Journal of Business and Management Studies
ISSN: 2709-0876
DOI: 10.32996/jbms
Journal Homepage: www.al-kindipublisher.com/index.php/jbms

JBMS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Explainable and Cost-Sensitive Machine Learning for Fraud Prevention in U.S. Banking: Threshold Optimization, SHAP Explainability, and Fairness Analysis

Kazi Shakhawat Hossain[1] ✉ **Arefin Ahmed Sajib**[2] **and Towsif Alam**[3]
[123]*Student, Business Analytics, Wright State University, Dayton, OH, United States*
Corresponding Author: Kazi Shakhawat Hossain, **E-mail**: hossain.145@wright.edu

| ABSTRACT
Credit card fraud is a major challenge for U.S. financial institutions. Banks need fraud detection systems that can identify fraudulent transactions, reduce false alarms for legitimate customers, and meet regulatory expectations. This study develops an explainable and cost-sensitive machine learning model for detecting credit card fraud using the publicly available Sparkov synthetic credit card transaction dataset from Kaggle. The dataset contains 1,296,675 transactions for training and 555,719 transactions for testing. The research creates and tests four kinds of machine learning models: XGBoost, LightGBM, random forest, and logistic regression. Instead of selecting the model based on accuracy, the study applies decision-oriented threshold optimization to minimize the total economic cost of classification errors. The cost is $1,000 for each false negative that misses a fraud case and $50 for each false positive that blocks a legitimate transaction. The results show that XGBoost has the lowest total economic cost of error with a threshold of 0.75 at $355,700 and catches 91.2% of fraud cases. The study employs SHAP values to identify the key factors influencing fraud prediction and enhance model transparency for banking decision-makers. Furthermore, gender-based fairness analysis is conducted using false negative rates, false positive rates, and true positive rates. The findings show that cost-sensitive optimization, explainable AI, and fairness monitoring can help create more responsible fraud detection systems in the banking sector.

| KEYWORDS
Fraud Detection, Cost-Sensitive Learning, XGBoost, LightGBM, Random Forest, Logistic Regression, SHAP, Explainable AI, Responsible AI, Threshold Optimization, Consumer Protection, U.S. Banking.

| ARTICLE INFORMATION
ACCEPTED: 18 July 2026 **PUBLISHED:** 23 August 2026 **DOI:** 10.32996/jbms.2026.8.9.11

1. Introduction

1.1. Background and Motivation

False fraud alerts can directly affect customer experience. For example, a legitimate customer may experience a declined card transaction at a supermarket because an automated fraud detection system incorrectly classifies the transaction as suspicious. This is not because the customer lacks sufficient funds, but because the bank's fraud detection systems incorrectly classify the transaction as suspicious. This kind of mistake is known as a False Positive. United States banks must deal with millions of False Positives every day. Now imagine someone using a stolen debit card to buy a home theater system at midnight. The bank's computer system does not detect anything out of the ordinary. So, the transaction is approved. This kind of mistake is called a False Negative. False Negatives cost customers and banks a lot of money thousands of dollars.

Credit card fraud has grown rapidly in the United States. The Federal Trade Commission reports that in 2023 United States consumers lost 10 billion dollars because of fraud. Credit card fraud is the common type of identity theft incident. To tackle this problem banks use systems that use intelligence to detect fraud. Most machine learning models try to get better at things like being accurate. Considering this, 99% of all transactions are legitimate. So, a model that approves all transactions will be 99% accurate, which sounds good but is not helpful. This study is not trying to make the model more accurate rather than wanting to

know what the real dollar cost of each decision is. Detecting a fraud case is more important than blocking a transaction. This way of thinking is called cost-sensitive learning. It is the core idea of this study. When a bank uses intelligence to make decisions that affect people's finances and lives those decisions need to be fair and explainable. This is the idea behind Responsible Artificial Intelligence. Computer science and ethics teach us how to build smart computer systems that work properly, explain their decisions and treat customers equally. Banks need to use Artificial Intelligence in a way that's fair and transparent so customers can trust the system.

1.2. Research Objectives

The research has four primary purposes. 1) To develop and compare a set of machine learning models in credit card fraud detection using a large-scale synthetic credit card transaction dataset. 2) To incorporate decision-oriented threshold optimization to reduce the overall costs of fraud detection errors in a machine learning model. 3) To use SHAP-based explainability analysis to discover key attributes for detecting fraudulent transactions and to make the model understandable to the banking industry decision-makers and regulators. 4) To examine the gender fairness of the machine learning model in producing disparate results for different groups of customers.

1.3. Methodology Overview

To create more challenging classification tasks, the research uses feature engineering based on the transaction date, customer demographics, and transaction account address. The study constructs four machine learning pipelines composed of various models from the Python scikit-learn library. The research evaluates models using the standard metrics accuracy, precision, recall, F1 score, ROC curve and AUC, precision-recall curve and AUC, and total cost function. Finally, for model explainability, the paper employed the SHAP Tree Explainer to understand the impact of each feature on the output of the group models.

1.4. Research Contribution

This study contributes to literature in three ways. Firstly, it applies cost-sensitive threshold optimization for fraud detection instead of depending only on default classification thresholds. This allows model selection to reflect the economic consequences of false positives and false negatives. Second, it uses SHAP to make the model clear and interpretable for banking decision-makers and explain the major drivers of fraudulent predictions. Third, it evaluates fairness by gender using false negative and false positive. It also highlights the importance of responsible AI monitoring in fraud detection systems using true positive rates.

1.5. Limitations

The study uses a hypothetical dataset of synthetic bank dat; the dataset may not entirely capture real-world complexity, account takeover fraud, or drifting data based on previous behavior of users over time. The cost of a false negative ($1,000) and the cost of a false positive ($50) are estimated costs, and these costs may vary by different banks and different types of transactions. Future work should extend the analysis to other attributes such as race, age groups, and income levels. Finally, this research does not involve deploying models in a real-world bank setting in terms of expectancy and degradation over time.

1.6. Future Research Directions

Future research can extend this study in several ways. First, researchers should apply the proposed methodology to actual bank data to explore if the conclusions are still applicable in a realistic financial context. Second, future work should study different fairness aware machine learning methodologies such group-specific threshold setting, fairness-controlled optimization, reweighting, and adversarial debiasing. Third, forthcoming research may explore additional aspects of demographic and geographic fairness such as age, income, race or ethnicity, and area if such data are legally, morally, and securely available. Fourth, upcoming research should also focus on real-time fraud detection model implementation and challenges such as model drift, latency, cybersecurity, privacy, and regulatory scrutiny. Finally, future research should investigate dynamic threshold instead of one constant threshold so that fraud detection decision can react to changing fraud trends, customer behavior, and institutional risk appetite.

2. Literature Review

2.1. Fraud Detection in Banking

Bank fraud has been around since banks were first established, but recent technological advancements have greatly increased their sophistication and prevalence. The transition from physical credit card use to online and contactless payment methods has introduced new fraud opportunities (Bolton & Hand, 2002). For decades, banks used rule-based systems to detect fraud, such as triggering alerts for transactions above $500 from an unusual location. As fraudsters learned to manipulate these rules, keeping transactions just below the limit, the industry realized the need for more sophisticated methods, leading to statistical and machine learning approaches (Dal Pozzolo et al., 2018).

This issue is particularly significant for U.S. banks because of the country's absolute size. Each year, Visa and Mastercard process billions of transactions. Among them, the fraud rates are estimated to be 0.1% of all transactions (Nilson Report, 2023). Although this seems low, a 0.1% fraud rate means that tens of millions of transactions are fraudulent each year. Any detection system must deal with this tremendous imbalance and maintain a very low false positive rate so as not to annoy legitimate customers.

2.2. Machine Learning in Fraud Detection

The application of machine learning techniques in fraud detection has seen significant growth since the early 2000s. Transactional fraud detection methods started to employ decision trees, logistic regression, and neural networks (Bhattacharyya et al., 2011). Currently, machine learners such as Random Forests and Gradient Boosting machines are widely used due to their high accuracy. Chen and Guestrin (2016) proposed a machine learning algorithm called XGBoost, which has become very popular in academic and industrial fraud detection studies due to its scalability and learning speed. Ke et al. (2017) proposed LightGBM, a highly efficient gradient boosting framework based on a novel leaf-wise tree growth strategy that is faster than the popular gradient boosting model (XGBoost) on small to medium-sized datasets.

Recurrent neural networks (RNNs), long short-term memory networks (LSTMs), and Transformer models, all based on deep neural networks, have been used extensively for sequential transaction data. But gradient boosting methods outperform deep learning models in terms of interpretability and computational costs in production systems (Fiore et al., 2019).

2.3. Limitations of Accuracy Based Evaluation

One of the most reported issues arising from evaluating fraud detection is the use of accuracy as the evaluation metric. Because legitimate and fraudulent transactions are usually ratios, there are 99 or more legitimate transactions for each fraudulent one, a classifier can achieve 99% accuracy by classifying all transactions as legitimate, although it cannot detect any fraud. Two models with the same accuracy can have very different effectiveness in detecting fraud (Provost and Fawcett, 2013).

Alternative metrics have been suggested by researchers. The Precision-Recall curve and its integrated metric (PR-AUC or Average Precision) are reported to be the most informative measure for imbalanced classification problems, as they focus on classifying the minority class, i.e., fraudulent transactions. The ROC-AUC curve is useful for comparing the performance of two models, but it can be over-optimistic in imbalanced problems because it considers the true negative rate, which is easy to maximize in an imbalanced dataset with millions of legitimate transactions. This research assessed the performance of all models using both ROC-AUC and PR-AUC, but to choose the best model, the study focuses on the total cost since it directly reflects the real-life decisions of a bank.

