Process Mining for Fraud Detection

van Dongen, B. (2017). BPI Challenge 2017 [Data set]. Eindhoven University of Technology.

Master: Intelligent Systems

Subject: SJK013 - Process Mining

Author: Pablo Muñoz Alcaide

GENERALITAT VALENCIANA

valgrAI Valencian Graduate School and Research Network of Artificial Intelligence

Table of Contents

01 Introduction & Motivation

02 Dataset

03 Process Discovery: PM4Py & Disco

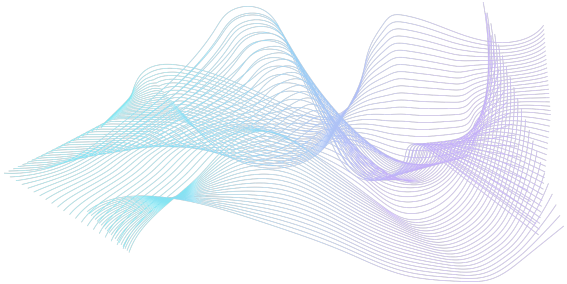
04 Experiments & Results: ProM

05 Conformance Checking

06 Conclusions

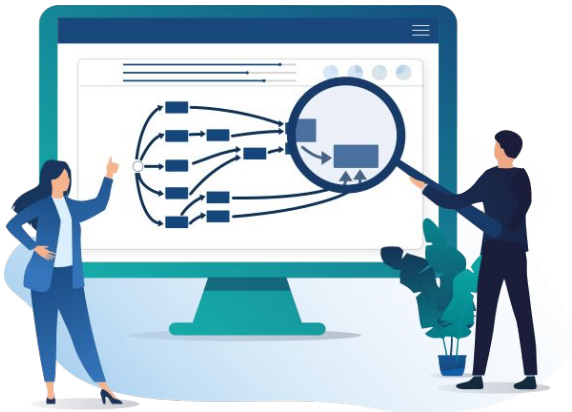
01

Introduction & Motivation



Introduction

The main objective of this project is to utilize process mining techniques to detect financial fraud of a loan application process of a Dutch Financial institution.



Objectives

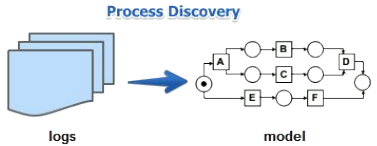
Application of acquired Knowledge in SJK013

ProM process mining workbench

PM4PY

Disco by Fluxicon

Test of tools

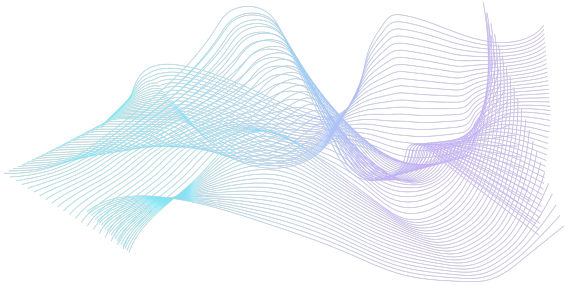


Analyze Event Log & Process Discovery

Identify anomalies, deviations, and suspicious activities within the transaction data.

Provide insights to financial institutions for improving fraud detection and prevention strategies.

02



Dataset

Introduction to the Dataset

BPI Challenge 2017

The event log pertains to a loan application process of a Dutch financial institute. The data contains all the applications filed through an online system in 2016 and their subsequent events until February 1st 2017, 15:11.

	Action	org:resource	concept:name	EventOrigin	EventID	lifecycle:transition	time:timestamp	case:LoanGoal	case:ApplicationType	case:concept:name	case:RequestedAmount	FirstWithdrawalAmount	NumberOfTerms	Accepted	MonthlyCost	Selected	CreditScore	OfferedAmount	OfferID
1	statechange	User_1	A_Submitted	Application	ApplState_1582051990	complete	2016-01-01 09:51:15.352000+00:00	Existing loan takeover	New credit	Application_652823628	20000.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	Created	User_1	W_Handle leads	Workflow	Workitem_1298499574	schedule	2016-01-01 09:51:15.774000+00:00	Existing loan takeover	New credit	Application_652823628	20000.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
9	Created	User_52	O_Create Offer	Offer	Offer_148581083	complete	2016-01-02 11:29:03.994000+00:00	Existing loan takeover	New credit	Application_652823628	20000.0	20000.0	44.0	True	498.29	True	979.0	20000.0	NaN
10	statechange	User_52	O_Created	Offer	OfferState_1514834199	complete	2016-01-02 11:29:05.354000+00:00	Existing loan takeover	New credit	Application_652823628	20000.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Offer_148581083
11	statechange	User_52	O_Sent (mail and online)	Offer	OfferState_2051164740	complete	2016-01-02 11:30:28.606000+00:00	Existing loan takeover	New credit	Application_652823628	20000.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Offer_148581083
12	Deleted	User_52	W_Complete application	Workflow	Workitem_1007505836	ate_abort	2016-01-02 11:30:28.621000+00:00	Existing loan takeover	New credit	Application_652823628	20000.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Overview of the entire event log.

Description of the Dataset (i)

Attribute (Name in Dataset)	Explanation	Example
caseID (case: concept: name)	The unique identifier of the application. The case identifier is necessary to distinguish different executions of the process.	Application_652823628
taskID (concept: name)	The name of the event. The name always starts with the initial of the event origin (ref. EventOrigin). There are three types of tasks: A: States of the application, O: States of the offer belonging to the application, W: States of the work item belonging to the application	A_Submitted
originator (org: resource)	The unique identifier of the person who executed the task.	User_1
Eventtype (lifecycle: transition)	The state of the task. There are seven possible values: schedule, start, suspend, resume, complete, withdraw and ate_abort.	complete
LoanGoal (case:LoanGoal)	The reason why the loan was applied for. There are fourteen possible values: Boat, Business goal, Car, Caravan / Camper, Debt restructuring, Existing loan takeover, Extra spending limit, Home improvement, Motorcycle, Not specified, Other, Remaining debt home, Tax payments and Unknown.	Existing loan takeover
RequestedAmount (case: RequestedAmount)	The requested loan amount (in EUR). The values vary between 0 and 450000.	20000.0

Description of the Dataset (ii)

Attribute (Name in Dataset)	Explanation	Example
ApplicationType (case: ApplicationType)	The type of the application. There are two possible values: Limit raise and New credit.	New credit
Action (Action)	The action made in the dataset when entering the data.	statechange
EventOrigin (EventOrigin)	The origin of the event. There are three possible values: Application, Workflow an Offer.	Application
EventID (EventID)	The unique identifier of the event.	ApplState_1582051990
Timestamp (time: timestamp)	The time at which the event occurred. The timestamp is used to put the events in the right order.	2016-01-01 09:51:15.352000+00:00
FirstWithdrawal Amount (FirstWithdrawalAmount)	Only filled in when taskID = O_Create Offer. The initial withdrawal amount.	20000.0
Accepted (Accepted)	Only filled in when taskID = O_Create Offer. Boolean that indicates whether an offer is still valid or not (based on the assessment of certain client information).	True

Description of the Dataset (iii)

Attribute (Name in Dataset)	Explanation	Example
Selected (Selected)	Only filled in when taskID = O_Create Offer. Boolean that indicates whether an offer is signed by the customer or not.	True
NumberOfTerms (NumberOfTerms)	Only filled in when taskID = O_Create Offer. The number of payback terms agreed to.	44.0
MonthlyCost (MonthlyCost)	Only filled in when taskID = O_Create Offer. The monthly costs to be paid by the customer to reimburse the loan.	Application
CreditScore (CreditScore)	Only filled in when taskID = O_Create Offer. The credit score of the customer. A high credit score provides high creditworthiness and vice versa.	979.0
OfferedAmount (OfferedAmount)	Only filled in when taskID = O_Create Offer. The loan amount offered by the bank.	20000.0
OfferID (OfferID)	Only filled in when taskID starts with O_ (except when taskID = O_Create Offer). The unique identifier of the offer. An application can have one or more offers.	Offer_148581083

