Machine Learning Approaches for Fraud Detection

Tiwari Abhishek vinod[1], Yash Kushagra Tiwari[2], Shubham verma[3], Vivek Awasthi[4]
Student, Babu Banarsi Das University, Lucknow, Uttar Pradesh

***Abstract*—Payments related fraud is a key aspect of cyber-crime agencies and recent research has shown that machine learning techniques can be applied successfully to detect fraudulent transactions in large amounts of payments data. Such techniques have the ability to detect fraudulent transactions that human auditors may not be able to catch and also do this on a real time basis. In this project, we apply multiple supervised machine learning techniques to the problem of fraud detection using a publicly available simulated payment transactions data. We aim to demonstrate how supervised ML techniques can be used to classify data with high class imbalance with high accuracy. We demonstrate that exploratory analysis can be used to separate fraudulent and non-fraudulent transactions. We also demonstrate that for a well separated dataset, tree-based algorithms like Random Forest work much better than Logistic Regression.**

***Index Terms*—Component, formatting, style, styling, insert.**

I. INTRODUCTION

Digital payments of various forms are rapidly increasing across the world. Payments companies are experiencing rapid growth in their transactions volume. For example, PayPal processed ~$578 billion in total payments in 2018. Along with this transformation, there is also a rapid increase in financial fraud that happens in these payment systems. Preventing online financial fraud is a vital part of the work done by cybersecurity and cyber-crime teams. Most banks and financial institutions have dedicated teams of dozens of analysts building automated systems to analyze transactions taking place through their products and flag potentially fraudulent ones. Therefore, it is essential to explore the approach to solving the problem of detecting fraudulent entries/transactions in large amounts of data in order to be better prepared to solve cyber-crime cases.

.

1.1Data and Sources of Data

Due to the private nature of financial data, there is a lack of publicly available datasets that can be used for analysis. In this project, a synthetic dataset, publicly available on Kaggle, generated using a simulator called PaySim is used. The dataset was generated using aggregated metrics from the private dataset of a multinational mobile financial services company, and then malicious entries were injected. (TESTIMON @ NTNU, Kaggle). The dataset contains 11 columns of information for ~6 million rows of data. The key columns available are – • Type of transactions • Amount transacted • Customer ID and Recipient ID • Old and New balance of Customer and Recipient • Time step of the transaction • Whether the transaction was fraudulent or not In the following figure, a snapshot of the first few lines of the data set is presented. Financial institutions generate millions of transactions every day. Among them, fraudulent transactions are extremely rare (often <0.5%), resulting in imbalanced datasets, making fraud detection a challenging task. Fraudsters continuously evolve their strategies, making static rule-based systems ineffective.

Problem Statement:

To design and implement a Machine Learning-based model capable of detecting fraudulent activities effectively from large-scale transaction data, ensuring high accuracy, low false positives, and adaptability to emerging fraud patterns.

Research Questions:

Which ML algorithms perform best for detecting fraud in imbalanced datasets?

How can anomaly detection techniques improve detection of new fraud types?

Can hybrid models combining supervised and unsupervised learning increase accuracy?

What preprocessing techniques are most effective in handling real-world financial datasets?

Constraints:

Highly imbalanced datasets

Noisy and unlabelled data

Need for real-time processing

II. OBJECTIVES

Evolving fraud techniques

The main objectives of this dissertation are:

To study existing fraud detection techniques and identify research gaps.

To analyze various machine learning algorithms for detecting fraudulent activities.
To preprocess and balance datasets using techniques like SMOTE and undersampling.
To implement supervised algorithms (Logistic Regression, Decision Tree, Random Forest, XGBoost).
To apply anomaly detection algorithms (Isolation Forest, One-Class SVM).
To build a hybrid ML model combining supervised and unsupervised approaches.
To evaluate model performance using accuracy, precision, recall, F1-score, ROC-AUC.
To propose an optimized ML architecture for fraud detection.
To validate the model with real-world or open-source datasets (e.g., Credit Card Fraud Dataset – Kaggle).

