DATA SCIENCE PROJECT

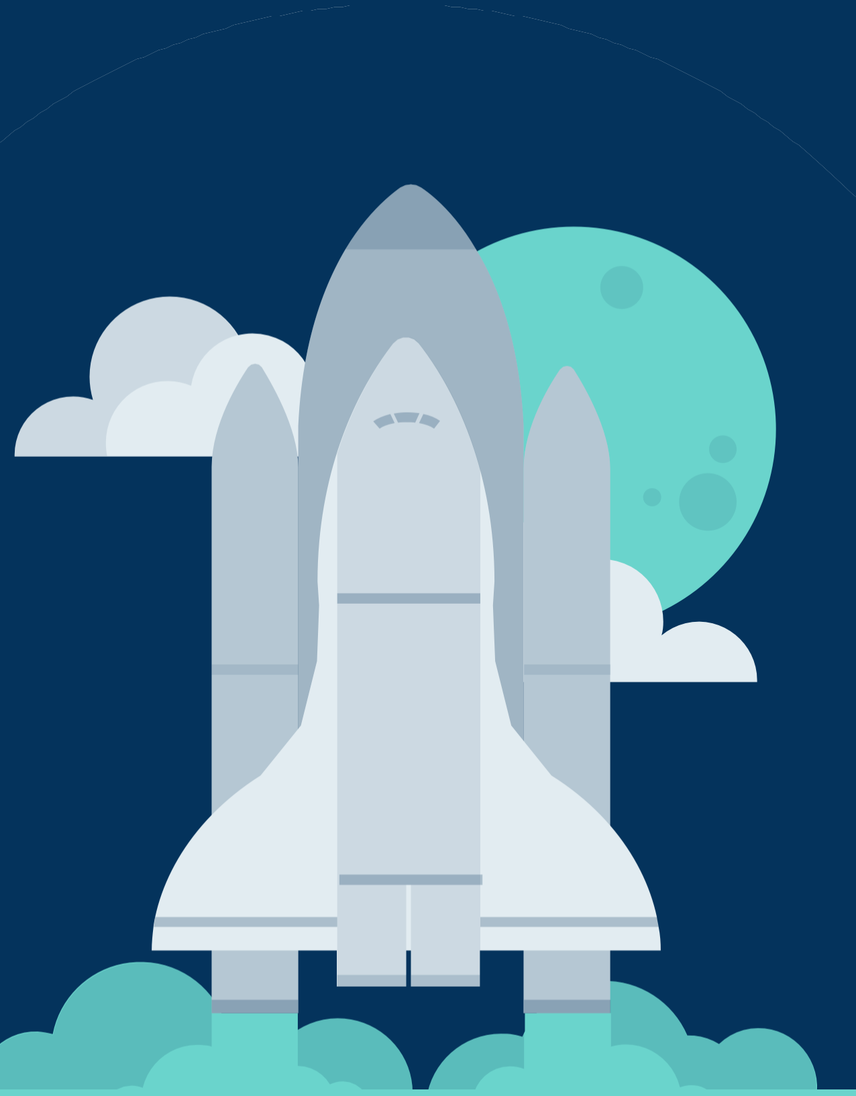
Loan Prediction Based on Customer Behavior

By: Haikal Zamzami

OUTLINE

1. Business Understanding
2. EDA, Insight & Visualization
3. Data Pre-Processing
4. Machine Learning, Modelling & Evaluation
5. Business Insight & Recommendation

BUSINESS UNDERSTANDING



WHO ARE WE?

KreditYuk adalah nama tim dari sekelompok data scientist yang mampu memberikan insights dan rekomendasi dari masalah loan company agar mengurangi dari resiko gagal bayar.

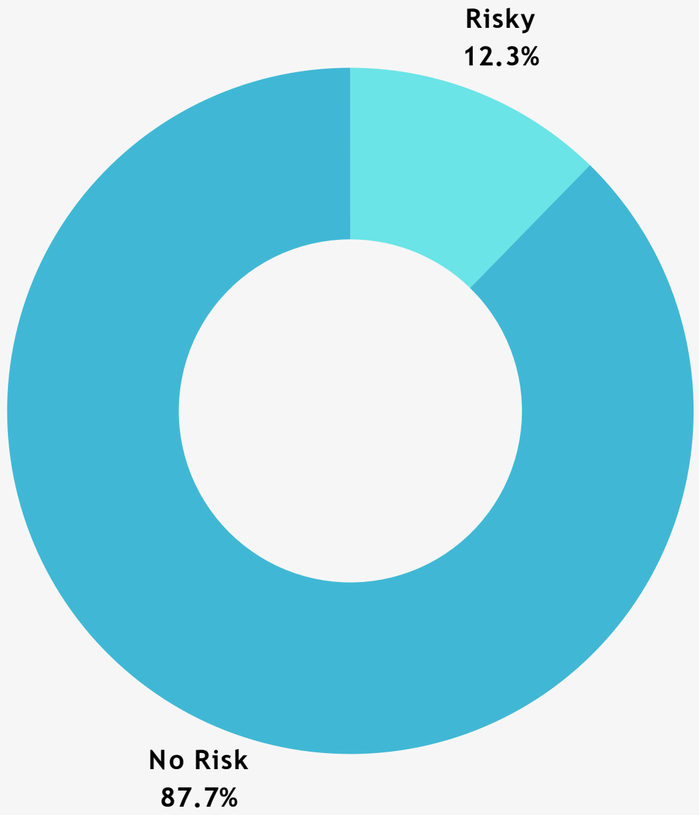
WHO & WHAT CASE?

