**Predicting Credit Card Default using XGBoost**

Alexander Indrajaya L, Novan Dwi Atmaja, Siti Fatimah

Abstract

Keywords: credit scoring, xgboost, smote

**Executive Summary**

Salah satu fungsi dari perbankan adalah menyalurkan dana yang telah terkumpul untuk disalurkan kepada masyarakat salah satunya dalam bentuk pemberian pinjaman. Bank dituntut untuk mampu mengelola pinjaman nasabah dengan cermat. Sehingga proses menentukan nasabah yang dinilai layak untuk mendapatkan pinjaman atau tidak menjadi hal yang penting. Dalam mengukur resiko credit pada umumnya perbankan menggunkan metode credit scoring.

In order to evaluate the accuracy of this method, various prediction techniques have been formulated and introduced. Several measurement methods such as expert systems, econometric models, artificial intelligence (AI) techniques and hybrid form have different techniques.

Pada analisa ini, kami akan melakukan pemodelan credit card default berdasarkan data yang sudah di provide oleh FinHack 2018. Akan akan menggunakan model XGboost sebagai model dasar. Kami juga akan menyelesaikan masalah unbalanced data dengan menggunakan algoritma SMOTE. Selanjutnya kami akan mengukur performa dan stabilitas model.

**The Dataset**

Data yang kami gunakan terdiri dari 15.493 data training dan 2.214 data test. Data tersebut berisi terkait history pemakaian kartu kredit beserta status apakah kredit tersebut lancer atau tidak (namun pada data test data status kredit tersebut tidak tersedia). Detail terkait data yang digunakan dapat dilihat pada dashboard finhack.

**Exploratory Data Analysis**

Data training terdiri dari 1.359 (8.77%) credit macet dan 14.134 (91.23%) credit lancar.

table(data\_train$flag\_kredit\_macet)

0 1

14134 1359

Berikut ini adalah summary dari data yang akan digunakan

summary(data\_train)

