# CCT College Dublin

## Assessment Cover Page

| **Module Title:** | Problem Solving for Industry |
| --- | --- |
| **Assessment Title:** | Capstone Pair Project |
| **Lecturer Name:** | Dr. Muhammad Iqbal |
| **Student Full Name:** | Caroline de Sa Teixeira, Heber Mota |
| **Student Number:** | 2020331, 2020317 |
| **Assessment Due Date:** | 17/05/2024 |
| **Date of Submission:** | Text |

### Declaration

| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |
| --- |

**Table of Contents**

[**CCT College Dublin**](#_olc80hl9baca) **0**

[Assessment Cover Page](#_a9h7agjn3vk1) 0

[Declaration](#_out22zwh2o13) 0

[Roles & Responsibilities](#_u0sbu71o7pxp) 3

[**Business Understanding**](#_qw8p6cbyadfe) **4**

[Determine Business Objectives](#_6wth4cvemhti) 4

[Assess Situation](#_enfjif3z2is) 4

[Inventory Of Resources](#_ysyuz5k4z7wp) 5

[Computing and Premisses](#_1iuemzuwwxqq) 5

[Personnel](#_9u482s3kf7jm) 6

[Evaluating Risks & Contingencies](#_d1k5xe9nkfzw) 6

[**Data Understanding**](#_edb6lsfr5gkd) **8**

[● Transaction Data](#_if2n57c7o07d) 8

[● User Behaviour Data](#_y1aaxahzmjlp) 8

[● Customer Demographics](#_7o0rus3y6qfn) 8

[● Dynamic Currency Conversion](#_x0tkfkqvib54) 8

[**Data Preparation**](#_ypc3iviyhudo) **9**

[**Deployment**](#_7y18id6upg4w) **10**

[**Conclusion**](#_8uy3gp1m5tve) **11**

[**Referencing**](#_kka0gg2s7aam) **12**

### 

# 

# HELP GUIDELINES FOR REPORT

# 

[RPubs - CRISPR-DM Case Study](https://rpubs.com/Argaadya/crispr_dm)

**report + project guideline:**

[**The Data Science Process (CRISP-DM) - Michael Fuchs Python (michael-fuchs-python.netlify.app)**](https://michael-fuchs-python.netlify.app/2020/08/21/the-data-science-process-crisp-dm/)

# Business Understanding

## Business Objectives

The core objective of the Banking Solutions project is to leverage machine learning techniques to detect and prevent fraudulent transactions within a provided dataset effectively. Our approach involves constructing a robust system adept at managing financial data while also integrating additional features such as a currency converter to enhance financial management capabilities. Furthermore, we aim to utilize the inherent properties of the dataset to generate insightful visualizations depicting customer behavior patterns. These visualizations will empower our clients to identify trends and valuable insights crucial for their business operations. Through these features, we anticipate the development of a comprehensive Financial Manager and Banking Solutions product that stands out in the market, offering competitiveness and completeness.

## Assess Situation

Before starting the implementation phase of our project, it's important to conduct a thorough assessment of the current situation. This assessment involves several key steps to ensure a clear understanding of the project's context, objectives, and constraints.

**Define Objectives:** The first step is to clearly define the objectives of the project. This involves understanding the overarching goals and desired outcomes. In the case of our Banking Solutions project, the primary objective is to detect and mitigate fraudulent transactions using machine learning techniques.

**Identify Stakeholders:** It's essential to identify our stakeholders involved in the project, including clients and end-users to understand their needs, expectations, and concerns will help in tailoring the project to meet their requirements effectively. In our case they are financial entities such as banks that seek to better understand their customer behaviour and protect them from fraudulent transactions.

**Assess Data Availability and Quality:** The success of our project relies heavily on the availability and quality of the data. We need to assess the accessibility of relevant datasets and evaluate their completeness, accuracy, and consistency. Additionally, we must consider any data privacy and security regulations that may impact our data collection and usage.

**Analyze Technical Requirements:** Next, we need to analyze the technical requirements for implementing our solution. This includes assessing the hardware, software, and infrastructure needed to support our machine learning models, data visualization tools, and other components of the solution.

**Consider Time and Resource Constraints:** Time and resource constraints play a crucial role in project planning and execution. We need to evaluate the project timeline, budget (which is not necessary for this project), and available resources to ensure feasibility and manage expectations effectively.

