

Credit Card Fraud Analysis

Zakra Chachar
Daniela Jaimes
Yohanes Nuwara

Fundamentals of Statistics



February 2023

Project

This project utilizes credit card data to analyze patterns linked to loan default risk. Our goal is to develop borrower profiles that can inform responsible lending decisions in the financial sector. Additionally, this analysis may offer valuable insights into the city's economic landscape, potentially allowing authorities to tailor social safety net programs and address potential financial risks.

Tools

We employed a diverse toolkit for this analysis, including:

- Data manipulation tools like **Excel** and **Azure**
- Programming languages like **Python** and **R**
- Machine learning platforms like **Orange**

300,000

Original dataset

Dataset obtained in
kaggle.com

15,000

Sample

Randomly drawn sample
used for running analysis

50%

Defaulters

Our sample consisted in 50%
defaulters and 50% non
defaulters.

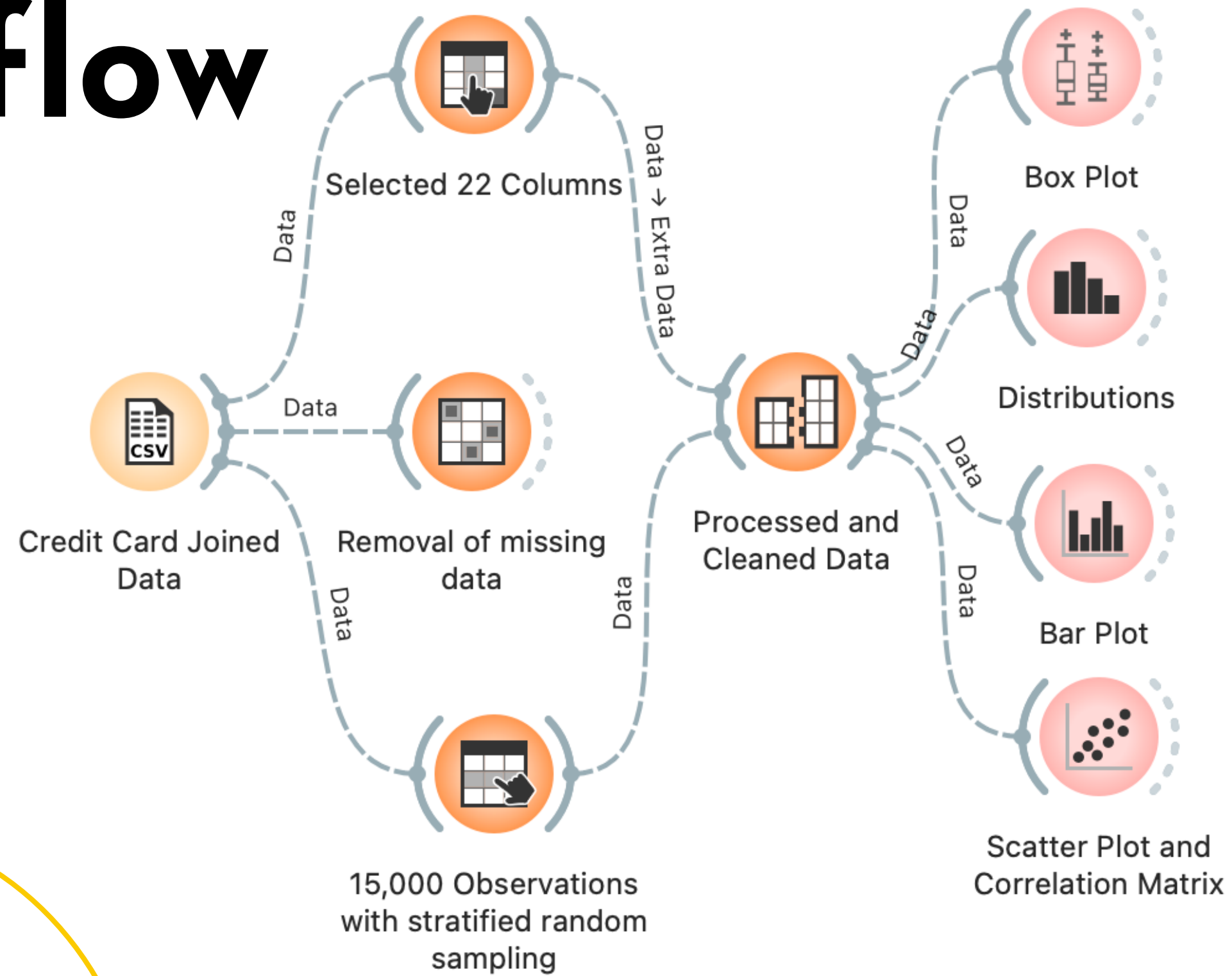
Research Questions

Bank X is experiencing its customers who has applied and used its credit cards doing suspicious fraudulent activity / defaulting. Our task as data analyst team is to perform statistical analysis to answer the following questions



1. Are there any differences in variables that can help us distinguish those customers who are defaulting and those who are not?
2. Are there any correlations between variables such as credit amount with other variables?
3. Can we identify groups of customers with similar characteristics?
4. Which loan type is most likely to default?
5. Are females more likely to default than males?
6. Are low-income groups more likely to default than higher-income groups?
7. Do variables such as age affect the amount of credit?

Workflow

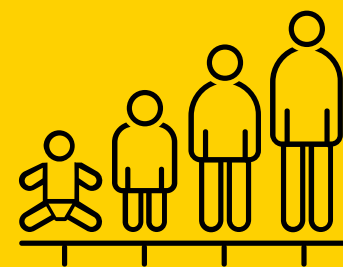


Key Observations



Contract type

Revolving loans have a lower risk of default.



Age

Young clients are more risky.
Most defaulters are in their 20's and 30's.



Gender

Men exhibiting a higher default rate compared to women.



Family Status

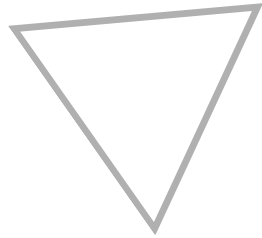
Married people exhibit lower tendency to default.

Answering research questions:

Question 1: Are there any differences in variables that can help us distinguish those customers who are defaulting and those who are not?

Identifying Profiles





Profiles prone to default

Table showcasing the characteristics associated with high and medium risk borrower profiles.

	Medium Risk	High Risk
AGE GROUP	30-39	20-29
GENDER	Women	Men
OCCUPATION	Sales Staff, Drivers	Laborers
EDUCATION	Secondary / Secondary special	Lower secondary
FAMILY STATUS	Civil marriage	Single
HOUSING TYPE	Municipal apartment	Rented apartment / With parents

Ideal Borrower

The analysis suggests a potential profile for a "perfect borrower". Based on this ideal borrower, personnel can easily identify what clients have higher probabilities of defaulting.



45-50 YEARS OLD

REVOLVING LOAN

MARRIED

\$160K - \$180K

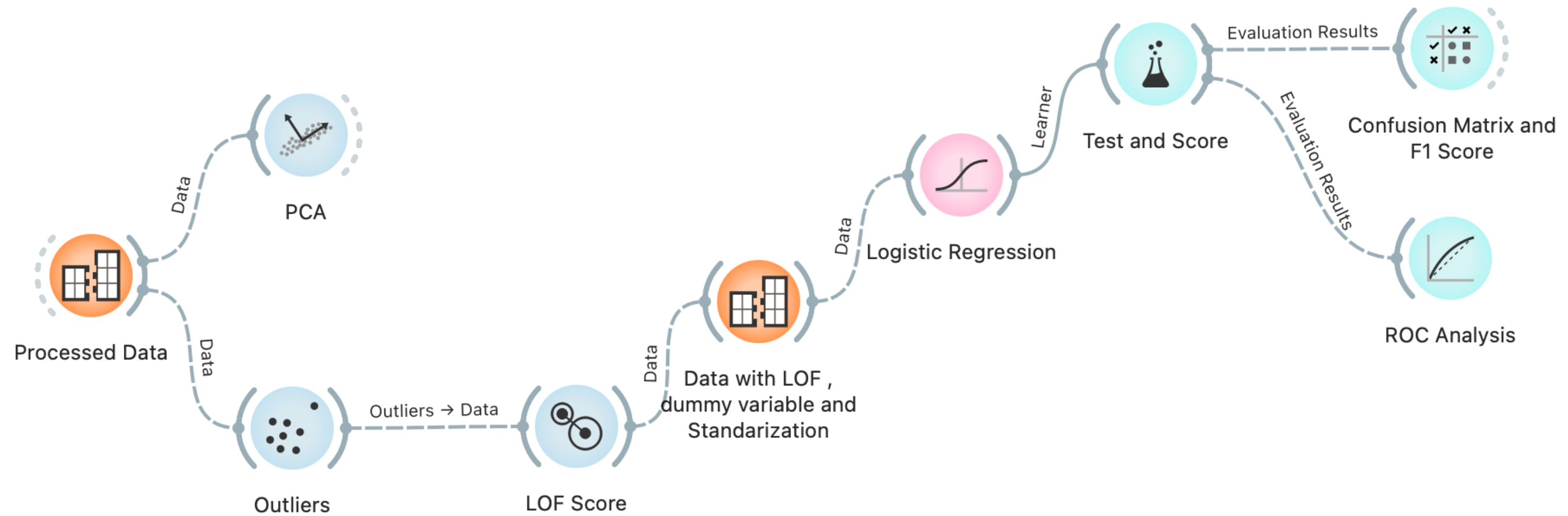
STABLE HOUSING SITUATION:

- Homeowner, ideally owning more than one property.
- Resides in their own home.
- Permanent address match contact address.

PROFESSIONAL STABILITY AND EDUCATION:

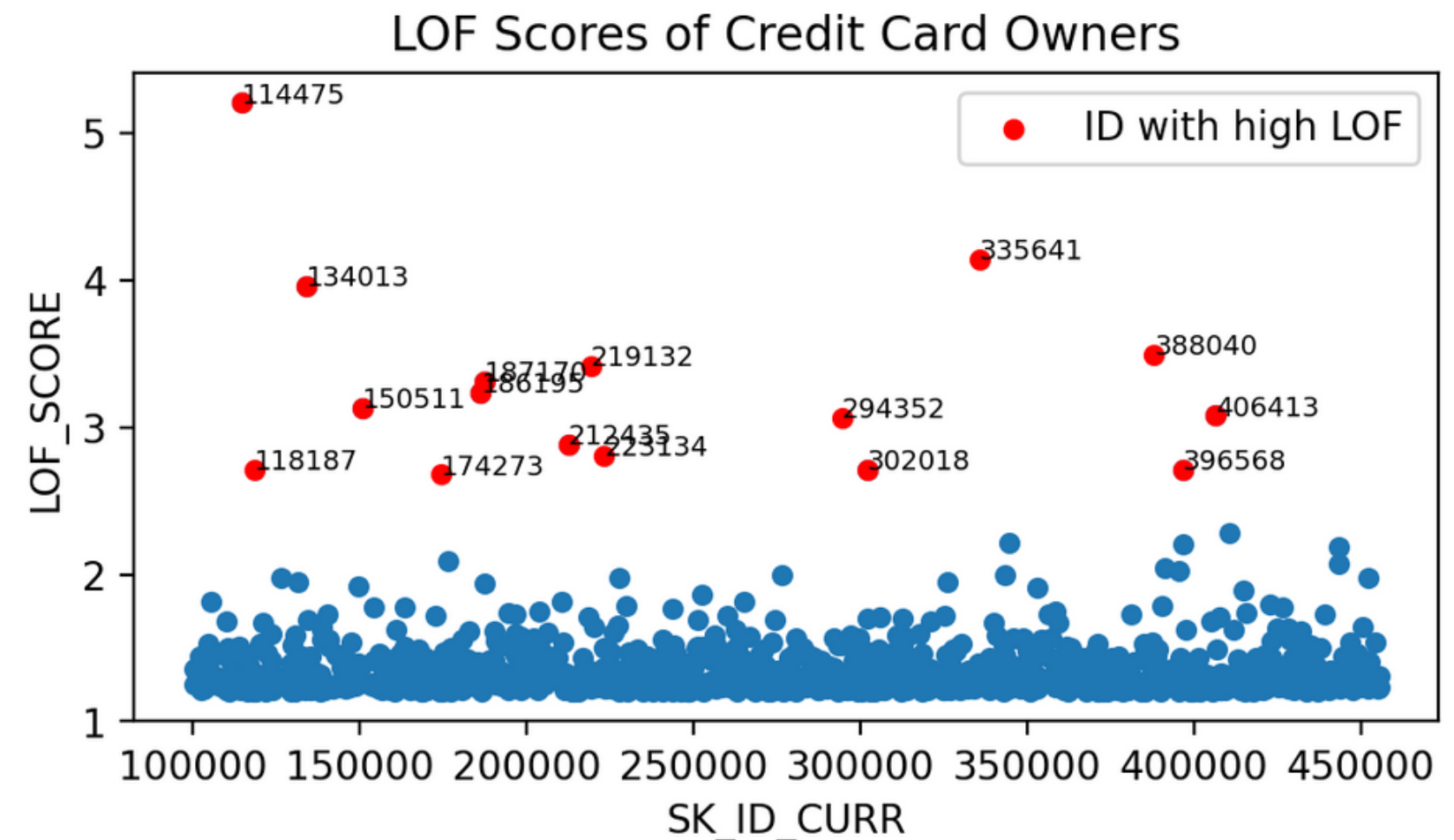
- Holds a completed higher education degree.
- Core staff member position within a business entity.

