novática
Revista de la Asociación de Técnicos de Informática

Novática, founded in 1975, is the oldest periodical publication amongst those especialized in Information and Communication Technology (ICT) existing today in Spain. It is published by **ATI (*Asociación de Técnicos de Informática*)** which also publishes **REICIS** (*Revista Española de Inovación, Calidad e Ingeniería del Software*).

<http://www.ati.es/novatica/>
<http://www.ati.es/reicis/>

ATI is a founding member of **CEPIS** (*Council of European Professional Informatics Societies*), the Spain´s representative in **IFIP** (*International Federation for Information Processing*), and a member of **CLEI** (Centro Latinoamericano de Estudios en Informática) and **CECUA** (*Confederation of EuropeanComputer User Associations*). It has a collaboration agreement with **ACM** (*Association for Computing Machinery*) as well as with diverse Spanish organisations in the ICT field.

Editorial Office, Advertising and Madrid Office
Plaza de España 6, 2ª planta, 28008 Madrid
Tlfn.914029391; fax.913093685 <novatica@ati.es>
Layout and Comunidad Valenciana Office
Av. del Reino de Valencia 23, 46005 Valencia
Tlfn. 963740173 <novatica_prod@ati.es>
Acounting, Subscriptions and Catalonia Office
Calle Avila 48-50, 3a planta, local 9, 08005 Barcelona
Tlfn.934125235; fax 934127713 <secregen@ati.es>
Andalucía Office <secreand@ati.es>
Galicia Office <secregal@ati.es>
Subscriptions and Sales <novatica.subscripciones@atinet.es>
Advertising Plaza de España 6, 2ª planta, 28008 Madrid
Tlnf.914029391; fax.913093685 <novatica@ati.es>
Legal depósit: B 15.154-1975 -- ISSN: 0211-2124; CODEN NOVAEC
Cover Page: "Mineral, Vegetable, Animal" - Concha Arias Pérez / © ATI
Layout Dising: Fernando Agresta / © ATI 2003

Special English Edition 2013-2014
Annual Selection of Articles

summary

Arjel Bautista, Lalit Wangikar, S.M. Kumail Akbar
CKM Advisors, 711 Third Avenue, Suite 1806, New York, NY, USA

<{abautista,lwangikar,sakbar}@ckmadvisors.com>

Process Mining-Driven Optimization of a Consumer Loan Approvals Process

Abstract: *An event log (262,200 events; 13,087 cases) of the loan and overdraft approvals process from a bank in the Netherlands was analyzed using a number of analytical techniques. Through a combination of spreadsheet-based approaches, process mining capabilities and exploratory analytics, we examined the data in great detail and at multiple levels of granularity. We present our findings on how we developed a deep understanding of the process, assessed potential areas of efficiency improvement and identified opportunities to make knowledge-based predictions about the 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.*

Keywords: *Big Data, Business Process Intelligence, Data Analytics, Process Mining.*

Authors

Arjel Bautista is a consultant at CKM Advisors, involved in the development of innovative process re-engineering and analytical research techniques within the firm. In his projects, Arjel has deployed a combination of state-of-the-art data mining tools and traditional strategic analysis to solve a variety of problems relating to business processes. He has also developed strategies for the analysis of unstructured text and other non-traditional data sources. Arjel holds Masters and Doctorate degrees in Chemistry from Yale University, and a Bachelor's degree in Biochemistry from UC San Diego.

Lalit Wangikar is a Partner at CKM Advisors. As a consultant, Lalit has advised clients primarily in the financial services sector, insurance, and payment services industries. His primary area of expertise is use of Big Data and Analytics for driving business impact across all key business areas such as marketing, risk, operations and compliance. He has worked with clients in North America, UK, Singapore and India. Prior to joining CKM Advisors, Lalit ran Decision Analytics practice for EXL Service / Inductis. Prior to that he worked as a consultant with Deloitte Consulting and Mitchell Madison Group where he advised clients in banking and capital markets verticals.

Syed M. Kumail Akbar is a Consultant at CKM Advisors where he is a member of the Analytics Team and assists in data mining, process mapping and predictive analytics. In the past, he has worked on strategy and operations projects in the financial services industry. Before joining CKM, Syed worked as a research assistant in both the Quantitative Analysis Center and the Physics Department at Wesleyan University. He also co-founded Possibilities Pakistan, a Non-Governmental Organization dedicated to providing access to college counseling for high school students in Pakistan. Syed holds a BA in Physics and Mathematics-Economics from Wesleyan University.

1. Introduction

As the role of Big Data gains prevalence in this information-driven era [1][2][3], businesses the world over are constantly searching for ways to take advantage of these potentially valuable resources. The 2012 Business Processing Intelligence Challenge (BPIC, 2012) is an exercise in analyzing one such data set 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 the following areas in detail:

- Develop a thorough understanding of the data and the underlying process.
- Understand critical activities and decision points.
- Map the lifecycle of a loan application from start to eventual disposition.
- Identify any resource-level differences in performance and opportunities for process interventions.

