NORTON UNIVERSITY
GRADUATE SCHOOL

KINGDOM OF CAMBODIA
NATION RELIGION KING

Machine Learning Course

Midterm Group Assignment

Academic Year: 2026

Project Title

Credit Card Fraud Detection

Submitted to

MR. SEK SOCHEAT

Machine Learning Lecturer

Submitted by

Group 7

Student Name	Student ID	Role	Responsibilities
Mr. Chhun Menghong	G20240115	Project Leader	Coordinate project, manage timeline, ensure integration of all components
Mr. Pheuy Tay	G20250010	Data Engineer	Dataset preprocessing, cleaning, feature engineering, exploratory data analysis
Mr. San Ratana	G20240109	Machine Learning Engineer	Model development, training, hyperparameter tuning, evaluation
Mr. Tang Kimhab	G20240085	System Developer & Documentation	Build Streamlit system, testing, documentation, deployment

Table Contents

1. Background and Problem Statement

1.1 Background

The rapid advancement of information technology has transformed the global financial industry, leading to significant growth in electronic payment systems. Credit cards have become one of the most widely used payment methods because they provide convenience, speed, and flexibility for both consumers and businesses. The increasing popularity of online shopping, mobile banking, and digital payment platforms has resulted in millions of financial transactions being processed every day. As digital payment adoption continues to expand, financial institutions are required to process large volumes of transaction data accurately and securely.

Despite these advantages, the widespread use of credit cards has also created new opportunities for fraudulent activities. Cybercriminals continuously develop sophisticated techniques to obtain unauthorized access to credit card information through methods such as phishing attacks, data breaches, card skimming, identity theft, and malware. These fraudulent activities cause significant financial losses to banks, merchants, and customers while reducing public confidence in electronic payment systems.

According to industry reports, global payment fraud continues to increase each year as the number of online transactions grows. Financial institutions invest substantial resources in fraud prevention technologies to reduce financial losses and maintain customer trust. Consequently, developing accurate and efficient fraud detection systems has become a critical challenge within the banking and financial services sector.

Traditional fraud detection systems primarily rely on manually defined business rules. For example, a transaction may be flagged if the amount exceeds a predetermined limit or if it occurs in an unusual geographic location. While rule-based systems are effective in detecting previously known fraud patterns, they have several limitations. They require continuous manual updates, struggle to identify newly emerging fraud strategies, and often produce a high number of false alarms. As fraud techniques become increasingly complex, traditional methods are no longer sufficient to provide effective protection against evolving cyber threats.

Machine learning has emerged as one of the most promising technologies for fraud detection because it enables computer systems to learn patterns directly from historical transaction data. Unlike rule-based approaches, machine learning algorithms can automatically recognize complex relationships among multiple transaction characteristics and classify new transactions as legitimate or fraudulent. These algorithms are capable of adapting to changing fraud patterns and improving prediction performance as additional data become available.

Supervised machine learning techniques have been successfully applied in various financial applications, including fraud detection, credit risk assessment, loan approval, and customer behavior analysis. Classification algorithms such as Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and Extreme Gradient Boosting (XGBoost) have demonstrated strong predictive capabilities for identifying fraudulent financial transactions. However, the performance of these algorithms varies depending on the characteristics of the dataset, feature representation, and preprocessing techniques used.

Therefore, selecting the most suitable machine learning algorithm requires systematic experimentation and objective performance evaluation using multiple evaluation metrics.

1.2 Research Gap

Numerous studies have investigated the application of machine learning techniques for detecting fraudulent credit card transactions. Most previous research has demonstrated that machine learning algorithms outperform traditional rule-based systems by learning complex transaction patterns from historical data. However, several limitations remain.

First, many existing studies focus on evaluating only one or two classification algorithms, making it difficult to determine which algorithm consistently performs best under different transaction characteristics. Second, some studies emphasize overall prediction accuracy while overlooking other important evaluation metrics such as precision, recall, F1-score, and ROC-AUC, which are particularly important for fraud detection due to the potential imbalance between legitimate and fraudulent transactions. Third, previous studies often provide limited discussion regarding the impact of data preprocessing techniques, including feature engineering, categorical encoding, and data preparation, on model performance.

These limitations indicate that further research is needed to conduct a comprehensive comparison of multiple machine learning classification algorithms using a consistent preprocessing pipeline and standardized evaluation metrics. Such a comparison will help identify the most appropriate algorithm for detecting fraudulent credit card transactions and provide practical guidance for future fraud detection system development.

1.3 Problem Statement

Credit card fraud continues to pose significant financial and operational challenges for financial institutions despite the implementation of traditional fraud detection mechanisms. Rule-based detection systems are increasingly unable to adapt to evolving fraud techniques and often generate inaccurate predictions due to their dependence on predefined rules. As transaction volumes continue to increase, manual investigation becomes impractical and costly.

To address these challenges, there is a need to develop an intelligent fraud detection system capable of automatically learning transaction patterns and accurately distinguishing between legitimate and fraudulent transactions. Machine learning offers a promising solution because it can analyze historical transaction data, discover hidden relationships among transaction characteristics, and improve prediction performance without relying solely on manually defined rules.

Therefore, this research proposes the development of a machine learning-based credit card fraud detection system that implements and compares multiple supervised classification algorithms. The study aims to identify the most effective algorithm for accurately detecting fraudulent credit card transactions while providing a practical prototype that demonstrates the application of machine learning techniques in financial fraud detection.

2. Objectives of the Project

2.1 General Objective

To develop a machine learning-based system for detecting fraudulent credit card transactions.

2.2 Specific Objectives

- **To collect and analyze credit card transaction data**

  This objective focuses on obtaining and examining the credit card transaction dataset to understand its structure, data types, feature distributions, and transaction

characteristics. Exploratory Data Analysis (EDA) will be performed to identify patterns, detect anomalies, examine data quality, and gain insights that support the development of machine learning models.

- **To pre-process and prepare data for machine learning**

  The collected dataset will undergo a comprehensive preprocessing process before model development. This includes handling missing values (if any), removing duplicate records, transforming the transaction date into useful time-related features, encoding categorical variables, scaling numerical features where appropriate, and splitting the dataset into training and testing subsets. Proper preprocessing is essential to improve data quality and enhance the performance of machine learning algorithms.

- **To implement and compare multiple classification algorithms**

  Several supervised machine learning algorithms will be implemented, including Logistic Regression, Decision Tree, Random Forest, and XGBoost. Each algorithm will be trained using the same preprocessed dataset and evaluated using consistent performance metrics. Comparative analysis will then be conducted to identify the algorithm that provides the most effective performance for detecting fraudulent credit card transactions.

3. Research Questions and Expected Outcomes

3.1 Research Questions

The primary purpose of this research is to investigate how machine learning techniques can improve the detection of fraudulent credit card transactions. Based on the research objectives, the study seeks to answer the following research questions:

Research Question 1: How can credit card transaction data be effectively prepared for machine learning-based fraud detection?

Machine learning models require high-quality and well-preprocessed data to achieve reliable prediction performance. This research investigates appropriate preprocessing techniques, including data cleaning, feature engineering, categorical encoding, and data transformation, to ensure that the transaction dataset is suitable for machine learning.

Research Question 2: Which machine learning classification algorithm provides the best performance for detecting fraudulent credit card transactions?