Workflow



Outliers

- Fraud credit card owners tend to have high Local Outlier Factor (LOF) scores
- We sample 5 credit card owners identified as outlier and Fraud and found information that does not make sense
- The income of the manager (USD 135K) is lower than the driver (USD 225K)
- 2 owners have extremely high income (USD 117M)
- The core staff in Kindergarten has higher income (USD 1.3M) than average Kindergarten staff is

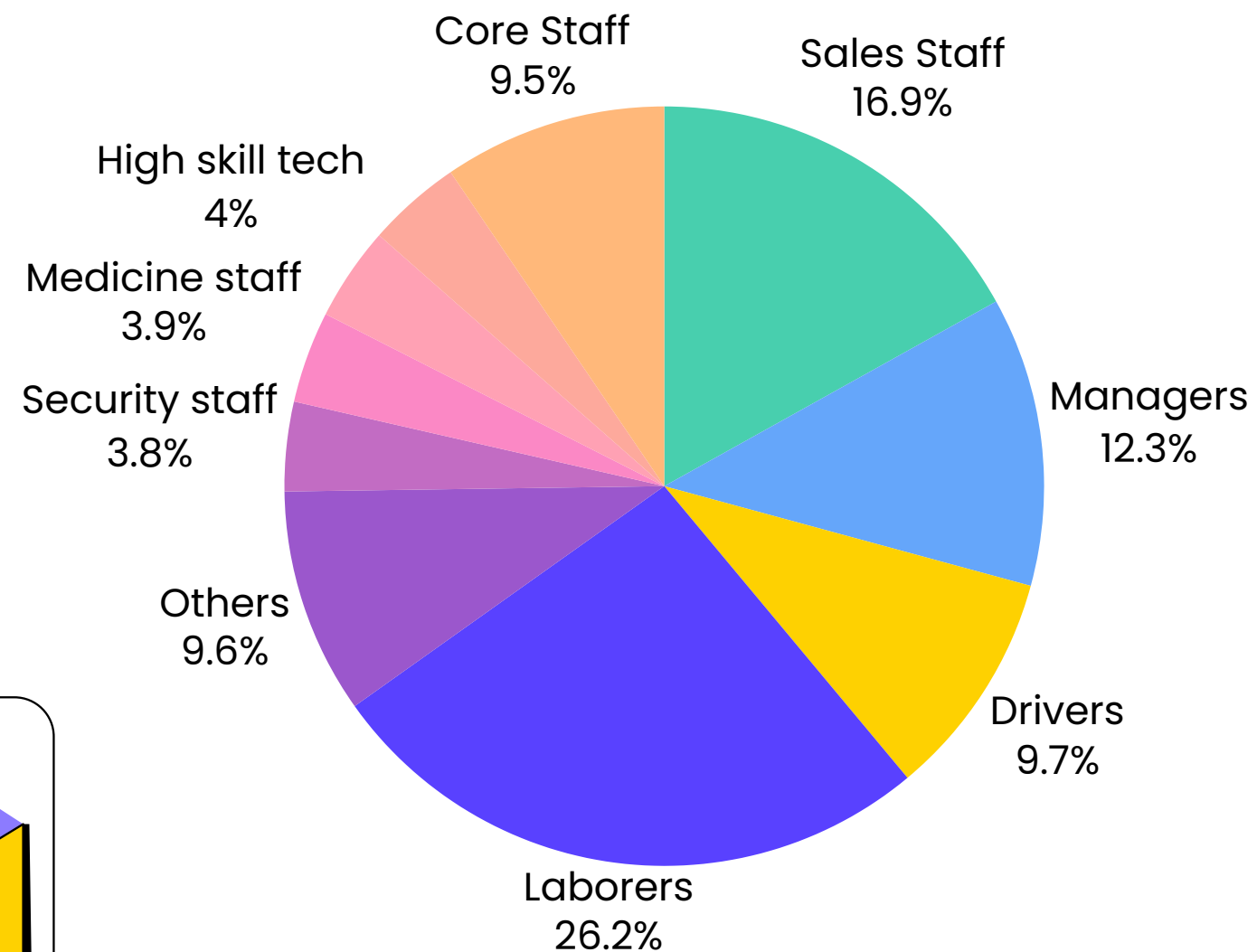


SK ID	GENDER	INCOME TOTAL	OCCUPATION TYPE	ORGANIZATION TYPE	LOF
174273	M	135000	Managers	Business Entity	2.678
335641	M	225000	Drivers	Business Entity	4.145
114967	F	117000000	Laborers	Business Entity	216.32
114967	F	117000000	Laborers	Business Entity	216.31
219132	F	1305000	Core staff	Kindergarten	3.414

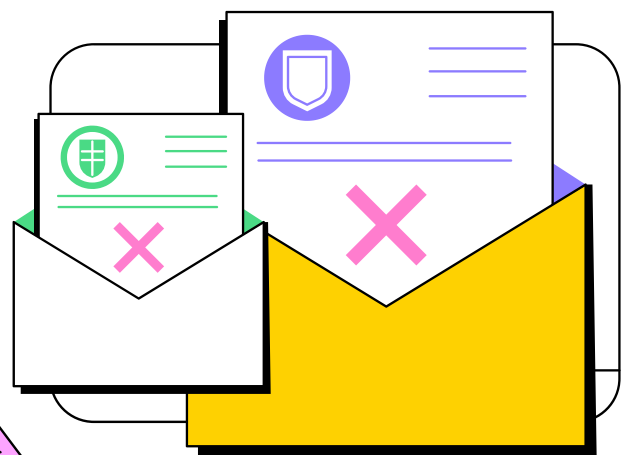
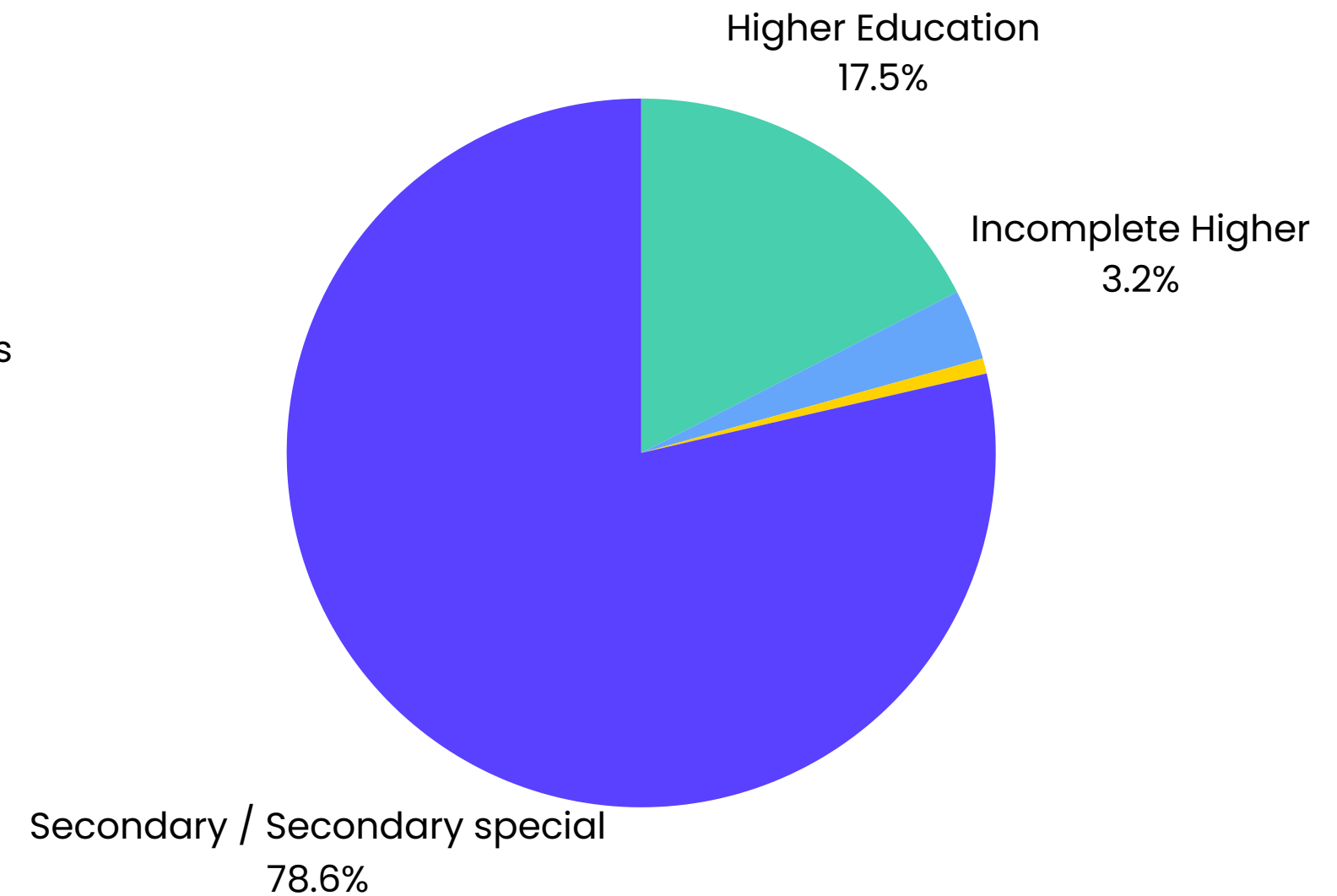
Outliers

Majority of the Fraud outliers are Laborers (26.2%) and having Secondary education (78.6%)
Income of some occupations of Fraud outliers are higher than normal ones, for example
Managers, Accountants, Waiters, Realty agents, and HR staffs

