Calibration of an Optimal Bidding Model for the Mobile Advertisement Markets

A Major Qualifying Project Report:

Submitted to the Faculty of the

WORCESTER POLYTECHNIC INSTITUTE

in partial fulfillment of the requirements for the

Degree of Bachelor of Science

by

Laura Antul
Advisors:

Stephan Sturm

Marcel Blais

Sponsor:
Chitika, Inc.

Abstract

One goal of every business is to save money by minimizing cost and maximizing profit. Cidewalk is a mobile advertisement company that wishes to accomplish this goal through implementing an optimal bidding strategy. By collecting and analyzing market data, it was possible to identify the distribution that best represents the second best bids, which is the price Cidewalk pays for an advertisement space by bidding in a Vickrey auction. Implementing a simulation for the optimal bidding model that utilizes Bayesian updating methods ensures that the model responds to any changes in the distribution parameters. The model's performance was evaluated by simulation and throughout one day the implemented model yielded 17% savings when compared to Cidewalk's current bidding model.

Acknowledgements

This project was conducted in collaboration with a Financial Mathematics Master's student, Anastasiia Parkhomenko, therefore, this report will be the same as or very similar to her Master's Project Report.

We would like to begin by thanking our advisors, Professor Marcel Blais and Professor Stephan Sturm, for providing invaluable insight whenever we encountered roadblocks in our project and always being there to provide vital feedback on our work.

We would also like to thank Cidewalk for sponsoring this project and specifically Venkat Kolluri, for not only providing us with his time and all available resources but also for providing critical feedback that helped us grow not only as students, but as young business professionals.

Finally, we would like to thank Joe Regan for showing us how to collect all the information we needed and sharing his valuable opinions and ideas with us during our weekly meetings.

Contents

List of Figures

1 Introduction

Cidewalk is a mobile advertisement company that helps small businesses grow by providing them with an opportunity to reach potential clients through placing their advertisements in popular mobile applications. Cidewalk's clients are small business owners that want to increase their clientele and make locals aware of their business, and at the same time do so within a budget. In order to provide advertisement spaces, Cidewalk buys the right to show clients' advertisements on exchanges that use Vickrey auctions to identify the bidder who wins a space. In these type of auctions, interested market participants bid on the auctioned good and the one who places the highest bid wins the auction; however the price they pay is equal to the second highest bid. Cidewalk, like most businesses, seeks to minimize their cost and maximize their profit which created a need for a model that can determine the optimal bid price strategy that minimizes the cost of purchasing advertisement spaces.

Hu and Macaluso [1] were focused on creating a theoretical optimal bidding model that determined an optimal bidding strategy for Vickrey auctions. Its focus was on understanding how to assess bidding strategies where the bids follow a specific distribution and how to use this distribution to minimize the overall cost required to win the necessary number of Vickrey auctions. Through working on developing strategies for specific distributions they were able to create a method that worked for any distribution, and that was the method that was eventually put to the test by the sponsor during a live test. However, due to assumptions made about the distribution that the data followed, the live simulation did not prove as effective as it was expected to be, and therefore, further analyze of sample data sets coming from the mobile advertisement market is necessary. In short, the distribution of competitors' bids used in the model needed to be calibrated to market conditions.

We begin by addressing the need to understand the actual distribution of second best bids, which represents the price at which an auction is won, in order for the the bidding model created by Hu and Macaluso [1] to produce accurate optimal bids. Since the overall actual distribution of second best bids does not follow any specific distribution it is necessary to understand how the sample data set is impacted by certain criteria such as which exchange the auction occurred on, or the geographic location of the auction. Analysis of recently collected sample data sets indicates that the distribution differs based on the exchange that the auction occurs on within the mobile advertisement market. Identifying the distribution that best fits each exchange, allows for the creation of a simulation to understand the performance of the model in a real-world setting. Since the model in [1] is heavily dependent on the parameters of the distribution of second best bids, it is necessary to determine a method for updating these values to ensure that the model produces accurate optimal bids as the mobile advertisement market is ever-changing. To address the volatile market, Bayesian updating is used on the distribution parameters to enhance the performance of the model created in [1]. Another issue comes from the exchanges deciding not to disclose the winning price of the auctions to the bidders. This results in missing sample data, since the second best bids is only known when Cidewalk wins the auction. Using the distribution of observed second best bids, it is possible through the acceptance-rejection method to create reasonable estimates of the missing values.

As a result of the final simulation, which is discussed in detail in Section 3.5.1, it is discovered that the model is effective at reducing the overall cost to Cidewalk when compared to the current model Cidewalk is using. The savings were significant enough to justify a live test of the model.

2 Background

This section outlines Cidewalk as a company, and explains how their business works from both the customer's perspective and from the company's perspective including necessary background information on the mobile advertisement market. Relevant information from previous work is also discussed.

2.1 Cidewalk's Mobile Advertisements

Cidewalk is an online advertisement company that helps small businesses reach potential local customers through different mobile applications such as Angry Birds, Accuweather, Pandora, etc. For many business owners this type of advertisement is advantageous over traditional advertisements such as journals, newspapers, billboards or marketing companies. The cost of Cidewalk's services allows businesses of all sizes and income levels to run their advertisement campaigns within their budgets. For example, right now business owners can get their advertisement seen by 10,000 potential customers for as low as $20 a month. Another advantage is that Cidewalk can target potential customers based on their zip code. If a business owner wants to advertise his/her café, or small shop, in Worcester, MA, it is more effective to show the advertisement to mobile users in Worcester, MA than in Westborough, MA. Hence, only potential local customers are targeted. Moreover, Cidewalk's clients can view the quality of their services by tracking how many people have seen their advertisement.

2.1.1 Cidewalk From a Customer's Perspective

We consider an illustrative example. Suppose Lisa recently started her own business. She is a good cook, so she decided to open a little bakery, which also serves tea and coffee. The bakery is new, so Lisa started to look for advertisement options that would suit her budget to raise awareness of the place and attract customers. After spending many long hours searching for a suitable option, she found an online advertisement company that provides services, the cost of which is within her budget. She registered online, specified the location where she wanted the advertisement to be shown, entered the content for the advertisement including the name of the shop, address, website, phone number, and shop logo. At that time Ryan, who lived a couple of blocks from Lisa's bakery, was checking the weather for the next day through his Accuweather application. As usual there was an advertisement that popped up at the bottom of the screen, and this time it showed Lisa's bakery. Ryan noticed that the bakery was very close to his house, so the next day he decided to visit it. During the next couple of weeks she noticed that the number of customers entering her bakery increased, and some even mentioned that they found out about the bakery through the advertisement on their phones. This is the type of service that Cidewalk provides.

2.1.2 Cidewalk From an Internal Perspective

The moment Lisa purchased the monthly plan from Cidewalk and submitted the advertisement information, Cidewalk's servers started to buy advertisement spaces within mobile applications via the online advertisement market to display Lisa's advertisement. This online advertisement market is an exchange with a series of auctions. For each auction a seller provides an advertisement space and buyers offer their price, or bid, for this space. Cidewalk bids in auctions along with other market participants to get 10,000 advertisement spaces for Lisa's order. Whenever Cidewalk wins an auction, Lisa's advertisement appears on the phone screen of the a mobile application user. However, there is one characteristic about online advertisement markets that is worth mentioning: these markets use Vickrey auctions to decide which participant gets the space to show the advertisement.

2.1.3 Vickrey Auctions

Vickrey auctions are a type of sealed-bid auction where all participants bid for the lot at the same time without knowing how their competitors are bidding. As a result, the winner of the auction is the bidder with the highest bid. However, the winner does not pay the highest bid, but the second-highest bid amount. This type of auction encourages the participants to bid exactly how much they value the good being sold. If participant A bids higher that his/her valuation of the auctioned good, then there might be another participant, participant B, whose bid is lower than that of participant A's, but higher than participant A's valuation. Hence, participant A will get the good for more than he/she values it to be worth. Now consider the opposite situation, where participant A's bid is lower that his/her valuation of the auctioned good. Again, in this case there might be another participant whose bid is higher than participant A's bid, but lower than participant A's valuation. As a result, another participant wins the auctioned good and pays less than participant A's valuation. Therefore, it is always optimal for participants to place the bid that represents their own valuation of the good. This causes the good to be received by the bidder who values the product the most, which benefits the seller as the good he/she sold is not undervalued[2].

Returning to our example, Cidewalk bids on an auction selling one advertisement space. Before Cidewalk bids, they receive all available information about the user of the application (e.g., their location, phone model, gender, age, etc.). If the user of the application is located near Lisa's shop, Cidewalk considers buying this space and bids in this auction. Suppose Cidewalk's bid is $2. After Cidewalk submitted their bid, there were 2 possible outcomes. In one case, assume the opponents bid $1.5, $0.75 and $1.15, respectively. Since Cidewalk's bid was the highest, they win the auction and pay the second highest bid, which is $1.5. In another case, assume the opponents bid $2.5, $1.78 and $1.77, respectively. Cidewalk lost this auction and the only information they received was that the winner paid a price that was higher than or equal to their own bid.

As a result, the main disadvantage for bidders participating in Vickrey auctions within the online advertisement market is the inability to obtain information about other market participants' bids. The only information that can be obtained is the value of the second-highest bid, if the auction was won, or that the highest bid was made by another unnamed market participant, if the auction was lost. There are also two other restrictions that should be taken into account. First, online advertisement markets do not allow one participant to win all available auctions. Therefore, there might be a situation when the participant's bid was the highest, but he/she did not win due to the amount of auctions he/she had already won. The second issue is that some applications set a minimum bid price for their auctions, also called a floor price. Therefore, any bidding strategy is influenced by this floor and can be detrimental when trying to determine an optimal bid.

2.1.4 Ad Impressions

This subsection intends to familiarize the reader with the notions that are common in the world of online advertisement and will be referenced throughout this paper. An ad impression is a unit of measurement indicating how many times an online advertisement appeared on a web page. We say that the online advertisement company delivers, and its client receives, 100 ad impressions if the advertisement appeared on web pages/mobile applications 100 times. It does not necessarily need to be seen or clicked on by the user who visited the web page, as sometimes an advertisement is shown at the bottom of the page and the user might not scroll down to see it. Another term that is used in the online advertisement market is CPM. CPM, or cost per thousand impressions, is the cost of delivering 1,000 ad impressions for the client. If Cidewalk spends 2 CPM, then it costs $2 for 1,000 ad impressions, or 0.2 cents per ad impression.

One of Cidewalk's monthly plans allows their clients to receive 10,000 ad impressions per month for $20. That means that clients who choose this plan will get 10,000 ad impressions within a month. Therefore, whenever the client's advertisement appears on the screen of a mobile application user, Cidewalk delivers one ad impression for the client. In the case of the bakery shop example, Lisa, as a client of Cidewalk, received one ad impression when Ryan saw her advertisement.

2.2 Review of the Theoretically Optimal Bidding Model

Hu and Macaluso [1] focus on discovering a theoretically optimal bidding strategy for Vickrey auctions. Due to the structure of Vickrey auctions it is important to understand the strategy the opponents use to bid on any given auction. Hu and Macaluso focus on first understanding how to assess bidding strategies where the bids follow a specific distribution and how to use this distribution to minimize the overall cost required to win the necessary number of Vickrey auctions. As these auctions happen within seconds, it is also important that they take into consideration the speed at which the distribution of the second best bids can be analyzed and utilized to find an optimal bid.

2.2.1 Analysis In the Case of Specific Distributions

Hu and Macaluso focus on three possible distributions that the second best bid can take to see if these distributions are accurate estimations of the actual trend of the second best bids based on the sample data collected. The specific distributions that they focus on are uniform, binomial, and multinomial distributions. Analysis of the uniform and binomial distributions result in a formula capable of determining the minimum cost of winning the required number of auctions promised to the client. However, analysis of the multinomial distribution results in a formula too complex to be utilized. Although these discoveries are useful, the discovery most pertinent to this project is the derivation of a bidding strategy that can be used to minimize cost regardless of the distribution that the second best bid follows.

2.2.2 Analysis for All Distributions

Hu and Macaluso determine how to bid optimally regardless of the distribution that the opponents' bids follow by understanding that a company will not place a bid on an auction if they will end up spending more to win that auction than they would to win the next auction. If the cost of winning an auction is lower than the expected cost of winning the next auction, then the obvious conclusion is for the company to bid to win. The following derivation shows how they attain the optimal bidding strategy.