As newcomers to process mining, we at CKM Advisors wanted to use this opportunity to put into practice our learning 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 increasingly driven by methods developed specifically for process mining.

Our attempt was to be as broad as possible in our analysis 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 areas that we did not cover (for example, social network analysis) are driven solely by our own comfort and familiarity with the subject matter, and not necessarily a limitation of the data.

2. Materials and Methods

2.1. Understanding the Data

The data captures process events for 13,087 loan / overdraft applications over a six month period, between October 2011 and March 2012. The event log is comprised of a total of 262,200 events within these cases, starting with a customer submitting an application and ending with eventual conclusion of that application into an approval, cancellation or rejection (declined). Each application contains a single attribute, AMOUNT_REQ, which indicates the amount requested by the applicant. For each event, the extract shows the type of event, lifecycle stage (Schedule, Start, Complete), a resource indicator and time of completion.

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

By itself, the event log is a complicated mass of information from which it is difficult to draw logical conclusions. Therefore, as other researchers have noted [4][5], it is necessary to subject the log to some degree of preprocessing in order to reduce its overall complexity, make visual connections between the steps contained within, and aid in analyzing and optimizing the business concepts at hand.

> "As other researchers have noted, it is necessary to subject the log to some degree of preprocessing in order to reduce its overall complexity"

Type	Description
"A_" Application Events	Refers to states of the application itself. After a customer initiates an application, bank resources follow up to complete the application where needed and facilitate decisions on applications.
"O_" Offer Events	Refers to states of an offer communicated to the customer.
"W_" Work 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. - *W_Afhandelen leads:* Following up on incomplete initial submissions - *W_Completeren aanvraag:* Completing pre-accepted applications - *W_Nabellen offertes:* Follow up after transmitting offers to qualified applicants - *W_Valideren aanvraag:* Assessing the application - *W_Nabellen incomplete dossiers:* Seeking additional information during assessment phase - *W_Beoordelen fraude:* Investigating suspect fraud cases *W_Wijzigen contractgegevens:* Modifying approved contracts

Table 1. Event Names and Descriptions.

Although we were provided a rigorously pre-processed event log that could be analyzed in process mining tools quiet readily, we processed the data further to build tailored extracts for various analytical purposes.

2.2. Tools Used for Analysis

■ *Disco*: We procured an evaluation version of Disco 1.0.0 (Fluxicon) and used it in the exportation of data into formats suitable for spreadsheet analysis. Disco was especially helpful in facilitating visualization of typical process flows and exceptions.

■ *Microsoft Excel*: We used Excel 2010 (Microsoft) to foster deeper exploration into the preprocessed data. Excel was especially helpful for performing basic and advanced mathematical functions and data sorting, two capabilities notably absent from the Disco application.

■ *CART*: We used an evaluation version of the CART implementation (Salford Systems) 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 most) with respect to the time frame of the case as a whole.

Initial analysis of the raw data in *Disco* 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

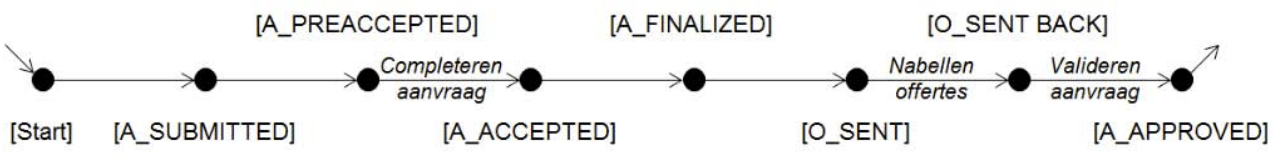

Figure 1. Standardized Case Flow for Approved Applications.

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.
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 interspersed among these events.

Table 2. Potential Redundancies in the Event Log.

significant reduction in the number of variants. This simplification is compounded further when the number of removed variants is multiplied by others occurring downstream of the initial event.

Additionally, we eliminated two O-type events (O_CANCELLED and O_DECLINED) which occur simultaneously with A_CANCELLED and A_DECLINED, respectively. 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 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 consolidate events that occur together and sequencing variations not critical for business analysis.

3.2. Determining Standard Case Flow

We next sought to determine the standard case flow for a successful application, against which all other cases could then be compared. We did this by loading the simplified 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 submission to approval (see **Figure 1**).

