ISSN-2320-7884 (ONLINE)
ISSN-2321-0257 (PRINT)

Machine Learning Approach to Handle Fraud Bids

D.S.L.Manikanteswari [1], M.Swathi [2], M.V.S.S.Nagendranath [3]

[1] Student, Sasi Institute of Technology and Engineering, Tadepalligudem,W.G(dt)

[2] Asst.professor, Sasi Institute of Technology and Engineering,Tadepalligudem,W.G(dt)

[3] Assoc.professor, Sasi Institute of Technology and Engineering, Tadepalligudem,W.G(dt)

Abstract: Online auctions propelled ecommerce activities since they remove the limitations of traditional auctions such as location, presence, time, space, and a small target audience. As of 2013, top two players of the online auction industry alone contributed 28% to the overall ecommerce finances. This increase in popularity has also opened doors for unlawful activities within an auction process. Existing counter measures such as seller validations and activity classification using human experts is not efficient enough with around 40% success rate. So we propose a Bayesian driven online model framework for the binary response. The framework is equipped with a well-known technique in statistical literature called the stochastic search variable selection (SSVS), to handle the dynamic evolution of the current activities with respect to prior activities of the seller. Involving selection bias process during classification can be helpful to classify both online and offline models effectively.

Keywords: Online Auction, Fraud Detection, Online Modeling, Online Feature Selection, Multiple Instance Learning.

I INTRODUCTION

Ecommerce in short for Electronic commerce consists of the buying and selling of products or services over electronic systems such as the Internet and other computer networks. The amount of trade conducted electronically has grown extraordinarily with widespread Internet usage. Electronic commerce that is conducted between businesses is referred to as business-to-business (B2B) or electronic commerce that is conducted between businesses and consumers, on the other hand, is referred to as business-to-consumer (B2C). Online auctions propelled ecommerce activities since they remove the limitations of traditional auctions such as location, presence, time, space, and a small target audience.

Figure 1 : List Of Top Auction Houses with their market cap[1]

Most Actives (dollar volume)		Last Trade	Change	Mkt Cap
eBay Inc	EBAY	57.38	+0.57 (1.00%)	74.47B
Amazon.com, Inc.	AMZN	308.69	+1.82 (0.59%)	140.53B
Mercadolibre Inc	MELI	109.50	+3.79 (3.59%)	4.83B
Sothebys	BID	40.77	+0.20 (0.49%)	2.78B
E Commerce China Dangdang	DANG	7.64	+0.64 (9.14%)	1.02B

Figure 2 : Industry Statistics and Market Size as of 2013[2]

As of 2013, top two players of the online auction industry alone contributed 28% to the overall ecommerce finances. Online Auction houses like ebay(e.g., an on-line trading/auctioning platform) is a market place that hosts both buyers and sellers who happens to be the varying entities in this virtual marketplace. There are two types of sellers: honest sellers always deliver high-quality products, whereas strategic sellers choose between delivering high or low quality products. Buyers are heterogeneous in the valuation of high-quality products; There are two types of sellers: honest sellers always deliver high-quality products, whereas strategic sellers choose

ISSN-2320-7884 (ONLINE)
ISSN-2321-0257 (PRINT)

between delivering high or low quality products. Sellers are heterogeneous in entry cost and production cost. If a seller delivers a low-quality product and the buyer reports it to the platform, the platform can reimburse the buyer and penalize the seller. Similar to any platform supporting financial transactions, online auction attracts criminals to commit fraud. The varying types of auction fraud are as follows.

- Products purchased by the buyer are not delivered by the seller.
- The delivered products do not match the descriptions that were posted by sellers.
- Malicious sellers may even post non-existing items with false description to deceive buyers, and request payments to be wired directly to them via bank-to-bank wire transfer.
- Furthermore, some criminals apply phishing techniques to steal high-rated seller's accounts so that potential buyers can be easily deceived due to their good rating.

Victims of fraud transactions usually lose their money and in most cases are not recoverable. As a result, the reputation of the online auction services is hurt significantly due to fraud crimes. Existing Counter measures includes the following:

- Seller identity verification through email, SMS, or phone.
- Setting up a rating system where buyers provide feedbacks, commonly used in e-commerce sites so that fraudulent sellers can be caught immediately after the first wave of victim complaints.
- Proactive moderation systems that allow human experts to manually investigate suspicious sellers or buyers.

Although the results are satisfactory the implementation of the moderation systems using manual approaches is tedious and needs intelligent automation.