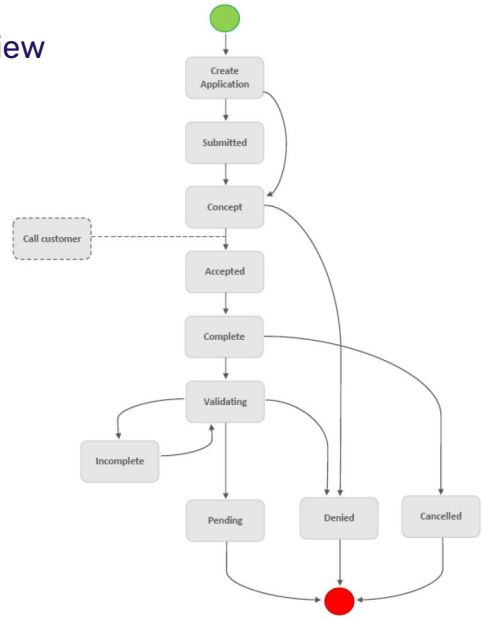
Dataset Characteristics

- **Number of Events**: 1.202.267
- **Number of Cases**: 31.509
- **Unique Activities**: 26
- **Most Frequent Activities**: W_Validate application, W_Call after offers, W_Call incomplete files, W_Complete application and W_Handle leads.
- **Least Frequent Activities**: A_Denied, W_Assess potential fraud, O_Sent (online only), W_Shortened completion, W_Personal Loan collection.
- **Average events per case**: 38.15
- **Average Case Duration**: 21 days.
- **Number of Different Paths**: 15.930

Dataset Understanding: Overview

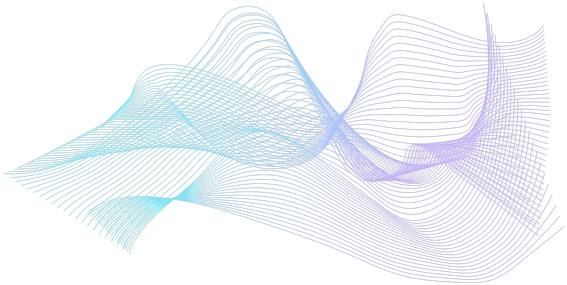
The loan application process involves multiple stages to ensure thorough evaluation and decision-making. Here is a simplified overview of this process:

1. **Create Application**: Initiate the process with a new application.
2. **Submitted**: Official submission by the applicant.
3. **Concept**: Initial checks and assessments.
    - Customer may be contacted for additional information.
4. **Accepted**: Application meets initial criteria.
5. **Complete**: All required information provided.
6. **Validating**: Detailed validation of the application.
    - If incomplete, additional information is required.
7. **Pending**: Awaiting final decision.
8. **Denied**: Application does not meet criteria.
9. **Cancelled**: Application process terminated.



Expected flow of the loan application process. (Blevi, Delporte, & Robbrecht, 2017)

03



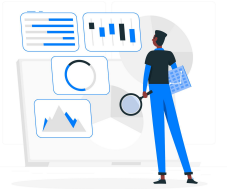
Proces Discovery: PM4Py & Disco

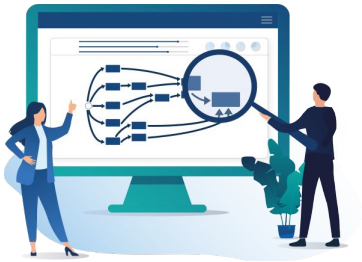
Section Overview

In this section, Disco and PM4Py will be used for process discovery to ensure a complete and deep understanding of the loan application process.

Disco by Fluxicon

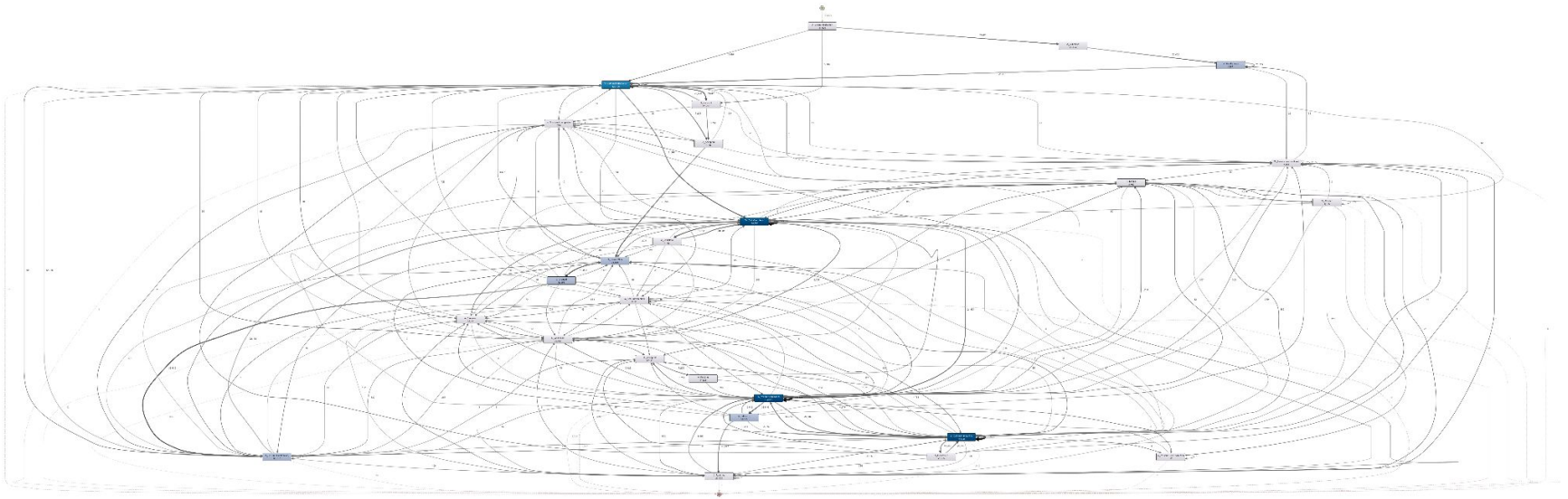
PM4PY





Process Discovery: Disco (i)

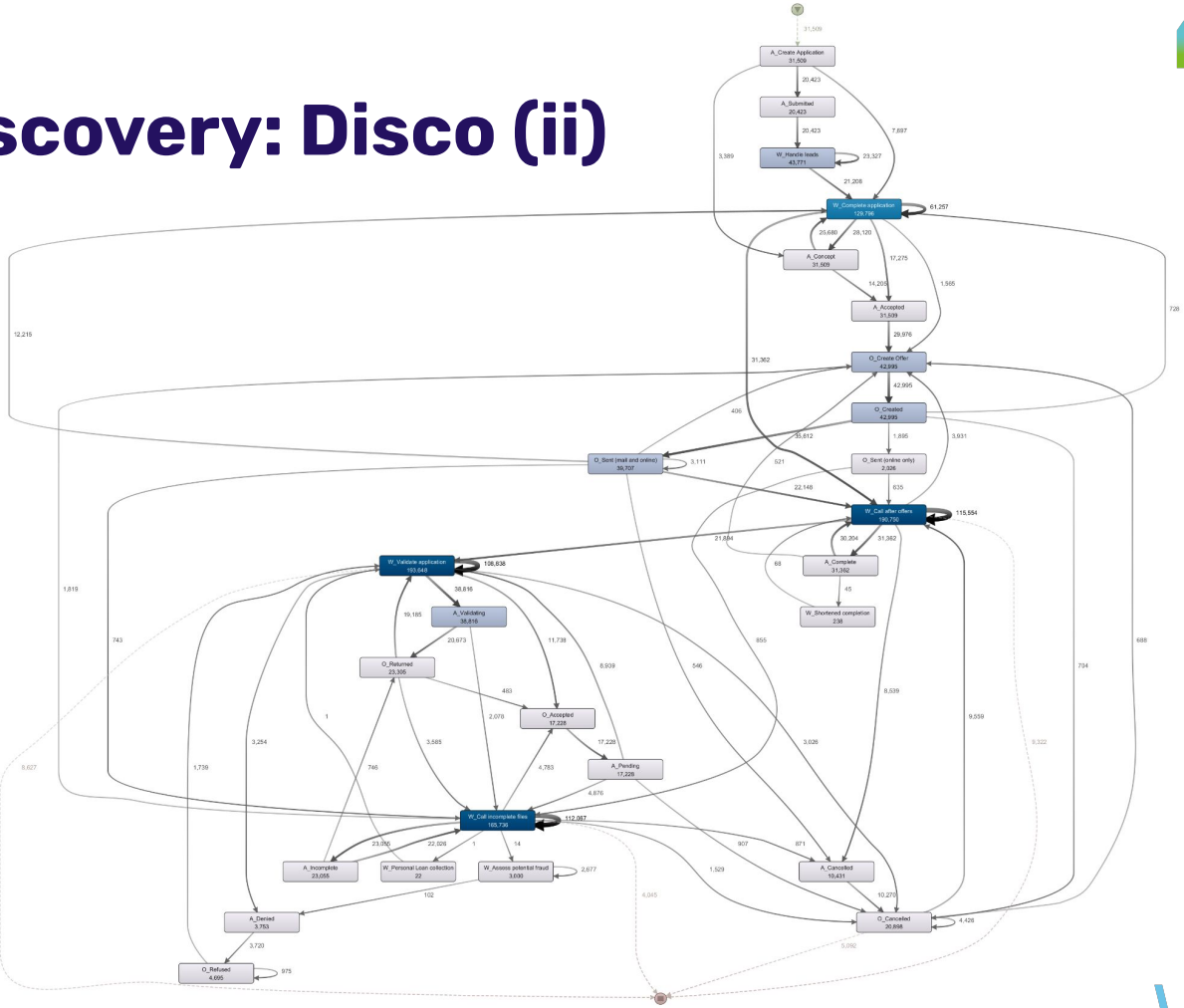
Although we have presented the dataset in a simplified way above, the reality is much more complex and the paths the process can take are multiple.