**Risk Assessment:** Identifying and assessing potential risks is essential for proactive risk management. We need to analyze potential risks related to data quality, technical challenges, regulatory compliance, and project dependencies. Developing mitigation strategies for these risks will help minimize their impact on the project.

**Define Success Metrics:** Finally, we need to define clear success metrics that will guide the evaluation of our project's outcomes. These metrics should align with the project objectives and provide measurable indicators of success.

## Determine Data Mining & Machine Learning Goals

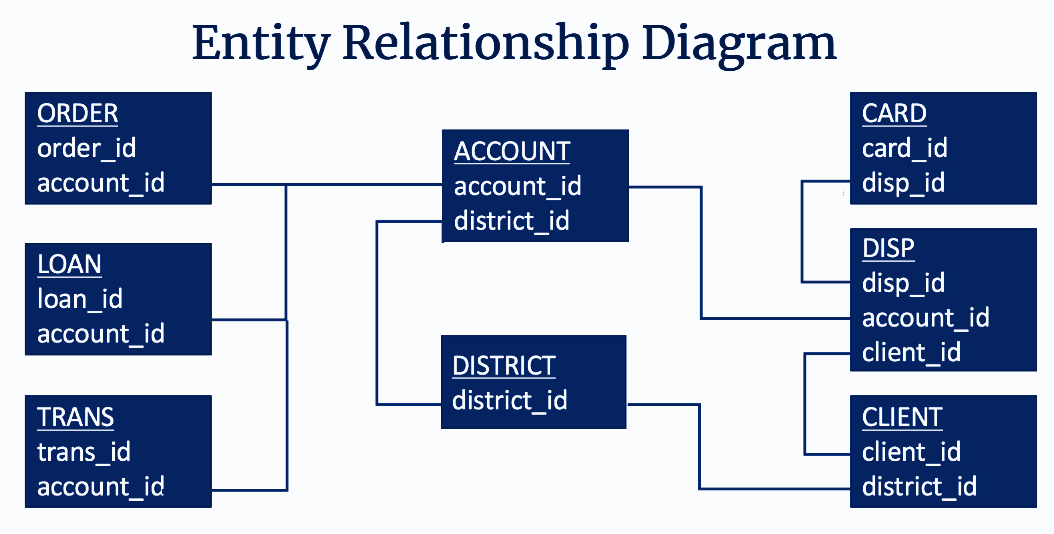
In determining data mining and machine learning goals, it is essential to align them with the overarching objectives of the project. This involves identifying specific tasks and outcomes that leverage data mining and machine learning techniques to extract actionable insights and drive decision-making processes. These goals may include:

* **Fraud Detection:** Utilize machine learning algorithms to identify patterns and anomalies in financial transactions, enabling proactive detection and mitigation of fraudulent activities.
* **Customer Segmentation:** Employ data mining techniques to segment customers based on their transaction behavior, demographics, and preferences, allowing for targeted marketing strategies and personalized services.

These goals aim to process data mining and machine learning capabilities to visualise valuable insights, enhance decision-making processes, and drive business growth. By defining clear and measurable objectives, the project can focus on implementing effective solutions that deliver tangible benefits and add value to the organization.

# Data Understanding

The database, curated by Petr Berka and Marta Sochorova, comprises financial data sourced from a Czech bank, encompassing details of over 5,300 bank clients and approximately 1,000,000 transactions. Furthermore, the dataset includes information on nearly 700 loans and close to 900 credit cards provided by the bank, all of which are represented within the data.

Regarding the data structure, each account in the dataset possesses both static attributes (e.g., creation date, branch address) delineated in the "account" relation, and dynamic attributes (e.g., transaction debits or credits, balances) provided in the "permanent order" and "transaction" relations. The "client" relation describes the attributes of individuals authorized to manage accounts, where one client may have multiple accounts, and multiple clients can manipulate a single account. The connections between clients and accounts are articulated in the "disposition" relation. Additionally, the "loan" and "credit card" relations outline services offered by the bank to its clients, with the possibility of multiple credit cards being issued to one account and a maximum of one loan being granted per account. Furthermore, the "demographic data" relation furnishes publicly available information about districts (e.g., unemployment rates), with potential insights into client demographics derived from this data.