MoneyFree adalah sebuah startup company P2P landing yang berada di India. Di tahun-tahun pertama nya, mereka mengalami beberapa kesulitan mendeteksi calon nasabah yang mengalami gagal bayar.

Untuk memberikan kenyamanan kepada investor, perlu dilakukan analisis kepada calon customer yang beresiko gagal bayar.

LET'S CHECK THE DATA

Risk Flag Chart



Sampai saat ini, MoneyFree memiliki total 252,000 Customer, dan sebanyak 12.3% merupakan Non Performing Loan.

Non Performing Loan (NPL) adalah kondisi pinjaman dengan kondisi debitur gagal melakukan pembayaran yang dijadwalkan untuk jangka waktu tertentu.

Penyebab terjadinya NPL :

- **Unsur tidak terduga (bencana)**
- **Analisis bank atau loan company tidak tepat**
- **Karakter dari debitur dll**

Reference https://www.rumah.com/panduan-properti/npl-non-performing-loan-53934

PROBLEM STATEMENT

252000
Customer

12.3%
Customer berpotensi NPL

₹ 620 Mio
Total kerugian

Dengan asumsi Debitur meminjam 10% dari income

NPL rate harus diturunkan agar perusahaan tidak mengalami kerugian besar dan bisnis dapat bertahan.

GOALS, OBJECTIVE & BUSINESS METRICS

GOALS

1. Menurunkan kemungkinan kerugian perusahaan.

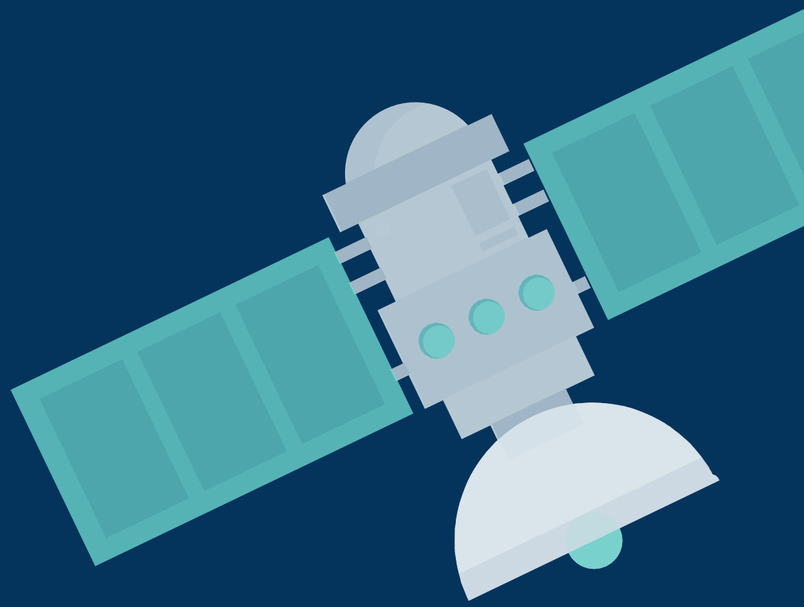
OBJECTIVE

1. Mem-*filter* nasabah yang berkemungkinan gagal bayar dan lancar.
2. Mencari faktor-faktor yang menyebabkan orang bisa gagal bayar.

BUSINESS METRICS

1. Jumlah nasabah *flagged* potensi gagal bayar
2. Conversion rate (jumlah nasabah yg diloloskan untuk mendapatkan hutang)

