NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Business Cases with Data Science

Case 4: Business Process Conclusion Prediction

David Psiuk	number: 20230818
Peter Falterbaum	number: 20230956
Luis Penteado	number: 20230441
Noah Campana	number: 20230996

Group A

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

May, 2024

Table of Contents

1 Executive Summary 1

2 Business Needs and Required Outcome 2

2.1 Business Objectives 2

2.2 Business Success Criteria 2

2.3 Situation assessment 3

2.3.1 Business History 3

2.3.2 Current Situation 3

2.4 Determine Data Mining Goals 4

3 Methodology 5

3.1 Data understanding 5

3.1.1 Business Process 5

3.1.2 Data Exploration 5

3.2 Data preparation 7

3.2.1 Data cleaning 7

3.2.2 Feature Engineering 9

3.2.3 Splitting the datasets 9

3.3 Modeling 10

3.3.1 Model Selection 10

3.3.2 Feature Selection 10

3.4 Evaluation 11

4 Results Evaluation 11

5 Deployment 12

5.1 Next Steps 12

5.2 Proposed Implementation Timeline 12

6 Conclusions 13

7 Bibliography 14

Appendix 15

1 Executive Summary

Millennium bcp, Portugal's largest private sector bank, is actively enhancing its operational efficiency and service quality through strategic predictive process monitoring. This initiative strategically targets optimizing business processes, focusing on reducing unnecessary procedural steps, improving sequence flows and minimizing lead times.

In 2023, the bank showcased significant resilience and strategic foresight, underscored by substantial upgrades in its credit ratings and a remarkable 124% increase in its stock price and now wants to invest into business intelligence services.

The bank's process dataset is collected from April 2022 to May 2024, encompassing 45,722 individual processes. This analysis is designed to identify inefficiencies and predict four potential outcomes for each process step. The objective is to significantly reduce delays and associated costs, as they currently missed their deadlines accumulated by 500 years across all processes in the last 1.5 years, offering an opportunity to save on 6.6 million € in labor cost.

The models employed include Decision Tree, Random Forest, and XGBoost, which have been evaluated on the basis of accuracy and weighted F1 score. The 4 different classifiers were benchmarked against each other, while Random Forest and XGBoost tend to outperform Decision Trees. Further refinement was applied to the best preforming models. Best weighted F1 Score was achieved with a Random Forest.

The implementation strategy for integrating these predictive models is methodically structured into a five-phase approach. This begins with setting up the necessary environments and extends to a full-scale rollout, encompassed with continuous performance monitoring. Each phase is designed to integrate seamlessly into the bank's existing systems, allowing for real-time data processing and enabling feedback loops for continuous improvements.

Our results indicate a significant potential to improve the process by accurately predicting the process outcome. This enhancement is expected to significantly boost Millenium bcp’s efficiency, demonstrating the value of the proposed data-driven methods and currently saves 3.6 mil. € in labor cost.

.

2 Business Needs and Required Outcome

2.1 Business Objectives

When it comes to Millenniums business objectives in the context of value creation through process optimization there are 5 categories:

Process step-related: Process step-related value creation potentials exist where a process contains unwanted, missing or repetitive steps. For example, the given process exhibits repetitive tasks like first 'Analyse and Resolve' a case and then 'Analyse accounting impact'. This way two different persons need to understand the request and invest time into it, rather than just one person doing both.

Sequence-related: Sequence-related value creation potentials are those in which the steps of the process are carried out in the wrong order. For example, if the process would seek payment approval before the corresponding compliance verification steps.

Lead time related: Throughput time-related value opportunities arise when the throughput time between process steps is too long or too short. For example, if it takes too long for third parties to complete their tasks.

Automation-related: A process that has the potential for greater automation represents an automation-related value opportunity. For example, a system that is configured to require manual communication with the client represents an automation-related value opportunity.

Attribute-related: Attribute-related value creation opportunities are those that are linked to a specific case characteristic. For example, you can discover a value creation opportunity that is only displayed in cases that relate to a specific third party or service area [1].

The business, however, decided to create value through predictive process monitoring, which can be associated with the field of lead time and automation-related value creation. Further to this, predictive process monitoring can be divided into three categories: remaining time predictions, next event predictions and outcome predictions. All these types of prediction would assist Millennium in enhancing their processes. However, the business has chosen to concentrate on the latter.

Concluding the objective is to predict the outcome of the process at specific stages, thereby reducing the time and money spent on incidents that should not require significant resources.

2.2 Business Success Criteria

As stated before, the main goal is to drive significant improvements across various dimensions of the given business processes. The implementation of a machine learning model to predict expected outcomes can significantly enhance various aspects of a business process by addressing key value creation potentials. These value creation potentials are process-related inefficiencies, which can be reduced by optimizing lead times and identifying automation opportunities, as well as leveraging insights specific to attributes. By doing so, Millennium bcp will be able to achieve higher levels of efficiency, reduced costs and an improved level of service quality.

2.3 Situation assessment

2.3.1 Business History

Millennium bcp, formally known as Banco Comercial Português, S.A., is Portugal's largest private sector bank. The bank was established on 17 June 1985, following the deregulation of the Portuguese banking industry. It was founded by a coalition of over 200 shareholders and seasoned banking professionals, to serve the Portuguese financial market more effectively.

Throughout its history, Millennium bcp has experienced significant growth, initially through organic expansion and subsequently through strategic acquisitions. One of the most notable acquisitions was the purchase of Banco Português do Atlântico in 1995, which was at the time the largest private sector bank in Portugal. Since then, the bank enjoys a strong presence in cultures which are historically close to Portugal, particularly Angola, Mozambique, the United States, Canada, France, Luxembourg and Macau (China).

In line with its strategy to concentrate on core business areas, Millennium bcp underwent significant restructuring in the early 2000s. This included the sale of its non-life insurance businesses to Ageas (formerly Fortis) in 2004, which allowed the bank to refocus on its primary financial services.

Further strategic refocusing led to the divestment of several international operations, including those in France, Luxembourg, the United States, Canada, Greece, Turkey, and Romania, although the bank retained commercial protocols to facilitate remittances from Portuguese emigrants in some markets. This move was part of a broader restructuring plan initiated in 2012, which included a management overhaul introducing a one-tier management and supervisory model, complying with EU regulations while reducing costs.

Recent developments in the bank's operations include the 2016 merger of Banco Millennium Angola with Banco Privado Atlântico, which resulted in the creation of Angola's second-largest private sector bank. Further consolidations included the incorporation of Banco de Investimento Imobiliário, S.A. into Millennium bcp in 2019, and the merger of Bank Millennium S.A. with Euro Bank S.A., which consolidated its operations under a single brand in Poland. In 2021, the bank sold Banque Privée BCP (Suisse) SA to Union Bancaire Privée, aligning with its strategy to focus on core geographies [2].

