Process Mining-Driven Optimization of a Consumer Loan Approvals Process

The BPIC 2012 Challenge

Arjel D. Bautista, Lalit Wangikar and Syed M. Kumail Akbar

CKM Advisors, 711 Third Avenue, Suite 1806, New York, NY, USA
{abautista,lwangikar,sakbar}@ckmadvisors.com

Abstract. A real life event log of the loan and overdraft approvals process from a bank in Netherlands is analyzed using process mining and other analytical techniques. The log consists of 262,200 events and 13,087 cases. We used a combination of traditional spreadsheet-based approaches, process-mining capabilities available in Disco and exploratory analytics using Classification and Regression Tree (CART), we examined the data in great detail and at multiple levels of granularity. In this report, we present our findings on how we developed a deep understanding of the process using the event log data, assessed potential areas of efficiency improvement within the institution's operations and identified opportunities to use knowledge gathered during process execution to make predictions about likely eventual outcome of a loan application. We also discuss unique challenges of working with such data, and opportunities for enhancing the impact of such analyses by incorporating additional data elements that should be available internally to the Bank.

1 Introduction

As the role of Big Data becomes increasingly prevalent in this information-driven era [1 – 5], businesses the world over are constantly searching for innovative ways to take advantage of these potentially valuable resources. The 2012 Business Processing Intelligence Challenge (BPIC 2012) is an exercise in analyzing one such set of real-world data using a combination of commercial, proprietary, and open-source tools, and combining these with creative insights to better understand the role of process mining in the modern workplace.

1.1 Approach and Scope

The situation depicted in BPIC 2012 focuses on the loan and overdraft approvals process of a real-world financial institution in the Netherlands. In our analysis of this information, we sought to understand the underlying business processes in great detail and at multiple levels of granularity. We also sought to identify any opportunities for improving efficiency and effectiveness of the overall process. Specifically, we attempted to investigate following areas in detail:

- Develop thorough understanding of the data
- Develop a detailed understanding of the underlying process
- Understand critical activities and decision points
- Understand and map life cycle of a loan application from start to eventual disposition as approved, declined or cancelled
- Identify any resource level differences in performance one can discern based on available data
- Identify opportunities for "process interventions": places in the process where one could change the effort investment from the bank's resources based on likelihood of success

In doing so, we combined the use of dedicated, state-of-the-art process mining technologies with traditional spreadsheet modeling techniques to identify crucial steps and discover important correlations in the data.

As new comers to the discipline of Process Mining, the CKM Advisors team wanted to use this opportunity to put to practice our learning to-date in this discipline. We also attempted to combine Process Mining tools with traditional analytical methods to build a more complete picture. We are certain that with experience, our approach will become more refined and more driven by methods being developed specifically for process mining.

Our attempt was to be as broad in our analysis as possible and delve deep where we could. While we have done detailed analysis in a few areas, we have not covered all possible areas of process mining in our analysis. Any area that we have not covered (for example, Social Network Analysis) is solely driven by our own comfort and familiarity with the subject matter, and not a limitation of the data.

2 Materials and Methods

2.1 Developing Thorough Understanding of the Data

The data captures process events for 13,087 loan / overdraft applications over a roughly six month period from October 2011 to March 2012. The event log is comprised of a total of 262,200 events within these 13,087 cases, starting with a customer submitting an application and ending with eventual conclusion of that application into an Approval, Cancellation or Rejection (Declined). Each case contains a single case level attribute, AMOUNT_REQ, which indicates the amount requested by the applicant. For each event, the extract shows the type of event, life

cycle stage (Schedule, Start, Complete), a resource indicator and the time of event completion.

The events themselves describe steps along the approvals process and are classified into three major types. Table 1 below shows the event types and our understanding of what the events mean.

Table 1. Names and Descriptions of Events

Type	Event Description
"A_" Application Events	Refers to states of the application itself. It appears that the customer initiations an application. Bank resources, then, follow up to complete the application where needed and also facilitate decisions on applications. Initial application submission: – A_SUBMITTED / A_PARTLYSUBMITTED Application pre-accepted but requires additional information: – A_PREACCEPTED Application accepted and pending screen for completeness: – A_ACCEPTED Application finalized after passing screen for completeness: – A_FINALIZED End state of successful (approved) applications: – A_APPROVED / A_REGISTERED / A_ACTIVATED End states of unsuccessful applications: – A_CANCELLED – A_DECLINED
"O_" Offer Events	Refers to states of an offer communicated to the customer: Applicant selected to receive offer: – O_SELECTED Offer prepared and transmitted to applicant: – O_PREPARED / O_SENT Offer response received from applicant: – O_SENT BACK End state of successful offer: – O_ACCEPTED End states of unsuccessful offers: – O_CANCELLED – O_DECLINED
"W_" Work item Events	Refers to states of work items that occur during the approval process. These events capture most of the manual effort exerted by Bank's resources during the application approval process. The events describe efforts during various stages of the application process.

Following up on incomplete initial submissions:
- W_Afhandelen leads

Completing pre-accepted applications:
- W_Completeren aanvraag

Follow up after transmitting offers to qualified applicants:
- W_Nabellen offertes

Assessing the application:
- W_Valideren aanvraag

Seeking additional information during assessment phase:
- W_Nabellen incomplete dossiers

Investigating suspect fraud cases:
- W_Beoordelen fraude

Modifying approved contracts:
- W_Wijzigen contractgegevens

Work items that take place within the approvals process (denoted by "W_" in the event log) are themselves associated with three transitions, each of which occur at distinct stages of the item's life cycle (Table 2):

Table 2. Names and Descriptions of Transitions in the Work Item Life Cycle

Transition	Description
SCHEDULE	Indicates a work item has been scheduled to occur in the future
START	Indicates the opening / commencement of a work item
COMPLETE	Indicates the closing / conclusion of a work item

By itself, the event log is an overwhelming, complicated mass of information from which it is extremely difficult to draw logical conclusions. Therefore, as previous researchers have noted [6,7], it is necessary to subject the log to a fair amount of preprocessing in order to reduce its overall complexity, make visual connections between the steps contained within, and aid in analyzing and optimizing the important business concepts at hand. We were provided a rigorously pre-processed event log in a format that could be analyzed in process mining tools quiet readily. However, we further processed this data to build tailored extracts for various analytical purposes.

2.2 Tools Used for Analysis

For this study we employed a combination of dedicated process mining tools and traditional spreadsheet-based analysis.

- ***Disco***: We procured an evaluation version of Disco (Version 1.0.0; Fluxicon, Eindhoven, The Netherlands) and loaded into it a project set created specifically for the BPIC 2012 exercise from the original XES / MXML files. This tool was especially helpful for the preprocessing and exportation of data into formats suitable for Microsoft Excel analysis (see below). Also of great value was the Disco process map generator (Figure 1), which greatly facilitated visualization of typical process flows and exceptions.

  We also found significant utility in Disco's built-in filtering algorithm, which allows us to include or exclude cases based on the appearance of one or more properties (Figure 2). Specifically, we used this tool to classify cases according to their endpoint behavior (explained in Section 2.3 below) and to examine outlier cases, as identified through the process map generator and in subsequent analyses.

