

4 Tech - Stage 1

Dokumen Laporan Final Project

(dipresentasikan setiap sesi mentoring)



Dataset



RangeIndex: 21000 entries, 0 to 20999

#	Columns (total 25 columns):	Non-Null Count	Dtype
0	ID	21000 non-null	
1	LIMIT_BAL	21000 non-null	int64
2	SEX	21000 non-null	int64
3	EDUCATION	21000 non-null	int64
4	MARRIAGE	21000 non-null	int64
5	AGE	21000 non-null	int64
6	PAY_0	21000 non-null	int64
7	PAY_2	21000 non-null	int64
8	PAY_3	21000 non-null	int64
9	PAY_4	21000 non-null	int64
10	PAY_5	21000 non-null	int64
11	PAY_6	21000 non-null	int64
12	BILL_AMT1	21000 non-null	int64
13	BILL_AMT2	21000 non-null	int64
14	BILL_AMT3	21000 non-null	int64
15	BILL_AMT4	21000 non-null	int64
	BILL_AMT5	21000 non-null	
17	BILL_AMT6	21000 non-null	
18	PAY_AMT1	21000 non-null	int64
19	PAY_AMT2	21000 non-null	
20	PAY_AMT3	21000 non-null	
21	PAY_AMT4	21000 non-null	
22	PAY_AMT5	21000 non-null	int64
	PAY_AMT6	21000 non-null	
24	default_payment_next_month	21000 non-null	int64
	es: int64(25) ry usage: 4.0 MB		

Hasil pengamatan:

- Dataset payment default prediction terdiri dari data 21.000 nasabah bank.
- Dataset ini memiliki data dengan jenis numerikal dan juga kategorikal dengan tipe data int.

Kategorikal:

- SEX
- EDUCATION
- MARRIAGE
- PAY_0
- PAY_2
- PAY_3
- PAY_4
- PAY_5
- PAY_6
- default_payment_next_ month

Variabel Target:

default_payment_next_month

Numerikal:

- LIMIT_BAL
- AGE
- BILL_AMT1
- BILL_AMT2
- BILL_AMT3
- BILL_AMT4
- BILL_AMT5
- BILL_AMT6
- PAY_AMT1
- PAY_AMT2
- PAY_AMT3
- PAY_AMT4
- PAY_AMT5
- PAY_AMT6





RangeIndex: 21000 entries, 0 to 20999

Data columns (total 25 columns): Non-Null Count Dtype Column ID 21000 non-null int64 LIMIT BAL 21000 non-null int64 SEX 21000 non-null int64 **EDUCATION** 21000 non-null int64 int64 MARRIAGE 21000 non-null AGE 21000 non-null int64 PAY 0 21000 non-null int64 PAY 2 21000 non-null int64 PAY 3 21000 non-null int64 PAY 4 21000 non-null int64 PAY 5 21000 non-null int64 PAY 6 21000 non-null int64 BILL AMT1 21000 non-null int64 BILL AMT2 21000 non-null int64 14 BILL AMT3 21000 non-null int64 BILL AMT4 21000 non-null int64 BILL AMT5 21000 non-null int64 BILL AMT6 21000 non-null int64 PAY AMT1 21000 non-null int64 PAY AMT2 21000 non-null int64 PAY AMT3 21000 non-null int64 21 PAY AMT4 21000 non-null int64 22 PAY AMT5 21000 non-null int64 23 PAY AMT6 int64 21000 non-null 24 default_payment_next_month 21000 non-null int64 dtypes: int64(25) memory usage: 4.0 MB

