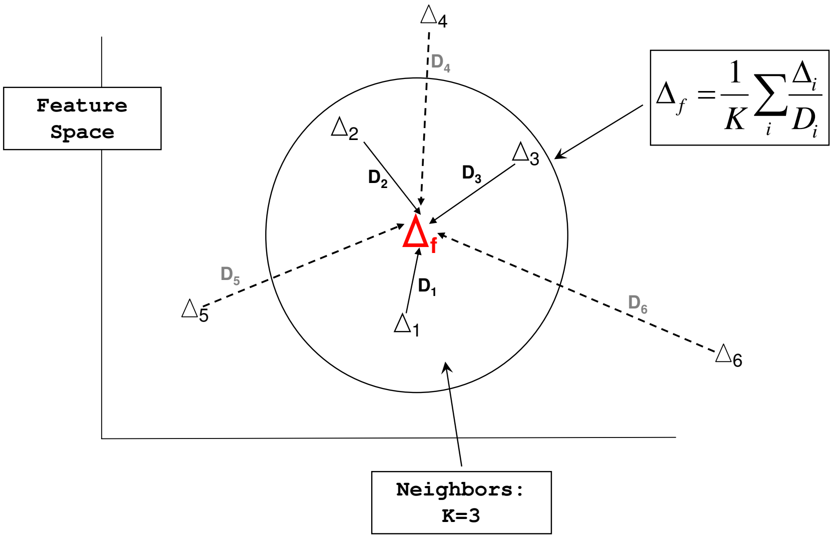


Figure 4.4: Illustration of forecasting scheme.

the weighted average of $\triangle_1 - \triangle_3$. More generally, we estimate $\triangle_f$ as

$$\triangle_f = \frac{\sum_{i=1}^{K} \triangle_i / D_i}{\sum_{i=1}^{K} 1/D_i}. \tag{4.3}$$

As we can see in equation (4.3), the two main elements of this approach are the choice of K and the choice of a distance metric D. We discuss these next.

4.4.2 Choice of Distance Metric

As pointed out earlier, online auction data comprise of three types of information: Static information captures information that does not change during the course of the auction, time-varying information that changes during the auction, and auction dynamics, which are captured and represented by functional data. Table

4.1 summarizes the three types and the specific variables for each data type. We now discuss distance metrics for both data-types.

Table 4.1: Summary of information sources characterizing online auctions

Data Type		Measurement Scale	Example
Non-functional	Static	Interval	opening price, screen size, process speed
		Binary	buy-it-now, reserve price, condition
		Categorical	brand
	Time-Varying	Interval	number of bids, current price
Functional		Functional	price-velocity, -acceleration

4.4.2.1 Static and Time-Varying Data

Static and time-varying information includes data measured on different scales (interval, binary and categorical). Following [53], we use separate metrics for each individual scale, and then combine the individual metrics into an overall distance metric for non-functional data.

For binary data x_B and x'_B (e.g., an auction with the buy-it-now option vs. an auction without that feature), we define the distance as

$$d^B = \mathbf{1}(x_B \neq x'_B), \tag{4.4}$$

where $\mathbf{1}$ denotes the indicator function and thus d^B equals 1 if and only if $x_B \neq x'_B$; otherwise it is 0.

We adopt a similar measure for categorical data. For instance, the categorical variable "brand" can assume 8 different levels (Dell, Fujitsu, Gateway, etc) which can be coded as a vector of 7 different binary variables. Thus, each categorical variable can be represented as a set of binary variables. Let $\mathbf{x}_C$ and $\mathbf{x}'_C$ denote two vectors representing categorical data, then we define their distance, similar to equation (4.4), as

$$d^C = \mathbf{1}(\mathbf{x}_C \neq \mathbf{x}'_C), \tag{4.5}$$

which takes the value of 1 if and only if $x_C \neq x'_C$, and 0 otherwise.

For interval-scaled data x_I and x'_I (e.g. two auctions with different opening prices), we use a scaled version of the Minkowski metric [47]:

$$d^I = \frac{|\tilde{x}_I - \tilde{x}'_I|}{\tilde{R}_I}, \tag{4.6}$$

where $\tilde{x}$ denotes the standardized value of x, and $\tilde{R}$ denotes the range of $\tilde{x}$. The advantage of the Minkowski metric is that it renders interval-scaled data onto the interval $[0, 1]$. Note that the maximum and minimum values of d^I are 1 and 0 respectively, which are also the values taken by the binary and categorical distance metrics in equations (4.4) and (4.5). Having metrics in comparable magnitudes makes it easier to combine individual distance metrics.

We combine individual distance metrics in the following way. Let $\mathbf{x} = \{x_1, x_2, ..., x_p\}$ be a vector of p non-functional features, including binary, categorical and interval data. We compute the overall distance between $\mathbf{x}$ and $\mathbf{x}'$ as

$$d(\mathbf{x}, \mathbf{x}') = \frac{1}{p} \sum_{i=1}^{p} d^*, \tag{4.7}$$

where d^* denotes the appropriate individual distance metric from equations (4.4)-(4.6).

As an example, let x and x' be two three-feature vectors. Specifically, $x =$\{w/ buy-it-now, dell, 1G memory\} and $x' =$\{w/o buy-it-now, IBM, 1G memory\}. The first, second and third features are binary, categorical, and interval scaled, respectively. Using equation (4.7), $d(x, x') = 1/3(d1 + d2 + d3)$, where $d1 = 1$ based on equation (4.4), $d2 = 1$ based on equation (4.5), and $d3 = 0$ based on equation (4.6). The overall distance between x and x' is therefore 2/3.

Note that the definition of d in (4.7) is flexible in the sense that one can use only subsets of the available information. For instance, d^{Static} would refer to the distance metric using only static information, while $d^{Time-Varying}$ would refer to the metric with only time-varying information. One problem with distance metrics of this type is that they may over-weigh different sources of information, depending on how elaborately each source is recorded. For instance, a data set with 100 different static features and only 10 time-varying features puts 10-times more weight on the information from static features. In order to overcome this potential bias, we follow the ideas of [14] and first scale each individual distance metric by its mean root square (MRS). MRS is a statistical measure of the magnitude of a vector. For a vector $x = \{x_1, ..., x_p\}$, MRS is defined as $\sqrt{\frac{1}{p}\sum_{i=1}^{p} x_i^2}$ [70]. We apply the same scaling to each individual distance metric and obtain

$$
\begin{aligned}
d_s^{Static} &= d^{Static}/\mathrm{MRS}(d^{Static}) && (4.8)\\
d_s^{Time-Varying} &= d^{Time-Varying}/\mathrm{MRS}(d^{Time-Varying}) && (4.9)\\
d_s^{Static\&Time-Varying} &= d_s^{Static} + d_s^{Time-Varying}. && (4.10)
\end{aligned}
$$

Note that the combined metric $d_s^{Static\&Time-Varying}$ now puts equal weight on both static and time-varying information.

4.4.2.2 Dynamics (Functional Data)

As shown in Section 4.3.3, we can measure the distance between two functional observations using the KL distance. Let (α, β) and (α', β') denote the Beta parameters for two different auction price paths, then their distance (when x is the focal auction) is defined as

$$d^F = |D_{\mathrm{KL}}(x, x')|, \qquad (4.11)$$

where $D_{\mathrm{KL}}(x, x')$ is defined in equation (4.2).

Note that d^F ranges within $[0, +\infty)$ as the KL distance assumes values on the real line. In order to make d^F comparable with the non-functional distance measures, we again scale it using the MRS transformation. Thus we obtain

$$d_s^{Dynamics} = d^F/\mathrm{MRS}(d^F). \qquad (4.12)$$

4.4.2.3 Optimal Distance Metric

To determine which combination of individual distance metrics leads to the best forecasting model, we investigate a series of different distance metrics. In particular, we investigate performance using five different metrics: d_s^{Static}, $d_s^{Time-Varying}$, $d_s^{Dynamics}$, $d_s^{Static\&Time-Varying}$ or d_s^{All}, where $d_s^{All} = d_s^{Static\&Time-Varying} + d_s^{Dynamics}$. We first determine the optimal metric based on a validation set and then investigate the predictive accuracy of the resulting fKNN forecaster on a test set.

4.4.3 Choice of K

The second important component of fKNN is the choice of K, the number of neighbors from which the forecasting is calculated. Too small a value will filter out relevant neighbors; too big a value will introduce noise and weaken the prediction.

[91] finds that the optimal value of K is data-dependent, and it usually grows with the sample size. In additional, K may also vary as different distance metrics are used. Therefore, we select the optimal value of K separately for each distance metric and each data set. To do so, we again select the best value of K based on a validation set; we then apply the resulting model to the test set.

4.4.4 Forecasting Scheme

Our complete forecasting process includes determining the optimal distance metric and the optimal value of K. We determine both based on a validation set. Then, using the optimal metric and K, we estimate the fKNN model based on the records in the training set. We investigate the performance of that model by predicting a new focal auction using auctions from a test set.

4.4.5 Comparison With Alternate Methods

We benchmark our fKNN forecaster against two other very popular prediction methods: parametric regression models, and nonparametric regression trees (CART).

In a linear regression model, the closing price is modeled as a linear function of the observed predictor information. This information can include some or all of the three types of data from Table (4.1). Note that in such models, all auctions from the training set are weighed equally when estimating the model coefficients.

CART forecasting takes a hierarchical approach. It recursively partitions the data into smaller sub-groups; the focal auction is then forecasted based on the average of the most relevant sub-group. While CART, like KNN, uses neighboring information from similar auctions, it weighs each auction equally, which is one major aspect in which it differs from KNN.

We discuss differences in prediction performance next.

4.5 Results

We now discuss the predictive performance of our functional KNN forecaster when applied to the two datasets of eBay auctions, and compare it with competing approaches. We also investigate the optimal distance metric and the optimal value of K. The two datasets, Palm PDAs and laptops, are different in their level of heterogeneity. While the Palm PDA dataset is very homogeneous, the laptop data are very heterogeneous. We also investigate different time horizons, that is, we investigate forecasting different distances into the future.

We split each of the two datasets into a training set (50% of the auctions), a validation set (25%) and a test set (25%). We split the data according to the temporal nature of our prediction task. That is, auctions in the training set transact prior to those in the validation set; and auctions in the test set transact after those in the validation set. Therefore, our experiments mimic the prediction task that real bidders face.

For the competing models (regression and CART), we train the models on the combined training and validation set, and then test their predictive performance on the test set.

We evaluate all models using the *mean absolute percentage error* (MAPE)

$$\text{MAPE} := \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \widehat{y}_i}{y_i}\right|, \tag{4.13}$$

where y_i and $\widehat{y_i}$ denote the true and estimated final price in auction i, respectively.

4.5.1 Selecting the Optimal K and the Optimal Distance Metric

We select the optimal value of K in the following way. Recall that we have 5 candidate metrics, $D \in \{d_s^{Static}, d_s^{Time-Varying}, d_s^{Dynamics}, d_s^{Static\&Time-Varying}, d_s^{All}\}$. For each metric, we select a value of K from the set $K \in \{1, 2, \ldots, 100\}$. For each combination of $(D \times K)$, we estimate the corresponding fKNN model on the training set, then measure its predictive accuracy (in terms of MAPE) on the validation set. Figure 4.5 shows the results. The left panel shows the results for the laptop auctions; the right panel shows the corresponding Palm PDA results. The top panel shows an overview, the bottom panel zooms-in on the most relevant part.

From the left panel in Figure 4.5 (laptop auctions) we can see that $d_s^{Dynamics}$ results in the worst model performance, regardless of the value of K. In other words, using only the dynamic information of the price path is not sufficient for achieving good prediction accuracy. We also see that, of the remaining 4 distance metrics, d_s^{All} yields the uniformly lowest prediction error. This suggests that for laptop auctions, due to their diversity in makes and models, every single piece of auction information is necessary to achieve good prediction accuracy. Moreover, we note that for d_s^{All}, the lowest prediction error is achieved at $K = 41$. We conclude that $D = d_s^{All}$ together with $K = 41$ results in the best predictions. The story is somewhat different for the Palm PDA data (right panel in Figure 4.5). For those data, $D = d_s^{Time-Varying}$ results in the uniformly lowest error (across all distances). Moreover, choosing K=2 optimizes that distance.

It is interesting that the two different data sets result in very different choices for K and D. While for the laptop data, we need all auction information (using the distance metric d_s^{All}) and a very large neighborhood (via K=41), the Palm PDA auctions require only the time-varying information of the auction process (using $D = d_s^{Time-Varying}$) and a very small neighborhood (via K=2). One possible explanation lies in the difference in heterogeneity between the two data sets. In the homogeneous

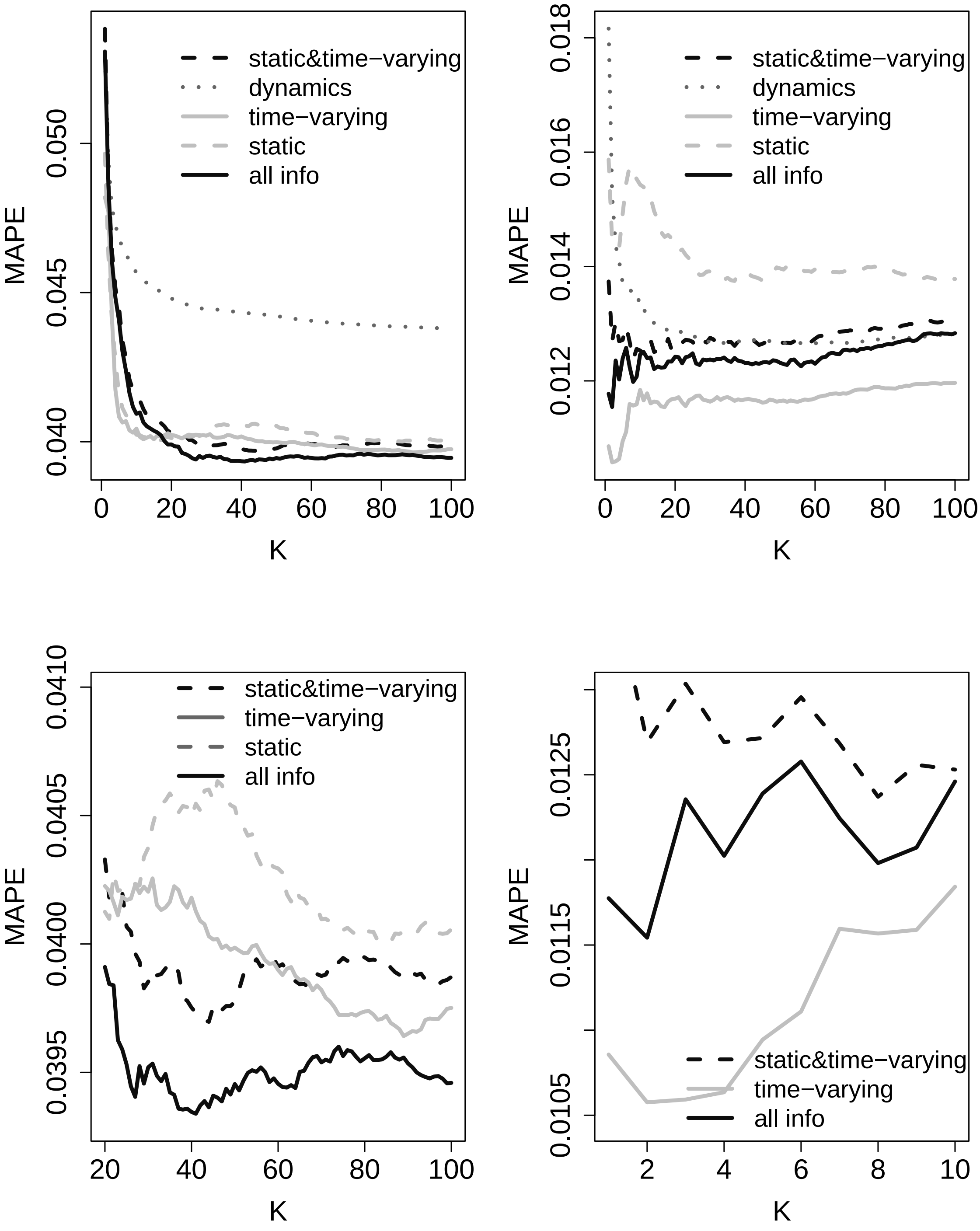


Figure 4.5: Optimal values of K and D. The left panel shows the results for the laptop auctions; the right panel shows the corresponding Palm PDA results. The top panel gives an overview, the bottom panel zooms-in on the most relevant part.

data (Palm PDA), all products are the same and differences in auction outcome will be mostly due to differences in the current price and the level of competition for that product. The competition level is reflected in the number of bids and bidders, which, together with the price level, are captured in $d_s^{Time-Varying}$. Moreover, since products are very homogeneous, we only need a very small neighborhood, thus K=2. This is different for the laptop auctions. In that data set, products are very heterogeneous, thus the forecaster needs all available information (in d_s^{All}) to distinguish between more relevant samples. Since the products are very different, the method also requires a larger neighborhood which leads to a larger value of K. This suggests that, as expected, forecasting more heterogeneous auctions is a more difficult task.

4.5.2 Robustness of Optimal D and K to the Time Horizon

In the previous section, we investigated the interplay of K and D for a fixed time horizon of 1 minute. That is, we assumed that we observe the auction until 1 minute before its close. We now investigate the robustness of this choice for different time horizons. Specifically, we investigate the robustness of K and D for different time horizons (δ_T) in the set $\delta_T \in \{\text{ 2h, 1h, 30 min, 15 min, 5 min, 1 min}\}$.

4.5.2.1 Robustness of Optimal K

Figure 4.6 investigates the robustness of K to the choice of δ_T. For a given value of K ($K \in \{20, 40, 60, 80, 100\}$ for the laptop data and $K \in \{2, 5, 10, 50, 100\}$

for the Palm PDA data), we investigate the prediction accuracy for different values of δ_T. We hold D fixed at $D = d_s^{All}$ for the laptop data and $D = d_s^{Time-Varying}$ for the Palm data. Figure 4.6 shows the *relative* prediction error $\text{Rel.MAPE}_K := \text{MAPE}_K / \text{MAPE}_{K^*}$, relative to a benchmark value ($K^* = 40$ for the laptop data, $K^* = 5$ for the Palm PDA data).

We see that for the laptop data (top panel in Figure 4.6), lower values of K (K=20) lead to poor performance. We also see that while K=40 generally leads to good forecasting accuracy, it is outperformed by higher K-values when forecasting time horizons of 30 or 15 minutes. This suggests that the value of K is not very robust to the time horizon. It is even less robust for the Palm PDA data (bottom panel in Figure 4.6), where K=5 leads to good forecasting performance only for very long time horizons (δ_T = 2h); in contrast, choosing K=2 leads to the best performance for very short horizons (δ_T = 1 min). This suggests that the choice of K should be a function of δ_T. Table 4.2 lists the optimal value of K for each combination of δ_T and D.

4.5.2.2 Robustness of Optimal D

We now investigate the impact of the time horizon δ_T on the choice of the distance metric D. Figure 4.7 shows the prediction accuracy as a function of the time horizon δ_T for different choices of D. Note that for each combination of D and δ_T, we use the optimal values of K from Table 4.2.