3.3. Understanding Application Outcomes

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 (see **Figure 1**), we determined that all applications are subject to one of four *fates* at each stage of the approvals process:

- ***Advancement to next stage***: The application proceeds to the next stage of the process.
- ***Approved***: Applications that are approved and where the customer has accepted the bank's offer are considered a success and are tagged as Approved, with the end point depicted by the event A_APPROVED.
- ***Cancelled***: The application is cancelled by the bank or at the request of the customer. Cancelled applications have a final endpoint of A_CANCELLED.
- ***Denied***: 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 possible endpoint behaviors. 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).

Figure 2 shows a high-level process flow that marks how the cases are disposed at each of the key process steps. This analysis provides us useful insights on the overall business impact of this process as well as overall case flow through critical process steps.

We observe several baseline performance characteristics from **Figure 2**:

- ~26% of applications are instantly declined (3,429 out of 13,087); indicating tight screening criteria for moving an application beyond the starting point.
- ~24% of the remaining (2,290 out of 9,658) are declined after initial lead follow up, indicating a continuous risk selection process at play.
- 754 of the 3,254 applications that go to validation stage (~23%) are declined, indicating possibilities for tightening upfront scrutiny at application or offer stages.

4. Assessing Process Performance

4.1. Case-Level Analysis

4.1.1. 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 excluded 3,429 cases that are instantly declined on initial application submission, as no effort is spent on these. We endeavored to visualize the point at which exertion of additional effort yields minimal or no return in the form of completed (closed) applications.

Figure 3 shows a lifecycle view of all applications, indexed to the time of submission. As shown in the figure, within the first seven days 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 (see **Figure 1**).

Approvals continue until ~Day 23, at which point >80% 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 inactive cases receiving no response from the applicant after stalling in the bottleneck stages *Completeren aanvraag* or *Nabellen offertes* are cancelled, likely according to bank policies.

This raises the interesting question of when the bank should stop any proactive efforts to convert an application to a loan, and whether the bank should treat customers differently based on behaviors that indicate likelihood of eventual approval. 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 on customer profitability or lifetime value and comparative cost of additional effort, one can determine an optimal point

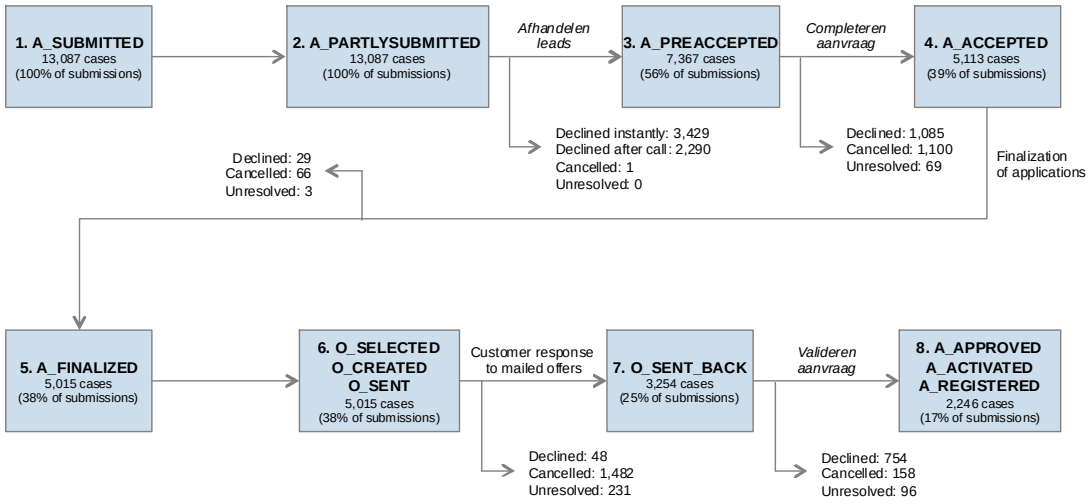

Figure 2. Key Process Steps and Application Volume Flow.

in the process where additional effort on cases that have not reached a certain stage carries no positive value.

4.1.2. 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. 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 (see **Figure 4**).

We immediately observed the highest approval percentages in deciles 3 and 6, whose cases contained request ranges of

