Bidding Wars: A Look into Consumer Tendencies on Ebay

Ryker Dolese

Introduction

The dataset at the heart of this project centers around eBay bidding wars, where users compete for the chance to win a product through online auctions. These auctions have a unique competitive environment where bidders place bids, and the item goes to the highest bidder when the auction concludes. The specific focus of this project is on eBay auctions for Palm Pilot M515 PDAs, which were a popular personal digital assistant (PDA) device from the early 2000s. A Palm Pilot M515 PDA was a handheld electronic device designed for various tasks, including organization, note-taking, and data management, and this project will explore the dynamics of eBay auctions for these devices.

Statement of Purpose:

The primary objective of this project is to examine the factors that influence the final selling price of Palm Pilot M515 PDAs in eBay auctions. To achieve this, we aim to address the following questions and goals:

1. **Price Prediction**: Can we develop a predictive model that accurately estimates the closing price of Palm Pilot M515 PDAs in eBay auctions based on key variables such as the bid, bidder feedback rating, and bid timing?

2. **Bidder Behavior Analysis**: What are the patterns and tendencies of eBay users in online auctions, and how do these behaviors impact the final selling price? Are there any effects of particular interest regarding bidder actions and outcomes?

Background and Plan:

To frame our project, we conducted extensive background research into the world of online auctions, eBay in particular, and the unique characteristics of Palm Pilot M515 PDAs. Our project's goals will be addressed through a combination of statistical analyses, including regression analysis, to predict selling prices. We will also utilize descriptive statistics and visualizations to explore bidder behavior and auction dynamics. Confidence intervals and hypothesis tests will help us draw meaningful conclusions about the factors that influence auction outcomes.

Data Collection and Assessment:

In our dataset, acquired from https://www.modelingonlineauctions.com, there were 7 initial fields:

1. auctionid - This field serves as a unique identifier for each auction in the dataset. It distinguishes one eBay auction from another and will be useful for tracking individual auction records.
2. bid - The "bid" field represents the proxy bid placed by a bidder in the eBay auction. This is the amount the bidder is willing to pay for the item.
3. bidtime - This field indicates the time, in days, when the bid was placed relative to the start of the auction. It provides insights into the timing of bids throughout the auction's duration.
4. bidder - "bidder" contains the eBay username of the bidder participating in the auction. It identifies who is placing the bids.

5. bidderrate - This field represents the eBay feedback rating of the bidder. eBay users typically provide feedback on their experiences with a bidder, and this rating can indicate their trustworthiness and history on the platform.
6. openbid - "openbid" denotes the opening bid set by the seller when the auction began. It represents the initial price at which the item was offered for bidding.
7. price - The "price" field indicates the closing price at which the item was sold in the auction. This closing price is calculated based on the second-highest bid plus an increment, as per eBay's auction rules.

