

Process Mining: Uncovering hidden opportunities to improve a collection process

Carolina Rocha da Silva

Master's Dissertation

Supervisor at FEUP: Prof. Eduardo Gil da Costa



Master in Industrial Engineering and Management

2022-02-27

Abstract

The growing complexity of processes is a challenge that many organizations are facing today. With the digital transformation and subsequent increased availability of process data, it has become gradually more relevant for companies to have techniques to make a good use of such assets, once the process awareness potentiates the identification of inefficiencies and supports improvement actions. Thus, the current scenario led to new opportunities for disciplines such as Process Mining, which uses event data to create a clear visualization of the process, and data analysis techniques to gain insights into business processes and support decision-making.

The present dissertation covers the implementation of a Process Mining application for a collections process, comprising also the theoretical knowledge gain on the topic and a final set of recommendations to improve the process. The complete implementation of the Process Mining tool to deal with part of an Order to Cash process of a service provider company, was delivered in order to support the company's process analysis, monitoring, and improvement.

The first major part of the document concerns the exploration on the topic of Process Mining, the comprehension of the domain framework, the current stage of its application at the companies, and possible benefits from its use.

The second fundamental part covers the Process Mining implementation itself. It comprehends the analysis of a real collections process, from a Business to Consumer service provider using the Process Mining tool Celonis. Furthermore, the identification of inefficiencies present in the process and recommendations for improvement are also here provided.

Lastly, a predictive model to support the analysis and reinforce recommendations regarding the next activities of the process, was implemented. Its core considerations, regarding the topic of this document, are here presented.

The goal of this document is to demonstrate the value of Process Mining in identifying and addressing inefficiencies in a process, ultimately leading to process improvement.

Keywords: Process Mining, Technology Transformation, Process Improvement

Resumo

A crescente complexidade dos processos é um desafio enfrentado por muitas organizações hoje em dia. Com a transformação digital e o subsequente aumento da disponibilidade de dados de processos, tornou-se cada vez mais relevante que as empresas façam uso de técnicas por forma a fazer um bom aproveitamento desses ativos, uma vez que a consciência acerca do processo potencializa a identificação de ineficiências e atua como suporte para ações de melhoria. Dessa forma, o cenário atual abriu novas oportunidades para disciplinas como *Process Mining*, que usa dados de eventos para criar uma visualização clara do processo, e técnicas de análise de dados para a melhor compreensão dos processos das empresas, e apoio na tomada de decisões.

A presente dissertação reflete a implementação de uma aplicação de *Process Mining*, abrangendo também o processo de aquisição de conhecimento teórico sobre o assunto, e a elaboração de um conjunto de recomendações finais por forma a melhorar o processo. A completa implementação de uma ferramenta de *Process Mining* para lidar com parte de um processo de *Order to Cash* de uma empresa fornecedora de serviços, foi entregue para apoiar a análise e monitorização do processo da empresa e promover melhoria ao mesmo.

A primeira parte do documento concentra-se na exploração do tema de *Process Mining*, na compreensão do enquadramento do domínio, da etapa atual da sua aplicação nas empresas e dos possíveis benefícios de seu uso.

A segunda parte fundamental reflete a implementação propriamente dita de *Process Mining*. Abrange aqui a análise de um processo real de cobrança, de uma empresa do segmento *Business to Consumer*, usando a ferramenta de *Process Mining* Celonis. Além disso, a identificação das ineficiências presentes no processo assim como recomendações para melhoria são também aqui fornecidas.

Por fim, foi implementado um modelo preditivo acerca da atividade seguinte do processo, para apoiar a análise e reforçar as recomendações. As suas componentes fundamentais, face ao tema deste documento, são aqui apresentadas.

O objetivo desta dissertação é demonstrar o valor da tecnologia de *Process Mining* na identificação e abordagem de ineficiências num processo, levando, finalmente, à melhoria do processo.

Palavras-chave: Process Mining, Transformação Tecnológica, Melhoria de Processos

Acknowledgments

This dissertation marks the end of a five-year journey that has been both challenging and rewarding. It is the culmination of an extensive personal and intellectual growth. It represents not just the closure of five years of hard work, dedication, and perseverance, but also a starting point for another path of learning and discoveries.

Is impossible to thank each and every one of who were part of it. Nevertheless, I would like to thank those who were a crucial support for the journey and for this project.

First, I would like to extend my deepest appreciation to my supervisor at Deloitte, João Santos, for the expertise and feedback which were invaluable, and I am deeply grateful for the provided opportunities. I would also like to thank Miguel Vale for his constant availability and will to help me during the project, even for the silliest questions.

I would like to thank my supervisor at Faculty of Engineering from the University of Porto, Eduardo Gil Costa, for being also so responsiveness to my doubts.

And because the journey would not be possible without them: I would like to express my deepest gratitude to my parents for their unwavering support and encouragement throughout all my life stages. For their constant guidance and love, and for always believing in me, even during the most challenging times, even when I didn't believe.

To the friends, who throughout the years have remained in touch. For the coffee, hugs, and the feeling of 'home'. To the friends who did not stick around, but who were part of the journey. For the shared tears, smiles, and panic over all the assessments.

To Bruno, for his constant support and understanding throughout all the moments. My anchor, my confidant, my calming breeze, and my best friend. For his love and tireless patience.

Your support has been essential, and I am incredibly grateful for all that you have done for me.

"Success is not the key to happiness. Happiness is the key to success. If you love what you are doing, you will be successful."

Albert Schweitzer

Contents

1	Introduction	1
1.1	Project motivation	1
1.2	Key project objectives	1
1.3	Process Mining at Deloitte Portugal	2
1.4	Project methodology	2
1.5	Document structure	3
2	Theoretical Framing	5
2.1	Technology Transformation	5
2.1.1	Digital Transformation	5
2.1.2	Data Burst	6
2.2	Linking process information with process efficiency	7
2.2.1	Data Science	7
2.2.2	Process Science	9
2.3	Process Mining	9
2.3.1	Concept	10
2.3.2	The emergence of Process Mining	11
2.4	Process Mining in organizations	12
2.4.1	Literature review on Process Mining in Business Management	12
2.4.2	Opportunities for Process Mining	13
2.4.3	Current state of Process Mining in organizations	15
2.4.4	Core processes in Process Mining	17
2.5	Predictive Business Process Monitoring	17
2.5.1	Recurrent Neural Networks	18
2.5.2	Long Short-Term Memory for Sequence Modeling	19
2.5.3	Loss Functions	19
3	Problem scope and description	21
3.1	Business context	21
3.2	Operations overview	22
3.3	Data Characterization	23
4	Project implementation	25
4.1	Data understanding and treatment	25
4.2	Data integration on Process Mining tool	26
4.2.1	Connection and Extraction	26
4.2.2	Transformations	26
4.2.3	Data Model	28
4.3	Analysis	29

4.3.1	Dashboards	29
4.3.2	Costs calculation	30
4.4	Predictive Model	30
4.4.1	Considerations	30
4.4.2	Implementation	31
5	Results	33
5.1	Process Variant and Process Explorer views	33
5.2	General Process analysis	35
5.3	Operational analysis	36
5.4	Financial analysis	38
5.5	Subject analysis	40
5.6	Claims analysis	42
5.7	Conformance view	43
5.7.1	Comparison between real process flow and ideal	43
5.7.2	Root Cause Analysis	43
5.8	Predictive Model outcome	46
5.9	Recommendations	46
6	Conclusions and future work	49
6.1	Conclusions	49
6.2	Future work	50
A	Dataset	55
B	Dashboards	57

Acronyms and Symbols

B2C	Business to Consumer
BPM	Business Process Management
ERP	Enterprise Resource Planning
IVR	Interactive Voice Response
KPI	Key Performance Indicator
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
O2C	Order to Cash
PD	Payment Deadline date
RNN	Recurrent Neural Networks
SMS	Short Message Service

List of Figures

2.1	Adoption of basic Process Mining types (adapted from Gartner)	11
2.2	Expectations on Process Mining (adapted from Deloitte)	14
2.3	Process Mining use cases (adapted from Gartner)	16
2.4	Areas of Process Mining application.	16
2.5	Recurrent Neural Network. Source: LeCun et al. (2015)	19
3.1	The two major phases of the process	22
3.2	Payment phase	23
4.1	Data Model	29
5.1	Most common process variant - Variant Explorer	34
5.2	Process with all activities and connections - Process Explorer	34
5.3	General Dashboard	35
5.4	Operational Dashboard	37
5.5	Financial Dashboard	39
5.6	Subject Dashboard	40
5.7	Claims Dashboard	42
B.1	General Dashboard	58
B.2	Financial Dashboard	59
B.3	Operational Dashboard	60
B.4	Subjects Dashboard	61
B.5	Claims Dashboard	62

List of Tables

5.1	Top 5 Subjects	41
5.2	Top 8 detailed Subjects	41
5.3	Top 4 Violations	44
5.4	Root causes for Issuing invoice followed by payment deadline date	45
5.5	Root causes for Subject creation	45
5.6	Subjects on clients list	48
A.1	Data set	56

Chapter 1

Introduction

The purpose of the first chapter of this dissertation is to succinctly present the motivation behind the project, as well as its key objectives, and the business context where the challenge took place.

1.1 Project motivation

The choice and further development of the project is built on two major motivations.

The first one is the fact that Process Mining is a new technique for the understanding and improvement of business processes, and is at an early stage of adoption and exploration in companies. This active research field has brought important developments, however, evidence on the implementation and use of Process Mining in companies has been substantially more restricted (Martin et al., 2021). That said, there seems to be a need to put into practice Process Mining techniques and verify the obtained results.

The second motivation to consider lies in the fact that Process Mining has an enormous potential for improving organizational processes. With the technological evolution and the use of process information as a tool for improving the way companies work, techniques such as Process Mining increasingly assume a key role in supporting firms to identify bottlenecks, opportunities for improvement and ways of reduction of inefficiencies and costs.

1.2 Key project objectives

Given the state of art on Process Mining, the growing adoption, and potential benefits of such technique, the project has three key objectives throughout its development.

Particularly due to the fact that the discipline is a relatively new domain under exploration, the first objective concerns, in a first phase, the better understanding of Process Mining. This encompasses the comprehension of the domain framework, application criteria, operating mode, current state of its use in the organizations, and benefits of the technique.

The second, and most extensive, objective covers the implementation of Process Mining on an Order to Cash process, with real data from a Deloitte's client. It entails the implementation

of the Process Mining tool, including the data interpretation and treatment, the model building, and the discovery of the process flow. Additionally, encompasses the analysis of the process and withdrawal of conclusions, aiming for the identification of inefficiencies among the process and potentially improvement suggestions, as well as the creation of dashboards in order to support the findings and serve as future control for the client.

Regarding the last objective, it concerns the application of a predictive model in order to complement the process analysis, described in the previous paragraph. The goal is to be able to predict the next activity that each client will perform in the process, and with this being capable of taking actions accordingly.

1.3 Process Mining at Deloitte Portugal

Deloitte Touche Tohmatsu Limited was founded in 1845 in London, and has, at the present day, presence in more than 150 countries and 345 000 professionals, providing audit & assurance, consulting, financial advisory, risk advisory and tax services. In Portugal, the firm member is Deloitte Central Services, S.A., and the project was developed in the subsidiary responsible for consulting services, Deloitte Business Consulting S.A., in the department of Enterprise Technology & Performance.

The service offer on Process Mining in Deloitte Portugal exists since 2021, so the company is in fact, at the time of this project, taking its first steps on the domain. Furthermore, the firm has not explored yet the full capabilities and opportunities of this new discipline.

1.4 Project methodology

The project theme was a new domain for which there was no personal knowledge of how to implement it in practice. With that being said, and after exploring what became the support of the theoretical framework component, a training program on a Process Mining tool was conducted. The training on Celonis included the Welcome Kit, for an introduction to the platform, the Create area, for the building of solutions, and some extra components to enrich the knowledge. Together, this learning phase took two weeks to complete.

Subsequently, the case for development of the tool was assigned, and the implementation began. After a better understanding of the data, the processing and integration of these on the platform occurred. This was followed by the necessary transformations and creation of the Data model. With everything lined up, the analysis of the process finally took turn, along with the creation of dashboards.

Lastly, a prediction model was applied toward the next activity in the process and the respective timestamp prediction. This application intends to complement the analysis conducted through the Process Mining tool.

1.5 Document structure

This thesis is structured according to the real timing of development of each part, resulting in a document with six main chapters.

This first chapter presents a brief summary of the thesis content, the major motivations and goals of the project, and the context in which it appears.

The second chapter provides the theoretical framework, presenting the academic concepts addressed in this project, as well as the applicability of the concept on organizations.

The third chapter presents the context in which the problem tackled here emerges, with the description of the process under analysis as well as the data used.

Chapter 4 concerns the project implementation, namely the procedures to implement the tool, the analysis and the predictive model.

The results and main outcomes from this implementation are presented and analyzed in Chapter 5.

Lastly, Chapter 6 presents a reflection on what was achieved and statements about further work that should be applied and developed on the sequence of this project.

Chapter 2

Theoretical Framing

The second chapter presents the state of art regarding Process Mining and its surroundings.

For a clear understanding on the project theme, it become relevant to start by exploring the current technological reality and business environment that led to the origin of Process Mining and its growing use tendency.

2.1 Technology Transformation

This section highlights some aspects of the technological evolution, that is, the Digital Transformation and the growing amount of available business data, which made Process Mining not just possible but also attractive for the companies to improve their processes.

2.1.1 Digital Transformation

The first article addressing the topic of Digital Transformation dates back to 1981. The number of published and cited articles has grown since then, with a growth of scientific publications since 2018 (Kraus et al., 2021). Digital transformation can be described as the digital technology changes that impact all aspects of the human existence (Stolterman and Fors, 2004), also concerning the acquisition and use of innovator and game changing technologies to boost productivity, gain value, and improve the social well-being (Ebert and Duarte, 2018). It is thus more than the digitization of existing processes or work products, it implies the revision of current processes with new perspectives coming from digital technology (Parviainen et al., 2017).

The performed research on Digital Transformation exhibits that the subject can be linked to two complementary fields. On one hand, it is correlated to technologies such as big data, artificial intelligence, and data analytics. On the other hand, the concept is similarly associated with digital strategy, supply chain management, leadership, value creation, and entrepreneurship. So technology can possibly be seen as a critical asset for a favorable position among the business competition via digital transformation (Kraus et al., 2021). Complementarily, the pandemic reinforced the importance of digital businesses. More digitalized companies have surpassed the others, thus this

revealed that having and executing a good digital business strategy is fundamental for nowadays firms success (Hunter and Smith, 2021).

Additionally, globalization is imposing digital processes and collaborative tools in order to carry out an efficient integration of businesses (Westerman et al., 2011). Organizations must constantly strive to innovate to better manage their processes complexity, to boost operational productivity and excellence, and this way create competitive advantage. The complexity of the firms, principally in those with a wide variety of processes, asks for innovative methods to increase transparency.

