CCT College Dublin

Assessment Cover Page

Module Title:	Data Visualisation Techniques
Assessment Title:	CA2 Individual
Lecturer Name:	David McQuaid
Student Full Name:	Simone Finelli
Student Number:	sba22524
Assessment Due Date:	28/05/2023
Date of Submission:	27/05/2023

Declaration

By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

Introduction

All the work performed is based on the “Shill Bidding” dataset from UCI. But what does Shill Bidding mean?

Basically, the only retailer eBay provides the option to bid an auction. The aim of the bidding process is to find the market price of a product based on the auction interaction and demand.

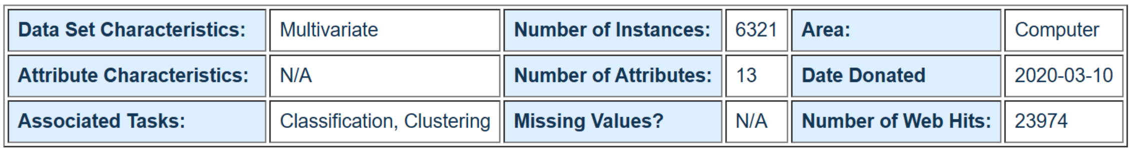
Data coming from eBay was scraped and processed by UCI to provide to the public a dataset where, based on different characteristics, a bid is identified either as normal (class 0) or as anomalous (1). The anomalous bid is also called a “Shilling Bid”:

Shill bidding in English auction is the deliberate placing bids on the seller’s behalf to artificially drive up the price of his auctioned item.

(Wang, W., Hidvégi, Z. and Whinston, A., 2001)

In this report I will show what visualizations I built, demonstrating that great visualizations can lead to an easier way to quickly understand the information we have.

Before moving to the next section, I recall a description of the dataset:



Data Set Characteristics:	Multivariate	**Number of Instances:**	6321	**Area:**	Computer
Attribute Characteristics:	N/A	**Number of Attributes:**	13	**Date Donated**	2020-03-10
Associated Tasks:	Classification, Clustering	**Missing Values?**	N/A	**Number of Web Hits:**	23974

UCI Dataset description

Attribute information:

- Record ID – Unique identifier of a record in the dataset.
- Auction ID – Unique identifier of an auction.
- Bidder ID – Unique identifier of a bidder.
- Bidder Tendency – A shill bidder participates exclusively in auctions of few sellers rather than a diversified lot. This is a collusive act involving the fraudulent seller and an accomplice.
- Bidding Ratio – A shill bidder participates more frequently to raise the auction price and attract higher bids from legitimate participants.
- Successive Outbidding – A shill bidder successively outbids himself even though he is the current winner to increase the price gradually with small consecutive increments.
- Last Bidding – A shill bidder becomes inactive at the last stage of the auction (more than 90\% of the auction duration) to avoid winning the auction.
- Auction Bids – Auctions with SB activities tend to have a much higher number of bids than the average of bids in concurrent auctions.
- Auction Starting Price – a shill bidder usually offers a small starting price to attract legitimate bidders into the auction.
- Early Bidding – A shill bidder tends to bid pretty early in the auction (less than 25\% of the auction duration) to get the attention of auction users.
- Winning Ratio – A shill bidder competes in many auctions but hardly wins any auctions.
- Auction Duration – How long an auction lasted.
- Class – 0 for normal behaviour bidding; 1 for otherwise.

Dataset can be found at: UCI Machine Learning Repository: Shill Bidding Dataset Data Set

Libraries used are:

- Pandas
- Matplotlib
- Sklearn
- Plotly
- Seaborn
- Numpy
- Statsmodels
- Streamlit

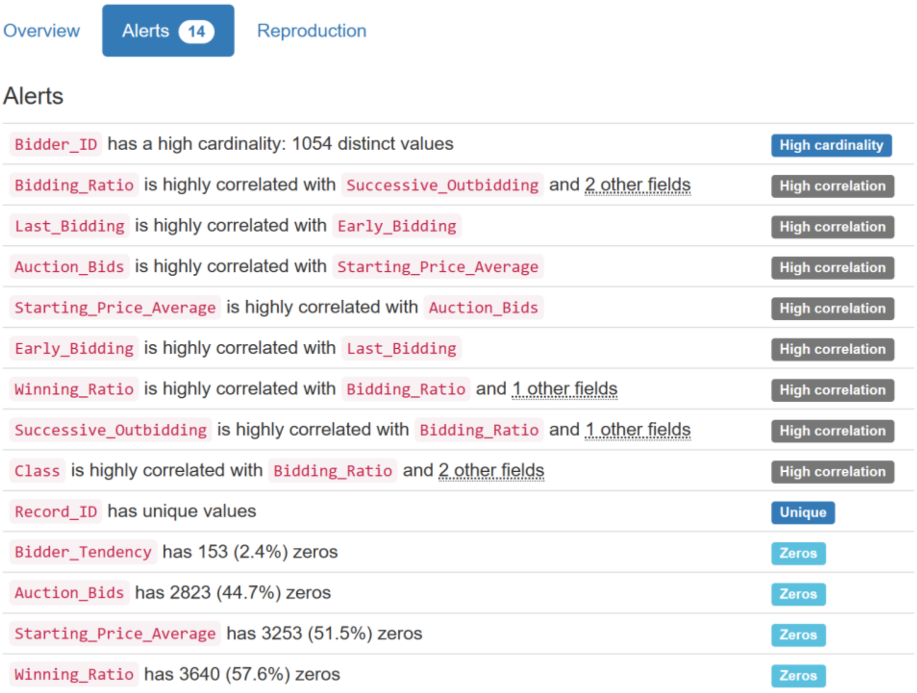
Rationale of data visualization techniques

