# C964: COMPUTER SCIENCE CAPSTONE

## Using Machine Learning to Identify Fraudulent Transactions

Melanie Akau
mkkakau@gmail.com

# Contents

Part A: Project Proposal for Business Executives ..... 3
Letter of Transmittal ..... 3
Project Recommendation ..... 4
Problem Summary ..... 4
Application Benefits ..... 4
Application Description ..... 4
Data Description ..... 5
Objectives and Hypothesis ..... 5
Methodology ..... 5
Funding Requirements ..... 6
Data Precautions ..... 6
Developer's Expertise ..... 6
Part B: Project Proposal ..... 7
Problem Statement ..... 7
Customer Summary ..... 7
Existing System Analysis ..... 8
Data ..... 8
Project Methodology ..... 10
Agile ..... 10
SEMMA ..... 11
Project Outcomes ..... 12
Implementation Plan ..... 12
General Strategy ..... 12
Rollout Phases ..... 12
Evaluation Plan ..... 13
Resources and Costs ..... 13
Timeline and Milestones ..... 14
Part C: Application ..... 15
Part D: Post-implementation Report ..... 16
Business Vision ..... 16
Datasets ..... 16
Data Product Code ..... 18
Objective Verification ..... 19
Effective Visualization and Reporting ..... 20

Accuracy Analysis ..... 22
Application Testing ..... 22
Application Files ..... 24
User Guide ..... 24
Summation of Learning Experience ..... 25
References ..... 26

10/5/2023

# Part A: Project Proposal for Business Executives

## Letter of Transmittal

Melanie Akau
PO Box 315
Kurtistown HI 96760

ABC Bank
Attn: Senior Leadership
123 Money Ln
Honolulu HI 96810

Subject: Enhancing Credit Card Fraud Detection with Advanced Machine Learning Techniques

Dear Senior Leadership Team:

In our rapidly evolving digital landscape, our fraud department is becoming overwhelmed by the increasing number of fraudulent transactions. Traditional manual fraud detection methods, while effective in the past, are struggling to keep pace with the sophisticated tactics employed by modern fraudsters. Recognizing this challenge, I propose developing a solution powered by machine learning. Unlike manual methods, machine learning can continuously learn and adapt to new patterns of fraud, making it a more dynamic and effective tool for identifying suspicious activities.

By integrating machine learning into our fraud detection process, we can significantly reduce financial losses and improve customer trust. This approach offers a more proactive stance, ensuring smoother transactions for our customers and reducing the strain on our fraud department.

To bring this vision to life, a preliminary cost assessment suggests an initial investment of approximately $114,000. The largest initial investment will be the cost of hardware at $66,000 since they will need to process many transactions in real time. The cost for personnel needed for the development of the program is around $41,000. We estimate that licensing and tools will cost $7,000.

With years of experience in both financial data science and cybersecurity, I've been part of teams that have worked to enhance our bank's operations and security. My background in analyzing data and understanding cybersecurity challenges provides a solid foundation for developing a machine learning solution to address our concerns with fraudulent transactions. By leveraging machine learning, we can stay ahead of potential threats, safeguard our customers, and reinforce our bank's reputation as a secure and forward-thinking institution. I am confident that with the collective support of our leadership and my team's expertise, we can usher in a new era of enhanced security and customer trust.

Sincerely,

*Melanie Akau*

Melanie Akau
Lead Data Scientist
ABC Bank

# Project Recommendation

## Problem Summary

As technology advances, so does the prevalence of cybercrime, with credit and debit card fraud becoming a significant concern. To address this, we propose a project to integrate machine learning into ABC Bank's fraud detection system. This initiative aims to harness the capabilities of machine learning, a computer technology that learns from data to make decisions or predictions, enhancing our bank's ability to identify and counteract fraudulent transactions.

In today's digital banking landscape, the instances of fraudulent activities are rising alarmingly. Traditional methods, while valuable, are becoming less effective against the sophisticated tactics of modern fraudsters. ABC Bank needs a solution that not only addresses the current challenges but also anticipates and adapts to future threats. Implementing machine learning in our fraud detection system is a proactive step towards this goal.

The proposed machine learning solution aligns directly with ABC Bank's need for enhanced security and efficiency. By automating the fraud detection process, we can swiftly identify potential threats, ensuring a safer banking environment. This not only safeguards our financial assets but also reinforces our commitment to customer trust and satisfaction.

Upon completion, the project will deliver a machine learning model integrated into ABC Bank's transaction system. This model will continuously analyze transaction patterns, flagging potentially fraudulent activities for review. The end goal is a significant reduction in undetected fraudulent transactions, a streamlined fraud detection process, and an enhanced customer banking experience. This proposal further delves into the advantages of implementing machine learning algorithms in our fraud detection system and the business benefits it promises.

## Application Benefits

The proposed machine learning solution is uniquely tailored to address the challenges of ensuring transactional security while maintaining operational efficiency. Unlike traditional fraud detection methods that rely heavily on predefined rules and manual oversight, our machine learning model can be trained on new data from transactions which have been verified as fraudulent. This continuous learning approach ensures that as fraudsters evolve their tactics, our system adapts in real-time, recognizing new patterns and threats. By staying ahead of these emerging tactics, ABC Bank ensures protection for its assets and upholds its reputation as a secure financial institution.

Implementing the machine learning solution offers ABC Bank multifaceted benefits. Financially, by reducing the number of undetected fraudulent transactions, the bank stands to decrease associated losses, leading to healthier bottom lines. Operationally, the automation of fraud detection will free up valuable resources, allowing the fraud department to focus on complex cases and strategic initiatives. For our customers, fewer false positives mean smoother transactions and a heightened sense of security. Collectively, these benefits not only enhance ABC Bank's operational efficiency but also solidify its position as a trusted and customer-centric financial institution.

## Application Description

The machine learning application we're developing will be trained to recognize patterns and relationships from a training dataset. The specific method we will be using is the Random Forest Classification algorithm. The algorithm will analyze the data based on the different features such as time, amount, location, etc. The application will use this algorithm to predict if a transaction is fraudulent or non-fraudulent. This prediction will be tested against the actual classification and the machine learning model will learn to identify correlations that may signal that a transaction is fraudulent. Once this

training phase is complete, the application will be capable of making predictions on new transactions, identifying potentially fraudulent activities.

## Data Description

The data was collected by Worldline and Machine Learning Group from Université Libre de Bruxelles. It is a collection of 284,807 transactions from European consumers. Out of those transactions, 492 are classified as fraudulent. The data has been collected from real credit card transactions made in 2013.