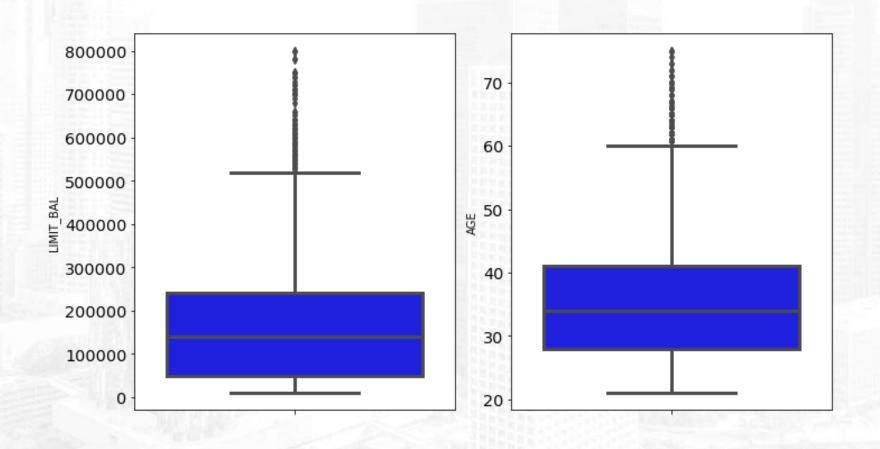
```
df.duplicated().any()

False
```

Tidak ada data yang duplicated dan missing value

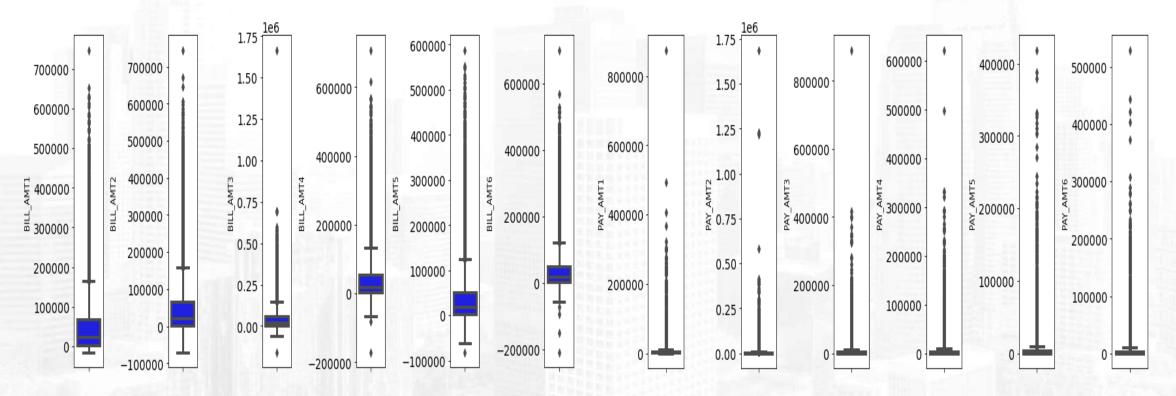
Memeriksa Data Outlier





Memeriksa Data Outlier



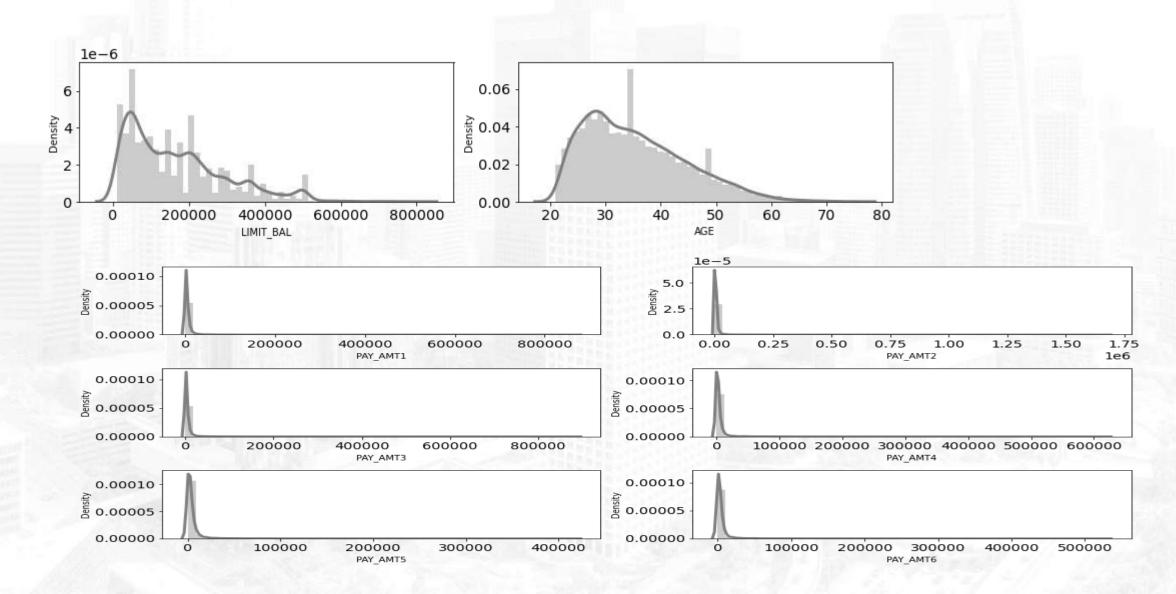


Hasil pengamatan:

- * Terdapat banyak outlier pada semua kolom dengan tipe data numerikal (kecuali data pada kolom ID)
- * Terlihat kolom tersebut distribusi skewed yang ditandai dengan lokasi box yang jauh dari daerah tengah sumbu Y
- * Terdapat nilai minus pada nilai min untuk kolom BILL_AMT1 BILL_AMT6, dimana seharusnya untuk jumlah bill statement tidak ada yang bernilai negatif.

