NOVA IMS Information Management School

MDSAA

Master's Degree Program in

Data Science and Advanced Analytics

Business Cases with Data Science

Case 4: **Business Process Predictive Monitoring Model**

Duarte Mendes, number: 20230494

Dzmitry Nisht, number: 20230776

Inês Silva, number: r20201580

José Marçal, number: r20201581

Ricardo Sousa, number: r20201611

Group E

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

March, 2024

INDEX

1. EXECUTIVE SUMMARY

Millenium BCP is one of the top-ranking Portuguese banks whose influence extends internationally. One approach to be able to stay ahead of the trends and have available what customers are looking for is the usage of a Business Process Predictive Monitoring Model which allows to have a competitive advantage.

No model is ideal and there are many factors influencing it. However, it does not mean it cannot add value to the business, as long as it is as tailor-made as possible so this study will revolve around finding an optimal choice within the available tools.

Throughout the project, a CRISP-DM approach was used which led to the selection of specific models for each business process. The recommendations vary according to the results obtained by Millenium BCP in their research.

Money is always a key concern to any business so at the end of this research, it is possible to find an estimate of costs associated with the creation of this concept. It showed that the company will reach revenues of over 1 million euros.

Finally, one can also find potential improvements to the company's data storage and organization.

2. BUSINESS NEEDS AND REQUIRED OUTCOME

The methodology to be implemented in this project is based on the CRISP-DM (Cross-Industry Standard Process for Data Mining) detailed approach. The methodology was employed throughout the project to guide the work and to describe and analyse the various key areas of the project, namely the Business Understanding, Data Understanding, Modelling, Evaluation and Deployment and Maintenance Plans steps.

2.1. BUSINESS INTRODUCTION

Millennium BCP is a leading Portuguese bank with over three decades of history. It commenced operations in 1985 and currently has over 1700 agencies and 21000 employees in a multitude of locations, serving over 5.4 million customers worldwide. Throughout time, the bank has become the largest privately held bank in Portugal, in addition to maintaining a strong presence in other core markets, including Angola, Mozambique, Macao, Poland and Switzerland. Its focus and commitment are based on delivering and creating value through retail banking, offering financial products and services to individuals and companies that comply with the highest standards of quality and corporate responsibility.

In this context, Millennium BCP is seeking to enhance its business processes further through the utilisation of process mining and machine learning techniques, to ensure efficient business processes and elevated customer satisfaction. For the analysed business process, the bank receives a considerable number of requests from its customers, which are processed by various departments within the organisation, involving either internal or outsourced employees. Each request is comprised of several interrelated tasks, the successful or not completion of which determines the outcome of the

request. Nevertheless, not all requests are successfully completed, and some are rejected or closed due to various reasons throughout the process.

2.2. BUSINESS OBJECTIVES

The main objective of this project is to develop machine learning models that can accurately predict the outcome of each request based on the information provided in the dataset for each task, after its necessary transformation into appropriate features that can be ingested by the models. This approach will enable Millennium BCP to identify any potential issues early in the process and take corrective action, thereby improving both the efficiency and effectiveness of its business processes.

Moreover, the enhancement of the predictive accuracy of business process requests' outcomes represents a significant business objective for this project. The employed models must be capable of accurately predicting the outcome of a request, allowing Millennium BCP to take proactive measures to ensure process completion without any issues.

Furthermore, a reduction in the number of incorrectly rejected requests is also considered a business objective. The identification of potential issues at an early stage in the process allows for the implementation of corrective action by Millenium BCP, thereby ensuring that each request receives an adequate assessment and analysis, resulting in a correct outcome.

The enhancement of business process efficiency is also a key business objective. The early prediction of a request's outcome allows the bank to proceed with a more efficient resource allocation process while ensuring that each request is completed promptly and encompasses less operational costs.

In addition, the enhancement of the customer experience is included in the set of defined business objectives. This is achieved through the improvement of the efficiency and effectiveness of the business processes, which allows Millennium BCP to provide better services to its customers. This, in turn, translates into an enhanced customer experience and, consequently, increased customer satisfaction and loyalty.

A comprehensive examination of the pertinent data about a process has the capacity to furnish insights into the variables that shape the outcome of a request, which represents one of the most pivotal defined business objectives. Consequently, the selected model must be capable of identifying the key features that influence the outcome of a request, thereby enabling Millenium BCP to implement accurate measures that can reduce inefficiencies and improve the business process in analysis.

Finally, the achievement of such business objectives can ultimately lead to the improvement of Millenium BCP’s business processes, namely by the cost reduction coming from efficient resource allocation, timely execution of processes and an adequate insight analysis of factors influencing outcomes. Furthermore, the customer experience can be enhanced as a consequence of the implementation of efficient and effective operational business processes conducted at the appropriate times, which will consequently improve Millennium BCP’s competitiveness and profitability abilities.

2.3. BUSINESS SUCCESS CRITERIA

It is of the utmost importance to define clear business criteria that align with the business objectives in order to ensure the successful implementation of this project. Consequently, the defined business

success criteria will provide meaningful insights into the efficacy of the implemented machine learning model, as well as ensure that the needs and expectations of Millennium's BCP stakeholders are met and satisfied.

One of the key business success criteria for this project is improved prediction accuracy. The predictive accuracy of the implemented model is crucial for ensuring reliable outcome predictions and extracting insights that can lead to trustworthy recommendations for process improvement. This, in turn, can facilitate high-impact decision-making regarding the business processes involved. In this context, the implementation of the aforementioned model will allow Millennium BCP to streamline the business process in analysis, predict potential issues that may arise during processing time, and implement corrective measures in a proactive manner. This will contribute to more strategic planning and optimisation of the business process. Consequently, the accuracy of the predictions will be evaluated using measures such as the F1-score, precision and recall. The objective is to achieve a minimum score of 0.85 in each of these measures, after model implementation.