Occupations of Fraud Outliers

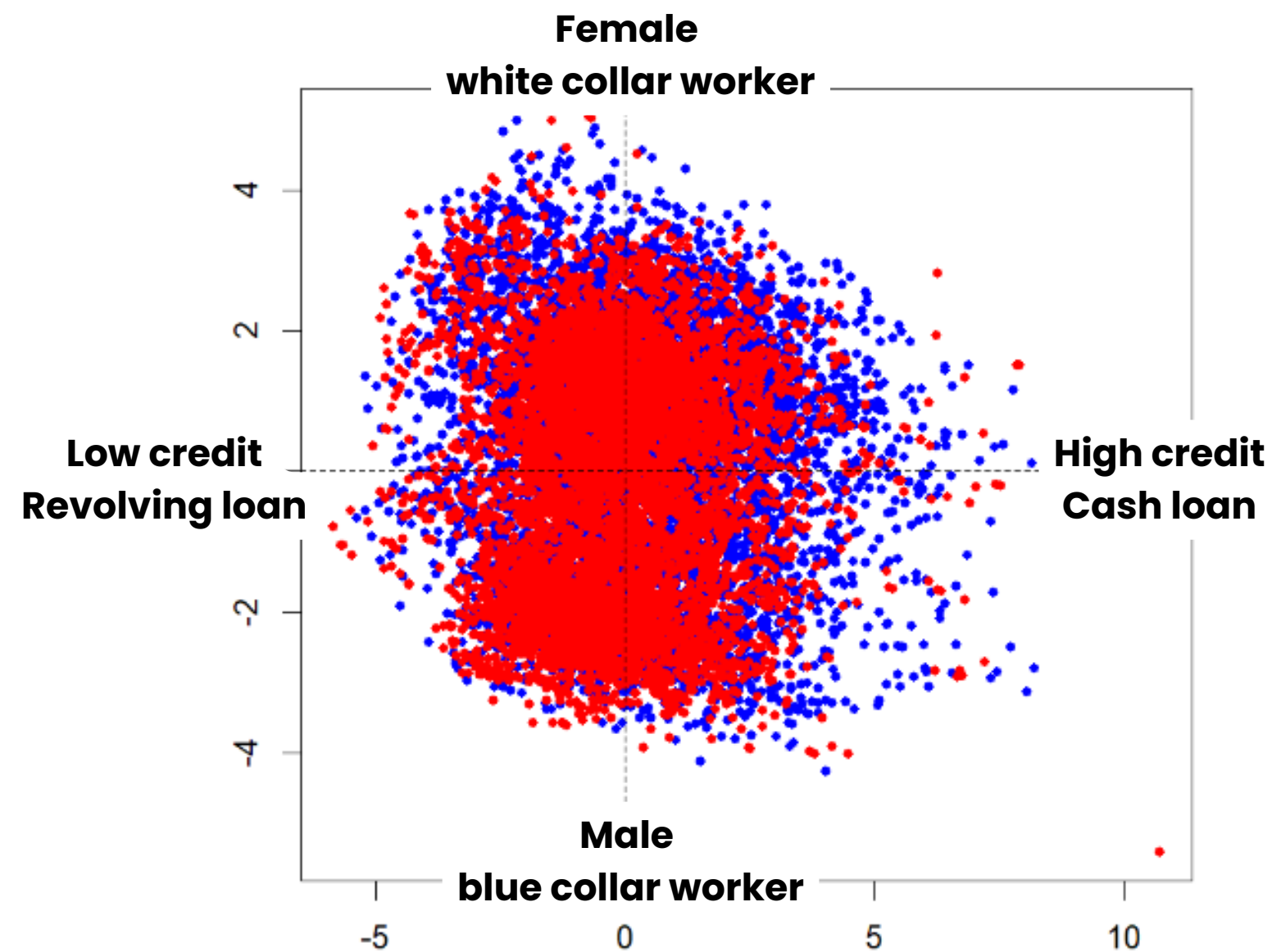


Education of Fraud Outliers

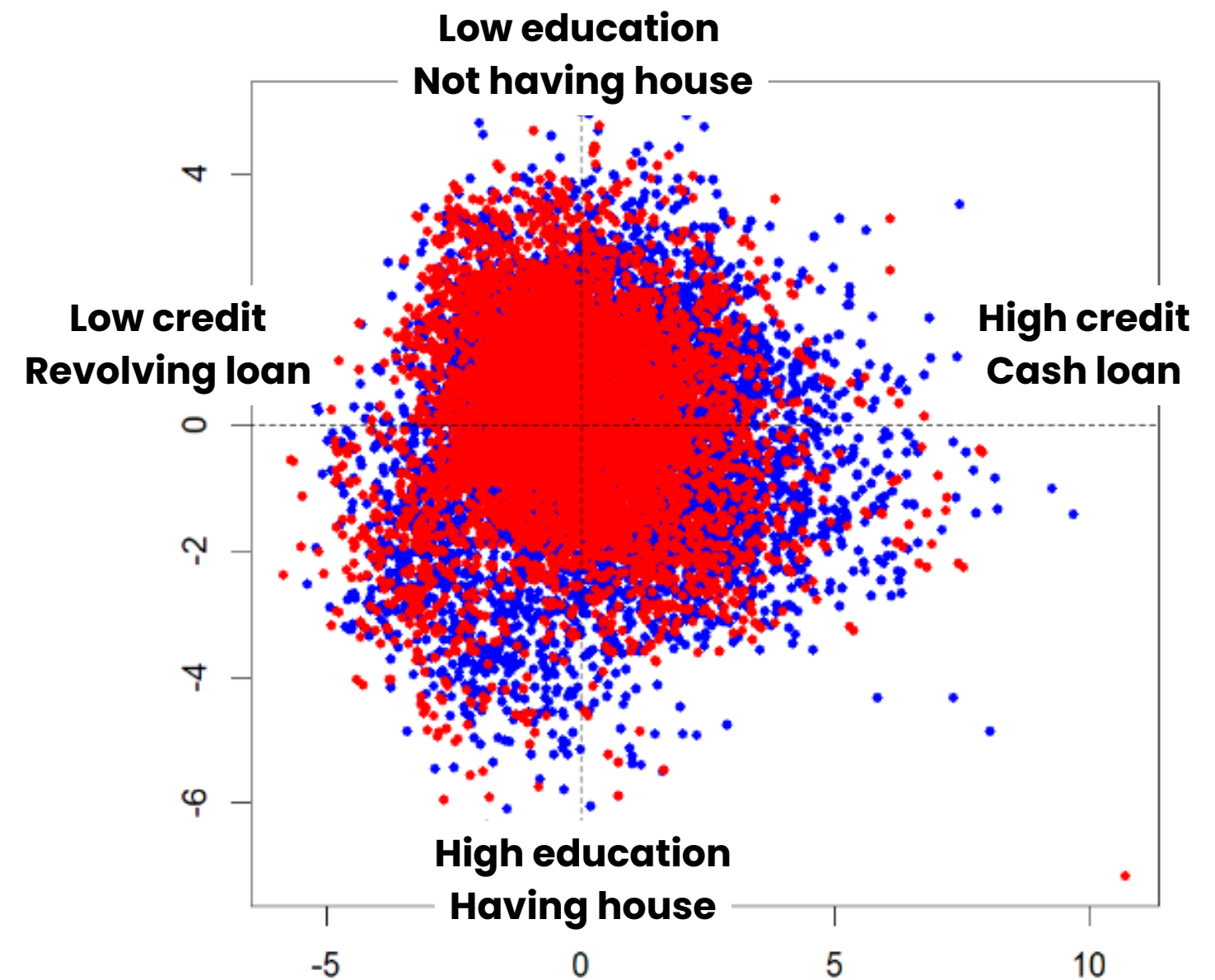


Group of Customers from PCA

We use PCA to identify clusters/groups of customers and few most important variables based on Loadings
Positive loading has characteristics (variables) opposite to negative loading



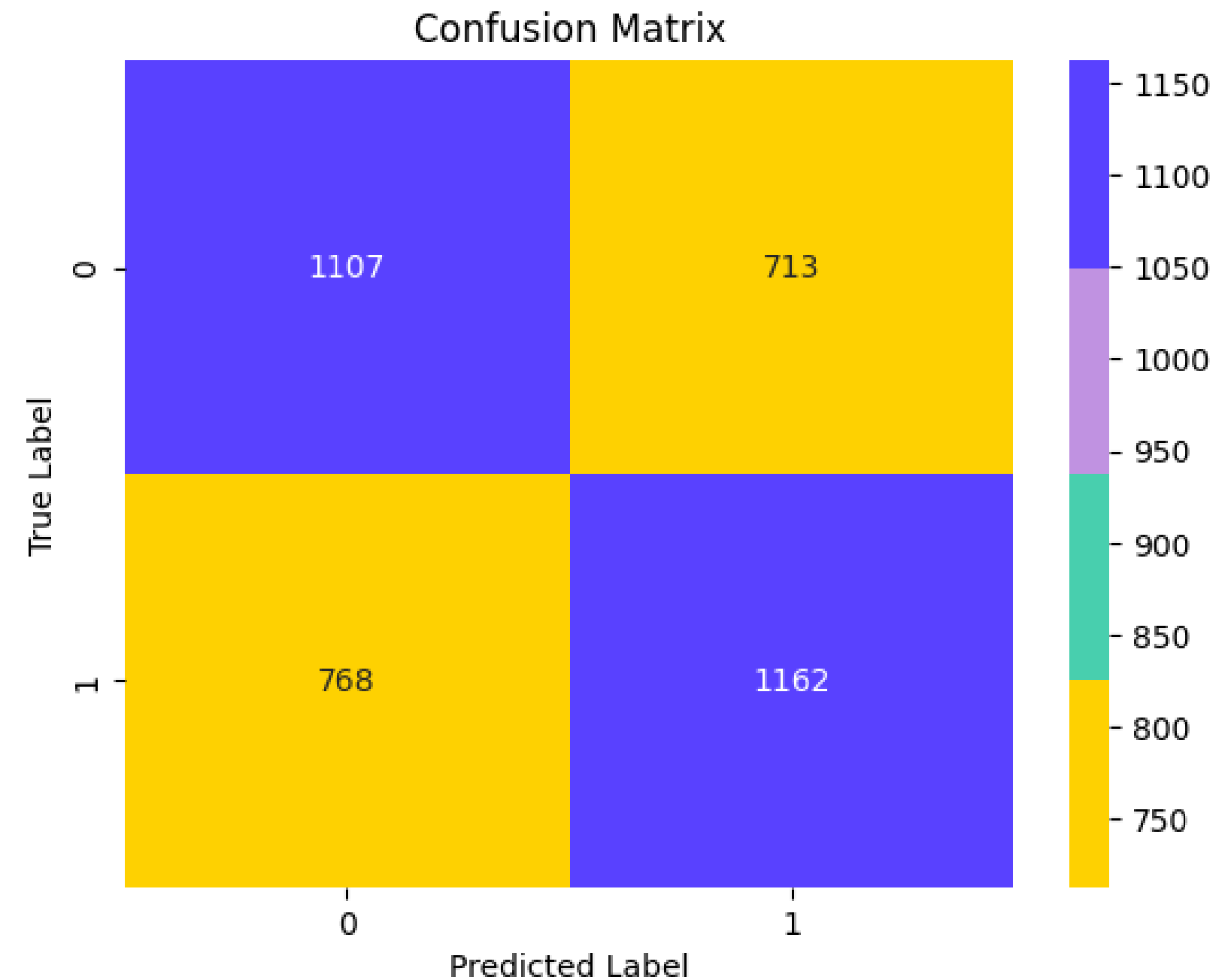
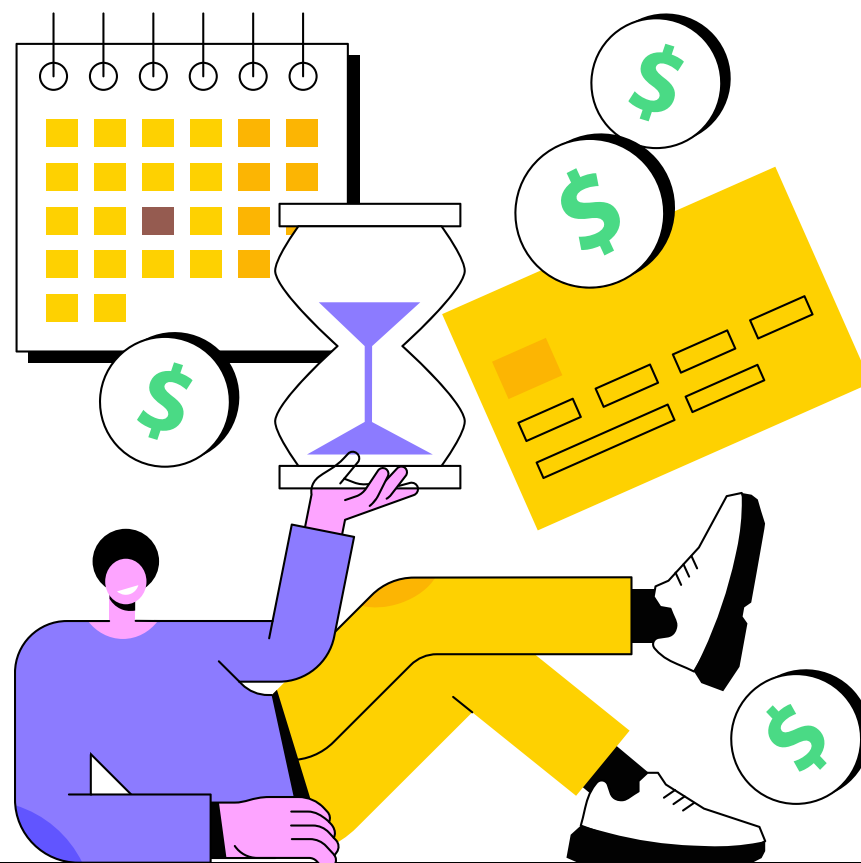
Biplot of PC1 and PC2