The left panel in Figure 4.7 corresponds to the laptop data; the right panel is

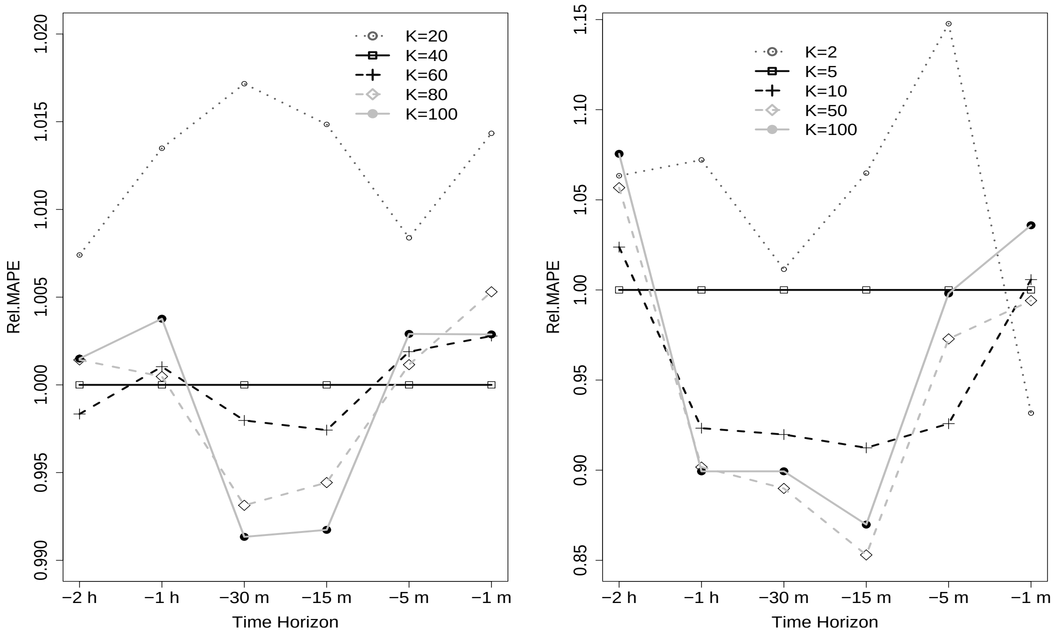


Figure 4.6: Relative prediction accuracy for different values of K at different time horizons δ_T. The left panel corresponds to the laptop data; and the right panel corresponds to the Palm PDA data.

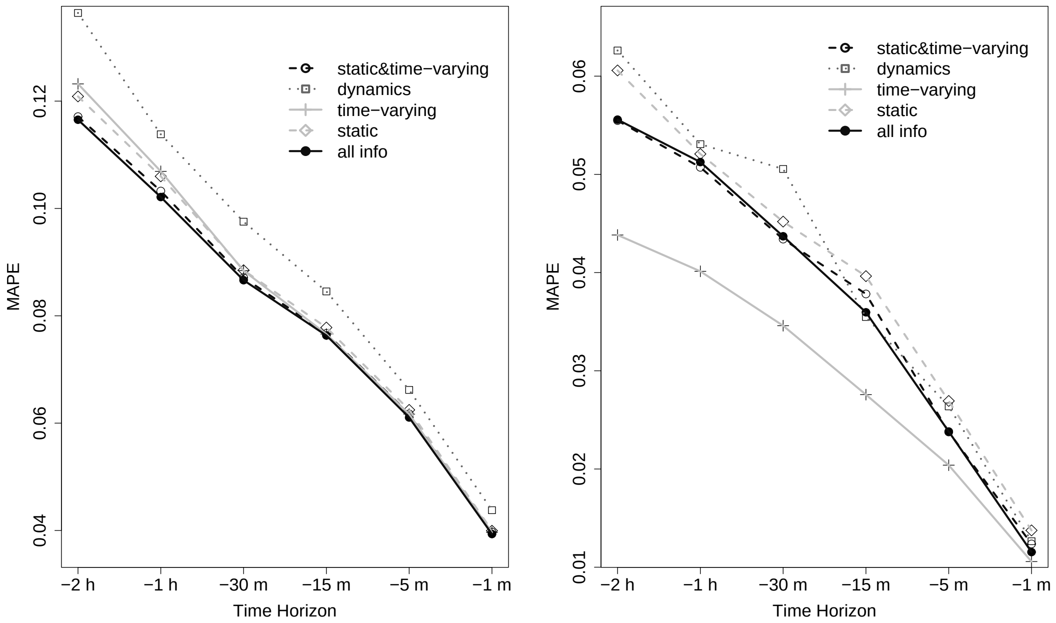


Figure 4.7: Comparison of different distance metrics. The left panel is for laptop auctions; and the right panel is for palm auctions.

Table 4.2: Optimal choice of K for different distance metrics D and different time horizons δ_T. The top panel corresponds to the laptop data; the bottom panel is for the Palm PDA data.

	Laptop Data					
Time Horizon	2h	1h	30min	15min	5min	1min
static	95	94	99	99	81	14
time-varying	31	27	97	100	91	89
dynamics	100	100	100	100	100	100
static&time-varying	40	79	100	96	47	44
all	33	77	98	100	44	41

	Palm PDA Data					
Time Horizon	2h	1h	30min	15min	5min	1min
static	52	69	63	63	61	37
time-varying	3	10	4	1	1	2
dynamics	94	95	95	29	29	68
static&time-varying	7	18	32	52	12	8
all	6	40	30	61	11	2

for the Palm PDA data. Each line corresponds to a distance metric $D \in \{d_s^{Static}$, $d_s^{Time-Varying}$, $d_s^{Dynamics}$, $d_s^{Static\&Time-Varying}$, $d_s^{All}\}$. We can see that, for each data set, a single distance metric yields the consistently best result across all values of the time horizon. That is, d_s^{All} results in the best prediction accuracy for the laptop data, regardless of the value of δ_T; similarly, $d_s^{Time-Varying}$ yields the best results for all values of δ_T in the Palm PDA data. This suggests that the choice of the distance metric is very robust to the forecasting horizon, at least for a given data set. We also note that while $d_s^{Time-Varying}$ significantly outperforms all other distance metrics for the Palm PDA data, for the laptop data most choices of D (except for $d_s^{Dynamics}$) yield very comparable results, at least for short time horizons ($\delta_T \leq 30$ min).

4.5.3 Comparison With Alternate Prediction Methods

We evaluate the performance of functional KNN by comparing its predictive accuracy to more classical approaches – linear regression models and tree (CART)[4].

We study the performance of all methods on the test set. Recall that we partitioned our data into a training set (50%), validation set (25%) and test set (25%). While we estimated the fKNN forecaster on the training set and optimized its parameters K and D on the validation set, we now compare its performance (using the optimal parameter values) on the test set. That is, for each time horizon δ_T, we use the optimal combination of K and D from the previous section. In order to make fair comparisons, we apply regression and CART using the same

[4] We used the software defaults for pruning in CART; that is, we used the defaults in the R package *rpart*.

information as for functional KNN.

Figure 4.8 shows the results. We display the *relative* prediction error between fKNN and the regression model (dotted line) and between fKNN and the tree model (dashed line). We can see that fKNN generally outperforms its two competitors. In particular, for the laptop data (left panel), fKNN outperforms the tree model by as much as 40%. While the gap between the regression model and fKNN is smaller, fKNN leads to improvements that range between 5% and 10%. The picture is similar for the Palm PDA data (right panel). While for this data set fKNN also leads to general improvements, it is curious to see that only for the longest time horizon ($\delta_T = 2h$), both alternate approaches are competitive.

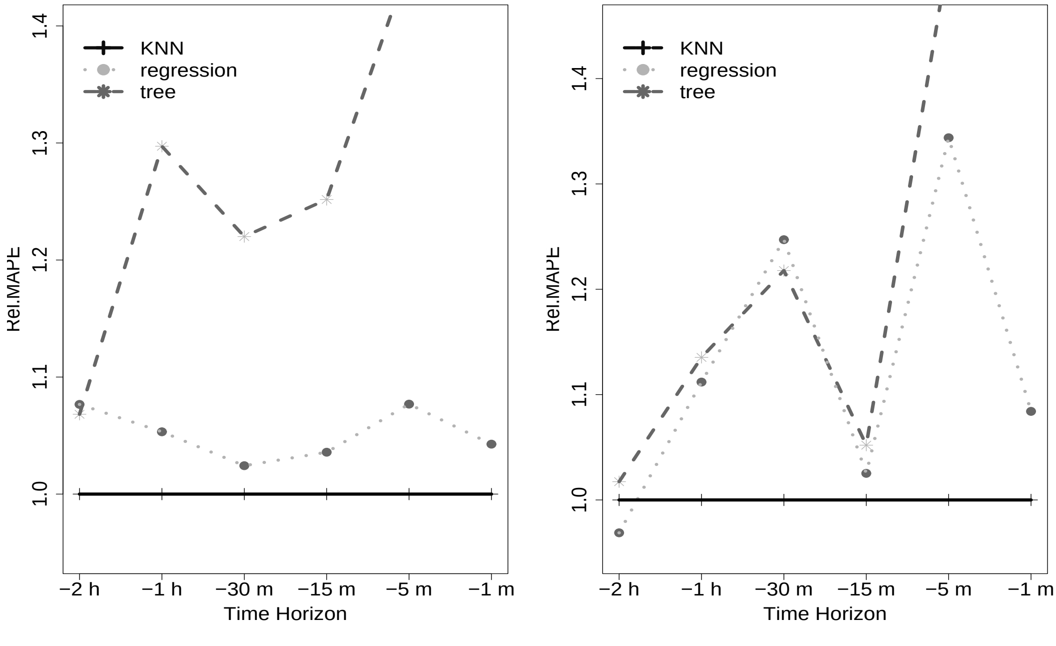


Figure 4.8: Comparison of different forecasting methods. The left panel corresponds to laptop auctions; the right panel is for the Palm PDA auctions.

It is revealing to compare performance on each of the two data sets. While for the laptop data, both fKNN and regression significantly outperform CART, the

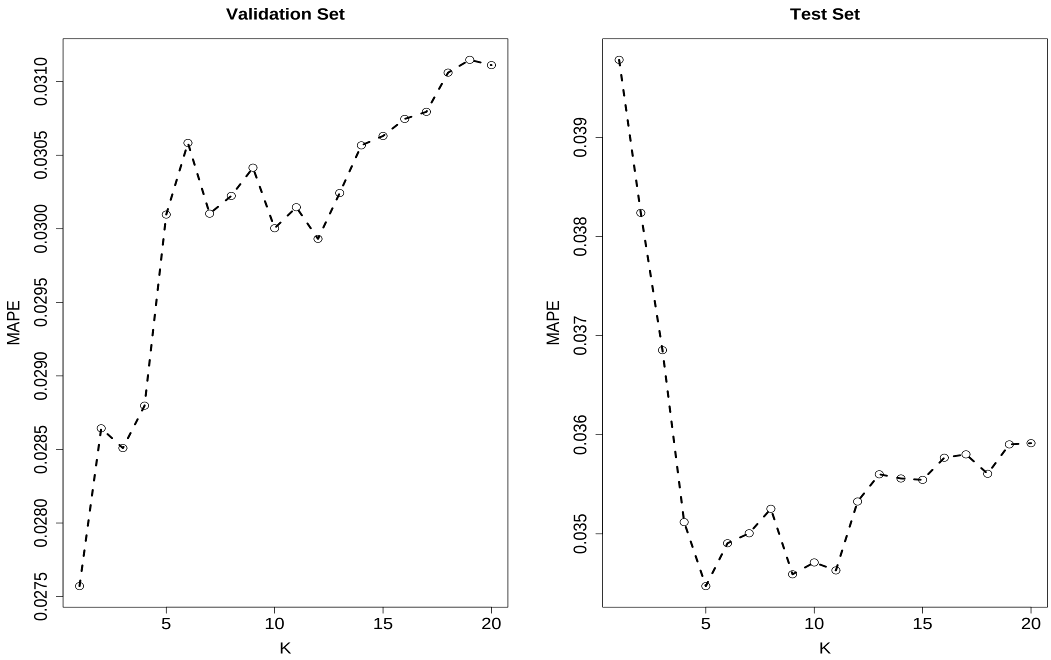


Figure 4.9: Optimal values of K for the Palm PDA data at $\delta_T = 15$ min. The left panel corresponds to the validation set; the right panel corresponds to the test set.

gap is not as large in the Palm PDA data; in fact, for the Palm PDA data, CART and regression are comparable for almost all time horizons. The poor performance of CART on the laptop data illustrates the general problem of the method with prediction: while it often fits the training set well, it has a tendency to over-fit and thus perform poorly on the test set, especially in situations like the laptop data where the underlying population is very heterogeneous. On the other hand, functional KNN can handle heterogeneous populations well by selecting only those neighbors that are most relevant for the focal auction; in particular, compared to regression, it performs especially well for forecasting longer time horizons (one hour, two hours), which is very relevant in practical situations.

Functional KNN also leads to improvements for less heterogeneous data sets such as the Palm PDA data. While the right panel in Figure 4.8 suggests that fKNN outperforms both competitors for every time horizon, there is a sharp drop for the competitors at $\delta_T \leq 15$ min. At this point, both regression and the tree model perform almost as well as fKNN. A closer investigation of this phenomenon reveals that for this time horizon, the optimal value of K (based on the validation set) equals one (see left panel in Figure 4.9); however, that value leads to very poor performance on the test set (right panel in Figure 4.9). This suggests that finding the right value of K is especially difficult for homogeneous data sets (such as the Palm PDA data). While the data-homogeneity suggests very small values of K, slight perturbation of the homogeneity can lead to weaker results. This was already implied by the lack of robustness seen in Section 4.5.2.1.

4.6 Conclusions

In this chapter, we propose a novel functional KNN forecaster for forecasting the final price of an ongoing online auction. Assuming that more similar auctions contain more relevant information for incorporation into forecasting models, we propose a novel dissimilarity measure that takes into account both static and time-varying features as well as the auction's price dynamics information. The latter is obtained via a functional representation of the auction's price path. We select both the optimal distance metric as well as the optimal number of neighbors based on a validation set. We find that weighting information unequally yields better forecasts compared to classical methods such as regression models or trees, and this result holds in auctions of varying levels of heterogeneity. Moreover, the proposed Beta model has many nice properties as a representation of auction price path besides providing distance measures for functional price curves. We explore those properties in further detail in the next Chapter.

Although we observe improvement of the KNN forecaster over regression and CART for auctions of varying levels of heterogeneity, our study shows that the improvement is bigger for heterogeneous data. This means that selecting the most useful information and making use of only most relevant neighbors is especially crucial for prediction accuracy in situations where objects are heterogeneous and information is noisy. This fact is true not only for forecasting online auctions but also in many other forecasting situations (e.g., weather forecasting). Another finding worth noticing is the robustness of the optimal distance to the time horizon. The fact that the same distance metric is optimal regardless of the time horizon implies that the most important information for making price prediction is time-invariant. This insight simplifies the process of decision making. To compute forecasts, we only need to find the optimal distance once, and this distance can then be re-used as the forecasting process proceeds.

There are several ways to extend this study. While we scale distance metrics for different information sources to achieve equal weighing across all metrics, one could alternatively assign individual weights to individual metrics and then optimize the weights. There are also alternative ways to define distances for different data types. For example, for categorical data we can define several levels of category "similarity", such as "US brand". Then, the distance between items can be set to 0.5 for "similar categories" (e.g., US brand) or 1 for categories more different.

Another way to complement this study is by investigating alternates to classical linear regression and trees, e.g. via weighted regression or tree models, which might lead to forecasting advantages especially for heterogeneous data.

Chapter 5

A Flexible Model for Price Dynamics in Online Auctions

5.1 Introduction

One stream of online auction research has focused on the *dynamics* of the auction price paths which leads to deeper understanding of price formation process. Descriptive studies have shown that price dynamics can be very heterogeneous, even for auctions of the same product (e.g. [51; 82]). Furthermore, statistical approaches, such as functional data analysis, has been developed and employed in providing insight into price dynamics [52; 10; 86].

Price dynamics have also been shown instrumental for price forecasting. [96] pioneered real-time forecasting models for ongoing auctions where price dynamics serve as important predictors. [57] recently expanded upon this idea. Both studies show that the inclusion of the dynamic information adds additional predictive power to the forecasting model compared to models that do not make use of such information.

In summary, the price path and the price dynamics are of special interest in online auctions, and therefore developing models that can capture them effectively are both important and useful. In Chapter 4, we have briefly introduced a Beta model for capturing the price path and its dynamics; and it is employed to help measure distances between auction price paths. The Beta model is parsimonious, flexible, and computationally efficient. In this Chapter we further explore the properties of the beta model, compare our model to alternative existing models and show its advantages both in terms of fit as well as forecast accuracy. This work has been submitted to the Journal of the Royal Statistical Society C for review.

The remainder of the paper is organized as follows: Section 5.2 presents existing models for capturing price path and dynamics in online auctions. In section 5.3, we introduce our Beta model and describes its properties, estimation, and advantages over alternate approaches. We then compare the Beta model to several competitors empirically in Section 5.4, in terms of fit as well as forecast accuracy. We conclude in Section 5.5.

5.2 Models for Auction Price Paths

Dynamics (e.g. velocity or acceleration) are typically computed as the first or second derivative of an underlying smooth function. However, observed bids create a non-decreasing step function with jumps at the times of bids (see e.g. Figure 2.1). Thus, in order to gauge an auction's price dynamics, one needs some smooth representation of the price path. There have been two general approaches to obtaining smooth auction price paths from observed bid data: non-parametric and parametric. Using a functional data analytic (FDA) approach, [54] employed penalized smoothing splines (p-splines) to generate smooth curves. An alternative to p-splines are monotone splines [78] which guarantee the monotonicity of the resulting curves. Finally, [45] proposed a parametric family of four distributions that capture an auction's most typical price path shapes. Each of these three approaches yield smooth price curves, and then price dynamics are computed by taking derivatives of the smooth curves. The first derivative captures price velocity (i.e. how fast the price is moving at any point in time); the second derivative captures price acceleration, and so forth.

In the following we describe each of these three approaches and discuss their strengths and weaknesses.

5.2.1 Smoothing Splines

Penalized smoothing splines (p-splines) [88] fit a polynomial of order p. In order to control the smoothness of the fitted curve, a penalty is imposed on the estimating function. Let $\tau_1, \tau_2, \ldots, \tau_L$ be a set of knots, then a polynomial spline of order p is given by

$$f(t) = \beta_0 + \beta_1 t + \beta_2 t^2 + \ldots + \beta_p t^P + \sum_{l=1}^{L} \beta_{pl}(t - \tau_l)_+^p \qquad (5.1)$$

where $u+ = uI(u \geq 0)$ is the positive part of the function u. Many functions of this type tend to fit the data too closely (and thus model noise in addition to the signal); therefore, a *roughness penalty* approach is often employed which takes into account the trade off between data-fit (i.e., minimizing $f(t) = \sum_j (y_j - f(t_j))^2$) and function smoothness. A popular measure of roughness, which measures degree of departure from a straight line, is of the form

$$PEN_m(t) = \int [D^m f(t)]^2 dt \tag{5.2}$$

where $D^m f, m = 1, 2, 3, \ldots$ denotes the m^{th} derivative of the function f. A highly variable function will yield a high value of $PEN_m(x)$. If the highest derivative of interest is m, then using $m + 2$ as the polynomial order will assure m continuous derivatives. The penalized smoothing spline f minimizes the penalized squared error

$$PENSSE_{\lambda,m} = \int (y(t) - f(t))^2 + \lambda PEN_m(t). \tag{5.3}$$

When the roughness parameter is set to $\lambda = 0$, the penalized squared error drops out, and the function fits the data perfectly. Larger values of λ penalize the function for being curvy, and as $\lambda \rightarrow \infty$, the fitted curves approach a linear regression.

Smoothing splines are widely used in functional data analysis. They are advantageous in terms of their flexibility which results in good data-fits, in terms of their ease of obtaining derivatives (i.e. dynamics) and in terms of their computational efficiency. However, there are several challenges when applying penalized smoothing splines to the online auction context. Firstly, although bidding data are non-decreasing over time, smoothing splines do not necessarily result in monotonically non-decreasing curves. Hence, they may not truly reflect the monotonic nature of the auction price process. Second, the fitted curves are often very variable, especially at their ends, even with a heavy smoothness penalty. This is problematic in the auction context, where the opening and closing prices are of special importance. Moreover, in a forecasting context it is crucial to obtain precise estimates at the final stages of the auction (see Chapter 2 and [96]). Finally, smoothing splines require the specification of many nuisance parameters (such as the smoothing parameter, the number and position of knots, and the polynomial order) which are often determined in an ad-hoc fashion. While one can estimate some of these parameters from the data (e.g., using cross-validation), their optimal choice is not always guaranteed.

