VIRTUAL INTERNSHIP PROGRAM

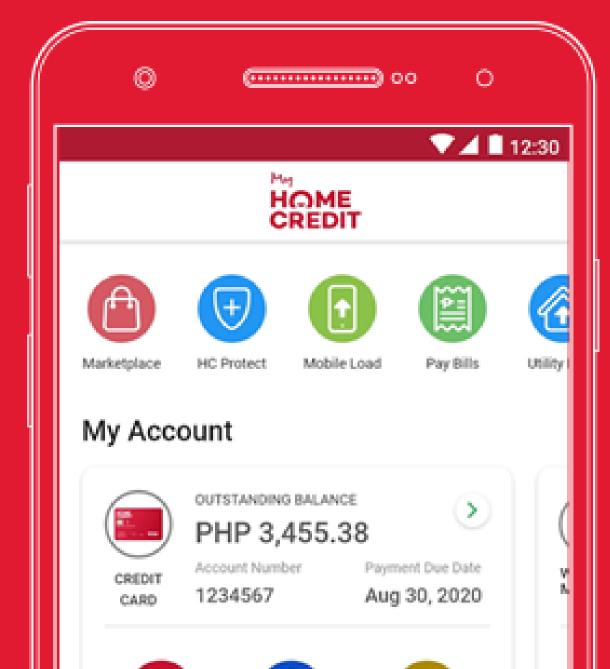
# HOME CREDIT

Oky Hariawan

**Get started** 

Let's Go

## HCME CREDIT Para sa life!



#### **Home Credit**

## **PROBLEM**

Many people struggle to get loans due to insufficient or non-existent credit histories. And, unfortunately, this population is often taken advantage of by untrustworthy lenders.

From the data Home Credit Indonesia has suffered Rp. 13.846.851.949 losses from customers that failed to pay their debt which covers 8.07% of total customers.

The reason for this is the wrong decision when reviewing the credit applications.



#### Virtual Internship Home Credit

## Overview

## 1.Role

As a Data Scientist Intern at Home Credit Indonesia, will provide data visualization to give insight and machine learning solutions

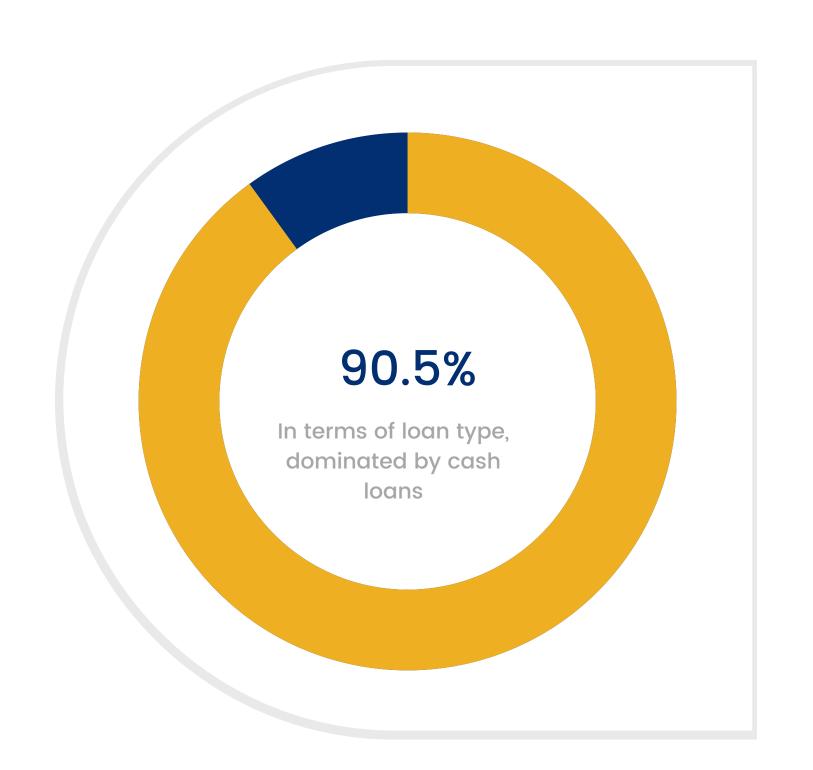
## 2. Objective

Using historical loan application data to predict whether or not an applicant will be able to repay a loan

## 3. Business Metrics

**Default Rate** 

Highlighted Goal



# Type of Loans

90.5% Cash Loans

A cash loan is a loan which is received by the borrower in cash

9.52% Revolving Loans

A committed loan facility allowing a borrower to borrow (up to a limit), repay and re-borrow loans

# Data Prepocessing



01

#### **Handle Missing Value**

Drop feature with missing value >30% & impute several feature

02

#### **Feature Creation**

Add 4 new feature. Debt to income ratio (DIR), Annuity ratio (AIR), Annuity to credit ratio (ACR), Days employed (DAR)

