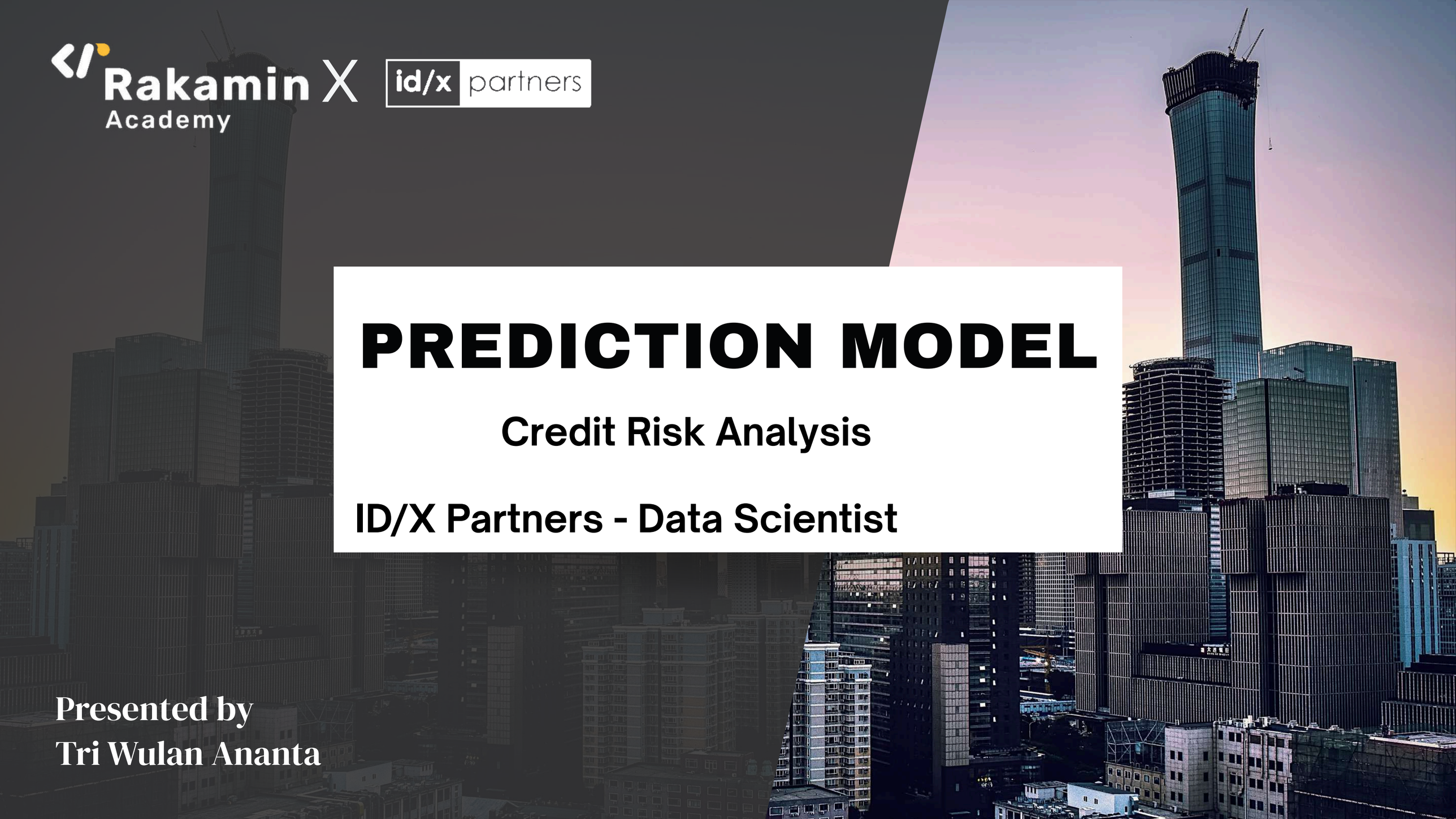
Rakamin Academy X id/x partners

PREDICTION MODEL

Credit Risk Analysis

ID/X Partners - Data Scientist



Presented by
Tri Wulan Ananta

Tri Wulan Ananta

Lulusan baru Sistem Informasi , Universitas Budi Luhur.

Saya memiliki ketertarikan dalam bidang data, bisnis, teknologi, pembelajaran bahasa.

Data analisis & data science.

Memiliki keterampilan dalam menggunakan Python (pustaka pandas, numpy, seaborn, matplotlib, sklearn) , SQL, Excel. Visualisasi data dengan tools (Power BI, Tableau).

Jenjang karir :
1. Business Development - PT Indo Sukses Pratama
Fulltime 2023 - present

Project based
Website engineer site : www.indosuksespratama.co.id

2. AI testing - Outlier
Freelance Dec 2024 - Jan 2025

wulananta08@gmail.com Triwulan Ananta

COURSES & CERTIFICATION

- Basic Computer Algorithm Competency - Univ. Budi Luhur, 2022
- MonsoonSIM ERP Courses - MonsoonSIM, 2022
- Generative AI for System Information Google Cloud - Google, 2024
- Data Visualization with Power BI - Unpad, 2025
- Fundamental Data Science - Digitalent, 2025
- Intermediate Data Science - Digitalent, 2025

BACKGROUND

Id/x partners, berdiri sejak 2002, perusahaan konsultan yang berfokus pada data analytics, manajemen risiko, dan solusi pengambilan keputusan (decisioning solutions). Bisnis inti mencakup:

- Credit & Risk Management – merancang strategi siklus kredit, membangun model scoring, dan kerangka kerja risiko.
- Data Analytics & Decisioning – mengubah data menjadi insight yang dapat ditindaklanjuti melalui model prediktif dan decision engine.
- Process & Performance Optimization – meningkatkan efisiensi, profitabilitas, dan kualitas pengambilan keputusan.
- Data-Driven Marketing – memperkuat akuisisi, retensi, serta pertumbuhan portofolio pelanggan.

