SHILL BIDDING FRAUD DETECTION

Submitted to the Office of Graduate and Professional Studies of

UNIVERSITY OF HOUSTON

EDS 6340 – INTRODUCTION TO DATA SCIENCE

MASTER OF SCIENCE IN ENGINEERING DATA SCIENCE

UNIVERSITY OF HOUSTON

Spring 2023

Professor: Dr. Amaury Lendasse

Authors:

Harshavardhan Reddy Lingammagari	**–**	**2251826**
Sravan Josh Koka	**–**	**2218808**
Viniktha Gadde	**–**	**2242893**
Venkata Sai Bharath Thoranala	**–**	**2245697**

Table of Contents

1. INTRODUCTION 3

1.1. About the Dataset: 3

2. DATA PRE-PROCESSING 3

3. RESULTS FOR ALL THE SINGLE MODELS 4

3.1. Logistic Regression 4

3.2. K-Nearest Neighbors Classifier (KNN) 4

3.3. Support Vector Machines (SVM) 4

3.4. Random Forest Classification: 5

3.5. Support Vector Machine – Non-Linear: 5

3.6. Extreme Learning Machines (ELM): 5

4. Variable Selection 6

4.1. Variable selection using Lasso and correlation. 6

4.2. BI-DIRECTIONAL ELIMINATION 7

5. Clustering 7

6. VISUALIZATION 8

7. ENSEMBLE MODELLING 8

8. GENERAL DISCUSSION 9

9. CONCLUSION 9

10. REFERENCES 10

List of figures

Figure 2–1 Figure showing the null values and the datatypes of features from the data. 3

Figure 4–1 Performance metrics of the two best performing models on selected features using primary slection method. 6

Figure 4–2 Execution of Bidirectional feature selection and the corresponding Confusion matrices for the best performing models 7

Figure 5–1 Figure showing inertia values for different values of k, and silhouette plots for different numbers of clusters. 7

Figure 6–1 Left: showing the variances of the principal components and right: visualizing the first three principal component 8

Figure 7–1 Confusion matrix and classification report for ensemble model on the data. 9

Figure 9–1 Comparing the performance of all models. 9

1. INTRODUCTION

Shill bidding fraud is a type of auction manipulation in which a seller or a third party artificially inflates the price of an item up for auction by placing bids on the item under false pretenses. The term "shill" refers to a person who is hired or enlisted to create the appearance of legitimate bidding activity in an auction, to drive up the price of an item. This can be done by either the seller themselves, or by someone acting on their behalf. Shill bidding fraud can be conducted in various ways, such as through multiple accounts, anonymous bidding, or fake identities. Shill bidding aims to make it appear as if there is high demand for an item, which can entice other bidders to raise their bids in response. Shill bidding is considered fraudulent because it misleads bidders into believing that an item is worth more than it is and can result in them paying more than they should.

1.1. About the Dataset:

The dataset was downloaded from the open source UCI machine learning repository https://archive.ics.uci.edu/ml/datasets/Shill+Bidding+Dataset#. The dataset has 6341 instances and 12 attributes. The dataset is structured and has a "Class" attribute, the target variable. Overall, the problem type is a binary classification where the target variable class "0" indicates a normal bid, and "1" indicates a fraudulent bid.

2. DATA PRE-PROCESSING

Data preprocessing refers to the steps taken to clean and prepare raw data before it can be used for analysis. This typically involves removing irrelevant or duplicate information, dealing with missing values, handling outliers, and transforming the data into a more usable format. Good data preprocessing can improve the accuracy and reliability of analytical results.

```
Auction_ID                0
Bidder_ID                 0
Bidder_Tendency           0
Bidding_Ratio             0
Successive_Outbidding     0
Last_Bidding              0
Auction_Bids              0
Starting_Price_Average    0
Early_Bidding             0
Winning_Ratio             0
Auction_Duration          0
Class                     0
dtype: int64
```

```
Auction_ID                int64
Bidder_ID                object
Bidder_Tendency         float64
Bidding_Ratio           float64
Successive_Outbidding   float64
Last_Bidding            float64
Auction_Bids            float64
Starting_Price_Average  float64
Early_Bidding           float64
Winning_Ratio           float64
Auction_Duration          int64
Class                     int64
dtype: object
```

Figure 2–1 Figure showing the null values and the datatypes of features from the data.

