VIRTUAL INTERNSHIP EXPERIENCE (VIX)

by Archie Citra Muhammad

PROBLEM RESEARCH

- Problem : the number of customers defaults reached 50,968 people and the company suffered a loss of $743,972,450 in 2007 to 2014

- Goal : reduce company losses by up to 30 percent of bad loan

- Objective : Create a ML system to help loan assessments automatically

- Business Metrics : cost

DATASET

01

Loan_data_2007_2014 have 466285 rows and 74 columns

- Float64 : 46
- Int64 : 6
- Object : 22

02

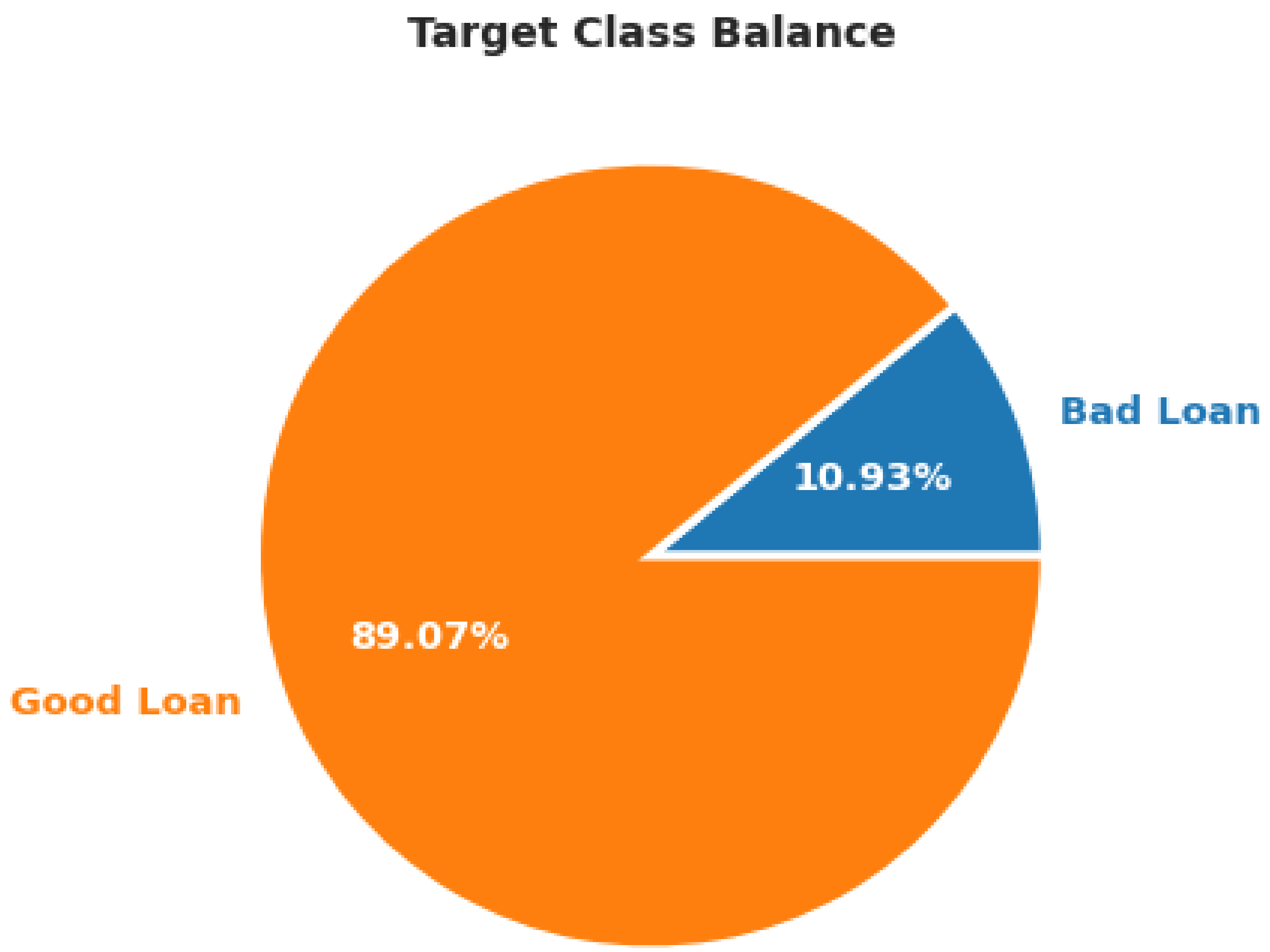
Target from loan_status column

- Current 0.480878
- Fully Paid 0.396193
- In Grace Period 0.006747
- Does not meet the credit policy. Status:Fully Paid 0.004263
- Late (16-30 days) 0.002612

