

# Home Credit Scorecard Model

Project Based Intern : Data Science



Created by:  
**Muhammad Ariq Arfina**

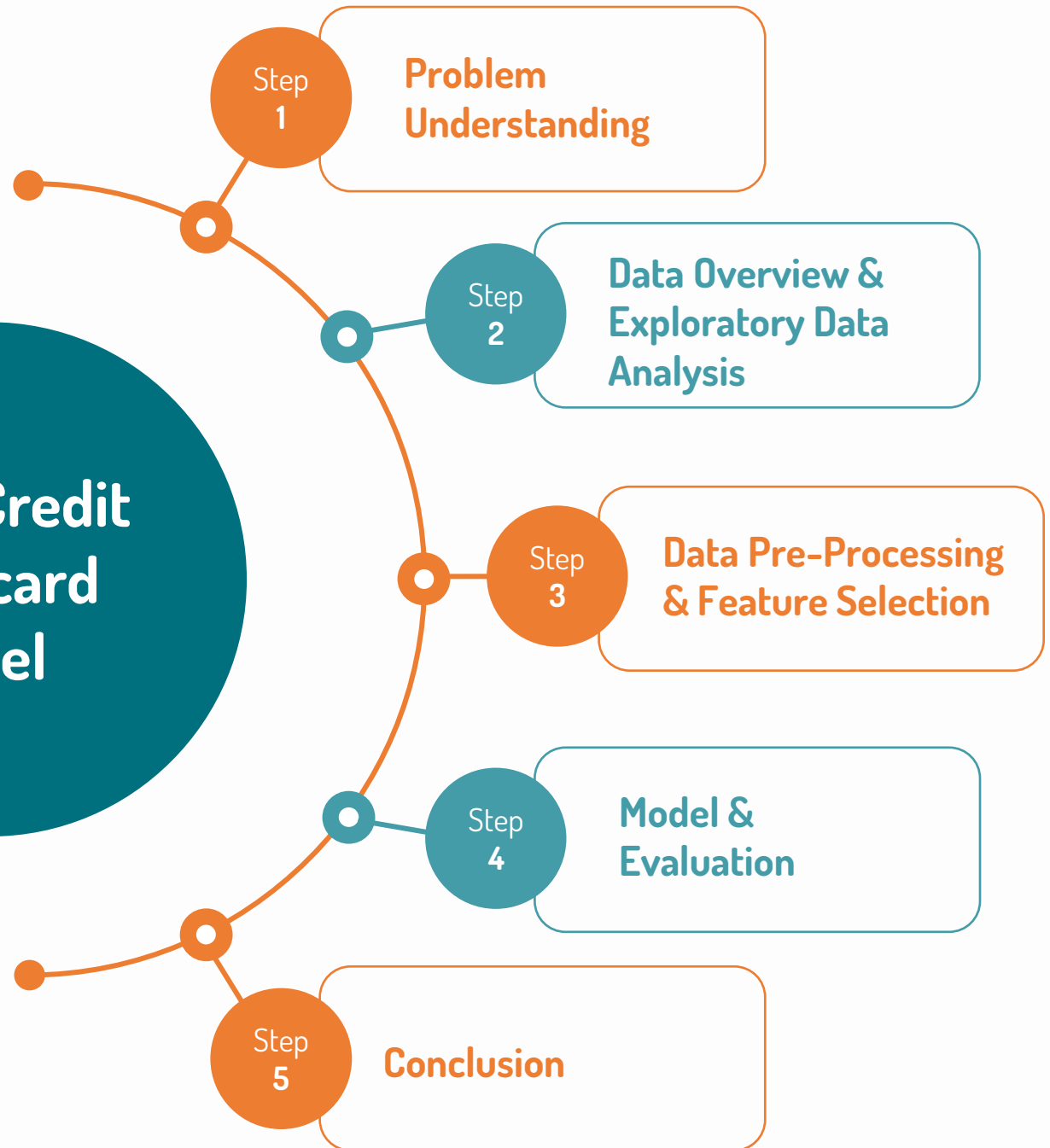
[ariqarfina05@gmail.com](mailto:ariqarfina05@gmail.com)

LinkedIn : [Muhammad Ariq Arfina](#)

Github : [ariqarfina](#)



# Outline



# Home Credit Scorecard Model

Project Based Intern : Data Science

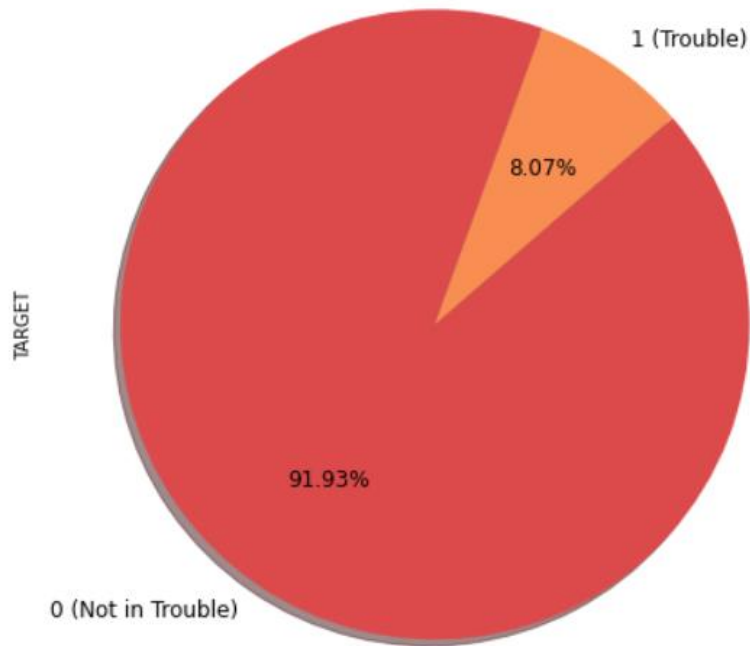
## Problem Understanding



# Problem Understanding

## 8.07% of the total borrowers have the Potential to Default

The '1' Label explain Client with payment difficulties  
ex: late payment



Problem statement:

**The loss generated by non-paying customers is quite high. (8,07%)**

Goals:

Predicting potential customers who can borrow money with minimum risk

Objective:

1. Help minimize these losses by **using predictive models to predict customers who are likely to default.**
2. Use AUC and Kolmogorov-Smirnov (KS) with target:
  - AUC : 0,7
  - KS : 0,3

Business Metrics:

Lost Given Default (LGD)