100 % Paths of the Process (Spaguetti Path)

Process Discovery: Disco (ii)

With 30% of the most common paths, the process still very interconnected and difficult to understand.



30 % Paths of the Process

Process Discovery: Disco (iii)

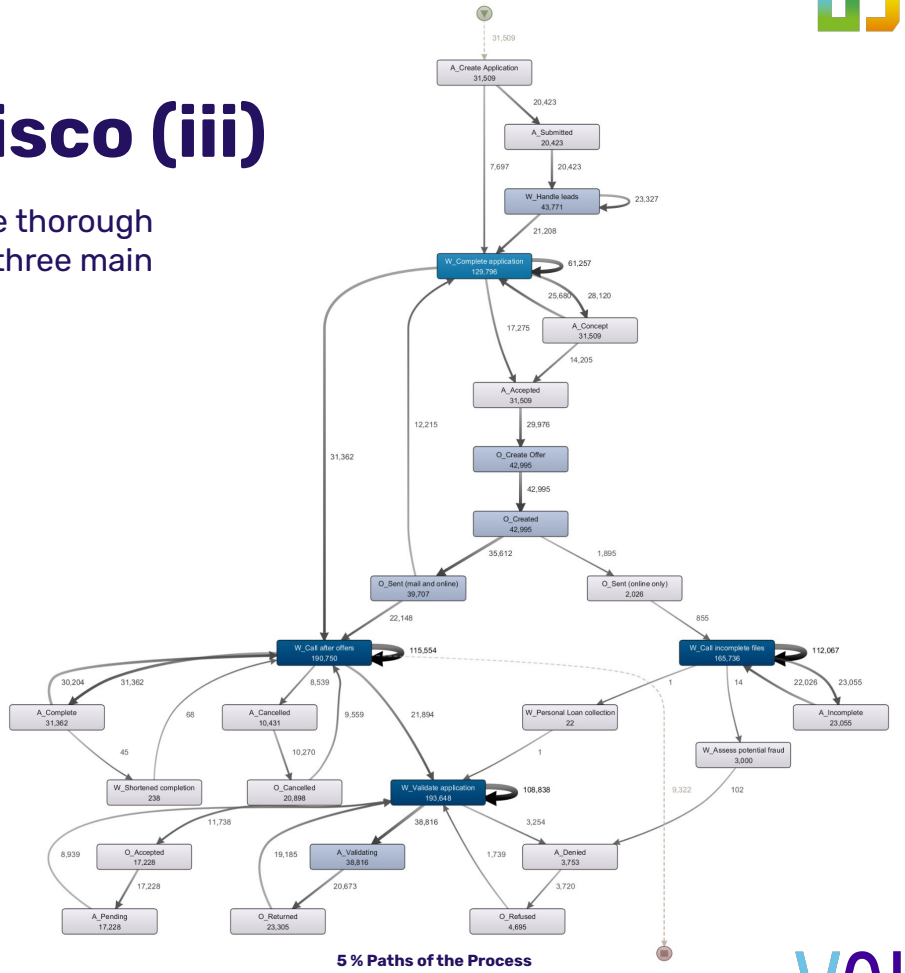
The loan application process is designed to ensure thorough evaluation and decision-making. It is divided into three main phases:

- Application Phase
- Offer Phase
- Workflow Phase.

Each phase plays a critical role in managing the loan application from submission to final decision.

However, the phases are interspersed with one another, making it difficult to understand the function of each.

Therefore, another phase structure can be deduce. (Jeong, Lim, & Bae, 2017)

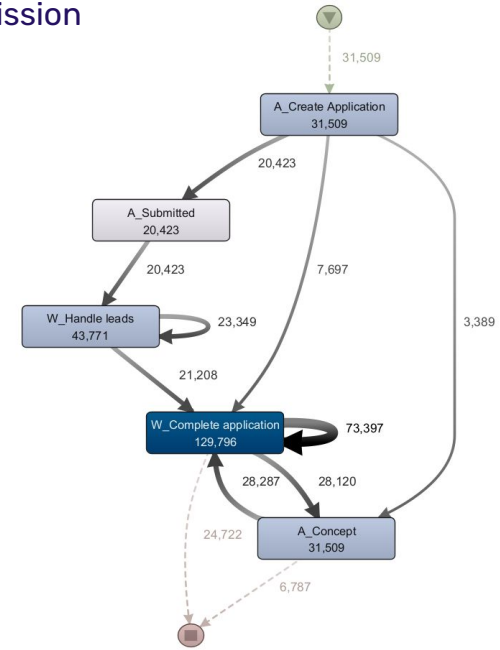


5 % Paths of the Process

PD Disco (iv): Application Phase

The application phase encompasses all activities related to the submission and initial processing of a loan application.

- **Submission**: The applicant creates and submits the loan application. (A_Create Application, A_Submitted)
- **Initial Processing**: Preliminary steps to assess the application. (W_Handle leads)
- **Automatic Assessment:** First assessment done it automatically. (A_Concept)
- **Ending:** Complete this part of the application. (W_Complete application)

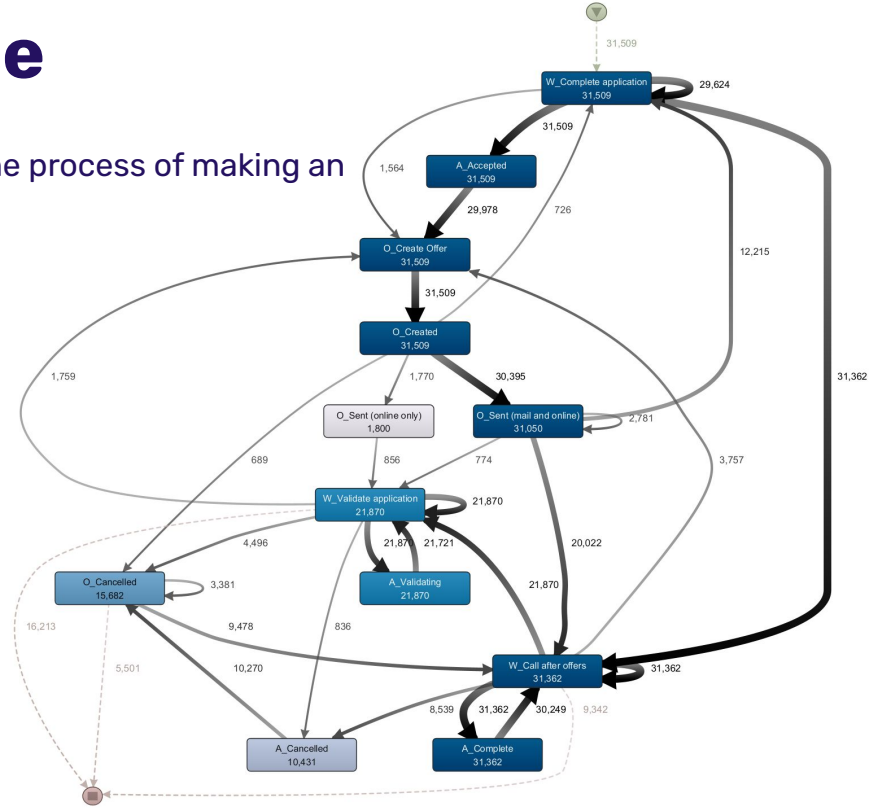


100% Paths of Application Phase

PD Disco (v): Offer Phase

The offer phase encompasses all activities related to the process of making an offer to the consumer and the delivery of the offer.

- **Offer Creation:** The offer is made and created in the system and is ready to send. (O_Create Offer, O_Created)
- **Send Offer**: The offer is sent via mail and via web or only via web. (O_sent)
- **Follow-up**: A bank employee contacts the customer to discuss the offer. They gather the customer's feedback. (W_Call after offers). The offers have been sent and the banks wait for the customer. (A_Complete)
- **Reject or continue**: If the customer isn't interested or doesn't respond, the offer and application will be canceled (A_Cancelled, O_Cancelled), otherwise the customer will submits documents and signature (W_validate application).
- Another reasons such us bad applications or lack of documents can lead to cancel.

