Rakamin Academy X HOME CREDIT

HOME CREDIT SCORECARD MODEL

Home Credit – Data Scientist

Presented by:

Mardio Edana Putra

Rakamin Academy

mardiopq@gmail.com

mardiopq99

Courses and Certification

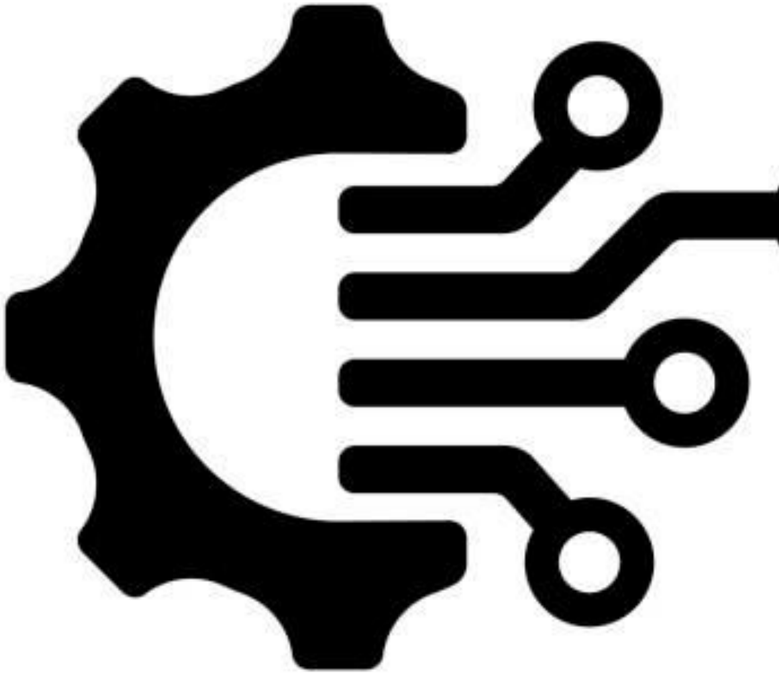
Mardio Edana Putra

Mardio Edana Putra

Data Analyst

I am an aspiring data analyst/scientist with a strong interest in data and its applications in business. I'm passionate about using data to uncover insights that can support better decision-making and drive growth. While I'm currently exploring opportunities, I am eager to apply my skills and knowledge in the world of data

Rakamin Academy



Introduction

Welcome to my portfolio!

Here, you'll find a selection of my projects related to data analysis. These projects showcase my interest in extracting insights from data and applying them to real-world problems, with the goal of driving informed decision-making and delivering impactful solutions.

Education

2017 - 2022
Bandung Institute of Technology
Industrial Engineering

- **GPA 3.38/4.00**
- **Thesis Title: Proposal for Product Quality Improvement of L14 Nails Using Six Sigma Methodology at PT Surabaya Wire**

2014 - 2017
8 Senior High School Jakarta
Math and Science

Experience

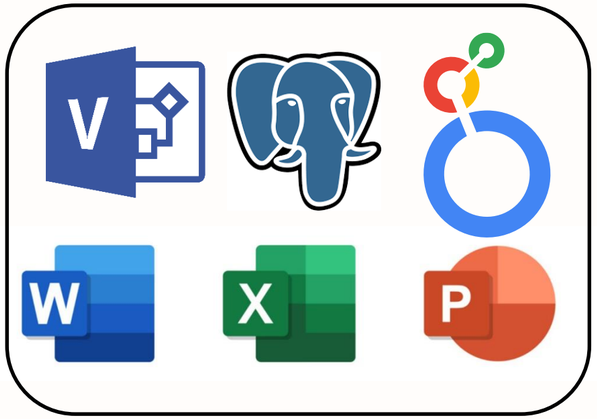
•**Leader of Final Project E-Commerce Data Scientist Bootcamp – Rakamin Academy**
August 2024 to January 2025
Led a team in developing machine learning models for an e-commerce case study, achieving the excellence grade of 89.2. Gained experience in SQL, Python, data preprocessing, statistics, machine learning, and data visualization.

