# 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.

## Produce Project Plan – EDIT

To address this issue effectively, our approach involves taking advantage of AWS services to operate the processing and analysis of our current static financial dataset. We have outlined specific requirements that will guide our development process, which may be refined as we progress:

**1. Develop Machine Learning Algorithm within AWS Lambda Function:**

* Integrate a Machine Learning Algorithm within an AWS Lambda Function.
* Utilize AWS S3 Storage Bucket with two main folders:
* Input Folder: Upload datasets for thorough processing using AWS Lambda Function.
* Output Folder: Store evaluation results for presentation to users in a user-friendly manner.

**2. Implement Amazon Simple Notification Service (SNS):**

* Configure SNS to notify clients about data processing overview and alerts for necessary actions.

**3. Utilize Javascript for User Interface Interaction:**

* Design User Interface using React to provide clients with intuitive access to application insights.

**4. Logging & Monitoring Mechanisms:**

* Implement robust logging and monitoring mechanisms as essential security measures.

**5. Transaction Navigation and Currency Conversion:**

* Enable users to navigate through transactions, distinguishing fraudulent from non-fraudulent ones.
* Implement Currency Conversion functionality for various currencies, adapting to user input.

**6. Summary of Scans:**

* Provide a summary of different scans conducted, offering users comprehensive insights into data analysis.

# Data Understanding

The database, curated by Petr Berka and Marta Sochorova, is commonly referred to as The Berka Dataset in this report. It 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.

## Explore Data

After carrying out Exploratory Data Analysis we have understood interesting features from our Dataset that will be discussed below:

Figure 1- The range of age between 18-60 is the most active, having a peak within the 30’s.

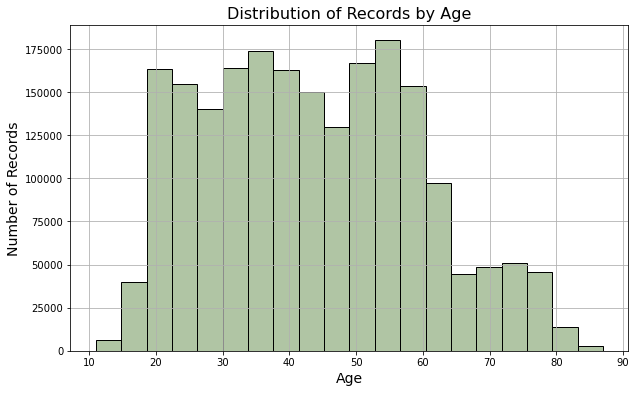


Figure 2- The distribution of average salaries indicates that there are more records around 330-360 euros on average, having a second peak of wealth around 500 euros as well (converted from Czech Koruna to Euro).

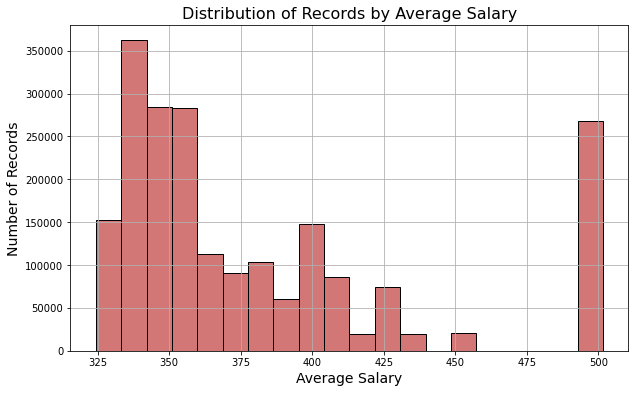


Figure 3- Debit cards were the most used for the transactions obtained, followed by Credit and Withdrawal in cash.

