# **Business Case: Implementing an ML-Based Fraud Detection System**

**Contents**

[**Business Case: Implementing an ML-Based Fraud Detection System** 1](#_Toc150515837)

[**Executive Summary** 1](#_Toc150515838)

[**Objectives** 1](#_Toc150515839)

[**Context** 1](#_Toc150515840)

[**Proposed Solution** 3](#_Toc150515841)

[**Literature Review** 4](#_Toc150515842)

[**Benefits** 5](#_Toc150515843)

[**ROI and Cost-Benefit Analysis:** 5](#_Toc150515844)

[**Risks and Mitigations** 6](#_Toc150515845)

[**Accuracy** 6](#_Toc150515846)

[**Conclusion** 7](#_Toc150515847)

[**Appendix** 7](#_Toc150515848)

[**References** 7](#_Toc150515849)

## **Executive Summary**

The financial landscape is evolving rapidly, with the increasing prevalence of online transactions, making fraud detection a critical concern for our institution. This business case proposes the implementation of a state-of-the-art machine learning (ML) based fraud detection system to enhance our ability to identify and mitigate fraudulent activities. Funbucks Financial Services is seeking bids from consulting organizations in streamlining their legacy operations surrounding fraud.

## **Context**

The global rise in online fraud, coupled with the dynamic nature of fraudulent activities, necessitates an initiative-taking and adaptive approach to fraud detection. Our institution seeks to leverage advanced technologies to stay ahead of emerging threats and provide a secure financial environment for our customers.

## **Client Statement**

As a leading technology institution with specialization in fintech services we seek to provide a bid to Funbucks Financial Services on assisting in streamlining fraud detection services for their clients based on the following objectives

.

## **Objectives**

1. **Improved Fraud Detection:**
   * Utilize ML algorithms to enhance the accuracy and efficiency of fraud detection in real-time transactions.
2. **Customer Experience:**
   * Minimize false positives to enhance customer satisfaction by reducing disruptions to legitimate transactions.
3. **Financial Security:**
   * Mitigate financial losses associated with undetected fraudulent transactions.

## **Literature Review**

**Fraud uses in AI:**

**Credit Card Fraud Detection using Machine Learning Algorithms:**

The journal article discusses the use of machine learning algorithms for detecting credit card fraud. Online transactions are becoming increasingly susceptible to fraudulent activities, partly due to the growing ubiquity of e-commerce. In response to these threats, different machine learning methods have been used to analyze and detect fraud in online transactions. The authors develop a novel method for fraud detection that groups customers based on their transactions and then applies different classifiers, such as decision trees and logistic regression. This method is then tested on a European credit card dataset, showing the effectiveness of the proposed system in detecting fraudulent behavior. Additionally, the paper addresses the problem of 'concept drift', employing a feedback mechanism to deal with dynamic changes in customer behavior. The paper concludes that logistic regression, decision tree and random forest methods offer the best results for fraud detection. (Dornadulaa & S, 2019)

**How is AI transforming fraud detection in banks?:**

Credit card fraud is a growing problem with a report from TransUnion found that globally, online fraud attempt rates for financial services rose 149% between Q4 of 2020 and Q1 of 2021 alone. Due to this total fraud losses have climbed to $56 billion in 2020, a whopping 83% gain from a year earlier. Due to the growing trends in technology and this increased financial burden banks are looking more towards AI with 64% believing it can help prevent fraud before it happens. (Telus International, 2022)

**American banks dominate AI transformation race:**

While the United States in leading the world in banking related AI research and transforming banking into an AI-First industry, most of that comes from the largest banks in the US like J.P Morgan Chase, with many small banks having no such skill set relying entirely on customer service to flag fraudulent transactions (Heath, 2023).

