Machine Learning Model Recommendation

Rohit Sharma

Paripath Inc
Milpitas, CA

FCB Detection

IDENTIFYING VALIDITY OF TRANSACTIONS.

Objective of this work is to perform exploratory data analysis on the financial transaction provided in the form of excel spread sheet and identify suitability of machine learning algorithms and propose the one that is likely to produce an effective model in the long run.

> *"There are some frauds so well conducted that it would be stupidity not to be deceived by them."* -Charles Caleb

Problem Statement

Financial fraud and baseless chargebacks have been a problem for businesses. Increasing dependence on new technologies such as cloud and mobile computing in recent years has compounded the problem. Traditional methods of data analysis have long been used to detect fraud. They require complex and time-consuming investigations that deal with different domains of knowledge like financial, economics, business practices and law. Fraud instances can be similar in content and appearance but usually are not identical. (ref: G.K. Palshikar, The Hidden Truth – Frauds and Their Control: A Critical Application for Business Intelligence, Intelligent Enterprise, vol. 5, no. 9, 28 May 2002, pp. 46–51.). **Machine learning algorithms** are in perfect spot to take charge and spot fraudulent activity before it happens as trends in financial transactions continue to evolve along with method, data, technology and consumer habits.

This work is focused on spotting fraud indicator and chargeback indicators given a sample of over 88k financial transactions.

Metrices

Financial fraud and chargeback indicators are labels given to a transaction, so it is a class for machine learning model. Confusion metrics is a standard technique for machine learning classification problems. A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known.

		Predicted class	
		P	*N*
Actual Class	*P*	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

Data Analysis

Work dataset contains over 88k financial transactions with couple of transaction identifier (customer id and transaction id), 11 transaction features (column-1 to colum-11) that play role in identifying validity of transaction and indicators (Approved transactions, fraud indicator and chargeback indicator).

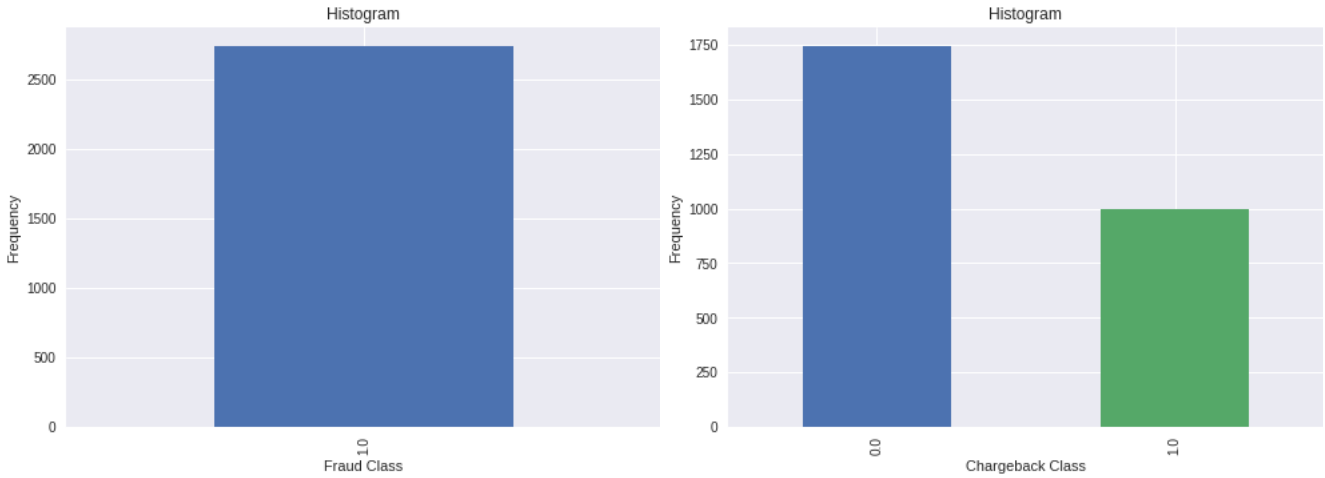
Data provided has 13 important features, 7 of which are partially filled anywhere from 17% to 79%. It is expected that this incomplete data will present a challenge in achieving the full potential of fraud detection algorithms. Moreover, format inconsistency of data was found unsuitable for direct consumption of Machine Learning algorithms. An indirect coding will be applied to continue with the data analysis.

Next two tables show basic statistics on these features and values.

