Title: Development of a Transaction Flagging System for Fraudulent Transactions Exceeding €1000

Author: Jan Strydom

Abstract

This report explores the development of a system to flag fraudulent transactions over €1000. Utilizing a dataset of credit card transactions, various data mining techniques were applied to identify key variables that indicate fraudulent activity. The focus is on selecting simple, linearly related variables suitable for a regulatory-compliant, logic-based flagging system.

Introduction

Fraudulent transactions pose significant risks to financial institutions and their clients. This project addresses the need to flag transactions exceeding €1000 for potential fraud. Data-driven approaches explored identify key indicators of fraud, emphasizing simplicity and regulatory compliance. This approach contrasts with more complex, often opaque machine learning models, aligning with the need for transparent and explainable financial practices.

Methodology

Data

The dataset, 'creditcard.csv', comprises various features of credit card transactions. Our analysis focuses on transactions above €1000, recognizing higher value transactions' susceptibility to fraud. Variables of interest are given without any identifying characteristics and named “V1” through ”V28”. Class imbalance was found to not be an issue in the dataset.

Methods

Three distinct models were utilized in this project: Logistic Regression, Random Forest, and XGBoost. Each model was chosen for its unique approach to data analysis, ranging from the simplicity and interpretability of Logistic Regression to the more complex ensemble methods employed by Random Forest and XGBoost. The rationale behind this selection strategy was to capture a comprehensive range of insights from the dataset, particularly focusing on key variables indicative of fraudulent transactions over €1000. Here's an expanded explanation incorporating the use of R and the relevant packages:

Logistic Regression: Emphasis on Simplicity and Interpretability

Due to its straightforwardness in modeling binary outcomes, Logistic Regression was selected as a baseline model. This approach is advantageous for its interpretability, which is crucial in the context of financial regulations. The `glm` function in R, a part of the base package, was employed for implementing logistic regression. Its robustness and ease of use make it suitable for binary classification tasks in financial data analysis.

Random Forest: Robustness through Ensemble Learning

Random Forest was chosen for its method of creating multiple decision trees and combining their outcomes to improve prediction accuracy. This technique enhances the model's robustness, enabling it to capture complex patterns in data more effectively.

The implementation of Random Forest was facilitated by the `randomForest` package in R, offering extensive functionality for model tuning and evaluation.

XGBoost: Advanced Performance through Gradient Boosting

For handling large datasets and complex variable interactions efficiently, XGBoost was selected. Known for its high performance in classification tasks, XGBoost is adept at feature selection and model optimization.

The implementation of XGBoost in this project was done using the `xgboost` package in R, which allows for extensive model parameter tuning.

Gathering Insights from Diverse Models

The use of these models was intended to provide a spectrum of insights into the dataset. Logistic Regression provides a linear perspective, while Random Forest and XGBoost offer more complex, non-linear views, ensuring a thorough analysis of potential fraud indicators.

Incorporation of Additional R Packages

Additional packages such as `pdp` for Partial Dependence Plots and `plot` for data visualization were also utilized. These tools were crucial in interpreting and visualizing the models' outcomes, enhancing the clarity and comprehensiveness of the findings.

Results and Discussion

Logistic Regression Analysis

The logistic regression model, despite convergence challenges, identified [V9] as a significant predictor of fraud. Only one variable exhibited a linear relationship with the fraud outcome. The rest of the variables were found to have non-linear relationships with the fraud outcome.

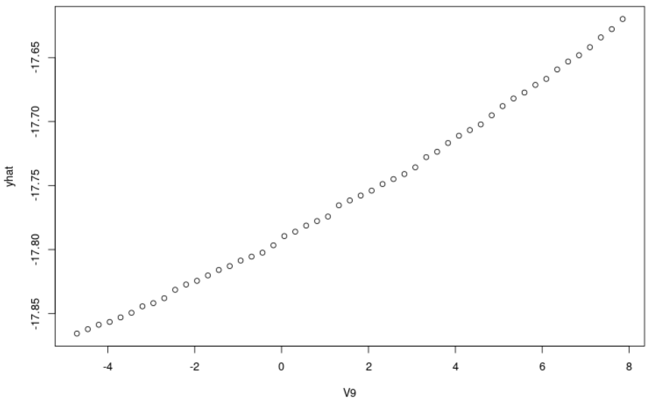

Figure 1: V9 Identified as having a linear relationship with the prediction [Logistic Regression].