# Home Credit Scorecard Model

Project Based Intern : Data Science

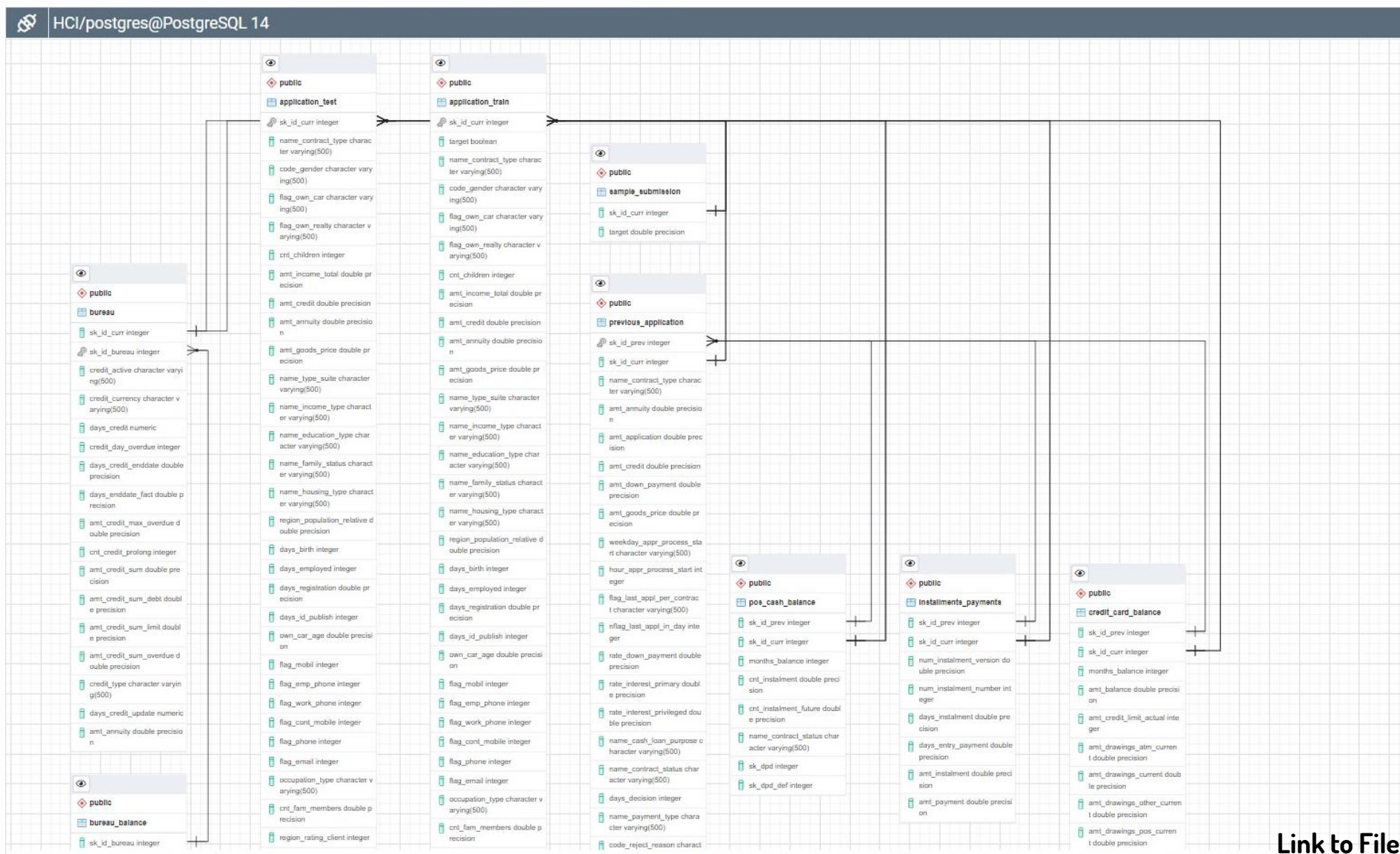
## Data Overview & Exploratory Data Analysis



# Data Overview

Table	Description		
application_{train test}.csv	<ul style="list-style-type: none"> <li>This is the main table, broken into two files for Train (with TARGET) and Test (without TARGET).</li> <li>Static data for all applications. One row represents one loan in our data sample.</li> </ul>	POS_CASH_balance.csv	<ul style="list-style-type: none"> <li>Monthly balance snapshots of previous POS (point of sales) and cash loans that the applicant had with Home Credit.</li> <li>This table has one row for each month of history of every previous credit in Home Credit (consumer credit and cash loans) related to loans in our sample – i.e. the table has (#loans in sample * # of relative previous credits * # of months in which we have some history observable for the previous credits) rows.</li> </ul>
bureau.csv	<ul style="list-style-type: none"> <li>All client's previous credits provided by other financial institutions that were reported to Credit Bureau (for clients who have a loan in our sample).</li> <li>For every loan in our sample, there are as many rows as number of credits the client had in Credit Bureau before the application date.</li> </ul>	credit_card_balance.csv	<ul style="list-style-type: none"> <li>Monthly balance snapshots of previous credit cards that the applicant has with Home Credit.</li> <li>This table has one row for each month of history of every previous credit in Home Credit (consumer credit and cash loans) related to loans in our sample – i.e. the table has (#loans in sample * # of relative previous credit cards * # of months where we have some history observable for the previous credit card) rows.</li> </ul>
bureau_balance.csv	<ul style="list-style-type: none"> <li>Monthly balances of previous credits in Credit Bureau.</li> <li>This table has one row for each month of history of every previous credit reported to Credit Bureau – i.e the table has (#loans in sample * # of relative previous credits * # of months where we have some history observable for the previous credits) rows.</li> </ul>	previous_application.csv	<ul style="list-style-type: none"> <li>All previous applications for Home Credit loans of clients who have loans in our sample.</li> <li>There is one row for each previous application related to loans in our data sample.</li> </ul>
		installments_payments.csv	<ul style="list-style-type: none"> <li>Repayment history for the previously disbursed credits in Home Credit related to the loans in our sample.</li> <li>There is one row for every payment that was made plus and one row each for missed payment.</li> <li>One row is equivalent to one payment of one installment OR one installment corresponding to</li> </ul>

