Semi- and non-parametric estimation of price dispersion and auction heterogeneity with eBay auctions

Author: Julian Valdman

Advisor: Bertel Schjerning

University of Copenhagen

June 2022

Abstract

Over the last couple of years, online auctions on marketplaces like eBay have gained popularity but have also given rise to information asymmetries. Investigating the prices of an iPhone 12 Pro Max on eBay, price dispersion contradicting The Law of One Price is immediately apparent. First, the paper explains why strategic behavior may lead to price dispersion and bidding behavior not following auction-theoretical predictions. Second, a semi-parametric Maximum Likelihood approach utilizing Hermite Series is used to estimate how auction features affect the auction's end prices. The auction features comprise auction-specific and seller-specific variables. It is found that longer auctions and many bidders yield higher prices while more reviews and negative feedback lower the end price. Overall, the estimation suggests that the auction features will affect the price between $1\% - 8\%$. Lastly, non-parametric Machine Learning methods (Neural Networks) are applied to classify the auction prices. The multi-class classification analysis backs up the initial conclusions from the Maximum Likelihood Estimation and shows that auction features partly do determine the auction price. In accordance with other existing literature, the results are not fully unambiguous; that is, having received many bids or being a seller with many reviews actually lowers the end price on average. In continuation, the predictive power of the auction features on the end price is limited alluding to that other factors also determine the auction price. The ambiguous results and limited predictive power may be due to the strategic behavior, outside factors having an influence on the price, such as concurrent or past auctions, or due to a limited data set with only one product.

Contents

1 Introduction

In recent years, online marketplaces like eBay and Amazon have experienced a vast increase in use and popularity. At these places, trading occurs in a very decentralized manner, including one-to-one trading and one-to-many auctions. Even though these marketplaces have gained widespread recognition, they have given rise to asymmetric information problems (Saeedi, 2014).

The Law of One Price predicts that identical homogeneous goods necessarily must be traded at the same price to avoid arbitrage opportunities. When empirically investigating the auction data on eBay, it is immediately apparent that the given product exhibits a vast price variation dependent on the auction's characteristics. This paper aims to explain the potential reasons behind the price dispersion and estimate how much predictive power the single auction's characteristics have on the final price.

The goal is to maximize the fraction of the price variability which can be explained using auction characteristics and heterogeneity. The use of Machine Learning (ML) is meant to solve two potential problems - which may arise in other econometric estimation methods – which are the omitted variable bias and functional form misspecification. Hence, the aim is not solely to identify individual parameter estimates to which causal demand interpretations may be attached. Rather, the author attempts in this paper to also determine what combinations of more ample auction characteristics and more flexible statistical models can exhaust predictive power for cross-sectional auction price differences (Bodoh-Creed, Boehnke, & Hickman, 2004).

Overall, this paper is divided into three parts.

The first part sheds a light on the auction dynamics on the eBay marketplace. With this in hand, it aims to describe which auction-theory perspectives and microeconomic problems may explain the price dispersion. In addition, the data set and data collection process are explained.

In **the second part**, Maximum Likelihood Estimation is applied to semi-parametrically identify the underlying distribution of the collected auction bids. By estimating the distribution based on different subsamples of the data set, an initial end-price effect from single auction attributes can be inferred.

In **the third part**, Machine Learning (ML) methods are used for multi-class classification. After dividing the auctions into 4 equally sized price classes, the ML model utilizes a flexible non-parametric statistical structure to find non-trivial, ample combinations of the auction characteristics to exhaust the predictive power on the end-price class.

After the main sections, the analysis and results will be discussed and concluded.

2 The eBay marketplace and data

The analysis and data in this paper are built upon data from the eBay marketplace. The eBay auctions are chosen as the auctions of interest since eBay provides a good setting for studying reputation and market friction effects. First, researchers can observe all seller characteristics observable by buyers. Second, buyer and seller have little or no interactions with each other outside the eBay platform. This nuanced selection of information enables researchers to estimate the impact of auction characteristics and adverse selection on the final price (Bajari & Hortacsu, 2004).

2.1 The dynamics in eBay auctions

In recent years, online marketplaces like eBay and Amazon have experienced a vast increase in use and popularity. At these places, trading occurs in a very decentralized manner, including one-to-one trading and one-to-many auctions (Saeedi, 2014).

In general, eBay provides three ways to sell a product. The seller can (i) offer a take-it-or-leave-it price as you have in regular shops and sell the product to the first buyer, (ii) engage in pair-wise negotiations with individual buyers – either sequentially or simultaneously – or, lastly, (iii) auction his/her goods using the eBay auction format (Gentry, Hubbard, Nekipelov, & Parsch, 2018).

An individual auction on eBay lasts between one and ten days and is chosen by the seller. All eBay auctions use a second-price, non-sealed, ascending-bid format and they have fixed ending-date and ending-time set by the seller. The eBay platform provides a so-called "proxy bidding" system that updates a buyer's bid incrementally up to the provided maximum bid (and the maximum remains secret until the bidder is outbid). In this paper, the fees the seller pays to eBay – an initial listing fee and a final value fee – are disregarded due to minor impact and simplicity. In addition, the analysis only focuses on sellers' experience (measured in the count of reviews) and disregards the buyers' experience even though this affects how strategically the given buyer acts (Roth & Ockenfels, 2002).

As many authors have pointed out (Bajari & Hortacsu, 2004), the eBay marketplace is plagued by information asymmetries, but the problem is somewhat alleviated by the eBay reputation system. After a completed transaction, the buyer can rate the seller, and the sum of positive ratings minus negative ratings will constitute a seller's rating.

The feedback system aims to ensure higher transparency and higher seller credibility through feedback scores. Thus, it is assumed that more reviews and higher feedback scores yield higher end-prices (Kalyanam & McIntyre, 2001). Even though the feedback system addresses and mitigates some of the adverse selection problems rising, there are economic reasons why the feedback might not have a significant impact on the auction price. First, any user can provide a rating to any user at any time. In other words, eBay does not require that a transaction between the rater and the rated has occurred. Second, there is a potential free-rider problem; when a buyer rates a seller, the first-mentioned gets little personal benefit for doing so – the public good benefits accrue to the people who later will see the rating (Eaton, 2005; Lucking-Reiley, Bryan, & Reeves, 2007).

Another concern related to the impact of the feedback scores is the possibility that those with negative feedback tend to be those with more eBay transactions. Thus, the number of negative feedback reviews acts as a proxy for more experienced sellers (Eaton, 2005). As more experienced sellers are assumed to yield higher prices ceteris paribus, negative feedback may, in turn, have a counter-intuitive, positive effect on the auction price.

These shortcomings of the reputation system are worth noticing when later evaluating the predictive power of the reputational variables.

2.2 Auction theory, strategies, and efficiency

The eBay auctions are especially interesting to investigate from an auction theoretic point of view as inconsistencies between theoretical prediction and empirical data are immediately apparent. Since the auctions follow a second-price, ascending English auction format, the equilibrium - and the only strategically dominant strategy – will be that a buyer's first bid is her only bid and equals her true valuation (Gonzalez, Hasker, & Sickles, 2004). In many of the observed auctions, bidders place several bids and exhibit strategic bidding decisions. This discrepancy, as an example, alludes to strategic incentives and behavior performed either deliberately or by mistake by the buyer (Roth & Ockenfels, 2002).

Several authors (Gonzalez et al., 2004; Roth & Ockenfels, 2002) claim and show that this deviation from theoretical predictions may be due to irrationality, inexperience, or simply procrastination. On the other side, they also point out that strategic behavior and psychological aspects also may play a factor: in order to protect their private information, bidders avoid bidding wars with other like-minded bidders. Hence, bidders may apply various bidding strategies like bid sniping, cross-bidding, or jump bidding. The first-mentioned concept covers that the majority of bids in fixed-length auctions are placed within the very last part of the auction. The second strategy refers to bidders who interact in several similar auctions simultaneously in order to increase the probability of winning and paying less in general (Anwar, McMillan, & Zheng, 2004). The latter strategy, Jump bidding, consists of bidding excessively high compared to what is needed to become the current winning bid. Experienced bidders tend to use this strategy more often than inexperienced ones.

The study conducted by Bajari and Hortacsu (2000) finds that it is the concept of the time-fixed auction end which encourages this strategic behavior and especially bid sniping. In addition, they find that the late bidding erodes the price effects coming from the duration, number of bidders, and number of bids. Ghani and Simmons (2004) report that 15 % of the sampled auctions are won in the last minute which increases the uncertainty of the end-price of a given auction.