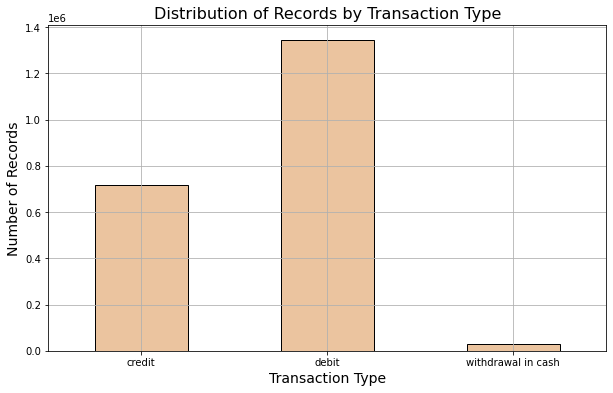


Figure 4- Distribution of gender within records show to be even between male and female.

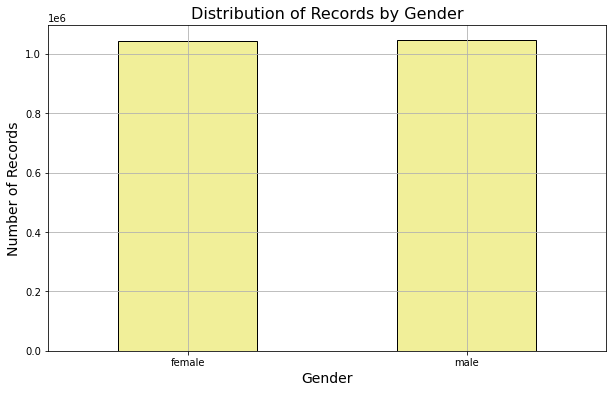


Figure 5- The most active districts in transactions are:

* The capital, Prague – with over 60 million in transactions,
* Followed by Ostrava, Mesto – around 18 million,
* And Frydek, Mistek – around 15 million.

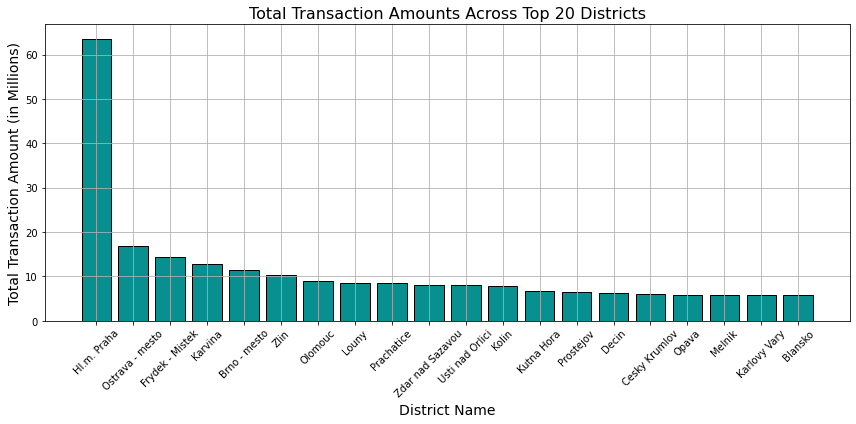


Figure 6- This visualization shows that a receiver in particular soars in order receivings in comparison to others.

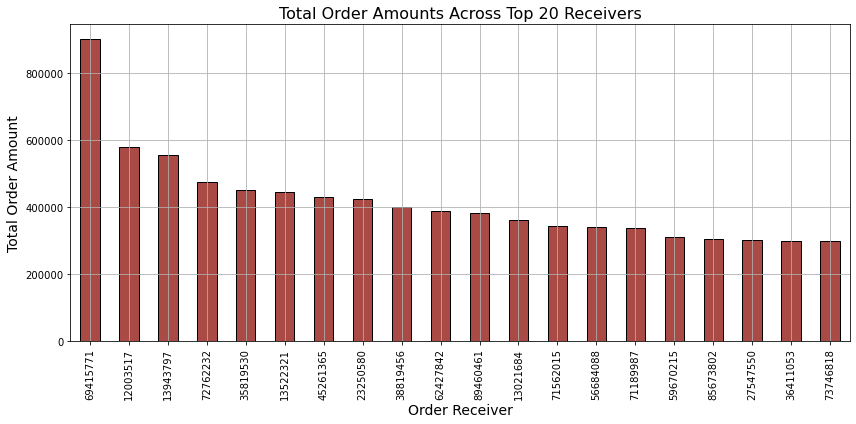


Figure 7- Additionally, we can see that most transactions are dedicated to Household Payments