Process Mining appears as one of the key technologies driving the Digital Transformation, being used both in separate divisions for operational improvement and as part of strategic plans to operate digital transformation and rise organizational efficiency. Large companies such as Siemens, BMW, and Telekom, are using Process Mining as a standard tool for digital transformation and other improvement activities (Reinkemeyer, 2020).

2.1.2 Data Burst

The availability of data is crucial for the existence of Process Mining, and fortunately, as a result of technological evolution on digital data acquisition and storage, there is an exponential growth of databases (Hand et al., 2001).

To conceptualize the data growth into real numbers, some key findings from RedGate (2021) that reports the IDC's Global DataSphere Forecast, 2021-2025, are provided. The previous report predicts a global data creation and replication going for a compound annual growth rate (CAGR) of 23% over the 5 years period, meaning a hit on 181 zettabytes in 2025. To note that a zettabyte corresponds to 10^{21} bytes.

Furthermore, the predictions on storage utilization and on the amount of storage available per year, from IDC's Global StorageSphere Forecast regarding the same period, reflects a global storage capacity of around 16 zettabytes for 2025. This reflects a CAGR of 19.2% since 2019. Thus, a huge amount of the generated and replicated data will not be saved, a great part of it will be kept in the cloud, but it still represents a considerable challenge for the companies to deal with this dimension of data, and to take the adequate value from it (RedGate, 2021).

Big Data

As formerly exposed, Digital Transformation is strongly related to Big Data. As a response to the challenge of dealing with such amount of data, Big Data can be found.

Big Data could be simply described as a compilation of extensive and complex data sets. Nonetheless the first times being properly used, in the sense of referring not just to a phenomenon of high volume of data but also to discipline that compass computer science and statistics/econometrics, draw back to mid 1990s. The author was John Mashey, at Silicon Graphics Inc. (SGI) with a slide titled "Big Data and the Next Wave of InfraStress" and shows the consciousness of the Big Data phenomenon (Diebold, 2019; Ishwarappa and Anuradha, 2015).

The development of activities that generate large amounts of data such as e-commerce, and activities which potentiate the exchange of data gave origin to a new characterization of Big Data. The previous factors have fiercely increased data management challenges along three elements which are volume, velocity and variety, designated as the 3V's of data by Laney (2001).

The volume represents the amount of collected data and the challenge of processing it, the velocity refers the increasing speed at which data is generated and thereby processed, and variety relates to the obstacle of creating a relational database with data that is not always structured. Over time, with the growing use and importance of Big Data, there was a need to add two more V's to the data characterization: Veracity and later Value. Veracity emerges due to the fact that not all data is completely correct, and Value shows up since not the entirely amount of data is important or useful, a lot of data carry costs so the importance of making a good and smart use of the data is emphasized (Ishwarappa and Anuradha, 2015).

However, there is a lot of valuable information hidden in the data, that cannot be accessed through the standard data processing (Kourla et al., 2020). In this domain, Process Mining can respond to the challenge of exploiting event data in a meaningful way, and can be used to perform extensive analysis activities on numerous computing resources on the Big Data infrastructures (van der Aalst, 2016).

2.2 Linking process information with process efficiency

Addressed the topic of technological evolution that enabled Process Mining, it becomes essential to analyze the other causes that lead to its existence, namely Data Science and Process Science, both targeting improvement in the business environment (Provost and Fawcett, 2013; van der Aalst, 2016).

Process Mining is portrayed by van der Aalst (2011, 2016) as the bridge between Data Mining and Business Process Management (BPM) and, in broader terms, the filling gap between Data Science and Process Science. It emerged as an alternative for process improvement.

With that being said, this section consists of a brief clarification of the previous mentioned concepts, with the aim to facilitate the latter explanation of Process Mining.

2.2.1 Data Science

In order to properly take the so spoken value from the Digital Transformation and growing available data, Data Science takes place (Provost and Fawcett, 2013), and it could be in the form of Process Mining, once the last is considered to be part of the former (van der Aalst, 2016).

The field of Data Science has been suffering changes in its concept for more than 50 years. It could be flexibly described as a science of learning from data, including the study of methods to analyse and process the data, and advances technology for empirical methods improvement (Donoho, 2017). Data Science is also defined by van der Aalst (2016) as a multidisciplinary domain, including Process Mining, focused on transforming data, that can be structured or not, in

large amount or not, static or streaming, into real value. It also includes data extraction, preparation, exploration, transformation, storage and recuperation. The extracted value can be presented in the form of predictions, automated decisions, intelligent learning models or any type of data visualization delivering insights and considering business context. Data Science must be then prepared to report, diagnose, predict and recommend, through the analysis of data.

Such an embracing domain has to be dependent on several disciplines, which normally do not have clear boundaries between themselves. In fact, fields such as statistics, data mining and machine learning are part of it (van der Aalst, 2016). For this project, it becomes valuable to slightly explore more the domains of Data Mining and Machine Learning, since Process Mining can be situated between both. Process Mining complements these two disciplines with the process perspective (van der Aalst, 2016).

Data Mining

Data Mining can be portrayed as the analysis of large scale data with the goal of finding relationships and extract useful information from it. The input data is usually previously available, for another purpose, and is normally provided in a table format (Hand et al., 2001). According to the same author, a possible categorization for data mining tasks can be: Exploratory data analysis (EDA), Descriptive Modeling, Predictive Modeling, Discovering Patterns and Rules, and Retrieval by Content.

Both Data Mining and Process Mining are data-driven, however the former is not typically process-centric, neither cover the end-to-end process (van der Aalst, 2016).

A simple distinction between Data Mining and Process Mining could be explained with the fact that the former is based on available data and seeks to find patterns on it, while the last is focused on finding how the information is created and how it fits the overall process. On top of it, Process Mining distinguishes the patterns deviations to understand what is occurring in the processes, whereas Data Mining rejects those irregularities once it represents an impediment to the pattern definition (Moore, 2020).

Machine Learning

Machine Learning is the field that is involved in creating computer programs that learn from experience, and consequently improve (Mitchell, 1997). The discipline appears from Artificial Intelligence (van der Aalst, 2016), and it uses data to generate learning algorithms, via computational methods, which create predictive models aiming to improve the system performance. The training set is used for the learning phase, where machine learning algorithms use the data on this set and build models. The model is then tested in a never processed data set: the testing dataset (Zhou, 2021).

There are already applications of Process Mining using Machine Learning aiming to not needing a user to analyse the processes and gain knowledge on it for improvement recommendations.

This technique runs the analysis automatically, and gives intelligent insights and recommendations to improve the business processes (Veit et al., 2017).

2.2.2 Process Science

On the other side of the origin and utilities of Process Mining, Process Science can be found. Process Science was portrayed by van der Aalst (2016) as a vast field bringing together the insights from information technology and the expertise in management sciences in order to improve and carry operational processes. Areas such as BPM, or Business Process Improvement can be considered as part of Process Science. And Process Mining, with its extraction of process-related insights from event logs, complements the existing approaches (van der Aalst, 2011).

A business process can simply be described as a set of actions that organizations do when delivering a product or service to their clients (Dumas et al., 2018). Processes are now extremely intertwined with digital technologies such as the Internet of Things, Machine Learning, and distributed ledger technology combined with data analytics. Business processes also lean on information systems and may be cross-organizational. As a result, process modeling has turned to be very important, and there is a need to integrate several disciplines to deal with processes, so, Process Science can be used to develop, implement, and manage them (Brocke, 2020; van der Aalst, 2013).

Business Process Management

BPM encompasses several techniques for the companies to hit their goals by supporting the design, execution, documentation, analysis, monitor and control business processes. This way, BPM manages complete chains of actions and decisions around the firm businesses, and is widely applied for process modeling (Weske, 2019; Reinkemeyer, 2020).

The methods applied by the majority of the disciplines in Process Science lack in capabilities to learn from the data, since, by being process-centric approaches, the focus is on modeling. However, with the growth of event data, those areas are being challenged since there is no useful use for model-based analysis and implementations without the proper study of the valuable information buried in information systems (van der Aalst, 2011). Thus, it seems reasonable to associate event data with process models, so the real processes can be identified, and existing process models can be analyzed and improved, which is exactly what Process Mining does. Whereas BPM permit to model To-Be processes, more concretely, the way the processes should be performed, Process Mining gives the visualization of the real working mode, the As-Is processes (Reinkemeyer, 2020).

2.3 Process Mining

Having already been given the conditions in which Process Mining appears, and its framework, this section serves to effectively explain in what Process Mining consists of, and demonstrate its applicability in the business context.

2.3.1 Concept

The roots of Process Mining date back to 1958 with the work of A. Nerode (1958) as cited by van der Aalst (2011), and the notion of Process Mining techniques in the environment of workflow management appeared in Agrawal et al. (1998). Nonetheless, the father of Process Mining is considered to be Wil Van der Aalst, by the end of the last millennium (Reinkemeyer, 2020).

Process Mining can be described as a discipline for process improvement, that uses event data to objectively and quickly discover and analyse business processes (van der Aalst, 2011). It joins process model-driven methods with data mining (van der Aalst, 2016) and provides holistic insights on the processes and their complexities with full transparency (Reinkemeyer, 2020). This takes place through the analyses of business data using event logs, and automatically visualizing the actual flow of business (Kang Yeong-sik et al., 2015 in Park and Kang (2016) ; van der Aalst, 2011). Important to state that an event log contains specifications of the executed operations that maintain the information systems, that can be provided by several different systems, and it has, at least, enough information to describe and order each activity (Suriadi et al., 2017). Thus, an event log contains, as a minimum, a Case ID for numeric identification, an Activity for specification of the operation, and a Timestamp recording the time of each activity (Reinkemeyer, 2020).

Nevertheless, the information of the event log is used not just to find and track down processes, but also to verify compliance, detect deviations, predict delays, examine bottlenecks, compare process variations, support decision making and propose improvements. Process Mining can be this way categorized within three techniques: Process Discovery, Process Conformance, and Process Enhancement (Kang Yeong-sik et al., 2013 in Park and Kang (2016); van der Aalst, 2011, 2016).

- **Process Discovery** builds a process model based on an event log information (van der Aalst, 2011), that can latter be used as input for improvement suggestions (Partington et al., 2015).
- **Process Conformance** performs a comparison between the activities and corresponding flow of the event log with the process model. This model may have been discovered with the use of Process Mining or constructed other way, in order to detect possible variability and irregularity. The outcome can provide valuable awareness for auditing and compliance situations. Besides, it can be useful to inspect the discovering phase and align the incongruent models with the real processes (van der Aalst, 2011).
- **Process Enhancement** uses the information of the event log and tries to suggest extensions or improvements on the existing process model (van der Aalst, 2011). New user-driver techniques for this Process Mining stage, add Machine Learning and Artificial Intelligence to evolve the process uncovering to an intelligent and automated process analysis. Besides unmasking problems and inefficiencies, it explores root causes and delivers advises to solve them (Veit et al., 2017).

This technology can significantly reduce the time and cost required to understand the current processes, as well as sustain evidence-based recommendations by being a data-driven method

(Kang Yeong-sik et al., 2015 in Park and Kang (2016) ; van der Aalst, 2011), and supporting companies to digital transformation (Zerbino et al., 2021).

Results from a survey from Kerremans et al. (2021), conducted by Gartner, and which was directed to process mining vendors, shows that the current implementation focus, even though decreasing, is still the process Discovery. Nevertheless, Enhancement seems to display a growing tendency, with forecast to get higher than Discovery in 2022. Figure 2.1 presents the combined data from Gartner's preceding surveys, since 2017 to 2020, with the values for 2022 as predicted numbers.

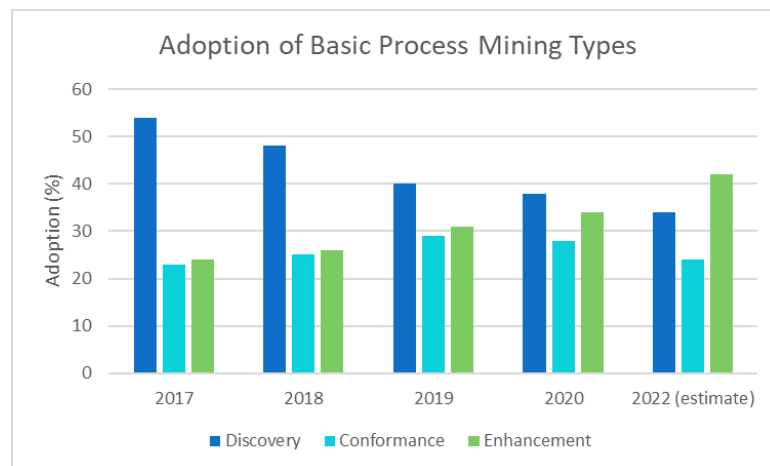


Figure 2.1: Adoption of basic Process Mining types (adapted from Gartner)

Furthermore, researches on state-of-art of Process Mining applications, support Discovery as the most explored Process Mining type, followed by Enhancement, and finally Conformance (Garcia et al., 2019; Zerbino et al., 2021).

2.3.2 The emergence of Process Mining

As a result of the current state of the technological transformation and process modeling, and management, described so far, it seems reasonable to consolidate the origins and uses of Process Mining.

In order to manage important operational processes, Process Mining is now considered as a central tool for modern organizations. The present is a result from the fact that event data has grown remarkably, furthermore processes and information must be perfectly aligned to meet requirements related to conformity, efficiency and customer service.

The barrier between digital and physical is nowadays faded, events are logged once they occur and processes are directed and controlled based on the event data. As already mentioned, Process Mining is thus considered to be the link between Data Science and Process Science. On one hand Data Science methods tend to not understand the actual underlying process, areas like Data Mining, statistics and Machine Learning usually do not consider end-to-end process models,

and Business Intelligence tools focus on simple dashboards and reports, also neglecting the business process insights. On the other hand Process Science approaches are more model-driven and process-centric but do not consider the information provided, in a hidden way, by the data, so they do not learn with the event data (van der Aalst, 2011, 2016).

2.4 Process Mining in organizations

As this project covers the implementation of Process Mining on top of a real collection process, which is part of the O2C (Order to Cash) process, it is pertinent to explore the state-of-art regarding Process Mining in Business Management, as well as the challenges felt in the companies while trying to improve their processes, and the current state of the technique in the firms.

2.4.1 Literature review on Process Mining in Business Management

Zerbino et al. (2021) present a research on literature review on Process Mining in Business Management. The starting point was given with 3600 papers on business management-orientated Process Mining literature, of those 145 manuscripts were selected. Aspects such as the type of contribution of those papers (empirical or conceptual), type of process mining algorithm, type of process mining, process mining perspective, sector, business function, and managerial focus were explored.