Assume we are aware of the cost we will be charged upon winning an auction when we have not yet bid on that auction. The cost will be denoted as $W(n)$, $E(r, n)$ denotes the expected value, or expected cost, of winning the required number of auctions, r represents the number of auctions that Cidewalk must win to provide the accurate number of advertisements to the client, and n is the total number of auctions available to bid on within an exchange over some fixed period of time.

In the case that the auction is lost there is now $n-1$ auctions available to bid on and the new expected cost is the cost of winning r out of $n-1$ auctions which we represent by:

$$E(r, n-1)$$

If an auction is won, the number of auctions that Cidewalk must win is reduced by 1 and the number of auctions available is also reduced by 1. Therefore, the expected cost is now the cost of winning $r-1$ out of $n-1$ auctions and the price that the specific auction was won for, $W(n)$, must be added to the equation:

$$W(n) + E(r-1, n-1).$$

Now applying what we know about the company and keeping its best interest in mind, if there was one auction less available to win this will be represented by:

$$W(n) + E(r-1, n-1) < E(r, n-1), \text{ which is equal to:}$$

$$W(n) < E(r, n-1) - E(r-1, n-1).$$

If the win price exceeds the expected overall cost of winning the necessary number of auctions, then it is more beneficial for the company to lose the auction.

$$E(r, n-1) < W(n) + E(r-1, n-1), \text{ which is equal to:}$$

$$E(r, n-1) - E(r-1, n-1) < W(n).$$

If the win price is equivalent to the expected overall cost of winning the necessary number of auctions, then it doesn't matter if the auction is won or lost.

$$E(r, n-1) = W(n) + E(r-1, n-1), \text{ which is equal to:}$$

$$W(n) = E(r, n-1) - E(r-1, n-1). \tag{1}$$

Equation (1) always takes the companies best interest into consideration by determining whether it is better to bid to win or to simply lose based on the expected cost of the overall number of auctions. Therefore, it represents the optimal bid.

The following piecewise function represents the various cases based on the number of auctions required to satisfy the customer and the number of advertisements available to bid on within an exchange. If the number of available auctions is equal to the number of auctions Cidewalk needs to win then the optimal bid is infinitely large. If no auctions must be won for Cidewalk to fulfill their required number of advertisements then the optimal bid is zero since they do not need to win any more auctions. In every other case the optimal bid is the expected cost of winning subtracted from the expected cost of losing since the difference of these two values would ensure that the company only bids to win when it is in their best interest.

$$X(r,n) = \begin{cases} 0 & r=0 \\ \infty & r=n \\ E(r,n-1)-E(r-1,n-1) & else \end{cases} . \tag{2}$$

From equation (2) a recursive formula was created to determine the minimum overall estimated cost required to meet the needs of the client. The use of order statistics then increased the speed at which the optimal bid price was computed since it did not require the estimated cost to be recalculated for every single auction.

They [1] also produced a computationally efficient algorithm to determine the optimal bids. This algorithm produces the amount that Cidewalk should bid in order to win r out of n auctions given that the second highest bids are independent from one another and identically distributed:

$$\int_0^\infty N((n-1)P(x),(n-1)P(x)(1-P(x)),r)dx \tag{3}$$

where $N(\mu,\sigma^2,r)$ is the normal cumulative distribution function with mean μ and variance σ^2, and $P(x)$ is the probability distribution of the best bid of the others.

2.2.3 Analysis of Computational Efficiency

Computational analysis [1] supports the fact that order statistics significantly reduces the run-time of the algorithm required to calculate the optimal bid. The formulas derived to find the optimal bid in the case of a uniform distribution, a binomial distribution, and using order statistics are used to calculate the number of iterations or run-time required to find the optimal bid in the worst case scenario. Order statistics has a constant run-time whereas analysis in the case of a uniform distribution even using data structures that reduce the run-time still has a run-time exponentially greater than order statistics. The run-time associated with analysis in the case of a uniform distribution increases as the number of auctions n available to bid on increases, therefore making this method even less efficient. In the case of a binomial distribution the run-time is still less efficient than the method utilizing order statistics, but more efficient than analysis in the case of a uniform distribution. Analysis of a binomial distribution has a run-time reliant on the number of required winning bids and the number of remaining available auctions and is equal to the multiple of these two values which is far less efficient than a constant run-time.

Due to the fact that the most computationally efficient algorithm utilizes order statistics it is not surprising that they suggest this method be used if Cidewalk decides to use their model. They then test four numerical methods to determine which method proves more efficient in calculating the optimal bid using the order statistics method. The numerical method that is most efficient after analyzing the test results is the Composite Trapezoidal Rule.

2.2.4 Test Results

Hu and Macaluso are then able to implement their methods by programming the aforementioned algorithms using the programming language preferred by Cidewalk, Erlang. After the algorithms are written they begin testing to see how effective their methods are in calculating the optimal bidding strategy in two areas, New York, and Boston. The results of the tests show that their methods are effective in the Boston marketplace since the cost of the campaign is reduced from the default strategy Cidewalk uses by 18.4%. Whereas the campaign for New York, they are unable to come close to fulfilling the number of target wins, only attaining 0.5% of their target number of wins.

2.3 Implementation of the Model

Hu and Macaluso [1] develop a theoretically optimal bidding strategy based on the format of Vickrey auctions. Due to the theoretical nature of their focus they make many assumptions throughout the project that may not accurately represent the real-world application of an optimal bidding strategy for this kind of auction. This could be the cause of their lack of success in New York. In order for Cidewalk to be able to utilize the model, the various assumptions regarding the distribution of the second best bids, and the fact that every bid placed within an auction, whether it is Cidewalk's bid or the bid of one of their many competitors, affects the overall trend of the second best bids, need to be addressed.

2.3.1 Addressing Assumptions Regarding Distribution of the Second Best Bids

The win price is what must be paid for the advertisement. In order to minimize cost, it is important to understand the trend the win price takes and what factors are influencing this trend. This project will determine the distribution of the second best bids, or win prices, of the auctions that Cidewalk bids on. It is unlikely that the second best bids will fit exactly into any known distributions, and therefore an approximation of the distribution is an accurate way to model the second best bids. This distribution can then be used to replace the assumed exponential distribution used in [1], which will provide more accurate calculations of the optimal bid for each auction.

2.3.2 Addressing Assumptions Regarding Bid Weight

Cidewalk has implemented a strategy that prevents them from bidding against themselves in auctions, when they have multiple orders from different customers to fill in the same geographical location. The fact that there are multiple, rather than a singular opponent within the exchanges, needs to be addressed as trends in the second best bid could reflect this. There is no way of telling who wins the auction or for how much unless Cidewalk wins, since this information is not disclosed if the auction is lost. Therefore, it is impossible to know how many opponents there are in the market and what strategies they use to bid on the auctions within the various exchanges. This project relies on the distribution of the second best bids to determine how much Cidewalk should bid on each auction to minimize the cost to the company. If the opponents are also using a strategy that relies on the distribution created from the sample data they are able to collect on the second best bids it could hinder Cidewalk's ability to minimize cost. Hu and Macaluso do not address this issue as they assume that the bids of the opposing companies do not rely on each other or our own bids. Therefore, to produce an effective model capable of real-world utilization this case would need to be addressed and will be taken into consideration in this project.

3 Methods

This project is focused on translating a theoretically optimal bidding model into a realistically optimal bidding model which involves meeting the following objectives:

1. Gathering various sample data sets with detailed information about past auctions from Cidewalk
2. Analyzing the actual distribution of second best bids, or win prices, and understanding what impacts these distributions
3. Utilizing updating Bayesian methods to ensure that the model can adapt to changes in the market
4. Simulating the mobile advertisement market to test how the model performs

This section outlines how to take the necessary intermediate steps to meet the objectives stated above which allow the production of results that justify testing the model live.

3.1 Building a model

Understanding Cidewalk as a company is important to accurately build the model for optimal bidding on exchanges to win the required number of auctions within a budget, while taking into consideration the business's goals. Therefore, the optimization problem will contain certain constraints, which will reflect Cidewalk's needs. Each business owner seeks to maximize profit and minimize cost, and that is why when modeling the bidding strategy, budget constraints will need to be taken into account. Cidewalk's goal is to deliver advertisement spaces of the highest quality to assist in expanding the customer base of their clients. To address these concerns, Cidewalk is willing to pay extra for advertisement spaces that generate additional interest from the public and in turn increase the income of their clients. The advertisement space is worth more to Cidewalk if there is a higher chance that the mobile application user will click on the advertisement or even purchase the product. Another constraint is related to the mobile advertisement market in which Cidewalk does business. Some exchanges and mobile applications set a floor, which impacts the optimal bid since only bids higher than the floor price can win the auction.

Building a model with the aforementioned constraints immediately would be complex. Thus, an attempt to model the optimal bid that minimizes the cost without taking into account any other constraints is developed first. As a result it will be easier to incorporate any other necessary constraints at a later time.

We can observe from Equation 3 that the calculation of the optimal bid involves knowing the distribution of the second best bid. The second best bid is only known if Cidewalk won the auction and is equal to the price that Cidewalk pays for the advertisement space[1].

3.1.1 Data Gathering

The optimal bidding formula created by Hu and Macaluso requires an estimation of the second best bid, or winning price, represented by $P(x)$. To find the best approximation for the distribution, we must first collect second best bids. Cidewalk maintains a database that records all the information on auctions that they bid on including the win price, bid price, exchange, mobile application identifier, geographic location, and any other information collected by the mobile application on the user. Access to this database enables the collection of win price values for various lengths of time and for a varying numbers of auctions.

The first sample set obtained from Cidewalk contains 13,351,897 auctions that occurred at the end of August, 2015. The set is then grouped according to the the hour upon which the auctions occurred, and each hour contains a count of the number of auctions won based on ten cent CPM intervals with the last interval counting all values above one dollar CPM. The sample data shows what hours during the day most auctions are won and what hours the highest number of auctions are won for the lowest cost.

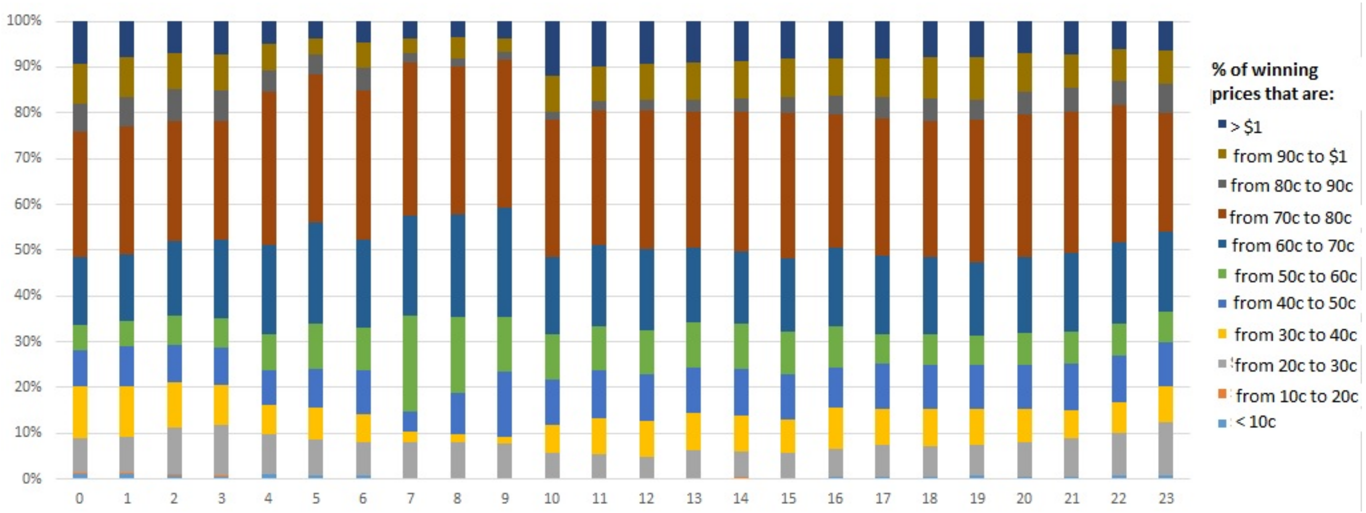


Figure 1: Percentage of auctions won for ten cent CPM intervals grouped by hour.

