BPIC 2017: Business process mining – A Loan process application

Dongyeon Jeong, Jungeun Lim, Youngmok Bae

Department of Industrial and Management Engineering, POSTECH(Pohang University of Science and Technology),
Pohang, Republic of Korea
{colobrother,je5719,ymbae}@postech.ac.kr

Abstract. The collection of a huge amount of loan process data in the financial industry has never been easier with the advent of sophisticated data collection technologies. This study uses sample data of a loan process generated by a major financial institute. The BPI Challenge 2017 provides the dataset of loan process from a financial institute, collected in the period of 2016.01-2017.02. This study first reviews the attributes of the dataset and the loan process, and then we analyze the data based on three questions that the company is interested in. In order to provide accurate results, many tools such as DISCO, MINITAB, MATLAB, R, etc. were utilized. This study is expected to help to enhance the loan process availability and provide a basis for a stable loan process that can support increasing of profit of financial institute in the future.

Key words: Process mining, Loan process, Data mining, Statistical analysis, Classification

1 Introduction

Understanding the loan process precisely is considered as necessary process because the lending money plays a key role in a financial institute. Traditionally, to improve the process of loan, finance managers had to check account books and manually check its process. However, the revolution of ICT has led to the collection and loading of a huge amount of data, and the changes provide opportunities to turn into process mining and statistical analysis.

The process log given from BPIC 2017 is a log extracted from the financial institutes loan process. The log is the loan process that is also an improved process through BPIC 2012. For this study, three questions were asked to identify process more specifically. The first is to compare the time the company waits for the customer and the time the customer waits for the company. The second problem is how incompleteness affects the final outcome. The last issue is about how the process varies with the number of offers. We used a variety of methodologies from various perspectives to answer these questions below. Furthermore, we applied several data mining techniques that can support to understand the

problem in loan process.

In Question 1, we did not compare the activity time and the time taken in the path simply, but analyzed the semantic unit by dividing the process as group of activities. In Question 2, we conducted a statistical analysis on the ratio of A_Pendig and A_Cancelled according to the ratio of A_Incomplete. And we used machine leargning technique to analyze the cause of the difference in each group. In Question 3, we made process map according to the number of offers by Disco, and analyze the difference in activities and paths between two groups. Finally, in addition to the three given problems, we analyze resources and the difference of pending and not pending case.

1.1 Used Tools

We used various tools to examine issues from various perspectives. We used tools on three purposes. First, we wanted to sort through of process by applying process mining techniques. Even though there are many tools to check processes precisely, we decided to utilize Disco and Celonis for this purpose. Statistical analysis is also a major purpose of our study. In order to provide accurate result in statistical analysis, we decided to utilize R, Excel, Minitab and Matlab. Lastly, we had to transforms the data into a format that is needed to be by removing or adding other information, and oracle 11g was utilized for the transforming the data.

Briefly speaking, Disco was the most useful tool to understand and visualize the overall process. Using Disco, we were able to identify process-related indicators rapidly. Celonis is also process mining tool like Disco, but we used Celonis for a slightly different purpose than Disco. Celonis was good to visualize the usage of resources. Also, we were able to analyze daily activities better. R was used to calculate process time and Excel was used to identify data quickly and visualize it. Oracle 11g was used to extract the data with specific condition we want. Minitab was used to do major statistical analysis and Matlab was good to apply machine learning techniques.

1.2 Data description

Two types of data for monitoring the process of loan company were aquired in this study. First is the data that represent the total event in each cases. The dataset contains activity, resource, time stamp, and additional information such as requested amount of money and offered amount of money. The other dataset is the offer-log data that hold the offers history that company made. The operation record dataset contains the exactly same variables with the event data. Besides, since the first data set contains the entire offers event that the second data set had, we decided to only analyze the loan process with the first data set.

In the first data set, there are 26 types of event that can be grouped into three category: Application state changes, Offer state changes, and workflow events. Each category has specific events in the event log. We tried to understand the meaning of each event briefly. Submitted event can be understood that customers has submitted a new application. Concept is the event when a first assessment has been done automatically. Accepted is the event can be understood that there is a possibility to make an offer. Complete can be understood as the offer have been sent to the customer. Validating is the event that the offer and necessary paper works are received and are checked. When the customer needs send in additional paper work, the incomplete will be marked. In particular, this event will be used in question 2. Pending will be marked when the loan is final and customer is payed. Denied is the event that could be marked when the application does not fit the acceptance criteria. Lastly, cancelled is the event that the customer calls to tell he/she does not need the loan any more or customers never send in their documents

Table 1. Basic statistics of numerical attributes

	Requested Amount	First Withrawal Amount	Monthly Cost	Number Of Terms	Offered Amount
Average	16587.57	8394.339	281.4033	83.04198	18513.72
StDev	15387.05	10852.44	192.5777	36.3862	13718.51
Zero value	4124	12786	0	0	0
N/A value	0	0	0	0	0

The data used in the present study come from the loan process of the financial institute. The event log contains all applications filed in 2016, and their subsequent handling up to February 2nd 2017. In total, there are 1,202,267 events pertaining to 31,509 loan applications. According to the dataset that we acquired, 42,995 offers were created. They are hiring 145 users include automatic system, and categorized the loan goal into 14 categories such as car and home improvement. In addition, we conducted basic statistics to the entire cases to understand overall process. Table 1. represents the basic statistics of numerical attributes.

2 Understanding Loan Process

We needed to understand the loan process before answering the questions. In this section, we will define the concept that is necessary for the basic analysis. In order to understand the loan process, we first only looked at the case of one offer because most cases are one offer case, and it is sufficient to understand the loan process in general. Figure 1. represents the loan process. When customers

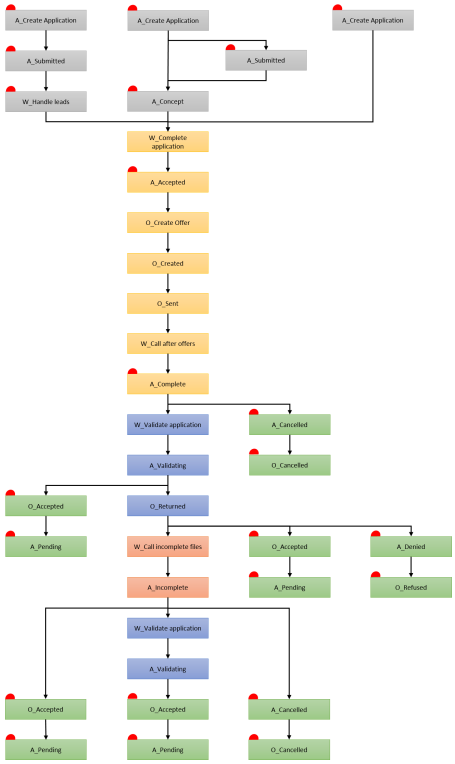

Fig. 1. Whole process

apply for the loan to lend money, the company suggest an offer, and customers submit documents and signature as next step. After that, the company make a decision whether they could lend money or not. In more detail, we divided the loan process into 3 parts: application part, making offer part and validating documents and making a decision part. The processes are as follows.