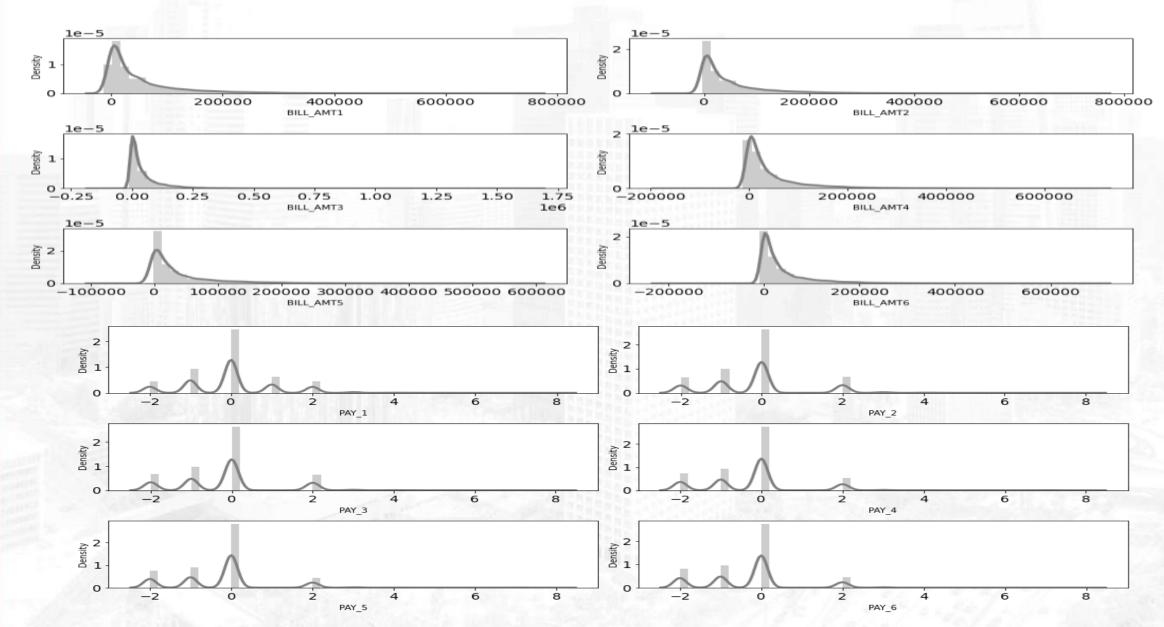
Memeriksa Distribusi Data





Memeriksa Distribusi Data

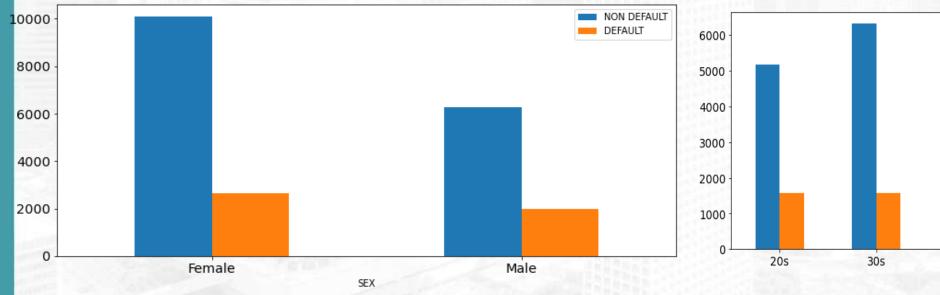


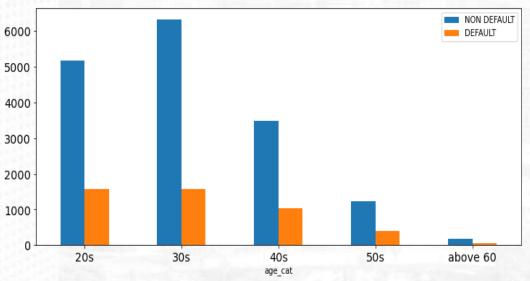




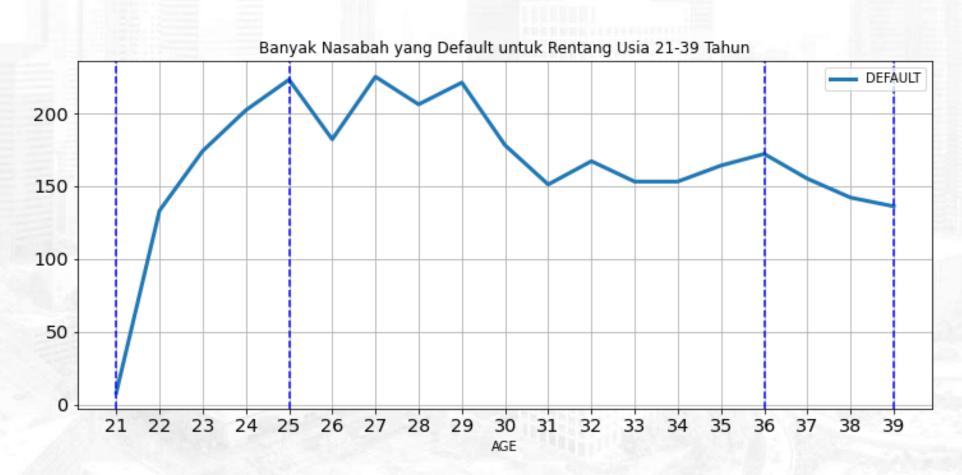


Umur





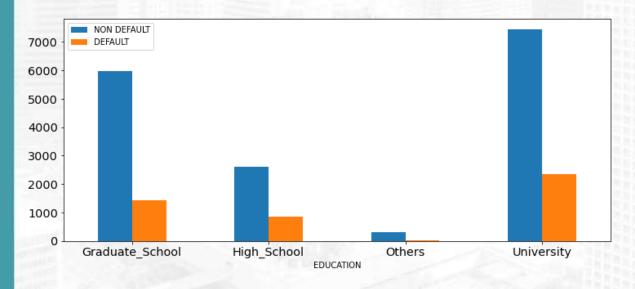


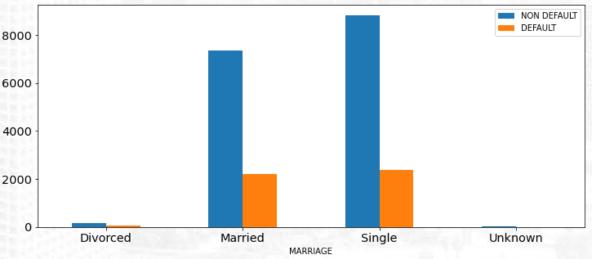




Pendidikan

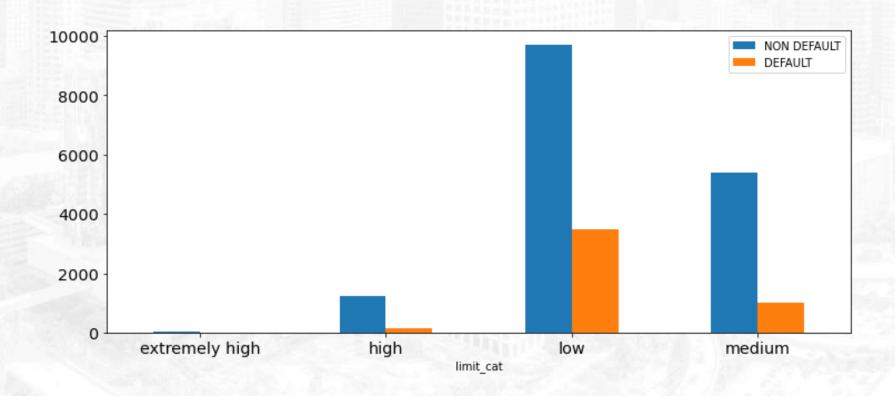
Pernikahan







Kategori Limit



Insight Terkait Default Percentage



- 1. Banyak nasabah bank yang mengalami default di bulan berikutnya adalah **4645 orang** atau sebanyak **22,12%** dari total keseluruhan nasabah bank (21.000).
- 2. Dalam 22,12% tersebut,
 - a. Berdasarkan gender, 9,44% (1983 orang) adalah laki-laki dan <mark>12,68% (2662 orang) adalah perempuan</mark>.
