Seventh International Business Process Intelligence Challenge (BPIC17)

Chang Hyup Oh , Min Gyu Choi and Mohammad Zahedy

Depart of Industrial and Management Engineering, Pohang University of Science and Technology, Korea,
{changhyup.oh, cmg306, mzahedy}@postsech.ac.kr

Abstract. Process mining is useful to look for improvement point in current system from process execution logs. Nowadays, many companies collect all kinds of their execution logs and try to find what they should enhance. In this challenge, we are provided data from a financial institute and the company's 3 particular interests. The first is about throughput times per part of the process. More precisely, the company wants to know the difference between the time spent by its systems and applicants. The company has been also interested in the influence on the frequency of incompleteness to the final outcome. Finally, the company wants to compare the conversion for a single offer cases to the conversion for multiple offers cases. In order to answer to the company's questions and give some significant insight from the company's execution logs, we analyze the company's execution logs with several process mining techniques such as bottle neck analysis, discovering simulation models. For applying process mining techniques, we use some computer software such as ProM, Disco, C# and so on. Furthermore, we do additional analysis which the company did not refer directly but we think it is meaningful to understand its execution logs and processes.

Keywords: BPI Challenge, financial institute, throughput time, incompleteness, multiple offers cases

1 Introduction

Process mining is well known tool for finding improvement points in real world systems. In this challenge, we are provided a real event log and 3 questions from a financial institute. Nowadays, almost all companies have a management information system (MIS), so they collect all kinds of executions. Unfortunately, many of them cannot use their event logs to improve their system although many problems or improvement points are described in the event logs. In this report, we firstly understand the company’s processes by interpreting event logs. After that, we answer to 3 questions and give additional insights by analyzing both event logs and processes. For analysis, we use several tools in Table 1. In order to answer questions, we need to do preprocessing for each question, since given event logs have too many

attributes to analyze easily. After preprocessing, we can get simplified event logs, then we analyze the event logs by using Disco and ProM in accordance with requirements of each question.

Table 1. Used tools in this report

Tool	Use
ProM 5.2 / 6.6	for overall analysis such as social network analysis and so on.
fluxicon Disco	drawing process map, filtering event logs, obtaining basic statistics.
Oracle SQL developer	preprocessing event logs, making groups.
Microsoft C#	preprocessing event logs, making groups.

1.1 Data description

The event logs provided contain all applications filed in 2016. In event logs, there are 31,509 loan applications and 42,995 offers with 1,202,267 events. All events are classified as three types, namely **Application state changes**, **Offer state changes** and **Workflow** events. There are two files of data, one is **Application event log file** which contains all events and the other is **Offer event log** file which contains only events related to offers. We use Application event log file to analyze, since we think events related to offers are strongly dependent on other attributes in data.

Table 2. Names and desciptions of activities for each type

Type	Event description
"A_" Application state changes	Refers to state changes of the applications. All customers should submit their applications to the company. – A_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. – A_Cancelled: If the customer never sends in his/her documents or calls to tell he/she doesn't need the loan, the application is cancelled. – A_Complete: The offers have been sent to the customer and the company waits for the customer to return a signed offer along with the rest of the documents (payslip, ID etc) – A_Concept: The application is in the concept state, it means that the customer just submitted it (or the company started it), and a first assessment has been done automatically. – A_Create Application: A customer creates his/her application on the website.

	– A_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 this status. – A_Incomplete: If documents are not correct or some documents are still missing, the status is set to incomplete. – A_Pending: If all documents are received and the assessment is positive, the loan is final and the customer is payed. – A_Submitted: A customer has submitted a new application from the website. A new application can also be started by the company, in that case this state is skipped. – A_Validating: The offer and documents are received and are checked.
"O_" Offer state changes	Refers to state changes of the offers. If the applications are accepted, the company creates one or more offers and sends to the customers. – O_Accepted: Similar meaning to accepted status in Application. – O_Cancelled: Similar meaning to cancelled status in application. – O_Create Offer: The company creates one or more offers for the customers. – O_Created: After O_Create Offer status, the status is automatically changed to O_Created. – O_Refused: The offers are refused after O_Returned – O_Returned: After validating, the offers are returned since it cannot fit the acceptance criteria. After this status, the customers can determine that submit additional documents to the company or refuse the offers. – O_Sent (mail and online) and O_Sent (online only): If the offers are created, the company sends the offers to the customers. Then, the status is O_Sent.
"W_" Workflow	Refers to work items. Some of work items are corresponding work items for call agents to process. These are the so called 'standard work items'. – W_Handle leads, W_Complete application, W_Validate application, W_Call incomplete files, W_Assess potential fraud. The others are the so called 'customer work items'. – W_Personal loan collection, W_Shortened completion. If the customers have certain profiles that define as a lower credit risk, then validating processes are skipped.

Table 2 shows names and descriptions of all activities in data. In the data, there are 149 originators (in data, namely user) who are employees or systems of the company. Moreover, there are additional information such as requested load amount, loan goal, offered amount and so on. If the additional explanations for questions or results, we refer them in following sections for each question. Fig. 1 shows the general process for taking loan. The customers send application to the company, then the company sends proper offers to the custoemers. After that, the customers accept one of offers then the company starts validating the customers.

Fig. 1. General process for taking loan

2 Question 1

2.1 Question description

Given question: 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.

This question requires about throughput times per part of process. Hence we look at each activity or part of process execution in two type time variable. The first one is the time that the activity execution takes which is called service time, and the second one is the time which activity is in standby mood waiting for user to execute that activity, called waiting time. Our focus is on waiting, as it is matter of interest for the company. The waiting time is caused for two reasons. It is either the time activity is waiting for a user inside the company to take action and execute that activity or it is waiting for applicant to provide required input for that activity to be executed. We look into waiting time and try to find out the cause and conditions it happens. For this, we go through the following steps to answer to the question 1.

2.2 Import data into Disco

We import Application event log dataset which contains all activities to Disco to generate process map. As the data is in xes format, Disco does not need to set case id, activity and other fields. In order to generate process maps, we set a bar for activities to 100% and a bar for paths to 0% to have all activities for further analysis regarding waiting time.

2.3 Find activities with their causes of waiting time

After importing data into Disco and generating a process map, we try to figure out how to find the cause of waiting time for activities. There are totally 26 activities. In order to find out whether activity is waiting for user action or applicant input, we use the process map from Disco with setting in subsection 2.2. Then looking in the process map and based on the activity explanation in ProM forum web site, we can determine causes of delay in each activity. While choosing the activities, we consider the following points:

- The activities which are the first activity for each application (e.g., A_Create Application, A_Concept, A_submitted) are not considered as waiting time by both user and applicant.
- The activities whose waiting time is almost 0 are removed.

Table 3. Major activities and their cause of waiting time

Activity	Cause of waiting time
A_Complete → W_Validate application	Applicant
A_Complete → A_Cancelled	Applicant
A_Complete → A_Validating	User
A_Incomplete → W_Validate application	Applicant
O_Returned → W_Incomplete files	User
A_Concept → W_Complete application	User
A_Concept →A_Accepted	User
W_Complete application → A_Accepted	Applicant
A_Validating → O_Accepted	User

2.4 Data segmentation

In order to betted find out when waiting time happens and when some activities have longer waiting time or shorter waiting time, we divide the data into three segment. The data is segmented in Disco by applying filter based on mean waiting time.

Table 4. Data segmentation based on mean waiting time

Segment	Mean waiting time (hours)	Cases (%)
Segment 1	0 – 8	3

Segment 2	8 – 72	75
Segment 3	72 –	22

Table 4 shows information for each segment. Segment 1 contains about 3% of all cases, and its mean waiting time is 0 to 8 hours. It means cases in segment 1 have very short waiting time and are executed quickly. These are the most efficient cases based on duration of waiting time. In segment 2, most of the cases are contained (75%) whose waiting time is between 8 and 72 hours. The last segment contains the longest waiting time among all cases. These cases need more than 72 hours of waiting time. These cases consist 22% of whole cases.