Data visualization is useful for data cleaning, exploring data structure, detecting outliers and unusual groups, identifying trends and clusters, spotting local patterns, evaluating modeling output, and presenting results.

(Unwin A., 2020)

Profile report

To decide which visualizations I had to build, the very first thing I did was a profile report. The alerts gave my insights about the kind of plot I wanted to build:



Alerts

Bidder_ID has a high cardinality: 1054 distinct values	High cardinality
Bidding_Ratio is highly correlated with Successive_Outbidding and 2 other fields	High correlation
Last_Bidding is highly correlated with Early_Bidding	High correlation
Auction_Bids is highly correlated with Starting_Price_Average	High correlation
Starting_Price_Average is highly correlated with Auction_Bids	High correlation
Early_Bidding is highly correlated with Last_Bidding	High correlation
Winning_Ratio is highly correlated with Bidding_Ratio and 1 other fields	High correlation
Successive_Outbidding is highly correlated with Bidding_Ratio and 1 other fields	High correlation
Class is highly correlated with Bidding_Ratio and 2 other fields	High correlation
Record_ID has unique values	Unique
Bidder_Tendency has 153 (2.4%) zeros	Zeros
Auction_Bids has 2823 (44.7%) zeros	Zeros
Starting_Price_Average has 3253 (51.5%) zeros	Zeros
Winning_Ratio has 3640 (57.6%) zeros	Zeros

Profile report of the dataset

Number of Bids and Shell Bids by Acution Duration

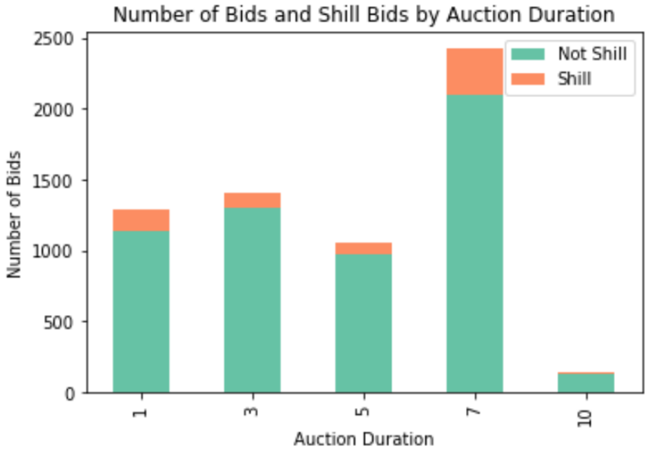
I had this assumption that we would find many more anomalous bids when looking at auctions that would last longer in time. My assumption was that, for shell bidders to go unnoticed, they would distribute successive outbidding across longer periods of time.

For this specific visualization I decided to use stacked bars to showcase number of normal and anomalous bids divided by auction duration.

"If you have to talk about both the overall total and the subcomponent breakdown, but your main goal is to focus on the total length of the bars, stacked bars could work."

(storytelling with data, 2022)

Below you can see the result I obtained:

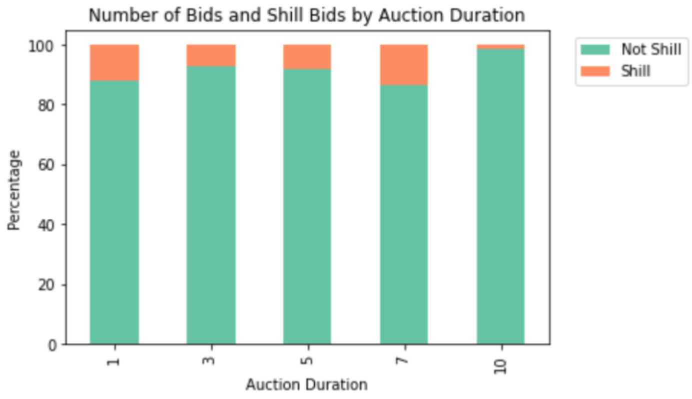


"Number of Bids and Shill Bids by Auction Duration" stacked bar

"Could work" is the tricky part about previous statement. In fact, from the plot I see that we have unbalanced classes across auction durations, and this wasn't giving me an easy way to compare the bars. For this reason, I decided to work on 100% stacked bar chart variation:

"In this variation, the bars extend from 0 to 100%. The benefit is that there are two baselines. Unlike the original, you can now compare two categories across bars with ease because of the consistent starting point."

