Chair of Information Systems and Business Process Management
TUM School of Computation, Information and Technology
Technical University of Munich

First Exercise - Business Process Prediction, Simulation, and Optimization

Data and Process Science on the BPIC-17 Event Log

First Author ✉

✉ moritz.hawener@tum.de

November 20, 2025

Abstract — This study analyzes the BPI Challenge 2017 event log containing loan applications from a Dutch financial institution to identify process inefficiencies. Process discovery revealed significant complexity with 15,930 variants, requiring extensive preprocessing to achieve balanced model quality metrics. Temporal analysis demonstrated process instability with significant throughput time variations across consecutive months. Clustering analysis identified four distinct process archetypes, showing that approval rates decrease as process complexity increases, from 75.0% for moderately complex cases to 66.9% for highly rework-intensive cases. The findings suggest that reducing unnecessary loops while maintaining structured engagement optimizes both approval rates and operational efficiency.

1 Introduction

The Business Process Intelligence (BPI) Challenge 2017 dataset originates from a Dutch financial institution that provides customer loans. As a continuation of the BPI Challenge 2012, the dataset captures the handling of loan applications from initial submission by customers, through evaluation and offer creation to the eventual acceptance or rejection of an offer. The financial institution wants to measure and compare throughput times in different parts of the process, investigate the impact of incompleteness of the data, analyze multi-offer behavior, and identify additional trends and dependencies that may reveal process inefficiencies [2].

Prior research by KPMG on this challenge revealed several patterns in the loan application process. The most critical bottleneck was found to be the waiting period after offers are sent, when customers need to decide or provide additional documents [1, p. 19]. In general this analysis demonstrated that proactive communication increases the conversion rate in the loan application process: calling clients about incomplete files not only reduced cancellation rates but also shortened processing times [1, p. 3], while presenting multiple offers to customers increased conversion rates compared to single-offers [1, p. 11]. Together, these findings highlight communication and customer engagement as key drivers of both process efficiency and successful outcomes.

2 Approach and Results

2.1 Technical setup and Preliminaries

In this analysis, the data was processed and analyzed using pm4py[1] and pandas[2], while matplot[3] was used to visualize. Additional I used the libraries sklearn[4] and scipy[5] for advanced analysis tasks. The final BPMN process model was created based on the previously discovered process models, reports, and common sense using SAP Signavio. The full implementation is available in the Github repository[6].

2.2 Basic Analysis

The event log covers 31,509 cases corresponding to loan applications, totaling 1,202,267 events. Across these cases, 15,930 distinct process variants were observed, reflecting a variety in how applications were handled. Each event includes timestamps, activity labels, and resource identifiers, along with other attributes such as offer IDs and loan characteristics.

The initial descriptive statistics (Table 1) reveal the structure and variability of the log. Case durations were calculated for all cases, showing the mean duration of approximately 22 days with a standard deviation of 13 days. Figure 1 highlights the wide variability among cases, with some cases taking longer than average. The chart clearly shows the spread of durations

[1] https://processintelligence.solutions/pm4py
[2] https://pandas.pydata.org/
[3] https://matplotlib.org/
[4] https://scikit-learn.org/stable/
[5] https://scipy.org/
[6] https://github.com/moritzhwnr/BPI2017.git

Statistic	Value
Number of cases	31,509
Number of events	1,202,267
Number of process variants	15,930
Number of case labels	31,509
Number of event labels	26
Mean case length	38.16
Std case length	16.72
Mean case duration	21d 21:35:25
Std case duration	13d 04:03:41
Number of categorical events	12
Number of traces with loops	31,509
Percentage of cases with >1 offer	27.16%

Table 1 Descriptive statistics of the event log.

and helps identify potential outliers that could distract the process analysis.

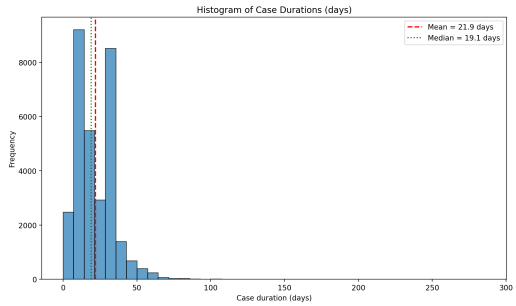

Figure 1 Case Durations (in days)

Furthermore the case length, measured as the number of events per case, was analyzed. With a mean of 38 and a standard deviation of 17 it shows a substantial variability. Figure 2 demonstrates that while many cases follow the average pattern, some cases are unusually long or short leading to irregular processes. Both duration and length histograms highlight the need for preprocessing.