2.3.2 Current Situation

In 2023, Millennium bcp demonstrated remarkable resilience and strategic foresight despite a challenging macroeconomic landscape. The bank's primary focus was on strengthening its financial stability, enhancing operational compliance, and focusing on the bank main revenue streams, namely retail banking. This was acknowledged through a series of upgrades by major credit rating agencies, i.e. DBRS Morningstar, Moody’s, Fitch Ratings and S&P Global Ratings. Millenium bcp also optimized its capital structure by repaying debt ahead of schedule, which resulted in the bank having more working capital and better compliance with EU regulations.

In conclusion, 2023 was a year of strategic triumphs for Millennium bcp, marked by significant upgrades in credit ratings, enhanced regulatory compliance, and recognition as a leader in sustainable and innovative banking practices [2].

All these developments led to the very positive evaluation of the company, with a performance of approximately 124% increase in the company's stock price, as seen in Figure 1 [3].

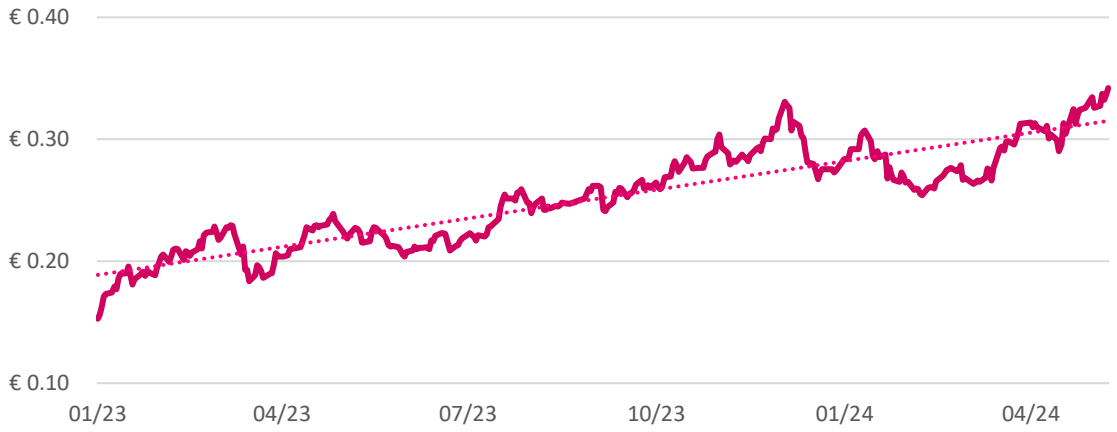

Figure 1: Banco Comercial Português stock price

2.4 Determine Data Mining Goals

By analyzing historical data we aim to find patterns that can lead to the possible outcome faster, increasing efficiency and effectiveness.

Regarding process step-related improvements identifying unwanted, missing, or repetitive steps, the model could provide a straightforward call to action outputs, thus saving time and resources.

Throughput time is critical in business processes. A machine learning model can predict and highlight steps where the lead time is either too long or too short. By identifying bottlenecks where third parties delay the process, a prediction of the outcome could speed up this process by only requiring validation instead of evaluating the process from scratch. Alternatively, if the model is sufficiently confident in its predictions, it may be possible to skip certain activities and provide the outcome to the process at an earlier stage than anticipated.

By integrating a machine learning model to predict expected outcomes, businesses can substantially enhance their process efficiency and effectiveness. This technology provides a powerful tool for identifying and eliminating redundant or missing steps, reducing unnecessary delays, which decreases Millennium's costs for unnecessary process steps.

3 Methodology

3.1 Data understanding

Understanding the provided Data such as the underlying process is crucial for the implementation of the predictions. 4 different datasets were provided consisting of task execution data, user information, specific request data and rejections.

3.1.1 Business Process

The topic of the specific Business Process was not shared due to confidential reasons. However, a replication of this process can be seen in Figure 2. The process consists of a starting point with 4 different outcomes: 'request cancelled', 'request finished', 'closed administratively requester rejects accounting impact' and 'closed administratively'. Every process starts with the activity 100 and continues either with 102 or 105. Afterwards there are several options in how the instance can proceed depending on which actions were taken during the process.

3.1.2 Data Exploration

3.1.2.1.1 Task execution data

The execution dataset represents the main table which consists of logs of the different process steps with a granularity of seconds. It links the additional datasets through foreign keys and the dates for the task arrivals range from 11.04.2022 to 7.05.2024. Furthermore, it consists of 45722 individual processes with 9 unique activities and 19 unique actions. The data provides detailed information about the instances which run through a chosen process of Millenium.

From the 9 unique activities 102, 104 and 100 are the ones with the highest frequency. 105, 108 and 106 have a relatively low frequency compared to the others. The overall frequency ranges from around 100 occurrences to 55000. The action with the most occurrences is 270 and leads to the activities 102 and 105.

The column Task Type consists of 4 different categories with the highest frequency for execution, followed by initial request, final task and requester response to rejection. The category final task is misleading since it does not represent the actual final task in a process.

3.1.2.1.2 Execution and Capturing Time

Some tasks include data regarding a timeframe in which they should be executed. Taking a deeper look at the data reveals that around 62% of the tasks are executed in time while 38% are executed with delay. When looking at the median we can determine that the execution time for all tasks are less than 11 minutes, whereby activity 105 represents the highest execution time. The average reveals a different distribution with significantly longer execution times indicating several activities with more than 1000 minutes (ca. 16,6 hours) and for activity 104 even 84 days.

The capture time represents the difference between the task arrival date and the capture date, thus the period in which a task is waiting to be executed. Utilizing the median, the capture time for different activities range up to 6 days (ca. 160 hours) for activity 108. The

high capture time for this activity might be caused by the dependency of third-party information. During this step Millenium possibly waits for a police report or similar information without exhibiting sufficient influence in that process. The average capture time peaks for activity 104 with approximately 39 days (ca. 940 hours). Supposedly in this activity the reimbursement of a fraud or similar issues get defined. Therefore, the different cases might need exceeding discussions with employees from different departments causing an increased delay in capturing time. The different scale for the two metrics illustrates the high quantity of outliers present in the dataset regarding the time component. Generally, it can be observed that the capture time tends to be significantly longer than the execution time leading to a high potential of increasing Millenium´s overall Task execution by optimizing capture processes and streamlining workflow efficiencies.

The total overtime accumulated over the entire timeframe in all processes is an incredible 499.8 years, calculated by subtracting the predicted end time from the actual execution time. Averaged across the entire workforce, the average bank worker earns approximately €1,000 per month [4]. Additionally, the company is required to pay an additional 11% in social security contributions [5], resulting in an average monthly cost of €1,100 per employee. As the process delay also includes delays that occur on weekends or holidays, this can be equaled out by simply taking the monthly salary, which already accounts for weekends and holidays. Consequently, multiplying 1,100 euros by 12 by 500 years yields a total of 6.6 million €, which represents the expenditure incurred due to the non-achievement of deadlines.