For this topic, it became relevant to understand in which type of companies (Sector) and type of processes (Managerial focus) it is possible to find Process Mining. Regarding the sector, the industry sectors found and clustered were: Energy and Materials, Industrials, Consumer Goods, Consumer Services, Healthcare, Utilities, Financial, and Technology. Respecting the Managerial focus, three hierarchical levels were defined: strategic, tactical, and operational. The conclusions for these two aspects were reunited considering only the empirical Process Mining contributions.

The results show that:

- Healthcare is the most seen sector on those papers, with Process Mining implemented mainly in processes that support operational functions. As regards the other sectors, the papers are more uniformly distributed over the business functions. Subsequent to Healthcare, comes the Financial, Utilities, and then the Technology sector.
- The most traced managerial focus is, by far, the operational processes, and it was not found any empirical papers with a purely strategic address.

Results from another mapping study of the state-of-art regarding Process Mining applications, from Garcia et al. (2019), out of existing journals and conferences publications, are presented for comparison. This research encompassed the analysis of 3713 manuscripts from 2002 to 2018, which with further exclusion criteria, become 1278. It was possible to conclude that:

- There is a notable increase of publications over the years.

- When grouping the papers by application area, 584 manuscripts were classified into 21 categories, leading to Healthcare proving to have the highest appearance. The following order, by occurrence, is Technology, Manufacturing and Industrial, Education, and latter Financial, with the remaining sectors representing only around 25% of the total.

Both studies declare Healthcare as the most covered sector, and despite the mismatch on the successive sectors, it is presented a clear similar disposal.

2.4.2 Opportunities for Process Mining

As a result of the current business environment reality, Process Mining has a lot of potential spots to act on. Herein is presented some of the results from a survey deducted by Forrester (2022), questioning 818 process owners and business execution decision-makers across industries world-wide, in order to better understand the current state of process improvement and data execution within organizations. Some key and valuable findings, in the context of this project, representing opportunities for Process Mining can be found next.

- **Organizations are going through difficulties while dealing with process complexity and visibility.** On one hand, with digital transformation, the growing activities and processes, along with their inter-connectivity, 71% of the respondents declare that they use at least 10 business applications for their primary processes. On the other hand, 72% are still using manual controls that do not facilitate process visibility.
- **Firms are lacking real-time visibility and immediate tools.** Slightly more than half of the daily actions associated with executing organizational processes are driven by data insights in more than a day old, and only 7% of the actions data-driven are dealt in real-time.
- **Organizations are having difficulties to incorporate all the department's processes systems.** Only 56% of the respondents state that they are capable of such integration and visualize the end-to-end processes. Without this proper view, companies do not have a clear understanding on how to improve the processes and what processes should be automated, which is making 44% of them to report higher organizational costs due to process-related issues such as incorrectly application of automation.
- **Decision-makers intend to implement process improvements technologies.** By the next year, 57% of organizations are planning to spend more on process improvement technology being that for companies already with highly interconnected and interdependent processes, the value rises to 70%. Increase productivity, reduce costs and improve customer satisfaction are the top three goals.

Moreover, in order to identify what problems are the entities trying to mitigate when applying Process Mining, results from the Global Process Mining Survey by Deloitte (Galic and Wolf, 2021) are presented. The study about Process Mining applications, combines answers from 106 IT and

business executives from 24 countries, that may or may not currently use the referred technology. The benefits that the respondents aim to obtain with Process Mining are listed below.

- Process Improvement is expected by 77% of the surveyees from the technology application, followed by Process Transparency and Cost Reduction, representing 57% and 46%, respectively, of the answers.
- The remaining top five of expected benefits encompasses Digital Transformation acceleration and Automation identification. Figure 2.2 displays all the listed expectations.



Figure 2.2: Expectations on Process Mining (adapted from Deloitte)

Also, Kerremans et al. (2021) highlight five main drivers for the adoption of Process Mining, that also represent opportunities for the method's employment:

- **Digital Transformation:** enables Process Mining and increases the consciousness on the gains of deeply understand the business processes.
- **Artificial Intelligence:** algorithms acting on immense quantity of data in order to find patterns and insights create new opportunities.
- **Task Automation:** Process Mining can be a complement of Robotic Process Automation (RPA) with improved insights on the processes and help in the task automation implementation.
- **Hyperautomation:** or "AI-based process automation"; can be described as the harmonised automation of the entire value chain. Process Mining enables the understanding of the processes before automation.

- **Operational resilience:** the capacity to change the current operations to meet the variable business scenario needs. Process Mining enables factual modeling of work, and the understanding of those models for everyone. Moreover, the technique provides insights for the improvement of the processes.

To conclude this section regarding opportunities for Process Mining, statements taken by EY (2019) corroborate and complement the adoption drivers previously indicated with:

- **Automation** reinforcement, mentioning the critical previous knowledge on the processes before RPA.
- **ERP consolidation** once Process Mining is able to compare each ERP (Enterprise Resource Planning) information and resolve errors before consolidation.
- **Operational efficiency** reinforcement, with the possibility of process transparency and identification of potential improvements.

2.4.3 Current state of Process Mining in organizations

For a more comprehensive framing on the use of Process Mining in the organizations of today, some conclusions by Forrester (2022) are bellow provided.

- 27% of the decision-makers use business process management or process mapping tools to manage their organization's processes. The correspondent with process mining are only 13%.
- In relation to the already Process Mining users, 90% of them are convinced about reaching their yearly process improvements goals.
- 61% of decision-makers will use, or are evaluating, Process Mining in the next 12 months, considering it the number one technology they plan to use to measure or improve business processes.

Furthermore, Kerremans et al. (2021), report data that state that Business Process Improvement represents the majority of use cases in process mining applications. However, when looking at Figure 2.3, it is detectable that Process Automation and Digital Transformation are increasing trends, corroborating the tendencies of Technology Transformation already mentioned.

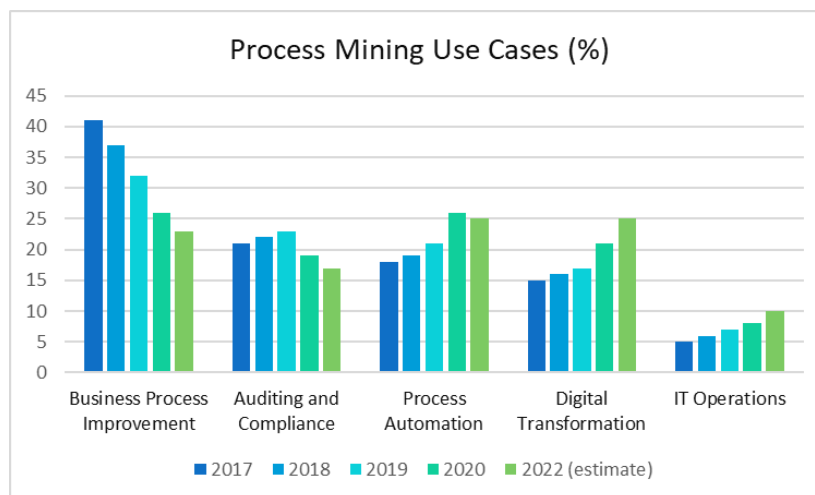


Figure 2.3: Process Mining use cases (adapted from Gartner)

The previous information can be complemented with other conclusions from the survey of Galic and Wolf (2021), which are presented below:

- 63% of the IT and business executives respondents are already working in Process Mining projects, and 22% are planning to use it in the future.
- 12% do not intend to adopt the technology, due to reasons such as Process Mining not being a priority, not having budget or management attention, not knowing what Process Mining is, or other.

Moreover, the areas of interest when using Process Mining, available on Figure 2.4, exhibit that the Activities of Purchase/Procurement, and subsequently, Accounting/Payment are on the uppermost of the activities list (Galic and Wolf, 2021).

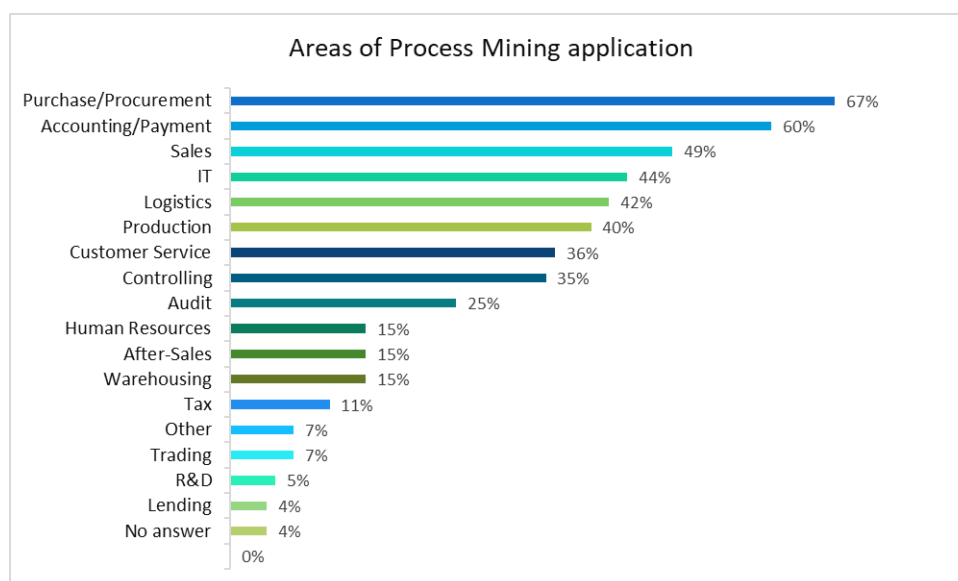


Figure 2.4: Areas of Process Mining application.

At last, 84% of the respondents affirm that Process Mining is delivering value to their businesses (Galic and Wolf, 2021).

2.4.4 Core processes in Process Mining

Process Mining can be applied to all areas inside the companies, from administration processes to production, passing by development actions, production control, financial activities, and sales. For that, the IT implementation must be "end to end" and cross-functional (Reinkemeyer, 2020).

The O2C process covers the complete customer journey, since the moment of the purchase, with the creation of an order, until the order invoice or another record settling the open balance (Deb, 2019; van der Aalst & Carmona, 2022). This process represents, according to the Figure 2.4, a predominant process in which Process Mining operates.

The steps of the O2C process are relatively easy to collect from the usual storage of event logs in ERP systems. The simple visualization of the activities flow, such as order entry, order check, delivery note, invoice, and cash collection, allows useful process insights, namely the identification of inefficiencies, and potential automation actions (Reinkemeyer, 2020).

KPIs Process Mining addresses process improvement, hence there is a need to investigate and quantify the changes supported by the discipline. Given that, performance measurement assists organizations to measure and to examine the effectiveness and efficiency of their processes (Meier et al., 2013b). Additionally, performance indicators support the service delivery management through the assessment and comparison of the previous state, or intended state, with the obtained outcomes (Meier et al., 2013a). Key performance indicators (KPIs) measure the companies' performance on their critical processes, providing evaluation and benchmarking on that (Badawy et al., 2016; Meier, Lagemann, et al., 2013).

2.5 Predictive Business Process Monitoring

In addition to what was possible to build for the Process Mining implementation, this project also covers the application of a predictive model for the business process. A Long Short-Term Memory (LSTM) model was applied to predict the next activity of the process, and respective timestamp. Thus, a brief explanation on the topic is here provided.

Recently, advanced deep learning techniques have begun to be used in the realm of predictive business process monitoring (Camargo et al., 2019). The improved business process visibility has facilitated the opportunity of predictive business process monitoring, which intends to predict the progression of running cases of a business process. The prediction is made through models using past event logs, and aims to use these prediction to anticipate problems and proactively manage, and mitigate them (Jalonen and Lönnqvist, 2009; Tax et al., 2017).

2.5.1 Recurrent Neural Networks

A neural network, which is a subset of machine learning, is a complex system made up of layers of interconnected cells. The network typically includes an input layer, one or more hidden layers, and an output layer. The cells within each layer are linked to the cells in the other layers through a series of weighted connections, which allows a wide range of network architectures. The output of each cell is determined by the weighted sum of its inputs (Evermann et al., 2017).

A neural network is a supervised learning technique. The key feature of supervised learning is the presence of labeled training data. The term "supervised" implies that there is a "supervisor" who guides the learning system by providing the correct labels for the training examples. This allows the system to learn how to correctly associate inputs with their corresponding outputs (Cunningham et al., 2008).

In a supervised learning setting, the network's output is compared to a target output using a loss function. To improve the network's performance, the gradients of the network's parameters are computed with respect to the loss function using backpropagation, and this information is then used to adjust the parameters using variations of gradient descent algorithms, with the goal of minimizing the loss function. The choice of the output layer cell and loss function are often made based on their computational properties with respect to backpropagation (Evermann et al., 2017).

During the 1990s, recurrent neural networks (RNNs) were a topic of significant research and development. These networks are specifically used to recognize patterns that change over time or occur in a sequence. A RNN is a type of neural network that includes feedback connections, creating a closed loop system. This allows the network to retain information from previous inputs and use it to inform its current processing (Medsker and Jain, 1999). One of the ways to train RNNs is by feeding it with real data sequences, one step at a time, and then training it to predict the next element in the sequence. This approach can be used for sequence generation tasks (Graves, 2013).

RNNs can be represented in an unrolled (or unfolded) way, into a full network, as shown in Figure 2.5. Each component is known as a time step, where x_t is the input at that particular time step, and s_t is the hidden state containing information extracted from all previous time steps. The hidden state is updated with the new input x_t after each time step: $s_t = Ux_t + Ws_{t-1}$, where U and W are weight vectors for the new inputs and the hidden state respectively. Function f , known as the activation function, is commonly the hyperbolic tangent or the logistic function, also known as the sigmoid function: $\text{sigmoid}(x) = \frac{1}{(1+\exp(-x))}$ (Tax et al., 2017).

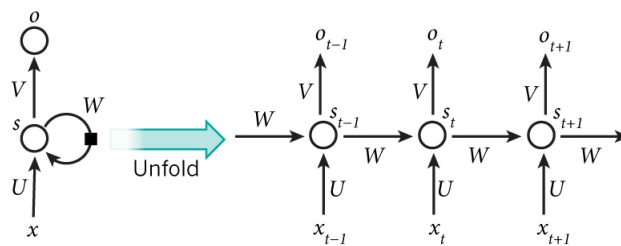


Figure 2.5: Recurrent Neural Network. Source: LeCun et al. (2015)

2.5.2 Long Short-Term Memory for Sequence Modeling

A Long Short-Term Memory (LSTM) is a variation of RNN architecture which was developed to more effectively handle the modeling of sequences and their dependencies over longer periods of time, as compared to traditional RNNs (Sak et al., 2014 ; Graves, 2013).

