IE 7374: Final Project Document

Shill Bidding Prediction

By – Group 13

Group Members:

Abhishek Shetty
Danish Bhatt
Eswar Balaji Jalaparthi

Table of Contents:

Abstract:

The Data was scraped from many eBay auctions of popular products. The dataset we used was obtained after this auction data was preprocessed.

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:**	7240

Figure 1: Data Abstract

Introduction:

Shill bidding in auctions is the practice of placing bids on behalf of the seller with a motive to drive up the price of the auctioned item. In auctions of luxury items like art and antiques where the bidders' valuations differ, and the payoffs sellers' get from fraud; Shill bidding is said to have occurred. We though our project wants to check the features and the extent to which they contribute to Shill Bidding. This project aims to help companies like eBay to predict which Bidders are taking part in this fraud called Shill Bidding.

Data Description:

We have the following attributes in our dataset:

1. **Record ID:** This is a unique identifier for records in the dataset.
2. **Auction ID:** This is a unique identifier of auctions.
3. **Bidder ID:** This is a unique identifier of bidders.
4. **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.
5. **Bidding Ratio:** A shill bidder participates more frequently to raise the auction price and attract higher bids from legitimate participants.
6. **Successive Outbidding:** A shill bidder successively outbids himself even though he is the current winner to increase the price gradually with small consecutive increments.
7. **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.
8. **Auction Bids:** Auctions with SB activities tend to have a much higher number of bids than the average of bids in concurrent auctions.
9. **Auction Starting Price:** a shill bidder usually offers a small starting price to attract legitimate bidders into the auction.
10. **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.
11. **Winning Ratio:** A shill bidder competes in many auctions but hardly wins any auctions.
12. **Auction Duration:** How long an auction lasted.

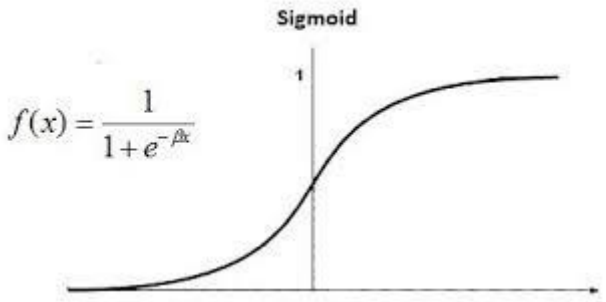
13. **Class:** 0 for normal behavior bidding; 1 for otherwise. This is our Target Attribute.

Methods:

Since this is a classification problem, we have used Logistic Regression, KNN and Feedforward Neural Network for our predicting our target. Each of them are popular models for classification problems in their own right and here we will talk about the pros and cons of each of these methods. These are the reasons which led to us using them:

1) Logistic Regression:

It is one of the most popular classification algorithms which is used to find how probable is the success or the failure of an event. It is used when the variable we want to predict is binary in nature. It allows the users to categorize the data into discrete classes by analyzing the relations from a set of labelled data. The linear relationship between the dataset is first studied and then we introduce the non-linearity in the form of a sigmoid function. It is also known as Binomial Logistic regression.



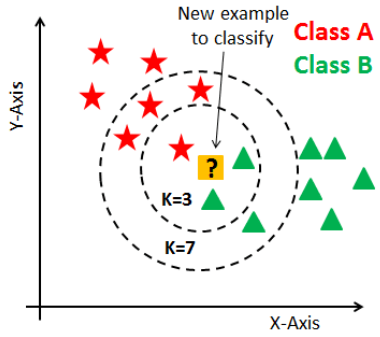
Advantages	Disadvantages
It is easy to implement, interpret, and efficient to train.	Cannot handle data where the number of features is more than the number of observations. This is because it leads to overfitting.
It not only provides a measure of how appropriate a predictor (coefficient size) is, but also its direction of association (positive or negative).	It can only be used to predict discrete functions. Hence, the dependent variable of Logistic Regression is bound to the discrete number set.
It is very fast at classifying unknown records.	Since it has linear decision surface non -linear problems cannot be solved using this algorithm.
The accuracy is good for simple data sets. It even performs well if the data is linearly separable.	Logistic Regression requires average or no multicollinearity between independent variables.
It can interpret model coefficients as indicators of feature importance.	It is difficult to get complex relationships using logistic regression.
Even though it is less inclined to over-fitting, it can still occur in high dimensional datasets. Regularization (L1 and L2) techniques can be	Logistic Regression requires the independent variables to be linearly related to the log odds (log(p/(1-p)).

considered to avoid over-fitting in these scenarios.	

