Rakamin Academy

Credit Card Payment Default Prediction

Dokumen Laporan Final Project

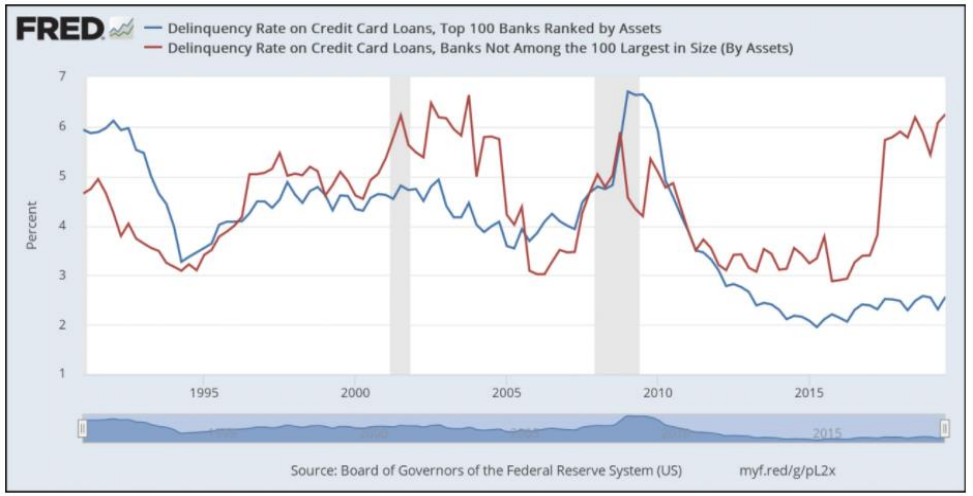
Cristanto – Steven Benny – Tri Setiawan – Ulva Dewiyanti

DataRider

Latar Belakang Masalah

Adapun latar belakang masalah adalah sebagai berikut,

- Jumlah customer pengguna layanan kartu kredit pada September 2005 sebanyak 21000 customer.
- Namun, 23% dari customer tersebut mengalami gagal bayar (*default*) pada September 2005.
- Sementara itu berdasarkan data FRED, default rate global dari tahun 1990 hingga 2005 berada dalam range 3% hingga 6,5%.



* Federal Reserve Economic Data (FRED) adalah database yang dikelola oleh divisi Riset Federal Reserve Bank of St. Louis yang memiliki lebih dari 816.000 deret waktu ekonomi dari berbagai sumber yang mencakup perbankan, bisnis/fiskal, indeks harga konsumen, nilai tukar, produk domestik bruto, suku bunga, dsb.

Lingkup Kerja

Lingkup kerja dari project adalah sebagai berikut,

1. **Problem Statement**

   Bagaimana cara untuk mengurangi default rate dengan menganalisis dan memprediksi debitur yang akan default pada bulan berikutnya berdasarkan data?

2. **Goal**

   Mengurangi default rate pada bulan berikutnya

3. **Objective**

   • Melakukan analisis prediktif menggunakan machine learning untuk memprediksi debitur yang default pada bulan berikutnya
   • Merekomendasikan strategi yang tepat untuk mengurangi default rate

4. **Business Metrics**

   Default rate

Informasi Dataset

* `ID`: Unique identifier untuk setiap klien/debitur
* `LIMIT_BAL`: Jumlah kredit yang diberikan dalam NT dollar
* `SEX`: Jenis kelamin (1=laki-laki, 2=perempuan)
* `EDUCATION`: (1=pascasarjana, 2=universitas, 3=SMA, 4=lainnya, 5=tidak diketahui, 6=tidak diketahui)
* `MARRIAGE`: Status pernikahan (1=menikah, 2=lajang, 3=bercerai)
* `AGE`: Usia dalam tahun
* `PAY_0`: Status pembayaran pada bulan September 2005 (-
1=pay duly, 1=payment delay for one month, 2=payment delay for two months, ... 8=payment delay for eight months, 9=payment delay for nine months and above)
* `PAY_2`: Status pembayaran pada Agustus 2005 (skala sama dengan `PAY_0`)
* `PAY_3`: Status pembayaran pada Juli 2005 (skala sama dengan `PAY_0`)
* `PAY_4`: Status pembayaran pada Juni 2005 (skala sama dengan `PAY_0`)
* `PAY_5`: Status pembayaran pada Mei 2005 (skala sama dengan `PAY_0`)
* `PAY_6`: Status pembayaran pada April 2005 (skala sama dengan `PAY_0`)
* `BILL_AMT1`: Jumlah tagihan tagihan pada bulan September 2005 (NT dollar)
* `BILL_AMT2`: Jumlah tagihan tagihan pada bulan Agustus 2005 (NT dollar)
* `BILL_AMT3`: Jumlah tagihan tagihan pada bulan Juli 2005 (NT dollar)
* `BILL_AMT4`: Jumlah tagihan tagihan pada bulan Juni 2005 (NT dollar)
* `BILL_AMT5`: Jumlah tagihan tagihan pada bulan Mei 2005 (NT dollar)
* `BILL_AMT6`: Jumlah tagihan tagihan pada bulan April 2005 (NT dollar)
* `PAY_AMT1`: Jumlah pembayaran sebelumnya di bulan September 2005 (NT dollar)
* `PAY_AMT2`: Jumlah pembayaran sebelumnya di bulan Agustus 2005 (NT dollar)
* `PAY_AMT3`: Jumlah pembayaran sebelumnya di bulan Juli 2005 (NT dollar)
* `PAY_AMT4`: Jumlah pembayaran sebelumnya di bulan Juni 2005 (NT dollar)
* `PAY_AMT5`: Jumlah pembayaran sebelumnya di bulan Mei 2005 (NT dollar)
* `PAY_AMT6`: Jumlah pembayaran sebelumnya di bulan April 2005 (NT dollar)
* `default.payment.next.month`: Pembayaran default di bulan berikutnya (1=yes, 0=no)

Eksplorasi Dataset

	count	mean	std	min	25%	50%	75%	max
ID	21000.0	14949.183667	8632.775153	1.0	7508.75	14939.5	22386.75	29998.0
LIMIT_BAL	21000.0	167214.746667	128965.188482	10000.0	50000.00	140000.0	240000.00	800000.0
SEX	21000.0	1.607571	0.488303	1.0	1.00	2.0	2.00	2.0
EDUCATION	21000.0	1.854190	0.791628	0.0	1.00	2.0	2.00	6.0
MARRIAGE	21000.0	1.551714	0.521176	0.0	1.00	2.0	2.00	3.0
AGE	21000.0	35.461619	9.206628	21.0	28.00	34.0	41.00	75.0
PAY_0	21000.0	-0.011190	1.123210	-2.0	-1.00	0.0	0.00	8.0
PAY_2	21000.0	-0.127238	1.198957	-2.0	-1.00	0.0	0.00	8.0
PAY_3	21000.0	-0.164857	1.198624	-2.0	-1.00	0.0	0.00	8.0
PAY_4	21000.0	-0.218190	1.172210	-2.0	-1.00	0.0	0.00	8.0
PAY_5	21000.0	-0.260952	1.141454	-2.0	-1.00	0.0	0.00	8.0
PAY_6	21000.0	-0.288667	1.151592	-2.0	-1.00	0.0	0.00	8.0
BILL_AMT1	21000.0	51501.542381	73453.641859	-14386.0	3564.75	22578.0	67876.25	746814.0
BILL_AMT2	21000.0	49463.502667	70866.586004	-69777.0	3000.00	21550.0	64918.25	743970.0
BILL_AMT3	21000.0	47232.577762	69539.883466	-157264.0	2686.25	20242.0	60826.75	1664089.0
BILL_AMT4	21000.0	43387.372476	64081.073110	-170000.0	2332.00	19158.5	55376.75	706864.0
BILL_AMT5	21000.0	40398.551095	60396.811177	-81334.0	1759.00	18266.5	50517.25	587067.0
BILL_AMT6	21000.0	38931.194000	59196.499234	-209051.0	1242.75	17203.5	49463.00	699944.0
PAY_AMT1	21000.0	5686.349333	16868.075695	0.0	998.25	2100.0	5023.25	873552.0
PAY_AMT2	21000.0	5923.003476	23909.526477	0.0	836.00	2011.0	5000.00	1684259.0
PAY_AMT3	21000.0	5202.325333	17006.416467	0.0	390.00	1811.5	4500.00	889043.0
PAY_AMT4	21000.0	4793.172000	15467.403159	0.0	284.00	1500.0	4002.25	621000.0
PAY_AMT5	21000.0	4797.012952	15270.031988	0.0	241.00	1500.0	4051.00	417990.0
PAY_AMT6	21000.0	5211.736762	17698.795697	0.0	102.00	1500.0	4000.00	528666.0
default_payment_next_month	21000.0	0.221190	0.415058	0.0	0.00	0.0	0.00	1.0

Pengamatan :

1. Terdapat penamaan kolom yang kurang selaras yaitu setelah kolom PAY_0 langsung ke PAY_2, sementara pada BILL_AMT dan PAY_AMT diawali dengan 1 bukan 0 dan definisi kolomnya urut sesuai dengan BILL_AMT dan PAY_AMT.

2. Terdapat nilai yang belum terdefinisi dalam dataset kolom PAY_0 - PAY_6, yaitu nilai 0 dan -2

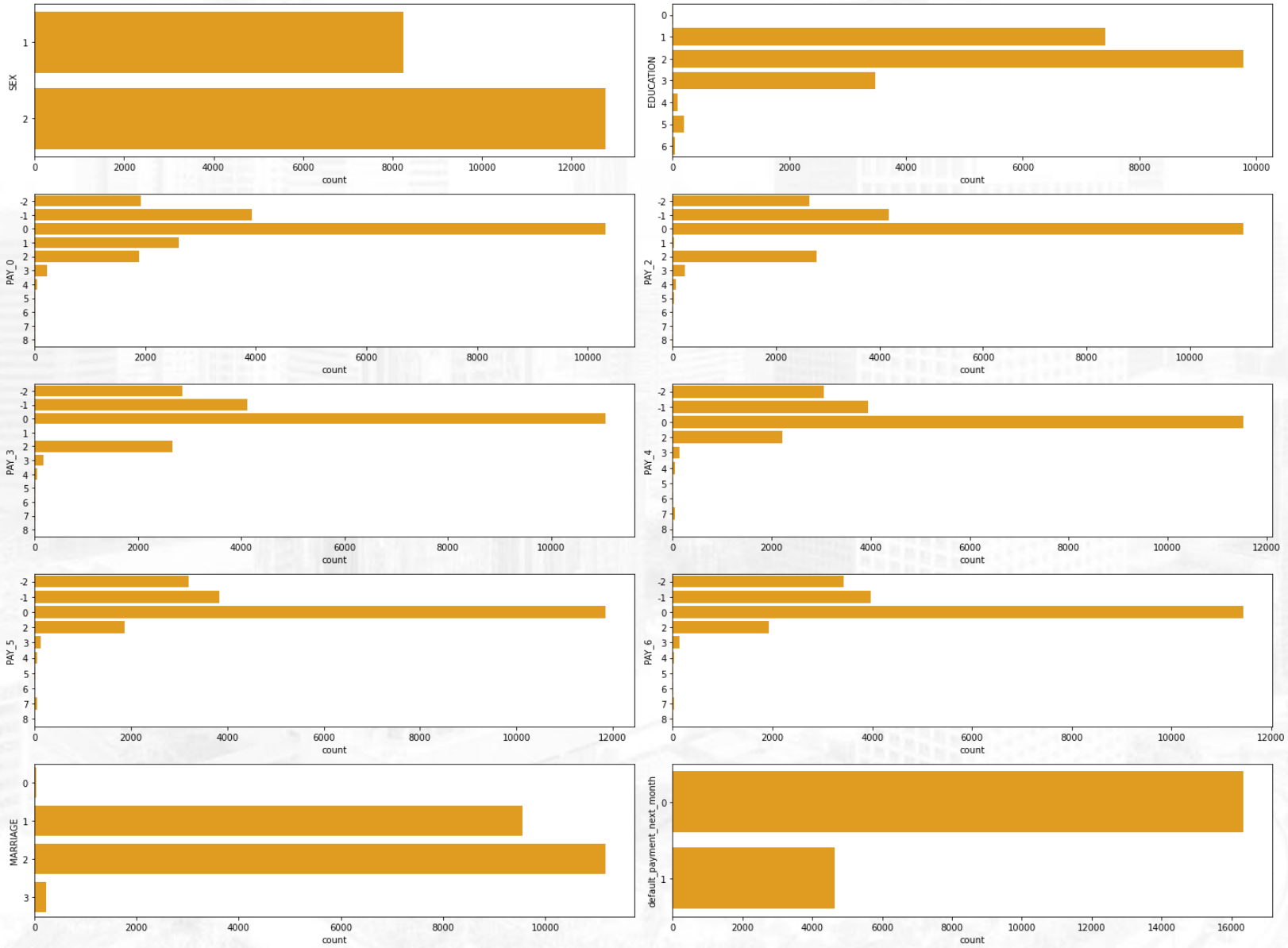
3. Terdapat nilai minus pada nilai min untuk kolom BILL_AMT1 - BILL_AMT6, dimana kami asumsikan sebagai kelebihan bayar.

4. Kolom SEX, EDUCATION, MARRIAGE, PAY_0 - PAY_6, default_payment_next_month merupakan data kategorikal yang direpresentasikan menggunakan numerik.

5. Hanya kolom AGE yang tampak sudah cukup simetrik distribusinya (mean dan median tak berbeda jauh)

6. Kolom LIMIT_BAL, BILL_AMT1 - BILL_AMT6, PAY_AMT1 - PAY_AMT6 sepertinya right skewed.
    1. Kolom LIMIT_BAL, PAY_AMT1 - PAY_AMT6 memiliki nilai mean > median dan selisih percentil 75 dengan max sangat jauh.
    2. BILL_AMT1 - BILL_AMT6 memiliki nilai mean > median.