Id/x partners memiliki pengalaman luas di Asia dan Australia di berbagai industri seperti jasa keuangan, telekomunikasi, manufaktur, dan ritel. Dengan menggabungkan konsultasi strategis dan solusi teknologi, Id/x partners hadir sebagai mitra terpadu (one-stop partner) untuk mendukung pengambilan keputusan bisnis yang lebih cerdas dan berbasis data.

PROJECT PORTFOLIO

Sebagai Data Scientist di ID/X Partners, klien meminta solusi untuk meningkatkan akurasi dalam menilai serta mengelola risiko kredit. Tujuannya adalah membantu mereka mengoptimalkan keputusan bisnis sekaligus meminimalkan potensi kerugian. Tugas saya adalah membangun model machine learning yang mampu memprediksi risiko kredit dengan memanfaatkan dataset pinjaman yang mencakup data pengajuan yang disetujui maupun ditolak.

Link code here!

Link code here!

Github

Project explanation video here!

DATA UNDERSTANDING

Data memiliki 466285 baris & 75 kolom dengan data tipe numerik float, int64. lalu kategorikal, object

STUKTUR DATASET

```
Jumlah baris dan kolom: (466285, 75)
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 466285 entries, 0 to 466284
Data columns (total 75 columns):
```

```
dtypes: float64(46), int64(7), object(22)
memory usage: 266.8+ MB
```

```
=== Kolom Full Missing ===
                      column  missing_count  missing_pct
54           annual_inc_joint         466285        100.0
55                    dti_joint         466285        100.0
56  verification_status_joint         466285        100.0
60                  open_acc_6m         466285        100.0
61                   open_il_6m         466285        100.0
62                  open_il_12m         466285        100.0
63                  open_il_24m         466285        100.0
64         mths_since_rcnt_il         466285        100.0
65                total_bal_il         466285        100.0
66                     il_util         466285        100.0
67                  open_rv_12m         466285        100.0
68                  open_rv_24m         466285        100.0
69                   max_bal_bc         466285        100.0
70                     all_util         466285        100.0
72                       inq_fi         466285        100.0
73                 total_cu_tl         466285        100.0
74                inq_last_12m         466285        100.0
```

```
=== Kolom Sebagian Besar Missing (>30%) ===
                      column  missing_count  missing_pct
20                      desc         340304    72.981975
29       mths_since_last_delinq         250351    53.690554
30       mths_since_last_record         403647    86.566585
48                 next_pymnt_d         227214    48.728567
51  mths_since_last_major_derog         367311    78.773926

=== Kolom Sedikit Missing (<=30%) ===
                      column  missing_count  missing_pct
11                  emp_title          27588     5.916553
12                 emp_length          21008     4.505399
14                 annual_inc              4     0.000858
22                      title             21     0.004504
26                delinq_2yrs             29     0.006219
27           earliest_cr_line             29     0.006219
28             inq_last_6mths             29     0.006219
31                   open_acc             29     0.006219
32                    pub_rec             29     0.006219
34                 revol_util            340     0.072917
35                  total_acc             29     0.006219
46               last_pymnt_d            376     0.080637
49         last_credit_pull_d             42     0.009007
50  collections_12_mths_ex_med            145     0.031097
57             acc_now_delinq             29     0.006219
58               tot_coll_amt          70276    15.071469
59                tot_cur_bal          70276    15.071469
71          total_rev_hi_lim          70276    15.071469
```

ANALISIS MISSING VALUE

Terdapat kolom

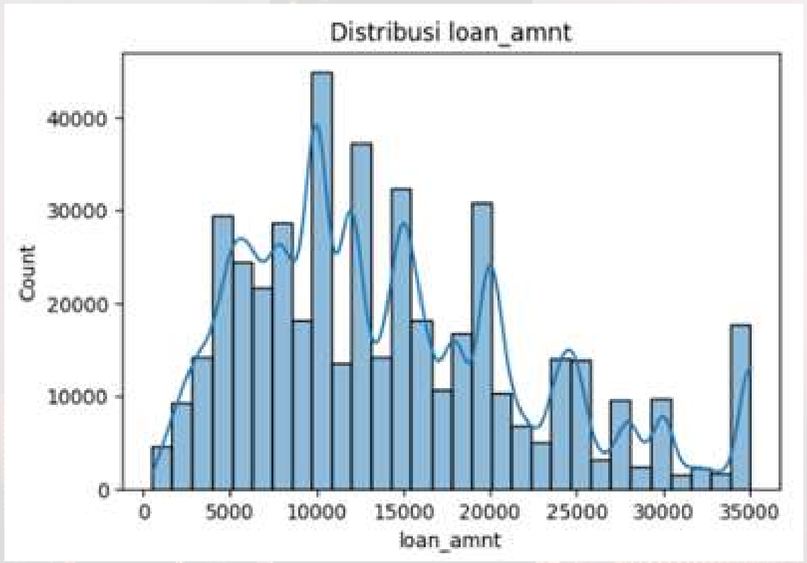
- **Full missing 100%**
- **Sebagian missing**
- **Sedikit missing**

EXPLORATORY DATA ANALYSIS (EDA)

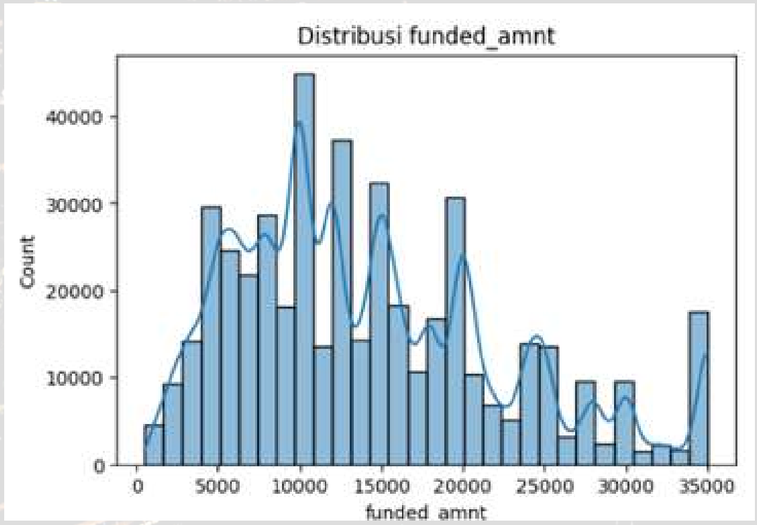
Distribusi variabel

Jumlah pinjaman > loan_amnt

- Sebagian besar pinjaman berada pada kisaran 5.000 – 20.000.
- Ada puncak sekitar 10.000 – 15.000, menunjukkan banyak nasabah mengajukan pinjaman di range tersebut.
- Distribusi tampak multimodal (ada beberapa puncak).
- Nilai ekstrim terlihat pada pinjaman > 30.000, tapi tidak terlalu jauh.



