DATA SCIENCE TOOLBOX: PYTHON PROGRAMMING

PROJECT REPORT

(Project Semester January-April 2025)

Analysis of Shill Bidding

Submitted by

Name: Kaja Revanth Sri Narasimha

Registration No: 12312200

Programme and Section: Data Science, K23EU

Course Code: INT375

Under the Guidance of

Dr. Tanima Thakur (UID:23532)

Discipline of CSE/IT

Lovely School of Computer Science

Lovely Professional University, Phagwara

CERTIFICATE

This is to certify that Kaja Revanth Sri Narasimha bearing Registration no. 12312200 has completed INT375 project titled, "Analysis of Shill Bidding" under my guidance and supervision. To the best of my knowledge, the present work is the result of her original development, effort and study.

Signature and Name of the Supervisor
Designation of the Supervisor
School of Computer Science
Lovely Professional University
Phagwara, Punjab.

Date:

DECLARATION

I, Kaja Revanth Sri Narasimha, student of Data Science, under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 12-04-2025 Signature

Registration No.12312200 Kaja Revanth Sri Narasimha

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my project guide, Tanima Thakur, for their valuable guidance and support throughout this project, *" Analysis of Shill Bidding"*. I am thankful to the Department of Computer Science and Engineering, Lovely Professional University, for providing the necessary resources and environment. I alsoacknowledge the Government of India for making the dataset publicly available, enabling this research.

Name: Kaja Revanth Sri Narasimha

Registration number: 12312200

Table of Contents

INTRODUCTION

Online auction platforms such as eBay have transformed how goods and services are bought and sold in the digital age. These platforms enable users to participate in competitive bidding environments from virtually anywhere in the world, making transactions faster, broader in reach, and more efficient. However, with the rise of online auctions, a range of fraudulent activities has also emerged—among which **shill bidding** stands out as one of the most deceptive and damaging to both buyers and sellers.

To address this issue, data-driven techniques have become increasingly important. **Exploratory Data Analysis (EDA)** plays a critical role in the early stages of fraud detection. By visually and statistically examining key patterns in auction behavior—such as timing of bids, frequency, and success rates—EDA allows analysts and researchers to uncover insights that can differentiate between legitimate users and suspected shill bidders.

This report presents a comprehensive EDA on a publicly available **Shill Bidding dataset**, with the goal of identifying behavioral patterns associated with fraudulent bidding. The dataset includes features related to bidding frequency, timing, auction outcomes, and more. By analyzing these variables, we aim to:

- Understand the distribution of legitimate vs. fraudulent activities
- Identify trends in how shill bidders behave compared to genuine users
- Explore correlations between different auction features
- Extract actionable insights that can inform fraud detection models

This investigation lays the groundwork for developing more robust machine learning models that can be trained to recognize and flag potential shill bidding activities in real time. The results not only contribute to a better understanding of auction fraud but also serve as a guide for designing safer and more transparent online marketplaces.

Source of Dataset

The dataset used is titled **"Shill Bidding Dataset.csv"**, and it contains a variety of attributes relevant to user behavior in online auctions. Each record represents a user's activity in a particular auction, along with a class label indicating whether the activity was **legitimate (0)** or **shill bidding (1)**.

Key features include:

- **Auction_Duration:** Length of the auction in days.
- **Auction_Bids:** Total number of bids placed by the user.
- **Winning_Ratio:** User's success rate in auctions.
- **Early_Bidding:** Extent to which the user bids early in the auction.
- **Last_Bidding:** Extent to which the user bids toward the end of the auction.
- **Class:** Binary label indicating the nature of the bidder.

EDA Process (Exploratory Data Analysis Process)

Exploratory Data Analysis (EDA) is a vital step in understanding the dataset, uncovering patterns, and identifying potential issues. For this report, EDA is used to identify key differences between legitimate bidders (Class 0) and shill bidders (Class 1).

