BUAN 6341.002 - Applied Machine Learning - S24

THE UNIVERSITY OF TEXAS AT DALLAS

HOME CREDIT DEFAULT RISK

Presented by Group 1

Akshay Varma Penmatsa
Suman Bar
Rucheek Rajeev Kashyap

Varun Golas
Jagannadh Bora

OBJECTIVE

Predicting whether a prospective borrower may default

Classification Problem

Positive Class – Borrower

Defaults (1)

Assessing Credit Worthiness

Risk Management

Access to Competitive Loans

DATA SOURCE

https://www.kaggle.com/competitions/home-credit-default-risk/data

https://www.homecredit.net/

Data on Mortgage Loans Financed by Home Credit

Data collected during application of Loan

Target Variable – Active (0) / Default (1)

Personally Identifiable Information Removed

DATA DESCRIPTION

Data from Loan Applications with Home Credit

122 Columns – 120 Features, 1 Identifier, 1 Target

307511 Rows – Each Record corresponding to a unique Loan Application

24,825 Record in Positive Class (Default) – About 8.07% of all Data

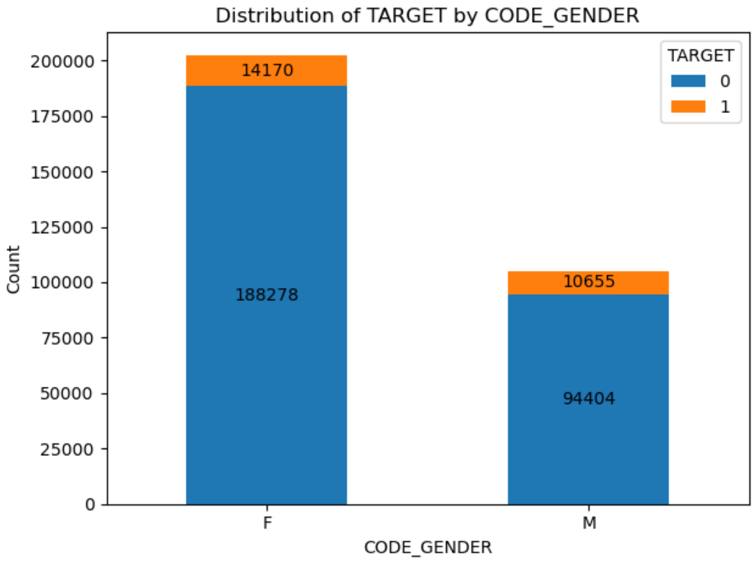
Demographic Information – Gender / Education / Family

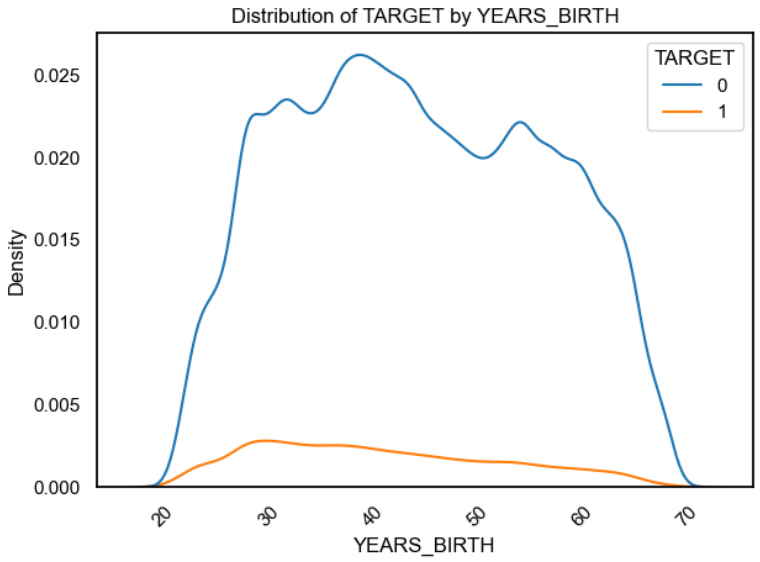
Employment Information – Occupation / Organization / Income

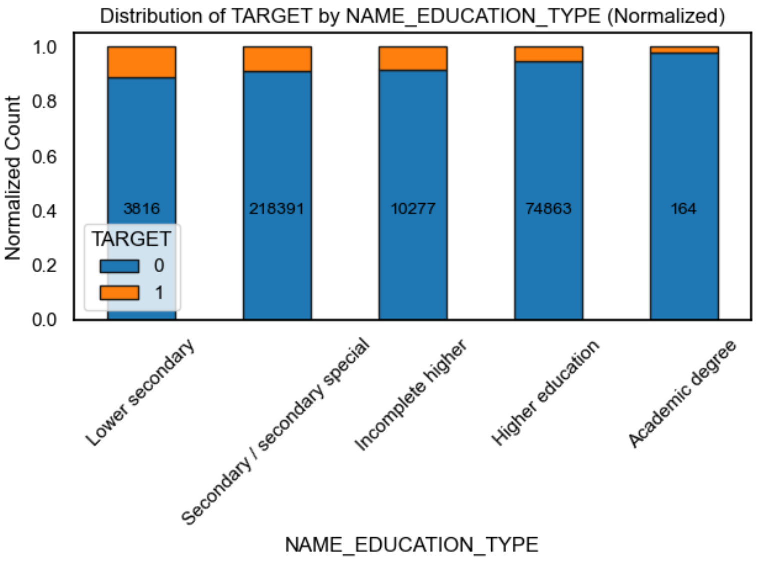
Geographic Information – Region / Living arrangement