2.4. Explainable Artificial Intelligence in Fraud Detection

As mentioned before, explainable AI is crucial for fraud detection because of the need for accurate, comprehensible, and fair AI systems in financial institutions. Phillips et al (2021), state that AI systems should provide reason, meaningful explanations, explanations, accuracy, and clear knowledge limits. These principles are highly relevant when AI systems make decisions about fraudulent behavior in banking. The High-Level Expert Group on Artificial Intelligence (2019) in their paper "Ethics Guidelines for Trustworthy AI" mention that trustworthy AI is lawful, ethical, and robust. AI systems used in fraud detection must protect customers from unfair or harmful decisions. In a paper more relevant to fraud detection, Zhou et al (2023) show that SHAP-based explanations help external stakeholders understand why a fraudulent case is predicted while maintaining competitive predictive performance. They provide the closest supporting argument for because they relate model accuracy with explanation, fairness, and accountability. The study discusses the importance of explainable AI in this paper because it reduces fraud losses, helps decision-makers trust the model, and reduces the risk of harming honest customers through unclear automated decisions. This study suggests that explainable fraud detection is not only a technical advantage but also an ethical and practical need for modern banking systems.

2.5. Cost-Sensitive Fraud Detection

Cost-sensitive learning is the process of instructing a learning algorithm about the presence of different misclassification costs. This is crucial in a fraud detection problem since the costs of misclassifying a fraudulent case as genuine and the cost of misclassifying a genuine customer as fraudulent are not the same. A **False Negative** is a system that fails to catch genuine fraud. This can lead to both bank and customer losses, regulatory consequences, and reputation impacts. A **False Positive** is a system that marks a genuine transaction as fraudulent. This can lead to unhappiness among good customers, operational costs, and in some cases, customer loss. The project shows that the cost of missing fraud is usually much higher than the cost of a false alarm (Verhelst et al., 2022).

There are at least two ways to implement cost-sensitivity in machine learning. One method is to consider fraud cases more heavily when training the model to ensure the model focuses more on detecting fraud cases. Another method is to adjust the

classification threshold. For example, instead of classifying a transaction as fraudulent with a prediction score greater than 0.50, the threshold can be moved up or down to find the optimal point where the overall cost is minimal. The study uses the threshold optimization method due to its simplicity, applicability, and no need for retraining of models.

2.6. Customer Impact and False Positives

The impact of fraud detection on customer experience has not received as much attention as detection accuracy. While some researchers focus on the system's ability to detect fraud accurately, customers' experience with (false) alerts has been largely overlooked. Industry reports show that false positives have a major impact on customer satisfaction, and customers switch banks. False positive occurs when a genuine transaction is mistakenly identified as suspicious transaction and blocked by the bank. According to J.D. Power (2023), customers experiencing false declines are more likely to switch banks within the next 90 days. Aite-Novarica Group (2022) estimates the risk U.S. banks face in losing $11.7 billion annually due to customers switching banks after too many false fraud alerts. This number is likely to double or even triple in the presence of a real fraud.

The study supposes that this is important as it supports the implementation of a cost-sensitive approach. One model that detects more fraud is not necessarily superior. For example, one model may detect 95% of fraud but generate 10,000 false alarms per day, whereas another model may detect only 85% of fraud but generate only 1,000 false alarms. The second model may be preferable due to the reduced harm it causes. This is why the research utilizes the total economic cost function in this paper to help banks balance the cost of false alarms and the loss from fraud to select the most suitable model parameters for their customers and business objectives.

2.7. Research Gap

Prior fraud detection studies often focus on fraud detection accuracy and cost-sensitive learning, or explainability separately. However, fewer studies combine them with Explainable AI and fairness checking into one comprehensive system. Most studies on fraud detection focus either on maximizing the amount of detected fraud or on minimizing the cost. They do not fully consider the explainability requirement for the fraud detection model (Bahnsen et al., 2016). But, in the current landscape, responsible AI is becoming increasingly important in the banking sector. Regulatory guidance from agencies such as the Consumer Financial Protection Bureau (CFPB) and the Office of the Comptroller of the Currency (OCC) suggests banks should develop predictive models that are not only accurate and cost-effective, but also explainable, auditable, and equitable to different customer groups. This paper addresses this gap by proposing a comprehensive framework that integrates accuracy, cost, explainability, and fairness.

3. Conceptual Framework

The foundation of this study comprises two pillars: the cost structure of fraud detection errors and the principles of responsible, explainable AI deployment.

The first pillar reflects the fact that each prediction a fraud model makes leads to real costs. If the model misses a fraud (False Negative), then the costs will be the amount of the stolen transaction plus the costs of the investigation and chargeback, and potential legal fees. The study assumes the total cost per False Negative to be $1,000. If the model incorrectly predicts a legitimate transaction as fraud (False Positive), then there will be costs from the call center and friction with the customer, and the customer may switch to a different bank. The study assumes the cost per False Positive to be $50. The total economic cost of a fraud detection system can therefore be expressed as:

$$\textit{Total Economic Cost} = (\textit{Number of False Negative} \times \$1{,}000) + (\textit{Number of False Positive} \times \$50)$$

This formula helps the model figure out the best idea for making decisions. The goal is to lower the total cost, not merely to get more accurate.

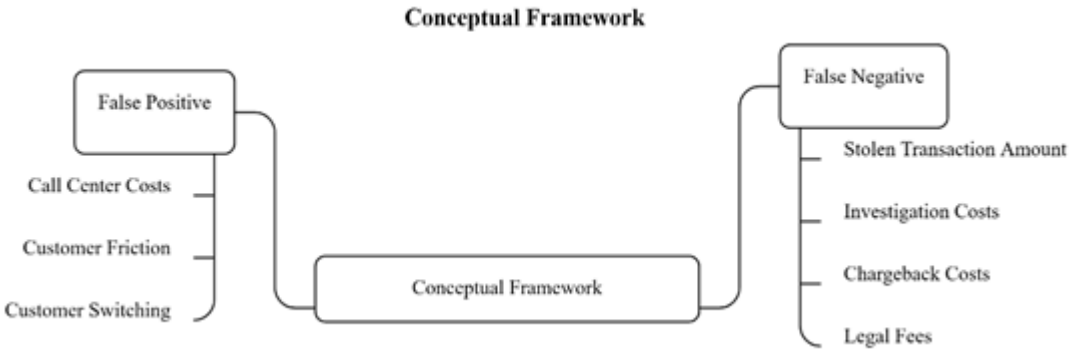


Figure 1: Conceptual Framework

The second main aspect of the proposed framework is Explainable AI (XAI) and Responsible AI (RAI). The fraud detection model should explain why they have marked a specific transaction as fraud. It should indicate which features in that transaction contributed to that classification and to what extent. This is necessary for bank officers since they need to explain and record their decisions. This is also necessary for customers to know why their transaction is stopped. SHAP values help in this regard. They provide a fair way of explaining an individual transaction.

Responsible AI means that the model should not produce unfair or disproportionate harm to any customer group. It treats all groups equally and causes no more damage to one group than to others. In the banking context, it can be discriminatory because of their gender, race, age, or other similar factors, which can also be against the law. Therefore, this framework will also test the fairness of the model. It compares the model's error rate with male and female clients. This gives a first idea of whether the models treat fairness towards male and female groups in the fair way.

4. Methodology and Data Description

4.1. Methodology

4.1.1 Data Sources and Tools

A public practice dataset of credit card transactions from Sparkov Data Generation (2020) serves as the foundation for this research. The study selected it for investigation, because this dataset is frequently employed in fraud detection studies. This dataset is publicly available on Kaggle. It has information about U.S. credit card transactions from January 2019 to December 2020. The authors divide the dataset into training and test data. The training data set has 1,296,675 transactions, and the test data set has 555,719 transactions.

This research uses Python 3.11. The Pandas and NumPy libraries are used for data processing and calculation. For this fraud detection model, this study relies on Scikit-learn, regression, random forest, XGBoost, and LightGBM. The SHAP library is used to explain the model predictions. Microsoft Power BI is used to make maps and charts, while Microsoft Excel is used to make tables and summarize the results.

Table 1: Total Dataset Records

Item	Value
Training records	1,296,675
Testing records	555,719
Total modeled records	1,852,394

4.1.2 Feature Engineering

The study has used the original transaction dataset, which offers a lot of information about each transaction. The transaction includes time, the credit card number, the merchant ID, the amount spent, the customer's name, birthday, gender, and job, as well as the customer's location and the merchant's location. The original data set also has another characteristic called Is_Fraud, which shows if anything is fraudulent or not.

Several variables were renamed for clarity and consistency before feature engineering. The transaction time was taken out so that the learning model could figure out when fraud is most likely to happen. It's helpful to remember that most fake transactions happen late at night when not many people are awake to spot transactions right away. The weekday and month were also taken from the date. The age was derived from the customer by matching the date of their birth with the date of the purchase. This attribute helps the model figure out if fraud happens more often to younger or older people.

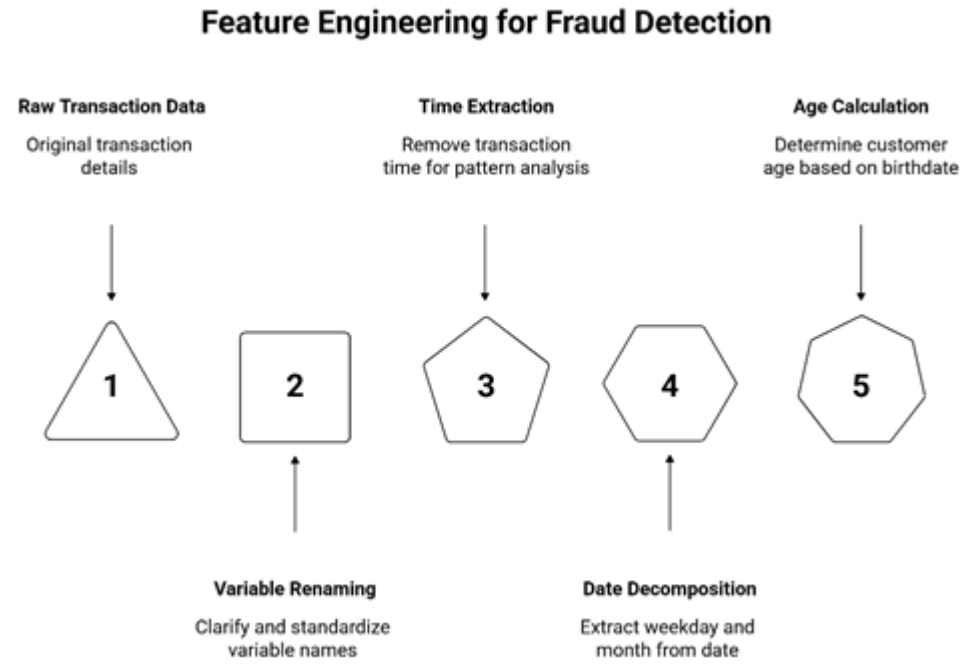


Figure 2: Feature Engineering

4.1.3 The Preprocessing Pipeline

The preprocessing pipeline separates the columns into two groups: Gender, Purpose, Merchant, City_Name, and State_Name in the category column, and Amount, Transaction_Hour, Transaction_Month, Age, Customer_Lat, Customer_Long, Merchant_Lati, and Merchant_Long in the numerical column. It uses the median value to fill in the gaps in the columns. It then adjusts the missing data, allowing the model to interpret them more easily. When a value is missing from a set of category columns, it uses the most common text to fill the gaps. To make them machine-readable, it encodes the words and breaks them down into many smaller columns. The two groups are then combined in a single preprocessor, which is linked to the logistic regression, random forest, XGBoost, and LightGBM models to clean data before learning.