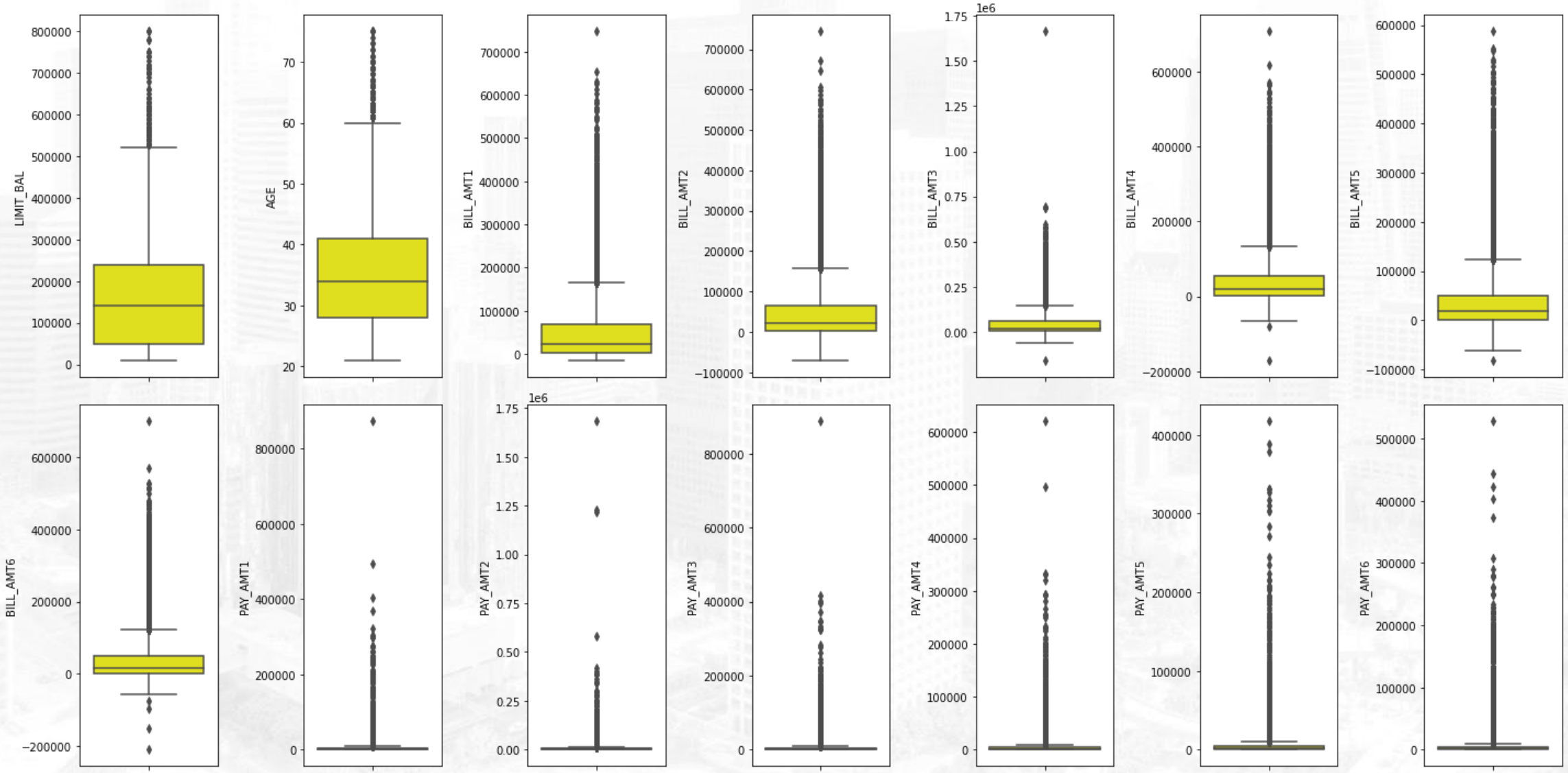
Eksplorasi Dataset



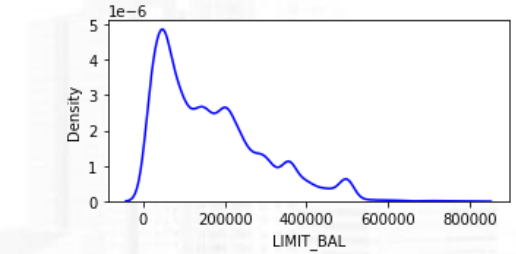
Pengamatan:

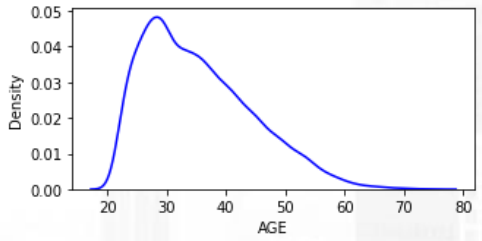
1. SEX didominasi oleh kategori 2 (female)
2. EDUCATION didominasi oleh kategori 2 (university)
3. PAY_0 - PAY_6 didominasi oleh kategori 0 (tidak terdefinisi di dataset) dan terdapat jumlah kategori yang cukup banyak
4. MARRIAGE didominasi oleh kategori 2 (single) dan disusul oleh kategori 1 (married)
5. default_payment_next_month didominasi oleh 0 (not default), tampak bahwa terdapat class imbalance pada label.

