MAHAPRAJA BANK

LOAN PREDICTION AUTOMATION

KINTSUGI

Image: Shutterstock

OUR TEAM

A SALSABILA I.

Data Science Intern

SUBHAN HAIKAL E.

Data Science Intern

ONANG SURYA N.

Data Science Intern

MUHAMMAD REZA R.

Data Science Intern

ANNETTE MICHELLE

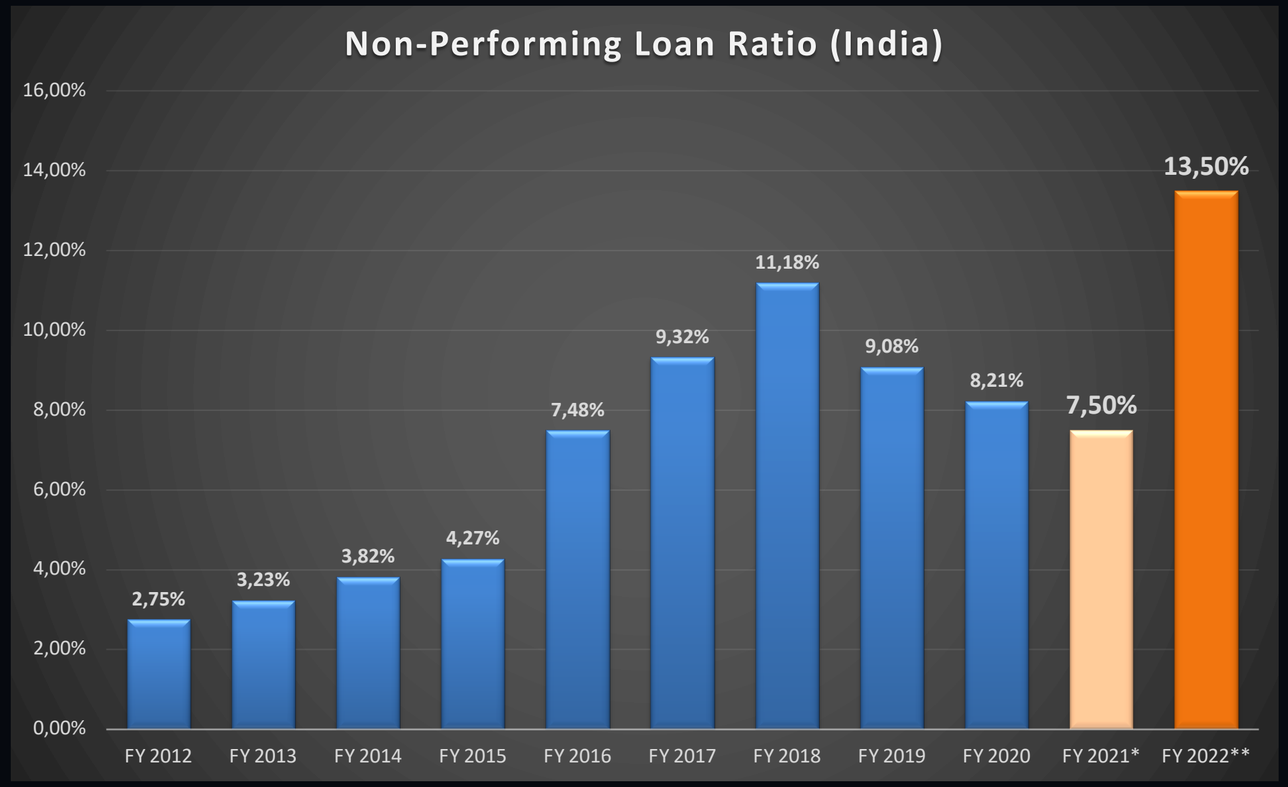
Data Science Intern

TABLE OF CONTENTS

Our presentation for today is divided into five sections.
Each parts are given short description as shown.

BACKGROUND

- Mahapraja Bank is based in India.
- One of their products is personal loan for consumptive purposes that gain high public interests.
- High number of loan application received in such a short time.
- After internal evaluation done by the end of FY 2020, they found out that they experienced a big loss just from this type of credit product because many debtors end up defaulting and their NPL rate was so high compared to other banks' NPL rate or average NPL rate for banks in India.
- On the other hand, because of the pandemic, Mahapraja Bank needs to tighten up their budget by reducing the number of employees but then the assessment would take a longer time to get done.