Furthermore, the improvement of business process efficiency is regarded as a fundamental criterion for the success of this project. The insights derived from the implemented model can be utilised by Millenium BCP to enhance the optimisation of tasks and workflow sequences, as well as an accurate allocation of process resources, ultimately leading to improvements in the efficiency of the business process. The reduction of the time required to complete each request can facilitate operational cost savings for Millenium BCP and enhance overall productivity in the context of the business process in analysis. This, in turn, can enable faster turnaround times for Millenium BCP's critical business operations, such as customer service responses. Such improvements can ultimately lead to higher customer satisfaction and retention rates. A reduction in processing time also permits bank employees to direct their attention to more significant tasks, rather than those that are time-consuming and of lesser importance. Furthermore, enhanced process efficiency will result in more effective resource management, as streamlined operations will require fewer resources to achieve the same outcomes. In order to ascertain the impact of the model implementation, a comparison will be made between the average time required to complete a process over a six-month period prior to the implementation and the same period following the implementation. In this context, the objective is to reduce the average completion time of processes by at least 20%.

The growth of the business as a consequence of the implementation of the model is also a key factor in the success criteria for this project. The model has the capacity to have a positive impact on Millennium BCP's business objectives, including cost reduction, customer satisfaction improvement, and revenue increase. In this context, the growth in question can be quantified by its capacity to drive tangible improvements in key performance indicators (KPIs) aligned with Millennium BCP's business objectives and strategy.

The reduction in processing costs also encompasses the set of defined business success criteria. This is because it can enable Millennium BCP to enhance the profitability of its services, namely by minimising operational expenses and therefore saving costs to the bank. In order to evaluate the reduction in processing costs, a comparison should be made between the average processing costs before and after the model was implemented. Furthermore, a target of at least a 10% reduction in average processing times should be set, comparing a six-month period before the system was implemented with the same period after implementation.

From another perspective, the improvement of customer satisfaction is linked to the defined business success criteria for this project. This is because such improvement can contribute to an increase in Millennium BCP's customer loyalty and retention rates, as well as the use of additional or complementary services from the bank or service recommendations to others. Ultimately, this can lead to organic growth. In order to ascertain the efficacy of the aforementioned measures, customer satisfaction will be gauged via customer satisfaction surveys, which will be employed to monitor progress. The average customer satisfaction scores will be compared before and after the implementation of the model, with a target of an increase of 5% in the average customer scores between a six-month period before system implementation and the same period after implementation.

Furthermore, an increase in Millennium BCP's revenue represents one of the defined business success criteria for the implementation of a business process outcome prediction model. In this context, the insights provided by the model can facilitate the unlocking of processing time gains and cost savings, which can provide a substantial financial boost for the bank, allowing it to increase margins in such a competitive market as banking. Such revenue growth is closely linked to enhanced process efficiency, which results in improved and more expedient service delivery, enhanced customer experience, and the capacity to manage a greater volume of processes with high standards of service quality and requiring less processing time. The revenue increase will be quantified by comparing Millennium BCP's revenue related to the business process between a six-month period before system implementation and the same period after implementation. The aim is to achieve a 2% increase in revenue for that specific type of revenue between both periods.

2.4. SITUATION ASSESSMENT

Situation assessment includes a comprehensive analysis of the resources required, a risk assessment, and a cost-benefit analysis, regarding the implementation of a business process outcome predictive model for Millennium BCP. This phase will play an active role in the preparation of the project implementation plan, as well as in the definition of business objectives and success criteria, and encompasses considerations on the dataset provided for analysis, as well as the intrinsic business context in which this project is situated.

Currently, Millennium BCP recognises the importance of BPM processes in the context of the bank's operations, with a significant impact on business process efficiency, customer satisfaction, and overall business performance. Consequently, an extensive BPM model for the Millennium BCP customer service process is already in place, establishing guidelines for the process's course and providing a detailed schema that includes the involved departments, possible paths of action to follow in specific situations, as well as the multiple possible outcomes for the requests. This enables all the involved departments to optimise process workflows, coordinate efforts and perform tasks efficiently. Nevertheless, Millennium BCP aspires to further enhance its BPM structure for the customer service process in analysis. This will be achieved by leveraging insights derived from the application of predictive analytics and process mining techniques. This will result in enhanced efficiency, accuracy and cost-effectiveness of the involved business process and its outcomes. Ultimately, this will translate into enhanced overall productivity, decreased operational costs and excellent standards of quality in the service provided to its customers.

The dataset analysed throughout this project was made available by Millennium BCP and consisted of the file 'Case4_UNL-IMS - Data - delivery v2.xlsx', which contained 4 different sheets: 'Q1 – Task execution data', 'Q2 – User information', 'Q3 – Specific request data' and 'Q4 – Rejections'. The 'Q1 – Task execution data' sheet initially consisted of 209017 rows, each representing a task that was part of a process and its respective information, and 12 columns: 'Task Id', referring to a unique identifier for each task; 'Request Identifier', consisting of a unique identifier for each request; 'Task Arrival date', representing the exact time a task arrives to the work queue to be captured by an executer; 'Task Capture Date', referring to the exact time a task is captured by an executer; 'Task execution end date', consisting of the exact time the task is completed; 'Task predicted end date', which refers to the time that the task is predicted to be completed; 'Activity Id', consisting of a unique code identifying the type of task involved; 'Task executer', representing an unique identifier for the employee performing the task; 'Task executer department', referring to a unique identifier for each department involved in the task; 'Task Type', consisting of the type of task performed; 'Action', which consists of the action associated to each task type; and 'idBPMApplicationAction', mentioning a unique code representing the task action in the context of the Business Process Management. Moreover, the 'Q2 – User information' consisted of 11370 rows, each of those representing information about specific employees involved in task execution, and 7 columns: 'Task Executer', containing a unique identifier for each Millennium BCP's employee possibly executing tasks; "Sex", regarding the gender of employees; "BirthYear", referring to the year of birth of each task executer; "Role ID", consisting of a unique code associated with the type of role each employee entitles; "Is Manager", a binary variable referring to if an employee is an organizational unit manager or not; "OrgUnitSince", consisting of the year from which the employee started working in the department; and "IsOutsourcer", representing if an employee is working in Millennium BCP, but is contractually linked to an external outsourcing company. The 'Q3 – Specific Request Data' sheet was composed of 297556 rows, each representing specific information about each customer request, and 3 columns: "idField", a unique identifier regarding specific fields present in the customer request; "Request Identifier", representing a unique identifier for each customer request, which is also included in the "Q1 – Task Execution Data" sheet; and "Value", consisting of specific data that is part of the customer request. Finally, the "Q4 – Rejections" sheet provided information on the requests that were rejected, with 4099 rows and 2 columns: "Task Id", referring to the unique identifier of the rejected task; and "idBPMRequirement", containing a unique identifier for each rejection motive, according to the BPM model.