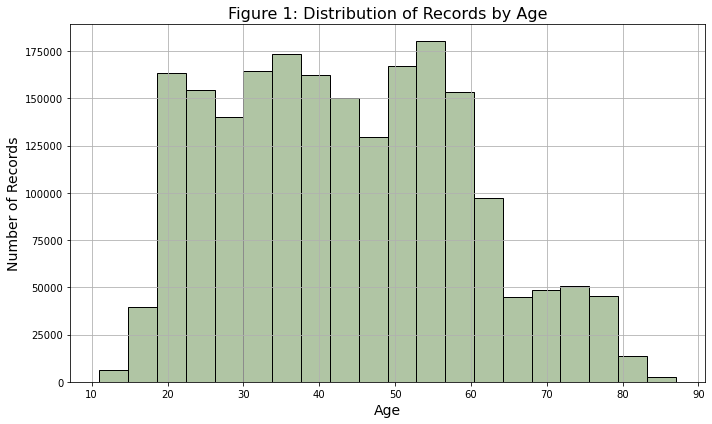
### 

### MIT Licensing

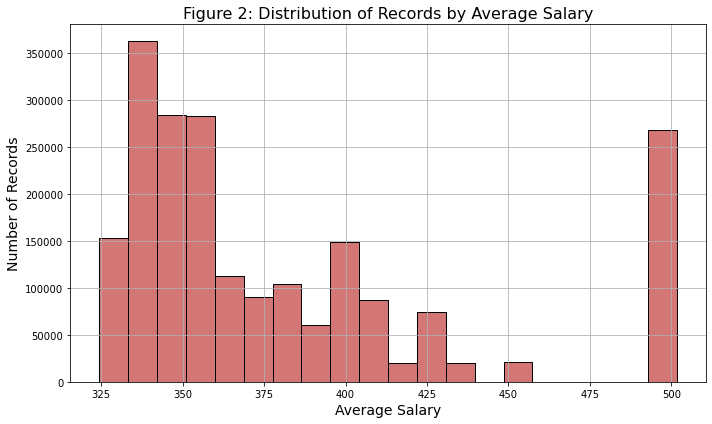
| *Copyright 2024, Heber Mota, Caroline de Sá,*  *Permission is hereby granted, free of charge, to any person obtaining a copy of this dataset and associated documentation files (the "Dataset"), to deal in the Dataset without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Dataset, and to permit persons to whom the Dataset is furnished to do so, subject to the following conditions:*  *The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Dataset.*  *THE DATASET IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE DATASET OR THE USE OR OTHER DEALINGS IN THE DATASET.*  Source: [1999 Czech Financial Dataset - Real Anonymized Transactions - Public Domain Dataset | data.world](https://data.world/lpetrocelli/czech-financial-dataset-real-anonymized-transactions) |
| --- |

## Exploring Data

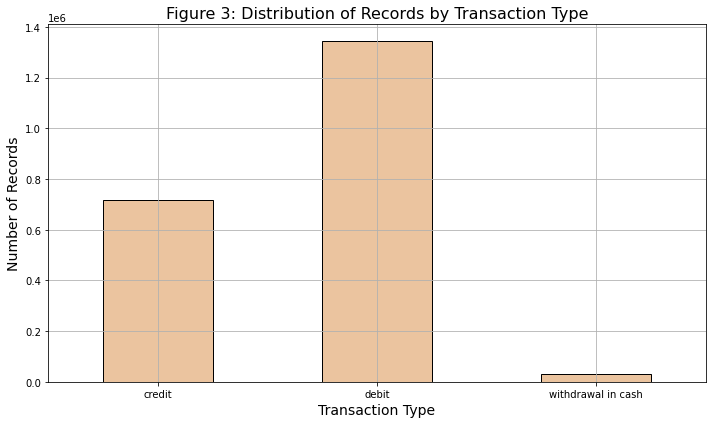
After conducting an extensive Exploratory Data Analysis (EDA), we have gained valuable insights into the intricacies of our dataset. This comprehensive analysis has shown interesting features such as the ones presented below:

**Figure 1:**  The graph portrays a notable concentration of individuals aged between 18 and 65, indicating a considerable level of activity within this demographic range. Furthermore, within this range, there is a distinct peak observed particularly within the age group between the 30s and 50s. 