3.1.2.1.3 Request Arrival and Path Analysis

Most instances initially arrive on Mondays, while Saturdays and Sundays barely any requests occur. When looking at the initial task arrivals per month, a low can be witnessed in April and a peak in august. Moreover, in the beginning and end of the given dataset the occurrences of tasks are significantly low with almost 0 frequency.

Most processes end with activity 107, followed by 104 and 101. Requests that end with the activities 102 and 105 are misleading since these processes cannot end at this stage considering the provided business process. Accordingly, these cases will be further excluded. The 10 most frequent activity sequences can be seen in Figure 3. Apparently the most occurring process starts at activity 100, followed by 102, 103 and finishes with activity 104. Accordingly, the most frequent sequences by action can be seen in Figure 4.

3.1.2.1.4 Target Analysis

The label “closed administratively requester rejects accounting impact” shows the highest frequency with almost 20000 cases followed by “closed administratively” and “request finished”, while “request cancelled” is the least occurring with around 1700 cases.

The distribution of early and delayed executions per outcome is demonstrated in Figure 5. One can observe that the outcome “closed administratively requester rejects accounting impact” is the only target which has on average more delayed than in-time executions. The ratio of early terminated executions is relatively high for the label “request finished”, suggesting a more effective workflow for processes ending in this outcome.

When analyzing the most frequent sequences per label as seen in Figure 6 and Figure 7, it is possible to detect important characteristics. For the 3 most frequent processes, the outcomes “closed administratively” and “request finished” share the identical processes and only differentiate each other by their final application action which is 288 for the first and 299 for the latter. Similarly, “closed administratively requester rejects accounting impact” shares the identical 3 most occurring sequences with the process finishing one step earlier at task 104. The label “request cancelled” shows a different distribution of frequent sequences.

Examining the average execution and capture time per target reveals that processes which terminate in “closed administratively requester rejects accounting impact” are significantly slower. These processes show an average of more than 80 days for the capture time and almost 200 days for the execution time. The other outcomes reveal execution and capture times which range from around 2 to 9 days. Moreover, these outcomes are characterized by higher capture times compared to execution times.

In Figure 8 the frequency of different labels over time is displayed based on the first occuring task of the process. Interestingly we can observe that after august 2023 the arrivals of tasks which end with "closed administratively requester rejects accounting impact" drastically drop and show no more frequency since October 19th of the same year.

3.1.2.1.5 Additional datasets

The user information dataset contains data about the different task executors such as their sex, birthyear, if outsourced or not and since when the executor is in the company. Around 64% of the tasks were executed by employees inside the company. The specific request dataset is linked to the executor dataset via the "Request Identifier" column. It also includes a "Value" column, the details of which remain confidential. Since the project's objective is to predict outcomes in various scenarios without the use of future data, the use of this dataset is problematic. Specifically, it's unclear at what stage in the process the values in the "Value" column are recorded, which compromises their reliability for our predictive analysis. Therefore, this dataset will further not be utilized. Analyzing the Rejections Dataset did not reveal substantive impacts on the outcome of the process which is why this dataset will be ignored for further analysis and predicting.

3.2 Data preparation

3.2.1 Data cleaning

3.2.1.1.1 Task execution data

As part of the data preparation process, we addressed the presence of null values across several columns within the dataset. Our handling strategies are detailed below to ensure transparency and reproducibility of our data cleaning steps.

“Task predicted end date” represents the time in which a task is supposed to terminate. It contains 45,785 null values. We opted to retain these as Not a Number (NaN) values, rather than converting them to zero, to maintain integrity for subsequent feature engineering steps. Moreover, several instances were identified where predicted end dates occur before the task

arrival date. To address these discrepancies, we calculated the most common timeframe for predicted end dates associated with each activity and applied it accordingly.

The column “Task Executor” has several instances with missing values, although corresponding 'Task Executer Department' data might be available. This typically indicates that the executor is no longer with the company. To handle this, we will assign a placeholder value of 9999 to represent such cases, acknowledging their previous contributions to the tasks. After applying this fix, 14,106 missing values remain unassociated with any discernible patterns linked to other columns. Given the substantial number of these cases, we will represent these remaining nulls with a distinct numeric value (0) after confirming its non-usage for other executors within the dataset. This approach allows us to preserve potential patterns or insights related to these data points.

There are 27,130 missing entries in the 'Task Executer Department' column, spread across 391 distinct departments. For instances where the task executor is known but the department isn't, we will impute the missing department data using the most recent department associated with the executor. The remaining missing values will be replaced with a placeholder value (0) to maintain data consistency and utility in subsequent analyses.

The 'Action' column, which corresponds to the content described by the 'idBPMApplicationAction,' often cannot be shared due to confidentiality constraints. Therefore, we will substitute missing values in this column with the term "Confidential." This ensures that the dataset remains informative while respecting privacy and confidentiality requirements.

3.2.1.1.2 User Information

In this dataset few instances were observed where the recorded enrollment date of an executor preceded their birth date. Such entries are logically impossible and were consequently removed from the dataset to prevent potential distortions in subsequent analyses. Furthermore, entries representing exceedingly young or unusually old individuals were identified, specifically those younger than 18 and older than 89. To align with realistic operational capacities, entries falling outside this age range were removed.

Several instances of duplicated task executors were found, each showing different starting dates for their organizational units. To ensure our dataset reflects the most current and relevant information, we streamlined the records by keeping only the latest entry for each executor.

The 'Sex' column contained empty fields, which, according to Millenium's operational definitions, indicate tasks performed by robots. Additionally, entries labeled as 'U' were identified under the same criteria. We standardized these fields by filling in such missing or unspecified values with 'R' to denote robotic involvement, thereby maintaining consistency with Millenium’s classification protocols.

These steps in data preparation are crucial for ensuring that our dataset is robust and reliable for further analysis, providing a solid foundation for predictive modeling and insights extraction.

3.2.2 Feature Engineering

To effectively use the date features provided in the dataset, they were transformed to predicted task and actual task duration in minutes. The new created columns were used to calculate a duration ratio to represent if a task was executed in time or not. Values below 1 represent tasks that were executed before the predicted end date while values above 1 describe tasks that were executed beyond the expected timeframe.

To effectively address outliers characterized by unusually long execution times in our dataset, we employed various binning strategies on the actual duration column. This approach helped standardize the duration data, making it more manageable and useful. The most effective binning method will be determined by subsequent feature selection.

The derived insights from the prior analysis led to the construction of the necessary target label to effectively predict the 4 different outcomes as these are not provided in the given dataset. To configure the label, the last appearing application actions of the process were analyzed as these deliver the very last information before an outcome. Table 1 effectively captures how the different activities and application actions lead to the outcome label. Processes that terminate in a different manner will be labeled as intermediate and ignored in the following analysis.