The provided dataset was then meticulously examined and prepared for the subsequent phase of the project. This was achieved through the utilisation of well-known Python libraries, including *pandas* and *numpy* for data preparation and *matplotlib* and *seaborn* for data visualisation. The selected libraries for the feature selection and modelling phases were *sklearn*, *scipy*, *xgboost*, *lightgbm*, *termcolor* and *catboost*.

The development of such a project also necessitates the consideration of potential obstacles or events that may affect the timeframe for its implementation, the associated costs, or the anticipated outcomes of the project. Furthermore, it requires the identification of resources and actions that can be taken to minimise such impacts. Consequently, a comprehensive risk assessment and contingency plan must be implemented to enhance the capacity and preparedness to address such scenarios.

Incomplete or poor-quality data represent a significant risk for the development of this project. This is due to the potential for inconsistencies in the data or the lack of relevant data for analysis in the

provided dataset. In light of these considerations, it was necessary to address situations involving an extremely high number of tasks or those subject to change over time. To this end, an extensive implementation of data cleaning and preprocessing techniques was employed to obtain a clean, accurate and business-oriented pre-processed dataset. This was intended to serve as a foundation for subsequent stages, including feature selection and modelling. The objective was to avoid major implementation issues and negative impacts on model accuracy and reliability. Furthermore, ensuring data quality, accuracy and consistency prevents the possibility of obtaining biased and unreliable results and contributes significantly to the extraction of improved insights from the predictive models, which can translate into significant savings in terms of processing time and operational costs.

Furthermore, issues about data diversity and representativeness also emerge as a risk for this project. Despite the provided dataset containing relevant information for outcome prediction, it still lacks relevant information that could be imputed to the model. This includes information such as the specific date on which an employee left the bank, for example, and its absence may limit the quality and accuracy of the final process outcome predictions and extracted insights provided by the implemented model. Consequently, ensuring data richness and diversity by requesting access to more complete data in the future may help to overcome this risk.

The selection of distinct predictive algorithms and suboptimal parameterisation may also have a considerable impact on the predicted outcomes and insights derived from this project. Inadequate algorithm selection or suboptimal parameter tuning in the modelling phase may result in suboptimal outcome prediction accuracy and the extraction of poor-quality insights from model predictions. Furthermore, this may lead to underperformance in meeting the expectations of Millennium BCP's stakeholders for the project. The selection and respective parameter tuning of the different selected algorithms to be applied, such as Logistic Regression, Support Vector Machines, Random Forest or Extreme Gradient Boosting, ensured the diversity of modelling approaches in order to gain insights into the most efficient process workflows and accurate outcome predictions for Millennium BCP's departments and remaining stakeholders. These insights can be turned into significant value for the company.

Furthermore, resistance to the adoption of the model and a lack of user training may also impact the insights and outcomes of the project. In this context, the development of an extensive and comprehensive training roadmap programme and the provision of detailed explanations on changing procedures and respective benefits to Millennium BCP's involved employees regarding the use of the newly implemented model are vital steps to the success of the project. Furthermore, it is essential to provide ongoing support for the utilisation of the model and to gather feedback from users regularly. This will enable any issues to be identified and addressed promptly, thereby enhancing the user experience over time.

The handling of sensitive and masked data throughout the project, derived from the sensitivity of customer and operational processing data, is particularly challenging in a sensitive activity sector such as banking. This is due to the need to comply with data regulations and to interpret and prepare the data, which can affect the quality and accuracy of the final results from model implementation. In this context, comprehensive and exhaustive exploratory data analysis can assist in mitigating this issue, including the presentation of pertinent visualisations and summary statistics on the diverse dimensions and characteristics of the data, thereby enabling the generation of actionable insights and the

establishment of relationships within the data. Such outcomes have the potential to be pivotal for enhanced data comprehension and can significantly enhance the data preparation and feature selection stages of the project, ultimately resulting in enhanced outcomes and insights from model implementation.

Furthermore, an inaccurate communication to Millennium BCP's stakeholders regarding the key findings and recommendations for process efficiency improvement of the project can impede the implementation of the project. Consequently, it is of extreme importance to ensure an effective and business-oriented presentation of the project in question, given that the success of the project hinges upon the clear and concise addressing of the actionable findings and resulting recommendations that emerge from the implemented model.