Table 1: Logistic Regression Pros and Cons

2) KNN

K-nearest neighbor algorithm is a very popular method to solve classification problems. This algorithm uses similarity measures like distance functions to classify new data points from the stored data. This means that when a new data point appears, K-NN algorithm can easily classify into a well-suited category.
The new data point is assigned the same class the majority of it's neighbors have measured by the distance function. It can be illustrated with the help of the following example:



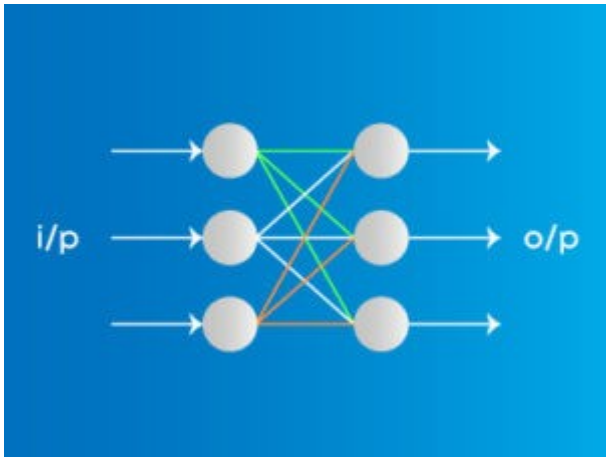
If we take if k = 3, the nearest three neighbors of the data point is found and based on the majority votes of its neighbors, the data point is classified into a class. In the case of k = 3, for the above diagram, it is Class B. Similarly, when k = 7, for the above diagram, based on the majority votes of its neighbors, the data point is classified to Class A.

Advantages	Disadvantages
It is quick calculation time.	The Accuracy Depends on the quality of data fed.
The algorithm is easy to interpret.	The prediction speed will slow down if the dataset is large.
This algorithm has high accuracy.	As there is a requirement to store the data,

	this algorithm requires high memory.
For a nonlinear data case it is an important algorithm. In this we do not require to make assumptions about data, tune parameters.	Since this algorithm stores all the training data, it can be computationally expensive.

Table 2:KNN pros and cons

3) Deep Neural Network:



Deep Neural networks are based on perceptron. In deep neural network we have more than one hidden layer. Initially the input is multiplied with the weights of the hidden layers and then transformed using activation functions in the forward pass and using the loss function we backpropagate through the whole network to update the weights of the layers with the objective of decreasing the loss.

Advantages:

1. Can transform the input into many dimensions.
2. Can fit any data.

Disadvantages:

1. Data and computation intensive.

Explanatory Data Analysis (EDA):