Good loan: 89%

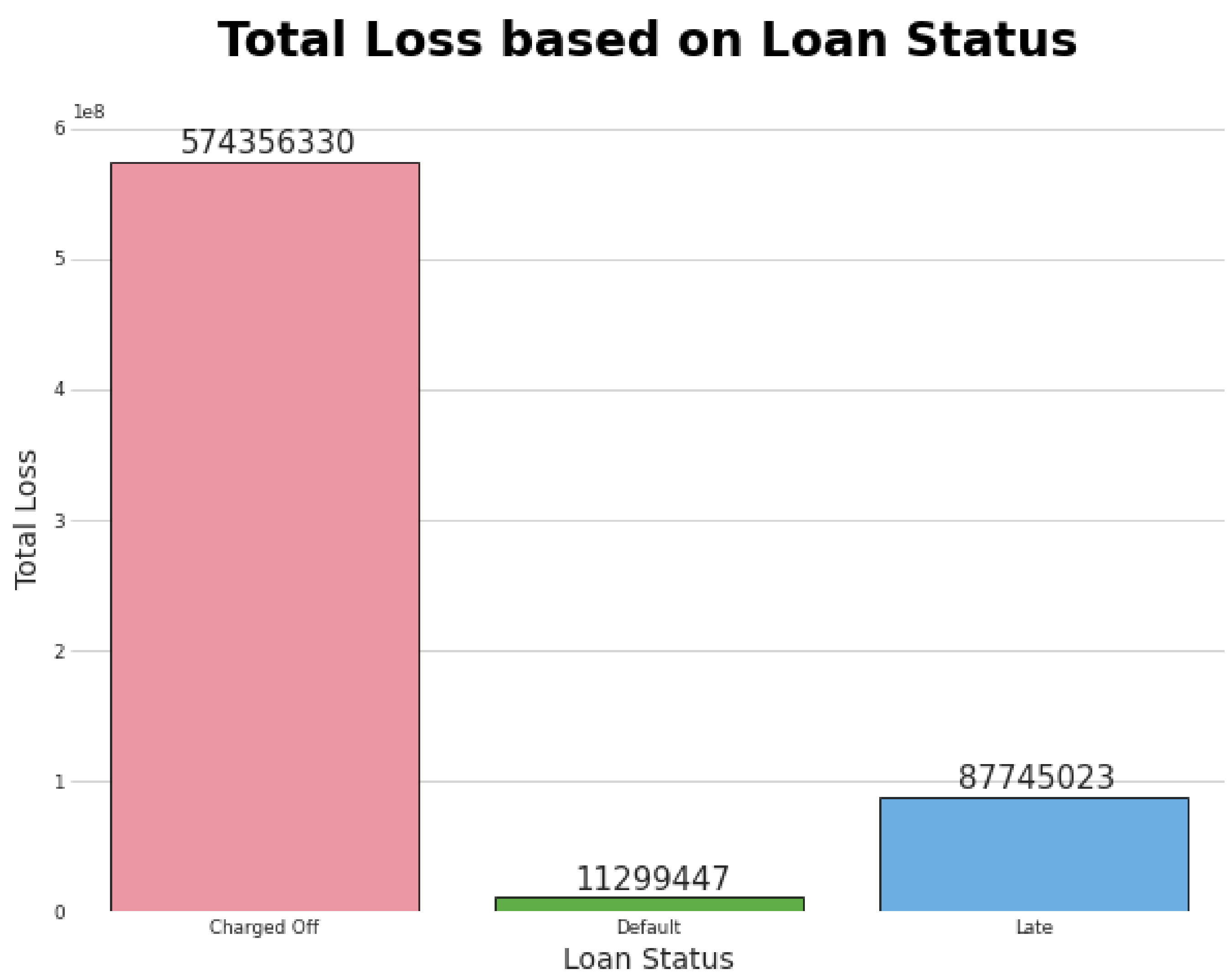
- Charged Off 0.091092
- Late (31-120 days) 0.014798
- Default 0.001784
- Does not meet the credit policy. Status:Charged Off 0.001632