II RELATED WORK

Online auction fraud is a major threat to ecommerce survival. There are articles on websites to teach people how to avoid online auction fraud (e.g. [3, 4]). [5] Categorizes auction fraud into several types and proposes strategies to fight them. Reputation systems are used extensively by websites to detect auction frauds, although many of them use naive approaches. [6] Summarized several key properties of a good reputation system and also the challenges for the modern reputation systems to elicit user feedback. Other representative work connecting reputation systems with online auction fraud detection include [7, 8, 9], where the last work [9] introduced a Markov random field model with a belief propagation algorithm for the user reputation.

Other than reputation systems, machine learned models have been applied to moderation systems for monitoring and detecting fraud. [10] proposed to train simple decision trees to select good sets of features and make predictions. [23] developed another simple approach that uses social network analysis and decision trees. Other approaches proposed an offline logistic regression modeling framework for the auction fraud detection moderation system which incorporates domain knowledge such as coefficient bounds and multiple instance learning.

In this paper we treat the fraud detection problem as a binary classification problem. The most frequently used models for binary classification include logistic regression [11], probit regression [12], support vector machine (SVM) [13] and decision trees [14]. Feature selection for regression models is often done through introducing penalties on the coefficients. Typical penalties include ridge regression [34] (L2 penalty) and Lasso (L1 penalty). Compared to ridge regression, Lasso shrinks the unnecessary coefficients to zero instead of small values, which provides both intuition and good performance. Stochastic search variable selection (SSVS) uses "spike and slab" prior so that the posterior of the coefficients have some probability being 0. Another approach is to consider the variable selection problem as model selection, i.e. put priors

ISSN-2320-7884 (ONLINE)
ISSN-2321-0257 (PRINT)

on models (e.g. a Bernoulli prior on each coefficient being 0) and compute the marginal posterior probably of the model given data. People then either use Markov Chain Monte Carlo to sample models from the model space and apply Bayesian model averaging, or do a stochastic search in the model space to find the posterior mode. Among non-linear models, a tree model usually handles the non-linearity and variable selection simultaneously. Representative work includes decision trees, random forests, gradient boosting and Bayesian additive regression trees (BART). Online modeling (learning) considers the scenario that the input is given one piece at a time, and when receiving a batch of input the model has to be updated according to the data and make predictions and servings for the next batch. The concept of online modeling has been applied to many areas, such as stock price forecasting, web content optimization, and web spam detection. Compared to offline models, online learning usually requires much lighter computation and memory load; hence it can be widely used in real-time systems with continuous support of inputs. For online feature selection, representative applied work include [11] for the problem of object tracking in computer vision research, and for content-based image retrieval. Both approaches are simple while in this paper the embedding of SSVS to the online modeling is more principled. Multiple instance learning, which handles the training data with bags of instances that are labeled positive or negative, is originally proposed by [12]. Many papers have been published in the application area of image classification. The logistic regression framework of multiple instance learning is presented in [12], and the SVM framework is presented in [13].

III PRELIMINARIES

ACTION/ PROBLEM AFFECTING TRUST OF ECOMMERCE ENTITIES

Problem/ action which affects	Trust factor of	Trust on
seller provides wrong detail or wrong product itself	buyer	seller & ecommerce service provider
buyer creates problem in payment part	seller	buyer & ecommerce service provider
security or service issue or problem	buyer & seller	ecommerce service provider

No matter whose action is affecting trust of whom, most hampered is success of ecommerce. In another words trust in ecommerce has to enhance or improve in order to wide acceptability of ecommerce instead of wide spread. Due to the limited expert human resources, only around 40% of the cases can be reviewed and labeled leading to inefficient handling. Since it is necessary to develop an automated pre-screening moderation system that separates suspicious cases from all cases for expert inspection. The moderation system using machine-learned models is proven to improve fraud detection significantly compared to prior approaches. Machine-learned models are classified into two types

- Offline models
- Online models

Offline models are constructed by using the previous 30 days transactional data to serve the next day. Since the response is binary (fraud or non-fraud) and the scoring function has to be linear, logistic regression is used. Applying expert knowledge, such as bounding the rule based feature weights to be positive and multiple-instance learning, can significantly improve the performance in terms of detecting more frauds and reducing customer complaints given the same workload from human experts. However, offline models often meet the following challenges:

- Large amount of historical training data is required since offline models tend to be fairly unstable compared to online models.
- Since the fraudulent sellers change their pattern very fast, it requires the model to also evolve dynamically which is again non-trivial for offline models compared to online models

Also, since the training data is from human labeling, the high cost makes it almost impossible to obtain a very large sample for offline models.

