Predicting Loan Default Using Various Machine Learning Algorithms

Dery Purnama Saefudin

Outline

Objective & Data

01

02

Exploratory Data Analysis

Data Preprocessing

03

04

Model Preparation & Modeling

Business Impact

05

06

Recomendation

Business Understanding

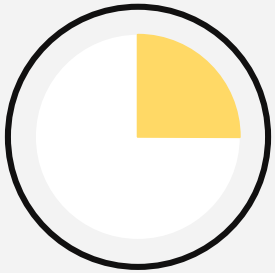
Loan Default is a *PROBLEM* for Financial Industry.

The *failure of a debtors to repay a loan* according to the terms and conditions outlined in the loan agreement.

Objective

Current

6.188 debtors are current



22%

Default

22.313 debtors are default

78%

Obtain Insight of Loan Status

Create Machine Learning Prediction

Data

Before Cleaning

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32581 entries, 0 to 32580
Data columns (total 12 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   person_age                  32581 non-null  int64
 1   person_income               32581 non-null  int64
 2   person_home_ownership       32581 non-null  object
 3   person_emp_length           31686 non-null  float64
 4   loan_intent                 32581 non-null  object
 5   loan_grade                  32581 non-null  object
 6   loan_amnt                   32581 non-null  int64
 7   loan_int_rate               29465 non-null  float64
 8   loan_status                 32581 non-null  int64
 9   loan_percent_income         32581 non-null  float64
 10  cb_person_default_on_file   32581 non-null  object
 11  cb_person_cred_hist_length  32581 non-null  int64
dtypes: float64(3), int64(5), object(4)
memory usage: 3.0+ MB
```

- Retrieved from https://www.kaggle.com/datasets/laotse/credit-risk-dataset/data

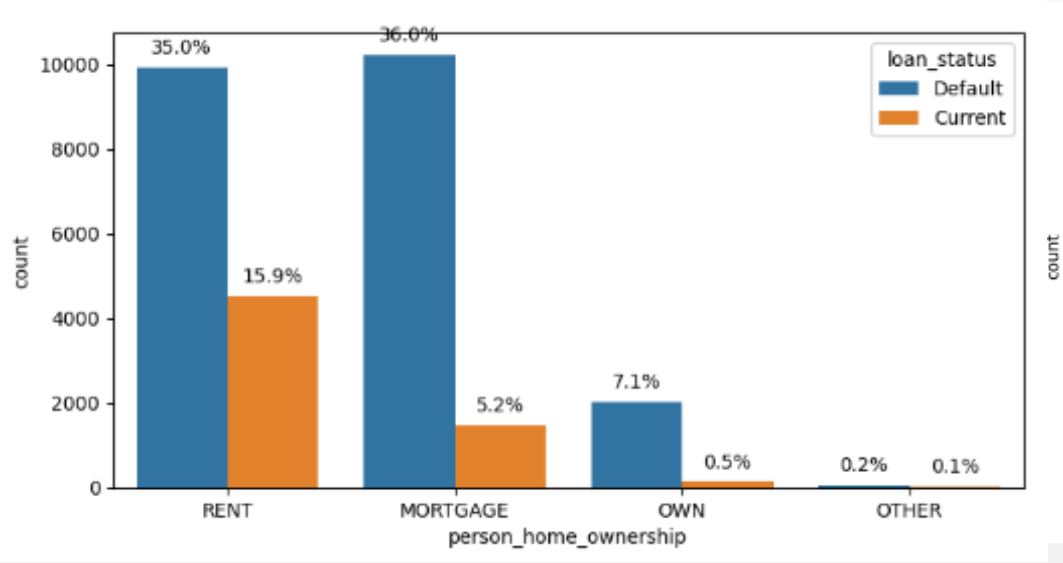
After Cleaning

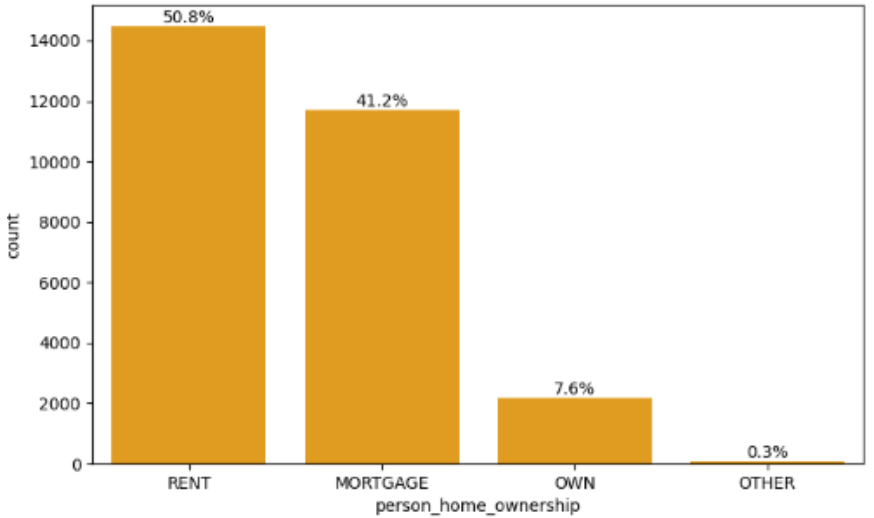
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 28501 entries, 0 to 32580
Data columns (total 12 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   person_age                  28501 non-null  int64
 1   person_income               28501 non-null  int64
 2   person_home_ownership       28501 non-null  object
 3   person_emp_length           28501 non-null  float64
 4   loan_intent                 28501 non-null  object
 5   loan_grade                  28501 non-null  object
 6   loan_amnt                   28501 non-null  int64
 7   loan_int_rate               28501 non-null  float64
 8   loan_status                 28501 non-null  int64
 9   loan_percent_income         28501 non-null  float64
 10  cb_person_default_on_file   28501 non-null  object
 11  cb_person_cred_hist_length  28501 non-null  int64
dtypes: float64(3), int64(5), object(4)
memory usage: 2.8+ MB
```

