Process mining on the loan application process of a Dutch Financial Institute

BPI Challenge 2017

Liese Blevi, Lucie Delporte, Julie Robbrecht
KPMG Technology Advisory, Bourgetlaan 40, 1130 Brussels, Belgium
[lblevi, ldelporte, jrobbrecht]@kpmg.com
https://home.kpmg.com/be/en/home/insights/2017/09/process-mining.html

Abstract. The BPI Challenge is an annual process mining competition, in which the participants are provided with a real-life event log. This year's event log includes all events related to the loan application process of a Dutch Financial Institute. The data includes three types of information, states of the application, states of the offer(s) belonging to the application and states of the workitem(s) belonging to the application. The process owner wants to gain insight in the throughput times per part of the process, the impact of information requests on the outcome of the process and the difference in process patterns based on the number of created offers. Therefore, we analyzed the event logs using a combination of process mining techniques and tools, including SQL, Power BI, Disco and ProM.

Keywords: BPI Challenge, Process Mining, Event logs, Loan Application Process

1. Introduction

One of the key challenges for financial institutes in the current economic environment is the increased competition from Financial Technology (FinTech) firms[1]. A way to resist this new breed of competitors is to enhance customer experience, using digitalization and automation techniques that streamline the loan processes[2].

Therefore, our goal is to analyze the loan application process with the specific objective of improving the customer experience and increasing revenue. We made use of KPMG's customer experience methodology (KPMG Nunwood) to identify the six essential pillars of customer experience: Personalisation, Integrity, Time & Effort, Expectations, Resolution and Empathy.

In this report we will focus on Time & Effort, because the other pillars cannot be evaluated based on the received data. Time & Effort states that an organization needs to:

a. Explain exactly what is needed
b. Effectively manage delays
c. Offer alternative solutions
d. Inform customers about issues
e. Use latest efficient technology

We will address these aspects by answering the process owner's questions[3]:

Nr.	Question	Aspect of Time & Effort
1	What are the throughput times per part of the process?	b, e
2	What is the influence of the frequency of incompleteness to the final outcome?	a, d
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)?	c
4	How does the conversion compare between applicants for whom a single offer is made and applicants for whom multiple offers are made?	c

2. Management summary

Based on our analysis, we made several observations that could give the process owner further insight in the loan application process. We made recommendations with the objective improving customer experience in mind.

Below, we provide an answer on the key questions as asked by the process owner:

1 What are the throughput times per part of the process?

We noted that the average lead time to process an application from submission to final decision is approximately 22 days. Half of the time is spent on work items, which we consider time spent by the bank employee. The other half of the time the bank is waiting on input from the customer. This waiting

time mainly occurs after the communication of the offer, when the customer needs to decide on taking the offer or not. Considering the related aspects of the Time & Effort pillar, we believe it might help to contact the customer more quickly and frequently after the communication. This way the bank can react faster on (changing) expectations or requirements of the client. Furthermore, we encourage the bank to not only do phone calls, but also make use of other technologies to remind the customer. For example e-mails, mobile app, etc.

2 *What is the influence of the frequency of incompleteness to the final outcome?*

In contradiction to what the process owner expected, we did not find a negative relationship between the number of requests for documentation and the decision of the customer. In our analysis we focused on activity W_Call incomplete files. However, we noted that this activity occurs after the decision of the customer. Whether an application is approved or not is at that moment the decision of the bank. Here we can see that the time spent in W_Call incomplete files has a positive impact on the final outcome, meaning that a loan is more likely to be granted if the bank has to call the customer to request additional files. In order to influence the decision of the customer it is important to focus on activity W_Call after offers.

3 *How many customers ask for more than one offer?*

In 27% of the cases, the customer asks for more than one offer. Most of the time in multiple conversions (meaning more than 1 day apart from each other). On average, a new offer is created +/- 7 days after the previous.

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

We calculated the conversion rate as the number of applications resulting in a signed offer divided by the total number of applications. We noted that the conversion rate increased as the number of offers increased. Furthermore, we noted that the conversion rate is almost 100% if the application is processed between 10 and 30 days.

Next to those 4 questions, we evaluated the 5 aspects of the Time & Effort pillar of customer experience:

a. Explain exactly what is needed

We noted that in 48% of the cases a request for completion has been done. This shows that is not fully clear for the customer which documentation

he/she has to send. Seeing that on average it takes 4 days to complete the dossier, the average lead time to process an application can be significantly increased when the instructions are better explained. Furthermore, it will have a positive effect on the customer effect as the customer will only have to put effort once (upon the initial request) in searching for documentation instead of twice or more.

b. Effectively manage delays

It is important to keep the customer informed in case of delays. For example in case of personal loan collection, the customer might have to wait a long time before he/she is made an offer by the bank. In these kind of situations, it is recommended to send an update to the customer on the status of his/her application.

c. Offer alternative solutions

We consider an offer to be a solution for the client and thus, additional offers are alternative solutions. Based on our test results, we noted that the customer appreciates additional offers from the bank. We noted that the conversion rate increases as more offers were made. Therefore, we recommend to keep this practice, but we suggest to faster respond to the customer's (changing) expectations and requirements by actively contacting him/her more frequently.

d. Inform customers about issues

This goes together with the effective management of delays. In case of issues, like for example technical errors in the system, the process might be delayed. It is important to communicate these issues towards the customer.

e. Use latest efficient technology

We noted that the bank uses an online channel, mail and telephone in order to contact the client. Offers are either communicated through the online channel only or through the online channel and mail. We recommend to also use these channels for sending out reminders, instead of calling only. Especially in the communication of delays and issues it is important to use these kind of channels in order to avoid a call overload, which might be a too intrusive way for the customer[4].

Another important aspect of customer experience is customer segmentation. Therefore, we mapped the customer journey based on loan goal. We focused on the 3 most profitable loan goals, being car, home improvement and existing

loan takeover. In our time analysis, we mainly noted a difference on the level of fraud assessment. For home improvement, this part of the process takes significantly longer than for other loan goals. In our process analysis, we noted that that for existing loan takeovers more requests for completion are done which is also visible in the time needed to validate the application. Especially for this type of loan, it might be profitable to improve the instructions on required documentation as depicted in the aspects of the Time & Effort pillar.

Customer segmentation can be performed on other attributes as well. In our predictive analysis we noted that credit score has a significant impact on the decision of the customer to take the offer or not. This might indicate that the bank has a more competitive offering for high-credit customers than for low-credit customers, which might be a basis for further investigation.

3. Our understanding

From the information provided by the BPI Challenge 2017, we understand that a Dutch Financial Institute is looking for new insight in its loan application process. In this section we describe our understanding of the process and the data delivered by the organization.

3.1 Data understanding

The following event logs were provided:

- Application: 10.4121/uuid:5f3067df-f10b-45da-b98b-86ae4c7a310b
- Offer: 10.4121/uuid:7e326e7e-8b93-4701-8860-71213edf0fbe

The Application event log contains all events with the application ID as case ID and the Offer event log contains all events with the offer ID as case ID.

