# 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:** | 17/05/2024 |

## 

## 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**

[Introduction 3](#_Toc166830347)

1. [Business Understanding 3](#_Toc166830348)

[Assess Situation 4](#_Toc166830349)

[Determine Data Mining & Machine Learning Goals 5](#_Toc166830350)

1. [Data Understanding 6](#_Toc166830351)

[MIT Licensing 7](#_Toc166830352)

[Exploring Data 8](#_Toc166830353)

1. [Data Preparation 16](#_Toc166830354)

[Data Cleaning & Integration 16](#_Toc166830355)

[Data Preparation & Construction 16](#_Toc166830356)

1. [Modelling 19](#_Toc166830357)

[Base Models: Random Forest & Decision Tree[7] 19](#_Toc166830358)

[Final Estimator: Logistic Regression[8] 19](#_Toc166830359)

[Ensemble Learning Techniques 20](#_Toc166830360)

1. [Evaluation 21](#_Toc166830361)
2. [Deployment 24](#_Toc166830362)

[Conclusion 26](#_Toc166830363)

[Peer Reflection 27](#_Toc166830364)

[Caroline 27](#_Toc166830365)

[Heber 27](#_Toc166830366)

[Referencing 29](#_Toc166830367)

### 

Word Count: 4,661

# Introduction

A computer with icons on it

Description automatically generated

# 

# Business Understanding

The core objective of CH Banking Solutions is to make use of machine learning to detect fraudulent transactions within a provided dataset effectively. Our approach involves constructing a robust system that will clean and prepare the given financial data while also integrating the machine learning models. Furthermore, we aim to utilize the properties of the dataset to generate insightful visualizations depicting customer behaviour patterns. These visualizations will empower our clients to identify trends and valuable insights crucial for their business operations. Through these features, we will develop a comprehensive Banking Solutions initial product that in the future could stand out in the market, offering competitiveness.

## 

## 

## 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. In order to address these constraints, we are going to make use of **CRISP-DM Framework**[**[1]**](#_kka0gg2s7aam)provided in class as the report guiding and documenting flow of this work. In other words, the following tasks will be achieved:

* **Define Objectives:** The first step is to clearly define the objectives of the project. This involves understanding the goals and desired outcomes. In the case of CH Banking Solutions, the primary objective is to detect and mitigate fraudulent transactions using machine learning techniques. Additionally, we aim to create a User Interface (UI) allowing users to upload datasets and access a dashboard displaying comprehensive statistics on their financial transactions, including fraud analysis, customer demographics, spending insights, and customizable currency conversion.
* **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. The Open-Source Dataset being used "1999 Czech Financial Dataset - Real Anonymized Transactions Dataset" is free to use and so for this work we do not need further permissions.
* **Analyse Technical Requirements:** Next, we need to analyse 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. That said, we are going to mainly code it in Python using Jupyter Notebook which are also free tools. Additionally, in the future we could also mention AWS Cloud Platform or any other Cloud Services Provider to host and process data with our developed models.
* **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 and available resources to ensure feasibility and manage expectations effectively. As we were given around 2 months for the completion, that is what we are going to work with, starting from data cleaning and preparation and then moving on to finding the best models suitable for our data and deploying it.
* **Risk Assessment:** Identifying and assessing potential risks is essential for proactive risk management. We need to analyse 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. After carrying out Exploratory Data Analysis (EDA) that can be accessed in the [Exploring Data](#_6dds1otopg11) section of this work we realised that the data is vast and complete, however it could also present a problem as it seems imbalance from the fraud detection perspective.
* **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. As a Machine Learning Project, our main goal is to clean and prepare the dataset for further development of a model capable of predicting and detecting fraud with the highest accuracy possible. Moreover, deploying it with a User Interface (UI) which will only be possible if we manage to finalise the model in time for the delivery timeframe of this project.

## Determine Data Mining & Machine Learning Goals

In determining data mining and machine learning goals, it is essential to align them with the 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 include Fraud Detection – utilizing machine learning algorithms to identify patterns and anomalies in financial transactions, enabling proactive detection and mitigation of fraudulent activities. By defining clear and measurable objectives, the project can focus on implementing effective solutions that deliver tangible benefits and add value to the organization.

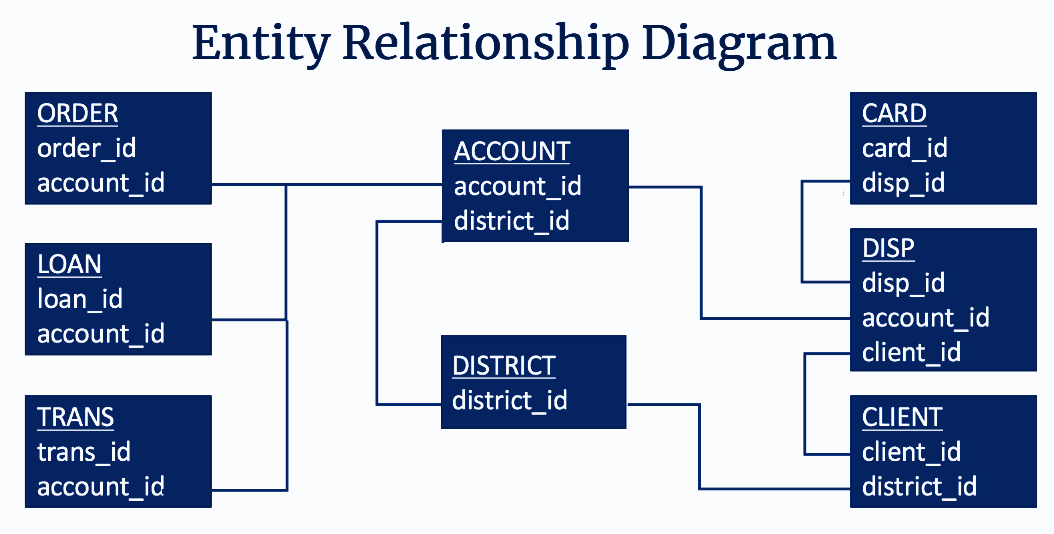
# 

# Data Understanding

The Public Domain 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. *(Petr Berka and Marta Sochorova 2002)* [[2, 3]](#_kka0gg2s7aam).