# Data Overview



Link to File ->

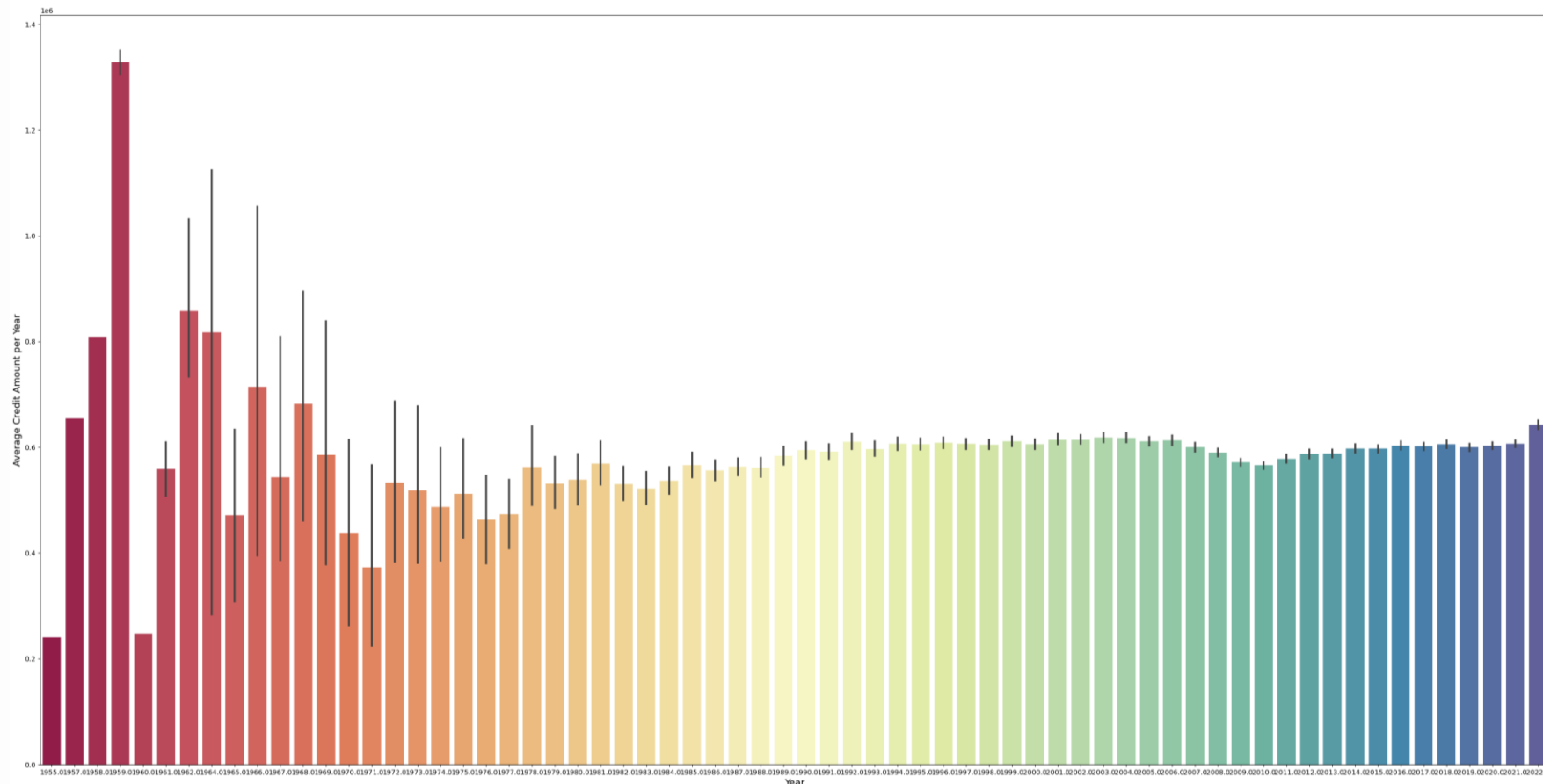




# Exploratory Data Analysis

## Generally, the nominal loan lending to customers is stagnant

The nominal credit chart has tended to stagnate since 1978 and is expected to do so until 2021. It appears to have increased by 2022. The company can investigate this further.

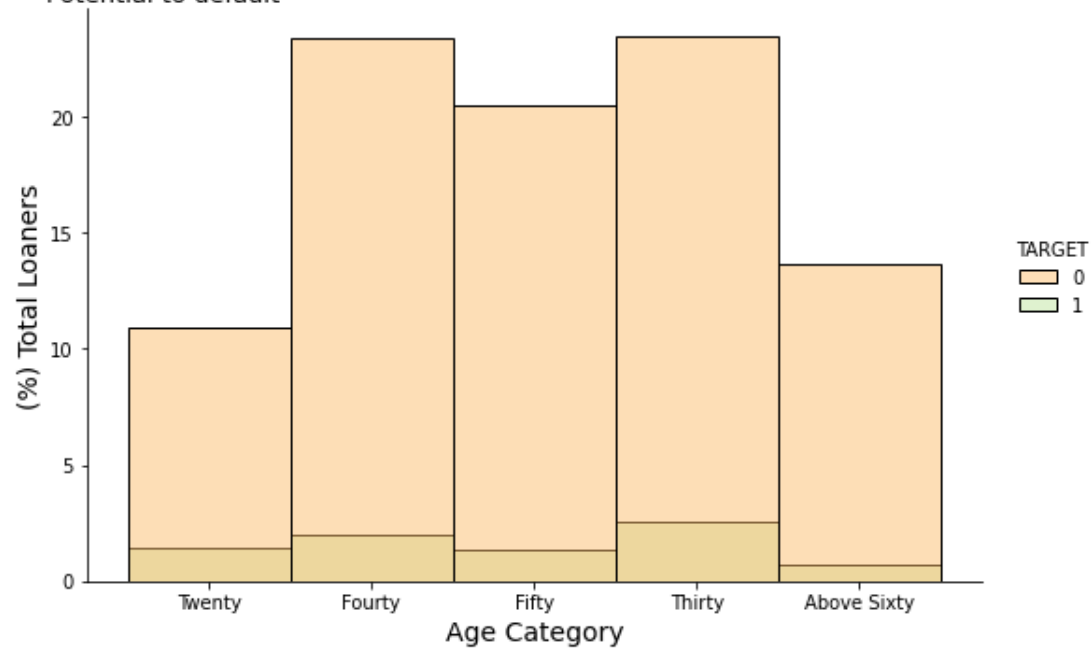




# Exploratory Data Analysis

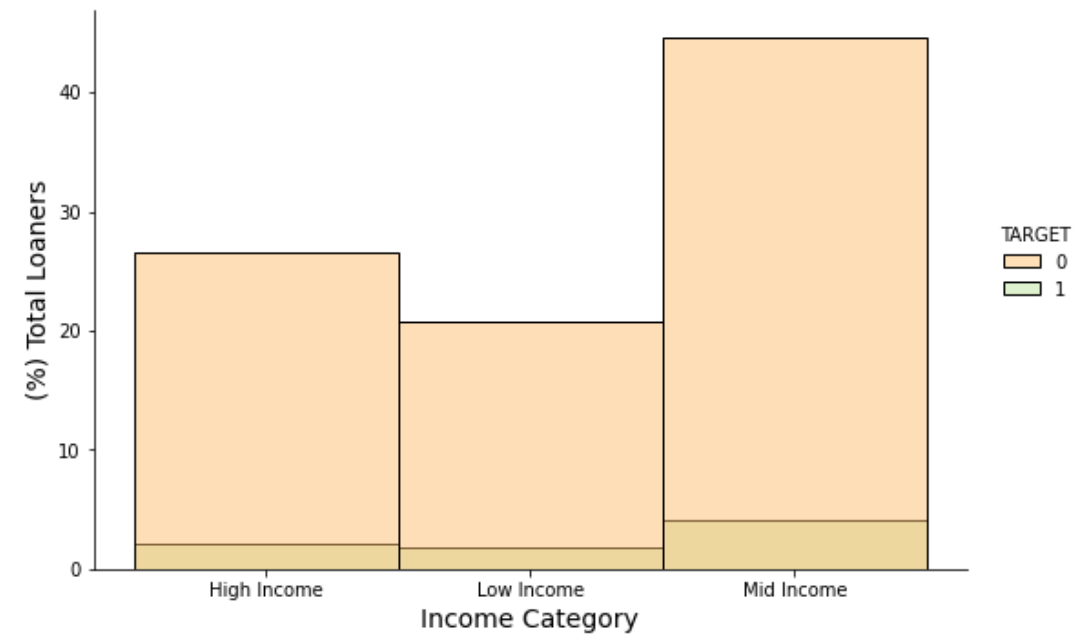
## 3% Loaners in Thirty more likely to default then others

In the graphic image below, Every Age Category is have Potential to default



## 4% Loaners in Mid Income more likely to default then others

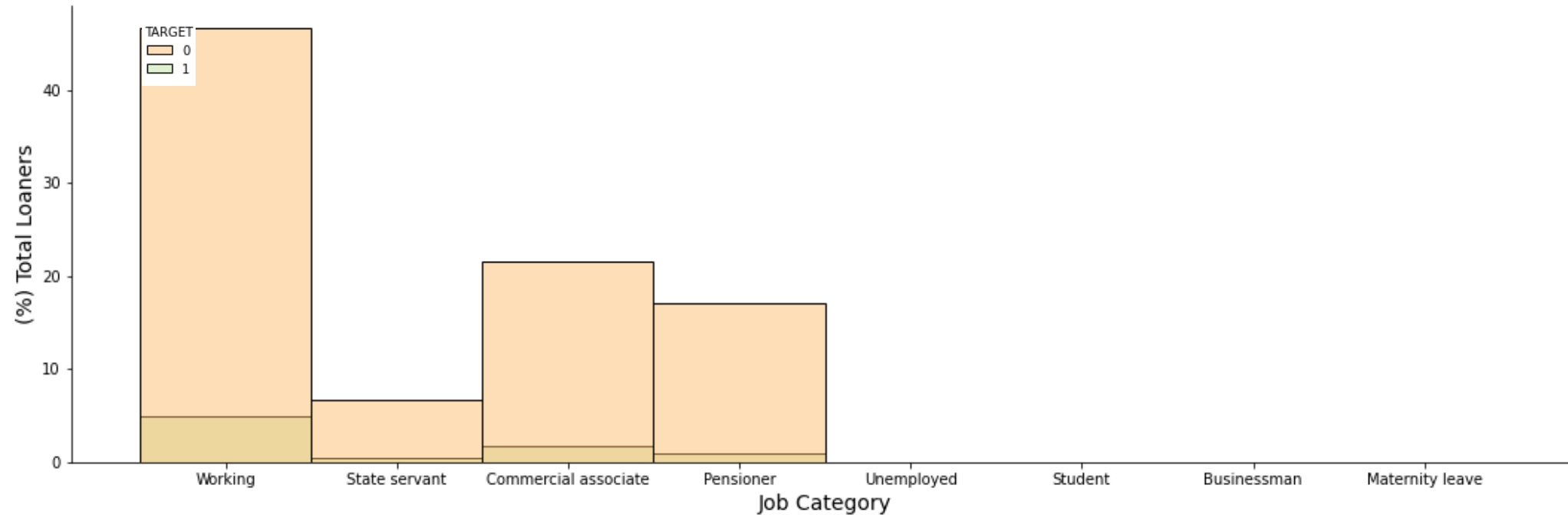
In the graphic image below, Every Income Category is have Potential to default



