|  |  |
| --- | --- |
| **Module Title (s):** | Problem Solving for Industry |
| **Assessment Title:** | Capstone Project |
| **Lecturer (s) Name:** | Muhammad Iqbal |
| **Student Full Name:** | Eliabe Baliero De Moura |
| **Student Number:** | 2022474 |
| **Issue Date:** | 26th Feb 2024 |
| **Assessment Due Date:** | 17th May 2024 (23:59 Irish Time) |
| **Date of Submission:** | 17th May 2024 |

**Declaration**

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

Eliabe Baliero de Moura n:2022474.

**Enhancing Bank Transaction Security: An Accurate Approach to Fraud Detection**

**GitHub:** Eliabe2022474/Eliabe2022474-Eliabe\_2022474\_capstone\_finalCA (github.com)

**Table of Contents**

[Acknowledgments 2](#_Toc1599056917)

[Abstract 2](#_Toc443004845)

[Introduction 3](#_Toc392184205)

[Business Goals 4](#_Toc952464707)

[CRISP-DM Framework 5](#_Toc1707861887)

[Business Understanding 6](#_Toc324575066)

[Data Understanding 6](#_Toc695522635)

[Data Understanding 7](#_Toc1839707093)

[Data Preparation 7](#_Toc1699914778)

[Data Preparation 7](#_Toc1664748128)

[Modelling 8](#_Toc1143366441)

[Cross-Validation 10](#_Toc1768913214)

[Cross Validation 10](#_Toc1031910014)

[Evaluate each model 11](#_Toc1556093909)

[Model Fitting 12](#_Toc167309036)

[RandomFlorestClassifier Diagram 12](#_Toc925288472)

[Model Evaluation 13](#_Toc337748220)

[Evaluation 14](#_Toc1279985412)

[Deployment 16](#_Toc1540440645)

[Conclusion 17](#_Toc1669965012)

[References: 17](#_Toc1289848657)

# **Acknowledgments**

I extend my deepest gratitude to my professors Dr. Muhammad Iqbal, David McQuaid, and Ken Healy for their invaluable guidance, support, and expertise throughout the duration of this project. Their mentorship and insights have been instrumental in shaping my approach and achieving my goals.

Additionally, I would like to express my heartfelt appreciation to my friend Luiz Eduardo Costa for his unwavering encouragement and understanding during this journey. His constant support and encouragement have been a source of strength and motivation. I am truly grateful for the contributions of each of these individuals, without whom this project would not have been possible.

Eliabe Baliero de Moura n:2022474.

# **Abstract**

To find a way to improve and control anomaly detection in Bank transactions, our fraud detection software was developed as a solution to better spot and manage unusual activities in bank transactions. Our application uses machine learning to detect anomalies in transaction data, helping to catch fraudulent behaviours or odd patterns that might signal risks. We start by carefully preparing the data, fixing any missing info, outliers, or mistakes. Then, we create new transaction features like amount, time, location, and frequency to help us spot any suspicious activity.

Our fraud detection software brings a differential of other applications, it brings innovation for our customers to make them feel safe, and sure that their transfer will be done in a security way. Fraud detection groups are always concerned about the money, and we know how you worked hard for it, that's why we will bring our solution for you and for those that have already had a bad experience with a fraudulent transaction.

To achieve this project, we will make use of some tools, Languages, and API for our application connection such as: HTML, CSS, and Java script, for the back-end Python language and we count with MySQL for our data. In this document Fraud detection Group will try to bring as much as possible information in a didactic way where we will provide 5 slide presentations to ensure that all questions that you may have in mind are answered.

Eliabe Baliero de Moura n:2022474.

# **Introduction**

In today's digital landscape, safeguarding financial transactions is crucial. Our team is dedicated to refining and managing anomaly detection within transfers. Our solution revolves the power of machine learning techniques, as machine learning techniques are crucial to avoid transaction fraud (Fariborz, 2023).

In 2023, Kartikay Goyle said that data preparation and cleaning enhance dataset quality, thinking that our approach begins with thorough data preprocessing, which involves addressing missing values, outliers, and inconsistencies. Following this, we engage in feature engineering to extract essential transaction attributes such as amount, time, location, and frequency.

Our software stands out due to our commitment to innovation and customer safety. We recognize the significance of securing your hard-earned money and strive to provide reassurance with every transaction. Our solution not only serves as a deterrent against fraudulent behaviour but also offers protection to those who have experienced fraudulent transactions in the past.