In a RNN, the value of the state s_t is the outcome of a function that combines the weighted average of s_{t-1} and x_t . However, in an LSTM, the state C_t is managed through controlling gates, o_t , i_t , and f_t , to access, write and clear the memory cell, respectively. When the input gate is activated, new information will be added to the memory cell. Additionally, the previous memory cell status C_{t-1} can be discarded if the forget gate f_t is activated. The information stored in C_t will be passed to the output h_t based on the activation of the output gate o_t .

LSTM neural networks are well suited for Predictive Business Process Monitoring because of their ability to process sequential data. LSTMs are a type of RNN which are able to maintain internal memory states, allowing them to understand context and dependencies in the data over time. This makes them particularly useful for tasks such as predicting the next step in a business process, where understanding the history of the process is important. Additionally, LSTMs are able to handle variable-length input sequences, which is important in business process monitoring where the length of the process can vary. LSTMs can also be trained on large datasets, which is often the case in business process monitoring where a large amount of data is generated. LSTM networks are widely used in various fields like natural language processing and time-series forecasting (Camargo et al., 2019; Tax et al., 2017).

2.5.3 Loss Functions

A loss function is a metric that quantifies the difference between the output data predicted by a model and the actual data in a given dataset, and which primary role is to direct the model towards convergence during the training phase. With the goal of optimizing the model during training, by minimizing the loss value, the model can fit the training data more accurately and achieve minimal test error. This ensures that the model is capable of accurately classifying new samples (Chen et al., 2012).

The following loss functions are relevant for the predictive model implementation on this project.

Mean Absolute Error

The intervals between events can vary significantly, spanning across different orders of magnitude. The mean absolute error (MAE) metric can be used as loss function, but also as an evaluation metric. If compared to another popular metric, the root mean squared error (RMSE), MAE is less sensitive to substantial errors on outlier data points Tax et al. (2017). MAE can be represented by the Equation 2.1 where x_i means the prediction rating; y_i means the true rating in testing data set; N is the number of rating prediction pairs between the testing data and prediction result (Ragab et al., 2020).

$$MAE = \frac{1}{n} \sum_{i=1}^n |x_i - y_i| \quad (2.1)$$

Cross Entropy

The cross-entropy loss function is widely used in deep neural network models, particularly for (multi-class) classification tasks (Gordon-Rodriguez et al., 2020). The predicted probabilities for each class are compared with the actual desired output of 0 or 1. Based on how far the predicted probabilities are from the actual values, a logarithmic penalty is imposed. This penalty yields a large loss score for large differences close to 1 and a small loss score for small differences close to 0.

During training, the cross-entropy loss is used to adjust the model weights. The goal is to minimize the loss, as a lower loss corresponds to a better model fit (Koech, 2020).

The Cross-entropy loss function can be represented by the Equation 2.2, with t_i as the truth label and p_i as the softmax probability for the i^{th} class.

$$Cross-entropy\ loss = - \sum_{i=1}^n t_i \log(p_i) \quad (2.2)$$

The softmax function is often used as the final activation function in neural networks. It is used to normalize the output of the network to a probability distribution over the predicted output classes, which is proportional to the exponential values of the input numbers. This ensures that the resulting values sum up to 1, thus creating a valid probability distribution over the predicted output classes (Martins and Astudillo, 2016).

Chapter 3

Problem scope and description

This chapter has the purpose of providing a logical comprehension on the addressed problem at this project. Firstly, the business context and a wide positioning of the company in which the challenge emerges is briefly presented, in order to better understand the relevance and scope of the process under analysis. Afterwards, a description of the process is provided, and finally, a description of the data that supports the process under analysis is presented.

3.1 Business context

The problem scope targeted by this project broadly contemplates the absence of knowledge and deep understanding of the processes of a company. The objective is to provide these insights, and to find inefficiencies in the processes, and in the best scenario, promote improvement actions.

Therefore, and as shortly presented in the introduction, this project aims to explore an Order to Cash process (O2C), of a mass market B2C (Business to Consumer) service provider. The focus here is on the events happening between the invoicing and payment, as the company has several notice mechanisms steps in order to properly execute the collection.

The company has already modeled the indicated process, according to what they believe to be the actual order of events. Nevertheless, the goal of the analysis covers all the Process Mining techniques: is focused on discovering the real process flow, that is, Process Discovery, on verifying the compliance of the events flow, that is, Process Conformance, and finally find inefficiencies and striving for a potential optimization, meaning Process Enhancement. The previous presented phases include objectives that are aligned with goals for the project:

- Discover the real process flow;
- Verify if the process is taking place as expected and envisioned by the company, but also to compare the real process with the ideal flow of activities;
- Identify bottlenecks or inefficiencies on the process;
- Identify patterns both in the type of customer and service, as well as in the form of invoicing and payment, which may be causing problems;

- Create control dashboards for the company;
- Suggest ways of improving the process.

3.2 Operations overview

As previously stated, the O2C process covers the complete customer journey, however, the present study only involves the process since the invoice issuance extending until the effective payment, or in a scenario without payment, a rigid measure to end the process, the submission for litigation.

The operating scheme for the process, modeled by the company itself, is divided in two phases, the Billing phase, represented in green in Figure 3.1, and the Payment phase, in blue.

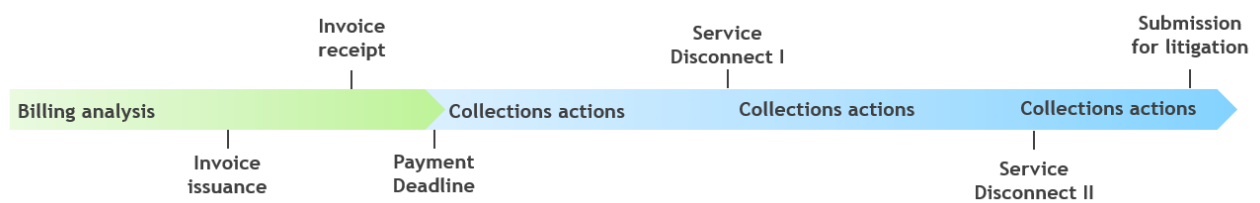


Figure 3.1: The two major phases of the process

- **Billing:**

- The invoice issuance is what triggers the process. It may be electronic or not, and has a payment deadline date (PD).
- Two days before the due date, the client receives a SMS (Short Message Service), or an IVR (Interactive Voice Response), with a warning on the payment deadline date.
- With the customer's proper payment, follows the Invoice receipt. The payment must be made until the due date, if not, the second phase of the process begins.
- This phase also deals with undelivered invoices, claimed invoices, and credits issued to customers. These activities do not enter on the main process, displayed on Figure 3.1, as they represent a deviation to the fundamental flow of activities.

- **Payments:**

- Passed the payment deadline date without the complete payment, a sequence of warning messages and sequential service cutting actions start.
- Whenever a client enters this phase, two days after the due date, an extra charge (€) is activated for the next invoice as a penalty.
- The warning messages are sent through different methods, namely SMS, email, letter, and IVR.

- Outbound actions also take place at this phase.
- The submission to litigation represents the last possible action.

Figure 3.2 displays the second phase in detail. To note that the first point of the timeline, prior to the payment deadline, are part of the Billing phase, nonetheless, it was placed here in order to make the sequence of actions more explicit.

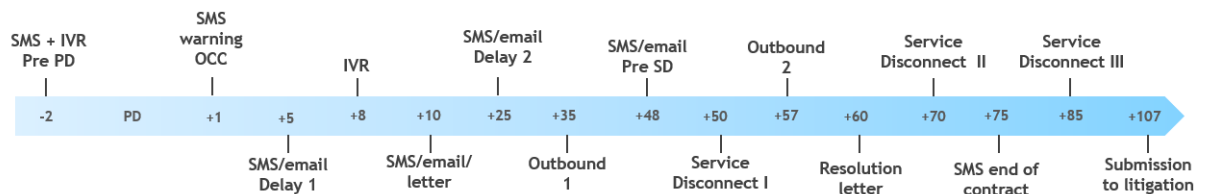


Figure 3.2: Payment phase

The numbers on the arrow, linked to each action, represent the number of days before or after the Payment Deadline date (PD).

3.3 Data Characterization

This final section has the intention to deliver an outline of the available data.

For the project development, the company has provided one month of records regarding issued invoices. This data entails information, on its core, about the invoicing and payment, which are complemented with information such as insights on clients and special lists of clients, and payment agreements.

Data Description - Tables

The provided data allowed to successfully interpret the process flow, as well as to get insights on the process. With this dataset, it was possible to demonstrate the process flow through the records and track of the issued invoices. Over all the available data, each invoice has a unique identifier, and is associated to a billing and client account. More information on the available data is below provided:

- **Issued Invoices:** This table can be considered as the most important one. It provides basic identification information about the invoice. It entails fields such as the dates of release and payment, concise description of the addressee, and provided service. It also demonstrates if the invoice is already paid (even if partial), the payment method, current debt action stage if already in debt, among others.
- **Undelivered Invoices:** This table saves the date from the undelivered invoices notifications. The invoice identifier and the notification date allows the extension of the process flow of actions.

- **Claimed Invoices:** Similar to the previous, this table ensures the invoice tracking whenever the invoice has a claim. It provides information on the creation and closure date of the claim, the motive, the amount of open and close claims for the invoice, among others.
- **Subjects:** This table reports information on the subjects raised by each client, meaning the topic that led the client to contact the company. It is the only table that does not have an invoice identifier, since each open subject is associated only to the client account.
- **Steps:** When the client does not pay the invoice until the predefined deadline, a flow of actions is triggered. Each invoice supposedly passes through the stages displayed in Figure 3.2, which are being recorded and displayed at this table. Hence, information regarding the type of process flow (which is related to the client category), dates of entering the stage, and debt amount on that stage are some of the information provided by this table. An example of step could be given as "Email or SMS or Warning Letter", or even "No Action".
- **Actions:** Identical to the previous, this table is also about the process after the due date passing by without the proper payment. It entails a more detailed characterization of the actions taken in order to collect the debt. The need to use this table has emerged from the fact that, without this kind of information, it is not possible to understand the real process, or to investigate the results from each action taken by the company to collect the debt. The actions from this table provide a more complete and useful distinction between the actual stages of the process.

A brief summary of the information on each table, as well as a few field examples, are provided in Appendix A, in order to provide a more visual and clear understanding of the content of the used data.

Chapter 4

Project implementation

The purpose of this chapter is to describe the key steps of the implementation and development of the Process Mining application. These key steps are in alignment with the objectives for the project, and are presented here in the order in which they were performed.

With that being said, the beginning of the journey was about understanding and dealing with the provided data, followed by its import and treatment in the selected Process Mining tool, Celonis process mining, and the construction of the Data Model. Afterwards, the dashboards were created alongside with the analysis of the process. The statements regarding the implementation steps were mainly supported by the Celonis Academy Training.

To complete the project implementation, a predictive model regarding process monitoring was implemented.

Throughout this document, figures regarding the project development are shown, however, the information on it is not always in english, since the data was provided in portuguese.

4.1 Data understanding and treatment

In order to structure a coherent analysis on Process Mining, as with any other type of data analysis, there was the need of a clear understanding of the data. As demonstrated in the previous chapter, a considerable amount of information about the process steps was provided, and only with further analysis was possible to better comprehend the meaning of it. Therefore, an analysis assisted by Power BI was carried out with the main objective of achieving a correct interpretation of each component of the provided data.

One of the most important findings regarding the exploration done, was the inference that the Steps table may not be complete. The inference was raised by the fact that the actions, from the Actions table, which description fits the step classification, are present in a higher number. Besides, the former has around half of the entries of the latter, and several steps are filled with "No Action" marks. The mentioned finding has led to the conclusion that the process stages would be more accurately defined by the actions rather than the steps. Nevertheless, the Steps table was

kept during the implementation, once the Issued Invoices has fields relating to the steps and may represent additional information for future analysis.

As formerly explained in subsection 2.2.1, Process Mining distinguishes the patterns deviations to understand how the processes are taking place. As a matter of fact, Celonis comprises several types of filters, but also routines and algorithms that perform the preprocessing of the event log. This platform makes use of Machine Learning to discover the source of divergences on the process, or matters due to inconsistent events or missing values (Marin-Castro and Tello-Leal, 2021). With the previous being stated, a great part of the preprocessing was done through the data integration and exploration phase.

Nonetheless, posterior to the analysis at Power BI, it was considered that for a more complete analysis some empty values should remain unchanged, as long as they did not represent a necessary field for the modeling.

Additionally, certain columns with dates in wrong format were also rectified using Excel. Also supported by this tool, it was created an association file, between the Steps and Actions of the respective two provided tables. Each step could be represented by several actions, and there are actions that do not have a representative step. The file supporting the connection had the intention to fill some logic and process gaps that could appear.

The remaining data treatment was processed at the Data integration phase, more precisely in the Transformations stage.

4.2 Data integration on Process Mining tool

The first step of the project implementation itself was the collection of the data, mentioned in the previous chapter. The typical Collection phase on Celonis comprehends the Data Integration process, which in turn includes the Connection to the transactional systems, the Extraction, the Transformations, and lastly and as final goal, the Set up and Load of the Data Model. These stages are described in the following sections.

4.2.1 Connection and Extraction

The data was provided in offline data sets, regardless of the transactional systems used by the company. Moreover, this project intends to analyze past events and to build a Process Mining application, so the connection phase for this project was conducted as a one-time upload of the files containing the information tables, also mentioned in Chapter 3.

4.2.2 Transformations

The Transformation phase happens after extracting all the necessary raw information. It encompasses the construction of necessary tables such as the Activity Table, with the relevant process activities, as well as other tables connecting important complementary information. The necessary

rework to the imported data also takes place at this stage, since it will impact the activities that are considered, and consequently the process flow.

All the transformations were performed using Vertica SQL, and made via Data Jobs, once the real time connection was not possible and there was the need to perform a Data Model Load.

- **Processing:** At this stage some data processing was done. Furthermore, it was found one duplicate invoice on the issued invoices table, which was removed along with some columns that did not represent crucial or interesting field for the analysis. Some other specific modifications had to be made in order to achieve a clear and real representation of the process.
- **Actions and Steps Tables:** The Action table had 129 different actions taking place, but some of these activities discrimination were due to the type of receiver of the action, or due to specifications regarding the stage on which they occur. By having the actions timing's, the majority of this kind of distinction becomes unnecessary. Thus, the number of actions was reduced to 66.

Similarly to the activities from Action table, the Steps table also had excessive categorization. The 29 steps were reduced to 22.

- **Activity Table:** One of the most important pieces that make Process Mining possible is the Activity Table or Event Log. The table was created with the necessary columns to downstream the process, already mentioned at Chapter 2. Thus, the table contains the Case Key (invoice ID), Activity, Event Time, plus the Sorting column, in order to distinguish activities that, regarding taking place in the same day, should have a predefined order for future proceedings. As an example, for cases in which the Payment has taken place in the Payment Deadline day, the payment should appear first in order to be considered as a timely payment.

