




Prediction Model

Analisis Risiko Kredit: Pemodelan dan Prediksi dengan Machine Learning

ID/X Partners - Data Scientist

Presented by
Destya Rosa Mardiana

Destya Rosa Mardiana

 **Elektronika dan Instrumentasi, Universitas Gadjah Mada**
 Data Science & Machine Learning Enthusiast
 Berpengalaman dalam Analisis Data, Machine Learning, dan IoT

SKILL :

- 1. Machine Learning & AI:** Autonomous driving ,
Computer Vision (CycleGAN, Pix2Pix, PSPNet)
- 2. Pemrograman:** Python (Pandas, Scikit-learn,
TensorFlow, PyTorch)
- 3. IoT & Embedded Systems:** ESP32, BLE, sensor
(Nitrogen, pH, Kelembaban, dll.)
- 4. Leadership & Research:** Asisten Praktikum,
Project Engineer di AMX UAV, Ketua & Project
Manager di Elins Research Club.



Daerah Istimewa Yogyakarta



destyarosa@gmail.com



Destya Rosa Mardiana

About Company

id/x partners didirikan pada tahun 2002 oleh **ex-banker** dan **konsultan manajemen** yang memiliki pengalaman luas dalam manajemen siklus dan proses kredit, pengembangan scoring, 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.

The logo for id/x partners, consisting of the text "id/x" in white on a blue background, followed by "partners" in white on a dark blue background.

id/x partners

Project Portfolio

Sebagai seorang Data Scientist di sebuah perusahaan multifinance memiliki tugas dalam meningkatkan keakuratan dalam menilai dan mengelola risiko kredit, sehingga dapat mengoptimalkan keputusan bisnis dan mengurangi potensi kerugian. Pada portofolio ini berisi pengembangan model machine learning yang dapat memprediksi risiko kredit (credit risk) berdasarkan dataset yang disediakan, yang mencakup data pinjaman yang disetujui dan ditolak. Analisis data ini bertujuan untuk memprediksi risiko kredit berdasarkan pola perilaku dan riwayat peminjam guna mengidentifikasi tingkat risiko (tinggi atau rendah).

Link code [here!](#)

[GitHub](#)

Project explanation video [here!](#)

Business **Understanding**

Latar Belakang



Tingginya permintaan kredit membutuhkan analisis risiko yang akurat.

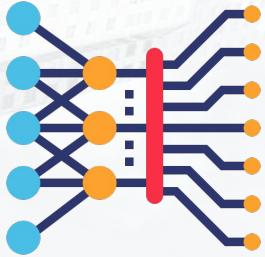


Risiko kredit yang tidak terkelola dapat menyebabkan kerugian finansial.



Data perilaku dan riwayat peminjam doapat menjadi indikator penting untuk memprediksi risiko.