Different machine learning algorithms have different strengths and weaknesses. This research compares several supervised classification algorithms to determine which model achieves the highest performance in detecting fraudulent transactions.

Research Question 3: How effective are machine learning classification algorithms in identifying fraudulent credit card transactions?

This research evaluates the effectiveness of machine learning models using multiple performance metrics, including Accuracy, Precision, Recall, F1-Score, and Confusion Matrix analysis. The evaluation aims to determine whether machine learning can provide accurate and reliable fraud detection.

3.2 Expected Outcomes

The successful completion of this project is expected to produce the following outcomes:

- **Comprehensive Understanding of Transaction Data**

  The research will provide a detailed analysis of the credit card transaction dataset, including data characteristics, feature distributions, transaction patterns, and potential issues affecting model performance.

- **High-Quality Preprocessed Dataset**

  A fully preprocessed dataset will be produced through data cleaning, feature engineering, categorical encoding, and data transformation. This dataset will be suitable for training and evaluating machine learning models.

- **Multiple Machine Learning Models**

  Several supervised machine learning classification models will be developed and trained using the prepared dataset.

  These include:

  - Logistic Regression
  - Decision Tree
  - Random Forest
  - Support Vector Machine (SVM)
  - XGBoost

- **Comparative Performance Analysis**

  The performance of all implemented algorithms will be compared using standardized evaluation metrics. The comparison will identify the strengths and limitations of each algorithm and recommend the most suitable model for credit card fraud detection.

- **Fraud Detection Prototype**

  A web-based prototype application will be developed using Streamlit. The prototype will allow users to enter transaction information and predict whether a transaction is legitimate or fraudulent using the trained machine learning model.

- **Academic Contribution**

  The research findings are expected to contribute to the field of financial fraud detection by providing a comparative analysis of multiple classification algorithms using a structured transaction dataset. The study may also serve as a reference for future research involving machine learning applications in cybersecurity and financial analytics.

4. Target Users or Beneficiaries

The proposed credit card fraud detection system is intended to benefit several stakeholders involved in financial transactions and cybersecurity.

4.1 Banks and Financial Institutions

Banks process millions of credit card transactions every day and face significant financial losses due to fraudulent activities. The proposed machine learning system can assist banks by automatically identifying suspicious transactions before they are approved. This reduces financial losses, improves operational efficiency, and enhances customer confidence in digital banking services.

4.2 Credit Card Issuers

Credit card companies continuously monitor customer transactions to prevent unauthorized card usage. By integrating machine learning into fraud detection systems, credit card issuers can improve transaction monitoring, reduce fraudulent claims, and strengthen payment security.

4.3 Payment Service Providers

Electronic payment providers require intelligent fraud detection mechanisms to ensure secure financial transactions. The proposed system can support payment gateways by analyzing transaction behavior and identifying potentially fraudulent activities in real time.

4.4 Online Merchants and E-Commerce Businesses

Online merchants are vulnerable to fraudulent purchases involving stolen credit card information. The proposed system can help identify suspicious transactions before payment authorization, reducing chargeback losses and protecting business revenue.

4.5 Fraud Analysts and Cybersecurity Professionals

Fraud investigators and cybersecurity analysts can use the prediction results generated by the system to prioritize suspicious transactions for further investigation. This improves investigation efficiency and reduces manual workload.

4.6 Researchers and Students

The project also serves as a valuable educational resource for researchers and students studying machine learning, artificial intelligence, cybersecurity, financial technology (FinTech), and data science. The implementation demonstrates a practical application of supervised machine learning techniques within the financial sector.

5. Dataset Source or Data Collection Plan

5.1 Dataset Source

This research uses a structured **Credit Card Fraud Dataset** containing approximately **100,000 transaction records**. The dataset was provided specifically for academic research and machine learning experimentation. It consists of historical transaction records that include both legitimate and fraudulent transactions, making it suitable for binary classification.

Each transaction contains several attributes describing transaction characteristics together with a fraud label indicating whether the transaction is fraudulent.

5.2 Dataset Summary

Item	Description
Dataset Name	Credit Card Fraud Dataset
Number of Records	Approximately 100,000
Number of Features	6 Predictor Variables
Target Variable	IsFraud
Machine Learning Task	Binary Classification

5.3 Dataset Attributes

Attribute	Data Type	Description
TransactionID	Integer	Unique identifier assigned to each transaction. This attribute is used only for identification and will not be included in model training.
TransactionDate	Date/Time	Records the date and time when each transaction occurred. This attribute will be transformed into useful temporal features.
Amount	Numeric	Monetary value of the transaction. It represents one of the most important indicators for fraud detection.
MerchantID	Categorical	Unique identifier representing the merchant where the transaction occurred.
TransactionType	Categorical	Types of transaction such as payment, purchase, withdrawal, or transfer.
Location	Categorical	Geographic location where the transaction occurred.
IsFraud	Binary	Target variable indicating whether the transaction is legitimate (0) or fraudulent (1).

5.4 Data Collection Plan

Although the dataset has already been collected, several preparation activities will be performed before machine learning model development.

The data preparation process includes:

- Importing the dataset into Python using the Pandas library.
- Inspecting the dataset structure and verifying data types.
- Identifying duplicate records.
- Detecting missing values and handling them appropriately.
- Performing exploratory data analysis (EDA).
- Transforming date and time information into meaningful features.
- Encoding categorical variables into numerical representations.
- Scaling numerical variables when required.
- Splitting the dataset into training and testing subsets using an 80:20 ratio.

These preprocessing activities are essential for improving model performance and ensuring reliable prediction results.

5.5 Data Preprocessing Workflow

The proposed preprocessing workflow consists of the following stages:

1. Dataset Import
2. Data Inspection
3. Data Cleaning
4. Duplicate Removal
5. Missing Value Handling
6. Feature Engineering
7. Categorical Encoding
8. Feature Scaling
9. Train-Test Split
10. Machine Learning Model Development

This workflow provides a systematic approach to preparing high-quality data for supervised machine learning.

6. Features and Target Variable

Machine learning algorithms require carefully selected input variables that provide meaningful information for prediction. The dataset contains six predictor variables and one target variable.

6.1 Predictor Variables (Features)

TransactionDate

The transaction date records when each transaction occurred. Fraudulent activities often occur during unusual hours, weekends, or holidays. Therefore, the TransactionDate attribute will be transformed into additional temporal features such as:

- Hour
- Day
- Month
- Day of Week

These derived variables enable machine learning models to recognize temporal fraud patterns more effectively.

Amount

The transaction amount represents the monetary value of each transaction. Fraudulent transactions may involve unusually large or unusually small amounts depending on the fraud strategy. This feature provides valuable numerical information for distinguishing normal and suspicious transactions.

Before model training, the Amount attribute may be standardized or normalized to improve algorithm performance.

MerchantID

MerchantID identifies the business where the transaction occurred. Some merchants may experience higher fraud rates than others due to business type or transaction characteristics. Since this variable is categorical, it will be encoded into numerical values before model training.

TransactionType

Different transaction types exhibit different fraud risks. For example, online purchases may have different fraud characteristics compared with ATM withdrawals or fund transfers.

This categorical feature will be encoded using an appropriate encoding technique before machine learning implementation.