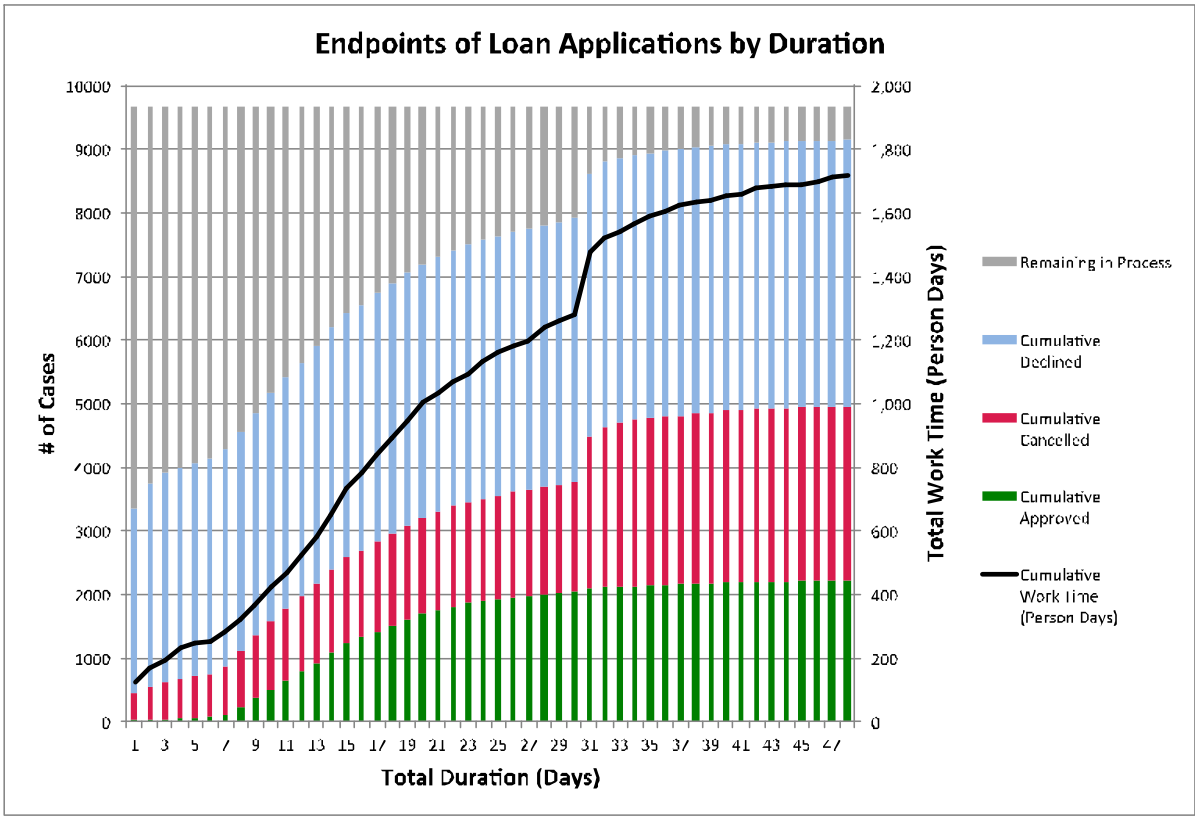

Figure 3. Distribution of Cases by Eventual Outcome and Duration, with Cumulative Work Effort. *Gray:* Remaining In Progress, *Blue:* Cumulative Declined, *Red:* Cumulative Cancelled, *Green:* Cumulative Approved. *Excludes 3,472 Instantly Declined Cases.*

> "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"

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 a certain risk threshold change in the bank's approval process causes a step change in approval percentages.

4.2. Event-Level Analysis

4.2.1. Calculating Event Duration

We sought to gain a detailed understanding of the work activities embedded in the approvals process, specifically those that contribute a significant amount of time or resources toward resolution. The format of 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 defined work time (presumably actual effort expended by human resources) for each event as the duration from start to finish (START / COMPLETE transitions, respectively). In contrast, wait time was defined 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:

As shown in **Table 3**, 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 numerous attempts to reach the applicant until contact is made.

On closer inspection of the data, we realized that the bank attempts to contact the customer multiple times per day until Day 30 in order to complete the application, as well as to follow up on offers that have been extended but not yet replied to.

4.2.2. 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 (see **Figure 5**).

Some differences in initial and follow-up instances are minimal (*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 non-responsive customers. One can leverage such event data to understand customer behavior