Exploratory Data Analysis

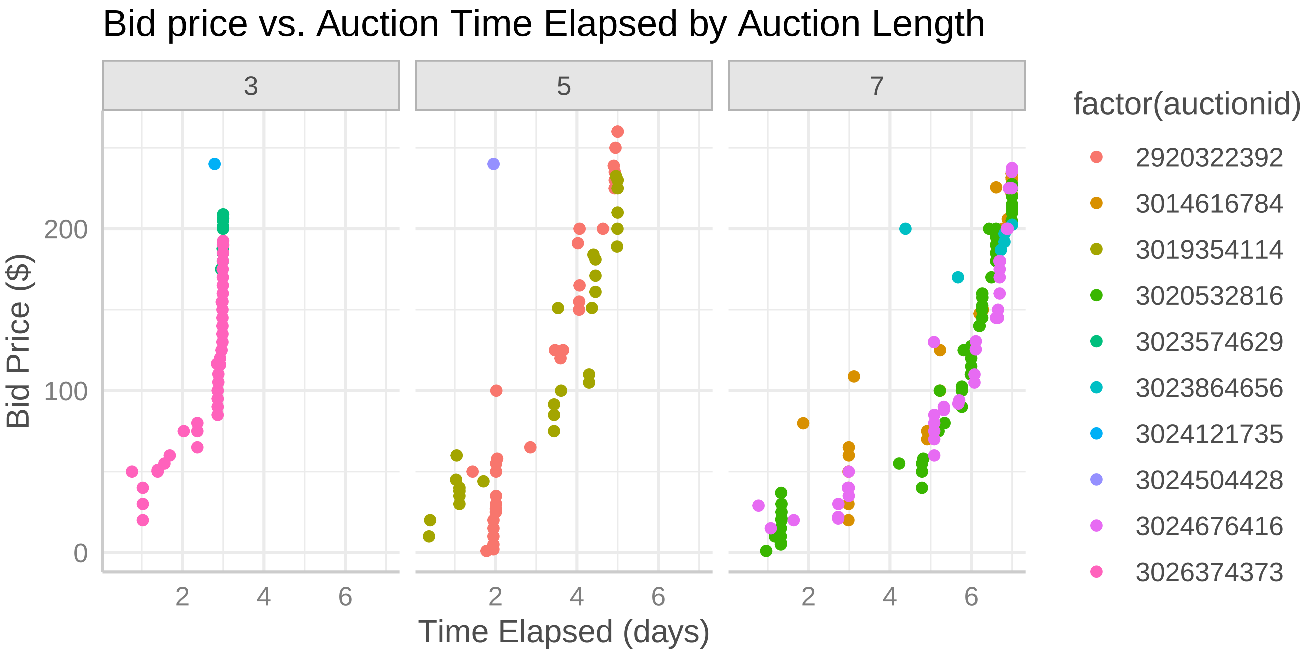

Figure 1: Price and Auction Length

Figure 1 provides a visual representation of a random sample of auctions and their bids over time. Notably, it reveals that, as auctions approach their final day, there is a significant and rapid increase in bidding activity, suggesting a distribution with exponential characteristics. Moreover, the plot indicates that auctions with shorter durations, such as 3 days and 5 days, experience this price increase much earlier in the bidding process.

This observation could imply that bidders are more motivated to participate as the auction deadline approaches, potentially due to a sense of urgency to secure the item they desire. Additionally, shorter auction durations may prompt quicker bidding and competitive behavior, leading to early price escalation.

Feature Engineering

As mentioned above, we have the data for every individual bid. However, it may be most ideal if we look at each auction wholistically. Therefore, we have grouped by 'auctionid' to create new metrics that could predict selling price. We use our previous data to identify these new features:

1. **AvgBid**: represents the average bid placed in the auction. It's calculated as the mean of all the individual bids.
    - Formula: `AvgBid = mean(bid)`
2. **TotalBids**: TotalBids is the count of the total number of bids in the auction. It provides the total bidding activity for all auctions.
    - Formula: `TotalBids = n()`

3. **BidsBeforeLast**: indicates the number of bids placed before the last day in an auction. It helps identify the level of bidding activity prior to the closing stages of an auction.

   - Formula: `BidsBeforeLast = sum(Biddate == "BeforeLast")`

4. **BidsAfterLast**: represents the number of bids placed after the last day in an auction. It reflects the activity that occurs during the final moments of an auction.

   - Formula: `BidsAfterLast = sum(Biddate == "Last")`

5. **AverageRating**: the mean eBay feedback rating of the bidders in the auction. It provides an overall assessment of the bidders' trustworthiness and performance on eBay.

   - Formula: `AverageRating = mean(bidderrate)`

6. **UniqueBidders**: counts the number of unique eBay usernames (bidders) participating in the auctions. It identifies the diversity of bidders involved in the bidding process.

   - Formula: `UniqueBidders = length(unique(bidder))`

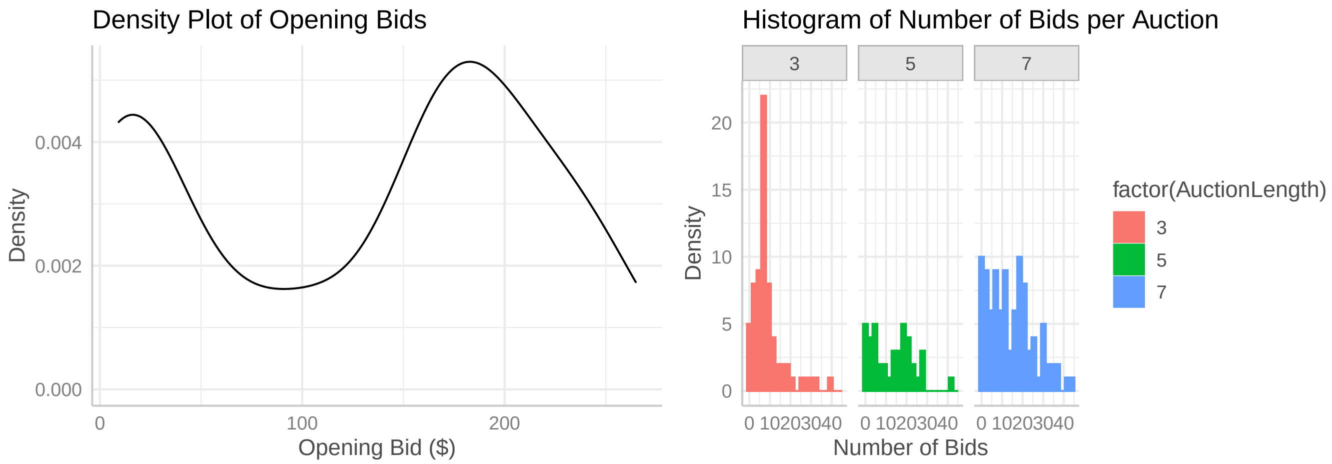

Figure 2: Distribution of Bids

Figure 2 provides a clear representation of the distribution of opening bids, revealing a bimodal pattern. The majority of opening bids fall into two distinct categories: one with very low bids, typically less than $10, and another with considerably higher opening bids, often around $200, approaching the eventual selling price.