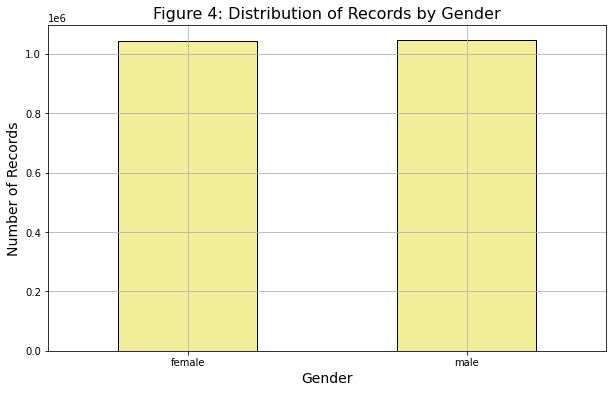
**Figure 2:** The majority of records cluster within the range of 330-360 euros on average, suggesting a prevalent income bracket among the sampled population. Interestingly, there is a secondary peak observed around 500 euros, indicating another significant concentration of wealth. It's worth noting that these figures have been converted from Czech Koruna to Euro, providing a comprehensive perspective on the financial landscape within the dataset. This observation underscores the existence of distinct salary groups within the population, potentially reflecting varying socio-economic factors or industry trends.

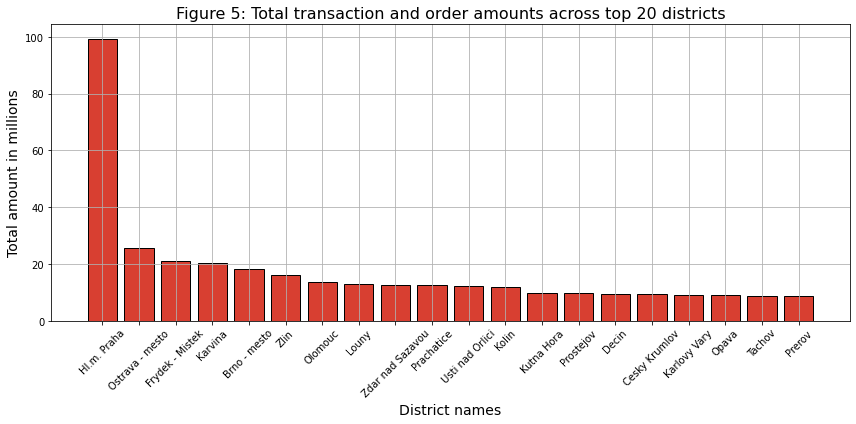


**Figure 3:** The graph illustrates a clear trend in transactional preferences, with debit cards emerging as the most commonly utilized payment method, followed by credit transactions and cash withdrawals. This observation underscores the significance of digital payment methods in modern financial transactions, reflecting a widespread reliance on debit cards for everyday purchases and transactions. Additionally, the prominence of credit transactions highlights the continued relevance of credit-based purchasing in consumer spending habits.