Loops in the process were analyzed by counting repeated activities within each case. Figure 3 counts loops per case and shows that while most cases contain at least one loop, some cases contain multiple repeated activities. Mean and median lines were added to the chart to highlight typical looping behavior and to identify extreme cases that may need special attention.

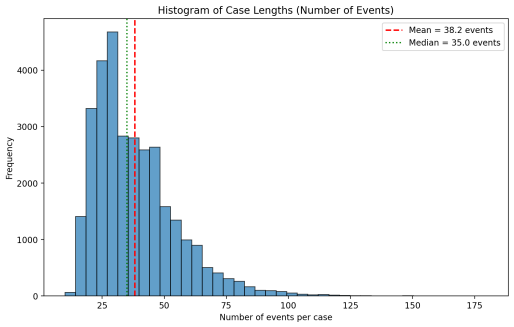

Figure 2 Case Length (in activities)

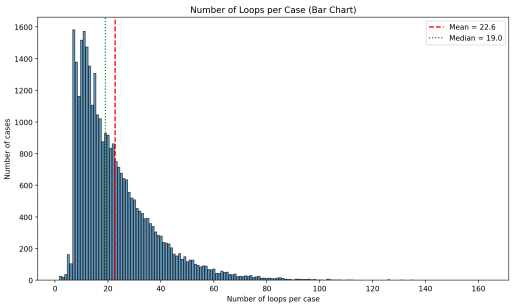

Figure 3 Loops in the processes

Lastly activity frequencies were examined to understand which activities dominate the process. A horizontal bar chart (Figure 4) of activity counts shows the most frequent steps in the process, providing a clear overview of the main operational tasks. Similarly, end activities were extracted and visualized as a horizontal bar chart (Figure 5), showing which activities most commonly conclude a case. These charts help identify rare or unusual activities that could be filtered during preprocessing.

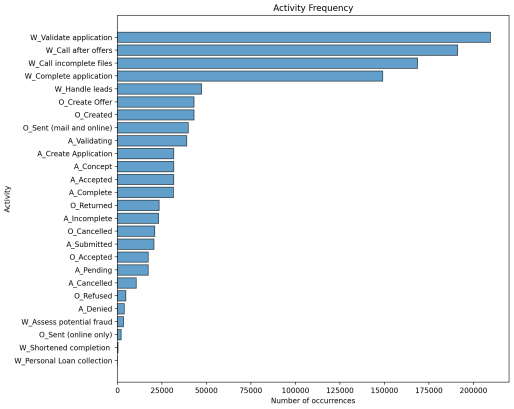

Figure 4 Case Activity Frequencies

Together, these statistics and visualizations offer an initial understanding of the process, highlighting potential outliers in case length and duration, iterative

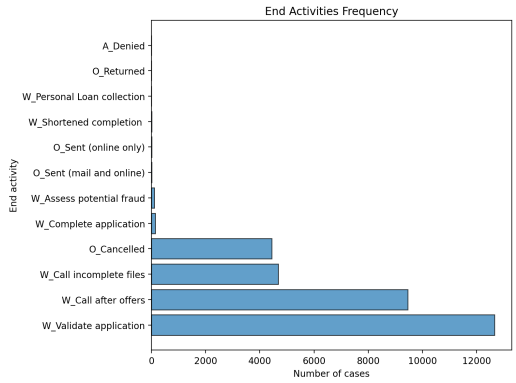

Figure 5 End Activitiy Frequency

loops, and activity frequencies. This shows the need for preprocessing to remove outliers.

2.3 Process Model Creation and Validation

To obtain an initial overview, the process was visualized using the inductive miner prior to any preprocessing actions. This initial model (Figure 6 shows significant complexity with numerous exceptional paths, revealing the presence of many outliers. While the model achieved perfect fitness of 100%, capturing all observed behavior, the precision was low at 14.06% (Table 6), indicating overgeneralization where the model allowed more behavior than actually occurred in the log.