	Column-1	**Column-2**	**Column-3**	**Column-4**	**Column-5**	**Column-6**	**Column-7**	**Column-8**
mean	8.537699	0.000666	2.074292	0.093514	11.02849	1.006885	0.265621	89.44171
std	1.315182	0.042358	0.407573	0.828265	11.71606	0.484698	0.936665	238.6015
min	-1	0	0	-1	-1	0	-1	-1
max	10	3	3	1	30	2	1	1115

	Column-9	Column-10	Column-11	Column-12	Column-13	Approved transactions	Fraud indicator	Chargeback indicator
mean	20.37915	0.913427	-0.60115	29.24471	880.2897	0.458136	1	0.363271
std	11.55154	1.004926	0.758187	37.52408	648.2519	0.498247	0	0.48103
min	0	-1	-1	-1	-1	0	1	0
max	33	2	1	130	1585	1	1	1

Data Visualization



Algorithm

Since we have over 88,000 transactions with different 13 features in the dataset with different distribution characteristics. This group is expected to give us a reasonable sense of a normal transaction free of fraud and chargeback. So we are tempted to apply outlier detection technique, traditionally popular in financial transaction fraud detections.

An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism [reference: Hawkins D 1980 Identification of Outliers Chapman and Hall]. Anomalies like transaction fraud and chargeback typically represent a different class (generating mechanism) of features, so there may be a large class of similar features that are the outliers.

we use a Density-Based Anomaly Detection algorithm. This algorithm assumes that normal data points occur around a dense neighborhood and abnormalities are far away. It finds μ (mean) and σ (standard deviation) of each feature in the set. These μ (mean) and σ (standard deviation) is used to compute probability for each sample. These probabilities are collected for every sample in the dataset to find local outlier factor (LOF).

We encode inconsistent feature set to provide numerical consistency to missing cells, which is represented as a category within the feature. Basis statistical properties, μ (mean) and σ (standard deviation), is computed for each feature. These properties are used to compute gaussian probability for each feature in the financial transaction. This probability is used to find a local outlier factor for each transaction, which determines if the transaction matches the properties of an outlier. Once, transaction is marked an outlier, it is evaluated for fraud and chargeback. [reference: Kriegel, H.-P., Kröger, P., Schubert, E., and Zimek, A. 2009a. LoOP: Local Outlier Probabilities. In Proc. ACM Conference on Information and Knowledge Management (CIKM), Hong Kong, China.]

Implementation

We used Colab as infrastructures for Python 3.0, numpy, pandas and other machine learning packages in order to implement our experiments and hypothesis.

Results

Our early results are encouraging given, we used unsupervised algorithm to detect these two labels.

Fraud prediction accuracy = 80%

Overall Fraud predictions = 55,889

Fraud prediction accuracy of 80% is an excellent start. However, large percentage of overall fraud prediction with a large False Positive metric is equally discouraging. Confusion matrices for the fraud is shown in the next picture. More work is needed to improve fraud prediction accuracy while reducing false positives.

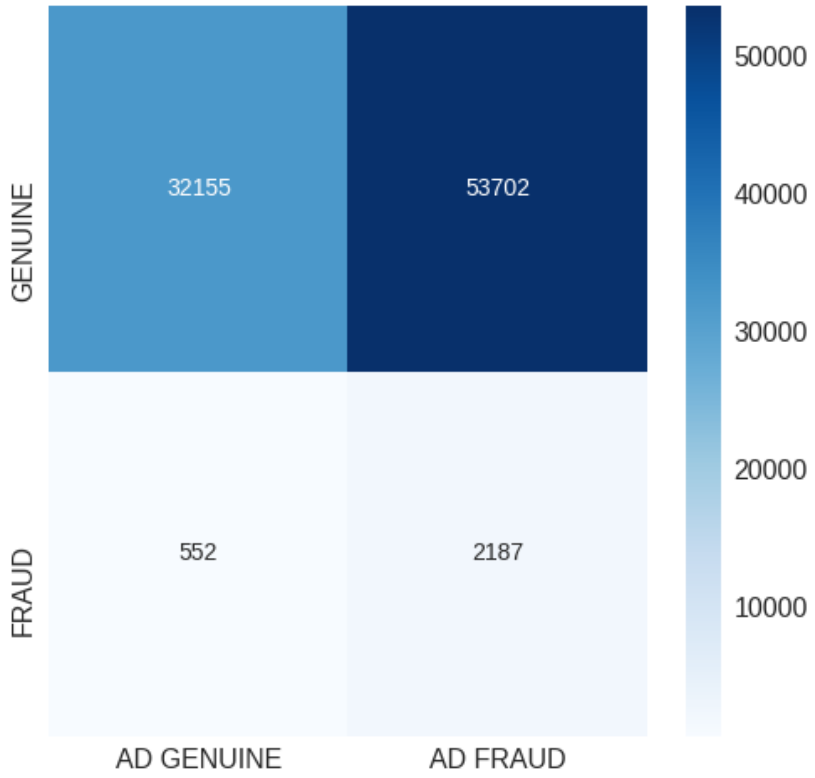


Figure 1: Fraud Prediction Confusion Matrix

Chargeback prediction results are like fraud prediction and indeed encouraging. This similarity gives us an indication that these two prediction labels follow similar underlying trends with minor differences.

