Paper 4218-2020

The Seven Most Popular Machine Learning Algorithms for Online Fraud Detection and Their Use in SAS®

Patrick Maher, SAS Institute, Inc.

ABSTRACT

Today, illegal activities regarding online financial transactions have become increasingly complex and borderless. This unlawful activity results in substantial economic losses for both customers and organizations. Many techniques have been proposed for fraud prevention and detection in the online environment. All these techniques have the same goal of identifying and combating fraudulent online transactions. However, each machine learning technique comes with its characteristics, advantages, and disadvantages. This session reviews the use of the most common machine learning algorithms used in online fraud detection, the strengths and weaknesses of these techniques, and how these algorithms are developed and deployed in SAS®. Types of fraud discussed include credit card fraud, financial fraud, and e-commerce fraud. Algorithms reviewed include neural networks, decision trees, support vector machines, K-nearest neighbor, logistic regression, random forest, and naïve Bayes.

INTRODUCTION

Fraud is a widespread and increasing issue in online transactions. These online transactions include credit card payments, financial fraud, and e-commerce/retail fraud. According to a global survey by the Association of Certified Fraud Examiners (ACFE) and SAS, nearly three-quarters of organizations (72%) are projected to use automated monitoring, exception reporting, and anomaly detection by 2021. Specific to the machine learning algorithms discussed in this paper, about half of organizations anticipate employing machine learning algorithms/predictive modeling (52%; up from 30%) (SAS Institute, Inc., 2019).

This paper examines several machine learning algorithms that can be used to detect anomalies indicative of potential fraud. This paper will discuss how these models work and compare their use on a publicly available data set constructed for fraud detection.

APPROACHES TO MODELING

When trying to identify fraud with machine learning, two approaches are commonly used. One approach is to use methods associated with unsupervised machine learning. These machine learning methods are referred to as unsupervised because there is no historical data about fraudulent cases used to train the model. Unsupervised methods are used to find outliers by locating observations within the data set that are separated from other densely populated areas of the data set (Gillespie, 2019). Machine-learning techniques intensively use math statistics, as well as knowledge and results from fields such as mathematics, psychology, neurobiology, information technology (Minastireanu & Mesnita, 2019). Supervised learning involves using historical data that contains examples of the type of fraud that the user is trying to find. The algorithm can then learn to detect the potentially fraudulent event by training a model using the examples of fraudulent and non-fraudulent cases (Gillespie, 2019).

MACHINE LEARNING ALGORITHMS

The machine learning algorithms selected for this paper represent some of the most commonly implemented techniques used in online fraud detection. Both recent academic research (Minastireanu & Mesnita, 2019) and industry publications (Mayenberger, 2019) highlight the algorithms covered in this paper. The unique twist is that we are doing this in SAS in a graphical user interface.

Neural network

Neural networks were initially developed by researchers who were trying to mimic the neurophysiology of the human brain. By detecting complex nonlinear relationships in data, neural networks can help make predictions about real-world problems.

Strengths: Neural networks can model non-linear and complicated relationships; no assumptions on input variables.

Weaknesses: It can be challenging to interpret the results of a neural network.

Decision tree

Tree models are built from training data for which the response values are known, and these models are subsequently used to classify response values for new data (SAS Institute, Inc., 2015). The leaves of binary trees produce output that is discrete for a classification tree.

Strengths: One of the most straightforward models one can use for binary targets or classification.

Weaknesses: Decision trees tend to overfit.

Support Vector Machines

Support vector machines are a geometric method of separating two classes by finding the best hyperplane that puts one class above it and the other below. A Support Vector Machine (SVM) model is a supervised machine-learning method that is used to perform classification and regression analysis.

Strengths: High accuracy, ability to deal with high-dimensional data, ability to generate non-linear decision boundaries

Weaknesses: SVMs do not perform well with large data sets due to training time.

k Nearest Neighbor

Classifies new cases according to the k most similar (closest) cases in the training set. No model is learned from the training data. Learning occurs when a test case needs to be classified. kNN is considered a lazy learning method (as opposed to eager learning methods). Uses distance or similarity function (e.g., Euclidean distance) to determine which are the closest cases.