(storytelling with data, 2022)



"Number of Bids and Shill Bids by Auction Duration" stacked bar variant using percentage

From previous plot, we can easily see that my assumption was wrong. Also, legend was moved outside of the box given that no space was made available.

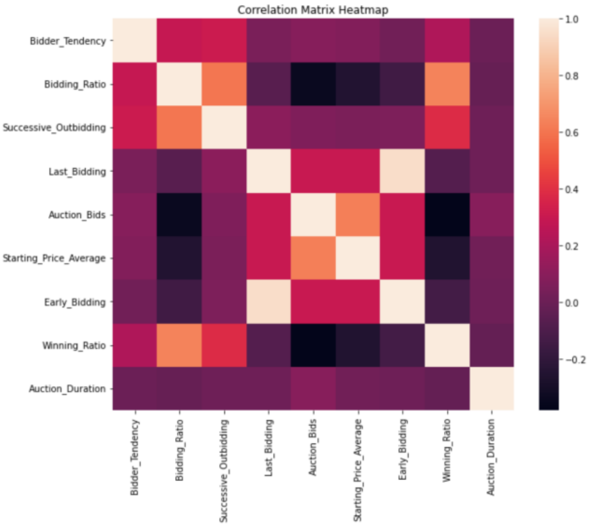
Correlation Matrix Heatmap

I decided to use this visualization because the profile report was showing many alerts regarding high correlation.

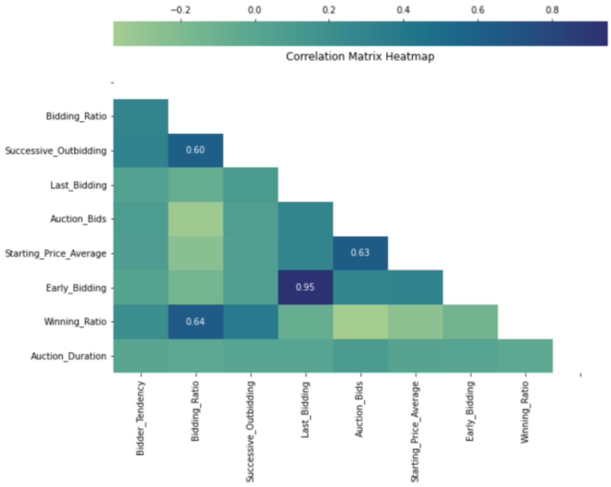
A correlation matrix shows the correlation between different variables in a matrix setting. However, because these matrices have so many numbers on them, they can be difficult to follow. Heatmap coloring of the matrix, where one color indicates a positive correlation, another indicates a negative correlation, and the shade indicates the strength of correlation, can make these matrices easier for the reader to understand.

(LOST)

I wanted to build an heatmap that not only was easy to read, but also with similar colours to the previous plot. Below the first version of the plot and the final one.



"Correlation Matrix Heatmap" simple plot



"Correlation Matrix Heatmap" optimized plot

Basically, I simplified the reading by:

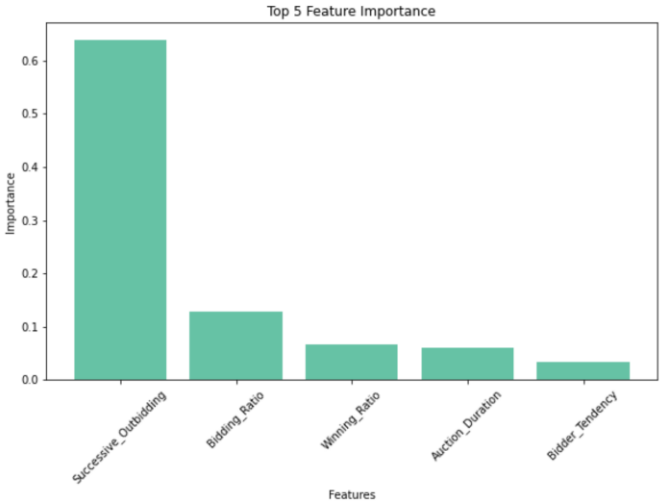
- Showing numbers only for correlations greater than 0.5
- Converted it into a triangle to avoid info repetition.

Top 5 Features Importance

I also wanted to show what were most important features in the classification of the bid. In this case, a random forest classifier was used for this purpose:

"Feature importance can be measured using a number of different techniques, but one of the most popular is the random forest classifier. Using Random forest algorithm, the feature importance can be measured as the average impurity decrease computed from all decision trees in the forest."

