CCT College Dublin

Assessment Cover Page

Module Title:	Data Visualisation Techniques
Assessment Title:	CA 2 - Individual
Lecturer Name:	David McQuaid
Student Full Name:	Victor Ferreira Silva
Student Number:	2021324
Assessment Due Date:	28th May 2023
Date of Submission:	29th May 2023

Declaration

Contents

Assessment Task

Victor Ferreira Silva

May 29, 2023

1 Introduction

The ability to predict normal and abnormal bidding behaviour of eBay users can help companies identify scams and other undesirable users on the platform. This work focuses on the Shill Bidding Data set (SBD) and aims to find insights and detect patterns that can help subsequent phases of analyses that can mitigate such behaviours. In turn, the SBD aims to facilitate academic research and support the advancement of machine learning techniques by providing a realistic and preprocessed auction data set (Alzahrani and Sadaoui 2018) and it consists of eBay auctions that have various features, including auction duration, bidder tendency and class.

This project demonstrates how the libraries *Plotly* and *Dash* can be used to provide data visualisation solutions to support general data analysis. These libraries can not only provide elementary data visualisation solutions but also interactive plots and a dashboard external to *Jupyter Notebook*.

2 Imports & Configurations

For this study, a *Jupyter Notebook* was created and some libraries and modules were imported for data analysis, clustering and visualisation. Firstly, the libraries *Pandas* and *Numpy* were selected for data manipulation and analysis and numerical operations. Secondly, the modules *MinMaxScaler* and *KMeans* were imported from *sklearn* library, respectively for data preprocessing and clustering tasks. Thirdly, two *Plotly* modules were added for creating and customising interactive visualisations.

Finally, for the dashboard, *JupyterDash* was installed for creating *Dash* applications and the modules *dcc* and *html* were added for constructing the layout and components of the Dash Application.

3 Data Understanding

The description of the data set columns can be visualised in Table 1 and neither duplicates nor missing data were detected in this data set. Also, it was observed that the data set is considerably imbalanced, with approximately 89% of Normal bids and only 11% Abnormal.

4 Insights

In the following subsections, three visualisations were implemented to reveal patterns and variations in bidding behaviour, supporting the identification of potential fraudulent activity.

To ensure all variables are on a consistent scale, a *MinMaxScaler* was used to scale all features, but the identifiers and the class. This action allows a fair a better comparison and interpretation of variable relationships and avoids distorted or skewed visualisations.

4.0.1 Distribution of Bids Per Bidder

Analysing the distribution of bids per bidder is useful for detecting shill bidding in online auctions, as outliers can serve as a starting point for further investigation into potential shill bidding practices.

Given this, the first visualisation presented in this work is a box plot that illustrates the distribution of bids among the 1054 distinct bidders identified using the *Bidder_ID* feature (Figure 1), which serves

Table 1: Description of Shill Bidding Data set Columns

Column	**Pandas.Dtype**	**Description**
Record ID	int64	Unique identifier of a record in the data set.
Auction ID	int64	Unique identifier of an auction.
Bidder ID	object	Unique identifier of a bidder.
Bidder Tendency	float64	Shill bidder's preference for auctions of few sellers.
Bidding Ratio	float64	Shill bidder's frequency of participation and raising the price.
Successive Outbidding	float64	Shill bidder's tactic of outbidding themselves gradually.
Last Bidding	float64	Shill bidder's inactivity in the last stage of the auction.
Auction Bids	float64	High number of bids in auctions with shill bidding activities.
Auction Starting Price	float64	Shill bidder's small starting price to attract bidders.
Early Bidding	float64	Shill bidder's tendency to bid early in the auction.
Winning Ratio	float64	Shill bidder's low success rate in winning auctions.
Auction Duration	int64	Duration of the auction.
Class	int64	0 for normal behaviour, 1 for shill bidding.

