

# LOAN PREDICTION BASED ON CUSTOMER BEHAVIOUR

WIN GAMES GROUP (kelompok 4)



# OUR TEAM

**Adam Ardiansyah**  
(Mentor)

**Wilhelmus M.**  
(Leader)

**Alfiyanti S.**

**Angina Dwi F.**

**Credenda M.**

**Dimas Fauzi P.**

**Mayang Indi G.**

**Sumayya**

# TABLE OF CONTENTS

1

## Business Understanding

Understand the problem to solve, define what success looks like for solving the problem.

2

## EDA (Exploratory Data Analysis)

Analyze and investigate data sets and get the insights.

3

## Data PreProcessing

Handles null, duplicate and 'weird' data. Choose which features to use to minimize errors

4

## Modeling & Evaluation

Select and develop the algorithm model to be used, and evaluate the results of the model.





1

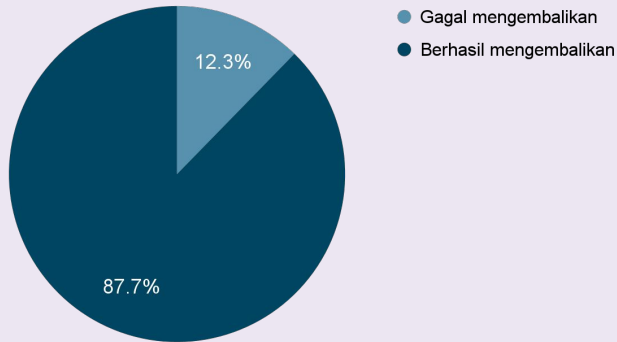
# Business Understanding

# Role & Problem Statement

## Role

**Win Games Group** merupakan tim Data Scientist di PT. Pinjamkan bertugas untuk mencari **solusi** dalam mengatasi masalah melalui dataset nasabah yang tersedia.

Persentase Kelompok Nasabah



## Problem Statement

PT. Pinjamkan memberikan pinjaman uang



12,3% nasabah gagal mengembalikan pinjaman



Jika nasabah yang berisiko bertambah maka dapat menyebabkan kerugian



Perusahaan perlu mendeteksi nasabah mana yang mungkin berisiko





# Goal & Business Metrics

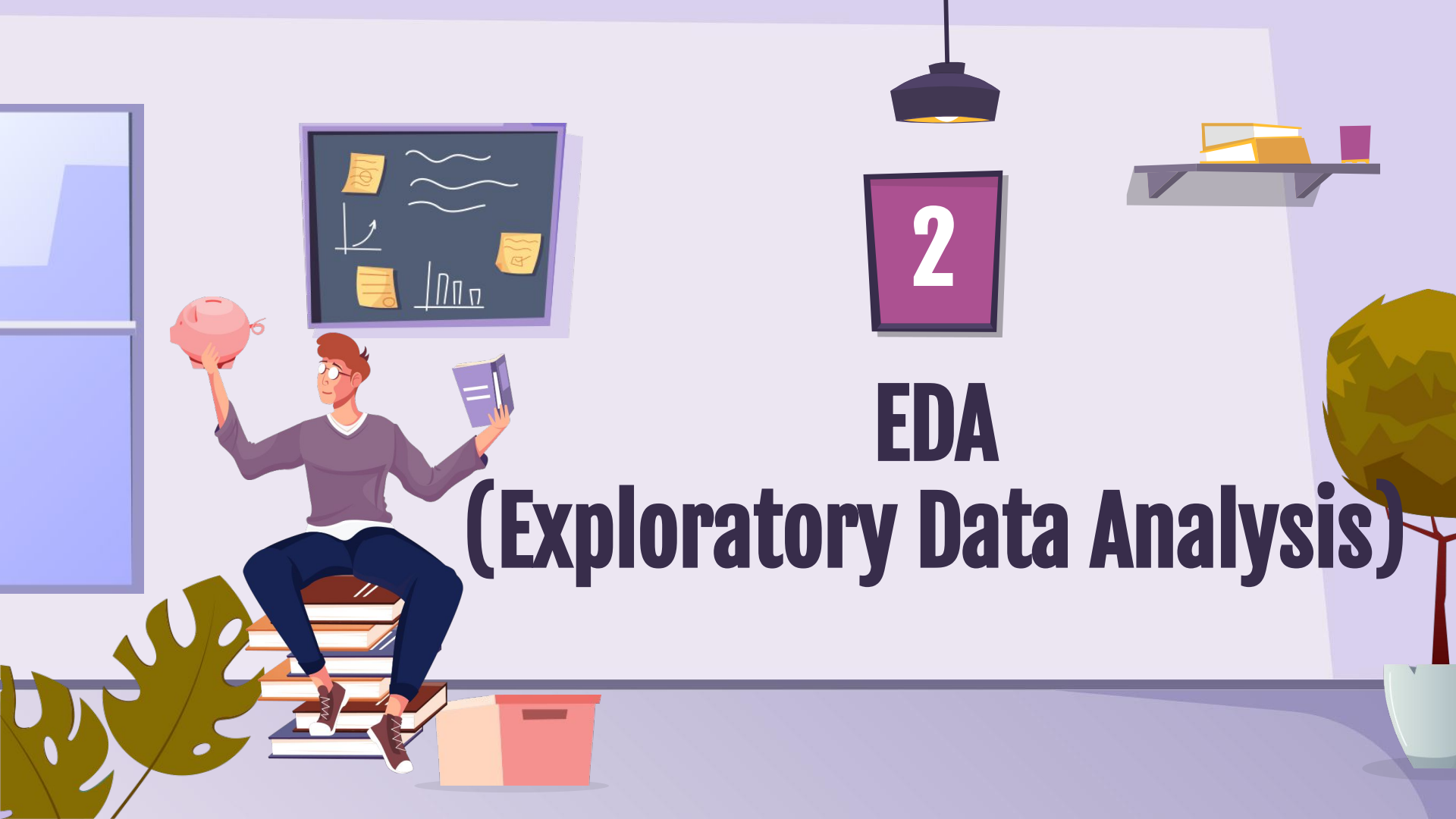


## Objective & Goal

- **Membuat predictive model** yang mampu memprediksi nasabah yang memiliki kemungkinan gagal membayar pinjaman, dan mengetahui fitur-fitur penting dalam memprediksi *risk customer*
- **Mengoptimalkan laba bisnis**, dengan cara mengklasifikasikan nasabah yang berisiko berdasarkan dataset yang ada untuk meminimalkan kerugian.