** FY 2021 as of September 2021*

*** Estimates for FY 2022 are estimates for September 2021 follow the baseline scenario of the Reserve Bank of India*

**** India's financial year starts in April and ends in March.*

For example, FY 2020 refers to data from April 2019 until March 2020.

SOURCE:

https://www.statista.com/statistics/1013267/non-performing-loan-ratio-scheduled-commercial-banks-india/

MAIN ISSUE

It takes

3 to 7

workdays

to manually assessed a single credit application

It needs

5 to 7

people

to complete all the assessment for each state

Mahapraja Bank's NPL rate is:

19,7%

which is way higher than average banks' NPL rate in India (7,5%)
*per September 2021

WHAT SHOULD WE DO?

PROBLEM STATEMENT

How can we help the credit analyst team assessing the loan applications and decreasing the probability of NPL to happen?

OBJECTIVE

Predict which potential debtor with high probability to become a defaulter and minimize the probability of NPL by using a Supervised Machine Learning model

GOAL

Create an automation system to identify potential debtor who have a high probability of default.

BUSINESS METRICS

> Shorten resolve time
> Decrease NPL rate

THE DATASET

> Contains details of all Maharaja Bank's debtors up to Sept 2021 (when the dataset was extracted)

No	Column Name
1	ID
2	Income
3	Age
4	Experience
5	Married/Single
6	House Ownership
7	Car Ownership
8	Profession
9	City
10	State
11	Current Job Years
12	Current House Years
13	**Risk Flag**

Load & Describe Data

- The proportion between unique values in categorical columns is imbalance
- High cardinality for column #8, #9, and #10.

Data Cleansing

> Initially had 250k+ observations but reduced to almost 50k+ because the rest was identified as duplicates.

Scaling

> Create new column whose values are scaled column #2 using MinMaxScaler().

Feature Encoding

> Recategorize column #8 based on the industry.

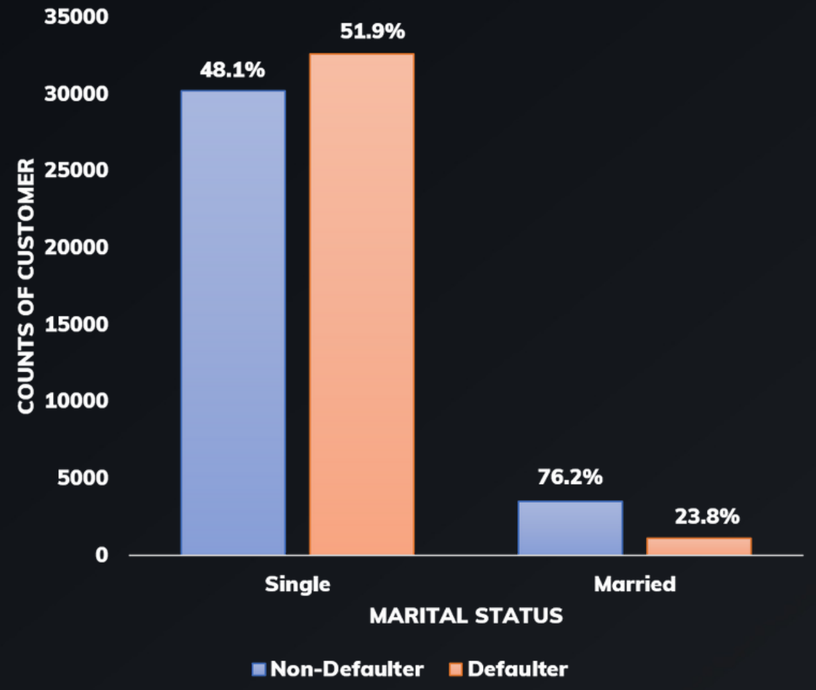
- Recategorize column #10 based on threshold of frequency.
- Apply LabelEncoder() for column #5 and #7.
- Apply OHE for for column #6, #8, and #10.
- Drop column #1, #2, #6, #7, #8, #9, and #10.

