Credit Card Approval Prediction

R. Bimo Mandala Putra

Ways Bank Make Money

- Interest income
- Capital markets income
- Fee-based income

Interest Income

[How Do Banks Make Money? - Overview, Forms, Examples (corporatefinanceinstitute.com)](corporatefinanceinstitute.com)

Problem

By assuming 1 person creditor lending Rp 1.000.000 with interest 5%.

Credit Loan	**: Rp 1000.000**	**Risk lose**	**: ± Rp 1000.000**
Interest	**: 5%**	**Profit**	**: Rp 50.000**

https://www.investopedia.com/terms/l/lossgivendefault.asp

Target & Goals

01 Problem

Loss because of Bad User

02 Goals

Prevent Potential Bad User aproval credit card

03 Objective

Identify Bad User and predict it using ML models

04 Result

Eliminate or reduce loss from Bad User

Data Pre-Processing & Analysis

application_record.csv

438.557 Rows

18 Columns

Duplicate data : 12 Rows
Missing Values :
134203 Rows on Occupation type

credit_record.csv

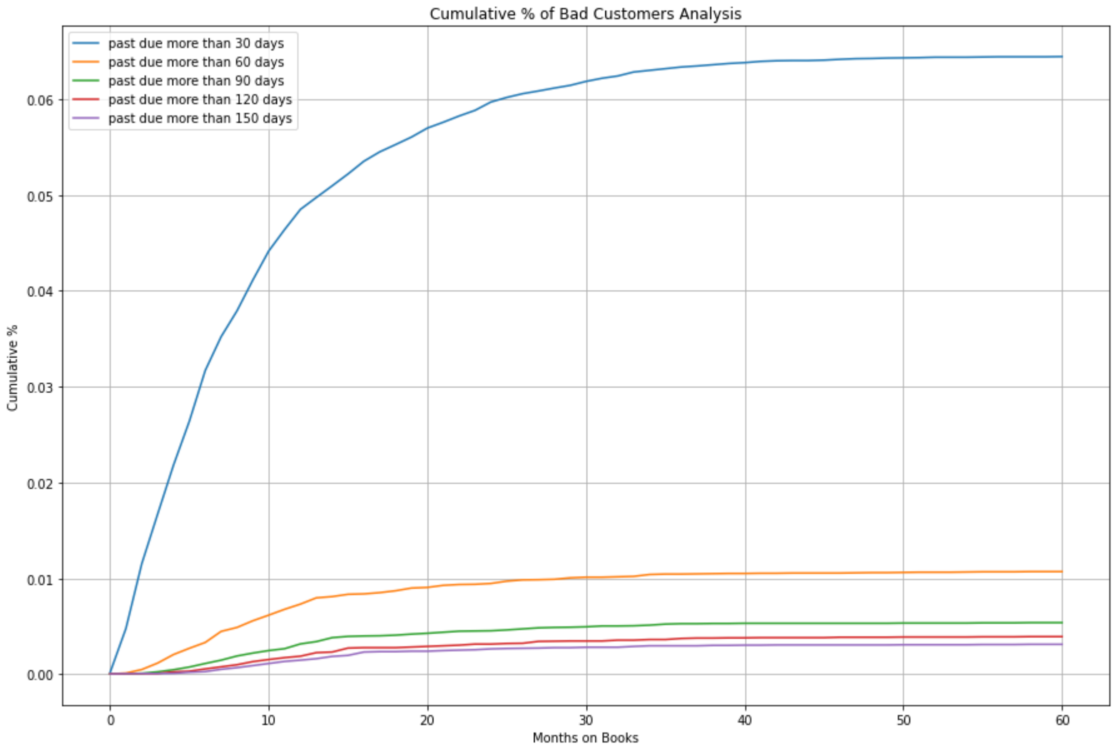
1.048.574 Rows

3 Columns

Duplicate data : 0 Rows
Missing Values : 0 Rows

Data Pre-Processing & Analysis

Determine Target with Vintage analysis



User 30 Days past due looks higher than other user, it happens more often.
We can’t just say user who 30 past days are Bad User

