Fact or Fraud: Identifying Factors to Predict Fraudulent Card Activity

Joshua Jay (joshua.l.jay-1@ou.edu)
Lily Rohrbach (rohrbachlf@ou.edu)
Shahzad Ansari (shahzad.s.ansari-1@ou.edu)
Zachary Burton (zachary.h.burton-1@ou.edu)
University of Oklahoma

EXECUTIVE SUMMARY

Fraud is an issue that has multiplied as many financial operations transition to the digital world, with recent focus on fraudulent transactions. In the best case, targets of credit card fraud lose the ability to withdraw their own funds from their account until a new card can be issued, and at worst, can find themselves out thousands of dollars, crippling credit card scores and financial independence. As a result, it is important to not only have measures on the back end to track attackers and restore funds, but also take preventative measures, stopping fraudulent transactions from initially occurring, aided by predictive models that can flag a transaction as potentially fraudulent, alerting the account owner.

A sample data set, downloaded from Kaggle, was used and is linked in the references portion of this report. Though thorough, the resulting size of this data presented a computational and time issue, resolved by using a small subset of the data when compiling. It is assumed the subset of data is representative of the data as a whole, which is supported by the use and comparison of multiple iterations to ensure results do not significantly differ each time.

The solution presented involves clustering, neural networks, and regression modeling. Clustering is used to separate the data into three distinct groups, based on their characteristics, helping to identify characteristics of individuals who are more likely to be victims of credit card fraud. Additionally, neural networking was done to map potential hidden correlations between data. Five different types of regression models were then created, including linear (lm), generalized linear (GLM) with a binomial fit, principal component regression (PCR), multivariate adaptive regression splines (MARS), and random forest (RF), and evaluated on goodness of fit. An iteration was run for each with every relevant variable to develop a baseline, before the selection of variables and interaction terms were fine tuned to achieve the best output

possible for each model, separately. All showed significant improvement from the baseline, with the MARS model achieving the best goodness of fit, with an average R^2 of 0.8975.

In summary, the resulting model provides analysis that identifies relevant variables and hidden correlations, groups data into clusters to better identify characteristics of fraudulent transactions, and trains regression models to predict fraudulent transactions. Before further steps are taken, a needs analysis is recommended to identify gaps between current fraud detection systems and this proposed model. Adaptations can then be made and run using a greater percentage of the data, using a more powerful computing system, with the goal of developing a comprehensive warning system to alert account owners of suspicious spending, minimizing the financial harm to both individuals and financial institutions.

PROBLEM BACKGROUND

Context

Credit card fraud occurs when an individual or group of individuals withdraw money from another's account without their knowledge or consent. There are multiple types of fraud: counterfeit card, lost/stolen card, and compromised account fraud [1]. For the purposes of this report, all types are simply treated as a binary; 0 if legitimate, 1 if fraudulent. Largely enabled by the advancement in technology and transition to digital banking, efforts to identify cases of fraud are becoming more important, with an emphasis on early identification to alert the targeted individual. This practice significantly reduces the financial loss to individuals and banks. At a cost of $28.65 billion, financial institutions are certainly invested in safeguarding their transactions[2]. Additionally, early and accurate fraud detection allows credit card companies such as Mastercard and Visa to more quickly cancel or freeze an individual's card and send them a new one. The resulting turnaround time is then shorter, reducing the inconvenience on targeted individuals. Financial institutions are heavily invested in this issue as it has become a primary factor in a customer's decision when selecting a company to bank with.

Data Description

The selected data set has been broken down into a training and test set, containing 1,296,675 entries and 555,719 entries, respectively. 21 variables are available to use as predictors, and the training

set includes whether or not each case was found to be fraud; of the 21 variables, 10 are numeric, and 11 are non-numeric, and are as follows: *...1* is an index serving as a primary key, *trans_date_trans_time* provides the data and time of the transaction, *cc_num*, is the credit card number, merchant is the provider of the product of service, *amt* is the transaction amount, *first* and *last* are the first and last name of the credit card owner, *gender* is assumed binary and is the gender of the credit card owner, *street*, *city*, *state,* and *zip* refer to the billing address of the credit card owner, *lat* and *long* are the latitude and longitude coordinates of the credit card owner's billing address, *city_pop* is the population of the city the credit card owner resides in, *job* gives the occupation of the credit card owner, *dob* is the credit card owner's date of birth, *trans_num* is a unique ID associated with the transaction, *unix_time* is a time stamp of the transaction, and *merch_lat* and *merch_long* are the latitude and longitude coordinates of the merchant. Variables will be referred to by their abbreviated terms for the remainder of this report.