- The data has at most *32,581 rows and 12 columns* before cleaning
- There are *28,501 rows and 12 columns* after cleaning

Exploratory Data Analysis

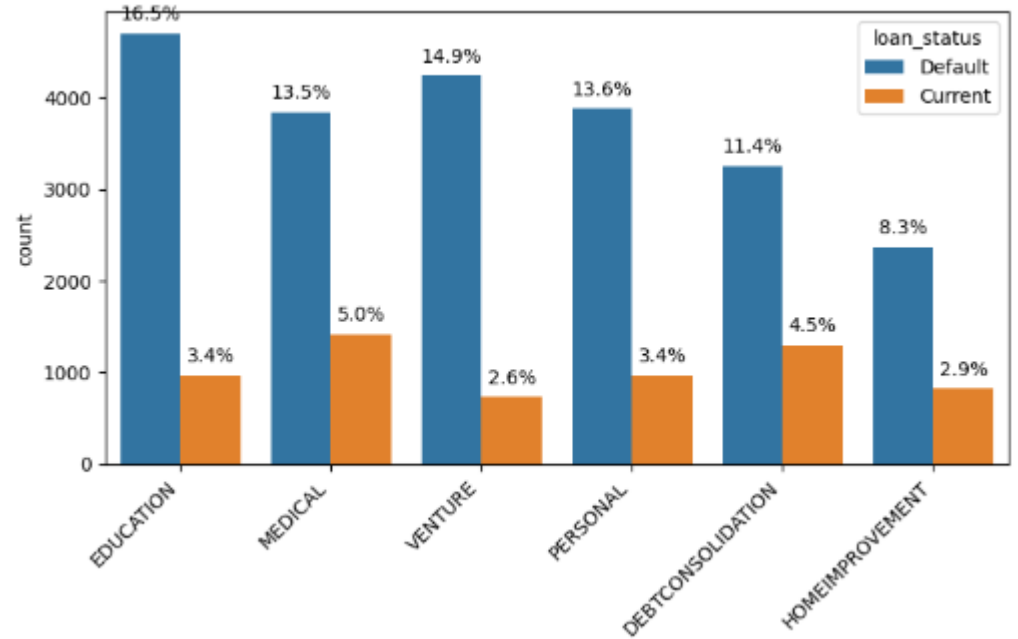
Home Ownership by Loan Status

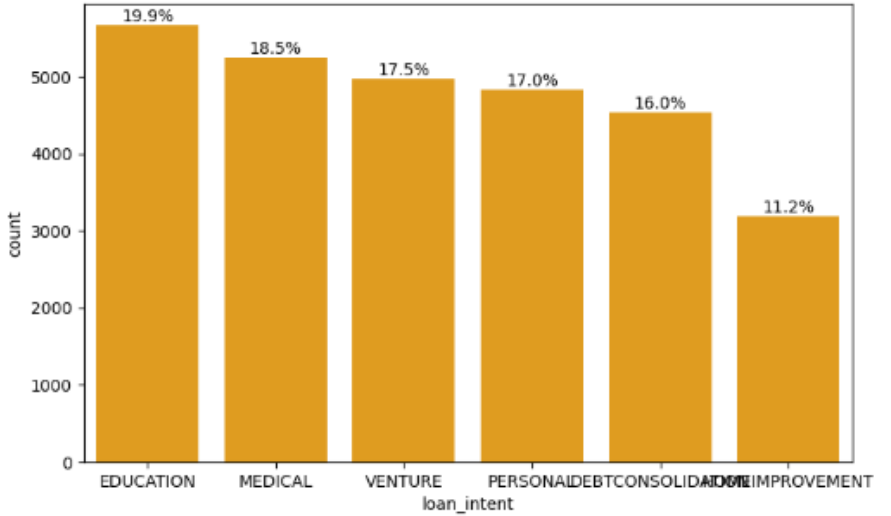




- Most of Debtors' home ownership are Rent (50.8%)
- Debtors who default the most are Occupier (15.9%)