•**Quality Control Intern – PT Surabaya Wire**
April 2021 to March 2022
Analyzed nail defects using the DMAIC methodology and recommended improvements to reduce defect rates.

•**Marketing Analyst Intern – PT Enciety Binakarya Cemerlang**
June 2020 to September 2020
Measured participant satisfaction using surveys and statistical methods, providing recommendations to improve training programs.

Skills and Expertise

1. Programming Language
2. Microsoft (Word, Excel, PPT, Visio)
3. Data Analysis



4. Mathematics
5. Data Visualization
6. Machine Learning

1. Problem Research

Home Credit Indonesia is a financial services company established in 2013, focusing on responsible lending for consumers across Indonesia. By leveraging data and technology, the company provides accessible and transparent credit solutions, including installment financing and cash loans. It serves both offline and online customers, promoting financial inclusion among the underbanked population. With millions of customers served, Home Credit continues to support economic empowerment through smart, fast, and secure financing. Its mission is to help people live better lives by making financial services simple and inclusive.

🔍 Problem Statement

Home Credit wants to improve its credit approval process by accurately identifying customers who are likely to default. Misclassifying good customers can result in lost revenue, while approving high-risk applicants increases the chance of bad debt.

🎯 Goal

To analyze and understand patterns that distinguish customers who are likely to default from those who are likely to repay loans on time.

🎯 Objective

To develop predictive models that can classify applicants based on their credit risk level, supporting more accurate and efficient credit decision-making.

