

Credit Risk Prediction Modeling

ID/X Partners - Data Scientist

Presented by Nindita Sari Sarasidya

 Kota Bekasi nindita.saras@gmail.com www.linkedin.com/in/nindita-sari-sarasidya

Nindita Sari Sarasidya

Data & Machine Learning Enthusiast

Nindita is a **fresh graduate in Industrial Engineering ITB** with a **strong passion for data science** and business analysis. She has developed **expertise in Python, SQL, Looker Studio, and Tableau**, with additional proficiency in **Machine Learning** techniques for advanced data modeling and analysis.

Her **experience as a business process analyst** in the procurement division has sharpened her **problem-solving** skills, particularly in applying **analytical thinking** and **attention to detail**. Nindita is eager to deepen her knowledge in data-related fields and broaden her understanding of the business landscape.

About Company

id/x partners didirikan pada tahun 2002 oleh mantan bankir dan konsultan manajemen yang memiliki pengalaman luas dalam manajemen siklus dan proses kredit, pengembangan skoring, dan manajemen kinerja. Pengalaman gabungan kami telah melayani korporasi di seluruh wilayah Asia dan Australia serta di berbagai industri, khususnya layanan keuangan, telekomunikasi, manufaktur, dan ritel.

id/x partners menyediakan layanan konsultasi yang mengkhususkan diri dalam memanfaatkan solusi analitik data dan pengambilan keputusan (DAD) yang dikombinasikan dengan disiplin manajemen risiko dan pemasaran terintegrasi untuk membantu klien mengoptimalkan profitabilitas portofolio dan proses bisnis.

Layanan konsultasi yang komprehensif dan solusi teknologi yang ditawarkan oleh id/x partners menjadikannya sebagai penyedia layanan terpadu.



id/x partners

Project Portfolio

Latar Belakang

ID/X Partners membantu perusahaan pemberi pinjaman mengembangkan **model machine learning untuk memprediksi risiko kredit**. Dengan menggunakan dataset pinjaman yang **disetujui dan ditolak**, proyek ini bertujuan **meningkatkan akurasi penilaian risiko**, mengoptimalkan keputusan bisnis, dan mengurangi kerugian.

Problem Statement

- Mengembangkan model untuk prediksi resiko kredit
- Meningkatkan akurasi model dalam menilai kelayakan pinjaman

Dataset

- Loan Dataset
- Data Dictionary

Tools



[Github Link](#)

Road Map

Handling Missing Value
Handling Duplicate Data

Data
Preprocessing 1



Exploratory Data
Analysis (EDA)

Descriptive
Univariate
Multivariate

Handling Outlier
Encode
Scaling
Split
Handling Imbalance

Data
Preprocessing 2



Modeling

Logistic Regression
Random Forest

[Github Link](#)

Data Understanding

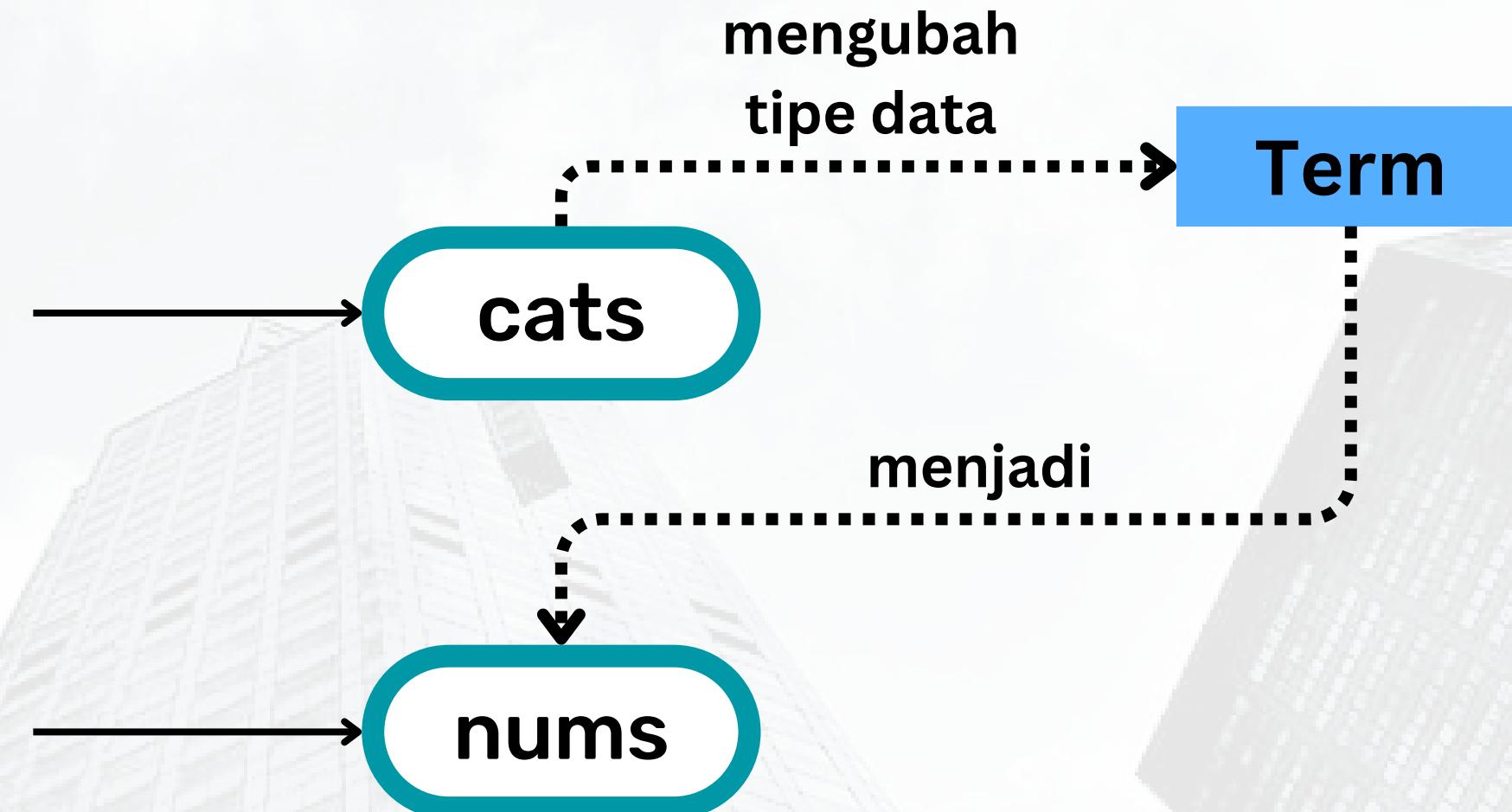
Dataset Information

Kolom
74

Float
22

numerik
33

Baris
466285



Jumlah kolom numerik + kategori tidak sama dengan kolom total. Hal ini merupakan dampak dari adanya kolom yang memiliki nilai kosong

Maka
cats = 21
nums = 34

Data Preparation 1

Database Flow

Handling Missing Value

Kolom tak relevan

15

```
# Let's define the irrelevant columns and missing columns
irrelevant_col = [
    'Unnamed: 0', 'id', 'member_id', 'pymnt_plan', 'url',
    'desc', 'title', 'zip_code', 'recoveries',
    'collection_recovery_fee', 'last_pymnt_d', 'last_credit_pull_d', 'policy_code', 'a]
```

Missing value

20

```
missing_col = [
    'mths_since_last_delinq', 'mths_since_last_record', 'annual_inc_joint', 'dti_joint',
    'open_acc_6m', 'open_il_6m', 'open_il_12m',
    'open_il_24m', 'mths_since_rcnt_il', 'total_bal_il',
    'il_util', 'open_rv_12m', 'open_rv_24m',
    'max_bal_bc', 'all_util', 'inq_fi', 'total_cu_tl',
    'inq_last_12m', 'mths_since_last_major_derog', 'next_pymnt_d'
```