5.2.2 Monotone Splines

Monotone splines [78] are a natural alternative to smoothing splines in the online auction context, since they guarantee a monotone path of the resulting price process. The idea behind monotone smoothing is that monotonously increasing functions have a positive first derivative. The exponential function has this property and can be described by the differential equation $f'(t) = w(t)f(t)$. This means that the rate of change of the function is proportional to its size. Consider the linear differential equation

$$D^2 f(t) = w(t) Df(t). \tag{5.4}$$

Here, $w(t) = \frac{D^2 f(t)}{Df(t)}$, which is the ratio between acceleration and velocity. The differential equation has the following solution:

$$f(t) = \beta_0 + \beta_1 \int_{t_0}^{t} exp\left(\int_{t_0}^{v} w(v)dv\right) du \tag{5.5}$$

where t_0 is the lower bound over which we are smoothing. After some substitutions (see [79]), we can write

$$f(t) = \beta_0 + \beta_1 e^{wt}. \tag{5.6}$$

and estimate β_0, β_1, and $w(t)$ from the data. Since $w(t)$ has no constraints it may be defined as a linear combination of K known basis functions (i.e., $w(t) = \sum_k c_k \phi_k(t)$). The penalized least squares criterion is thus

$$PENSSE_\lambda = \sum_i [y_i - f(t)]^2 + \lambda \int_0^T [w^2(t)]^2 dt. \tag{5.7}$$

For capturing online auction price paths and dynamics, monotone smoothing indeed solves the excess variability of penalized smoothing splines and their non-monotonicity problems. The resulting curves are better representations of a continuous non-decreasing price path, and dynamics can be computed via curve derivatives. However, some challenges remain and new ones arise. First, monotone smoothing is computationally intensive, as it relies on an iterative fitting process where several passes have to be made through the data. Therefore, fitting a dataset of even tens or hundreds of auctions can take a long time. Second, like smoothing splines, there are many nuisance parameters to be determined (the number and location of knots and the smoothing parameter). Hence, while, at least conceptually, monotone splines are preferable over smoothing splines, they can be slow to implement even with medium-sized datasets.

5.2.3 Parametric Growth Models

To overcome the disadvantages of smoothing splines and monotone splines, [44] proposed a family of four growth models for representing the price process. They find that the shape of auction price paths can be categorized into four main types: exponential growth, logarithmic growth, logistic growth, and reflected-logistic growth. These four models not only provide parametric fit of monotone data, but they also have appealing interpretations, and are easy to estimate. We describe each of the four models next.

5.2.3.1 The Exponential Model

Exponential growth has been used for describing a variety of natural phenomena including the dissemination of information, the spread of disease, and the multiplication of cells in a petrie dish. In exponential growth the rate of growth is proportional to a function's current magnitude; that is, growth follows the differential equation

$$Y'(t) = rY(t), \tag{5.8}$$

or the equivalent equation

$$Y(t) = Ae^{rt}, \tag{5.9}$$

where t denotes time, and $r > 0$ is the growth constant. Equivalently, exponential decay, when $r < 0$, can model phenomena such as the half-life of an organic event. In an online auction context, exponential growth describes a price process with gradual price increases until mid-to-late auction, and a heavy price jump towards the end.

5.2.3.2 The Logarithmic Model

Logarithmic growth is technically the inverse of the exponential function,

$$Y(t) = \frac{1}{r} \ \ ln(\frac{t}{A}), \tag{5.10}$$

The resulting curves are reflections of exponential growth over the line $x = y$. In the online auction context, such behavior occurs when early bidding quickly increases the price during the opening stages of the auction, but because of market constraints (e.g. a market value or budget constraints), price flattens out for the remainder of the auction. This type of price behavior tends to be rare, as most bidders do not wish to reveal their valuations early in the auction. However, inexperienced bidders who may not completely understand eBay's proxy bidding mechanism, may place high early bids.

5.2.3.3 The Logistic Model

Logistic growth is useful for describing processes which reach a limit or a "carrying capacity". In the context of auction prices, in many cases there are competing online and brick-and-mortar markets for the auctioned item, thereby creating a "market value" for the item.

The logistic model is given by

$$Y(t) = \frac{L}{1 + Ce^{rt}}, \tag{5.11}$$

and the differential equation is

$$Y'(t) = rY(t)(\frac{Y(t)}{L} - 1), \tag{5.12}$$

where L is the carrying capacity, t is time, r is the growth rate, and C is a constant. Logistic growth can also be explained in the auction context as a stretched-out "S"-shaped curve, where the price increases slowly early, jumps up during mid-auction, and levels off towards the end of the auction. The resulting closing price is analogous to the carrying capacity L in the logistic growth function.

5.2.3.4 The Reflected-Logistic Model

Another common price process in online auctions is a reflected "S" shaped curve. Such behavior can be captured by the inverse of logistic growth, or reflected-logistic growth, given by the function

$$Y(t) = \frac{ln(\frac{L}{t} - 1) - ln(C)}{r}. \tag{5.13}$$

In the online auction context, this type of growth occurs when there is some early bidding that results in a price increase, followed by little to no bidding in the middle of the auction, and then another price increase as the auction approaches its close. In particular, price spikes near the end may be caused by sniping.

5.2.3.5 The 4-member growth model family

The set of four growth models is used to approximate price paths as follows: For a dataset of auctions, each of the four models is fitted to each auction. Then, for each auction, the four estimated models are compared in terms of fit, and the best fitting model is chosen (for more on the fitting process, see [45]). Hence, the fitting process is a two-stage process.

Since the family is entirely parametric, no nuisance parameters require determination. Moreover, since the family is monotonic, it is well-suited for capturing auction price processes. Moreover, the 4-member family of growth models is computationally efficient compared to monotone splines, and ordinary least squares functions can be used for estimation.

The main disadvantage of the four-model family is it is limited to only four basic shapes – exponential, logarithmic, logistic, and reflected-logistic – which may be overly simplistic for some auction scenarios. Moreover, because the four models are not nested within a single model, comparing fit (for choosing the best model) is nontrivial. Finally, when fitting the exponential and logistic models via least squares, the models minimize error in the bid *amount* space. In contrast, when fitting the two reflected models (logarithmic and reflected-logistic growth) using least squares, the error minimization is done in the bid *time* space. A comparison is therefore more complicated.

5.2.4 Comparison

To illustrate the difference between penalized splines, monotone splines and the 4-member family of growth models, consider Figure 5.1, which displays the fit of the three methods to two sample auction price paths. The solid lines represent p-splines, the dashed lines represent monotone splines, and the dotted lines represent the best fit of the 4-member growth model family. We see that while p-splines can fit the data very well, they are very variable and do not capture the monotonic nature of the price path. While the 4-member growth family results in a monotonous price path, it does not fit the data well. Monotone splines appear to provide the best fit in this example; however, it takes on average almost 7 seconds to fit a single monotone spline (compared to 0.02 seconds for one p-spline and and 0.04 seconds for one 4-member growth model).

5.3 A New Model for Auction Price Paths: The Beta Model

In light of the shortcomings of existing models for online auction price paths and dynamics, we introduce a single parametric model that is flexible yet parsimonious for approximating price paths and their dynamics. We have briefly introduced the model in Chapter 4 based on which Kullback-Leibler distance is used to measure the distance between two auction price paths. We now discuss model estimation and its properties in detail.

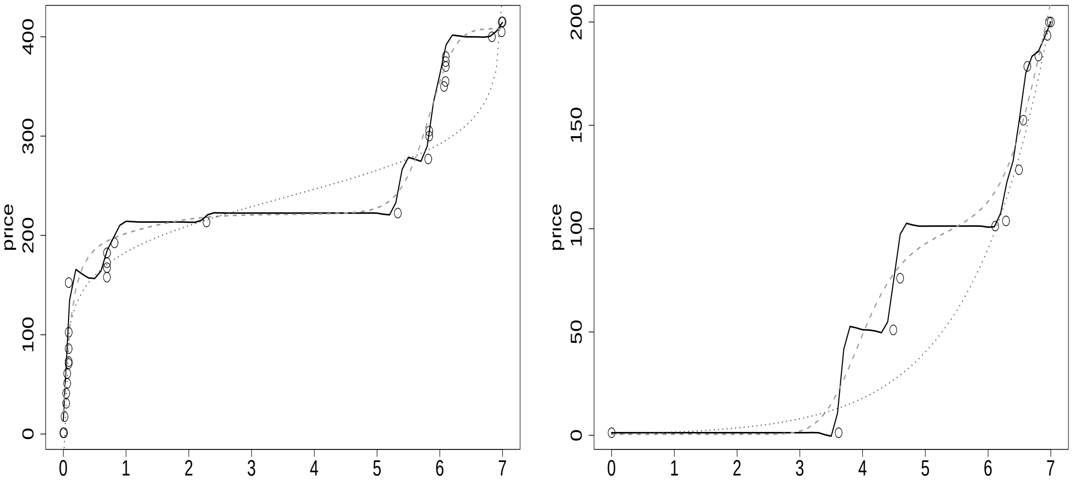


Figure 5.1: Illustration of the three existing smoothing methods. The solid lines represent p-splines, dashed lines are for monotone splines, and the dotted lines are for growth models.

Our proposed model is based on the Beta cumulative distribution function (CDF). The Beta distribution is a continuous probability distribution defined on the interval $[0, 1]$ with two shape parameters (α and β) that fully determine the distribution. Its CDF can be written as

$$\mathrm{F}(x, \alpha, \beta) = \frac{\int_0^x u^{\alpha-1}(1-u)^{\beta-1}du}{B(\alpha, \beta)}, \tag{5.14}$$

where $B(\alpha, \beta)$ is the *beta* function[1], which serves as a normalization constant in the CDF to ensure that $F(1, \alpha, \beta) = 1$.

We use the Beta model to capture auction price paths in the following way. Let p denote the sequence of observed bids with associated time-stamps t. Since auctions can be of varying durations, we thus normalize the time sequence to a 0-1

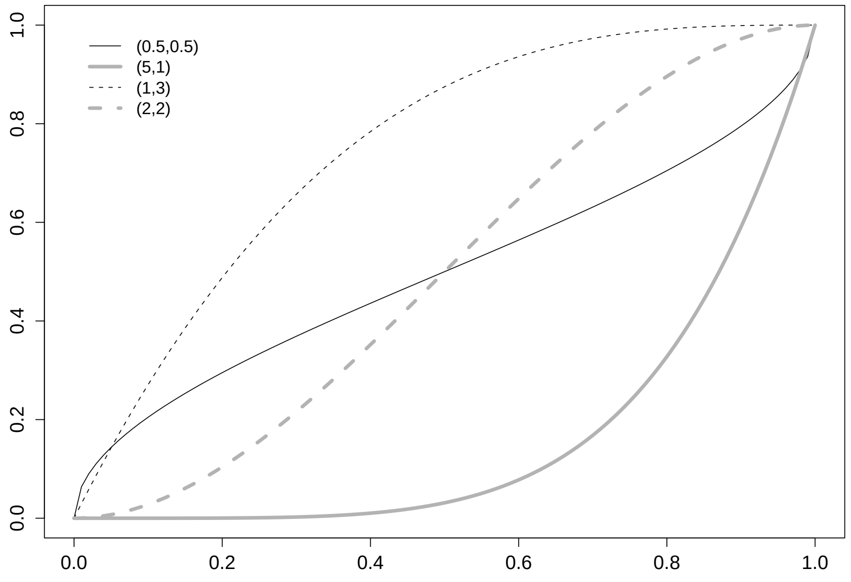
[1] $B(\alpha, \beta) = \int_0^1 u^{\alpha-1}(1-u)^{\beta-1}du$

scale by using the transformation $t_n = t/Duration$. t_n are time-stamps between 0 and 1. Similarly, because auctions close at different prices, we normalize the observed bids to a 0-1 scale by using the transformation $p_n = p/ClosingPrice$. p_n are bid values between 0 and 1. The goal is then to find the values of α and β that satisfy $p_n = \int_0^{t_n} u^{\alpha-1}(1-u)^{\beta-1}du/B(\alpha,\beta)$ for every element of p_n and t_n. An algorithm for achieving this goal efficiently is described in Section 5.3.1.

The Beta model is very flexible in the types of curves that it can produce. It includes as special cases the four shapes of the 4-member growth model family of [45]. The top panel in Figure 5.2 shows the Beta model curves for different values of α and β. The solid line represents the case where price grows rapidly at the auction beginning and at the end, but not in the middle, corresponding to logit growth. The long-dashed line represents the situation where rapid growth only occurs at the end, corresponding to exponential growth. The short-dashed line shows early rapid growth, corresponding to logarithmic growth. And finally, the dotted-dashed line captures a rapid increase in price somewhere in the middle of the auction, corresponding to the reverse-logit growth pattern.

5.3.1 Fitting the Beta Model

Fitting the Beta model to bid data can be done in a way that results in curves that fit well in two dimensions: bid time and bid amount. In the auction context both dimensions are important. In particular, a good fit in terms of the bid timing is necessary in order to accurately capture points of different bidding activities.



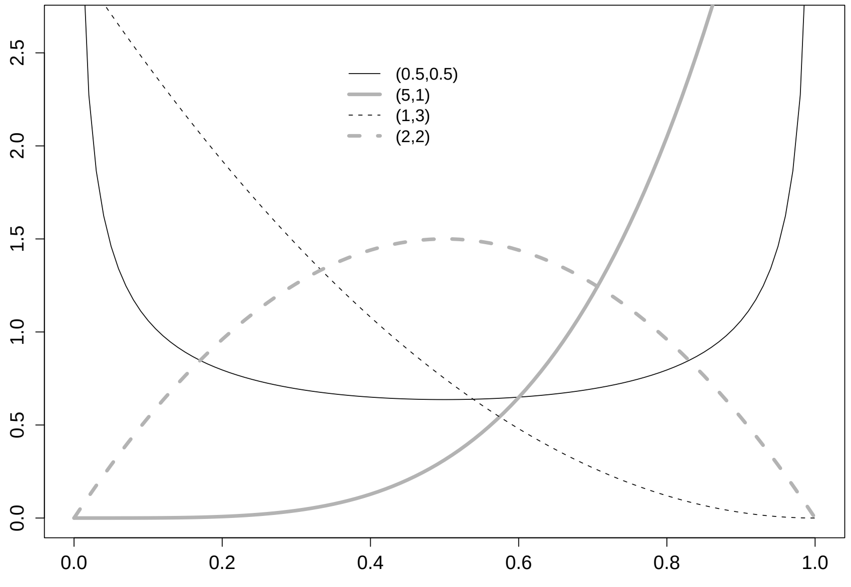


Figure 5.2: Beta CDF (top panel) and corresponding PDF (bottom panel) with different shape parameters (α, β)

Periods of vastly different biding activity, e.g. early or last-minute bidding, have been documented well in the auction literature (e.g. [86]), and they are important to capture adequately. In terms of bid amounts, a model that adequately captures the bid amounts (i.e., the price at that point of the auction), is necessary for generating accurate forecasts of an auction's final price. Auction price forecasting is of practical interest and different forecasting models have been suggested in the literature [32; 33; 96; 57; 53; 21].

The only inputs required for fitting the Beta CDF are the observed bid amounts and their associated time stamps. The resulting price path representation is characterized by only two parameters. The simplicity and parsimony of the Beta model distinguish it from alternative approaches. Our algorithm for fitting the Beta CDF minimizes residuals in both bid amount and bid time dimensions simultaneously.

5.3.1.1 Beta-Fitting Algorithm

For a given auction, we estimate α and β from the observed bids as follows:

Step 1: Standardize bid amounts and bid times

Since the range (y) as well as the domain (x) of the Beta CDF is $[0, 1]$, we first standardize the bid amounts and bid times by the following two transformations.

$$y \leftarrow \frac{bid - min(bid)}{max(bid) - min(bid)}$$

and

$$x \leftarrow \frac{time - min(time)}{max(time) - min(time)}.$$

x and y are now bid times and bid amounts standardized within $[0,1]$.

Step 2: Compute α_0 and β_0, the initial values of $\hat{\alpha}$ and $\hat{\beta}$

Since we treat x as a Beta-distributed random variable, it is reasonable to assume that the empirical average and variance of x are close to their theoretical mean and variance. That is, $mean(x) \simeq \frac{\alpha}{\alpha+\beta}$ and $var(x) \simeq \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$. Therefore, the initial values of α and β are found by solving the minimization problem:

$$(\alpha_0, \beta_0) = \left\{(\alpha^*, \beta^*) | DIST^A(\alpha^*, \beta^*) = min(DIST^A(\alpha, \beta))\right\},$$

where $DIST^A(\alpha,\beta) = \left(mean(x) - \frac{\alpha}{\alpha+\beta}\right)^2 + \left(var(x) - \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}\right)^2$.

Step 3: Compute $\hat{\alpha}$ and $\hat{\beta}$

In order to capture both the bid levels as well as the bid times, our model minimizes error both in y and x directions simultaneously. Specifically, we choose to minimize the sum of the squared residuals in y and x directions. With the initial values α_0 and β_0 from Step 2, we solve for $\hat{\alpha}$ and $\hat{\beta}$ through the following minimization problem:

$$(\hat{\alpha}, \hat{\beta}) = \left\{(\alpha^*, \beta^*) | DIST^B(\alpha^*, \beta^*) = min(DIST^B(\alpha, \beta))\right\}$$

where $DIST^B(\alpha,\beta) = \sum(y - pbeta(x,\alpha,\beta))^2 + \sum(x - qbeta(y,\alpha,\beta))^2$; and *pbeta* and *qbeta* represent the cumulative distribution function and the inverse of the cumulative distribution function of the beta distribution respectively.

The above algorithm is computationally very efficient. It takes, on average, 0.0489 seconds to fit the Beta model to one auction (using the above algorithm), which compares favorably to the 4-member growth family (0.0362 seconds). Unsurprisingly, penalized splines fare better (0.0190 seconds) since they do not encounter any iterations. Conversely, fitting monotone splines, which do require iterative passes through the data, result in 150 times larger computing times (on average of 6.9726 seconds per auction).

5.3.2 Properties of the Beta Model

The Beta model shares the main properties of competing methods (p-splines, monotone splines and the 4-family growth model), but it also has several additional properties that set it apart. Like all competing methods, the derivatives of the continuous Beta curves can be used to capture price dynamics. The Beta model produces monotonically non-decreasing curves, yet it is computationally fast (using the algorithm from Section 5.3.1). Unlike non-parametric approaches, fitting the Beta model does not involve any nuisance parameters.

Like the 4-member parametric growth model, the two-parameter Beta model can be used to characterize auction types (e.g. exponential, logarithmic, logistic or reflected logistic) in terms of price dynamics.

The Beta model has two additional unique properties, which make it especially advantageous in the online auction context: (1) Because both of its dimensions (bid time and bind amount) are derived from a probability function, the Beta summary statistics can be used to learn about the bid timing distribution, and (2) there is an easy and straightforward way to measure pairwise distances between price paths. The latter is especially useful in the context of pairwise comparisons and dynamic forecasting as shown in Chapter 4. We discuss those model properties in detail next.

5.3.2.1 Representing Price Dynamics

The Beta CDF representation of the price paths means that price velocity, which is the first derivative of the price curve, is given by the Beta probability density function (PDF). In particular, at any given time T, the price velocity of an auction with shape parameters α and β can be computed as:

$$Vel(t, \alpha, \beta) = \frac{t^{\alpha-1}(1-t)^{\beta-1}}{B(\alpha, \beta)}, \tag{5.15}$$

where t is the normalized T on a scale of $[0, 1]$ ($t = \frac{T}{duration}$) and $B(\alpha, \beta)$ is the *beta* function.