1. HISTOGRAM OF NUMERICAL COLUMNS

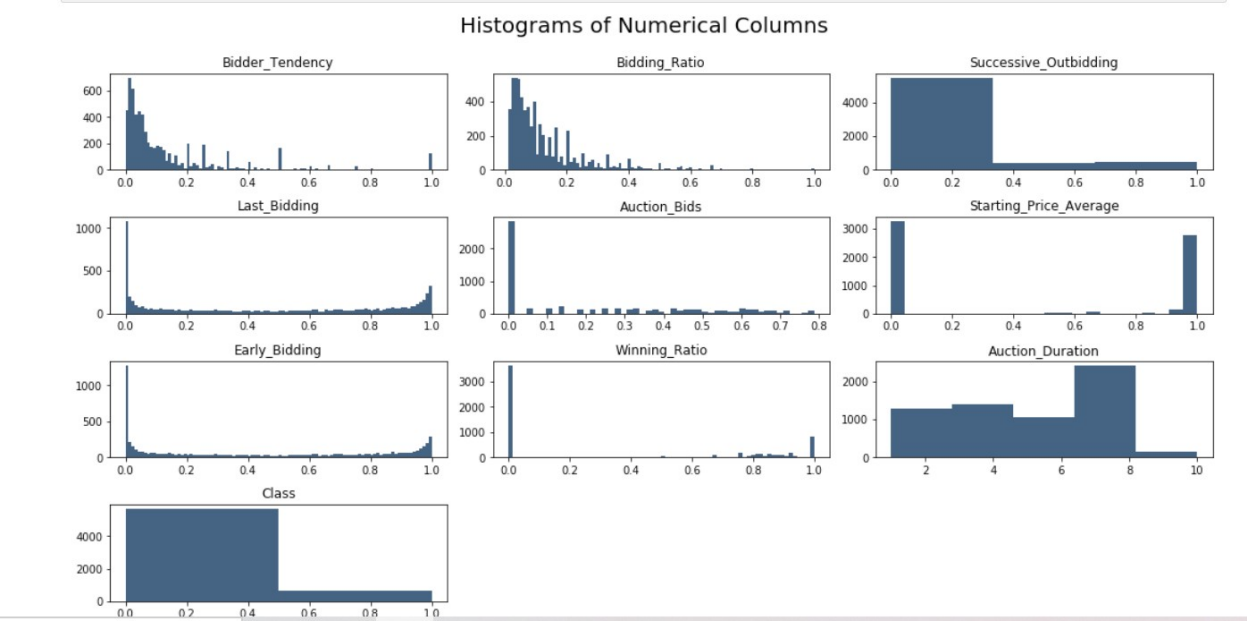


Figure 2: Histogram

Bidder Tendency: From the histogram we see that most of the people have the value of 0 - 0.2. Between 0.2 - 0.4, we see the next cluster of Bids.

Bidding Ratio: The histogram implies that most people are in the Bidding Ratio of 0 - 0.2 which means in most cases, there is no bidding ratio.

Successive Outbidding: There is a high percentage of people in the 0 category and very less people in the in 1 category which implies that most records show that they do not indulge in successive outbidding.

Last Bidding: A Shill Bidder becomes inactive at the last stage. There is not an evidence of Last Bidding but in the end, the data shows a spike in the number of people in last bidding.

Auction Duration: The data shows that the auction duration is mainly between 1 - 6 for most number of biddings and between 6-8, the number of bidding is the highest.

Class: The histogram suggest in most of the bidding cases (6000), the data is 0 which implies the shill bidding did not take place while the answer is 1 for around less than 1000 cases.

