Pacmann

Credit Scoring Project

Kelas Credit Scoring
Juniarto Kurniawan - Batch 9

Outline

- Introduction/Background
- Workflow
- Conclusion
- Recommendation
- References

Introduction

Business Problem

- Usaha perbankan meliputi tiga bagian yaitu melakukan penghimpunan dana dari nasabah dalam bentuk simpanan, kedua yaitu menyalurkan dana yang dihimpun kepada debitur dalam bentuk pinjaman, dan yang ketiga adalah memberikan jasa bank lainnya.
- Salah satu risiko yang sering dijumpai adalah risiko kredit yang merupakan risiko kegagalan debitur dalam memenuhi kewajibannya sehingga dapat menggerus profit perbankan.

- Penerapan manajemen risiko kredit yang baik dapat meminimalisir risiko pada level tertentu sehingga bisnis menjadi lebih optimal. Pada bisnis konsumer yaitu bisnis kartu kredit manajemen risiko dapat dilakukan dengan menggunakan metode credit scoring yaitu sebuah penilaian yang dijadikan dasar pertimbangan bagi pemberi pinjaman sebelum menyalurkan dana pinjaman ke peminjam.
- Sehingga Objective dari analisis ini adalah dapat memberikan model yang menghasilkan skor optimal yang dapat menurunkan tingkat kredit bermasalah.

Workflow

Data Gathering

- Data Understanding

Sample Splitting

- Split Train Test

EDA

- Univariate
- Multivariate

Initial Character Analysis

- Binning
- WOE and IV

Design Scorecards

- Logical Trends and Business Consideration
- Independence Test
- Preprocessing Data
- Initial logistic Regression
- Best Selection

Model Evaluation

- Predict on Test and Train data
- Model Adjustment

Scorecards Development

- Create Scorecard
- Predict Credit Score
- Setting Cutoffs

Data Understanding

Data terdiri dari bebrapa variabel yang meliputi karakteristik dan demografi dari nasabah yang beserta status kreditnya. Data terdiri drari.

- **Variabel Response**
  - Status credit : Accepted dan Rejected

- **Variabel Predictor Categorical**
  - No of Dependents : jumlah tanggungan yang dimiliki
  - Education : Graduate atau Bukan Graduate
  - Self emplyed : Yes or No

- **Variabel Predictor Numerical**
  - Income Annum : pendapatan yang dieproleh selama satu tahun
  - Loan amount : Jumlah pinjaman yang diajukan
  - Loan Term : lama atau jangka waktu pinjaman
  - residential assets value : jumlah asset tetap (harta tidak bergerak/perumahan)
  - commercial assets value : jumlah aset produktif
  - luxury assets value : jumlah asset tersier/mewah
  - bank asset value : jumlah aset kas yang terdapat pada bank

Sample Splitting

Selanjutnya, split data menjadi training dan testing untuk masing-masing variabel prediktor (X) and response (y).

- Set stratify = y for splitting the sample with stratify, based on the proportion of response y.
- Set test_size = 0.2 for holding 20% of the sample as a testing set.
- Set random_state = 42 for reproducibility.

```
X train shape : (3415, 10)
y train shape : (3415,)
X test shape  : (854, 10)
y test shape  : (854,)
```

```
# Check Proportion for train predictor
y_train.value_counts(normalize = True)
```

```
1    0.622255
0    0.377745
Name: loan_status, dtype: float64
```

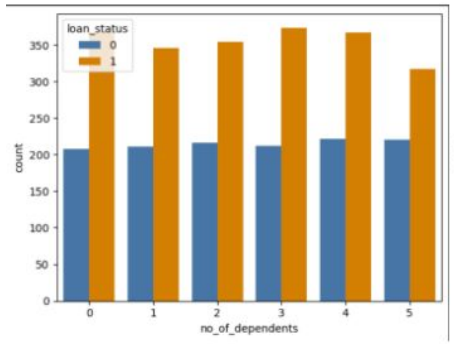
```
# Check Proportion for test predictor
y_test.value_counts(normalize = True)
```

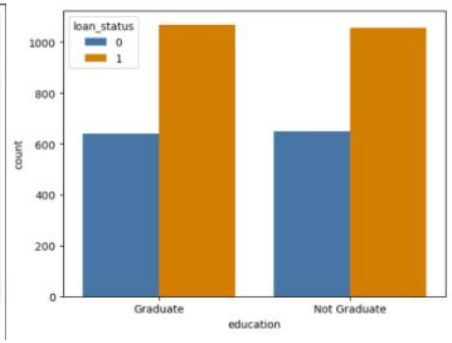
```
1    0.62178
0    0.37822
Name: loan_status, dtype: float64
```

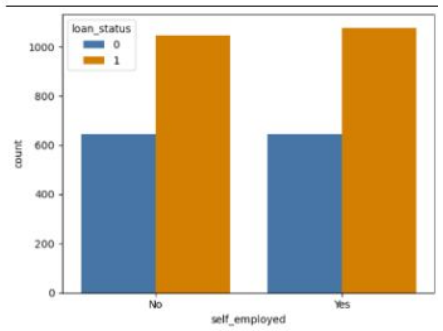
- Masing-masing data berjumlah 3145 (train data) dan 854 (test data)
- Proporsi rejected dan accepted untuk masing-masing train dan test data di variabel response relatif sama.