"loan_status akan menjadi fitur utama analisis"



Distribusi variabel

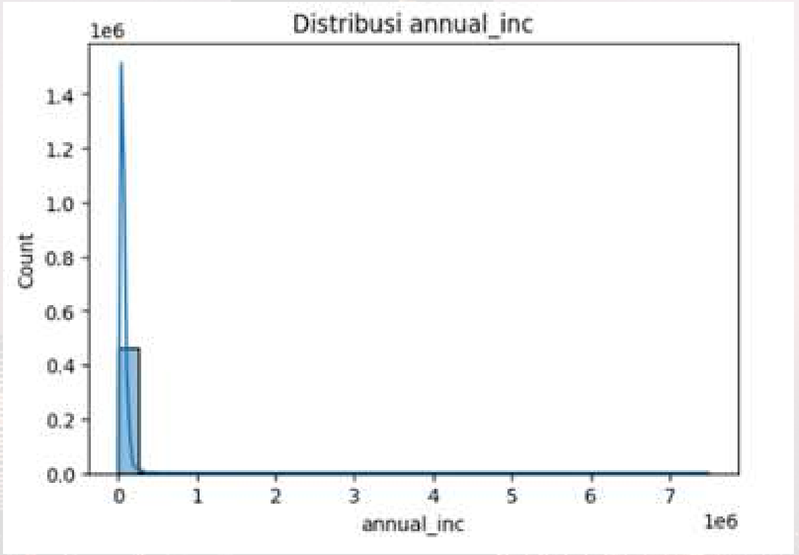
Jumlah pinjaman yang disetujui > funded_amnt

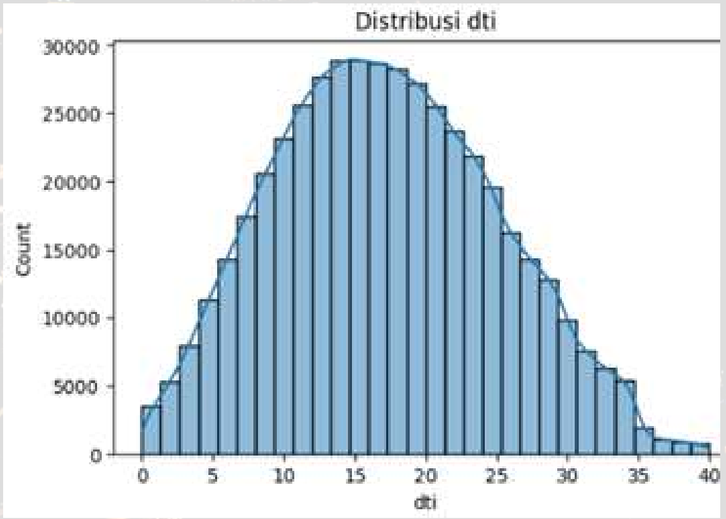
- Bentuk distribusinya hampir identik dengan loan_amnt, karena biasanya jumlah yang diajukan mendekati jumlah yang disetujui.
- Puncak tetap di sekitar 10.000 – 15.000.

Distribusi variabel

Pendapatan tahunan > annual_inc

- Mayoritas pendapatan tahunan nasabah berada pada rentang < 200.000.
- Namun terlihat ada nilai sangat besar hingga lebih dari 7 juta.
- Ini jelas merupakan outlier ekstrem, karena jauh berbeda dengan mayoritas distribusi.
- → Bisa jadi data error (typo) atau memang ada segelintir nasabah berpendapatan sangat tinggi.



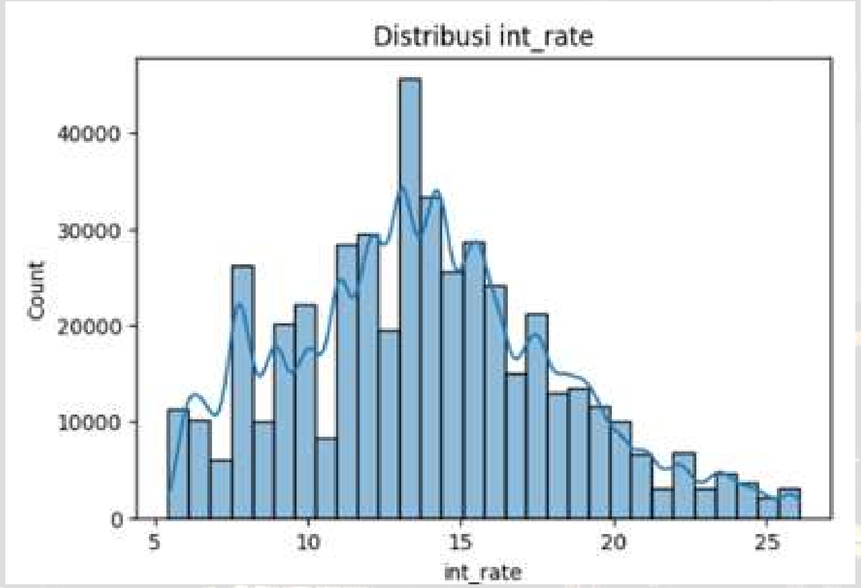


Distribusi variabel

DTI > Debt-to-Income ratio

- Distribusi berbentuk seperti kurva lonceng (mendekati normal).
- Sebagian besar nilai dti berada pada rentang 10 – 25.
- Ada beberapa nasabah dengan dti mendekati 40, ini agak jarang tapi masih mungkin terjadi → outlier moderat.

