Machine Learning Strategies for Fraud Prevention in Financial Data

N.Sreeyutha[1], G.Karthik[2], S.Prashanthi[3], Mrs.Revathi[4]

[1,2,3]*Department of Artificial Intelligence and Machine Learning, Sphoorthy Engineering College, Nadergul, India*

[4]*M.tech., Asst. Professor, Department of CSE(AIML), Sphoorthy Engineering College*

Abstract— The rapid expansion of the E-Commerce industry has led to an exponential surge in credit card usage for online transactions. Unfortunately, this growth has also resulted in an increase in fraudulent activities. Detecting fraud within credit card systems has become increasingly difficult for banks. Machine learning techniques play a pivotal role in identifying credit card fraud during transactions. To predict these fraudulent activities, banks employ various machine learning methodologies, leveraging historical data and incorporating new features to enhance predictive accuracy. In this study, we evaluate the effectiveness of three machine learning models—*Logistic Regression, **Decision Tree, and **Support Vector Machine (SVM)*—for credit card fraud detection. Our dataset comprises 3,925,159 credit card transactions sourced from Kaggle. Transactions are labeled as either "genuine" (denoted by "0") or "fraudulent" (denoted by "1"). With 3,921,920 genuine transactions and 3,239 fraud cases, the dataset is imbalanced. To address this, we create a new balanced dataset with 3,239 samples for training and testing the models. Our evaluation focuses on accuracy. The results indicate the following accuracy rates for the three models:- Logistic Regression: 92.47%, Decision Tree: 99.21%, SVM : 85.57% Comparatively, the Decision Tree outperforms both Logistic Regression and SVM. This research contributes to the ongoing efforts to combat credit card fraud using predictive analytics, artificial intelligence, and machine learning in real-time applications.

Keywords: Predictive Analytics, Artificial Intelligence, Machine Learning, Streaming Data, Real-Time Applications, Deep Learning

I. INTRODUCTION

Credit card fraud events take place frequently and then result in huge financial losses. The number of online transactions has grown in large quantities and online credit card transactions hold a huge share of these transactions. Therefore, banks and financial institutions offer credit card fraud detection applications much value and demand. Fraudulent transactions can occur in various ways and can be put into different categories. This paper focuses on four main fraud eccasions in real-world transactions. Each fraud is addressed using a series of machine learning models and the best method is selected via an evaluation. This evaluation provides a comprehensive guide to selecting an optimal algorithm with respect to the type of the fraud and we illustrate the evaluation with an appropriate performance measure. Credit card generally refers to a card that is assigned to the customer (cardholder), usually allowing them to purchase goods and services within credit limit or withdraw cash in advance. Credit card provides the cardholder an advantage of the time, i.e., it provides time for their customers to repay later in a prescribed time, by carrying it to the next billing cycle. Credit card frauds are easy targets. Without any risks, a significant amount can be withdrawn without the owner's knowledge, in a short period. Fraudsters always try to make every fraudulent transaction legitimate, which makes fraud detection very challenging and difficult task to detect, in 2017, there were 1,579 data breaches and nearly 179 million records among which Credit card frauds were the most common form with 133,015 reports, then employment or tax-related frauds with 82,051 reports, phone frauds with 55,045 reports followed by bank frauds with 50,517 reports from the statics released by FTC [10].

A. AI: The key to business decisions:

One of the driving factors which is greatly responsible for transforming the banking industry is the artificial intelligence technology and the powerful analytics leading to most informed decisions. Almost every financial company is investing in this growing technology to identify fraud risks boost revenue and value for its customers. Big data is another powerful tool that the power to analyze millions of records and data becomes easy and time saving for the organizations to maintain their profits in the market.

There are various key features of AI technology that leads to advancements and smarter results such as:

- Anomaly Detection: Anomaly detection is basically a technique to detect unusual or undefined patterns in the data that do not direct towards the expected outcome. This technique has been widely used in predictive analytics for analyzing the data, critical or raw points in data and identifying features to extract the best results.
- Contribution Analysis: Contribution analysis provides the users with the reason and context in which the event has occurred and further with its results. It helps to analyze the data, and points that have resulted in that form of data.