Bad loan: 11%



https://github.com/archie-cm/Credit_Score_Home_Credit_Indonesia

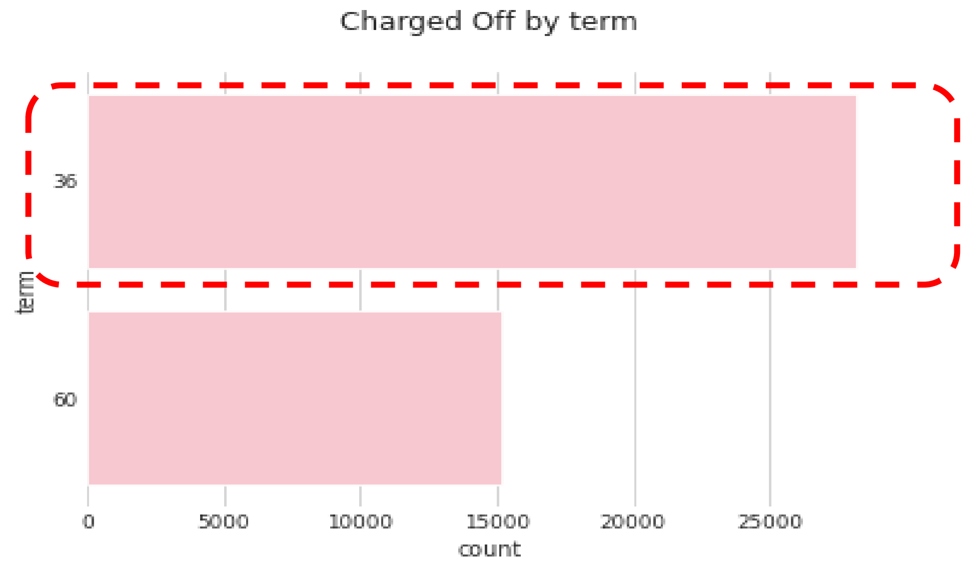
DATA VIZ AND BUSINESS INSIGHT

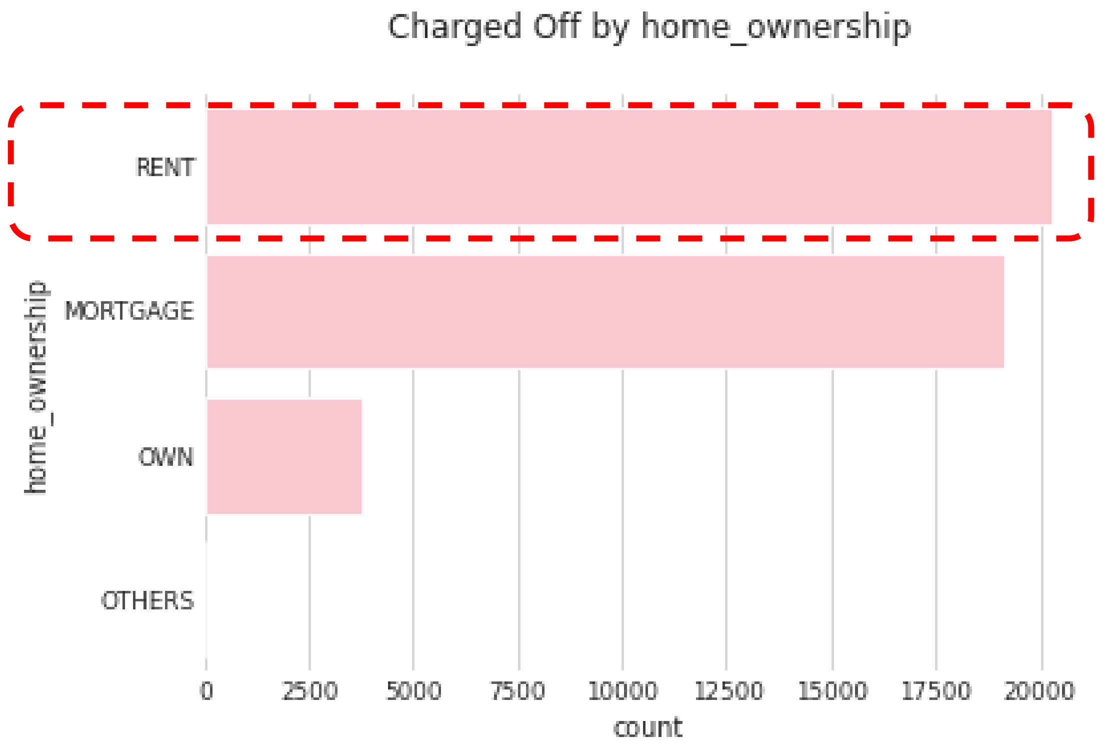


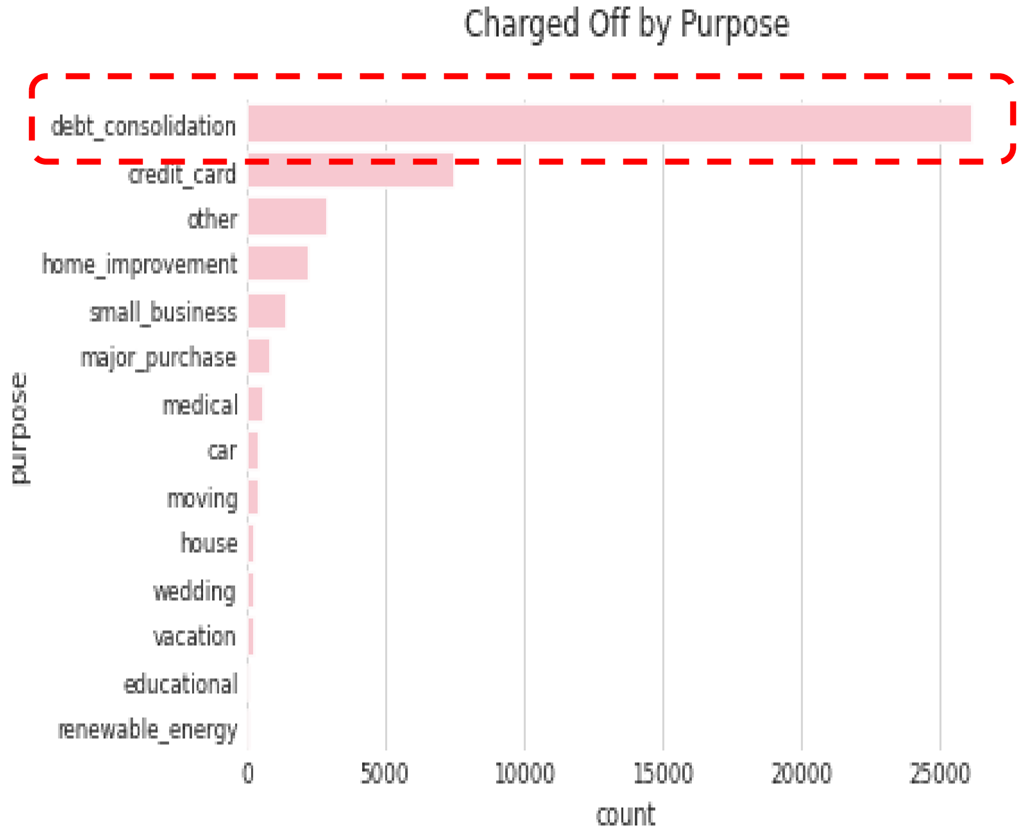
loan status	total loss	total applicant	% total loss	avg loss
