**BPI2017 Challenge: Business process mining–A Loan process application Introduction- Roei Michael , Avichay Hayne and Asaf Koren**

Process mining, the practice of extracting knowledge from event logs recorded by systems, plays a pivotal role in comprehending and optimizing complex operational processes across various fields, including banking. This project, in particular, focuses on the application of process mining to the BPI2017 Challenge dataset. This dataset, made available by the BPI Challenge 2017, comprises various application traces within a banking environment. The project aims to glean insights, identify patterns, and propose enhancements that could potentially revolutionize the operational workflows of financial institutions.

The BPI2017 dataset originates from the system logs of a financial institution. Through a meticulous examination of this data, we aim to map/discover all process flows and investigate any possible inefficiencies. By focusing on the frequency of events, we anticipate identifying points of improvements, thus, enhancing the overall efficiency and effectiveness of banking operations. The project also entails searching for behavior patterns that might enable the institution to perform more in-depth analysis, suggesting changes, improvements, corrections, and learning from its processes.

In the contemporary economic environment, financial institutes face stiff competition, especially from emerging Financial Technology (FinTech) firms. One strategy to withstand this competition is by enhancing customer experience through the application of digitalization and automation techniques that streamline loan processes. Consequently, this project's goal is not only to analyze the loan application process, but also to focus on improving the customer experience and potentially increase revenue.

By addressing these aspects, our project seeks to answer critical questions raised by the process owners. These include queries related to throughput times for various parts of the process, the influence of the frequency of incompleteness on the final outcome, the number of customers requesting more than one offer, and the comparison of conversion between applicants who receive a single offer and those who receive multiple offers.

By intertwining process mining with the banking sector, we hope to unlock profound insights that could help financial institutions improve their operational processes, customer experience, and ultimately, their bottom line. Through this project, we wish to demonstrate the untapped potential and wide applicability of process mining in contemporary financial practices.

**Preprocessing**

The preprocessing phase aimed to prepare the dataset for a smooth and effective process mining activity. It was performed in a stepwise manner, systematically reducing the size and complexity of the dataset, while ensuring to maintain and highlight the most critical and informative parts of the process.

Trace Frequency Filter:

Our first step was to filter out all traces that occurred only once in the log file. The assumption behind this action was that infrequent traces, specifically those that occurred only a single time, were less likely to be representative of the common process pathways in our application system and hence, could be removed without significant loss of information.

This can be expressed as follows:

Let T be a set of all traces in the log file and n(t) be the number of occurrences of a specific trace t in T. Then, the reduced set of traces T' after this preprocessing step is given by:

Event Redundancy Filter:

We noticed through a visual inspection of the event log that some events seemed to contribute little to no new information about the processes. These included duplicated events and others that were not essential to the process. We removed such events, simplifying the event logs while still maintaining the overall process information.

Creation of Offer Dataset:

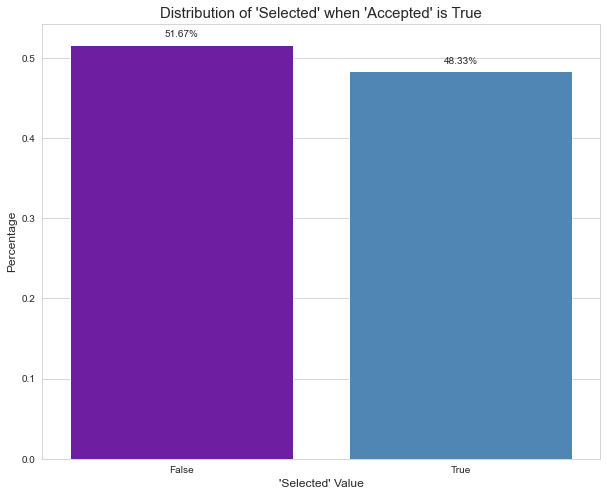
In order to enrich our understanding of the processes and enhance our analysis, we created a separate dataset focused on the offers part of the applications. This included important information such as the loan goal, requested amount, offered amount, and credit score. This dataset, while being valuable on its own, also served as an additional dimension to our primary event log data, allowing us to derive deeper insights about the process.

