Topic : Bank Fraud

Presented by:

GANDEMA Rafissatou
NIKIEMA Elodie Marie Laureine
OUEDRAOGO Abdoul Fataze
OUEDRAOGO Jonathan Steve

Teacher's name: Miss Otema YIRENKYI
Miss Kweyakie BLEBO

Predictive Fraud Analysis Banking

Analysis by Python

Problematic :

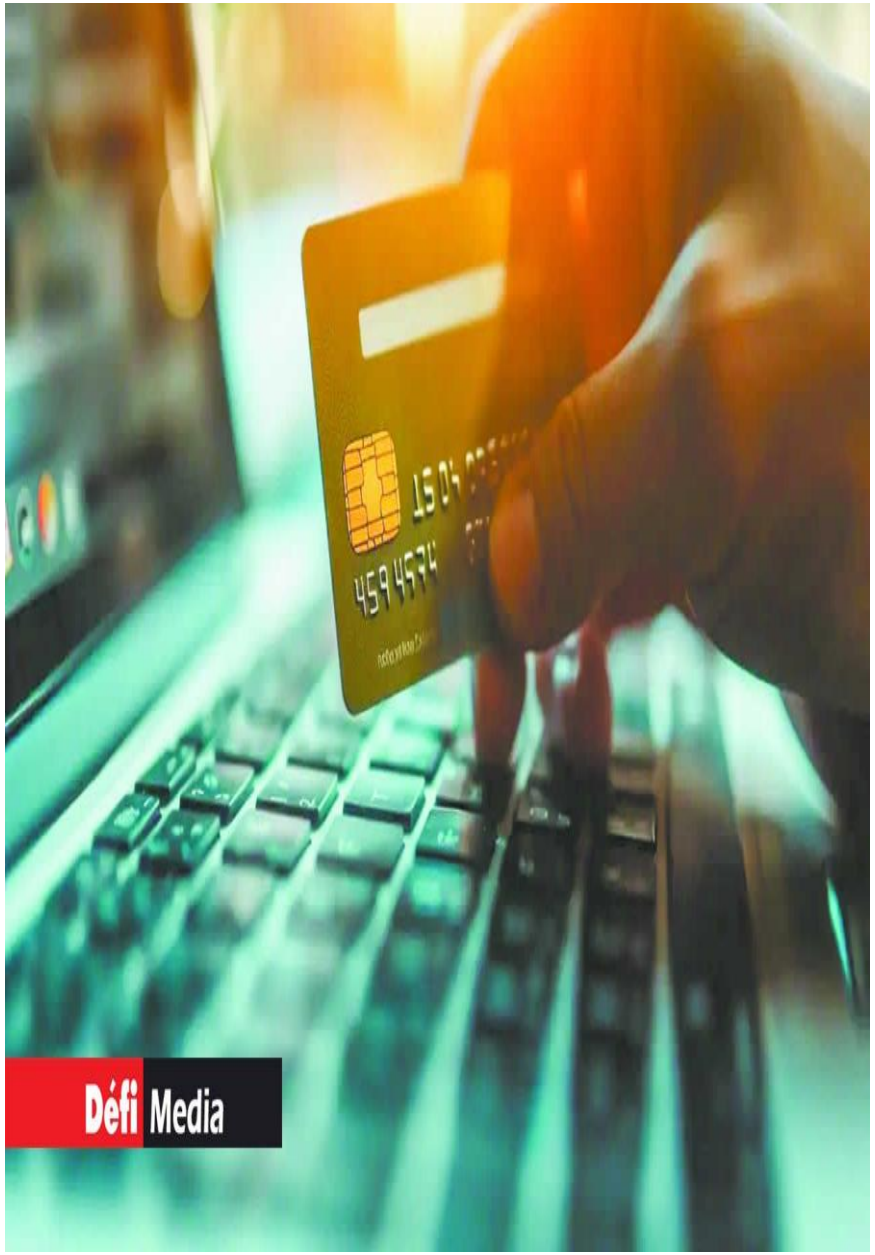
Modern Bank Fraud

Context current

Bank fraud represents a major challenge for financial institutions in the digital age. With the proliferation of online transactions, fraudulent techniques are constantly evolving.