Below, the Entity Relationship Diagram depicts the relationships of its features as shown:



*Diagram created by Heber Mota*

## MIT Licensing

In ethical and legal terms, it's crucial to acknowledge that the dataset utilized for this project was sourced from an open repository and is governed by the MIT License, the details of which are provided below and in the [Deployment Section](#_7y18id6upg4w) of this work. However, in a real-world scenario, it's imperative for companies to obtain explicit consent from customers before accessing and utilizing their data.

|  |
| --- |
| 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 "1999 Czech Financial Dataset - Real Anonymized Transactions 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) |

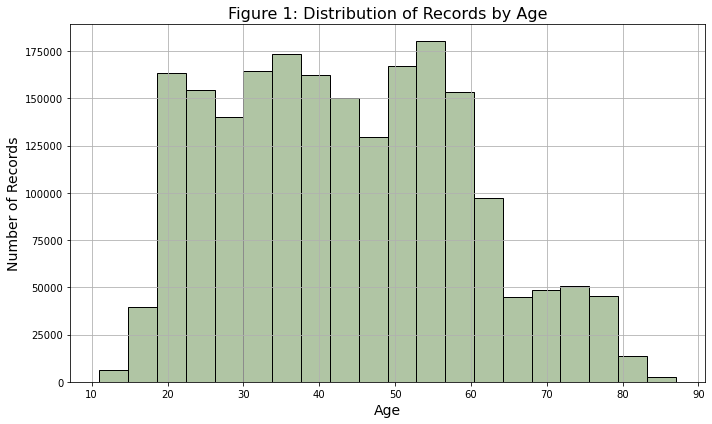
Additionally, adherence to GDPR (General Data Protection Regulation) practices is mandatory to ensure the privacy and security of individuals' personal information. As such, any deployment or utilization of similar models in a commercial setting should involve robust measures to obtain consent, protect data privacy, and comply with relevant regulations and ethical guidelines.

By prioritizing ethical and legal considerations, companies can keep trust and integrity while obtaining data-driven solutions for fraud detection and other purposes.

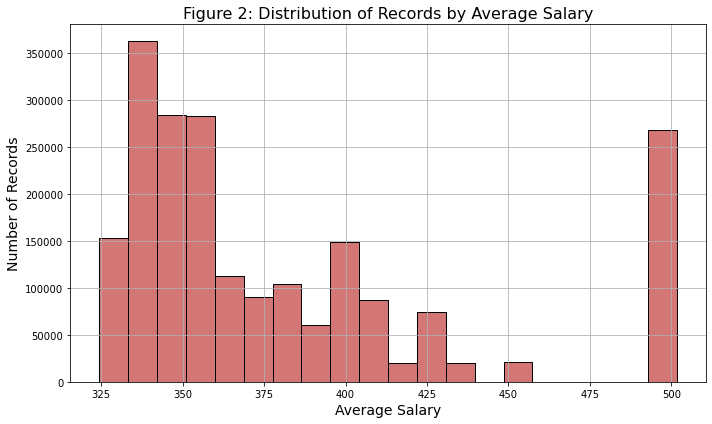
## 

## Exploring Data

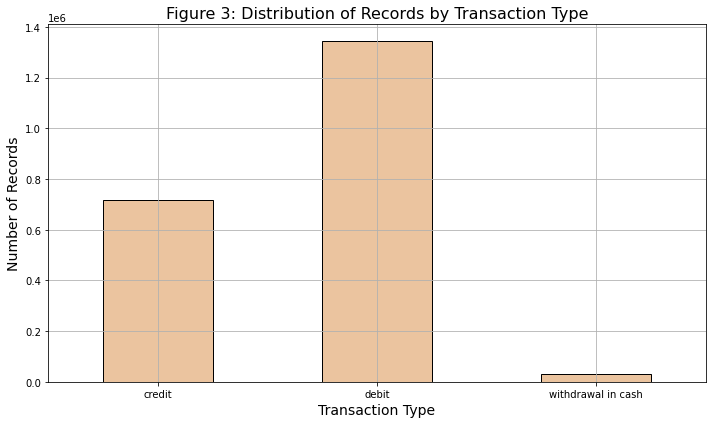
After conducting an extensive Exploratory Data Analysis (EDA), we have gained valuable insights into the intricacies of our dataset. This comprehensive analysis has shown interesting features such as the ones presented below:

**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.

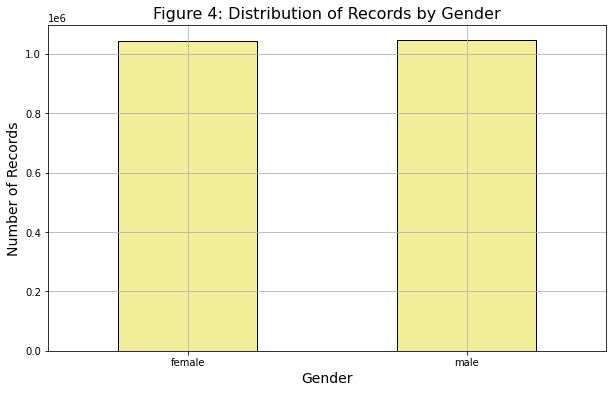
**Figure 2:** The majority of records cluster within the range of 325-370 euros on average, suggesting a prevalent income bracket among the sampled population. Interestingly, there is a secondary peak observed around 500 euros, indicating another significant concentration of wealth. It's worth noting 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.



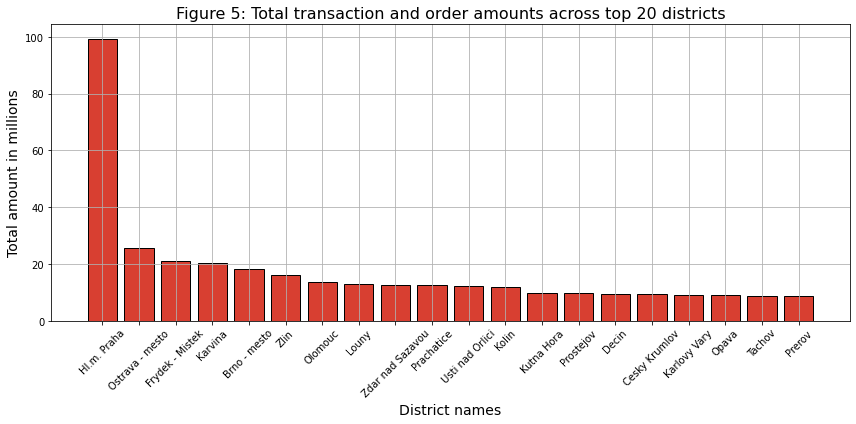
**Figure 3:** The graph illustrates a clear trend in transactional preferences, with debit cards emerging as the most utilized payment method, followed by credit transactions and cash withdrawals. This observation underscores the significance of digital payment methods in modern financial transactions, 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.

