Data Mining I

Should We Loan?

Ana Sofia Teixeira

André Manada

Problem Definition

To elevate customer service standards, our bank faces a critical challenge: distinguishing between good clients—those eligible for enhanced services—and bad clients—requiring vigilant oversight to minimize potential risks. With an extensive repository of client data encompassing account details, transaction histories spanning multiple months, prior loan records, and issued credit cards, this project aims to leverage data mining techniques.

Our objective is twofold: to provide a comprehensive understanding of our clients through descriptive analysis and to address the predictive challenge of determining the success of a loan. While the descriptive aspect explores open-ended insights into client behavior, the predictive facet focuses on anticipating the outcome of loan ventures. The goal is to empower bank managers with actionable insights to identify ideal candidates for credit card services.

Data Understanding and Preparation

To understand the data, we had to analyze every table in the dataset. After this, we had to decide which data is important for our task, and which data to ignore.

We concluded that to decide whether a loan will end successfully or not, we had to have this into consideration:

- **Client:** we decided that we don't need any information about the client
- **District:** although this table has a lot of information, the only features that make sense for us are: the average salary, the number of crimes in 95 and 96 and the amount of unemployment in 95 and 96
- **Account:** in this table we discarded the feature "frequency"
- **Disp:** since only owners can make a loan, we decided that the feature "type" of this table is helpful
- **Card:** since most of the account don't have a card associated, we decided to also not use this table
- **Trans:** this table was very helpful to understand the history of each account, so we decided to use the main features (account id, date, type, amount and balance)
- **Loan:** since the goal is to determine whether a loan is going to be successful or not, all the data in this table is important

Data Understanding and Preparation

After selecting which attributes to use, we had to prepare it:

- For each account, we calculated the cash flow per month. The cash flow uses the transactions table, for every entry belonging to the same month that has a “credit” type we add it to the cash flow, and for every entry that has a “withdrawal” type, we subtract
- Calculated the age in month for each account when the loan was requested
- We also calculated the mean cash flow for each account
- Created a map with the amount of money in the account every month
- Calculated the mean rate of unemployment between 95 and 96 and the mean rate of crimes committed between the same years
- Counted the number of “good loans” for each account (loans with a status of 1) – this feature was later dropped because no account has more than 1 loan

When gathering all this information, we where able to create the final table which we used for the basis of our project.

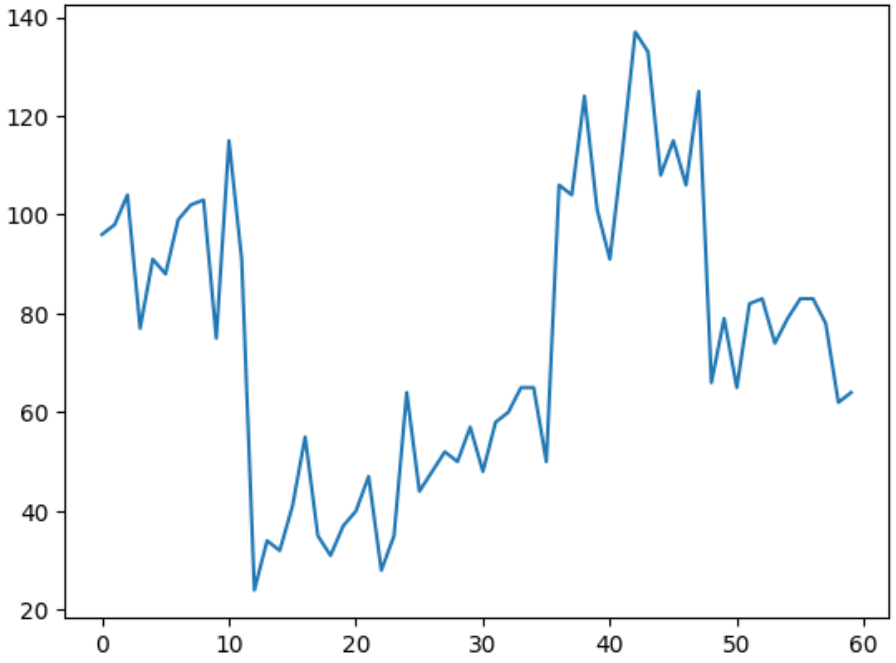
Data Understanding and Preparation

Our training table has the following data for every account that has a loan (either with a status of 1 or -1):