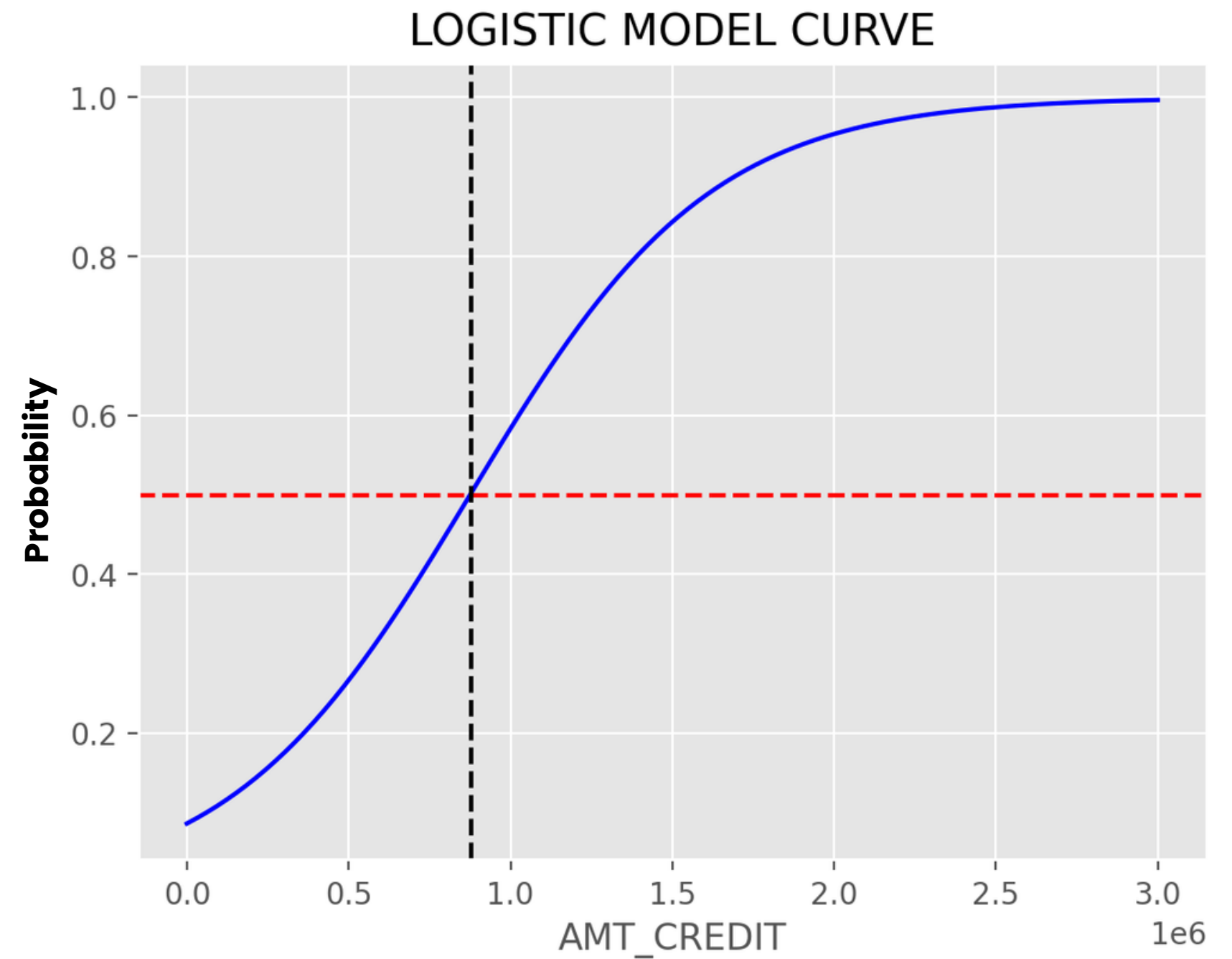
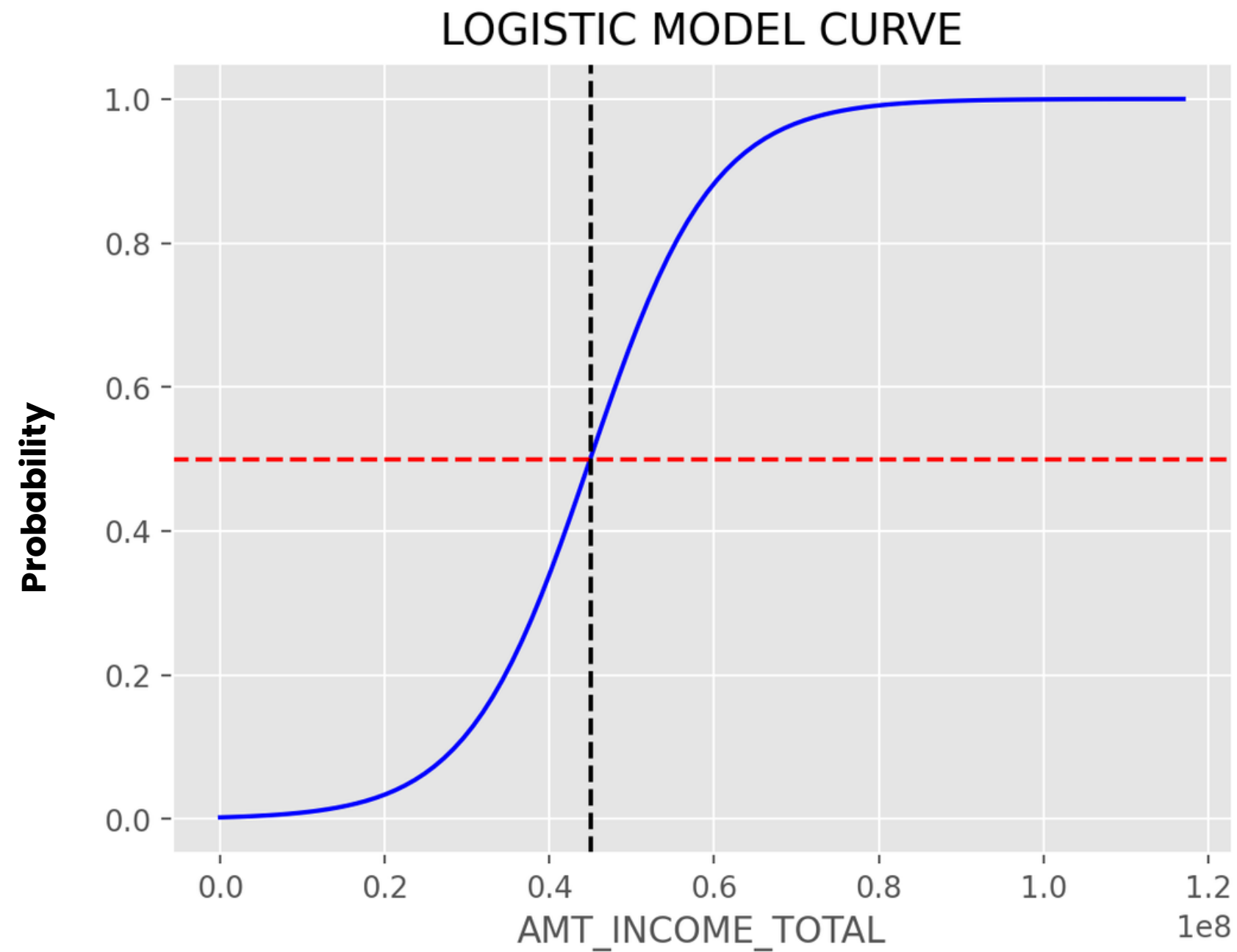
Biplot of PC1 and PC4

Logistic Regression

- Model Accuracy (F1): 61%
- Hyperparameters: C=10, solver = newton-cg
- Validation set approach for Cross Validation (75:25)
- Data was standardized and regularized
- Multicollinearity was checked and no significant collinearity was found



Logistic Model Curve

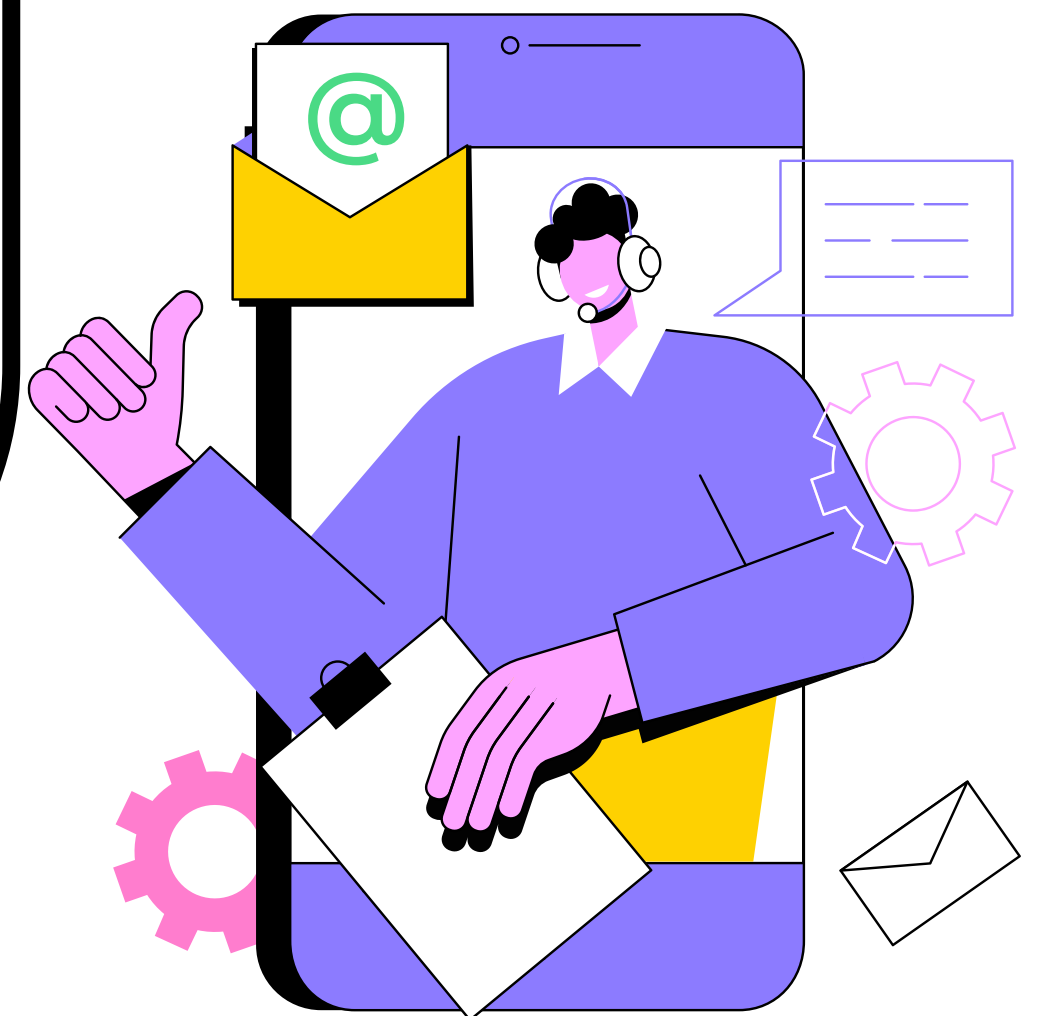


Advice and Future Strategies

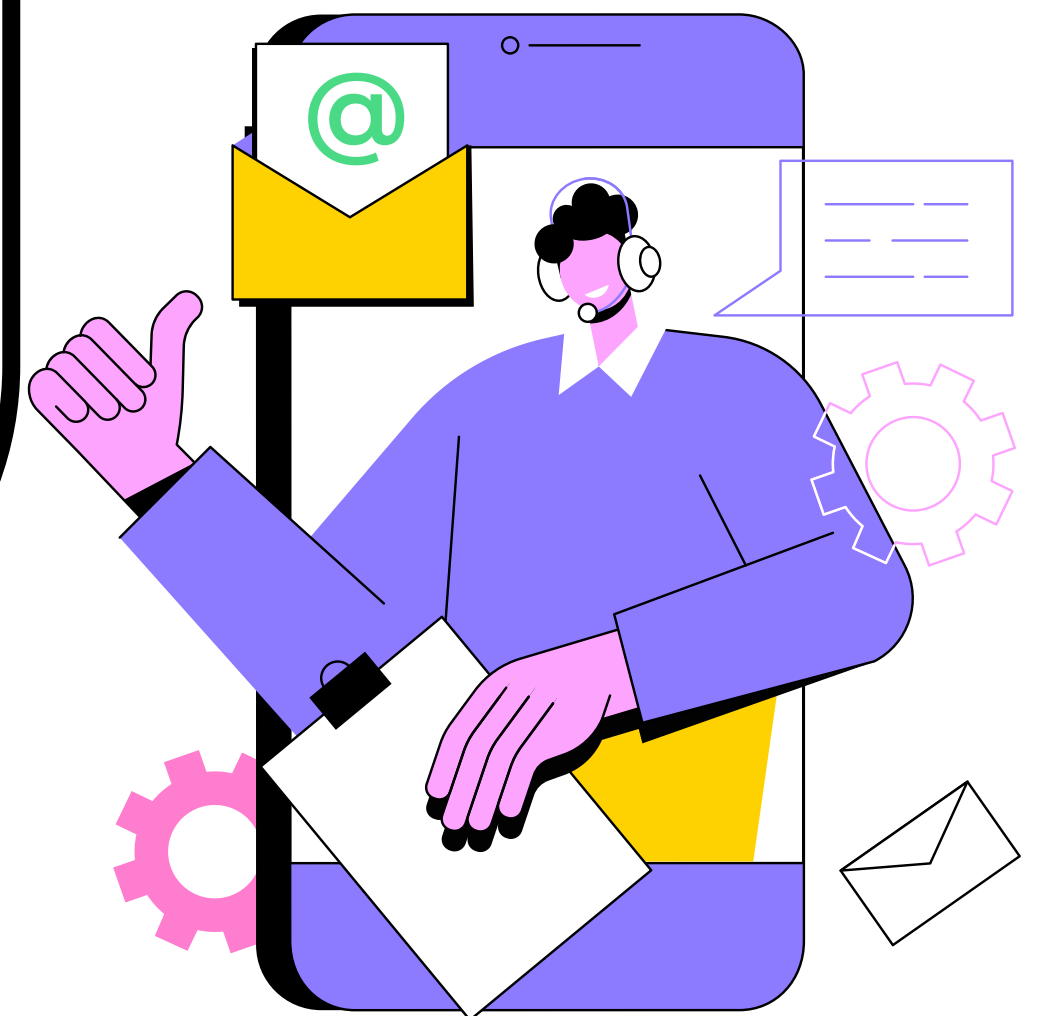
- Develop new products more suitable to profiles prone to more default (as they are the core customers of the bank)
- Initiate a knowledge sharing program with the government to develop stronger social security nets for the risky profiles
- Run distinct “loyalty/reward” program for customers with ideal profiles
- Assistance for risky profiles as they cross the threshold of the “risk” numbers by the government to provide debt relief
- Conduct Behavioural Analysis to improve the accuracy