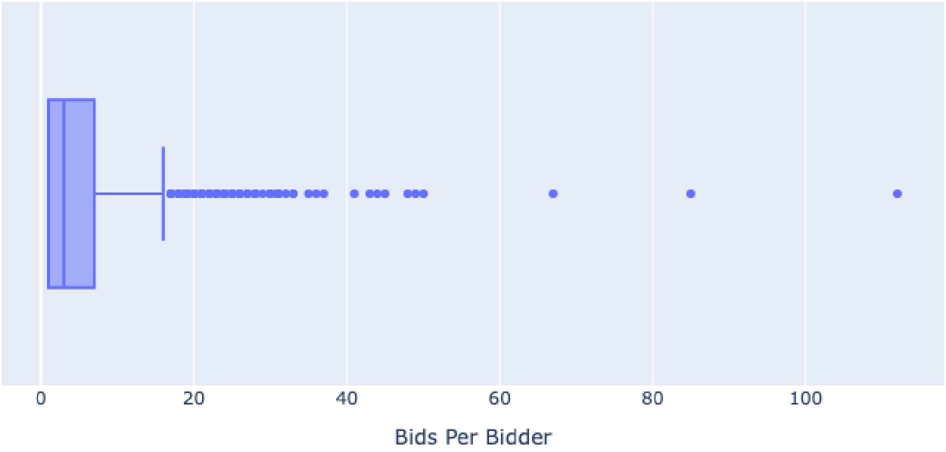


Figure 1: Distribution of Bids Per Bidder, illustrating the variability in the number of bids made by individual bidders

as a unique identifier for each bidder. Also, the feature *Record_ID* can be used to count the amount of distinct bids for each bidder.

As indicated, three quarters of bidders made 7 bids or fewer, with some outliers making up to 112 bids. In contrast, it is worth noting that the most frequently occurring number of bids per bidder was 1.

4.0.2 Distribution of Bids Per Auction

By analysing this distribution pattern, it is possible to identify an abnormally high amount of bids in a specific auction, which may indicate fraudulent activity. Thus, detecting such anomalies helps in identifying potential instances of shill bidding, where fake bids are placed to artificially inflate prices.

Considering that, the histogram visible in Figure 2 shows the distribution of bids among the 807 distinct auctions identified using the *Auction_ID* feature, which serves as a unique identifier for each auction, and *Record_ID* to count the number of distinct bids for each auction. Besides that, half of the auctions received 7 bids or fewer, with a few outliers receiving 26 bids. In the other hand, it is worth mentioning that the most frequently occurring value - or the mode - was 5.

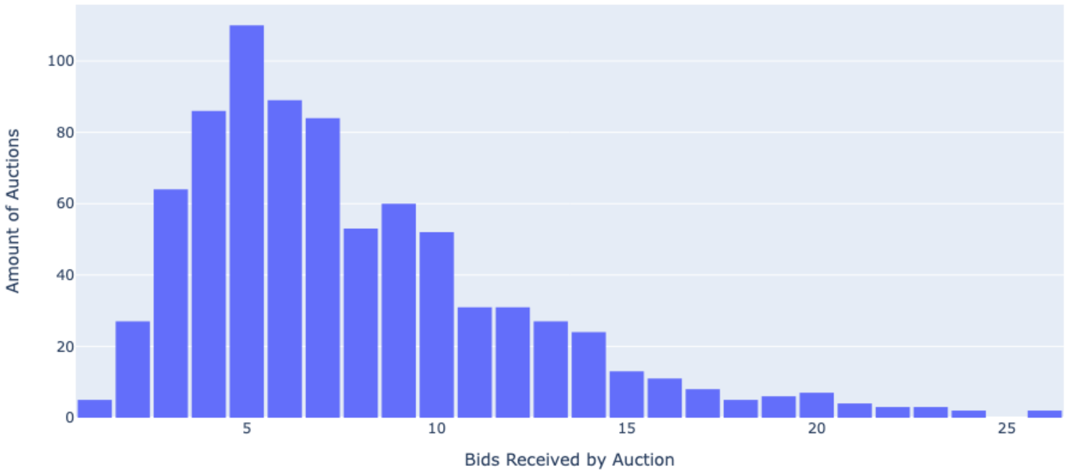


Figure 2: Distribution of Bids per Auction, demonstrating the frequency of bids received by each auction, with bins representing the count of bids