To realize our objectives, we utilize a range of tools, programming languages, and APIs for seamless application development. From front-end technologies like HTML, CSS, and JavaScript to back-end essentials such as Python and data management with MySQL, our technological toolkit is comprehensive and reliable.

In this document, our group aims to offer comprehensive insights in a clear and accessible manner. Through a poster presentation, we will try to address any questions or concerns you may have, ensuring transparency and understanding throughout our project journey.

:

Eliabe Baliero de Moura n:2022474.

# **Business Goals**

“The innovations can be responses to new needs in today’s market,”

(Tom B.W, 2012).

Our goal is to bring an innovative system to detect potential fraud, avoiding losses to banks or organizations by reducing chargebacks. The differential of our application system is because as we know the security of your transactions and protecting your money is our priority. Most of the system when they suspect the transaction is a fraud, normally they will just message you to confirm and ask you to say yes or not, where could be a problem if someone else has access to your Phone and just say yes. Our system will bring you the choice where you can customize the system with a personal question, for example let’s suppose you are travelling and lose your phone/card and then someone find it and then try to do a big transaction straightway instead it just ask to confirm yes or not, it will ask a specific answer that you set where could be your name surname and so on.

**Open-Source Tools**

Our group chose Python for the backend language and MySQL. As opensource is free for all users and this makes OSS classic public good (Anamika Sen, 2022, p.1). Let's break down these choices:

Python:

1. Python is a popular programming language known for its simplicity and readability.

We chose It, as it has many libraries and tools for machine learning, which are crucial for building anomaly detection models.

Python for sure is widely used in data science and machine learning communities, so there's plenty of resources and support available.

An alternative could be R, another language for statistics and machine learning, but Python is more versatile and commonly used.

MySQL:

1. MySQL is an open-source database system used for storing structured data.

It's scalable, reliable, and performs well, making it suitable for handling large volumes of transaction data.

MySQL has features like ACID compliance and security mechanisms to ensure data integrity and safety.

Being open source means it's free and has a large community for support.

An alternative could be Microsoft SQL Server. It has advanced features but comes with licensing costs and may not have as much community support as MySQL.

Eliabe Baliero de Moura n:2022474.

# **CRISP-DM Framework**

By following CRISP-DM framework steps, we can effectively develop and deploy a robust fraud detection solution for bank transactions, enhancing security and customer trust. Frank Lee (2023)

Diagrama

Descrição gerada automaticamente

**Business Understanding:** In this step, my goal is to fully understand the business problem. By defining clear project objectives, I can then transform them into specific data.

**Data Understanding:** Here I will be preparing it for analysis. This involves thorough data collection, exploration, and initial cleaning steps.

**Data Preparation:** Now, I focus on getting the data ready for modelling. I select the necessary data subset, clean it up, transform it as needed, and fit it into the models.

**Modelling:** This phase is all about building the models. We construct the models, and then evaluate their performance rigorously using some validation techniques.

**Evaluation:** Here, I assess the models against predefined success criteria, then I analyse their performance and draw valuable insights from the results obtained.

**Deployment:** This step, I move to the deployment phase, where I put the selected model into action. This involves integrating it into the existing system.

Eliabe Baliero de Moura n:2022474.

# **Business Understanding**

**Business Understanding:**

**Problem Identification:** The primary goal is to improve and control anomaly detection in bank transactions using fraud detection software. This is crucial for enhancing security, preventing financial losses due to fraudulent activities, and instilling confidence in customers.

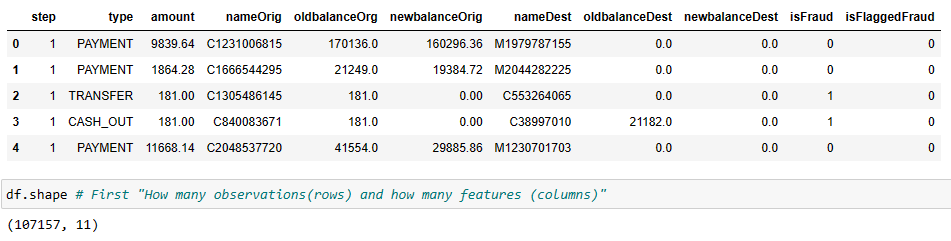
**Objectives:** Develop a machine learning-based fraud detection software.