- Checking for outliers in the dataset, with the help of box plots. After plotting the box plots, fortunately, there were no outliers in the dataset.
- After all the data preprocessing, the conclusion was that the dataset was almost clean. This might be because it was meant to be downloaded for end users for research purposes on UCI website.
- For further analysis, we dropped the attribute "Bidder_ID" as it contains alphanumeric digits and each entity unique, so it was difficult to encode it.

3. RESULTS FOR ALL THE SINGLE MODELS

We applied many different Machine learning models on our dataset, our workflow was to first we split our data into training and testing datasets. Later, we implemented classification models on the training and validation set and calculated the performances of each model. As we used k-fold cross validation we did not take validation set. Then we did K-fold cross-validation on the model as our model structure selection and used GridSearchCV and RandomizedSearchCV wherever required to tune the hyper parameters of the classification models we have chosen.

3.1. Logistic Regression

As a part of Linear Model, we used Logistic Regression (Logistic Regression, n.d.) as it is considered as a generalized linear model, and it performs well for classification problems. We trained the model on training dataset and the hyperparameters were found using GridSearchCV. Best hyperparameters are {C: 2.559, max iterations: 150, Penalty: 'l2'}. Performs very well on the data with an accuracy of 98 percent.

	precision	recall	f1-score	support
0	0.98	1.00	0.99	1110
1	0.97	0.88	0.92	155
accuracy			0.98	1265
macro avg	0.98	0.94	0.96	1265
weighted avg	0.98	0.98	0.98	1265

Figure 3–1 Classification report for logistic regression

3.2. K-Nearest Neighbors Classifier (KNN)

The K-Nearest Neighbor (KNN, n.d.) Classifier is a non-parametric algorithm used for classification and regression tasks. It works by finding the K nearest data points to a new input and classifying it based on the majority class of those neighbors. KNN is simple to implement and doesn't require assumptions about the underlying data distribution but can be computationally expensive for large datasets.

KNN Report:

	precision	recall	f1-score	support
0	0.99	0.99	0.99	1133
1	0.91	0.95	0.93	132
accuracy			0.99	1265
macro avg	0.95	0.97	0.96	1265
weighted avg	0.99	0.99	0.99	1265

Figure 3–2 Results for KNN

3.3. Support Vector Machines (SVM)

Support Vector Machine (SVM, n.d.) is a powerful machine learning algorithm that can be used for both classification and regression tasks. It works by finding the best possible boundary (i.e., hyperplane) that can separate different classes of data points in a high-dimensional space. We tried using linear kernel. After hyperparameter tuning we found that the best hyper parameters are {'C': 1.0, 'gamma': 0.1, 'kernel': 'linear'}. It performed well with an accuracy of 98 percent.

	precision	recall	f1-score	support
0	0.99	0.98	0.99	1137
1	0.87	0.91	0.89	128
accuracy			0.98	1265
macro avg	0.93	0.95	0.94	1265
weighted avg	0.98	0.98	0.98	1265

Figure 3–3 Confusion matrix for SVM classifier

3.4. Random Forest Classification:

Random Forest (Anlytics Vidhya, n.d.) uses the combination of multiple random decision trees each trained on a subset of data. The use of multiple trees gives stability to the algorithm and reduces the variance. After tuning the hyperparameters the best parameters are {max depth: 20, min samples leaf: 1, minimum samples split: 5, no. of. estimators: 100}. It performs well on our dataset with 97 percent accuracy with best hyper parameters outperforms the baseline model by nearly 9 percent.