In terms of the cost-benefit analysis of the project, the potential direct and indirect costs include computing resources, data processing and modelling, system maintenance, training on how to use and interpret the results of the predictive model, as well as implementation costs of the changes in the process operational paradigm for Millennium BCP. On the other hand, the benefits of the project include an increase in overall productivity and process efficiency, increasing the volume of handled requests at the same time. This allows for enhanced and accurate resource allocation during the business process. Additionally, improved service quality is provided to customers, with more focus on relevant tasks and less propensity to errors. This leads to higher customer satisfaction and retention rates through enhanced customer experience. The optimisation of the bank's operational strategy, including business processes and respective workflows, is promoted through the efficient allocation of resources, operational cost savings and gains in processing time. Furthermore, the competitive advantage gained through the implementation of a predictive model can bring the previously referred positive impacts to Millennium BCP, which can provide the bank with a distinctive competitive advantage, especially in an activity sector where competition is extremely tight and cost savings can improve significantly revenue margins.

2.5. Determine Data Mining goals

The definition of data mining goals represents a pivotal stage in the development of this project, as it outlines the technical steps that will facilitate the pursuit of the previously defined business objectives. Consequently, the data mining goals should establish the technical guidelines and expected results for the work developed throughout the project. This will contribute to the creation of robust and reliable predictive models for Millennium BCP, which will boost overall productivity, drive the optimisation of the business process and increase revenue margins by saving operational costs.

The primary data mining goal of this project is to enhance the process outcome prediction accuracy, with the F1-score serving as the primary metric for assessment and the highest possible score being the goal. Maximising the F1 score improves the prediction accuracy of the different outcomes for the processes, as well as the ability to provide insightful outputs from those predictions. These outputs can be further used to boost Millennium BCP's operational efficiency and to allow for savings of both processing time and operational costs. In this context, the implementation of algorithms is designed to maximize prediction accuracy and the use of diverse evaluation metrics to assess the performance of selected predictive models, including the F1 score, precision, recall, hit rate, and ROC-AUC. Furthermore, by fine-tuning the model parameters and algorithms to optimise outcome prediction accuracy based on cross-validation results, it is possible to achieve significantly higher F1 scores, which

can lead to more precise predictions on the outcome of each process and useful insights regarding the business process. This, in turn, improves business process efficiency for Millennium BCP's involved departments, which benefits the entire organisation as a whole.

In addition, the integration and cleansing of data from multiple sheets to create a consolidated and high-quality dataset is identified as a significant data mining goal for this project. The merging of the disparate dataframes corresponding to each sheet into a unified, consistent dataset, facilitated by the use of common identifiers such as "Task Executer", "Request Identifier" and "Task Id", represents a pivotal step in the project, as it ensures the reliability and consistency of the data for subsequent analysis, thereby enabling the appropriate application of process mining and predictive modelling techniques. This approach enables Millennium BCP to gain insights that can inform decision-making regarding process optimisation, based on the outcome predictions.

The definition of a suitable target variable for outcome prediction is also considered a vital data mining goal for the project. This is because it ensures that the target accurately reflects the diverse outcomes for the business process, thus facilitating the development of precise and meaningful models in the subsequent modelling step. In order to achieve this, it is first necessary to gain an initial understanding of the business context and to map out the entire process lifecycle, identifying the key decision points and conditions associated with each of the possible outcomes. This will then allow for an appropriate definition of a target variable in a process mining context. Furthermore, an accurate and meticulous encoding of the target variable is essential to preserve the meaningful distinctions between different outcomes and ensure compatibility with the chosen models. This enables the construction of predictive models on solid foundations, which provide precise and actionable predictions and process insights.

Furthermore, performing prefix extraction on the merged dataset also constitutes one of the defined data mining goals for the project. Such a step will allow to generate sequences of activities with different lengths, up to a certain point in time, which will be used in order to analyse process data and apply the chosen predictive models. Prefix extraction emerges as a crucial process mining technique for the project, allowing for a better understanding of partial process executions in the context of the whole process, as well as improving the predictions of process outcomes based on extracted partial process executions. An accurate and effective prefix extraction phase allows to leveraging of partial process data for predictive modelling and analysis steps, which can enable pattern identification in the early stages of processes, prediction of future outcomes, and ultimately, optimization of business process management.

Moreover, the project has identified the objective of performing bucketing on the merged dataset as a defined data mining goal. In this context, bucketing enables the segmentation of the merged dataset into meaningful groups (or buckets) based on stipulated criteria. The definition of clear bucketing criteria, along with the accurate segmentation of the data, the detailed analysis of each bucket and the development of tailored predictive models, collectively contribute to a targeted and effective process analysis, thereby improving business process efficiency within Millennium BCP.

Another crucial data mining goal of the project is to perform encoding to the previously defined buckets, given the diversity of encoding approaches in the context of process mining. In order to achieve this, a number of different techniques can be employed in isolation or combination. These include frequency-based encoding, index latest payload encoding and complex index encoding. The

objective of this encoding process is to transform the data in order to more effectively capture the underlying process characteristics and specificities. This is done in order to facilitate outcome prediction by effectively representing the information contained in the process data, as well as to enhance the predictive power of the chosen models. In this context, the encoding of data from each bucket by employing accurate and appropriate techniques to capture various aspects of the process data can ultimately contribute to the generation of accurate and actionable outcome predictions.

The utilisation of diverse modelling approaches during the modelling phase is also identified as a crucial objective within the data mining framework of the project. This is due to the fact that it contributes to the enhancement of the predictive accuracy of the models, through the utilisation of a range of machine learning algorithms. This, in turn, allows for the identification of the most effective and robust models capable of predicting the outcomes of Millennium BCP's business processes. In this context, it is recommended that different modelling approaches, such as logistic regression, decision trees, random forest, extreme gradient boosting (XGBoost), catboost, and support vector machines (SVM), be employed. The objective is to identify the most effective models and potentially combine their strengths through ensemble techniques. This strategy allows for a comprehensive exploration of different modelling techniques, intending to leverage the unique strengths of each algorithm to improve overall model performance. Moreover, it is also important to perform proper parameterisation and identify the optimal hyperparameters for each of the employed models, in order to achieve the best possible model performance and reach the maximum potential prediction accuracy.