EDA, INSIGHTS & VISUALIZATION



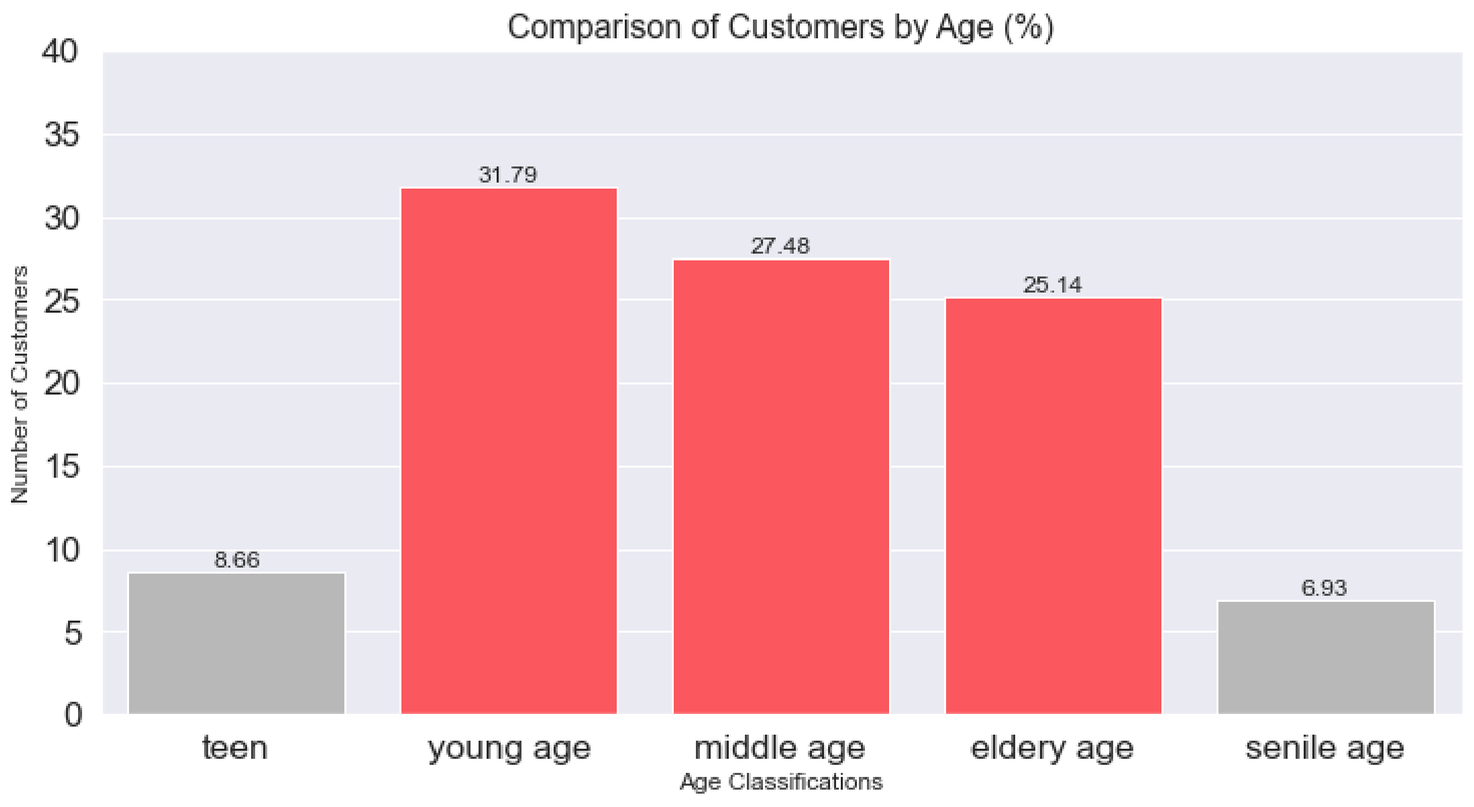
DATA OVERVIEW

Column	Description
income	Income of the user
age	Age of the user
experience	Professional experience of the user in years
profession	Profession
married	Whether married or single
house_ownership	Owned or rented or neither
car_ownership	Does the person own a car
risk_flag	Defaulted on a loan
current*job*years	Years of experience in the current job
current*house*years	Number of years in the current residence
city	City of residence
state	State of residence

Load Data

- Target Label : Risk Flag
- 11 feature
- Categorical features
  a. Married/Single
  b. House Ownership
  c. Car Ownership
  d. Profession
  e.CITY
  f.STATE

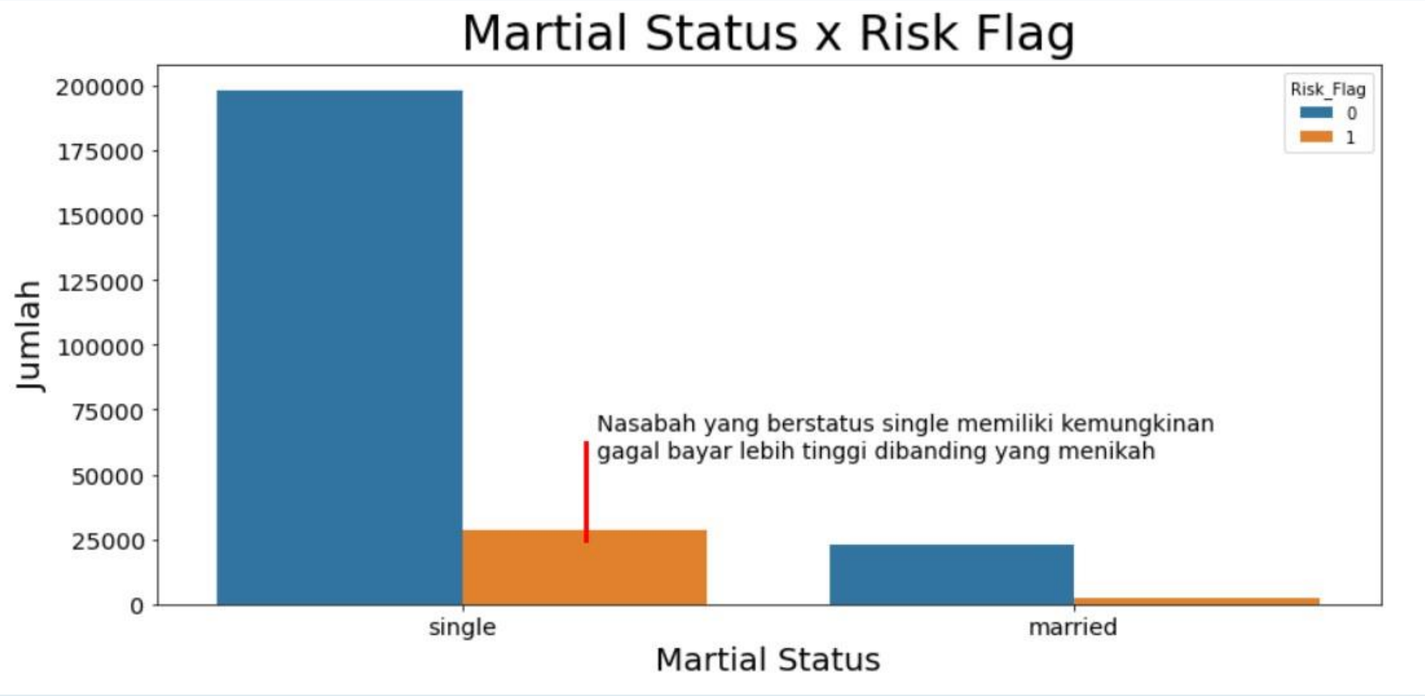
EXPLORATORY DATA ANALYSIS (EDA)