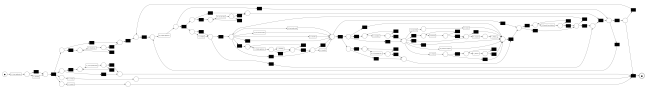

Figure 6 Initial Petri net (inductive miner)

Metric	Value	Metric	Value
Fitness	100%	Precision	14.06%
Simplicity	62.83%	Generalization	94.84%
Arc Density	0.9%	Control-Flow Complexity	10

Table 2 Metrics of Figure 6

The low precision combined with high control-flow complexity suggested that preprocessing was essential to extract a process model. By filtering out rare end activities and extremely long or short traces the variants were reduced from 31,509 to 25,398 cases. This initial filtering step led to small improvements in precision (16.50%), simplicity (63.41%), and generalization (97.5%), while maintaining near-perfect fitness of 99.99% (Table 7). However, the model remained complex with a control-flow complexity of 10, indicating that more aggressive filtering was necessary.

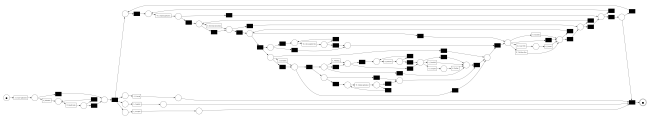

Figure 7 Petri net after initial filter (inductive)

Metric	Value	Metric	Value
Fitness	99.99%	Precision	16.50%
Simplicity	63.41%	Generalization	97.5%
Arc Density	1,2%	Control-Flow Complexity	10

Table 3 Metrics of Figure 7

An additional preprocessing step involved filtering based on the conditional probability of activity sequences, which identifies and removes unlikely activity transitions that likely represent outliers. This reduced the dataset from 25,398 to 22,837 cases. Applying a noise threshold of 0.11 within the inductive miner (Figure 8) achieved a result with 99% fitness and significantly improved precision of 50.37%. Additionally the control-flow-complexity decreased to only 4 (Table 4) suggesting a more interpretable model.

To explore alternative modeling approaches, the same preprocessed data was applied to the heuristic miner. The heuristic miner (Figure 9) produced a model with substantially higher precision of 76.81%, though at the cost of reduced fitness (94.22%). Interestingly, the control-flow complexity increased again to 10 (Table 5), while simplicity decreased to 58.73%.

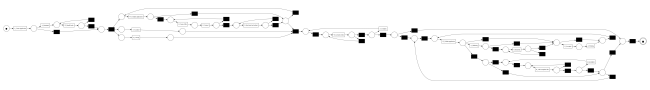

Figure 8 Petri net after conditional filtering (inductive)

Metric	Value	Metric	Value
Fitness	99.48%	Precision	50.37%
Simplicity	70.99%	Generalization	99.14%
Arc Density	1.29%	Control-Flow Complexity	4

Table 4 Metrics of Figure 8

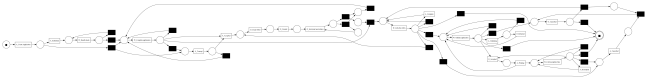

Figure 9 Petri net filter conditional filtering (heuristic)

Metric	Value	Metric	Value
Fitness	94.22%	Precision	76.81%
Simplicity	58.73%	Generalization	99.21%
Arc Density	1.82%	Control-Flow Complexity	10

Table 5 Metrics of Figure 9

These results reveal a trade-off between fitness and precision in process discovery. The inductive miner prioritized log fitness, capturing most behavioral patterns at the cost of overgeneralization, while the heuristic miner achieved higher precision by filtering noise but struggled with complete trace representation. The developed BPMN model (Figure 10) achieved a fitness of 92.83% and precision of 50.08%, positioning it between the two discovered Petri nets in terms of the fitness-precision trade-off. The simplicity score of 71.9% was notably higher than the heuristic miner result, and critically, the control-flow complexity reduced to 0, indicating a well-structured model without problematic control-flow patterns.