The dataset is structured in a file in tabular format. It contains many features of a transaction, and all fields are quantitative except for the fraudulent or non-fraudulent classification. The dataset contains various factors, such as the time and amount of each transaction which help us determine the nature of the transaction. Many of the factors are renamed to variables labeled V1-V28 and processed through filters in order to protect customer privacy. The dataset encompasses various independent variables, such as transaction time and amount, that provide insights into each transaction's details. The machine learning algorithm will use these variables to identify patterns. The key outcome we're focused on is the nature of the transaction, labeled as either fraudulent or non-fraudulent. This is the dependent variable.

A significant aspect to note is the rarity of fraudulent transactions; they represent a very small portion of the dataset. This imbalance necessitates further data transformation and optimization to ensure accurate predictions. Although some original data points have been adjusted for privacy considerations, the dataset still contains the vital information necessary for our analysis, and the algorithm can effectively utilize this transformed data, albeit on a different scale.

## Objectives and Hypothesis

While we aim for our system to be 80% accurate, accuracy alone doesn't give us the full picture because of the unique nature of our data. Instead, we're focusing on two key measures: precision and recall, both targeted at 70%. In simple terms, precision ensures that when our system flags a transaction as suspicious, it's correct 70% of the time. Recall ensures that our system catches 70% of all actual fraudulent transactions. Achieving these targets will strike the right balance between safeguarding our customers' finances and ensuring they have uninterrupted access to their accounts.

## Methodology

For the development and implementation of this project, we have chosen the Agile methodology. Agile is ideal for our fraud detection project due to its emphasis on iterative progress and adaptability. As we delve deeper into the complexities of fraud detection, requirements may evolve. Its emphasis on collaboration, adaptability, and continuous feedback ensures that we remain aligned with the project's objectives while being responsive to new insights or challenges.

### *Project Methodology Outline*

#### Concept

Stakeholders from the fraud department, cybersecurity team, and product owners collaborate to define the scope and priorities of the fraud detection system, aligning with the bank's security objectives.

#### Inception

The team is assembled, roles are assigned, and essential tools are provided. The primary focus during this phase is on developing a machine learning model tailored for fraud detection, ensuring it can effectively identify patterns and anomalies in transaction data.

#### Iteration

The team works through the product backlog in development sprints, building the fraud detection system iteratively, allowing for a foundational system that can adapt to evolving fraud patterns.

### Testing

Comprehensive testing ensures the system's reliability, from evaluating individual components like fraud alert triggers to verifying the system's compatibility with the bank's existing infrastructure and ensuring it meets the fraud department's requirements.

### Release

After thorough testing, the fraud detection system is prepared for integration into the bank's operations. Training sessions ensure the fraud department can utilize the system effectively.

### Review

Post-integration, the focus is on continuous improvement. Feedback from the fraud department is gathered, potential refinements are identified, and the team provides ongoing support to ensure the system's optimal performance.

## Funding Requirements

| | |
|---|---|
| Environment<br><br>Dedicated to top-tier cloud solutions, computing resources, and secure servers, ensuring our system operates at its best and remains secure. | 66,000.00 |
| Personnel<br><br>Allocated for our specialized team, encompassing data scientists, cybersecurity experts, and developers, driving the project to success. | 41,000.00 |
| Licensing & Tools<br><br>Covers essential software licenses and collaborative tools, allowing us to leverage the best in technology and maintain efficient teamwork. This also includes security software licenses for user control and security. | 7,000.00 |
| Total | $114,000.00 |

## Data Precautions

The dataset we're utilizing contains transactional data of European cardholders, which, even though transformed for confidentiality, remains sensitive due to its financial nature. When handling this data, it's imperative to ensure encrypted storage and transmission, limit access to only authorized personnel, and regularly monitor data access and usage. While our dataset has undergone transformations for privacy, it's worth noting that public datasets, like those from platforms such as Kaggle.com, typically don't have the same restrictions. Irrespective of the data's source, the bank is mandated to uphold the highest standards of customer privacy and data security.

## Developer’s Expertise

With specialized training and a degree in Computer Science, coupled with hands-on experience in financial data science and cybersecurity, I've consistently contributed to enhancing our bank's operations and security. My expertise, rooted in analyzing complex data patterns and understanding cybersecurity nuances, is directly aligned with the demands of this project. Reinforcing our customers' trust and our bank's reputation is paramount. For this project, understanding the nuances of financial transactions and the potential security vulnerabilities is essential. My qualifications, deeply rooted in data science and cybersecurity, are crucial in navigating these intricacies.

# Part B: Project Proposal

## Problem Statement

In today's interconnected world, the convenience of digital transactions has transformed the way consumers interact with financial institutions. Credit and debit cards have become ubiquitous, allowing consumers to seamlessly conduct transactions across the globe. However, this digital revolution has also opened the door to new challenges. Cybercriminals, equipped with evolving techniques, are finding innovative ways to breach security barriers and access sensitive card information.

The digital transformation of banking and commerce has inadvertently made it easier for fraudsters to access and misuse sensitive information. While ABC Bank has implemented stringent security measures, encrypting data, and ensuring secure transactions, the vast digital landscape presents numerous opportunities for potential breaches. The Consumer Sentinel Network Data Book 2022 paints a concerning picture, highlighting that fraud accounted for a staggering 46% of all reported cases, amounting to 2.2 million reports.

As technology continues to advance, so does the toolkit of the cybercriminal. They are no longer relying solely on brute force or hacking methods. Instead, they're innovating, finding new vulnerabilities in systems, and exploiting them before institutions can patch them up. Moreover, they've recognized the weakest link in the security chain: the customers themselves. Through sophisticated social engineering tactics, fraudsters manipulate customers into revealing their confidential information, bypassing even the most robust security systems.

Despite the evolving nature of cyber threats, many banks, including ABC Bank, still predominantly rely on traditional, manually programmed rules for fraud detection. These static systems, while effective in the past, are now struggling to keep up with the dynamic and innovative methods employed by fraudsters. The limitations of these systems become evident when they fail to recognize new patterns of fraud.

The challenges posed by cybercriminals in this digital age necessitate a paradigm shift in how banks approach fraud detection. Traditional methods are no longer sufficient on their own. Machine learning offers adaptability and the ability to learn from past patterns to predict future threats. By integrating machine learning into our fraud detection systems, ABC Bank can stay a step ahead of cybercriminals, ensuring the security of transactions and maintaining customer trust.

## Customer Summary

ABC Bank, a leading financial institution in Hawaii, serves a vast clientele through its extensive network of over 50 branches spread across the islands. With its headquarters in Honolulu, the bank employs over 500 individuals, managing a myriad of services from checking accounts to credit products. Given its expansive operations and the sheer volume of transactions processed daily, the bank faces an ever-increasing challenge: ensuring the security of these transactions. While ABC Bank has always prioritized its customers' financial safety, the evolving landscape of cyber threats necessitates a more advanced approach. Traditional fraud detection systems are becoming less effective against the sophisticated tactics employed by modern fraudsters. Moreover, with the bank's commitment to customer trust and its reputation at stake, there's an urgent need to strengthen its defenses against potentially fraudulent activities.

