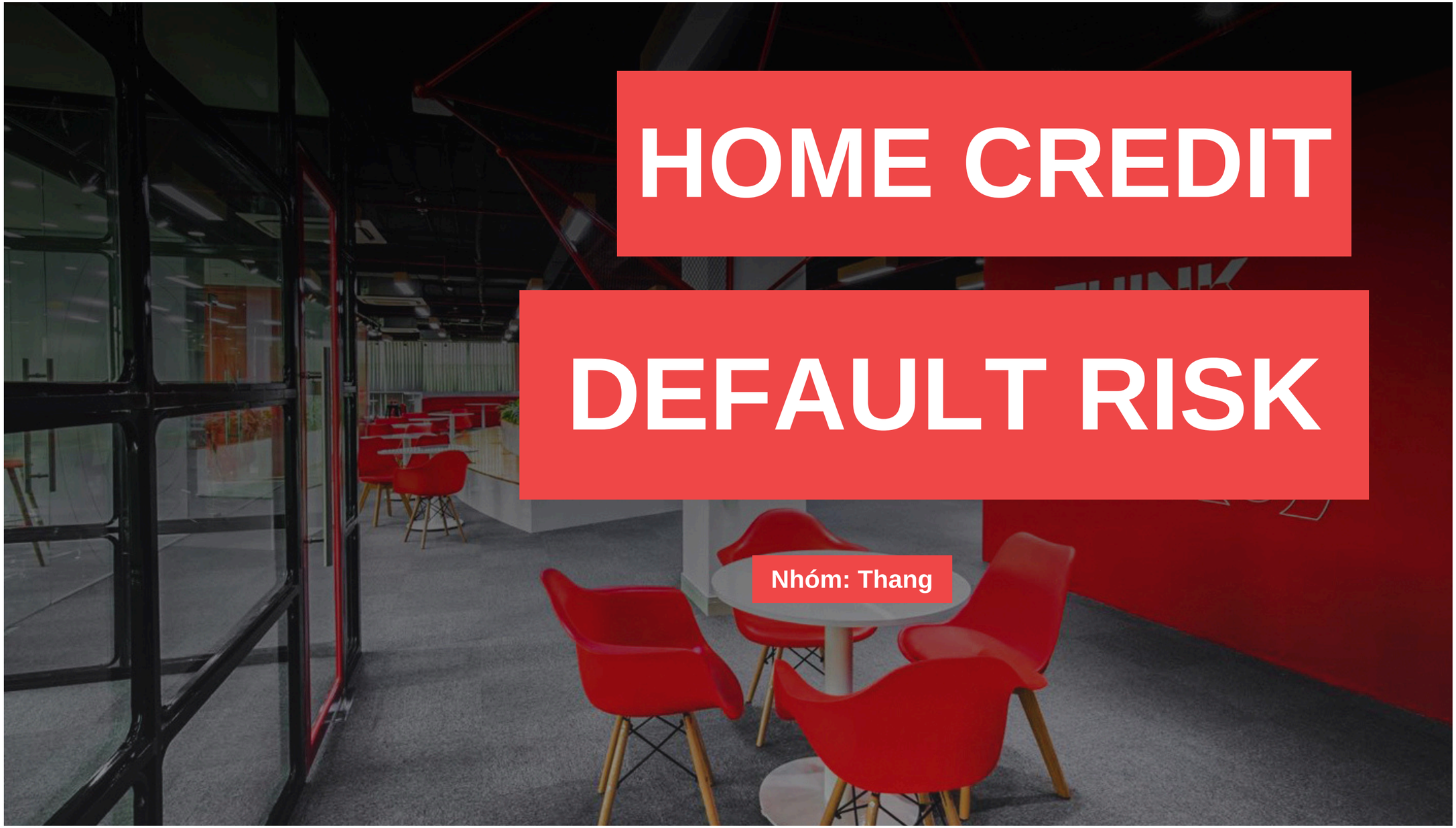
THANG LONG UNIVERSITY



HOME CREDIT

DEFAULT RISK

Nhóm: Thang

PIPELINE

01 **INTRODUCTION**
Introduce about HOME CREDIT, how they work and problem statement

02 **DATA PROCESSING**
Cleaned, encoded, and engineered features to prepare data

03 **ANALYSIS AND RESULT**
Using model to identify patterns and correlations in default risk and model evaluation

INTRODUCTION

ABOUT HOME CREDIT

Home Credit is a global financial services company that specializes in consumer finance, particularly in emerging markets.

Founded in 1997 in the Czech Republic, Home Credit operates in various countries across Asia, Europe, and beyond

HOW THEY WORK

Customer choose a product

Home Credit evaluates the customer’s creditworthiness

Loan Disbursement and Purchase

Customer pay the money

MAIN GOAL

Be able to predicts the probability of a customer defaulting on a loan

Because the economy is quite difficult at the moment, high recall index will be prioritized

AUDIENCE

- Risk Management Department
- Business Strategy and Product Development Department
- Senior Management

DATA PROCESSING

OVERVIEW

01

Exploring the dataset

02

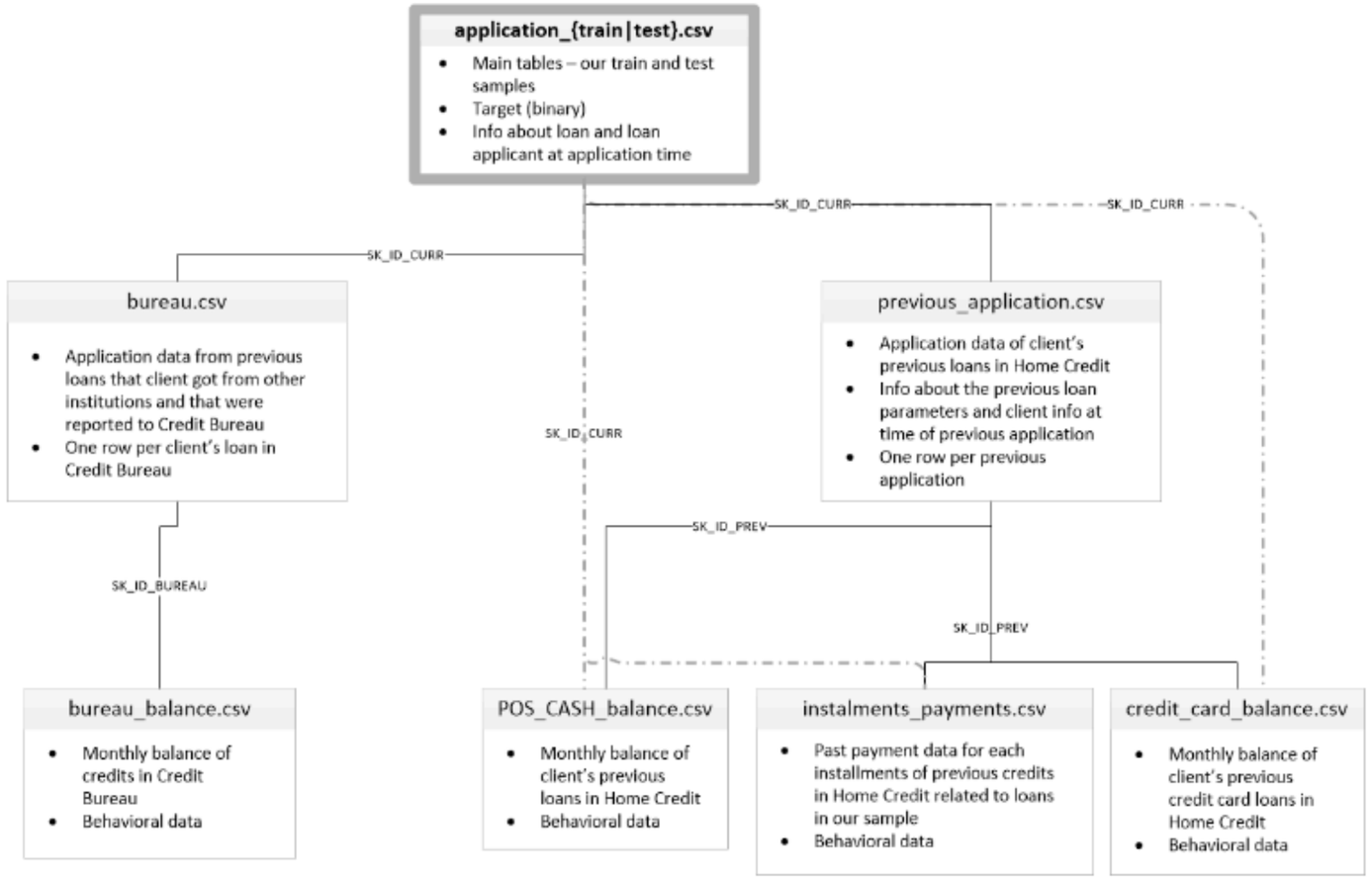
Data cleaning

03

Feature engineering

Home Credit Default Risk dataset

The Home Credit Default Risk dataset is provided by Home Credit Group, a global consumer finance company. It originates from real-world lending data, where Home Credit assesses loan applicants, many of whom lack formal credit histories.



Data exploration

application_train.csv

Main table, Static data for all applications

- SK_ID_CURR: Unique identifier for each loan applicatio
- TARGET: Binary label indicating if the client defaulted on the loan. 0 = good, 1 = bad
- AMT_CREDIT: The amount of credit for the loan
- DAYS_BIRTH: Client’s age in days

bureau.csv

All client's previous credits provided by other financial institutions

- SK_ID_CURR: Unique identifier for the client
- CREDIT_ACTIVE: The status of the credit at the time of data collection
- CREDIT_DAY_OVERDUE: Number of days the credit payment is overdue

previous_application.csv

All previous applications for Home Credit loans

- NAME_CONTRACT_STATUS: Status of the previous application
- AMT_APPLICATION: The amount of credit applied for in the previous application
- AMT_CREDIT: The amount of credit approved for the previous application

Data cleaning

- Calculate missing percentage in every columns
  Remove columns with high missing percentage
- Calculate medium, median to fill in some missing columns
- With some missing columns that’s unable to fill in by median, fill -1 in to not affect the model training

Feature engineering

SK_ID_CURR	AGE	CREDIT_INCOME_RATIO	ANNUITY_INCOME_RATIO	CREDIT_GOODS_RATIO
100007	55	4.222222222222222	0.17996296296296296	1.0
100007	55	4.222222222222222	0.17996296296296296	1.0
100007	55	4.222222222222222	0.17996296296296296	1.0
100007	55	4.222222222222222	0.17996296296296296	1.0
100007	55	4.222222222222222	0.17996296296296296	1.0
100007	55	4.222222222222222	0.17996296296296296	1.0
100008	47	4.9545	0.27795454545454545	1.0791980198019802
100008	47	4.9545	0.27795454545454545	1.0791980198019802
100008	47	4.9545	0.27795454545454545	1.0791980198019802
100008	47	4.9545	0.27795454545454545	1.0791980198019802

only showing top 10 rows

- Create AGE colums from ‘DAYS_BIRTH’
- Create ‘CREDIT_INCOME_RATIO’ from “AMT_CREDIT" and "AMT_INCOME_TOTAL”
- Create ‘ANNUITY_INCOME_RATIO’ from “AMT_ANNUITY" and "AMT_INCOME_TOTAL”
- Create ‘CREDIT_GOODS_RATIO’ from “AMT_CREDIT" and "AMT_GOODS_PRICE”

ANALYSIS AND RESULT

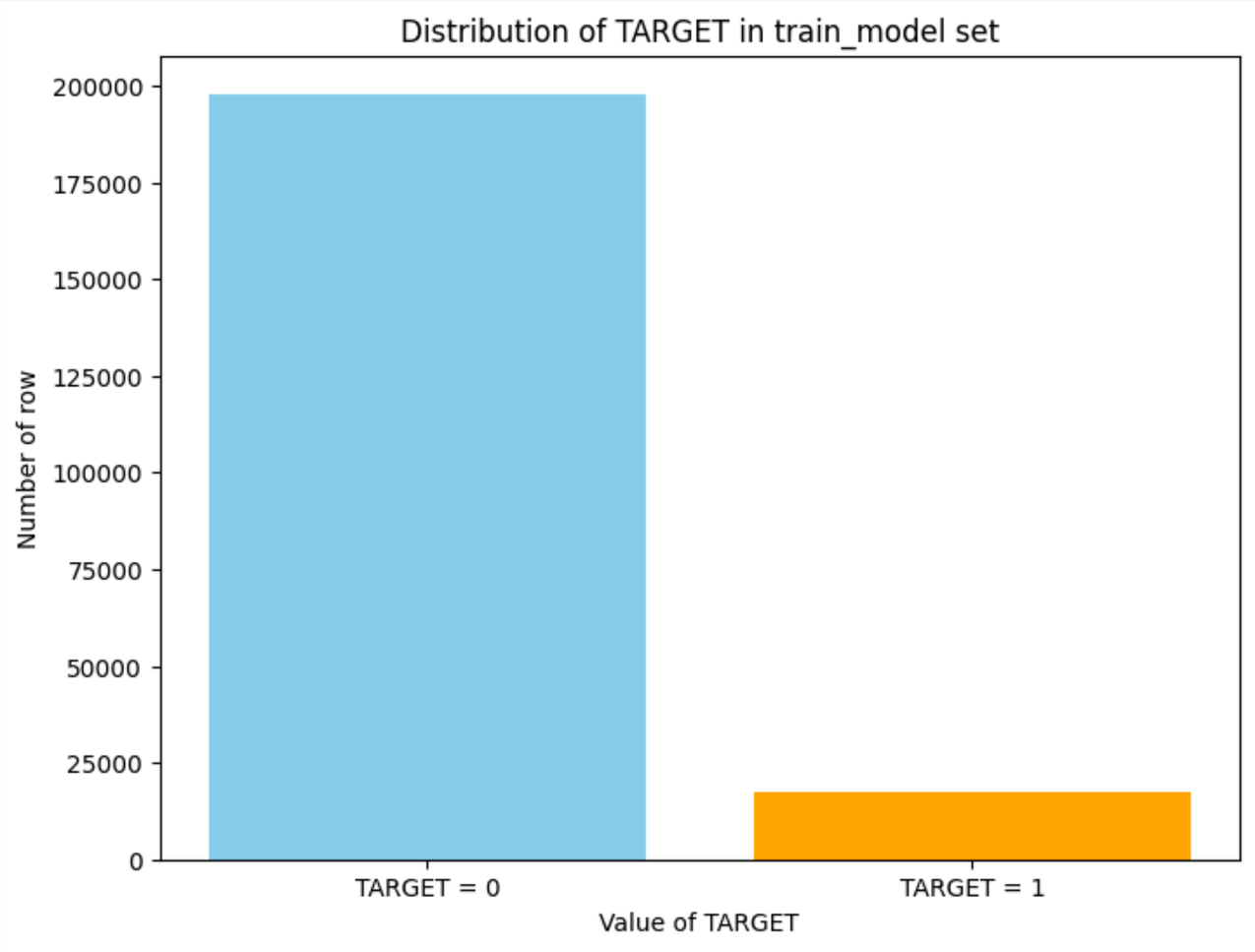
	Feature	IV
18	total_amt_credit	0.645132
3	AMT_ANNUITY	0.496698
17	total_amt_application	0.496602
20	total_amt_goods_price	0.495993
25	CREDIT_INCOME_RATIO	0.441936
15	avg_loan_duration	0.420824
1	DAYS_BIRTH	0.392690
5	EXT_SOURCE_3	0.387644
23	avg_days_decision	0.386442
14	total_amt_credit_sum_debt	0.349437
10	DAYS_REGISTRATION	0.334067
16	total_active_debt	0.325334
11	DAYS_ID_PUBLISH	0.271829
2	AMT_CREDIT	0.271500
8	DAYS_EMPLOYED	0.270522
19	total_amt_annuity	0.246783
13	total_amt_credit_sum	0.234204
26	CREDIT_GOODS_RATIO	0.156827
21	total_amt_down_payment	0.156583
4	AMT_GOODS_PRICE	0.128161
22	avg_rate_down_payment	0.115597
12	total_amt_credit_max_overdue	0.103094
24	AGE	0.090262
0	REGION_POPULATION_RELATIVE	0.074612
9	ORGANIZATION_TYPE	0.073368
6	NAME_INCOME_TYPE	0.058322
7	NAME_EDUCATION_TYPE	0.050836

USE WOE TO FIND OUT WHICH COLUMNS IS IMPORTANT

SET IV < 0.7 TO AVOID OVERFITTING
USE TOP 20 FEATURE TO TRAIN THE MODEL

ONE-HOT ENCODING THE ORGANIZATION_TYPE

HUGE GAP IN TARGET = 1 AND TARGET = 0



Use Random OverSampling to balance the data

Number of target = 0: 197880
Number of target = 1: 17347
Oversampling rate : 11.41

MODEL TRAINING

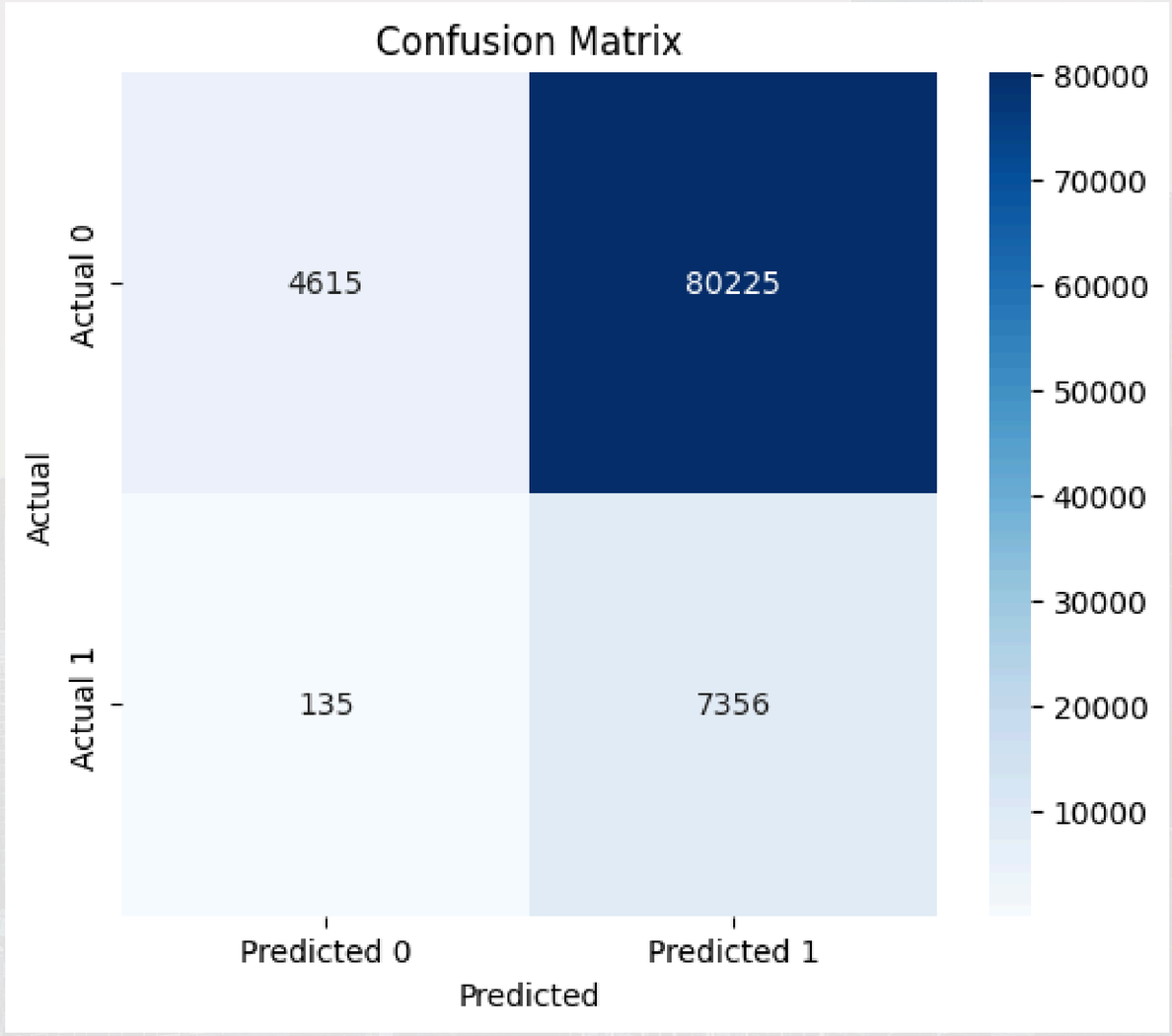
LOGISTIC REGRESSION: WHY?

Model	Ease of Use	Interpretability	Deployment Cost
Logistic Regression	Very Easy	Very Easy	Low (minimal resources)
Decision Tree	Easy	Easy	Medium (light resources)
Random Forest	Fairly Easy	Less Interpretable	High (more resources)
K-Nearest Neighbors	Moderate	Easy	High (computationally intensive for large data)

RESULT OF MODEL

Recall score: 0.9820

ROC-AUC score: 0.6513



COMPANY PROFIT

Assume that a company have:

- 2,500,000 customer
- Percentage of high-risk customers: ~10% (250,000 customers)
- Cost to investigate each customer: 10 USD.
- Loss incurred for each undetected high-risk customer (False Negative): 10,000 USD.

	Cost before using model	Cost after using model
Customers investigated	2500000 customers	87581 customer, 2412419 fewer
Investigation cost	25000000 USD	875810 USD, 24124190 USD saved
Loss from undetected risks (USD)	2500000000 USD	1350000 USD 2498650000 USD saved

Time saved: 96.50%
Money saved: 99.91%

Thanks you for listening

Any question?