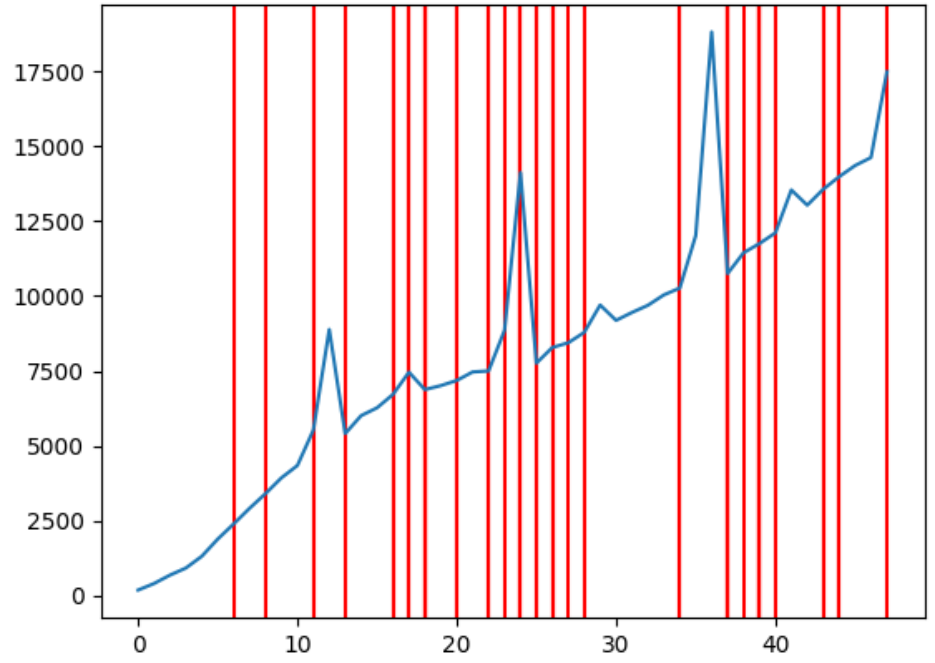
- **Loan amount** – the amount of the loan (present in the loans table)
- **Loan payment** – the monthly payments amount (present in the loans table)
- **Account cash** – the amount of cash in the account when the loan was requested
- **Account cash flow** – the mean cash flow of the account when the loan was requested
- **Account age** – the account age (in months) from its creation to the date of the requested loan
- **Average Salary** – the account's district average salary (present in the district table)
- **Crime rate** – the mean crime rate of the account's district
- **Unemployment rate** – the mean unemployment rate of the account's district
- **Status** – the status of the loan (present in the loans table)

Data Understanding and Preparation

To better understand the dataset, we chose to visualize the following data:

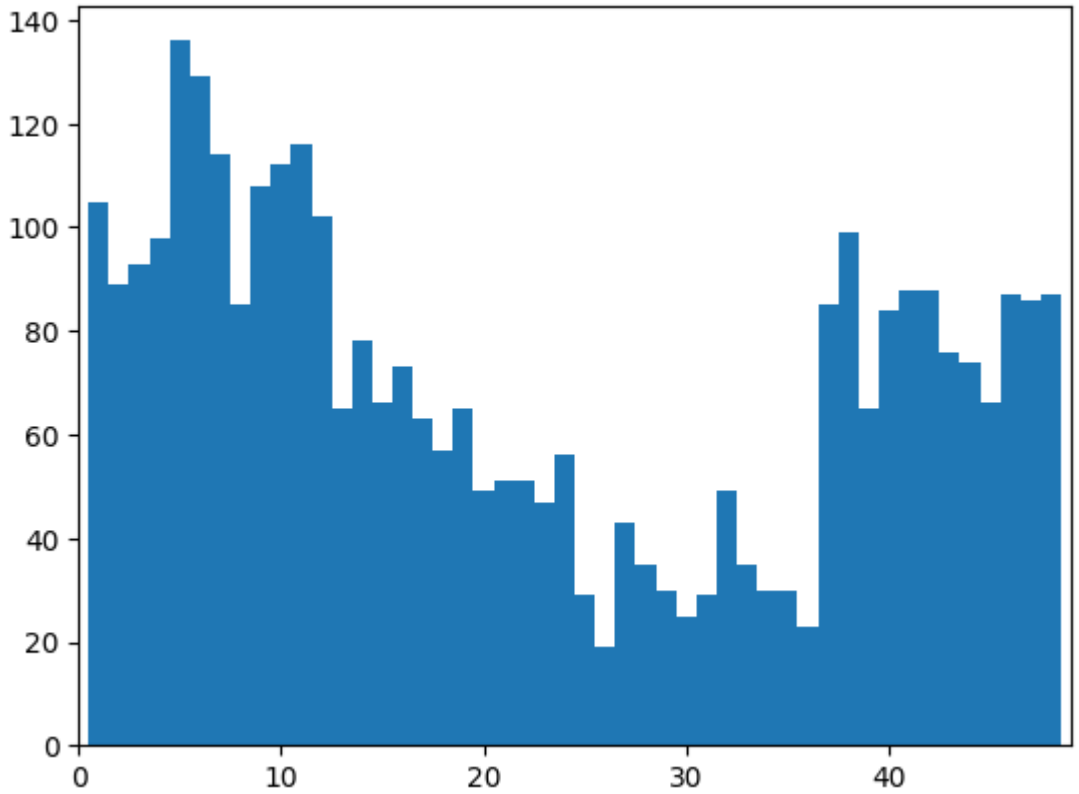


Number of accounts joining per month

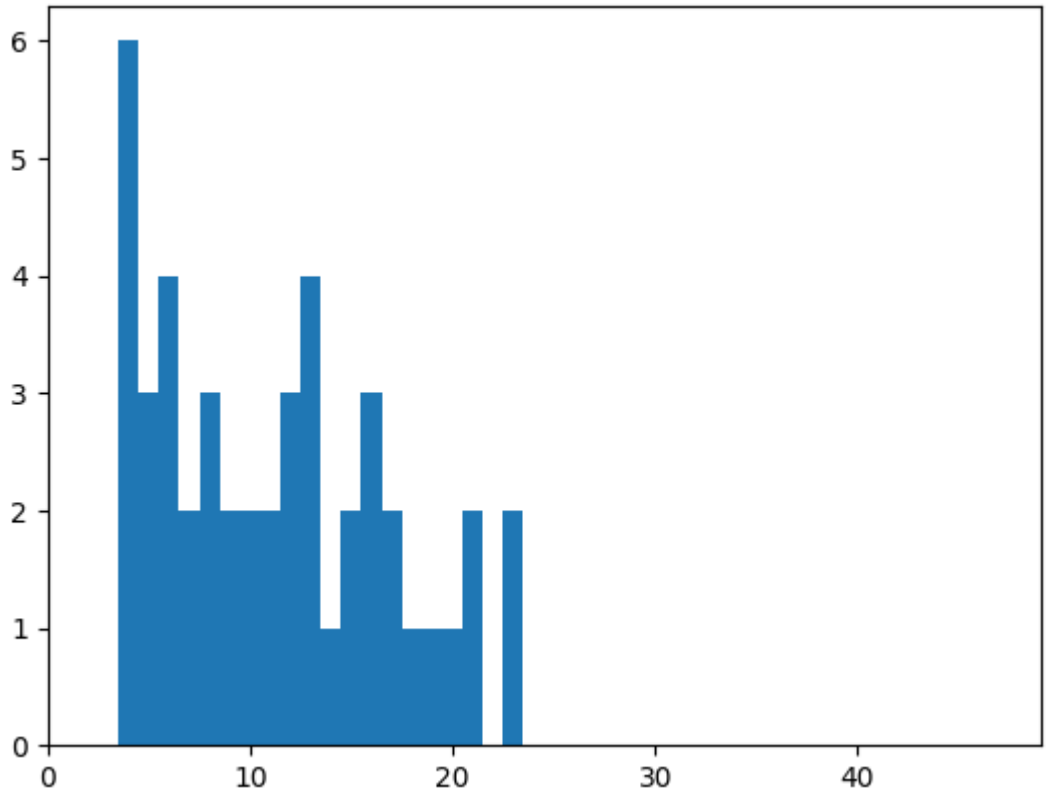


Number of transactions done per month
Red line: months that have a fraudulent loan