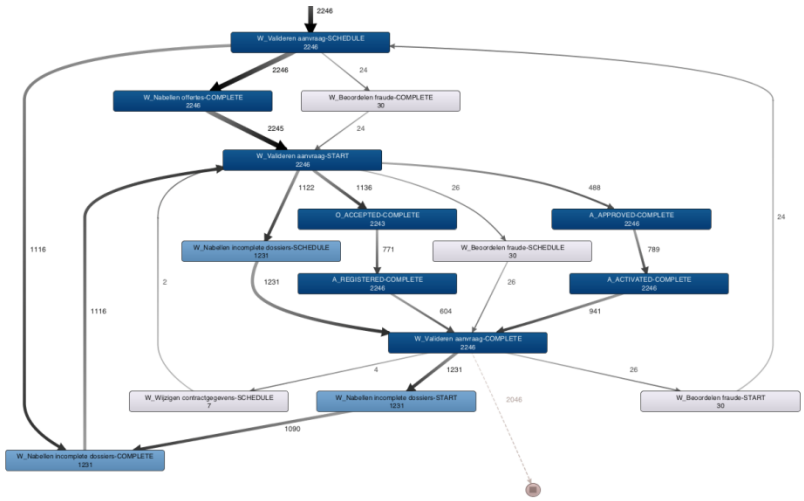

Fig. 1. Fluxicon Disco Process Map Generator

- ***Microsoft Excel***: We used Microsoft Excel (Microsoft Office 2010; Microsoft Corporation, Redmond, WA, USA) to foster deeper exploration into the preprocessed data. In many cases, we used Excel alongside Disco, which helped us visualize, rationalize and refine observations in real time. Excel was especially helpful for performing basic and advanced mathematical functions and data sorting, two capabilities notably absent from the Disco application.
- ***CART Implementation from Salford Systems***: We used evaluation version of the CART implementation from Salford Systems (www.salfordsystems.com) for conducting preliminary segmentation analysis of the loan applications to assess opportunities for prioritizing work effort.

3 Understanding the Process in Detail

3.1 Simplifying the Event Log

Upon obtaining the BPIC 2012 event log, we first attempted to reduce its overall complexity by identifying and removing redundant events. For the purposes of this analysis, an event is considered *redundant* if it occurs concurrently with or subsequently after another event, such that the time between the two events is minimal (a few seconds at best) with respect to the time frame of the case as a whole.

Initial analysis on *Disco* of the raw, unfiltered data revealed a total of 4,366 event order variants among the 13,087 cases represented. We surmised that removal of even one sequence of redundant events could result in a significant reduction in the number of variants, as depicted below (Figure 3). This potential simplification is compounded further when the number of removed variants is multiplied by others occurring downstream of the initial event.

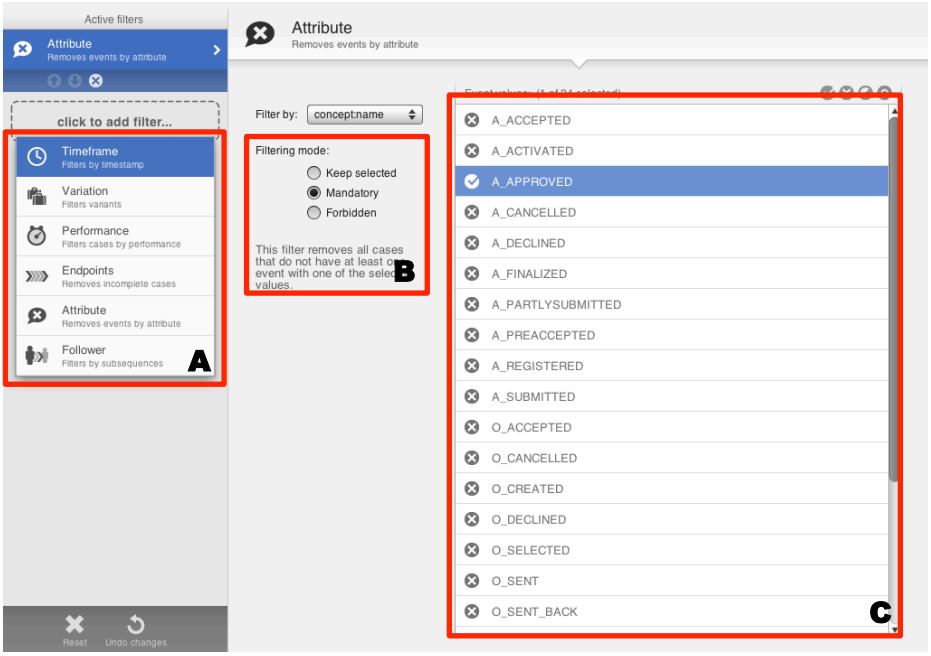

Fig. 2. Fluxicon Case Filter.
(A) Filter selector – Allows user to define a filter based on timeframe, variation, performance, endpoints, attribute, or follower (order); (B) Filtering mode (C) Property selector – Defines the properties upon which the filter is constructed.

With this in mind, we identified six potential redundancies for removal (Table 3):

Table 3. Potential Redundant Events in the Process Event Log

Redundant Events	Occurrence
A_PARTLYSUBMITTED	Immediately after A_SUBMITTED in all 13,087 cases
O_SELECTED O_CREATED	Both in quick succession prior to O_SENT for the 5,015 cases selected to receive offers. – In certain cases, O_CANCELLED (974 instances), A_FINALIZED (2,907 instances) or W_Nabellen offertes-SCHEDULE (1 instance) occur between O_SELECTED and O_CREATED in the offer creation process. All these occur within a few seconds of each other and we believe different sequences represent variations in how employees mark the events in the workflow. With this, removing this redundancy will not impact the overall process understanding.
O_ACCEPTED A_REGISTERED A_ACTIVATED	All three occur, in random order, with A_APPROVED for the 2,246 successful applications. – In certain cases, O_ACCEPTED is also interspersed among these events.

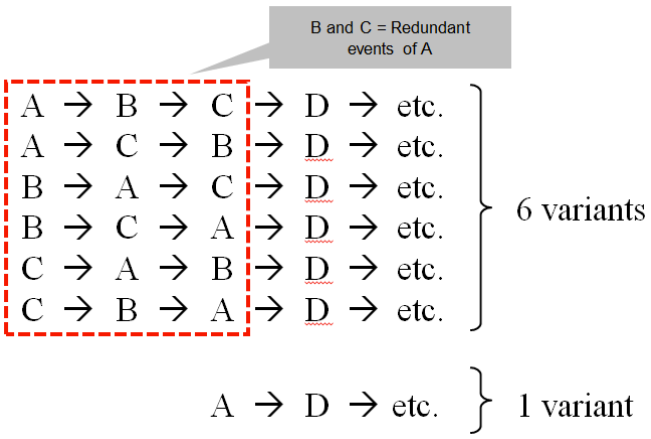

Fig. 3. Removal of Redundant Events

Figure 3 illustrates this approach pictorially. Here, events B and C are redundancies of A that occur in various permutations and proceed immediately to D and downstream events.

