UNIVERSITY OF TURKU

ABSTRACT

- [ ] Bachelor's thesis
- [ ] Master's thesis
- [ ] Licentiate's thesis
- [ ] Doctoral dissertation

Subject	Information Systems Science - IMMIT	Date	04.07.2021
Author(s)	José Faustino Martínez-Delgado Rubio	Number of pages	90+appendices
Title	**Perspectives On Data-Driven failure diagnosis - With a case study on failure diagnosis at an Payment Service Provider**		
Supervisor(s)	Dr. E.A.M. Caron		

Abstract

Data-driven failure diagnosis aims to extract relevant information from a dataset in an automatic way. In this paper it is being proposed a data driven model for classifying the transactions of a Payment Service Provider based on relevant shared characteristics that would provide the business users relevant insights about the data analyzed.

The proposed solution aims to mimic processes applied in industrial organizations. However, the methods discussed in this paper from these organizations does not directly deal with the human component in information systems. Therefore, the proposed solution aims to offer the relevant error paths to help the business users in their daily tasks while dealing with the human factor in IT systems. The built artifact follow the next set of steps:

- Categorization of variables following data mining techniques.
- Assignation of importance for variables affecting the transaction process using predictive machine learning method.
- Classification of transactions in groups with similar characteristics.

The solution developed effectively and consistently classify more than 90% of the faults in the database by grouping them in paths with shared characteristics and with a relevant failure rate. The artifact does not depends in any predefined fault distribution and satisfactorily deal with highly correlated input variables. Therefore, the artifact has a scalable potential if previously, a data mining categorization of variables is performed. Specially, in companies that deals with rigid processes.

Key words	Data Driven, Fault Management, Incident Management, Machine Learning.

UNIVERSITY OF TURKU
Turku School of Economics

PERSPECTIVES ON DATA-DRIVEN FAILURE DIAGNOSIS

With a case study on failure Diagnosis at an E-Money Institute

Master's Thesis
in IMMIT (Information Management = major)

Author(s):
José Faustino Martínez-Delgado Rubio

Supervisor(s):
Dr. E.A.M Caron

04.07.2021
Luxembourg

The originality of this thesis has been checked in accordance with the University of Turku quality assurance system using the Turnitin OriginalityCheck service.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONIMS

ATM – Automated Teller Machine
IT – Information and Technology
PSP – Payment Service Provider
PSD – Payment Service Directive
SCA – Strong Customer Authentication
FD – Failure Diagnosis
RS – Research Question

dd-FD – Data Driven Failure Diagnosis
NP – Nondeterministic Polynomial (time)
ITIL – IT infrastructure Library
OSI – Open System Interconnection
ISO – International Standards Organization
DoS – Denial of Service
PM-FD – Process Monitoring and Failure Diagnosis
HROs – High Reliable Organizations
BI&A – Business Intelligence & Analytics
OLAP – Online Analytical Processing
ETL – Extract Transform and Load
TTS – Trouble Ticket Systems
MTS – Multivariate Time Series
TE – Tennessee Eastman
PCA – Principal Component Analysis
ICA – Independent Component Analysis
PLS – Partial Least Squares
FDA – Fisher Discriminant Analysis
MCA – Multiple Correspondence Analysis
EDA – Exploratory Data analysis
PA – Payment Authorization
CP – Capture
CC – Credit Card
DC – Debit Card
CDA – Confirmatory Data Analysis
AUC – Area Under the Curve
TP – True Positive
TN – True Negative
FP – False Positive
FN – False Negative
CRISP-DM – Cross Industry Standard Process for Data Mining
CPU – Central Processing Unit
ROC – Receiver operating characteristic
FPR – False Positive Rate
TPR – True Positive Rate
EUR – Euro
DKK – Danish Krone
GBP – Pound Sterling
CSV – Comma Separated Document
EEA – European Economic Area

ID – Identifier
HUF – Hungarian Forint
EMI – Electronic Money Institute
MVS – Minimum Variance Sample
JSON – JavaScript Object Notation
SEPA – Single Euro Payments Area
IBAN – International Bank Account Number
BIN – Bank Identification Number

1 INTRODUCTION

In recent history financial institutions have been early-adopters regarding the introduction of technology in their processes. Specially regarding the characteristics required to ensure high quality banking procedures as traditionally new technologies have helped to ensure the integrity of the transactions, avoid human error and increase the overall reliability of their procedures.

An early adopter example of a specific technology of financial institutions would be the commercial adoption of cryptography standards for withdrawing money from ATMs, or the introduction of IT systems two ensure integrity of transactions.[1]

However, this initial environment where technology played an important role ensuring integrity, availability and security of transactions has become a norm already and therefore, financial institutions are shifting into the use of technology in more open-minded ways. Especially since the appearance of disruptive Fintech and Insurtech companies. Fintech is defined as:

"A new financial industry that applies technology to improve financial activities" (Schueffel, P. 2016)[2]

Nowadays, technology not only has a supportive role from a system perspective, but also helps managers in decision making, gather and process data to acquire marketing insights, or help automate all kinds of processes in a company. This evolution or introduction of new technologies ultimately have to deal with the well-established standards and procedures of the systems already in place in the financial sector, that is the concept of "Legacy System". These older systems are normally considered as a liability for traditional companies that would offer them less flexibility when competing against the "flexible" Fintech. However, when there is no possibility to change these older systems, companies still need and want to introduce the latest technological developments or trends in their processes.

---

[1] Anderson, R. "*Security Engineering*". P. 25

[2] Schueffel, P. "*Taming the Beast: A Scientific Definition of Fintech*". P. 45.

The roles of technology mentioned in the paragraph above: Automatization, extraction of relevant information from data and a supportive role in decision making processes, are highly related with Big Data Analytics (Provost, F. 2012). Dealing with a high volume of data could lead business users to not perceive relevant information. In the financial sector, due to the rigidness of the transaction processes, not being able to effectively spot and diagnose faults in their processes affects the reliability of the company and the trust of their users. Therefore, improving fault management or incident management processes from a data driven perspective would help the company to be more efficient in their daily operations and help them sustain a business-as-usual environment.

This paper proposes a state-of-the-art data decision making procedure and is automatically applying it to a young financial institution part of a traditional automotive company in order to improve their ability to spot failure trends and ultimately be more efficient in handling faults in a highly dynamic environment. The next paragraphs shortly discuss an introduction on the research question, methodology, relevance and scope of this thesis to provide an overview of the research performed.

1.1 Background

To properly understand this research there is the need of briefly clarifying the typology of the company and the intended objective of the tool developed.

The company is a young financial institution based in Luxembourg that manage payments transactions for a large corporation in Europe. The company is legally considered as an E-money Institute that acts as a Payment Service Provider (PSP) settling and allocating transactions in their merchant's accounts. The company aims to alleviate the cash flows potential problems as well as providing an e-commerce standardized tool for the merchants to capture the transactions of the customers.

In this environment there was an important event that did shock the banking sector in Europe during the end of December 2020 and the beginning of August 2021. The implementation of the new Payment Service Directive (PSD2). This directive obliges the different actors in the financial market to adopt new measures in order to ensure the application of new technical protocols based on the Strong Customer Authentication (SCA).

Taking into account this situation, the aim is to provide a tool to the Company that would effectively classify these faults in their transactions processes to effectively spot and extract explanatory information from them. These insights are ultimately focused on providing relevant information to the operations teams in the company to address incidents and faults as quickly as possible to ensure a "business-as-usual" environment throughout a new technological implementation.

1.2 Research question

As briefly introduced before, business intelligence would typically focus on generalizations rather than exceptions in order to gain knowledge from data. Other approaches such as failure diagnosis would ultimately be more focused on spotting and dealing with these exceptions. Normally one expects that failure diagnosis precedes the business intelligence analysis. (Kavulya, S. P. 2012). In this research the precedence is clear, based on the need of the company, a Payment Service Provider in Luxembourg, to understand the reasoning of receiving erroneous responses from transaction processing. These standardized error codes of payment transactions, based on the sample database that was analyzed, do not offer enough insights to understand why transactions failed in the context of the newly implemented risk-based authentication methods. Therefore, being able to categorize these failures by commonly shared characteristics shall help to understand where (in which markets) these rejection rates are higher and why they are higher. For example, Denmark may have higher rejection rates than Germany, this proposes that the reason for errors may be related with the characteristics of the country and that Germany performs better than Denmark. However, when further analyzing the data and adding business context, one is able to notice that the error codes from Denmark correspond to a more strict implementation of the new European directive. Consequently, Denmark may well outperform Germany once this country is obliged to implement them too. At the same time, the data reveals that Germany has a significantly higher rejection rate than Denmark when using a specific payment process flow. Thus, the outcome can reveal an important process issue between the payment system of the company and the issuing banks in specific countries,

The insights that are gathered by the application of business intelligence approaches such as OLAP or data mining, may as well support the diagnosis of the failure for rigid

processes, such as financial transaction processes, in an automatic way. Therefore, this thesis has the research question (RS):

How to develop and evaluate models for data-driven failure diagnosis?

Before being able to explore the different perspectives that a data-driven approach could offer to support failure diagnosis, one needs to define what exactly is understood as "failure diagnosis". Therefore, the first relevant sub question is:

1. *What is Failure Diagnosis*?

The next step is to understand the context of this research, especially the artifact, where the research took place. The thesis then reviews the concepts and possibilities of different models, both predictive and exploratory that support failure diagnosis to answer the next sub question:

2. *What are appropriate models for data-driven failure diagnosis (dd-FD)*?

This application of data driven decision making applied to rigid processes such as transactions in a financial institution is not focused on the prediction of failure rates but rather in how these errors are connected with each-other in order to discover interesting paths for operations teams to focus further investigations. Therefore, a relevant sub-question is:

3. How to develop an explanatory model for data-driven FD for data from a company in the financial sector?

Once relevant data points are retrieved according to the specific research environment, to be able to provide useful insights for the company there is the need of effectively running analysis taking these variables into account. Therefore, this research focuses on the identification of a suitable model and test its usability and fit for purpose to better understand transaction failure by answering the question:

4. How to evaluate and interpret these models?

1.3 Research Relevance

As introduced before, this study allows the company to automate a classification of data that provides relevant insights of their raw data. By automatically classifying their error codes and the obtainment of basic insights reports the company can avoid time consuming processes such as systematic navigation through data in order to find details on low succession rates of transactions. At the same time, this solution offers a clearer picture of the company's current situation as well as offering insights of data that could easily be compared through time. Consequently, this solution is a more efficient and less time consuming alternative to standard data analysis processes that are already in place.

From an academic point of view, the application of the rigid scopes from industrial production environments when dealing with faults or incidents in an IT system in order to support decision making is considered to be a literature niche. Therefore, this innovative approach not only offers a case study using up-to-date technologies, but it also increases the available literature corpus. The research deals with insightful workarounds when dealing with NP problems.[3] such as determining an optimal decision tree and, specifically in this case, extracting knowledge from a decision tree considered to be of high data volume with a relevant number of multivalued categorical variables.

Finally, this research serves as a case study on how to apply predictive methods to failure diagnosis, as an exemplification on how the application of flexible approaches could be successfully implemented in rigid contexts and processes with descriptive or exploratory purposes.

1.4 Research Methodology

This research uses a design science research methodology to ultimately develop an artifact that assists the company to achieve their objectives of obtaining insights from their

[3] A non-technical definition of complete NP problems are basically those for which, based on a given dataset, the processing power required to solve a mathematical problem increases exponentially in relation with increments in the inputted data. Few examples would be finding an optimal decision tree or extracting the random walk centrality measures of a huge graph. The P vs NP is one of the well-known “The Millennium Prize Problems” from the Clay Mathematics Institute of Cambridge, Massachusetts (CMI) and there is plenty of available information about them on the web.

data. Consequently, in order to extract these insights, the research is following the CRISP-DM approach.

1.4.1 Design Science

The main objective is to automate the process in order to effectively offer insights to the business. This research follows the framework and terminology provided by Alan R. Hevner et al. (2004). It incorporates the environment of the research, especially the structure of the organization and the technological infrastructure of the company to develop an artifact that can easily be used by the company. In addition, it considers different frameworks and tools such as basic statistics, decision trees and event trees as its knowledge base. The analysis uses only open-source software, runs in python environments and is thoroughly documented to ensure replication of the processes performed by the artifact. Finally, the thesis evaluates the solution not only by using statistical measures, but also by receiving feedback from experts in the field.

1.4.2 CRISP-DM

We use this process to understand and develop our solution. The first step of our framework corresponds with the business understanding part, where the objectives and needs of the company are described, as well as the goal of our project. Secondly, the typology and quality of the data that we are working with is understood. In the following chapters we describe the data preparation and modeling steps. Finally, we evaluate the result of our analysis.[4]

1.4.3 Predictive and Descriptive Methods

During our data mining modeling in order to create a powerful artifact to classify errors in the company in an automatic way. We are using technologies related to machine learning and data interpretation techniques using python libraries. The following methods are introduced in our analysis: Binary Classification Trees, Logistic Regressions, Chi squared correlation, Gradient Descent, Bayesian Boosting and basic statistics considering our database as a matrix.

[4] CHAPMAN, P.(et. al) "*CRISP-DM 1.0 - Step-by-step data mining guide*". SPSS. 2000.

1.5 Scope

The scope of this project is delivering an artifact that allows the company to automatically classify the error codes of the transactions in a way that is easily understandable for a specific day. Running this analysis on a daily basis allows us to produce comparable data that could easily be interpreted using most data visualization software. It is not in the scope of this analysis to automatize that visualization but to automatically procure the relevant and transformed data to effectively do so. The evaluation of the artifact would not extremely focus on the extreme reliance of the algorithms but on the utility of them for business intelligence purposes. Therefore, effectively dealing with the NP problem nature of decision trees in order to provide insights for the diagnosis of failed processes is the main objective of the project.

1.6 Thesis Outline

The thesis is structured as follows. In Chapter Two, the researcher explains what exactly “Failure Diagnosis” is to effectively understand the objective and implications of this research. After that, in Chapter Three, the different implications of the models and methods used for failure diagnosis are explained. This chapter specially, emphasizes the difference between predictive and descriptive models, along with an understanding of the theoretical principles of the tools that are going to be used in this research. In the fourth Chapter, all the gathered knowledge is applied to the business environment to effectively develop a tool for automatically classifying errors based on the knowledge extracted from the database at hand. Finally, the conclusion summarizes the findings and evaluates the effectiveness of the designed tool.

2 INTRODUCTION TO FAILURE DIAGNOSIS

This chapter provides an overview on failure diagnosis by comparing relevant literature on related subjects such as fault and incident management, failure definition and diagnostic and finalized with a summary on a comparison of failure diagnostic approaches.

2.1 Fault Management and Incident Management

Analyzing literature related to fault management and incident management, one observes that depending on the focus of the research, either managerial or technical, researches articles chose to use the concept of "incident" or "fault" respectively.

The IT infrastructure Library (ITIL) 4 defines "incident management" as:

"*The practice of minimizing the negative impact of incidents by restoring normal service operation as quickly as possible.*"[5]

ITIL does also offer a definition of "incident" as:

"*An unplanned interruption to a service or reduction in the quality of a service.*"

Therefore, the main objective of incident management is "minimizing" the effect of interruptions or reductions in delivery of a service to ensure business as usual operations.

According to the ISO/EIC 7498-4: 1989 for Information processing systems, specifically Open System Interconnection (OSI), fault management is understood as a process "*encompassing fault detection, isolation and the correction of abnormal operations in the OSI environment*". While faults, either persistent or transient, cause "open systems to fail to meet their operational objectives". Thus, the focus is on the recovery of proper functionality of the communication in a system rather than in ensuring a quick recovery of business-as-usual operations.

[5] "*ITIL Foundation. ITIL 4 Edition."* Axelos Global Best Practice. 2019.

When incidents and faults are transient, they are considered as isolated events. Such as, a Denial of Service (DoS) Attack that impedes the customer to complete the purchasing journey. However, when these events are persistent in time with the same cause, it results in problem management or failure management. ITIL does define problem management objectives as being able to identify root causes and try to prevent incidents in the future. Therefore, it has two approaches: reactive and proactive. Failure management does not necessarily need to be proactive in analyzing trends of faults or incidents and assessing if these incidents could recur. However, as fault management does generally comprise *fault detection, fault diagnosis and fault recovery* (Lewis, L. 1993), the conclusions extracted from the failure diagnosis process can support to base any future intent of assessing the likeliness of incidents to reoccur. Especially, if the historical data extracted from the failure diagnosis process is properly stored and accessible.

Both failure detection and failure diagnosis are the basis of the concept of the research approach discussed in this thesis. Even if the detection concept is straightforward, the diagnosis definition becomes a tricky concept as it implies a causality relationship between fault and failure such as Dr. Kavulya (2016) defines it:

"*the process of identifying the causes of impairment in a system's function based on observable symptoms, i.e., determining which fault led to an observed failure. Since multiple faults can often lead to very similar symptoms, failure diagnosis is often the first line of defense when things go wrong a prerequisite before any corrective actions can be undertaken*"

In this paper, the terminology of fault and failure management is used. However, these concepts can be replaced by incident and problem management respectively. Further on in this chapter, data driven approaches are discussed on how these lead to automatic detection and diagnostics of failures or incidents in a system to effectively support decision makers offering relevant information to act based on previous research. Before these approaches are discussed, a definition of failure is provided for our specific context.

2.1.1 Our Definition of Failure

Failure, error and incident, all of these terms refer to the same idea: A process did not finish as expected. However, as Sebastian Kunert, author of "Strategies in Failure Management Scientific Insights, Case Studies and Tools"[6], states:

"*the closer you get to the concept of failure, the more it becomes fuzzy, blurred, and difficult to grasp.* (*it becomes*)"

Therefore, "failure" in the environment in scope needs to be defined first. Considering the characteristics of the sector and the company (a financial service provider part of a big European automotive group), a failure is defined as:

"*A failed subprocess in a transaction process*"

The concept of failure does not refer to natural disasters or bad luck. It is a subjective concept that requires individuals to assess the severity of the fault. As Kurnet, S. 2018 does mention: Failure does require "individuals, who set goals, make plans, act and evaluate."[7] For example, one observes an elevated number of failed calls to run a certain process in an IT system, this characteristic does not offer enough information to assess if it is effectively a failure or if in this case it could be an asynchronous process that generally returns failed requests. Therefore, the research's definition of failure is modified to include this characteristic:

"*A failed subprocess in a transaction process, that interrupts the normal or known behavior of the transaction*"

Acknowledging the relativeness of failures in a system requires knowledge and information to effectively benchmark it. This is one key concept that differentiates "fault" from "failure" in the same way as "incident" is differentiated from "problem" (ITIL 4), as problems are the cause of incidents . For example, if an asynchronous process it's been

---

[6] "*Strategies in Failure Management Scientific Insights, Case Studies and Tools*". KURNET, S. (E.). Springer. Humboldt-University Berlin. Berlin, Germany. 2018. Pg. 2.
[7] *Ibidem*. Pg.5.

triggered unsuccessfully several times, we could encounter from the perspective of the business user that there is a certain process with an elevated fault or error rate. However, if the presence of these failed calls is expected in a business-as-usual environment, it would not become a failure as it does not interrupt the expected behavior of the process.[8]

This research focuses on IT-systems and uses the explanation of IT-systems used by Anderson in his book "Security Engineering" (2008). A system could be at the same time:

- Non-human intervened systems:
    - A technical component;
    - A set of technical components with an operating system (the base of the organization infrastructure);
    - The above set of components with one or more programs running.

- Human intervened systems:
    - Any not human system plus IT Staff;
    - Any not human system plus internal users and management;
    - Any not human system plus customer and external users.

In this research, we look at human intervened systems in the case study represented in Chapter 4. The type of system into consideration incorporates the human factor for different reasons:

- Taking into account customers and business users it is required in the chosen framework for value creation that it is explained in the following section.
- Customers and the unproper alignment between different payment providers is a known source of errors.
- Even if the communication among the different non-human intervened sub systems is automatic. For example, after the order is created a request is created to

[8] Synchronous processes do follow an individual thread. The next process would not start till the previous one is fulfilled. However, Asynchronous processes are multithreaded. Therefore, processes that could run in parallel, such as when you are playing a shooter video game and you are able to shoot and move at the same time. Therefore, following the example provided, even if you trigger the "shooting" process by pressing a button in your controller, there is the need for the process "being alive in the game" to be able to actually run that process. Therefore, If you are not able to "shoot" in the menu screen it would not be considered a fault but the expected behavior of the program.

the billing engine. When the system does confront a problem or failure, human intervention is needed to mitigate the negative effects.