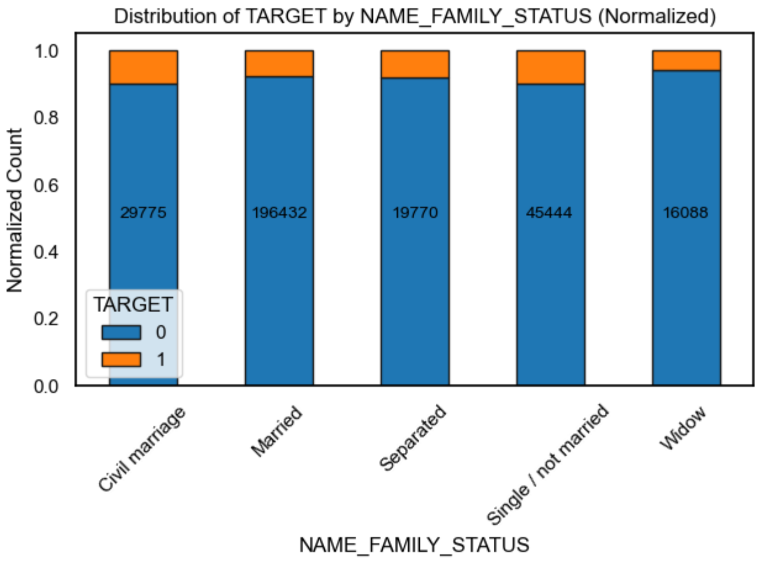
Existing Credit Scores if any – 3 External Rating Agencies

