Analyzing Loan Process with Process Mining Techniques

Kyounghoon Park , Heeje Lee, Hyemin Jung

Department of Industrial Management Engineering, POSTCH,
77 Cheongam-ro, Nam-gu, Pohang 790-784, Gyeongsangbuk-do, Republic of Korea
{pkh8820, zed, jhm0237}@postech.ac.kr

Abstract. Process mining is a way to analyze processes based on the event logs of the systems to support them. In general, process mining techniques are used for process discovery, conformance checking, predictive analytic of certain process. In this report, we used a real-life event log of the loan process in Dutch bank. Applying several process mining techniques to log data of loan process, we discovered process flow and several meaningful insights related to throughput times per part of the process, frequency of incompleteness to the final outcome and impact of the number of offer in which the company is particularly interested. We used Disco, ProM and other tools for analysis.

Keywords: Process mining, Loan process, Process discovery, ProM, Disco, BPIC 2017

1 Introduction

Until now, the data from the complex process were not usefully utilized. So company did not know what part of the process had problem and how to improve them. However, as process mining area has been growing since 1990s, it can give a chance to improve their process using data. Process mining used real-life log data for discovering process model, conformance checking and so on in order to analyze the complex real-life process.

In this challenge, we analyzed data from complex loan process. Dutch bank provided data for us and ask their main concerns in BPIC 2017 website. Actually, bank will know better about the main issues. However this is limited to their experience and if they use appropriately their data, in-depth and hidden knowledge could be extracted.

To solve three main concerns of bank, we analyzed data using several process mining techniques. After we deeply understood the data, we discovered process model of loan process using tools such as ProM and Disco. And also we performed further studies to solve three main concerns using tools.

2 Materials and Methods

2.1 Understanding of the data

Data was provided by Dutch banks and had many columns. The number of columns is 22. Raw data was categorized by case ID which means the customer who was involved in loan process. Each case ID had a serial process and the activities and other information were entered to bank system. Among the columns of data, there are five important columns when we analyzed the process.

First, 'Accepted' column is true/false binary item and means that if it is true, customer enable to select offers of bank and the bank changed the status of column depending on their validation results of customer. Basic status is false but if the certain offer was reasonable after validating customer's financial condition, the status is changed to true. Second, 'First withdrawal amount' column means that initial amount of withdrawal from customer. Third, 'Request amount' means amount of money that customers applied to the bank. Forth, 'Monthly cost' means amount owed by customers to monthly payment. Fifth, the number of terms means the number of months to repay the loan. 'Resource' column means the employ responsible for the work and it can be system of bank like PC. 'Loan goal' columns is purpose of customer's application of loan. 'Selected' means that customer finally select certain offer which the bank suggested. Remaining columns which were not mentioned this chapter were easy to understand or not important to analyze process.

2.2 Tools used for analysis

To analyze the process, we used many tools which were Disco, ProM, Excel and Python. Excel is a spreadsheet program for Windows environments developed by Microsoft. Python is open source advanced programming language based on C language. We used both tools for preprocessing which were extracting subset of entire data and deleting unnecessary data columns. Also we used python to perform data mining techniques for in-depth studies. Disco and ProM are widely used in process mining area developed by process mining group in Eindhoven university. We used Disco to see the rough process model and to determine the time spent in each process. And we used ProM to analyze process for discovery, conformance checking.

3 Understanding the Process

3.1 Understanding of Activity

Loan process were very complex and there are many exception and difference according to human beings. So a clear understanding of 'Activity' columns is important. Activity columns means each one step of loan process.

Application activity begins with the letter A and refer to states of the application itself and follow up of bank resources to complete the application where needed and also facilitate decisions on application.

Offer activity begins with the letter O and refer to states of an offer which is communicated to the customer.