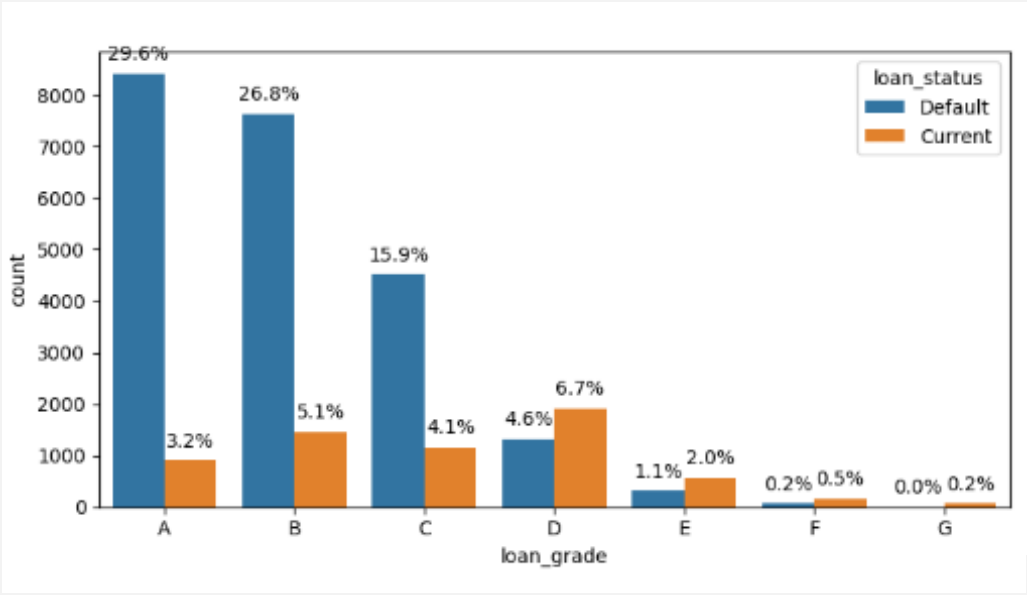
Loan Intent by Loan Status

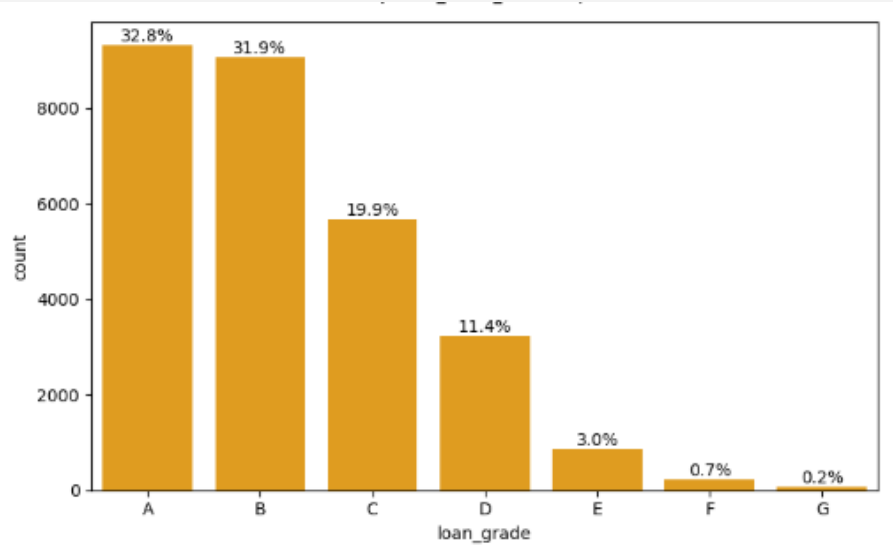




- The most common purpose of loan is Education (19.9%).
- Debtors who the most default are those with Education as the purpose (16.5%)

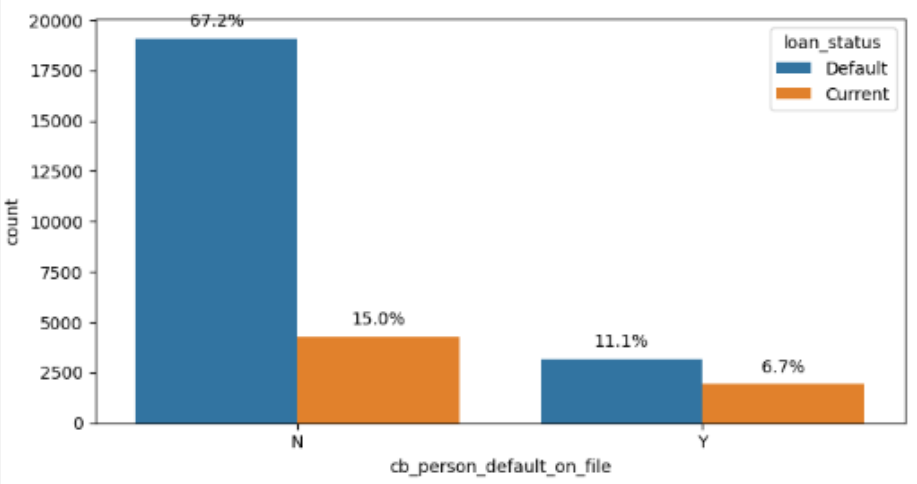
Loan Grade by Loan Status

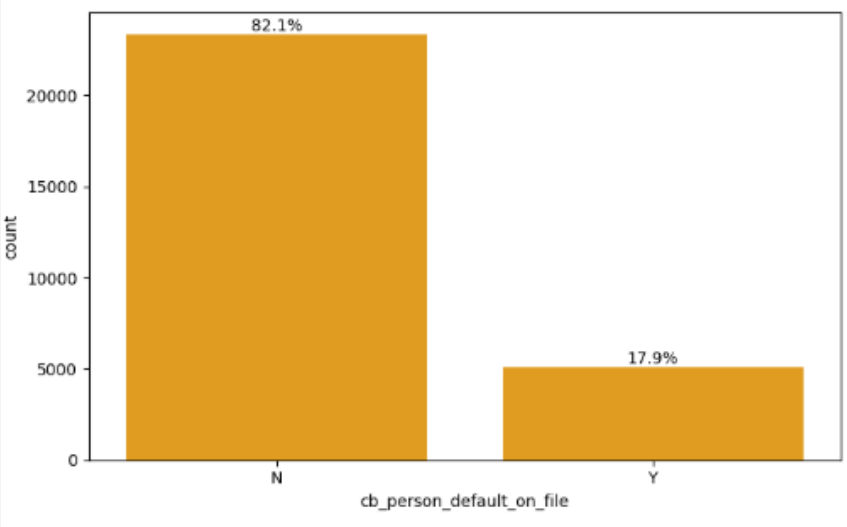




- The most common loan grade is Grade A (32.8%).
- Debtors who the most default are those with Loan Grade A (29.6%)