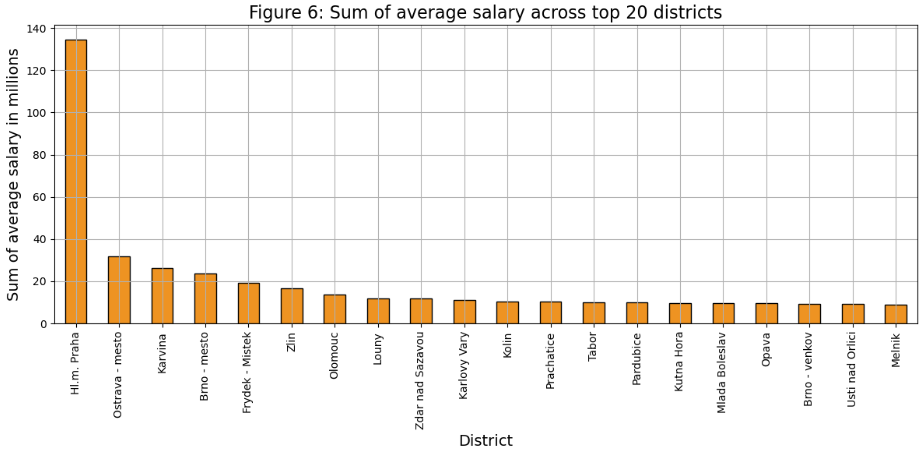


**Figure 4:** The distribution of gender within the records demonstrates a balanced representation between male and female individuals. The even distribution suggests a fair and equitable representation of both genders, which is essential for ensuring unbiased analyses and decision-making processes.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

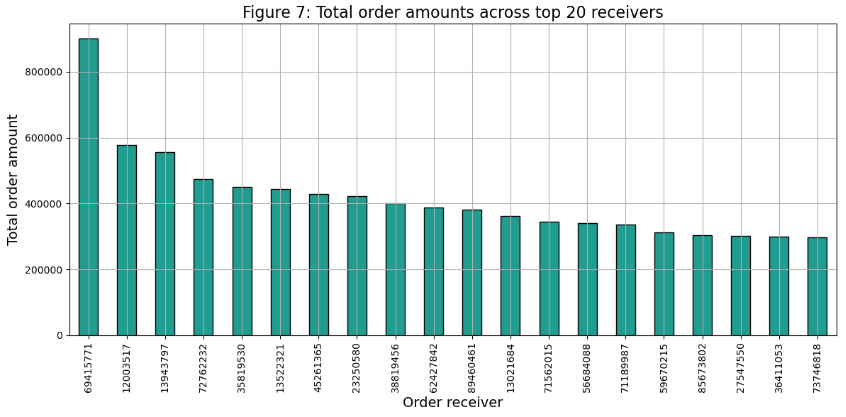


**Figures 5 & 6:** Both graphical analysis reveals notable patterns in transaction activity across different districts and standpoints. Figure 5 focuses on the total transaction and order amounts whereas Figure 6 focuses on the average salaries which shows that they are directly connected. The capital city, Prague, with the highest average salary is demonstrating a significantly high transaction volume reaching 100 million units. Following behind is Ostrava, Mesto, with transaction activity hovering over the 20 million marks, indicating a substantial level of economic activity in the region.

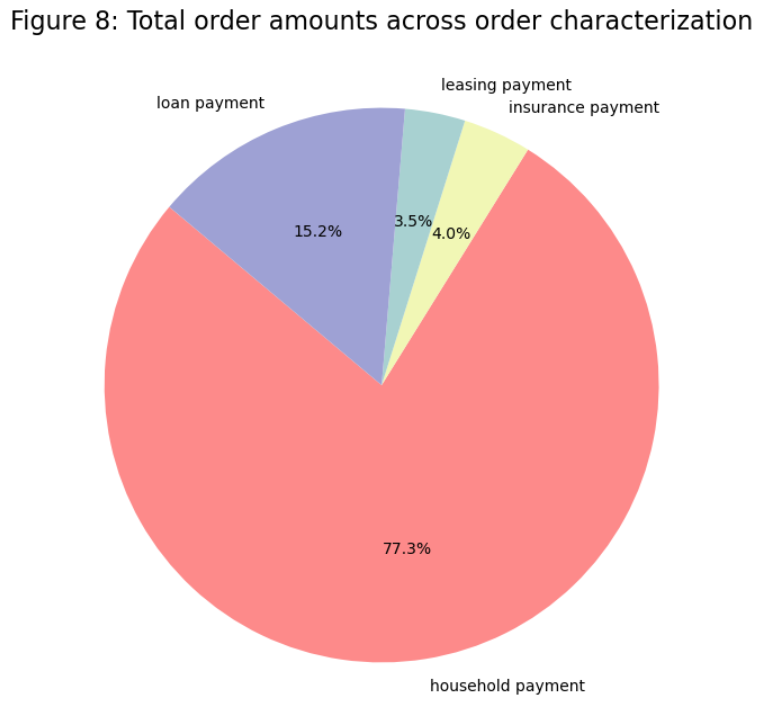




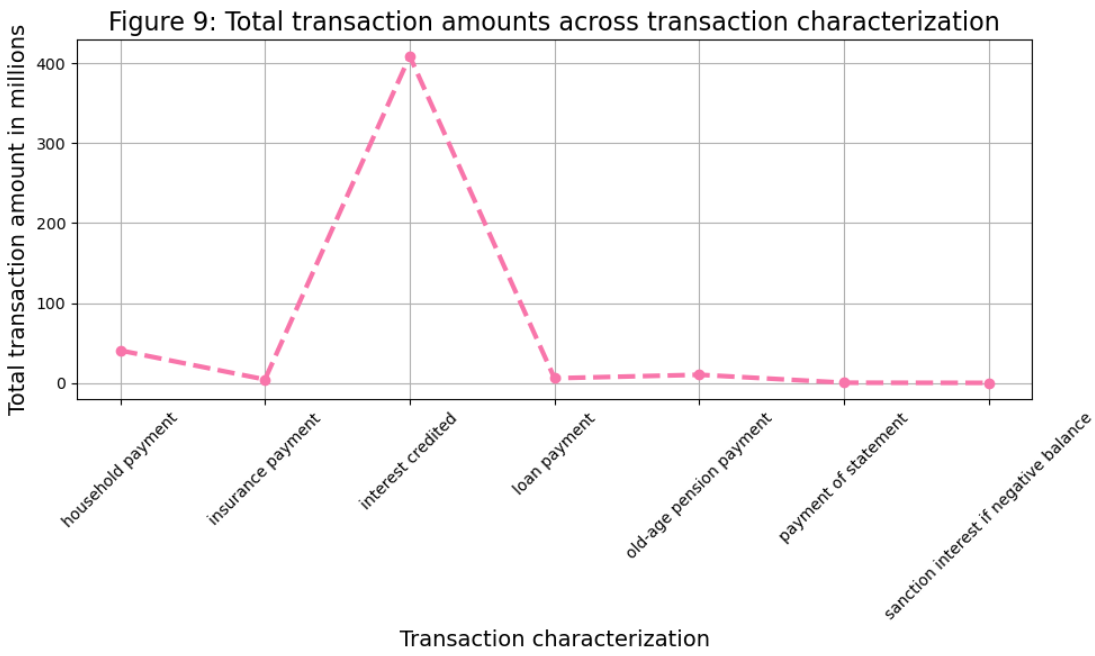
**Figure 7:** Upon examining the graph depicting the total order amount across the top 20 receivers, a striking observation emerges. It becomes evident that a particular receiver stands out significantly, commanding over 800k orders, which contrasts with the second and third place receivers, each with approximately 600k orders. The remaining receivers in the top 20 exhibit a gradual decline in order volumes, ranging from 500k to 300k orders. This discrepancy emphasizes the dominance of the top-ranking receiver and highlights the considerable gap between their order volumes.