Additionally, we eliminated two O-type events – O_CANCELLED and O_DECLINED – which occur simultaneously with A_CANCELLED and A_DECLINED, respectively. Work item (W-type) events were not considered for removal, as their transition phases are crucial for calculating work time spent per case. With the redundant events removed from the event log, the number of variants was reduced to 3,346 – an improvement from the unfiltered data set of nearly 25%. Such event consolidation can aid in simplifying the process data and facilitating quicker analysis. The variant complexity could be further reduced by interviewing process experts at the bank to help further consolidate events that occur together and sequencing variations are not critical for business analysis.

3.2 Determining Standard Case Flow

We next sought to determine the standard case flow for a successful application, to which all other cases could then be compared. We did this by loading the simplified, pre-rendered project into Disco and filtering all cases for the attribute A_APPROVED. We then set both the activities and paths thresholds to the most rigorous level (0%), which resulted in an idealized depiction of the path from initial submission to loan approval (Figure 4).

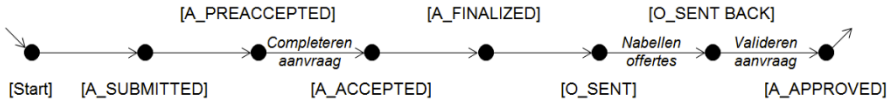

Fig. 4. Standardized Case Flow for Approved Applications

3.3 Understanding Eventual Outcomes for Each Application

Before launching into a more detailed review of the data, we found it necessary to define endpoint outcomes for all 13,087 applications. Using the standardized case flow defined in Figure 4, we determined that all applications are subject to one of four *fates* at each stage of the approvals process:

- ***Advancement to next stage, still under process***: The application proceeds to the next stage of the approvals process.
- ***Approved***: Applications that are approved and where customer has accepted the bank's offer are considered a success and are tagged as Approved, with the end point being depicted by the event A_APPROVED.
- ***Cancellation***: The application is cancelled by the Bank (presumably based on set rules) or at the request of the customer (customer did not like the offer or changed her/his mind). Cancelled applications have a final endpoint of A_CANCELLED.

- ***Denial***: The applicant, after having been subject to review, is deemed unfit to receive the requested loan or overdraft. Denied applications have a final endpoint of A_DECLINED.

We leveraged Disco's filtering algorithm to define a set of twelve possible endpoint behaviors, as listed below (Table 4). An additional 399 cases were classified *unresolved* as they were in progress at the time the data was collected (i.e., did not contain endpoints of A_DECLINED, A_CANCELLED or A_APPROVED).

Table 4. Possible End Points In the Application Approval Process

Endpoint Status	Mandatory Attributes	Forbidden Attributes
0 – Declined Instantly	– A_SUBMITTED – A_DECLINED	– All others
1A – Declined after *W_Afhandelen leads*	– A_SUBMITTED – A_DECLINED	– A_PREACCEPTED
1B – Cancelled after *W_Afhandelen leads*	– A_SUBMITTED – A_CANCELLED	– A_PREACCEPTED
2A – Declined after *W_Completeren aanvraag*	– A_PREACCEPTED – A_DECLINED	– A_ACCEPTED
2B – Cancelled after *W_Completeren aanvraag*	– A_PREACCEPTED – A_CANCELLED	– A_ACCEPTED
3A – Passed initial screen, declined before application was finalized	– A_ACCEPTED – A_DECLINED	– A_FINALIZED
3B – Passed initial screen, cancelled before application was finalized	– A_ACCEPTED – A_CANCELLED	– A_FINALIZED
4A – Declined after customer did not respond to sent offer	– O_SENT – A_DECLINED	– O_SENT BACK
4B – Cancelled after customer did not respond to sent offer	– O_SENT – A_CANCELLED	– O_SENT BACK
5A – Declined after application was assessed	– O_SENT BACK – A_DECLINED	– A_APPROVED
5B – Cancelled after application was assessed	– O_SENT BACK – A_CANCELLED	– A_APPROVED

6 – Loan / overdraft approved (application successful)	– A_APPROVED	– A_DECLINED – A_CANCELLED

Figure 5 below shows a high-level process flow and also marks how the 13,087 cases are disposed at each of the key process steps. Figure 6 shows distribution of the 13,087 cases by their end status. This analysis provides us useful insights on overall business impact of this process (what % applications are declined instantly, cancelled, approved, eventually declined etc.) as well as overall case flow through critical process steps.

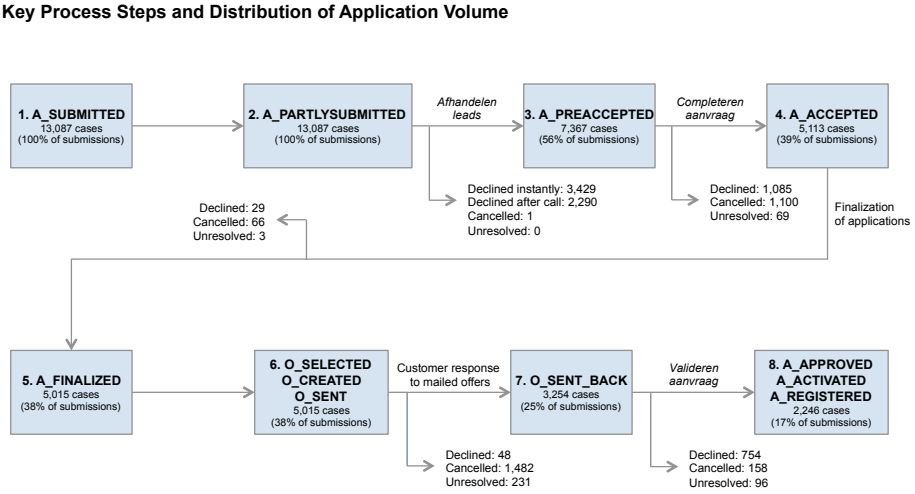

Fig. 5. Key Process Steps and Application Volume Flow

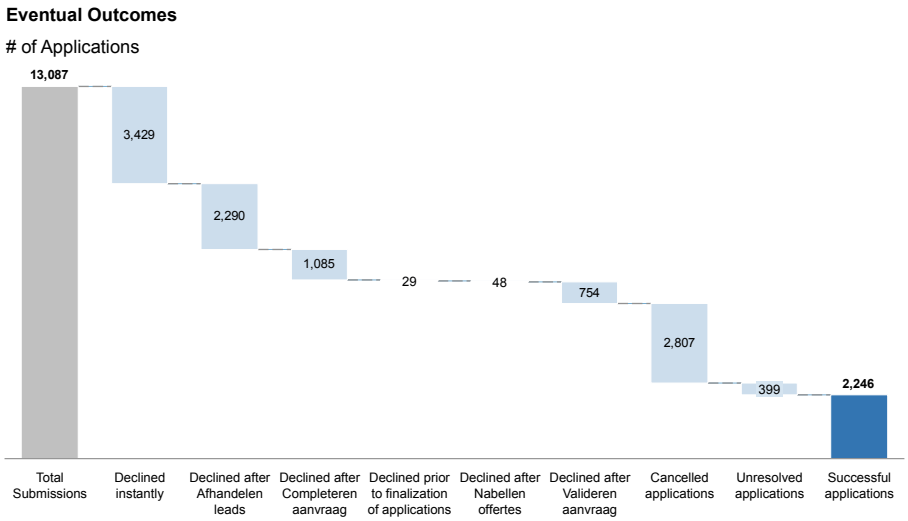

Figure 6: End Status based Distribution of Applications

We observe several baseline performance characteristics from Figures 5 and 6:

- About a quarter of applications are instantly declined (3,429 out of 13,087); indicating tight screening criteria for moving an application beyond the starting point (Figure 5, after process step #2)
- Nearly a quarter of the remaining (2,290 out of 9,658) are declined after initial lead follow up; indicating a continuous risk selection process at play (Figure 5, after process step #2)
- Nearly 23% or 754 of the 3,254 applications that go to validation stage (Figure 5, process step #7) are declined, indicating possibilities for tightening upfront scrutiny at application or offer stage

4 Assessing Process Performance

4.1 Case-Level Analysis

We also evaluated the event log from a macro standpoint, examining the overall fate of each case and segmenting applications according to a wide variety of attributes.

Case Endpoint vs. Overall Duration

In an effort to evaluate how the fate of a particular case changes with overall duration, we prepared a plot of these two variables and overlaid upon it the cumulative amount of work time amassed over the life of these cases. We performed this analysis by excluding 3,429 cases that are instantly declined on initial application submission as no effort is spent on these. We strive to visualize the point at which exertion of additional effort yields minimal or no return in the form of completed (closed)

applications. Figure 7 shows lifecycle view of all application, indexed to the time of starting the application. The figure shows applications grouped by "fates" for the first 50 days since the start of the application. As shown in the figure, within the first seven days from initial submission, applications continue to move forward or are declined. At Day 7, the number of approved cases begins to rise, suggesting this is the minimal number of days required to fulfill the steps in the standard case flow (Figure 4). Approvals continue until approximately Day 23, at which point over 80 percent of all cases that are eventually approved have been closed and registered. There is a significant jump in the number of cancelled applications at Day 30, as those inactive cases receiving no response from the applicant after stalling in the bottleneck stages *Completeren aanvraag* or *Nabellen offertes* are likely cancelled as per Bank's policies.

This raises the interesting question of what is the right duration after which the bank should stop any proactive efforts to convert an application to a loan and whether the bank should treat customers differently based on behavior that might indicate likelihood of eventual approval and acceptance. For example, the bank exerts an additional 380+ person days of effort between Days 23 and 31, only to cancel a majority of pending cases at the conclusion of this period. With additional data about customer profitability or lifetime value and comparative cost of additional effort, one can determine an optimal point on the process where additional effort on cases that have not reached a certain stage in the application process is assessed to be of no positive value.

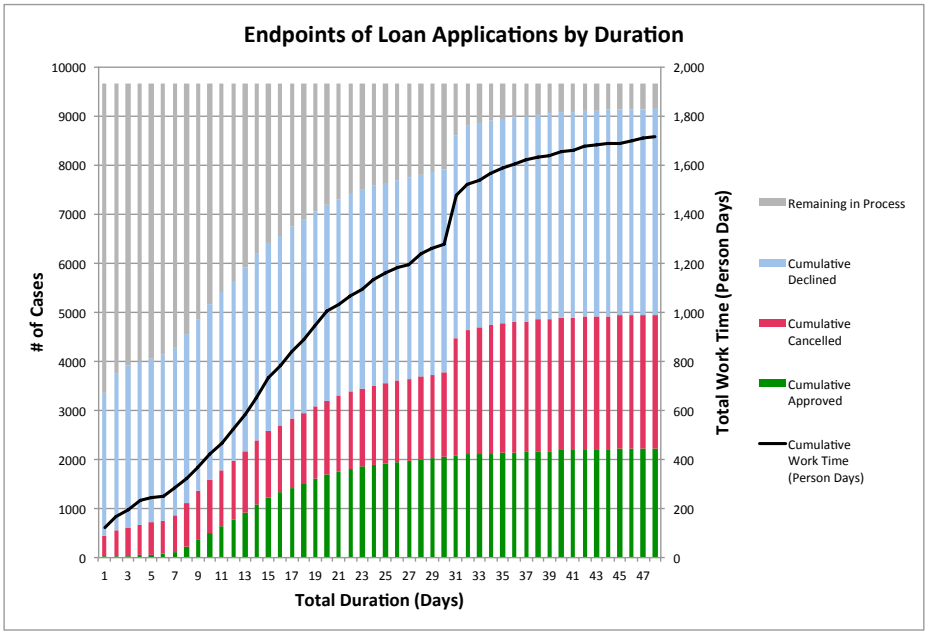

Fig. 7. Distribution of Cases by Eventual Outcome and Duration, with Cumulative Work Effort (Excludes 3,472 Instantly Declined Cases)

Segmenting Cases by Amount Requested

As each case is associated with an amount requested by the applicant, we found it appropriate to arrange them into segments of roughly equal number, sorted by total requested value. We first removed the instantly declined cases by filtering them through Disco, as these are immediately resolved upon submission and do not have any additional effort or steps in the process. The resultant 9,658 cases (which include those in progress) were then split into deciles of 965-966 cases each (Table 5). Each decile was further segmented by classifying the cases according to eventual outcome, and the ensuing trends were examined for correlation of approval percentage with amounts requested (Figure 8).

Table 5. Breakdown of Cases by Decile on Application Amount

Decile	Total Cases	Request Range	Average Work Time per Case (Minutes)	Declined	Cancelled	Approved	In Progress
1	965	0 – 4,000	23.72	68.2%	25.0%	5.0%	1.9%
2	966	4,000 – 5,000	15.92	62.1%	34.8%	1.8%	1.3%
3	966	5,000 – 6,500	80.07	37.8%	17.8%	38.7%	5.7%
4	966	6,500 – 8,000	62.87	42.3%	29.3%	23.9%	4.5%
5	965	8,000 – 10,000	62.62	49.7%	37.6%	10.5%	2.2%
6	966	10,000 – 14,000	90.61	31.9%	22.6%	38.9%	6.6%
7	966	14,000 – 16,000	65.59	32.2%	30.4%	33.5%	3.8%
8	966	16,000 – 23,000	82.42	34.8%	33.2%	27.4%	4.6%
9	966	23,000 – 30,000	90.16	35.7%	32.2%	26.8%	5.3%

10	966	30,000 – 99,999	78.49	40.8%	27.7%	26.0%	5.5%
Total	**9,658**			**4206**	**2807**	**2246**	**399**

We immediately observed the highest approval percentages in Deciles 3 and 6, whose cases contained request ranges of 5,000 – 6,000 and 10,000 – 14,000, respectively. The exact reason for this pattern is unclear; however, we speculate that typical applicants will often choose a "round" number upon which to base their requests (indeed, this is reflected in the three most frequent request values in the data set: 5,000, 10,000 and 15,000). Perhaps certain risk threshold change in the bank's approval process causing a step change in approval percentages.

