

Credit Risk Prediction

Project Based Virtual Internship

Presented by
Clavino Ourizqi Rachmadi

ID/X Partners - Data Scientist



 Depok, West Java

 clavinorach@gmail.com

 linkedin.com/in/clavinorachmadi

Clavino Ourizqi Rachmadi

Data Scientist

I'm a Motivated Computer Science student at Mercu Buana University with a strong foundation and deep knowledge in Information Technology, specializing in Web Development and Artificial Intelligence. Dedicated to mastering web technologies and AI, I excel in dynamic environments, demonstrating adaptability and a strong capacity for rapid learning. Eager to explore emerging technologies and contribute to cutting-edge projects in Web Development and AI.

Courses and Certification

Alibaba Cloud Academy Big Data | [link certificate](#)

August, 2024

Oracle Database Programming with SQL | [link certificate](#)

April, 2024

Oracle Database Design | [Link Certificate](#)

March, 2024

Alibaba Cloud Academy Cloud Computing

October, 2023

Alibaba Cloud Certified Developer -1

October, 2023

[Visit My Portofolio Here!](#)

About Company

id/x partners was established in 2002 by ex-bankers and management consultants who have vast experiences in credit cycle and process management, scoring development, and performance management. Our combined experience has served corporations across Asia and Australia regions and in multiple industries, specifically financial services, telecommunications, manufacturing and retail.



id/x partners provides consulting services that specializes in utilizing data analytic and decisioning (DAD) solutions combined with an integrated risk management and marketing discipline to help clients optimize the portfolio profitability and business process.

Comprehensive consulting service and technology solutions offered by **id/x partners** makes it as a one-stop service provider.

Business Understanding

Proyek **Credit Risk Prediction** bertujuan meningkatkan akurasi penilaian risiko kredit di industri multifinance, yang menghadapi tantangan dari kegagalan peminjam memenuhi kewajibannya. Perusahaan ingin mengoptimalkan pengambilan keputusan pinjaman dan meminimalkan kerugian melalui model machine learning yang efektif. Dataset yang digunakan mencakup data historis pinjaman (disetujui dan ditolak) dengan fitur penting seperti jumlah pinjaman, suku bunga, durasi, riwayat kredit, dan tanggal-tanggal penting. Tantangan utama adalah mengembangkan model yang dapat mengidentifikasi peminjam berisiko tinggi dengan tepat, menggunakan pendekatan eksplorasi data, visualisasi, penyeimbangan kelas (SMOTE), serta algoritma seperti Logistic Regression, Random Forest, dan XGBoost.

Link code [here!](#)

Project explanation video [here!](#)

Data Understanding

df.head()															
Unnamed: 0		id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	grade	...	total_bal_il	il_util	open_rv_12m	open_rv_24m
0	0	1077501	1296599	5000	5000	4975.0	36 months	10.65	162.87	B	...	NaN	NaN	NaN	N
1	1	1077430	1314167	2500	2500	2500.0	60 months	15.27	59.83	C	...	NaN	NaN	NaN	N
2	2	1077175	1313524	2400	2400	2400.0	36 months	15.96	84.33	C	...	NaN	NaN	NaN	N
3	3	1076863	1277178	10000	10000	10000.0	36 months	13.49	339.31	C	...	NaN	NaN	NaN	N
4	4	1075358	1311748	3000	3000	3000.0	60 months	12.69	67.79	B	...	NaN	NaN	NaN	N

5 rows x 75 columns

Preview lima baris pertama mencakup **informasi dataset**

[] #Meliiat dimensi pada dataset
df.shape

→ (466285, 75)

Pada dataset, terdapat **466285 baris data**

dengan **75 fitur / kolom**

#	Column	Non-Null Count	Dtype
0	Unnamed: 0	466285	non-null int64
1	id	466285	non-null int64
2	member_id	466285	non-null int64
3	loan_amnt	466285	non-null int64
4	funded_amnt	466285	non-null float64
5	funded_amnt_inv	466285	non-null object
6	term	466285	non-null float64
7	int_rate	466285	non-null float64
8	installment	466285	non-null object
9	grade	466285	non-null object
10	sub_grade	466285	non-null object
11	emp_title	438697	non-null object
12	emp_length	445277	non-null object
13	home_ownership	466285	non-null object
14	annual_inc	466281	non-null float64
15	verification_status	466285	non-null object
16	issue_d	466285	non-null object
17	loan_status	466285	non-null object
18	pymnt_plan	466285	non-null object
19	url	466285	non-null object
20	desc	125981	non-null object
21	purpose	466285	non-null object
22	title	466264	non-null object
23	zip_code	466285	non-null object
24	addr_state	466285	non-null object
25	dti	466285	non-null float64
26	delinq_2yrs	466256	non-null float64
27	earliest_cr_line	466256	non-null object
28	inq_last_6mths	466256	non-null float64
29	mths_since_last_delinq	215934	non-null float64
30	mths_since_last_record	62638	non-null float64
31	open_acc	466256	non-null float64
32	pub_rec	466256	non-null float64
33	revol_bal	466285	non-null int64
34	revol_util	465945	non-null float64
35	total_acc	466256	non-null float64
36	initial_list_status	466285	non-null object
37	out_prncp	466285	non-null float64
38	out_prncp_inv	466285	non-null float64

Data Understanding

	0
Unnamed: 0	0.000000
id	0.000000
member_id	0.000000
loan_amnt	0.000000
funded_amnt	0.000000
...	...
all_util	1.000000
total_rev_hi_lim	0.150715
inq_fi	1.000000
total_cu_tl	1.000000
inq_last_12m	1.000000
75 rows × 1 columns	
dtype: float64	

	missing_proportion
desc	0.729820
mths_since_last_delinq	0.536906
mths_since_last_record	0.865666
mths_since_last_major_derog	0.787739
annual_inc_joint	1.000000
dti_joint	1.000000
verification_status_joint	1.000000
open_acc_6m	1.000000
open_il_6m	1.000000
open_il_12m	1.000000
open_il_24m	1.000000
mths_since_rcnt_il	1.000000
total_bal_il	1.000000
il_util	1.000000
missing_proportion	
inq_last_12m	1.0
total_bal_il	1.0
dti_joint	1.0
verification_status_joint	1.0
annual_inc_joint	1.0
...	...
total_pymnt	0.0
total_pymnt_inv	0.0
total_rec_prncp	0.0
total_rec_int	0.0
out_prncp	0.0
[75 rows × 1 columns]	

Dataset memiliki beberapa kolom dengan nilai hilang signifikan, seperti kolom **desc** dan **dti_joint**, memerlukan pembersihan atau imputasi agar dapat digunakan secara optimal dalam pemodelan.

