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

Guoping Zeng

Potential competing interests: No potential competing interests to declare.

This paper uses various machine learning classifiers to predict transactions flagged as fraud in mobile money transfers. Since data for this paper came from real-time transactions that stimulate a well-known mobile transfer fraud scheme, this paper has a merit in banking and finance industry. To be more convincing, the author is suggested to add some theatrical comparisons beyond simulation comparisons.

As overall comments, equations should be centered on the line and the number label enclosed in parentheses. When referencing an equation in text, simply refer to it using the same label, also enclosed in parentheses.

In the following, please find my comments section by section.

Section 2.2.1. Logistic Regression

Descriptions preceding eq. 1 and after "Where": What do you mean by "derivative"?

"e is the mathematical constant" is better be "e is the Euler constant".

Descriptions preceding eq. 2 and after "Where":

"Y= values between 0 and 1": Y is discrete. It has only 2 values 0 and 1.

"eβ0+β1+X represents the independent features, and": The two "+" should not appear as a subscript but the addition operator. The Same applies to eq. 2.

Why are there only 2 coefficients β0 and β1? Unless it is a simple logistic regression with only one independent variable, it should have (p + 1) coefficients for p independent variables.

When you mention drawbacks of logistic regression in the last paragraph of 2.2.1, you may need to add separation 9complete separation and quasi-separation) and multicollinearility. Please refer to the following paper:

Zeng G, Zeng E. (2021). On the Relationship between Multicollinearity and Separation in Logistic Regression. Communications in Statistics - Simulation and Computation. 50(7): 1989-1997. doi: 10.1080/03610918.2019.1589511

2.2.2. Decision Tree

Only Gini Index (Impurity function) is considered. You may also add Entropy (Impurity function).

Description preceding eq. 4 and after "Where": k in fk should appear as a subscript.

You may add explanation for m such as "m is the number of levels of the dependent variable Y".

2.2.3. Gradient Descent

0 and 1 in 0 and 1 should appear as a subscript.

3.2. Data Description and Variables

"target variables" before Table 1 should be singular not plural. You should keep the term consistent. So, target is the

same as dependent.

Table 1

I do not understand how nameOrig (Customer who started the transaction) can be a continuous variable. Same for nameDest (Recipient of transaction).

The equation at the end of this section should be aligned. The eq. 1 should be eq. 5.

Do you use any categorical variables? If yes, you need to convert them into continuous ones for logistic regression. You may refer to the following paper:

Zeng G. (2021). On the analytical properties of category encodings in logistic regression. Communications in Statistics - Theory and Methods, Advance online publication. doi: 10.1080/03610926.2021.1939382

3.3. Data Cleaning and Preprocessing

Feature isFlaggedFraud is not in Table 1.

3.3.1. Feature Scaling

The normalization in eq. 5 is unreadable with numerator X - X (= 0). The mean in the denominator should be min. So, the normalization should be like

(X - min(X))/ (max(X) - min(X))

3.3.3. Fraud Model Performance Measures

For the concept of confusion matrix, you may refer to the following paper:

Zeng, G. (2020). On the Confusion Matrix in Credit Scoring and Its Analytical Properties. Communications in Statistics - Theory and Methods, 49(9), 2080-2093. doi: 10.1080/03610926.2019.1568485

4.1. Descriptive Results

The table should be labeled as Table 3 and explained.

4.2.1. Baseline Logistic Regression

The logit regression results after the first paragraph should be labeled. Variables such as type_CASH_IN, type_DEBIT and type_PAYMENT have very large P-Value. You should drop them.

The referred Table 1 in this section should be for the logit regression results.

Table 1 in Section 3.2 has nothing to do with the logit regression results.

The second table in this section (about Features, Odds, Change_Odds%) should be labeled.

The 3rd table in this section should also be labeled.

4.2.2. Performance Accuracy and Matthew Correlation Coefficient

Please label the table at the end of this section.

4.2.3. Classification Measures

The table for Classification Metrics needs to be labeled.

The referred Table 1 in this section is again confusing. It does not seem to be Table 1 in Section 3.2.

Please also label the decision tree figure at the end of this section.

4.2.4. The ROC Curve

Please label the ROC curve at the end of this section.

4.3. Feature Importance

Please label the figure for feature importance at the end of this section.