Data Mining for Fraud Detection in Large Scale Financial Transactions

Knox Kamusweke, Mayumbo Nyirenda and Monde Kabemba

October 21, 2019

Data Mining for Fraud Detection in Large Scale Financial Transactions

Knox Kamusweke[1]
Dept. of Computer Science
University of Zambia
P.O. Box 32379, Lusaka. Zambia
Email: knox.kamusweke@cs.unza.zm

Mayumbo Nyirenda[2]
Dept. of Computer Science
University of Zambia
P.O. Box 32379, Lusaka. Zambia
Email: mayumbo.nyirenda@cs.unza.zm

Monde Kabemba[3]
Dept. of Computer Science
University of Zambia
P.O. Box 32379, Lusaka. Zambia
Email: monde.kalumbilo@cs.unza.zm

Abstract— Financial Institutions are involved with generating and handling million records of transactions across their platforms. These transactions contain significant patterns and trends which are hidden but needed for knowledge discovery and actionable insight into either fraudulent or non-fraudulent events. Uncovering these patterns and trends has always been a challenge to most financial institutions as they are ever-changing at an unknown frequency and handling the large scale financial transactions is not easy. This study uses data mining to uncover significant patterns and trends in large scale financial transactions and construct a model to predict and forecast fraud on the basis of uncovered patterns and trends.

Keywords: Data Mining; Fraud; Patterns and Trends

I. INTRODUCTION

Financial service providers (FSPs) are involved with generating and handling million records of transactions daily which are stored in large databases. These transactions have continued to grow exponentially in the financial institutions across multiple platforms. One major challenge faced by FSPs is lack of knowledge discovery and actionable insight about the nature or patterns of transactions taking place and their trends. Uncovering these patterns and trends has been a challenge due to the unknown frequency of transactions and ever changing patterns, not forgetting the million records of transactional data which demands scalable intelligent learning systems to analyse and handle. This research paper highlights the use of data mining for financial fraudulent pattern and trend discovery particularly in the financial and banking sector and then constructs a model for predicting and forecasting fraud on the basis of uncovered patterns. In addition, it presents cases in which data mining techniques were successfully implemented to detect fraud. It further discusses the methodology used in our research before discussing our findings. Before going into the details, a brief discussion of fraud and data mining is introduced to give an insight.

II. FRAUD

There are many definitions for fraud that vary depending on the perspective of interested parties [1] [2] [3]. In its simplest definition, fraud been defined as any intentional act or omission designed to deceive others, resulting in the victim suffering a loss and/or the perpetrator achieving a gain [3] or criminal activity of misrepresenting information to achieve unjust gains [4]. Fraud commonly includes activities such as theft, corruption, conspiracy, embezzlement, money laundering, bribery and extortion [2]. The Association of Certified Fraud Examiners defines fraud as “the use of one’s occupation for personal enrichment through the deliberate misuse or misapplication of the employing organization’s resources or assets” [5]. Frauds are perpetrated by parties and organizations to obtain money, property, or services; to avoid payment or loss of services; or to secure personal or business advantage”.

The process of detecting and preventing fraud can be very challenging due to the fraud patterns that keep changing. Calculating the actual costs of fraud for businesses and corporations is a formidable task because of the stealthy nature of fraud. The detected fraud cases represent only the "tip of the ice berg" of all the frauds that go unnoticed [6]. In addition, fraud can be very complex and has some temporal characteristics and patterns that have to be identified.

The Reserve Bank of India – RBI maintains data on frauds on the basis of area of operation under which the frauds have been perpetrated [7]. According to such data pertaining, top 10 categories under which frauds have been reported by banks are as follows;

1) Credit Cards
2) Deposits – Savings A/C
3) Internet Banking
4) Housing Loans
5) Term Loans
6) Cheque /Demand Drafts
7) Cash Transactions
8) Cash Credit A/C (Type of Overdraft A/C)
9) Advances
10) ATM/Debit Cards

Studies and research have proven that traditional fraud detection techniques like random checks, targeted audits, internal control systems, external audits, risk management systems and whistleblowing hotlines, might not be effectively applicable for the new trends in fraud [8]. To analyse data and determine various kinds of fraud-like patterns on a large scale transaction data, data mining techniques have merged to make it less vulnerable and provide reliable solutions to business [9] [8]. Although detecting fraud is considered a high priority for many financial institutions, the current literature lacks for an up-to-date, comprehensive and in-depth review that can help firms with their decisions of selecting the appropriate data mining technique [10] .The next section provides a brief discussion of data mining and highlights the most commonly recognized fraud detection techniques in data mining.

III. DATA MINING

A. Data Mining the core of Knowledge Discovery Process.

[11] Larose et al, defined data mining as the process of discovering useful patterns and trends in large data. The process must be automatic and the patterns discovered must be meaningful in that they lead to some advantages, usually an

economic advantage. Also known as Knowledge Discovery Data or Knowledge Mining, Data Mining (DM) involves the analysis of data from different perspectives and summarizing it into useful information [12] [13]. Data mining is the core of the knowledge discovery in databases (KDD) process, thus, data mining and KDD are often used interchangeably [14, 15]. Fig 1 shows an illustration of the process of extracting knowledge from data using data mining. The first three processes, that are data selection, data preprocessing and data transformation, are considered as data preparation processes. The last three processes including data mining, pattern evaluation and knowledge representation are integrated into one process called data mining [16].

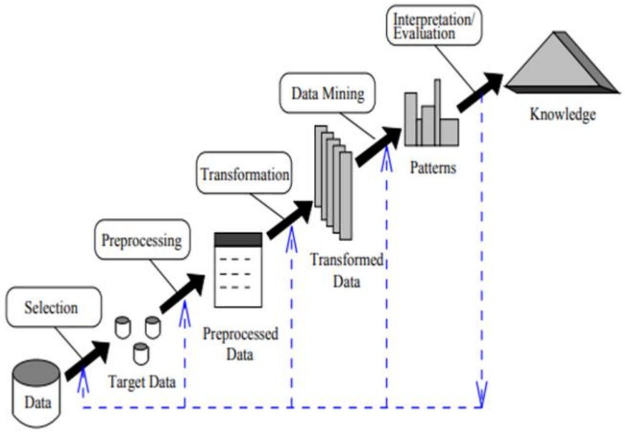


Fig 1. The process of data mining transitioning from raw data to valuable knowledge.