Table 1 Explanation of acitivity of application, 'A_' and 'O_'

Activity	Definition
A_submit	Submit a new application to the website
A_Concept	When the customer has just finished handling in the bank, the bank notices that application documents have been received
A_Accept	With the application fully received, the stage where the bank can generate offer and customer wait
A_Complete	Offer is sent to the customer, bank waits for the customer to accept the offer
A_Validating	Customer's documents are submitted to the bank, the documents are been evaluating, and the customer is waiting
A_Incomplete	Status that customer's documents are not enough to evaluate
A_Pending	The documents submitted by the customer have been completed and the assessment has been completed. The evaluation is positive so the loan is prior to the execution or the customer has received a loan
A_Denied	The evaluation by bank is negative, so bank does not suggest other offer or conclude that they do not arrange the loan
A_Cancelled	Customers receive the offer that the bank offers, but they do not accept the offer or do not send the necessary documents to process the loan
O_Create offer	Generate offer within the bank to provide to the customer
O_Created	Generate offer to provide to the customer
O_Sent	Channel for suggesting the offer to customer
O_Refused	One of the offer which was suggested by bank was excluded due to customer's poor evaluation results
O_Cancelled	Customer cancelled the offer which was provided by bank
O_Accepted	Customer accepted offer which was provided by bank

Work item activity begins with the letter W and refers to states of work item that occur during the approval process and capture manual effort exerted by bank's resources

Table 2 Explanation of acitivity of application, 'W_'

Activity	Definition
W_Assess potential fraud	Bank decide whether the customer is a fraud during the assessment process
W_Call after offers	Bank call the customer to inform that the offer was suggested
W_Call incomplete files	Bank call the customer to request additional documents due to missing documents
W_Complete application	Handling the customer's application
W_Handle leads	Handling the customer's information
W_Personal loan collection	Collecting the customer's loan record
W_Shortened completion	Request simplification of process to system

3.2 Process Discovery

We used the Disco and ProM to discover the process model of loan data. We concluded that if there is A_Pending at least one time, the case succeed to a loan. There are two big categories of process depending on application types. One is new credit process and the other is limit raise process. However, almost process after O_Sent is similar in both cases. The process is classified into two part. One is before creating offer and the other is validating part. Process before the creating offer is same as almost loan process. After that, there is a great deal of cases depending on personal problem, personal financial condition and so on.

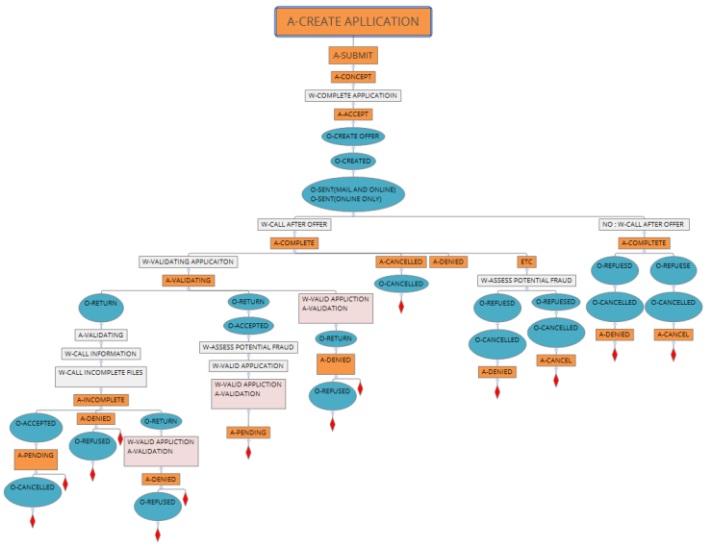

Figure 1 Discovered loan process model by Disco and ProM

4 Analysis of Process

4.1 Analysis of Performance

Question of this chapter is 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 application as this is currently unclear?

In order to find the solution of the above question, we try to compare the time the customer waits and the time the bank staff waits. First, we identify the entire process by analyzing each case in disco tool. So, we find the process in all cases. Second, we decided in the process where the customer waits and where the bank staff waits. We think throughput time means the sum of execution time and waiting time. Also, we assume the throughput time means the sojourn time in Prom tool.