The bottom panel in Figure 5.2 plots the price velocities corresponding to the price paths in the top panel. The solid black line shows rapid dynamics at the beginning and end, but not much price activity during the middle; in contrast, the dashed gray line signals heightened price dynamics during mid-auction. Similarly, the solid gray line captures increased price-velocity towards the end while the dotted dashed line captures early price spurts.

Higher order price dynamics can also be readily obtained by taking higher order derivatives. For example, price acceleration can be computed as

$$Acc(t, \alpha, \beta) = \frac{t^{\alpha-1}(1-t)^{\beta-1}}{B(\alpha, \beta)} \left(\frac{\alpha - 1}{t} - \frac{\beta - 1}{1 - t} \right)$$

Price dynamics carry important information about the auction process [10]. Therefore accurate approximations of price dynamics are beneficial across multiple applications. In section 5.4 we show that the price dynamics generated via the Beta model lead to more accurate price forecasts compared to competing approaches.

5.3.2.2 Characterizing Growth Patterns

Similar to the 4-member growth family of [45], the Beta model provides a tool for characterizing price process types. In fact, there is a one-to-one mapping between the Beta model and the four growth models via the shape parameters α and β. For example, if both α and β are smaller than 1, then the price curve is similar to the reflected-logistic model. Table 5.1 lists the relationship between the Beta model and the 4-member growth family. The implication of this relationship is that it allows us to easily characterize auctions in terms of their type of price dynamics, without the need of more specialized techniques such as functional clustering (e.g. [54]) or via laborus visual examination (e.g. [44]).

5.3.2.3 Characterizing Bid Timing

The estimated Beta parameters α and β can be used to compute summary statistics which capture bid timing information. Table 5.2 gives the formulas for the

Table 5.1: Correspondence between the Beta model and the four growth models

Growth Models	Beta Model	
Exponential	$\alpha = 1$ $\alpha > 1$	$\beta < 1$ $\beta \leq 1$
Logarithmic	$\alpha < 1$ $\alpha = 1$	$\beta \geq 1$ $\beta > 1$
Logistic	$\alpha > 1$	$\beta > 1$
Reflected-logistic	$\alpha < 1$	$\beta < 1$

variance, mode, and skewness. The variance gives information about the dispersion of the bid arrivals; the mode, which is the peak of the price velocity curve, tells us about the time during the auction when the price moved fastest. Finally, skewness measures the level of asymmetry in the bid timings. Online auctions tend to see either high bidding activity at the start and/or at the end.

Table 5.2: Beta distribution summary statistics and their auction meaning

	Formula	Explanation
Variance	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$	Dispersion of the bid arrivals
Mode	$\frac{\alpha-1}{\alpha+\beta-2}$	The peak of the velocity curve; price increases fastest at this point.
Skewness	$\frac{2(\beta-\alpha)\sqrt{\alpha+\beta+1}}{(\alpha+\beta+2)\sqrt{\alpha\beta}}$	Asymmetry of the bid arrivals.

5.4 Empirical Comparison

In this section we compare the proposed Beta model with competing methods for fitting auction price paths. In particular, we compare it to p-splines, monotone splines and the 4-member growth model family. Comparisons are made on two different dimensions: In terms of fit, we compare the different models' ability to generate accurate price representations of observed auction data; and in terms of prediction, we compare the forecast accuracy of the four methods in predicting the final price of a set of ongoing auctions. We will see that in the forecasting context, it is especially important not only to have an adequate approximation of the auction's price path, but also of its price dynamics. The data we use is the Palm PDA data set (Appendix A).

5.4.1 Model-fit Comparison

To evaluate goodness of fit of a model, we examine the residual error both in terms of bid amount (y) and bid time (x), because it is important to accurately capture not only the times when bids are placed but also the resulting price. For this purpose, we define the residual error for the ith auction with n bids as

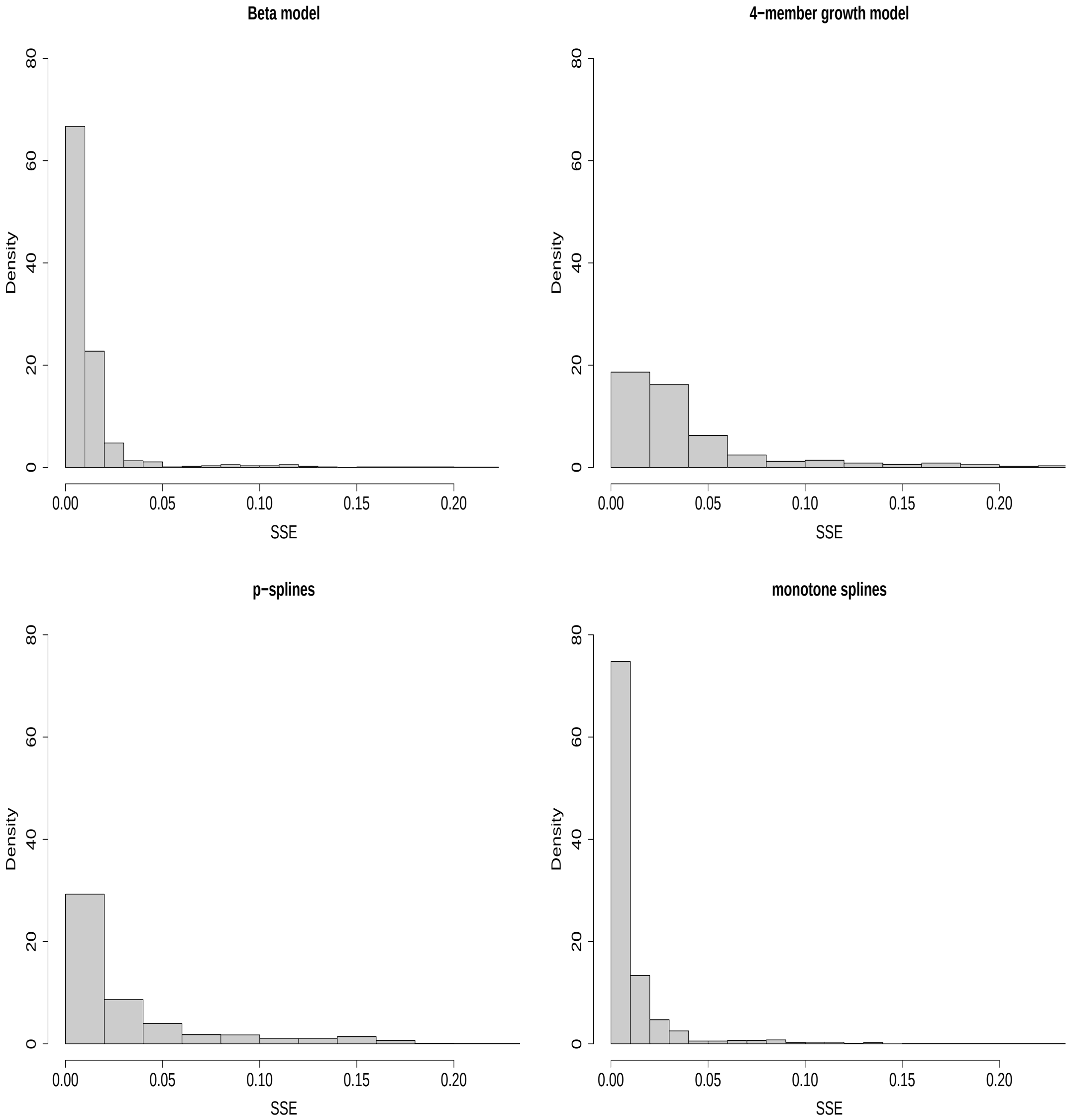
$$\text{resid}_i = \frac{1}{n}\sum_{k=1}^{n}[0.5(y_k - \hat{y_k})^2 + 0.5(x_k - \hat{x_k})^2],$$

where $(x_k,\ y_k)$ and $(\hat{x_k},\ \hat{y_k})$ are the observed and fitted values, respectively. Note that since both p-splines and monotone splines only minimize errors in terms of bid amount (y), we set $x_k = \hat{x_k}$, which may result in an overly optimistic view of these two methods.

We apply all four methods to the Palm PDA auction data set. That is, for each auction we estimate a p-spline, a monotone spline, the best of the 4-member growth model family, and our Beta model. For each auction, we first normalize the observed bids p and associated times t into a [0,1] scale via the transformations $p_n = p/ClosingPrice$ and $t_n = t/Duration$. We then fit each of the models[2] to the normalized data $(t_n,\ p_n)$. Normalization results in an equal weighing of the residuals in both bid time and bid amount dimensions, since they are measured on equal scales. We repeat the process for all 380 auctions in our data set.

The distributions of the absolute residuals are shown in Figure 5.3. We can see that the Beta model (top left panel) results in the second-best fit (average error = 0.0125), surpassed only slightly by the fit of monotone splines (bottom right panel; average error = 0.0112). Both, p-splines and the 4-family growth models, result in a much worse representation of the data (average error of 0.0326 and 0.0434, respectively). But also recall the much longer estimation time for monotone splines: it takes a total of 2,650 seconds (or 44 minutes) to fit the 380 auctions; this compares to only 19 seconds for the Beta model!

[2] For p-splines and monotone splines, the smoothing parameters are determined using leave-one-out cross-validation; for the 4-member growth family, we fit each of the 4 growth models via least squares and then select the model with the best fit [45]; fitting Beta models is summarized in Section 5.3.1.1.



Beta model

4−member growth model

p−splines

monotone splines

Figure 5.3: Residuals for the four models: Beta model (top left), 4-member growth models (top right), p-splines (bottom left) and monotone splines (bottom right).

5.4.2 Forecasting Accuracy Comparison

We now compare the four methods in terms of their capability of producing accurate forecasts of an auction's final price. Information of the final price ahead of time has advantages for all auction participants. Bidders can use this information to make more informed bidding decisions (see Chapter 2 or [57]). Sellers can use predictions to identify times when the market is more favorable to sell their products (e.g., higher demand, lower supply). We pay particular attention to the role of price dynamics: how different smoothing methods result in different dynamics and the subsequent effect on predictions.

The following model incorporates the price path and price dynamics in a linear fashion for predicting the final price. Although one can use a wide range of model-formulations, such as complicated regression models (see Chapter 2 and 3) or tree models (Chapter 4), we choose the simple linear model for simplicity. Linear regression models have also been the main tool for investigating price in online auctions. More formally, to forecast the final price at time T during the ongoing auction, we use the model:

$$FinalPrice = \beta_1' X + \beta_2' Price_T + \beta_3' Velocity_T, \tag{5.16}$$

where $Price_T$ and $Velocity_T$ correspond to the price and its velocity (i.e. first derivative) at time T, estimated using one of the four methods, and X includes control variables that describe the seller, the product, and the auction features. Such variables include the opening price, auction duration, shipping fee, seller's feedback score, whether or not the auction features a picture, whether the seller is an eBay store, whether the seller is a powerseller, and the number of bids, and average bidders rating (measured at time T). The inclusion of the control variables accounts for wide variability in price that results from different product features, seller credentials, and auction setting. Since this information is observable, we include it in our forecasting model.

Our goal is to compare the impact of different methods for approximating price and price dynamics (i.e. $Price_T$ and $Velocity_T$) on the forecast accuracy. We therefore estimate model (5.16) four times, each time only exchanging the price path estimation method, but leaving everything else the same[3].

To compare the forecasting performance, we use a holdout set. In particular, we split the auctions into a training and a holdout set, each consisting of 50% of the auctions. Model parameters are estimated using the training set, and then predictive accuracy is measured by the Mean Absolute Percentage Error (MAPE) of the auctions in the holdout set which is defined to be $MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i-\widehat{y_i}}{y_i}\right|$, where y_i and $\widehat{y_i}$ denote the true and estimated final price in auction i, respectively. We also study the robustness of our results to different forecasting windows by changing T.

Figure 5.4 shows the results. We see that the Beta model and monotone splines produce the most accurate forecasts. Their accuracy also improves fastest as the time horizon shortens (i.e. as T becomes smaller and smaller, closer to the auction-

[3] Recall that the Beta model operates on normalized bid times and amounts. Thus, in order to estimate $Price_T$ and $Velocity_T$, we first normalize the bid amounts and bid timings to $[0, 1]$ scale, fit the Beta model, and then obtain the predictions via reverse-transformation.

end). While both the Beta model and monotone splines produce forecasts of similar quality, recall the extra computational burden necessary for monotone splines (44 minutes vs. 19 seconds).

Both p-splines and the 4-member growth model family result in poorer forecasting performance. One explanation is that both methods result in poorer model-fit (as outlined in Section 5.4.1), and as a consequence, the predictors for the forecasting model do not accurately reflect the true price and dynamics at time T.

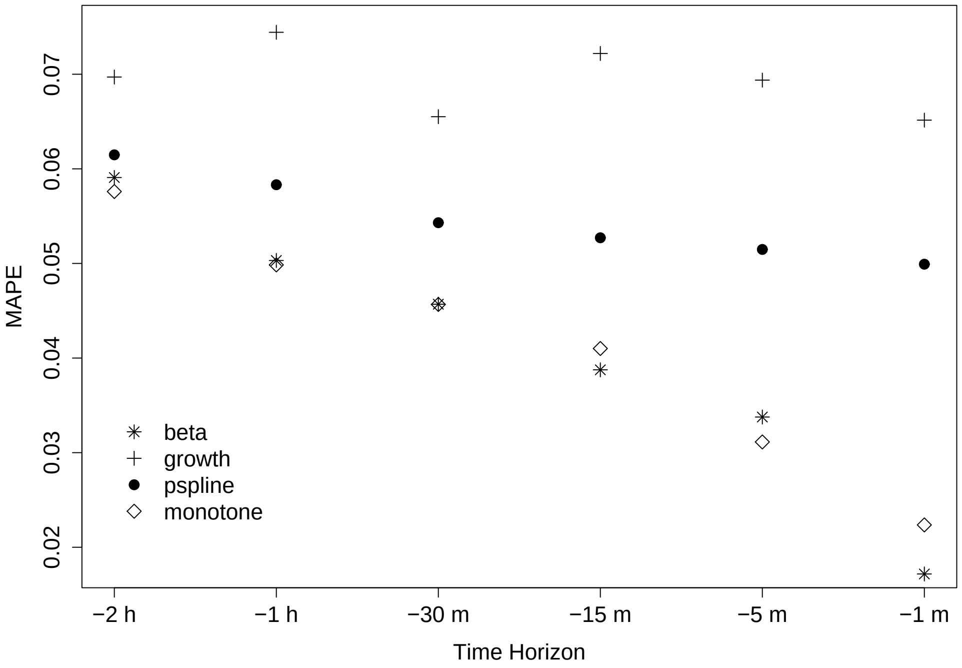


Figure 5.4: Comparison of forecasting accuracy for different time horizons.

5.5 Conclusion

We have introduced a two-parameter model for approximating price paths in online auctions. The Beta model combines the strengths of p-splines or monotone splines models, such as monotonicity and computational efficiency, with additional properties that make it especially useful in the online auction context. It adequately captures the price path and its dynamics and measures pairwise distances between price paths or price dynamics curves in a straightforward manner as shown in Chapter 4. Moreover, the Beta model can be used to characterize auctions in terms of price growth and to summarize the bid timing distribution.

The Beta model is parsimonious, yet very flexible for capturing a wide range of price paths. It is computationally cheap to estimate, and provides good fit both in terms of the bid amount and the bid timing. Our empirical comparison with competing models shows the advantages of the Beta models for model fitting as well as for accurately forecasting the final price of ongoing auctions.

The Beta model can be used for several practical purposes. We have discussed and illustrated the power of the Beta model for forecasting ongoing auctions. We show in Chapter 2 that accurate and efficient forecasting models can be used for making automated bidding decisions. The Beta model has the potential to increase the accuracy of forecasting models significantly due to its ability to measure similarity between price paths. For instance, Chapter 4 show that using similarities between auction price paths can lead to much improved forecasts using a K-nearest neighbor context. But the Beta model can also lead to innovations beyond forecasting. [21] provide evidence that price dynamics can proxy for competition between bidders across auctions. We are curious to learn about additional applications of the Beta model in the near future.

Chapter 6

Pricing and Sales Person Decision Making: An Exploratory Analysis

6.1 Introduction

As the rapid development in company which accompanies increasing production size and customer population, setting an appropriate price which guarantees customer satisfaction yet yields acceptable profit margin has been a difficult task for many business. AMR Research (2004) stated that improved pricing can yield 20%-35% reduction in waste or unused inventory, 2%-4% increase in corporate revenues, and 1%-3% increase in profit. Data mining tools are been employed in this offline business setting for help understand business problems and make informed decisions. One example is the employment of decision support tools (DST).

Retail chains (such as apparel retailers *The Gap*), airlines and hotels often face an extremely difficult task when selecting prices for hundreds of products/services over hundreds of stores nationally and/or internationally. In such complex Business-to-Consumer (B2C) pricing environments, DSTs have proven themselves to be extremely valuable in aiding firms and improving their profits. For instance, *Marriott International Hotels* uses information technology (IT) and DSTs in demand forecasting and scientific pricing optimization to determine the price of every bed in each of their properties; [76] mentions Marriott's annual profit increase for individual hotels totaled $86 million after the rollout of their in-house developed pricing and revenue management system in 2004.

The computation power used to collect vast amounts of data and run in real time large statistical analysis and optimization routines is all being done to help uncover the holy grail of pricing: a customer's maximum willingness-to-pay (WTP). Customer WTP is often endogenously determined - part of the process of determining it relies on observable traits (e.g., price of comparable products, customer's purchase history including past transaction prices and purchase quantities, market indicators such as seasonal effects), which can be captured and modeled in a decision tool. Other parts of the WTP formation process depends on unobservable traits that speak to how a customer perceives/ internalizes a price quote and reacts to it (e.g., concepts of fairness, [65], anchoring and adjustment, [64], framing of the price quote, [94]). While both observable and unobservable factors may exist and hence be useful in determining customer WTP in B2C markets, the relatively small dollar spend of each customer coupled with the large number of customers present in the market generally imply that firms can ignore the unobservable traits and still make reasonable pricing decisions that are implementable. The same cannot be said for B2B markets; and it is on these markets that our research is focused.

Pricing in B2B settings typically is done by sales people (henceforth 'salesreps'), who are in charge of managing the (relatively) large accounts of and relations with several business customers. Hence, in contrast to B2C settings, the typical B2B pricing environment relies more heavily on personal relations and human interactions, whereby the salesrep is entrusted with determining the impact of the unobservable customer traits on the customer's WTP. To emphasize the human involvement, we hereby refer to B2B settings by H2H (Human-to-Human) hereafter. For example, the salesrep must assess if a customer will find a price to be fair (whether or not it is a price that is justified by current market conditions), how and on what the customer anchors his willingness to pay (e.g., the past price paid or possibly a competitor's current price), the strength of the relationship between salesrep and customer and hence whether a customer will trust a quoted price as being reasonable, how customer reacts to price increases, etc. These intangible pieces of information are only known to the salesrep and as such are very hard to incorporate into a DST. Thus, while a DST can gather information across hundreds of sales people, products and markets, and is able to make better aggregate predictions about demand, an experienced sales person may have a better "sense" for individual customers and hence may rightfully reject DST price recommendations as inappropriate for a particular customer.

If we were able to view salesreps rejection of/deviation from price recommendations as only improving our knowledge of customer WTP, it is possible to adjust the demand forecast and pricing algorithms in DSTs to properly incorporate the salesrep's better informed action. However, it is well documented that being an "expert" does not always imply better decisions [93]. Salesreps themselves are human, and hence are subject to their own decision biases and judgment heuristics (e.g., memory bias, [93], satisficing behavior, [77], status quo bias, [66]). A salesreps tacit knowledge of the customer (demand), coupled with his own decision heuristics can be significantly different than those that of a scientific price formation process. As a consequence, it is not clear whether pricing recommendations to salesreps in

H2H markets as they have in B2C e-commerce.