jumlah\_kartu outstanding

Min. : 1.000 Min. : 0

1st Qu.: 2.000 1st Qu.: 2000551

Median : 2.000 Median : 4726943

Mean : 2.507 Mean : 11598331

3rd Qu.: 3.000 3rd Qu.: 10648482

Max. :16.000 Max. :798058574

limit\_kredit tagihan

Min. : 3000000 Min. : 20043

1st Qu.: 5000000 1st Qu.: 818250

Median : 9000000 Median : 3145857

Mean : 20798328 Mean : 8078663

3rd Qu.: 22000000 3rd Qu.: 7404991

Max. :1000000000 Max. :628000000

total\_pemakaian\_tunai total\_pemakaian\_retail

Min. : 0 Min. :-15667200

1st Qu.: 0 1st Qu.: 0

Median : 0 Median : 201042

Mean : 74533 Mean : 2025857

3rd Qu.: 0 3rd Qu.: 1398000

Max. :28840572 Max. :285000000

sisa\_tagihan\_tidak\_terbayar kode\_cabang

Min. : 0 A :8705

1st Qu.: 0 F :1697

Median : 2750923 B :1351

Mean : 8140875 I : 991

3rd Qu.: 7143198 E : 592

Max. :444000000 G : 566

(Other):1591

rasio\_pembayaran persentasi\_overlimit

Min. :-18138.00 Min. : 0.000

1st Qu.: 0.00 1st Qu.: 0.000

Median : 27.00 Median : 0.000

Mean : 64.65 Mean : 3.344

3rd Qu.: 100.00 3rd Qu.: 1.060

Max. : 68983.00 Max. :190.000

rasio\_pembayaran\_3bulan rasio\_pembayaran\_6bulan

Min. :-57792.34 Min. :-77056.00

1st Qu.: 20.20 1st Qu.: 21.60

Median : 50.00 Median : 65.60

Mean : 50.53 Mean : 81.67

3rd Qu.: 91.37 3rd Qu.: 100.00

Max. : 75575.00 Max. : 54899.00

skor\_delikuensi flag\_kredit\_macet

0:15451 Min. :0.00000

1: 29 1st Qu.:0.00000

2: 4 Median :0.00000

3: 2 Mean :0.08772

4: 4 3rd Qu.:0.00000

5: 3 Max. :1.00000

jumlah\_tahun\_sejak\_pembukaan\_kredit

Min. : 0.750

1st Qu.: 2.920

Median : 5.667

Mean : 6.634

3rd Qu.: 9.330

Max. :34.417

total\_pemakaian sisa\_tagihan\_per\_jumlah\_kartu

Min. :-15667200 Min. : 0

1st Qu.: 0 1st Qu.: 0

Median : 248000 Median : 1209318

Mean : 2100421 Mean : 2968060

3rd Qu.: 1512235 3rd Qu.: 3088477

Max. :314000000 Max. :148000000

sisa\_tagihan\_per\_limit total\_pemakaian\_per\_limit

Min. :0.0000 Min. :-0.6320

1st Qu.:0.0000 1st Qu.: 0.0000

Median :0.3350 Median : 0.0221

Mean :0.4682 Mean : 0.1094

3rd Qu.:0.9300 3rd Qu.: 0.1280

Max. :7.2400 Max. : 4.6000

pemakaian\_3bln\_per\_limit pemakaian\_6bln\_per\_limit

Min. :-0.1260 Min. :-0.5690

1st Qu.: 0.0352 1st Qu.: 0.0329

Median : 0.1080 Median : 0.1170

Mean : 0.1674 Mean : 0.2026

3rd Qu.: 0.2390 3rd Qu.: 0.2850

Max. : 3.4900 Max. : 8.1100

utilisasi\_3bulan utilisasi\_6bulan

Min. :0.000288 Min. :0.0000

1st Qu.:0.189000 1st Qu.:0.1550

Median :0.575000 Median :0.4960

Mean :0.571102 Mean :0.5337

3rd Qu.:0.917000 3rd Qu.:0.8630

Max. :8.125671 Max. :9.7300

Dari summary di atas diperoleh bahwa tidak ada data yang missing sehingga pada analisa ini tidak dilakukan proses data cleansing.

**Key Finding**

Pertama kami akan mengukur predictive power of characteristic dengan menggunakan information value. Berdasarkan Naeem siqqidi , one rule of thumb regarding IV is:

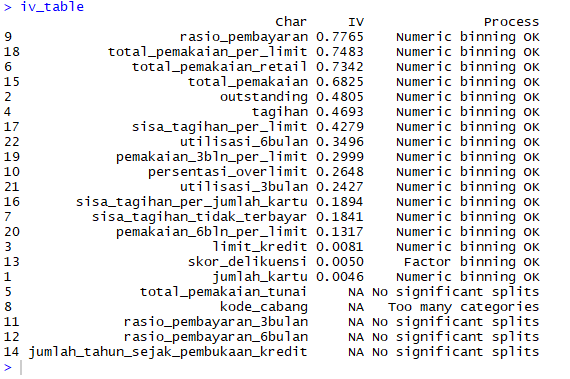
• Less than 0.02:unpredictive

• 0.02 to 0.1:weak

• 0.1 to 0.3:medium

• 0.3 +:strong

Kami mengukur IV menggunakan library(smbinning). Algoritma ini akan mengkategorikan data numerik kedalam bins tertentu berdasarkan Conditional Inferences Tree.



Berdasarakan nilai IV, diperoleh bahwa terdapat 8 feature: rasio\_pembayaran, total\_pemakaian\_per\_limit, total\_pemakaian\_retail, total\_pemakaian, outstanding, tagihan, sisa tagihan per\_limit, dan utilisasi\_6bulan merupakan predictor yang kuat untuk mengukur kredit macet.

Dari table di atas diperoleh bahwa hanya terdapat 4 features yang signifikan tidak mempengaruhi status kredit nasabah. Namun dengan mempertimbangkan jumlah feature yang tidak terlalu banyak dan jumlah yang signifikan hanya sedikit, maka pada proses pemodelan akan menggunakan semua features.