Exploratory Data Analysis

The training and test data frames were reduced to a smaller percentage of themselves, allowing models to run in an appropriate amount of time. Because this model is to be used for demonstrative purposes, as opposed to realistic implementation, 0.01% of the data was kept, equating to approximately 1300 points for the training set and 560 for the test set. This is easily adjusted, allowing the model to be adapted to include a greater percentage of the data. Additionally, each time the code is run, it selects a different portion of the data, however a seed may be set if desired.

To provide background information and better understand the dataset, data was separated into numeric and non-numeric data frames for initial analysis. Figure 1 summarizes the numeric data.

variable	n	missing	missing_pct	unique	unique_pct	mean	min	Q1	median	Q3	max	sd
X	1296675	0	0	1296675	100.000	648337.000	0	324168.50	648337.0	NA	1296674.0	374317.974
cc_num	1296675	0	0	983	0.076	417192042079726784.000	60416207185	180042946491150.00	3521417320836166.0	NA	4992346398065154048.0	1308806447000240640.000
amt	1296675	0	0	52928	4.082	70.351	1	9.65	47.5	NA	28948.9	160.316
zip	1296675	0	0	970	0.075	48800.671	1257	26237.00	48174.0	NA	99783.0	26893.222
lat	1296675	0	0	968	0.075	38.538	20	34.62	39.4	NA	66.7	5.076
long	1296675	0	0	969	0.075	-90.226	-166	-96.80	-87.5	NA	-68.0	13.759
city_pop	1296675	0	0	879	0.068	88824.441	23	743.00	2456.0	NA	2906700.0	301956.361
unix_time	1296675	0	0	1274823	98.315	1349243636.726	1325376018	1338750742.50	1349249747.0	NA	1371816817.0	12841278.423
merch_lat	1296675	0	0	1247805	96.231	38.537	19	34.73	39.4	NA	67.5	5.110
merch_long	1296675	0	0	1275745	98.386	-90.226	-167	-96.90	-87.4	NA	-67.0	13.771
is_fraud	1296675	0	0	2	0.000	0.006	0	0.00	0.0	NA	1.0	0.076

Figure 1. Numerical Data Quality Report

Certain variables, such as credit card number, will not be used as predictors, but instead can be used to aggregate data points. Each of the variables are either highly unique or highly similar. For instance, it is seen that for variables *X, unix_time, merch_lat,* and *merch_long*, upwards of 95% of the observations are unique; these will need to be collapsed into categories. A correlation matrix was created to see which numeric variables correlate with one another as well as with the is_fraud variable within the training set, shown in Figure 2.

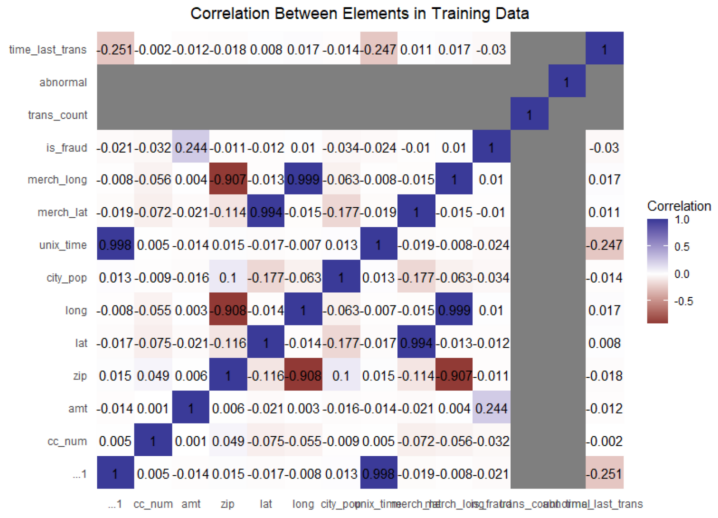

Figure 2. Correlation Matrix

The only variable that directly corresponds to *is_fraud* is the amount of money spent, but several other variables correlate with one another, demonstrating the need to adjust the model accordingly with the use of interaction terms.

Categorical data was also analyzed to better determine the initial parameters of the model. Similar to numerical data, the number of unique instances was found for each of the non-numeric variables, shown in Figure 3.