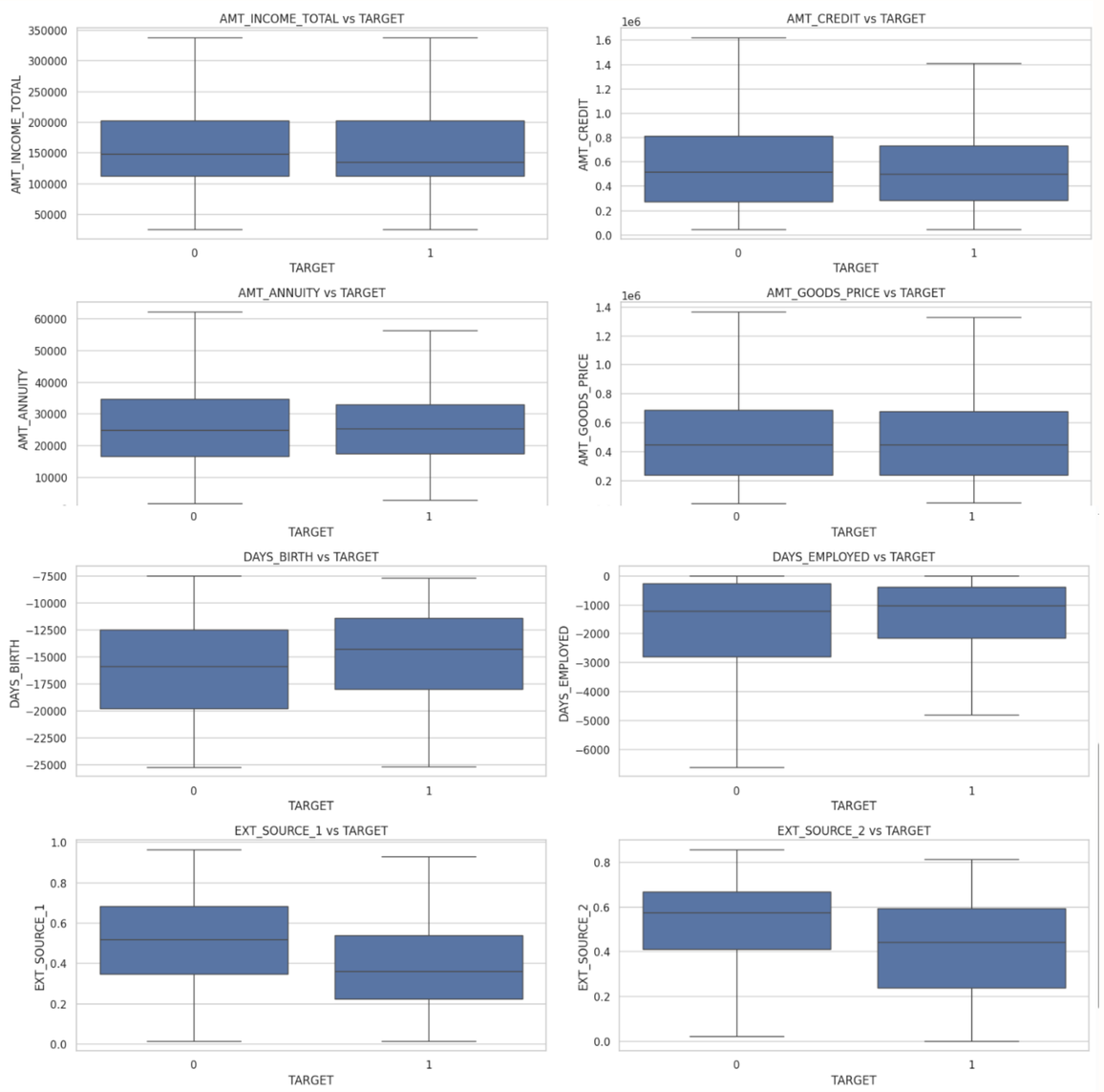
📊 Business Metrics

- Default Rate Reduction
- Approval Rate of Low-Risk Applicants
- Model Performance (Accuracy, Precision, Recall)

Data Description

Category	Example Features	Brief Description
Borrower Identity	SK_ID_CURR, TARGET, CODE_GENDER	Unique ID, default status, gender
Financial Status	AMT_INCOME_TOTAL, AMT_CREDIT	Income and loan amount
Family & Housing	CNT_CHILDREN, NAME_HOUSING_TYPE	Number of children, housing type
Employment & Social	NAME_INCOME_TYPE, OCCUPATION_TYPE	Job type and income source
Time-related Info	DAYS_BIRTH, DAYS_EMPLOYED	Age & employment duration (in days)
Communication	FLAG_PHONE, FLAG_EMAIL	Ownership of communication devices
Location & Region	REGION_RATING_CLIENT, LIVE_CITY_*	Region rating & domicile-workplace match
Property Details	OWN_CAR_AGE, *_AVG, *_MODE	Property ownership & building statistics
Loan Application	WEEKDAY_APPR_PROCESS_START	Loan application timing
Credit History	AMT_REQ_CREDIT_BUREAU_*	Frequency of credit requests to bureau

2. Exploratory Data Analysis (1)



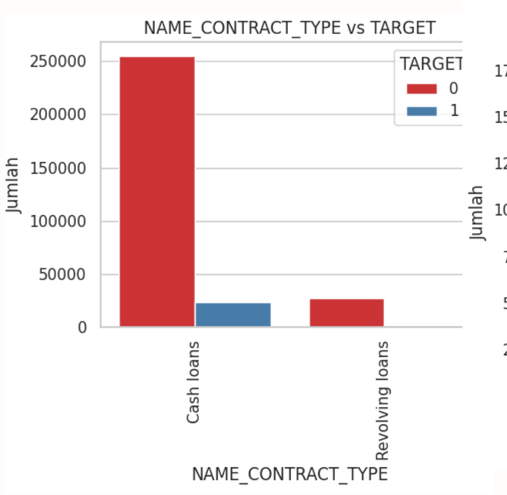
🔍 Key Business Insights from Numerical Features