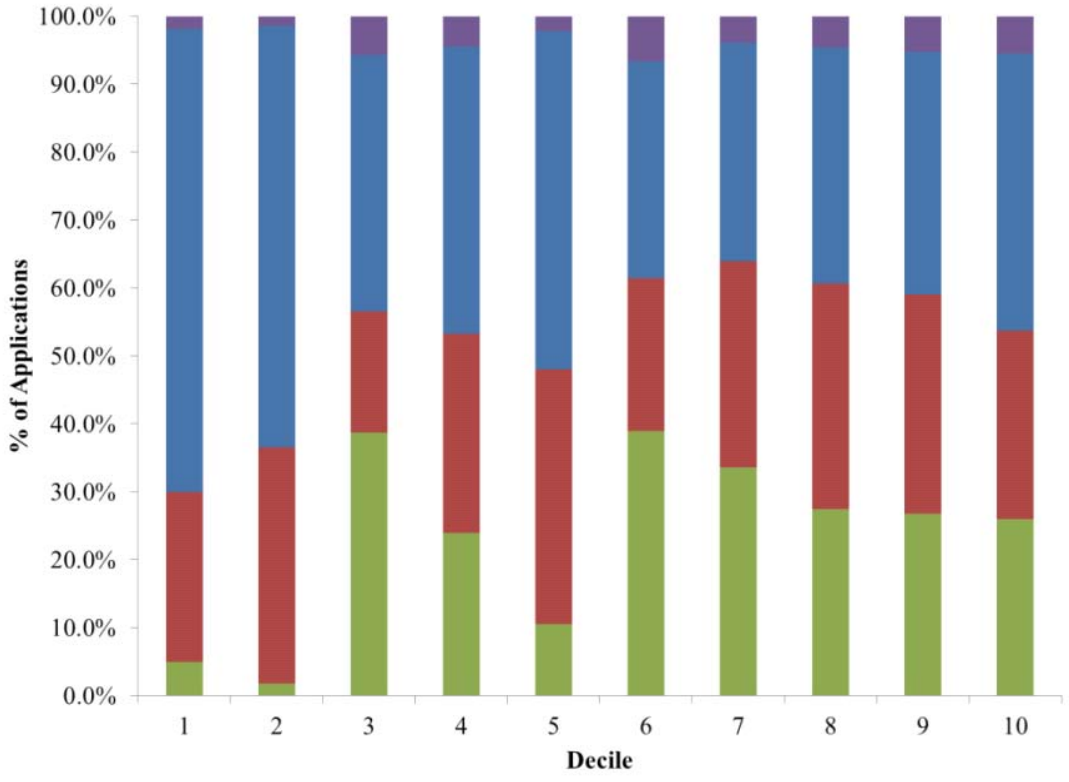

Figure 4. Endpoints of Cases (Left Axis), Segmented by Amounts Requested by the Applicant. *Green:* Approved, *Red:* Cancelled, *Blue:* Declined, *Violet:* In Progress.

> "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"

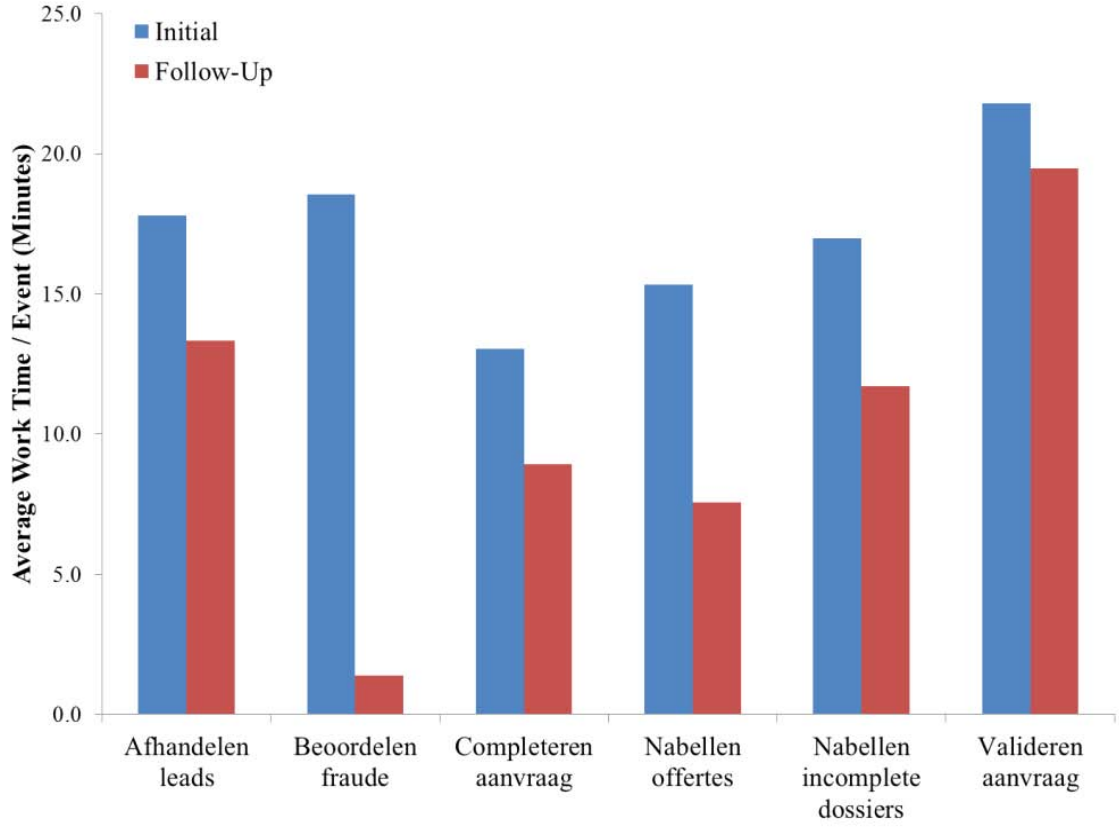

Figure 5. Comparison of Average Work Times, Initial vs. Follow-Up Event Instances.

and assess potential usefulness of such data for work prioritization.