Data Understanding and Preparation

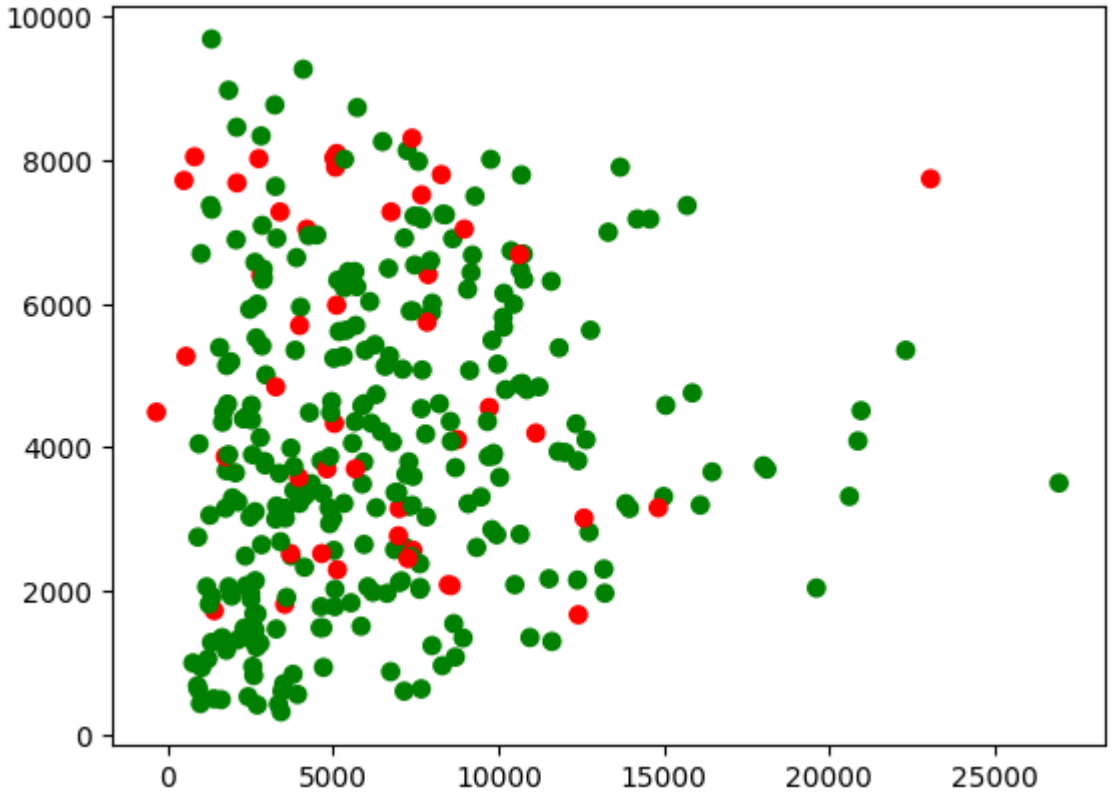


Distribution of account ages (in months)