The activities that were added to this table were taken from all the provided data tables except the Steps table, once the Actions table would provide further distinction. The previous operation was done always taking into account that the activities should only be considered for the invoices present in the Issued Invoices table, and when the time in which it occurred is available. On top of it, in order to consider each the actions from the Actions table, it was assumed that the activity only took place when the field regarding the state of action was "Processed/Discarded". Besides, when the actions required manual services, other specific field would have to be marked as "Processing".

Beyond the actual activities, the Issuing and Payment Deadline dates were also included as activities for a clear process flow and deeper analysis. With this, all the process flows happening at the month of the provided data reflect almost 6.7 million of activities, representing 62 distinct activities.

- **Case Table:** Another crucial table for the analysis is the Case Table which contains more information regarding each case. In this circumstance, the Issued Invoices Table, from the provided data, fulfills the requirements.
- **Invoice-Subject Table:** As mentioned in the previous chapter, every topic from the Subject table is associated to a client account, but there is no invoice linked to it. Resulting from an analysis to those same topics, it was found several levels of description of it. The major and first level of those descriptions is concerned with Invoicing or Invoice. Given this, and to all appearances, each subject should be linked to an invoice.

Consequently, in order to perform the prior connection, an assumption was made. Every subject that was raised by a client, after the issuing data of an invoice of that same client, and, at the limit of 30 days after the issuing date, was considered to be the subject raised for that specific invoice. One potential concern with this approach could be that a topic can be incorrectly associated to more than one invoice. Nevertheless, the analysis only concerns one month of invoices, and the typical situation, covering approximately 90% of the cases, is when a client receives just one invoice per month. With that being said, it was considered to be more aligned with the reality to perform the described method rather than to associate the subject to just one of the invoices, potentially the incorrect one, and leaving the invoice that actually triggered the subject with no connection to it.

4.2.3 Data Model

Once all the data needed was uploaded and ready for use, the next and last step of this first stage of Collection is to build the Data Model. The Data Model is composed by a collection of tables and the relationships between those tables. It is at this stage that occurs the definition of the data that can be accessed by the analyses built on top of it. As already mentioned, most of the assumptions and actions taken for the proper definition of the model were made throughout a process with recurrent cycles between the data integration phase and the data model design and upload. This was due to the fact that only with the process model under analysis it was possible to identify some points that required some type of treatment.

The Data Model for this project is composed of nine tables, from which seven are directly linked to the Case Table. These tables are connected via N:1 relationship to the Case table, meaning that each invoice can have several claims, actions, notifications for non-receipt, or topics raised. In the same way, the Activity Table is linked to the Case Table through a N:1 relationship, once each invoice has several activities throughout the process. The remaining tables provide complementary information to other tables. The constructed Data Model is displayed in Figure 4.1.

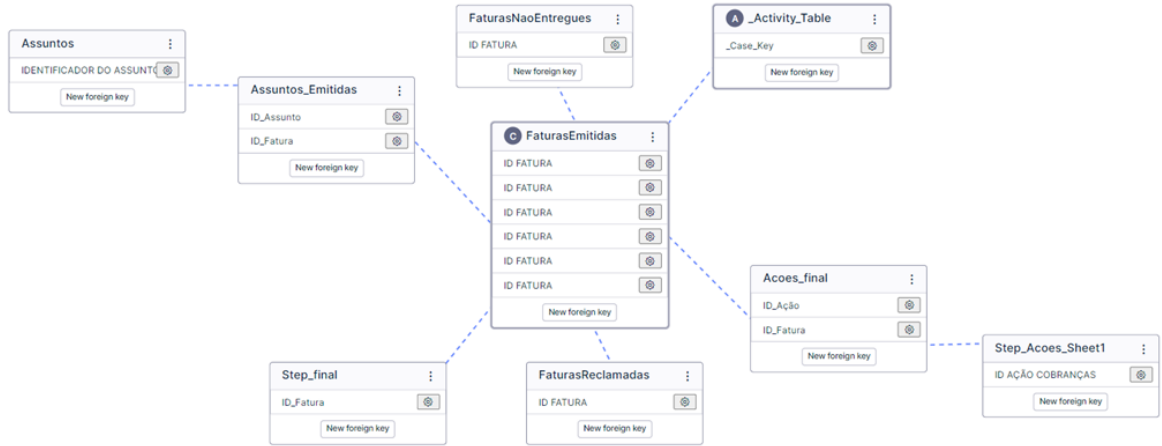


Figure 4.1: Data Model

4.3 Analysis

After the data model has been successfully loaded into Celonis, the data becomes accessible in the Process Mining platform front end. These analyses were built through an own programming language entitled Process Query Language (PQL), which is very similar to SQL, but centered on the process language rules.

The company has defined that the analysis focus should be on a specified set of client types, and also on active client accounts. Thus, a filter for each one of the conditions was kept during the entire analysis, and consisted of attribute selections on the Case Table.

Part of the analysis has resulted from the Celonis platform itself, through specific features, since it provides the identification of the violations on the process taken as the ideal one, and subsequently determine possible root causes for it. The Pyclois Root Cause Analysis explores the specified tables in order to find dimensions or field values that are highly correlated to the defined KPI associated to the violation. This function uncovers the root causes by calculating the percentage of violations and the absolute number of violating cases for each unique value of a dimension (Celonis, 2022). The correlation metric used is the lift, and its formula is bellow presented at Equation 4.1, with A as the dimension, and B as root cause under analysis (Vu et al., 2019). A lift value higher than 1 indicates that the dimension is performing worse than the average.

$$Lift(A,B) = \frac{P(A|B)}{P(B)} \quad (4.1)$$

4.3.1 Dashboards

In order to understand the process, identify opportunities for improvement, and at the same time create a set of visual boards for the company itself to control their processes, there were built nine dashboards.

The dashboard construction had, as principle, the objective to always contain essential information for the proper analysis on the course of the business or process, and important KPIs. Additionally, while performing the inefficiencies identification, some factors stood out as potentially root causes, so it was considered that they should also be somehow demonstrated in the main views. Therefore, components that were capable of communicating the previous stated were added to the dashboards, in order to transmit it to the accountables for the process analysis.

Some filters are also available in all the tabs so it would be easier to quickly move the analysis focus through the same dashboard, in an efficient way. Nevertheless, in order to keep the visualization simple avoiding confusing and overwhelming analysis, other filters are just kept in specific tabs.

4.3.2 Costs calculation

For the purpose of measuring the costs associated to the actions performed by the company, a few assumptions had to be done.

According to Associação Portuguesa de Contact Centers (APCC), in Expresso (2022), in 2021 (year of the data under analysis), the outbound calls had 5.4 minutes of average duration. To complete the calculations, a wage of 7.5€/h for the personal in charge of performing the outbounds was considered, based on the typical wages for call-center jobs in Portugal. The expenses associated to investigation and implementation of the analytical model that define the outbound calls and the IVR were not considered.

The SMS cost was defined as 0.025€ based on CTT (2023), and the letter as 1.74€, according to CTT (2021).

4.4 Predictive Model

The Process Mining tool Celonis provides an embedded Machine Learning environment in order to leverage the projects. PyCelonis, a Python client library for Celonis can be used to interact with various resources of the platform, such as the data pools, tables and analyses. However, the version used for the development of this project did not allow the use of this feature. Nonetheless, an attempt to complement the analysis described so far through other means was made.

This section provides the main considerations taken regarding the intentions and applicability of the model, as well as the central topics concerning the implementation itself.

4.4.1 Considerations

The choice of the predictive model, for a more complete analysis and insights on the process, was conducted considering two key aspects:

- Adequacy of the model to the objective of improving the process. As mentioned at the Section 1.2, concerning the key project objectives, it was reasoned that a prediction of the next

activity from a process trace would enrich the analysis and allow higher quality recommendations.

- Availability of support documentation to the implementation, once the predictive model is not the primary focus of this project and is here used as complement to the application of Process Mining.

Therefore, the implemented model aims to predict the next activity and timestamp of the cases of the process. This implementation can be seen as a complement to the Process Mining application, having the goal to use the predictions to support recommendations, which are in turn also based on the preceding analysis and conclusions.

As input data, it was created a CSV file containing the invoice ID, the activity ID and the timestamp of each activity, for each invoice. This procedure was done through the Process Mining tool.

The algorithm used for the machine learning predictive model was the LSTM, and the implementation was based on the work of Tax et al. (2017). The mentioned study performs a practical analysis and comparison to the performance of several techniques for the next activity prediction and respective timestamp, using real-life data sets, in which the selected technique outperforms.

4.4.2 Implementation

Model

The method for predicting the next activity and respective timestamp was performed using distinct functions for the activity prediction function and time prediction. Each event, or in this case 'step', of the process trace is transformed into a feature vector, which is used as input for the LSTM.

The feature vector includes one-hot encoding of the activity identification, and three time components. The time-based features are added to the model in order to learn dependencies between time differences at different points in the process. These features include the time between the previous and current events, the time within the day, and the time within the week.

The last two time features are included to account for activities that can only occur during office hours and to learn that if an event occurred at the end of the working day or week, the time until the next event is expected to be longer.

During the learning process, the target output for the activity at a specific time step is set to the one-hot encoded representation of the activity that occurs one time step later. However, if the case ends at that time step, there is no new event to predict, so an extra element is added to the output one-hot encoded vector which is set to 1. Additionally, a second target output is set for the time difference between the current and the next event. The weights of the neural network are optimized using an algorithm to minimize the cross-entropy and mean absolute error (MAE) between the predicted and ground truth outputs.

As demonstrated by Tax et al. (2017), the architecture with better performance was the hybrid option, containing shared layers (regarding both activity and time prediction), followed by layers that were dedicated to a specific prediction task, with each layer specializing in one of the two predictions. Therefore this was the architecture applied, containing two layers, with one shared layer.

Implementation environment

The model was implemented with Python, using as base the available scripts by Tax et al. (2017).

Due to processing and memory limitations, certain conditions had to be considered:

- The number of epochs on the learning phase had to be reduced from 500 (in the existing model) to 25;
- The used dataset was a subset of all the activities for all the available invoices. As the complete dataset has a considerable large size, it was considered that the subset may represent a starting point to test the feasibility of the model.

The number of neurons per layer was kept as 100 neurons, once the tests on 75 and 150 neurons resulted in worst performance.

For the evaluation of the predictions, the accuracy of the activity prediction was calculated, and the error metric used for the time prediction was MAE. This choice was based on the fact that this metric is less susceptible to substantial errors on outlier data points, and there are time differences between two events in the log that are considerably large (Tax et al., 2017). The performance of predicting the next activity and its timestamp was evaluated by using the first 2/3 of the traces as training data and evaluating the predictions on the remaining 1/3 of the traces. The predictions were made on all prefixes of all traces in the set of test traces for except for the trace prefixes of size one due to the lack of data to base the predictions on.

Chapter 5

Results

The present chapter intends to demonstrate the project outcomes, from the dashboards building and analysis components, until the conclusions and recommendations taken from it. It is although crucial to keep in mind that the analysis was conducted only considering one month of invoices, whereby the insights cannot be considered as a standard behaviour.

Throughout this project, the creation of both the analysis components and the findings regarding the process were two components that constantly intersect and promote the development of each other. The outcomes are here demonstrated through the presentation of the dashboards, one by one, with a brief description of part of their components, and the most relevant insights and conclusions taken from each one of them. These views could not be presented in english due to the fact that the provided data was in portuguese, as mentioned in the previous chapter.

The presence of the dashboards visualization throughout the results presentation is crucial as they provide support to those outcomes. Notwithstanding, the substantial dimensions of the images necessitated their presentation in a reduced scale, consequently causing challenges in comprehending and analyzing them. In order to ensure the proper interpretation, the dashboards are presented at their full resolution in the Appendix B.

For the sake of the anonymity of the company, an hypothetical number of issued invoices within the scope of the analysis was considered as 1 million. The remaining numbers and calculations were achieved through the previous one.

5.1 Process Variant and Process Explorer views

Celonis has two powerful tools capable of showing how the process is actually functioning, which are the Variant Explorer and Process Explorer. The former allows the visualization of the different paths (or variants) of the process, while the latter is a visual representation of the process, allowing users to see the flow of the process and analyze it in a more intuitive way, with the option to increase or decrease the activities and connections shown. Figure 5.1 was obtained through the Variant Explorer and reflects the most common process flow occurring, which represents 51% of the cases.

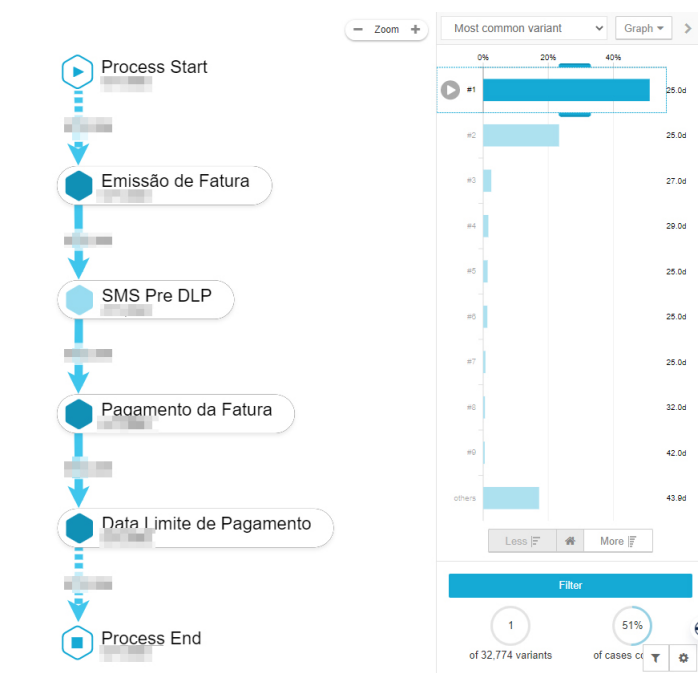


Figure 5.1: Most common process variant - Variant Explorer

Figure 5.2 reflects the complexity of the relationships between activities, displayed at the Process Explorer filtered by case frequency. The feature allows users to see the flow of the process and analyze it in a more intuitive way. All activities and possible connections between them are here displayed, but it is possible to drill down into specific process steps and identify patterns and anomalies.