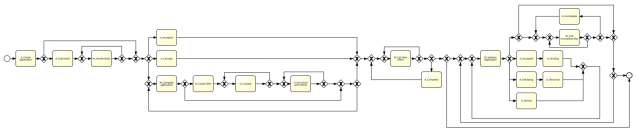

Figure 10 BPMN

Metric	Value	Metric	Value
Fitness	93.07%	Precision	50.08%
Simplicity	72.88%	Generalization	72.17%
Arc Density	1.37%	Control-Flow Complexity	0

Table 6 Metrics of Figure 10

2.4 Advanced Analysis

2.4.1 Concept Drift

Following the initial process discovery phase, an analysis was conducted to determine whether the loan application process shows temporal variations. Prior analyses indicated that the mean case duration was approximately 21 days and 21 hours, with a relatively low standard deviation of 13 days and 4 hours, suggesting a stable process. Despite this stability, it is important to examine whether small shifts in approval behavior or throughput time occurred across consecutive months. This could reveal underlying patterns or temporal variability. The hypotheses were formulated as follows:

- Null Hypothesis (H_0): The approval process and throughput time distributions have remained stable over time.
- Alternative Hypothesis (H_1): The approval process and throughput time distributions have changed significantly over time.

To assess potential drift, a concept drift analysis was performed, analyzing monthly approval rates alongside Kolmogorov-Smirnov (KS) tests to compare throughput time distributions between months. The analysis revealed that approval rates increased progressively over time, correlating with an increase in average loan request amounts, while throughput times exhibited month-to-month fluctuations (Figure 11).

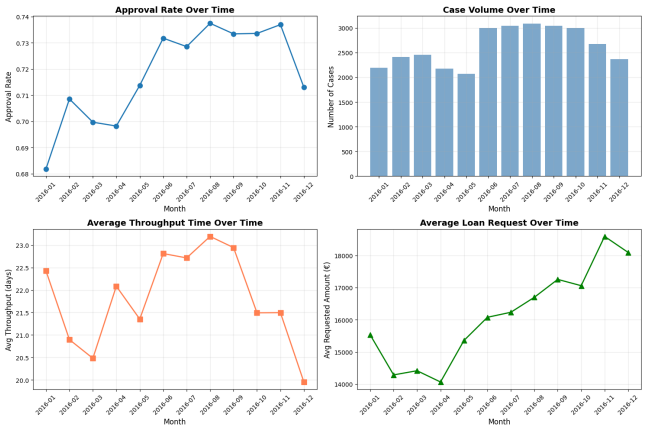

Figure 11 Concept Drift Statistics

The KS-test results supported these observations: applying a significance threshold of $p < 0.05$, ten out of twelve month-to-month comparisons demonstrated significant drift in throughput time distributions. This indicates that the process is not static and experiences measurable temporal changes (Figure 12.

These findings lead to the rejection of the null hypothesis, confirming that the loan approval process shows temporal instability characterized by variations in both approval rates and throughput time distributions across months. This represents a critical consideration for simulation modeling, as any accurate representation must account for these time-dependent variations in process behavior.

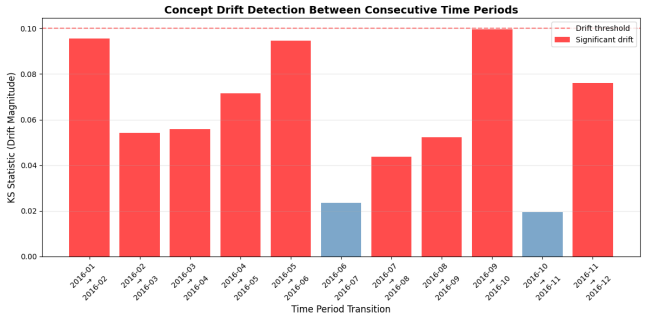

Figure 12 KS-Test

2.4.2 Clustering

Beyond temporal drift analysis, understanding whether distinct process archetypes exist within the loan application journey provides valuable insights into process heterogeneity. This analysis addresses whether certain variant clusters correlate with specific outcomes and whether "golden paths" can be distinguished from "problematic paths" marked by loops and rework. The hypotheses were formulated as follows:

- Null Hypothesis (H_0): Process variant clusters do not exhibit significant differences in approval rates.
- Alternative Hypothesis (H_1): Process variant clusters exhibit significant differences in approval rates, with complexity negatively correlated with approval success.