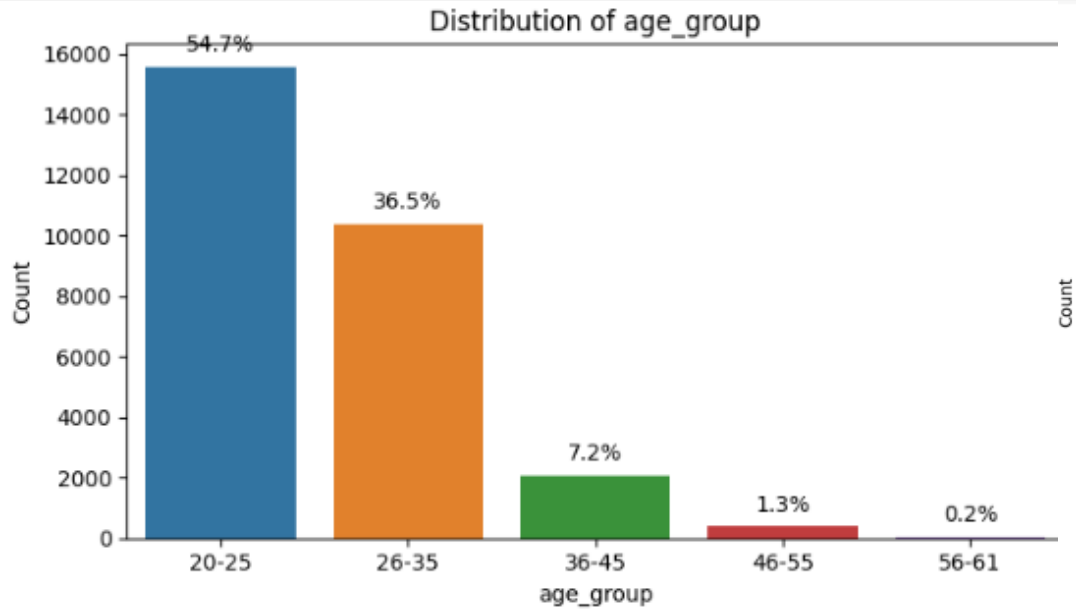
Default History by Loan Status

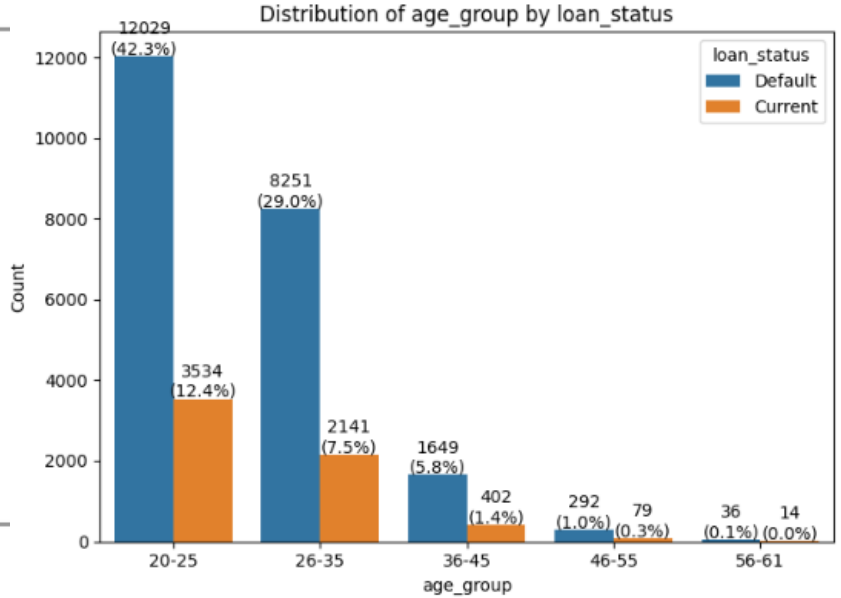




- The most debtors do not have loan default history (82.1%)
- Debtors who the most default are those without loan default history (67.2%)