Figure 3: Processing Pipeline

4.1.4 Model Specification

Four supervised machine learning models were trained and compared in this study: Logistic Regression, Random Forest, XGBoost, and LightGBM. All four models processed the data through the same sequence of preprocessing steps within the pipeline, which enabled us to judge the performance of the models fairly. The first model named Logistic regression is the simplest and most straightforward model of the lot. The models fit logistic regression with class_weight = 'balanced' to account for class imbalance and max_iter = 1000 to control the maximum number of repetitions. Logistic regression is the benchmark model in this comparison. The second model is Random Forest. It consists of 200 decision trees trained with class_weight = 'balanced'. Random Forest is a popular benchmark fraud detection algorithm. It is capable of capturing nonlinear interactions between features and does not overfit.

The third model is XGBoost, a model trained using the eXtreme Gradient Boosting library, comprising 300 trees (estimators) with a maximum depth of 6 and a learning rate of 0.08. To balance for the imbalanced dataset, the model used scale_pos_weight = the number of legitimate transactions to the number of fraud transactions (approximately 172:1), which instructs XGBoost to penalize errors made on fraudulent transactions more than legitimate transactions during training. The fourth model is LightGBM. It is a Microsoft gradient boosting framework with 300 estimators, a maximum depth of 8, and a learning rate of 0.05. It used the balanced class weight setting. LightGBM grows trees leaf-wise, contrary to the level-wise strategy used by XGBoost. The results were that it converged faster and attained higher accuracy than XGBoost when the dataset was large.

4.1.5 Evaluations Metrics

The study evaluates each model using standard classification metrics, including accuracy, precision, recall, F-1-score, ROC-AUC, and PR-AUC. Accuracy that is the proportion of correct predictions out of all predictions, Precision, which is the proportion of predicted fraudulent cases that are truly fraud, Recall is the proportion of truly fraudulent cases that are detected, F1 Score that is harmonic mean of precision and recall, ROC-AUC measures the model's discrimination ability across thresholds, while PR-AUC

is especially useful for imbalanced classification problems. In addition to the above metrics, the study evaluated all models based on the Total Economic Cost function defined in section 3, which has the greatest operational impact of model performance for a bank.

4.1.6 Decision-Oriented Threshold Optimization

By default, in scikit learning a transaction is fraudulent when the model's predicted probability more than is 0.50. If the threshold is higher than default value, the model is less like labeling a case as fraud and more likely to treat it as non-fraud. If the threshold is lower than the default value, the model is more likely labeling a case as fraud. The study performs the following steps to identify the optimal threshold for each model: (i) extract the predicted probability scores from each model on the test set; (ii) resample these predicted scores 99 times by an increment of 0.01, starting from 0.01 up to 0.99; (iii) generate a confusion matrix for each resampled threshold value; (iv) calculate the cost of confusion matrix using cost formula (Cost = (FN × $1,000) + (FP × $50)) with the assumed cost parameters; (v) identify the threshold with the minimum cost and consider it as the optimal threshold to resample an instance into fraud class for the corresponding model.

4.1.7 SHAP Explainability

SHAP is a tool that helps us understand why a machine learning model calls a transaction fraud or not fraud. It shows how each feature, such as amount, time, or location, affects the model's decision for one transaction. In XGBoost, TreeExplainer can calculate these SHAP values quickly and accurately. A positive SHAP value means the feature pushes the prediction more toward fraud. A negative SHAP value means the feature pushes the prediction more toward non-fraud or legitimate. In this way, SHAP explains which factors increase or decrease the fraud score. The SHAP summary plot ranks features based on how strongly they affect the model on average. This helps bank analysts quickly see the most important warning signs of fraud in the dataset.

4.1.8 Fairness Analysis

The study analyzed the fairness of the XGBoost model for male and female responders. It used three indicators to do this. There are false positive rates, the false negative rate, and the true positive rate. The false positive (FP) rate quantifies the frequency with which non-fraud transactions are falsely predicted to be fraudulent. The false negative (FN) rate tells you how often fraud transactions are wrongly thought to be non-fraud. The true positive (TP) rate tells you how well the model finds fraudulent transactions. These three criteria should have values that are close to the same for both men and women. If there are big variations among these indicators for men and women, it could mean that one group is being treated unfairly.

4.2 Data Description

4.2.1. Dataset Overview

It presents a summary of the dataset structure. The training set contains 1,296,675 transactions and the test set contains 555,719 transactions. It is a combined total of 1,852,394 observations. The 30 columns in the raw dataset were cleaned and renamed for clarity. The target variable, Is_Fraud, is binary: 1 indicates a fraudulent transaction and 0 indicates a legitimate one.

Feature	Training Set	Test Set
Total Transactions	1,296,675	555,719
Fraudulent Transactions	7,506 (0.578%)	2,145 (0.386%)
Legitimate Transactions	1,289,169 (99.42%)	553,574 (99.61%)
Time Period	Jan 2019 – Dec 2020	Jun 2020 – Dec 2020
Unique Merchants	High Cardinality	High Cardinality
Unique States	51 (50 + DC)	51 (50 + DC)
Avg. Transaction Amount	$67.62	$67.66
Avg. Fraud Transaction Amount	$527.44	$530.12
Total Features (after engineering)	30 raw + derived	Same

Table 2: Dataset Overview

4.2.2. Target Variable

The target variable Is_Fraud has a value of 1 to indicate fraud and a value of 0 to indicate a normal transaction. Frauds are very rare in the dataset. There are around 172 normal transactions for every 1 fraud case. Hence, the dataset is highly imbalanced. Because of this, the model might bias towards normal transactions while detecting fraud cases. To avoid this problem, the study used a few balancing techniques such as the class_weight parameter, the scale_pos_weight parameter in the XGBoost model, and

oversampling or using an appropriate classification threshold. These methods force the model to pay extra attention to the minority classes (fraud cases).

4.2.3. Key Feature Description

The key features used in the model training, their types, and their relevance to fraud detection are summarized below.

Feature Name	Type	Relevance to Fraud
Amount	Numeric	Fraudulent transactions average $527 vs $67 for legitimate
Transaction Hour	Numeric	Fraud peaks between 12 AM and 4 AM
Transaction Month	Numeric	Seasonal patterns in fraud activity
Age	Numeric	Older customers more targeted in certain fraud types
Gender	Categorical	Used in fairness analysis; minor predictive value
Purpose (Category)	Categorical	Certain merchant categories have higher fraud rates
Merchant	Categorical	Specific merchants associated with fraud patterns
City Name	Categorical	Some cities show 100% fraud rates in small-sample data
State Name	Categorical	NY, TX, PA show highest absolute fraud volumes
Customer Lat/Long	Numeric	Distance from home signals unusual purchase location
Merchant Lat/Long	Numeric	Used together with customer location for distance features

Table 3: Feature Description

4.2.4. Geographical Distribution of Fraud

A geographic location analysis of fraudulent transactions shows important regional patterns. It shows the top 10 states by total fraud volume in the training dataset. New York State leads with 555 fraudulent transactions, followed by Texas (479), Pennsylvania (458), California (326), and Ohio (321). These states are also among the most populous in the U.S. It suggests that total fraud volume is partially driven by transaction volume.

Table 4: Geographical Distribution of Top 10 Fraud States

Rank	State	Fraud Count
1	New York (NY)	555
2	Texas (TX)	479
3	Pennsylvania (PA)	458
4	California (CA)	326
5	Ohio (OH)	321
6	Florida (FL)	281
7	Illinois (IL)	248
8	Michigan (MI)	238
9	Alabama (AL)	215
10	Minnesota (MN)	207