EXPLORATORY DATA ANALYSIS

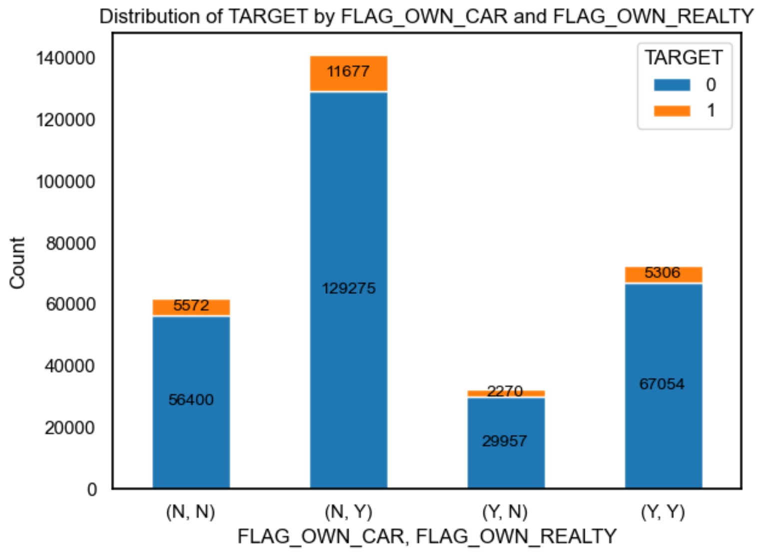


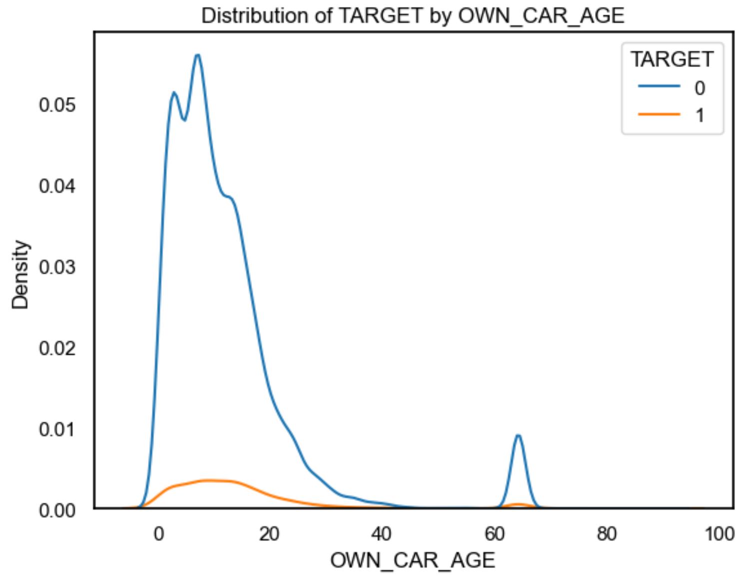


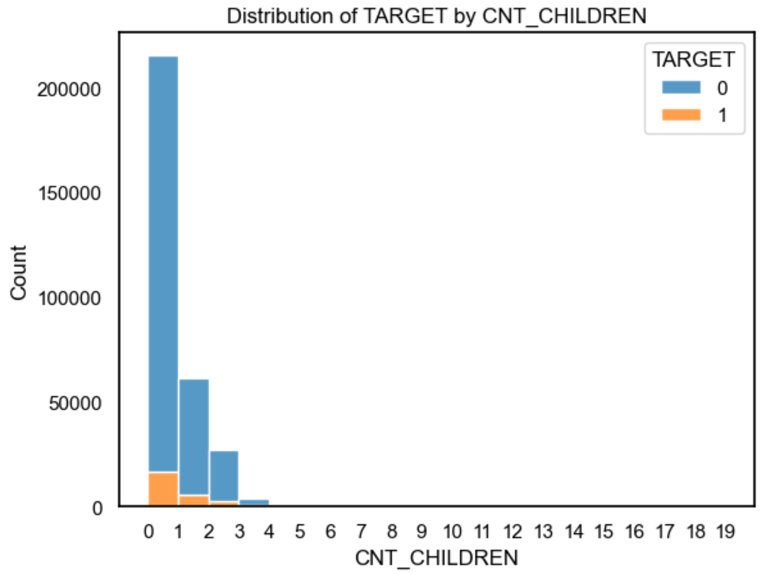


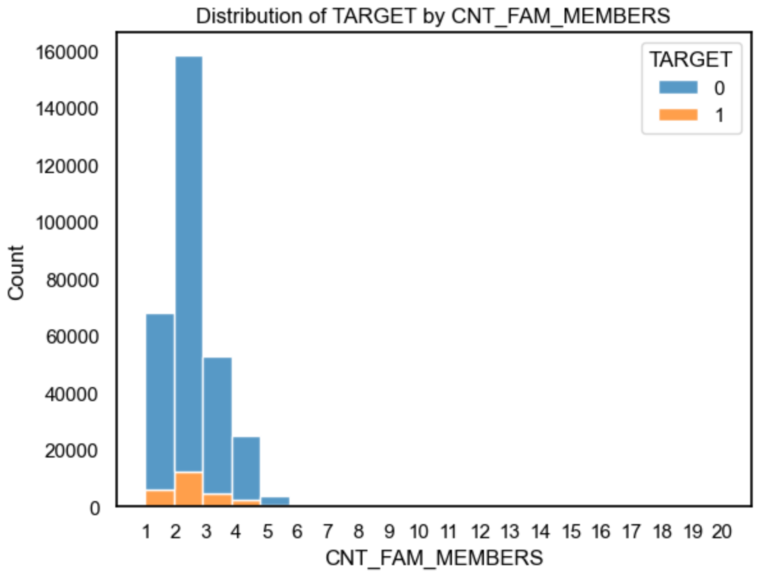


EXPLORATORY DATA ANALYSIS

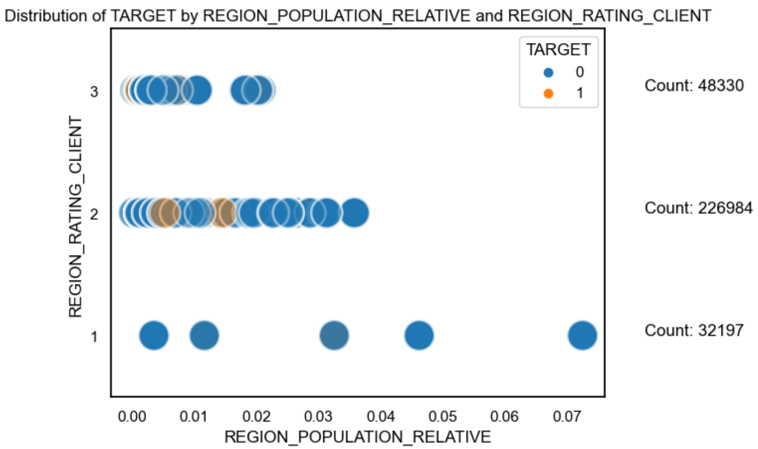


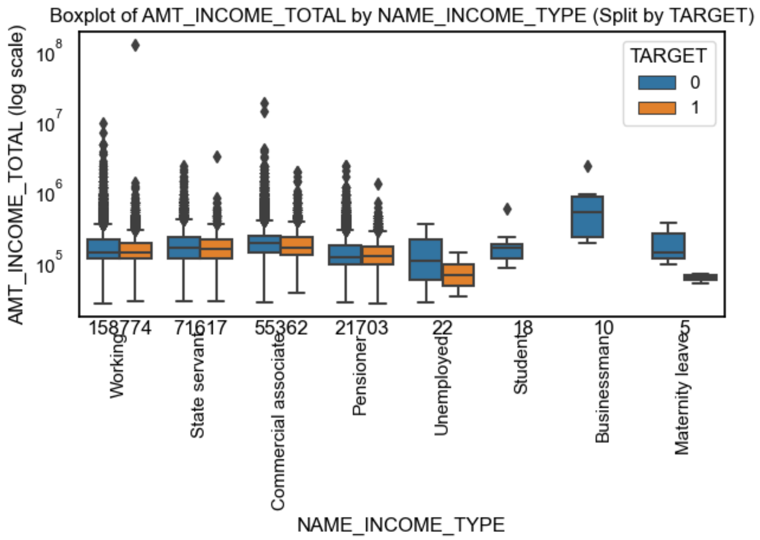


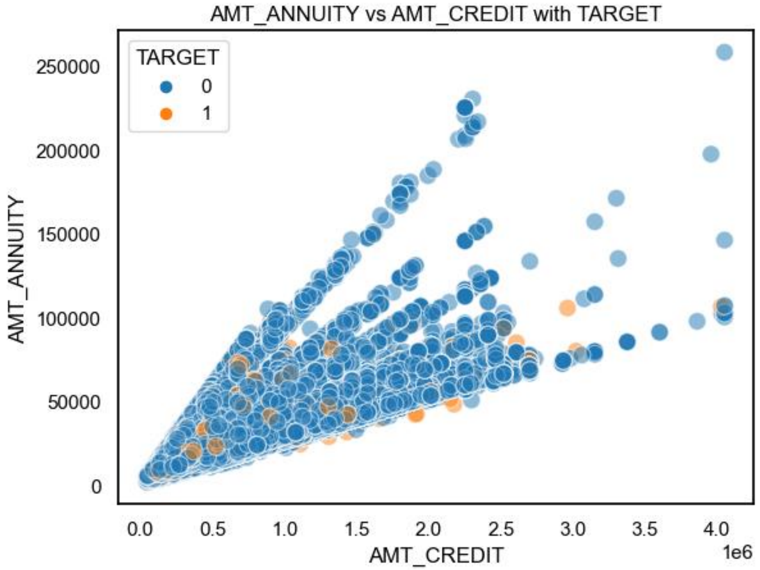


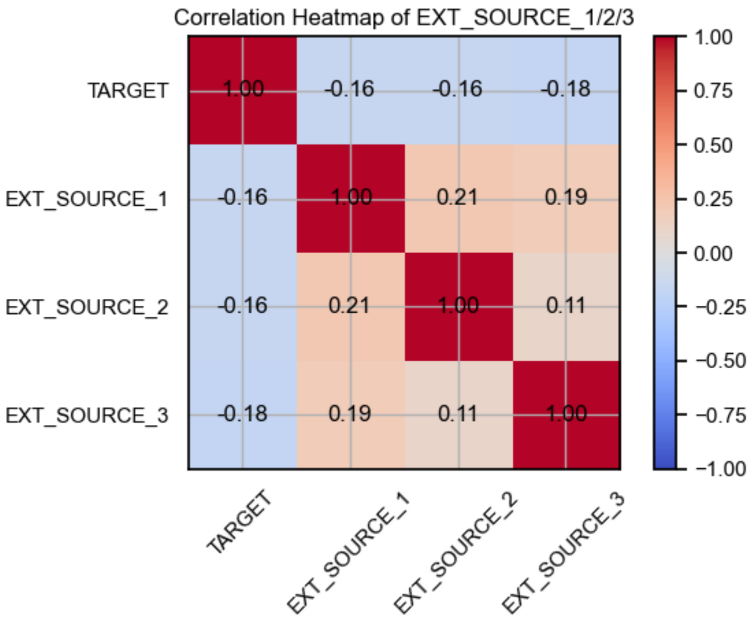


EXPLORATORY DATA ANALYSIS









FEATURE LIST (shortened)

DEMOGRAPHIC FEATURES	
CODE_GENDER	Gender of the client
FLAG_OWN_CAR	Flag if the client owns a car
FLAG_OWN_REALTY	Flag if client owns a house or flat
CNT_CHILDREN	Number of children the client has
NAME_EDUCATION_TYPE	Level of highest education the client achieved
NAME_FAMILY_STATUS	Family status of the client
NAME_HOUSING_TYPE	What is the housing situation of the client (renting, living with parents, ...)
DAYS_BIRTH	Client's age in days at the time of application
OWN_CAR_AGE	Age of client's car
CNT_FAM_MEMBERS	How many family members does client have
4 Features for Social Circle	How many observation of client's social surroundings with observable DPD

INCIDENTAL FEATURES.	
NAME_CONTRACT_TYPE	Identification if loan is cash or revolving
AMT_CREDIT	Credit amount of the loan
AMT_ANNUITY	Loan annuity
AMT_GOODS_PRICE	For consumer loans it is the price of the goods for which the loan is given
NAME_TYPE_SUITE	Who was accompanying client when he was applying for the loan
DAYS_REGISTRATION	How many days before the application did client change his registration
DAYS_ID_PUBLISH	How many days before the application did client change the identity document with which he applied for the loan
6 Features for Contact info	Did client provide x (1=YES, 0=NO)
WEEKDAY_APPR_PROCESS_START	On which day of the week did the client apply for the loan
2 Features for Application Time	On which day of the week / hour did the client apply for the loan