# Exploratory Data Analysis

## 7% Loaners in State Servant more likely to default then others

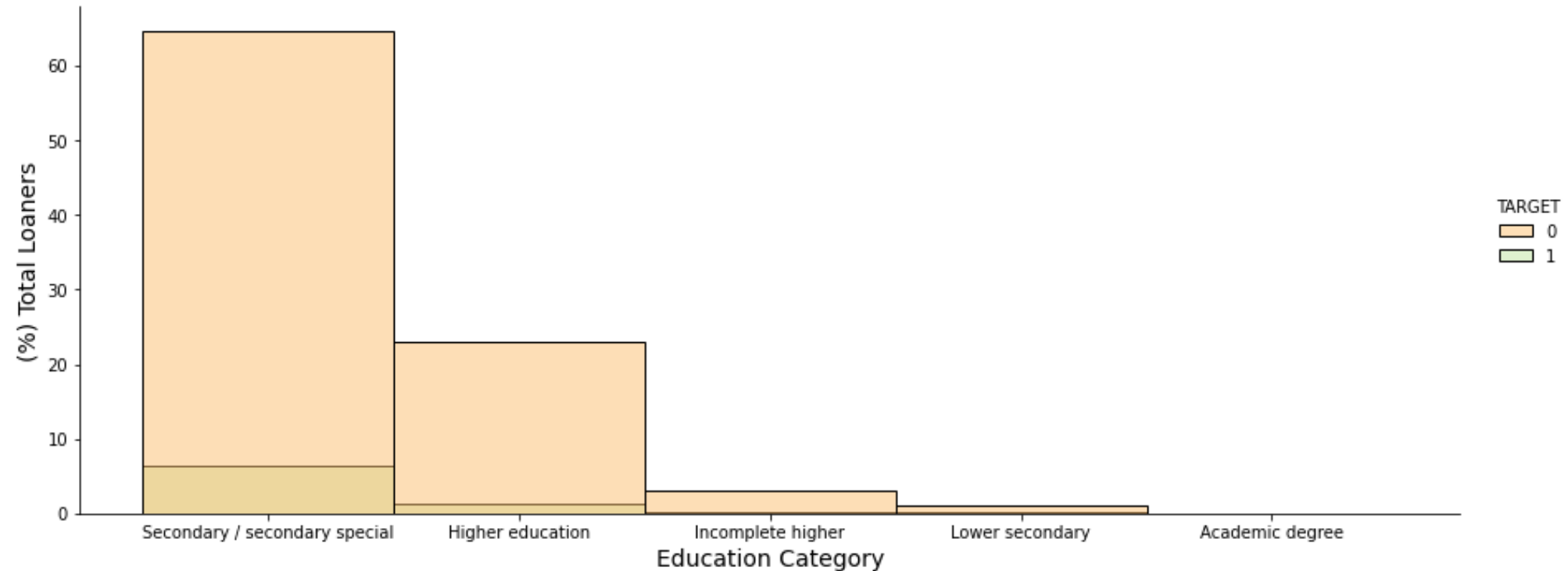
In the graphic image below, Some of loaner by job category is have Potential to default



# Exploratory Data Analysis

## 6% Loaners in Secondary / secondary special more likely to default then others

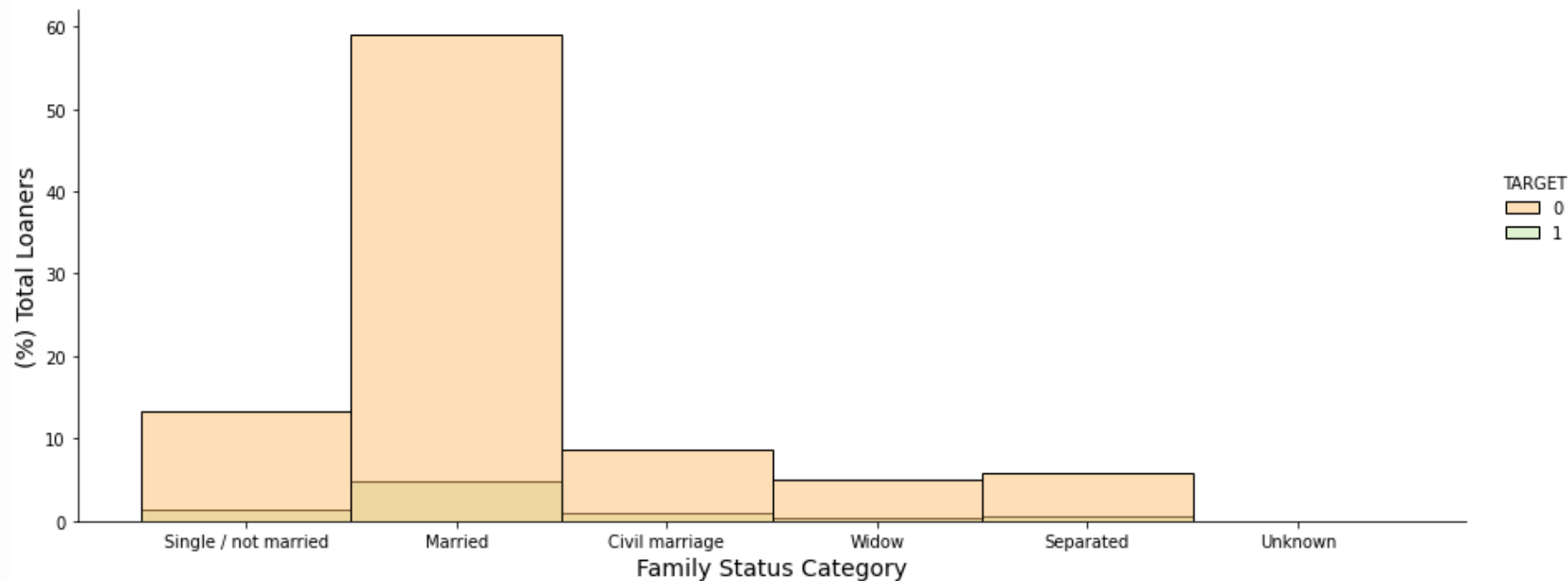
In the graphic image below, Some of loaner by edu category is have Potential to default