In general, data mining techniques can be classified into four categories according to the type of the machine learning techniques as;

1) Supervised Learning for Fraud Detection: This method uses supervised learning in which all the available records are classified as “fraudulent” and “non-fraudulent”. Then machines are trained to identify records according to this classification. However, these methods are only capable of identifying frauds that has already occurred and about which the system has been trained [8] [17].

2) Unsupervised Learning for Fraud Detection: This method only identifies the likelihood of some records to be more fraudulent than others without statistical analysis assurance [8] [17].

3) Semi-supervised Learning for Fraud Detection: This method addresses fraud in a similar way as supervised learning. However, in semi-supervised learning, the machine is provided with labeled data along with additional data that is not labeled with predetermined outcome [18].

4) Reinforcement Learning for Fraud Detection: This method differs from the ones above because models do not get trained with sample data but through trial and error to detect fraud [19].

B. Data Mining Models.

A data mining model is created by applying an algorithm to data. A data mining model gets data from a mining structure and then analyzes that data by using a data mining algorithm. The data mining models are of two types [16] [20]: Predictive and Descriptive. The predictive model makes prediction about unknown data values by using the known values. Examples include Classification, Regression, Time Series Analysis, Prediction etc. The descriptive model identifies the patterns or relationships in data and explores the properties of the data examined. Examples are Clustering, Summarization, Association rule and Sequence discovery etc.

C. Summary of the Most Significant Fraud Detection Techniques/Models.

Table 1 gives a summary of some of the most widely known data mining techniques for fraud detection in banking/financial sector. Choosing the right method for a specific implementation depends on trial and error [10] [21].

DM Technique Employed	Description	Fraud Type
Neural Network	It can be taught to implement sophisticated machine learning algorithms and pattern recognition to make predictions from historical data.	Financial statement fraud, forecasting cash withdrawals in the ATM, fraud in credit card transactions
Fuzzy Neural Network	Use traditional association rules	Financial Reporting
Rule-Learning	The extraction of useful if-then rules from data based on statistical significance	Credit Card Fraud
Decision Tree Analysis	A predictive model that is based on a sequence of decision	Credit Card Fraud
Fuzzy C-Means Clustering	Relates known fraudsters to other individuals	Financial and automobile insurance fraud detection
Peer Group Analysis	Detects individual objects that begin to behave in a different way	Credit Card Fraud
Break-Point Analysis	Observation where irregular behavior for a particular account is detected	Credit Card Fraud
Predictive Analysis	Integrates a variety of techniques from mathematics,	Financial Statement Fraud

	statistics with data mining to analyze current and historical facts to make predictions about future events	
K-Means Clustering	Forms clusters or patterns of transactions to describe the data	Refund fraud, financial statement fraud and Credit card fraud
Logistic Regression	Uses predictive analysis to model the probability of a certain class or event existing such as pass/fail	Financial fraud statements.

Table 1. Summarized work for detecting financial fraud with DM techniques
Sources: [10] [21] [22] [23] [24] [25] [26] [27] [28]

IV. RELATED WORKS

Despite being a relatively new technology that has not fully matured, there are a number of industries such as banks, insurance companies and retail store that are already using data mining on a regular basis [29] [30]. Data mining techniques like k-means clustering, logistic regression, decision tree, support vector machine (SVM), naïve Bayes , random forest etc. have been used in detecting financial fraud [10] [31]. One system that has been successful in detecting fraud is Falcon's "fraud assessment system" being used by nine of the top ten credit card issuing banks, where it examines the transactions of 80% of cards held in US [7]. Mellon Bank is also using data mining for fraud detection and is able to protect itself and its customers' funds from potential credit card fraud.

In their study, Vaishali [28], were able to detect fraud in credit card by using clustering approach. Their results were found by using K-Means clustering algorithm. Their algorithm formed four clusters being low cluster, high cluster, risky cluster and high risky cluster. They then tested transactions on five credit card numbers by applying K-Means clustering. Their results showed that some fraudulent transactions done in India, Ukraine and Ecuador with credit card numbers 456723, 234562, 176345 belonging to low cluster, high cluster and high risky cluster.

In the United Kingdom (UK), Provident Financial's Home credit Division had no system to detect and prevent fraud. After applying data mining techniques, they have reduced frequency and magnitude of agent and customer fraud, saved money through early fraud detection, and saved investigator's time and increased prosecution rate [29].

Peer Group Analysis, an unsupervised DM technique for monitoring customer behaviors over period of time is used in Australia [32] to monitor customer behaviors. For each individual that has a credit card account, a "Peer Group" of accounts are created that exhibit similar behavior. Over time, the behavior of an account is tracked by those accounts in its peer group. If an account has subsequent behavior which deviate strongly from its peer group, it is thus considered to have behaved anomalously and flagged as a potential fraudulent. Also in Australia, a "Break-Point" Analysis DM technique is being used to distinguish spending activities supported from transaction information in a single account. Current transactions are matched up with prior spending activities to spot features, such as rapid spending and an increase in the level of spending, which would not essentially be captured without data mining [33].

In their study, S. Thiprungsri [34] used K-Means as a clustering procedure on the 40,080 claims that were paid in the first quarter of 2009. For the first set of clusters using 2 attributes, eight 8 clusters were formed. About 90% of claims were clustered into cluster 7 and 6% in cluster 0. Three clusters (1, 2 and 5) had membership of less than 1% and the numbers of claims in those clusters were 54, 84 and 31 respectively. Examining the characteristics of those less populated clusters, a couple of suspicious characteristic were mentioned. Claims in those clusters had high interest/beneficiary payment percentage and/or claims with a long period time from death dates to payment dates. The results showed that the total number of claims identified as possible anomalies from clusters was 169. Clusters with larger membership have higher numbers of possible anomalies.