B. Real-time predictive analytics

Real-time predictive analytics, which is performed on an aggregated predictive model, is deployed for a run-time prediction of continuous streaming data, leading to a real-time powerful decision. Here every single observation from the raw data, that continuously arrives as streaming data, is used to compute the final predicted outcome and affects business decisions accordingly. The most informed decisions come with real-time financial data and this is fulfilled with the power of streaming analytics which is efficient to picking a bulk of streaming data and perform efficient analytics in a fraction of seconds, helping business decisions. Various challenges and previous methodologies have been described in further sections. Feature selection and predictive modelling have been discussed in the next section, focusing on some selective algorithms commonly used for predictions in real-time financial data. The discussion compares and analyses the studies and research of other authors in the field of prediction models and emphasizes the various combinations of different financial indicators used as predictors in the models and compares the results of the realized study with results of other studies based on different calculation methods.

II. PROPOSED SYSTEM

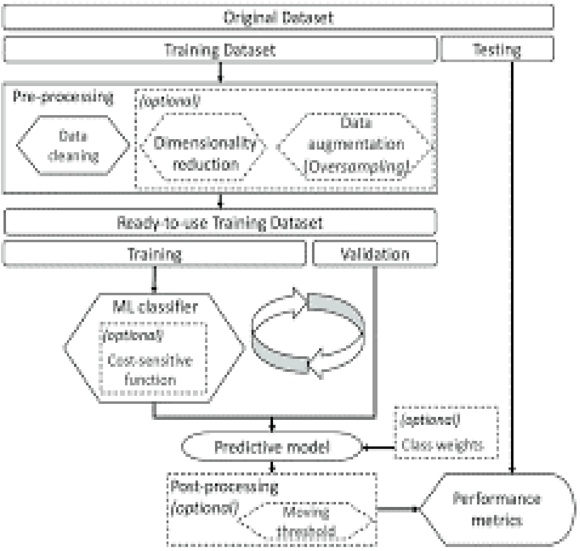

Fig1.System Design

The proposed system for predictive analytics with machine learning for real-time fraud detection in financial data involves a comprehensive architecture. It starts with the ingestion of real-time financial data, followed by preprocessing and feature engineering. Data is processed through a real-time streaming platform, and machine learning models, including algorithms like Random Forests and neural networks, are trained on historical data for fraud detection. Deployed models are containerized and managed with technologies like Docker and Kubernetes. Real-time scoring is implemented, triggering alerts through an alerting system when potential fraud is detected. Continuous monitoring, model interpretability using techniques like SHAP values, and compliance with security standards and regulations are integrated into the system. A feedback loop captures user feedback for model improvement, ensuring ongoing adaptability to evolving fraud patterns. The system's technologies include Apache Kafka for data streaming, Scikit-learn, TensorFlow, and PyTorch for machine learning, Docker for containerization, and Prometheus, Grafana, and ELK Stack for monitoring and logging. The overall design prioritizes accuracy, interpretability, security, and compliance in real-time fraud detection for financial data.

III. LITERATURE REVIEW

Nowadays, the development of technology is rapidly increasing, including the credit card fraud. The credit card fraud (CCF) is one of the problem our banking system is facing today. Fraudsters used many methods to attack the customer. Conventional method of identification based on possession of pin and password are not all together reliable. Machine learning has been proved contributory to solve problems containing sensitive data, such as email spam detection, accurate product recommendation, accurate medical diagnosis, etc. like some technological solutions to the digital world. Lot of organizations are approaching towards machine learning. These technologies solutions, equipped with real- time detection capabilities are enforcing and helping to establish strong supervisory controls helping banks to quickly respond to any type of suspicious activity or pattern and control the financial risk efficiently and at the earliest. Organizations currently using machine learning and predictive analytics are sown in the table below:

TABLE I. FEATURING ORGANIZATIONS USING MACHINE LEARNING AT PRESENT

Organization name	Year	Technological aspects using predictive analytics
Dataiku	2013	Developed ML techniques to analyze raw data and historical patterns for transactions
Teradata	2014	Claim to use an advanced AI platform for analytics
DataRobot	2012	Offers predictive analytics in Finance using the automated machine learning platform
RapidMiner	2007	Data science team to build efficient predictive models using machine learning

IV. SOFTWARE REQUIREMENTS AND SPECIFICATIONS

With digital strategies coping up with banks and financial institutions, enormous data passed to these sectors, business transactions are becoming more prone to frauds and threats resulting in data leakage and personal details exposed to fraudsters leading to huge loss to organizations as well as to customers. This makes organizations adapt to high-level security and data handling technology solutions like machine learning, deep learning and predictive analytics which are efficient enough to deal with highly sensitive data, predict frauds and unwanted behavioral patterns in this data Machine learning has been proved contributory to solve problems containing sensitive data, such as email spam detection, accurate product recommendation, accurate medical diagnosis. like some technological solutions to the digital world. Research in this domain has found a technique, semantic analytics, machine learning task that allows a dataset to be analyzed throughout whether the data is structured, unstructured on tabular form. Lot of organizations are approaching predictive analytics, artificial intelligence, and big dat analytics. These technologies solutions, equipped with real-time detection capabilities an enforcing and helping to establish strong supervisory controls helping banks to quickly respor to any type of suspicious activity or pattern and control the financial risk efficiently and at th earliest.

V. FUNCTIONAL REQUIREMENTS

The functional requirements of this system are mainly based on the user's ease of access.

The system should allow users to create an account.

The system should accept payments made through credit cards. Customer information should be maintained. As soon as the transaction is done, the data should be added to the database to generate the shopping pattern of the user. The algorithm used to create patterns should be efficient. The patterns should be created as soon as the user wishes to purchase a product.

1. Features capturing properties of transactions, including cards used, identifying email addresses used, and location.

2. Features pertaining to customer behaviour, including frequency of orders, how the customer navigates a page before placing an order, and time elapsed between orders

With all the variety of fraudulent schemes involving credit cards, they can be roughly divided into two large groups - identity theft and transaction laundering.

Identity theft

Credit card fraud is the most common form of identity theft, affecting more than 10.7 million people annually. It occurs when someone steals a card or snatches personal information to perform so-called card-not-present (CNP) transactions.

Most commonly, ID thieves use a victim's identity and payment credentials to make purchases a cardholder doesn't authorize, withdraw money from a victim's existing account (account takeover), apply for a new credit card (fraudulent application), or open a new account.

Transaction laundering

This relatively new and advanced method of money laundering is also known as undisclosed aggregation, factoring, or credit card laundering. The fraud involves a legitimate merchant whose credentials are used to process payments for illicit or illegal products and services through a payment card network.

VI. NON-FUNCTIONAL REQUIREMENTS

Performance Requirements

- The system should ensure that user has to be registered user failing which, he will not be given access to any kind of transaction.
- The system should provide fast access to the users.
- Once the fraud is detected, the user should get the confirmation mail within 10 minutes.

At the end of every transaction the system should provide a detailed log of the user's profile from the database. This should be a quick access.

- The shopping pattern should be generated and saved in the database as soon as the user makes a transaction with the credit card

Safety and Security Requirements

User authentication for any transaction can be a safety issue. The system will ensure that every user is authenticated.

At the time of registration, every user's profile should be first checked for all constraints and only then confirmed.

- The system will make sure, no user is given access to any other user's profile and personal data including the transaction details.

Software Quality Attributes

- Reliability: The system will be reliable from the customer's point of view. No transaction will take place without the user login
- Portability: This website will work on all the majorly used operating systems and from all browsers.
- Reusability: The system can be reused. Additional features can be added to it for future scope.
- Flexibility: As mentioned above, the system can be changed or upgraded as per future requirements.
- Robustness: It will be built to be a robust system. The developers should ensure it is error free and can sustain failures if any.

VI. IMPLEMENTATION PROCESS

Data Ingestion:

Utilize data connectors to ingest real-time financial data from multiple sources, including transactions, user activity logs, and historical data repositories. ETL processes involve extracting data from source systems, transforming it to meet the desired structure, and loading it into a target system.