Table 5. Waiting time in each segment

Segment 1	**Activity**	**Cause**	**Mean waiting time**
	A_Complete → W_Validate application	Applicant	24.1 h
	A_Complete → A_Cancelled	Applicant	14.2 h
	A_Incomplete → W_Validate application	Applicant	15.7 h
	O_Returned → W_Call incomplete files	User	19.7 h
Segment 2			
	A_Complete → W_Validate application	Applicant	7.1 d
	A_ Complete → A_Cancelled	Applicant	30.6 d
	A_Incomplete → W_Validate application	Applicant	23.2 h
	O_Returned → W_Call incomplete files	User	25.7 h
	A_Concept → W_Complete application	User	20.2 h
	A_Concept → A_Accepted	User	11.4 h
	W_Complete application→ A_Accepted	Applicant	5.2 h
	A_Validating → O_Accepted	User	6.0 h
Segment 3			
	A_Complete → W_Validate application	Applicant	18.9 d
	A_Complete → A_Cancelled	Applicant	30.8 d
	A_Complete → A_Validating	User	17.9 d
	O_Returned → W_Call incomplete files	User	1.9 d
	A_Concept →W_Complete application	User	22.4 h
	A_Concept → A_Accepted	User	15.7 h

2.5 Analysis

As shown in Table 5, segment 2 and 3 have longer waiting time. We can notice two bottlenecks. One is A_Complete bottleneck which is caused by applicant (A_Complete → W_Validate application, A_Complete → A_Cancelled) and the other is O_Returned bottleneck which is caused by user (O_Returned → W_Call incomplete files).

In all segments, the number of activities delayed by user and applicant are different. For each segment, there are 4 (User 1, Applicant 3), 8 (User 4, Applicant 4) and 6 (User 4, Applicant 2) major delayed activities respectively.

There are some waiting time which happens in all segments. For example, the process (A_Complete → W_Validate application) appears in all segments, but the waiting times of that process are different for each segment. It indicates that overall cases' waiting time are affected by these activities' waiting time. If they spend shorter time, the average waiting time of those cases goes down. Fig. 1 and 2 show that A_Complete bottleneck for segment 2 and 3 which is included in all segments. The process (A_Complete → W_Validate application) spends more time in segment 3. It can occur longer mean waiting time for segment 3. In all segments, the process (A_Complete → W_Validate application) takes waiting time to complete submitting some documents from applicants to the company. This is one of the bottleneck, so it makes overall waiting time go higher. There are another bottleneck caused by user. It is the process (O_Returned to W_Call incomplete files).

Fig. 2. A_Complete bottleneck for segment 2

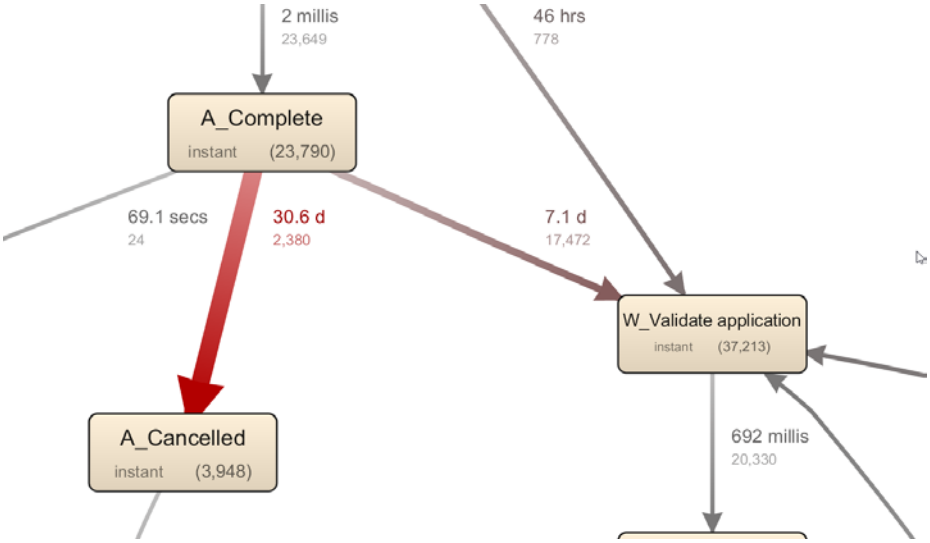
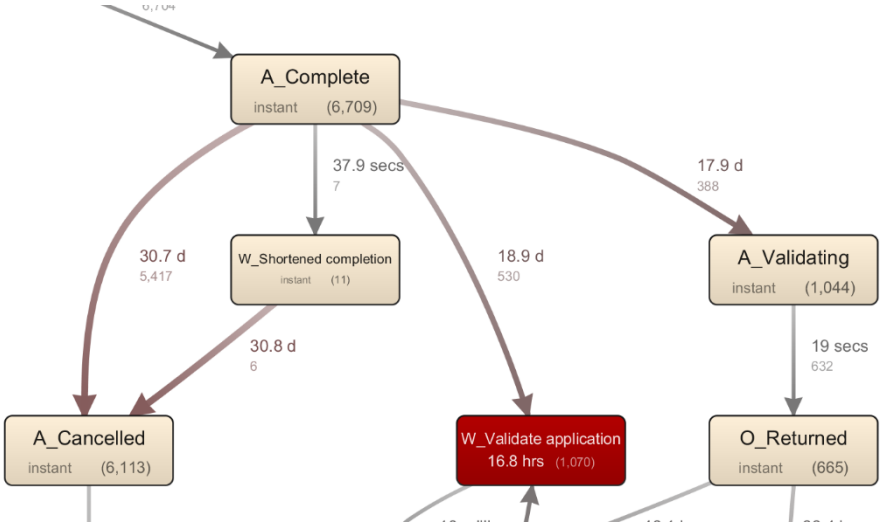


Fig. 3. A_Complete bottleneck for segment 3



In addition to waiting time, we also consider a service time in each segment. We find out that in segment 2 which has the most cases, so it spends more service time and waiting time than the others. For all cases, both waiting time and service time occur in some specific processes which are around completion/validation of application, returning offers for incompleteness or assessing potential fraud.

3 Question 2

3.1 Question description

Given question: 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 not to accept the final offer.

The main purpose of this question is to observe the final outcome when the number of incompleteness is different. Thus, we firstly need to define meanings of incompleteness and the final outcome.

In the process, there is the activity whose name is A_Incomplete. It appears when documents are not correct or some documents are still missing so the customers need to submit more documents to the company. After A_Incomplete, if the customers submit additional documents, then a process for validating will begin. Otherwise, the customers do not submit additional documents, the offers should be cancelled. In the analysis for this question, we define frequency of incompleteness as the number of A_Incomplete activities in each cases.

In case of defining final outcomes, the 14 activities are observed as end point, but considering the frequency of each activity and information form ProM forum web site, there are 5 main end points.

Table 6. Main end points in data

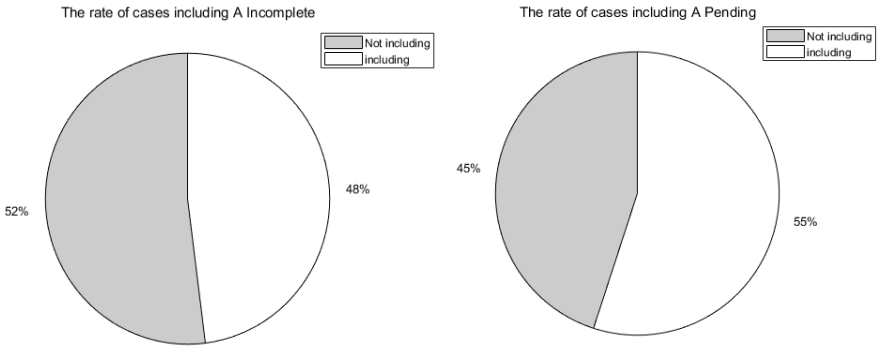
End-point	Frequency
O_Cancelled	14,707
A_Pending	12,791
O_Refuesd	3,719
A_Cancelled	161
A_Denied	33

Table 6 shows the 5 main end points and the frequency of each activity. However, as the main purpose of making offer is to take a loan from the company, the result of each case can be classified whether the customers can take a loan or not. A_Pending is the key activity to judge whether the customers are paid or not. If all documents are submitted and all assessments are positive, the company determines to pay. Then, A_Pending appears in the cases. Therefore, if the cases include A_Pending, we can regard the customers in those cases taking a loan. In order to answer to this question, we find out the influence on the frequency of A_Incomplete to the appearance of A_Pending.

3.2 Basic statistics and information about A_Incomplete and A_Pending

The rate of cases including A_Incomplete at least once and A_Pending is in Fig.3.

Fig. 4. The rate of cases including A_Incomplete and A_Pending



48% of cases include A_Incomplete at least once. It means about half of cases' submitted documents are incorrect or missing. And 55% of cases include A_Pending. It means more than half of cases are accepted and paid by the company. From these

information, we start to analyze the influence of the frequency of A_Incomplete to the appearance of A_Pending.