2.1 Application part

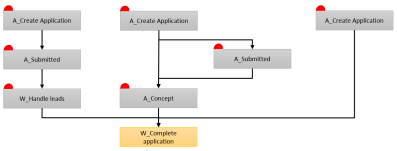

Fig. 2. Application part of the loan process.

Figure 2. represents the application part. In this part, customers apply for the loan. Therefore, this part is defined as the time spent waiting on input from

the applicant for Question 1. This part has 4 types of path: A_Create Application - A_Submitted - A_Concept, A_Create Application - A_Concept, A_Create Application - A_Submitted -W_Handle leads and A_Create Application. Of these, A_Create Application - A_Submitted - A_Concept and A_Create Application - A_Concept are the mostly.

2.2 Making offer part

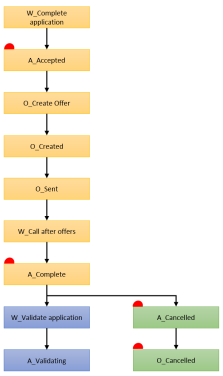

Fig. 3. Making offer part of the loan process.

Figure 3. represents the making offer part. In this part, the company makes an offer and sends the offer to a customer. Thus, the process from W_Complete application to A_Complete path is defined as the time spent in the company's systems waiting for processing by a user for Question 1. After that, there are two path: W_Validate application - A_Validating and A_Cancelled - O_Cancelled. The first case starts with W_Validate application is the time when a customer submits documents and signature. The second case starts with A_Cancelled is the activity where a customer cancels the loan. Thus, two paths are defined as the time spent waiting on input from the applicant.

2.3 Validating documents and Making a decision part

Figure 4. represents the validating documents and making a decision part. W_Validate application - A_Validating is to validate documents and signature of a customer. This part is defined as the time spent in the company's systems waiting for processing by a user. If documents and signature are insufficient, then the company may ask the more documents to a customer. So, the process from W_Validate application to A_Incomplete path is defined as the time spent in the company's systems waiting for processing by a user. After the company ask documents, a customer resubmit documents. So, A_Incomplete - W_Validate

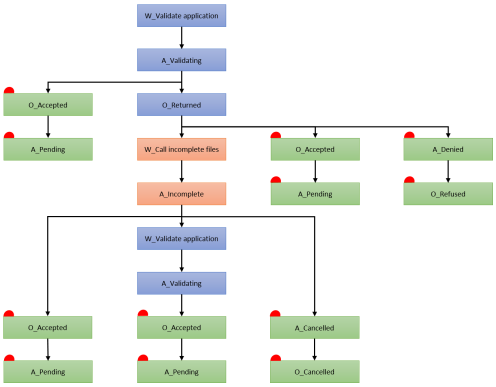

Fig. 4. Validating documents and making a decision part of the loan process.

application path is defined as the time spent waiting on input from the applicant. Finally, we can find the green activities as validating process in Figure 4. Except A_Cancelled, the paths to do O_Accepted and A_Denied is defined as the time spent in the company's systems waiting for processing by a user because these activities are done by the company. On the contrary, Since A_Cancelled is done by a customer, the paths to do A_Cancelled is defined as the time spent waiting on input from the applicant.

3 Question 1: What are the throughput times per part of the process, in particular the difference between the time spent in the company's systems waiting for processing by a user and the time spent waiting on input from the applicant as this is currently unclear

In this section, we analyzed the Question 1 to compare the time spent in the company's systems waiting for processing by a user and the time spent waiting on input from the applicant. For computing each time, we defined where paths belong in previous section. We calculated each time based on the definition. For accuracy, we subtracted a start time of first start activity from a complete time of final activity of each path in each path. We analyzed the time spent waiting on input from the applicant first and then the time spent in the company's system waiting for processing by a user.

3.1 The time spent waiting on input from the applicant

In this section, we analyzed the time spent waiting on input from the application. Table 2. is the time per path. A_Create application - W_Complete application and A_Create application - A_Accepted are the application part. The medians of these are 4.6 and 4.3 hours, respectively. The averages on these are 16.6 and

Table 2. Time spent waiting on input from the applicant(d is day and h is hour).

Path	Median	Average
A_Create application - W_Complete application	4.6h	16.6h
A_Create application - A_Accepted	4.3h	19.5h
A_Complete - W_Validate applicaition	7.1d	6d
O_Sent(mail and online) - W_Validate application	6.6d	7.6d
W_Call incomplete files - W_Validate application	23.4h	59.15h
W_Call after offers - A_Cancelled	30.6d	26.9d
A_Complete - O_Create Offer	3.8d	6.3d
A_Incomplete - O_Create Offer	4.9h	43.3h

19.5 hours, respectively. This can be understood that there are some cases that are extremely long. That is, although customers applied for a loan quickly, some cases spent long time for application. That is why we could see the difference between the median and average.

A_Complete - W_Validate application and O_Sent(mail and online) - W_Validate application are the time that a customer submits documents and signature after the company suggested offer and asked that. The medians of these are 7.1 and 6.6 days, respectively. The averages on these are 6 and 6.3 days. That is, it usually takes a week.

W_Call incomplete - W_Validate application is the time that a customer resubmits additional documents after the company validated previous documents.. The median and average on this are 23.4 and 59.15 hours, respectively. It usually takes less than 3 days.

W_Call incomplete files - A_Cancelled is the time a customer makes a decision not to borrow or the company cancels the offer because the customer does not respond the offer. The averages and median of this are 30.6 and 26.9 days. It usually takes a very long time.

A_Complete - O_Create Offer and A_Incomplete - O_Create Offer are the time that a customer asks another offer or the company suggests it. The medians of these are 3.8 days and 4.9 hours, respectively. The averages of these are 6.3 days and 43.3h. As shown, there is difference between two paths. A_Complete O_Create Offer is that the company suggest another offer without a customers documents and signature after first offer. Otherwise, A_Incomplete O_Create Offer is that the company suggest another offer after the company validated the documents and signature. Thus, due to these reason there are differences of average and median between two paths.