Handling Duplicate Value

Tidak ditemukan data yang terduplicasi

Before & After

df

74

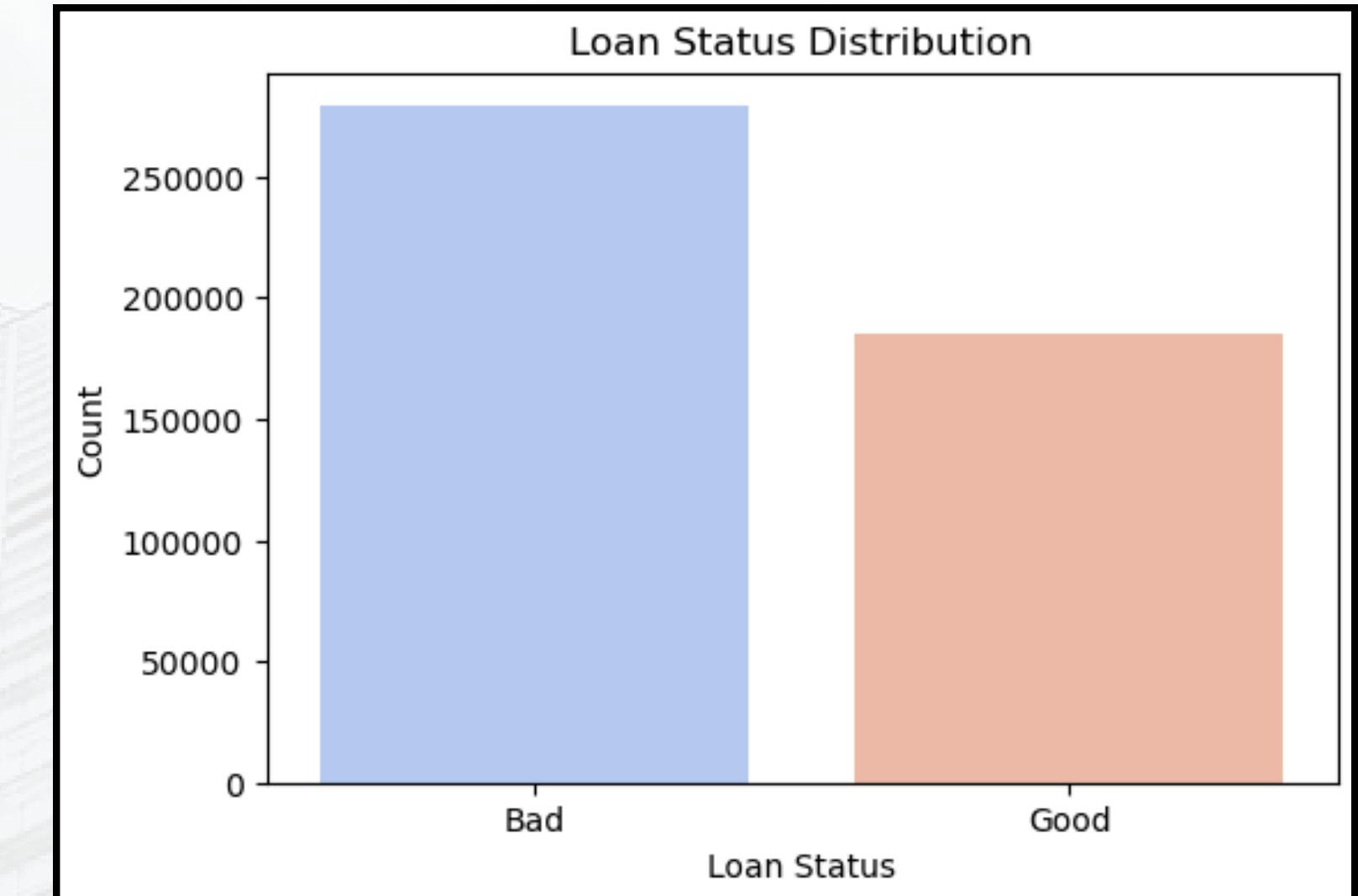
df_cleaned

39

Exploratory Data Analysis (EDA)

Loan Status Distribution

Bad	Good
Charged Off	
Default	
Current	Fully Paid
In Grace Period	
Late (31-120 days)	
Late (16-30 days)	



INSIGHT!

Jumlah pinjaman dengan status "Buruk" jauh lebih tinggi dibandingkan "Baik". Hal ini dapat mengindikasikan **risiko kredit yang signifikan** atau perlunya evaluasi lebih ketat dalam proses persetujuan pinjaman.

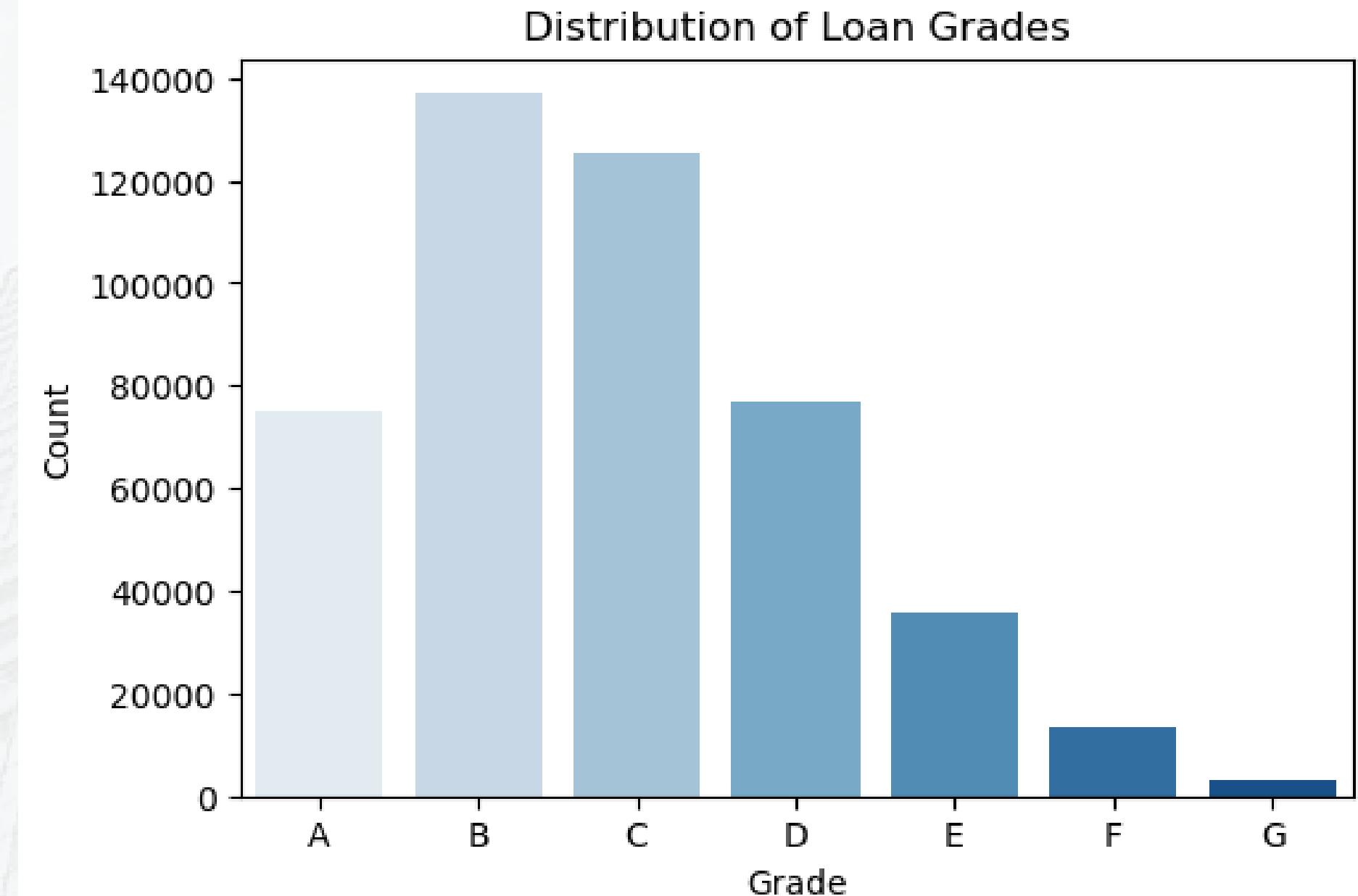
Exploratory Data Analysis (EDA)

Loan Grade Distribution

INSIGHT!

Distribusi grade pinjaman menunjukkan **majoritas pinjaman berada di grade B dan C**, sedangkan **grade F dan G memiliki jumlah paling sedikit**.

Hal ini mengindikasikan **kebanyakan pinjaman memiliki kualitas sedang**, namun pinjaman berkualitas rendah tetap ada meski proporsinya kecil.



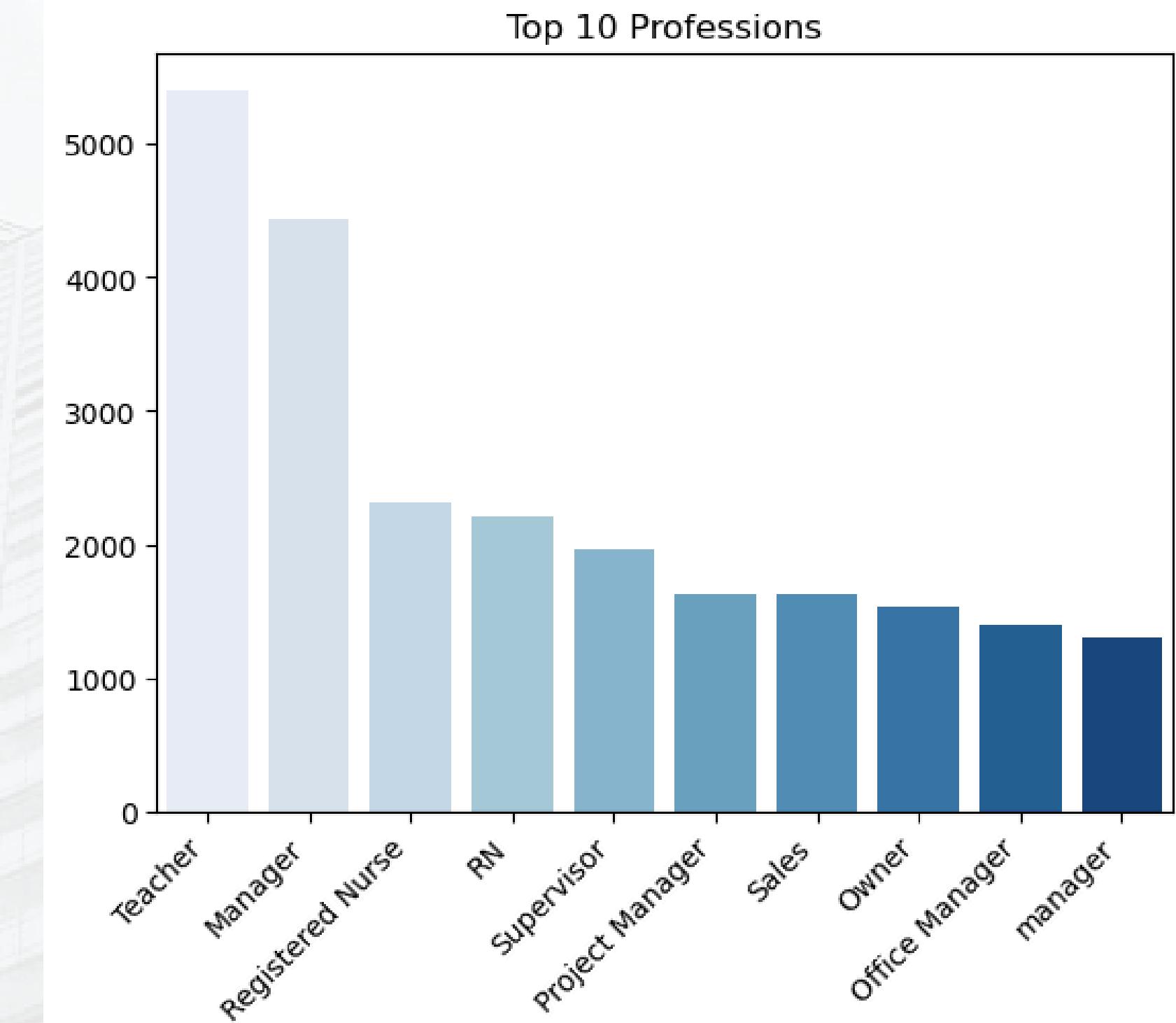
Exploratory Data Analysis (EDA)

Top 10 Professions

INSIGHT!