The Application log contains all applications filed in 2016, and their subsequent handling up to February 2nd 2017. For all applications a number of attributes is available containing additional information about the application and the related offer(s). In table 1 a full overview of the attributes is available:

Table 1. Overview of the available data attributes

Attribute	Our understanding
caseID	The unique identifier of the application. The case identifier is necessary to distinguish different executions of the process.
taskID	The name of the event. The name always starts with the initial of the event origin (ref. EventOrigin). There are three types of tasks: – A: States of the application – O: States of the offer belonging to the application – W: States of the work item belonging to the application
originator	The unique identifier of the person who executed the task.
Eventtype	The state of the task. There are seven possible values: schedule, start, suspend, resume, complete, withdraw and ate_abort.
LoanGoal	The reason why the loan was applied for. There are fourteen possible values: Boat, Business goal, Car, Caravan / Camper, Debt restructuring, Existing loan takeover, Extra spending limit, Home improvement, Motorcycle, Not specified, Other, Remaining debt home, Tax payments and Unknown.
RequestedAmount	The requested loan amount (in EUR). The values vary between 0 and 450000.
ApplicationType	The type of the application. There are two possible values: Limit raise and New credit.
Action	The attribute is not clear.
EventOrigin	The origin of the event. There are three possible values: Application, Workflow and Offer.
EventID	The unique identifier of the event.
Timestamp	The time at which the event occurred. The timestamp is used to put the events in the right order.
Offer related attributes	
FirstWithdrawal Amount	*Only filled in when taskID = O_Create Offer.* The initial withdrawal amount.
Accepted	*Only filled in when taskID = O_Create Offer*.

	Boolean that indicates whether an offer is still valid or not (based on the assessment of certain client information).
Selected	*Only filled in when taskID = O_Create Offer.* Boolean that indicates whether an offer is signed by the customer or not.
NumberOfTerms	*Only filled in when taskID = O_Create Offer.* The number of payback terms agreed to.
MonthlyCost	*Only filled in when taskID = O_Create Offer.* The monthly costs to be paid by the customer to reimburse the loan.
CreditScore	*Only filled in when taskID = O_Create Offer.* The credit score of the customer. A high credit score provides high creditworthiness and vice versa.
OfferedAmount	*Only filled in when taskID = O_Create Offer.* The loan amount offered by the bank.
OfferID	*Only filled in when taskID starts with O_ (except when taskID = O_Create Offer).* The unique identifier of the offer. An application can have one or more offers.

The dataset is suitable to perform different types of analyses. On the one hand it can be used for descriptive and exploratory analysis, on the other hand for predictive analysis. We will start with the first type of analysis in order to visualize and discover unknown distributions and relationships between the data attributes and assess assumptions for confirmatory analysis. In a later phase we will try to build a predictive model in order to test our assumptions[5].

3.2 Process understanding

Based on the information received from the process owner (Prom Forum), we understand that each application should go through a number of states. Figure 1 shows the expected path, including the main states. We refer to table 2 for the explanation of these states[6].

We see that there are multiple paths possible. On the one hand, an application can be submitted by physically going to the bank. In that case the bank employee will enter the application in the system and the status Submitted

will be skipped. On the other hand, a client can submit his application online via a webpage. In that case the status will be set to Submitted. From then on, the states are the same for both scenarios.

The process has three possible outcomes. Two of them are triggered by the client. Either the client selects or refuses the offer. The application will be set to status Pending and Cancelled respectively. The other outcome is triggered by the bank itself in case the application is not suitable to make an offer for. The status is then set to Denied.

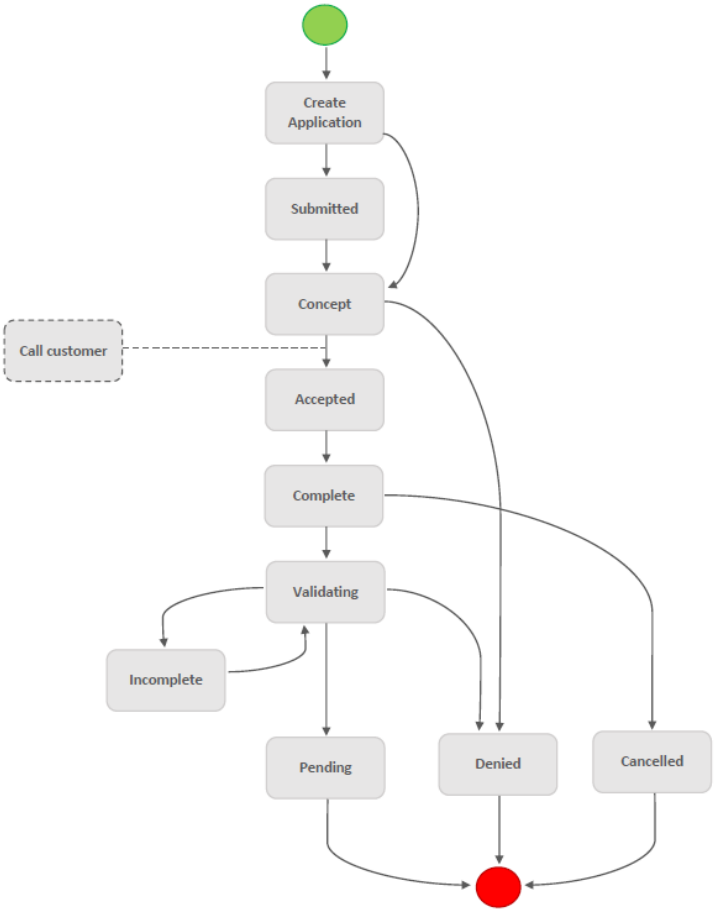


Figure 1. Expected flow of the loan application process

As requested by the process owner, we will further investigate which parts of the process are handled by the client and the bank. Furthermore, we will analyze if certain process paths are more likely to result in a successful outcome than others.

Table 2. Overview of the application states

Activity	Our understanding
Submitted	A customer has submitted a new application via the website.
Concept	A first assessment on the submitted application is done automatically.
Accepted	After contact with the customer and completion of the application, the status is accepted. The bank can now make an offer.
Complete	The offer has been sent to the customer and the bank waits for the customer to send a signed offer and the rest of the required documentation.
Validating	The signed offer and documents are received and checked by the bank.
Incomplete	The documents are not correct or some documents are still missing.
Pending	All documents are received and the assessment is positive, the loan is final and the customer is paid.
Denied	The application does not fit the acceptance criteria.
Cancelled	The customer did not send the documents or he/she called to tell he/she does not need the loan anymore.

4. Our analysis

We will look at the dataset from a process-based view. First of all we will perform an exploratory analysis by visualizing the dataset in Power BI and combining it with process mining results from Disco and ProM. This way we will try to identify distributions and relationships between events and attributes. In the next step we will perform a predictive analysis to predict the outcome of certain process flows.

4.1 Exploratory analysis

We first loaded the data into ProM 6 in order to transform the provided .xes file into a .csv file. Then we loaded the .csv file into a SQL database in order to perform some simple transformations and calculations, including event duration. We calculated the duration of an event by subtracting the

timestamp of the event from the timestamp of the next event. The calculation is explained on the basis of an example in figure 2.

caseID	Date	taskID	originator	eventtype	#Sec
Application_1000158214	2016-06-02 12:14:26.000	A_Create Application	User_1	complete	0
Application_1000158214	2016-06-02 12:14:26.000	A_Submitted	User_1	complete	1
Application_1000158214	2016-06-02 12:14:27.000	W_Handle leads	User_1	schedule	69
Application_1000158214	2016-06-02 12:15:36.000	W_Handle leads	User_1	withdraw	0

12:14:26.000 – 12:14:26.000 = 0

12:15:36.000 – 12:14:27.000 = 69

Figure 2. Calculation of event duration

Next, we visualized the data set in a Power BI report with multiple views. The time buckets and average # days as can be seen in figure 3 –view 1 are based on the calculated column #Sec shown in figure 2.

From the dashboard we can see that the dataset consists of applications created in the period between January 1th 2016 and December 31th 2016. The line graphs shows a significant increase in the number of submitted applications in June 2016 and as from November 2016 the number decreases again. The average lead time KPI shows that it takes approximately 22 days to process an application. Accordingly, the biggest part of the applications was finished between 10 and 20 days.