The data from this sample set also provides the win rate for the time period in which the sample data represents, which was approximately twenty to thirty percent of all auctions Cidewalk bid on across all exchanges. The information that results from this sample data set shows that the time periods between 1am and 4am EST has a higher winning rate[1]. However, this is a result of the overnight testing that Cidewalk performs on its systems.

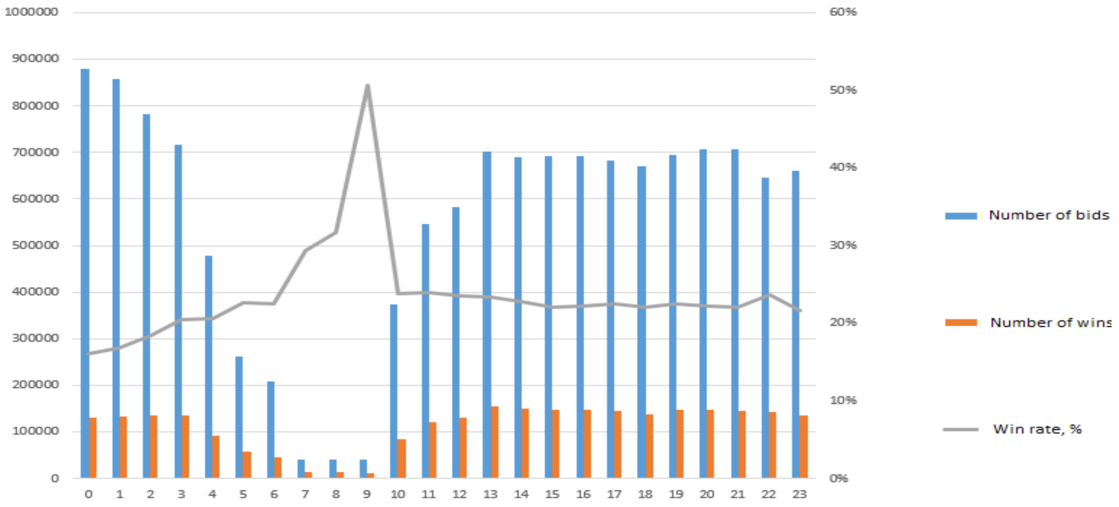


Figure 2: Number of auctions grouped by hour with win rate trend displayed.

[1] Hours in the graph are listed in the GMT format. To convert to EST - subtract 5 hours.

This collection of sample data was useful on a larger scale, but as the project requires a precise distribution representing the win prices grouping the win prices in a ten cent CPM interval, is not accurate enough for analysis. As the model must be effective despite the time of day, it is important to perform further analysis to provide more detail on the win prices and a better picture of the sample data with no hourly restrictions.

The second sample set of data represents auctions that occurred in the first half of September, 2015 and contains 9,326,675 auctions. The collected data is not grouped according to the time of day, but instead focuses on only auctions won by Cidewalk. No other restrictions are put on the collection which causes all of the information on each auction to be present including the time at which the auction occurred, the win price, the bid price, the exchange, the geographic location, the mobile application identifier, and any information the mobile application required from the user upon registration. The following scatter plot of this sample set shows the bid prices versus the win prices by assigning the bid price as the x-coordinates and the win price as the y-coordinates. The graph displays how the bid price differs from the actual win price and what the range for the bid and win prices are for the sample data it represents. The insight this graph provides on the range of win prices is useful, but it causes the realization that restricting the sample data collection to only auctions won by Cidewalk is disregarding a large portion of important information, since this sample data is based on strategies currently in use by the company. Despite whether or not an auction is won the bid price is still recorded and if we include these bid prices it could help provide us with a larger picture of the mobile advertisement market as a whole.

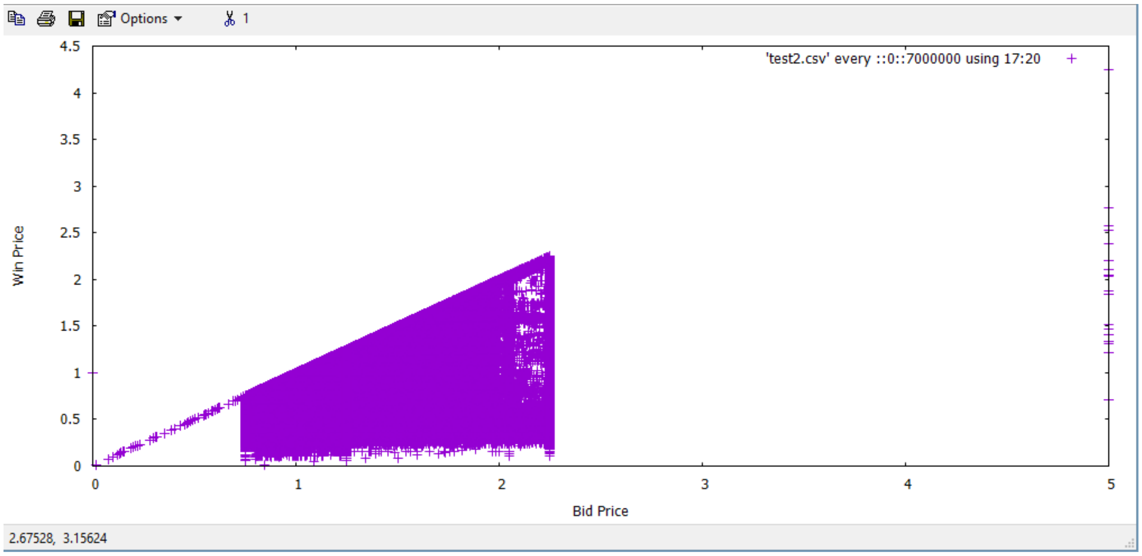


Figure 3: Bid versus win price for auctions held from the second sample data set.

The third sample set of data represents auctions that occurred at the end of September, 2015. The second sample set of data was neglecting a vast majority of data by focusing only on winning auctions therefore this set includes all available information on the auctions Cidewalk won and lost. The sample data provides information regarding the overall trend of the winning prices through the use of a histogram depicting the win prices and the corresponding frequencies of the win prices with an accuracy of $0.01 CPM.

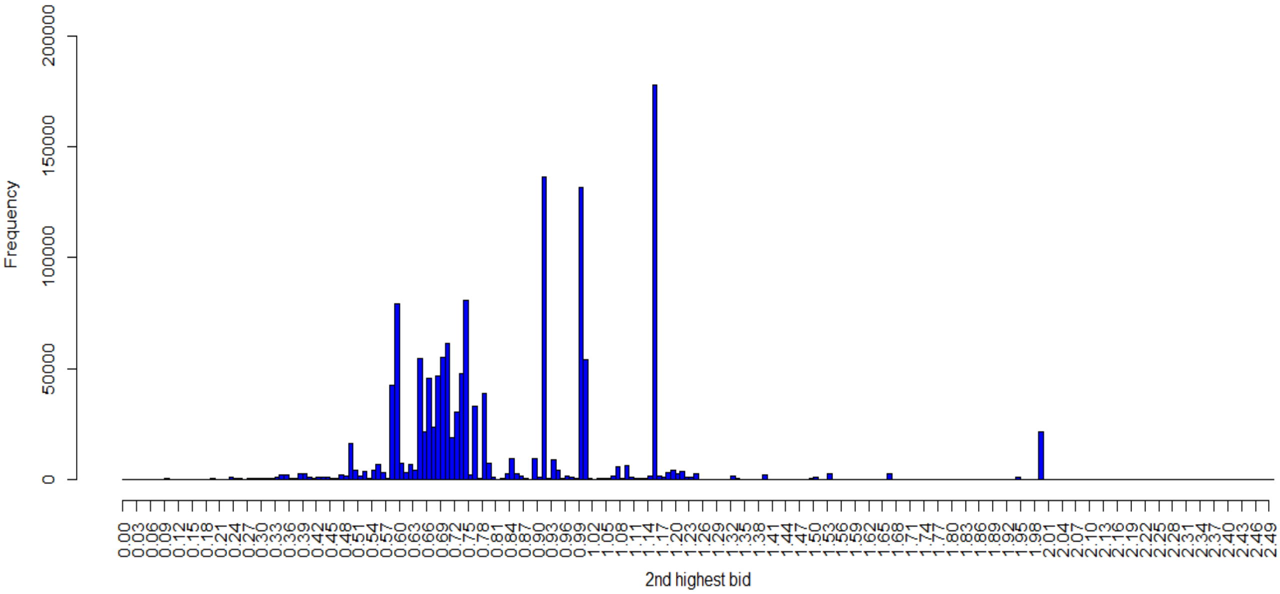


Figure 4: Distribution of win prices for all auctions from the third sample data set.

We can observe a trend in the histogram which prompts further examination into how the exchanges Cidewalk uses impact the overall trend. During this time period Cidewalk primarily bid on auctions through the Adexchange 1, where Adexchange 1 accounts for 8,222,486 auctions out of 11,005,702. Adexchange 2 and Adexchange 3 account for 1,264,367 auctions and 1,507,097 auctions respectively, out of the total 11,005,702 auctions. Creating histograms depicting the win prices and the corresponding frequencies of the win prices with an accuracy of $0.01 CPM shows how these exchanges influence the overall distribution and motivates the focus of the analysis to shift to Adexchange 1 as it was the most frequently used exchange.

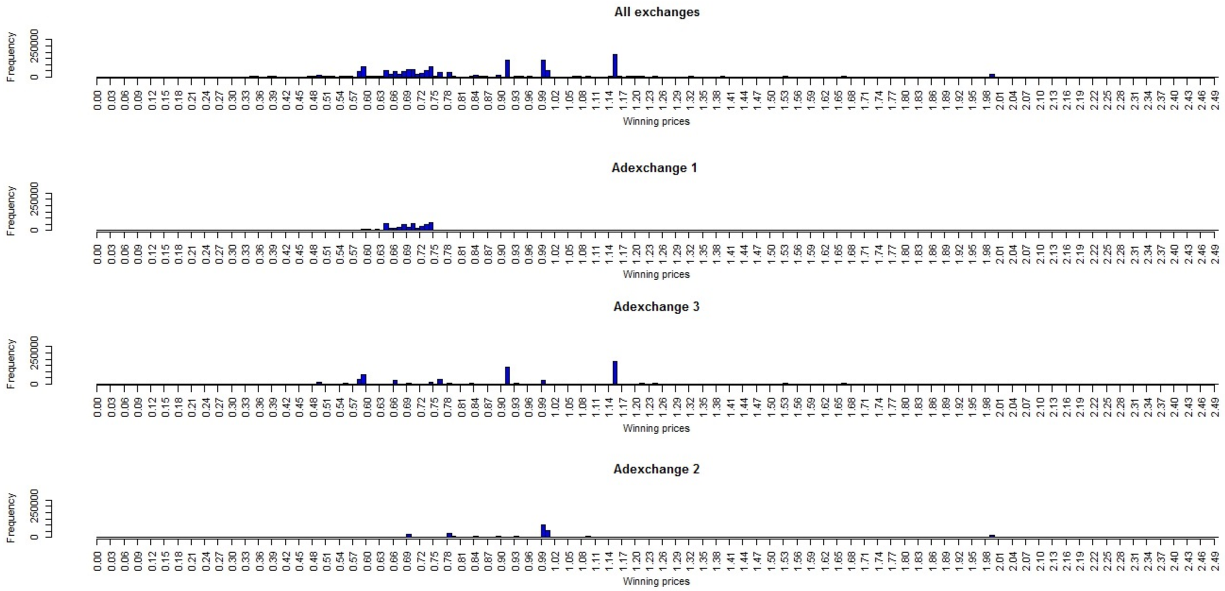


Figure 5: Comparison of each exchange's second best bid distribution to the distribution for all exchanges.

The fourth sample set of data represents auctions that occurred in October 2015. This sample data set is then split in two for analysis purposes to verify that the the distribution of the second best bids is the same for two subsets, just with different parameters. After discovering that Cidewalk no longer uses the Adexchange 1 it was evident that new data set that better represents the exchanges Cidewalk currently bids on had to be collected. This set of sample data focuses on the exchanges that Cidewalk currently bids on the most which are the Adexchange 2 and Adexchange 3 respectively, where Adexchange 2 accounts for 6,640,384 auctions during this time period and Adexchange 3 accounts for 2,176,627 auctions out of the 8,839,678 total auctions. Histograms depicting the win prices and corresponding frequencies of the win prices with an accuracy of $0.01 CPM for both of these main exchanges, help us to understand how these two main exchanges impact the overall trend.