Data Cleaning

Kolom: term

Nilai unik: ['36 months' '60 months']

Kolom: emp_title

Nilai unik: [nan 'Ryder' 'AIR RESOURCES BOARD' ... 'Mecânica' 'Chief of Interpretation (Park Ranger)' 'Server Engineer Lead']

Kolom: emp_length

Nilai unik: ['10+ years' '< 1 year' '1 year' '3 years' '8 years' '9 years' '5 years' '6 years' '2 years' '7 years' nan]

Kolom: issue_d

Nilai unik: ['Dec-11' 'Nov-11' 'Oct-11' 'Sep-11' 'Aug-11' 'Jul-11' 'Jun-11' 'Apr-11' 'Mar-11' 'Feb-11' 'Jan-11' 'Dec-10' 'Nov-10' 'Oct-10' 'Sep-10' 'Aug-10' 'Jul-10' 'Jun-10' 'May-10' 'Apr-10' 'Mar-10' 'Feb-10' 'Jan-10' 'Dec-09' 'Nov-09' 'Oct-09' 'Sep-09' 'Aug-09' 'Jul-09' 'Jun-09' 'May-09' 'Apr-09' 'Mar-09' 'Feb-09' 'Jan-09' 'Dec-08' 'Nov-08' 'Oct-08' 'Sep-08' 'Aug-08' 'Jul-08' 'Jun-08' 'May-08' 'Apr-08' 'Mar-08' 'Feb-08' 'Jan-08' 'Dec-07' 'Nov-07' 'Oct-07' 'Sep-07' 'Aug-07' 'Jul-07' 'Jun-07' 'Dec-13' 'Nov-13' 'Oct-13' 'Sep-13' 'Aug-13' 'Jul-13' 'Jun-13' 'May-13' 'Apr-13' 'Mar-13' 'Feb-13' 'Jan-13' 'Dec-12' 'Nov-12' 'Oct-12' 'Sep-12' 'Aug-12' 'Jul-12' 'Jun-12' 'May-12' 'Apr-12' 'Mar-12' 'Feb-12' 'Jan-12' 'Dec-14' 'Nov-14' 'Oct-14' 'Sep-14' 'Aug-14' 'Jul-14' 'Jun-14' 'May-14' 'Apr-14' 'Mar-14' 'Feb-14' 'Jan-14']

Kolom: loan_status

Nilai unik: ['Fully Paid' 'Charged Off' 'Current' 'Default' 'Late (31-120 days)' 'In Grace Period' 'Late (16-30 days)' 'Does not meet the credit policy. Status:Fully Paid' 'Does not meet the credit policy. Status:Charged Off']

Kolom: pymnt_plan

Nilai unik: ['n' 'y']

Kolom: title

Nilai unik: ['Computer' 'bike' 'real estate business' ... 'LoanGetter' 'Consolidation 01' 'Paying off the car and some bills']

Kolom: zip_code

Nilai unik: ['860xx' '309xx' '606xx' '917xx' '972xx' '852xx' '280xx' '900xx' '958xx' '774xx' '853xx' '913xx' '245xx' '951xx' '641xx' '921xx' '067xx' '890xx' '770xx' '335xx' '799xx' '605xx' '103xx' '150xx' '326xx' '564xx' '141xx' '080xx' '330xx' '974xx' '934xx' '405xx' '946xx' '445xx' '850xx' '604xx' '292xx' '088xx' '180xx' '029xx' '700xx' '010xx' '441xx' '104xx' '061xx' '616xx' '947xx' '914xx' '765xx' '980xx' '017xx' '752xx' '787xx' '077xx' '540xx' '225xx' '440xx' '437xx' '559xx' '912xx' '325xx' '300xx' '923xx' '352xx' '013xx' '146xx' '074xx' '786xx' '937xx' '331xx' '115xx' '191xx' '114xx' '908xx' '902xx' '992xx' '750xx' '950xx' '329xx' '226xx' '614xx' '802xx' '672xx' '083xx' '100xx' '926xx' '931xx' '712xx' '060xx' '707xx' '342xx' '895xx' '430xx' '919xx' '996xx' '891xx' '935xx' '801xx' '928xx' '233xx' '927xx' '970xx' '211xx' '303xx' '070xx' '194xx' '263xx' '403xx' '301xx' '553xx' '993xx' '312xx' '432xx' '602xx' '216xx' '151xx' '971xx' '305xx' '334xx' '050xx' '129xx' '925xx' '483xx' '760xx' '961xx' '200xx' '085xx' '981xx' '601xx' '117xx' '063xx' '920xx' '543xx' '775xx' '570xx' '038xx' '221xx' '985xx' '113xx' '275xx' '236xx' '148xx' '028xx' '450xx' '532xx' '729xx' '321xx' '959xx' '941xx' '955xx' '217xx' '880xx' '660xx' '062xx' '193xx' '761xx' '857xx' '306xx' '271xx' '142xx' '956xx' '983xx' '945xx' '109xx' '112xx' '187xx' '630xx' '435xx' '488xx' '287xx' '705xx' '592xx' '318xx' '549xx' '212xx' '347xx' '274xx' '265xx' '785xx' '027xx' '089xx' '813xx' '069xx' '260xx' '201xx' '349xx' '322xx' '075xx' '124xx' '940xx' '967xx' '111xx' '773xx' '997xx' '076xx' '538xx' '021xx' '304xx'

Kolom: zip_code

Nilai unik: ['860xx' '309xx' '606xx' '917xx' '972xx' '852xx' '280xx' '900xx' '958xx' '774xx' '853xx' '913xx' '245xx' '951xx' '641xx' '921xx' '067xx' '890xx' '770xx' '335xx' '799xx' '605xx' '103xx' '150xx' '326xx' '564xx' '141xx' '080xx' '330xx' '974xx' '934xx' '405xx' '946xx' '445xx' '850xx' '604xx' '292xx' '088xx' '180xx' '029xx' '700xx' '010xx' '441xx' '104xx' '061xx' '616xx' '947xx' '914xx' '765xx' '980xx' '017xx' '752xx' '787xx' '077xx' '540xx' '225xx' '440xx' '437xx' '559xx' '912xx' '325xx' '300xx' '923xx' '352xx' '013xx' '146xx' '074xx' '786xx' '937xx' '331xx' '115xx' '191xx' '114xx' '908xx' '902xx' '992xx' '750xx' '950xx' '329xx' '226xx' '614xx' '802xx' '672xx' '083xx' '100xx' '926xx' '931xx' '712xx' '060xx' '707xx' '342xx' '895xx' '430xx' '919xx' '996xx' '891xx' '935xx' '801xx' '928xx' '233xx' '927xx' '970xx' '211xx' '303xx' '070xx' '194xx' '263xx' '403xx' '301xx' '553xx' '993xx' '312xx' '432xx' '602xx' '216xx' '151xx' '971xx' '305xx' '334xx' '050xx' '129xx' '925xx' '483xx' '760xx' '961xx' '200xx' '085xx' '981xx' '601xx' '117xx' '063xx' '920xx' '543xx' '775xx' '570xx' '038xx' '221xx' '985xx' '113xx' '275xx' '236xx' '148xx' '028xx' '450xx' '532xx' '729xx' '321xx' '959xx' '941xx' '955xx' '217xx' '880xx' '660xx' '062xx' '193xx' '761xx' '857xx' '306xx' '271xx' '142xx' '956xx' '983xx' '945xx' '109xx' '112xx' '187xx' '630xx' '435xx' '488xx' '287xx' '705xx' '592xx' '318xx' '549xx' '212xx' '347xx' '274xx' '265xx' '785xx' '027xx' '089xx' '813xx' '069xx' '260xx' '201xx' '349xx' '322xx' '075xx' '124xx' '940xx' '967xx' '111xx' '773xx' '997xx' '076xx' '538xx' '021xx' '304xx'

Dilakukan pengecekan **nilai unik** pada kolom bertipe 'object' dan 'bool' untuk memahami distribusi data, seperti kolom **term** yang memiliki dua kategori durasi pinjaman: "**36 months**" dan "**60 months**". Proses ini mempersiapkan data untuk pembersihan lebih lanjut.