To address the growing concerns of cyber threats and to stay ahead of potential vulnerabilities, the introduction of a machine learning-based solution is key. Unlike traditional systems that rely on pre-defined rules, machine learning continuously learns and adapts from transactional data. This dynamic nature allows it to identify emerging patterns of fraud, even those that haven't been previously encountered. By implementing this machine learning application, ABC Bank can proactively detect and

counteract fraudulent activities, ensuring the security of transactions and reinforcing customer trust. This not only positions the bank as a technologically advanced institution but also ensures that it remains resilient against the ever-evolving threats in the digital age.

## Existing System Analysis

ABC Bank currently employs manually programmed fraud detection algorithms to safeguard its transactions. While these algorithms have been effective in identifying numerous fraudulent activities, they also produce a significant number of false positives. This has led to a surge in call center volumes, with customers either having to phone in or visit a branch whenever their card is mistakenly flagged. The repercussions are twofold: customers face prolonged wait times, both in branches and over the phone, leading to dissatisfaction, and bank employees, particularly in customer service and the fraud department, are overwhelmed with the increased workload.

Furthermore, when a fraudulent transaction slips through undetected, the onus falls on the customer to raise a dispute. Adhering to Federal Regulation E, the bank is pressed for time to resolve these disputes. The process involves the fraud department liaising with merchants to ascertain the legitimacy of the transaction. If the investigation doesn't favor the customer and they've already utilized the funds, the bank risks a potential loss, especially if the customer opts to close their account.

The rising tide of fraud, coupled with the inefficiencies of the current system, is not just a drain on the bank's resources in terms of employee hours but also poses a threat to customer retention. As dissatisfaction grows, the bank faces the risk of losing valuable deposit accounts, which could have long-term implications for its market position and profitability.

In the rapidly evolving landscape of financial transactions, traditional fraud detection methods are increasingly proving inadequate. Manually programmed algorithms, while effective to a degree, lack the adaptability and scalability required to combat the sophisticated tactics employed by modern fraudsters. Machine learning, with its ability to learn and adapt from vast amounts of data, offers a solution tailored to address the unique challenges posed by today's financial fraud scenarios.

Machine learning models, once trained on historical transaction data, can identify intricate patterns and anomalies that might be overlooked by manual systems. Unlike static algorithms, these models continuously evolve, refining their detection capabilities with each transaction. This dynamic nature drastically reduces false positives, ensuring that genuine transactions aren't mistakenly flagged. Machine learning algorithms can help to reduce strain on customer service channels, improve operational efficiency, and enhance customer satisfaction. Customers no longer need to endure long wait times or unnecessary visits to branches, and bank employees can focus on tasks that add more value.

The predictive capabilities of machine learning can proactively identify potential threats, allowing the bank to implement preventive measures before any fraudulent activity occurs. This not only safeguards the bank's assets but also strengthens its reputation as a secure and customer-centric institution. By transitioning to a machine learning-based fraud detection system, ABC Bank can ensure it remains at the forefront of financial security, offering its customers a seamless and secure banking experience while optimizing its internal operations.

## Data

### *Raw Data Set Description*

The data set we'll be utilizing for this project is sourced from Kaggle, a platform renowned for its contributions to the data science community through competitions and collaborations. This particular dataset is the result of a collaborative effort between the Machine Learning Group of ULB (Université Libre de Bruxelles) and Worldline, a European leader in the payment and transactional services industry.

This specific dataset encapsulates transactions made by European cardholders in September 2013. Within a span of two days, it recorded 492 fraudulent transactions out of a total of 284,807, highlighting the challenge of imbalance with frauds accounting for a mere 0.172% of all transactions. The data is primarily numerical, derived from a PCA transformation. This transformation has been applied to protect sensitive information, and as a result, the features are labeled from V1 to V28. The only exceptions to this transformation are the 'Time' and 'Amount' features. The 'Class' feature is our target variable, indicating whether a transaction is fraudulent or not.

### *Data Collection*

The data set we'll be utilizing for this project is sourced from Kaggle, a platform renowned for its contributions to the data science community through competitions and collaborations. This particular dataset is the result of a collaborative effort between the Machine Learning Group of ULB (Université Libre de Bruxelles) and Worldline, a European leader in the payment and transactional services industry.

For the initial testing and development phase of our machine learning model, data will be sourced from the aforementioned Kaggle dataset. This dataset is provided in a CSV (Comma-Separated Values) file format, which is advantageous for our purposes. The CSV format is not only universally recognized but also seamlessly integrates with the Pandas library in Python, a toolset we'll be leveraging extensively during the model's development and testing stages.

Once our model has been tested and refined using the Kaggle dataset, its real-world application will involve processing transactions from ABC Bank's customer base. The model will be designed to continuously export and learn from ABC Bank's transaction data. This ongoing data collection is pivotal; as fraudsters evolve and devise new techniques, our model must adapt in tandem. By continuously retraining the model with fresh data, we ensure that it remains up to date, enhancing its ability to identify and counteract emerging fraudulent tactics. This dynamic approach not only enhances the model's accuracy but also fortifies ABC Bank's defenses against evolving cyber threats.

### *Data Processing*

The dataset has already undergone a significant degree of preprocessing. A notable advantage is the absence of null values, which often require additional handling or imputation techniques. Furthermore, the majority of the variables in the dataset have been scaled using Principal Component Analysis (PCA) transformation. PCA is a dimensionality reduction technique that ensures variables are on a similar scale, making them more amenable to machine learning algorithms without compromising the variance in the data.

Before feeding the data into our machine learning model, it's essential to understand the relationships between variables. To achieve this, we'll employ a correlation matrix, which provides a visual representation of how variables interact with one another. While most variables have been scaled using PCA, the 'Amount' and 'Time' variables have not. To ensure consistency and optimal performance of our model, these two variables will be scaled using the Robust Scaler tool from the scikit-learn library. This scaler is particularly effective when dealing with outliers, ensuring that extreme values don't unduly influence our model. A significant challenge with our dataset is its imbalance, with fraudulent transactions being a small fraction of the total. To address this, we'll utilize the SMOTE (Synthetic Minority Over-sampling Technique) to oversample the minority class, ensuring our model is trained on a more balanced dataset and can effectively discern between legitimate and fraudulent transactions.

### *Data Management Throughout Development Lifecycle*

Throughout the application's development lifecycle, the data will be stored securely in a version-controlled environment, ensuring that any changes or iterations can be tracked and reverted if

necessary. Regular backups will be scheduled to prevent data loss. Access to this data will be restricted to authorized personnel only, ensuring data integrity and security. As the application transitions from design to development and then to maintenance, the data will be periodically reviewed for relevancy and accuracy, ensuring that the machine learning model remains updated and effective.