Strengths: Simple to implement; Flexible to feature/distance choices; Naturally handles multi-class cases

Weaknesses: The user needs to determine the number of nearest neighbors). Computation cost is quite high because we need to compute the distance of each query instance to all training samples (Subramanian, 2019).

Logistic regression

Logistic regression is used when the response variable is categorical. In many cases, the response variable is a binary, typically coded as 0 and 1, with 1 representing a positive

case. Logistic regression measures the relationship between the response variable and one or more independent variables.

Strengths: Widely used; highly interpretable; calculated relationship for each predictor variable

Weaknesses: Underlying math might be considered complicated.

Random forest

The Random Forest model is a predictive model that consists of several decision trees that differ from each other in two ways. First, the training data for a tree is a sample without replacement from all available observations. Second, the input variables that are considered for splitting a node are randomly selected from all possible inputs. In other respects, trees in a forest are trained like standard trees.

For many data sets, it produces a highly accurate classifier.

Strengths: A random forest is an ensemble model that improves predictive power and reduces over-fitting.

Weaknesses: Random forest models can be challenging to interpret.

Naïve Bayes

The basic idea is simple: Each input variable, all by itself, has something to say about whatever one is trying to predict. No relationship between input variables exists. All connections are between the target variable and the input variables. All input variables are assumed to be conditionally independent.

Strengths: Widely used; calculated relationship for each predictor variable

Weaknesses: Strong assumption that features are independent.

ALGORITHM PERFORMANCE

How do we identify a "good" model for use in fraud detection? Reviewing several research studies and working with our customers, we consider the area under the receiver operating curve (AUROC), the false-negative rate (FNR), and the false discovery rate (FDR).

Performance testing applies to all models to determine how good the model is performing in its objective. A commonly used metric for evaluating model performance is the area under the receiver operating curve (AUROC) that measures the probability that a randomly positive outcome will be ranked higher than a randomly chosen negative one (Mayenberger, 2019).

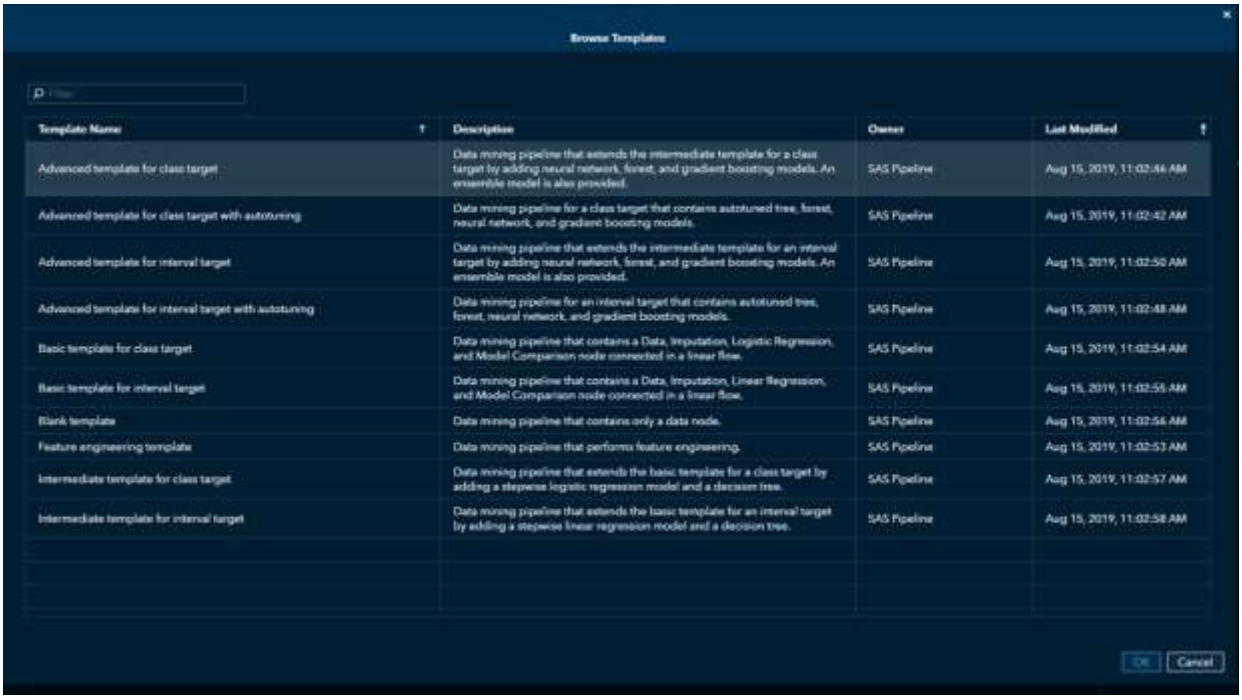
According to Mayenberger (2019), fraudulent transactions that are not detected are much more critical than legitimate transactions that are stopped. Therefore, an essential measure is the false-negative rate (FNR). The objective is to minimize the FNR. However, this will lead to a high false-positive rate (FPR). Mayenberger argues that the FPR must be limited to a certain tolerance level.