Data Cleaning

```
# Daftar kolom yang perlu dibersihkan (ordinal)
col_need_to_clean = ['term', 'emp_length', 'issue_d', 'earliest_cr_line', 'last_pymnt_d', 'next_pymnt_d', 'last_credit_pull_d']

# Tampilkan nilai unik pada kolom 'term' sebelum pembersihan
print("Nilai unik 'term' sebelum pembersihan:", df['term'].unique())

Nilai unik 'term' sebelum pembersihan: [' 36 months' ' 60 months']

# Pembersihan kolom 'term'
# Menghapus string ' months' dan mengubah nilai menjadi numerik
df['term'] = pd.to_numeric(df['term'].str.replace(' months', '', regex=False))

# Tampilkan kembali nilai unik pada kolom 'term' setelah pembersihan
print("Nilai unik 'term' setelah pembersihan:", df['term'].unique())

Nilai unik 'term' setelah pembersihan: [36 60]

# Pembersihan kolom 'term'
# Menghapus string ' months' dan mengubah nilai menjadi numerik
df['term'] = pd.to_numeric(df['term'].str.replace(' months', '', regex=False))

# Tampilkan kembali nilai unik pada kolom 'term' setelah pembersihan
print("Nilai unik 'term' setelah pembersihan:", df['term'].unique())

Nilai unik 'term' setelah pembersihan: [36 60]
```

Kolom **term** yang awalnya berisi nilai string (misalnya "36 months") dibersihkan dengan menghapus kata "**months**" dan mengubahnya menjadi tipe numerik. Proses ini memastikan bahwa kolom **term** hanya berisi angka **(36, 60)**, yang lebih sesuai untuk analisis dan pemodelan. Setelah pembersihan, nilai unik pada kolom term menjadi **[36, 60]**.

Data Cleaning

	emp_length
0	10.0
1	0.0
2	10.0
3	10.0
4	1.0
...	...
466280	4.0
466281	10.0
466282	7.0
466283	3.0
466284	10.0

466285 rows × 1 columns

dtype: float64

```

df['emp_length'] = df['emp_length'].fillna(0.0)

# Berikut menghilangkan data 'NaN' dan merubah keseluruhan data menjadi integer
df['emp_length'] = df['emp_length'].astype(str)
df['emp_length'].unique()

array(['10.0', '0.0', '1.0', '3.0', '8.0', '9.0', '4.0', '5.0', '6.0',
       '2.0', '7.0'], dtype=object)

# Kolom 'emp_length' akan dirubah kedalam bentuk numerik dan akan dihilangkan stringnya

df['emp_length'] = df['emp_length'].str.replace(' years', '')
df['emp_length'] = df['emp_length'].str.replace(' years', '')
df['emp_length'] = df['emp_length'].str.replace('< 1 year', '0')
df['emp_length'] = df['emp_length'].str.replace(' year', '')

df['emp_length'].fillna(value = 0, inplace=True)

```

Kolom **emp_length** yang berisi nilai NaN diisi dengan 0, kemudian data diubah menjadi tipe string untuk memudahkan pembersihan nilai. Selanjutnya, string seperti "years", "year", dan "< 1 year" dihapus atau diganti dengan angka yang sesuai. Setelah pembersihan, kolom **emp_length** berisi angka yang merepresentasikan tahun pengalaman kerja dalam format numerik, dan nilai yang hilang diisi dengan 0.

Data Cleaning

Data pada kolom-kolom tanggal:

```
issue_d earliest_cr_line last_pymnt_d next_pymnt_d last_credit_pul Jumlah missing values pada setiap kolom bertipe tanggal:  
0 Dec-11 Jan-85 Jan-15 NaN Jan issue_d: 0  
1 Dec-11 Apr-99 Apr-13 NaN Sep earliest_cr_line: 29  
2 Dec-11 Nov-01 Jun-14 NaN Jan last_pymnt_d: 376  
3 Dec-11 Feb-96 Jan-15 NaN Jan next_pymnt_d: 227214  
4 Dec-11 Jan-96 Jan-16 Feb-16 Jan last_credit_pul_d: 42  
... ... ... ...  
466280 Jan-14 Apr-03 Jan-16 Feb-16 Jan issue_d - Jumlah missing values setelah penghapusan: 0  
466281 Jan-14 Jun-97 Dec-14 NaN Jan earliest_cr_line - Jumlah missing values setelah penghapusan: 0  
466282 Jan-14 Dec-01 Jan-16 Feb-16 Dec last_pymnt_d - Jumlah missing values setelah penghapusan: 0  
466283 Jan-14 Feb-03 Dec-14 NaN Apr next_pymnt_d - Jumlah missing values setelah penghapusan: 0  
466284 Jan-14 Feb-00 Jan-16 Feb-16 Jan last_credit_pul_d - Jumlah missing values setelah penghapusan: 0  
[466285 rows x 5 columns]
```

Kolom **bertipe tanggal** diperiksa untuk nilai yang hilang (NaN), dan baris dengan NaN dihapus. Setelah penghapusan, tidak ada nilai yang hilang pada **kolom tanggal**, memastikan data bersih dan siap digunakan.

Exploratory Data Analysis

→ 239071

		count	unique	top	freq
	loan_performance	238121	100	Feb-16	200649
0		8950	2	Feb-16	7744

Berdasarkan analisis, diketahui bahwa terdapat **239.071** kreditur yang akan membayar pada pembayaran berikutnya (**next_pymnt_d**).

→ Statistik dasar untuk 'last_pymnt_d':

```
count    465909
unique     98
top      Jan-16
freq    179620
Name: last_pymnt_d, dtype: object
```

Distribusi 'last_pymnt_d' berdasarkan 'loan_performance':

	count	unique	top	freq
loan_performance	414848	98	Jan-16	178744
0	51061	94	Jul-15	2426

Statistik dasar untuk 'last_credit_pull_d':

