Application of Process Mining in a Consumer Loan Company Process

Minji Kwon[1], Minchul Lee[1], Hokyum Kim[1]

[1] Pohang University of Science and Technology, Industrial and management engineering, 77 Cheongam-ro Pohang, Republic of Korea,
minjkwon@postech.ac.kr
bonoyap@postech.ac.kr
variety2001@postech.ac.kr

Abstract. The incorporation of data analysis in business is imperative. In this report, a real data from a financial institution is used to analyze business process. The event log has 561,671 events, 31,509 cases, and 26 activities. We used Disco and ProM tool for exploratory analysis and traditional process-mining approaches including throughput times, bottleneck, and process centered questions listed in Business Process Intelligence Challenge(BPIC) homepage. From those analysis, we aim to find any process improvement or suggestion to increase customer loan success case.

Keywords: Petri nets, Process mining, Dotted chart, Bottleneck analysis, Performance analysis, Social network analysis

1 Introduction

As the role of data increases, there are various trials to apply innovative methods to real world data to find value. Business process mining, or process mining for short, aims at the automatic construction of models explaining the behavior observed in the event log[1]. The event logs capture information about activities performed. Each event records the execution of an activity instance by a given resource at a certain point in time along with the output produced[2]. In the area of process discovery, emphasis in the field has been on comparatively basic performance metrics (e.g. throughput times, working times, and waiting times) and not so much on analysis of more advanced performance-related behavior of processes[3]. The 2017 Business Process Intelligence Challenge (BPIC 2017) is one of the example of them. It aims to analyze the data of financial institution in the Netherlands. Their domain is online based consumer credit. Starting from the understanding of the data in respect of process, we tried to find business insights to apply.

1.1 Understanding of the data

The company is a financial institute which is located in Netherlands. This company is specialized in customer credit and online based. The data offered from the company is a log data which is about customer loan process. The company evaluates the customer credit and provides more than one offer. Customers receiving that offer decides whether accept or deny the offer. There are whole data set and reduced data set which is related to loan offer information. The whole data has 561,671 events, 31,509 cases, 26 activities, and 4,047 variants. The offer data contains 193,849 events, 42,995 cases, 8 activities, and 16 variants. Activity name which starts with 'A' means application events, 'O' means offer event, and W means work item events. The detailed description we guess is in Table 1.

Table 1. Activity description

Activity	Frequency	Description
A_Accepted	31,509	application finalized after passing screen for completeness
A_Cancelled	10,431	never sends in his documents or calls to tell he doesn't need the loan
A_Complete	31,362	automatic application assessment is positive
A_Concept	31,509	first assessment has been done automatically
A_Create Application	31,509	customer write application
A_Denied	3,753	application is declined/denied by the bank
A_Incomplete	23,055	documents are not correct or some documents are still missing. customer need to send in documents
A_Pending	17,228	offer was accepted by the applicant
A_Submiteed	20,423	initial application submission
A_Validating	38,816	bank validating application
O_Accepted	17,228	end state of successful offer
O_Cancelled	20,898	offer was sent to applicant who did not in reply in time
O_Create Offer	42,995	bank creates offer
O_Created	42,995	offer is created
O_Refused	4,695	the offer is refused by the bank
O_Returned	23,305	The customer returns the offer
O_Sent(mail and online)	39,707	An employee of the financial institute sends the offer
O_Sent(online only)	2,026	An employee of the financial institute sends the offer
W_Assess potential fraud	355	Investigating suspect fraud cases
W_Call after offers	31,485	creating offers, notify customer with call
W_Call incomplete files	23,218	seeking additional information during assessment phase
W_Complete application	29,918	completing pre-accepted applications
W_Handle leads	3,727	following up on incomplete initial

		submissions
W_Personal Loan collection	4	applications for a personal loan
W_Validate application	39,444	assessing the application
W_Shortened completion	76	shortened process

Analyzing the data, the big picture of the process is in Fig. 1.