In line with this idea of different end-users' behavior, there is the fact that the company financial system is incorporated in a network of tightly coupled systems. The payment system is integrated in two ways: Front End Side to the web shops of merchants (the company's business customers) as well as directly to the end-customer over the wallet user interface and Back End Side to their external providers such as the payment gateway, the acquiring bank, PayPal etc. This means the payment platform acts as a medium between various external and internal systems. Concrete examples are provided in the case study in chapter 4. However, as an initial idea, the flow of a transaction starts on the merchant's side when a customer is trying to purchase a service or a good. At this point the customer is redirected to the company platform to where the processes of "payment authorization" and "capture" of the transaction take place in close collaboration with the external systems and the merchants' side. The company has more than 50 web shops to whom they provide services to and 12 external systems integrated in their back end.

Consequently, this means that even if the faults that are identified in the payment system are not statistically relevant for the company, there is the chance that one of the merchants is not able to process payment transactions at all. That edge failing in this chain of systems may not lead to a systemic failure of the company. However, it may lead to significant damage to one node of the chain, in this case the merchant. For the merchant this results in losing customers and revenues.. Considering all the above, this thesis finally uses the following definition of failure:

"A failed subprocess within the processing of a payment transaction that interrupts the normal and known behavior of one or several nodes within the payment processing chain."

2.2 Data Driven Failure Diagnosis (dd-FD)

Process Monitoring and Failure Diagnosis (PM-FD) has been an active research field as it is relevant for many industrial applications.[9] It is mainly used in industrial scenarios to ensure that the processes are operated properly spotting abnormal behavior or variations in industrial plants (Mei, J. 2014).

According to the classification PM-FD proposed by Mei, J. 2014, there exist two differentiated categories: Model driven approaches and Data driven approaches. Model does rely on a priori knowledge of the environment. However, the data driven approach does only rely on measuring the data that could be extracted. Specifically, Data Driven PM-FD algorithms detect and diagnose faults by measuring data normally on real time from different sensors and components.[10] (Mei, J. 2014). Therefore, there is a clear scalability advantage compared to the model-based approach.

One of the assumptions of this paper is that this industrial approach to failure diagnosis could be effectively extrapolated to other sectors, in which the scope is based on reliability and ensuring quality, when providing a service. When dealing with complex and highly coupled systems, ensuring an efficient and effective communication within the system became both a technical and governance endeavor. The investigations on High Reliable Organizations (HROs) offer companies a set of common characteristics to be able to effectively manage and operate in complex systems[11] from a governance point of view. An example of these characteristics is: a system of rewards for discovering errors, the respect and nurturing of the workers' skill set and the delegation of authority in stressful situations based on their expertise. Therefore, from a more technical point of view, the methods that belong to the PM-FD are useful for organizations with a higher fault tolerance than industrial plants to excel in their treatment and coping techniques with faults and incidents. Applying this restricted view on fault tolerance and the methods for failure diagnosis to analyze the performance of the payment system in scope is expected

[9] MEI, J. (et. Al)"*A Novel Data-Driven Fault Diagnosis Algorithm Using Multivariate Dynamic Time Warping Measure*". 2014.
[10] Ibidem.
[11] G.I. ROCHLINL "*Reliable Organizations: Present Research and Future Directions*". 1996

to increase the time to spot, understand and deal with failures in the tightly coupled network of the company that collaborates to process a transaction in a better way than the traditional ticketing systems in place in the company.

The research chose the dd FD as the extraction of information is performed from an available data set. Also, the research is based on the application of predictive technologies to increase the identification and understanding of faults and incidents in the system of the company. However, this research does not focus on a component scale interaction where the failure root tends to be technical, such as in production plants, as in the system in scope there are users and other actors in the network chain that cannot be controlled. Therefore, traditional industrial approaches do not completely apply to this research and it is at the core of this paper to effectively apply approaches in a more flexible environment. As an example, a user is trying to purchase some services in the payment platform, he/she is redirected to his bank to insert the payment credentials, in the case that the user inputs the wrong data the transaction process would collapse not due to a technical failure but to a human error. This would indeed be an incident and in the payment system, it is presented as a faulty transaction.

The industrial approach of dd FD is not strictly able to deal with external users´ faults as described above as they rely on independent and isolated technical structures under the control of the operators. However, the methods used for this system remain useful for identifying the source of the errors and possibly some explanatory characteristics of these faults if the data exchanged between the different actors in the network is structured and standardized. In the same way as sensors and components exchange data in a production plant.

Actually, these automatic fault detection and diagnosis tools are expected to be used by big corporations and service providers include them in their tools such as the ITIL 4 describes it:

"*Modern IT service management tools can provide automated matching of incidents to other incidents, problems, or known errors, and can even provide intelligent analysis of incident data to generate recommendations for helping with future incidents.*"

Further in this chapter we combine the industrial dd-FD approach and business data driven approach. However, to effectively accomplish this objective in the next section notions regarding the creation of value for Big Data from a business point of view is discussed.

2.2.1 A Data Driven Approach

According to the general literature, there is no creation of value for a company if data is not interpreted in its context and information is extracted from it. This research used the definition of Data Science as: "*a set of fundamental principles that support the principled extraction of information and knowledge from data*"[12] (Provost, J. 2012)

This definition stresses the importance of information and knowledge in data science, however it could be misinterpreted depending on the definition of what we understand as "data", "information" and "knowledge". Therefore, this thesis uses the following definitions provided by Aamodt, A. (1995) for these terms:

- Data: Is a syntactic entity with no meaning. It would be the input to be interpreted. In this case, a single transaction entry with its attributes.
- Information: Is the interpretation of the data in its context. It is at the same time the output of the data interpretation process, as well as the input to the knowledge-based decision-making process. An example suffices to illustrate the term in our context: Specific operations have been performed in the database and it encounters a combination of attributes that corresponds to a rate of unsuccessful transactions. That would be the information extracted from data.
- Knowledge: Is the information incorporated into our body of knowledge. Ready to be used in accordance with our resources and experience with the ultimate objective of guiding and supporting the knowledge-based decision-making process. Therefore, and following our former example, the relevance of this combination of variables is assessed when it is compared to our historical data and the benchmarks extracted from it. Therefore, the information is interpreted according to the

[12] PROVOST. "*Data Science and Its Relationship to Big Data and Data-Driven Decision Making"*. 2012

existing body of knowledge, and it is relevant if the rate of unsuccessful transactions is significantly higher than the average rate of the benchmarks.

The connotations of the terminology data, information and knowledge needs to be complemented by the process on how value can be extracted from data. The well-known 3V's of Big Data[13]: Volume, Velocity and Variety, does correspond to the nature of the data we are handling in our research. The variety dimension does not fully represent the nature of the database in scope as it is dealing with financial transactions. However , the extent of the database and the need for faster analysis to improve efficiency of the operations, does clearly relate to the concepts of Volume and Velocity. This definition of Big Data, when the Variety dimension does not necessarily define the characteristic of the dataset, is compliant with the definition used by Chen, H. (2012) when referring to the concept of "(Big) Data Analytics".

To effectively extract value from the data in scope, the first step is to understand the needs for the business. Following the big data value creation framework provided by Verhoef P. C. (2016), in its essence the value that data analysis brings to a company is considered as "Value to the firm" and "Value to the customer".

- Value to the customer: In the case of this research, the customers are merchants (business-to-business customers) and end-users of the company, and a quicker problem resolution/identification of a problem increases not only their satisfaction but also avoids economic loss to our merchants. Since the company is a payment provider.
- Value to the firm: The ability to easily spot error paths allow the company to understand changes in the electronic payment environment. With this the company can proactively address customers before problems escalate and ultimately increase the satisfaction of their customers.

According to Provost (2012) and in consonance with his definition of Data Science, Data Mining is the actual extraction of knowledge from data. In addition, the tools assisting in this data mining endeavor are numerous. Therefore, the next paragraphs provide an

[13] Taylor, Cowls, Schroeder, & Mayer, 2014; Leeflang et al., 2014. Mentioned in "*Creating Value with Big Data Analytics: Making smarter marketing decisions*". Verhoef, P. C. (2016)

overview on the state of the are methods for Business Intelligence & Analytics (BI&A). Based on the review of Chen, H (2012) of the evolution of BI&A during the last decades, there are three types of BI&A:

- BI&A 1.0: Technologies currently adopted in the enterprises, using more structured data. The foundational technologies used in this typology of BI&A are mostly related with database management, such as: Extract Transform and Load (ETL) processes, online analytical processing (OLAP), data mining, statistical analysis and Dashboards.
- BI&A 2.0: It is more closely related with web-based sources of information and the data analyzed is more unstructured than before. The technologies used in this typology vary from Social media and network analysis to opinion mining and question answering.
- BI&A 3.0: It is related to a usual suspect in data analysis, the "internet of things", based in mobile and sensor-based content.

Due to the characteristic of the database in scope and the intended analysis regarding the aim to detect anomalies/interesting patterns, the scope can be categorized in BI&A 1.0. The research framework provided by Chen in the same paper mentions emerging research trends for "(Big) Data Analysis": statistical machine learning, sequential and temporal mining and process mining for example. Therefore, in order to extract value from the analyzed database, the researcher executes an analysis incorporating the intrinsic characteristics of the data. To accomplish the objective of delivering value both to the customer and the firm, the research uses state of art methods, such as machine learning algorithms for categorical variables and ETL, as well as well-known and tested data mining methodology: CRISP-DM.

2.2.2 Dd-FD and the value creation

ITIL 4 reviews how incident management supports the value chain activities similar to the data driven approach that creates value both for the customer and the company. In this section it is proposed a combination the industrial dd-FD approach and the business data driven approach focused on creating value based on the ITIL 4 framework on how incident management supports value chain activities as follows:

- Improve the service quality: Incidents or faults are key data inputs to improve activities of a company. These effects are especially visible in an operational environment.
- Engage key actors in the system: The different actors in the tightly coupled networked system should establish an efficient communication to cope with faults or incidents effectively. To be able to quickly communicate with the specific actors we need to comprehend the root causes of the faults or at least know in which part of the process or market these errors are occurring more frequently, to know to whom to communicate.
- Design and build: both in test and operational environments, the analysis of the faults and incidents assist a company to assess which "components" in their systems are more prone to faulty behavior and act in consequence.
- Support and effective delivery of the service: Just like traditional ticketing systems, fault detection and diagnosis support resolving incidents as well as faults in a timely manner, and even being able to detect potential failure issues before the customers' claims arrive to the operators

2.3 Overview of approaches supporting failure diagnosis

There are multiple examples in the literature of data driven approaches to incident management. However, as mentioned before the trend is not to use the nomenclature failure management but problem or incident management. In this section the different papers presented deal with the extraction of information and knowledge from data supporting the identification of faults/incidents as well as their root causes.

An early example of this approach is the article from Lundy Lewis in 1993[14] based on the classification of tickets attending to their shared similar attributes. This approach effectively defines paths where the processes of a certain company are compromised and offers help desk workers with information enough to draw causal conclusions. However, the artifact developed does not detect the faults, but rather reflects a knowledge base of previous customer complaints or support requests.

[14] LEWIS, L. (et. Al)"*Extending Trouble Ticket System to Fault Diagnostics*". 1993.

Indeed, supporting trouble-ticket systems, incident management solutions or simply "tickets"[15] are one of the most influential outcomes of the artifact. Both from a preventive perspective, being able to realize which set of incidents or failures are going to be dealt with before the ticket arrives to the business users; and a corrective perspective, being able to spot commonalities of the tickets in an easier way. Bartolini's paper "*SYMIAN: Analysis and Performance Improvement of the IT Incident Management Process*"[16] establishes a consistent overview, on a simulator approach to detect the performance of IT systems in a company with his data driven model approach, as well as on how ticket systems work in "real-life" organizations: one helpdesk entry point, different sets of skills of individuals, ticket escalations… All these characteristics imply that normally, there exists a considerable extension of time till relevant tickets arrive to the people that are qualified to solve them or that have the clearance enough to make decisions and address relevant problems that may be detected during the analysis of these tickets. This time gap between receiving a request and treating a request appropriately can impact the customer volume. Consequently, this shows how a data-driven solution supporting failure diagnosis can benefit both the end-customers and the company to create value, as the customers´ incident requests are treated in a shorter period of time.

Other approaches focus on machine learning methods to extract information from a raw data source , such as the decision tree for failure diagnosis that systematically identifies errors from the logs of a company in a predefined classification of error typology (Chen, M. 2004)[17]. Even if this approach is consistent with the approach in scope and offers numerous insights on how decision trees can support failure diagnosis, this paper is mainly focusing on technical faults and the pre-classification of errors based on technical faults. In addition, this approach is mainly focusing on offering insights to system end-users in the company to effectively draw conclusions based on understandable concepts for business experts. Therefore, the categories used are more related to Business Intelligence traditional variables, such as market location, merchant, payment type or currency, rather than, single host fault, software bug or database fault.[18]

---

[15] Bartolini, C. "*Analysis and Performance Improvement of the IT Incident Management Process*" 2010 pg.133
[16] Bartolini, C. "*Analysis and Performance Improvement of the IT Incident Management Process*" 2010
[17] CHEN, M. "*Failure Diagnosis Using Decision Trees*" 2004
[18] Ibidem. P. 4.

Other studies, which are more technically oriented offer data mining models based on binarization of variables and modeling techniques such as the research of Kwon, J-H (2017) “Association Rule-based Predictive Model for Machine Failure in Industrial Internet of Things”. Again, it is observed that the typology of these researches is industrially oriented. These are further discussed to retrieve additional understanding of fault and incident in the following chapter.

The different studies mentioned and the ones that would be further discussed in the following chapter are summarized in the table below with their main characteristics:

Table 1 - Literature Review

Tittle	Methods	Main findings or insights	Author
Extending Trouble Ticket Systems to Fault Diagnosis	Rule base Reasoning Case Based Reasoning Fuzzy Logic Framework	Trouble Ticket Systems (TTS) information and its analysis could be used to provide insights into failure diagnosis of network problems and behavior	LEWIS, L. DREO, G. (1993)
Failure Diagnosis Using Decision Trees	C4.5 Decision Tree with a tailored "pruning" for finding error paths vs Association Rule mining	Decisions trees are applicable for the task of failure diagnosis. While there are other classifiers with perhaps better failure prediction performance, decision trees return easily interpretable lists of suspicious system components that would allow us to define error paths	CHEN, M. ZHENG, A. X. LLOYD, J. JORDAN, M. I. BREWER, E. (2004)
SYMIAN: Analysis and Performance Improvement of the IT Incident Management Process	Statistical analysis and inference. Building a stochastic transition model-based on escalation of tickets between different support groups.	The performance optimization of large-scale IT support organizations can be extremely complex and lends itself to being tackled with decision support tools. open queuing network models could reproduce the behavior of real-life IT support organizations with a very high degree of accuracy. The SYMIAN decision support tool could be exploited by commercial applications.	BARTOLINI, C. STEFANELLI, C. TORTONESI, M. (2010)

Association Rule-based Predictive Model for Machine Failure in Industrial Internet of Things	Association rule-based predictive model	Three major steps were considered in developing the model: 1) binarization, 2) rule creation, 3) visualization. The implementation verified that the proposed predictive model for machine failure can provide realistic prediction results	KWON, J-H. LEE S-B. PARK, J. KIM, E-J. (2017)
A Novel Data-Driven Fault Diagnosis Algorithm Using Multivariate Dynamic Time Warping Measure	Multivariate Time Series (MTS) Tennessee Eastman (TE) benchmark	One contribution of this paper is that it firstly uses MTS pieces to represent the measurement signals as dynamic features in the measurement signals can provide more information than static features.	MEI, J. HOU, J. KARIMI, H. R. HUANG, J. (2014)
A comparison study of basic data-driven fault diagnosis and process monitoring methods on the benchmark Tennessee Eastman process	PCA, ICA PLS, FDA (Fisher Discriminant Analysis). TE benchmark.	The design parameters in PCA, PLS and ICA related approaches will (considerably) influence the PM–FD performance. different results can be obtained according to different values of N. In practice, the large scale industrial plants are generally com- plex dynamic systems and the process measurements will not strictly follow Gaussian distribution as shown in TE process.	YIN, S. DING, S. X. HAGHANI, A. HAO, H.ZHANG, P. (2012)

3 METHODS FOR DATA DRIVEN FAILURE DIAGNOSIS

In this chapter the state-of-the-art dd-FD methods are discussed, by introducing them in a business environment, thanks to data mining, and by explaining the chosen relevant methods for developing the tool in the business environment in scope. A review from a technical point of view of the current literature supporting failure diagnosis is also provided as closure for this chapter.

As already mentioned, dd FD is mainly industrially focused. The methods most widely used for fault detection and diagnosis are Principal Component Analysis (PCA), Partial Least Squares (PLS) and Independent Component Analysis (ICA), see (Yin, S. 2012. Mei, J. 2014. Qin, J. S. 2009 and Nor, N. 2018.) All these data driven techniques are focused on assessing confidence intervals and establishing thresholds to effectively detect failures in complex systems using statistical tools that belong to the concept of Exploratory Data analysis (EDA). However, other tools such as Neural Networks, Decision Trees or Association Rule mining could be employed to find those components or relationships relevant to spot and diagnose faults in a system, see (Liu, W. 2018. Kwon, J-H. 2017. and Nor, N. 2018.). This distinction could be further understood and corresponds to the approach of the researcher or the company when dealing with fault in a system, either if the objective of the research is descriptive/explanatory or predictive. From a more managerial point of view, we could further connect this industrial fault management approach with the needs of the company by recurring to the most common data mining methodology: CRISP-DM.[19]

Data Mining integrates different methods to be able to classify large amounts of data from different fields. such as Machine Learning, statistics or pattern recognition.[20] Depending on the scope of a study, exploratory or predictive, these tools can be applied differently:

- Being the main objective predictive, the models developed are focused on providing an "output" that could be assessed as accurately as possible. A feasible example would be a machine learning model that is built to predict the value of the

[19] CHAPMAN, P.(et. al) "*CRISP-DM 1.0 - Step-by-step data mining guide*". SPSS. 2000.
[20] SUJATHA, M (et all.) "*A Survey of Classification Techniques in Data Mining*". Pg.7

square meter price in a city depending on the input values related with the characteristics of the house, the location of the house in the city… In this case, the evaluation of the tool, taking into account the mentioned objective, is not excessively related to assessing which variables are more relevant in our model. However, the success of the algorithm would be evaluated according to the capacity to predict as accurately as possible this price.

- The same problem from an exploratory perspective would shift the interest in discovering which variables strongly determine the shift on these prices. Not necessarily due to a causality relationship. However, with enough evidence to establish statistical significance. Same as the well-known case of "red meat" producing "cancer"[21]. There is not a causality relationship between these variables. However, people that normally do consume red meat on a regular basis may have other behaviors associated that allow the authors of this study to assert that there is a statistically significant correlation between chances of having cancer and usually eating red meat.

Both approaches could effectively support decision making. The descriptive example would support understanding if it is a good time to invest in real estate. The exploratory insights offer information on how to maximize an investment if we are planning to sell our home, for example deciding to repair the house or divide a big bedroom into two.

Most of the tools used to develop studies supporting these different approaches could effectively be combined to develop consistent models independently of the ultimate objective. For example, in order to choose the variables relevant to a decision tree that aims to predict which customers are more likely to abandon a company an Exploratory Data Analysis to search for the correlation of the input variables becomes relevant to decide which variables to keep in our model, as for classification or regression trees input variables are supposed to be independent from each other.

In the next section it is briefly described the technicalities of the problem it is dealt with in this research. Further explanation of the theoretical implications of these characteristics are presented as follows:

[21] FERGUSON, L. R. "*Meat and Cancer*" (2020)

- Model for explanatory purposes (Section 3.2).
- Models for predictive purposes. (Section 3.3).
- Evaluation of performance measures. (Section 3.4)

Finally, this chapter is concluded with a discussion and comparison of the different approaches to failure diagnosis with the research, based on the available literature.