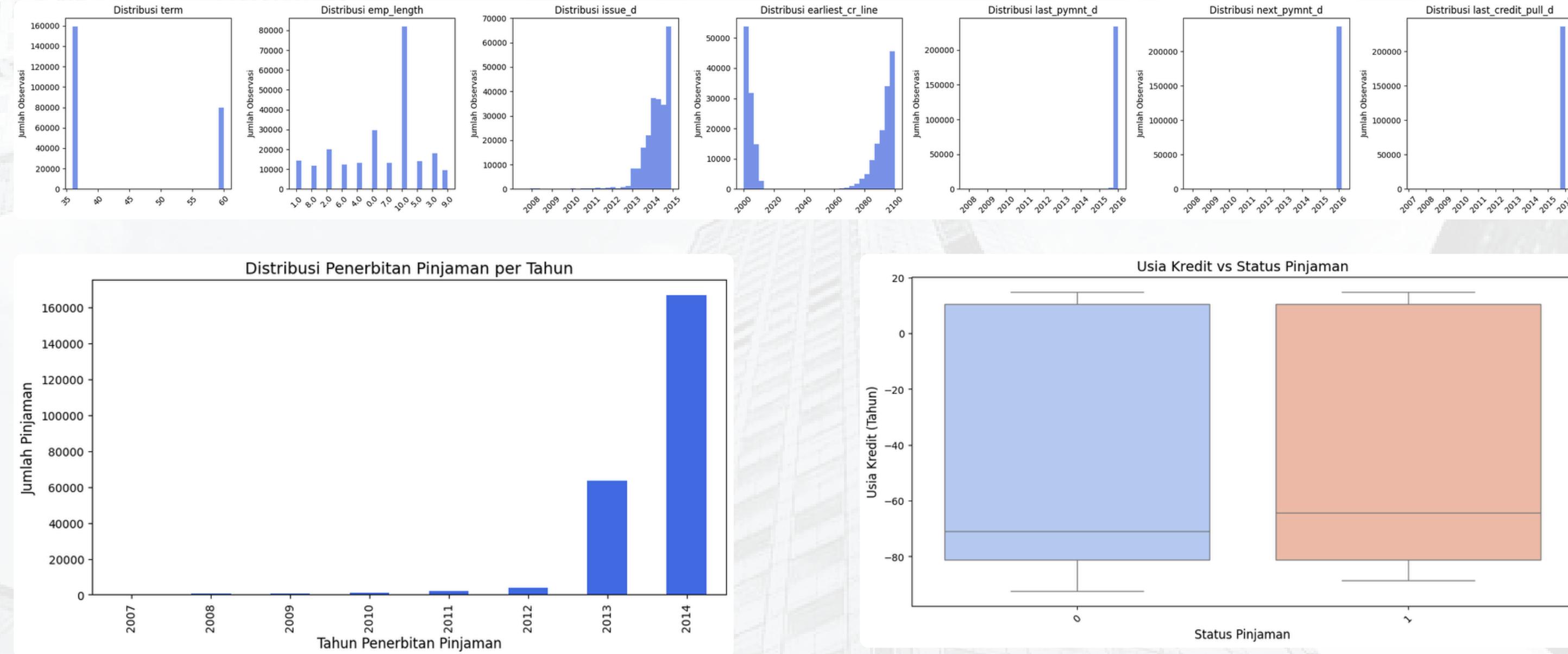
```
count    466243
unique    103
top      Jan-16
freq    327699
Name: last_credit_pull_d, dtype: object
```

Distribusi 'last_credit_pull_d' berdasarkan 'loan_performance':

	count	unique	top	freq
loan_performance	414824	103	Jan-16	300881
0	51419	79	Jan-16	26818

Kolom **last_pymnt_d** maupun **last_credit_pull_d** menunjukkan konsentrasi besar pada Januari 2016, dengan mayoritas pinjaman berstatus 'good' memiliki pembayaran terakhir dan penarikan laporan kredit pada bulan tersebut.

EDA Univariate Analysis



Sebagian besar pinjaman berdurasi **36 bulan** dan diterbitkan antara **2010-2014**, dengan puncak pada **2013-2014**. Pemilik pinjaman '**good**' memiliki usia kredit lebih lama dibandingkan '**bad**', terlihat dari boxplot usia kredit berdasarkan status pinjaman.