 - **b.** Berdasarkan kategori umur, 7,495% (1574 orang) berumur 20-an tahun, 7,481% (1571 orang) berumur 30-an tahun, 4,986% (1047 orang) berumur 40-an tahun, 1,871% (393 orang) berumur 50-an tahun, dan 0,286% (60 orang) berumur di atas 60 tahun.
 - c. Berdasarkan kategori pendidikan, 6,82% (1433 orang) lulusan graduate school, 11,167% (2345 orang) lulusan universitas (S1), 4,04% (849 orang) lulusan SMA, 0,086% (18 orang) lainnya.
 - d. Berdasarkan status pernikahan, 0,01% (2 orang) tidak diketahui statusnya, 10,56% (2218 orang) sudah menikah, 11,25% (2362 orang) belum menikah, dan 0,3% (63 orang) sudah bercerai.
 - e. Berdasarkan kategori limit, 16,61% (3488 orang) memiliki limit di bawah NTD 200.000, 4,729% (993 orang) memiliki limit di range NTD 200.000 NTD 400.000, 0,748% (157 orang) memiliki limit di range NTD 400.000 NTD 600.000, dan 0,033% (7 orang) memiliki limit di atas NTD 600.000.
 - f. Berdasarkan **repayment status** di bulan April September 2005, <mark>mayoritas nasabah memiliki status 0</mark>.

Insight Terkait Default Percentage



	gender	age_cat	marriage	education	mean_limit	count
51	2	20s	2	2	95049.261084	406
50	2	20s	2	1	118777.429467	319
5	1	20s	2	2	65536.332180	289
60	2	30s	1	2	137832.167832	286
71	2	40s	1	2	138846.153846	208

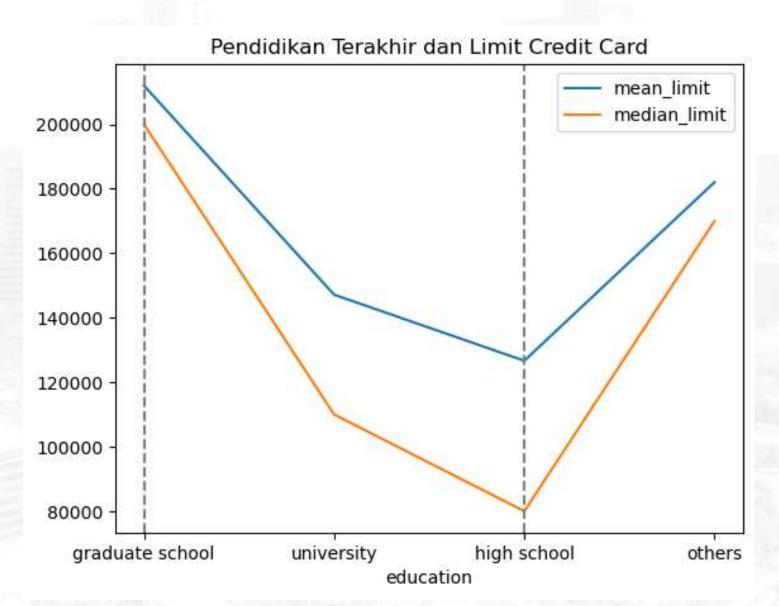
3. Nasabah yang paling banyak default adalah perempuan single lulusan universitas berumur 20an dimana rata-rata limitnya adalah NTD 95.049.