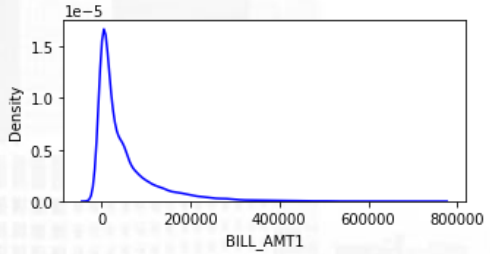
Eksplorasi Dataset

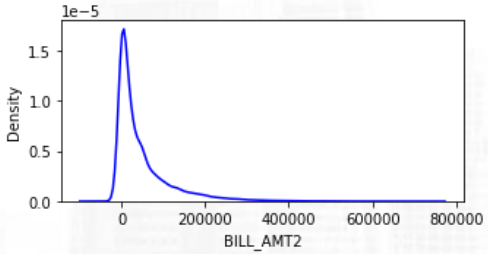


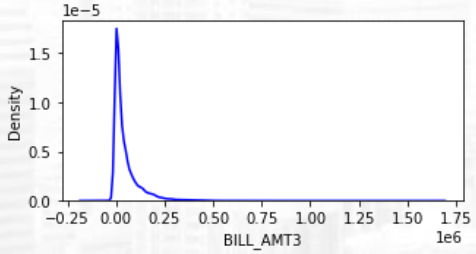
Eksplorasi Dataset

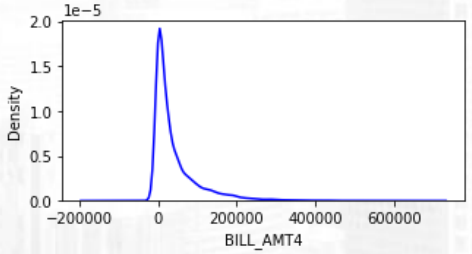


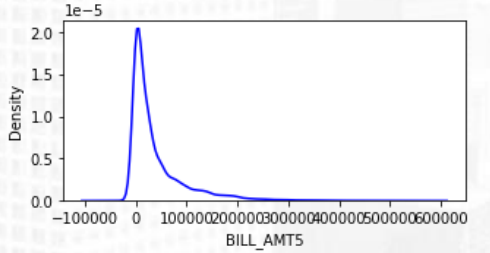


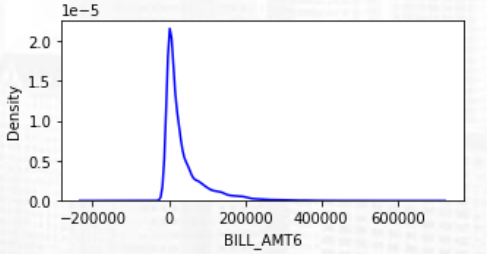


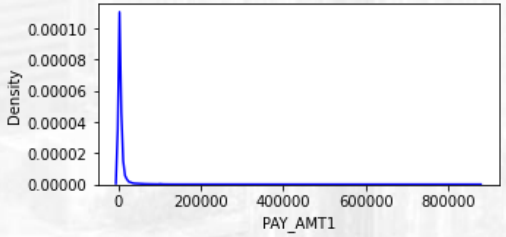


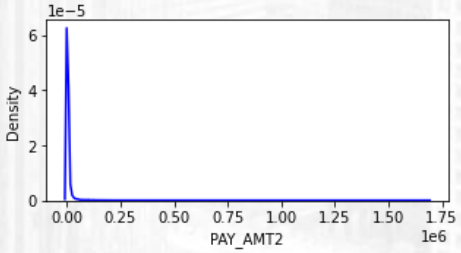


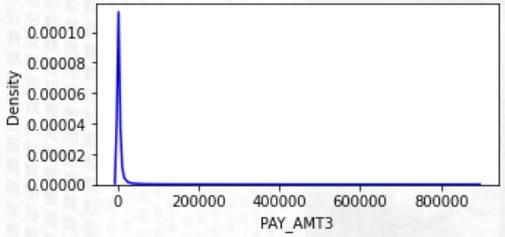


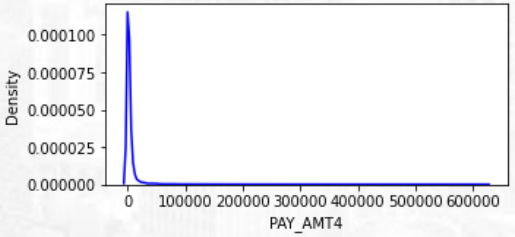


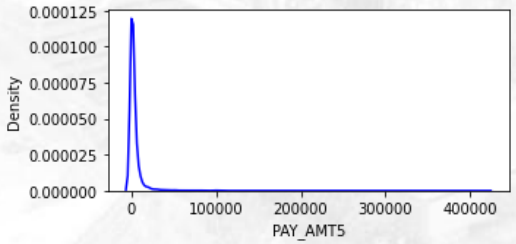


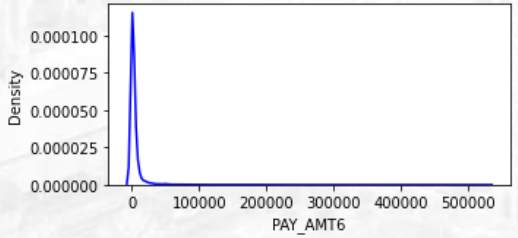




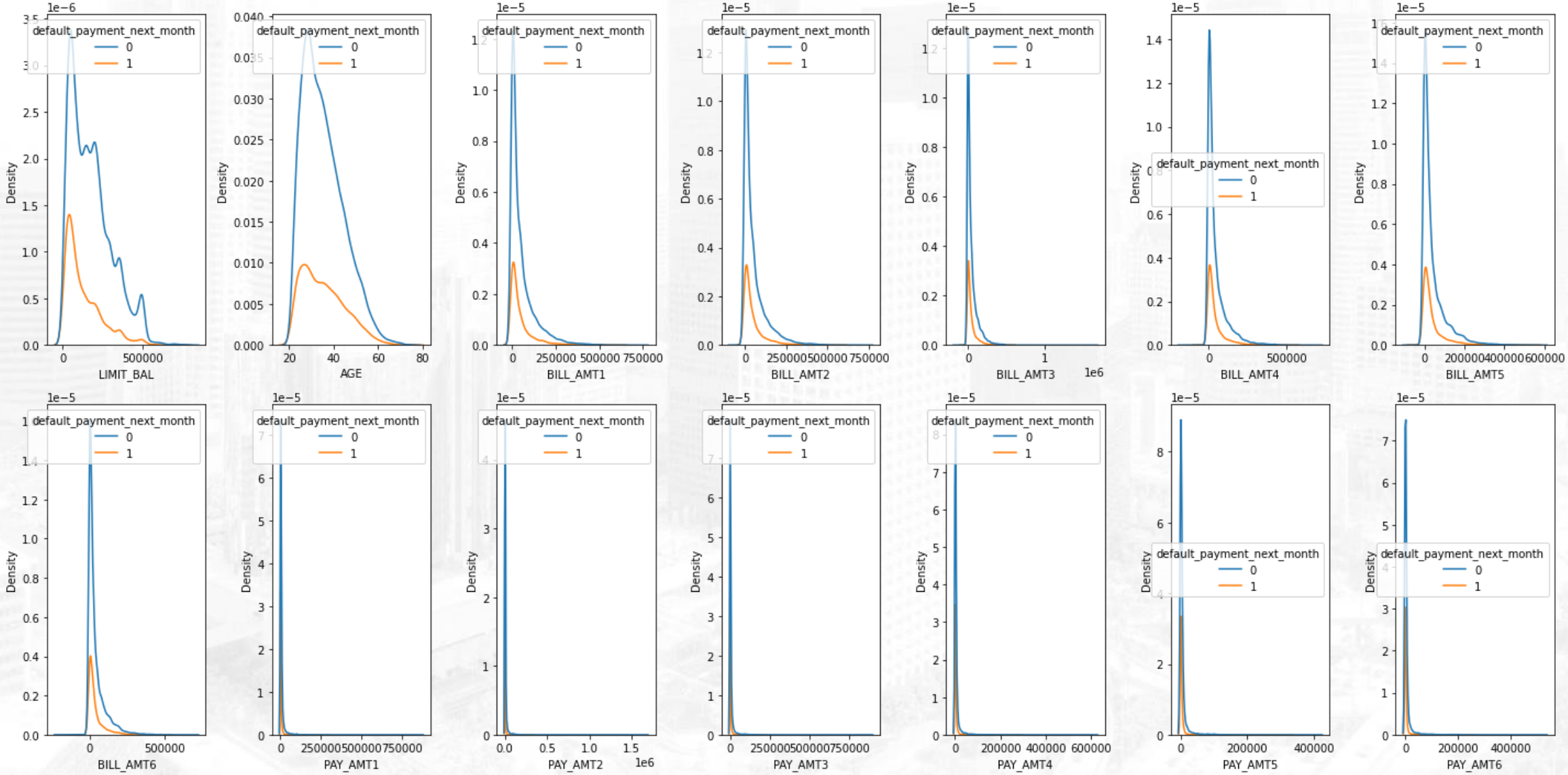




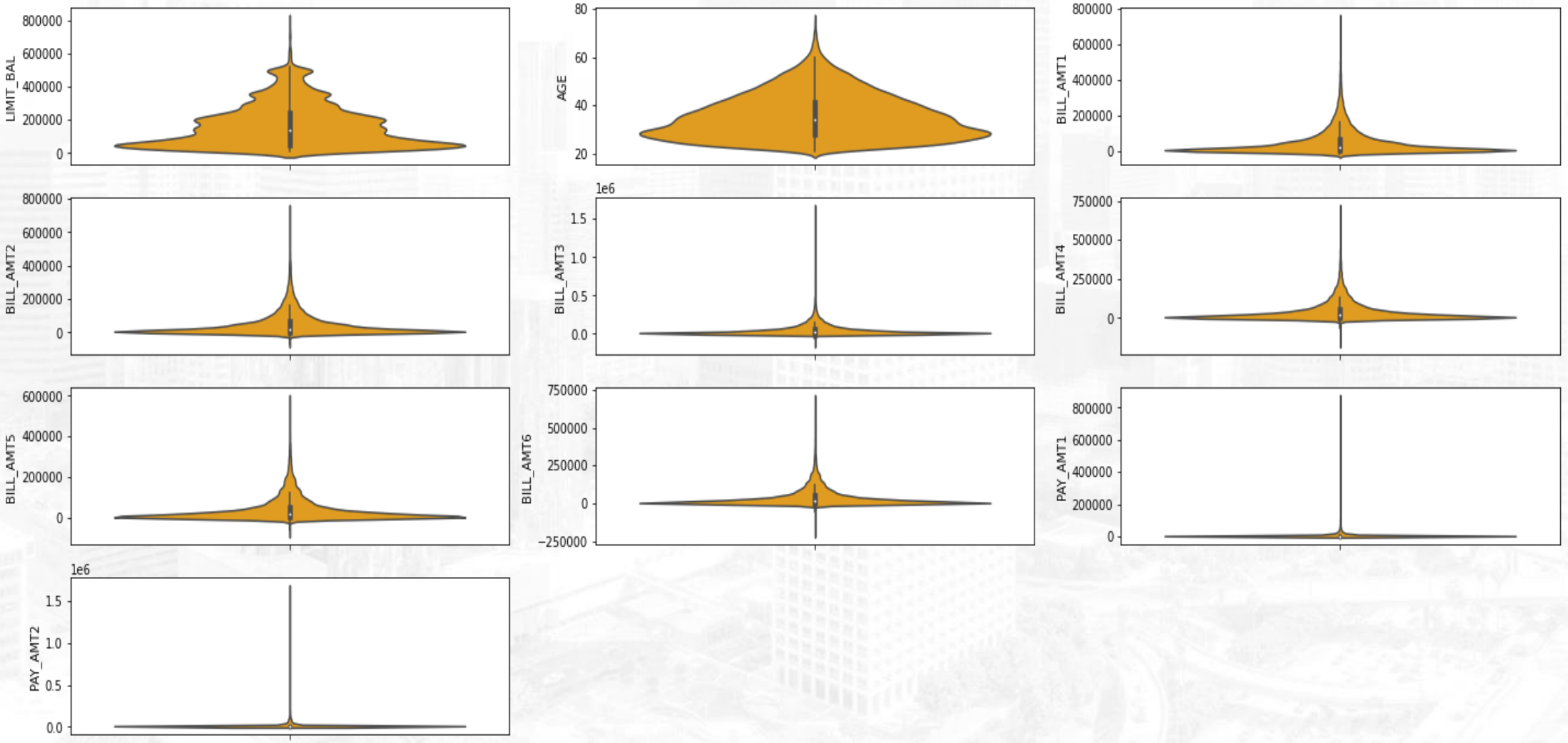




Eksplorasi Dataset



Eksplorasi Dataset

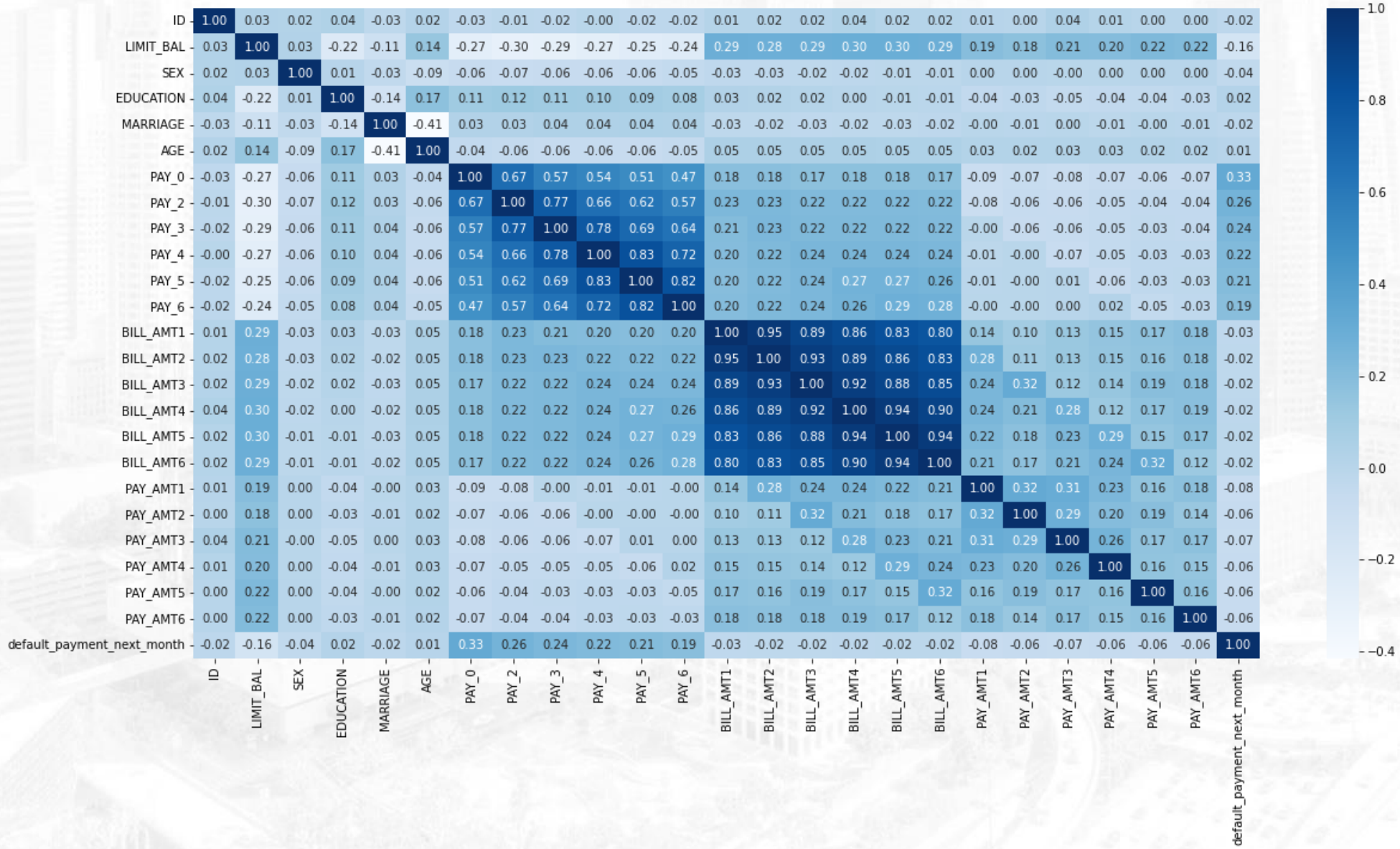


Eksplorasi Dataset

Pengamatan:

1. Tampak disetiap numerikal data pada dataset memiliki outlier semua. (untuk penanganan yang dilakukan y aitu dengan memfilter outlier dengan menggunakan z score, kemudian akan dilakukan pengecekan distribus i dari data setelah difilter dengan z score)
2. `BILL_AMT3 - BILL_AMT6` memiliki outlier dikedua sisinya baik di bagian positif atau negatif. (seperti dijelaskan diatas untu k `BILL_AMT` memang terdapat keanehan dari datanya yaitu ada cukup banyak data yang bernilai negatif. Ole h karena itu, kami asumsikan nilai minus tersebut adalah kelebihan bayar dari jumlah tagihan yang seharusnya)
3. Berdasarkan visualisasi, semua numerical data pada dataset tampak skew ke kanan semua (penanganannya akan dilakukan filtrasi outlier)
4. Hanya kolom AGE yang tampak sudah cukup simetrik distribusinya (mean dan median tak berbeda jauh)

Eksplorasi Dataset



Eksplorasi Dataset

Pengamatan:

1. Target default_payment_next_month
   * memiliki korelasi positif lemah dengan `LIMIT_BAL, Pay_0, Pay_1, Pay_2, Pay_3, Pay_4, Pay_5, Pay_6` yang s elanjutkan akan digunakan pada model.
   * tidak memiliki korelasi positif cukup kuat terhadap feature apapun.

2. korelasi antar-feature
   * Terdapat banyak sekali feature yang saling berkolerasi positif cukup kuat (`BILL_AMT` dengan `BILL_AMT` dan `PAY_AMT` dengan `PAY_AMT`) dan ada juga featur yang berkorelasi negatif lemah (`AGE - MARRIAGE`)

Eksplorasi Dataset

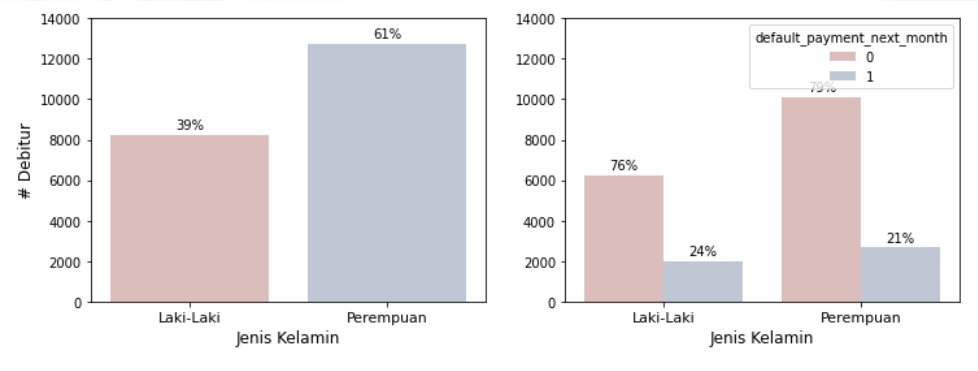
Feature Selection dengan SelectKBest Library

```
[ ]  X= df_train_20D8GL3.drop(columns=["default_payment_next_month", "AGE_BIN","ID"])
     y= df_train_20D8GL3['default_payment_next_month']
```

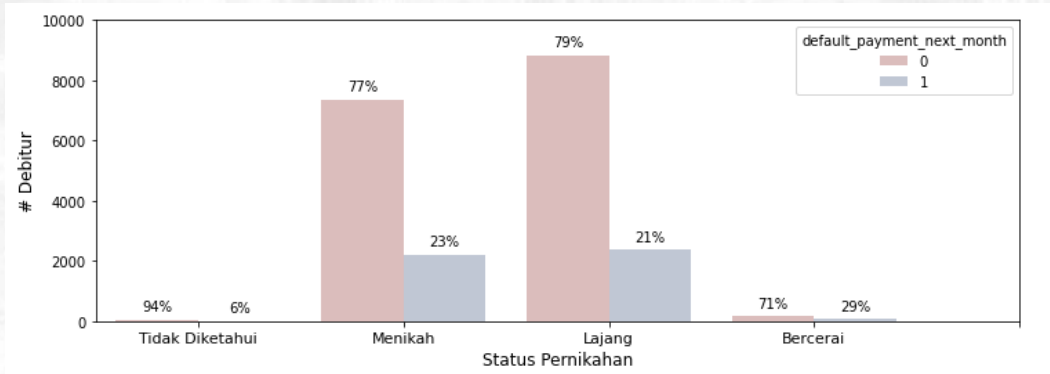
```
[ ]  uni = SelectKBest(score_func = f_classif, k = 23)
     fit = uni.fit(X, y)
     X.columns[fit.get_support(indices=True)].tolist()
```

```
['LIMIT_BAL',
 'SEX',
 'EDUCATION',
 'MARRIAGE',
 'AGE',
 'PAY_1',
 'PAY_2',
 'PAY_3',
 'PAY_4',
 'PAY_5',
 'PAY_6',
 'BILL_AMT1',
 'BILL_AMT2',
 'BILL_AMT3',
 'BILL_AMT4',
 'BILL_AMT5',
 'BILL_AMT6',
 'PAY_AMT1',
 'PAY_AMT2',
 'PAY_AMT3',
 'PAY_AMT4',
 'PAY_AMT5',
 'PAY_AMT6']
```

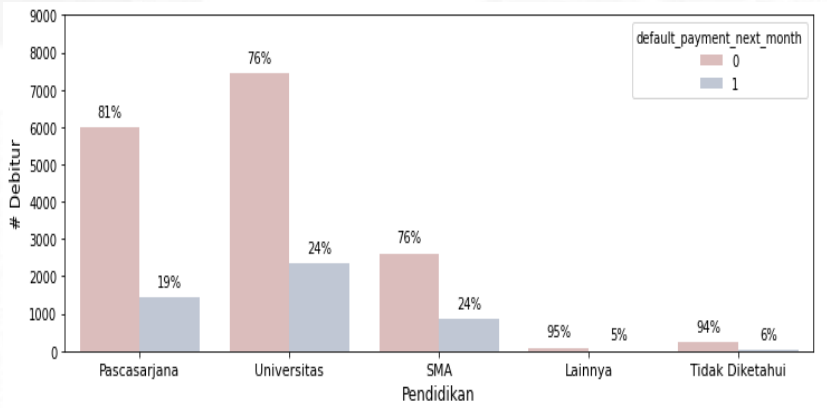
Eksplorasi Dataset



61% debitur adalah perempuan,namun berdasarkan grafik diatas, tampak laki-laki memiliki peluang gagal bayar yang sedikit lebih tinggi yaitu 24% dibandingkan perempuan yang hanya 21%

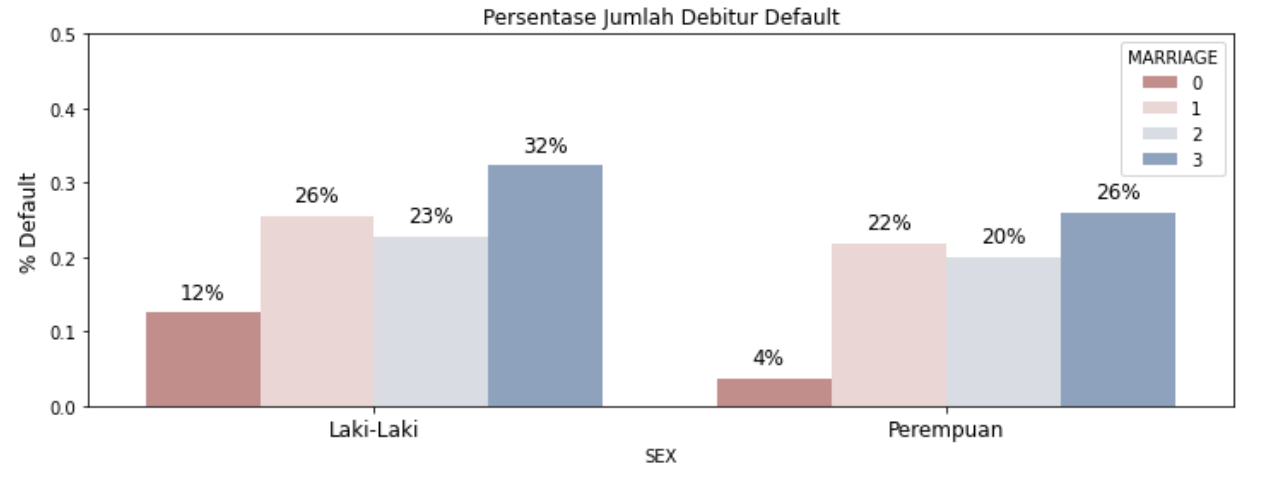


- Sebagian besar debitur berasal dari kategori Menikah dan Lajang.
- Sementara itu, debitur yang memiliki status pernikahan bercerai memiliki peluang untuk gagal bayar lebih besar dibandingkan kategori lainnya , disusul oleh status menikah, dan kemudian lajang