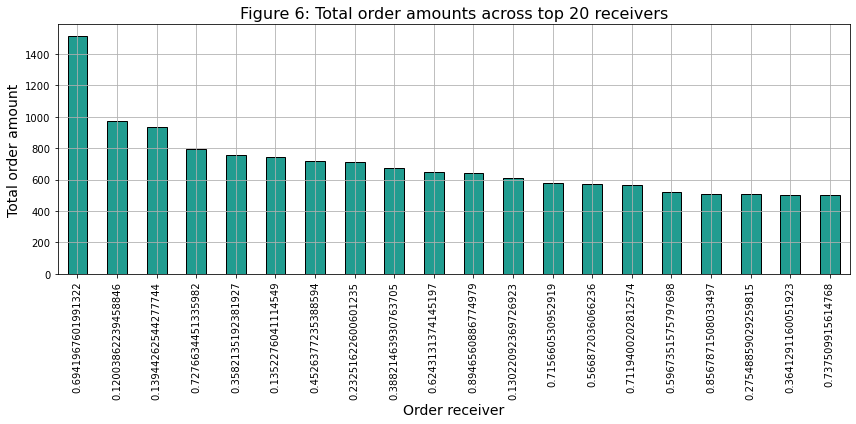


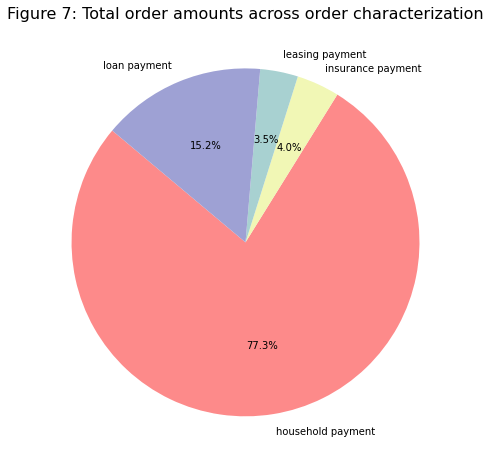
**Figure 4:** The distribution of gender within the records demonstrates a balanced representation between male and female individuals. The even distribution suggests a fair and equitable representation of both genders, which is essential for ensuring unbiased analyses and decision-making processes. This gender balance within the dataset provides a solid foundation for conducting comprehensive and inclusive analyses, ultimately contributing to more robust insights and outcomes.



**Figure 5:** The graphical analysis reveals notable patterns in transaction activity across different districts. The capital city, Prague, is demonstrating a significantly high transaction volume exceeding 90 million units. Following behind is Ostrava, Mesto, with transaction activity hovering over the 20 million mark, indicating a substantial level of economic activity in the region. Additionally, Frydek, Mistek, emerges as another active district, with transaction volumes reaching approximately 20 million units. These observations underscore the diverse economic landscapes and transactional dynamics present across various districts, with Prague standing out as a major hub of financial activity.

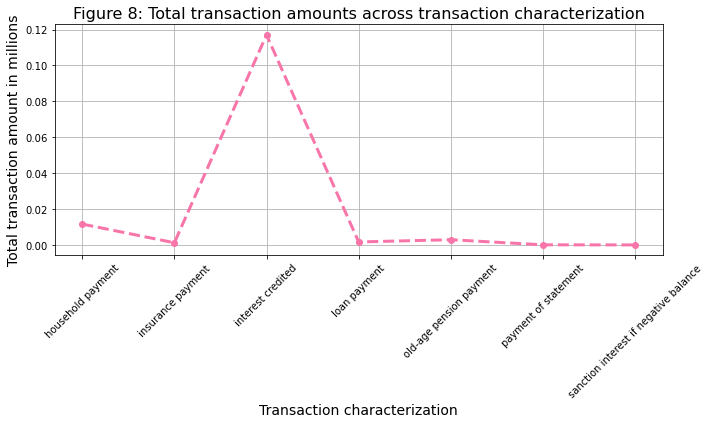
**Figure 6:** Upon examining the graph depicting the total order amount across the top 20 receivers, a striking observation emerges. It becomes evident that a particular receiver stands out significantly, commanding over 1400 orders, which contrasts with the second and third place receivers, each with approximately 900 orders. The remaining receivers in the top 20 exhibit a gradual decline in order volumes, ranging from 800 to 500 orders. This discrepancy emphasizes the dominance of the top-ranking receiver and highlights the considerable gap between their order volumes.



**Figure 7:** Household payments constitute the predominant share, comprising 77.3% of the total order amounts. Following behind, loan payments account for 15.2% of the total, indicating a significant but comparatively lesser portion. Meanwhile, leasing payments and insurance payments represent around 4% of the total order amounts each. This distribution highlights the substantial contribution of household payments to the overall transaction volume, underscoring their importance within the analyzed dataset.

## 

**Figure 8:** While the previous visualization emphasized household payments in terms of appearances, a notable contrast emerges with the category of interest credited in values. In this instance, the transaction amounts significantly surpass those of other categories. This observation underscores a distinctive pattern where interest credited transactions exhibit substantially higher values compared to other types of transactions depicted in the graph.