	precision	recall	f1-score	support
0	0.97	1.00	0.98	1975
1	0.97	0.76	0.85	238
accuracy			0.97	2213
macro avg	0.97	0.88	0.92	2213
weighted avg	0.97	0.97	0.97	2213

Figure 3–4 Confusion matrix for Random Forest classifier

3.5. Support Vector Machine – Non-Linear:

Support Vector Machine Non-Linear (SVM, n.d.), we use Kernels to make non-separable data into separable data. We map data into high dimensional space to classify. Kernel type is the most important parameter for SVC. It can be linear, polynomial or gaussian SVC. We tried both ‘poly’ and ‘rbf’ (a gaussian type) kernel for Non-linear model. After using RandomizedSearchCV for hyper parameters we found the best hyperparameters to be {'C': 18.52669390630025, 'gamma': 0.2563640674119393, 'kernel': 'rbf'}.

	precision	recall	f1-score	support
0	0.99	0.98	0.99	1133
1	0.87	0.95	0.91	132
accuracy			0.98	1265
macro avg	0.93	0.97	0.95	1265
weighted avg	0.98	0.98	0.98	1265

Figure 3–5 Classification report for Support Vector Machine

3.6. Extreme Learning Machines (ELM):

Extreme Learning Machine (ELM) is a machine learning algorithm that is used for supervised learning tasks, such as classification and regression. It works by randomly initializing the input weights of a single hidden layer neural network and then solving for the output weights analytically. ELM is known for its fast-training speed and good generalization performance.

	precision	recall	f1-score	support
0	1.00	0.99	0.99	1133
1	0.91	0.97	0.94	132
accuracy			0.99	1265
macro avg	0.95	0.98	0.97	1265
weighted avg	0.99	0.99	0.99	1265

Figure 3–6 Classification report for Extreme Learning Machine

4. Variable Selection

Variable selection is the process of identifying a subset of relevant features from a larger set of potential predictors in a dataset. This is an important step in many machine learning and statistical modeling tasks because including irrelevant or redundant features can lead to overfitting and poor generalization performance. There are two main types of variable selection: filter and wrapper methods. Filter methods involve selecting features based on their statistical properties, such as correlation or mutual information with the target variable. Wrapper methods, on the other hand, use a predictive model to evaluate the performance of different subsets of features and select the best performing one. Other types of variable selection methods include embedded methods, which incorporate feature selection directly into the model training process, and hybrid methods, which combine multiple variable selection techniques. The choice of variable selection method depends on the specific problem and dataset characteristics.

4.1. Variable selection using Lasso and correlation.

Lasso and correlation are valuable techniques for feature analysis, aiding in the development of more accurate and interpretable models. Lasso, a regularization method in linear regression, helps prevent overfitting and enhances model generalization by adding a penalty term to the loss function. This term pushes less important features' coefficients to zero, leading to automatic feature selection and reduced model complexity. On the other hand, correlation analysis measures the strength and direction of the relationship between variables. By identifying highly correlated features, we can reduce multicollinearity issues, further simplifying the model and enhancing interpretability. Employing Lasso and correlation analysis in tandem allows for a more focused feature set, promoting better model performance and easier interpretation of the underlying relationships within the data.

RF Report:
[[1114 18]
[2 131]]

	precision	recall	f1-score	support
0	0.99	0.99	0.99	1132
1	0.90	0.92	0.91	133
accuracy			0.98	1265
macro avg	0.94	0.96	0.95	1265
weighted avg	0.98	0.98	0.98	1265

NN Report:
[[1118 14]
[10 123]]

	precision	recall	f1-score	support
0	1.00	0.98	0.99	1132
1	0.88	0.98	0.93	133
accuracy			0.98	1265
macro avg	0.94	0.98	0.96	1265
weighted avg	0.99	0.98	0.98	1265

Figure 4–1 Performance metrics of the two best performing models on selected features using primary slection method.