3.1 Technical Description of the dd-FD problem

During this section it is introduced the assumptions and technicalities of the models implemented in the research.

The data corresponds to and extracted dataset from the payment gateway of the company, where most of the transactions are processed with the exception of SEPA payments, that follows a different payment flow, are not included for privacy reasons and are gathered and processed in a different payment gateway than the one from which the data was extracted from. The dataset used for this research has a fault rate of 8.17% and a total number of entries 15246 taking into account Credit transactions and PA and CP subprocesses. Once we had properly filtered the data that represent unique transactions, the new fault or error rate is 13.72% and the number of items is 8861.

From a general point of view the proposed tool does classify the mentioned data in two differentiated steps that effectively combines the predictive capabilities of a model that assess the importance of the dependent variables to assess if a transaction it is likely to fail or not, with the simplicity of looping through the mentioned dataset to retrieve the most relevant paths. Therefore, we are using the predictive power of supervised machine learning algorithms to retrieve important values in the dataset that would offer insights from a descriptive point of view about the commonalities of faulty transactions in our dataset.

Regarding the first part of the analysis, the objective variable could only have two values: Being “1” a failed transaction and “0” a successful one. Therefore, the problem for the first part of the modelling is a binary classification one. The models chosen in the research are both supervised machine learning algorithms known as Binary Classification Tree and Binary Logistic Regression. However, as mentioned before, the classification tree method does present a NP problem when trying to determine which is indeed an

optimal classification tree. Therefore, a library that implements a gradient descent and that assigns weights to the different values of the dependent variables in a process known as boosting, has been selected to overcome this issue. This model produces 1000 binary classification trees every time it is run and chooses the model that offers a higher degree of predictive precision based on the evaluation metric that the business user chose. In this case either F1 or AUC "Area Under the Curve". This process is known as cross-validation.

Regarding the dependent variables, the dataset consists of mostly multivalued variables that correspond with characteristics of the transaction as it is further explained in the case study. It is relevant to mention that only the variables related to the amount of the transaction present a continuous typology. It is also important to mention that, even if both models assume that the dependent variables are independent from each other, these variables present a certain correlation between each other after the dataset has been properly cleaned and the researcher chose to maintain some of them attending to:

- Their business relevance based on the descriptive potential of the variables that ultimately are inputted in the classification tool.
- The researcher also assumes that introducing this characteristic against the backbone assumptions of both models creates an opportunity to assess which model does provide better insights on our data.

In relation with the combination of both approaches, the predictive and the extraction of the failure paths. The researcher assumes that for rigid processes that are not prone to fail, the expected error paths present a significantly higher failure rate than the overall dataset. Therefore, looping through the database based on the values that represent the importance of the features to predict if a transaction would fail or not, while establishing a fault threshold significantly higher than the average of the dataset would suffice to extract most of the relevant error paths. Regarding the potential of this tool, it is also worth mentioning that it supports the ability to use parallel processing and it does not depend on a predefined fault frequency or distribution to extract these error paths.

3.2 Descriptive or Exploratory Models

It is useful for understanding what is EDA to explain it along the concept of Confirmatory Data Analysis (CDA):

- CDA is a "*set of statistical procedures aiming to confirm a pre-formulated hypothesis using either p-values or confidence intervals*".[22]
- EDA's primary objective is to discover hypotheses about the relationships between variables, based on the concept of "detective work" developed by Tukey (1977).

Therefore, in EDA the researcher takes the position of a detective trying to find those "clues" while investigating a dataset. While, in CDA the researcher acts as a judge of the discovered insights.

It is not in the aim of this chapter to review the large set of different tools and methods used for EDA or CDA. However, it is important for the researcher to discuss the implications of the used models in this research and in the similar researches mentioned in the previous chapter:

All the mentioned above basic methods of dd FD: PCA, PLS and ICA are part of EDA analysis, as their aim is to discover relevant paths and reduce dimensional complexity to ultimately help to choose the data that is relevant to be processes in our model, as most of them would need independent variables to be selected. This independence of the variables is a requirement to develop a statistically robust classification and regression trees; and generally, the absence of multicollinearity between predictors is considered a requirement to develop predictive models.

Methods such as PCA and PLS are intended to reduce dimensional complexity in the data set in scope. These methods are intended to be used with continuous variables and not with discrete ones. In this research, categorical variables are used, thus a Multiple Correspondence Analysis (MCA) is appropriate. According to the available literature,

[22] HO YU, S. "*Exploratory data analysis in the context of data mining and resampling*."

there is a significant outcome when determining the correlation among categorical variables as described by Nishishato, S. 2006 when proposing a MCA based on comparing two categorical variables and the span between them. The coefficient of the MCA is the same as the Cramer's V coefficient. (Nishishato, S. 2006.)[23]

Acknowledging the typology of the data is required to properly assess the correlation between different variables. Normally, measures like the Pearson correlation coefficient are intended to be used to check the relationship between variables with continuous values. In other cases, such as in the research, the variables to be analyzed are multivalued. Therefore, the Cramer's V coefficient has been chosen to assess this correlation.

Being X_i and Y_j two different variables for $i = 1, \dots, r\,; j = 1, \dots, k$ and $n_{ij} = \sum of\ the\ same\ values\ of\ X_i\ and\ Y_j\ (X_i\,, Y_j)$, for a n sample, the Pearson Chi squared is defined as:

$$X^2 = \sum_{i,j} \frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}}$$

Therefore, the Cramer's V coefficient is defined as:

$$V = \sqrt{\frac{X^2}{n \min(k-1, r-1)}}$$

3.3 Predictive Models

Based on the different methods reviewed in Liu, W. 2018 and Nor, N. 2018 it could be observed that there are plenty of dd-FD methods that could be effectively implemented following machine learning principles. Liu, W. 2018 offers an initial classification of these methods based on the concepts of supervised, unsupervised and "semi-supervised" dd-FD learning methods. However, most of the methods proposed are based on the analysis of rates or intervals. Therefore, based on continuous variables. However, in the paper

[23] NISHISATO, S. "*Correlational Structure Of Multiple-Choice Data*" P.170-173

of Nor, N. 2018 we could observe that approaches such as decision trees C4.5, genetic algorithms and Bayesian network-based methods have been already implemented in industrial environments. However, there is not a specific model mentioned for dealing with categorical variables.

Therefore, even if there is also an extensive number of predictive methods such as Bootstrapping, Neural networks or weighted moving average, also related with IT systems.[24] In order to choose a set of models we need to clarify the objective variable of the research. If the selected variable is a continuous one, models such as weighted moving average, or single exponential smoothing could be considered. On the other hand, if the objective variable is a categorical one, either binary or multivalued, it is appropriate to choose a model that allows the classification of the output. Therefore, this section would briefly discuss the following concepts and methods dividing them in evaluation methods and algorithms.

Regarding the models that are able to predict a binary classification of their output we could find neural networks, logistic regressions, decision trees or random forest. However, other methods could also be successfully applied such as association rule mining.[25] This last method based on the declaring if-then typology rules for different variables in relation with the studied output is a similar approach as the one we have developed to go through the relevant paths in the last step of our artifact. However, our approach does not intend to be predictive but descriptive in nature. We are going to extend our knowledge of the most commonly used prediction models for binary classification: Binary Logistic Regression and Classification Tree. Then we discuss the nature of underfitting and possible solutions to these NP complete problems. Finally, we explain the theoretical implications of our predictive model of our artifact:

3.3.1 Binary Logistic Regression:

It is one of the most widely used models for binary classification as it does allow to combine different typology of input variables, both binary and continuous. In its core it is

[24] FURTAK, S. "*Sensing the Future: Designing Predictive Analytics with Sensor Technologies*"
[25] KWON. J-H. "*Association Rule-based Predictive Model for Machine Failure in Industrial Internet of Things*"2017

based in a linear function where each predictor or independent variable (X) has a parameter (B) assigned in order to predict the probability of a dependent variable (Y), such as a standard regression model:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$$

However, in this case Y could only have binary values [0,1]. Therefore, the Y would follow a Bernoulli Distribution and we need a logistic function that generates values from 0 to 1:

$$Y = log_b \frac{p(x)}{1 - p(x)} = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$$

This is known as a "logit", and we can extract the odds or probability of this function as follows:

$$p(x) = \frac{b^{\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n}}{1 + b^{\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n}}$$

Finally, the objective of using this logistic regression model is to assess which values of the parameters i maximize the likelihood expression[26]:

$$Y(\beta_0, \beta_1, \ldots, \beta_n) = \prod_{i:y_i=1} P(x_i) \prod_{i':y'_i=0} (1 - p(x'_i))$$

Therefore, the objective is to assess what is the weight for the independent variables to predict the objective variable.

3.3.2 Binary Classification Tree:

One of the most widely used models for prediction and explanatory purposes. Even if it is used for prediction purposes, by plotting the tree we could further understand how these

[26] ABDULLAHI, A. I. "*Comparison of the CatBoost Classifier with other Machine Learning Methods*". 2020.

"decision paths" are generated. Therefore, this is one of the methodologies more commonly found in similar scoped research.

Decision trees are a supervised learning method that could be used for regression or classification. Following the decision tree algorithm from "scikit learn" or "sklearn" one of the most known libraries for python to develop decision tree models. Given a set of training vectors $x_i \in R^n, i = 1, 2, \ldots, l$ and a label vector $Y \in R^l$, the decision tree groups together different classes based on the outputs of the dependent variable or target variable recursively partitioning the dataset in different nodes m. For a Dataset (D) being the data at a particular node m be denoted as D_m and $N_m$the number of samples at that node. For each splitting procedure of a candidate $\theta = (j, t_m)$, being j the feature and t_m the threshold of the recursive partitioning, we obtain the different subsets of a node:

$$D_m^{left}(\theta) = \{(x, y) | x_j <= t_m\}$$
$$D_m^{right}(\theta) = D_m \setminus D_m^{left}(\theta)$$

The quality of this split for each individual node in our tree structure is computed using an impurity function or loss function $H()$ depending on the scope of the algorithm, either classification or regression. For further information the researcher recommends: Abdullahi, A. I. "Comparison of the CatBoost Classifier with other Machine Learning Methods". Apart from the official Catboost[27] and Skikit Learn[28] documentation.

3.3.3 Underfitting and Overfitting: Implications and techniques.

On the one hand, the problem of underfitting is easily understood, as not using all the information of a dataset could lead to imprecise conclusions. On the other hand, using all the data from the dataset could lead to an overfitting model, a concept that is more difficult to interpret. Both mentioned algorithms are prone to find local minimums in a large dataset, especially for an increasing number of independent variables. Not being able to find the global maximum that minimizes the impurity of our predictor would not allow our model´s conclusions to be extrapolated with other datasets. This is what is considered to have an overfitted model, a model that would not be able to classify previously unseen

[27] Official Catboost docs. Available at: https://catboost.ai/docs
[28] Official Scikit Learn docs. Available at: https://scikit-learn.org/stable/modules/tree.html

records once they are included in our dataset.[29] Noisy data could also produce overfitting effects, as noisy training data could lead to a misclassification of nodes in a certain dataset. Therefore, as finding the global maximum in a NP problem for decision trees we need to compare different trees and assess which one does interpret the data in a more precise way. We have already discussed different evaluation metrics to assess this fitting of a model: AUC and F1. There are different methodologies to compare different predictive algorithms and assess which one is better. In the case of Logistic Regressions, gradient descend is a commonly used approach in order to minimize the cost function $L()$ that we will discuss later. When dealing with decision trees there are multiple approaches: Bagging, Occam's Razor, Boosting or using Random Forest.[30] Boosting is especially interesting for our model as it sequentially assigns weights to incorrectly predicted previous instances for classifying. There is an approach that combines both the Gradient Descend and the Boosting technique known as Gradient Boosting that we are using in our model. More specifically, the Catboost classifier that uses binary trees as the source of its gradient boosting algorithm.[31] Let's define our data set as before in our decision tree without focusing on the data in the nodes but the hole dataset:

$$D = \{(X_j, y_j)\}, j = 1, \ldots, m$$

Taking into account that our possible values of the target value $y_j \in R$, could only be [0,1], following the reasoning from before were k=1 or k=0, the subset $D_k \subseteq D$ for each node; and that $X_j = (x_j^1, x_j^2, \ldots, x_j^n)$ is a vector on n features or independent variables. The goal of the algorithm is to train a function $F: R^n \rightarrow R$, therefore a collection of functions $F^0, F^1, \ldots, F^t, \ldots, F^n$, for a given loss function $L(y_j, F^t)$. Assuming that we have constructed a function F^t we can improve our estimates by finding another function (h):[32]

$$F^{t+1} = F^t + h^{t+1}(x)$$

That minimizes the value of the Loss function:

[29] SUJATHA, M. "*A Survey of Classification Techniques in Data Mining*" (2013)
[30] SUJATHA, M. "*A Survey of Classification Techniques in Data Mining*" (2013) pg.90-91
[31] IBRAHIM, A. A. (et Al.) "*Comparison of the Catboost Classifier with other Machine Learning Methods* pg. 742
[32] HANCOCK, J. T. KHOSHGOFTAAR, T. M. "*CatBoost for big data: an interdisciplinary review*". p. 6

$$\mathbb{E}L(y, F^{t+1}) = \mathbb{E}L(y, F^t + h^{t+1})$$

Now that we have a clearer idea on how this gradient descent works, let's check out the exact algorithm that we used for our model; Log Loss and Cross Entropy:

The Log Loss used for fitting a logistic regression is:

$$\mathcal{L}_\theta(x, y) = -y\log(\tilde{y}) - (1 - y)\log(1 - \tilde{y})$$

Where, $\mathcal{L}_\theta(x, y)$ is the Loss function for the parameters θ for the x inputs of our model and y labels of our inputs x. If $y = 1$, x would belong to a certain class. And if $y = 0$, x would not belong to that class. $\tilde{y}$ is the estimate of y in our logistic regression. Therefore, our logit:[33]

$$(y = 1|x) = \tilde{y} = \sigma(z) = \frac{e^z}{e^z + 1} = \frac{1}{1 + e^{-z}}$$

Log loss is used for binary classification while cross entropy is used for multi classification. Therefore, if our target value could have only values [0, 1], "success" or "fault" then we would normally opt for choosing the Log Loss. However, if we reduce the number of classes of the cross-entropy algorithm, we would observe that it is the same as the one for logistic regression.[34] The only difference would reside in the way of updating the model parameters, as the cross entropy would use two logits while the log loss would use only one. Therefore, as it has more parameters, Cross Entropy algorithm is expected to tend to overfit the model more than the Log Loss.[35] According to Catboost Library our Log Loss and Cross Entropy would be defined as follows:[36]

Log Loss:

$$\frac{-\sum_{i=1}^{N} w_i \left(c_i \log(p_i) + (1 - c_i)\log(1 - p_i)\right)}{\sum_{i=1}^{N} w_i}$$

[33] MAO, L. "*Cross Entropy Loss VS Log Loss VS Sum of Log Loss*". 2020.
[34] *Ibidem*
[35] *Ibid*.
[36] Catboost official docs: https://catboost.ai/docs/concepts/loss-functions-classification.html

Cross Entropy:

$$\frac{-\sum_{i=1}^{N} w_i \left(t_i \log(p_i) + (1 - t_i)\log(1 - p_i)\right)}{\sum_{i=1}^{N} w_i}$$

We could indeed observe that formulas are similar except for the c_i and t_i components. As explained before $c_i = 1$ would only be considered, as if it is 0, x would not belong to that class. However, for t_i both cases $t_i = 1$ and $t_i = 0$ would be assessed when choosing the parameters, or feature importance.

On the other hand we could also observe that both Catboost Functions are similar to the Log Loss for logistic regression. However, the summation operators and the multiplication with weights, both in the divisor and dividend would correspond with the "Boosting" technique mentioned earlier. According to Catboost library documentation, these weights are predefined to have a value of "1" for the first run of our algorithm. Afterwards, these weights are recalculated according to the Bayesian bootstrap procedure. Please refer to the official documentation for further information and to D. Rubin "The Bayesian Bootstrap", 1981, Section 2.[37]

3.4 Evaluation Measures

The main goal of predictive evaluation measures is to assess if a certain method would be able to successfully predict the real or actual result of an objective variable by automatically analyzing the input variables from a dataset. To develop such an analysis, normally the dataset is divided into a training and a testing subset used for different purposes. The training dataset would be used by the algorithm to "learn" how the input variables relate to the objective variable. Once this learning process is fulfilled, the input variables that belong to the testing subset are processed by the algorithm to predict the result variables. After this process has been carried out we can compare the predicted output with the actual test data that we had. Therefore, we would be able to assess and evaluate the effectiveness of the prediction qualities of our model.

[37] Catboost official docs: https://catboost.ai/docs/concepts/algorithm-main-stages_bootstrap-options.html and https://catboost.ai/docs/concepts/parameter-tuning.html

The confusion matrix in the appendix would help us to graphically understand these concepts.[38] It could be observed in the Figure 1 that if the output of our model for a certain set of input variables corresponds to the real data from the testing data sample. Then we would obtain either a True Positive (TP) or a True Negative (TN). Therefore, we would be able to define the different concepts that would help us understand the effectiveness of our model in Table 2.

All the metrics presented in Table 2 are relevant to assess the effectiveness of the binary classification of the different algorithms. There are some important remarks between the concepts Accuracy and F1. As even if both of them do assess the quality of our prediction process their importance varies depending on the distribution of the "result class", that is the distribution of errors (result = 0) and successful transactions (result =1). Therefore, we are expected to rely more on F1 as a more valuable evaluation metric than Accuracy.

The case of failure diagnosis is especially interesting regarding this distribution, as the results classes are not evenly distributed due to the nature of IT systems with synchronous processes, it would not be possible to have similar failure and success distributions in a system expected to maintain high success rates, such as 90% or 98% of all transactions succeeding. As an example, trying to enter a webpage could not fail 2 times out of every 10.

However, if a new technology is quickly implemented in the process workflow, the distribution of these error rates could significantly vary to achieve higher error rates. This is a key concept that guides our research. As we assume that this increase of the error distribution would allow us to implement predictive methods to find relevant paths that would allow us to spot significant error trails in an ecosystem where a new technology has been implemented.

There is still one evaluation method that is being used in our analysis and that is related with the mentioned above concepts. The "Area Under the Curve" (AUC) concept. This concept is defined as *the probability of a random observation from a positive result*

[38] Figure 1

class to be classified higher than a negative positive class. This concept is intrinsically related with the ROC space. Which is the curve created by plotting the values of the probability of obtaining TP (P(TP) or Recall) and the probability of obtaining false positives (P(FP) or FPR). Graphically we could intuitively better understand this concept.[39]

Being $f(x_0)$ the distribution of the negative class and $f(x_1)$ the distribution of the positive class. In the same graph the dotted line would represent the threshold of 0.5. Therefore, if the probability distribution does not share any area and the positive distribution is on the right of the 0.5 threshold; the value of AUC is equivalent to 1. And our model would be capable of perfectly differentiate the positive and negative class. AUC would always have a value between 0 and 1, and our model would be not able to differentiate at all between the positive and negative class when both of the distributions share the same space around the 0.5 threshold.

3.5 Technical overview of approaches supporting failure diagnosis

In this section we would review the selected literature from a more technical approach and compare it with our proposed solution for dd FD following the table presented in chapter 2. As already mentioned, there is a lack of literature trying to apply these industrial approaches to other sectors such as the financial one. Therefore, we have added extra studies. Therefore, we have added extra studies (Yin, S. 2012 and Mei, J. 2014) to the literature review table that applies dd FD to chemical plants.

All of these studies' intent is to identify and classify errors in different environments and on different datasets. The newly added technical ones (Yin, S. 2012 and Mei, J. 2014) provide information on how in an industrial environment the data extracted from sensors in a chemical plant is used by means of different data driven methods to whether assess if there is indeed a failure in a certain component in the production process. It is especially interesting the paper of Mei, J. 2014 as its scope on dynamic features taking into account the time dimension using a multivariate time series (MTS) method, does offer better results than the traditional approaches. This is an interesting approach for future research,

[39] Figure 2

as comparing how these relevant fault paths vary in a specific period of time could greatly increase our understanding of the behavior of faults in our system.