3.3 Analysis

Table 7. The rate of cases including A_Pending for each segment

Frequency of A_Incomplete	Number of cases	Number of cases including A_Pending	Rate of cases including A_Pending
0	16,506	4,581	28%
1	9,317	7,665	82%
2	3,970	3,471	87%
More than 3	1,716	1,511	88%
At least 1	15,003	12,647	84%

In Table 7, we compare the rate of cases including A_Pending among segments that each segment has different frequency of A_Incomplete. We can easily check the rate of A_Pending increases when the frequency of A_Incomplete increases. In order to the question 2, we compare the segment whose frequency of A_Incomplete is 0 to the segment whose frequency of A_Incomplete is at least 1. The rate difference between those segments is 56%. This dramatic difference means when the frequency of incompleteness goes higher, the probability of taking a loan also goes higher. However, when considering results from question 1, many cases include A_Cancelled after A_Complete which means the customers do not submit any documents for validating. In those cases, there are no chance to appear A_Incomplete. Furthermore, the customers in those cases can be regarded that they do not have intentions to take loans. Therefore, we define those customers as under-motivated customers who do not need to take loans from the company and except the cases related to under-motivated customers.

Table 8. The rate of casese including A_Pending for each segment except under-motivated customers

Frequency of A_Incomplete	Number of cases	Number of cases including A_Pending	Rate of cases including A_Pending
0	8,502	4,581	54%
1	9,317	7,665	82%
2	3,970	3,471	87%
More than 3	1,716	1,511	88%

At least 1	15,003	12,647	84%

Table 8 shows the result with above treatment. About 8,000 cases are related to under-motivated customers and are removed. As a result, the rate of A_Pending in the segment whose frequency of A_Incomplete is 0 is changed from 28% to 54%. However, after removing the under-motivated customers, there are still huge difference of rate of A_Pending between the segments whose frequencies of A_Incomplete are 0 and at least 1. Therefore, we conclude that the frequency of incompleteness and the probability of taking a loan have weak positive correlation which means that A_Incomplete can increase the rate of taking a loan. Our answer to question 2 is that the hypothesis is not always true.

3.4 Additional analysis – Duration time

Additionally, we check the relationship between duration time and frequency of incompleteness. According to Table 9, we can check the relationship between mean case duration time and frequency of A_Incomplete. The mean case duration time also increases when frequency of A_Incomplete goes higher.

Table 9. Relationship between duration time of cases and frequency of A_Incomplete

Frequency of A_Incomplete	Number of cases	Mean Case Duration Time (Days)
0	8,502	17.1
1	9,317	18
2	3,970	23.2
More than 3	1,716	29.4
At least 1	15,003	20.7
Whole cases	31,509	21.9

In this analysis, a presence of A_Incomplete is not significant to the mean case duration time. The segment who includes no A_Incomplete has 17.1 days as its mean case duration time and the segment who includes only 1 A_Incomplete has 18 days as its mean case duration time. The difference between both segments is less than 1 day. Moreover, the mean case duration time for cases who have at least 1 A_Incomplete is 20.7 days and is not significantly differ from the mean case duration time for cases with no A_Incomplete. Therefore, we conclude the presence of A_Incomplete is not significant to mean duration time. However, when the number of A_Incomplete goes higher, the mean duration time increases significantly (e.g., the time difference between cases who include A_Incomplete 1 and 2 is 5.3 days, 2 and more than 3 is 6.2 days).

3.5 Additional analysis – Considering loan goal

In this subsection, we compare the influence on the frequency of incompleteness to final outcome for top 3 categories of loan goal. Top 3 categories of loan goal are Car, Home improvement and existing loan takeover. All categories show that the similar rates of cases including A_Pending and increasing trends of duration time. However, the cases whose loan goal is Car present interesting result that the mean case duration times for the cases without A_Incomplete and the cases with only 1 A_Incomplete are same, but the pending rates are 53%, 82% respectively. If we regard a time is cost for the company, two segments of cases have same cost but more revenue can occur by the cases with only 1 A_Incomplete. The precise information is presented in following tables.

Table 10. Relationship between frequency of A_Incomplete and mean duration time for cases whose loan goal is Car

Frequency of A_Incomplete	Total cases	Cases including A_Pending	Mean Case Duration Time (Days)
0	2,658	1,416 (53%)	16.4
1	2,691	2,212 (82%)	16.4
2	1,007	874 (87%)	21
More than 3	339	285 (88%)	25.9
At least 1	4,037	3,371 (84%)	18.4

Table 11. Relationship between frequency of A_Incomplete and mean duration time for cases whose loan goal is Home improvement

Frequency of A_Incomplete	Total cases	Cases including A_Pending	Mean Case Duration Time (Days)
0	2,077	1,246 (60%)	17.7
1	2,384	1,994 (84%)	19.1
2	1,020	893 (88%)	23.8
More than 3	409	358 (88%)	28.3
At least 1	3,813	3,245 (85%)	21.3

Table 12. Relationship between frequency of A_Incomplete and mean duration time for cases whose loan goal is Existing loan takeover

Frequency of A_Incomplete	Total cases	Cases including A_Pending	Mean Case Duration Time (Days)
0	1,421	649 (45%)	18.9
1	1,647	1,343 (82%)	19.7
2	831	729 (88%)	25
More than 3	390	353 (91%)	30.4
At least 1	2,868	2,425 (85%)	22.7

3.6 Additional analysis – Requested amount

Similar tasks are executed by this subsection. We classify all cases to 3 segments. Each segment has requested amount range € 0 – 20,000, € 20,000 – 40,000 and € more than 40,000 respectively. The results of each segment are presented in following tables.

Table 13. Relationship between frequency of A_Incomplete and mean duration time for cases whose requested amount is less than 20,000

Frequency of A_Incomplete	Total cases	Cases including A_Pending	Mean Case Duration Time (Days)
0	6,427	3,399 (53%)	16.8
1	7,002	5,764 (82%)	17.6
2	2,795	2,456 (88%)	22.2
More than 3	1,130	986 (87%)	28.3
At least 1	10,927	9,206 (84%)	19.9

Table 14. Relationship between frequency of A_Incomplete and mean duration time for cases whose requested amount is between 20,000 and 40,000

Frequency of A_Incomplete	Total cases	Cases including A_Pending	Mean Case Duration Time (Days)
0	1,622	942 (58%)	17.9
1	1,732	1,461 (84%)	19.2
2	891	786 (88%)	24.6
More than 3	420	380 (90%)	30.8
At least 1	3,043	2,627 (86%)	22.4

Table 15. Relationship between frequency of A_Incomplete and mean duration time for cases whose requested amount is more than 40,000

Frequency of A_Incomplete	Total cases	Cases including A_Pending	Mean Case Duration Time (Days)
0	492	240 (49%)	19.2
1	583	441 (76%)	20
2	284	229 (81%)	28.2
More than 3	166	144 (87%)	33.5
At least 1	1,033	814 (79%)	24.4

3.7 Conclusion

As the frequency of incompleteness and final outcome, the rate of customers taking loans show positive correlations. The hypothesis that referred by the company, if applicants are confronted with more requests for completion, they are more likely not

to accept the final offer, is not always true. Furthermore, the difference between the mean case duration times of frequency of A_Incomplete is 0 and 1 is small, but the difference between the pending rates of both segments is significant (i.e., at least 24%). Therefore, it is effective that the company positively consider making incompleteness once.

4 Question 3

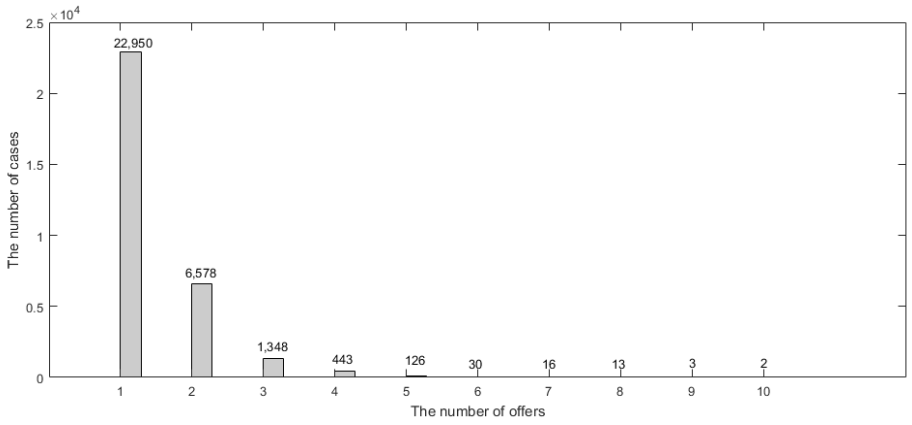
4.1 Question description

Given question: 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?