	age	gender	marriage	education	mean_limit	count		
37	24	2	2	2	85652.173913	69		

4. Lebih spesifik, nasabah yang paling banyak default berumur 24 tahun dengan rata-rata limitnya adalah NTD 85.652.

Insight terkait Limit Credit Card





Hasil pengamatan:

Semakin rendah tingkat pendidikan terakhir nasabah bank, semakin rendah juga limit kartu kredit nasabah tersebut.

Korelasi

Koefisien Korelasi	Kuat-lemahnya Korelasi
±0.81 sampai ±1.00	Sangat kuat
±0.61 sampai ±0.80	Kuat
±0.41 sampai ±0.60	Sedang
±0.21 sampai ±0.40	Lemah
±0.00 sampai ±0.20	Tidak ada hubungan

		Ď		EDÍ	M								BILL	18	B	H	BIL	118	M.	ă.	ž.	PAY	PA	8	t_payment_next_month
	- 0	UMIT_BAL -	SEX.	EDUCATION -	MARRIAGE -	AGE -	PAY_0	PAY 2	PAY 3	PAY 4	PAY 5	PAY_6	LAMT1 -	BILL AMT2 -	BILL AMT3 -	BILL AMT4 -	BILL AMTS -	BILL_AMT6 -	PAY_AMT1 -	PAY_AMT2 -	PAY_AMT3 -	Y AMT4 -	PAY_AMT5 -	PAY_AMT6 -	Cmonth -
nt_next_month -	0.02	0.16	0.04	0.02	-0.02	0.01	0.33	0.26	0.24	0.22	0.21	0.19	-0.03	-0.02	-0.02	-0.02	0.02	0.02	-0.08	-0.06	0.07	0.06	0.06	0.06	1.00
PAY_AMT6 -	0.00	0.22	0.00	-0.03	-0.01	0.02	-0.07	0.04	0.04	-0.03	-0.03	-0.03	0.18	0.18	0.18	0.19	0.17	0.12	0.18	0.14	0.17	0.15	0.16	1.00	-0.06
PAY_AMT5 -	0.00	0.22	0.00	-0.04	-0.00	0.02	-0.06	0.04	-0.03	-0.03	-0.03	-0.05	0.17	0.16	0.19	0.17	0.15		0.16	0.19	0.17	0.16	1.00	0.16	-0.06
PAY_AMT4 -	0.01	0.20	0.00	-0.04	0.01	0.03	-0.07	0.05	0.05	0.05	0.06	0.02	0.15	0.15	0.14	0.12		0.24	0.23	0.20	0.26	1.00	0.16	0.15	0.06
PAY_AMT3 -	0.04	0.21	0.00	-0.05	0.00	0.03	-0.08	0.06	0.06	-0.07	0.01	0.00	0.13	0.13	0.12		0.23	0.21	0.31	0.29	1.00	0.26	0.17	0.17	0.07
PAY_AMT2 -																		100000	-			0.20	0.19	0.14	0.06
PAY AMT1 -													and the same	Table and	Name and Address	A CONTRACT	EN HARACI	Treasure.	NAME OF TAXABLE PARTY.	-		0.23			
BILL AMT6 -																		-							****