4.2. BI-DIRECTIONAL ELIMINATION

"Bi-directional elimination" is used to iteratively choose and eliminate features based on their performance both individually and collectively. To find the most pertinent features for shill bidding activity detection in this research, we employed the bi-directional elimination technique to the Shill Bidding Dataset.

Forward selection and backward elimination are the two processes that make up the bi-directional elimination procedure. We begin the forward selection stage with a blank set of selected features and incrementally add characteristics that enhance the model's performance. In the backward elimination process, we begin with all characteristics present and gradually remove those that don't significantly affect the performance of the model.

RF Report:

```
[[1131    2]
 [   1  131]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      1133
           1       1.00      0.98      0.99       132

    accuracy                           1.00      1265
   macro avg       1.00      0.99      1.00      1265
weighted avg       1.00      1.00      1.00      1265
```

NN Report:

```
[[1133    0]
 [   2  130]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      1133
           1       0.98      0.99      0.99       132

    accuracy                           1.00      1265
   macro avg       0.99      1.00      0.99      1265
weighted avg       1.00      1.00      1.00      1265
```

Figure 4–2 Execution of Bidirectional feature selection and the corresponding Confusion matrices for the best performing models

5. Clustering

Clustering can be helpful for exploring data and identifying patterns, but it may not be sufficient for developing a predictive model for shill bidding identification. The goal of this task is to categorize events as either shill or real bidding, and clustering groups similar cases based on features but does not provide labels for categorization. Labeled data is necessary to create a supervised learning model that can identify shill bidding. Clustering can be used in combination with supervised learning methods to aid in data exploration and understanding, but it should not be used alone for shill bidding identification.

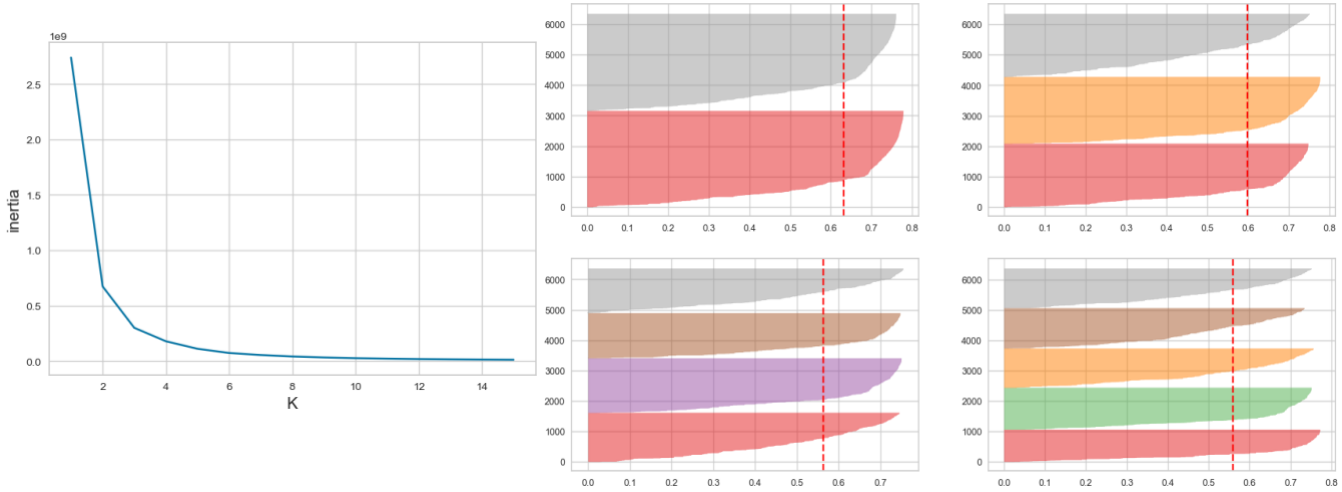

Figure 5–1 Figure showing inertia values for different values of k, and silhouette plots for different numbers of clusters.

