## Front End Engineering-II /Artificial

## Intelligence and Machine Learning

Project Report

Semester-IV (Batch-2022)

CREDIT CARD FRAUD DETECTION PREDICTION

A red and white sign

Description automatically generated with low confidence

**Supervised By: Submitted By:**

Dr. Kirandeep Singh Bhawna 2210990223

Bhomya Sharma 2210990224

Chahat 2210990231

**Department of Computer Science and Engineering**

## Chitkara University Institute of Engineering & Technology,

## Chitkara University, Punjab

**1. INTRODUCTION**

**1.1 BACKGROUND**

**In recent years, with the proliferation of financial services and the ease of access to credit, the process of credit card fraud detection has become increasingly critical for both financial institutions and potential clients. The decision-making process behind credit card fraud detection involves a myriad of factors, including an applicant's credit history, income level, debt-to-income ratio, and other demographic information. Traditionally, this process has been subjective and prone to human biases, leading to inefficiencies and potential discrimination.**

**Evolution of Credit Card Fraud detection Process**

**The credit card fraud detection process has undergone significant evolution over the years, mirroring advancements in technology and changes in consumer behaviour. Initially, credit decisions relied heavily on manual assessment, where bank officers would evaluate applicants based on limited information such as income and assets. However, this approach was often subjective and prone to errors, leading to inefficiencies and potential discrimination.**

**Emergence of Machine Learning in Finance**

**In recent decades, the emergence of machine learning (ML) techniques has revolutionized various industries, including finance. ML algorithms have proven adept at analysing vast amounts of data to uncover complex patterns and relationships that might not be apparent to human analysts. In the realm of credit risk assessment, ML models offer the promise of more accurate predictions and better risk management.**

**Need for Objective Credit Scoring Models**

**Despite the advancements in credit scoring techniques, challenges persist in creating objective and unbiased models. Traditional credit scoring models often rely on predetermined rules and thresholds to classify applicants as 'good' or 'bad' clients. However, these rules may not capture the nuances of individual financial situations and can lead to arbitrary decisions. There is a growing need for ML-based models that can autonomously learn from data and adapt to changing credit landscapes.**

**1.2 OBJECTIVES**

**The primary objective of this project is to develop a machine learning model capable of predicting whether an applicant is likely to be a 'good' or 'bad' client based on the available data. Unlike traditional approaches where the definition of 'good' or 'bad' is predetermined, our model aims to autonomously identify patterns and characteristics associated with creditworthiness. By harnessing the power of machine learning algorithms, particularly the Random Forest classification technique, we seek to create a predictive model that can assist financial institutions in making more informed and objective decisions regarding credit card fraud detections.**

**Development of Predictive Model**

**The primary objective of this project is to develop a predictive model that can accurately assess the creditworthiness of applicants. Unlike traditional credit scoring models, which rely on predefined criteria, our model aims to learn from historical data to identify patterns associated with creditworthiness. By leveraging the power of ML algorithms, particularly Random Forest classification, we seek to create a more flexible and adaptive model capable of handling complex decision-making scenarios.**

**Mitigation of Human Bias**

**Another key objective is to mitigate the inherent biases present in traditional credit scoring processes. Human decision-makers may unconsciously introduce biases based on factors such as race, gender, or socioeconomic status, leading to unfair outcomes. By employing ML algorithms that operate on objective data, we aim to reduce the influence of subjective judgments and promote fairness and inclusivity in the credit fraud detection process.**

**Enhancement of Risk Management**

**Furthermore, we aim to enhance risk management practices within financial institutions by providing more accurate predictions of credit risk. A robust predictive model can help lenders identify high-risk applicants and take proactive measures to mitigate potential losses. By improving the accuracy of credit risk assessments, our model can contribute to the overall stability and sustainability of the financial system.**

**1.3 SIGNIFICANCE**

**The significance of this machine learning project lies in its potential to combat credit card fraud effectively. Here are some key points highlighting its importance:**