[35] Kirkos et al. tried to identify firms that published fraudulent financial statements using Decision Trees, Artificial Neural Networks and Bayesian Networks. In the study, 76 financial tables with half of them fraudulent were used. As a result of training and tests with 10 selected features, the best classifier was Bayes Networks with 90.3% success.

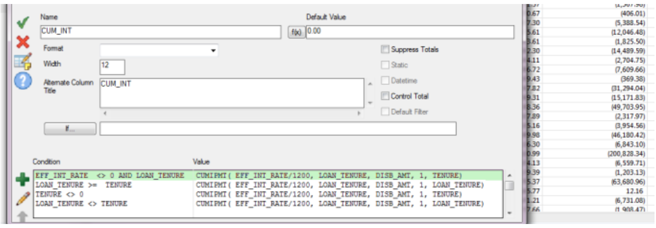
[36] A large regional bank was looking to improve revenues, and improve the customer experience. As part of their growth efforts, they put increasing emphasis on reducing the occurrence of fraudulent debit card transactions. The bank was losing about $100,000 per month and risked negatively impacting customers through rejected transactions and reissued debit cards. After employing data mining through predictive analytics, the bank quickly reclaimed an average of $2 per account in fraudulent transactions. Given the bank's large customer base, this quickly added up to substantial savings.

V. METHODOLGY

The aim of this research is to develop a model for fraud detection in large scale financial transaction using Data Mining (DM). To achieve this, the first thing we did was to conduct a study and review existing literature in order to investigate the challenges which financial institutions face during financial fraud detection. Based on this study we then propose a data mining model for fraud detection. To better understand the usage scenarios, we will discuss our proposed approach based on the currently used approach. We decided to withhold the name of the bank for privacy and also due to the sensitivity of data involved. We therefore name the bank as Bank X.

A. Case of Bank X

Like in most financial institutions, fraud detection through audits checks are targeted and sometimes following whistle blowing or engagement with external inspectors from government institutions such as Central Bank, Ministry of Finance and Auditor General's office. The current approach usually employs techniques such as sampling and random checks with very much predefined patterns or characters to look out for in the transactions. The auditors and inspectors usually start by planning for the targeted audits checks. They set the audit objectives, which in many cases is involve defining the patterns or characteristics to look out for in the data. The expected resulting patterns are therefore known for example, to reconcile or check if loan recovery amounts, interest bookings and settlements were correctly passed on customer savings and loans accounts. The officers then identify the available data from the system through requests with information staff considering the period under examination. After receiving the data, the officers carry out manual integrity tests on the data using ACL analytics for which bank maintains a license [37] and sometimes manually using excel. The test data validity are done to check against any errors, missing values, ensure only numeric data is in numeric fields, if data is in fields where it is expected, confirming control totals and ensure calculated fields deliver correct values using statistic commands. After completing their validity tests, the auditors then perform data analysis. They perform the data analysis through tests necessary to achieve their objectives. During data analysis, the auditors create tables after which they create expressions, filters, and computed fields (unconditional computed filters, conversion computed fields, conditional computed fields) to apply in their analysis (Fig 2). They use expressions to normalize the data. After concluding their analysis, they report the significant findings developed in response to each audit objective. The current approach is biased, error prone due to heavy reliance on labour from control officers, takes longer periods to conclude therefore, the results are not attainable within the shortest possible time. The current approach is purely static-rule based because it looks for specific, predetermined patterns or set of characters in the data through targeted audits. It requires manual intervention at every stage of the process on manipulating variables and experts need to specify the set of patterns or characters to look for before starting.



Name: CUM_INT

Default Value: 0.00

Format:

Width: 12

Alternate Column Title: CUM_INT

□ Suppress Totals
□ Static
□ Datetime
□ Control Total
□ Default Filter

Condition	Value
EFF_INT_RATE <> 0 AND LOAN_TENURE	CUMIPMT(EFF_INT_RATE/1200, LOAN_TENURE, DISB_AMT, 1, TENURE)
LOAN_TENURE >= TENURE	CUMIPMT(EFF_INT_RATE/1200, LOAN_TENURE, DISB_AMT, 1, LOAN_TENURE)
TENURE <> 0	CUMIPMT(EFF_INT_RATE/1200, LOAN_TENURE, DISB_AMT, 1, TENURE)
LOAN_TENURE <> TENURE	CUMIPMT(EFF_INT_RATE/1200, LOAN_TENURE, DISB_AMT, 1, LOAN_TENURE)

0.67	(406.01)
7.30	(5,388.54)
5.61	(12,046.48)
8.61	(1,825.50)
2.30	(14,489.59)
4.11	(2,704.75)
6.72	(7,609.66)
9.43	(369.38)
7.82	(31,294.04)
9.31	(15,171.83)
8.36	(49,703.95)
7.89	(2,317.97)
5.16	(3,954.56)
9.98	(46,180.42)
6.30	(6,843.10)
0.99	(200,828.34)
4.13	(6,559.71)
9.39	(1,203.13)
5.37	(63,680.96)
5.77	12.16
1.21	(6,731.08)
7.66	(1,908.47)

Fig 2. Current approach to find patterns on loans showing conditions being set on what specific patterns to look for in the expressions. Once set, these conditions remain static and only focus on getting specific results.

B. Proposed Approach: A DATA MINING MODEL

We observed that one of the challenges with the current approach is that it remains a traditional packaged fraud detection approach because it is custom tuned for the particular data that is available for the bank. The current approach is static-rule based which looks for specific predefined patterns or set of characters in the data and leads to longer periods of time to yield the results especially on large scale financial transactions. This leaves a gap as fraud patterns are ever changing without known frequency in the growing data. It is for this reason that we base our model on the currently used approach. However, we incorporate data mining to replace the static rules set in the data analysis of the current approach. Therefore, preceding our steps in our approach with data mining was understanding of the application domain based on prior knowledge obtained from current approach.