Mayoritas pengguna pinjaman berasal dari **profesi Teacher dan Manager**, menunjukkan permintaan tinggi dari sektor pendidikan dan manajemen.

Profesi kesehatan, seperti Registered Nurse dan RN, juga signifikan, mengindikasikan kebutuhan pembiayaan di kalangan tenaga medis. **Strategi pemasaran dapat difokuskan pada tiga sektor utama ini.**



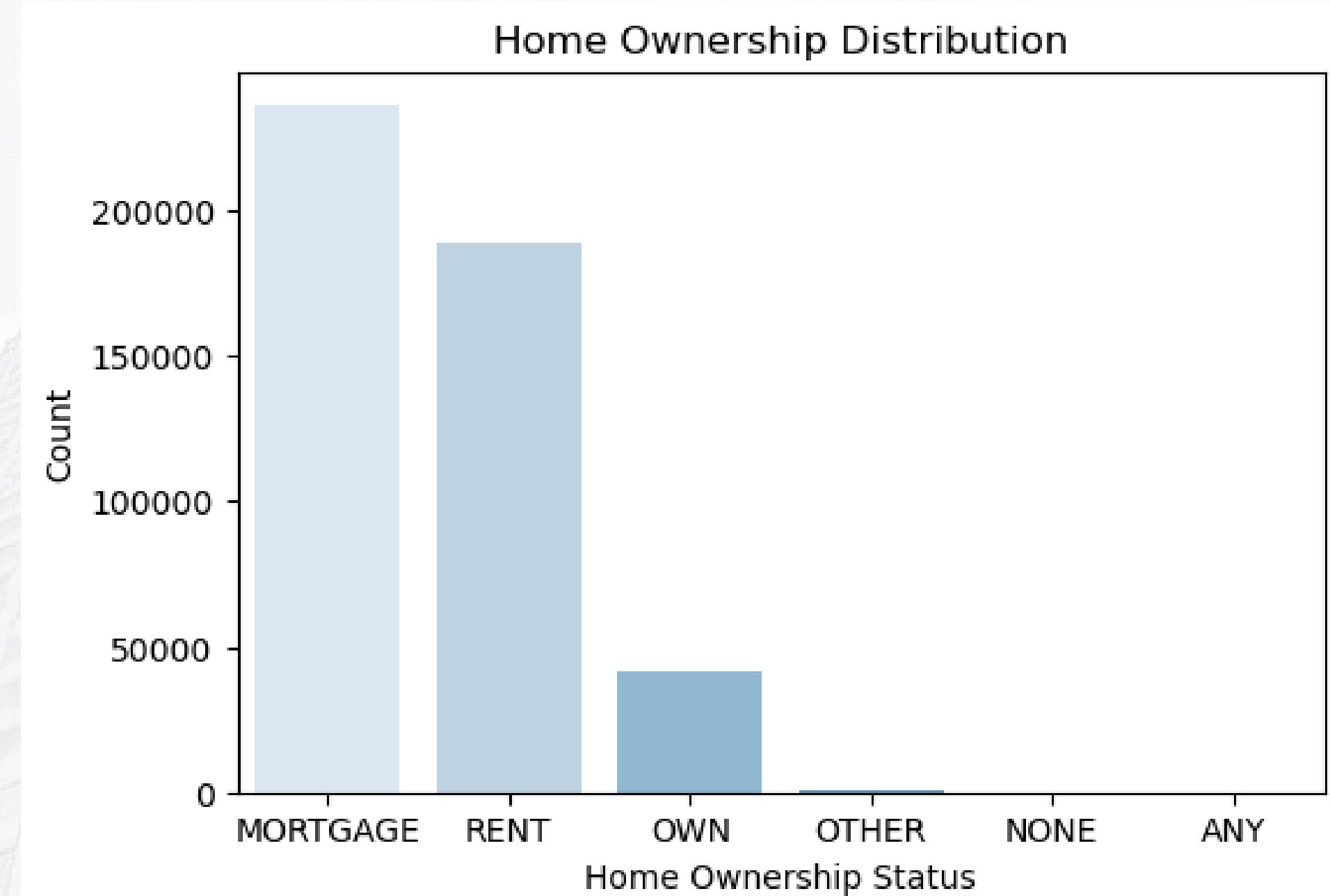
Exploratory Data Analysis (EDA)

Home Ownership Type

INSIGHT!

Mayoritas pengguna pinjaman memiliki status kepemilikan rumah berupa MORTGAGE atau RENT, menunjukkan bahwa mereka memiliki kewajiban finansial terhadap perumahan.

Pengguna dengan **status OWN jauh lebih sedikit**. Strategi dapat difokuskan pada penyediaan produk keuangan yang mendukung pembayaran cicilan atau kebutuhan penyewa.

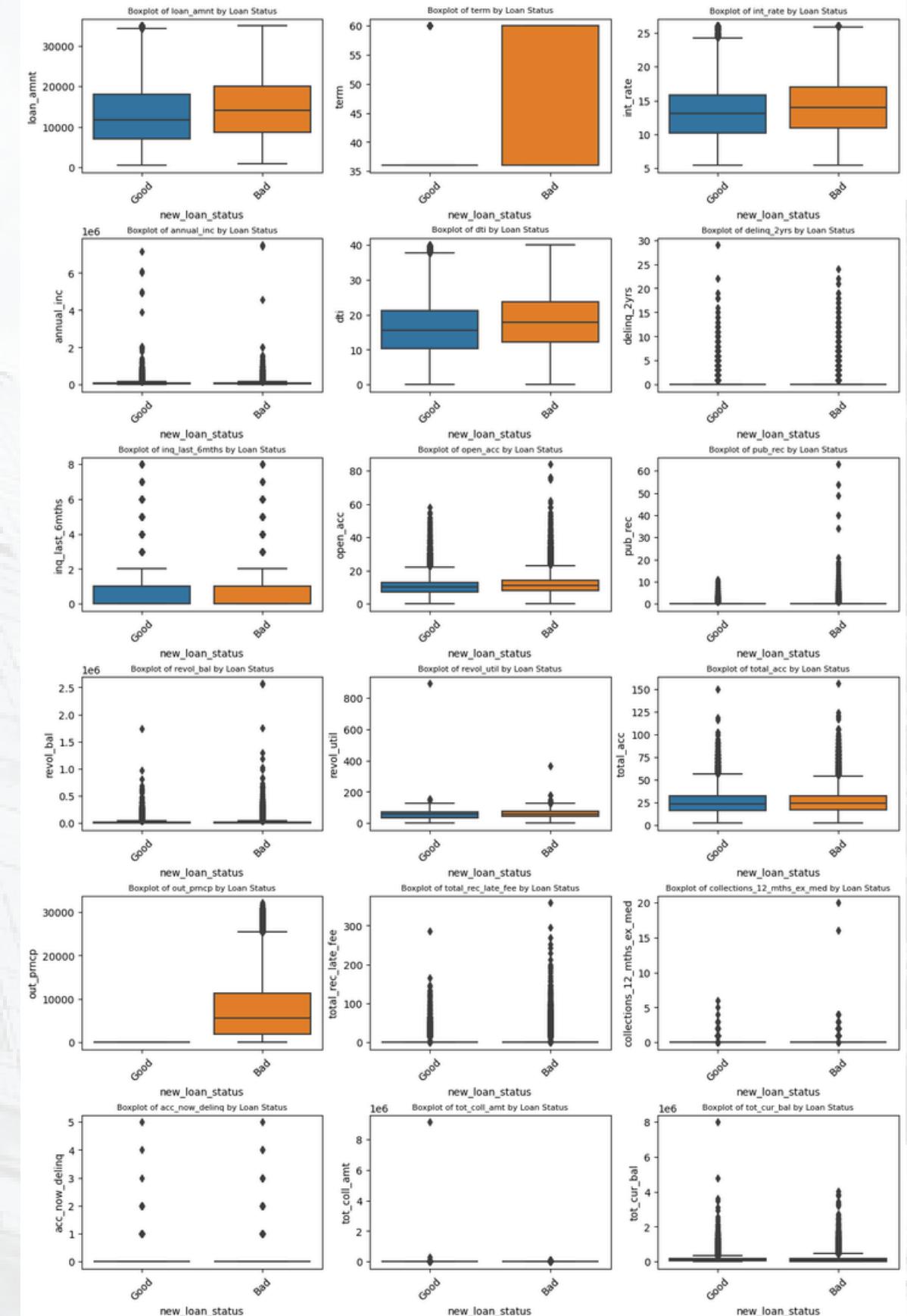


Exploratory Data Analysis (EDA)

Boxplot

Banyak data yang memiliki nilai outlier pada boxplot disamping.