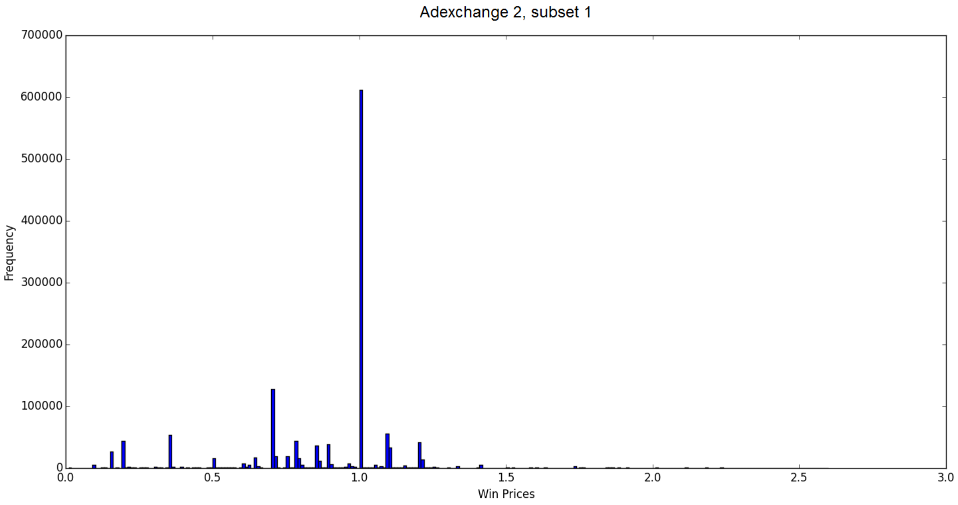


Figure 6: Distribution of the winning prices (second best bids) from auctions held in October, 2015, (first subset) on the Adexchange 2.

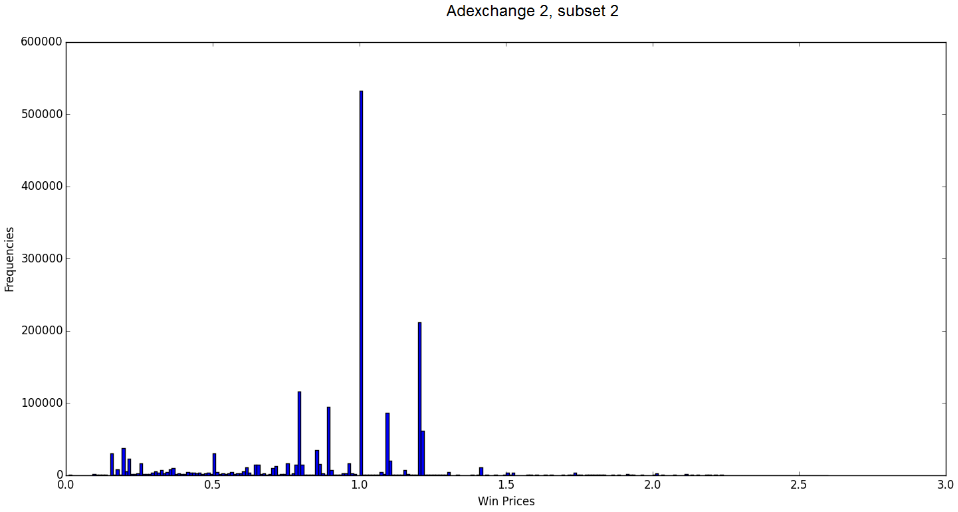


Figure 7: Distribution of the winning prices (second best bids) from auctions held in October, 2015, (second subset) on the Adexchange 2.

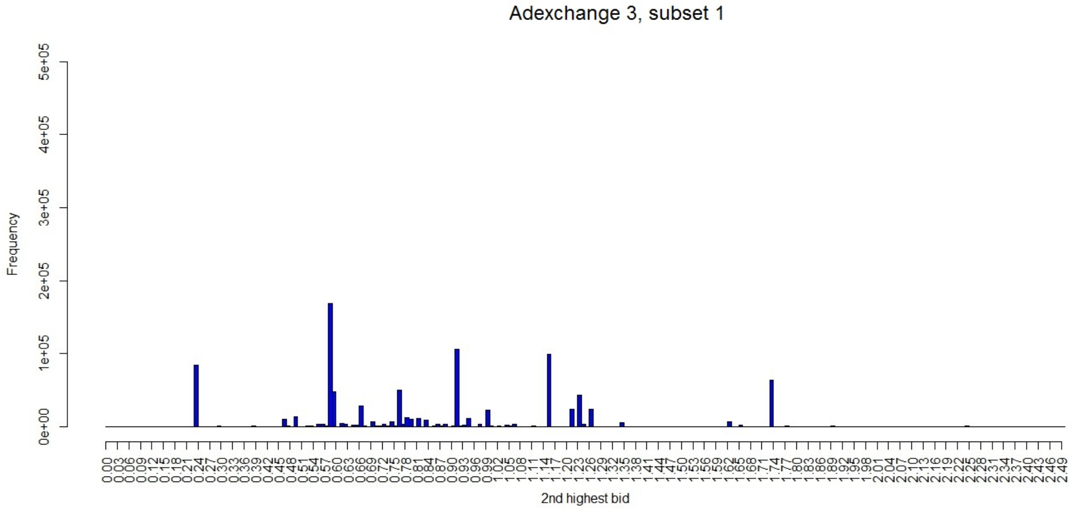


Figure 8: Distribution of the winning prices (second best bids) from auctions held in October, 2015, (first subset) on the Adexchange 3.

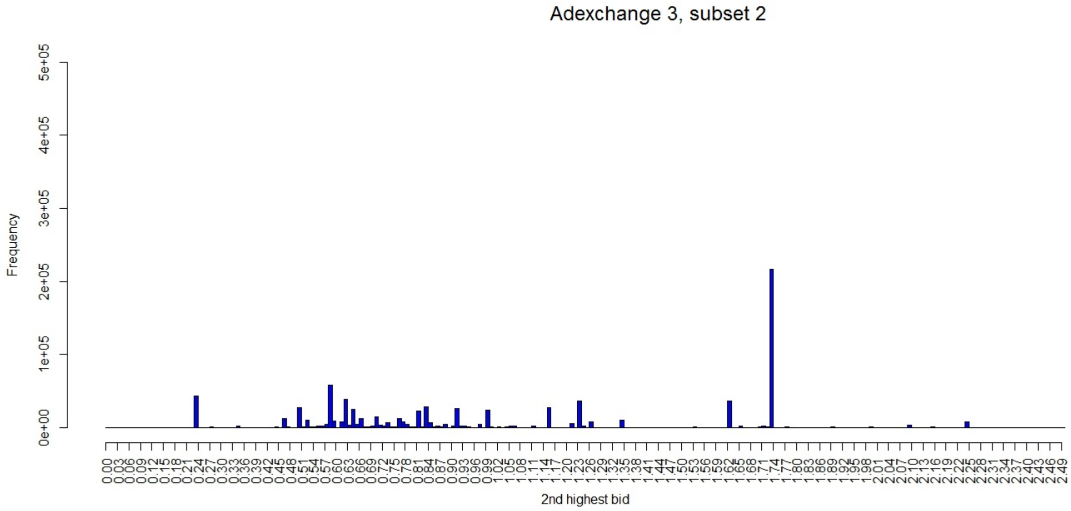


Figure 9: Distribution of the winning prices (second best bids) from auctions held in October, 2015 (second subset) on the Adexchange 3.

3.1.2 Data Analysis Methods

Analyzing the sample data to discover the best approximation for the distribution of second best bids or winning prices requires the use of various methods, since simply looking at a graph of the sample data can cause one to believe it fits a certain distribution when in reality it does not. One method we can use to analyze these distributions is QQPlots.

QQ Plot

A QQ Plot, or Quantile-Quantile Plot, is a method that uses a visual aid in the form of a graph to compare two sets of data to determine how likely it is that the two are from the same distribution. According to NIST/SEMATECH, "A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30 percent) quantile is the point at which 30 percent of the data fall below and 70 percent fall above that value. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line."[3] This method of analysis is useful in estimating the distribution which represents the best fit for the sets of sample data we collect from Cidewalk.

3.1.3 Choice of the distribution

The second best bid is the win price that Cidewalk pays when an auction is won. However, due to the nature of Vickrey auctions it is not possible to get the winning prices for the auctions that Cidewalk lost. This will result in missing sample data when modeling the distribution. This issue will be addressed in detail in Section 3.2.

The histograms representing the win price on all exchanges from the third sample data set show that reducing the width of the bins makes it harder to detect a proper distribution that best fits the data.

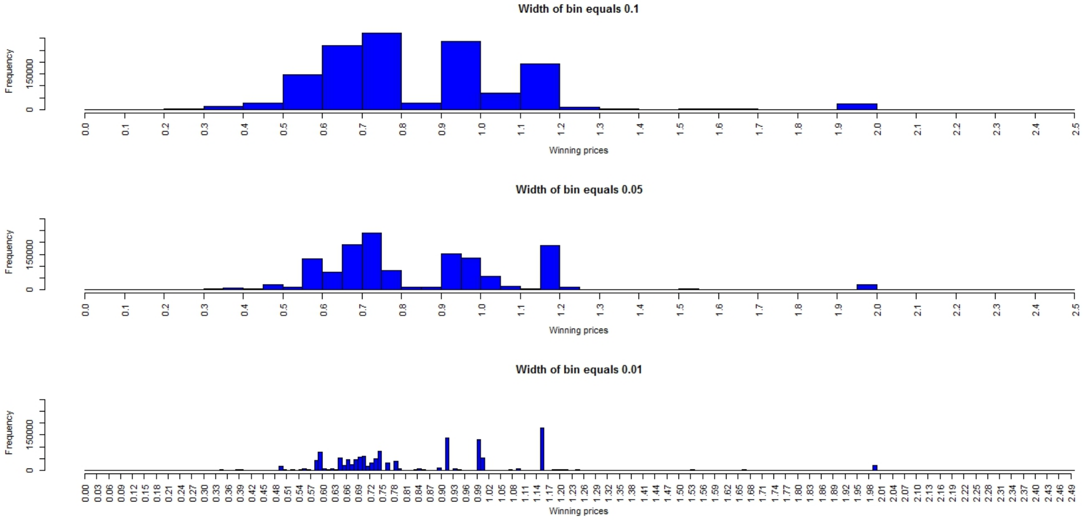


Figure 10: Distribution of the winning prices (second best bids) for auctions from third data set for different bin widths.

The existence of several peaks on the graphs implies that there is no natural parametric distribution that perfectly fits the set of sample data. It is possible that the sample data set is a combination of distributions or the sample data set needs to be decomposed into different categories. By dividing the whole set of sample data into separate groups, where each group represents an exchange Cidewalk bids in, it can be seen that Adexchange 1 is the major exchange that they bid on during this time period. We can also observe that Adexchange 1 has the lowest winning rate within the date range based on the number of auctions Cidewalk bid on. Adexchange 4 is the exchange that they bid on the least in comparison to the other exchanges. This exchange was excluded from analysis after the discovery that Cidewalk uses this exchange for testing purposes only.

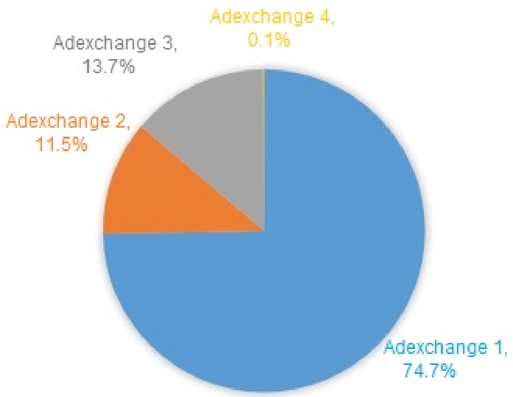


Figure 11: Exchanges Cidewalk bid on based on the total number of auctions by percentage.