**Role of Artificial Intelligence in Financial Fraud Detection:**

The article underscores the increasing importance of Artificial Intelligence (AI) in the banking sector for real-time fraud prevention and detection. It notes the substantial growth of AI-driven solutions, predicting a 300-billion-dollar impact by year 2030. Despite the costs, studies suggest a favorable cost-benefit equation, emphasizing both financial and non-financial returns. However, the article raises concerns about potential financial fraud amid the government's push for inclusion and digitalization, particularly in the absence of basic financial literacy. It advocates for a careful balance between embracing technology and safeguarding stakeholders' interests. A highlight of the article is the need for a blend of AI capabilities with human intervention working for a better customer experience. The article concludes by calling for further studies to assess the economic effects of AI implementation across business for evidence-based analysis. (Mishra, Mishra, & Mohanty, 2023)

**The Role Artificial Intelligence in Modern Banking:**

The article explores the transformative impact of Artificial Intelligence (AI) on banking fraud prevention and risk management. AI enables a shift from reactive to proactive measures, providing real-time fraud detection by analyzing vast data streams. Traditional static methods are replaced by AI's ability to recognize complex patterns, particularly through deep learning and neural networks. AI enhances 'Know Your Customer' processes through Natural Language Processing and graph analytics, ensuring rigorous customer verification. In credit scoring, AI considers diverse parameters, offering a more holistic risk profile. The integration of AI-powered chatbots facilitates immediate reporting of suspicious activities. Biometric verification, when coupled with AI, enhances accuracy and adaptability, preventing unauthorized access. AI-driven geospatial analysis, behavioral biometrics, and self-learning mechanisms further contribute to dynamic and holistic fraud detection. The article emphasizes AI's role in omni-channel analysis, stress testing, and ensuring compliance with evolving regulatory standards in the banking sector. Finally, as the Internet of Things (IoT) becomes more integrated with banking, AI continues to play a crucial role in monitoring and ensuring the security of interactions. (Aziz & Andriansyah, 2023)

**How AI is Used in Fraud Detection:**

This article discusses the benefits and risks of using AI in Fraud Detection with the Benefits being firstly, AI processes incoming data swiftly, blocking new threats in milliseconds, ensuring dynamic and fast security. Secondly, the more data AI receives, the better its predictions become. It continuously improves over time, especially when instances share knowledge globally. Lastly, AI reduces the need for reactive measures by swiftly identifying threats. This enables employees to spend less time on investigating threats and reviewing information, allowing them more time for projects that drive business growth. Additionally risks of using AI in fraud detection are explained primarily by Automated threats are not the only concern; social fraud, such as phishing and social engineering, remains a challenge. These non-automated threats are difficult for AI to combat, requiring ongoing employee education to mitigate risks. Secondarily, the extensive data processing by AI, especially when combined with machine learning and neural networks, can make it challenging to understand its inner workings. Despite this "black box" aspect, top fraud detection software allows customization of rules for better control. Finally, While AI minimizes false positives, completely eliminating them is impossible. Occasionally, genuine users, especially those employing uncommon browsers and VPNs, may be mistakenly blocked by AI. (How AI is Used in Fraud Detection – Benefits & Risks, 2022)

**How is artificial intelligence used in fraud detection?**

The role of Artificial Intelligence (AI) in fraud detection, cybersecurity, and crime prevention Is emphasizes by the global economic impact of fraud and cyber breaches. The projecting digital fraud losses exceed $343 billion between 2023 and 2027. The benefits of AI in fraud detection include enhanced accuracy, real-time monitoring, reduced false positives, increased efficiency, and cost reduction. However, potential risks of AI in fraud detection are acknowledged in this article, such as biased algorithms, negative results, and lack of transparency. Explainable AI is suggested as a solution to mitigate these risks by providing interpretable explanations for AI decisions. (Bassi, 2023)

**Machine Learning Models, Scaling and Performance:**

**Effects of dataset size and interactions on the prediction performance of logistic regression and deep learning models:**

The authors discuss how in their medical test data they modeled performance at 3 thresholds: 1,000, 10,000, and 100,000 observations. Overall performance did not improve with a penalization technique in place, whereas non penalized linear regression models saw diminishing returns .76, .79, and .80 respectively for accuracy. Across all models more data did improve AUC (Bailly, et al., 2022).