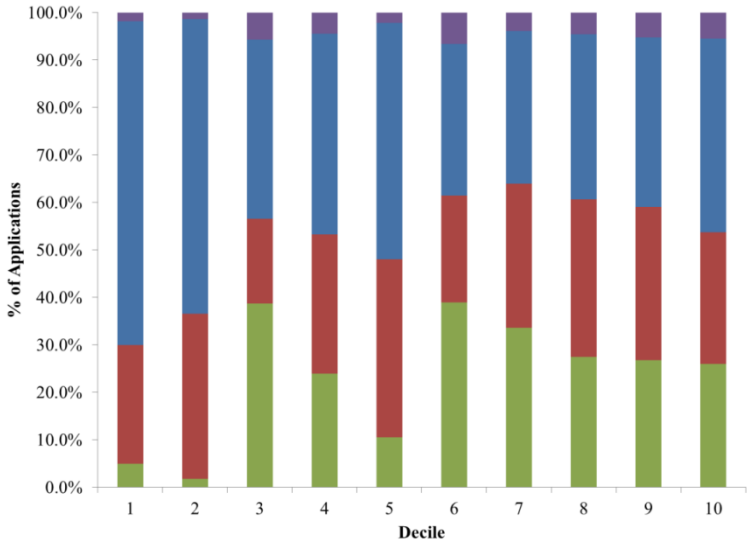

Fig. 8. Endpoints of cases (left axis), as segmented by amounts requested by the applicant. *Green*: Approved cases, *Red*: Cancelled cases, *Blue*: Declined cases, *Violet*: Cases in progress.

4.2 Event-Level Analysis

Calculating Event Duration: Wait vs. Work Time

We sought to gain an intimate understanding of the work activities embedded in the approvals process, specifically those that contribute a significant amount of time or resources toward case resolution. The format of event data made available in this case was not readily amenable to this analysis. We used Excel to manipulate the event level data as provided and applied the following logic to compute work time (presumably actual effort expended by human resources) for each work event.

We defined work time as the duration of events from start to finish (START / COMPLETE transitions, respectively), and wait time as the latency between event scheduling and commencement (SCHEDULE / START), or the time elapsed between two instances of a single activity type as well as between COMPLETE of one event and START of another (**Tables 6-7**).

Table 5. Total Work Time by Event Type (in Minutes)

	Afhandelen Leads	*Beoordelen Fraude*	*Completeren aanvraag*	*Nabellen Offertes*	*Nabellen Incomplete Dossiers*	*Valideren Aanvraag*
Approved Cases	13,659	23	45,909	68,473	89,204	121,099
Cancelled Cases	14,601	2	119,497	94,601	25,633	7,775
Declined Cases	67,560	2,471	63,052	30,870	26,993	29,946

Table 6. Total Wait Time by Event Type (in minutes)

	Afhandelen Leads	*Beoordelen Fraude*	*Completeren aanvraag*	*Nabellen Offertes*	*Nabellen Incomplete Dossiers*	*Valideren Aanvraag*
Approved Cases	198,916	8,456	1,873,537	34,972,224	5,980,887	10,537,938
Cancelled Cases	300,062	28,763	16,582,465	42,630,195	2,006,774	678,105
Declined Cases	986,421	236,115	3,294,367	13,542,054	1,001,354	3,227,252

As shown above, two activities, *Completeren aanvraag* and *Nabellen Offertes*, contribute a significant amount to the total case time represented in the event log. The accumulated wait time attributed to each of these two events can reach as high as 30+ days per case, as the bank presumably makes several attempts to reach the applicant until contact is made. On closer inspection of event logs (Figure 9), we realized that the bank attempts to contact the customer every day, many times a day, until day 30 for completing the application as well as for following up on offers extended to close the application.

Case ID	Activity	Resource	Complete Timestamp
173742	W_Completeren aanvraag-SCHEDULE	10939	10/1/11 8:56 AM
173742	W_Afhandelen leads-COMPLETE	10939	10/1/11 8:56 AM
173742	W_Completeren aanvraag-START		10/1/11 9:43 AM
173742	W_Completeren aanvraag-COMPLETE		10/1/11 9:50 AM
173742	W_Completeren aanvraag-START	11180	10/3/11 6:55 AM
173742	W_Completeren aanvraag-COMPLETE	11180	10/3/11 6:56 AM
173742	W_Completeren aanvraag-START	11201	10/3/11 9:53 AM
173742	W_Completeren aanvraag-COMPLETE	11201	10/3/11 9:55 AM
173742	W_Completeren aanvraag-START	11169	10/4/11 5:30 AM
173742	W_Completeren aanvraag-COMPLETE	11169	10/4/11 5:35 AM
173742	W_Completeren aanvraag-START	11179	10/4/11 7:25 AM
173742	W_Completeren aanvraag-COMPLETE	11179	10/4/11 7:26 AM
173742	W_Completeren aanvraag-START	11122	10/4/11 11:26 AM
173742	W_Completeren aanvraag-COMPLETE	11122	10/4/11 11:27 AM
173742	W_Completeren aanvraag-START	11180	10/5/11 3:31 AM
173742	W_Completeren aanvraag-COMPLETE	11180	10/5/11 3:32 AM
173742	W_Completeren aanvraag-START		10/5/11 7:27 AM
173742	W_Completeren aanvraag-COMPLETE		10/5/11 7:28 AM
173742	W_Completeren aanvraag-START		10/5/11 12:13 PM
173742	W_Completeren aanvraag-COMPLETE		10/5/11 12:14 PM
173742	W_Completeren aanvraag-START		10/5/11 12:51 PM
173742	W_Completeren aanvraag-COMPLETE		10/5/11 12:53 PM
173742	W_Completeren aanvraag-START	11201	10/6/11 3:59 AM
173742	W_Completeren aanvraag-COMPLETE	11201	10/6/11 4:00 AM
173742	W_Completeren aanvraag-START	11180	10/6/11 2:53 PM
173742	W_Completeren aanvraag-COMPLETE	11180	10/6/11 2:54 PM
173742	W_Completeren aanvraag-START	11181	10/7/11 12:33 PM
173742	W_Completeren aanvraag-COMPLETE	11181	10/7/11 12:34 PM
173742	W_Completeren aanvraag-START	10913	10/8/11 4:49 AM
173742	W_Completeren aanvraag-COMPLETE	10913	10/8/11 4:51 AM
173742	W_Completeren aanvraag-START	11181	10/8/11 5:27 AM
173742	W_Completeren aanvraag-COMPLETE	11181	10/8/11 5:30 AM

Fig. 9. Illustrative Event Log

Initial vs. Follow-Up Activities

The average work time spent performing each event changes whether the bank is conducting it for the first time, or following up on a previous step in a particular case (Figure 10). Some differences in initial and follow-up instances are minimal (such as that for *Valideren aanvraag),* while others are more pronounced (*Beoordelen fraude*). In the case of *Valideren aanvraag*, the bank is likely to be as thorough as possible during the validation process, regardless of how many times it has previously viewed an application. On the other hand, when investigating suspect cases for fraud, the

bank may already have come to a preliminary conclusion regarding the application and is merely using the follow-up instance to justify its decision.