III. METHODOLOGY / PLANNING OF WORK

This research follows a clear and step-by-step method to study how Machine Learning can help detect fraud. The aim is not to build a full project, but to understand, compare, and analyze different techniques used by researchers.
1. Research Approach
This study uses a quantitative research approach, meaning it focuses on numerical data and model performance. The research compares different machine learning methods to see which ones work best for fraud detection.
2. Review of Existing Studies
A detailed literature review is done to understand what previous researchers have already discovered. This includes learning about:
How fraud is detected
Which ML methods are commonly used
What problems exist (like imbalanced data)
What gaps still need more research
This helps the study know what new value it can add.
3. Dataset Used
The research uses publicly available datasets, such as the Credit Card Fraud Dataset. These datasets already contain real examples of normal and fraudulent transactions. They are chosen because they are reliable and widely used in academic research.
4. Preparing the Data
Before testing the algorithms, the data is cleaned and prepared. This includes:
Removing wrong or missing values
Standardizing the data
Handling imbalance (because fraud cases are very few)
Selecting the most important features
This helps make the results more accurate.
5. Choosing the Algorithms for reading
Several commonly used machine learning models are selected for comparison, such as:
Logistic Regression
Decision Tree
Random Forest
XGBoost
Isolation Forest
One-Class SVM
These models are chosen because they are widely used in fraud detection studies.
6. Running the Experiments
Each model is tested under the same conditions. This means:
All models use the same data
The same evaluation methods are applied
Models are tuned properly to work their best
This makes the comparison fair and trustworthy.
7. Evaluation of Models
The models are evaluated using simple and meaningful measures like:
Precision
Recall
F1-Score
ROC-AUC
These metrics show how well each model identifies fraud without making too many mistakes.
8. Comparison and Discussion
After testing, the models are compared to see:
Which one performs best
How data imbalance affects results
Whether combining multiple methods gives better accuracy
The results are then discussed in relation to previous research findings.
9. Conclusion
Finally, the study summarizes what was learned and suggests future improvements, such as using deep learning or real-time fraud detection systems.

IV. EXPECTED OUTCOMES

The expected outcomes of this project include:
A working ML model capable of detecting fraudulent transactions.
A comparative study of ML algorithms for fraud detection.
A hybrid intelligent detection system with improved accuracy and reduced false positives.
Practical understanding of handling large, imbalanced datasets.
A framework that organizations can use as the basis for real-time fraud detection.
Academic contribution through analysis of modern ML techniques and their applicability in cybersecurity and fintech industries.

References
Ngai et al., "Application of data mining in fraud detection," Expert Systems with Applications, 2011.
Dal Pozzolo et al., "Credit card fraud detection using machine learning," IEEE Symposium on Computational Intelligence, 2015.

3.3 Theoretical framework
Variables of the study contains dependent and independent variable. The study used pre-specified method for the selection ofvariables. The study used the Stock returns are as dependent variable. From the share price of the firm the Stock returns are calculated. Rate of a stock salable at stock market is known as stock price.

Systematic risk is the only independent variable for the CAPM and inflation, interest rate, oil prices and exchange rate are the independent variables for APT model.

Consumer Price Index (CPI) is used as a proxy in this study for inflation rate. CPI is a wide basic measure to computeusualvariation in prices of goods and services throughout a particular time period. It is assumed that arise in inflation is inversely associated to security prices because Inflation is at lastturned into nominal interest rate andchange in nominal interest rates caused change in discount rate so discount rate increase due to increase in inflation rate and increase in discount rateleads todecreasethe cash flow's present value (Jecheche, 2010). The purchasing power of money decreased due to inflation, and due to which the investors demand high rate of return, and the prices decreased with increase in required rate of return (Iqbal et al, 2010).
Equations
The equations are an exception to the prescribed specifications of this template. You will need to determine whether or not your equations should be typed using either the Times New Roman or the Symbol font (please, no other font). To create multilevel equations, it may be necessary to treat the equation as a graphic and insert it into the text after your paper is styled.

Number equations consecutively. Equation numbers, within parentheses, are top-positioned flush right, as in In Eq. 1, the tab stop is positioned to the right. To make your equations more compact, you may use the solidus (/), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in

Note that the equation is centered using a center tab stop. Be sure that the symbols in your equation have been defined before or immediately following the equation. Use "Eq. 1" or "Equation 1," not "(1)," especially at the atthebeginningofasentence: "Equation1is..."

V. RESEARCH METHODOLOGY

The methodology section outline the plan and method that how the study is conducted. This includes Universe of the study, sample of the study,Data and Sources of Data, study's variables and analytical framework. The detailsare as follows;

3.1Population and Sample
KSE-100 index is an index of 100 companies selected from 580 companies on the basis of sector leading and market capitalization. It represents almost 80% weight of the total market capitalization of KSE. It reflects different sector company's performance and productivity. It is the performance indicator or benchmark of all listed companies of KSE. So it can be regarded as universe of the study.Non-financial firms listed at KSE-100 Index (74 companies according to the page of KSE visited on 20.5.2015) are treated as universe of the study and the study have selected sample from these companies.
The study comprised of non-financial companies listed at KSE-100 Index and 30 actively traded companies are selected on the bases of market capitalization.And 2015 is taken as base year for KSE-100 index.