2. PAIRPLOT

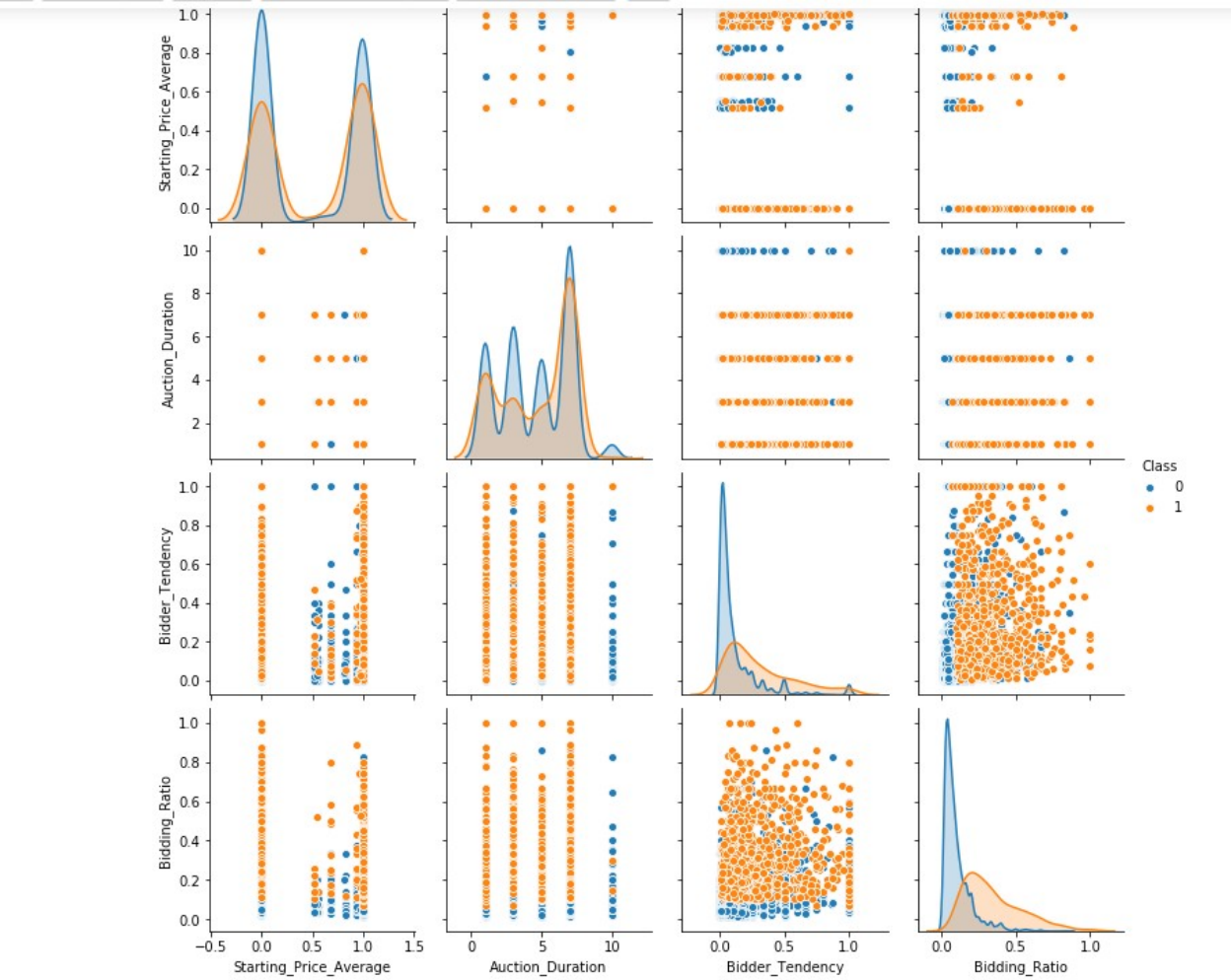


Figure 3: Pair plot

The blue portion shows the Class 0 while as Orange portion shows the Class 1.

Class 0: Represents that the shill bidding did not take place. The graph in every case shows that class 0 has a higher peak.

In Starting Price Average, Class 0 has a higher peak compared to Class 1 which suggests a higher influence of Starting Price Average on Class 0.

Bidder Tendency: The Bidder Tendency has a Left ward skew and a higher peak for Class 0. The same graph is for Bidding Ratio.

Bidding Ratio and Bidder Tendency: In the graph of the two, there is a clear distinction between the two classes. Class 0 has lower Bidding Ratio whereas Class 1 Bidding Ratio ranges from 0.2 TO 1 equally.

In the pair plot for Bidder Tendency and Auction Duration, It can be seen that Class 0 has the highest Auction Duration of 10 while Class is evenly distributed. It can also be seen that Class 0 is not in the lower Auction Duration.

In the Pair plot of Starting Price Average and Bidding Tendency, Bidding Ratio, The two classes separated can be clearly seen. Class 0 is between Starting Price Average of 0.5 and 1. The Bidding Ratio for Class 0 is low up to 0.2-0.3.

Auction Duration for Class 0 has several Peaks up to 4 while as Class 1 has two smaller peaks widely spread.