1) *Data Selection*

The first step in our approach was to identify and select the target data. We selected and created the target data in CSV format. This involved selecting and creating target dataset and subset of data samples or variables useful for our data mining (Fig 3). This was very important because not all data collected may be needed. This was historical data created and consisted of transactions on customer savings and loan accounts.

2) *Data Cleaning and Pre-processing*

The data we collected was not clean and contained errors, missing values etc. We placed the target data onto the staging area to perform data cleansing and preprocessing. This was necessary in order for us to apply techniques to get rid of the anomalies found in the data and ensure to have complete and consistent data which is suitable for our data mining model. One challenge we faced at this stage was how to handle missing values. Missing data are a problem that continues to plague data analysis methods and this is common in data sources with large number of variables. The absence of information is rarely beneficial but all things being equal, more information is always better for data mining. Therefore, we had to be careful about handling the thorny issue of missing values. To achieve this, we used a criteria of replacing the missing values with a value generated at random from our observed distribution of the variables.

3) *Data Transformation*

The variables in our cleansing and preprocessed data had ranges varying from each other. Therefore, we needed to normalize our variables in order to standardize the scale of effect each variable was to have on our results and achieve normality. In general, learning algorithms such as k-means clustering benefit from normalization and standardization of the data set. We normalized our data with MinMax scaling function provided by “sklearn.preprocessing” package and scikit-learn machine learning library for python. At this stage, we also coded the categorical feature into numerical features so as to implement our data mining model easily. Some of the notable categorical variables we coded in our data are: ***ACCT_STATUS*** where Active=1, Inactive=2, Dormant=3. ***TRAN_DATE*** took format of DDMMYYYY and HH24MISS for ***TIME***. The ***DEBIT/CREDIT INDICATOR*** had DR=0 and CR=1. ***PAYOFF_FLG*** for loans had Y=1 and N=0 etc. We coded the data to also maintain privacy due to its sensitivity and nature.

After transforming our data, we then had to perform dimensionality reduction because our data was highly dimensional which was going to have performance impact on our model. For this, we used unsupervised dimensionality reduction technique, Principal Component Analysis (PCA).

4) *Data Mining*

This is the core step in our approach. At this point, we were ready to apply our data mining algorithms. This involved first of all choosing a suitable data mining task. In our approach, we employed three data mining tasks namely clustering, regression and time-series analysis. These were suitable because our aim was to find unknown hidden patterns in our data involving transactions by grouping similar features and then be able to predict continuous-value attributes in the data. To achieve clustering task, we chose and employed k-means clustering algorithm (Fig 4). K-means clustering was used to give different patterns due to its simplicity and ease of implementation as well as versatility. For the regression task we used logistic regression algorithm so as to give predictive analysis and model the probability of certain classes of transactions found. For time-series analysis, we incorporated the exponential smoothing function to find regular patterns of unusual transactions in measures that are likely to be continued into the future.

5) *Interpretation/Evaluation*

At this stage of our approach we focused on interpreting and evaluating the mined patterns. This involved us performing visualization our results from the patterns of transactions in (Fig 5, 6, 7, 8, 9 and 10).

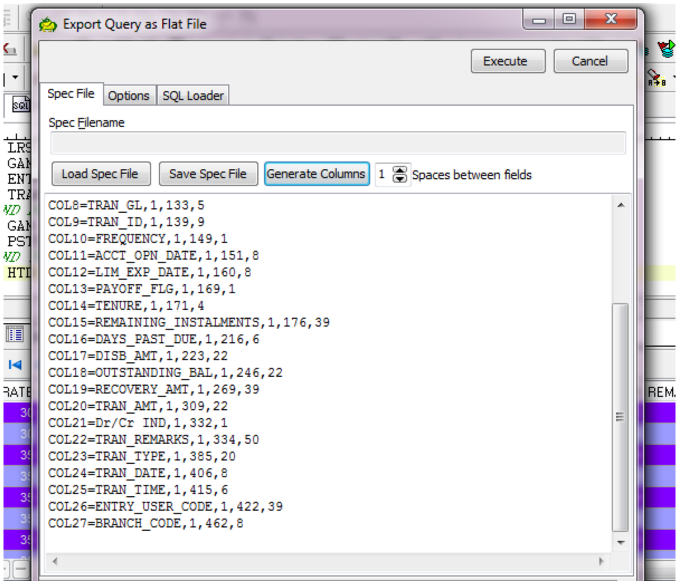
6) *Using Discovered Knowledge*

This step helped us make use of the knowledge acquired to make better decisions. The discovered knowledge can be used for different purposes by interested parties and/or integrated with another models for further action. In our case, we use the discovered knowledge from patterns for further action such as classifying the transactions and also getting the trend analysis for the given patterns of transactions to predict the likelihood of fraud repeating.

VI. EXPERIMENT AND RESULTS

A. *Data set*

To do our experimentation, we extracted the data from the core banking system using the data extraction tool which we developed in PL/SQL block of statements. Discussing details such as system name, operating system platform and version of our data source system is out of scope for this study.



```
COL8=TRAN_GL,1,133,5
COL9=TRAN_ID,1,139,9
COL10=FREQUENCY,1,149,1
COL11=ACCT_OPN_DATE,1,151,8
COL12=LIM_EXP_DATE,1,160,8
COL13=PAYOFF_FLG,1,169,1
COL14=TENURE,1,171,4
COL15=REMAINING_INSTALMENTS,1,176,39
COL16=DAYS_PAST_DUE,1,216,6
COL17=DISB_AMT,1,223,22
COL18=OUTSTANDING_BAL,1,246,22
COL19=RECOVERY_AMT,1,269,39
COL20=TRAN_AMT,1,309,22
COL21=Dr/Cr IND,1,332,1
COL22=TRAN_REMARKS,1,334,50
COL23=TRAN_TYPE,1,385,20
COL24=TRAN_DATE,1,406,8
COL25=TRAN_TIME,1,415,6
COL26=ENTRY_USER_CODE,1,422,39
COL27=BRANCH_CODE,1,462,8
```

Fig 3. Generating PL/SQL query to extract our data from source involving transactions on loans, savings and fixed deposits.