Tujuan Analisis



**Membangun Model
Prediktif**



**Identifikasi Faktor
Risiko**



Optimasi Keputusan

Exploratory Data Analysis

Data Understanding

466285 Row dan 75 Feature

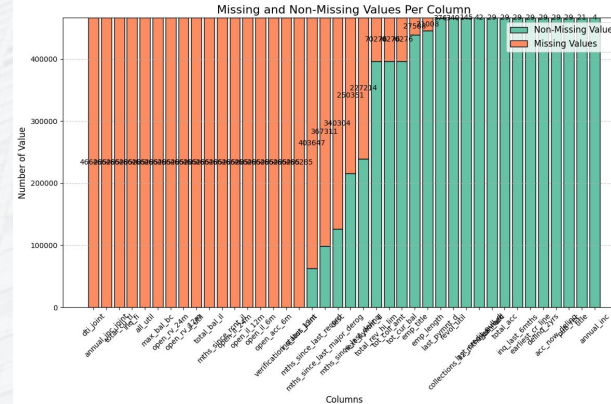
```
data.shape
✓ 0.0s
(466285, 75)
```

Analisis Missing Value

```
data.head()
```

✓ 0.0s

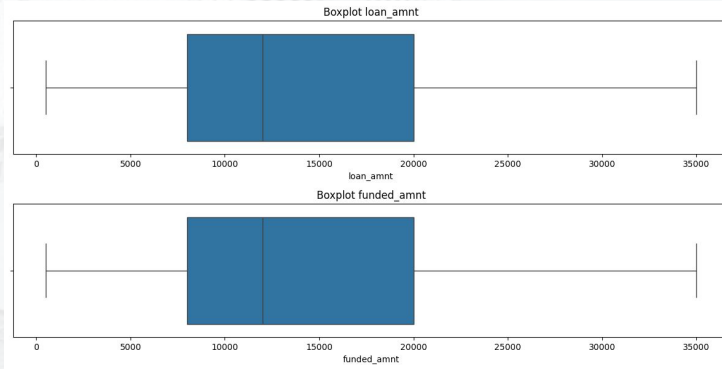
Unnamed: 0	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment
0	0	1077501	1296599	5000	5000	4975.0	36 months	10.65
1	1	1077430	1314167	2500	2500	2500.0	60 months	15.27
2	2	1077175	1313524	2400	2400	2400.0	36 months	15.96
3	3	1076863	1277178	10000	10000	10000.0	36 months	13.49
4	4	1075358	1311748	3000	3000	3000.0	60 months	12.69



```
51 mths_since_last_major_derog 98974 non-null float64
52 policy_code 466285 non-null int64
53 application_type 466285 non-null object
54 annual_inc_joint 0 non-null float64
55 dti_joint 0 non-null float64
56 verification_status_joint 0 non-null float64
57 acc_now_delinq 466256 non-null float64
58 tot_coll_amt 396089 non-null float64
59 tot_cur_bal 396089 non-null float64
60 open_acc_6m 0 non-null float64
61 open_il_6m 0 non-null float64
62 open_il_12m 0 non-null float64
63 open_il_24m 0 non-null float64
64 mths_since_rcnt_il 0 non-null float64
65 total_bal_il 0 non-null float64
66 il_util 0 non-null float64
67 open_rv_12m 0 non-null float64
68 open_rv_24m 0 non-null float64
69 max_bal_bc 0 non-null float64
70 all_util 0 non-null float64
71 total_rev_hi_lim 396089 non-null float64
72 inq_fv 0 non-null float64
73 total_cu_tl 0 non-null float64
74 inq_last_12m 0 non-null float64
dtypes: float64(46), int64(7), object(22)
```

Feature "loan_status" akan menjadi fitur utama pada analisis ini

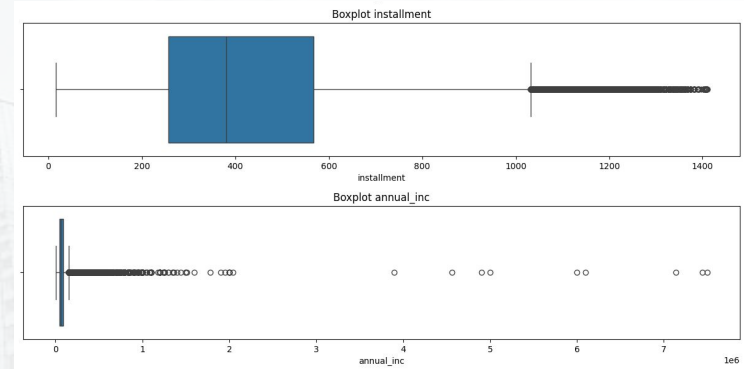
Data Distribution



Sebagian fitur memiliki distribusi yang cukup simetris dengan rentang nilai yang luas. Hal ini menunjukkan data tersebar secara normal dalam batas wajar. Ex :

loan_amnt

Funded_amnt



Sebagian fitur memiliki distribusi yang sangat skewed dengan banyak outlier di bagian atas. Ex :