6. VISUALIZATION

For understanding the underlying structure of the dataset and getting insights into the correlations between variables, visualization employing dimensionality reduction techniques, such as Principal Component Analysis (PCA) or t-SNE, might be helpful. However, visualization by itself might not immediately contribute to model development in the specific context of creating a predictive model for shill bidding identification in the current dataset.

In conclusion, while visualization utilizing dimensionality reduction techniques can provide information about the structure of the dataset, it should be used in conjunction with the right predictive modeling methods to create a model that is effective for shill bidding identification in the current dataset. Although the visualization aids in interpreting the data, choosing useful features, and evaluating the model's outputs, it is not a substitute for developing a predictive model on its own.

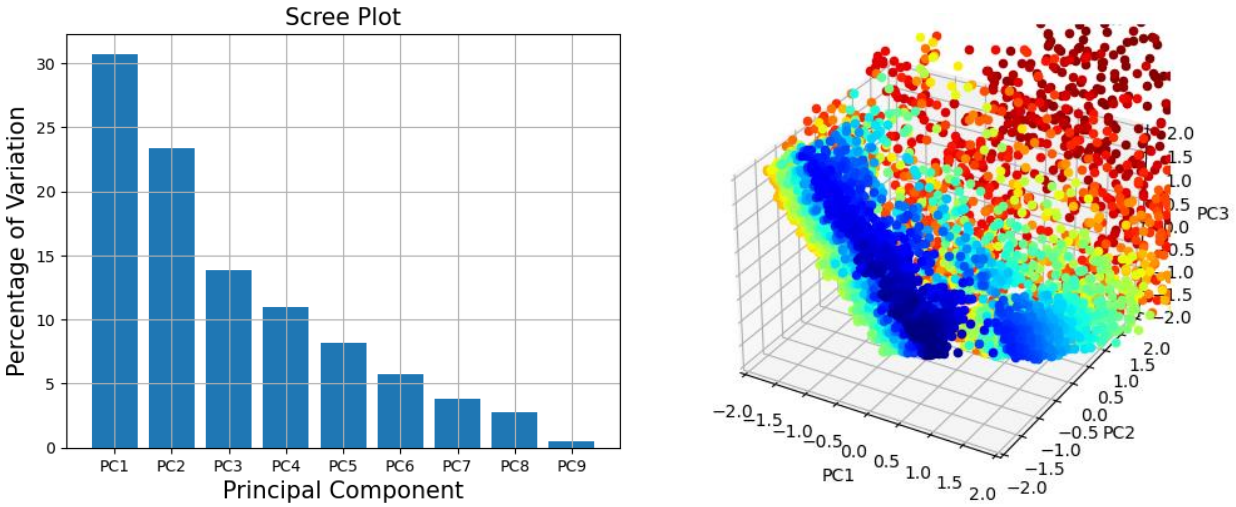

Figure 6–1 Left: showing the variances of the principal components and right: visualizing the first three principal component

7. ENSEMBLE MODELLING

This is a strong method for combining different models to increase overall predicted accuracy. Ensemble modeling can help to overcome the shortcomings of individual models and produce more reliable and accurate predictions in the context of shill bidding detection in the supplied dataset. We have created the ensemble model with the help of scikit library and used the stacking model to develop the ensemble. The following are the results that were achieved by the ensemble.

As we can observe from the results of Figure 7–1, in our dataset, ensemble modeling proved to be a useful method for shill bidding detection. As a result, the risk of overfitting was decreased and the model's capacity for generalization was enhanced. Ensemble approaches give a more accurate prediction by merging the results of different models, and they might be a useful tool for spotting shill bidding in online auctions.

```
[[1131    2]
 [   3  129]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      1133
           1       0.98      0.98      0.98       132

    accuracy                           1.00      1265
   macro avg       0.99      0.99      0.99      1265
weighted avg       1.00      1.00      1.00      1265
```

Figure 7–1 Confusion matrix and classification report for ensemble model on the data.

