**CCT College Dublin**

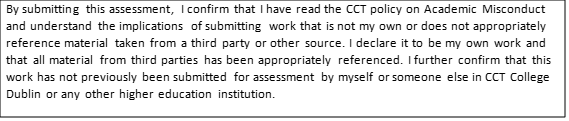
**Assessment Cover Page**

Form

|  |  |
| --- | --- |
| **Module Title:** | Strategic Thinking |
| **Assessment Title:** | Analysis of Financial Fraud |
| **Lecturer Name:** | James Garza |
| **Student Full Name:** | Eriton Delgado Pereira |
| **Student Number:** | 2023457 |
|  |  |
|  |  |
|  |  |
|  |  |
| **Assessment Due Date:** | 17-11-2023 |
| **Date of Submission:** | 17-11-2023 |

Form

**Declaration**



## 

## **Prediction of Fraudulent Financial Transactions.**

### *By*

#### Eriton Delgado Pereira – Student ID: 2023457

*Higher Diploma in Science in Data Analytics for Business Strategic Thinking*

*James Garza CCT College Dublin, Ireland*

# **Abstract:**

In the Strategic Thinking class, taught by Professor James Garza, we received the mission of producing our first work in data analysis. After some extracurricular discussions, we decided to look for a topic related to the financial sector, as we believe that the application of the knowledge of statistics, data preparation, machine learning and strategic thinking that are being acquired in the course, applied to topics relevant to this sector, will bring a greater alignment with the vacancies and job opportunities that we intend to seek in the market. And it is adherent with the business area we intend to follow.

After some effort evaluating possibilities and searching for datasets, in sources disseminated in the classroom such as Kaggle, World Bank, Eurostat, Covid Google among others, we decided to use a database of studies made available in Kaggle to develop a machine learning model to identify fraudulent transactions.

We believe that this issue is a serious, recurrent, and highly relevant problem in the global financial system. It is common knowledge that fraud in financial transactions occurs all the time, generating numerous losses not only for financial institutions, but also for the victims of fraud and for society, since these losses are somehow passed on to customers in the existing costs to have a bank account.

This report aims to deep dive into a huge a dataset of financial transactions and based in its information predict with applying machine learning models, what transaction was fraudulent, note that, in the dataset we have this information (fraude or not) and I will pretend to use it to test and train my model and compare its accuracy against the information registered in the dataset.

# Finally, the report will follow the Cross-Industry Standard Process for Data Mining (CRISP-DM). It is an approach to project management and goes through the stages: Business Understanding, Data Understanding, Data Preparation, Modelling and Evaluation and Deployment.

# **Key-words: CRISP-DM, financial transaction, fraud.**

# **Code available at** [**GitHub**](https://github.com/eriton2023457/CA2-Strategic-Thinking)**.**

**TABLE** **OF** **CONTENTS**

[**Abstract:** 3](#_Toc153725178)

[**Business Understanding:** 5](#_Toc153725182)

[**Technologies Used:** 5](#_Toc153725183)

[**Libraries:** 5](#_Toc153725184)

[**Ethical Considerations:** 6](#_Toc153725185)

[**Data Undestanding:** 6](#_Toc153725186)

[**Data Preparation:** 8](#_Toc153725188)

[**Modelling:** 9](#_Toc153725192)

[**Evaluation:** 11](#_Toc153725199)

[**Deployment:** 11](#_Toc153725200)

# 

# **Table of Figures**

[**Figure 1 - Shape of the data set 6**](#_Toc153727120)

[**Figure 2 - Data of the dataset 6**](#_Toc153727121)

[**Figure 3 - Head of the dataset 7**](#_Toc153727122)

[**Figure 4 - Importing Libraries 8**](file:///C:\Users\erito\Downloads\CA%202%20Strategic%20Thinking\CA2%20Strategic%20Thinking%20Eriton%20Delgado%20Pereira%202023457.docx#_Toc153727123)

[**Figure 5 - Missing Values 8**](#_Toc153727124)

[**Figure 6 - Amout by type 9**](#_Toc153727125)

[**Figure 7 - Fraudulent Transactions by type 9**](#_Toc153727126)

[**Figure 8 - Coding to develop the chart with amount of fraud transactions by type 10**](#_Toc153727127)

[**Figure 9 - Amount of Fraud Transaction by type 10**](#_Toc153727128)

# **Business Understanding:**

In this initial phase of applying the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology, the goal is to understand the scenario and comprehend its needs, meaning the objectives and requirements for this project. The financial market recognizes that in financial transactions, whether online or physical through ATMs, a considerable number are fraudulent or attempted frauds. The objective of this project is to study a dataset with over 6 million financial transactions and predict which of these financial transactions were fraudulent, comparing the results obtained by the machine learning model versus the results recorded in the dataset. It is important to note that the financial transactions targeted in this study include bank transfers, credit, debit, bill payments, cash deposits, or cash withdrawals. More sophisticated financial transactions and/or those involving investments, interest payments, or dividends, for example, are not the focus of this study.