3. Correlation Plot

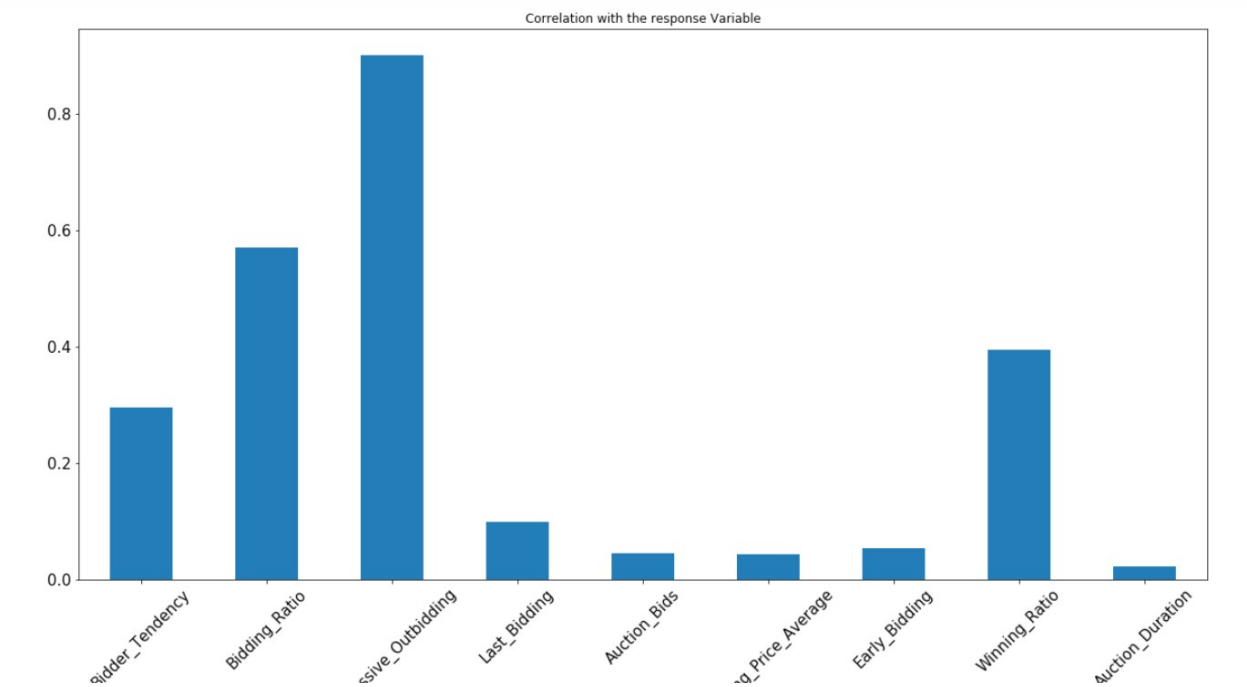


Figure 4: Correlation Plot

Successive Outbidding: A shill bidder successively outbids himself even though he is the current winner to increase the price gradually with small consecutive increments.

The co-relation plot for Successive Outbidding and response variable is highest. It suggests that if there is successive outbidding, then the co-relation of outbidding is the highest.

Bidding Tendency, Bidding Ratio, Winning Ratio also have high correlation which show that if their value is high, so is the probability of the Bidder indulging in Shill Bidding.

The correlation of Last Bidding, Auction Bids, Starting Price Average, Early Bidding and Auction Duration has a very low correlation with the response variable. It suggests that these variables have very less influence on the response variable.