4.0.3 Distribution of Bidding Ratio By Class

This analysis aims to highlight the differences in bidding patterns between abnormal and normal bids. Here, the ridge line plots provide insights into the distribution of bidding ratios for both abnormal and normal classes.

As can be visualised in Figure 3, the range of bidding ratios for abnormal bids is considerably wider, indicating a higher variability compared to normal bids. Also, the plot suggests that abnormal bids have a higher bidding ratio, as the mean line for abnormal bids is approximately 0.34 while the mean line for normal bids is around 0.10. Additionally, the median bidding ratio is higher for abnormal bids (0.29) in comparison to normal bids (0.07).

Ultimately, although there is some overlap between the distributions, the ridge line plots show that abnormal bids tend to have higher bidding rations on average in comparison to normal bids.

5 Clustering Visualisation

K-Means Clustering was selected for this section as it groups data points based on their similarity in a multi-dimensional feature space. In this work, four features were provided to the algorithm, namely *Bidder_Tendency*, *Bidding_Ratio*, *Winning_Ratio*, and *Successive_Outbidding* and after some analysis, three clusters were defined. The distribution of these clusters in a 3D scatter plot can be visualised in Figure 4. In this 3D scatter plot, the three axes are represented by three features *Successive_Outbidding*, *Bidding_Ratio*, *Winning_Ratio*.

6 Interactive Plots

In the following subsections, the two interactive plots created for this work will be demonstrated. They should empower exploratory data analysis by providing tools to make this activity more dynamic and effective.

6.0.1 Interactive 2D Histogram Contour Plot

An interactive 2D contour plot was implemented using Plotly and it can be visualised in Figure 5. After selecting two variables from the drop-down menus in this plot, the relationship between these

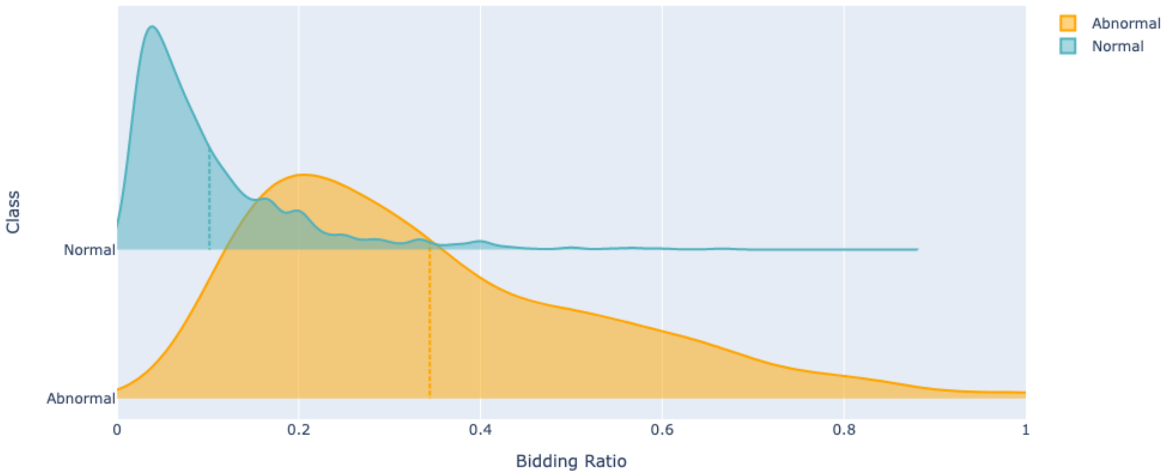


Figure 3: Distribution of Bidding Ratio by Class, illustrating the comparison between abnormal and normal bidding behaviour