The company wants to compare conversions between applicants for whom a single offer is made and applicants for whom multiple offers are made through this question. In the Application event log file, there are 31,509 cases. Among them, 22,950 cases receive a single offer and 8,559 cases receive multiple offers from the company. In order to answer to this question, we firstly divide the data into at least 2 segments, one is cases with single offer and the other is cases with multiple offers. However, for more precise analysis, we make 3 segments. In following subsections, we answer to this questions and give other insights through additional analysis.

4.2 Basic statistics and information

Fig. 5. The number of cases for 10 different kinds of applications which receive 1 – 10 offers



We make 3 segments based on Fig. 4. The segment 1 is set of cases which receive only 1 offer, the segment 2 is set of cases which receive two offers and the segment 3 is set of cases with more than 2 offers from the company.

Table 16. Basic statistics about the number of cases and pending rate for each segment

Segment	Total cases (31,509)	Cases with A_Pending (17,228)	Pending rate (%)
Single offer (Segment 1)	22,950 (72.84%)	12,178 (70.69%)	53.06
2 offers (Segment 2)	6,578 (20.88%)	3,775 (21.91%)	57.39
≥3 offers (Segment 3)	1,981 (6.29%)	1,275 (7.4%)	64.36

Fig. 4 and Table 16 show that the exact number of cases for each segment. It is easily shown that the pending rate increases when the cases have more offers. The pending rate for segment 3 is more than 2 is 64.36%. It is significantly higher than others. In following subsections, we analyze each segment about several perspectives.

4.3 Segment 1 – Single offer cases

Table 17. Basic statistics for segment 1

	- 20,000	- 40,000	40,000 -	Top 3 loan goals
Total	16,124 (70.26%)	5,179 (22.57%)	1,647 (7.18%)	• Car (27.48%) • Home Improvement (26.25%) • Existing loan takeover (17.69%)
Pending	8,428 (69.21%)	2,937 (24.12%)	813 (6.68%)	

In segment 1, there are 22,950 cases and 53.06% of them, 12,178 cases, include A_Pending activity. Practically, all segments' process maps are very similar to each other. In order to find difference among them, we focus 3 work items, W_Validate appliation, W_Call incomplete files and W_Assess protential fraud. There are several reasons of choosing them. First, all cases which include A_Pending activity also include W_Validate application. We think the cases which do not pass W_Validate application are not significant because the customers of those cases do not submit required documents to the company. Second, those 3 activities are executed by employees and take relatively long time to execute. The mean duration times of those 3 activities are 19.2, 6.5 and 41.2 hours respectively. We think those 3 activities require more duration time when the more offers are made.

Fig. 6. Process map for segment 1 and focused area on process map

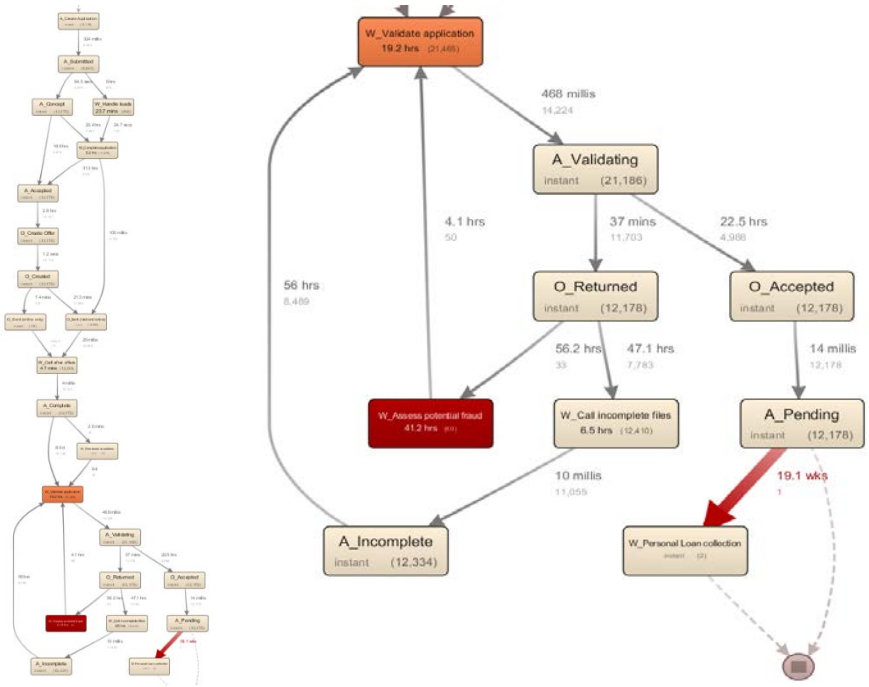


Fig. 5 shows that whole process map and focused area for segment 1. There are 12,178 cases including W_Validate application, 8,448 cases including W_Call incomplete files and only 62 cases including W_Assess potential fraud. We additionally separate cases to 3 different groups whose requested amounts are less than 20,000, between 20,000 and 40,000 and more than 40,000 respectively. In additional analysis, we consider requested amount as the key factor of duration time difference between each group.

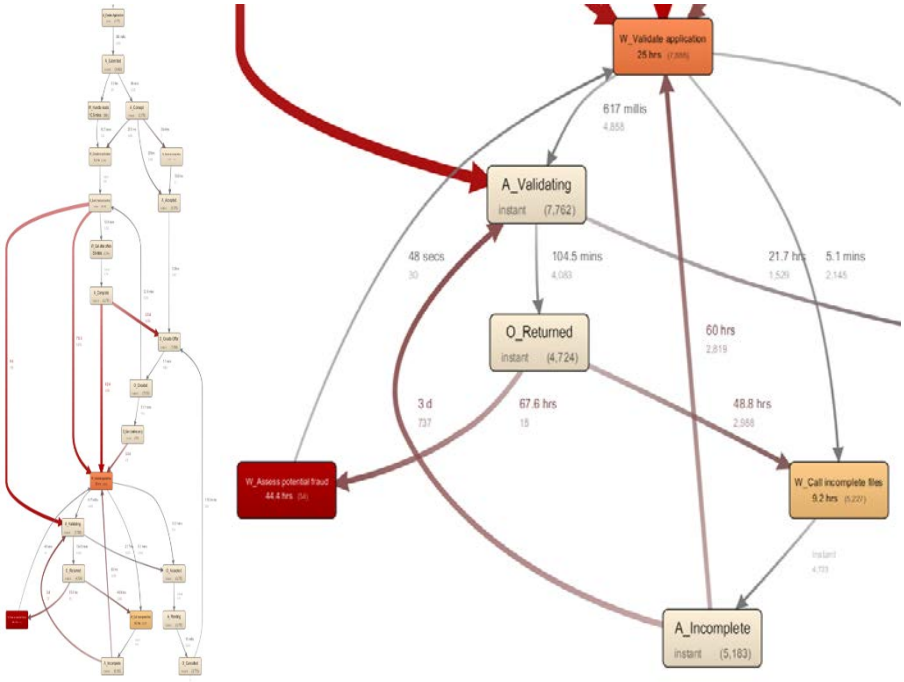
4.4 Segment 2 – 2 offers cases

Table 18. Basic statistics for segment 2

	- 20,000	**- 40,000**	**40,000 -**	**Top 3 loan goals**
Total	4,369 (66.42%)	1,611 (24.49%)	598 (9.09%)	• Car (26.92%) • Home Improvement (25.26%) • Existing loan takeover (20.06%)
Pending	2,431 (64.4%)	955 (26.36%)	349 (9.25%)	

In segment 2, there are 6,578 cases and 57.39% of them, 3,775 cases, can take loans from the company. We also focus the 3 activities in segment 2. Mean duration times for 3 activities are 25.0, 9.2 and 44.4 hours respectively.

Fig. 7. Process map for segment 2 and focused area on process map



In segment 2, 3,775 cases include W_Validate application, 3,076 cases include W_Call incomplete files and only 30 cases include W_Assess potential fraud. We can check that 2 offers cases' pending rate is higher than single offer cases' (53.06% for segment 1 and 57.38% for segment 2).