Figure 3. Visualization of the dataset in Power BI (from left to right, view 1 to 4)

The doughnut graph at the bottom left indicates that more than half of the events are workflow items. This can be attributed to the different event types, which is clearly visible in the second tab of the dashboard as shown in figure 3 – view 2. When selecting Workflow in the bottom right histogram, we see that 100% of the event types other than “complete” are being highlighted. This means that the other status types (Application and Offer) do not have this kind of event type distinction.

In figure 3 – view 3, the third tab of the dashboard is shown. It contains a view on loan goal. The biggest part of the applications are for car loans. This type of application is handled slightly faster than the average. From the bubble chart we can also see that car loans are part of the smaller loans (avg. 13k). This might indicate that the evaluation process is less strict and therefore certain activities take less time. The reverse is also visible. Debt restructuring (avg. 19k) and remaining debt home loans (avg. 23k) are part of the bigger loans and take the longest to process. However, these kind of loans only occur in 3% of the cases and are less significant for the analysis. In section 4.3.2, we will analyze these processes more in detail.

The last view is made on the basis of the offer related data, shown in figure 3 – view 4. The dashboard shows different graphs focusing on the number of offers created per application and the conversion rate of those applications. We calculated the conversion rate as the number of applications resulting in a signed offer (i.e. Selected = 1) divided by the total number of applications. This shows us that approximately 7 out of 10 applications are converted.

However, when we select the converted offers in the # Offers by Outcome – graph we noted that some of them are not accepted by the bank. Only 55% of all applications are actually resulting in the granting of a loan, meaning that the offer is selected by the customer but not accepted by the bank.

We noted that in 73% of the cases only 1 offer was created. In the remaining 27% of the cases, the customer requested one or more extra offers. In these last cases we made the distinction between applications where the extra offer was requested in 1 conversation or in multiple conversations. When offers for the same application are created less than 8 hours apart from each other, we consider them to be requested during 1 conversation. The conversion rate increases when more offers are created and is higher in case of multiple conversations. Multiple offers resulting from 1 conversation have a conversion rate of 65,34%, while these resulting from multiple conversations have a conversion rate of 81,75%.

The Power Bi dashboard makes it easy to look at data, slice it up and look at it again in different ways. We made the dashboard[7] publicly available for the process owner to perform further exploratory analysis on. You can find it on our website.

Our main observations, taking into account the process owner's questions, are the following:

- The most time-consuming activity is W_Personal loan collection. On average, it takes 18 days to move to the next status. The average lead time per case increases to 273 days, which is more than 10 times the overall average. However, the activity only occurs in 2 cases.
- In 33% of the cases the application is cancelled, in 12% denied and in the remaining 55% approved. The average time spend on these applications is 30 days, 17 days respectively 22 days.
- In 100% of the cases at least one suspend action occurred, 28% of the work item events are suspend actions.
- There is no clear distinction between types of users. The distribution between application, offer and workflow events is quite similar across the users. User_1 performs significantly more activities than all other users.
- In the most time-consuming cases (> 50 days), the O_Sent activities take 3 times longer than the overall average.
- Activity W_Call incomplete files occurs in 48% of the cases and takes on average half a day to move to another status.
- The conversion rate is almost 100% when the processing time of the application is between 10 and 30 days.
- Extra offers have a positive impact on the conversion rate. The average conversion rate for applications with one offer is 69,09% while it is 73,24% for applications with 2 or more offers.
- When extra offers are created upon multiple conversations the impact is even bigger. The overall conversion rate increases to 81,75%.
- In case of multiple conversations, the average time between 2 offers is around 11 days.
- There is a significant drop in the conversion rate when it takes more than 30 days to process the application.
- The average number of days is the highest for loan goals 'debt restructuring' and 'Remaining debt home'. Since 'debt restructuring' only

consists of 2 cases, it is not relevant to analyze this. The sub process for ‘remaining debt home’ will be discussed later.

4.2 Process mining

Process mining techniques allow for extracting information from event logs. It creates a process model from the data in the traces and represents the current-state operation. We applied process mining techniques for process discovery and to evaluate the process performance and patterns. We made use of Disco and ProM.

4.2.1 Process discovery

We imported the event log in ProM and ran the fuzzy miner in order to get a first high-level view on the process. We selected the best edges in order to preserve the best incoming and outgoing edge for each activity[8].

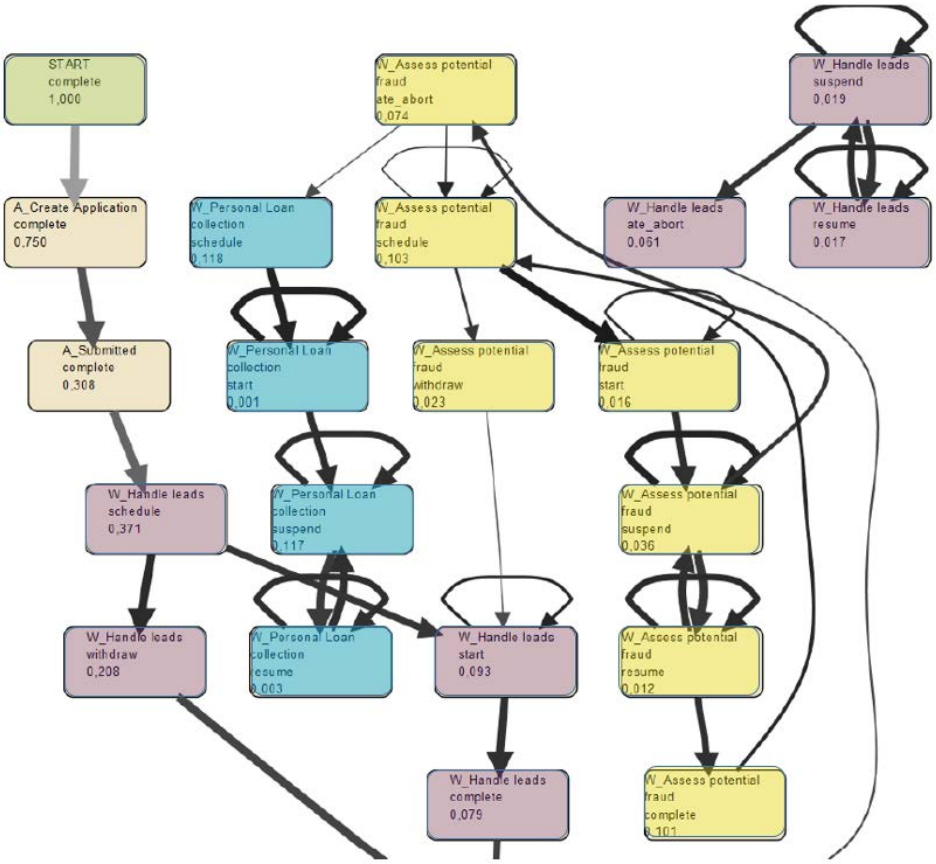


Figure 4. Process map (part 1)

Because of the size of the process map, we decided to split up the flow in different parts. In the first part of the process map we see the process steps between the submission of the application and the acceptance of the application. It can be considered as the review of the application. The review consist of 3 activities:

- W_Handle leads (purple): This activity includes the first assessment of the application. Normally the activity is performed automatically, unless there is a technical error. We noted the following:
    - The suspend, resume and abort actions are disconnected from the other actions. This might point to the manual interference in case of technical error. We clearly see ping-pong behavior between the suspend and resume actions in this case.
    - There is a thick arrow towards the withdrawn action, indicating that it is a significant process path. It might indicate that the first assessment is skipped a lot.
- W_Assess potential fraud (yellow): In the assessment of potential fraud we also see the ping-pong behavior between the suspend and resume actions.
- W_Personal loan collection (blue): As seen from our exploratory analysis this activity is less significant, as it occurs in 2 cases only. We noted that there is no complete action on this activity, which strengthens our observation of the high lead time of those cases.