Additional Filtration Methods:

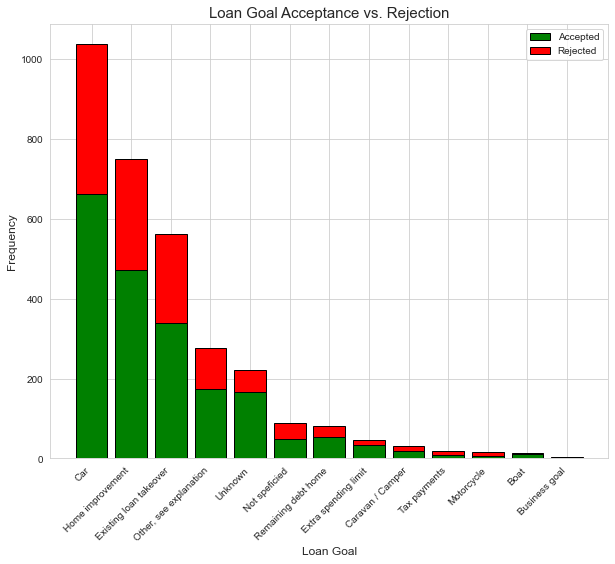
Finally, we devised and implemented several other filtration methods that were more oriented towards improving the visual analysis and exploration of the dataset. These included:

* **Top X Trace Filter**: This filter reduced the dataset to include only the top X most frequent traces, thereby focusing on the most common process paths.
* **Application Outcome Filter**: This filter selectively included only those traces that ended in either 'O\_accepted' or 'O\_rejected', i.e., the processes which resulted in a definitive outcome for the application.
* **Workflow Segment Filter**: This filter included only the events that occurred between 'W\_complete' and 'A\_accepted', thereby focusing on the main workflow of the application process.
* **Outliers filtration**: by using the Offers dataset we created we looked through the information about the offers created in the bank and filtered out outliers by examining columns like amount of offers, monthly payment, number of terms, etc.

In addition to the filtration we conducted on the data set we crafted a few graphs that helped us understand the data better:

the first analysis we had to do was showing the amount of accepted vs rejected offers :

From it we saw that we have almost a balanced dataset with a slight skewness to rejected offers,

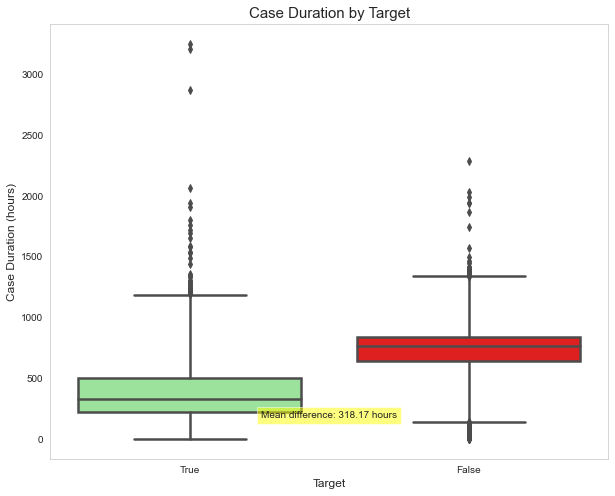
the next analysis we performed was based on the Loan Goal for each application , we measured the time and result of the application and shown them in a form of bar charts:

A chart with different colored bars

Description automatically generated

We noticed that most of the applications are related to cars and home while the time they take isn’t as  
less common loan goals, considering that we made an assumption that goals like remaining house debt might lead to a long trace with multiple events that is dragged across a lot of time and could cause bottlenecks.

we than looked at each case duration and compared them in regards to the way the offer ended.