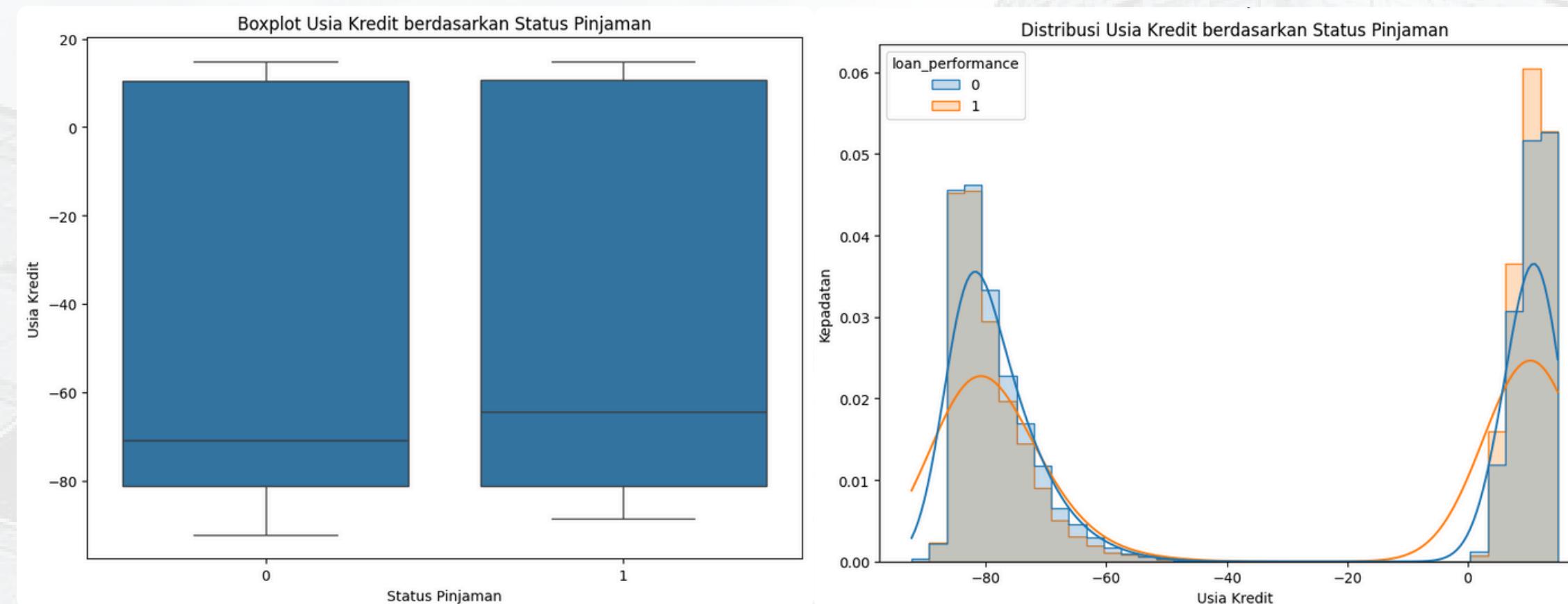
EDA Univariate Analysis

F-statistic: 82.03010173122637

P-value: 1.349992613280189e-19

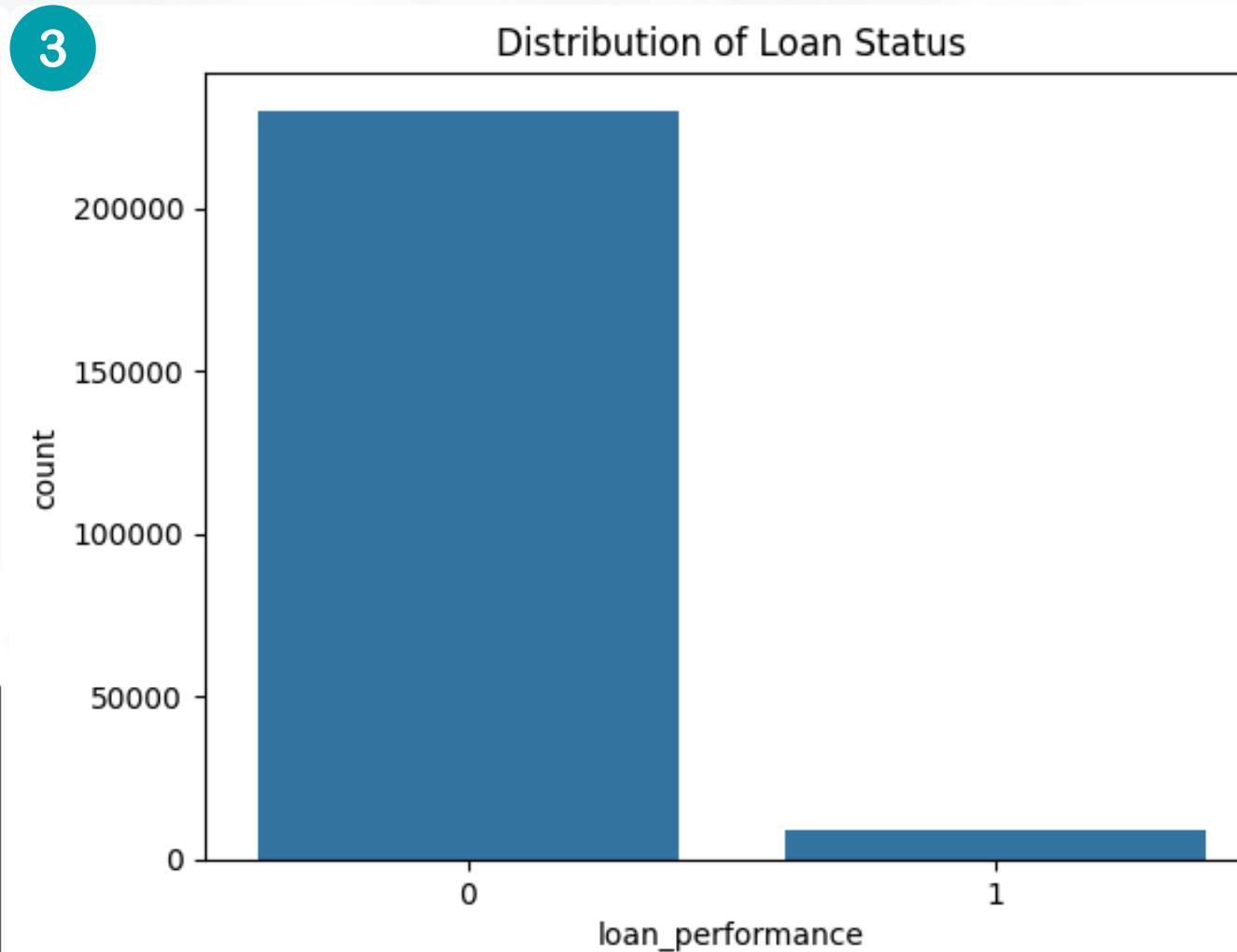
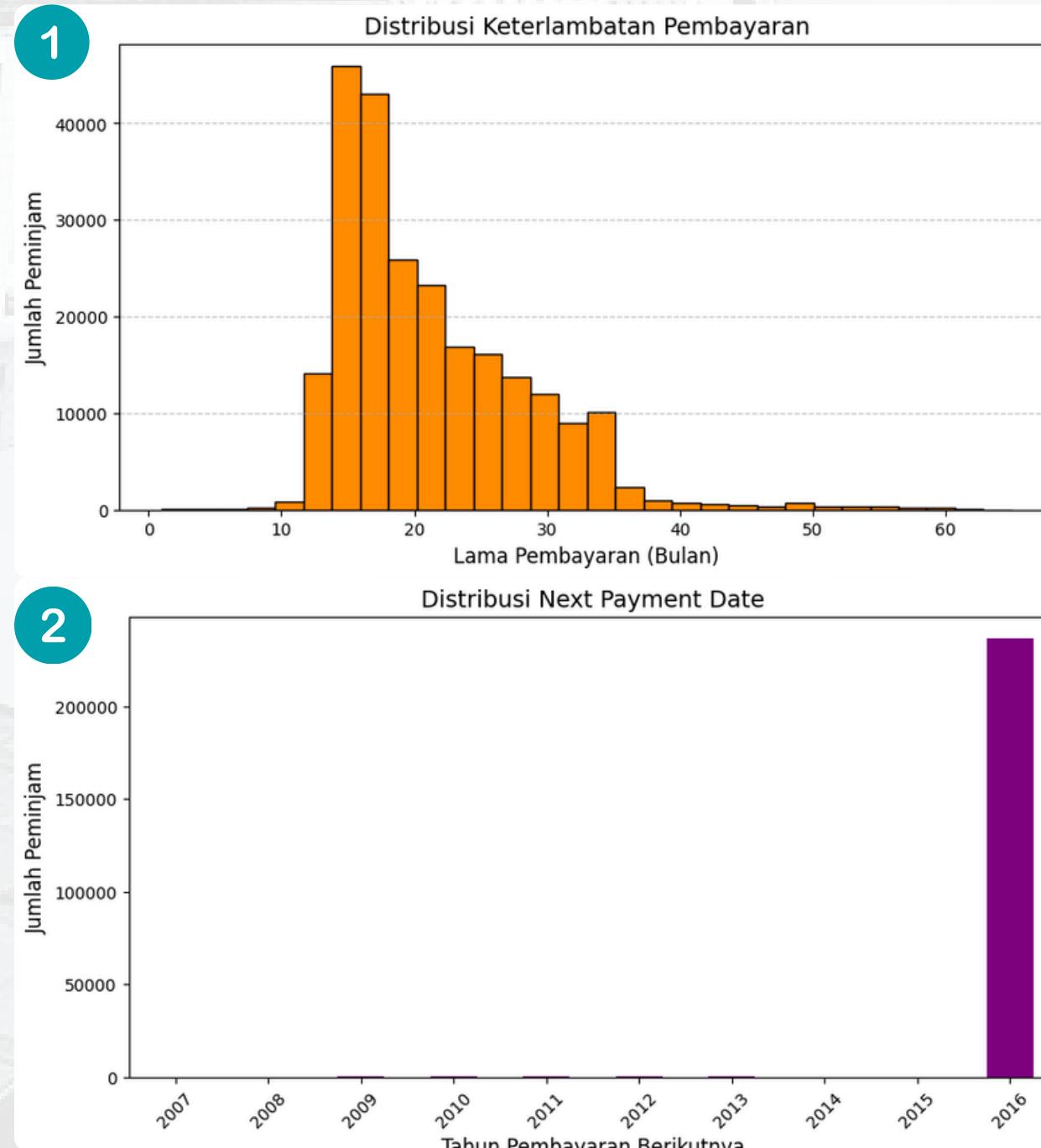
Terdapat perbedaan signifikan antara usia kredit peminjam yang gagal bayar dan yang tidak.

Nilai **F-statistik** dan **p-value** menunjukkan adanya perbedaan **signifikan** antara usia kredit peminjam yang **gagal bayar** dan **yang tidak**, dengan p-value yang sangat kecil.