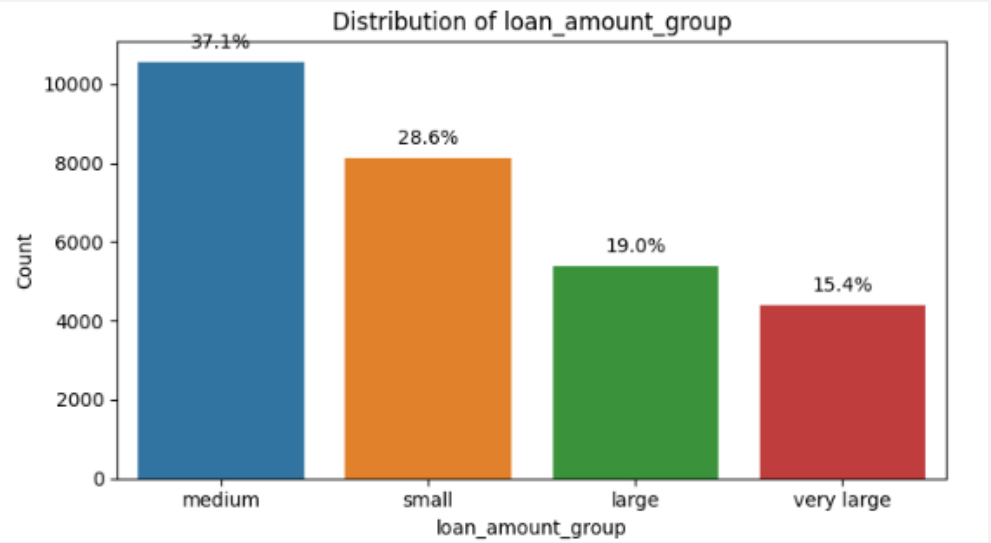
Age Group by Loan Status

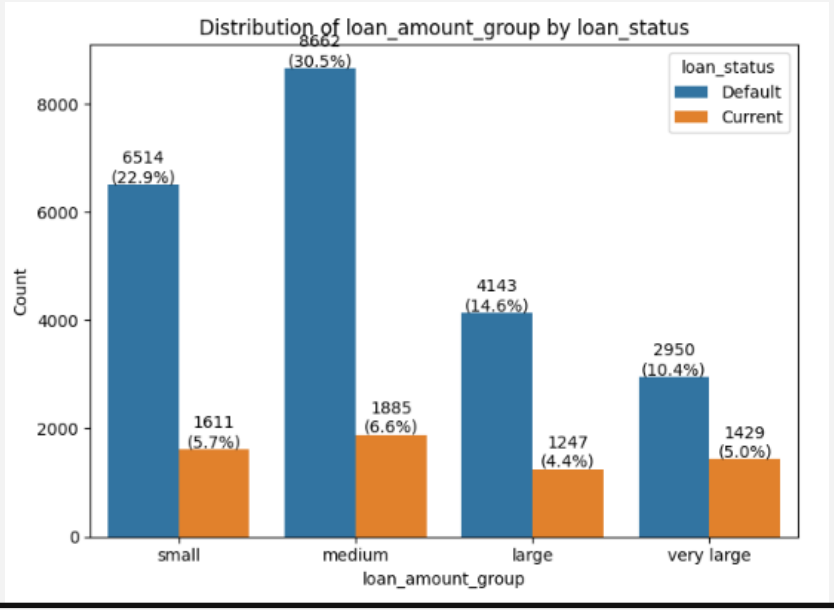




- The mostdebtors fall within the **20-25 age range (54.7%)**
- Debtors who the **most default are those** within the **20-25 age range (42.3%)**