This bimodal distribution could suggest two different types of client behavior:

1. Low-Budget Bidders: Bidders in the first group with very low opening bids may be looking for bargains or lower-priced items. They might be more price-sensitive and cautious in their initial bids, preferring to start with minimal offers.
2. Competitive Bidders: The second group with high opening bids approaching the final selling price might consist of more competitive or confident bidders. They might be willing to make larger initial bids to assert their interest in securing the item and to deter potential competition.

The second plot in Figure 2 illustrates the distribution of the number of bids per auction, revealing distinct patterns among different auction lengths. It's apparent that the number of bids tends to be right-skewed, particularly in shorter auctions like the 3-day auctions, where a significant proportion of auctions have fewer than 10 total bids. In contrast, the 5 and 7-day auctions exhibit a somewhat more uniform distribution but still maintain right-skewed tendencies.

Short auction durations appear to foster more rapid bidding, with a concentration of auctions having a limited number of bids. This could indicate a sense of urgency among bidders to participate and secure items quickly. Longer auctions display a somewhat more balanced distribution of bids, potentially reflecting a more gradual bidding process with less urgency. Bidders might take their time to assess the value of items and place bids at a steadier pace.

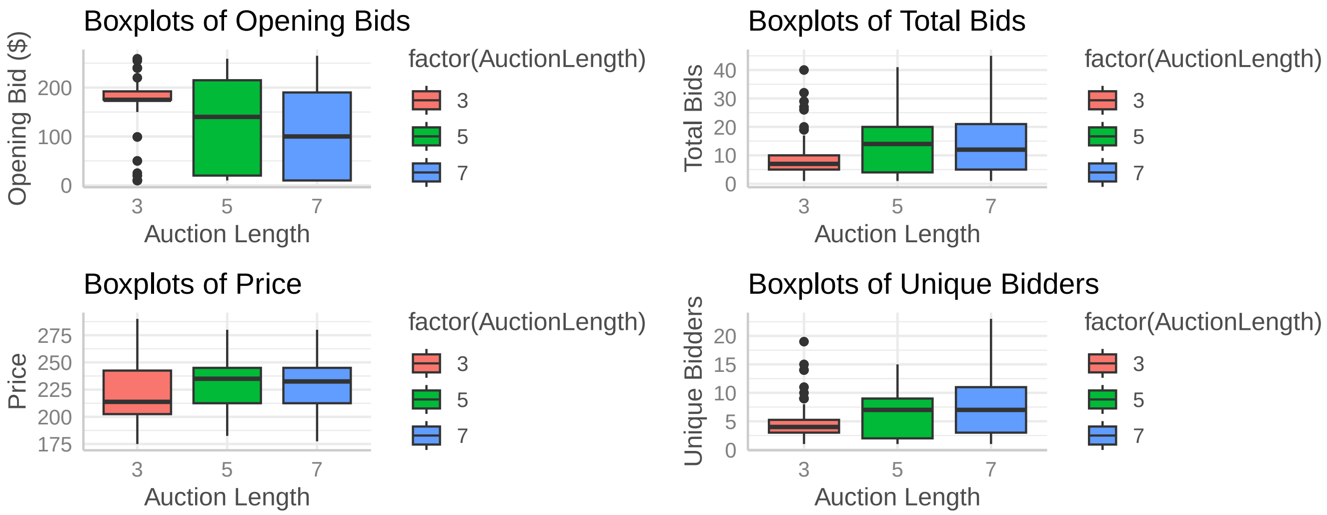

Figure 3: Boxplot of Numerical Variables

In this collection of boxplots (Figure 3), we observe varying trends across different measures for distinct auction lengths. Notably, the 5 and 7-day auctions exhibit similar characteristics, while the 3-day auctions display more sporadic patterns. In the realm of opening bids, the 3-day auctions show a broader range of values with pronounced outliers, indicating greater variability.

Conversely, the 5 and 7-day auctions reveal more tightly clustered opening bid values. The number of unique bidders tends to rise as auction length increases, with longer auctions showcasing a wider spread in the distribution of unique bidders. In 3-day auctions, there's a concentrated distribution with occasional extreme outliers. Similar patterns emerge with total bids, mirroring unique bidders. Remarkably, 3-day auctions maintain a wider range of total bid counts. Selling prices, on the other hand, appear relatively consistent across auction lengths, but 3-day auctions notably feature lower median selling prices compared to their longer counterparts, suggesting unique pricing dynamics in shorter auctions.