Thank You

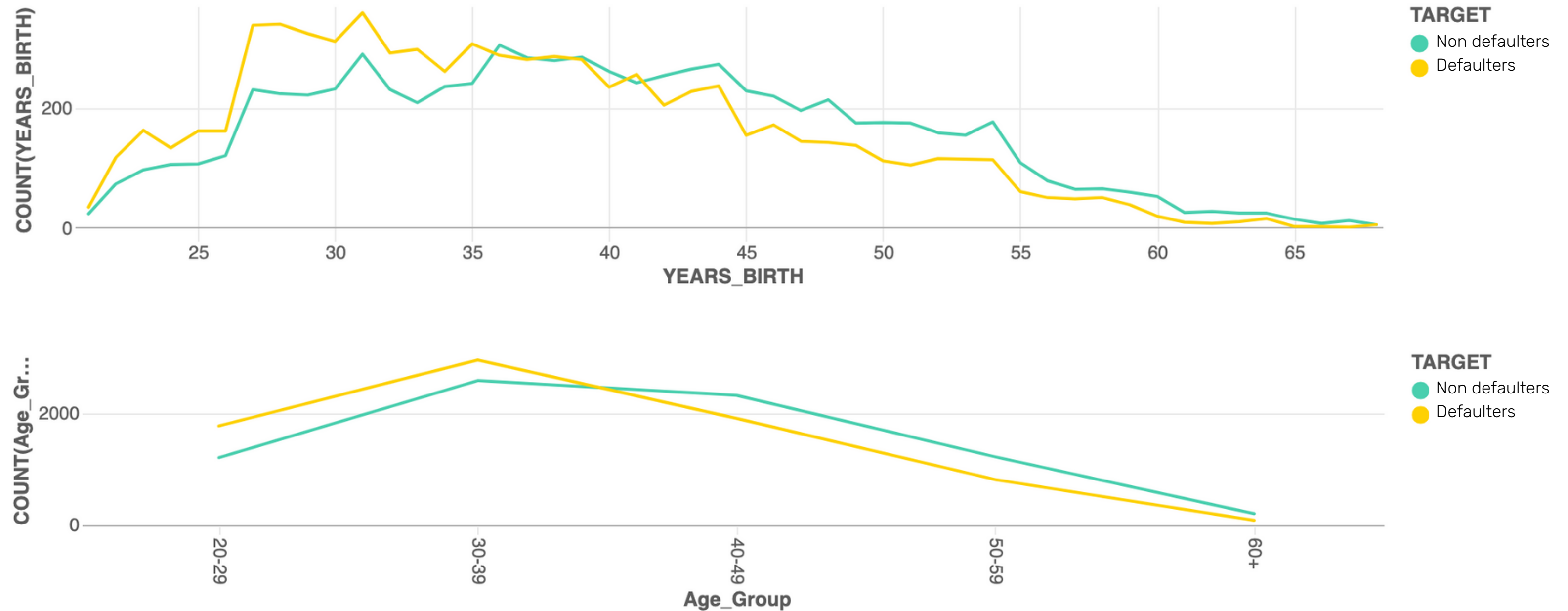


APPENDIX



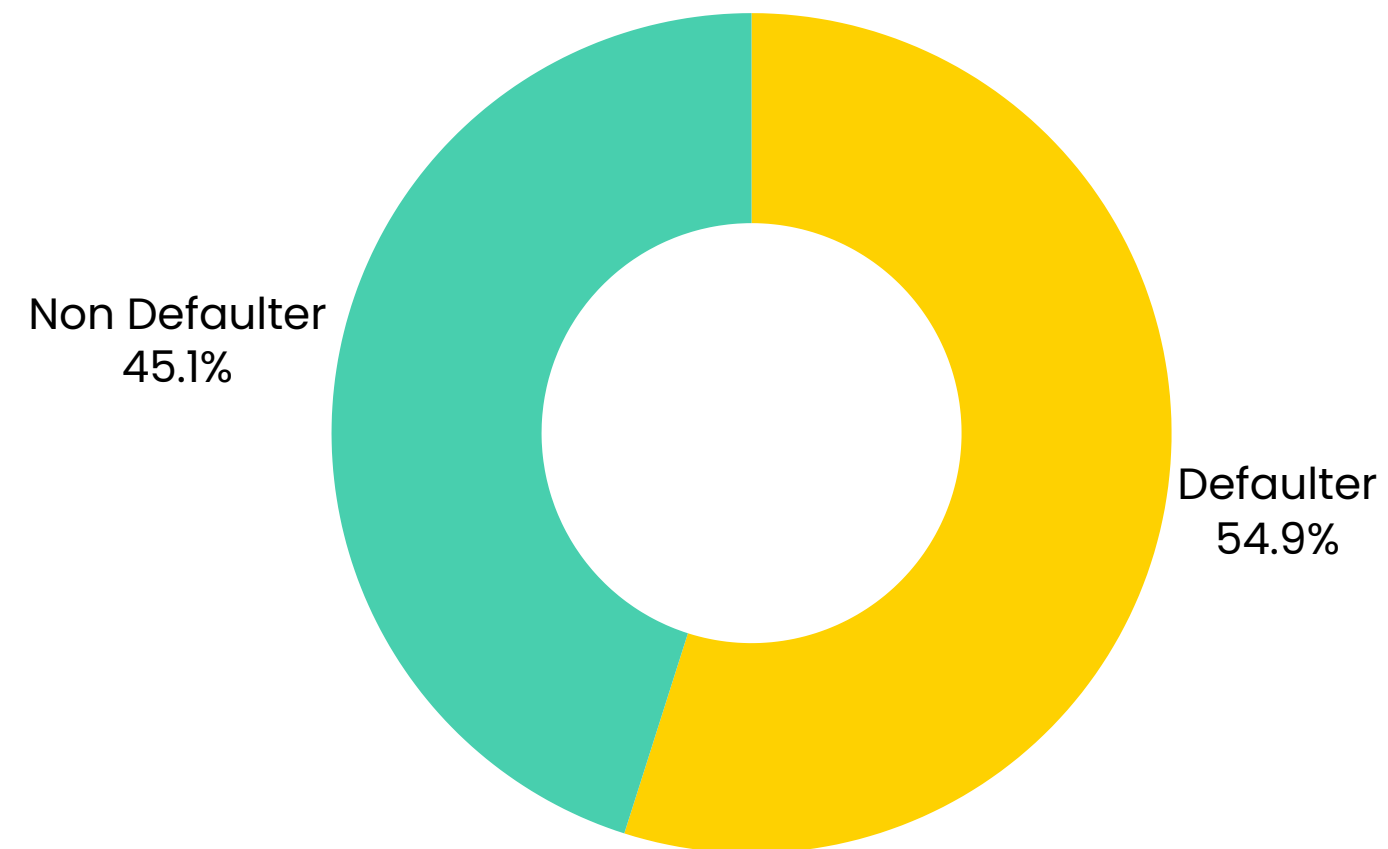
Key Observations

Default rate based on age



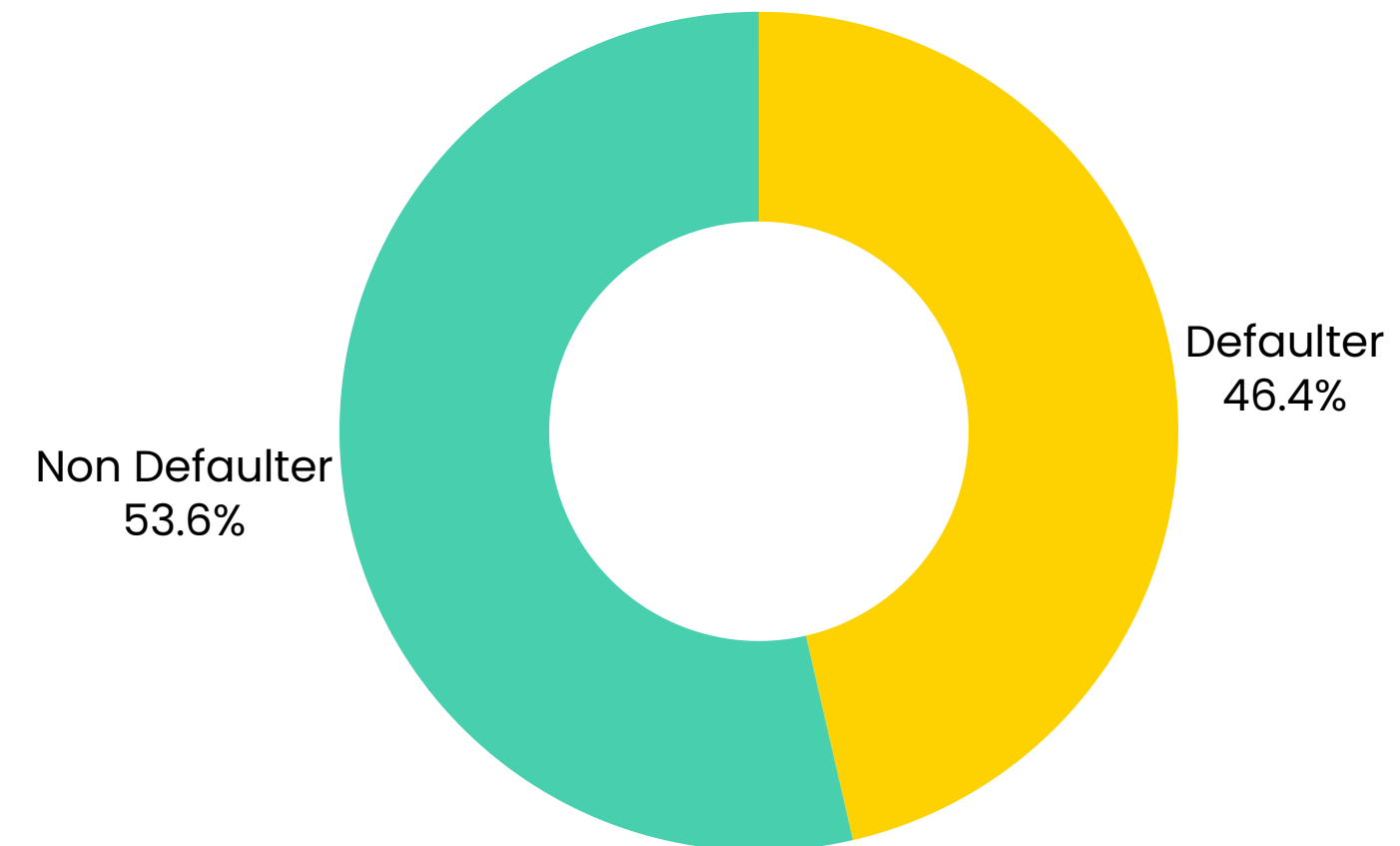
Key Observations

Default rate based on gender



Men

Percentage of defaulters in men

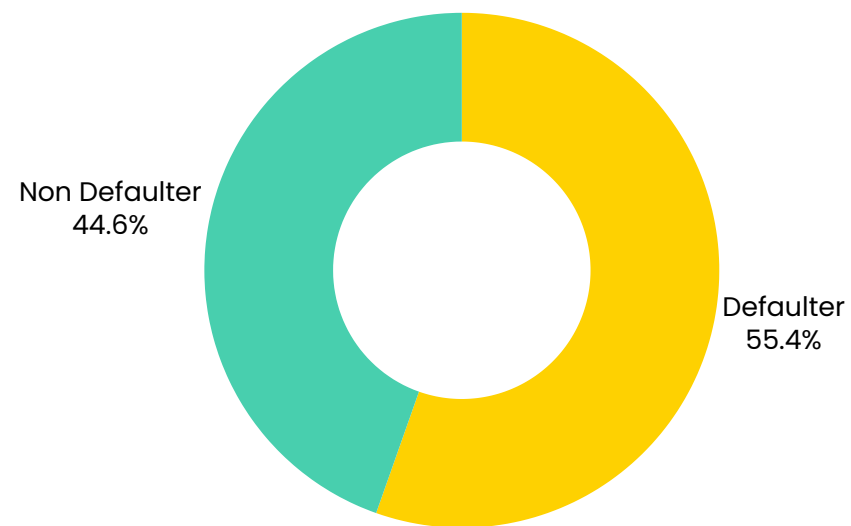


Women

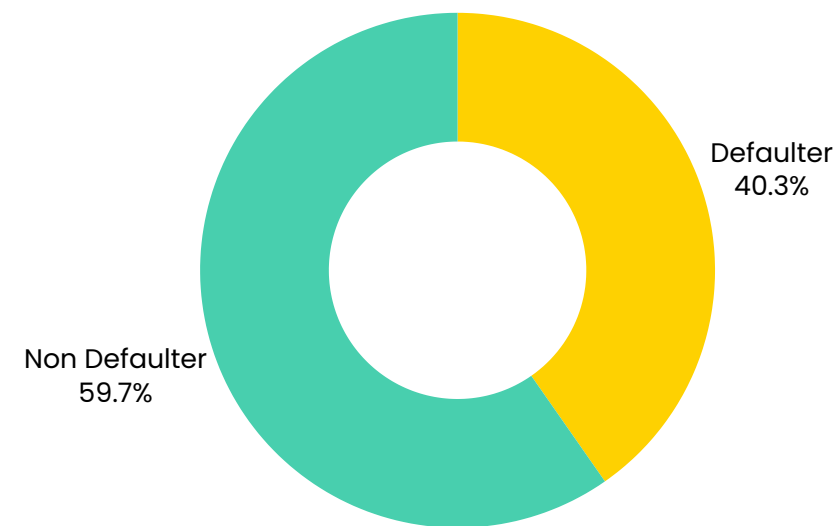
Percentage of defaulters in women