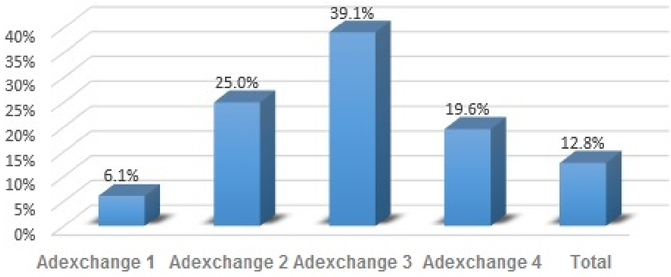


Figure 12: Winning rate for each exchange.

The collection of histograms in Figure 5 confirms that each exchange contributes to the peaks seen on the histograms in Figure 10. Therefore, a separate optimal bidding strategy will be modeled for each exchange.

As Adexchange 1 is the most bid on exchange within the data set and is the most promising when it comes to finding a distribution that fits the sample data most accurately, so analysis will begin with this exchange.

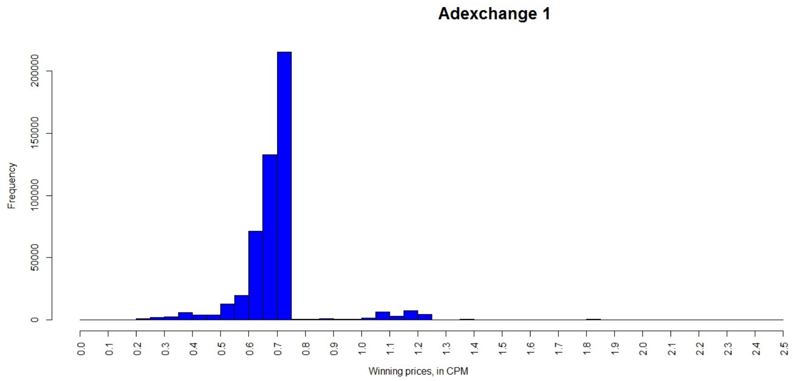


Figure 13: Second best bid distribution on Adexchange 1.

It can be seen from the graph that the second best bid follows a negatively skewed distribution which means the sample data does not fit any common distributions. One method for addressing this issue is to flip the distribution to create a positively skewed distribution which can be fitted to more commonly used distributions. We can observe from Figure 13 that the highest peak occurs at approximately $0.75, therefore we can flip the distribution around this value to accurately fit the values to a positively skewed distribution. The distribution that best fits our now positively skewed distribution is a gamma distribution, based on the log-likelihood value and QQplots in Figure 14.

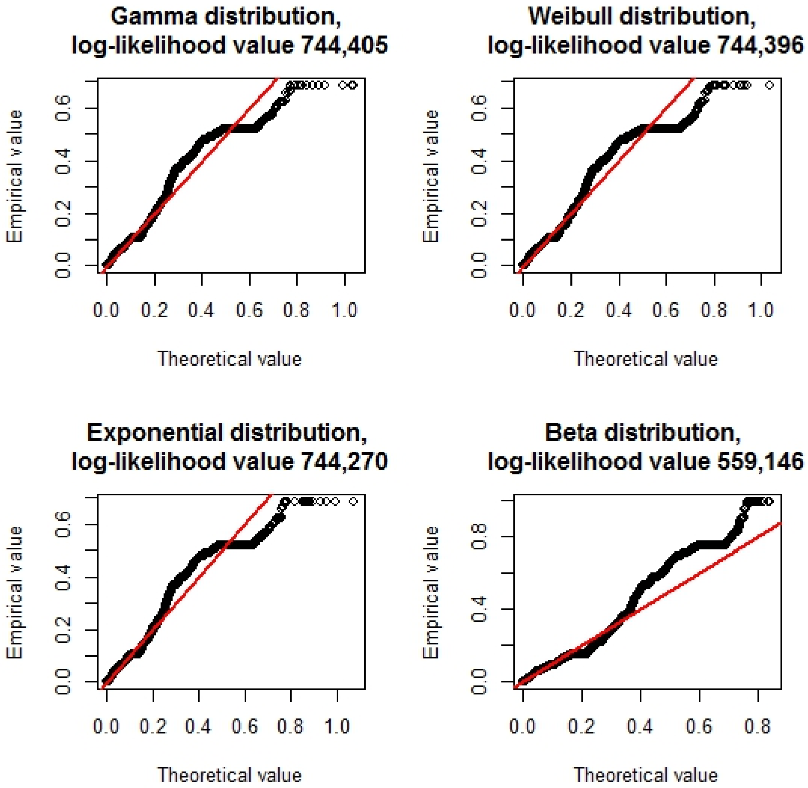


Figure 14: QQplots for distributions that the Adexchange 1 data set fits most accurately. Log-likelihood values provided for each distribution.

Adexchange 1 is no longer used by Cidewalk due to the low winning rate and lower quality advertisement spaces on the exchange which is not in line with the company's strategy to increase the conversion rate (increase the number of responses to the customer's advertisements and convert site visitors into paying customers). Sample data that represents the exchanges Cidewalk currently bids on is necessary to move forward towards the implementation of the optimal bidding strategy.

The fourth sample data set that is split into two subsets shows that the exchanges Cidewalk currently bids on the most are the Adexchange 2 and Adexchange 3, respectively. Figure 15 and Figure 16 display the overall trend of the two weeks of sample data.

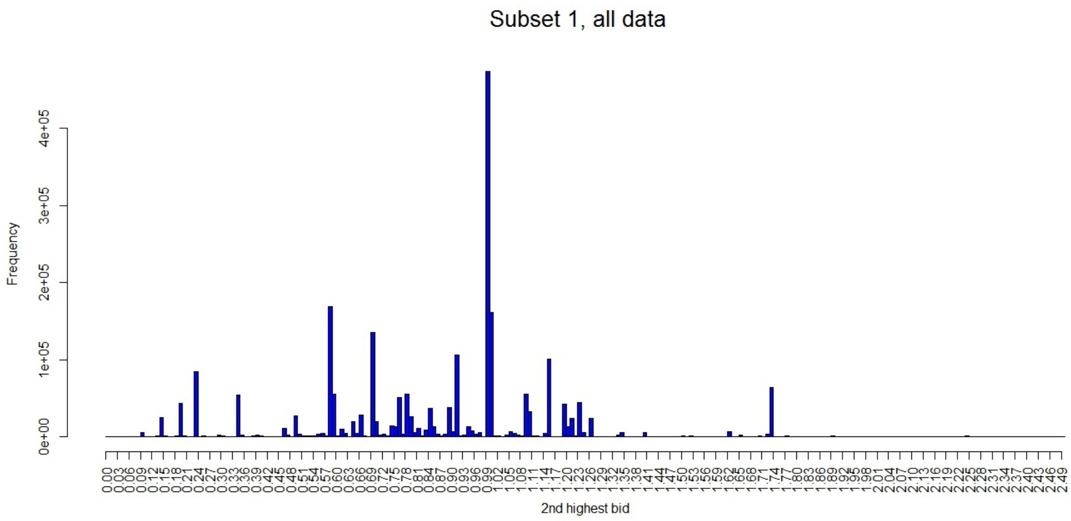


Figure 15: Distribution of the winning prices (second best bids) from auctions from subset 1.

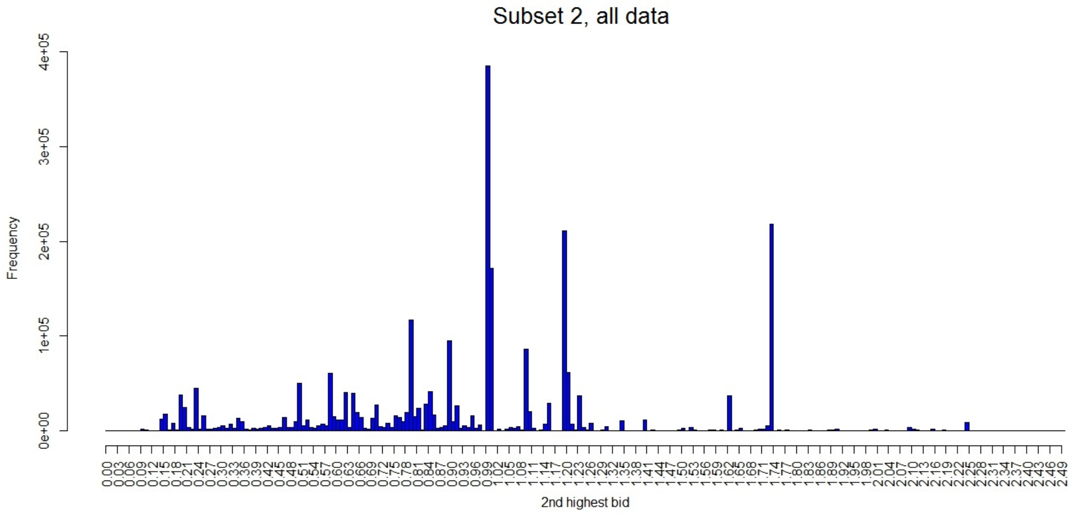


Figure 16: Distribution of the winning prices (second best bids) from auctions from subset 2.

To verify that the best fitting distribution stays consistent for a given exchange we begin by finding the distribution that best fits the sample data of subset 1. We then fit this distribution to the data in subset 2 to determine whether it most accurately represents subset 2's sample data in comparison to other distributions.

From the graphs below it can be seen that a normal distribution is a reasonable approximation of the second best bid on Adexchange 2 based on the log-likelihood value, where a higher value implies a better fit. Data in the tails in the graphs below is caused by special promotions Cidewalk runs and overnight testing that is performed on Cidewalk's systems. Therefore, such sample data will be neglected, since it represents special circumstances.

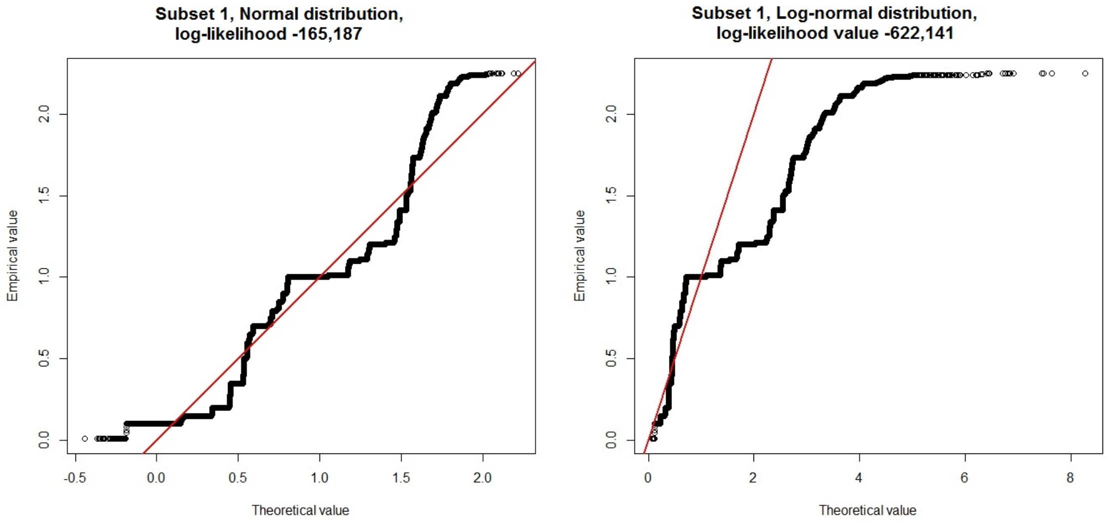


Figure 17: QQplots for the auctions from the subset 1 on Adexchange 2.