Regarding Yin, S. 2012 study, it does state an important conclusion regarding the design parameters to be chosen to model, as depending on the criteria for parameter selection the results vary. In our case, the reasoning behind the tools depicted in this study would support the selection of variables to be modeled. Therefore, it is necessary to carefully select the variables to input in our model.

There are important differences between these technical papers and our approach. Firstly, both studies focus on fault diagnosis as detecting if an abnormal behavior is taking place in their respective systems' components. However, our focus does not reside in identifying which component is prone to fail, but which combination of input variables does offer a significant error rate and from these variables which are the most relevant one to effectively understand the source of the errors. In our case we also have incomplete information as depending on the characteristics of the faults in the transaction process some data fields could not be properly retrieved and they were assigned "null" values. According to the current literature, it could be observed that when dealing with incomplete information machine learning methods are suggested to be used for failure diagnosis. (Nor, N. 2018)

Secondly, both study methods are applicable when dealing with continuous variables. Therefore, their methods could not be directly extrapolated to our analysis as we are dealing with categorical variables. Also, both authors recognize, and in the case of Yin, S. 2012 explicitly in its conclusions, that in order to benchmark their values they are assuming a Gaussian distribution of the faults in their system that does not necessarily apply to the real-life experience. The proposed tool does not rely on any predefined distribution of faults when assessing the feature importance of the different input variables. As these variables' values are assigned a weight that is recalculated in the learning process of our algorithm.

Regarding the studies of Lewis, L. 1993 and Bartolini, C. 2010, at a first glance it could be observed that the data in which they based their papers is the one retrieved from the ticketing systems and not from the raw dataset of failed and successful processes. However, these studies do offer interesting insights from a managerial point of view. In

the case of Bartolini, C. 2010, the main objective is which support group should be chosen to assign a ticket depending on a set of variables such as the time of resolving tickets or the number of tickets from the different support groups. This approach is interesting as its main focus is to support decision making in a company using a stochastic statistical model. Even if our objective of classifying fault data could be considered similar it is not the objective of the research to automatically assign tickets by typology to different support teams, but to effectively spot fault trends useful for the different members of the operations, support and technical teams. Therefore, the classification approach is similar to the one intended by Lewis, L. 1993. However, in the current research it is not the objective to deal with ticketing system data and nor it is to specifically deal with network availability problems to classify similar tickets. The proposed approach is more abstract and there are no predefined reasons for a fault to occur. Quite the opposite, the researcher uses all the available information in our dataset in an intent to better understand these faults.

Finally, the two papers which offer solutions to dd-FD with a higher degree of resemblance to our proposed tool are Chen, M. 2004 and Kwon, J-H. 2017 Regarding the Kwon, J-H research paper, even if the association rule mining has been proven to be useful in an industrial set up. The way of classifying these variables using one hot encoding is not properly explained. This issue presents an important drawback for two reasons. First, the mathematical implications of using one hot encoding instead of treating each categorical variable independently as the tool proposed in our research does, this may look like a minor drawback. However, the most widely used library for classification trees in python, "Scikit Learn", does not handle categorical variables as our tool does.[40] Secondly, the binarization of variables, as for example the decision explained in the paper of dividing the CPU usage ratio into 3 categorical variables is an assumption created by the researcher that bias the diagnosis of failures if it does not attend to a business or a system rule. Our proposed model deal with each categorization of variables providing a proper explanation by attending to business rules and the environment of the system following the CRISP-DM methodology.

[40] Please, for further information check the official documentation at: https://scikit-learn.org/stable/modules/tree.html

Regarding the paper of Chen, M. 2014, even if the researcher profoundly agrees with the decision of using decision trees to treat failure diagnosis tasks, especially regarding the availability of post processing the tree structure in search of the relevant paths for fault diagnosis. There is no discussion regarding the optimization of such decision trees avoiding local maximums. A characteristic that profoundly biases the research even if the results seem satisfactory enough. In the case of our proposed tool a gradient boosting descend has been implemented to secure optimal output trees.

APPENDIX CHAPTER 3

Confusion Matrix	Predicted Output		
Actual Output		Result = 1	Result =0
	Result = 1	TRUE POSITIVE (TP)	FALSE NEGATIVE (FN)
	Result = 0	FALSE POSITIVE (FP)	TRUE NEGATIVE (TN)

Figure 1 - Confusion Matrix

Table 2 - Predictive Performance Measures

Name	Description	Formula
Precision	Ratio of positive observations correctly predicted from the total positive predictions	$TP/(TP+FP)$
Recall or TPR (True Positive Rate)	Ratio of positive observations correctly predicted from the total actual positive results	$TP/(TP+FN)$
Accuracy	Ratio of true and negative observations correctly predicted from the whole population	$(TP+TN)/(TP+FP+TN+FN)$
F1 Score	The harmonic mean of Precision and Recall. Therefore, a weighted average of the precision and recall.	$2 * \frac{Precision * Recall}{(Precission + Recall)}$
Weighted F1	F1 weighted by the number of items belonging to each class	$\frac{F1_{class=1} * (TP+FN) + F1_{class=0} * (TN+FP)}{TP+FN+TN+FP}$
(FPR) False Positive Rate	Ratio of positive observations correctly predicted from the total actual positive results	$FP/(TN+FP)$

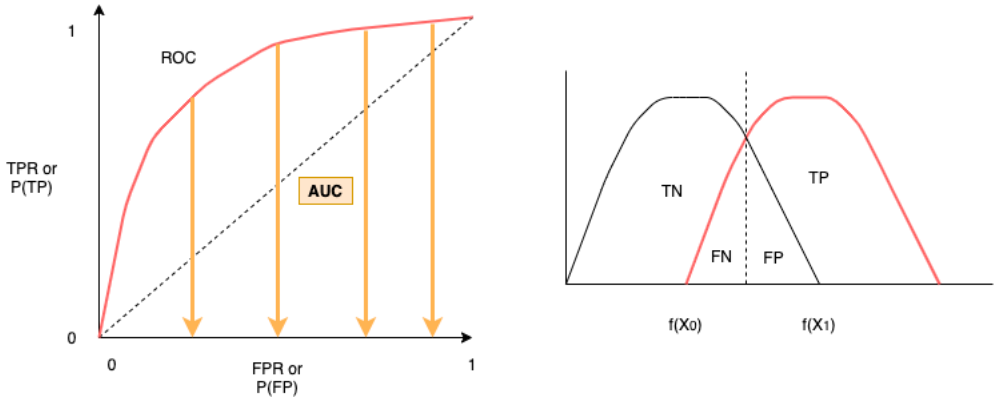


Figure 2 - AUC representation and fault distribution

4 CASE STUDY ON FAILURE DIAGNOSIS AT AN PAYMENT SERVICE PROVIDER: DATA PERSPECTIVE

The case study is discussed in two separate chapters according to a split of the steps within CRISP-DM:

- In Chapter 4 the business and data understanding phase as well as the data related issues of preparation and selection of the data to be modeled is outlined (CRISP-DM phases 1-3)
- In Chapter 5, the models implemented and the structure of the tool is presented. (CRISP-DM phases 4-6)

During the following chapters, an overview of the company's problematic is presented as well as the proposed solution. In a nutshell, the company is dealing with a technological change in their processes and needs a way of quickly identifying faults that may correspond to a problematic or failure in their processes. Our proposed solution does not only identify these faulty paths, but also offers these insights in a easily understandable way to the business users in the company. Therefore, the tool allows the company's business users to quickly identify and offer insights of failures in their processes to effectively manage them. This approach helps reducing the time dealing with faults and incidents and, consequently, it has a positive impact for the customers as explained in Chapter 2.

4.1 Case Study background

Before describing the approach of the analysis, the environment and the enterprise where the study has been conducted needs to be discussed. The company is a start-up that is a subsidiary of an automotive enterprise and it aims to be the payment service provider (PSP) of the whole group. Therefore, the company has an Electronic Money Institute (EMI) license to provide PSP related services.

Regarding the environment, that the next factor that needs to be taken into consideration is that on January 1st 2021 the new European Payment Service Directive (PSD2) has been implemented. As briefly introduced before, this directive aims to increase the complexity of payment validation procedures to avoid fraud in transactions in E-commerce for the whole European Union by the use of Strong Customer Authentication (SCA). This new payment procedure obliges customers, companies and payment providers to exchange

multiple data points in order to ensure that the customer purchasing a product or a service has not a fraud intent. Therefore, to successfully perform a payment process there is an exchange of different data between the mentioned actors such as MAC address, location, IMEI, used networks from the user, as well as the personal banking passwords or biometrical data such as the finger prints. As an example, in order to buy products or services online, a customer may not be able to finish the purchasing process without performing a Two-Factor Authentication using a Token, a banking app or other means

Failures can happen in multiple steps of the payment process, therefore, the objective of the tool developed, as mentioned before, is to effectively classify these faults to spot and extract explanatory insights from them in an automatic way. These insights ultimately focus on providing relevant information to the operations teams in the company to address incidents and faults as quickly as possible to ensure a "business-as-usual" environment throughout a new technical implementation.

4.2 Business Understanding

The payment service directive (PSD2) is implemented within the European Economic Area (EEA). This directive guides financial institution to implement solutions to increase fraud detection in transactions. However, at the same time it it reduced customer satisfaction by introducing an additional verification step in the online purchase process. For guest payers in e-commerce shops the additional verification also called 3D Secure is always applied. However, for registered payers, meaning that customers have a stored a payment option with the webshop, only during the storing of the payment option the 3D Secure verification needs to be performed.

The key of for the implementation of 3DS and PSD2 measures are data. Since the introduction of the new directive, more data of customers is needed to validate the payment – address, device information, etc. This is to assess the potential risk of the customer to perform a fraudulent transaction.

In the case, where the risk is assessed as high, the issuer could invoke a strong authentication method that may vary from a verification code sent to a mobile phone till the introduction of biometric input, such as a fingerprint. As this measures have only been introduced in the beginning of 2021, there is a lack of academic literature regarding the

effects of this adoption. However, different sources indicate that this system reduces both customer abandonment when making a transaction and the fraud rates.[41]

From a general point of view, we could already realize that in order to effectively make a transaction a cooperation between the vendors and payment providers and the user would be needed, as a successful exchange of information is required between the actors implicated in the transaction process. A successful transaction depends on the successful execution of tasks to fulfil a process, between the different actors or nodes in the network. In this process chain, the focus is on the position of the payment provider, as the company where this analysis is performed is an Electronic Money Institute (EMI) and a payment service provider (PSP). This company manages the payment services of different merchants and end users from Europe, cooperating with other payment service providers with international presence. Regarding its general structure, this company is a subsidiary of an automotive group for whom they manage the payments system. The company as such also has a subsidiary that serves as technical provider for their platform.

The company serves different customers such as merchants, end-customers and marketplace operators. Market operators are companies that provide a webshop (marketplace) to other companies. Merchants are Business-to-Business clients (legal entities) that receive the collected funds from the payment service. End-customers in this context are Business-to-Customer (B2C) clients, who are physical persons.

The core business of the company is the processing of payment transactions. The transactions follow different flows depending on the payment method. We can differentiate three types: SEPA Direct debit, credit card or alternative payment solutions. For clarification purposes, this study assumes a linear communication of the different actors for explaining the different typology of the payment process. For all processes the flow begins in the merchant's or *marketplace operator's backend* when the transaction is initiated. After that the *company's payment platform*, authorizes or rejects the transaction using the payment gateway, which is checking different data depending on the type of process (eg. CVC code, IBAN, expiry date of a credit card, etc).

The different flows of the transactions is divided as follows:

[41] "*Preparing for PSD2 SCA*". VISA. 2018. P.17

- SEPA payments: the payment platform and gateway check the correctness and validity of the IBAN. Then the payment gateway sends a response to the technical platform, if positive, the payment request is sent to the end-user bank, else the transaction is rejected. Once the payment is approved by the end-user bank, the payment is sent to the gateway and received by the company. [42]

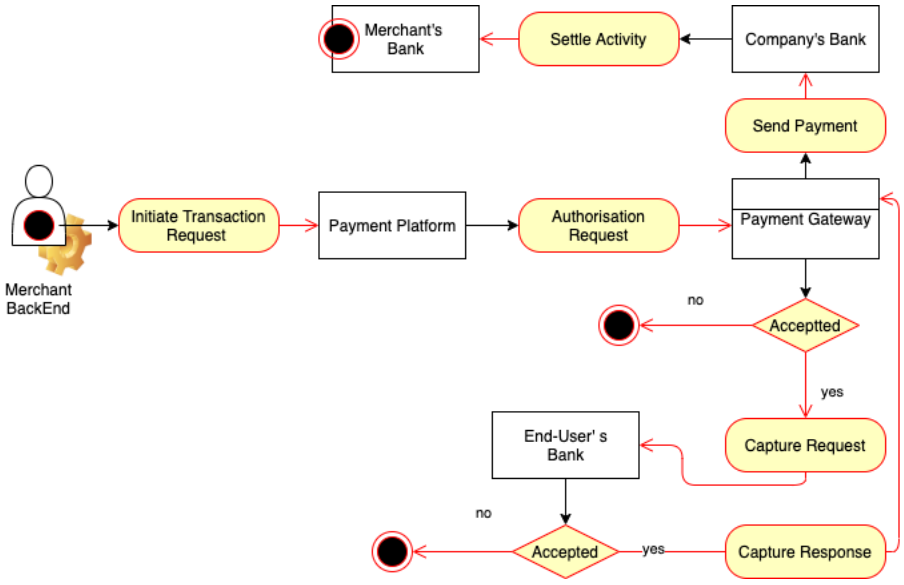


Figure 3 - SEPA payments Process Flow

- Alternative payment methods: the payment gateway would check that the data provided corresponds to the one necessary to perform a successful transaction with the alternative payment method's system. such as checking if the email address for PayPal payments. After the transaction is authorized, the payment process is initiated and a payment request is sent to the alternative payment service platform, which would approve and send the payment directly to the company bank.[43]

[42] Figure 3
[43] Figure 4

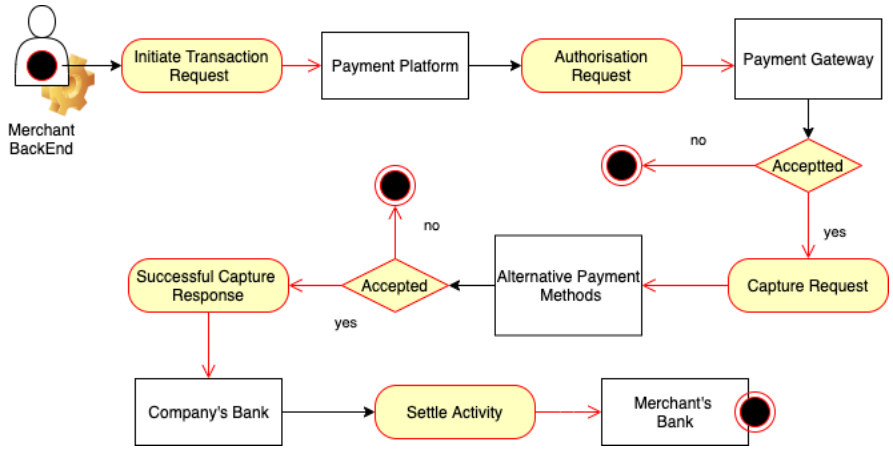


Figure 4 - Alternative Payments Process Flow

- Credit Card payment (cc): the payment gateway would perform validity checks such as the correctness of the credit card data. After the authorization of the transaction is positive, the payment process begins and the acquirer requests the payment to the end-user's bank. Once the bank approves the payment, the payment would be sent to the acquirer who ultimately would redirect this payment to the company bank. After the payment is received by the company bank in all process types, the company would finally send the payment to the merchant's bank.[44]

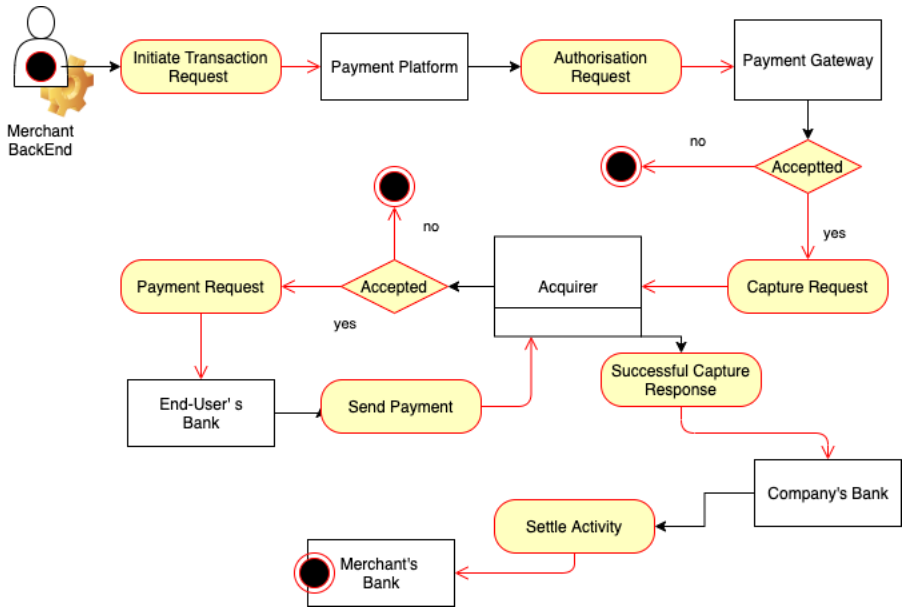


Figure 5 - Credit Card Payment Process Flow

[44] Figure 5

The adoption of this European directive obliged different payment providers and actors in the payment process to cooperate changing the protocols of the payment industry. As any implementation of a new technology, these protocols had incremented the rejection rate of transfers in Europe. In this situation, the company's objective is to reduce the rejection rates to the minimum possible by using the means under their control. This rejection rate could be decreased following diverse procedures, such as: by quickly informing other vendors of the nature of the problems that their current solution have or spotting which vendors are failing to accomplish a smooth enough transaction process and contacting them.

Therefore, to effectively convey information to the other actors in the environment the company needs to understand why the transactions are failing. That is: What may be the reason that affects the integrity of the transaction according to these new protocols? In addition, the location where these transactions are failing is important such as: a geographical market, a specific channel (e.g. Visa Channel, MasterCard Channel), or in a specific step of the payment flow.

Finally, the tool has to be able to prioritize according to the importance of these failed transactions, relying on the number of failed transactions or the percentage of failed transactions for a specific path. All these concerns relate to the descriptive or explanatory dimension of data.

The solution developed is required to be general enough to find error paths but specific enough to be illustrative to effectively spot where the problems are and, to some extent, offer insights about why these transactions failed. Therefore, the data and insights that are received from the analysis should not only be reliable but also pragmatic to fulfil the purpose of understanding why these transactions fail.

4.3 Data Understanding

Transaction data is stored on the servers of the company and it is accessed over a VPN connection using company credentials. The data can be downloaded from the vendor's website choosing the columns required for the analysis in a comma separated document (.csv) format. The data table contains 35 columns and more than 20,000 rows, each

corresponding to a unique transaction process. Each transactions (line item) has a corresponding channel. These channels represent different merchants and are also separated per payment method for these merchants. Relevant fields are observed such as: the ones indicating if a transaction was successful or rejected or the error code in case of rejection.