Random Forest Analysis

Random Forest, an ensemble learning method, builds multiple decision trees and merges their outcomes to improve prediction accuracy. This method contrasts with logistic regression's single-model approach. By aggregating the predictions of numerous trees, Random Forest effectively captures complex patterns and interactions among variables that might be overlooked by logistic regression. The variables [V20], [V9], and [V4] were identified as significant in the Random Forest model due to the algorithm's ability to assess feature importance. In Random Forest, variable importance is gauged by measuring how much each variable improves the purity of the node splits in the decision trees. This is a different approach compared to logistic regression, which evaluates variables based on their coefficients in a linear equation. The overlap observed with logistic regression variables suggests that these variables are robust indicators across different model types. However, the unique variables identified by Random Forest highlight its ability to uncover additional layers of information, potentially offering a more comprehensive understanding of factors contributing to transaction fraud.

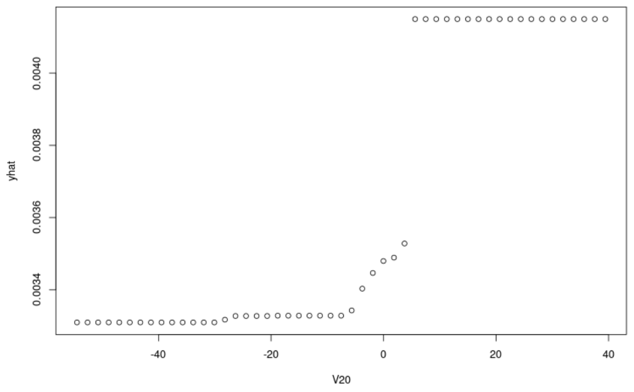

Figure 2: V20 Identified as having a linear relationship with the prediction [Random Forest].

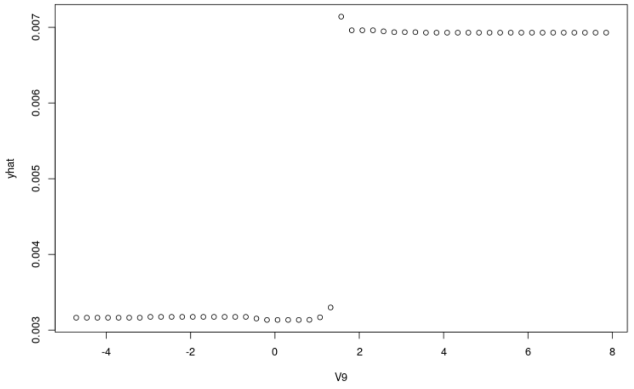

Figure 3: V9 Identified as having a linear relationship with the prediction [Random Forest].

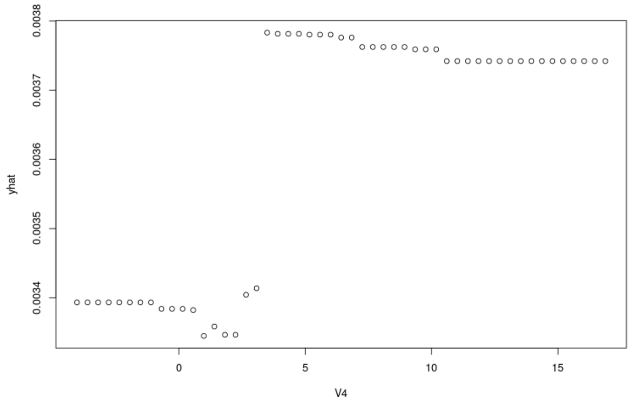

Figure 4: V4 Identified as having a linear relationship with the prediction [Random Forest].

XGBoost Analysis

XGBoost employs gradient boosting, an advanced ensemble technique that sequentially builds models, with each new model focusing on the errors of the previous one. This iterative refinement helps in identifying subtle, yet important patterns in the data, potentially leading to the selection of variables like [V26], [V19], and [V18]. To ensure these variables conform to the project's requirements of simplicity and linearity, partial dependence plots are employed. PDPs visualize the marginal effect of a variable on the predicted outcome, holding other variables constant. This analysis helps in understanding whether the relationship between these variables and the occurrence of fraud is straightforward and linear, as desired for the fraud flagging system. The combination of their effectiveness in the XGBoost model and the clarity in their impact on fraud prediction, as evidenced by the PDPs, makes these variables excellent candidates for the fraud flagging system. Their selection is grounded not just in their predictive power within a sophisticated machine learning model, but also in their ability to fit into a system that values simplicity and transparency.