Academic literature on pricing features surprisingly little research on H2H pricing, and even less so on behavior of sales people in this context. While there is a large literature on pricing in economics, marketing, or operations management for B2C markets [94; 20; 101], the human element as a final decision maker and its influence on future customer demand is often neglected. We set out to study what salesreps are considering when determining sales price with a particular customer in this lack of study. In particular, we are curious if salesreps will incorporate a price recommendation when it is presented to them; in other words, are these price recommendations having an impact?

Consider a sales person in charge of selling a single product to a specific customer. The pricing process of the sales person can be expressed as a "mental model", by which we mean the entire thought process that the sales person uses to arrive at the price decision. This thought process may be driven in part by factual data such as the unit cost of the product at the time of transaction, the the price recommended to the sales person for this customer. The thought process, and thus the mental model, may also involve other factors, some observable and some unobservable, such as customer's purchase history, a sales person's attitude towards risk in closing a deal (vs. losing the sales), information obtained from the customer during a sales transaction, a target sales quota self-selected by the sales person, tendency of the customer to negotiate, and so on. We aim at identifying important factors that determine a sales person's price and thus salesreps' mental decision model in a H2H setting. This study helps us understand the pricing process, in general, and outcome of a sales transaction, in particular.

This paper is organized as follows. In Section 6.2, we introduce the data used in this study and discuss the necessity of data reduction procedure prior to the analysis. In Section 6.3, we discuss factors that are potentially important for salesreps' mental models and select the mental model that best mimics salesrep' price decisions. We emphasize the significance of the recommendation from DST in Section 6.4. In Section 6.5, we summarize our findings and discuss future research directions.

6.2 Data Processing

6.2.1 Data Description

The data used in this study is the authentic transaction records of one of the leading grocery product distributors during a 20 month period (January 2007 to August 2008). Each transaction record contains information about the involved sales division ID, salesrep, customer, and product ID, product category, commodity flag, quantity ordered, invoice date, cost, transacted price, and for many transactions, the recommended price. We explain each piece of information in detail.

The grocery product distributor consists of 17 sales (geographic) divisions across the country. Each division is assigned a unique sales division ID, has its own bonus system and provides training to salesreps regarding making price decisions and using DST.

Within each division, each salerep is assigned a unique salesrep ID. Salesreps have direct interactions (e.g. phone calls or personal visits) with customers (e.g. hotels, restaurants, etc.) each of whom is assigned a unique customer ID, responding to inquiries for certain products and making final price decision for each transaction. All salesreps and customers are referred to by IDs; no personal information about them is available to us.

Each transaction record includes a product ID, the category and commodity flag for the transacted product, and the quantity ordered. Customers may order several products during one interaction with salesreps, each of which will generate a separate transaction record in the data set. Product category is a higher hierarchy of products, examples for some product categories include fruit or frozen cheese. Commodity flag is a binary flag used in the system to distinguish perishable products (commodity) from non-perishable products (non-commodity).

We also know the unit cost and transacted price for each transaction. Cost is the unit production cost for each product plus a certain margin. The transacted price determined by the salesrep is and should always be at least as high as the cost, which guarantees a positive profit margin.

The information above is available for all transactions. In addition, for some transactions (approximately 50% of all transactions in the data set), price recommendation is available to the salesreps and recorded in the data set. This recommendation is generated by DST as the result of a complicated dynamic optimization process and made specifically for the salesrep-customer-product triplet (hereafter referred to as *triplet*) under the specific market condition. Every weekend, DST makes price recommendations for triplets who are expected to experience new transactions in the following week. Transactions occurred in the following week involving those triplets are then provided with such price recommendation; yet salesreps can ignore/overwrite the recommendation. For all other transactions, price recommendation is absent.

6.2.2 Data Reduction

Not all the transactions in the data set contribute to our analysis. For reasons explained below, we perform significant data reduction procedure and focus our analysis on the remaining part.

In the complete data set, each triplet has made 25 transactions in average in the 18 months period. Nevertheless, the number varies greatly across triplets, ranging from 1 to 446. For triplets with too few transactions, a mature mental model might have not been established in the salesrep's mind. Due to this concern, we only keep triplets with more than 10 transactions in the analysis.

The remaining transactions involves 132 product categories which generate total profits of $20,107,234. Based on ABC analysis [95], we find that the top 88 product categories take up 99% of the total profit as well as 99% of the transactions. To focus on profitable products, we exclude transactions for the other 44 categeries from the analysis.

The goal of our study is to understand the salesreps' mental model for making price decisions. The mental model is influenced by the magnitude of cost change. Previous studies [72] show that in case of small or no change in the cost, little price adjustment should be made. This conclusion is also supported by our data. We find that salesreps usually make no price adjustment when the cost change is between -2% and 3%[1]in two consecutive transactions for a customer regarding a certain product. In other words, salesreps' mental model is simple in case of small or no cost change; that is $Price = the\ Last\ Transacted\ Price$. On the contrary, when the cost change is outside the range of [-2%, 3%], salesreps tend to make price adjustments in response. In this study, we focus on salesreps' mental model when facing cost changes in non-negligible scale, and thus exclude transactions with cost change between -2% and 3% from the analysis.

Finally, we exclude the first two transactions for each triplet from the analysis. The first two transactions for each triplet correspond to transactions with new customers or existing customers but new products for which salesreps may have different mental models in order to win the business [18]. Excluding these transactions removes the impact of the complication.

As a result of the data reduction procedure, we end up with 962,650 transactions which includes 1,167 salesreps, 13,863 customers, 36,538 products from 88 product categories. We further divide the data set into two: one consists of transactions with recommended price, and the other one consists of those without. The first data set takes up 41.34% transactions and includes 180,244 triplets while 123,065 triplets are there in the other data set. We make use of the first data set to investigate how sales people make pricing decisions in existence of price recommendation in H2H business (Section 6.3). And we use the second data set for comparison analysis

[1] This means the cost increase is less than 2% or the cost decrease is less than 3%.

based on which we try to understand how salesreps make price decisions differently with and without price recommendation (Section 6.4). Details about the two data sets can be found in Table A.3 in Appendix A.

6.3 Mining Mental Models

6.3.1 Building Mental Models

We aim at identifying important factors that determine salesreps' price, i.e. their mental models. In other words, our research questions is: what factors might influence the current price P_t a salesrep offers given that the last price he charged (to the same customer and product) was P_{t-1}. As we can see, our basic research object is each *triplet*.

As there is no existing research on the topic, we use a combination of economic theory (to develop candidate variables) and data mining (to elicit the most reasonable model). Salesreps' decisions on price decisions are influenced by many conceptually important factors, some observable to us from the data and some not. For example, we observe whether the cost of the product has changed; how frequently a customer has purchased the product from a salesrep; and what is the total value of the goods the customer is buying. At the same time, there are unobservable factors (unobservable from the data available to us) such as how the customer has reacted to price increases; or whether the salesreps trust the recommended price. In the following, we create a variety of features to account for these observable factors and then build a sequence of regression models that digs deeper and deeper into the sales persons mind. In particular, we build a sequence of mental models to explain the observed price adjustment salesreps make in two consecutive transactions, i.e. $P_t - P_{t-1}$, for a triplet by adding in potential predictors sequentially. The change in models' explanatory and prediction capability reflects the importance of associated predictor to the mental model.

The first predictor for our mental model is the cost change in two consecutive transactions for the triplet. Economic theories state that cost change is the fundamental drive for and explains most part of price change. We use this model as the baseline model, referred as model (1).

Model (2) includes commodityflag in addition to cost change. This predictor takes the value of 1 for commodity and 0 for non-commodity. Commodities are perishable products, for which Salesreps are expected to make appropriate price adjustments for timely sells. For instance, they might keep price stable in case of cost increases to make a sell.

Salesreps behave differently for upwards or downwards cost change. Salesreps are motivated to charge high prices (or margins) since the revenue (or profit) they generated has an impact on their salary (or bonus). We are informed by some salesreps that smaller price decrease is usually made for a certain level of cost decreases while larger price increase is often associated with the cost increases of the same amount. Model (3) in the sequence therefore includes the predictor - the sign of cost change which takes the value of 1 in case of cost increase and 0 otherwise - in addition to predictors in (2).

Besides the sign of cost change, size of cost change can also have an impact on salesreps' price decisions. Salesreps might not make price adjustments when facing a 5% cost increase, but they are much more likely to respond to a cost increase of 50%. From our data, we find 10% is an appropriate cutoff for cost change, above which (in either direction) salesreps have great chance to make corresponding price adjustments. We therefore refer to cost increase or decrease bigger than 10% as big cost change, and cost change less than that as medium cost change. Note that Transactions with cost changes that are between -2% and 3% have already been excluded in the data reduction step. The size of cost change is the additional predictor for model (4).

Salesreps should look beyond the two consecutive transactions when making price decisions. If the cost has been rising continuously and the salesreps have made price increase accordingly, customers might be scared away in fear of future higher prices. Experienced salesrep therefore are expected to take this into consideration. We define cost trend to be the existing of cost change in the same direction in two consecutive periods (three consecutive transactions) for a triple. For example, if the cost goes down continuously in three consecutive transactions for a triplet, the last two transactions are on a downward cost trend. Model (5) includes cost trend which takes the value of 1 if there exists cost trend in either direction and 0 otherwise as an additional predictor to capture the effect.

Salesreps might consider giving a long time customer lower price to compensate his/her loyalty or higher price since they're not as afraid of losing him as for new customers. We use repeated purchase for a triplet, which is measured by the cumulative number of transaction, to capture the length of relationship, and include it as a predictor in model (6).

Besides loyal customers, incentives are possibly also offered to bundles of large value. By bundles, we mean the transactions that occur during one interaction (phone call or personal visits) between a salesrep and a customer. Since transactions for different products are recorded separately even if the customers ordered them altogether, it makes sense to treat them as a bundle whose total value or quantities might influence salesrep's price decision. To capture this effect, we include total dollar value (in log scale) of the bundle that each transaction belongs to as a predictor in model (7).

We have proposed a sequence of 7 models. Each model include one additional predictor comparing to the previous model. Moreover, in each model, we include all one-way interaction among its predictors. The inclusion of the interaction terms helps capture the interplay of two factors.

6.3.1.1 Effect of Price Recommendation

Besides influential factors discussed above, price recommendations are also available to salesreps for transactions in our first data set (see Section 6.2 for detailed explanation). On one hand, as the output from DST which gathers information across hundreds of sales people, price recommendation is a theoretically optimized price. On the other hand, however, experienced salesreps possibly have a better "sense" for individual customers and hence may rightfully reject price recommendations from DST and make more appropriately personalized price decisions.

Therefore, it is not very clear if salesreps will and should incorporate a price recommendation when it is presented to them; and if the recommendations do have an impact, how big it is.

Because price recommendation is available in this data set, we can investigate the impact of price recommendation on salesreps' mental models in the following way. We define recommended price adjustments to be the difference between the current price recommendation and the last transacted price, i.e. $Recom_t - P_{t-1}$. For each model we propose, we build a paired model which includes, in addition to previous predictors, the recommended price adjustment as well as one way interaction between it and all other predictors. Then we compare each model to its counterpart. This comparison answers the questions that whether the recommended price adjustment has an impact on salesreps' mental models and if so, how much the impact is. We refer to the model without recommendations as model 1(a)-7(a), and their counterparts as model 1(b)-7(b).

So far, we have build 14 models (7 pairs). We summarize the models in Table 6.1[2] . To find the true mental models of salesreps and the factors that salesreps truly anchor on, we compare models' explanatory (explaining the observed behaviors of salesreps' price adjustments) and prediction (predicting price adjustments for future transactions) capability. We discuss model comparison and selection in detail in the next subsection.

[2] We have also investigated models by adding in predictors in different orders, all of the alternative analysis gives identical conclusion as our current analysis.

Table 6.1: Summary of All Models

Model	a	b
1	cost change	(a) + recommended price adjustment
2	+ commodityflag	
3	+ sign of cost change	
4	+ size of cost change	
5	+ cost trend	
6	+ repeated purchase	
7	+ values of bundle	

6.3.2 Model Selection

In order to identify factors that are included in salesreps' mental model, we compare above models in terms of their explanatory and predictive capability. High R^2 and low Bayesian Information Criterion (BIC) values infer that the model fits the data well and predictors in the model well explains observed price changes. However, good model-fit can possibly be the result of overfitting which weakens the generalizability of the model onto new data. Therefore, we also study models' capability of predicting salesreps' price adjustments on "future" transactions.

To that end, we divide our data set into a training set (70% of the transactions), and a holdout set (remaining 30% of the transactions). we first estimate our models on the training set using ordinary least square method; results of model-fit measured by R^2 and BIC are discussed below. We then apply the estimated models to the holdout set to gauge their predictive capabilities measured by $RMSE$ which is defined as $RMSE =: \sqrt{avg(|PC_i - \hat{PC}_i|^2)}$ where PC_i and $\hat{PC}_i$ represent observed and predicted price adjustment for the i^{th} transaction respectively.

A model's $RMSE$ value measures the average difference in dollars between the predicted and observed price adjustment. Moreover, the changes in R^2, BIC and $RMSE$ between any pair of models measure the additional explanatory and prediction power brought by the inclusion of recommended price adjustments on the salesreps' mental models when the recommendation is given.

6.3.2.1 Model Fit Comparison

We compare the explanatory capability in terms of R^2 and BIC on the training set. The top panel in Figure 6.1 plots the R^2 for all models and the bottom panel plots model BICs. The blue lines represent models without price recommendation, i.e. model 1(a)-7(a); and the red lines represent model with recommended price adjustments, i.e. model 1(b)-7(b).

We can see that the predictor cost change itself explains 77.5% of salesreps' price adjustments, which justifies our intuition that cost change is the most fundamental factor influencing the mental model. Moreover, as more predictors added into the model, the value of R^2 increases and BIC goes down in either blue or red line, both indicate a better model-fit. R^2 increases and BIC decreases at a steady speed as we add in commodityflag, sign and size of cost change sequentially, implying the first four predictors matter. After that, any additional predictor (cost trend, repeated purchase, and values of bundle) has only marginal effect on the model-fit. We also see that the red line lies above the blue lines for any model in sequence, which clearly indicates that recommended price adjustment is a valuable predictor for the mental model.

6.3.2.2 Prediction Capability Comparison

In addition to models' explanatory capability, we also compare models' capability to predict salesrep's price adjustments on the holdout set. We can see from Figure 6.2 that model 4 which includes cost change, commodityflag, sign and size of cost change can best predict salesreps' price decisions. The increasing in prediction accuracy brought by other predictors is negligible.

Comparing the red with the blue lines, we also see that the inclusion of recommended price adjustments makes us 8 cents (from \$1.65 to \$1.57 in model 4) closer to the observed price adjustments. This implies that the salesreps are taking into consideration of the price recommendation when it is presented to them.

6.3.3 Model Interpretation

From the discussion above, we can see that model 4(b) from Table 6.1 has the best explanatory and predictive capability, thus best mimics salereps' mental model. Moreover, DST price recommendation has an impact on salesreps' price decisions when presented to them. Specifically, our analysis indicates that:

• Cost change is the most important determinant of price adjustments.

• Sales people behave differently when making price adjustments for different product types (commodities vs. non-commodities, which correspond to items with very short vs. longer shelf lives).

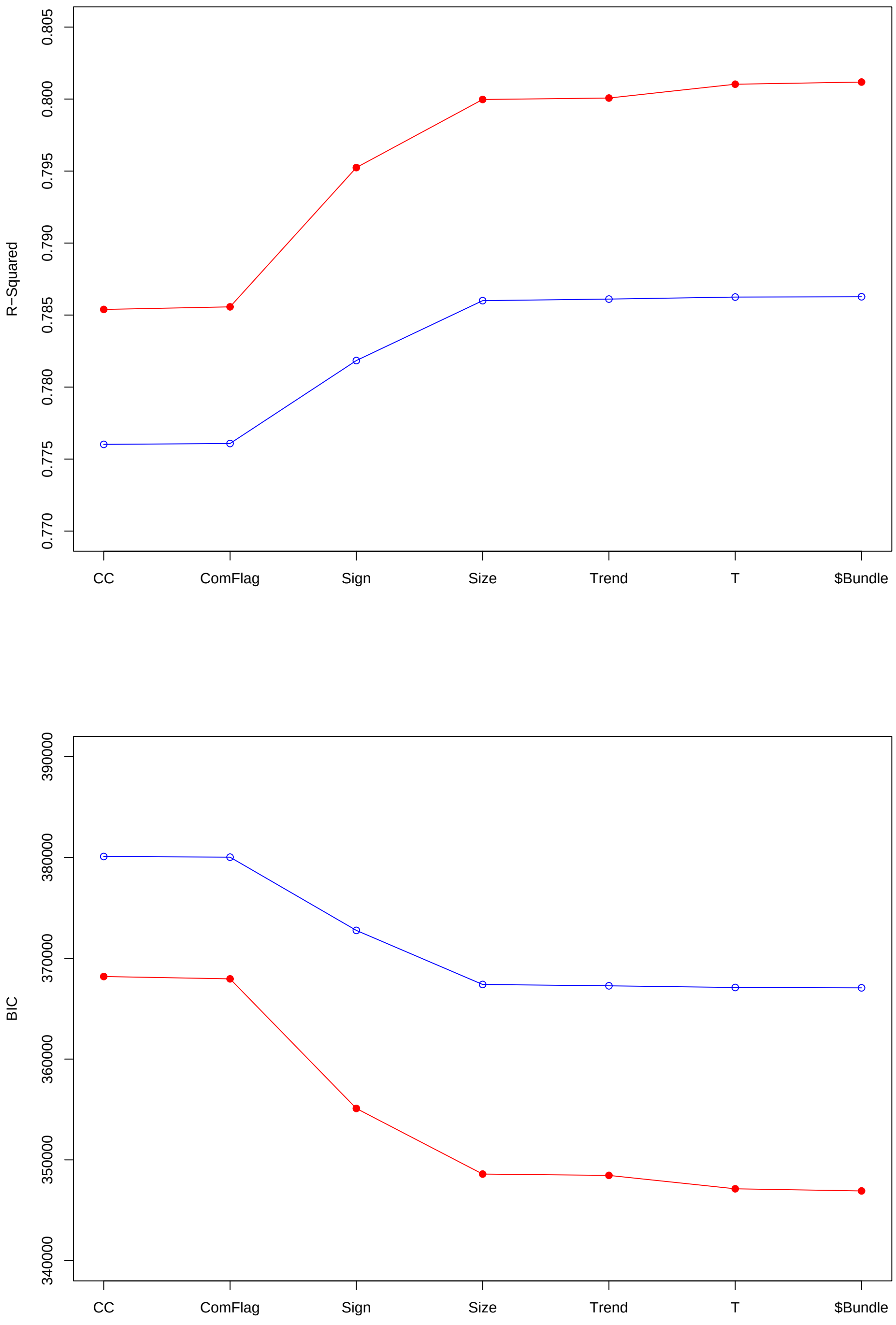


Figure 6.1: Comparison of R^2 and BIC. The blue line represents model 1(a)-7(a) and the red line represents model 1(b)-7(b).