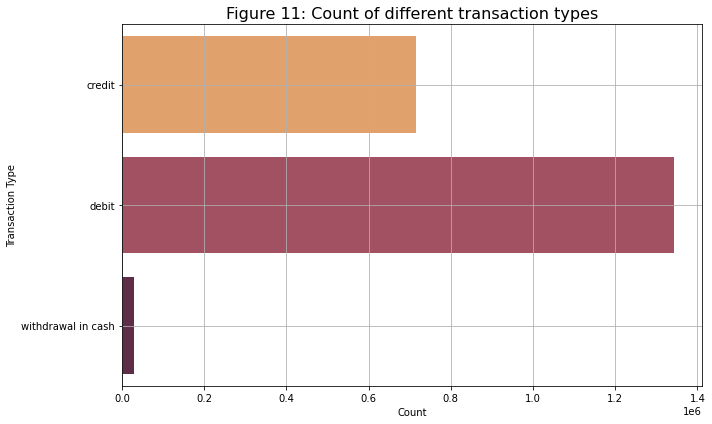


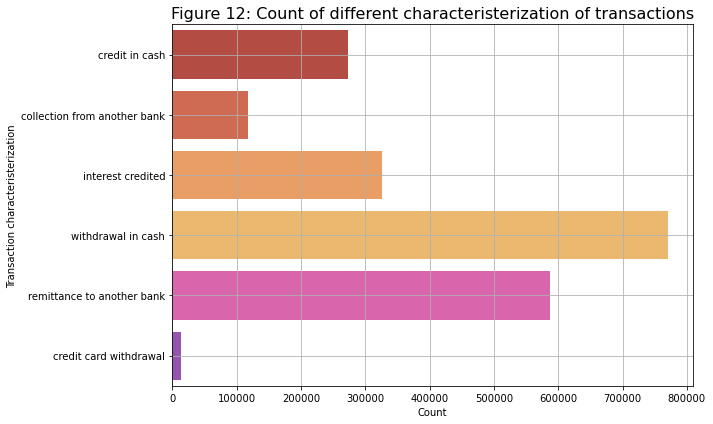
**Figure 9:** The data illustrates that credit card transactions predominantly stand within the lower spending range, typically not exceeding 750 transactions. In contrast, transactions made with debit cards extend beyond this threshold, reaching up to 2.000 in transactions. Additionally, a small proportion of transactions involve cash withdrawals exceeding 2.000. This observation underscores the differing spending behaviors associated with each transaction type, with credit card transactions skewed towards lower values, while debit card transactions encompass a broader range of expenditure.

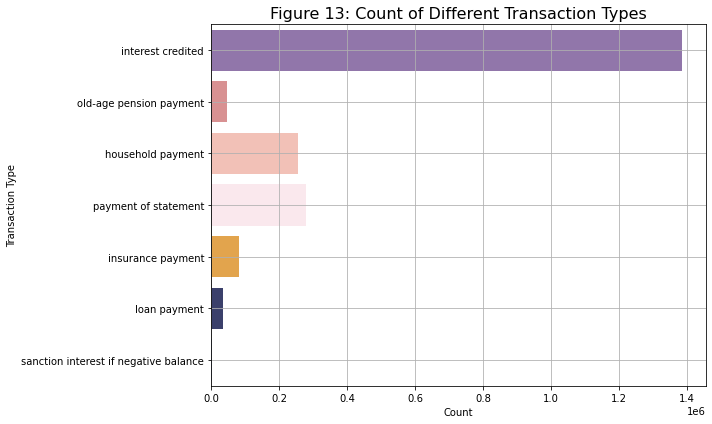
## 

## Figure 10: T

## 







## Verify Data Quality

a

# Data Preparation

In our project workflow, we made a choice to exclude the card and loan datasets from our analysis. These datasets were irrelevant as they did not provide any pertinent information necessary for the application of our machine learning model. However, the disp dataset played a crucial role in facilitating the integration of the account and client datasets. Yet, after this integration was achieved, all columns from the disp dataset were promptly discarded, as they no longer served a purpose in our analysis.

Before merging the datasets, we took a series of preparatory steps to ensure data integrity and coherence. Firstly, we designated the account\_id column as the index, facilitating seamless merging operations. Subsequently, any columns that did not contribute to the merging process or the model application were removed from consideration. To enhance clarity and comprehension, most columns underwent a renaming process to provide better insight into the data.

Furthermore, we added to the dataset additional information from the client's date of birth, specifically creating two new columns: gender and age. Additionally, to standardize the data and facilitate interpretation, transaction operation and characterization values were translated from Czech to English.

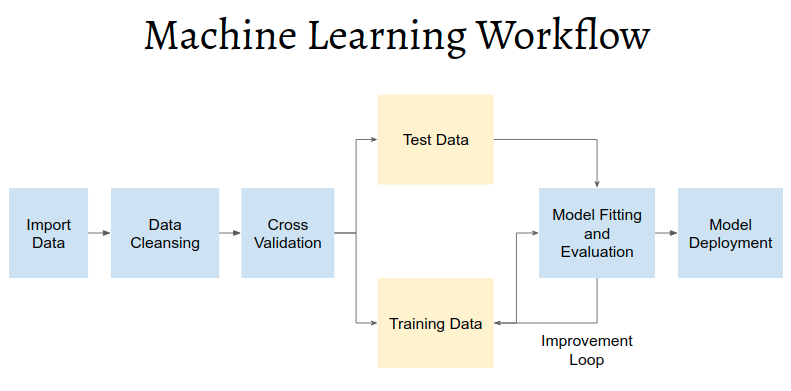
Moreover, currency-related columns underwent conversion from Czech Koruna to Euros to maintain uniformity and better understanding. To address missing data, null values were imputed using information sourced from other datasets, employing statistical measures such as mean and mode values.