3.1 Dataset Overview

- **Records:** 6,321 total
- **Features:** 11 numerical features
- **Target:** Class (0 = Legitimate, 1 = Shill Bidding)

The dataset has no missing values, making it ready for analysis.

3.2 Class Distribution

The dataset shows a **high imbalance** with ~89.3% legitimate bidders and ~10.7% shill bidders. This skew must be addressed in future models, potentially using resampling or anomaly detection methods.

Visualization: A bar chart illustrating the class distribution shows the significant imbalance between the two classes.

3.3 Auction Bidding Behavior

- **Shill Bidders** tend to place **more bids** than legitimate users, possibly to manipulate the auction.
- **Legitimate Bidders** place fewer, more strategic bids.

Visualization: A bar chart comparing the average number of bids by class shows that shill bidders are more aggressive in bidding.

3.4 Early Bidding Patterns

- **Shill Bidders** often bid early to create a false sense of demand.
- **Legitimate Users** have a more evenly spread bidding pattern over time.

Visualization: A histogram shows that shill bidders have concentrated early bidding activity, while legitimate users' early bids are more distributed.

3.5 Last-Minute Bidding (Sniping)

- **Legitimate Users** tend to place bids towards the end of the auction.
- **Shill Bidders** show erratic late bidding behavior or avoid last-minute bidding.

Visualization: A histogram for last bidding behavior shows clear differences between legitimate and shill users.

ANALYSIS ON DATASET

4.1 Legitimate vs Shill Bidding Distribution

Overview:

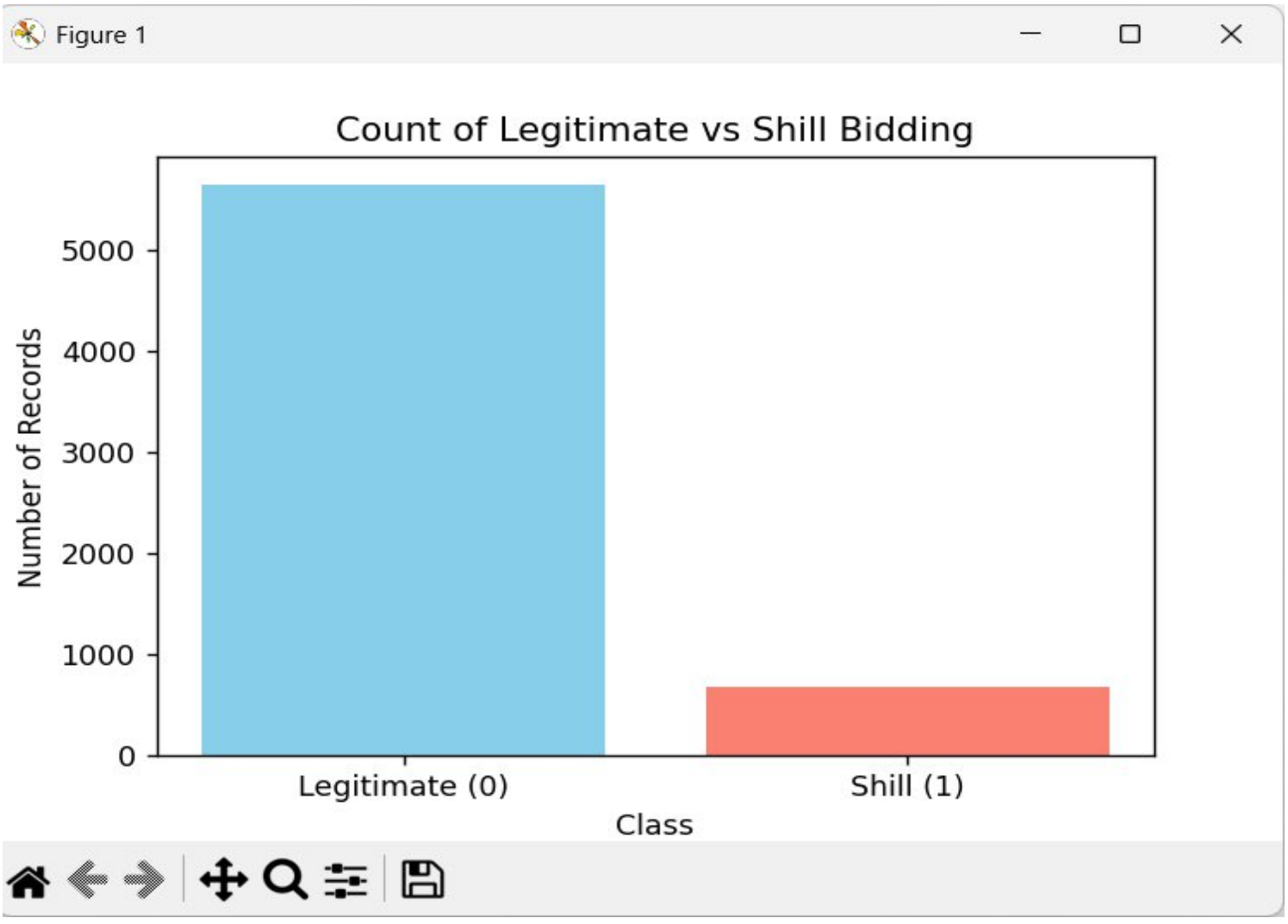
The first step in understanding the data is to assess how many records correspond to legitimate activity and how many are associated with shill bidding. This distribution provides insight into the **balance or imbalance** in the dataset, which is crucial for developing fair predictive models.