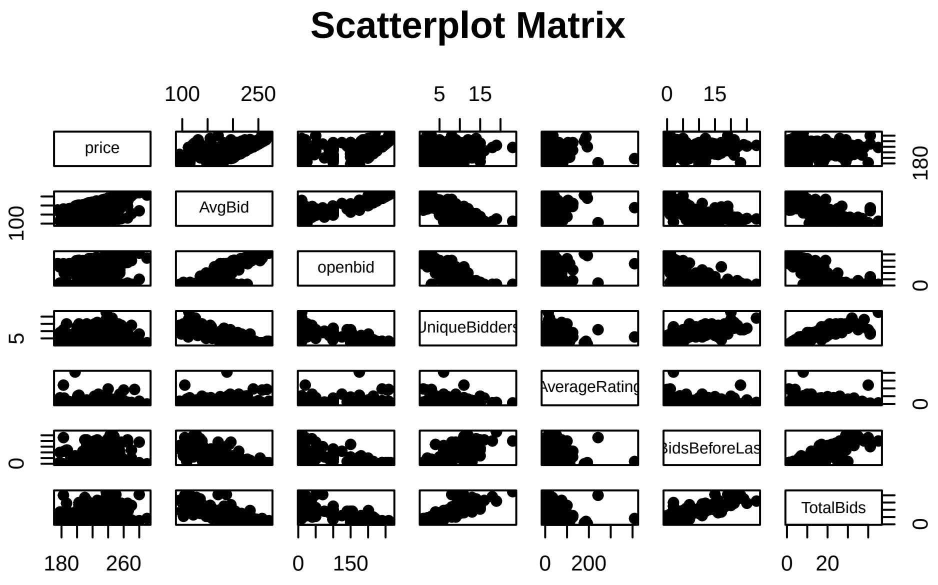

Figure 4: Correlations

Table 1: Correlation Matrix

	price	AvgBid	openbid	UniqueBidders	AverageRating	BidsBeforeLast	TotalBids
price	1.00	0.42	0.11	0.04	-0.05	0.10	0.10
AvgBid	0.42	1.00	0.88	-0.79	-0.07	-0.74	-0.76
openbid	0.11	0.88	1.00	-0.84	-0.10	-0.80	-0.83
UniqueBidders	0.04	-0.79	-0.84	1.00	0.11	0.80	0.90
AverageRating	-0.05	-0.07	-0.10	0.11	1.00	0.13	0.09
BidsBeforeLast	0.10	-0.74	-0.80	0.80	0.13	1.00	0.87
TotalBids	0.10	-0.76	-0.83	0.90	0.09	0.87	1.00

The correlation matrix (Table 1) and pairplot (Figure 4) of potential regressors reveal significant insights into the relationships among the variables in the dataset. Notably, a strong positive correlation exists between the selling price and the average bid (0.4211), indicating that, on average, higher bid amounts correspond to higher selling prices. Conversely, shorter auction lengths, such as 3-day auctions, exhibit higher negative correlations with total bids, open bid, unique bidders, and average bid. The most notable correlation in the matrix is the high negative correlation (-0.8266) between total bids and the opening bid, signifying that as the total number of bids increases, the initial bid amount tends to decrease. Furthermore, the number of unique bidders demonstrates a substantial negative correlation with total bids, open bid, and average bid, implying that an increase in unique bidders is associated with reduced values for these variables. These findings shed light on the complex interplay between auction dynamics and the variables at play, providing a solid foundation for further analysis and modeling. Of note, some variables even appear to have exponential relationships with each other; for instance, looking at the scatterplot between AvgBid and openbid, the relationship appears more exponential rather than linear. As such, we might want to look into regressor transformations when we begin modeling.

Modeling

Now, we will begin producing our model to predict selling price. We will begin with a relatively simple linear regression model where we see the effect of average bid and auction length on price.