The above-mentioned inefficiencies allude to that the auction price dispersion may not only be due to auction heterogeneity or seller characteristics (Bodoh-Creed et al., 2004), but also behavioral economics theory and irrational microeconomic agents. The study performed by Kalyanam and McIntyre (2001) takes the stance that the observed bidding patterns cannot be explained by rational buyers. These underlying economic aspects are important to address as the model and estimation exclusively include simple observable auction characteristics. Hence, the estimation will not include the above inefficiencies and a part of the price dispersion will not be estimable in the current setting.

2.3 Auction data collected from eBay

The data in this paper comes from auctions completed on eBay between March and June 2022. The data set is collected using a web scraper built in Python which is a high-level programming language. The scraping script logs onto eBay, traverses the displayed, completed auctions in order, and for each auction, it enters different sub-pages and collects a diverse set of auction-specific characteristics.

The list of auction characteristics to collect is inspired by Lucking-Reiley et al. (2007) and their stated hypotheses. The analysis variables comprise auction-specific variables, like duration or final bid price, and seller-specific variables, like feedback rating. For example, Lucking-Reiley et al. (2007) include the following variables in their modeling of price determinants: auction duration, the positive and negative seller ratings, number of bids, and whether the auction ended on weekend. In this paper, additional analysis variables are included but restricted by the availability on the eBay page. Among the features, some are directly available on the eBay page, and some are derived thereof: e.g., a dummy whether the auction possesses an additional item condition description.

Overall, the raw data set consists of $N = 1551$ auction listings each with $T = 20$ attributes.

In this paper, the product being subject to analysis is an iPhone 12 Pro Max by Apple which is chosen for several reasons.

Firstly, this product is similar to the products being analyzed in other similar papers, e.g., an iPod or a monitor (Saeedi, 2014; Bodoh-Creed et al., 2004).

Secondly, the product exhibits high selling and bidding volume on the eBay platform which has various positive economic and econometric implications. A bigger data set allows estimations with more consistent estimators and eliminates spurious correlations. In addition, the high number of bidders decreases the possibility of bidder collusions (and thereby suggests a competitive market) which enables a simpler analysis ignoring the market power effects (Gonzalez et al., 2004).

Thirdly, by selecting a reasonably homogenous good (Saeedi, 2014), the author aspires to isolate the impact of the seller's reputation and other characteristics on the price (Eaton, 2005).

Furthermore, some assumptions are imposed to ensure the validity of the estimated densities of the winning bids. For the information structure, independent private values (IPV) are assumed which means a bidder's bidding strategy does not depend on any other bidder's action or private information (Kim & Lee, 2014). This problem could arise in the case of the earlier-mentioned strategic behavior such as jump bidding. In addition, the bidders are assumed to be symmetric implying that their valuations are drawn from the same distribution, $F(\cdot)$, and this distribution is fixed across all auctions (Song, 2004).

2.3.1 Data cleaning and preparation

The raw data set, denoted by $\mathbf{X}$, collected by the web scraper consists of $N = 1551$ auction listings each with $T = 20$ attributes/characteristics, $\mathbf{X} \in \mathbb{R}^{N \times T}$.

In order to remove the data outliers, Interquartile Ranges (IQR) filtering is applied. Basing the filtering on the price attribute for each auction, define q_1 and q_3 as first and third quartile, respectively, of the price column, `price_dkk`. In addition, define $\text{IQR} \equiv q_3 - q_1$. Then the data, $\mathbf{X}$, is subsetted with all entries where the price, p, is in the interval $[q_1 - \alpha \cdot \text{IQR}; q_3 + \alpha \cdot \text{IQR}]$ where $\alpha = 2$ in this setting (Rousseeuw & Hubert, 2011). Denote the new, IQR-filtered data set by $\mathbf{X}^{IQR}$.

Denote the price (bids) column of $\mathbf{X}^{IQR}$ by $\mathbf{b}$. This is the winning bids of the auctions. The distribution of the bids are shown in Figure 4 in the Appendix. This price vector will be used in the upcoming sections as the dependent variable to be explained by the other auction characteristics.

First, the Maximum Likelihood Estimation will estimate the underlying distribution of $\mathbf{b}$. Second, $\mathbf{b}$ will be the left-hand-site variable to classify by the Neural Network. When classifying $\mathbf{b}$, only a subset of the T attributes in $\mathbf{X}$ will be regressed upon due to relevance. In specific, the columns stated in Table 1 are removed as these have no predictive power related to the price or are transformed into another variable.

Removing some of the auction attributes, the final data set, $\mathbf{X}^F$, has $N = 1434$ entries and $T = 14$ attributes. The list of independent variables to regress upon and their summary statistics are shown in Table 9 in the appendix. Only the numerical variables are shown in this table. In addition, the list of auction attributes also comprises Item Condition and Seller's Country.

Column Name	Reason for removal
Winning bid	The dependent variable
Title	No predictive power
Time of auction end	No predictive power
Seller's location, city	No predictive power
Date when seller's user was created	Transformed into numeric variable
Condition description	Transformed into dummy

Table 1: Some columns are removed from the raw data set due to lacking relevance or other reasons.

2.4 Section Conclusion - eBay marketplace and data

The eBay marketplace has over the years experienced a vast increase in popularity among auctioneers, but also among researchers due to its good setting for studying auction heterogeneity and seller reputation. The reputation feedback system on eBay is supposed to mitigate the informational asymmetries arising with online auctions, but the significance is questionable: when estimating the effect of the feedback scores it is important to consider the potential free-rider problems and that negative feedback counter-intuitively may increase the end price.

On eBay, there are immediately apparent inconsistencies between auction theory predictions and the empirical bidding data. These deviations can be a result of agent irrationality or, oppositely, strategic behavior. Such aspects, especially bid sniping, can in turn erode the price effects coming from simpler auction features, namely auction duration and the number of bidders or bids. And because the following estimation models do not include these inefficiencies but only simpler auction features, a part of the price dispersion will not be estimable.

In order to analyze the reputation system empirically, data is collected from the eBay marketplace using a scraper built in Python. For each auction, the price is collected together with auction-specific and seller-specific variables. The list of analysis variables is inspired by various similar papers. The item of interest in the following analysis is an iPhone 12 Pro Max due to different economic reasons, namely due to its homogeneity, trading volume, and precedence in other papers. After removing outliers and only including relevant regressors, the prepared data set contains 1434 auctions each with 14 auction features.

3 Maximum Likelihood Estimation

In the previous section, it was described that various auction characteristics affect the price due to information asymmetry and behavioral aspects. Later in this section, the data will be subsampled based on the above characteristics. For each subsample, the density of the subsample's prices will be estimated and the mean and variance will be compared to the baseline density estimate. This will, at first glance, indicate which auction attributes affect the end auction price.

In order to estimate the underlying distribution of the empirical winning auction bids, Maximum Likelihood Estimation (MLE) is used. The MLE will result in a density estimate, $\hat{f}(\cdot)$, and, in turn, the associated CDF, $\hat{F}(\cdot)$. In the end of this section, the CDF will be used to divide the auction bids into bins - based on the quartiles of the identified fitted distribution - used for multi-class classification in a later section.

The following empirical analysis processes the price data vector described in Section 2.3 denoted in this section by $\mathbf{b}$.

The two upcoming estimation follow the general Maximum Likelihood framework where identification of the true parameters, θ_o, implies:

$$\begin{aligned}\theta_o &= \arg\min_{\theta \in \Theta} E\left[q(\mathbf{b}, \theta)\right] \\ &= \arg\min_{\theta \in \Theta} E\left[-\ln f(\mathbf{b}; \theta)\right]\end{aligned}$$

Here, the parameter space, Θ, will vary dependent on the estimation below.

The empirical parameter estimates will hence solve the following sample problem.

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \frac{1}{N} \sum_{i=1}^{N} -\ln f(b_i; \theta) \tag{1}$$

Using different parametric structures of the density $f(\cdot)$, the aim will be to identify the best fitting density.

The following estimations are developed in the programming language R utilizing the built-in BFGS-B algorithm for solving these multi-variable nonlinear optimization problems where some parameters are constrained. The implementation of the below semi-nonparametric MLE procedure is inspired by Foster (2021).