There are two processes where the customer waits. First, customers wait until 'O-SENT' activity is over after 'A-SUBMIT' activity starts. The waiting time at the part of process is time from when the customer submits the loan application to receiving an offer from the bank.

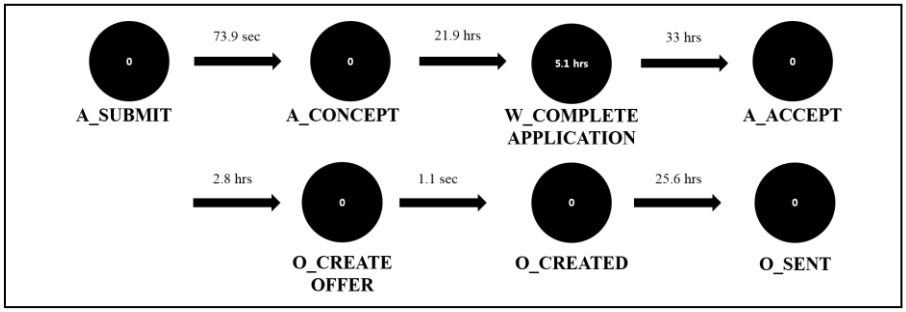

Figure 2 Execution time and waiting time in Disco

We found that customers waited the longest time in the process from W_Complete and W_Application activity to A_Accept activity

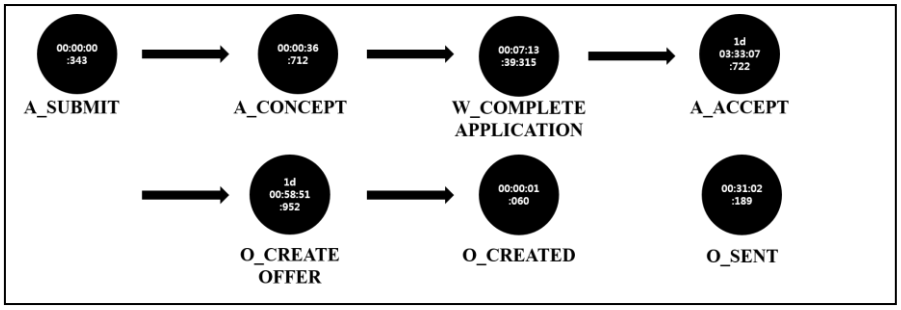

Figure 3 The sojourn time in ProM

We found that customers waited the longest time in the A_Accept activity.

Second, customers wait until A_validating activity is over after W_validating activity starts. This is time when the customer submits the application and then the bank requests evaluation and proceeds with the evaluation.

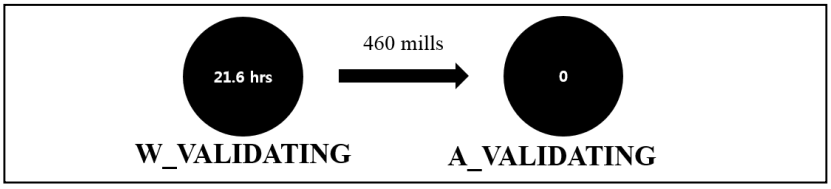

Figure 4 Execution time and waiting time in Disco

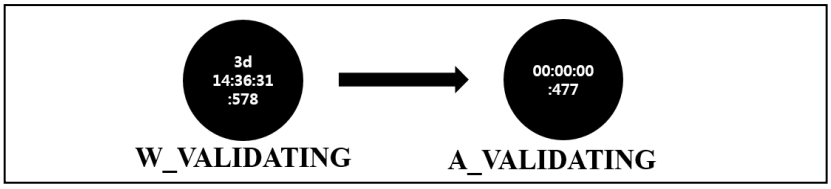

Figure 5 The sojourn time in ProM

There are two processes where the bank waits. First, bank waits until A_Complete activity is over after O_Sent activity starts. This is the time for all offers to be delivered to the customer and then wait for the customer to select an offer.