The four clusters exhibited different characteristics (Table 7 & 8). Cluster 0 represents a moderately complex path (46.27 events, 28.70 loops, 18.70 days throughput) with the highest approval rate at 75.0%, suggesting that iterative engagement contributes positively to success. Cluster 1 represents the "golden path" with minimal complexity (23.94 events, 11.42 loops) but high throughput time (33.31 days), reflecting customer-driven delays rather than process inefficiency, with a 70.6% approval rate. Cluster 2 exhibits rapid resolution (11.20 days) with moderate complexity (29.26 events, 14.28 loops) and 71.4% approval rate. Cluster 3 represents the most problematic archetype with extensive complexity (74.56 events, 56.43 loops, 33.29 days) and the lowest approval rate at 66.9%.

The observed pattern of decreasing approval rates with increasing complexity provides evidence against the null hypothesis. The 8.1% difference between the highest-performing Cluster 0 and lowest-performing Cluster 3 demonstrates that process complexity significantly impacts approval outcomes. This relationship between complexity and approval success results in two interpretations: complex cases may involve higher-risk applications requiring additional verification, or excessive iterations may cause customer disengagement, supporting KPMG's finding that proactive communication reduces cancellations [1, p. 3]. Additional throughput time doesn't correlate with trace complexity as e.g. Cluster 1 shows low complexity but high throughput. This indicates that calendar time is influenced more by external factors such as customer response times than intrinsic process complexity, supporting the detected *W_Call after offers* bottleneck identified by KPMG [1, p. 19]. These findings lead to the rejection of the null hypothesis confirming that process archetypes exhibit distinct approval rate patterns.

Cluster	Avg Len	Avg Unique	Avg Loops	Avg Throughput	Avg Events
0	46.27	17.58	28.70	18.70	46.27
1	23.94	12.52	11.42	33.31	23.94
2	29.26	14.98	14.28	11.20	29.26
3	74.56	18.13	56.43	33.29	74.56

Table 7 Cluster characteristics from k-means on TF-IDF encoded variants.

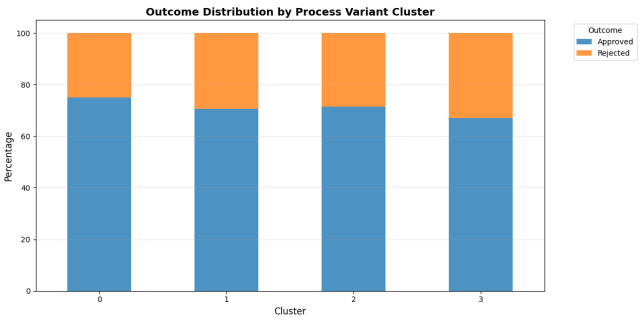

Figure 13 Approval and rejection rates per cluster

Cluster	Approved (%)	Rejected (%)
0	75.0	25.0
1	70.6	29.4
2	71.4	28.6
3	66.9	33.1

Table 8 Approval and rejection rates per cluster.

3 Disclaimer

In preparing this paper, I used DeepL to paraphrase sentences and Claude AI to assist with Python code debugging. All content, ideas, and analysis presented are my own work.

References

[1] Liese Blevi, Lucie Delporte, and Julie Robbrecht. Process mining on the loan application process of a dutch financial institute: Bpi challenge 2017. https://home.kpmg.com/be/en/home/insights/2017/09/process-mining.html, 2017. Accessed: 2025-11-16.

[2] BPI Challenge 2017 Organising Committee. Business process intelligence challenge 2017. https://ais.win.tue.nl/bpi/2017/challenge.html, 2018. Accessed: 2025-11-16.

A Process Models

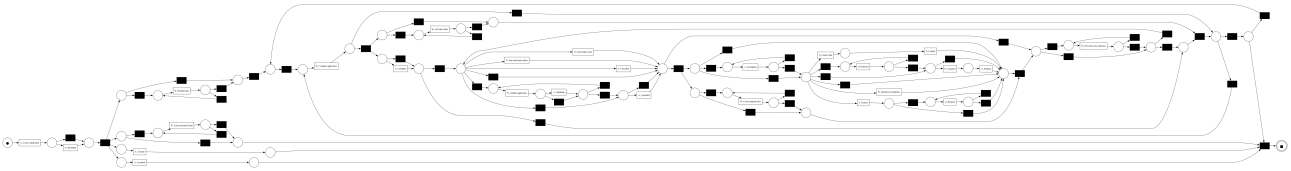

Figure 14 Initial Petri net discovered using the inductive miner.