**Success Criteria:** Accurately detect and manage fraudulent transactions, minimizing financial losses.

# **Data Understanding**

1. **Data Collection:** Transaction data from bank records is collected for analysis. This includes information such as transaction amount, timestamp, location, and frequency.
2. **Data Sources:** Sources of transaction data may include internal bank databases, third-party financial platforms, and historical transaction records.
3. **Data Preparation:** Data is prepared for analysis by cleaning, transforming, and enhancing it as needed. This may involve handling missing values, encoding categorical variables, and creating new features to improve the accuracy of anomaly detection.

Here are some samples of the data:



Eliabe Baliero de Moura n:2022474.

# **Data Understanding**

The collected data consists of (107157, 11) 107157 rows and with 11 variables. The description of each column/variable can be seen below:

* **step**: The unit of time in the simulation.
* **type:** Type of transaction.
* **amount:** The amount of money involved in the transaction.
* **nameOrig:** The original name associated with the transaction's origin.
* **oldbalanceOrg:** The original balance before the transaction for the origin account.
* **newbalanceOrig:** The new balance after the transaction for the origin account.
* **nameDest:** The name of the destination account.
* **oldbalanceDest:** The original balance before the transaction for the destination account.
* **newbalanceDest:** The new balance after the transaction for the destination account.
* **isFraud:** Indicates whether the transaction is fraudulent (1) or not (0).
* **isFlaggedFraud:** Indicates whether the transaction is flagged as fraudulent (1) or not (0).

# **Data Preparation**

Data preparation is a crucial step in the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology. Here's the approach that I used to do it.

**1** Irrelevant Data: Remove data that doesn't contribute to the analysis.

**2** Duplicates: Eliminate identical records to avoid bias in analysis.

**3** Type Conversion: Convert data into appropriate formats for analysis

**4** Syntax Errors: Correct any formatting inconsistencies.

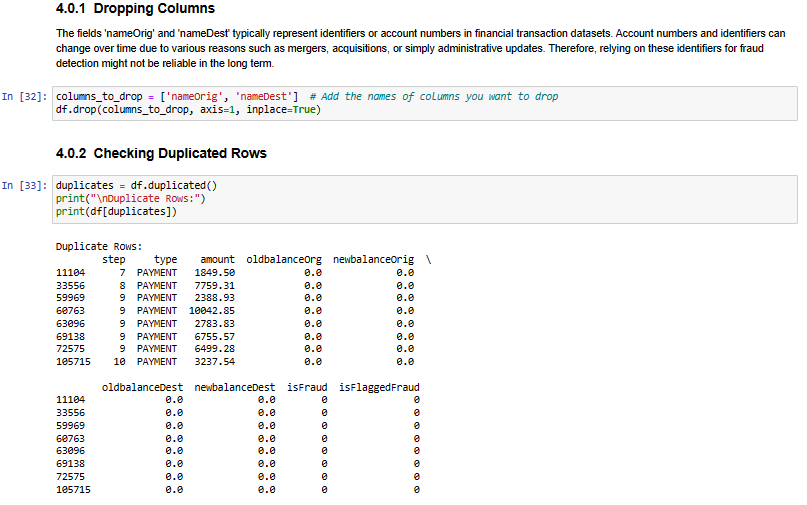
**5** Missing Values: Address missing data through imputation or removal.

**6** Outliers: Detects and handles extreme values that skew analysis. **7** Standardize: Scale numerical variables for consistent analysis.

Eliabe Baliero de Moura n:2022474.

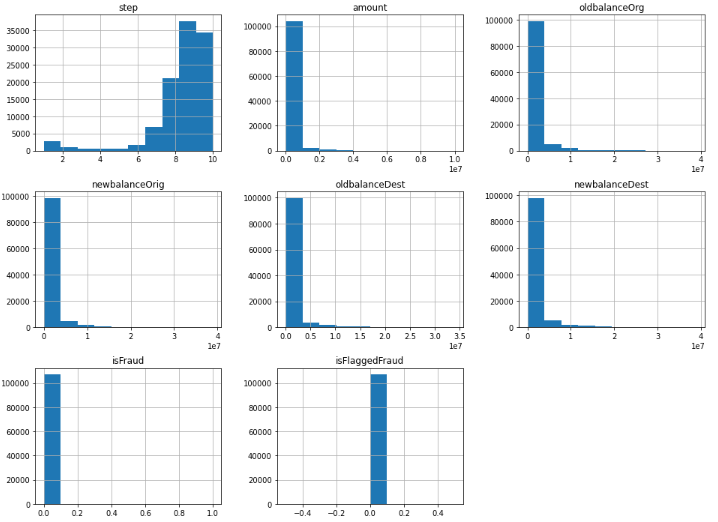
# **Data Preparation**

For example, in the image below you can see that I am checking my data to eliminate/drop Irrelevant data like: “nameOrig”, “nameDest”, and checking duplicate rows.