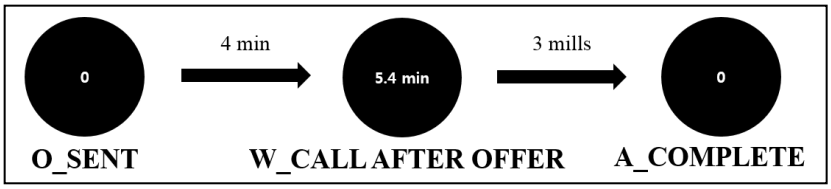

Figure 6 Execution time and waiting time in Disco

Figure 7 The Sojourn time in ProM

We found bank waits the longest time in the W_call after offer activity.

Second, bank waits until W_validating activity starts after A_Incomplete activity starts. This is the time to instruct the customer to submit.

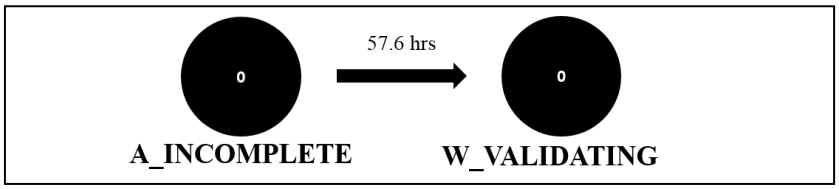

Figure 8 Execution time and waiting time in Disco

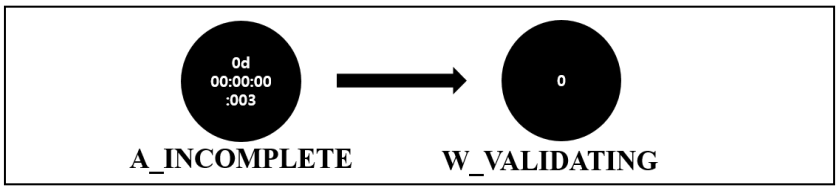

Figure 9 The sojourn time in ProM

4.2 Analysis of Loan Failure Factor

Question of this chapter is what is the influence on the frequency of incompleteness to the final outcome? The hypothesis here is that if applications are confronted with more requests for completion, they are more likely to not accept the final offer.

In order to find the solution of the above question, we assume that 'incompleteness' in the question means that the process is incomplete or 'A_Incomplete' activity. And we assume that 'frequency' in the question means that the frequency of case or the frequency of the activity.

First, we assume that 'the frequency of incompleteness' means the frequency at which the loan failed. We think that the case without 'A_Pending' was the case where the loan was not concluded. Approximately 55% of clients' loan are approved. We use Python to extract 100 subsets of successful loans and 100 subsets of unsuccessful loan. Next, we import the subsets to the ProM and use the 'Add identities to log' plug in. By the plug in, we find that the frequency of 'W_Validating application' and 'A_VALIDATING' is high when the loan is concluded. So the more the bank asks the customer, the better the loan is concluded. From the previous analysis, we find that the number of validating does not affect the success of the loan. To ensure that the results of the subset are the same in all data, we analyze rate of the success of the loan by separating the entire data based on the number of evaluations. The table below shows the analysis results. We conclude that the higher the validating, the

higher the likelihood of a loan being made and we conclude that once validation happens, the likelihood of approval of a loan to be increased.

Table 3 Result across the whole data

	Multiple Validation(11669)	Single Validation(10201)	No Validation(9639)
Loan Success (A_Pending)	10071 (86.31%)	7157 (70.16%)	0
Loan Failed (No A_Pending)	1598 (13.7%)	3044 (29.84%)	9639

Second, we assume that 'incomplete' in the question means 'A_Incomplete'. Under this assumption, we try to find that the frequency of 'A_Incomplete' affects loan fulfillment. In the whole data, 50% of the loans were succeed, but the ratio of loan success was higher when the process was passed 'A_Validating' or 'A_Incomplete'

Table 4 Result about the influence of A_Incomplete

	Total Number	Number of Loan Success
One A_Incomplete	9317	7666 (82.29%)
Two or more A_Incomplete	5686	3044 (87.60%)

And we try to find out what causes the process to end up in 'A_Incomplete'. As a result of filtering using disco, there are a total of 29 cases ending with 'A-Incomplete'. Since this number is very few, we did not undergo further analysis.