Objective of our Analysis

Our analysis consists of identifying the discriminating elements and variables in credit applications that allow us to distinguish fraudulent transactions from legitimate ones.

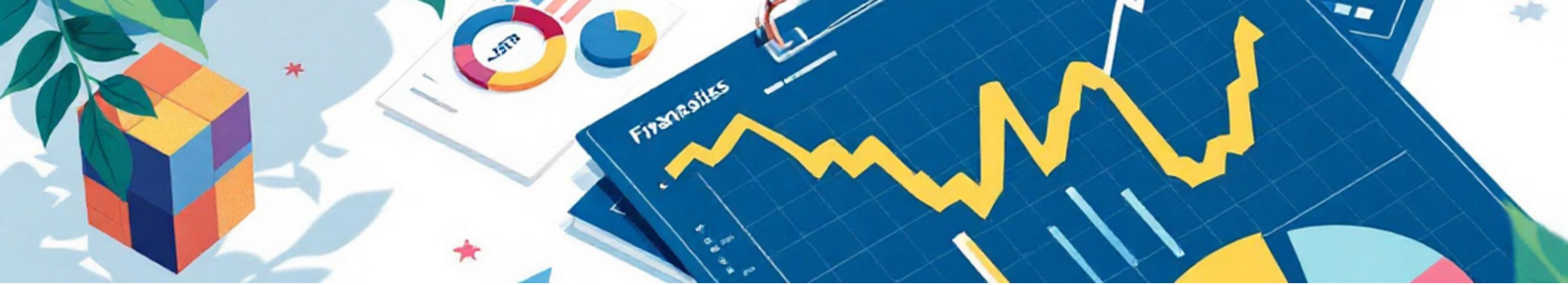


Our goal: To analyze data to identify indicators of fraud in loan applications and define simple rules for detecting bank fraud.

Critical issues

- Losses financial estimated at several billion CFA francs annually
- Need of a detection fast and accurate
- Impact
- Compliance regulatory strict

How to identify indicators of bank fraud by analyzing the variables from our data?



Dataset: Bank Account Fraud (NeurIPS 2022)

1M

Observations

32

Variables

2022

Publication

Presented at the NeurIPS conference, a leading event in machine learning

Source: https://www.kaggle.com/datasets/sgpjesus/bank-account-fraud-dataset-neurips-2022

Data Cleaning and Preparation

01

Initial inspection

Review of data types, identification of values missing .

02

Handling missing values

Management strategic depending on the nature of the variables.

03

Outlier detection

Value identification and contextual validation

04

Suppression

Column removal useless data duplicated and null

Result: Clean dataset ready for in-depth exploratory analysis

Exploratory data analysis

- In this section, we conduct an exploratory analysis of the data to better understand the structure of the dataset and the differences between fraudulent and non-fraudulent requests.

- We use descriptive statistics and groupby on the variable fraud_bool to compare the means of key variables such as the proposed credit limit and income.

- This step allows us to identify the most relevant variables for fraud detection.

Metric: Creating a new characteristic

Our composite metric relies on three key indicators, each identifying a specific risk signal in the customer profile.

high_income_flag

Income declared abnormal higher compared to the standards.