Distribusi variabel

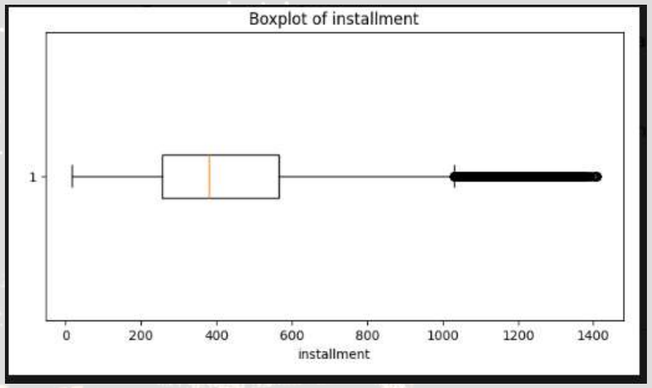


Suku bunga pinjaman > int_rate

- Sebagian besar bunga pinjaman berada di kisaran 10% – 15%, dengan puncak tertinggi sekitar 13% – 14%.
- Distribusi berbentuk mendekati normal tetapi condong ke kanan (right-skewed).
- Ada nilai bunga di atas 25%, yang bisa dianggap outlier signifikan, karena jauh dari mayoritas distribusi.

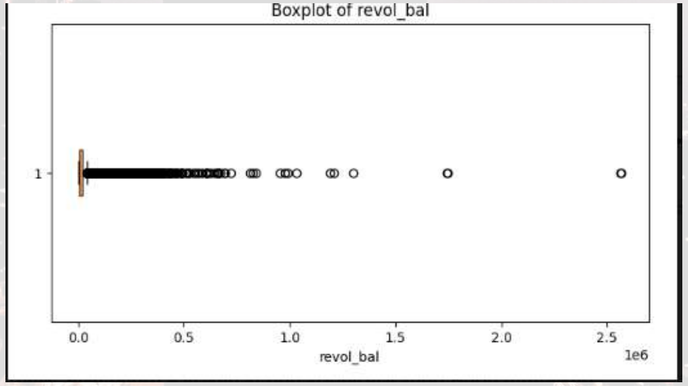
Boxplot

Boxplot of installment

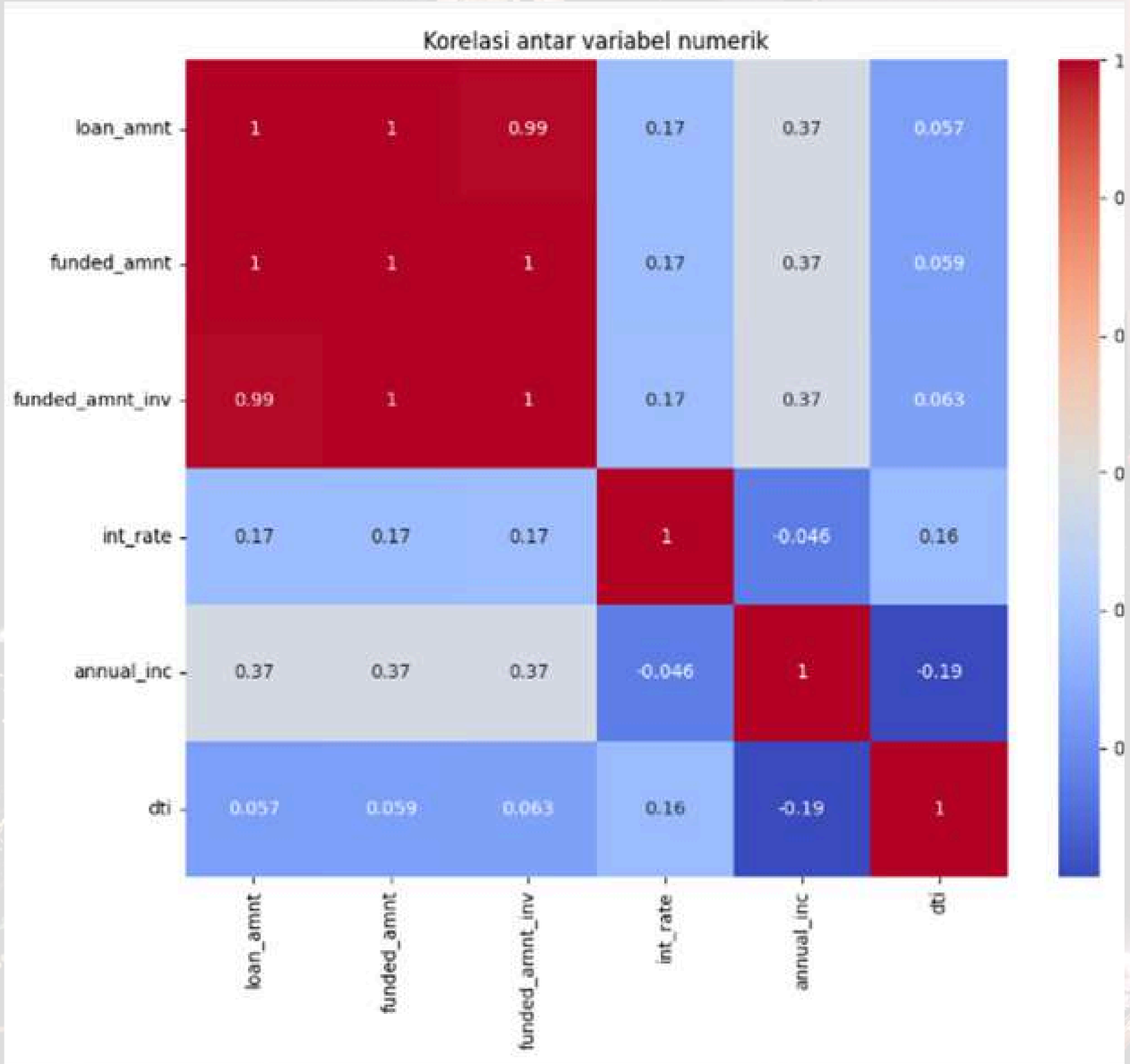


- Boxplot ini menunjukkan distribusi nilai installment (angsuran bulanan).