## Business Metrics

1. **Default rate** (persentase pelanggan yang tidak dapat mengembalikan pinjaman).
  2. **Revenue, Profit, dan Cost** perusahaan berdasarkan sebelum dan sesudah hasil model.
- 
- 



2

# EDA (Exploratory Data Analysis)



# ABOUT DATASET

## Loan Dataset



**12 Features  
in dataset**

### 5 Numerical

- Income
- Age
- Experience
- Current job years
- Current house years

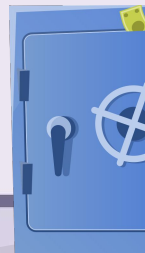
### 7 Categorical

- Profession
- Married/Single
- House Ownership
- Car Ownership
- City
- State
- Risk Flag

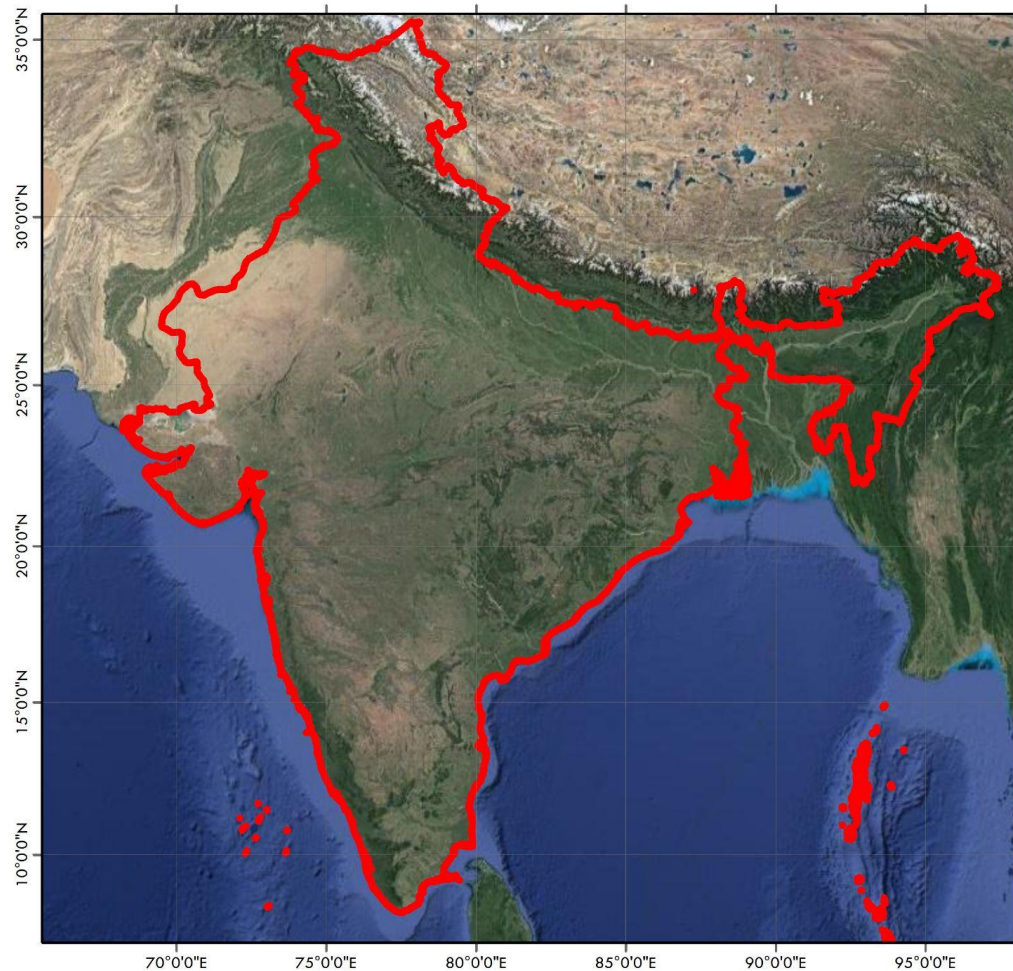
**252000 rows**

**0 null**

**0 duplicates**







## Peta Situasi Area Kajian Negara India

1:18.000.000



Kilometers  
0 130 260 520 780 1,040

Proyeksi : World Mecrator  
Sistem Grid : Grid Geografi  
Sphereoid : WGS 1984 World Mecrator

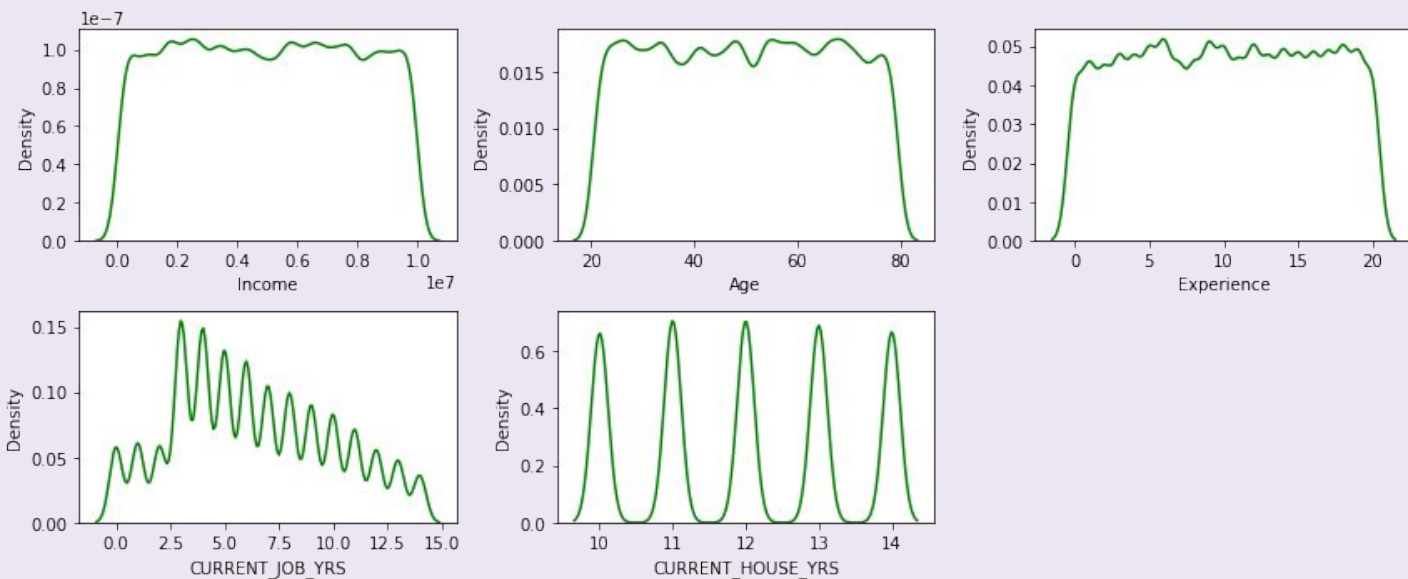
Sumber data : GADM database ([www.gadm.org](http://www.gadm.org))

