UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR
DE INGENIEROS DE TELECOMUNICACIÓN

GRADO EN INGENIERÍA DE TECNOLOGÍAS Y SERVICIOS DE TELECOMUNICACIÓN

TRABAJO FIN DE GRADO

PROCESS MINING FOR OPTMIZATION OF A LOAN APPROVAL PROCESS IN A FINANCIAL INSTITUTION

Eduardo Varas Gándara

2017

TRABAJO FIN DE GRADO

Título: Aplicación de Process Mining a la optimización de un proceso de solicitud de prestamos de una entidad financiera

Título (inglés): Process Mining for Optimization of a Loan Approval Process in a Financial Institution

Autor: Eduardo Varas Gándara

Tutor: Carlos A. Iglesias Fernández

Departamento: Ingeniería de Sistemas Telemáticos

MIEMBROS DEL TRIBUNAL CALIFICADOR

Presidente:

Vocal:

Secretario:

Suplente:

FECHA DE LECTURA:

CALIFICACIÓN:

Resumen

Las instituciones financieras están experimentando cambios drásticos tras la crisis financiera mundial de 2008. Por un lado, las instituciones financieras deben adaptarse rápidamente a las nuevas regulaciones que conllevan una adaptación rápida de sus procesos internos. Por otro lado, las Fintechs están cambiando las reglas bancarias tradicionales introduciendo procesos financieros innovadores.

Por lo tanto, las instituciones financieras deben ser capaces de mejorar sus ineficiencias operativas, estas mejoras suponen un business case para la aplicación de las técnicas de minería de procesos. En este trabajo de fin de grado, nos enfrentamos a un caso real de busqueda de mejoras operativas a traves de diferentes técnicas de minería de procesos y de minería de datos.

Los datos utilizados en este trabajo fueron proporcionados en el BPIC 2017. Este desafío propone un caso de uso real en el que una institución financiera holandesa proporciona registros de eventos del proceso de aprobación de préstamos, formado por 1.202.267 eventos relacionados con 31.509 solicitudes. Para el enfoque, combinamos la herramienta de minería de procesos de Disco con herramientas de análisis de datos y visualización basadas en Python como Pandas para combinar diferentes técnicas de inspección en distintos niveles de granularidad para proporcionar respuestas a las preguntas dadas.

En particular para el estudio nos centramos en las principales demandas del desafío 2017, que son: tiempos del proceso por fases del mismo, influencia de la frecuencia de incompletitud en la documentación en el resultado final y frecuencia de los clientes que solicitan más de una oferta. Nuestro enfoque ha consistido en identificar las fases del proceso y analizar cada pregunta por fase del proceso buscando comportamientos extrapolables al funcionamiento global del proceso. Finalmente, se discuten las observaciones finales y el trabajo futuro.

Palabras clave: Minería de procesos, Minería de datos, Python, Disco, Fintech, Proceso, Análisis

Abstract

Financial institutions are experiencing drastic changes after the global financial crisis of 2008. On the one hand, financial institutions need to quickly adapt to new compliance regulations that require adapting their internal processes. On the other hand, Fintechs are changing the traditional banking rules by introducing innovative financial processes.

Therefore, financial institutions need to be able to improve their operational inefficiencies, and represents a business case for the application of process mining techniques. In this final degree project a real case is faced through different process and data mining techniques.

The data used was provided in the BPIC 2017. This challenge proposes a real use case where a Dutch financial institution provides event logs of the loan approval process, with 1.202.267 events pertaining to 31.509 applications. For the approach, we leverage the Disco process mining tool with Python-based data analysis and visualization tools such as Pandas in order to combine different granularity inspection techniques to provide answers to the given questions.

In particular, we focus on the main requests from the BPIC 2017 challenge, which are: throughput times per part of the process, influence on the frequency of incompleteness to the final outcome and the frequency of customers asking for more than one offer. Our approach has consisted in identifying the process phases and analyzing each question by phase and then globally. Finally, we discuss concluding remarks and future work.

Keywords: Process Mining, Data Mining, Python, Disco, Fintech, Process, Analysis

Agradecimientos

Gracias a mis padres por motivarme a empezar esta carrera y apoyarme durante estos años hasta acabarla.

Contents

List of Figures

CHAPTER 1

Introduction

Financial institutions are experiencing drastic changes after the global financial crisis of 2008 [11]. On the one hand, financial institutions need to quickly adapt to new compliance regulations that require adapting their internal processes. On the other hand, Fintechs are changing the traditional banking rules by introducing innovative financial processes. Therefore, financial institutions need to be able to improve their operational inefficiencies, and represents a business case for the application of Process Mining (PM) techniques. PM [13] is a set of techniques which allows the discovery, conformance and enhancement of business processes, thanks to the record of business process events.

PM has three principal branches, which main difference is their aim. This branches could be linked for a complete process enhancement. The first one is the process discovery. This branch applies different algorithms to the finding of the process behavior. In this branch different algorithms are used looking for the process model, bottlenecks, and different interesting characterisitcs that could help in the process understanding. Some of the most known algorithms for the process model discovery are the alpha algorthim, the heuristic algorithm or the fuzzy miner. This discovery branch is really helpful to understand the real process performance and the understanding of the real workflow. The next branch is the conformance one. This branch aim is the conformance between the theorical process model and the real process behavior. For this phase the main task is the finding of the possible differences and deviations in the process by comparing it with the dicovery results. And the final PM branch is the enhancement branch. This branch looks for the process optimization in base of the discovery and the conformance analysis results. After the analysis result from the first branches in this last one improvements, and problem solutions are proposed.

For this final degree project data from a real use case was studied. This data was obtained from the Business Process Intelligence Challenge (BPIC) 2017, where a Dutch financial institution provides event logs of the loan approval process, with 1.202.267 events pertaining to 31.509 loan applications filed in 2016 and their handling up to February 2017. This log is from a Loan Application Process and, deliver us the following information about the different events: resource who execute the event, beginning and completion date, application type (new credit or limit raise), requested amount or loan goal. In the log, events are classified into three groups depending on the type of activity distinguishing between Application type, Offer type, and Workflow Type. Application type events are those which represent the state of the application process, while Offer type events are the ones where an offer changes its state, and Workflow type events represent the state of the different work items. It is intended to understand the business process in detail and try to find the different opportunities for optimization in the process, focusing on the topics requested by the financial institute:

- Throughput times per part of the process. Particularizing in the difference between the time spent in the company's systems waiting for processing by a user and the time spent waiting on input from the applicant
- The influence on the frequency of incompleteness to the final outcome. The hypothesis here is that if applicants are confronted with more requests for completion, they are more likely to not accept the final offer.
- How many customers ask for more than one offer. Keeping in mind if the different offers are asked for in a single conversation or in multiple conversations.

For most of the questions, there is a need to understand the interactions among the application and the loan system, and their impact of the outcome of the process. The reminder of the work is organized as follows. Chap. 2 presents our strategy to answer the proposed questions, and reviews previous works in the same application domain. Chap. 3 describes briefly the Extract, Transfor and Load (ETL) process and addresses the identification of phases in the provided process. As a result of this, we have identified two phases: *application creation* and *application validation*, which are detailed in Chap. 4 and Chap. 5, respectively. Finally, Chap. 6 draws overall conclusions.

CHAPTER 2

Methodology and Background

The methodology followed to analyze the problem is shown in Fig. 2.1. First, we have analyzed the articles from BPIC 2012, since they used a dataset of the same business process as detailed in Chap. 2.1 and have selected the tools to analyze the data (Chap. 2.2). Then, we have preprocessed the data (Chap. 3) and identified the main phases of the process, *application creation* and *application validation*, which are detailed in Chap. 4 and Chap. 5. Based on this, we have provided answers to the proposed questions in Chap. 6.

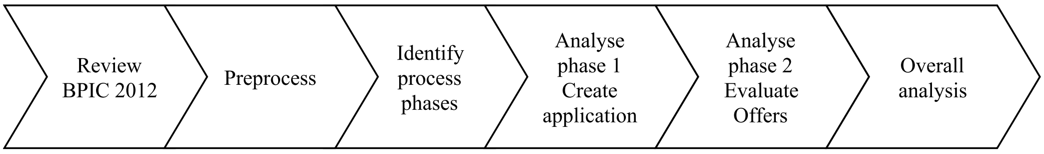

Review BPIC 2012
Preprocess
Identify process phases
Analyse phase 1 Create application
Analyse phase 2 Evaluate Offers
Overall analysis

Figure 2.1: Methodology

2.1 Related Work

It is important to highlight that the data for the analysis is provided by the same institute that previously provided the event log for the BPIC 2012, where six submissions were judged as shown in Table 2.1. Each submission from 2012 made different approaches to the case and achieve different conclusions. As the data taken into study comes from a newer version of the BPIC 2012 process was interesting to study and compare the different reports in order to visualize the methods and conclusions reached in the challenge. From this analysis, we have drawn the following conclusions. Most participants have used the PM tools Disco and Prom, and their use impact in their analysis (e.g. use Heuristic Miner if Prom is used). Nonetheless, some authors have used other tools, such as Excel or databases to carry out a detailed analysis of the event logs. Regarding the analysis objective, a number of researchers aimed at providing a process map, where the process is decomposed into phases. Some relevant aspects that have been studied are (i) activity paths that lead to a successful final state; (ii) study if resources impact on the final outcome of the project and possibility of process automation; and (iii) analyze the performance of the process to understand potential bottlenecks in the process.

Table 2.1: Review previous challenges

Ref.	PM Tool	Analysis Types	Techniques Event Analysis	Process insights
[1]	ProM	Process Discovery	Heuristic Miner Algorithm with manual	Process map
		Performance Analysis	Study over throughput times	Understanding the level of automation of events based on their execution times and obtaining bottlenecks
[4]	Disco, Excel, Cart	Case Level Analysis	Throughput times segmentation	Longer cases have less probability of success

Table 2.1: Review previous challenges

Ref.	PM Tool	Analysis Types	Techniques Event Analysis	Process insights
		Event Level Analysis	Quantity requested segmentation	Some amounts are more likely to be requested
		Resource Level Analysis	Wait vs Work time calculation	Specialists are far more efficient than minor players
		Leveraging Behavioral Data for Work Effort schedule	Specialist vs. Generalist-Driven Work Activities and performance	Find the paths with more chances to be accepted
			Study of the most likely paths to success	
[3]	ProM	Resource Perspective Analysis	Filters and cases analysis	Find of the process automatism
		Control Flow Analysis	Heuristic algorithms used to find the heuristic net and process map	Obtain the resources capabilities
		Process Map Discovery		Process map
[7]	Disco, Excel, ProM	Process Discovery	Obtaining of the process map	Process map

Table 2.1: Review previous challenges

Ref.	PM Tool	Analysis Types	Techniques Event Analysis	Process insights
		Highest Activation Resources	Looks for the resources with more satisfactory applications using disco utilities and a ProM Social Network Plugin	Process segmentation through their execution time
[10]	ProM, Disco	Process general analysis through dotted charts	Time analysis through a dotted chart generated in ProM	Find of the most common cancellation points
			Resources performance through a dotted chart generated in ProM	Resources performance
[15]	ProM	Process study through application states definition	Build of the application states model through ProM plugin	Find of the different states and their characteristics

2.2 Tools Used

For the study three tools were used. These tools have helped us to combine the insights of process mining tools with the power of data analysis tools.

- *Disco [12]*: This process mining tool was really helpful for the understanding of the process. It provides many functionalities which allows the discovery and the filter of the log easing the scope and the process of the data.
- *Excel [16]*: Used to filter and to accommodate the data for its study on Pandas toolkit and to check the results from other applications.

- *Pandas* [2]: This python data analysis toolkit result very useful at the time of analyzing and visualizing the data.

Below are explained with a little more detail the different tools and their use within the project

2.2.1 Disco

Disco is a commercial PM tool. It is a user friendly wich helps with the Process Discovery. It implements a fuzzy algorithm [6] for the model discovery and gives a really helpful overview of the process statistics. It also contains different features for the data filtering which were used for the preprocess of the data. It was chosen among other PM tools because of their ease for the configuration. Another similar PM tools studied were ProM [14] and Celonis [9]. On one hand ProM is an academic tool with much more features, but at the same time with more conflicts which complicate the final study. On the other hand Celonis is another commercial tool, this tool provides more features than disco too, but at the same time the discovery of the process model was not clear enough, for this reason it was discarded.

Disco has been used along all the study phases, although its main function was the process discovery and the preprocess of the data. Besides maps usually have no executable semantics and due to this, deviations cannot be analyzed accurately [9] but this limitation was overcome thanks to Pandas. For the first approach to the process, the data was loaded in the tool. When a XES [5] file is loaded it automatically generates a process model in a few seconds. Once the general process flow was clear some cases were filtered as it is explained in the Sect. 2.1.

All the process model figures in the project were obtained from Disco.