In next example I use a Histogram which is a graphical representation of the distribution of numerical data and grouped it by Transaction type, once transaction type has five different class which is:

['CASH\_IN', 'CASH\_OUT', 'tDEBIT', 'PAYMENT', 'TRANSFER'] , so The Histogram looks like this:



Eliabe Baliero de Moura n:2022474.

# **Modelling**

In the **Modelling** phase, we will start creating models to find patterns inside our data and to make future predictions for business purposes.

Machine learning involves different statistical models that learn and recognize patterns in data. These datasets, known as Training Datasets, serve as the foundation for model training and can be classified based on their objectives:

**1 Supervised Learning**: In this paradigm, models are guided during training. Supervised use labelled data, a target variable, representing what we aim to predict, is provided. Think of the model as a student learning from data and attempting predictions. If the model's predictions are incorrect, it adjusts itself until errors are minimized.

**. Regression:** Predicting numerical variables like house prices, car prices, or energy consumption.

**. Classification:** Predicting categorical variables such as customer churn probability, cancer cell detection from images, or credit scoring.

**2 Unsupervised Learning**: Here, no target variable is provided. Models autonomously uncover patterns within the data.

**3 Reinforcement Learning**: Models learn through interaction with their environment. This approach is often used for simulations and decision-making processes.

Eliabe Baliero de Moura n:2022474.

**Modelling**

For this data and in this context of machine learning we are working with **Supervised learning** because it offers us labelled data. **Labelled** data refers to data instances where the target variable

(in this case, the "**isFraud**" column), is our target variable, and labelled one.

Columns like: **Payment type** or **isFlaggedFraud** could be a consideratetarget variable as well, once they are categorical variables.

On the other hand, our **unlabelled** is represented as "Destination Port","Bytes,

"etc. as they do not directly indicate the action taken by the transaction, because

they are not categorical data; therefore, they cannot categorize a decision.

We are using **Supervised learning;** this is the most common approach in machine learning because:

**.** Labelled data is often readily available, making it practical for various tasks.

**.** Tasks have clear objectives, simplifying both training and evaluation.

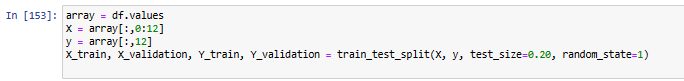
**.** A wide range of established algorithms and techniques exist, making implementation easier.

**.** Standard metrics enable straightforward performance evaluation and model selection.

# **Cross-Validation**

Create a validation dataset

Now we will split the loaded dataset into two, 80% of which we will use to train, evaluate and select among our models, and 20% that we will hold back as a validation dataset as shown below.



Eliabe Baliero de Moura n:2022474

# **Cross Validation**

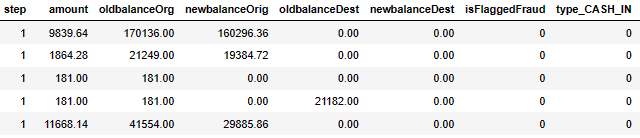
**Remember** that I deleted 2 columns ('nameOrig', 'nameDest') from my data and converted “Payment type” variable to numerical, as my Payment type variable has 5 type of payment it will generate five columns in my dataset, therefore my original dataset had 11 columns, as I deleted 2 and as payment type is no longer a column because it was converted to 5 different types of payments it become 8, therefore 8 + 5 payment types is now 13 columns, and as an array start from index 0 our rangers 12, that's why we use this range:

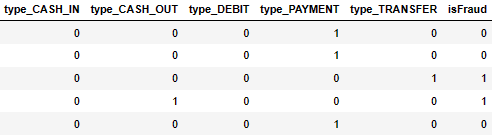
array = df.values

X = array[:,0:12]

y = array[:,12]

Below you can see how the dataset looks like.