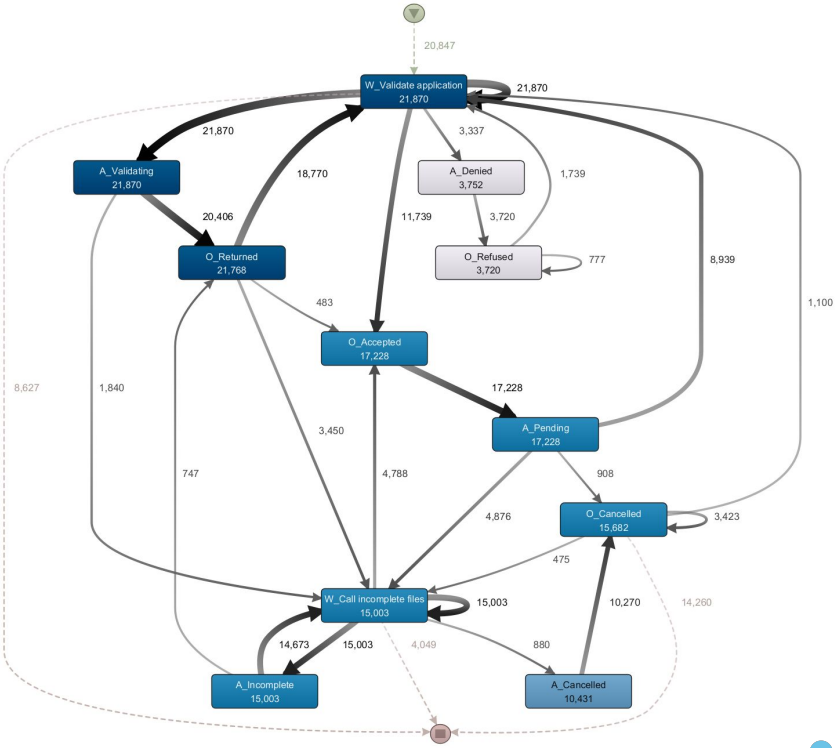


30% paths of offer phase

PD Disco (vi): Decision Process

The decision phase encompasses all activities related to the process of requiring the necessary documents and making the final decision.

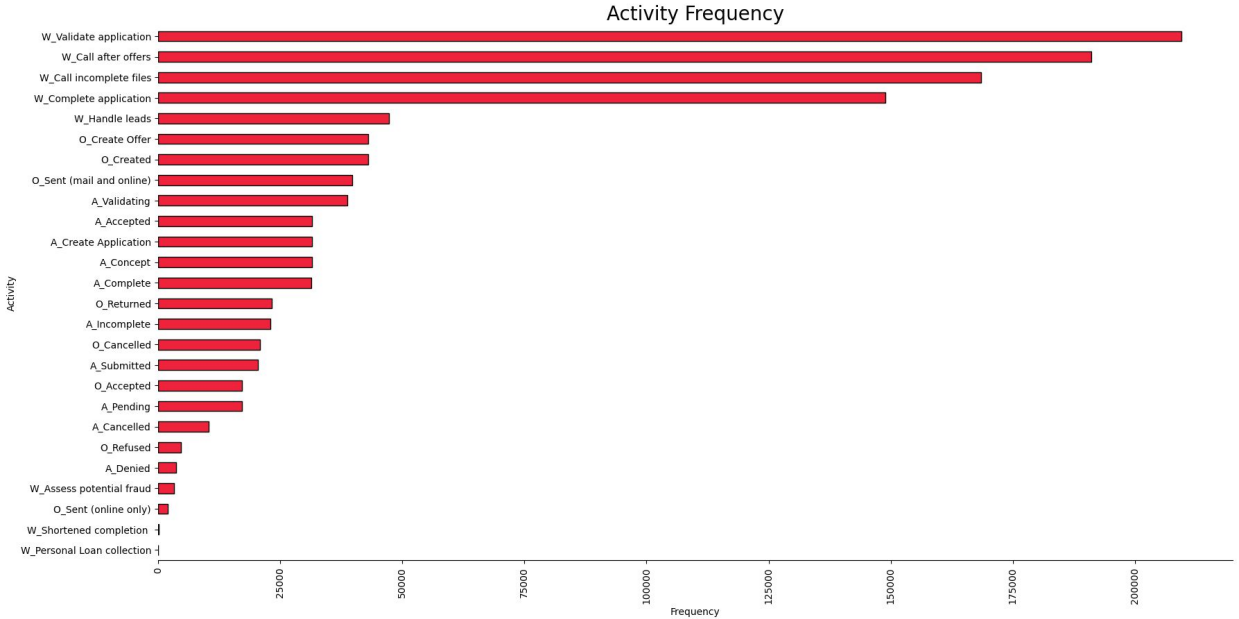
- **Documents Validation:** The bank evaluates the documents and signature of the customer (A_Validating).
- **Offer Acceptance**: The offer is returned to the customer (O_returned), requiring more documents (W_Call Incomplete files, A_incomplete), refusing the offer (A_Denied, O_refused) or waiting for the customer's final decision.
- **Customer Decision:** After providing further documents or not, the customer chooses whether to accept the offer (O_Accepted) and end the process with the payment (A_pending) or to reject the offer (O_Cancelled) and end the application (A_cancelled).



40% paths of decision attributes phase

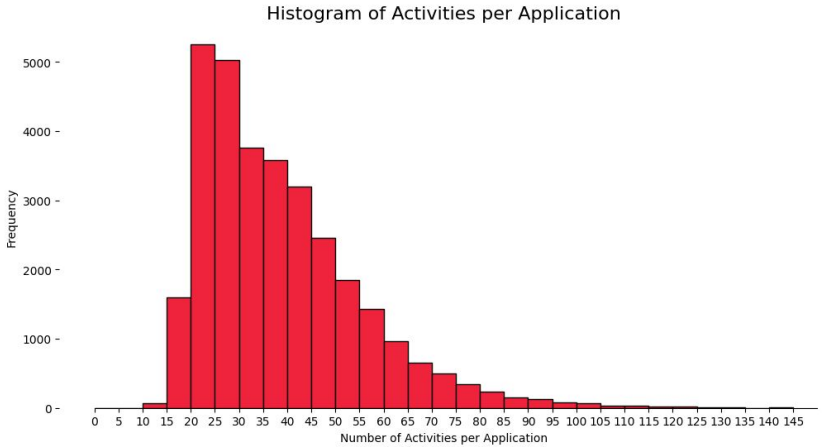
Process Discovery: PM4Py (i)

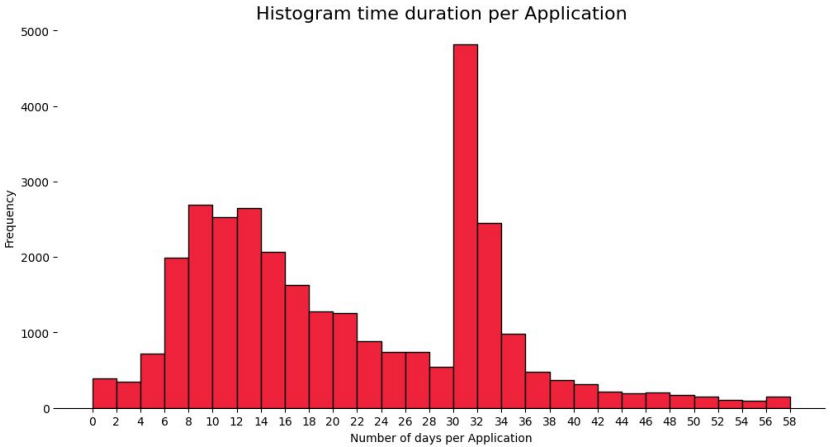
After reading the event log through the pm4py library, some graphics are generated in order to have a better overview of the data.