installment

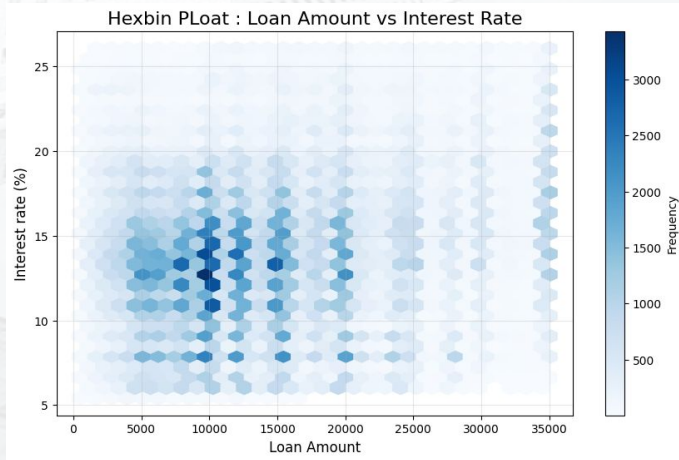
Int_rate

dti

annual_inc

revol_bal

Analisis Interaksi Data



Insight



Semakin tinggi jumlah pinjaman, suku bunga cenderung lebih stabil.

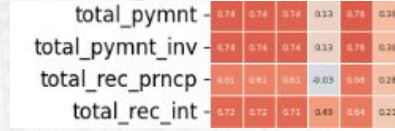
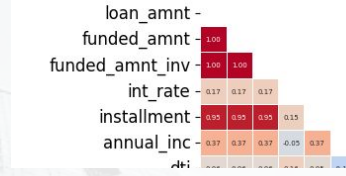
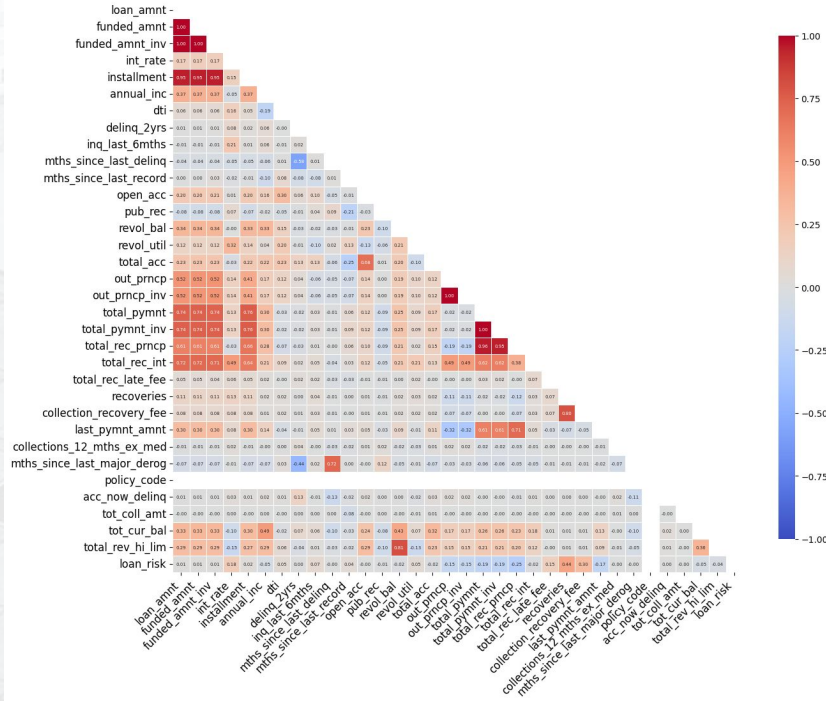
Pinjaman dengan jumlah kecil memiliki variasi suku bunga yang lebih luas.

Bisa digunakan untuk mengidentifikasi pola risiko kredit berdasarkan jumlah pinjaman.

- ❑ Warna yang lebih gelap menunjukkan area dengan kepadatan data yang lebih tinggi.
- ❑ Mayoritas pinjaman berada pada kisaran **\$5,000 - \$15,000**.
- ❑ Suku bunga sebagian besar berkisar antara **7% - 18%**.
- ❑ Terdapat beberapa pinjaman dengan jumlah lebih dari **\$25,000**, tetapi relatif jarang.