Key Observations

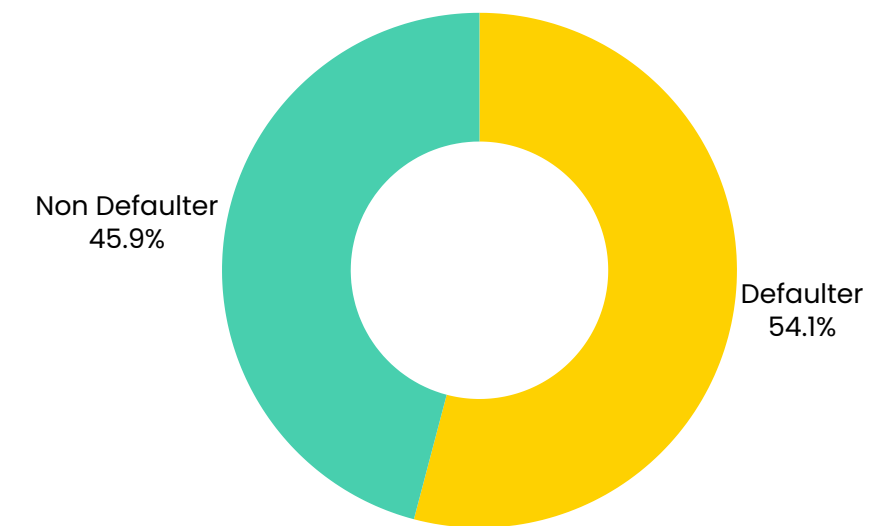
Default rate based on family status



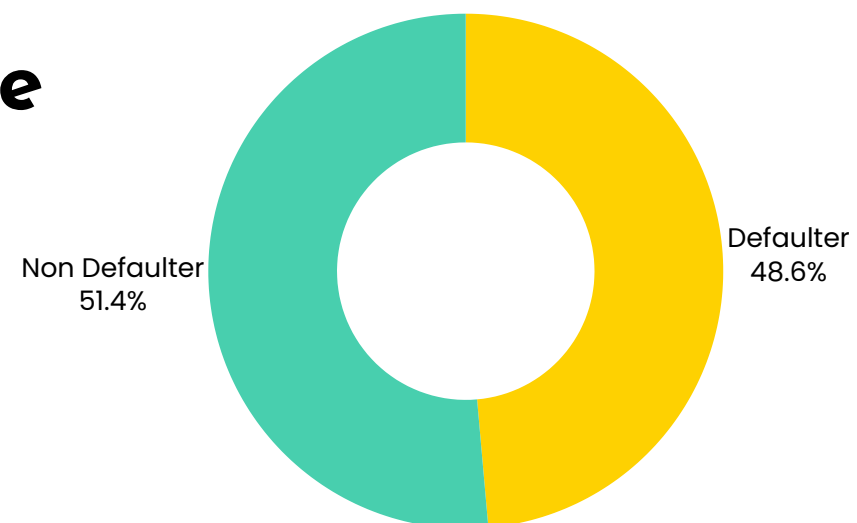
Civil Marriage



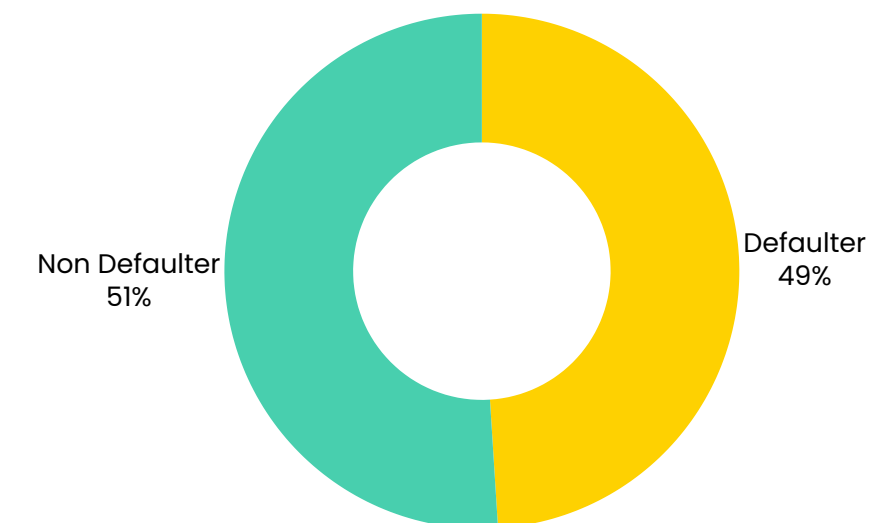
Widow



Single



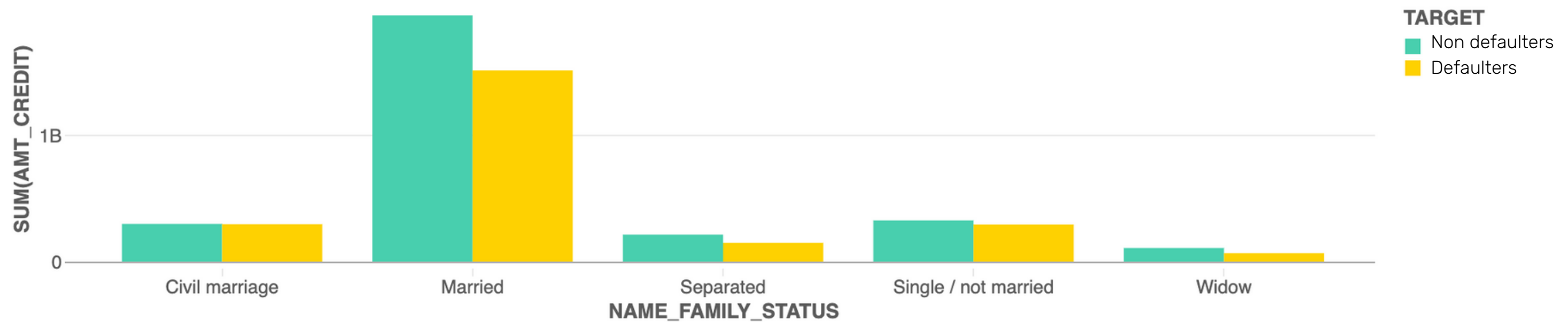
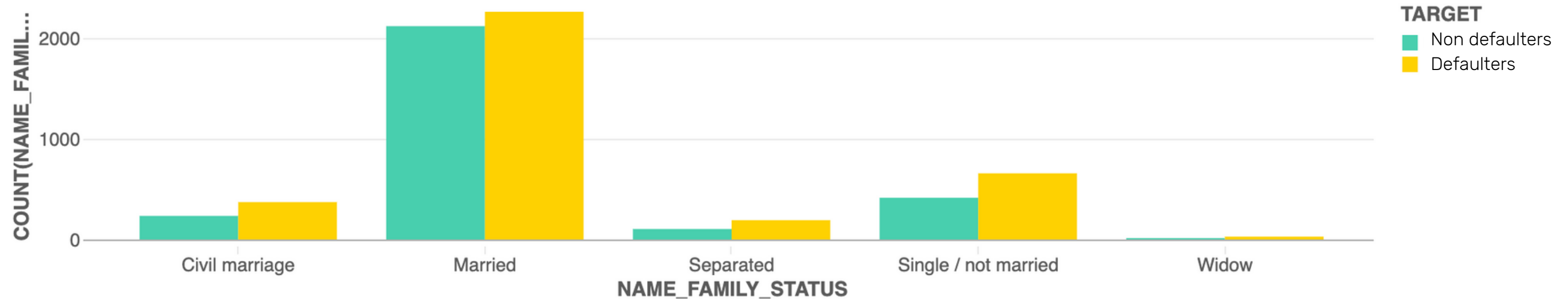
Married



Separated

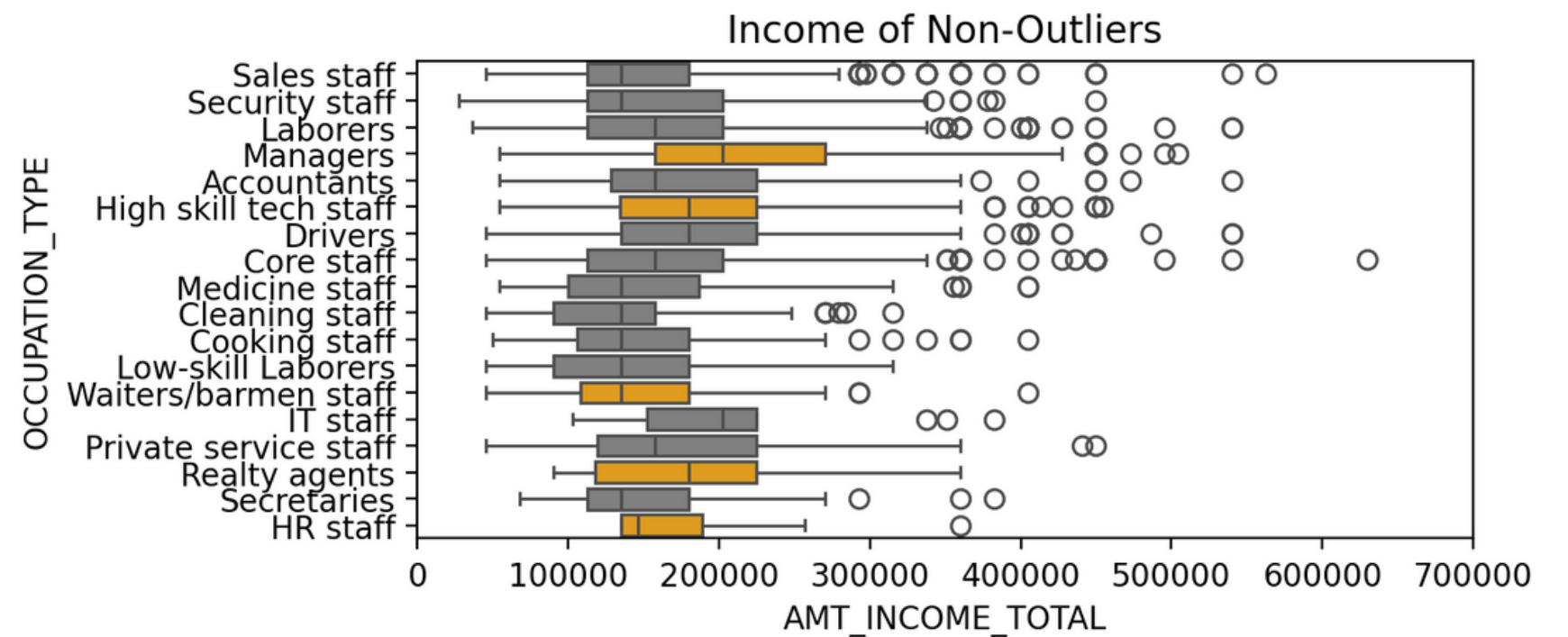
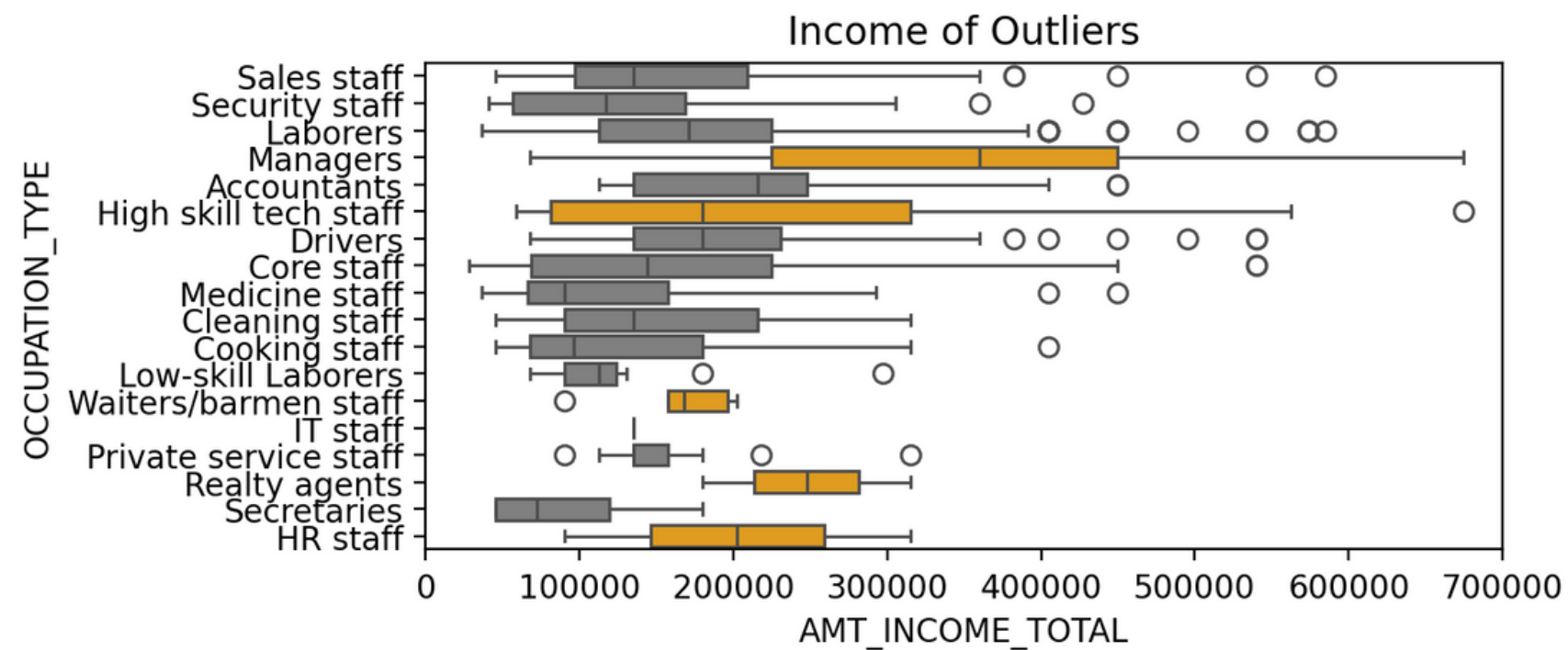
Key Observations

Default rate based on gender and family status



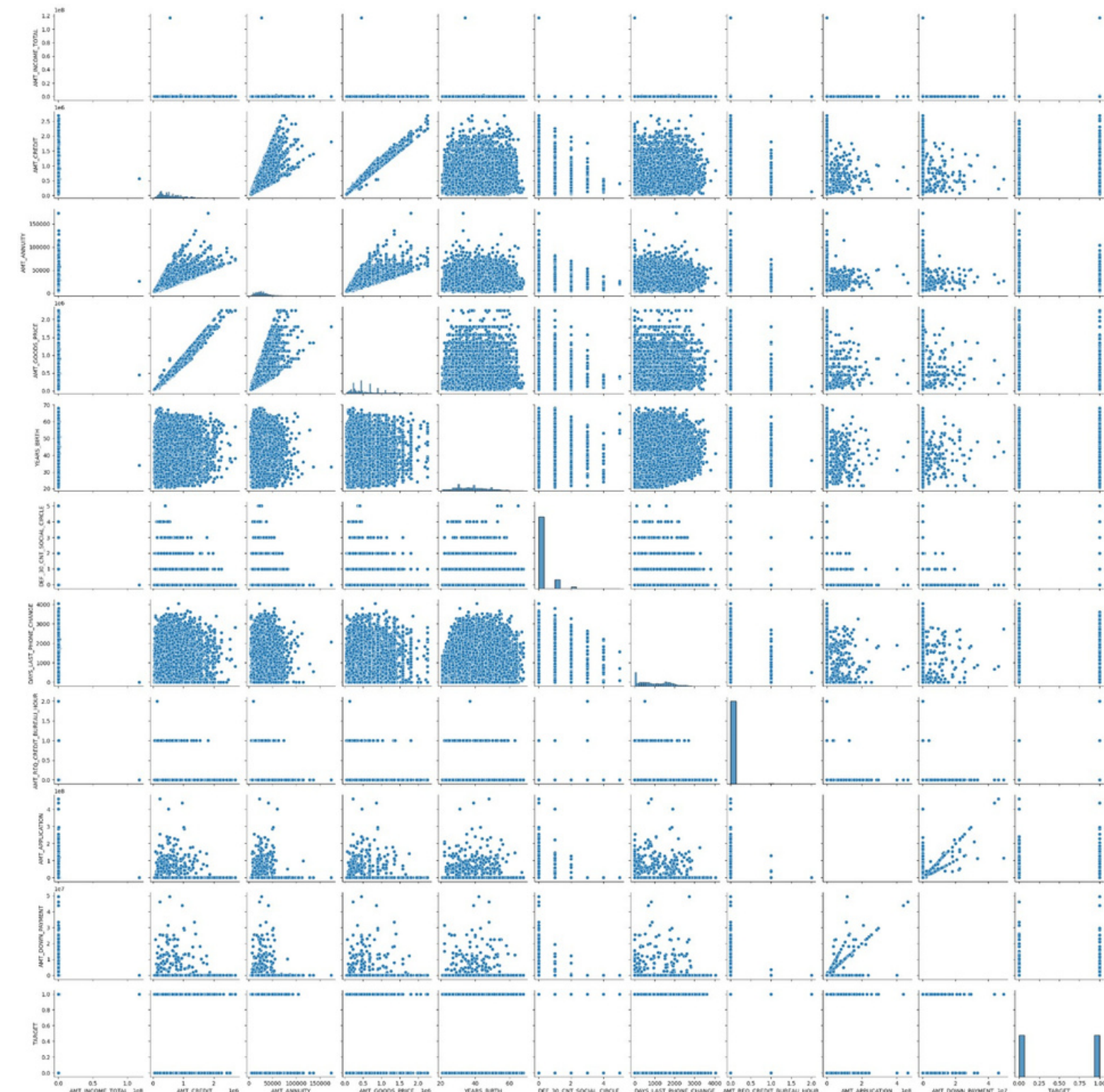
Key Observations

BOX PLOT – INCOME OF OUTLIERS AND NON OUTLIERS



Correlation

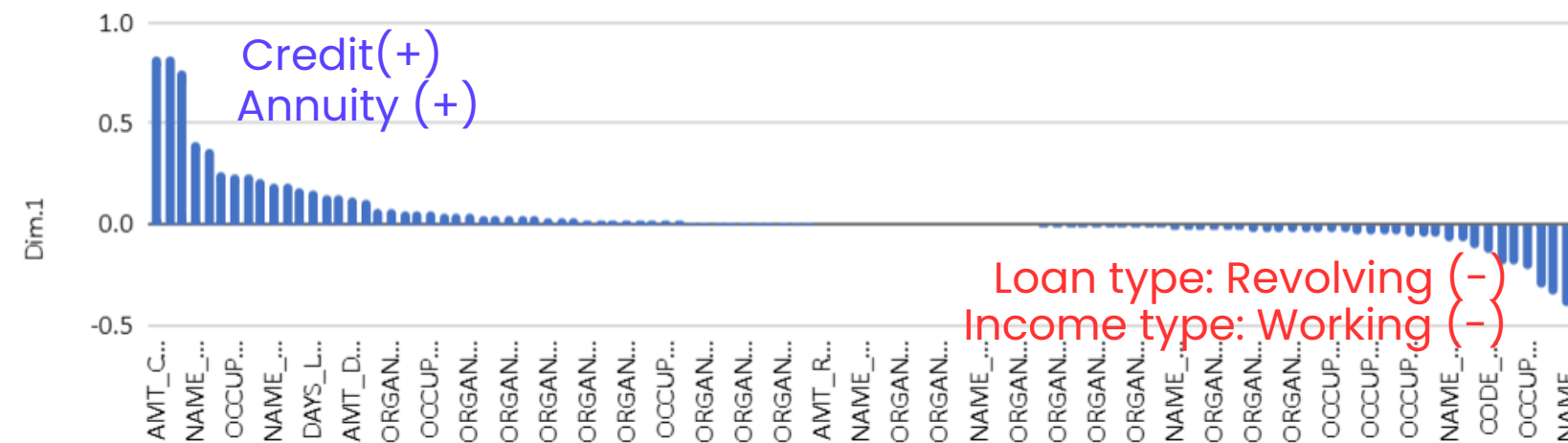
- All correlations present in this graph were not significant after standardization, and after applying regularization



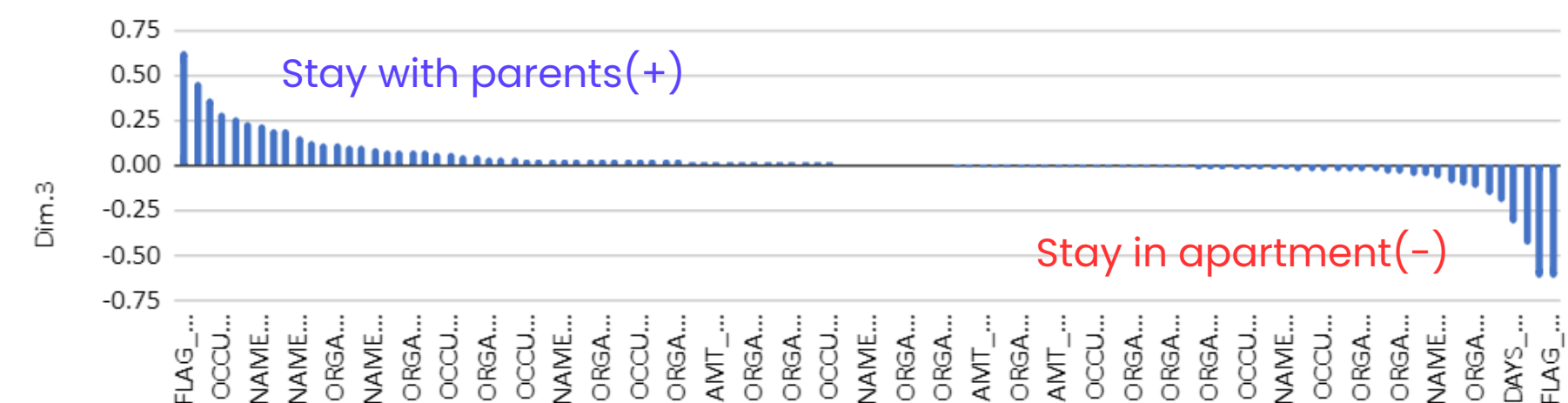
Principal Component Analysis

- We use PCA to identify clusters/groups of customers and few most important variables based on Loadings
- Positive loading has characteristics (variables) opposite to negative loading

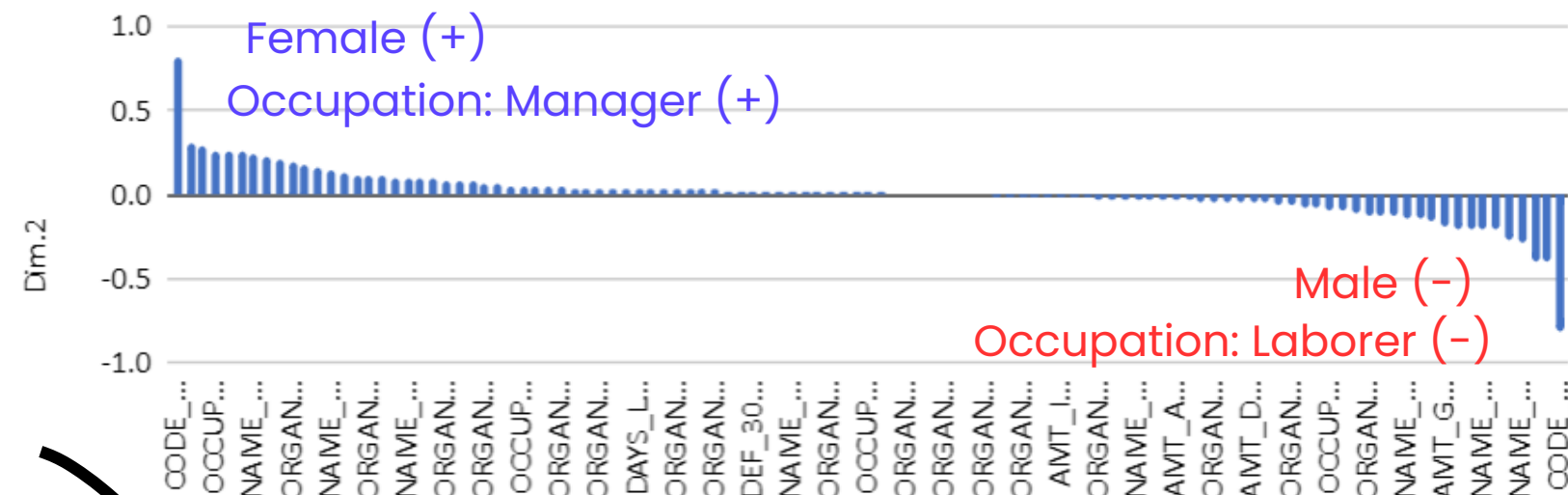
Loadings of 1st Principal Component



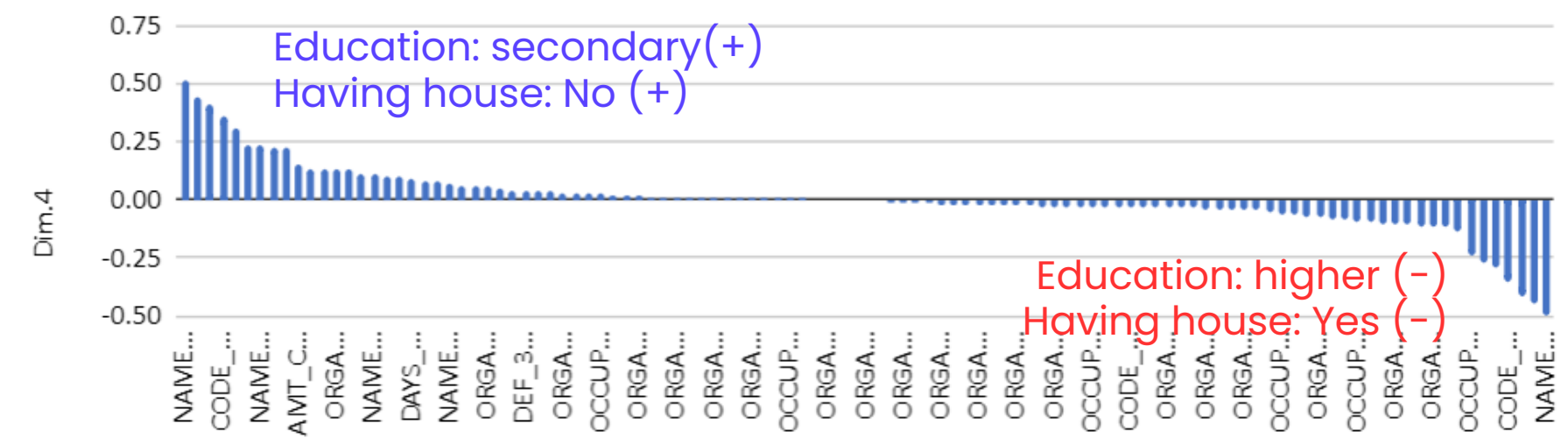
Loadings of 3rd Principal Component

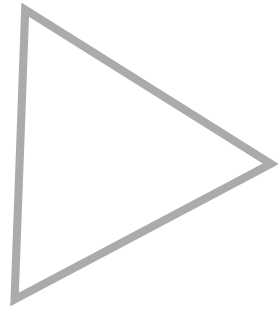


Loadings of 2nd Principal Component

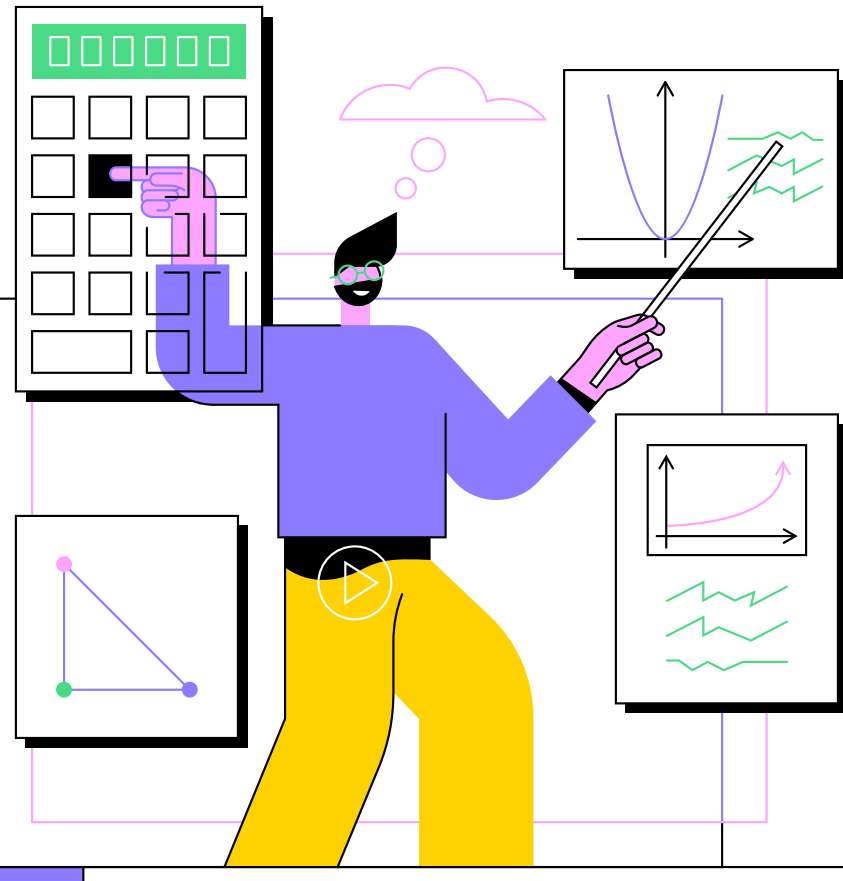
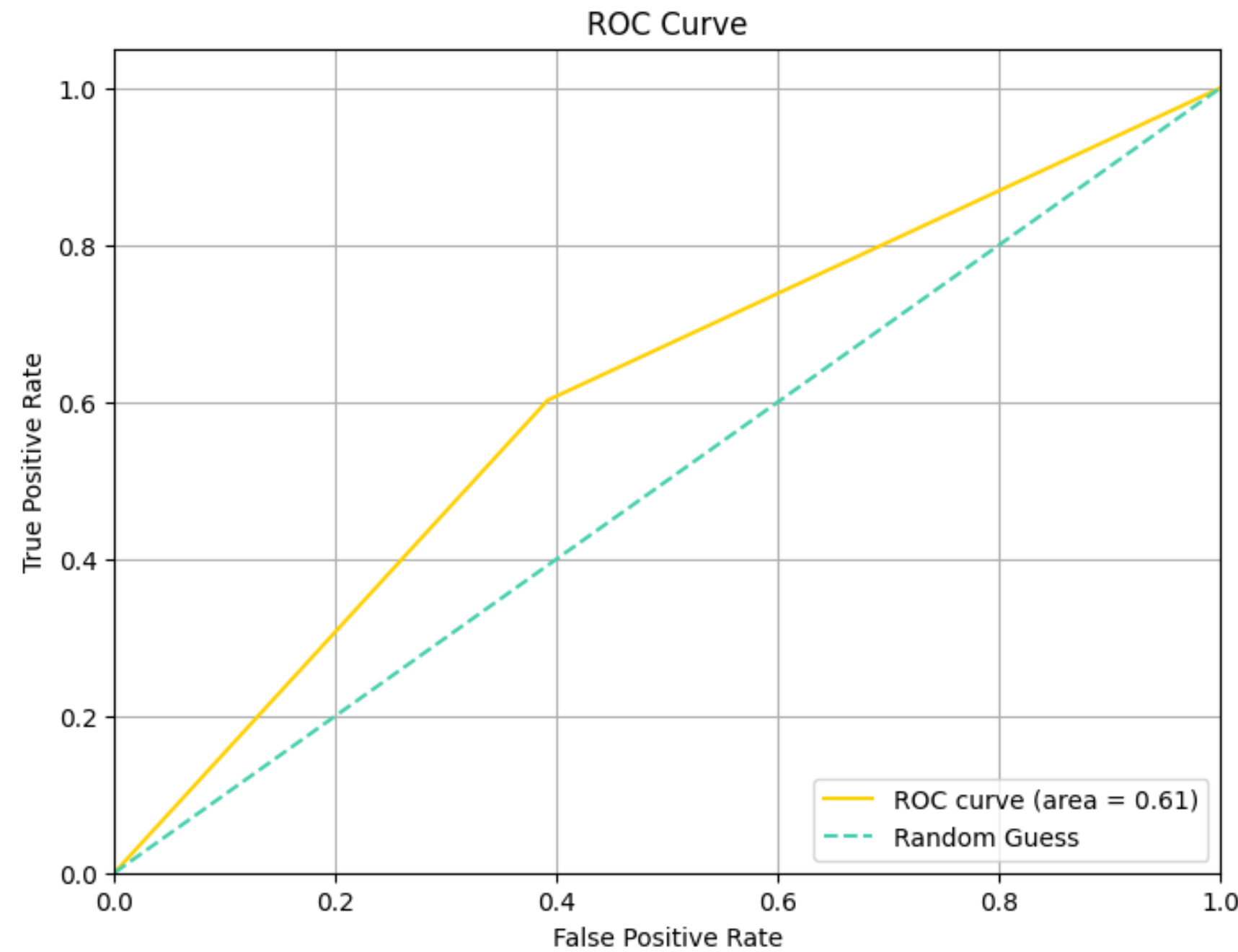


Loadings of 4th Principal Component





Logistic Regression



Project Overview (WIP)

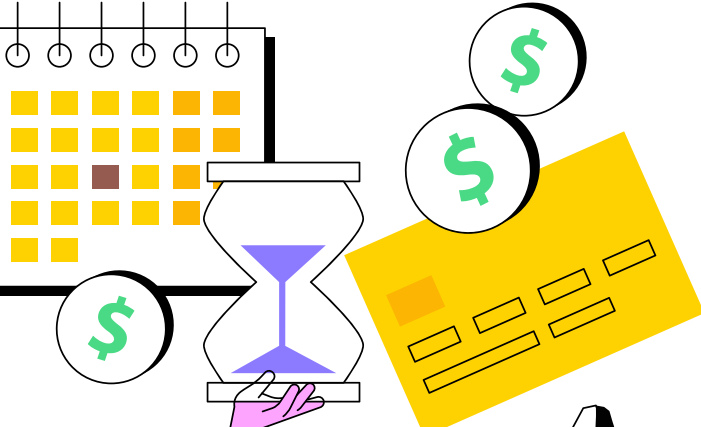
This project delves into exploring the characteristics of credit card defaulters by analyzing a rich dataset containing various attributes.

For analyzing this case, we made a balanced selection that encompasses both qualitative and quantitative descriptors, enriching our understanding of individual profiles.



Data Preprocessing

- Dataset contains of 300,000 observations with 122 variables
- Further sampling down to 2,000 observations
- Dimensionality Reduction: Selection of variables 22 variables, based on the following methodology:
 - Remove variables that contribute towards the same metric (eliminating noise)
- Data cleansing
 - Removing observations where variable "Occupation Type" is missing

Categorical	Numerical
AMT_INCOME_TOTAL	TARGET
AMT_CREDIT	NAME_CONTRACT_TYPE
AMT_ANNUITY	CODE_GENDER
AMT_GOODS_PRICE	FLAG_OWN_REALTY
DAYS_BIRTH	NAME_INCOME_TYPE
DEF_30_CNT_SOCIAL_CIRCLE	NAME_EDUCATION_TYPE
DAYS_LAST_PHONE_CHANGE	NAME_FAMILY_STATUS
AMT_REQ_CREDIT_BUREAU_HOUR	NAME_HOUSING_TYPE
	OCCUPATION_TYPE
	REG_REGION_NOT_LIVE_REGION
	ORGANIZATION_TYPE
	AMT_APPLICATION
	AMT_DOWN_PAYMENT

Variables (WIP)

By analyzing these variables and exploring their relationships through various statistical methods, we aim to uncover valuable patterns and associations that contribute to a more comprehensive picture of credit card defaulters within the dataset.

Analysis (WIP)

Analysis	Tools	Objective
Exploratory Data Analysis (EDA)	Boxplot, histogram, and more as necessary	Identify variables that differentiate card owner that does fraud from ones that does not
Multivariate Analysis	Correlation plot and more as necessary	Identify correlation between dependent and independent variables, identify outliers of data
Clusters	PCA / Classification Tree	Find clusters based on similarity of characteristics