3.1 Estimation 1: Normal Density

First, the data is fitted applying Maximum Likelihood Estimation using the normal density. The normal distribution is deemed a reasonable distribution given the initial plotting of the bidding data. See Figure 4 in the Appendix. Hence, the following sample likelihood function is considered:

$$L = \frac{1}{N}\sum_{i=1}^{N} -\ln\phi(b_i;\sigma,\mu) = \frac{1}{N}\sum_{i=1}^{N} -\ln\left[\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{b_i-\mu}{\sigma}\right)^2}\right]$$

The parameters, $\hat{\theta} = (\hat{\mu},\hat{\sigma})'$, will be found solving sample problem:

$$\hat{\theta} = \arg\min_{\theta\in\Theta}\frac{1}{N}\sum_{i=1}^{N} -\ln\left[\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{b_i-\mu}{\sigma}\right)^2}\right]$$

The parameter space is here the 2-dimensional space, $\Theta \subset \mathbb{R}^2$ where $\sigma \in \mathbb{R}_{>0}$ and $\mu \in \mathbb{R}$.

The results of the parameter estimation can be seen in Figure 4. Comparisons between the fitted CDF and the empirical CDF are shown in Figure 2 in the right panel.

Due to lack of accuracy and inflexibility being the result of a parametric approach, the next section introduces a Hermite Series which allows for more flexible and less parametric estimation of the sample data.

3.2 Estimation 2: Hermite Series

The following estimate utilizes a semi-nonparametric approach developed by Gallant and Nychka (1987). In this estimation, it is used that a parameterized orthogonal polynomial sequence (in this paper, a Hermite series of degree k is chosen) can replace the unknown density of interest.

The density of the observed winning bids, $\mathbf{b}$, will hence be specified as follows:

$$f(x;a_1,...,a_k,\mu,\sigma) = \frac{\left[1+a_1(\frac{x-\mu}{\sigma})+\cdots+a_k(\frac{x-\mu}{\sigma})^k\right]^2\phi(x;\mu,\sigma,\underline{x},\bar{x})}{\int_{\underline{x}}^{\bar{x}}\left[1+a_1(\frac{x-\mu}{\sigma})+\cdots+a_k(\frac{x-\mu}{\sigma})^k\right]^2\phi(x;\mu,\sigma,\underline{x},\bar{x})dx} \tag{2}$$

where $a = a_1,...,a_k$ are the Hermite parameters and $\phi(x;\mu,\sigma,\underline{x},\bar{x})$ is the density of $N(\mu,\sigma)$ with truncated support over the interval $[\underline{x},\bar{x}]$ (where $\Phi(\cdot)$ denotes the CDF of the unconstrained normal density):

$$\phi(x;\mu,\sigma,\underline{x},\bar{x}) = \frac{\phi(x;\mu,\sigma)}{\Phi(\bar{x};\mu,\sigma)-\Phi(\underline{x};\mu,\sigma)}$$

To ensure that $\int_{\underline{x}}^{\bar{x}} f(x)dx = 1$, the integral in the denominator is added. The support interval will be defined as:

$$\underline{x} = \min_t b_t - \epsilon \qquad\qquad \bar{x} = \max_t b_t + \epsilon$$

where ϵ is set arbitrarily small relative to the data. The estimation parameters will then be $\theta = (a_1,...,a_k,\mu,\sigma)'$ and will be chosen through maximum likelihood.

The parameter space is a subspace of the entire $k+2$ space, $\Theta \subset \mathbb{R}^{k+2}$, with the following restrictions; $\sigma > 0$, $|a_i| < 10^6$ for $i = 1, ..., k$ and $|\mu| < 10^6$. The optimal hermite series length, k^*, can be found using a cross-validation strategy employing the Integrated Squared Errors criteria (Coppejans & Gallant, 2002). However, in the below estimations, $k = 4$ is chosen due to simplicity (Song, 2004). Similar results were found with $k = 2$.

Using the above general setting from Equation 1, the parameters, $\hat{\theta}$, will be found solving:

$$\begin{aligned}\hat{\theta} &= \arg\min_{\theta\in\Theta} \frac{1}{N}\sum_{i=1}^{N} -\ln f(b_i;\theta)\\ &= \arg\min_{\theta\in\Theta} \frac{1}{N}\sum_{i=1}^{N} -\ln\left[\frac{\left[1 + a_1(\frac{b_i-\mu}{\sigma}) + \cdots + a_k(\frac{b_i-\mu}{\sigma})^k\right]^2 \phi(b_i;\mu,\sigma,\underline{x},\bar{x})}{\int_{\underline{x}}^{\bar{x}} \left[1 + a_1(\frac{x-\mu}{\sigma}) + \cdots + a_k(\frac{x-\mu}{\sigma})^k\right]^2 \phi(x;\mu,\sigma,\underline{x},\bar{x})dx}\right]\end{aligned}$$

Noting that the density being estimated only returns a positive number in the truncated interval $[\underline{x};\bar{x}]$, the CDF of the density is defined as follows.

$$\hat{F}(x;\hat{\theta}) = \int_{\underline{x}}^{x} \hat{f}(u;\hat{\theta})du \tag{3}$$

3.3 Monte Carlo simulation - Hermite Series

In order to observe the behavior of the estimated Hermite series densities, a Monte Carlo simulation is performed. Here, the Hermite estimators are compared to the true parameters of simulated data. The Monte Carlo setup is inspired by the method of Song (2004).

3.3.1 Data generating process

First, a simple model is set up for a set of fictional auctions. It is assumed that the bid from bidder i, B_t^i, is a linear combination of exogenous attributes for auction t, α_t and β_t, and her private valuation, v_t^i.

$$B_t^i = a_1\alpha_t + a_2\beta_t + a_3v_t^i$$

where $\alpha_t \sim N(0,1), \beta_t \sim Exp(1)$ and $v_t^i \sim \Gamma(9,3)$. In addition, $a_1 = 2, a_2 = -3$, and $a_3 = 1$. For simplicity, assume that there is only one bidder per auction, hence the superscript can be omitted.

$$B_t = a_1\alpha_t + a_2\beta_t + a_3v_t \tag{4}$$

Setting the sample size, $N = 1000$, each Monte Carlo iteration consists of $N = 1000$ auctions (or, equivalently, bids) $B_1, ..., B_{1000}$. The true density mean and standard deviation, denoted by stars, will as reference point then be:

$$\mu^* \equiv E[B_t] = E[a_1\alpha_t + a_2\beta_t + a_3 v_t] = 0 - 3 \cdot 1 + 3 = 0$$
$$\sigma^* \equiv \sqrt{V[B_t]} = \sqrt{a_1^2 \cdot V[\alpha_t] + a_2^2 \cdot V[\beta_t] + a_3^2 \cdot V[v_t]} = \sqrt{4+9+1} \approx 3.7417$$

Above, it is used that the random variables are independent.

3.3.2 Hermite density estimators

Given the N bids - and for a given Monte Carlo iteration - the above Maximum Likelihood Estimation is performed leading to the fitted parameters of the hermite series of degree k, $\hat{\theta} = (\hat{a_1}, ..., \hat{a_k}, \hat{\mu}, \hat{\sigma})'$. Hence, for each iteration, a density estimate, $\hat{f}(\cdot) = \hat{f}(x; \hat{\theta})$, is obtained. Here $k = 4$ as above.

Then, the mean and standard deviation of the estimated density, denoted by tilde, are computed. Let X be a random variable drawn from $\hat{f}(\cdot)$, $X \sim \hat{f}(\cdot)$.

$$\tilde{\mu} \equiv \mathbb{E}[X] = \int_{-\infty}^{\infty} x\hat{f}(x)dx$$
$$\tilde{\sigma} \equiv \sqrt{\text{Var}[X]} = \sqrt{\int_{-\infty}^{\infty} (x - \tilde{\mu})^2 \hat{f}(x)dx}$$

In this simulation, there will be $S = 360$ iterations leading to S pairs of $(\tilde{\mu}, \tilde{\sigma})$ named $\{(\tilde{\mu}_s, \tilde{\sigma}_s)\}_{s=1}^{S}$. In all S iterations, all the hermite parameters, $\hat{\theta}$, are statistically significant with p-values being close to 0.

Lastly, the estimator values, $\tilde{\mu}_s, \tilde{\sigma}_s$ for $s = 1, .., S$, are plotted in a histogram. As seen in Figure 1, the estimator collapses around the true density mean and standard deviation (μ^* and σ^*) and, thereby, exhibit consistency.