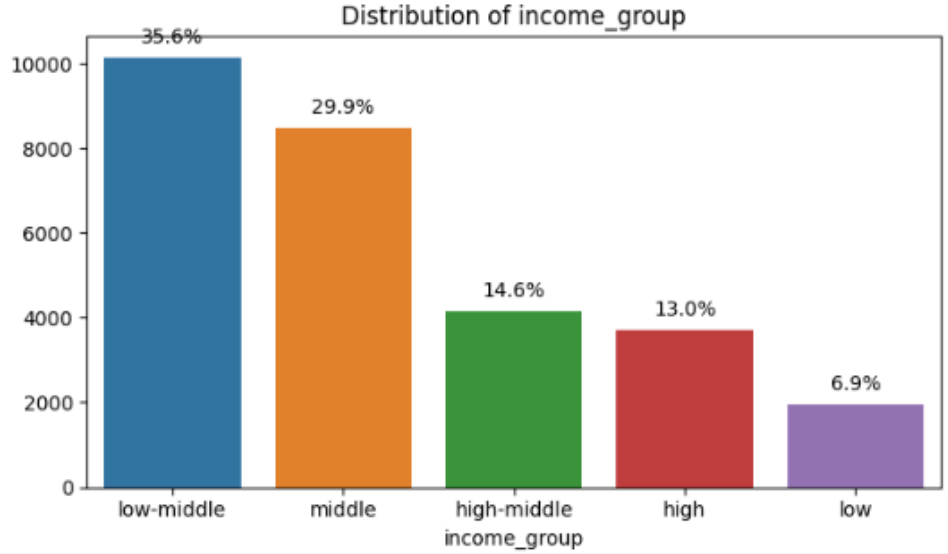
Loan Amount Group by Loan Status

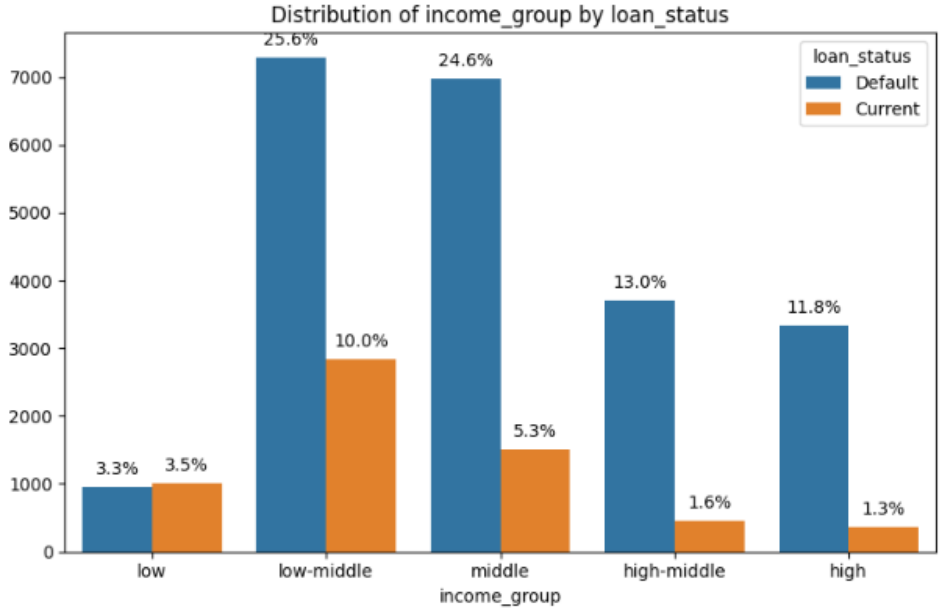




- The most common debtors are those with loan in the range of **5,000 - 10,000 / medium (37.1%).**
- Debtors who the **most default are those** with loan in the range of **5,000 - 10,000 / medium (22.9%)**

Income Group by Loan Status





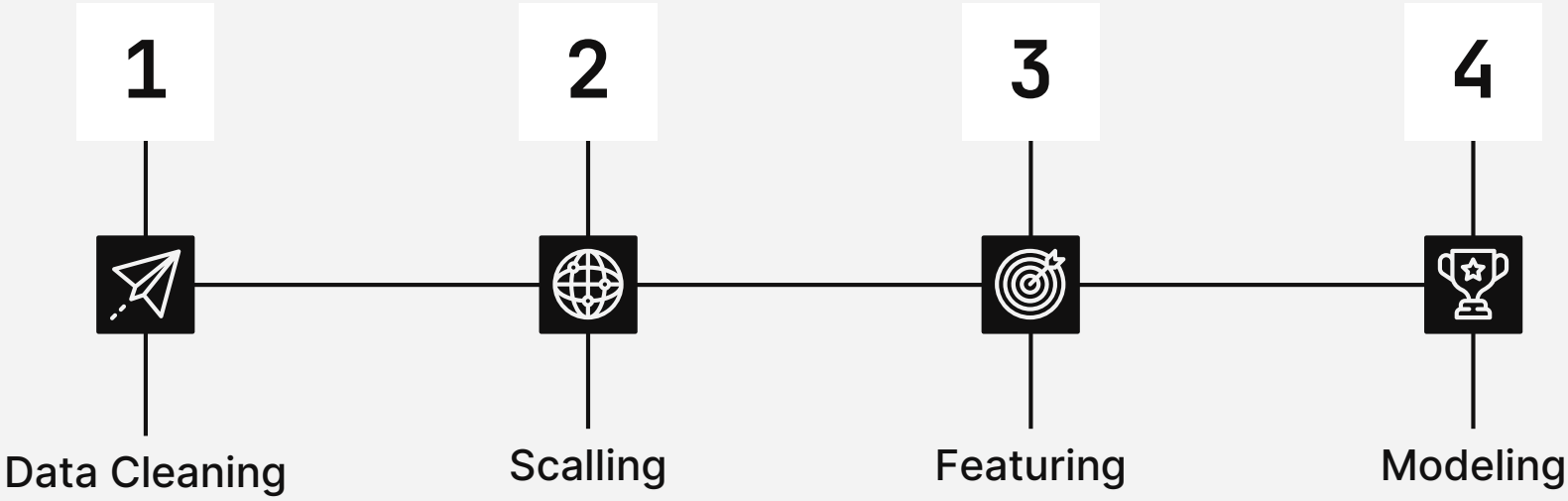
- The most common debtors are those with income in the range of **5,000 - 10,000 / medium (35.6%).**
- Debtors who the **most default are those** with loan in the range of **5,000 - 10,000 / medium (25.6%)**