# Exploratory Data Analysis

## Loaner who is married is more likely to default then others

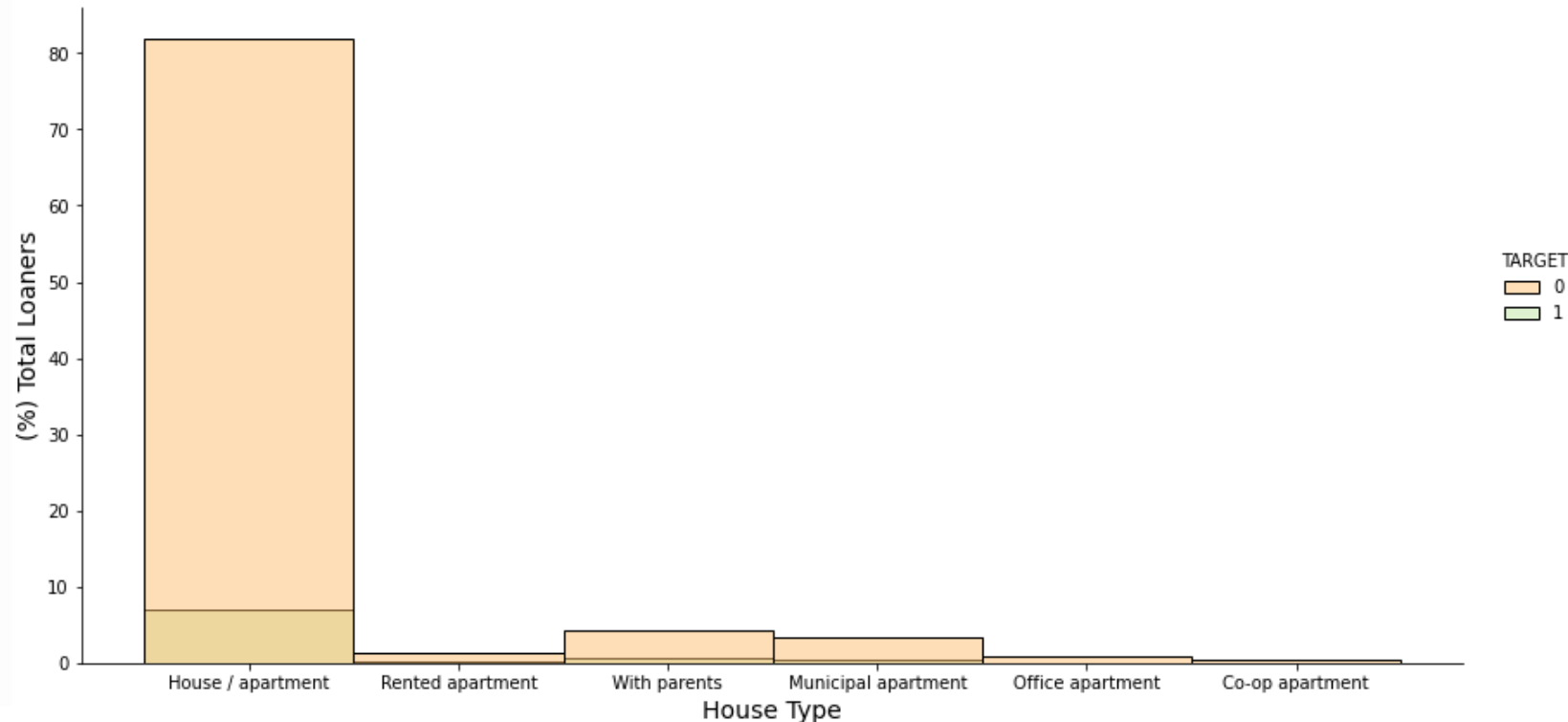
In the graphic image below, Some of loaner by family status category is have Potential to default



# Exploratory Data Analysis

## Loaner who have house is more likely to default then others

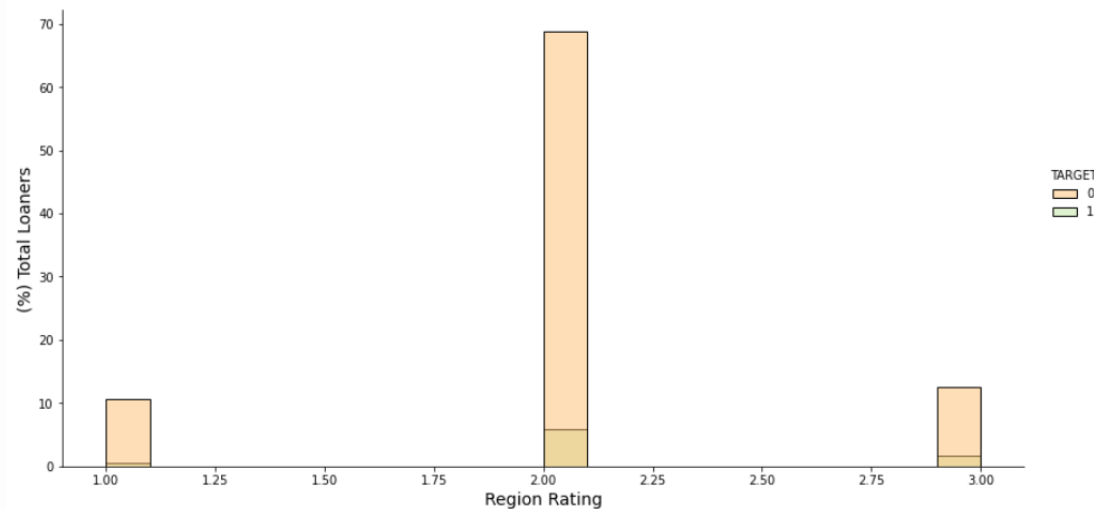
In the graphic image below, Some of loaner who have house (7%) is have Potential to default



# Exploratory Data Analysis

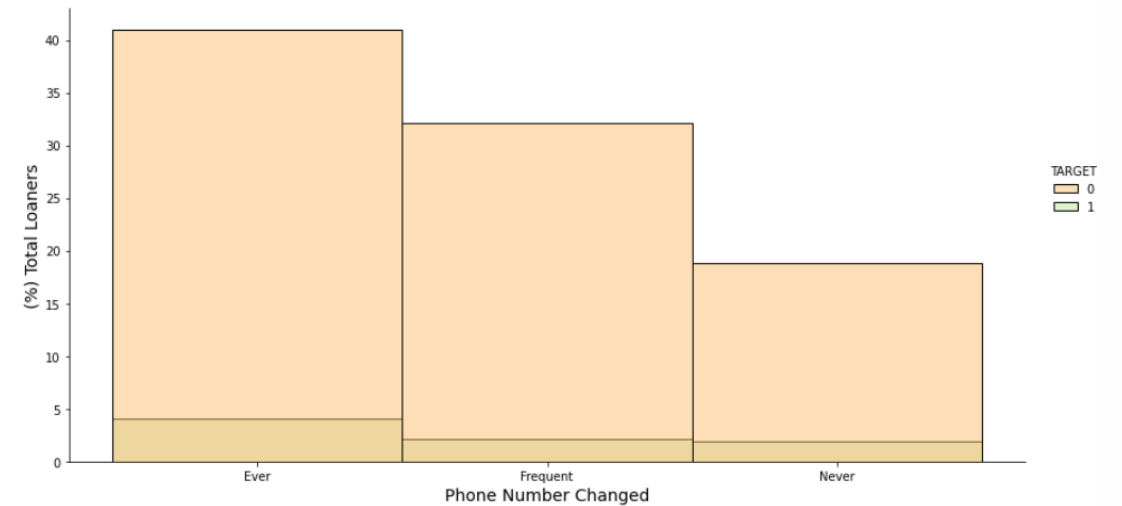
## Loaner who in 2 more likely to default then others

In the graphic image below, Some of loaner with rating 2 (6%) is have more Potential to default