Heatmap

Heatmap Korelasi



Insight

Installment dan *in_rate* memiliki korelasi positif dengan *loan_amnt*, hal ini dapat menjadi fitur penting

Fitur yang berkorelasi tinggi dapat dihilangkan untuk mencegah redundansi

Pada cuplikan tersebut, selain fitur *in_rate*, memiliki hubungan yang tinggi

total_pymnt, *total_pymnt_inv*, dan *total_rec_prncp* memiliki korelasi tinggi dengan *loan_amnt*, menandakan bahwa jumlah pinjaman berhubungan langsung dengan total pembayaran.

Data Preprocessing

Pembersihan Data

Penghapusan feature dengan data 0 yang banyak

```
df = df.dropna(axis=1, how='all')
```

✓ 1.8s

Penghapusan feature data yang tidak terlalu informatif untuk proses analisis

application_type

zip_code

desc

title

pymnt_plan

member_id

id

Unnamed: 0

url

```
51 mths_since_last_major_derog 98974 non-null float64
52 policy_code 466285 non-null int64
53 application_type 466285 non-null object
54 annual_inc_joint 0 non-null float64
55 dti_joint 0 non-null float64
56 verification_status_joint 0 non-null float64
57 acc_now_delinq 466256 non-null float64
58 tot_coll_amt 396009 non-null float64
59 tot_cur_bal 396009 non-null float64
60 open_acc_6m 0 non-null float64
61 open_il_6m 0 non-null float64
62 open_il_12m 0 non-null float64
63 open_il_24m 0 non-null float64
64 mths_since_rcnt_il 0 non-null float64
65 total_bal_il 0 non-null float64
66 il_util 0 non-null float64
67 open_rv_12m 0 non-null float64
68 open_rv_24m 0 non-null float64
69 max_bal_bc 0 non-null float64
70 all_util 0 non-null float64
71 total_rev_hi_lim 396009 non-null float64
72 inq_fi 0 non-null float64
73 total_cu_tl 0 non-null float64
74 inq_last_12m 0 non-null float64
dtypes: float64(46), int64(7), object(22)
```



```
27 total_acc 466256 non-null float64
28 initial_list_status 466285 non-null object
29 out_prncp 466285 non-null float64
30 out_prncp_inv 466285 non-null float64
31 total_pymnt 466285 non-null float64
32 total_pymnt_inv 466285 non-null float64
33 total_rec_prncp 466285 non-null float64
34 total_rec_int 466285 non-null float64
35 total_rec_late_fee 466285 non-null float64
36 recoveries 466285 non-null float64
37 collection_recovery_fee 466285 non-null float64
38 last_pymnt_d 465909 non-null object
39 last_pymnt_amnt 466285 non-null float64
40 next_pymnt_d 239071 non-null object
41 last_credit_pull_d 466243 non-null object
42 collections_12_mths_ex_med 466140 non-null float64
43 mths_since_last_major_derog 98974 non-null float64
44 policy_code 466285 non-null int64
45 acc_now_delinq 466256 non-null float64
46 tot_coll_amt 396009 non-null float64
47 tot_cur_bal 396009 non-null float64
48 total_rev_hi_lim 396009 non-null float64
dtypes: float64(29), int64(4), object(16)
```


Converting Datetime

Beberapa fitur waktu diubah dalam bentuk numerik dengan hanya mengekstraksi **bulan**.

Issue_d

last_pymnt_d

next_pymnt_d

last_credit_pull_d

earliest_cr_line

	issue_d	earliest_cr_line	last_pymnt_d	next_pymnt_d	last_credit_pull_d
count	463536	463536	463172	236322	463496
unique	91	664	97	3	102
top	14-Oct	Oct-00	16-Jan	16-Feb	16-Jan
freq	38782	3650	179617	208390	326939



	issue_d_month	last_pymnt_d_month	next_pymnt_d_month	last_credit_pull_d_month	earliest_cr_line_month
0	12	1.0	NaN	1.0	NaN
1	12	4.0	NaN	9.0	NaN
2	12	6.0	NaN	1.0	11.0
3	12	1.0	NaN	1.0	NaN
4	12	1.0	2.0	1.0	NaN
...
466280	1	1.0	2.0	1.0	4.0
466281	1	12.0	NaN	1.0	NaN
466282	1	1.0	2.0	12.0	12.0
466283	1	12.0	NaN	4.0	2.0
466284	1	1.0	2.0	1.0	NaN