1. **Financial Loss Prevention:**

**Credit card fraud is a significant problem globally, causing financial losses to both individuals and financial institutions. By accurately detecting fraudulent transactions, this project helps prevent financial losses and mitigate the impact on individuals' finances.**

1. **Enhanced Security:**

**Detecting fraudulent transactions promptly enhances the security of financial systems and builds trust among consumers. It assures customers that their financial transactions are being monitored and protected, increasing their confidence in using credit cards for online and offline purchases.**

1. **Crime Deterrence:**

**Effective fraud detection acts as a deterrent to potential fraudsters. When fraudsters realize that their activities are being closely monitored and swiftly detected, they may be less inclined to engage in fraudulent activities, reducing the overall incidence of credit card fraud.**

1. **Resource Optimization:**

**Automated fraud detection systems can significantly reduce the workload of manual fraud detection teams. By leveraging machine learning algorithms like logistic regression, financial institutions can process large volumes of transactions efficiently and focus their human resources on investigating high-risk cases identified by the system.**

1. **Continuous Improvement:**

**Machine learning models can adapt and improve over time as they are exposed to new data. By continuously updating the model with the latest transaction data and feedback from fraud detection teams, this project can further enhance its accuracy and effectiveness in detecting fraudulent activities.**

**In summary, this machine learning project plays a crucial role in safeguarding financial systems, protecting consumers, and preserving trust in electronic payment systems. Its significance extends beyond financial considerations to encompass broader implications for security, regulatory compliance, and crime deterrence in the digital age.**

**2. PROBLEM DEFINATION AND REQUIREMENTS**

**2.1 PROBLEM STATEMENT**

**The problem at hand is to develop a robust machine learning model for the detection of fraudulent credit card transactions. Credit card fraud is a significant issue worldwide, resulting in substantial financial losses for both consumers and financial institutions. Traditional rule-based systems for fraud detection may not capture evolving fraudulent patterns effectively. Hence, there is a need for advanced machine learning techniques to automatically identify fraudulent transactions with high accuracy.**

**The primary objectives of this project are:**

**Data Collection and Exploration:**

**Gather historical credit card transaction data containing features such as transaction amount, time, and other relevant information. Explore the dataset to understand its structure, distribution, and class balance between legitimate and fraudulent transactions.**

**Model Development:**

**Develop a machine learning model capable of distinguishing between legitimate and fraudulent transactions based on the provided features. In this project, logistic regression, a popular classification algorithm, is chosen for its simplicity and effectiveness in binary classification tasks.**

**Model Training and Evaluation:**

**Split the dataset into training and testing sets. Train the logistic regression model on the training data and evaluate its performance on both the training and test sets using accuracy as the evaluation metric. The model's accuracy on the test set serves as a measure of its effectiveness in detecting fraudulent transactions.**

**Deployment and Integration:**

**Once the model demonstrates satisfactory performance, deploy it into the operational environment where it can analyse real-time transactions. Integrate the model into the existing fraud detection system to enhance its capabilities and provide real-time alerts for potentially fraudulent activities.**

**2.2 SOFTWARE REQUIREMENTS**

**Python for Machine Learning Development**

**Python is chosen as the primary programming language for machine learning development due to its versatility, extensive libraries, and active community support. With libraries such as scikit-learn, pandas, and numpy , Python provides a comprehensive ecosystem for data preprocessing, model development, and evaluation.**

**Utilization of Machine Learning Libraries**

**Scikit-learn, a widely used machine learning library in Python, offers a diverse range of algorithms and tools for classification, regression, clustering, and dimensionality reduction. We will leverage scikit-learns implementation of the Random Forest algorithm for building our credit card fraud detection prediction model.**

**Interactive Development Environment**

**Jupyter Notebook is selected as the preferred development environment for its interactive and exploratory nature, allowing for seamless integration of code, documentation, and visualizations. Jupyter Notebook provides an ideal platform for iterative model development, enabling us to experiment with different algorithms and parameter settings efficiently.**