Location

The transaction location represents where the transaction occurred. Geographic information is useful because fraudulent transactions frequently occur in unfamiliar or unexpected locations.

Location will also be encoded into numerical values before model training.

6.2 Excluded Feature

TransactionID

Although every transaction has a unique TransactionID, this attribute serves only as an identifier and does not contain predictive information. Therefore, it will be excluded from machine learning model training to avoid introducing unnecessary noise into the dataset.

6.3 Target Variable

The dependent variable in this research is IsFraud, which indicates whether a transaction is fraudulent.

Value	Description
0	Legitimate Transaction
1	Fraudulent Transaction

This makes the project a binary classification problem, where the machine learning model learns to classify each transaction into one of two possible classes.

6.4 Feature Engineering Strategy

To improve prediction performance, several feature engineering techniques will be applied:

- Extract temporal information from **TransactionDate**.
- Encode categorical variables (**MerchantID**, **TransactionType**, and **Location**).
- Normalize or standardize the **Amount** feature when appropriate.
- Remove non-informative variables such as **TransactionID**.
- Evaluate feature importance using tree-based models to understand the contribution of each predictor.

These steps help the machine learning algorithms learn more meaningful patterns from the transaction data and improve fraud detection performance.

7. Proposed Machine Learning Algorithms

7.1 Introduction

Machine Learning (ML) has become one of the most effective technologies for detecting fraudulent financial transactions because it enables computers to automatically learn hidden patterns from historical data and make predictions on new transactions. Unlike traditional rule-based systems, which rely on manually defined conditions, machine learning algorithms continuously improve their predictive capability by learning from transaction characteristics.

Since the objective of this research is to determine the most suitable algorithm for detecting fraudulent credit card transactions, several supervised classification algorithms will be implemented and compared using the same dataset and evaluation criteria. The comparative analysis ensures that the selected model is based on objective performance rather than assumptions.

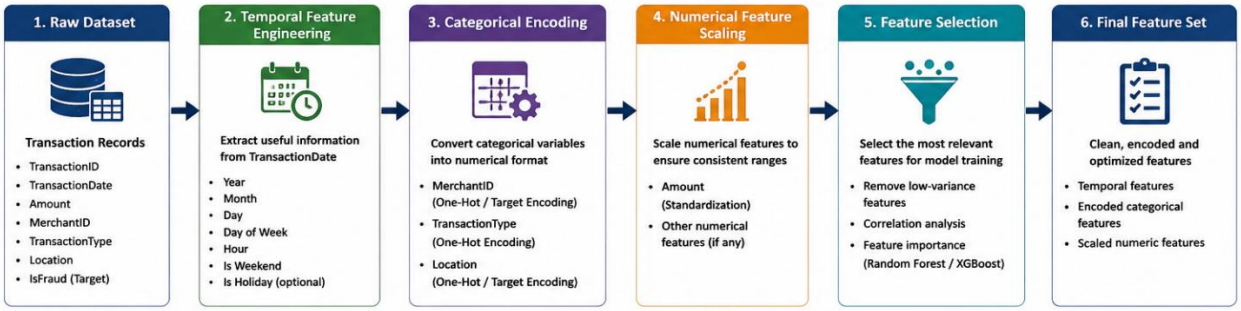
The proposed algorithms include:

- Logistic Regression (LR)
- Decision Tree (DT)
- Random Forest (RF)
- Support Vector Machine (SVM)
- Extreme Gradient Boosting (XGBoost)

Each algorithm has different characteristics, advantages, and limitations. Therefore, implementing multiple models allows this research to identify the most effective approach for fraud detection.

7.2 Research Methodology

The overall machine learning methodology adopted in this research consists of six major phases.



The selected model will be integrated into a Streamlit web application for fraud prediction.

7.3 Logistic Regression

Introduction

Logistic Regression is a statistical classification algorithm widely used for binary classification problems. Since fraud detection requires classifying transactions into legitimate (0) and fraudulent (1), Logistic Regression provides an appropriate baseline model.

Unlike linear regression, Logistic Regression estimates the probability that a transaction belongs to the fraud class using the Sigmoid (Logistic) function.

Working Principle

The algorithm calculates a weighted combination of input variables and transforms the result into a probability between 0 and 1.

If the predicted probability exceeds a predefined threshold (typically 0.5), the transaction is classified as fraudulent.

Advantages

- Simple implementation
- Fast training
- Easy interpretation
- Produces probability scores
- Effective baseline classifier

Limitations

- Assumes linear relationships
- Limited capability for complex nonlinear data
- Sensitive to multicollinearity

Reason for Selection

Logistic Regression serves as the benchmark model against which more sophisticated algorithms can be compared.

7.4 Decision Tree

Introduction

Decision Tree is a supervised classification algorithm that predicts outcomes by recursively splitting data according to decision rules. The algorithm constructs a tree structure in which internal nodes represent feature-based decisions and leaf nodes represent prediction classes.

Working Principle

The algorithm selects the feature that best separates legitimate and fraudulent transactions using impurity measures such as:

- Gini Index
- Information Gain (Entropy)

The process continues until stopping criteria are met.

Advantages

- Easy to understand
- Easy to visualize
- Handles nonlinear relationships
- Handles numerical and categorical data
- Minimal preprocessing

Limitations

- Easily overfits
- Sensitive to noisy data
- Can produce unstable trees

Reason for Selection

Decision Trees provide interpretable classification models and help identify important decision rules contributing to fraud detection.

7.5 Random Forest

Introduction

Random Forest is an ensemble learning algorithm that combines multiple Decision Trees to improve prediction accuracy and reduce overfitting.

Instead of relying on a single tree, Random Forest generates many trees using random subsets of both data and features.

Working Principle

Each tree independently predicts whether a transaction is fraudulent.

The final prediction is determined through majority voting.

This approach reduces model variance and improves generalization.

Advantages

- High prediction accuracy
- Resistant to overfitting
- Handles large datasets
- Measures feature importance
- Handles nonlinear relationships

Limitations

- Larger model size
- Longer training time
- Less interpretable

Reason for Selection

Random Forest is one of the most widely used fraud detection algorithms because it consistently performs well on structured financial datasets.**7.6 Support Vector Machine (SVM)**

Introduction

Support Vector Machine is a supervised learning algorithm designed to identify the optimal boundary separating different classes.

The objective is to maximize the margin between fraudulent and legitimate transactions.

Working Principle

SVM identifies support vectors that define the decision boundary.

Kernel functions may be used when the data are not linearly separable.

Common kernels include:

- Linear
- Polynomial
- Radial Basis Function (RBF)

Advantages

- Excellent binary classifier
- Strong generalization
- Effective in high-dimensional spaces
- Robust against overfitting

Limitations

- Computationally expensive
- Difficult parameter tuning
- Longer training time

Reason for Selection

Since fraud detection is fundamentally a binary classification problem, SVM provides an important comparison against tree-based methods.

7.7 Extreme Gradient Boosting (XGBoost)

Introduction

Extreme Gradient Boosting (XGBoost) is an advanced ensemble learning algorithm based on gradient boosting.

Instead of building independent trees, XGBoost sequentially builds new trees that correct the prediction errors of previous trees.