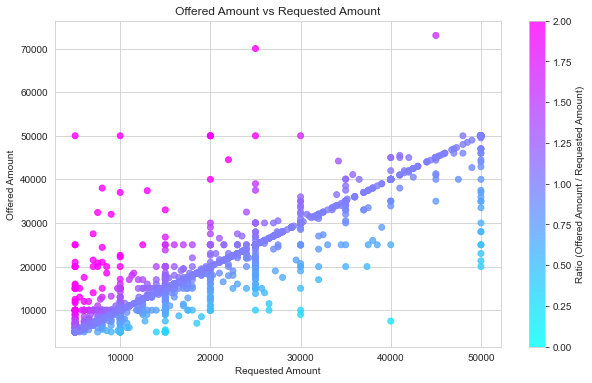


A graph with orange and blue bars

Description automatically generated

We saw from the graphs that across the months there wasn’t a real difference in rejected offers but there were a bit more accepting of offers in the summer months. Another thing to notice is the big gap in average time for case duration comparing between accepted and rejected offers, it did make sense that rejected applications probably got dragged out to multiple offers and we could consider applications with more than a single offer to be more time consuming for the bank and focus our analysis a bit more in that section.

A graph of a bar chart

Description automatically generatedWe than looked at the requested to offered amount in each offer to the bank to make predictions on how an offer is going to go based on the ratio of requested to offered amount:

A graph showing a number of bars

Description automatically generatedA graph of a bar chart

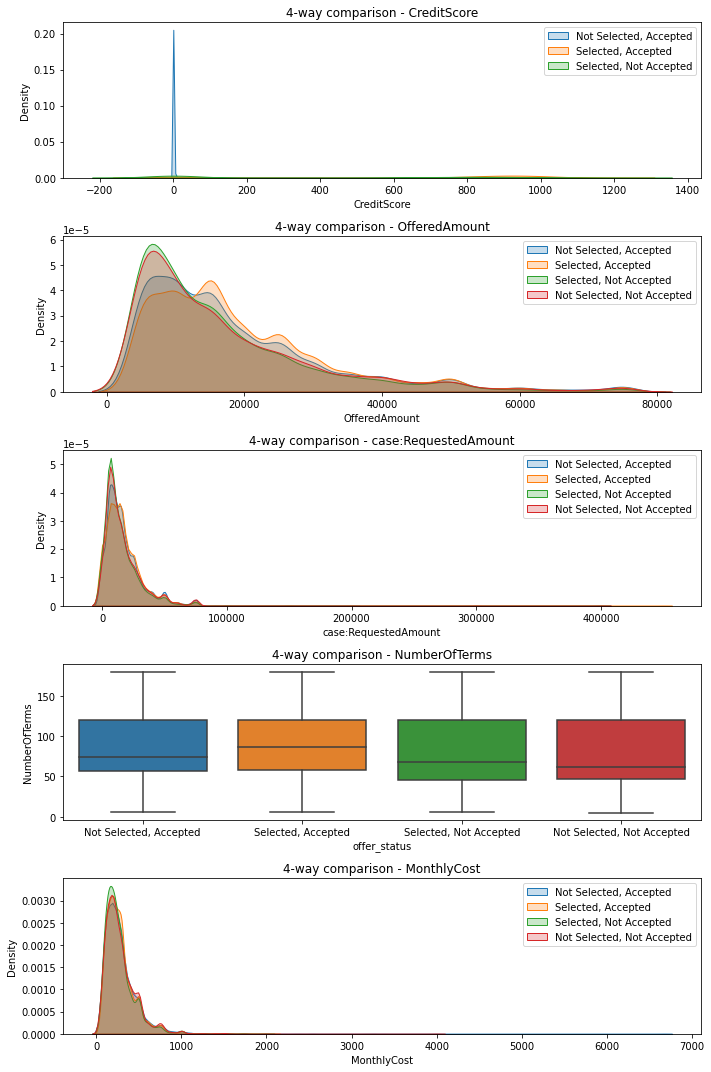
Description automatically generated

What we saw from these graphs is that most offers are met with the same amount of offer to requested basically giving us the linear purple line in the first graph, than we have 3 zones to analyze , when the offered amount is equal to the requested amount which we saw that ends up in accepted offers most of the time, and the next to zones are when the client and the bank are at a disagreement either the bank offers less or more to the client which led us to seeing how it affects the result of the application. Where when the offered amount is less than requested we see a rise in the percentage of rejected offers drastically.