B. *Setup*

We then applied the procedure outlined earlier on in section V (B). We employed K-means clustering algorithm to mine the hidden patterns, Principal Component Analysis (PCA) technique to reduce number of dimensions in the data and Logistic Regression algorithm to classify the transactions for predictive analysis. For implementation of k-means, PCA and logistic regression we used Python v3.7.0 (64-bit) with PyCharm v2018.2.4 Integrated Development Environments (IDE), Jupyter notebook v5.7.8 embedded in Anaconda-Navigator IDE which we run with Tensorflow2.0 Beta. We then used the scikit-learn free machine learning library. To get the time series analysis for forecasting, we used Tableau Desktop 2018.1.17.

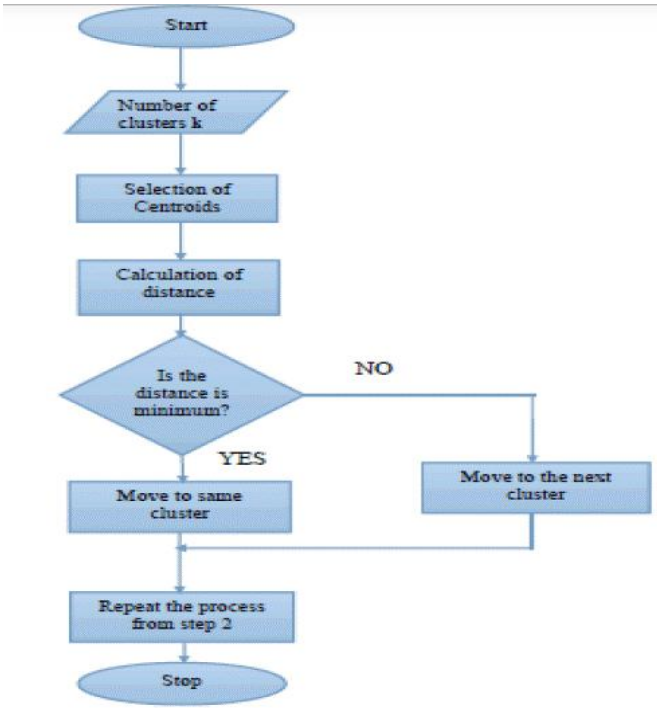
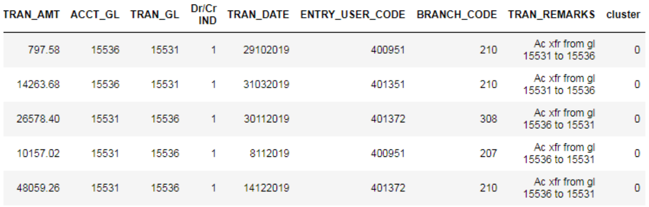


Fig 4. Workflow of our k-means clustering algorithm

The kmeans clustering algorithm was run on transactions for customer loans, then savings and lastly fixed deposit accounts.

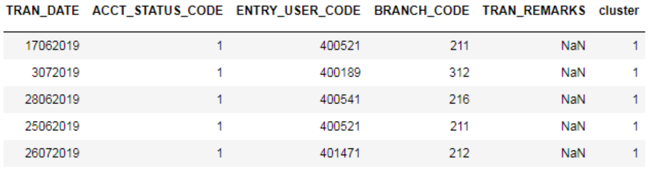
C. *Discussion and Results*

After applying our k-means algorithm, clusters were formed in the truncations which gave interesting patterns of transactions. Our results gave clusters with prevalence of transactions on paid-off loans, transactions on inactive/dormant accounts, balance transfers between loans, transactions without remarks, transactions with similar amounts passed on inactive/dormant accounts etc



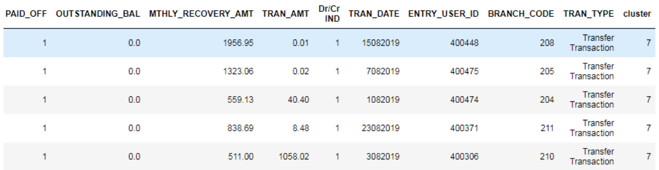
TRAN_AMT	ACCT_GL	TRAN_GL	Dr/Cr IND	TRAN_DATE	ENTRY_USER_CODE	BRANCH_CODE	TRAN_REMARKS	cluster
797.58	15536	15531	1	29102019	400951	210	Ac xfr from gl 15531 to 15536	0
14263.68	15536	15531	1	31032019	401351	210	Ac xfr from gl 15531 to 15536	0
26578.40	15531	15536	1	30112019	401372	308	Ac xfr from gl 15536 to 15531	0
10157.02	15531	15536	1	8112019	400951	207	Ac xfr from gl 15536 to 15531	0
48059.26	15531	15536	1	14122019	401372	210	Ac xfr from gl 15536 to 15531	0

Fig 5. Cluster 0 with balance transfers between loans, having different account and transaction GL. These patterns are unusual because a loans should only be settled or fully paid off unlike transferring a balance to other accounts.



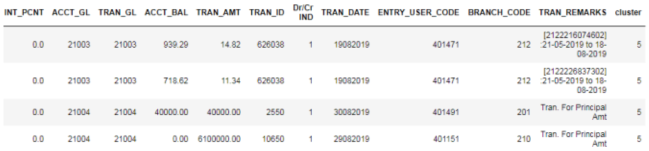
TRAN_DATE	ACCT_STATUS_CODE	ENTRY_USER_CODE	BRANCH_CODE	TRAN_REMARKS	cluster
17062019	1	400521	211	NaN	1
3072019	1	400189	312	NaN	1
28062019	1	400541	216	NaN	1
25062019	1	400521	211	NaN	1
26072019	1	401471	212	NaN	1

Fig 6. Cluster 2 with patterns of transactions on savings accounts without remarks or narrations. Normal transaction should have sufficient details including narrations or remarks.