Process Discovery: PM4Py (ii)

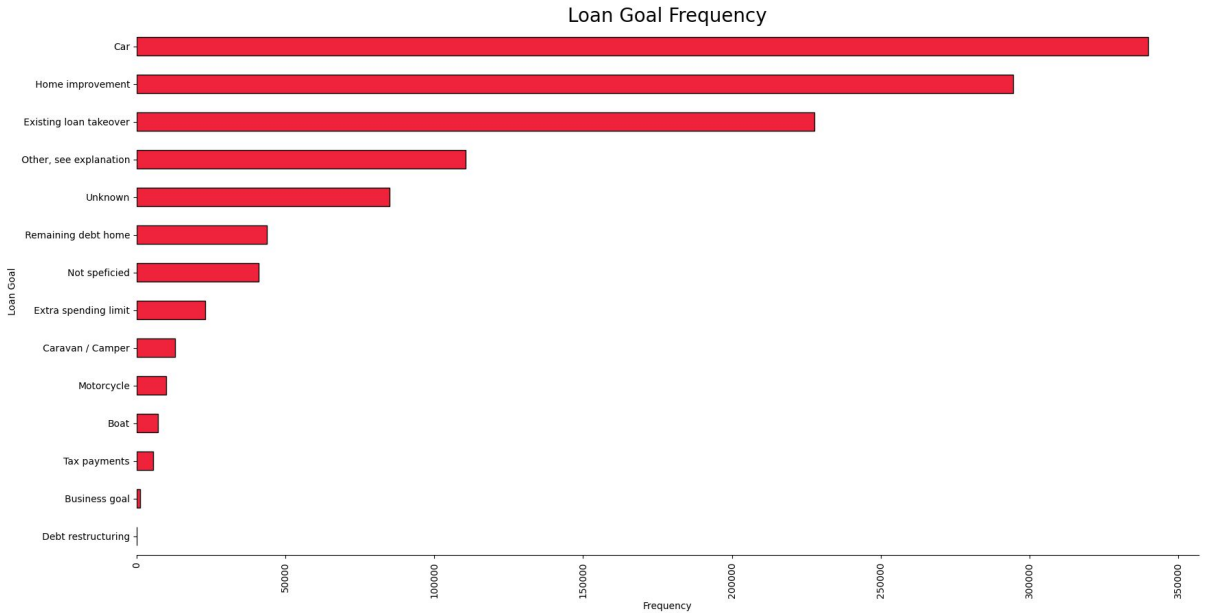
After reading the event log through the pm4py library, some graphics are generated in order to have a better overview of the data.





Process Discovery: PM4Py (iii)

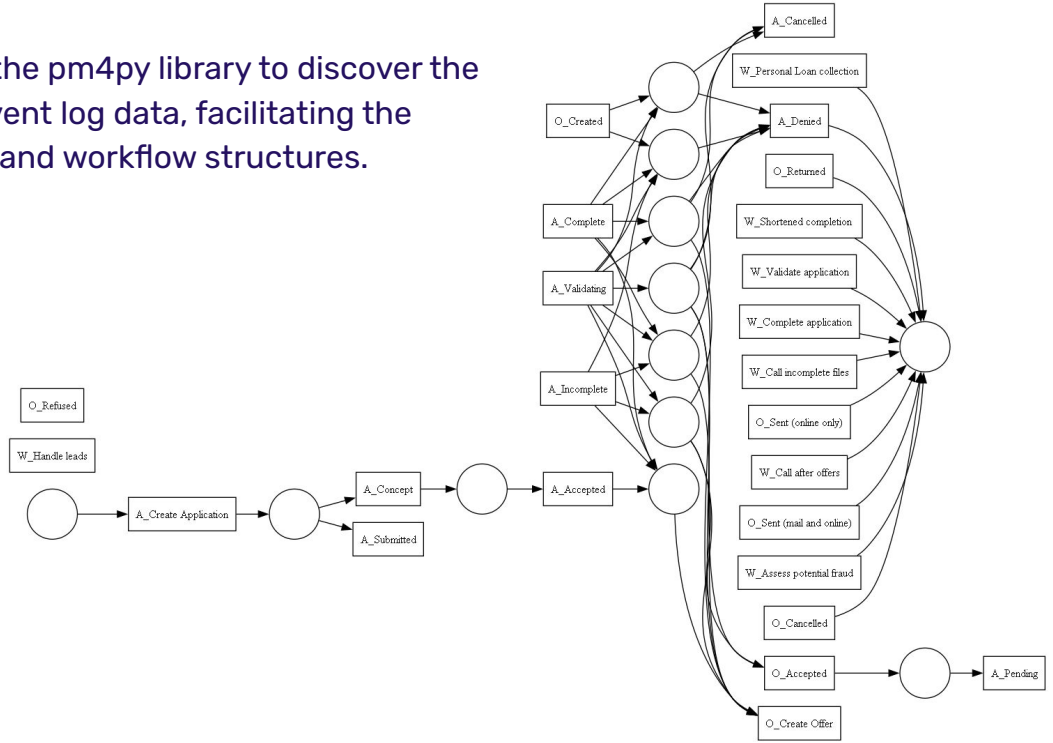
After reading the event log through the pm4py library, some graphics are generated in order to have a better overview of the data.



Process Discovery: PM4Py (iv)

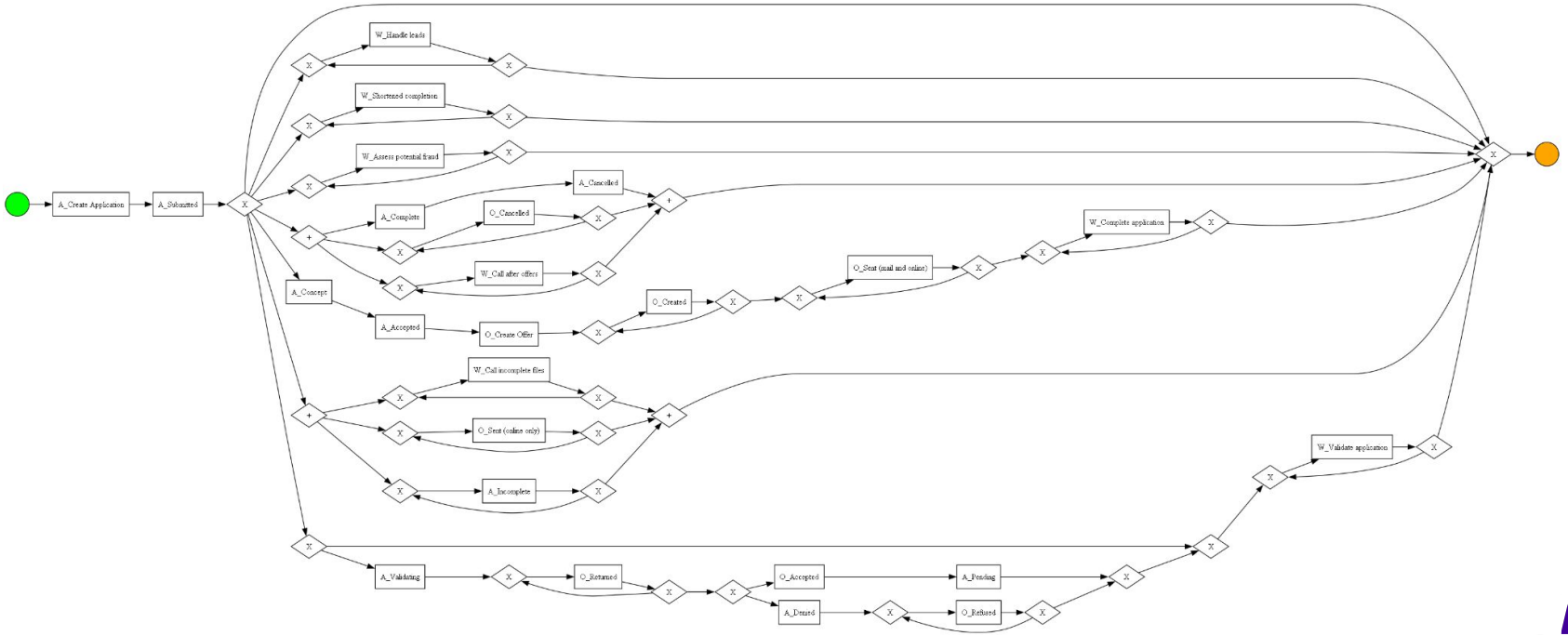
Utilized the Alpha algorithm with the pm4py library to discover the underlying process model from event log data, facilitating the identification of process patterns and workflow structures.