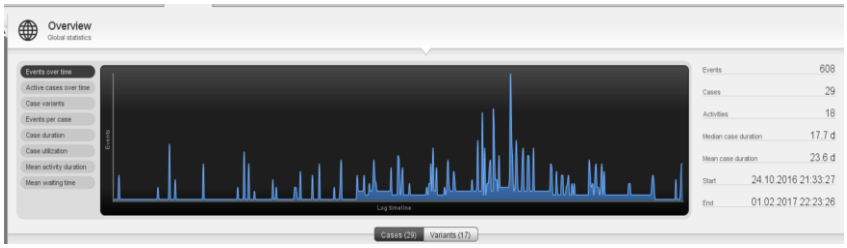

Figure 10 Result of filtering using Disco

4.3 Analysis of performance

4.3.1 How many customers ask for more than one offer?

We assume that the number of 'O-CREATE OFFER' is the number of offers offered by the bank to the customer

Table 5 The number of customers who receive one offer or more offers

	Number(people)
Customer who receive an offer	22950
Customer who receive multiple offers	8559

As a result of the data analysis, we find that the number of customers who received one offer was more.

4.3.2 How does the conversion compare between applications for whom a single offer is made and applicants for whom multiple offers are made?

In order to find the solution to the question, we compare the loan success rates of the customers who were offered one offer and those who were offered several offers. And we assume that there are two meanings of 'conversion' in the question. First, 'conversion' means process switching. Second, 'conversion' means 'A_Pending'.

Under the first assumption, we try to compare processes of one offer and multiple offer. We separate the whole data into one offering cases and several offering cases by using Python. We figure out the processes of the two cases by using 'Inductive miner' in Prom. As a result of analysis, we find that 'Application activity' is the same regardless of the number of offers which customers are offered. This means that 'Application activity' is a process that must be performed in all cases. Also, we find that the process of 'O_Create offer'→'O_Created'→'O_Sent'→'O_Refused' or 'O_Returned' is repeated in the process of providing several offers to the customer.

Under the second assumption, we find that the success of the does not depend on loan the number of offers.

Table 6 The number of customers who receive one offer or more offers and ratio

	Total Number of case	Number of case with A_Pending(ratio)
Total Case	31509	17228(54.67%)
One Offer	22950	12178(53.06%)
Multiple Offer	8559	5050(59.00%)

And we identify the process by which the loan is concluded. In both cases of single offer and multiple offer, the 'ACCEPTED value' is 'False' when offer is first created. Next, the bank changes the 'ACCEPTED value' of the successful offer to 'True' as the valuation progresses. An offer with True of ACCEPTED means the customer can select the offer. If the customer chooses the offer, the SELECTED value becomes 'true' and the loan is succeeded. It can be seen from the analysis of Python that the above procedure is the same regardless of the number of offers.

Offer	Accepted
Offer1	False
Offer1	False
Offer3	False
Offer4	False

↓

Offer	Accepted
Offer1	True
Offer2	**True**
Offer3	False
Offer4	False

↓

Offer	Accepted	Selected
Offer1	False	False
Offer2	**True**	**True**
Offer3	False	False
Offer4	False	False

Figure 11 Procedure of loan Success in multiple offer case

Offer	Accepted
Offer1	False

↓

Offer	Accepted
Offer1	True

↓

Offer	Accepted	Selected
Offer1	True	True

Figure 12 Procedure of loan success in single offer case

But there is exception. The exception is that 3012 cases are the value of 'Accepted' is False, but bank can ignore it and give the customer the opportunity to choose an offer, and if the value of 'Selected' is True, then the loan is succeeded. In such exception, there is always 'A_Incomplete'. We assume that this can be seen as a case in which the bank once completes the loan and then validates later.