IV PROBIT FRAMEWORK

Our application is to detect online auction frauds of an auctioning site where new auction cases are posted every day. Every new case is sent to our proactive

ISSN-2320-7884 (ONLINE)
ISSN-2321-0257 (PRINT)

anti-fraud moderation system for a pre-screening score to assess the risk. The current system is featured by:

- Rule-based features: Human experts with years of experience created many rules to detect whether a user is fraud or not. An example of such rules is "blacklist", i.e. whether the user has been detected or complained as fraud before. Each rule can be regarded as a binary feature that indicates the fraud likeliness.

- Linear scoring function: The existing system only supports linear models. Given a set of coefficients (weights) on features, the fraud score is computed as the weighted sum of the feature values.

- Selective labeling: If the fraud score is above a certain threshold, the case will enter a queue for further investigation by human experts. Once it is reviewed, the final result will be labeled as boolean, i.e. fraud or clean. Cases with higher scores have higher priorities in the queue to be reviewed. The cases whose fraud score are below the threshold are determined as clean by the system without any human judgment.

- Fraud churn: Once one case is labeled as fraud by human experts, it is very likely that the seller is not trustable and may be also selling other frauds; hence all the items submitted by the same seller are labeled as fraud too. The fraudulent seller along with his/her cases will be removed from the website immediately once detected.

- User feedback: Buyers can file complaints to claim loss if they are recently deceived by fraudulent sellers.

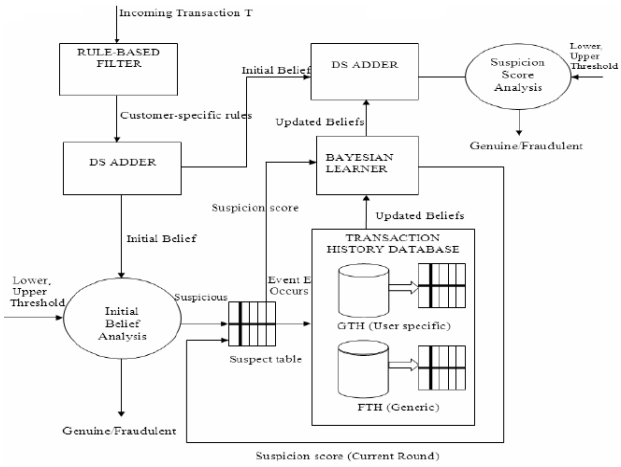


Figure 4.1. Block diagram of the proposed fraud detection system

Using these specific attributes in our proactive moderation system for fraud detection, we build our Bayesian online modeling framework with details of model fitting via Gibbs sampling. Also extending it features with a selection bias fitting increases its adaptation to offline models too. The selection bias fitting approach is represented in Initial Belief Analysis.

Online Probit Regression

Consider splitting the continuous time into many equal- size intervals. For each time interval we may observe mul- tiple expert-labeled cases indicating whether they are fraud or non-fraud. At time interval t suppose there are n_t obser- vations. Let us denote the i-th binary observation as y_{it}. If $y_{it} = 1$, the case is fraud; otherwise it is non-fraud. Let the feature set of case i at time t be x_{it}. The probit model [3] can be written as

$$P[y_{it} = 1 | x_{it}, \beta_t] = \Phi(x' \beta_t),$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution $N(0, 1)$, and β_t is the unknown regression coefficient vector at time t. Through data augmentation the probit model can be ex- pressed in a hierarchical form as follows: For each observation i at time t assume a latent random variable z_{it}. The binary response y_{it} can be viewed as an indicator of whether $z_{it} > 0$, i.e. $y_{it} = 1$ if and only if $z_{it} > 0$. If $z_{it} <= 0$, then $y_{it} = 0$. z_{it} can then be modeled by a linear regression

ISSN-2320-7884 (ONLINE)
ISSN-2321-0257 (PRINT)

$$z_{it} \sim N(x_{it}' \beta_t, 1)$$

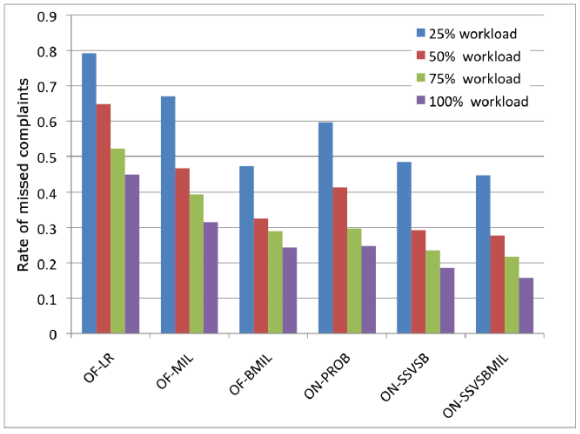
In a Bayesian modeling framework it is common practice to put a Gaussian prior on β_t,

$$\beta_t \sim N(\mu_t, \Sigma_t),$$

where μ_t and Σ_t are prior mean and prior covariance matrix respectively.

