UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
ESCOLA DE INFORMÁTICA APLICADA

Using Process Mining Techniques to Support Improvement in a Financial Institution Process

JESSICA ARAÚJO LEITE
MARCELLA TAVARES SANT'ANNA

Supervisors
FLÁVIA MARIA SANTORO
GEIZA MARIA HAMAZAKI DA SILVA

Catalogação informatizada pelo(a) autor(a)

Leite, Jessica
LLei52 Using Process Mining Techniques to Support
5Us84 Improvement in a Financial Institution Process /
Jessica Leite. -- Rio de Janeiro, 2017.
79

Orientadora: Flávia Santoro.
Coorientadora: Geiza Hamazaki.
Trabalho de Conclusão de Curso (Graduação) -
Universidade Federal do Estado do Rio de Janeiro,
Graduação em Sistemas de Informação, 2017.

1. Process Mining. I. Santoro, Flávia, orient.
II. Hamazaki, Geiza, coorient. III. Título.

RIO DE JANEIRO, RJ – BRASIL

JULHO DE 2017

Using Process Mining Techniques to Support Improvement
in a Financial Institution Process

JESSICA ARAÚJO LEITE
MARCELLA TAVARES SANT'ANNA

Graduation Project presented to the Escola de Informática Aplicada of the Universidade Federal do Estado do Rio de Janeiro (UNIRIO) to obtain the Bachelor's degree in Information Systems.

Approved by:

_____

FLÁVIA MARIA SANTORO (UNIRIO)

_____

GEIZA MARIA HAMAZAKI DA SILVA (UNIRIO)

_____

KATE CERQUEIRA REVOREDO (UNIRIO)

RIO DE JANEIRO, RJ – BRASIL.

JULHO DE 2017

Special Thanks

I would like to thank the support of my family and friends during this time that I had to concentrate on the graduation project, to my friend Jessica Leite who at all times kept the animation and confidence in this work, making it more joyful and pleasurable, to our supervisors: Geiza Hamazaki who introduced us to an idea of research and investigation from information systems skills and Flávia Santoro who presented us with the idea of participating in an international challenge, and also to Professor Kate Revoredo who taught about process mining, giving enlightening explanations, to the Master's Degree student Valdemar Confort who clarified doubts regarding process mining softwares, and for both of them always encourage us in relation to the theme and the challenge.

Marcella Tavares.

Firstly I would like to thank Marcella Tavares for the opportunity of working in such an interesting project and for all the support given. I would also like to thank our supervisor professors Geiza Hamazaki and Flávia Santoro, for all the support, time, knowledge, insights and patience dedicated to us, so we could elaborate a great graduation project.
I cannot fail to mention the extremely important support given by the Process Mining Professor Kate Revoredo, and her Master's Degree student, Valdemar Confort, who taught us about the project's subject and were always available to solving our doubts.
Finally, I would like to thank all my loved ones for understanding my absence during this project elaboration.

Jessica Leite.

ABSTRACT

This paper presents the analysis of the system logs of a financial institution, made available to BPI Challenge 2017, through process mining techniques, which is the activity of extracting knowledge from event logs that are recorded by systems. With the aim of mining this content, the following softwares were used: Disco and Celonis. Through these techniques, all process flows were meticulously mapped/discovered and investigated, in order to identify possible inefficiencies with the focus on the frequency of events, and thus it was possible to indicate points to be improved. In addition, we searched for some relevant behavior patterns that might allow the company to perform further analysis, suggesting changes, improvements, corrections and/or learning its processes.

Key words: ANALYSIS, BPI, PROCESS MINING, LOGS, TECHNIQUES, IMPROVEMENTS, BEHAVIOR PATTERNS, EXTRACTING KNOWLEDGE, INEFFICIENCIES, FREQUENCY OF EVENTS.

Index

Tables Index

Figures Index

1 Introduction

1.1 Motivation

Business Process Management (BPM) is a method that “aims to discover, monitor and improve real processes by extracting knowledge from event logs”, according tovan der Aalst (2014).
It “starts from events and the output is related to an end-to-end process model”, according tovan der Aalst (2015). In the last years, process mining has increased as a new decision support method for companies. These methods techniques analyze systems, event logs and help to improve business process. In addition, the results can guide to conclusions about other aspects of the business.

The BPIC makes part of the International Workshop on Business Process Intelligence that occurs annually, held in conjunction with BPM International Conference that provides the most prestigious forum on Business Process Management area. BPI is about the application of data and process mining techniques in the field of Business Process Management.

BPI Challenge 2017 [1] provided a dataset from a financial institute, the same company from BPI Challenge 2012[2]. There were changes during this time and they decided that a new participation would be great to answer some questions. Three questions were required and also extra analyses were demanded. Two log files were available and some explanations about the data given.

[1] https://www.win.tue.nl/bpi/doku.php?id=2017:challenge
[2] http://www.win.tue.nl/bpi/doku.php?id=2012:challenge

1.2 Aims

The aims of this project are the construction of quantitative and qualitative analyses regarding the logs made available for the BPI Challenge 2017 by the financial institution. These analyzes were guided by the script of questions of the challenge, and in turn, based on the results found through the filters performed in the logs, led us to choose the ideal algorithm.
Therefore, it was possible to obtain more information regarding the scenarios that we identified, such as those that influence in a more relevant way the positive conclusion of the process as a whole.
The process of mining the logs in question was oriented to the construction of a more critical vision:

- Visualization of process mining results;
- Process discovery;
- Analyze the data;
- Suggest a process remodeling in order to improve it;
- Business process quality;
- Look for gaps in the process

1.3 Organization of the text

This dissertation is structured in chapters and, in addition to this introduction, will be developed as follows:

- Chapter II: Context: Specification of the proposed challenge, the environment to be studied and the analysis items to be developed.
- Chapter III: Methods, Algorithm and Tools: Definition of the methods and algorithm that will be used to develop the results and the tools that will help this goal.
- Chapter IV: Analysis: Detailed exploration of the analysis that will be developed and results obtained from what will be found. Graphs and images of the tools will be used to present results.

- Chapter V: Conclusions – Final conclusions and considerations, project limitations, suggestions.
- Chapter VI: Appendix – Descriptions of what were seen to find the results of the analysis of throughput times per part of the process and all the correspondents results.

2 Context

2.1 About BPI Challenge 2017

The International Workshop on Business Process Intelligence (BPI) takes place every year in conjunction with the International Conference on Business Process Management (BPM). BPM [3] is an annual conference that provides the most prestigious forum for researchers and practitioners in the field of Business Process Management (BPM). Over the past decade, the conference has built its reputation by showcasing leading-edge research of the highest quality together with talks, tutorials and discussions by the most renowned thought leaders and innovators in the field. The BPM conference series embraces the diversity and richness of the BPM field and serves as a melting pot for experts from a mix of disciplines including Computer Science, Information Systems Management, Services Science and Technology Management.

The BPI Workshop [4] refers to the application of data- and process-mining techniques in the field of Business Process Management. BPI is an area that spans process mining, process discovery, conformance checking, predictive analytics and many other techniques that are all gaining interest and importance in industry and research. In practice, BPI is embodied in tools for managing process execution by offering several features such as analysis, prediction, monitoring, control, and optimization.

The Business Process Intelligence Challenge (BPIC) is held as part of BPI Workshop. BPIC provides real-life event logs to participants and receives the papers submissions that are documenting the results of the analysis of these data, using any methods, techniques and tools available. There is no need to be open-source tools.

There are 3 categories of participants - as shown below.

[3] https://bpm2017.cs.upc.edu/
[4] http://www.win.tue.nl/bpi/doku.php?id=2017:start

The *Student category* targets Bachelor, Master and PhD students or student teams. Focus: "The originality of the results, the validity of the claims and the depth of the analysis of specific issues identified." [5]

The *Academic category* targets academics. Focus: "Is much more on the novelty of the techniques applied than the actual results." [6]

The *Professional category* targets professionals. Focus: Completeness and usefulness of analysis, broader range of aspects and level of professionalism.

Then, a jury decides which are the best ones and the winners of each category are announced at the annual International Workshop on Business Process Intelligence.

In 2017, the challenge is concerned to answer questions from a financial institute thathad already provided logs to the BPIC in 2012. After the success of this log and the implemented changes by the company, also with the financial crisis, the case volume increased. In 2017, there are 3 process owner's questions and the company is also interested in other unique insights captured from the event logs. This year it was sponsored by Minit[7] and Celonis[8] that made available trial versions of their softwares for the participants.

2.2 About the Data

According to PROM organization web site[9], to be able to apply process mining techniques it is essential to extract *event logs* from data sources (e.g., databases, transaction logs, audit trails, etc.). XES is the standard format for process mining supported by the majority of process mining tools.

[5] https://www.win.tue.nl/bpi/doku.php?id=2017:challenge
[6] https://www.win.tue.nl/bpi/doku.php?id=2017:challenge
[7] https://www.minit.io/
[8] https://www.celonis.com/
[9] http://www.processmining.org/logs/start

Process mining assumes the existence of an event log where each event refers to a case, an activity, and a point in time. An event log can be seen as a collection of cases and a case can be seen as a trace/sequence of events.

Event data may come from a wide variety of sources:

- a database system (e.g., patient data in a hospital),
- a comma-separated values (CSV) file or spreadsheet,
- a transaction log (e.g., a trading system),
- a business suite/ERP system (SAP, Oracle, etc.),
- a message log (e.g., from IBM middleware),
- an open API providing data from websites or social media,
- …

In BPI Challenge 2017, the financial company involved event logs regarding their main business activities.

The data contains 1,202,267 events divided in three types:

1. Application state changes (A);
2. Offer state changes (O) and
3. Workflow events (W).

These events correspond to 31,509 loan applications with 42,995 offers. Also there are 149 originators (employees or systems of the company) in the data.
The data was provided in two files. The application event log and the offer event log. An application can have multiple offers. An offer is always related to one application. If the application has multiple offers and one of them is accepted, automatically the others are cancelled.

For all applications, the following data is available:
– Requested load amount (in Euro),
– The application type,
– The reason the loan was applied for (Loan Goal), and
– An application ID.

For all offers, the following data is available:
– An offer ID,
– The offered amount,
– The initial withdrawal amount,
– The number of payback terms agreed to,
– The monthly costs,
– The credit score of the customer,
– The employee who created the offer,
– Whether the offer was selected, and
– Whether the offer was accepted by the customer. (BPIC 2017)[10]

2.3 Significant words for the analyses

- Application: Is a request of a loan made for a customer and can have different reasons. It is always related with, at least, one offer. If the application has more than one offer and one of them is accepted, the others are automatically cancelled. It has some attributes related and, for this reason, one of the logs provided is about the applications;
- Bank: The financial institution that provided the logs;
- Customer: Also mentioned as "Customer", is the applicant, who asks for a loan application;
- Conversation: When there are a contact with the customer;
- Conversion: When the payment is released to the customer;
- Incompleteness: When the bank verifies that something is missing in the application to proceed;

[10] https://www.win.tue.nl/bpi/doku.php?id=2017:challenge

- Mandatory: Type of attribute filter in Disco that removes all cases that do not have at least one event with one of the selected value(s);
- Offer: Options of loan made/analyzed by the customer or a bank employee. It is always related to one application. The offers of an application signalized the negotiation between the bank and the customer. It has attributes related to and a log of the offers was provided by the bank.
- Requests: When there is a case of incompleteness and the bank demands something of the customer;
- State: Is the status of the application. It tells if it was accepted, for example;
- User: Someone who works at the bank and uses the bank systems to proceed with the processes of the applications.

For a better understanding of the terms explained above, a visual example from de log provided can be found in Figure 54 (Annex).

2.4 Proposed Questions

The following topics were proposed by the BPIC 2017:

1. What are the throughput times per part of the process, in particular 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 as this is currently unclear,
2. What is 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,
3. How many customers ask for more than one offer (where it matters if these offers are asked for in a single conversation or in multiple conversations)? How does the conversion compare between applicants for whom a single offer is made and applicants for whom multiple offers are made?
4. Any other interesting trends, dependencies etc. (BPIC 2017)

3 Methods, Algorithm and Tools

3.1 Process Mining

Through process mining techniques it is possible to analyze business processes based on event logs, extract information and to monitor processes. It is seen as a "bridge between data mining and business process modeling". According to van der Aalst (2016).

With the aid of mining algorithms, the main goals of process mining are to understand the processes, identify behaviors and trends, upgrade the efficiency, the improvement of business processes and enrich the information systems. Thus, various perspectives are addressed through mining.

The main types of process mining are described below:

Play-in (Process Discovery): Produces a process model using an event log.
Replay (Conformance Checking): It is a comparison between an existing process model and an event log of the same process. Aims to check the reality with the model and vice versa and to check deviations.
Replay (Enhancement): Use the event log to improve and extend the process model. Used to find bottlenecks, for example.
Play-out: Shows a simulation of the model, a workflow automation to generate an event log.

So, these techniques are used to make analyses extracting relevant information using data in a process context and with the results propose improvements and provide knowledge about the business processes.

To exemplify the data, you can observe in Figure 1 a fragment of an event log. In this example is possible to observe that for the cases (specified by a case id), some information are recorded. Each case has the event id's that set it. And each event id has the properties that set it too, as the timestamp, the activity name that occurred, the resource that made it and the cost associated.

Case ID	Activity	Resource	Start Timestamp	Complete Timestamp	Variant	Variant index	(case) Application Type	(case) LoanGoal	(case) Requested Amount	Accepted	Action	CreditScore	EventID	EventOrigin
Application_652823628	A_Create Application	User_1	2016/01/01 07:51:15.304	2016/01/01 07:51:15.304	Variant 2	2	New credit	Existing loan takeover	20000.0		Created		Application_652823628	Application
Application_652823628	A_Submitted	User_1	2016/01/01 07:51:15.352	2016/01/01 07:51:15.352	Variant 2	2	New credit	Existing loan takeover	20000.0		statechange		ApplState_1582051990	Application
Application_652823628	A_Concept	User_1	2016/01/01 07:52:36.413	2016/01/01 07:52:36.413	Variant 2	2	New credit	Existing loan takeover	20000.0		statechange		ApplState_642383566	Application
Application_652823628	W_Complete application	User_17	2016/01/02 08:45:22.429	2016/01/02 08:45:22.429	Variant 2	2	New credit	Existing loan takeover	20000.0		Obtained		Workitem_1875340971	Workflow
Application_652823628	A_Accepted	User_52	2016/01/02 09:23:04.299	2016/01/02 09:23:04.299	Variant 2	2	New credit	Existing loan takeover	20000.0		statechange		ApplState_99568828	Application
Application_652823628	O_Create Offer	User_52	2016/01/02 09:29:03.994	2016/01/02 09:29:03.994	Variant 2	2	New credit	Existing loan takeover	20000.0	TRUE	Created	979	Offer_1485801083	Offer
Application_652823628	O_Created	User_52	2016/01/02 09:29:05.354	2016/01/02 09:29:05.354	Variant 2	2	New credit	Existing loan takeover	20000.0		statechange		OfferState_1514834199	Offer
Application_652823628	O_Sent (mail and online)	User_52	2016/01/02 09:30:28.606	2016/01/02 09:30:28.606	Variant 2	2	New credit	Existing loan takeover	20000.0		statechange		OfferState_2051164740	Offer
Application_652823628	W_Call after offers	User_52	2016/01/02 09:30:28.631	2016/01/02 09:30:28.631	Variant 2	2	New credit	Existing loan takeover	20000.0		Obtained		Workitem_719909876	Workflow
Application_652823628	A_Complete	User_52	2016/01/02 09:30:28.633	2016/01/02 09:30:28.633	Variant 2	2	New credit	Existing loan takeover	20000.0		statechange		ApplState_946455804	Application
Application_652823628	W_Validate application	User_117	2016/01/13 11:10:55.145	2016/01/13 11:10:55.145	Variant 2	2	New credit	Existing loan takeover	20000.0		Obtained		Workitem_1641716416	Workflow
Application_652823628	A_Validating	User_117	2016/01/13 11:10:55.973	2016/01/13 11:10:55.973	Variant 2	2	New credit	Existing loan takeover	20000.0		statechange		ApplState_752879093	Application
Application_652823628	O_Returned	User_117	2016/01/13 11:11:03.569	2016/01/13 11:11:03.569	Variant 2	2	New credit	Existing loan takeover	20000.0		statechange		OfferState_1310330551	Offer
Application_652823628	W_Call incomplete files	User_115	2016/01/14 07:16:20.972	2016/01/14 07:16:20.972	Variant 2	2	New credit	Existing loan takeover	20000.0		Obtained		Workitem_2031858979	Workflow

Figure 1 – Fragment of the log provided by BPIC 2017

Note that, in this log event fragment (Figure 1), it is possible to observe the activities that happen, the resource of the activity in the case, and the timestamp. Each column represents a type of data related to the process and it can be in presented in different format types (the data in a same column should present the same format). Through all the information provided by this log, it was possible to perform the necessary filters and, consequently, to answer the questions proposed by the challenge.

It is important to remember that "having high-quality event logs, process mining techniques can be used to improve business processes and predict problems.", according to van der Aalst (2011).