	Column Name	Unique Levels
1	trans_date_trans_time	1274791
2	merchant	693
3	category	14
4	first	352
5	last	481
6	gender	2
7	street	983
8	city	894
9	state	51
10	job	494
11	dob	968
12	trans_num	1296675

Figure 3. Number of Unique Instances for Categorical Data

There is also concern over the potential for skew, especially looking at how fraud in a given state may be more prevalent with a higher population. Figure 4 shows the number of fraudulent transactions per state, using the training set.

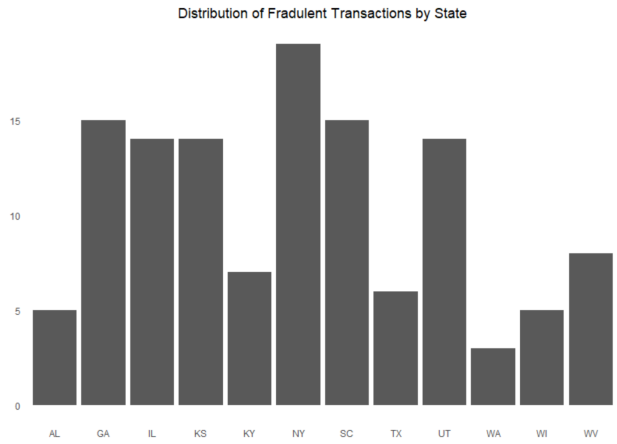

Figure 4. Count of Fraudulent Transactions Per State

It is clear from the graph some states like New York and Texas have the highest number of fraudulent cases where others like Hawaii and Rhode Island have the fewest number, but it cannot be assumed a transaction in one of these states is more likely to be fraudulent, just looking at this raw data. This issue can be solved through normalization, factoring in each state's population with the number of fraudulent transactions.

METHODOLOGY

Feature Selection and Data Cleaning

Factors were selected based on numerical correlation as well as through a comparison of goodness of fit scores, manually attempting combinations until a maximum score was reached. The interactions between variables with high correlation to one another were also considered. Factor collapsing was also done for the job and state variables, combining states into regions, and jobs with similar titles, such as “teacher-elementary” and “teacher-junior high” into single categories such as “teacher.”

There are no missing values in this data set, so missing value imputation was not required. To eliminate outliers, box plots were first made for each variable of interest to determine which had outliers. Quantiles were then created for variables containing outliers and points that fell outside of the interquartile range were eliminated.

Some elements were transformed in an attempt to better represent the data during predictive modeling. The first was encoding transactions as abnormal based on the time of day which the transaction occurred. All of the transaction times are given in the time zone of the purchaser, so any transaction between the time 22:00:00 and 4:00:00 are flagged as abnormal. The next step in feature engineering was evaluating the frequency of transactions in the last x days. This can help track patterns in purchasing, and can be supplementary to the time transactions take place. For example, an individual consistently makes purchases during the late evening but suddenly has a charge on his account mid-morning. This could be an instance where a notification could be sent to confirm or deny the purchase. Lastly, the time between transactions was measured. This may be helpful in specific situations where transactions occur abnormally close together, such as within seconds. These transformations ultimately provide greater insight on the data and can be used to more effectively predict fraudulent behavior.

Modeling Choices

Clustering was performed to identify groupings within the data. This can help find similar groups of subjects and potentially identify groups who may be more at risk for being victims of credit card theft

or types of transactions that are likely to be fraudulent. Multiple methods were used including k-means, k-medoids, and hierarchical clustering. A neural network was also created to show hidden patterns within the data. Lastly, several models were developed and compared, using various measures of goodness of fit, including GLM with binomial fit, lm, PCR, MARS, and RF.

Model Validation Plan

By selecting a percentage of the dataset to use each time the code is run, due to machine capability constraints, the effectiveness of the predictive models may change, evaluated by shifts in the goodness of fit metric. These will be compared across multiple iterations, with the average and standard deviation used to measure how the model differs with new subsets of data. 5-fold Cross validation is also being used for necessary models, such as PCR and RF.

RESULTS

Clustering Performance Summary

Clustering shows the separation of groupings within the data, found through different methods; in this case, k-means, k-medoids, and hierarchical clustering were all used. To find the optimal k value, a WSS plot and Hartigan's plot were both made, shown in Figure 4 and Figure 5.