- Mayoritas data angsuran berada di kisaran 200–600, terlihat dari panjang box (IQR).
- Ada outlier di sisi kanan (nilai lebih besar dari ~1000), yang ditandai titik-titik hitam.
- Artinya, sebagian kecil peminjam memiliki cicilan jauh lebih tinggi dibanding mayoritas.

Boxplot of revol_bal



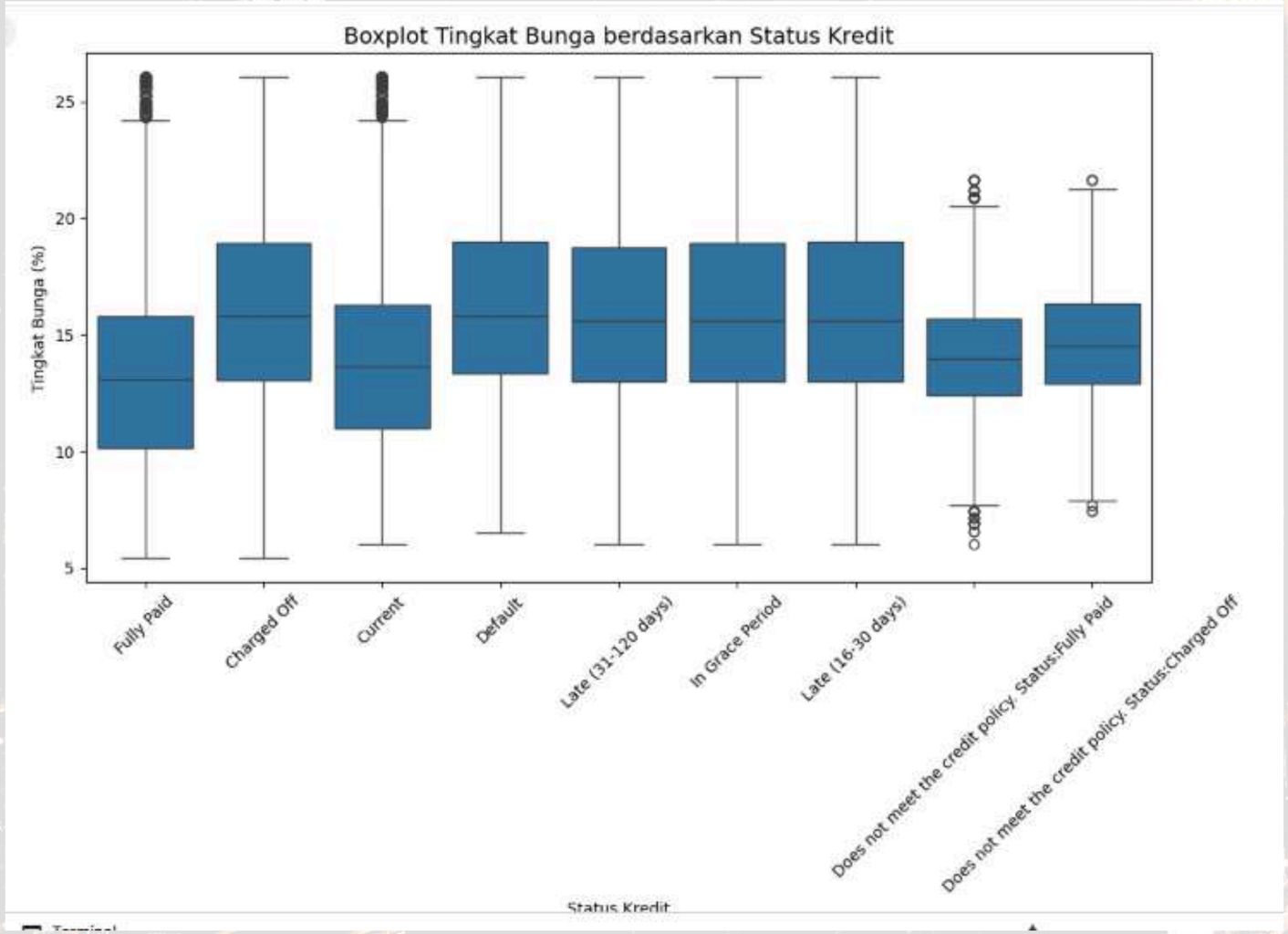
- Sebagian besar bunga pinjaman berada di kisaran 10% – 15%, dengan puncak tertinggi sekitar 13% – 14%.
- Distribusi berbentuk mendekati normal tetapi condong ke kanan (right-skewed).
- Ada nilai bunga di atas 25%, yang bisa dianggap outlier signifikan, karena jauh dari mayoritas distribusi.