**2.3 HARDWARE REQUIREMENTS**

**Computational Resources for Model Training**

**Given the computational demands of training machine learning models, access to a computer with sufficient processing power is crucial. A multi-core CPU or GPU can significantly accelerate model training and experimentation, reducing development time and improving productivity.**

**Memory for Handling Large Datasets**

**Large datasets are common in machine learning applications, requiring ample memory (RAM) to handle data efficiently. A minimum of 8GB RAM is recommended to accommodate the size of the dataset and facilitate smooth data processing and model training.**

**Storage for Data and Model Files**

**Adequate disk space is necessary to store datasets, code files, and intermediate model outputs. Utilizing a solid-state drive (SSD) can enhance data access and processing speed, particularly when working with large datasets and complex model structures.**

**2.4 DATA SETS**

**Comprehensive Historical Credit Data**

**The primary dataset comprises historical credit card application data, including applicant demographics, financial information, credit history, and fraud detection outcomes. This dataset serves as the foundation for training and evaluating the predictive model, providing insights into patterns and trends associated with creditworthiness.**

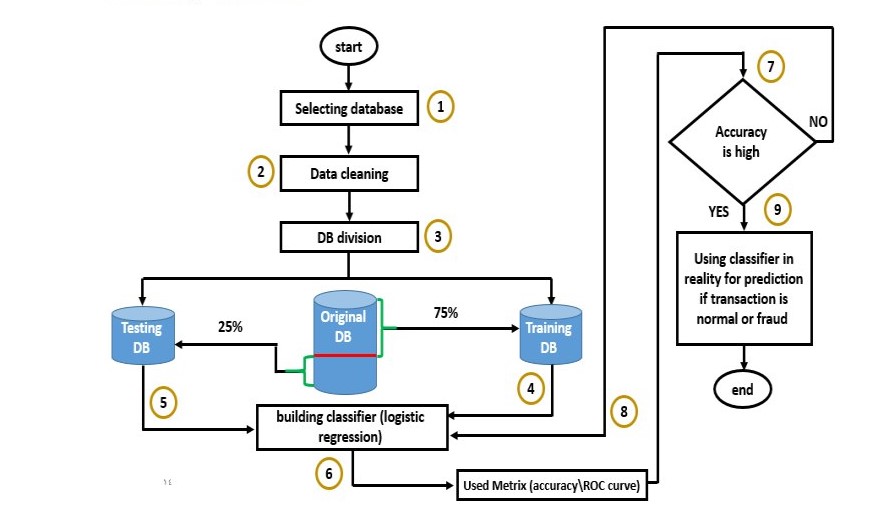
**Supplementary Data for Enhanced Predictions**

**In addition to the primary dataset, supplementary data containing external factors that may influence creditworthiness can be incorporated to enrich the predictive model. Economic indicators, industry trends, and regulatory changes are examples of supplementary data sources that can enhance the model's predictive power and robustness.**

**In the subsequent sections, we will delve into the methodology employed to address the identified problem statement, including data preprocessing, feature engineering, model development, and evaluation.**