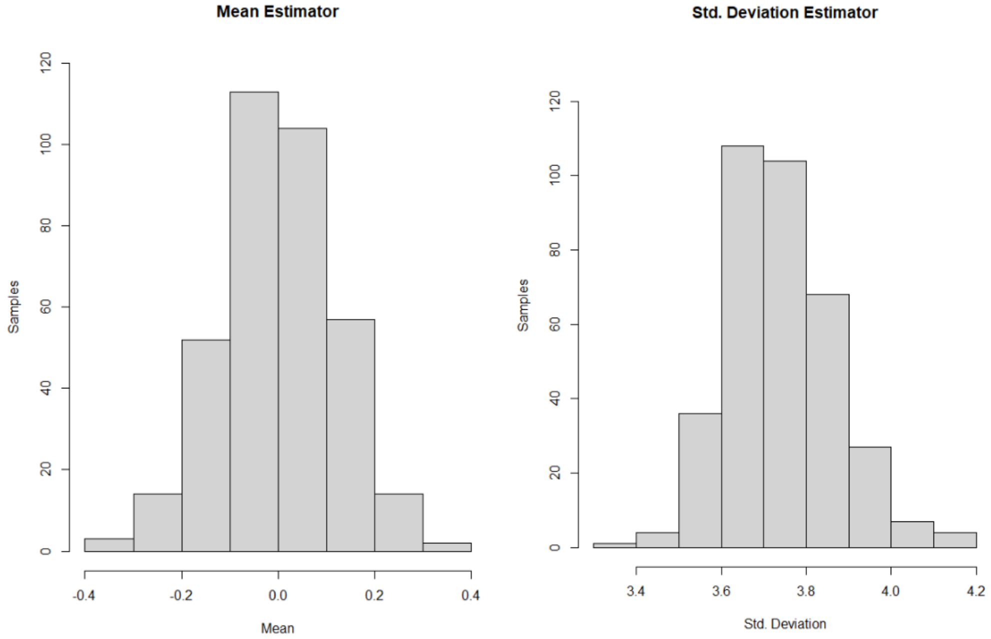
To compare the estimator values to the true parameters, define the mean estimator value.

$$\bar{\tilde{\mu}} = \frac{1}{S}\sum_{s=1}^{S} \tilde{\mu}_s \qquad , \qquad \bar{\tilde{\sigma}} = \frac{1}{S}\sum_{s=1}^{S} \tilde{\sigma}_s$$

All the Monte Carlo results are given in Table 3 given the setting in Table 2.

Parameter	**Name**	**Value**
Sample Size	N	1000
Number of samples	S	360

Table 2: The parameters of the The Monte Carlo of the Hermite Series. There are $S = 360$ samples with each $N = 1000$ observations. observations.



(a) The mean estimator

(b) The std. deviation estimator

Figure 1: The estimators from the Hermite series collapse around the true values; the true mean is $\mu^* = 0$ and the true std. deviation is $\sigma^* = 3.74$. Data is coming from the Monte Carlo process.

True Distribution		**Hermite Estimator**	
Mean μ^*	Std. Deviation σ^*	Mean $\bar{\bar{\mu}}$	Std. Deviation $\bar{\bar{\sigma}}$
0.0000	3.7417	0.0006	3.7420

Table 3: Estimation results from the Monte Carlo Simulation. The estimator means almost equal the true parameter values. There are $S = 360$ samples with each $N = 1000$ observations.

3.4 Hermite Series estimation on eBay data

Now as the Hermite Series estimator is shown to be consistent, the estimator will be applied on the empirical eBay bidding data. The estimation is performed on the whole data set and on the various subsamples of $\mathbf{X}^F$ from Section 2.3. By performing the same estimation on different subsamples, the aim is to isolate the effects from different auction attributes (Wan, 2001). As in Song (2004), it is investigated whether the seller's reputation affects the mean and variance of the underlying distribution. Below, the subsamples are described. The results of the parameter estimation for all subsamples are shown in Table 4.

In **Subsample ALL** (baseline), the whole data set $\mathbf{X}^F$ and its prices are used for estimation.

In **Subsample A**, only entries where there have been at least 1 negative review are considered.

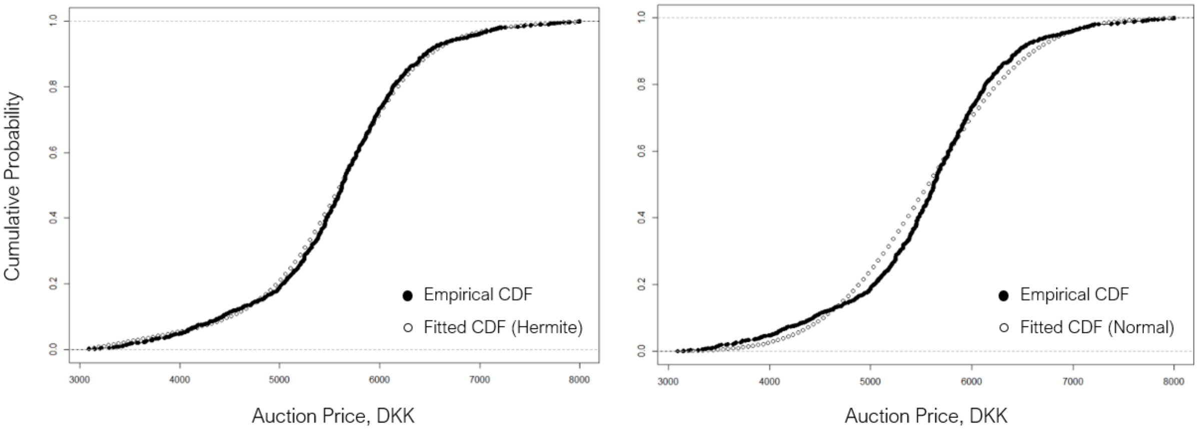


Figure 2: Comparison between the empirical CDF of the eBay data and the CDF of the Hermite Series, $k = 4$, (left) and the Normal Density (right). The Hermite Series is used as this is fitting more accurately.

In **Subsample B**, only entries where there have been at least 1 review in total for seller are considered.

In **Subsample C**, only entries where the seller has more than 10 existing reviews in total are considered.

In **Subsample D**, only entries where there have been no more than 10 bidders are considered.

In **Subsample E**, only entries where there have been at least 30 bids are considered.

In **Subsample F**, only entries whose duration is 1 or 3 days are considered.

In all estimations in Table 4, all the parameters are statistically significant at 1% or 5% confidence interval. The price premium are calculated as the relative mean difference of the squared values of a given subsample and the baseline. E.g., $\frac{5077.94^2}{5548.97^2} - 1 = -8.49\%$.

The results show that receiving a single non-positive review may affect the auction price negatively. Especially, having received at least one negative review lowers the price more significantly. Furthermore, the seller does not experience a lower price even though she has not received any feedback before. Actually, being a seller with more reviews may lower the price for the seller which contradicts the purpose of the feedback system and initial hypothesis.

The auctions having fewer than median bidders (below 10) experience on average lower end price. But opposite to the conclusion from Lucking-Reiley et al. (2007), having more bids may on average lower the end price. Furthermore, having longer auctions on 7 or 10 days on average leads to higher end price for the seller.

	Normal	Hermite						
	$\mathbf{p}^{\text{ALL}}$	$\mathbf{p}^{\text{ALL}}$	$\mathbf{p}^{A}$	$\mathbf{p}^{B}$	$\mathbf{p}^{C}$	$\mathbf{p}^{D}$	$\mathbf{p}^{E}$	$\mathbf{p}^{F}$
N	1434	1434	234	1252	742	746	440	408
$\hat{\mu}$	5565.4*** (0.001)	5565.4*** (0.000)	5143.2*** (0.000)	5569.9*** (0.000)	5363.6*** (0.000)	5505.5*** (0.000)	5486.2*** (0.000)	5484.7*** (0.000)
$\hat{\sigma}$	813.5*** (0.002)	813.9*** (0.000)	1480.5*** (0.000)	815.1*** (0.000)	1038.6*** (0.000)	7740.2*** (0.000)	1052.5*** (0.000)	888.5*** (0.000)
$\hat{a_1}$		0.098*** (0.021)	0.301*** (0.062)	0.131*** (0.023)	0.177*** (0.029)	0.082** (0.029)	0.125** (0.043)	0.155*** (0.047)
$\hat{a_2}$		-0.185*** (0.019)	-0.265*** (0.061)	-0.197*** (0.021)	-0.242*** (0.027)	-0.180*** (0.027)	-0.165*** (0.042)	-0.156*** (0.039)
$\hat{a_3}$		-0.034*** (0.006)	-0.117*** (0.021)	-0.045*** (0.007)	-0.062*** (0.008)	-0.029*** (0.008)	-0.042** (0.014)	-0.055*** (0.013)
$\hat{a_4}$		0.035*** (0.004)	0.068*** (0.014)	0.038*** (0.004)	0.048*** (0.005)	0.035*** (0.005)	0.038*** (0.009)	0.033*** (0.008)
Mean	5565.4	5548.97	5077.94	5552.51	5338.99	5486.95	5472.61	5469.74
Std. Dev.	813.5	836.87	1557.36	844.43	1060.15	793.57	1064.47	913.27
Premium	/	/	-8.49 %	0.06 %	-3.78 %	-1.12 %	-1.38 %	-1.43 %

Table 4: The parameter values when applying the Normal Density and Hermite Series on the eBay data. For Hermite, estimation of the whole data set and subsamples are included. $k = 4$. The standard errors in parenthesis. *** denotes statistic significance at 1 % level. ** denotes statistic significance at 5 % level.