Exploratory Data Analysis (EDA)

Univariate Analysis

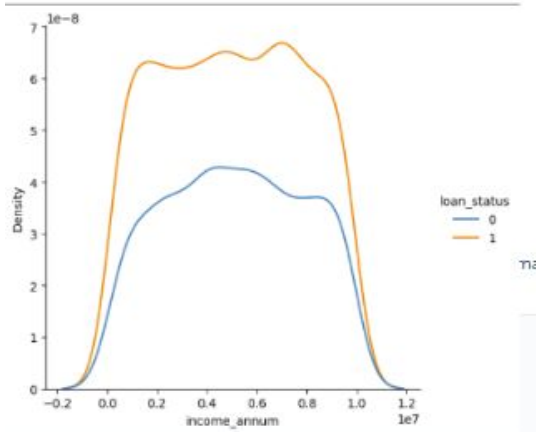


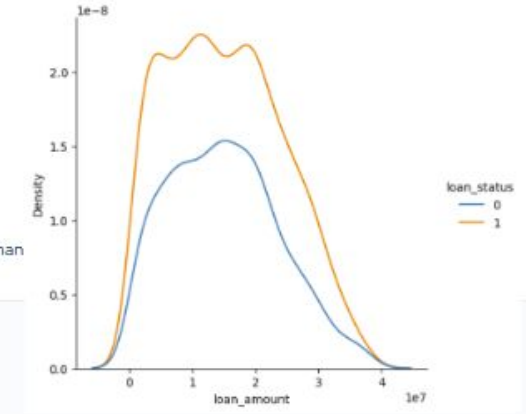




Berdasarkan grafik dari tiga prediktor yang dipilih yaitu (no of dependents, education, dan self employed)

- memiliki karakter yang sama diantara masing masing kategorinya. Contoh untuk education, proporsi accepted dibanding rejected untuk graduate memiliki besar perbandingan yang sama dengan not graduate.
- hal ini berlaku untuk masing-masing kategori pada no of dependents dan self employed.