Table 1: Label Construction Overview

Activity ID	idBPMApplicationAction	Outcome label	Comment
101	2981	Request cancelled	-
101	298	Request cancelled	-
103	8888 or 888	closed administratively	does not happen in the dataset
104	8888	closed administratively Requester rejects Accounting Impact	assumption as this is the only value for finished processes in 104
106	8888 or 888	closed administratively	does not happen in the dataset
107	888 or 8888	closed administratively	-
107	299	request finished	only if 278 does not directly appear after

Further findings revealed that the outcome label “closed administratively requester rejects accounting impact” does not appear after October 19, 2023. To incorporate this into the predictive modeling, we introduced a new feature which categorizes if a request arrived before or after the recognized date, thereby enhancing the model´s ability to accurately predict relevant outcomes.

3.2.3 Splitting the datasets

A common issue in predictive process monitoring is the inappropriate use of future data and the determination of an appropriate prefix size. In order to avoid the use of future information and the development of specialized models, it is necessary to split the dataset according to the frequency with which the process has passed through activities 102 or 105, respectively. Consequently, the first dataset contains very little information, as only two activities have been performed at this point in time, namely 100 and either 102 or 105. In considering the second occurrence of 102 or 105, it is important to note that information on previous process steps is already available, including data on activities such as 107 or 108. Subsequently, we constructed nine datasets. One dataset was created for the activity 100, as no loops are

possible for this activity, and four datasets were created for the activities 105 and 102. Four datasets were created as the number of available rows decreased significantly with each reduction in the prefix size. For instance, with four occurrences of activity 105, there were only 11 rows left in the dataset.

3.3 Modeling

3.3.1 Model Selection

When it comes to predictive process monitoring methods, various researcher have experimented with different models. It has been identified that the Decision Tree, Random Forest and XGBoost have been the best performing models so far [6]. There were also some attempts to use support vector machines or generalized boosted regression models, but it was found, that they were inferior to Random Forest [7].

As the dataset contained 33 features after the preprocessing, we needed to make a preselection of features to evaluate the different models against each other, therefore features which correlated strongly were removed from the dataset. We applied the kbest algorithm based on mutual information with the target variable and a k value of 10. Therefore, the model performance is more comparable as all models receive data frames suited to their tree-based characteristics. With the preselected features the models were trained on all datasets and for each dataset the best performing model was selected and further tuned by selecting the most appropriate features based on forward and backwards sequential feature selection as well as recursive feature elimination with the corresponding classifier, like explained in the following.

3.3.2 Feature Selection

As different models performed better in different positions in the process, we needed to select the best features for those models. Therefore, we proceeded to apply feature selection techniques for selected models. We applied Chi-Square Test of Independence, Cramer's V, Analysis of Variance, Pearson and Spearman correlation to decrease the number of features used for the model selection. Recursive Feature Elimination, Sequential Forward and Backward Feature Selection were used to further refine the corresponding models.

Pearson and Spearman correlation analyses among numerical features revealed a strong correlation between features. As a result, we decided to drop features with a correlation above 0.8 in order to avoid collinearity.

Chi-Square Test of Independence revealed that all categorical features had a significant relationship with the target variable.

Cramer's V statistic was used to evaluate the strength of association between each pair of categorical features. In cases where a strong association was found (Cramer's V close to 1), it suggested redundancy between the features. To avoid multicollinearity and to simplify the training data we deleted features that showed values above 0.7.

Analysis of Variance (ANOVA) exhibited high F-values and very low P-values, indicating strong statistical evidence against the null hypothesis. The H0 of the Anova test is, that there is no difference in the means of numerical features and categorical variables. Therefore, no features were dropped in this step.

Recursive Feature Elimination (RFE) with the corresponding classifier was employed to identify the optimal number and mixture of features. This was achieved by utilising the weighted F1 score as the performance metric.

Sequential Feature Selection, both forward and backward, was used to enhance model performance and to refine the feature set further. The amount of features to choose for forward and backward sequential feature selection were set to 10 making the results comparable to each other.

To select the final features for each model we applied forward and backward sequential feature selection as well as RFE and extracted the features which were at least selected from two methods. An example of this process can be seen in Figure 9.

As the the dataset for activity 102 and 105 with occurrence 4 are too small for a train test validation split we were obliged to not continue working with these datasets, as no reliant predictions could be made with this few data.

3.4 Evaluation

To evaluate and compare the models, they were trained on all datasets, comparing them based on the weighted F1 score, as the target label was highly imbalanced. Table 2 summarizes which model was chosen for which datasets including its performance. It can be observed that Random Forest performs very well on the activity 100 dataset and for the dataset where we already have data from recurrent processes. XGBoost tends to perform better on the dataset, where we only have few data, and the activity is passed through the first time. Decision tree, the only non-ensemble learner, continuously exhibits the lowest value for each dataset.

After feature selection and hyperparameter tuning we achieved an average weighted F1 Score of 0,57 and 0,60 overall accuracy. The best model was the Random Forest with a weighted F1 Score of 0,68.

4 Results Evaluation

As previously stated, the process has accumulated 500 years in over time over the two years of provided data. By taking the model into consideration and following the outcome, it is possible to speed up the process and save 300 years overtime, resulting in a recovery of 3.6 million euros, assuming an average bank worker earns 100 euros, and the employer pays 11% of additional social contribution. This illustrates the significant potential for implementing the model into the process and transitioning from an analysis of each case individually to a validating process where an employee simply needs to determine whether the outcome prediction is accurate or if it still requires reevaluation.

5 Deployment

5.1 Next Steps

In order to improve the current process and reduce the throughput time, further decisions must be made. The first step requires identifying the optimal process, which can then be evaluated to determine whether the current state is indeed optimal or if the most current common paths are suboptimal. From this point on further analysis can be conducted like analyzing variants or the current conforming rate (processes that go through the exact ideal sequence vs. all processes).

Secondly, a workshop should be conducted to validate the business objective and, if necessary, to pivot from predictive process monitoring to a more analytical approach. This could involve identifying process bottlenecks or large idle times. Additionally, the proposed timeline and milestones must be discussed and adapted to the current data resources and predictive tools. Consequently, it is recommended that further data needs to be gathered, e.g. client type (business vs. private) or fraud patterns. In addition, the disclosed information should be shared with the project team in order to enable further model fine-tuning.

5.2 Proposed Implementation Timeline

The first step for the implementation will be the Preparation and Environment Setup. The full process will take one week, where the main objective will be to ensure all necessary infrastructure is in place for deployment, setting up environments for development, testing, and production. The cost estimations are €19,500 considering servers, cloud services, software, data pipeline for real time data ingestion, security controls and consultancy company.