Observation & Insight: Debitur terbanyak terdapat di kategori usia aktif, yaitu young age, middle age dan eldery.

Age categorize refers to Dyussenbayev, A. (2017). Age Periods Of Human Life. Advances in Social Sciences Research Journal.

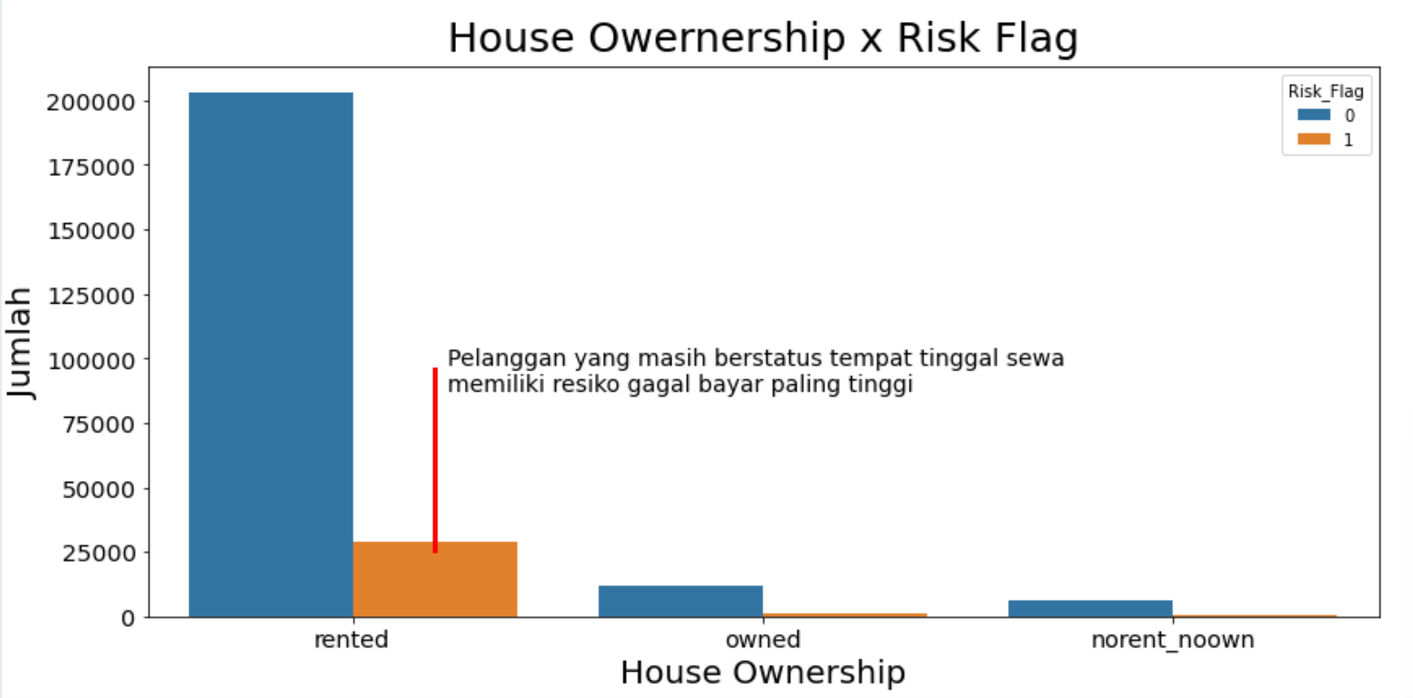
EXPLORATORY DATA ANALYSIS (EDA)



Observation & Insight :

Debitur dengan status single memiliki potensi gagal bayar yang lebih tinggi daripada status married

EXPLORATORY DATA ANALYSIS (EDA)



Observation & Insight:
Debitur yang masih berstatus tempat tinggal sewa memiliki resiko gagal bayar paling tinggi