Insights:

It was observed that legitimate bidding accounts for a **significantly higher proportion** of the dataset compared to shill bidding. This imbalance suggests that fraudulent behavior is relatively rare, which aligns with real-world conditions. However, it also indicates that any machine learning approach will need to address **class imbalance** through techniques like resampling or anomaly detection.

Visualization:

A bar chart was used to display the count of each class side by side, making the imbalance clear and visually impactful.



4.2 Bidding Behavior by Class

Overview:

This section explores how often shill bidders participate in auctions compared to legitimate users. The number of bids placed by a participant can reveal aggressive or abnormal bidding behavior—often a red flag for fraud.

Insights:

The analysis showed that **shill bidders tend to place more bids** per auction on average than legitimate bidders. This may be an intentional tactic to drive up the auction price or maintain control over bidding progression. In contrast, legitimate bidders generally place fewer bids, possibly indicating a more conservative and strategic approach.

Visualization:

A horizontal bar chart was employed to compare the average bid count for both classes. The difference in bid volume was clearly distinguishable, reinforcing the idea that bidding intensity is a key indicator.

4.3 Early Bidding Pattern

Overview:

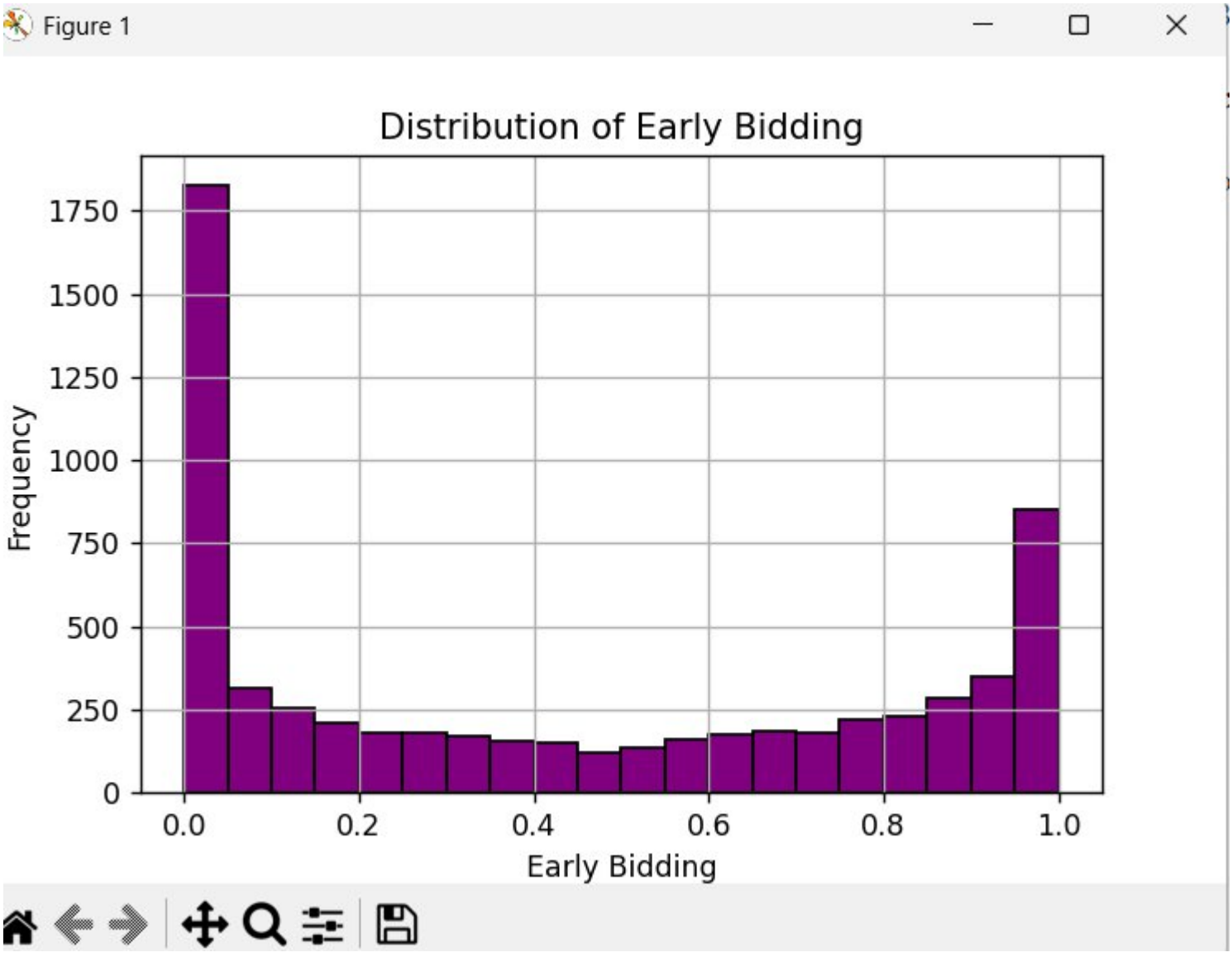
Timing plays a critical role in auction strategies. This section examines how frequently participants bid during the **early phase** of an auction, which can sometimes be used to create momentum or manipulate perception.

Insights:

The histogram of early bidding values revealed that shill bidders often show concentrated activity at specific early intervals. This behavior might be aimed at creating false excitement or misleading genuine users into believing the item is in high demand. On the other hand, legitimate bidders displayed a broader and more balanced distribution of early bids.

Visualization:

A histogram was used to represent the frequency distribution of early bidding across the dataset. The presence of peaks within certain value ranges suggests non-random behavior that may be indicative of fraudulent activity.



4.4 Relationship Between Early and Late Bidding

Overview:

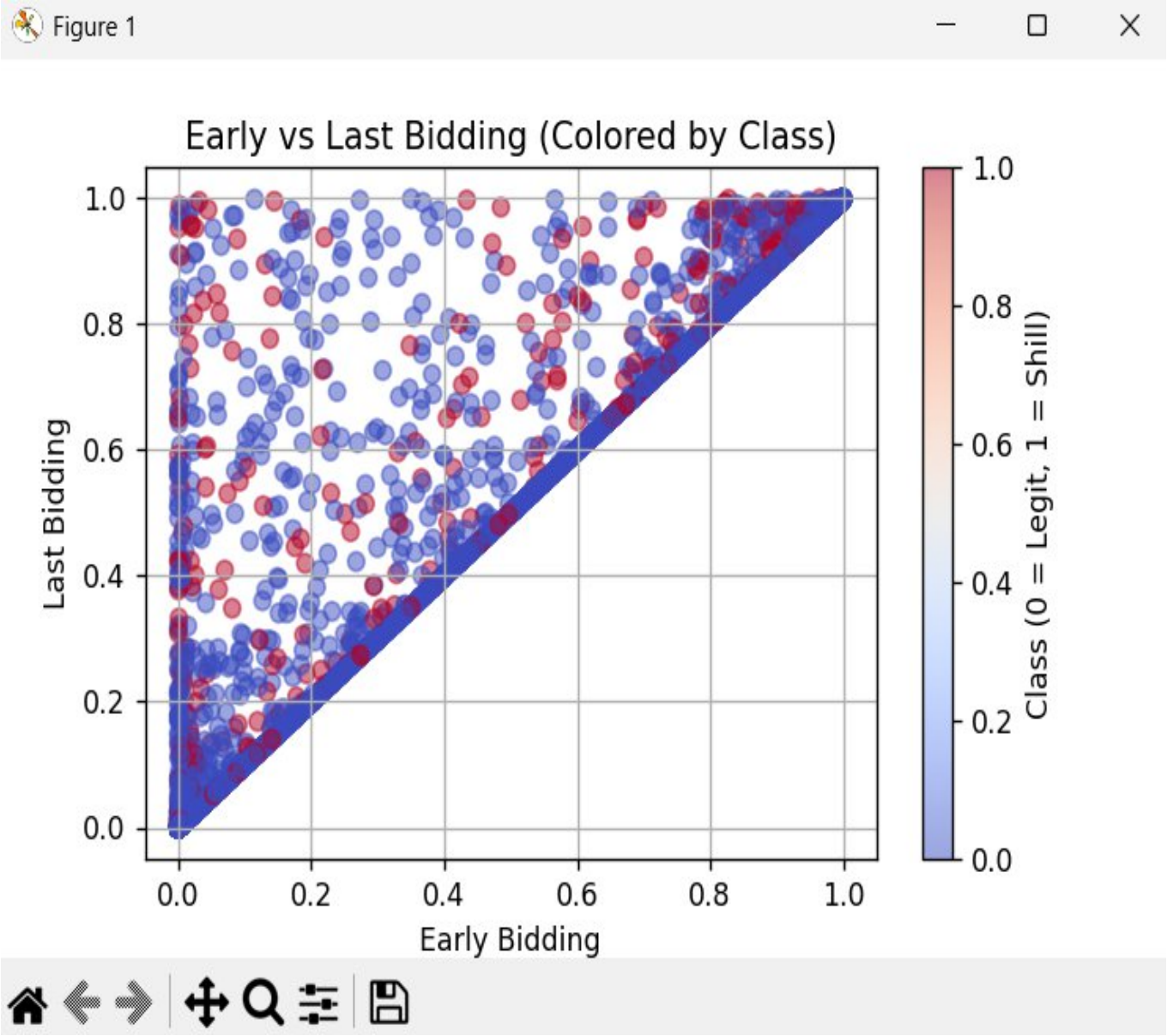
Analyzing the relationship between early and last-minute bidding provides a dynamic view of how users engage with the auction lifecycle. The idea is to see whether a user spreads out their bidding or shows suspicious concentration at either end.

Insights:

When early bidding was plotted against last bidding, two distinct patterns emerged. Legitimate bidders showed scattered behavior, participating throughout the auction. In contrast, shill bidders exhibited more **clustered or extreme** values—either heavily early, late, or both. This suggests a strategy of controlling key phases of the auction, either to set the price early or block competitors near the end.

Visualization:

A scatter plot with color-coded classes was used, helping distinguish between the two bidding styles and revealing the behavioral footprint of shill participants.