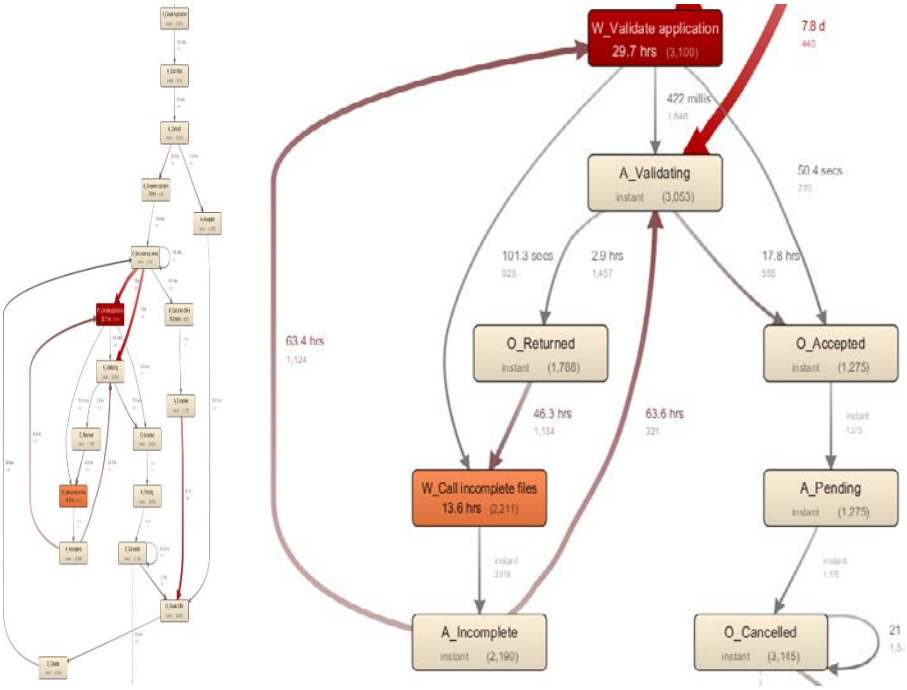
4.5 Segment 3 – more than 2 offers cases

The last segment is set of cases which have more than 2 offers (maximum 10 offers) cases. In this segment, there are 1,981 cases and 64.36% of them, 1,275 cases, include A_Pending activity. The cases in this segment have the highest pending rate among all segments.

Table 19. Basic statistics for segment 3

	- 20,000	- 40,000	40,000 -	Top 3 loan goals
Total	1,254 (63.3%)	529 (26.7%)	198 (9.99%)	• Car (26.32%) • Home Improvement (23.8%) • Existing loan takeover (21.06%)
Pending	791 (62.04%)	357 (28.0%)	127 (9.96%)	

Fig. 8. Process map for segment 3 and focused area on process map



In segment 3, there are no cases including W_Assess potential fraud, 1,275 cases including W_Validate application with 29.7 hours mean duration time and 1,123 cases including W_Call incomplete files with 13.6 hours mean duration time. Before analyzing, we think that the company seriously ask assessment for potential fraud to cases in segment 3, but no case is asked for that activity.

4.6 Analysis

Table 20. Summary of results

	- 20,000	**-40,000**	**40,000 -**	**W_Validate application**	**W_Call incomplete files**	**W_Assess potential fraud**
Seg. 1	8,428 (69.21%)	2,937 (24.12%)	813 (6.68%)	19.2 h	6.5 h	41.2 h
Seg. 2	2,431 (64.4%)	955 (26.36%)	349 (9.25%)	25.0 h	9.2 h	44.4 h
Seg. 3	791 (62.04%)	357 (28.0%)	127 (9.96%)	29.7 h	13.6 h	-

For all cases in all segments, there are not significant difference from process map. Because, both single offer and multiple offers cases should follow standard process to take loans. The differences can occur in our focused area which includes W_Validate application, W_Call incomplete files and W_Assess potential fraud. The more offers are made, the more number of those activities are required. However, we can find several trends among 3 segments. First, the pending rate increases when the number of offers increases, 53.06%, 57.39% and 64.36% respectively for each segment. We can interpret this phenomenon that the more number of offers can increase the probability that the customers accept one of those offers. Second, the requested amount also increases when the number of offers increases. According to Table 20, in segment 1, 69.21% of cases ask less than €20,000 and only 6.68% of cases ask more than €40,000. However, in segment 3, 62.04% of cases ask less than €20,000 and 9.96% of cases ask more than €40,000 to the company. We think that if the requested amount goes higher, then the company will be able to create the more number of portfolio of offers and suggest to their customers. Finally, the mean duration time of W_Validate application, W_Call incomplete files and W_Asseses potential fraud also increase. It is reasonable that the more number of offers need the longer time to validate documents, credit risk, and so on. For all segments, top 3 loan goals are same, Car, Home improvement and Existing loan takeover.

4.7 Additional analysis – requested amount

For the company, the requested amount of each case is very important since the requested amount is related to potential revenue and loss. Therefore, the company wants to thoroughly validate the customers whose requested amount are high. We think it can make differences in process.

In Application event log file, there are 31,509 cases with 701 different requested amounts. The requested amounts are between 0 and 450,000. Our purpose of this additional analysis is checking differences in process among cases whose requested amounts are different.

In order to analyze, we also do preprocessing. We create 3 different segments, one is set of cases whose requested amounts are between 0 and 20,000, another is set of cases whose requested amounts are between 20,000 and 40,000 and the other is set of cases whose requested amounts are more than 40,000 (maximum 450,000).

Fig. 9. The number of cases for each segment

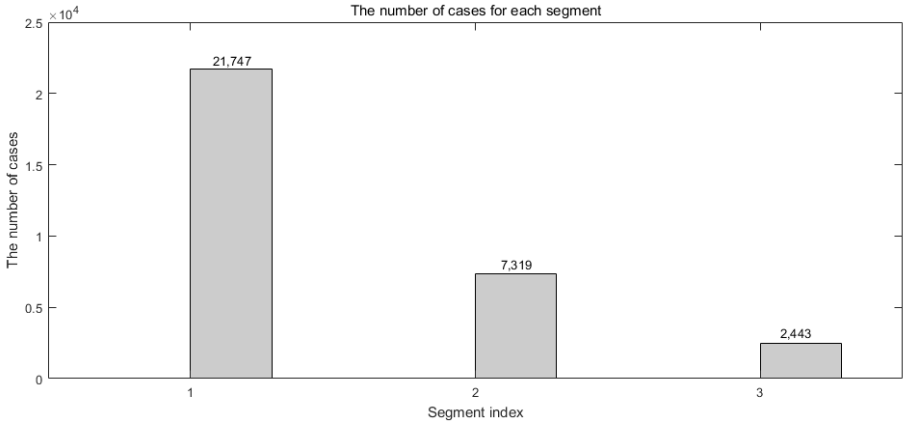
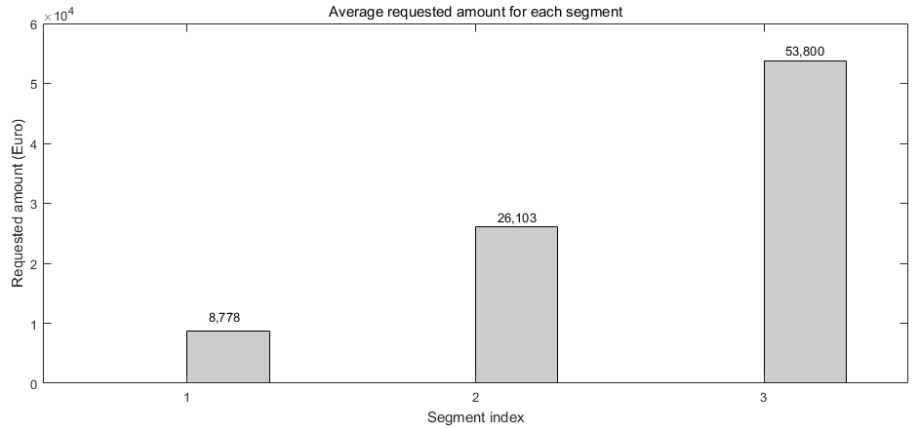


Fig. 8 shows the number of cases for each segment. 21,747 cases have requested amount less than 20,000, 7,319 cases' requested amount are between 20,000 and 40,000 and 2,443 cases' requested amount are more than 40,000. Fig. 9 shows the average requested amount for each segment.

Fig. 10. The average requested amount for each segment



According to Table 21, we can check same result of analysis in subsection 4.6. The customers ask the more requested amount, the company makes the more offers to the customers. Reject rates (100 – Pending rates) are not significantly different among all segments. And, mean duration time and median duration time increase when the customers ask the more requested amount to the company. It is also same result in subsection 4.6.

Table 21. Basic statistics for each segment

	Single	**2 offers**	**≥ 3**	**Reject**	**Mean**	**Median**

	offer (%)	(%)	offers (%)	rate (%)	duration	duration
Seg. 1	74.1	20.1	5.8	42	21.3 d	18.4 d
Seg. 2	70.8	22.0	7.2	38	22.5 d	20.0 d
Seg. 3	67.4	24.	8.1	40	24.5 d	22.9 d

For segment 1 and 2, top 3 loan goals are Car, Home improvement and Existing loan takeover, but for segment 3, top 3 loan goals are Existing loan takeover, Home improvement and **Other, see explanation**.