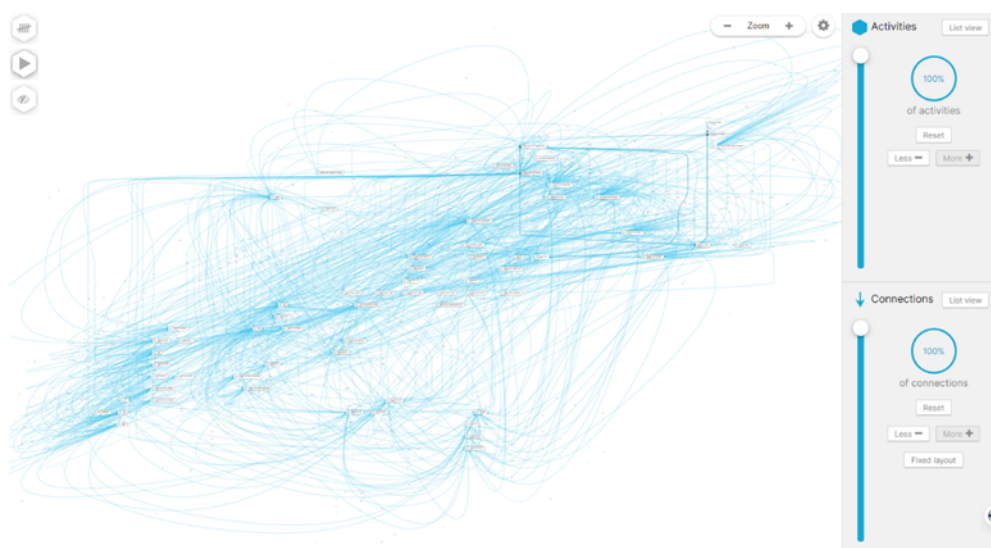


Figure 5.2: Process with all activities and connections - Process Explorer

5.2 General Process analysis

The visual components intend to not just extract insights on the business and identify inefficiencies, but also to serve for future monitoring. Thus, a dashboard with the most general aspects of the business was considered as essential, and shown in Figure 5.3. The image with more quality is addressed at Figure B.1 in the Appendix B.

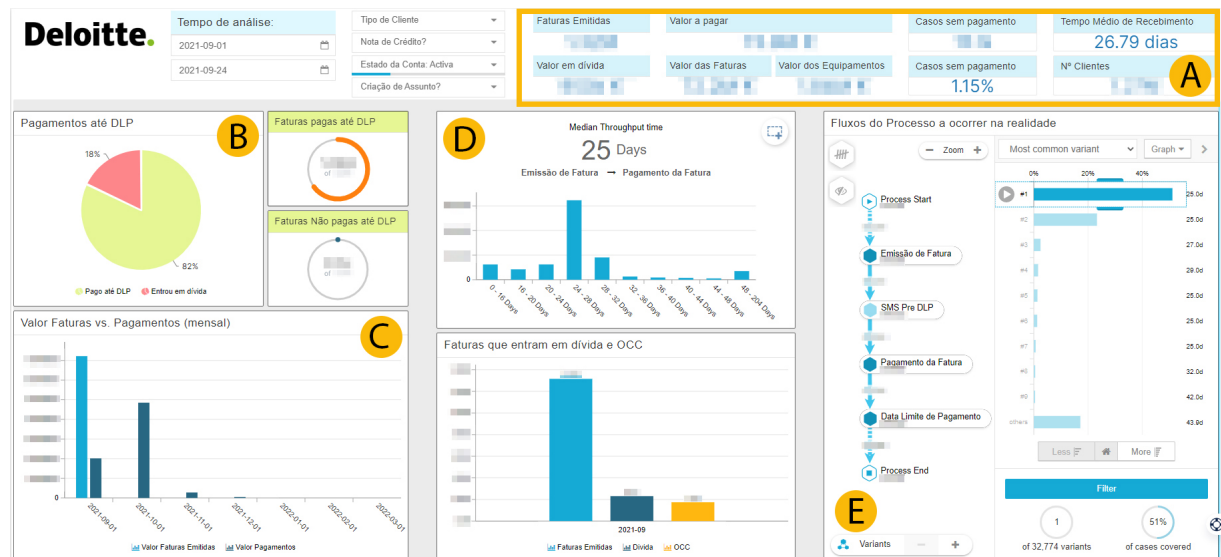


Figure 5.3: General Dashboard

View components

The dashboard displays some important numbers representative of the state of the business, such as:

- Numbers of issued invoices and corresponding sum of monetary values (€), clients, current debt, and cases without payment. These KPIs, alongside with the following point, are presented in the section A in Figure 5.3.
- Visualization of the number and percentage of invoices paid until the payment deadline date, present in the pie chart B in Figure 5.3.
- Visual comparison between the values on the invoices and the values that were actually paid, displayed at the section C in Figure 5.3.
- Analysis on the payment days, in the D section in Figure 5.3.
- Possibility to explore the several process flows that actually happened during the analysis period, on the small variant explorer represented by the letter E in Figure 5.3.

- The dashboard also allows to filter the cases according to the client type, the account state, or cases with the characteristic of credit payment and the subject creation, once they both could be useful to deepen the analysis.

View insights

Some of the insights on the business process, which were achieved by the aforementioned dashboard, are here provided.

Regarding the payment dates, it was found that 18% of the invoices were not paid until the payment deadline date, representing this way invoices in debt. Also, the average time between the issuance of the invoice until the payment was 26.79 days, while the median, found in the interval with more occurrences, was 25 days.

If analysing the first partial payment, by selecting the flow from "Invoice Issuance" until "First partial payment", the period until the final payment extends, since less than 1% of these cases has a payment between 2 and 41 days after the issuance.

To finish, customers without direct debit tended to postpone the payment. There were only 3% of the cases with direct debit that did not paid until the deadline date, while the number raises for 28% for the cases without this payment option.

5.3 Operational analysis

The main purpose of the operational view, displayed at Figure 5.4, is to provide a mean to keep the process under control regarding actions taken by the company. A larger scale of the dashboard image is provided in the Figure B.3 at the Appendix B. This dashboard aims for the supervision of the efficiency and effectiveness of the actions, such as the payment warnings before and after the payment deadline date, as well to provide an analysis to the penalty collection and some invoice details.

The supplementary filters intend to make more accessible the selection between invoices that have been paid or not paid by the due date. The option of the payment warnings enables a more detailed analysis regarding specific warnings.

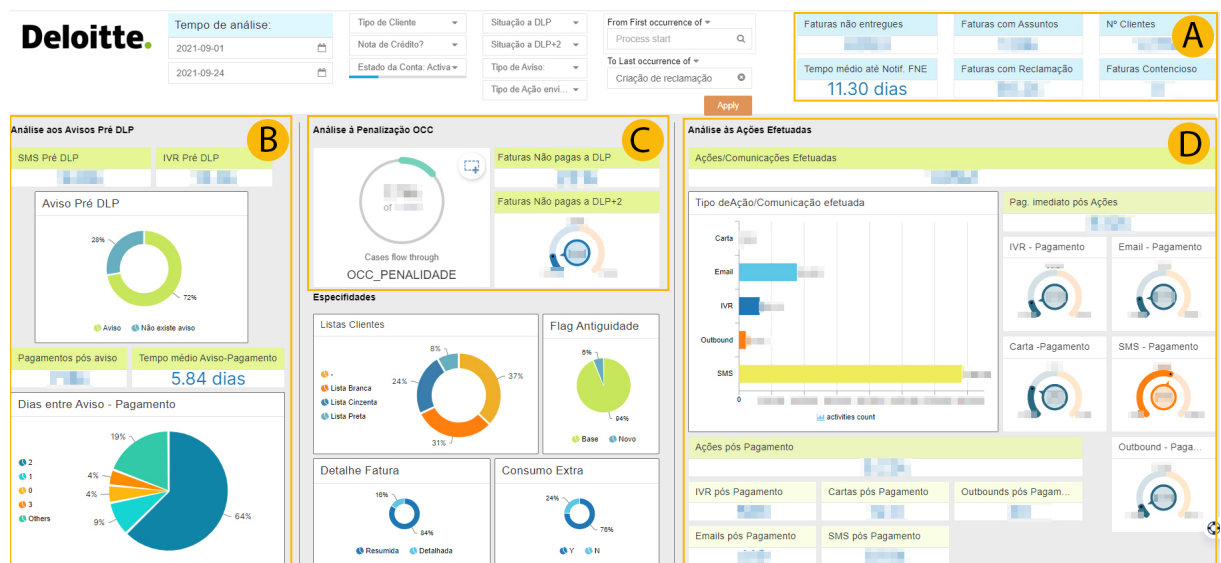


Figure 5.4: Operational Dashboard

View components

The emphasis of this view goes to the next components mentioned:

- The number of invoices that were not delivered, and time until the corresponding notification, both present on the KPIs in the area A in Figure 5.4.
- Analysis to the payment warnings and their efficiency, which can be found in the the section B in Figure 5.4.
- Analysis to the penalty collection, at the section C in Figure 5.4.
- Analysis to the actions made after the payment deadline date, available on the fragment D in Figure 5.4.

View insights

The dashboard in Figure 5.4 has allowed the discovery of pain points in the process. Some of them are here presented.

The postal invoice was not delivered 0.5% of the times. Even though it represents a residual number of cases, the number of invoices not delivered allows to take some insights on the benefits of using electronic invoice instead of postal. When looking into these 4.59k cases, 68% of them have electronic invoice, resulting in 24% of them not being paid until the due date. On the other side, regarding the cases with no electronic invoice, 42% were not paid on time. By receiving the digital invoice, there are more chances that the client will pay on time.

Regarding the warnings related to the limit date for the payment, 72% of the cases received this warning, and the majority of them was made via SMS. It was also noticed that 93% of the

invoices with direct debit received the warning and it took them 2.76 days to do the payment, coinciding with the deadline date. On the other side, the cases which do not have direct debit as payment mode, delayed the payment until 11 days past warning.

Concerning the process itself, the analysis indicated that there were mistakes on the penalty collection. The lack of payment until 2 days after the deadline represented 13.4% of the invoices, meaning that the respective clients should be charged with a penalty in the next invoice. However the analysis has shown that only 13.2% passed by the activity of that penalty being charged. So it is known that part of the clients are not being charged with this penalty when they should, but this kind of statement leads to the false belief that 0.2% of the invoices are not being collected at this stage.

While analyzing the invoices that were not paid until the referenced date, only 91% of it pass through the penalty. Translating the number to the most general scenario, 1.2% of the total invoices were not charged with the penalty, instead of the 0.2% mentioned above. The difference in these numbers represents cases that were paid up to the limit date for not receiving the extra charge, so they were indeed wrongly charged.

Also respecting the process, it was observed that missing payment actions were sent after payment. Almost 7k of the actions performed or sent as reminders for the clients to pay were sent after the payment was done. The costs associated may represent more than 1k€. These actions were performed at a fixed constant number of days after the payment deadline, so even in situations where the payment was performed between those activities, the actions were sent anyway.

5.4 Financial analysis

The financial view, available in Figure 5.5 and later in the Figure B.2 at the Appendix B, encompasses characteristics of the invoices and payments, actions costs and important numbers such as the actual debt, cases without payment and cases with breach of contract.

The available filters, along with the previous ones described, allow the selection of the payment type or the filtering of processes between the occurrence of two activities. These kind of analysis could help the user to more easily get insights on the process.

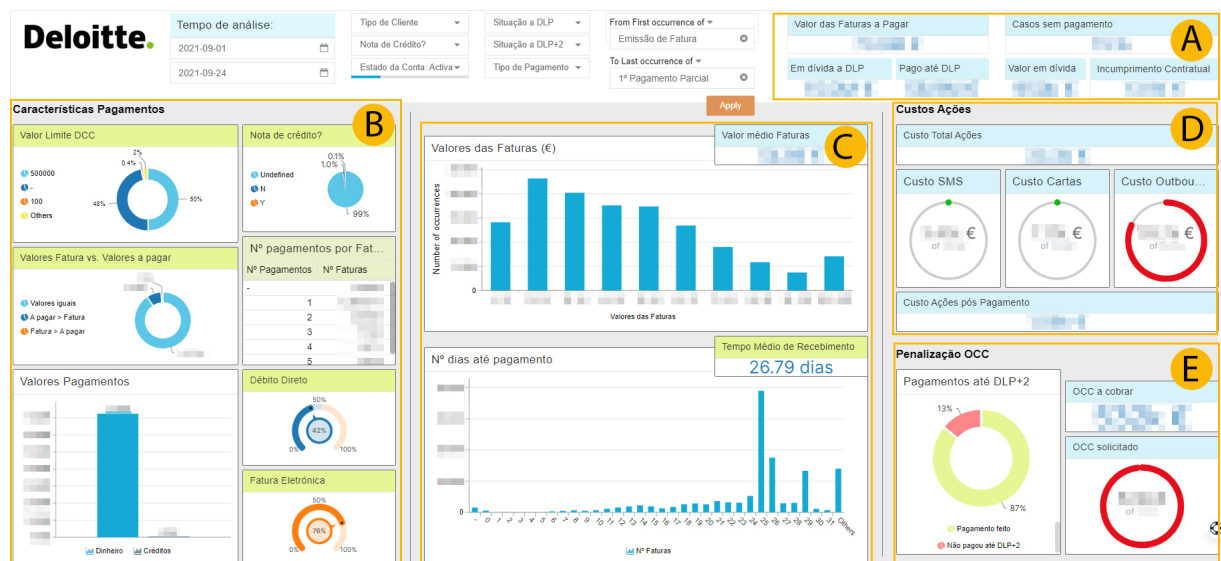


Figure 5.5: Financial Dashboard

View components

Some of the constituents of the financial dashboard are here presented:

- Discrimination of the total value of the invoices and the total value to actually be paid, presented with the remaining KPIs in the section A in Figure 5.5.
- Characteristics of the payments, occupying the fraction B in Figure 5.5.
- Distribution of the invoice values and days until payment, visible in the division C in Figure 5.5.
- Costs associated with the actions of the company, occupying the section D in Figure 5.5.
- Values regarding the penalty to collect and the value actually received, similar to the operational view but with monetary values, which are presented in the area E in Figure 5.5.

View insights

With this dashboard, it was possible to achieve some insights, mostly regarding payments. A few of them are provided below.

It was found that there was a difference between the total value of the invoices and the total value to actually be paid, and that this dissimilarity is mainly due to equipment charges.

Examination also revealed that customers with credit had demonstrated tendency to postpone the payment. Despite the fact that only 1.34k invoices have credit note, and therefore any insight should represent poor strength, the percentage of invoices paid until the deadline drop from 82% to 64%, also leading to an increase of 1 day in the average payment date.

5.5 Subject analysis

The Subject dashboard explores the subjects raised by the client and it is presented in Figure 5.6, and with more quality at Figure B.4 in the Appendix B. This topic could be associated with the customer satisfaction and the quality of the company's process, and therefore depicts an important topic to analyse and keep under control. It is important to remind that only subjects raised until 30 days after the invoice was issued were considered as a subject of the invoice.