03

#### **Feature Extraction**

Extract new feature to reduce the dimensionality. AVG\_FLAG\_DOCS & AVG\_EXT\_SOURCE

04

#### **Feature Selection**

Select the most important feature from huge number of features

05

#### **Feature Encoding**

Encode categorical feature into binary using one-hot encoding & label encoding

# Modelling

Model	Accuracy	AUC	Recall	Prec.	Fl
Logistic Regression	0.6790	0.7224	0.6444	0.1511	0.2448
Random Forest Classifier	0.9187	0.6923	0.0059	0.3156	0.0115
Ada Boost Classifier	0.8643	0.6534	0.1613	0.1606	0.1605
K Neighbors Classifier	0.7016	0.5843	0.3952	0.1134	0.1762
Decision Tree Classifier	0.8376	0.5292	0.1613	0.1209	0.1382

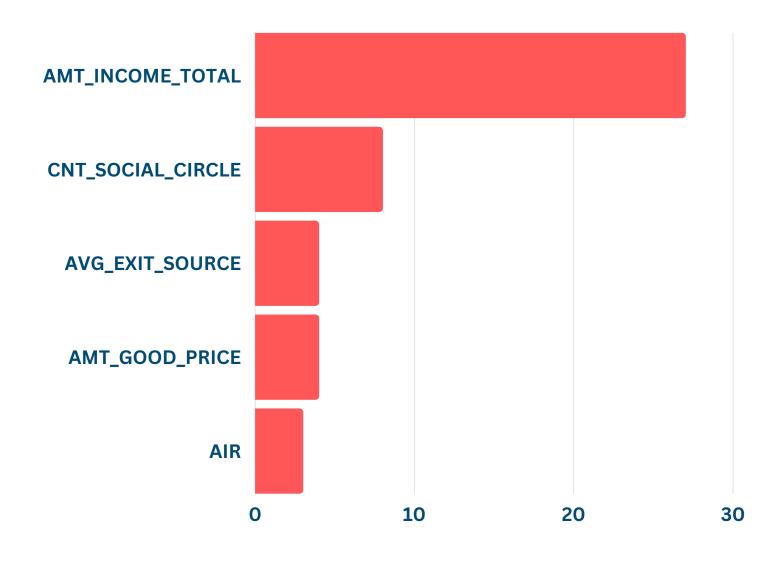


Metrics

**ROC AUC** 

#### **Business Product Overview**

Logistic Regression have the best result in terms of AUC score with 0.72 & area under ROC (It tells how much model is capable of distinguishing between classes)



# Feature Importance

After modelling we can see that these 5 feature have importance to model performance.

01 AMT\_INCOME\_TOTAL

This feature contains Income of the client

02 DEF\_30\_CNT\_SOCIAL \_CIRCLE

How many observation of client's social surroundings defaulted on 30 DPD (days past due) 03 AVG\_EXIT\_SOURCE

Feature extraction contains normalized score from external data source

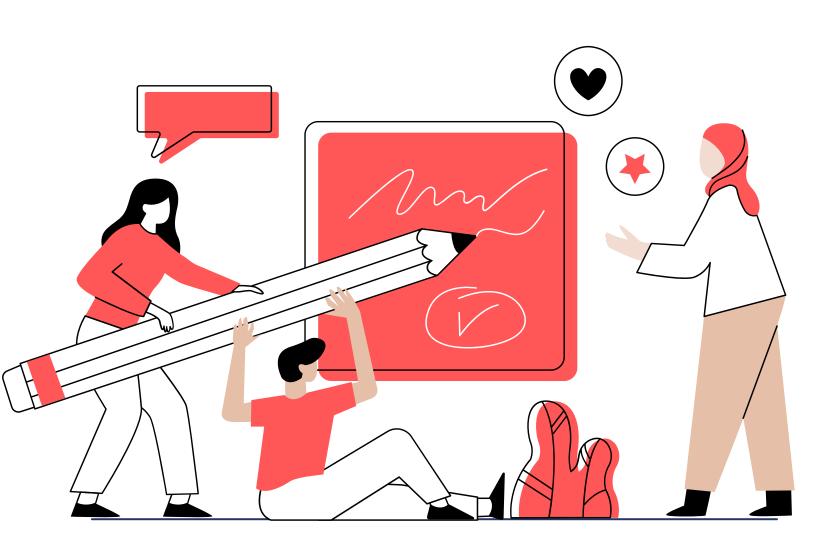
04 AMT\_GOOD\_PRICE

Goods price of good that client asked for (if applicable) on the previous application

**05** AIR

AIR (Annuity to income ratio).
Feature creation contains
calculation of Loan Annuity/
Total Income

# Business Recommendation



## **Technical Finding**

On this dataset only small features contains informative value & the others are still undefinied

#### Recommendation

Use model machine learning & feature importance for guidance in future project also implementation

### **Next Step**

More feature engineering could be done. Advance technique SMOTE could be implement to handle class imbalance. Try on complexed model such as Neural Network



## THANK YOU





**GitHub Repositories**