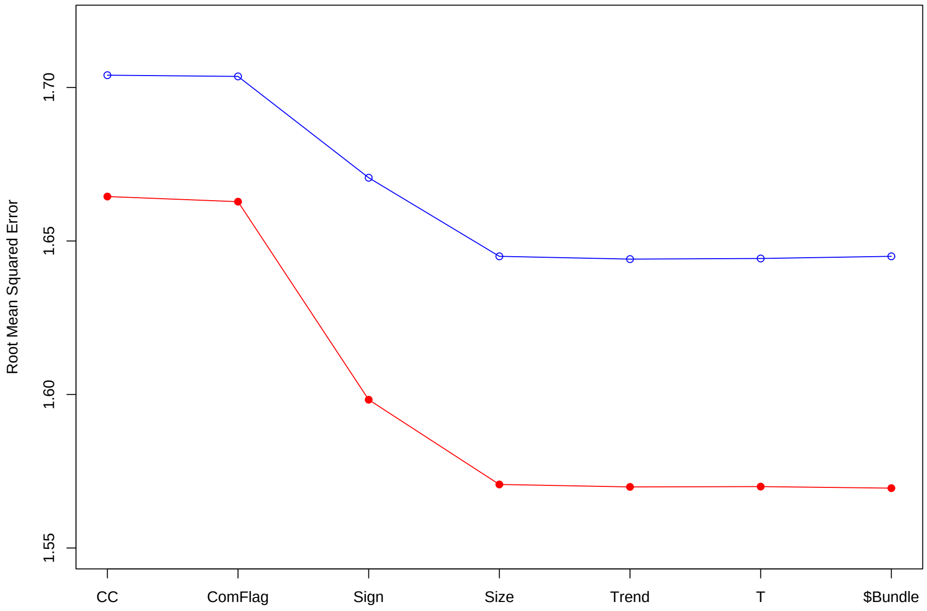


Figure 6.2: Comparison of RMSE. The blue line represents model 1(a)-7(a) and the red line represents model 1(b)-7(b).

- Salesreps' price adjustments are different facing cost increases from facing cost decreases. Furthermore, they also differ for cost change in different sizes (big vs. medium).

- In the presence of cost changes and commodity flag, the impacts of repeat purchases, markets trends in cost, and the value of the purchase bundle are very small.

- Salesreps anchor on price recommendations from DST. Incorporating recommended price adjustments into models helps explaining observed salesreps' price decision and predicting future prices. In particular, when present, recommended price adjustment helps us better predict future price adjustments in the amount of 8 cents per transaction.

6.4 Effects of Existence of Price Recommendation

Every weekend, DST updates the price recommendation for triplets that are expected to experience new transactions in the following week. For other triplets, the price recommendation is left blank.

From our analysis in the last section, we can see that salesreps anchor on recommended price adjustments if presented when they make price decisions. However, the fact that price recommendations have an impact on salesreps' mental models does not necessarily imply the existence of price recommendation leads to better price decision. The reason is as follows. No matter how experienced salesreps are, they are all human beings who are subject to cognitive bias[3].It has been found that people's decision or judgement is generally affected by irrelevant information [40]. Consequently, it is possible that price recommendation is irrelevant information for making appropriate price decisions, and salesreps mistakenly takes that into consideration.

To investigate whether or not price recommendation leads to better mental models, we conduct the following comparison study. First, we study salesreps' mental model for the setting where price recommendation is absent. Then we compare the results with settings where price recommendation exists.

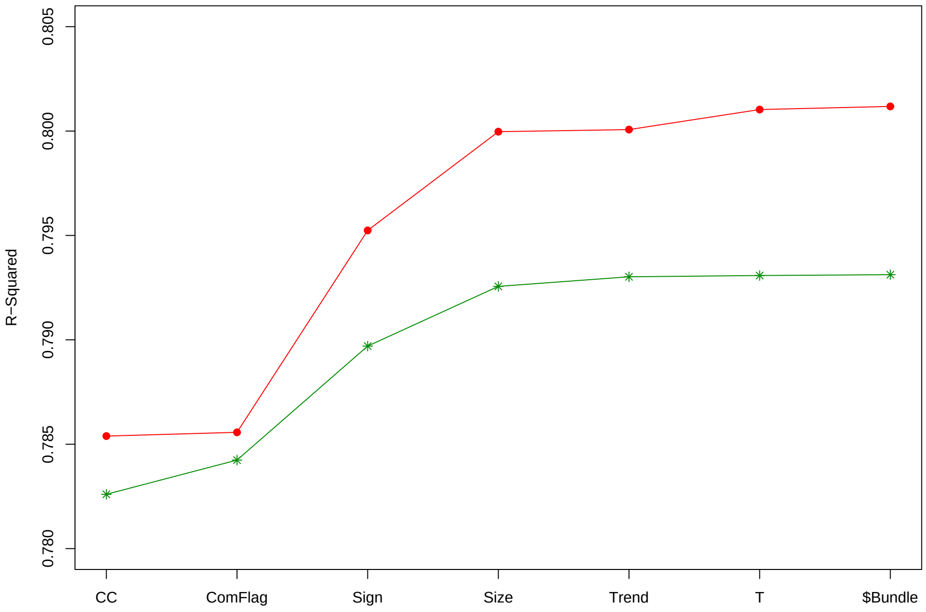
We have two data sets. The first one consists of transactions with price recommendation and has been used for previous analysis. The second data set includes transactions without price recommendation (see Section 6.2 for detailed explana-

[3] A cognitive bias is a person's tendency to make errors in judgment based on cognitive factors.

tion). The two data sets take up 41.33% and 58.67% of combined transactions respectively, whose sizes are somewhat comparable. Moreover, there are 88,441 common triplets in the two data sets, which are 50% and 72% of all triplets in them respectively. The large portion of common triplets guarantees the fairness of our comparison. Any difference in the results should not be caused by the intrinsic difference in the objects involved in the transactions but due to heterogeneity in salesreps' mental models.

For transactions in the second data set, salesreps make price decisions in absence of DST price recommendation. We feed the data into model 1(a)-7(a) from Table 6.1, and compare models' R^2 and $RMSE$ to find the best model, i.e. salesreps' mental model, in this setting. The green line in Figure 6.3 plots R^2 (top panel) and $RMSE$ (bottom panel) for the seven models. We can clearly see that model 4(a) is the best model in terms of both model's explanatory and predictive capability. Remember we have found previously that salesreps' mental model given price recommendation is model 4(b). Therefore, we claim that the existence of the price recommendation does not change salesreps' mental model but adding in one additional anchoring factor which is recommended price adjustments. In other words, except for price recommendation, salesreps anchor on the same factors when making price decision no matter whether DST provides them with recommended price information or not. The other important factors that salesreps anchor on are cost change, commodityflag, sign and size of cost change.

To make a clear comparison, we plot R^2 and $RMSE$ for model 1(b)-7(b) for transactions from the first data set (where price recommendation is given) using



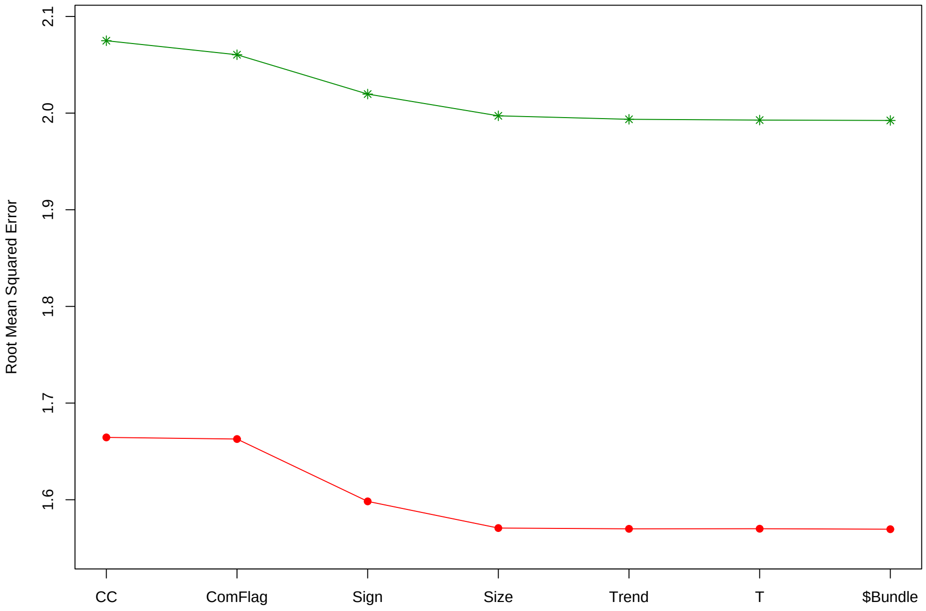


Figure 6.3: Model comparison for transactions with or without price recommendation. The red line plots model 1(b)-7(b), and represents transactions from the first data set which includes price recommendation information. The green line plots model 1(a)-7(a), and represents transactions from the second data set where price recommendation is not given.

red lines on the same graph. Compare R^2 (top panel) and $RMSE$ (bottom panel) for the red line with that for the green line, we can see that the inclusion of the price recommendation improves both model-fit/explanatory and predictive capability. The R^2 for model 4 in the red line is slightly higher (0.007) than that in the green line; and our prediction for salesreps' price adjustments are on average \$0.45 ($\$2 - \$1.55$ for model 4) closer to the truth in the red line which corresponds to the case with DST price recommendation. The higher $RMSE$ in the green line implies that it is more difficult to describe how Salesreps make decisions in the absence of price recommendation. This is because the existence of recommended price tunes down the roles played by unobservable factors in salesreps' decision making process. In the absence of recommendation, unobservable factors, such as customers' reaction to price increase or competitors' price adjustments, weigh more heavily in their price decision process.

6.5 Conclusions

Different from B2C settings where DST has been adopted and proven to be extremely valuable in aiding firms and improving their profits, pricing in H2H settings relies more heavily on personal relations and human interactions. In such settings, salesreps are entrusted with taking charge of managing the large accounts of and relations with business customers and making final price decisions. Although DST has been adopted to provide price recommendations in such setting in practice by a small number of companies, it is not very clear whether salesreps are under the influence of such recommendation and if so, whether it helps salesreps make better price decisions.

Salesreps are influenced by many conceptually different factors when making price decisions, some observable and some not. In this study, we build a sequence of mental models that dig deeper and deeper into salesreps' mind, and use model selection procedure to identify key (observable) factors that influence their pricing decisions. We find that salesreps anchor on cost change, commodity flag, sign and size of cost change, and price recommendation if presented, when making price decisions. We also find that price recommendation weakens the influence played by unobservable factors. Without recommendation, salesreps are influenced more by unobservable factors, which makes their price decisions more difficult to explain and predict. All these findings suggest there is a future for pricing tools in these H2H settings!

Chapter 7

Future Research

In the following sections, we describe future research directions for the studies presented in this dissertation.

7.1 Data Driven Bidding Strategy

In Chapter 2, we propose an automated and data-driven bidding strategy that provides bidders with complete decision guides. Our current approach only consider auctions that close within the given prediction window $[T,(T+1)]$, and we only predict the final price of auctions that end within that interval. To relax the restriction, one can roll the model one additional time period forward to make predictions at $T+2$, based on the predicted values at $T+1$; so that bidding on auctions that close later is allowed. However, predictions two time periods into the future (i.e. $T \rightarrow T+2$) are more uncertain than predictions only one step forward (i.e. $T \rightarrow T+1$). It is not quite clear how to discount the additional prediction uncertainty in our decision framework.

We consider only single unit auctions in this research, one could expand the scope of our bidding strategy to multi unit auctions. Let us assume that a seller sells n items (of identical product specification and quality) in the same auction; then the bidders with the top n bids each win one item. In order to apply our bidding strategy to this scenario, bidders need to know the lowest transacting bid; that is, they need to predict the price at which the n^{th} item sells. Given a set of relevant bidding records, one solution would be to apply our model to the price of the n^{th} item; that is, we would train our model to predict the lowest transacting bid.

A related issue is the purchasing of more than one unit at a time. If a bidder has demand for more than one unit, then the current bidding strategy could still be employed if the the bidder has no time constraints and decides to bid sequentially and if the bidder's WTP is the same for all units (assuming that there is unlimited supply which is realistic for many of the items sold on eBay). However, if the bidder needs to purchase n units within a short period of time and places m bids simultaneously, then each bid should be discounted relative to the size of m; on the other hand, bids may be inflated with decreasing time periods to assure that all units are available on time. This calculation may change further for varying WTP distributions. All-in-all, there are many opportunities for future research and we hope to inspire some of it with this study.

7.2 Model Selection for Improved Forecasting

For bidders, making informed bidding decision requires forecasting models that works well across an entire range of time-increments. In Chapter 2, we investigate model selection criteria to find such a forecasting model. We make use of various summary statistics of conventional model selection criteria, AIC and BIC, over a time window to select models, and comparing models' prediction capability of models selected under different criteria.

Our study is conducted in the context of online auctions, which are characterized by events that arrive at very irregularly-spaced time intervals: since sellers determine the end of an auction, bidders have to make decisions about events that are sometimes very dense (i.e. several auctions closing within only seconds of each other) and other times very sparse (e.g. at night when only very few auctions close). It is this irregular spacing that calls for forecasting models that perform well in the short-term as well as in the long-run; in fact, as we have pointed out, it calls for models that perform well over a continuous distribution of time-increments.

While we derive the market-model within the context of eBay auctions, there are other examples where similar models are called for. In fact, similar models could be useful for markets that have similar characteristics (i.e. competition between individual market occurrences that are unevenly spaced and that exhibit different dynamics). Examples include other C2C auctions (e.g. uBid, Prosper, Overstock), auctions for fine art [82], B2B auctions such as `govdeals.com` or `liquidation.com`. Similar characteristics can also be found on traditional stock markets (in particular, derivatives markets) or virtual markets (e.g. [90]), and *Yahoo! Movies* or *CNET.com* where user ratings or blog postings are often marked by time periods of little activity, followed by times of very dense information arrival. It would be interesting to compare the performance of different model selection criteria, and to see whether our conclusion, which states that the extreme of AIC or BIC selects the best model, holds for those settings.

We can also extend our study to model selection criteria other than AIC and

BIC. Traditional statistical theories have developed many model selection criteria, such as R^2, Mallows' Cp, and Deviance information criterion. A more complete investigation over those criteria may help find better forecasting models. Besides the list of model selection criteria, an extension can also be made to ways of summarizing those criteria over an interval. In this study, we summarize via summary statistics including mean, median, standard deviation, minimum and maximum. One can also use other statistics, such as quartiles or interquartile range, to summarize the distribution of criteria over a given interval.

7.3 Weighted Forecasting of Closing Prices

Chapter 4 proposes a novel functional KNN forecaster for forecasting the final price of an ongoing online auction. To accomplish this, we first introduce a functional representation of the auction's price path which allows measuring distances between two paths via KL distance. Then, we define different distance metrics for other data-types and combine them and KL distance into a single distance metric. Finally, we apply K-Nearest Neighbor algorithm with carefully selected distance metric and number of neighbors for making forecasts. There are many ways to expand upon this area of research.

One extension is to search for alternative ways of defining distance metrics. Currently, we scale distance metrics for different information sources to achieve equal weighing across all metrics, one could alternatively assign individual weights to individual metrics and then optimize the weights; Or we can construct the overall distance metric by applying principle component analysis to all individual distance metrics. There are also alternative ways to define distances for different data types. For example, for categorical data we can define several levels of category “similarity”, such as “US brand”. Then, the distance between items can be set to 0.5 for “similar categories” (e.g., laptops of a US brand) or 1 for categories that are more different.

We can also investigate alternate ways for making weighted forecasts. One possibility is to expand upon classical linear regression and regression trees, i.e. to develop weighted regression or weighted tree models, which might lead to forecasting advantages especially for heterogeneous data. This extension, however, requires defining weights for each sample, thus should be combined with the other extensions suggested above.

7.4 A Flexible Model for Price Dynamics in Online Auctions

Section 5 explores various properties of a parsimonious parametric Beta model as a representation of auction price paths. We develop an algorithm to estimate the model by minimizing residual errors in both bid time and bid amount dimensions simultaneously. In our current definition of a residual, we weigh the x and y directions equally (the weight is 0.5) because we have no particular reason to prioritize either direction. Alternatively, one could overweigh the x or y direction if the bidding time or price level, correspondingly, is of special interest. Comparing price paths resulting from different weights in residuals can help gain insights about roles played by bidding time and bidding amounts in determining price paths.

While we show that the beta model has overall better model-fit compared to p-splines, monotone splines, and 4-member growth models, the others might be better for individual auction. One future research direction could be to investigate which kind of auctions can be best described by which model. One possible way to get this done is to run all models for every auction, categorize them based on the model that fits them the best, find the common characteristics of auctions in each group, and link that to the properties of corresponding model.

7.5 Pricing and Sales Person Decision Making

Pricing in B2B settings is typically done by salesreps, thus we refer to such setting by H2H. Salesreps rely on their expertise, knowledge of individual customers, many observable and non-observable information, and possibly price recommendations from DST, to make price quotes. In Chapter 6, we investigate factors influential to salesreps' price formation process with special attention to the impact of DST price recommendations. We find that cost related information, including cost, sign and size of cost change, and types of products (perish commodities or non-commodities), are the most important predictors for salesresps' price decision. Moreover, price recommendation, whenever provided, influence salesreps' decisions in a positive way.

There are many unanswered question, thus research opportunities, for H2H pricing. In a H2H setting, sales people are the ones that interact with the customers and quote them prices. By providing a first look into how salesreps form prices and respond to price recommendations in H2H markets, we do not only show the value of DST, but also open the door to research about designers of DSTs. For instance, one may turn to questions of how we can incorporate our findings into design of DSTs and pricing processes to counter salesrep biases (similar to as is done in [31]).

Another extension to this study is to study the heterogeneity of the salesreps. Our results in this study apply to the general case, or an *average* salesrep; and we should expect very different price formation process for different salesreps. One way to investigate salesreps' heterogeneity is to repeat the analysis on a subset of data which only includes a certain type of salesreps. For example, we can investigate the mental model of salesreps in some sales division. Because each sales division provides its own training regarding DST and has its own bonus system, we expect to see that salesreps' attitudes towards DST price recommendation and anchoring factors are different across divisions.

Appendix A

Data Sets Used in the Study

A.1 eBay Bids Level Data

A.1.1 Palm Pilot M515 PDA data

This dataset includes the complete bidding records for 380 auctions for new Palm Pilot M515 handheld PDA that transacted on eBay between March and May, 2003. Each bidding record includes the auction ID, the starting and closing times and prices, all bids with associated time stamps, and other information such as auction duration, shipping fee, seller's feedback score, whether the seller is a power seller, whether the product is from an eBay store, and whether the auction descriptions include a picture. Table A.1 presents summary statistics for these variables.

A.1.2 Laptop Data

The data set contains information on 4,965 laptop auctions that took place on eBay between May and June, 2004. Table A.2 summarizes the data which include products of a wide variety of makes and models. We can see that the data include over 7 different brands, and for each brand laptops differ further in terms of their memory size, screen size, processor speed, whether they are a new or used product, and whether or not they include an Intel chip or a DVD player.

Table A.1: Description of the Palm auctions. The top panel reports statistics for all continuous variables; the bottom panel reports statistics for all discrete variables.

Variable	Mean (Stdev)	Median	Min	Max
OpeningPrice	\$76.67 (92.45)	\$9.99	\$0.01	\$265
ClosingPrice	\$229.45 (22.00)	\$232.50	\$172.50	\$290
AuctionLength	5.74 (1.79)	7	3	10
NumberOfBids	17.45 (11.23)	17.50	1	54
NumberOfBidders	8.92 (5.13)	9	1	23
ShippingFee	\$15.44 (5.51)	\$15	\$0	\$50
SellerFeedback	545.73 (1787.47)	44	0	27652

Variable	Yes	No
PowerSeller	121(31.84%)	231(60.79%)
eBayStore	117(30.79%)	235(61.84%)
Picture	332(87.37%)	20(5.26%)

The data set also contains information regards auction setting. For instance, Buy-It-Now auctions are listings that have the option of a fixed-price transaction and thus forego the auction mechanism. Over 20% of the laptop auctions included that feature. Moreover, a secret reserve price is a floor price below which the seller is not required to sell. This feature is particularly popular for high-value auctions. Roughly 30% of all laptop auctions make use of the secret reserve price feature.