User 60 Past due or more will be categorized as Bad User

Feature Engineering

New feature column

Age

Define Age of user, Transformed from `days_birth` column divided by 365 days

Months_in_books

Define How many Months user join Credit service, Pivoted from `Credit.csv`

Status (Target)

Define whatever user are Bad User or Good user,

Score =< 60 past due =Bad score
Score < 30past due = Good Score

Good Score>Bad Score = Good User
Good Score<Bad Score = Bad User

Feature Engineering

Encoding

Binary Convert the negative value to 0 and the positive value to 1

More than 2 Convert the Column with One Hot Encoding Technique

Imbalance

Extremely Imbalance Data between Good & Bad User

Over Sampling with Smooth

Scaling

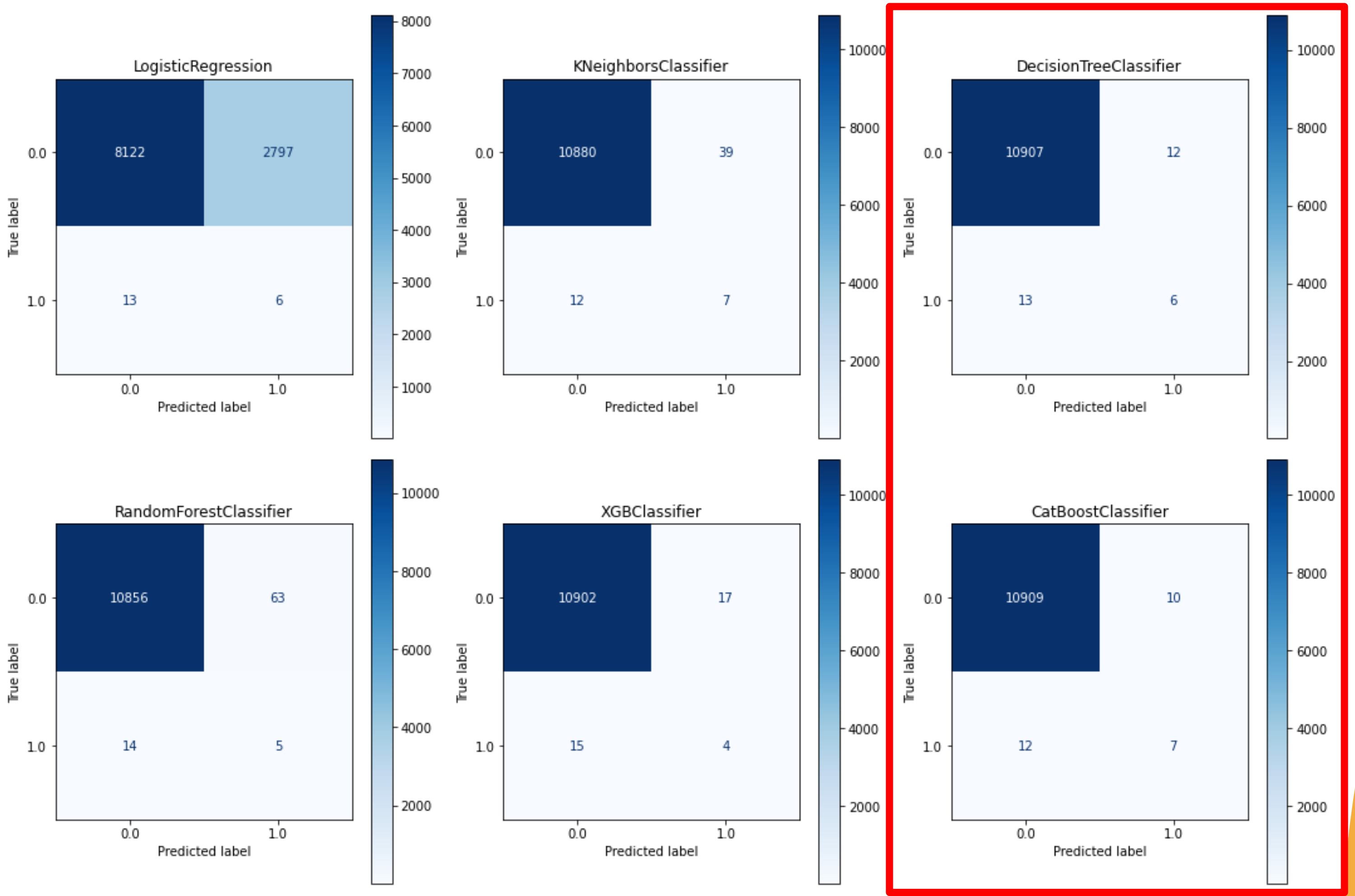
Non-Categorical / Numeric don’t have normal distribution