20 Features for FLAG_DOCUMENT	Did client provide document

GEOGRAPHIC FEATURES	
REGION_POPULATION_RELATIVE	Normalized population of region where client lives (higher number means the client lives in more populated region)
REGION_RATING_CLIENT	Our rating of the region where client lives (1,2,3)
REGION_RATING_CLIENT_W_CITY	Our rating of the region where client lives with taking city into account (1,2,3)
6 Features for Address whether same	Flag if client's x address does not match y address (1=different, 0=same, at region/city level)
47 Features for Living Arrangement	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor

EMPLOYMENT FEATURES	
AMT_INCOME_TOTAL	Income of the client
NAME_INCOME_TYPE	Clients income type (businessman, working, maternity leave,...)
DAYS_EMPLOYED	How many days before the application the person started current employment
OCCUPATION_TYPE	What kind of occupation does the client have
ORGANIZATION_TYPE	Type of organization where client works

CREDIT HISTORY FEATURES	
3 Features for EXT_SOURCE	Normalized score from external data source
6 Features for Credit Enquiries	Number of enquiries to Credit Bureau about the client

HANDLING MISSING VALUES

	Missing Values	% of Total Values
COMMONAREA_MEDI	214865	69.9
COMMONAREA_AVG	214865	69.9
COMMONAREA_MODE	214865	69.9
NONLIVINGAPARTMENTS_MEDI	213514	69.4
NONLIVINGAPARTMENTS_MODE	213514	69.4
NONLIVINGAPARTMENTS_AVG	213514	69.4
FONDKAPREMONT_MODE	210295	68.4
LIVINGAPARTMENTS_MODE	210199	68.4
LIVINGAPARTMENTS_MEDI	210199	68.4
LIVINGAPARTMENTS_AVG	210199	68.4
FLOORSMIN_MODE	208642	67.8
FLOORSMIN_MEDI	208642	67.8
FLOORSMIN_AVG	208642	67.8
YEARS_BUILD_MODE	204488	66.5
YEARS_BUILD_MEDI	204488	66.5
YEARS_BUILD_AVG	204488	66.5
OWN_CAR_AGE	202929	66.0
LANDAREA_AVG	182590	59.4
LANDAREA_MEDI	182590	59.4
LANDAREA_MODE	182590	59.4

New Shape of Training Data: (307511, 122)

Out of 122 columns, 67 columns that have missing values.