(Data Analytics, 2020)



"Top 5 Features Importance" using RF

Even though this was any easy visualisation, it needed the development of a ML model and provides very powerful info about what are the most important features in the determination of the bid class.

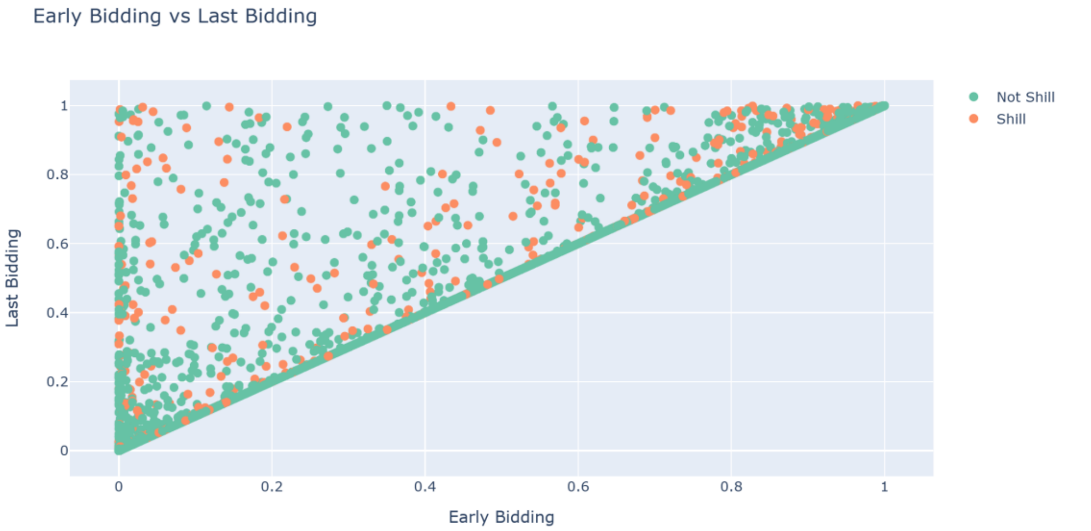
Early Bidding vs Last Bidding

From the heatmap I could see a high correlation between the two features in the title. For this reason, I wanted to better understand what kind of correlation is happening between them by building an interactive scatter visualization:

"Scatter plots are excellent charts for showing a relationship between two numerical variables across a number of unique observations. We see them in business communications from time to time, although they're much more commonly used in the "exploration" part of the process—when we're still trying to understand our data and find the important insights."

(storytelling with data, 2022)

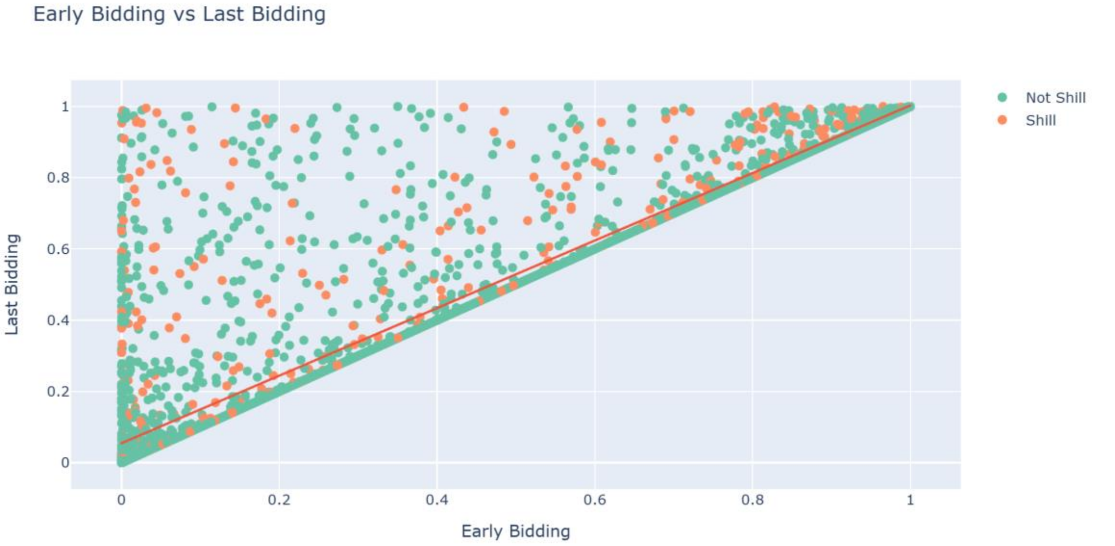
First result is shown below:



"Early Bidding vs Last Bidding" simple plot

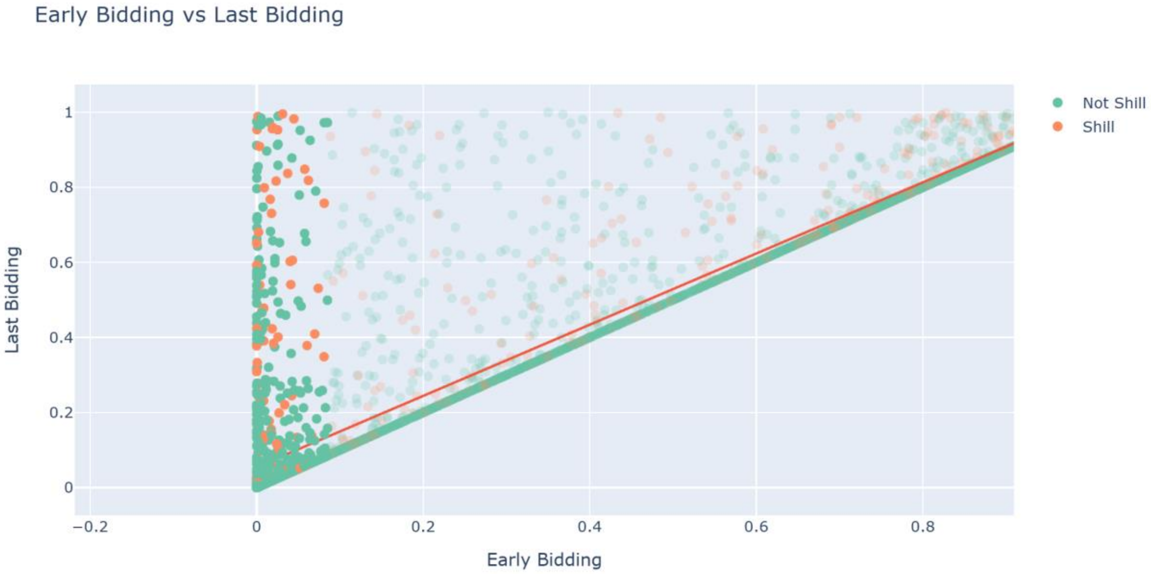
From the graph it seems that there is a strong linear correlation with an imaginary line connecting points (0,0) and (1,1). This is a sign of positive linear correlation, telling us that an increase of early bidding is also a sign of an increasing last bidding.

Given that many plots are accumulated towards the extremes previously quoted, I also wanted to make sure that those were a small portion with respects to the ones accumulated across the imaginary line. To prove this, I computed the trendline equation performing an ordinary least squares (OLS) regression. The line was then plotted on the graph:



"Early Bidding vs Last Bidding" optimized plot with OLS

The fact the I have a quota different from 0 means that there is a higher number of situations where early bidding is null but it’s not the case for last bidding:



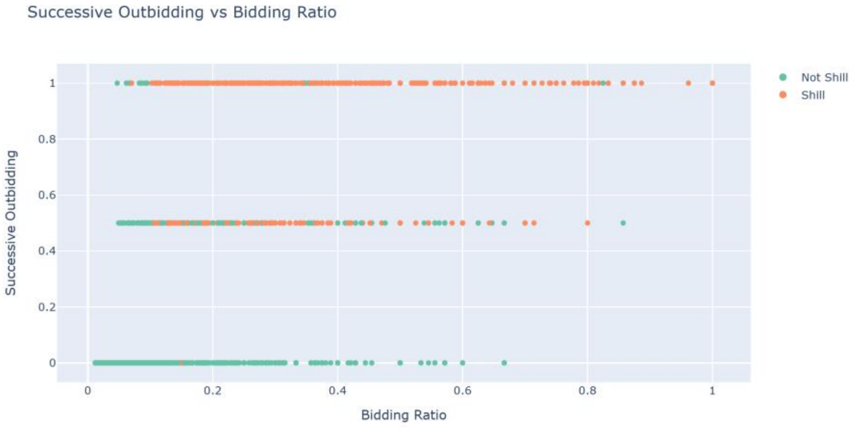
“Early Bidding vs Last Bidding” optimized plot focusing on low “Early Bidding”

A third dimension is also plotted on the graph, and this is the class of the bid where two different colours were used for a better visualization. Unfortunately, no insights could be obtained on how those classes relate to the two features analysed.

Successive Outbidding vs Bidding Ratio

The aim of this visualization was to see if the presence of successive bidding (discrete value) may relate to a higher or lower bidding ratio. Again, I also added a third dimension to show the class of points in the scatter plot.