**How Much Data Is Required for Machine Learning?**

Here the author explains a few key pieces of information that is necessary to determine how much data a model needs and the general effects of model performance with data. The main factors are the complexity of the model and learning algorithm, the number of needed labels and the margin of error that is considered acceptable. While this article recommends generally following the 10x rule, Dinesh R. Pai, Ph.D. says 20x (Pai, 2022). There are methods to deal with lack of data, such as synthetic data and data augmentation (Dorfman, 2022).

However, an argument for limiting data comes from the Harvard Business Review. There they state that the phenomenon regarding diminishing returns in regard to the learning curve is justified theoretically too. Consider this, the performance of a machine learning model depends on two kinds of error - Bias error and Variance error. The bias error decreases with more data, but the variance error increases. This happens because, for the same quantity of data, the model can learn a certain pattern accurately (thus reducing bias) but cannot perfectly predict each data point (leading to more variance).

The reason why variance increases because the model starts learning the noise along with the signal in the data. The model tries to catch every data point so minutely that it ends up learning the error/noise in the data too. And this error/noise keeps increasing as we add more data. So, after a certain point, the increase in variance error outpaces the decrease in bias (McAfee, 2019).

**Class Imbalance and Evaluation Metrics:**

**SMOTE-WENN: Solving class imbalance and small sample problems by oversampling and distance scaling**

The article discusses the challenges of imbalanced data classification in practical applications. It highlights that imbalanced datasets, where one class is significantly smaller than the other, can lead to a deterioration in the performance of traditional classification methods, particularly affecting the recognition rate of the minority class. The problem of imbalanced classification is attributed to two main factors: inappropriate optimization metrics in traditional learning algorithms, which favor the majority class, and the imbalanced data itself, characterized by an imbalance ratio (IR) that measures the degree of imbalance. “Research shows that a larger IR leads to worse performance. The complexity of distribution characteristics in imbalanced datasets includes small disjuncts, overlapping between classes, rare cases, and outliers in the minority class space, making the correct identification of examples, especially in the minority class, challenging.” (Guan, Zhang, Xian, Cheng, & Tang, 2021) The classification of small-sample imbalanced datasets is even more difficult, with fewer training examples leading to poorer performance. To address these challenges, both algorithm-level and data-level methods have been proposed. Algorithm-level methods modify traditional classification algorithms to handle imbalanced datasets by assigning different misclassification attributes to the minority and majority classes. Data-level methods, like resampling, are used to modify the dataset itself. Resampling techniques include undersampling, oversampling, and hybrid sampling, each with its own advantages and applications. In the context of small-sample imbalanced datasets, oversampling and hybrid sampling methods, particularly those that combine synthetic minority oversampling **(SMOTE)** with data cleaning, have shown promising results. The authors introduce the weighted edited nearest neighbor rule **(WENN)** to perform data cleaning on balanced datasets after SMOTE, preserving positive examples in overlapping areas. A new hybrid resampling method, SMOTE-WENN, is proposed to address the challenges of small-sample imbalanced data classification. This method keeps a high percentage of safe positive and negative examples, leading to improved classification performance. (Guan, Zhang, Xian, Cheng, & Tang, 2021)

**Effective Class-Imbalance learning based on SMOTE and Convolutional Neural Networks**

This article explores the challenge of imbalanced data (ID) in machine learning models, where one class significantly outnumbers the other, causing bias in model training. Various solutions, including synthetic data generation and data reduction, have been proposed to tackle this issue, and deep neural networks (DNNs) and convolutional neural networks (CNNs) have been mixed with established imbalanced data techniques like oversampling and undersampling. To evaluate these methods, experiments were conducted on imbalanced datasets, with shuffled data distributions. The results showed that the mixed Synthetic Minority Oversampling Technique (SMOTE)-Normalization-CNN model outperformed other methodologies. This mixed model can be applied to other real imbalanced binary classification problems. The article also emphasizes the importance of addressing imbalanced datasets in machine and deep learning, particularly when the minority class is more significant and crucial. It discusses the use of oversampling and undersampling methods, such as SMOTE, NearMiss, ROS, and RUS, combined with deep learning models. (Joloudari, Marefat, Nematollahi, Oyelere, & Hussain, 2023)