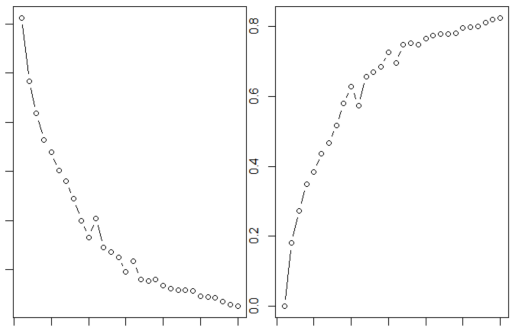

Figure 5. WSS Plot

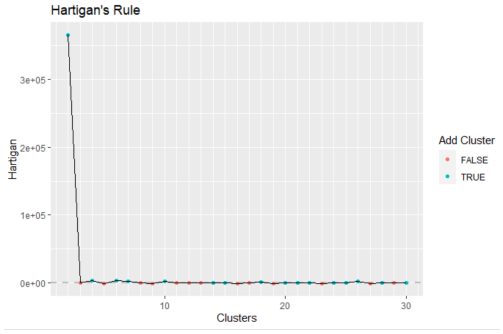

Figure 6. Hartigan's Rule

While the WSS plot shows a bend, it is not distinct enough to select a value. Looking at Figure 6, a steep drop can be seen, and at k=3, the legend shows a cluster should be added. It was decided to move forward using this.

The locations of the cluster centroids for k-means and k-medoids are shown in Table 1.

Table 1. Location of Centroids for K-Means and K-Medoids

	Cluster 1		Cluster 2		Cluster 3	
	k-means	k-medoid	k-means	k-medoid	k-means	k-medoid
cc_num	-0.1016	-0.3044	0.2725	-0.3072	-0.0096	-0.3072
amt	0.0271	-0.2224	0.0676	0.0693	-0.0852	-0.1397
zip	-0.6314	-0.2108	1.2680	0.7082	0.2061	-0.8370
lat	0.5507	0.2559	0.4064	-0.1141	-1.1269	0.1227
long	0.6141	0.5364	-1.5630	-0.7912	0.0060	0.7781
city_pop	-0.0504	-0.2974	-0.0070	-0.2896	0.0842	-0.2961
unix_time	0.0169	0.6915	-0.0120	-0.4209	-0.0193	-0.9775
merch_lat	0.5538	0.2939	0.4047	-0.1894	-1.1307	0.1249
merch_long	0.6143	0.5624	-1.5624	-0.7838	0.0053	0.8399
is_fraud	0.0010	-0.0788	0.1231	-0.0788	-0.0788	-0.0788

These clusters were created using scaled values and numbered based on the output. The k-medoid cluster is visualized below in Figure 7.

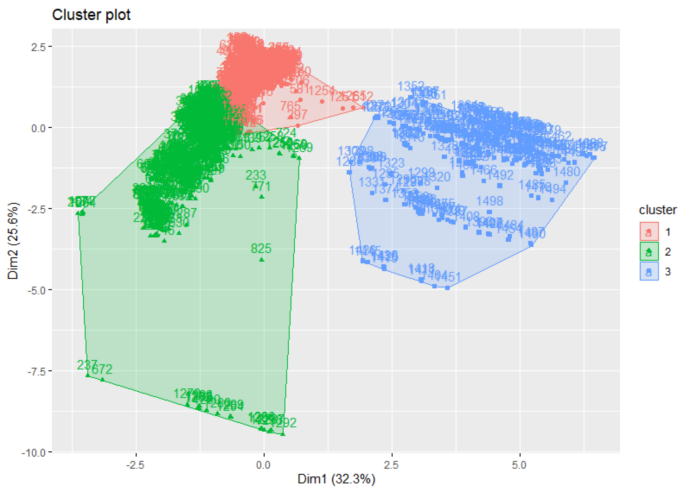

Figure 7. K-Medoid Plot

The centroid locations between the k-means and k-medoids methods differ for nearly all variables, the exception being the amount variable. There appears to be some overlap between Cluster 1 and Cluster 2, but Cluster 3 has no overlap with either of the other two. Given this, it's possible that providing two clusters, instead of three may yield results that better separate the groupings.

Lastly, hierarchical clustering was performed with four different linking methods: single cluster, complete cluster, average cluster, and ward cluster, but the first and last of these were decided to be the most significant and are shown below in Figures 8 and 9.