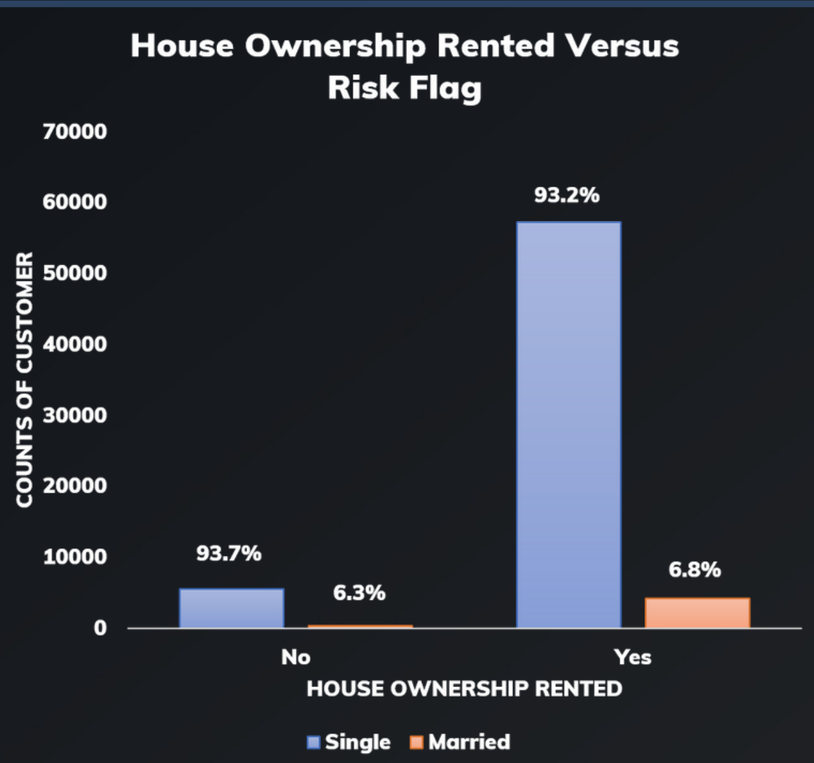
Class Imbalance

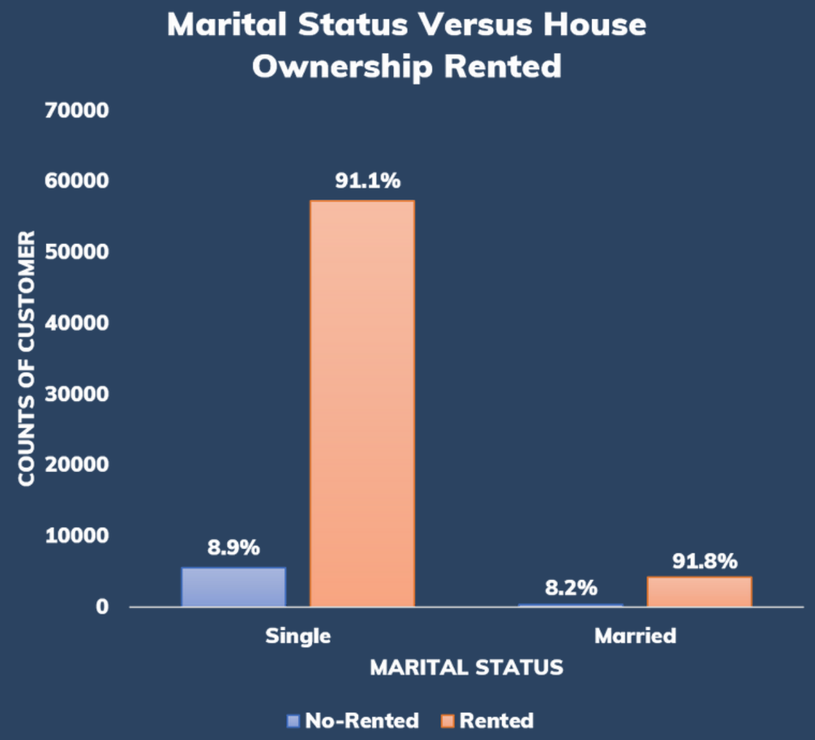
> Oversampling SMOTE with ratio = 1.