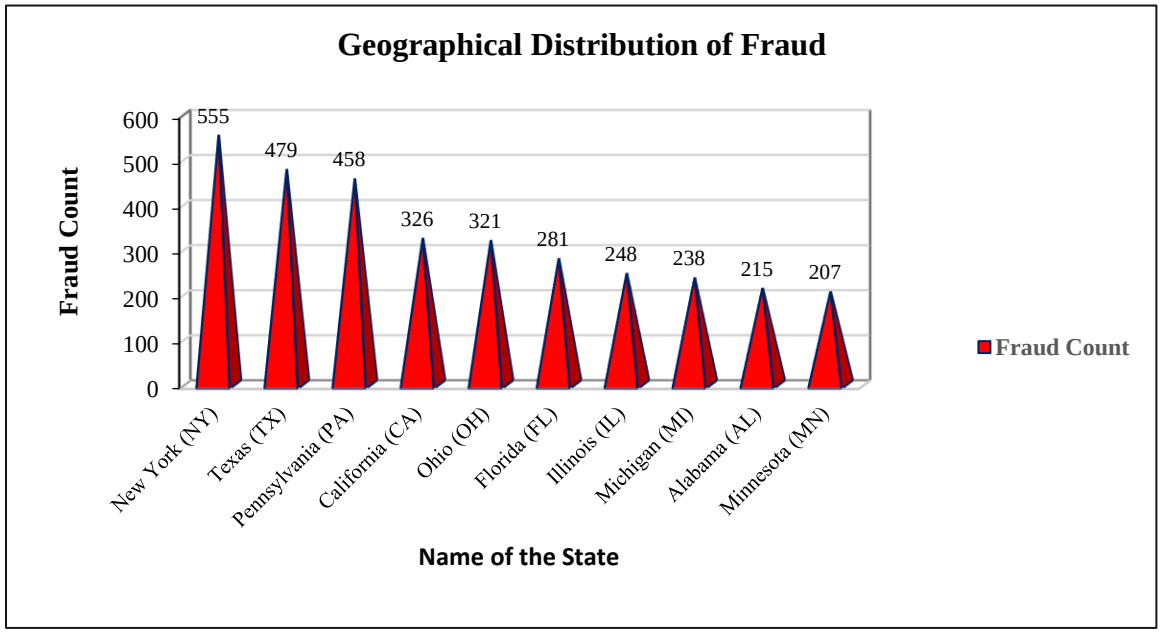


Figure 4: Geographical Distribution of Fraud

4.2.5 Temporal Patterns

The analysis of fraud rates by hour of the day shows a distinct temporal pattern. The fraud rate is highest in the early morning hours from 12 AM to 4 AM when most cardholders are asleep and unlikely to notice suspicious activity. It decreases to its lowest level during regular business hours from 9 AM to 6 PM and increases again during late evening hours. This temporal pattern is also observed in the literature (Chan et al., 1999 and Dal Pozzolo et al., 2018) and confirms that the hour of the transaction is a useful feature.

The average transaction amount across the two classes is different. The values are $67 and $527 for normal and fraudulent transactions in the training dataset, respectively. It means that fraudulent transactions are almost eight times larger than normal ones. Hence, the amount is one of the most relevant features in the dataset. This is because when they gain access to card fraud will attempt to make them spend as much money as possible before the card is blocked.

5. Results

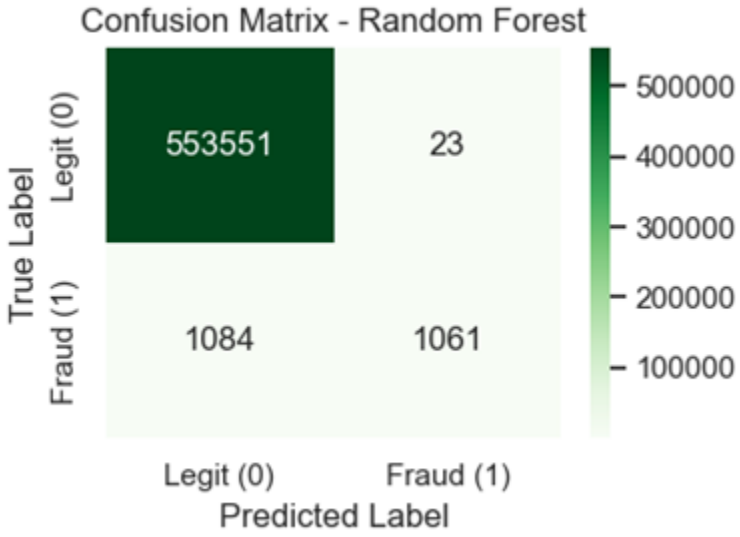
5.1 Model Performance at Default Threshold (0.50)

Model	True Negative (TN)	False Positive (FP)	False Negative (FN)	True Positive (TP)	Recall
Logistic Regression	504,138	49,436	1,522	623	29.04%
Random Forest	553,551	23	1,084	1,061	49.46%
XGBoost	547,348	6,226	114	2,031	94.69%
LightGBM	545,791	7,783	98	2,047	95.43%

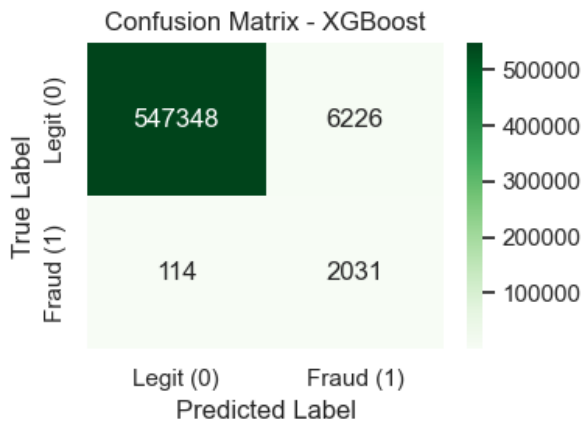
Table 5: Confusion Matrices at Default Threshold (0.50)

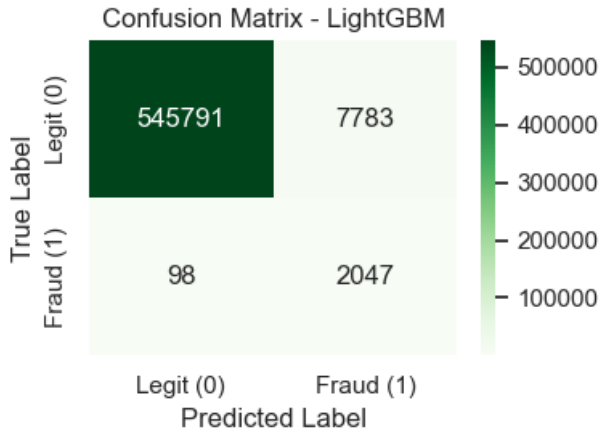
The confusion matrix of the four models at the default decision threshold of 0.50 is shown in Table 5. With logistic regression, the model detects only 623 out of 2,145 fraud transactions at the default decision threshold of 0.50. That means that its recall is 29.04% with 49,436 false alarms. This result is not unexpected, as without any class weight balance a simple logistic regression model will fail when faced with this kind of severely imbalanced data.

The Random Forest model produces the fewest false positives, that is only 23, but at the cost of missing 1,084 fraud cases with recall is 49.46%.



In terms of recall, the best models are XGBoost and LightGBM at the default decision threshold of 0.50. XGBoost detects 2,031 out of 2,145 fraud cases with 94.69% recall, and LightGBM detects 2,047 of them with 95.43% recall though both produce several thousand false positives.





5.2. Standard Metrics Comparison

The evaluation results for the four models with the default threshold are shown in Table 5. XGBoost performed the best on the two most important metrics for imbalanced fraud detection among the models tested with a ROC-AUC of 0.9969 and a PR-AUC of 0.8415. Then LightGBM came next with the same ROC-AUC of 0.9969 but a slightly lower PR-AUC of 0.8286. In terms of the F1 Score, Random Forest outperforms the other three models with a value of 0.6572, due to its extremely low false positive rate, but its recall rate is only 49.46%, which means half of the fraudulent transactions would be missed. This is unacceptable from the consumer protection perspective. Logistic Regression performs the worst on all the metrics with a ROC-AUC of 0.6067 and a PR-AUC of 0.0146. It indicates that the probability of fraud predicted by this model is not well calibrated.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	ROC-AUC	PR-AUC
XGBoost	98.86	24.59	94.69	39.05	0.9969	0.8415
LightGBM	98.58	20.82	95.43	34.19	0.9969	0.8286
Random Forest	99.80	97.88	49.46	65.72	0.9691	0.8151
Logistic Regression	90.83	1.24	29.04	2.39	0.6067	0.0146