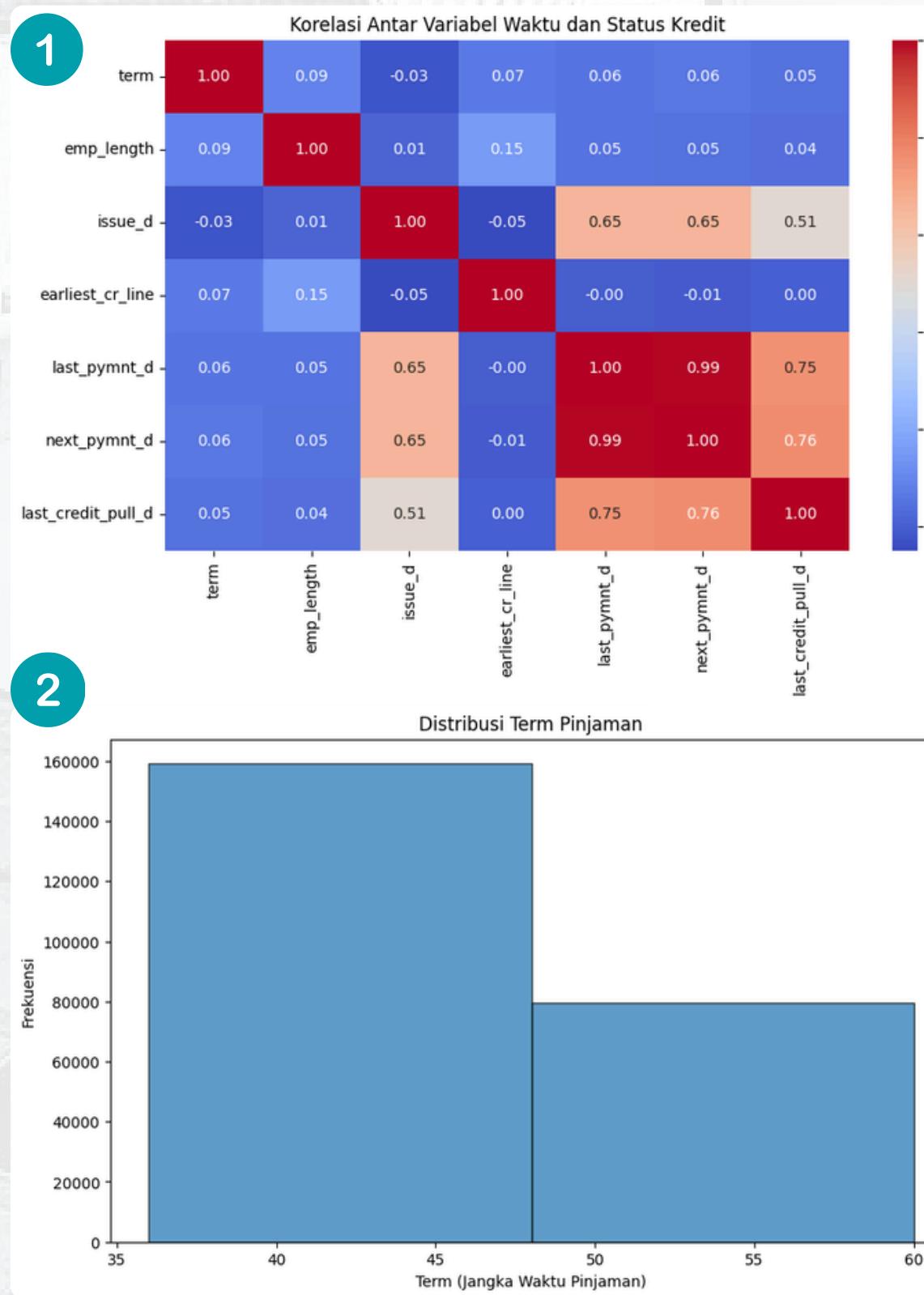
Peminjam yang **gagal bayar** cenderung memiliki **usia kredit** yang **lebih rendah** atau **negatif**, sementara peminjam yang **tidak gagal bayar** memiliki **usia kredit** yang **lebih tinggi** dan **lebih tersebar**.

EDA Univariate Analysis



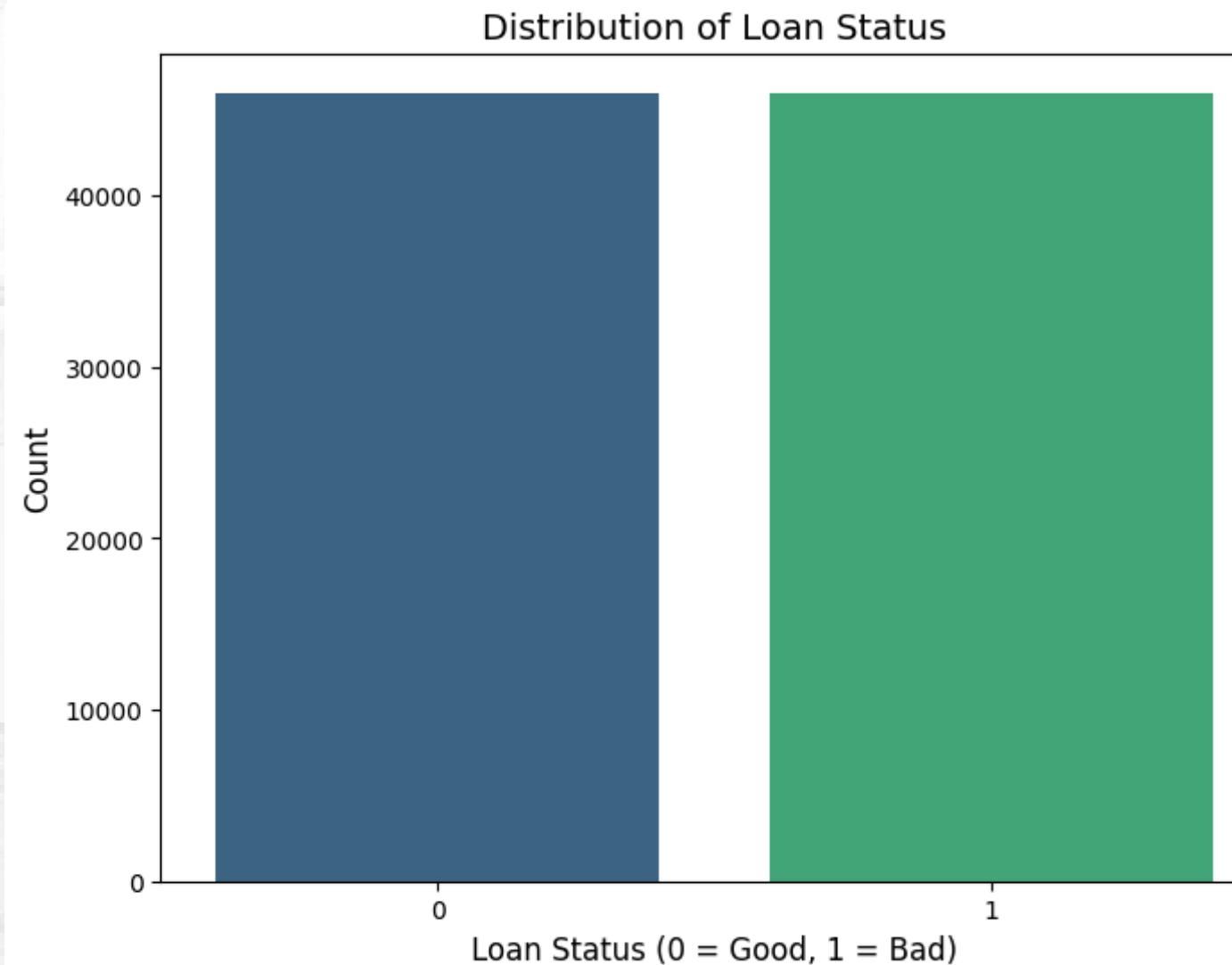
- Distribusi Keterlambatan Pembayaran:** Mayoritas pinjaman memiliki keterlambatan 0-10 bulan.
- Distribusi Next Payment Date:** Sebagian besar pembayaran berikutnya dijadwalkan pada tahun 2016.
- Distribusi Status Pinjaman:** Terdapat ketidakseimbangan besar, dengan lebih banyak pinjaman berstatus '0' (pembayaran tepat waktu).