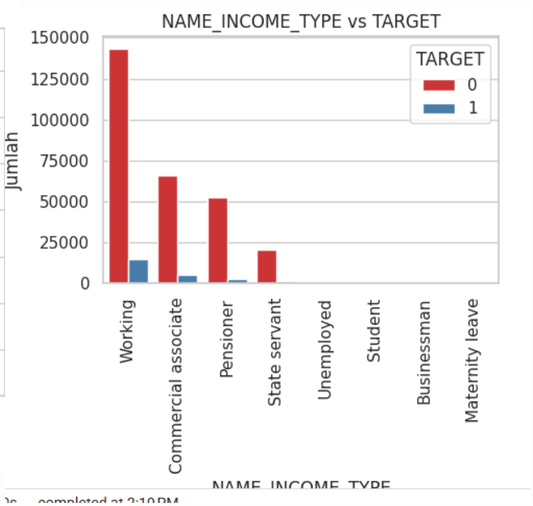
- **Most informative features:** EXT_SOURCE_1, EXT_SOURCE_2, EXT_SOURCE_3 → clear differences in median and distribution between TARGET classes.
- **Moderately informative features:** DAYS_BIRTH (younger borrowers are more likely to default), DAYS_EMPLOYED (shorter employment duration = higher risk).
- **Less informative features:** AMT_INCOME_TOTAL, AMT_ANNUITY, CNT_FAM_MEMBERS → similar medians and distributions across TARGET values → **do not significantly distinguish TARGET.**

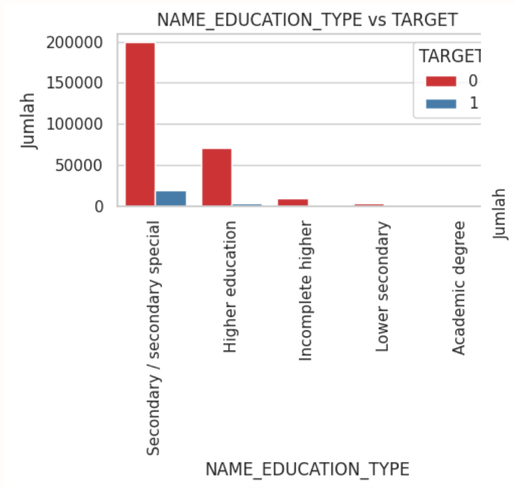
Would you like this summary added to the previous table or formatted for a slide as well?

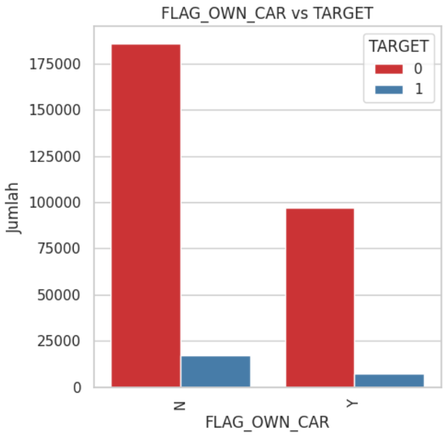
2. Exploratory Data Analysis (2)

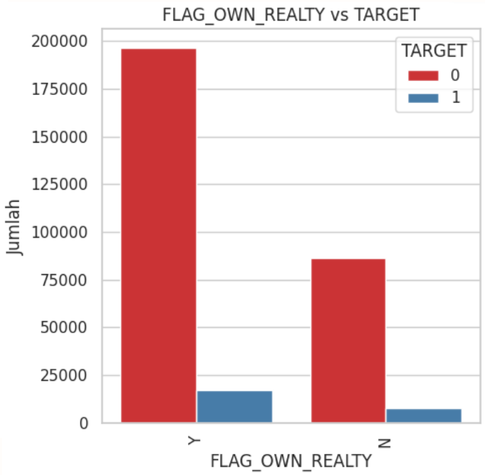
Rakamin Academy











📊 Summary of Categorical Feature Insights

Analysis shows that **cash loans** are more commonly chosen and have a **higher chance of repayment issues** compared to revolving loans. **Gender differences** in default are minimal, with slightly more issues among female applicants. **Applicants without cars or property** tend to struggle more with repayments, suggesting that **asset ownership relates to financial stability**.