3.2 Data and Sources of Data
For this study secondary data has been collected. From the website of KSE the monthly stock prices for the sample firms are obtained from Jan 2010 to Dec 2014. And from the website of SBP the data for the macroeconomic variables are collected for the period of five years. The time series monthly data is collected on stock prices for sample firmsand relative macroeconomic variables for the period of 5 years. The data collection period is ranging from January 2010 to Dec 2014. Monthly prices of KSE -100 Index is taken from yahoo finance.

3.3 Theoretical framework
Variables of the study contains dependent and independent variable. The study used pre-specified method for the selection ofvariables. The study used the Stock returns are as dependent variable. From the share price of the firm the Stock returns are calculated. Rate of a stock salable at stock market is known as stock price.

Systematic risk is the only independent variable for the CAPM and inflation, interest rate, oil prices and exchange rate are the independent variables for APT model.

Consumer Price Index (CPI) is used as a proxy in this study for inflation rate. CPI is a wide basic measure to computeusualvariation in prices of goods and services throughout a particular time period. It is assumed that arise in inflation is inversely associated to security prices because Inflation is at lastturned into nominal interest rate andchange in nominal interest rates caused change in discount rate so discount rate increase due to increase in inflation rate and increase in discount rateleads todecreasethe cash flow's present value (Jecheche, 2010). The purchasing power of money decreased due to inflation, and due to which the investors demand high rate of return, and the prices decreased with increase in required rate of return (Iqbal et al, 2010).

Exchange rate is a rate at which one currency exchanged with another currency. Nominal effective exchange rate (Pak Rupee/U.S.D) is taken in this study.This is assumed that decrease in the home currency is inverselyassociated to share prices (Jecheche,2010). Pan et al. (2007) studied exchange rate and its dynamic relationship with share prices in seven East Asian Countries and concludethat relationshipof exchange rate and share prices varies across economies of different countries. So there may be both possibility of either exchange rate directly or inverselyrelated with stock prices.Oil prices are positively related with share prices if oil prices increase stock prices also increase (Iqbal et al, 1012).Ataullah (2001) suggested that oil prices cause positive change in the movement of stock prices. The oil price has no significant effect on stock prices (Dash & Rishika, 2011).Six month T-bills rate is used as proxy of interest rate. As investors arevery sensitive about profit and where the signals turn into red they definitely sell the shares. And this sensitivity of the investors towards profit effects the relationship of the stock prices and interest rate, so the more volatility will be there in the market if the behaviors of the investors are more sensitive. Plethora (2002)has tested interest rate sensitivity to stock market returns, and concluded an inverse relationship between interest rate and stock returns. Nguyen (2010) studies Thailand market and found thatInterest rate has aninverse relationship with stock prices.

KSE-100 index is used as proxy of market risk. KSE-100 index contains top 100 firms which are selected on the bases of their market capitalization. Beta is the measure of systematic risk and has alinear relationship with return (Horn, 1993). High risk is associated with high return (Basu, 1977, Reiganum, 1981 and Gibbons, 1982). Fama and MacBeth (1973) suggested the existence of a significant linear positive relation between realized return and systematic risk as measured by β. But on the other side some empirical results showed that high risk is not associated with high return (Michailidis et al. 2006, Hanif, 2009). Mollah and Jamil (2003) suggested thatrisk-return relationship is notlinear perhaps due to high volatility.

3.4Statistical tools and econometric models

This section elaborates the proper statistical/econometric/financial models which are being used to forward the study from data towards inferences. The detail of methodology is given as follows.

3.4.1 Descriptive Statistics

Descriptive Statics has been used to find the maximum, minimum, standard deviation, mean and normally distribution of the data of all the variables of the study. Normal distribution of data shows the sensitivity of the variables towards the periodic changes and speculation. When the data is not normally distributed it means that the data is sensitive towards periodic changes and speculations which create the chances of arbitrage and the investors have the chance to earn above the normal profit. But the assumption of the APT is that there should not be arbitrage in the market and the investors can earn only normal profit. Jarque bera test is used to test the normality of data.