DATA PRE-PROCESSING

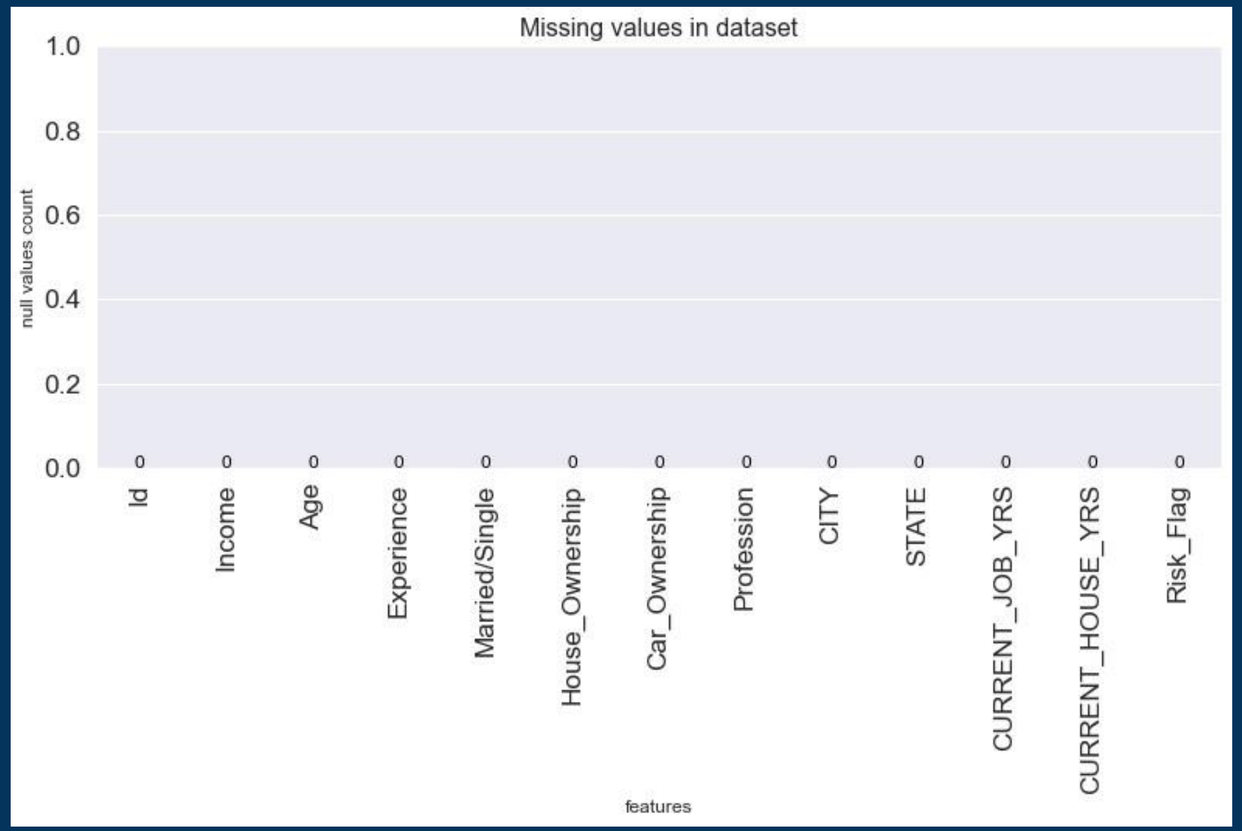
DATA CLEANING

OUR DATA SET

Check Missing Values -> 0 missing value

Check Duplicates -> 0 duplicate

Check Outliers -> 0 outlier



DATA PRE-PROCESSING

1. CATEGORICAL FEATURE

- One Hot Encoding
- Label Encoding
- Frequency Encoding

2. SCALING & TRANSFORMATION

- StandardScaler

3. BALANCING DATA

- SMOTE

5. MODELLING

- Catboost

4. DATA TRAIN & TEST SET

Train : Test = 70 : 30

MACHINE LEARNING, MODELLING & EVALUATION

Model After Hyperparameter Tuning

Model	Accuracy	Precision	Recall	F1-score	AUC	Gini Scoring
Logistic Regression	0.55	0.54	0.59	0.57	0.56	0.12
Decision Tree	0.91	0.88	0.95	0.91	0.93	0.86
Random Forest	0.92	0.88	0.96	0.92	0.92	0.84
XGBoost	0.65	0.66	0.64	0.65	0.72	0.44
CatBoost	0.94	0.91	0.96	0.94	0.96	0.92

CatBoost menjadi model yang dipilih karena:

Model Catboost yang paling fit dengan nilai persentase matrix tertinggi, terutama pada nilai precision dan AUC nya.

Fokusnya pada nilai precision untuk meminimalisir salah duga pada orang yang sebenarnya gagal bayar tapi diprediksi mampu bayar.