4.3. Resource-Level Analysis

4.3.1. Specialist vs. Generalist-Driven Work Activities

We profiled 48 resources that handled at least 100 total events (excluding resource 112, as this resource does not handle work events outside of scheduling) and computed work volume by number of events handled by each. We observed nine resources that spent >50% of their effort on *Valideren aanvraag*, and a distinct group that mostly performed activities of *Completeren aanvraag, Nabellen offertes* and *Nabellen incomplete dossiers*. It appears 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* might require similar skills that are performed by another specialized group.

We next examined the performance of resources identified as specialists (>50% of work events of one single type) or contributors (25-50%) 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 6**, 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.

4.4. Leveraging Behavioral Data for Work Effort Prioritization

One of the objectives of process mining is to identify opportunities for driving process effectiveness; that is, achieving better business outcomes for the same or less effort in a shorter or equal time period. In particular, we sought to 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 was submitted.

To do this, we created an application-level data set for 5,255 cases that lasted >4 days and where the end outcome is known. For these applications, we captured all events from submission until the end of day 4 and used them to calculate the following:

- What stage the application had reached, and if it had been completed.
- How much effort had already gone into the application.
- How many events of each kind had already been logged.
- If the application required lead follow up.

We attempted to find key segments in this

	Afhandelen Leads	*Beoordelen Fraude*	*Completeren aanvraag*	*Nabellen Offertes*	*Nabellen Incomplete Dossiers*	*Valideren Aanvraag*
Work Time:						
Approved	13,659	23	45,909	68,473	89,204	121,099
Cancelled	14,601	2	119,497	94,601	25,633	7,775
Declined	67,560	2,471	63,052	30,870	26,993	29,946
Wait Time:						
Approved	198,916	8,456	1,873,537	34,972,224	5,980,887	10,537,938
Cancelled	300,062	28,763	16,582,465	42,630,195	2,006,774	678,105
Declined	986,421	236,115	3,294,367	13,542,054	1,001,354	3,227,252

Table 4. Potential Time Savings Associated with Conversion of Current Generalists to Single-Activity Specialists. (*) None of the resources performing *Nabellen offertes* were identified as specialists; therefore mean efficiency for area contributors was used instead.

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 the Classification and Regression Tree (CART) technique (see Figure 7).

The partial tree above shows two segments with <6% approval rates: Terminal Nodes 1 and 14, consisting of a total of 1,018 cases with only 49 eventual approvals. Node 14, consisting of 818 cases, shows incomplete applications where the bank could not prepare an offer for the customers by the end of Day 4. Such "slow-moving" applications had a <6% chance of being approved, compared to an average of 42% for the entire group of 5,255. Node 1 has applications that are touched by 3 or fewer resources; with 112 being one of them. This might be another indicator for a slow-moving application. Such applications have virtually no likelihood of being approved 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, and additional information on resources who work on such cases will help refine the findings and suggest specific action steps to improve process effectiveness.

5. Discussion

5.1. Working with Data Challenges

5.1.1. Managing Event Complexity

The optimization of the loan approvals process 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 with automated process mining tools is the amount and complexity of data captured. If such data is not used with accompanying business judgment, one can get lost in apparent complexity (>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 good business judgment, by performing additional data pre-processing steps.

It is also critical to scrutinize event data up front to understand all quirks and to build ways of addressing these. For example, a comparison of the number of START and COMPLETE transitions for W-type events in the data set 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 "correct" 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 the overall file size (which becomes important as the volume of data grows), but also reduce the complexity of the initial log.

5.2. Potential Benefits of Resource

5.2.1. Re-Deployment Recasting Generalists as Specialists

As mentioned previously, 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 4, we show potential gains to be made through such restructuring. If the bank can improve performance of everyone 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 excluding resource 112), we estimate opportunity to reduce the work effort by 35%:

5.3. The Power of Additional Information Additional

5.3.1. Case-Level Attributes

In its raw form, the BPIC 2012 event log is a gold mine of information that, once decoded, provides a 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 attribute – the amount requested