Charged Off	$574,356,330.27	43,236	85.29	$13,284.22
Late	$87,745,022.55	6,900	13.03	$12,716.67
Default	$11,299,446.58	832	1.68	$13,581.07

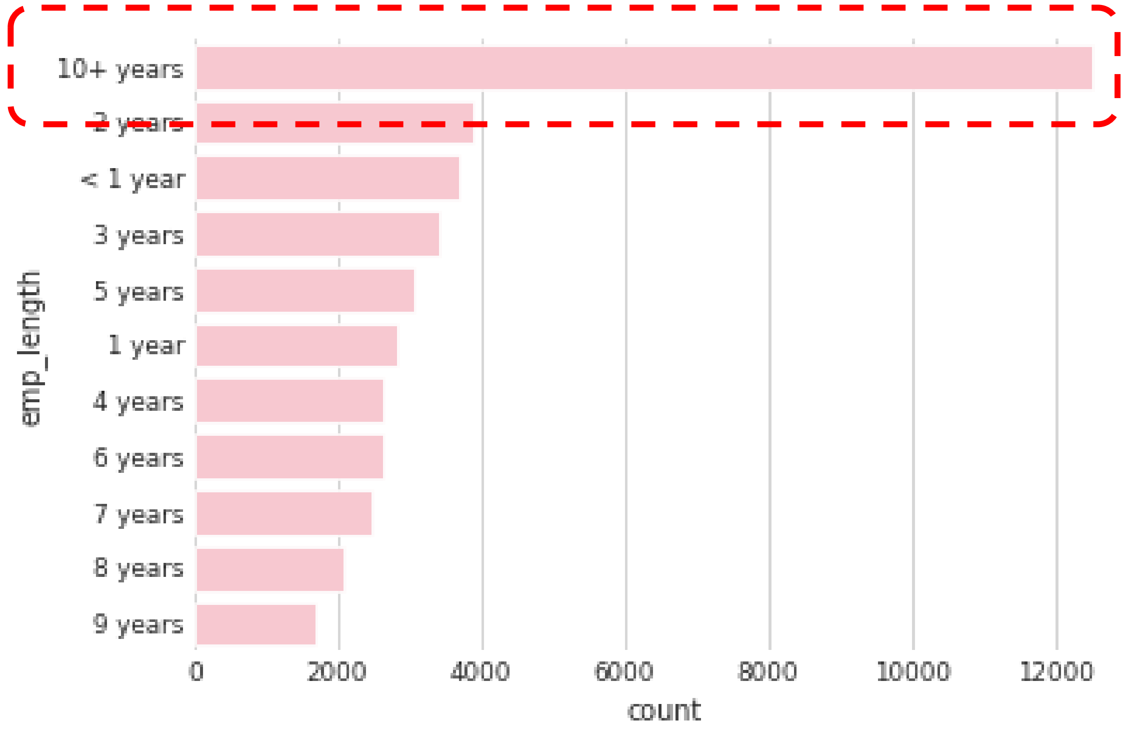
A company has lost a total of $574,356,330 to customers who were charged off, with an average loss of $13,284.22 per person. Further analysis will be conducted.