The different columns can be classified according to the nature of the information they provide regarding the transaction. The classification are as follows:

Table 3 - Data Description

Question type	Type of Data	Variables	Instance	Unique values	Null Values
Where	Location Data – identifies where the transaction origins and where it is performed	Customer country	ES	35	2931
		Account country	IT	35	486
		Channel	PIT98_Merchant_ES_CC	38	0
Who	Customer Specific Data	Email	somebody@server.com	Not taken into account for privacy reasons	
		Customer name	Name Surname		
		Account holder	Name holder		
		IP address	***.***.***		
How	The typology of the transaction	Payment Type	PA	7	0
		Payment Method	CC	4	0
	The current status of the transaction flow	Payment Type	CP	7	0
		Return codes	800.200.159	31	0
	Merchant or Marketplace	Channel	PIT98_Merchant(s)_ES_CC	38	0
	Parties involved in the process	Brand	VISA	10	49
		Channel	PIT98_Merchant_ES_CC	38	0
	Description of the product or service	Usage	CartID:KSIDJFUjnufu7s87f	622	6356
Which	Basic transaction data	Debit	***€	Real Number	19978
		Credit	***€	Real Number	475
		Currency	EUR	14	4732
	Data that indicates if the transaction	Result	ACK	2	0
		Status Code	60	4	0

	have failed and extra data about that result	Reason Code	90	10	0
		Return Codes	800.200.159	31	0
	From where the technical service has retrieved the data	Source	OPP_system	3	0
	Data and relationships that are related with the integrity of the transaction and the risk-based Authentication	Unique ID	Kfjdufud****KGIfid	20453	0
		Transaction ID	Kfjdufud****KGIfid	7387	6356
		Merchant ID	Kfjdufud****KGIfid	104	55
		Channel ID	Kfjdufud****KGIfid	38	0
		The account number	**** **** **** ****	2368	939
		BIN number	******	705	12220

In Table 3 there is an presentation of the different variables that the dataset contains. In this project, the variable “Result” is the objective variable. When the value of result is “ACK”, the transaction’s sub-process is successful. On the contrary, when the value is “NOK”, the transaction’s sub-process failed. The variable “Payment Type”, indicates the sub-processes of the transaction, such as Payment Authorization or Capture.

There are only two variables: “Credit” and “Debit”, whose values are numerical. The rest of the data is categorical or multivalued and are formatted as a string. Attending to the classification goal of the model, there are some data fields or variables that of relevance to the fault diagnosis objective. For example, unique IDs or Account number cannot be categorized. However, there is still information to be extracted from these variables such as: Does a valid bin number exists in the transaction process? Do account holder and customer name fields have the same value? Does the currency used in the customer account country or the currency from the country where the transaction has been coincidental with the currency field?

An especially interesting field is the result code, as it provides relevant insights for analysis by clearly stating the reason of the rejection. For example, the result codes can be very explicit like 800.200.159 account or user is blacklisted (card stolen), to more general descriptions of or 100.380.401 “User Authentication Failed”. Therefore, these error codes would greatly support business users to better understand why the transaction

has failed. It is of great importance to interpret if the paths that the tool identified present a similar typology of these errors for business user evaluation and assessment of the situation.

For privacy reasons the data sensitive variables have been excluded from the analysis. Notwithstanding, these variables are too specific to be able to offer information about trends in the data to classify or group transactions according to its values for a classification model.

Regarding the data quality, it is needed to acknowledge that in order to extract all the available information of the dataset, the undefined or null values could not be simply eliminated from the dataset, as their presence could constitute an indicative on why a transaction has failed. For example, a failed transaction using credit card (CC) with a valid BIN and an inexistent CustomerCountry value. This transaction may fail as it does not specify from which country the customer comes from, represented with a *null* value instead of a valid code country as a value. Therefore, it is compulsory to introduce this data typology in the model as another value, in this case it has been chosen the String value: "DUMMYNA" to avoid the wrong interpretation of different algorithms when trying to introduce strings such as "na", "nan", "null" or their capitalized values.

The dataset in scope has a failure rate of 10.21% of 20453 lines items. As already mentioned before, each of the line-items correspond to a certain sub-process of a transaction. The company is focused on retrieving two different sub-processes of a transaction: Payment Authorization (PA) and Capture (CP), as these are the main processes involved in the flow of collecting money from customers or merchants. Therefore, these are the possible values from "Payment Type" taken into account for this project. The resulting dataset consists of 15246 entries with a fault rate of 8.17%. In this newly filtered dataset, depending on the value of the rest of the variables it is observable certain interesting fault distributions:

- Credit: Most of the payments processes that have a credit field specified, have a low value. According to the violin plot of Figure 6, it can be observed that failed sub-processes not depend on the credit amount. The shape of "NOK" or failed

transactions in the plot in Figure 6 is more concentrated than the “ACK” or successful processes, due to the inferior number of processes for each category based

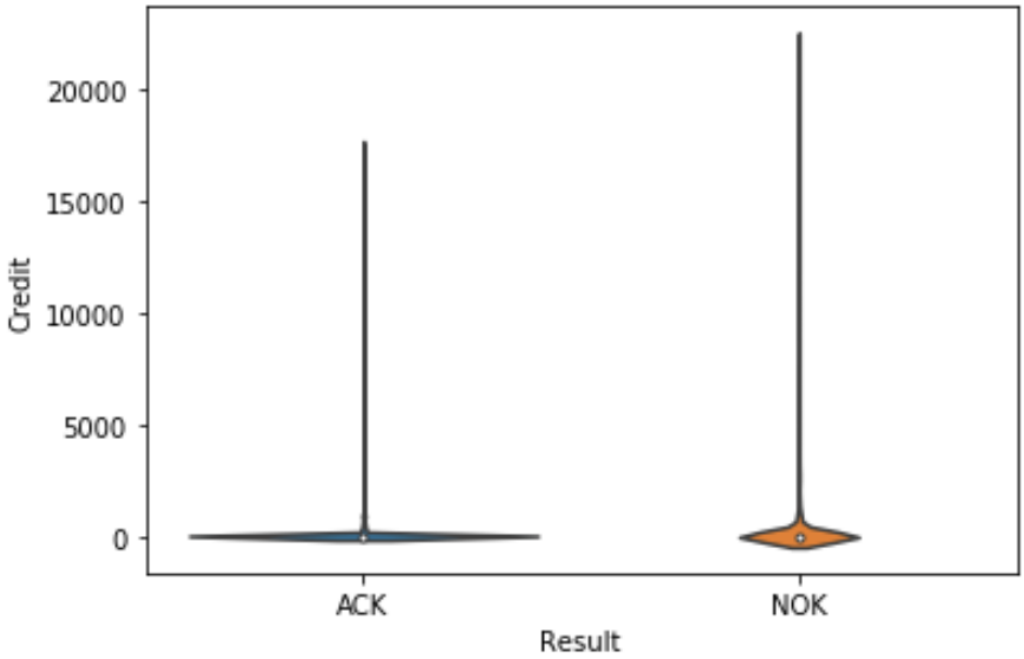
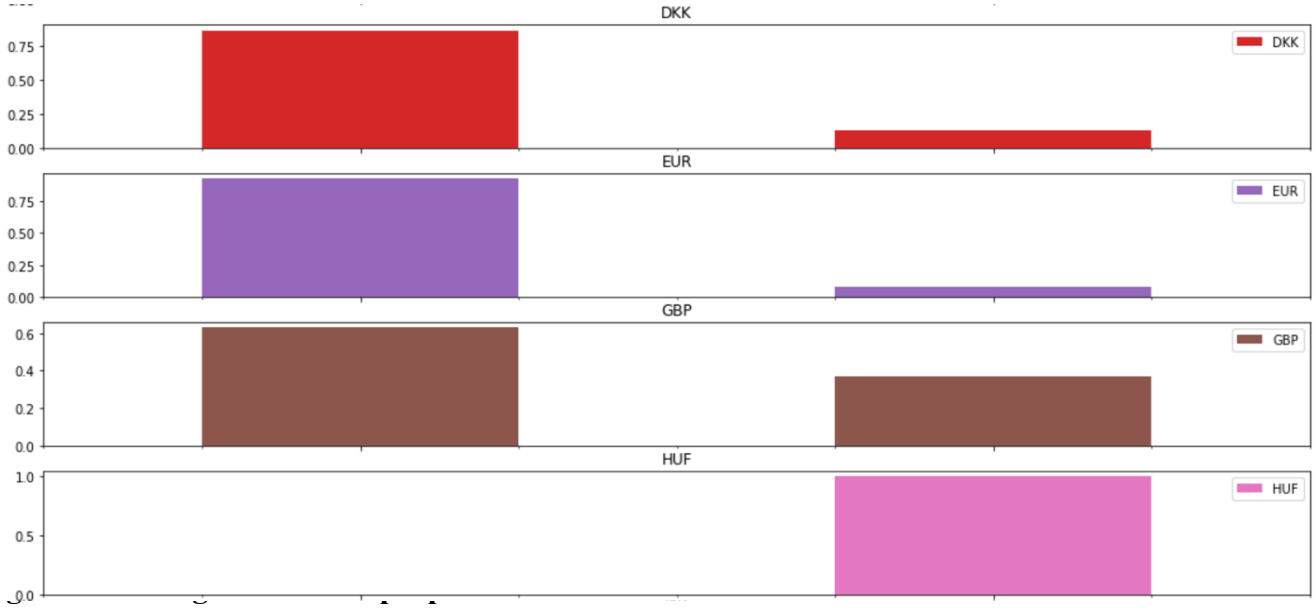


Figure 6 - Violin Plot : Error distribution based on the Credit amount

- Other variables such as Channel, Currency, Customer Country, Account Country or Brand do vary in their associated fault distribution. As an exemplification Figure 7 for the Currency variable is showing an extreme effect. In the plot the left bar represents successful transactions and the right bar represents unsuccessful transaction. These bars represent which share of the transactions either failed or were successful. The different currencies in the example are: DKK (Danish



Krone), EUR (Euro), GBP (Pound Sterling) and HUF (Hungarian Forint). As shown in the Figure 7 both DKK and EUR have less than a 20% of a fault rate, GBP transactions nearly reach a 40% percent error rate and HUF has no successful transaction at all.

- Specially relevant is the error distribution for the "Payment Type", both for the CP process and the PA process. As shown in Figure 8 most of the CP processes are successful. While the errors are mostly produced in the PA process.

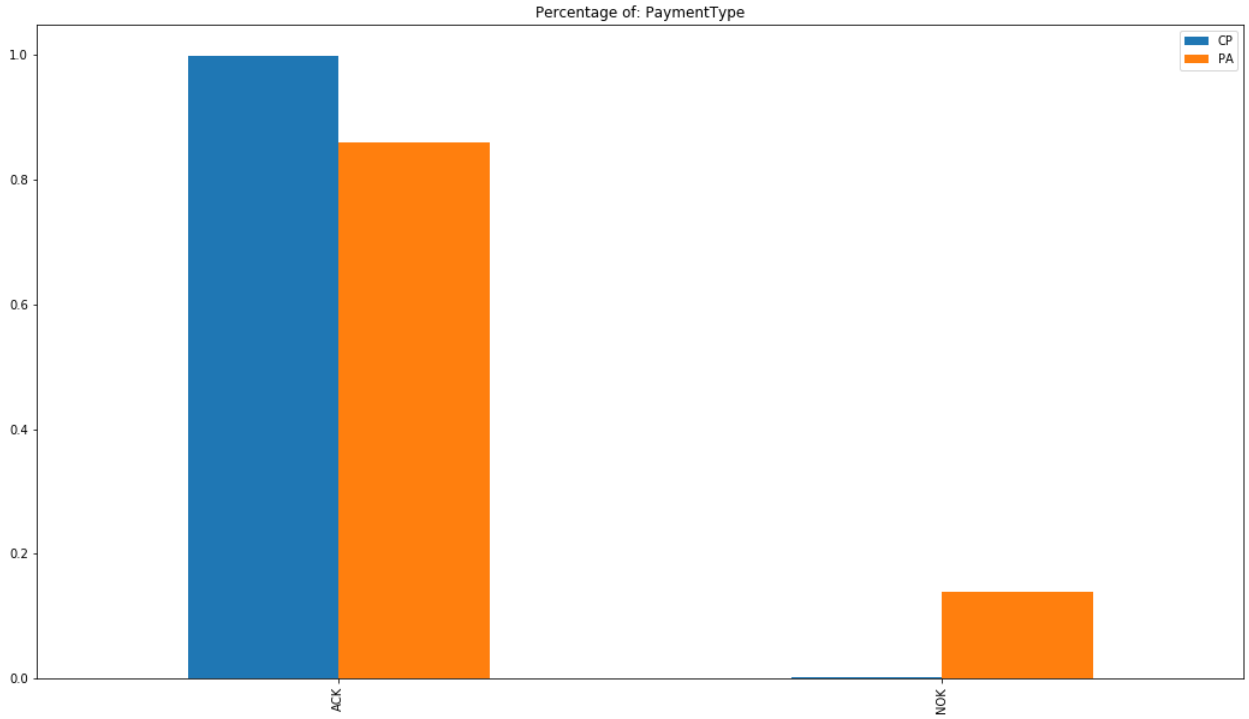


Figure 8 - Histogram of fault distribution based on the values of the PaymentType variable

However, attending to the simplified process flows in Figure 3, 4 and 5 from the section above. There is no "Capture" process initiated if there is not already a successful finished PA process. This simple process mining concept guide the next filtering of the database the successful PA requests that were followed by a CP request. There is a data field allows so: TransactionID, which refers to a unique transaction process. By deleting the obsolete PAs followed by a successful CP request, the dataset finally consists of 8861 rows and an fault rate of 13.72%.

Once understood the basics of the dataset structure and the data characteristics, the data it is prepared and a descriptive analysis is performed to further comprehend the big picture of the company situation at this stage and what to expect to retrieve from our automatic model.

4.4 Data Preparation

4.4.1 Selecting variables and Exploratory Data Analysis

In the data understanding phase, a brief selection of variables has been performed by excluding data that do not provide relevant information to be extracted by the model of the data mining project. Therefore, summarizing the previous section:

- The dataset to be modeled has a fault rate of 13.72% and a total number of 8861 entries, taking into account PA and CP subprocesses and once we had properly filtered the data that represent unique transactions, the new fault or error rate is 13.72% and the number of items is 8861.
- Due to the objective of our tool, we are interested in credit processes as we are focusing our efforts in understanding the payment process that is started by the user. Therefore, Debit processes such as Refund would be excluded from our analysis as mentioned before.
- In order to be compliant with data privacy regulations and also due to the variety of the data values from the personal detail fields, we could not directly categorize these input fields and would be excluded from our analysis for this research project. Therefore, we would exclude: Email, IP address, Account holder, Customer name and Account Number.

In addition, from the different variables mentioned earlier in Table 3 there are some of them that are not added to the analysis as they do not provide consistent information:

- Depending on the merchant, the “Usage” type does provide different values or no values at all. Therefore, we would not be able to include this variable in our analysis.
- MerchantID, TransactionID, Bin Number and UniqueID, are considered to have to many unique values as multivalued variables to be effective classified. Therefore, their information it is not directly retrieved as a raw input in our model. These

variables are transformed in the next chapter in order to model the information that it could be retrieved from them.

- Regarding the value of the MerchantID, as is the variable with less unique values. It has been also excluded as a non-transformed variable for another reason: The researcher is not entitled to extract and present the list of clients from the company, and the values from these fields are not human readable. Therefore, it does not serve to the main purpose of the tool, offer to the business users an automatic classification of error paths to easily retrieve insights from the fault trends.

Apart from the selection and cleansing of the original dataset based on model requirements and limitations. A Exploratory Data Analysis (EDA) is performed to assess the correlation between the different variables, both independent variables and the objective variable. As mentioned in chapter 3, for the models to be used in the project, independent variables are expected to be indeed independent from each other or at least present a low correlation. The Cramer's V coefficient analysis indicates as observed in the heatmap in Figure 9 that:

- Variables like: Result Code, status code and reason Code. Were obviously highly correlated with our objective variable "Result", as they only have values different from 0 or 000.000.000 in the case of "Result Code" when a process fails. Therefore, we excluded these variables to be modelled by our tool.
- The variable "Source" was highly correlated with most of the variables except currency and the payment type. After realizing the meaning of this variable's possible values attending to the internal documentation of the company, we decided to exclude this variable as well, as these are values not related with the transaction itself but with the differentiation of operational and testing environments and when the data is retrieved and updated in our database.
- It can be observed that there is a high correlation with the variables Currency, Account Country, Customer Country. However, these variables are useful from a business descriptive point of view. Therefore, even if theoretically it could be argued to express this information just in one variable, these variables are being maintained to check if the different models are capable of choosing which of them is more relevant, and therefore, help us better diagnose the source of these failures.

- The same situation applies both to the correlation between Payment Method and Brand and the correlation between Channel and Payment Method. As Payment Method has values such as CC (Credit Card) and Brand is the commercial brand of these payment methods. Such as, Visa or Mastercard. Also, Channel, most of the times include a differentiation among different payment methods such as: PIT85_Merchants_PaymentMethod.

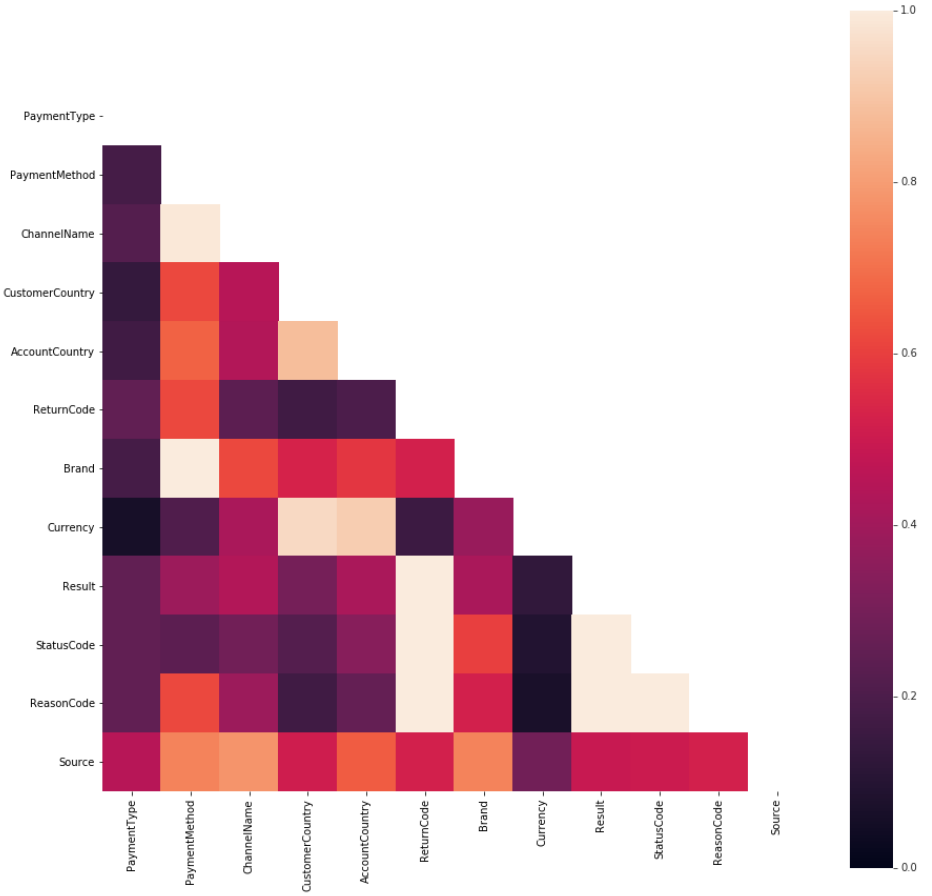


Figure 9 – Heat Map - Cramer's V coefficient

Therefore, after cleaning the dataset for each row to represent a unique credit transaction, the following variables have been selected for to be modeled as independent or input variables: “PaymentType”, “PaymentMethod”’ , “ChannelName”, “CustomerCountry”, “AccountCountry”, “Brand”, “Credit” and “Currency”. In the next section, it is explained how new variables are created to represent the information that is not possible to be directly modeled in our project.

4.4.2 Transforming and formatting Data

Once decided what variables are relevant to be modeled based on the interpretation of the business environment and data characteristics. It is needed to apply these learnings and focus on extracting as much information as possible from our data. Therefore, there are several variables and data structures that would be transformed in order to be easily understood not only for the different modeling methods, but also for the users that would interpreted these models:

Some of these variables have been developed attending to the characteristics of 3DS2 Secure requirements. As the payment gateway needs to check the integrity of transactions depending on the typology of the payment (ex/BIN check for international Credit card payments) or simply attending if the IDs of these transactions or processes exist. The values for the variables such as BIN, TransactionID and MerchantID in the dataset does not offer any relevant information for modeling using predictive or explanatory approaches. The information they provide is based on their presence when needed, that is why we have created new variables to extract their information.