Table 6: Full Model Performance Comparison

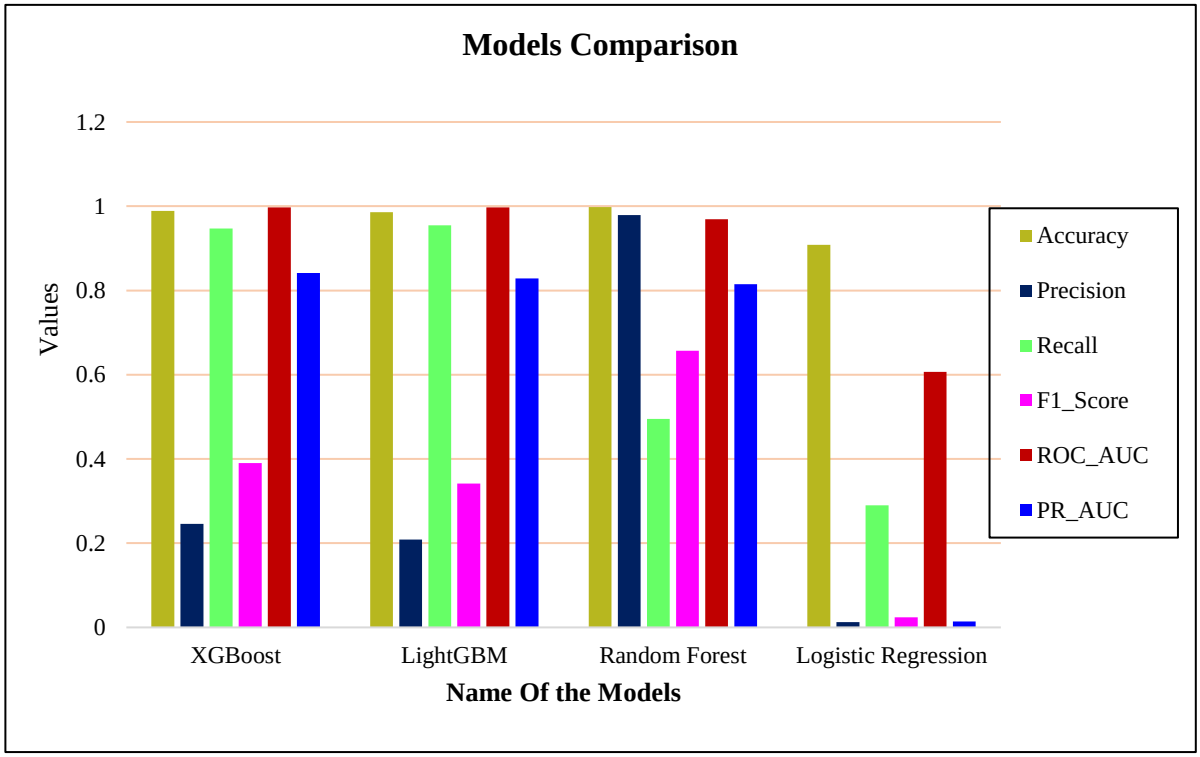


Figure 5: Full Model Performance Comparison in Bar Chart

5.3. Threshold Optimization Result

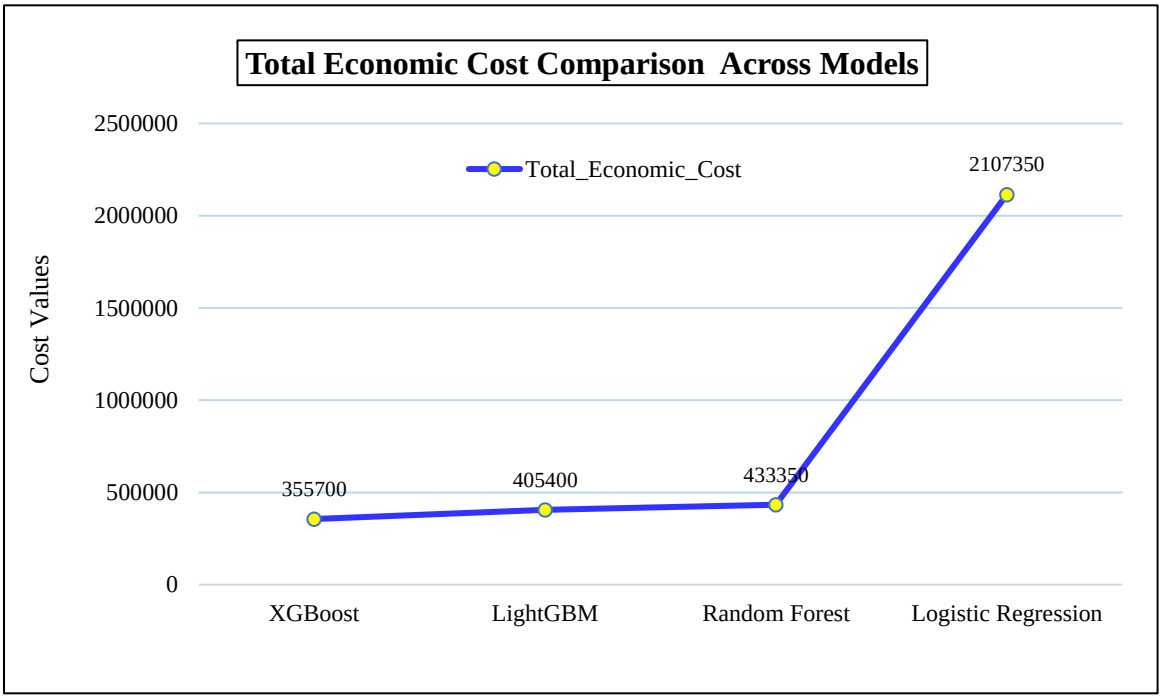
A threshold optimization procedure determines the optimal cutoff threshold for each model. Here, the economic cost function is exercised to identify the threshold value associated with the minimum total economic cost. In this study, the cost per false negative error is set at $1,000, while the cost per false positive error is $50. Table 7 below summarizes the economic cost-related results for the default threshold and the optimal threshold for all models. It shows that the total economic cost is lower at the optimal threshold than at the default threshold for all models. At optimal threshold is 0.75, XGBoost calculates the lowest total economic cost is $355,700. At the optimal threshold, XGBoost detects 1,956 fraud cases out of 2,145 fraud cases with 3,334 false positives. LightGBM has the second overall performance with $405,400 total cost. The optimal threshold of the Random Forest is 0.10 much smaller than the default threshold. This implies that the probabilities of fraud detection provided by the Random

Forest model are mostly with a low value, and a low threshold is required to detect more fraud cases. Logistic Regression has the poorest performance even at the optimal threshold. Its optimal threshold is 0.94, and its total economic cost is $2,107,350, which is almost six times that of XGBoost. It is not a proper choice for an actual fraud detection system in this study.

Model	Best_Threshold	FP	FN	TP	Total_Economic_Cost
XGBoost	0.75	3334	189	1956	355700
LightGBM	0.75	3848	213	1932	405400
Random Forest	0.10	1111	378	1767	433350
Logistic Regression	0.94	3607	1927	218	2107350

Table 7: Results After Decision-Oriented Threshold Optimization

Figure 6: Decision Oriented Threshold



The graph illustrates the overall economic cost of XGBoost for 99 different threshold values tested between 0.01 to 1. It shows that the economic cost of the model reduces when the threshold increases from 0.01 and reaches its minimum value at 0.75. After this, it increases again when the threshold becomes too high and it starts misclassifying more fraud cases. This figure depicts the cost-sensitive nature of the model and helps the bank risk managers see the impact of slight variations in the threshold level on the overall economic cost of fraud detection.

5.4. ROC Curve Comparison for All Models

The ROC curve comparison shows once more how the linear model and the ensemble-based models do things differently. Both XGBoost and LightGBM have the greatest AUC of 0.997, which shows that they are very good at telling the difference between real and fake transactions across all types of forecasts. With an AUC of 0.969, Random Forest comes in second. This means it is not as good as the two models. On the other hand, Logistic regression has a much lower AUC of 0.606, which shows that it can't tell fraud patterns apart very well. The ROC curve shows that the non-linear models, especially XGBoost and LightGBM, are far better at finding fraud in this dataset than logistic regression.

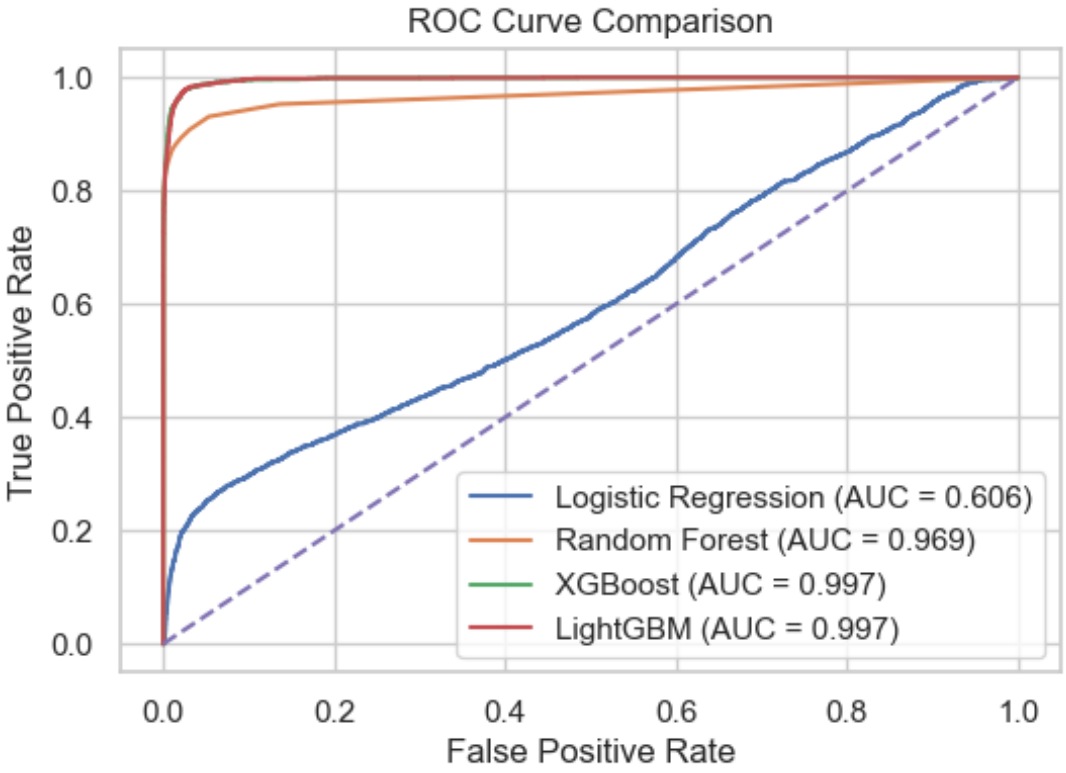


Figure 7: ROC Curve Comparison for All Models

5.5. Precision-Recall Curve Comparison for All Models

The Precision-Recall curve can better quantify how well fraud detection works because the percentage of fraudulent transactions is very low in the whole data set. The chart shows that XGBoost is the best of the four models, with the highest PR AUC of 0.841. This indicates that XGBoost is the best model for finding the most fraudulent transactions while making the fewest valid transactions look like they were fraudulent. LightGBM has a PR AUC of 0.829, which is close to the best, and Random forest has a PR AUC of 0.815, which is also good. This shows that the ensemble tree-based models work well for this fraud detection task. In the end, Logistic regression had the worst outcome, with a PR AUC of 0.015. This shows that a basic linear model can't capture the complicated and non-linear correlations that fraud detection needs to find. The picture reveals that XGBoost is the best model for finding fraud, with LightGBM and Random Forest as close second and third best models.