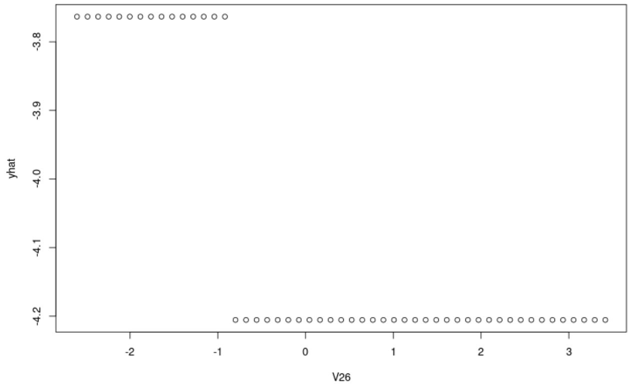

Figure 5: V26 Identified as having a linear relationship with the prediction [XGBoost].

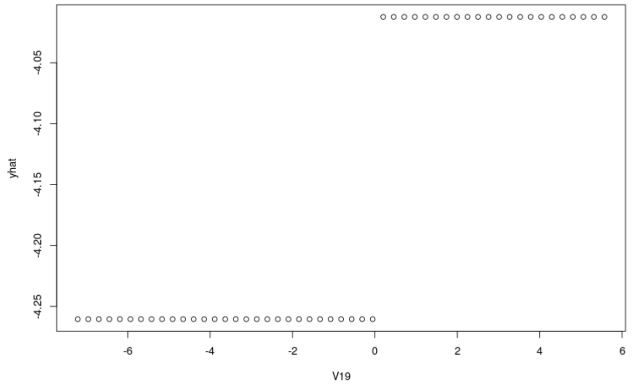

Figure 6: V19 Identified as having a linear relationship with the prediction [XGBoost].

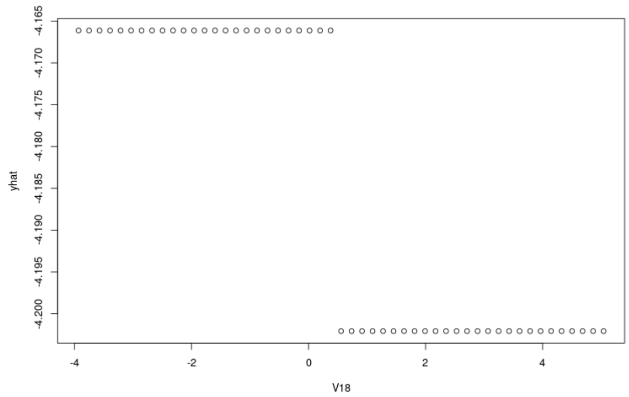

Figure 7: V18 Identified as having a linear relationship with the prediction [XGBoost].

Conclusion

Based on the analysis from the three models, we propose a set of variables for the transaction flagging system. Using the Random Forest Model, three variables [V20,V9,V4] were found to be good candidates for an early fraud flagging system. In addition, the three variables [V19,V18,V26] were found to be a good fit for the project's needs given by the XGBoost model. Refer to the following section regarding the exclusion of the logistic regression model.

Comparative Analysis

It was difficult to simply use the logistic regression model parameters and coefficients as an indicator of fraud, as only one of the variables showed a clear linear relationship. As the project's needs dictate a variable with a linear relationship with the predicted class, the use of the logistic regression model was excluded.

Please note that references were used purely to gain an understanding of existing approaches to fraud detection.

References

Soltani, M., Kythreotis, A., & Roshanpoor, A. (2023). "Two decades of financial statement fraud detection literature review; combination of bibliometric analysis and topic modeling approach." Journal of Financial Crime, Vol. 30 No. 5, pp. 1367-1388.

"Financial Fraud Detection Based on Machine Learning: A Systematic Literature Review" by Abdulalem Ali, Shukor Abd Razak, Siti Hajar Othman, Taiseer Abdalla Elfadil Eisa, Arafat Al-Dhaqm, Maged Nasser, Tusneem Elhassan, Hashim Elshafie, and Abdu Saif.