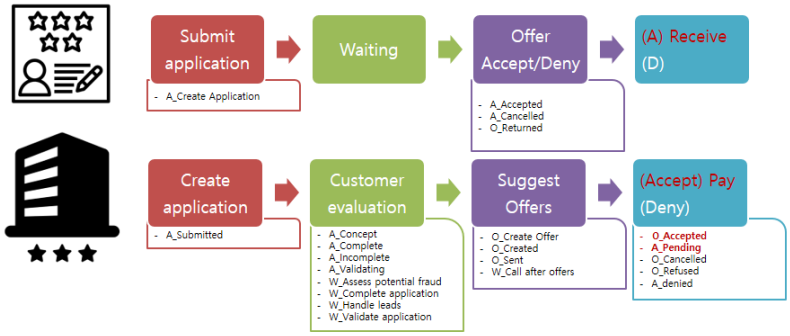

Fig. 1. Process of the company

2 Basic Analysis

In this section, we introduce basic analysis to understand the process further.

2.1 Xdotted Chart

We first performed an Xdotted Chart analysis using a given log. We can observe the whole flow through the Xdotted Chart. The basic XDotted Chart analysis yielded the Fig. 2.

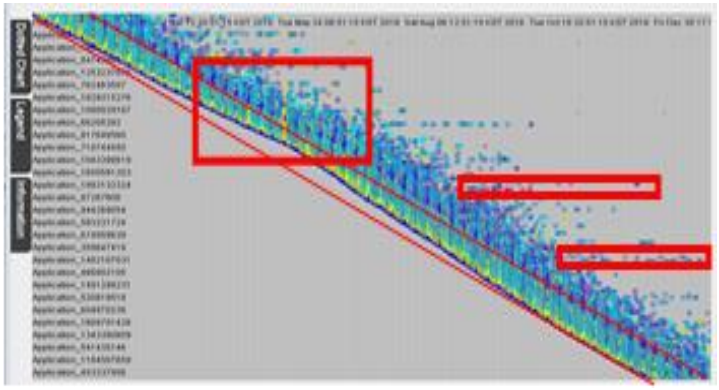

Fig. 2. Basic Xdotted Chart

This picture shows the followings. Firstly, In some cases (applications), the process proceeds for a long time. Secondly, The arrival rate changes at some point. Thirdly, In some cases, the same activity occurs at a specific time. However, there is a limit to the information that we can get with the Xdotted Charts that we see now. Therefore, we extended the Xdotted Chart by feature. Therefore, Let me introduce some of the same characteristic pictures.

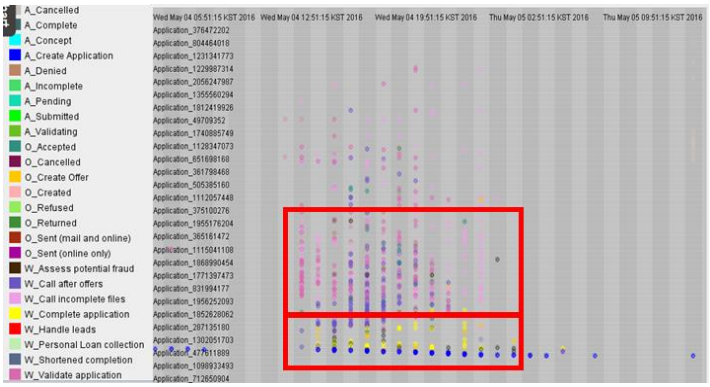

Fig. 3. Xdotted Chart by feature

First, in the case of the above Fig. 3, the timestamp of the activity is displayed (in the minimum unit) at 1 hour intervals. In addition, based on the color, it can be confirmed that the event progresses in the 'blue → yellow → pink' as a whole. Among them, the activity of pink series takes a relatively long time and many events are performed.