DATA VIZ AND BUSINESS INSIGHT









Customers who have been charged off have certain attributes or characteristics.

- Loan with term 36 or short
- Home ownership status rent
- Purpose charged off is debt consolidation
- Employer length > 10 years

Insight

- Making customers have long loan terms
- Targeting customers who have home ownership status own
- Have employer length of more than 2 years and less than 10 years

Preprocessing & Modelling

Data Cleaning

Check Data Duplicate
Check Missing Data
Check Data Type

Feature Selection

Split Data Train (80:20)
Categorical (Chi Square)
Numerical (ANOVA)

Feature Engineering

WoE Binning
Information Value (IV)

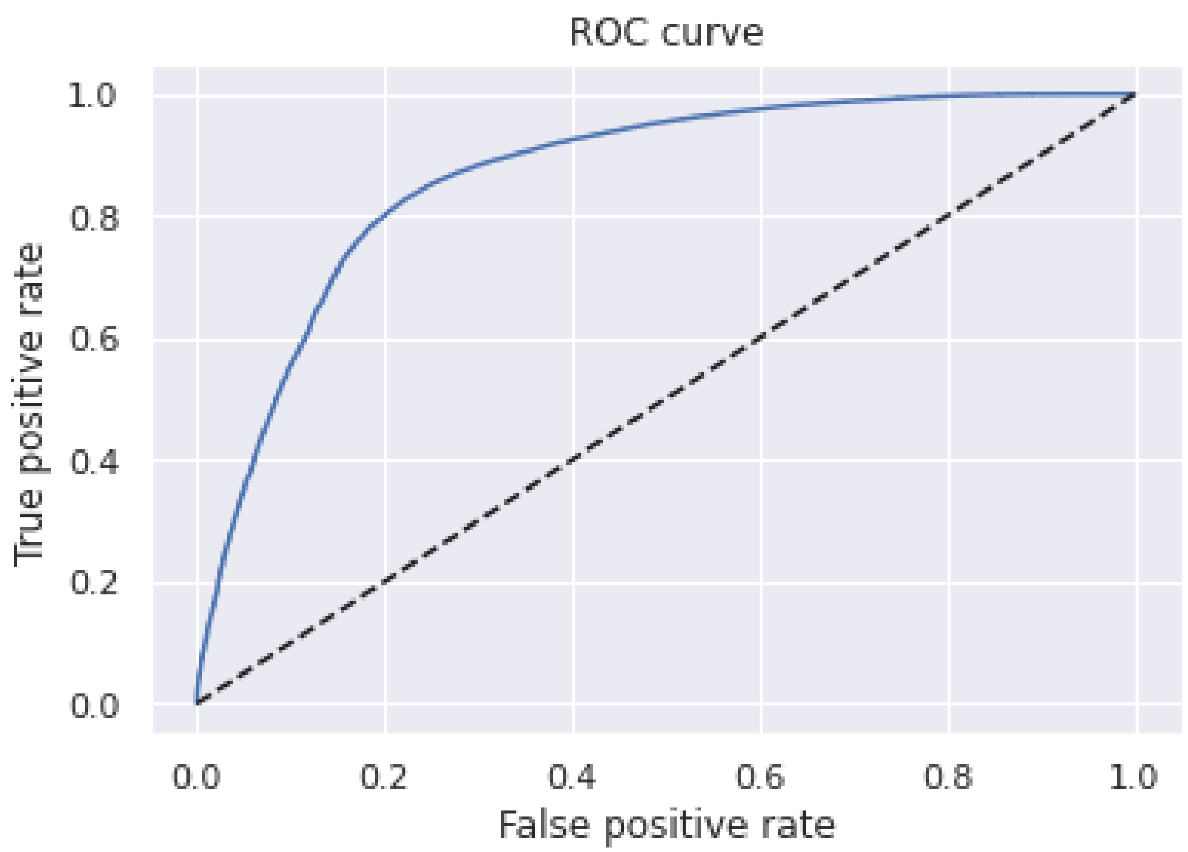
Modelling & Cross Validation

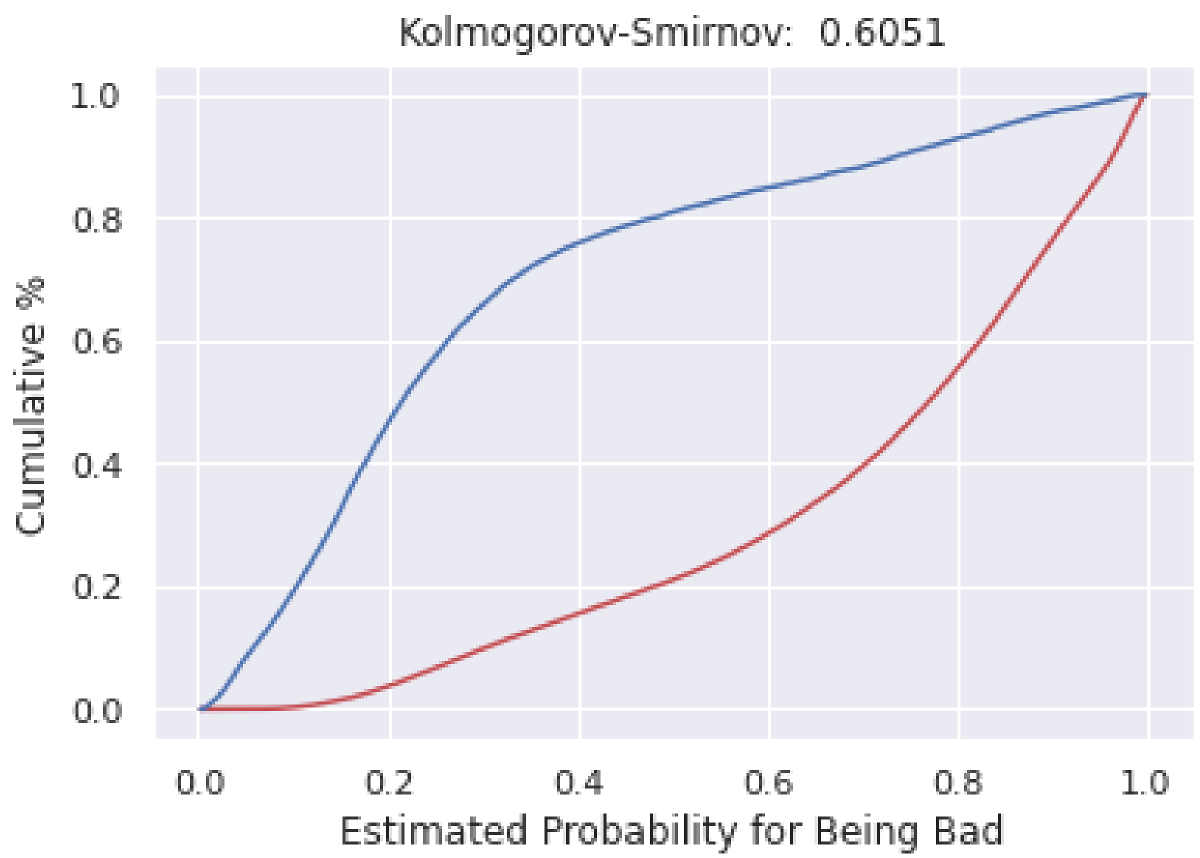
Class_Weight = Balance
Woe_transform
RepeatedStrafieldKfold

https://towardsdatascience.com/feature-selection-and-eda-in-python-c6c4eb1058a3
https://towardsdatascience.com/how-to-develop-a-credit-risk-model-and-scorecard-91335fc01f03
https://medium.com/@finntanweelip/feature-selection-in-credit-scoring-b0eee604cd51

Evaluation

Models	MEAN AUROC	GINI
Decision Tree	0.71	0.41
Logistic Regression	0.87	0.73





`AUROC = 0.866` and `KS = 0.61`. In the world of credit risk modeling, generally AUC above 0.7 and KS above 0.3 is considered good performance