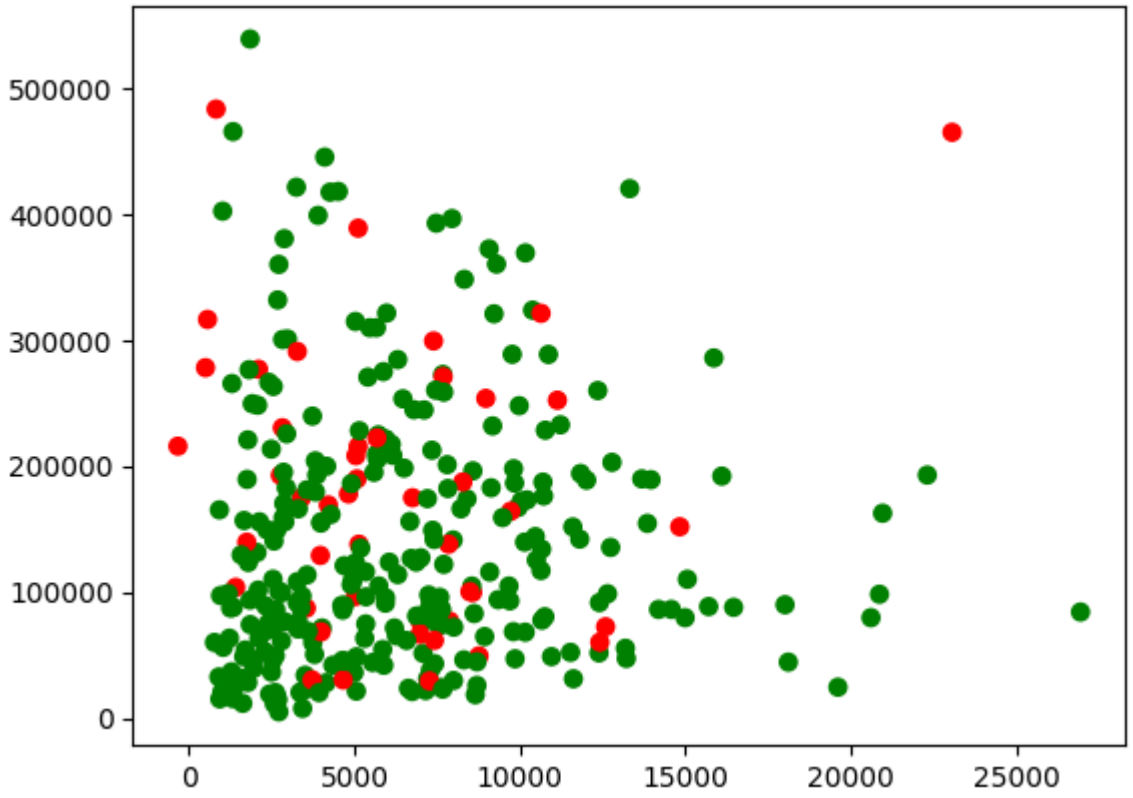


Age of accounts (months) that have fraudulent loans

Data Understanding and Preparation



Loan payment per account mean cash flow



Loan amount per account mean cash flow

Descriptive Modelling

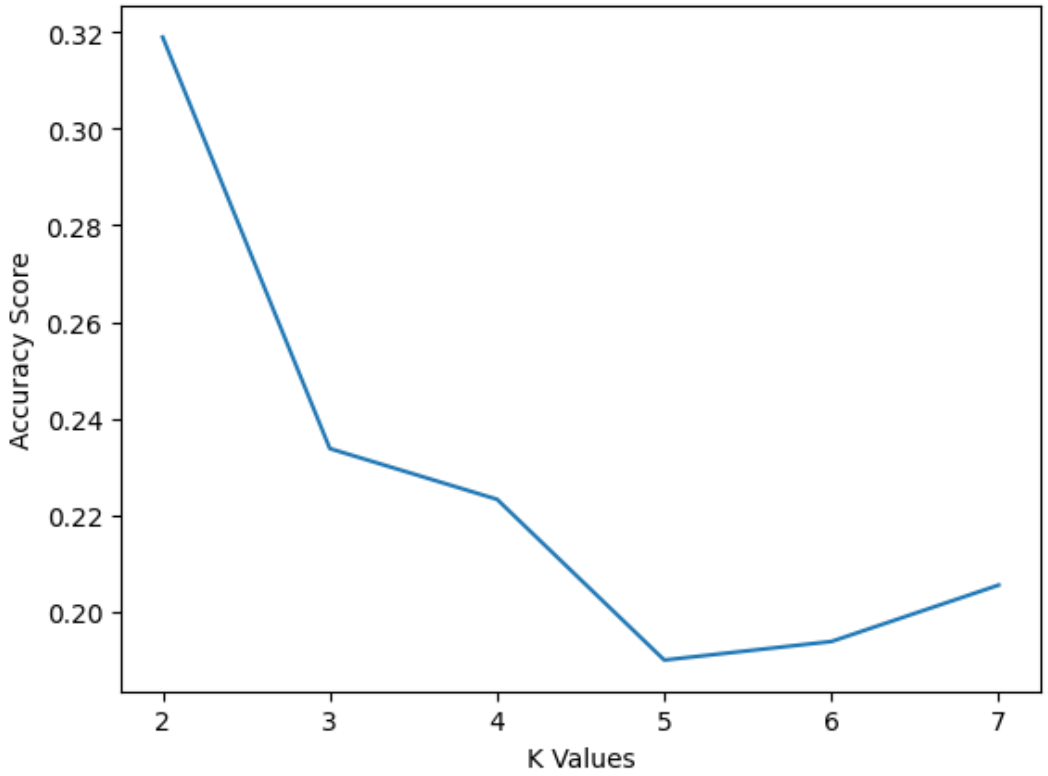
For the descriptive modelling, we chose to use the K-means algorithm.

K-Means

When given the final table without the feature "status", the K-means algorithm enables us to understand better the natural groupings or similarities among accounts based on their financial attributes, district information, and account history.

We chose to use k values between 2 and 7. The graphic shows the silhouette score for each of those k values.

Note: when the card type feature was used, the k value with the best performance jumped between 2 and 4.