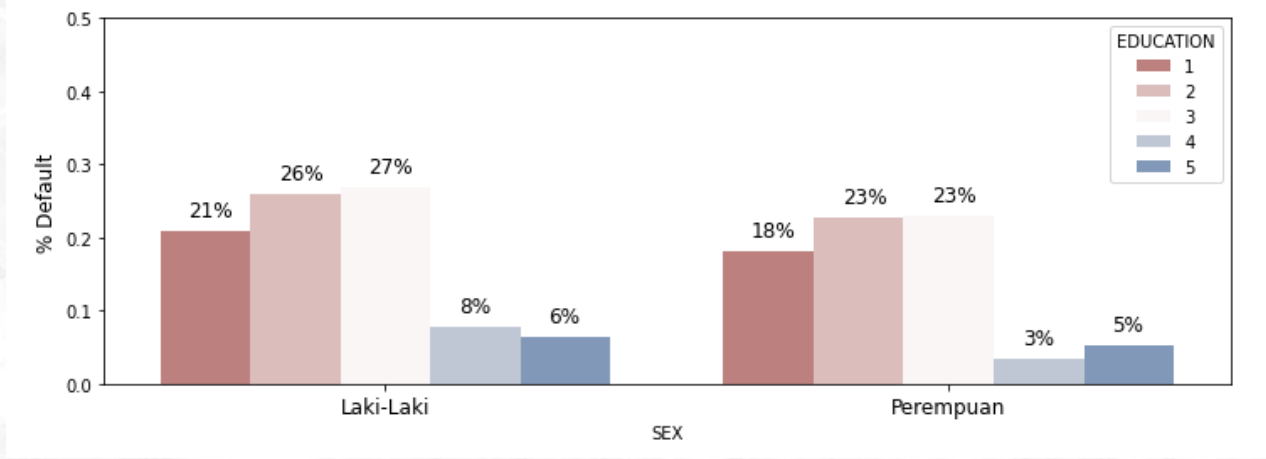


- Tingkat pendidikan didominasi oleh Universitas, diikuti oleh Pasca sarjana, SMA, Tidak Diketahui dan Lainnya.
- Apabila melihat 3 kategori pertama dalam grafik (pascasarjana, u niversitas, dan SMA), tampak bahwa semakin tinggi pendidikan de bitur maka peluang gagal bayar menjadi lebih rendah, hal ini mun gkin dikarenakan pemahaman mengenai rencana keuangan yang lebih baik seiring dengan meningkatnya tingkat pendidikan debitu r.
- Namun kategori Tidak Diketahui dan Lainnya memiliki peluang ga gal bayar yang jauh lebih rendah dari tiga kategori lainnya, tapi unt uk kedua kategori tersebut tidak dapat ditentukan apakah merup akan pendidikan yang lebih tinggi dari pascasarjana atau lebih ren dah dari SMA.

Eksplorasi Dataset

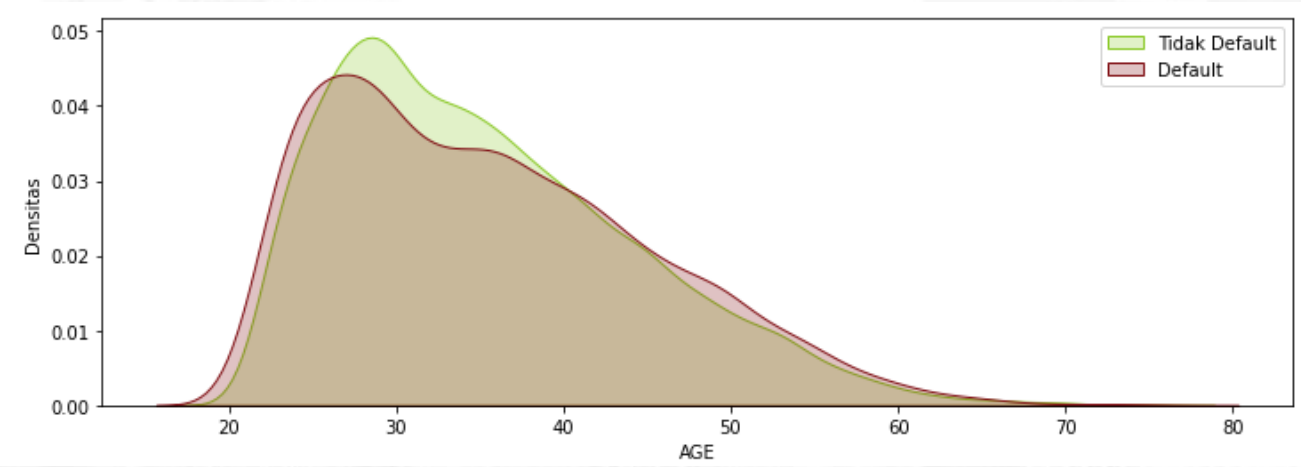


- di setiap kategori MARRIAGE ternyata dapat disimpulkan bahwa laki-laki dengan status Bercerai lebih banyak yang gagal bayar, diikuti oleh laki-laki yang Menikah kemudian Lajang
- Jika melihat dari Perempuan yang memilki kesamaan grafik dengan laki-laki, hanya berbeda di nilai besarannya

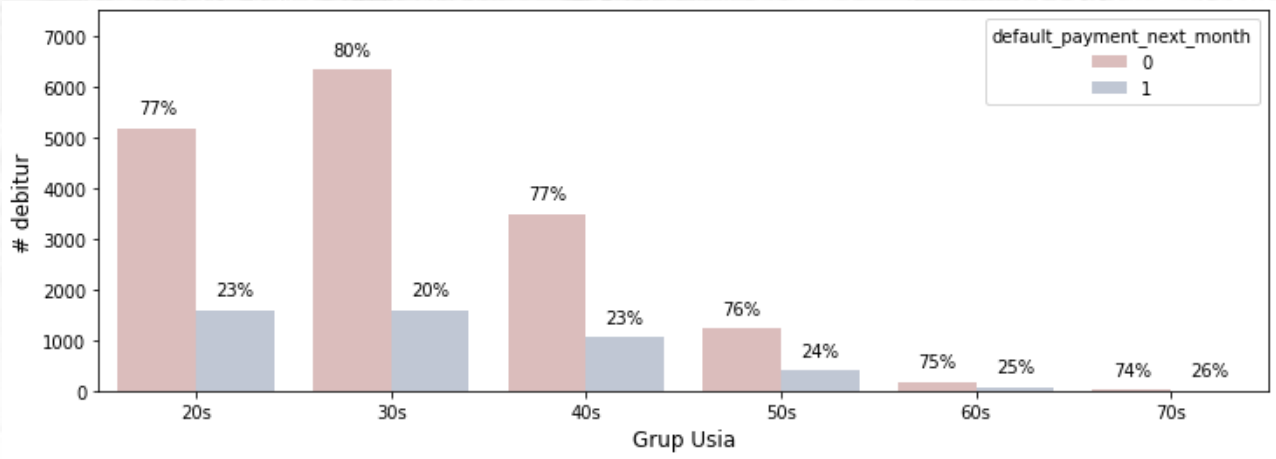


- di setiap kategori EDUCATION ternyata dapat disim pulkan bahwa memang laki-laki SMA lebih cenderung gagal bayar diikuti oleh la ki laki yang berkuliah di universitas.
- Nilai perempuan yang SMA dan Berkuliah di univer sitas memiliki kesamaan nilai gagal bayar

Eksplorasi Dataset

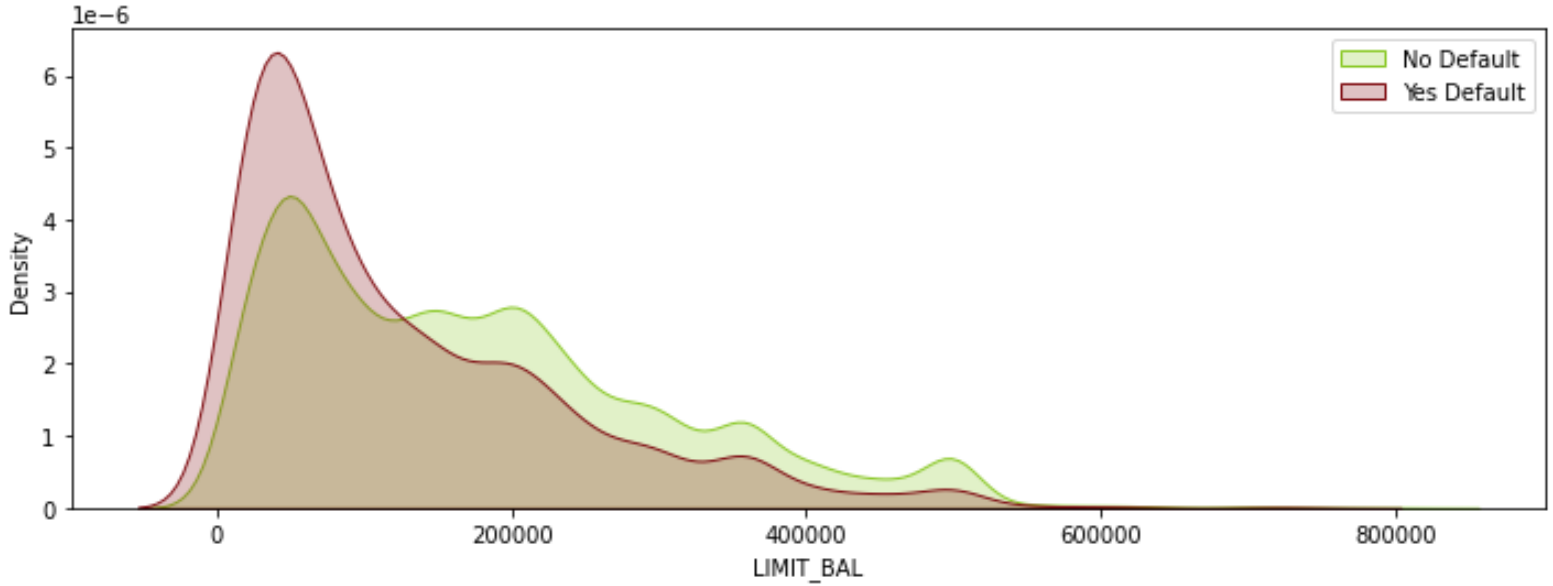


Nilai min untuk AGE adalah 21 tahun, dan AGE maks 75, serta rata-rata AGE adalah 35 tahun.



- Debitur didominasi oleh usia antara 25 sampai 40 tahun, berdasarkan grafik diatas juga terlihat bahwa dalam range usia 25 - 40 tahun tersebut memiliki peluang default yang lebih rendah
- Peluang default paling rendah yaitu debitur dengan usia 30an (30-39 tahun), sementara default yang tinggi berada pada usia-usia lanjut yaitu default tertinggi pada range usia 70-79 tahun, disusul oleh 60-69 tahun, kemudian 50-59 tahun.

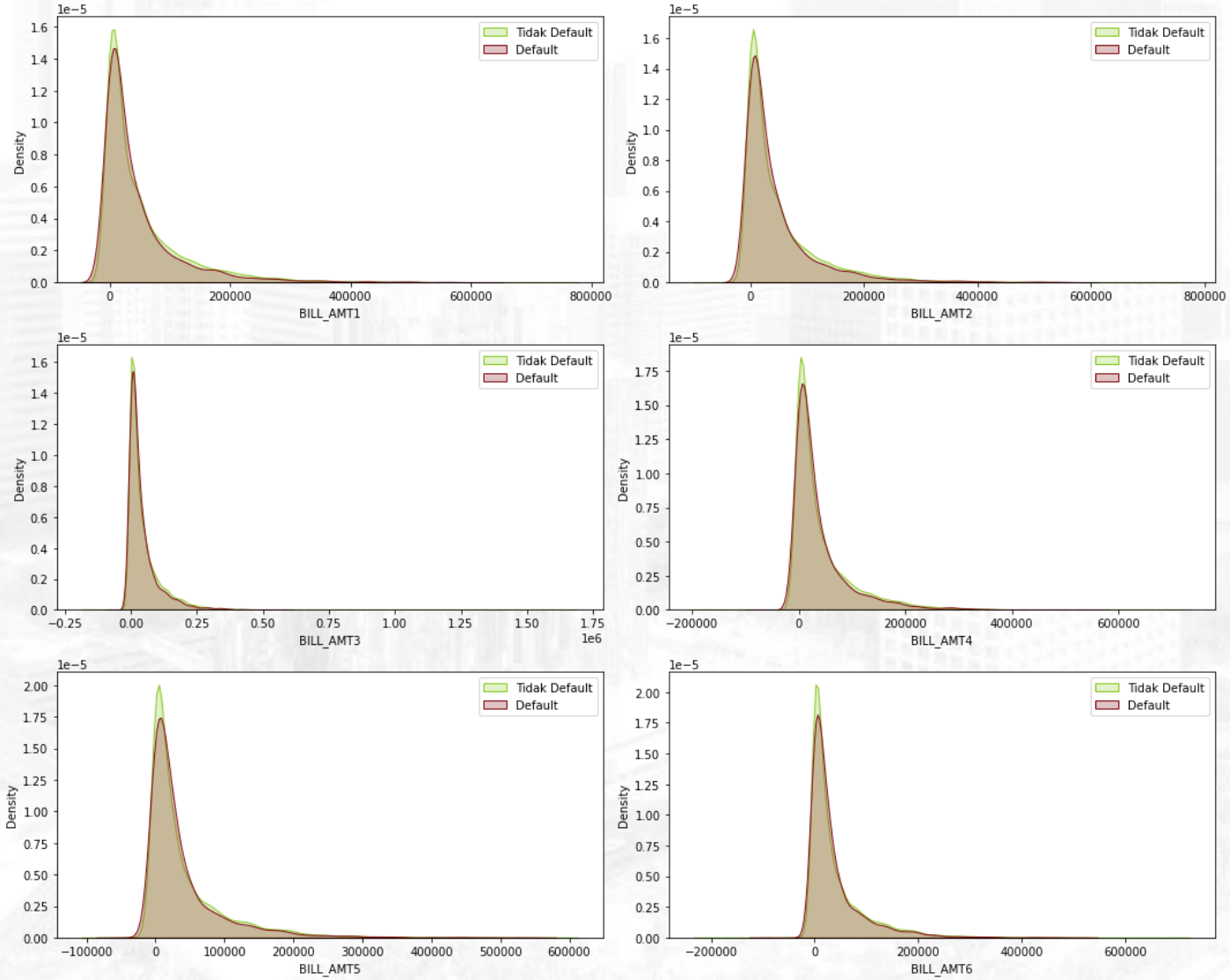
Eksplorasi Dataset



Pengamatan:

1. Dapat disimpulkan bahwa data pada kolom LIMIT_BAL memiliki distribusi yang right-skewed dan berdasarkan grafik terdapat keberadaan beberapa outlier dengan nilai ekstrim positif.
2. Terlihat bahwa terdapat sejumlah kecil debitur dengan LIMIT_BAL yang lebih dari 600000, yaitu sebanyak 48 orang, dima na dari 48 orang tersebut terdapat 5 orang yang default.
3. Sebagian besar debitur memiliki limit kredit sebesar 200000 atau kurang, dan tampak dalam range tersebut terdapat ju mlah debitur default yang tinggi dibandingkan limit kredit lainnya.