The log files analyzed are in the .XES format, which stands for extensible event stream. This kind of file format is based on a unified and extensible methodology to capture systems behaviors by means of event logs and event streams defined in its standards'[11]. From this kind of files, using tools such as ProM, Celonis and Disco, process models can be extracted.

Once we have the .XES log file, when importing it to ProM there are many data-mining algorithms can be applied in the event log dataset according to the analysis objective. For example, if the main objective was to generate an ideal net (process model), the genetic miner could be used. If the process log had process models of less structured process, the Fuzzy Miner algorithm could be used to emphasize graphically the most frequent behavior and avoid a "spaghetti" model. However, in order to answer to the challenge questions, it was decided that an algorithm focused on the frequency analysis of the events: the Heuristic Miner.

In the initial phase of the process discovery, there are no models defined yet. By mapping the event logs, a new model is built or discovered based on low-level events, thus generating a Petri Net process model (example in Figure 2). A Petri Net is a bipartite graph, in which the nodes represent transitions (events that may occur, represented by bars) and places (i.e. conditions, represented by circles). The directed arcs describe which places are pre- and/or post conditions for which transitions (signified by arrows). The example in Figure 2 shows the possible ways that a case can assume. The squares represents the activities that occurs, and is always preceded and followed by transitions (circle) - that can indicate the start or the end too. So, the arcs has the mission of show where to go. Is good to note that an activity can wait for one or more transitions to start its procedures and a transition can wait for anyone, one or more activities to start its procedures.

[11] http://www.xes-standard.org/

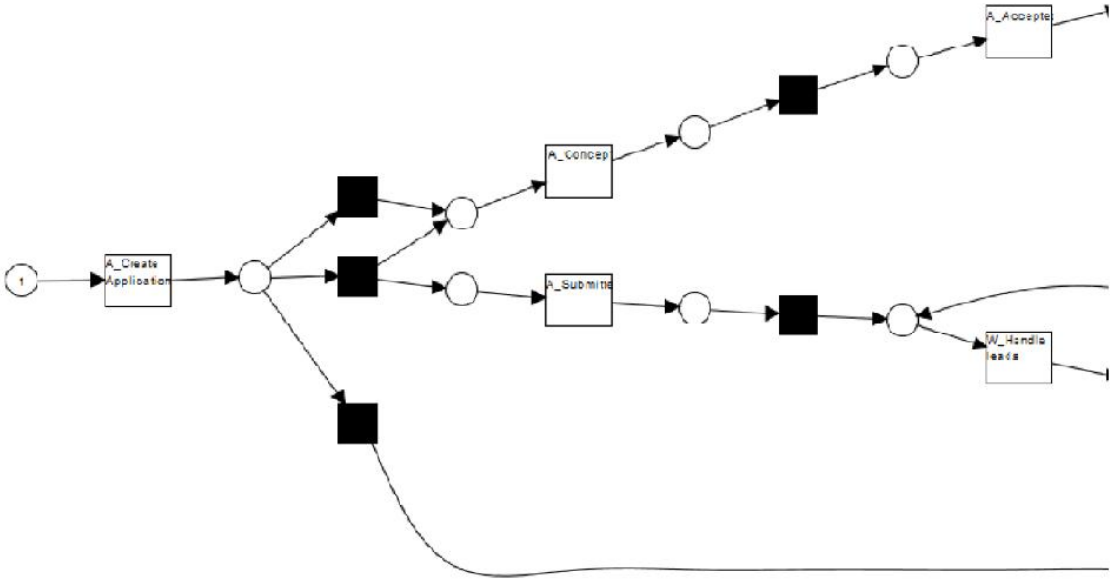


Figure 2 – Heuristic Net fragment from the BPIC 2017 log, turned into a Petri Net

In the petri net fragment represented in the picture above, it can be observed the beginning of the process. In the cases evaluated, the process always starts with the activity "A_CreateApplication" which is always followed by a decision point, represented by circles. After the decision point, three possible events are represented by the squares.

3.2 Heuristic Miner Algorithm

The Heuristic Miner Algorithm aims to mine processes, using statistics in the "dependency relationships between activities represented by logs. It focuses on the control flow perspective and generates a process model in the form of a Heuristic Network" [12]. It is based on frequency and, therefore, is less sensitive to noise and to the incompleteness of the records. Besides that: Detects small loops, Detects skipping activities and it does not guarantee sound process models.

[12] http://www.memoireonline.com/01/16/9357/m_Comparison-of-process-mining-techniques-application-to-flexible-and-unstructured-processes2.html

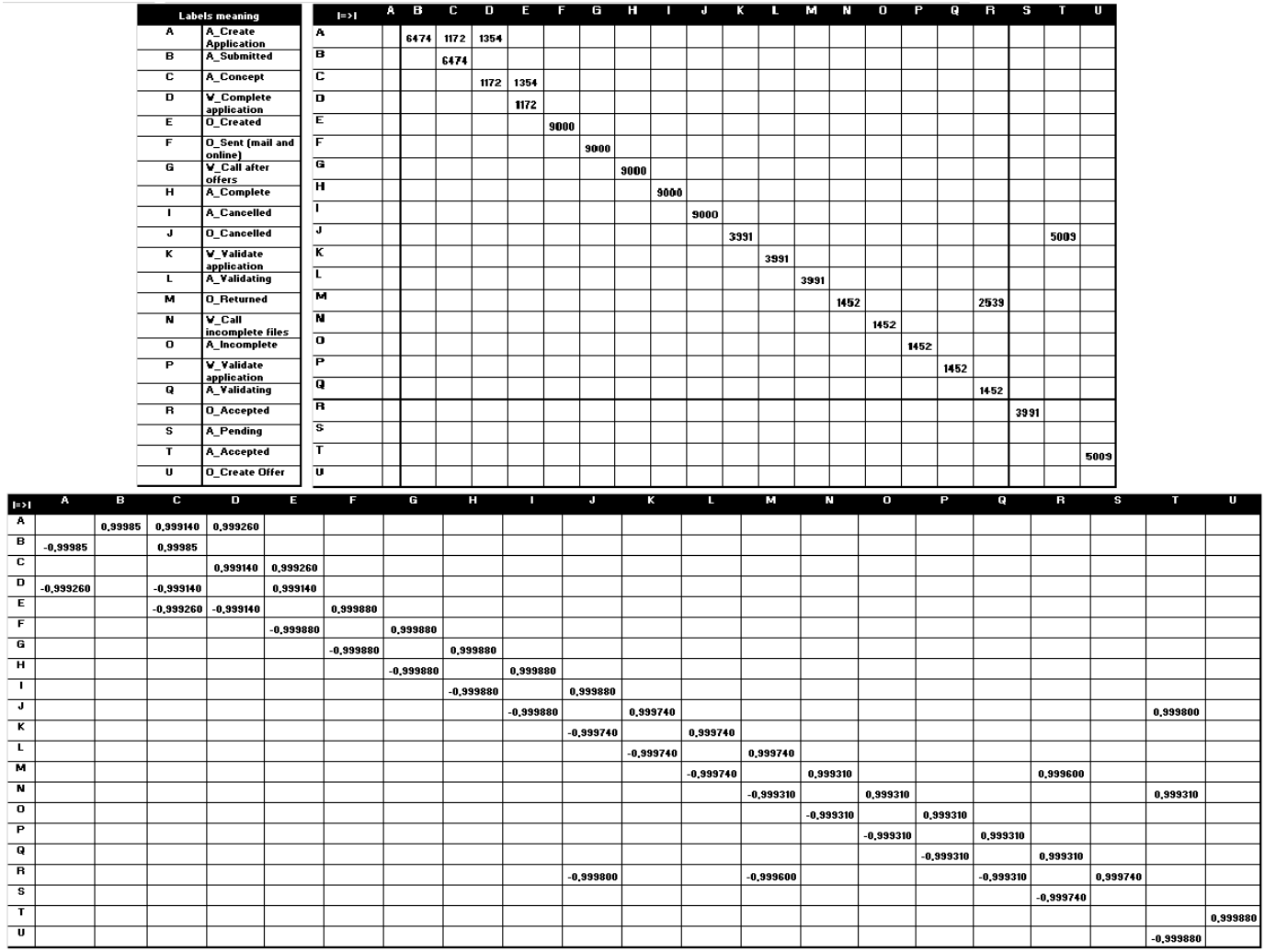
The algorithm creates a frequency matrix of direct tracking (see Figure 4) using the frequency of events - as can be seen in Figure 3. In this example the possible ways are represented by traces - represented by < ... >; it is just another way to show what was seen in Figure 1 and 2. To explain what it is, let's take the first trace to observation: It is possible to see that it occurred in total 3655 times. Besides, we can identify the activities that occurred and the sequence. This trace starts with activity "A_Create Application", followed by "A_Submitted", followed by "A_Concept", followed by "W_Complete application", followed by "A_Accepted", followed by "O_Create Offer", followed by “O_Created", followed by "O_Sent(mail and online)", followed by "W_Call after offers", followed by "A_Complete", followed by "A_Cancelled" and ends in "O_Cancelled".

Frequency	3655	1452	1367	1354	1172
Activities Flux	A_Create Application	A_Create Application	A_Create Application	A_Create Application	A_Create Application
	A_Submitted	A_Submitted	A_Submitted	W_Complete application	A_Concept
	A_Concept	A_Concept	A_Concept	A_Concept	W_Complete application
	W_Complete application	W_Complete application	W_Complete application	A_Accepted	A_Accepted
	A_Accepted	A_Accepted	A_Accepted	O_Create Offer	O_Create Offer
	O_Create Offer	O_Create Offer	O_Create Offer	O_Created	O_Created
	O_Created	O_Created	O_Created	O_Sent (mail and online)	O_Sent (mail and online)
	O_Sent (mail and online)	O_Sent (mail and online)	O_Sent (mail and online)	W_Call after offers	W_Call after offers
	W_Call after offers	W_Call after offers	W_Call after offers	A_Complete	A_Complete
	A_Complete	A_Complete	A_Complete	A_Cancelled	W_Validate application
	A_Cancelled	W_Validate application	W_Validate application	O_Cancelled	A_Validating
	O_Cancelled	A_Validating	A_Validating		O_Returned
		O_Returned	O_Returned		O_Accepted
		W_Call incomplete files	O_Accepted		A_Pending
		A_Incomplete	A_Pending		
		W_Validate application			
		A_Validating			
		O_Accepted			
		A_Pending			

Figure 3 - The traces that will be analyzed

The cases listed above, represent the top 5 variant cases in the whole log, without any kind of filtering.

For a better visualization, the traces observed in the image above will be converted into letter, as it can be seen in the following image.



Labels meaning	
A	A_Create Application
B	A_Submitted
C	A_Concept
D	W_Complete application
E	O_Created
F	O_Sent (mail and online)
G	W_Call after offers
H	A_Complete
I	A_Cancelled
J	O_Cancelled
K	W_Validate application
L	A_Validating
M	O_Returned
N	W_Call incomplete files
O	A_Incomplete
P	W_Validate application
Q	A_Validating
R	O_Accepted
S	A_Pending
T	A_Accepted
U	O_Create Offer

I=>I	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
A		6474	1172	1354																	
B			6474																		
C				1172	1354																
D					1172																
E						9000															
F							9000														
G								9000													
H									9000												
I										9000											
J											3991									5009	
K												3991									
L													3991								
M														1452				2539			
N															1452						
O																1452					
P																	1452				
Q																		1452			
R																			3991		
S																					
T																					5009
U																					

I=>I	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
A		0,99985	0,999140	0,999260																	
B	-0,99985		0,99985																		
C				0,999140	0,999260																
D	-0,999260		-0,999140		0,999140																
E			-0,999260	-0,999140		0,999880															
F					-0,999880		0,999880														
G						-0,999880		0,999880													
H							-0,999880		0,999880												
I								-0,999880		0,999880											
J									-0,999880		0,999740									0,999800	
K										-0,999740		0,999740									
L											-0,999740		0,999740								
M												-0,999740		0,999310				0,999600			
N													-0,999310		0,999310					0,999310	
O														-0,999310		0,999310					
P															-0,999310		0,999310				
Q																-0,999310		0,999310			
R										-0,999800			-0,999600				-0,999310		0,999740		
S																		-0,999740			
T																					0,999880
U																				-0,999880	

Figure 4 - The generated matrix - Significance and Frequency

Figure 4 shows how the significance and frequency matrix are generated. After we see the traces, two square matrices are created using the activities; each one to a type of record. First, let's explain the construction of the frequency matrix that uses the frequency presented in Figure 3. Here, the direct tracking is observed. This matrix (second matrix in Figure 4) is read from the vertical column to the horizontal line. For example, from the activity "a" to activity "a" there is anything; but, from activity "a" to activity "b", it occurs 6474 times (3655 x in trace 1, 1452 x in trace 2 and 1367 x in trace 3 => 3655 + 1452 + 1367 = 6474). It is made to all possible combinations. Now,

explaining the significance matrix (the first one in Figure 4), after constructing the frequency matrix and observing the results, mathematical calculations are made and the results are generated. Let's use the marked example in the Figure 4: To calculate the significance from "a" to "b", first we take the result on frequency matrix from "a" to "b" and its reverse "b" to "a". So, applying the formula descript in Figure 4, we can find the significance result from "a" to "b". It is made to all possible combinations too.

Figure 5 depicts a complete example of application of the Heuristic Miner Algorithm. So, first we have the traces of the log represented by "L". Each trace "< ... >" here is above followed by the frequency on the log. Then, by mapping the traces of the log the matrix of frequency is built. After that, the matrix of significance and the process model are generated. The process model presented is a heuristic network that is composed of nodes and edges. The nodes represents the activities (in this case, the activities are: a, b, c, d and e) and edges represents the transitions between activities, so, if “a” is directly followed by “b” in one or more traces, there is a transition between “a” and “b” - an arrow that goes from “a” to “b”. The frequency and significance are next to the transitions; the second is presented in parentheses.

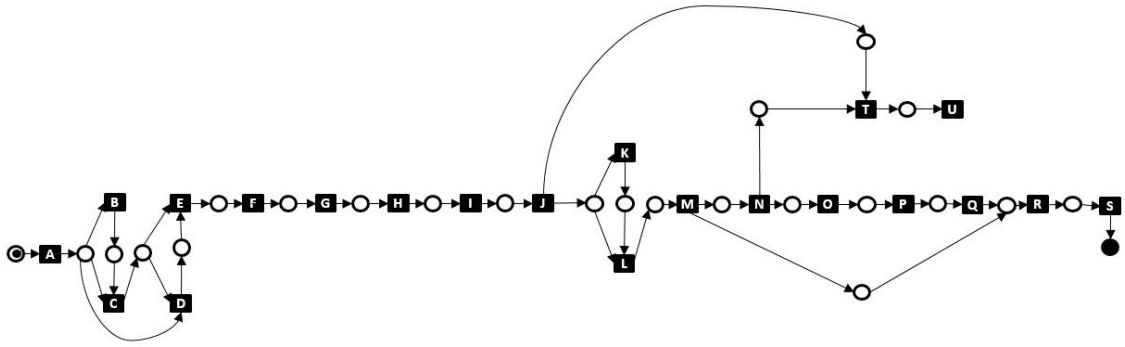


Figure 5 - Complete application of the heuristic miner algorithm (Control Flow Process Discovery Presentation)

3.3 Disco Miner Algorithm

The Disco software uses the Disco Miner[13] to generate processes models - as shown in Figure 6 - to logs. Is based on Christian's Fuzzy Miner that was the first one to introduce the "map metaphor" (that are advanced features like process simplification and highlighting of frequent activities and paths). It was combined with user testing and an extensive experience of their own practice. It is also suggested to be used when the log data is complex and unstructured, or when it is desired to simplify the process model in an intuitive way. Besides that, a new set of process metrics and modeling strategies were implemented on it. Disco base algorithm focuses on simplifying the process model at the desired level of abstraction, which means that it can eliminate or hide activities that might be considered less important than others, basing on a significance/correlation metrics.

The priorities on development were usability, fidelity to the data and performance. To explain Figure 6: the process starts in the green circle and ends in the red circle. The activities are represented by the rectangles and the transitions by the arrows - about these two: the darker the more often it happened (map metaphor). The frequency of the activity is inside the rectangle and the frequency of the transition is next to the arrow.

[13] https://fluxicon.com/disco/files/Disco-Tour.pdf

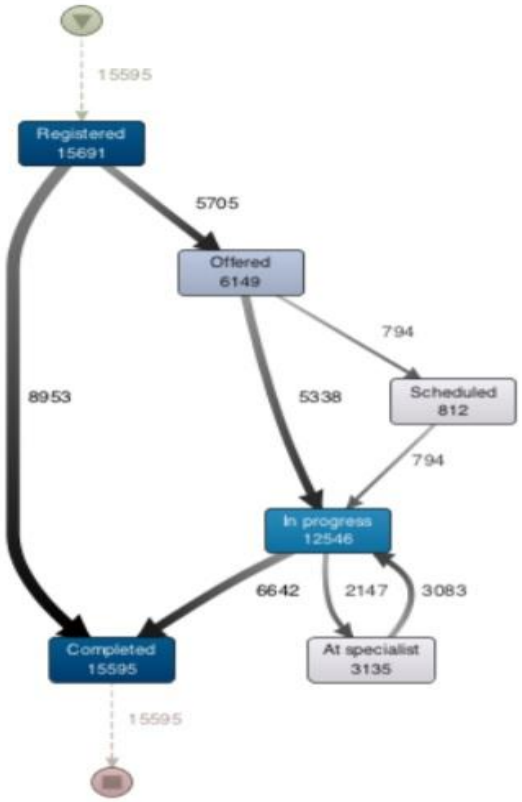


Figure 6 - Example of the process model generated by disco miner algorithm (Fluxicon)

3.4 "Celonis Miner Algorithm"

Celonis software uses advanced algorithms to reconstruct and visualize processes end-to-end[14]. It does not have a specific name, but it is also based on Fuzzy Miner. It works in a very similar way to Disco. It is able to handle large event logs efficiently and can work with real time process discovery technique while information systems are in operation. In Figure 7, there is an example of the process model generated by Celonis that uses the same logic to draw processes. But the activities are represented by rounded rectangles and the start and the end by hexagons each one with their descriptions.