Heatmap Korelasi antar Variabel Numerik

- **Beberapa variabel sangat multikolinear (loan_amnt, funded_amnt, funded_amnt_inv), sangat berkorelasi tinggi (hampir 1.0), artinya informasi di antaranya hampir sama/redundan.**

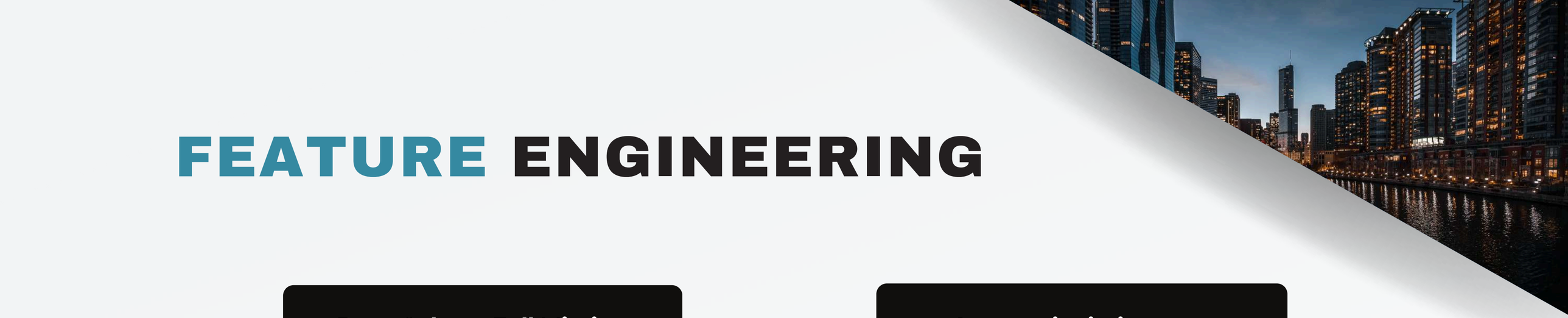
Boxplot Tingkat Bunga berdasarkan Status Kredit



- **Pinjaman dengan status bermasalah (Charged Off, Default, Late) cenderung memiliki tingkat bunga lebih tinggi.**

- **Pinjaman yang lancar/lunas (Fully Paid, Current) lebih sering memiliki bunga lebih rendah.**

Tingkat bunga bisa menjadi predictor penting dalam memodelkan risiko kredit, karena terlihat ada pola antara tinggi bunga dan kemungkinan gagal bayar



FEATURE ENGINEERING

Drop Coloms Full Missing

Imputasi missing

Defining Target Variabel

Data cleaning

DATA PREPROCESSING

Project 1

- Drop Full Missing Coloms

```
Shape setelah drop: (466285, 58)
```

setelah di hapus shape menjadi 466285 baris, 58 kolom.

Project 2

- Drop >60% Missing

```
Shape setelah drop >60% missing: (466285, 55)
```

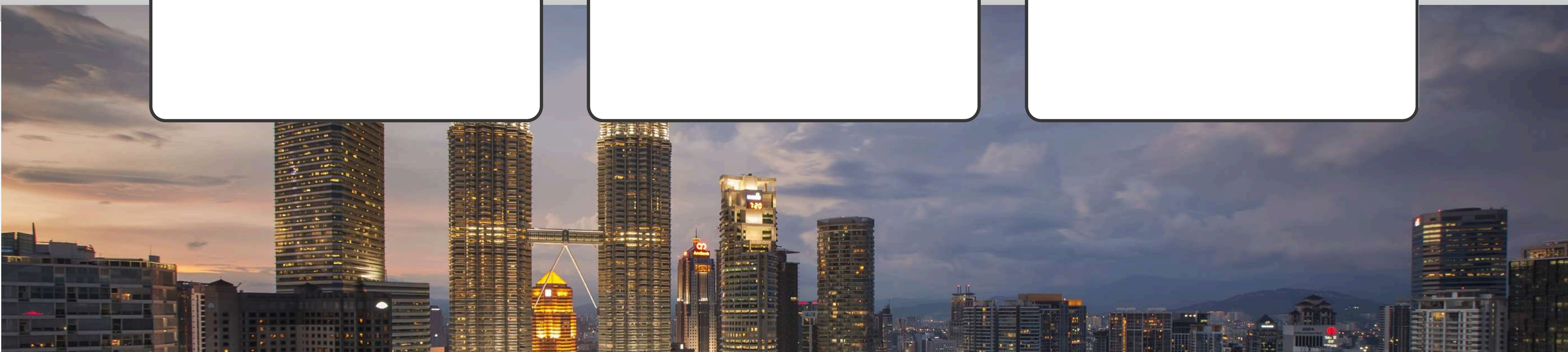
setelah di hapus shape menjadi 466285 baris, 55 kolom.

Project 3

- Imputasi Missing

Numerik → median

Kategorikal → "Unknown"



DATA PREPROCESSING

Project 4

Tangani Outlier

pada outlier_cols

"annual_inc"
"dti"
"revol_bal"
"loan_amnt"
"installment"
"int_rate"

Project 5

Tangani Skewness

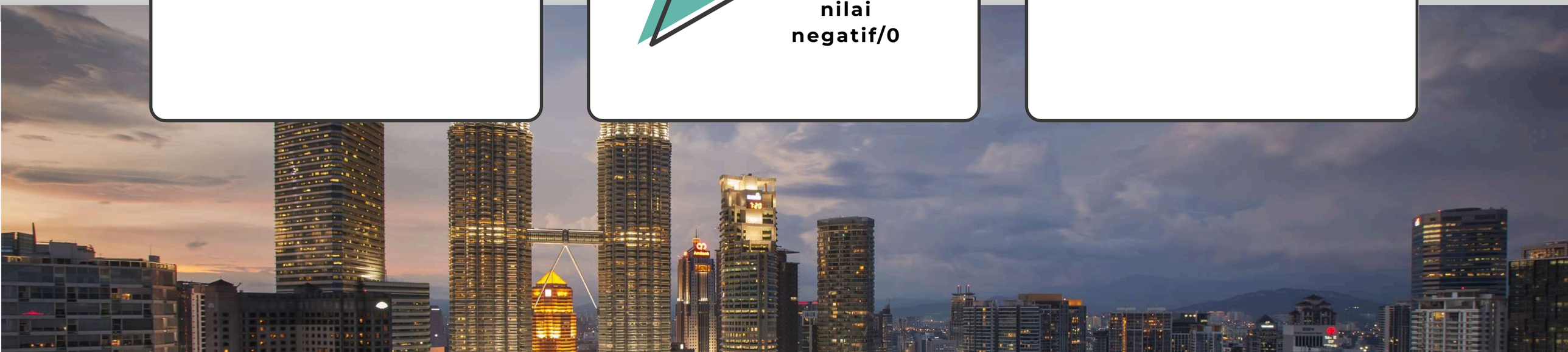
log1p kalau semua nilai positif

Yeo-Johnson kalau ada nilai negatif/0

Project 6

Hapus Redundan

pada ["funded_amnt", "funded_amnt_inv"]