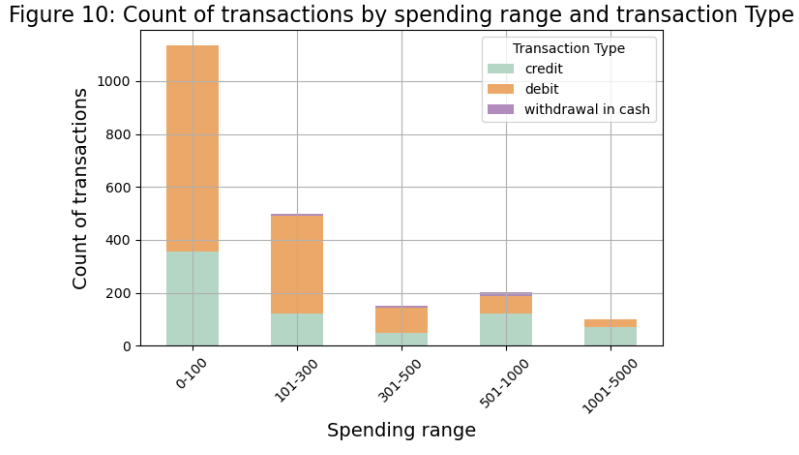
**Figure 8:** Household payments constitute the predominant share, comprising 77.3% of the total order amounts. Following behind, loan payments account for 15.2% of the total, indicating a significant but comparatively lesser portion. Meanwhile, leasing payments and insurance payments represent around 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 analysed dataset.



**Figure 9:** 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.

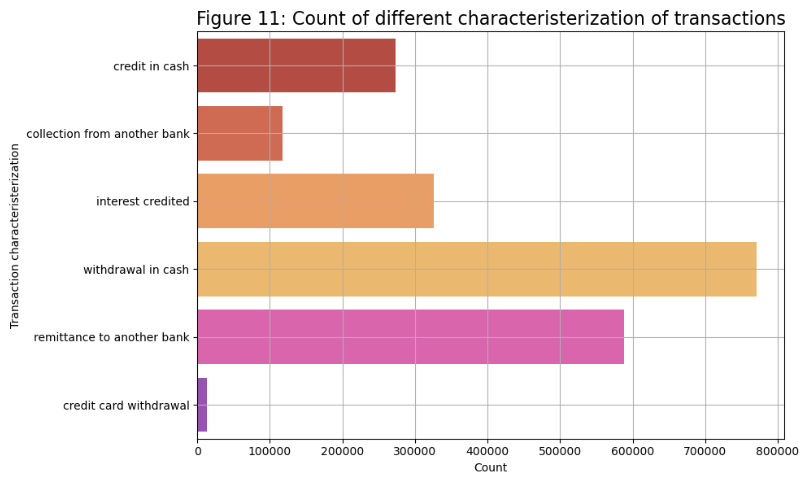


**Figure 10:** The data illustrates that credit card transactions predominantly stand within the lower spending range, typically not exceeding 750 transactions. In contrast, transactions made with debit cards extend beyond this threshold, reaching up to 2.000 transactions. Additionally, a small proportion of transactions involve cash withdrawals exceeding 2.000. This observation underscores the differing spending behaviours associated with each transaction type, with credit card transactions skewed towards lower values, while debit card transactions encompass a broader range of expenditure.

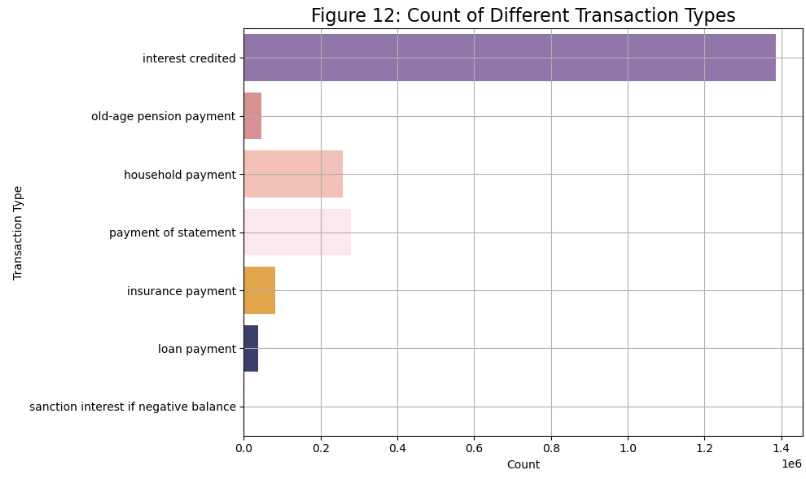


Now, we are going to explore the Categorical variables:

**Figure 11:**  This figure shows that cash withdrawals emerge as the most prevalent transaction characteristic, comprising nearly 800,000 records. Following closely behind is remittance to another bank, with approximately 600,000 instances, while interest credited transactions surpass 300,000. Moreover, credit in cash transactions and collections from another bank are each recorded nearly 300,000 and 100,000 times, respectively.

****

**Figure 12:**  In this graph we can notice that interest credited emerges as the most prevalent transaction type, boasting the highest count. Following this, old-age pension payment, household payment, payment of statement, insurance payment, and loan payment are also observed at a considerable distance from interest credited.



# Data Preparation

## Data Cleaning & Integration

In our project workflow, we made a choice to exclude the card and loan datasets from our analysis. These datasets were considered irrelevant as they did not provide any pertinent information necessary for detecting fraudulent transactions. 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, dropping from 43 columns altogether to 15.