## Loaner who ever change their phone more likely to default then others

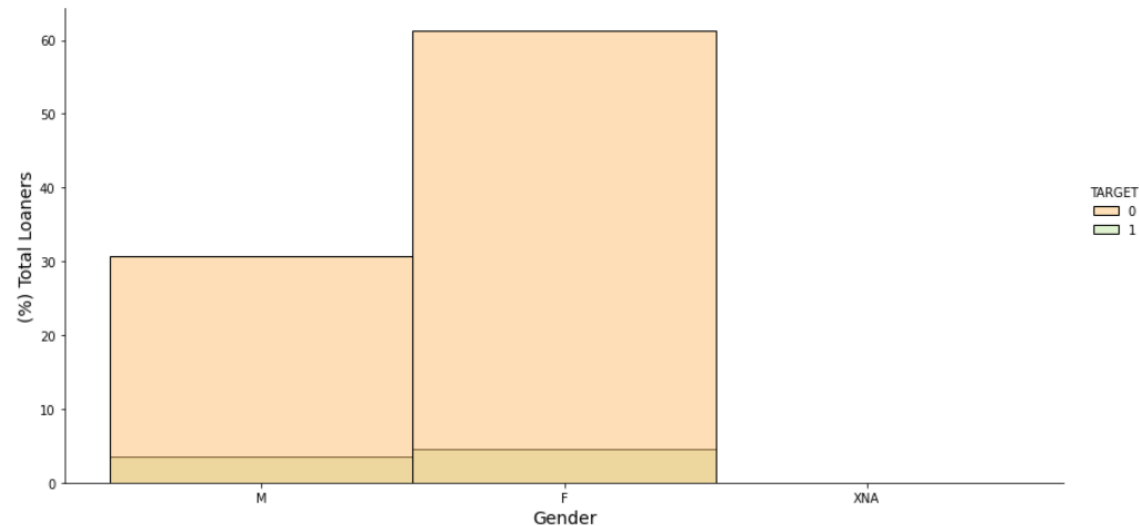
Some of loaner who ever change their phone (1-3 times) is have more Potential to default (4%)



# Exploratory Data Analysis

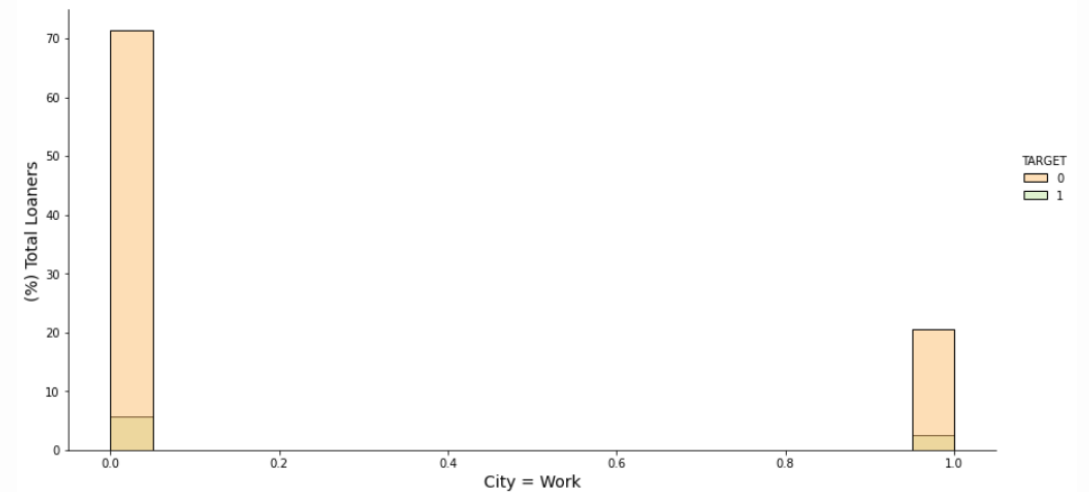
## Female Loaners is have more potential to default

14170 Female Loaners is have more Potential to default (5%) then the Male Loaners



## Loaners who have an address that is not the same as their place of work is have more potential to default

17305 Loaners who have an address that is not the same as their place of work is have more Potential to default (6%)

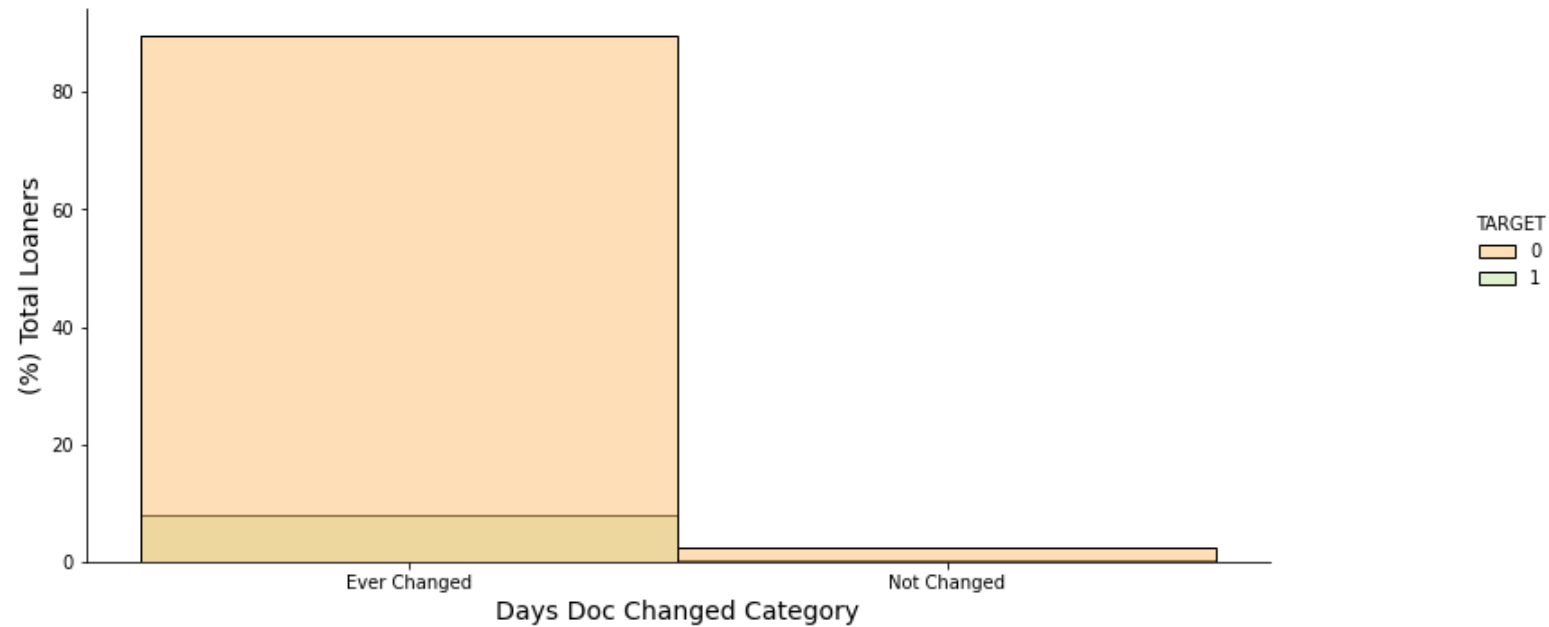




# Exploratory Data Analysis

## Loaner who changed the document in 'Ever Changed' more likely to default then others

24035 Total Loaners who ever changed the identity Document  
have more Potential to default (8%)



# Home Credit Scorecard Model

Project Based Intern : Data Science

## Data Pre-Processing & Feature Selection



# Data Pre-Processing & Exploratory Data Analysis

## Drop Missing Value (<68.4% from Dataset)

	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



## Input Missing Value with Single Imputer

```
#create two DataFrames, one for each data type
data_numeric = df_1[nums]
data_categorical = pd.DataFrame(df_1[cats])

from sklearn.impute import SimpleImputer
imp = SimpleImputer(missing_values=np.nan, strategy='median', verbose=0)
data_numeric = pd.DataFrame(imp.fit_transform(data_numeric), columns = data_numeric.columns)

cimp = SimpleImputer(missing_values=np.nan, strategy='most_frequent', verbose=0)
data_categorical = pd.DataFrame(cimp.fit_transform(data_categorical), columns = data_categorical.columns)

#you could do something like one-hot-encoding of data_categorical here

#join the two masked dataframes back together
df_1 = pd.concat([data_numeric, data_categorical], axis = 1)
```