EDA Multivariate Analysis



- 1. Korelasi Antar Variabel Waktu dan Status Kredit:** Menunjukkan korelasi tinggi antara variabel terkait status pinjaman.
- 2. Distribusi Term Pinjaman:** Mayoritas pinjaman berdurasi 36 bulan.
- 3. Hubungan antara Tanggal Penerbitan dan Tanggal Pembayaran:** Pembayaran sebagian besar terjadi setelah penerbitan pinjaman, terutama pada 2013-2014.

Data Preparation

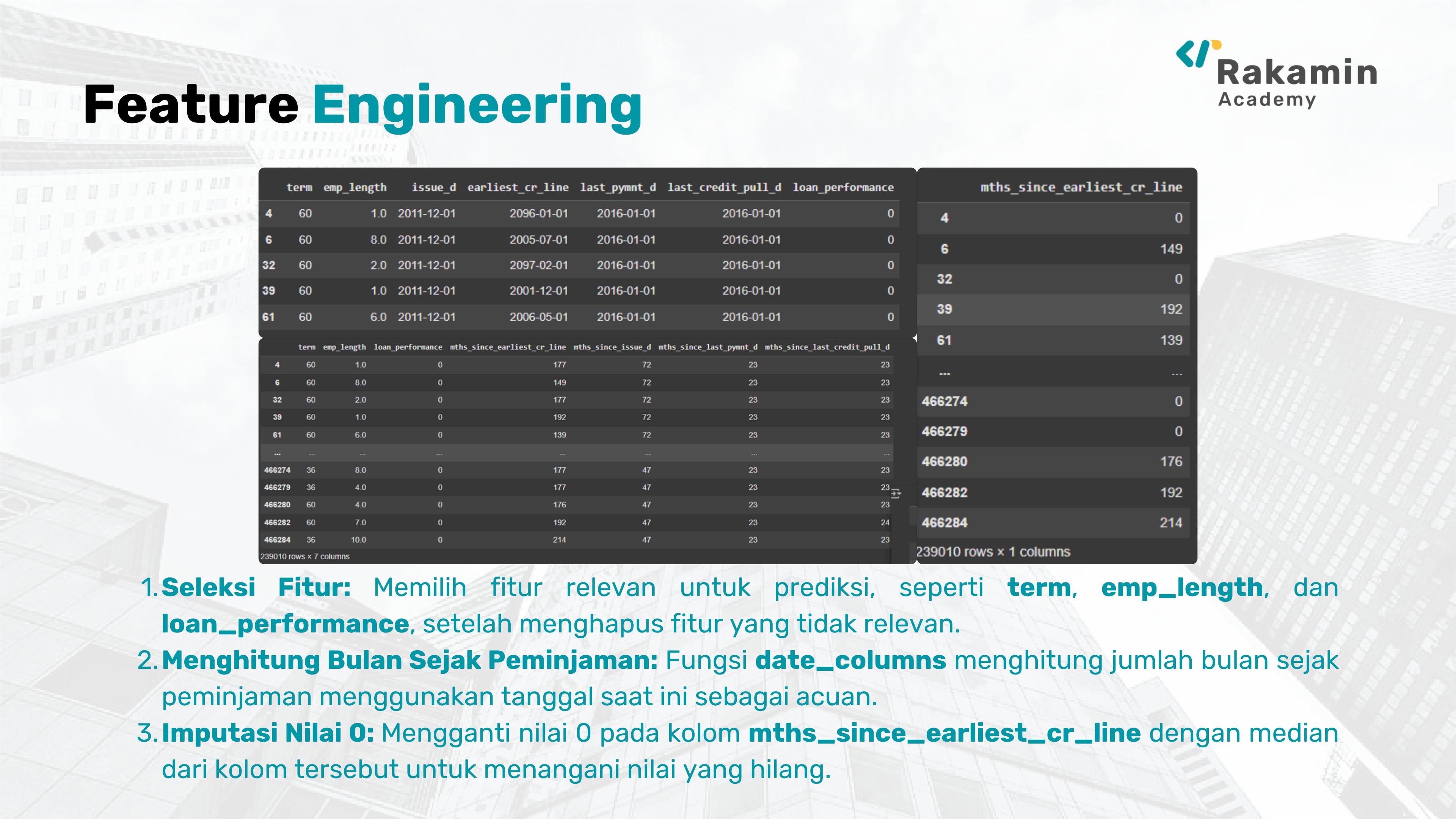


```
Distribusi pada Test Set:  
loan_performance  
1 0.5  
0 0.5  
Name: proportion, dtype: float64  
  
Distribusi pada Train Set:  
loan_performance  
0 0.5  
1 0.5  
Name: proportion, dtype: float64  
  
Distribusi sebelum SMOTE: Counter({0: 230060, 1: 8950})  
Distribusi setelah SMOTE: Counter({0: 230060, 1: 230060})
```

y_test.shape	(92024,)
y_train.shape	(368096,)
X_test.shape	(92024, 6)
X_train.shape	(368096, 6)

Pada tahap Data Preparation, distribusi status pinjaman ditampilkan, kemudian **SMOTE** diterapkan untuk menangani ketidakseimbangan kelas dengan menghasilkan sampel sintetik pada kelas minoritas. Data kemudian dibagi menjadi train set dan test set dengan proporsi **80:20**, memastikan distribusi kelas seimbang pada kedua set menggunakan teknik stratified sampling.

Feature Engineering



The background of the slide features a grayscale image of a computer keyboard.

	term	emp_length	issue_d	earliest_cr_line	last_pymnt_d	last_credit_pull_d	loan_performance
4	60	1.0	2011-12-01	2096-01-01	2016-01-01	2016-01-01	0
6	60	8.0	2011-12-01	2005-07-01	2016-01-01	2016-01-01	0
32	60	2.0	2011-12-01	2097-02-01	2016-01-01	2016-01-01	0
39	60	1.0	2011-12-01	2001-12-01	2016-01-01	2016-01-01	0
61	60	6.0	2011-12-01	2006-05-01	2016-01-01	2016-01-01	0
<hr/>							
	term	emp_length	loan_performance	mths_since_earliest_cr_line	mths_since_issue_d	mths_since_last_pymnt_d	mths_since_last_credit_pull_d
4	60	1.0	0	177	72	23	23
6	60	8.0	0	149	72	23	23
32	60	2.0	0	177	72	23	23
39	60	1.0	0	192	72	23	23
61	60	6.0	0	139	72	23	23
...
466274	36	8.0	0	177	47	23	23
466279	36	4.0	0	177	47	23	23
466280	60	4.0	0	176	47	23	23
466282	60	7.0	0	192	47	23	24
466284	36	10.0	0	214	47	23	23

239010 rows × 7 columns

mths_since_earliest_cr_line	
4	0
6	149
32	0
39	192
61	139
...	...
466274	0
466279	0
466280	176
466282	192
466284	214

239010 rows × 1 columns

- Seleksi Fitur:** Memilih fitur relevan untuk prediksi, seperti **term**, **emp_length**, dan **loan_performance**, setelah menghapus fitur yang tidak relevan.
- Menghitung Bulan Sejak Peminjaman:** Fungsi **date_columns** menghitung jumlah bulan sejak peminjaman menggunakan tanggal saat ini sebagai acuan.
- Imputasi Nilai 0:** Mengganti nilai 0 pada kolom **mths_since_earliest_cr_line** dengan median dari kolom tersebut untuk menangani nilai yang hilang.