The second step is using the defined environment, integrating and evaluating the model functionality in the existing bank's systems. That would take two more weeks, where the developer would implement the API's endpoints, perform load testing to ensure the system can handle expected traffic and address any bugs or performance issues identified during testing. The cost estimations are €36,200 considering labor, cloud services and data integration, consultancy company.

The third step consists in implementing a pilot phase during three weeks and takes feedback from stakeholders and final users. Set a subset of the process and apply the predicted outcomes to measure the efficiency now in real life. Use those results to collect information to improve the systems or the model. It is important to submit the employees to the new way to work and make sure to utilize human judgment while capitalizing on the advantages offered by the model. The cost estimates are €66,000 considering the consultancy company to provide the training, labor of employees, data integration and software.

After the pilot phase and ensuring quality of services, rollout the model to all users and start collecting the business benefits. Gradually increase the number of users and processes accessing the model. At this point it may be necessary to conduct additional training sessions for the employees. This process will take between four and six weeks. The cost estimations

here are €121,000 for consultancy developers, deployment, employees labor for deployment, software, data integration, consultancy company for the training. [8] [9] [10] [11] [12] [13]

The final step is ongoing, where starting between the week number 10 or 12, reporting the main kpis where the company will be able to continuously monitor the system performance, provide regular reports on the impact and efficiency, and find opportunities to optimize the model periodically. The cost estimations are €5,100 dashboard software, data integration, business analyst labor. [14] [15]

This revised plan ensures a structured and phased approach to deploying the model, incorporating employee training to enhance manual processes and focusing on continuous monitoring and reporting for sustained efficiency improvements.

6 Conclusions

The main goal for the project was to enhance Millennium bcp's operational efficiency by leveraging predictive process monitoring models. The methodology employed involved the collection, cleansing, and analysis of data in order to identify inefficiencies within the bank's operations and to propose practical solutions for improvement.

Our analysis covered a comprehensive range of tasks and process steps, based on a 45,000 individual processes dataset. By utilizing sophisticated algorithms like Decision Tree, Random Forest, and XGBoost, we effectively predict the outcomes of different process steps. The most effective model delivered the highest weighted F1 score, highlighting its applicability in scenarios with imbalanced target labels.

The rollout of the project was strategically divided into five stages to ensure smoothly integration with Millennium bcp's current systems. This phased approach facilitates continuous monitoring and iterative enhancements.

By focusing on predictive process monitoring, Millennium bcp is well-positioned to optimize its business processes, reduce unnecessary procedural steps, and improve overall service quality. The project highlights the value of data-driven decision-making in achieving operational excellence and sets a precedent for future initiatives within the bank, potentially saving the bank up to €6.6 million annually.

In conclusion, this initiative has proven that substantial gains in business process efficiency are attainable with the appropriate data, analytical tools, and approaches. The successful implementation of these predictive models represents a crucial advancement for Millennium bcp towards a more streamlined and cost-efficient operational model.

7 Bibliography

[1] Celonis, "What is Process Mining?," 29 May 2024. [Online]. Available: https://www.celonis.com/process-mining/what-is-process-mining/. [Accessed 29 May 2024].

[2] Millenium bcp, "Millenium bcp," 25 May 2024. [Online]. Available: https://ind.millenniumbcp.pt/en/Institucional/investidores/Pages/RelatorioContas.aspx. [Accessed 25 May 2024].

[3] Yahoo!, "yahoo!finance," 25 May 2024. [Online]. Available: https://finance.yahoo.com/quote/BPCGF. [Accessed 25 May 2024].

[4] Glassdoor , "How much does a Banker make in Portugal?," 28 May 2024. [Online]. Available: https://www.glassdoor.com/Salaries/portugal-banker-salary-SRCH_IL.0,8_IN195_KO9,15.htm. [Accessed 28 May 2024].

[5] AMA - Agência para a Modernização Administrativa, "Social Security contributions – paying contributions as an employee," 28 May 2024. [Online]. Available: https://eportugal.gov.pt/en-GB/servicos/obter-informacoes-sobre-as-contribuicoes-para-a-seguranca-social-pagamento-de-entidade-empregadora-por-trabalhador-por-conta-de-outrem. [Accessed 28 May 2024].

[6] I. Teinemaa, M. Dumas, M. La Rosa and F. M. Maggi, "Outcome-Oriented Predictive Process Monitoring: Review and Benchmark," *ACM Transactions on Knowledge Discovery from Data (TKDD),* vol. 13, no. 2, pp. 1-57, 2019.

[7] A. Leontjeva, R. Conforti, C. D. Francescomarino, M. Dumas and F. M. Maggi, "Complex symbolic sequence encodings for predictive monitoring of business processes," *Springer,* p. 297–313.

[8] Glassdoor, "Glassdoor," 28 May 2024. [Online]. Available: https://www.glassdoor.com/Salaries/lisbon-senior-data-analyst-salary-SRCH_IL.0,6_IM1121_KO7,26.htm. [Accessed 29 May 2024].

[9] Glassdoor, "Glassdoor," 28 May 2024. [Online]. Available: https://www.glassdoor.com/Salaries/developer-salary-SRCH_IM1121_KO0,9.htm. [Accessed 29 May 2024].

[10] Glassdoor, "Glassdoor," 28 May 2024. [Online]. Available: https://www.glassdoor.com/Salaries/senior-business-analyst-salary-SRCH_IM1121_KO0,23.htm. [Accessed 29 May 2024].

[11] Tableau, "Tableau," 29 May 2024. [Online]. Available: https://www.tableau.com/pricing/teams-orgs. [Accessed 29 May 2024].

[12] NordLayer, "All-round network security for the modern workforce," 29 May 2024. [Online]. Available: https://nordlayer.com/pricing/. [Accessed 29 May 2024].

[13] databricks, "Workflows & Streaming: Delta Live Tables Pricing," 29 May 2024. [Online]. Available: https://www.databricks.com/product/pricing/delta-live. [Accessed 29 May 2024].

[14] Amazon AWS, "Create estimate: Configure Amazon Kinesis Data Streams," 29 May 2024. [Online]. Available: https://calculator.aws/#/createCalculator/KinesisDataStreams. [Accessed 29 May 2024].

[15] Amazon, "Preço do Amazon Lightsail," 29 May 2024. [Online]. Available: https://aws.amazon.com/pt/lightsail/pricing/?pg=ln&sec=hs. [Accessed 29 May 2924].