#### *Handling Data Anomalies*

As previously highlighted, the dataset we're working with is already devoid of null values, thanks to the preprocessing efforts of the Machine Learning Group of ULB. This eliminates a common challenge in data preparation. When it comes to outliers, our choice of the Robust Scaler is deliberate. This scaler is designed to mitigate the influence of outliers, ensuring that they don't disproportionately affect our model's training. The dataset's imbalance, characterized by a stark difference in the number of legitimate and fraudulent transactions, presents another anomaly. To rectify this, we'll employ the SMOTE technique, which oversamples the underrepresented class. This ensures that our model receives adequate exposure to both types of transactions during training, enhancing its ability to accurately detect fraud in real-world scenarios.

## Project Methodology

### Agile

For the development and deployment of our machine learning-based fraud detection system, we'll be leveraging the Agile methodology. Agile's iterative approach is ideal for projects like ours, where requirements might evolve as we delve deeper into the complexities of fraud detection. Its emphasis on collaboration, adaptability, and continuous feedback ensures that we remain aligned with the project's objectives while being responsive to new insights or challenges.

#### *Concept*

At the onset, stakeholders, product owners, and the IT team will collaboratively define the project's scope, objectives, and priorities. Given the dynamic nature of fraud detection, this phase will ensure that we have a clear roadmap, understanding the challenges and setting the goals for our machine learning model.

#### *Inception*

During this phase, we'll assemble our dedicated team, which will include data scientists, cybersecurity experts, software developers, and IT specialists. The team will delve into the Kaggle dataset, understanding its intricacies, and laying the groundwork for the development of the machine learning model using Random Forest Classification. This phase will also involve setting up the necessary IT infrastructure, ensuring seamless data integration and processing.

#### *Iteration/Development*

This phase will involve a series of sprints. Each sprint will focus on a specific aspect of the model development, from data preprocessing to model training and refinement. The IT team will play a crucial role in ensuring that the development environment is optimized, tools are available, and any technical challenges are addressed promptly. Feedback from each sprint will be used to inform subsequent sprints, ensuring continuous improvement and alignment with IT best practices.

#### *Testing*

Post-development, the model will undergo a comprehensive testing phase. The IT team will ensure that the testing environment replicates real-world conditions. We'll conduct unit tests to validate individual components and integration tests to ensure the entire system works harmoniously. The model's

performance in detecting fraudulent transactions will be critically evaluated, ensuring it meets the set benchmarks.

### *Release*

Once testing affirms the model's effectiveness, we'll move to the release phase. The IT team will be instrumental in deploying the model into the bank's live environment, ensuring seamless integration with existing systems and databases.

### *Review*

After deployment, the focus will shift to monitoring and maintenance. The IT department will continuously monitor the model's performance, ensuring it operates optimally. Feedback loops will be established, allowing for real-time insights and adjustments if needed. Regular reviews will ensure that the model remains updated, catering to evolving fraud patterns and techniques.

## SEMMA

In addition to the Agile methodology, the SEMMA (Sample, Explore, Modify, Model, Assess) methodology will be employed to guide the development and evaluation of the machine learning model. SEMMA provides a structured approach to data mining, ensuring that each step in the process is methodically executed.

### *Sample*

The foundation of our project will be a dataset sourced from a single database, encompassing over 200,000 credit card transactions conducted over a span of two days. This dataset will offer a comprehensive view of transactional behaviors, both legitimate and potentially fraudulent.

### *Explore*

Upon acquiring the dataset, we will utilize Python libraries to delve into its intricacies. This exploration phase will involve visualizing the data, identifying patterns, and discerning relationships between various data attributes.

### *Modify*

Recognizing that raw data often requires refinement, the modify stage will address any imbalances or irregularities in the dataset. Specifically, we will employ the SMOTE technique to balance the data and the Robust Scaler tool from the scikit-learn library to scale the necessary variables, ensuring that the data is primed for accurate modeling.

### *Model*

With the data prepared, the modeling phase will commence. The chosen algorithm for this task is the Random Forest Classifier, renowned for its efficacy in classification tasks. This algorithm will be trained on the dataset, learning from the patterns and relationships identified in the exploration phase.

### *Assess*

Post-modeling, the assessment phase will evaluate the model's performance. Initially, the model's predictions on the test data will be gauged using metrics like recall, accuracy, and precision. Once the model is deployed in a real-world scenario, ABC Bank's fraud department will play a crucial role in this phase, reviewing the model's classifications of fraudulent transactions and providing feedback on its accuracy.

# Project Outcomes

### *Finished Application*

The primary deliverable of this project will be a fully functional machine learning application designed to detect fraudulent transactions. This application will utilize the Random Forest Classifier algorithm. Users will be able to input data via a CSV file, and the application will subsequently output prediction results, indicating potentially fraudulent activities.

Beyond the core prediction functionality, the application will be essential for processing the data, ensuring it's appropriately scaled using the Robust Scaler from the scikit-learn library, and balanced using the SMOTE technique. While the application will be developed using the Kaggle dataset, it will be retrained on ABC Bank's transactional data, ensuring the application remains relevant and effective against evolving fraud patterns.

### *User Guide*

Accompanying the application will be a comprehensive user guide. This guide will detail the step-by-step process of using the application, supported by screenshots. It will outline the system requirements, ensuring that users have the necessary environment to run the application seamlessly.

# Implementation Plan

## General Strategy

Our primary objective is to seamlessly integrate a machine learning model that can proficiently detect fraudulent transactions into ABC Bank's existing systems. To achieve this, we've devised a comprehensive strategy that leverages both the Agile and SEMMA methodologies, ensuring continuous feedback and iterative refinement throughout the project's lifecycle.

## Rollout Phases

The initial phase will focus on data acquisition and preprocessing. We'll source our dataset from Kaggle and embark on an exploratory data analysis to understand its structure and inherent characteristics. This will be followed by the application of various preprocessing techniques, including scaling and balancing, to prepare the data for modeling.

Once our data is primed, we'll transition into the model development phase. Here, the Random Forest Classifier will be trained on the processed dataset. A subset of the data will be reserved to evaluate the model's initial performance, ensuring we're on the right track.

As we move forward, integration with ABC Bank's systems will be paramount. We'll work closely with the bank's IT department to develop interfaces or APIs, allowing our model to access real-time transaction data. It's crucial that the model's outputs, specifically the flags for fraudulent transactions, are easily accessible and actionable for the bank's fraud department.