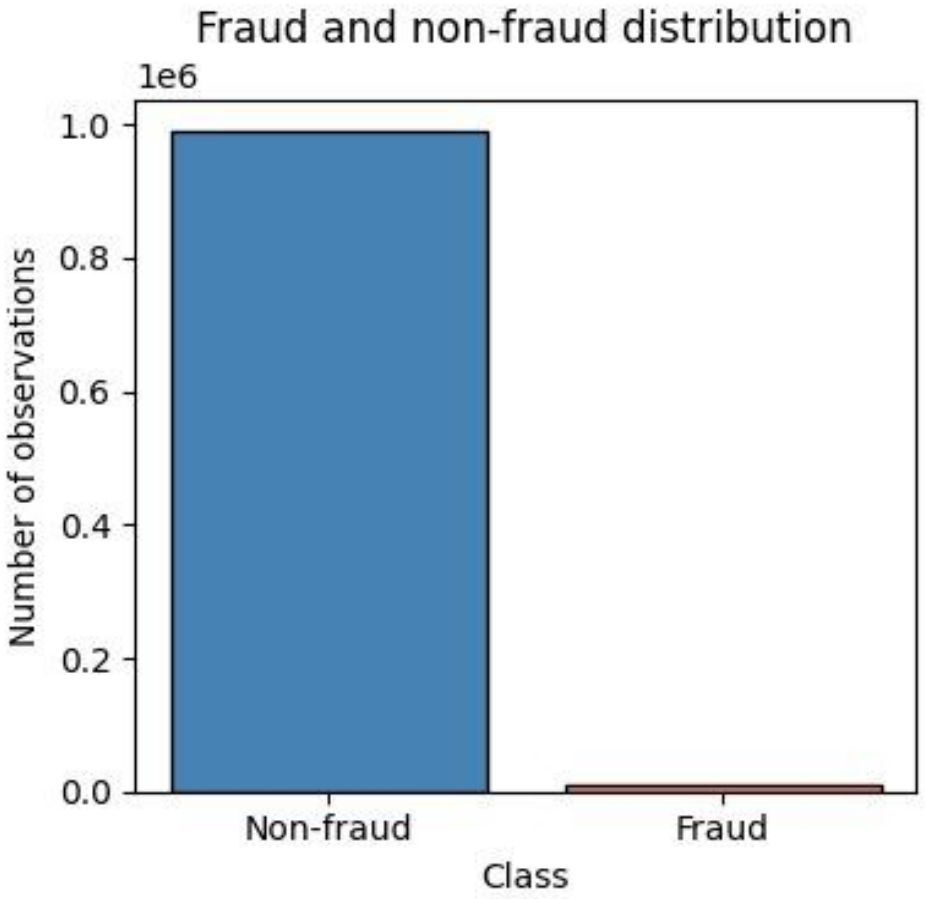
new_client_flag

client with limited history, potentially suspicious profile

high_credit_risk_flag

Profile financial already exhibiting high-risk characteristics

Advantage : Instead of three separate variables, a single metric synthesizes the overall risk profile.



This graph shows that the dataset is extremely unbalanced between fraud and non-fraud.

