Advanced Process Mining - 12.25111

Practical Assignment; Part II

Deadline: 15.07.2022, 23:59 CET

T. Brockhoff M.Sc., T. Huang M.Sc.
V. Peeva M.Sc., Dr. ir. S.J. van Zelst

Chair of Process and Data Science
RWTH Aachen University

Introduction

In this assignment, you will deal with a real-life event log provided by a financial institute for its *loan application* process. The system administrator provided you with the event log "loanApplications.xes.gz" (hereafter denoted by L) and a description of the process. Based on the information, the financial institute wants you to analyze the process and answer various business questions, which are listed below.

The Data

A short description of the process is as follows. In the log, every case represents a loan application. An application is filed through the website or by an employee from the institute upon the customer's request. Then, the application goes through several checks and corrections. If the institute can offer loans to the customers, the employees send the offer(s) to the customers. The customer can decide whether to accept an offer and send the necessary documents to proceed. These documents are then verified by employees from the financial institute. The customers are contacted if they do not respond. After all the necessary documents are received, the institute makes the final decision whether to approve or reject the application.

There are three types of events: events that record **A**pplication state changes, **O**ffer state changes, and **W**orkflow events. This is indicated by the prefix (A, O, W) in each event. The process has the following case attributes:

Attribute Name	Attribute Description
LoanGoal	the reason the loan was applied for
ApplicationType	new credit or limit raise
RequestedAmount	requested loan amount

A few event attributes are related to the offer events, i.e., events star with the prefix *O_*.

Attribute Name	Attribute Description
OfferID	ID of the offer
FirstWithdrawalAmount	the initial withdrawal amount
NumberOfTerms	the number of payback terms agreed to
MonthlyCost	the monthly cost
Accepted	indicates whether the offer is still acceptable* by the customer
Selected	whether the offer was selected by the customer
CreditScore	the credit score of the customer
OfferAmount	the offered amount

*If Accepted is *True*, the customer may still aceept the offer. However, a customer is evaluated several times during the process (e.g., based on the provided documents). Depending on the results of the checks, the bank may one-sidedly withdraw an offer. In that case, *Accepted* is set to *False*, and the customer cannot accept the offer anymore. Nevertheless, the bank may still create a new offer.

To help you better understand the data, the company provides the following information for the events related to application state changes.

- *A_Submitted*: a customer has submitted a new application via the website. If the application is started by the bank, this state is skipped.
- *A_Concept*: the customer just submitted the application (or the bank started it), and a first assessment has been done automatically.
- *A_Accepted*: after a call with the customer, the application is completed and assessed again. If there is a possibility to make an offer, the status is accepted. The employee now creates one or more offers.
- *A_Complete*: the offers are sent to the customer. The bank waits for the customer to return a signed offer along with the remaining documents (payslip, ID etc).
- *A_Validating*: the offer and documents are received and checked. During this phase, the status is validating.
- *A_Incomplete*: if documents are not correct or some documents are still missing, the status is set to incomplete, which means the customers needs to send missing/updated documents.
- *A_Pending*: if all documents were received and the assessment is positive, the loan is final and the customer is payed. The application state is set to pending, which indicates a successful application.
- *A_Denied*: if the loan cannot be offered to the customer, because the application doesn't fit the acceptance criteria, the application is declined, which results in the status 'denied'.
- *A_Cancelled*: if the customer never sends the requested documents or calls to tell that the loan is no longer needed, the application is cancelled.

Assignment Details

- Total number of points achievable: 100 (20 % of final course grade)
- Group size: 2-3
- Input: event logs, LaTeX report template
- Deliverables: PDF report, figures folder, .pnml files of potential models.

Tools

In this assignment, if not otherwise stated in the task description, you may choose whichever tools you want to use. These can be the tools introduced in the lecture, the previous assignment or other tools (e.g., commercial) as long as this is under an appropriate license (i.e., we can also have access). Examples for commercial tools with an academic license are:

- Disco: `https://fluxicon.com/disco/`
- Celonis: `https://www.celonis.com`
- UiPath Process Mining: `https://www.uipath.com/product/process-mining`

We should be able to follow your analysis even if we do not know the particular method or tool you are using. Therefore, whichever tool or method you use, ensure that your report clearly documents the inputs and the results you obtain.

Deliverables

The deliverables comprise a report, a folder with high resolution figures, and .pnml files of potential models.