2.2.2 Excel

Excel was used as a support tool for some particular cases understanding. It also ease the obtaining of particular cases figures used for some patterns illustration.

2.2.3 Pandas

The python based Pandas toolkit has been the principal tool used along the study. Once the process flow was clear thanks to Disco, the data filtered was exported to a CSV file and loaded in Pandas. Thanks to its DataFrame structures different patterns could be obtained and sorted for the different process aspects illustration. For the work with Pandas different Jupyter Notebooks [8] were created. This notebooks preformed a playground function. Within this notebooks different sorted were realized In the first place the original DataFrame with the 25.890 cases was iterated for the obtaining of the following information: the different times per phase, the different paths followed, the different work activities performed and the different endings. All of this information was obtained per case. After this thanks to different functions in Pandas for each study the different cases were filtered for the pattern obtaining.

CHAPTER 3

Preprocessing and understanding of the data

First of all, it is needed to understand the process as well as the data supplied. As it was mentioned before, the data consists of two files in eXtensible Event Stream (XES) format [5]. The first one, Application Event log, contains 31,509 application processes with 1,202,267 events. The second file, Offer event log, contains 193,849 events related to the 42,995 offers created, and can been obtained by extracting the Offer type events, explained below, from the first file.

These events belong to one of the next three types:

- (A) Application State: This type of events or activities are used to represent the state in which the application process is. They are useful to follow the process through its different steps.
- (O) Offer State: Offer State events represent the offer possible states, from its creation to its acceptance/declination. These events also contain offer information, such as the amount requested, the number of terms, or its acceptance, among others.
- (W) Workflow State: These events are useful to calculate workflow times, and to understand the work accomplished.

In the previous mentioned .xes files, the information about the events is provided as a list of events ordered by date and case ID, being consecutive all the events from the same application, and ordered by the start date. From the event log, we have selected the following fields: *case ID*, *offerID*, *activity*, *resource*, *start timestamp*, *complete timestamp*, *variant* and *event origin*. The rest of fields have not been used.

3.1 Process Endpoints

Once the information provided by the financial institute has been understood, our next step has been to understand the possible endings of the loan process.

After loading the information into Disco, without any previous preprocessing, fourteen possible endings appear. But, after a detailed analysis, only three cases seemed to represent the normal process flow: A_Pending, O_Cancelled and O_Refused, adding up the 99% of the total cases behavior, and following an understandable workflow. Based on this, we have removed the cases that do not match one of these final events, trying to keep only the normal cases under scope. After filtering the applications that conclude in one of the above listed valid final states (A_Pending, O_Canceled and O_Refused), we have reduced the total number of applications from 31,509 to 31,217. Thus, we have filtered a small number of applications that did not follow the standard process.

In order to understand the normal application endings, we have checked the different process flows with Disco. Our main conclusion is that the last event of an application does not determine the application result. In particular, when there is more than one offer, the last event could be canceled, since one offer is accepted but the rest ones are canceled. Thus, we have identified the events that truly determine the final application result, as shown in Table 3.1.

This understanding will help us to obtain the application ending at the time to process the information with the python Pandas toolkit.

Table 3.1: Final events

Final Event	**Description**
O_Accepted	If one application fires this event it means that one offer has been finally accepted and it is a successful case. After this event the application always goes to A_Pending where it is closed, unless there were more than one offer in the process. In these cases the application has to cancel the rest of the offers before its final end.
A_Canceled	This event represents the final cancellation of an application. Once an application arrives this point it can only be canceled. This end represents a cancellation by the financial institution side, and the reason for cancellation is the reach of a thirty days timeout before receiving an answer from the applicant.
A_Denied	This event also represents an unsuccessful application, but this time it is due an offer rejection by the applicant. As in the previous point, once an application is denied it will set the rest of offers as canceled and will arrive to its final state.

3.2 Event Filtering

After a quick look over the process there is a couple of events that for its relevance inside the process and their function have been excluded from the study:

- *Process Loan Application*: This event is only executed two times in one application from the 31,217 applications left, and it is not clear its function in the process lifecycle since its name does not clarify its function and there are no more cases to compare. For this reason it has been eliminated from our study.
- *Assess Potential Fraud:* This event only appears 354 times in 300 different cases. Even though those cases have a higher probability of being refused than the average, since the result of this assessment is not provided, we have excluded them. Nonetheless, they could be interesting in case the study would be focused in fraud.

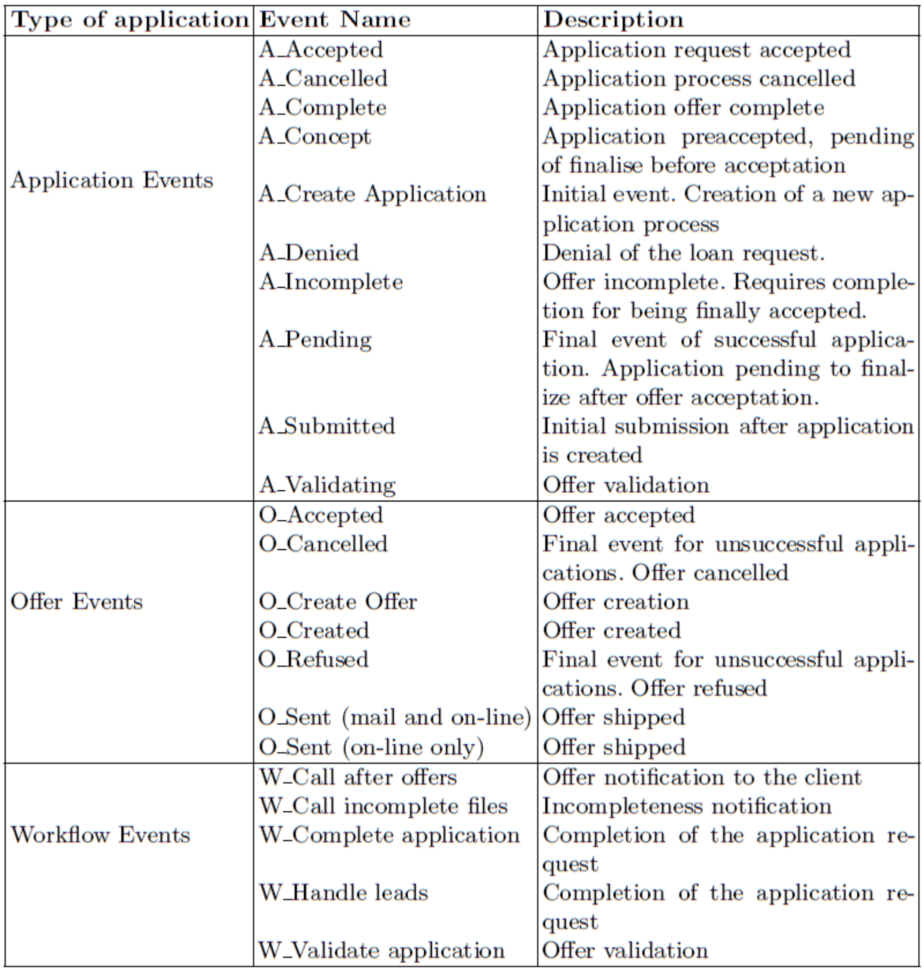
After these filters were applied, from the previous 31,217 cases only 30,917 remain.

3.3 Noise reduction

Apart from the filters explained before, we have also used a feature provided by Disco that allows the elimination of the most uncommon paths. Thanks to this, Pareto principle is applied keeping the 80% of the cases but just the 6% of the different paths reducing the noise of the data by deleting a huge amount of variations and special cases, looking for the normal process flow and simplifying the study. The application of this filter reduces the cases from 30,917 to 25,272. The application of this filter also drives the elimination of the W_Shortened Completion event, which was executed 74 times in 72 cases which represents less than 0,25% of the total cases, so its removal does not bring any consequences to the general process study. After the noise reduction filter, a deeper look inside the process is required in order to get insight of the application process, as detailed in Table 3.2.

Finally, during the analysis of the process to decompose it into phases, another filter has been used. If we exclude the cases where the event A_Complete is not present, we reduce from the 25,272 to 25,190. These are the cases that will be further analyzed.

After this preprocessing stage, a deeper insight of the process has been achieved. For the study of the three proposed questions mentioned in the introduction, a detailed analysis is carried out for each process phase and a final study considers a holistic perspective based on the findings obtained through the phase analysis. In the following points, we discuss the



Type of application	Event Name	Description
Application Events	A_Accepted	Application request accepted
	A_Cancelled	Application process cancelled
	A_Complete	Application offer complete
	A_Concept	Application preaccepted, pending of finalise before acceptation
	A_Create Application	Initial event. Creation of a new application process
	A_Denied	Denial of the loan request.
	A_Incomplete	Offer incomplete. Requires completion for being finally accepted.
	A_Pending	Final event of successful application. Application pending to finalize after offer acceptation.
	A_Submitted	Initial submission after application is created
	A_Validating	Offer validation
Offer Events	O_Accepted	Offer accepted
	O_Cancelled	Final event for unsuccessful applications. Offer cancelled
	O_Create Offer	Offer creation
	O_Created	Offer created
	O_Refused	Final event for unsuccessful applications. Offer refused
	O_Sent (mail and on-line)	Offer shipped
	O_Sent (on-line only)	Offer shipped
Workflow Events	W_Call after offers	Offer notification to the client
	W_Call incomplete files	Incompleteness notification
	W_Complete application	Completion of the application request
	W_Handle leads	Completion of the application request
	W_Validate application	Offer validation

Table 3.2: Events understanding

proposed questions and our main hypotheses and approach. The company is interested in understanding:

- The different times spent along the process. In particular, the company is interested in knowing: the time spent in the company's system waiting to be processed by an user, and the time spent waiting on input from the applicant. The main difference between these two times is the responsible for the next step in the process. Two cases are considered: an internal user from the company, in this case the waiting time will be classified as waiting time inside the company's system to be processed, or, on the other hand could be the applicant, who for some reasons could be delaying the answer causing another waiting time for the final application resolution. It is also interesting to see the difference in times between the different possible paths and try to analyze why there is such a great variation in times between different applications.

- The influence on the frequency of incompleteness to the final outcome. The hypothesis here is that if applicants are confronted with more requests for completion, they are more likely not to accept the final offer. This point will be clarified in the second phase of the process, where offers are processed and validated, and the requests for completion are demanded.

- How many customers ask for more than one offer. Keeping in mind if the different offers are asked for in a single conversation or in multiple conversations. For this study, both phases of the process should be analyzed. Even though most of the offers are created in the first process phase, offers belonging to multiple conversations are created in the second phase. Thus, we should draw out our conclusions for this topic in the final overall process assessment, based on the compiled findings when analyzing every process phase.

In order to obtain these different times and to simplify the analysis, the loan process has been divided into two phases according to the different nature of the pursued objectives.

The first phase, from A_Create Application to A_Complete is where the application is created, and the majority of the offers are defined. This phase has been divided into two different subphases. The first subphase consists of the application creation, where the information about the applicant is collected and submitted. It encompasses the events from A_Create Application to A_Accepted, being A_Accepted the starting event of the next subphase. The second subphase happens between A_Accepted and A_Complete. In this second phase the majority of the offers are created, it is said the majority because in the cases where the application has more than one offer, this happens in the second phase.

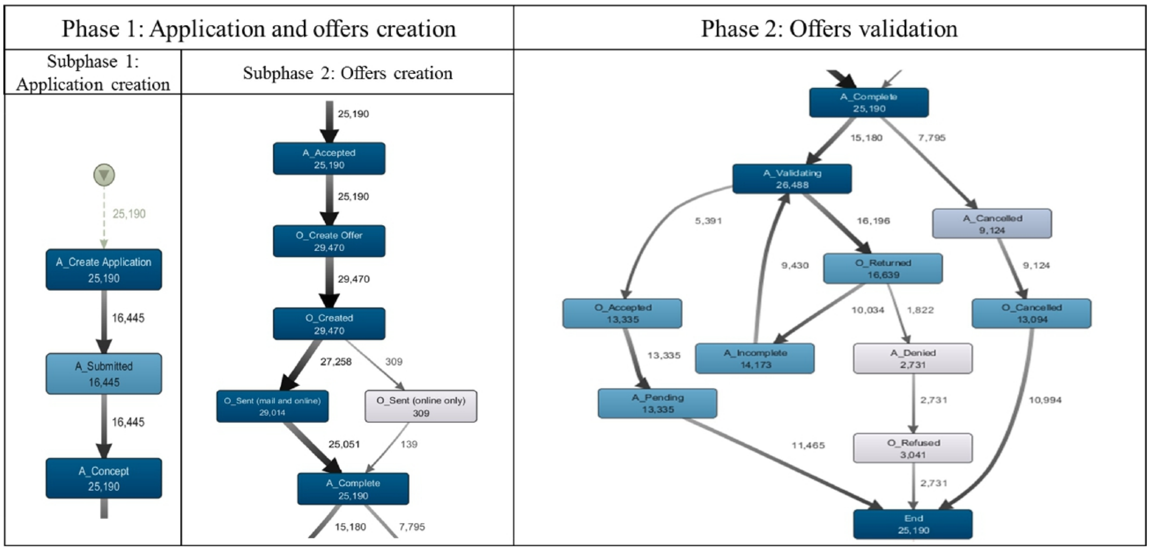


Figure 3.1: Phase division

The second phase consists in the validation of the application and the creation of multiple offers through multiple conversations. As a result, applications not providing the required information in time will be canceled or validated otherwise. The phase includes the events between A_Complete and the end of the process. This phase is the most complex because there is not one single final event, it intersects with the creation of more offers, in the cases where more than one offer is created in multiple conversations. In addition, this phase has many possible loops. In the Fig. 3.1 it is shown the process map division per phases and subphases with the intention to clarify the division and the process flow.