DATA PREPROCESSING

Project 7

- Menentukan Data Target

Dari kolom loan_status, dibuat kolom baru bernama target.

nilai 1 = berisiko/gagal bayar
"Charged Off"
"Default"
“Late (31-120 days)”
“Late (16-30 days)”
“In Grace Period”

nilai 0 = tidak berisiko
“Fully Paid”
“Current”

Project 8

- Split Data

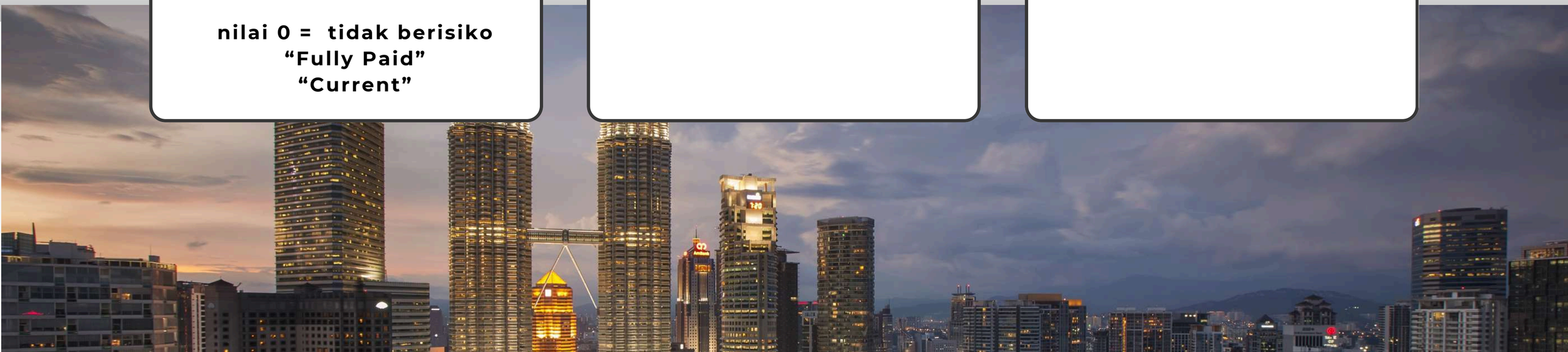
```
Train shape: (326399, 74)
Test shape: (139886, 74)
```

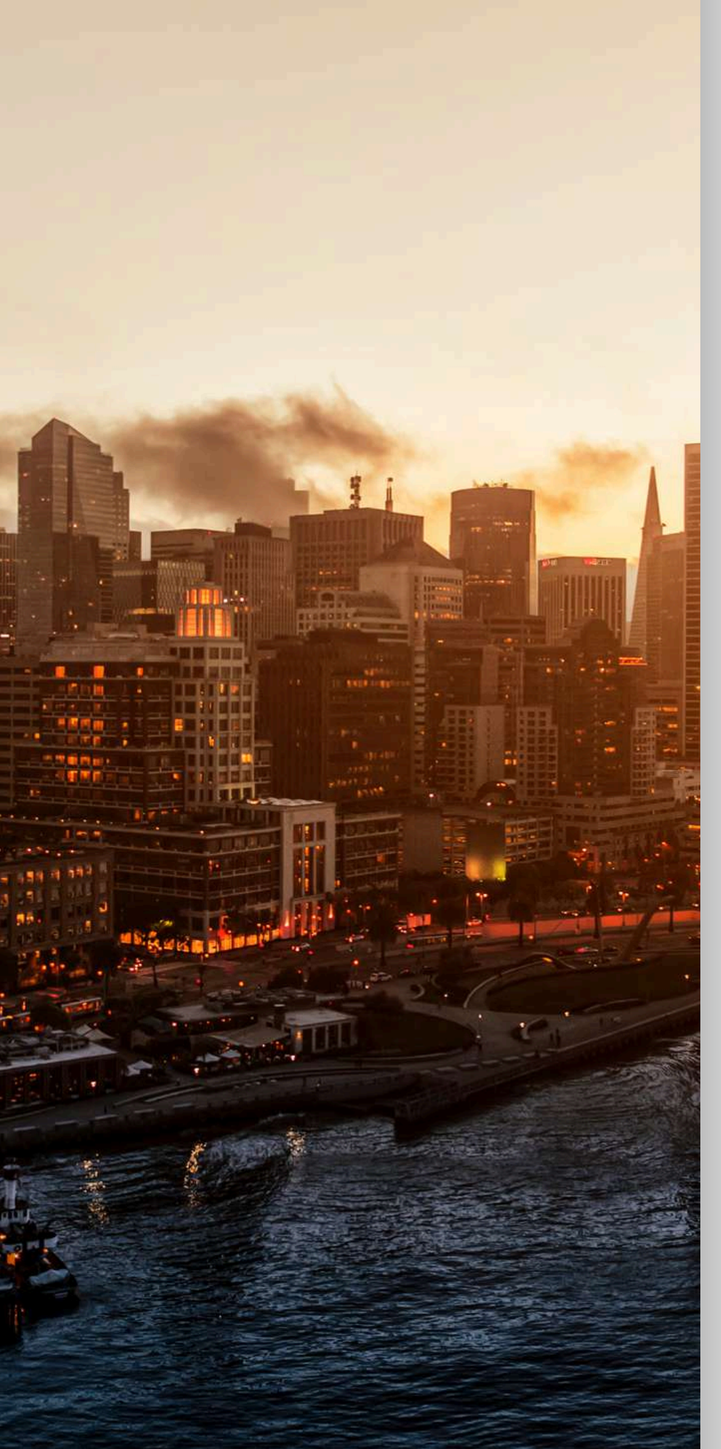
Project 9

- scaling / normalisasi

OneHotEncoder
StandardScaler
ColumnTransformer

```
Shape sebelum preprocessing: (326399, 74)
Shape sesudah preprocessing: (326399, 526924)
```