We have created the variable “Logical_TransactionId” to check if the subprocess has a identifier or else it is a null value. Therefore, this is a binary described by:

$Logical\_TransactionID_i = 1$; There is a Transaction ID for the transaction process.
$Logical\_TransactionID_i = 0$; The transaction process does not have an ID

The variable “Logical_MerchantID” indicates if the transaction process has an identifier. Therefore it is a binary variable:

$Logical\_MerchantID_i = 1$; There is a Merchant ID for the transaction process.
$Logical\_MerchantID_i = 0$; The transaction process does not have a Merchant ID

The variable “Logical_Bin” indicates if the subprocess does have a valid BIN value. and it is needed to successfully fulfill the transaction process. Therefore it is a binary variable described by:

$Logical\_Bin_i = 1$; There is a BIN number for the transaction process

$Logical\_Bin_i = 0$; The transaction process does not have a BIN number

On the other hand, to be able to effectively distinguish comparable data from both a business point of view and for modeling purposes, we need to be able to understand exactly what the value for the "Credit" field expresses. In this case, it is an amount of a certain currency. Therefore, to be able to compare the value of the different amounts represented by the "Credit" field we need to create a new variable that transforms those amounts to the same currency. We have chosen € as the currency of our variable "AccountingAmount". However, due to the intrinsic characteristics of the company we are carrying out the study, there is a significant number of 0.01 amounts for different currencies that are not part of a "traditional" transaction process. These amounts correspond to a process included to be compliant with the new PSD2 directive, as it is a requirement to authorize a new payment option in order to be able to perform "frictionless" payments in the future. Therefore, we have created a new variable from the "AccountingAmount" that classifies the different values to differentiate and represent this fact. This new variable is called "CreditClass" it is multivalued variable that behaves as follows:

$$CreditClass_i = "Validation"$$

If the value of the AccountingAmount variable is less than 0.05 EUR we would assume that it is part of a validation process. As there is a little number of transactions that corresponds with a value of those characteristics that it is not part of a of this process of validation, according to historical data of the company. All of them free services

$$CreditClass_i = "Micro"$$

If the value of the AccountingAmount variable is more than 0.05 EUR and less than 10EUR we would name it as a "Micro" transaction. Following best practices of companies in the sector, such as Paypal.

$$CreditClass_i = "Macro"$$

If the value of the *AccountingAmount* variable is more than 10 as a "Macro" transaction. This differentiation does not only obey the encompassing of best practices of the sector in our model. But also attend to the compliance requirements of the PSD2 Secure Directive. The company implements the highest level of security requirements for all the different transaction process, however as we belong to a system in a network with different actors with different requirements there could exist a node that consider payments above a certain amount, would be "double checked" or more "strongly checked" by the different Payment Providers in Europe. In a process known as "Soft Decline".

Finally, the different variables to be modeled in our tool are represented in the following table:

Table 4 - Variables after preparation

Variable	Variables	Data Type	Original Instances	Transformed Instances	Original Unique values	Transformed Unique values
Objective or Dependent	Result	Binary	NOK, ACK	0, 1	2	2
Input or independent	Payment Type,	Category, Binary	PA, CP	PA, CP	7	7
	Payment Method	Category, Multivalued	CC, DC	CC, DC	4	4
	Channel	Category, Multivalued	PIT98_Merchant _ES_CC	PIT98_ Merchant_ ES_CC	38	38
	Brand,	Category, Multivalued	VISA	VISA	10	10
	Credit	Category, Multivalued	***€	MACRO, MICRO, VALIDATION	Real Number	3
	Currency	Category, Multivalued	EUR	EUR	14	14
	Logical_ Transaction ID	Binary	Kfjdufud**** KGIfid	0 ,1	7387	2
	Logical_ Merchant ID	Binary	Kfjdufud**** KGIfid	0 ,1	104	2

	Logical_ BIN number	Binary	******	0 ,1	705	2

In the next chapter these variables are processed by the proposed model to spot and describe the relevant fault paths in the database.

APPENDIX CHAPTER 4

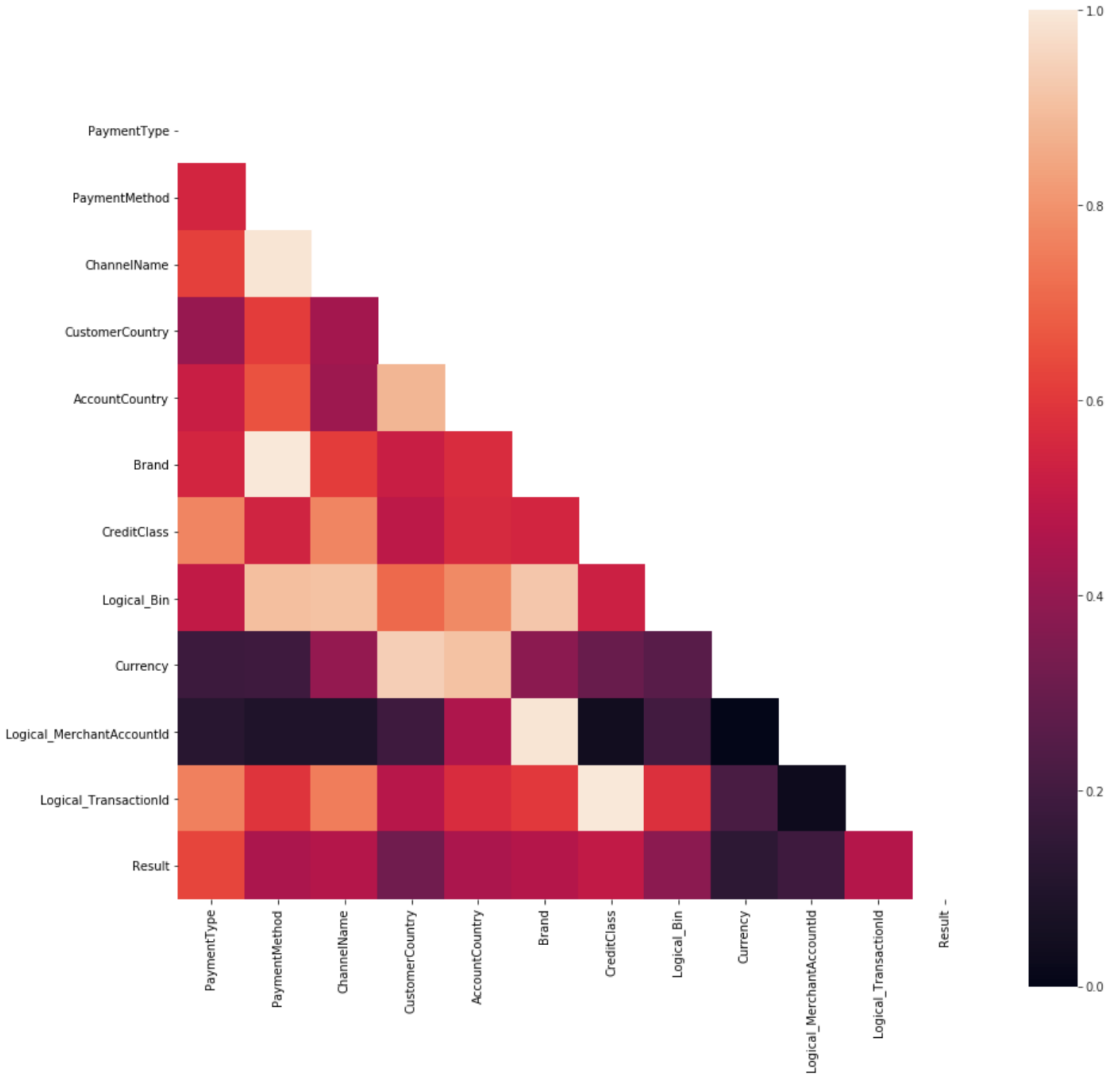


Figure 10 - Heat Map - Cramer's V of prepared data.

5 CASE STUDY ON FAILURE DIAGNOSIS AT THE E-MONEY INSTITUTE: MODELLING PERSPECTIVE

In this chapter we discuss if the case study sub question is being answered with our proposed model:

How to develop an explanatory model for data-driven FD for data from a company in the financial sector?

The developed tool identifies faulty paths in an automatic way following differentiated steps. From a global importance point of view, it is important to mention that if our data is properly structured as multivalued or categorized variables. The artifact, independently of the inputted dataset, would be able to identify discrepancies between the different values of the independent variables and the result of the values of the dependent or objective variable. Thanks to this characteristic, the developed model is considered useful for any company after performing a brief data mining project focused on the preparation of the data. This scalability dimension is the approach that guided the development of our artifact as it does not rely on any predefined distribution of errors.

Regarding the importance of this artifact for the company, a tool has been developed which is able to offer, depending on the model run, both specific fault paths and general fault paths. In addition to this characteristic, it does append the new information to a modified excel file that allows the business user, in this case the head of operations, to quickly filter according to these error paths the characteristics retrieved assigned to them. Such as the total number of errors or the percentage of errors per path. The designed outcome from this artifact does provide the company another way to look at the data, not focused on just detecting the errors in their system or comparing general aggregated data that would not be able to detect relevant errors such as little merchants outperforming.

The proposed artifact processes data and offer insights about it in two differentiated steps: First the extraction of relevant features using a classification algorithm based on decision trees or a binary logistic regression, then the output variables from this process are retrieved to find the most relevant combination of values for faulty behavior, finding in this way the most relevant fault or error paths.

The random extraction of a group of transactions (the sample) is expected to follow a similar error distribution than the whole dataset used in this study (for simplicity, the population). However, this model as a whole is based on the assumption of the researcher that financial transactions are rigid processes. Therefore, it is expected that groups of transactions attending to their shared characteristics would have either a significantly higher or lower error rate. This behavior could be already inferred from the Figure 7 and Figure 8 from Chapter 4. In Figure 7, after classifying by currency type, it is shown that HUF transactions have a 100% fault rate for a dataset with a 8% fault rate. An example presenting a lower fault distribution could be observed attending to Figure 8, where the fault rate of CP transactions is way lower than 8%.

Therefore, depending on the value of the categorical variables, the distribution tends to vary significantly. In this situation, assessing which variables do have more weight when "predicting" if a transaction would fail or not, in combination with an algorithm that would loop through the different combinations of values from these "predictive" variables, while identifying significantly higher fault distributions, generates "Relevant error or faulty paths", as it is further explain in this chapter.

Other considerations taken into account to select this stepped approach are:

- Interest in being able to compare different models: The weight in percentage that each model gives to each variable or feature could be compared in the case of the binary logistic regression and the binary classification tree models.
- Avoiding NP hard and NP complete problems: Classification trees are easy to produce but difficult to read by a machine as Binary classification trees are in essence directed graphs. Graphs have associated well known NP-hard and NP-complete problems. Therefore, we have chosen the stepped approach for not increasing unnecessarily the complexity of our solution. For a thoughtful presentation of examples on NP-hard and NP-complete problems in trees, the researcher recommends Bryant, D. "*Building Trees, Hunting for trees and comparing trees*" 1997.
- The ability to clearly spot significant fault distributions attending to a little number of characteristics. Therefore, this model increases the usability and helps to quickly spot fault distributions with instruments already known by business users instead of interpreting extremely long decision trees.

In the following sections we will discuss the assumptions behind the modelling decisions made and its relevance to answer the research questions and objectives of the project. This chapter is divided according to this stepped approach and the CRISP-DM proposed structure. Therefore:

- STEP 1: Automatic Extraction of Relevant Variables – Predictive models. (Section 5.1)
    - Binary Logistic Regression (Section 5.1.1)
    - Binary Classification Trees (Section 5.1.2)
- STEP 2: Database Relevant Path Retriever (Section 5.2)

5.1 STEP 1: Automatic Extraction of Relevant Variables – Predictive models.

According to the guidelines of CRISP-DM this section is divided according to the different models used. As the technical procedures have been already discussed in Chapter 3, in this case the two different predictive models are presented, the testing design is explained, the models are built and their results are assessed.

Two different models have been developed:

- Binary Logistic Regression, using the Scikit Learn library, and
- Binary Classification Tree, using the CatBoost library.

By applying instruments normally related with predictive analysis, such as Logistic Regressions and Decisions Trees, the artifact detects the features that would better "predict" the behavior of a transaction and would assign to them a relative importance regarding our objective variable "Result". Therefore, both models' main objective in the context of the tool is to successfully extract the most relevant variables or features for predicting if a transaction is likely to fail. Both models' effectiveness would be assessed with the same evaluation metrics to measure their predictive quality. These metrics are the ones explained in Table 2 of Chapter 3.

5.1.1 Binary Logistic Regression

The assumptions behind the Logistic Regression Model are:[45]

- The response variable must be binary. - In the project, the objective variable' values could only be "1" failed or "0" success.
- It does not require a linear relationship between the independent variables and the objective one.
- A large number of observations is often required. – In this project the cleaned dataset is of 8861entries.
- The categories must be mutually exclusive and exhaustive. – In the project's dataset all unique transactions are represented, and the categories for each variable are exclusive.
- The correlation between the independent variables must be low. – This requirement is not completely fulfilled. However, the reasons behind maintaining this correlation are already explained in Chapter 4.

Therefore, it could be observed that most of the basic assumptions are fulfilled for the data used to be modeled.

The data is separated in the train and test split exactly in the same way for both models: 70% for training and 30% for testing. In addition, the exact same data entries conform both the training and testing sets. The binary logistic regression model is developed based on the library Skicit Learn. This library has an inbuilt function that performs a gradient descend in order to choose the optimal values for the β_i mentioned in Chapter 3. These parameters represent the weight for deciding if a transaction fails or not. By squaring the absolute value of these parameters the odds are obtained. Finally, expressing these odds in a percentage would allow us to compare the weight assigned for the different variables with the other model.

Regarding the tuning of parameters to perform this gradient descent, the Python library chosen implements the *Liblinear* library[46] that support both the L1-loss and L2-loss

[45] ABDULLAHI, A. I. "*Comparison of the CatBoost Classifier with other Machine Learning Methods*". 2020.
[46] CHANG, K-W. HSIEH, C-J. WANG, X-R. "*LIBLINEAR: a library for large linear classification*". 2008.

function. Both L1-Loss and L2-Loss implementation does offer similar outcomes as the variables importance follows a very similar order of importance. However, and even if the predictive quality is similar for both Loss Functions, the percentual weight of variables differ greatly, as it is discussed in the next section 5.2. The results could be observed in the following tables (Table 5 and 6) after choosing the best model running the algorithm 5 times with 100 rounds of the gradient descent:

Table 5 - Logistic Regression Predictive Quality

Model Name	**Loss Function**	**Precision [0,1]**	**Recall [0,1]**	**F1 [0,1]**	**Weighted F1**	**AUC**
Final Model 5 - LOGIGTIC REG	L1 loss function	[0.94 , 0.65]	[0.95 , 0.62]	[0.95, 0,64]	0.91	0.79
Final Model 6 - LOGIGTIC REG	L2 loss function	[0.95 , 0.63]	[0.94 , 0.64]	[0.94, 0,64]	0.90	0.79

Table 6 - Output Feature Importance - Logistic Regression

FEATURE IMPORTANCE	**Final Model 5 - LOGIGTIC REG**	**Final Model 6 - LOGIGTIC REG**
Loss Function	L1 loss function	L2 loss function
PaymentType	85.79%	67.76%
Logical_MerchantAccountId	4.72%	5.41%
CreditClass	1.80%	4.48%
PaymentMethod	1.70%	4.46%
Brand	0.92%	2.48%
Currency	0.89%	2.48%
AccountCountry	0.86%	2.39%
CustomerCountry	0.84%	2.36%
ChannelName	0.83%	2.34%
Logical_Bin	0.83%	2.58%
Logical_TransactionId	0.83%	3.35%

5.1.2 Binary Classification Tree

Classification trees do not have distributional model assumptions.[47] [48] However, data should still adhere to some requirements such as the training and testing data set to be independently and identically distributed.[49] In the specific case of a binary classification tree the objective variable must be binary. A requirement that, as mentioned before, is satisfied.

The implementation of binary classification trees chosen, using the Catboost library, does learn following a gradient descent of binary classification trees.[50] To avoid overfitting during this process, the produced symmetric trees are evaluated based on a given evaluation metric; in this case either AUC or F1. As mentioned before in chapter 3, this process follows a gradient descent combined with bootstrap sampling. Therefore, the training dataset is re-sampled and evaluated with the test dataset in each split assigning weights to the variables of the model based on the Minimum Variance Sample (MVS)[51], is a process known as bootstrapping with cross validation.

In total, four different models have been developed attending to different parameter settings, all of them based on the bootstrap sampling algorithm MVS. These models split the training dataset in a 1:5 proportion in each tree production. Therefore, 80% of the data is used for training and evaluated with the test dataset in each training round. For 5 times, once the model is developed after 1000 learning rounds, one last prediction of the test dataset is performed with the best model acquired. Then, the best result is chosen. The results of the predictive models are displayed in the following table:

Table 7 – Classification Trees Predictive Quality

Model Name	Loss Function	Evaluation Metric	Precision [0,1]	Recall [0,1]	F1	Weighted F1	AUC
Final Model 1 - TREE	Log Loss	AUC	[0.96 , 0.65]	[0.94 , 0.73]	[0.95 , 0.69]	0.91	0.83
Final Model 2 - TREE	Cross Entropy	AUC	[0.96 , 0.64]	[0.93 , 0.76]	[0.95 , 0.69]	0.91	0.85

[47] CHATTERJEE, D.R. "*All the annoying Assumptions*". Towards Data Science. 2019.
[48] VAYSSIÈRES, M. P. "*Classification Trees: An Alternative Non-Parametric Approach for Predicting Species Distribution*". 2000.
[49] ROTH, D. "*Decision Trees*". University of Pennsylvania. 2016
[50] See Figure 12 and Figure 11 for the AUC and an example tree of model 1 for support information.
[51] Catboost official docs: https://catboost.ai/docs/concepts/algorithm-main-stages_bootstrap-options.html

Final Model 3 - TREE	Log Loss	F1	[0.96 , 0.66]	[0.94 , 0.76]	[0.95 , 0.70]	0.92	0.85
Final Model 4 - TREE	Cross Entropy	F1	[0.96 , 0.66]	[0.94 , 0.75]	[0.95 , 0.70]	0.92	0.85

These metrics would be discussed in the next subsection, comparing them with the predictive metrics of the Logistic regression. However, it could be already appreciated that for the models in table 7 their predictive power is fairly similar. Even if, once again, the importance or relevance of the different features or variables does vary, as shown in Table 8.

Table 8 - Output Feature Importance - Classification Tree

FEATURE IMPORTANCE	Final Model 1 - TREE	Final Model 2 - TREE	Final Model 3 - TREE	Final Model 4 - TREE
Loss Function	Log Loss	Cross Entropy	Log Loss	Cross Entropy
Evaluation Metric	AUC	AUC	F1	F1
PaymentType	51.52%	55.03%	57.38%	65.99%
AccountCountry	9.27%	6.15%	2.74%	2.70%
PaymentMethod	6.55%	7.02%	8.75%	8.00%
CreditClass	6.45%	5.42%	8.50%	4.71%
ChannelName	6.18%	5.60%	5.18%	4.03%
CustomerCountry	5.69%	5.37%	6.02%	3.44%
Currency	4.80%	6.35%	2.45%	3.03%
Brand	4.59%	3.97%	3.32%	2.91%
Logical_Bin	3.97%	4.33%	3.55%	4.32%
Logical_TransactionId	0.75%	0.66%	1.99%	0.74%
Logical_MerchantAccountId	0.23%	0.09%	0.11%	0.11%

5.1.3 Assessment of predictive models

In this section, the comparison of these models in table 5 and 7 would be the base of our evaluation of the predictive power of these models. In addition, in the next section the repercussions of the different assignment of weights would be further developed.

Before comparing the results it is worth mentioning that precision, recall and F1 score are extracted both for the class “0” (successful transactions) and the class “1”(unsuccessful transactions). On the other hand, weighted F1 takes into account the number of classes

in “0” and in “1” to create a single metric for the performance of the algorithm. Briefly observing the metrics mentioned, it is observed that successful transactions could be predicted better that the failed ones. These results could obey to different reasons, such as: the lower number of failed transactions to extrapolate relevant conclusions, extremely reliable specific values for certain variables that indicates that a transaction would succeed (In figure 8 we can observe an example with the Capture (CP) value for the variable “PaymentType”) or more vague characteristics for determining if a transaction is likely to fail that are not represented in the database. The last two interpretations support the initial understanding of transactions being a rigid processes and the humans as a source of faults respectively. Transactions are not prone to fail, as it is observed in CP process. However, when the human intervenes as the source of faults, the ability for the machine to predict if a transaction is going to fail it is reduced. Still, it is able to predict faults in the case of decision trees-based approaches with an F1 of 70%. Therefore, the human nature of our system is not the only source of errors as it is being described in Chapter 6.