Hal ini menandakan perlu dilakukan handling outlier

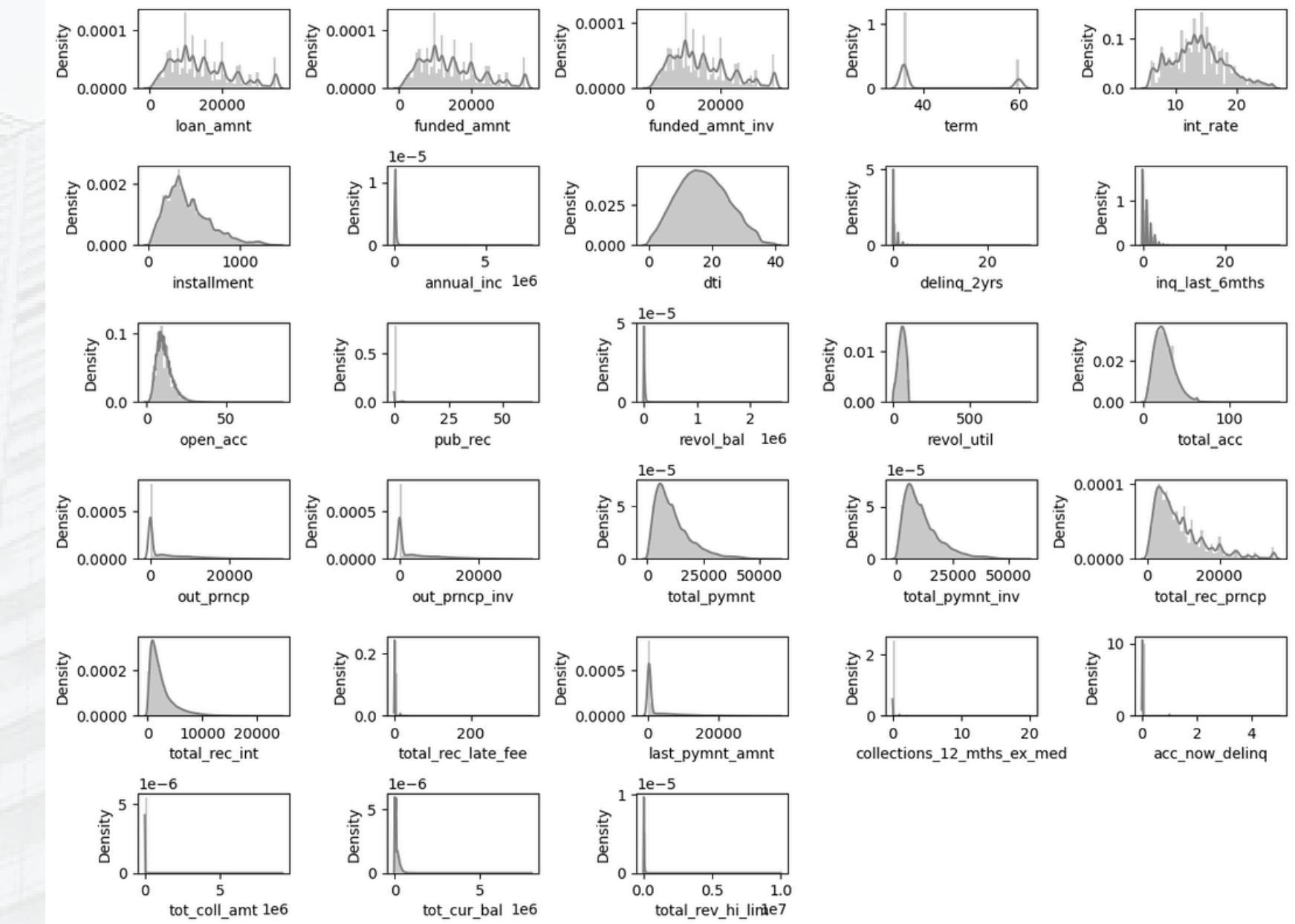


Exploratory Data Analysis (EDA)

Density Plot

Dengan menggunakan density plot,
kita bisa melihat distribusi tiap kolom
numerik yang ada.

Temuan ini menjadi bahan
pengambilan keputusan untuk
standardisasi maupun penggunaan
metode pengangan outlier



Exploratory Data Analysis (EDA)

Numerical Data Correlation

Before

28 Features

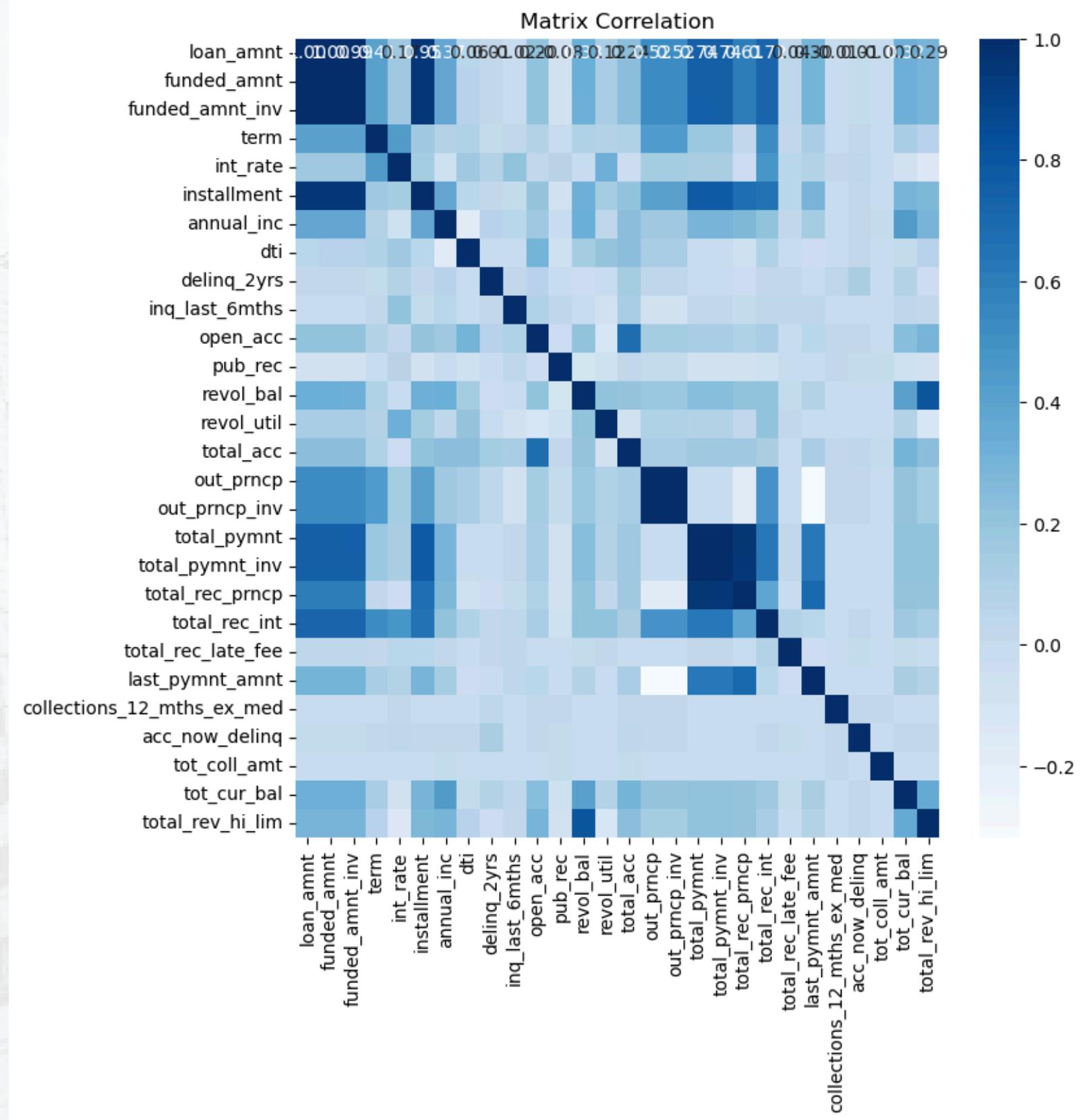
After

18 Features

```
# Find columns with correlation above 0.7
upper_tri = correlation_matrix.where(np.triu(np.ones(correlation_matrix.shape), k=1).a
must_drop = [column for column in upper_tri.columns if any(upper_tri[column] > 0.7)]
print(must_drop)

# Drop the highly correlated columns
df_cleaned.drop(must_drop, axis=1, inplace=True)

['funded_amnt', 'funded_amnt_inv', 'installment', 'out_prncp_inv', 'total_pymnt', 'tot
al_pymnt_inv', 'total_rec_prncp', 'total_rec_int', 'last_pymnt_amnt', 'total_rev_hi_li
m']
```



Exploratory Data Analysis (EDA)