Testing and refinement will be an ongoing process. We'll conduct rigorous testing in a sandbox environment, replicating the bank's real-world systems. This will allow us to ascertain the model's accuracy and its integration capabilities. Feedback from these tests will be invaluable, guiding our refinements to optimize performance.

Upon satisfactory testing, we'll proceed to the deployment phase. The model will first be introduced to a limited set of real-world transactions, serving as a litmus test for its performance in a live environment. As we gain confidence in its reliability, we'll gradually increase the volume of transactions it analyzes.

Post-deployment, our work doesn't end. We'll be continuously monitoring the model's performance, ensuring it remains effective against evolving fraudulent tactics. Periodic retraining sessions with new

data will be scheduled, ensuring our model remains updated and adept at identifying new fraud patterns.

## Evaluation Plan

Due to the iterative nature of the Agile methodology, each stage of the development will be evaluated on usability, accuracy, recall, and precision. Every sprint will be evaluated using this criterion. Upon completion of the project, these metrics will be used to measure the effectiveness of the machine learning model and user interface.

| Objective | Success Criteria |
|---|---|
| Usability | The program will be utilized by the fraud department and must be simple to use by non-technical users. User surveys will be conducted to measure usability. |
| Accuracy | Accuracy is measured by the number of true positive and true negative transactions divided by the total transactions. The algorithm must score higher than a value of 80%. |
| Recall | Recall is the number of correctly identified fraud transactions divided by the total number of transactions classified as fraud. The algorithm must score higher than a value of 70%. |
| Precision | Precision is the number of correctly identified fraud transactions divided by the number of correctly identified frauds plus the incorrectly identified normal transactions. The algorithm must score higher than a value of 70%. |

## Resources and Costs

The budget allocation for the project is divided into three primary categories:

### *Hardware*

An investment in state-of-the-art cloud solutions and high-volume servers. This ensures our system not only operates at peak performance but also ensures data integrity through cloud storage backups.

### *Software*

This portion is dedicated to procuring essential software licenses and tools for security and collaboration during the development process.

### *Labor*

A significant allocation for our team of professionals, which includes data scientists, cybersecurity experts, and seasoned developers. Their combined expertise and dedication are pivotal, propelling the project towards its envisioned success.

| | |
|---|---|
| **Hardware** | **66,000.00** |
| Server | 50,000.00 |
| Cloud Back up | 10,000.00 |
| Computers for Development | 6,000.00 |
| **Software** | **7,000.00** |

| | |
|---|---|
| Version Control | 300.00 |
| User Control | 3,500.00 |
| Firewall and Encryption | 3,200.00 |
| **Labor Time and Cost (including maintenance)** | **41,000.00** |
| Software / Data Scientist Development Team | 30,000.00 |
| Cybersecurity Team | 11,000.00 |
| **Grand Total** | **114,000.00** |

## Timeline and Milestones

| | Start | End | Tasks |
|---|---|---|---|
| 1 | 11/01/23 | 11/30/23 | Proposal evaluated and accepted |
| 2 | 12/01/23 | 12/22/23 | Hardware team: Gather hardware |
| 3 | 12/01/23 | 2/29/24 | Software team: Development |
| 4 | 3/01/24 | 3/31/24 | Testing |
| 5 | 4/01/24 | 4/30/24 | Revisions |
| 6 | 5/01/24 | 06/30/24 | Project evaluation |

# Part C: Application

The application can be found at:

https://colab.research.google.com/drive/1GW3Efc_JBUA_5y5aVuBghhB6rgE-sPsC?usp=sharing

Example data:

https://media.githubusercontent.com/media/mkkakau/cc-fraud/main/example.csv

# Part D: Post-implementation Report

## Business Vision

### *Problem Description*

ABC Bank faces an increasing challenge with credit card fraud. While traditional, rule-based fraud detection systems are in place, they often result in high false positives, straining both bank resources and inconveniencing customers.

### *Application Solution*

To address this challenge, we developed a machine learning application using the Random Forest Classification algorithm. This application is designed to analyze transactional data and accurately identify potential fraudulent activities. Unlike the previous systems, this model can adapt and learn from historical data, making it more effective against evolving fraud techniques. The application accepts data in CSV format, seamlessly integrating with the bank's existing data infrastructure.

### *User Interaction*

Users input transaction data in CSV format into the application. Upon processing, the system analyzes each transaction against learned patterns and flags potential fraudulent activities. The result is a list of flagged transactions for further investigation. The interface is designed for ease of use, ensuring that even those unfamiliar with machine learning can operate it efficiently. This streamlined approach significantly reduces false positives and enhances the bank's fraud detection capabilities.

## Datasets

### *Raw and Processed Data Description*

The raw data sourced from the Machine Learning Group of ULB in collaboration with Worldline is a comprehensive dataset detailing credit card transactions made in September 2013 by European cardholders. This dataset, available in CSV format, presents transactions spanning two days, with 492 frauds out of 284,807 transactions. The data is primarily numerical, derived from a PCA transformation, with the exception of 'Time' and 'Amount'. The 'Class' feature indicates whether a transaction is fraudulent (1) or legitimate (0).

The processed data, on the other hand, was a result of several preprocessing steps, including scaling of the 'Time' and 'Amount' features and addressing the data imbalance using the SMOTE technique. This ensured that the data was in an optimal format for the machine learning model to train and test effectively.

### *Data Processing*

The raw data underwent minimal preprocessing, given its already transformed nature. However, to make it accessible to our algorithm, the data was split into two halves. The first half was used for both training and testing the machine learning model, ensuring its accuracy and reliability. The second half served as a demonstrative set, used to showcase the application's capabilities to the fraud team.

The 'Time' and 'Amount' features, which were not transformed using PCA, were scaled using the Robust Scaler tool from the scikit-learn library. This ensured uniformity across all features, making the data more conducive for the Random Forest Classification algorithm.