### Keterangan :

 Batas Area Kajian

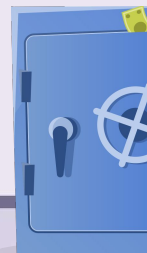
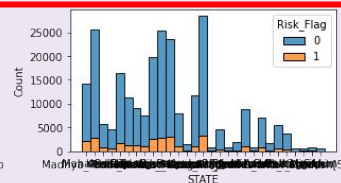
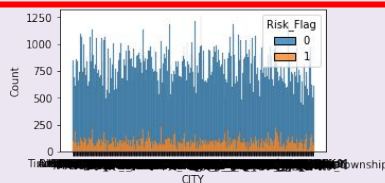
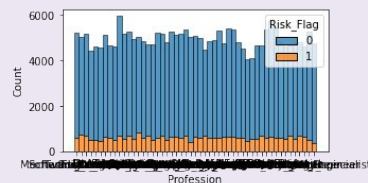
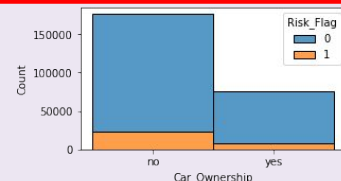
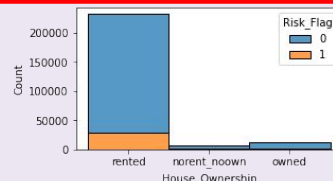
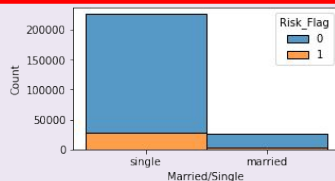
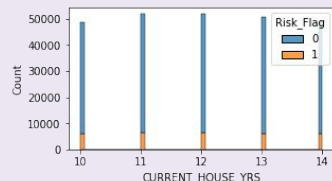
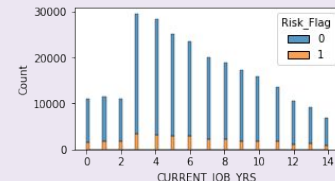
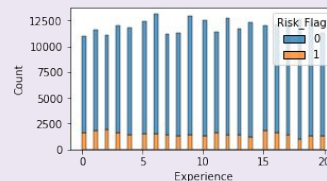
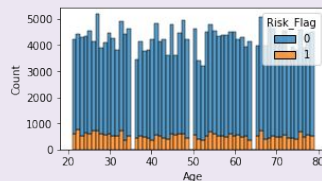
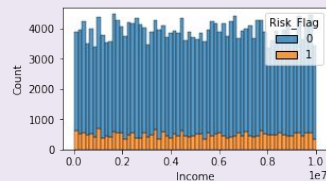
# ABOUT DATASET

## DATA VISUALIZATION



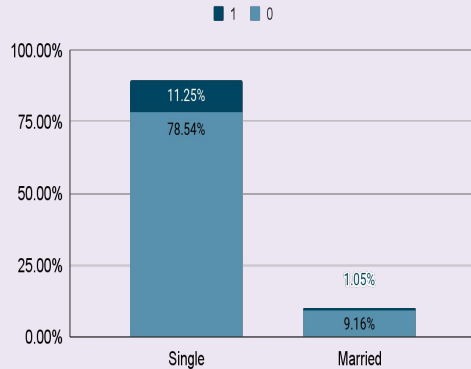
# ABOUT DATASET

## DATA VISUALIZATION



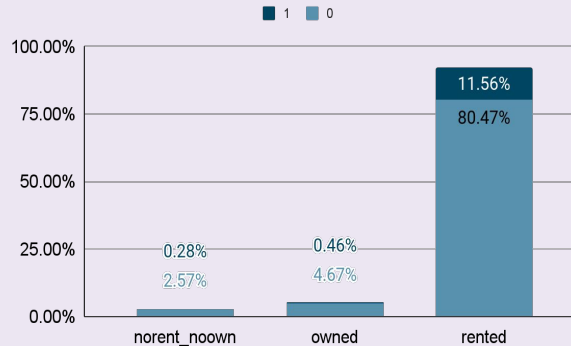
# ABOUT DATASET

Marital Status



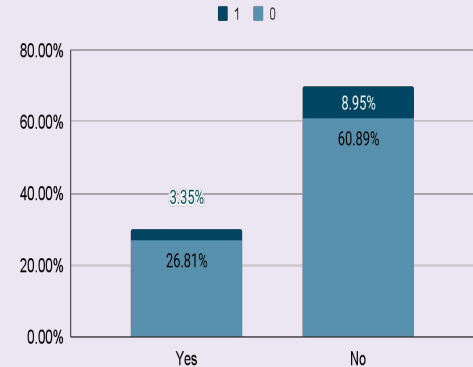
**89,7%** dari Total  
Customer berstatus  
**Single**

Own House



**92,02%** dari Total  
Customer masih  
**menyewa rumah**

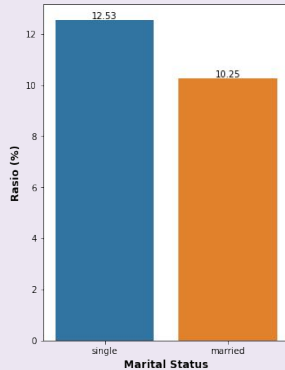
Own Car



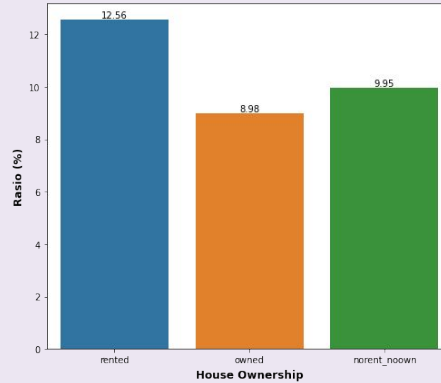
**69,84%** dari Total  
Customer **tidak**  
**memiliki mobil**

# ABOUT DATASET

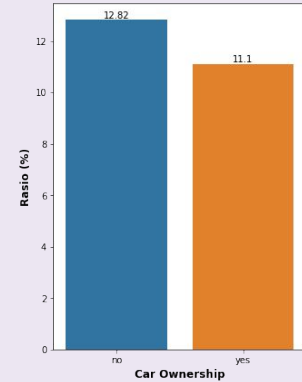
Distribusi risk customers berdasarkan marital status



Distribusi risk customers berdasarkan house ownership



Distribusi risk customers berdasarkan car ownership



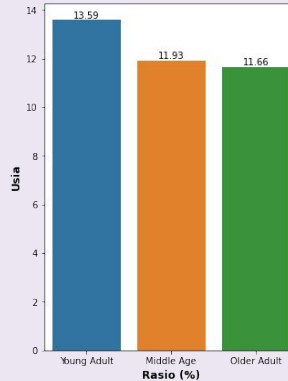
Rasio *risk customers* berdasarkan marital status, house ownership, dan car ownership cukup seimbang, hal ini berbeda dengan distribusi data secara keseluruhan