The last analyzation we performed was separate our data to 4 “classes”:

* Accepted , Selected
* Accepted , Not Selected
* Not Accepted , Selected
* Not Accepted, Not Selected

\*\*Accepted- by the bank , Selected- by the client

And than comparing the offered amount , requested amount , number of months for payment and amount of money needed to be paid each month as follows:

We noticed that there wasn’t much of a change in the 4 classes and it seems that there isn’t a quick determination that could be done to early classify if the offer will be selected/accepted based on the early information like the offered and requested amounts.

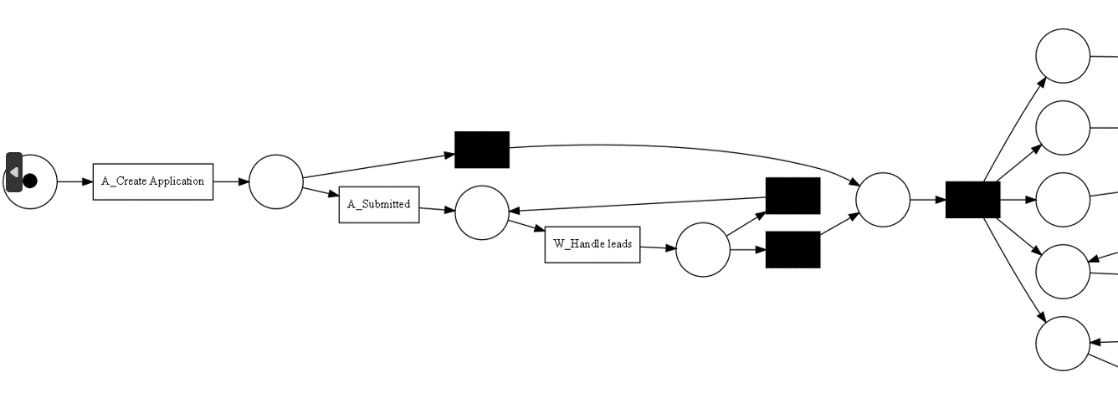
Overall, the preprocessing stage involved careful and systematic curation of the dataset, rendering it more manageable and comprehensible, while still capturing the essential dynamics of the process flows.

**Analysis Results**

Our analysis unearthed notable correlations and repeating patterns within the process, hinting towards potential areas for process optimization.

We used the filters mentioned in our preprocessing phase, in order to better handle and inspect various aspects of the dataset and the process. We divided our analysis into 3 main parts: Graphical representation and analysis of various aspects of the process using a representing data set, process mining discovery of selected datasets which describe different parts of the loan application process, and throughput analysis of frequently visited activities which we suspected to be bottlenecks.

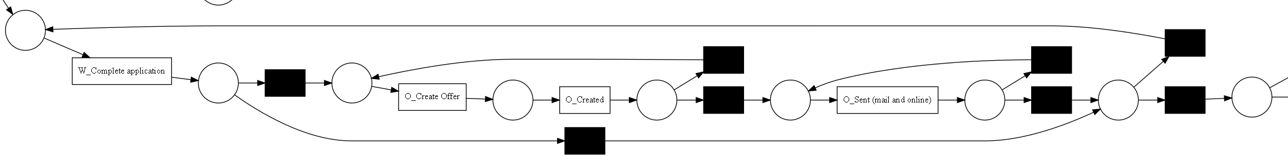
**Process petri-net building and analysis:**