Report Your written report will be the main basis for grading. In the report, you should present your methods, motivation, results, and explanations. Doing so, **clearly indicate which answer belongs to which question.** Moreover, the report should be **self-contained** (i.e., it should not require references to the figures folder). The length should be at **most 20 pages**, including the title page, excluding appendices. Large, repetitive or less relevant figures as well as raw data/results can be placed in appendices. However, a reference should be included in the main report. Also, do not use appendices to create a longer report. The report should be understandable and readable by itself, appendices are only included to further support and allow for additional insights. Make sure to include **all group members' names and student IDs on the title page**. If you do not want to use the provided LaTeX template, create a document mimicking its structure in any text editor of your choice. Nevertheless, we expect the report itself to be handed in in PDF format. In total, 10 points are reserved for the style of your report. The following criteria (amongst others) are considered when judging the style of the report:

- Proper spelling, punctuation, readability
- Comprehensive structure
- Use of **adequate** visualizations for showing (aggregated) results or illustrating methods
- Figures have captions, axes have labels, diagrams have headers

- Figure quality (e.g., resolution and relevance)
- All figures and tables are numbered and referred to in text
- Any figure or table that is not mentioned in the text is, by definition, irrelevant
- Contents of the appendix are clearly labeled and referred to

Hints

When answering the questions, document what you did and carefully describe and explain your results. In particular, explain how you derived your results, on which facts you base your claims, and motivate the methods you used. To improve your discussion and explanation, results from previous questions can (and should) be referred to. **Note, that this assignment is a lot more open and less guided than Assignment I. You are asked to decide on a lot of details of the applied methods and the direction of your investigation by yourself. The provided real-life data may exhibit data quality issues (e.g., noise, missing data, contradictions, or irrelevant parts) and will not always allow for a clear and straightforward result. Therefore, grading will mainly be based on reasonable decisions, proper appliance of adequate methods, and your understanding and interpretation of the applied methods and results obtained.**

Optional Resources

- Social networks video: `https://www.youtube.com/watch?v=RnAggRd4xVE`
- Jupyter: `https://jupyter.org/index.html`
- PM4Py installation guide: `https://pm4py.fit.fraunhofer.de/install`
- PM4Py documentation: `https://pm4py.fit.fraunhofer.de/docs`
- PM4Py source code: `https://github.com/pm4py/pm4py-source/tree/release/pm4py`
- ProM: `http://www.promtools.org/doku.php`
- PMTK: `https://pmtk.fit.fraunhofer.de/`

Question 1: Preliminary Analysis

The goal is to get an overview over the event log L and to gather information to direct and facilitate further analysis.

(a) (3 points) Compute the following basic statistics for L:

- Number of traces and trace variants
- Number of events and average trace length
- Timespan covered by the log
- Number of distinct activities
- Number of distinct resources.
- Duration of the most time-consuming case, the shortest case, as well as the median case duration

(b) (7 points) Additionally, investigate the following application-specific aspects:

- What is the ratio of successful applications (traces where *A_Pending* occurs)?
- What is the ratio of unsuccessful applications (traces with the occurrence of *A_Denied* or *A_Cancelled*)?
- What is the number of cases for each *ApplicationType*? What is the ratio of successful applications for each *ApplicationType*?
- What are the numbers of cases for each *LoanGoal*? What is the ratio of successful applications for each *LoanGoal*?
- How many offers are created in log L?
- What is the ratio of applications that have more than one loan offer?
- What is the ratio of *offers* that are refused (the occurrence of *O_Refused* indicates an offer is refused)?
- What is the potential conceptual problem when computing the ratio on a log containing incomplete cases? Considering your findings in Q1 d, can we neglect this problem?

Hint: "A picture is worth a thousand words." Considering the style points, you are encouraged to use figures to answer the questions when suitable.

(c) (5 points) Using L, create dotted charts as specified below. Describe, explain, and briefly discuss any patterns you can find (e.g., case arrival rate, batch processing, etc.).

- The first dotted chart visualizes the events for every case over the event timestamp. Use the timestamp as the x-axis, the case ID for the y-axis, timestamp of the first event for sorting, and the activity name for the color attribute.
- The second dotted chart visualizes the events for every case over time since case start. Use the time since case start for the X-axis, the case ID for the y-axis, the duration of the trace for sorting, and the activity name for the color attribute.
- If you followed the description correctly, you should see "a block of cases with sparse events" in the second dotted chart. Create a filtered log L_{sparse} by filtering log L to keep only these cases and create another dotted chart for L_{sparse} using the same configuration as the first one. Discuss what you see from the chart.

Hint: Dotted Charts are a very helpful tool for preliminary data exploration. Playing with additional perspectives might be helpful when making design decisions or looking for explanations in other tasks of this assignment as well.

(d) (3 points) The log may contain incomplete cases, which have either not terminated yet, have not been logged since their start, or where the logging has likely been incomplete. Design a reasonable filtering strategy and create a log L_{complete}, which contains only complete cases. You are free to make reasonable assumptions on the data. Justify your decisions for the filters.

Total for Question 1: 18

Question 2: Process Discovery and Conformance

In this question, you will discover multiple process models covering different facets of the process and analyze the conformance. The company provides you with the filtered log "loanApplicationsFiltered.xes.gz" (hereafter L_{filtered}) to conduct the following analysis.