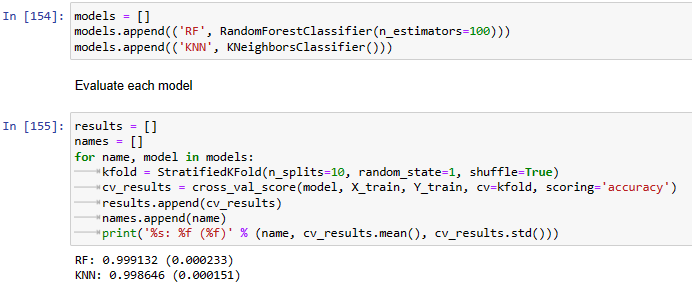




Eliabe Baliero de Moura n:2022474.

# **Evaluate each model**

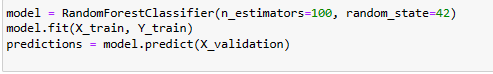
Evaluate each model to choose the best one to fit:



As we can see above, I chose two algorithms, **RandomFlorestClassifier** and **KNeighborsClassifier**.

**RandomFlorestClassifier** is giving us a better accuracy over KNeighborsClassifier.

# **Model Fitting**



The random forest algorithm involves randomness in the process of building individual decision trees and making splits at each node. Setting random\_state ensures that the random forest model will produce the same results each time it is trained on the same data. This is important for reproducibility and consistency in the behaviour of the model.

Eliabe Baliero de Moura n:2022474.

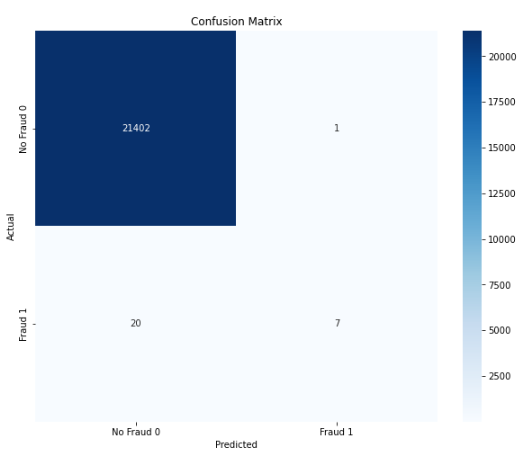
# **RandomFlorestClassifier Diagram**

# 

The n\_estimators parameter specifies the number of decision trees that will be used in Random Florest model. In the example provided (n\_estimators=100), it means that the Random Florest will consist of 100 decision trees.

Eliabe Baliero de Moura n:2022474.

# **Model Evaluation**



Confusion Matrix:

The matrix is divided into four area:

**True NoFraud (TN)**: The model predicts correctly if transaction is Not a Fraud

**False NoFraud (FN)**: The model predicts incorrectly if transaction is Not a Fraud

**True Fraud (TF): The** model predicts correctly if transaction is a Fraud

**False Fraud (FF)**: The model predicts incorrectly if transaction is a Fraud

Those values in the matrix if we some we can see it's representing 20% of our data:

No Fraud 0: The classifier correctly predicted 21402 instances as "no Fraud" and misclassified only 1 as another class.

Fraud 1: The classifier correctly predicted 7 instances as "Fraud", but it misclassified 20 instances as other classes.

Eliabe Baliero de Moura n:2022474.

# **Evaluation**

In the **Evaluation** phase, we will further evaluate the model into the context of the business problem.

The deliverable or result of this phase should include:

* Model business assessment
* Defining Cost and Benefit

**How Much Do Chargeback Fees Cost?**

Chargeback fees cost between $20 and $100, depending on the merchant's agreement with their acquirer.

Chargebacks can also come with many expenses that merchants often don’t think about–or even know exist–until it’s too late:

**Transaction fees** are indeed a significant expense for merchants, covering fees to payment processors and interchange fees to issuing banks, typically ranging from 1.5 to 4% per transaction. Chargebacks result in wasted money as these fees are essentially lost.

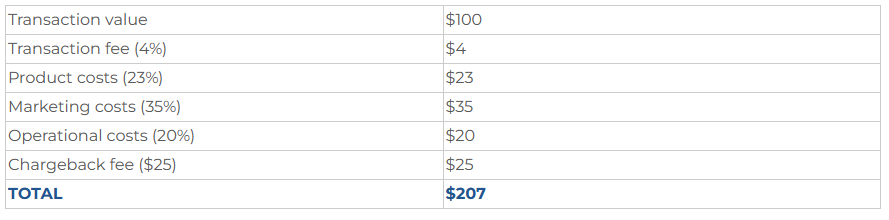
**Operational costs**, such as order processing, inventory management, and shipping, typically consume around 20% of merchant revenue. Chargebacks directly impact this revenue stream when orders are reversed.