*Example of Raw and Processed Data*

Raw Data

```
Class Counts
====================
Non-Fraudulent: 142158
Fraudulent: 246
```

| | Time | V1 | V2 | V3 | ... | V27 | V28 | Amount | Class |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 41505 | -16.526507 | 8.584972 | -18.649853 | ... | -2.018575 | -1.042804 | 364.19 | 1 |
| 1 | 44261 | 0.339812 | -2.743745 | -0.134070 | ... | 0.040996 | 0.102038 | 520.12 | 0 |
| 2 | 35484 | 1.399590 | -0.590701 | 0.168619 | ... | 0.011409 | 0.004634 | 31.00 | 0 |
| 3 | 167123 | -0.432071 | 1.647895 | -1.669361 | ... | -0.237386 | 0.001934 | 1.50 | 0 |
| 4 | 168473 | 2.014160 | -0.137394 | -1.015839 | ... | -0.078043 | -0.070571 | 0.89 | 0 |

Scaled Data

```
Class Counts
====================
Non-Fraudulent: 142158
Fraudulent: 246
```

| | V1 | V2 | V3 | V4 | ... | V28 | Class | new_time | new_amount |
|---|---|---|---|---|---|---|---|---|---|
| 0 | -16.526507 | 8.584972 | -18.649853 | 9.505594 | ... | -1.042804 | 1 | -0.510887 | 4.799299 |
| 1 | 0.339812 | -2.743745 | -0.134070 | -1.385729 | ... | 0.102038 | 0 | -0.478506 | 6.986255 |
| 2 | 1.399590 | -0.590701 | 0.168619 | -1.029950 | ... | 0.004634 | 0 | -0.581628 | 0.126227 |
| 3 | -0.432071 | 1.647895 | -1.669361 | -0.349504 | ... | 0.001934 | 0 | 0.965014 | -0.287518 |
| 4 | 2.014160 | -0.137394 | -1.015839 | 0.327269 | ... | -0.070571 | 0 | 0.980875 | -0.296073 |

Split Data

```
Train Class Counts
====================
Non-Fraudulent: 113730
Fraudulent: 193

Test Class Counts
====================
Non-Fraudulent: 28428
Fraudulent: 53
```

Final Oversampled Data

```
Train Class Counts
======================
Non-Fraudulent: 113730
Fraudulent: 113730
```

| | V1 | V2 | V3 | V4 | ... | V27 | V28 | new_time | new_amount |
|---|---|---|---|---|---|---|---|---|---|
| 0 | -0.925069 | 0.133531 | 0.868665 | -2.026892 | ... | -0.075717 | 0.018547 | 0.482830 | 0.322581 |
| 1 | 1.988999 | 0.006200 | -1.873828 | 0.560928 | ... | -0.017739 | -0.026335 | 0.542328 | 0.298036 |
| 2 | -0.395544 | -0.390356 | 2.324042 | -2.002338 | ... | 0.220649 | -0.016454 | 0.103169 | -0.168303 |
| 3 | 1.890122 | -0.596596 | 0.119191 | 0.399677 | ... | -0.047943 | -0.053818 | 0.186506 | 0.238429 |
| 4 | -0.406713 | -1.301137 | 1.087018 | -2.979571 | ... | -0.052586 | -0.013761 | 0.599041 | 3.282609 |

### *Access to Datasets*

The dataset used for this project can be accessed directly from Kaggle at the following URL: https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud/

## Data Product Code

### *Processing Raw Data*

The raw data provided had already undergone significant preprocessing using a PCA transformation and null values were removed from the dataset. The PCA transformation not only scaled the data but also ensured the privacy of the customers, a crucial aspect given the sensitive nature of financial transactions. However, two variables from the raw data required additional scaling to ensure consistency and improve the model's performance. To address this, the Robust Scaler from the scikit-learn library was employed, given its efficacy in handling outliers. Recognizing the class imbalance inherent in the data, the SMOTE technique was subsequently applied to oversample the minority class, ensuring a more balanced dataset for effective model training.

### *Descriptive Methods and Visualizations*

Descriptive statistics and visualizations were employed to understand the inherent characteristics of the data.

- A pie chart was used as a primary visualization tool to illustrate the significant imbalance present in the data. This visualization highlighted a critical challenge: the overwhelming majority of transactions were non-fraudulent. Such an imbalance, if unaddressed, could lead the machine learning model to assume that most transactions are legitimate, thereby compromising its ability to detect fraudulent activities.
- A correlation matrix was utilized to understand the relationships between the various features in the dataset. This matrix revealed that while the V1-V28 features, resulting from the PCA transformation, largely remained uncorrelated with each other, the 'Time' and 'Amount' variables did exhibit correlations with some of these V-features. Recognizing these correlations was essential, as it provided insights into potential patterns and dependencies within the data.

- A confusion matrix was later used post-model training to visualize the performance of our machine learning model, showcasing true positives, false positives, true negatives, and false negatives.

### *Non-Descriptive Methods*

The primary analytic method applied was the Random Forest Classifier, a machine learning algorithm known for its ability to handle large datasets with higher dimensionality. The choice of the Random Forest Classifier was deemed appropriate for this project due to its inherent ability to handle imbalanced datasets. Its ensemble nature, which builds multiple decision trees during training and outputs the mode of the classes for classification, makes it particularly suited for fraud detection where false negatives (fraudulent transactions classified as legitimate) can have severe implications.

The model was trained using the oversampled dataset to ensure it learned from an equal representation of both fraudulent and legitimate transactions. It was then tested on a separate dataset to evaluate its performance. The key metrics used for this evaluation were accuracy, precision, and recall. Accuracy provided a general measure of how often the model was correct. Precision gave insights into how many of the transactions flagged as fraudulent were actually fraudulent, and recall indicated how many of the actual fraudulent transactions the model was able to catch.

The data analysis, especially the insights from the descriptive methods like the pie chart illustrating data imbalance and the correlation matrix, played a pivotal role in choosing and refining the Random Forest Classifier. The understanding of the data's nature and its intricacies, as revealed by the descriptive methods, ensured that the non-descriptive method was tailored to address the specific challenges posed by the dataset, leading to a more effective fraud detection system.

## Objective Verification

### *Project's Objective*

The primary objective of our project was to develop a machine learning system capable of detecting fraudulent transactions with a high degree of accuracy. While the initial aim was to achieve an accuracy of 80%, we recognized that accuracy alone wouldn't provide a comprehensive assessment due to the unique nature of our data. Consequently, we shifted our focus to two pivotal metrics: precision and recall, both with a target of 70%. The rationale behind this was twofold:

- Precision: To ensure that when our system identifies a transaction as suspicious, it's correct about 70% of the time.
- Recall: To guarantee that our system detects 70% of all genuine fraudulent transactions.

This approach was designed to strike an optimal balance, ensuring both the security of our customers' finances and their seamless access to their accounts.

### *Objective Achievement*

Upon evaluating the model, we found that it surpassed our expectations in terms of accuracy, achieving a score of 0.999. Furthermore, the model attained a precision of 0.88 and a recall of 0.75. While the precision exceeded our target, the recall was slightly above the set benchmark.

In essence, the system was highly accurate in flagging genuine fraudulent transactions, ensuring minimal false alarms. Moreover, it was adept at identifying a significant majority of actual fraudulent activities, thereby safeguarding customer assets.

In conclusion, the objective was largely met. The slight deviation in recall underscores the ever-evolving nature of fraudulent techniques and emphasizes the need for continuous model training and adaptation.

However, the results are promising and indicate a substantial improvement over traditional fraud detection methods.