4. HEAT MAP

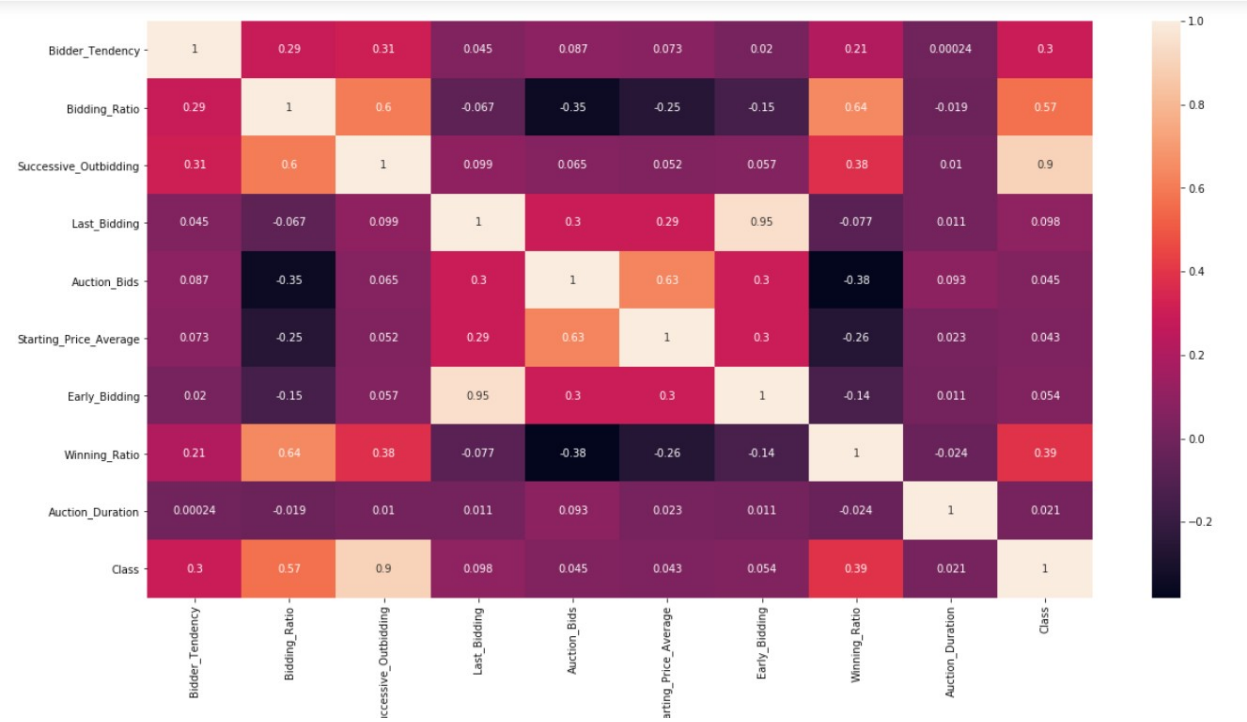


Figure 5: Heat Map

In the heat map, we see a 0.9 correlation between successive bidding and Class. We also see a 0.95 correlation between Early Bidding and Last Bidding.

5. Correlation Matrix

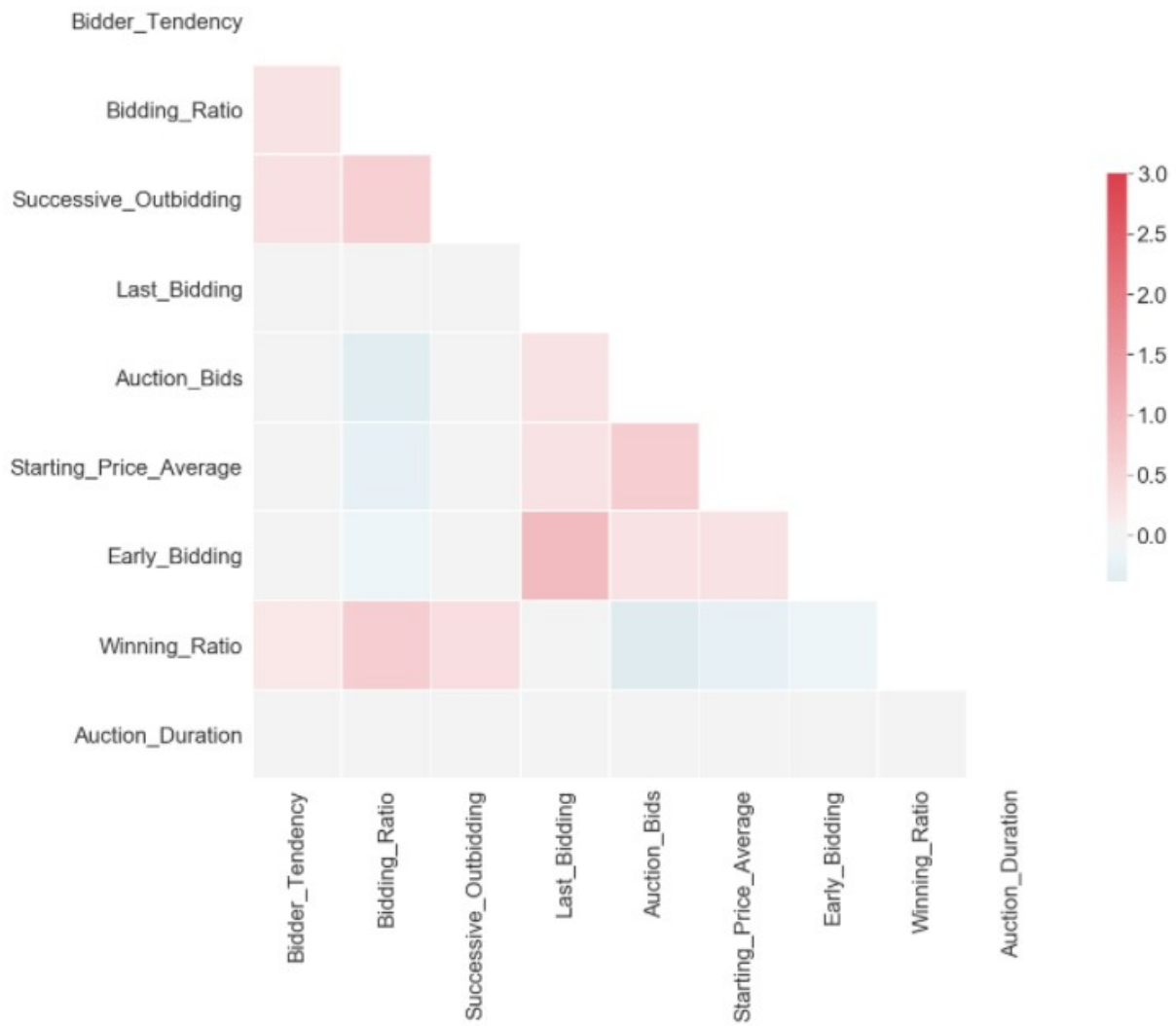


Figure 6: Correlation Matrix

Positive Correlation:

- Bidder Tendency and Bidding Ratio, Successive Outbidding, Winning Ratio have high correlation.
- Successive Outbidding and Winning ratio have high correlation.
- Last Bidding and Starting Price Average, Early Bidding, and Auction Bids have positive correlation.

Negative Correlation:

- Bidding Ratio and Auction Bids, Starting Price Average, Early Bidding have negative correlation.
- Auction Bids and Winning Ratio have negative corelation.
- Starting Price Average and Winning Ratio have negative correlation

6. Data Summary

```
In [21]: shill.describe()
```

Out[21]:

	Record_ID	Auction_ID	Bidder_Tendency	Bidding_Ratio	Successive_Outbidding	Last_Bidding	Auction_Bids	Starting_Price_Average	Early_Bidc
count	6321.000000	6321.000000	6321.000000	6321.000000	6321.000000	6321.000000	6321.000000	6321.000000	6321.000
mean	7535.829457	1241.388230	0.142541	0.127670	0.103781	0.463119	0.231606	0.472821	0.430
std	4364.759137	735.770789	0.197084	0.131530	0.279698	0.380097	0.255252	0.489912	0.380
min	1.000000	5.000000	0.000000	0.011765	0.000000	0.000000	0.000000	0.000000	0.000
25%	3778.000000	589.000000	0.027027	0.043478	0.000000	0.047928	0.000000	0.000000	0.026
50%	7591.000000	1246.000000	0.062500	0.083333	0.000000	0.440937	0.142857	0.000000	0.360
75%	11277.000000	1867.000000	0.166667	0.166667	0.000000	0.860363	0.454545	0.993593	0.826
max	15144.000000	2538.000000	1.000000	1.000000	1.000000	0.999900	0.788235	0.999935	0.999

Figure 7: Data Summary

The dataset has a total of 6321 records. The data shows that Starting Price Average, Winning ratio, Early bidding and Last bidding have the highest Standard Deviation. Bidding Tendency, Bidding Ratio have the least Standard Deviation.

Auction Duration is for an average of 4.6 hrs. and a standard deviation of 2.466hrs.

Results:

To understand the data more, we have transformed the data using principal component analysis and the first two principal components account for 62% of the variation. Utilized the first two principal components and the class labels for a scatter plot.