Drop columns with more than 60% missing values.

New Shape of Training Data: (307511, 105)

New Data Frame has 105 columns.

DATA PREPROCESSING

Through error analysis from running models without pre-processing we found few anomalies we need to take care of

- Label Encoding

```
NAME_CONTRACT_TYPE   FI
                 0
                 0
                 1
                 0
                 0
               ...
                 0
                 0
                 0
                 0
                 0
```

- One-hot Encoding

```
WALLSMATERIAL_MODE_Mixed  WALLSMATERIAL_MODE_Monolithic  \
                   False                          False
                   False                          False
                   False                          False
                   False                          False
                   False                          False
                     ...                            ...
                   False                          False
                   False                          False
                   False                          False
                   False                          False
```

To handle the remaining missing values, we used Imputer library

Categorical Data

```
NAME_CONTRACT_TYPE             2
CODE_GENDER                    3
FLAG_OWN_CAR                   2
FLAG_OWN_REALTY                2
NAME_TYPE_SUITE                7
NAME_INCOME_TYPE               8
NAME_EDUCATION_TYPE            5
NAME_FAMILY_STATUS             6
NAME_HOUSING_TYPE              6
OCCUPATION_TYPE               18
WEEKDAY_APPR_PROCESS_START     7
ORGANIZATION_TYPE             58
HOUSETYPE_MODE                 3
WALLSMATERIAL_MODE             7
EMERGENCYSTATE_MODE            2
dtype: int64
```

DATA MODELING

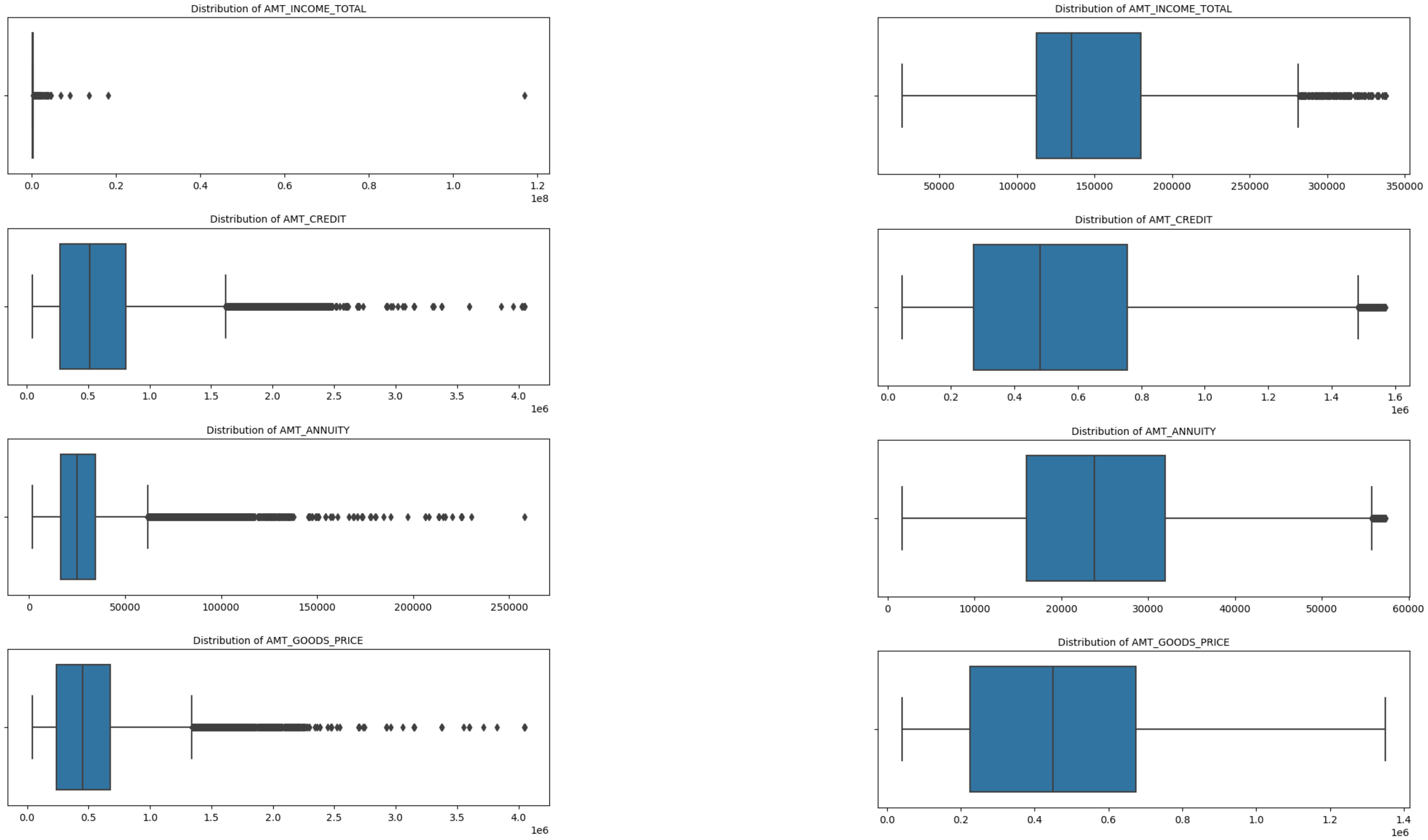
We used Logistic regression initially

Accuracy received: 0.9195

Can we do better ?

Is the accuracy we got reliable ?