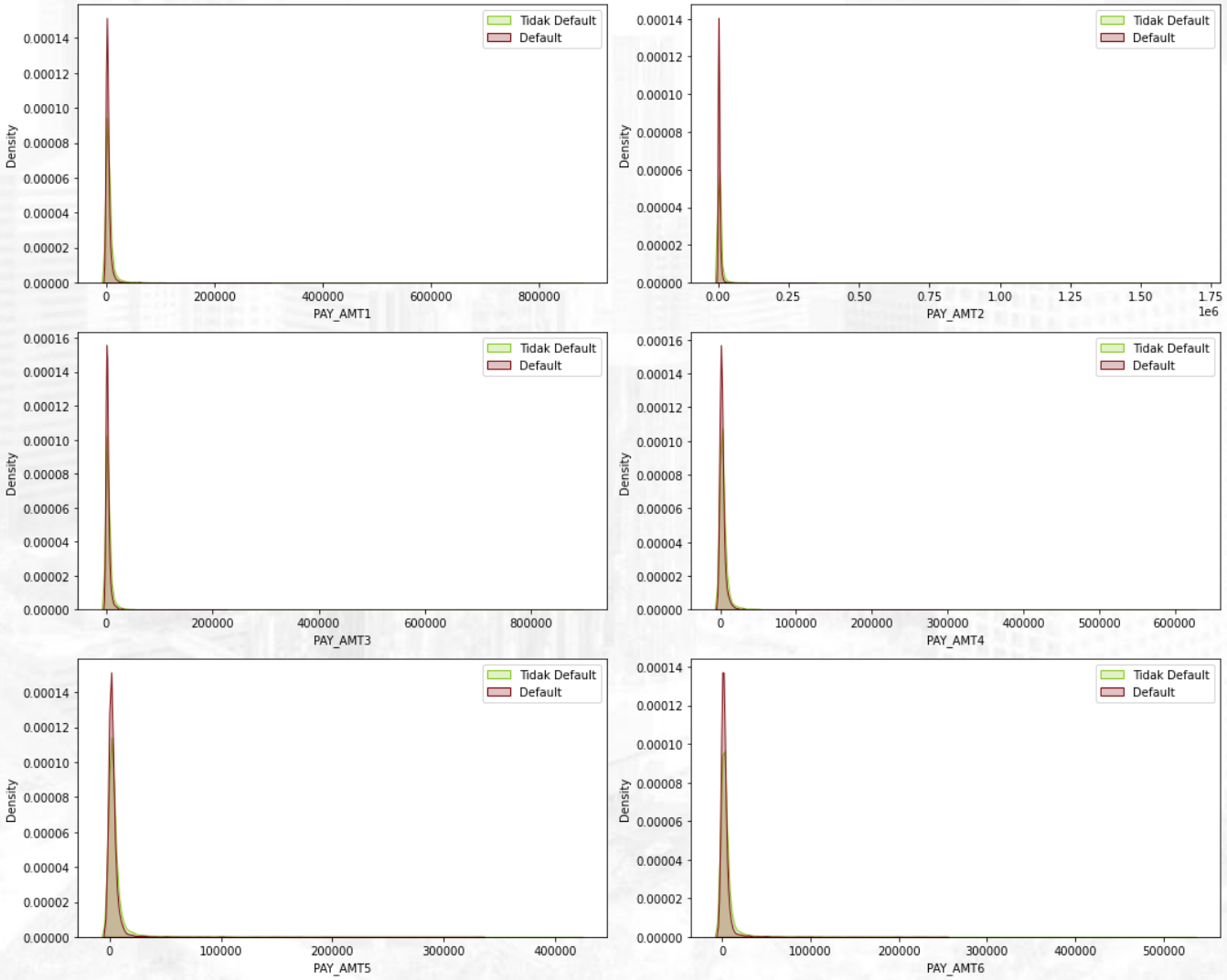
Eksplorasi Dataset



Pengamatan:

1. Distribusi Bill amount juga right skewed, terdapat beberapa pembayaran yang lebih dibulan sebelumnya

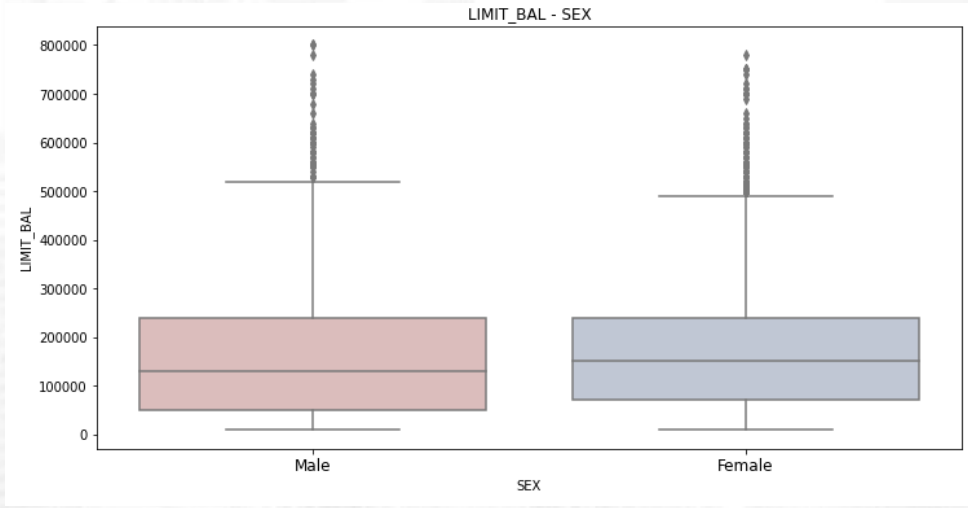
Eksplorasi Dataset

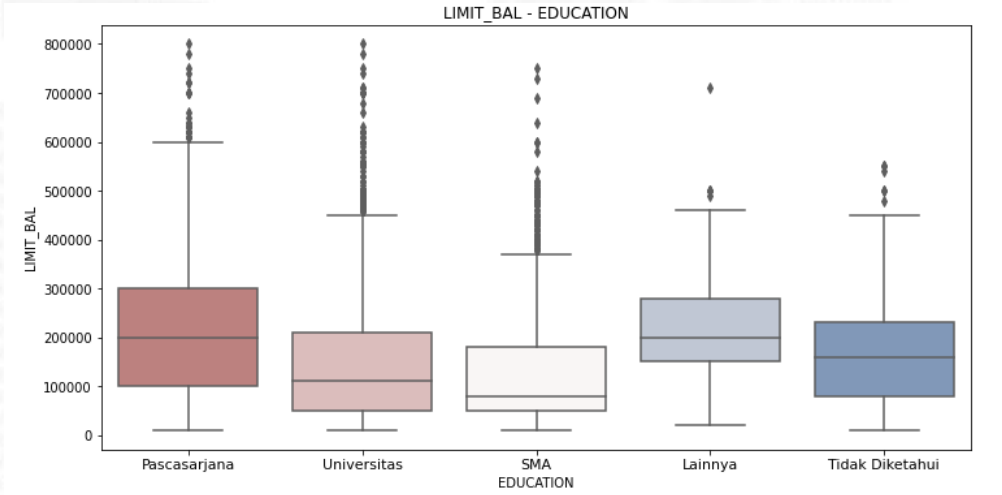


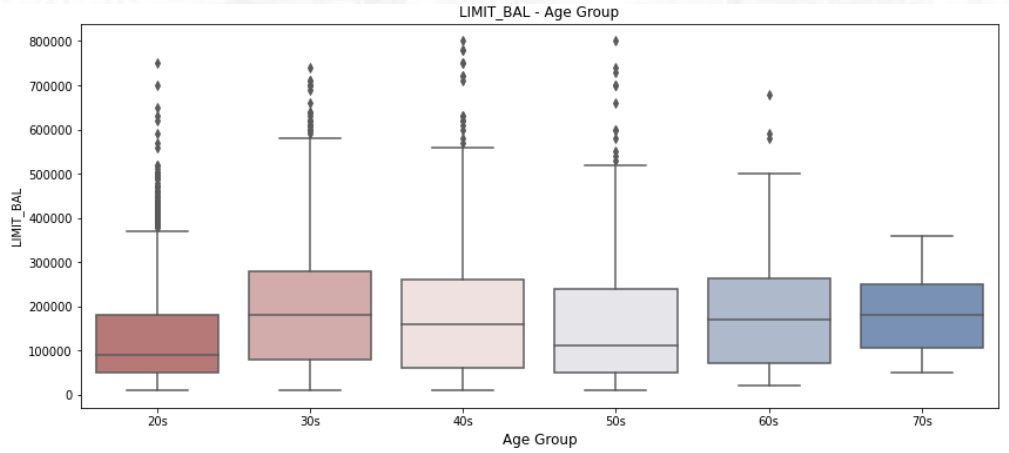
Pengamatan:

1. Terlihat bahwa untuk setiap PAY_AMT memiliki distribusi yang hampir sama yaitu right skewed dengan nilai ekstrem positif
2. Pay amount didominasi oleh pembayaran yang kurang dari 100000
3. Tampak juga bahwa semakin kecil nilai PAY_AMT maka peluang untuk default cendurung lebih besar terutama bagi debitur yang pembayarannya 0 (tidak membayar)

Eksplorasi Dataset







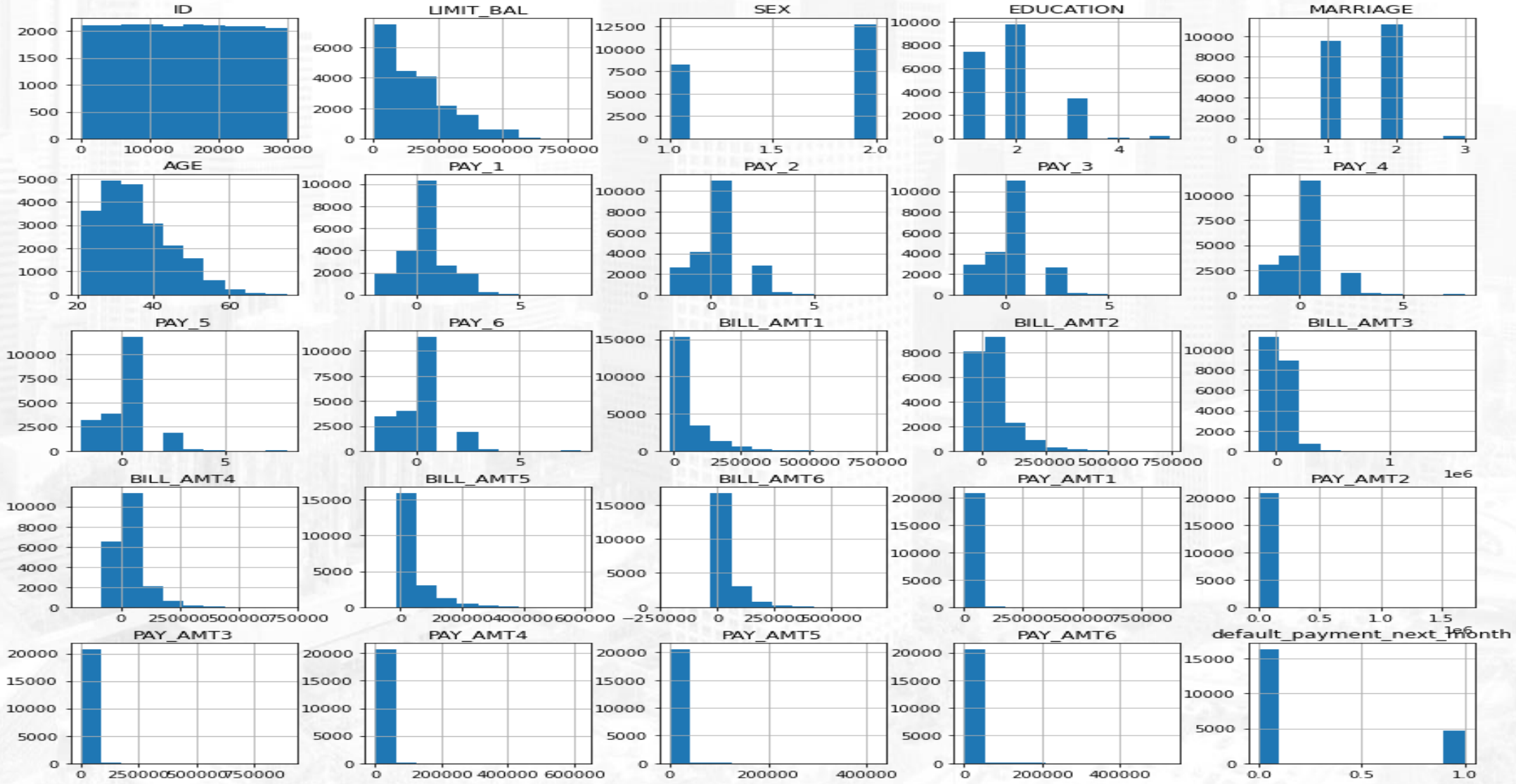
Eksplorasi Dataset

Pengamatan:

1. Secara keseluruhan dari grafik-grafik antara Limit credit dengan data demografis, terlihat bahwa sebagian besar pemberian limit kredit berbanding lurus dengan probabilitas default pada kategori tersebut, maksudnya adalah terihat bahwa semakin kecil peluang default yang ditunjukkan pada bagian univariate analysis untuk categorical data, maka limit kredit yang diberikan juga akan semakin besar

2. Namun terdapat pengecualian yaitu pada kategori usia 60 tahun keatas dimana walaupun memiliki peluang gagal bayar yang tinggi dibandingkan kategori usia lainnya, namun kategori ini tetap mendapatkan limit kredit yang tinggi

Eksplorasi Dataset



Eksplorasi Dataset

	default_payment_next_month	0	1
ID	mean	15026.017670	14678.651884
LIMIT_BAL	mean	178153.592174	128699.177610
SEX	mean	1.617365	1.573089
EDUCATION	mean	1.846408	1.885038
MARRIAGE	mean	1.556405	1.535199
AGE	mean	35.428921	35.576749
PAY_1	mean	-0.206237	0.675565
PAY_2	mean	-0.295628	0.465662
PAY_3	mean	-0.315561	0.365770
PAY_4	mean	-0.355671	0.265877
PAY_5	mean	-0.387955	0.186222
PAY_6	mean	-0.404647	0.119699
BILL_AMT1	mean	52616.872944	47574.474273
BILL_AMT2	mean	50324.411312	46432.251668
BILL_AMT3	mean	48078.405381	44254.426911
BILL_AMT4	mean	44076.679976	40960.327449
BILL_AMT5	mean	40906.776276	38609.095156
BILL_AMT6	mean	39419.521553	37211.797417
PAY_AMT1	mean	6369.907368	3279.548116
PAY_AMT2	mean	6679.975176	3257.713455
PAY_AMT3	mean	5816.066769	3041.347686
PAY_AMT4	mean	5284.227270	3064.171152
PAY_AMT5	mean	5315.105717	2972.813348
PAY_AMT6	mean	5789.031122	3179.088913