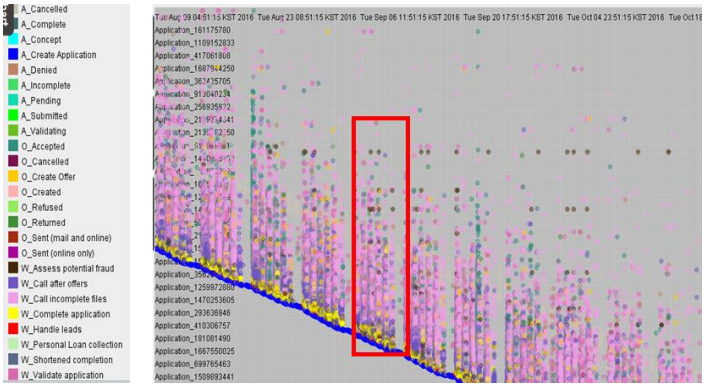

Fig. 4. Xdotted Chart by time

You can get more information from the following Fig. 4. it can be assumed that the empty space means the weekend because the empty space is formed at intervals of about 7 days on the horizontal axis. Also, on weekends, there is very little activity going on, and even if it goes on, very few or A_Create applications (activities that receive customers) will be processed.

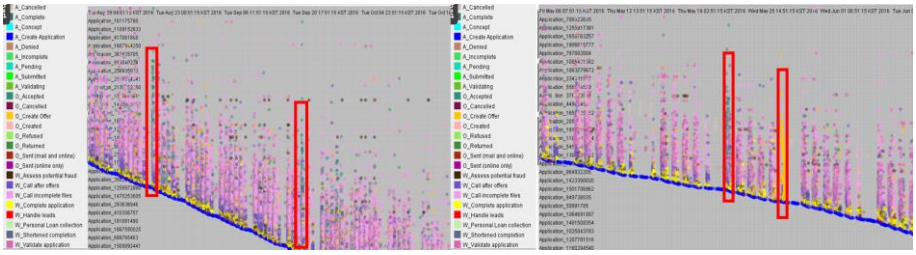

Fig. 5. Xdotted Chart by time

Finally, in the above picture, we can see that one activity proceeds in almost all cases at a certain time. One of the highlighted areas here looks closer to A_Complete (colors closer to green). In other words, although enough precomputation is made for A_Complete, it can be deduced that computation processing is performed at the same time in a certain period, and it is close to one month interval in terms of interval. Also, the orange highlight means O_Create Offer. In other words, the A_Complete and O_Create Offer mentioned above are only performed at a specific time. Therefore, we have conducted additional analysis through Disco to increase the credibility of this analogy. When you check with Disco, you can actually increase the credibility of the hypothesis that A_Complete is once a week and O_Create Offer is once a month through Max, Min duration.

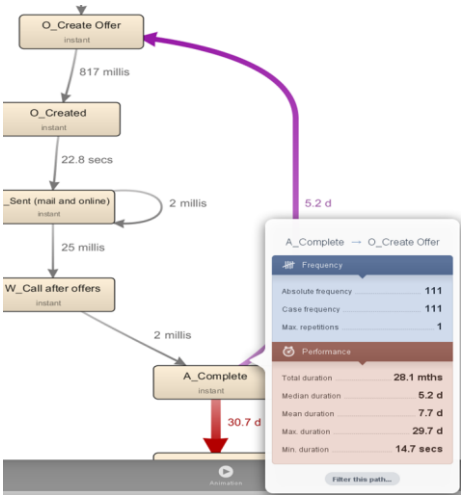

Fig. 6. Disco process map

2.2 VIP Analysis

The next analysis is VIP analysis. The VIP analysis is an analysis of the process that the customer goes through, literally creating a customer with a high loan volume. The starting point of the analysis is that 'it would be better for banks to attract a few VIP customers than a lot of ordinary customers', and 'if we improve the VIP process, it would be more efficient for the institution.' Therefore, this analysis assumes that the process throughput time of a customer with a high credit score or a customer with a large loan volume is important for a bank. For this analysis, the loan volume is filtered based on the credit score because it provides a loan volume that is higher than the actual credit score. And this Loan volume threshold is based on more than 30,000. Then, only the customers exceeding the reference point are filtered to derive the process map.