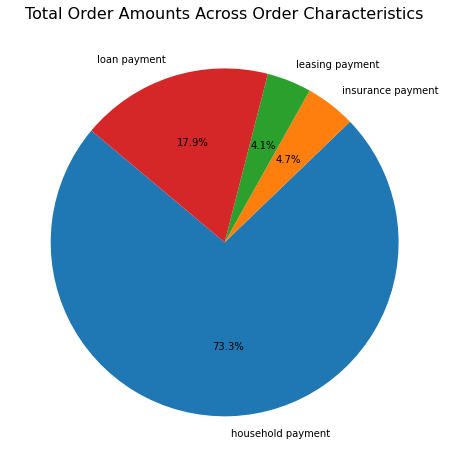
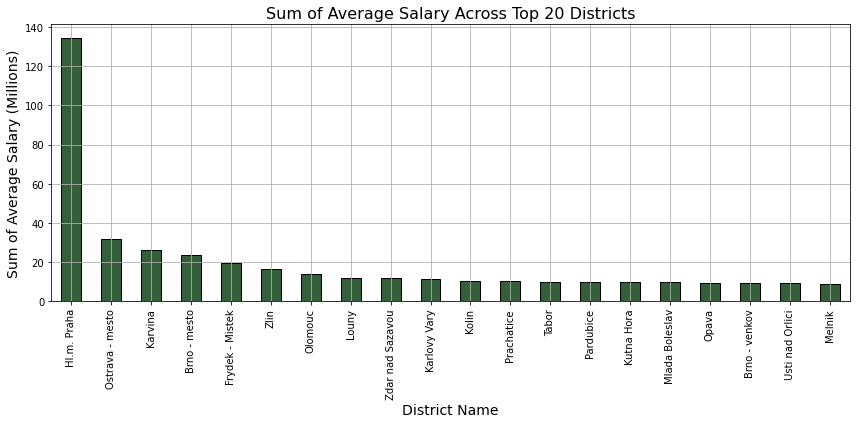


Figure 8- The average salaries per district shows that Prague is the highest by far, followed by Ostrava, and Karvina which both present a similar figure that descends gradually to the following districts from the graph.



## 

Figure 9- By analysing spending range and transaction type we can notice that although most transactions are made using debit cards, the transactions above 500 euros are mainly made using credit cards.

## 

## 

## 

## Verify Data Quality

a

Data Collection Strategy will be used in order to obtain the required data that our software needs to ensure a smooth functioning of its features, here is how we intend to collect and consider the data used and processed for our CH Banking Solution Application:

#### Transaction Data

Once a dataset is uploaded to our S3 bucket, one of the crucial steps is to verify the transaction logs and collect information such as transaction amount, date, time and purchase history so as to identify possible fraudulent activities, transaction patterns and their similarities. We will ensure to comply with Data Privacy Regulations by using the safe AWS Cloud technologies and frameworks.

#### User Behaviour Data

By collecting transaction data we can also define user behaviour which will help understanding and defining what features differ from a regular and a fraudulent behaviour and their patterns, preferences and mechanisms. By doing it, we also will ensure data confidentiality and that Personally Identifiable Information (PII) is strictly being safe.

#### Customer Demographics

By collecting data such as age, location, industry, among other features from customers we can better identify not only customer behaviour but also demographics, creating a robust set of parameters to assist identify fraudulent transactions while still respecting regulations.