Common ETL tools include Apache NiFi, Apache Airflow, Talend, and Informatica.

Data Preprocessing:

Implement data cleaning processes to handle missing values, outliers, and data anomalies. Conduct feature engineering to extract relevant features such as transaction frequency, user behavior patterns, and historical trends. Identify and handle missing values using techniques such as imputation (replace missing values with a statistical measure) or removal.

Real-time Data Streaming:

Employ a robust data streaming platform like Apache Kafka to process and manage incoming data in real-time. Streaming data refers to a continuous flow of information that is generated and transmitted in real time. It is often characterized by high volume, velocity, and variety.

Machine Learning Models:

Develop and train machine learning models, such as Random Forests, Gradient Boosting Machines, and neural networks, on historical data to identify patterns indicative of fraud.

Model Training and Validation:

Split the data into training and testing sets, utilizing cross-validation techniques to ensure model robustness.

Feature Scaling and Normalization:

Standardize numerical features to ensure uniform scales for accurate model training and prediction.

Model Deployment:

Containerize the trained models using Docker for efficient deployment.

Orchestrate the deployment with Kubernetes for scalability and resource management.

Real-time Scoring:
Implement a real-time scoring mechanism to apply trained models to incoming data streams and generate predictions for potential fraud.

Alerting System:
Set up an alerting system to trigger notifications (email, Slack, etc.) when suspicious activities are detected beyond a certain threshold.

Model Monitoring and Maintenance:
Implement continuous monitoring of model performance using tools like Prometheus and Grafana.
Schedule regular model retraining using updated data to maintain accuracy.

Explainability and Interpretability:
Incorporate SHAP (Shapley Additive exPlanations) values to provide explanations for model predictions, enhancing interpretability.

Compliance and Security:
Ensure SSL/TLS encryption for data in transit and implement access controls to protect sensitive information.
Adhere to financial regulations and security standards, regularly auditing and updating security protocols.

Feedback Loop:
Establish a feedback loop to capture and analyze user feedback, fraud outcomes, and false positives/negatives to iteratively improve model performance.

VII. CONCLUSION AND FUTURE SCOPE

In fraud detection, we often deal with highly imbalanced datasets. For the chosen data set (Paysim), we show proposed approaches can detect fraudulent transactions with very high accuracy and low false positives. Machine learning techniques offer promising opportunities for the prevention and mitigation of fraud in financial markets. In this study, we took an iterative approach by consistently applying three established machine algorithms to identify a relationship between selected features and fraud that would probably not be detected with a human-centered approach to market surveillance.
Here we cannot train the model using an imbalanced data set Random Sampling is done on the data set to make it balanced. The new data set contains an equal number of fraud transactions and legit transactions Now the data set is divided into train data set and a test data set Machine learning techniques like Logistic Regression, Decision Tree, and SVM were used to detect the fraud in the credit card system. Precision, recall, and f1-score used to evaluate the performance of the proposed system. The accuracy for logistic regression, Decision tree, and SVM are 92.47, 99.21, and 85.57 respectively. By comparing all three methods, found that the decision is better than the logistic regression and SVM. This also emphasizes the usefulness of conducting rigorous exploratory analysis to understand the data in detail before developing machine learning models. Through this exploratory analysis, we derived a few features that differentiated the classes better than the raw data.
This is helping companies make better and more informed decisions, reducing loss factors. A lot of applications and developments are empowering Al in financial services and helping organizations serve better to their clients better. Future research in this area and development old's to use automated machine learning and predictive analytics techniques for much better results and cost-saving architecture.

Future Scope:
Advanced Machine Learning Models:
Explore and implement more advanced machine learning models such as deep learning architectures (e.g., neural networks) to capture complex patterns in financial data.

Ensemble Methods:
Experiment with ensemble methods like stacking or blending multiple models to improve overall fraud detection performance.

Continuous Model Improvement:
Implement mechanisms for continuous model improvement by retraining models with incoming data to adapt to evolving fraud patterns.