PAID_OFF	OUTSTANDING_BAL	MTHLY_RECOVERY_AMT	TRAN_AMT	Dr/Cr IND	TRAN_DATE	ENTRY_USER_ID	BRANCH_CODE	TRAN_TYPE	cluster
1	0.0	1956.95	0.01	1	15082019	400448	208	Transfer Transaction	7
1	0.0	1323.06	0.02	1	7082019	400475	205	Transfer Transaction	7
1	0.0	559.13	40.40	1	1082019	400474	204	Transfer Transaction	7
1	0.0	838.69	8.48	1	23082019	400371	211	Transfer Transaction	7
1	0.0	511.00	1058.02	1	3082019	400306	210	Transfer Transaction	7

Fig 7. Cluster 7 with transaction patterns on paid-off loans. These transactions are unusual because once a loan is paid off, there should not be any other transaction taking place as the balance is 0.



INT_PCNT	ACCT_GL	TRAN_GL	ACCT_BAL	TRAN_AMT	TRAN_ID	Dr/Cr IND	TRAN_DATE	ENTRY_USER_CODE	BRANCH_CODE	TRAN_REMARKS	cluster
0.0	21003	21003	939.29	14.82	626038	1	19082019	401471	212	[2122216074602] :21-05-2019 to 18-08-2019	5
0.0	21003	21003	718.62	11.34	626038	1	19082019	401471	212	[2122226837302] :21-05-2019 to 18-08-2019	5
0.0	21004	21004	40000.00	40000.00	2550	1	30082019	401491	201	Tran. For Principal Amt	5
0.0	21004	21004	0.00	6100000.00	10650	1	29082019	401151	210	Tran. For Principal Amt	5

Fig 8. Cluster 5 containing fixed deposits which had 0 applicable Interest percent but interest amounts were paid.

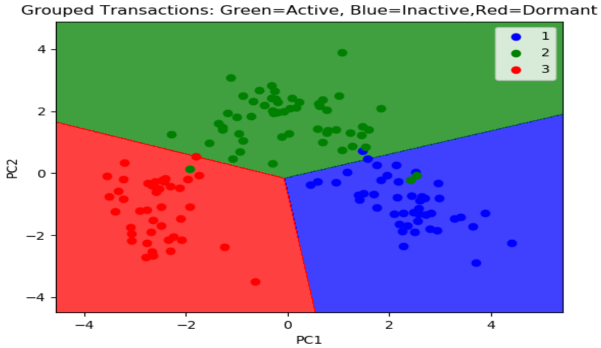


Fig 9. Patterns showing presence transactions on active(green), inactive(blue) and dormant(red) accounts

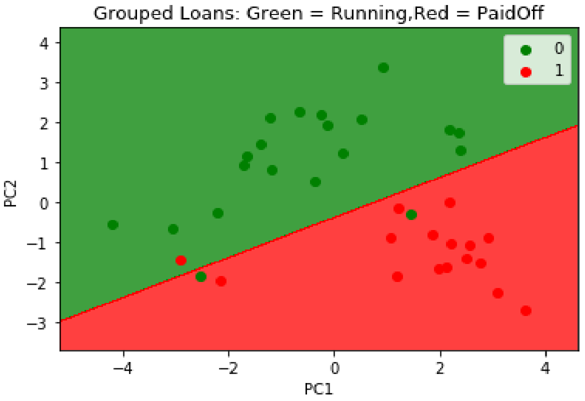


Fig 10. Patterns showing ppresence of transactions on both running loans (green) and paid-off loans (red).

Table 2. Summary of most significant patterns discovered

Pattern Description	**Observed on**
Transactions involving fully paid off loans	Loans
Transactions on inactive/dormant accounts	Savings
Balance transfers between loans with different account GL and transaction GL	Loans
Transactions without remarks or narrations	Savings
Similar amounts passed on dormant accounts	Savings
Interest amounts on accounts with 0 interest	Term Deposits
Same customer operative accounts being used for recoveries on different loans	Loans
Different recovery amounts from expected monthly repayment amount	Loans
Transactions on insufficient balances	Savings

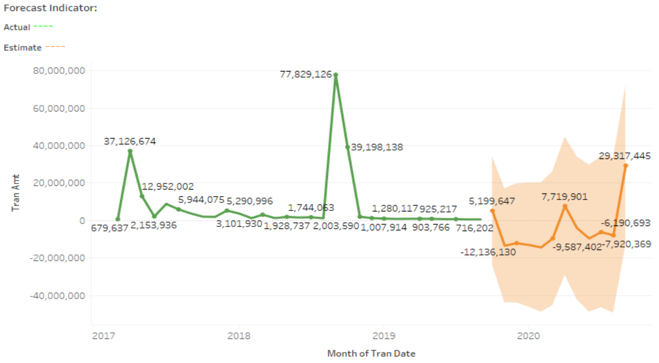


Fig 11. Trend and Seasonality of the unusual transactions happening on loans with summed up values of tran_amts monthly. Green gives actual trend whereas orange is estimate of patterns of unusual transactions in measures that are likely to be continued into the future (uptrend).

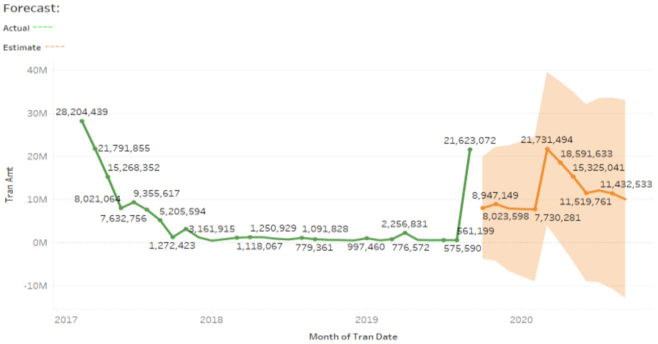
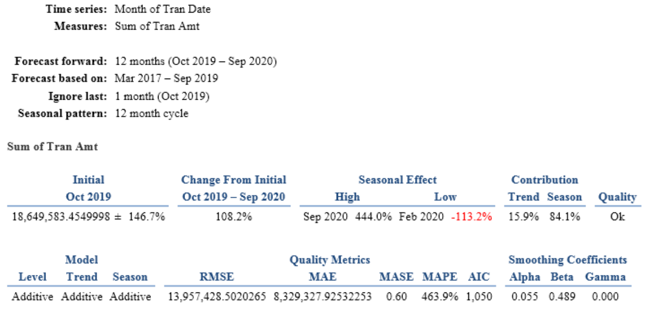


Fig 12. Trend and Seasonality of the unusual patterns of transactions on savings with summed up values of tran_amts monthly. Green is actual trend whereas orange is trend estimate in measures that are likely to be continued into the future with an uptrend.