Similarly, **stable jobs** (like pensioners or government employees) show better repayment behavior than informal jobs (like laborers or sales staff).

Lastly, **higher education levels** are linked to **better repayment performance**, while those with only basic education default more often.
Overall, variables such as **loan type, job type, asset ownership, and education** are useful indicators for credit risk and should be prioritized in risk scoring and policy decisions.

3. Data Pre-Processing Summary

Rakamin Academy

Step	Description
1. Missing Value Analysis	- Identify features with low missing values → impute (median/mode) - High missing values (>50%) → consider dropping the columns
2. Data Imputation	- Numeric imputation: median - Categorical imputation: mode or use label like "Unknown" (e.g., for OCCUPATION_TYPE)
3. Drop Irrelevant Columns	- Drop features with excessive missing values and statistically redundant features (*_AVG, *_MODE, *_MEDI, etc.)
4. Duplicate Handling	- Check for and remove duplicate rows
5. Binary Value Conversion	- Map binary categorical values: CODE_GENDER, FLAG_OWN_CAR, FLAG_OWN_REALTY → 0/1
6. Categorical Encoding	- Label Encoding: for ordinal feature NAME_EDUCATION_TYPE - One-Hot Encoding: for nominal features (e.g., NAME_INCOME_TYPE, WEEKDAY_APPR_*)
7. Drop High Cardinality	- Drop columns like OCCUPATION_TYPE and ORGANIZATION_TYPE due to too many unique categories and high missing rate
8. Feature Selection	- Apply Mutual Information to select top 30 most informative features related to the target (TARGET)

4. Modeling Result

Model	Accuracy Train	Accuracy Test	Precision Train	Precision Test	Recall Train	Recall Test	F1 Train	F1 Test	Time Taken (s)
Logistic Regression	0.68140	0.68357	0.89316	0.89378	0.68140	0.68357	0.75352	0.75517	27.79
Decision Tree	1.00000	0.86026	1.00000	0.86247	1.00000	0.86026	1.00000	0.86136	5.36
Random Forest	0.99995	0.91929	0.99995	0.88667	0.99995	0.91929	0.99995	0.88128	97.55
XGBoost	0.92180	0.91938	0.91241	0.88815	0.92180	0.91938	0.88814	0.83881	4.32