The requirements for this project are:

* Data preparation to support data analysis and prepare it for the machine learning model;
* Building a robust machine learning model for categorize which is fraud and which is not, specifically a "Decision Tree”.
* Comparing the final prediction results with the original data.

It is important to emphasize that this analysis will be conducted on a single dataset. The application of the study to other databases with the same type of business information would bring more robustness to the analysis of the success/accuracy obtained through this model.

The success criteria and the project's objective are to predict with a high degree of accuracy, subjectively defined here as at least 85% accuracy, which operation was fraudulent, comparing it with the actual result in the dataset.

**Technologies Used:**

All the work was carried out in the Python language, using the jupyter platform for the creation and execution of codes and commands. The original database has the CSV format, "comma separated value".

# **Libraries:**

For this step we will only use the Pandas and Matplotlib libraries and the Skelearn machine learning model to apply “Decision tree” model.

# **Ethical Considerations:**

Our dataset of financial transactions was extracted from Kaggle (<https://www.kaggle.com/datasets/chitwanmanchanda/fraudulent-transactions-data>), customer data is encrypted and respects the General Data Protection Law, in the source of this dataset there is also no information about the financial institution, so we cannot confirm if the dataset is real or fictitious.

# **Data Undestanding:**

The second stage of the CRISP-DM process requires you to acquire the data (or access to the data) listed in the project resources, it also will contains an initial data collect report and data description report by examining the ‘gross” or “surface” of the dataset and reporting its results:

Tabela

Descrição gerada automaticamente com confiança média

Figure 1 - Shape of the data set

Tabela

Descrição gerada automaticamente com confiança média

Figure 2 - Data of the dataset

In the figure 1 – shape of the data set, it is possible to see the dataset’s size and it contains 6.362.620 millions of rows and 11 columns.

In the figure 2 – info of the data set shows 11 columns with: 3 columns type int64, 3 columns type object and 5 columns type float64.

Note that: The data type object which returned in 3 columns: “type”, “NameOrig”, “NameDest” is a point of attention in order to implement some machine learning models indicates that we will need to use some techniques of data transformation.

Tabela

Descrição gerada automaticamente

Figure 3 - Head of the dataset

This step introduce us to the information inside the dataset, more specifically in the first 5 rows where we can take an overall idea about the dataset by analysing informations such as “type” which is the transaction method”, amount, NameOrig and NameDest which is the account holder involved in the transaction, old balance, new balance and so on.

Note that, we also have the collum “IsFraud” which indicates us as a target to apply the Machine Learning model.

**Business definition for the columns.**

**step**: Indicates the time of the transaction.

**type**: The type of the transaction, which can be "PAYMENT," "TRANSFER," "CASH\_OUT," "DEBIT," etc.

**amount**: The amount of the transaction.

**nameOrig**: The account from which the transaction originated.

**oldbalanceOrg**: Balance of the originating account before the transaction.

**newbalanceOrig**: Balance of the originating account after the transaction.

**nameDest**: The target account for the transaction.

**oldbalanceDest**: Balance of the target account before the transaction.

**newbalanceDest**: Balance of the target account after the transaction.

**isFraud**: A binary column that indicates whether the transaction is fraudulent (1) or not (0).

**isFlaggedFraud**: A column that most likely indicates whether the transaction is flagged as fraudulent (1) or not (0).

# **Statistical analysis of the dataset.**

Tabela

Descrição gerada automaticamente

By the statiscal analysis of the dataset we can confirm that, all columns has the same amount of line by count equals to 6.362.620, indicating that there is no missing value in any column. Beside it is interesting to see the min and max amount transactioned.

# **Data Preparation:**

In the 3d stage of CRISP-DM we are going to work on the data preparation,it is all about how we organize the data to modelling.

In this stage we are going to import librarires, looking for missing values and duplicate values.

**Importing the required libraries:**

Text

Auto-generated descriptionAs previously mentionated in this project we imported pandas, matplotlib and skelan in order to execute the models.