Categorical Data Correlation

Before

13 Features

After

3 Features

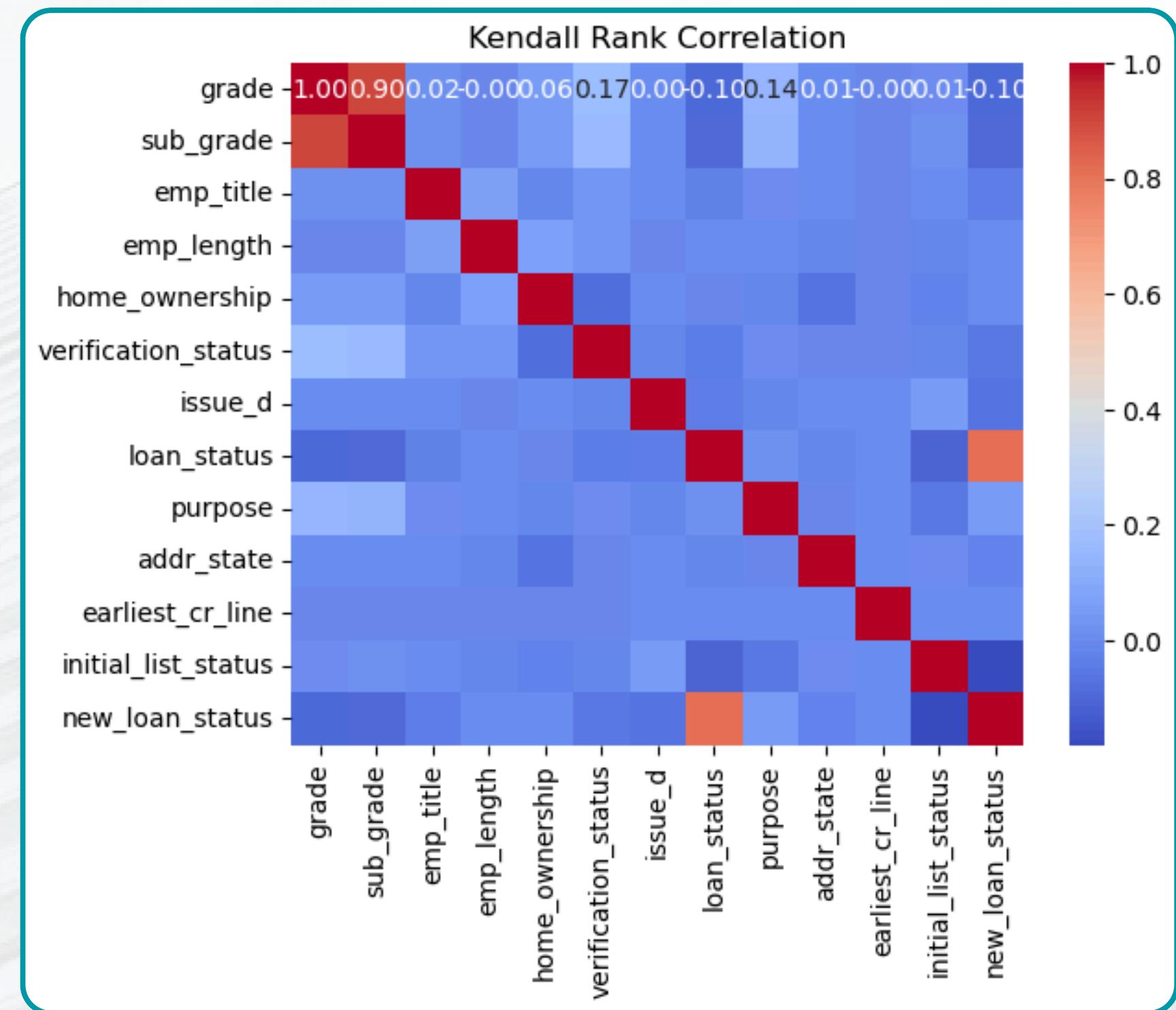
```
# Tentukan threshold untuk korelasi lemah
threshold = 0.1

# Ambil korelasi new_loan_status dengan fitur lainnya
new_loan_status_corr = kendall_corr['new_loan_status']

# Filter fitur dengan korelasi lemah terhadap new_loan_status
low_corr_features = new_loan_status_corr[(new_loan_status_corr > -threshold) & (new_loan_status_corr < threshold)]
print("Fitur dengan korelasi lemah atau tidak berkorelasi dengan new_loan_status:")
print(low_corr_features)

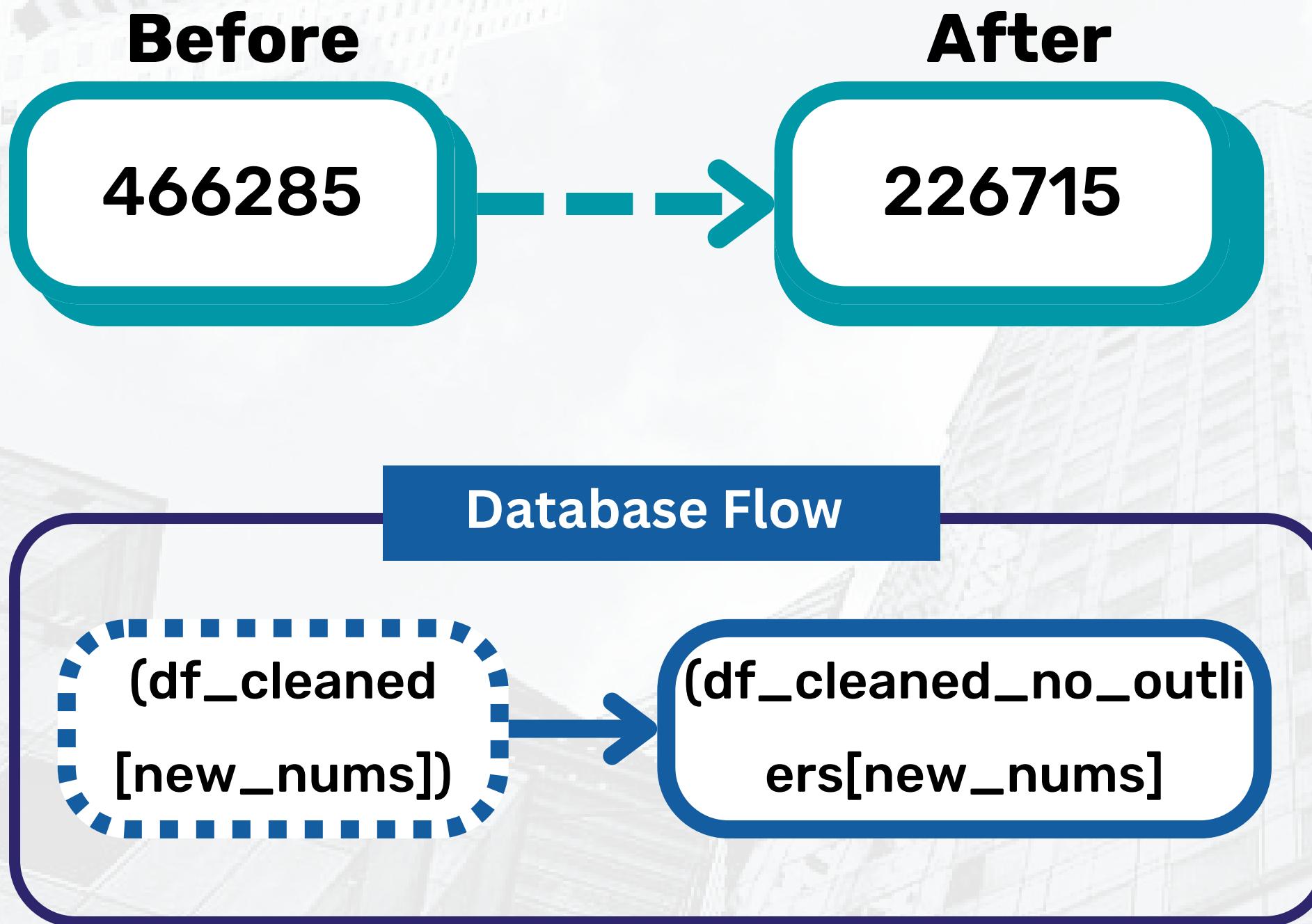
# Drop the highly correlated columns
df_cleaned.drop(low_corr_features, axis=1, inplace=True)

Fitur dengan korelasi lemah atau tidak berkorelasi dengan new_loan_status:
['sub_grade', 'emp_title', 'emp_length', 'home_ownership', 'verification_status', 'issue_d', 'purpose', 'addr_state', 'earliest_cr_line']
```