Another useful performance measure is the false discovery rate (FDR). The FDR is provided as assessment criteria in SAS Model Studio along with many other assessment measures. The FDR is the ratio of false positives overall positives (false and true). The FDR has gained notoriety in academic literature in the conversation over the fact that many studies that have been published have not been replicable in follow up research. The FDR provides a balance top-values (i.e., $p \leq .05$) that may be too lenient in testing. Colquhoun (2014) argues that maintaining $p = 0.05$ one will be wrong at least 30% of the time. He goes on to say that if one wishes to keep a false discovery rate below 5%, one would have to insist on $p \leq .001$ in testing. The FDR relates to our current topic in that many software defaults are

set at p = 0.05 for algorithm building and testing of significance. Consider this a topic for further discussion and research as it relates to fraud detection in general.

MACHINE LEARNING ALGORITHMS IN SAS

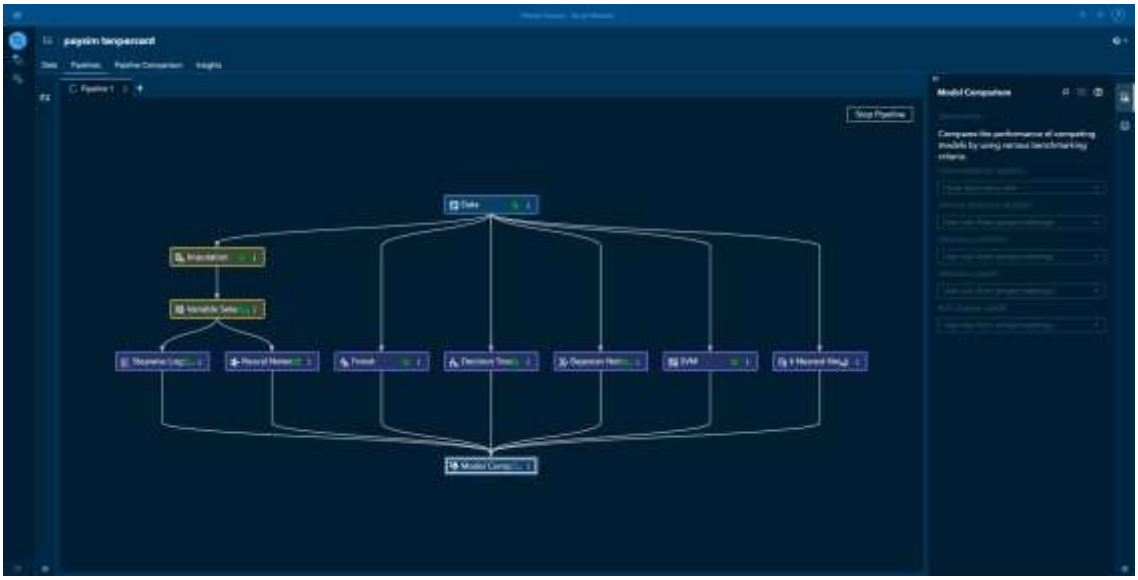
SAS Model Studio allows one to build each of these seven models within a drag and drop interface. Model Studio offers the user the choice of creating custom data process flows, also called pipelines, or selecting template flows. For this paper, an advanced template for a class target was chosen. By default, the advanced template adds an ensemble model choice, which will combine the best fit from multiple models. For this paper, we have excluded the ensemble model from the comparison of models.