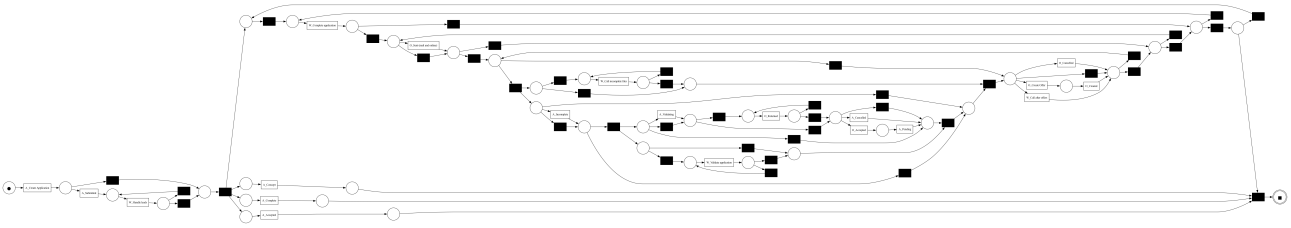

Figure 15 Petri net after applying the initial filter (inductive miner).

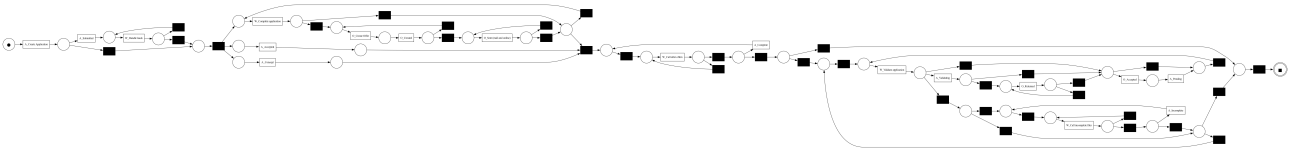

Figure 16 Petri net after conditional filtering (inductive miner).

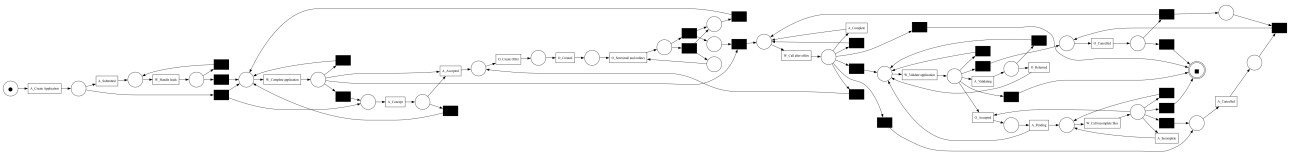

Figure 17 Petri net after conditional filtering using the heuristic miner.

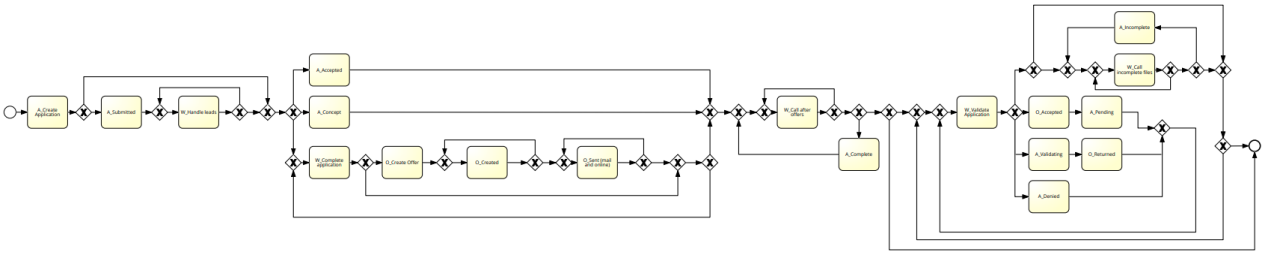

Figure 18 BPMN representation of the process.