Working Principle

Each new decision tree focuses on transactions that previous trees classified incorrectly.
The model gradually minimizes prediction errors using gradient optimization.

Advantages

- Excellent predictive performance
- Handles nonlinear relationships
- Built-in regularization
- Fast computation
- Handles missing values
- Prevents overfitting

Limitations

- More parameters to tune
- Higher computational cost
- More complex implementation

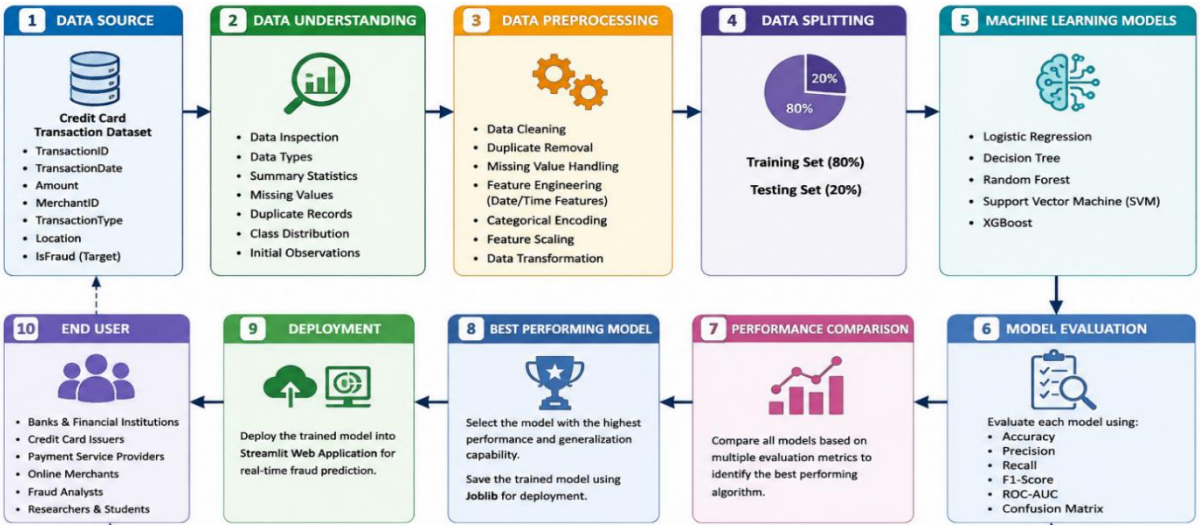
Reason for Selection

XGBoost has demonstrated outstanding performance in numerous machine learning competitions and financial fraud detection studies, making it an excellent candidate for this research.

7.8 Comparison of Proposed Algorithms

Algorithm	Strengths	Weaknesses
Logistic Regression	Fast, interpretable, simple	Limited nonlinear capability
Decision Tree	Easy to explain and visualize	Overfitting risk
Random Forest	High accuracy and robust	Higher computational cost
SVM	Strong binary classifier	Slow on large datasets
XGBoost	Excellent predictive performance	Requires parameter tuning

7.9 Proposed Research Framework



8. Proposed Evaluation Metrics

Evaluating fraud detection models requires more than measuring overall accuracy. Because fraudulent transactions may represent only a small proportion of all transactions, relying solely on accuracy can produce misleading conclusions. Therefore, this research will use multiple evaluation metrics to provide a comprehensive assessment of model performance.

8.1 Accuracy

Accuracy measures the proportion of correctly classified transactions among all transactions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Although accuracy provides an overall performance indicator, it may not adequately reflect a model's ability to detect rare fraudulent transactions.

8.2 Precision

Precision measures the proportion of transactions predicted as fraudulent that are actually fraudulent.

$$Precision = \frac{TP}{TP + FP}$$

A high precision value reduces false alarms and minimizes unnecessary investigations of legitimate transactions.

8.3 Recall

Recall measures the proportion of actual fraudulent transactions that are correctly identified.

$$Recall = \frac{TP}{TP + FN}$$

Recall is particularly important in fraud detection because failing to identify fraudulent transactions may result in financial losses.

8.4 F1-Score

F1-Score combines Precision and Recall into a single metric.

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

It provides a balanced evaluation, especially when dealing with imbalanced datasets.

8.5 ROC-AUC

The Receiver Operating Characteristic (ROC) curve illustrates the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR).

The Area Under the Curve (AUC) summarizes the model's ability to distinguish between legitimate and fraudulent transactions. A higher ROC-AUC value indicates better discrimination capability.

8.6 Confusion Matrix

The Confusion Matrix provides a detailed summary of prediction outcomes.

Actual / Predicted	Legitimate	Fraudulent
Legitimate	True Negative (TN)	False Positive (FP)
Fraudulent	False Negative (FN)	True Positive (TP)

The Confusion Matrix helps identify the types of prediction errors made by each algorithm and supports a more detailed performance analysis.

8.7 Evaluation Metric Summary

Metric	Purpose
Accuracy	Measures overall prediction correctness
Precision	Evaluates false positive reduction
Recall	Measures fraud detection capability
F1-Score	Balances Precision and Recall
ROC-AUC	Measures class discrimination ability
Confusion Matrix	Provides detailed classification analysis

The combination of these metrics ensures that the selected model is evaluated comprehensively rather than relying on a single performance indicator.

9. Programming Language, Framework, and Tools

The project will be implemented using Python and its associated machine learning ecosystem because of its flexibility, extensive libraries, and strong support for data science applications.

Tool / Framework	Purpose	Justification
Python	Core programming language	Widely used in machine learning, easy to learn, extensive library support.
Pandas	Data manipulation and preprocessing	Efficient handling of structured datasets and missing values.
NumPy	Numerical computation	Fast mathematical operations and array processing.
Scikit-learn	Machine learning algorithms	Provides reliable implementations of Logistic Regression, Decision Tree, Random Forest, SVM, preprocessing tools, and evaluation metrics.
XGBoost	Gradient boosting implementation	High-performance ensemble learning library widely used for structured datasets.
Matplotlib	Data visualization	Creates charts, confusion matrices, ROC curves, and feature importance plots.
Pickle	Model serialization and persistence	Saves and loads trained machine learning models efficiently.
Jupyter Notebook	Experimentation and documentation	Supports interactive development, exploratory data analysis, and result presentation.
Visual Studio Code	Integrated Development Environment (IDE)	Provides a productive environment for coding, debugging, and project management.
Streamlit	Web application framework	Enables rapid development of an interactive fraud detection system for end users.

9.1 Development Environment

The proposed development environment includes:

- **Operating System:** Windows 11 (or Linux)
- **Programming Language:** Python 3.12 or later

- **IDE:** Visual Studio Code
- **Notebook Environment:** Jupyter Notebook
- **Machine Learning Libraries:** Scikit-learn, XGBoost
- **Deployment:** Streamlit

This integrated environment supports the complete machine learning lifecycle, from data preprocessing and model development to evaluation, deployment, and maintenance.

10. Expected Outputs of the System

The successful completion of this research is expected to produce several academic, technical, and practical outputs that demonstrate the effectiveness of machine learning techniques in detecting fraudulent credit card transactions.

10.1 Academic Outputs

The project will produce several academic deliverables required by the Machine Learning course and future research activities.