# Data Pre-Processing & Exploratory Data Analysis

## Feature Encoding

```
cats_le = df_for_model[['NAME_CONTRACT_TYPE', 'CODE_GENDER', 'FLAG_OWN_CAR',  
                        'FLAG_OWN_REALTY', 'EMERGENCYSTATE_MODE',  
                        'NAME_EDUCATION_TYPE', 'AMT_INCOME_CAT',  
                        'DAYS_ID_PUBLISH_CAT', 'PHONE_CHANGE_CAT']]  
  
cats_oh = df_for_model[['NAME_TYPE_SUITE', 'NAME_INCOME_TYPE',  
                        'NAME_FAMILY_STATUS', 'OCCUPATION_TYPE',  
                        'WEEKDAY_APPR_PROCESS_START', 'ORGANIZATION_TYPE',  
                        'WALLSMATERIAL_MODE', 'HOUSETYPE_MODE',  
                        'AGE_CATEGORY', 'NAME_HOUSING_TYPE']]  
  
for i in cats_le:  
    df_for_model[i] = LabelEncoder().fit_transform(df_for_model[i])  
for cat in cats_oh:  
    onehots = pd.get_dummies(df_for_model[cat], prefix=cat)  
    df_for_model = df_for_model.join(onehots)
```

## MinMaxScaler

```
from sklearn.preprocessing import MinMaxScaler  
for i in numss:  
    df_for_model[i] = MinMaxScaler().fit_transform(df_for_model[i])
```

## Feature Selection using SelectKBest

```
x_k = df_for_model[[col for col in df_for_model.columns if col not in ['TARGET']]]  
y_k = df_for_model['TARGET']  
  
print("Feature data dimension: ", x_k.shape)
```

Feature data dimension: (307511, 236)

```
select = SelectKBest(score_func=f_regression, k=15)  
z = select.fit_transform(x_k, y_k)  
  
print("After selecting best 15 features:", z.shape)
```

After selecting best 15 features: (307511, 15)

Selected best 15:

```
['DAYS_BIRTH' 'DAYS_ID_PUBLISH' 'REGION_RATING_CLIENT'  
 'REGION_RATING_CLIENT_W_CITY' 'REG_CITY_NOT_WORK_CITY' 'EXT_SOURCE_1'  
 'EXT_SOURCE_2' 'EXT_SOURCE_3' 'DAYS_LAST_PHONE_CHANGE' 'AGE'  
 'DAYS_LAST_PHONE_CHANGE_FIXED' 'DAYS_ID_PUBLISH_FIXED' 'CODE_GENDER'  
 'NAME_EDUCATION_TYPE' 'NAME_INCOME_TYPE_Working']
```

# Home Credit Scorecard Model

Project Based Intern : Data Science

## Model & Evaluation



# Model & Evaluation

## Train Test Split & Handling Imbalance Data

### Feature :

'DAYS\_BIRTH'  
'DAYS\_ID\_PUBLISH'  
'REGION\_RATING\_CLIENT'  
'REGION\_RATING\_CLIENT\_W\_CITY'  
'REG\_CITY\_NOT\_WORK\_CITY'  
'EXT\_SOURCE\_1'  
'EXT\_SOURCE\_2'  
'EXT\_SOURCE\_3'  
'DAYS\_LAST\_PHONE\_CHANGE'  
'CODE\_GENDER'  
'NAME\_EDUCATION\_TYPE'  
'NAME\_INCOME\_TYPE\_Working'

### Target :

'TARGET'

### Imbalance using SMOTE

```
X, y = over_sampling.SMOTE(0.5).fit_resample(X, y)

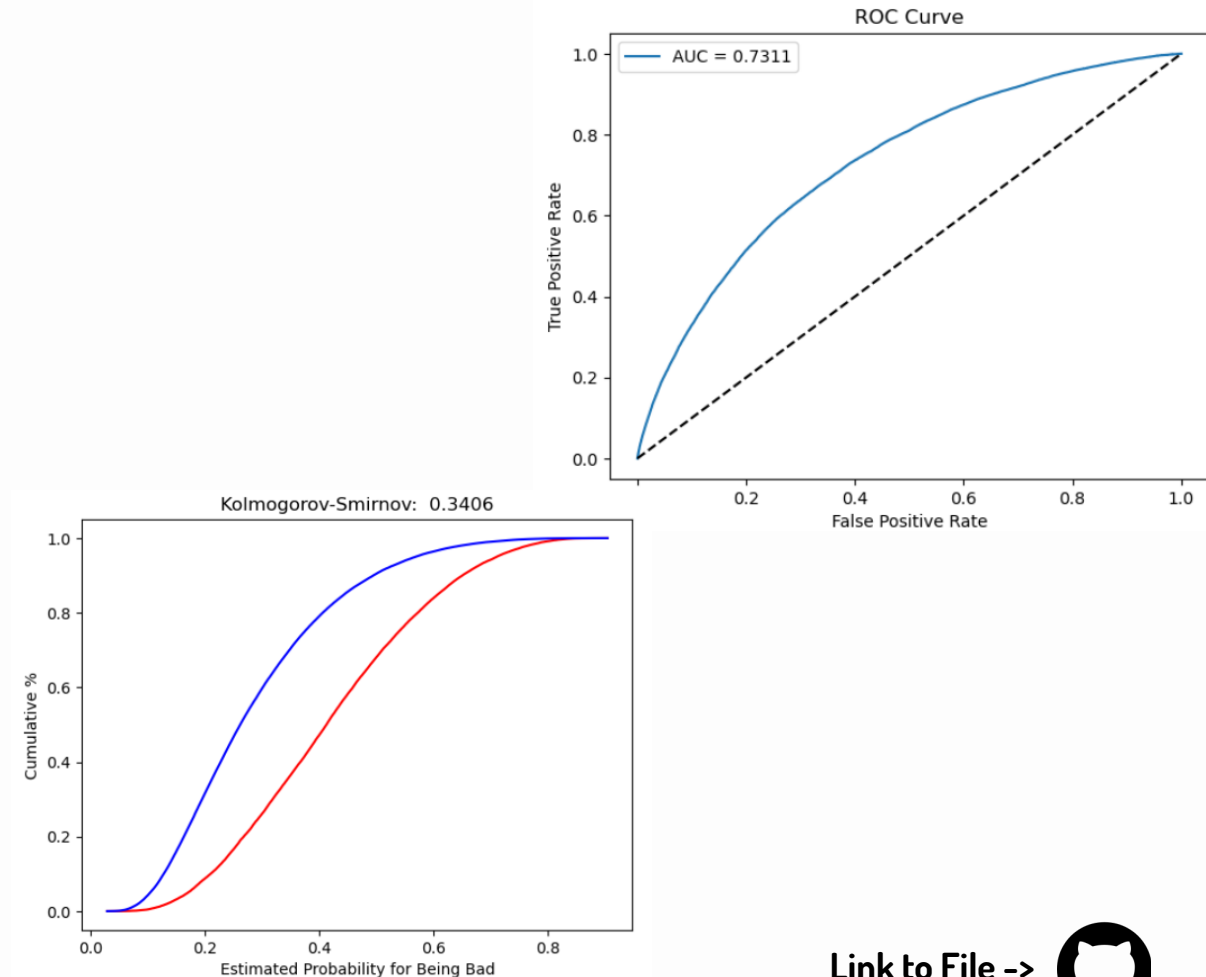
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
```