Fig. 11. Process maps for each segment

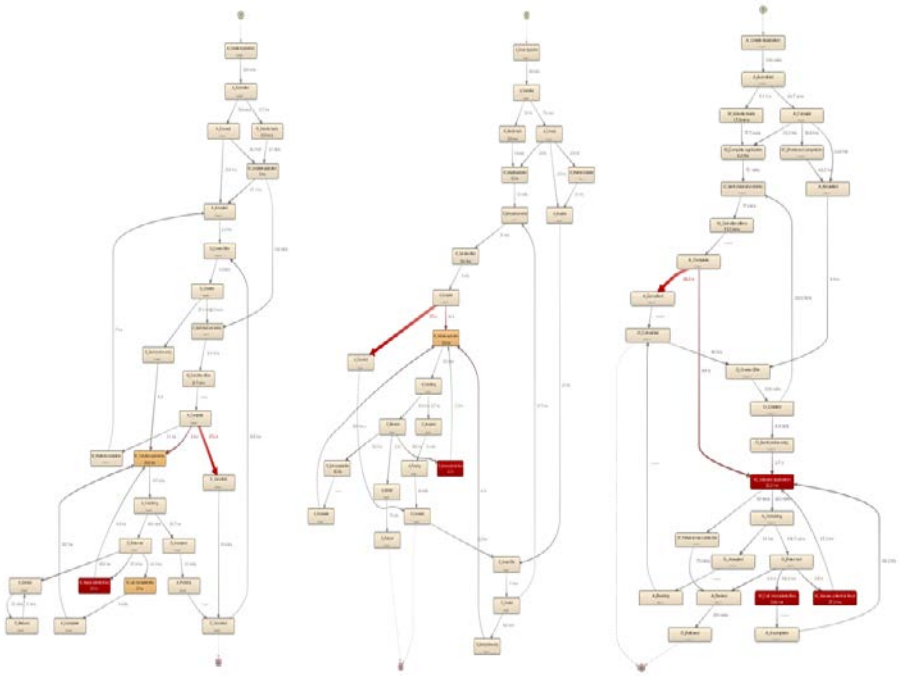
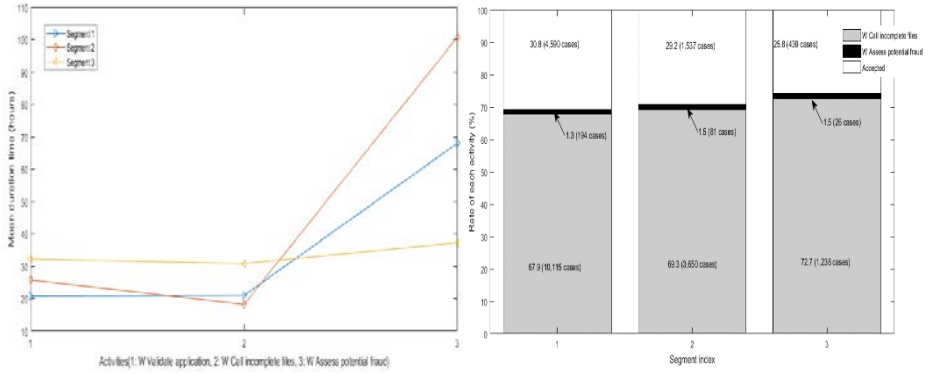


Fig. 10 shows that whole process maps for each segment. Positions of activities are different, but kinds of activities and sequences are very similar to each other. Thus, in this analysis, we also focus 3 kinds of activities, W_Validate application, W_Call incomplete files and W_Assess potential fraud. According to process maps, there are 3 main paths after passing W_Validate application. First one is the path following <A_Validating, O_Accepted. ...>, second one is the path following <A_Validating, O_Returned, W_Call incomplete files, ...> and the last one is the path following <A_Validating, O_Returned, W_Assess potential fraud, ...>. We firstly calculate mean duration times of W_Validate application ($d_{validate}$), W_Call incomplete files (d_{call}) and W_Assess potential fraud (d_{assess}) for each segment. Then, sum of 3 mean duration times can distort the results, because W_Assess potential fraud is very rare activity in all segments but it takes very long time to pass. In order to prevent distortion, we calculate weighted sum of duration times for 3 standard work items (T),

$$T = d_{validate} + \frac{n_{call}}{n_{call} + n_{assess}} \times d_{call} + \frac{n_{assess}}{n_{call} + n_{assess}} \times d_{assess}$$

for each segment and compare.

Fig. 12. Mean duration times and rates/the number of cases for each activities



The left figure in Fig.11 shows mean duration times of W_Validate application, W_Call incomplete files and W_Assess potential fraud. The right one shows the rates and numbers of cases for each activities in 3 segments.

Table 22. Statistics for each segment

	W_Validate app.		**W_Call incom.**		**W_Assess pot. fraud**	
	$d_{validate}$	$n_{validate}$	d_{call}	n_{call}	d_{assess}	n_{assess}
Seg. 1	20.8 h	14,899	21.0 h	10,115	68.0 h	194
Seg. 2	25.8 h	5,268	18.2 h	3,650	100.8 h	81
Seg. 3	32.2 h	1,703	30.8 h	1,238	37.3 h	26

By using data in Table 22, we can calculate weights for obtaining weighted sum (T) for each segment. For segment 1, the rate of W_Call incomplete files is 0.9812 and the rate of W_Assess potential fraud is 0.0188. Therefore, the weighted sum of mean duration times of segment 1 is 42.68 hours. For segment 2, the rates are 0.9783 and 0.0217. Then, the weighted sum of mean duration times of segment 2 is 45.79 hours. Finally, for segment 3, the rates are 0.9794 and 0.0206. The weighted sum is 63.13 hours. This result is reasonable because the company probably need more time for validating the customers who ask more requested amount. Moreover, for segment 3, the rate and the mean duration time of W_Call incomplete files are the biggest among 3 segments, 72.7% and 30.8 hours respectively. It means that the company wants to validate thoroughly the customers who apply more loans. The rate of W_Assess potential fraud is very small for each segment (less than 1.5%), so the mean duration time of that activity cannot affect significantly.

5 Organizational Analysis

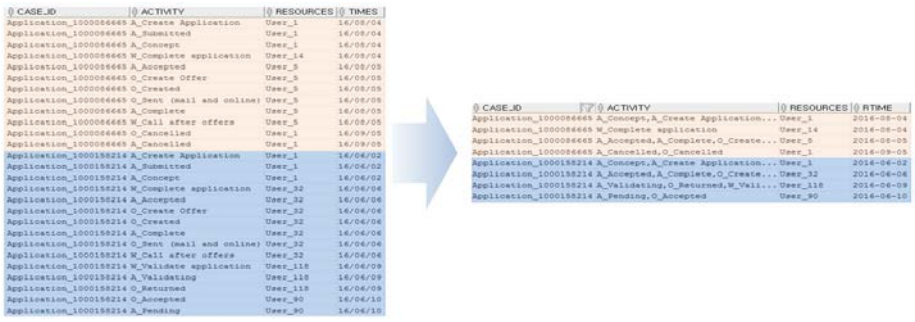
In addition to analysis about questions, we carry out organizational analysis about users in the whole data. We try to find kinds of teams or functional groups between users and derive several social networks. Unfortunately, there are too many resources and activities in data, precisely 145 users and 26 activities. Moreover, as most of resources execute more than half of activities. Averagely, one user deals with 16 activities. Thus, it is very hard to find some significant organizational insights among users when just exploiting raw data. It is necessary to take some approach to overcome those problems for carrying out organizational analysis.

In this section, we suggest two approach for organizational analysis. One is using preprocessed data for social network mining plug-in in ProM, the other is just concentrating on each single activity and its associated resources and making groups. From the former approach, we derive hand-over, subcontracting and working together social networks, and find several functional groups. Using later approach, we segment the activities with their associated resources and make 6 groups based on the type of activity.

5.1 Approach 1 - preprocessing

When looking into each cases in the data set, we can find some simple features that each user carries out successive activities. In other words, each user takes multiple activities consecutively when joins the cases. Considering that feature, we do preprocessing that we combine the activities which are consecutively carried by single user. Fig. 12 shows the dataset before preprocessing and after preprocessing. By the preprocessing, we can make suitable data for social network mining in ProM.

Fig. 13. Combining consecutive activities which are dealt with same user



5.2 Approach 1 – Social Network Analysis (Hand-over, Subcontracting)

By exploiting the preprocessed data, we derive hand-over, subcontracting and working together social networks. We except user 1 since user 1 is regarded as a

system of the company and it participates almost all cases and works both week day and weekend. In addition, we also remove the users who appear very rarely.