**The impact of class imbalance in classification performance metrics based on the binary confusion matrix**

The perplexity of dealing with imbalanced datasets when evaluating machine learning models is addressed in this article. The problem involves finding the most suitable performance metrics for assessing these models. The research uses simulations with binary classifiers to analyze how imbalance impacts performance metrics based on the binary confusion matrix. They introduce a new measure called the "Imbalance Coefficient," which helps characterize the class disparity better than previous methods like the Imbalance Ratio. Different clusters of performance metrics were identified throughout the study, such as the Geometric Mean and Bookmaker Informedness, which are good choices when focusing on classification successes. The authors also propose a set of null-biased multi-perspective Class Balance Metrics that extend the concept of Class Balance Accuracy to other performance metrics, providing a guide for selecting metrics in the presence of imbalanced classes. (Luque, Carrasco, Martín, & Heras, 2019)

**Literature Review**

1. **Class Imbalance**
   * The analysis reveals that there is an 8.74% rate of fraud in the dataset. This indicates that class imbalance should be implemented in the ML model.
   * The literature is suggesting a class imbalance. Using techniques like over sampling or synthetic sampling can improve the recall values.
2. **False Positives and False Negatives**
   * It is important to understand the impacts of both false positives and false negatives on the customer and the financial institution. The strategy will prioritize decreasing false negatives rather than false positives as false negatives hold a higher risk to both the institution and customers. False positives will not be ignored but indivertibly mitigated as well.
3. **Continuous Monitoring and Adaptation**
   * A key system that can be implemented is a continuous monitoring system to adapt and notice fraud patterns as they change.
4. **Technology and Fraud Detection**
   * Customers are relying on banks to detect fraud on their account and either prevent, stop, and or revert fraud that has taken place. It is important for banks to have advanced technology, mechanisms, and systems in place to detect fraud.

## **Proposed Solution**

The following are proposed components and techniques that can be implemented in an ML fraud detection system.

1. **Data Collection and Preprocessing**
   * This component will gather data and address the class imbalance and preprocessing data for model training
2. **Feature Engineering**
   * Development of a system that stays relevant to improve the ML model’s ability to detect false negatives.
3. **Model Selection and Training**
   * Selecting appropriate ML algorithms and training models on proper datasets. Key metrics to consider are recall and precision to name a few.
4. **Threshold Setting**
   * Creating a threshold that sets the balance between false positives and negatives.
5. **Feedback Mechanism**
   * To prevent overfitting, a mechanism to provide feedback on new data and fraud patterns so that the ML model can train the new data correctly.
6. **Monitoring and Reporting**
   * Implementation of a monitoring and reporting system to get the model's effectiveness vs previous data and false negatives that are reported.

## **Benefits**

1. **Enhanced Fraud Detection**
   * Improved accuracy of detecting fraudulent transactions.
2. **Customer Satisfaction**
   * Reduction of false negatives will increase the trust between the customer and the financial institution, increasing customer loyalty. While also mitigating financial stress.
3. **Financial Security**
   * Reduce the financial loss from paying out false negatives.
4. **Adaptive System**
   * Adapt and learn from evolving fraud techniques, keeping the institution ahead of the curve amongst its competitors.

## **ROI and Cost-Benefit Analysis**

1. **Costs:**
   * Investment in a new AI system, model training, and development.
   * Costs to maintain and host the model and data.
2. **Benefits**
   * Increased fraud detection for both negative and positive false flags.
   * Increased customer satisfaction and trust.