3.4.2 Fama-Mcbeth two pass regression

After the test statistics the methodology is following the next step in order to test the asset pricing models. When testing asset pricing models related to risk premium on asset to their betas, the primary question of interest is whether the beta risk of particular factor is priced. Fama and McBeth(1973)develop a two pass methodology in which the beta of each asset with respect to a factor is estimated in a first pass time series regression and estimated betas are then used in second pass cross sectional regression to estimate the risk premium of the factor. According to Blum (1968) testing two-parameter models immediately presents an unavoidable errors-in-the variables problem.It is important to note that portfolios (rather than individual assets) are used for the reason of making the analysis statistically feasible.Fama McBeth regression is used to attenuate the problem of errors-in-variables (EIV) for two parameter models (Campbell, Lo and MacKinlay, 1997).If the errors are in the β (beta)of individual security are not perfectly positively correlated, the β of portfolios can

be much more precise estimates of the true β (Blum, 1968).

The study follow Fama and McBeth two pass regressionto test these asset pricing models.The Durbin Watson is used to check serial correlation and measures the linear association between adjacent residuals from a regression model. If there is no serial correlation, the DW statistic will be around 2. The DW statistic will fall if there is positive serial correlation (in worst case, it will be near zero). If there is a negative correlation, thestatistic will lie somewhere between 2 and 4. Usually the limit for non-serial correlation is considered to be DW is from 1.8 to 2.2. A very strong positive serial correlation is considered at DW lower than 1.5 (Richardson and smith, 1993).

According to Richardson and smith(1993) to make the model more effective and efficient the selection criteria for the shares in the period are: Shares with no missing values in the period, Shares with adjusted R2 < 0 or F significant (p-value) >0.05of the first pass regression of the excess returns on the market risk premium are excluded. And Shares are grouped by alphabetic order into group of 30 individual securities (Roll and Ross, 1980).

3.4.2.1 Model for CAPM

In first pass the linear regression is used to estimate beta which is the systematic risk.

$$R_i - R_f = (R_m - R_f)\beta \qquad (3.1)$$

Where RiisMonthly return of thesecurity, Rf isMonthly risk free rate, Rm isMonthly return of market and βis systematic risk (market risk).

The excess returns Ri - Rf of each security is estimated from a time series share prices of KSE-100 index listed shares for each period under consideration. And for the same periodthe market Premium Rm - Rfalso estimated. After that regress the excess returns Ri - Rf on the market premium Rm - Rfto find the beta coefficient (systematic risk).

Then a cross-sectional regression or second pass regression is used on average excess returns of the shares and estimated betas.

$$\hat{R}_i = \gamma_0 + \gamma_1\beta_1 + \epsilon \qquad (3.2)$$

Where λ0= intercept, $\hat{R}$Iis average excess returns of security i,βIisestimated be coefficient of security I and Є is error term.

3.4.2.2 Model for APT

In first pass the betas coefficients are computed by using regression.

$$R_i - R_f = \beta_i f_1 + \beta_{i2} f_2 + \beta_{i3} f_3 + \beta_{i4} f_4 + \epsilon \qquad (3.3)$$

Where Ri is the monthly return of stock i,Rf is risk free rate, βi is the sensitivity of stock i with factors and ϵ is the error term.

Then a cross-sectional regression or second pass regression is used on average excess returns of the shares on the factor scores.

$$\hat{R} = \gamma_0 + \gamma_1\beta_1 + \gamma_2\beta_2 + \gamma_3\beta_3 + \gamma_4\beta_4 + \epsilon_i \qquad (3.4)$$

Where$\hat{R}$ is average monthly excess return of stock I, λ = risk premium, β1 to β4 are the factors scores and εi is the error term.

3.4.3 Comparison of the Models

The next step of the study is to compare these competing models to evaluate that which one of these models is more supported by data.This study follows the methods used by Chen (1983), the Davidson and Mackinnon equation (1981) and the posterior odds ratio (Zellner, 1979) for comparison of these Models.

3.4.3.1 Davidson and MacKinnon Equation

CAPM is considered the particular or strictly case of APT. These two models are non-nested because by imposing a set of linear restrictions on the parameters the APT cannot be reduced to CAPM. In other words the models do not have any common variable. Davidson and MacKinnon (1981) suggested the method to compare non-nested models. The study used the Davidson and MacKinnon equation (1981) to compare CAPM and APT.

This equation is as follows;

$$R_i = \alpha R_{APT} + (1-\alpha) R_{CAPM} + e_i \qquad (3.5)$$

WhereRi= the average monthly excess returns of the stock i, RAPT= expected excess returns estimated by APT, RCAPM= expected excess returns estimated by CAPM and α measure the effectiveness of the models. The APT is the accurate model to forecast the returns of the stocks as compare to CAPMif α is close to 1.