[14] https://www.celonis.com/

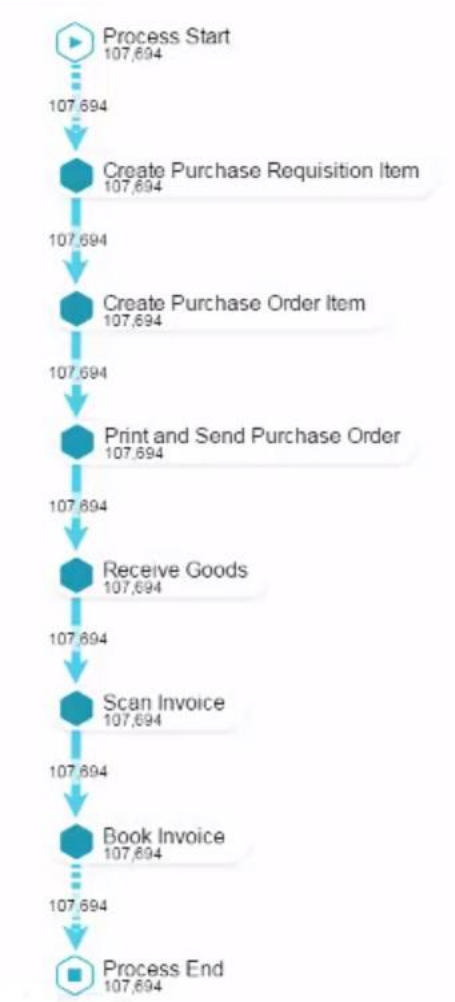


Figure 7 - Example of the process model generated by Celonis miner algorithm (Celonis)

3.5 Applications used

Celonis[15] is a process mining application. This tool is very visual and intuitive. It was helpful to analyze, explore and visualize the process in general and with filters applied. Celonis focuses on qualitative analysis approach of the logs, providing the user a general overview, according to the filters applied.

[15] https://www.celonis.com/

Disco[16] is a process mining application. We used a set of technologies available in this software. Basically from the tools actions we apply an automated process discovery, filtering and analyzing cases and detailed statistics. Each step taken using this tool was explained in the answers.

Prom[17] is a process mining software that has an extensible framework which supports a wide variety of process mining techniques in the form of plug-ins. Each plug-in represents a different type of mining algorithm. For the study developed in this paper, it was necessary to download the Heuristic Miner algorithm plug-in.

Excel[18] It was used to generate graphs according to the results, helping with visualization.

Paint[19] Editions of some images were made with this tool in order to improve the original ones.

[16] https://fluxicon.com/disco/
[17] https://fluxicon.com/disco/
[18] https://products.office.com/en-us/excel
[19] The software Paint is part of Windows.

4 Analyses

4.1 Analyzing Throughput Times per Part of the Process

Overview:

The question regards the time spent in the processing per parts, in particular, the difference between the time spent in the company systems in standby waiting for the user processing and the time spent by an applicant input.

Investigation:

To reach a conclusion, at first which events belonged to the application log and by whom they wait (customer or system/internal bank user). In a topic[20] of the BPI Challenge 2017 category in the Eindhoven University of Technology PROM FORUM the following data was observed:

Submitted: a customer has submitted a new application from the website. A new application can also be started by the bank, in that case this state is skipped.

Concept: the application is in the concept state, that means that the customer just submitted it (or the bank started it), and a first assessment has been done automatically. An employee calls the customer to complete the application.

Accepted: after the call with the customer, the application is completed and assessed again. If there is a possibility to make an offer, the status is accepted. The employee now creates 1 or more offers.

[20] http://www.win.tue.nl/promforum/discussion/755/brief-event-descriptions

Complete: the offers have been sent to the customer and the bank waits for the customer to return a signed offer along with the rest of the documents (payslip, ID etc).

Validating: the offer and documents are received and are checked. During this phase the status is validating.

Incomplete: if documents are not correct or some documents are still missing, the status is set to incomplete, which means the customer's needs to send in documents.

Pending: if all documents are received and the assessment is positive, the loan is final and the customer is payed.

Denied: if somewhere in the process the loan cannot be offered to the customer, because the application doesn't fit the acceptance criteria, the application is declined, which results in the status'denied'.

Cancelled: if the customer never sends in his documents or calls to tell he doesn't need the loan, the application is cancelled.

As exposed above, the conclusion is exposed in Table 1.

Table 1 - Activities Standby

CUSTOMER	SYSTEM
Cancelled	Accepted
Complete	Concept
Incomplete	Denied
Submitted	Pending
	Validating

Based on Table 1, it is possible to observe that few events from the log are missing. So, it means that we do not have the explanations about the others activities that were found in the log. With that said, the others definitions constructed and considered for us are:

Other definitions:

Regarding the offer status change:

Create - Create proposal

Created - Proposal created

Refused - Proposal rejected

Returned - Proposal returned

Sent - Proposal sent

Regarding the event flow:

Assess potential fraud

Call after offers

Call incomplete files

Complete application

Handle leads

Personal loan collection

Shortened completion

Validate application

Analyzing the remaining activities and their definitions, the Table 2 is created based on what we could observe by the processes and we think they could be. Is important to remember that if something was not well done here in this table, it could change some results found in this analysis.

Table 2 - Remaining Activities

CUSTOMER	SYSTEM
Refused	Create
Returned	Created
	Sent
	Assess potential fraud
	Call after offers
	Call incomplete files
	Complete application
	Handle leads
	Personal loan collection
	Shortened completion
	Validate application

After that, specifically the activities average times and the standby average time by another activity so that its procedure can begin, were analyzed. All the procedure was realized using filters in the Disco Software.

1 - Evaluating the activities average time:

It was observed the time spent by the activities in the models - time without dependence on another activity.

In the software Disco - it was chosen because of the fact that it is easy to make the filters and see the draw results - and as said before, using the application log, the activity analyzed is set as an attribute at the attribute filter, so only the cases that contain the activity can be seen.

Then, we set the paths in 100% in order to visualize all the cases that contain the activity analyzed. And thus, see how much time the activity takes - it can be noted inside the rectangle between parentheses.

It was created for all the activities presented in the log.

In the example below - figure 8, the “W_Call after offers” (that is when a new contact is made with the customer) is being analyzed (in the attribute filter “W_Call after offers” was set as attribute and the paths were set in 100% - as explained before) and it was detected that it takes in average 23.4 minutes from the beginning until its conclusion (analyzed in red).

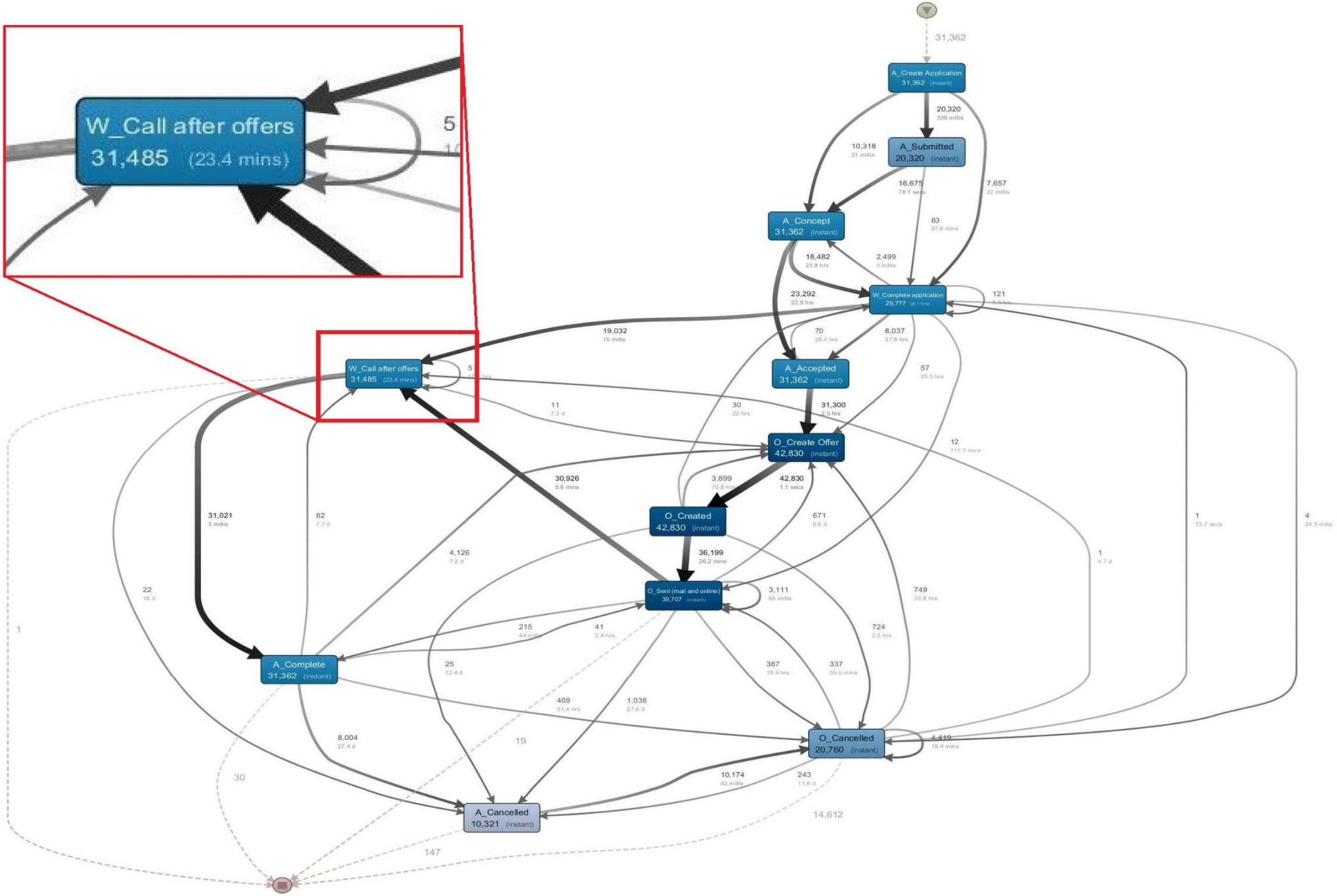

W_Call after offers
31,485 (23.4 mins)
31,362
A_Create Application
31,362 (instant)
20,320
339 millis
A_Submitted
20,320 (instant)
10,318
21 millis
16,675
78.1 secs
7,657
22 millis
A_Concept
31,362 (instant)
83
97.8 mins
18,482
23.8 hrs
2,499
5 millis
23,292
23.9 hrs
W_Complete application
29,777 (6.1 hrs)
121
19,032
15 millis
70
28.4 hrs
8,037
37.8 hrs
W_Call after offers
31,485 (23.4 mins)
A_Accepted
31,362 (instant)
57
25.5 hrs
11
7.3 d
30
22 hrs
31,300
2.5 hrs
O_Create Offer
42,830 (instant)
12
117.7 secs
30,926
5.8 mins
3,899
70.8 mins
42,830
1.1 secs
31,021
3 millis
82
7.7 d
O_Created
42,830 (instant)
671
6.6 d
1
73.7 secs
4
24.5 mins
4,126
7.2 d
36,199
26.2 mins
1
4.7 d
22
18 d
O_Sent (mail and online)
39,707 (instant)
3,111
65 millis
749
20.8 hrs
1
215
44 mins
41
2.4 hrs
724
2.5 hrs
A_Complete
31,362 (instant)
25
12.4 d
387
18.9 hrs
337
35.5 mins
469
51.4 hrs
1,038
27.6 d
8,004
27.4 d
19
O_Cancelled
20,760 (instant)
4,419
18.4 mins
30
10,174
42 millis
243
11.8 d
A_Cancelled
10,321 (instant)
14,612
147

Figure 8 - W_Call after offers average time

2 - Analyzing the standby average time by another activity so that the analyzed activity can begin its procedure:

For this analysis, Disco was used for the same reason - because it is easy to make the filters and see the draw results.

At first, we filtered by attributes: the activity analyzed is set as attribute, so only the paths that contain the activity can be seen. Besides that, we set the paths in 100% in order to visualize all the paths that lead to the activity analyzed. And thus, see how long it takes between waiting for one activity to another. It was also made for all the activities presented in the log. For example, in Figure 9 “O_Returned” (that is when the status of the application returns to the bank) is being analyzed and is possible to see that it waits “A_Incomplete” (that is when the bank is waiting for the customer to answer/send documents) - marked in red with number 5 - 1,167 times and this wait lasts, on average, 31 hours.

In the example below - figure 9, the activity “O_Returned” is analyzed. The incoming activities, along with the frequencies and its times until “O_Returned” can occur. In this case, seven predecessors activities regarding the one that is being analyzed. So, each one of these predecessors were observed.

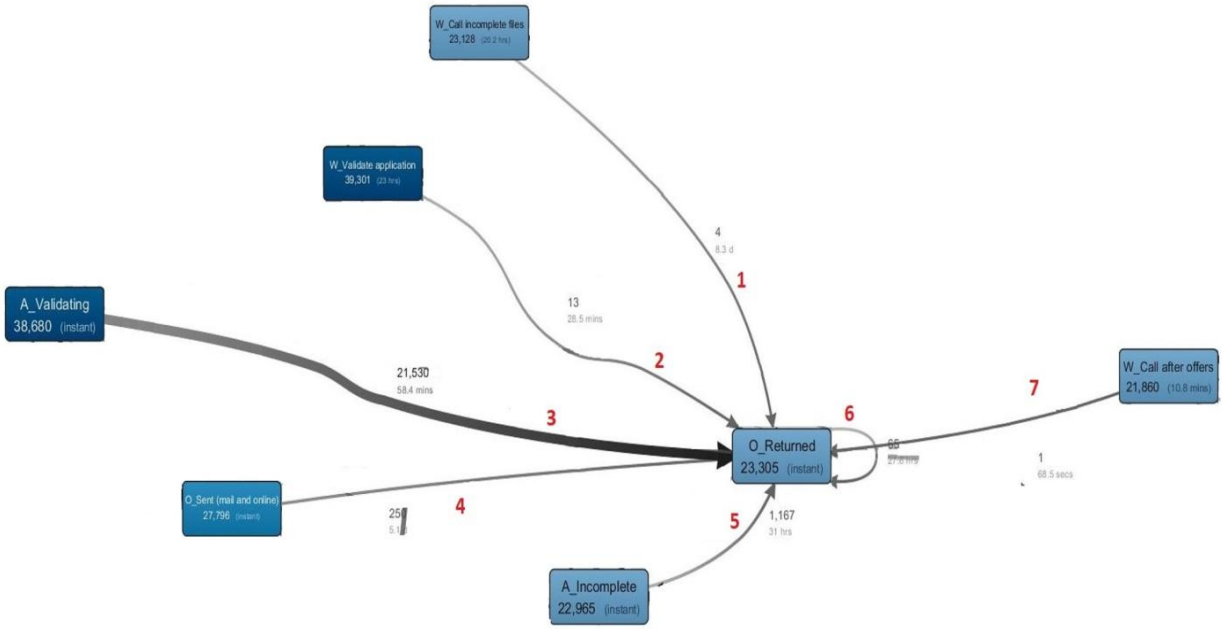


Figure 9 - O_Returned standby average time by another activities

With the results found in the tools, we analyzed some categories of information:

Total Average Time Activity - Shows how much time the analyzed activity takes from its beginning to its end - independent of others activities;

Standby Average Time From Another Activity - Shows how much time the analyzed activity waits for until it can begin, after the end of an incoming activity. Specified for each of the incoming activities;

Total Occurrences Regarding the Standby Average Time - After analyzing the standby times from other activities, it was verified the frequency from each analysis from an incoming activity;

Total Time (approximate value) - The total time multiples the two previous results for each incoming activity analysis. Besides that, all the values were set in hours, and for this reason an approximate value is considered;

∑ Total Time / ∑ Total Occurrences (approximate value) - This category regards the division of the sum of the total time values (previous category) by the sum of the occurrences values (third category), analyzing the total time from all the incoming activities and dividing by the total activities frequency. And finding a general average value that could be compared with the others.

Approximate values are considered as well.

All the information found by the analysis, that can be seen in the Appendix, returned a lot of responses that can be seen below in the results section.

Results:

Regarding the activities time:

As the study shows, it can be affirmed that the only activities that have a significant time are: W_Assess Potential Fraud, W_Validate Application, W_Call Incomplete Files, W_Complete Application, W_CallAfter Offers e W_Handle Leads; ranging from the highest to lowest time (from days to minutes) respectively, as shown in Figure 10.