DATA PREPROCESSING



DATA PREPROCESSING

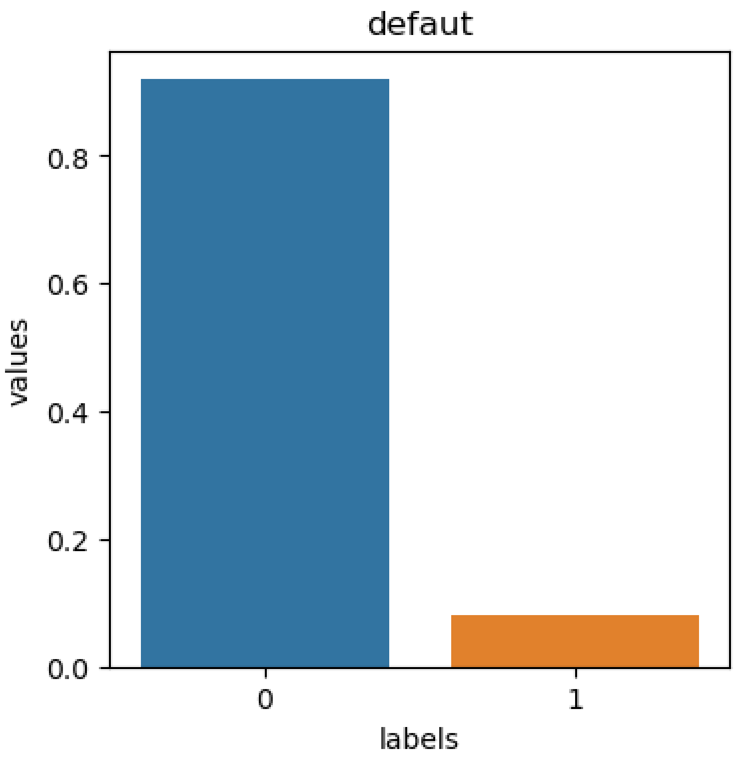
CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	...	FLAG_DOCUMENT_18	FLAG_DOCUMENT_19	FLAG_DOCUMENT_20	FLAG_
0	202500.0	406597.5	24700.5	...	0	0	0	
0	270000.0	1293502.5	35698.5	...	0	0	0	
0	67500.0	135000.0	6750.0	...	0	0	0	
0	135000.0	312682.5	29686.5	...	0	0	0	
0	121500.0	513000.0	21865.5	...	0	0	0	

After scaling the data using MinMaxScaler

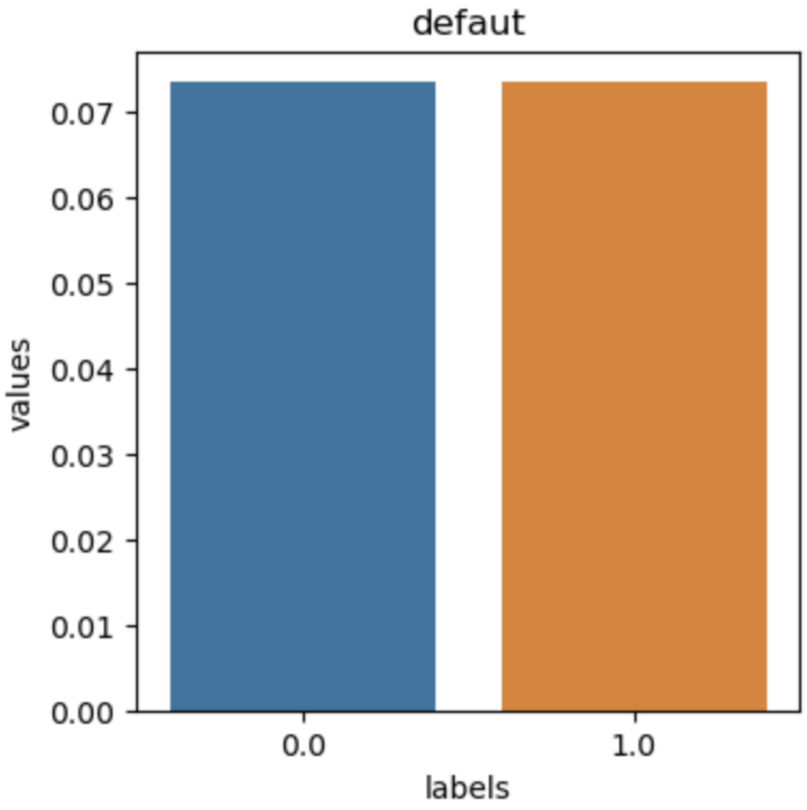
_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	...	ORGANIZATION_TYPE_XNA	HOUSETYPE_MODE_sp ho
1.0	0.0	0.567100	0.237262	0.414478	0.237113	...	False	
0.0	0.0	0.783550	0.819205	0.611942	0.831615	...	False	
1.0	0.0	0.134199	0.059053	0.092187	0.072165	...	False	
1.0	0.0	0.350649	0.175640	0.503999	0.195876	...	False	
1.0	0.0	0.307359	0.307078	0.363578	0.360825	...	False	