# ABOUT DATASET

	Income
Mean	4,997,116
Min	10,310
Max	9,999,938

Pendapatan Rata-Rata  
dari Total Customer  
adalah **5 Juta**

Distribusi risk customers berdasarkan usia



Rasio tertinggi **risk customers** terdapat  
pada kelompok usia  
**Young Adult** (21-35 th)

	Experience
Mean	10
Min	0
Max	20

Rata-Rata Nasabah  
memiliki **experience**  
selama **10 Tahun**

# ABOUT DATASET

## State rank :

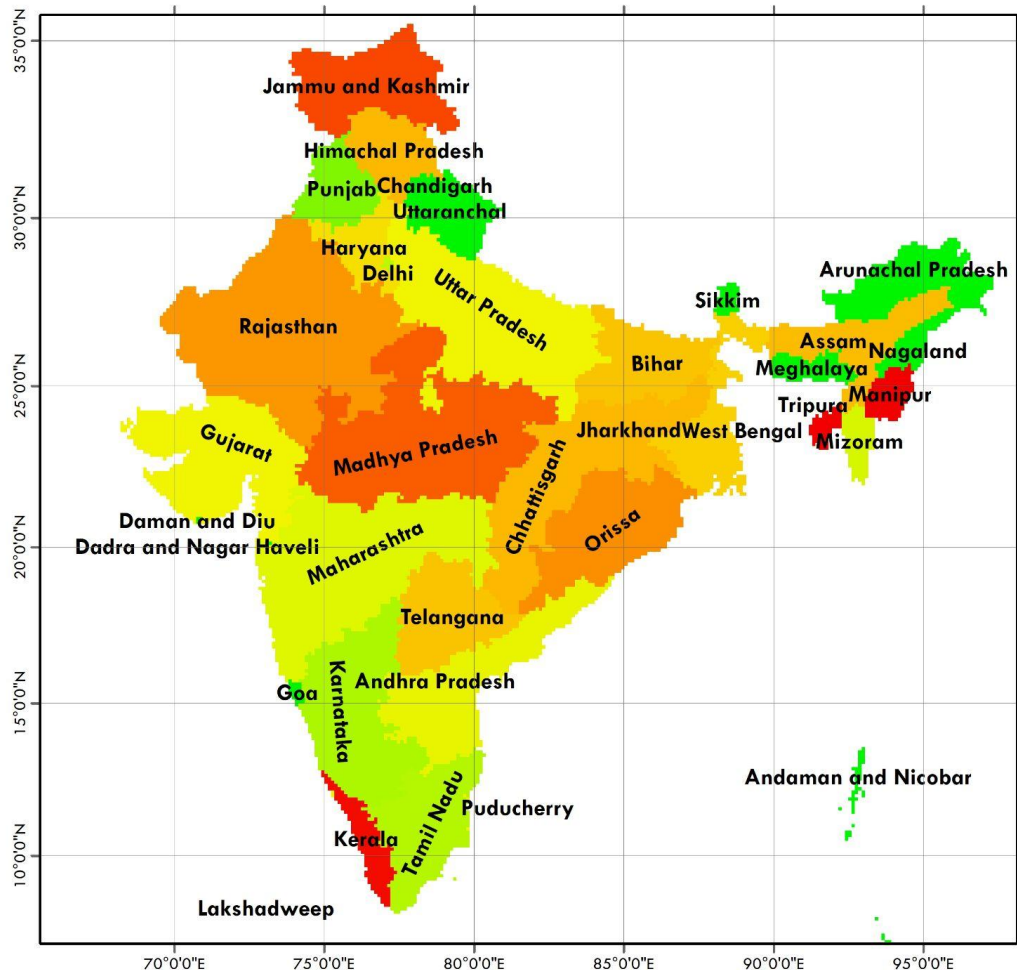
5 State **teratas** (berdasarkan *risk customers*)



Manipur is the **third poorest state in India**.

Poverty in Manipur is around 36.89%.

The social and economic structure of Manipur is one of the lowest in India.



## Peta Negara Bagian India Berdasarkan Rasio Customer yang Beresiko Gagal Bayar

1:18.000.000



Kilometers  
0 130 260 520 780 1.040

Proyeksi : World Mecrator  
Sistem Grid : Grid Geografi  
Sphereoid : WGS 1984 World Mecrator

Sumber data : GADM database ([www.gadm.org](http://www.gadm.org))

### Keterangan :

Nilai Rasio (fraksi desimal)

High : 0,215548



Low : 0

Semakin tinggi  
persentase resiko  
suatu state maka  
warnanya akan  
cenderung merah.

State	Ratio(%)
Manipur	21,55
Tripura	16,81
Kerala	16,70
Jammu Kashmir	15,89
Madya Pradesh	15,43



# ABOUT DATASET

## Profession rank :

5 Profesi **teratas** (berdasarkan risk customer masing-masing profesi)

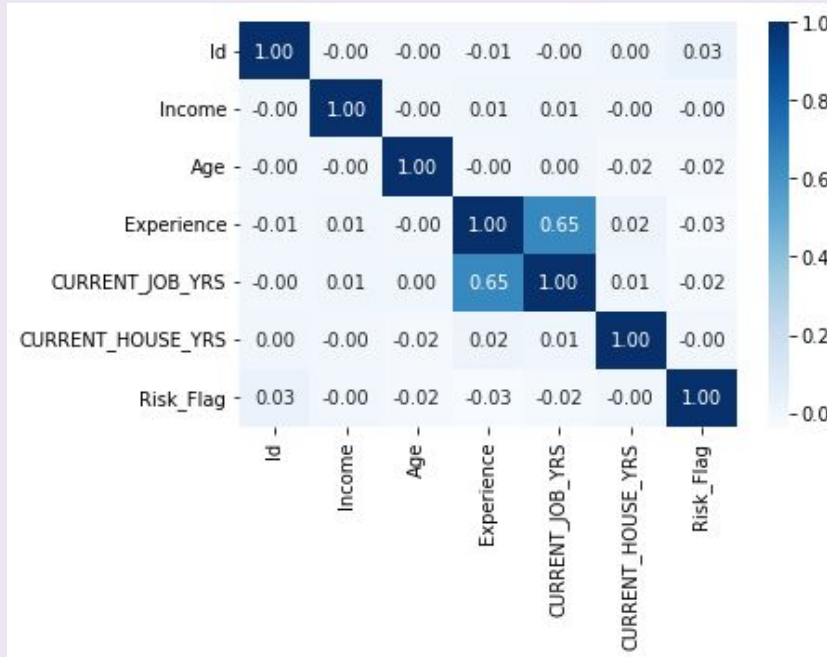


Police officers have a fairly **low salary** compared to other professions

0 - 2 Years		9,920 INR
2 - 5 Years	▲ +34%	13,200 INR
5 - 10 Years	▲ +48%	19,600 INR
10 - 15 Years	▲ +22%	23,900 INR
15 - 20 Years	▲ +9%	26,000 INR