**Part I:**In order to capture the essence of the loan process inside the organization we need to ask ourselves what the bread and butter of loans for the bank are? To answer that question, we took the top 50 variants of traces and built a petri net using the inductive miner, in order to understand the process to the fullest we have decided to explain the petri nets in sections:

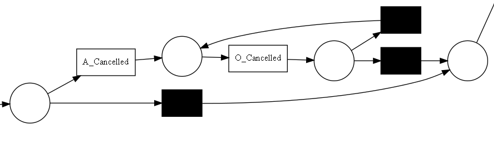
The first -

in this part we can see the application being processed inside the bank, the difference between the upper and the lower branches is in the way of submitting the application, for the upper branch the application was submitted via the bank itself, so a banker saw the application and submitted it himself. As for the lower branch the application was submitted online and there was a need for a banker to handle it. After the submission we can see that there are 5 conditions that has to be met in order to continue:

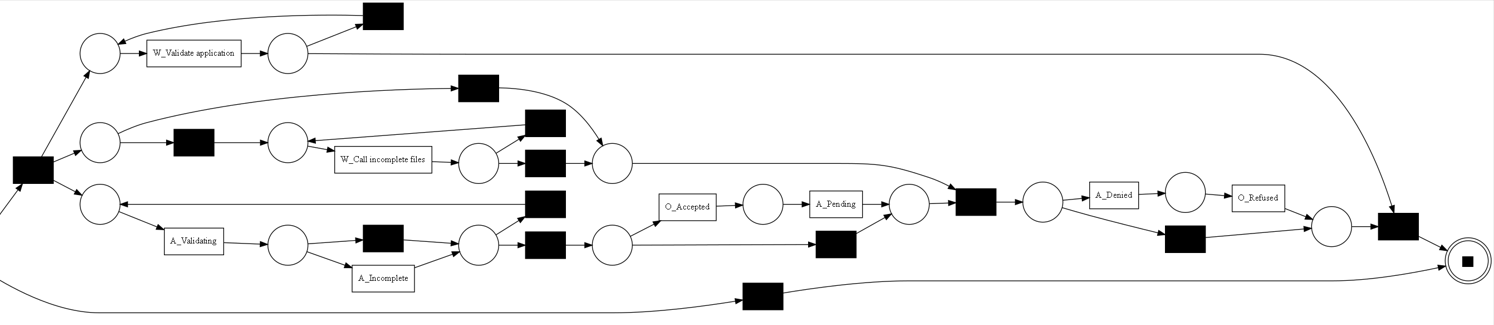
the first condition is for the bank to accept the application, the second is for the bank to create an initial concept for the loan. The third is the offer has been sent to the customer and the bank waits for the customer to send a signed offer and the rest of the required documentation. The fourth, A bank employee contacts the customer to follow-up on the offer. He/she gathers the thoughts of the customer about the offer. If the customer indicates he/she is not interested in the offer or does not answer the inquiries, both the offer and the application will be cancelled. And the fifth, for the actual offer to be created and or canceled:



Fifth branch



After those 5 conditions had been met there is now the second section of the process, if the application was canceled, we jump to the output instantly, otherwise we have a lot more to the process.   
there are again 3 criteria that need to be met for the process to finish, the first is to validate the application with the bank, the second is to call about incomplete files, the third is to validate the files, the second and third branches are combined into a single branch that decides the fate of the application. Whether it will be accepted of refused.