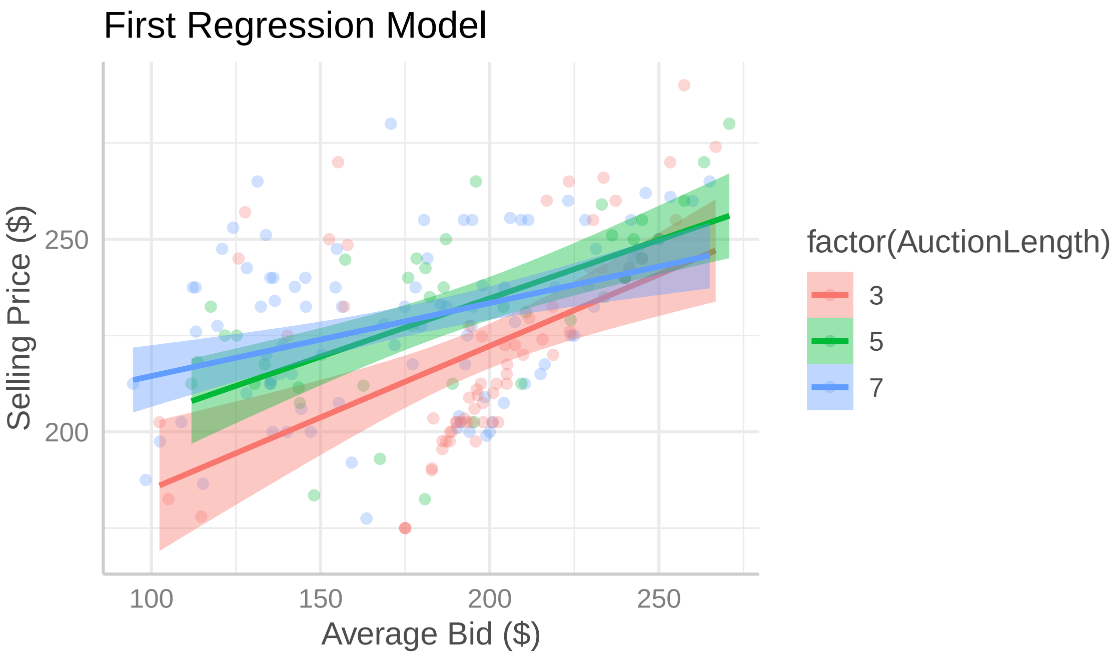

Figure 5: Regression Modelling

In this regression model plot (Figure 5), several key observations stand out. Firstly, there's a clear and positive linear correlation between the average bid and the selling price. As the average bid increases, the selling price consistently rises, indicating a direct relationship between these two factors. Notably, the plot showcases varied slopes for different auction lengths, emphasizing that the strength of this correlation differs across auction durations. While auction length does exert some influence on the correlation, it doesn't significantly alter the overall positive trend. In essence, this plot underscores the robust positive relationship between average bid and selling price, with some nuanced variations in the influence of auction length.

Table 2: First Model Summary

term	estimate	std.error	statistic	p.value
(Intercept)	162.0252514	8.6645926	18.699697	0.0000000
AvgBid	0.2592163	0.0349663	7.413324	0.0000000
AuctionLength	3.1455342	0.8363027	3.761239	0.0002211

Table 3: Additional Statistics

r.squared	adj.r.squared	sigma	statistic	p.value	MSE	det	AIC	BIC	Cp
0.2308912	0.2233137	20.85947	30.47092	0	443.8637	0.9665323	1254.555	1264.539	374.8016

The coefficients in Table 2 reflect the estimated effects of the independent variables on the dependent variable. The intercept indicates the expected selling price when all other variables are zero, which might not have a practical interpretation. The coefficient for AvgBid suggests that for each one-unit increase in the average bid, the selling price is estimated to increase by approximately 0.2592 dollars. Likewise, the coefficient for AuctionLength implies that a one-unit increase in auction length is associated with an estimated selling price increase of approximately 3.1455 dollars. The accompanying statistics (Table 3) reveal that about 23.09% of the variation in selling price is explained by the model, indicating a statistically significant model. The degrees of freedom for residuals and the number of observations in the dataset are also provided.

New Model: MLR

Our previous model performed alright. Both coefficients had p-values well below the alpha = 0.05 threshold. However, only 23% of our variation can be explained by our model, which is not ideal. With the additional variables at our disposal, we can implement a more robust MLR model and see how it performs.

Table 4: Step-Wise Selection

Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
	NA	NA	205	114845.73	1304.635
+ AvgBid	-1	20361.294	204	94484.44	1266.433
+ TotalBids	-1	46387.723	203	48096.72	1129.337
+ openbid	-1	9018.974	202	39077.74	1088.559
+ BidsBeforeLast	-1	4197.303	201	34880.44	1067.152
+ UniqueBidders	-1	3192.627	200	31687.81	1049.377
+ AverageRating	-1	1001.911	199	30685.90	1044.759

Although we understand the limitations of a stepwise algorithm approach (including biased and high r-squared values, f statistics and p-values being overestimated due to multiple-comparison, and the inability to find optimal subsets of regressors), we have used step-wise regression as a start to identify some possible regressors. We have decided to use a step-wise forward algorithm to find the best MLR model by AIC. The results are in Table 4. As one can see, the best combination of regressors to minimize AIC is AvgBid, openbid, UniqueBidders, AverageRating, BidsBeforeLast, and TotalBids.