These include:

- Project Proposal
- Final Research Report
- Presentation Slides
- Literature Review
- Project Documentation
- User Manual

These documents will comprehensively describe the research methodology, experimental results, discussion, and conclusions.

10.2 Technical Outputs

The research will develop a complete machine learning solution consisting of the following components.

Machine Learning Models

Five supervised learning models will be implemented and evaluated:

- Logistic Regression
- Decision Tree
- Random Forest
- Support Vector Machine
- XGBoost

Each model will be trained using the same dataset and evaluated using identical evaluation metrics.

Best Performing Model

Following comparative analysis, the model demonstrating the highest predictive performance will be selected as the final fraud detection model.

The trained model will be exported and saved using Joblib for future deployment.

Fraud Detection Prototype

A web-based prototype application will be developed using **Streamlit**.

The application will enable users to:

- Upload transaction CSV files.
- Perform fraud prediction.
- View fraud analysis results.
- Download a fraud-only transaction report

The prototype demonstrates the practical application of machine learning within financial fraud detection.

Visualization Dashboard

Several visualizations will be produced to support data analysis and model interpretation.

These include:

- The system provides fraud-versus-normal analysis
- fraud probability distribution
- fraud transactions by location
- fraud-only transaction report

These visualizations assist researchers in understanding dataset characteristics and evaluating model performance.

10.3 Research Outputs

The project is expected to produce several significant research findings.
These include:

- Identification of the most suitable machine learning algorithm.
- Performance comparison of multiple classification algorithms.
- Analysis of important transaction features influencing fraud detection.
- Recommendations for implementing machine learning in financial fraud detection systems.

These findings may contribute to future academic research and practical financial security applications.

11. Work Plan and Timeline

The project is planned to be completed within **12 weeks**. The work plan follows the standard machine learning development lifecycle, ensuring systematic progress from proposal preparation to final system implementation.

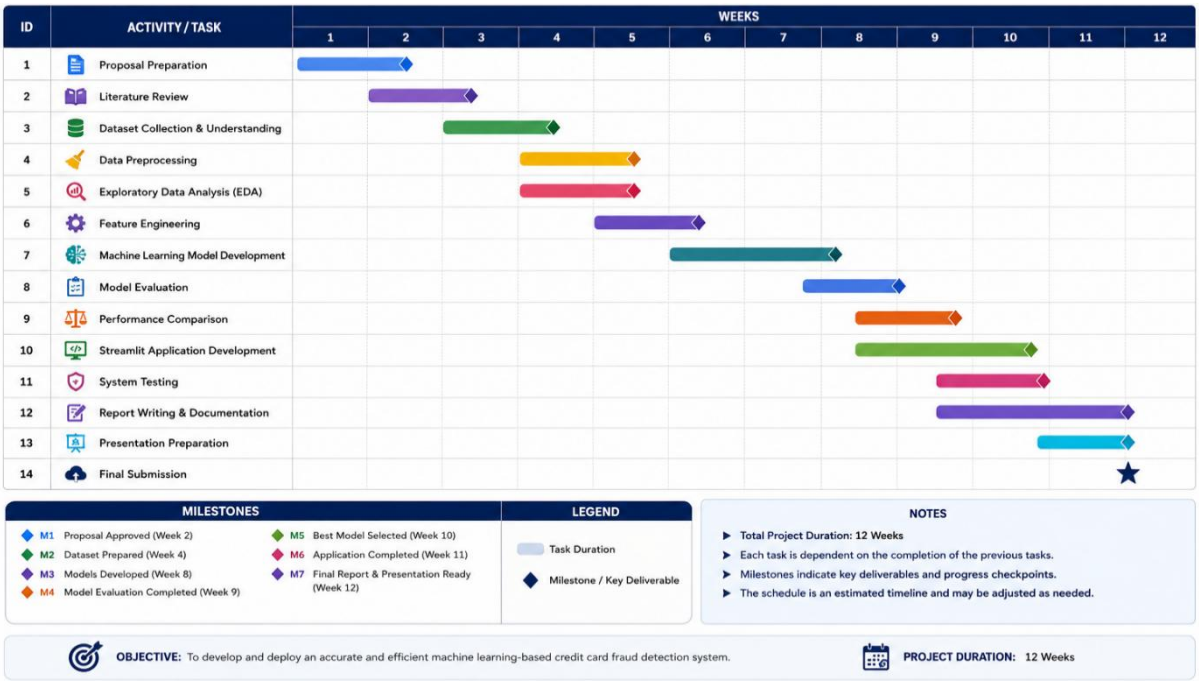
11.1 Project Activities

The major project activities include:

1. Proposal Preparation
2. Literature Review
3. Dataset Collection and Understanding
4. Data Preprocessing
5. Exploratory Data Analysis

6. Feature Engineering
7. Machine Learning Model Development
8. Model Evaluation
9. Performance Comparison
10. Streamlit Application Development
11. System Testing
12. Report Writing
13. Presentation Preparation
14. Final Submission

11.2 Project Timeline (Gantt Chart)



11.3 Project Milestones

Milestone	Expected Outcome
Milestone 1	Proposal Approved
Milestone 2	Dataset Successfully Prepared
Milestone 3	Machine Learning Models Developed
Milestone 4	Model Evaluation Completed
Milestone 5	Best Model Selected
Milestone 6	Streamlit Application Completed
Milestone 7	Final Report Submitted

The milestone-based approach ensures effective project monitoring and timely completion.

12. Possible Limitations or Risks

Although machine learning has demonstrated significant success in fraud detection, several limitations and risks may affect the performance and applicability of the proposed system.

12.1 Class Imbalance

Fraudulent transactions often represent only a small proportion of all transactions. This imbalance may cause machine learning models to favor legitimate transactions, leading to poor fraud detection performance.

Mitigation Strategy:

- Analyze the class distribution before training.
- If necessary, apply resampling techniques (e.g., SMOTE) or use class weighting.
- Evaluate models using Precision, Recall, and F1-Score instead of relying solely on Accuracy.

12.2 Limited Dataset Attributes

The dataset used in this research contains a limited number of transaction attributes. Real-world financial institutions may use additional information such as customer profiles, device identifiers, IP addresses, browser fingerprints, and historical behavioral patterns.

Mitigation Strategy:

- Clearly acknowledge this limitation.
- Recommend incorporating richer datasets in future research.

12.3 Model Overfitting

Some algorithms, particularly tree-based methods, may fit the training data too closely and perform poorly on unseen transactions.

Mitigation Strategy:

- Use train-test splitting.
- Apply k-fold cross-validation.
- Tune hyperparameters systematically.
- Compare training and testing performance.

12.4 Computational Requirements

Ensemble algorithms such as Random Forest and XGBoost require greater computational resources than simpler models.

Mitigation Strategy:

- Optimize hyperparameters.
- Use efficient implementations.
- Select appropriate hardware for model training.

12.5 Data Privacy and Security

Credit card transaction data are highly sensitive and must be protected throughout the research process.

Mitigation Strategy:

- Use anonymized datasets.
- Store data securely.
- Ensure compliance with institutional ethical guidelines.

- Avoid exposing personally identifiable information (PII).

12.6 Generalization to Real-World Environments

A model trained on one dataset may not achieve the same performance on data collected from different financial institutions due to variations in customer behavior and fraud patterns.