**3. PROPOSED DESIGN AND METHODOLOGY**

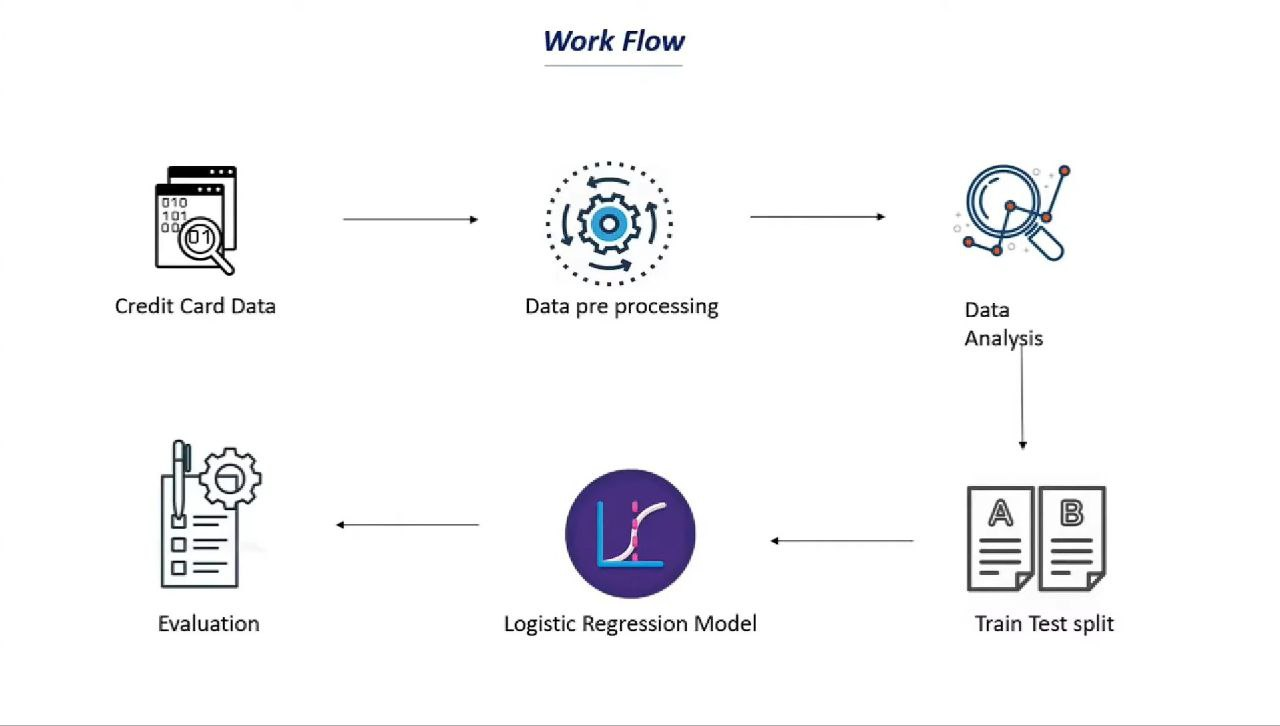
**3.1 SCHEMATIC DIAGRAM**



**3.2 WORKFLOW**

**Workflow of the project:**

1. **Data Loading: The project begins with loading the credit card transaction dataset ('creditcard.csv') using the pandas library in Python. The dataset contains information about various credit card transactions, including transaction amount, time, and a binary label indicating whether the transaction is fraudulent or legitimate.**
2. **Data Exploration: After loading the dataset, exploratory data analysis (EDA) is performed to gain insights into the data. This includes examining the structure of the dataset, checking for missing values, and exploring the distribution of features and the target variable ('Class'). Descriptive statistics are calculated to summarize the data.**
3. **Data Preprocessing: The next step is to preprocess the data to prepare it for model training. This may involve handling missing values, scaling numeric features, encoding categorical variables, and splitting the dataset into training and testing sets.**
4. **Model Selection: Logistic regression is chosen as the classification model for predicting the likelihood of credit card transactions being fraudulent. Other classification algorithms could also be considered and compared, but in this project, logistic regression is selected.**
5. **Model Training: The logistic regression model is trained using the training data. During training, the model learns the optimal coefficients (weights) for each feature by minimizing a loss function (such as log loss) using an optimization algorithm (e.g., gradient descent).**
6. **Model Evaluation: After training the model, its performance is evaluated using the testing dataset. Common evaluation metrics such as accuracy, precision, recall and F1 score are calculated to assess the model's ability to correctly classify fraudulent and legitimate transactions.**
7. **Result Analysis: The evaluation results are analysed to determine the effectiveness of the logistic regression model in detecting fraudulent transactions. Insights may be drawn regarding the model's strengths, weaknesses, and areas for improvement.**