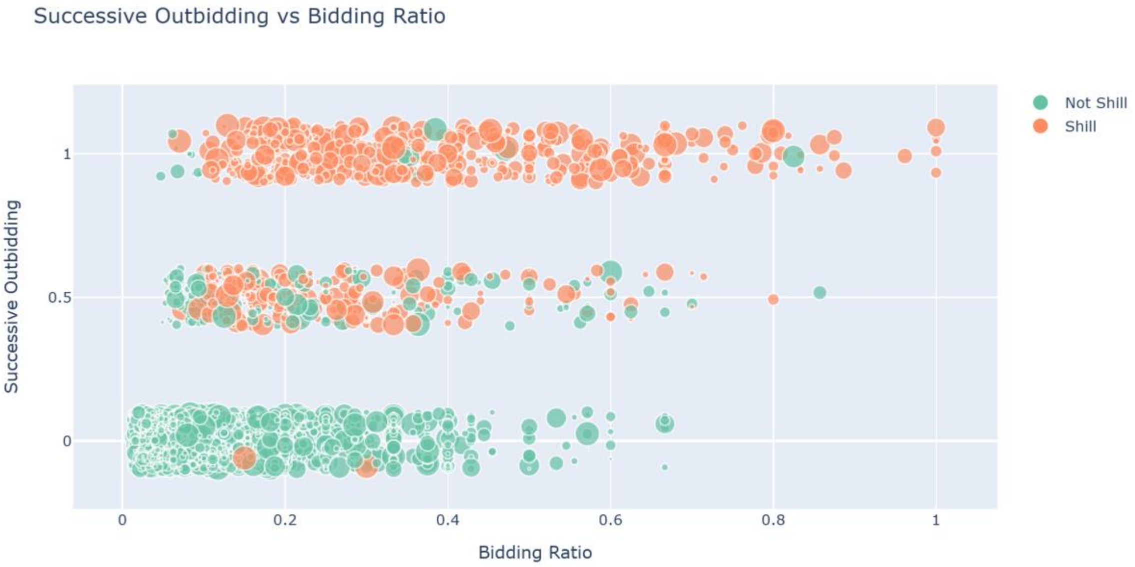
Below is the result of the first attempt:



“Successive Outbidding vs Bidding Ratio” simple plot

It is very clear that visualization doesn't look good, for this reason I improved it by:

- Increasing the size of the points based on 'Bidder Tendency'.
- Creating some noise to vertically distribute points and avoid too much overlap.



"Successive Outbidding vs Bidding Ratio" optimized plot randomizing vertical distribution

Insights obtained are:

- The higher the 'Successive Outbidding', the higher the probability to have a Shill Bid. This also confirms the fact that random forest placed 'Successive Outbidding' as the most important feature.
- The higher the 'Bidding Ratio', the higher the 'Successive Outbidding'. In fact, the heatmap previously showed us a 0.6 correlation between those variables. Note how there is no bidder with a high 'Bidding Ratio' but low 'Successive Outbidding'.

3D bids clustering using PCA

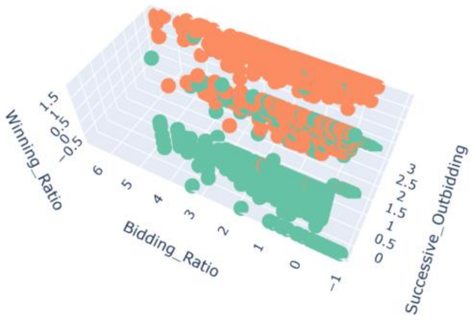
For this visualization, my initial idea was to demonstrate that we could distinguish normal and anomalous bids analysing features at our disposal.

For this purpose, I wanted to build a 3D visualization but I could only take into consideration 5 features for the plot:

- 3 axes, 1 feature for each.
- Point dimension, 1 feature.
- Point colour, 1 feature.

Given the constraints of the dimensions I could plot, I decided to only take into consideration top 4 important features from RF and the bid class.

Below a screenshot of the scatterplot result using above considerations:



3D scatterplot using most important features

As we can see, results are not that great. This is because:

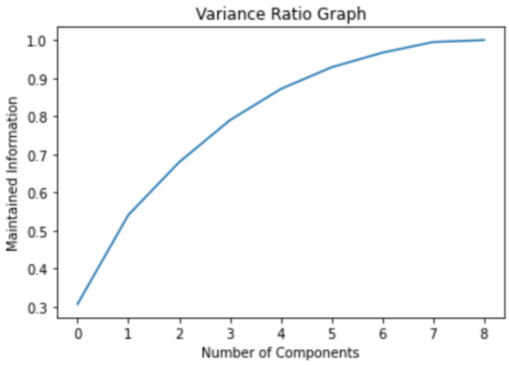
- "Successive_Outbidding" is a discreate feature, meaning that it only can assume 3 different values.
- We couldn't make use of the feature "Auction_Duration" because the scaling made it assume negative values that the plot doesn't accept.

To obtain better results I wanted a way to reduce the dimension of features (excluding class) to only the 3-axis dimension. PCA did the trick:

"The Principal Component Analysis is a popular unsupervised learning technique for reducing the dimensionality of data. It increases interpretability yet, at the same time, it minimizes information loss. It helps to find the most significant features in a dataset and makes the data easy for plotting in 2D and 3D."