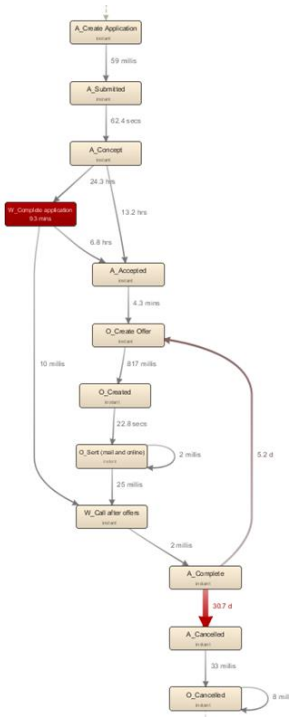

Fig. 7. Disco process map ; VIP

This can yield the following results. Disco analyzes the throughput time of two activities, which is not the only throughput time. As a result, we judged that the ratio of throughput time to the time spent in the path is about 110 times when the activity is not computed immediately. Therefore, we have assumed the actual throughput time through this ratio as in Table 2.

Table 2. Waiting and throughput time

Previous Activity	Next Activity	Waiting Time
A_Concpet	A_Accepted	13.2 hrs
A_Concpet	W_Complete application	24.3 hrs
W_Complete application	A_Accepted	6.8 hrs
A_Complete	A_Create offer	5.2 days
A_Complete	A_Cancelled	30.7 days
Previous Activity	Next Activity	(Expected)Throughput time
A_Concpet	A_Accpeted	7.2 mins
A_Concept	W_Complete application	13.2 mins
W_Complete application	A_Accepted	3.7 mins
A_Complete	O_Create Offer	1.1 days

A_Complete	A_Cancelled	6.7 days

We (generally adopted median) have determined that the path that takes less than 1hr is not an important consideration in terms of improvement. Then, we can see that the Loan volume is larger in the parts that are not in the automatic processing as much as the whole, which is larger than the average. Among these, 'A_Concept → W_Complete application' and 'A_Complete → O_Create Offer' parts, which are not responsibilities by customers, can be regarded as time consuming parts on the system. In other words, it can be described as a part that can be improved from the standpoint of the company. The conclusion of the VIP analysis is as follows. Loan volume is a major customer for a bank that is lending a large sum. Therefore, if we only improve the bottleneck from the total throughput time, it may not be more effective than the financial indicator in terms of customer segmentation. However, if the path time required for a VIP (a customer with a large loan volume) is further improved, a certain financial benefit can be expected.

3 Process centered Analysis

In this part, unlike the previous part, we will answer our comments on the three questions (1 to 3) that are explicitly given in this BPI Challenge.

3.1 Throughput times

The first question is about throughput times per part of the process, in particular the difference between the time spend in the company's systems waiting for processing by a user and the time spent waiting on input from the applicant. We first analyzed the question itself. "What are the throughput times per part of the process?" I tried to think about throughput time. Throughput time means the time it takes for the computer to process the data when it occurs. Therefore, it means only rough time in activity when viewed from the engineering definition. Therefore, we decided to proceed with the additional analysis if the results of the analysis are different from the expected results. Basically, throughput time is calculated as median value, and additional analysis is possible through mean, max, and min values.
Most of the throughput time is derived instantly (median application) unlike the expectation. Therefore, it is possible to consider whether instant is appropriate or not, depending on the meaning of the activity. The analysis was similar to the VIP analysis of the previous part. First, from the result of basic statistical analysis and process map analysis Fig. 8, we can see that the throughput time of the activity, which is yellow and red, is not 0.