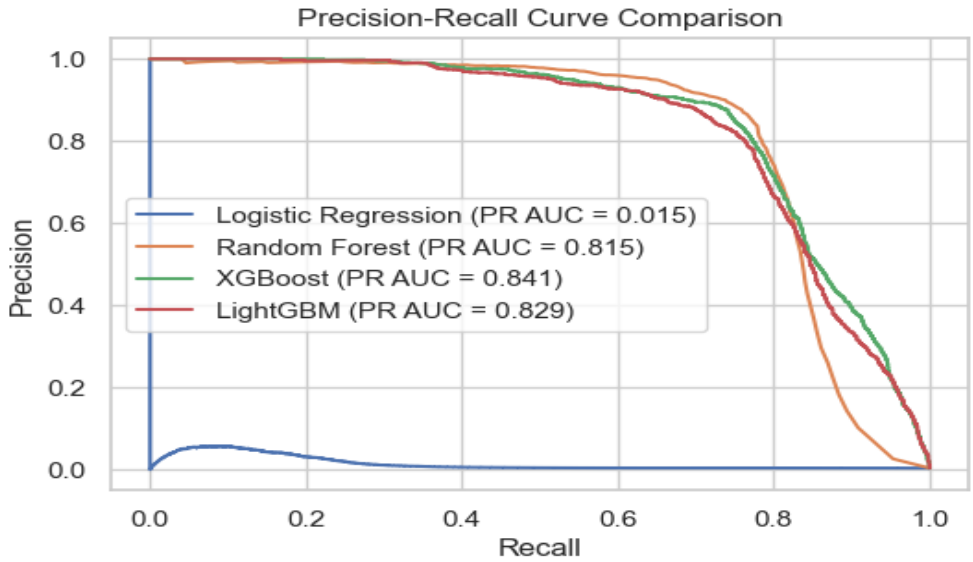


Figure 8: Precision-Recall Curve Comparison for All Models

5.6. SHAP Explainability Analysis

The Summary Plot provides more insights into the prediction of fraud by the XGBoost model. According to this SHAP summary plot, the most important feature in influencing the prediction of the XGBoost model is the "Transaction Month". It appears that fraudulent transactions have a seasonal trend such that some months are more vulnerable to fraud than others. The second most important feature in predicting the fraud by the XGBoost model is "the transaction amount". The research considers very large transactions are suspected in the eyes of the model. The third important factor for the XGBoost model is the transaction purpose, is the buyer carrying petrol or going to a restaurant or going shopping? This information helps the XGBoost model to gain insight into which types of transactions are the most vulnerable to fraud. The time of the transaction is also an important feature. Transactions performed during odd hours are suspect in the eyes of the XGBoost model. Overall, this plot shows that the XGBoost model suspects the behavior mainly based on when the transaction was performed, the transaction amount, and the type of purchase.

SHAP Summary Plot:

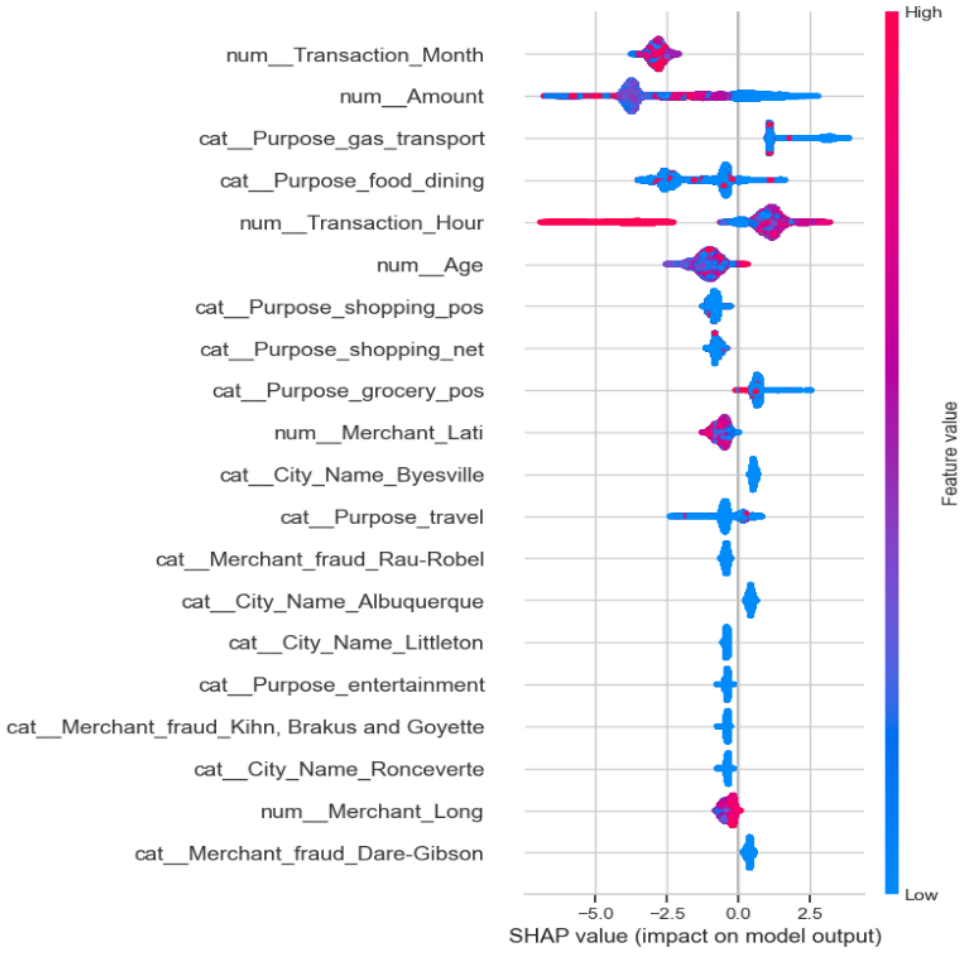


Figure 9: SHAP Explainability Analysis-Summary Plot

SHAP Bar Plot:

The Bar Plot depicts the features' importance using a SHAP bar plot. Transaction Month has the highest mean impact on the model prediction, followed by Transaction Amount, Gas Transport, Food Dining, Transaction Hour, and Age. These features are the most important to the model in predicting whether a transaction is fraudulent. Other features such as city name, merchant name, and merchant location also contribute to the prediction of fraudulent transactions, but to an insignificant extent. As shown in the SHAP summary plot and SHAP bar plot, the most important features that push the model to predict fraudulent transactions are when the transaction is performed, how much money is spent with the transaction, and what purchase the consumer is taking. These results suggest that the XGBoost model is learning from the behavior of transactions rather than the identity or location of the consumer.

SHAP Bar Plot:

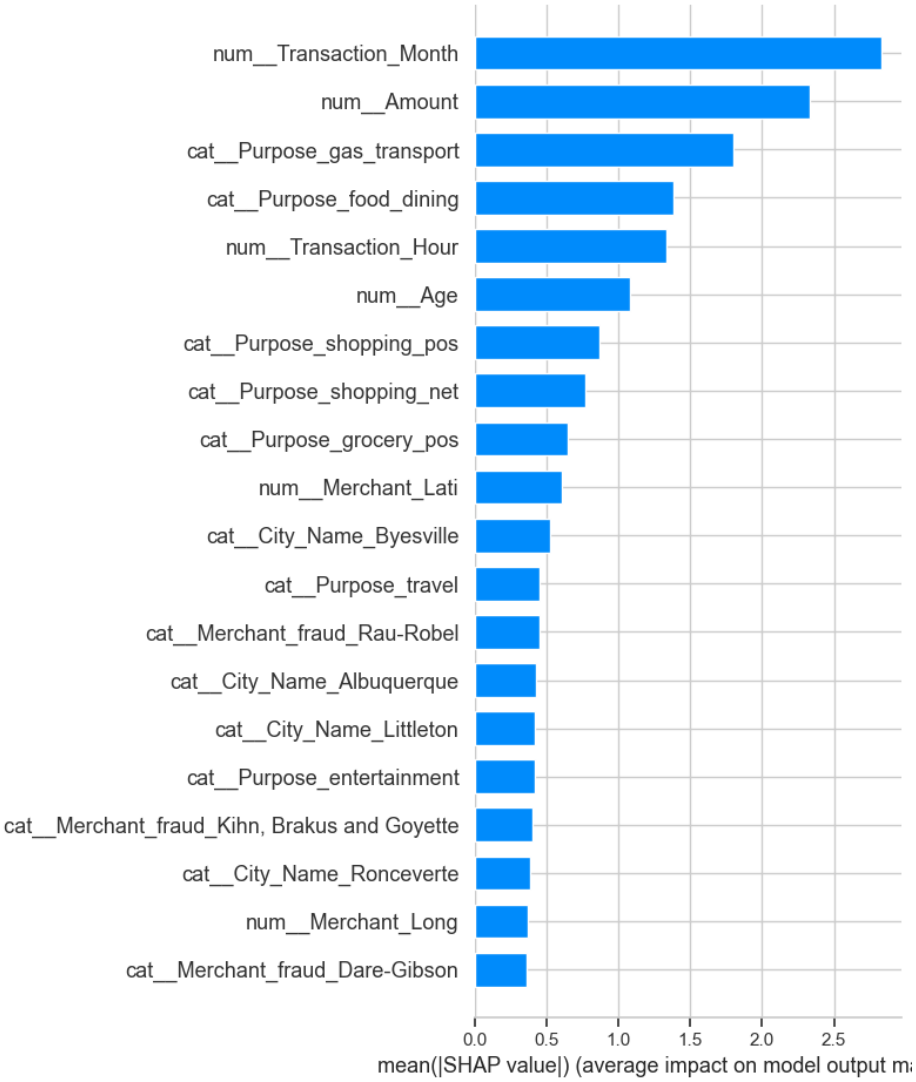


Figure 10: SHAP Explainability Analysis-Bar Plots

5.7. Fairness Analysis by Gender

Gender	False Positive Rate (FPR)	False Negative Rate (FNR)	True Positive Rate (TPR/Recall)
Male (M)	0.48%	4.49%	95.51%
Female (F)	0.70%	12.46%	87.54%