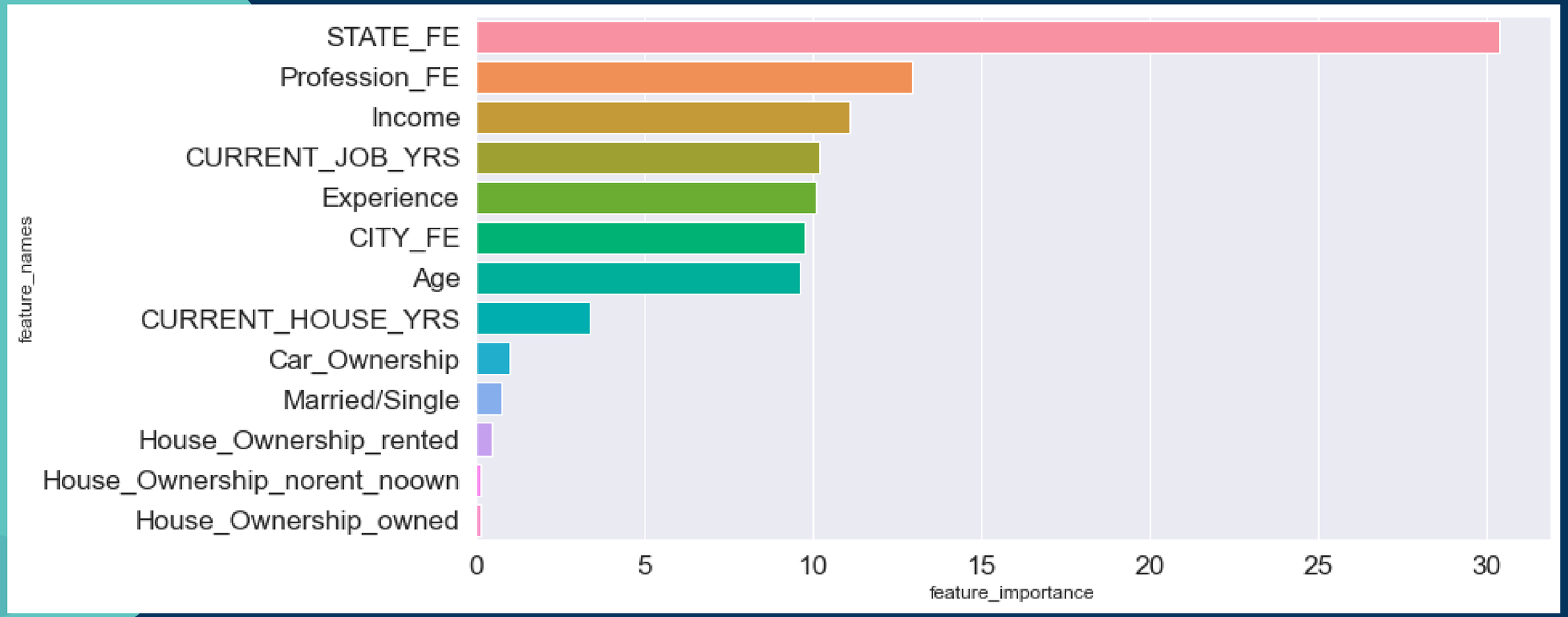
Confusion Matrix

PRIMARY MATRIX: PRECISION
SECONDARY MATRIX: RECALL

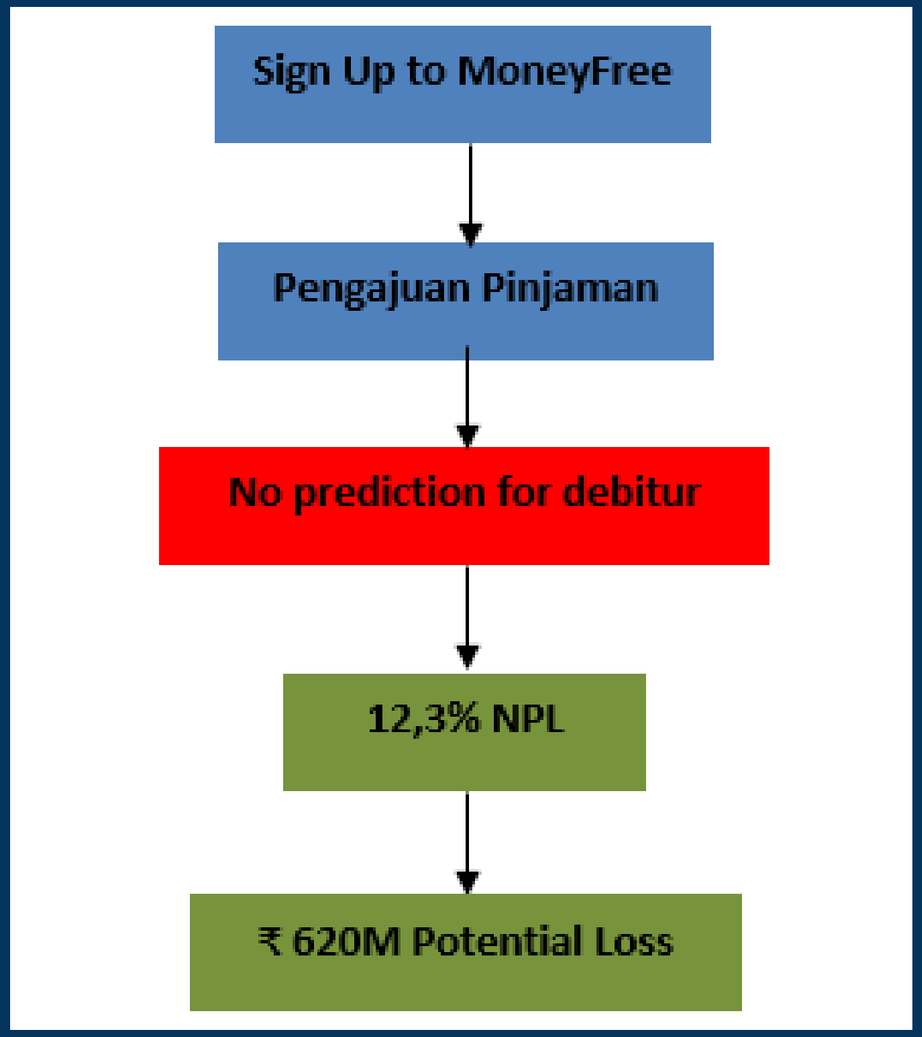
Kita tidak memperbolehkan false positif yang besar karena yang sebenarnya gagal bayar namun dianggap tidak gagal bayar pada model prediksi.