True chargeback prediction accuracy = 78.39 %

False chargeback prediction accuracy = 21.61 %

Combined Chargeback prediction accuracy = 40.78 %

Chargeback prediction accuracy of 78% is an excellent start. However, large percentage of false chargebacks prediction with a large False Positive metric is equally discouraging. Confusion matrices for the chargeback detection is shown in the next picture. More work is needed to improve fraud prediction accuracy while reducing false positives.

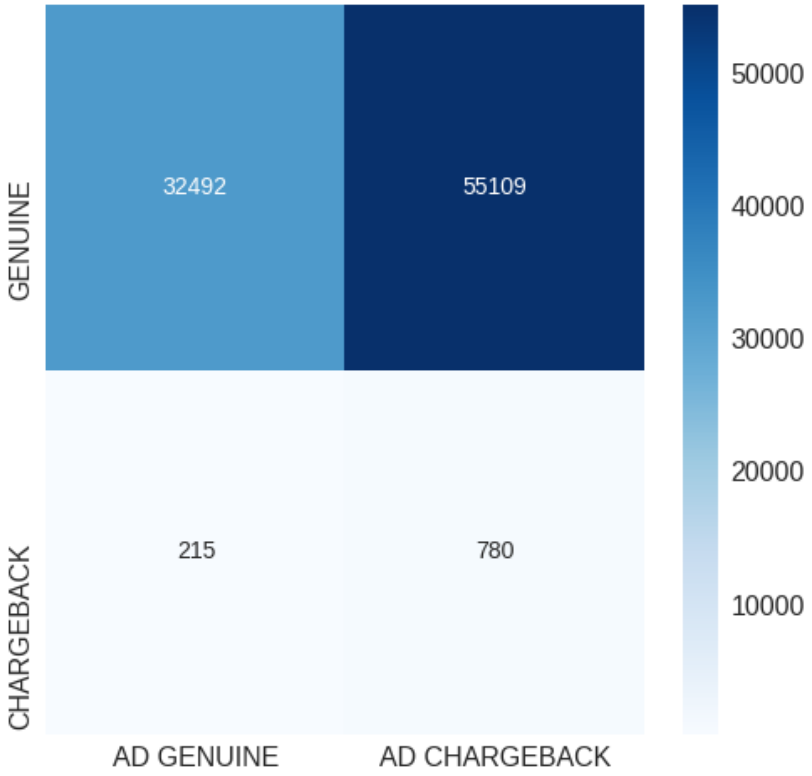


Figure 2: Chargeback Confusion Matrix

Recommendation

Density based algorithm for fraud and chargeback detection gave us a good start with around 80% accuracy. Large percentage of false positives makes this approach burdensome to put effort on. Next step of this work should focus on supervised anomaly detection and classifiers in search for further improvements in accuracy and reduction in false positives.

Single Class Support Vector Machine (SCSVM) [1] appears best solution for the dataset, since SVM does not require many labels in the dataset to find a large margin boundary by translating features in to high dimensional space. SCSVM learns a hyperplane in some feature space that divides the data points from the origin with maximum-margin. For translation-invariant kernel matrices, both approaches are equivalent. It has been shown that the classical SCSVM is a special case of a general class of density level set estimators that minimize a convex risk functional and give a general dual criterion for this class.

References

1. West J, Bhattacharya M and Islam R (2014) Intelligent Financial Fraud Detection Practices: An Investigation", in Proceedings of the 10th International Conference on Security and Privacy in

Communication Networks (SecureComm 2014), Vol. 153, 2015, LNICS, Springer, ISBN: 978-3-319- 23801-2 (Print) 978-3-319-23802-9 (Online), pp. 186-203.

2. N. G¨ornitz, M. Kloft, K. Rieck, and U. Brefeld. Toward supervised anomaly detection. Journal of Artificial Intelligence Research (JAIR), 46:235–262, 2013.
3. G. Blanchard, G. Lee, and C. Scott. Semi-supervised novelty detection. Journal of Machine Learning Research, 11:2973–3009, 2010
4. Campos, Guilherme O.; Zimek, Arthur; Sander, Jörg; Campello, Ricardo J. G. B.; Micenková, Barbora; Schubert, Erich; Assent, Ira; Houle, Michael E. (2016). "On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study". Data Mining and Knowledge Discovery. 30 (4): 891. doi:10.1007/s10618-015-0444-8. ISSN 1384-5810