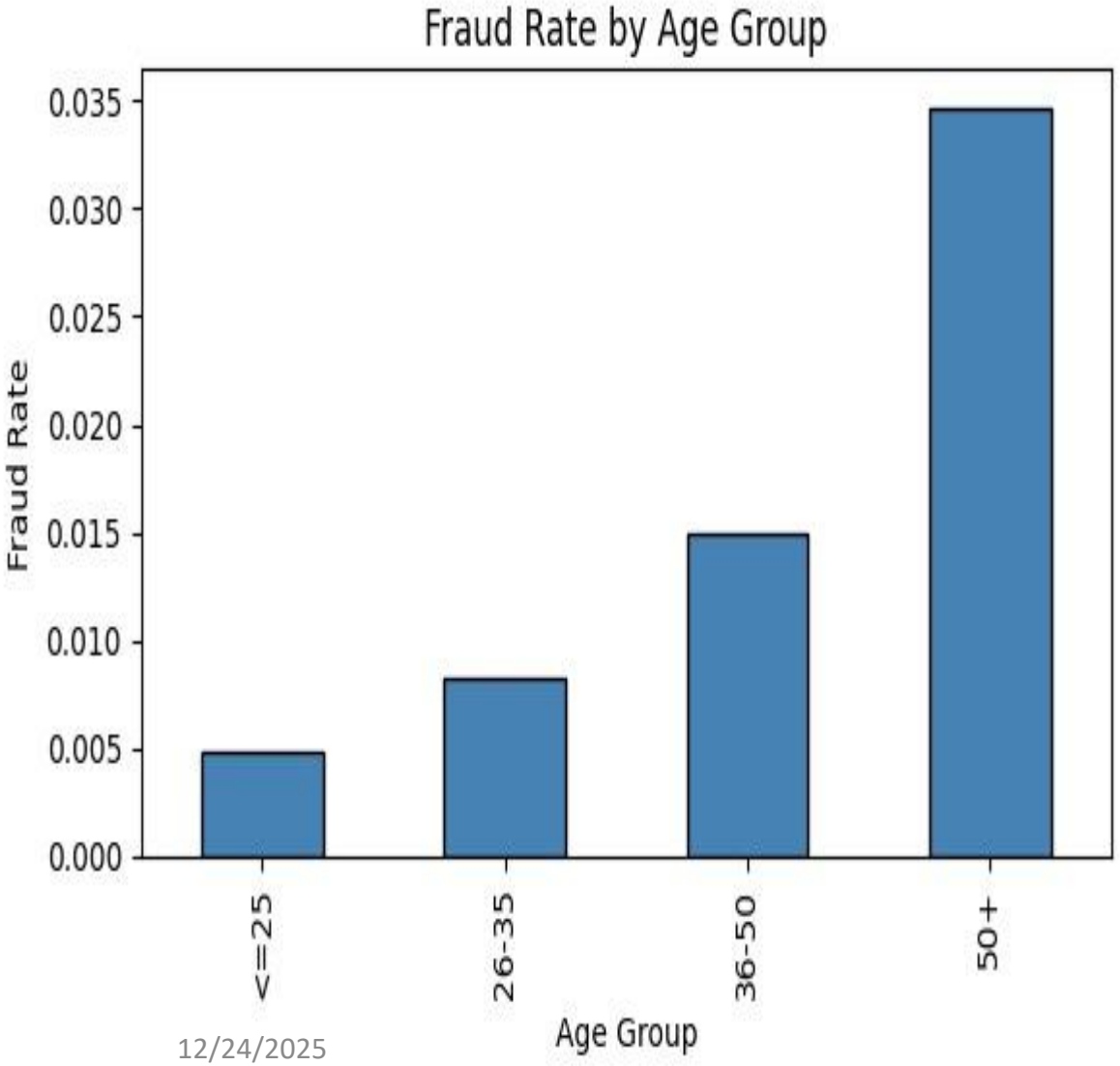
The vast majority of observations are not fraudulent. This class imbalance is typical of fraudulent data and makes detection more difficult. It is therefore necessary to balance the data.

Final Analysis :

At the end of our work, data analysis allowed us to identify the characteristics related to fraud cases, namely: age, type of payment, credit score, and the amount requested.