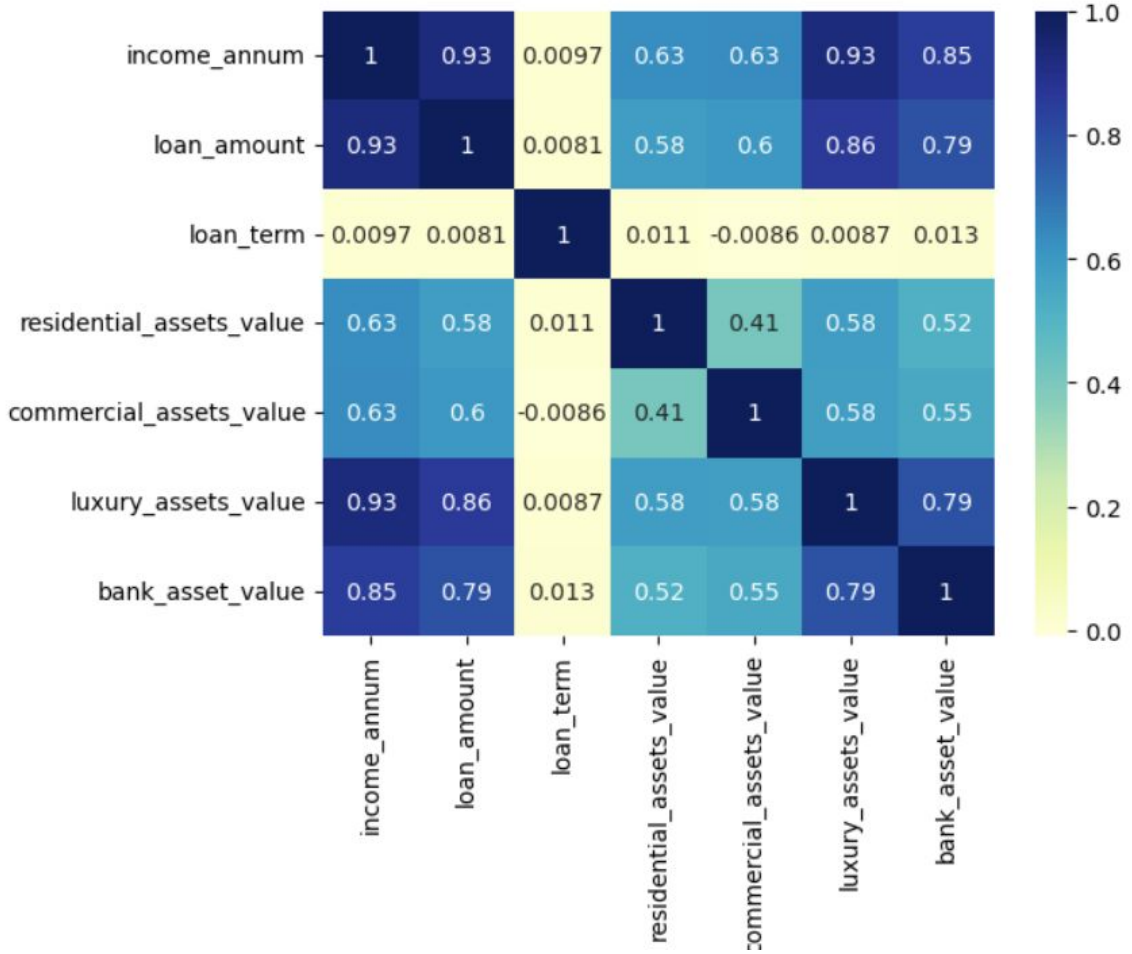


Grafik distribusi dari dua prediktor yang dipilih yaitu (income annum dan loan amount) juga :

- memiliki karakter yang sama diantara masing masing kategorinya. Contoh untuk income annum, pola distribusi untuk status rejected memiliki kemiripan atau identik sama dengan accepted.
- Begitu pula untuk kategori pada loan amount.

Exploratory Data Analysis (EDA)

Multivariate Analysis



Terdapat beberapa korelasi yang cukup kuat antar variabel yaitu diantaranya:

- Income Annum mempunyai korelasi yang cukup kuat dengan beberapa variabel lain.
- Loan amount mempunyai korelasi yang cukup kuat dengan beberapa variabel lain.
- Residential asset value mempunyai korelasi yang cukup kuat dengan beberapa variabel lain.
- Commercial asset value mempunyai korelasi yang cukup kuat dengan beberapa variabel lain Luxury asset dan bank assets.
- Luxury asset value mempunyai korelasi yang cukup kuat dengan bank assets.
- Hal yang unik adalah variabel loan term memiliki hubungan yang rendah dengan 9 variabel lainnya.

Initial Character Analysis

Binning

Binning atau discretization atau kategorisasi diberikan untuk mempermudah dalam melakukan interpretasi variabel kedepannya. Binning dilakukan pada numerical kategori dengan mengelompokkannya ke dalam beberapa grup yang ditentukan.

income_annum_bin	loan_amount_bin	loan_term_bin	residential_assets_value_bin	commercial_assets_value_bin	luxury_assets_value_bin	bank_asset_value_bin
(2700000.0, 5100000.0]	(7800000.0, 14600000.0]	(1.999, 6.0]	(5600000.0, 11200000.0]	(-0.001, 1300000.0]	(7500000.0, 14600000.0]	(2400000.0, 4500000.0]
(5100000.0, 7400000.0]	(14600000.0, 21300000.0]	(16.0, 20.0]	(5600000.0, 11200000.0]	(7600000.0, 19400000.0]	(14600000.0, 21600000.0]	(2400000.0, 4500000.0]
(7400000.0, 9900000.0]	(21300000.0, 38800000.0]	(16.0, 20.0]	(11200000.0, 29100000.0]	(7600000.0, 19400000.0]	(21600000.0, 39200000.0]	(7000000.0, 14700000.0]
(2700000.0, 5100000.0]	(7800000.0, 14600000.0]	(1.999, 6.0]	(11200000.0, 29100000.0]	(3700000.0, 7600000.0]	(14600000.0, 21600000.0]	(4500000.0, 7000000.0]

Initial Character Analysis

WOE dan IV

- WOE digunakan untukk memberikan bobot yang prorposional pada setiap kategrori di semua variabel prediktor
- IV adalah Nilai informasi yang dapat menjelaskan variabel tersebut yang sudah di kategorikan menjadi (unpredictive, weak, medium, strong)

	Characteristic	Attribute	WOE
0	income_annum_bin	(199999.999, 2700000.0]	-0.118668
1	income_annum_bin	(2700000.0, 5100000.0]	0.084854
2	income_annum_bin	(5100000.0, 7400000.0]	0.007535
3	income_annum_bin	(7400000.0, 9900000.0]	0.028359
0	loan_amount_bin	(299999.999, 7800000.0]	-0.043268
1	loan_amount_bin	(7800000.0, 14600000.0]	0.056442
2	loan_amount_bin	(14600000.0, 21300000.0]	0.137854
3	loan_amount_bin	(21300000.0, 38800000.0]	-0.155484