(a) (3 points) The financial institute wants to understand the process for the three types of events (i.e., application state changes, offer state changes, and workflow events) separately. Therefore, based on L_{filtered}, create three more event logs using the event prefixes (i.e., A, W, O), hereafter referred to as L_{A}, L_{W}, and L_{O} respectively. Also, for each of the three sublogs L_A, L_W, and L_O, compute the following basic statistics:

- Number of cases and trace variants
- Number of events and average number of events per case
- Number of distinct activities

(b) (2 points) For your convenience, the company also provides you with an additional log "loanOffers.xes.gz" (hereafter L'_O) that contains only the offer events (events with prefix O). This is similar to the log L_O you created previously. The difference is that it uses OfferID (instead of ApplicationID) as the case ID. Create a Directly-Follows Graph (DFG) for both L_O and L'_O without any filter. Compare the two DFGs and discuss which logs are preferable for discovering the process of offer events?

(c) (9 points) For the logs L_{A}, L_{W}, and L'_{O}, apply at least two different discovery techniques with various parameter settings (i.e., at least three different parameters if there are parameters) to discover Petri nets. Select one "best" model for each log, hereafter M_{A}, M_{W}, and M_{O}, and explain your choice. To this end, especially consider the quality dimensions precision, fitness, generalization, and simplicity. As there are no good "hard" metrics for simplicity and generalization, assess these based on your current understanding of the process.

Hint: Do not waste space including all discovered models in your report. Include the chosen models (and possibly representatives or snippets of rejected models) and put the rest in appendices.

(d) (3 points) Describe the process that M_{A}, M_{W}, and M_{O} model.

(e) (3 points) Investigate and describe the deviations of L_{A}, L_{W}, and L'_{O} with respect to their corresponding models M_{A}, M_{W} and M_{O}. For example, you may provide fitness scores and deviations projected onto models. In particular, you should also discuss a few alignments. Can you detect systematic problems?

(f) (6 points) Based on the process description and what you have found so far, can you manually create a model (denoted as M_{hand}) for the overall process i.e., for L_{filtered}? Elaborate on the major design decisions and describe the behavior of the model.

Hint: You are free to use any tools to create the model. Some options to create a Petri net could be (1) Use the "Create Petri net (Text language based)" Plugin in ProM. For your convenience, you can use the provided M_hand.txt file to start with. (2) Create a BPMN model and convert it to Petri net (e.g. in ProM or PM4Py `https://pm4py.fit.fraunhofer.de/documentation#item-13-4`). (3) Create the Petri net in PM4Py: `https://pm4py.fit.fraunhofer.de/documentation#item-4-3`.

(g) (3 points) In addition, use a discovery algorithm of your choice to automatically discover a "best" (in terms of quality dimensions. To this end, apply the same procedures as in Q2(c)) to discover a process model M_{auto} for L_{filtered}.

(h) (6 points) Use L_{filtered} to evaluate the quality of models M_{hand} and M_{auto} in terms of precision and fitness. Investigate and describe the deviations you identify (Follow the same procedures in Q2(e)). Compare the two models and explain the differences. In your opinion, does M_{hand} describe the process better? If so, provide a few reasons why does the algorithm you used had difficulties discovering a better model. If not, discuss possible improvements for M_{hand}.

Total for Question 2: 35

Question 3: Investigation of Attributes and Social Networks

In this question, you will investigate the attribute "org:resource" contained in the log "loanApplicationsWorkflows.xes.gz" (hereafter L_{AW}) given by the company. The log contains events with prefixes (A,W).

(a) (2 points) Create a dotted chart for L_{AW} using resource and activity name for the two axes. Describe, explain, and briefly discuss any patterns you can find.
Hint: Note that the visualization for large logs in dotted chart is sampled, i.e., not every point is shown until you zoom in to a certain degree.

(b) (6 points) Create a Similar-Task Social Network in ProM to investigate the social network for log L_{AW}. Try different settings of the parameters to find the most expressive network. Describe, explain, and discuss the discovered network. Are there activities that are specific for certain clusters? Describe any patterns you can find.
Hint: (1) The distance metrics have a significant impact on the resulting networks. (2) The tools you used are not limited to social networks. Feel free to combine other analyses to support your findings.

Total for Question 3: 8

Question 4: Performance

In this question, you will analyze the performance of the process to discover bottlenecks and inefficiencies.

(a) (6 points) Analyze time performance based on the process models (M_{A}, M_{W} and M_{O}) and the corresponding logs (L_{A}, L_{W}, and L'_{O}) created in Q2. To this end, project performance metrics on the models and discuss the results.

(b) (8 points) The Performance Spectrum is a useful tool to complement traditional performance analysis, which typically maps aggregated measures on process models. Use the Performance Spectrum to enrich the analysis from Q4(a). Are you able to identify any interesting patterns (e.g., batch processing, FIFO patterns)? Pick four interesting findings and discuss them.
Hint: You may use the implementation in PMTK or ProM.

(c) (3 points) Based on your findings in Q4(a) and Q4(b), which advice would you give to the stakeholders to improve their time performance?

Total for Question 4: 17

Question 5: Summary and Recommendations

Summarize your findings for each question. Based on your results, indicate possible improvements for the stakeholder. Give ideas and directions for a potential follow-up project.

Total for Question 5: 12