Based on the QQplots it is apparent that the distribution of the second best bid on Adexchange 2 remains the same for subset 2:

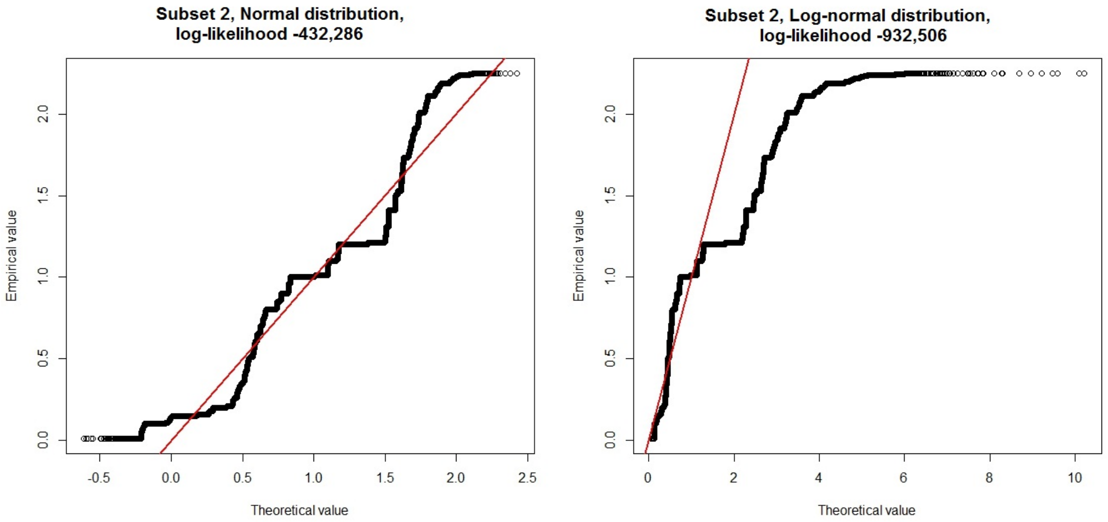


Figure 18: QQplots for the auctions from the subset 2 on Adexchange 2.

The distribution of the second best bid that fits Adexchange 3 differs from the distribution that fits Adexchange 2 best. Figure 8 and Figure 9 show the distribution of the second best bid for two subsets from October, 2015.

In order to find the distribution that describes the behavior of the second best bid on Adexchange 3 best, it is necessary to examine several distributions. From Figure 19 and Figure 20 we can observe that out of the four distributions the distribution that best describes the behavior of the second best bid is a beta distribution for both subsets.

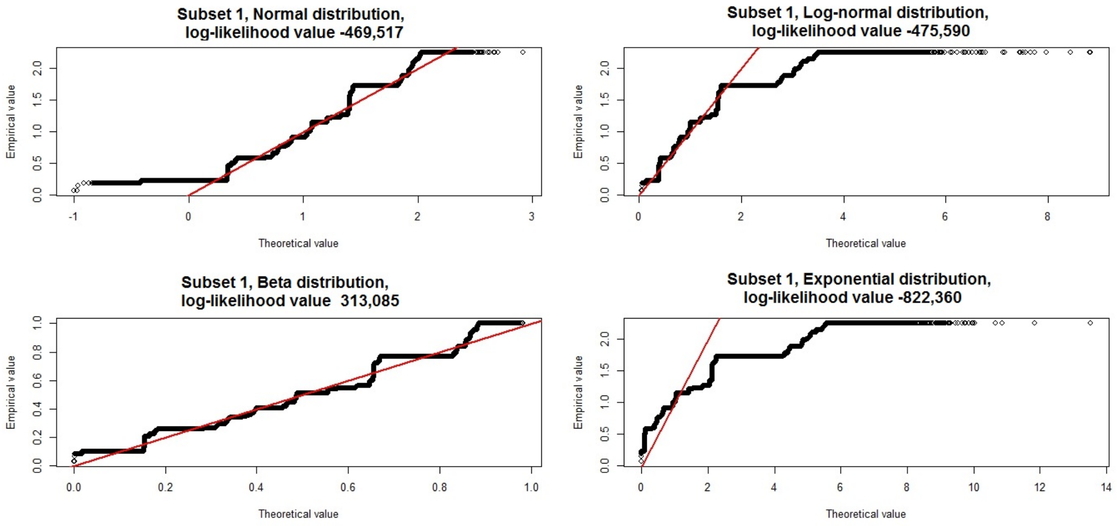


Figure 19: QQplots for the auctions on Adexchange 3 from subset 1.

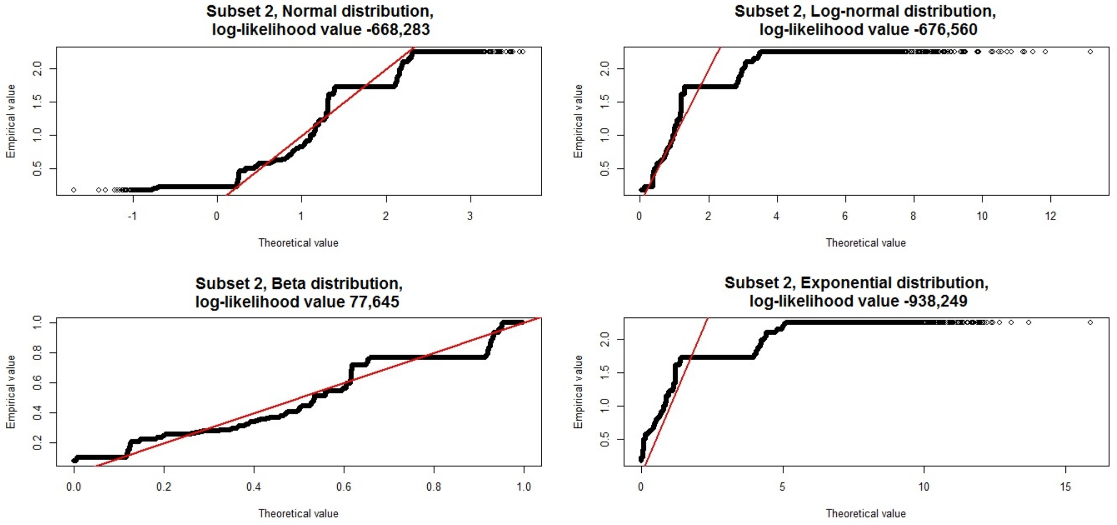


Figure 20: QQplots for the auctions on Adexchange 3 from subset 2.

As a result, for any further analysis or when building the model, we will use a normal distribution for Adexchange 2 and a beta distribution for Adexchange 3.

3.2 Missing values in the data

Due to the nature of Vickrey auctions it is impossible to get the whole picture when trying to determine the distribution of the best bid of the competitors, since the best bid of the competitors is only observed when Cidewalk wins the auction. Therefore, only part of the whole picture is provided and it is necessary to address the missing values in the sample data set. There were two obvious ways of dealing with these missing values, either not take them into account at all or find a way to estimate the values to fill in the gaps. Estimation of the missing values can be accomplished through an educated guess, using average of the data, performing regression analysis, or other methods. The following describes the method that addresses the missing values for this model's purposes.

It is assumed that the distribution of the best bid of the competitors is modeled by a random variable X with cumulative distribution function (CDF) F_X and probability density function f_X. However, besides the random variable X, there are two additional random variables present. The random variable Y, which represents the best bid of the others given that the value was higher than the corresponding bid b, which represents the case where Cidewalk won the auction. The other random variable Z represents the best bid of the competitors given that Cidewalk lost, which means that the best bid of the others was equal to or higher than b. This random variable Z is the aforementioned missing value in the set of sample data.

In mathematical terms, the distribution of random variable Z is represented as follows:

$$F_Z(z) = \mathbb{P}[Z \leq z] = \mathbb{P}[X \leq z \mid X \geq b] = \frac{\mathbb{P}[X \leq z, X \geq b]}{\mathbb{P}[X \geq b]},$$

and the density is

$$f_Z(z) = \frac{\partial F_Z(z)}{\partial z}.$$

To obtain this random variable Z one needs to sample from the distribution that X follows and discard those values that are smaller than b. This is easily done by utilizing the acceptance-rejection method. However, this technique requires knowing the distribution of X.

3.3 Estimation of the distribution of the best bid of others

The distribution of X is of the most interest as it cannot be observed from the market. It turns out that the distribution may be estimated using the information that we observe from the distribution of Y.

Assume that X and Y follow the same distribution just with different parameters. Consider the distribution of Y as a conditional distribution of X given that the auction was won at a bid b, which means that $Y \sim X|X \leq b$. Hence, the CDF of Y is as follows:

$$F_Y(y) = \mathbb{P}[Y \leq y] = \mathbb{P}[X \leq y|X \leq b] = \frac{\mathbb{P}[X \leq y, X \leq b]}{\mathbb{P}[X \leq b]}$$

with the corresponding density

$$f_Y(y) = \frac{\partial F_Y(y)}{\partial y}.$$

Here the distribution of Y is known, hence it is possible to estimate the distribution of X using the distribution of Y.

3.3.1 Case of Adexchange 2

The distribution of Y is observed when Cidewalk wins and is considered to be normal on the Adexchange 2 (see Section 3.1.3). One can obtain the explicit formulas to find the distribution of X given that $Y \sim X|X \leq b$ and using the formulas presented in the previous subsection.

Consider the CDF of Y:

$$F_Y(y) = \frac{\mathbb{P}[X \leq y, X \leq b]}{\mathbb{P}[X \leq b]} = \begin{cases} \frac{F_X(y)}{F_X(b)}, & y \leq b \\ 1, & y > b \end{cases} = \begin{cases} \frac{\Phi\left(\frac{y-\mu_X}{\sigma_X}\right)}{\Phi\left(\frac{b-\mu_X}{\sigma_X}\right)}, & y \leq b \\ 1, & y > b \end{cases}$$

with a density

$$f_Y(y) = \frac{\partial F_Y(y)}{\partial y} = \frac{f_X(y)}{F_X(b)} \mathbb{1}_{\{y \leq b\}} = \frac{1}{\sqrt{2\pi}\sigma_X F_X(-b)} e^{-\frac{(y-\mu_X)^2}{2\sigma_X^2}} \mathbb{1}_{\{y \leq b\}}$$

where $\Phi(x)$ and $F_X(x)$ are the CDFs of a standard normal and normal distribution respectively, and $f_X(x) = \frac{1}{\sqrt{2\pi}\sigma_X} e^{\frac{(x-\mu_X)^2}{2\sigma_X^2}}$ is a probability density function of a normal distribution.

Moreover, the mean and variance of Y is also known:

$$\begin{aligned} \mu_Y = \mathbb{E}[Y] = \frac{1}{F_X(b)} \int_{-\infty}^{b} y f_X(y) dy &= \frac{1}{\sqrt{2\pi}\sigma_X \Phi(\frac{b-\mu_X}{\sigma_X})} \int_{-\infty}^{b} y e^{-\frac{(y-\mu_X)^2}{2\sigma_X^2}} dy \\ &= \mu_X - \frac{\sigma_X}{\sqrt{2\pi}\Phi(\frac{b-\mu_X}{\sigma_X})} e^{-\frac{(b-\mu_X)^2}{2\sigma_X^2}}, \end{aligned} \tag{4}$$

$$\mathbb{E}[Y^2] = \frac{1}{F_X(b)} \int_{-\infty}^{b} y^2 f_X(y) dy = \frac{1}{\sqrt{2\pi}\sigma_X \Phi(\frac{b-\mu_X}{\sigma_X})} \int_{-\infty}^{b} y^2 e^{-\frac{(y-\mu_X)^2}{2\sigma_X^2}} dy$$

$$= \mu_X^2 + \sigma_X^2 - \frac{\sigma_X(b-\mu_X) + 2\mu_X\sigma_X}{\sqrt{2\pi}\Phi(\frac{b-\mu_X}{\sigma_X})} e^{-\frac{(b-\mu_X)^2}{2\sigma_X^2}}$$

$$\sigma_Y^2 = \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 = \sigma_X^2 - \frac{\sigma_X(b-\mu_X)}{\sqrt{2\pi}\Phi(\frac{b-\mu_X}{\sigma_X})} e^{-\frac{(b-\mu_X)^2}{2\sigma_X^2}} - \frac{\sigma_X^2}{2\pi\Phi^2(\frac{b-\mu_X}{\sigma_X})} e^{-\frac{(b-\mu_X)^2}{\sigma_X^2}} \tag{5}$$

Equations (4) and (5) form a nonlinear 2x2 system of equations with two unknowns. Since mean μ_Y and variance σ_Y^2 are known, μ_X and σ_X^2 can be recovered numerically from the system of equations.