Time series: Month of Tran Date
Measures: Sum of Tran Amt

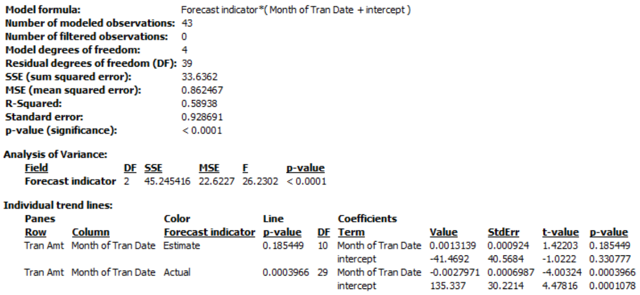
Forecast forward: 12 months (Oct 2019 – Sep 2020)
Forecast based on: Mar 2017 – Sep 2019
Ignore last: 1 month (Oct 2019)
Seasonal pattern: 12 month cycle

Sum of Tran Amt

Initial	Change From Initial	Seasonal Effect				Contribution		
Oct 2019	Oct 2019 – Sep 2020	High		Low		Trend	Season	Quality
18,649,583.4549998 ± 146.7%	108.2%	Sep 2020	444.0%	Feb 2020	-113.2%	15.9%	84.1%	Ok

Model			Quality Metrics					Smoothing Coefficients		
Level	Trend	Season	RMSE	MAE	MASE	MAPE	AIC	Alpha	Beta	Gamma
Additive	Additive	Additive	13,957,428.5020265	8,329,327.92532253	0.60	463.9%	1,050	0.055	0.489	0.000

Fig 13. Summary description of our forecasting model on loan transactions. All forecasts were computed using exponential smoothing. Our Mean Absolute Scaled Error (MASE) was 0.60. MASE is a measure of the accuracy of forecasts and anything below 0.8 = OK, above 0.8 =POOR, below 0.4 = GOOD and above 1 = BAD.



```
Model formula:                       Forecast indicator*( Month of Tran Date + intercept )
Number of modeled observations:      43
Number of filtered observations:     0
Model degrees of freedom:            4
Residual degrees of freedom (DF):    39
SSE (sum squared error):             33.6362
MSE (mean squared error):            0.862467
R-Squared:                           0.58938
Standard error:                      0.928691
p-value (significance):              < 0.0001

Analysis of Variance:
   Field              DF  SSE        MSE      F        p-value
   Forecast indicator 2   45.245416  22.6227  26.2302  < 0.0001

Individual trend lines:
   Panes                         Color               Line                Coefficients
   Row       Column              Forecast indicator  p-value     DF  Term                Value       StdErr     t-value   p-value
   Tran Amt  Month of Tran Date  Estimate            0.185449    10  Month of Tran Date  0.0013139   0.000924   1.42203   0.185449
                                                                     intercept           -41.4692    40.5684    -1.0222   0.330777
   Tran Amt  Month of Tran Date  Actual              0.0003966   29  Month of Tran Date  -0.0027971  0.0006987  -4.00324  0.0003966
                                                                     intercept           135.337     30.2214    4.47816   0.0001078
```

Fig 14. A linear trend model is computed for natural log of sum of Tran Amt (actual & forecast) given Tran Date Month. The model may be significant at p <= 0.05. The factor Forecast indicator may be significant at p <= 0.05. Our model gave p-value = 0.0354183 as our significance.

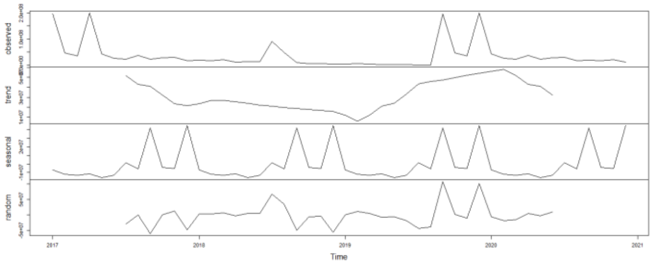


Fig 15. Decomposition of additive time series in our forecasting

VII. CONCLUSION AND FUTURE WORK

In this study, we developed a data mining model for fraud prediction and forecasting based on the discovered hidden patterns in large scale transactions. We employed data mining tasks such as clustering, regression and time series analysis from which we were able to get the hidden patterns of transactions as well as getting the trend and seasonality of discovered patterns. Some transactions may be legitimate and others not therefore, depending on the severity and business rules of the bank, sound decisions should be made and action taken accordingly.

We therefore, recommend that financial institutions should adopt and use Data Mining to uncover hidden patterns and trends in transactions which are very useful for knowledge discovery. Data mining will find answers to business problems and identify causes of business problems thereby providing actionable insight and highly profitable results through uncovering of hidden patterns of transactions.

For future work, we plan to incorporate Text Mining implementation in our model in order to also discover the common words that are being used when carrying out transactions.

REFERENCES

[1] P. Barson and R. Frank. *Fraud Auditing and Forensic Accounting*. John Wiley & Sons, 2nd edition, 1998.

[2] Fraud and Risk Management Working Group. *Fraud Risk Management: A Guide to Good Practice*. CIMA, 2008.