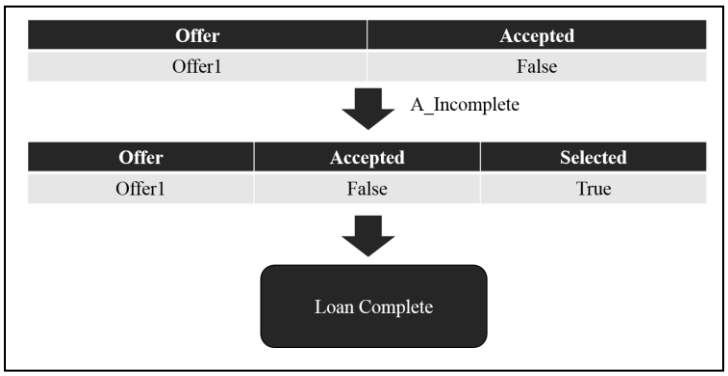

Figure 13 Procedure of expectation of single offer case

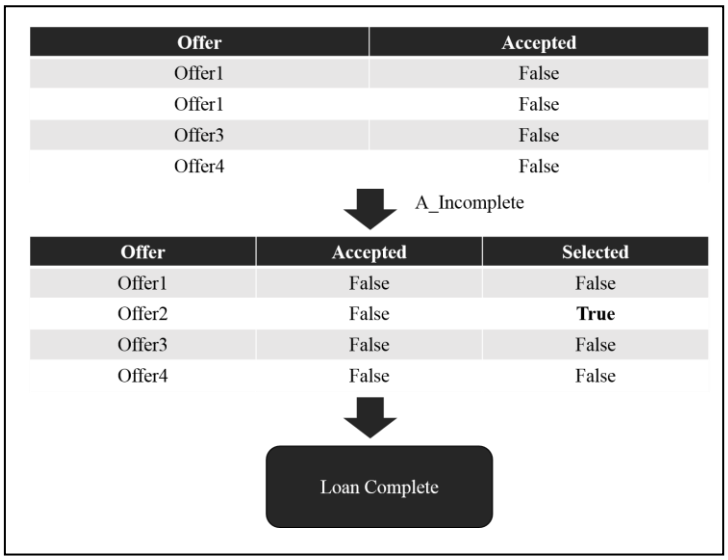

Figure 14 Procedure of exception of multiple offer case

5 Conclusion

With no domain knowledge, it was difficult to understand the data. And also provided three main questions from bank in BPIC 2017 website was hard to grasp the meaning. So, we had to undergo many trials and errors.

In order to solve first problem regarding time, we have identified thoroughly at where customer took time or where bank took time. First we discovered process model of loan process and identified which activity was performed by customer or bank. We concluded that time from A_Submit to O_Sent and from W_Validating to A_Validating are customer waiting time. And we concluded that time from O_Sent to A_Complete and from A_Incomplete to W_Validating are bank waiting time. Bank waiting time was depending on customer so there is rare room for improvement. To increase customer satisfaction, bank should improve customer waiting process.

In order to solve second and third problem, we had to understand thoroughly the meaning of problem but it is not easy. So we analyzed process all potential meaning of problem. We split the number and counted it by on piece. Using Python and Excel, we tried to find what factor impact on the incompleteness. As result of us, as the number of validation activity goes up, success of loan process goes up.

To solve third question, we calculated basic statistics for checking correlation between the number of offer and A_pending which means success of loan. But there was no correlation so we tried to another factors which impact on success of loan.

References

1. Minseok Song, Will M.P. van der Aalst.: Toward comprehensive support for organizational mining. Decision Support System (2008)
2. A.Rozinat, R.s. Mans, M.Song, W.M.P van der Aalst.: Discovering simulation models. Information Systems (2009)
3. Van der Aalst, W., Adriansyah, A., Alves de Medeiros, A.K., Arcieri, F., Baier, T. et al: Process Mining Manifesto. In: Business Process Management Workshops 2011, Lecture Notes in Business Information Processing, vol. 99, Springer-Verlag (2011)