Follow-up instances for those events in which the bank must contact the applicant often have smaller average work times than their initial counterparts, as these activities are those most likely to become trapped in repeating loops, perhaps due to no-responsive customers. One can leverage such event data to understand customer behavior and assess potential usefulness of such behavioral data for work prioritization.

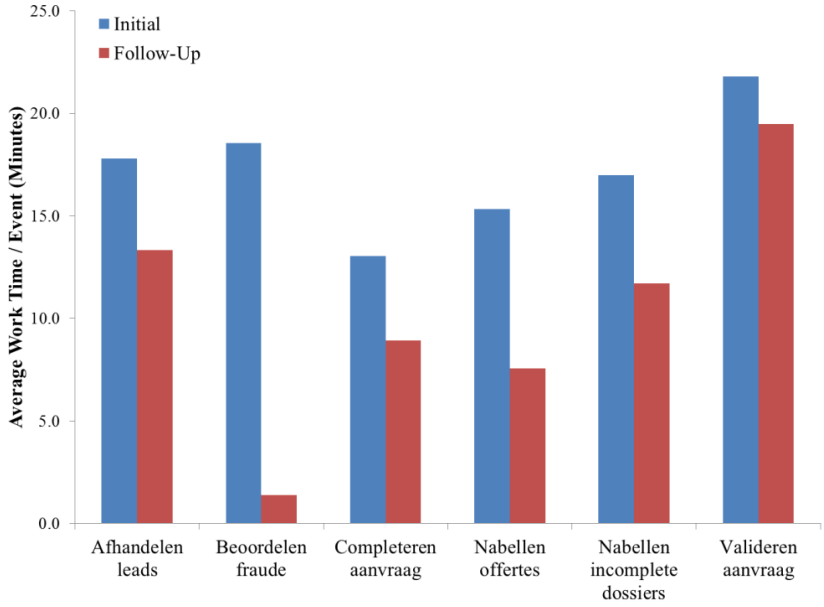

Fig. 10. Comparison of average work times, initial vs. follow-up event instances

4.3 Resource-Level Analysis

Specialist vs. Generalist-Driven Work Activities

We profiled 48 resources that handled at least 100 total events (Figure 11). We excluded resource 112, as this resource does not handle work events outside of scheduling and seems to represent a system). We computed work volume by number of events handled by each of these resources. We observed nine resources that spent more than 50% of their efforts on *Valideren aanvraag*. We also observed a distinct group of resources that mostly performed activities of *Completeren aanvraag, Nabellen offertes* and *Nabellen incomplete dossiers*. It appears application validation is performed by a dedicated team of specialists focused on this work type. While

customer facing activities such as *Completeren aanvraag, Nabellen offertes* and *Nabellen incomplete dossiers* , that might require similar skills are performed by another specialized group.

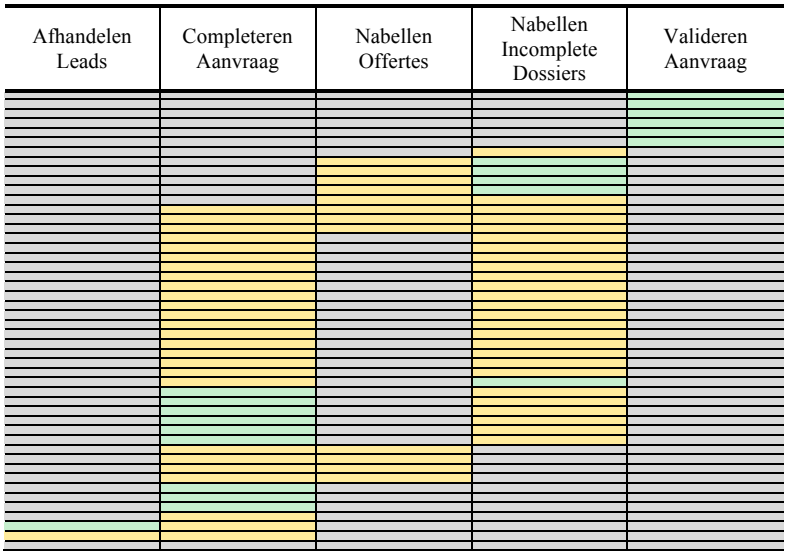

Fig. 11. Graphical representation of total effort exerted by 48 resources across the 5 major work activities. Each row represents one resource. *Green*: >50% of total resource events, *Yellow:* 25-50% of total resource events. *Gray:* <25% of total resource events.

Performance of Specialists and Contributors vs. Minor Players

Next, we examined the performance of those resources identified as specialists (more that 50% of work events of one single type) or contributors (between 25 – 50% of events in one type) and compared them with those who played only minor roles in similar activities. To do this, we took the total work time accumulated in an activity by resources belonging to a particular category and calculated averages based on the total number of work events performed in that category. Two activities, *Nabellen offertes* and *Valideren aanvraag*, did not contain specialists and contributors, respectively, and so these categories were omitted from the comparisons for these activities.

As depicted in Figure 12, area specialists spent less time per event instance than their counterparts, in some cases performing tasks up to 80% more efficiently than minor

players. The performance of contributors is far less consistent, however, exhibiting average work times / case that are both higher (*Afhandelen leads, Nabellen offertes*) and lower (*Completeren aanvraag*, *Nabellen incomplete dossiers*) than those of the minor players. These results suggest that an office of specialists performing single activities may be better suited to handle a larger amount of cases than an army of resources charged with a myriad of tasks.

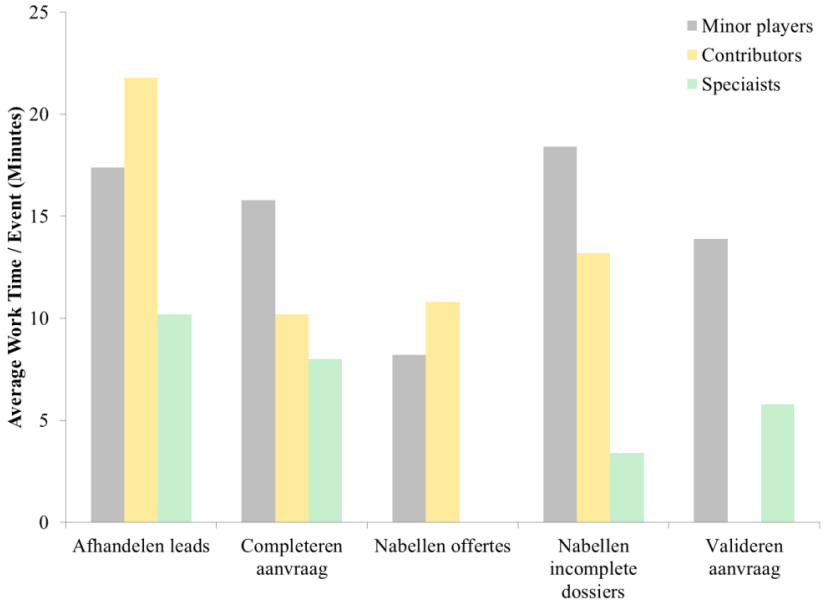

Fig. 12. Comparison of work time spent per event, specialists / contributors vs. minor players

Performance of the Top 5 Resources based on Time Spent

We also compared the performance of the five players who spent the most amount of time in each area (those amassing the highest number of event instances) with the performance of all participating resources (Figure 13). In all areas except *Afhandelen leads*, the leaders exhibited a pattern that mirrored that of the activity specialists (though, notably, not all leaders were specialists themselves). This again points to the possible optimization of the loan approvals process by recasting current generalist resources as single-area specialists.