Data Preprocessing 2

Handling Outlier

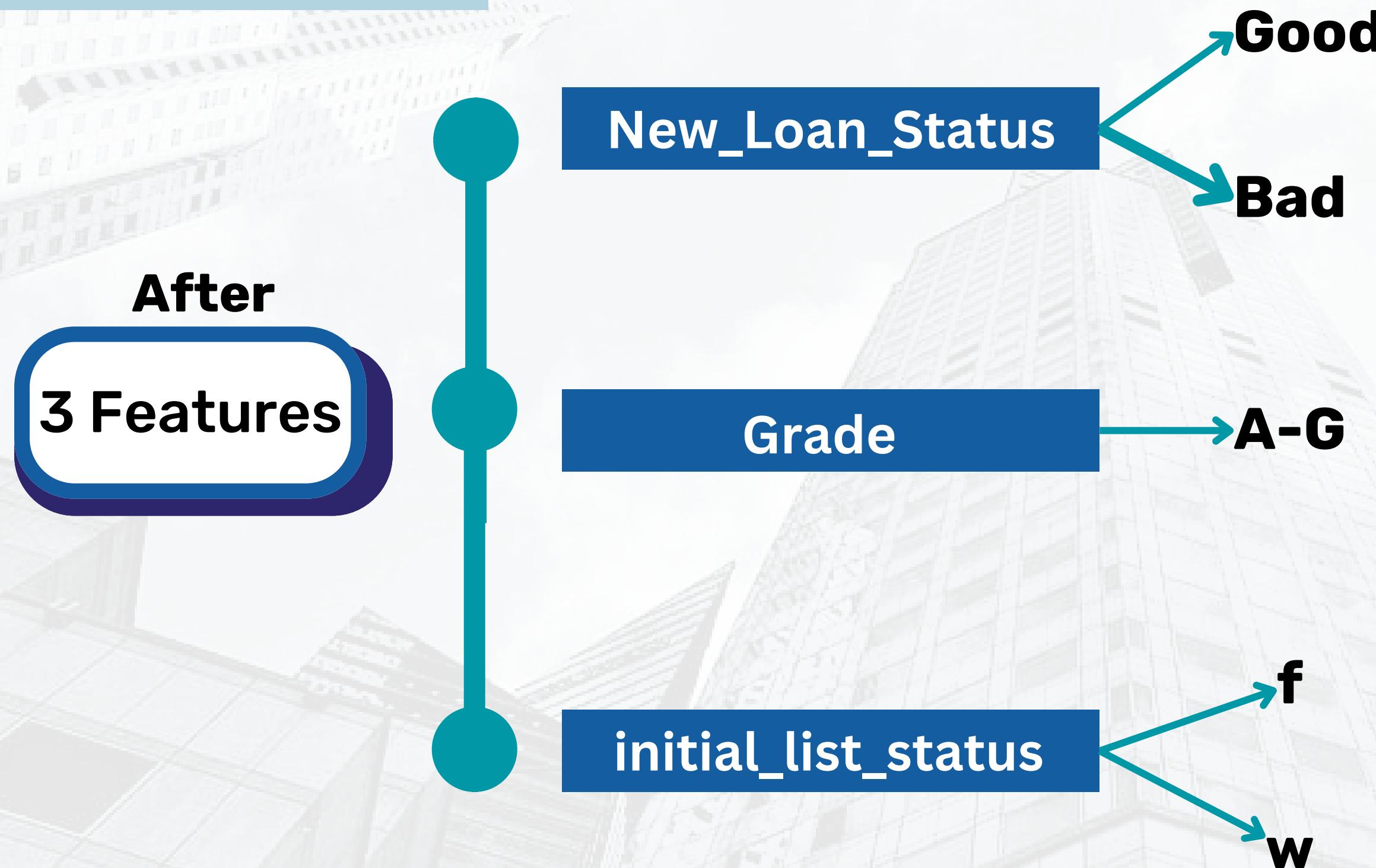


Permasalahan: Distribusi data menunjukkan banyak skewness, mengindikasikan adanya outlier yang dapat memengaruhi analisis dan performa model.

Pendekatan: Menggunakan metode Interquartile Range (IQR) untuk mengidentifikasi dan menghapus outlier.

Data Preprocessing 2

Encoding



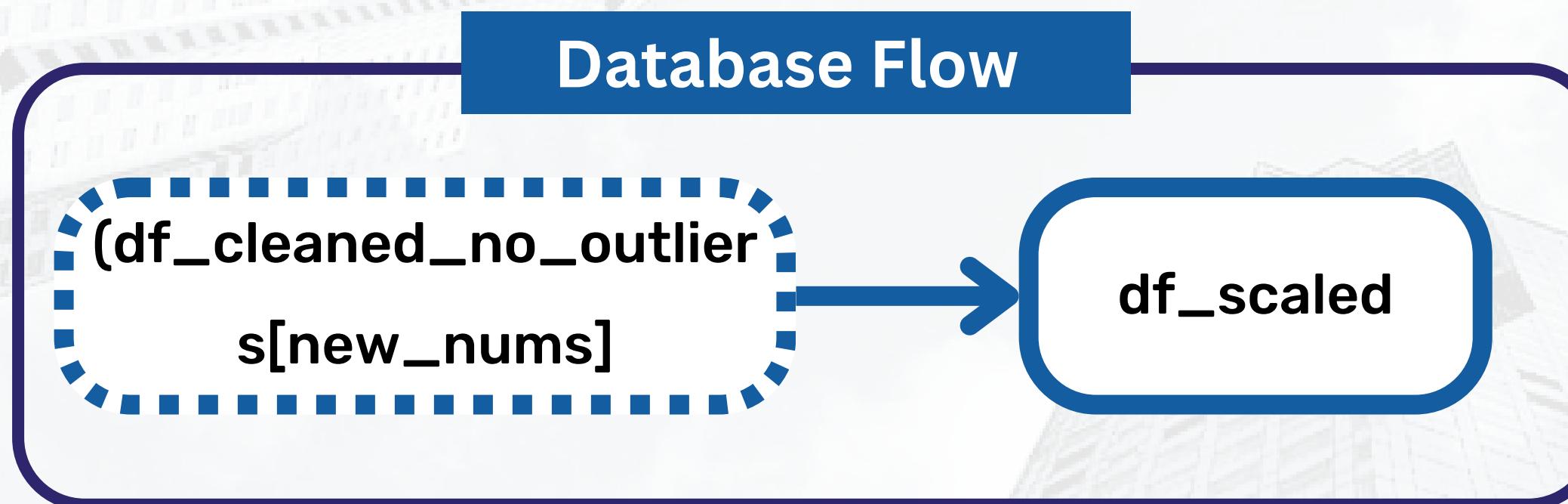
label
encoding

one hot
encoding

label
encoding

Data Preprocessing 2

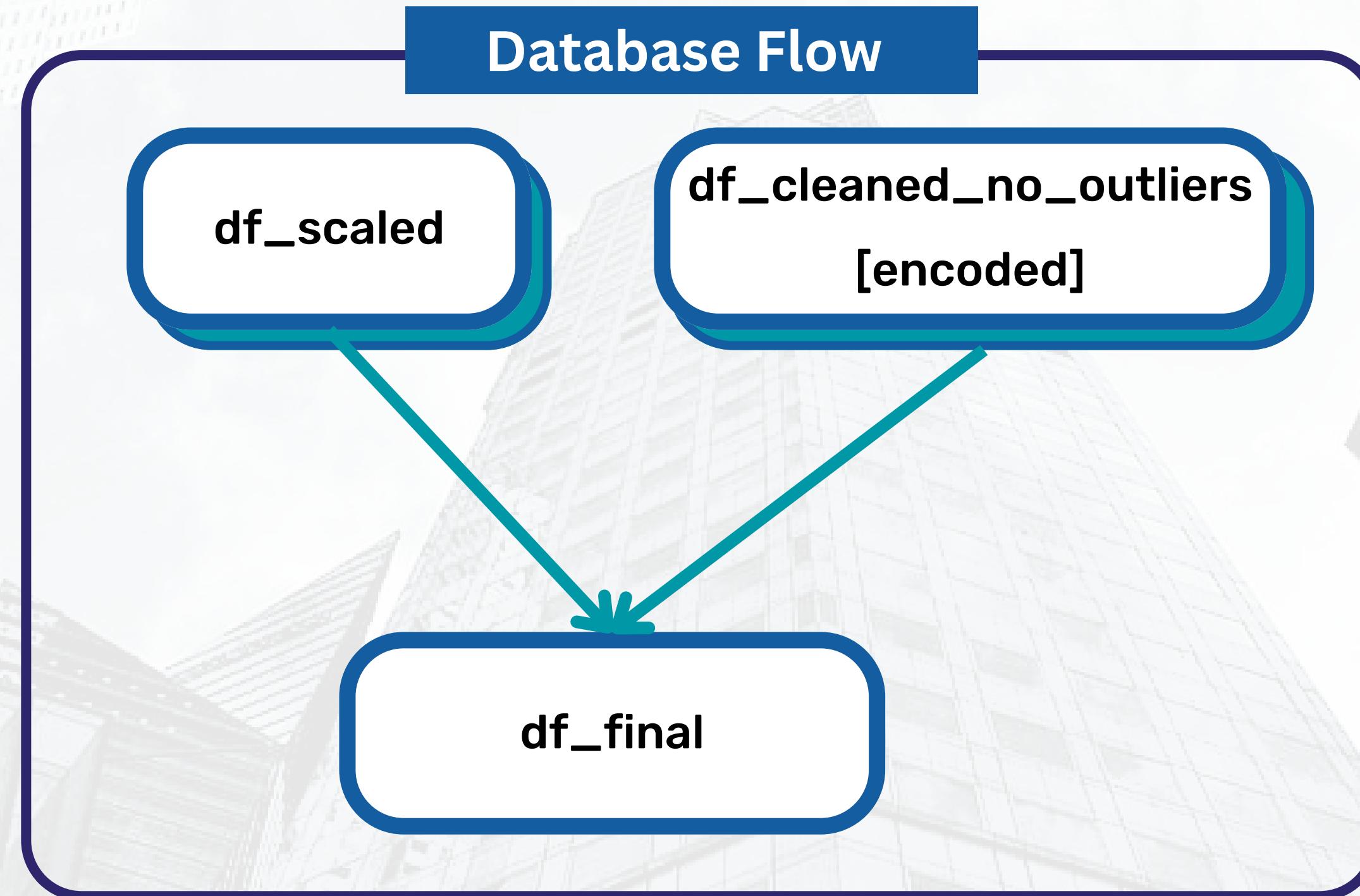
Scalling



	loan_amnt	term	int_rate	annual_inc	dti	delinq_2yrs	inq_last_6mths	open_acc	pub_rec	revol_bal	revol_util	total_acc	out_prncp	total_rec_late_fee	collections_12_mth
0	-1.128362	-0.558588	-0.586402	-1.396056	1.334819	0.0	0.643142	-1.725316	0.0	-0.017721	1.103457	-1.322554	-0.669343	0.0	
1	-1.486482	-0.558588	0.667868	-1.830154	-1.088492	0.0	2.057621	-1.961219	0.0	-1.200738	1.718519	-1.222502	-0.669343	0.0	
2	-1.403839	1.790229	-0.104535	0.673187	0.091800	0.0	-0.771337	1.105514	0.0	1.546247	-0.134978	1.578973	-0.506458	0.0	
3	-0.852885	1.790229	0.667868	-0.546040	0.804840	0.0	0.643142	-0.781706	0.0	0.433489	1.182418	-1.122449	-0.268099	0.0	
4	-1.403839	-0.558588	1.300909	-0.509238	-1.519901	0.0	2.057621	-1.489413	0.0	-0.618192	1.261378	-1.822818	-0.669343	0.0	