3.4.1 Divide auction data set into price bins

Now let $\hat{f}(x;\hat{\theta})$ be the estimated density based on the whole data set, $\mathbf{X}^F$. The fitted parameters, $\hat{\theta}$, are shown in the third column in Table 4. Then the CDF will be derived using equation 3.

$$\hat{F}(x;\hat{\theta}) = \int_{\underline{x}}^{x} \hat{f}(u;\hat{\theta}) du$$

Having the CDF in hand, the auctions are now divided into $K = 4$ bins. First, the 3 quartiles of the price vector, $\mathbf{b}$, belonging to $\mathbf{X}^F$ are calculated using the inverse CDF function, $\hat{F}^{-1}(q)$.

$$\begin{aligned}
\hat{q_1} &= \hat{F}^{-1}(0.25; \hat{\theta}) = 5099.56 \\
\hat{q_2} &= \hat{F}^{-1}(0.50; \hat{\theta}) = 5597.65 \\
\hat{q_3} &= \hat{F}^{-1}(0.75; \hat{\theta}) = 6059.80
\end{aligned}$$

Class 1 will then comprise all auctions i for $i = 1, ..., 1434$ in $\mathbf{X}^F$ whose price is $\min \mathbf{b} < p_i \leq \hat{q_1} = 5099.56$. Class 2 will be all auctions whose price is $\hat{q_1} < p_i \leq \hat{q_2} = 5597.65$, and so forth. The comprehensive list of classes can be found in Table 5.

Price Bin	Lower Price	Upper Price
Class 1	$\min \mathbf{b}$	5099.56
Class 2	5099.56	5597.65
Class 3	5597.65	6059.80
Class 4	6059.80	$\max \mathbf{b}$

Table 5: The auctions in the whole data set are divided into $K = 4$ classes based on the auction price and the empirically estimated quartiles using $\hat{F}^{-1}(q)$.

3.5 Section Conclusion - Maximum Likelihood Estimation

The previous section alluded to that various auction attributes may affect the auction's end price. After showing the Hermite Series density estimator was consistent, it was applied to the eBay data and subsamples thereof. The mean of the different smaller data sets, subsampled based on auction attributes and other papers' hypotheses, show that both seller-specific and auction-specific attribution on average affect the end price either direction. Especially, will negative reviews lower the average price, and having many reviews as seller will too lower the expected price by more than 3 %. The preliminary conclusions from the MLE are somewhat aligned and deviating from existing literature contradicting the wanted effects of the feedback score.

Finally, the original auctions data set was divided into $K = 4$ classes based on the 3 quartiles from the Hermite Series cumulative density function, $\hat{F}(\cdot)$. Now, a non-parametric approach will try to predict the auction end price class based on the auction characteristics.

4 Artificial Neural Network and Multi-class classfication

4.1 Introduction and structure

In this section, the author applies Machine Learning (ML) methods to classify the auction end-prices based on auction and seller characteristics. The goal of this section is to investigate how big predictive power the auction attributes have on the end-prices. In the previous section, the auctions were divided into bins using the empirically estimated underlying distribution.

Due to its relative simplicity together with its widespread and diverse real-world applications, an Artificial Neural Network (ANN) is developed and applied for this classification problem (Abiodun, 2018). Being constructed of a network of non-linear functions, the algorithms are capable of tuning themselves and finding non-trivial relationships in the given data.

The upcoming analysis method is partly inspired by Ghani and Simmons (2004) whose purpose, too, is to classify eBay auctions based on characteristics. In addition, the below ANN setup is aligned with existing literature about Neural Networks and exhibit standard properties (Dreiseitl & Ohno-Machado, 2003). Hence, the below definitions of, e.g., loss function or activation function follow standard definition. The choice of hyperparameters and general structure will however be elaborated and discussed further when deemed relevant and needed. The extensive list of hyperparameters can be seen in Section 4.4 and in Table 7.

The ANN comprises 26 input neurons, 20 neurons in the hidden layer, and 4 neurons in the output layer. See Figure 3 for a visual representation of the ANN structure. Here, the number of input neurons will equal the number of regressors (after being One Hot Encoded) described in Section 2.3, and the number of output neurons will equal the number of bins the data initially is divided into. Specifically, the auctions are split into $K = 4$ price classes.

4.2 Predicting

When predicting a price-bin given the parameters at the current network state, the ANN forward propagates the data through the network of functions which results in so-called activations in the next layer, and, in turn, in the output layer. For example, will the neuron values of the hidden layer be dependent on (a non-linear function f of) the input values and the set of appurtenant parameters. In general, the activations in layer i, $a^{(i)}$, depends on the previous layer, $i - 1$. In this setting with 1 hidden layer, the activation in the hidden layer, $a^{(1)}$, and in the output layer $a^{(2)}$ will be:

$$\begin{aligned} a^{(1)} &= f(a^{(0)}) = \sigma(W^{(0)}a^{(0)} + b^{(0)}) \\ a^{(2)} &= f(a^{(1)}) = \sigma^{M}(W^{(1)}a^{(1)} + b^{(1)}) \end{aligned}$$

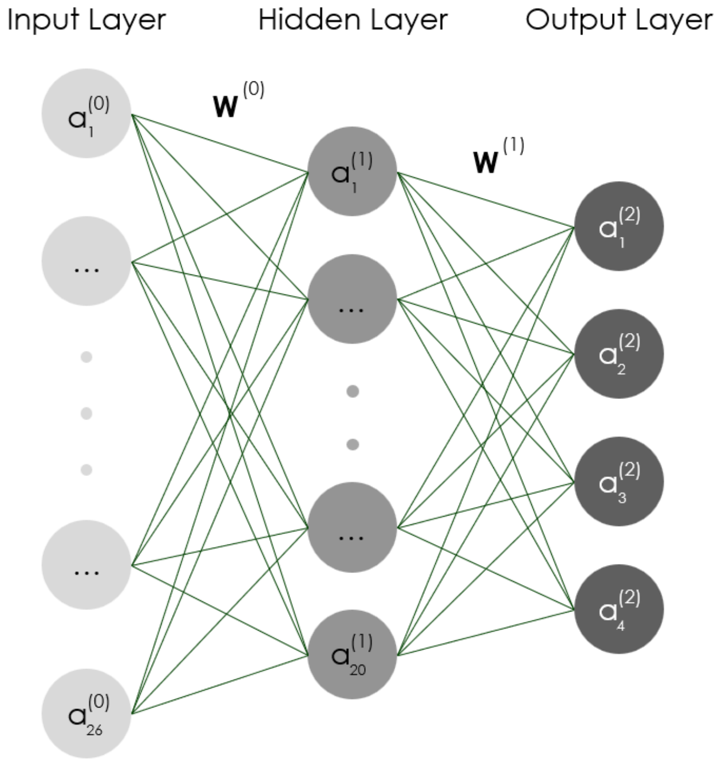


Figure 3: The ANN consists of 26 input neurons, 20 hidden neurons, and 4 output neurons which are connected by weights and biases. Fully connected neural network with 624 weights and biases in total.

where W refers to the weight matrix (node connections) at the given layer i, b denotes the bias vector at the given layer i and σ is the sigmoid function.

$$\sigma(x) = \frac{1}{1+e^{-x}} \in (0;1)$$

The activations in the input layer, $a^{(0)}$, are by definition the input values. For clarity, the dimensions of the above parameters and variables are seen in Table 6. As it is a dense neural network will fully connected layers, the network possesses 624 weights in total which is the sum of entries in both W's and b's, thereof 24

	$W^{(0)}$	$W^{(1)}$	$b^{(0)}$	$b^{(1)}$	$a^{(0)}$	$a^{(1)}$	$a^{(2)}$
Rows	26	20	20	4	26	20	4
Columns	20	4	1	1	1	1	1

Table 6: The shapes of the weight matrices, bias vectors, and activation vectors are aligned with number of input, hidden and output neurons.

bias terms and 600 node connections.