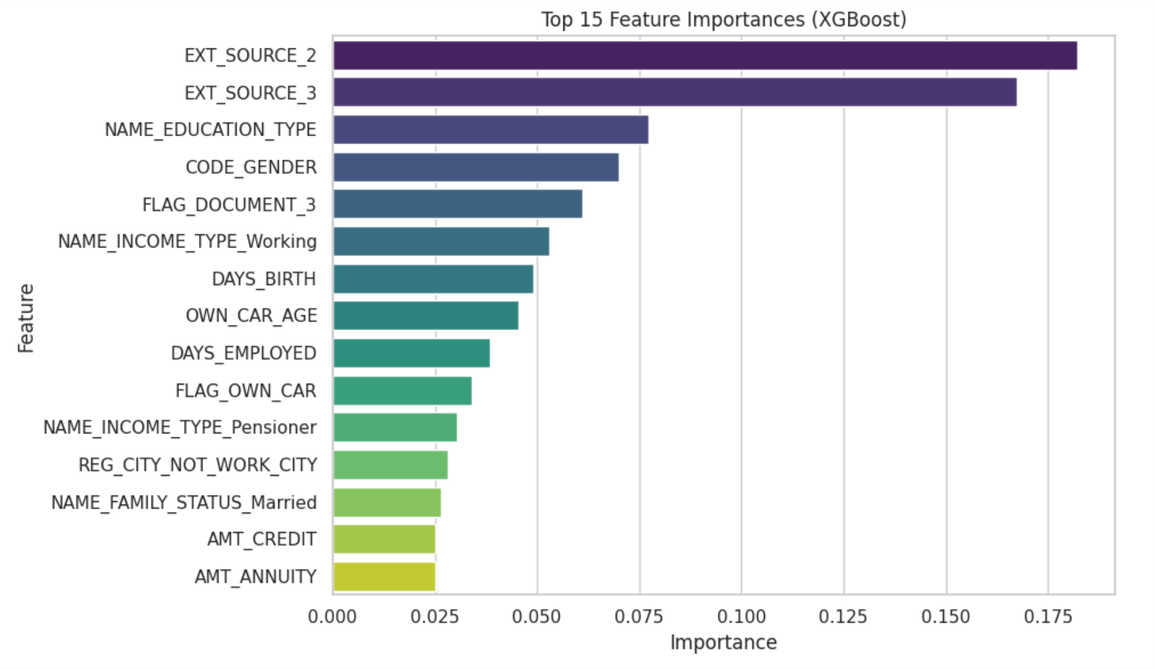
✅ **Model Selection Justification**

•**XGBoost** achieves an excellent balance between train and test scores across all key metrics (Accuracy, Precision, Recall, F1), suggesting **strong generalization without overfitting**.

•While Random Forest performs similarly on test metrics, its training metrics are nearly perfect (overfitting), and its computation time is **over 20× longer** than XGBoost.

•XGBoost combines **high performance and low computation time (4.32s)**, making it optimal for both **accuracy and efficiency**.

5. Feature Importance



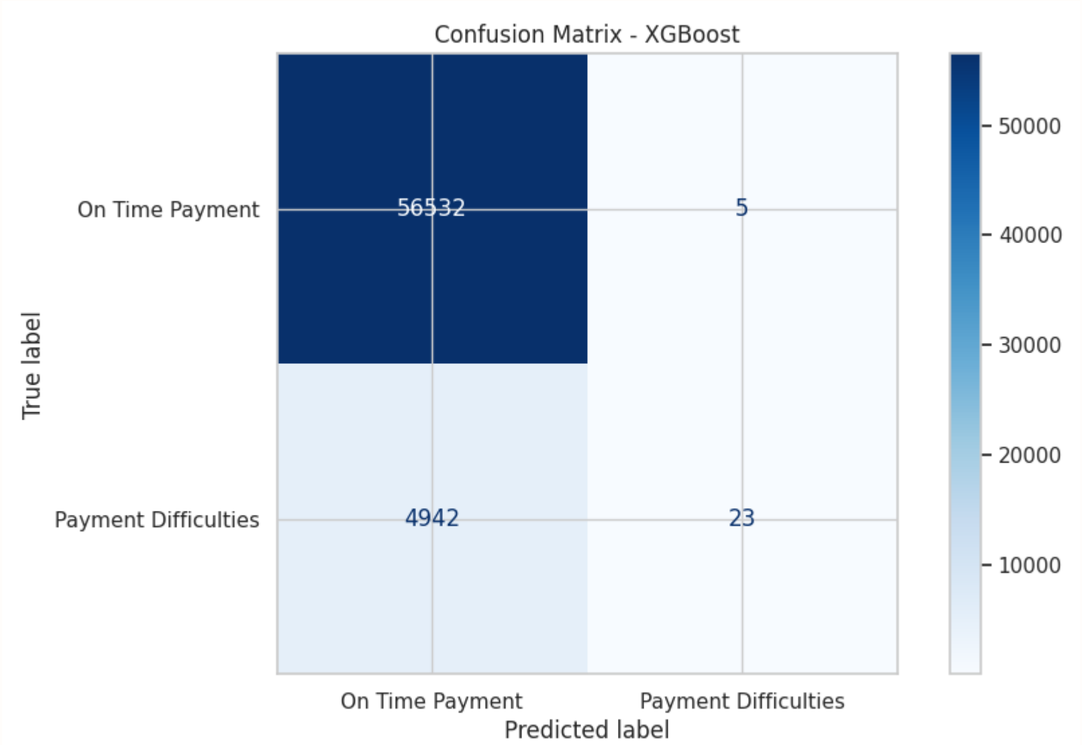
🎯 Feature Importance Summary (XGBoost Model)