Mitigation Strategy:

- Evaluate the model on unseen test data.
- Recommend periodic retraining using updated transaction data.
- Suggest external validation using additional datasets where available.

12.7 Ethical Considerations

Incorrect fraud predictions can affect legitimate customers by delaying or blocking genuine transactions. Therefore, the model should be viewed as a decision-support tool rather than a replacement for human judgment.

Mitigation Strategy:

- Use the system to assist fraud analysts.
- Include manual review for high-risk transactions.
- Continuously monitor model performance after deployment.

13 Conclusion

The increasing volume of digital financial transactions has made credit card fraud detection an important challenge for banks, payment service providers, and online merchants. Traditional rule-based fraud detection methods are becoming less effective because they require continuous manual updates and cannot easily adapt to emerging fraud patterns.

This research proposes the development of a **machine learning-based credit card fraud detection system** capable of automatically learning transaction patterns from historical data and identifying fraudulent activities with greater accuracy. By implementing and comparing **Logistic Regression, Decision Tree, Random Forest, and XGBoost**, the study aims to determine the most suitable classification algorithm for the provided transaction dataset.

The research also emphasizes systematic data preprocessing, feature engineering, and comprehensive model evaluation using multiple performance metrics. In addition to producing a trained fraud detection model, the project will develop a **Streamlit-based prototype application** to demonstrate the practical implementation of the selected model.

The outcomes of this study are expected to benefit financial institutions, payment service providers, researchers, and students by providing a practical and comparative evaluation of machine learning techniques for fraud detection. Furthermore, the findings will contribute to the growing body of knowledge on intelligent financial security systems and provide recommendations for future improvements in machine learning-based fraud detection.

14 References

Books

1. Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
2. Géron, A. (2022). *Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow* (3rd ed.). O'Reilly Media.
3. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.
4. Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann.
5. Provost, F., & Fawcett, T. (2013). *Data Science for Business*. O'Reilly Media.
6. Kuhn, M., & Johnson, K. (2019). *Feature Engineering and Selection*. CRC Press.
7. Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5–32.
8. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.
9. Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning, 20*(3), 273–297.
10. Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

សាកលវិទ្យាល័យនរតុន
NORTON UNIVERSITY

Machine Learning Course

Midterm Group Assignment

Topic: **Credit Card Fraud Detection**

Date: 13 Aug 2026

Lecturer: MR. SEK SOCHEAT

Group 7:

- Mr. Chhun Menghong
- Mr. Pheuy Tay
- Mr. San Ratana
- Mr. Tang Kimhab

Agenda

1. Background
2. Objectives
3. Dataset
4. Feature Engineering
5. Machine Learning Algorithms
6. Evaluation Metrics
7. System Architecture
8. Expected System Outputs
9. Conclusion

Background

- **Credit cards** are widely used financial tools that enable cashless and online transactions, providing convenience, speed, and global accessibility for customers. However, their widespread use also increases exposure to fraud risks such as unauthorized transactions, card theft, and identity fraud.
- **Credit Card Fraud Detection,** use supervised machine learning techniques, where models are trained on historical labeled transaction data (fraud vs. non-fraud). These models learn patterns of legitimate and fraudulent behavior to automatically identify suspicious transactions in real time.

Objectives

General Objective

Develop a Machine Learning-based system for detecting fraudulent credit card transactions.

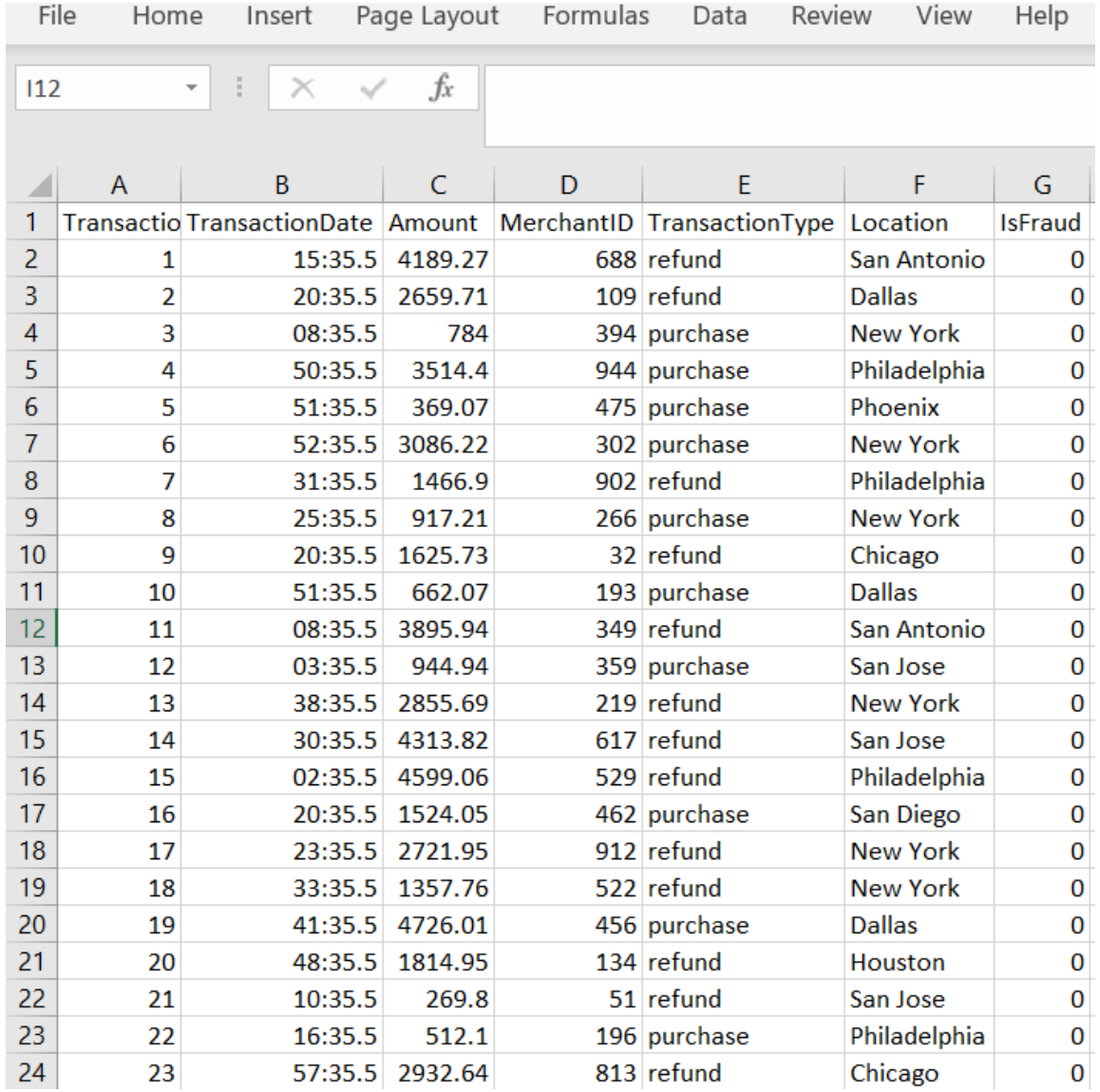
Specific Objectives

- ✔ Collect and analyze transaction data
- ✔ Preprocess data
- ✔ Implement multiple ML algorithms
- ✔ Compare model performance
- ✔ Develop an interactive fraud detection web application.