BILL AMTS -							0.18					allenters.										0.29			
BILL_AMT3 -							0.17					ACCOUNT.						10000							
BILL_AMT2 -							0.18						1					-177.5							
BILL_AMT1 -							0.18						and the					100000							
PAY_6 -						20110			_				· ·				E-1001								
PAY_5 -	0.02	-0.25	0.06	0.09	0.04	-0.06	0.51	0.62	0.69	0.83	1.00	0.82	0.20	0.22	0.24	0.27	0.27	0,26	-0.01	-0.00	0.01	0.06	0.03	0.03	0.21
PAY_4 -	0.00	0.27	0.06	0.10	0.04	-0.06	0.54	0.66	0.78	1.00	0.83	0.72	0.20	0.22	0.24	0.24	0.24	0.24	-0.01	-0.00	-0.07	0.05	0.03	0.03	0.22
PAY_3 -	0.02	-0.29	0.06	0.11	0.04	-0.06	0.57	0.77	1.00	0.78	0.69	0.64	0.21	0.23	0.22	0.22	0.22	0.22	0.00	0.06	0.06	0.05	0.03	0.04	0.24
PAY_2 -	0.01	-0.30	0.07	0.12	0.03	-0.06	0.67	1.00	0.77	0.66	0.62	0.57	0.23	0.23	0.22	0.22	0.22	0.22	-0.08	-0.06	-0.06	0.05	0.04	0.04	0.26
PAY_0 -	-0.03	-0.27	0.06	0.11	0.03	-0.04	1,00	0.67	0.57	0.54	0.51	0.47	0.18	0.18	0.17	0.18	0.18	0.17	-0.09	-0.07	-0.08	0.07	0.06	0.07	0.33
AGE -	0.02	0.14	-0.09	0.17	-0.41	1.00	0.04	0.06	-0.06	-0.06	-0.06	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.03	0.02	0.03	0.03	0.02	0.02	0.01
MARRIAGE -	0.03	0.11	-0.03	0.14	1.00	0.41	0.03	0.03	0.04	0.04	0.04	0.04	-0.03	0.02	-0.03	0.02	0.03	0.02	-0.00	-0.01	0.00	0.01	0.00	0.01	0.02
EDUCATION -	0.04	0.22	0.01	1.00	-0.14	0.17	0.11	0.12	0.11	0.10	0.09	0.08	0.03	0.02	0.02	0.00	0.01	0.01	0.04	€0.0	0.05	0.04	0.04	0.03	0.02
SEX -	0.02	0.03	1.00	0.01	-0.03	0.09	-0.06	0.07	0.06	-0.06	0.06	0.05	-0.03	-0.03	-0.02	-0.02	0.01	0.01	0.00	0.00	-0.00	0.00	0.00	0.00	-0.04
LIMIT_BAL -	0.03	1.00	0.03	-0.22	-0.11	0.14	-0.27	-0.30	-0.29	-0.27	-0.25	-0.24							0.19	0.18	0.21	0.20	0.22	0.22	-0.16
ID -	1.00	0.03	0.02	0.04	-0.03	0.02	-0.03	0.01	0.02	-0.00	-0.02	-0.02	0.01	0.02	0.02	0.04	0.02	0.02	0.01	0.00	0.04	0.01	0.00	0.00	-0.02