Single customers are prone to be defaulters. Why?

About 90% of single-status customers have rented a house and about a half tend to be defaulters. The correlation between house ownership and marital status are plausible as it makes single customers prone to have lower net income after paying the rent bills.

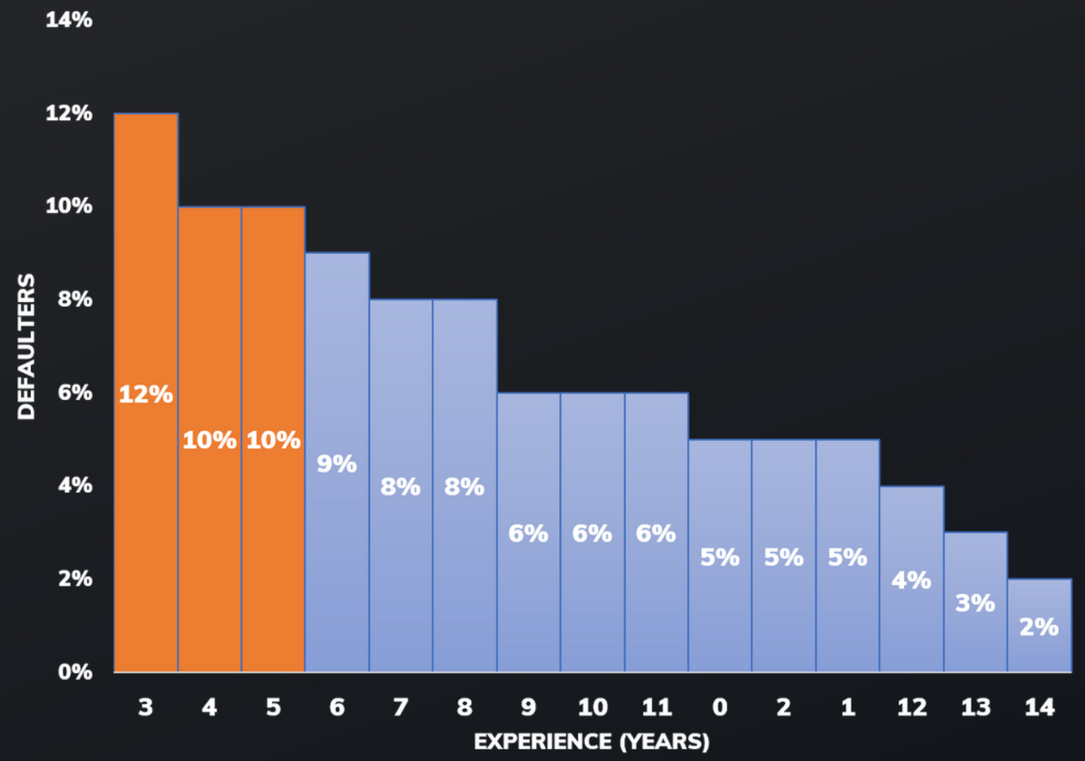






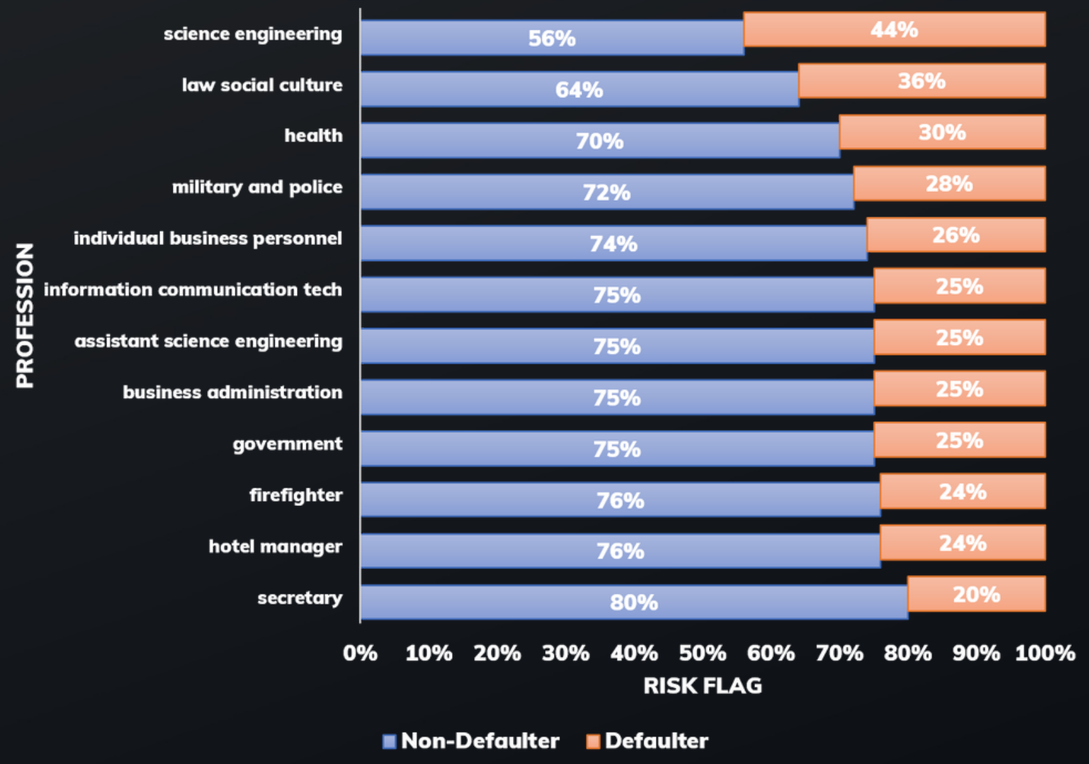
Current Job Years vs Risk Flag

Customer with working period between 3 to 5 years tend to be a defaulter, accounted for more than one-to-ten of the total population.



Current Job Years vs Risk Flag

Customer who work in science engineering, law social culture, and health industry is prone to be a defaulter with the proportion for each industry is at least 30%.



Referring to multivariate analysis, it was found that all features did not show strong linear correlation to target.

Thus, the data must be treated using classification model

MODELING

We chose KNN
(no hyperparameter tuning and/or drop some feature) as the best model because:

- High recall value but still get a good number and proportion for accuracy & precision (not too imbalance)
- Short execution time

MODEL	ACCURACY	PRECISION	RECALL	EXECUTION TIME
DECISION TREE	0,77	0,76	0,79	393 ms
DECISION TREE (HYPERPARAMETER TUNING)	0,77	0,78	0,76	65,5 sec