# ABOUT DATASET

## CORRELATION





# Pre-Processing

# DATA CLEANSING

Handling missing value



Handling duplicated data



Handling outlier

```
train['STATE'].unique()  
  
array(['Madhya_Pradesh', 'Maharashtra', 'Kerala', 'Odisha', 'Tamil_Nadu',  
      'Gujarat', 'Rajasthan', 'Telangana', 'Bihar', 'Andhra_Pradesh',  
      'West_Bengal', 'Haryana', 'Puducherry', 'Karnataka',  
      'Uttar_Pradesh', 'Himachal_Pradesh', 'Punjab', 'Tripura',  
      'Uttarakhand', 'Jharkhand', 'Mizoram', 'Assam',  
      'Jammu_and_Kashmir', 'Delhi', 'Chhattisgarh', 'Chandigarh',  
      'Uttar_Pradesh[5]', 'Manipur', 'Sikkim'], dtype=object)
```

Terdapat beberapa value yang 'aneh' atau berbeda dari value yang lain pada kolom STATE dan CITY



# DATA CLEANSING

## FEATURE TRANSFORMATION

### Normalized

- Age
- Income
- Experience
- Current house years

## FEATURE ENCODING

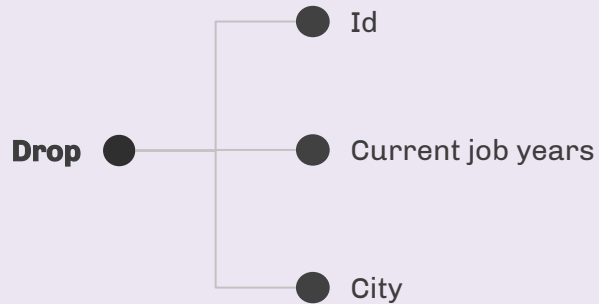
### Label Encoding

- Married/Single
- House Ownership
- Car Ownership
- Profession
- State

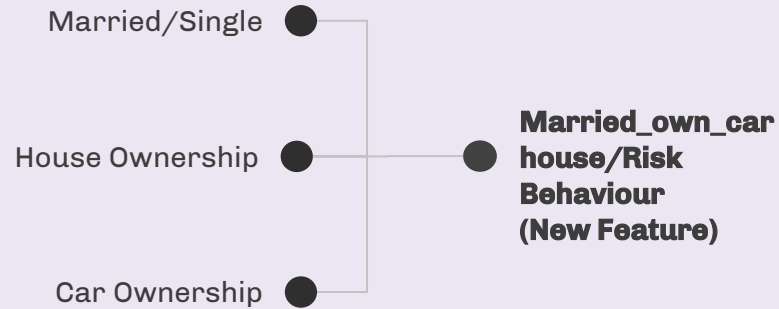


# FEATURE ENGINEERING

## FEATURE SELECTION

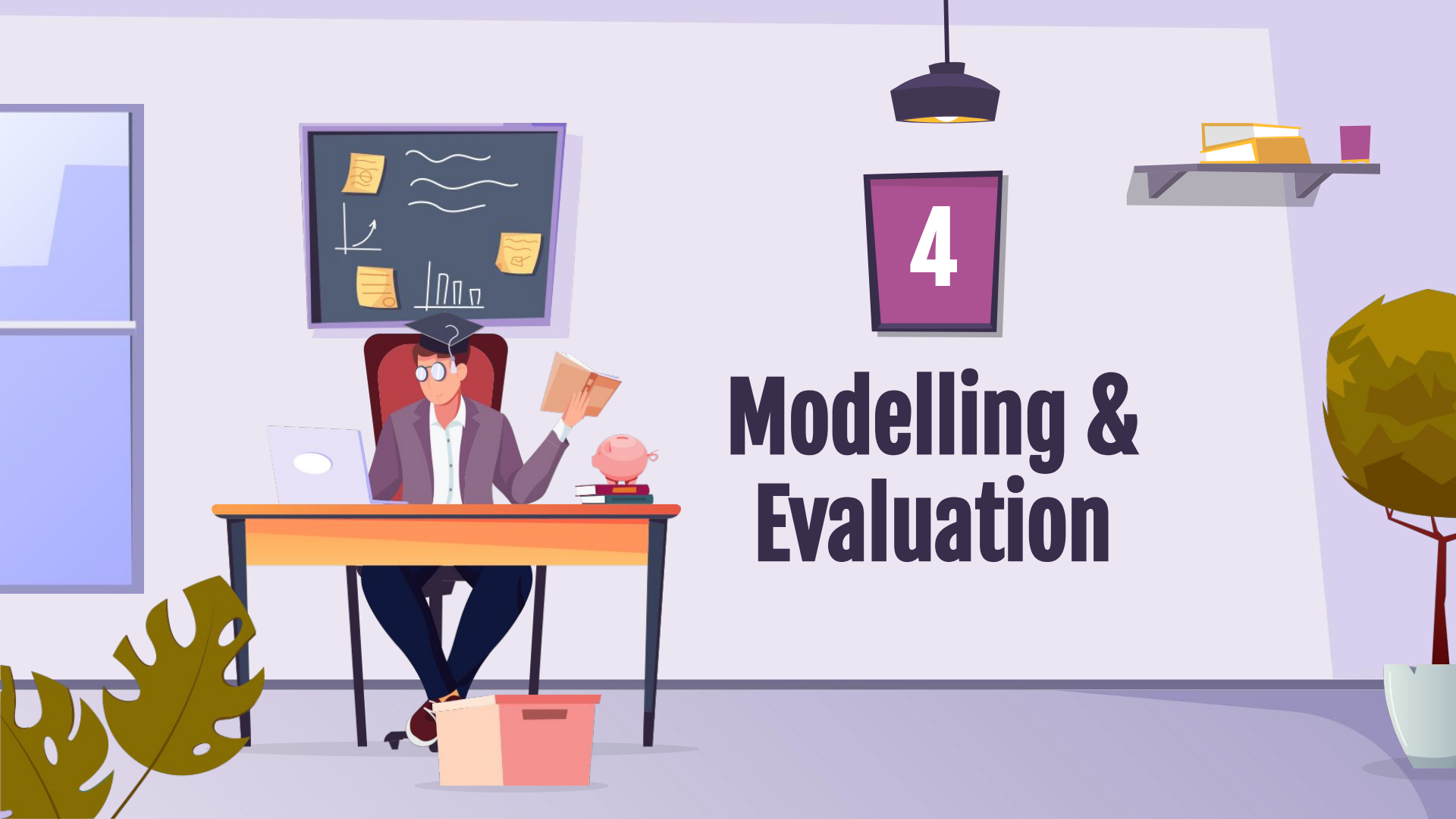


## FEATURE EXTRACTION



Nasabah yang telah menikah, memiliki rumah dan mobil akan diberikan nilai 1, sedangkan lainnya diberi nilai 0. Nasabah dengan karakteristik tersebut cenderung memiliki peningkatan resiko gagal membayar.