3.4.3.2 Posterior Odds Ratio

A standard assumption in theoretical and empirical research in finance is that relevant variables (e.g stock returns) have multivariate normal distributions (Richardson and smith, 1993). Given the assumptionthat the residuals of the cross-sectional regression of the CAPM and the APT satisfy the IID (Independently and identically distribution) multivariate normal assumption (Campbell, Lo and MacKinlay, 1997), it is possible to calculate the posterior odds ratio between the two models.In general the posterior odds ratio is a more formal technique as compare to DM equation and has sounder theoretical grounds (Aggelidis and Maditinos, 2006).

The second comparison is done using posterior odd radio. The formula for posterior odds is given by Zellner (1979) in favor of model 0 over model 1.
The formula has the following form;

$$R = [ESS_0/ESS_1]^{N/2} N^{K_0 - K_1/2} \quad (3.6)$$

WhereESS0iserror sum of squares of APT, ESS1iserror sum of squares of CAPM, Nisnumber of observations, K0is number of independent variables of the APT and K1 isnumber of independent variables of the CAPM.As according to the ratio when;
R> 1 means CAPM is more strongly supported by data under consideration than APT.
R < 1 means APT is more strongly supported by data under consideration than CAPM.

IV. RESULTS AND DISCUSSION

4.1 Results of Descriptive Statics of Study Variables

Table 4.1: Descriptive Statics

Variable	Minimum	Maximum	Mean	Std. Deviation	Jarque-Bera test	Sig
KSE-100 Index	-0.11	0.14	0.020	0.047	5.558	0.062
Inflation	-0.01	0.02	0.007	0.008	1.345	0.510
Exchange rate	-0.07	0.04	0.003	0.013	1.517	0.467
Oil Prices	-0.24	0.11	0.041	0.060	2.474	0.290
Interest rate	-0.13	0.05	0.047	0.029	1.745	0.418

Table 4.1 displayed mean, standard deviation, maximum minimum and jarque-bera test and its p value of the macroeconomic variables of the study. The descriptive statistics indicated that the mean values of variables (index, INF, EX, OilP and INT) were 0.020, 0.007, 0.003, 0.041 and 0.047 respectively. The maximum values of the variables between the study periods were 0.14, 0.02, 0.04, 0.41, 0.11 and 0.05 for the KSE- 100 Index, inflation, exchange rate, oil prices and interest rate.
The standard deviations for each variable indicated that data were widely spread around their respective means.
Column 6 in table 4.1 shows jarque bera test which is used to checkthe normality of data. The hypotheses of the normal distribution are given;
H0 : The data is normally distributed.
H1 :The data is not normally distributed.
Table 4.1 shows that at 5 % level of confidence, the null hypothesis of normality cannot be rejected. KSE-100 index and macroeconomic variables inflation, exchange rate, oil prices and interest rate are normally distributed.
The descriptive statistics from Table 4.1 showed that the values were normally distributed about their mean and variance. This indicated that aggregate stock prices on the KSE and the macroeconomic factors, inflation rate, oil prices, exchange rate, and interest rate are all not too much sensitive to periodic changes and speculation. To interpret, this study found that an individual investor could not earn higher rate of profit from the KSE. Additionally, individual investors and corporations could not earn higher profits and interest rates from the economy and foreign companies could not earn considerably higher returns in terms of exchange rate. The investor could only earn a normal profit from KSE.

Figures and Tables
Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table captions should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation “Fig. 1” in the text and “Figure 1” at the beginning of a sentence.

Use 10-point Times New Roman for figure labels. Use words rather than symbols or abbreviations when writing figure-axis labels to avoid confusing the reader. As an example, write the quantity “Magnetization” or “Magnetization, M,” not just “M.”

Table 1 Table Type Styles

TableHead	TableColumnHead		
	Tablecolumnsubhead	Subhead	Subhead
copy	Moretablecopya		

VI. ACKNOWLEDGMENT

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g.” Avoid the stilted expression “One of us

(R.B.G.) thanks...” Instead, try “R.B.G. thanks.” Put applicable sponsor acknowledgments here; DO NOT place them on the first page of your paper or as a footnote.

REFERENCES

[1] Ali, A. 2001.Macroeconomic variables as common pervasive risk factors and the empirical content of the Arbitrage Pricing Theory. Journal of Empirical finance, 5(3): 221–240.
[2] Basu, S. 1997. The Investment Performance of Common Stocks in Relation to their Price to Earnings Ratio: A Test of the Efficient Markets Hypothesis. Journal of Finance, 33(3): 663-682.
[3] Bhatti, U. and Hanif. M. 2010. Validity of Capital Assets Pricing Model.Evidence from KSE-Pakistan.European Journal of Economics, Finance and Administrative Science, 3 (20).