DECISION TREE (NO PROFESSION)	0,73	0,72	0,75	344 ms
DECISION TREE (NO PROFESSION, HYPERPARAMATER TUNING)	0,72	0,71	0,75	65,7 sec
K-NEAREST NEIGHBORS	**0,77**	**0,73**	**0,87**	**4,2 sec**
K-NEAREST NEIGHBORS (HYPERPARAMETER TUNING)	0,74	0,67	0,95	40,7 sec
K-NEAREST NEIGHBORS (NO PROFESSION)	0,76	0,75	0,79	4,33 sec
K-NEAREST NEIGHBORS (NO PROFESSION, HYPERPARAMETER TUNING)	0,75	0,69	0,90	30,7 sec

MODELING

- Its total percentage of observations that predicted correctly is the highest among all models.
- It has one of the lowest percentage of false negative among all the models.
- The proportion between false negative & false positive is not too jarring.

NOTE:

The testing data was consisted of 20.234 observations

MODEL	TRUE POSITIVE (%)	TRUE NEGATIVE (%)	FALSE POSITIVE (%)	FALSE NEGATIVE (%)
DECISION TREE	39,47	37,42	12,56	10,53
DECISION TREE (HYPERPARAMETER TUNING)	38,02	38,98	11,00	11,98
DECISION TREE (NO PROFESSION)	37,68	35,14	14,84	12,32
DECISION TREE (NO PROFESSION, HYPERPARAMATER TUNING)	37,70	34,73	15,25	12,30
K-NEAREST NEIGHBORS	**43,75**	**33,71**	**16,27**	**6,25**
K-NEAREST NEIGHBORS (HYPERPARAMETER TUNING)	47,27	26,73	23,25	2,73
K-NEAREST NEIGHBORS (NO PROFESSION)	39,70	36,40	13,58	10,30
K-NEAREST NEIGHBORS (NO PROFESSION, HYPERPARAMETER TUNING)	45,13	29,58	20,40	4,87