3.2 The time spent in the company's systems waiting for processing by a user

Table 3. Time spent in the company's systems waiting for processing by a user(d is day and h is hour). The case containing * path without A_Incomplete.

Path	Median	Mean
W_Complete application - W_Call after offers	0.68h	23.26h
W_Validate application - A_Pending or O_Refused*	1.95d	2.54d

In this section, we analyzed the time spent in the company's systems waiting for processing by a user. Table 3. is the time per path. W_Complete application - W_Call after offers is the time that the company makes offers and suggests the offers to a customer. The average and median of these are 0.68 hours and 23.26 hours, respectively. It usually takes less than one hour. The reason that the mean is longer than the median seems to be that there are weekends, holidays and the time that the work is postpone to the next days.

W_Validate application - A_Pending or O_Refused is the time that the company validate the documents and makes decision whether to lend money or not. The path does not contain A_Incomplete between the activities because if there was A_Incomplete, there can be cases such that a customer should submit another documents and the company should wait for it. The median and average on this is 1.95 days and 2.54 days. It usually takes less than 3 days.

We computed the times for Question 1. After that, we analyzed each time by dividing path. The followings are conclusion. First, the company has a lot of time waiting for the customer to wait. Second, it takes a long time for a customer to verify an offer and give the signature and documents to the company. Third, it takes a long time for a customer to verify an offer and ask another offer. Fourth, it takes a long time for a customer to decline a final offer or not to response. Thus, the company needs an activity that allow customers to make a final decision quickly. Finally, customer does not wait long compared to company

4 Question 2: What is the influence on the frequency of incompleteness to the final outcome. The hypothesis here is that if applicants are confronted with more requests for completion, they are more likely to not accept the final offer

4.1 Frequency of incompleteness Meaning of frequency of incompleteness

In order to see what influence on final outcome, we needed to define the meaning of frequency of Incomplete to the final outcome first. The meaning of frequency of Incomplete to the final outcome can be understood as the number of activity A_Incomplete. In this process, A_Incomplete can be occurred if documents are not correct or some documents are still missing. Then, the status is set to incomplete, which means the customers needs to send in back up documents. In particular, in this process, the influence of the activity A_Incomplete should be understood since more than half cases have the A_Incomplete. To check the influence on the frequency of A_Incomplete to the final outcome, we checked the number of the activity A_Incomplete. In the process we analyzed, each cases had zero to eight A_Incomplete. Furthermore, some of the cases which contains zero A_Incomplete can be deleted since those cases did not even have validation step in the process. This processes can be understood as the fake or unprepared cases which make noises when we check the influence on the frequency of Incomplete to the final outcome. To manage these issues, we only use the process that contains at least one validation activity in a process.

In the data we analyzed, 69% of cases had at least one validation activity in a process since the cases that never had any validation activity cannot be concerned as a normal process. Based on the cases we selected, the number of the activity A_Incomplete were checked, and the result were grouped into four categories. Brief explanation of the categories is summarized in Figure 5. and fuller explanations follow. As we mentioned above we grouped the cases based on the number of the activity A_Incomplete. However, in particular, the cases which contains 3 or above times of A_Incomplete are inadequate to construct independent categories since there are small amount of cases which exceed 3 or above times of A_Incomplete. That is, we put all cases which contains 3 or above times of A_Incomplete into a group.

In order to check the hypothesis that the company has believed, we decided to see percentage change in activity A_Pending and A_Cancelled. We believe that these two activity can be the evidence converging to support or reject the hypothesis. We checked the number of cases which contains A_Pending and A_Cancelled by using DISCO. First of all, in the perspective of ratio of A_Pending, once applicants get the A_Incomplete, the ratio of A_Pending is getting higher. For example, 67% of applicants who never had A_Incomplete got A_Pending. From this point, the ratio of A_Pending is getting higher up to 87%. The detail information and the trend of ratio will be explained in Table 4. Second, in other

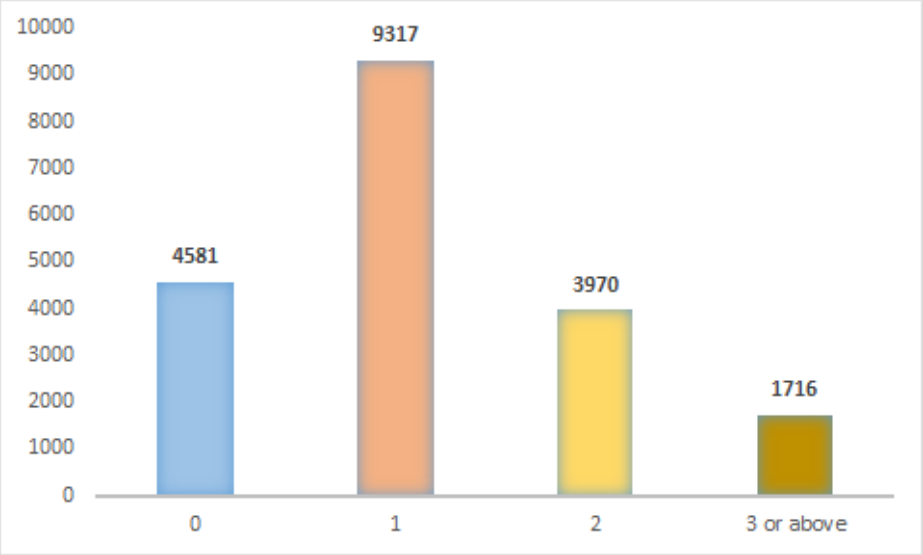

Fig. 5. Number of cases in each group

hands, the ratio of A_Cancelled provide interesting features. Unlike the ratio of A_Pending, the ratio of A_Cancelled is getting lower. In particular, once the A_Incomplete was happened, the ratio of A_Cancelled is reduced slightly. However, if the applicants do not get any A_Incomplete from company, the applicants rarely conduct the activity A_Cancelled in the process. The detail information and the trend of ratio will be also explained in Table 4. According to the result of the analysis, hypothesis that the company has believed must be rejected.