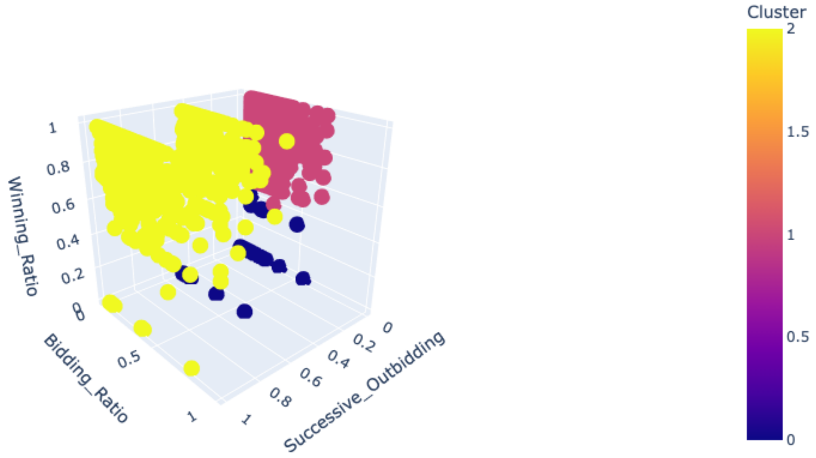


Figure 4: Visualisation of K-means Cluster Analysis in a 3D Space, illustrating the grouping patterns and relationships between successive outbidding, bidding ratio, winning ratio, and cluster assignments

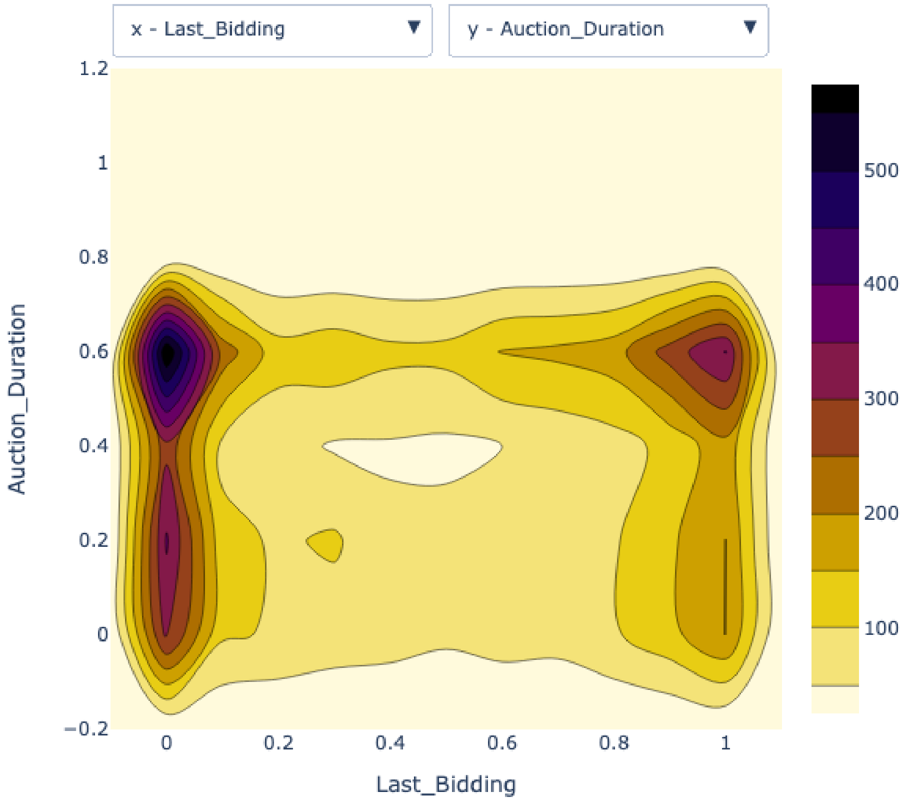


Figure 5: Interactive Contour Plot showcasing the relationship between Last Bidding and Auction Duration

two continuous variables can be visualised as the contours of their joint probability density.