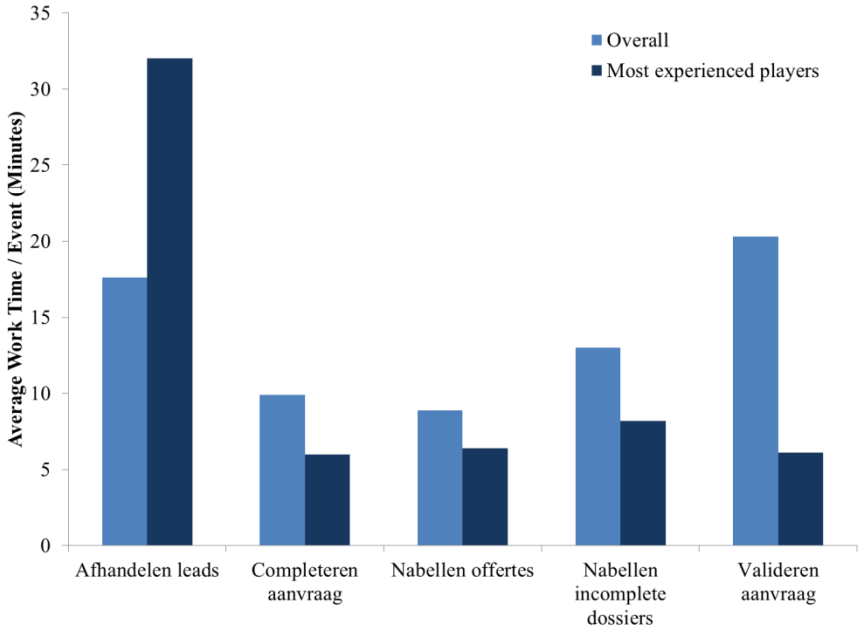

Fig. 13. Comparison of work time spent per event, most experienced players vs. all resources

While additional data, information and analysis would be needed to draw definitive conclusions; we conclude that event level data can provide significant insight in to resource performance.

4.4 Leveraging Behavioral Data for Work Effort Prioritization

One of the objectives of process mining is to identify opportunities for driving process effectiveness: achieving better business outcomes (as opposed increasing efficiency to improve process outcomes; such as turn around time, quality, resource productivity) for same or less effort; in a shorter or equal time period. We framed a specific question that we attempted to answer towards this objective: can we use process event data collected on an application to better prioritize work efforts. Specifically, we set out to understand if this could be done on the fifth day since the application submission.

To perform this analysis, we created a data set at application level for 5,255 cases that lasted more than 4 days and where we know the end outcome for the application. We only captured banks experience on these applications until the end of day 4. Specifically, we calculated the following variables for each such application:

- What stage had the application reached: lead follow up, completed application, offer stage
- How much effort had already gone in

- How many events had already been logged of various different kinds
- Did the application require lead follow up
- Had a complete application been already submitted

We tried to assess if we could find key segments in this population that were highly likely to be approved and accepted OR highly likely to be cancelled or declined. We did this by subjecting the data to segmentation using CART: Classification and Regression Tree technique. Figure 14 below shows partial output of one such segmentation scheme.

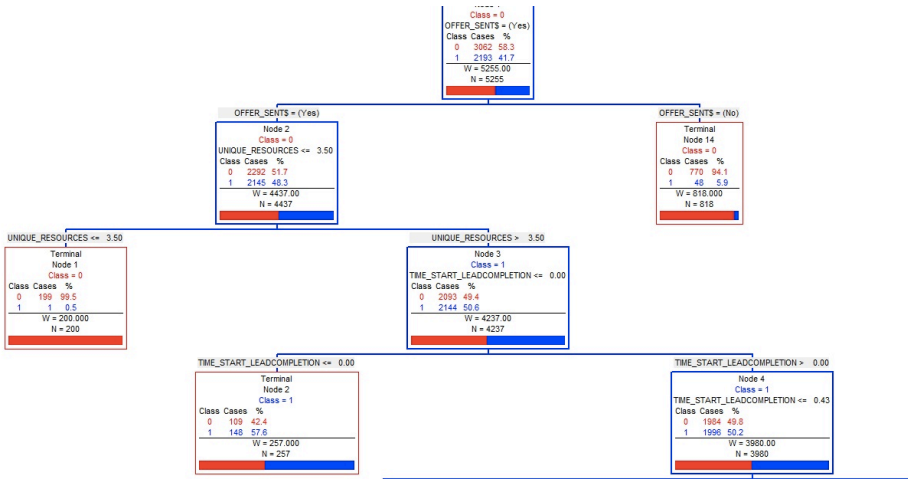

Fig. 14. Partial View of a CART Based Segmentation Tree

The tree above shows two segments with less than 6% approval rates: Terminal Node 1 and Terminal Node 14 consisting of a total of 1,018 cases with only 49 total eventual approvals. Terminal Node 14, consisting of 818 cases, shows applications that were not complete and the Bank could not complete an offer to send to the customers by the end of day 4. Such "slow moving" applications had a less that 6% chance of getting to approval compared to an average of 41.7% for the entire group of 5,255. Terminal Node 1 has applications that are touched by 3 or fewer resources; with 112 being one of them. This might also be an indicator for a "slow moving" application. Such applications have virtually no likelihood of getting to "approved" status in the end.

One could repeat this analysis at different stages in the lifecycle of the application to help with effort prioritization. This preliminary analysis indicates significant potential to reduce effort on cases that might not reach the desired end state. Further analysis with customer demographics, application details, more information on resources who work on such cases will help refine the findings and will suggest specific action steps to improve process effectiveness.

5 Discussion

5.1 Working with Data Challenges: Building a Robust Methodology

Managing Event Complexity in Data

The optimization of the loan approvals process highlighted in this challenge is an exercise in streamlining each step of the end-to-end operation. One notable point that creates challenges in building a streamlined process view using automated process mining tools is amount and complexity of data captured. If such data is not used with accompanying business judgment, one can get lost in apparent complexity (more than 4,000 process variants for a process that has 6 – 7 key steps). We illustrated this point above in our discussion regarding "redundant" events. We recommend dealing with such complexities at the time of analysis, using process knowledge and using good business judgment, by performing additional data pre-processing steps.

It is also critical to spend upfront time studying the event data in great detail to understand all quirks and build ways of addressing these. For example, we compared number of START and COMPLETE transitions that appear for work events in the data set. A simple count of these instances reveals the existence of 1,037 more COMPLETE transitions than START transitions. As the time stamps for these events are unique with respect to others in the same Case ID, they have the potential to greatly confuse the summation of work and wait times for a particular case and for resources within the institution. We denoted these as systems errors and worked with the first COMPLETE following a START as the right one; for a given work event type. In a real project, we would validate our assumption by deeper review of how such instances arise in the system and using that understanding to treat these observations correctly in our analysis.