8. GENERAL DISCUSSION

If the project were to be conducted again, it would be beneficial to allocate additional resources to feature engineering and data exploration to improve the accuracy of capturing shill bidding behavior. To enhance the performance of the model, more sophisticated machine learning techniques such as neural networks and advanced feature selection techniques, including genetic algorithms, could be utilized. The project highlights the importance of understanding human behavior and online auction platforms when developing a shill bidding detection model. Accurate and reliable predictive models require comprehensive feature engineering and data research. Ensemble modeling, which combines multiple models, can improve the accuracy and dependability of the predictive model. The project provides valuable hands-on experience in machine learning, data analysis, and predictive modeling in a real-world setting, culminating in the successful development of a shill bidding detection model.

9. CONCLUSION

The results obtained by training different machine learning models are discussed here.

Models	Accuracy	Precision	Recall	F1-Score	Area under ROC Curve
SVM	0.99	0.92	0.99	**0.96**	0.98
LR	0.97	0.87	0.89	**0.88**	0.94
Random Forest	0.99	0.93	0.99	**0.96**	0.98
KNN	0.99	0.91	0.95	**0.93**	0.97
ELM	0.98	0.86	0.98	**0.91**	0.95
NN	0.99	0.96	0.99	**0.97**	0.99
Ensemble	0.99	0.98	0.98	**0.98**	0.99
MLP Bidirectional	1	0.98	0.99	**0.99**	1
MLP Lasso	1	0.99	0.99	**0.99**	1

Figure 9–1 Comparing the performance of all models.

From the above results shown in Figure 9–1, it can be concluded that the MLP model performed exceptionally well with an accuracy of almost 100%. This suggests that the MLP model was able to correctly predict each instance after being trained on the relevant variables selected by the Bi-directional elimination method and the Lasso method. The other models also performed well, with an accuracy of more than 93%.

The Shill Bidding Dataset used in the experiment is a specialized dataset used to detect fraudulent activity in online auction systems. The project demonstrated how machine learning methods can be used to detect shill bidding actions and provided valuable insights into feature selection, ensemble modeling, and model validation. The project also showed how combining various machine learning models can increase the accuracy and robustness of the predictive model. Additionally, visualization and clustering techniques were used to gain a better understanding of the underlying patterns in the data.

Overall, the project adds to the growing body of research on online fraud detection and highlights the effectiveness of machine learning in solving complex problems in the field. The findings of this project can be useful in developing better fraud detection systems for online auction platforms, thereby reducing the negative impact of shill bidding activities on honest buyers.

Bibliography

Anlytics Vidhya. (n.d.). Retrieved from https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

Huang., G. (n.d.). Retrieved from ELM: https://towardsdatascience.com/introduction-to-extreme-learning-machines-c020020ff82b

KNN. (n.d.). Retrieved from https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761

Logistic Regression. (n.d.). Retrieved from https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc

SVM. (n.d.). Retrieved from https://www.google.com/search?client=safari&rls=en&q=Support+vector+machine&ie=UTF-8&oe=UTF-8

Malerba, D., Esposito, F., and Semeraro, G. "*A Further Comparison of Simplification Methods for Decision-Tree Induction."* In D. Fisher and H. Lenz (Eds.), "*Learning from Data: Artificial Intelligence and Statistics V*", Lecture Notes in Statistics, Springer Verlag, Berlin, 1995.

Esposito F., Malerba D., & Semeraro G. *Multistrategy Learning for Document Recognition. Applied Artificial Intelligence,* 8, pp. 33-84, 1994.

Steven Eschrich and Nitesh V. Chawla and Lawrence O. Hall. *Generalization Methods in Bioinformatics* BIOKDD. 2002