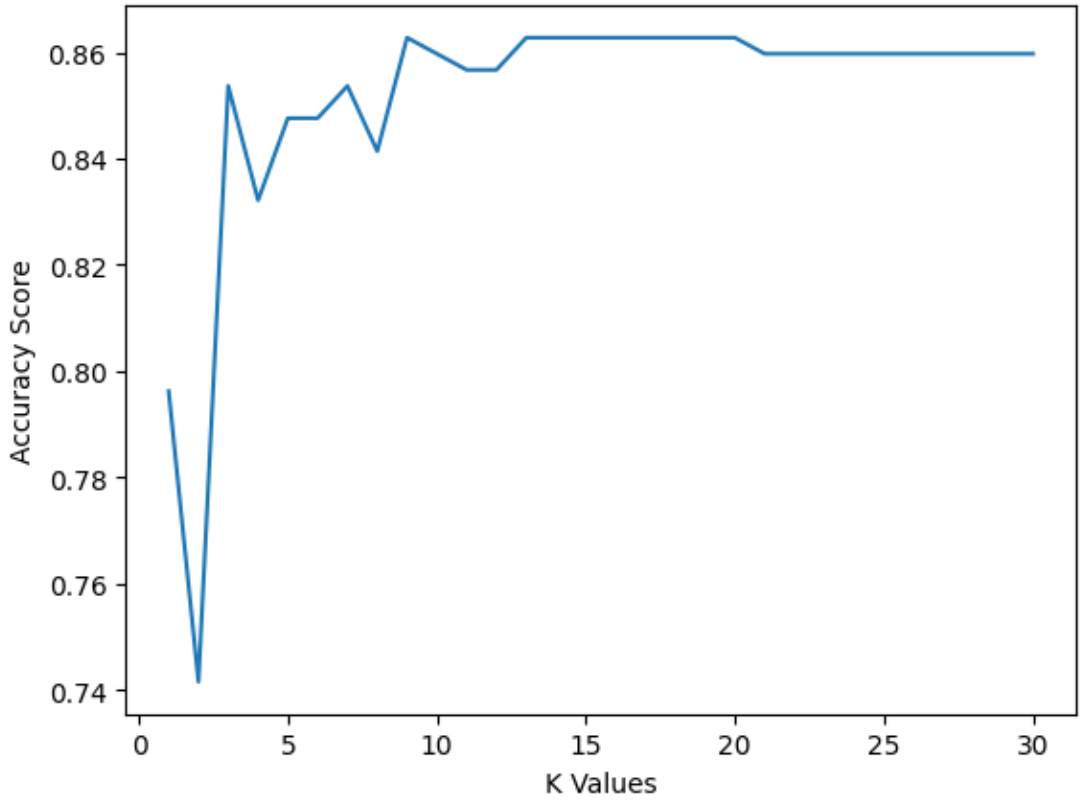


Predictive Modelling

For this part of the project, we were unsure about which algorithm to use, so we tested four: kNN, Naive Bayes, Support Vector Classification and Random Forest.

kNN

For the kNN algorithm, we chose to hyper tune the k by iterating through different values of K (1 to 30). For each iteration we initialize the NeighborsClassifier model with the number of neighbors equal to the k for that iteration and perform cross-validation with 10 folds. This function splits the data into 10 parts, uses 9 for training and 1 for testing, repeating this process 10 times.
The graphic shows the cross-validation score for each iteration of this algorithm.



Predictive Modelling

Naive Bayes
We chose to not hyper tune the Naive Bayes algorithm. The evaluation was also done with a 10-fold cross-validation method.

Support Vector Classification (SVC)
For this algorithm we used different kernels (linear, rbf, sigmoid, poly) for hyper tunning, and we got about the same score. Once again, we are using cross-validation with cv=10.

Random Forest
For this algorithm, we initialized a HistGradientBoostingClassifier model with a specified maximum number of iterations. We chose this number of maximum iterations because after hyper tunning it, we found out that 300 iterations was the option with most accuracy.

Conclusions

Following thorough data understanding, preparation, and the application of various algorithms, this project has made significant strides in addressing the challenges faced by the bank in enhancing customer service quality.

The exploration of KNN, Naive Bayes, SVC, and Random Forest algorithms for loan success prediction has yielded comprehensive perspectives. These algorithms have offered varying accuracies and predictive capabilities, presenting diversified approaches to determine successful loan outcomes. Since the accuracies obtained for each algorithm were very similar (between 0.84 and 0.86), none of the algorithms stood out.

For future work, maybe we should explore another vast number of algorithms to make our prediction even more accurate and do a more detailed analysis on transactions and districts to obtain better results.