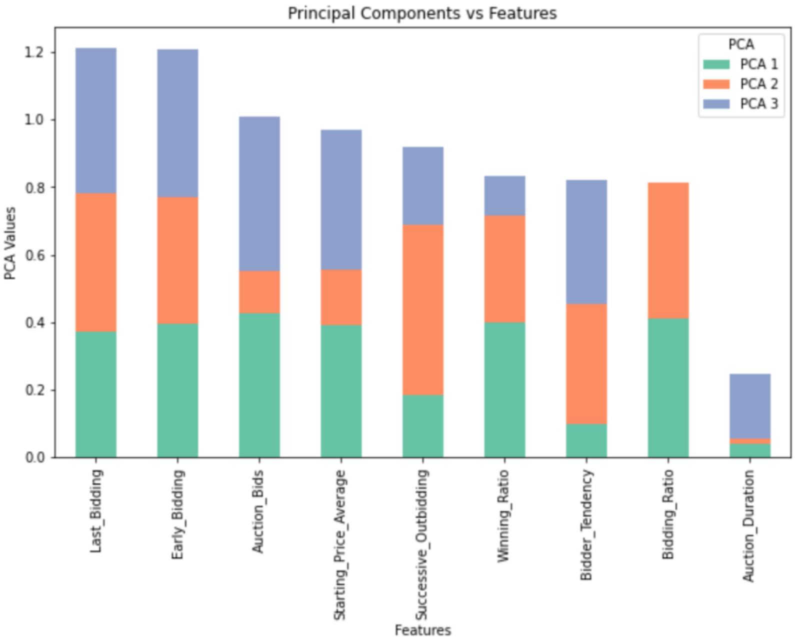
(Simplilearn.com, 2023)

After performing PCA on all features, I plotted the explained variance ratio graph to realize that by reducing the dataset to only 3 dimensions I would still be able to keep about 67% of the original information:



Variance Radio Graph representing information loss of PCA

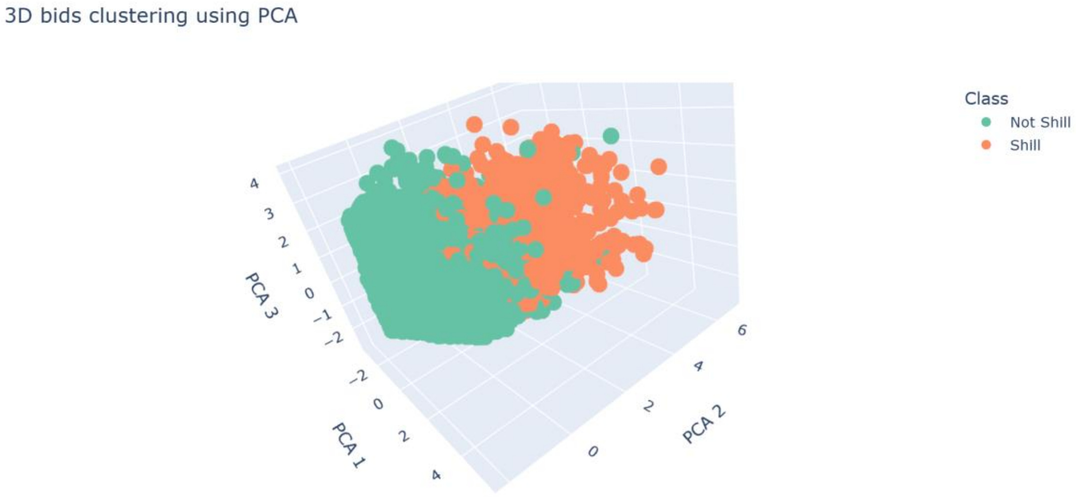
Being happy with the 67%, I proceeded with the dimensionality reduction and built a graph showing the contribution of each original feature for the 3 PCAs built:



Stacked bar of features weights on PCAs

It's important not to confuse this graph with the previously built most important features. One tries to keep as many information as possible of the features, the other tries to identify the most important features for the wanted prediction.

Below, result of the final plot:



"3D bids clustering using PCA" showing class clustering

It is clearly visible that, excepts for some points, by using the interactive visualization we see two distinguishable clusters for the classes.

Possible reasons of not having the two clusters perfectly separated are:

- We lost about 30% of the original features information applying PCA.
- We didn't manipulate and clean the data apart from the scaling.

This shows us that the features we have do allow us to create two separate clusters and, eventually, a machine learning model to predict shill bids.

Conclusion

To create appealing visualizations, you can notice that I mainly used two colours: green and orange.