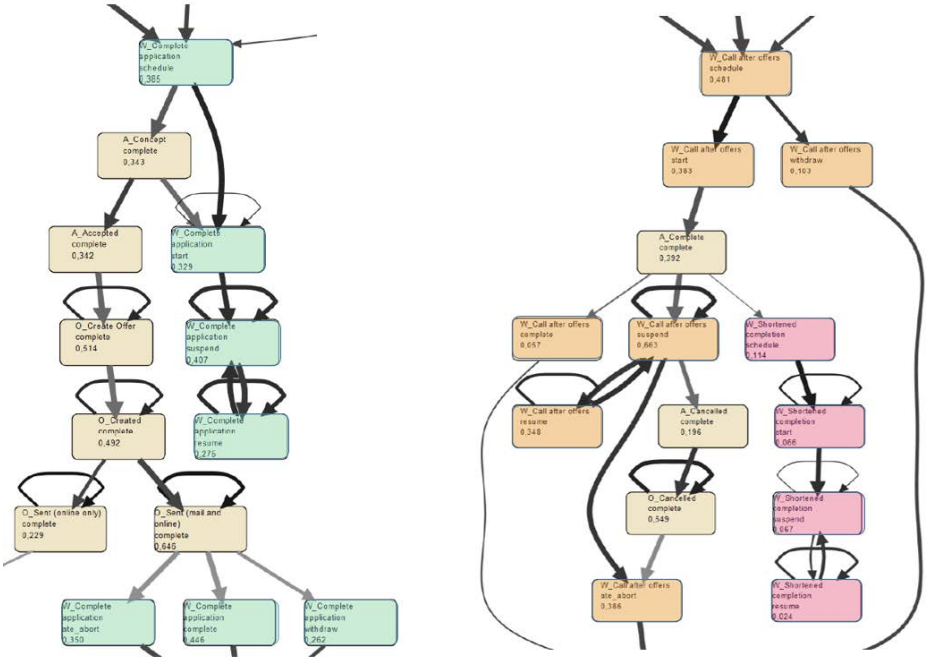


Figure 5. Process map (left: part 2; right: part 3)

When the review is completed, the first draft of the application is saved and a new work item is scheduled to complete the application. Furthermore, an order is created and send to the customer. Just like the other work items, W_Complete application (green) again shows ping-pong behavior.

After the application completion, there are two possible paths:

- W_Call after offers (orange): A bank employee contacts the customer to follow-up on the offer. He/she gathers the thoughts of the customer about the offer. If the customer indicates he/she is not interested in the offer or does not answer the inquiries, both the offer and the application will be cancelled.
- W_Shortened completion (pink): There is a small arrow towards shortened completion, which indicates that in a minority of the cases a shortened completion is scheduled.

Both work items show the same ping-pong behavior as noted before.

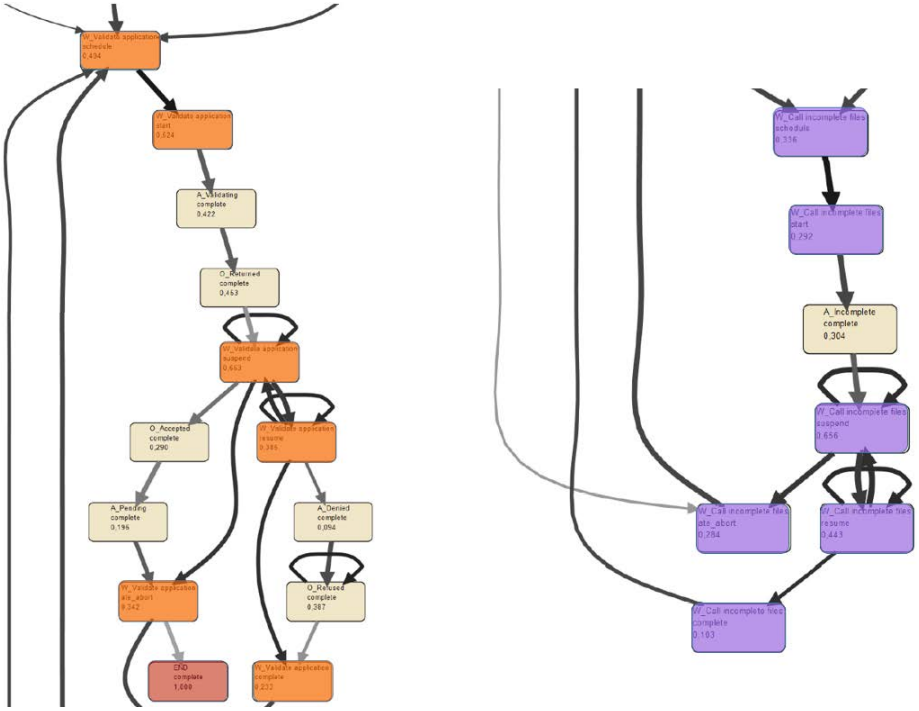


Figure 8. Process map (left: part 4; right: part 5)

When the customer is interested in the offer, the bank starts validating the application. The validation of the application either result in an accepted offer, meaning that the customer meets the requirements to grant the loan, or in a denied offer, meaning that the bank withdraws the offer because the customer does not meet the requirements for the loan (e.g. insufficient income, medical issues).

During the validation it occurs that not all information is available to decide on the suitability of the customer. In that case the bank needs to call the customer in order to request the missing documentation. When all the documentation is gathered, a new work item for application validation is scheduled. Both in W_Complete application (orange) and W_Call incomplete files (purple) there is ping-pong behavior like in the other work items.

Our main findings are:

- All work items demonstrate ping-pong behavior between the suspend and resume actions. This indicates that none of the work items are executed without interruption. It is a form of waste that has a negative impact on the throughput time of the process and thus, also on the customer experience.

- The first assessment of the application should be performed automatically, but there is indication of manual interference due to technical errors. It may point to misconfiguration of the system, which also can have a negative impact on the customer experience. The Time & Effort pillar points out the importance of the use of efficient technology.
- The first assessment of the application seems not to be sufficient to determine whether a customer is suitable for an offer or not. We would expect that applications are denied at the beginning of the process, but we could see in the process map that applications are denied quite late in the process (part 4) when the customer already expressed his/her interest in the offer.
- Calling for incomplete files is an important aspect of the process, as indicated by the thickness of the arrows in the process map. It might indicate that it is not clearly explained to the customer which documentation he/she has to deliver, which is also an aspect of Time & Effort pillar in customer experience.

4.2.2 Time analysis

In order to answer the first question of the process owner we have performed a time analysis. We understand that the process owner want to gain insight in the difference between the time spent by the bank's employees and the time spent waiting on input from the customer.

Workflow items indicate time that is spent in the company. These events have a status, such as start, suspend, resume, complete … which makes it possible to evaluate the throughput time. We calculated this throughput time for the different tasks per case by creating a start and end date.

We calculated the throughput time in different manners. We first calculated it as the time between the start event (i.e. event type = start) and the end event (i.e. event type = complete, withdraw or abort) of a work item. Work items without start or end event were filtered out. Next, we calculated the total duration between the first time that a work item occurred in the process and the last time that a work item occurred in the process. If a work item occurred twice for example, we took the first event of the first work item and the last event (irrespective of whether it is an end event) of the second work item. We refer to figure 9 for a schematic explanation.