When computing the 4 activations in the output layer, $a^{(2)}$, the Softmax function, σ^M, ensures that the activations sum to one (and are positive) such that the final activations for each class can be interpreted as the probability that a given auction is belonging thereto. The Softmax function is defined as follows.

$$p_k = \sigma^M(\mathbf{z})_k = \frac{e^{z_k}}{\sum_{j=1}^{K} e^{z_j}}$$

where $\mathbf{z}$ here is the vector of $a^{(2)}$ and $k = 1, ..., K$ is the given class whose probability is calculated.

Lastly, in the forward prediction phase, the model's loss will be calculated. Here, the cross-entropy loss for multi-class classification is applied. The loss will be a function of data and a given set of model parameters, θ.

$$L(\mathbf{X}, \theta) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} y_{ki} \log p_{ki} \tag{5}$$

where y_{ki} denotes a scalar dummy $\mathbb{1}[y_i = k]$ and p_{ki} denotes the Softmax-transformed probability that auction i belongs to class k. As the loss, L, is an indirect measure of the neural network's precision, the goal is to minimize the loss.

4.3 Learning using Backpropagation

The power of Neural Network is the ability to self-adjust the 624 model weights such that the loss, defined above, is minimized. Here, the Backpropagation algorithm is used utilizing the Gradient Descent technique (Kostadinov, 2019).

Formally, training a neural network with gradient descent requires the calculation of the gradient of the loss function $L(\cdot, \theta)$ from Equation 5 with respect to the weights, $W^{(i)}$ and the biases, $b^{(i)}$ for $i = 1, 2$. According to the learning rate, η, each iteration of gradient descent updates the weights and biases (collectively denoted θ) according to:

$$\theta_{t+1} = \theta_t - \eta \frac{\partial L(\cdot, \theta_t)}{\partial \theta} \tag{6}$$

where θ_t denotes the parameters of the neural network at iteration t in the gradient descent algorithm. The gradient descent parameter update continues until one of the following conditions are met: (i) the maximum number of iterations (called *epochs*) are reached, $t = 1, ..., E$, or (ii) a convergence criterion - based on change in model loss - is met (here L is the loss function from Equation 5 and δ_L is a threshold hyperparameter):

$$\frac{L(\theta_t) - L(\theta_{t-1})}{L(\theta_{t-1})} \leq \delta_L = 0.001$$

To avoid that the weights in the network explode (become excessively large) and may overfit the data, regularization in the parameter update is applied (Chekka, 2018). The update rule stated in Equation 6 remains unchanged for the bias terms, $b^{(i)}$ but for the weight matrices, it will now be the following.

$$W_{t+1}^{(i)} = W_t^{(i)} - \eta \frac{\partial L(\cdot, \theta_t)}{\partial W_t^{(i)}} - 2\lambda W_t^{(i)} = (1 - 2\lambda) W_t^{(i)} - \eta \frac{\partial L(\cdot, \theta_t)}{\partial W_t^{(i)}} \tag{7}$$

where λ is the regularization hyperparameter.

4.4 Hyperparameters

When setting up an ANN, the overall structure of the model must be considered. The full list of hyperparamters and their values are shown in Table 7. The parameter values are chosen based on (i) existing literature and (ii) hyperparameter tuning through testing. That is, testing the model accuracy, robustness, and speed when applying a range of different parameter values applying a simple grid search.

Parameter	Name	Value
Hidden Neurons		20
Learning Rate	η	4
Number of Epochs	E	100
Loss Convergence Threshold	δ_L	0.0001
Regularization Constant	λ	0.0002

Table 7: The hyperparameters of the ANN. Values based on literature and parameter tuning

4.5 Monte Carlo simulation - ANN

To ensure that the Neural Network behaves as intended - namely that the forward feeding and backpropagation processes work correctly - a Monte Carlo simulation is performed. This will be an expansion of the previous Monte Carlo simulation of the Hermite Series in Section 3.3. Here, a series of N bids were generated from Equation 4:

$$B_t = a_1\alpha_t + a_2\beta_t + a_3 v_t \tag{8}$$

where $a_1 = 2, a_2 = -3$ and $a_3 = 1$. Recall that $\alpha_t \sim N(0,1), \beta_t \sim Exp(1)$ and $v_t \sim \Gamma(9,3)$. Now assume that in each auction the seller has set a secret reservation price, R_t. This price will, too, be a linear combination of the auction-specific characteristics. The following formulation of the reservation price is, as the previous Monte Carlo, inspired by Song (2004).

$$R_t = a_1\alpha_t + a_2\beta_t + \omega + a_1 \cdot \mathbb{1}\left[\beta > 1\right] - a_2\frac{\beta+1}{\alpha} \tag{9}$$

where $\omega \sim \Gamma(9,3) - 2$. As in real world eBay auctions, the bid is censored if the bid is below the secret reservation price. Hence, define the dummy $c_t \equiv \mathbb{1}\left[V_t < R_t\right]$.

Letting the sample size still be $N = 1000$, c_t will on average be $\bar{c} = 0.447$. That means, the baseline accuracy of guessing whether a bid is censored is $1 - \bar{c} = 0.553$ (always guessing that $c = 0$ as this is the biggest group).

The data is now tested on the ANN to investigate whether the Neural Network is able to identify the pattern for which bids are censored. The input data consists of the 4 explanatory variables in B_t and R_t and the variable to classify, $\mathbf{c}$, for $t = 1, ..., N$.

$$\mathbf{X} = \begin{bmatrix} \alpha_1 & \beta_1 & v_1 & \omega_1 \\ \alpha_2 & \beta_2 & v_2 & \omega_2 \\ ... & ... & ... & ... \\ \alpha_{1000} & \beta_{1000} & v_{1000} & \omega_{1000} \end{bmatrix} \qquad \mathbf{y} \equiv \mathbf{c} = \begin{bmatrix} c_1 \\ c_2 \\ ... \\ c_{1000} \end{bmatrix}$$

This Monte Carlo-adjusted Neural Network consists of 4 input neurons, 3 hidden neurons and 2 output neurons. This sizes of the weight matrices and bias vectors are changed accordingly. The train and test data set consists of 800 and 200 records, respectively.

With the test set, the prediction accuracy is on average across all S Monte Carlo iteration 97.5 % which is significantly higher than the baseline accuracy. The high accuracy shows that the Neural Network is able to identify the obscure pattern telling whether a bid is censored or not. And, hence, that the forward feeding and backpropagation works correctly.

4.6 The ANN classification applied to eBay data

The below data manipulation procedure is common practice and this was also applied in the Monte Carlo Neural Network to the fictional data.

As mentioned above in Section 2.3, 14 collected variables from the eBay site will be used as regressors in the Neural Network. As the computations exclusively can be performed on numeric variables, the data set's categorical variables (that is, the product's condition, the seller's country, and the weekday of auction end) will be transformed into numerical dummies applying One Hot Encoding.

Equivalently, the labels indicating the price class, will be one hot encoded, too. An example with $K = 4$ classes where the left vector is the auction labels vector would look like the following.

$$\begin{pmatrix} 1 \\ 3 \\ 2 \\ 4 \\ \dots \end{pmatrix} \xrightarrow[\textbf{Transform}]{\text{One Hot Encoding}} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ \dots & \dots & \dots & \dots \end{pmatrix}$$

After the columns are one hot encoded, there are 26 regressors. This encoded label matrix will then be size compatible in the calculations of the loss, $L(\cdot)$, as stated in Equation 5.

Then, the input data is standardized before initiating the model training process. To ensure steady and faster learning process (having non-exploding gradients), the data columns are scaled using the column mean and standard deviation (Brownlee, 2020). Hence, for all columns $i = 1, .., 26$ in the input, $\mathbf{X}$, the standardized input data is

$$\tilde{\mathbf{X}}_i \equiv \frac{\mathbf{X}_i - \mu_i}{\sigma_i} \sim N(0, 1) \quad \text{where} \quad \mu_i \equiv N^{-1} \sum_{j=0}^{N} \mathbf{X}_{ji} \qquad \sigma_i \equiv \sqrt{N^{-1} \sum_{j=0}^{N} (\mathbf{X}_{ji} - \mu_i)^2}$$

Lastly, the cleaned data set is split into a training and test set. The training process will be based on $\delta = 80\%$ of the data, and model accuracy will be validated and compared using the designated test data.