**Marketing and acquisition expenses** are also substantial, with merchants investing significant resources in advertising, sales teams, and promotions, often amounting to 30 to 40% of revenue. Chargebacks diminish the return on these investments as sales generated through marketing efforts are lost.

Eliabe Baliero de Moura n:2022474.

**Evaluation**

Chargeback Gurus (2024).



In this table available in Chargeback Gurus company website [Chargeback Recovery & Prevention Experts | Chargeback Gurus](https://www.chargebackgurus.com/). we can see that a chargeback for a transaction of $100 easily equates to more than double the lost sale: over $200 lost on a $100 transaction.

Therefore:

Confusion Matrix:

[[14386 1]

[ 4 8]]

noFraud Fraud

The first thing we do is to define the cost and benefit of each decision. We will define it similarly with the previous confusion matrix. The cost will depend on the merchant's agreement with their acquirer between $20 to $100. We need to consider some additional fees cited above

True Positive (TN): If the model predicts the transaction is noFraud and the prediction is correct (customer actually buy), we will not have any loss

False Positive (FN): If the model predicts transaction is noFraud and the prediction is incorrect the loss can be between 20 $100 plus additional fees per transaction.

True Negative (TF): If the model predicts a transaction is Fraud and the prediction is correct, we are fine.

False Negative (FF): If the model predicts the transaction is Fraud and the prediction is incorrect nothing happens, once the system can message to try the customer and confirm the purchase.

Eliabe Baliero de Moura n:2022474.

# **Deployment**

Indeed, deploying a machine learning model into production is where the real value of data mining and model development is realized. Here's a common method for releasing a machine learning model into production:

1 REST API: Expose the model as a RESTful API (Application Programming Interface) endpoint. This allows other applications and systems to interact with the model over HTTP protocols, making predictions based on input data provided by clients.

2. Documentation and Training: Provide comprehensive documentation and training materials for end-users, developers, and stakeholders on how to interact with and use the deployed model effectively.

# **Conclusion**

The project follows the CRISP-DM framework, starting with Business Understanding to define objectives, followed by Data Understanding and Preparation to ensure data quality and relevance. Modelling involves building machine learning models for anomaly detection, and Evaluation assesses their performance against success criteria. Finally, Deployment integrates the selected model into the existing system, ensuring customer safety and innovation in combating fraudulent transactions.

Eliabe Baliero de Moura n:2022474.

# **References:**

Anamika Sen. (2022, p.1) Open SourceSoftware [Online] Available at: <https://thecommonsjournal.org/articles/10.5334/ijc.1176>

[Accessed 08 March 2024].

Kartikay Goyle. (2023, p.1) A Machine Learning Approach [Online] Available at:

[2307.07119.pdf (arxiv.org)](https://arxiv.org/pdf/2307.07119.pdf)

[Accessed 10 March 2024].

Frank Lee (2023) Using Supervised Machine Learning and CRISP-DM [Online]Available at:

[EJ1403991.pdf (ed.gov)](https://files.eric.ed.gov/fulltext/EJ1403991.pdf)

[Accessed 24 April 2024].

Chargeback Gurus (2024) How much Do Chargeback fees cost [Online]Available at:

[2024 Chargeback Fees - The True Cost of Your Chargebacks (chargebackgurus.com)](https://www.chargebackgurus.com/blog/chargeback-fees-the-true-cost-of-your-chargebacks)

[Accessed 10 May 2024].

Tom Brzustowski (2012., p.1) Innovation in Canada [Online] Available at: [Innovation in Canada : Why We Need More and What We Must Do to Get It (ebscohost.com)](https://viewer.ebscohost.com/EbscoViewerService/ebook?an=3323869&callbackUrl=https%3a%2f%2fresearch.ebsco.com&db=e020mww&format=EB&profId=eds&lpid=&ppid=&lang=en&location=https%3a%2f%2fresearch.ebsco.com%2fc%2fjzntuu%2fsearch%2fdetails%2fg43pyfx4g5%3flimiters%3dFT%253AY%26q%3dinnovation&isPLink=False&requestContext=&profileIdentifier=jzntuu&recordId=g43pyfx4g5)

[Accessed 12 March 2024].

Eliabe Baliero de Moura n:2022474.