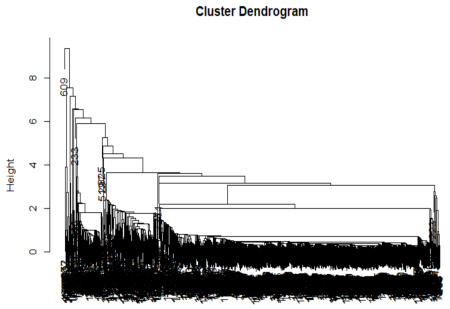

Figure 8. Single Cluster

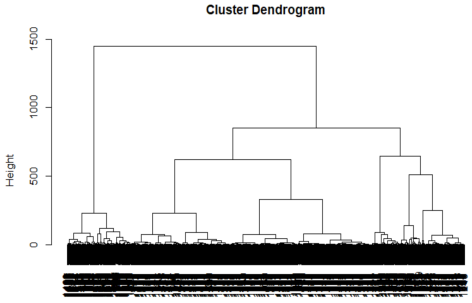

Figure 9. Ward Cluster

The single cluster linkage is the most unbalanced, shown with the long chains. Looking at the ward cluster, heights of approximately 600-800 have three lines (counted across the x axis), representing the three clusters. Above that, from heights of approximately 800-1500, there are two lines counted. There is a greater range in height for two clusters than for three, meaning the two clusters would be more distinct from one another, but with adding another, there still shows an appropriate level of separation, supporting the decision to separate into three clusters.

Neural Network Performance Summary

Neural networks are a form of machine learning and can better identify relationships between variables, especially those that may be hidden from other analysis methods. Several models were attempted, with the final model being a 3x3 deep neural network with a learning rate of .0001 and a threshold of .04. The input variables used were merch_lat, merch_long, lat, long, and amt. A visualization is shown in Figure 10.

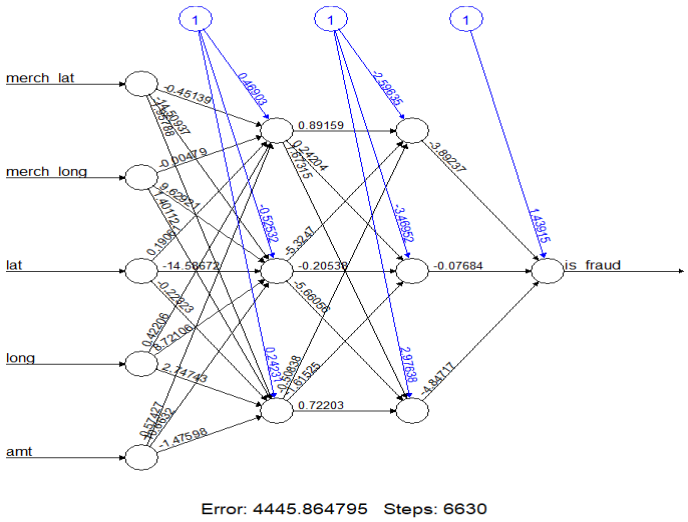

Figure 10. Neural Network

The final model had an accuracy of 92.8 percent with a specificity of 89.41 and a sensitivity of 93.46. This model was constructed using a training set of 25000 rows and tested on 2500 rows with a quarter of the rows indicated as fraudulent.

Regression Model Performance Summary

To establish a baseline, each model was run with every variable of interest, including *cc_num, merchant, category, amt, gender, state, zip, lat, long, city_pop, dob, unix_time, job, merch_lat,* and *merch_long*. Table 2 summarizes the initial results.

Table 2. Initial Summary of Models

	Goodness of Fit
GLM	1290 (AIC)
lm	-0.005 (adjusted R^2)
PCR	51.57% (variance explained) at 661 comps
MARS	0.480 (R^2)
RF	0.0059 (spearman R^2)

This provides an initial benchmark to evaluate the success or lack thereof of adapted models containing more specified variables and interaction terms. Models were modified with final results shown in Table 3.

Table 3. Final Summary of Models

	Goodness of Fit 1	Goodness of Fit 2	Goodness of Fit 3	Average	SD
GLM	801.86 (AIC)	716.87 (AIC)	665.33 (AIC)	728.02	56.293
lm	0.7167 (R^2)	0.7417 (R^2)	0.3681 (R^2)	0.6088	0.1705
PCR	50% at 11 comps 100% (variance explained) at 24 comps	50% at 11 comps 100% (variance explained) at 24 comps	50% at 11 comps 100% (variance explained) at 24 comps	*NA*	*NA*
MARS	0.8964 (R^2)	0.9039 (R^2)	0.8922 (R^2)	0.8975	0.0048
RF	0.4096 (spearman R^2)	0.4230 (spearman R^2)	0.3677 (spearman R^2)	0.4001	0.02355

All models improved. Most notably, the lm, PCR, and MARS models resulted in the best fits, with RF also demonstrating significant improvement, though still not resulting in a correlation value high enough to be significant. Some of the models varied by which variables were used in the prediction, in order to find the variables and interactions that yielded the best results for each particular model. The lm, MARS, and RFt models used the following variables: *lat, dob, amt*job, amt*category, amt*state, long*merch_long, gender*job, gender*category*. The GLM model used the following variables: *amt, lat, dob, amt*state*. Lastly, the PCR model used: *category, amt, lat, long, city_pop, job, merch_lat, merch_long*. It appears the MARS model may best fit this particular dataset.