Regarding the models in table 5, Model 5 and Model 6 offer virtually the same results for predictive quality and a similar distribution of weights for each variable, even if Model 6 does distribute this weight more evenly. For the models in table 7, the ones developed using the library Catboost, as explained in Chapter 3: it is expected for the models using *Loss Log* as algorithm for defining the Loss function of the classification to overfit less than the ones using *Cross Entropy*. However, as the same dataset is being used it is not possible to prove that this overfitting exists. In a similar way as Loss Log and Cross Entropy could lead to different outcomes regarding the level of “overfitness” based on taking into account one or two logits. The same logic applies for the difference between F1 and AUC as an evaluation metric. As already explained AUC would only focus on the distribution of either 0 values or 1 values. However, the F1 coefficient would deal at the same time with precision and recall. Therefore, with the ability to take into account both possible values of the objective variable the F1 as evaluation metric is expected to be more restrictive when deciding the feature importance, as it would assess which tree does perform better both for failures and successes.

Models based on F1 as evaluation metric, independently of the algorithm used, do perform slightly better than the models produced based on AUC metric. Also, Cross Entropy Model 2 does offer better results in their predictive metrics than model 1. However,

once again it is worth mentioning that Cross Entropy is expected to overfit more than the models developed with Log Loss. In addition, all models in table 7 do give more relative importance to the variable "PaymentType". However, the rest of the more relevant variables differs depending on the model. In Table 8, the green fields correspond with the four most relevant variables per model above a 5 %. We could observe that these variables are related either with the "PaymentMethod", such as Credit Card "CC"; the specific market where the transaction has taken place "AccountCountry", "Currency" or "CustomerCountry"; or the payment typology "CreditClass", which have the following instances: "Macro", "Micro" or "Validation".

Comparing the predictive metric from the decision tree based and logistic regression based models. It could be immediately observed that the decision tree based approach does offer better results than the Model 5 and 6. Specially regarding the Recall of the failed transactions and the AUC metric. Therefore, the decision tree models are ranked above the logistic regression based ones. In addition, for both logistic regression (Table5) and decision tree (Table 7) approaches, there is not a clear model performing better than the rest after the parameter tuning. Therefore, it is not possible to clearly rank which models are better inside both tables, attending just to the predictive metrics. However, when observing the assignment of importance for the relevant variables, the different models do perform this task differently. Thanks to the tools developed and explained in the next section, the activity of comparing the different assignments of importance is performed.

5.2 STEP 2: Database Relevant Path Retriever – Automatic Classification of Faulty Paths

A tailored solution has been developed to proceed with the analysis once the most relevant variables have been automatically retrieved. With this solution, several loops through the database are made in order to obtain all the possible paths attending to the combinations of the different values from these relevant variables. In this section we will discuss the modelling technique used, the test design, the build of the model and its assessment. Following the CRISP-DM guidelines.

5.2.1 Modelling technique

The objective of the complete process is to classify the faulty paths in a given dataset. For this purpose an specific solution has been developed to loop through all the possible combinations of the values of the categorical variables attending to their fault distribution. The code developed does only require a binary objective variable and a set of categorized variables in the dataset to effectively run and append to the transformed dataset information about the most relevant paths regarding their error distribution. As an input for this code snippet it is only required to input 2, 3 or 4 variables that could easily be retrieved both from a Logistic Regression or a Binary tree Classifier. The mathematical formulation of the processes performed by the code snippet it is defined as follows:

Taking into account that our database could be considered a matrix (X) with (j) columns, (i) entries and values that correspond to a known set. The Matrix would be notated as a standard matrix as follows:

$$X = (x_{ij})$$

Every row in this dataset correspond to a vector:

$$V_1 = x_{1j}\,, for\ j = 1, \dots, J$$

Looping through this database would allow us to find vectors for which certain values for the given relevant variables (j or columns) are coincident. allowing us to define a path as follows:

$$Path = \{V_n\}\,, where\ x_{1j} = x_{2j} = \cdots = x_{nj}$$

However, it should be taken into account that we would only take the most relevant variables to loop through the database. Therefore, the path is redefined as follows:

$$Path = \{V_n\}\,, where\ x_{1j} = x_{2j} = \cdots = x_{nj}\,, for\ j = \{relevant\ variables\}$$

Let's not forget that the maximum length of the relevant variables set could not exceed 4 members. Not all paths that we could obtain from looping through the database are relevant for our objective of classifying error paths. Therefore we should establish a

threshold to consider whether a path is relevant or not. In our case we would attend both to the business and the database characteristics to set this threshold at 20% of error rate. Note that this threshold could be dynamically retrieved attending to the fault distribution of the dataset inputted. However, in this project as the fault distribution is 13.72%, it has been already established a 20% which is roughly a 50% bigger. Therefore, in order to be considered a relevant path and taking into account that the "Result" variable could be considered as a binary variable with a value 1 when there is an error and 0 when there is not an error, the set of vectors should:

$$for\ j = "Result"\ , C_i = \ x_{ij} = 1, then:$$

$$Relevant\ Path\ (rP) = \ \{V_n\}, where\ \frac{\sum_{i=1} C_i}{n} > 0.2$$

Therefore, these relevant paths would correspond to a certain set of values of the different chosen relevant values for which the error proportion of their number of processes (n) would be bigger than 20%.

After these relevant paths has been obtained, another column is added to the dataset named "OUTPUT" where the values of the different relevant variables are concatenated as the data field for the errors, such as: If the relevant variables are: Payment Type and Customer Country, the possible paths would be: PAES (Payment Authorize and Spain), CPCZ (Capture and Czech Republic), etc. The failed transactions that could not be classified as they do not correspond with an error path of more than 20% would be classified as "Exception".

Thanks to this procedure it is possible to effectively classify the errors on the dataset according to the relevance of the typology of the paths the different entries belong to. Apart from being able to also append relevant data for the business users, such as the number of errors of the specific path, the number of processes in the path, the failure rate or even a specific indicator of relevance of the error path by weighting the number of errors and the proportion in the path. This data is relevant for business users to successfully interpret these paths or filter them out using any software that could handle .csv or .xlsx files in this specific example.

5.2.2 Test Design and Build of the model.

The dataset used has a fault rate of 8.17% and a total number of entries 15,246 taking into account Credit transactions and PA and CP subprocesses. Once the data has been properly filtered representing unique transactions, the new fault or error rate is 13.72% and the number of items is 8,861. The model developed loops through this entire database extracting the relevant faulty paths mentioned. Depending on the importance of the variables assigned in the first step by means of the predictive models, these paths typology would change and different errors would be classified. Therefore, in order to compare the quality of the different outputs from this process several metrics has been extracted to compare how relevant the classification of the error paths is performed:

- Important features: The features selected for looping the database in search of the relevant paths. We set a maximum of 4 relevant variables all of them if a feature importance is bigger than 5%.

- Number of paths: Number of different combinations of important feature values that have a fault rate higher than 20%. The bigger the number of paths the more useful would be to consider the information retrieved from them. However, the lesser the number of paths, the bigger the value for the company, as they would be able to quickly spot the relevance of these fault paths.

- Average length of paths: Average number of items in the error paths extracted from the tool. If the number of paths is similar, we expect that the bigger the number is, the less specific would be our tool to spot faults. However, the bigger the number is, the more information our variables' values offer to the decision makers in the company to take action.

- Average error per path: Average number of errors in all the paths identified by the tool. If the number of paths and the average length of paths are similar, we expect that the bigger the number, the more relevant information these paths would provide.

- Total Frequency of Error Rates (f_m) in the retrieved paths: f_i is the error frequency or fault percentage of a single error path and l_i is the number of vectors or the length of the array of vectors that conforms the path. The bigger the number is, the more relevant information the tool provides. The formula is expressed as:

$$f_m = \frac{\sum f_i l_i}{\sum l_i}$$

- Number of Exceptions: Value that represents the faults not included in an error path from our cleaned database. The lesser the value, the better our tool classifies errors or faults.

- Percentage of unclassified errors (from the number of errors): Relative frequency of unclassified errors. The lesser the value, the better our tool classifies errors or faults.

- Percentage of unclassified errors (from the complete transaction dataset): Frequency of unclassified faults in our dataset. The lesser the value, the better our tool classifies errors or faults.

Regarding the building of the model. The script runs in any Python environment and follows these differentiated steps:

- Choose the most relevant variables: The input of the process is the array of percentages assigned to the variables after the predictive model. A combination of 2, 3 or 4 variables is retrieved according to the individual percentual importance given. A maximum of 4 variables has been set to ensure readability of the path for the business user and the threshold of 5% for an individual variable to be considered as a valid input. The reasoning behind this selection is that in a theoretically equally distributed frequency of faults per independent variable, as we have 10 of them, the importance would be 10%. Therefore, half of this “perfect” importance is considered as a minimum requirement to be considered as a relevant variable for retrieving the paths.

- Retrieve the unique values of these variables: After 2 to 4 variables are selected, the retrieval of their unique values[52] is performed using an inbuilt function from the Pandas library of Python for Series objects.

- Group transactions according to all the possible combinations of values: After retrieving these unique values the artifact searches all the possible combinations using the *meshgrid* function from the library NumPy[53], that is automatically inputted as a filter in the database. In this way it is achieved the initial classification of possible paths.

- Check the error distribution of all paths: While retrieving all the possible paths, the fault distribution for that group of transactions is checked. If it is above a certain threshold, in this case 20%, the path name and extra information such as the error number, distribution and the number of items in the path is stored in a JSON object. It has been decided to establish a 20% fault rate in the paths for 3 different reasons: Firstly, after testing the tool it has been discovered that generally paths would have a surprisingly high error rate compared to the error distribution of the cleaned dataset (13.72%). Secondly, it does represent a 50 percent increase of the fault rate mentioned. There is the possibility to retrieve this value dynamically based on the 50% percent increase, However, the last reason that inclined the decision of setting this percentage as 20% is that, from a business point of view, it is a relevant percentage of failure transactions for business users to focus their attention on these paths.

- Append the selected combination of values as the name of the path and extra information of these faulty paths to the transformed dataset: Once this JSON object is retrieved, the relevant path's names and their characteristics are appended to the transformed database, in this case an Excel file.

[52] Pandas official documentation: https://pandas.pydata.org/docs/reference/api/pandas.Series.unique.html
[53] NumPy official documentation: https://numpy.org/doc/stable/reference/generated/numpy.meshgrid.html

5.2.3 Assessment of the model

Attending to the characteristics mentioned for tuning our model mentioned before, such as: 5% importance threshold, between 2 or 4 relevant variables and an error rate for the paths higher than 20%. The output of the models that fulfill these characteristics are presented in table 9:

Table 9 - Automatic Fault Path Classification

PATHS CLASSIFICATION	Final Model 1 - TREE	Final Model 2 - TREE	Final Model 3 - TREE	Final Model 4 - TREE	Final Model 6 - LOGISTIC REG
Loss Function	Log Loss	Cross Entropy	Log Loss	Cross Entropy	L2 Loss Function
Evaluation Metric	AUC	AUC	F1	F1	
Important features	PaymentType AccountCountry PaymentMethod CreditClass.	PaymentType, PaymentMethod, Currency, AccountCountry.	PaymentType, PaymentMethod, CreditClass, CustomerCountry	PaymentType, PaymentMethod.	PaymentType, Logical_MerchantAccountId
Number of paths	56	36	53	2	2
Average length of paths	35.29	55.92	37.32	1015.0	1238.0
Average error per path	20.20	31.61	21.39	569.0	603.5
Total error frequency of retrieved paths (in percentage)	57.24%	56.53%	57.33%	56.06%	48.75%
Number of Exceptions	85	78	82	78	9
Percentage of unclassified errors (from the number of errors)	6.99%	6.41%	6.74%	6.41%	0.74%
Percentage of unclassified errors (from the complete transaction dataset)	0.56%	0.51%	0.53%	0.51%	0.06%

Only one of the logistic regression models does assign a second feature an importance of 5%, in this case the L2 loss one (Model 6). Comparing the Logistic regression model's output with the ones retrieved from the decision tree-based approach. It is clearly shown that this model does outperform the rest in the metrics: "Number of Exceptions" and "Percentage of Unclassified Errors". However, even if Model 6 does classify a higher number of faults, the quality of the relevant faulty paths retrieved is significantly lower than the rest of models. This lower quality is represented by the metric "Total error frequency of retrieved paths". Therefore, it is considered that model 6 performs worse than the rest. After carefully reviewing the theory behind these models, seems that this different behavior is caused by intrinsic basic characteristics of the predictive models, as the logistic regression one would treat each of the variables as independent ones, while the decision tree approach would take into consideration how the combination of the different values performs together every time another decision node is split. Therefore, the decision tree approach outperforms the logistic regression one for different reasons:

- As mentioned before, based on the metrics retrieved during this second step of the model, the fault distribution of the retrieved paths is roughly ten percent less.
- Regarding the assignation of importance given to the variable "Logical_MerchantAccountId", the decision tree approach give a minimum weight while for both logistic regression tree models is the second more important variables as shown in Table 6 and 8. Therefore, after reviewing it is concluded that the faults related with a value of "0" in the for the "Logical_MerchantAccountId" does correspond with the error code 100.150.200 "registration does not exist". This error is related with the 3D secure verification and it is classified by the rest of the models at least as "PA" Payment Authorize and "CC" Credit Card. Therefore, the decision tree approach effectively spot this relationship and do not assigned importance to this variable, while the logistic regression does only improve the classification by gathering most of the errors under the "PA" and "Logical_MerchantAccountId = 1", which does not offer any extra information compared to the rest of the models.

Regarding the Tree based models, it could be observed that both models 1 and 3 have similar outcomes independently of the evaluation metric chosen. Model 1, does output the same paths and metrics every time. However, model 3 does vary more in the

outcomes, slightly outperforming model 1 or offering the same values as the feature importance of *CustomerCountry* and *AccountCountry* tend to exchange positions in the list of ordered features importance. Also *Creditclass* does fight for the 3rd or 4th position for the model 3. While in model 1 the order of importance is respected. Therefore, we could consider that even if the results are similar, the model that evaluates according to F1 score does offer slightly more inconsistent results.

For model 2 and 4, the ones that use Cross Entropy as a Loss Function the outcomes tend to be different. Model 4 does only consider two variables as relevant for all the runs, while Model 2 in all the runs has selected the maximum allowed number of variables: 4, and all of them assign the same order of importance to the selected variables. Therefore, it is considered that even if both models do not offer the same insights, the results are consistent. Model 4 is especially interesting as the third variable of the output feature relevance list, does not get values above the 5% threshold. At the same time, this method gives *PaymentType* the biggest importance (66%) among the decision tree-based models. A relevant characteristic for the model that, according to the theory in chapter 2, would be the most inclined to overfit (Cross Entropy) and the most generalized in its evaluation (F1 as evaluation metric).

Regarding the measures extracted to compare the models it is observed that even if the results for Model 1, Model 2 and Model 3 are similar, there are important differences. As mentioned before the differences after extracting the paths on the model 1 and 3 are really narrow. However, we could notice a difference depending on the selected variable "*CustomerCountry"* or "*AccountCountry*". These variables share a high correlation as the Cramer's V analysis showed in Figures 9 and 10. However, due to business reasons it has been decided to maintain them as it is explained in Chapter 3. When a comparison is made between both model 1 and model 3 with model 2, it is observed that these models do offer different outcomes:

- Model 2 offers a higher average number of errors in longer paths. Therefore, not relevant conclusions could be extracted from this comparison.
- In Model 2 there are fewer paths for the same number of features, a characteristic that offers interesting insights to business users as there are fewer variables to compare. However, a bigger number of paths *ceteris paribus* do offer more specificity when spotting these failures based on these features.

- The rest of the extracted measures (Percentage of unclassified errors, number of exceptions, total error frequency of retrieved paths) do not vary in a significant way for these models.

Model 4 does offer us interesting insights regarding its evaluation metrics. As its results offer us a similar number and percentages of faults not classified. However, as it only produces 2 paths, the other metrics that depend on the number of paths does not give us a proper picture of the tool, except from the total frequency that weights this size effect. The paths extracted by this model are: (PA, CC), with a number of items of 2,029 and (PA, DC) with one transaction. PA stands for "Payment Authorization" and it is a value of "*PaymentType*". While "CC" means "Credit Card" and "DC" Debit Card. For the CC path there is a 56% of failure rate, while for the "DC" path there is only one transaction. In this case, this model gives us an important insight from a general point of view. The transactions that failed are not captured transactions. Therefore, there is an issue with the payment authorization process. On the other hand, the other models offer us a more specific view where we can observe that specific markets have failure rates from 80% to 100%.

APPENDIX CHAPTER 5

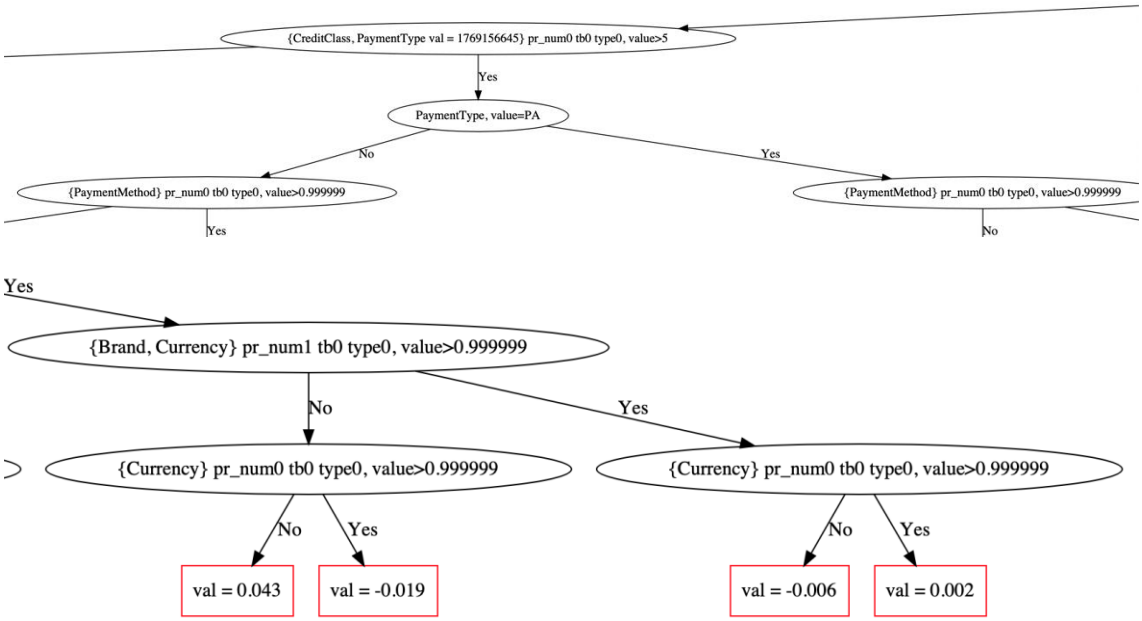


Figure 11 - Extract of Classification Tree - Model 1

This is an extraction of one of the multiple decision trees created in the model 1. Terminal leaf "val" are computed to optimize the Loss value. Please, also note that the Features between { } are sets of features values. Please refer to Catboost official docs for more information: https://catboost.ai/docs/concepts/python-reference_catboost_plot_tree.html and to the Github repo of Catboost: https://github.com/catboost/catboost/issues/1389

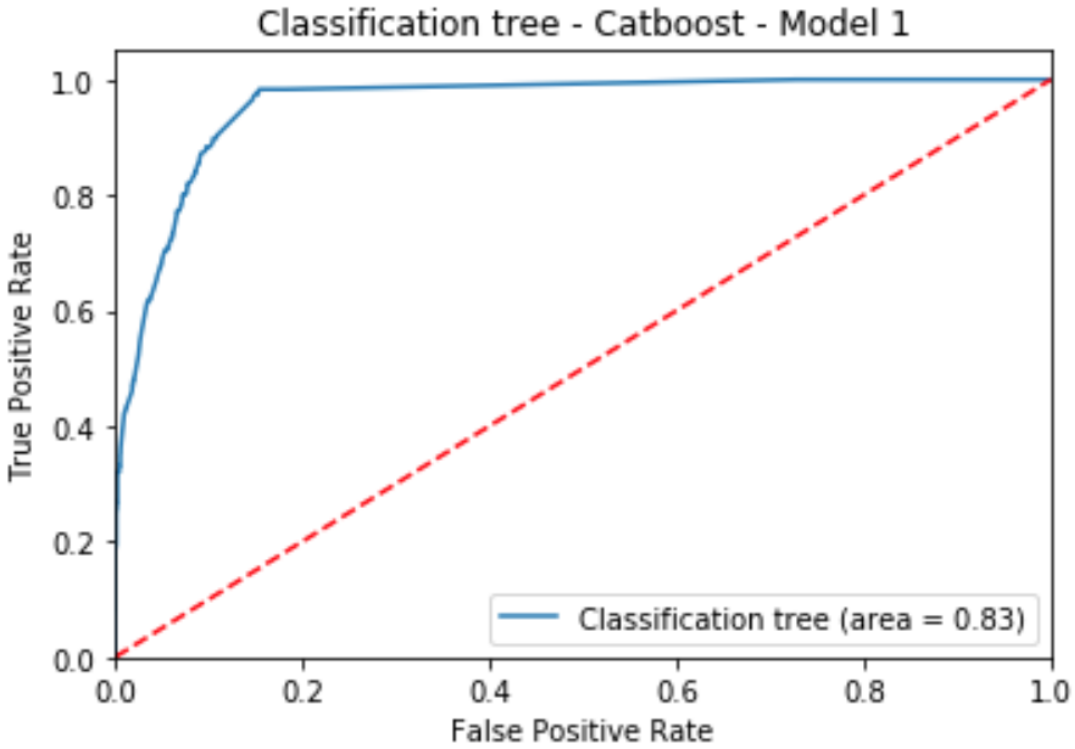


Figure 12 - Roc Curve - Model 1

6 EVALUATION

In this chapter we evaluate the proposed tool trying to answer the sub question:

"*How to evaluate and interpret these models?*"

In order to effectively achieve this objective, the chapter focuses on understanding the efficacy of the tool. According to the CRISP-DM framework we would assess the validity of the different models used and argue about the processes involved. Therefore, it is being discussed during the following lines if the objectives of the data mining project have been fulfilled from a business point of view and according to our research question.