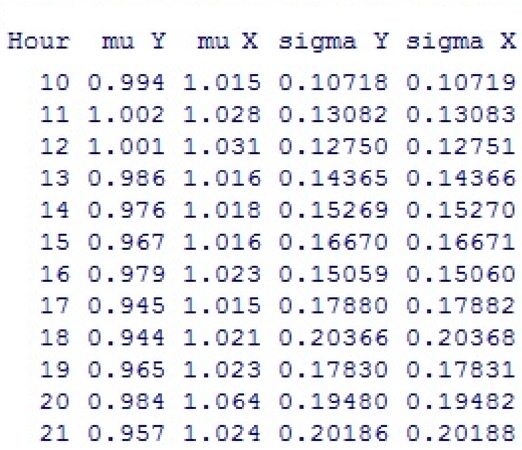
Now to shift to the multi-dimensional setting. Consider another random variable Y_i that represents the outcome of the i^{th} auction. Then by the Central Limit Theorem the random variable $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ describes the distribution of the best bid of the others that we observe, where $Y_i \sim X|X \leq b_i$. Hence, updated formulas (4) and (5) are the following

$$\mu_{\bar{Y}} = \mathbb{E}[\bar{Y}] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[Y_i] = \mu_X - \frac{\sigma_X}{\sqrt{2\pi}} \sum_{i=1}^{n} \frac{1}{\Phi(\frac{b_i-\mu_X}{\sigma_X})} e^{-\frac{(b_i-\mu_X)^2}{2\sigma_X^2}} \tag{6}$$

$$\begin{aligned} \sigma_{\bar{Y}}^2 = \mathbb{V}ar[\bar{Y}] = \frac{1}{n^2}\mathbb{V}ar[Y_i] = \sigma_X^2 - \frac{1}{n^2}\frac{\sigma_X}{\sqrt{2\pi}} \sum_{i=1}^{n} \frac{(b_i-\mu_X)}{\Phi(\frac{b_i-\mu_X}{\sigma_X})} e^{-\frac{(b_i-\mu_X)^2}{2\sigma_X^2}} \\ - \frac{1}{n^2}\frac{\sigma_X^2}{2\pi} \sum_{i=1}^{n} \frac{1}{\Phi^2(\frac{b_i-\mu_X}{\sigma_X})} e^{-\frac{(b_i-\mu_X)^2}{\sigma_X^2}}. \end{aligned} \tag{7}$$

There exists numerical methods that solve such systems of nonlinear equations in R and Python. The table below contains values for μ_Y, μ_X, σ_Y, and σ_X for several hours within one day. As we expect, μ_X is always higher than μ_Y, since only part of the distribution is observed when Cidewalk wins[2].

[2] While calculating μ_X and σ_X, it was discovered that there was a promotional campaign in Texas, which drove the mean higher. Therefore, auctions from Texas were not taken into account during the calculations.



```
Hour  mu Y  mu X sigma Y sigma X
  10 0.994 1.015 0.10718 0.10719
  11 1.002 1.028 0.13082 0.13083
  12 1.001 1.031 0.12750 0.12751
  13 0.986 1.016 0.14365 0.14366
  14 0.976 1.018 0.15269 0.15270
  15 0.967 1.016 0.16670 0.16671
  16 0.979 1.023 0.15059 0.15060
  17 0.945 1.015 0.17880 0.17882
  18 0.944 1.021 0.20366 0.20368
  19 0.965 1.023 0.17830 0.17831
  20 0.984 1.064 0.19480 0.19482
  21 0.957 1.024 0.20186 0.20188
```

Figure 21: Mean and standard deviation values calculated for random variables X and Y for Adexchange 2 for a sample set of data considered in this study.

3.4 Updating Methods

For any given data set it is easy to calculate the mean and standard deviation, but using a mean and standard deviation from a previous data set may not accurately represent a new data set even if the data is coming from the same source. This is especially true in a market as volatile as the mobile advertisement marketplace, and this is why it was important to create a strategy that will be able to adapt to the ever-changing market through periodic updating. Through the use of a common statistical learning method, Bayesian updating, the bidding strategy covered in Section 3.1 can utilize a mean and standard deviation that represents not only the previous data set but all data sets that came before it.

3.4.1 Bayesian Updating

Bayes Theorem was discovered by Thomas Bayes in 1763 and has been used by statisticians and in probability for hundreds of years. Bayesian updating uses Bayes Theorem to predict a future event based on past events [4]. Therefore, we can use Bayesian updating to predict the next best bid of others based on the previous best bids of others, but as the auctions in the mobile advertisement market occur rapidly there is no way a singular best bid of the others can be calculated before the end of a given auction. Given the bidding strategy detailed in Section 3.1, there is another value that can be updated less frequently that will still provide a picture of the best bids of others based on the previous, or prior best bids of others, and more generally, the parameters of the distribution of the best bid of the others. Equation 3, utilizes the normal cumulative distribution function with a mean and standard deviation from the distribution of the best bids of others therefore, updating the mean and standard deviation of this distribution periodically will also change the value of this model's bid accordingly. Due to the high volume and velocity of the auctions the mean and standard deviation will be updated on an hourly basis. After each hour passes the calculated mean and standard deviation of the real distribution, which is found by solving the system of non-linear equations listed in Section 3.3, is then used to predict, or produce the posterior mean and standard deviation of the real distribution for the following hour using the following equations:

Posterior Mean:

$$\frac{\frac{\mu_{prior}}{\sigma^2_{prior}} + \frac{\mu_{new}}{\sigma^2_{new}}}{\beta} \tag{8}$$

Posterior Variance:

$$\frac{1}{\beta} \tag{9}$$

Posterior Precision

$$\frac{1}{\sigma^2_{prior}} + \frac{1}{\sigma^2_{new}} \tag{10}$$

where μ and σ^2 represent the mean and variance of the corresponding distribution, and β represents the posterior precision of the distribution of the best bids of others[5].

The following figure depicts Bayesian updating on the mean and standard deviation of the real distribution of the best bid of others for Adexchange 2 for a sample set of data considered in this study. At hour 0, we learn from the sample data that is available by assuming that the real distribution of the best bids of others is the same as the distribution of the winning prices during this hour. However, at hour 10 new sample data about the winning prices becomes available and since it is different from the sample data observed during hour 0, we update our perception of the real distribution of the best bid of others by incorporating this new information into the distribution, which results in the shift to the left. This process remains the same for consecutive hours. This procedure enables updating of the distribution parameters of the best bid of others based on the arriving information, and it also helps to address the changing environment on the exchanges.

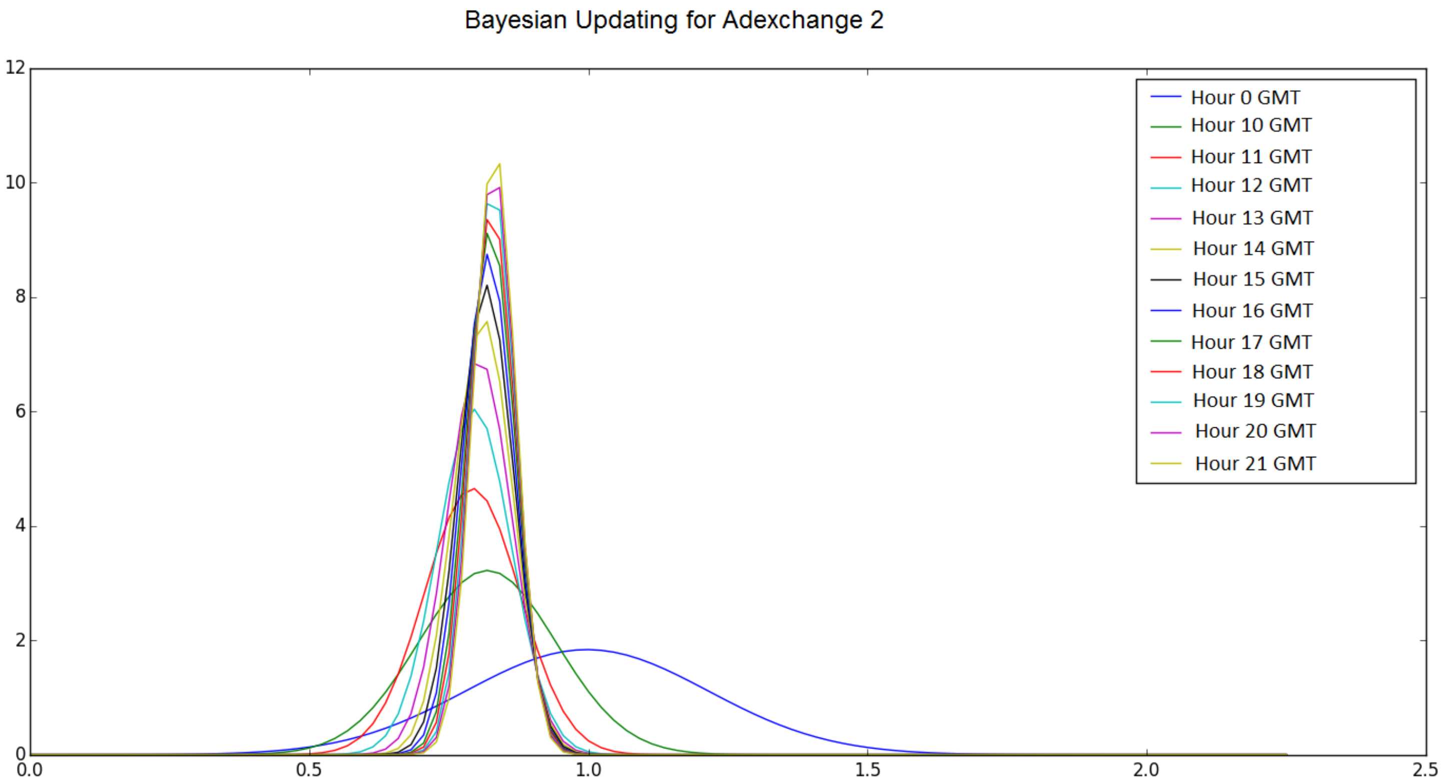


Figure 22: Bayesian Updating For Real Distribution Of Best Bid Of Others For Adexchange 2

3.5 Simulation

Before implementing any new method it is important to understand how the method would perform in a real-life setting, which requires the method to be tested on sample data. Simulating the environment in which the method will be used can provide valuable insight into how the method will perform and can allow for analysis of the results while preventing any potential repercussions of using the method immediately. Although the results of the simulation may not be exactly the same as the results obtained from utilizing the method in a real-world setting, simulation can allow for easy detection of any errors in the method and provide the necessary time to address any errors or issues.

3.5.1 Final Simulation

Simulation prior to implementation is key for this project as this method is used to bid on auctions and therefore, could result in Cidewalk losing money rapidly if there are any errors or issues that exist in the method, as the auctions occur within seconds. In order to simulate the mobile advertising market, a previous set of sample data is used for Adexchange 2 that contains all of the relevant information for all of the auctions that occurred in the time period that the sample data represents. The simulation is ran for 13 hours during a single day, and the hours selected are chosen based on the bidding activity level of Cidewalk, as Cidewalk does not bid on auctions that occur during the late evening hours. This procedure is done to simulate how the day's auctions would really happen. The first hour, or hour 0, is used for the model to gain initial estimates of the mean and standard deviation of the real distribution of second best bids. After the model has access to an initial mean and standard deviation of the second best bids the model can begin to simulate bids and for every auction that Cidewalk bid on during hours 10 to 21 inclusive, a bid will be produced using the model based on the number of auctions Cidewalk won and the number of auctions Cidewalk bid on. If there was a win price available for the auction and the bid price the model produces is higher than the win price, then the model would have won the action and that win is recorded. If a win price is not available, then a win price is simulated based on the real distribution of second best bids using Cidewalk's bid price as a lower bound, and if the model's bid price is higher than the simulated win price the model would have won the auction and that win is recorded. In all other cases the model lost and the loss is recorded.