The cases where the Total Average Time Activity is the only analyzed must not be considered because they do not represent the reality of the events in the whole process. Those cases simply do not represent the workflow reality. There will be always a natural standby time by some activities, no matter if this time is too low or too high, but it will never be instantaneous.

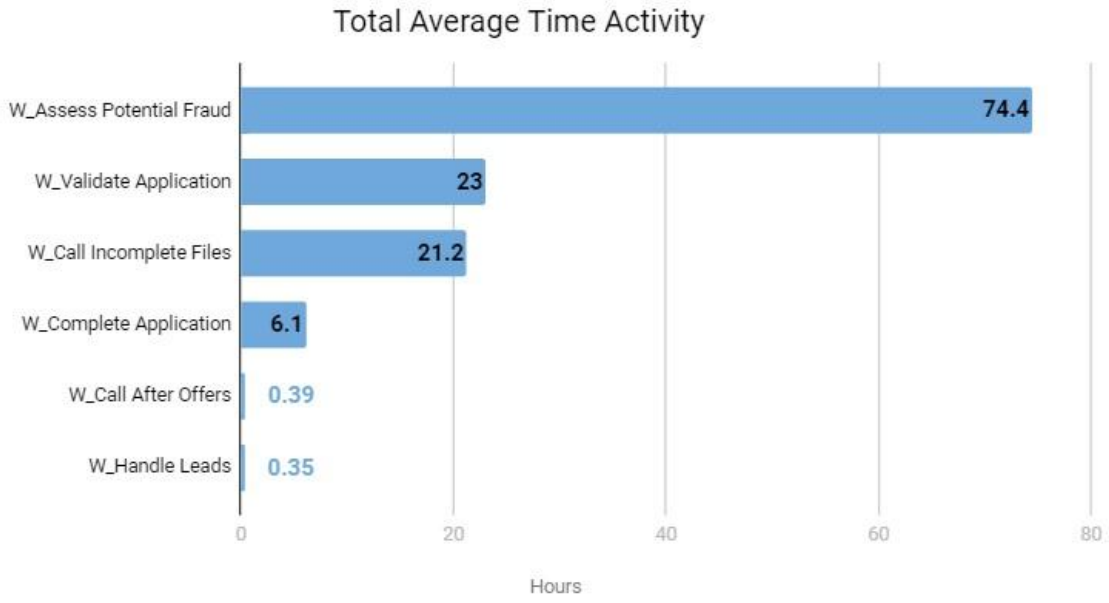


Figure 10 - Total Average Time Activity

Regarding the activity standby time:

Observing the results found - each colored trace in Figure 11 and 12 shows a result found for the respective activity, if the general average of all activities from each category is done, we found a customer waiting average time of 91.74 hours - see the activities results in Figure 11 that are shown from the activity that more time waits for customers to the activity that lasts wait for customers - and the system waiting average of 25.76 hours - see the activities results in Figure 12 that are shown from the activity that more time waits for systems to the activity that lasts wait for systems. So, it can be concluded that delay per customer is a lot greater than the waiting per system.

However, it must be emphasized the fact that an activity in particular has a waiting per customer quite significant. It is the "A_Cancelled - complete" activity. If this activity is not considered, the average customer waiting would dramatically decrease to about 2 hours. Looking at the median, we can observe that the value we find as a result is 2.12 hours. This activity has a waiting time really high compared to the others, both the ones that wait for the customer, and the ones that wait for the system. For this reason, a deeper study should be done to understand why this time is so high.

Figure 11 - Time Waiting For Customers

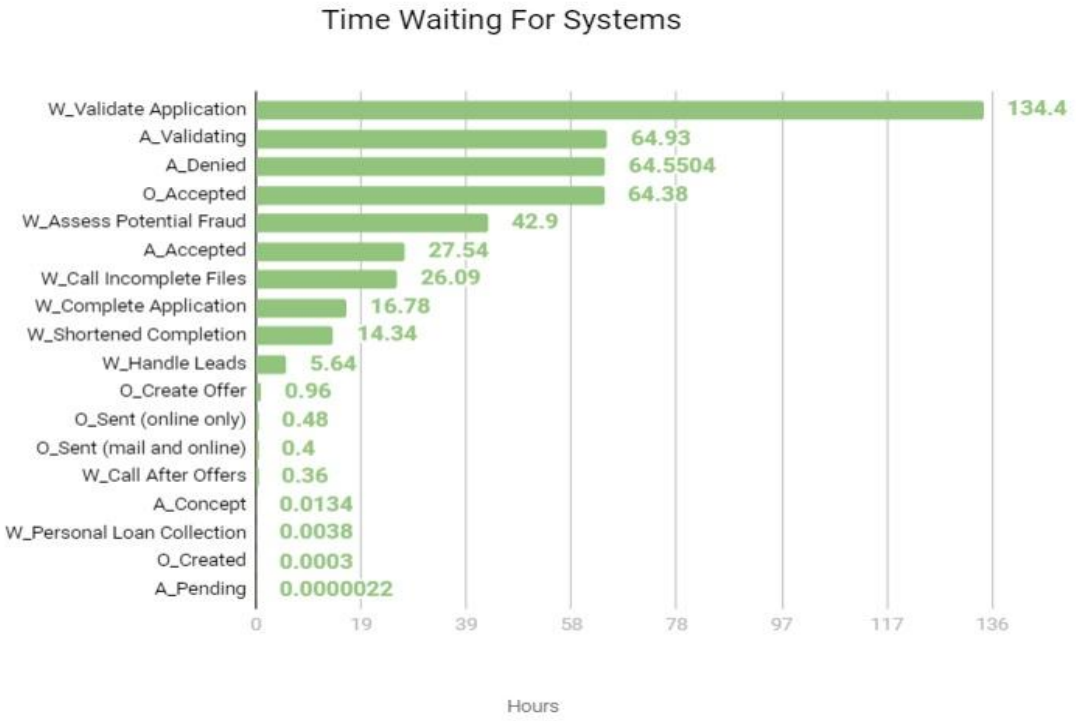


Figure 12 - Time Waiting For Systems

4.2 Analyzing the frequency of incompleteness

Overview:

Evaluating the applicants that are confronted with more requests to the completion and the hypothesis that they should be more inclined to not accept the final offer, the following facts were observed:

Verifying the offers log and using Disco that provides a good draw for processes and without applying any filters, we have what is observed in Figure 13:

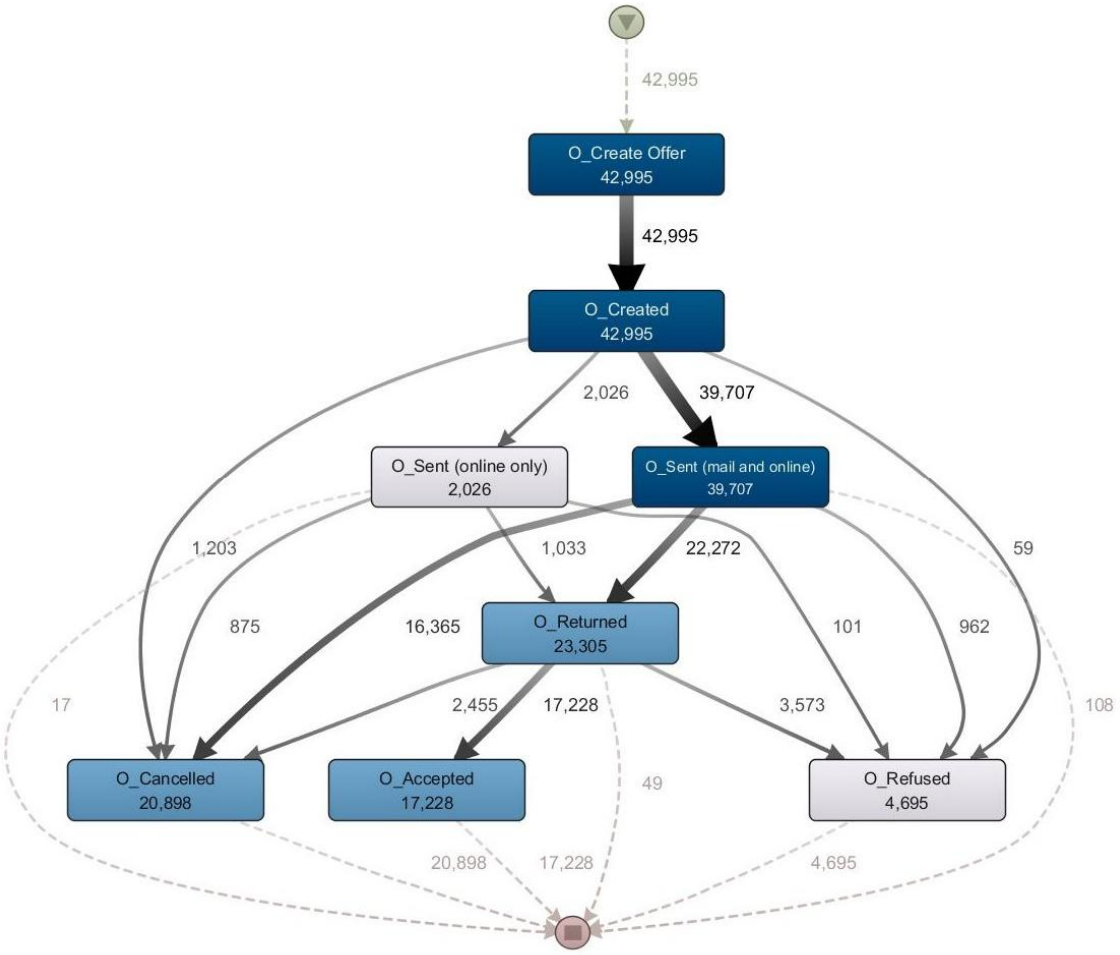


Figure 13 - All offers

Here it was observed the whole log, so we could see the whole process. From a total of 42,995 offers: 20,898 were Cancelled, 17,228 were Accepted and 4,695 were Rejected.

Investigation:

By the hypothesis given we verified the cases in which, in order for the offer to be made, there were more requests to the customer.
The mandatory requests were disregarded - those that all cases must change at some point, from the offer made - only the cases where, after the company's validation, were verified and recorded that items were missing and / or answered by the customer. That is, the cases in which there was the need to change the application's change status to

“A_Incomplete-complete”.
For this it is necessary to verify in the other file given (the application log), because it is where that status change is indicated and it is marked in this log.

Besides that, search for the final states:

– O Accepted - complete - To evaluate the cases where the proposal was accepted despite the requests;
– O Cancelled - complete - To evaluate the cases where the documents were not sent and the cases in which the customer did not need the loan anymore and
– O Refused - complete - To evaluate the cases where the proposal was rejected by the customer.

Analyzing the application log file and using Celonis that has a fast and efficient manner to show numeric results of flows, the following data were found - Figures 14, 15, 16, 17, 18 and 19. In Celonis, we used the "Activity Selection", and the "Case Flows Through" selection with the attributes required, so we could get the searched results.

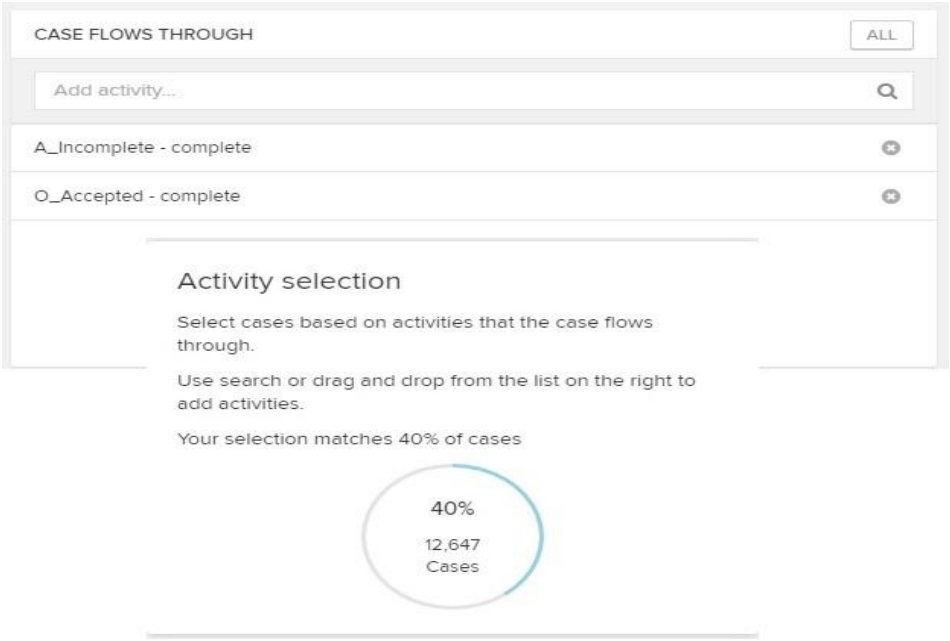


Figure 14 - The cases in which the customer accepts the proposal, despite the multiple requests, total 12,647 cases

In Figure 14, we were searching for how much cases that passed for incompleteness and in the final were accepted. The attributes used were: "A_Incomplete - complete" and "O_Accepted - complete".

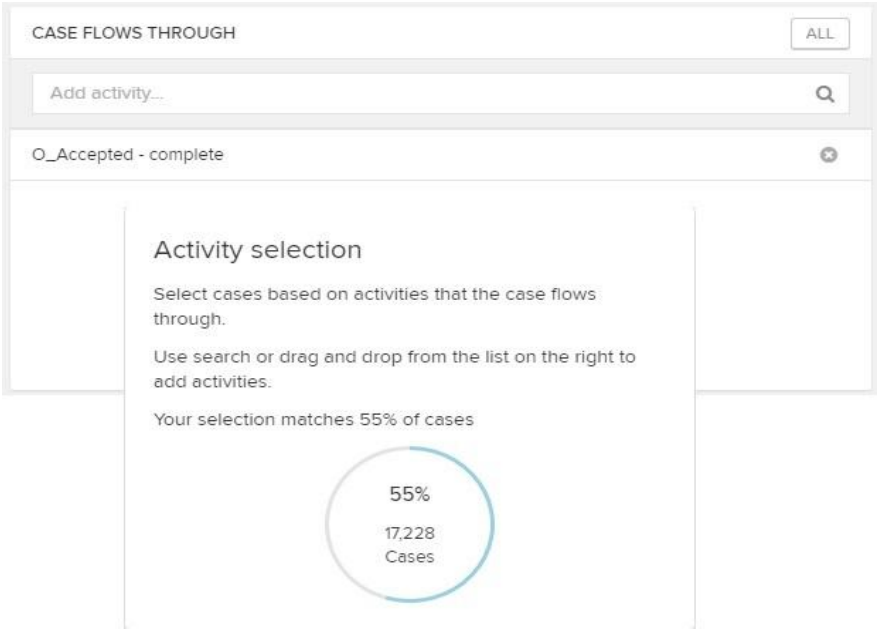


Figure 15 - Altogether 17,228 were accepted, including or not more requests

In Figure 15, we were searching for how much cases that were accepted. The attribute used was: "O_Accepted - complete".

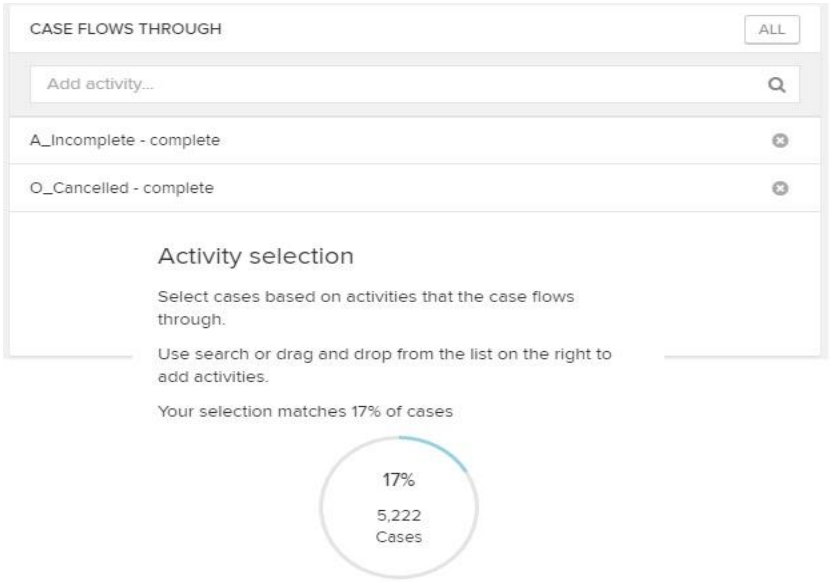


Figure 16 - The cases where the proposal were cancelled and there were multiple requests total 5,222

In Figure 16, we were searching for how much cases that passed for incompleteness and in the final were cancelled. The attributes used were: "A_Incomplete - complete" and "O_Cancelled - complete".