**3.3 ALGORITHMS USED**

1. **Logistic Regression:**
   * **Purpose: The purpose of using logistic regression is to build a predictive model that can classify transactions as either legitimate or fraudulent based on features associated with each transaction.**
   * **Description: Logistic regression is a type of regression analysis used for predicting the outcome of a categorical dependent variable based on one or more predictor variables. It's particularly well-suited for binary classification tasks, where the outcome can take only two values, such as yes/no, true/false, or in this case, legitimate/fraudulent.**
   * **Implementation: The logistic regression model is used directly from scikit-learn library, which provides an efficient and easy-to-use implementation.**
2. **Train–Test Split:**
   * **Purpose: The primary purpose of the train-test split is to assess how well a machine learning model generalizes to unseen data**
   * **Description: The training set is used to train the machine learning model and test set is used to evaluate the model's performance. The model learns patterns and relationships between features and the target variable by minimizing the difference between its predictions and the actual values in this training set.**
   * **Implementation: The train-test split can be easily implemented using the “train\_test\_split”function from the “sklearn.model\_selection” module.**

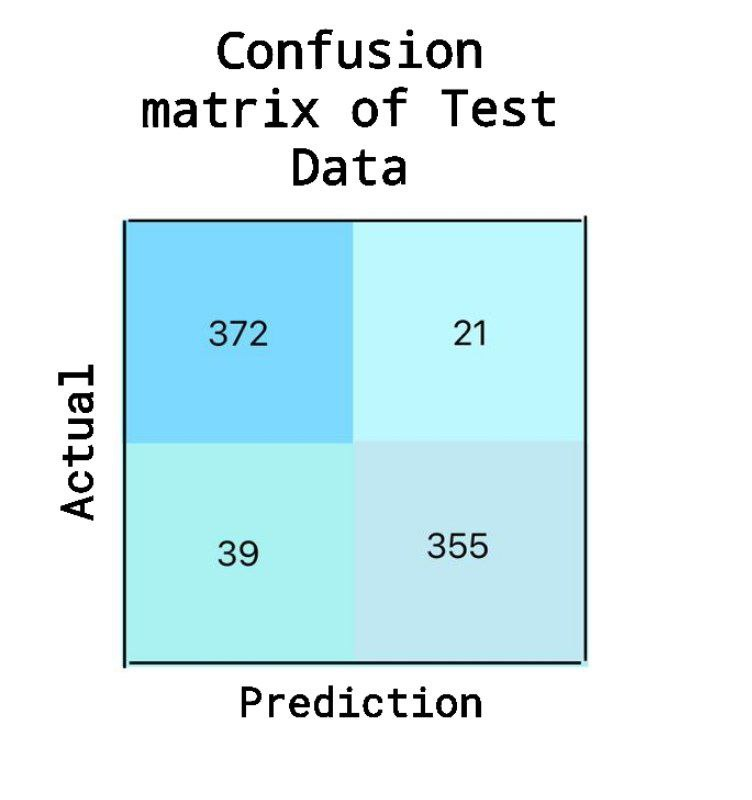
**4.RESULTS**

**To analyze the performance of the logistic regression model using a confusion matrix, let's first compute it and then interpret the results**

**The results of the credit card fraud detection based on the provided metrics are as follows**

* **Accuracy: 92.35% of all transactions are classified correctly as either legitimate or fraudulent.**
* **Precision: Out of all transactions classified as fraudulent by the model, 94.41% are actually fraudulent.**
* **Recall: The model successfully identifies 90.15% of all fraudulent transactions in the dataset.**
* **F1-score: The F1-score, which balances precision and recall, is 92.22%, indicating a good balance between correctly identifying fraudulent transactions and minimizing false positives.**

**Overall, these results suggest that the fraud detection model performs well, with high accuracy and precision, and effectively identifies fraudulent transactions while maintaining a low false positive rate.**