In order to analyze the waiting times, the events have been classified into two types: those that require and input from the applicant, and those that need an input from an internal user. This classification is based on the nature of the events as follows: there are only two events that require an input from an applicant: A_Cancelled and A_Validating. A_Cancelled is fired when the 30 days timeout expires, which is the maximum waiting time for an applicant response. A_Validating state fires after the event A_Complete within the event W_Validate application. This can be observed in Fig. 3.2, extracted from Disco [12]. Red paths represent the longest paths in time. As it will be discussed later in Chap. 5, they represent the most relevant times in the overall process duration.

With this in mind, the first thing to be done is to separate the application process into different independent phases and carry out an exhaustive analysis per phase. In addition, we also aim to identify relationships among these variables or other possible factors that could be decisive in the final result of the process or in an extra delay in the process flow.

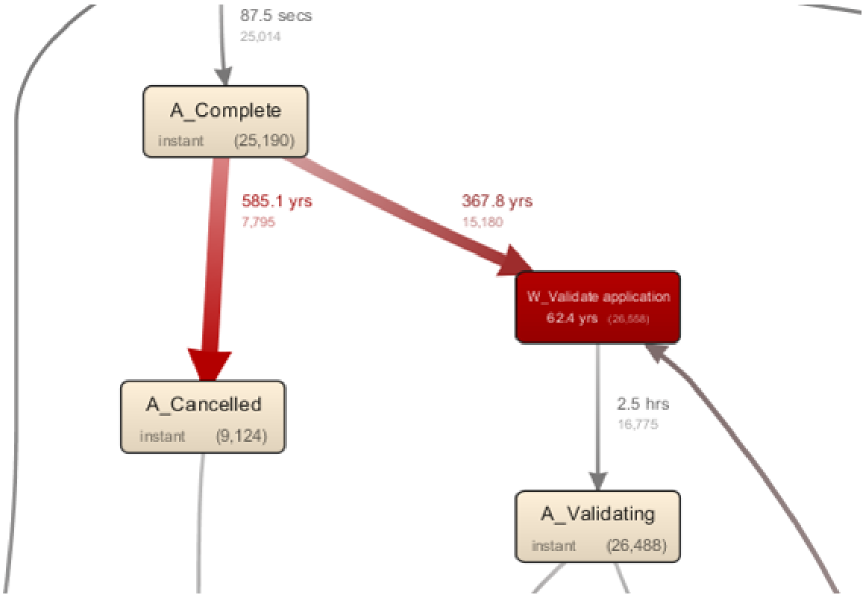


Figure 3.2: Events waiting for an external input

Events have been separated as follows. First of all, the original dataset has been converted to a CSV file through Disco [12]. This conversion is done for enabling the data load on Pandas [2] toolkit. Next the CSV file has been loaded in Pandas [2] as a DataFrame, which is a 2-dimensional labeled data structure with columns of potentially different types. This DataFrame contains the same information than the original XES file. The Dataframe has been iterated to create one Dataframe per phase and subphase to simplify the analysis. This transformation aims at grouping case events for each phase and subphase in one row, and has been performed by iterating the original dataframe and including the following information: start and complete timestamp of each phase, phase duration, phase working time, phase waiting time for an external input, phase waiting time for an internal input, path followed by the case, workflow type events within the case, and the final application result, codified as “0” for accepted cases, “1” for canceled cases, and “2” for denied cases.

In this munging process, working time is calculated as the adding up of the workflow time events duration. Waiting times are calculated as the time difference between the start from one event and the completion of the previous one, excluding from this calculation the time spent simultaneously than a workflow type event. The difference for the external or internal classification within the waiting times is done in function of the event starting as it was explained above. The phase duration calculation is as simple as the difference between the starting date and the completion date. For the paths calculation every possible event was codified with an alphabetic character. While a case was iterated, every event was codified with its appropriate character, and added to a characters chain which represents the path as a string [2] element. The workflow type events were registered in the same way. Finally the different endings were obtained by checking the decisive events that every case must contain as it was explained in the Chap. 3.1.

Thus, the use of Pandas data structured have provided us a very powerful data analysis tool to understand and query the event information. Moreover, the proposed structure for dataframes has proven to be useful for their analysis.

CHAPTER 4

Phase Application Creation

In this phase the application and offers are created. It has a clear and easy understandable path. Despite of this, for a deeper explanation, it has been divided into two subphases which fully clarify the process. This separation has be based on functionality and automation perspectives. A first subphase has been identified as the creation of the application itself, while a second subphase collects the offer creation. The first subphase is comprised by the following events: A_Create Offer, A_Submitted and A_Concept, while the second one comprises the events between A_Accepted and A_Complete as it was shown in Fig. 3.1. In the reminder of this chapter we analyse first both phases separately and provide then a general conclusion for the whole phase.

4.1 Subphase 1.1: Create application

This subphase is the shortest and simplest one. It consists of four possible events, three "A" type events: A_Create Application, A_Submitted, and A_Concept; and one "W" type: W_Handle leads. It is interesting to highlight that these steps are fully automated as they can be executed by user_1, which after observing its working ratios it could be instead an autonomous system. Fig. 4.1 shows the possible state paths followed in this subphase

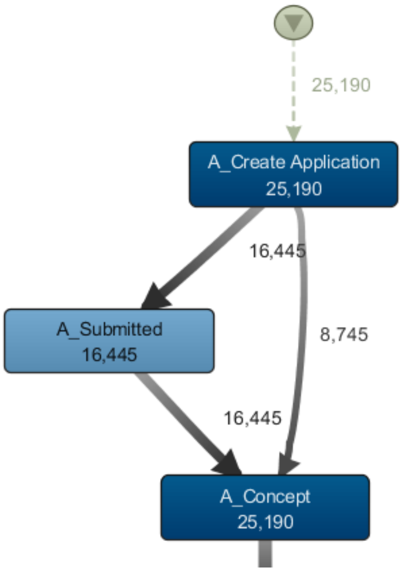
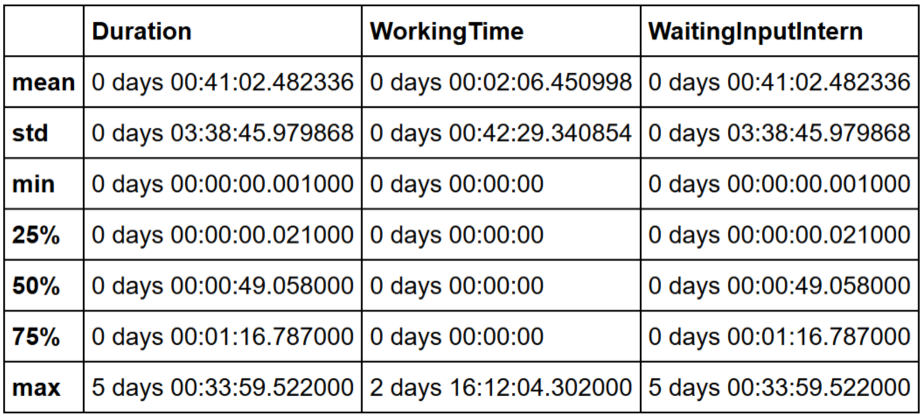


Figure 4.1: Path diagram of subphase 1.1

obtained from Disco [12].

In the state path diagram we can see that this subphase is over when A_Concept ends. But the decision of shattering the complete phase in that point was taken after observing the pattern followed by the W type activity W_Complete application. This activity starts at the same time than A_Concept, and when its working times are not equal to zero it follows the application until A_Complete is reached. This is the final point for the overall Application Creation phase. After realizing this situation it is deducted that, the first states in this phase are executed by an autonomous system. When it arrives to the end of the application creation, it automatically launches both events A_Concept and W_Complete Application. In A_Concept is registered the final data, and in W_Complete application the control is handled to a human resource for the final acceptance and the offer creation. This offer creation is registered in the next subphase. Following with this subphase, as it is shown in Fig. 4.1 there are two different paths for the state flow. The only difference between these paths is whether they go through the activity A_Submitted or not. This difference may be caused in function of the applicant previous relation with the institute. This hyphotesis is based on the following evidence: if the company has already registered and evaluated the applicant data, it seems plausible that the internal users just read the results of the previous submission in spite of recalculating it all again by submitting him/her to further investigation. Apart from this difference, there is an additional variation related with the W type events. This difference is the fire of W_Handle leads in a few cases. These cases have always fired A_Submitted before. Our hypothesis is that this variation corresponds to



	Duration	WorkingTime	WaitingInputIntern
mean	0 days 00:41:02.482336	0 days 00:02:06.450998	0 days 00:41:02.482336
std	0 days 03:38:45.979868	0 days 00:42:29.340854	0 days 03:38:45.979868
min	0 days 00:00:00.001000	0 days 00:00:00	0 days 00:00:00.001000
25%	0 days 00:00:00.021000	0 days 00:00:00	0 days 00:00:00.021000
50%	0 days 00:00:49.058000	0 days 00:00:00	0 days 00:00:49.058000
75%	0 days 00:01:16.787000	0 days 00:00:00	0 days 00:01:16.787000
max	5 days 00:33:59.522000	2 days 16:12:04.302000	5 days 00:33:59.522000

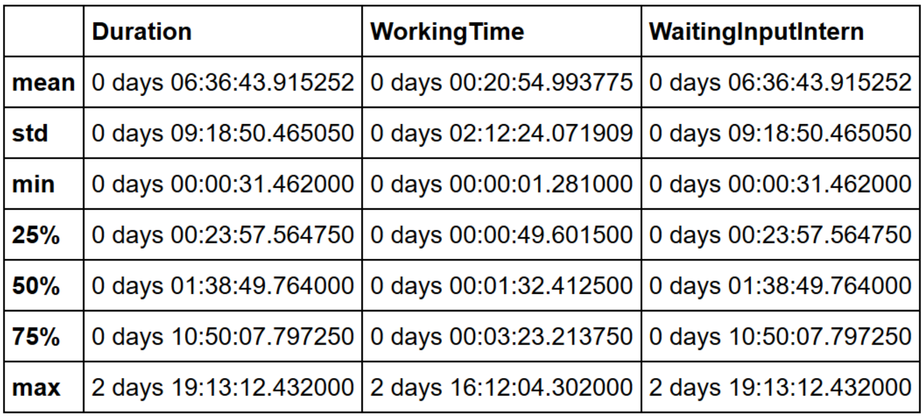
Table 4.1: Times obtained in Subphase 1 Create Application

a special process for loans with high business value, such as exclusive clients or loan with a high requested quantity.

In the time aspect, in this phase there are no activities requiring an input from the applicant so there are not waiting times due an external input from the applicant. Working and waiting times for the global actuation in this subphase are shown below in Table 4.1.

As it can be appreciated in Table 4.1, this subphase has very short times in general. The majority of cases has no workflow type events.This absence makes global working times equal to zero in a 90% of the cases . In addition, except for the cases which indeed have working events, waiting times are also very low. It is important to appreciate that the longest waiting times between application type events are lower than a minute and happened before A_Submitted. This makes us suppose that A_Submitted although is automatized requires longer processing times. In the next Table 4.2 are shown the times comparison for those cases with working time events. These cases are the ones with the biggest processing times in the subphase as it was expected, being those that need a human interaction (e.g. manual tasks) represented by the event W_Handle leads.

In conclusion, the possible paths are pretty simple and understandable as it was pointed out above. Our conclusions have been supported by evidences for this path division, and the time spent in this phase in relation to the overall process duration is meaningless. It has no major issues due to its high level of automation. This is supported by the fact that almost all the activities are realized by user_1, which could be possibly an automatic system after studying its working ratios. It is responsible of 63287 events wasting just 2 min and 32 seconds in it. It is important to notice that any application is ending during this phase in contrast with the data provided in BPIC 2012. Thus, this represents an improvement within the process in the recent years.