Second section

**Part II:**

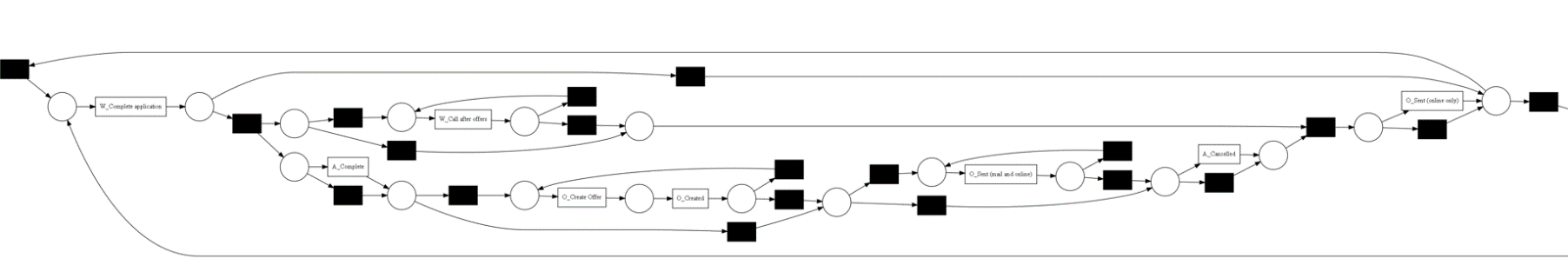
In this section, instead of discussing the essence of the process, its “ideal petri net”

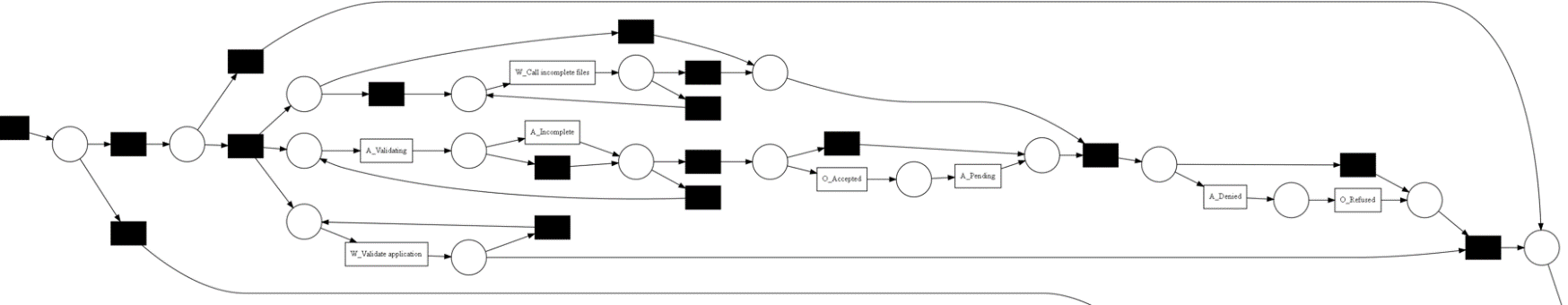
We chose to apply a more forgiving type of filters, first we have created a file that contained the information about each application without writing any additional info like events or type of events, from there we took built a normal distribution graph of the ‘mothlyCost’ and ‘numberOfTerms’ individually and then we took values that helped us take out outliers applications from the event log, finally we took the top 250 variance from the log file and only then we use the inductive miner. In the new Petri net, we can notice a few changes some are significant and some are less. The most notable change It's not in the events that are taking place inside the process but it's in the way that they are organized inside the Petri net. For example, in part one, we had two sections to our Petri net whereas in this Petri net It's more like a one long process rather than two separate sections we can see that instead of the five initial conditions from petri net 1, there are 3 primary conditions and one more branch if the application get canceled.

A diagram of a diagram

Description automatically generated

The place with the blue arrow pointed at it is the heart of the process because from there the actual process begins. Most of the process from the blue arrow onwards is the same as the fifth branch continued by the second section of the top50 variants with minor changes: more events are linear to each other rather than being parallel to one another. There are for example instead of calling to the client in parallel to finishing the application we now first finish the application and only then start communicating with him via phone calls.





The results of our filtering in combination with the inductive miner are remarkably good in our opinion:

Average trace fitness: 0.99775474.

Log fitness: 0.997390.

Percentage of fitting traces: 86.134.