Explainable AI (XAI):
Incorporate explainable AI techniques to enhance transparency and interpretability of the machine learning models, addressing the need for trust in financial applications.

Integration with External Data Sources:
Enhance fraud detection capabilities by integrating with external data sources, such as consortium databases or industry-specific threat intelligence feeds.

Dynamic Feature Engineering:
Implement dynamic feature engineering techniques that adapt to changing data patterns and characteristics.

Real-Time Data Governance:
Strengthen data governance practices to ensure data quality, integrity, and compliance with regulations in real-time processing.

Scalability and Performance Optimization:
Optimize the system for scalability to handle growing transaction volumes and improve performance to meet low-latency requirements.

Blockchain Technology:
Explore the integration of blockchain technology for secure and transparent transaction verification, enhancing the trustworthiness of financial data.

User Feedback and Collaboration:
Collect user feedback and collaborate with stakeholders to understand real-world challenges and iteratively improve the system based on practical insights.

REFERENCE

[1] K. Nagaraj and A. Sridhar, "A Predictive System for Detection of Bankruptcy Using Machine Learning Techniques", International Journal of Data Mining & Knowledge Management Process, vol. 5, no. 1, pp. 29-40, 2015. doi: 10.5121/ijdkp.2015.5103.

[2] H. Rezaie Doolatabadi, S. Mohsen Hosseini, and R. Tahmasebi, "Using Decision Tree Model and Logistic Regression to Predict Companies Financial Bankruptcy in Tehran Stock Exchanges", International Journal of Emerging Research in Management &Technology, pp. 7- 16, 2013.

[3] R. Guha, S. Manjunath, and K. Palepu, "Predictive Analytics For Insurance Fraud Detection - Wipro", Wipro.com.

[4] S. Aziz and M. Dowling, "Machine Learning and AI for Risk Management", Disrupting Finance, pp. 33-50, 2018. doi: 10.1007/978-3-030-02330-0_3.

[5] T. Le, M. Lee, J. Park, and S. Baik, "Oversampling Techniques for Bankruptcy Prediction: Novel Features from a Transaction Dataset", Symmetry, vol. 10, no. 4, p. 79, 2018. Available: 10.3390/ sym10040079.

[6] M. Li and P. Miu, "A hybrid bankruptcy prediction model with dynamic loadings on accounting-ratio-based and market-based information: A binary quantile regression approach", Journal of Empirical Finance, vol. 17, no. 4, pp. 818-833, 2010. doi: 10.1016/j.jempfin.2010.04.004.

[7] S. Ounacer, M. Amine, S. Ardchir, A. Daif, and M. Azouazi, "A New Architecture for Real-Time Data Stream Processing", International Journal of Advanced Computer Science and Applications, vol. 8, no. 11, 2017. doi: 10.14569/ijacsa.2017.081106

[7] S. Ounacer, M. Amine, S. Ardchir, A. Daif, and M. Azouazi, "A New Architecture for Real-Time Data Stream Processing", International Journal of Advanced Computer Science and Applications, vol. 8, no. 11, 2017. doi: 10.14569/ijacsa.2017.081106

[8] K. Valaskova, T. Kliestik, L. Svabova, and P. Adamko, "Financial Risk Measurement and Prediction Modelling for Sustainable Development of Business Entities Using Regression Analysis", Sustainability, vol. 10, no. 7, p. 2144, 2018. doi: 10.3390/su10072144.

[9] A. Thennakoon, C. Bhagyani, S. Premadasa, S. Mihiranga and N. Kuruwitaarachchi, "Real-time Credit Card Fraud Detection Using Machine Learning", 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2019. doi: 10.1109/confluence.2019.8776942.

[10] A. Abdallah, M. Maarof and A. Zainal, "Fraud detection system: A survey", Journal of Network and Computer Applications, vol. 68, pp. 90-113, 2016. doi: 10.1016/j.jnca.2016.04.007

[11] "4 Major Challenges Facing Fraud Detection; Ways to Resolve Them using Machine Learning", Medium. [Online].