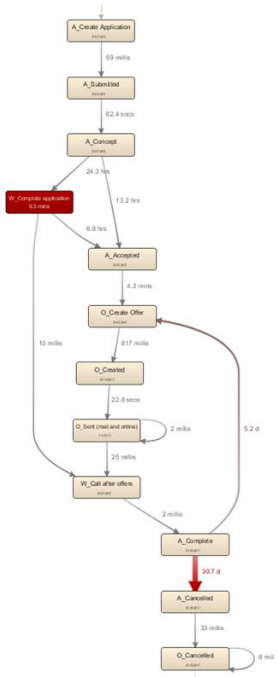

Fig. 8. Disco process map

Table 3. Throughput time of the activity

Case exceeds 100	
W_Handle leads	92.1 secs
W_Complete application	7.5 mins
Case exceeds 50	
H_Handle leads	93.3 secs
W_Complete application	7.5 mins

It is not significantly different from the previous 100 standards. This is becaus e most of the statistics are determined in variants with more than 100 cases.H ere we are wondering how many activities are instant, and some parts assume that some of the waiting time should be included in the throughput time. In a ddition, we conclude that the path that takes less than 1hr (generally = media n) is not an important consideration in terms of improvement.In order to make it the standard for this, we confirmed the case of W_Handle leads and W_Co mplete application which is the only non-instantaneous throughput time. In this case, the waiting time was 50-160 times the throughput time. Therefore, it is assumed that the waiting time of existing path and the throughput time of the activity have a ratio of about 110 times (54-164).

Table 4. Waiting and throughput time

Previous Activity	Next Activity	Waiting Time
A_Submitted	W_Handle leads	84 mins
A_Concpet	W_Complete application	20.5 hrs
A_Complete	W_Validate application	7.1 days
A_Complete	A_Cancelled	30.7 days
A_Validating	O_Accepted	6 hrs
O_Returned	W_Call incomplete files	26.4 hrs
Previous Activity	**Next Activity**	**(Expected)Throughput time**
A_Submitted	W_Handle leads	(real value) 92.1 secs
A_Concpet	W_Complete application	(real value) 7.5 mins
A_Complete	W_Validate application	1.6 hrs
A_Complete	A_Cancelled	6.7 hrs
A_Validating	O_Accepted	0
O_Returned	W_Call incomplete files	14.4 mins

If the throughput time is not instant, the waiting time in the remaining paths except the first and the second may be the time including the throughput time. In addition, since O is automatic processing using company resources, the throughput time can be set to instant. Therefore, among the Bottleneck activities seen in the results, W_Validate application and W_Call incomplete files are responsible for bank system rather than customer, and A_Cancelled is responsible for customer.
The conclusion is as follows. The above-mentioned throughput times are relatively long compared to other activities that are fast due to automation or simple tasks. Especially, in case of 'A_Complete → W_Validate application' and 'A_Complete → A_Cancelled', the throughput time is more time than the sum of all other median throughput times. Therefore, it can be concluded that if the path (or activity) is improved, the time cost of the whole process can be drastically reduced.

3.2 The influence on the frequency of incompleteness to the final outcome

The second question from the BPIC is "What is the influence on the frequency of incompleteness to the final outcome?" and the hypothesis from that question is If applicants are confronted with more requests for completion, they are more likely to not accept the final offer. To analyze the impact of the frequency of incompleteness to the final outcome, we define incompleteness and final outcome. The incompleteness means the status of 'A_Incompleteness'. So we count the frequency of A_incompleteness for each case. The range of this value is 0 to 7.

Table 5. Case ID and Frequency of incompletness

Case ID	Frequency of incompleteness

Application_652823628	1
Application_1691306052	0
Application_428409768	1
Application_1746793196	2
Application_828200680	0
⋮	⋮

To define final outcome, we analyzed the variants of Offer log data. All applicants get at least one offer and from the variants we figure out the end points.