*Screenshot 1: columns being dropped and renamed*

## Data Preparation & Construction

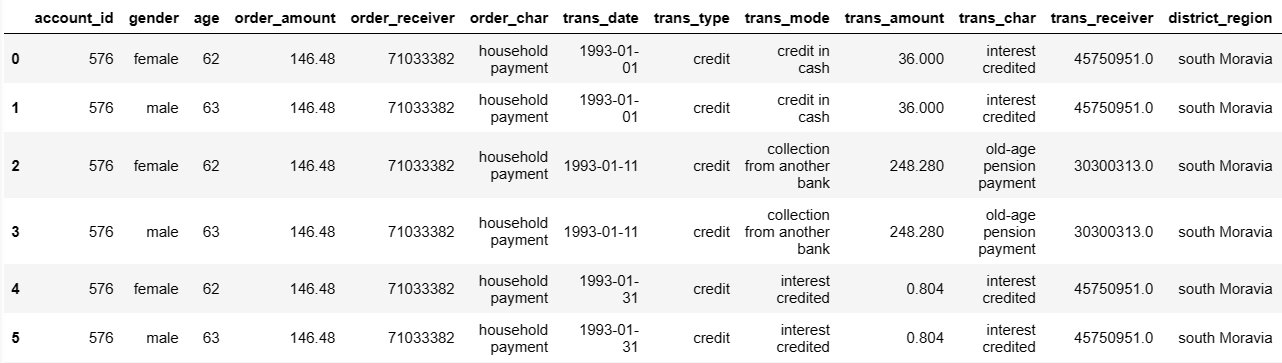
Before merging the datasets, we took a series of preparatory steps to ensure data integrity and coherence. Firstly, we renamed all the features to facilitate readability and manipulation. We also added to the dataset additional information from the client's date of birth by creating two new columns: gender and age. Additionally, to facilitate interpretation, transaction operation and characterization values were translated from Czech to English, while currency-related columns underwent conversion from Czech Koruna to Euros to maintain uniformity and better understanding *(Screenshot below)*. To address missing data, we opted to fill numerical values with the median and categorical columns with the mode, as well as information sourced from other datasets.



*Screenshot 2: date formatting, birthday calculation*

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.

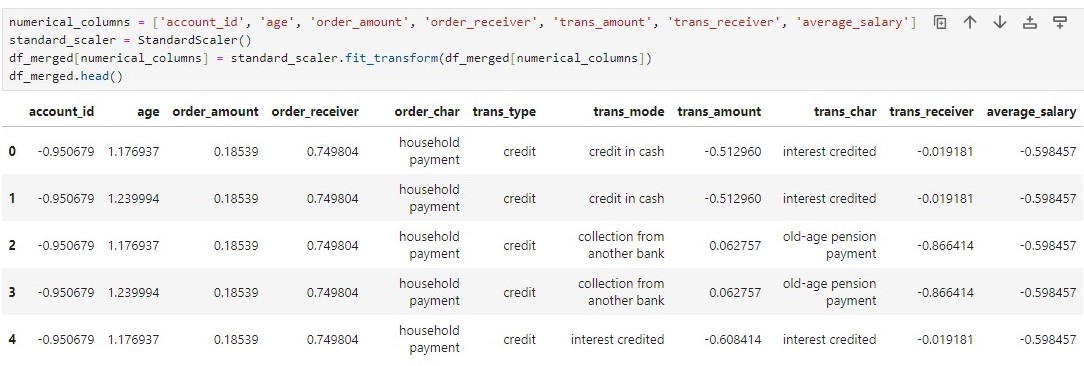
By meticulously executing these preparatory steps, we ensured that our datasets were appropriately curated and structured, laying a solid foundation for subsequent analysis and model application.



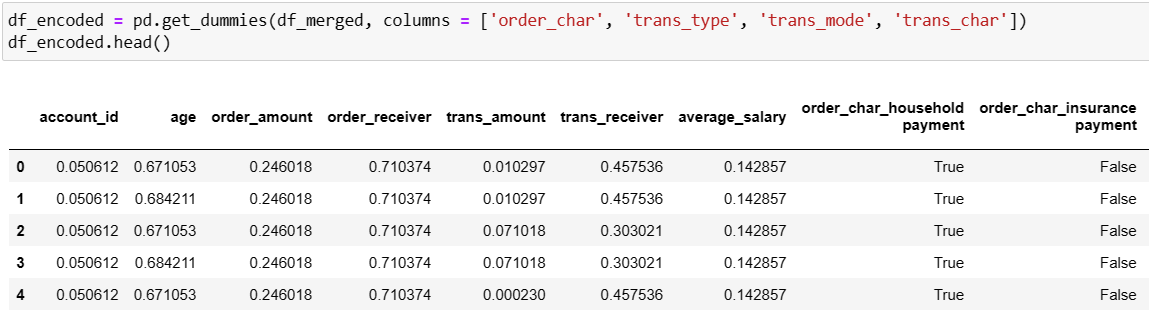
*Screenshot 3: dataset post cleaning, renaming, currency conversion*

**Scaling and Encoding**

I've employed the **StandardScaler[[5]](#_kka0gg2s7aam)** to normalize all **numerical** attributes, ensuring they share a common scale with a mean of 0 and a standard deviation of 1. This process involves subtracting the mean value from each attribute and dividing it by the standard deviation. Standard Scaler alters the data to follow a Gaussian distribution with a mean of 0 and a standard deviation of 1. This practice aids in avoiding bias during dataset training and enhances algorithm optimization.

*Screenshot 4: StandardScaler being applied to merged dataframe*

**One-Hot Encoding**[**[6]**](#_kka0gg2s7aam)is a method to transform **categorical** columns into binary sub-columns containing numeric data. This process involves encoding each unique category within a categorical column into its own binary sub-column. For instance, if a transaction's type is identified as "credit," the corresponding binary sub-column, such as "trans\_type\_credit," will be assigned a value of 1 (true). Similarly, if the transaction type falls under a different category, the "trans\_type\_credit" column will hold a value of 0 (false). This encoding technique is applied consistently across all categorical features, ensuring that only one of the binary sub-columns per feature contains a value of 1, while the rest hold values of 0.

*Screenshot 5: Data converted using One-Hot Encoding*

# Modelling

The selection of base models and the final estimator for stacking relies on a variety of factors, including the characteristics of the dataset, the complexity of the problem at hand, and the individual performance of each model. Let's assess the appropriateness of the chosen models for fraud detection:

## Base Models: Random Forest & Decision Tree[[7]](#_kka0gg2s7aam)

Random Forest (RF) and Decision Tree (DT) are frequently utilized classifiers within ensemble methodologies like stacking. RF is recognized for its resilience against overfitting and its capability to capture intricate relationships within the data. DT, on the other hand, is straightforward and easily interpretable, rendering it suitable for capturing straightforward decision rules. Both models possess the capacity to capture different facets of the data, potentially complementing each other within a stacked ensemble framework.

## Final Estimator: Logistic Regression[[8]](#_kka0gg2s7aam)

Logistic Regression (LR) serves as a linear classifier commonly employed as the final estimator in stacking due to its simplicity and interpretability. LR is well-suited for binary classification tasks such as fraud detection and can furnish insights into the significance of various features. By setting parameters such as max\_iter to a high value like 1000, LR ensures ample iterations to converge and determine optimal coefficients. In summary, the chosen base models (RF and DT) and final estimator (LR) appear well-suited for fraud detection:

1. RF and DT possess complementary strengths, adept at capturing both complex and simple data patterns.
2. LR provides interpretability and can quickly combine the predictions from the base models within the stacked ensemble.