Visualization using matplotlib

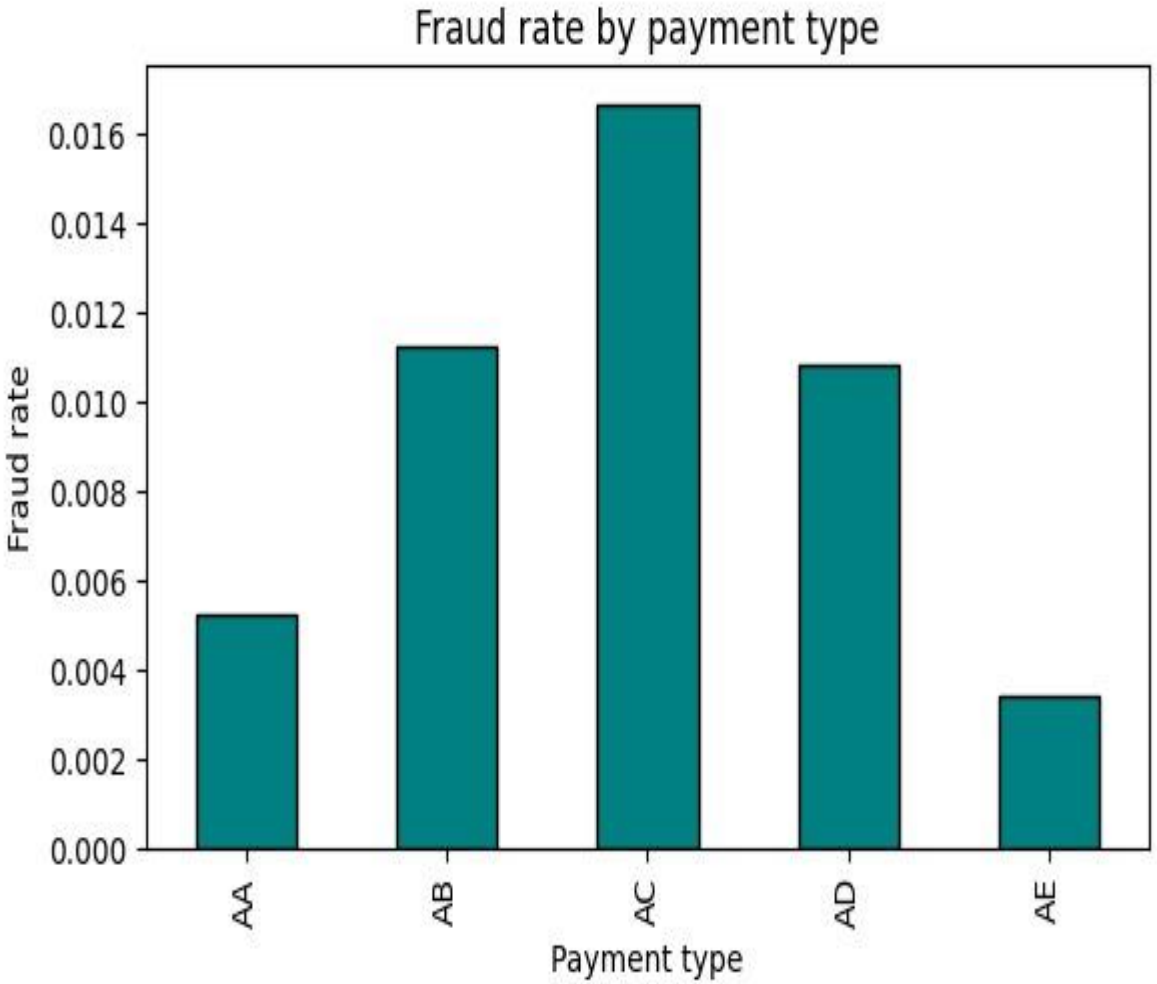
Visualization of fraud rates by age



This graph shows that the fraud rate increases with age.
Customers aged 50 and over have the highest fraud rate, much higher than other groups.
The groups ≤25, 26-35 and 36-50 have lower rates .

Analysis by age group shows a progressive increase in the fraud rate among customers aged 50 and over. This result suggests that advanced age is a factor associated with a higher risk of fraud and should be taken into account in customer risk assessment.