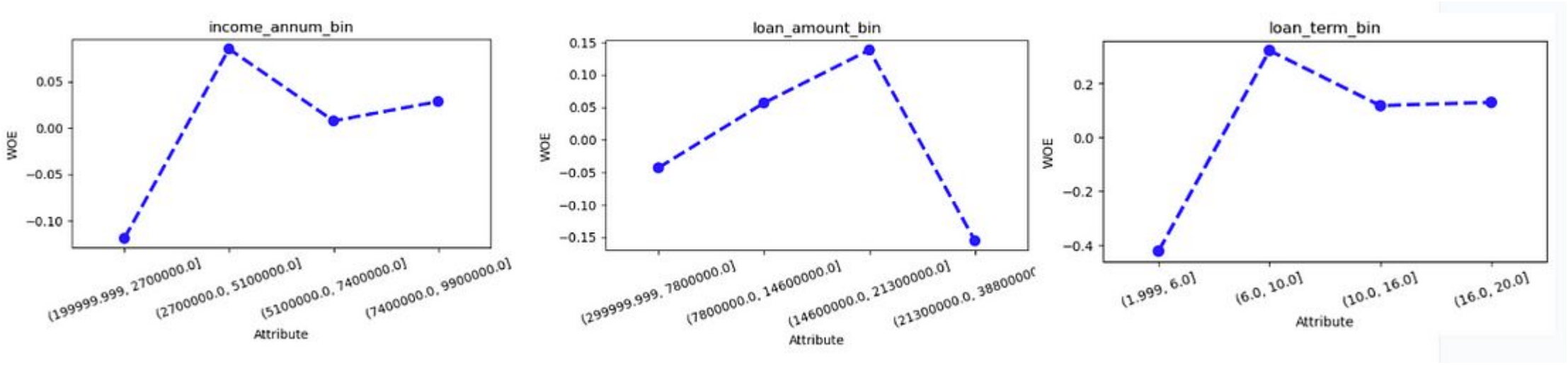
	Characteristic	Information Value	Strength
8	education	0.000152	Unpredictive
9	self_employed	0.000260	Unpredictive
3	residential_assets_value_bin	0.002138	Unpredictive
4	commercial_assets_value_bin	0.003740	Unpredictive
7	no_of_dependents	0.004661	Unpredictive
0	income_annum_bin	0.005621	Unpredictive
6	bank_asset_value_bin	0.006690	Unpredictive
5	luxury_assets_value_bin	0.007832	Unpredictive
1	loan_amount_bin	0.011949	Unpredictive
2	loan_term_bin	0.081267	Weak

Berdasarkan infermasi IV hanya loan term yang mempunyai nilai lebih memberikan value sisanya melum mencukupi unutk memberikan informasi.

Design Scorecards

Logical Trends and Business Consideration

Kita akan melihat logical trends dengan membuat plot WOE, apakah pembobotan telah diterapkan sesuai dengan kepentingan bisnis.



Secara umum berdasarkan WOE plot yang cukup bervariasi dapat diartikan sudah sesuai dengan business consideration.

Design Scorecards

Perform Independet Test

	Characteristic	Chi-stat	P-value	Conclusion
0	income_annum_bin	4.48	2.138383e-01	Independent
1	loan_amount_bin	9.56	2.274506e-02	Not Independent
2	loan_term_bin	63.19	1.222404e-13	Not Independent
3	residential_assets_value_bin	1.72	6.324813e-01	Independent
4	commercial_assets_value_bin	3.01	3.900086e-01	Independent
5	luxury_assets_value_bin	6.22	1.012851e-01	Independent
6	bank_asset_value_bin	5.39	1.452223e-01	Independent
7	no_of_dependents	3.77	5.832113e-01	Independent
8	education	0.10	7.537234e-01	Independent
9	self_employed	0.18	6.730536e-01	Independent

Hanya loan amount dan loan term yang memiliki hasil not independent sedangkan sisanya adalah independent.

Design Scorecards

Preprocessing Data

Preprocessing dilakukan pada train dan test data dengan langkah mereplace data train dan test menjadi data yang sudah terboboti (WOE)

	no_of_dependents	education	self_employed	income_annum	loan_amount	loan_term	residential_assets_value	commercial_assets_value
1877	-0.003555	-0.012320	0.016245	0.084854	0.056442	-0.423007	0.068047	0.020852
1729	-0.068540	-0.012320	-0.016034	0.007535	0.137854	0.129449	0.068047	-0.055867
164	0.005111	-0.012320	-0.016034	0.028359	-0.155484	0.129449	0.017200	-0.055867
2298	0.005111	0.012319	-0.016034	0.084854	0.056442	-0.423007	0.017200	0.090702
2461	0.138391	-0.012320	0.016245	0.007535	-0.155484	0.321754	0.068047	-0.055867
...	...	...	...	...	...	...	...	...