Standard Scaler

Model Comparison - Classifier

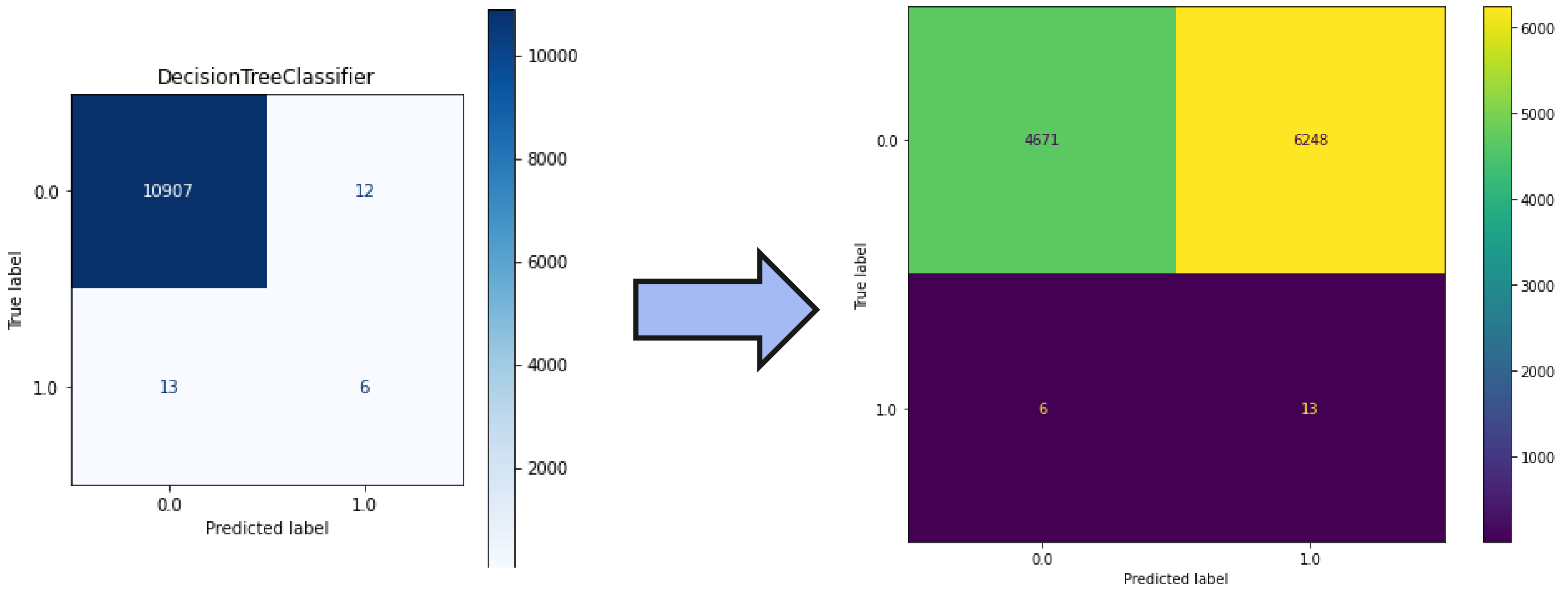
Choose model with highest **recall** with tolerable **accuracy**, we want false negative **smallest** as possible