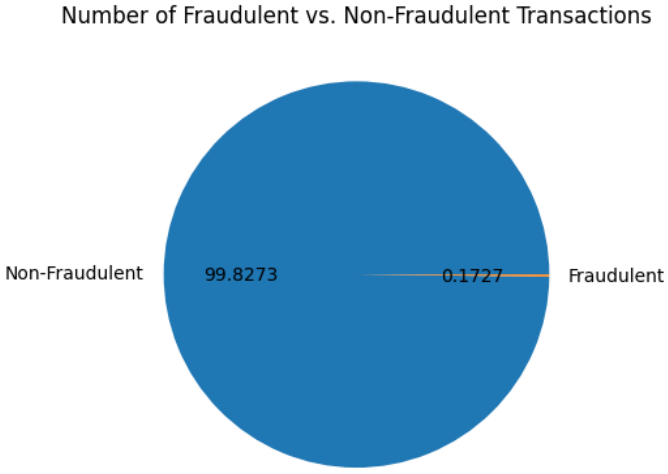
## Effective Visualization and Reporting

### *Data Exploration*

Before diving into the development of the machine learning model, it's imperative to understand the nature and structure of the data. This involves getting a sense of the distribution, potential correlations, and anomalies within the dataset. Descriptive methods, such as visualizations, play a pivotal role in this phase, offering a clear, visual representation of the data's characteristics.

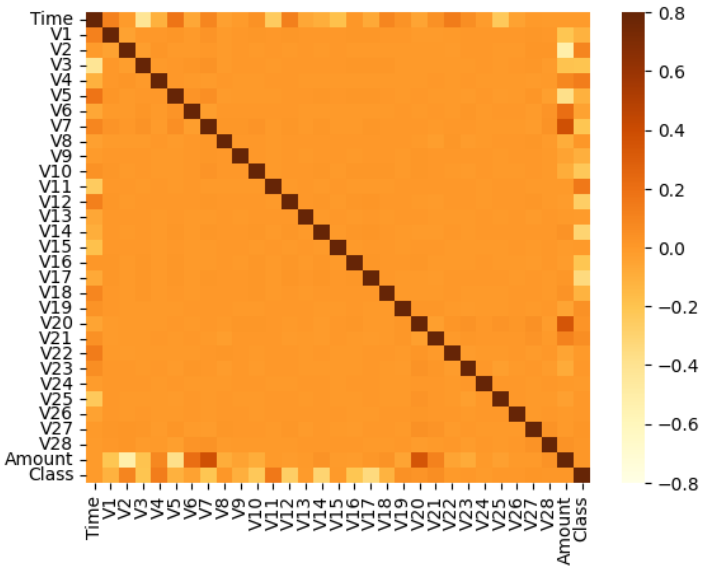
### *Data Analysis*

#### Pie Chart




The first step was to understand the distribution of our target variable. A pie chart was employed to illustrate the stark imbalance in the dataset. This visualization underscored the challenge at hand, emphasizing the need for techniques like SMOTE to address the data imbalance before model training.

## Correlation Matrix

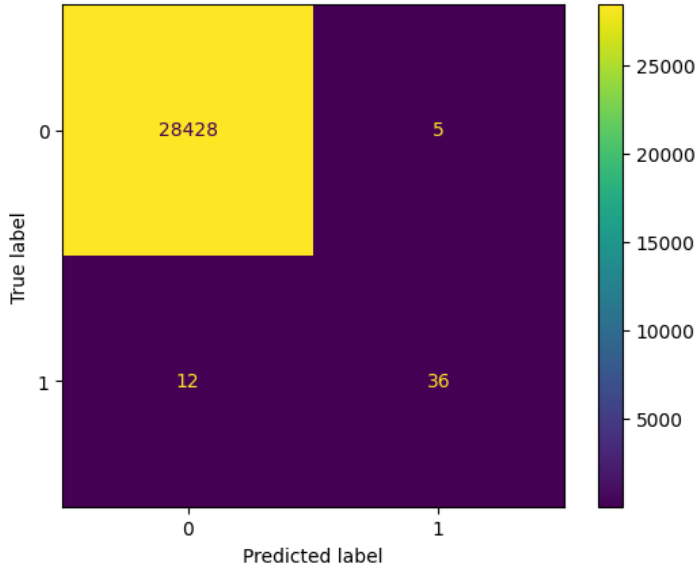



To understand the relationships between different features, a correlation matrix was utilized. Key observations from the matrix included:

- The V1 - V28 features, resulting from PCA transformation, showed no inter-correlation, which is expected from principal components.
- The 'Amount' feature exhibited inverse correlations with V2 and V5 and direct correlations with V7 and V20.
- The 'Time' feature was inversely correlated with V3.

These insights helped in understanding potential feature interactions and their relevance in predicting fraudulent transactions.

## Confusion Matrix




Post-model training, it's crucial to evaluate its performance on unseen data. A confusion matrix was used to visually represent the model's predictions against actual values. The matrix revealed:

This visualization provided a clear picture of the model's precision and recall, crucial metrics for our project's success.

### *Data Summary*

Through these visualizations, we were able to succinctly summarize the data's characteristics, its challenges (like imbalance), and the model's performance. The pie chart highlighted the need for data augmentation, the correlation matrix informed potential feature engineering, and the confusion matrix provided a clear assessment of the model's efficacy.

### *Analysis Application of Visualizations*

The visualizations served as foundational pillars throughout the project. They not only informed the development process but also facilitated communication with stakeholders, ensuring that both technical and non-technical team members were aligned in understanding the data's nuances and the model's performance.

## Accuracy Analysis

To assess the efficacy of our model, we didn't solely rely on the traditional accuracy metric, given the imbalanced nature of our dataset. Instead, we emphasized two pivotal measures: precision and recall. Precision measures the correctness of our model when it flags a transaction as suspicious, ensuring that false alarms are minimized. Recall, on the other hand, gauges the model's capability to identify and catch actual fraudulent transactions. Our objective was to achieve a precision and recall rate of 70%, ensuring a balanced approach that both protects our customers' finances and guarantees seamless access to their accounts.

### *Random Forest Classifier*

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 1.00 | 1.00 | 1.00 | 28433 |
| 1 | 0.88 | 0.75 | 0.81 | 48 |
| accuracy | | | 1.00 | 28481 |
| macro avg | 0.94 | 0.87 | 0.90 | 28481 |
| weighted avg | 1.00 | 1.00 | 1.00 | 28481 |

accuracy score: 0.999403110845827
average precision score: 0.6589579188864464

These scores indicate that our model not only correctly classifies transactions with a 99.9% accuracy rate but, more importantly, when it predicts a transaction as fraudulent, it's correct 88% of the time. Furthermore, it successfully identifies 75% of all actual fraudulent transactions, meeting our targeted recall rate. From the provided results, one can observe the average precision score, which is a harmonized measure of precision and recall. This score gives a comprehensive view of the model's performance across different thresholds, especially vital for imbalanced datasets like ours.