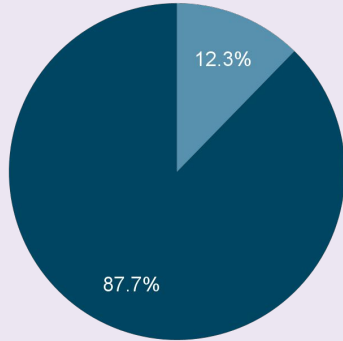


4

# Modelling & Evaluation

# MODELLING

Persentase Kelompok Nasabah



- Gagal mengembalikan
- Berhasil mengembalikan

1. Dataset awal,  
rasio **87 : 13** (252.000)
2. Handle Class Imbalance,  
rasio **66 : 33** (265.203)
3. Split Data,  
rasio **70 : 30** (265.203)

## Dataset awal

Total 252.000, customer beresiko (**30.996**) dan customer non resiko (**221.0004**)



## Handle Class Imbalance

SMOTE (0,4) & Under Sampling (0,5),  
customer beresiko (**88.401**) dan  
customer non resiko (**176.802**)



## Split Train & Test

Sampling strategy 70:30  
(**185.642 : 79.561**)



# MODELLING

## 7 Features

- Income\_norm
- Age\_norm
- Experience\_norm
- House year\_norm
- State Rank
- Profession Rank
- Married\_own\_carhouse

## Target

- Risk Flag

## Model

- Logistic
- K-Nearest Neighbour
- Decision Tree
- Random Forest
- XGBoost
- AdaBoost

# MODELLING

## MODEL EVALUATION

Model	Accuracy (%)	Precision (%)	Recall (%)	AUC (%)	F1 (%)
<b>Logistic Regression</b>	66,41	48,67	0,21	50,05	0,41
<b>K-Nearest Neighbor</b>	87,80	78,72	87,28	87,68	82,78
<b>Decision Tree</b>	88,65	78,67	90,84	89,19	84,32
<b>Random Forest</b>	<b>90,77</b>	<b>83,56</b>	<b>90,29</b>	<b>90,66</b>	<b>86,80</b>
<b>XGB</b>	85,47	85,61	68,20	81,20	75,92
<b>AdaBoost</b>	67,71	90,13	4,34	52,05	82,83

# TUNING HYPERPARAMETERS

```
# list dari hyperparameter
#n_estimators = [int(x) for x in np.linspace(100, 800, num = 8)]

# mengumpulkan semua hyperparameter pada dictionary
hyperparameters = dict(
    n_estimators = [int(x) for x in np.linspace(100, 1000, num=10)],
    criterion = ['gini', 'entropy'],
    max_depth = [int(x) for x in np.linspace(100, 1000, num=10)],
    min_samples_split = [int(x) for x in np.linspace(2, 10, num=5)],
    min_samples_leaf = [int(x) for x in np.linspace(1, 5, num=5)],
    max_features = ['auto', 'sqrt', 'log2']
)

# Fit model
model = RandomForestClassifier(random_state=42)
tune_clf = RandomizedSearchCV(model, hyperparameters, cv=5, scoring='recall')
```

# BEST MODEL

## TUNING HYPERPARAMETERS

Hasil tuning model  
RANDOM FOREST  
meningkatkan nilai  
metrik model cukup  
baik, terutama pada  
bagian precision.

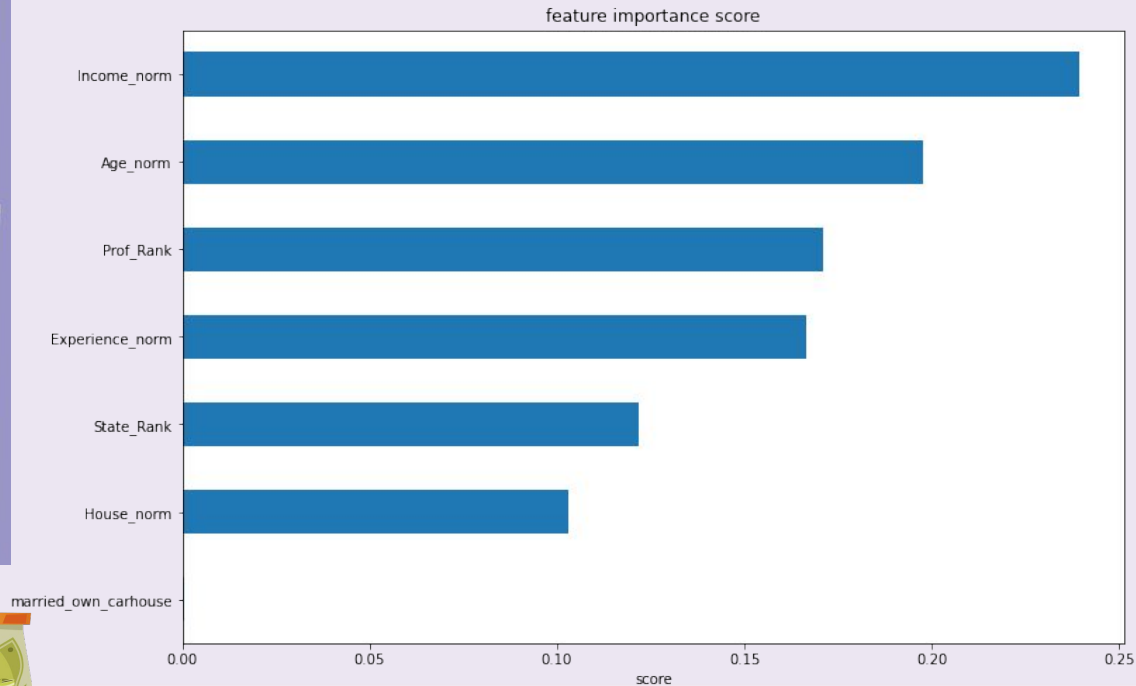
Metrics	Before Tuning (%)	After Tuning (%)	Perbedaan (%)
<b>Accuracy</b>	90,77	90,85	0,08
<b>Precision</b>	83,56	83,73	0,17
<b>Recall</b>	90,29	90,30	0,01
AUC	90,66	90,71	0,05
F1	86,80	86,89	0,09

# BEST PARAMETERS

```
# Iterasi model berdasarkan best parameter dan feature selection
best_model = RandomForestClassifier(bootstrap=True,
                                   ccp_alpha=0.0,
                                   class_weight=None,
                                   criterion='entropy',
                                   max_depth=200,
                                   max_features='auto',
                                   max_leaf_nodes=None,
                                   max_samples=None,
                                   min_impurity_decrease=0.0,
                                   min_samples_leaf=1,
                                   min_samples_split=2,
                                   min_weight_fraction_leaf=0.0,
                                   n_estimators=300,
                                   n_jobs=None,
                                   oob_score=False,
                                   random_state=42,
                                   verbose=0,
                                   warm_start=False)

check_scoring(best_model)
```

# FEATURE IMPORTANCE



## FEATURE IMPORTANCE

- **Income** merupakan feature terpenting dalam model.
- **Risk behaviour / married\_own\_car house** memiliki pengaruh paling rendah (tidak mencapai 0.01).

