Shill bidding detection

Unsupervised clustering of bidding records on eBay

Xin Zhao

Problem statement

- Shill bidding is a common auction fraud on eBay
- Shill bidding is hard to detect
- Unsupervised clustering to detect shill bidding

Data exploration

- Auction records on eBay from UCI

Auctions	807
Records	6321
Features	8
Bidding Durations	1, 3, 5, 7, 10

Durations	1 day	3 days	5 days	7 days	10 days	Total
Records	1289	1408	1060	2427	137	6321

- In the confidence level of 95%, there are significant differences in features of each bidding duration

- We can consider to detect shill bidding separately for each bidding duration

Experiment

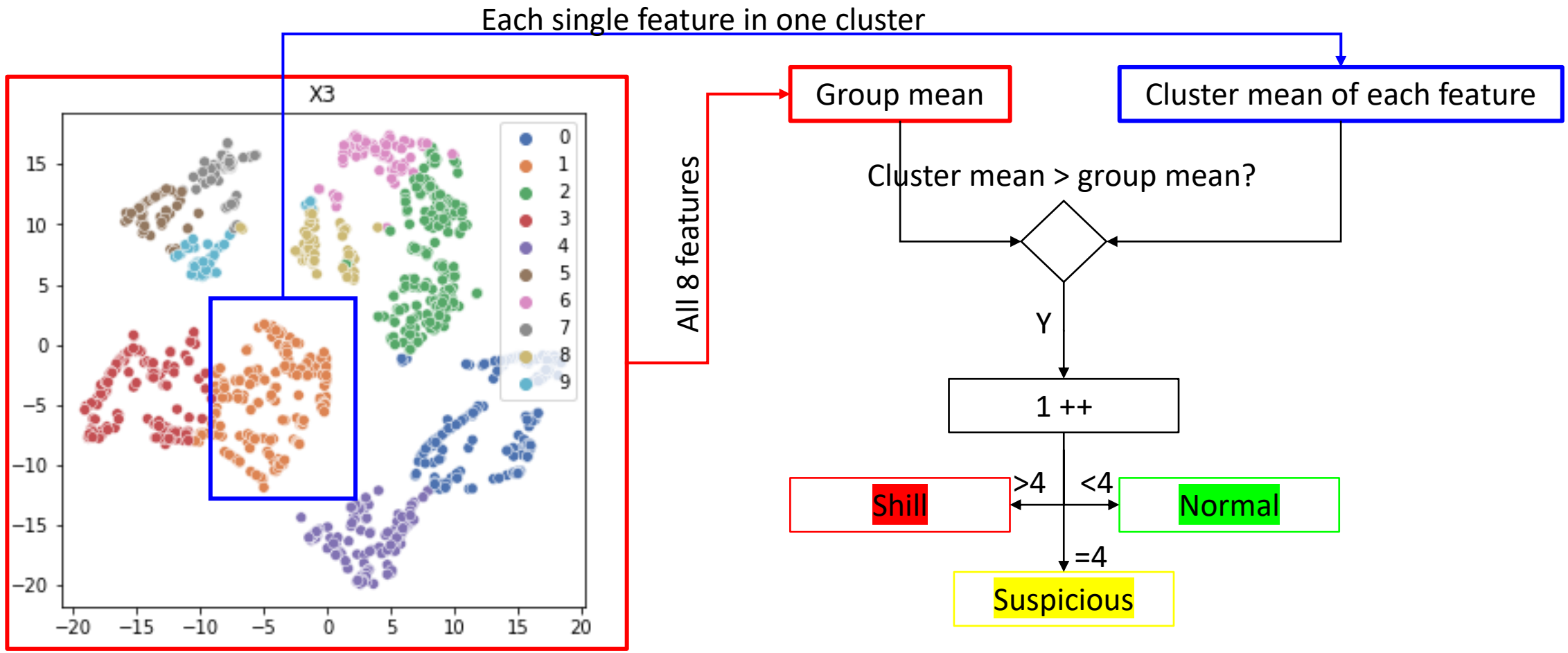
- Data preparation
  - Separate data based on auction duration into 5 groups

Durations	1 day	3 days	5 days	7 days	10 days	Total
Records	1289	1408	1060	2427	137	6321