Berdasarkan grafik di slide sebelumnya dan table di sampling, disimpulkan bahwa

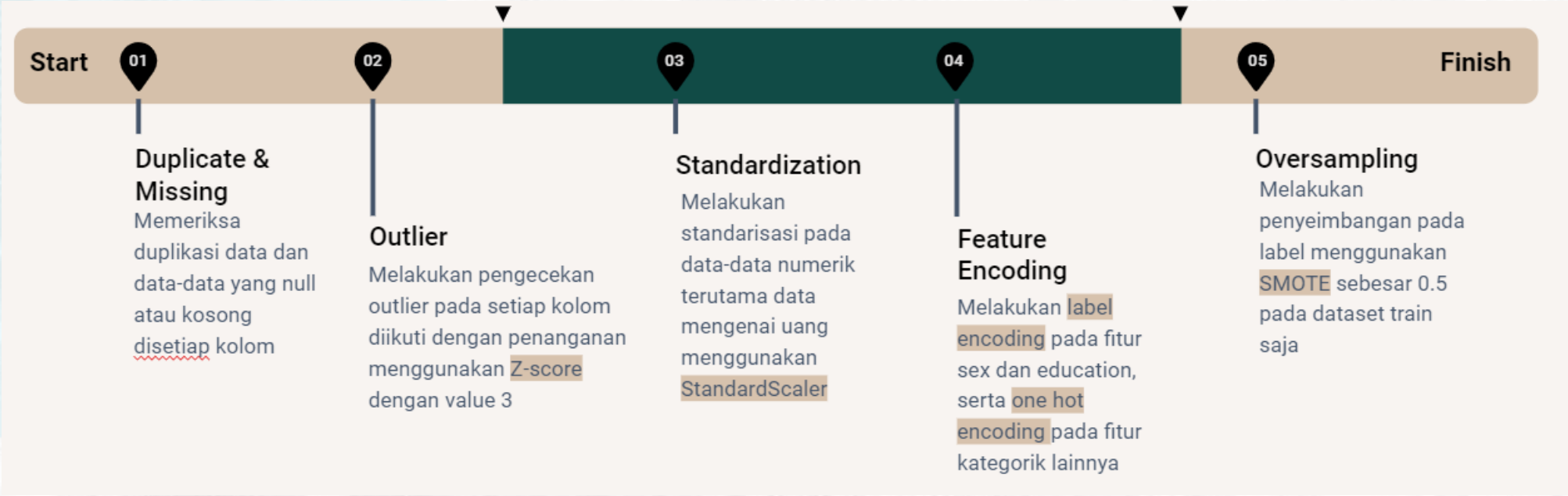
- Terdapat 4645 (22%) dari 21000 entries orang yang gagal bayar bulan depan. Sisanya tidak gagal bayar berjumlah 16355 (78%) d ari 21000
- Jumlah kredit yang diberikan (LIMIT_BAL) memiliki rata-rata sekitar 167.214
- Sebagian besar pelanggan adalah wanita
- Dari pendidikan, yang tertinggi adalah dari universitas, dan yang kedua adalah dari sekolah pascasarjana, dan yang ketiga adalah sekolah menengah
- Mayoritas sudah menikah, dan tertinggi kedua adalah single
- Usia rata-rata adalah sekitar 35 tahun
- Tedapat kecenderungan customer default memiliki buying power yang besar namun kemampuan bayarnya kecil, terlihat dari rata-rata pada PAY_AMT dan BILL_AMT disamping

Insights

- Data didominasi oleh jenis kelamin perempuan dengan persentase sebesar 61%, namun default didominasi oleh laki-laki dengan persentase sebesar 24%
- Pendidikan didominasi oleh universitas sebanyak 47% dan diikuti oleh pascasarjana sebanyak 35%, namun default didominasi oleh universitas dan SMA sebesar 24%
- Status pernikahan didominasi oleh lajang sebanyak 53%, namun default didominasi oleh status bercerai sebanyak 29%
- Customer didominasi oleh usia 30an yaitu usia antara 30 tahun – 39 tahun yaitu sebanyak 37,6% dari total customer yang ada, disusul oleh usia 20an dan kemudian 40an. Namun, jika dilihat dari persentase default, terlihat bahwa customer dengan usia lansia yaitu 70an dan 60an memiliki nilai yang lebih besar jika dibandingkan dengan customer pada kategori usia lainnya.
- Customer default dan tidak default memiliki rata-rata limit balance yang sesuai dengan kecenderungan default customer, dimana customer default diberikan limit yang lebih rendah. Namun jika dilihat dari *bill amount*/tagihannya, terlihat bahwa rata-rata tagihan customer default dan tidak hampir sama nilainya. Sementara itu, rata-rata *pay amount*/pembayaran terlihat jauh berbeda dimana kemampuan bayar customer default hanya ½ kali dari customer yang tidak default.

Pre-processing

Pada tahap ini dilakukan beberapa proses terhadap dataset sehingga dataset menjadi bersih ketika digunakan untuk melatih model. Berikut alur preprocessing yang dilakukan secara general sebelum proses modelling dilakukan,



Data Cleansing

Pertama kita cek apakah ada data yang kosong dan duplicate.

```
df_train_20D8GL3.isna().sum()
```

```
ID                          0
LIMIT_BAL                   0
SEX                         0
EDUCATION                   0
MARRIAGE                    0
AGE                         0
PAY_1                       0
PAY_2                       0
PAY_3                       0
PAY_4                       0
PAY_5                       0
PAY_6                       0
BILL_AMT1                   0
BILL_AMT2                   0
BILL_AMT3                   0
BILL_AMT4                   0
BILL_AMT5                   0
BILL_AMT6                   0
PAY_AMT1                    0
PAY_AMT2                    0
PAY_AMT3                    0
PAY_AMT4                    0
PAY_AMT5                    0
PAY_AMT6                    0
default_payment_next_month  0
AGE_BIN                     0
dtype: int64
```

```
df_train_20D8GL3.duplicated().sum()
```

```
0
```

- untuk mengecek data yang kosong, menggunakan df.isna()sum. Dari hasil tersebut ditemukan bahwa tidak ada data yang null.
- untuk mengecek data yang duplicate, menggunakan df.duplicates().sum(). Dari hasil tersebut ditemukan bahwa tidak ada data yang duplicate.

Outlier Handling

Kemudian setelah data cleansing, dilakukan handling outlier dengan menggunakan metode Z-score dengan value sebesar 3.

```
[ ]  print(f'Jumlah baris sebelum memfilter outlier: {len(df_train_20D8GL3)}')

     filtered_entries = np.array([True] * len(df_train_20D8GL3))

     for col in numericals:
         zscore = abs(stats.zscore(df_train_20D8GL3[col])) # hitung absolute z-scorenya
         filtered_entries = (zscore < 3) & filtered_entries # keep yang kurang dari 3 absolute z-scorenya

     df_train_20D8GL3 = df_train_20D8GL3[filtered_entries] # filter, cuma ambil yang z-scorenya dibawah 3

     print(f'Jumlah baris setelah memfilter outlier: {len(df_train_20D8GL3)}')
```

```
Jumlah baris sebelum memfilter outlier: 21000
Jumlah baris setelah memfilter outlier: 19027
```

Standardization

Setelah itu, standarisasi pada dataset menggunakan StandardScaler pada fitur-fitur yang berkaitan dengan nominal uang seperti limit bal, pay amount, dan bill amount.

```
[ ]  ftrs = ['LIMIT_BAL', 'BILL_AMT1','BILL_AMT2','BILL_AMT3','BILL_AMT4','BILL_AMT5','BILL_AMT6', 'PAY_AMT1','PAY_AMT2','PAY_AMT3','PAY_AMT4','PAY_AMT5','PAY_AMT6']
     for f in ftrs:
       df_train_20D8GL3[f] = StandardScaler().fit_transform(df_train_20D8GL3[f].values.reshape(len(df_train_20D8GL3), 1))

     df_train_20D8GL3.head()
```

Feature Encoding

Setelah itu, dilakukan label encoding pada kolom SEX dan EDUCATION dengan tujuan agar kolom SEX menjadi data biner saja dan kolom EDUCATION menjadi data ordinal dimana pascasarjana mendapatkan point terbesar, universitas dibawah pascasarjana, dst.

```
[ ]  # mengubah 1 -> 0 dan 2 -> 1, biar lebih general dan jadi biner
     mapping_gender = {
         1: 0,
         2: 1
     }

     mapping_education = {
         1: 5,
         2: 4,
         3: 3,
         4: 2,
         5: 1
     }

     df_train_20D8GL3['SEX'] = df_train_20D8GL3['SEX'].map(mapping_gender)
     df_train_20D8GL3['EDUCATION'] = df_train_20D8GL3['EDUCATION'].map(mapping_education)
     df_train_20D8GL3.head()
```

Kemudian, one hot encoding untuk kolom MARRIAGE, PAY_0 hingga PAY_6 dengan tujuan untuk melihat tingkatan antar data lebih representatif

```
[ ]  ftrs = ['MARRIAGE', 'PAY_1','PAY_2','PAY_3','PAY_4','PAY_5','PAY_6']

     for f in ftrs:
       onehots = pd.get_dummies(df_train_20D8GL3[f], prefix=f)
       df_train_20D8GL3 = df_train_20D8GL3.join(onehots)

     df_train_20D8GL3.info()
```

Drop Outdated Columns & Handling Data Bermasalah Lainnya

Kemudian, dilakukan drop kolom untuk kolom yang telah di encode atau tidak digunakan pada modelling,

yaitu kolom ID, AGE_BIN, PAY_0 hingga PAY-6

```
[ ]  df_train_20D8GL3 = df_train_20D8GL3.drop(columns=['ID', 'AGE_BIN', 'MARRIAGE','PAY_1','PAY_2','PAY_3','PAY_4','PAY_5','PAY_6'])
```

Proses handling untuk data bermasalah lainnya,

1. Merubah nama kolom `PAY_0` menjadi `PAY_1`, karena meilhat dari deskripsi kolom yang sudah berurutan

2. Mendefinisikan nilai-nilai yang belum terdefinisi pada kolom EDUCATION sebagai tidak diketahui (kategori 5), PAY_0 hingga PAY_6 pada nilai 0 sebagai tepat waktu, -1 kelebihan membayar selama 1 bulan, -2 sebagai kelebihan bayar selama 2 bulan

```
[ ]  df_train_20D8GL3.loc[:,'EDUCATION'] = df_train_20D8GL3.loc[:,'EDUCATION'].replace(0,5)
     df_train_20D8GL3.loc[:,'EDUCATION'] = df_train_20D8GL3.loc[:,'EDUCATION'].replace(6,5)
```

Oversampling

dilakukan oversampling menggunakan metode SMOTE dengan nilai sebesar 0.5 hanya pada dataset train saja.

```
[ ]  # balance dataset
     X = df_skenario4[[col for col in df_skenario4.columns if (str(df_skenario4[col].dtype) not in ['object', 'category']) and col not in ['default_payment_next_month']]]
     Y = df_skenario4[['default_payment_next_month']]
     X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size = 0.3, random_state = 42)
```

```
[ ]  # balancing train dataset
     x_balance, y_balance = over_sampling.SMOTE(0.5).fit_resample(X_train.values,y_train.values)
```

```
[ ]  print(y_train['default_payment_next_month'].value_counts())
     print(pd.Series(y_balance).value_counts())
```

```
0    10229
1     3089
Name: default_payment_next_month, dtype: int64
0    10229
1     5114
dtype: int64
```

Modelling Experiments

Secara general dalam tahap pemodelan dilakukan beberapa eksperimen terhadap dataset, algoritma model, dan fitur-fitur yang digunakan dalam model.

Dalam tahap pemodelan akan dibuat 4 skenario yaitu:

- menggunakan dataset original
- menggunakan dataset setelah outlier dihapus
- menggunakan dataset yang distandarisasi
- menggunakan dataset setelah outlier dihapus, distandarisasi, dan adanya one hot encoding

Pemodelan menggunakan 6 algoritma yaitu:

- Logistic Regression
- KNN
- Decision Tree
- Random Forest
- AdaBoost
- XGBoost

Pemdelan menggunakan kombinasi fitur yaitu:

- Menggunakan semua fitur
- Fitur berkorelasi cukup kuat pada heatmap yaitu LIMIT_BAL, PAY_1 hingga PAY-6
- Fitur PAY_AMT dan BILL_AMT saja karena nilai rata-rata antara yang default dan tidak cukup besar
- Fitur hasil SelectKBest (f_classif) yaitu LIMIT_BAL , SEX, EDUCATION, MARRIAGE, AGE, PAY_1 hingga PAY_6

Kondisi lainnya:

- Data di oversampling
- Data Imbalance
- Tuning Hyperparameters

Modelling Experiments

Skenario 1: Dataset original

Algoritma yang digunakan:
- Logistic Regression
- KNN
- Decision Tree
- Random Forest
- AdaBoost
- XGBoost

Feature yang digunakan:
- Semua feature

Algoritma	Original										Tuning Hyperparameters									
	Accuracy		Precision		Recall		AUC		F1		Accuracy		Precision		Recall		AUC		F1	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Logistic Regression	0.78	0.78	0.5	1	0	0	0.67	0.66	0	0	0.78	0.78	0.33	0	0	0	0.67	0.65	0	0
KNN	0.82	0.74	0.67	0.35	0.34	0.18	0.83	0.6	0.45	0.24	0.79	0.78	0.62	0.51	0.07	0.06	0.72	0.65	0.12	0.11
Decision Tree	0.99	0.69	0.99	0.33	0.97	0.35	1	0.58	0.98	0.34	0.81	0.76	0.64	0.43	0.29	0.19	0.82	0.65	0.4	0.26
Random Forest	0.99	0.78	0.99	0.52	0.98	0.2	1	0.7	0.98	0.29	0.82	0.78	0.9	0.56	0.23	0.12	0.86	0.72	0.37	0.19
Ada Boost	0.79	0.78	0.61	0.53	0.16	0.13	0.75	0.71	0.26	0.21	0.79	0.78	0.64	0.58	0.15	0.12	0.76	0.72	0.24	0.19
XGBoost	0.8	0.78	0.66	0.56	0.19	0.14	0.78	0.72	0.29	0.23	0.85	0.78	0.85	0.55	0.39	0.2	0.91	0.71	0.54	0.29

* Tidak dapat menjelaskan secara terperinci mengenai tuning hyperparameters karena eksperimen yang dilakukan cukup banyak dan parameter yang dituning juga berbeda

Modelling Experiments

Skenario 2: Dataset setelah outlier dihapus

Algoritma yang digunakan:
- Logistic Regression
- KNN
- Decision Tree
- Random Forest