Display 1. Template choices for SAS Model Studio

DATA

To demonstrate the use of these supervised machine learning techniques, we will evaluate them using a data set that contains synthetic payment transactions, known as PaySim (Lopez-Rojas, Elmir, & Axelsson, 2016). This data set contains transaction data for a variety of transactions, with 11 variables and 6,362,620 observations. We will use the 'isFraud' variable as a target for our supervised algorithms' ability to detect the fraudulent observations. As inputs to the model, we will use the amount, oldbalanceOrg, newbalanceOrig, oldbalanceDest, and newbalanceDest variables (Gillespie, 2019).



Display 2. Pipeline Interface for SAS Model Studio

RESULTS

The decision tree model was selected as the champion model for the seven supervised learning algorithms used in this example. This selection is based on the default criteria using the false discovery rate (FDR). In this example, we get an FDR of zero for a 10% event. Also, the decision tree has the best misclassification rate at .06. One might decide on using the decision tree based on this result. To further test this outcome, the models were run against a one percent sample (fraud occurring in 1 percent of cases). When the positive occurrence of fraud is rarer, as would be the case with real-world data, the decision tree shows a higher false discovery rate of .0167.

Display 3. Decision tree results in SAS Model Studio

CONCLUSION

This paper demonstrates the use of some of the most popular machine learning algorithms used in online fraud detection implemented in SAS. Our classification criteria were chosen based on some of the most common assessment criteria used in academia and by industry. The tools available within SAS offer these typical assessment criteria, along with other measures and visual assessment techniques. In this specific example, the decision tree algorithm offered the lowest misclassification rate along with the lowest false discovery rate (FDR).

REFERENCES

Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of p-values. *Royal Society Open Science, 1:140216*. Retrieved from rsos.royalsocietypublishing.org.

Gillespie, R. (2019). Detecting fraud and other anomalies using isolation forests. *Proceedings of SAS Global Forum 2019* (pp. 1-6). Dallas, TX: SAS Institute, Inc.

Lopez-Rojas, E., Elmir, A., & Axelsson, S. (2016). PaySim: A financial mobile money simulator for fraud detection. *The 28th European Modeling and Simulation Symposium-EMSS.* Larnaca, Cyprus.

Mayenberger, D. (2019). How to overcome modelling and model risk management challenges with artificial intelligence and machine learning. *Journal of Risk Management in Financial Institutions, 12, 3*, 241-255.

Minastireanu, E.-A., & Mesnita, G. (2019). An analysis of the most used machine learning algorithms for online fraud detection. *Informatica Economica*, 5-16.

SAS Institute, Inc. (2015, July). *SAS/STAT®14.1 users guide: The HPSPLIT procedure.* Retrieved from support.sas.com: https://support.sas.com/documentation/onlinedoc/stat/141/hpsplit.pdf

SAS Institute, Inc. (2019, June 24). *PRNewswire.* Retrieved from www.prnewswire: https://www.prnewswire.com/news-releases/study-ai-for-fraud-detection-to-triple-by-2021-300872958.html

SAS Institute, Inc. (2020, February 9th). *Predictive analytics: What is it and what it matters*. Retrieved from www.sas.com: https://www.sas.com/en_us/insights/analytics/predictive-analytics.html

Subramanian, D. (2019, June 6). *A simple introduction to k-nearest neighbor algorithms*. Retrieved from SAS Communities Library: https://communities.sas.com/t5/SAS-Communities-Library/A-Simple-Introduction-to-K-Nearest-Neighbors-Algorithm/ta-p/565402

ACKNOWLEDGMENTS

Thank you to Tony Cooper, Sam Edgemon, and Brion Maher for their feedback.

CONTACT INFORMATION

Comments and questions are valued and encouraged. Contact the author at:

Patrick Maher
SAS Institute, Inc.
Patrick.maher@sas.com
www.sas.com