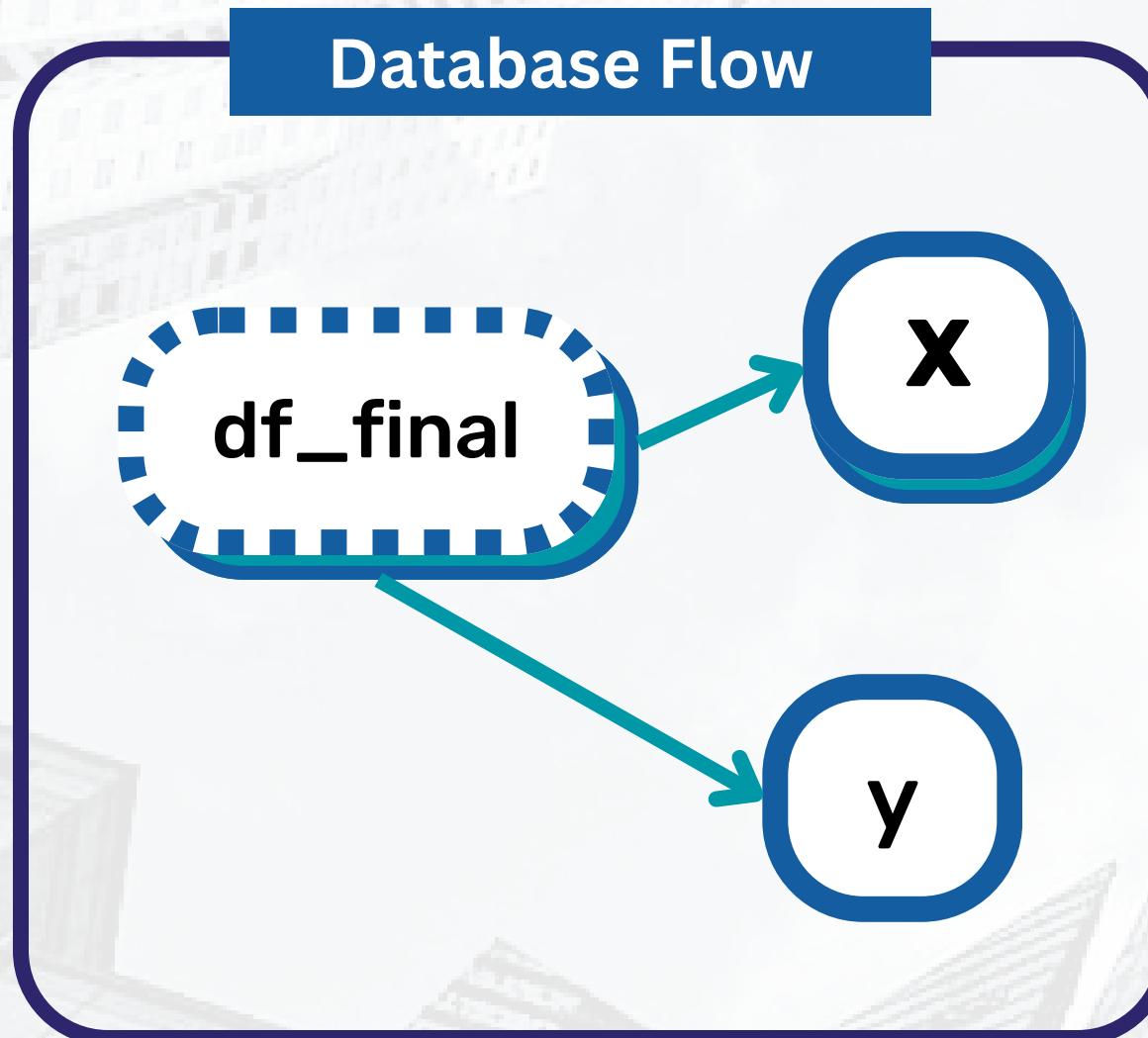
Data Preprocessing 2

Combine Data



Split & Class Imbalance

Train and Test Data



Data dibagi menjadi variabel independen (x) dan target (y/status pinjaman), lalu dipisahkan 80% untuk pelatihan dan 20% untuk pengujian.

SMOTE

	before	after
0	98310	98310
1	82821	98310

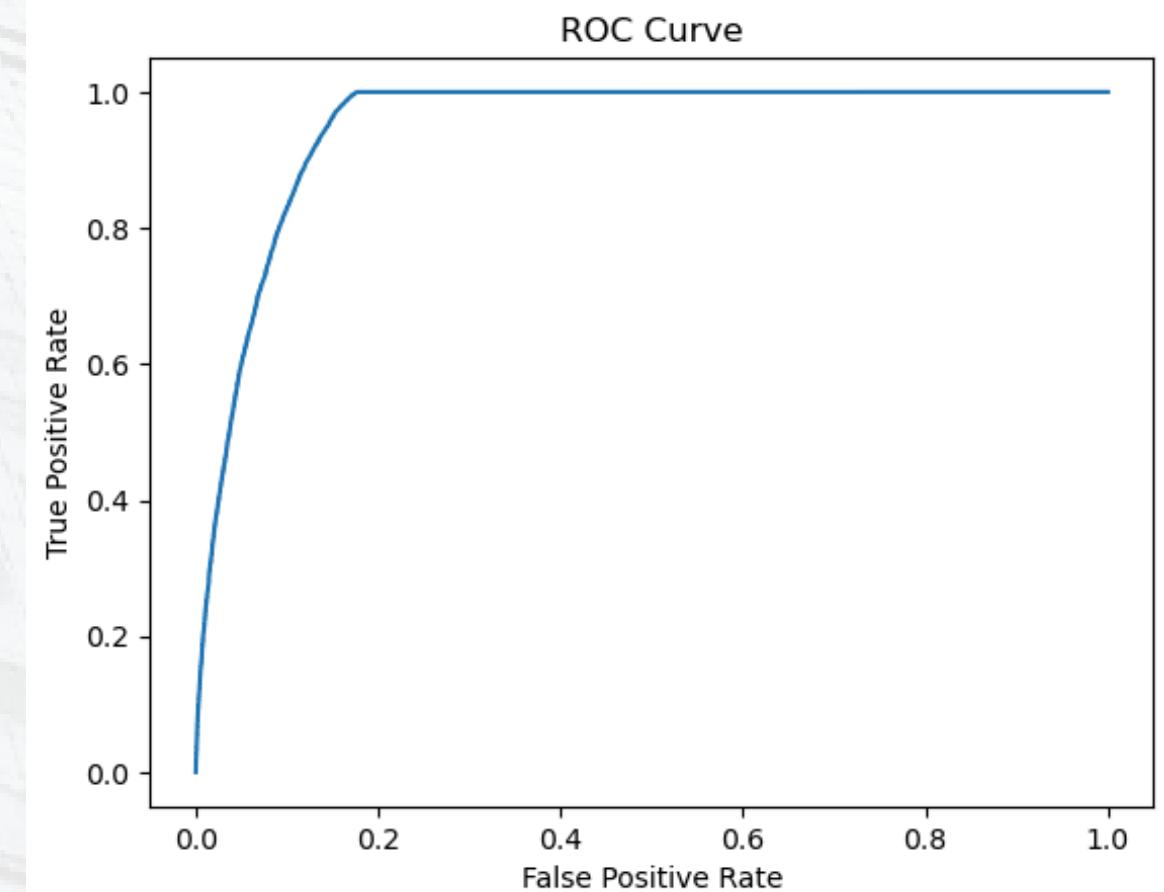
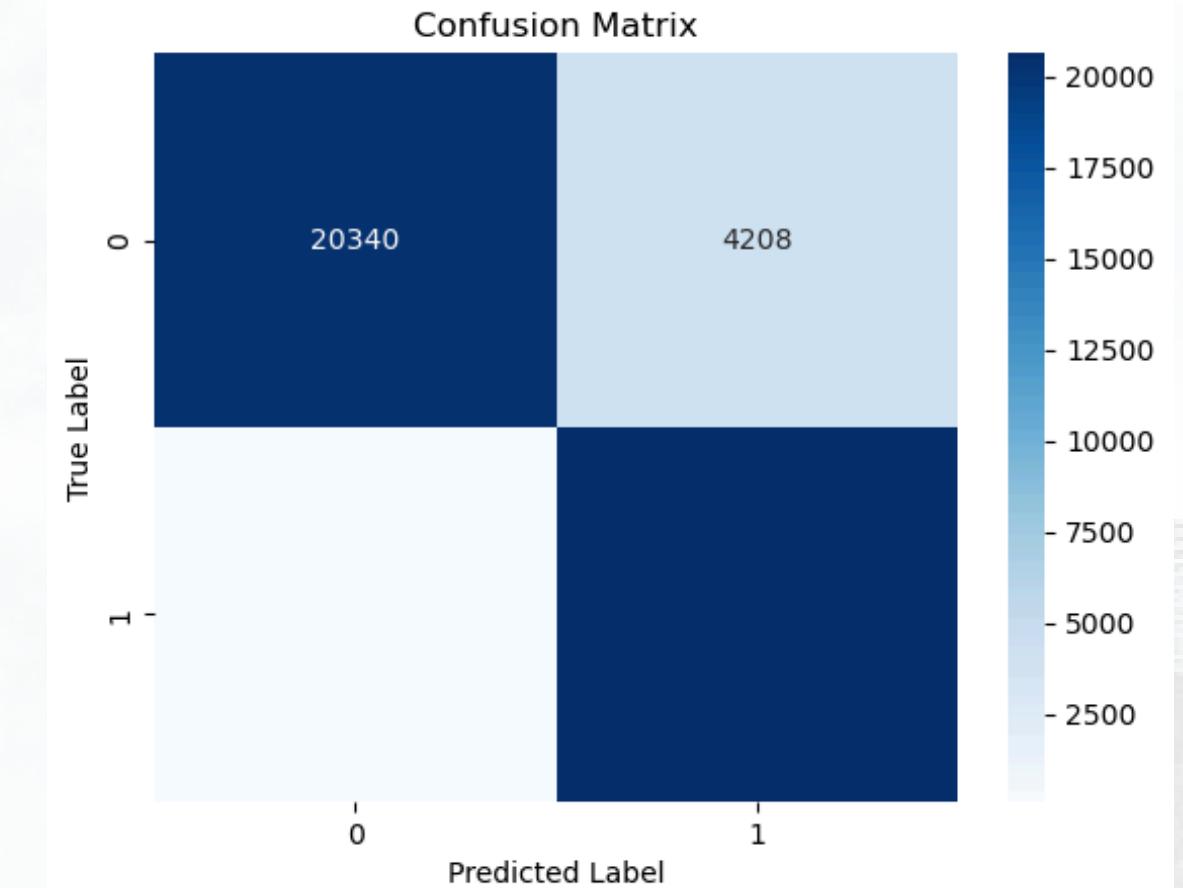
Data train (new_loan_status)/status pinjam di-oversample dengan SMOTE untuk menyeimbangkan kelas.