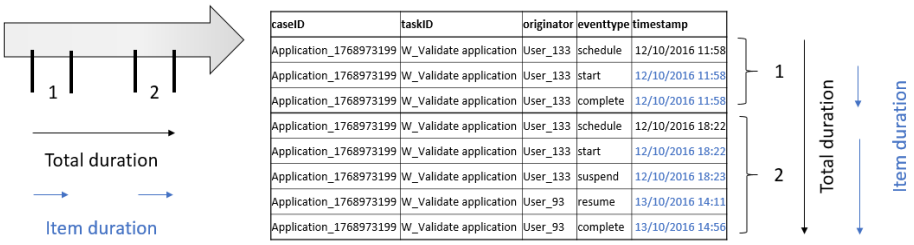


Figure 9. Calculation work item throughput time

As can be seen from table 4, there are significant differences in the results for work items W_Call incomplete files and W_Validate application. This indicates that these activities are more likely to occur more than once in the process. For the other work items the difference is less significant.

Work items W_Call after offers and W_Call incomplete files are the most time-consuming activities in both calculations. These are the two work items that require interaction with the customer and therefore might be biased by waiting time. In order to further reduce this bias, we further split up the work item based on user.

In the second work item of figure 2 there are 2 users involved, User_133 and User_93. Taken this into account, we created a new table with start and end timestamps for the work items as shown in table 3. The throughput time of the second work item is then equal to 45 minutes. Note that the results might still be biased due to work items that are fully executed by one user.

Table 3. Format new table

caseID	taskID	originator	timestamp_start	timestamp_end	minutes
Application_1768973199	W_Validate application	User_133	12/10/2016 11:58	12/10/2016 11:58	0
Application_1768973199	W_Validate application	User_133	12/10/2016 18:22	12/10/2016 18:23	0,02
Application_1768973199	W_Validate application	User_93	13/10/2016 14:11	13/10/2016 14:56	45

Table 4. Overview throughput times

Work items	Item duration (days)	Total duration (days)	User duration (days)
W_Assess potential fraud	3,45	4,17	3,99
W_Call after offers	13,98	15,31	1,00
W_Call incomplete files	3,98	6,82	3,15
W_Complete application	0,97	1,57	0,12
W_Handle leads	0,02	0,05	0,00
W_Validate application	1,63	5,19	2,35
W_Personal Loan collection	No end event	190,34	0,00
W_Shortened completion	No end event	7,51	0,00

W_Call after offers is no longer the most time-consuming activity. The top 2 are now W_Call incomplete files and W_Validate application. The total amount of days spent on work items is approximately 11 days. If we combine this with the average lead time that we found in our exploratory result, being +/- 22 days, we can say that almost half of the time the bank is waiting on input of the customer.

However, seeing the big difference between the item duration and user duration for W_Call after offers we would recommend to call the customer more quickly after the offer has being sent. This way the bank may gain input a lot sooner, respond better to the client's expectations and send an alternative offer more quickly. Furthermore, it might be an option to send reminder e-mails.

3.2.3 Customer behavior analysis

We will now consider the process in more detail. Is the process significantly different for a specific loan goal? Is the throughput time of specific tasks longer when only part of the cases are taken into account? Do some events become irrelevant? We will answer these questions in the first part of this section. Next, we will determine the amount of customers that ask for more than one offer (hereby making a difference between whether this is done in a single conversation or multiple conversation) and the effect that this has on the characteristics of the complete process.

In order to start the analysis, we refer back to figure 5 where we made a visualization on the different loan goals of the process, such as the average number of days per loan goal, the number of cases for each loan goal, etc.

Based on certain characteristics in this figure, that are of high importance to the bank, we selected the most important sub processes.

- Most cases: The number of cases is an important attribute of a sub process considering that this type of loan is mostly asked by customers and so the process should be standardized in the best way possible to avoid delay. The most popular products are the loans for **car**, **home improvement** or **existing loan takeover**.
- Highest request amount: The amount of money that is asked by the customer is important, since this has a high impact on the turnover of the bank. Note however that we also took into account the number of cases, since it is not relevant to evaluate a sub process that the company rarely

encounters. Existing loan takeover has a high average requested amount, but was already selected as part of 'most cases'. Next to existing loan takeover, **remaining home debt** is the loan goal for which the highest amount is requested. There is a significant difference with the rest of the processes.

- Most events: We will also briefly discuss **debt restructuring**, since this sub process consists of the highest number of events.

We will compare the time spent in the processes between the different loan goals. For each loan goal selected above, we calculated the total duration and user duration (as explained in section *3.2.2)* of the work items.

Table 5. Average total duration (in days)

	Assess personal fraud	Call after offer	Call incomp-lete files	Complete applica-tion (5)	Handle Leads (5)	Shorten-ed comp-letion (5)	Validate applica-tion
Car	3.6	15.25	5.43	1.57	0.04	4.32	4.23
Home improvement	5.51 (3)	15.64	6.66	1.70	0.03	11.34	4.70
Existing loan takeover	3.24	15.90	7.35	1.75	0.07	0.00	5.83
Remaining home debt	0.88	14.76	16.21 (2)	1.54	0.03	NA	15.12 (4)

Table 6. Average user duration (in hours)

	Assess personal fraud	Call after offer	Call incomp-lete files	Complete applica-tion	Handle Leads	Shorten-ed comp-letion	Validate applica-tion
Car	78.47	25.15	52.91	2.21	0.04	0.05	37.18
Home improvement	134	24.31	55.43	3.04	0.04	0.02	42.47
Existing loan takeover	58.94	21.98	79.04	4.10	0.08	0.02	62.68
Remaining home debt	11.40	21.60	340.40	3.33	0.02	NA	321.52
Debt restructuring	NA	324.25	NA	0.04	0.02	NA	NA

Furthermore, we also imported the process in Disco (we filtered the path slicer on 12,5%). Note that we imported our new table as shown in table 3, so for the workflow items, a loop indicates that another user continues the activity. The application and offer events have only one timestamp (completeness of application), so the start and end date are the same and only the waiting time between two activities can be evaluated for this.

We filtered on the specific loan goal and created a table with the frequency of the workflow event (going from -- to ++) based on the frequency view in Disco.

Table 7. Frequency of workflow events

	Assess personal fraud	Call after offer	Call incomp-lete files	Complete applica-tion	Handle Leads	Shorten-ed comp-letion	Validate applica-tion
Car	--	++	+	++	+-	--	+
Home improvement	-	++	+	++	+-	--	+
Existing loan takeover	--	++	++	++	+-	--	++
Remaining home debt	--	+	++	+	-	--	+

Based on these results, our main findings are:

(1)

Call after offers is a big problem in the process of debt restructuring. Since the other workflow items take a very limited amount of time to complete, the average number of 31 days to walk through the process will be mostly caused by this problem. The number of hours spend on this by the employees is 130 which is a lot higher than for the other loan goals. In the original dataset we see that this loan goal is only applicable to two cases. One user withdraws one of the applications a few months later, which causes the high number of days. Since this considers only a few cases, it is not necessary to further investigate this. An example of this behaviour is shown in table 8.

Table 8. Example of withdrawal

Application_1165780533	W_Call after offers	User_1	ate_abort	4/12/2016	7:32:06
Application_1165780533	W_Call after offers	User_1	schedule	4/12/2016	7:32:06
Application_1165780533	W_Call after offers	User_1	withdraw	31/12/2016	8:00:13

Also for the other types of loans call after offers is the biggest problem in the process, with a total duration of around 15 days and a rather high number of hours to complete ‘call after offers’. The high number of hours spent on this is probably due to the fact that the customer does not pick up his phone immediately or needs multiple reminders before sending in the documents. We recommend to have a policy on how many times the customer should be called before cancelling the application. With an SQL query we see that a customer is called on average 2 times when the application is not cancelled

(which means that the customer is interested and there is an application event 'A_Validating'), so we recommend to call approximately 2-3 times before cancelling the application. This will reduce the time spent on this activity. The company could for example implement a procedure as shown in figure 10.