4.5 Winning Ratio Trends Over Auction Duration

Overview:

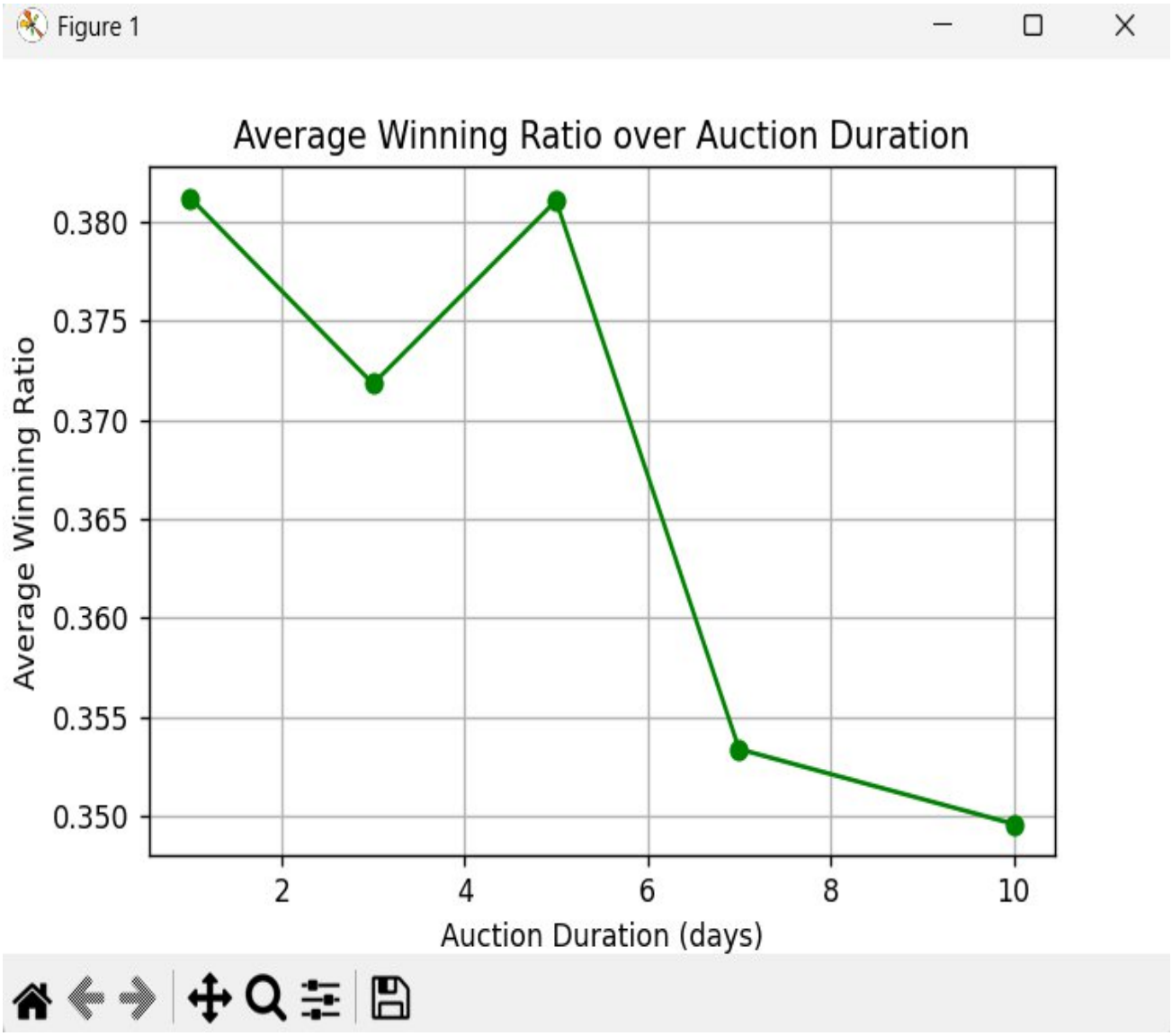
This section investigates whether auction duration has any impact on the participant's chance of winning. Specifically, it looks at how the **average winning ratio** changes with different auction durations.