Table 4. Ratio of A_Pending and A_Cancelled

Number of A_Incomplete	Number of cases	Ratio of A_Pending	Ratio of A_Cancelled
0	6867	67%	1%
1	9317	82%	6%
2	3970	87%	5%
3 or above	1716	87%	5%

4.2 Frequency of incompleteness Cause analysis

We tried to analyze not only the reason why the frequency of Incomplete can give huge impact on final outcome but also the reason why each process has the activity A_Incomplete. In order to figure out what is causing the activity A_Incomplete, we analyzed the difference in loan goal at first. Since the number

of cases in each categories is different each other, definite number of loan goal cannot provide any insights. That is, we compared the loan goal by using each ratio of loan goal. According to the result of analysis, there did not show any strong difference of loan goal. However, one thing that we discovered is that the cases that have Other, see explanation as their loan goal never get A_Incomplete during the process. In the same context, the cases that have Business as loan goal never get A_Incomplete more than three times. Although these pattern also remarkable, domain knowledge is also needed to understand the difference in loan goal deeply.

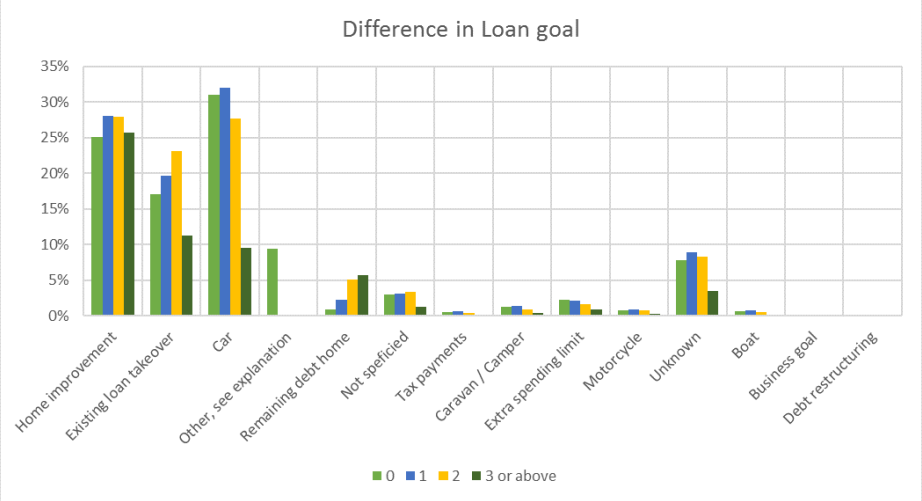

Fig. 6. Difference in loan goal for each category

4.3 Statistical methodologies

As a next step, we compared requested amount and offered amount among the categories. Since we could not compare every single requested amount and offered amount in each case, we computed representative values by applying the mean value to improve the perceived ease of use. In particular, even though some cases contain multiple offered amount in a case, we use the average value for those issues as well. In particular, we applied ANOVA that is a collection of statistical model. This methodology can analyze the difference among group means. As a result, we found that the mean values of requested amount of each category are increased as number of A_Incomplete increased. The offered amount also increased when the number of A_Incomplete increased. This might be understood that resources put more effort to check the status and return the application in high probabilities when the applicants wanted to have bigger money. The result of ANOVA test for requested amount is as follows.

Table 5. Result for ANOVA test

Level	N	Mean	StDev
0	8631	15712	14413
1	12732	16378	15882
2	6071	18023	16059
3 or above	3032	20057	17299

Source	DF	SS	MS	F	P
Factor	3	5.35E+10	1.78E+10	72.69	0
Error	30462	7.48E+12	2.34E+08		
Total	30465	7.53E+12			

Without any domain knowledge, we want to understand classification rule in machine learning perspective. We use MATLAB to figure out classification rule for entire cases. Since MATLAB provides various types of machine learning methodologies (i.e., Support Vector Machine, Decision Tree, Random Forest, etc.), we decided to apply all of the methodologies what the MATALB has provided. Offered Amount, Requested Amount, Resource, First Withdrawal Amount, Number of terms, and Monthly cost were chosen as input variable, and the four types of categories were chosen as response variable in each classification model. After check multi-collinearity, we compared all the result that we got. As a result, all methodologies provide less than 60% of accuracy. The result of decision tree is followed as an example.

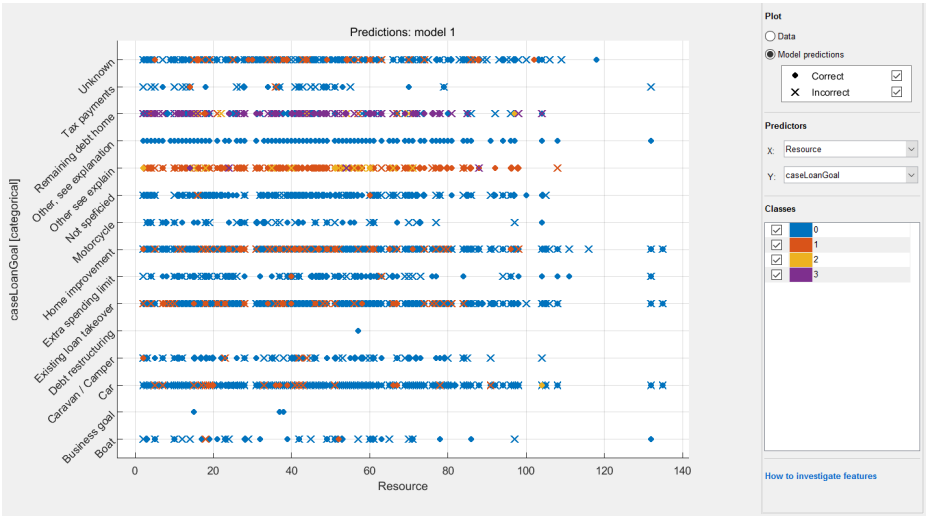

Fig. 7. Result of Machine Learning(example)

To sum up, we rejected the hypothesis that the higher frequency of A_Incomplete tend to cancel final outcome a lot. By checking trend of the A_Pending and A_cancelled above, we showed that the result what we had is significant. However, we could not figure out the reason why people had A_Incomplete during the process. This can be interpreted in three different ways. Even though the company had specific rules when they return the applications to applicants, resources who conducted returning process did not follow the rules that they have. Otherwise, resources who may be concerned about the process did not record appropriate information. Last but not least, the returning application process can be determined by other factors that we did not get from the company. That is, to understand the reason why applicants got A_Incomplete, additional information must be provided.

5 Question 3: How many customers ask for more than one offer (where it matters if these offers are asked for in a single conversation or in multiple conversations)? How does the conversion compare between applicants for whom a one offer is made and applicants for whom multiple offers are made?