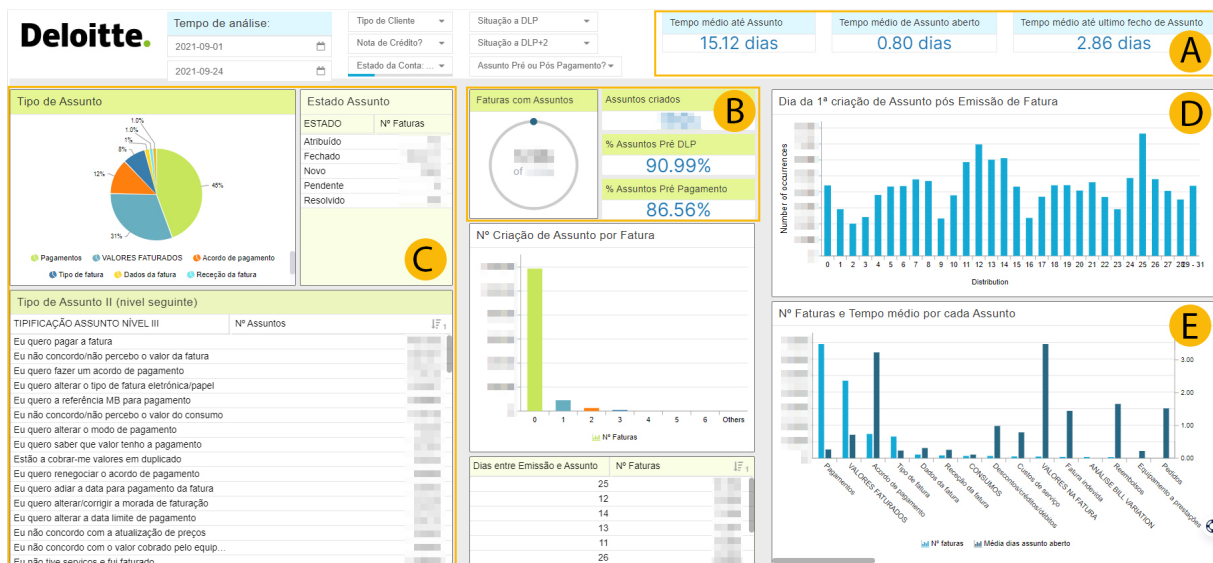


Figure 5.6: Subject Dashboard

View components

The principal components present on the view are bellow described:

- Average time to open, duration, and average time to closure the last subject of the invoice. All reflected on the part A of the dashboard in Figure B.4.
- Quantity of subjects raised and invoices with subjects, reflected on the area B in Figure B.4.
- Analysis on the subjects types and their state, on the section C in Figure B.4.
- Analysis to the creation dates, on the histogram at the component D in Figure B.4.
- Comparison between quantity of subjects of each type and their resolution times, displayed in the segment E of Figure B.4.

View insights

Throughout this dashboard creation and exploration, it was possible to achieve insights related to the subject creation, which impact the business process. The most important discoveries are here provided.

The invoices with subjects represented almost 10% of the total invoices. As regards the raised subjects, the five most requested topics together represent almost 93% of the total subjects. These subjects and their respective % are shown in Table 5.1.

Table 5.1: Top 5 Subjects

Subjects	
Payments	44%
Invoice values	30%
Payment agreements	12%
Invoice type	8%
Invoice data	1%

There were subjects raised due to missing or unclear information. The most common raised subjects in detail, are displayed in Table 5.2. When looking into these subjects, it seems that they were mainly originated by lack of available or easily accessible information. The simple requests "I want to pay", "I want the payment reference" and "Want to know what is the value to pay", appear to be themes that should not appear in such quantity on invoices, since they represent 32% of the total raised subjects.

Table 5.2: Top 8 detailed Subjects

Subjects	Nr. of subjects
'I want to pay the invoice'	34.2k
'I don't understand/agree with the value'	27.9k
'Want to make a payment agreement'	13.1k
'Want to change the invoice type'	10.1k
'Want the payment reference'	9.8k
'I don't understand/agree with the consumption'	8.7k
'Want to change the payment method'	5.7k
'Want to know what is the value to pay'	2.9k

Concerning the time periods in which the topics were created, available in section D of Figure 5.6, there were two periods with higher incidence of contacts. The day with more raised subjects was 25 days after the issuance, followed by day 12 and the involving period. At the twelfth day, the topics "I don't agree/understand the invoice value" and "I want to make a payment agreement" represented the most requested matters. Regarding day 25, the most seen subjects were "I want to pay" with almost half of the total here, followed by "I don't agree/understand the invoice value". The fact that the majority of clients (71%) received a warning SMS about the payment deadline at day 24 could be part of the cause of the peak on day 25.

By analysing the quantity of subjects of each type and their resolution times, displayed at section E in Figure 5.6, it was observed that "Payment Agreement" is the third most common

issue with a considerable number of 11.1k invoices, and has an average time of 3.23 days to be solved, a value that is 4 times higher than the average of 0.8 days.

5.6 Claims analysis

Similarly to the subject topics, also the claims analysis provide insights on the quality of the process. For this reason, a dashboard for the analysis of the previous was created, and is displayed in the Figure 5.7, and in larger scale in the Figure B.5 at the Appendix B.



Figure 5.7: Claims Dashboard

View components

A few of the components displayed at this dashboard are:

- Quantity of raised claims, and some time metrics regarding the opening and closure of the claims, with the remaining KPIs at the section A in Figure 5.7.
- Percentage of cases in each claim state at pie chart B in Figure 5.7.
- Analysis on the claims creation and closure motives, with the former on the chart column C, and the latter represented by the list on the component D Figure 5.7.

View insights

Similar to the previous described dashboards, also the Claims View has brought some useful findings which are here provided.

There were 71.3k raised claims, with an average of 13 days for the closure. By looking at the mode of time between the creation and predicted closure, it seems that the company has set by

default that complaints are closed 90 days after opening. It does not do justice to reality since the average time is 13 days, causing an average difference of -156.42 days between the forecast and the actual closing.

The complaint indicator in the table of issued invoices is not updated. There are 302 invoices with this indicator as “closed complaints”, however they had not passed through neither by the creation or closing of the claim. Moreover, there are 13.1k invoices with the ‘Open Claimed’ indicator, reflecting 31.2k claims, however, the table of claimed invoices indicates that only 386 claims are open, and that the remaining claims are already closed, a factor corroborated by the passage in the claim closing activity.

To conclude, there were 0.1% of invoices that generated a subject about the lack of receipt of the invoice, however there is no way to validate it. In the operational analysis, in section 5.3, it was noted that the postal invoice was not delivered for 0.5% of the cases.

5.7 Conformance view

The Conformance view of Celonis shows the extent to which actual business processes align with the ideal or intended process model. It compares the process model to actual process instances, highlighting deviations and non-compliant areas, and providing insights on how to improve the process (Celonis, 2022). Therefore, the presence of this dashboard is crucial and the main insights are provided in this section.

5.7.1 Comparison between real process flow and ideal

In order to compare the real process flow to the ideal one, this is, to perform the conformance phase of Process Mining, it was established that the optimal flow would be the Payment after the invoice being issued and before the payment deadline. With this input, the Conformance automatically computes the percentage of conforming cases as 23%. The extremely low number is derived from the warnings regarding the payment deadline which are sent in the middle of the ideal process flow, as shown in Figure 5.1. As this action does not represent something out of the expected, neither a barrier for the payment until the deadline, these actions were considered as part of the allowed list of the platform, rising the conforming cases to 75%.

5.7.2 Root Cause Analysis

As already described, the Celonis platform is able to identify the violations on the process taken as the ideal one, and subsequently determine possible root causes for it. In the Conformance overview is possible to see that, besides the already mentioned 75% of conforming cases, the throughput time for the violating cases increased by 16.1 days, passing from an average of 26 to 42.1 days to complete the process. The average number of activities also increased, from 3.7 to 9.2 days.

The top 4 violations detected by the platform are displayed in Table 5.3. The first violation represents a sequence of the process that should not happen, meaning the payment deadline date occurring before the payment. Therefore this violation deserves further exploration. Regarding the following two violations, the Penalty Warning and the Penalty, they are inevitable consequences for the excessive throughput time, being automatically raised by the days passing without payment. For this reason they do not represent a target for further development. Finally, the Subject Creation represents a different type of violation, by being an actual activity initiated by the client at any time that should be avoided. It is, for this reason, a topic that is here better analysed.

Table 5.3: Top 4 Violations

Violations	% of cases
Invoice issuance followed by Payment Deadline date	18%
SMS Warning penalty	16%
Penalty	13%
Subject creation	10%

Invoice issuance followed by payment deadline date

The most common, and important violation of the process flow is the fact that the invoice issuance was followed by payment deadline date, meaning that the invoice was not paid on time, and it happened on 18% of the cases. By performing the root cause analysis for this violation, the results exhibited in Table 5.4 were obtained, which are sorted by correlation. The factors that had a correlation superior to the lift value of 1.5, are here displayed in order to facilitate the analysis and only consider the most important factors for the violation.

Examples on the interpretation of the table are next provided:

- The first two rows of the table represent the cases where no payment was done.
- The Claim indicator, as demonstrated previously, is not updated with the real state of the claim, therefore the open state of the claim would get lower correlation, since there are fewer open claims and more closed ones. The correlation of both states is indeed related to the existence of claims, and not so much about the respective state.
- The Debt Step (and Debt max step) indicate that for the cases with this violation, usually the client did not extend the process to more than 1 or 2 actions performed by the company to achieve the payment.
- The payment method 'Other' positively impacted this root cause, meaning for example that the direct debit favored the payment on time.

Table 5.4: Root causes for Issuing invoice followed by payment deadline date

Possible root causes for violation		Lift	Nr. Cases
Payment real date	-	5.706	11.4k
Payment system date	-	5.706	11.4k
Claim indicator	Open	5.703	13.2k
	Closed	2.641	13.7k
Debt Step	1	5.392	29.3k
	2	5.223	11.9k
Debt max step	1	5.392	29.3k
	2	5.223	11.9k
Client colored list	Grey	4.19	80.3k
	Black	1.945	58.5k
Number of payment	2	4.112	15.8k
Type of payment	Another	1.586	160k
Client payment type	Another	1.585	157.7k
Direct debit activated?	N	1.583	160k
Special colored list?	N	1.504	164.6k

- The clients belonging to the grey list seem to be considerably present on the cases where the payments did not happen before the deadline date, with a lift of 4.19.

Subject creation

Regarding the violation concerning the subject creation, the possible root causes, with highest lift values, are displayed in Table 5.5.

Table 5.5: Root causes for Subject creation

Possible root causes for violation		Lift	Nr. Cases
Credit payment indicator	No	8.311	7.792k
Payment real date	-	6.490	7.263k
Payment system date	-	5.703	7.263k
Claim indicator	Open	5.637	16.363k
	Closed	6.298	8.128k
Number of payments	2	5.030	10.793k
	-	4.512	5.225k
First payment date	-	4.512	5.225k
Last payment date	-	4.512	5.225k
Credit payments value	-	4.512	5.225k
Total payments value	-	1.945	5.225k
Seniority flag	New	2.199	12.359k
Client colored list	Grey	1.407	16.363k
	Black	2.042	8.128k

The interpretation is similar to the preceding statements about the first violation. Nevertheless, a few insights on the process obtained by this analysis are next provided:

- The dimension with more correlation to the subject creation is the absence of the credit payment indicator. Its lift indicates that the 'N' option to this feature is performing 8.311 worse than the average. In fact, this option was only present in 1% of the total cases, while it represents 8% of the invoices with subjects, manifesting a strong correlation.
- The absence of payment, which is present directly or indirectly in 7 rows of the provided table, is also correlated to the subject creation.
- The clients from the Black list also may present higher tendency of creating subject.

5.8 Predictive Model outcome

Regarding the prediction of the next activity, by considering all the traces predictions, as described in the previous chapter, the model has shown an accuracy of 78.62%.

Considering only the instances where the activity prediction was correct, the average of days between the predictive time and real activity time was 2.64 days.

The conclusions that can be drawn from here focus on decision-making about actions to be taken considering the next activity prediction. A few examples are presented on the following section, as recommendations.

5.9 Recommendations

Some of the process findings and the predictive model have culminated in a set of recommendations in order to improve the process, and which are presented in this last section of the Results chapter. The recommendations provided in this document are based solely on what is currently known about the process so they may be subject to change as more information becomes available.

Revise actions schedule

As described in Section 5.3, there were warnings about missing payments that were sent after the payment was done. This kind of scenario results in not just extra costs, but also may potentiate client churn.

- It is suggested that the system confirms the payment status more frequently in order to avoid such situation.

Furthermore, without considering the cases with direct debt, since the payment is automatically done 2 days after the warning message, only 28% of the cases have paid in the day or in the day after the message, and 38% have paid 3 or more days after.

- In order to respond to it, a warning could be sent in more advance. Also considering that the SMS had demonstrated better performance, the advice would be to sent this warning via SMS.

Schedule actions based on the predictive model

- If the prediction addresses the "Payment" as next activity, happening before the scheduled date for some action by the company in order to collect the payment, the system should cancel these warning actions.

Furthermore, this question was addressed following the operational analysis about the actions submitted, on Section 5.3. It was verified that some warning actions about missing payments were carried out after the client paid and it was recommended to confirm the existence of the payment more frequently. The insight of the predictive model is aligned with the previous recommendation, and reinforces the recommendations approach.

- If the prediction addresses the "Partial payment" as next activity, a more intensive approach should be adopted, regarding warning actions. As shown on previous analysis, in Section 5.7, a number of payments equals to 2 was considered as a possible root cause for the violation of payment after deadline date, and also subject creation, reinforcing this way that more efforts should be taken in order to avoid this scenario.
- If the prediction addresses the "Subject creation" as next activity, which is also an activity that deviates from the ideal process flow, a faster response could be prepared. Knowing that the customer may create a subject, a preliminary analysis of the customer's situation can be performed in order to better respond and, by better, it is meant faster to their question. This analysis may involve automatically aggregating all the necessary data that is normally required to respond to subjects raised by customers that have already been analyzed here.

Revise the invoice data

In Section 5.5, it was identified that at least 32% of the raised subjects appear to happen due to missing or unclear information.

- Regarding that finding, the suggestion goes through reviewing the invoice description and to assure that all the information needed for the proper payment in time is provided and in an accessible and understandable way.

Pre flag biased invoices to certain subjects

Cases that are aligned with the possible root causes will be more likely to violate the ideal process. It is in accordance to the previous statement that the following recommendation is based on.

During the analysis, it was noted that the Payment Agreement is the subject, among the most raised subjects, that has the longest time to closure. This statement can be confirmed in the section C of the Subject Dashboard, which can be better visualized in the Figure B.4 in the Appendix B. Following, it was discovered that the average invoice value has raised 16% when considering only

the invoices with subjects, and almost 47% when filtering for the invoices with the mentioned subject. In fact, comparisons on the raised subjects on invoices with monetary values between zero and around the average value, and the invoices with higher values, showed that the subject Payment agreement had gained force passing from 8% of the total to 17%, and also the subject Invoice values, going from 27% from 35%.