Upon completion of these preparatory measures, the six remaining datasets were consolidated through merging operations, primarily utilizing the account\_id column as the key identifier. Subsequently, further redundant columns were eliminated from the merged dataset, and a systematic order was established for the retained columns.

By meticulously executing these preparatory steps, we ensured that our datasets were appropriately curated and structured, laying a solid foundation for subsequent analysis and model application.

**Modelling**

Modeling: select modeling technique; generate test design; build model; assess model

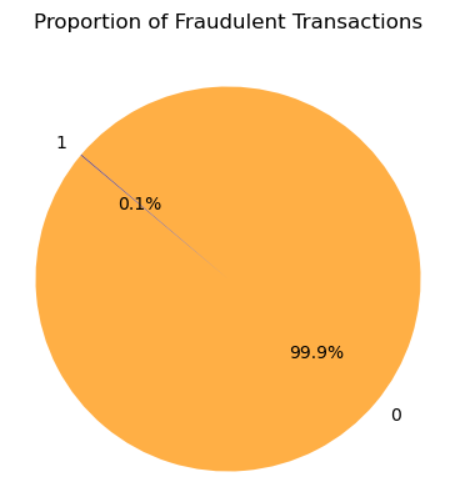
https://rpubs.com/Argaadya/crispr\_dm

**Evaluation**

Evaluation: evaluate results; review process; determine next steps

# Deployment

Deployment: plan deployment; plan monitoring and maintenance; produce final report; review project

****

# Conclusion

Text

Appendix 1: group reflection; Appendix 2: capturing evidence of group work.