	Duration	WorkingTime	WaitingInputIntern
mean	0 days 06:36:43.915252	0 days 00:20:54.993775	0 days 06:36:43.915252
std	0 days 09:18:50.465050	0 days 02:12:24.071909	0 days 09:18:50.465050
min	0 days 00:00:31.462000	0 days 00:00:01.281000	0 days 00:00:31.462000
25%	0 days 00:23:57.564750	0 days 00:00:49.601500	0 days 00:23:57.564750
50%	0 days 01:38:49.764000	0 days 00:01:32.412500	0 days 01:38:49.764000
75%	0 days 10:50:07.797250	0 days 00:03:23.213750	0 days 10:50:07.797250
max	2 days 19:13:12.432000	2 days 16:12:04.302000	2 days 19:13:12.432000

Table 4.2: Times obtained for the manual cases in Subphase 1 Create Applications

4.2 Subphase 1.2: Create Offer

This subphase encompasses those events reached between A_Accepted and A_Complete, which are entirely from "O" type. This fact seems reasonable since the functionality of this subphase is the creation and submission of the offers. It contains nine different events, but, there are not any cases with all of them in the same path. Fig. 4.2 shows the state flow in this subphase.

The nine possible events are the starter and the final, A_Accepted and A_Complete respectively. Several "O" type events are fired after the A_Accepted event. In the first place O_Create Offer event comes, which contains all the offer characteristics. It is followed by the event O_Created, which assigns an OfferID. In case there are issues with the offer, the O_Cancelled is executed next, before its submission to the applicant. Then several event paths can be followed. The most frequent path is that O_Sent (mail & online) is submitted, although in some few cases, O_Sent (online only) is the one submitted. Next, A_Complete is reached unless there is a loop for more offers creation. This event is the final state for this phase.

In this phase two workflow type events can be executed: W_Call after offers and W_Complete application. W_Call after offer is typically fired after the submission of an offer. This event is used to represent the notice of the offer to the applicant. Regarding W_Complete Application, it represents the work realized from A_Concept to A_Complete. This event is used to finish the process when A_Complete is reached. There are a few cases where the application never enters in the W_Complete application process. It is supposed that these cases with no W type events, as well as the cases with no working time defined, are caused by an incorrect application use. In Fig. 4.3 below it is shown one example for

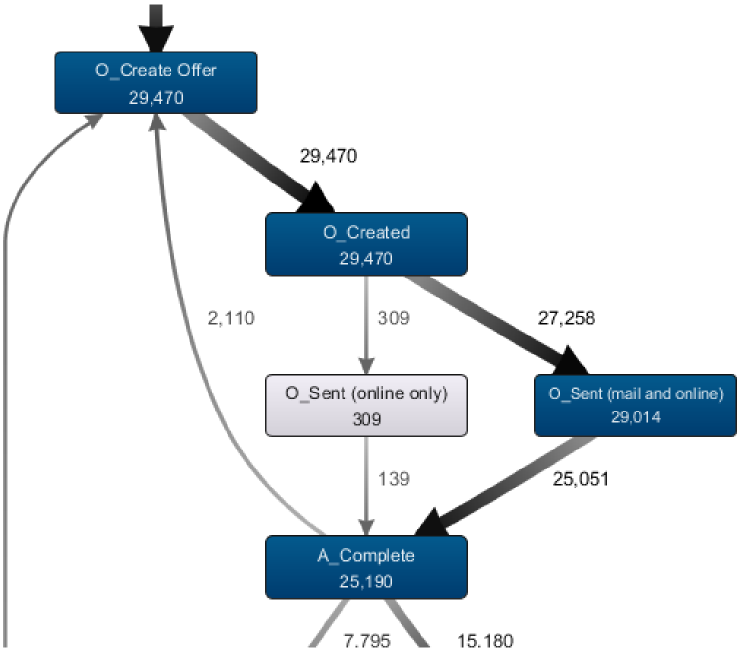
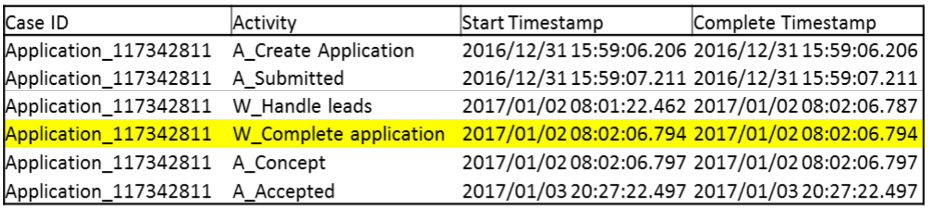


Figure 4.2: Path diagram of sub-phase 1.2 Create Offer



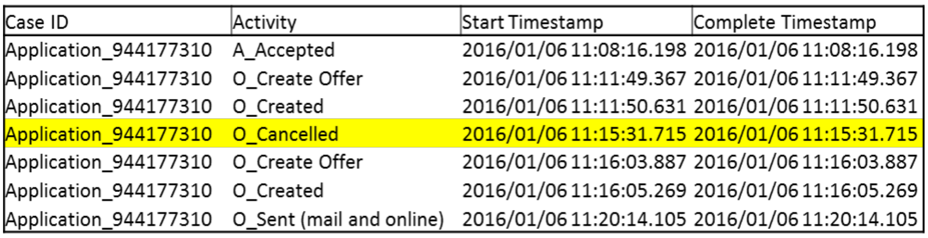
Case ID	Activity	Start Timestamp	Complete Timestamp
Application_117342811	A_Create Application	2016/12/31 15:59:06.206	2016/12/31 15:59:06.206
Application_117342811	A_Submitted	2016/12/31 15:59:07.211	2016/12/31 15:59:07.211
Application_117342811	W_Handle leads	2017/01/02 08:01:22.462	2017/01/02 08:02:06.787
Application_117342811	W_Complete application	2017/01/02 08:02:06.794	2017/01/02 08:02:06.794
Application_117342811	A_Concept	2017/01/02 08:02:06.797	2017/01/02 08:02:06.797
Application_117342811	A_Accepted	2017/01/03 20:27:22.497	2017/01/03 20:27:22.497

Figure 4.3: Example of case without working time

one of these cases. This wrong pattern could be due to a user who does not register the work amount correctly, or because the internal users delays to registering the completion of the task.

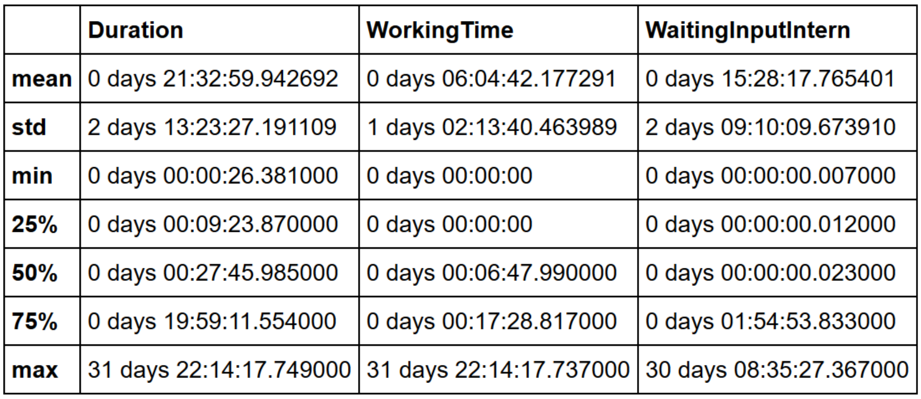
In this phase, as in the previous one, there are not many path variations, and the paths are easy to understand, as previously shown in Fig. 4.2. Note that O_Cancelled, although it is mentioned above, it does not appear in the flow. This is because it happens in a small number of cases (109 cases), as the one shown in Fig. 4.4.

Finally, five different paths can be appreciated. The main reason for this number of variations is the existence of loops, as it was mentioned before. These loops are the consequence of the creation of different offers for the same application. In this phase it is recorded a maximum of three offer creations. Despite of it, it does not represent the total amount of offers created, because in the last phase there are also some offers created within



Case ID	Activity	Start Timestamp	Complete Timestamp
Application_944177310	A_Accepted	2016/01/06 11:08:16.198	2016/01/06 11:08:16.198
Application_944177310	O_Create Offer	2016/01/06 11:11:49.367	2016/01/06 11:11:49.367
Application_944177310	O_Created	2016/01/06 11:11:50.631	2016/01/06 11:11:50.631
Application_944177310	O_Cancelled	2016/01/06 11:15:31.715	2016/01/06 11:15:31.715
Application_944177310	O_Create Offer	2016/01/06 11:16:03.887	2016/01/06 11:16:03.887
Application_944177310	O_Created	2016/01/06 11:16:05.269	2016/01/06 11:16:05.269
Application_944177310	O_Sent (mail and online)	2016/01/06 11:20:14.105	2016/01/06 11:20:14.105

Figure 4.4: Example of a case with the O_Cancelled event

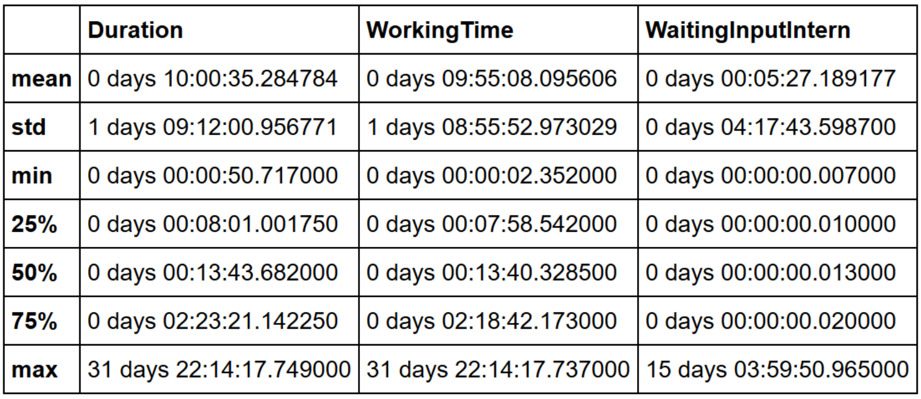


	Duration	WorkingTime	WaitingInputIntern
mean	0 days 21:32:59.942692	0 days 06:04:42.177291	0 days 15:28:17.765401
std	2 days 13:23:27.191109	1 days 02:13:40.463989	2 days 09:10:09.673910
min	0 days 00:00:26.381000	0 days 00:00:00	0 days 00:00:00.007000
25%	0 days 00:09:23.870000	0 days 00:00:00	0 days 00:00:00.012000
50%	0 days 00:27:45.985000	0 days 00:06:47.990000	0 days 00:00:00.023000
75%	0 days 19:59:11.554000	0 days 00:17:28.817000	0 days 01:54:53.833000
max	31 days 22:14:17.749000	31 days 22:14:17.737000	30 days 08:35:27.367000

Table 4.3: Times obtained in Subphase 2 Create Offer

the validation process. These offers still represent the majority of them (around 90% of the above). Apart from the loop existence, there are two more variation sources, one is the multichannelity for reception submission, and the other is the offer cancellation possibility mentioned above. Multichannelity creates different variations since offers can be submitted on-line or combining e-mail and on-line. Nevertheless, this does not affect to the process understanding. The other source of variability is the offer cancellation. It is thought that the main apparent reason to this, it is the possibility of finding an error after creating an offer. This is supported by the fact that all paths with offer cancellation include more than one offer.

For the times aspect, in this subphase the global times start to grow in comparison to the previous one. This increment is due to the higher number of not automatized manual tasks. This increment of manual tasks drives to an increment in the working times. Despite of this increment, there are still many cases with no working time registered, although they have workflow type events. Within this subphase, working times are mainly represented by the W_Complete application event. Regarding waiting times, as there are no events needing an input from the applicant, it are due to waiting times inside the application waiting for a user to continue with the process. The times for this phase are shown in Table 4.3.



	Duration	WorkingTime	WaitingInputIntern
mean	0 days 10:00:35.284784	0 days 09:55:08.095606	0 days 00:05:27.189177
std	1 days 09:12:00.956771	1 days 08:55:52.973029	0 days 04:17:43.598700
min	0 days 00:00:50.717000	0 days 00:00:02.352000	0 days 00:00:00.007000
25%	0 days 00:08:01.001750	0 days 00:07:58.542000	0 days 00:00:00.010000
50%	0 days 00:13:43.682000	0 days 00:13:40.328500	0 days 00:00:00.013000
75%	0 days 02:23:21.142250	0 days 02:18:42.173000	0 days 00:00:00.020000
max	31 days 22:14:17.749000	31 days 22:14:17.737000	15 days 03:59:50.965000

Table 4.4: Times in subphase 2 with working time equal zero cases filtered

It is important to explain a few details about the results presented in Table 4.3. First, we should highlight that the 39% of the cases has working times equal to zero in this phase. Several hypotheses could justify this fact. One hypothesis is that internal users of the financial institution do not record properly the timing of the tasks and register the work as done after processing the offers. In case this happens, these use cases would be useless for the waiting time analysis of in this subphase, because these events main function is to represent the amount of work delivered. The second hypothesis is that there is an automatic system developing the work. In order to analyze this hypothesis, we have analyzed the resources patterns. The evidences have not supported this second hypothesis since there were 113 different users performing these tasks. Thus, we have rejected the second hypothesis since the number of automatic users seems too high. As a result, our hypothesis is that there is an issue with timing registration and offer creation tasks require human work. The evidences that support this conclusion are shown in Table 4.4, where applications with worktimes equal to zero have been filtered. In Table 4.4 we can appreciate that the main weight for the global times is in the worktimes, discarding a time wasting while waiting for process an application.

