

Visualisation

Report

Shill Bidding Fraud Detection

Analysis Report

2023

Prepared by:

Bobbi McDermott

Contents

[Introduction 3](#_Toc136207931)

[Section 1: Business Understanding 4](#_Toc136207932)

[Section 2: Data Understanding 4](#_Toc136207933)

[Section 3: Data Preparation 5](#_Toc136207934)

[Exploratory Data Analysis 5](#_Toc136207935)

[Outliers 5](#_Toc136207936)

[Descriptive Statistics 8](#_Toc136207937)

[Visualizing correlations 9](#_Toc136207938)

[Feature Selection 10](#_Toc136207939)

[Feature Importance’s 12](#_Toc136207940)

[PCA 13](#_Toc136207941)

[Kmeans 14](#_Toc136207942)

[Dashboards 17](#_Toc136207943)

[Section 4: Conclusion 18](#_Toc136207944)

[Bibliography 19](#_Toc136207945)

# Introduction

This dataset was created via eBay shill bids (fraudulent bids) for a well-known product. It includes bids from several different bidder accounts, and the purpose of this report is to enable academics to forecast future bids.

Shill bidding is a concern for eBay since it reduces the buyer's surplus, which causes customers to switch to another platform (Ingebretsen Carlson & Wu, 2022).

# Section 1: Business Understanding

Ebay is an online marketplace for buying and selling which also supports auctioning.  Shill bidding can alleviate the problem of the winner's curse by lowering the price and it can, thus, provide benefits to bidders **CITE the effect of shill bidding on prices.** It is therefore essential for eBay to predict patterns in this bidding to endorse measures to maintain integrity of the business and not deter sellers from using the platform.

# Section 2: Data Understanding

This is an entirely numeric dataset with 6321 observations and 13 features. The ‘Class’ feature being the target variable to determine anomalous bids apart from the normal bids. Throughout the prescriptive and descriptive phase of the analysis, the same features recurred in aspects from balance, outliers, ratio, correlation and feature importance. This is reassuring in that it can determine a pattern that will help us in machine learning for the predictive (University of Bath, 2021)

# Section 3: Data Preparation

## Exploratory Data Analysis

To gain insight to a dataset, EDA is crucial as nothing can be achieved without data understanding to perform cleaning, preparation and modelling. It’s a numerical collection of aspects that attribute to bidding tendencies in eBay that can help predict normal or fraudulent bids

The normal to abnormal ratio is quite disproportionate. To view this is important as it gives relativity to what needs to be predicted. Seeing a graph is more relatable

## Outliers

A crucial part of data preparation Is viewing and determining outliers in features and whether they are relevant to the data frame as a whole Outliers are difficult to understand in a statistical context so visualizations support data manipulation towards the predictive process (Ansoleaga, 2021).

An interactive plot lends more weight in this category as it is difficult to communicate the information through this plot without translated guidance (Ansoleaga, 2021).

Here we can see the features containing outliers and distinct values that are anomalous

A picture containing text, screenshot, diagram, line

Description automatically generated

Interactive plot

A picture containing text, screenshot, diagram, line

Description automatically generated

A picture containing text, screenshot, font, line

Description automatically generated

A picture containing text, font, screenshot, line

Description automatically generated

A picture containing text, font, line, screenshot

Description automatically generated

Here we can see that we can easily depict the exact values and requires not visual analysis on the viewers part as the information is translated effortlessly

## Descriptive Statistics

Visualizations are probably the most relatable in descriptive and predictive process of data analytics. Descriptive give us a compass to what the data is showing us (Leard Statistics, 2018). The following graph shows initial insights to the rations of the features against the normal and abnormal bids. It indicates a strong leaning for specific features to be more inclusive in abnormal bidding, being Bidding\_Ratio, Successive\_Outbidding and Winning\_Ratio

A picture containing text, screenshot, diagram, colorfulness

Description automatically generated

This plot looks at the ratios of different attributes to the desired variables. The findings indicate a clear propensity for some qualities, including "Bidding\_Ratio," "Successive\_Outbidding," and "Winning\_Ratio," which demonstrate a higher level of involvement in atypical bidding behaviors. These characteristics are important markers in this analysis and offer useful insights into fraudulent bidding habits.

A plot of this kind, the selected visualization technique, completely matches the goals of the research. It allows us to develop a clear understanding of the characteristics that lead to fraudulent bids. Color is strategically used to distinguish between typical and anomalous bids, making interpretation simple. Additionally, the different line lengths clearly highlight the wide range of ratios

## Visualizing correlations

I initially viewed correlating features and then used density heatmap or a more communicative approach. I pin pointed features that were corelating to communicate the results to a non-technical audience. It reduces complexity to realize the relevant patterns and relationships

A screenshot of a computer screen

Description automatically generated with low confidence

This plotly express graph captures in fine detail the correlated aspects of these two features and communicates with easy the values being represented. This is an interactive plot in the notebook. I also performed this on Bidding\_Ratio and Winning\_Ratio also

A picture containing colorfulness, screenshot, rectangle, square

Description automatically generated

## Feature Selection

Viewing important features can support the predictive nature of a data source as it highlights contributors to the outcome. It is also imperative in the data preparation process as it can indicate if dimensionality reduction is necessary

A picture containing diagram, plan, map

Description automatically generated

I utilized the dtreeviz library as it is effective in communicating the decision tree breakdown in a more interpretable way, while also communicating the underlying logic (Stalling, 2021). This plot is dynamic in communicating the information both aesthetically and in detail. It uses Feature Space so that the intended audience can understand how the decision tree partitions the data (Parr and Grover, 2020)

## Feature Importance’s

A picture containing text, screenshot, line, plot

Description automatically generated

Feature importance’s were very obvious in their importance. Therefore I wanted a palatable consistency for viewing audience and have an aesthetic appeal as to not distort the front runners from the lesser valued features (Cotroneo, 2021). I also used percentages as they are much more understandable in terms of context of the balance of the important features.