We can see that auctions in this dataset are of a wide variety in terms of product features and auction setting. This is also reflected in the wide range of number of bids (between 6 and 115), bidders (6 and 30), and closing prices (between \$445 and \$1,000).

Table A.2: Summary statistics of the laptop auctions. The top two panels report statistics for auction features. The bottom three panels report summary statistics on the product characteristics.

Variable	Mean(Stdev)	Median	Min	Max
OpeningPrice	93.31(159.54)	9.99	0.01	900
ClosingPrice	499.22(210.26)	445	200	999.99
AuctionLength	5.00(1.81)	5	3	7
NumberOfBids	21.13(11.05)	19	6	115
NumberOfBidders	9.94(4.20)	9	1	30

Variable	Yes	No
BuyItNow	1027(20.68%)	3938(79.32%)
ReservePrice	1529(30.80%)	3436(69.20%)

Variable	Category
Brand(count)	Dell(1622); Fujitsu(15); Gateway(165); HP(1347); IBM(705); Sony(307); Toshiba(535); Other(229)

Variable	Mean(Stdev)	Median	Min	Max
MemorySize	269.12(157.78)	256	64	2000
ScreenSize	14.03(0.92)	14	12	21
ProcessSpeed	1125.05(728.83)	850	133	3200

Variable	Yes	No
NewProduct	628(12.65%)	4337(87.35%)
IntelChip	4863(97.95%)	102(2.05%)
DvdPlayer	2992(60.26%)	1973(39.74%)

A.2 Transactions Data from A Grocery Products Distributor

We use the transaction data from one of the leading grocery products distributors during in our study. The data set includes all transactions that took place during January 2007 to August 2008. Each transaction record contains information about the involved sales division ID, sales representative (referred to as "salesrep") ID, customer ID, product ID, product category, commodityflag (perishable commodities or non-commodity), quantity ordered, invoice date, cost, transacted price, and for many transactions, the recommended price.

We divide the remaining transactions after performing series of data reduction procedures into two data sets (see Section 6.2.2 for details). The first one (referred to as Set.1) includes transactions with DST price recommendations, and the other data set (referred to Set.2) has no price recommendation information provided for corresponding transactions. We provide summary statistics for both data sets below.

Table A.3: Descriptive statistics of important variables.

Variable	Set.1	Set.2
Num.Transactions	397,939	564,711
Num.Triplet	123,065	180,244
Num.Salesreps	1,138	1,146
Num.Customers	12,461	13,252
Num.Products	14,694	19,834
Num.Categories	88	88
Num.SalesDivisions	17	17
Tot.Revenue	$17,849,496	$25,793,360
Tot.Profit	$1,710,092	$3,037,315
Percentage.Com- -modity.Transactions	66.54%	47.90%
Percentage.Com- -modity.Products	32.98%	27.11%
Percentage.Com- -modity.Categories	19.32%	19.32%

Variable	Set.1	Set.2
Quantity/transaction	1.93(2.82)	1.78(3.42)
Cost/transaction	21.15(18.62)	22.99(18.42)
price/transaction	23.69(20.14)	26.52(20.34)
RecommendedPrice/transaction	25.03(20.95)	26.49(19.73)
Revenue/transaction	$44.85($89.24)	$45.68($117.05)
Profit/transaction	$4.30($9.06)	$5.38($12.91)
Num.transactions/triplet	3.23(3.78)	3.13(3.60)
Revenue/triplet	$145.04($396.33)	$143.10($561.76)
Profit/triplet	$13.90($36.88)	$16.85($57.43)

Appendix B

Simulation Results for Alternate Bidding Schemes

B.1 Alternate Bidding Heuristics

In this section we investigate alternative bidding heuristics to those discussed in Sections 2.5.2.1 (early bidding) and 2.5.2.2 (last minute bidding). These heuristics are based either on price trends that bidders observe in auctions that closed recently, or on strategies to shade bids below what one believes a good is worth. Table B.1 shows the results.

The top panel in Table B.1 shows the results of using recent price trends for making bidding decisions. Assume that a bidder wants to place a bid and that s/he has monitored prices of the n auctions that closed most recently (we choose n=10 here but the results do not change much for different values of n). The bidder then bids the minimum (we also investigate the mean or the maximum) of the n closing prices (as long as the minimum is smaller than his/her WTP). The bidder can place the bid any time before the auction closes. We can see from the table that this heuristic performs worse than early or last minute bidding. It is curious that bidding the mean results in the highest expected surplus (as it increases the chances of winning).

The bottom panel shows the result of early bidding (i.e. bidding on day one), but shading one's bid below what one really thinks the item is worth. In other words, rather than bidding one's WTP, one only bids a fraction, e.g. 90% or 80%. The results show that while shading increases the *average* surplus, it reduces the *expected* surplus as the probability of winning decreases. In fact, shading at 70% or below (not shown here) results in zero expected surplus.

Table B.1: Alternative bidding heuristics.

	RECENT PRICE TRENDS		
Heuristic	p.win	avg.sur	exp.sur
mean	44% (2%)	$8.86 ($0.28)	$3.90
min	11% (0.2%)	$35.11 ($0.22)	$3.86
max	57% (40%)	$2.75 ($2.35)	$1.57

	BID SHADING		
Heuristic	p.win	avg.sur	exp.sur
100%WTP	53% (2%)	$18.85 ($0.57)	$9.90
90%WTP	20% (1%)	$19.20 ($0.50)	$3.80
80%WTP	3.5% (0.4%)	$19.13 ($0.58)	$0.67

B.2 Robustness of Last-Minute Bidding

Last-minute bidders place an incremental bid over the current high-bid and we assume in Section 2.5.2.2 that this increment equals 2%. In practice, this increment could be larger or smaller; it could also be that some last-minute bidders increment not by a percentage of the current price but rather by a fixed amount. Table B.2 investigate that robustness of last-minute bidding to different increment strategies. We can see that the expected surplus is rather unaffected by the increment strategy. Moreover, regardless of the actual strategy chosen, the expected surplus is significantly lower than that of our automated bidding strategy in Table 2.6.

Table B.2: Robustness of last-minute bidding to different increments.

Increment	p.win	avg.sur	exp.sur
$0.50	86.26%	$18.73	$16.16
$1	89.38%	$18.61	$16.63
$2	92.50%	$18.44	$17.06
$5	96.43%	$17.39	$16.77
0.5%	89.82%	$18.58	$16.69
1%	92.34%	$18.47	$17.06
2%	95.19%	$17.97	$17.11
5%	98.03%	$15.97	$15.65

Appendix C

Model Selection Results for Palm Data Set from eBay

In the following we list the complete model selection results for Chapter 3. We refer to each variable according to its index from Table 3.1. We arrange the results by the total number of variables in the model. We start with models containing only one parameter (Table C.1) followed by 2-parameter models (Table C.2 and C.3) and so on. The first column either refers to the variables entering the model ("Var-In") or to the variables leaving the model ("Var-Out"). For instance, Var-In=1 means that the model contains only one variable, variable #1, i.e. Price; on the other hand Var-Out={1,6} means that all variables but #1 and #6 enter the model. We report result for both AIC and BIC. The highlighted number refers to the best model in each column.

Table C.1: 1-Parameter Models

	BIC					
Var-In	BIC_{avg}	BIC_{sd}	BIC_{med}	BIC_{min}	BIC_{max}	$\text{BIC}_{mean+sd}$
1	***-25.776***	25.551	***-18.542***	***-85.033***	***2.344***	***-0.225***
2	76.872	***7.207***	76.907	63.877	86.821	84.079
3	75.497	7.426	75.649	61.191	85.605	82.923
4	77.419	7.341	77.445	63.896	87.132	84.761
5	76.316	8.178	76.137	62.176	87.097	84.494
6	63.389	9.109	64.236	46.861	75.340	72.498
	AIC					
Var-In	AIC_{avg}	AIC_{sd}	AIC_{med}	AIC_{min}	AIC_{max}	$\text{AIC}_{mean+sd}$
1	***-27.690***	25.708	***-20.259***	***-87.215***	***0.564***	***-1.982***
2	74.958	***6.999***	74.967	62.332	84.639	81.958
3	73.584	7.219	73.753	59.646	83.423	80.803
4	75.506	7.134	75.556	62.351	84.950	82.640
5	74.402	7.971	74.204	60.631	84.915	82.373
6	61.475	8.914	62.347	45.316	73.158	70.390

Table C.2: 2-Parameter Models

	BIC					
Var-In	BIC$_{avg}$	BIC$_{sd}$	BIC$_{med}$	BIC$_{min}$	BIC$_{max}$	BIC$_{mean+sd}$
1,2	***-114.085***	57.220	***-93.854***	***-239.185***	***-49.978***	***-56.865***
1,3	-24.199	24.837	-16.351	-82.753	0.632	0.638
1,4	-26.161	27.113	-18.200	-92.598	3.717	0.952
1,5	-30.775	25.800	-25.350	-88.866	-3.988	-4.975
1,6	-29.101	25.695	-22.984	-87.808	-0.685	-3.406
2,3	77.037	***7.158***	77.618	63.960	86.975	84.195
2,4	79.770	7.286	79.843	66.646	89.812	87.056
2,5	78.050	7.849	79.038	62.783	87.630	85.899
2,6	62.095	7.572	61.824	48.378	73.069	69.667
3,4	78.014	7.554	78.307	63.956	88.499	85.569
3,5	76.641	8.712	77.027	62.426	88.427	85.353
3,6	64.874	9.077	66.068	47.237	76.699	73.951
4,5	79.075	8.232	79.092	64.633	89.766	87.308
4,6	65.688	8.650	67.086	49.282	76.466	74.338
5,6	64.892	9.820	64.682	48.410	78.348	74.712

Table C.3: 2-Parameter Models (*Continued*)

	AIC					
Var-In	AIC$_{avg}$	AIC$_{sd}$	AIC$_{med}$	AIC$_{min}$	AIC$_{max}$	AIC$_{mean+sd}$
1,2	***-116.956***	57.478	***-96.688***	***-242.458***	***-52.295***	***-59.478***
1,3	-27.069	25.074	-18.926	-86.026	-2.039	-1.995
1,4	-29.032	27.338	-20.776	-95.871	1.046	-1.694
1,5	-33.646	26.041	-28.184	-92.139	-6.975	-7.605
1,6	-31.971	25.935	-25.895	-91.081	-3.356	-6.036
2,3	74.166	***6.845***	74.707	61.642	83.702	81.012
2,4	76.900	6.974	76.933	64.328	86.539	83.874
2,5	75.179	7.539	76.147	60.465	84.357	82.718
2,6	59.225	7.273	59.072	46.061	69.796	66.497
3,4	75.144	7.243	75.397	61.638	85.226	82.387
3,5	73.770	8.402	74.116	60.108	85.154	82.173
3,6	62.003	8.788	63.235	44.919	73.426	70.791
4,5	76.205	7.921	76.202	62.315	86.492	84.125
4,6	62.817	8.360	64.253	46.964	73.193	71.177
5,6	62.021	9.520	61.849	46.092	75.075	71.541

Table C.4: 3-Parameter Models

	BIC					
Var-In	BIC$_{avg}$	BIC$_{sd}$	BIC$_{med}$	BIC$_{min}$	BIC$_{max}$	BIC$_{mean+sd}$
1,2,3	***-208.007***	87.681	***-181.948***	***-408.558***	***-112.873***	***-120.326***
1,2,4	-113.451	56.865	-92.923	-238.490	-48.223	-56.586
1,2,5	-111.473	57.089	-91.846	-236.673	-47.215	-54.384
1,2,6	-112.476	57.021	-92.295	-237.843	-49.259	-55.455
1,3,4	-24.159	27.511	-15.797	-92.154	3.368	3.352
1,3,5	-28.341	25.398	-22.487	-85.895	-4.015	-2.942
1,3,6	-28.030	25.109	-21.658	-86.657	-2.782	-2.922
1,4,5	-32.652	29.324	-23.882	-102.382	-5.608	-3.328
1,4,6	-28.732	26.553	-21.165	-92.875	0.533	-2.178
1,5,6	-33.836	26.719	-29.757	-92.583	-5.850	-7.117
2,3,4	79.712	7.398	80.456	66.714	90.001	87.110
2,3,5	78.373	7.843	79.864	64.719	88.678	86.216
2,3,6	61.821	7.383	61.630	48.368	71.884	69.204
2,4,5	80.921	7.926	82.007	65.501	90.502	88.847
2,4,6	64.915	7.610	64.528	51.142	75.998	72.525
2,5,6	63.645	***6.939***	64.148	50.369	72.967	70.585
3,4,5	79.198	8.845	79.581	65.182	91.357	88.043
3,4,6	66.870	9.061	68.611	49.299	78.472	75.931
3,5,6	65.736	10.052	65.628	48.849	79.785	75.788
4,5,6	67.182	9.321	67.475	50.786	79.550	76.503

Table C.5: 3-Parameter Models (*Continued*)

	AIC					
Var-In	AIC$_{avg}$	AIC$_{sd}$	AIC$_{med}$	AIC$_{min}$	AIC$_{max}$	AIC$_{mean+sd}$
1,2,3	***-211.835***	88.028	***-185.726***	***-412.922***	***-115.963***	***-123.807***
1,2,4	-117.279	57.211	-96.701	-242.854	-51.313	-60.068
1,2,5	-115.300	57.432	-95.623	-241.038	-50.305	-57.868
1,2,6	-116.304	57.365	-96.072	-242.207	-52.349	-58.939
1,3,4	-27.987	27.810	-19.231	-96.519	-0.194	-0.176
1,3,5	-32.168	25.722	-26.265	-90.259	-7.998	-6.446
1,3,6	-31.858	25.432	-25.435	-91.021	-6.343	-6.426
1,4,5	-36.479	29.637	-27.659	-106.746	-9.169	-6.843
1,4,6	-32.559	26.863	-24.942	-97.239	-3.028	-5.696
1,5,6	-37.664	27.045	-33.534	-96.947	-9.833	-10.618
2,3,4	75.884	6.983	76.575	63.624	85.637	82.868
2,3,5	74.545	7.429	75.984	61.628	84.313	81.975
2,3,6	57.993	6.982	57.852	45.278	67.520	64.975
2,4,5	77.094	7.513	78.127	62.411	86.138	84.607
2,4,6	61.088	7.210	60.859	48.052	71.634	68.298
2,5,6	59.818	***6.539***	60.268	47.279	68.603	66.357
3,4,5	75.370	8.434	75.701	62.092	86.993	83.804
3,4,6	63.042	8.676	64.833	46.209	74.108	71.718
3,5,6	61.909	9.648	61.850	45.759	75.421	71.556
4,5,6	63.354	8.920	63.697	47.696	75.185	72.274

Table C.6: 4-Parameter Models

	BIC					
Var-Out	BIC$_{avg}$	BIC$_{sd}$	BIC$_{med}$	BIC$_{min}$	BIC$_{max}$	BIC$_{mean+sd}$
1,2	67.729	9.895	67.866	50.852	81.559	77.625
1,3	66.448	6.995	67.066	53.088	75.785	73.443
1,4	63.532	***6.833***	64.522	50.471	72.770	70.365
1,5	64.586	7.548	64.471	51.048	74.975	72.135
1,6	80.839	7.899	82.326	67.264	91.005	88.738
2,3	-34.928	29.551	-28.119	-103.651	-6.496	-5.376
2,4	-31.774	26.342	-26.982	-90.227	-6.580	-5.431
2,5	-27.139	27.215	-18.980	-93.707	-0.039	0.075
2,6	-30.190	29.613	-21.378	-100.547	-3.206	-0.576
3,4	-109.798	56.842	-90.000	-235.007	-46.523	-52.956
3,5	-111.785	56.642	-91.335	-237.113	-47.604	-55.142
3,6	-110.999	56.643	-91.755	-235.400	-45.450	-54.356
4,5	-207.081	87.945	-181.266	***-407.868***	-111.583	-119.136
4,6	***-210.277***	85.124	***-191.714***	-405.526	***-115.430***	***-125.153***
5,6	-208.289	87.163	-184.485	-405.962	-110.488	-121.126

Table C.7: 4-Parameter Models (*Continued*)

	AIC					
Var-Out	AIC$_{avg}$	AIC$_{sd}$	AIC$_{med}$	AIC$_{min}$	AIC$_{max}$	AIC$_{mean+sd}$
1,2	62.945	9.392	63.144	46.989	76.104	72.337
1,3	61.663	6.493	62.216	49.225	70.329	68.156
1,4	58.747	***6.327***	59.800	46.608	67.315	65.074
1,5	59.802	7.045	59.749	47.185	69.520	66.847
1,6	76.055	7.385	77.476	63.401	85.549	83.439
2,3	-39.712	29.953	-32.842	-109.106	-10.948	-9.759
2,4	-36.558	26.756	-31.704	-95.682	-10.443	-9.802
2,5	-31.924	27.603	-23.702	-99.162	-4.490	-4.321
2,6	-34.974	30.005	-26.100	-106.002	-7.658	-4.969
3,4	-114.583	57.271	-94.723	-240.462	-50.386	-57.311
3,5	-116.569	57.075	-96.058	-242.568	-51.467	-59.495
3,6	-115.783	57.076	-96.477	-240.855	-49.313	-58.707
4,5	-211.865	88.378	-185.988	***-413.323***	-115.446	-123.487
4,6	***-215.061***	85.559	***-196.436***	-410.981	***-119.293***	***-129.502***
5,6	-213.073	87.601	-189.207	-411.417	-114.351	-125.472

Table C.8: 5-Parameter Models

	BIC					
Var-Out	BIC$_{avg}$	BIC$_{sd}$	BIC$_{med}$	BIC$_{min}$	BIC$_{max}$	BIC$_{mean+sd}$
1	66.099	***6.793***	67.370	53.059	74.726	72.891
2	-32.716	30.131	-25.356	-102.939	-4.412	-2.585
3	-109.312	56.516	-89.763	-234.093	-44.836	-52.797
4	-209.338	85.305	-190.885	-405.091	***-114.107***	-124.033
5	-207.639	87.338	-184.281	***-405.246***	-109.120	-120.302
6	***-210.303***	84.335	***-193.161***	-402.879	-112.897	***-125.968***
	AIC					
Var-Out	AIC$_{avg}$	AIC$_{sd}$	AIC$_{med}$	AIC$_{min}$	AIC$_{max}$	AIC$_{mean+sd}$
1	60.357	***6.185***	61.703	48.423	68.180	66.543
2	-38.457	30.614	-31.022	-109.486	-9.754	-7.843
3	-115.054	57.035	-95.430	-240.639	-49.471	-58.019
4	-215.080	85.827	-196.551	-411.637	***-118.743***	-129.252
5	-213.381	87.864	-189.947	***-411.792***	-113.755	-125.517
6	***-216.044***	84.864	***-198.827***	-409.426	-117.532	***-131.180***

Table C.9: 6-Parameter Models

	BIC					
Var-Out	BIC$_{avg}$	BIC$_{sd}$	BIC$_{med}$	BIC$_{min}$	BIC$_{max}$	BIC$_{mean+sd}$
None	-209.820	84.406	-192.966	-402.228	-111.517	-125.415
	AIC					
Var-Out	AIC$_{avg}$	AIC$_{sd}$	AIC$_{med}$	AIC$_{min}$	AIC$_{max}$	AIC$_{mean+sd}$
None	-216.519	85.023	-199.577	-409.865	-116.925	-131.495

Bibliography

[1] Abramowitz, M. and Stegun, I. A. (eds), "Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables", *Dover, New York*, 1972.

[2] Adamowicz, W., Bhardwaj, V. and Macnab, B., "Experiments on the Difference between Willingness to Pay and Willingness to Accept", *Land Economics*, Vol.69(4), pp.416-427, 1993.

[3] Anwar, S., McMillan, R. and Zheng, M., "Bidding Behavior at Competing Auctions: Evidence from eBay", *European Economic Review*, Vol.50(2), pp.307-322, 2006.