In Table 4.4 it is appreciated that the duration for a high number of the cases is extremely low, just a few minutes, despite of the existence of some exceptional cases with higher times. It is thought that these higher times are caused by problems with the offer calculation system, or by the need to introduce more variables implying a manual solution. In Table 4.4 it can also be appreciated that once the cases with no registered working time are deleted, the total working time represents almost the overall phase duration. Our most plausible explanation is that the overall phase, from the creation of the application to the submission of an offer, is followed by short and quick stops in some exceptional cases. Finally, Fig. 4.5, and Fig. 4.6 show the relation between the different paths and the phase duration. It is

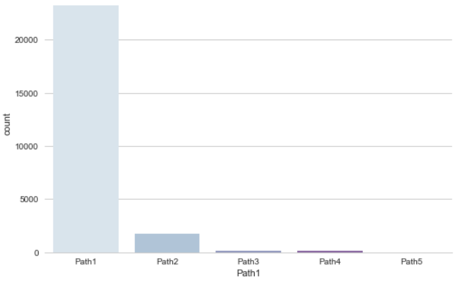


Figure 4.5: Subphase 2 Create Offer Amount of cases per path

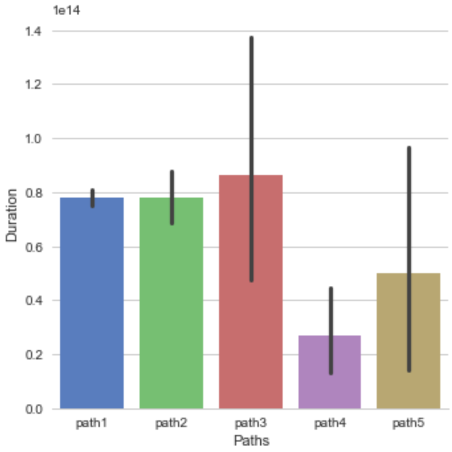


Figure 4.6: Relation between paths and duration

observed that although may exist some loops inside the phase these cases can be considered outliners and have no major impact in the phase duration.

In Fig. 4.5 and Fig. 4.6 the different paths have been numbered, and the different numbers refer to the next path explanation:

- Path 1 is the most common path. It refers to the cases where there is only one offer created and submitted via mail and online.
- Path 2 refers to the cases where two offers are created and submitted via mail and online.
- Path 3 refers to those cases where two offers were created but one of them was canceled before its submission so the applicant receives just one of them.
- Path 4 encompasses the cases with one offer submitted just via online. As it is normal because of the possibility of submitting only online, these are the shortest cases in time.

- Path 5 cases are those with three offers created and submitted via mail and online.

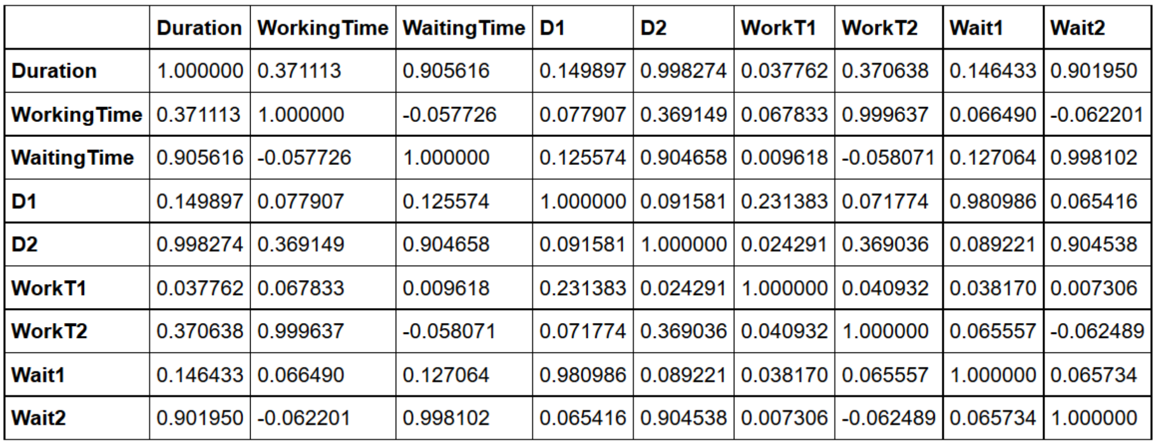
But as we can appreciate, only two cases have an interesting amount of cases, being the rest special cases for the application process.

4.3 Overall Application Creation Phase conclusion

Finally, for the complete phase after joining the two different subphases we obtained the following Pearson correlation coefficient between the times shown in Table 4.5. In this Table 4.5, it is appreciated that the subphase 2, Create Offer, is the one with a higher weight in the overall phase duration. There is a correlation close to one between this subphase 2 duration and the overall phase time. This was expected because it has a higher manual work amount. Despite of this, the main correlation its found between the global phase duration and the waiting times, which is normal with the high number of cases without registered working times.

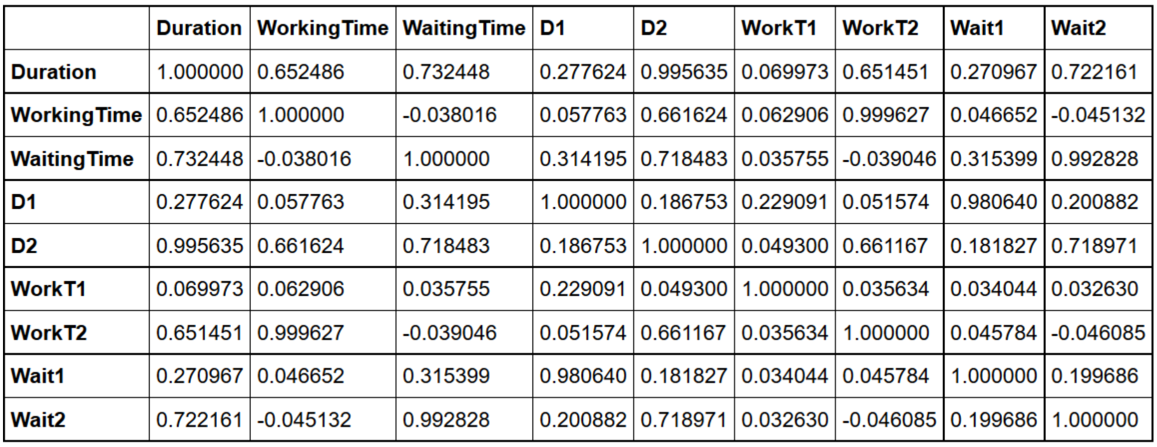
In Table 4.6 the cases with total working time equal to zero have been deleted to understand better the process. In the final cases the correlation between waiting times and the duration is reduced at the same time than the correlation with the working times growth. This balance the correlation between the waiting and the working times, although the waiting ones keep leading being the main weight in the overall duration.

Regarding path variation, as shown in Fig. 4.7, the number of paths is only relevant for Paths 1-1 and 2-1. This was expected based on the amount of cases per path in the second subphase.



	Duration	**WorkingTime**	**WaitingTime**	**D1**	**D2**	**WorkT1**	**WorkT2**	**Wait1**	**Wait2**
Duration	1.000000	0.371113	0.905616	0.149897	0.998274	0.037762	0.370638	0.146433	0.901950
WorkingTime	0.371113	1.000000	-0.057726	0.077907	0.369149	0.067833	0.999637	0.066490	-0.062201
WaitingTime	0.905616	-0.057726	1.000000	0.125574	0.904658	0.009618	-0.058071	0.127064	0.998102
D1	0.149897	0.077907	0.125574	1.000000	0.091581	0.231383	0.071774	0.980986	0.065416
D2	0.998274	0.369149	0.904658	0.091581	1.000000	0.024291	0.369036	0.089221	0.904538
WorkT1	0.037762	0.067833	0.009618	0.231383	0.024291	1.000000	0.040932	0.038170	0.007306
WorkT2	0.370638	0.999637	-0.058071	0.071774	0.369036	0.040932	1.000000	0.065557	-0.062489
Wait1	0.146433	0.066490	0.127064	0.980986	0.089221	0.038170	0.065557	1.000000	0.065734
Wait2	0.901950	-0.062201	0.998102	0.065416	0.904538	0.007306	-0.062489	0.065734	1.000000

Table 4.5: Times correlation within phase 1



	Duration	WorkingTime	WaitingTime	D1	D2	WorkT1	WorkT2	Wait1	Wait2
Duration	1.000000	0.652486	0.732448	0.277624	0.995635	0.069973	0.651451	0.270967	0.722161
WorkingTime	0.652486	1.000000	-0.038016	0.057763	0.661624	0.062906	0.999627	0.046652	-0.045132
WaitingTime	0.732448	-0.038016	1.000000	0.314195	0.718483	0.035755	-0.039046	0.315399	0.992828
D1	0.277624	0.057763	0.314195	1.000000	0.186753	0.229091	0.051574	0.980640	0.200882
D2	0.995635	0.661624	0.718483	0.186753	1.000000	0.049300	0.661167	0.181827	0.718971
WorkT1	0.069973	0.062906	0.035755	0.229091	0.049300	1.000000	0.035634	0.034044	0.032630
WorkT2	0.651451	0.999627	-0.039046	0.051574	0.661167	0.035634	1.000000	0.045784	-0.046085
Wait1	0.270967	0.046652	0.315399	0.980640	0.181827	0.034044	0.045784	1.000000	0.199686
Wait2	0.722161	-0.045132	0.992828	0.200882	0.718971	0.032630	-0.046085	0.199686	1.000000

Table 4.6: Times correlation without cases with no working time

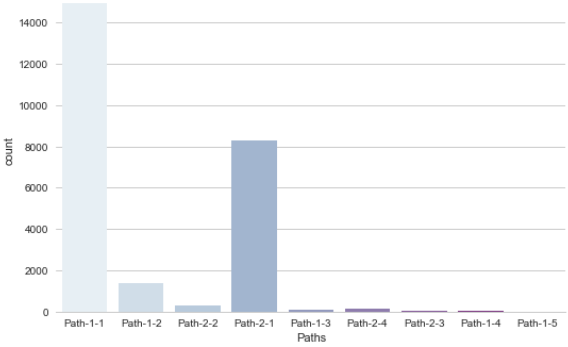


Figure 4.7: Cases amount per path in Phase 1

Fig. 4.8 shows the times variations between the most relevant paths, after deleting the paths with the lowest amounts of cases. In these figures (Fig. 4.7 and Fig. 4.8) the legend for the paths goes as follows: the first number refer to the path in the first subphase: Application Creation. In this subphase there were just two different paths: the first one where the application went trough A_Submitted which is represented with the number 1, and the second where the application never went through A_Submitted, referred with a 2. The second number refers to the second subphase path, and it has the same relation between number and path than the one explained in Fig. 4.5.