DATA PREPROCESSING

There is a lot of imbalance in our Target variable



After using SMOTE



Accuracy now is 0.83

HYPERPARAMETER TUNING AND MODELING

COMPARISION OF MODELS

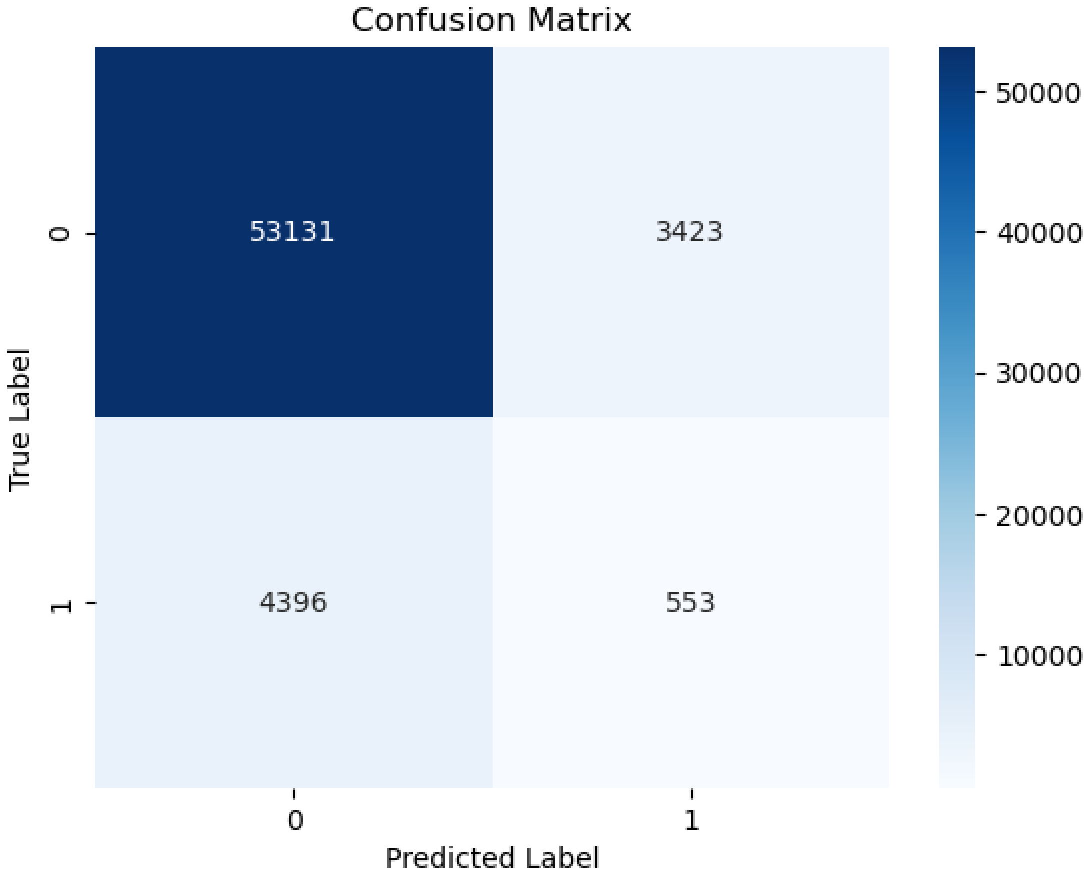
```
Best parameters for Logistic Regression: {'C': 1}
Best score for Logistic Regression: 0.8669318009807675
Best parameters for KNN: {'n_neighbors': 5}
Best score for KNN: 0.8172400105985096
Best parameters for Decision Tree: {'max_depth': 20, 'min_samples_leaf': 1, 'min_samples_split': 2}
Best score for Decision Tree: 0.8952918632477509
Best parameters for SVM: {'C': 0.001}
Best score for SVM: 0.5003161863067007
```

MODEL EVALUATION

Decision Tree

ROC AUC Score: 0.5797
Accuracy: 0.8729

	precision	recall	f1-score	support
0.0	0.92	0.94	0.93	56554
1.0	0.14	0.11	0.12	4949
accuracy			0.87	61503
macro avg	0.53	0.53	0.53	61503
weighted avg	0.86	0.87	0.87	61503