Table 6. Final outcome of the offer process

Variant	Percent(%)	Process
1	38.1	Create Offer ⟶ Created ⟶ Sent(mail and online) ⟶ Cancelled
2	37.9	Create Offer ⟶ Created ⟶ Sent(mail and online) ⟶ Returned ⟶ Accepted
3	8.2	Create Offer ⟶ Created ⟶ Sent(mail and online) ⟶ Returned ⟶ Refused
4	5.6	Create Offer ⟶ Created ⟶ Sent(mail and online) ⟶ Returned ⟶ Cancelled
5	2.8	Create Offer ⟶ Created ⟶ Cancelled
6	2.2	Create Offer ⟶ Created ⟶ Sent(mail and online) ⟶ Refused
7	2.2	Create Offer ⟶ Created ⟶ Sent(online only) ⟶ Returned ⟶ Accepted
8	2.0	Create Offer ⟶ Created ⟶ Sent(online only) ⟶ Cancelled
9	0.3	Create Offer ⟶ Created ⟶ Sent(mail and online)
10	0.2	Create Offer ⟶ Created ⟶ Sent(online only) ⟶ Refused
11	0.1	Create Offer ⟶ Created ⟶ Sent(online only) ⟶ Returned ⟶ Cancelled
12	0.1	Create Offer ⟶ Created ⟶ Refused
13	0.1	Create Offer ⟶ Created ⟶ Sent(mail and online) ⟶ Returned
14	0.1	Create Offer ⟶ Created ⟶ Sent(online only) ⟶ Returned ⟶ Refused
15	0.04	Create Offer ⟶ Created ⟶ Sent(online only)
16	0.0	Create Offer ⟶ Created ⟶ Sent(online only) ⟶ Returned

From that data we define Final outcome into two categories which are success and unsuccess. Among the variants success is the process which contains activity of 'Accepted' as the end point. For example, variants 2 and 7. And the others are the case of Unsuccess. Our guesses are that the end point of 'Cancelled' means applicants didn't reply in time, 'Refused' means the offer is refused by the bank, 'Sent' means applicants stop replying, and 'Returned' means bank stop replying. We concatenate those above results for each applicants.

Table 7. Concatenated result

Case ID	Frequency of incompleteness	Offer result
Application_652823628	1	Success
Application_1691306052	0	Unsuccess
Application_428409768	1	Success
Application_1746793196	2	Success
⋮	⋮	⋮

And categorize the result with the respect of frequency of incompleteness.

Table 8. Categorized result

Frequency of incompleteness	Success	Unsuccess	Total
0	4581	11925	16506
1	7623	1638	9261
2	3463	506	3969
3	1113	151	1264
4	319	44	363
5	97	14	111
6	25	0	25
7	7	3	10

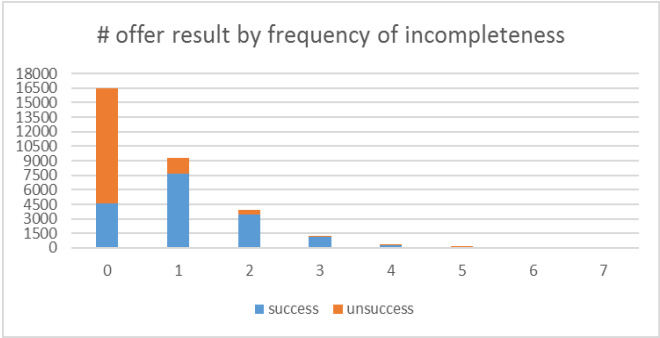

Fig. 9. Offer result by frequency of incompleteness

To unify the total frequency of each case, we normalized the result diving with the total number of corresponding incompleteness frequency.

Table 9. Normalized result

Frequency of incompleteness	Success	Unsuccess
0	0.2775	0.7225
1	0.8231	0.1769
2	0.8725	0.1275

3	0.8805	0.1195
4	0.8788	0.1212
5	0.8739	0.1261
6	1	0
7	0.7	0.3

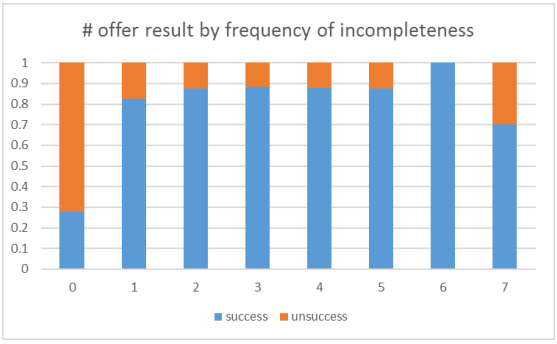

Fig. 10. Normalized offer result by frequency of incompleteness

The case of incompleteness 0 has lower success rate compared to the others which have about 70% success rates. So we can conclude that more requests for completion does not match with failure of the final outcome.