## PCA

Dimensionality reduction is a logical approach before machine learning in this dataset to filter out noise and re-express the data (Uzila, 2023). I performed pca for kmeans clustering. I initially view the clusters with mglearn discrete scatter as it provides clear grouping

A picture containing map, screenshot, text

Description automatically generated

I viewed the first and second component, the first to capture the variance and the second to identify the patterns (Zakarai, 2023)

A picture containing screenshot, line, diagram, colorfulness

Description automatically generated

I used this colour bar as it is successful t highlighting patterns and relationships

## Kmeans

I decided to perform kmeans as the remit was a scatter plot and kmeans is affective an anomaly detection.

A picture containing text, diagram, screenshot, map

Description automatically generated

A picture containing text, plot, screenshot, diagram

Description automatically generated

Silhouette Cluster allows me to see the density and separation of the data, with the red dotted line indicating the average silhouette score. This is a good plot as it is easier to demonstrate the sizes and separation of the clusters

I also employed InterClusterDistance, but it was not as successful at presenting the data accurately

A picture containing text, screenshot, circle, diagram

Description automatically generated

Finally, I utilized the Sankey Library to create this flow chart

A picture containing screenshot, text, graphic design, graphics

Description automatically generated

The Sankey chart is a straightforward communicator in that it is a roadmap to the answer a stakeholder is seeking. It focuses on a single aspect while visually linking relationships. It is interactive and provides buttons such as download, box select, lasso select, reset view and source

## Dashboards

Python dashboards are an incredible tool and as they have even more capabilities than that of Power BI and Tableau. They are also more flexible. However, it excludes non technical personnel from creating them and are quite difficult to implement

Unfortunately, I failed in this remit. I tried to run Dash and managed to get the environment working but the graph was unsuccessful. I think ran Panel and the graph was also unsuccessful so I installed Voila and created a dashboard from there

A screenshot of a computer

Description automatically generated with medium confidence

# Section 4: Conclusion

The objective is to communicate the truth. However, human perception, on occasion cannot discount bias. It is this balance when it comes to the math, the business objective and the data truth (Learning Data Analysis, 2020) that we heavily rely on to make decisions, albeit social, political, environmental etc. Visualizations can also be a force for negative representation however, while numbers are hard to represent, shapes and colors can be misrepresented for personal gain.

Overall, this project was both fun and frustrating at the same time. I discovered libraries and processes that intrigued me, however time limitations restricted my exploration and experimenting.

There are multiple considerations when producing these graphs. One must consider the object, audience, method, understanding, visual appropriation, colors and information to name a few. They are a vital part in data analytics as having results is pointless without the ability to communicate them

# Bibliography

Ansoleaga, U.L., 2021. How to Detect, Handle and Visualize Outliers [WWW Document]. Medium. URL https://towardsdatascience.com/how-to-detect-handle-and-visualize-outliers-ad0b74af4af7 (accessed 5.28.23).

Cotroneo, A., 2021. 12 Data Visualization Color Palettes for Telling Better Stories with Your Data [WWW Document]. URL https://www.heavy.ai/blog/12-color-palettes-for-telling-better-stories-with-your-data (accessed 5.28.23).

Leard Statistics, 2018. Understanding Descriptive and Inferential Statistics | Laerd Statistics [WWW Document]. URL https://statistics.laerd.com/statistical-guides/descriptive-inferential-statistics.php (accessed 5.28.23).

Learning Data Analysis, L., 2020. Understanding truths - Learning Data Analytics: 1 Foundations Video Tutorial | LinkedIn Learning, formerly Lynda.com [WWW Document]. LinkedIn. URL https://www.linkedin.com/learning/learning-data-analytics-1-foundations/understanding-truths (accessed 5.28.23).

Parr, T., Grover, P., 2020. How to visualize decision tree [WWW Document]. URL http://explained.ai/decision-tree-viz/index.html (accessed 5.28.23).

Stalling, J., 2021. dtreeviz: a great decision tree visualization and model interpretability tool [WWW Document]. URL https://programmer.group/dtreeviz-a-great-decision-tree-visualization-and-model-interpretability-tool.html (accessed 5.28.23).

University of Bath, 2021. Descriptive, predictive and prescriptive: three types of business analytics [WWW Document]. The University of Bath. URL https://online.bath.ac.uk/content/descriptive-predictive-and-prescriptive-three-types-business-analytics (accessed 5.27.23).

Uzila, A., 2023. K-means Clustering and Principal Component Analysis in 10 Minutes [WWW Document]. Medium. URL https://towardsdatascience.com/k-means-clustering-and-principal-component-analysis-in-10-minutes-2c5b69c36b6b (accessed 5.28.23).

Zakarai, J., 2023. Principal Component Analysis (PCA) Explained | Built In [WWW Document]. URL https://builtin.com/data-science/step-step-explanation-principal-component-analysis (accessed 5.28.23).