It is also important to develop models that generalise well to new and unseen data, to avoid overfitting to the provided dataset, and to scale the chosen models to handle both finalised and ongoing requests, which are also critical data mining goals for this project. In this context, the application of techniques such as cross-validation and ensemble learning ensures the development of robust predictive models that generalise well to new data and perform reliably across different stages of process completion (pre-mortem and post-mortem), thereby contributing to increased predictive accuracy of outcomes in different process contexts. Furthermore, it is essential to implement continuous monitoring and assessment of model performance to guarantee that the selected models remain effective and stable following their deployment in real-time applications.

Finally, the protection of the confidentiality and integrity of customer data and the processing of data in a secure and accountable manner is a further data mining objective for this project. This is achieved by ensuring that the data utilised is anonymised and does not disclose sensitive information on Millennium BCP's client base or operational structure. In order to fulfil this objective, it is necessary to maintain a continuous collaboration with the Data Governance and Compliance teams, in order to provide guidance and oversight. This ensures that the project adheres to all the data regulatory and compliance requirements of the bank.

3. METHODOLOGY

3.1. DATA UNDERSTANDING

For this project, an excel file was given with four sheets, each with information on different facets of the project: one regarding the execution of the task, another with the user information, the third with data on specific requests and the last with rejections data.

The task execution sheet has each row as a different task and displays information on the ID of the case for the task being performed, the arrival date, capture data, execution end date and predicted end date for that task, who performs the task and to what department they belong to, and the type of task that is being done. Each task is also associated with an activity, an action and a BPM application action. Overall, there were 209017 tasks, corresponding to 45772 cases, meaning on average a case is around 4-5 tasks long. From this sheet, information was extrapolated about the process being executed. These cases were composed of 9 distinct activities, and their median time from start to finish is about 12 days. Then the whole process flow was identified, as presented on image Figure 1. However, this includes lots of points of repetitions, backtracks or execution of tasks that do not add value to the business. The ideal flow for this process is represented in Figure 2.

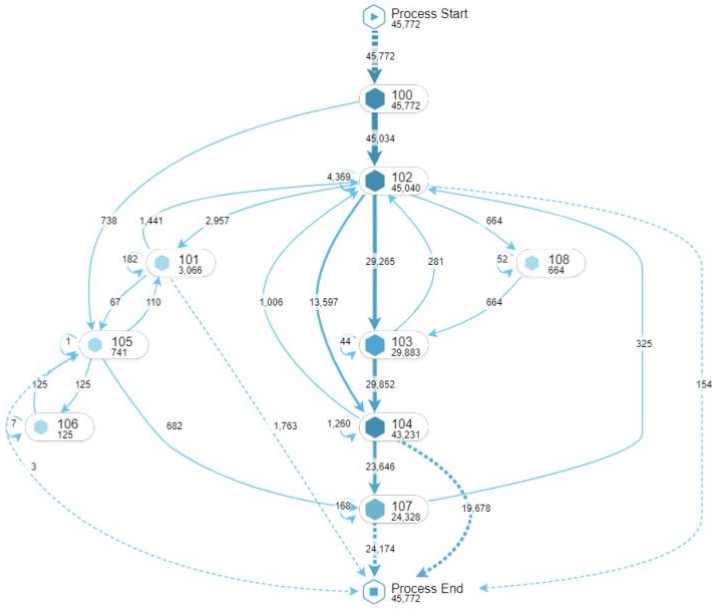

Figure 1 -Process Flow

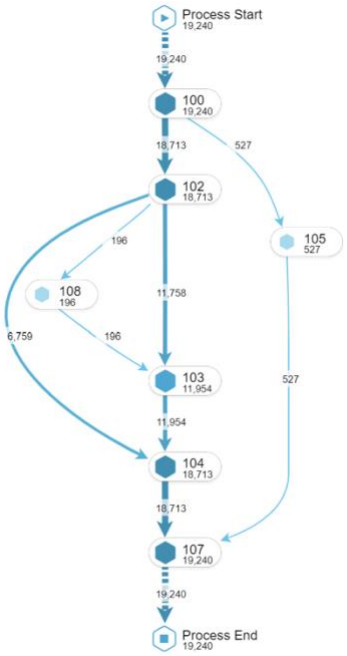

Figure 2 -Ideal Process Flow

Finally, the likelihood of each outcome happening for the process was explored. In the dataset, the most probable outcome is the case being closed due to the requester rejecting the accounting impact, followed by the case being closed administratively, then the request being finished successfully, and finally, the request being cancelled altogether, as shown in Figure 3.