4.7 Classification Results

After successful parameter fitting of the Neural Network, the model has a classification accuracy at 73.9% (see Table 11 in the appendix). That is, the model is in 3 out of 4 cases able to predict which of the four price bins (classes) a given auction belongs to – this is three times higher than if someone were to randomly guess among the four price bins. If the same test is performed on the test data (which the model has not encountered yet), the classification accuracy is 54.1%. This difference between the training and test accuracy suggests that the model somewhat overfits the given training data. Even though the overfitting is still apparent, its significance has been lowered by reducing the number of hidden neurons and by applying regularization of the parameter weights as shown in Equation 7.

Having a classification accuracy around 55% - 75%, it is seen that auction attributes – listing heterogeneity and seller characteristics – explain a great fraction of price variation. This includes the listing attributes themselves in isolation and the non-trivial, subtle combinations of them. At the same time, the fact that the predictive power is not higher alludes to that strategic behavior and earlier-mentioned aspects in Section 2, too, may be a part of the price determinants. This will be discussed in the next section.

Looking at the analysis of the outcome prediction also gives some insight. The following results are based on the hypotheses used in Section 2 and 3 and specifically the subsampling in Section 3.4. See the results in Table 8. E.g. the first entry in the table shows that auctions having at most 10 bidders are predicted, by the ANN, to have a -4.27% lower price class (ranging from $1-4$, 1 being the lower price) on average. Simultaneously, the MLE estimates from Section 3.4 show that auctions with at most 10 bidders, on average, have -1.12% lower price.

First, it is seen that the sign of all the hypothesis- / subsample-estimations are identical comparing the MLE and ANN. Hence, the ANN model backs up the initial MLE tests.

Summing up the results in Table 8; it is favorable for the seller to have many bidders and longer auctions yielding higher end prices (higher price class). At the same time will fewer auction bids and having fewer reviews as seller ensure, on average, higher end prices.

		Model premiums	
Auction Feature	Subsample	ANN Class Prediction	MLE Price
Number of bidders	≤ 10 bidders	-4.27%	-1.12%
Duration	≤ 3 days	-2.45%	-1.43%
Number of bids	≥ 30 bids	-2.98%	-1.38%
Number of reviews	≥ 10 reviews	-4.37%	-3.78%
Number of negative reviews	> 0 reviews	-0.74%	-8.49%

Table 8: Using the hypotheses stated in Section 3.4. The ANN Class Prediction Premium indicates how the auction class prediction depends on the auction feature values. E.g., auctions having few bidders are predicted to have a lower price class. The MLE premiums are taken directly from Section 3.4.

Even though the ANN model predicts higher price classes due to auction attributes, the results are not fully unambiguous. Considering longer duration or having 0 negative reviews, the *average* price class increases, but the probability of belonging to the highest price class ($k = 4$) decreases. The same is experienced regarding having more than 30 bidders, just opposite sign. And the conclusions being equal to the MLE's, having more reviews or more bids are counter-intended lowering the average auction price implying further ambiguity of the feedback system.

This analysis does not yield explicit regression coefficients from which you can conclude or infer attributes' direct contribution to price determination. Instead, the results inform researchers that the price variation, indeed, partly is a complex non-linear function of auction characteristics with no given parametric structure.

5 Discussion and Conclusion

5.1 Discussion

Even though the results from the Maximum Likelihood and Artificial Neural Network estimations allude to that auction features affect, or partially determine, the end price, the results are somewhat not decided or economically significant. This ambiguity is also obvious across studies on the price effects of auction features. See Table 10 in the appendix for a literature overview; taking into account that the papers use different methods and different data, there is no clear pattern which and how much different auction features affect the auction price. Hence, this report's estimations align with comparable study results.

This paper aims to describe and estimate the price dispersion given auction heterogeneity and adverse selection. Due to the rise of high-speed online marketplaces and costless search possibilities, bidding behavior may not solely be a function of characteristics of a single auction anymore but, too, be dependent on outside options: buy-it-now prices of other products, completed auctions, and concurrent auctions of similar items. As this analysis does not include or evaluate the mentioned outside options, the correlation between end price and auction attributes may also be due to outside factors (Anwar et al., 2004).

Even though the size, scope, and level of detail of the data set collected in this paper correspond to other papers', the above estimations may be limited by the data size. Applying Neural Networks - which is a big data estimation method requiring great loads of data - the analysis may be prone to overfitting and spurious correlations. This is directly apparent in the variation between the train and test data accuracy.

In addition, the analysis is solely based on one product. Including more products in the analysis would lead to more resistant and robust conclusions (Bodoh-Creed et al., 2004) but was left out due to simplicity and scope. And as this paper indirectly builds upon the analysis originally performed by Akerhof in 1970 (Akerlof, 1970) shedding light on the adverse selection problem in the used car market, including other product categories might be of high relevance.

Experiencing the same as Lucking-Reiley et al.(2007), the collected data set exhibits low variation in the seller feedback percentage across the auctions. Having a mean feedback score well above 90 percent, isolating the effects of high and low feedback scores will be difficult. Hence, the data composition complicates the adverse selection analysis and other features such as number of bids or bidders are preferred in the analysis.

As mentioned earlier, a substantial part of the apparent price dispersion – taking aside information asymmetry - may be due to behavioral aspects or irrational market agents. These behavioral angles could be included in further analysis. This could be done by firstly quantifying the strategic behavior such that it quantitatively would be included in an econometric model, e.g. creating dummies whether *jump biding* or *bid sniping* was performed in a given auction.

5.2 Conclusion

Over the last couple of years, online auctions on marketplaces like eBay have gained popularity but have also given rise to information asymmetries. Contradicting the Law of One Price, it is apparent that the product of interest in this paper, an iPhone 12 Pro Max, exhibit a lot of price dispersion. Referring to other papers and micro-founded auction theory, the auction inefficiencies can be explained by bidders' strategic behavior, information asymmetries appearing on online platforms, and/or lack of agent rationality.

First, a list of completed auctions consisting of various auction features is collected from the eBay marketplace using a web scraper. Applying Maximum Likelihood Estimation on the data set, it is seen that there are price differences when subsampling the data based on the above auction characteristics, listing seller-specific variables and auction heterogeneity. For example, on average, receiving negative feedback will lower the end price significantly, and having longer auctions or many bidders will increase the end price significantly. The estimation suggests that a given auction's features can affect the average price between $1\% - 8\%$.

Going a step further using Machine Learning methods, the model is able to predict a given auction's price class three times more accurately than at random. This confirms the hypotheses from the Maximum Likelihood Estimation (MLE) that the auction attributes, and complex combinations of them, *do* have some explanatory power on the price. The Neural Network further backs up the estimates of the MLE stating that the number of bidders positively and the number of reviews negatively affect the end price. The last-mentioned contradicts the results of other papers and the purpose of eBay's feedback system. The results regarding the adverse selection and feedback scores may be ambiguous due to (a) low variance in the feedback scores across auctions and (b) microeconomic inefficiencies in the feedback system such as free-rider problems or negative reviews acting as proxies.

The predictive power and significance of price determination are also limited by inestimable behavioral aspects and strategic behavior, such as bid sniping, eroding the effects of auction features. And due to the rise of the internet (which also enables online marketplaces), bidding behavior may also depend on outside options and cannot be modeled statically and in isolation as in this paper. Performing the same analysis over time or cross-sectional including other products would increase the robustness of the analysis conclusions.

References

Abiodun, O. I. (2018). *State-of-the-art in artificial neural network applications: A survey.* Heliyon.

Akerlof, G. A. (1970). *The market for "lemons": Quality uncertainty and the market mechanism.* The MIT Press.

Anwar, S., McMillan, R., & Zheng, M. (2004). *Bidding behavior in competing auctions: Evidence from ebay.*

Bajari, P., & Hortacsu, A. (2000). *Winner curse, reserve prices and endogenous entry: Empirical insights from ebay auctions.* Stanford Institute for Economic Policy Research.

Bajari, P., & Hortacsu, A. (2004). *Economic insights from internet auctions.* Journal of Economic Literature.