Offer created when application is completed. The company sends offers to customers depending on the application. However, offer does not end with the company sending it to the customer only once. Multiple offers can be possible depending on the customers request.

In Question 3, we want to find how many cases of multiple offers take up and what the difference is between groups which have different number of offers if a loan is completed (the process including A_Pending).

5.1 Number of cases by number of offers

In all 31509 cases in the given data, customers receive at least one offer. It means each case contains O_Create offer at least once. Regardless of what the end activity of the process, if you categorize a case by the number of offers (the number of O_Create offer) only, the case with only one offer is 22950, which accounts for 72.8% of the total cases. As the number of offers increases, the ratio decreases. The number of cases received 2 offers is 6578, which is 20.9% of the total, and for the 3 offers, 1348 cases which account for 4.3% of the total. If there are more offers than 3, it will be less than 1.5% and it does not take up much of the total case.

5.2 Basic Statistics of Process of pending cases

According to the number of offers, when the offer is once, the pending case accounts for 53.1%. it occurs 12178. The case takes from at least 7 minutes to 152

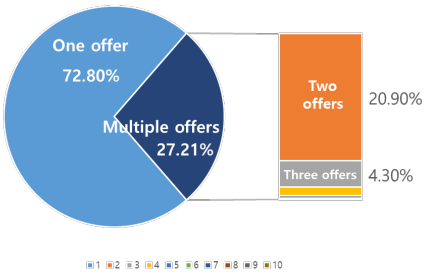

Fig. 8. Case ratio by offers

Table 6. Number of cases by offers

Number of offers	Frequecy	Percentage%
1	22950	72.836
2	6578	20.877
3	1348	4.278
4	443	1.406
5	126	0.400
6	30	0.095
7	16	0.051
8	13	0.041
9	3	0.010
10	2	0.006
Total	31509	100

days, with an average of 16.1 days and a median of 13.7 days. Activites which occurs in one case are from at least 14 to 40 with an average 18.1. However, 95% of cases just contains less than 25 activities.

Of the 8559 total multiple offers, 5050 pending cases account for 59%. The cases take from at least 16 minutes to 145 days, with an average of 23 days and a median 19.6 days. Activities which occurs in one case are from at least 17 to 61 with an average 25.6 days. However, 95% of cases contains less than 36 activities.

5.3 Difference between one offer and multiple offers

The overall process map of the original offer and the multiple offers is shown in the following Figures. The Figure 9. and 10. are made through Disco with the 100%. of activity and path ratio

Table 7. Basic Statistics of Process of pending cases

	One offer	Multiple offer
Case	12178 (53.1%)	5050 (59%)
Activity	Min: 14 Max: 40 Mean: 18.1	Min: 61 Max: 61 Mean: 25.6
Case Duration	Mean: 16.1 days Median: 13.7 days	Mean: 23 days Median: 19.6 days

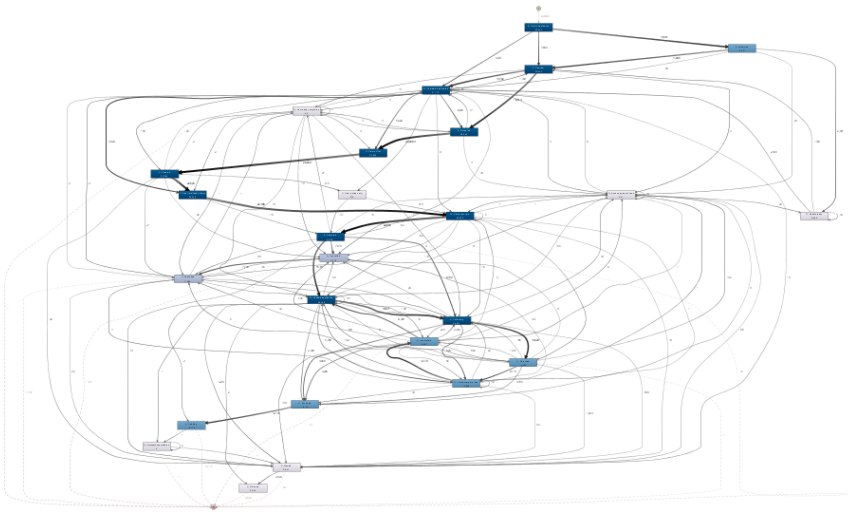

Fig. 9. process map of one offer

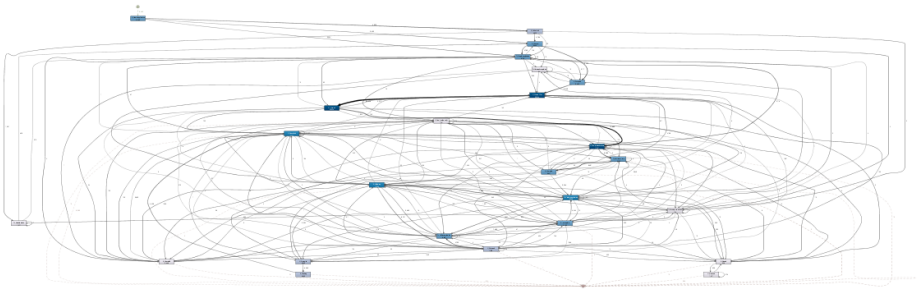

Fig. 10. process map of multiple offers

The reason for this difference in the overall process is due to the difference in the number of activities and the differece in path between the two processes.

Activity There is a difference in the number of activities on average of seven and eight between the one offer and multiple offers. This is because of the activities that are repeated in multiple offers. Repeated activities include O_Cancelled, O_Create Offer, O_Created, O_Sent (mail and online), O_Sent (online only) and O_Returned.

Process In the one offer process, activity occurs only after A_Accepted and W_Complete application. Among them, there are 12141 cases (99.70%) in which O_Create offer is generated after A_Accepted, and 35 cases (0.29%) in which O_Create offer is generated after W_Complete application. In most cases, O_Create offer is generated after A_Accepted. However, this rule is broken in multiple offers process. In the multiple offers process, 7 additional points are found in addition to the two points in the one offer process. In each case, there are 4906 cases after A_Accepted (42.86%), 11 cases after W_Complete application (0.10%), 2331 cases after A_Complete (20.37%), 1893 cases after O_Created (16.54%), 1412 cases after A_Incomplete (12.34%), 398 cases after O_Sent (mail and online) (3.20%), 43 cases after A_Validating (0.38%), 13 cases after W_Call incomplete files (0.11%).