Furthermore, the colored client list had interference both on the invoice issuance being followed by payment deadline, and Subject creation, as demonstrated by the root cause analysis in Section 5.7.2. The distribution of clients into their lists according three different filters, being them the general analysis, the invoices that had subject creation, and lastly the invoices which created a subject on Payment agreements, is displayed in Table 5.6.

Table 5.6: Subjects on clients list

Colored clients list	General	Subject	Payment agreement
-	37%	24%	2%
White	31%	27%	12%
Grey	24%	33%	50%
Black	8%	16%	37%

The difference between the average invoices values inside each colored list, is not significant in order to consider that there is a connection between the list and the invoices values.

- All aspects being considered, the cases belonging to the Black or Grey list and with higher invoice values, could be flagged in order to speed up the process by, for example, having the conditions for the payment agreement more easily available.

The company has provided the information that the payment agreement is only possible for cases where there are at least two paid invoices. As there is only data from 1 month of invoices, it is not possible to verify whether these subjects are being created for invoices that do not fulfill the condition.

- Nevertheless, an impediment to the creation of this subject for cases where the condition is not met should be applied, in order to not waste time with the situation and directly send an alternative solution.

The factors on the root cause analysis for the violation Issuing invoice followed by payment deadline date, shown in Table 5.4, are indicative to what factors may lead to the absence of in time payment.

- These dimensions could together represent the requirements for a special behaviour of the company. Customers on this list could have, for example, more frequent notices, or in case of not intending to create more warnings, redirect some of the predefined warnings to these cases.

Chapter 6

Conclusions and future work

In this last chapter, a final and brief reflection on the project from the optic of the results obtained is performed. Future potential improvements are also presented, in order to evidence the project limitations and support further developments.

6.1 Conclusions

The present dissertation has provided a comprehensive examination of the process flow of the collections process of a service provider company. Through the application of Process Mining techniques, this project has been able to shed light on the inefficiencies within the company's processes, such as the delayed warning actions and the failure to impose penalties on certain clients. These findings highlight the importance of a thorough analysis of process flow and the need for continuous improvement.

To support this analysis and the company, a Process Mining environment was created, through the building of dashboards that enable the company to monitor and track the key aspects of their process. This will provide the company with valuable insights into the efficiency of their operations, enabling them to make informed decisions and take appropriate actions to improve their processes.

The previous mentioned improving actions are sustained by the final part of this project which was to make recommendations that will assist the company in predicting the next activity and taking the necessary steps to ensure the continued improvement of their processes.

The results of this dissertation demonstrate the potential of Process Mining in improving process efficiency and contributing to the success of a business. Overall, this research has provided valuable insights into the potential of Process Mining as a tool for process improvement and enhancement.

6.2 Future work

In the course of this project, several points for improvement were identified, which could potentially bring value to the business. These points are worth mentioning as they deserve attention.

The first, and most simple, component that should be changed for future work would be an increased dataset. In order to extract meaningful insights and identify inefficiencies in a more accurate way, it would be crucial to have more than one month of data.

Second, the project demonstrated part of the capabilities of the tool, however, others remained unused, on the majority due to the absence of real timing input data. Recently, there has been a growing interest in using Process Mining for online operational support, driven by the increasing availability of real-time data and powerful computing capabilities. While offline analysis focuses on analyzing past process data, online operational support aims to monitor and shape processes as they are happening. Predictive business process monitoring is a key component of this approach, as it allows for proactive and corrective actions to be taken, based on real-time process information. These actions may include recommendations for task and resource allocation, as well as notifications about potential risks Park and Song (2020). Therefore for future work, the implementation should be on the systems of a company instead of using offline files.

Regarding the predictive model, in order to achieve more accurate results, other models and techniques should be tested. While this document has explored the topic of predictive business process monitoring, there have been more recent and advanced studies on the subject that were not covered due to limitations in time and resources, namely the memory capacity and processing power. Additionally, this document's main focus was not on the implementation of a predictive model, and its implementation had the main goal of complementing the Process Mining environment. Therefore, an analysis on the work of Pasquadibisceglie et al. (2021), which has improved the current model used in this project, could be used to leverage future work.

Also regarding the predictive model, the available capacity only allowed to use a subset of the eventlog for the model. This may not be enough to train a predictive model. It is recommended, for future work, to incrementally increase the amount of data used for training as more computational resources become available.

Bibliography

- Agrawal, R., Gunopulos, D., and Leymann, F. (1998). *Mining process models from workflow logs*. Advances in Database Technology — EDBT’98. EDBT 1998. Lecture Notes in Computer Science, vol 1377. Springer, Berlin, Heidelberg.
- Brocke, J. v. (2020). *Chapter Abstracts of Keynotes in Business Process Management : 18th International Conference, BPM 2020, Seville, Spain, September 13-18, 2020, Proceedings*. Springer Nature Switzerland AG.
- Camargo, M., Dumas, M., and González-Rojas, O. (2019). Learning accurate lstm models of business processes. In Hildebrandt, T., van Dongen, B. F., Röglinger, M., and Mendling, J., editors, *Business Process Management*, pages 286–302, Cham. Springer International Publishing.
- Celonis (2022). Celonis academy training.
- Chen, N., Zhu, J., Sun, F., and Xing, E. P. (2012). Large-margin predictive latent subspace learning for multiview data analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(12):2365–2378.
- CTT (2021). Preços 2021. <https://www.ctt.pt/dA/c63a5fba43/Tarifario-Janeiro-2021.pdf>.
- CTT (2023). Sms e email. <https://www.ctt.pt/empresas/marketing-publicidade/cttads/sms-email>.
- Cunningham, P., Cord, M., and Delany, S. J. (2008). *Supervised Learning*, pages 21–49. Springer Berlin Heidelberg, Berlin, Heidelberg.
- Diebold, F. X. (2019). On the origin(s) and development of ‘big data’: The phenomenon, the term, and the discipline. *University of Pennsylvania*.
- Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, 26:745–766.
- Dumas, M., La Rosa, M., Mendling, J., and Reijers, H. A. (2018). *Fundamentals of Business Process Management*. Springer-Verlag GmbH Germany, 2 edition.
- Ebert, C. and Duarte, C. H. C. (2018). Digital transformation. *IEEE Software*, 35:16–21.
- Evermann, J., Rehse, J.-R., and Fettke, P. (2017). A deep learning approach for predicting process behaviour at runtime. pages 327–338.
- Expresso (2022). Número de trabalhadores de ‘contact center’ em portugal aumentou 19% em 2021. *Jornal Expresso*.
- EY (2019). Process mining and its impact on bpm.

- Forrester (2022). Trends in process improvement and data execution. Forrester consulting.
- Galic, G. and Wolf, M. (2021). Global process mining survey. Deloitte.
- Garcia, C. d. S., Meincheim, A., Faria Junior, E. R., Dallagassa, M. R., Sato, D. M. V., Carvalho, D. R., Santos, E. A. P., and Scalabrin, E. E. (2019). Process mining techniques and applications – a systematic mapping study. *Expert Systems with Applications*, 133:260–295.
- Gordon-Rodriguez, E., Loaiza-Ganem, G., Pleiss, G., and Cunningham, J. P. (2020). Uses and abuses of the cross-entropy loss: Case studies in modern deep learning.
- Graves, A. (2013). Generating sequences with recurrent neural networks. *CoRR*, abs/1308.0850.
- Hand, D., Mannila, H., and Smyth, P. (2001). *Principles of Data Mining*. Mit Press.
- Hunter, R. and Smith, S. (2021). The business of digital business. Gartner.
- Ishwarappa and Anuradha, J. (2015). A brief introduction on big data 5vs characteristics and hadoop technology. *Procedia Computer Science*, 48:319–324. International Conference on Computer, Communication and Convergence (ICCC 2015).
- Jalonen, H. and Lönnqvist, A. (2009). Predictive business – fresh initiative or old wine in a new bottle. *Management Decision*, 47:1595–1609.
- Kerremans, M., Srivastava, T., and Choudhary, F. (2021). Market guide for process mining. Gartner.
- Koech, K. E. (2020). Cross-entropy loss function.
- Kourla, S. R., Putti, E., and Maleki, M. (2020). Importance of process mining for big data requirements engineering. *International Journal of Computer Science and Information Technology*, 12:1–12.
- Kraus, S., Jones, P., Kailer, N., Weinmann, A., Chaparro-Banegas, N., and Roig-Tierno, N. (2021). Digital transformation: An overview of the current state of the art of research. *SAGE Open*, 11.
- Laney, D. (2001). 3d data management: Controlling data volume, velocity, and variety. *META Group*.
- LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521:436–444.
- Marin-Castro, H. M. and Tello-Leal, E. (2021). Event log preprocessing for process mining: A review. *Applied Sciences*, 11(22).
- Martin, N., Fischer, D. A., Kerpedzhiev, G. D., Goel, K., Leemans, S. J. J., Röglinger, M., van der Aalst, W. M. P., Dumas, M., La Rosa, M., and Wynn, M. T. (2021). Opportunities and challenges for process mining in organizations: Results of a delphi study. *Business & Information Systems Engineering*, 63:511–527.
- Martins, A. and Astudillo, R. (2016). From softmax to sparsemax: A sparse model of attention and multi-label classification. In Balcan, M. F. and Weinberger, K. Q., editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1614–1623, New York, New York, USA. PMLR.

- Medsker, L. and Jain, L. (1999). *Recurrent Neural Networks: Design and Applications*. International Series on Computational Intelligence. CRC Press.
- Meier, H., Lagemann, H., Morlock, F., and Rathmann, C. (2013a). Key performance indicators for assessing the planning and delivery of industrial services. *Procedia CIRP*, 11:99–104. 2nd International Through-life Engineering Services Conference.
- Meier, H., Morlock, F., and Dorka, T. (2013b). *Delivery Management of Industrial Product-Service Systems - Challenges for a Performance Measurement*, pages 455–460. Advances in Manufacturing Technology XXVII. Cranfield University.
- Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill Science, Engineering, & Math.
- Moore, L. (2020). Process mining vs. data mining: What's the difference?
- Park, G. and Song, M. (2020). Predicting performances in business processes using deep neural networks. *Decision Support Systems*, 129:113191.
- Park, S. and Kang, Y. S. (2016). A study of process mining-based business process innovation. *Procedia Computer Science*, 91:734–743.
- Partington, A., Wynn, M., Ouyang, C., Suriadi, S., and Karnon, J. (2015). Process mining for clinical processes: A comparative analysis of four australian hospitals. *ACM Transactions on Management Information Systems*.
- Parviainen, P., Tihinen, M., Kääriäinen, J., and Teppola, S. (2017). Tackling the digitalization challenge: how to benefit from digitalization in practice. *International Journal of Information Systems and Project Management*, 5:63–77.
- Pasquadibisceglie, V., Appice, A., Castellano, G., and Malerba, D. (2021). Leveraging multi-view deep learning for next activity prediction.
- Provost, F. and Fawcett, T. (2013). *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking*. O'Reilly Media, Inc.
- Ragab, M., Jadid Abdulkadir, S., Aziz, N., Al-Tashi, Q., Alyousifi, Y., Alhussian, H., and Alqushaibi, A. (2020). A novel one-dimensional cnn with exponential adaptive gradients for air pollution index prediction. *Sustainability*, 12:10090.
- RedGate (2021). The real-world challenges of growing server estates insights from the 2021 state of database monitoring survey part 1.
- Reinkemeyer, L., editor (2020). *Process Mining in Action: Principles, Use Cases and Outlook*. Springer Cham, 1 edition.
- Sak, H., Senior, A., and Beaufays, F. (2014). Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition.
- Stolterman, E. and Fors, A. C. (2004). Information technology and the good life. In *Relevant Theory and Informed Practice*.
- Suriadi, S., Andrews, R., ter Hofstede, A. H. M., and Wynn, M. T. (2017). Event log imperfection patterns for process mining: Towards a systematic approach to cleaning event logs. *Information Systems*, 64:132–150.

- Tax, N., Verenich, I., Rosa, M. L., and Dumas, M. (2017). Predictive business process monitoring with LSTM neural networks. In *Advanced Information Systems Engineering*, pages 477–492. Springer International Publishing.
- van der Aalst, W. (2016). *Process Mining Data Science in Action*. Springer Berlin, Heidelberg.
- van der Aalst, W. M. P. (2011). *Process Mining : Discovery, Conformance and Enhancement of Business Processes*. Springer Berlin, Heidelberg.
- van der Aalst, W. M. P. (2013). Business process management: A comprehensive survey. *ISRN Software Engineering*, 2013:1–37.
- Veit, F., Geyer-Klingenberg, J., Madrzak, J., Hang, M., and Thomson, J. (2017). The proactive insights engine: Process mining meets machine learning and artificial intelligence. In *15th International Conference on Business Process Management*.
- Vu, K., Clark, R. A., Bellinger, C., Erickson, G., Osornio-Vargas, A., Zaïane, O. R., and Yuan, Y. (2019). The index lift in data mining has a close relationship with the association measure relative risk in epidemiological studies. *BMC Medical Informatics and Decision Making*, 19.
- Weske, M. (2019). *Business Process Management : Concepts, Languages, Architectures*. Springer-Verlag GmbH Germany, 3 edition.
- Westerman, G., Calm ejane, C., Bonnet, D., Ferraris, P., and McAfee, A. (2011). Digital transformation: A roadmap for billion-dollar organizations. , Capgemini Consulting & MIT Center for Digital Business.
- Zerbino, P., Stefanini, A., and Aloini, D. (2021). Process science in action: A literature review on process mining in business management. *Technological Forecasting and Social Change*, 172:121021.
- Zhou, Z.-H. (2021). *Machine learning*. Springer Nature Singapore.

Appendix A

Dataset

Table A.1: Data set

Table	Description	# columns	Some table fields
Issued Invoices	Varied information on the issuance of each invoice	54	ID Invoice ID Billing Account ID Client Account Type of client Payment deadline date Electronic payment? (...)
Undelivered Invoices	Tracking of the invoices that do not arrive to the clients	3	ID Invoice ID Billing Account Undelivered invoice notification date
Claimed Invoices	Information on the creation, and remaining path, of clients' claims	14	ID Invoice ID Billing Account ID Client Account Claim date Number of open claims (...)
Subjects	Information about the topics that led the clients to interact with the company	19	ID Subject ID Client Account Date Type of subject (...)
Steps	Information and tracking on the invoices that were not paid until the due date	17	ID current step ID Invoice ID Billing Account ID Client Account Date Debt value at each step (...)
Actions	Similar to the Steps table, but with increased detail	12	ID Action ID Invoice ID Billing Account Date State of Action State of External Action (...)

Appendix B

Dashboards



Figure B.1: General Dashboard



Figure B.2: Financial Dashboard



Figure B.3: Operational Dashboard



Figure B.4: Subjects Dashboard



Figure B.5: Claims Dashboard