$$Precision = \frac{TP}{TP + FP}$$

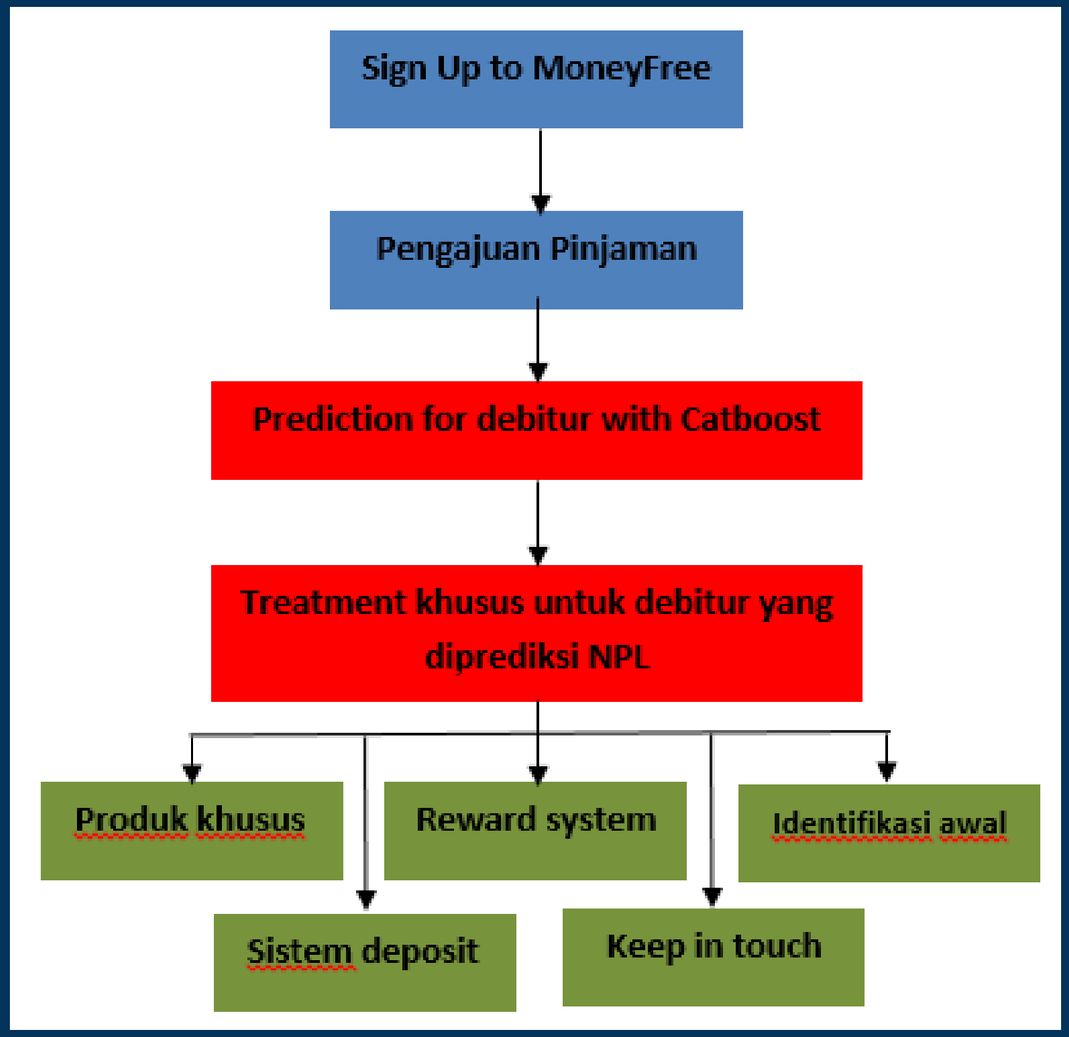
FEATURE IMPORTANCE



BEFORE MACHINE LEARNING MODEL



AFTER MACHINE LEARNING MODEL

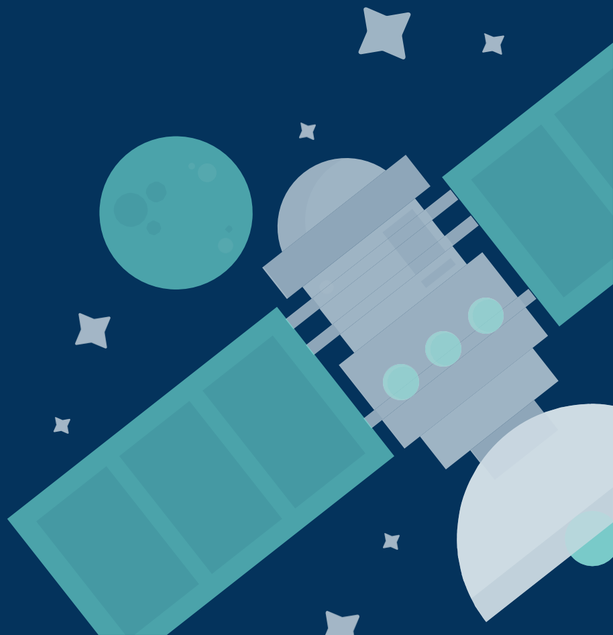


The Decrease of Potential Loss

Before Using Model	After Using Model (Precision: 0,91)
Total Debitur 252.000 (30.996 debitur NPL)	Remaining Debitur: 221004 (debitur tidak NPL)
	Total debitur based on prediction: prediction x remaining debitur 91% x 221004 = 201114 (19890 debitur NPL)
	Setelah dilakukan treatment dan tindak lanjut kepada para debitur NPL, asumsi debitur gagal bayar berkurang menjadi : 20% x 19890 = 3978 (15912 NPL)
(initial) NPL = 12.3%	Jika diasumsikan 15912 debitur NPL, maka rasio NPL : 15912/221004 = 7,2% (turun **5,1%**)
Potential loss & profit	Predicted potential loss akibat NPL yang terselamatkan : ₹ 20000 x 15912 = **₹ 318.240.000**
	Predicted potential profit setelah treatment : ₹ 20000 x 3978 x 7,9% = **₹ 6.285.240** (asumsi bunga pinjaman 7,9%)

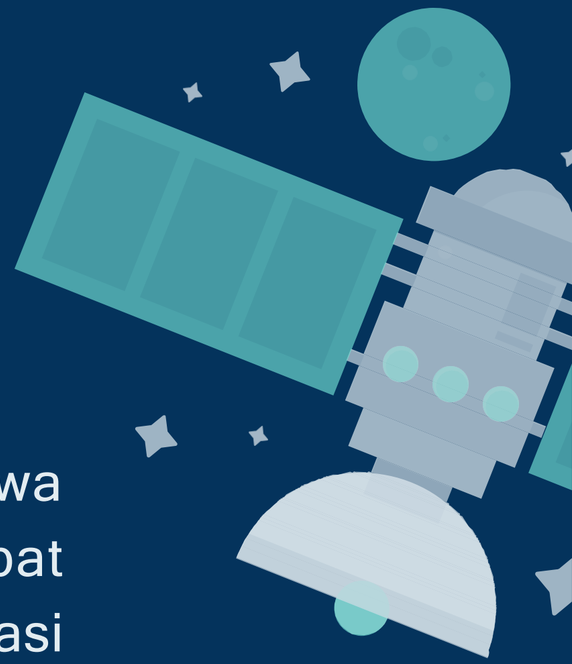
Reference:
https://www.bankbazaar.com/personal-loan.html