## **Risks and Mitigations**

1. **Model Accuracy**
   * Accuracy will be ensured by regularly updating the model and adapting the evolving fraud techniques.
2. **Regulatory Compliance**
   * Comply with all laws and regulations, especially data protection. Implement a transparent model to provide insights on its decisions.

## **Accuracy**

**False Positives**

False positives are transactions labeled as fraud, but legitimate.

1. **Cost Implications**
   * There are no cost implications on a false positive flag. The customer will be prevented from making a purchase and will be inconvenienced and annoyed. This may cause the customer to end its relationship with the financial institution if it happens regularly.
2. **Financial Implications for Users**
   * It is unlikely a user will face financial issues from a false positive.
3. **Operational Implications for the Bank**
   * Decrease in customer satisfaction. If too many false positives are flagged the institution can see fewer customers.

**False Negatives**

False negatives are transactions labeled as legitimate but are fraudulent.

1. **Cost Implications**
   * The bank will be expected to pay the cost of the fraudulent charge.
2. **Financial Implications for Users:** 
   * Customers may end up with many issues such as immediate loss of funds, overdrawn, no access to their accounts, financial strain.
3. **Operational Implications for the Bank**
   * The institution can be expected to lose out on a lot of money if false negatives are making it through. They can also see a drop in customers.

**Minimizing One Error Over the Other**

In fraud detection, the cost of false negatives (missing actual fraud) is often considered more severe than that of false positives (inconveniencing users). Therefore, the focus may be on minimizing false negatives even at the expense of a slightly higher false positive rate.

Banks might prioritize recall over precision in their models. Recall is the process of including both relevant and irrelevant cases which will help correctly call our relevant cases.

**Balancing Act**

Creating a balance between false negatives and false positives is crucial to the model. Overall false negatives need to be stopped more so than a false positive. By using false positive cases the model can reduce false negatives.

**Conclusion**

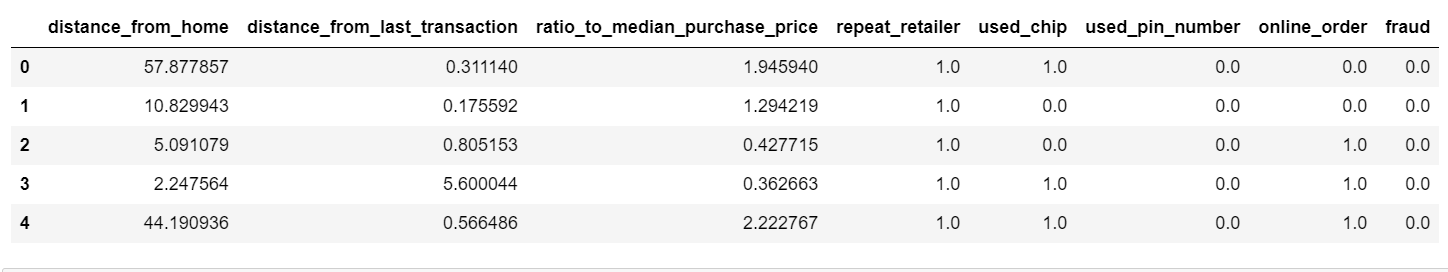
By implementing an ML based fraud detection system, we can provide innovative technology to minimize customer impact, reduce financial loss, and ensure adaptability to evolving threats.

It is important that accuracy and costs are factored into the ML model. When designing the threshold for false positives and negatives the model must align with the financial institutions risk tolerance, objectives, and mission statements.