Feature yang digunakan:
- LIMIT_BAL, PAY_X, BILL_AMTX, PAY_AMTX

Metrik Evaluasi Data Imbalance

Algoritma	Original										Tuning Hyperparameters									
	Accuracy		Precision		Recall		AUC		F1		Accuracy		Precision		Recall		AUC		F1	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Logistic Regression	0.77	0.77	0.5	0.5	0	0	0.67	0.66	0	0										
KNN	0.81	0.74	0.67	0.36	0.35	0.2	0.83	0.6	0.46	0.26										
Decision Tree	1	0.7	1	0.37	1	0.43	1	0.61	1	0.39										
Random Forest	1	0.81	1	0.64	1	0.37	1	0.76	1	0.47	1	0.81	1	0.64	1	0.37	1	0.76	1	0.47

Metrik Evaluasi Data Setelah Oversampling

Algoritma	Original										Tuning Hyperparameters									
	Accuracy		Precision		Recall		AUC		F1		Accuracy		Precision		Recall		AUC		F1	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Logistic Regression	0.62	0.61	0.62	0.32	0.63	0.61	0.66	0.65	0.63	0.42										
KNN	0.84	0.58	0.77	0.28	0.95	0.54	0.95	0.59	0.85	0.37										
Decision Tree	1	0.67	1	0.34	1	0.48	1	0.61	1	0.4										
Random Forest	1	0.79	1	0.53	1	0.5	1	0.76	1	0.51	1	0.79	1	0.53	1	0.5	1	0.76	1	0.51

* Tidak dapat menjelaskan secara terperinci mengenai tuning hyperparameters karena eksperimen yang dilakukan cukup banyak dan parameter yang dituning juga berbeda

Modelling Experiments

Skenario 3: Dataset yang distandarisasi

Algoritma yang digunakan:
- Logistic Regression
- KNN
- Decision Tree

Feature yang digunakan:
- Semua data numerik

Algoritma	Original										Tuning Hyperparameters									
	Accuracy		Precision		Recall		AUC		F1		Accuracy		Precision		Recall		AUC		F1	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Logistic Regression	0.77	0.77	0.38	0	0	0	0.67	0.67	0	0	0.77	0.77	0.38	0	0	0	0.67	0.67	0	0
KNN	0.81	0.75	0.65	0.42	0.41	0.25	0.84	0.64	0.5	0.31	0.77	0.78	0.68	0.74	0.06	0.04	0.72	0.67	0.1	0.08
Decision Tree	0.99	0.7	0.99	0.33	0.95	0.34	1	0.58	0.97	0.34	0.79	0.78	0.6	0.52	0.26	0.22	0.77	0.69	0.36	0.31

* Tidak dapat menjelaskan secara terperinci mengenai tuning hyperparameters karena eksperimen yang dilakukan cukup banyak dan parameter yang dituning juga berbeda

Modelling Experiments

Skenario 4: Dataset setelah outlier dihapus, distandarisasi, dan adanya one hot encoding

Algoritma yang digunakan:
- Logistic Regression
- KNN
- Decision Tree
- Random Forest
- AdaBoost
- XGBoost

Feature yang digunakan:
- Semua feature
- Berdasarkan nilai correlation heatmap
- PAY_AMTX dan BILL_AMTX only
- SelectKBest Features (sklearn library)

* Tidak dapat menjelaskan secara terperinci mengenai tuning hyperparameters karena eksperimen yang dilakukan cukup banyak dan parameter yang dituning juga berbeda

Modelling Experiments

Skenario 4: Dataset setelah outlier dihapus, distandarisasi, dan adanya one hot encoding

Metrik Evaluasi Data Imbalance

Algoritma	Features	Imbalance										Tuning Imbalance									
		Accuracy		Precision		Recall		AUC		F1		Accuracy		Precision		Recall		AUC		F1	
		Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Logistic Regression	All	0.81	0.82	0.68	0.69	0.37	0.33	0.78	0.77	0.48	0.45	0.81	0.81	0.68	0.69	0.36	0.33	0.78	0.77	0.47	0.45
	Corr Heatmap	0.81	0.82	0.68	0.67	0.36	0.37	0.77	0.76	0.47	0.47	0.81	0.82	0.68	0.67	0.36	0.37	0.77	0.76	0.47	0.48
	PAY_AMT & BILL_AMT only	0.77	0.77	0.38	0	0	0	0.67	0.67	0	0	0.77	0.77	0.38	0	0	0	0.67	0.67	0	0
	SelectKBest	0.81	0.82	0.68	0.67	0.37	0.37	0.77	0.76	0.48	0.48	0.81	0.82	0.68	0.67	0.36	0.37	0.77	0.76	0.47	0.47
KNN	All	0.83	0.78	0.74	0.52	0.43	0.29	0.87	0.69	0.54	0.37	0.8	0.79	0.69	0.71	0.22	0.15	0.78	0.73	0.33	0.25
	Corr Heatmap	0.83	0.79	0.7	0.54	0.46	0.36	0.79	0.69	0.55	0.43	0.81	0.81	0.68	0.66	0.35	0.34	0.78	0.75	0.46	0.45
	PAY_AMT & BILL_AMT only	0.81	0.75	0.65	0.42	0.41	0.25	0.84	0.64	0.5	0.31	0.77	0.78	0.75	0.89	0.02	0.01	0.69	0.68	0.04	0.03
	SelectKBest	0.84	0.79	0.74	0.54	0.45	0.33	0.88	0.7	0.56	0.41	0.8	0.8	0.69	0.65	0.26	0.25	0.78	0.75	0.38	0.36
Decision Tree	All	1	0.69	1	0.35	1	0.44	1	0.6	1	0.39	0.8	0.8	0.65	0.59	0.31	0.36	0.76	0.71	0.41	0.45
	Corr Heatmap	0.87	0.79	0.87	0.54	0.54	0.35	0.9	0.66	0.67	0.42	0.81	0.81	0.69	0.67	0.35	0.34	0.77	0.75	0.47	0.45
	PAY_AMT & BILL_AMT only	0.99	0.7	0.99	0.33	0.95	0.34	1	0.58	0.97	0.34	0.79	0.77	0.61	0.5	0.22	0.16	0.76	0.66	0.32	0.24
	SelectKBest	0.98	0.72	1	0.39	0.91	0.41	1	0.61	0.95	0.4	0.81	0.8	0.67	0.62	0.38	0.35	0.78	0.73	0.49	0.45
Random Forest	All	1	0.81	1	0.68	1	0.33	1	0.76	1	0.44	0.84	0.81	0.79	0.7	0.41	0.31	0.9	0.78	0.54	0.43
	Corr Heatmap	0.87	0.79	0.83	0.54	0.58	0.37	0.89	0.71	0.68	0.44	0.82	0.82	0.7	0.68	0.37	0.36	0.8	0.76	0.48	0.47
	PAY_AMT & BILL_AMT only	0.99	0.78	0.99	0.52	0.95	0.2	1	0.7	0.97	0.29	0.82	0.78	0.89	0.61	0.24	0.13	0.85	0.72	0.38	0.21
	SelectKBest	0.98	0.79	0.99	0.54	0.92	0.4	1	0.73	0.95	0.46	0.82	0.82	0.71	0.68	0.37	0.35	0.81	0.77	0.49	0.46
Ada Boost	All	0.81	0.81	0.69	0.67	0.35	0.32	0.79	0.76	0.47	0.43	0.81	0.81	0.69	0.67	0.34	0.34	0.77	0.77	0.45	0.45
	Corr Heatmap	0.81	0.81	0.68	0.66	0.37	0.37	0.77	0.76	0.48	0.47	0.81	0.81	0.69	0.66	0.34	0.34	0.77	0.76	0.45	0.45
	PAY_AMT & BILL_AMT only	0.78	0.78	0.61	0.63	0.13	0.12	0.73	0.71	0.21	0.2	0.78	0.78	0.62	0.62	0.11	0.1	0.74	0.72	0.19	0.17
	SelectKBest	0.81	0.82	0.69	0.67	0.36	0.37	0.77	0.76	0.47	0.48	0.81	0.81	0.69	0.67	0.34	0.34	0.77	0.77	0.45	0.45
XGBoost	All	0.82	0.82	0.71	0.67	0.38	0.37	0.79	0.77	0.49	0.48	0.86	0.81	0.82	0.64	0.49	0.38	0.91	0.78	0.62	0.47
	Corr Heatmap	0.82	0.82	0.69	0.66	0.38	0.37	0.78	0.77	0.49	0.48	0.82	0.82	0.7	0.67	0.38	0.37	0.79	0.77	0.49	0.48
	PAY_AMT & BILL_AMT only	0.79	0.79	0.65	0.6	0.18	0.15	0.77	0.72	0.28	0.24	0.85	0.78	0.87	0.53	0.42	0.22	0.91	0.71	0.57	0.31
	SelectKBest	0.82	0.82	0.71	0.67	0.38	0.37	0.79	0.77	0.49	0.48	0.81	0.81	0.7	0.67	0.33	0.32	0.79	0.77	0.45	0.43

* Tidak dapat menjelaskan secara terperinci mengenai tuning hyperparameters karena eksperimen yang dilakukan cukup banyak dan parameter yang dituning juga berbeda

Modelling Experiments

Skenario 4: Dataset setelah outlier dihapus, distandarisasi, dan adanya one hot encoding

Metrik Evaluasi Data Setelah Oversampling

Algoritma	Features	Balance										Tuning Balance									
		Accuracy		Precision		Recall		AUC		F1		Accuracy		Precision		Recall		AUC		F1	
		Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Logistic Regression	All	0.76	0.81	0.73	0.62	0.45	0.45	0.79	0.77	0.55	0.52	0.76	0.81	0.73	0.62	0.43	0.44	0.77	0.76	0.54	0.51
	Corr Heatmap	0.76	0.81	0.73	0.61	0.44	0.44	0.77	0.76	0.55	0.51	0.76	0.81	0.73	0.62	0.44	0.44	0.77	0.76	0.55	0.51
	PAY_AMT & BILL_AMT only	0.67	0.77	0.52	0.25	0	0	0.68	0.67	0.01	0	0.67	0.77	0.53	0.29	0	0	0.68	0.67	0.01	0
	SelectKBest	0.76	0.81	0.72	0.61	0.43	0.44	0.76	0.76	0.54	0.51	0.76	0.81	0.73	0.62	0.43	0.44	0.77	0.76	0.54	0.51
KNN	All	0.84	0.73	0.75	0.41	0.79	0.48	0.92	0.7	0.77	0.44	1	0.77	1	0.48	1	0.44	1	0.72	1	0.46
	Corr Heatmap	0.79	0.76	0.73	0.47	0.58	0.43	0.82	0.69	0.64	0.45	0.77	0.8	0.73	0.58	0.47	0.45	0.79	0.75	0.57	0.51
	PAY_AMT & BILL_AMT only	0.82	0.7	0.74	0.35	0.71	0.36	0.9	0.63	0.73	0.35	1	0.78	1	0.56	1	0.14	1	0.68	1	0.22
	SelectKBest	0.84	0.75	0.76	0.44	0.74	0.48	0.91	0.7	0.75	0.46	0.98	0.77	1	0.5	0.95	0.42	1	0.7	0.97	0.46
Decision Tree	All	1	0.71	1	0.37	1	0.44	1	0.61	1	0.4	0.79	0.78	0.75	0.51	0.57	0.42	0.85	0.73	0.65	0.46
	Corr Heatmap	0.86	0.78	0.9	0.52	0.65	0.38	0.91	0.66	0.75	0.44	0.76	0.81	0.74	0.6	0.45	0.43	0.79	0.75	0.56	0.5
	PAY_AMT & BILL_AMT only	0.98	0.66	0.95	0.31	1	0.4	1	0.57	0.97	0.35	0.75	0.72	0.66	0.38	0.52	0.35	0.81	0.67	0.58	0.36
	SelectKBest	0.98	0.72	1	0.39	0.95	0.42	1	0.61	0.97	0.41	0.81	0.78	0.79	0.53	0.59	0.41	0.87	0.72	0.68	0.46
Random Forest	All	1	0.81	1	0.61	1	0.4	1	0.76	1	0.49	1	0.81	1	0.61	0.99	0.4	1	0.76	0.99	0.49
	Corr Heatmap	0.86	0.78	0.87	0.52	0.67	0.41	0.9	0.71	0.76	0.46	0.79	0.81	0.78	0.59	0.52	0.45	0.83	0.76	0.62	0.51
	PAY_AMT & BILL_AMT only	0.98	0.77	0.95	0.48	1	0.33	1	0.69	0.97	0.39	0.98	0.77	0.95	0.48	0.99	0.32	1	0.7	0.97	0.39
	SelectKBest	0.98	0.78	0.99	0.53	0.94	0.41	1	0.73	0.97	0.46	0.95	0.8	0.97	0.57	0.86	0.42	0.99	0.74	0.92	0.48
Ada Boost	All	0.8	0.81	0.8	0.64	0.53	0.39	0.84	0.76	0.64	0.48	0.8	0.82	0.8	0.66	0.52	0.4	0.84	0.77	0.63	0.5
	Corr Heatmap	0.76	0.81	0.74	0.62	0.45	0.43	0.77	0.76	0.56	0.51	0.76	0.81	0.74	0.61	0.45	0.43	0.78	0.76	0.56	0.5
	PAY_AMT & BILL_AMT only	0.71	0.77	0.66	0.5	0.29	0.27	0.73	0.7	0.41	0.35	0.71	0.78	0.68	0.53	0.27	0.24	0.74	0.71	0.38	0.33
	SelectKBest	0.79	0.81	0.79	0.64	0.5	0.4	0.81	0.76	0.62	0.49	0.79	0.82	0.79	0.65	0.5	0.41	0.82	0.77	0.61	0.5
XGBoost	All	0.82	0.82	0.84	0.66	0.56	0.39	0.87	0.78	0.68	0.49	1	0.8	1	0.57	1	0.4	1	0.75	1	0.47
	Corr Heatmap	0.77	0.81	0.75	0.62	0.47	0.43	0.79	0.77	0.58	0.51	0.82	0.8	0.81	0.57	0.6	0.43	0.86	0.74	0.69	0.49
	PAY_AMT & BILL_AMT only	0.73	0.78	0.71	0.53	0.34	0.29	0.77	0.71	0.46	0.37	0.98	0.75	0.95	0.43	0.99	0.35	1	0.68	1	0.68
	SelectKBest	0.82	0.82	0.84	0.66	0.55	0.38	0.84	0.77	0.67	0.48	0.83	0.81	0.86	0.67	0.58	0.34	0.87	0.77	0.69	0.46

* Tidak dapat menjelaskan secara terperinci mengenai tuning hyperparameters karena eksperimen yang dilakukan cukup banyak dan parameter yang dituning juga berbeda

Modelling Experiments

Terdapat lebih dari 100 eksperimen yang dilakukan dari skenario 1 -4 termasuk eksperimen dengan melakukan tuning hyperparameter, setelah dianalisis secara menyeluruh pengaruh tuning hyperparameter pada model dapat disimpulkan bahwa penggunaan tuning dapat membuat model menjadi tidak terlalu overfit maupun underfit, namun pengaruhnya tidak terlalu signifikan antara score evaluasi hasil tuning dengan tidak.