Scorecard

General

BASE(intercept) = 598, Min score = 300, Max score = 850

Feature Name	Details	Score
term	36	-2.0
	60	0.0
total_pymnt	10,000-15,000	-101.0
	15,000-20,000	-73.0
	20,000-25,000	-47.0
	<10,000	-155.0
	>25,000	0.0
total_rec_int	1,089-2,541	94.0
	2,541-4,719	66.0
	4,719-7,260	34.0
	<1,089	109.0
	>7,260	0.0
total_rev_hi_lim	19,144-25,525	5.0
	25,525-35,097	2.0
	35,097-54,241	1.0
	54,241-79,780	-0.0
	6,381-19,144	8.0
	<6,381	9.0
	>79,780	0.0
	missing	5.0
verification_status	Not Verified	0.0
	Source Verified	-7.0

Feature Name	Details	Score
annual_inc	102,606-120,379	5.0
	120,379-150,000	4.0
	28,555-37,440	11.0
	37,440-61,137	10.0
	61,137-81,872	9.0
	81,872-102,606	8.0
	28,555	12.0
	150K	0.0
	missing	15.0
dti	1.6-5.599	12.0
	10.397-15.196	6.0
	15.196-19.195	3.0
	19.195-24.794	1.0
	24.794-35.191	1.0
	5.599-10.397	8.0
	<=1.6	13.0
	>35.191	0.0
grade	A	24.0
	B	20.0
	C	15.0
	D	12.0
	E	8.0
	F	5.0
	G	0.0

Feature Name	Details	Score
inq_last_6mths	0	10.0
	1-2	9.0
	3-4	6.0
	>4	0.0
	missing	3.0
int_rate	10.374-13.676	1.0
	13.676-15.74	1.0
	15.74-20.281	-0.0
	7.071-10.374	7.0
	<7.071	24.0
	>20.281	0.0
mths_since_earliest_cr_line	125-167	-2.0
	167-249	-0.0
	249-331	-0.0
	331-434	1.0
	<125	-4.0
	>434	0.0
	missing	3.0
mths_since_issue_d	100-122	-3.0
	79-89	-8.0
	89-100	-8.0
	<79	-9.0
	>122	0.0
home_ownership	MORTGAGE	0.0
	OTHER_NONE_RENT	-3.0
	OWN	-1.0

Feature Name	Details	Score
mths_since_last_credit_pull_d	56-61	30.0
	61-75	7.0
	<56	16.0
	>75	0.0
	missing	15.0
out_prncp	1,286-6,432	-27.0
	10,291-15,437	-7.0
	6,432-9,005	-16.0
	9,005-10,291	-10.0
	<1,286	-88.0
	>15,437	0.0
purpose	credit_card	-6.0
	consolidation	-8.0
	educ__ren_en__sm_b__mov	-12.0
	major_purch__car__home_impr	0.0
	vacation__house__wedding__med__oth	0.0
revol_util	0.1-0.2	-14.0
	0.2-0.3	11.0
	0.3-0.4	-3.0
	0.4-0.5	8.0
	0.5-0.6	10.0
	0.6-0.7	-6.0
	0.7-0.8	1.0
	0.8-0.9	-1.0
	0.9-1.0	15.0
	<0.1	-4.0
	>1.0	0.0
	missing	-2.0

https://github.com/archie-cm/Credit_Score_Home_Credit_Indonesia

Setting Loan Approval Cut-Offs

General
BASE(intercept) = 569
Min score = 300
Max score = 850

Load_data_2007_2014
466285 applicants

Model
Logistic Regression
AUC 0.87
Recall 0.97
KS 0.61

Threshold
0.5

Best Threshold
0.186574

Threshold = 0.5

Accept Score	N Approved	N Rejected	Approval Rate	Rejection Rate
525.0	67559	25698	0.724439	0.275561

Best Threshold

Accept Score	N Approved	N Rejected	Approval Rate	Rejection Rate
488.0	86384	6873	0.92630	0.07370

Threshold 0.5 would result in a very high rejection rate with a corresponding loss of business.
Accordingly, we will stick with our ideal threshold and the corresponding Credit Score of 488