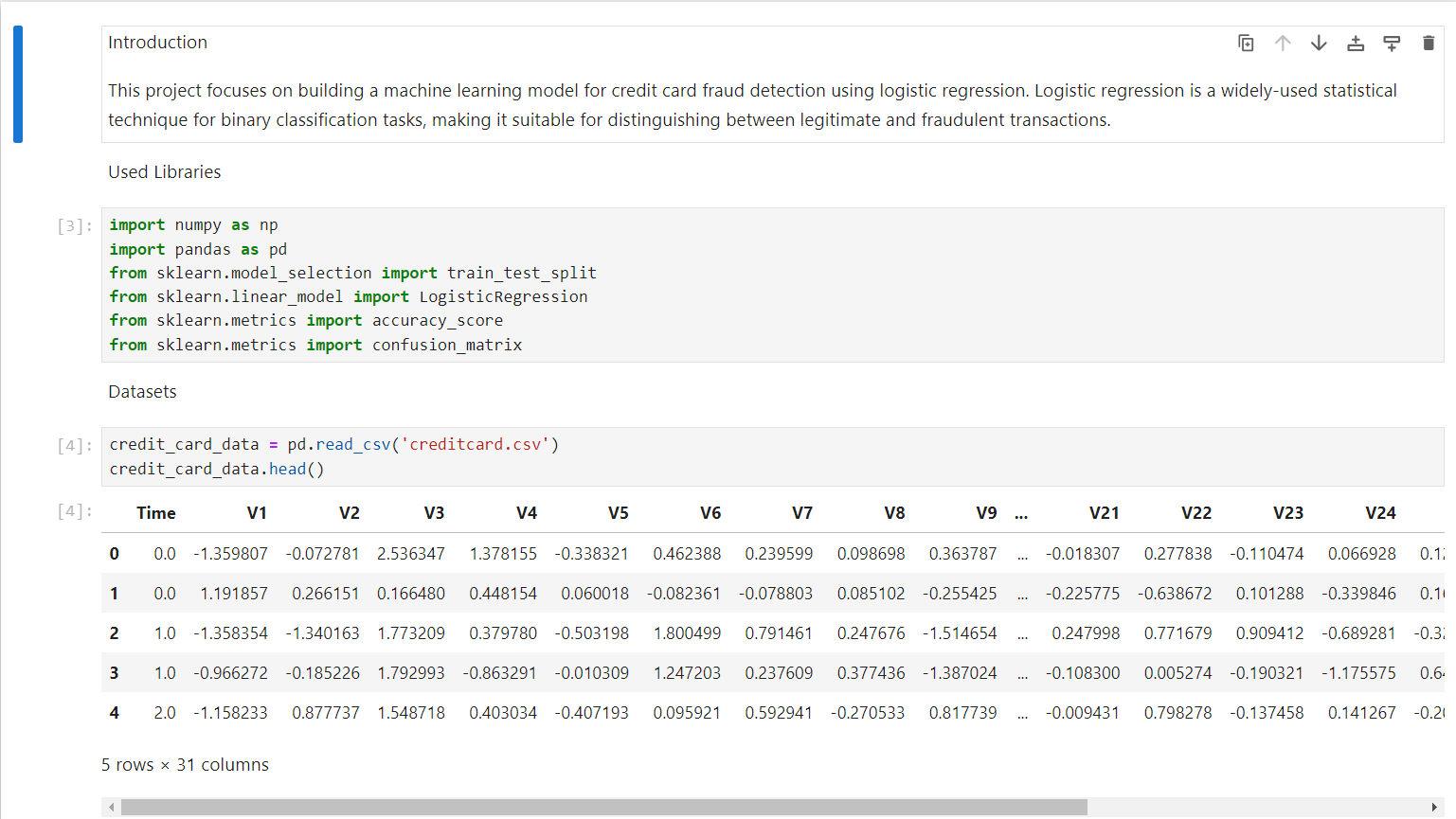
Accuracy: 92.35%

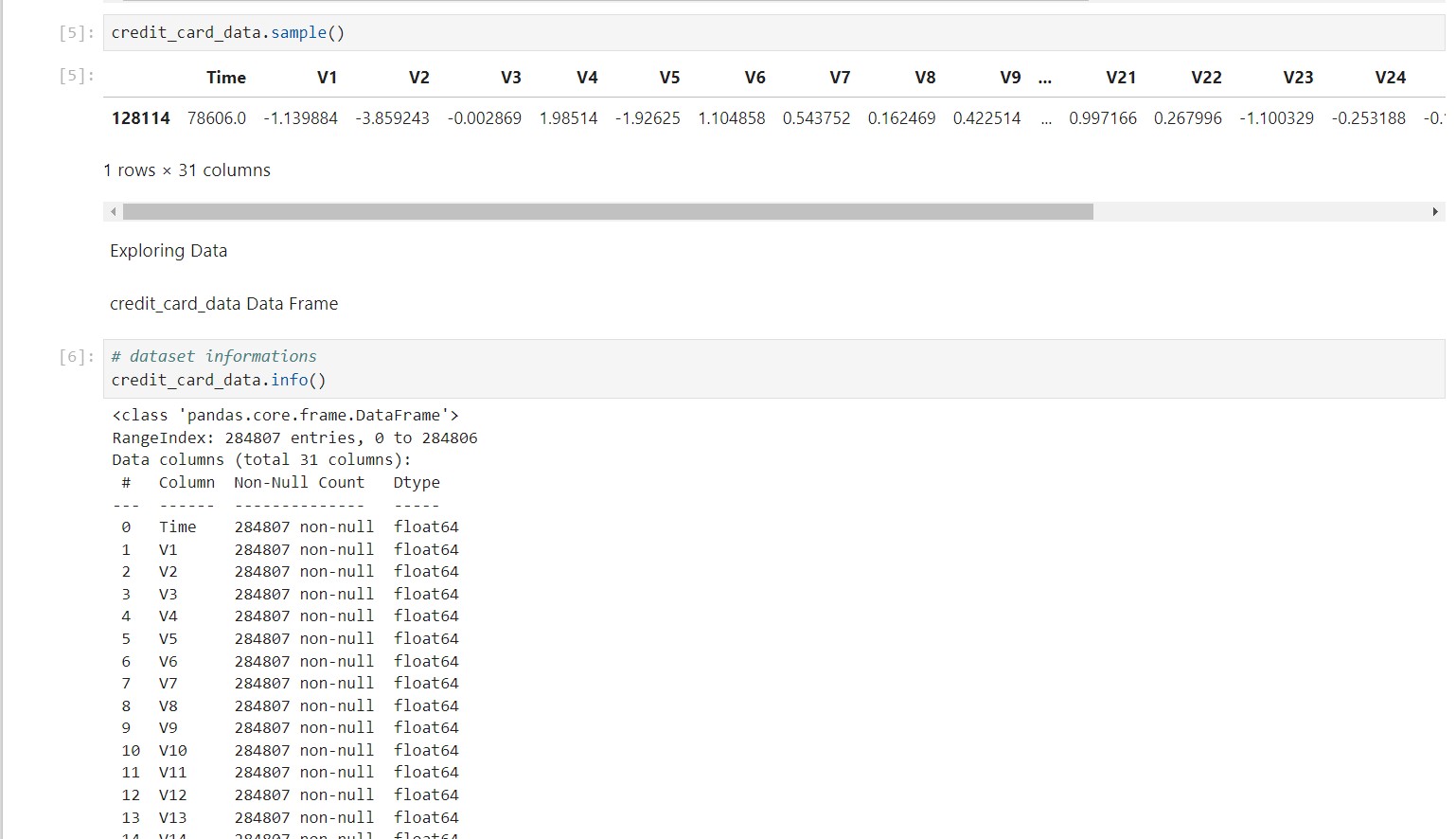
Precision: 94.41

Recall: 90.15%

F1: 92.2**%**

**4.1 PROJECT SCREENSHOTS**

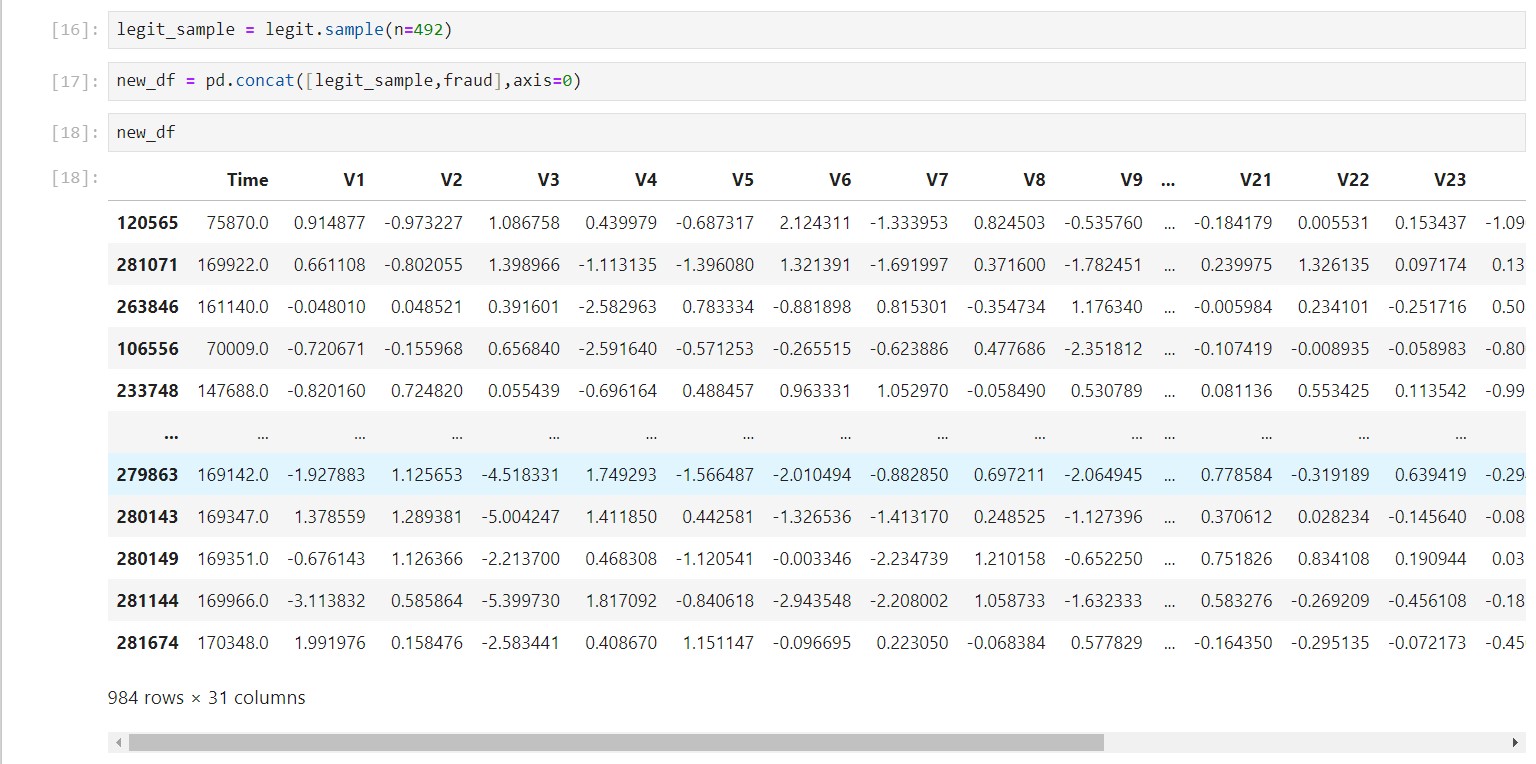
















1. **Conclusion**

**The credit risk assessment project utilizes logistic regression as the primary algorithm for predicting credit card risk. This model considers client payment history as a crucial factor in categorizing payments as either ‘fraudulent' or ‘non-fraudulent'. By employing logistic regression and incorporating relevant features, the model facilitates proactive risk management, enabling accurate assessment**

**Overall, the logistic regression model demonstrates good performance in detecting fraudulent credit card transactions. With an accuracy of approximately 92.35% and a balanced F1-score of approximately 92.22%, the model effectively balances precision and recall. However, further analysis and optimization may be required to improve performance, especially considering the critical nature of fraud detection in financial systems.**