HOME CREDIT DEFAULT RISK

TEAM „FUTURE WORK“

Daniel Büttner
Meteorologist

Jorrik Fulda
PhD Pol Science / Law

Dennis Aschenbrenner
Meteorologist

Overview

1. Introduction
2. Data Analysis
3. Predictive Modeling
4. Recommendations
5. Future Work

HC Business model - Provide loans to the unbanked!

Focus on underserved borrowers (blue collar and junior white collar) with regular income, who are less likely to access financing from banks and other traditional lenders.“

3 most offered Types of loans:

- **POS loans (in store or online, mobile phones)**
- **Cash loans (for any consumer goods, wedding)**
- **Revolving loans (credit cards, offered to existing cust.)**

Pic Source: Home Credit Website, moneyinc.com

9 Countries of Operation

HC started in Czechoslovakia in 1997 and enlarged from Eastern Europe to Asia (incl. Russia, China and India). Half of the worlds population lives in these countries.

Outstanding gross loans of all 116 Million customers at the end of March 2019 were nearly 20 Billion Euros!

Czech Billionaire Petr Kellner owns 91% of Home Credit!

Pic Source: Home Credit Website

Losses vs. Income

Title	2Q 2019	2Q 2018	2018	2017	2016
(MEUR)					
Net interest income	1,829	1,557	3,296	2,463	1,552
Operating income	2,102	1,942	4,025	3,152	2,014
Credit risk costs[1]	(863)	(988)	(1,711)	(1,126)	(562)
Operating expenses[2]	(833)	(855)	(1,700)	(1,583)	(1,079)
Net profit after tax	318	87	571	318	263
Net profit attributable to equity holders of the parent	318	863	571	321	266

1. Credit risk costs represent impairment losses
2. Operating expenses comprise general administrative and other operating expenses

- Loss of roundabout 40% = 1,7 Billion Euros in 2018

- If HC can better predict the „black sheep“ within the customers the company can diminish their losses!

Pic Source: Home Credit Website

Dataset

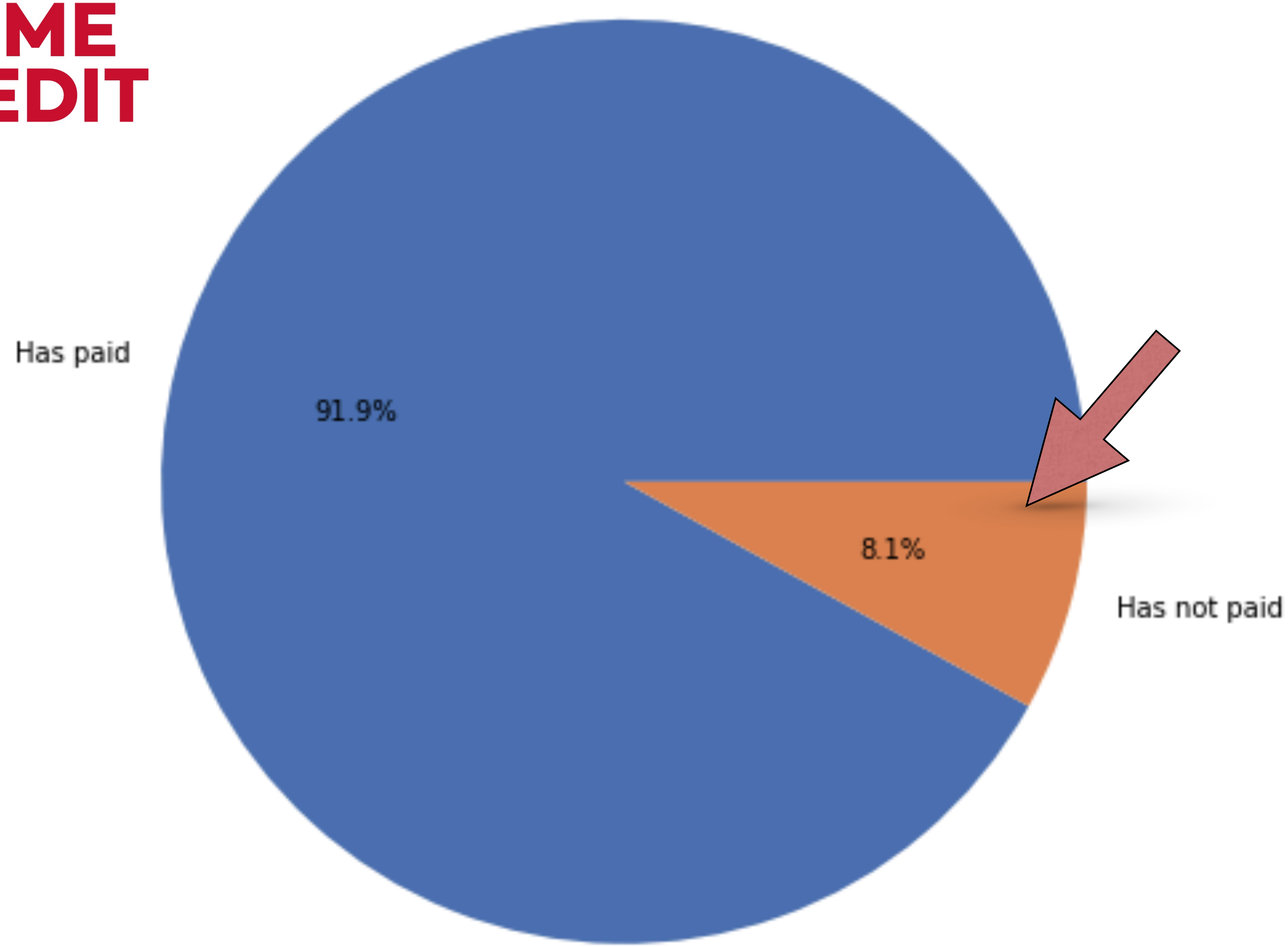
- Customer/Borrower = 300.000 (of 116 Million)
- 121 Features
- Credit Type: Cash, Revolving
- Overall Credit Amount = 170 Billion (Rupees?Ruble?)
- Median Credit Amount p. C. = 500.000
- Median Monthly Annuity p. C. = 25.000
- Average Monthly Income = 150.000

From which country is our dataset? Based on each currency rate it could be Russia (7000/353€), India (6200/314€), Philippines (8800/440€) or Kazakhstan (1100/58€).

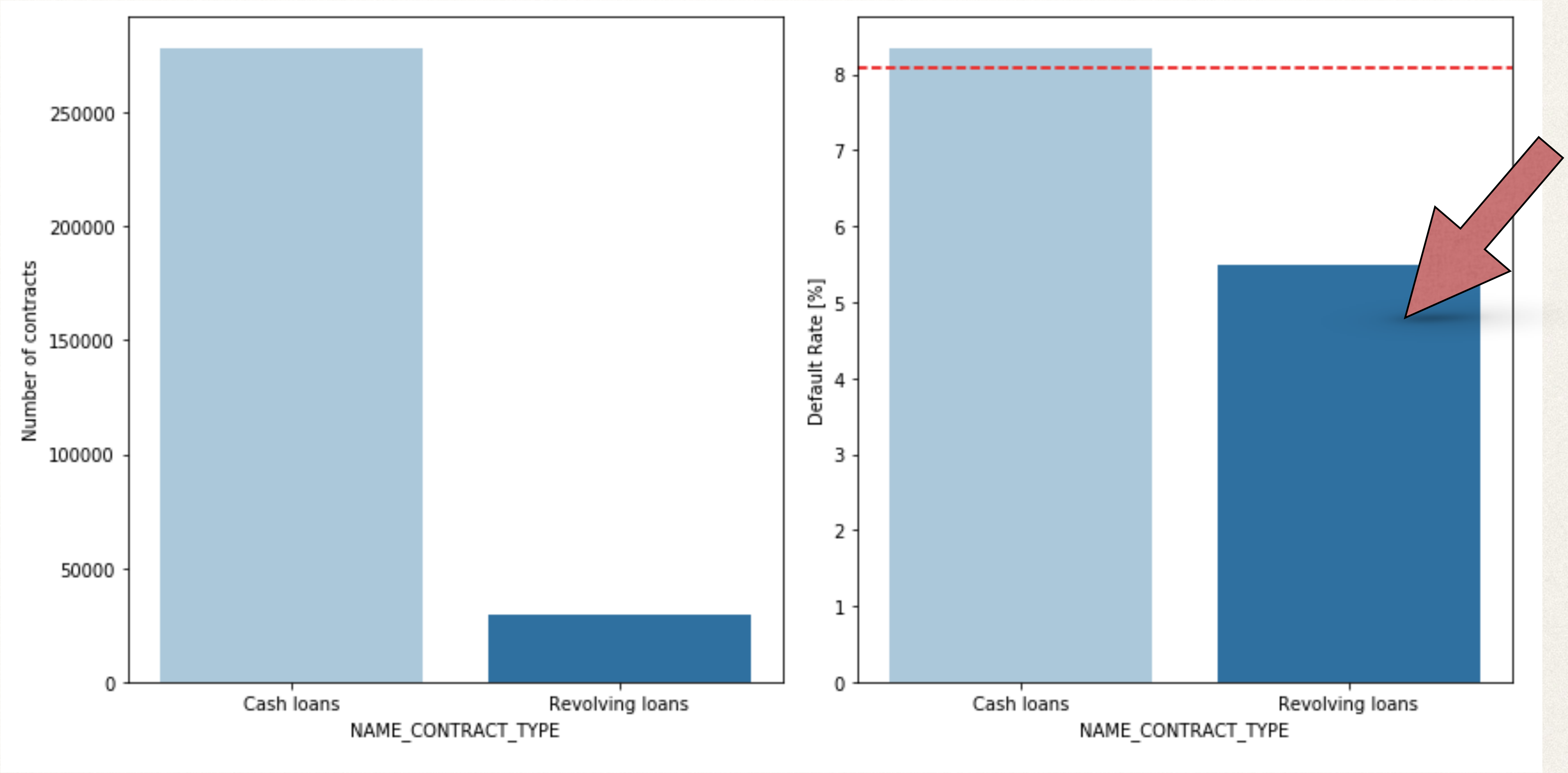
Business Objective

- Binary Classification Problem
- Predicting Credit default or not
- Goal: avoid false negatives, achieved by a model with a high recall rate
- Finding high-impact Features

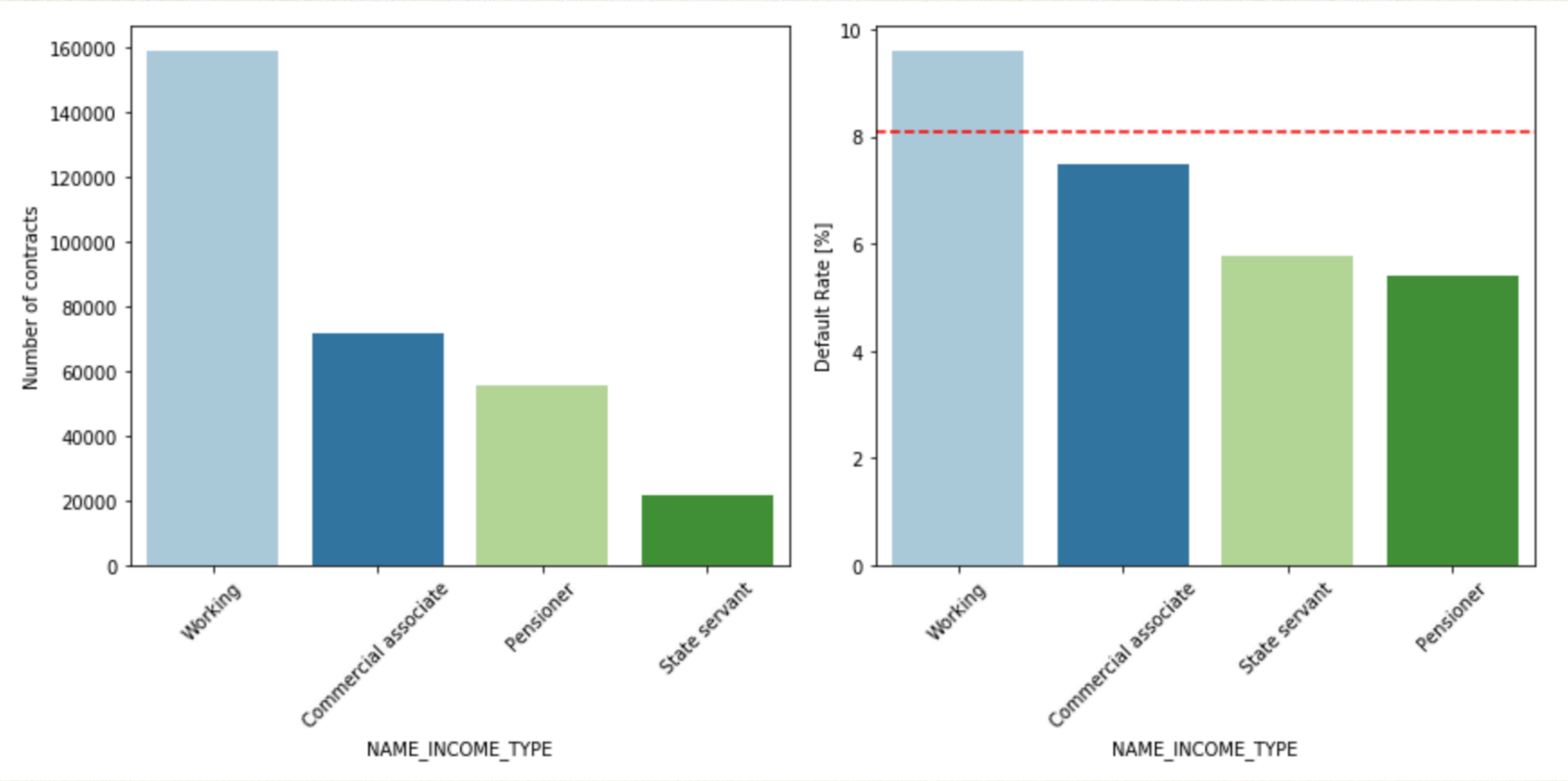
DATA ANALYSIS

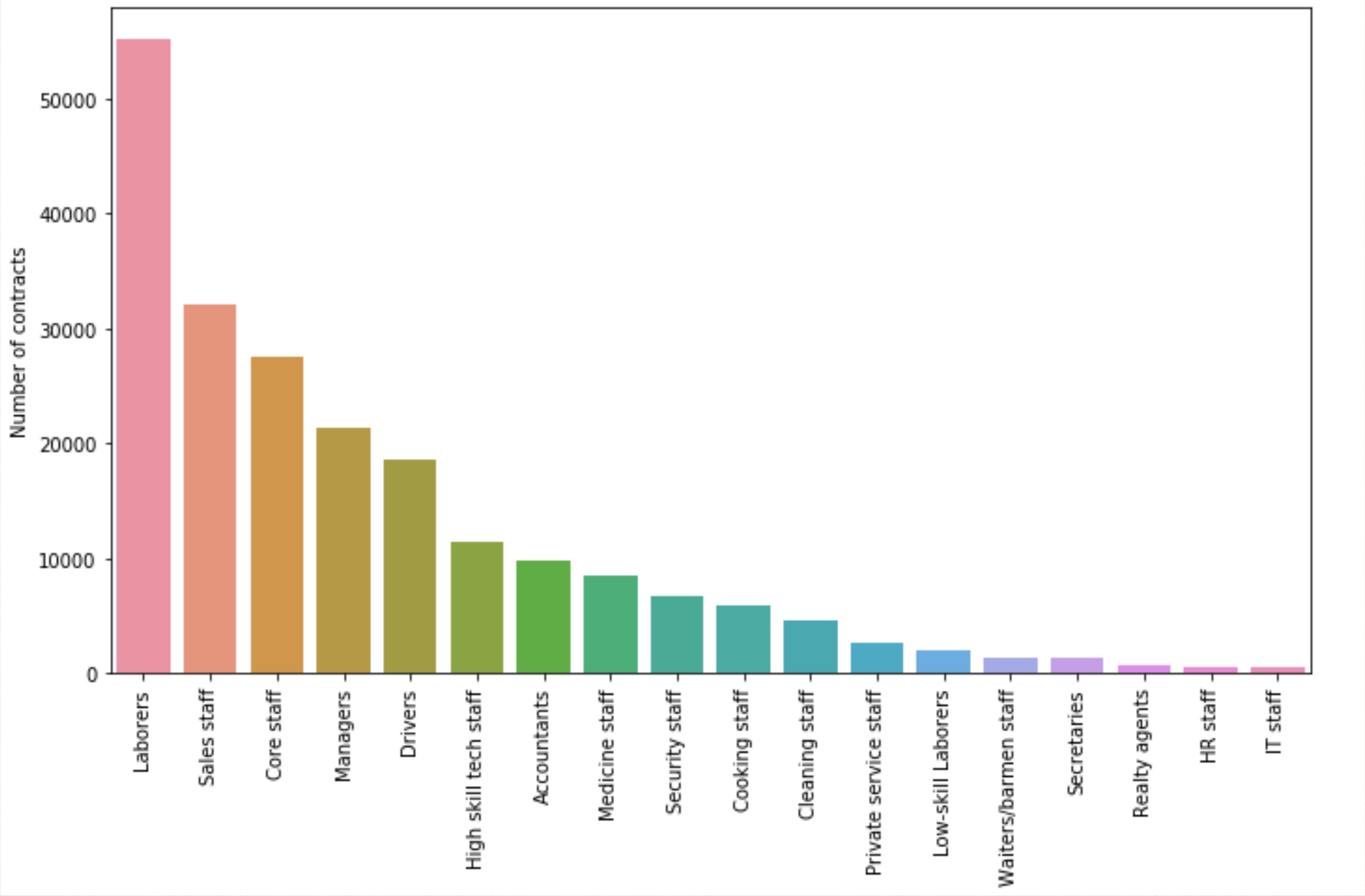


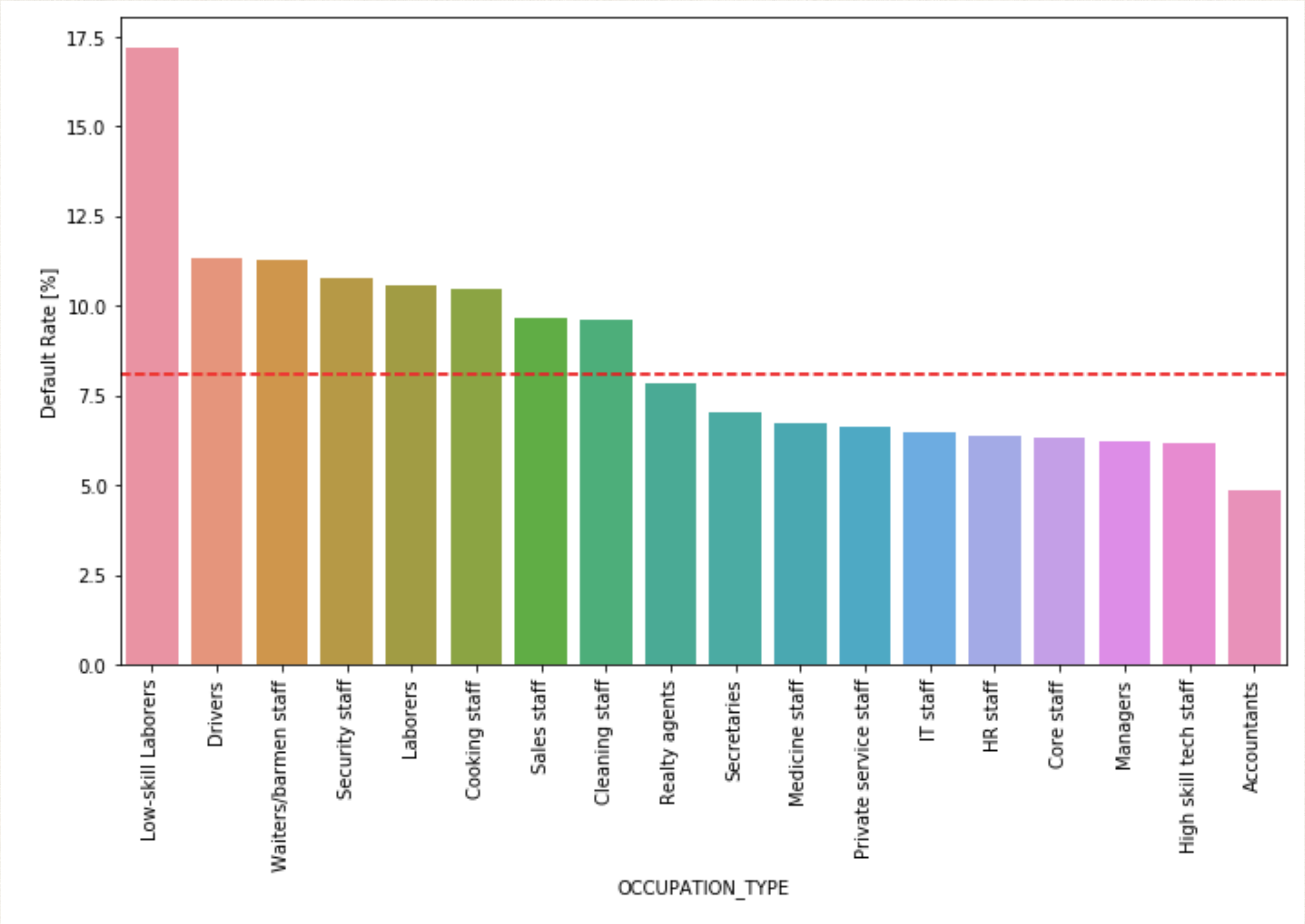
TARGET VARIABLE



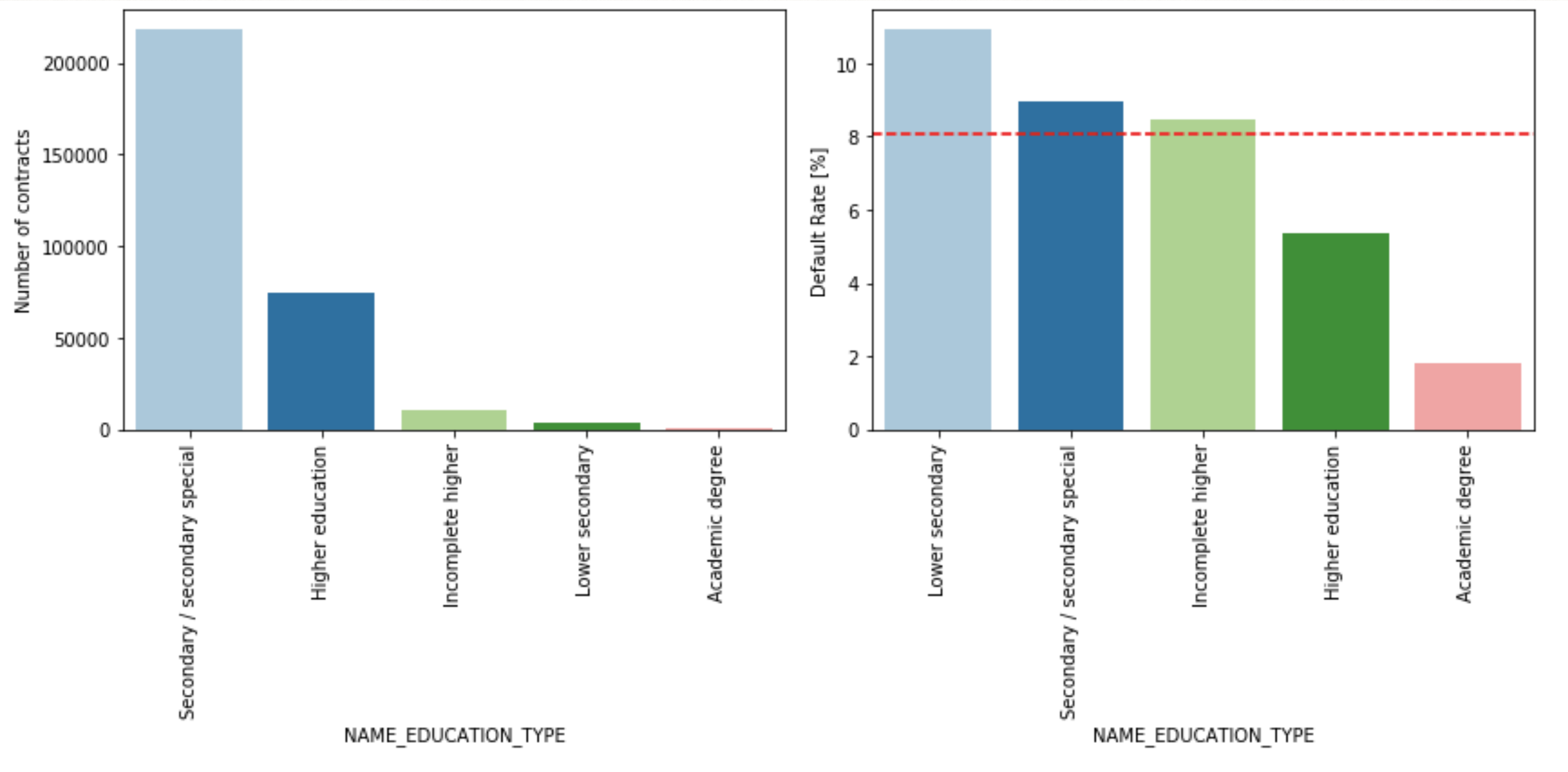
The bulk of the loans are Cash loans. Despite that, Revolving loans tend to have a higher default rate!


160000
140000
120000
100000
80000
60000
40000
20000
0
Number of contracts
Working
Commercial associate
Pensioner
State servant
NAME_INCOME_TYPE
10
8
6
4
2
0
Default Rate [%]
Working
Commercial associate
State servant
Pensioner
NAME_INCOME_TYPE

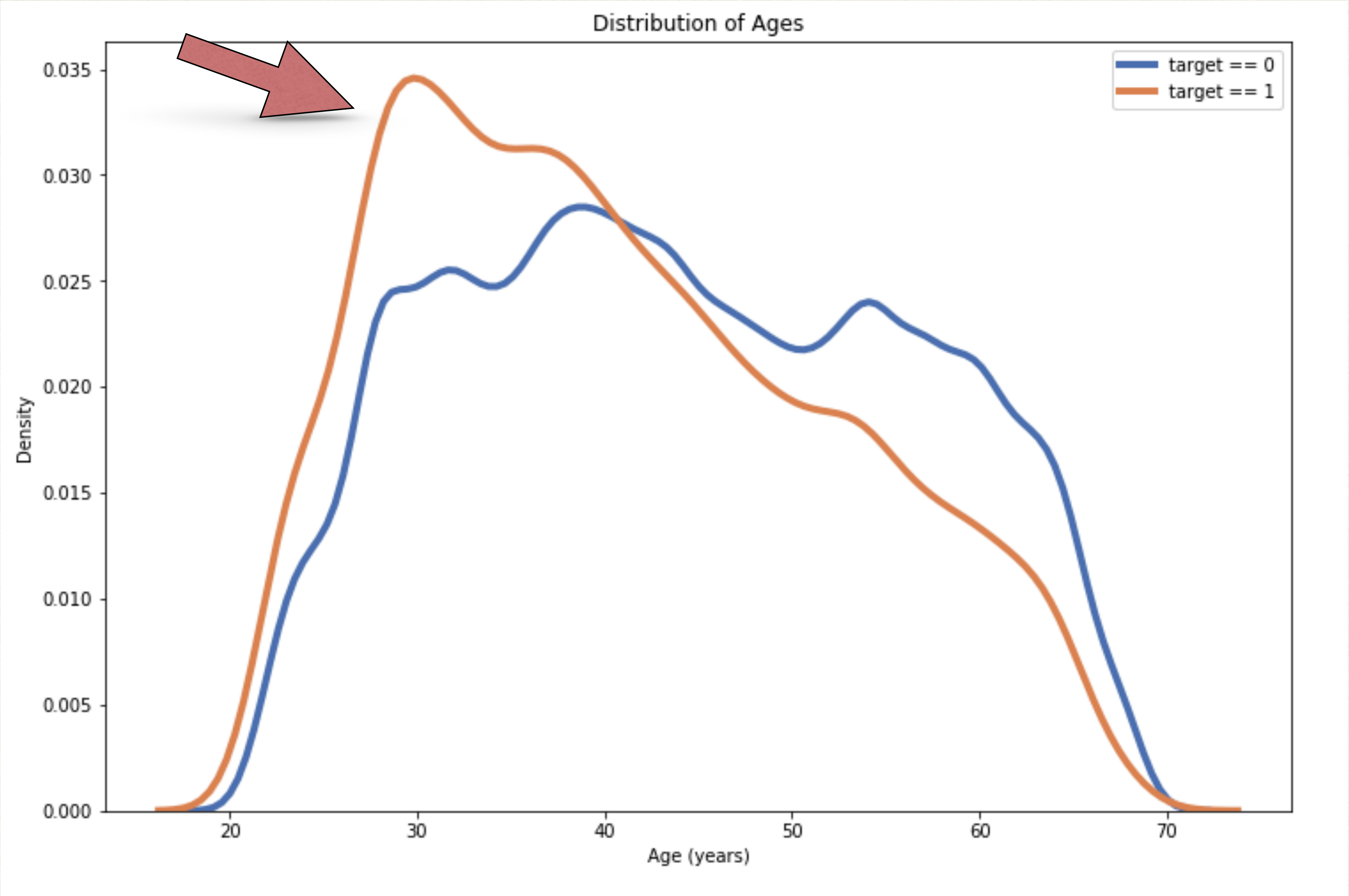

50000
40000
30000
20000
10000
0
Number of contracts
Laborers
Sales staff
Core staff
Managers
Drivers
High skill tech staff
Accountants
Medicine staff
Security staff
Cooking staff
Cleaning staff
Private service staff
Low-skill Laborers
Waiters/barmen staff
Secretaries
Realty agents
HR staff
IT staff



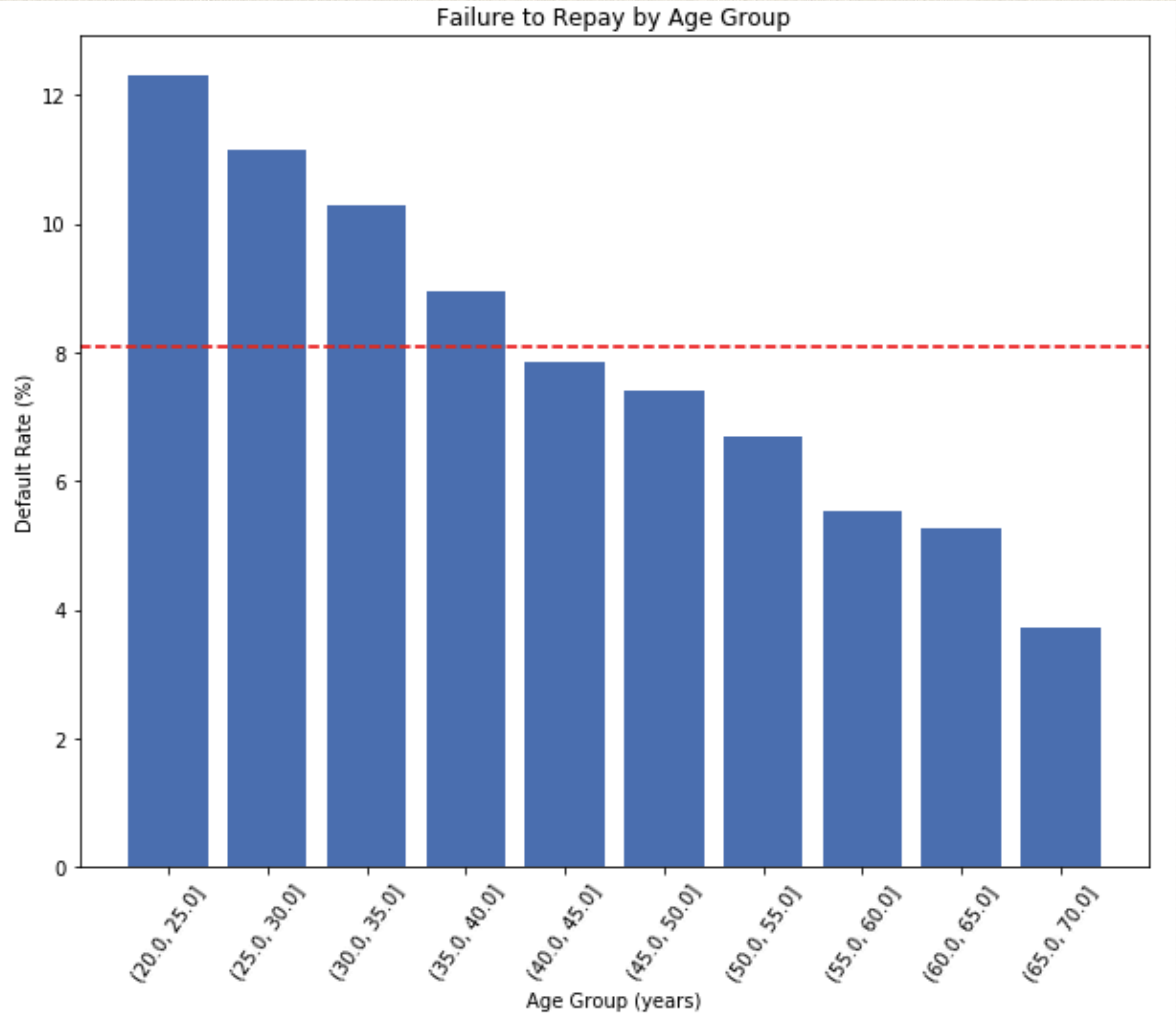
HOME CREDIT DEFAULT RISK



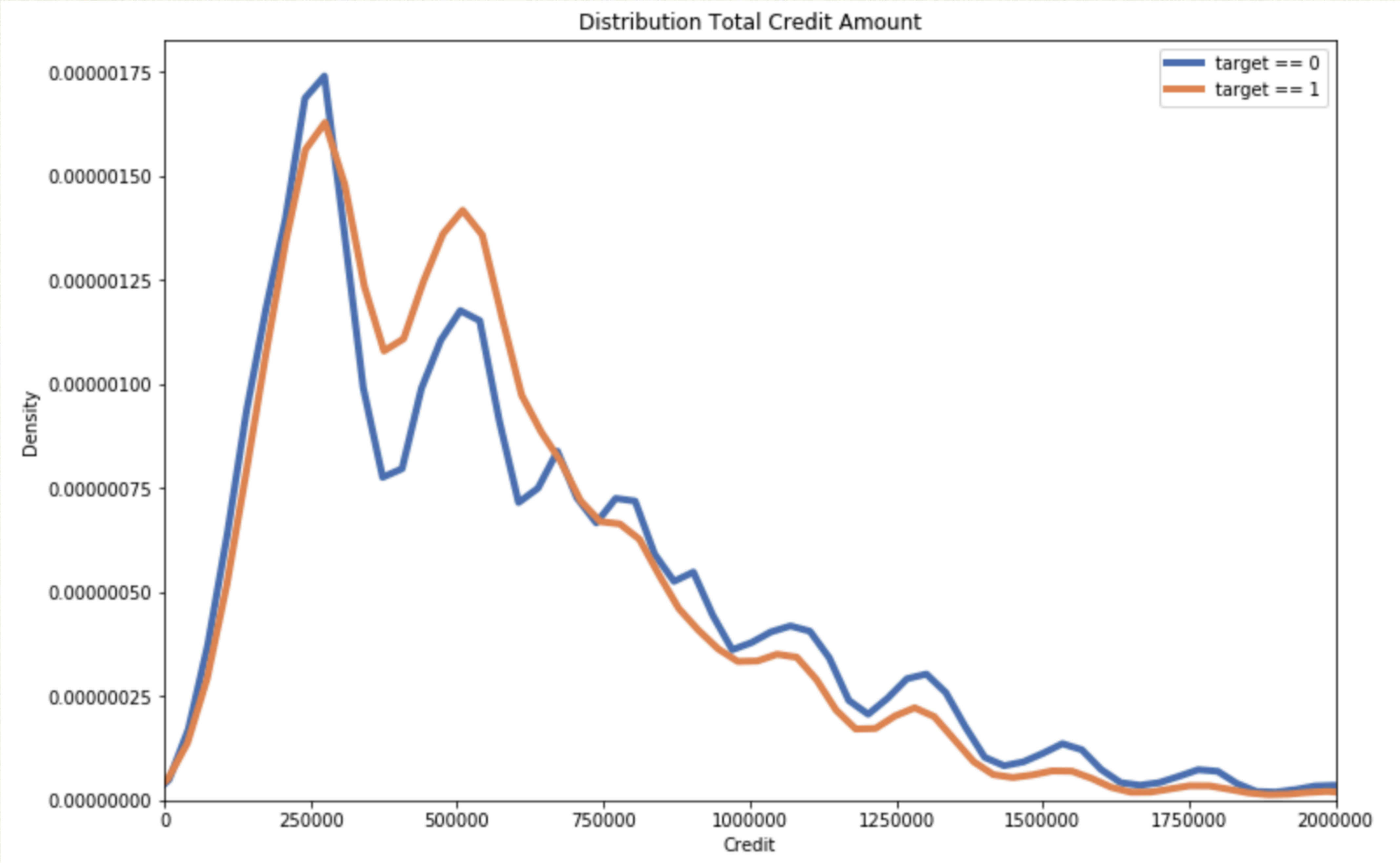
Most customers have secondary school as education. Lower secondary have the highest default rate.



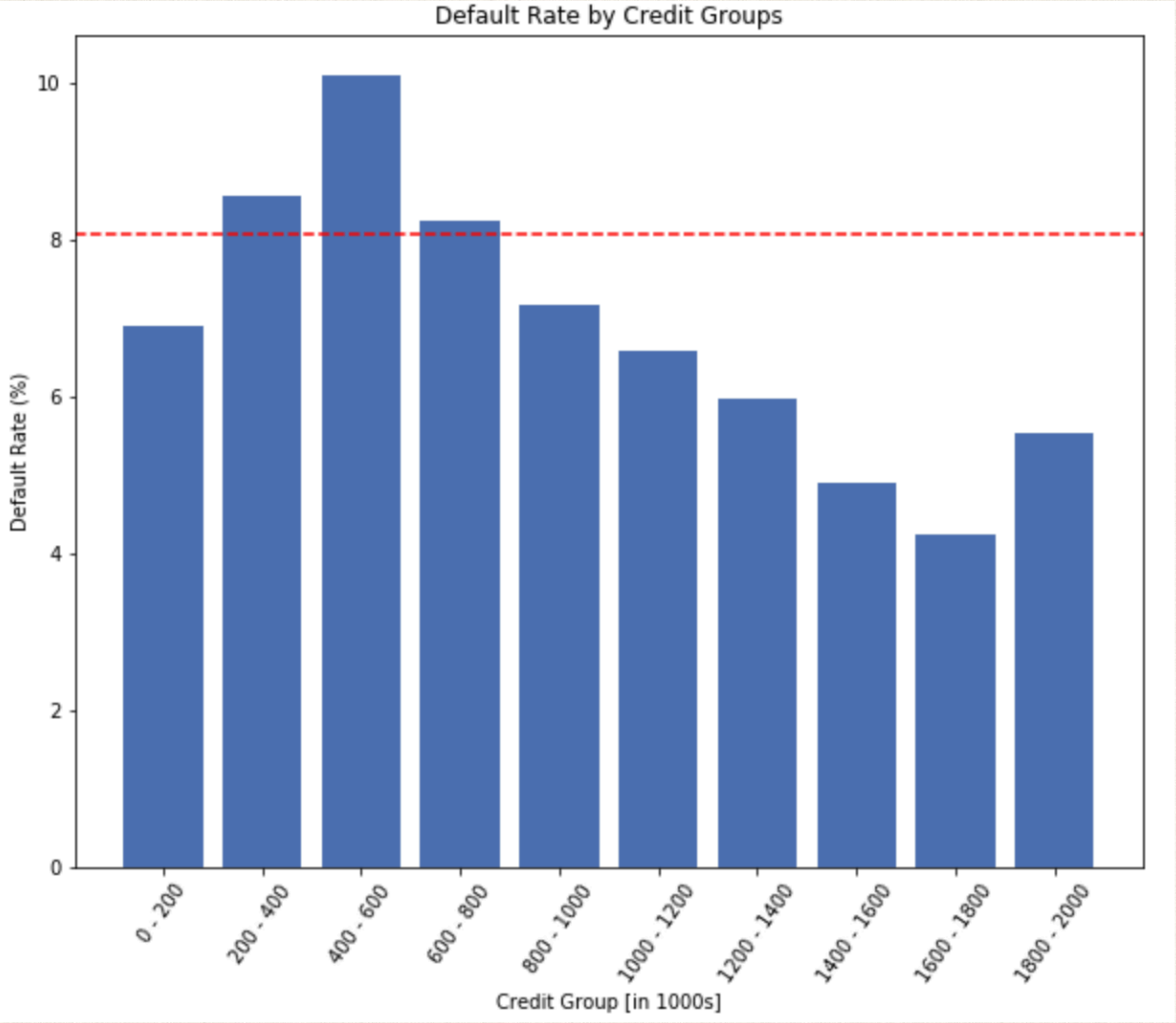
HOME CREDIT DEFAULT RISK



Younger people between 20 to 35 years tend to have a higher default rate than older people.

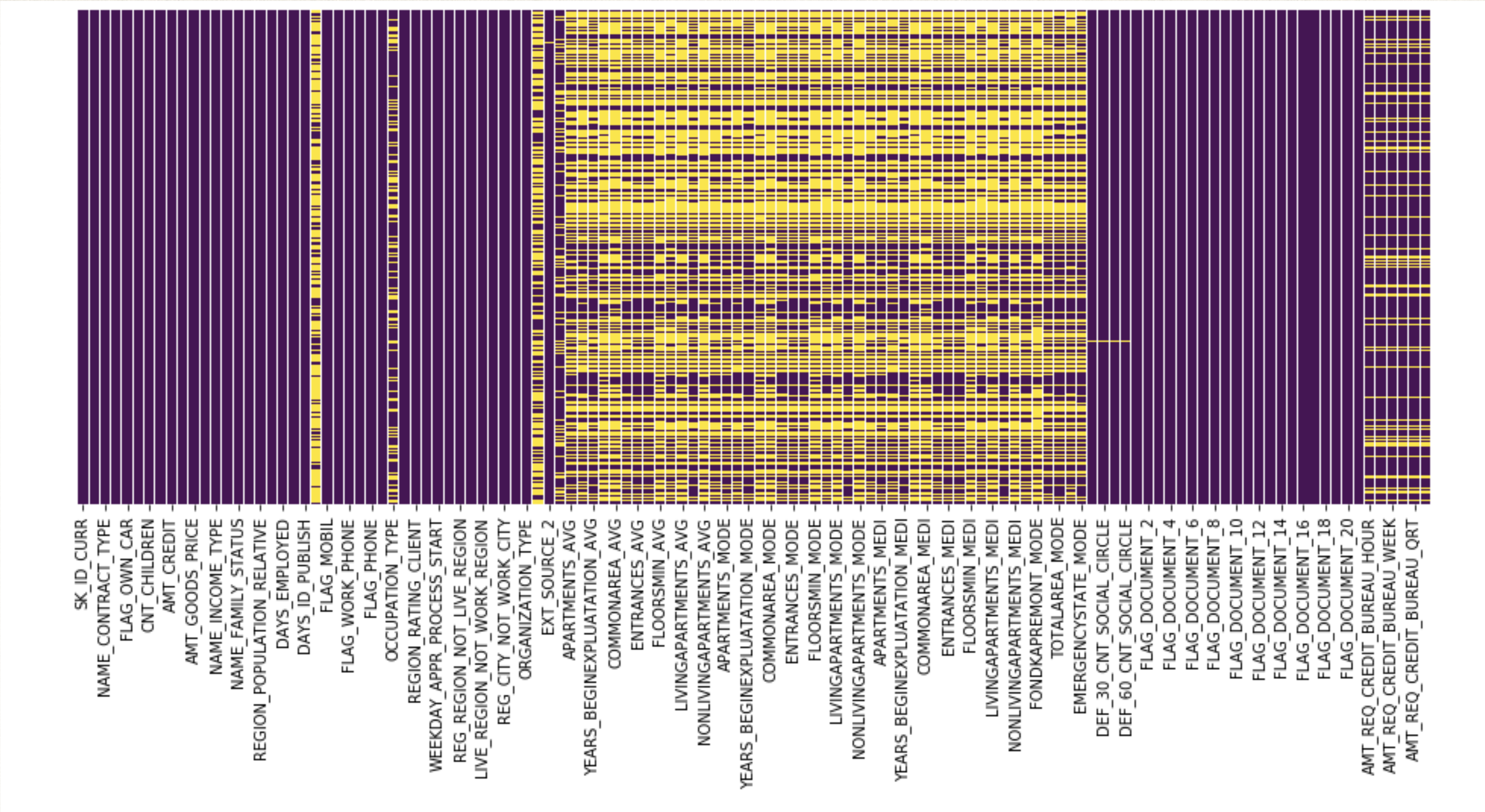


Distribution Total Credit Amount
target == 0
target == 1
0.00000175
0.00000150
0.00000125
0.00000100
0.00000075
0.00000050
0.00000025
0.00000000
Density
0
250000
500000
750000
1000000
1250000
1500000
1750000
2000000
Credit


Default Rate by Credit Groups
Default Rate (%)
10
8
6
4
2
0
0 - 200
200 - 400
400 - 600
600 - 800
800 - 1000
1000 - 1200
1200 - 1400
1400 - 1600
1600 - 1800
1800 - 2000
Credit Group [in 1000s]

PREDICTIVE MODELING

Missing Values



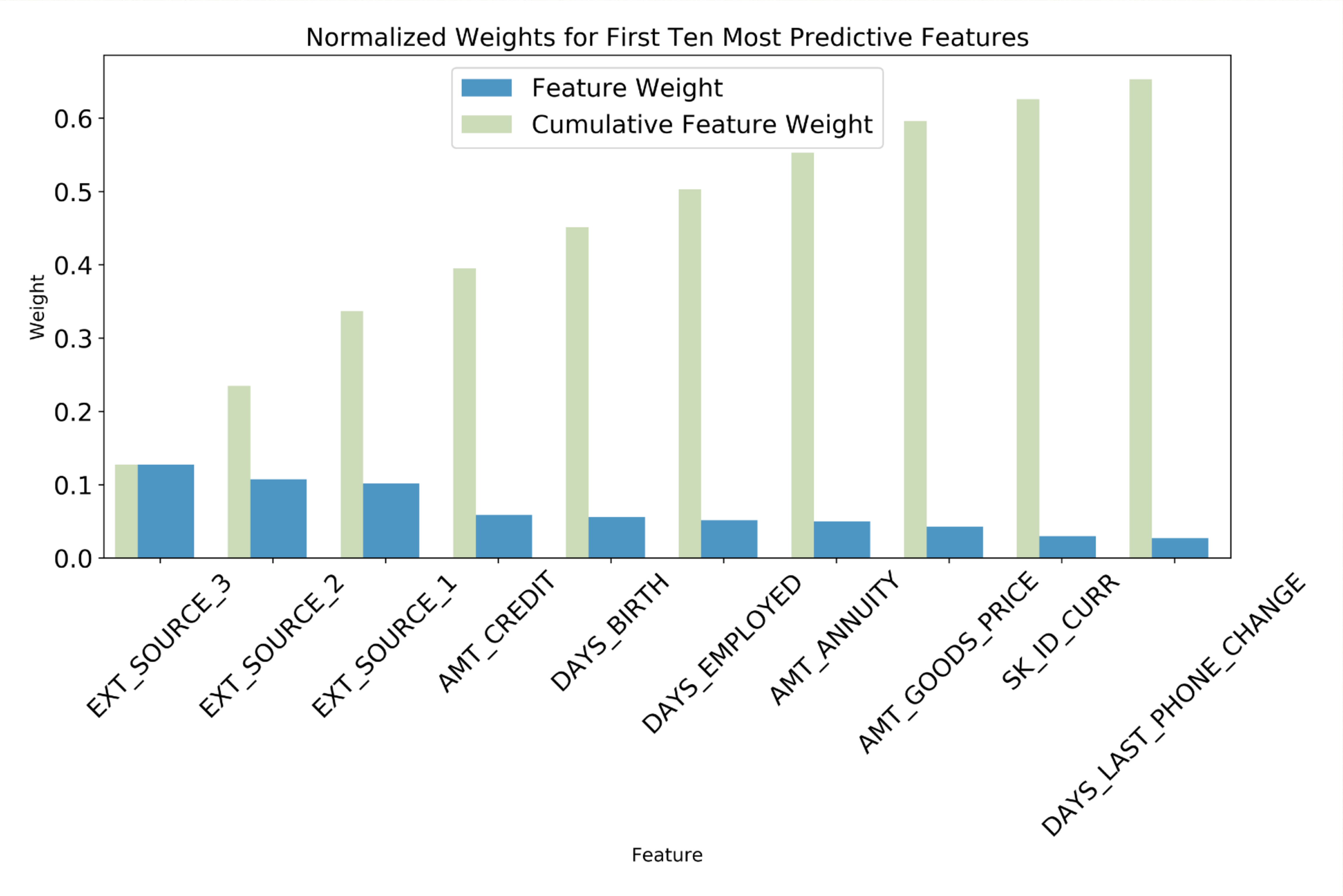
Overall missing values were high in the housing features! We dropped 47 unimportant Features.

FEATURE ENGINEERING

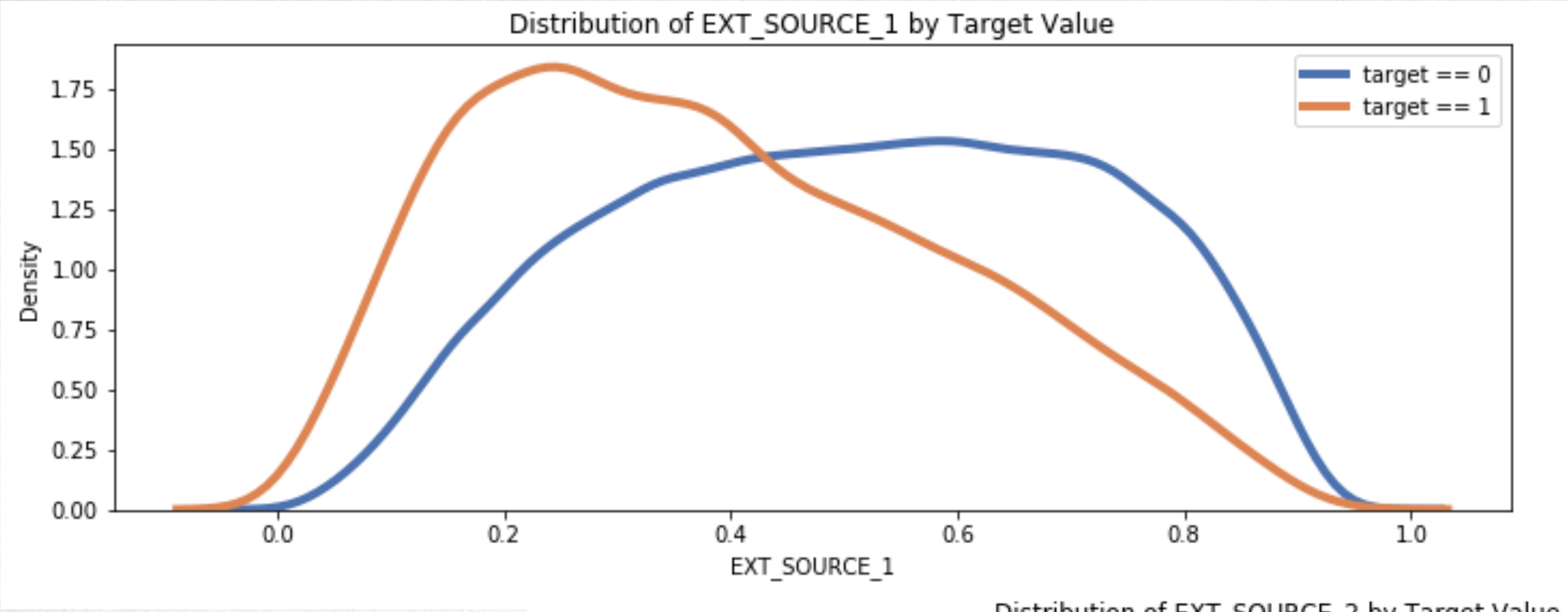
Dummies for Categorical Variables
Imputation with Median
Scalar MinMax
Balancing

	RECALL	PRECISION	F1-SCORE	ROC	TIME
XG Boost	67.3 +- 0.4%	68.1 +- 0.4%	67.7 +- 0.4%	0,765	21 s
ADA Boost	67.2 +- 0.1%	68.4 +- 0.9%	67.8 +- 0.4%	0,756	7 s
Random Forest	66.2 +- 0.9%	67.2 +- 0.5%	66.7 +- 0.4%	0,745	197 s 11 s
Logistic Reg	66.8 +- 0.8%	67.9 +- 0.7%	67.4 +- 0.7%	0,747	1.4 s

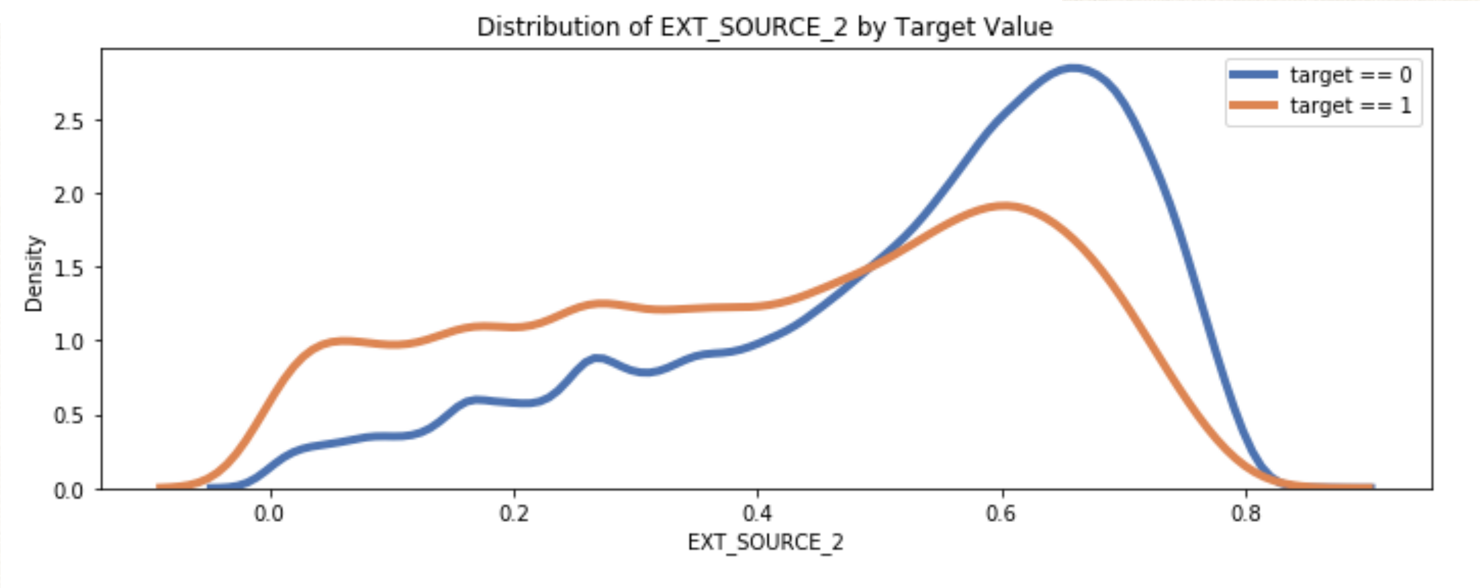
XG Boost



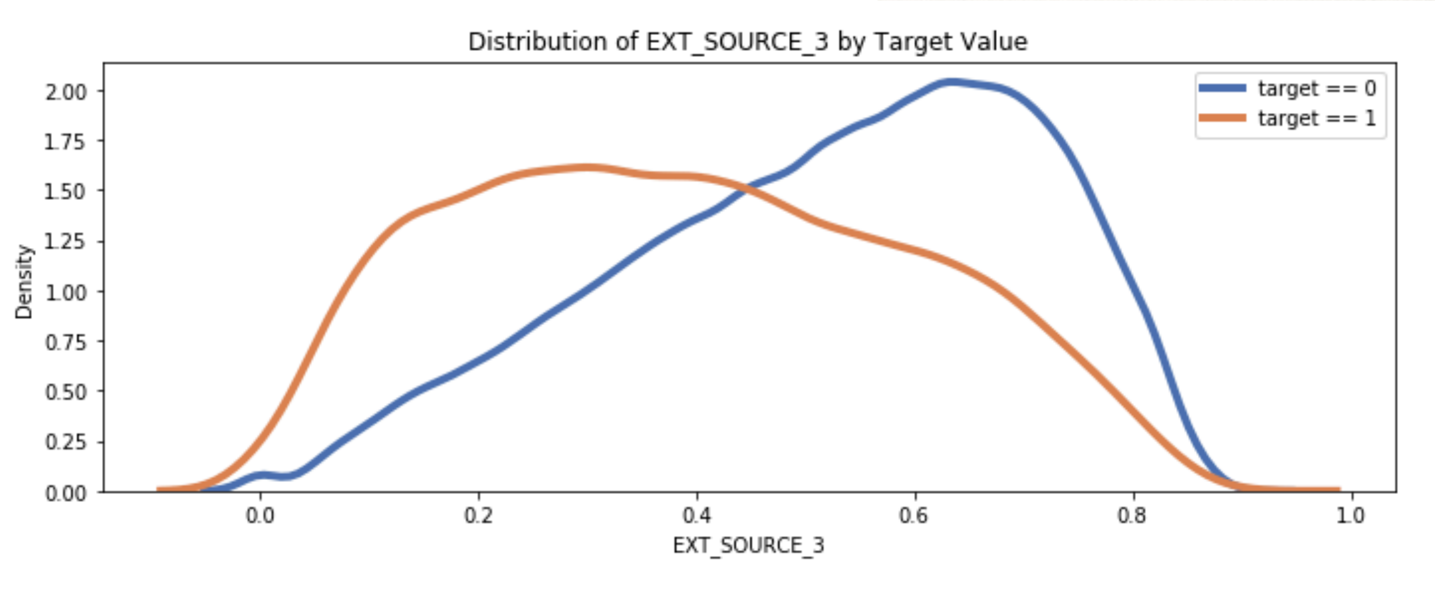
Normalized Weights for First Ten Most Predictive Features
Feature Weight
Cumulative Feature Weight
Weight
0.0
0.1
0.2
0.3
0.4
0.5
0.6
EXT_SOURCE_3
EXT_SOURCE_2
EXT_SOURCE_1
AMT_CREDIT
DAYS_BIRTH
DAYS_EMPLOYED
AMT_ANNUITY
AMT_GOODS_PRICE
SK_ID_CURR
DAYS_LAST_PHONE_CHANGE
Feature



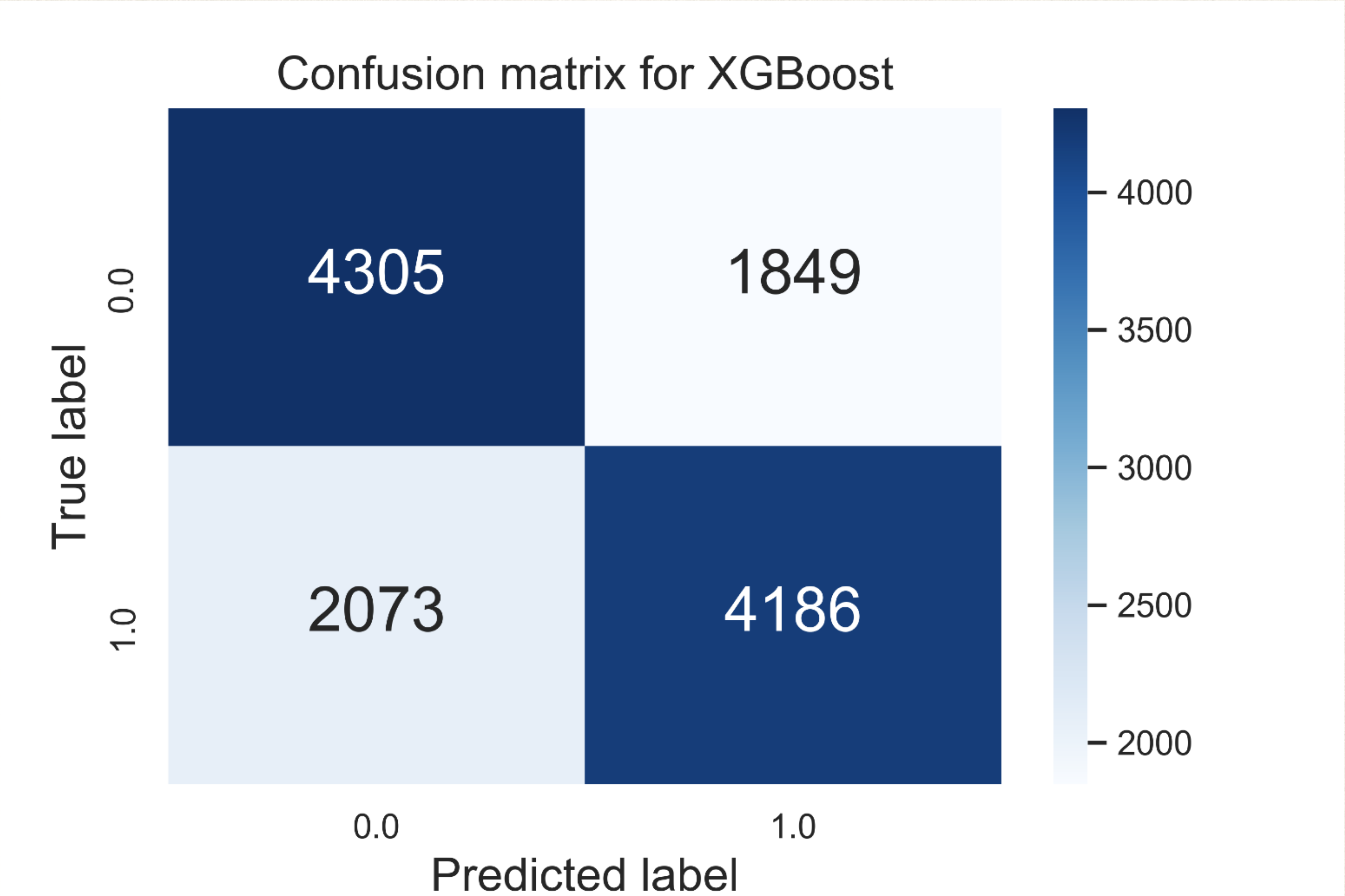
Distribution of EXT_SOURCE_1 by Target Value
target == 0
target == 1
Density
EXT_SOURCE_1



Distribution of EXT_SOURCE_2 by Target Value
target == 0
target == 1
Density
EXT_SOURCE_2

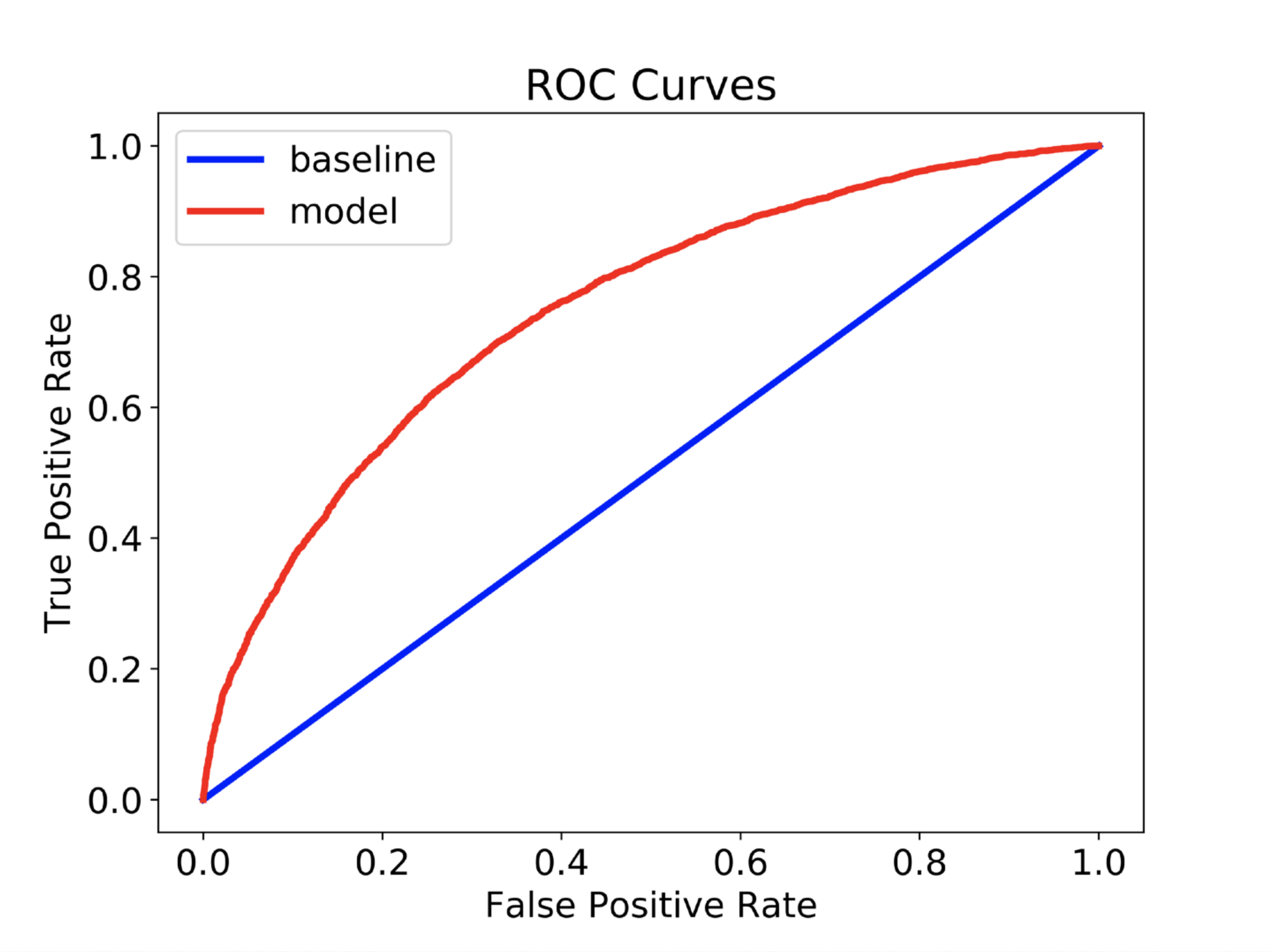


Distribution of EXT_SOURCE_3 by Target Value
target == 0
target == 1
Density
EXT_SOURCE_3



Confusion matrix for XGBoost
4305
1849
2073
4186
True label
0.0
1.0
0.0
1.0
Predicted label
4000
3500
3000
2500
2000

XG Boost


ROC Curves
baseline
model
1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate

BUSINESS RECOMMENDATIONS

1. **Extended Search for external Sources (EXT_SRC)**
2. **Stricter Validation-Checks of younger Customers**
3. **Higher Attention for Credit amounts around average**

FUTURE WORK

- Use Principal Component Analysis
- Use LightGBM (Parameter Tuning)
- Include other relevant datasets (Bureau, …)
- Analysis of all countries/markets
- Analysis of default rates for different credit types
- DAYS_EMPLOYED: fill Values with MAX, Pensioners

„Thank you for your Attention!“