Design Scorecards

Performing Initial Logistics Regression

Yang dimaksud dengan initial ini adalah melakuan pemodelan dengan menggunakan regresi logistik dan memilih prediksi terbaik dari setiap kombinasi variabel prediktor.

	Predictors	Recall
0	[]	0.000000
1	[0]	0.521412
2	[0, 1]	0.585882
3	[0, 1, 6]	0.586353
4	[0, 1, 6, 7]	0.581176
5	[0, 1, 6, 7, 2]	0.563765
6	[0, 1, 6, 7, 2, 9]	0.554353
7	[0, 1, 6, 7, 2, 9, 3]	0.540706
8	[0, 1, 6, 7, 2, 9, 3, 8]	0.528941
9	[0, 1, 6, 7, 2, 9, 3, 8, 4]	0.520000
10	[0, 1, 6, 7, 2, 9, 3, 8, 4, 5]	0.518118

```
Best index           : 3
Best Recall          : 0.5863529411764705
Best predictors (idx) : [0, 1, 6]
Best predictors      :
['no_of_dependents', 'education', 'residential_assets_value']
```

Dengan menggunakan pemilihan secara forward maka diperoleh model dengan 3 prediktor terbaik yaitu (no_of_dependents, education, dan residental assets value)

Design Scorecards

Performing Best Predictor Logistics Regression

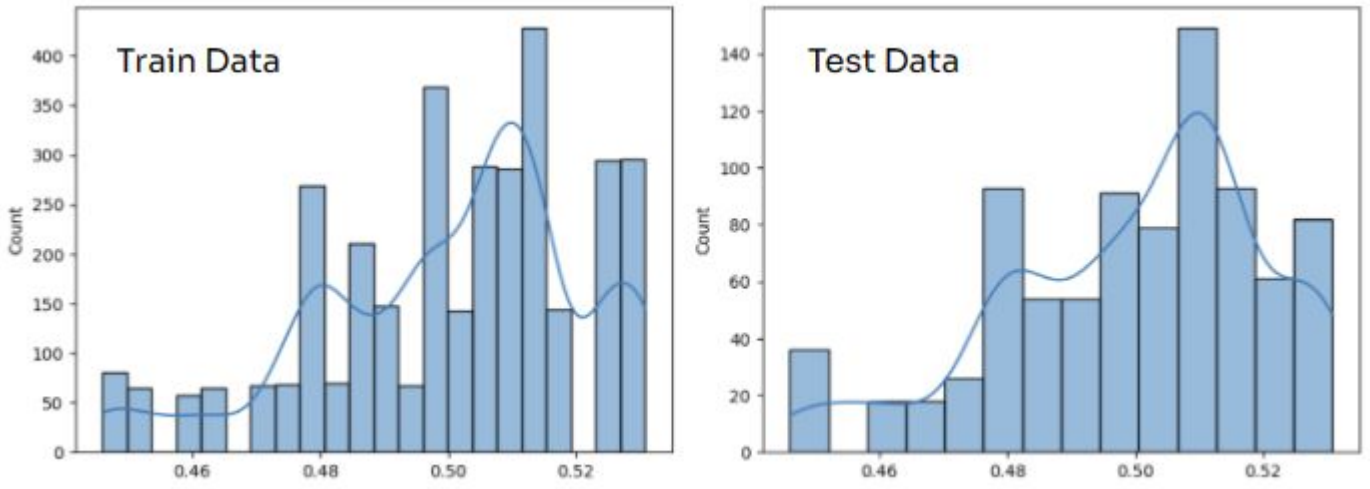
Selanjutnya kita gunakan ketiga prediktor terbaik ini untuk membuat model regresi logistik.

	Estimate
Intercept	-9.693952e-07
no_of_dependents	-9.943471e-01
education	-9.815531e-01
residential_assets_value	-9.897752e-01

Model Evaluation

Predict On Train and Test Data

Kita gunakan best predictor untuk melakukan prediksi pada train dan test data. Hasilnya sebagai berikut:



Dari model selection, kita mendapat model terbaik dengan 3 variabel prediktor yaitu (no_of_independents and education dan residential_assets_value).
Sensitivity/recall score dari best model adalah 0,56 untuk train data dan 0,55 untuk test data:
Tartinya model dapat memprediksi sekitar 51% rejected applicant.

Kita masih terdapat 44% dalam misklasifikasi yang artinya cukup besar.
Best model dapat dipertanyakan mengingat variabel yang diambil hanya 3 dari 10
kita dapat melakukan prediksi pada 10 model dan membanding apa terdapat perbedaan recall yang soignifikan.

Model Evaluation

Model Adjustment

Scorecard dengan sedikit karakteristik secara umum tidak dapat diterapkan dalam jangka waktu yang panjang:

- Mereka rentan terhadap perubahan kecil pada profil pelamar.
- Prediktor yang baik adalah prediktor yang bukan hanya melihat berdasarkan 3 dari 10 variabel yang tersedia.