The next Petri net is discovered:



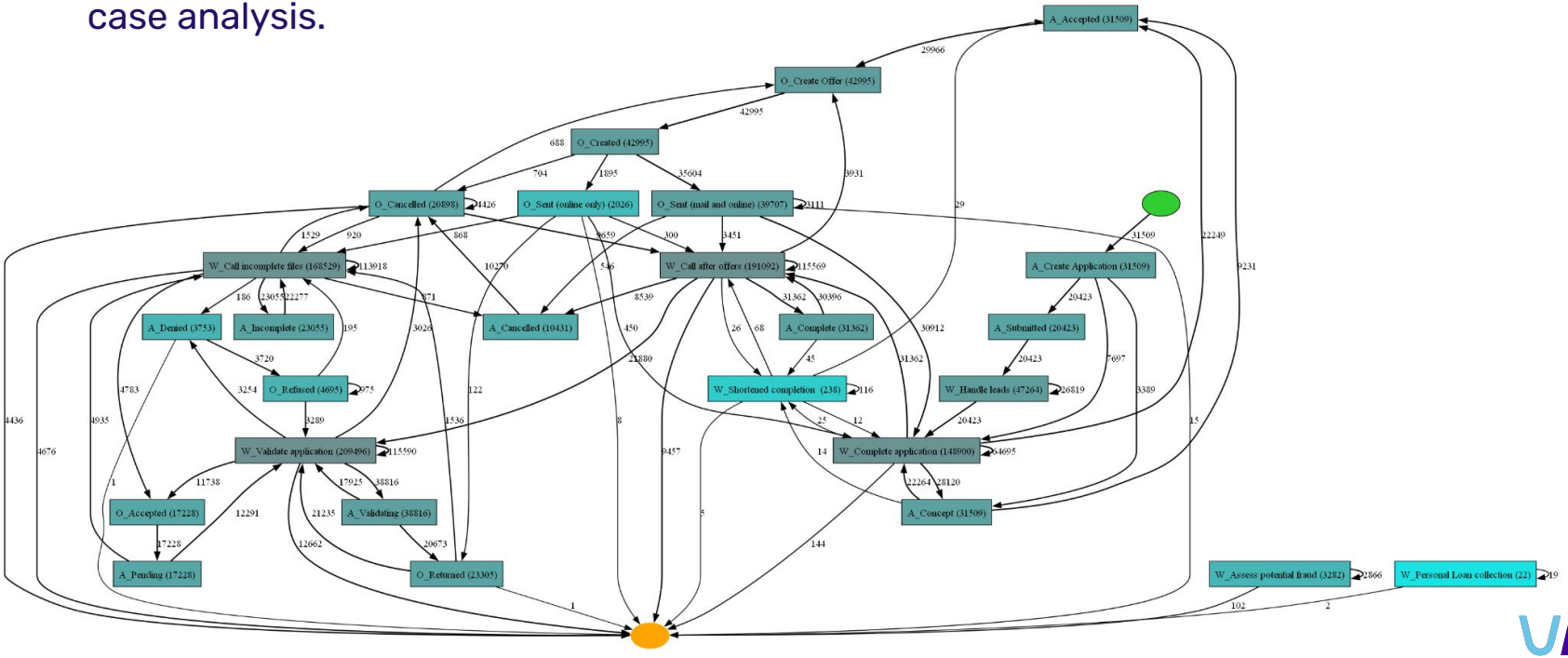
Process Discovery: PM4Py (v)

Other algorithms, such as the Inductive Miner, can also be employed for process discovery, effectively generating BPMN models from event log data.



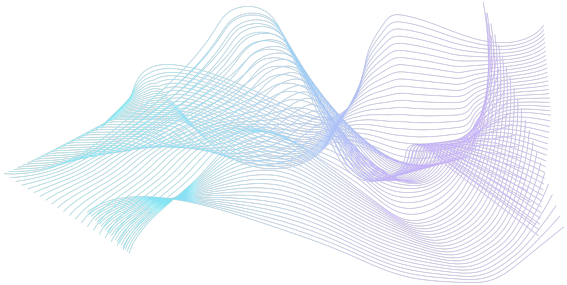
Process Discovery: PM4Py (vi)

Applied heuristic net discovery using pm4py with dependency threshold set to 0.8, and threshold to 0.7, and loop two threshold to 0.5, optimizing activity and case analysis.



04

Experiments & Results



ProM Experiment: Fraud Detection (i)

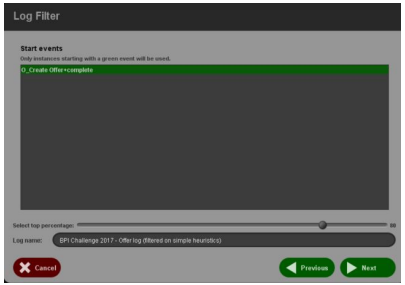
Fraud detection in process mining consists of analyzing event logs to uncover anomalies and deviations from standard process flows. By visualizing these processes, deviations that may indicate fraudulent activities can be identified and investigated.

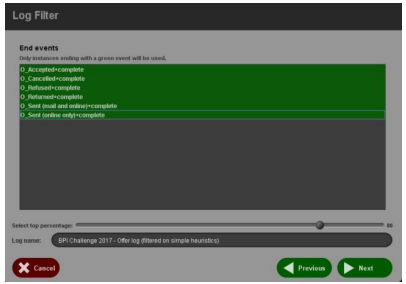
In our case, the next steps were followed:

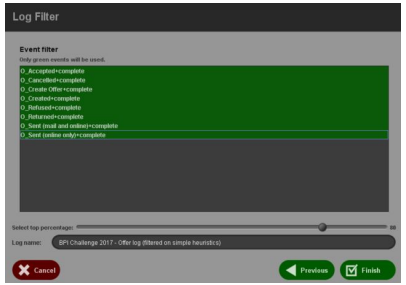
- **Event Log Filter:** The event log was filtered to only include activities related to the offer (O_*) because they are considered the most relevant for identifying potential deviations and anomalies indicative of fraud. (With Disco)
- **Heuristic Filter:** An heuristic filtering is applied to simplify the event log and focus on the most relevant data. This step ensures that only complete cases are analyzed, which increases the reliability of the results.
- **Inductive Visual Miner Visualization:** The inductive visual miner is used to visualize the process models, allowing us to see the typical paths and any deviations in a clear and interactive manner.
- **Tool Platform**: ProM 6.13

ProM Experiment: Fraud Detection (ii)

Filter by Simple Heuristics in order to simplify the event log and focus on the most relevant data. This step ensures that only complete cases are analyzed, which increases the reliability of the results. 80% of top percentage is selected in each case, to ensure that the analysis remains focused on the most frequent and significant paths, enhancing the ability to detect meaningful deviations and potential fraud.



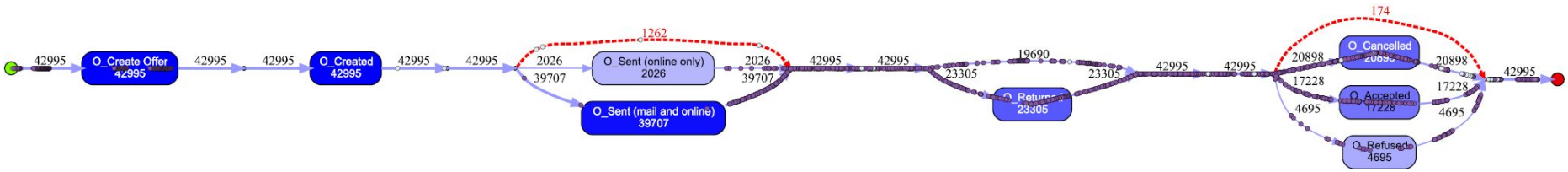




Heuristic Mining Filter Configuration.

ProM Experiment: Fraud Detection (iii)

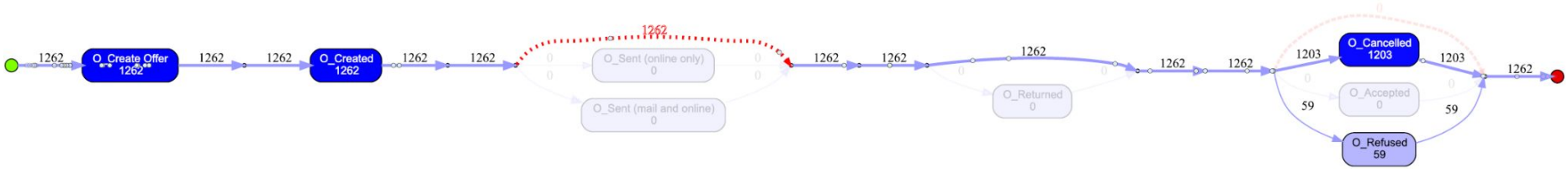
Inductive Visual Miner allows us to detect two deviate path, those who skip twice relevant process steps, such as the sending of the offer or the decision on the offer, and could jump directly to the payment (A_Pending). This could indicate a clear case of fraud inside of the institution.