**Data regarding mean and median duration of direct bank process parts**(in days)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| FROM | TO | MEDIAN DURATION | MEAN DURATION | INSTANCE FREQUENCY |
| A\_Concept | A\_Accepted | 0.86 | 1.41 | 31.5 |
| A\_Validating | O\_Accepted | 0.25 | 0.91 | 7.14 |
| O\_Returned | O\_Accepted | 1.09 | 2.05 | 5.29 |
| O\_Returned | A\_incpomlete | 1.08 | 2.00 | 13.76 |
| A\_incomplete | O\_accepetd | 3.70 | 5.90 | 4.78 |

The analysis of the loan application process specifically the bank related procees parts, reveals valuable insights and conclusions. Firstly, the examination of the median and mean durations indicates the presence of exceptional cases with longer durations, resulting in a higher mean value. Thus, the median duration is considered a more robust measure for further analysis.

Secondly, specific stages within the loan application process exhibit notable waiting times. For instance, optimizing the A\_Concept -> A\_Accepted process segment can significantly impact the overall throughput time due to its high frequency. Additionally, distinguishing between applications submitted through different channels, such as the website or in-person at the bank, highlights the need for expedited processing to minimize waiting periods.

Thirdly, implementing strategic measures can effectively address the waiting time issue. Exploring options like staffing enhancements for offer creation or adopting automation and AI technologies for streamlined processes and acceptance checks can yield substantial reductions in waiting times. Furthermore, incorporating systems to ensure the submission of complete documentation and adherence to acceptance criteria can further expedite the process.

By leveraging these insights, financial institutions can enhance the loan application process through targeted interventions such as optimizing waiting times, leveraging automation and AI, and emphasizing complete documentation. This can result in improved customer satisfaction, streamlined operations, and faster loan processing.

**Data regarding activities where the bank waits for a client response(in days)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| FROM | TO | MEDIAN DURATION | MEAN DURATION | INSTANCE FREQUENCY |
| A\_Complete | A\_Validating | 7.1 | 8.8 | 31.509 |
| A\_Incomplete | A\_Validating | 1 | 2.5 | 10.504 |
| A\_Complete | A\_Cancelled | 30.3 | 27.7 | 8.034 |
| A\_Complete | O\_Create Offer | 4.1 | 7.2 | 4.135 |

When analyzing the loan application process, two process segments stand out for further improvement: the segment involving completion of the application (A\_Complete) and subsequent validation by the bank (A\_Validating).

In the first process segment, A\_Complete - A\_Validating, the median duration is 7.1 days, with an instance frequency of 31,509. This indicates that once the bank has sent the offer to the applicant, there is a considerable waiting period for the applicant to sign the documents and send them back. To enhance this segment, the bank can implement measures to expedite the document signing process, such as providing clear instructions and reminders to the applicant. Streamlining communication channels and offering convenient options for document submission can significantly reduce the waiting time in this segment.

The second process segment, A\_Incomplete - A\_Validating, involves cases where the bank is waiting for the applicant to submit missing documents for validation. With a median duration of 1 day and an instance frequency of 10,504, it presents an opportunity for improvement. To minimize delays, the bank can implement proactive measures, such as automated reminders and a streamlined document submission process. Clear communication regarding the required documents and their submission deadlines can ensure a smoother validation process and reduce waiting times.

While these process segments may have relatively shorter waiting times compared to others, optimizing them can have a significant impact on the overall efficiency of the loan application process. By focusing on improving communication, streamlining document submission, and implementing proactive measures,the bank can enhance customer satisfaction, reduce waiting times, and create a more seamless application experience.

Discussion:

The analysis of the loan application process using process mining techniques has provided valuable insights into the operational workflows of financial institutions. Through the application of various preprocessing steps and filters, we were able to curate the dataset and extract meaningful patterns and correlations. These findings have significant implications for process optimization and enhancing the customer experience in the banking sector.

One key finding of our analysis is the presence of inefficiencies and bottlenecks within the loan application process. By examining the throughput times of different process parts, we identified stages such as A\_Concept to A\_Accepted and A\_Complete to A\_Validating as areas that can be improved to reduce waiting times and enhance overall efficiency. These insights suggest the need for targeted interventions, including staffing enhancements, automation, and streamlined communication channels, to expedite these process segments.