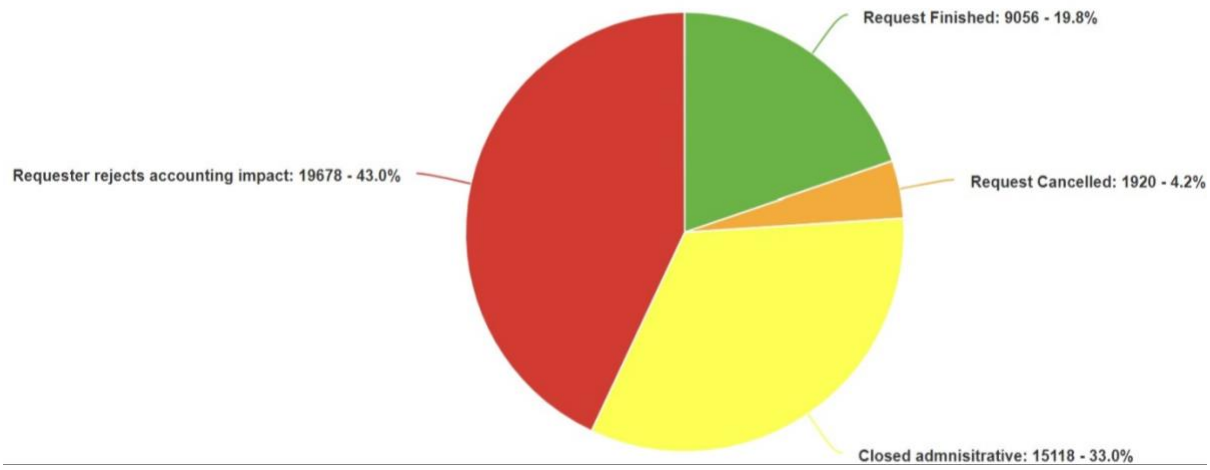

Figure 3- Processes' Outcome

The user information dataset has information on the task executer's sex, birth year, role within the company, date of joining the company and a binary column on if they are a manager and another binary column on if they are an outsourcer. Here, it was found that the binary columns were colinear, thus one of them should be removed and the other kept. It was also discovered that some rows were repeated, likely due to an update in the role and the previous entry mistakenly was not deleted, as well as odd values in the gender column that were assumed to be referring to bots.

The specific request sheet has columns referring to a field, a case number and a value. From what was told by the company, this table reflects how clients answered the form to ask for their request, so value refers to the answer they gave to each question. The final table has data on rejected tasks and their BPM requirements.

3.2. DATA PREPARATION

In order to do the modelling, all tables were joined together. To achieve that, some transformations were performed. The first step was in the rejections table, where all the BPM requirements were grouped for each task in a list. Then, the specific requests table was transformed into a pivot table where each row was a request identifier, the columns were each field, and the values were from the column value. Then, for the user information table, the binary column related to the manager dropped, replaced the birth year column with an age column, removed the duplicate rows and kept the most recent entry, and replaced the missing and odd values in the gender column with a value indicating they were bots.

These treated tables were then merged into our untreated task execution table. Then all columns referring to dates were converted into the datetime formats and a column "Time Surpassed" to show how much longer a task took to be executed than predicted, with it showing 0 if it was faster than the predicted time. Regarding the missing values existent in this table, different approaches were taken according to the situations: removed the action column since it had their corresponding ID's in the "idBPMApplicationAction" column; there were no alterations in the "Task predicted end date" column since its missing values mostly corresponded to the initial task; since the BPM actions only appear for

rejected tasks, most rows will have this column empty so those suffered no change; task executer and task executer department also show up empty for most rows, but need to be filled out. This was done by assigning the missing department the value “-1” and by assigning a missing executer a negative value according to the department they belong to. With the missing values taken care of, the last task was used and action to map the outcome of each case: If the action was 299, the case was finished successfully, mapped as 1; If the final task was 104, the request was closed because the requester rejected the accounting impact, mapped as 4; if the final task was 101, 102 or 105 the request was cancelled, mapped as 3; other combinations led to the request being closed administratively, mapped as 2. An ongoing case was also found so it was removed.

It was now time to move on to prefix extraction and bucketing. Prefix extraction consists of taking into account only a set number of tasks for each case. For the first prefix, at the moment of the first task, a table was created for each case, only was taken into account the month and hour when the task arrived merged with the specific requests pivot table, as well as an additional column calculating the time difference between the form being filled by a client and the task arriving, and another column with the outcome. For the second prefix, at the time of the second task, information was added about the first task, such as the time surpassed, the number of departments involved, the mean age of executors, the mean time on department, if outsourcing was involved, the duration in hours, the number of executers and who executed it, and if there are any BPM requirements. From the third to the eight prefixes, columns were added with the ID of the activities and actions that have been concluded thus far, as well as a column that indicates if the action 290 has taken place. If a case ends in a number of activities smaller than what is being considered for a certain prefix, that case is removed from that and any future prefix table.

Although a direct outlier removal step was not performed, the prefix bucketing process involved using the data of processes up to the 8th task. In this context, outliers were implicitly handled as any processes or data extending beyond the 8th task. Therefore, in cases where a process has more than 8 tasks, only the first 8 tasks were considered. This approach helped in organizing the data and highlighting the most important part of the data, as processes with a long duration are rarer.

Finally, feature selection was performed on each prefix table by evaluating the variance, Spearman correlation (evaluates the correlation between numerical variables), Cramer V (evaluates association between categorical variables), ANOVA (evaluates the variance between all variables) and feature importance (using a decision tree classifier). As a result, the following columns were removed: from prefix_1 and prefix_2, column “ID_1602”, from prefix_3, columns "ID_1602" and "Total Time surpassed (hours)"; from prefix_4, columns "ID_1602", "Total Time surpassed (hours)" and "Number of Executers"; from prefix_5, columns "ID_1602", "W/ BPM Requirements", "Activity_2", "Total Time surpassed (hours)" and "Action_2"; from prefix_6, columns "ID_1602", "W/ BPM Requirements", "Activity_2", "Total Time surpassed (hours)" and "Out-Sourced Involved"; from prefix_7, columns "Action_2", "Activity_2", "Total Time surpassed (hours)", "Out-Sourced Involved"; from prefix_8, columns “Activity_6", "Activity_2", "Total Time surpassed (hours)", "Out-Sourced Involved", "Activity_5", "Activity_4".

3.3. MODELLING

With all the datasets ready to be used, it was time for the modelling stage. Here the goal was to use each prefix to predict the outcome of the case. To achieve this, the following models were tested on all of our datasets: Decision Tree, Random Forest, Gaussian Naïve Bayes, Logistic Regression, Gradient Boosting Classifier, Light GBM Classifier, Extreme GBM Classifier, Cat Boost Classifier, Ada Boost Classifier. Subsequently, for each prefix bucket iterations were made with the goal to find the best parameters of the models that presented the best scores in the initial stage. Most datasets worked better with Random Forest and variations of Gradient Boosting model (light, cat, extreme).