As described in Section 3.1, the event log would also benefit from consolidation of events that happen concurrently, such as those that occur when successful applications are approved (A_APPROVED, A_REGISTERED and A_ACTIVATED). This would not only decrease overall file size (which becomes important when volume of data grows), but also reduce the complexity of the initial log.

Working with "Outliers"

A significant number of work tasks (W_ events) in the event log show unusually long gaps between START and COMPLETE events, with 191 lasting between 500 – 1,000 minutes, and an additional 102 lasting 1,000 – 5,000 minutes. This might occur if work events are not properly closed at the conclusion of a task (via execution of the appropriate log-off procedure(s)), and continue indefinitely until noticed and corrected. While the most feasible explanation for these atypically long events is the work instance on the supporting process execution software being left open

unintentionally overnight or through the weekend, their occurrence serves to skew trends in the data and can cause discrepancies in analysis. One can suggest system level changes to rectify this going forward. For a project that looks back at data, one needs to develop the right approach to treat such "outliers". In this case as well, the recommended approach would be to leverage detailed process knowledge, systems understanding and business judgment to select the right outlier treatment. Lacking specific information, we did not use any outlier treatment and have used values as observed in the data in conducting the analysis for this paper.

5.2 Assessing Potential Benefits: Illustrative Example for Resource Deployment

Recasting Generalists as Specialists

As mentioned previously and depicted in Figure 11, the tasks involved in the loan approvals process are performed by a mixture of "specialists" and "generalists". Through our analysis we concluded that the bank might benefit from specialization of labor, whereby current resources are reassigned to single posts in order to maximize efficiency. In Table 8 below, we show potential gains to be made through such restructuring. If the bank can improve performance of every one executing a task to the same levels as "specialists", we estimate a substantial overall time saving.

We also evaluated the potential savings associated with downsizing the overall pool of resources assigned to these tasks. Using the average amount of work time for resources handling >100 total events (approximately 16,000 minutes; again, this excludes resource 112 as highlighted previously), we estimate opportunity to reduce the work effort by 35%. (Table 8).

Table 7. Potential time savings associated with conversion of current generalist resources to single-activity specialists. *-None of the resources performing *Nabellen offertes* were identified as specialists; therefore mean efficiency for area contributors was used instead.

	Afhandelen Leads	*Completeren aanvraag*	*Nabellen Offertes*	*Nabellen Incomplete Dossiers*	*Valideren Aanvraag*
Total Work Time (Minutes)	88,905	205,588	133,768	171,668	158,566
Total # of Tasks	5,041	20,830	10,426	19,748	7,819
Mean Specialist Efficiency (Minutes / Event)	10.2	8.0	10.8*	3.4	5.8

Total Work Time Under Mean Specialist Efficiency (Minutes)	51,418	166,640	112,600	67,143	45,350
Projected Time Savings (Minutes)	**37,487**	**38,949**	**21,167**	**104,525**	**113,216**

5.3 The Power of Additional Information

Beyond Loan Request Amounts: Additional Case-Level Attributes

In its raw form, the BPIC 2012 event log is a goldmine of information that, once decoded, provides an extremely detailed view of a consumer loan approvals process. However, this information would be greatly strengthened by the addition of a few key data points. As each case carries with it a single lone attribute – the amount requested by the applicant – we have no way of knowing why certain cases are approved while others with identical request amounts and paths are rejected. Therefore it would be useful to know customer demographics, application details, any current or past relationships with the customers, additional details about the resources that execute these processes. With this information in hand, we can build very specific recommendations for changing the process and also more accurately estimate likely benefits of such changes.

Customer Profitability and Operating Costs for the Application Process

A final set of data notably absent from the provided BPIC 2012 log are the overall costs associated with the loan approvals process and value of each of the loan applications to the bank. It would be worthwhile to understand how much it costs to operate each resource, and whether this cost varies based on the activity they perform or the number of work events they participate in. This information would also allow us to calculate an average acquisition cost for each applicant, and subsequently understand the minimum threshold below which it does not make economic sense to approve an incoming loan request.

6 Conclusions

Through comprehensive analysis of the BPIC 2012 event log, we managed to convert a data set containing 262,200 events and 13,087 cases into a clearly interpretable, end-to-end workflow for a loan and overdraft approvals process. We examined the data at multiple levels of granularity, discovering interesting insights at the event, resource, and case levels. Through our work we also uncovered potential improvements at all three levels, including revision of automated processes,

restructuring of key resources, and evaluation of current case handling procedures. Indeed, more extensive work in this area would be greatly aided by the inclusion of additional data points, such as customer information, policies that govern the process, operating costs for the process and eventual customer value.

As part of our analysis, we performed a rudimentary predictive exercise whereby we determined the current status of cases at various days in the approvals process and quantified their chances of approval, cancellation, or denial. This allowed us to estimate the fate of a case based on its performance and tailor the overall process to minimize stalling at traditional case bottlenecks. While preliminary in its nature, this surely opens the door to more elaborate future modeling exercises, perhaps driven by sophisticated computer programs and algorithms.

While we covered several areas of exploration in this exercise, there are others where we did not conduct detailed analysis. The bank would find significant additional benefits from exploring such additional areas, for example, social network analysis.

In conclusion, the procedures highlighted by the 2012 Business Process Intelligence Challenge elaborate the role and importance of process mining in the modern workplace. Steps that were previously elucidated only after years of practice and painstaking observation can now be examined using a sample set of existing data. As the era of Big Data continues its march toward the business world, we foresee process mining as a central player in the charge toward turning questions into solutions and problems into sustainable profit.

Acknowledgements We are grateful to the Bank from Netherlands that made this invaluable data available for study. The authors are grateful to Dr. Anne Rozinat and Dr. Christian Gunther (Fluxicon) for providing us with an evaluation copy of Disco and an accompanying copy of the BPIC 2012 data set. We also thank Dr. Tom Metzger, Dr. Nicholas Hartman, Rolf Thrane and Pierre Buhler (CKM Advisors) for helpful discussions and insights. Our special thanks also to Salford Systems, who make their software available in a demonstration version.

Note: Author names appear in alphabetical order based on first names.

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. Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A.H.: Big Data: The Next Frontier for Innovation, Competition, and Productivity. Technical report, McKinsey Global Institute (2011)
3. Brown, B., Sikes, J.: McKinsey Global Survey Results: Minding your Digital Business. In: McKinsey Quarterly, McKinsey & Company (2012)
4. Adduci, R., Blue, D., Chiarello, G., Chickering, J., Mavroyiannis, D. *et al*: Big Data: Big Opportunities to Create Business Value. Technical report, Information Intelligence Group, EMC Corporation (2011)
5. The Economist Intelligence Unit: The Deciding Factor: Big Data & Decision Making. Technical report, Capgemini (2012)
6. Bose, R.P.J.C., van der Aalst, W.M.P.: Analysis of Patient Treatment Procedures: The BPI Challenge Case Study. In: First International Business Process Intelligence Challenge (2011)
7. Van der Aalst, W.M.P.: Process Mining: Discovery, Conformance and Enhancement of Business Processes. Springer, Heidelberg (2011)