Kita akan memasukkan seluruh karakteristik dalam final model.

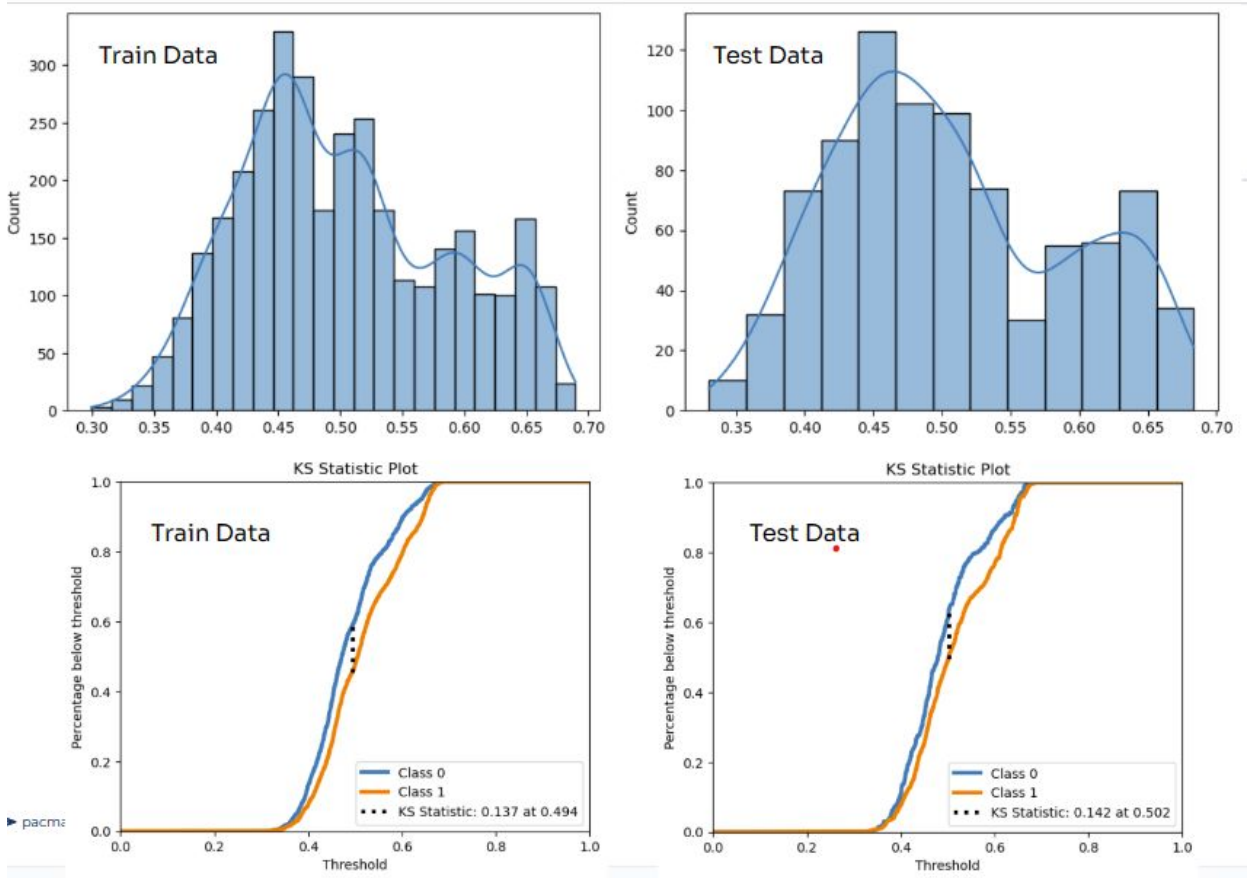
- Berdasarkan independent test, terdapat 8 variabel yang tidak independent. Namun salah satu asumsi regresi logistik menyatakan bahwa independensi variabel tidak mutlak disyaratkan
- Secara umum, best practice untuk final scorecards terdiri dari 8 hingga 15 variabel.

	Characteristic	Estimate
0	Intercept	-0.000105
1	no_of_dependents	-1.216861
2	education	-1.169949
3	residential_assets_value	-0.335929
4	commercial_assets_value	-0.829924
5	self_employed	-1.488469
6	bank_asset_value	-0.495299
7	income_annum	0.454696
8	luxury_assets_value	-0.844206
9	loan_amount	-0.900625
10	loan_term	-1.036899

Model Evaluation

Model Adjustment

- Hasil dari prediksi data untuk model dengan seluruh variabel adalah sebagai berikut:



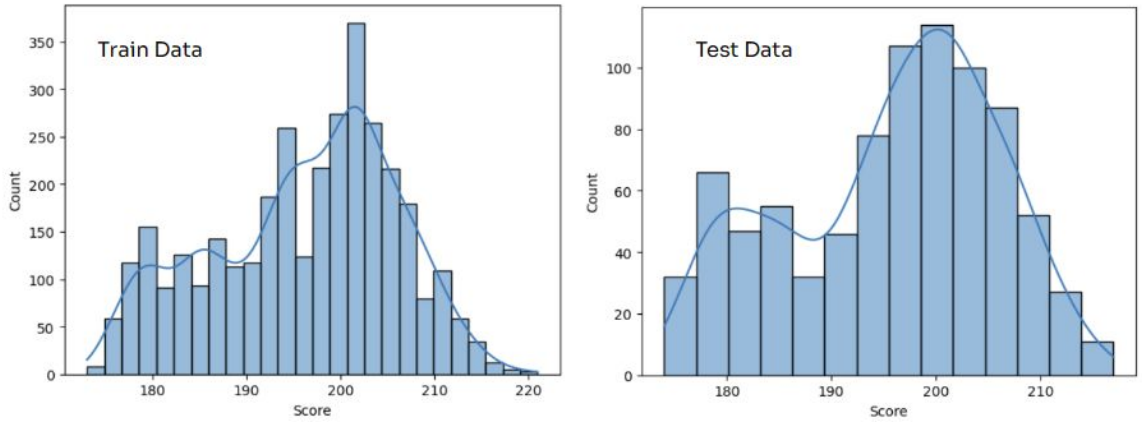
- Sensitivity/recall score dari adjusted model atau final model sedikit menurun yaitu is 0,52 on the train set and 0.51 on the test set.
- AUC dari adjusted model sedikit lebih tinggi ; 0.59 pada train set dan 0.58 pada the test set.
- KS Statistik plot baik dari data train maupun data test memiliki karakteristik yang sama yaitu selisihnya tidak cukup signifikan sehingga perbedaan data sulit diidentifikasi.

Scorecards Development

Predict Score

Langkah melakukan scorecard adalah dengan memberikan points pada setiap kategori dari proedictor yang telak dilakukan pembobotan (WOE)

	Characteristic	Attribute	WOE	Estimate	Points
0	income_annum_bin	(199999.999, 2700000.0]	-0.118668	0.454696	21
1	income_annum_bin	(2700000.0, 5100000.0]	0.084854	0.454696	19
2	income_annum_bin	(5100000.0, 7400000.0]	0.007535	0.454696	20
3	income_annum_bin	(7400000.0, 9900000.0]	0.028359	0.454696	19
4	loan_amount_bin	(299999.999, 7800000.0]	-0.043268	-0.900625	19
5	loan_amount_bin	(7800000.0, 14600000.0]	0.056442	-0.900625	21
6	loan_amount_bin	(14600000.0, 21300000.0]	0.137854	-0.900625	23
7	loan_amount_bin	(21300000.0, 38800000.0]	-0.155484	-0.900625	16