3.4. EVALUATION

In the evaluation phase of our predictive modeling project for Business Process Management (BPM) processes, we meticulously assessed the performance of various machine learning models. Our objective was to predict the end outcomes of BPM processes with accuracy and reliability, ultimately aiming to enhance efficiency and decision-making within our organization.

After conducting rigorous data preprocessing and model development, we evaluated the performance of each model using cross-validation techniques. Our analysis focused on key metrics such as accuracy, precision, recall, and F1-score to gauge the models' effectiveness in predicting BPM outcomes.

To compare these results across different datasets, we assessed the models by their recall, precision, and F1-score for all outcomes. Additionally, cross-validation scores (with accuracy) and comparisons of accuracy on the training and validation datasets were utilized to evaluate overfitting and rate the models in a more global way.

Based on our evaluation, the following models were chosen for each dataset prefix: Prefix_1: Light GBM and XGBoost; Prefix_2: XGBoost and Random Forest; Prefix_3: XGBoost and Light GBM; Prefix_4: XGBoost and Light GBM; Prefix_5: Random Forest, Light GBM, and Gradient Boosting; Prefix_6: XGBoost, Light GBM, and Random Forest; Prefix_7: Random Forest and CatBoost; Prefix_8: CatBoost and XGBoost.The scores obtained can be found in figure 4.

Predicting Final Outcome before the process is picked up by a team *[prefix 1]*

Model	Request Finish	Closed Administrative	Request Canceled	Closed Ad. Rquester Rej. Acc. Impact	*Cross-Validation Score*	*Overfit*
Light GBM	0.84	0.46	0.95	0.73	0.698	1.17%
XGBoost	0.84	0.47	0.95	0.73	0.695	1.46%

Predicting Final Outcome after first task ended *[prefix 2]*

Model	Request Finish	Closed Administrative	Request Canceled	Closed Ad. Rquester Rej. Acc. Impact	*Cross-Validation Score*	*Overfit*
XGBoost	0.86	0.70	0.95	0.81	0.793	6%
Forest Forest	0.84	0.49	0.94	0.74	0.698	0.52%

Predicting Final Outcome after the second task ended *[prefix 3]*

Model	Request Finish	Closed Administrative	Request Canceled	Closed Ad. Rquester Rej. Acc. Impact	*Cross-Validation Score*	*Overfit*
Light GBM	0.86	0.71	0.96	0.81	0.793	13.7%
XGBoost	0.86	0.69	0.96	0.81	0.787	7.55%

Predicting Final Outcome after the third task ended *[prefix 4]*

Model	Request Finish	Closed Administrative	Request Canceled	Closed Ad. Rquester Rej. Acc. Impact	*Cross-Validation Score*	*Overfit*
Light GBM	0.86	0.68	0.97	0.82	0.799	2.88%
XGBoost	0.87	0.69	0.97	0.82	0.804	3.45%

Predicting Final Outcome after the forth task ended *[prefix 5]*

Model	Request Finish	Closed Administrative	Request Canceled	Closed Ad. Rquester Rej. Acc. Impact	*Cross-Validation Score*	*Overfit*
Random Forest	0.89	0.89	0.86	0.94	0.905	~0%
Light GBM	0.91	0.91	0.91	0.96	0.925	3.63%
Gradient Boosting	0.90	0.91	0.88	0.95	0.919	0.05%

Predicting Final Outcome after the fifth task ended *[prefix 6]*

Model	Request Finish	Closed Administrative	Request Canceled	Closed Ad. Rquester Rej. Acc. Impact	*Cross-Validation Score*	*Overfit*
Random Forest	0.95	0.93	0.78	0.73	0.919	1.16%
Light GBM	0.96	0.94	0.84	0.80	0.931	1.94%
XGBoost	0.96	0.94	0.90	0.80	0.931	1.90%

Predicting Final Outcome after the sixth task ended *[prefix 7]*

Model	Request Finish	Closed Administrative	Request Canceled	Closed Ad. Rquester Rej. Acc. Impact	*Cross-Validation Score*	*Overfit*
Random Forest	0.83	0.82	0.67	0.63	0.814	3.40%
CatBoost	0.88	0.85	0.73	0.77	0.856	4.85%

Predicting Final Outcome after the seventh task ended *[prefix 8]*

Model	Request Finish	Closed Administrative	Request Canceled	Closed Ad. Rquester Rej. Acc. Impact	*Cross-Validation Score*	*Overfit*
CatBoost	0.89	0.87	0.82	0.76	0.861	3.89%
XGBoost	0.88	0.87	0.86	0.75	0.858	5.49%

4. RESULTS EVALUATION

On the grounds that not all data was made available and real-life application always differs from theory, it is necessary to evaluate before its application so the company must make its own assessment of the models so that it is robust, accurate and deployment ready. The metrics recommended were made available in the Notebook, which ,essentially, are accuracy, precision, recall, F1-Score, and the Confusion Matrix.

Nevertheless, its actual quality is only verified once putted to use. Therefore, it is crucial that after a period of 3 months the company performs once again these calculations. Subsequently, once again adapt the model to execute better. Afterwards, it is believed that it should be checked and maintained periodically, by semester or annually.

5. DEPLOYMENT AND MAINTENANCE PLANS

5.1. GENERAL IDEA

The process starts with the company selecting the right projects to them, share it with the responsible entities within the company, to then develop the model the most accurate as possible to retain as much value as possible out of it. Subsequently, a continuous checkup on the model is required because models are affected by a wide range of factors that may change the necessary characteristics of the model.