Fig. 14. Hand-over network with threshold > 0.0015 and Subcontracting network with threshold > 0.0004

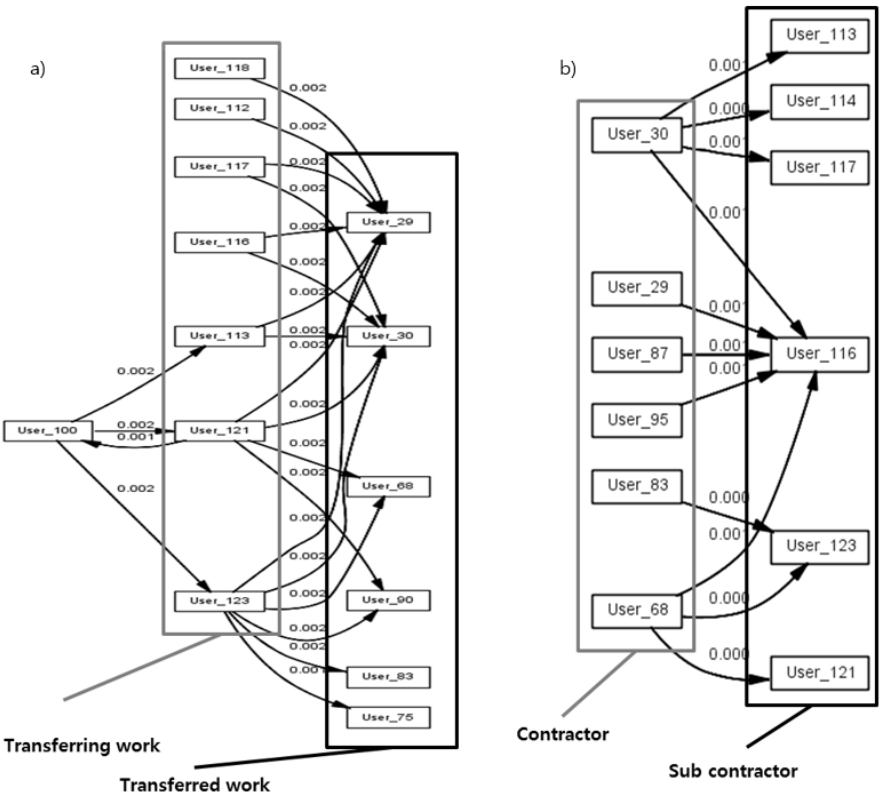


Fig. 13 is the hand-over network and subcontracting network between users whose frequency is between 200 and 10,000. From the hand-over network, we can observe two groups, transferring works and transferred group, while in the subcontracting network, we can observe the contractor and subcontractor. As the transferring group in the hand-over network and subcontractor group in subcontracting network are similar. Likewise, the transferring group and contractor group are also very similar. From this insight, we can infer that users in contractor groups may be superviosr of subcontractor groups or just teammate. No matter what relationship there is, those two groups might have some relationship in the process. When just checking the loan goal of each case in hand-over network, we can recognize the percentage of loan goal in each group is different.

Table 23. The number of cases in each loan goal including the hand-over relation

Transferred work group's member	Car	Home Improvement	Existing Loan Takeover
User_30, User_68, User_90	486	418	348
User_29, User_75, User_83	414	437	383

Table 23 show the number of cases in each loan goal including the relationship in the hand-over network based on transferred work. The groups including User_30, User_68 and User_90 carry out the cases whose the most loan goals are Car, while the other groups have Home improvement as the most loan goal. Although the difference between the number of each case by loan goals is large, considering Car is the most frequent loan goal for the whole cases, later groups' feature that Home improvement is more than others might be meaningful.

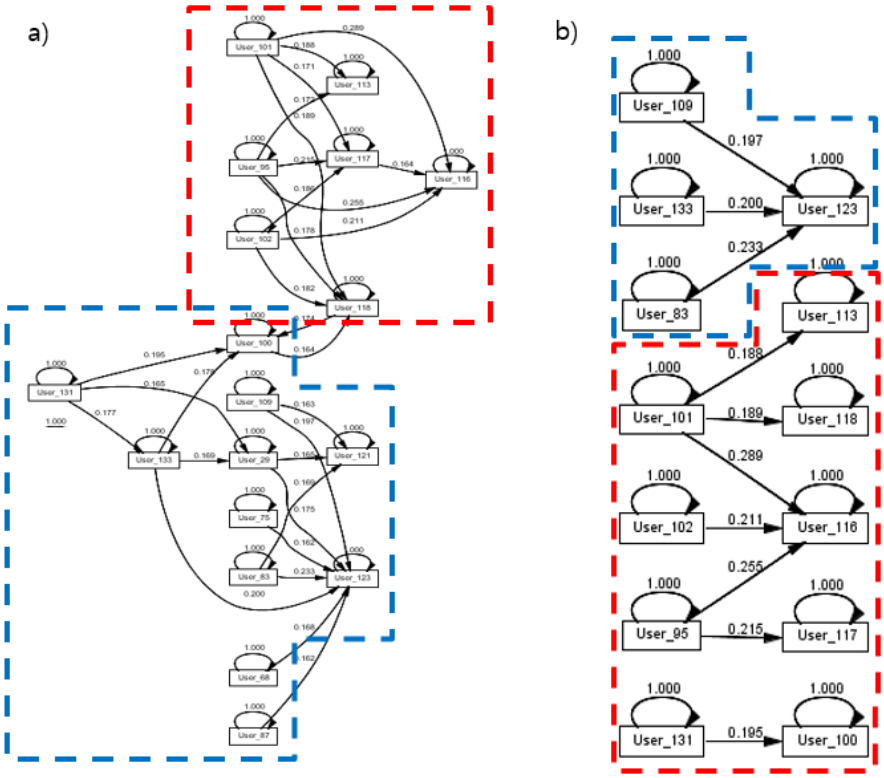
5.3 Approach 1 – Social Network Analysis (Working together)

This stage, we construct the working together network which shows the relationship between users who work together. Fig.14 is the working together network whose frequency is between 1,000 and 10,000. From the Fig.14, we can identify two groups one is red, the other is blue. Users in red group deal with cases related to Car more, while users in blue group do cases related to home Improvement. Although the number of cases related to Home Improvement is smaller than Car in blue group when confirming the Table 24, the difference between Home Improvement and Car is low comparing to red groups. Furthermore, when increasing the threshold higher, Home Improvement cases overcome those of Car. Thus, we can infer that blue group is team which tends to take more Home Improvement cases, while red group do more Car cases.

Table 24. The number of each loan goal for each group in working together network

> 0.161	Groups	Car	Home Improvement	Existing Loan Takeover
	Red	**3,031**	2,478	1,703
	Blue	3,649	**3,510**	2,822
> 0.188				
	Red	3,739	3,201	2,150
	Blue	1,726	**1,744**	1,380

Fig. 15. Working together network with threshold > 0.161 and > 0.188



5.4 Conclusion of Approach 1

From the social network analysis, we recognize some hierarchical relationship and two kinds of team whose main loan goals are different. Although it is hard to assure there are teams which treat the cases with such loan goals, we can find such trend form social network analysis.

5.5 Approach 2 – Data segmentation

As most of resources are involved in most of activities, finding organizational perspective become challenging. Therefore, we first have to find out the patterns and see which resources are more involved with which activities. We think that if a resource execute an activity more than others and in higher frequency, it indicates that the resource's job is doing that activity than those who do it less frequently. Moreover, we think that the group of resources who are doing an activity more than others, it indicate that those resources are in same group and work together. Table 25

shows 3 activities and how some activities are similar in terms of resource involvement frequency.

Table 25. Resources (Users) for 3 activities and frequency

Activity	User	Freq	Activity	User	Freq	Activity	User	Freq
A_ Accepted	49	1200	A Complete	49	1242	O Accepted	29	1710
	3	1081		3	1131		109	1235
	10	1080		10	1100		90	1147
	42	849		42	864		75	1036
	61	799		18	829		68	971
	18	783		61	801		30	970
	5	774		28	788		102	943
	37	762		37	754		100	914
	. . .			. . .			. . .	

After finding out the frequency of activity execution for each resources, as mentioned above, some activities were executed by same resources with high frequency. Therefore, we clustered those activities together based on similarity of resources and frequency of execution. Although in all of these activities, a high number of resources were involved, we tried to ignore the resources which executed an activity not frequency. This helped us to only consider the main resources for each activity.

After clustering resources based on frequency for each activity execution, we derived 5 different groups and one group which included activities executed by specific group of resources that were not same any of other activities (Table 26). Looking at type of activities in each cluster it can be seen that:

Cluster 1: It includes 7 different activities, the type of activities in this cluster shows that the resources in this cluster are common (low ranking) resources, and they perform daily operative activities of the organizations.