DATA PREPROCESSING

Converting Datetime

	issue_d_month	last_pymnt_d_month	next_pymnt_d_month	\
0	0	0	0	
1	0	0	0	
2	0	0	0	
3	0	0	0	
4	0	0	0	

	last_credit_pull_d_month	earliest_cr_line_month
0	0	1
1	0	4
2	0	0
3	0	2
4	0	1

Ekstrak bulan (1-12), missing jadi 0

Beberapa fitur waktu diubah dalam bentuk numerik agar modelling dapat memprediksi

DATA MODELLING

```
=== Logistic Regression ===
[[123353    162]
 [  1842  14529]]
              precision    recall  f1-score   support

           0       0.99      1.00      0.99    123515
           1       0.99      0.89      0.94     16371

    accuracy                           0.99    139886
   macro avg       0.99      0.94      0.96    139886
weighted avg       0.99      0.99      0.99    139886

ROC-AUC: 0.986885254300783
```

- Accuracy: 0.99
- ROC-AUC: 0.986
- Kelas 0: precision 0.99, recall 1.00
- Kelas 1: precision 0.99, recall 0.89 (sangat baik, recall cukup tinggi)

Kesimpulan: Model seimbang, generalisasi bagus. Cocok sebagai baseline yang stabil. Tidak overfit/underfit.

DATA MODELLING

DecisionTree Classifier

- Best params: min_samples_split=5, min_samples_leaf=1, max_depth=5
- Accuracy: 0.95
- ROC-AUC: 0.945
- Kelas 0: precision 0.95, recall 1.00 (sangat bagus untuk kelas mayoritas)
- Kelas 1: precision 0.97, recall 0.61 → banyak miss pada prediksi positif (loan default), recall rendah artinya banyak false negative.

Kesimpulan: Model ini gagal menangkap pola minoritas. Walaupun accuracy tinggi, f1-score untuk kelas minoritas rendah → underfitting.

```
Best params: {'model__min_samples_split': 5, 'model__min_samples_leaf': 1, 'model__max_depth': 5}
              precision    recall  f1-score   support

           0       0.95      1.00      0.97    123515
           1       0.97      0.61      0.75     16371

    accuracy                           0.95    139886
   macro avg       0.96      0.81      0.86    139886
weighted avg       0.95      0.95      0.95    139886

ROC-AUC: 0.9454115010940566
```

DATA MODELLING

RandomForest Classifier

```
Best params: {'model__max_depth': 10, 'model__max_features': 'sqrt', 'model__min_samples_leaf': 1, 'model__min_samples_split': 10, 'model__n_estimators': 171}
              precision    recall  f1-score   support

           0       0.97      0.92      0.95    123515
           1       0.57      0.77      0.66     16371

    accuracy                           0.91    139886
   macro avg       0.77      0.85      0.80    139886
weighted avg       0.92      0.91      0.91    139886

ROC-AUC: 0.9268066588186957
```

- **Best params:**
max_depth=10, max_features='sqrt', min_samples_leaf=1, min_samples_split=10, n_estimators=171

- **Accuracy: 0.91**
- **ROC-AUC: 0.926**
- **Kelas 0: precision 0.97, recall 0.92**
- **Kelas 1: precision 0.57, recall 0.77 → precision sangat rendah (banyak false positive).**

Kesimpulan: Random Forest underfitting ringan → meskipun cukup menangkap pola, tapi gagal pada kelas minoritas, terlihat dari precision rendah di kelas 1.

DATA MODELLING

XGB Classifier

```
  bst.update(dtrain, iteration=i, fobj=obj)
Best params: {'model__subsample': 1.0, 'model__n_estimators': 200, 'model__max_depth': 5, 'model__learning_rate': 0.1, 'model__colsample_bytree': 1.0}
              precision    recall  f1-score   support

           0       0.99      1.00      0.99    123515
           1       0.99      0.89      0.94     16371

    accuracy                           0.99    139886
   macro avg       0.99      0.95      0.97    139886
weighted avg       0.99      0.99      0.99    139886

ROC-AUC: 0.9922445590268675
```

Best params: subsample= 1.0, n_estimators=200, max_depth=5, learning_rate=0.1, colsample_bytree=1.0

- **Accuracy: 0.99**
- **ROC-AUC: 0.992**
- **Kelas 0: precision 0.99, recall 1.00**
- **Kelas 1: precision 0.99, recall 0.89 → bagus, mirip Logistic Regression tapi sedikit lebih tinggi performanya.**

Kesimpulan: XGBoost paling powerful di sini, memberikan keseimbangan precision & recall untuk kedua kelas. Hampir tidak ada indikasi overfitting (Train-Test konsisten).

EVALUASI MACHINE LEARNING

Hasil evaluasi model dengan akurasi, presisi, recall, atau ROC-AUC, di temukan :

Best Model

- **XGBoost (paling stabil, performa tinggi, generalisasi bagus).**
- **Alternatif: Logistic Regression (ringan, hasil mendekati XGBoost).**

Kurang Baik

- **Random Forest → underfitting, bisa ditingkatkan dengan tuning (lebih banyak estimators, depth lebih dalam).**
- **Decision Tree → overfitting, bisa diatasi dengan pruning atau ubah ke ensemble lainnya.**

	Model	Train Accuracy	Test Accuracy	Train ROC-AUC	Test ROC-AUC
0	XGBoost	0.988195	0.986925	0.994776	0.992245
1	Random Forest	0.906274	0.905931	0.930763	0.926807
2	Logistic Regression	0.988385	0.985674	0.997078	0.986885
3	Decision Tree	1.000000	0.982779	1.000000	0.941606

THANK YOU!

Rakamin Academy x id/x partners