There is a clearer difference between the cases than in the first subphase. Cases not including the A_Submitted event have shorter times than the rest of cases. After looking for the source of this behavior, it was found that these shorter cases are always processed from the beginning to the end of the phase by just one resource. This pattern is shown in the Table 4.9 below with an example. This behavior allows avoiding the internal waiting times from the ending of one resource to the start of the next one. Thus, in these cases the

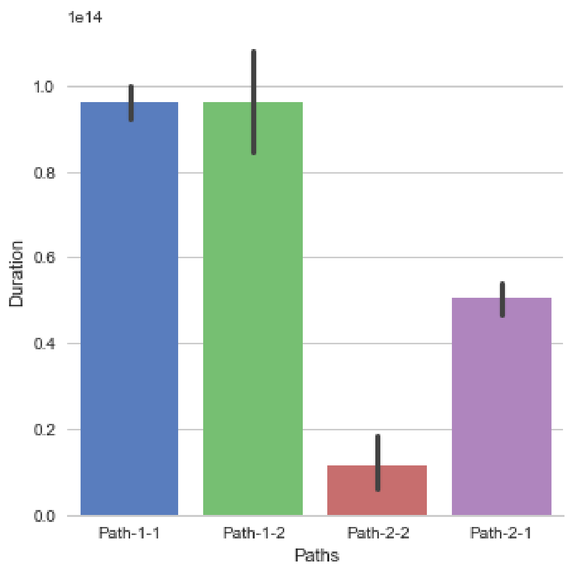
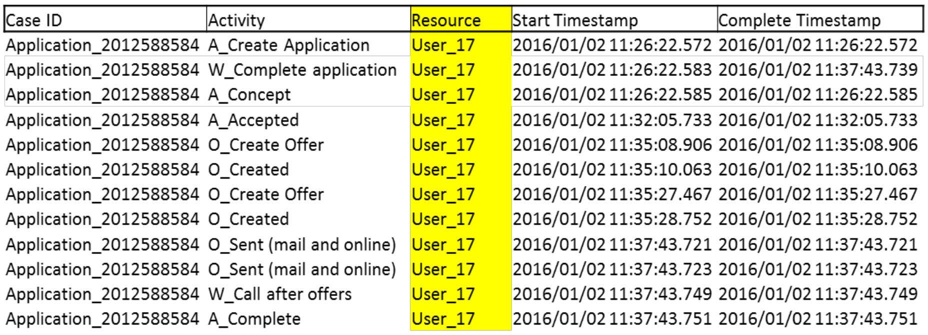


Figure 4.8: Relation between duration and path for the most relevant paths in Phase 1



Case ID	Activity	Resource	Start Timestamp	Complete Timestamp
Application_2012588584	A_Create Application	User_17	2016/01/02 11:26:22.572	2016/01/02 11:26:22.572
Application_2012588584	W_Complete application	User_17	2016/01/02 11:26:22.583	2016/01/02 11:37:43.739
Application_2012588584	A_Concept	User_17	2016/01/02 11:26:22.585	2016/01/02 11:26:22.585
Application_2012588584	A_Accepted	User_17	2016/01/02 11:32:05.733	2016/01/02 11:32:05.733
Application_2012588584	O_Create Offer	User_17	2016/01/02 11:35:08.906	2016/01/02 11:35:08.906
Application_2012588584	O_Created	User_17	2016/01/02 11:35:10.063	2016/01/02 11:35:10.063
Application_2012588584	O_Create Offer	User_17	2016/01/02 11:35:27.467	2016/01/02 11:35:27.467
Application_2012588584	O_Created	User_17	2016/01/02 11:35:28.752	2016/01/02 11:35:28.752
Application_2012588584	O_Sent (mail and online)	User_17	2016/01/02 11:37:43.721	2016/01/02 11:37:43.721
Application_2012588584	O_Sent (mail and online)	User_17	2016/01/02 11:37:43.723	2016/01/02 11:37:43.723
Application_2012588584	W_Call after offers	User_17	2016/01/02 11:37:43.749	2016/01/02 11:37:43.749
Application_2012588584	A_Complete	User_17	2016/01/02 11:37:43.751	2016/01/02 11:37:43.751

Figure 4.9: Example of cases with just one resource executing completely the Phase 1

automatic system (user_1) is not used. An additional human resource study could check if the use of human resources for these events is useful. Our conclusion is that the process can be optimized if the same user process the full application, but we should also analyze if this could bring as a side effect delays or the appearance of bottlenecks in this or other processes.

Thus, as we have commented before, our main conclusion is that the best performance happens when the same human resource handles all the work from the beginning. As it was pointed before this could be generating issues in other phases or processes. It also does not look useful since there is an autonomous system available for developing those tasks. It is also important to highlight that there is a huge amount of cases without any worktime registered, complicating this way the times study. This fact could indicate a wrong use of the system by the financial institution which should be fixed. In addition, the majority of offers are created in this phase and a single conversation is used for creating multiple offers.

CHAPTER 5

Phase Application Validation

In this phase is where the different offers are validated. For this purpose, after the offer is submitted to the client, a response and some extra information is expected. In case that any response is received, applications are ended after a thirty days timeout.

When a response arrives, it is validated. Then it can take two different paths. The first one drives to the final offer acceptation. The application goes through O_Accepted event first, and in the cases where no more offers were created it ends in the A_Pending event. If more cases were created, it would finally go to the O_Cancelled event were the rest of the offers are deleted, and it will end in this event.

Regarding the second path, if the validation is not successful, it could be for two different reasons: the first reason is because the applicant has refused the offer, and the second reason is the lack of some files or information. In the refused cases the offer goes through the O_Returned event, where it chooses the final path for a refused offer. This path consists in two different events, first the A_Denied event, where the application is finally established as denied, and O_Refused where the offer is established as refused. At this point the application process finally ends. In the cases where more information is requested, this information is requested to the applicant. Then the offer goes through a path which drives the application to the beginning of the validation phase, where the response and the files are checked again.

This path contains three events which flow as follows. In the first place the application fires the O_Returned event, then it starts an "W" type event, W_Call incomplete files where it is supposed that the applicant is informed about the need of additional information, and the application passes to the state A_Incomplete. This path is the main cause for the loop existence in this phase, with cases that register even five loops through the validation events.

In this phase, as it was pointed in Chap. 4, some offers are created. The offers creation registered in this phase answers to two different reasons which can be generated in two points. The first point is before the validation events start. It is not clear the reason for the creation of these offers once the A_Complete is reached, but they seem to be done after the previous round and by a different resource. This lead us to think that these offers are created in different conversations than the offers from the Phase 1. The second point refers to the creation of an offer in the middle of the validation process. This could be due to a negotiation within the validation of a previous offer. This drives the creation of a new offer trying to satisfy the client request. These offers may be generated in multiple conversations.

This phase, unlike the previous one, is not as simple for the paths study. The main reason for this is the huge variations amount (76) within the phase. In order to improve our understanding, we have analyzed each type of application (cancelled, accepted and denied applications), as described in the following sections.

5.1 Canceled Applications

First of all, we analyze paths referring to the canceled applications. 93% of canceled applications do never receive an applicant response. The remaining canceled applications do not follow this pattern. For our study, we filter the applications including A_Cancelled as explained in Table 3.1. With this filter, we reduce the complexity from 76 to 19 different variations. From these variations it is appreciated that the simplest one responds to the offers with no response from the applicant. In these cases the application dies after the timeout is reached. This path has the 85% of the cases, in three variants depending on the number of offers previously sent. These variations have been grouped in the graphics under the name "C1". For these cases there is a similar duration which main weight comes from the 30 days waiting for a response.

There is also another similar path with a relevant amount of cases. This happens when there is an offer creation at the beginning of this phase but canceled after the timeout is reached. These cases have similar waiting times, although they exhibit a longer duration, which is normal due to the creation offer process. These cases have been grouped with the

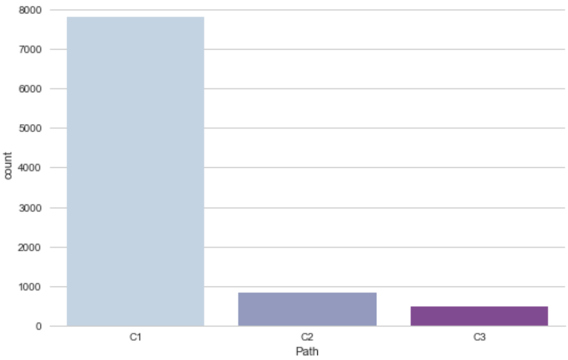


Figure 5.1: Amount of cases per path

name C2. The rest of the different cases refer to the applications that, after being validated, are returned and finally is canceled The main reason for this pattern seems to be that the application does not match the company requirements. These cases are grouped under the "C3" name.

With the cases grouped in three final paths: those canceled after the completion due to the timeout; those started with an offer creation before a cancellation due to the timeout; and those validated before the final cancellation. Times and the amount of cases per path are shown in Fig. 5.1 and Fig. 5.2.

Nevertheless, we have not found a complete justification to explain why some applications are cancelled after a validation step. This would require to interact with the financial institution to get higher insight about these cases.

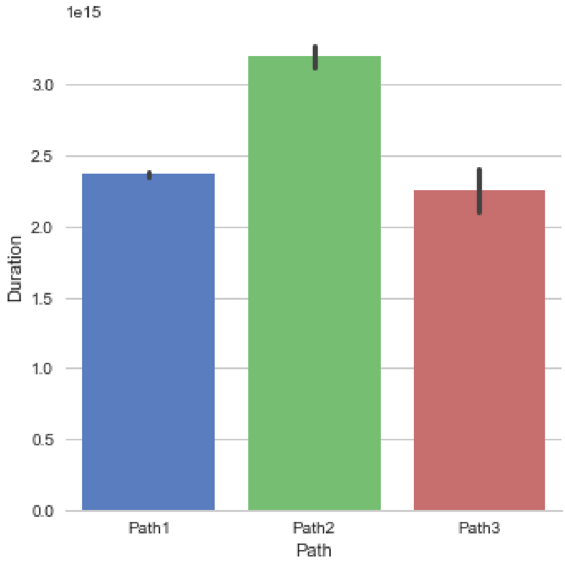


Figure 5.2: Cases duration per path

5.2 Accepted Applications

The next kind of offers studied in this phase are the approved ones. These offers have as a common join point their pass through the O_Accepted event. After filtering these applications in Pandas we can appreciate 44 different paths, with different weights like is shown in the Fig. 5.3.

The main reason for this huge number of variations is the amount of loops that the application passes through before the final acceptation. Other sources are the order for the events registration or the extra offers creation within the validation process. Those variations have shorts amounts of cases, and for that reason are not treated as a difference value.

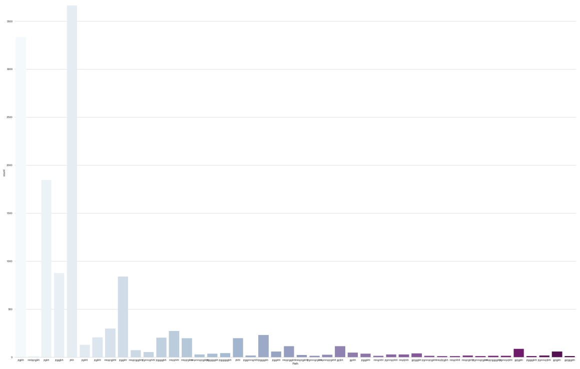


Figure 5.3: Accepted applications possible paths

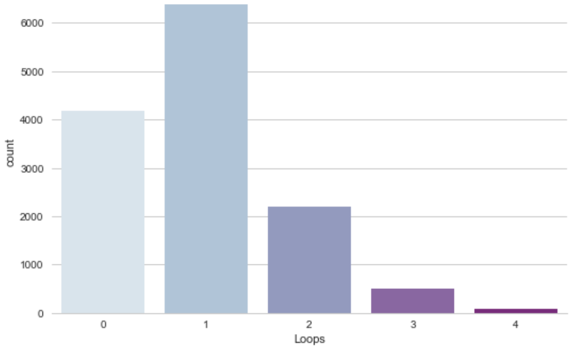


Figure 5.4: Accepted cases per amount of loops

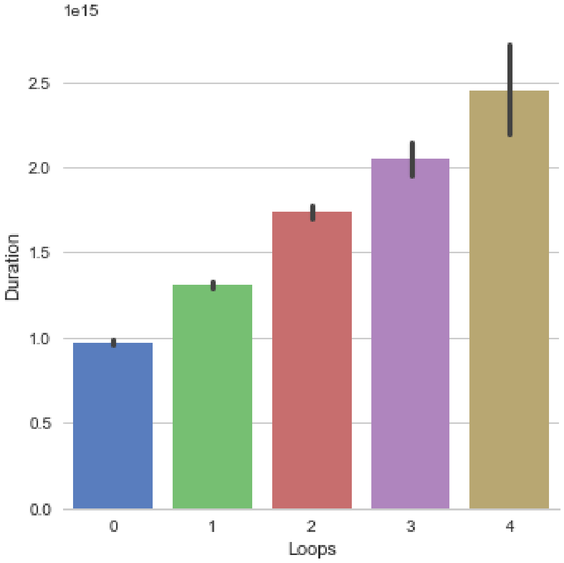


Figure 5.5: Duration in function of the amount of loops

Having this in mind the cases have been grouped in function of their loops requesting for extra information obtaining the cases distribution shown in Fig. 5.4. It has been considered the amount of loops, as the number of times where the application fires the O_Returned event.

It is important to highlight the loops distribution as shown in Fig. 5.4. There are more cases where the offer is accepted after one validation than with no more information requested. This could reveal a possible misinformation in the applicants which can have repercussions in the company efficiency as shown in Fig. 5.5. The main waiting time source is again the waiting time for the applicant response, as shown in Fig. 5.5. Moreover, since in every loop there is a waiting time for an applicant, this drives the increment in the overall process duration.