Insights:

The trend analysis revealed that auctions with shorter durations had a **higher average winning ratio**, especially among shill bidders. This could imply that fraud is more common or effective in shorter auctions, where there is less time for legitimate users to react. Longer auctions showed more moderate winning ratios, possibly due to a higher chance of genuine competition balancing the outcome.

Visualization:

A line plot was used to show how average winning ratios evolve with auction duration. The trend line helped identify auction lengths that are more prone to manipulation.



Conclusion

Utilizing NumPy, Pandas, Matplotlib, and Seaborn for Python analysis projects provides a robust foundation for data manipulation, visualization, and exploration. These libraries are essential tools for any data analyst or scientist, enabling efficient data handling and insightful visualizations.

- NumPy offers powerful numerical operations, making it ideal for complex computations.
- Pandas simplifies data manipulation and analysis with its versatile DataFrame structure.
- Matplotlib and Seaborn provide comprehensive visualization capabilities, helping to uncover patterns and trends in data.

By leveraging these libraries, you can create comprehensive data analysis projects that effectively extract insights from data. The resources provided earlier can further enhance your skills and help you tackle more complex projects in the future.

Share

Export

Rewrite

FUTURE SCOPE

The current analysis provides a solid foundation for understanding the behavioral traits of shill bidders in online auctions. However, there are several directions in which this work can be extended and improved to build more robust and real-time fraud detection systems:

1. Machine Learning-Based Detection Models

Building classification models using machine learning algorithms like **Random Forest, XGBoost, SVM, or Neural Networks** can enhance detection accuracy. With the EDA insights as a guide, these models can be trained to identify suspicious bidding patterns more effectively.

2. Feature Engineering and Automation

Further **feature extraction**—such as session patterns, time intervals between bids, or user history across auctions—can improve model performance. Automating this process using scripts or pipelines would also allow for scalability in larger datasets.

3. Real-Time Detection Systems

Integrating fraud detection with **real-time auction platforms** would allow platforms to flag or block shill bidders while the auction is ongoing. This would require a lightweight, fast, and highly accurate model deployment setup.

4. User Behavior Profiling

Developing user profiles based on long-term activity across multiple auctions can help distinguish between new, irregular, and suspicious behavior. This longitudinal analysis can provide deeper context.

5. Adaptation to Other Fraud Types

The techniques used here can be adapted to detect other online auction frauds such as **bid shielding, bid siphoning, or identity spoofing**, making the system more comprehensive.

6. Larger and Real-World Datasets

Applying these methods to larger or real-world auction data (from platforms like eBay or simulated environments) would make the findings more generalizable and applicable in practice.

7. Collaborative Filtering for Fraud Networks

Future research could explore detecting **collusive groups** of shill bidders using network analysis or graph-based methods, revealing hidden relationships between fraudulent users.

REFERENCES

If you've been using NumPy, Pandas, Matplotlib, and Seaborn for your Python analysis projects, here are some webpage names with links that might be helpful for further learning or project ideas:

1. **NumPy Documentation**
   https://numpy.org/doc/
2. **Pandas Documentation**
   https://pandas.pydata.org/docs/
3. **Matplotlib Documentation**
   https://matplotlib.org/stable/index.html
4. **Seaborn Documentation**
   https://seaborn.pydata.org/documentation.html
5. **DataCamp's Pandas Tutorial**
   https://www.datacamp.com/tutorial/pandas-tutorial-python
6. **Real Python's Matplotlib Tutorial**
   https://realpython.com/python-matplotlib-guide/
7. **LinkedIn**
   https://www.linkedin.com/posts/revanth-kaja-4ab19b2a2_dataanalysis-python-matplotlib-

activity-7317210962938212352-tV-v?utm_source=share&utm_medium=member_desktop&rcm=ACoAAEj_kgMBVznEUDMUr1EcYJtvC8OATnePwkk