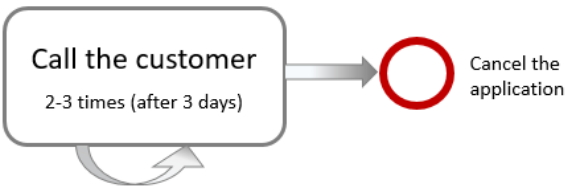


Figure 10. Recommended procedure for W_Call after offers

Note that this activity also contains waiting time since input from the customer is required before this activity can be finished.

(2)

The same problem arises for 'W_Call incomplete files'. Especially for remaining home debt, the number of hours spend on this is very high. Since this is also mostly waiting for input of the client and calling again if no answer is received, it is difficult to optimise this.

(3)

Assess personal fraud is very time-consuming for home improvements and to a lesser extent, cars. In the dataset, we see that this activity takes place in less than 1% of the cases for both loan goals, which makes the activity rather unimportant to evaluate further. The high number of hours may also be required to perform this activity. If deemed necessary, the company could develop a standardised method for assessing this personal fraud.

(4)

'Validate application' has a high throughput time in the case of remaining home debt.

(5)

The other activities, such as 'handle leads', 'complete application' and 'shortened completion' do not seem to cause an issue for the company. Shortened completion takes a long time to complete for home improvements and for cars, while employees did not perform many hours on this. We see in the database that there are only 72 cases in which shortened completion takes place and that there are some outliers, which cause the high completion time. An example is shown in table 8.

Table 9. Outlier in W_Shortened completion

CaseID	TaskID	Originator	Eventtype	Timestamp		LoanGoal
Application_1068646658	W_Shortened completion	User_43	start	5/07/2016	11:19:37	Car
Application_1068646658	W_Shortened completion	User_43	suspend	5/07/2016	11:19:39	Car
Application_1068646658	W_Shortened completion	User_1	resume	25/10/2016	16:32:05	Car
Application_1068646658	W_Shortened completion	User_75	suspend	25/10/2016	17:33:29	Car

Next, we evaluate the time between two activities, which can be considered waiting time.

First interesting fact is that the time between call after offers and application cancelled is always very long and happens for quite a few cases.

Table 10. Average time between call after offers and cancelled

Loangoal	Time between call after offers and cancelled (in days)	Frequency of this path
Car	19.8	2900 (of 9300 applications)
Home improvement	20.4	1900 (of 7600 applications)
Existing loan takeover	20.4	1451 (of 5600 applications)
Remaining home debt	18.3	203 (of 842 applications)

As already mentioned, we recommend to implement a standard process that indicates when to cancel an application (how much time after the last call).

Also the time after sent mail and validate application is very long for almost all of the loan goals. This cannot be changed easily, since this is the time that the bank is waiting on input from the client. Note that there are a lot more offers sent by mail than only online, therefore we consider the path between sent online and validating is less important.

Table 11. Average time between sent online and sent only and by mail

	Sent online – A validating (in days)	Sent online and by mail – A validating (in days)
Car	2.6	4.9
Home improvement	6.1	3.7
Existing loan takeover	4.2	3
Remaining home debt	2.8	4.2

Furthermore, we will discuss the waiting times between two activities that have not been discussed yet based on the performance view in Disco. We will only discuss the path frequency when the waiting time is high. We will not discuss the parts of the processes that have no critical waiting time.

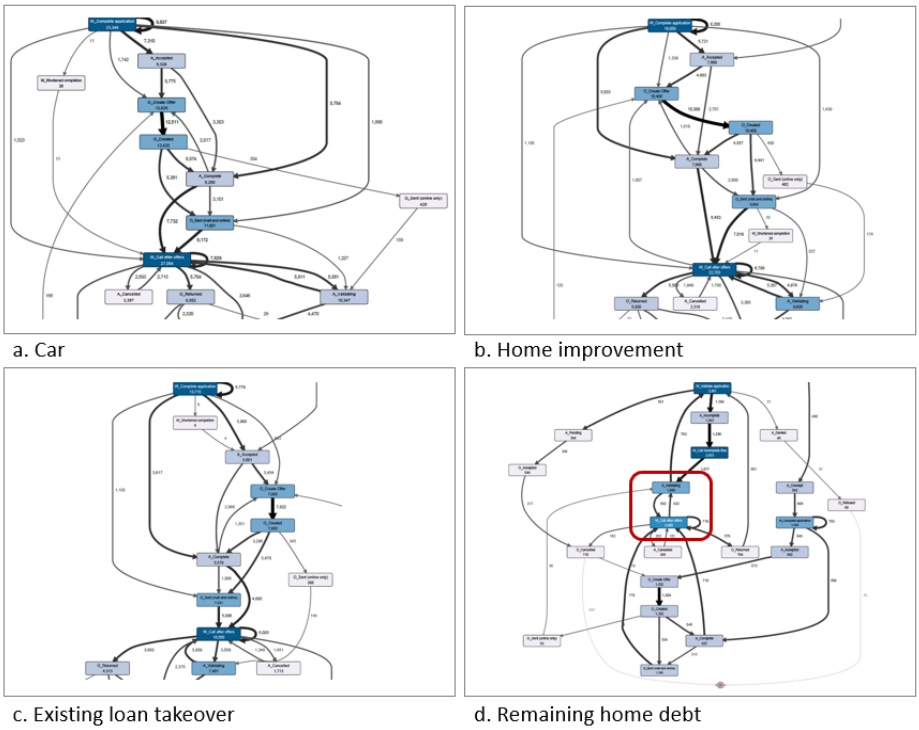


Figure 11. Process maps per loan goal (created with Disco)

The 3 most occurring loan goals show great similarity to the overall process as described in section 4.2.1. Process discovery. The main difference we see is the use of Shortened completion in case of car loans, but not in the 2 other loan goals.

For all 3 loan goals, we note that if the order was sent online only no call after offers is executed. We assume that offers are sent online only in case the bank does not dispose of the clients contact information, like e-mail address, telephone number, etc. If we further assume that these are the cases that are submitted through the web application, it indicates that the online form does not ask for contact information or that the related fields are not mandatory to fill in.

We also inspected the process map of remaining home debt and noted that it is significantly different from the other 3 maps. The activities are shown in a different sequence. The order creation seems to take place after call after offers and call after offers seems to take place after the order validation. However, when following the arrows properly we can see that the process is

actually similar. This bias might be caused by our table transformation. In a lot of cases the W_Call after offers is aborted and at the exact same time W_Validate application is started by another user than the one who performed the other actions (schedule, start, suspend, etc.) on the W_Call after offers. Furthermore, there is a higher demand for additional offers than in the other loan goals. Which might also explain a part of the W_Call after offers before the offer creation. If the bank employee has to put more effort on the call after the initial offer than the additional offers, the significance is higher earlier in the process.

In summary, we assume that there is a lot of switch-over between users and a high demand for one or more additional offers. This might indicate that the initially assigned user did not have the correct knowledge in order to evaluate the application and to make an appropriate offer for the customer's requirements. This can be a point for the bank to further investigate.

4.3 Predictive analysis

In order to determine whether certain attributes or process behaviors have an impact on the process outcome, we have performed predictive analysis. On the one hand, we performed predictive analysis on the basis of the offer related attributes and on the other hand on the basis of the occurrence of certain activities. For this last analysis, we had to create a new table.