Table 8: Fairness Analysis by Gender — XGBoost at Threshold 0.75

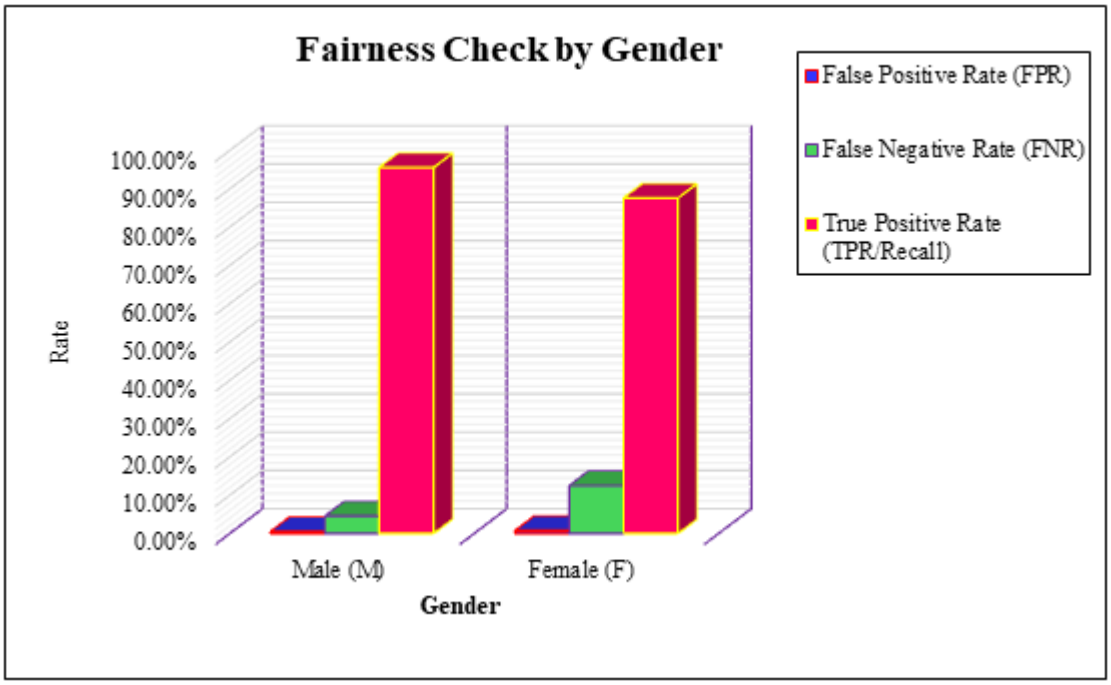


Figure 11: Fairness Analysis

The fairness analysis results at the best threshold 0.75 of the XGBoost model for male and female customers are provided in Table 8. The results suggest a potential gender-based performance difference within the synthetic dataset. The false positive rate for male customers is 0.48%, which means 1 in every 208 legitimate male customers' transactions is incorrectly blocked by the fraud detection system. For female customers, the false positive rate is 0.70%, which means 1 in every 142 female customers' legitimate transactions is incorrectly blocked by the fraud detection system, indicating a slightly higher friction among female customers.

The differences in the rates of false negatives are more notable from a fairness perspective. The recall (true positive rate) for male customers is 95.51%, meaning that the model detects 95.5% of fraud against male cardholders. The recall for female customers is 87.54%, meaning that the model misses about 12.5% of fraud against female cardholders. This suggests that fraud against female customers has a slightly different pattern than that of the training data, perhaps because female customers use different types of merchants, or purchase at different times or in different amounts compared to their male counterparts, and the model has not learned the signature of fraud in such cases. This result demonstrates the importance of continuous fairness testing of live fraud detection systems and is a direction of future work to test whether such differences continue to exist after model tuning for each demographic group.

5.8. Fraudulent Transaction Locations

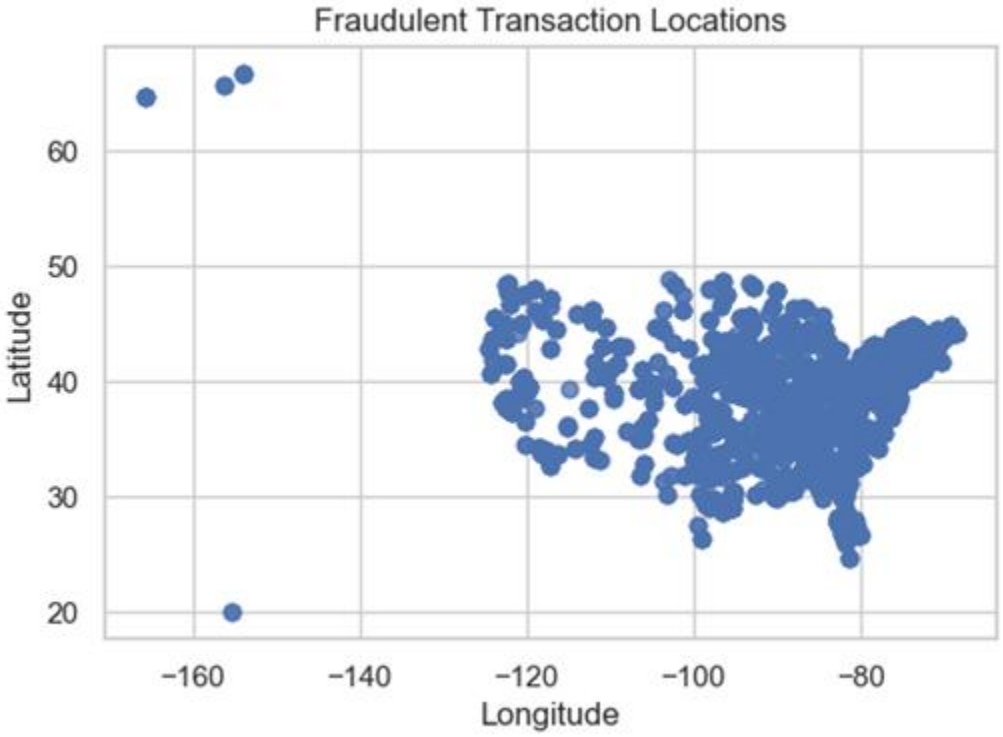


Figure 12: Fraudulent Transaction Locations

Most of the fraudulent transactions scattered plot points are between the longitude of -125 and -70 and the latitude of 30 and 48 in the mainland US. It is expected as it contains many of the most populated areas and probably where most of the transactions are taking place. A few outliers are seen in Alaska and Hawaii, indicating fraudulent transactions also happen in these regions, but with a much lower frequency. The study cannot conclude that a region is more prone to fraudulent transactions just because it sees many fraudulent transactions in the region where most of the transactions happen.

5.9. Total Fraud Counts by Class

is_Fraud	Class	Quantity	Percentage (%)
0	Non-Fraud	1289169	99.42
1	Fraud	7506	0.58

Table 9: Fraud Counts by Class

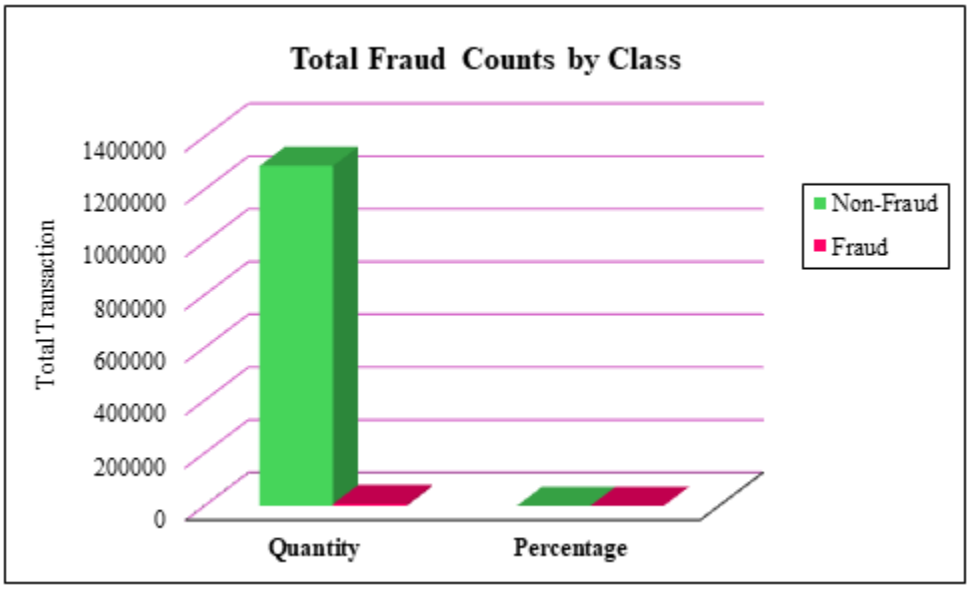


Figure 13: Fraud Counts by Class

The classes of non-fraud and fraud are highly imbalanced. The percentage of non-fraud and fraud transactions in the training set is 99.42% and 0.58%, respectively. Fraud transactions are rare when compared with normal transactions. Therefore, in the figure, due to the extreme imbalance of the classes, the fraud bar is almost invisible. The imbalance of the classes should be taken into consideration in designing a good model because a model can easily achieve high accuracy by labeling most transactions as non-fraud. Therefore, the research needs to evaluate the recall, precision, PR-AUC, and cost-performance of the model.

5.10. Fraud by Transaction Hour

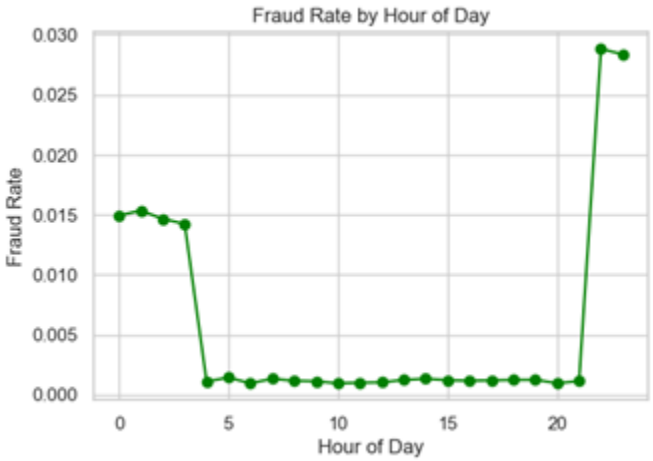


Figure 14: Fraud by Transaction Hour

The graph shows that the risk of fraud changes a lot during the day. When fraud happens the most, between 10 p.m. and 1 a.m., the rate goes up a lot. From 4 AM to 9 PM, fraud stays very low and almost flat. This pattern shows that fraud is more likely to happen when things don't go as planned. This could be because customers aren't as active or it takes them longer to notice things that are different. The time of the transaction is important for finding fraud because it helps the model find risky behavior, especially when there aren't many customers around, like in the early morning or late at night when fraud is more likely to happen.