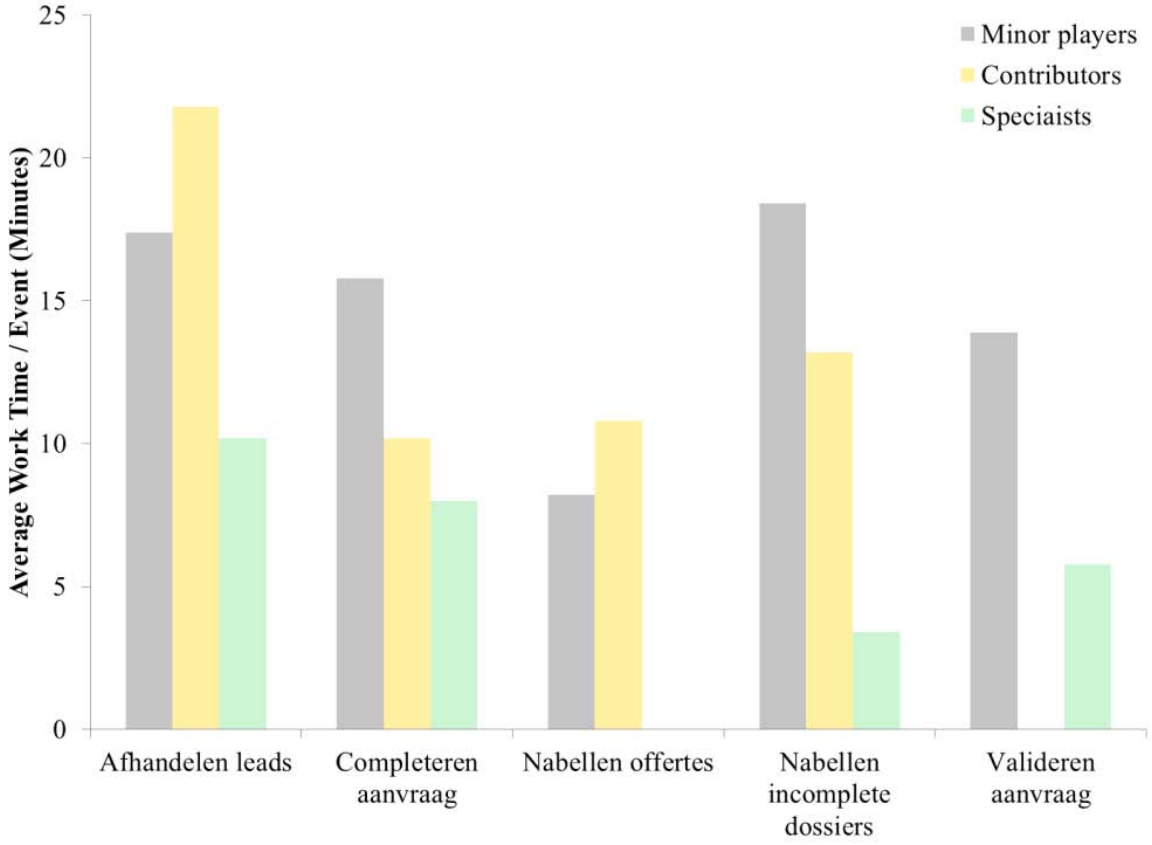

Figure 6. Work Time per Event, Specialists / Contributors vs. Minor Players.

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, current or past relationships with the customers, and additional details about the resources that execute these processes. With this information, we can build specific recommendations for changing the process and more accurately estimate likely benefits of such changes.

5.3.2. Customer Profitability and Operating Costs

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 loan application 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 activities they perform or the number of 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 converted a fairly complex data set into a clearly interpretable, end-to-end workflow for a loan and overdraft approvals process. We examined the data at multiple levels of granularity, uncovering interesting insights at all levels. Through our work we uncovered potential improvements in a number of areas, including revision of automated processes, restructuring of key resources, and evaluation of current case handling procedures. Indeed, future analysis would be greatly aided by the inclusion of additional data, such as customer information, governing policies, operating costs and relative 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 opens the door to more elaborate future modeling exercises, perhaps driven by sophisticated computer algorithms.

While we covered several areas 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 BPIC 2012 elaborate the role and importance of process mining in the modern workplace. Steps that were previously elucidated only after years of practice and 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 financial institution that made this data available for study. Special thanks to Fluxicon for providing us with an evaluation copy of Disco and an accompanying copy of the BPIC 2012 data set. We also thank Tom Metzger, Nicholas Hartman, Rolf Thrane and Pierre Buhler for helpful discussions and insights. Our thanks also to Salford Systems, who made their software available in a demonstration version.