Decision Tree & **Cat Boost** have good performance



Hyperparameter tuning

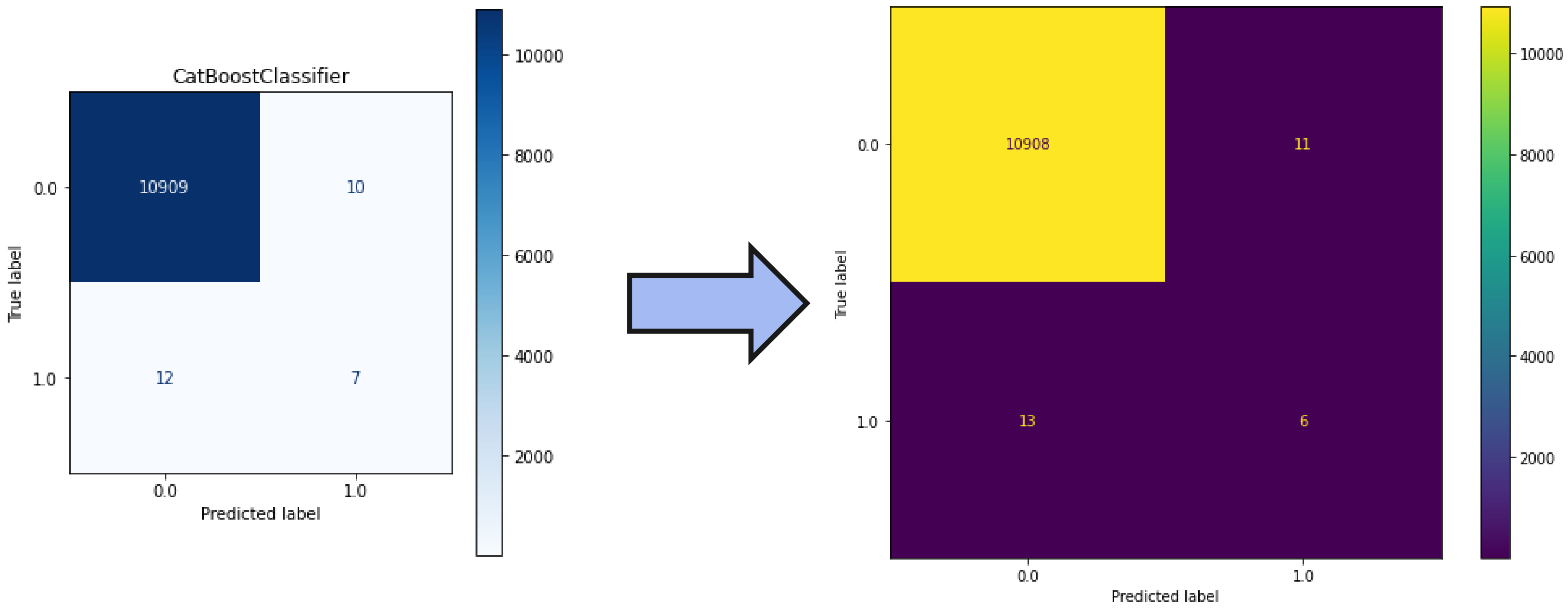
Decision Tree



Although the recall score increases, the false negative in the model is huge. More than half of the test data was incorrectly predicted, making the model unreliable.