Modeling

Logistic Regression

	precision	recall	f1-score	support
0	1.00	0.83	0.90	24548
1	0.83	1.00	0.91	20735

- **Akurasi: 90,5%** menunjukkan bahwa model ini **cukup baik** dalam mengklasifikasikan data.
- **Precision (83,1%)**: Model **cukup baik dalam mengurangi prediksi positif palsu**.
- **Recall (99,5%)**: Sangat tinggi, berarti **hampir semua kasus positif berhasil terdeteksi**. Cocok untuk masalah di mana mendeteksi semua kasus positif lebih penting dibandingkan risiko positif palsu.
- **F1 Score (90,5%)**: Menunjukkan **keseimbangan yang baik** antara precision dan recall.

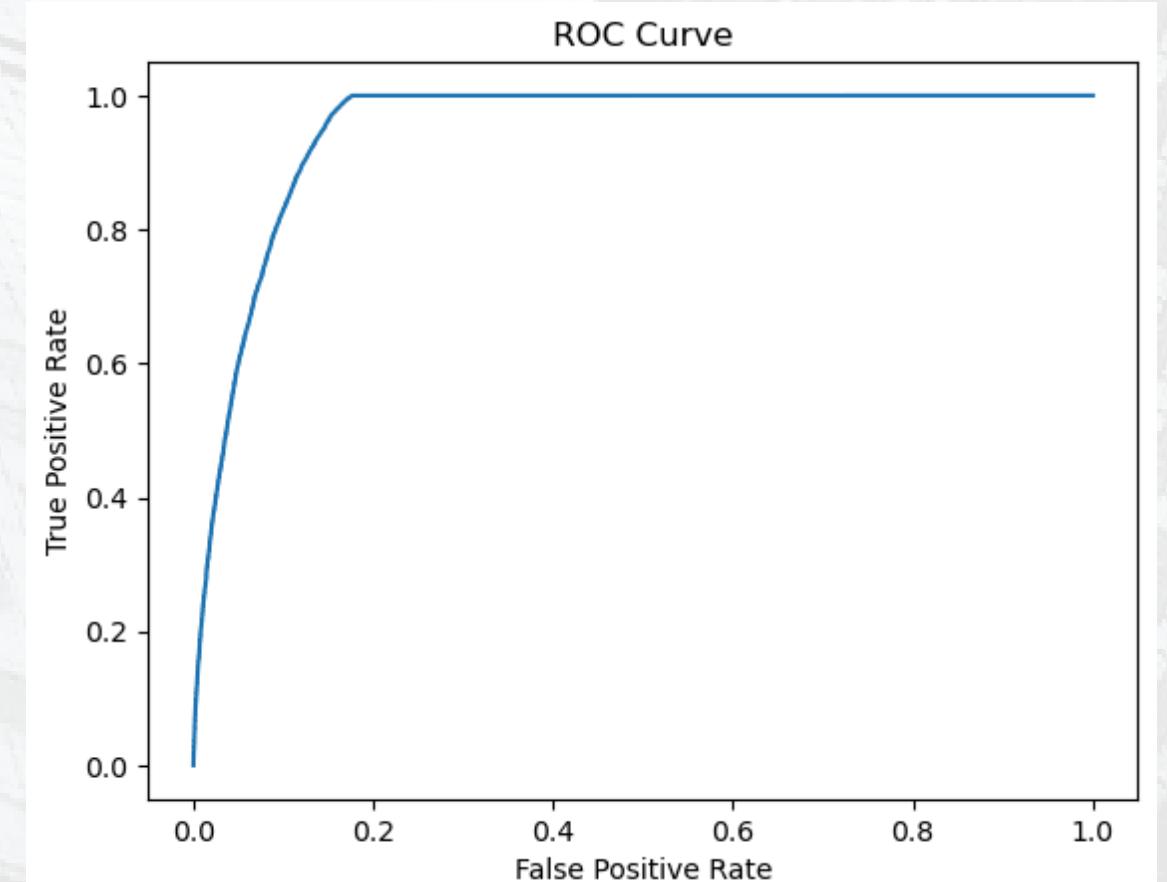
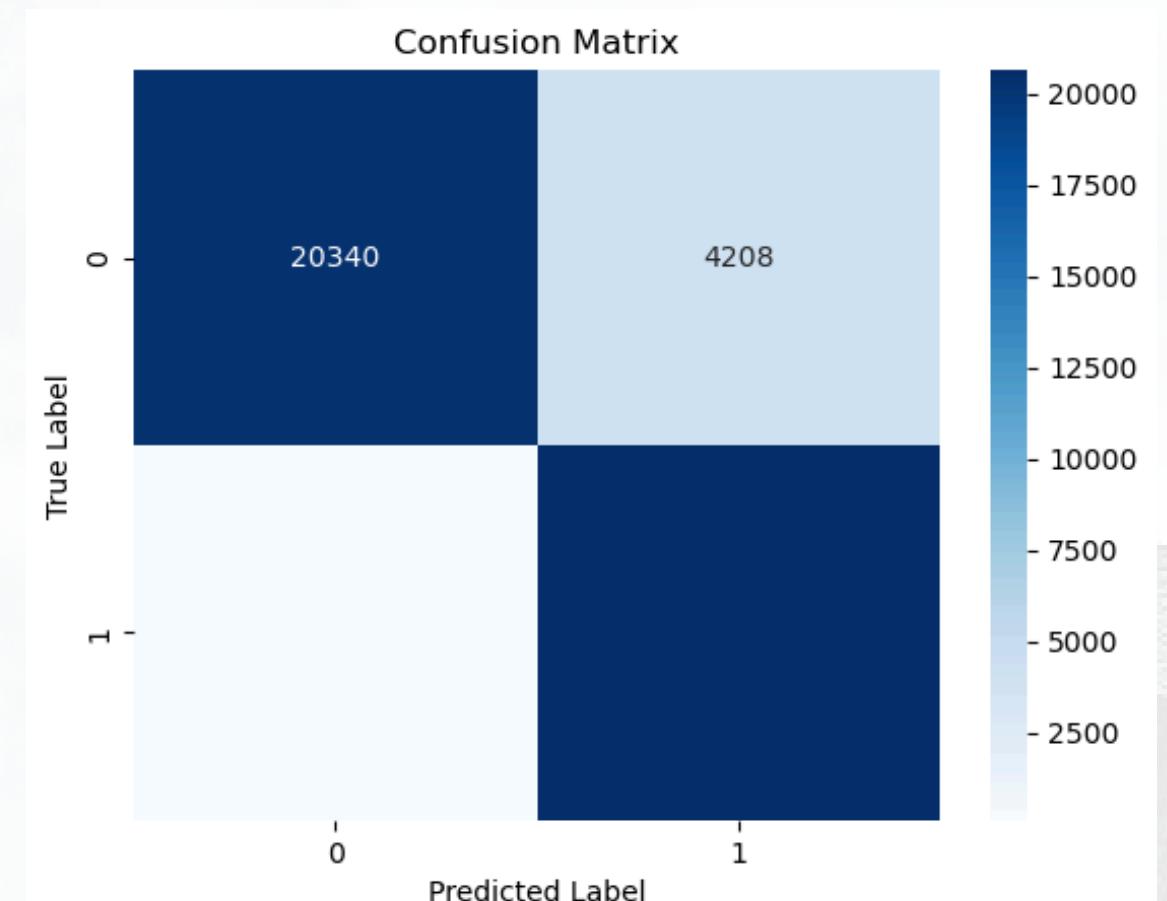


Modeling

Random Forest

	precision	recall	f1-score	support
0	0.99	0.84	0.91	24548
1	0.84	0.99	0.90	20735

- **Akurasi (90,5%):** Sama dengan Logistic Regression, menunjukkan kinerja model yang konsisten.
- **Precision (83,5%):** Sedikit lebih baik dibanding Logistic Regression dalam menghindari positif palsu.
- **Recall (98,8%):** Sedikit lebih rendah dibanding Logistic Regression, tetapi tetap sangat tinggi.
- **F1 Score (90,5%):** Sama dengan Logistic Regression, menunjukkan keseimbangan precision dan recall yang serupa.



Conclusion

	Model	Akurasi	Precision	Recall	F1 Score
0	Logistic Regression	0.905	0.831	0.995	0.905
1	Random Forest	0.905	0.835	0.988	0.905

- **Kedua model menunjukkan performa yang kuat dan konsisten**, dengan perbedaan yang sangat minimal pada precision dan recall. Logistic Regression dan Random Forest sama-sama bisa diandalkan untuk berbagai skenario klasifikasi, dan **pemilihan model lebih bergantung pada kebutuhan spesifik** dari masalah yang ingin diselesaikan.

• Logistic Regression cocok untuk **kasus yang sangat sensitif terhadap recall**, seperti diagnosis medis, di mana mendekripsi semua kasus positif (**false negative minimal**) adalah prioritas utama.

• Random Forest unggul dalam aplikasi yang membutuhkan **precision lebih tinggi**, seperti deteksi penipuan atau spam, di mana mengurangi **false positive menjadi prioritas**.

Thank You



Rakamin
Academy



id/x partners