Another difference in the process is due to the paths that only appear in multiple offers. The path that occurs only in the multiple-offers is A_Pending - O_Cancelled, O_Sent (mail and online) - W_Validate application, A_Complete O_Create Offer, O_Created - O_Create Offer, O_Cancelled - O_Cancelled, O_Sent (mail and online) - O_Sent (mail and online) and A_Incomplete - O_Create offer. It is A_Complete - O_Create Offer and A_Incomplete - O_Create offer that makes difference between one offer and multiple offers. A_Complete - O_Create Offer takes 6.3 days on average and a median is 3.8 days. A_Incomplete - O_Create offer takes 43.3 hours on average and a median is 4.9 hours. Considering that in one offer process, after A_Complete, W_Validate application occurs reviewing the documents form customer, the case duration becomes longer adding the process that create new offer between A_complete and W_Validate application in multiple offers. This is same in A_Incomplete cases.

5.4 The reason for difference

We looked at Credit Score, First Withdrawal Amount, Requested amount, Loan goal and so on to see which makes the difference in the number of offers. In other items, it was hard to find significant difference, but the difference between the Requested amount and Offered amount was found to be different. We assume that if Offered amount is different, it is different case. Because in one case, Offered amount changes although Requested amount was constant.

The results were as follows. In the one offer, the cases with requested amount = offered amount was 77.3%, the cases with requested amount < offered amount was 18.7% and the cases with requested amount > offered amount was 4.0%. In the multiple offers, the cases with requested amount = offered amount was 55.9%, the cases with requested amount < offered amount was 28.9% and the cases with requested amount > offered amount was 15.2%. We can see that the case with requested amount = offered amount decreases and other cases increase. Since we assumed that each time the offered amount is changed the new case is created, we can see from this result that if there is difference between requested amount and offered amount, the cases to regenerate new offers increase.

To make the process more efficient, its a good idea to manage the case which have different requested amount and offered amount so that they do not interfere with other processes.

6 Additional work - Pending vs. Not Pending

In the perspective of applicants, the main purpose of this process is to lend money from the company. On the other hands, in the perspective of financial institutes, the main purpose of this process is to choose right person who will repay their money. That is, getting the activity A_pending is one of the most important part of the process. In order to find out the major differences between the applications that got A_Pending and the applications that did not get A_Pending, we applied several statistical methodologies.
We wanted to know the classification rule for the process. As we mentioned before, we grouped the data into two group that represent complete process and incomplete process. Complete process can be understood as the cases that contains A_Pending. There are some cases that is end with A_Cancelled even though the cases had A_Pending in the process. We defined this type of cases as complete process as well since the A_Cancelled can be understood as a cancellation for other offer that they already made. Incomplete process can be understood

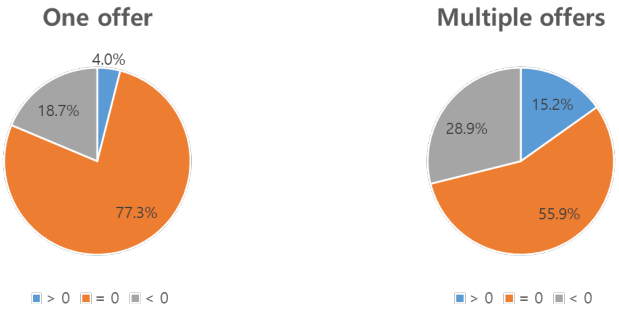

Fig. 11. Difference of Requested amount and Offered amount

as the cases that end with the activity O_Cancelled, O_Refused and contain A_cancelled in the process. The cases that is not fit into this two classification standard are not concerned in this analysis. As a result of classification, the complete process group contains 17228 cases and incomplete process contains 10269 cases respectively.

In order to check the average event per cases, we applied ANOVA test to two groups. In the cases of complete process, each cases contains 20.32 per cases on average. On the other hand, in the cases of incomplete process, each cases contains 13.69 event per cases on average. The result of ANOVA test was significant, which means the complete process contains more activities than incomplete process. MINI TAP showed the appropriate P-value (¡0.01) and R2 (0.3).

Table 8. Basic information of complete/incomplete process

	Case per day	Case per user	Throughput time	Users per case
Complete Process	43(888)	23(119)	17 days	5
Incomplete Process	26(359)	20(72)	30 days	3

Even though we noticed that the complete process contains more events, we still do not know the reason why people get A_Pending from the company. In this respect, we checked average and standard deviation of some variables that contains numerical data such as credit score, requested amount and offered amount. Those data are really important because numerical data is easy to do objective assessment of status. In general, we could not get any remarkable insight from this statistical information. However, we found two facts that we need to concern. First, the average value of offered amount is bigger than average value of requested amount. In normal process, offered amount could not be bigger than offered amount, but it was happened in this dataset. Multiple offers or missing data that was made by users might be the reason why offered amount is bigger than requested amount. Those multiple offers and missing data can cause erroneous result in average value. Second, all applicants who did not get A_Pending from the company have zero credit score. This also might be missing critical piece of information. The detailed result of each group is as follows.

In the perspective of process, we found an interesting fact. The initiating processes in the groups are very similar each other. This can be understood that the company has specific manual when they got an application. By applying fuzzy mining, most of processes are start with A_Create Application and end with A_complete as an initiating process. That is, most of applicants can have a phone call from the company whether they could borrow money from them or not. However, when they got validation procedure, it became complete process with

Table 9. Basic statistics of complete/incomplete process

Process		Requested Amount	Credit Score	First Withrawal Amount	Monthly Cost	Number Of Terms	Offered Amount
Complete	Mean	16532.3	594.2	7981.6	282.8	85.4	19029.4
	StDev	15693.1	434.9	10137.1	192.7	36.1	13353.0
Incomplete	Mean	15915.9	0.0	8413.7	278.3	77.9	17161.5
	StDev	14860.4	0.0	11621.0	198.4	36.9	133773.6

A_Pending. On the other hands, most of incomplete processes have A_Cancelled rather than validation procedure without any validation procedure in general. Figure 12. Represents the common process for each groups. Although we found process-based and statistical-based differences from the data, the reason why applicant fail to borrow money is not specified yet.

Fig. 12. Common process in complete/incomplete process

As a last analysis, we applied machine-learning methodology again to find out classification rules that we might miss. Loan goal, requested Amount, first

withdrawal amount, monthly cost, number of terms and offered amount were used as predictor variables. Decision tree, random forest, support vector machine and logistic regression were used to find out classification rules. Even though we conducted Principal Component Analysis (PCA), the result was not good to get any insight from it. The range of accuracy of methodologies are 55.5% to 61.3%. We could understand that there might be other predictor variables such as external environment. Otherwise, the loan company did not have specific criteria when the company decide whether it would cancel the application or not.