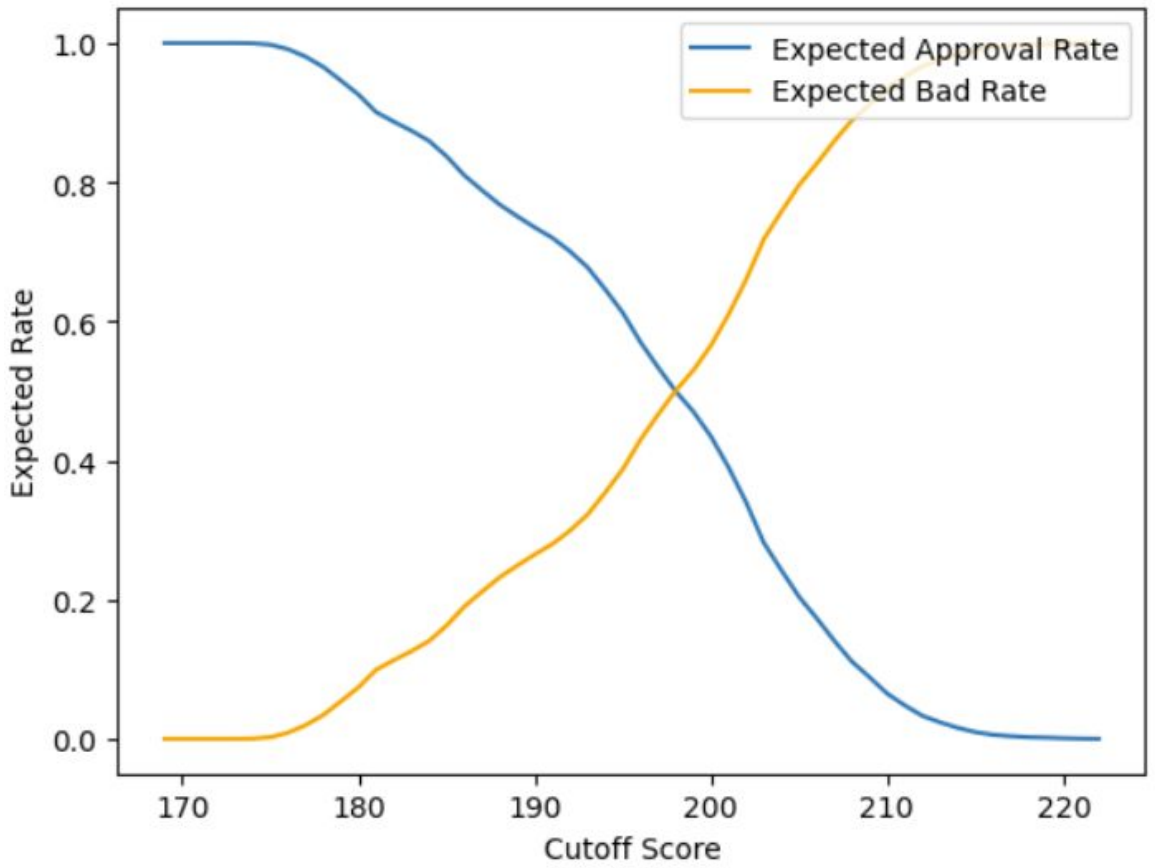


Berdasarkan distribusi total score diperoleh penjelasan bahwa baik distribusi pada data train maupun data test memiliki pola yang sama membentuk distribusi normal.

Scorecards Development

Setting Cutoffs

Cara melakukan setting cutoff adalah dengan membuat grafik perbadingan expected approval terhadap expected bad rate sebagai berikut:



Jika melihat pada grafik di atas, terdapat persimpangan/ titik potong. yang artinya telah terjadi perubahan dari approve menjadi rejected. Sehingga Cutoff yang optimal ada di sekitar 200.

Scorecards Development

Test in Real Case

Selanjutnya kita implementasikan, model di atas dengan cara melakukan input pada model sebagai berikut :

```
input = {
    'bank_asset_value_bin': 9000000,
    'commercial_assets_value_bin': 18000000,
    'education': 'Graduate',
    'loan_amount_bin': 25000000,
    'loan_term_bin': 20,
    'luxury_assets_value_bin': 7000000,
    'no_of_dependents': 2,
    'residential_assets_value_bin': 7500000,
    'self_employed': 'No',
}
```

Maka Score yang akan kita dapat adalah sebagai berikut:

```
input_score = predict_score(raw_data = input_table,
                            cutoff_score = 200)
```

```
Credit Score :  172
Recommendation : REJECT
```

Conclusion

Conclusion

Setelah melakukan serangkaian analisa kita memperoleh beberapa kesimpulan sebagai berikut :

- Kita melakukan prediksi menggunakan metode regresi logistik dengan 10 variabel yang terdiri dari 3 variabel prediktor kategorik dan 7 variabel prediktor numerik

- Berdasarkan pemilihan model terbaik dengan menggunakan metode forward diperoleh model terbaik yaitu model dengan 3 variabel prediktor yaitu no_of_dependents, education, dan residential assets value.

- Akan tetapi model dengan 3 variabel berdasarkan best practice dirasa terlalu sedikit dan sensitif terhadap perubahan data. Oleh karena itu digunakan adujstment menggunakan ke sepuluh variabel predictor dengan hasil recall yang tidak terlalu signifikan yaitu berkisar 50%. Selanjutnya model dengan seluruh prediktor akan digunakan dalam membangun scorecards.

- Hasil dari scorecards yang telah dibuat, diperoleh cutoffs yang terbaik atau direkomendasikan yaitu dengan score di sekitar 200.

Recommendation

Analisa atau pemodelan ini memiliki banyak kelemahan diantaranya:

- Jumlah loan approval dengan rejected tidak terpaut cukup jauh seperti yang common terjadi pada scoring rejected rate hingga 35%.

- Berdasarkan Information values hampir seluruh prediktor memberikan hasil unpredictive kecuali loan term dengan hasil weak.

- Berdasarkan independence test hampir seluruh predictor memberikan hasil independent, kecuali loan_amount dan loan_term. Berdasarkan referensi memang regresi logistik tidak mensyaratkan variabel yang berhubungan tapi alangkah lebih baik apabila seluruh variabel memiliki hasil yang tidak independen.

- Hal yang dapat dilakukan perbaikan pada data adalah dengan melakukan penambahan jumlah data sehingga hasil lebih representatif. Sedangkan dalam proses analisa dapat dilakukan perbaikan pada proses binning dan pembobotan dilakukan dengan cara expert adjustment (manual berdasarkan rekomendasi para ahli).

Reference

- 16481–53839–1-ED.docx (live.com)
- Regresi Logistik.docx (live.com)

Thank You