5.11. Transaction Amount by Class "is_Fraud"

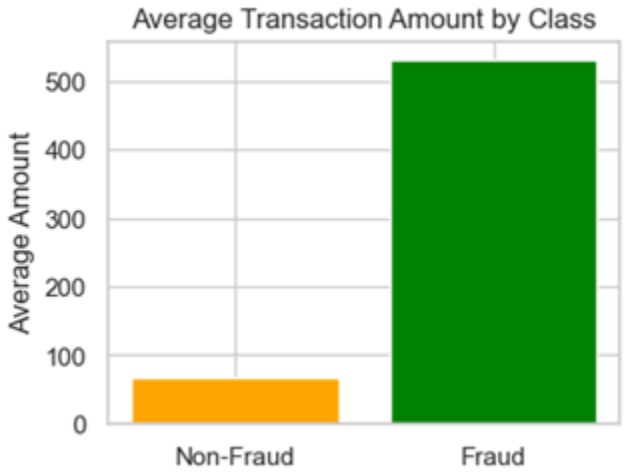


Figure 15: Transaction Amount by Class "is_Fraud"

The bar graph shows that transaction amount is higher for fraud transactions than for non- transactions. There is only $65 to $70 that is non-fraud, while fraud transactions have an average amount above $500. This means that most of the time, fraud transactions in this dataset are for big things, like buying something expensive or getting a service that is known to be a target for fraud. The big difference means that checking the amount of the transaction is a good way to find fraud. If the amount is higher than normal, the model might think the transaction is more suspicious.

6. Discussion

6.1 The Case for Cost-Sensitive Evaluation

The results indicate that banks should not consider the models to detect fraud solely based on accuracy. At the default decision threshold of 0.50, XGBoost model already has a high accuracy of 98.86% and a PR-AUC score of 0.8415, which might be satisfactory for many banks. But accuracy here fails to capture the economic impact of fraud detection. It is the customization of the decision threshold based on an economic cost that makes the models attractive to banks. The optimized XGBoost model provides a better balance between the costs of missing fraud cases and false alarm cases compared to the other models.

The total economic cost in the test data case for the optimized XGBoost model is $355,700. There are 555,719 transactions in the test data set, among which 2,145 are fraud. The optimized XGBoost model detects 1,956 fraud cases and misses 189 fraud cases. It also generates 3,334 false alarms so some legitimate transactions are falsely flagged as fraud.

This result may not suit all the banks. Each bank may have different optimum threshold as per the cost it associates with each error. A bank worrying about customer experience may be considering false alarms more costly because it doesn't want to irritate valid customers by blocking valid transactions. That type of bank would prefer a relatively higher threshold value to keep the false alarm rate marginally low, even if it means detection of fraud cases will be compromised to a certain extent. The threshold optimization methodology will be very useful to customize the balance between these two parameters for each bank depending upon its business objectives, likelihood of fraud and its inclination to serve its customers in the best possible way.

6.2 Explainability as a Business and Regulatory Requirement

XGBoost was selected as the best model because it achieved the lowest total economic cost, and SHAP analysis was used to explain its predictions. It gives more importance to the features which would provide the bank's risk management and compliance team with more information. The bank may be required to explain to the regulator how the model for detecting fraud works and which information it uses. SHAP facilitates the regulator's understanding of the estimates for each transaction.

If a customer calls the bank to ask why their transaction was denied, the bank worker can use SHAP to explain why their transaction was denied. The bank teller might say that the transaction was flagged because it was for a lot of money compared to usual transactions, it was late at night, and the store is far away from the customer's home. Such explanations may improve customer communication compared to standard fraud alert responses. Thus, SHAP demonstrates value beyond fraud detection, extending to customer service and regulatory communication. Additionally, the SHAP model is significant for the bank because increasing attention is being paid by financial regulators to banks' automated decision-making systems. Banks are required to

deploy AI systems under adequate supervision and explicability. Using the SHAP method, banks can maintain a record of why a fraud prediction is made for each transaction. This helps in understanding and auditing the models and ensures regulatory compliance.

6.3 Responsible AI and Fairness Gaps

As the model indicates a lower false negative rate for male customers than for female customers, the fairness analysis shows a potential gender-based performance inequality. The false negative rate is 12.46% for women and only 4.49% for men. Although gender was not the main predictive driver, the results suggest that the model missed fraudulent transactions for female customers at a higher rate than for male customers within this dataset. This type of unintentional disparate impact is a common issue in machine learning called proxy discrimination. Gender has a low SHAP importance, which means it doesn't have a direct effect on predictions for individuals. It could also be related to other things, like the type of store or the time of day. The model may have unintentionally picked up on these links. The ECOA and its implementing regulation (Regulation B) say that it is against the law to treat people differently in credit transactions because of their gender. Fraud detection isn't a credit decision in the strictest sense, but the spirit of these protections applies to any financial service that affects people.

Researchers refer to this as "fairness-aware machine learning," which means that the model training will penalize differences in FNR and FPR among various demographic groups. This is the best way to go. Some ways to do this are adversarial debiasing, changing the weights of training examples based on demographic group, or using different thresholds for different groups to change the model's scores after they are made. This is also called threshold calibration by groups. Each method has its own pros and cons when it comes to fairness and overall performance. The best choice will depend on what the bank thinks is right and what its rules say.

6.4 Practical Implication for U.S banks

For bank practitioners, this study offers several actionable takeaways. First, gradient boosting models particularly XGBoost and LightGBM, significantly outperform simpler models for fraud detection on large transaction datasets. Second, every institution should perform threshold optimization using its cost estimates before deploying a fraud detection model. Third, SHAP-based explainability infrastructure should be built to deliver significant value of regulatory compliance and customer service. Fourth, fairness monitoring should be a regular process because customer behavior and fraud patterns can change over time. Banks that recognize and address all four dimensions to maintain customer trust, avoid regulatory censure, and build financially resilient operations.

7. Conclusion

This study demonstrates that cost-sensitive machine learning can improve credit card fraud detection by aligning model evaluation with the economic impact of false positives and false negatives. XGBoost outperformed all other models in this study, yielding the minimum total economic cost of $355,700 at an optimal decision threshold of 0.75. The results show that banks should not rely on accuracy because false negatives and false positives create different economic consequences. SHAP analysis shows that the three most important things that affect fraud prediction are the amount of the transaction, the patterns of where it took place, and the time of the transaction. This improves model interpretability for banking decision-makers and risk managers. The fairness results show that men and women customers exhibit distinct transaction behavior patterns, which underscores the need for continuous fairness monitoring. Effective fraud detection should be cost-efficient, explainable, and fair.

Funding: This research received no external funding.
Conflicts of Interest: The author declares no conflict of interest.
Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] Aite-Novarica Group. (2022). *False declines and the cost of friction in the U.S. payments market*. Aite-Novarica Group.
[2] Bahnsen, A. C., Aouada, D., Stojanovic, A., & Ottersten, B. (2016). Feature engineering strategies for credit card fraud detection. *Expert Systems with Applications, 51*, 134–142. https://doi.org/10.1016/j.eswa.2015.12.030
[3] Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. *Decision Support Systems, 50*(3), 602–613. https://doi.org/10.1016/j.dss.2010.08.008
[4] Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. *Statistical Science, 17*(3), 235–255. https://doi.org/10.1214/ss/1042727940
[5] Chan, P. K., Fan, W., Prodromidis, A. L., & Stolfo, S. J. (1999). Distributed data mining in credit card fraud detection. *IEEE Intelligent Systems, 14*(6), 67–74. https://doi.org/10.1109/5254.809570
[6] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785

[7] Consumer Financial Protection Bureau. (2022). *Consumer Financial Protection Circular 2022-03: Adverse action notification requirements and the Equal Credit Opportunity Act*. https://www.consumerfinance.gov/

[8] Dal-Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. (2018). Credit card fraud detection: A realistic modeling and a novel learning strategy. *IEEE Transactions on Neural Networks and Learning Systems, 29*(8), 3784–3797. https://doi.org/10.1109/TNNLS.2017.2736643

[9] Federal Trade Commission. (2024). *Consumer Sentinel Network data book 2023*. https://www.ftc.gov/reports/consumer-sentinel-network

[10] Fiore, U., De Santis, A., Perla, F., Zanetti, P., & Palmieri, F. (2019). Using generative adversarial networks for improving classification effectiveness in credit card fraud detection. *Information Sciences, 479*, 448–455. https://doi.org/10.1016/j.ins.2017.12.030

[11] Harris, B. (2020). *Credit card transactions fraud detection dataset* [Data set]. Kaggle. https://www.kaggle.com/datasets/kartik2112/fraud-detection

[12] High-Level Expert Group on Artificial Intelligence. (2019). *Ethics guidelines for trustworthy AI*. European Commission. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[13] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems, 30*.

[14] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems, 30*.

[15] Office of the Comptroller of the Currency. (2011). *Supervisory guidance on model risk management*. Board of Governors of the Federal Reserve System and Office of the Comptroller of the Currency.

[16] Phillips, P. J., Hahn, C. A., Fontana, P. C., Yates, A. N., Greene, K., Broniatowski, D. A., & Przybocki, M. A. (2021). *Four principles of explainable artificial intelligence* (NIST Interagency Report 8312). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.IR.8312

[17] Power J.D. (2023). *U.S. retail banking satisfaction study*. J.D. Power.

[18] Provost, F., & Fawcett, T. (2013). *Data science for business*. O'Reilly Media.

[19] The Nilson Report. (2023). *Card fraud losses worldwide* (Issue 1232). HSN Consultants.

[20] Verhelst, T., Vanderschueren, T., Verbeke, W., & Baesens, B. (2022). Profit-oriented fraud detection: A prescriptive analytics perspective. *Decision Support Systems, 161*, Article 113831. https://doi.org/10.1016/j.dss.2022.113831

[21] Zhou, Y., Li, H., Xiao, Z., & Qiu, J. (2023). A user-centered explainable artificial intelligence approach for financial fraud detection. *Finance Research Letters, 58*(PA), Article 104309. https://doi.org/10.1016/j.frl.2023.104309