[3] JD. Ratley, BC. Melancon, DA. Richards. *Managing the Business Risk of Fraud: A Practical Guide*. IIA, AICPA, ACFE.

[4] J. F. Elder G. Miner R. Nisbet, J. Elder. Handbook of statistical analysis and data mining applications. In *London: Academic Press*, 2009.

[5] M. Sheetz H. Silverstone, S. Pedneault and F. Rudewicz. *Forensic Accounting and Fraud Investigation - CPE Edition*. The CPE Store, 3rd edition, 2012.

[6] C. Albrecht W. Albrecht, C. Albercht and M. Zimbelman. *Fraud Examination*. South-Western Cengage Learning, Mason, USA, 2009.

[7] B. Rajdeepa and D. Nandhitha. "Fraud Detection in Banking Sector using Data Mining". *International Journal of Science and Research (IJSR)*, 4(7), July 2013.

[8] A. Shmais and Hani R. Data Mining for Fraud Detection. *Prince Sultan University, Saudi Arabia*.

[9] J. Agrawal S. Agrawal. Survey on Anomaly Detection using Data Mining Techniques. *Procedia Computer Science*, (60):708–713, 2015.

[10] M. Albashrawi. Detecting Financial Fraud Using Data Mining Techniques: A Decade Review from 2004 to 2015. *Journal of Data Science*, (14):553–570, December 2016.

[11] DT. Larose and CD. Larose. *DISCOVERING KNOWLEDGE IN DATA: An Introduction to Data Mining*. John Wiley & Sons, Inc, second edition, July 2014.

[12] J. Vohra and E. Jyoti. Data Mining Approach for Retail Knowledge Discovery. *International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE)*, 6(3), March 2016.

[13] D.Bansal and L. Bhambhu. Execution of APRIORI Algorithm of Data Mining Directed Towards Tumultuous Crimes ConcerningWomen. *International Journal of Computing and Technology (IJCAT)*, 3(9), September 2013.

[14] M. Mwanza. FRAUD DETECTION ON BIG DATA USING BUSINESS INTELIGENCE, DATA MINING TOOL. A CASE OF ZAMBIA REVENUE AUTHORITY. Master’s thesis, School of Engineering, The University of Zambia, January 2017.

[15] Y. Fu. Data Mining: Tasks, Techniques, and Applications. *University of Missouri - Rolla*.

[16] WideSkills. Data Mining Tutorial: Data Mining Architecture and Process [online]. *Available: http://www.wideskills.com/data-mining-tutorial*, 2019.

[17] R. Bolton and D. Hand. Statistical Fraud Detection: A Review. *Statistical Science*, 3(17):235–255, 2002.

[18] K. Nevala. *The MACHINE LEARNING Primer*. SAS Institute Inc, 100 SAS Campus Drive, Cary, NC 27513-2414, USA, 2017.

[19] J. Hurwitz and D. Kirsch. *Machine Learning: For Dummies*. John Wiley & Sons, Inc, ibm limited edition edition, 2018.

[20] S. Deshpande and V. Thakare. DATA MINING SYSTEM AND APPLICATIONS: A REVIEW. *International Journal of Distributed and Parallel Systems (IJDPS)*, 1(1), September 2010.

[21] A. Berson, S. Smith, and K. Thearling. An Overview of Data Mining Techniques. 2005.

[22] N. Prasad R. Dâ€™Souza S. Singh, A. Karnwal and A. Shenoy. Fraud Detection using Neural Network. *Department of Information & Communication Tenchnology*.

[23] S. Nemeshaev and A. Tsyganov. Model of the forecasting cash withdrawals in the ATM network.

Procedia Computer Science 7th Annual International Conference on Biologically Insipired Cogntive Architectures, BICA 2016, 88:463–468, 2016.

[24] S. Lotfi and S. Boumediene. Using Neural Network to Detect Financial Statements Fraud of Tunisian Bansks.

[25] S.Goele and N. Chanana. DATA MINING TREND IN PAST, CURRENT AND FUTURE. *International Journal of Computing & Business Research*.

[26] S. Subudhi and S. Panigrahi. Use of optimized Fuzzy C-Means Clustering and supervised classifiers for automobile insurance fraud detection. *Journal of King Saud University- Computer and Information Sciences*, September 2017.

[27] S. Kumari and A. Choubey. Credit Card Fraud Detection Using HMM and K-Means Clustering Algorithm. *International Journal of Scientific Research Engineering & Technology (IJSRET)*, 6(6), 2017.

[28] Vaishali. Fraud Detection in Credit Card by Clustering Approach. *International Journal of Computer Applications*, 98(3):0975–8887, July 2014.

[29] B. Ramageri. DATA MININNG TECHNIQUES AND APPLICATIONS. *Indian Journal of Computer Science and Engineering*, 1(4):301–305.

[30] C. Priyadharsini and Dr. A Thanamani. An Overview of Knowledge Discovery Database and Data Mining Techniques. *International Journal of Innovative Research in Computer and Communication Engineering*, 2(1), March 2014.

[31] E.W.T. Ngai, Y. Hu, Y.H. Wong, Y. Chen, X. Sun. The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems*, (50):559–569, August 2016.

[32] D. Hand C. Whitrow D. Weston, N. Adams and P. Juszczak. Plastic Card Fraud Detection using Peer Group Analysis. In *Marchel Publication*, Sydney, Australia, 2007.

[33] D. Montague. Fraud Prevention Techniques for Credit Card Fraud. In *Spring-field Press*, New York, 2006.

[34] S. Thiprungsri. "Cluster Analysis for Anomaly Detection in Accounting Data". *Collected Papers of the 9th Annual Strategic and Emerging Technologies Research Workshop*, July 2010.

[35] E. Kirkos, C. Spathis, Y. Manolopoulos. Data Mining techniques for the detection of fraudulent financial statements. *Expert Systems with Application*, (32):995–1003, 2007.

[36] MapR. Fraud Detection Solution at a Large Regional Bank. 2016.

[37] *ACL Software License Agreement - Galvanize Terms.*