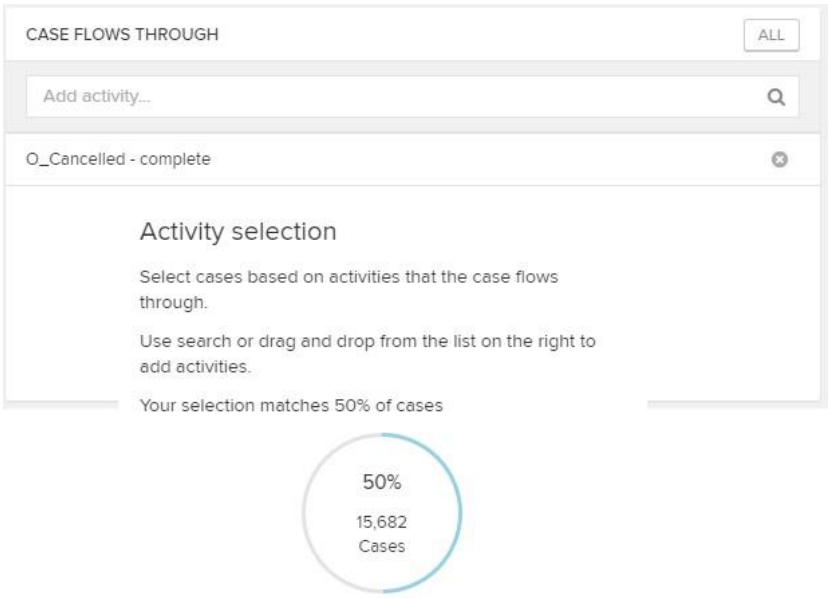


Figure 17 - Altogether 15,682 were cancelled, including more requests or not

In Figure 17, we were searching for how much cases that were cancelled. The attribute used was: "O_Cancelled - complete".

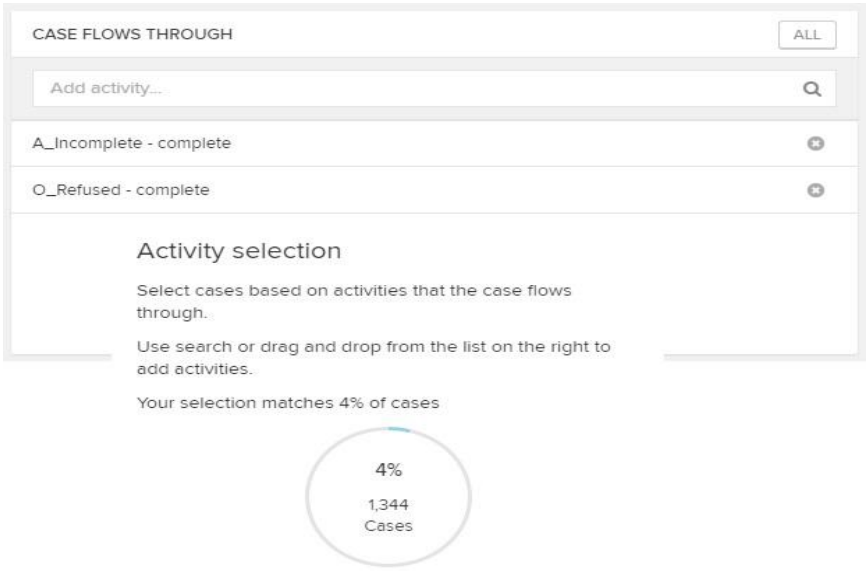


Figure 18 - The cases where the proposal is refused after multiple requests total 1,344 cases

In Figure 18, we were searching for how much cases that passed for incompleteness and in the final were refused. The attributes used in Celonis were: "A_Incomplete - complete" and "O_Refused - complete".

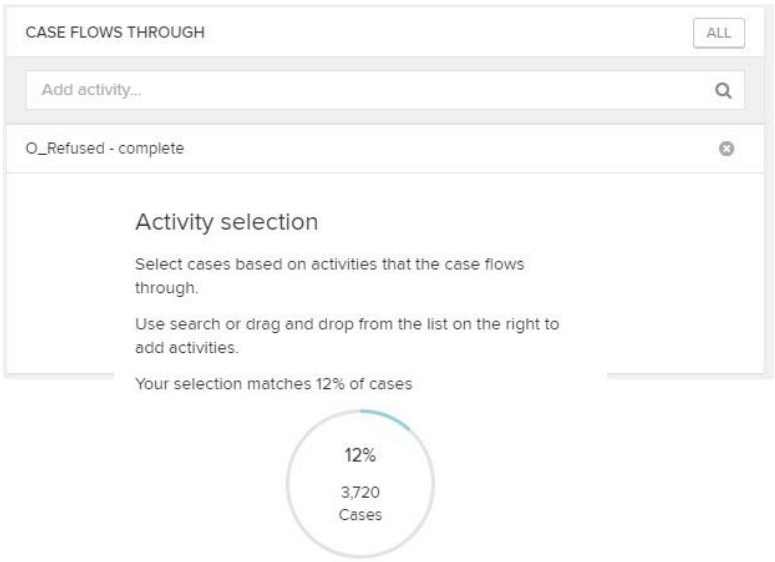


Figure 19 - Altogether 3,720 were refused, including more requests or not

In Figure 19, the goal of the search was to find how many cases were refused. The attribute used was: "O_Refused - complete".

Results:

Based on the data presented above - Figures 14 to 19, the following results - Table 3 - were found. So, percentages were calculated based on the total and the cases with more requests.

Table 3 - Table with consolidated results

EVALUATED CASE	TOTAL	TOTAL WITH MORE REQUESTS (+ A_Incomplete - complete)	PERCENTAGE OF CASES WITH MORE REQUESTS (APPROXIMATED VALUES)
O Accepted - complete	17,228	12,647	73%
O Canceled - complete	15,682	5,222	33%
O Refused - complete	3,720	1,344	36%

Figure 20 - Cases with more requests compared with its total

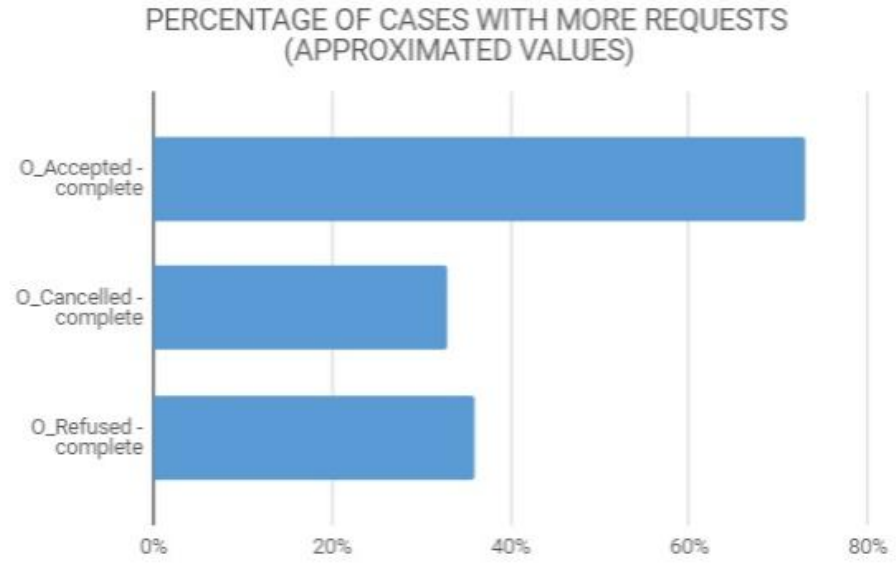


Figure 21 - Percentage of requests in cases

Observing the results depicted in - Figures 20 and 21 (that are shown the results of Table 3 in a graphic way to a better visualization of the results), it can be identified that applicants which received more requests had a high acceptance percentage to the final conclusion of the offer. Therefore, we concluded that the hypothesis raised is not true.

More Analysis:

We can also see that almost 66% of all the cases with more requests were accepted. And that the percentages of the cases with more requests cancelled or refused are quite similar.

4.3 Analyzing offers, conversations and conversions

In the attempt to answer the third question, we decided to split it in two parts for a better understanding after looking at the forum about the challenge[21]. The used log to reach the answers to follow was the application log file. It is worth remembering that it has 31509 cases, 4047 variants and all the cases have offers.

Overview - First part:

The first part of the third question asks how many customers ask for more than one offer and if these offers are made in one or more conversations.

Investigation - First part:

Scenario 1 - Evaluating how many customers ask for more than one offer:

In this analysis, we used Disco for a better visualization of the number of cases and variants. Just applying the filters, we already have these numbers as results. Here, cases are the traces of events that occurred and the variants are the cases with distinct traces. For each evaluation, we applied a filter.

Filtering the log by “Follower”, where the activity “O_Create Offer” ("Reference event values") was followed in some point by another “O_Create Offer” ("Follower event

[21]http://www.win.tue.nl/promforum/discussion/comment/2236/#Comment_2236

values") activity and the same resource was required. So, we have all the cases where there were two or more offers per application, once that an activity "O_Create Offer" signalizes that an offer was created. Note on Figure 22:

Figure 22 - Customers who asked for more than one offer

The result shows 4449 cases and 1729 variants that are cases with different traces, what we mean is that cases with equals traces are "combined", "seen" as one same variant.

Scenario 2 - Evaluating if the offers occurred in one or more conversations:

After what was done in scenario 1, it was also verified if the offers were made in one or more conversations. To check if more than one conversation had occurred, we focused on the “W_Call after offers” activity.

Thus, another “Follower” filter was made in the log - besides the filter made in scenario 1- , where the “W_Call after offers” ("Reference event values") activity was followed at some point by an “O_Create Offer” ("Follower event values") activity and where the same resource was required. This shows that new offers had been made after a new conversation. A new conversation is marked with the activity "W_Call after offers". This can be seen in the Figure 23:

Figure 23 - Offers that occurred in one or more conversations

The result shows 850 cases and 501 variants.

Overview - Second part:

The second question concerns about the conversion that occurs when the case in which the "A_Pending" status is reached, and the proposal receives the payment release. Besides, it can be related to the behavior which regards the cases where only one offer is made and with respect to cases where more than one offer is made.

Investigation - Second part:

Scenario 3 - Evaluating the cases that came to conversion:

Considering that all the cases that arrived at the conversion state, from the "Attribute" filter by "Activity", by the "Mandatory" mode with event value "A_Pending", as shown

in Figure 24. What signalizes that the process get into the conversion (the payment is released) state is the event value "A_Pending".

Figure 24 - Cases that came to conversion

17228 cases and 2575 variants were found.

Scenario 4 - Evaluating the cases that came to conversion with more than one offer:

Besides what was made in scenario 3 to filter, now, the cases with more than one offer were analyzed using the filter of the scenario 3 and the filter “Follower” in the log, where an activity “O_Create Offer” ("Reference event values") was followed in some point by another “O_Create Offer” ("Follower event values") activity and where the same resource was required, note Figure 25.

Figure 25 - Cases that came to conversion with more than one offer

2309 cases and 1092 variants were found.

Scenario 5 - Evaluating the cases that came to conversion with more conversations:

We also decided to analyze the cases where the offers were made from new conversations, using the filters of the scenarios 3 and 4 and also using the “Follower” filter in the log as well, where the activity “W_Call after offers” ("Reference event values") was followed, in some point, by another activity “O_Create Offer” ("Follower event values"), and where the same resource was required - Figure 26:

Figure 26 - Cases that came to conversion with more conversations

491 cases and 319 variants were observed.

Results:

After all that was seen, Table 4 and Figure 27 were generated:

Cases with conversation followed by an offer creation: those are the cases where there is an activity “W_Call after offers” followed, in some point, by a activity “W_Call after offers”.

Table 4 - Table with results about offers, conversations and conversions

RESTRICTIONS	ALL CASES	CASES WITH MORE THAN ONE OFFER
Cases with no restrictions	31,509	4,449
Cases with conversations followed by an offer creation	856	850
Cases that reached the "A_Pending"	17,228	2,309
Cases with conversation followed by an offer creation and that reached the "A_Pending" status	494	491

Table 4 shows the results found, comparing the cases in total in each restriction and the cases with more than one offer in each restriction.

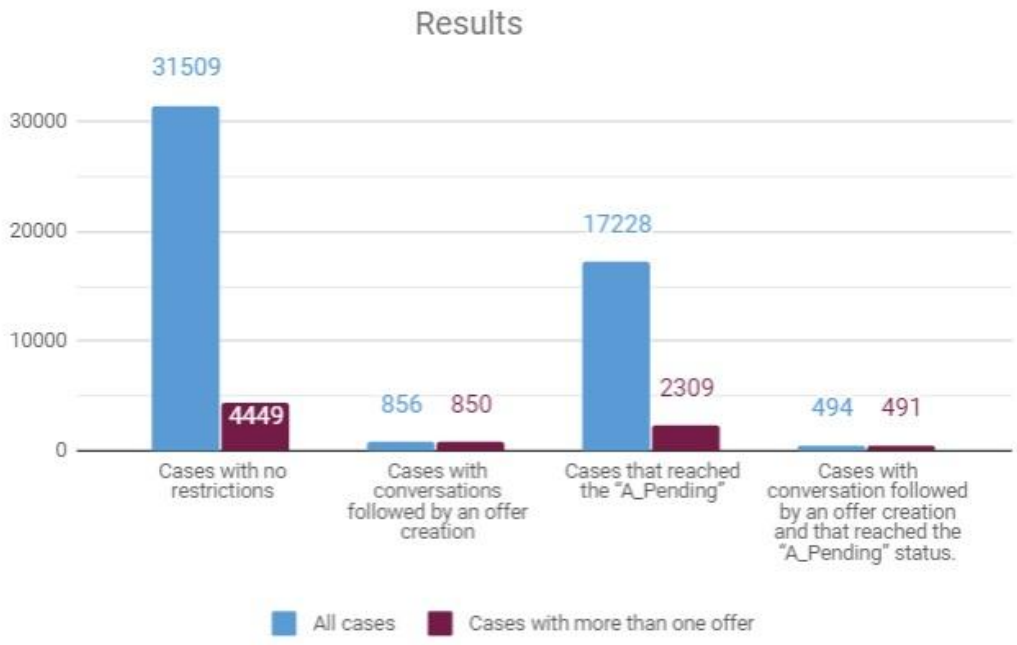


Figure 27 - Visualizing the results

The Figure 27 shows what was seen in Table 4 in a graphic way for a better visualization and comparison of the results.

About the first part and question, we can conclude that most of the cases 85.88% -

received just one offer.

Besides, when there is a conversation, in 99.30% of the cases there are more than one offer compared with the cases with just one offer.
We can see that, comparing cases with more than one offer, cases with conversation(s) correspond to 19.10%.

About the second part and question, we can observe that 54.68% of all the cases turn into a conversion.
Then, we can note that less than 14% of the cases that turn into a conversion receive more than one offer.

Further Analysis:

We can also conclude that almost 100% of the cases with conversation(s) followed by the creation of an offer and that turn into a conversion receive more than one offer.
Note that about 51.90% of the cases with more than one offer turn into a conversion.
The 57.71% of all the cases with conversation(s) followed by the creation of an offer turn into a conversion.
And the 57.76% of the cases with conversation(s) followed by the creation of an offer and with more than one offer turn into a conversion.

4.4 Additional Discoveries

Considering that only 40% of the offers were accepted, the waiting time by the customer is one of the main reasons for quitting. Therefore, we tried to analyze a customer pattern behavior regarding the offer acceptance and the time for the process conclusion.

It was identified that most of the offers that were made in a few contacts with the customer, or in only one, were the ones that had more acceptance. The values that were

most requested in a contact were $10,000.00, $5,000.00, $15,000.00 and $20,000.00.

Assuming that the customer is approached with the loan service (or credit limit increase - minority of the cases), without any explicit quantified offer (11,05% of the cases in which the offers were accepted), and this in turn, when interested, requests the desired amount, which is accepted by the bank without any further negotiations.

Based on what has been reported before, we have come to the conclusion that the core business of the process is the sale of loan services estimated in the range of $5,000.00 - $20,000.00.

The creation of several offers means a negotiation process with the customer, to reach to a common denominator regarding the value to be available. However, we realized that if the customer received an offer with a lower value than the one requested, the tendency is for the offer to be canceled. In the minority of the cases where the offered value was higher than the one requested by the customer, the offers also ended up being canceled.

When analyzing the number of contacts made with the customers, it was also observed that the activities directly related to the required documents validation process, when not complete for more than one cycle, led to the customer giving up, even when his request was answered immediately.

As indicated above, some workflows activities have a high duration range, and in turn, when they reach their peak, directly impact negatively on the offer acceptance by the customer. It was verified that in most cases where the “W_Asses Potential Fraud” and “W_Validate application” activities have an abnormally high execution time, bidding is canceled. The “W_Validate application” activity may cause more impact, as it may occur more than once during the process.

After all the analyzes carried out on the proposed log, we concluded that the subprocesses of the activities that take the longest time in the process as a whole, should be reviewed in order to drastically reduce their negative impact on the acceptance of offers by customers. It should also be taken into consideration, if indeed such activities are extremely important to the completion of the process and if they are being performed in the most correct and efficient manner.

5 Conclusions

5.1 Conclusions about the results obtained

Through the use of the tools Disco and Celonis in the analysis of the proposed log, it was possible to answer the questions made by BPI Challenge 2017 succinctly. With Disco, it was possible to generate several scenarios of activity flows through the use of filters in the log. Yet, Celonis allowed a quantitative analysis of the information collected in the flow scenarios drawn.

