Review of: "Predicting Mobile Money Transaction Fraud using Machine Learning Algorithms"

Siyu Zhou[1]

1 University of Pittsburgh

Potential competing interests: No potential competing interests to declare.

This manuscript gives a good introduction on mobile money service, mobile money fraud and the promising use of machine learning methods. However, the statistical analysis is not clear.

1. In the section for reviewing machine learning methods, gradient descent is listed in parallel to logistic regression, decision tree and random forests. However, gradient descent is an optimizing algorithm that can be incorporated in many models. The author mentioned the use of a probabilistic formula in Sec 2.2.3, but that is just one example. Gradient boosting is another model that makes use of gradient descent and decision trees.
2. Precision and recall are not correctly explained. The true positive rate (TPR) and the false positive rate (FPR) are not defined.
3. In section 4.1, the author gives a description of the data and the implication. Considering the skewness in these variables, for the related discussion, it is better to use statistics robust to outliers or extreme values such as median and inter-quartile range instead of mean and standard deviation.
4. Regarding the analysis in Sec 4.2
    1. It is unclear how the training and test sets are divided.
    2. Feature space is not rich enough. It may be good to explore more possible valuable features
    3. It is unclear how tuning of these models are carried out, or whether they are tuned.
    4. Some existing features, such as oldbalanceOrg and newbalanceOrg, seem not to be used in logistic regression.
    5. To interpret any coefficient of some feature in logistic regression, it should be emphasized that all other features are hold fixed. Apart from this, a decrease of 0.5 in odds should not be interpreted as 50% decrease.
    6. Features with p-value larger than 0.05 in the logistic regression are not statistically significant when other features are included in this logistic model, but they can be useful for other models. It can be beneficial to include them for other models.
5. Feature importance in Sec 4.3 is not clearly defined.
6. Other minor comments
    1. Cross references for equations and tables are unclear. Several equations are labelled as Equation 1 and several tables are labelled as Table 1. Captions are missing.
    2. Equation 2 should be improved so that the exponential term is clear.