- 0.8

. 0.6

- 0.

-02

1201

Hasil Pengamatan Korelasi



- * Korelasi paling kuat terjadi antara kolom PAY_X dengan kolom PAY_X lainnya (X = 1,2,3,4,5,6) dan antara kolom BILL_AMTX dengan kolom BILL_AMTX lainnya (X = 1,2,3,4,5,6).
- * Dari nilai korelasi antara BILL_AMT6 dengan BILL_AMT5, BILL_AMT5 dengan BILL_AMT4, BILL_AMT3 dengan BILL_AMT2, dan BILL_AMT2 dengan BILL_AMT1, didapat bahwa semakin besar jumlah tagihan kartu kredit nasabah pada bulan sebelumnya, semakin besar pula jumlah tagihan kartu kredit nasabah pada bulan berikutnya (misal jumlah tagihan pada bulan April besar, maka tagihan pada bulan Mei akan semakin besar). Begitu pula sebaliknya.
 - * Hal ini dapat dipengaruhi oleh adanya bunga kartu kredit dan biaya keterlambatan pembayaran.
- * Nilai korelasi antara PAY_6 dengan PAY_5 adalah 0.82, PAY_5 dengan PAY_4 berkorelasi sebesar 0.83, PAY_4 dengan PAY_3 berkorelasi sebesar 0.78, PAY_3 dengan PAY_2 berkorelasi sebesar 0.77, PAY_2 dengan PAY_1 berkorelasi sebesar 0.67.
- * Dari sini dapat dilihat bahwa hubungan antara repayment status seorang nasabah dari bulan satu ke bulan lainnya semakin lama cenderung semakin lemah (dari PAY_6 ke PAY_1).
- * Dari nilai-nilai korelasi ini didapat juga bahwa apabila di bulan sebelumnya nasabah menunggak pembayaran tagihan kartu kredit, maka di bulan berikutnya nasabah tersebut juga kemungkinan besar akan kembali menunggak. Begitu pula sebaliknya. Jika di bulan sebelumnya nasabah melakukan pembayaran tepat waktu / tidak menunggak, maka kemungkinan besar di bulan berikutnya nasabah juga tidak akan menunggak.
- * Di antara kolom-kolom lainnya, kolom yang paling berkorelasi dengan variabel target adalah kolom PAY_1 yang kemudian diikuti dengan kolom PAY_2, PAY 3, PAY 4, PAY 5, dan PAY 6.
- * Limit nasabah memiliki korelasi positif yang lemah (sampai dengan 0.3) dengan jumlah tagihan nasabah (BILL_AMT).
- * Ada kemungkinan semakin besar limit nasabah, semakin besar juga jumlah tagihannya. Begitu pula sebaliknya.
- * Limit nasabah memiliki korelasi negatif yang lemah dengan repayment status (PAY_1 sampai PAY_6).
- * Ada kemungkinan semakin besar limit nasabah, semakin kecil nilai pada status repaymentnya. Begitu pula sebaliknya.

Rekomendasi Bisnis



Dari penggalikan insight bisnis yang kami lakukan rekomendasi yang kami berikan ke perusahaan adalah :

1. Memberikan kemudahan untuk mengajukan peningkatan limit kepada pemegang kartu kredit yang melakukan payment tepat waktu.

2. Nasabah yang terdeteksi berpotensi akan gagal bayar segera dihubungi lebih dulu dan ditawarkan solusi.

Kesimpulan



- 1. Setiap kolom memiliki tipe data yang sesuai
- 2. Dataset terdiri dari 25 kolom dan 21000 baris tidak terdapat missing value dan tidak ada data yang duplikat
- 3. Penamaan kolom yang tidak sesuai pada kolom PAY_0 langsung ke PAY_2, sementara kolom BILL_AMT dan PAY_AMT diawali dengan angka 1 bukan 0
- 4. Nilai 0 kita asumsikan sebagai tepat bayar , -1 sebagai cepat bayar satu bulan dan -2 cepat bayar dua bulan
- 5. Terdapat nilai minus pada nilai kolom BILL_AMT1 BILL_AMT6
- 6. Kolom SEX, EDUCATION, MARRIAGE, PAY_0 PAY_6, default_payment_next_month tampaknya sudah dilabel encoding
- 7. Kolom LIMIT_BALL,AGE,BILL_AMT1-BILL_AMT6,PAY_AMT1-PAY_AMT6 distribusi datanya right skewed

Rekomendasi Pre-Processing



- 1. Kolom ID akan di drop karena merupakan identifier dari tiap baris yang nilainya unik dan tidak dapat memberikan informasi apa-apa dalam analisis.
- 2. Merubah kolom PAY_0 menjadi PAY_1
- 3. Terdapat nilai yang belum terdefinisi pada kolom EDUCATION akan diubah menjadi unknown
- 4. Terdapat nilai minus pada kolom BILL_AMT1 BILL_AMT6 hal itu wajar terjadi tetapi nilai tersebut akan di drop karena outlier
- 5. Pada visualisasi boxplot terdapat banyak outlier di kolom PAY_AMT1 PAY_AMT 6 dan BILL_AMT1-BILL_AMT6, untuk penanganannya akan memfilter outlier dengan menggunakan z-score, kemudian akan dicek distribusi datanya setelah difilter
- 6. Melakukan sedikit experiment perbandingan antara z-score dengan IQR untuk pengamatan perbandingan hasil akurasinya
- 7. kolom PAY, BILL_AMT, PAY_AMT memiliki hubungan sebab akibat, kita akan memilih salah satu nya atau akan kami pertimbangkan lebih lanjut untuk tidak drop kolom tersebut sebagai bahan pertimbangan akurasi model kedepannya