Dataset

Source Data: A widely used dataset is available on Kaggle: Credit Card Fraud Detection (https://www.kaggle.com/datasets/bhadramohit/credit-card-fraud-detection).

It contains 100,000 anonymized transactions with features such as transaction amount, time, and engineered PCA components, along with a binary fraud label.

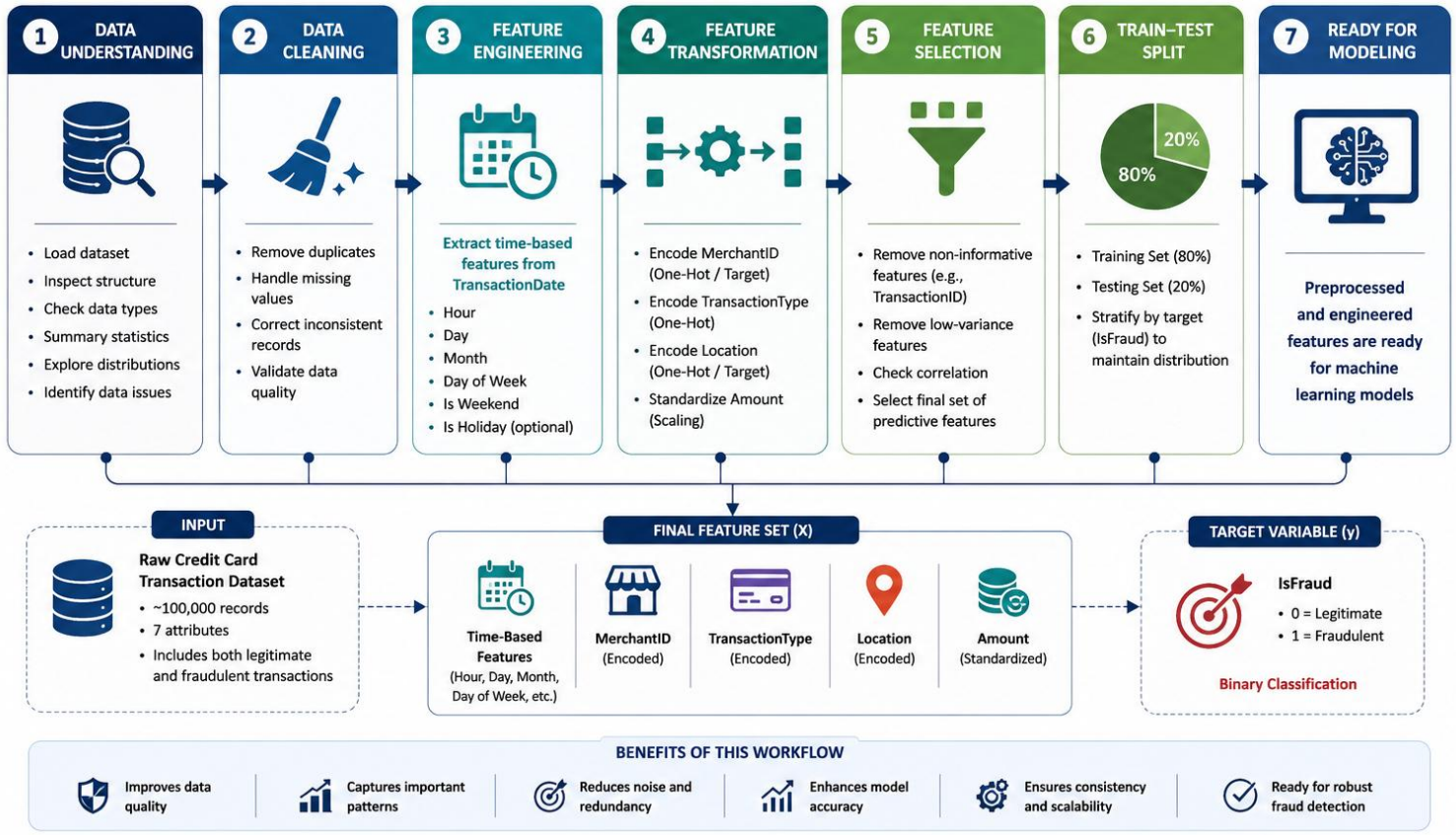


File Home Insert Page Layout Formulas Data Review View Help

I12

	A	B	C	D	E	F	G
1	Transactio	TransactionDate	Amount	MerchantID	TransactionType	Location	IsFraud
2	1	15:35.5	4189.27	688	refund	San Antonio	0
3	2	20:35.5	2659.71	109	refund	Dallas	0
4	3	08:35.5	784	394	purchase	New York	0
5	4	50:35.5	3514.4	944	purchase	Philadelphia	0
6	5	51:35.5	369.07	475	purchase	Phoenix	0
7	6	52:35.5	3086.22	302	purchase	New York	0
8	7	31:35.5	1466.9	902	refund	Philadelphia	0
9	8	25:35.5	917.21	266	purchase	New York	0
10	9	20:35.5	1625.73	32	refund	Chicago	0
11	10	51:35.5	662.07	193	purchase	Dallas	0
12	11	08:35.5	3895.94	349	refund	San Antonio	0
13	12	03:35.5	944.94	359	purchase	San Jose	0
14	13	38:35.5	2855.69	219	refund	New York	0
15	14	30:35.5	4313.82	617	refund	San Jose	0
16	15	02:35.5	4599.06	529	refund	Philadelphia	0
17	16	20:35.5	1524.05	462	purchase	San Diego	0
18	17	23:35.5	2721.95	912	refund	New York	0
19	18	33:35.5	1357.76	522	refund	New York	0
20	19	41:35.5	4726.01	456	purchase	Dallas	0
21	20	48:35.5	1814.95	134	refund	Houston	0
22	21	10:35.5	269.8	51	refund	San Jose	0
23	22	16:35.5	512.1	196	purchase	Philadelphia	0
24	23	57:35.5	2932.64	813	refund	Chicago	0