Interactive Inductive Visual Miner discovered process

ProM Experiment: Fraud Detection (iv)

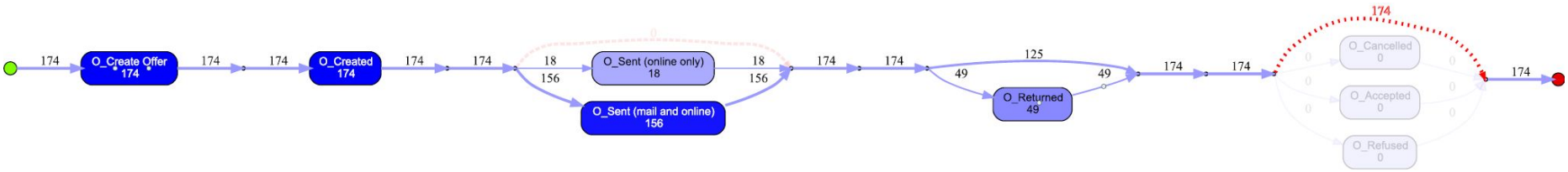
However, the interactive IVM allow us to check the entire path of the deviates processes. Therefore, it can be concluded that possible fraudulent events that bypass the communication process are in all cases cancelled or rejected. In conclusion, they should not be treated as fraud but as a failure in the process.



Interactive Inductive Visual Miner Analysis (i)

ProM Experiment: Fraud Detection (v)

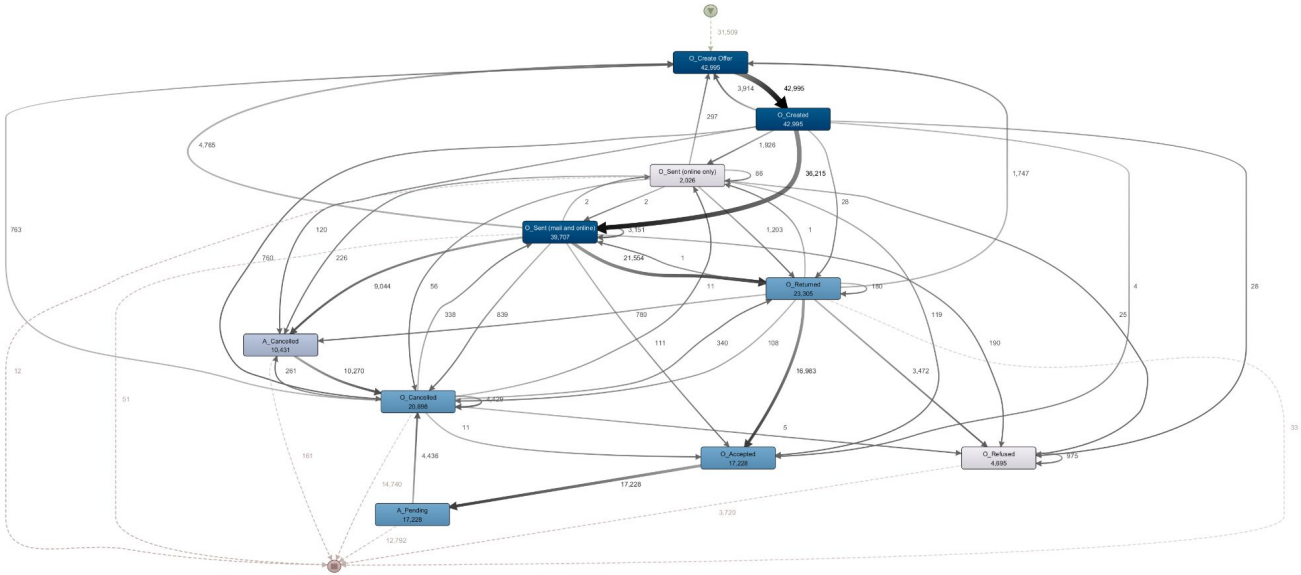
With the Interactive IVM visualisation it is difficult to determine that the second possible fraudulent path is indeed a fraud. However, it is much more suspicious of fraud because it skips the last activities regarding the offer process. Indicating that it could end with direct payment (A_pending), without going through the process of accepting the offer.



Interactive Inductive Visual Miner Analysis (ii)

Fraud Detection (vi)

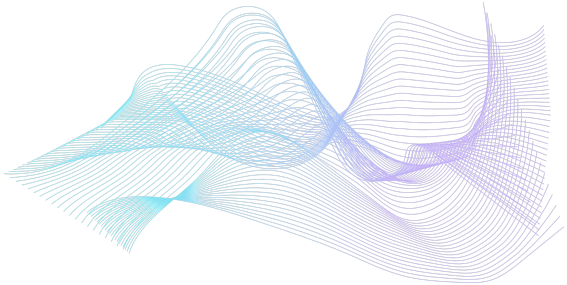
However, thanks to the analysis done previously, we can see how all the offers paid (17,228) have been previously accepted (17,228). Therefore, it can be concluded that the possible cases of fraud are only incomplete cases,



Application Phase Paths

05

Conformance Checking



Conformance Checking (i)

Conformance checking is a core aspect of process mining that involves comparing the actual behavior recorded in event logs with a predefined process model. This comparison helps in detecting deviations from the standard process, which can be crucial for identifying fraudulent activities (Stoop & Oezer, 2012).

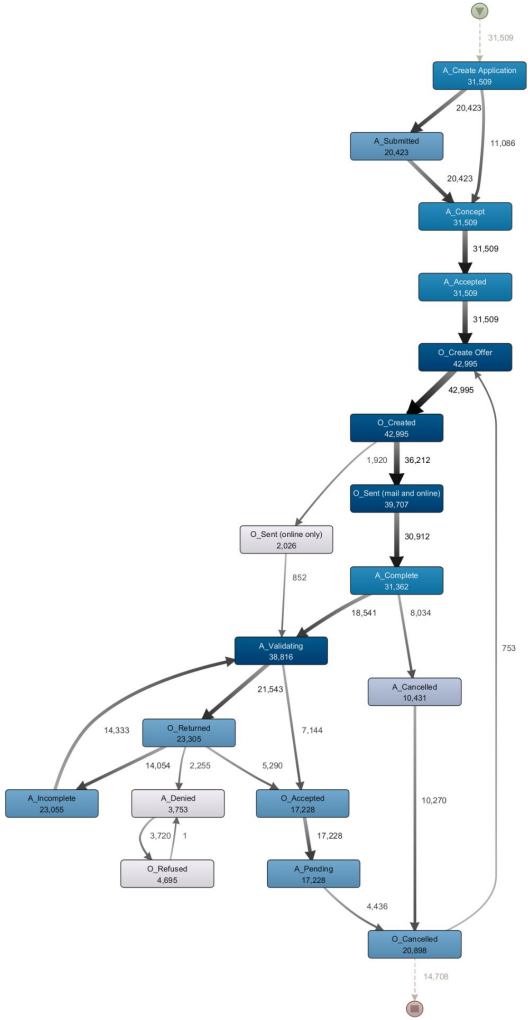
Key Reasons for Conformance Checking in Fraud Detection:

- **Deviation Detection:** Pinpoints where the actual process diverges from the expected model, revealing suspicious behaviours.
- **Process Compliance:** Verifies that processes are followed correctly, spotting non-compliant actions indicating fraud.
- **Quantitative Metrics:** Provides objective measures to assess deviations, prioritizing investigation efforts.
- **Real-Time Monitoring:** Allows prompt intervention, reducing the impact of fraudulent activities.

Conformance Checking (ii)

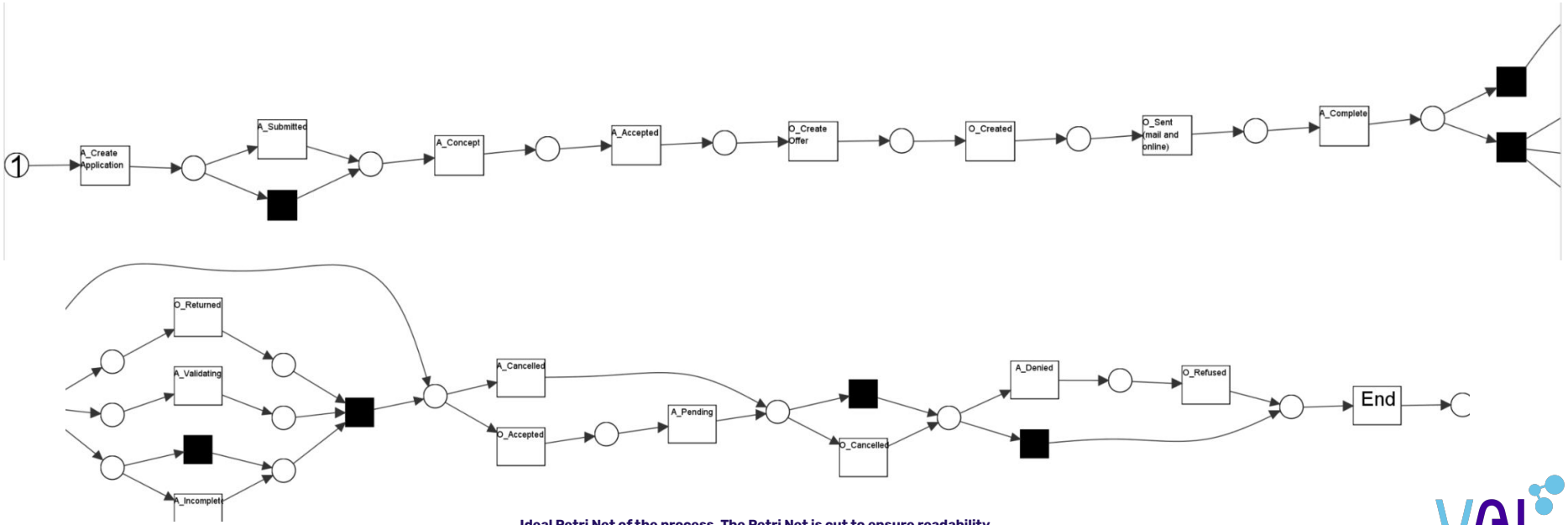
Conformance Checking is applied in this case to the filtered event log, which include only activities related with the application and the offer (O_*, A_*).