## 

# Referencing

### 

1. 1999 Czech Financial Dataset - Real Anonymized Transactions - dataset by lpetrocelli (2019) data.world. Available at: <https://data.world/lpetrocelli/czech-financial-dataset-real-anonymized-transactions> (Accessed: 12 March 2024).
2. PKDD’99 Discovery Challenge (no date) Financial Data description: Available at: <https://sorry.vse.cz/~berka/challenge/pkdd1999/berka.htm> (Accessed: 19 March 2024).
3. Adams, P. (no date) Credit Card Analysis of Czech Bank: Domain - Transactions, Credit Card Analysis of Czech Bank | Domain - Transactions. Available at: <https://webpages.charlotte.edu/mirsad/itcs6265/group1/transaction_domain.html> (Accessed: 19 March 2024).
4. (No date). Available at: <https://escholarship.org/content/qt18d8w42r/qt18d8w42r_noSplash_cb3c2ba106c7a9be7ae91dd4c0de33e4.pdf?t=rrv8zi> (Accessed: 2 March 2024).
5. Philip, A. (2023) Deploying Machine Learning Models with AWS Lambda and SageMaker for Prediction Storage in S3, Medium. Medium. Available at: <https://medium.com/@annmariaphilip8/deploying-machine-learning-models-with-aws-lambda-and-sagemaker-for-prediction-storage-in-s3-c3328e654ec8> (Accessed: 2 March 2024).
6. MS, S. (2023) Read File Data from S3 using Python AWS Lambda, Medium. Medium. Available at: <https://shreyasms.medium.com/read-file-data-from-s3-using-python-aws-lambda-4b3eb515285c> (Accessed: 9 March 2024).
7. Tavasoli, S. (2024) Essential Machine Learning Algorithms: Top 10 to Master in 2024: Simplilearn, Simplilearn.com. Simplilearn. Available at: <https://www.simplilearn.com/10-algorithms-machine-learning-engineers-need-to-know-article> (Accessed: 10 March 2024).
8. (No date). Available at: https://aws.amazon.com/pt/solutions/implementations/fraud-detection-using-machine-learning/ (Accessed: 19 March 2024).
9. Symbl, T. (2023) Performing Anomaly Detection in Python, Symbl.ai. Available at: <https://symbl.ai/developers/blog/performing-anomaly-detection-in-python/> (Accessed: 19 March 2024).
10. What Is Anomaly Detection? (no date) IBM. Available at: <https://www.ibm.com/topics/anomaly-detection#:~:text=Some%20of%20the%20most%20common,values%20of%20the%20selected%20feature>. (Accessed: 19 March 2024).
11. Tuychiev, B. (2023) A Comprehensive Introduction to Anomaly Detection, DataCamp. DataCamp. Available at: <https://www.datacamp.com/tutorial/introduction-to-anomaly-detection> (Accessed: 19 March 2024).
12. GfG (2023) Principal Component Analysis with Python, GeeksforGeeks. GeeksforGeeks. Available at: <https://www.geeksforgeeks.org/principal-component-analysis-with-python/> (Accessed: 19 March 2024).
13. Python - Replace Missing Values with Mean, Median & Mode, Analytics Yogi. Available at: <https://vitalflux.com/pandas-impute-missing-values-mean-median-mode/> (Accessed: 19 March 2024).
14. Real Python (2023) Pythonic Data Cleaning With pandas and NumPy, Real Python. Real Python. Available at: <https://realpython.com/python-data-cleaning-numpy-pandas/> (Accessed: 19 March 2024).
15. (No date) GitHub. Available at: <https://github.com/Kusainov/czech-banking-fin-analysis/blob/master/Czech-banking-customer-trans-analysis.ipynb> (Accessed: 19 March 2024).
16. Kosourova, E. (2023) Types of Data Plots and How to Create Them in Python, DataCamp. DataCamp. Available at: <https://www.datacamp.com/tutorial/types-of-data-plots-and-how-to-create-them-in-python> (Accessed: 19 March 2024).
17. Brownlee, J. (2023) Your First Machine Learning Project in Python Step-By-Step, MachineLearningMastery.com. Available at: <https://machinelearningmastery.com/machine-learning-in-python-step-by-step/> (Accessed: 19 March 2024).
18. Long, A. (2020) Machine Learning with Datetime Feature Engineering: Predicting Healthcare Appointment No-Shows, Medium. Towards Data Science. Available at: <https://towardsdatascience.com/machine-learning-with-datetime-feature-engineering-predicting-healthcare-appointment-no-shows-5e4ca3a85f96#:~:text=Basically%20you%20can%20break%20apart,begins%20in%20the%20prior%20year>. (Accessed: 22 March 2024).
19. MadhuBabu (2021) How to Handle Date and Time Columns in Python: Feature Engineering: Machine Learning, Medium. Analytics Vidhya. Available at: <https://medium.com/analytics-vidhya/how-to-handle-date-and-time-columns-in-python-feature-engineering-machine-learning-254e45a031ae> (Accessed: 22 March 2024).
20. Quick explanation: One-hot encoding (2023) YouTube. YouTube. Available at: <https://www.youtube.com/watch?v=G2iVj7WKDFk> (Accessed: 22 March 2024).
21. Brownlee, J. (2020) Why One-Hot Encode Data in Machine Learning?, MachineLearningMastery.com. Available at: <https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/> (Accessed: 22 March 2024).
22. Team, G. L. (2023) Label Encoding in Python - 2024, Great Learning Blog: Free Resources what Matters to shape your Career! Available at: <https://www.mygreatlearning.com/blog/label-encoding-in-python/#:~:text=Label%20encoding%20is%20a%20technique,only%20operate%20on%20numerical%20data>. (Accessed: 22 March 2024).
23. Ali, M. (2023) Handling Machine Learning Categorical Data with Python Tutorial, DataCamp. DataCamp. Available at: <https://www.datacamp.com/tutorial/categorical-data> (Accessed: 22 March 2024).
24. GfG (2023) *Random Forest Regression in Python*, *GeeksforGeeks*. GeeksforGeeks. Available at: <https://www.geeksforgeeks.org/random-forest-regression-in-python/> (Accessed: 30 March 2024).
25. [RPubs - CRISPR-DM Case Study](https://rpubs.com/Argaadya/crispr_dm)
26. [The Data Science Process (CRISP-DM) - Michael Fuchs Python (michael-fuchs-python.netlify.app)](https://michael-fuchs-python.netlify.app/2020/08/21/the-data-science-process-crisp-dm/)