Rank	Feature	Insight
1	EXT_SOURCE_2, EXT_SOURCE_3	External credit scores — strongest predictors of loan default risk.
2	NAME_EDUCATION_TYPE, CODE_GENDER	Education level and gender play a significant role in loan decisions.
3	FLAG_DOCUMENT_3	Document completeness likely reflects creditworthiness.
4	NAME_INCOME_TYPE_Working, DAYS_BIRTH, OWN_CAR_AGE	Indicators of income stability and borrower maturity.
5	FLAG_OWN_CAR, REG_CITY_NOT_WORK_CITY	Reflect lifestyle and employment-residence consistency.
6	AMT_CREDIT, AMT_ANNUITY, NAME_FAMILY_STATUS, REGION_RATING_CLIENT	Financials have moderate to low impact.

📌 Business Insights

- ✅ Prioritize **external scores** and **document completeness** in the credit screening process.
- ✅ **Demographic and employment stability data** are critical and should be captured accurately.
- ⚠️ Financial figures like **loan amount** and **annuity** are **less decisive** in default prediction — creditworthiness is more tied to behavior and stability, not just numbers.

6. Confusion Matrix



📌 Key Takeaways

•True Positives (56,532):
The model **accurately identified the vast majority** of customers who paid on time.
→ Suggests strong performance in recognizing **low-risk borrowers**.

•False Positives (5):
Very **few customers were incorrectly predicted** as defaulters.
→ Indicates the model is **conservative in flagging risk**, minimizing lost opportunities due to misclassification.

•False Negatives (4,942):
A **notable number of customers who defaulted** were predicted as low risk.
→ Highlights a common challenge in **imbalanced datasets**, where minority class detection may require **further enhancement**.

•True Negatives (23):
A small number of true defaulters were correctly predicted.
→ Shows the model's **potential to learn from risky profiles**, which can be expanded with **class balancing or more data**.

7. Business Recommendation

No.	Insight (Finding)	Business Action	Expected Impact
1	**Cash loans** are the most common loan type and show **higher default rates** compared to revolving loans.	Reevaluate cash loan terms, such as tightening credit score requirements or applying higher interest for high-risk cases.	Reduce default risk from more vulnerable loan segments.
2	**Slightly higher default rates** are observed among **female** applicants.	Develop financial literacy programs for female borrowers on managing budgets after taking loans.	Improve financial knowledge and reduce defaults in this segment.
3	Applicants with **formal and stable jobs** (e.g., pensioners, state servants) show **lower default rates** than those in informal employment.	Prioritize applicants with stable job types for loan approval, or offer special deals (e.g., lower interest rates).	Increase profitability with lower-risk customers.
4	Applicants who **do not own a car or property** are more likely to default.	Use asset ownership as an additional credit scoring factor.	Support more accurate creditworthiness assessments.
5	Surprisingly, **higher education** (e.g., university level) does **not guarantee better payment behavior**, with some defaults higher than secondary education.	Avoid assuming higher education equals lower risk; apply a data-driven approach instead.	Achieve more objective risk profiling and avoid bias.
6	External scores (EXT_SOURCE_1/2/3) strongly influence creditworthiness.	Use them as core features in credit scoring systems.	More accurate and reliable loan decisions.

Conclusion

In this **Home Credit scorecard modeling project**, **four machine learning models**, **Logistic Regression**, **Decision Tree**, **Random Forest**, and **XGBoost** were built to **predict repayment behavior** of loan applicants. Among them, **XGBoost achieved the highest recall**, making it the **most suitable** for detecting **high-risk customers**, while **Logistic Regression** was valuable for its **simplicity and interpretability**. This supports Home Credit's objective to **reduce default rates** and **increase approval** among **low-risk applicants**. Insights from **categorical features** and **feature importance** further help the business **avoid rejecting capable payers** and allow more **personalized loan planning**. Moving forward, **XGBoost** can be adopted for **deployment** due to its **strong predictive performance**, supported by ongoing **monitoring and refinement**.

LINK FILES

THANK YOU!