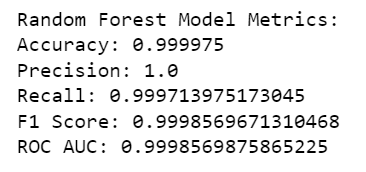
## **Appendix**

**Tables:**

**Table 1- Credit Card Fraud Detection Database:**

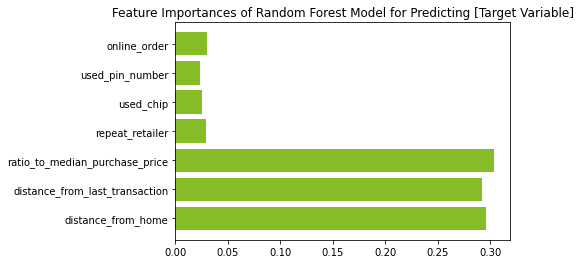


**Table 2- Random Forest Model Metrics**

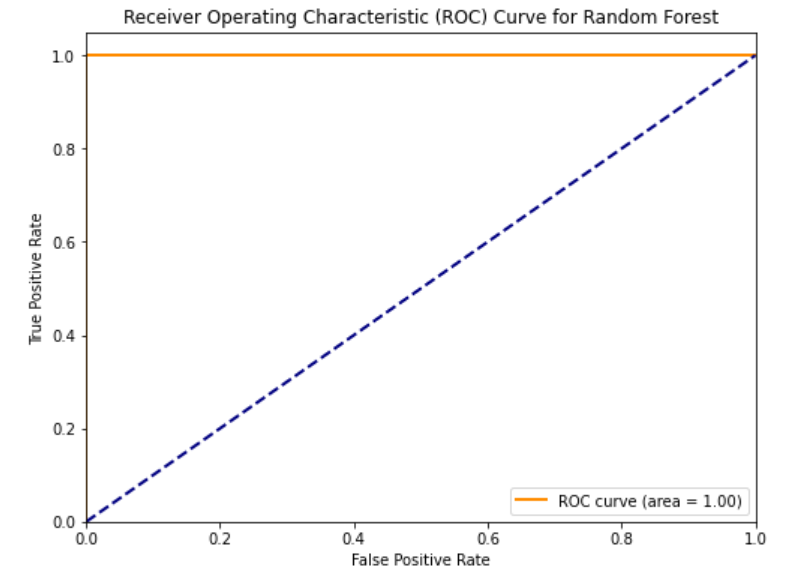


**Graphs:**

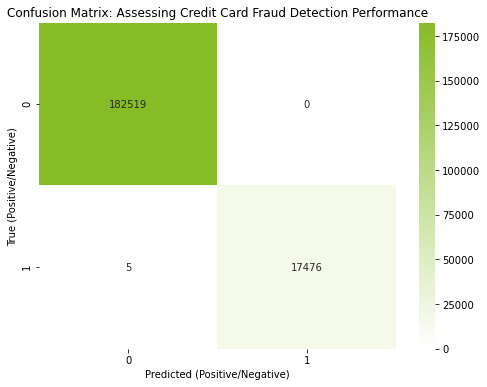
**Graph 1 - Weighing what variables are most important for Random Forest Model Predicting the Target Variable**



**Graph 2 - Evaluation metrics of Random Forest Model**



**Graph 3 – Assess the rates of the True Positive / True Negative / False Positive / False Negative**



# **References**

Dorfman, E. (2022, March 25). *Is Required for Machine Learning?* Retrieved from Postindustria: https://postindustria.com/how-much-data-is-required-for-machine-learning/

Dornadulaa, V. N., & S, G. (2019). Credit Card Fraud Detection using Machine Learning Algorithms. *Procedia Computer Science*, 632-641.

Heath, R. (2023, August 2). *American banks dominate AI transformation race*. Retrieved from Axios: https://www.axios.com/2023/08/02/american-banks-ai-transformation-race

McAfee, A. (2019). *Why More Data Isn’t Always Better in Machine Learning*. Retrieved from Harvard Business Review.

Pai, D. R. (2022, March). BUS 510: Business Analytics and Data Modeling - Lecture on data quality.

Telus International. (2022, February 17). *How is AI transforming fraud detection in banks?* Retrieved from Telus International: https://www.telusinternational.com/insights/trust-and-safety/article/ai-fraud-detection-in-banks