3.3 Frequency of offers and conversion

"How many customers ask for more than one offer?" To find an answer, we tried to find the number of offers for each Case ID. Activity named "O_Create Offer" means that customers are offered their loan. To eliminate customers who do not receive the offer, we filtered "Activity" column using Excel (Figure 1). This extracts "Case IDs" which have offer by the bank.

	A	B	C	D	E	F	G	H	I	J	K	L
1	Case ID	Activity	Resource	Start Ti	Comple	Variant	Variant	(case) A	(case) L	(case) R	Accepte	Action
7	Application_652823628	O_Create Offer	User_52	13:54.5	13:54.5	Variant 2	2	New credi	Existing lc	20000	TRUE	Created
26	Application_1691306052	O_Create Offer	User_38	03:40.8	03:40.8	Variant 38	388	New credi	Home imp	10000	FALSE	Created
40	Application_428409768	O_Create Offer	User_19	43:44.4	43:44.4	Variant 20	209	New credi	Home imp	15000	TRUE	Created
45	Application_428409768	O_Create Offer	User_19	43:44.4	43:44.4	Variant 20	209	New credi	Home imp	15000	TRUE	Created
64	Application_1746793196	O_Create Offer	User_19	46:52.0	46:52.0	Variant 25	256	New credi	Car	5000	FALSE	Created
69	Application_1746793196	O_Create Offer	User_12	46:52.0	46:52.0	Variant 25	256	New credi	Car	5000	FALSE	Created
91	Application_828200680	O_Create Offer	User_19	47:21.2	47:21.2	Variant 1	1	New credi	Home imp	35000	TRUE	Created
103	Application_1085880569	O_Create Offer	User_17	54:31.8	54:31.8	Variant 19	19	New credi	Existing lc	13000	TRUE	Created
108	Application_1085880569	O_Create Offer	User_8	54:31.8	54:31.8	Variant 19	19	New credi	Existing lc	13000	TRUE	Created
119	Application_1266995739	O_Create Offer	User_3	20:45.2	20:45.2	Variant 30	30	New credi	Existing lc	7000	FALSE	Created
136	Application_1878239836	O_Create Offer	User_3	21:25.2	21:25.2	Variant 9	9	New credi	Home imp	15000	TRUE	Created
151	Application_619403287	O_Create Offer	User_19	53:25.6	53:25.6	Variant 6	6	New credi	Car	15000	TRUE	Created
168	Application_1710223761	O_Create Offer	User_3	28:24.6	28:24.6	Variant 1	1	New credi	Car	11000	TRUE	Created
180	Application_1529124572	O_Create Offer	User_31	40:11.5	40:11.5	Variant 2	2	New credi	Other, see	5000	FALSE	Created
199	Application_387012864	O_Create Offer	User_17	40:57.3	40:57.3	Variant 1	1	New credi	Other, see	5000	TRUE	Created
210	Application_1120819670	O_Create Offer	User_17	57:00.2	57:00.2	Variant 23	23	New credi	Car	6850	TRUE	Created

Fig. 11. Log data filtered by O_Create Offer

After filtering, we deleted duplicated "Case IDs" and count the number of "O_Create Offer" activities using "countif" function in Excel. Table 10 shows the result of the number of "Case IDs" and the answer of the question is 8559.

Table 10. Frequency of Offer

Total number of case ID	Single conversion	Multiple conversion
31509	22950	8559

"How does the conversion compare between applicants for whom a single offer is made and applicants for whom multiple offers are made?"
In BPI forum, conversion means that the application gets to the endstate "A_pending", where the loan is actually paid out to the customer. So, we eliminated "Case IDs" which do not have "A_pending" activity using filter in Excel and divided "Case IDs". One is "Case IDs" which have single offer and the other is "Case IDs" which have multiple offers.
The ratio of success of loan is compared in table 11. This shows that multiple offer has more success in loan but the difference is small.