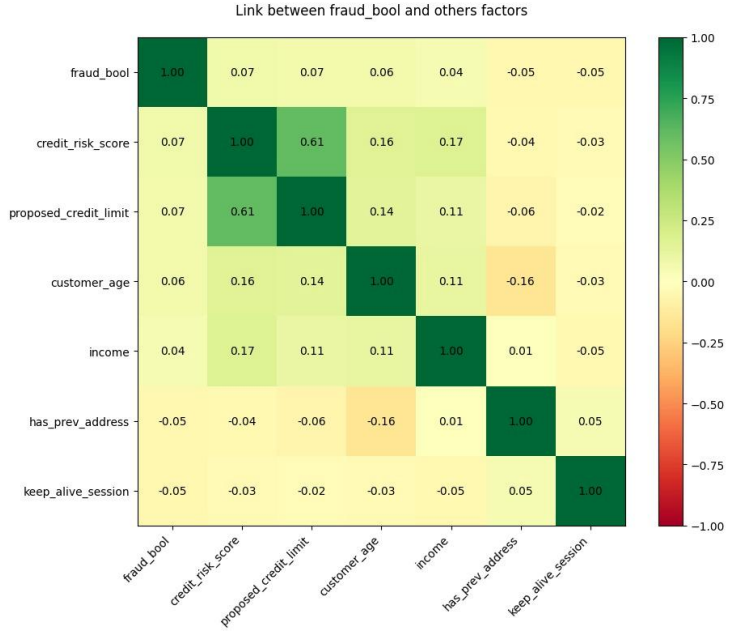
View by payment type

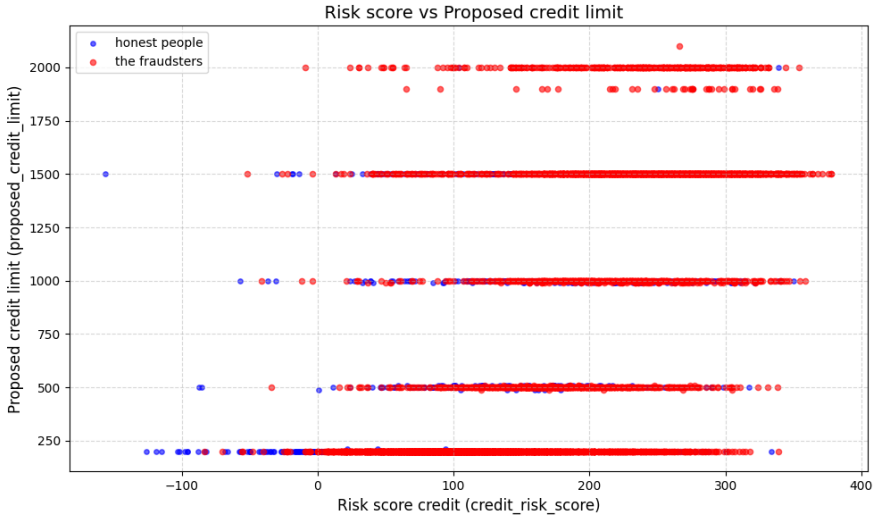


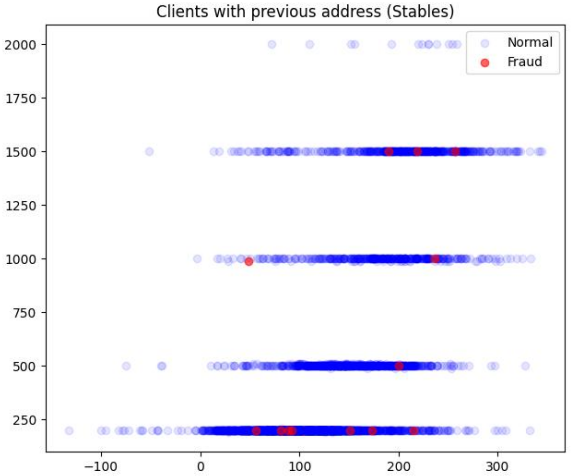
Analysis by payment type reveals that some channels are much more exposed than others.

Type AC concentrates the highest proportion of frauds, while AA and AE remain relatively secure.

The type of payment thus appears as an important indicator to integrate into fraud detection.







We can notice that:

- There is a strong correlation between credit_risk core and proposed_credit_limit based on the heatmap
- Aso notice that we get there is another variable called «having_previous_address» that really affect
  the primary factors and the faud_bool.

Improvement of the fight against fraud based on data analysis

- To watch in Priority is given
  pupil especially When She comes from a lower-

- Monitor certain types of payment more closely
  (such as type AC), which have a rate of
  fraud is significantly higher than through other channels.

- Use the client_risk_score to prioritize checks:
  require manual control
  or additional supporting documents for customers with
  a score of 2 or 3 before granting credit

Future Perspective

Develop a fraud detection model using the variables

The most influential factors (credit limit, risk score, age, payment type,has previous address) .

Implement a real-time alert system as soon as a new request accumulates several risk signals (high score, risky payment type, recent customer).

Enrich the data with new behavioral indicators (connection history, request frequency, inconsistencies in declarations) to further refine fraud detection in the future.


Thank you