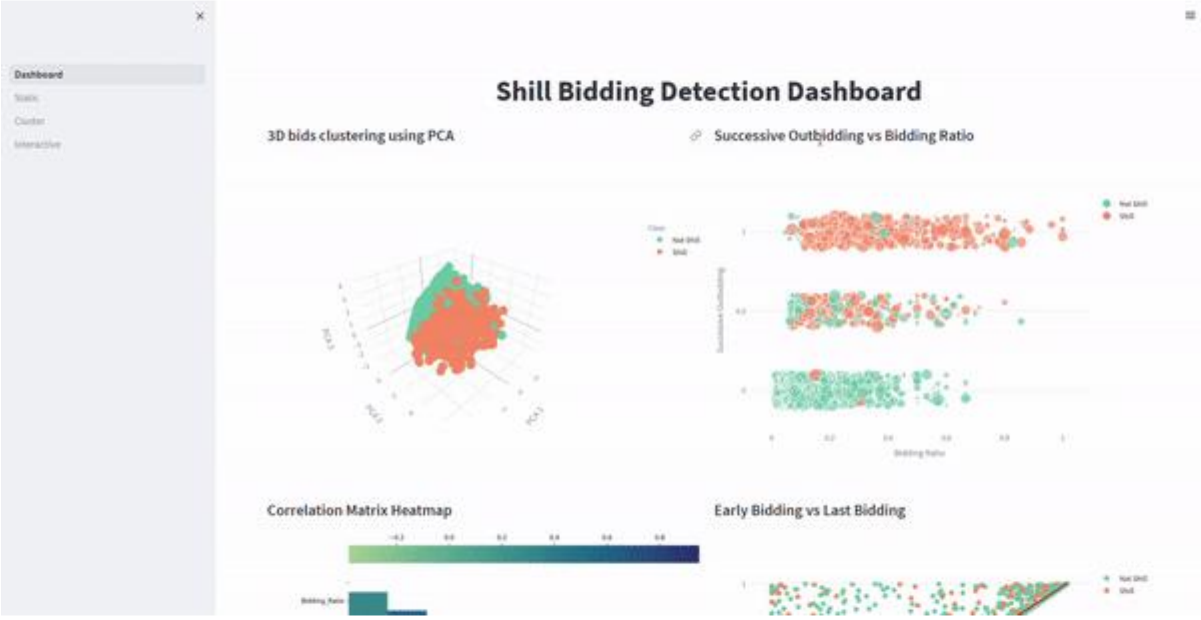
I decided to use them because they are easily differentiable by most people and create a harmonious high contrast to distinguish between the categories, which is essential for all visualizations implemented.

For a final outlook of the work, I built a dashboard using Streamlit. The choice of this library comes from the fact that I already had some basic knowledge on how to use it and is very easy to deploy as a web app to share content broadly.

The dashboard is made of different pages:

- Dashboard -> to collect what I believe are most important visualizations.
- Static -> to represent static visualizations.
- Cluster -> to show on a larger screen the 3D interactive cluster visualization.
- Interactive -> to visualize all interactive plots.

Below a GIF of the submitted video "dasboard_video":



Shill Bidding Detection Dashboard

3D bids clustering using PCA

Successive Outbidding vs Bidding Ratio

Correlation Matrix Heatmap

Early Bidding vs Last Bidding

GIF showing dashboard built using Streamlit library

Total word count: 1351

Referencing

Wang, W., Hidvégi, Z. and Whinston, A. (2001). Shill Bidding in English Auctions. [online] Available at: https://oz.stern.nyu.edu/seminar/fa01/1108.pdf [Accessed 27 May 2023].

Unwin, A. (2020). Why Is Data Visualization Important? What Is Important in Data Visualization? Harvard Data Science Review. [online] Available at: https://doi.org/10.1162/99608f92.8ae4d525 [Accessed 27 May 2023].

storytelling with data. (2022). what is a stacked bar chart? [online] Available at: https://www.storytellingwithdata.com/blog/stacked-bars [Accessed 27 May 2023].

LOST. (n.d.). Heatmap Colored Correlation Matrix. [online] Available at: https://lost-stats.github.io/Presentation/Figures/heatmap_colored_correlation_matrix.html [Accessed 27 May 2023].

Data Analytics. (2020). Feature Importance using Random Forest Classifier - Python. [online] Available at: https://vitalflux.com/feature-importance-random-forest-classifier-python/ [Accessed 27 May 2023].

storytelling with data. (2022). how to make a scatter plot in Excel. [online] Available at: https://www.storytellingwithdata.com/blog/how-to-make-a-scatter-plot-in-excel [Accessed 27 May 2023].

Simplilearn.com. (2023). Principal Component Analysis in Machine Learning | Simplilearn. [online] Available at: https://www.simplilearn.com/tutorials/machine-learning-tutorial/principal-component-analysis#:~:text=The%20Principal%20Component%20Analysis%20is [Accessed 27 May 2023].