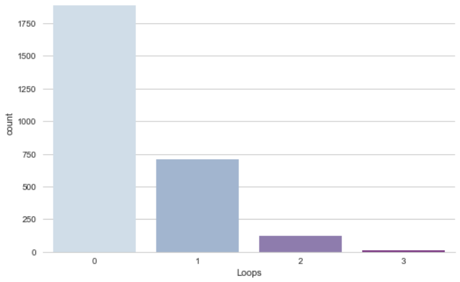


Figure 5.6: Denied cases per amount of loops

5.3 Denied Applications

Finally, within this phase the refused cases are filtered. For these cases it is proposed the same approach than for the accepted cases due to its multiple similarities. In the first place, although there are fewer cases than in the previous phase, we can appreciate that the main cause of variations is the loop existence. Once again the overall phase time increases with every loop. Following the global phase characteristics, some offers are created within these cases. Nevertheless, this is not relevant for the time analysis, because there is a great difference between the waiting time for an applicant response times and the time to create an offer. In these cases, unlike in the accepted ones, there are not more cases requesting for more information than being refused directly, like it is shown in Fig. 5.6. Our hypothesis is that applicants refuse an offer based on their characteristics, and the requirement of extra information has a limited impact.

Regarding time analysis, this phase follows the same pattern than the accepted cases, increasing times with the number of loops, as shown in Fig. 5.7.

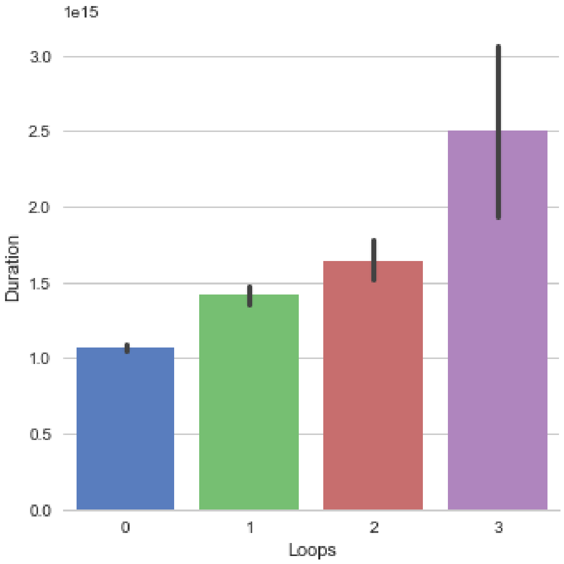


Figure 5.7: Duration of Denied cases per amount of loops

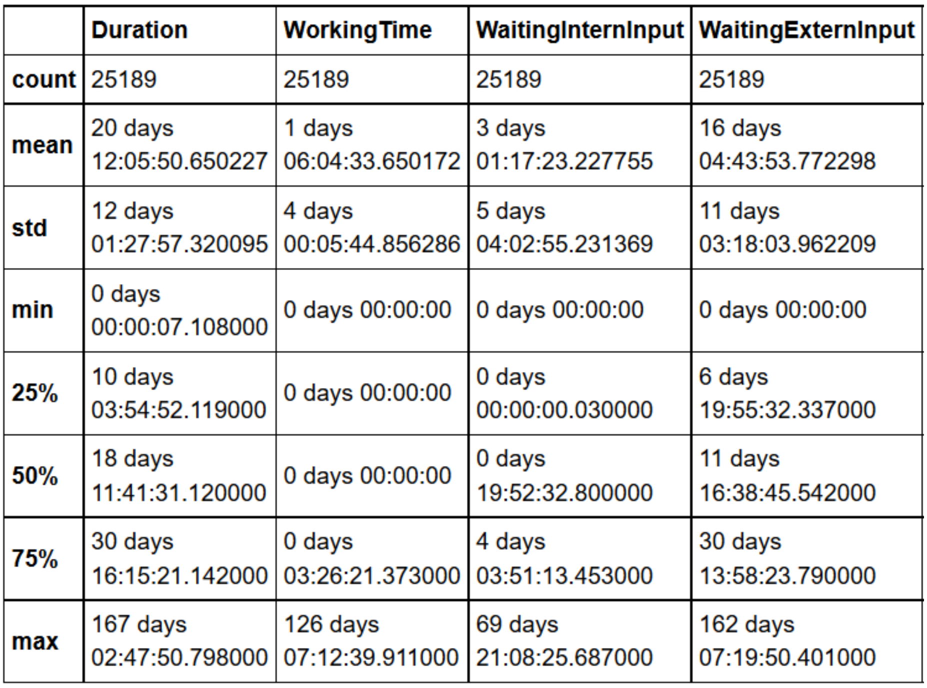
5.4 Overall phase conclusion

After dividing the phase in their different paths in order to understand and find out the main delaying causes.

Regarding time analysis, there are many cases with working time equal to zero as in the previous phases. For these cases is important to point out that there are paths like the ones intermediately cancelled where it makes sense these null working times, because those cases are cancelled after waiting for a response that never arrives and no work is required. For the rest of the cases with workflow type events, we have not found an evidence based explanation. Our plausible explanation is a wrong time registration of times in the company.

Waiting times have been classified into waiting times due to an external applicant from those waiting in the application, as shown in Table 5.1 below. Waiting times due an applicant tend to be higher for almost all the cases than working and waiting times inside the application. This is reasonable since the company has to wait for the applicant final response after sending an offer. This answer normally is delayed for a couple of days in the best cases, and it can last until the 30 days timeout is reached for the worst ones. These are the cases where no answer is received. If there is more than one loop inside the application process this 30 days limit could be exceeded.

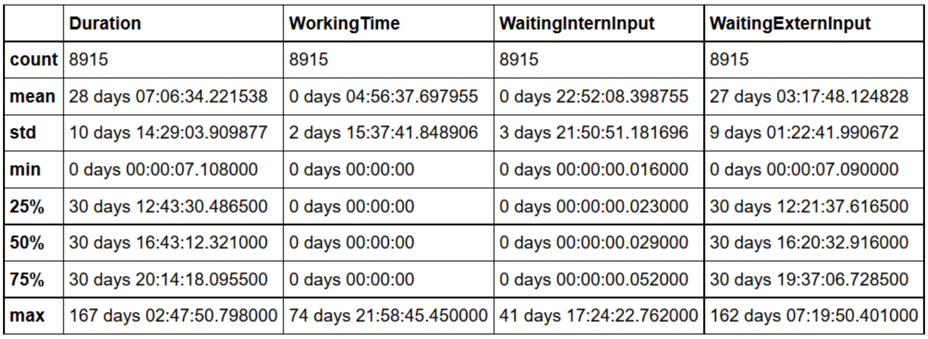
A comparison between the times in function of their ends is shown in Table 5.2, Table 5.3 and Table 5.4. These tables illustrate that the longest cases use to be the cancelled ones,



	Duration	WorkingTime	WaitingInternInput	WaitingExternInput
count	25189	25189	25189	25189
mean	20 days 12:05:50.650227	1 days 06:04:33.650172	3 days 01:17:23.227755	16 days 04:43:53.772298
std	12 days 01:27:57.320095	4 days 00:05:44.856286	5 days 04:02:55.231369	11 days 03:18:03.962209
min	0 days 00:00:07.108000	0 days 00:00:00	0 days 00:00:00	0 days 00:00:00
25%	10 days 03:54:52.119000	0 days 00:00:00	0 days 00:00:00.030000	6 days 19:55:32.337000
50%	18 days 11:41:31.120000	0 days 00:00:00	0 days 19:52:32.800000	11 days 16:38:45.542000
75%	30 days 16:15:21.142000	0 days 03:26:21.373000	4 days 03:51:13.453000	30 days 13:58:23.790000
max	167 days 02:47:50.798000	126 days 07:12:39.911000	69 days 21:08:25.687000	162 days 07:19:50.401000

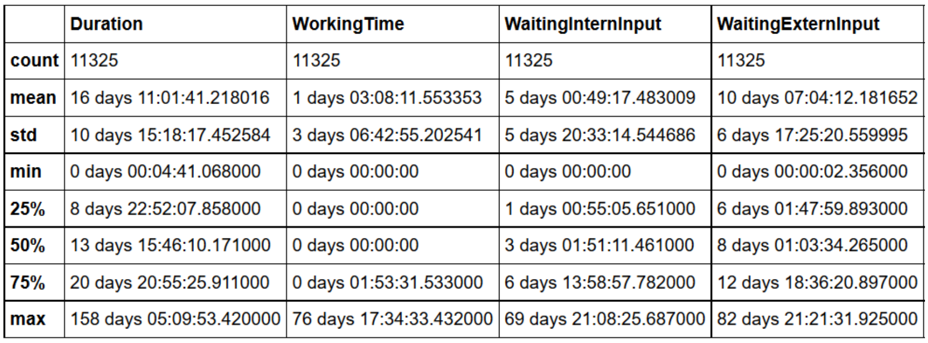
Table 5.1: Times calculation for the overall phase

mainly for its timeout dependency. The next longer cases are the accepted ones. The main appearance reason seems to be that when an offer satisfies the applicant expectation, the applicant is willing to fix every file an inconvenience that could appear. This do not happened in the cases where the offer does not satisfy them. They usually refuse it without trying to fix the files, providing an evidence to conclude that the main cause for a refused offer is not the requirements for completion but the offer itself.



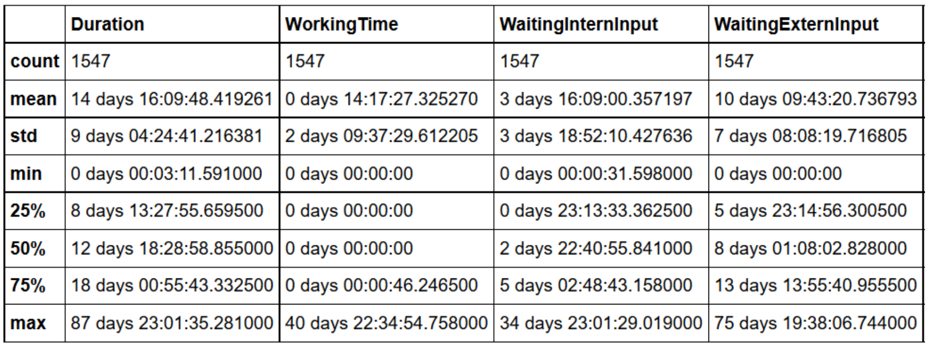
	Duration	WorkingTime	WaitingInternInput	WaitingExternInput
count	8915	8915	8915	8915
mean	28 days 07:06:34.221538	0 days 04:56:37.697955	0 days 22:52:08.398755	27 days 03:17:48.124828
std	10 days 14:29:03.909877	2 days 15:37:41.848906	3 days 21:50:51.181696	9 days 01:22:41.990672
min	0 days 00:00:07.108000	0 days 00:00:00	0 days 00:00:00.016000	0 days 00:00:07.090000
25%	30 days 12:43:30.486500	0 days 00:00:00	0 days 00:00:00.023000	30 days 12:21:37.616500
50%	30 days 16:43:12.321000	0 days 00:00:00	0 days 00:00:00.029000	30 days 16:20:32.916000
75%	30 days 20:14:18.095500	0 days 00:00:00	0 days 00:00:00.052000	30 days 19:37:06.728500
max	167 days 02:47:50.798000	74 days 21:58:45.450000	41 days 17:24:22.762000	162 days 07:19:50.401000

Table 5.2: Canceled cases times description



	Duration	WorkingTime	WaitingInternInput	WaitingExternInput
count	11325	11325	11325	11325
mean	16 days 11:01:41.218016	1 days 03:08:11.553353	5 days 00:49:17.483009	10 days 07:04:12.181652
std	10 days 15:18:17.452584	3 days 06:42:55.202541	5 days 20:33:14.544686	6 days 17:25:20.559995
min	0 days 00:04:41.068000	0 days 00:00:00	0 days 00:00:00	0 days 00:00:02.356000
25%	8 days 22:52:07.858000	0 days 00:00:00	1 days 00:55:05.651000	6 days 01:47:59.893000
50%	13 days 15:46:10.171000	0 days 00:00:00	3 days 01:51:11.461000	8 days 01:03:34.265000
75%	20 days 20:55:25.911000	0 days 01:53:31.533000	6 days 13:58:57.782000	12 days 18:36:20.897000
max	158 days 05:09:53.420000	76 days 17:34:33.432000	69 days 21:08:25.687000	82 days 21:21:31.925000

Table 5.3: Accepted cases times description



	Duration	WorkingTime	WaitingInternInput	WaitingExternInput
count	1547	1547	1547	1547
mean	14 days 16:09:48.419261	0 days 14:17:27.325270	3 days 16:09:00.357197	10 days 09:43:20.736793
std	9 days 04:24:41.216381	2 days 09:37:29.612205	3 days 18:52:10.427636	7 days 08:08:19.716805
min	0 days 00:03:11.591000	0 days 00:00:00	0 days 00:00:31.598000	0 days 00:00:00
25%	8 days 13:27:55.659500	0 days 00:00:00	0 days 23:13:33.362500	5 days 23:14:56.300500
50%	12 days 18:28:58.855000	0 days 00:00:00	2 days 22:40:55.841000	8 days 01:08:02.828000
75%	18 days 00:55:43.332500	0 days 00:00:46.246500	5 days 02:48:43.158000	13 days 13:55:40.955500
max	87 days 23:01:35.281000	40 days 22:34:54.758000	34 days 23:01:29.019000	75 days 19:38:06.744000

Table 5.4: Canceled cases times description

CHAPTER 6

Overall process study and conclusions

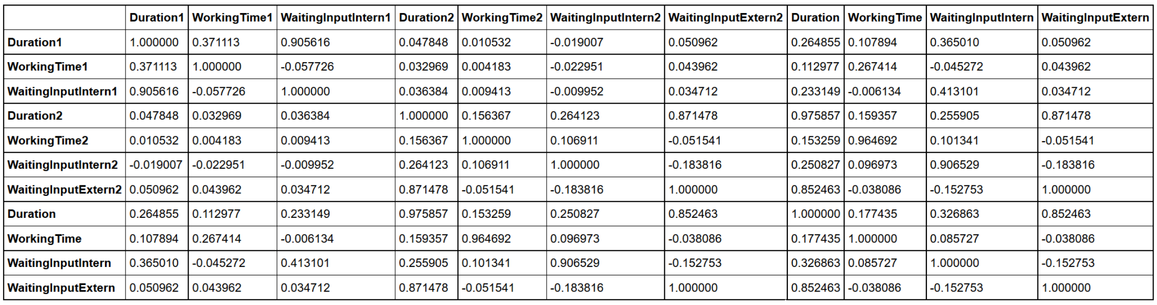
Finally, after the study of the two phases separately, we provide conclusions analyzing the overall process.