## Application Testing

Multiple machine learning algorithms were employed to ascertain which one delivered the best results. The primary metric we focused on was the average precision score, as it provides a balance between precision and recall, especially crucial given the imbalanced nature of our dataset.

The Random Forest Classifier clearly outperformed the other algorithms with a significantly higher average precision score. As a result, it was chosen as the primary algorithm for the application. The other algorithms, despite their lower scores, provided valuable insights into the nature of the data and the challenges associated with it.

*K-Nearest Neighbors Results*

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 1.00 | 1.00 | 1.00 | 28438 |
| 1 | 0.31 | 0.84 | 0.46 | 43 |
| accuracy | | | 1.00 | 28481 |
| macro avg | 0.66 | 0.92 | 0.73 | 28481 |
| weighted avg | 1.00 | 1.00 | 1.00 | 28481 |

accuracy score: 0.9969804431024192
average precision score: 0.26232868991936836

*Logistic Regression Results*

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 1.00 | 0.97 | 0.98 | 28432 |
| 1 | 0.05 | 0.96 | 0.09 | 49 |
| accuracy | | | 0.97 | 28481 |
| macro avg | 0.52 | 0.96 | 0.54 | 28481 |
| weighted avg | 1.00 | 0.97 | 0.98 | 28481 |

accuracy score: 0.967627541167796
average precision score: 0.046690318068386 84

*Naïve Bayes Results*

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 1.00 | 0.97 | 0.99 | 28441 |
| 1 | 0.04 | 0.85 | 0.08 | 40 |
| accuracy | | | 0.97 | 28481 |
| macro avg | 0.52 | 0.91 | 0.53 | 28481 |
| weighted avg | 1.00 | 0.97 | 0.99 | 28481 |

accuracy score: 0.9733155436957972
average precision score: 0.036885793663849635

*Random Forest Classifier*

```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     28433
           1       0.88      0.75      0.81        48

    accuracy                           1.00     28481
   macro avg       0.94      0.87      0.90     28481
weighted avg       1.00      1.00      1.00     28481

accuracy score:  0.999403110845827
average precision score:  0.6589579188864464
```

## Application Files

The application was developed in Google Colaboratory which allows you to run Python code with zero configuration required. This notebook contains both the code and the markdown cells that explain each step of the process. All required libraries are imported at the beginning of the notebook for clarity and ease of understanding.

*Python Libraries Used*

numpy: This library is fundamental for numerical operations and handling arrays.

pandas: Used for data manipulation and analysis, especially for structured data operations.

matplotlib: A plotting library that provides a MATLAB-like interface for creating visualizations.

sklearn (scikit-learn): This is a machine learning library that provides simple and efficient tools for data analysis and modeling.

seaborn: A statistical data visualization library based on matplotlib, offering a higher-level interface for drawing attractive and informative statistical graphics.

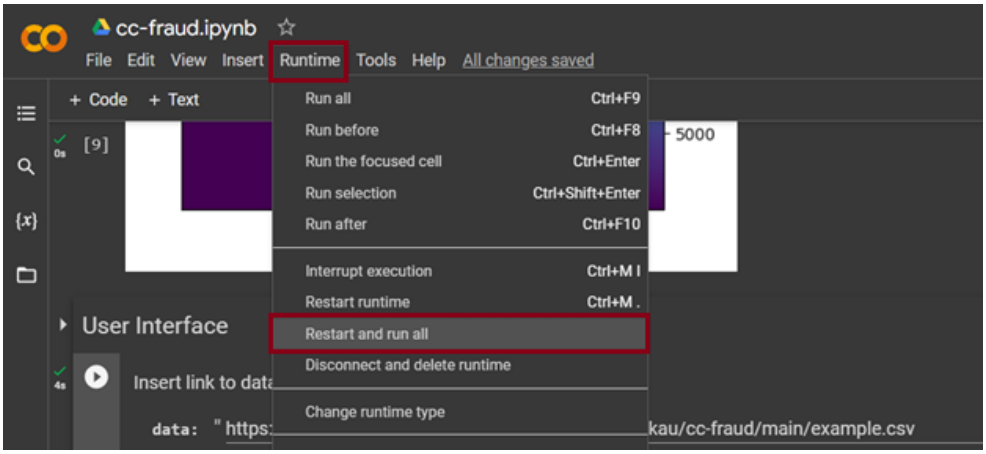
imblearn: A library that provides tools when dealing with classification with imbalanced classes, especially useful for the SMOTE technique in our application.

*Files Used*

| File | URL |
|---|---|
| **cc-fraud.ipynb** | https://colab.research.google.com/drive/1GW3Efc_JBUA_5y5aVuBghhB6rgE-sPsC?usp=sharing |
| **example.csv** | https://media.githubusercontent.com/media/mkkakau/cc-fraud/main/example.csv |

## User Guide

1. Access the application at the following url: https://colab.research.google.com/drive/1GW3Efc_JBUA_5y5aVuBghhB6rgE-sPsC?usp=sharing
2. Execute the application from the menu:
3. Runtime > Restart and run all.




4. Wait until all code segments are executed.
5. Scroll to the bottom of the notebook to the User Interface section and enter the URL to the example csv file. Press the run cell button.

   Example CSV file URL: https://media.githubusercontent.com/media/mkkakau/cc-fraud/main/example.csv


6. Fraudulent transaction totals and a preview of the transactions will be displayed under the Results.

## Summation of Learning Experience

My academic background in computer science laid a solid foundation for understanding computational problems and their solutions. However, the realm of machine learning, especially its application in fraud detection, was relatively new to me at the outset of this project. While I had a general understanding of algorithms and data structures, diving deep into machine learning algorithms was both challenging and enlightening.

I quickly realized the depth and breadth of knowledge required to effectively implement and optimize these algorithms. To bridge the gap in my understanding, I delved into various resources, from research papers to online courses, to grasp the intricacies of techniques like the Random Forest Classifier and the SMOTE method for handling imbalanced datasets. The process of learning, experimenting, and iterating was rigorous but immensely rewarding.

This project underscored the importance of continuous learning in the tech industry. While my foundational knowledge was crucial, the ability to adapt and acquire new skills on-the-fly proved equally vital. The journey through this project has reinforced my belief in the value of lifelong learning. It's clear to me that to stay relevant and effective in the ever-evolving world of technology, one must maintain a persistent curiosity and a willingness to learn and adapt.

# References

Ritchie, J. N. & A., & Staff in the Bureau of Competition & Office of Technology. (2021, August 24). Consumer Sentinel Network Data Book 2020. Federal Trade Commission. https://www.ftc.gov/reports/consumer-sentinel-network-data-book-2020