BUSINESS INSIGHTS & RECOMMENDATION



Recommendation

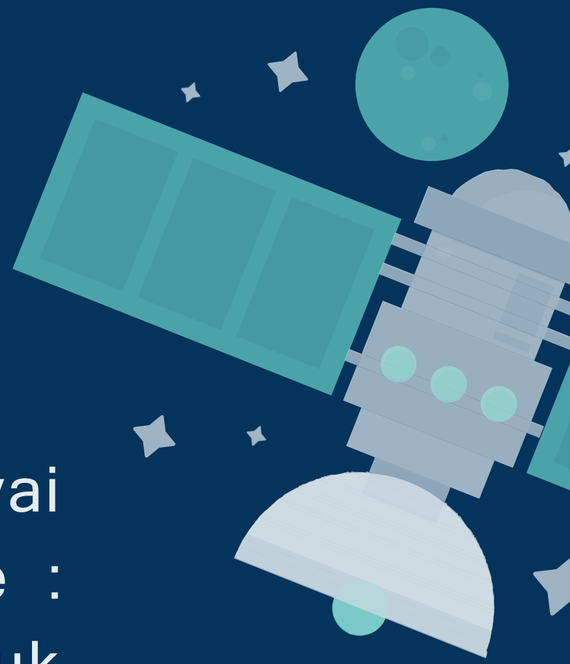
For Business Team



- Dari analisa feature importance dan EDA, dapat disimpulakan bahwa lokasi merupakan faktor yang menjadi risk flag pada debitur. Bank dapat membuat **produk khusus** yang disegmentasikan berdasarkan lokasi debitur (cth : limit rendan dan *payback* fleksibel untuk debitur yang berlokasi di kota-kota besar)
- Memberikan diskon pembayaran atau bonus limit untuk debitur yang membayar tepat waktu (**reward system**).
- Memberlakukan **sistem deposit** untuk pinjaman dengan nilai 10% dari pendapatan debitur

Recommendation

For Account Manager



- **Mengidentifikasi dari awal** debitur-debitur yang mempunyai *risk flag* yang signifikan (berdasarkan feature importance : pendapatan, umur dan lokasi) , agar dapat diarahkan untuk mengambil produk pinjaman khusus.
- Selalu ***stay in touch*** dengan debitur yang telah mengambil pinjaman dengan memberikan info promosi dan *reward* pada produk pinjaman yang telah mereka ambil.

“Data is a precious thing and will last longer than the systems themselves.”