Hyperparameter tuning

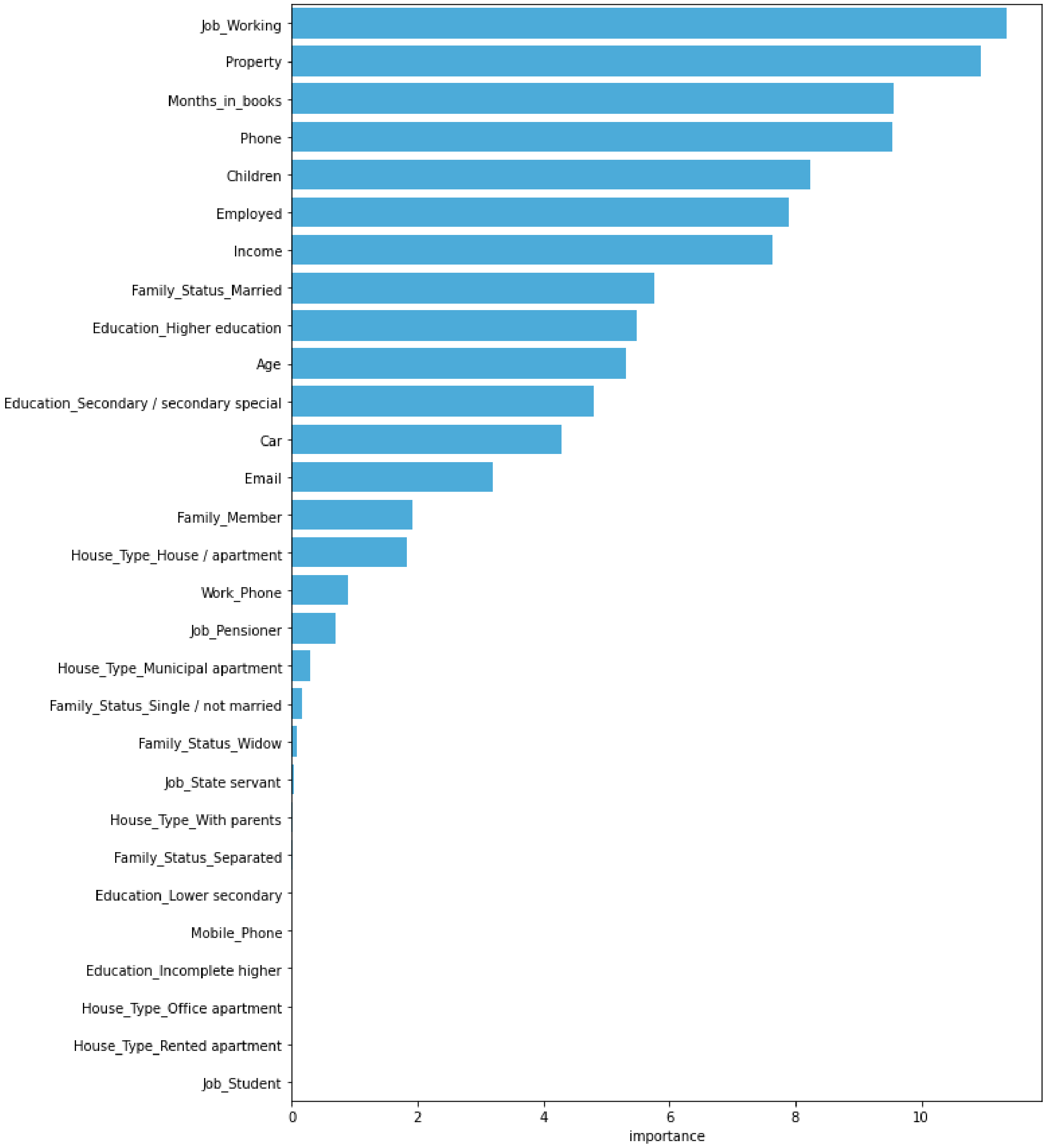
Cat Boost Classifier



The new parameter found using grid search hyperparameter tuning has a worse score than the default parameter. Therefore, we keep the last parameter as our model.

How Model Interpret the Data

Job_Working, Property, Months_in_books are features with the most contributions



Model’s accomplishments

Model successfully reduce loss from Defaulters

Rp 1.250.000 ▼ 26%

Assumption : Before using model all defaulter approval credit are approved, returned 75% value, 5% interest and Rp 1.000.000 loan money each person

Before

- Defaulters : 19
- Prevented : 0
- Loss : Rp 4.750.000

After (with Model)

- Defaulters : 19
- Prevented : 7
- Wrong predict : 10
- Loss : Rp 3.500.000

Conclusion

- The Cat Boost model **successfully reduce loss** from defaulters, with 26,3% percentage there are more room for improvement
- Job with working attributes are the most **important feature**. By analysis, it is found that good user able to pay their dept from working as worker with their stable income than other job.
- The Extreme **Imbalance** of data set make it hard to create model, example if model predict all value as good user model still have 99% accuracy but cannot predict even a single Bad User
- For **Profit** purpose User predicted as Bad User by the model still have a chance actually a Good User, we must still consider user economics value with their property or other valuable value for loss recovery in credit risk management.
- **Failure** in Hyper Parameter tuning because there is a chance wrong with chosen parameter, more trial with new parameter value for finding best parameter.

Thanks!

Do you have any questions?