Table 5: Second Model Summary

term	estimate	std.error	statistic	p.value
(Intercept)	44.1689023	8.2956389	5.324352	0.0000003
AvgBid	0.9356855	0.0446241	20.968154	0.0000000
openbid	-0.1460181	0.0257121	-5.678966	0.0000000
UniqueBidders	2.3295070	0.4974129	4.683246	0.0000052
AverageRating	-0.0512631	0.0201110	-2.549011	0.0115567
BidsBeforeLast	1.5461870	0.2744460	5.633848	0.0000001
TotalBids	0.3828934	0.2436707	1.571356	0.1176886

Table 6: Additional Statistics

r.squared	adj.r.squared	sigma	statistic	p.value	MSE	det	AIC	BIC	Cp
0.7328077	0.7247516	12.41775	90.96364	0	154.2005	0.002376	1044.759	1068.054	7.689019

Table 7: VIFs

	x
AvgBid	4.754943
openbid	6.596297
UniqueBidders	6.312976
AverageRating	1.022501
BidsBeforeLast	4.565991
TotalBids	7.850924

Our improved model is notably strong, with an R-squared value of approximately 73.5%, adjusted to 72.4% when accounting for predictors, as seen in Table 5 and 6. The MSE is around 154 and the Residual Standard Error is around 12.4 (much lower than the previous model), reflecting typical prediction errors. The model is highly statistically significant, as indicated by an F-statistic with a p-value of approximately 0. Lower p-values for individual predictors suggest their significance. Overall, the model effectively explains variation in the dependent variable, attaining a robust R-squared and a significant F-statistic. The AIC and BIC are also significantly lower than that of the first model. The determinant of correlation is relatively low suggesting some multicolinearity, but it is above the threshold of 0.0001, so the model can still be considered valid. The VIFs reflect a similar result with openbid, UniqueBidders, and TotalBids each having VIFs greater than 5.

Transforming the MLR Model: Accounting for Multicolinearity

To account for the multicolinearity issue, we have several options: remove variables, produce linear combinations of existing regressors, or identify other forms of regression (such as LASSO or Ridge). We have first opted for the linear combination approach. Because openbid and TotalBids are inherently related (for instance, if someone has a high opening bid we would expect less total bids in the auction), we have decided to add these two predictors to form 1 new predictor: x2 = openbid+TotalBids. Therefore, we have opted for a MLR model with these the same predictors but have consolidated openbid and TotalBids in hopes to reduce multicolinearity. The results of such regression are seen in Tables 8 and 9.

Luckily, our approach worked to an extent. The R-squared and R-squared adjusted both remained relatively high, the MSE hardly changed, and the high VIF values fell considerably. Only the new “x2” regressor retains a VIF greater than 5, but just barely. For this reason, our model appears to be especially robust.

Table 8: Third Model Performance

r.squared	adj.r.squared	sigma	statistic	p.value	MSE	det	AIC	BIC	Cp
0.726389	0.7195488	12.53456	106.193	0	157.9048	0.02234	1047.649	1067.616	10.48604

Table 9: VIFs

	x
AvgBid	4.754935
x2	5.426999
UniqueBidders	3.943519
AverageRating	1.018707
BidsBeforeLast	3.193711

Regularization with Lasso

While we are particularly satisfied with our previous approach, another more systematic strategy is using a regularization model. Here, we will employ LASSO (Least Absolute Shrinkage and Selection Operator) regression, which is a linear regression technique used for feature selection and regularization. It's a modification of ordinary least squares (OLS) regression particularly apt in accounting for multicolinearity.

$$\text{minimize}\left(\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\lambda\sum_{j=1}^{p}|\beta_j|\right)$$