## 

## Ensemble Learning Techniques

## 

Bagging, Stacking and Boosting[[9]](#_kka0gg2s7aam) are ensemble learning techniques used to improve machine learning model performance. Bagging involves training multiple models independently on different subsets of the data and then combining their predictions, typically through averaging or voting. For our project, we’re using the Decision Tree Classifier for bagging. [[10]](#_kka0gg2s7aam)

In contrast, stacking trains multiple models sequentially on the entire dataset and then uses a meta-model to combine their predictions. Stacking encourages model diversity and can utilize various algorithms as base models and meta-models. We have used both Random Forest Classifier and Decision Tree Classifier, with Logistic Regression serving as the final estimator. Both bagging and stacking improve model generalization and robustness by leveraging multiple models to capture different aspects of the data, reduce overfitting.

AdaBoost, short for Adaptive Boosting, is a machine learning technique that combines multiple weak learners to create a strong classifier. It operates iteratively, adjusting the weights of misclassified data points in order to prioritize them in subsequent training rounds. Each weak learner focuses on the instances that were difficult to classify by previous learners, gradually improving the overall accuracy. By aggregating the predictions of these weak learners, AdaBoost constructs a model capable of handling complex classification tasks.

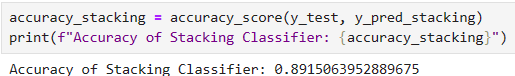
Because AdaBoost showed better results – presenting 89,31% accuracy – we have decided to utilize it to create the fraud column in our analysis. We also decided to focus on this model because it showed better efficiency timewise. By taking advantage of the diversity of multiple models and their ability to capture different nuances of the data, AdaBoost enabled us to achieve superior results compared to the other ensemble techniques.

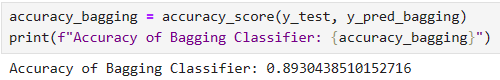
## 

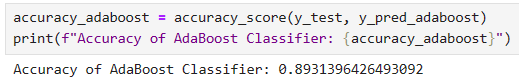
## 

# Evaluation

In the evaluation section of our report, we present the performance of all the algorithms used in our analysis. After applying the ensemble learning techniques, the stacking classifier achieved an accuracy of 0.8915, demonstrating its effectiveness in identifying fraudulent activities. Similarly, the bagging classifier showed strong performance with an accuracy of 0.8930, while the AdaBoost classifier held an accuracy of 0.8931. These results highlight the reliability of ensemble learning techniques used in our fraud detection algorithm, mixing multiple models to enhance predictive accuracy as shown below:

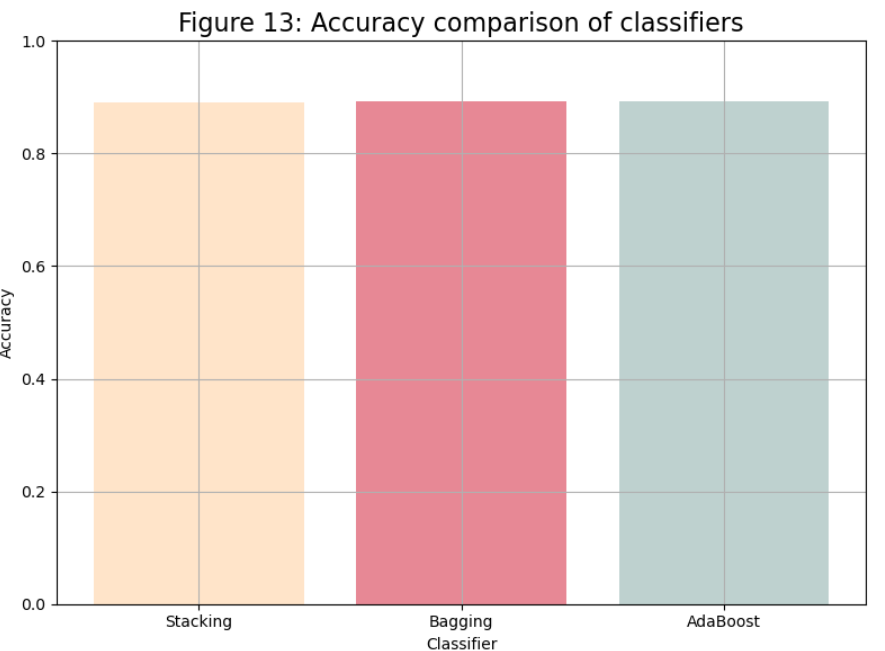


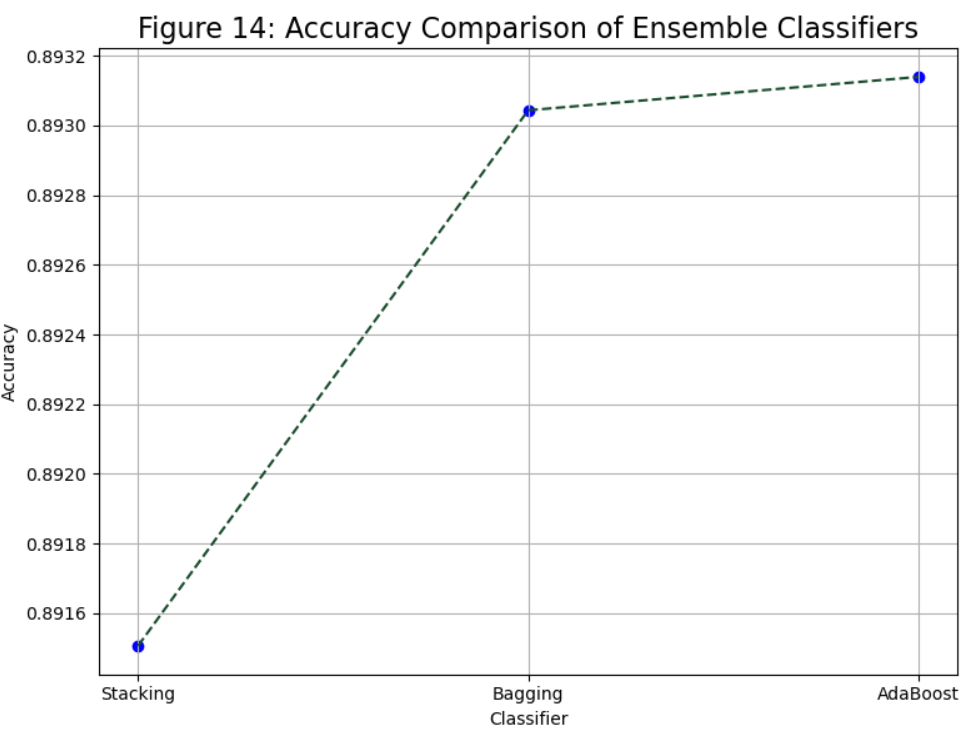




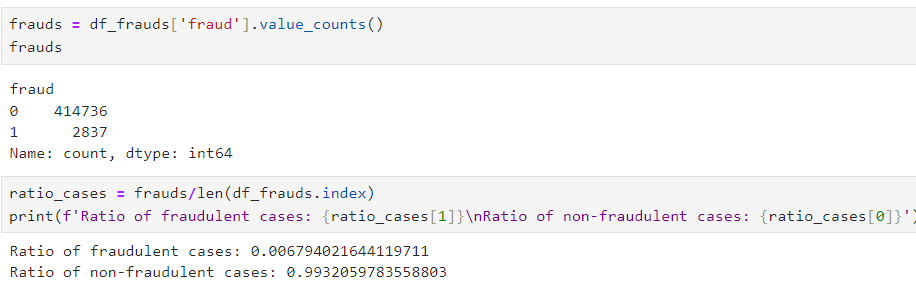
*Screenshot 6, 7 and 8: Accuracy of Stacking (89,15%), Bagging (89,30%), AdaBoost (89,31%) Classifiers*

Similarly, to the screenshots shown above, we can also note their high performance from these graphs below. As we can see, the difference between them is very small:





In concluding our evaluation, it's essential to examine the final figures from all instances of fraud within the dataset, totalling 2,837 cases. This accounts for approximately 0.07% of the dataset, with non-fraudulent cases constituting the remaining 99.3%. Given that the model achieves an accuracy of 89.31%, we can confidently reinforce the reliability of these figures. Such insights could significantly benefit potential customers of CH Banking Solutions, providing better understanding and identifying of customer data, as well as discerning when to be vigilant in cases of fraudulent activity.

*Screenshot 9: Frauds detected after model is applied.*

# Deployment

When it comes to finalizing the project, several considerations arise. Initially, our intention was to utilize AWS Cloud Services (Such as S3 Bucket and Lambda), as well as JavaScript, for online service provision. However, as time progressed, we realized that achieving this within our timeframe was unfeasible. Consequently, we changed our focus towards developing a highly efficient model.

The deployment was done by providing the final resources: df\_merged.csv, df\_encoded.csv, and df\_frauds.csv (final file) at the end of the project, along with all the tree models that were used. These can be imported and loaded for further analysis and implementation.

This project could be further refined and developed by us. Additionally, we aimed to create a User Interface (UI) allowing users to upload datasets and access a dashboard displaying comprehensive statistics on their financial transactions, including fraud analysis, customer demographics, spending insights, etc. While the UI remains a work in progress, we are satisfied with the outcomes of our efforts. Despite the challenges, we are eager to continue refining and potentially deploying this application, even if solely for demonstration purposes.

In compliance with ethical and legal standards, it is important to address the transparency and consent requirements regarding the system's capability to track transactions. Prior to utilizing the services provided by CH Banking Solutions, customers must explicitly agree to the terms and conditions, acknowledging their awareness of the system's tracking capabilities.

To ensure compliance with regulatory frameworks such as the General Data Protection Regulation (GDPR), rigorous measures would be implemented to ensure user data privacy and security. Any data collected or processed by the system will be done so in accordance with GDPR guidelines, with a commitment to maintaining confidentiality and integrity. In the previous report with the brainstorming of this work we highlighted some points that should be adhered:

|  |  |  |
| --- | --- | --- |
| **Legal & Ethical Concerns Regarding CH Banking Solutions Application** | | |
| **Compliance Concerns** | **Legal Consideration** | **Ethical Consideration** |
| **Data Usage & Compliance** | Ensure alignment with licensing terms; communicate it's for educational purposes. | Be transparent about the intent for prototype and educational use. |
| **Data Privacy**  **& Anonymization** | Comply with data privacy laws, anonymize personally identifiable information. | Prioritize robust anonymization techniques and avoid using sensitive information. |
| **AWS Services Compliance** | Understand and comply with terms of service and usage policies of AWS services. | Use AWS services responsibly, considering environmental impact and security. |
| **Open Source Libraries**  **& Licensing** | Verify licensing terms for open-source tools. | Acknowledge and credit developers of incorporated open-source tools. |
| **User Consent**  **& Transparency** | Obtain explicit consent for user data usage. | Transparently communicate data usage and purpose. |
| **Prototype Limitations**  **& Disclaimer** | Clearly state prototype limitations and purpose. | Be transparent about capabilities and potential inaccuracies. |

*Legal & Ethical Concerns Regarding CH Banking Solutions Application (taken from CA1)*

Furthermore, the deployment of the system on AWS platforms will follow the platform's security protocols and standards. AWS offers robust security features and compliance certifications, providing assurance regarding data protection and regulatory compliance. By incorporating transparency, informed consent, and adherence to legal and regulatory requirements, CH Banking Solutions could seek to maintain ethical standards while delivering valuable services to its customers.

# Conclusion

The fraud detection project has been a challenging journey marked by numerous obstacles and opportunities for personal and professional growth. From data cleaning and preparation to navigating the wide and deep availability of models to select from and evaluate, each phase of the project presented its own unique challenges that required dedication and resilience from us.

One of the most difficult challenges we encountered was the huge scale and complexity of the dataset. With 8 distinct datasets to merge and pre-process, the data cleaning and preparation process demanded continued effort, with adjustments and refinements being made until the completion of the project. Similarly, the process of model selection posed its own set of challenges. As we thought we had found out the perfect algorithm, we then faced their limitations in accurately detecting fraudulent activities and overfitting. This led to a deeper exploration of alternative methodologies and the refinement of our strategies to better align with our objectives.

Furthermore, our efforts are evidenced by the remarkable results achieved through the implementation of various ensemble learning techniques. All the ensemble models enabled us (CH Banking Solutions) to achieve an impressive accuracy rate of at least 89,30% for each of them. The utilization of those methods, including bagging, stacking, and boosting, played an important role in enhancing the predictive capabilities of our models. These techniques not only facilitated the fusion of diverse predictive models but also fixed the flaws of individual algorithms, resulting in a robust and highly accurate fraud detection system that potential customers could benefit from.

# 

# Peer Reflection

## Caroline

I have learned a lot about data cleaning and data preparation during the two months of working on this project. It took extensive research, constant trial and error and many hours of debugging to achieve the final results present in this project. My main tasks involved cleaning and preparing the dataset, as well as researching and applying the machine learning models to the training and testing data. The work wasn’t linear, as I always had to swap between those three tasks to achieve the current outcome. Many decisions that we made at the beginning of the project were altered along the way, and dozens of models were applied to the data until we finally found what works.