Per application ID, we calculated the number of occurrences per activity (Freq.) and the total time spent per activity (Time (sec)) as shown in figure 12. The time per event is calculated as explained in figure 2 of the exploratory analysis.

CaseID	TaskID	Originator	Eventtype	Timestamp		Time
Application_2142897957	A_Create Application	User_1	complete	22/09/2016	13:34:18	0
Application_2142897957	A_Submitted	User_1	complete	22/09/2016	13:34:18	0
Application_2142897957	W_Handle leads	User_1	schedule	22/09/2016	13:34:18	66,58333
Application_2142897957	W_Handle leads	User_1	withdraw	22/09/2016	13:35:26	0
Application_2142897957	W_Complete Application	User_1	schedule	22/09/2016	13:35:26	0
Application_2142897957	A_Concept	User_1	complete	22/09/2016	13:35:26	18769,65
Application_2142897957	W_Complete Application	User_91	start	22/09/2016	18:54:55	191,9176
Application_2142897957	W_Complete Application	User_91	suspend	22/09/2016	18:58:11	1932,875
Application_2142897957	W_Complete Application	User_91	resume	22/09/2016	19:31:05	21,54167
Application_2142897957	W_Complete Application	User_91	suspend	22/09/2016	19:31:27	35,25
Application_2142897957	W_Complete Application	User_91	resume	22/09/2016	19:32:03	554,2083
Application_2142897958	A_Accepted	User_91	complete	22/09/2016	19:41:29	197,7917
Application_2142897959	O_Create offer	User_91	complete	22/09/2016	19:44:51	0,979166
Application_2142897960	O_Created	User_91	complete	22/09/2016	19:44:52	106,7292
Application_2142897961	O_Sent (mail and online)	User_91	complete	22/09/2016	19:46:41	0
Application_2142897962	W_Complete Application	User_91	complete	22/09/2016	19:46:41	0
Application_2142897963	W_Call after offers	User_91	schedule	22/09/2016	19:46:41	0
Application_2142897964	W_Call after offers	User_91	start	22/09/2016	19:46:41	0
Application_2142897965	A_Complete	User_91	complete	22/09/2016	19:46:41	83,22918
Application_2142897966	W_Call after offers	User_91	suspend	22/09/2016	19:48:06	313099,3
Application_2142897967	W_Call after offers	User_132	resume	26/09/2016	12:37:27	20,5625
Application_2142897968	W_Call after offers	User_132	suspend	26/09/2016	12:37:48	2267924
Application_2142897969	A_Cancelled	User_1	complete	23/10/2016	8:00:46	0
Application_2142897970	O_Cancelled	User_1	complete	23/10/2016	8:00:46	0
Application_2142897971	W_Call after offers	User_1	ate_abort	23/10/2016	8:00:46	0

A_Create Application		A_Submitted		W_Handle leads		W_Complete application		A_Concept	
Freq.	Time (sec)	Freq.	Time (sec)	Freq.	Time (sec)	Freq.	Time (sec)	Freq.	Time (sec)
1	0	1	0	2	66,58	7	2735,79	1	18769,65

A_Accepted		O_Create Offer		O_Created		O_Sent (mail and online)		W_Call after offers	
Freq.	Time (sec)	Freq.	Time (sec)	Freq.	Time (sec)	Freq.	Time (sec)	Freq.	Time (sec)
1	197,79	1	0,98	1	106,73	1	0	6	2581044,17

A_Complete		A_Cancelled		O_Cancelled	
Freq.	Time (sec)	Freq.	Time (sec)	Freq.	Time (sec)
1	83,23	1	0	1	0

Figure 12. Data transformation for predictive analysis

4.3.1. Predictive analysis based on offer related attributes

We try to predict the value of "Selected", which indicates whether the customer signs the offer or not. Hence, we use this variable as the response variable. It is a categorical variable, meaning that the only possible values are TRUE or FALSE.

We first cleaned the dataset and filtered out all offers for which the "Selected" column was not filled in. Furthermore, we noted that a significant amount of offers did not have a credit score. In order to minimize the bias of this variable we replaced the null-values with the median of remaining values. Our final dataset contains 42 995 offers.

We calculated 2 extra variables which might have a significant influence on whether an application is selected or not. These variables are:

- FrequencyOfIncompleteness: indicates whether the received documents were incomplete (based on the occurrence of activity A_Incomplete)
- Duration_days: the time it takes from creating the application until the final decision regarding the application (approved, denied, cancelled)

We used 2 different methods for the predictive analysis: logistic regression and random forest. To this end, we made use of R, a free tool for statistics and data modeling.

1 Logistic regression

Logistic regression is a simple classification algorithm. We used the glm() function to fit generalized linear models and the predict() function to predict the probability that an application is selected or not.

We used a random split of 80/20 to divide the data in a training set and a test set. The model is fitted to the data in the training set. The fitted model is then used to predict whether the offers in the test set have TRUE or FALSE in the "Selected" column. In table 12 the result of the fitting is shown.

Table 12. Results logistic regression

Variable	**Estimate**	**Std. Error**	**Pr(>|z|)**	**signif. Code**
(Intercept)	2,86	0,251	5,75E-30	***
ApplicationTypeNew credit	-0,993	0,0494	6,77E-90	***
AcceptedTRUE	0,203	0,0272	8,75E-14	***
MonthlyCost	-0,00064	0,00014	5,18E-06	***
CreditScore	-0,00124	0,000203	1,04E-09	***
FrequencyOfIncompleteness	0,645	0,0133	0	***
Duration_days	-0,0642	0,00103	0	***

We only considered the most statistically significant variables. From these variables, application type has the lowest p-value suggesting a strong association of the application type with the probability of the offer being selected. The negative coefficient for this predictor suggest that all other variables being equal, new credit loans are less likely to be selected than limit raise loans. Also the monthly cost, credit score and duration have a negative impact on the outcome. This last one confirms the importance of the Time & Effort pillar for the customer.

We then checked the predicted values with the actual values and created a confusion matrix which displays the number of correctly and incorrectly predicted (classified) observations. There are 2789 observations correctly classified as FALSE and 3333 observations are correctly classified as TRUE. On the other hand, 2477 (1014 + 1463) observations are incorrectly predicted. The fitted model has an accuracy of 71,2%, which means that the predictive value is limited.

Table 13. Confusion matrix

		Selected	
		FALSE	**TRUE**
Prediction	**FALSE**	2789	1014
	TRUE	1463	3333

2 Random forest

In order to get a higher predictive value, we tried the random forest method. The random forest method has different advantages over logistic regression. One main advantage is that it does not expect linear features or even features that interact linearly. Furthermore, logistic regression can hardly handle categorical features while this is no problem for random forest. In the random forest approach, a large number of decision trees are created. For every observation, all of these decision trees are executed and the most common outcome is used as the final output of the model. We used the same training and test set as for the logistic regression.

The results of the fitting are presented in table 14. The model has an accuracy of 89,8%, which is a significant improvement with respect to the logistic regression. The random forest method also provides an indication of the importance of each variable. The importance is measured with the Mean Decrease Accuracy which is based on the decrease in accuracy when removing the variable from the model. Variables with a large mean decrease in accuracy are more important for the classification than the others. The top 3 most influencing variables are the credit score, the duration and the frequency of incompleteness.