The model is generated with Disco, only considering 5% of the most common paths.



Conformance Checking (iv)

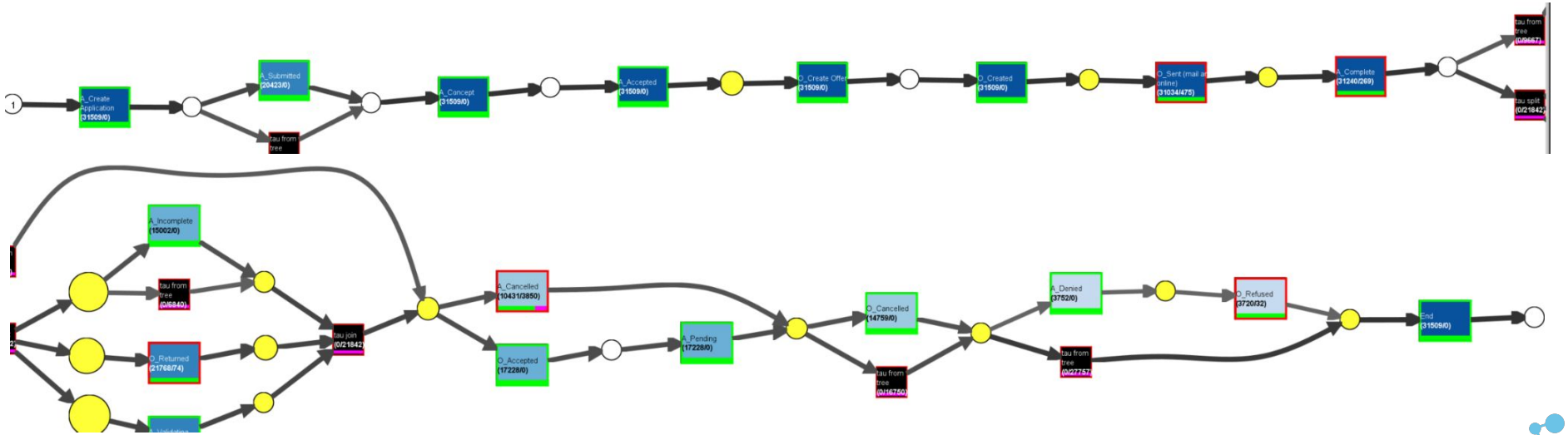
Then the ideal process is transformed into a Petri Net with the inductive miner, with a noise threshold of 0.2 in order to filter out infrequent and potentially irrelevant behaviors, thereby creating a more accurate and simplified model that highlights the most significant and frequent process flows, making deviations indicative of potential fraud more apparent.



Ideal Petri Net of the process. The Petri Net is cut to ensure readability

Conformance Checking (v)

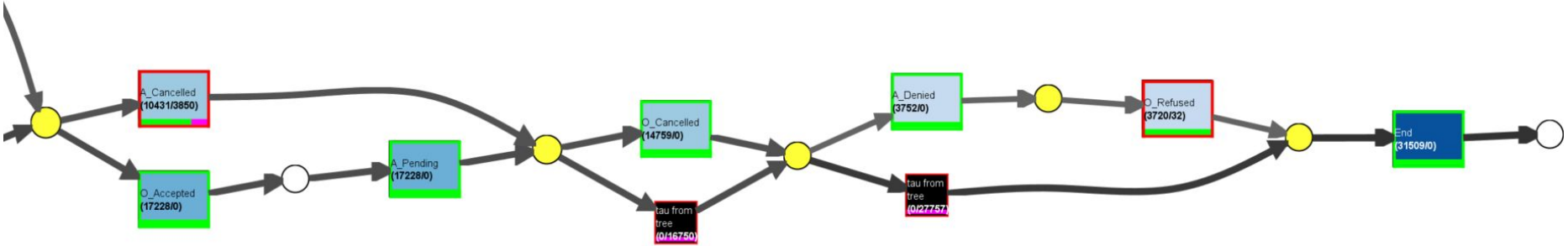
Finally, Replay a Log on Petri Net for Conformance Analysis is chosen for its accurate, visual method to identify process deviations, crucial for detecting potential fraud. The Event Name Classifier is selected, which ensures correct event-to-activity mapping, enhancing analysis accuracy. Additionally, the ILP- based replayer method offers precise, quantifiable insights into the fitness of the event log against the process model, optimizing deviation detection.



Petri Net discovered from the Conformance Checking. The image is cut to ensure readability.

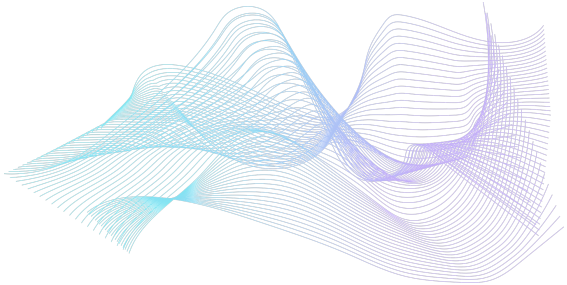
Conformance Checking (vi)

Finally, the only strange behaviour we can see in the process discovered by the conformance checking, is that the A_Cancelled activity is skipped 3850 during the execution time (purple colour).



This could be a case to analyze, however, observing the Petri Net, it can be deduced easily that this Process are finally cancelled, denied or just incomplete cases. Therefore, they would not constitute a case of fraud.

06



Conclusions

Conclusions

The analysis and implementation of process mining techniques for process discovery and detecting fraudulent credit card transactions have yielded several key findings and insights:

- **Process Discovery**: Thanks to the use of the tools provided by Disco, PM4Py and ProM, the process has been almost completely understood.
- **Fraud Detection**: Process mining facilitated the identification of anomalies and deviations within the transaction data. These deviations are critical indicators of potential fraud. However, all cases were eventually ruled out as fraud, by focusing on these anomalies, the system can flag suspicious activities that deviate from the norm, enhancing the overall fraud detection capabilities.
- **Actionable Insights**: The identified patterns and anomalies can be used to refine and improve existing fraud detection and prevention strategies. Financial institutions can implement more targeted measures to monitor and mitigate fraud based on the detailed transaction patterns uncovered.

Bibliography

1. Blevi, L., Delporte, L., & Robbrecht, J. (2017). Process mining on the loan application process of a Dutch Financial Institute. BPI Challenge. KPMG Technology Advisory, Bourgetlaan 40, 1130 Brussels, Belgium.

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

3. Jeong, D., Lim, J., & Bae, Y. (2017). BPIC 2017: Business process mining - A Loan process application. Department of Industrial and Management Engineering, POSTECH(Pohang University of Science and Technology), Pohang, Republic of Korea.

4. Scheithauer, G., Henne, R., Kerciku, A., Waldenmaier, R., & Riedel, U. (2017). Suggestions for Improving a Bank's Loan Application Process based on a Process Mining Analysis. metafinanz Informationssysteme GmbH, 80804 Munich, Germany.

5. Stoop, J.J., & Oezer, T. (2012). Process Mining and Fraud Detection.

6. Broer Bahaweres, R., Trawally, J., Hermadi, I., & Suroso, A. I. (2021, February). Forensic Audit Using Process Mining to Detect Fraud. Journal of Physics: Conference Series, 1779(1), 012013.

Thanks!

Do you have any questions?

CREDITS: This presentation template was created by **Slidesgo**, and includes icons by **Flaticon**, and infographics & images by **Freepik**