The dataset itself showed to be one of the most difficult challenges. Apart from being very hard to find data for this project given that we are working with very sensitive information, the vast financial data split across 8 distinct datasets needed to go through various processes. A merging process was needed to consolidate them into a cohesive, single dataset. The cleaning, preparation, and encoding stages accounted for a big part of the project's code, persistently requiring adjustments and refinements up until completion.

Finding the right algorithm also took a lot of time, research, and resilience. That said, the most rewarding part was making it work. I can say for myself and for Heber that we learned a lot during the process, both through research and from errors. It was very energy and time-consuming, but seeing it work with very good results in the end is very satisfying.

## Heber

Working on this fraud detection project has provided me with a valuable opportunity to deeper understand the intricacies of machine learning and its methodologies as it brought a series of challenges, moments of being stuck and requiring extensive research, making us rethink our approaches. Another significant challenge emerged during the model selection phase. Despite initial enthusiasm for a particular algorithm, we soon realized its limitations in accuracy to detect fraudulent activities. This experience enriched my understanding of the field and enhanced my skill set, particularly in collaboration. Throughout the project, my focus was:

I assumed the responsibility of meticulously documenting various aspects of our work that served as our deliverables apart from the code itself, including writing this report, designing the academic poster, and creating the PowerPoint presentation. While I carried out these tasks, Caroline offered her insights through revisions, ensuring the clarity and coherence of our documentation but focusing on the data preparation and modelling.

I also conducted extensive exploratory data analysis (EDA), which laid the foundation for our subsequent modelling efforts. By meticulously examining the dataset, identifying patterns, and understanding its structure, I contributed to the formulation of effective strategies for fraud detection.

This project not only increased my technical skills but also reinforced the importance of effective communication and collaboration in achieving our goals. The challenges faced served as valuable learning experiences, highlighting the complexities in programming and the importance of constant research, adaptability, and collaborative problem-solving.

# 

# Referencing

### 

1. CRISP-DM Framework. (2024). Moodle. [Online]. Available at: https://moodle.cct.ie/mod/folder/view.php?id=33135 (Accessed: 13 May 2024).
2. 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).
3. PKDD’99 Discovery Challenge (no date) Financial Data description: Available at: <https://sorry.vse.cz/~berka/challenge/pkdd1999/berka.htm> (Accessed: 19 March 2024).
4. 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).
5. Geek Python. (2023). How to Use sklearn StandardScaler. Available at: <https://geekpython.in/how-to-use-sklearn-standardscaler> (Accessed: 14 May 2024)
6. 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).
7. Analytics Vidhya. (2020). Decision Tree vs Random Forest Algorithm. Analytics Vidhya. Retrieved May 14, 2024, from [https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/#:~:text=A%20decision%20tree%20is%20more,%2C%20healthcare%2C%20and%20deep%20learning.](https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/#:~:text=A%20decision%20tree%20is%20more,%2C%20healthcare%2C%20and%20deep%20learning)
8. Murali, V. (2023). Ensemble Learning: Combining Models for Better Predictions. Medium. Retrieved May 14, 2024, [from https://vivek-murali.medium.com/ensemble-learning-combining-models-for-better-predictions-53a4c80cb18a](https://vivek-murali.medium.com/ensemble-learning-combining-models-for-better-predictions-53a4c80cb18a)
9. Medium. (2023). Ensemble Models: How to Make Better Predictions by Combining Multiple Models with Python Codes. Medium. Retrieved May 14, 2024, from <https://aryanbajaj13.medium.com/ensemble-models-how-to-make-better-predictions-by-combining-multiple-models-with-python-codes-6ac54403414e>
10. Maheshhkanagavell. (2023). Exploring Decision Trees, Random Forest, Logistic Regression, KNN, Linear Regression, SVM, RNN, LSTM, and LightGBM for Effective Data Analysis. Medium. Available at: <https://medium.com/@maheshhkanagavell/exploring-decision-trees-random-forest-logistic-regression-knn-linear-regression-svm-rnn-bf52fa94a066> (Accessed: 15 May 2024).
11. 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).
12. OpenAI. (n.d.). ChatGPT. Retrieved May 14, 2024, from <https://openai.com/chatgpt> *(uses include: brainstorming and refining ideas; fine tuning research; drafting an outline to organise thoughts and checking grammar).*
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. 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).
15. 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).
16. Argaadya. (n.d.). CRISPR/Cas9 Genome Editing: Principle, Applications, and Challenges. RPubs. Retrieved May 14, 2024, from https://rpubs.com/Argaadya/crispr\_dm
17. 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).
18. 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).
19. 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).
20. Sahagian, G. (2020) What is Random State 42?, Medium. Medium. Available at: <https://grsahagian.medium.com/what-is-random-state-42-d803402ee76b> (Accessed: 15 April 2024).
21. Top difference between training data and testing data (2023) Testsigma Blog. Available at: <https://testsigma.com/blog/difference-between-training-data-and-testing-data/> (Accessed: 18 April 2024).
22. Bajaj, A. (2023) Ensemble Models: How to Make Better Predictions by Combining Multiple Models with Python Codes..., Medium. Medium. Available at: <https://aryanbajaj13.medium.com/ensemble-models-how-to-make-better-predictions-by-combining-multiple-models-with-python-codes-6ac54403414e> (Accessed: 11 May 2024).
23. GeeksforGeeks (2022) Online Payment Fraud Detection using Machine Learning in Python, GeeksforGeeks. GeeksforGeeks. Available at: <https://www.geeksforgeeks.org/online-payment-fraud-detection-using-machine-learning-in-python/> (Accessed: 11 May 2024).
24. Ensemble methods in Python: stacking (no date) Educative. Available at: <https://www.educative.io/answers/ensemble-methods-in-python-stacking> (Accessed: 13 May 2024).
25. AI, T. (2022) Ensemble Learning - AdaBoost with Python, Medium. Medium. Available at: [https://tirendazacademy.medium.com/ensemble-learning-adaboost-with-python-8332778fbb61#:~:text=A%20guide%20on%20how%20to,)%20algorithm%20with%20scikit%2Dlearn.&text=AdaBoost%20is%20a%20boosting%20method,starting%20point%20for%20understanding%20boosting.](https://tirendazacademy.medium.com/ensemble-learning-adaboost-with-python-8332778fbb61#:~:text=A%20guide%20on%20how%20to) (Accessed: 13 May 2024).
26. AI, T. (2022) Ensemble Learning - AdaBoost with Python, Medium. Medium. Available at: [https://tirendazacademy.medium.com/ensemble-learning-adaboost-with-python-8332778fbb61#:~:text=A%20guide%20on%20how%20to,)%20algorithm%20with%20scikit%2Dlearn.&text=AdaBoost%20is%20a%20boosting%20method,starting%20point%20for%20understanding%20boosting.](https://tirendazacademy.medium.com/ensemble-learning-adaboost-with-python-8332778fbb61#:~:text=A%20guide%20on%20how%20to) (Accessed: 13 May 2024).