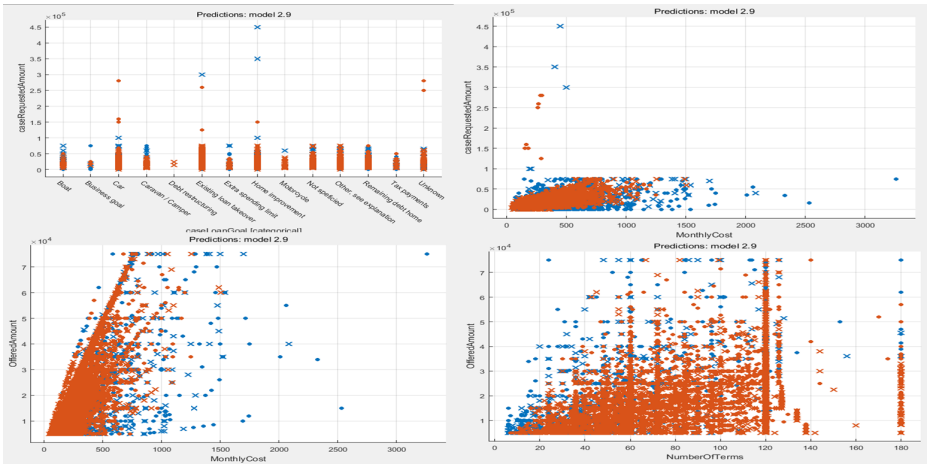

Fig. 13. Result of machine learning

7 Additional work - Clustering with numerical attributes

We did K-menas clustering with Offered amount, Requested amount, Number of terms, First withdrawal amount and Monthly cost. 3 were chose as K in this clustering. The medoids of each group is like as follows.

We can find that the value of each fields increases from C1 to C3. In order to see the difference in each group, we analyzed the ratio of A_Pending cases and O_Cancelled cases, average activity number, case duration, and resources. The ratio of A_Pending cases and O_Cancelled cases was similar to about 50%. The average number of activities is from 17 to 18, and the average case duration is from 21 to 24 days. There was not that big difference. There is difference only in resources. When comparing to 10 resources in each group, C2 and C3 have same resources and C1 has different resources. Each resource item is shown in the table below, and it can be seen that the resources that handle the high cost are distinguished.

Table 10. The medoids of three groups

	C1	C2	C3
Requested amount	8741.36	2381.96	335.36
Credit score	335.36	319.26	272.10
First withdrawal amount	5352.71	11192.67	23202.44
Monthly cost	206.20	362.36	621.20
Number of terms	73.10	97.95	108.81
Offered amount	10846.85	26964.17	51722.67

Table 11. Top 10 resouces of each group

C1	C2	C3
User_10	User_10	User_10
User_100	User_123	User_123
User_121	User_29	User_29
User_123	User_3	User_3
User_27	User_30	User_30
User_28	User_49	User_49
User_29	User_5	User_5
User_3	User_68	User_68
User_42	User_75	User_75
User_49	User_99	User_99

8 Additional work - Resource Analysis

In this section, we analyzed the resources to look at the features. First, we did clustering to check whether there are the groups of resources. Since the log data does not represent the groups, we needs to do that. After that, we compared the performance of resources. We looked at activities related Workflow events because other activities do not have execution time.

8.1 Resource Clustering

For clustering, we used originator task as data set. Originator task data is a numerical matrix that counts activities for each resource. It can be obtained using ProM. Next, we preprocessed the data set to be normalized for better performance before clustering. After we normalized the data set for preprocessing, we excluded User_1. Because User_1 seems to be system. We did k-means clustering by using MATLAB software for $k = 4$. Table 12. is the result of clustering.

Group 1 has 93 resources and do 12 main activities. The main jobs of Group 1 seem to receive applications and suggest offers. Group 2 has 21 resources and do 7 main activities. The main jobs of Group 2 seem to validate documents and

Table 12. The result of k-means clustering by using Originator task table except User_1.

Group Number	Resource									Activity
1	2	3	4	5	6	7	8	9	10	A_Create Application
	11	12	13	14	15	16	17	18	19	A_Concept
	20	21	22	23	24	25	26	28	31	A_Accepted
	32	33	34	35	36	37	38	39	40	A_Complete
	41	42	43	44	45	46	47	48	49	A_Cancelled
	50	51	52	53	54	55	56	57	58	O_Create Offer
	59	60	61	62	63	65	66	67	69	O_Created
	70	71	72	73	74	76	77	78	79	O_Cancelled
	80	81	82	84	85	86	88	89	91	O_Sent (mail and online)
	92	94	96	97	98	103	104	105	108	W_Call after offers
	110	132	135							W_Complete application
										W_Handle leads
2	29	30	68	75	83	87	90	93	95	W_Validate application
	99	100	101	102	106	107	109	115	124	W_Call incomplete files
	128	129	136							A_Incomplete
										A_Pending
										A_Denied
										O_Accepted
										O_Refused
										O_Cancelled
3	27	112	113	114	116	117	118	119	120	W_Validate application
	121	122	123	125	126	127	130	131	133	A_Validating
	134	137	139	140						O_Returned
4	64	111	138	141	142	143	144	145		W_Assess potential fraud

make a decision whether to lend money or not. Group 3 has 22 resources and do 3 main activities. The main jobs of Group 3 seem to validate documents and return offers. Group 2 and 3 do similar work. Two groups seem to be one team. Thus, we can guess that there may be hierarchy between two groups. Group 4 has 8 resources and do one main activity. The main job of group 4 seem to assess potential fraud. Although there are some resources to do that in group 4, the number of entire activities of some resources is very low. In next section, we will analyze W_Assess potential fraud and other workflow events.

8.2 Execution time of resources

In this section, we will analyze workflow events. Before analyzing that, although they were not properly logged such that there are execution time is zero because start time of some events is same to complete time, we assume that workflow

Table 13. The basic statistic for workflow events(d is day, h is hour, m is minute and s is second). By using Disco software, this table can be obtained.