8 Appendix

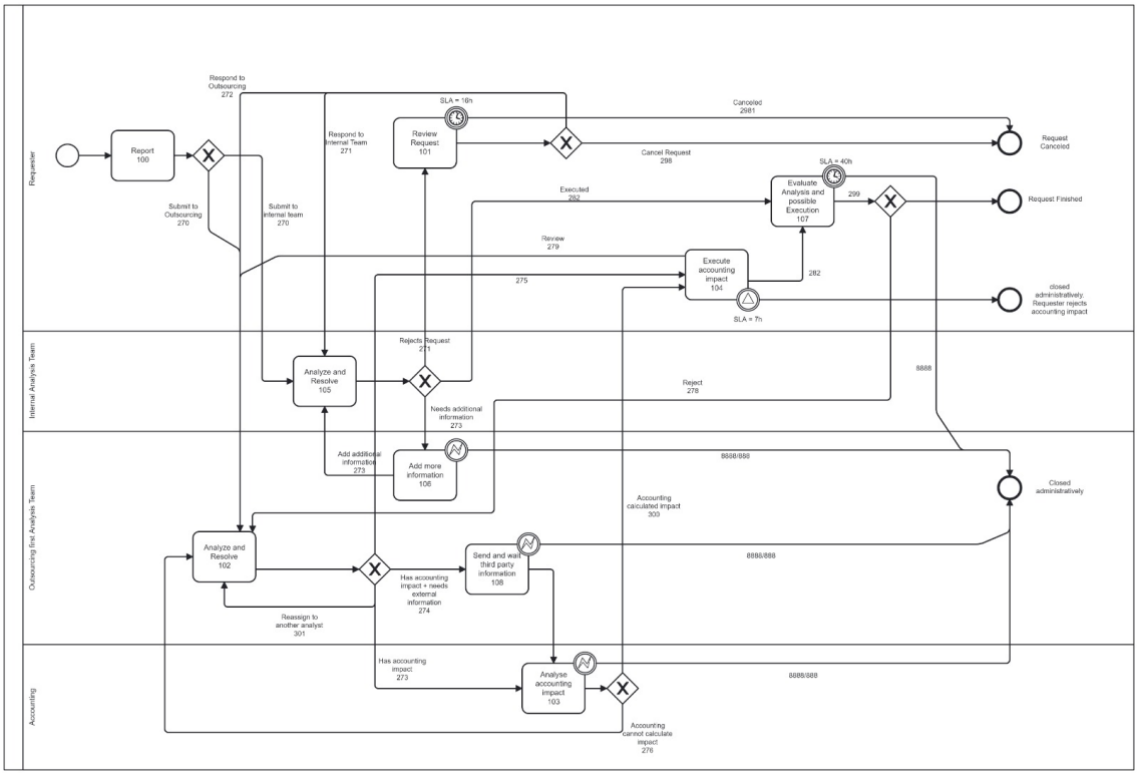

Figure 2: Business Process Modell Overview

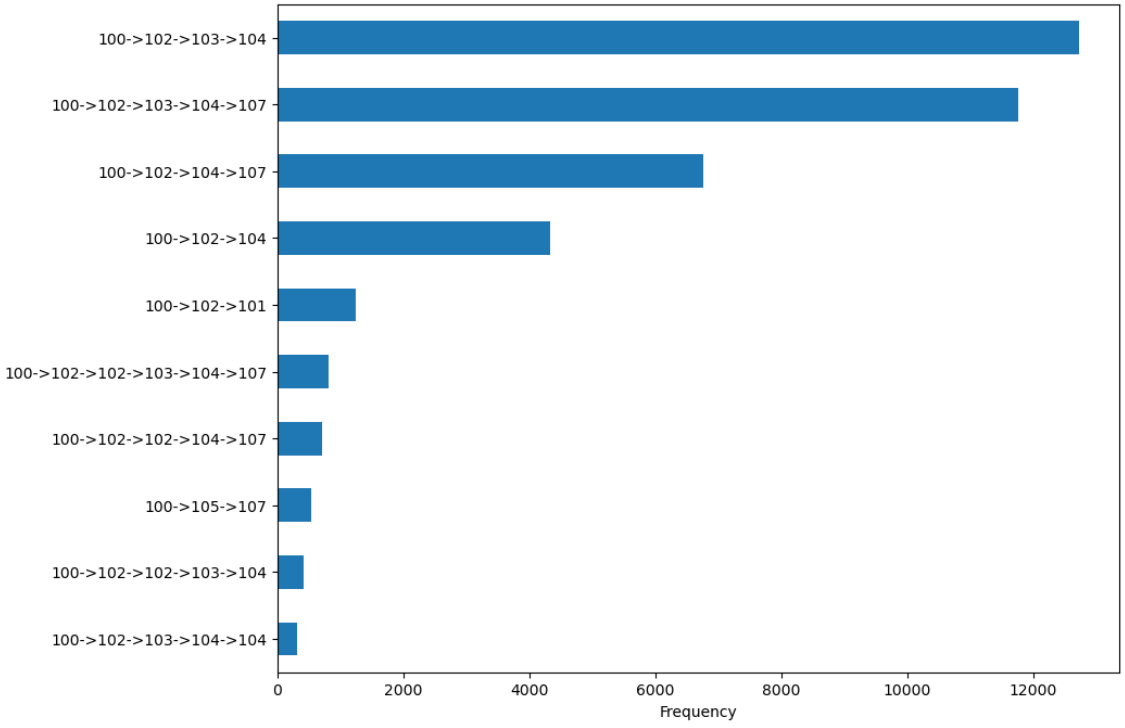

Figure 3: Top 10 most common activity sequences

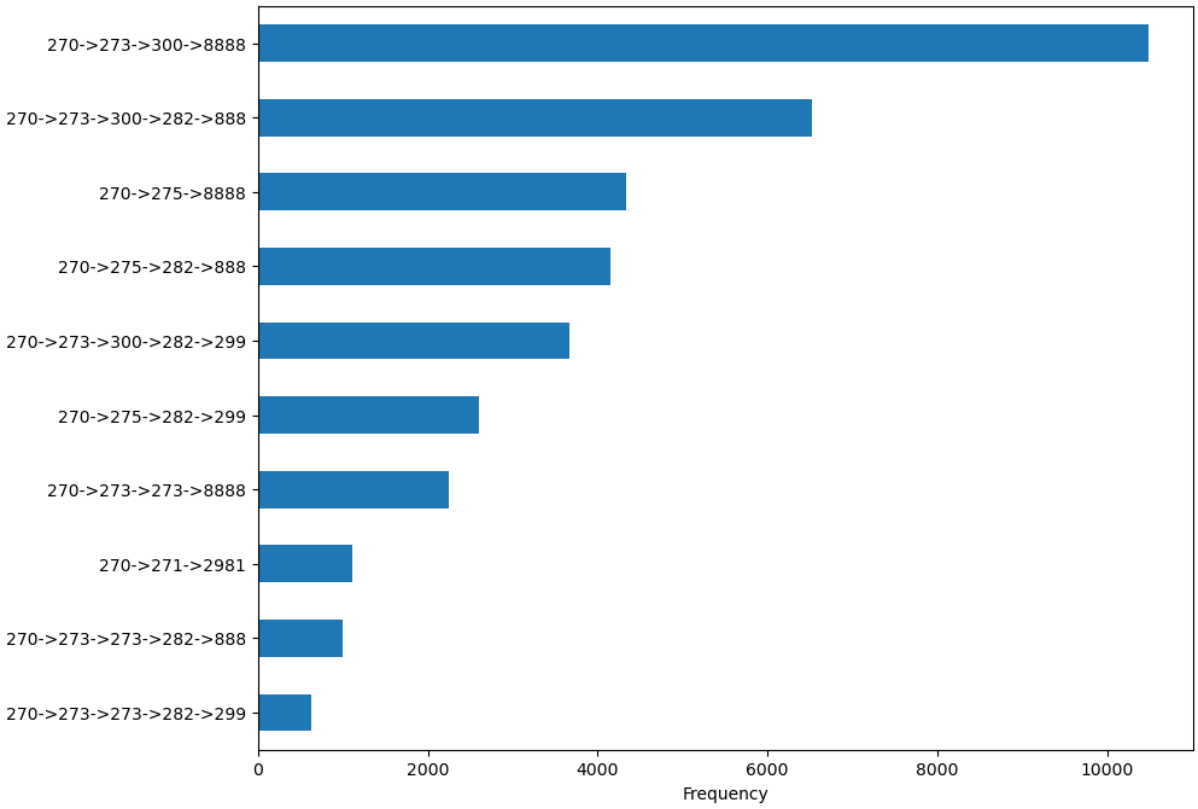

Figure 4: Top 10 most common action sequences

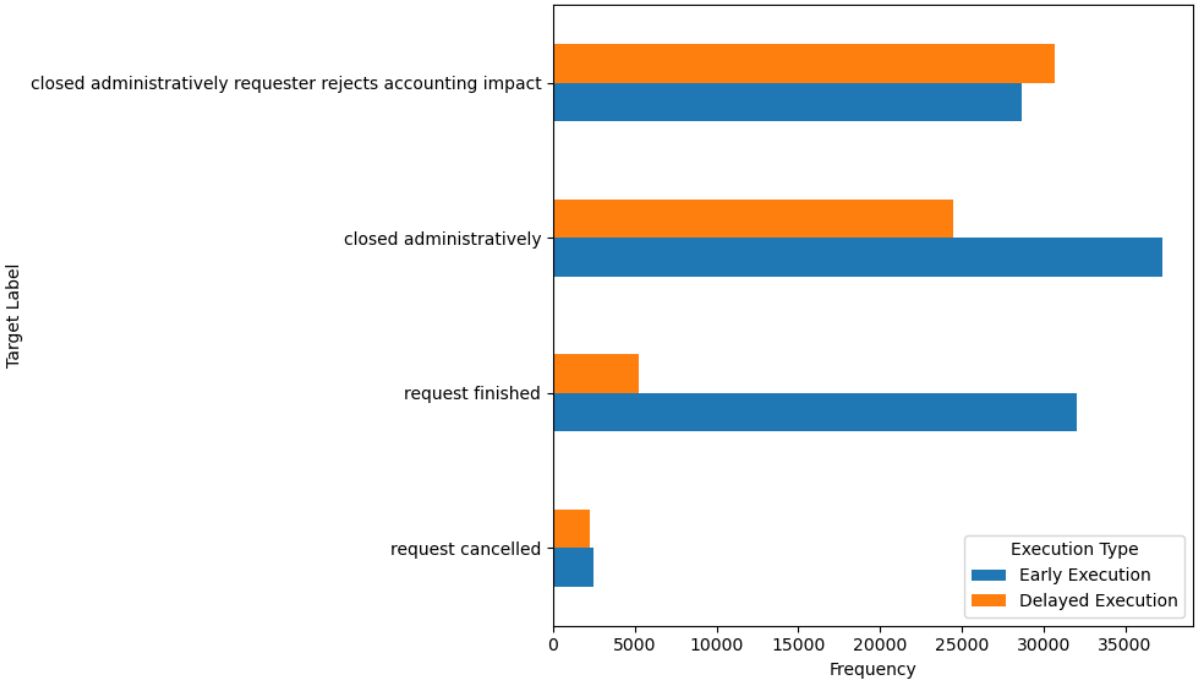

Figure 5: Distribution of early and delayed executions by target

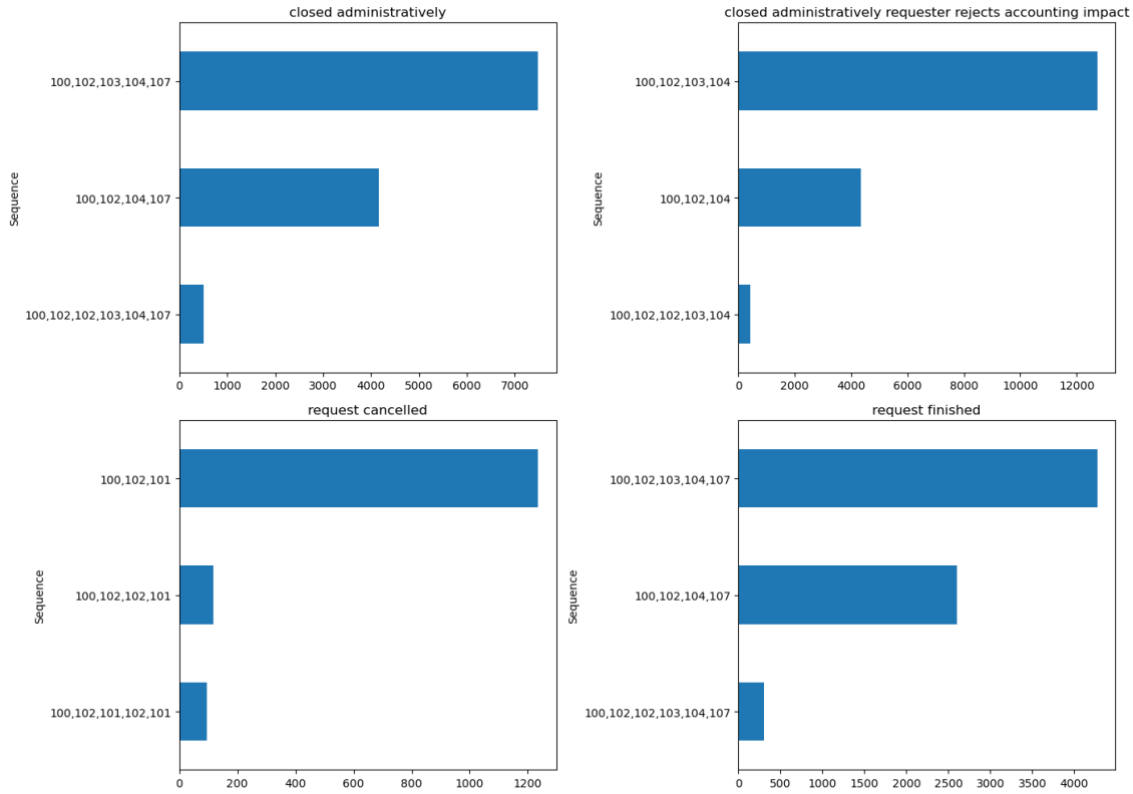