Table 14. Results of the random forest algorithm

Variables	MeanDecreaseAccuracy
Creditscore	443,23877
Duration_days	209,65442
FrequencyOfIncompleteness	163,06181
ApplicationType	160,78845
Accepted	110,31862
FirstWithdrawalAmount	86,01613
MonthlyCost	76,40768
OfferedAmount	72,75993
NumberOfTerms	65,08257
RequestedAmount	61,39399
LoanGoal	37,87916

		Selected	
		FALSE	TRUE
Prediction	FALSE	3963	585
	TRUE	289	3762

These results might indicate that the bank has a more competitive offering for high-credit customers than for low-credit customers. In the light of customer experience, it is important to further investigate this as both types of customers belong to a different segment and thus require a different approach.

4.3.2. Predictive analysis based on the occurrence of certain activities

We try to predict the occurrence of A_Pending, which is the activity that occurs when a loan was effectively granted to the customer (i.e. offer selected = true and offer accepted = true). We use 3 different methods for the predictive analysis: logistic regression, random forest and neural network.

For the first 2 methods, we made again use of R. For the neural network, we made use of Microsoft Azure Machine Learning Studio.

1 Logistic regression

We reduced our dataset to the following activities: O_Created, O_Sent (mail and online), O_Sent (online only), W_CallIncompleteFiles, W_CallAfterOffers, W_AssessPotentialFraud, W_HandleLeads, W_ValidateApplication and A_Submitted. The other activities were filtered out because they might create a bias in correlation (e.g. A_Pending and A_Cancelled are perfectly correlated, as only one of them can occur).

For the logistic regression, we made use of the glm() and predict() functions as explained before. The results of the fitting are shown in table 15.

Table 15. Results of the logistic regression method

Variable	Estimate	Std. Error	Pr(>|z|)	Signif. code
(Intercept)	1,18	0,0525	< 2e-16	***
freq_A_Submitted	-0,326	0,0354	< 2e-16	***
freq_O_Sent_mailAndOnline	0,667	0,0771	< 2e-16	***
freq_O_Sent_onlineOnly	0,908	0,101	< 2e-16	***
freq_W_CallAfterOffers	-1,84	0,136	< 2e-16	***
freq_W_HandleLeads	-0,849	0,0628	< 2e-16	***
time_O_Sent_mailAndOnline	-6,6E-07	4,82E-08	< 2e-16	***
time_W_CallAfterOffers	-1,3E-06	2,14E-08	< 2e-16	***
time_W_CallIncompleteFiles	2,95E-07	3,17E-08	< 2e-16	***
time_W_ValidateApplication	2,25E-06	9,12E-08	< 2e-16	***
time_W_HandleLeads	-8,3E-07	1,47E-06	0,57162	
time_A_Submitted	0,0178	0,0174	0,30596	
time_O_Created	-4,9E-07	2,89E-07	0,09331	.
time_O_Sent_onlineOnly	-5E-07	1,33E-07	0,00016	***
time_W_AssessPotentialFraud	-2,1E-06	5,24E-07	4,37E-05	***
freq_W_AssessPotentialFraud	-0,852	0,183	3,21E-06	***
freq_W_CallIncompleteFiles	-0,25	0,0536	3,14E-06	***
freq_O_Created	-0,36	0,0708	3,58E-07	***
freq_W_ValidateApplication	0,187	0,0235	1,68E-15	***

We identified 9 variables with the lowest p-value, including 5 frequency- and 4 time-related variables. For activities O_Sent (mail and online) and W_Call

after offers, we noted that both the frequency- and time-related variables have a significant influence. In case of O_Sent the frequency has a positive impact on the outcome, and the time has a negative impact. This indicates that the offer is more likely to be selected when the customer is notified more frequently. In case of W_Call after offers both variables have a negative impact, which points in the same direction as the finding regarding O_Sent.

2 Random forest

We also applied the random forest algorithm on the activity behavior and noted that the time spent on activity W_Validate application is the most important variable. The remaining variables have little to none impact on the outcome.

3 Neural network

A neural network is a set of interconnected layers, in which the inputs lead to outputs by a series of weighted edges and nodes[9].

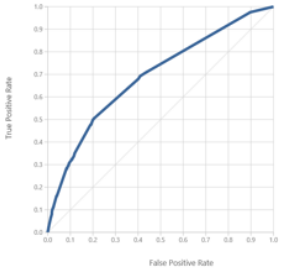
We uploaded the newly created frequency table into Microsoft Azure Machine Learning Studio and selected same activities as for the logistic regression and random forest. In the next step we normalized the data and applied the filter based feature selection, which automatically removes all irrelevant columns from the model based on correlation. Thereafter, we split the data to train and evaluate the neural network.

In the selection of the columns, we noted that the tool automatically filtered out the 4 activities as presented in table 15.

Table 16. Overview of automatically filtered columns

Activity	**Correlation with A_Pending**
O_Sent (mail and online)	0,033732
O_Created	0,057063
W_CallAfterOffers	0,010425
W_AssessPotentialFraud	0,034858

These figures indicate that the number of offers made to the client has no significant impact on the successful outcome of the process. The most important activities seem to be W_ValidateApplication, A_Submitted (i.e. application via website or not) and W_HandleLeads (i.e. first assessment of the application).



True Positive	False Negative	Accuracy	Precision
3663	1486	0.642	0.659
False Positive	**True Negative**	**Recall**	**F1 Score**
1893	2409	0.711	0.684
Positive Label	**Negative Label**	**AUC**	
1	0	0.692	

Figure 13. Results of the neural network

In figure 13 the Receiver Operator Characteristic is shown, which plots the true positives (Recall) against the false positives (precision). The predictive accuracy of the model is calculated as the Area Under the Curve (AUC). This metric ranges from 0,5 to 1, with 1 being perfect classification and 0,5 being pure luck. Our Two-Class neural network has an AUC of 0,69 which means that the frequency of the abovementioned activities have a limited predictive value.

KPMG

Process Mining

Find and present process alternatives to real life processes to make them more efficient and controls more effective.

Contact us

Anthony Van de Ven
Partner
KPMG Advisory
T +32 3 821 18 59
E avandeven@kpmg.com

Peter Van den Spiegel
Director
KPMG Advisory
T +32 2 708 37 79
E pvandenspiegel@kpmg.com

Liese Blevi
Senior Advisor
KPMG Advisory
T +32 3 821 19 59
E lblevi@kpmg.com

home.kpmg.com/be/en/home/insights/2017/09/process-mining.html

5. Bibliography

1 Vítor Constâncio (July 1, 2016). Challenges for the European banking industry. European Central Bank (ECB). https://www.ecb.europa.eu/press/key/date/2016/html/sp160707_1.en.html

2 Hervé Leasage (May 23, 2016). How can a better customer experience with loan application process help support banks revenue growth? https://www.linkedin.com/pulse/how-can-better-customer-experience-loan-application-process-lesage

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

4 Larry Myer (October 27, 2015). How To Capture Customer Attention In A World Of Information Overload. https://www.forbes.com/sites/larrymyler/2015/10/27/how-to-capture-customer-attention-in-a-world-of-information-overload/#72e45bb162aa

5 Jerry A. Smith (February 5, 2013). Six Types Of Analyses Every Data Scientist Should Know. https://datascientistinsights.com/2013/01/29/six-types-of-analyses-every-data-scientist-should-know/

6 BPIC17. Activity Explanation. http://www.win.tue.nl/promforum/discussion/764/activity-explanation

7 KPMG Publication. https://home.kpmg.com/be/en/home/insights/2017/09/process-mining.html

8 http://www.processmining.org/online/fuzzyminer

9 https://msdn.microsoft.com/en-us/library/azure/dn905947.aspx

We would also like to thank our predictive analysis experts, Koen Van Eijk (kvaneijk@kpmg.com) and Toon Declerck (toondeclerk@kpmg.com) for their input.