Data Modeling

```
# Import model machine learning yang akan digunakan

from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from xgboost import XGBClassifier

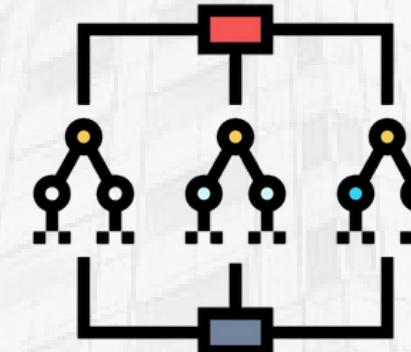
#Dibentuk variabel untuk model Logistic Regresi, yaitu 'model'

model = LogisticRegression()
model_2 = RandomForestClassifier()
model_3 = XGBClassifier()
```



Logistic Regression

Akurasi Model: 85.02%



Random Forest

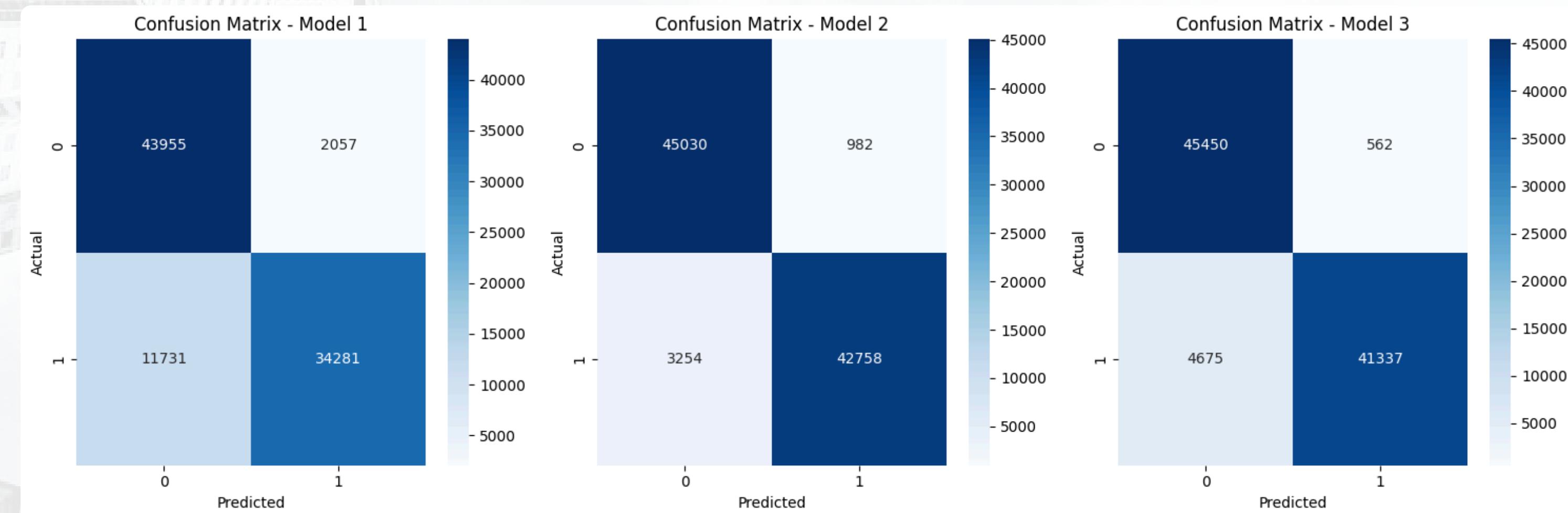
Akurasi Model: 95.40%

XGBoost

XGboost

Akurasi Model: 94.31%

Evaluation



1. **Logistic Regression:** Banyak false negatives—kurang efektif mendeteksi peminjam berisiko tinggi.
2. **Random Forest:** Meningkatkan true positives dibanding Logistic Regression, namun menghasilkan lebih banyak false positives daripada XGBoost.
3. **XGBoost:** Paling optimal—memiliki false negatives dan false positives paling rendah, sehingga paling andal dalam mengidentifikasi risiko kredit.

Classification Report

Classification Report for Logistic Regression:				
	precision	recall	f1-score	support
0	0.79	0.96	0.86	46012
1	0.94	0.75	0.83	46012
accuracy			0.85	92024
macro avg	0.87	0.85	0.85	92024
weighted avg	0.87	0.85	0.85	92024

Classification Report for Random Forest:				
	precision	recall	f1-score	support
0	0.93	0.98	0.96	46012
1	0.98	0.93	0.95	46012
accuracy			0.95	92024
macro avg	0.96	0.95	0.95	92024
weighted avg	0.96	0.95	0.95	92024

Classification Report for XGBoost:				
	precision	recall	f1-score	support
0	0.91	0.99	0.95	46012
1	0.99	0.90	0.94	46012
accuracy			0.94	92024
macro avg	0.95	0.94	0.94	92024
weighted avg	0.95	0.94	0.94	92024

- **Logistic Regression:** Memiliki akurasi 0.85, dengan recall tinggi untuk kelas 0 (96%), tetapi recall untuk kelas 1 (75%) lebih rendah, menunjukkan model kurang baik dalam mendeteksi peminjam berisiko tinggi.
- **Random Forest:** Mencapai akurasi 0.95, dengan recall tinggi pada kedua kelas (98% untuk kelas 0 dan 93% untuk kelas 1), lebih baik dalam mendeteksi kedua kategori, meskipun menghasilkan sedikit false positives.
- **XGBoost:** Memiliki akurasi 0.94, dengan recall yang sangat baik pada kelas 0 (99%) dan kelas 1 (90%), menjadikannya model yang paling seimbang dan efektif untuk mendeteksi risiko kredit tinggi.

Conclusion

Berdasarkan **Exploratory Data Analysis (EDA)**, ditemukan bahwa jumlah pinjaman meningkat tajam pada tahun 2014, dengan sebagian besar pinjaman mengalami keterlambatan 10-30 bulan. Hasil uji ANOVA menunjukkan bahwa usia kredit berpengaruh terhadap status pinjaman yang buruk, dengan peminjam yang lebih muda cenderung berisiko lebih tinggi.

Dari segi pemodelan, **XGBoost** menunjukkan kinerja terbaik dalam hal akurasi dan evaluasi metrik seperti False Positives dan False Negatives. Untuk meminimalkan kesalahan dalam mendeteksi peminjam berisiko tinggi, **XGBoost atau Random Forest lebih direkomendasikan** daripada Logistic Regression, sementara XGBoost lebih disarankan jika ingin mengurangi False Positives secara ketat.

Thank You