CONCLUSION

Problem Summary, Approach, & Findings

As the banking world continues to transition to a digital interface, protecting individuals' financial accounts is crucial. Fraudulent attacks are only likely to increase as attackers find new ways to access money electronically. While fraud prevention efforts continue to adapt and evolve, response to fraudulent attempts is proven to be just as important, alerting individuals of potential fraud early on with the goal of prohibiting a transaction from going through and locking account services until the account owner can properly secure their finances.

After base analysis and data cleaning, clustering was performed, using Hartigan's rule to evaluate how many clusters the data should be grouped into, resulting in a total of three. K-means, k-medoids, and hierarchical clustering were then performed and compared. The centroids of k-means and k-medoids show little similarity, with the exception being the variable *amt*, meaning this may be the best form of separation between the three clusters. Hierarchical clustering then shows a sufficient separation of clusters, supporting the decision to generate three, but also supports the possibility of only using two. The neural network was also able to achieve a high degree of accuracy, using five input variables to identify correlations.

Variables were then selected and included in regression model variants, evaluated based on their goodness of fit. Different combinations were run, with the goal of maximizing the first of these metrics, and minimizing the latter. Visuals such as the correlation matrix were used to determine which variable combinations should be used as an interaction effect within the regression model.

Initial models, including nearly all of the variables, reflect low correlation and high levels of error, but provide a benchmark to surpass by integrating more specific variables and including interaction effects. Each model was separately run with varying combinations of variables in an attempt to find the best output possible. The lm, MARS, and RF models shared variables and interaction terms, while GLM and PCR each had their own variables and interaction terms. Each showed an improvement over the benchmark, with MARS resulting in the best output, with an R^2 average of 0.8975, compared to the benchmark of 0.480.

Issues & Limitations

Due to the size of the dataset, reducing the number of data points within both the training and test set was required. This was done by creating new data frames only containing a percentage of the original data. Because of this, it is possible the data points selected may be a skewed portion of the original data and affect variables such as time and date data, removing necessary points that would indicate a transaction is abnormal.

Similarly, the implementation of this code in a realistic situation would require significantly more computing power to accommodate the millions of purchases being made daily. Consumers would also need to opt into a fraud detection program and be willing to provide some of the information used for prediction as well.

Factor collapsing was done manually because it was important specifical factor levels fell into specific categories, as opposed to using a lumping technique that groups based on the n^{th} largest factor size. Because of this, the introduction of new factor levels outside of the ones in the given training sets will result in an error. If a variation of this code were to be implemented in a realistic situation, users would need to be given a selection of jobs to choose from to guarantee a finite number of factor levels with no accidental overlap due to slight description variation or typos.

The neural network was subject to significant computational and time constraints, resulting in a higher threshold value than what is optimal. Being able to dedicate a higher level of power and time would allow more input variables to be used, thus resulting in a more accurate model.

Recommendations

A needs analysis should be conducted to evaluate current fraud detection systems and better understand how to bridge the gap between the current and desired future states. Expanding the model to encompass a larger portion of data, using a more powerful computing system, would help prevent potential skew through data selection and result in better trained models. In addition, running more iterations of each model would allow the evaluation metrics (average and standard deviation) to better represent the goodness of fit and potential variation.

REFERENCES

[1] Lee, N. (2021, February 1). *Credit card fraud will increase due to the Covid pandemic, experts warn*. CNBC. https://www.cnbc.com/2021/01/27/credit-card-fraud-is-on-the-rise-due-to-covid-pandemic.html

[2] RB, A., & KR, S. K. (2021). Credit card fraud detection using artificial neural network. *Global Transitions Proceedings*, *2*(1), 35–41. https://doi.org/10.1016/j.gltp.2021.01.006

[3] Kaggle (2021). Credit card transactions fraud detection dataset. https://www.kaggle.com/kartik2112/fraud-detection