[4] Ariely, D. and Simonson, I., "Buying, Bidding, Playing, or Competing? Value Accessment and Decision Dynamics in Online Auctions", *Journal of Consumer Psychology*, Vol.13(1&2), pp.113-123, 2003.

[5] Bajari, P. and Hortacsu, A., "The Winner's Curse, Reserve Prices and Endogenous Entry: Empirical Insights From eBay Auctions", *The RAND Journal of Economic*, Vol.34(2), pp.329-355, 2003.

[6] Bajari, P. and Hortacsu, A., "Economic Insights from Internet Auctions", *Journal of Economic Literature*, Vol.42(2), pp.457-486, 2004.

[7] Bapna, R., Goes, P., Gupta, A., and Karuga, G., "Predictive Calibration of Online Multi-unit Ascending Auctions", in Proceedings of *WITS-2002*, Barcelona, Spain, December 2002.

[8] Bapna, R., Goes, P., Gupta, A. and Jin, Y., "User Heterogeneity and Its Impact on Electronic Auction Market Design: An Empirical Exploration", *MIS Quarterly*, Vol.28(1), pp.21-43, 2004.

[9] Bapna, R. Goes, P., Gopal, R., and Marsden, J., "Moving from Data-constrained to Data-enabled Research: Experiences and Challenges in Collecting, Validating, and Analyzing Large-scale e-Commerce Data", *Statistical Science*, Vol.21, pp.116-130, 2006.

[10] Bapna, R., Jank, W. and Shmueli, G., "Price Formation and Its Dynamics in Online Auctions", *Decision Support Systems*, Vol.44(3), pp.641-656, 2008.

[11] Bapna, R., Jank, W. and Shmueli, G., "Consumer Surplus in Online Auctions", *Information Systems Research*, Vol.19(4), December Issue, 2008.

[12] Basseville, M., "Distance Measure for Signal Processing and Pattern Recognition", *Signal Processing*, Vol.18(4), pp.349-369, 1989.

[13] Berk, K., "Comparing Subset Regression Procedures", *Technometrics*, Vol.20(1), pp.1-6, 1978.

[14] Becker, R. A., Chambers, J. M. and Wilks, A. R., "The New S Language", *Pacific Grove, CA: Wadsworth & Brooks*, 1988.

[15] Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J., "Classification and Regression Trees", *Pacific Grove, CA: Wadsworth*, 1984.

[16] Caccetta, L., Chow, C., Dixon, T. and Stanton, J., "Modelling the Structure of Australian Wool Auction Prices", In Zerger, A. and Argent, R.M. (eds) *MODSIM 2005 International Congress on Modelling and Simulation*, 2005.

[17] Chan, Tat Y., Kadiyali, V. and Park, Young-Hoon, "Willingness to Pay and Competition in Online Auctions", *Journal of Marketing Research*, Vol.44(2), pp.324-333, 2007.

[18] Yongmin Chen, "Paying Customers to Switch", *Journal of Economics & Management Strategy*, Vol.6(4), pp. 877-897, 2004.

[19] Cover, T. M. and Hart, P. E., "Nearest Neighbor Pattern Classification", *IEEE Transactions On Information Theory*, Vol.IT-13, pp.21-27, 1967.

[20] Dai, Q. and Kauffman, R. J., "Business Models for Internet-Based B2B Electronic Markets", *International Journal of Electronic Commerce*, Vol.6, pp.41-72, 2002.

[21] Dass, M., Jank, W., Reddy, S., Shmueli, G. and Wang, S., "Dynamic Price Forecasts in Online Indian Art Auctions", In the proceeding of *the Third Symposium on Statistical Challenges in eCommerce Research*, University of Connecticut, May 19-20,2007.

[22] Dellarocas, C., "The Digitization of Word of Mouth: Promise and Challenges of Online Reputation Mechanisms", *Management Science*, Vol.49(10), pp.1407-1442, 2003.

[23] Devroye, L. P., "On the Almost Everywhere Convergence of Nonparametric Regression Function Estimates", *Ann. Stat.*, Vol.9, pp.1310-1319, 1981.

[24] Devroye, L. P., "Necessary and Sufficient Conditions for the Pointwise Convergence of Nearest Neighbor Regression Function Estimates", *Z Wahrscheinlichkeitstheorie verw. Gebiete*, Vol.61, pp.467-481, 1982.

[25] Easley, R. F. and Tenorio, R., "Jump Bidding Strategies in Internet Auctions", *Management Science*, Vol.50(10), pp.1407-1419, 2004.

[26] Escabias, M., Aguilera, A. M. and Valderrama, M. J., "Modeling Environmental Data by Functional Principal Component Logistic Regression", *Environmetrics*, Vol.16(1), pp.95-107, 2004.

[27] Faraway, J. J., "Regression Analysis for a Functional Response", *Technometrics*, Vol.39, pp.254-261, 1997.

[28] Fix, E., "Discriminatory Analysis: Small Sample Performance", *USAF School of Aviation Medicine, Randolph Field, Texas*, Project 21-49-004, Report 11., August 1952.

[29] Fix, E. and Hodges, J. L. Jr., "Discriminatory Analysis, Nonparametric Discrimination", *USAF School of Aviation Medicine, Randolph Field, Texas.*, Project 21-49-004, Report 4, Contract AF41(128)-31, Feb. 1951.

[30] George, E., "The Variable Selection Problem", *Journal of the American Statistical Association*, Vol.95(452), pp.1304-1308, 2000.

[31] George, J., Duffy, K., and Ahuja, M., "Countering the Anchoring and Adjustment Bias with Decision Support Systems", *Decision Support Systems*, Vol.29, pp.195-206, 2000.

[32] Ghani, R. and Simmons, H., "Predicting the End-Prices of Online Auctions", *Workshop on Data Mining & Adaptive Modeling Methods for Economics & Management*, ECML/PKDD 2004, 2004.

[33] Gneezy, U., “Step-Level Reasoning and Bidding in Auctions”, *Management Science*, Vol.51(11), pp.1633-1642, 2005.

[34] Goldstein, M., “K-Nearest Neighbor Classification”, *IEEE Transactions On Information Theory*, Vol.IT-18(5), pp.627-630, 1972.

[35] Gupta, A. and Bapna, R., “Online Auctions: A Closer Look”, Book Chapter in Lowry, P (ed) *Handbook of Electronic Commerce in Business and Society*, CRC Press, 2001.

[36] Hart, P. E., “An Asymptotic Analysis of the Nearest-Neighbor Decision Rule”, *Technical Report 1828-2, SEL-66-016*, Stanford Electron. Lab., Stanford, California, May 1966.

[37] Haruvy, E., Katok, E. and Popkowski Leszczyc, Peter T. L., “Consumer Optimization, Switching and Search in Pairs of Online Auctions for Identical Items”, *Working paper*, 2007.

[38] Haruvy, E., Katok, E. and Popkowski Leszczyc, Peter T. L., “Individual Choice among Charity Auctions”, *Working paper*, 2007.

[39] Haruvy, E., Popkowski Leszczyc, P., Carare, O., Cox, J., Greenleaf, E., Jap, S., Jank, W., Park, Y. and Rothkopf, M., “Competition Between Auctions”, Special issue of *Marketing Letters*, Vol.19(3-4), pp.431-448, 2008.

[40] Haselton, M. G., Nettle, D. and Andrews, P. W. , “The Evolution of Cognitive Bias.”, in D. M. Buss (Ed.),*Handbook of Evolutionary Psychology*, Hoboken: Wiley., pp.724-746, 2005.

[41] Hasti, T. J. and Tibshirani, R. J., “Generalized Additive Models”, *Chapman and Hall, London*, 1990.

[42] He, Y. and Popkowski Leszczyc, Peter T. L., “Jump Bidding in Online Auctions: A Double-edged Sword”, *Working paper*, University of Alberta, 2007.

[43] Heyman, J., Orhun, Y. and Ariely, D., “Auction Fever: The Effect of Opponents and Quasi-Endowment on Product Valuations”, *Journal of Interactive Marketing*, Vol.18(4), pp.7-21, 2004.

[44] Hyde, V., Jank, W. and Shmueli, G., "Investigating Concurrency in Online Auctions Through Visualization", *The American Statistician*, Vol.60(3), pp.241-250, 2006.

[45] Hyde, V., Jank, W. and Shmueli, G., "A Family of Growth Models for Representing the Price Process in Online Auctions", In Jank and Shmueli (eds.) *Statistical Methods in eCommerce Research*, Wiley & Sons., 2008.

[46] Isaac, M., Salmon, T. C. and Zillante, A., "A Theory of Jump Bidding in Ascending Auctions", *Journal of Economic Behavior & Organization*, Vol.62(1), pp.144-164, 2004.

[47] Jain, A., Murty, M. and Flynn, P., "Data Clustering: A Review", *ACM Computing Surveys*, Vol.31(3), pp.264-323, 1999.

[48] James, G. M., "Generalized Linear Models with Functional Predictors", *Journal of the Royal Statistical Society, Series B*, Vol.64(3), pp.411-432, 2002.

[49] James, G. M., and Sugar, C. A., "Clustering Sparsely Sampled Functional Data", *Journal of the American Statistical Association*, Vol.98, pp.397-408, 2003.

[50] Jank, W. and Shmueli, G., "Visualizing Online Auctions", *Journal of Computational and Graphical Statistics*, Vol.14(2), pp.299-319, 2005.

[51] Jank, W. and Shmueli, G., "Profiling Price Dynamics in Online Auctions Using Curve Clustering", *Working paper*, Robert H. Smith School of Business, University of Maryland, http://ssrn.com/abstract=902893, 2005.

[52] Jank, W. and Shmueli, G., "Functional Data Analysis in Electronic Commerce Research", *Statistical Science*, Vol.21(2), pp.155-166, 2006.

[53] Jank, W. and Shmueli, G., "Modelling Concurrency of Events in Online Auctions via Spatio-temporal Semiparametric Models", *Journal of the Royal Statistical Society: Series C*, Vol.56(1), pp.1-27, 2007.

[54] Jank, W. and Shmueli, G., "Studying Heterogeneity of Price Evolution in eBay Auctions via Functional Clustering", In Adomavicius, G. and Gupta, A. (eds.) *Handbook of Information Systems Series: Business Computing*, Elsevier, 2008.

[55] Jank, W., Shmueli, G., and Wang, S., "Modeling Price Dynamics in Online Auctions via Regression Trees", Book chapter in *Statistical Methods in eCommerce Research* (Forthcoming), Wiley & Sons., 2007.

[56] Jank, W., Shmueli, G., and and Zhang, S., "A Flexible Model for Price Dynamics in Online Auctions", *Working Paper*, University of Maryland, 2009.

[57] Jank, W. and Zhang, S., "An Automated and Data-Driven Bidding Strategy for Online Auctions", *Working Paper*, University of Maryland, 2008.

[58] Jank, W. and Zhang, S., "Model Selection for Online Auction Forecasting", *Working Paper*, University of Maryland, 2009.

[59] Jap, S. and Naik, P., "BidAnalyzer: A Method for Estimation and Selection of Dynamic Bidding Models", *Marketing Science*, Vol.27, pp.949-960, 2008.

[60] Jaynes, E. T., "Information Theory and Statistical Mechanics", *Physical Review*, Vol.106, pp.106-620, 1957.

[61] Ku, G., Malhotra, D. and Murnighan, J. K., "Towards a Competitive Arousal Model of Decision-making: A Study of Auction Fever in Live and Internet Auctions", *Organizational Behavior and Human Decision Processes*, Vol.96, pp.89-103, 2005.

[62] Kulkarni, Sanjeev R., Lugosi, Gabor and Venkatesh, Santosh S., "Learning Pattern ClassificationA Survey", *IEEE Transactions On Information Theory*, Vol.44(6), 1998.

[63] Kulkami, Sanjeev R. and Posner, Steven E., "Rates of Convergence of Nearest Neighbor Estimation Under Arbitrary Sampling", *IEEE Transactions On Information Theory*, Vol.41(4), 1995.

[64] Kahneman, D. and Tversky, A., "Prospect Theory: An Analysis of Decision Under Risk", *Econometrica*, Vol.XLVII, pp.263-291, 1979.

[65] Kahneman, D., Knetsch, J. L., and Thaler, R. H., "Fairness as a Constraint on Profit Seeking: Entitlements in the Market", *American Economic Review*, Vol.76, pp.728-741, 1986.

[66] Kahneman, D., Knetsch, J. L., and Thaler, R. H., "Anomalies: The Endowment Effect, Loss Aversion, and Status Quo Bias", *The Journal of Economic Perspectives*, Vol.5, pp.193-206, 1991.

[67] Kassardjian, E., Gamble, J., Gunson, A., and Jaeger, S. R., "A New Approach to Elicit Consumers' Willingness to Purchase Genetically Modified Apples", *British Food Journal*, Vol.107(8), pp.541-555, 2005.

[68] Kullback, S. and Leibler, R. A., "On Information and Sufficiency", *Annals of Mathematical Statistics*, Vol.22, pp.79-86, 1951.

[69] Langdoc, S. and Newmark, E., "Retail Lifecycle Price Management: Blending Optimization and Execution Modernizes Retail Pricing", *AMR Research*, 2004.

[70] Levinson, N., "The Wiener RMS (Root Mean Square) Error Criterion in Filter Design and Prediction", *Journal of Mathematics and Physics*, Vol.25, pp.261-278, 1946.

[71] Levy, D., Chen, H., Ray, S., and Bergen, M. E., "Asymmetric Price Adjustment in the Small: An Implication of Rational Inattention", *Working Paper*, T.C. Koopmans Research Institute, http://ssrn.com/abstract=563867, 2005.

[72] Daniel Levy, Georg Müller, Haipeng Chen, Mark E. Bergen, and Shantanu Dutta, "Holiday Price Rigidity and Cost of Price Adjustment", *Working Paper*, http://ssrn.com/abstract=389640, 2008.

[73] Liu, B and Mueller, H.-G., "Functional Data Analysis for Sparse Auction Data", Book Chapter in Jank, W. and Shmueli, G. (eds.) *Statistical Methods in eCommerce Research*, Wiley and Sons, New York, pp.269-290, 2008.

[74] Lucking-Reiley, D., Bryan, D., Prasad, N. and Reeves, D., "Pennies from eBay: the Determinants of Price in Online Auctions", *Journal of Industrial Economics*, Vol.55(2), pp.223-233, 2007.

[75] Nalewajski, Roman F. and Parr, Robert G., "Information Theory, Atoms in Molecules, and Molecular Similarity", in Proceedings of *the National Academy of Sciences of the United States of America*, Vol.97(16), pp.8879-8882, 2000.

[76] Overby, S., "The Price Is Always Right", *CIO Magazine*, June 13, 2007.

[77] Radner, R., “Satisficing”, *Journal of Mathematical Economics*, Vol.2, pp.253-262, 1975.

[78] Ramsay, J. O., “Estimating Smooth Monotone Functions”, *Journal of the Royal Statistical Society B*, Vol.60, pp.365-375, 1998.

[79] Ramsay, J. O. and Silverman, B. W., “Functional Data Analysis”, *Springer-Verlag, New York*, 2nd edition, 2005.

[80] Ramsay, J. O. and Silverman, B. W., “Applied Functional Data Analysis: Methods and Case Studies”, *Springer-Verlag, New York*, 2002.

[81] Raubera, T. W., Braun, T. and Berns, K., “Probabilistic Distance Measures of the Dirichlet and Beta Distributions”, *Pattern Recognition*, Vol.41(2), pp.637-645, 2008.

[82] Reddy, S. K. and Dass, M., “Modeling Online Art Auction Dynamics using Functional Data Analysis”, *Statistical Science*, Vol.21, pp.179-193, 2006.

[83] Roth, A. E. and Ockenfels, A., “Last-minutes Bidding and the Rules for Ending Second Price Auctions: Evidence from eBay and Amazon on the Internet”, *American Economic Review*, Vol.92, pp.1093-1103, 2002.

[84] Shmueli, G. and Jank, W., “Modeling the Dynamics of Online Auctions: A Modern Statistical Approach”, Book Chapter in R. Kauffman and P. Tallon (eds.) *Economics, Information Systems and Ecommerce Research: II. Advanced Empirical Methods*, part of Advances in Management Information Systems Series, M.E. Sharpe, 2008.

[85] Shmueli, G., Jank, W., Aris, A., Plaisant, C. and Shneiderman, B., “Exploring Auction Databases through Interactive Visualization”, *Decision Support Systems*, Vol.42(3), pp.1521-1538, 2006.

[86] Shmueli, G., Russo, R. and Jank, W., “The barista: A Model for Bid Arrivals in Online Auctions”, *The Annals of Applied Statistics*, Vol.1(2), pp.412-441, 2007.

[87] Short, R. D. and Fukunaga, K., “The Optimal Distance Measure for Nearest Neighbor Classification”, *IEEE Transactions On Information Theory*, Vol.IT-27(5), pp.622-627, Sep. 1981.

[88] Simonoff, J. S., “Smoothing Methods in Statistics”, *Springer-Verlag, New York*, 1996.

[89] Simonsohn, U. and Ariely, D., “When Rational Sellers Face Non-Rational Buyers: Evidence from Herding on eBay”, *Working Paper*, http://ssrn.com/abstract=722484, 2007.

[90] Spann, M. and Skiera, B., “Internet-Based Virtual Stock markets for Business Forecasting”, *Management Science*, Vol.49(10), pp.1310-1326, 2003.

[91] Stone, C. J., “Consistent Nonparametric Regression”, *Ann. Stat.*, Vol.5, pp.595-645, 1977.

[92] Sun, E., “The Effects of Auctions Parameters on Price Dispersion and Bidder Entry on eBay: A Conditional Logit Analysis”, *Working paper*, Stanford University, 2005.

[93] Tversky, A. and Kahneman, D., “Judgment under Uncertainty: Heuristics and Biases”, *Science*, Vol.185, pp.1124-1131, 1974.

[94] Tversky, A. and Kahneman, D., “The Framing of Decisions and The Psychology of Choice”, *Science*, Vol.211, pp.453-458, 1981.

[95] Thomas Vollmann, William Berry, David Clay Whybark, and F. Robert Jacobs, “Manufacturing Planning and Control Systems for Supply Chain Management”, *McGraw-Hill*, 5th edition, 2004.

[96] Wang, S., Jank, W. and Shmueli, G., “Explaining and Forecasting Online Auction Prices and their Dynamics using Functional Data Analysis”, *Journal of Business and Economic Statistics*, Vol.26(2), pp.144-160, 2008.

[97] Wang, S., Jank, W., Shmueli, G. and Smith, P., “Modeling Price Dynamics in eBay Auctions Using Differential Equations”, *Journal of the American Statistical Association*, Vol.103(483), pp.1100-1118, 2008.

[98] Wang, Xin, Montgomery, Alan L., and Srinivasan, Kannan, “When Auction Meets Fixed Price: A Theoretical and Empirical Examination of Buy-it-Now Auctions”, *Quantitative Marketing and Economics*, Vol.6(4), pp.339-370, 2008.

[99] Zbaracki, M., Ritson, M., Levy, D., Dutta, S., and Bergen, M., "Managerial and Customer Costs of Price Adjustment: Direct Evidence From Industrial Markets", *The Review of Economics and Statistics*, Vol.86, pp.514-533, 2004.

[100] Zeithammer, R., "Forward-looking Bidding in Internet Auctions", *Journal of Marketing Research*, Vol.43, pp.462-476, 2006.

[101] Zeng, Y., Wen, H.J., and Yen, D. C., "Customer Relationship Management (CRM) in Business-to-Business (B2B) e-Commerce", *Information Management & Computer Security*, Vol.11, pp.39-44, 2003.

[102] Zhang, S., Jank, W. and Shmueli, G., "Real-Time Forecasting of Online Auctions via Functional K-Nearest Neighbors", *Working paper*, University of Maryland, 2009.