Features Engineering

After Encoding

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 28441 entries, 0 to 28440
Data columns (total 44 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   person_age                      28441 non-null  int64
 1   person_income                   28441 non-null  int64
 2   person_emp_length               28441 non-null  float64
 3   loan_amnt                       28441 non-null  int64
 4   loan_int_rate                   28441 non-null  float64
 5   loan_status                     28441 non-null  int64
 6   loan_percent_income             28441 non-null  float64
 7   cb_person_cred_hist_length      28441 non-null  int64
 8   loan_to_income_ratio            28441 non-null  float64
 9   loan_to_emp_length_ratio        28441 non-null  float64
 10  int_rate_to_loan_amt_ratio      28441 non-null  float64
 11  person_home_ownership_MORTGAGE  28441 non-null  int64
 12  person_home_ownership_OTHER     28441 non-null  int64
 13  person_home_ownership_OWN       28441 non-null  int64
 14  person_home_ownership_RENT      28441 non-null  int64
 15  loan_intent_DEBTCONSOLIDATION   28441 non-null  int64
 16  loan_intent_EDUCATION           28441 non-null  int64
 17  loan_intent_HOMEIMPROVEMENT     28441 non-null  int64
 18  loan_intent_MEDICAL             28441 non-null  int64
 19  loan_intent_PERSONAL            28441 non-null  int64
 20  loan_intent_VENTURE             28441 non-null  int64
 21  loan_grade_A                    28441 non-null  int64
 22  loan_grade_B                    28441 non-null  int64
 23  loan_grade_C                    28441 non-null  int64
 24  loan_grade_D                    28441 non-null  int64
 25  loan_grade_E                    28441 non-null  int64
 26  loan_grade_F                    28441 non-null  int64
 27  loan_grade_G                    28441 non-null  int64
 28  cb_person_default_on_file_N     28441 non-null  int64
 29  cb_person_default_on_file_Y     28441 non-null  int64
 30  income_group_low                28441 non-null  int64