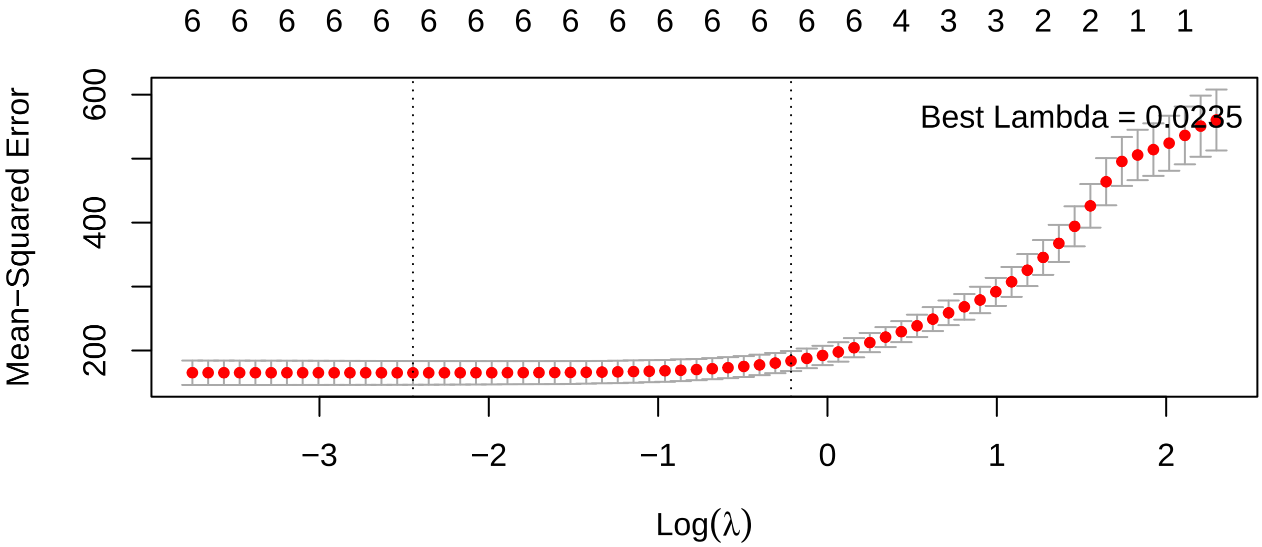

Figure 6: Lasso Regularization

Table 10: Summary of Lasso Regression

	Estimate
(Intercept)	45.0054594
AvgBid	0.9301959
openbid	-0.1442501
UniqueBidders	2.3190214
AverageRating	-0.0503891
BidsAfterLast	-1.5206000
TotalBids	1.9150019

In our Lasso Summary, we can see how our coefficients changed when compared to model 2. While none were entirely eliminated, the coefficient for openbid shrunk, while that of TotalBids increased (Table 10). This suggests the model is not veering entirely away from using openbid as criteria, but that it prefers TotalBids when considering the colinearity issue.

Table 11: Lasso Results

Lambda	MSE	AIC	BIC	R_squared
0.0235075	153.4481	1044.785	1068.08	0.7327735

Solving for our regularization parameter lambda using k-fold cross-validation, we find that lambda equals around 0.023. This value produces the lowest mean squared error. Moreover, the small lambda value suggests the LASSO effects will not be particularly strong. This is good and suggests the colinearity isn't an overbearing issue.

Notably, The outcomes of Lasso regression have yielded an R-squared value that is slightly higher than our conventional MLR model (model 3) and an MSE that is even lower, as seen in Table 11. The AIC and BIC are nearly identical. With the LASSO model's ability to minimize multicolinearity, this is likely the most suitable model and we will move forward for validation. Before we move to further validation, however, it is important to acknowledge the inclusion of bias within the LASSO algorithm: Lasso regression trades off an increase in bias with a decrease in variance. Lasso regression goes to an extent where it can enforce beta coefficients to become 0.

In the scatter plot of price vs. fitted values (Figure 7), it is evident that our model performs well in predicting selling prices. Additionally, there no discernible pattern among the residuals as seen in the residual plot. The residuals are scattered evenly around the horizontal line at zero, indicating that they exhibit no systematic relationship with the fitted values. This observation suggests that a linear model is appropriate for the data, as there are no indications of heteroscedasticity or other issues that might violate the assumptions of linear regression. The absence of a clear pattern in the plot affirms the model's validity for the given dataset. The Q-Q plot shows that the quantiles follow a relatively normal distribution (same with the histogram) with not much deviation from the line, although there are some influential points in the dataset as seen in the Residuals vs. Leverage. Overall, the model seems valid.

Table 12: Modified Levene Test of Residuals

statistic	p_value
0.0087194	0.9933257

From a hypothesis testing perspective, we use a Modified Levene Test to determine whether the variance of residuals is homogeneous. With a high p-value as shown in Table 12, we conclude that the residuals appear to have a homogeneous distribution, further suggesting that a LASSO regression model is a good fit for this data.