Figure 6: Distribution of most frequent activity sequences per label

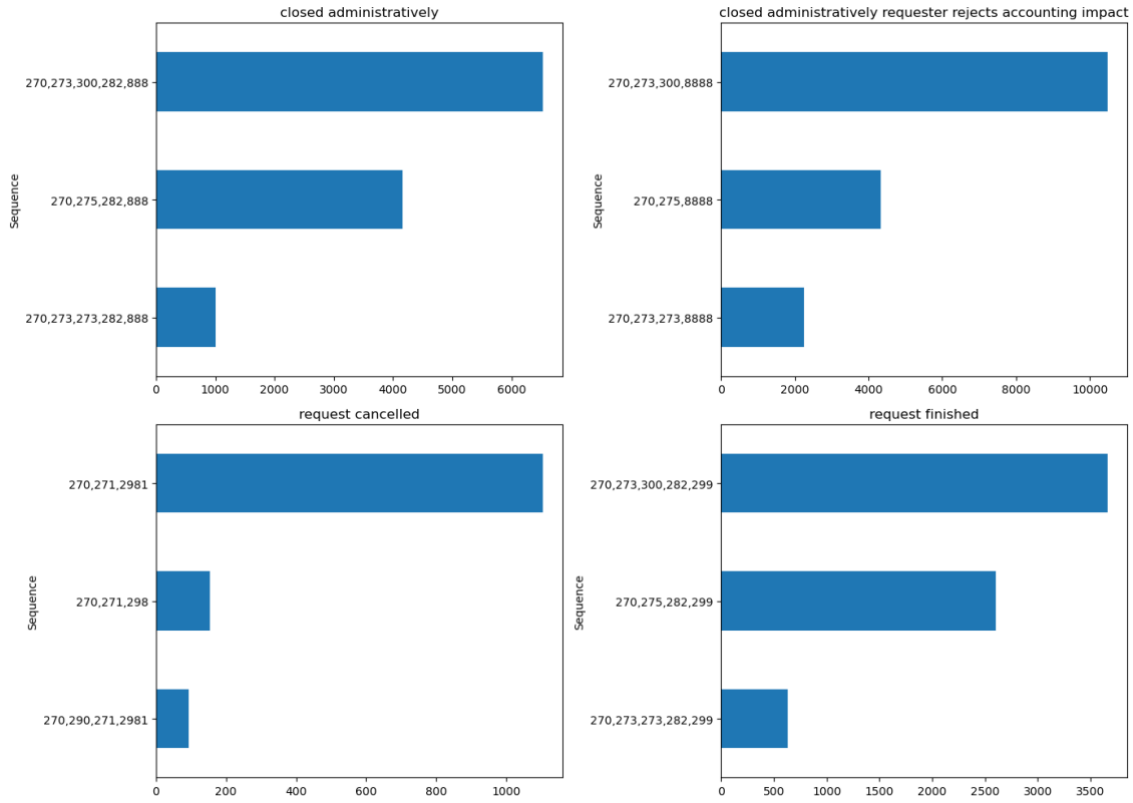

Figure 7: Distribution of most frequent action sequences per label

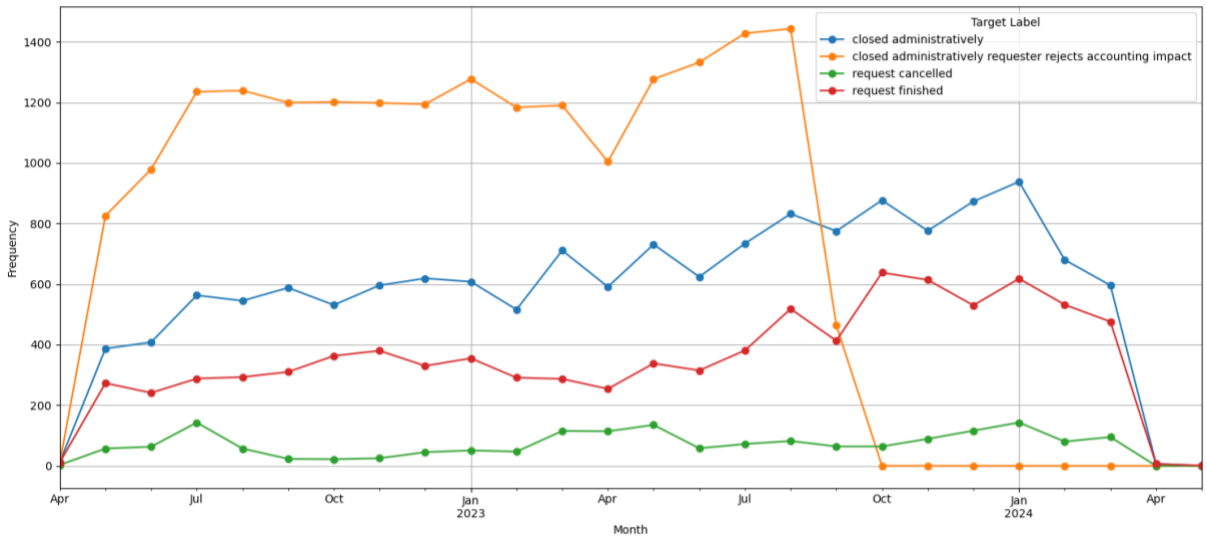

Figure 8: Frequency of the outcome labels over time

Figure 9: Example of Majority Vote Process for Dataset 105 Occurrence 2 with Random Forest

Table 2: F1 Scores of Trained Models

	Decision Tree		Random Forest		XGBoost	
	Weighted F1	Accuracy	Weighted F1	Accuracy	Weighted F1	Accuracy
Activity 100	0.68	0.68	0.68	0.69	0.54	0.59
Activity 102 occurrence 1	0.45	0.45	0.46	0.46	0.48	0.56
Activity 102 occurrence 2	0.45	0.44	0.50	0.50	0.47	0.54
Activity 102 occurrence 3	0.47	0.48	0.52	0.53	0.47	0.54
Activity 102 occurrence 4	0.46	0.46	0.51	0.51	0.54	0.55
Activity 105 occurrence 1	0.50	0.50	0.47	0.50	0.52	0.57
Activity 105 occurrence 2	0.55	0.56	0.67	0.68	0.62	0.62
Activity 105 occurrence 3	0.42	0.57	0.65	0.71	0.42	0.57