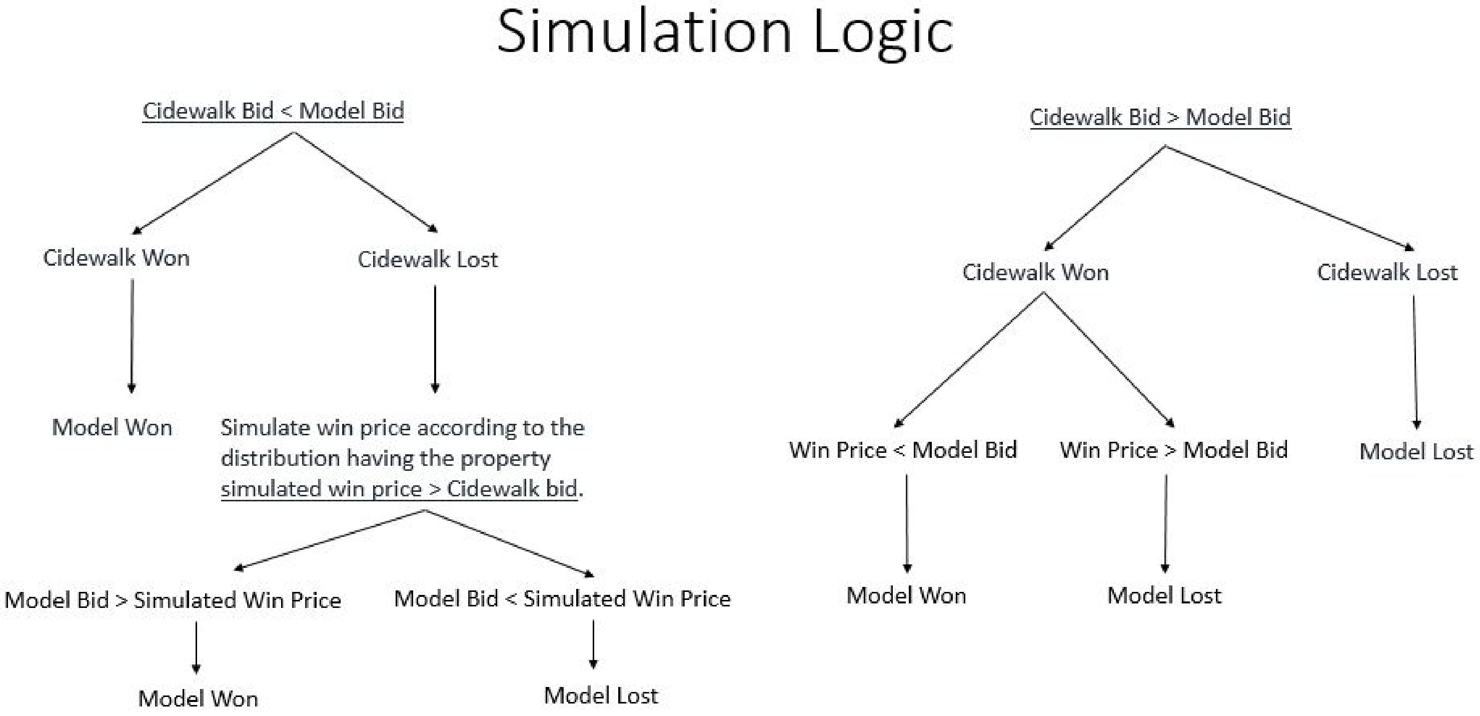


Figure 23: Decision Tree For Simulation.

This process is repeated until the number of auctions completed reaches the number of auctions Cidewalk bid on and then the total number of wins for Cidewalk and the model are compared along with the total cost of auctions won by Cidewalk and the model. These resulting values are then compared to see how the model performs when compared to Cidewalk's current method. For the sample set of data considered in this study the total number of auctions Cidewalk and the model won is 130,991, and the amount lost using the model was approximately 6%. Upon realization that the model will be bidding on more auctions than Cidewalk does in a real-life setting the number of auctions available was then extended to 16 times the number of auctions Cidewalk bid on, as Cidewalk only bids on 1 auction for every 16 auctions. For a sample set of data considered in this study the results with the extended number of available auctions is then, 130,991 wins for both the model and Cidewalk, and the overall savings using the model rather than Cidewalk's method was approximately 17%. The case, where the model lost 6%, is not accurately estimating the total number of available auctions, or n, properly. Therefore, the actual real-life savings would have been

[2] From the simulation logic it can be observed that the model is not bidding against Cidewalk, which represents a real-life scenario as only one model would run at one time.

approximately 17%.

4 Future Work and Considerations

This project takes a theoretically optimal bidding strategy that finds an optimal bid for Vickrey auctions and adapts the strategy to be functional in the mobile advertisement market. However, there are still many ways in which the strategy can be improved and things to take into consideration. The model is currently unable to differentiate the quality of one auction from the next. Therefore it does not recognize that winning some auctions may be more beneficial to Cidewalk in the long-run. It is also unable to adapt the optimal bid strategy if the real distribution of second best bids changes. The model does not currently have an user interface or dashboard set-up to help monitor its performance throughout the day, and the optimal time period for performing Bayesian updating on the mean and standard deviation of the real distribution of second best bids was not determined. These are all concepts that could be built upon if future improvements are made to the current method and some gaps in the current method that could be addressed.

4.1 Distribution Detection

Currently the model is capable of producing an optimal bid price for a given auction, but to do so it requires the distribution of best bids of the competitors to be known prior to calculation. Therefore, the data would have to be analyzed and a distribution would need to be determined separately through data analysis using tools such as QQ plots and log-likelihood as this functionality is not built into the model or the algorithms created to perform the calculations for the model. The mobile advertisement marketplace is incredibly volatile which makes the odds of having the same distribution or even the same family of distributions all of the time highly unlikely. To improve overall usability and increase automation building a method within the algorithms created for this project that detects what distribution best represents the given data set periodically and then performs calculations based on the determined distribution would decrease the chance of errors due to poor assumptions of the real distribution of second best bids.

4.2 Quality Recognition

Implementation of a method or methods that recognizes the quality of impressions would prove to be useful for Cidewalk, although it could in some cases increase the overall cost. For example, creating a way to filter out auctions where the mobile advertisement would show up on a phone with some form of ad prevention as with this software it would be less likely that the potential customer would even be able to see the advertisement making the quality of auctions like these low. Another example, would be devising a way to filter auctions based on the number of times that a given individual has clicked on advertisements in the past, since getting a potential customer to click on an advertisement promotes the business, which in turn makes Cidewalk's clients pleased with their service. Although both of these examples would increase the satisfaction of Cidewalk's customers, which benefits Cidewalk as this means the customers are more likely to continue using their service, they could cause the overall cost of the model to increase since it is possible that the auctions in both of these cases are higher priced.

4.3 User Interface

The development of a user interface for the model could improve the overall efficiency and effectiveness of the model, as it would allow the user to make important decisions that impact how the model runs and performs. For example, implementing a pop-up window at the end of updating period that reports to the user numerous QQ plots comparing different distributions and allowing the user to pick the one that represents the best fit could prevent the model from choosing a distribution that may not be the best. Another example, would be the development and design of a functional dashboard that tracks how the model is bidding and winning auctions to ensure that the model does not overbid or underbid and that the model is winning enough auctions to fulfill Cidewalk's quota. Within the dashboard there could also be a notification that alerts the user to any of the aforementioned abnormalities to ensure that potential errors are addressed before they have negative repercussions for Cidewalk.

4.4 Optimal Updating

Currently the model utilizes Bayesian updating on a hourly basis to update the mean and standard deviation of the real distribution of second best bids. In this project it was determined that an hour would be a sensible time period because the sample data differed within an hour interval, which would make updating on a daily basis not effective at grasping the changes that occur throughout the day. Multiple tests were not performed to ensure that this time period is optimal. Therefore, if the model was updated in the future it could benefit from comparing how updating the mean and standard deviation of the real distribution of second best bids at different time intervals effects the overall cost for the model. Real-time updating after every auction would be far too computational expensive and therefore, when testing the time intervals one should try and balance computational efficiency and overall cost to find an optimal solution that takes into consideration the run-time of the algorithms used in the model.

5 Conclusions

This project aims to implement a bidding model to help Cidewalk determine optimal bids, and in turn decrease the cost of the online advertisement spaces that they win in the exchanges. This objective is met by identifying a reasonable approximation of the distribution of the second best bid, finding a method to deal with missing values in the collected sets of sample data, and incorporating Bayesian updating, which dynamically updates the parameters of the distribution to adapt to the ever-changing environment in the exchanges. The exchanges' decision not to disclose the winning prices makes it impossible to know the best bid of the others when Cidewalk loses an auction. However, this information is necessary to recreate the whole picture of how the second best bid behaves and if neglected, may lead to incorrectly determining bid prices. It is possible to develop a method that models the distribution of second best bids using the distribution of observed win prices, which helps give a more accurate estimate of the optimal bids. As a result, the implemented model saves almost 17% when compared to Cidewalk's current bidding strategy, and this number can potentially be improved by correctly identifying the optimal updating period. This model allows Cidewalk to minimize their cost and as a result, increase their profit. The model's strategy also makes it more difficult for a competitor to reverse engineer Cidewalks bidding strategy and exploit this information to their advantage as Cidewalk's bids are constantly changing.

In other words, in this project we estimate the distribution of the second best bids, which is then used to calculate optimal bids. To address changes in the environment, the Bayesian updating is used to allow the model to learn on new data as it arrives to grasp the changing trends and update optimal bids accordingly. Thus, this is a robust approach that adapts to the dynamic marketplace and can be used to effectively bid on auctions optimally for the foreseeable future.

Appendices

A Appendix A

Analyzing the sample data to discover the best approximation for the distribution of second-highest bids or winning prices required the use of various methods and tools due to the sheer volume of sample data collected in all of the various sets of sample data. When working with data of a high volume, programs that are normally used for data analysis such as Excel cannot be used due to size limitations built into these programs. Even when using tools and programs that do not have built-in size restrictions data sets of high volume can exceed the amount of memory built-in to the computer or laptop being used. Therefore, the method in which files containing large quantities of data are read into memory must be considered. The tools used for analysis of the data collected for this project were Python, and R.

A.1 R

R is a programming language and software used widely to perform statistical analysis on data. According to the official web site ”R provides a wide variety of statistical (linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering,...) and graphical techniques, and is highly extensible ... One of R’s strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulas where needed.”[6] It is a free software that can be downloaded from the official web site, which makes it popular among statisticians and data analysts. Popularity of the R language is also related to the variety of different user-created packages that one can use to satisfy their own needs. For this project R was used to quickly assess the data and find the proper distribution. Usage of the log-likelihood values allowed quantitative determination of the best fitted distribution. Values were obtained using the function $logLik()$ that is commonly used for model fitting by maximum likelihood. QQPlots were used to visualize how well the data fits a chosen distribution. In order to process the data different packages were used such as ff, MASS, fitdistrplus, stats, pracma, nleqslv, rgl, mosaic.

A.2 Python

Python is an object oriented programming language, which according to The American Heritage Science Dictionary, is "A schematic paradigm for computer programming in which the linear concepts of procedures and tasks are replaced by the concepts of objects and messages. An object includes a package of data and a description of the operations that can be performed on that data. A message specifies one of the operations, but unlike a procedure, does not describe how the operation should be carried out."[7] This specific programming language has multiple add-ins which have been created to help provide pre-defined functions for the user to utilize. These add-ins make Python a great programming language to analyze data statistically and that is what this project specifically used the language for. The add-ins scipy and matplotlib were found to contain the most beneficial functions for analyzing data for this project through scipy's functions fit, and probplot and matplotlib's ability to produce visually appealing graphs. The scipy add-in's fit function allows the user to fit a given set of data to a distribution whose attributes can be found within the same add-in and produces the parameters that allow the best fit to the given data set. The scipy add-in's probplot function allows the comparison of the fit of a given set of data to a pre-defined distribution through the use of a QQPlot. When using Python it is important to consider the amount of built-in memory in the computing device being used and with extremely large data sets like the one's this project is dealing with it was best to read in the files line-by-line rather than all at once to avoid running out of memory.

References

[1] Youwei Hu, Jeremy Macaluso *Optimal Bid Pacing for Online Ad Impression Vickrey Auction Markets*, WPI, MQP report, 2015.

[2] Jonathan Levin, *Auction theory*, Oct. 2004, available at http://web.stanford.edu/ jdlevin/Econ%20286/Auctions.pdf

[3] NIST/SEMATECH *e-Handbook of Statistical Methods*, http://www.itl.nist.gov/div898/handbook/, 10/30/2013

[4] Fienberg, Stephen E. *"When Did Bayesian Inference Become Bayesian?"* Bayesian Analysis: 1-40. 2006. Web. 13 Mar. 2016. http://www.stat.cmu.edu/ fienberg/fienberg-BA-06-Bayesian.pdf

[5] Burkett, John. "Bayesian Statistics" University of Rhode Island, Department of Economics. Web. 25 Dec. 2015. http://www.uri.edu/artsci/ecn/burkett/545lect3.pdf.

[6] *R official website*, https://www.r-project.org/about.html

[7] Object-oriented programming. (n.d.). The American Heritage Science Dictionary. Retrieved March 13, 2016 from Dictionary.com websitehttp://www.dictionary.com/browse/object-oriented-programming