Figure 4 - Importing Libraries

**Identify if there are any missing values in any columns:**

Text

Auto-generated description with medium confidence

Figure 5 - Missing Values

**Check for any duplicate rows in the database:**

Graphical User Interface, Text, Application

Auto-generated description

We identified that there were no missing values in any column of the dataset and also no duplicate rows in the entire dataset. It is good, because avoid us to manipulate the dataset by dropping a set of information in order to execute the models properly and make the result more accurate.

Hopefully, we don’t need in this case.

# **Modelling:**

In this stage of the Cross-Industry Standard Process for Data Mining (CRISP-DM) we are going to apply the modelling technique to be used in order to achieve the goal of this project, is important to mention we have already selected the specific modelling technique during the Business Undestanding stage. Which will be Decision Tree.

The reasons that I choose “Decision Tree” model for this project are various but I believe is important to highlight a few of the them.

* My problem is a Classification problem: My goal predict which transaction is fraud or not, decision trees are well-suited to tasks where the goal is categorize the date into differente classes.
* Mixed Data Types, In my dataset I am working with 3 different data types, decision trees model can handle both, categorical and numerical features without requiring an extensive use of processing.

**Before we apply the model, we are going to analyze a good bit deep the dataset to take some insights that can be used in the next steps.**

**Calculating the mean of amount by type of financial transaction**

Graphical User Interface, Text, Application

Auto-generated description

Figure 6 - Amout by type

**Fraudulent transactions by type**

Graphical User Interface, Text, Application

Auto-generated description

Figure 7 - Fraudulent Transactions by type

In this code, we filter out the rows where the "isFraud" column equals 1 to get the fraudulent operations. We then use the value counts() method to count the number of occurrences of each type of operation. The result is the count of fraudulent operations for each transaction type.

Despite we have 5 different types of transactions, according to the dataset only cash out and transfer has occurrence of fraud. It indicates an linear relationship between transactions type and fraud which can be further explored.

**Chart:**

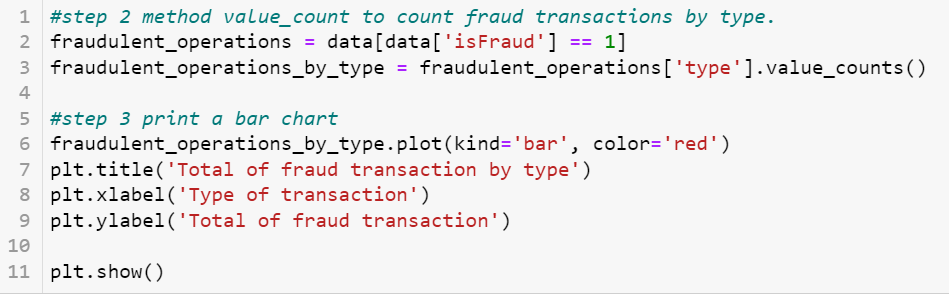


Figure 8 - Coding to develop the chart with amount of fraud transactions by type

Chart, Bar Chart

Auto-generated description

Figure 9 - Amount of Fraud Transaction by type

# **Decision Tree model**

Text

Auto-generated description

Table

Auto-generated description

Addressing the basis for correcting an identified error: Conversion of string fields to float64.

Text

Auto-generated description

**Confirmation**Table

Auto-generated description

# **Result:**

Tela de celular com texto preto sobre fundo branco

Descrição gerada automaticamente

# **Evaluation:**

The model demonstrated an accuracy of 99.97%. Furthermore, the high precision of recall, F1-score, and support values suggest that the selected model for this dataset exhibits exceptional accuracy and appears ready for real-world application if the data collected in this case (sourced from Kaggle) accurately reflects real-world scenarios. The substantial quantity of rows indicates a robust sample, and the nearly 100% accuracy suggests that predicting fraudulent transactions with this set of characteristics is entirely feasible.

However, for additional assurance and confidence in the model, in the subsequent stages of the project, we will apply cross-validation techniques and attempt to predict fraud using alternative machine learning models. This approach will enhance our confidence as we proceed with the project, aiming to develop a prototype for real-time prediction of fraudulent transactions.

# 

# **Deployment:**

Although Successfully developing and evaluating my machine learning model for predicting fraudulent transactions, I believe I need to validate it more. In order to do it, In the next stages of this project I am planning to apply cross validation techniques, change the sample of training and testing and apply other models with the goal to develop a prototype for real-time prediction of fraudulent transactions and most important, understand well exactly how the model selected works in the dataset to be able to explain it to anyone.

However the results are very satisfactory based on what was developed and, in addition the substantial size of the dataset.