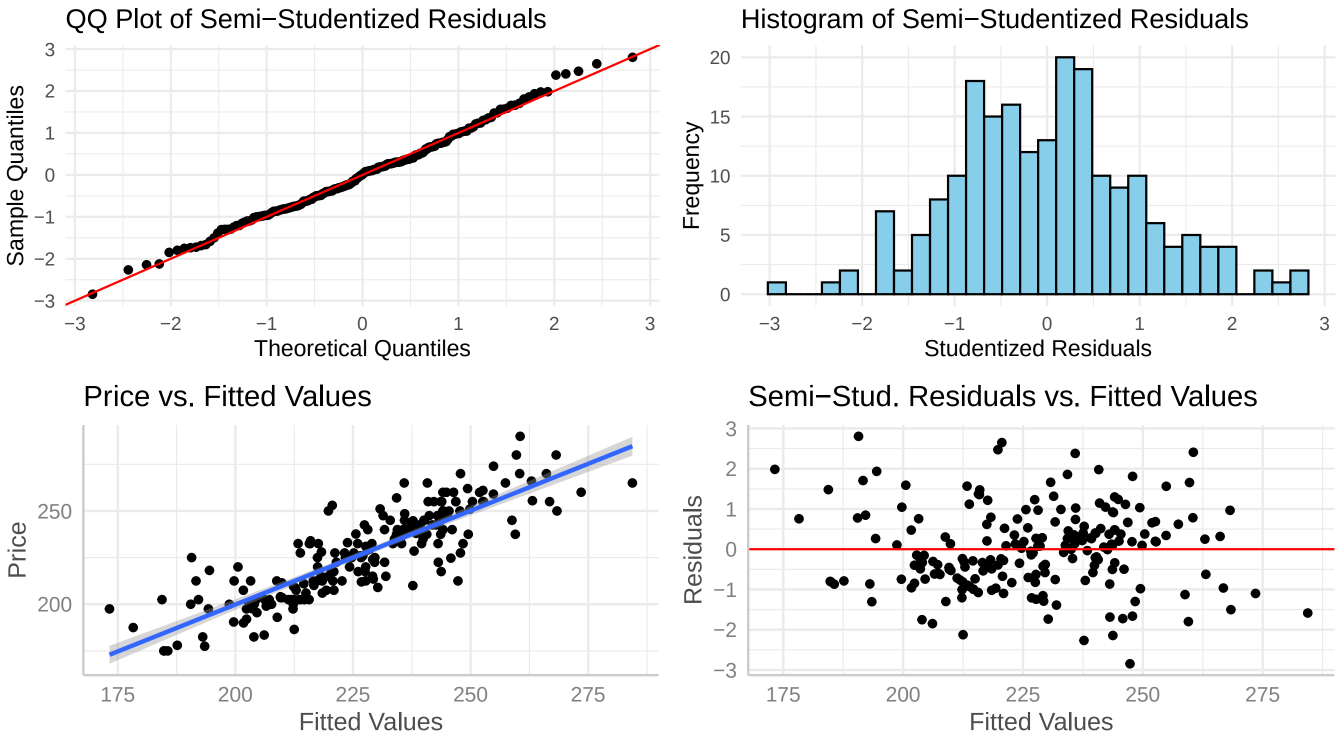

Figure 7: Model Validation

Discussion

The project began by using visualizations, including density plots, boxplots, and histograms, provided insights into bidder behavior and the influence of auction length on various factors like opening bids, total bids, and unique bidders. A correlation matrix was used to explore relationships between various variables, such as price, total bids, open bid, unique bidders, average bid, and auction length. This analysis revealed important insights into which factors are most strongly correlated with selling prices. Notably, shorter auction lengths exhibited more variability in opening bids, while longer auctions tended to have more unique bidders. Subsequently, the project examined the effectiveness of a multiple linear regression (MLR) model in predicting eBay auction selling prices for Palm Pilot M515 PDAs. After implementing 3 MLR models, we settled for a LASSO Regression model to account for the multicolinearity of predictors, acknowledging the bias-variance trade-off.

In terms of future work, one promising avenue is exploring alternative prediction models beyond multiple linear regression. Time series analysis could be a valuable approach, particularly for price prediction, as it can capture the temporal patterns and seasonality often present in auction data. Additionally, a deeper examination of bidder behavior patterns, including when most bids occur during the auction, can provide insights into eBay auction dynamics. This analysis can help optimize auction strategies for sellers and enhance predictions. Given the lower median selling prices observed in 3-day auctions, it's worth investigating the specific price dynamics in shorter auctions, such as the impact of auction parameters like start times or seller ratings on final selling prices. Experimenting with feature engineering to identify additional predictor variables and thorough model validation, including cross-validation, will contribute to the robustness of predictive models. Expanding the dataset by collecting additional eBay auction data can further enhance the prediction models.

Another consideration could be expanding these findings to a variety of products, not just Palm PDAs. It is unreasonable to expand the current results and models to all ebay products – but this is an emphasis for future research.

Appendix

Link to Code

References

Boaz Shmueli, 2020. “Modeling Online Auctions Datasets.” Retrieved from https://www.modelingonlineauctions.com/datasets

Self-Reflection

Overall, this assignment was incredibly fulfilling. I got to apply many of the concepts I learned in STAT 410 in a very practical way. Witnessing the successful application of regression techniques was particularly satisfying. As I delved into the project, I found myself utilizing various skills, especially in the context of regressor transformations, model validation, and statistical tests to enhance the model’s suitability. The project also addressed some long-standing curiosities I had about online auction dynamics. Answering questions like whether to start auctions at low or high prices and when to place bids provided practical insights. Maybe I’ll have a advantage when bidding on ebay items now! While some aspects remain unclear, the experience has ignited a curiosity to delve deeper into these observations. This project has reinforced my fascination with the world of statistics, and I eagerly anticipate applying these analytical skills in the future. The ability to draw meaningful conclusions from data and make informed decisions is a powerful aspect of statistical analysis that continues to captivate me.