Activity	Frequency	Median	Mean	Duration range
W_Assess potential fraud	355	15h 33m	3d 1h	88d 4h
W_Validate application	39444	0	23h 1m	83d 1h
W_Call incomplete files	23218	0	21h 11m	158d 21h
W_Complete application	2918	7m 23s	6h 4m	30d 22h
W_Call after offers	31485	0	23m 23s	96d 27m
W_Handle leads	3727	1m 24s	20m 58s	2d 16h

execution time is correct. Table 13. shows the execution time of workflow events. We focused on the top 3 activities because they take a long time.

W_Assess potential fraud Table 14. is a basic statistic of the top 4 resources who mainly do the activity W_Assess potential fraud. This activity is done by User_138, 143 and 144. They belong to group 4 of clustering which usually takes a long time.

Table 14. Basic statistic of the top 4 resources who mainly do the activity W_Assess potential fraud (d is day, h is hour, m is minute and s is second).

Resource	Frequency	Mean	Median
User_138	130	5.25d	1d
User_143	65	2.95d	1.12d
User_144	109	1.9d	0.81d
User_55	5	0.38d	31s

W_Validate application Table 15. shows basic statistic of top 4 resources who mainly do the activity W_Validate application. Top 4 resources who mainly do the activity W_Validate application are User_67, 99, 90 and 109. They belong to group 2. They take a long time to do it as well. Since User_123 did 2474 times and took an average on 1394 seconds to do it, the company needs to investigate how he handled it fast.

W_Call incomplete Table 16. shows basic statistic of top 4 resources who mainly do the activity W_Call incomplete. Top 4 resources who mainly do the activity W_Call incomplete are User_69, 26, 2 and 58. They belong to group 1. They take a long time to do it. Since User_100 did 2269 times and took a average on 1387 seconds to do it, the company needs to investigate how he handled it fast. Except for a few resources, the mean execution time tends to be lower as the execution number of resources is increased.

Table 15. The basic statistic of top 4 resources of mean for W_Validate application(d is day, h is hour, m is minute and s is second).

Resource	Frequency	Mean	Median
User_68	1818	3.51d	2.95d
User_99	1677	3.27d	2.84d
User_90	313	3.14d	1.97d
User_109	295	3.01d	2.12d

Table 16. The basic statistic of top 4 resources of mean for W_Call incomplete(d is day, h is hour, m is minute and s is second).

Resource	Frequency	Mean	Median
User_69	2	40.63d	40.63d
User_26	37	17.31d	13.91d
User_2	229	13.84d	4.01d
User_58	9	12.75d	1.06d

8.3 Waiting time of resource

In this section, we will analyze the waiting time. We defined the waiting time that the time is from 0 workload to first work arrived. Table 17. shows the basic statistics of top 8 resources of waiting time. The mean execution time tends to be longer as the execution number of resources is decreased. The User_138, 143 and 144 has a long waiting time because they do W_Assess potential fraud. We need to consider the median not the mean because resources spend weekends and holidays. So, there are resources that has less than 1.7s median of the waiting time.

Table 17. The basic statistic of top 4 resources of mean for W_Call incomplete(d is day, h is hour, m is minute and s is second).

Resource	Frequency	Mean	Median
User_142	13	35.64d	0.081s
User_103	73	6.99d	1.508s
User_141	54	6.96d	10.344s
User_82	46	5.90d	0.523s
User_144	245	5.17d	0.84d
User_111	47	4.13d	0.69s
User_143	177	3.52d	1.40h
User_138	343	2.92d	1.91d

9 Conclusion

The data used in this study were loan process log data that come from actual financial institute and its loan process. Process mining methodologies and statistical methodologies were applied to the actual loan process log data, and addressed the result of the analysis. In order to get reliable result of analysis, we used many tools such as DISCO, MINI TAP, R, MATLAB which provide many applications. By using these tools, we analyzed current loan process based on three questions that the BPIC 2017 provided.

Briefly, in question 1, we defined 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. As a result, we found out that waiting time of the financial institute is longer than the waiting time of applicants. In addition, we found bottleneck process which extend the total time of process. In question 2, we checked the power of frequency of A_Incomplete to final outcome. At first, the company thought that the number of A_Incomplete lead to increase of A_Cancelled, but as a result of analysis the ratio of A_Pending were increased and ratio of A_Cancelled were decreased. In question 3, we divided the data set into number of activity O_Create offer, and addressed the process differences among groups. Even though number of offers hardly affect each activity, the time that offers are created become various and this makes the process complicated. And offers that are not first one make the time to be needed more to make another new offer. Those number of offers give huge impact on case duration.

The contributions of this study are as follows. First, according to the existing studies, they have focused on process. In this study, however, by using the statistical methodologies we could check the problem in various perspectives. In particular, machine learning methodologies could be worthy when the financial institute has more reliable process log data. Second, quick validation is possible. For example, if the new application has arrived, then the financial institutes could check the process based on the result of our analysis. Since the insight from the process mining and statistical methodologies could support the check process robustly.

References

1. 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)
2. Song, M., & Aalst, W. M. P. Van Der. (2008). Towards comprehensive support for organizational mining. https://doi.org/10.1016/j.dss.2008.07.002
3. Song, M., Christian, W. G., & Aalst, W. M. P. Van Der. (2009). Trace Clustering in Process Mining, 109120.
4. Aalst, W. M. P. Van Der, Reijers, H. A., Weijters, A. J. M. M., & Dongen, B. F. Van. (2007). Business process mining : An industrial application, 32, 713732. https://doi.org/10.1016/j.is.2006.05.003
5. Berry, M. J., & Linoff, G. (1997). Data mining techniques: for marketing, sales, and customer support. John Wiley & Sons, Inc.. Integration. Technical report, Global Grid Forum (2002)
6. Suykens, J. A., & Vandewalle, J. (1999). Least squares support vector machine classifiers. Neural processing letters, 9(3), 293-300.
7. Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics), 28(1), 100-108.

Appendix

Table 18. Points that O_Create Offer occurs

One offer	Multiple offers
A_Accepted - O_Create Offer (12141, 99.70%)	A_Accepted - O_Create Offer (4906, 42.86%)
W_Complete application - O_Create Offer (35, 0.29%)	W_Complete application - O_Create Offer (11, 0.10%)
	A_Complete O_Create Offer (2331, 20.37%)
	O_Created O_Create Offer (1893, 16.54%)
	A_Incomplete O_Create Offer (1412, 12.34%)
	O_Cancelled O_Create Offer (439, 3.84%)
	O_Sent (mail and online) O_Create Offer (398, 3.20%)
	A_Validating O_Create Offer (43, 0.38%)
	W_Call incomplete files O_Create Offer (13, 0.11%)