Table 11. Frequency of Offer and success

	Success	Unsuccess
Single offer (Total 22950, 72.8%)	12178 (53.1%)	8559 (46.9%)
Multiple offer (Total 8559, 27.2%)	5050 (59.0%)	3509 (41.0%)

We also compared single offer with multiple offer in terms of loan attributes. Table 12 shows 2 sample T-Test. We figure out that multiple offer shows more values for every attributes and we guess more requested amount, more offers for customers.

Table 12. T-test

	Requested amount	First withdrawal amount	Monthly cost	Number of terms	Offered amount
Single offer (average)	16466	8064	277	84.9	18682
Multiple offer (average)	17798	9282	286	87.5	19828
Difference (95%)	955.90	960.55	4.52299	1.72006	815.44
P-Value	0.000	0.000	0.000	0.000	0.000

We supposed that if customers require high amount of money, it would take more time. However, goal of loan is not related to process time and Fig. 10 shows the results.

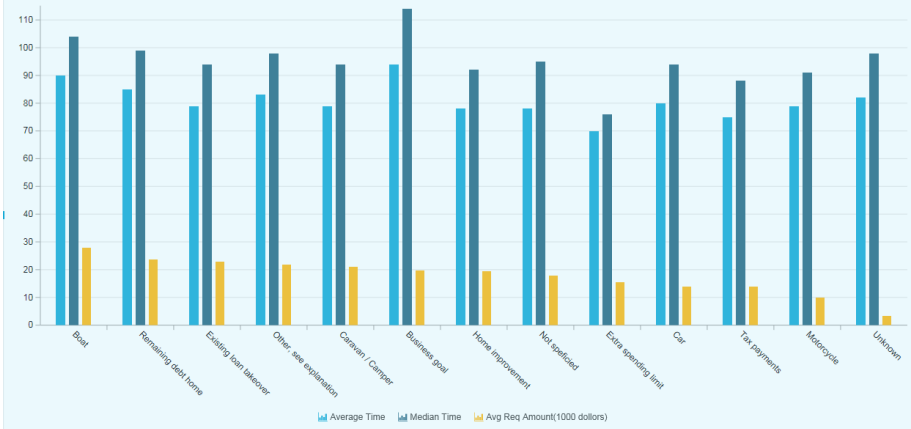

Fig. 12. Goal of loan

4 Conclusion

Through basic and process centered analysis of BPIC 2017 event log data, we dealt data with 561,671 events, 31,509 cases. We used Disco and ProM as tool for exploratory and process centered analysis.

As part of analysis, we analyzed Xdotted chart process map to find potential bottleneck in the process. If the path is improved, the throughput time of the whole process can be drastically reduced.

From the analysis of relation between frequency of incompleteness and final outcome, we found that applicant who got one offer has the lowest rate of success. and from the relation between frequency of offer and final outcome, there's little gap between success rate of applicant who got single offer and multiple offers.

In conclusion, from the BPIC 2017 we could verify how process centered approach can be applied to the real life example. Thanks to the useful tools that are Disco and ProM we could easily load and analyze data.

References

1. W.M.P. van der Aalst, H.A. Reijers, A.J.M.M. Weijters, B.F. van Dongen, A.K. Alves de Medeiros, M. Song, H.M.W. Verbeek. : Business process mining : An industrial application. Information Systems, vol. 32, pp. 713--732 (2007)
2. Massimiliano de Leoni, W.M.P. van der Aalst, Marcus Dees. : A general process mining framework for correlating, predicting and clustering dynamics behavior based on event logs. Information Systems, vol. 56, pp. 235--257 (2016)

3. Suriadi Suriadi, Chun Ouyang, W.M.P. van der Aalst, Arthur H.M. ter Hofstede. : Event interval analysis : Why do processes take time? Decision Support Systems, vol. 79, pp. 77--98 (2015)