Besides that, the darker colours represent the denser regions of the plot, while the data points are more sparse in the lighter regions. Thus, this interactive plot provides insights into the distribution and density of data points in the two-dimensional feature space.

6.0.2 Interactive 3D Scatter Plot

An interactive 3D scatter plot was implemented using Plotly and it is visible in Figure 6, that is, a three-dimensional plot allows the visualisation of relationships and patterns among three distinct feature simultaneously.

Regarding the data points colours, using orange for abnormal and blue for normal provides clear visibility and differentiation. Besides that, the former symbolises caution and aligns with potential fraud, while the latter is often associated to trust and stability.

As a result, this plot provides flexibility in analysing combinations of variables and it aids in identifying and understanding how abnormal and normal data points are distributed within a three-dimensional space. Ultimately, users can rotate or zoom the plot to visualise it from many angles, obtaining new insights and perspectives and identifying clusters, outliers, trends, and patterns to support data-driven decision making.

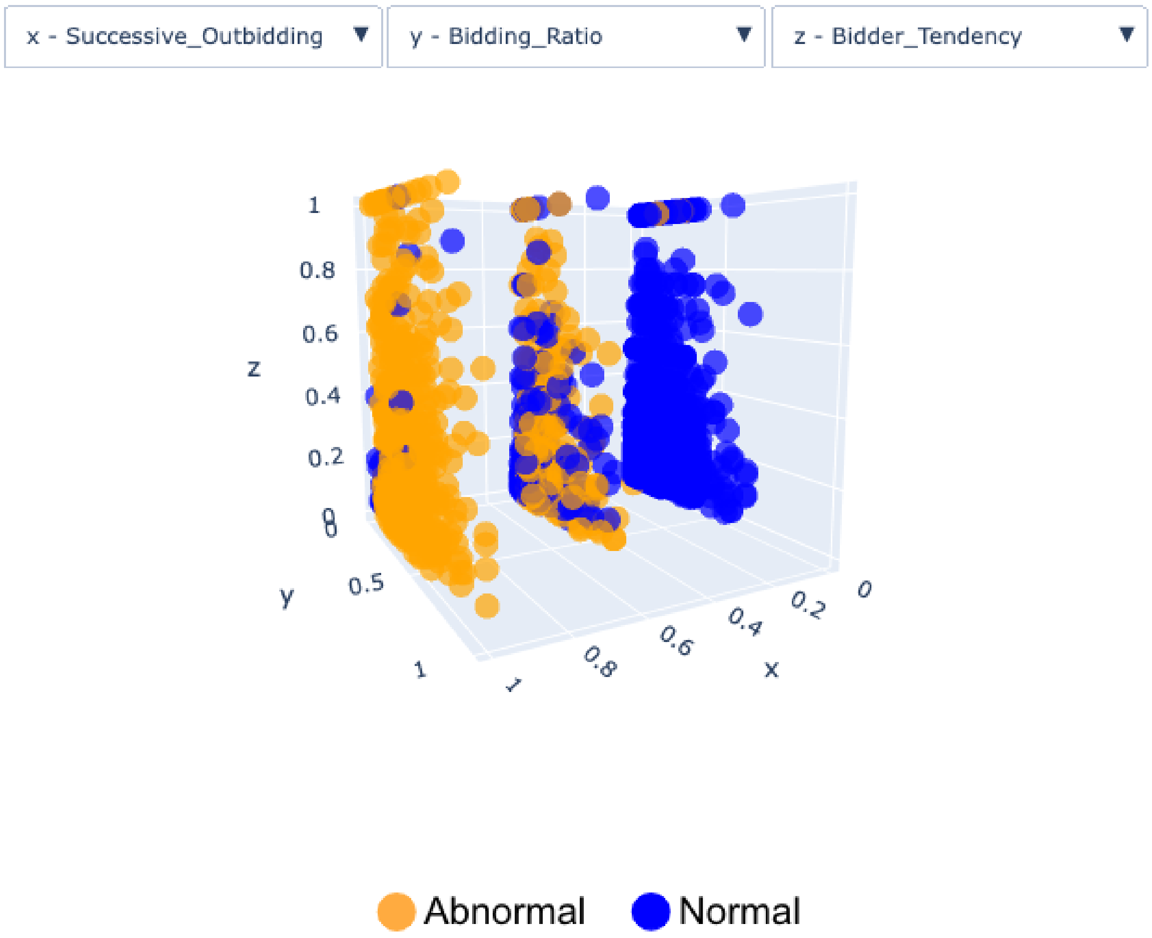


Figure 6: 3D Scatter Plot displaying bid-related features with adjustable dimensions and class markers.

7 Dashboard

A dashboard featuring the interactive plots discussed earlier in this project was implemented using *Dash* library, executed from a Jupyter Notebook and deployed in a localhost. A screenshot of this Dash app can be visualised in Figure 7.