Bodoh-Creed, A., Boehnke, J., & Hickman, B. (2004). *Using machine learning to predict price dispersion.* University of British Columbia.

Brownlee, J. (2020). *How to use data scaling improve deep learning model stability and performance.* Retrieved 2022-06-01, from `https://machinelearningmastery.com/how-to-improve-neural-network-stability-and-modeling-performance-with-data-scaling/`

Chekka, V. (2018). *Regularization in machine learning: Connect the dots.* Retrieved 2022-06-01, from `https://towardsdatascience.com/regularization-in-machine-learning-connecting-the-dots-c6e030bfaddd`

Coppejans, M., & Gallant, A. R. (2002). *Cross-validated snp density estimates.* Journal of Econometrics.

Dreiseitl, S., & Ohno-Machado, L. (2003). *Logistic regression and artificial neural network classification models: a methodology review.* Journal of Biomedical Informatics.

Eaton, D. H. (2005). *Valuing information: Evidence from guitar auctions on ebay.* Murray State University.

Foster, J. (2021). *Semi-nonparametric estimation of secret reserve prices in auctions.* Ivey Business School, Western University.

Gallant, A. R., & Nychka, D. W. (1987). *Semi-nonparametric maximum likelihood estimation* (Vol. 55) (No. 2). Econometrica.

Gentry, M. L., Hubbard, T. P., Nekipelov, D., & Parsch, H. (2018). *Structural econometrics of auctions: A review* (Vol. 9) (No. 2-4). Foundations and Trends in Econometrics.

Ghani, R., & Simmons, H. (2004). *Predicting the end-price of online auctions.* Accenture Technology Labs.

Gonzalez, R., Hasker, K., & Sickles, R. C. (2004). *An analysis of strategic behavior in ebay auctions.*

Kalyanam, K., & McIntyre, S. (2001). *Return on reputation in online auction markets.* Santa Clara University.

Kim, K. I., & Lee, J. (2014). *Nonparametric estimation and testing of the symmetric ipv framework with unknown number of bidders.*

Kostadinov, S. (2019). *Understanding backpropagation algorithm.* Retrieved 2022-06-01, from `https://towardsdatascience.com/understanding-backpropagation-algorithm-7bb3aa2f95fd`

Lucking-Reiley, D., Bryan, D., & Reeves, D. (2007). *Pennis from ebay: the determinants of price in online auctions.*

Roth, A., & Ockenfels, A. (2002). *Last-minute bidding and the rules for ending second-price auctions: evidence from ebay and amazon auctions on the internet.* American Economic Review.

Rousseeuw, P. J., & Hubert, M. (2011). *Robust statistics for outlier detection.* John Wiley and Sons, Inc.

Saeedi, M. (2014). *Reputation and adverse selection: Theory and evidence from ebay.* The Ohio State University.

Song, U. (2004). *Nonparametric estimation of an ebay auction model with an unknown number of bidders.* University of British Columbia.

Wan, H.-H. T. W. (2001). *An examination of auction price determinants on ebay.* University of Singapore.

6 Appendix

6.1 Appendix A: Auction Summary Statistics

Below, the summary statistics are shown for the auctions collected on the eBay marketplace. Only the numerical variables are shown. In addition there are the following auction attributes: *Item Condition* and *Seller's Country.*

	Type	Mean	Std. Dev.	Minimum	25%	Median	75%	Maximum
Auction Price, DKK	Continuous	5436.1	1288.3	416.0	5057.0	5604.0	6045.5	15203.0
Auction End Weekday	Categorical	3.2	2.1	0.0	1.0	3.0	5.0	6.0
Auction End Weekend	Dummy	0.3	0.5	0.0	0.0	0.0	1.0	1.0
Seller's Feedback Score, Pct	Continuous	98.6	5.7	33.3	100.0	100.0	100.0	100.0
Seller's Positive Reviews	Discrete	169.9	1098.8	0.0	5.0	14.0	38.0	10866.0
Seller's Neutral Reviews	Discrete	0.5	3.2	0.0	0.0	0.0	0.0	29.0
Seller's Negative Reviews	Discrete	0.4	1.8	0.0	0.0	0.0	0.0	22.0
Seller's Membership duration	Discrete	3772.0	2438.2	6.0	1649.0	3583.0	5944.3	8841.0
Number of bidders	Discrete	11.1	6.3	1.0	6.0	10.0	15.0	43.0
Number of bids	Discrete	25.4	18.7	3.0	12.0	21.0	34.0	113.0
Auction Duration, Days	Discrete	5.4	2.4	1.0	3.0	7.0	7.0	10.0
Has extra item description	Dummy	0.1	0.3	0.0	0.0	0.0	0.0	1.0

Table 9: Summary statistics of the auction attribute variables collected from eBay marketplace between May and June 2022. $N = 1551$ auctions were collected in total. Only numerical variables are shown. Some variables are directly accessible on eBay and some are derived thereof.

In Figure 4, the distribution of the auction prices are shown. Since the price distribution resembles a bell curve, a normal density is applied in the Hermite Series estimation.

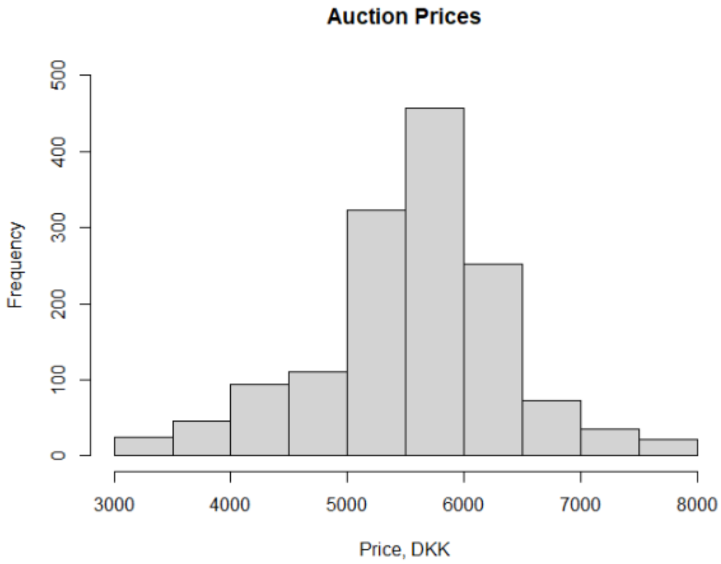


Figure 4: The auction prices collected from eBay. From looking at the distribution, a Normal Density is used in the Hermite Series for fitting the prices.

6.2 Appendix B: Literature Overview of price determinants

In the economic literature there is an obvious ambiguity regarding how or whether auction attributes actually affect the auction price. Below in Table 10, different studies are included and their conclusions are shown.

Auction Feature	**Affects price**	**Does not affect price**
Auction duration	(Lucking-Reiley et al., 2007)	(Gonzalez et al., 2004) (Wan, 2001)
Feedback Scores	(Lucking-Reiley et al., 2007) (Kalyanam & McIntyre, 2001)	(Eaton, 2005)
Negative reviews	(Eaton, 2005) (Positively)	(Bajari & Hortacsu, 2000)
Number of bidders	(Wan, 2001) (Bajari & Hortacsu, 2000)	
Number of bids	(Bajari & Hortacsu, 2000)	(Gonzalez et al., 2004)
Ending on weekend		(Wan, 2001) (Lucking-Reiley et al., 2007)
Reserve Price	(Lucking-Reiley et al., 2007) (Wan, 2001)	

Table 10: In the economic literature, there is a lot of ambiguity whether some auction features actually affect the auction's end price.

6.3 Appendix C: ANN Prediction Accuracy

To test whether the auction features do have predictive power, the Neural Network prediction is performed with the full model and restricted models. Comparing the results of the full model with models omitting some auction features, is it illustrated whether an auction feature enhances the price prediction.

As seen in Table 11, the full model has, as it should, higher prediction accuracy. By omitting the given auction attributes (single at a time) listed in the table, the predictive power falls by 3-5 % which can be seen as a significant drop.

		Omitted features			
Data Set	Benchmark	Feedback Score	Duration	Bids	Number of bidders
Train set	73.9%	70.3%	68.6%	70.6%	71.2%
Test set	54.1%	55.3%	54.9%	58.5%	55.3%

Table 11: The ANN's accuracy for the full model (Benchmark) and robustness checks omitting some auction features. The full mode has better accuracy but also higher overfitting.