V PERFORMANCE

In this paper we adopt an evaluation metric introduced that directly reflects how many frauds a model can catch: the rate of missed complaints, which is the portion of customer complaints that the model cannot capture as fraud. Note that in our application, the labeled data was not created through random sampling, but via a pre-screening moderation system using the expert-tuned coefficients. This in fact introduces biases in the evaluation for the metrics which only use the labeled observations but ignore the unlabeled ones. This rate of missed complaints metric however covers both labeled and unlabeled data since customers do not know which cases are labeled, hence it is unbiased for evaluating the model performance. A comparative results chart differentiating our approach(ON-SSVSBMIL) and prior approaches validates our claim.



Model	Rate of Missed Complaints	Batch Size	Best δ
Expert	0.3466	–	–
OF-LR	0.4479	–	–
OF-MIL	0.3149	–	–
OF-BMIL	0.2439	–	–
ON-PROB	0.2483	Day	0.7
ON-SSVSB	0.1863	Day	0.7
ON-SSVSBMIL	0.1620	Day	0.7
ON-SSVSBMIL	0.1330	1/2 Day	0.8
ON-SSVSBMIL	0.1508	1/4 Day	0.9
ON-SSVSBMIL	0.1581	1/8 Day	0.95

Table 5.1: The rates of missed customer complaints for all the models given 100% workload rate.

VI CONCLUSION

We proposed and built an proactive model framework for handling the auction fraud moderation and detection system designed for typical online auction website. By empirical experiments on a real world online auction auction events fraud data, we showed that our proposed online probit model framework, which combines online feature selection, bounding coefficients from expert knowledge, selective biasing and multiple instance learning yeilds results that has significant performance gains over rule based baselines or the manual human-tuned models. Note that this online modeling framework can be easily extended to many other applications, such as web spam detection, content optimization and so forth which can be regarded as future research to generalize the framework.

VII REFERENCES

[1]http://www.ibisworld.co.uk/market-research/e-commerce-online-auctions.html?partnerid=prweb

[2]http://www.google.com/finance?catid=us-TRBC:5720103013&ei=zIDSUfiZFJ6wlgOABQ

[3]USA Today. How to avoid online auction fraud. http://www.usatoday.com/tech/columnist/2002/05/07/yaukey.htm, 2002.

ISSN-2320-7884 (ONLINE)
ISSN-2321-0257 (PRINT)

[4] Federal Trade Commission. Internet auctions: A guide for buyers and sellers. http://www.ftc.gov/bcp/ conline/pubs/online/auctions.htm, 2004.

[5] C. Chua and J. Wareham. Fighting internet auction fraud: An assessment and proposal. Computer, 37(10):31–37, 2004.

[6] P. Resnick, K. Kuwabara, R. Zeckhauser, and E. Friedman. Reputation systems. Communications of the ACM, 43(12):45–48, 2000.

[7] P. Resnick, R. Zeckhauser, J. Swanson, and K. Lockwood. The value of reputation on ebay: A controlled experiment. Experimental Economics, 9(2):79–101, 2006.

[8] D. Gregg and J. Scott. The role of reputation systems in reducing on-line auction fraud. International Journal of Electronic Commerce, 10(3):95–120, 2006.

[9] H. Chipman, E. George, and R. McCulloch. Bart: Bayesian additive regression trees. The Annals of Applied Statistics, 4(1):266–298, 2010.

[10] D. Chau and C. Faloutsos. Fraud detection in electronic auction. In European Web Mining Forum (EWMF 2005), page 87.

[11] P. McCullagh and J. Nelder. Generalized linear models. Chapman & Hall/CRC, 1989.

[12] C. Bliss. The calculation of the dosage-mortality curve. Annals of Applied Biology, 22(1):134–167, 1935.

[13] N. Cristianini and J. Shawe-Taylor. An introduction to support Vector Machines: and other kernel-based learning methods. Cambridge university press, 2006.

[14] J. Quinlan. Induction of decision trees. Machine learning, 1(1):81–106, 1986.