> "It is also critical to scrutinize event data up front to understand all quirks and to build ways of addressing these"

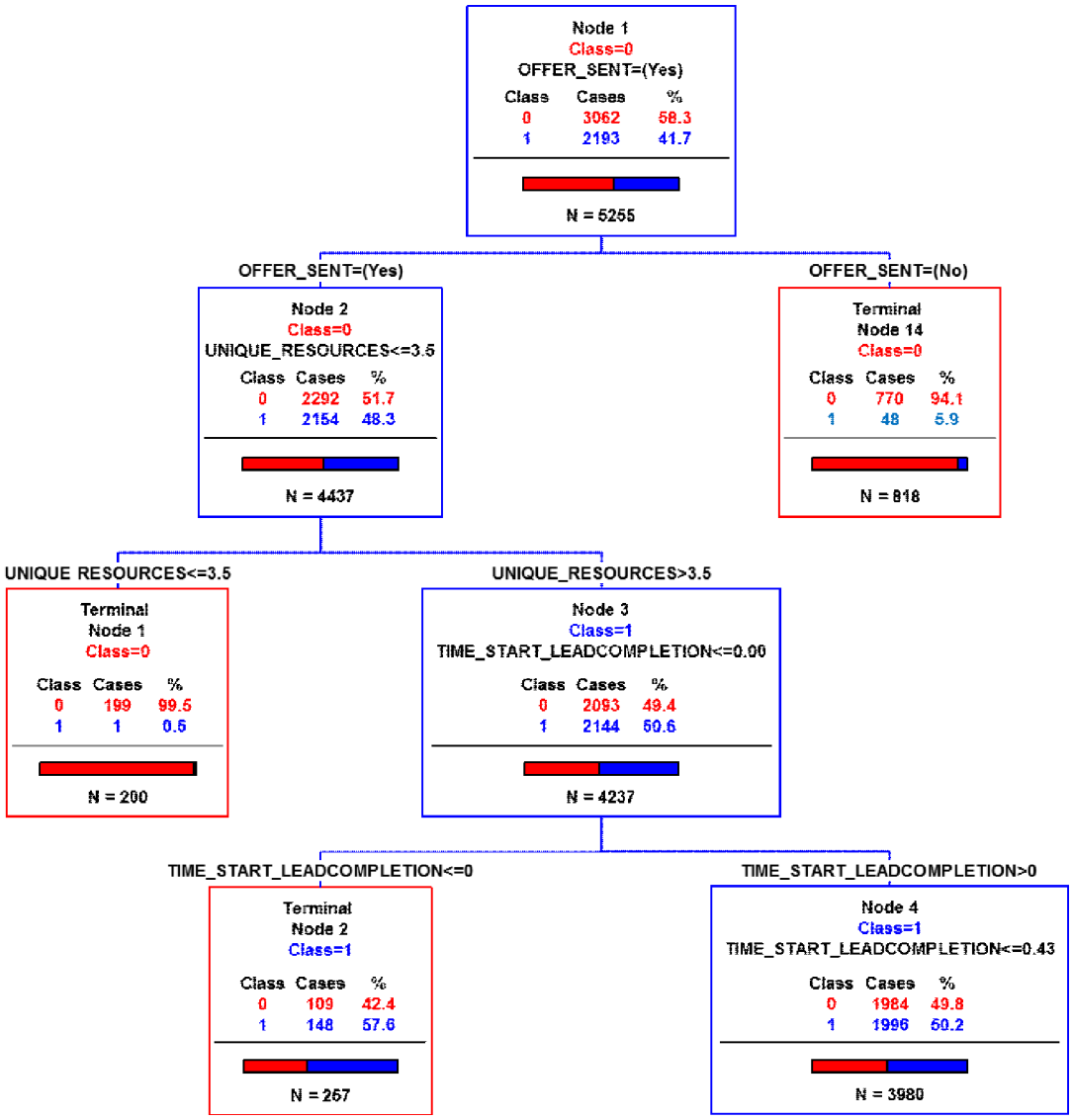

Figure 7. Partial View, CART Segmentation Tree.

References

[1] W. Van der Aalst, A. Adriansyah, A.K. Alves de Medeiros, F. Arcieri, T. Baier *et al.* Process Mining Manifesto. *Business Process Management Workshops 2011, Lecture Notes in Business Information Processing, vol. 99*. Springer-Verlag, 2011.

[2] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, A. Byers. Big Data: The Next Frontier for Innovation, Competition, and Productivity. McKinsey Global Institute, 2011. <http://www.mckinsey.com/insights/business_ technology/big_data_the_next_frontier_for_innovation>.

[3] R. Adduci, D. Blue, G. Chiarello, J. Chickering, D. Mavroyiannis *et al.* *Big Data: Big Opportunities to Create Business Value*. Technical report, Information Intelligence Group, EMC Corporation, 2011.

[4] R.P.J.C. Bose, W.M.P. van der Aalst. Analysis of Patient Treatment Procedures: The BPI Challenge Case Study. *First International Business Process Intelligence Challenge*, 2011. <http://bpmcenter.org/wp-content/uploads/reports/2011/BPM-11-18.pdf>.

[5] W.M.P. van der Aalst. *Process Mining: Discovery, Conformance and Enhancement of Business Processes*. Springer, 2011. ISBN-10: 3642193447.