Using Dash in this work allowed a relatively simple integration with the previous plots developed with Plotly library, as Dash was created by the parent company Plotly. Besides that, Dash provides flexibility in designing and customising the dashboard layout, allowing the arrangement of multiple interactive plots and components.

Given this, a Dash application for a dashboard to analyse correlations and patterns in SBD was set up. It consists of layout with a title and two interactive plots. Also, the implementation contained two callback functions: *update_contour_plot* and *update_scatter_plot*, which are triggered when the contour plot and scatter plot figures are updated, respectively.

Finally, this application runs in an external server mode, thus the dashboard is accessible through a web browser, allowing users to interact with the plots displayed in this application, explore the data and visualise bidding patterns.

8 Conclusion

This project successfully applied data visualisation techniques and clustering algorithms to analyse the proposed data set. Firstly, the plots provided valuable insights into the distribution of bids per bidder and auction, and the bidding ratio by class. Moreover, the interactive plots enhanced the exploration and understanding of the data. Finally, the integration of the interactive plots into a dashboard provided a comprehensive interface for further analysis. Overall, this work demonstrated the effectiveness of Dash and Plotly in creating effective visualisations for bidding prediction and anomaly detection.

References

Alzahrani, Ahmad and Samira Sadaoui (June 2018). *Scraping and preprocessing commercial auction data for fraud classification*. URL: https://arxiv.org/abs/1806.00656.