# Model & Evaluation

**Model :**  
**Logistic Regression**

**Model Evaluation:**  
**AUC & KS**

Evaluation	Model
	Logistic Regression
AUC (Train Set)	0.73
AUC (Test Set)	0.73
KS (Test Set)	0.3406





# Home Credit Scorecard Model

Project Based Intern : Data Science

## Conclusion



# Conclusion

## Before and After using model:

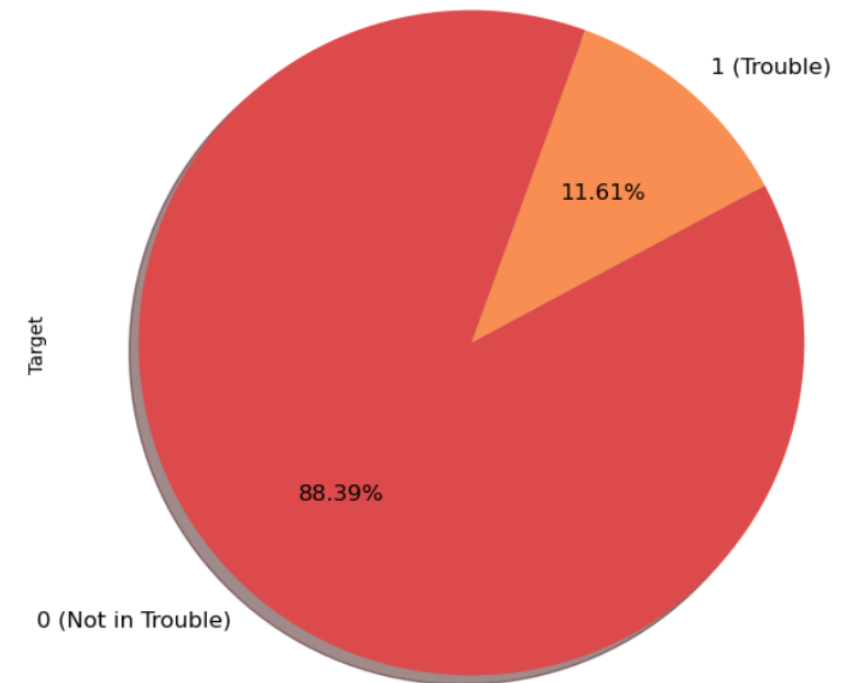
Before we use the model, we know that **8,07 %** of Customers indicated failure from data train.

We can predict that **11.61 %** of the Test Data will default based on our modeling. As a result, customers who are expected to fail to pay can have their **applications canceled**.

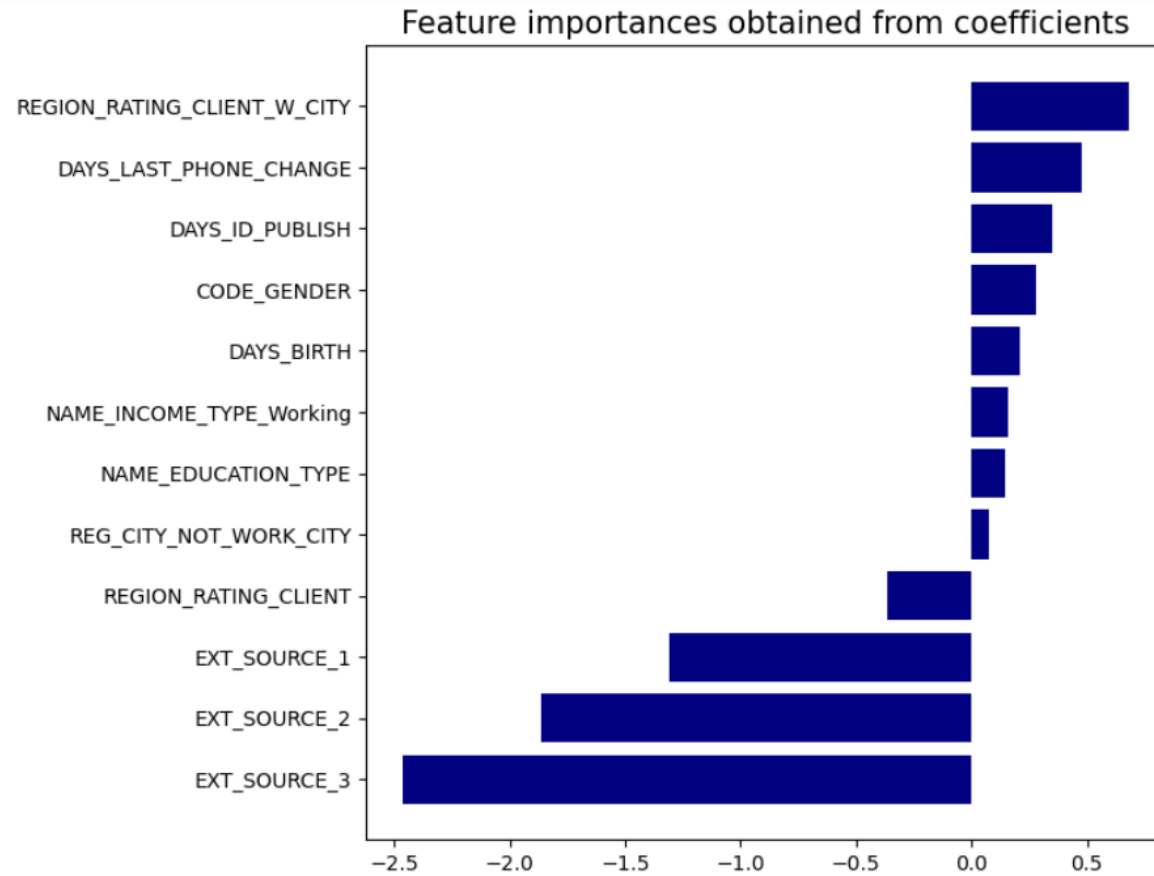
Loss generate (Before using Model) Using Previous Data	Loss generate (After using Model) Using Data Test
8,07% Default	11,61% Predicted Default

## After using Logistic Regression Model

We can reduce 11,61% potential to default



# Feature Importance



**Top 3 Feature Importance that give more impact is:**

- Region Rating Client W City
- Days Last Phone Change
- Days ID Publish

# Credit Score Card

## Credit Score Card with Logistic Regression Model (Based on Feature Importance and EDA):

1. The most impactful feature in the following model is "**Region Rating Client City**." According to EDA, the value in this column represents a rating of the city in which the customer resides. It should be noted that those who live in a "2" rating area are more vulnerable to credit default.
2. The "**Days Last Phone Change**" feature is the second most important. In EDA, we divide the customer's phone number changes into three categories: "Ever," "Once," and "Never." According to the EDA, people who ever changed their phone number are more likely to default.
3. One of the top three features in this dataset is "**Days ID Publish**." We categorize people into two groups: those who have not changed their identity documents (Not Changed) and those who have (Ever Changed). This needs to be looked into further because it could be a scam.
4. If some of the top three features are insufficient, features that are not in the top three become additional features.



# Thank You

Created by:

**Muhammad Ariq Arfina**

[ariqarfina05@gmail.com](mailto:ariqarfina05@gmail.com)

LinkedIn : [Muhammad Ariq Arfina](#)

Github : [ariqarfina](#)

## Home Credit Scorecard Model

Project Based Intern : Data Science