**Predicting**

Data yang tersedia menunjukan adanya kasus imbalanced data,

However, most of the existing state-of-the-art classification approaches are well developed by assuming the underlying training set is evenly distributed. Thus, they are faced with a severe bias problem when the training set is a highly imbalanced distribution (i.e., the data comprises two classes, the minority class C and the majority class C). The resulting decision boundary is severely biased to the minority class, and thus leads to a poor performance according to the receiver operator characteristic (ROC) curve analysis. The synthetic minority oversampling technique (SMOTE) is an important approach by oversampling the positive class or the minority class.

Pada tahap ini, kami akan memprediksi credit macet dengan menggunakan XGBoost. Kami akan membandingkan pengaruh smote pada data training terhadap performa model. Berikut ini adalah tahapan yang kami lakukan:

1. Create new data training using SMOTE algorithm
2. convert categorical factor into one-hot encoding
3. construct XGBoost object dengan xgb.DMatrix
4. Tentukan parameter (kami memilih iterasi sebanyak 100 kali)
5. Construct XGBoost model
6. Using model to predict credit default
7. Model evaluation

Berikut ini adalah data hasil SMOTE

> table(data\_smote$flag\_kredit\_macet)

0 1

8154 4077

> 100\*prop.table(table(data\_smote$flag\_kredit\_macet))

0 1

66.66667 33.33333

> summary(data\_smote[,c(1:23)])