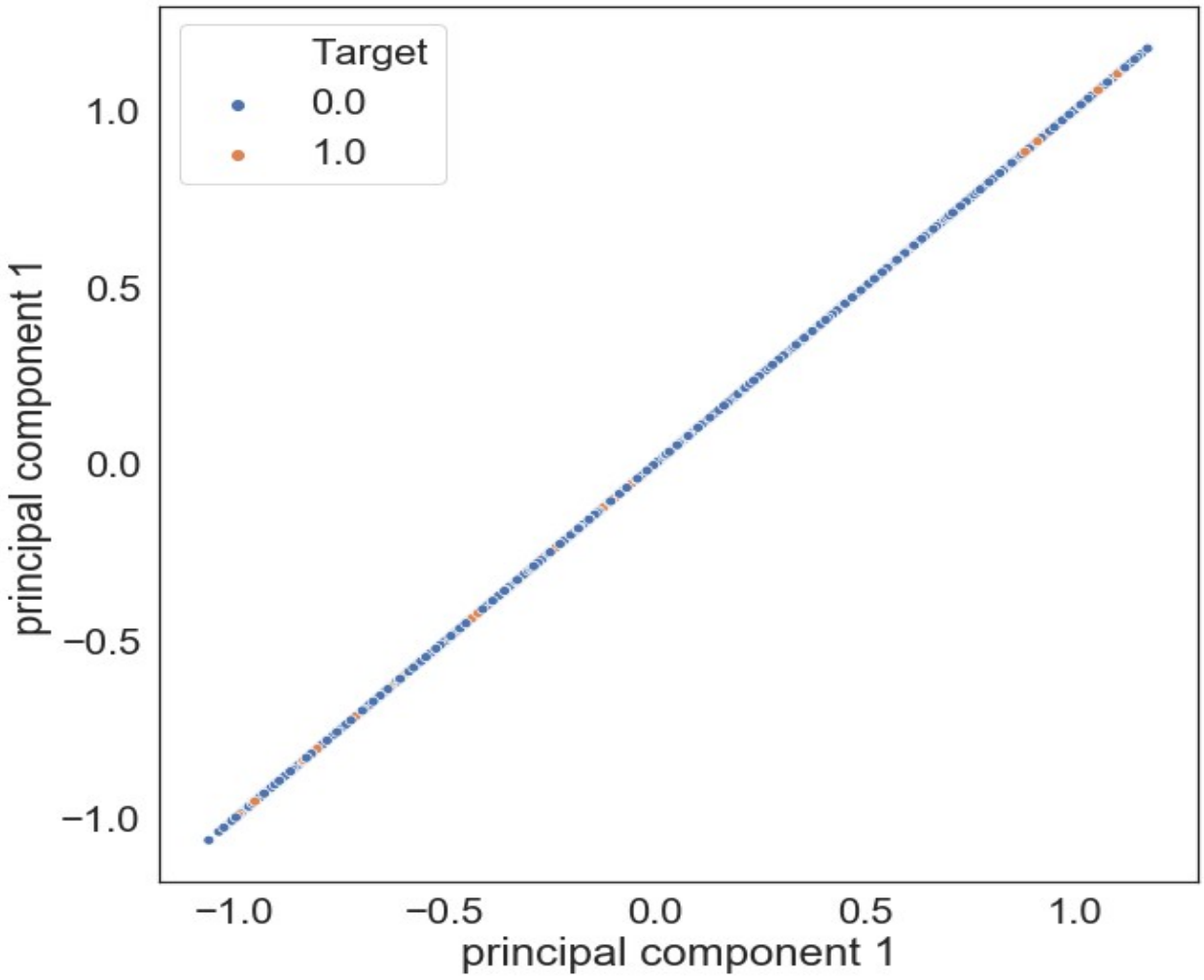


Figure 8: PCA Transformed scatter plot

From the above plot we can see that there is no linear boundary or hyperplane that can be drawn to split the data points and we observe that nearest neighbors dictate the class of a point and deep learning can find a decision boundary by transforming the input.

Because of the above reasons we opted to model using logistic regression as the base model and KNN, DNN.

We split the data into three parts as train, validation and test data to build machine learning models.

Train data has 70% of the data, validation has 15% of the data and test has 15% of the data. The data is shuffled before splitting. The target labels are imbalanced in 8:1 ratio, but the models are performing very well without having to over sample or under sample the data.

Accuracy

	Logistic Regression	KNN	Deep Neural Network
Train	0.96	0.997	0.999
Test	0.95	0.998	0.997

Recall

	Logistic Regression	KNN	Deep Neural Network
Train	0.81	0.977	1
Test	0.76	0.99	0.99

Precision

	Logistic Regression	KNN	Deep Neural Network
Train	0.80	0.994	1
Test	0.84	0.99	0.99

From the above three tables we see that Logistic regression has performed very poorly on the train and test data, the recall and precision are very less.

Using validation data, we have observed that k=5 works best for our data. KNN with 5 nearest neighbors significantly improved all the performance metrics for both train and test data with no signs of overfitting.

Finally, to improve performance further, we fit the data using deep neural network with one hidden layer and “Relu” and “sigmoid” as the activation functions. This further improved the performance metrics on train and test data.

For this data Deep Neural network with one hidden layer worked best.

Discussion:

- In our dataset the number of records with class 0 is around 8 times more than the records with class 1. So, in the future we could try oversampling records with class 1 or under sample data with class 0.
- In the future we can add regularization to detect which features are more important for predicting our class.
- We can try collecting more data for our prediction.
- We have tried SVM, Naïve Bayes using normal distribution, but the performance metric was low for those algorithms. We can try gradient boosting trees to further improve the accuracy.

References:

- https://keras.io/
- https://numpy.org/doc/
- https://pandas.pydata.org/docs/
- https://pandas.pydata.org/docs/
- https://archive.ics.uci.edu/ml/datasets/Shill+Bidding+Dataset
- http://towardsdatascience.com/

Table of Figures: