RESPONSIBLE DATA SCIENCE FINAL PROJECT

Larry Li
zl2902

Kayan Shih
ks5250

PART 1: BACKGROUND AND GOAL(S)

ADS BACKGROUND

1. Banking Industry

2. Risk Control

3. Predict the likelihood of future defaults and credit card borrowing

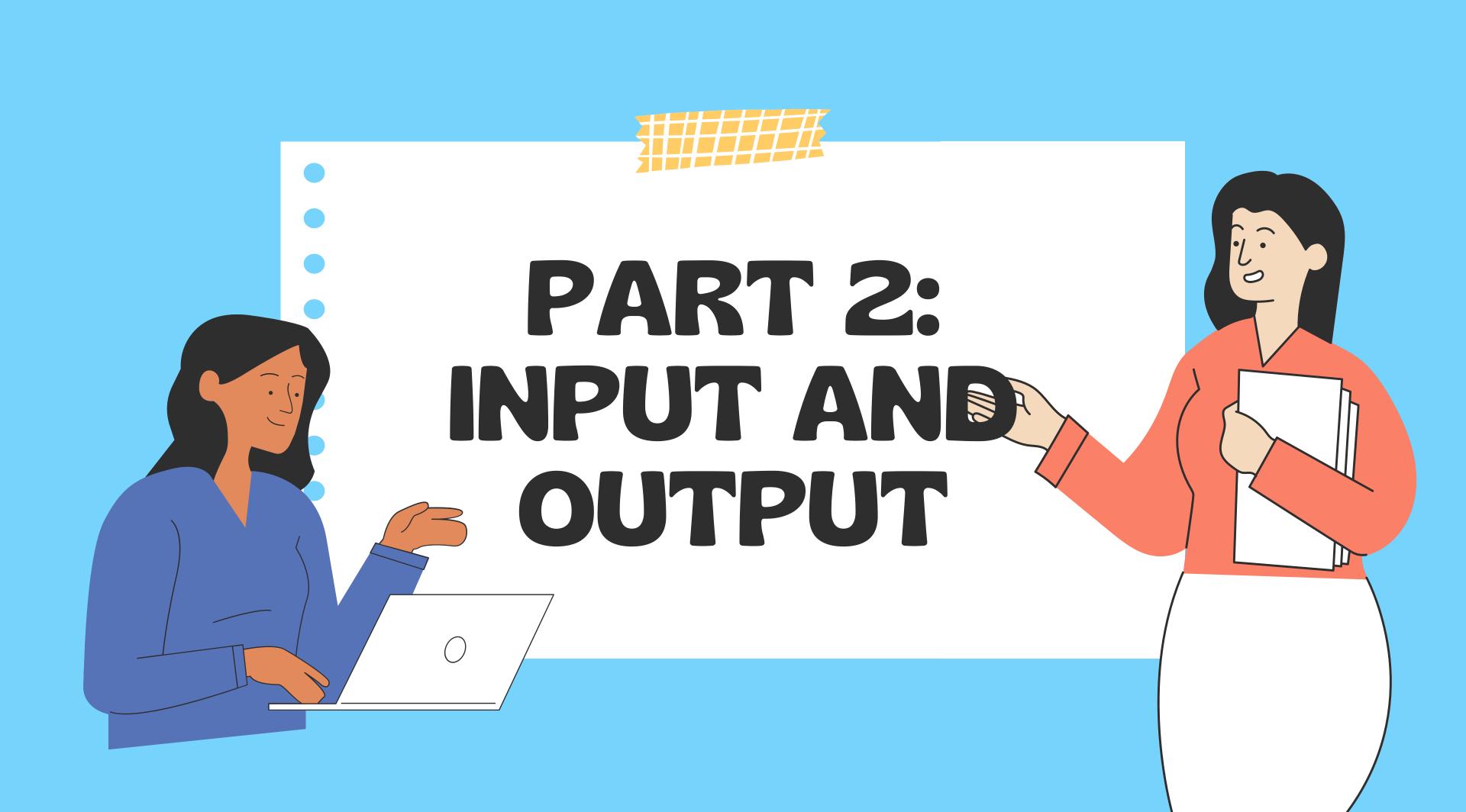## ADS: Credit Card Approval Prediction

Predict if an applicant is a 'good' or 'bad' client and decides whether to issue a credit card to the applican

Goal(s)

Improve the accuracy of the ADS results while mitigating its bias to avoid discriminatory outcome

PART 2: INPUT AND OUTPUT

INPUT

Table1: Applicant Dataset

438557 rows and 18 columns

Table2: Credit Dataset

1048575 rows and 3 columns

Feature Name	Datatype
ID	ID
CODE_GENDER	Categorical (Binary)
FLAG_OWN_CAR	Categorical (Binary)
FLAG_OWN_REALTY	Categorical (Binary)
CNT_CHILDREN	Numerical
AMT_INCOME_TOTAL	Numerical
NAME_INCOME_TYPE	Categorical
NAME_EDUCATION_TYPE	Categorical
NAME_FAMILY_STATUS	Categorical
NAME_HOUSING_TYPE	Categorical
DAYS_BIRTH	Numerical
DAYS_EMPLOYED	Numerical
FLAG_MOBIL	Categorical (Binary)
FLAG_WORK_PHONE	Categorical (Binary)
FLAG_PHONE	Categorical (Binary)
FLAG_EMAIL	Categorical (Binary)
OCCUPATION_TYPE	Categorical
CNT_FAM_MEMBERS	Numerical

Table 1: Input Features for Applicant Dataset

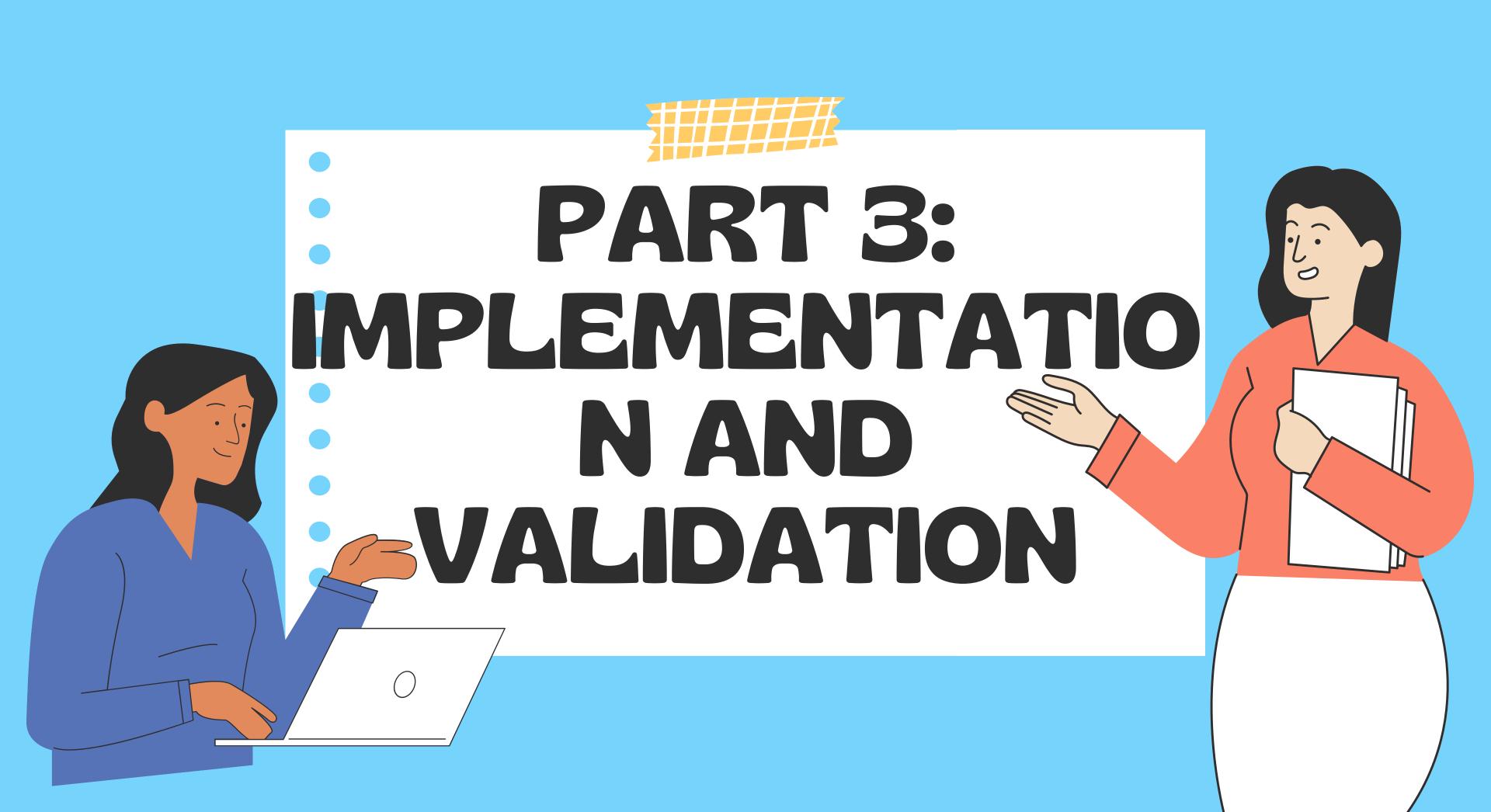| Feature Name	Datatype
ID	ID
MONTHS_BALANCE	Numerical
STATUS	Categorical

Table 2: Input Features for Credit Dataset

OUTPUT (BINARY)

0 Indicate that the consumer has good credit

1 Indicate that the consumer has bad credit

PART 3: IMPLEMENTATION AND VALIDATION

DATA CLEANING AND PREPROCESSING

ACTION 1

Define users who are overdue for more than 60 days as target risk users and the other users as good credit users

ACTION 2

Employ Synthetic Minority Over-Sampling Technique (SMOTE) to overcome the sample imbalance problem

ACTION 3

Drop all the null values in the dataset

ACTION 4

Transform the data using Weight of Evidence (WOE) to help calculate the Information Value of each feature

IMPLEMENTATION OF THE SYSTEM

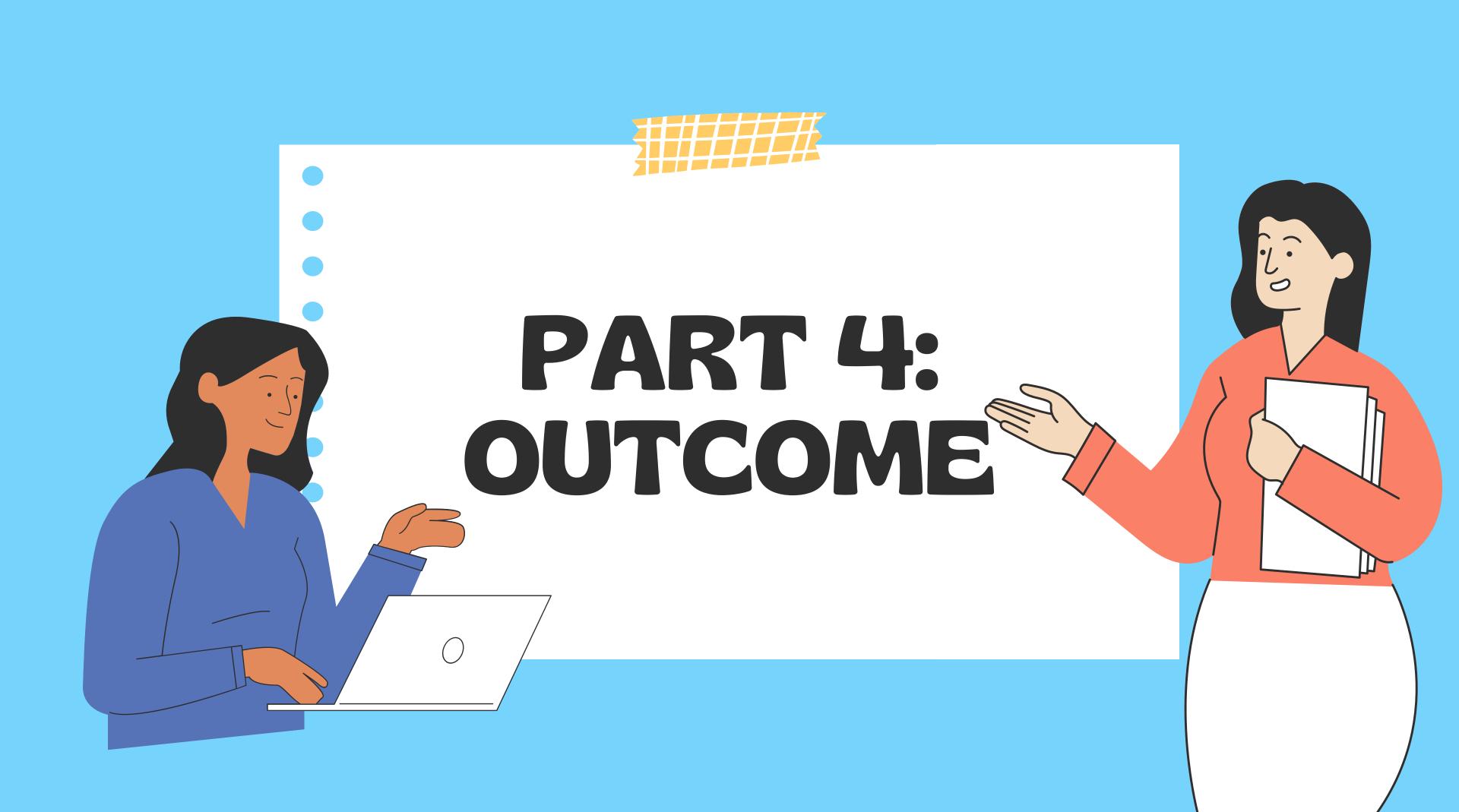Select features to be used and break these features into distinct column

Fit these selected features to the XGBoost model and assign a series of hyperparameter

VALIDATION OF THE SYSTEM

Use accuracy as the validation metric and show an accuracy score of 87.8%

Can better examine the credit status of their applicants while applying risk controls

PERFORMANCE BASED ON GENDER

Number of Female Applicants: 8975

Number of Male Applicants: 5853

	Overall	Female	Male
Accuracy	0.8734	0.8697	0.8790
PPV	0.8459	0.8460	0.8459
FPR	0.1663	0.1553	0.1844
FNR	0.0869	0.1040	0.0622

PERFORMANCE BASED ON AGE

Number of Young Applicants: 2416

Number of Middle Age and Senior Applicants: 11970

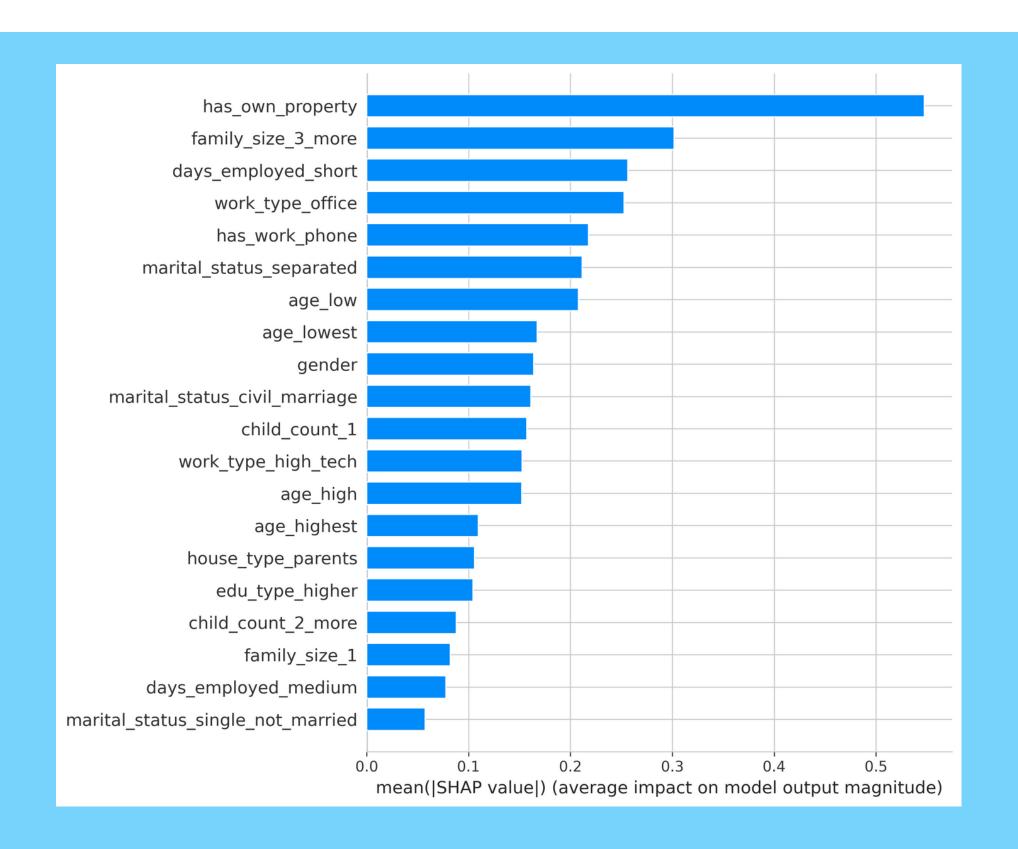|	Overall	Young	Middle Age and Senior
Accuracy	0.8734	0.8816	0.8671
PPV	0.8459	0.8569	0.8327
FPR	0.1663	0.1577	0.1680
FNR	0.0869	0.0801	0.0950
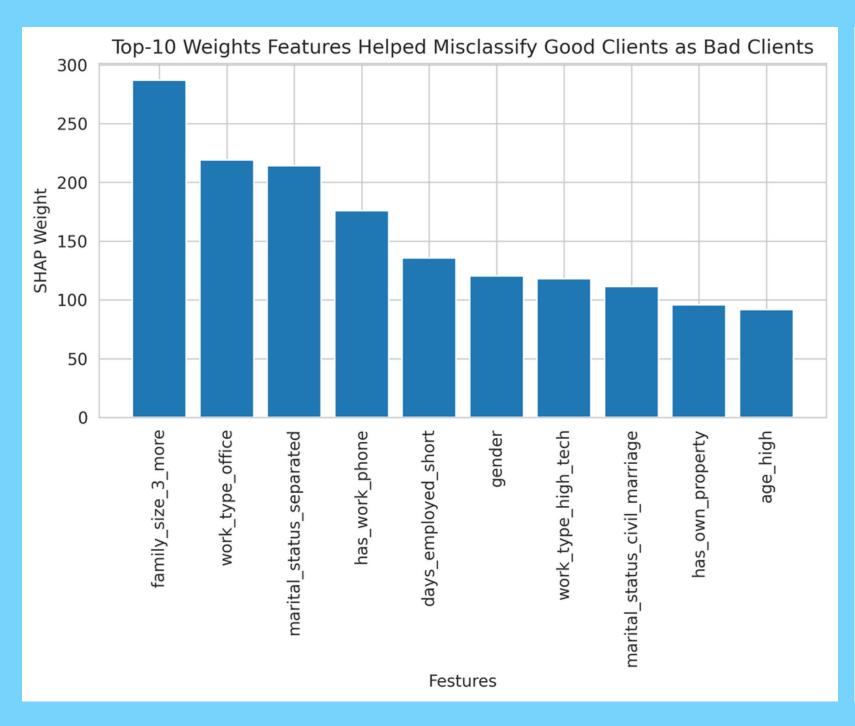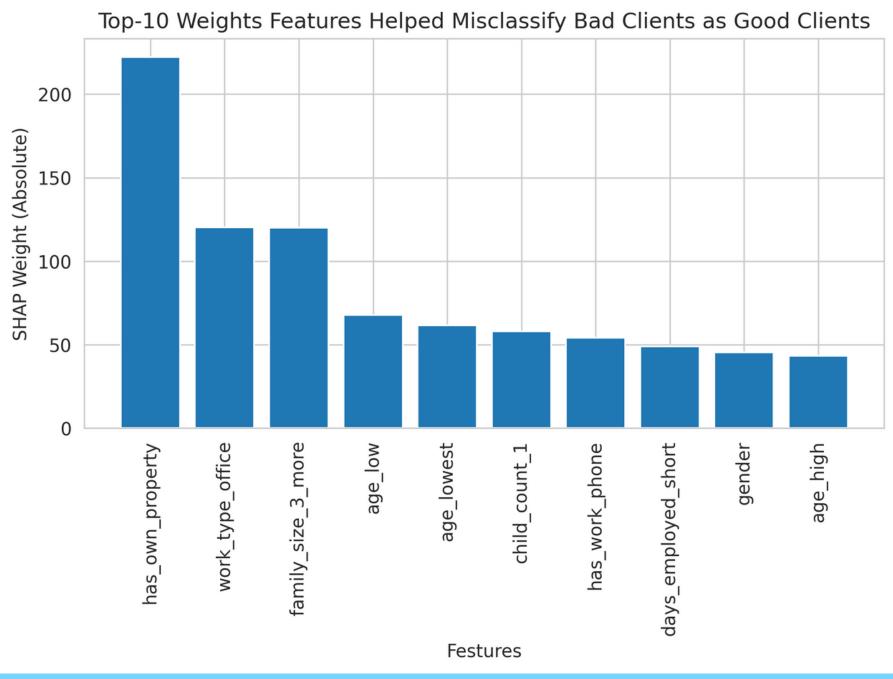
FEATURE IMPORTANCE

has_own_property	
family_size_3_more	
days_employed_short	
work_type_office	
has_work_phone	
marital_status_separated	
age_low	
age_lowest	
gender	
marital_status_civil_marriage	
child_count_1	
work_type_high_tech	
age_high	
age_highest	
house_type_parents	
edu_type_higher	
child_count_2_more	
family_size_1	
days_employed_medium	
marital_status_single_not_married	

mean(|SHAP value|) (average impact on model output magnitude)

THANK YOU!

Have a great day ahead.