Due to the absence of information concerning the main procedures, the in-depth analysis carried out was limited. In order to proceed with the analysis, it was necessary to carry out the investigation of certain information that was identified by patterns in the possible flows. Thus, based on the evidence found, some analyzes were oriented from these deductions.

In this study, in addition to identifying which of the activities offered by the financial institution in question represents the “heart of the business” of the company, we also succeeded in highlighting the main indicators that contribute to the poor performance of the conclusion of offers, thus proposing a critical review relative to the performance of such activities.

Considering an increasingly globalized world with enormous amounts of information, with this study it is possible to note the relevance of process mining; Since from the past data we were able to extract a great amount of information and results for the company and, noting that there would be possibility of further studies.

5.2 Final Considerations

In the analysis, the business rules, of the company that provided the logs, are not considered. The company did not give it. And, it is important to say that if it was taken into consideration, we could conclude more. At some parts of the analysis, we catch information in the forum category[22] opened to the challenge and we deduce some others. If the company applied the business rules in the analysis we made, maybe they could see other aspects about the results that they would like to find or maybe they could do deeper analysis or other questionings with the results that we found and see if the questions, that they are worried about, are the important ones.

5.3 Project Limitations

Basically, the limitations in this project are about information.
First, the logs were not so explained in the challenge and we have to search and try to understand a lot of information.
Second, the company provided the logs in an anonymous way. So, if we had some knowledge about the company, we could do deeper analysis and/or deeper conclusions.

Some of the analyses built in this paper were made from assumptions that were considered by us as obvious facts, by only paying attention to patterns in the log. That had to be done, because some information were missing in the log provided by the financial institution.

This was observed in the cases where the customer was approached with an offer, without any previous contact, or at least that was not reported in the log. As analyzed before, those were the cases that had more acceptance by the customers. Therefore, this scenario brings questions such as: "Was the customer chosen randomly?", "Had the customer already made business with the financial institution before?", "Was it verified

[22]http://www.win.tue.nl/promforum/categories/-bpi-challenge-2017

if the customer had any depts?" The answers to those questions could bring more insights to the process improvement.

5.4 Future Work

During the analysis of the main activities in this process flux we could identify as the main key process indicators the following:

-Reduction of the customer waiting time;
-Find out more about why proposals that receive more than one offer and turn into a conversion, in most cases don't reach to an end of the proposal conclusion;
-Invest in all the activities that contribute to the proposal acceptance by the customer, when it is approached without any explicit quantified offer and this in turn, when interested, requests the desired amount, which is accepted by the bank without any further negotiations. In other words, the institution should focus more on the activities workflow process scenarios where the probability to reach the desired end is higher.

Focusing on the cases where there where the activities regarding the documents verification process occurred too many times (too many loops), the analysis could be oriented to verify if there was an attempt of document fraud.
In order to make performance improvements in the AS IS business process model, the Genetic Miner algorithm could be applied to event log, thus generating the ideal process model thought recombination and mutation of parts of the existing process.
If the AS IS business process model is known, it should be possible to make conformance checking analysis.

References

van der Aalst, W.M.P.: “Process Mining: Data Science in Action”. 2nd ed. Springer Verlag, ISBN 978-3-662-49850-7 (2016). 467 p.

van der Aalst, W.M.P., Mans, Ronny, Vanwersch, Rob: “Process Mining in Healthcare: Evaluating and Exploiting Operational Healthcare Processes”. Springer International Publishing, ISBN 978-3-319-16070-2 (2015). Abstract.
https://link.springer.com/chapter/10.1007%2F978-3-319-16071-9_3

van der Aalst, W.M.P.: “Process Mining: Discovery, Conformance and Enhancement of Business Processes”. Springer Verlag, ISBN 978-3-642-19344-6 (2011) Page 99.

van der Aalst, W.M.P.: “Process Mining in the Large: A Tutorial ”. In E. Zimanyi, editor, Business Intelligence (eBISS 2013), volume 172 of Lecture Notes in Business Information Processing, pages 33-76. Springer-Verlag, Berlin, 2014.
http://www.processmining.org/_media/blogs/pub2014/p775.pdf

van der Aalst, W.M.P., van Hee, Kees M., van Werf, Jan Martijn and Verdonk, Marc: “Auditing 2.0: Using Process Mining to Support Tomorrow’s Auditor”. Vol. 43, No. 3, Pages 9093. IEEE Computer (March 2010).
http://www.processmining.org/_media/blogs/pub2010/ieeecomputer-wvdaalst.pdf

International Workshop on Business Process Intelligence, 13th, 2017:
https://www.win.tue.nl/bpi/doku.php?id=2017:challenge

International Workshop on Business Process Intelligence, 8th, 2012:
http://www.win.tue.nl/bpi/doku.php?id=2012:challenge

PROM FORUM “BPI Challenge 2017”
http://www.win.tue.nl/promforum/categories/-bpi-challenge-2017

PROM FORUM “brief event descriptions”
http://www.win.tue.nl/promforum/discussion/755/brief-event-descriptions

PROM FORUM “Clarifications about question three”
http://www.win.tue.nl/promforum/discussion/comment/2236/#Comment_2236

FUTURE LEARN COURSES “Introduction to Process Mining With Prom” By Eindhoven University of Technology
https://www.futurelearn.com/courses/process-mining/0/steps/15639

MEMOIRE ONLINE “Comparison of process mining techniques application to flexible and unstructured processes”
http://www.memoireonline.com/01/16/9357/m_Comparison-of-process-mining-techniques-application-to-flexible-and-unstructured-processes2.html

SLIDE PLAYER “Process Mining Control flow process discovery” Elaborated by Fabrizio Maria Maggi
http://slideplayer.com/slide/7092717/

BPM.COM “What is bpm?”
https://bpm.com/what-is-bpm

BPM 2017 “15th International Conference on Business Process Management”
https://bpm2017.cs.upc.edu/

XES - Extensible Event Stream
http://www.xes-standard.org/

EXCEL - Spreadsheets software by Microsoft
https://products.office.com/en-us/excel

MINIT - Process mining software by Minit
https://www.minit.io/

CELONIS - Process mining software by Celonis
https://www.celonis.com/en/
https://www.celonis.com/en/product/

DISCO - Process mining software by Fluxicon
https://fluxicon.com/disco/
https://fluxicon.com/disco/files/Disco-Tour.pdf
https://fluxicon.com/disco/files/Disco-User-Guide.pdf

6 Appendix

6.1 Introduction

This section presents the details of descriptions about the results of the analysis of throughput times per part of the process and also all the correspondent results.

6.2 Descriptions to understand the results

Analyzed Activity and Frequency - The name of the analyzed activity and its total frequency.

Total Average Time Activity - Shows how much time the analyzed activity takes from its beginning to its end - independent of others activities;

Standby Average Time From Another Activity - Shows how much time the analyzed activity waits for until it can begin, after the end of an incoming activity. Specified for each of the incoming activities;

Total Occurrences Regarding the Standby Average Time - After analyzing the standby times from other activities, it was verified the frequency from each analysis from an incoming activity;

Total Time (approximate value) - The total time multiples the two previous results for each incoming activity analysis. Besides that, all the values were set in hours, and for this reason an approximate value is considered;

$\sum$ **Total Time / $\sum$ Total Occurrences (approximate value) -** This category regards the division of the sum of the total time values (previous category) by the sum of the occurrences values (third category), analyzing the total time from all the incoming activities and dividing by the total activities frequency. And finding a general average value that could be compared with the others. Approximate values are considered as well.

6.3 Results of the analysis of throughput times per part of the process

Analyzed Activity and Frequency	Total Average Time Activity	Standby Average Time from Another Activity	Total Occurrences Regarding the Standby Average Time	Total Time (approximate value - hours)	Σ Total Time / Σ Total Occurrences (approximate value - hours)
A_Accepted (31,509)	Instant	24 hrs	23,405	561,720	27.54
		37.8 hrs	8,070	305,046	
A_Cancelled (10,431)	Instant	27.4 d	8,004	5,263,430.4	630.35
		27.6 d	1,038	687,571.2	
		11.5 d	259	71,484	
		6.4 d	116	17,817.6	
		18 d	22	9,504	
		10 d	3	720	
A_Complete (31,362)	Instant	3 millis	31,021	0.025850833	0.00000091
		44 millis	215	0.002627778	
A_Concept (31,509)	Instant	78 secs	16,753	362.9816667	0.0134
		21 millis	10,342	0.0603283333	
A_Denied (3,753)	Instant	3 d	2,165	155,880	64.5504
		47.3 hrs	1,084	51,273.2	
		5.2 d	104	12,979.2	
		28 hrs	88	2,464	
		78.1 mins	25	32.5416667	
		6.4 d	15	2,304	
		9.7 mins	2	0.3233333	
		24.9 hrs	1	24.9	
		99.7 secs	1	0.02769444	
A_Incomplete (23,055)	Instant	3 millis	20,262	0.016885	5.7935
		54.5 hrs	2,380	129,710	
		9.9 hrs	373	3,692.7	
		2.5 hrs	30	75	
		23 hrs	2	46	
A_Pending (17,228)	Instant	8 millis	17,228	0.0382844444	0.0000022
A_Submitted (20,423)	Instant	339 millis	20,423	1.923165833	0.000094
A_Validating (38,816)	Instant	469 millis	23,146	3.0154094444	64.93
		8.7 d	9,120	1,904,256	
		65.2 hrs	3,831	249,781.2	
		7.9 d	1,599	303,170.4	
		36 hrs	449	16,164	
		3.6 d	42	3,628.8	
		4 d	40	3,840	
		6.4 d	16	2,457.6	

		36.3 secs	2	0.02016667	
O_Accepted (17,228)	Instant	22 hrs	7,072	155,584	64.38
		48.7 hrs	5,227	254,554.9	
		5.9 d	4,764	674,582.4	
		5 d	232	27,840	
		4.4 d	16	1,689.6	
		6.9 d	4	662.4	
		20.9 mins	2	0.6966667	
O_Cancelled (20,898)	Instant	4.2 millis	10,270	0.011981667	2.12
		18.4 mins	4,429	1,358.226667	
		2.4 hrs	754	1,809.6	
		51.4 hrs	469	24,106.6	
		18.9 hrs	387	7,314.3	
		24.7 mins	6	2.47	
O_Create Offer (42,995)	Instant	2.5 hrs	31,447	78,617.5	0.96
		7.2 d	4,126	712,972.8	
		6.6 d	671	106,286.4	
		25.5 hrs	57	1,453.5	
		7.3 d	11	1,927.2	
O_Created (42,995)	Instant	1.1 secs	42,995	13.13736111	0.00030
O_Refused (4,695)	Instant	46 millis	3,720	0.0475333333	0.000011
		12 millis	975	0.00325	
O_Returned (23,305)	Instant	58.4 mins	21,530	20,955.86667	3.92
		31 hrs	1,167	36,177	
		5.1 d	250	30,600	
		27.6 hrs	65	1,794	
		28.5 mins	13	6.175	
		8.3 d	4	796.8	
		68.5 secs	1	0.01902778	
O_Sent (mail and online) (39,707)	Instant	26.2 mins	36,199	15,806.896667	0.40
		65 millis	3,111	0.0561708333	
		35.5 mins	337	199.3916667	
		2.4 hrs	41	98.4	
		117.7 secs	12	0.392333333	
O_Sent (online only) (2,026)	Instant	30.4 mins	1,919	972.2933333	0.48
		4 millis	85	0.0000944445	
		48.8 secs	11	0.14911111	
		95.2 mins	6	9.52	
		50.8 secs	2	0.028222222	
		7.3 mins	1	0.121667	
		5.7 mins	1	0.095	
		40.8 secs	1	0.01133333	
W_Call After Offers (31,485)	23.4 mins	5.8 mins	30,926	2,989.5133333	0.36
		15 millis	19,032	0.0793	
		7.7 d	82	15,153.6	

		10.1 hrs	5	50.5	
		4.7 d	1	112.8	
W_Complete Application (29,918)	6.1 hrs	23.8 hrs	18,551	441,513.8	16.78
		22 millis	7,697	0.0470372222	
		9.1 hrs	122	1,110.2	
		95.9 mins	85	135.858333	
		28.4 hrs	70	1,988	
		23.4 hrs	36	842.4	
		4 d	3	288	
W_Assess Potential Fraud (355)	3.1 d	63.8 hrs	184	11,739.2	42.90
		88.6 mins	130	191.96667	
		49.1 hrs	77	3,780.7	
		17.1 hrs	24	410.4	
		8.9 d	17	3,631.2	
		3.1 hrs	14	43.4	
		11.7 hrs	8	93.6	
		11 millis	7	0.0000213889	
		30.2 mins	4	2.0133333	
		5.9 d	1	141.6	
		45.8 mins	1	0.7633333	
W_Handle Leads (3,727)	21 mins	5.7 hrs	3,670	20,919	5.64
		2.6 mins	38	1.646667	
W_Call Incomplete Files (23,218)	21.2 hrs	47.5 hrs	13,831	656,972.5	26.09
		7.2 mins	10,964	1,315.68	
		24.7 hrs	8,811	217,631.7	
		36.5 hrs	142	5,183	
		25.2 hrs	42	1,058.4	
		4.8 hrs	16	76.8	
		18 hrs	9	162	
W_Personal Loan Collection (4)	Instant	4.6 secs	2	0.00255556	0.0038
		31.8 secs	1	0.008833333	
W_Shortened Completion (76)	Instant	98.1 mins	33	53.955	14.34
		36.6 hrs	26	951.6	
		3.3 hrs	8	26.4	
		14.5 hrs	5	72.5	
		74.9 secs	2	0.041611111	
		11.5 secs	2	0.00638889	
		17.3 secs	1	0.004805556	
W_Validate Application (39,444)	23 hrs	8.8 d	18,466	3,900,019.2	134.40
		59.7 hrs	14,226	849,292.2	
		7.6 d	3,242	591,340.8	
		3.2 hrs	2,238	7,161.6	
		38.9 hrs	1,713	66,635.7	
		26.2 hrs	284	7,440.8	
		35.9 hrs	171	6,138.9	

		15.4 mins	47	120.63333	
		5.4 d	36	4,665.6	

7 Annex

7.1 Introduction

This section presents the images of what was seen in Disco to find the results of the analysis of throughput times per part of the process. Figures 28 to 52 shows the activities average time marked with a red rectangle and the paths of the standby average time by another activity numbered in red. Besides being elaborated using Disco, the images were adapted using Paint to show just the relevant informations.