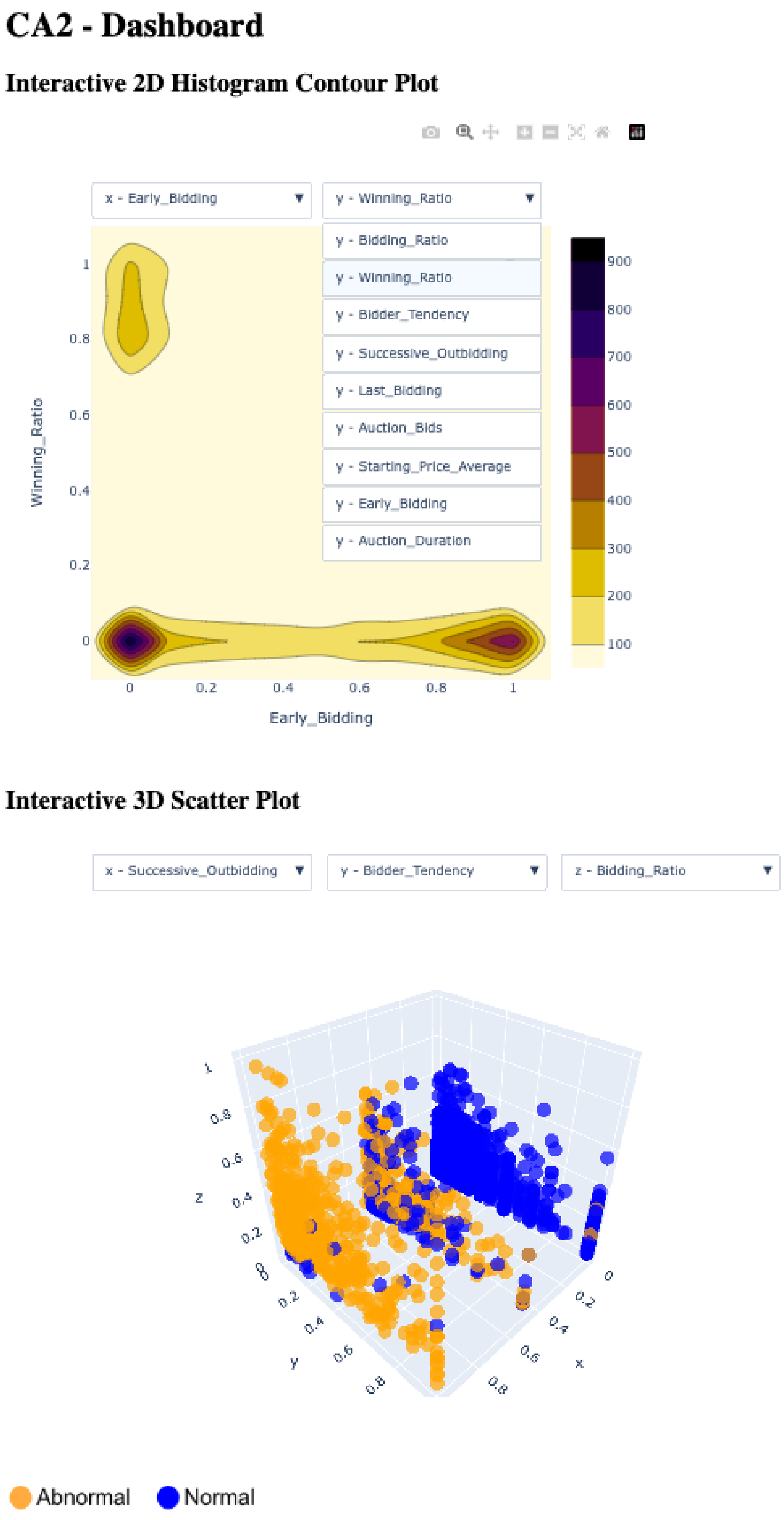
Pedregosa, F. et al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Plotly (2023a). *Dash Documentation & User Guide*. URL: https://dash.plotly.com/.

— (2023b). *Plotly/jupyter-dash: Develop dash apps in the Jupyter Notebook and JupyterLab*. accessed on May 2023. URL: https://github.com/plotly/jupyter-dash.

UCI (2018). *UCI Machine Learning Repository: Shill bidding dataset data set*. URL: https://archive.ics.uci.edu/ml/datasets/Shill+Bidding+Dataset.

Wilke, Claus O. (2020). *Fundamentals of data visualization a primer on making informative and compelling figures*. O'Reilly.



CA2 - Dashboard

Interactive 2D Histogram Contour Plot

x - Early_Bidding

y - Winning_Ratio

Interactive 3D Scatter Plot

x - Successive_Outbidding

y - Bidder_Tendency

z - Bidding_Ratio

Figure 7: Screenshot of a dashboard powered by Dash library, featuring an Interactive 2D Histogram Contour Plot and Interactive 3D Scatter Plot.