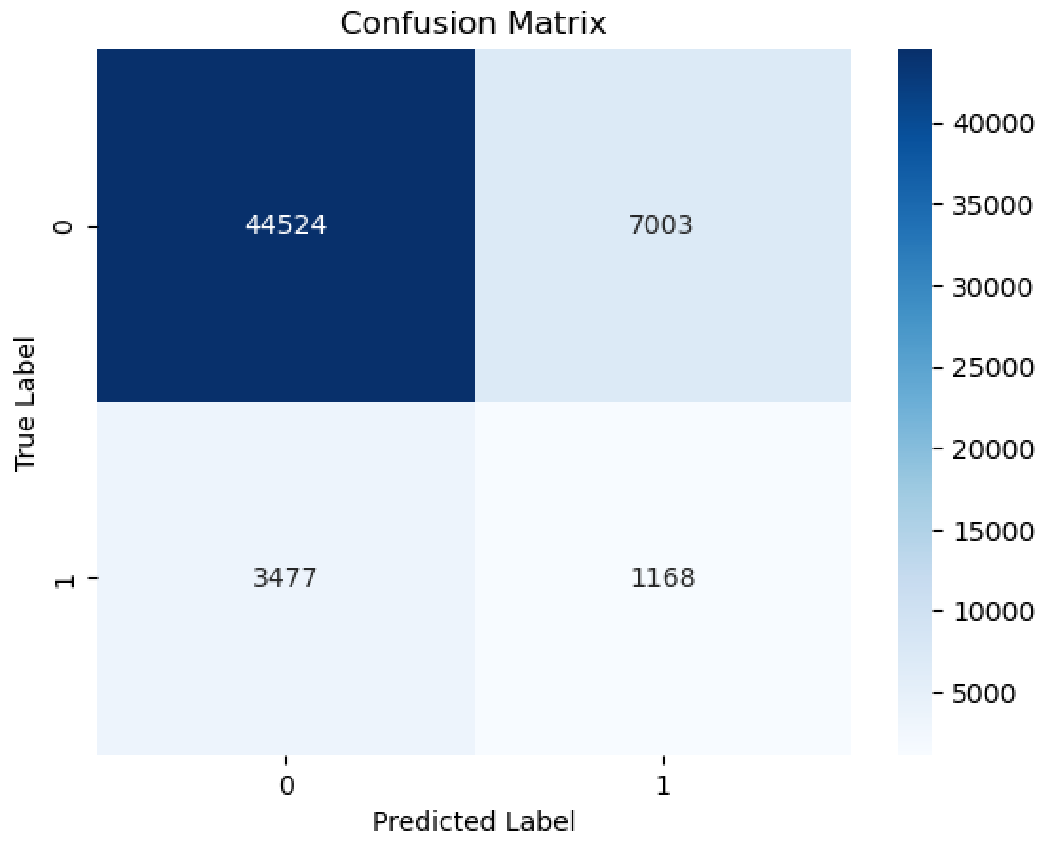


MODEL EVALUATION

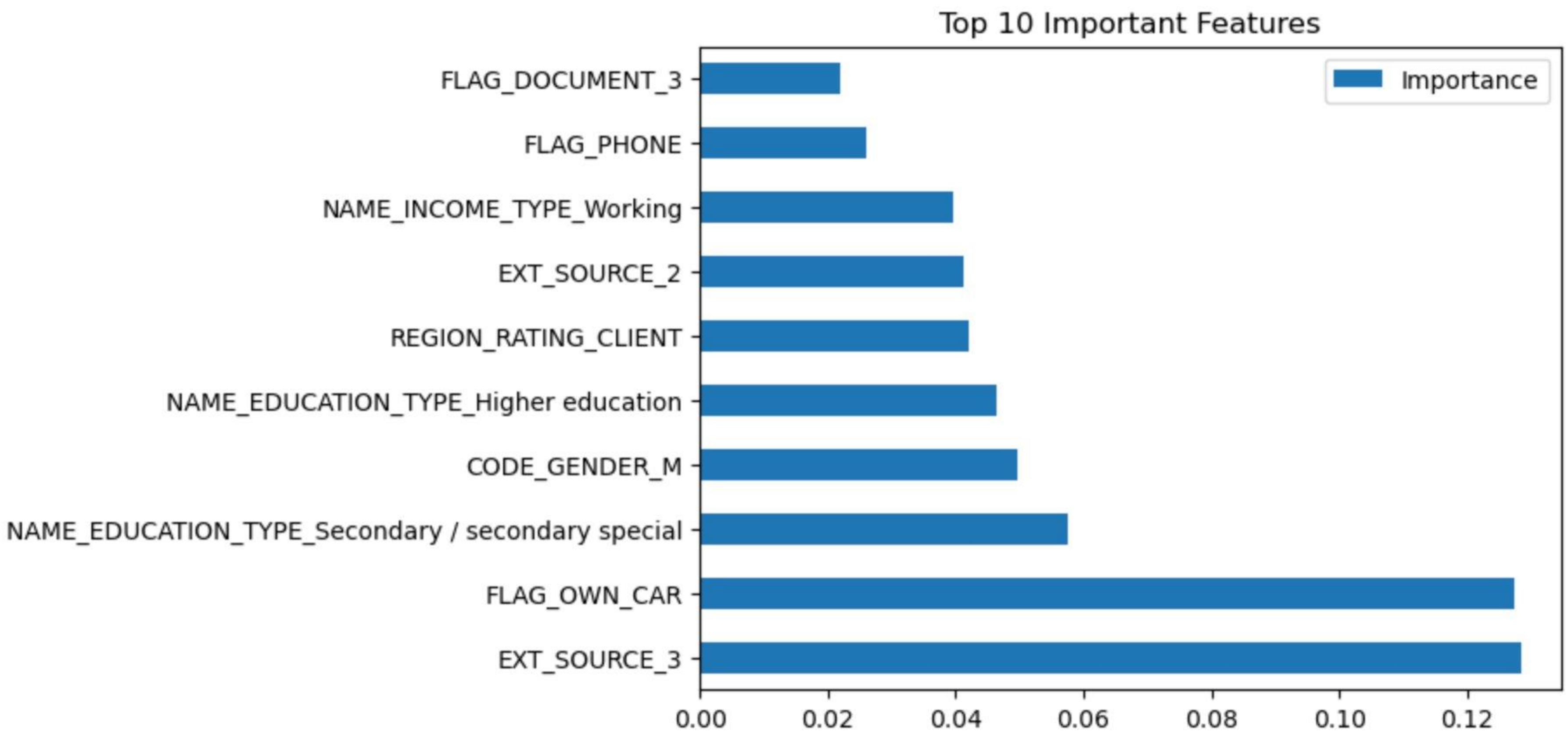
Logistic Regression

ROC AUC Score: 0.6168
Accuracy: 0.8134

	precision	recall	f1-score	support
0.0	0.93	0.86	0.89	51527
1.0	0.14	0.25	0.18	4645
accuracy			0.81	56172
macro avg	0.54	0.56	0.54	56172
weighted avg	0.86	0.81	0.84	56172



FEATURE IMPORTANCE



INSIGHTS

Key Insights: 1. The most critical predictors of loan default are EXT_SOURCE_3.

2. Applicants who are educated and weather they own a car or not are big contributing factors.

3. Implementing the model can give insights for people who don’t default.

Future Directions:

1. Enhance data collection strategies to include more relevant features.

2. Explore more sophisticated models such as Gradient Boosting or Neural Networks.

Thank You!