7.2 Images

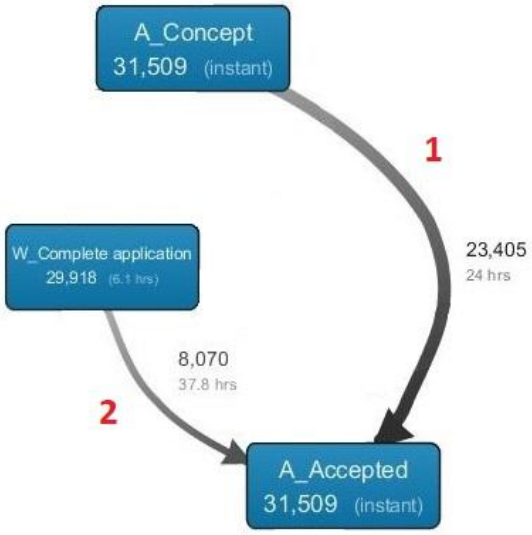


Figure 28 - Visualizing A_Accepted

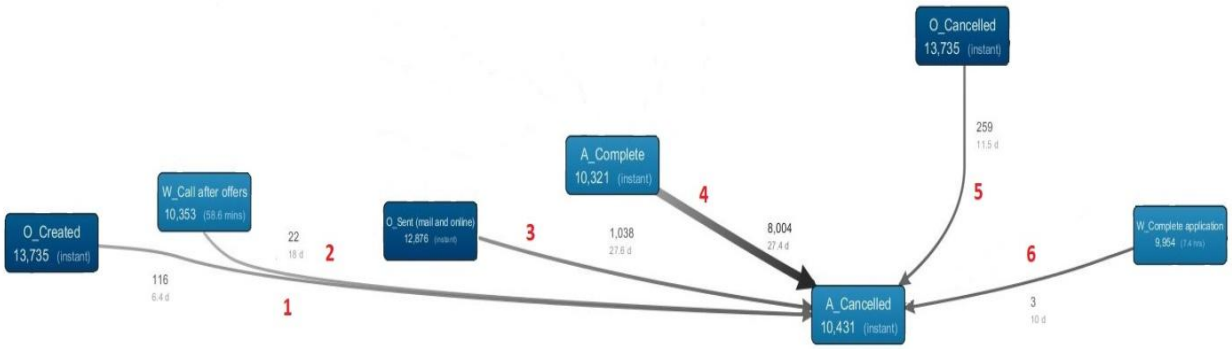


Figure 29 - Visualizing A_Cancelled

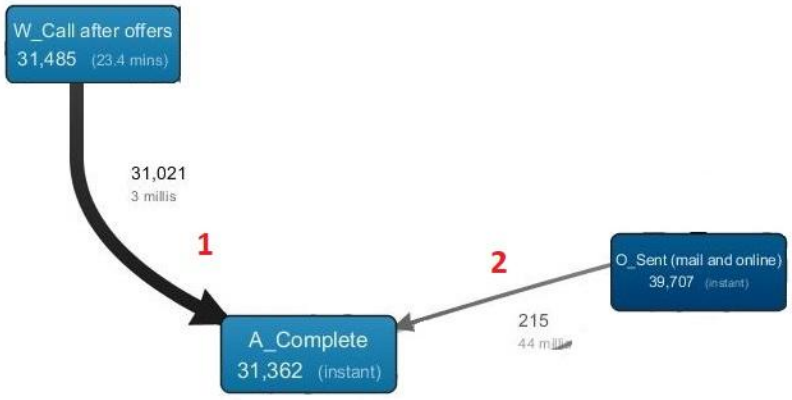


Figure 30 - Visualizing A_Complete

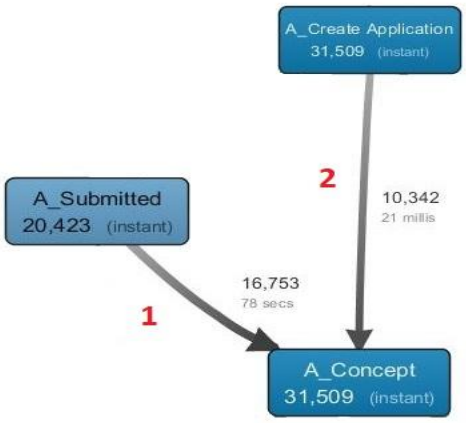


Figure 31 - Visualizing A_Concept

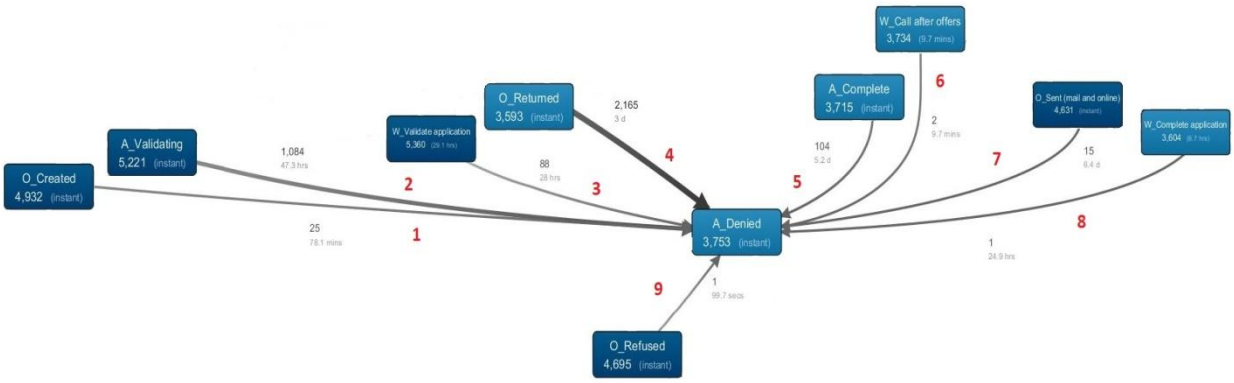


Figure 32 - Visualizing A_Denied

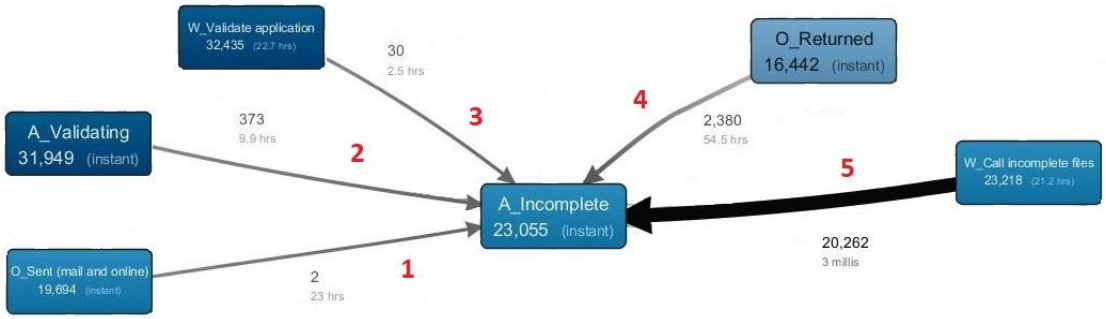


Figure 33 - Visualizing A_Incomplete

Figure 34 - Visualizing A_Pending

Figure 35 - Visualizing A_Submitted

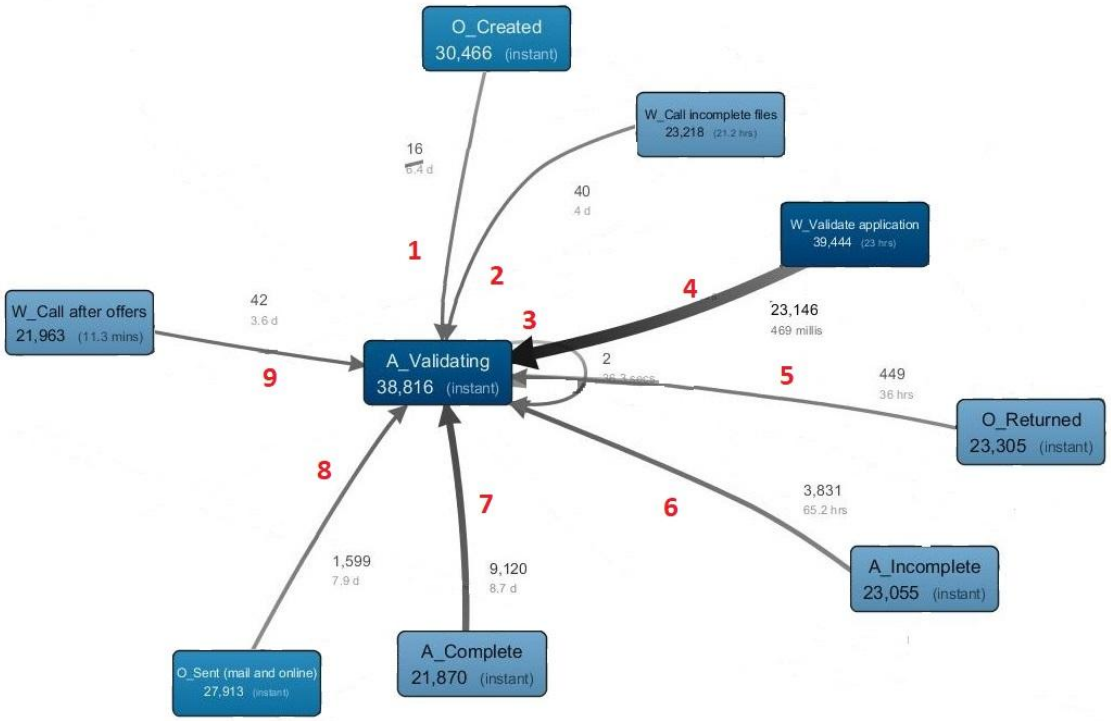


Figure 36 - Visualizing A_Validating

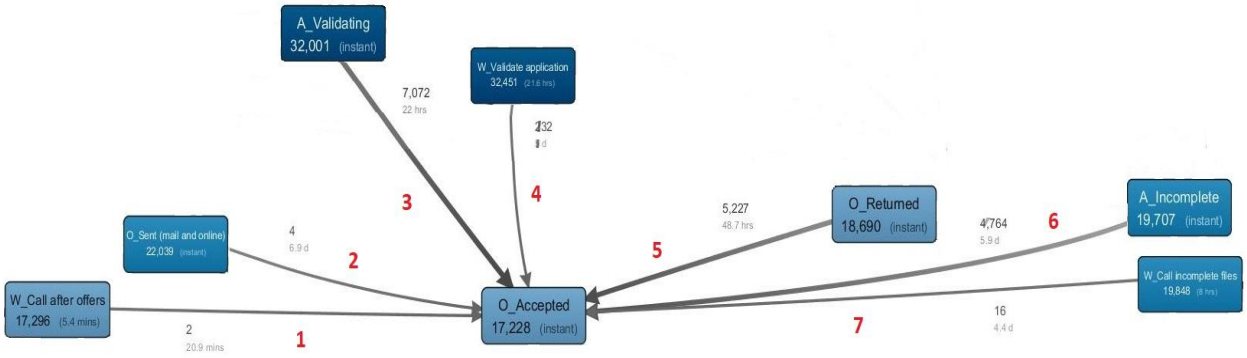


Figure 37 - Visualizing O_Accepted

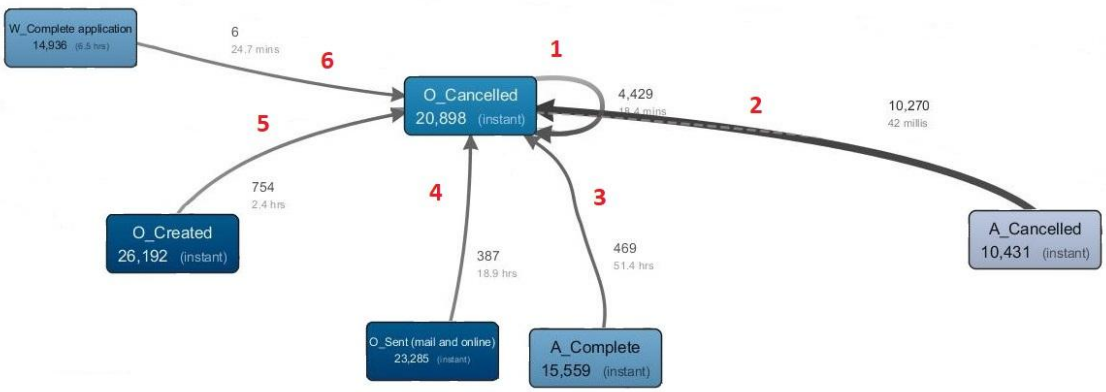


Figure 38 - Visualizing O_Cancelled

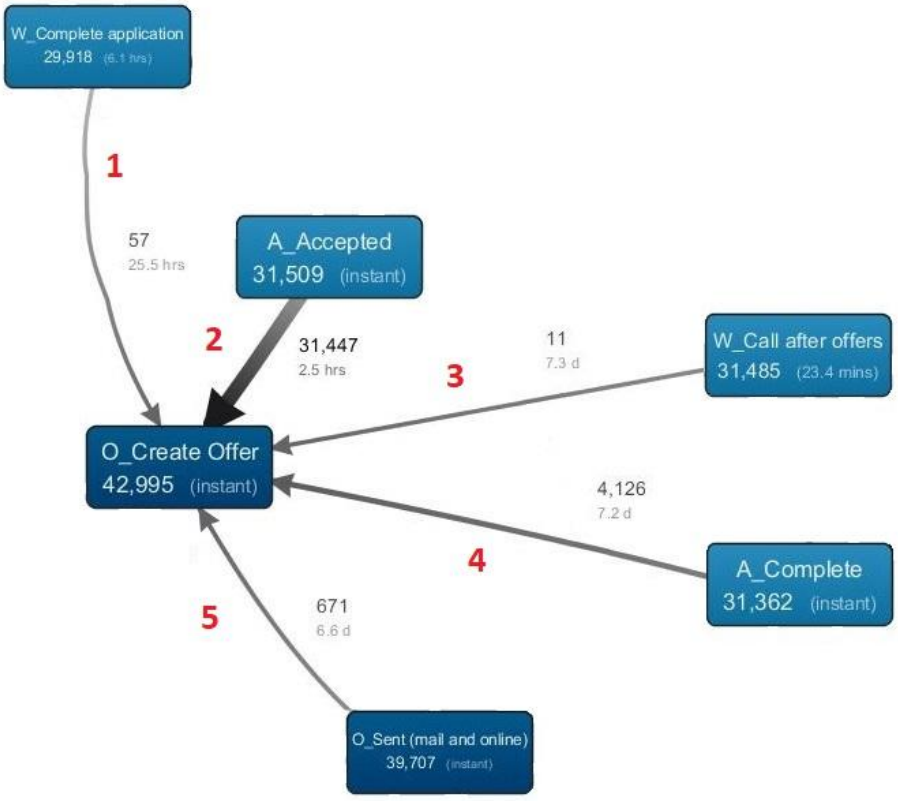


Figure 39 - Visualizing O_Create Offer

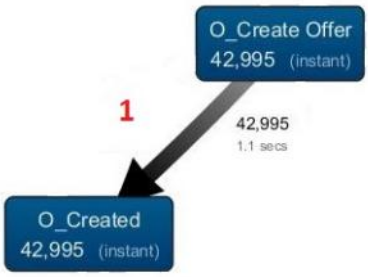


Figure 40 - Visualizing O_Created

Figure 41 - Visualizing O_Refused

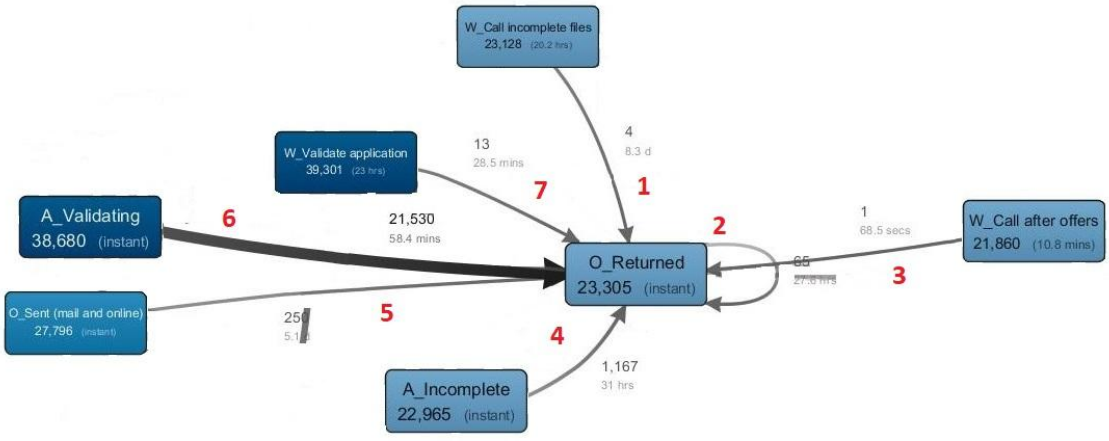


Figure 42 - Visualizing O_Returned

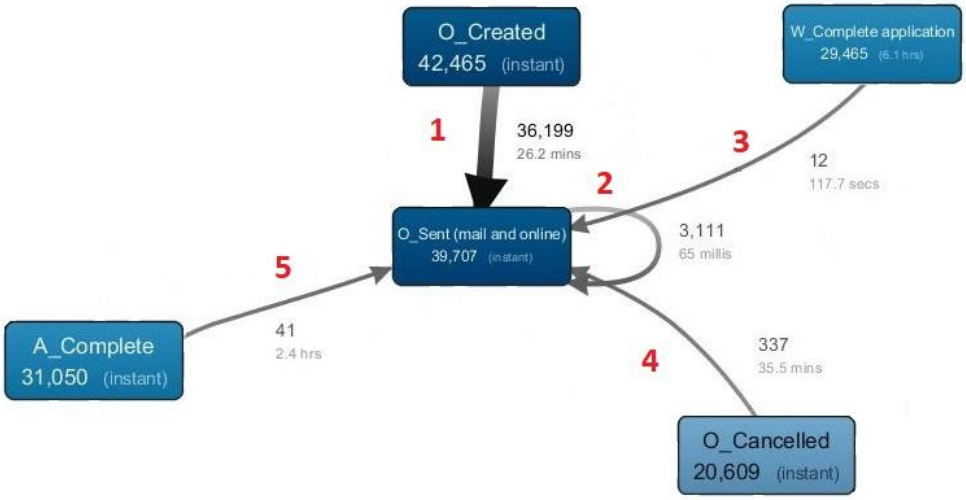


Figure 43 - Visualizing O_Sent (mail and online)

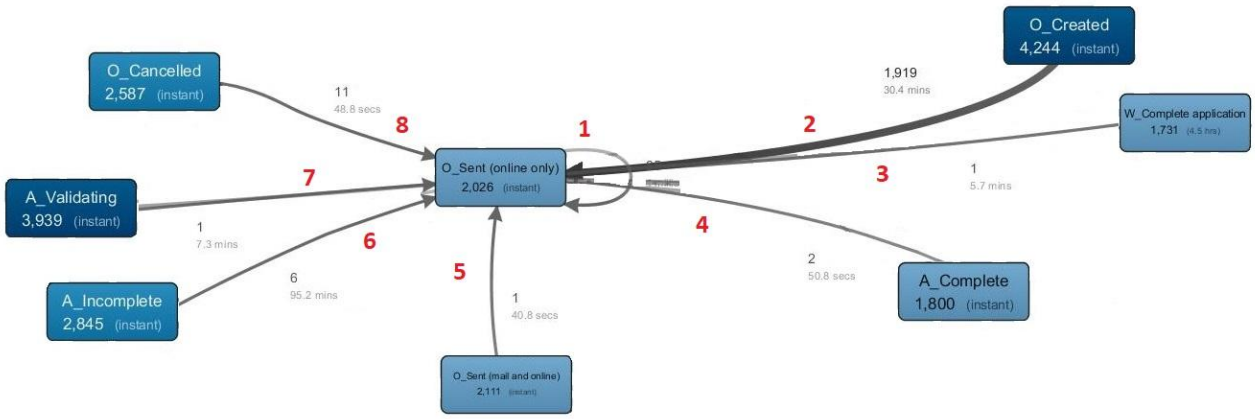


Figure 44 - Visualizing O_Sent (online only)