POTENTIAL IMPACT

LET'S ASSUME:
Loan application received for one state= 50/day (in average)
Number of analysts per state = 5 people
Avg. salary for Credit Analyst in India = 601.663 rupee/year (Sept 2021)
Workdays in a month = 25 days
Workhour per week = 40 hours per week
Effective workdays = 48 weeks per year

RESOLVE TIME

BEFORE

ESTIMATION FOR NUMBER OF RESOLVED APPLICATION:
- The fastest = (25-3) workdays x 50 application = 22 x 50 = 1.100 application/month
- The slowest = (25-7) workdays x 50 application = 18 x 50 = 900 application/month
- In average, the team has to process **1.000 application/month**

EXPENSE:
- Annual salary cost = 5 x 601.633 = **3.008.315 rupee/year**
- Productivity rate @ analyst = (number of application assessed / man hour)
  = (1/000 x 12) / (40 x 48) = 12.000 / 1.920 = **6,25 application/hour (expected)**

AFTER

ESTIMATION FOR NUMBER OF RESOLVED APPLICATION:
- The fastest = (25-1) workdays x 50 application = 24 x 50 = 1.200 application/month
- The slowest = (25-2) workdays x 50 application = 23 x 50 = 1.150 application/month
- In average, the team has to process **1.175 application/month**.

EXPENSE:
- Annual salary cost = 3 x 601.633 = **1.804.989 rupee/year**
- Productivity rate @ analyst = number of application assessed / man hour
  = (1.175 x 12) / (40 x 48) = 14.100 / 1.920 = **7,34 application/hour (expected)**

POTENTIAL IMPACT

NPL RATE

LET'S ASSUME:
Loan application approved for a year = 10.000 (est. for FY 2022)
Loan balance @ application = 7.500.000 rupee (average)
NPL rate per FY 2020 = 19,7%
NPL rate model (predicted) = 6,25%
Interest (fixed rate) for 10 years = 10% / year
Debtor fail to pay around the 5th year

EXPECTED PROFIT:

- Total interest payable by the 10th year = 10% x 7.500.000 = 750.000 rupee/debtor
- Total expected profit from the interest by the 10th year = 750.000 x 10.000 = **7,5 billion rupee**

BEFORE

POTENTIAL LOSS:

- Number of defaulters = 19,7% x 10.000 = **1.970 defaulters**
- Total interest payable by the 5th year = 50% x 7.500.000.000 = 3,75 billion rupee
- Total expected profit from the interest after the 5th year = (50% x 750.000) x (10.000 – 1.970)
  = 3.011.250.000 rupee

MARGIN: 7.500.000.000 - (3.750.000.000 + 3.011.250.000) = **738.750.000 rupee**

AFTER

POTENTIAL LOSS:

- Number of defaulters = 6,25% x 10.000 = **625 defaulters**
- Total interest payable by the 5th year = 50% x 7.500.000.000 = 3,75 billion rupee
- Total expected profit from the interest after the 5th year = (50% x 750.000) x (10.000 – 625)
  = 3.515.625.000 rupee

MARGIN: 7.500.000.000 - (3.750.000.000 + 3.515.625.000) = **234.375.000 rupee**

RECOMMENDATION

Based on the insights obtained at the EDA stage and the model's outcome, we would like to recommend several things, namely:

1

To cover possible losses due to loss of potential customers, we can forward our findings from the EDA process to the marketing team so that they can focus on promoting credit products (especially the individual consumer loans type) to people with the following criteria:

- Married
- Own a house
- Minimum of 9 years working period in their current company
- Work for government or in hospitality field

2

To improve the performance and accuracy of the model, we strongly recommend the integration of basic data, credit history, and credit scores of potential customers.

THANK YOU

KINTSUGI DATA TEAM