# FEATURE SELECTION

Feature **Risk Behaviour (married\_own\_carhouse)** di **drop**, kemudian iterasi model dengan best parameter model.

Metrics	Tuned Random Forest (%)
<b>Accuracy</b>	90,85
<b>Precision</b>	83,73
<b>Recall</b>	90,30
AUC	90,71
F1	86,89

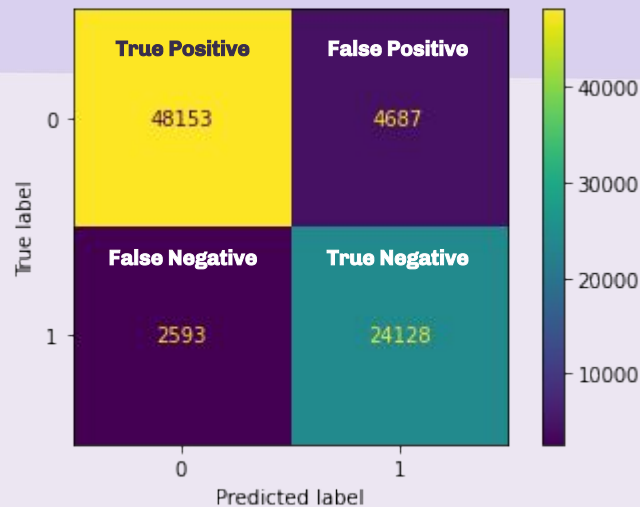
# EXTERNAL SOURCE

- Menurut situs **BankBazaar**, pinjaman di India digolongkan menjadi 2 yakni pinjaman tanpa jaminan (biaya murni untuk pribadi) dan pinjaman dengan jaminan (biaya lainnya).
- Jumlah pinjaman personal yang dapat diberikan oleh Bank di India **minimum sebesar 1 lakh** (1 lakh = 10.000 rupee; 1 rupee = Rp. 189,21).
- **Diasumsikan setiap orang** diberikan **pinjaman 10.000 rupee**, kemudian dengan menggunakan **data test** sebanyak **79,561 nasabah** akan dilakukan perhitungan **Default Rate, Cost, Revenue, dan Profit** sebagai berikut :





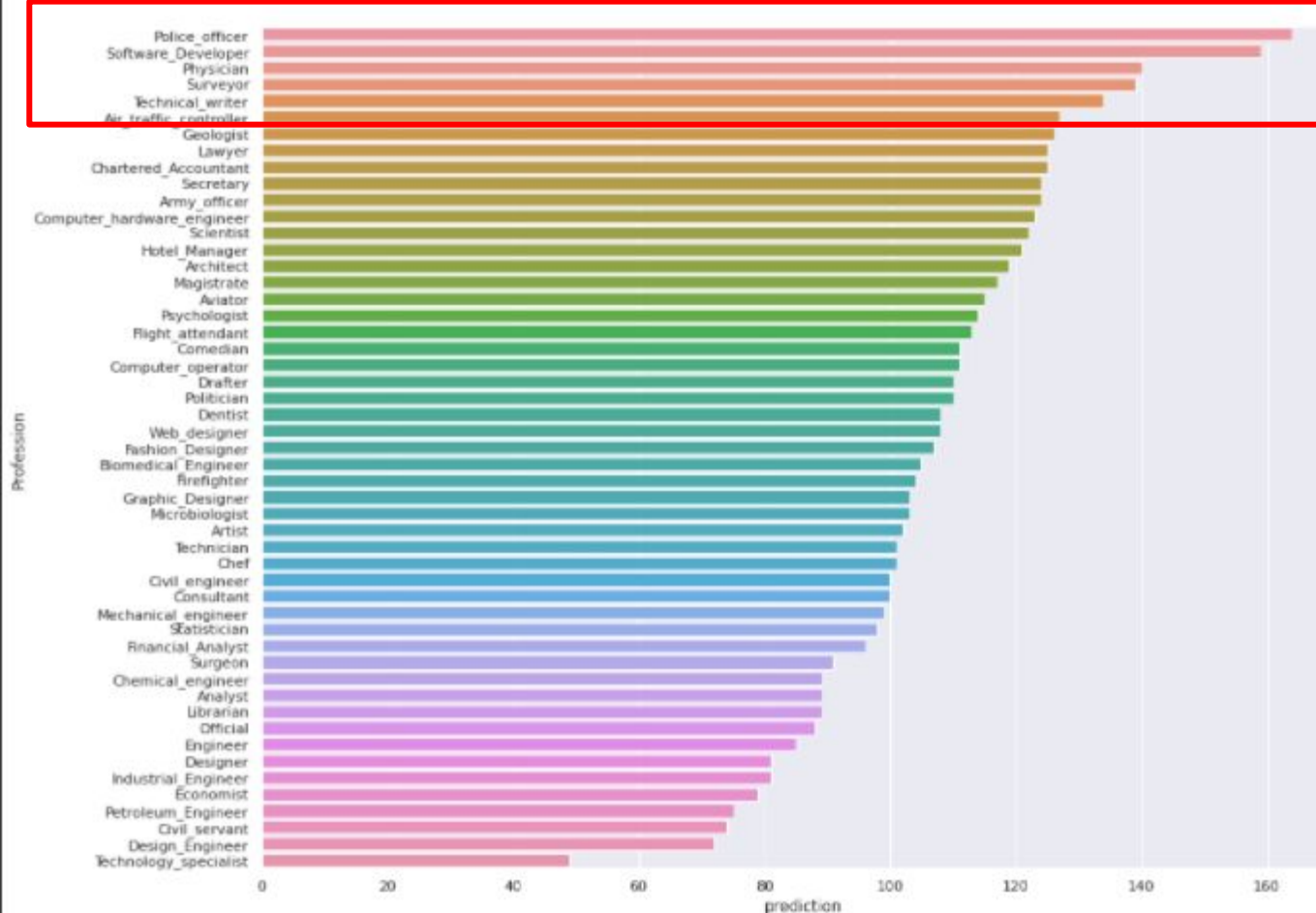
Metrics	Tuned Random Forest (%)
Accuracy	90,85
Precision	83,73
Recall	90,30



Kondisi	Cost	Revenue	Profit	Net Profit Margin	Return of Investment	Default Rate
Ideal	INR 795,610,000	INR 875,171,000	INR 79,561,000	9.09%	10.00%	0.00%
Aktual	INR 795,610,000	INR 581,240,000	-INR 214,370,000	-36.88%	-26.94%	33.59%
Prediksi Model	INR 507,460,000	INR 529,683,000	INR 22,223,000	4.20%	4.38%	5.11%