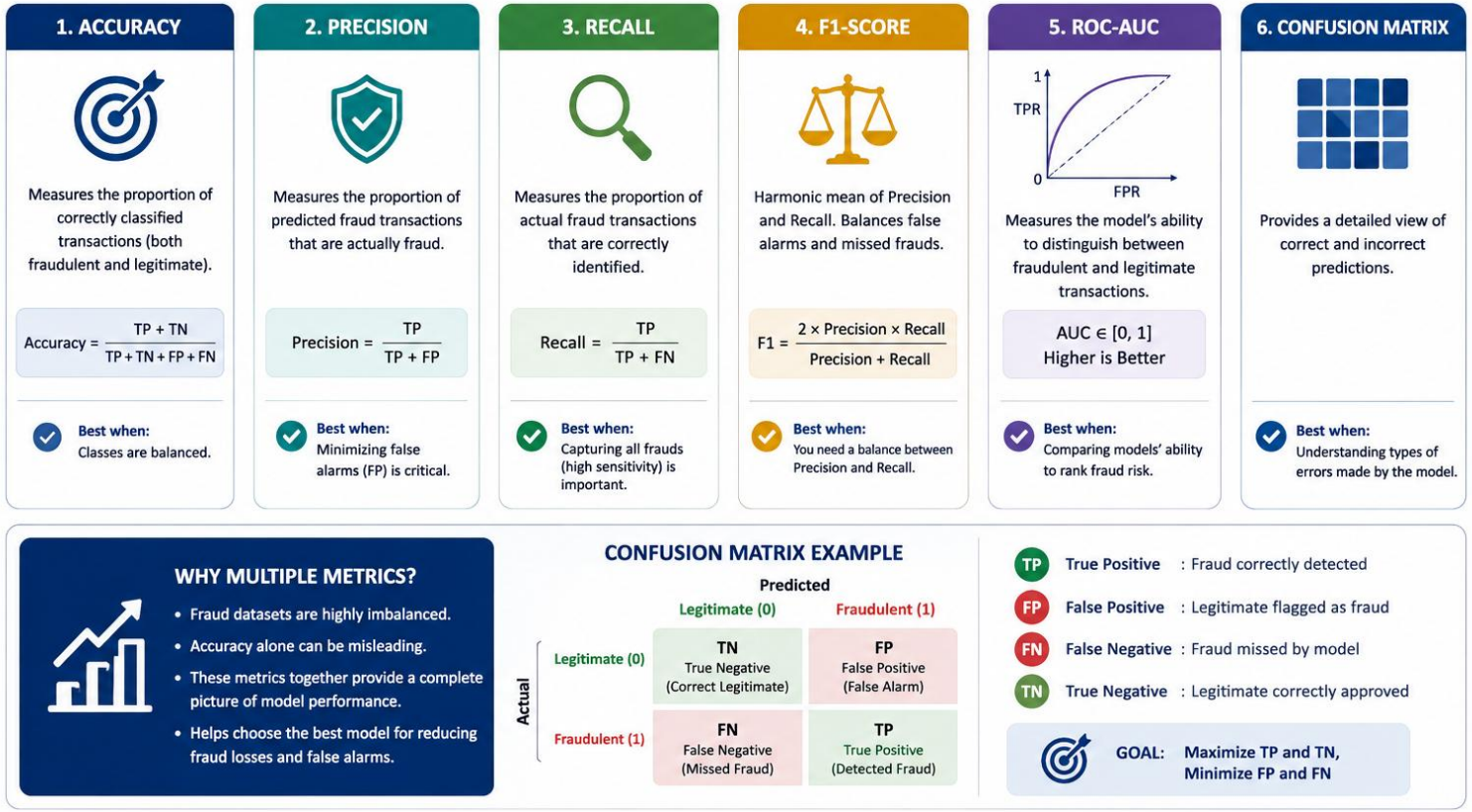
Feature Engineering



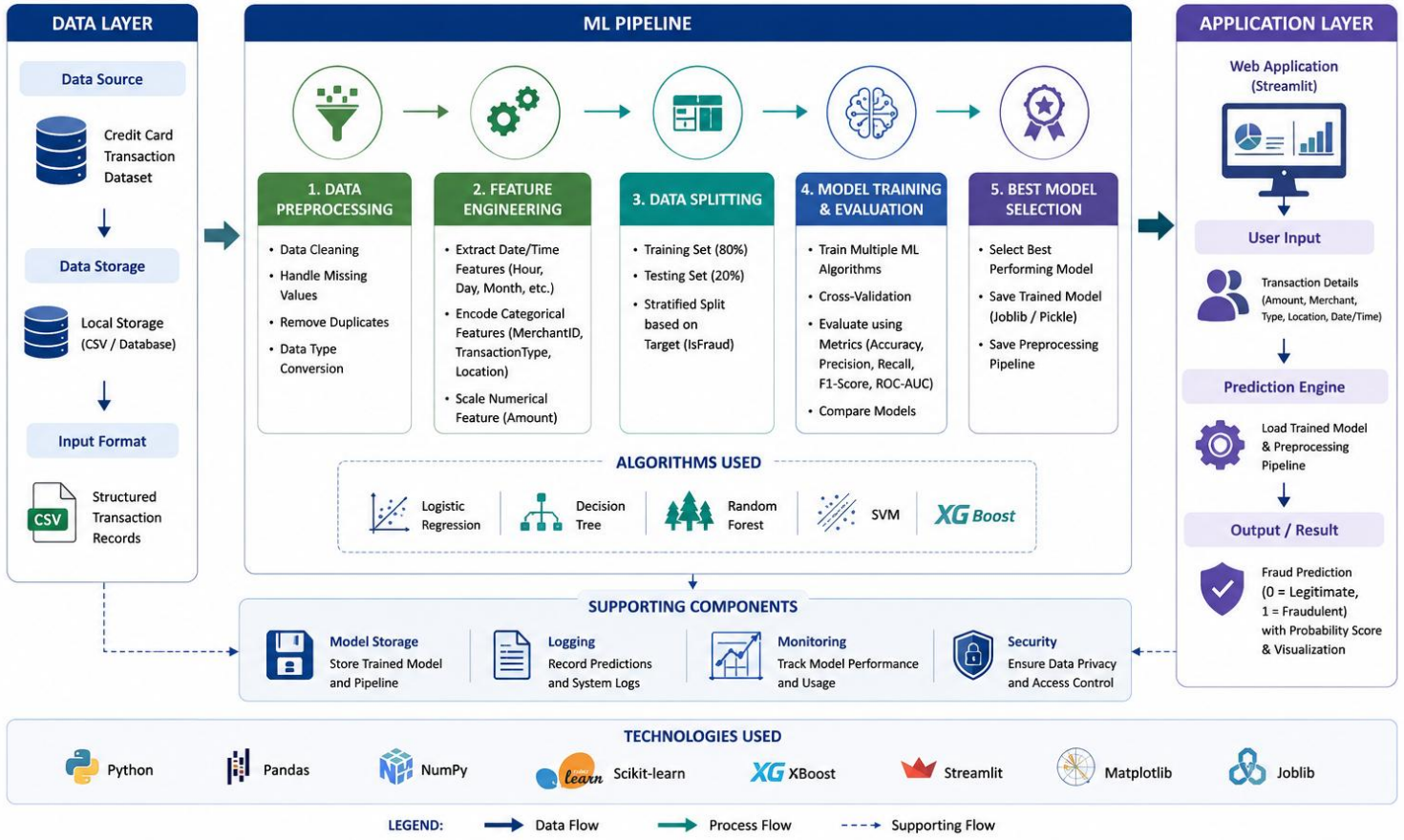
Machine Learning Algorithms

Algorithm	Purpose
Logistic Regression	Baseline Model
Decision Tree	Rule Learning
Random Forest	Ensemble Learning
XGBoost	Gradient Boosting

Evaluation Metrics



System Architecture



Expected System Outputs

1. Fraud Prediction

Fraud / Normal classification

2. Fraud Probability

Probability score for each transaction

3. Fraud Analysis

Fraud vs Normal distribution

Probability distribution

Fraud by location

4. Fraud Report

Fraud-only transaction table

Downloadable CSV report

Conclusion

Machine learning models were evaluated for credit card fraud detection.

The best-performing model was selected using F1 Score.

Feature engineering and preprocessing were applied to transaction data.

A Streamlit web application was developed for fraud prediction and analysis.

The system can generate a fraud-only transaction report.

Thank you