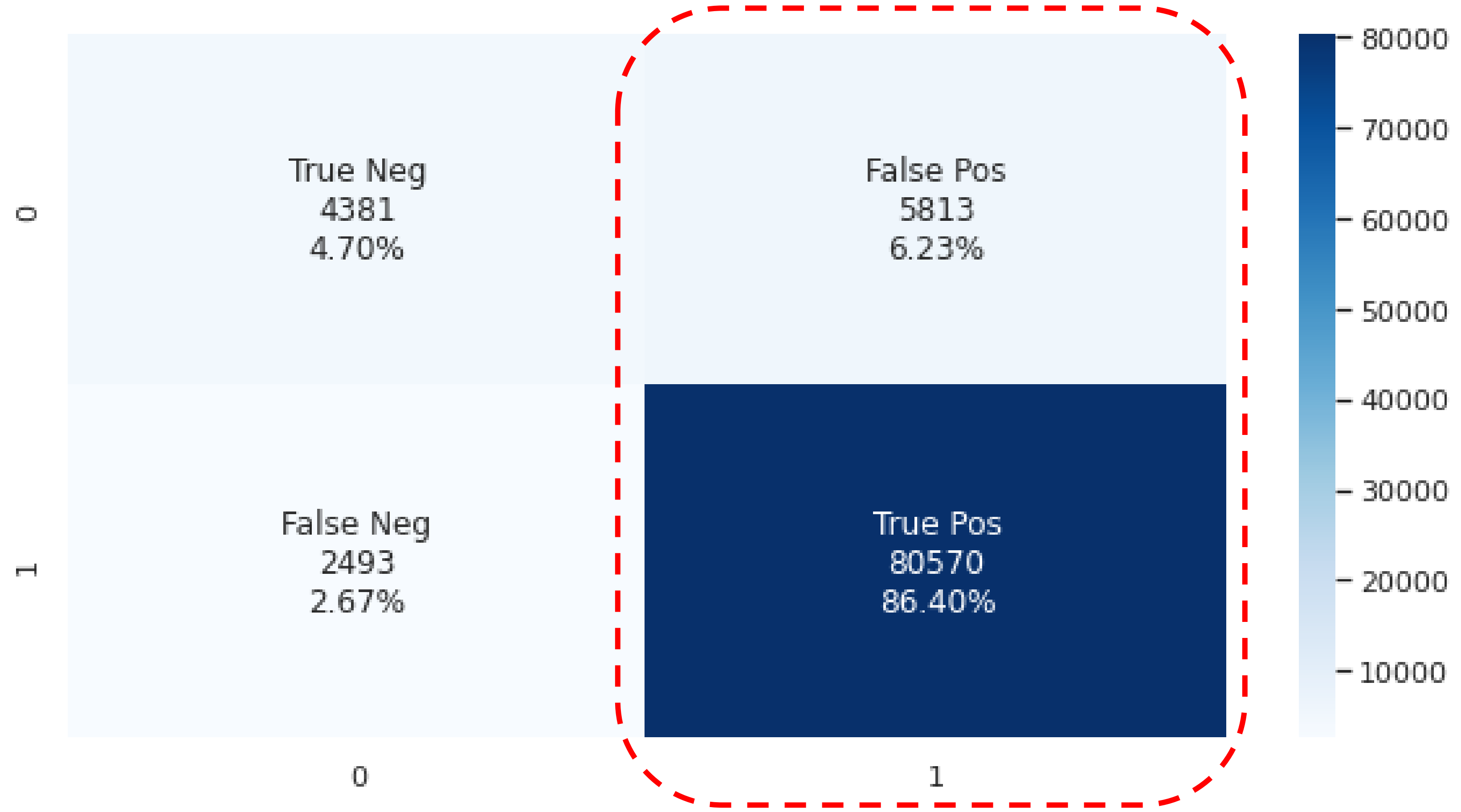
Simulation

Profile of Andi		Score calculation
Base		598
annual_inc	60,000	9
Dti	10	8
Grade	B	20
home_ownership	Own	-1
inq_last_6mths	3	6
int_rate	8	7
earliest_cr_line	120	-2
issue_d	70	-8
last_credit_pull_d	70	7
out_prncp	12,000	-7
purpose	education	-12
revol_util	0.4	-3
term	60	0
total_pymnt	20,000	-47
total_rec_int	3,500	66
total_rev_hi_lim	15,000	8
verification_status	verified	-7
Total		642

If [Accept Score > Score] -> Approve & [Accept Score < Score] -> Reject

Accept Score is 488 and Andi Score is 642, So Andi loan is Approve

Conclusion



Without ML

N Customer	N Approved	Loss
466,285	466,285	$743,972,450

With ML

N Customer	N Approved	Loss
466,285	86,384	$77,225,705

With machine learning, Company can reduce loss from bad loan reached 90% and can focus on customer have long loan terms (60), customers who have home ownership status own and Have employer length of more than 2 years and less than 10 years to maximize revenue.