jumlah\_kartu outstanding limit\_kredit

Min. : 1.000 Min. : 0 Min. : 3000000

1st Qu.: 2.000 1st Qu.: 1680140 1st Qu.: 5000000

Median : 2.000 Median : 4629947 Median : 10000000

Mean : 2.501 Mean : 12166773 Mean : 21238642

3rd Qu.: 3.000 3rd Qu.: 11142612 3rd Qu.: 25000000

Max. :16.000 Max. :798058574 Max. :1000000000

tagihan total\_pemakaian\_tunai

Min. : 20043 Min. : 0

1st Qu.: 703806 1st Qu.: 0

Median : 3223929 Median : 0

Mean : 8775385 Mean : 71512

3rd Qu.: 7933172 3rd Qu.: 0

Max. :628000000 Max. :28840572

total\_pemakaian\_retail sisa\_tagihan\_tidak\_terbayar

Min. :-15667200 Min. : 0

1st Qu.: 0 1st Qu.: 1342

Median : 41401 Median : 2952105

Mean : 1512171 Mean : 9268437

3rd Qu.: 881808 3rd Qu.: 7894910

Max. :285000000 Max. :444000000

kode\_cabang rasio\_pembayaran persentasi\_overlimit

A :6751 Min. :-2019.00 Min. : 0.000

F :1346 1st Qu.: 0.00 1st Qu.: 0.000

B :1100 Median : 13.78 Median : 0.000

I : 844 Mean : 51.71 Mean : 4.065

E : 514 3rd Qu.: 100.00 3rd Qu.: 3.892

G : 391 Max. :35711.00 Max. :189.000

(Other):1285

rasio\_pembayaran\_3bulan rasio\_pembayaran\_6bulan

Min. :-45469.00 Min. :-77056.00

1st Qu.: 14.20 1st Qu.: 17.10

Median : 37.50 Median : 50.00

Mean : 35.42 Mean : 60.24

3rd Qu.: 75.36 3rd Qu.: 100.00

Max. : 38828.00 Max. : 54899.00

skor\_delikuensi flag\_kredit\_macet

0:11588 Min. :0.0000

1: 526 1st Qu.:0.0000

2: 108 Median :0.0000

3: 1 Mean :0.3333

4: 6 3rd Qu.:1.0000

5: 2 Max. :1.0000

jumlah\_tahun\_sejak\_pembukaan\_kredit total\_pemakaian

Min. : 0.750 Min. :-15667200

1st Qu.: 2.920 1st Qu.: 0

Median : 5.580 Median : 56345

Mean : 6.563 Mean : 1583695

3rd Qu.: 9.330 3rd Qu.: 1011052

Max. :34.400 Max. :314000000

sisa\_tagihan\_per\_jumlah\_kartu sisa\_tagihan\_per\_limit

Min. : 0 Min. :0.000000

1st Qu.: 544 1st Qu.:0.000051

Median : 1327157 Median :0.411000

Mean : 3321153 Mean :0.496165

3rd Qu.: 3471681 3rd Qu.:0.966000

Max. :148000000 Max. :5.530000

total\_pemakaian\_per\_limit pemakaian\_3bln\_per\_limit

Min. :-0.273000 Min. :-0.05160

1st Qu.: 0.000000 1st Qu.: 0.02102

Median : 0.004523 Median : 0.09010

Mean : 0.085656 Mean : 0.15236

3rd Qu.: 0.085850 3rd Qu.: 0.21600

Max. : 3.330000 Max. : 3.18000

pemakaian\_6bln\_per\_limit utilisasi\_3bulan utilisasi\_6bulan

Min. :-0.07820 Min. :0.000288 Min. :0.0000

1st Qu.: 0.02508 1st Qu.:0.163000 1st Qu.:0.1370

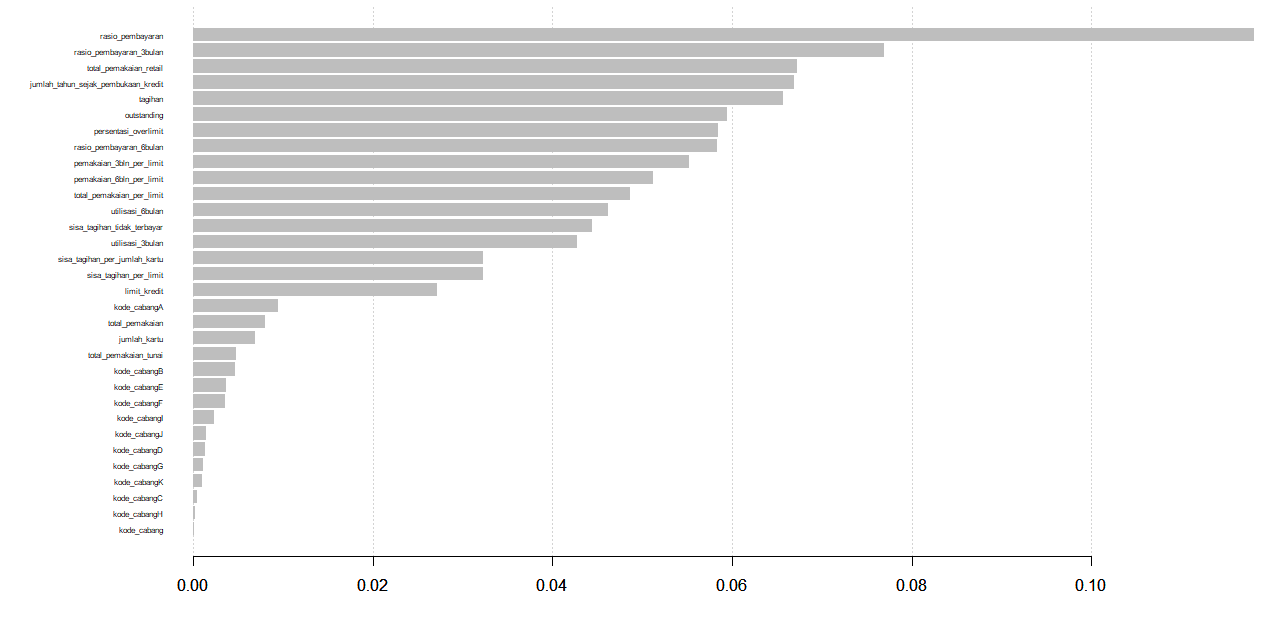
Median : 0.10800 Median :0.591000 Median :0.5060