6.1 Times per process phase conclusions

After analyzing every phase, our first conclusion is that Offer Validation phase has the principal weight in the application process time. We provide evidences for this conclusion in Table 6.1, which shows the Pearson correlation coefficients between different phase times.

Going through this phase, the waiting time for an applicant answer is the most relevant. As discussed previously, it is important to highlight that it is not clear and easy to determinate the real weight for the time spent inside the application, since there exists a big number of cases with working time equal to zero. But, despite of this situation, the biggest cause in the delay of an application is the time waiting for an applicant input, as can be appreciated in Table 6.1.

Some potential reasons to explain this wrong time registration are bad training of human resources or that the application does not provide a suitable support for the real process. For both of these potential reasons, a conformance check between the reality and the application



	Duration1	WorkingTime1	WaitingInputIntern1	Duration2	WorkingTime2	WaitingInputIntern2	WaitingInputExtern2	Duration	WorkingTime	WaitingInputIntern	WaitingInputExtern
Duration1	1.000000	0.371113	0.905616	0.047848	0.010532	-0.019007	0.050962	0.264855	0.107894	0.365010	0.050962
WorkingTime1	0.371113	1.000000	-0.057726	0.032969	0.004183	-0.022951	0.043962	0.112977	0.267414	-0.045272	0.043962
WaitingInputIntern1	0.905616	-0.057726	1.000000	0.036384	0.009413	-0.009952	0.034712	0.233149	-0.006134	0.413101	0.034712
Duration2	0.047848	0.032969	0.036384	1.000000	0.156367	0.264123	0.871478	0.975857	0.159357	0.255905	0.871478
WorkingTime2	0.010532	0.004183	0.009413	0.156367	1.000000	0.106911	-0.051541	0.153259	0.964692	0.101341	-0.051541
WaitingInputIntern2	-0.019007	-0.022951	-0.009952	0.264123	0.106911	1.000000	-0.183816	0.250827	0.096973	0.906529	-0.183816
WaitingInputExtern2	0.050962	0.043962	0.034712	0.871478	-0.051541	-0.183816	1.000000	0.852463	-0.038086	-0.152753	1.000000
Duration	0.264855	0.112977	0.233149	0.975857	0.153259	0.250827	0.852463	1.000000	0.177435	0.326863	0.852463
WorkingTime	0.107894	0.267414	-0.006134	0.159357	0.964692	0.096973	-0.038086	0.177435	1.000000	0.085727	-0.038086
WaitingInputIntern	0.365010	-0.045272	0.413101	0.255905	0.101341	0.906529	-0.152753	0.326863	0.085727	1.000000	-0.152753
WaitingInputExtern	0.050962	0.043962	0.034712	0.871478	-0.051541	-0.183816	1.000000	0.852463	-0.038086	-0.152753	1.000000

Table 6.1: chapters division

should be executed.

For the waiting times reduction it would be interesting to check if the user experience is well defined, marking a clear path to follow for the applicant in the different channels. An investment in improving the web services, and designing an omnichannel service could reduce these waiting times. An omnichannel service refers to the possibility of commuting the same application between different channels.

6.2 Request for additional information drives refused offers

Our objective was validating if data analysis and mining could support the suggested hypothesis, this is, that multiple completion requirements drive a bigger amount of denied applications.

It was concluded in the Chap. 5 that this hypothesis was not supported. Analyzing the accepted cases, there are more cases with more than one request for completion than cases completed at the first validation. Regarding denied offers, the majority of cases were denied at the first validation. The evidences for these conclusions are shown in Fig. 5.6 and Fig. 5.4

From this behavior, two conclusions can be stated. First, the applicants does not seem to understand clearly the information that has been requested to them. This misunderstanding drives to a time wasting due the multiple requirements. Second, the main reason of applicants for refusing an offer is the offer itself, not the requests for more information. This is concluded after seen that the majority of the refused cases has a direct refusal.

It could be interesting to find the origin for the refusal offers. One hypothesis could be the misunderstanding of the offer and the original requirements. Since there are many accepted cases with more information requested, we can suppose that it is not completely clear the offer or the information requested. This lack of comprehension of information requests could be a reason for offer refusal.

Thus, we would recommend simplifying and clarifying the loan procedure and the information requested. This would lead to waiting time reduction and improving the business KPIs by reducing offers refusal.

6.3 Offers amount, and single or multiple conversation conclusion

Finally, we provide the conclusions for the last point requested for the institute, that is, how many customers ask for more than one offer.

As discussed before, offers can be asked in multiple conversations or in a single one. Single conversation offers are those who happened within the Phase 1, before the A_Completed state. Multiple conversation offers are those cases with more than one offer, where some of those offers were created during the validation process. Thanks to Disco [12] filtering capabilities, we obtained that 4,012 cases have more than one offer. Then, we proceed to filter these cases using Pandas for both single and multiple conversations study. Fig. 6.1 shows the evidences for our conclusions. We can appreciate that there are more cases with multiple conversation offers than multiple offer in single conversation. It is also appreciated that the multiple conversation cases have higher chances to be finally accepted.

A plausible reason to explain this fact is that after an offer is validated for the client, if he is not satisfied, he could renegotiate. This would drive the creation of an additional offer more aligned to the applicant objectives. Nevertheless, receiving more than one offer at the same time can be confusing for the applicant and make him to reject them.

After this study, our recommendation is maintaining only one active offer, that could be modified during the process.

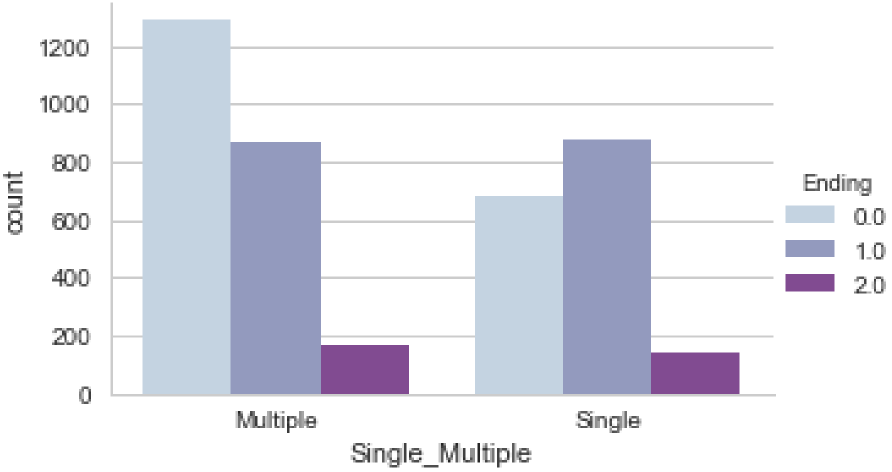


Figure 6.1: Comparative between cases with offers created in multiple or single conversations

CHAPTER 7

Conclusions

In this chapter we will describe the conclusions extracted from this final project, the achievements and thinkings about future work.

7.1 Conclusions

In this project different process mining tools have been applied to a real case from a Dutch Financial Institute. This case data consists of a log from a Loan Approval process within the Financial intitute. Data mining tools were also used for the answer of the principal hypothesis and questions from this institute. With these tools a deep study about different process characteristics has been accomplished. The three aspects under scope finally were: The different times per process phase, the influence of incompleteness in the final application result, and the amount of offers submitted to the applicant in different conversations. Finally different conclussions were obtained and possible improvements were suggested showing the scope accomplished for an external Process Mining study without contact with the institution.

This project has been developed looking for the possible conclusions and improvements enabled by a PM study in combination with python based Data Mining tools without more information than the log itself. This objective was setted looking for the capabilities of these tools and approcahes in a real situation.

During this project it has been realised that mnay conclussions and information can be obtained with the logs, but in this process many assumptions has been taken in order to explain different patterns. Having a relation with the different stakeholders many of these assumpion could be clarified faster and with a better approach to the real situation within the institute. A combination between these techniques and a classic business process reengineering could enhance faster and better the multiple aspects included in a company operative.

7.2 Achieved goals

The goals achieved from this analysis of a real process engineering case through process and data mining tools are the following:

- Deep understanding of the process and finding of the main issues in the fields analyzed. With the tools described the objective was to understand the process and being able to answer the principal questions and hypothesis from the institute. After the study we could extract the main patterns within the process with the chosen approach.
- Scope of an analysis with PM and Data mining tools without more information than the data log itself. After the study we find the main strenghts and weaknesses of an extern analysis of a company process with PM and Data mining tools.
- The submission of a paper to the 13th International Workshop on Business Process Intelligence 2017 congress for the BPIC 2017.

7.3 Future Work

In the following points some of study or improvement are presented to continue the development:

- Test of the application of these techniques in an overall company process reengineering.
- Application of different machine learning algorithms to simplify and improve the data mining results.

- Detection of errors in the log registration process for the following studies improvement.
- Test and application of extra PM capabilities in deep like the resource performance study.

Bibliography

[1] Arya Adriansyah and Joos C. A. M. Buijs. *Mining Process Performance from Event Logs*, pages 217–218. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[2] Joey Bernard. Python data analysis with pandas. In *Python Recipes Handbook*, pages 37–48. Springer, 2016.

[3] R. P. Jagadeesh Chandra Bose and Wil M. P. van der Aalst. *Process Mining Applied to the BPI Challenge 2012: Divide and Conquer While Discerning Resources*, pages 221–222. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[4] Juliana Baptista dos Santos França, Joanne Manhães Netto, Rafael Gomes Barradas, Flávia Santoro, and Fernanda Araujo Baião. *Towards Knowledge-Intensive Processes Representation*, pages 126–136. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[5] C Günther. Extensible event stream xes standard definition. 2009.

[6] Christian W Günther and Wil MP Van Der Aalst. Fuzzy mining–adaptive process simplification based on multi-perspective metrics. In *International Conference on Business Process Management*, pages 328–343. Springer, 2007.

[7] Chang Jae Kang, Chul Kyu Shin, Eun Sang Lee, Ju Hui Kim Kim, and Min Ah An. Analyzing application process for a personal loan or overdraft of dutch financial institute with process mining techniques, 2012.

[8] Thomas Kluyver, Benjamin Ragan-Kelley, Fernando Pérez, Brian Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, Jessica Hamrick, Jason Grout, Sylvain Corlay, et al. Jupyter notebooks—a publishing format for reproducible computational workflows. *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, page 87, 2016.

[9] Sander JJ Leemans, Dirk Fahland, and Wil MP van der Aalst. Exploring processes and deviations. In *International Conference on Business Process Management*, pages 304–316. Springer, 2014.

[10] Thomas Molka, Wasif Gilani, and Xiao-Jun Zeng. *Dotted Chart and Control-Flow Analysis for a Loan Application Process*, pages 223–224. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[11] Fariborz Moshirian. The global financial crisis and the evolution of markets, institutions and regulation. *Journal of Banking & Finance*, 35(3):502–511, 2011.

[12] Anne Rozinat. *Disco's User Guide.*

[13] Wil M P van der Aalst. *Process Mining.* Springer Berlin Heidelberg, Berlin, Heidelberg, 2016.

[14] Boudewijn F Van Dongen, Ana Karla A de Medeiros, HMW Verbeek, AJMM Weijters, and Wil MP Van Der Aalst. The prom framework: A new era in process mining tool support. In *International Conference on Application and Theory of Petri Nets*, pages 444–454. Springer, 2005.

[15] H. M. W. (Eric) Verbeek. *BPI Challenge 2012: The Transition System Case*, pages 225–226. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[16] Wayne Winston. *Microsoft Excel data analysis and business modeling.* Microsoft press, 2016.