5.2. ROADMAP

Phase 1: Initiation and Planning

- **Project Evaluation (3rd June – 17th June) -** From all projects made available select the one with the best quality.
- **Pitch the Models to all Stakeholders (17th June) -** Show the results obtained and show them to the people that hold the decision power and everyone that would have to be involved so that they could agree or not with the projects feasibility.
- **Project Acceptance (18th June – 21st June) -** Period of time to reflect on the project and whether or not, the project should be continued.
- **Resource Allocation (24th June – 28th June) -** Assign teams and allocate resources.

Phase 2 - Data Collection and Preparation

- **Data Inventory (1st July – 5th July) -** Look into the data available useful for the model.
- **Data Cleaning and Preparation (8th July – 19th July) -** Preprocessing data (missing values, outliers). Check if it is necessary any external data
- **Data Integration (22nd July – 2nd August) -** Put together all the data, organise it and format it.
- **Initial Data Analysis (5th August) -** Understand the final data available.

Phase 3 - Model Development

- **Feature Engineering (5th August – 16th August) –** Develop useful features for the model.
- **Model Selection (19th August – 23rd August) –** Check which algorithms best fit the model to reach the desired outcome.
- **Model Training (26th August – 6th September) –** Using data already available to put the model into work.
- **Model Evaluation (9th September – 13th September) –** Evaluate if the model is working as expected or if it needs tuning.

Phase 4 - Model Deployment

- **Model Improvement (16th September – 20th September) –** Considering the results obtained in the previous phase some parameters always have to be improved for better results.
- **Implementation Plan (23rd September – 27th September) –** Adaptation of the proposed plan into the norms of Millenium BCP.
- **Model Deployment (30th September – 4th October) –** Application of the model into the business.
- **Model Monitoring (7th October – 11th October) –** Check if it works correctly in real life usage.

Phase 5 - Maintenance

- **Tuning (24th October – 25th October) –** Final tweaks to improve the quality even more.
- **Employee Training (28th October – 8th November) –** Train users to use it in their daily functions.
- **Routine Maintenance (11th November – 22nd November) –** Establish the routine procedures.

5.3. COSTS

This process implies costs, namely, infrastructure, Software, Human Resources, Training and Maintenance, and Data storage. Taking into consideration that it is very likely that the company already has the needed infrastructures and software, the only possible costs that these can have, are the licenses which were estimated to be around 2 000€ (for the 6 months of the model design). Regarding Human resources IT people already employed can contribute to the project without implicating added costs but hiring a new data scientist could be of great help for the company since not only this project includes maintenance needs, but also would make him available for any other projects and decrease the Data Science team workload. Consequently, it could cost around 12 500€, for the 6-month period. Moreover, training could be performed by this extra employee since not only this person would be the one more acquainted with the system but also because it is the team member with least responsibilities as that person is new in the team. Subsequently, it should also be this person's job to perform maintenance on the model. However, if any issues happen, it can imply some costs that it is why the budget includes a 10% contingency plan. Lastly, data storage should not be a problem, since banks have a great deal of data available to them and so they are already prepared for this type of storage need.

In conclusion, the final budget should be, approximately, 16 000 € (2 000€ + 12 500€ + 14 500*0.1 = 15 950€).

5.4. BENEFITS

Although this project implies high costs to the company, it is believed that the advantages outweigh the disadvantages. The implementation of a Business Process Predictive Monitoring Model allows to prevent and detect frauds, increases efficiency, improves the quality of risk management and customer retention and satisfaction as well as decreases costs in regulatory compliance. Therefore, once the model starts working on an optimal quality for Millennium BCP, the bank should be able to save up at least one million euros.

6. CONCLUSIONS

In this report, a comprehensive data mining project was undertaken to predict the end outcomes of Business Process Management (BPM) processes within Millenium BCP. Through rigorous data collection, preprocessing, modelling, and analysis, valuable insights were gained into the factors influencing the final results of these processes, and predictive models were developed to forecast these outcomes.

The findings indicate that predictive modelling holds significant promise in enhancing the efficiency, reliability, and performance of BPM workflows within the bank. Leveraging advanced machine learning algorithms and techniques, satisfactory levels of predictive accuracy were achieved, enabling the forecast of end outcomes of BPM processes with confidence.

The predictive models demonstrated robust performance, surpassing baseline benchmarks in terms of accuracy rates. Rigorous parameter testing and validation established the reliability and efficacy of these models in forecasting the final results of BPM processes across different stages of the process lifecycle.

In conclusion, the data mining project has provided actionable insights and predictive capabilities that can drive tangible benefits for Millenium BCP. By harnessing the power of predictive analytics, the bank can make informed decisions, optimize operations, mitigate risks, and ultimately achieve its business objectives in a dynamic and competitive landscape. However, theory and practise are different so the team would like to know how can it help in the next steps?

6.1. CONSIDERATIONS FOR MODEL IMPROVEMENT

Even though the Millenium BCP has available high-quality data, it still does not have some processes mapped which makes data preprocessing more time consuming and less accurate since for certain processes assumptions were made and it is impossible to be assured if it is an error or not. Consequently, it is advised that the bank maps these to not only save money but also to increase accuracy of the model.

7. REFERENCES

OpenAi. (2022b, November 30). *ChatGPT*. Chatgpt.com; OpenAI. https://chatgpt.com

A nossa História - Millennium BCP. (n.d.). Ind.millenniumbcp.pt. Retrieved May 30, 2024, from https://ind.millenniumbcp.pt/pt/Institucional/Pages/historia.aspx

John, B. (2022, July 22). When to choose CatBoost over XGBoost or LightGBM [practical guide]. Neptune.Ai. https://neptune.ai/blog/when-to-choose-catboost-over-xgboost-or-lightgbm

(N.d.). Datacamp.com. Retrieved May 30, 2024, from https://www.datacamp.com/blog/classification-machine-learning