Berikut adalah 4 nilai metrik evaluasi dengan AUC terbesar

Algoritma	Features	Condition	Accuracy	Precision	Recall	AUC	F1
XGBoost	All features	Imbalance	0.82	0.67	0.37	0.77	0.48
AdaBoost		Imbalance & Tuning Hyperparameters	0.81	0.67	0.34	0.77	0.45
XGBoost		Oversampling	0.82	0.66	0.39	0.78	0.49
	SelectKBest	Oversampling & Tuning Hyperparameters	0.81	0.67	0.34	0.77	0.46

Best model yaitu XGBoost dengan semua fitur dan telah di oversampling, dimana model tersebut memiliki nilai AUC dan Recall tertinggi.

Evaluation Metric

Evaluation metric yang digunakan adalah AUC sebagai primary metric, recall & precision sebagai secondary metric dengan alasan sebagai berikut,

- Technical Reasons
  - Dataset imbalance dengan proprosi hampir mencapai 8:2
  - Berfokus untuk mengurangi jumlah false negative dan false positive
  - Nilai accuracy menjadi kurang representatif karena banyak data sintetis hasil dari oversampling

- Business Side Reasons

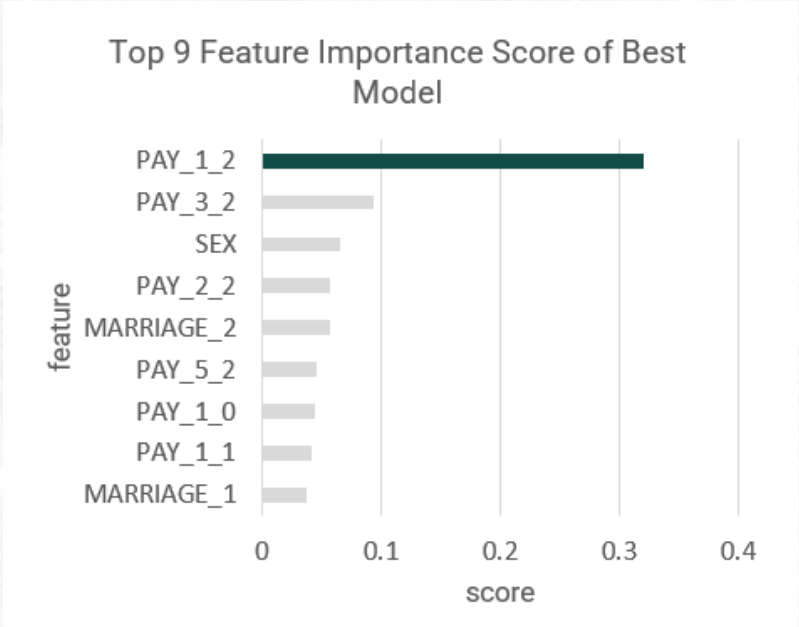
1.Tujuan utama dari model adalah untuk memprediksi jumlah default (yang benar-benar default) sebanyak-banyaknya sehingga dapat diberikan tindakan penanganan, sehingga diperlukan memperhatikan recall score

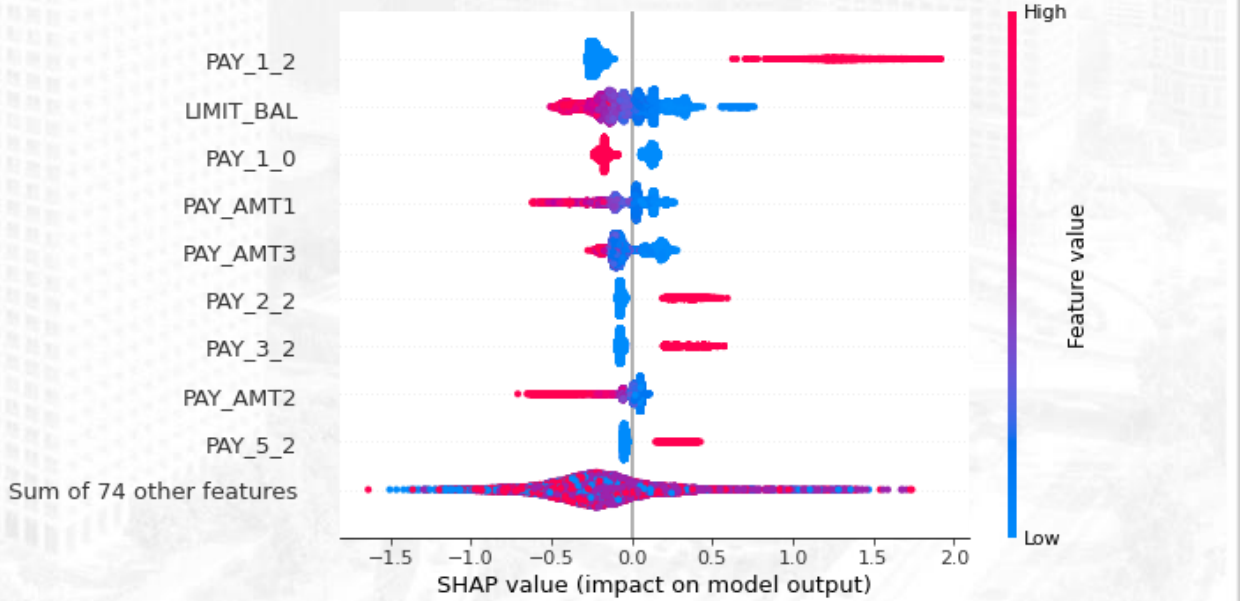
2. Namun, kita juga harus memperhatikan debitur yang seharusnya tidak default namun terprediksi default (false positive), ka rena jika terprediksi default namun tidak, maka debitur akan diberikan tindakan penanganan sebagai seorang default, hal terse but tentunya dapat memicu ketidaknyamanan debitur tersebut yang dapat menyebabkan komplain atau churn, sehingga score precision juga perlu diperhatikan

Feature Importance Best Model

9 Features importance tertinggi dari model terbaik berdasarkan SHAP value dan feature importance score didominasi fitur status pembayaran, sehingga dapat disimpulkan bahwa status pembayaran merupakan fitur yang berkontribusi cukup besar dalam menentukan customer akan default atau tidak di bulan berikutnya.

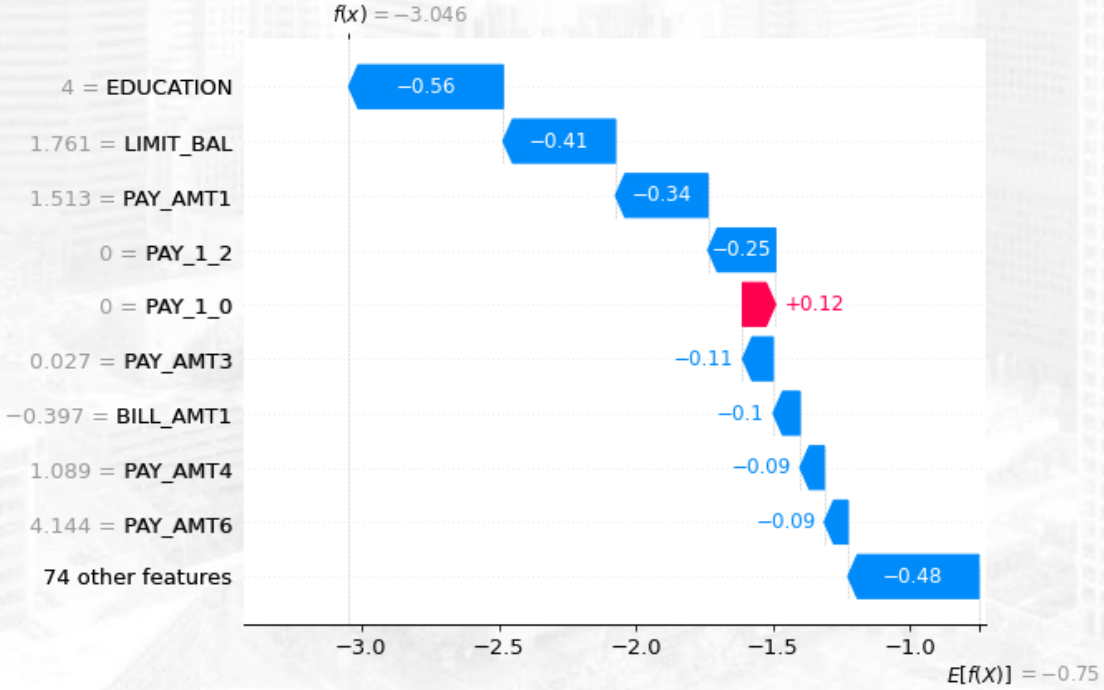
Feature importance tertinggi pada grafik feature importance score yaitu **PAY_1_2** yang merupakan status pembayaran di bulan September, dimana customer gagal bayar selama 2 bulan berturut-turut. Fitur ini nantinya dapat digunakan dalam penentuan customer prioritas yang akan dimitigasi untuk menurunkan default rate pada bulan berikutnya.



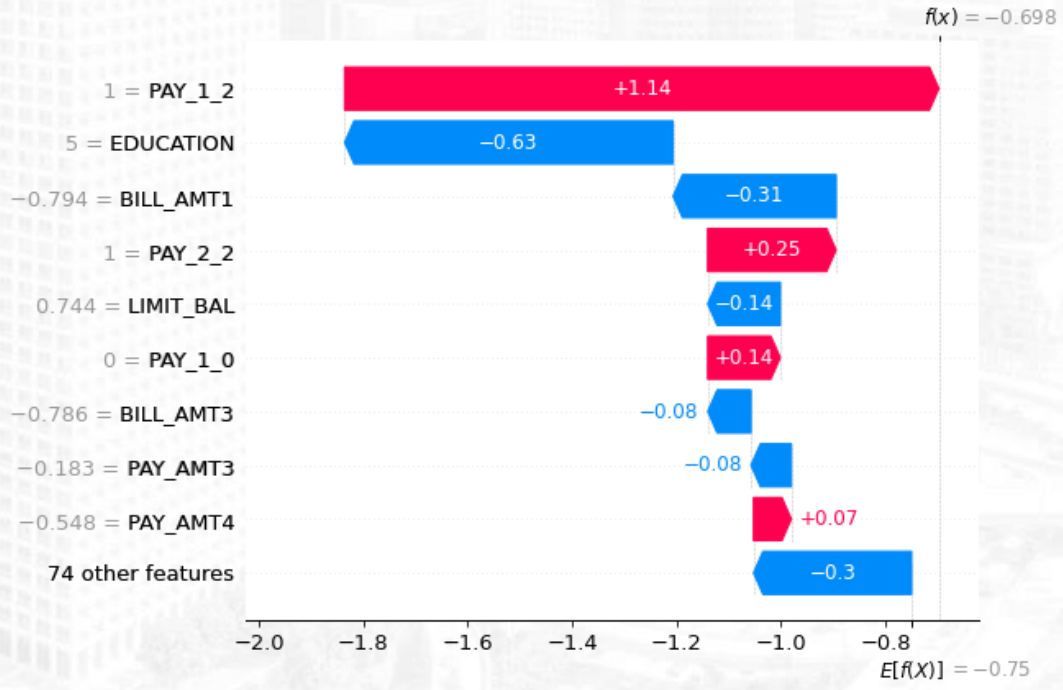


Contoh SHAP Value Customer

Grafik di atas adalah salah satu contoh SHAP value untuk customer yang tidak default.



Grafik di atas adalah salah satu contoh SHAP value untuk customer yang default, terlihat bahwa PAY_1_2 memberikan kontribusi yang besar dalam penentuan customer tersebut default.



Iterasi Feature Importances ke Best Model

metric evaluation yang dihasilkan oleh model yang menggunakan 9 features dengan kepentingan tertinggi tidak memberi kan score yang lebih bagus dibandingkan dengan model sebelumnya.

```
[ ]  # balance dataset
     X = df_skenario4[['PAY_1_2', 'PAY_3_2', 'SEX', 'PAY_2_2', 'MARRIAGE_2', 'PAY_5_2', 'PAY_1_0', 'PAY_1_1', 'MARRIAGE_1', 'PAY_6_2', 'EDUCATION']]
     Y = df_skenario4[['default_payment_next_month']]
     X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size = 0.3, random_state = 42)
```

```
[ ]  # balancing train dataset
     x_balance, y_balance = over_sampling.SMOTE(0.5).fit_resample(X_train.values,y_train.values)
```

```
[ ]  print(y_train['default_payment_next_month'].value_counts())
     print(pd.Series(y_balance).value_counts())
```

```
0    10229
1     3089
Name: default_payment_next_month, dtype: int64
0    10229
1     5114
dtype: int64
```

```
[ ]  XGB = XGBClassifier(random_state=42)
     XGB.fit(x_balance, y_balance)
     eval_classification(XGB, x_balance, y_balance, X_test.values, y_test.values)
     # show_feature_importance(XGB)
```

```
Accuracy (Test Set): 0.81
Precision (Test Set): 0.60
Recall (Test Set): 0.44
F1-Score (Test Set): 0.51
AUC (Test Set): 0.75

Accuracy (Train Set): 0.75
Precision (Train Set): 0.72
Recall (Train Set): 0.43
F1-Score (Train Set): 0.54
AUC (Train Set): 0.76
```

Rekomendasi & Simulasi

Berikut adalah strategi dan simulasi terhadap rekomendasi yang diberikan,

21000
TOTAL CUSTOMER

39%
RECALL MODEL

Rekomendasi Bisnis

1. Customer default dapat diberikan penawaran berupa,
    - Mengubah struktur pembiayaan seperti besar tagihan, bunga pinjaman, besar limit balance, dsb
    - Mengubah jadwal pembayaran tanpa mengubah pokok pinjaman secara signifikan

2. Proses mitigasi customer default dapat dimulai dari customer yang memiliki status pembayaran yang terlambat selama 2 bulan pada September 2005

Pembagian Tugas

Stage 0: Dikerjakan bersama, setiap individu memberikan hasil pemikiran masing-masing

Stage 1: setiap individu memberikan hasil pengerjaan masing-masing dan kemudian dikumpulkan menjadi satu

Stage 2: setiap individu memberikan hasil pengerjaan masing-masing dan kemudian dikumpulkan menjadi satu

Stage 4: scenario 1 oleh Steven Benny, scenario 2 oleh Cristanto, scenario 3 oleh Tri Setiawan, scenario 4 oleh Ulva Dewiyanti

Laporan Mentoring: Steven Benny

Ketua Tim : Tri Setiawan

Laporan Progress Mingguan: Bersama

Laporan & PPT Final : Ulva Dewiyanti