Kondisi	Nasabah		Jumlah	Pinjaman	Bunga	Cost	Expected Revenue	Revenue Realization	Expected Profit	Profit Realization	Net Profit Margin	Default Rate
a	b		c	d	e	f = c x d	g = f x (1 + e)	h	i = g - f	j = h - f	k = j / h	l
Ideal	Semua		79,561	10,000	10%	INR 795,610,000	INR 875,171,000	INR 875,171,000	INR 79,561,000	INR 79,561,000	9.09%	0%
Aktual	Tidak berisiko		52,840	10,000	10%	INR 528,400,000	INR 581,240,000	INR 581,240,000	INR 52,840,000	INR 52,840,000		
	Berisiko		26,721	10,000	10%	INR 267,210,000	INR 293,931,000	-	INR 26,721,000	- INR 267,210,000		
Total			79,561	10,000	10%	INR 795,610,000	INR 875,171,000	INR 581,240,000	INR 79,561,000	-INR 214,370,000	-38.88%	33.59%
Prediksi Model	Tidak berisiko 50,746	Tidak berisiko	48,153	10,000	10%	INR 481,530,000	INR 529,683,000	INR 529,683,000	INR 48,153,000	INR 48,153,000		
		Berisiko	2,593	10,000	10%	INR 25,930,000	INR 28,523,000	-	INR 2,593,000	- INR 25,930,000		5,11%
	Berisiko (tolak) 28,815	Tidak berisiko	4,687	0	10%	-	-	-	-	-		
		Berisiko	24,128	0	10%	-	-	-	-	-		
Total			79,561	10,000	10%	INR 507,460,000	INR 558,206,000	INR 529,683,000	INR 50,746,000	INR 22,223,000	4,20%	5,11%

### Prediksi urutan profesi berdasarkan false negative



5 Profesi **teratas**  
(berdasarkan  
profesi terhadap  
jumlah customer  
pada false negative)

	Profession	prediction
0	Police_officer	164
1	Software_Developer	159
2	Physician	140
3	Surveyor	139
4	Technical_writer	134

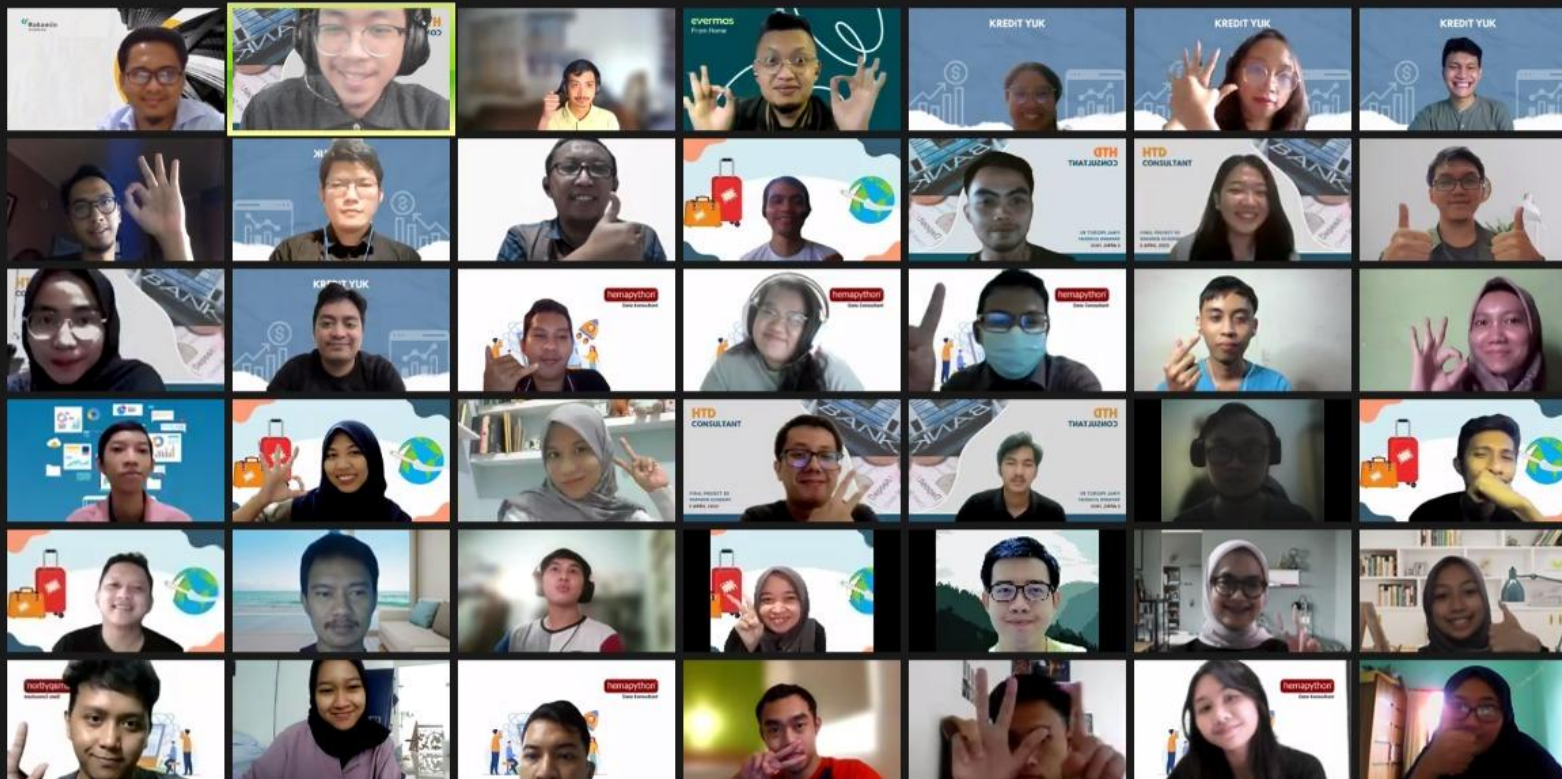
# BUSINESS RECOMMENDATION

1. Berdasarkan karakteristik dataset, perusahaan dapat **menawarkan program pinjaman modal** untuk **menikah, membeli rumah**, ataupun **cicilan mobil** kepada nasabah.
2. Perusahaan dapat **mengimplementasikan model** ini dalam **prediksi** nasabah tergolong dalam status beresiko atau non resiko.
3. Perusahaan perlu **menambah informasi finansial nasabah** agar dapat menambah feature lebih banyak dalam mengembangkan model.
4. Perusahaan dapat lebih **memperketat syarat pinjaman** bagi nasabah, terutama yang **profesinya** termasuk dalam lima peringkat teratas nasabah gagal bayar. (Sebagai contoh pada nasabah yang terklasifikasi False Negative, 5 profesi terbanyak adalah **polisi, software developer**, fisikawan, **surveyor**, dan penulis teknik).



# THANK YOU!

**Are there any questions?**



1/2

1/2

Mute

Stop Video

Participants 54

Q&A

Polls

Chat 1

Share Screen

Raise Hand

Record

Leave