Moreover, our analysis revealed the influence of certain factors on the loan application outcomes. By examining the requested to offered amount ratio, we observed that applications with a lower offered amount compared to the requested amount had a higher likelihood of being rejected. This finding emphasizes the importance of aligning the offered amount with the customer's expectations to increase the chances of acceptance.

Additionally, we explored the durations associated with the bank waiting for client responses. Notably, the A\_Complete to A\_Validating process segment exhibited a considerable waiting period, indicating a potential area for improvement. Implementing measures to expedite document signing and submission, as well as clear communication regarding required documents and deadlines, can help minimize delays and enhance the applicant's experience.

In conclusion, our analysis of the BPI2017 Challenge dataset using process mining techniques has provided actionable insights for enhancing the loan application process in the banking sector. By addressing inefficiencies, optimizing waiting times, and aligning the offered amount with customer expectations, financial institutions can streamline their operations, improve customer satisfaction, and increase the efficiency of loan processing. These findings demonstrate the untapped potential of process mining in revolutionizing operational workflows in contemporary financial practices.

Conclusions:

Through the application of process mining techniques on the BPI2017 Challenge dataset, we have gained valuable insights into the loan application process within the banking sector. Our analysis has highlighted areas of improvement and provided actionable recommendations for optimizing the operational workflows and enhancing the customer experience.

By utilizing various preprocessing steps and filters, we curated the dataset to focus on relevant process flows and extract meaningful patterns. The derived Petri nets and process mining results demonstrated strong alignment with the observed data, validating their representativeness and reliability. This allows financial institutions to make informed decisions and allocate resources effectively for process optimization.

The analysis revealed specific process segments that exhibit notable waiting times, such as the completion of the application and subsequent validation by the bank. By implementing strategic measures, such as streamlining communication channels, automating reminders, and improving document submission processes, financial institutions can significantly reduce waiting times and expedite the loan application process.

Furthermore, our findings highlighted the influence of factors such as the requested to offered amount ratio on the loan application outcomes. By aligning the offered amount with the customer's expectations, financial institutions can increase the likelihood of acceptance and improve conversion rates.

Overall, this project demonstrates the untapped potential of process mining in revolutionizing operational processes in the banking sector. The insights derived from our analysis provide a roadmap for financial institutions to optimize their loan application processes, enhance customer satisfaction, and drive operational efficiency.

By leveraging the power of process mining, financial institutions can stay competitive in the contemporary economic environment, particularly against emerging FinTech firms. The findings from this project underscore the importance of digitalization, automation, and customer-centric approaches in streamlining loan processes and ultimately increasing revenue.

In conclusion, process mining offers a comprehensive and data-driven approach to uncovering inefficiencies, identifying patterns, and proposing enhancements in the loan application process. Through our project, we have demonstrated the applicability and value of process mining in the banking sector, paving the way for future advancements in operational optimization and customer experience.

**References**

List all the resources you used or referred to in your project.

1. Gregor Scheithauer, Roman Henne, Arled Kerciku, Robert Waldenmaier, Ulrich Riedel. "Suggestions for Improving a Bank's Loan Application Process based on a Process Mining Analysis." metafinanz Informationssysteme GmbH, 80804 Munich, Germany.

2. Liese Blevi, Lucie Delporte, Julie Robbrecht. "Process mining on the loan application process of a Dutch Financial Institute BPI Challenge 2017." KPMG Technology Advisory, Bourgetlaan 40, 1130 Brussels, Belgium. URL: [https://home.kpmg.com/be/en/home/insights/2017/09/process-mining.html]

3. PM4Py. "PM4Py Documentation." URL: [https://pm4py.fit.fraunhofer.de/documentation]

4. Parisa Khademolhossieni. "BPI Challenge 2017." Prezi presentation. URL: [https://prezi.com/8pa\_nlaoikol/bpi-challenge-2017/]