- Unsupervised clustering
  - K-means
  - Agglomerative
  - Gaussian Mixture

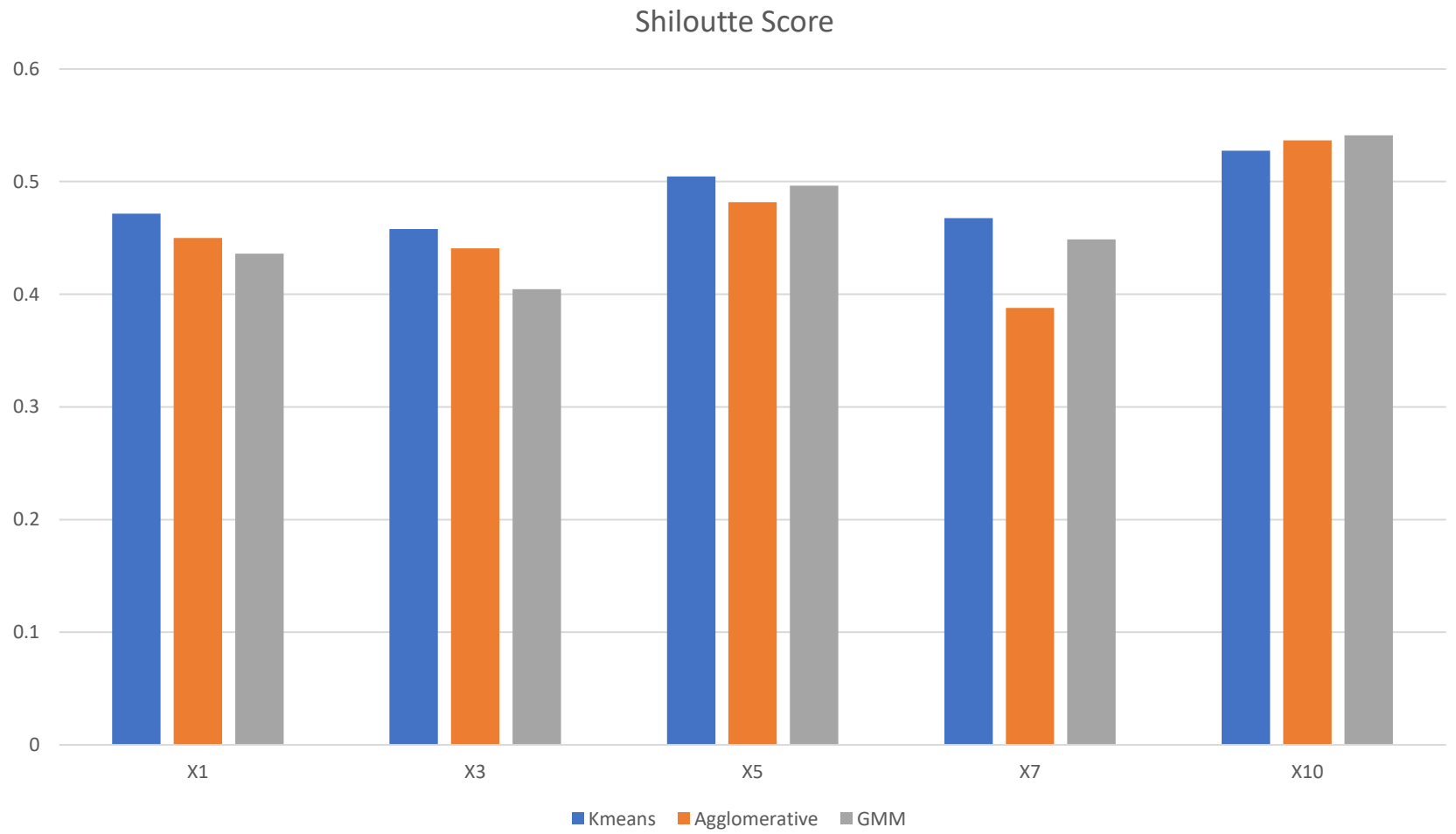
- Clustering labeling
  - Voting algorithm

Voting based on mean value of each feature



Using clusters in “3 days” group as a example.

Kmeans has best clustering performance



Voting for shill bidding

Durations	1 day	3 days	5 days	7 days	10 days	Total
Records	1289	1408	1060	2427	137	6321
Clusters	10	10	10	10	8	48
Normal	866	1026	640	1573	83	4188
Shill	120	58	74	122	6	380
Shill %	9%	4%	7%	5%	4%	6%
Suspicious	303	324	346	732	48	1753

	Bidder_Tendency	Bidding_Ratio	Successive_Outbidding	Last_Bidding	Auction_Bids	Starting_Price_Average	Early_Bidding	Winning_Ratio
13	0.041322	0.208333	1.0	0.286045	0.250000	0.000000	0.286025	0.818182
6231	0.166667	0.076923	0.0	0.253399	0.000000	0.000000	0.253399	0.000000
6285	0.281690	0.740741	1.0	0.651391	0.333333	0.993593	0.651339	1.000000

Observations and discussions

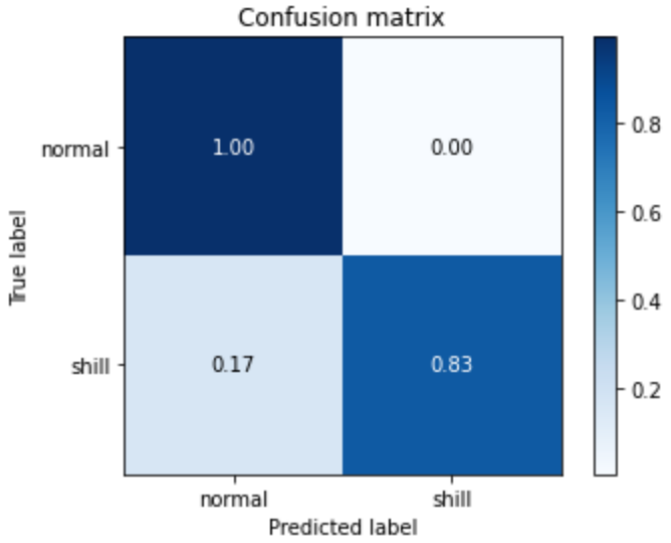
- 6% bidding are recognized as shill bidding using unsupervised clustering

- Shill bidding is more detected in short bidding duration
  - the longer the bidding durations, the easier to mimic the shill bidding?

- Shill bidding is tricky to detect, noticing we still have 28% suspicious biddings
  - Including more information should be helpful
  - Classification using supervised modeling

Classification using supervised learning

- Using recognized normal and shill bidding for modeling
- Predicting shill bidding in the suspicious records
  - Precision for normal bidding is 100%
  - Precision for shill bidding is 83%



	Total	Normal	Shill	Suspicious
KMeans	6321	4188	370	1753
SVM	6321	5732	569	0

Conclusion

- After combining unsupervised clustering and supervised modeling, I can detect 9% of biddings on eBay could be shill bidding

- After being fed more data, this model should be able to predict shill bidding on eBay more efficiently