#### Dynamic Currency Conversion

As we plan to make our software a strong product being used by clients across the globe, we also intend to use Currency Conversion technologies to enable fast and reliable access to currency conversions from data input and output values in Euro, Dollars and other currencies as desired while considering fluctuations, and financial regulations.

# Data Preparation

We have decided not to use the card and loan datasets since they do not provide any information needed for applying the machine learning model. The disp dataset is used only for merging the account dataset with the client dataset, and all its columns are dropped afterwards.

Many steps were needed before merging all the datasets together:

1. The account\_id column is set as the index.
2. Any columns that are not needed for either merging the datasets or applying the model are dropped.
3. Most of the columns are renamed for better understanding of the data.
4. Two additional columns were created from the client’s date of birth: gender and age.
5. Transaction operation and characterization values were replaced in order to translate them from Czech to English.
6. Columns including currency data were converted from Czech Koruna to Euros.
7. Null values were replaced using data taken from other datasets, mean and mode values.

After the steps above, the six remaining datasets were merged together, mostly by the account\_id column. More unnecessary columns were dropped from the merged dataset, and a new order was established for the remaining columns.

We have decided to not use the card and the loan datasets, since they don’t provide any information needed for applying the machine learning model. The disp dataset is used only for merging the account dataset with the client dataset, and all its columns are dropped afterwards.

Many steps were needed before merging all the datasets together:

1. The account\_id column is set as the index.
2. Any columns that are not needed for either merging the datasets or applying the model are dropped.
3. Most of the columns are renamed for better understanding of the data.
4. Two additional columns were created from the client’s date of birth: gender and age.
5. Transaction operation and characterization values were replaced in order to translate them from Czech to English.
6. Columns including currency data were converted from Czech Koruna to Euros.
7. Null values were replaced using data taken from other datasets, mean and mode values.

After the steps above the 6 remaining datasets were merged together, mostly by the account\_id column, more unnecessary columns were dropped from the merged dataset and a new order was established for the remaining columns.

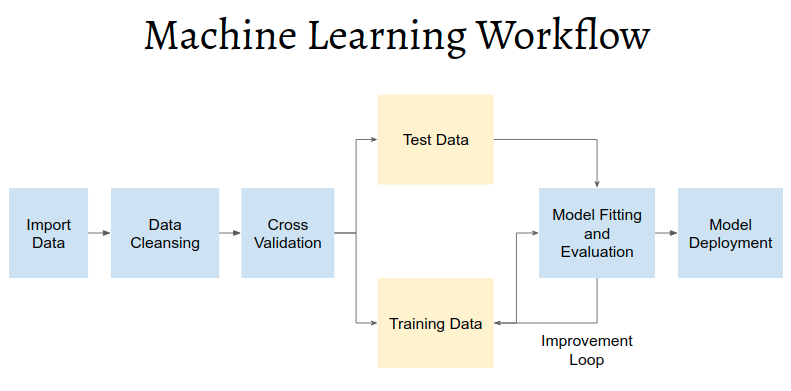
Data Preparation (generally, the most time-consuming phase): select data; clean data; construct data; integrate data; format data

**Modelling**

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

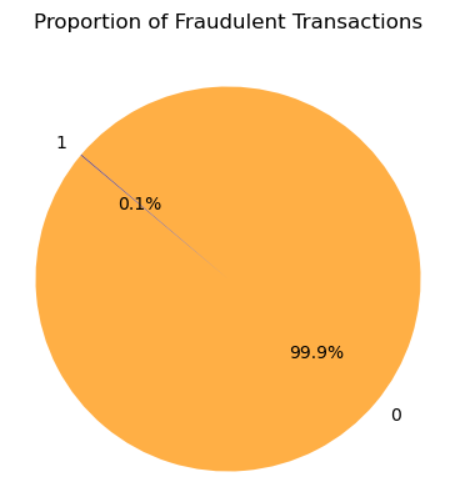
**Evaluation**

Evaluation: evaluate results; review process; determine next steps

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

# 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/)