Mean : 0.19873 Mean :0.577869 Mean :0.5405

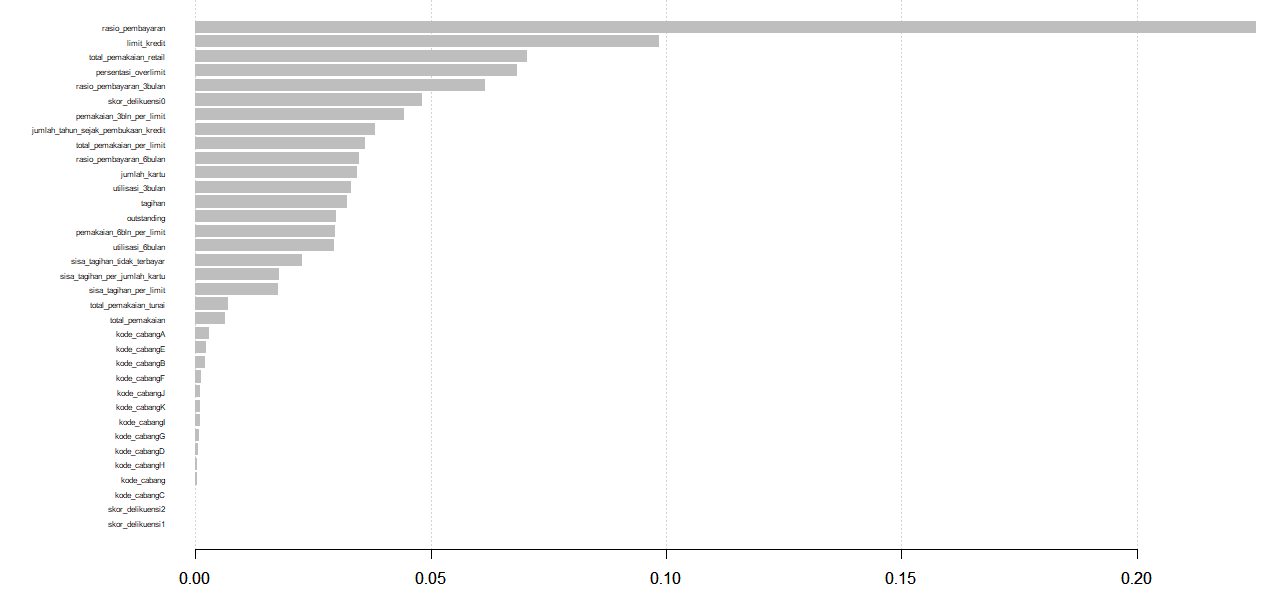
3rd Qu.: 0.27400 3rd Qu.:0.944157 3rd Qu.:0.8815

Max. : 8.11000 Max. :8.125671 Max. :9.7300

Berikut important variable yang diperoleh dari model

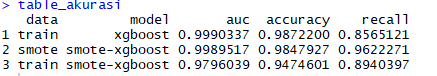


Berikut importance variable dengan menggunakan smote



Diperoleh bahwa rasio\_pembayaran (Rasio perbandingan jumlah yang dibayar dengan tagihan pada bulan terakhir) merupakan features yang plg penting, hal ini konsisten dengan IV features ini yg paling besar.

Berikut ini perbandingan yg diperoleh



Proses smote sedikit mengurangi tingkat akurasi, namun dapat meningkatkan recall. Dengan mempertimbangkan nilai akurasi yang masih tetap besar. Maka model dengan smote akan dipilih sebagai model prediksi.

Kami juga mengukur stabilitas distribusi model terhadap data testing. Berikut ini adalah langkah-langkahnya

1. Hitung PD data training berdasarkan model xgboost dan smote-xgboost
2. Untuk setiap model, group data PD ke dalam 5% percentile. Simpan batas masing-masing klas.
3. Hitung PD data test berdasarkan model xgboost dan smote-xgboost
4. Klasifikasikan nilai PD berdasarkan batas yang diperoleh pada langkah 2
5. Hitung nilai PSI (population stability index)

Berikut ini hasil perbandingan PSI



Dari table di atas diperoleh bahwa model smote-xgboost lebih stabil.

Referensi

<https://pdfs.semanticscholar.org/afb7/08f63a504513eb3cedcf76a1ae935da6509e.pdf>

<http://sci2s.ugr.es/keel/pdf/specific/congreso/04129201.pdf>