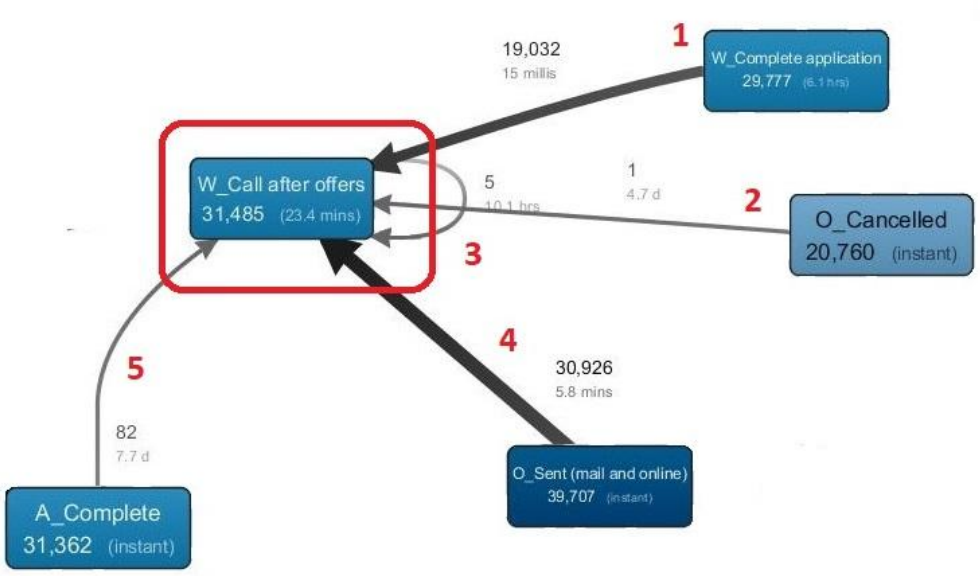


Figure 45 - Visualizing W_Call after offers

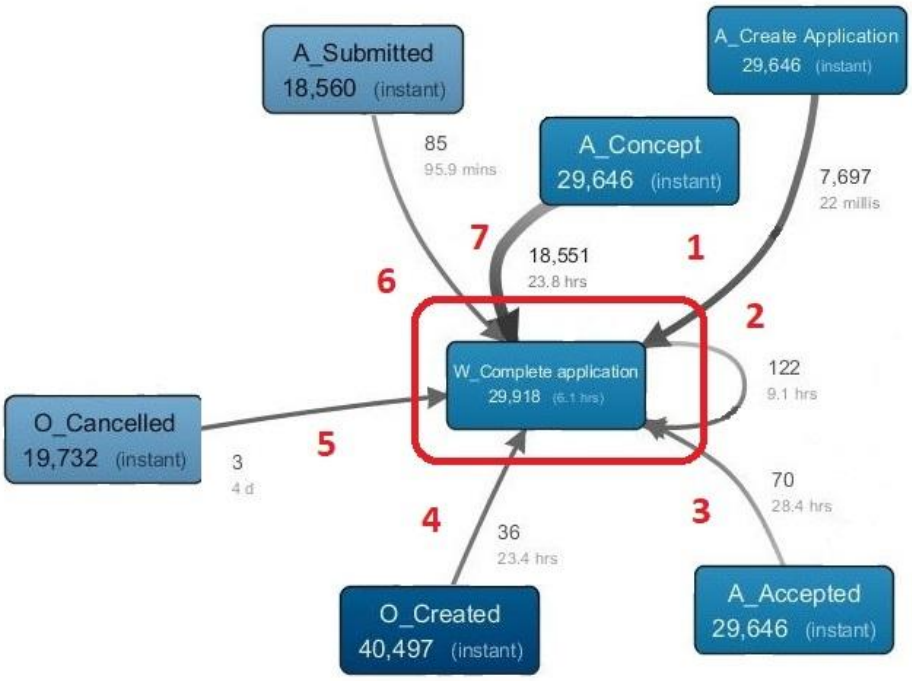


Figure 46 - Visualizing W_Complete application

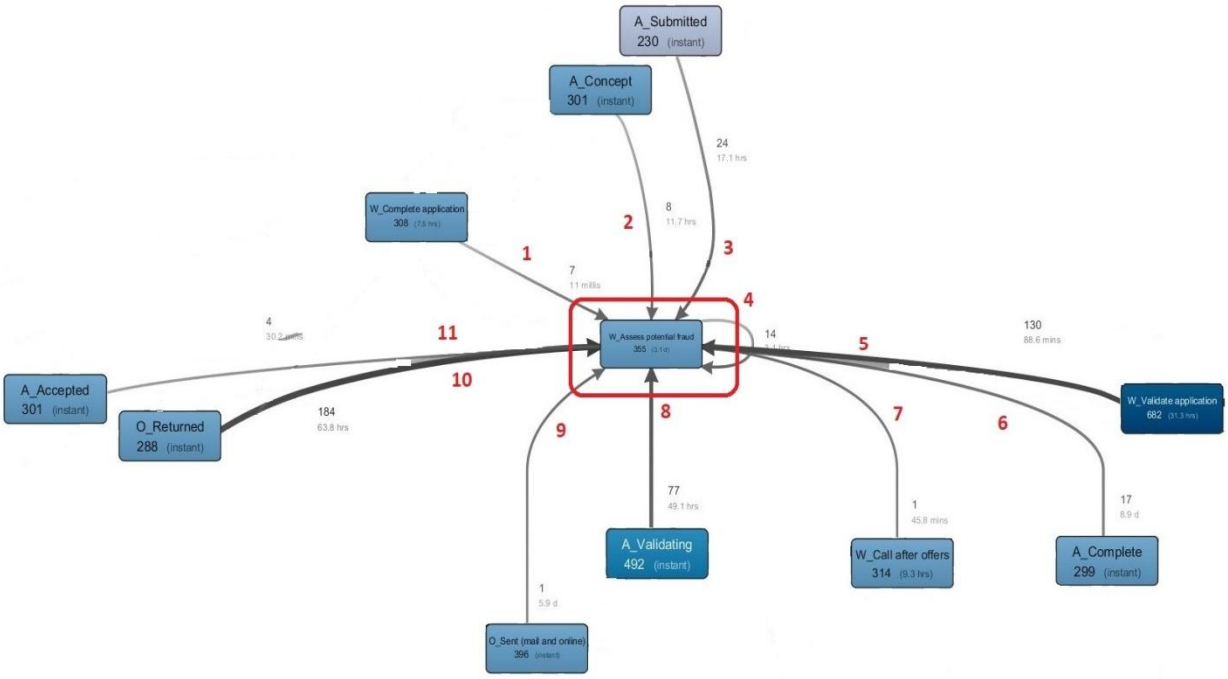


Figure 47 - Visualizing W_Asses potential fraud

Figure 48 - Visualizing W_Handle leads

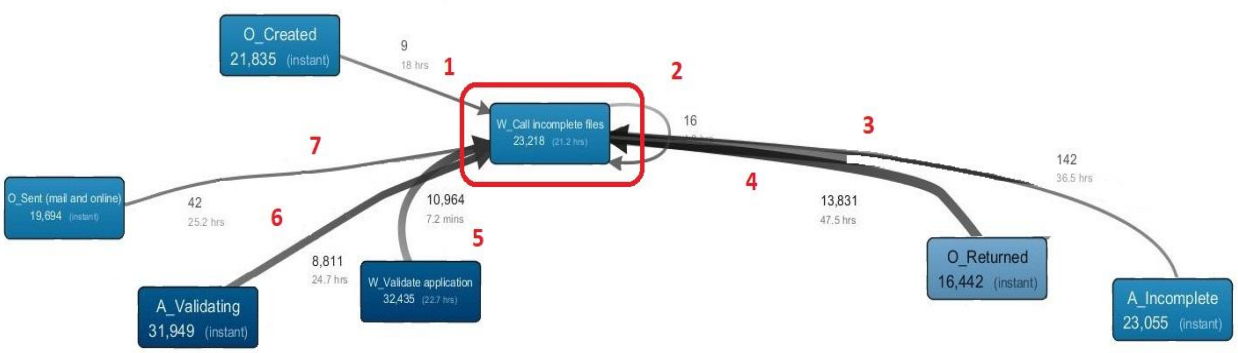


Figure 49 - Visualizing W_Call incomplete files

Figure 50 - Visualizing W_Personal Loan collection

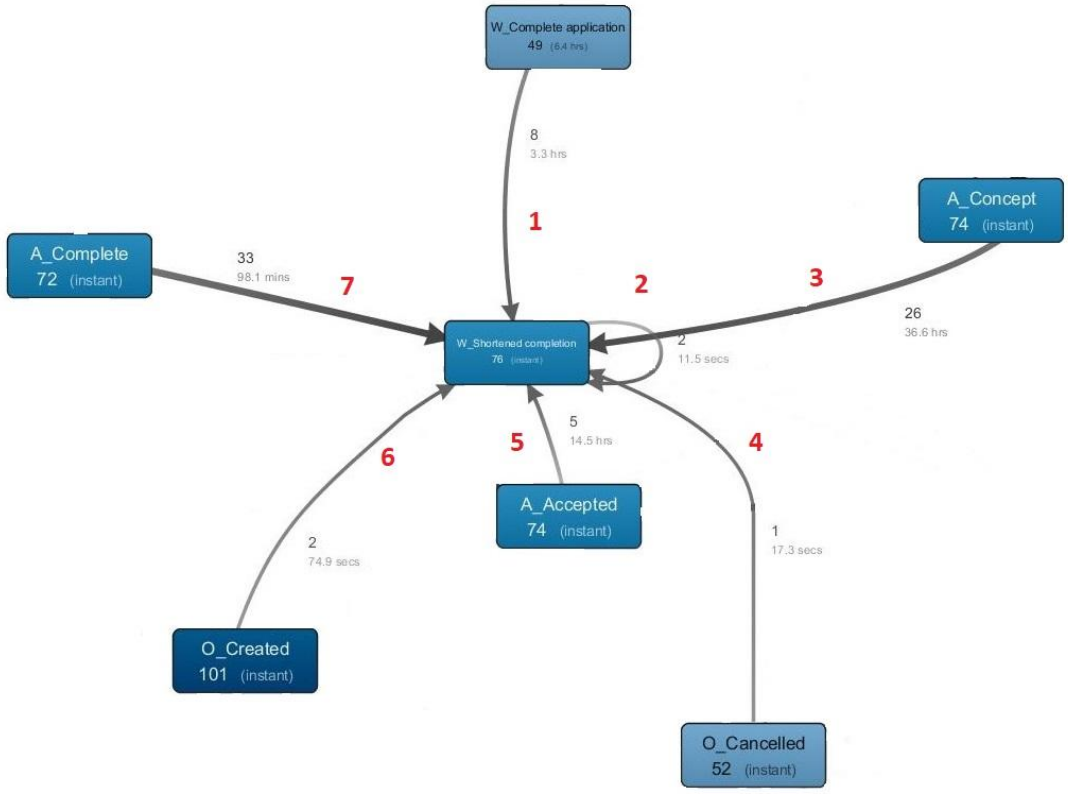


Figure 51 - Visualizing W_Shortened completion

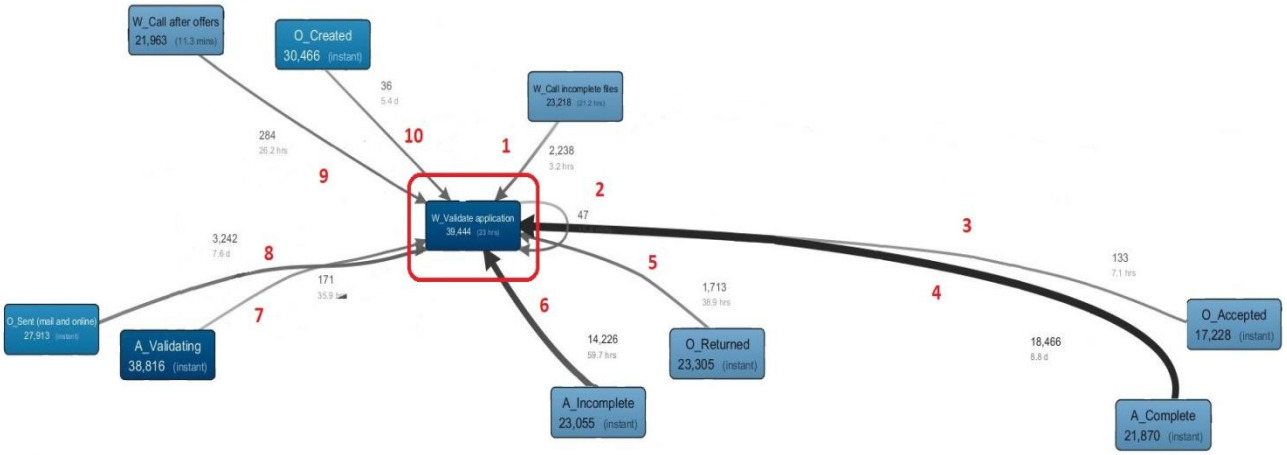


Figure 52 - Visualizing W_Validate application

Case ID	Activity	Resource	Start Timestamp	Complete Timestamp	Variant	Variant index	(case) Application Type	(case) LoanGoal	(case) Requested Amount	Accepted	Action	CreditScore	EventID	EventOrigin	FirstWithdrawalAmount	Monthly Cost	NumberOfTerms	OfferID	Offered Amount	Selected	lifecycle:transition
Application_652823628	A_Create Application	User_1	2016/01/01 07:51:15.304	2016/01/01 07:51:15.304	Variant 2	2	New credit	Existing loan takeover	20000.0		Created		Application_652823628	Application							complete
Application_652823628	A_Submitted	User_1	2016/01/01 07:51:15.352	2016/01/01 07:51:15.352	Variant 2	2	New credit	Existing loan takeover	20000.0		statechange		ApplState_1582051990	Application							complete
Application_652823628	A_Concept	User_1	2016/01/01 07:52:36.413	2016/01/01 07:52:36.413	Variant 2	2	New credit	Existing loan takeover	20000.0		statechange		ApplState_642383566	Application							complete
Application_652823628	W_Complete application	User_17	2016/01/02 08:45:22.429	2016/01/02 08:45:22.429	Variant 2	2	New credit	Existing loan takeover	20000.0		Obtained		Workitem_1875340971	Workflow							start
Application_652823628	A_Accepted	User_52	2016/01/02 09:23:04.299	2016/01/02 09:23:04.299	Variant 2	2	New credit	Existing loan takeover	20000.0		statechange		ApplState_99568828	Application							complete
Application_652823628	O_Create Offer	User_52	2016/01/02 09:29:03.994	2016/01/02 09:29:03.994	Variant 2	2	New credit	Existing loan takeover	20000.0	TRUE	Created	979	Offer_148581083	Offer	20000.0	498.29	44		20000.0	TRUE	complete
Application_652823628	O_Created	User_52	2016/01/02 09:29:05.354	2016/01/02 09:29:05.354	Variant 2	2	New credit	Existing loan takeover	20000.0		statechange		OfferState_1514834199	Offer				Offer_148581083			complete
Application_652823628	O_Sent (mail and online)	User_52	2016/01/02 09:30:28.606	2016/01/02 09:30:28.606	Variant 2	2	New credit	Existing loan takeover	20000.0		statechange		OfferState_2051164740	Offer				Offer_148581083			complete
Application_652823628	W_Call after offers	User_52	2016/01/02 09:30:28.631	2016/01/02 09:30:28.631	Variant 2	2	New credit	Existing loan takeover	20000.0		Obtained		Workitem_719909876	Workflow							start
Application_652823628	A_Complete	User_52	2016/01/02 09:30:28.633	2016/01/02 09:30:28.633	Variant 2	2	New credit	Existing loan takeover	20000.0		statechange		ApplState_946455804	Application							complete
Application_652823628	W_Validate application	User_117	2016/01/13 11:10:55.145	2016/01/13 11:10:55.145	Variant 2	2	New credit	Existing loan takeover	20000.0		Obtained		Workitem_1641716416	Workflow							start
Application_652823628	A_Validating	User_117	2016/01/13 11:10:55.973	2016/01/13 11:10:55.973	Variant 2	2	New credit	Existing loan takeover	20000.0		statechange		ApplState_752879093	Application							complete
Application_652823628	O_Returned	User_117	2016/01/13 11:11:03.569	2016/01/13 11:11:03.569	Variant 2	2	New credit	Existing loan takeover	20000.0		statechange		OfferState_1310330551	Offer				Offer_148581083			complete
Application_652823628	W_Call incomplete files	User_115	2016/01/14 07:16:20.972	2016/01/14 07:16:20.972	Variant 2	2	New credit	Existing loan takeover	20000.0		Obtained		Workitem_2031858979	Workflow							start

Figure 53 – event log fragment example (with all the attributes), based on the logs provided by the BPI Challenge 2017