 31  income_group_low-middle         28441 non-null  int64
 32  income_group_middle             28441 non-null  int64
 33  income_group_high-middle        28441 non-null  int64
 34  income_group_high               28441 non-null  int64
 35  age_group_20-25                 28441 non-null  int64
 36  age_group_26-35                 28441 non-null  int64
 37  age_group_36-45                 28441 non-null  int64
 38  age_group_46-55                 28441 non-null  int64
 39  age_group_56-61                 28441 non-null  int64
 40  loan_amount_group_small         28441 non-null  int64
 41  loan_amount_group_medium        28441 non-null  int64
 42  loan_amount_group_large         28441 non-null  int64
 43  loan_amount_group_very large    28441 non-null  int64
dtypes: float64(6), int64(38)
memory usage: 9.5 MB
```

Model Preparation



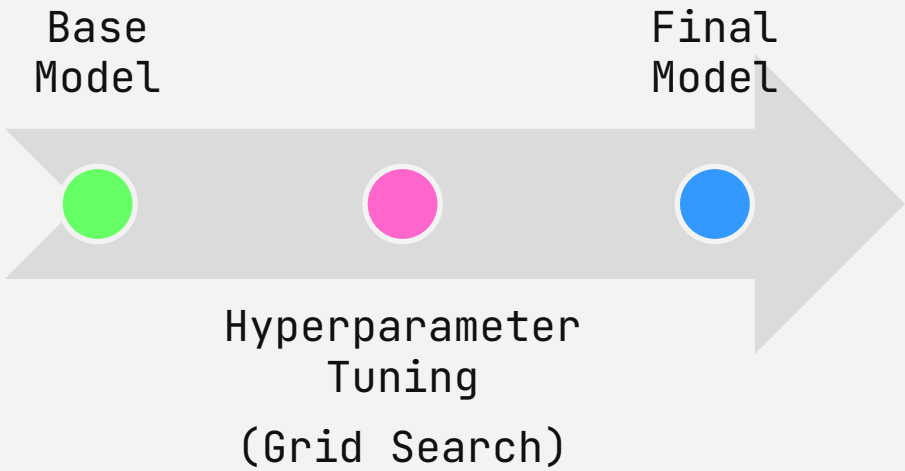
Modeling

Random Forest

Logistic Regression

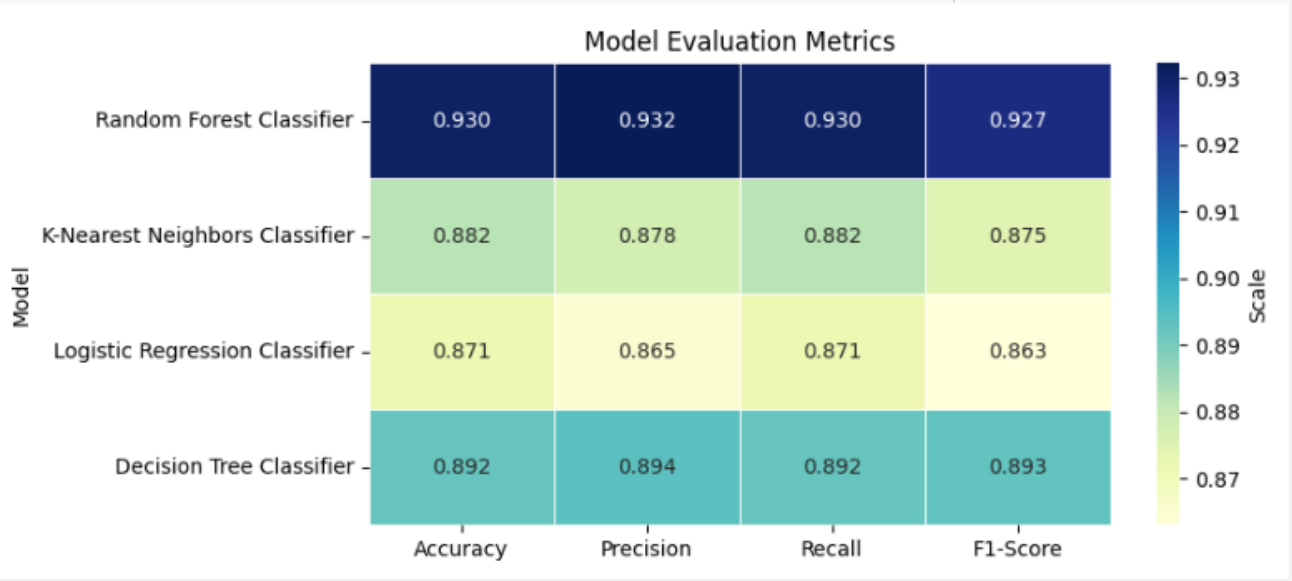
K-Neighbour

Decission Tree



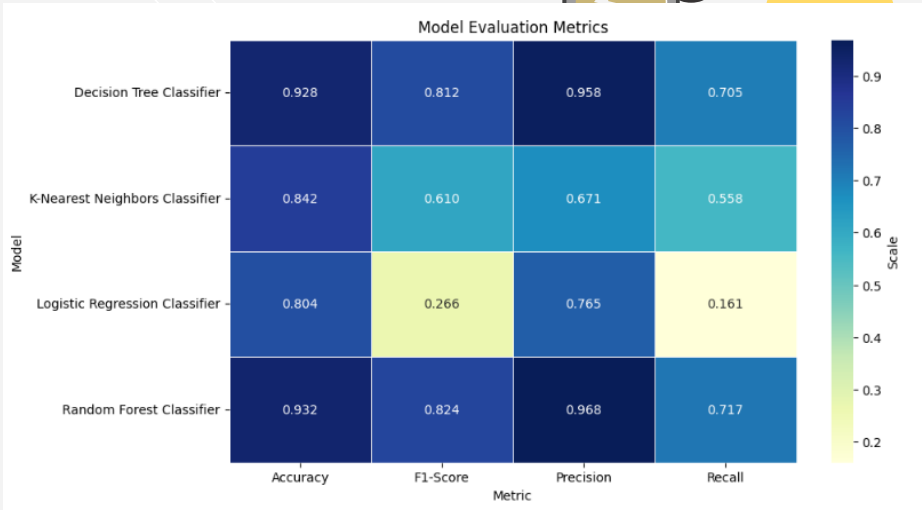
Model

Before Tuning



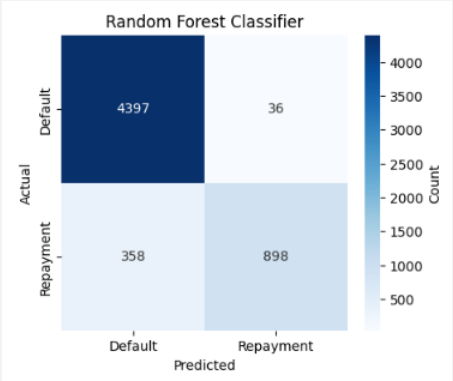
Random Forest
Precision (0.932)
Recall (0.892)

After Tuning



96 %

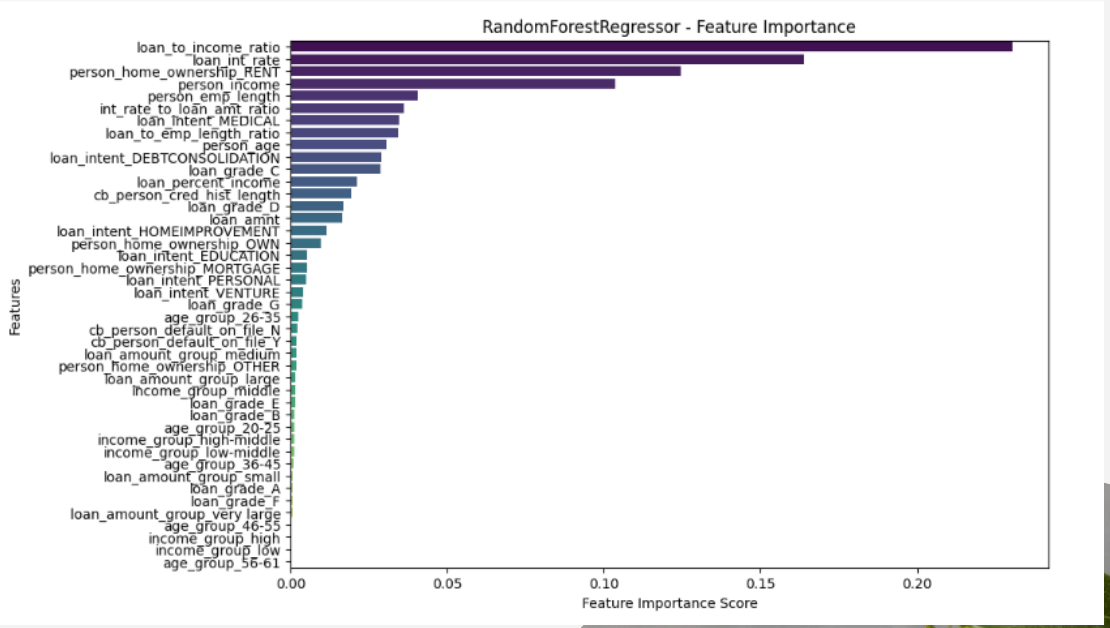
Model can predict the true default debtors correctly



Business Impact

- The model can correctly predict 96 of the 100 default case and 29 default case can not be predicted correcly.
- If on average the bank's loss due to the default case is Rp 100 million per case & month, then using this model, the total potential losses that could be prevented are Rp 4.2 billion.

Feature Important



Loan to Income Ratio (LTI)

is dominant factor for defaults

Interest Rate, Occupier and Person Income

are significant contributors to defaults.

Business Recomendation

Improved risk management strategies :

- Adjusting loan principal limits (LTI < 35%)
- Adjusting interest rates
- Requesting additional documentation
- Adding additional approval processes
- Targeting Low Risk Market Segment