The automatic classification extracting relevant paths, given the constraints imposed by the researcher, is performed by Model 1, Model 2, Model 3, Model 4 and Model 5. Therefore, one of the Logistic regression models (number 6) and all the models based on decision trees (number 1, 2, 3 and 4).

It has being discussed in the model assessment part that, even if the predictive quality could be considered similar, Model 6 does not perform as the rest of the models when applied the script for classifying error paths. While this model assigns a relatively higher importance the variable "*Logical_MerchantAccountId*", the rest of the models do spot the characteristics in the real life scenario that made the transactions to fail. Roughly, the characteristics identified where having a PA "Payment Authorize" process with a "CC" credit card payment. For this group of transactions, after eliminating the ones that were captured "CP" later in that specific day, the fault distribution was 56% percent. In addition Model 1, 2 and 3 offered more specific insights such as "DUMMYNA" value for *AccountCountry* or *CustomerCountry* with a fault rate around 80% and related with 3D secure failures in PA and CC transactions. All these models do also identify problems with currencies either explicitly with "*Currency*", such as model 2, or indirectly through *AccountCountry* or *CustomerCountry*. Also, Model 1 and Model 3 identified explicitly the importance of the variable *CreditClass* when its value is "Validation" for PA and CC. Ultimately one of the highest sources of errors for specific markets due to the different currency payments mentioned in the business understanding phase. Therefore, from a

business point of view, decision tree-based models are ranked as more effective to identify these paths than the Logistic Regression one.

From the decision tree-based models mentioned above we can conclude that there is no model better than the other but characteristics that would incline us to use each of them. If we want to understand what is failing from a general point of view, we are more inclined to use the model 4. On the contrary, If we intend to acquire more specific insights we are inclined to use any of the other 3 models. From these models, model 2 proved to be less consistent in their feature importance assignment than the others. This characteristic turns out to be interesting because it includes among its features the variable "*Currency*" that is relevant from a business point of view as it is being discussed in this section.

The situation of the business when the data used for the analysis was extracted (29 of December, 2020) it is not a business-as-usual environment. The company was in the middle of the adoption of new protocols to be compliant with the already mentioned PSD2. Therefore, the proportion of failed transactions increased during that period of time and increased while the rest of the merchants, banks and payment service providers started adopting or increasing the rigidness of these protocols in their own adoption phase during January 2021. The causes of the failed transactions were several:

- The new payment authorization required the "purchase" of one cent to verify the payment method used (Debit card, Credit Card) before being able to purchase a product or a service. This method proves to be complicated because users pay in different currencies and there were problems for some of them.
- Mismatching payment service providers implement the new 3D secure systems.
- End users not being able to finish the purchase journey: mainly not inputting the right data or not accepting to pay one cent to completely authorize their first purchase.

The models developed did recognize these patterns based on the limited data field that were inputted. "Payment Type" is by far the most relevant variable in all the models and the paths created did only retrieve "PA" Payment Authorization as a relevant value. "Payment Method" was also in all the models, and the field "CC" Credit Card the field one more repeated in the extracted paths. Therefore, it is observed that the failed processes

corresponded to the CC flow of the business, while DC (Debit Card) transactions and alternative payment providers transaction flows did not get strongly affected by this implementation. Model 1, Model 2 and Model 3 did also identified the paths with a high fault distribution that were related with the validation process and the currency. However, they did so in different ways. Model 1 and Model 3 identified these paths attending to the value "Validation" inside the relevant variable "*CreditClass*". Meanwhile, model 2 specifically identified "*Currency*" as a relevant field while "*CreditClass*" received the fifth position on the model. Therefore it was not able to specifically detect the "validation" process. However, Model 1 and Model 3 did mention either *AccountCountry* or *CustomerCountry*, variables that are highly correlated with the variable "*Currency*". Independently of their variance on these selection all of them classified most of the fault processes effectively and offered us specific and relevant information about the faulty paths.

Therefore, the artifact does effectively classify fault paths and, assuming the business users have knowledge from the company processes, the information retrieved from the models is easily interpretable.

In addition, the tool does not only fulfill the preliminary objective of the modeling. It also provides data both graphically and appending extra information to csv and excel files with relevant data from the decision paths. After running more analysis with the head of operations of the company, these are the outcomes from the tool:

- It takes less than 2 minutes to perform the required operations and offer the information to the user.
- It is easy to install and use. As it does not require any programming language.
- The output information, especially the graphical one, is really useful for spotting right away if there is a problem and where.
- Normally, the operations team deal with error codes and specific error logs to try to understand the big picture. However, this tool does provide them insights the other way around.

The highly positive impact has been possible thanks to the collaboration of the business users, specially the head of operations, who had recurrent meetings with the researcher in order to assist in the business understanding phase and exposing the demands and requirements for this tool to be considered useful for the company. As a summary: “It should not take us a lot of time to analyze and it should show us relevant and easily interpretable information”. Therefore, we could conclude that not only our data mining model has been fulfilled, but also our design science approach for the company.

7 LIMITATIONS, FURTHER RESEARCH AND CONCLUSION

In this chapter we discuss the different limitations of the research as well as further directions from improvement. Finally we will conclude this paper by clearly showing the results of the research.

7.1 Limitations and Further Research

From a general point of view, the research has been performed in 3 months due to the unavailability of starting the contract earlier. This time constraint limited the research scope by focusing on the development of the tool for dealing with datasets on a daily basis and with data related to a specific problem that the company had at that time: the implementation of new payment protocols that affected the number of fault transactions. Therefore, there was not time enough to perform analysis including a business as usual scenario following formal academic procedures. However, after briefly testing these models in an informal context, signs are found indicating that the models proposed do have that potential to effectively classify these errors. Even if this business as usual environment was not clearly among the initial objectives, during this research there have been discovered paths for further research apart from the tool developed. Such as, the potential in the implementation of the Crammers' V modelling to perform dd-FD analysis in this specific sector for time series analysis, such as the ones performed in the industrial sector with analysis such as the MCA.

The Crammers' V metrics, mentioned in the theoretical chapters and that is automatically extracted by the developed tool, have the potential to imitate the processes from industrial companies dealing with high quality requirements in their production chain, once data has been properly prepared and enough historical data is provided. In the case study, as we were dealing with an exceptional situation, where information was not always properly retrieved from the different systems due to the newly implemented protocols. It has been chosen a more flexible approach based on machine learning dd-FD methods. The decision of opting for this flexible approach does also correspond with an important difference that limits our study. In our definition of "system" customers and partners are included, the system is not just a set of sensors and components from where data is retrieved in a controlled environment.

From a development point of view, this paper has also been affected by time constraints. A rapid prototyping production methodology has been selected by the researcher in order to validate and test assumptions in the most efficient way. Different machine learning models, data preparation techniques have been tested with a different degree of success for trying to assess the potential of retrieving relevant information from the dataset, in a learning loop that ultimately made the researcher to choose to transform into multivalued variables the inputs to the model and to select both the Catboost Classification library and the Logistic regression Scikit Learn library to extract the features importance in the model. In this paper it has been only presented the last outcomes from the research regarding the tested models. However, the learning loop could still be improved. For example, in further research we could consider this time that we will create another variable that combines currency and country in a composite key style to avoid the high correlation of these variables. The possibilities of the testing and learning loop could be further improved.

Apart from the mentioned limitations and further research possibilities, there is one clear outcome from this tool that allows the researcher to indicate an important future research path: this tool does not rely on any predefined fault distribution. Therefore, if the inputted set of variables is formed by multivalued ones, the artefact would be able to automatically detect and find these relevant fault paths independently of the sector as long as it does correspond to a binary classification problem.

7.2 Conclusion

The RS in the paper is:

How to develop and evaluate models for data-driven failure diagnosis?

In this paper we have developed a tool that relies only in the provision of a categorized dataset and does not depends on a predefined fault distribution. In order to achieve this objective a definition of what it is understood as a fault in the case study and what is failure diagnosis has been provided in Chapter 2. According to the sub question (1): “What is Failure Diagnosis”.

As a result, fault and incident as the ITIL 4 described it, are considered synonyms for this paper and, it also has been presented a link between dd-FD and the ITIL incident management creation of value. Consequently, in the case study, it has been included the human factor both from a customer point of view and from a business user view. Therefore, the tool should deal with the fault originated by customers and facilitate and support the tasks of the operations team to effectively deal with failures in the system.

In chapter 3 a more technical overview of the application of data driven techniques has been presented answering the sub question (2): *What are appropriate models for data-driven failure diagnosis (dd-FD).* To be able to develop the highly reliable models the focus is on industrial environments like chemical industries where there is no room for failure. Based on the different techniques applied, the ones dealing with uncomplete data and used for binary classification of categorical variables has been explained deeply. In this case decision tree-based and logistic regression models.

Thanks to the information retrieved in the chapter above the researcher has been able to answer the sub question (3): "*How to develop an explanatory model for data-driven FD for data from a company in the financial sector".* The tool developed has demonstrated that it is capable of effectively classifying faulty transactions by using automated means and attending to characteristics that are useful from a business point of view. The stepped approach that made this objective possible to achieve, and that has provided another example on how predictive models could effectively we applied for extracting explanatory insights, consists of:

- A data preparation phase were the different variables are categorized in order to extract relevant information from them,
- The application of predictive models, mainly the systemic creation of classification trees in order to find an optimal one using gradient descent and boosting techniques, to extract the most relevant features,
- Once these variables are retrieved, a custom made script loops through the dataset extracting relevant fault paths according to the values of these selected relevant variables.

Finally, in order to answer the last sub question (4): *How to evaluate and interpret these models?.* Different evaluation measures has been produced, explained and compared. Arriving to the conclusion that from the different decision tree-based models, there

is not a model that clearly outperforms the rest. However, the tool is it also evaluated according to the functionality that offers to the company. In this case, once these paths are extracted, the information is presented to the business users in a graphical way and provides them an enriched dataset both in .csv and .xlsx format that could be easily understood by any business intelligence tool for further visualization of the data.

The tool developed is scalable for the business not only in the amount of data that could be processed but in how it could be easily applied to other datasets after a data preparation phase. The artifact that also shows potential to support daily operations in a business as usual environment. However, more testing is needed. Also, the output of the data could be assessed and compared taking into account the time dimension. However, it would be extremely interesting how this characteristic could be effectively automatically modeled in the future.

8 REFERENCES

AAMODT, A. NYGÅRD, M. "*Different roles and mutual dependencies of data, information, and knowledge - An AI perspective on their integration*". Data and Knowledge Engineering, vol 16. El Sevier. The Netherlands. 1995. pp 191-222

ANDERSON, R. J. "*Security Engineering – A Guide to Building Dependable Distributed Systems*". Wiley. Second Edition. 2008.

BARTOLINI, C. STEFANELLI, C. TORTONESI, M. "*SYMIAN: Analysis and Performance Improvement of the IT Incident Management Process*". IEEE transactions on network and service management, Vol. 7, No. 3, 2010.

CHANG, K-W. HSIEH, C-J. WANG, X-R. "*LIBLINEAR: a library for large linear classification*". Journal of Machine Learning Research 9. 1871-1874. 2008.

CHAPMAN, P. CLINTON, J. KERBER, R. KHABAZA, T. REINHARTZ, T. SHEARER, C. WIRHT, R. "*CRISP-DM 1.0 - Step-by-step data mining guide*". SPSS. 2000.

CHATTERJEE, D.R. "*All the annoying Assumptions*". Towards Data Science. 2019.(Online) Available at: https://towardsdatascience.com/all-the-annoying-assumptions-31b55df246c3 . Last visited on: 19-07-2021.

CHEN, H. CHIANG, R. H. L. STOREY, V. C. "*Business Intelligence And Analytics: From Big Data To Big Impact*". MIS Quarterly Vol. 36 No. 4. pp. 1165-1188. 2012.

CHEN, M. ZHENG A. X. LLOYD, J. JORDAN, M. I. BREWER, E. "*Failure Diagnosis Using Decision Trees*". IEEE Computer Society. University of California at Berkeley and eBay Inc. United States of America, 2004.

FERGUSON, L. R. "*Meat and Cancer*". Meat Science. Volume 84, Issue 2. Pp. 308-313. 2010.

FURTAK, S. AVITAL, M. PEDERSEN, R. U. "*Sensing the Future: Designing Predictive Analytics with Sensor Technologies*". Association for Information Systems. AIS Electronic Library (AISeL). ECIS 2015 Completed Research Papers. 2015.

GÜLÜK, O. KALAGNANAM, J. LI, M. MENICKELLY, M. and SCHEINBERG, K. "*Optimal Decision Trees for Categorical Data via Integer Programming*". Cornell University. In cooperation with: IBM Research, Google and Oxford University. 2019.

HANCOCK, J. T. KHOSHGOFTAAR, T. M. "*CatBoost for big data: an interdisciplinary review*". Springer. Journal of Big Data. 2020.

HEVNER, A. R. MARCH, S. T. PARK, J. RAM, S. "*Design Science in Information System Research*". Mis Quarterly Vol. 28 No. 1. 2004. pp. 75-105

HO YU, S. "Exploratory data analysis in the context of data mining and resampling." International Journal of Psychological Research 3. 2010.

IBRAHIM, A. A. RIDWAN, R. L. Muhammed, M .M. Abdulazid, R. O. SAHEED, G. A. "*Comparison of the Catboost Classifier with other Machine Learning Methods*". International Journal of Advanced Computer Science and Applications, Vol 11. African Institute for Mathematical Sciences. University of Tlemcen. University of Silesia. 2011.

"*ITIL Foundation. ITIL 4 Edition."* Axelos Global Best Practice. 2019.

"*ISO/IEC 7498-4:1989*" Information processing systems. Open systems Interconnection. Basic Reference Model. Part 4: Management Framework. ISO. 1989. (Last Reviewed: 2006)

KAVULYA, S. P. JOSHI, K. DI GIANDOMENICO, F. NARAMSIMHAM, P. "*Failure Diagnosis of Complex Systems*". Carnegie Mellon University, AT&T Labs Research, ISTI-CNR. United States of America and Italy, 2012.

KURNET, S. (E.). "*Strategies in Failure Management Scientific Insights, Case Studies and Tools*". Springer. Humboldt-University Berlin. Berlin, Germany. 2018.

KWON, J-H. LEE, S-B. PARK, J. KIM, E-J. "*Association Rule-based Predictive Model for Machine Failure in Industrial Internet of Things*". IOP Publishing. Conf. Series: Journal of Physics: Conf. Series 892. 2017.

LEWIS, L. DREO, G. "*Extending Trouble Ticket System to Fault Diagnostics*". IEEE Network. 1993.

LIU, W. PENG, Z. CUI, D. "*A Review of Industrial Fault Diagnosis Based on Data-driven Methods*". Atlantis Press. Advances in Intelligent Systems Research, volume 147 International Conference on Network, Communication, Computer Engineering (NCCE). 2018.

MAO, L. "*Cross Entropy Loss VS Log Loss VS Sum of Log Loss*". (Online). Available at: https://leimao.github.io/blog/Conventional-Classification-Loss-Functions/ . Last Updated: 18-07-2020. Last visited: 30-06-2021.

MEI, J. HOU, J. KARIMI, H.R. HUANG, J. "*A Novel Data-Driven Fault Diagnosis Algorithm Using Multivariate Dynamic Time Warping Measure*", Abstract and Applied Analysis, vol. 2014. 2014.

NISHISATO, S. "*Correlational Structure Of Multiple-Choice Data*". GREENACRE, M. (ed.) "*Multiple Correspondence Analysis and Related Methods*". Statistics in The Social and Behavioral Sciences Series. 2006.

NOR, N. M. HUSAIN, M. A. CHE HASSAN, C. R. "*A review of data-driven fault detection and diagnosis methods: applications in chemical process systems*". De Gruyter. Rev Chem Eng. Iowa State University. 2019.

PROVOST, F. FAWCETT, T. "*Data Science and its relationship to Big Data and data-driven decision making*". Leonard N. Stern School of Business New York University. NY, USA. 2012. Pp. 19.

QIN, S. J. "*Data-driven Fault Detection and Diagnosis for Complex Industrial Processes*". IFAC. Proceedings of the 7th IFAC Symposium on Fault Detection, Supervision and Safety of Technical Processes. Barcelona, Spain. 2009.

ROCHLINL, G.I. "*Reliable Organizations: Present Research and Future Directions*". of Contingencies and Crisis Management. Volume 4, Issue 2 Journal. 1996. p. 55-59

ROTH, D. "*Decision Trees*". (Online) CS 446 Machine Learning. University of Pennsylvania. 2016. Available at: https://www.cis.upenn.edu/~danroth/Teaching/CS446-17/LectureNotesNew/dtree/main.pdf Last visited on: 19-07-2021

SCHUEFFEL, P. "*Taming the Beast: A Scientific Definition of Fintech*". Journal of Innovation Management. SSRN Electronic Journal. Switzerland. 2016. Pp. 32-54

SUJATHA, M. PRABHAKAR, S. DEVI, G. L. "*A Survey of Classification Techniques in Data Mining*". International Journal of Innovations in Engineering and Technology (IJIET). Vol. 2 Issue 4. 2013.

VERSCHUREN, P. HARTOG R. "*Evaluation in Design-Oriented Research*". Quality & Quantity (2005) 39. Springer. 2005. pp.733-762

VAYSSIÈRES, M. P. PLANT, E. R. ALLEN-DIAZ, B. H. "*Classification Trees: An Alternative Non-Parametric Approach for Predicting Species Distribution*". Wiley. Journal of Vegetation Science. 2000.

YIN, S. DING, S. X. HAGHANI, A. HAO, H. ZHANG, P. "*A comparison study of basic data-driven fault diagnosis and process monitoring methods on the benchmark Tennessee Eastman process*". El Sevier. Journal of Process Control 22 1567–1581. 2012.