Cluster 2: In this clusters, activities are related on granting loan, accepting offer, canceling offer or checking if application is complete or not. This indicates that, the users in this clusters are decision makers and probably higher ranking than first cluster resources.

Cluster 3: This cluster includes those activities that are related to validity of application. Activities like A_Validating or W_Validate application are among these activities. This indicate that resources in this cluster are related to audit section and responsible for application validity.

Cluster 4: Refuse of offer and denial of application is in this cluster. In the event log the frequency of these two activity is not high, meaning that, these resources handle rare cases of offer denial or application denial or offer refuse.

Cluster 5: Activities in this cluster are in the beginning of application process. Resources in this cluster are involved in early concept and checking of application and see whether these application can go further for offer receiving.

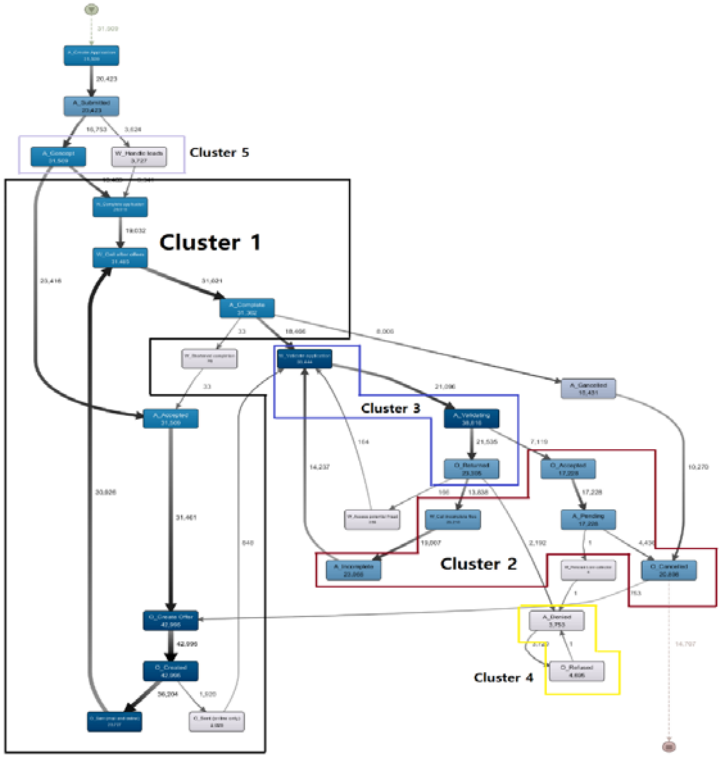
Cluster 6: In this cluster also some exceptional activities are included. However resources in this cluster are not same with one another and for each activity resources are different. Shortening application procedure, assessing frauds and application cancel have different resources each. Again the low frequency of each of these activities in dataset implies that resources in this cluster deal with exceptional activities.

Table 26. 6 Clusters based on activities

Cluster1 (Normal users)	**User**	**Cluster2 (Decision makers)**	**User**	**Cluster3 (Checking validity)**	**User**
A Accepted **A_ Complete** **O Create Offer** **O Created** **O_ Sent(mail and online)** **'W Call after offers** **W_Complete application**	User 49 User 3 User 10 User 42 User 61 User 18 User 5 User 37 . . .	**A Pending** **O Accepted** **O Cancelled** **A Incomplete** **W Call incomplete files**	User 29 User 109 User 90 User 75 User 68 User 30 User 102 User 100 . . .	**O Returned** **W Validate application** **A Validating**	User1 23 User1 21 User1 18 User1 16 User1 13 User1 26 User1 12 User1 17 . . .
Cluster4 (rejection of applications and offers)	**User**	**Cluster5 (Beginning of process)**	**User**	**Cluster6 (Special cases)**	**User**

A Denied O Refused	User 68 User 87 User 75 User 30 User 99 User 83 User 109 User 95 . . .	**W Handle leads A Concept**	User 14 User 2 User 5 User 16 User 28 User 26 User 18 User 46 . . .	**W Assess potential fraud W Shortened completion A Cancelled**	Each of these activities are unique in terms are resource associati on

Fig. 16. 4 main activities of each cluster in process map

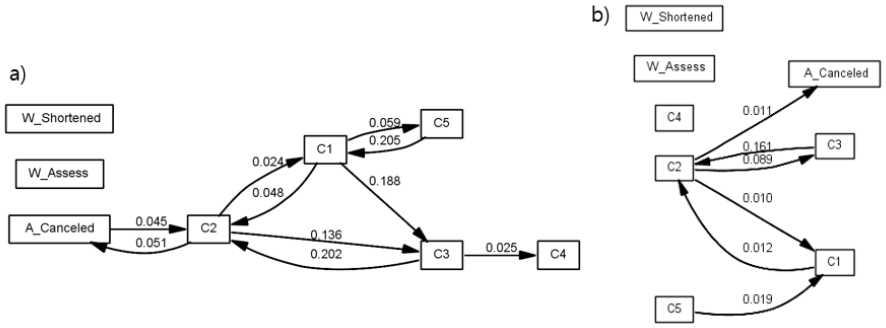


5.6 Approach 2 - Analysis

After we do data preprocessing same as approach, we make hand-over network and subcontracting network. Fig. 16 is the social network about hand-over and subcontracting. From the hand-over network, {"A_Canceled", C2(rejection team)}, {C1(normal users), C2(rejection team)}, {C2(rejection team), C3(Checking validity)} and {C1(Normal users), C5(Beginning of process)} show that they exchange their work each other. C2 shows hand-over relationship with "A_Canceled", C1(normal users) and C3(Checking validity). It means that decision making is usually done after normal users' works, beginning of process and checking validity, and after decision making those clusters' works can appear again. In short, it simplifies complex relationship of multiple offering process. From the relationship between C3 and C4, we can recognize that rejection of application and offers, after checking validity. Considering the subcontracting network c4 is independent clusters which doesn't show much relationship with other clusters. Thus, we inferred c4 as the activity of system as the activity seem to appear automatically.

From the subcontracting social network, we can confirm the result of hand-over again, decision make interact with "A_Canceled", "C3", and "C1". As C2 is contractor of "A_Canceled" and "A_Canceled" doesn't show any significant relationship with other clusters, users of "A_Canceled" is just confirm the result of decision workers. C5 is also contractor of C1 which means users in C5 may be the supervisor of C1.

Fig. 17. Hand-over network with threshold > 0.011 and subcontracting network with threshold > 0.01



5.7 Conclusion of Approach 2

By segmenting users based on their activity, we can simplify the relationship in process model. C2, decision makers, play a prominent part of process that it shows located in center between three main clusters, "A_Cancelled", C1(Normal users) and C3 (Checking validity).

6 Conclusion

In this challenge, we try to answerw 3 questions given by the company and analyze additional analysis for several insights. For question 1, we divide activities whether they are executed by the company (Users) or the customers (Applicants). After that, we calculate waiting times for each activity. In addition, we create 3 segments which include cases with particular waiting time ranges and analyze which activities take long time in each segment. For question 2, we identify the hypothesis which is referred by the company about the influence on the frequency of incompleteness. According to the company, the frequency of incompleteness goes higher, then the more number of customers withdraw taking loans. However, according to our analysis, the frequency of incompleteness affects positively to the pending rate of the customers. For question 3, we compare conversions from single offer cases and multiple offers cases. There are not significant difference between single and multiple offers cases, since all customers should pass standard process to take loans from the company. The only differecne is the number of validating processes which is caused by difference in the number of offers. Finally, we also do organizational analysis such as social network analysis and so on. We can find some groups thorugh this analysis.

7 Reference

van der Aalst, W. M., Reijers, H. A., Song, M. “Discovering social networks from event logs.” Computer Supported Cooperative Work (CSCW), Vol. 14, No. 6, pp. 549-593, 2005.

Song, M., van der Aalst, W. M. “Towards comprehensive support for organizational mining.” Decision Support Systems, Vol. 46, No. 1, pp. 300-317, 2008.

Song, M., van der Aalst, W. M. “Towards comprehensive support for organizational mining.” Decision Support Systems, Vol. 46, No. 1, pp. 300-317, 2008.

Song, M., van der Aalst, W. M. “Towards comprehensive support for organizational mining.” Decision Support Systems, Vol. 46, No. 1, pp. 300-317, 2008.

van der Aalst, W. M., Weijters, A.J.M.M., Maruster, L. “Workflow mining: discovering process models from event logs.” IEEE Transactions on Knowledge and Data Engineering 16 (9) (2004) 1128–1142