463536 rows x 5 columns

Proses Labelling

Fitur **loan_status** dilakukan labelling untuk penentuan target prediksi. Pengklasifikasiannya dilakukan sebagai berikut :

Risk

- Charged Off
- Default
- Late (31-120 days)
- Late (16-30 days).

👉 **Kategori ini dianggap memiliki risiko kredit tinggi**, karena keterlambatan atau kegagalan pembayaran.

Non-Risk

- Fully Paid,
- Current
- In Grace Period.

👉 **Kategori ini dianggap aman**, karena peminjam menunjukkan kepatuhan dalam pembayaran.

Features Engineering

Ordinal Encoding

Mengubah kategori menjadi angka berdasarkan urutan tertentu.

1. **term**
2. **grade**
3. **sub_grade**
4. **emp_length**
5. **verification_status**

Label Encoding

Mengubah setiap kategori menjadi angka unik (tanpa mempertimbangkan urutan).

1. **home_ownership**
2. **purpose**
3. **addr_state**
4. **initial_list_status**

Pembuatan **Fitur Baru**

loan_to_income

Mengukur seberapa besar jumlah pinjaman dibandingkan dengan pendapatan tahunan peminjam.

credit_utilization

Mengukur sejauh mana peminjam telah menggunakan batas kredit yang tersedia.

installment_to_income

Mengukur proporsi cicilan bulanan terhadap pendapatan bulanan peminjam.

high_risk_delinquency

Menandai peminjaman yang memiliki lebih dari satu keterlambatan pembayaran dalam laporan kredit mereka

revolving_utilization

Mengukur seberapa besar revolving balance (saldo berjalan) terhadap batas kredit yang tersedia

Train Test Split dan Fitur Scaling

Train Test Split

Membagi dataset menjadi data test sebanyak 20% dan data train sebanyak 80%

Fitur Scaling

Fitur scaling yang digunakan adalah Standardization untuk menyamakan skala fitur agar model tidak terpengaruh oleh perbedaan unit dan rentang nilai.

Data Modeling

Training Model

Prediction

Evaluation

Logistic Regression

```
==== Logistic Regression ====
Accuracy: 0.9888
Classification Report:

```

	precision	recall	f1-score	support
0	0.99	0.99	0.99	82423
1	0.95	0.95	0.95	10285
accuracy			0.99	92708
macro avg	0.97	0.97	0.97	92708
weighted avg	0.99	0.99	0.99	92708

```

Confusion Matrix:
[[81904  519]
 [ 523 9762]]

```

Data Training

```
c:\Users\ASUS A456UR\anaconda3\envs\tf_env\lib\site-packages\sk
warnings.warn(
Akurasi Logistic Regression pada testing: 0.6523600983733874

```

Ini menunjukkan bahwa Logistic Regression memiliki performa yang buruk pada data testing, jauh lebih rendah dari pada data training. Ini adalah indikasi kuat bahwa model ini **underfitting** atau **tidak cocok dengan data**.

Data Modeling

Training
Model

Prediction

Evaluation

Random Forest
Classifier

```
==== Random Forest Classifier ====
Accuracy: 0.9801
Classification Report:
      precision    recall  f1-score   support

     0       0.99       0.99       0.99      82423
     1       0.89       0.94       0.91      10285

 accuracy          0.98      92708
 macro avg       0.94       0.96       0.95      92708
weighted avg       0.98       0.98       0.98      92708

Confusion Matrix:
[[81233  1190]
 [  654  9631]]
```

Data Training

Akurasi Random Forest pada testing: 0.9801095914052724

Performa yang baik dan stabil, dengan akurasi yang tinggi pada data testing. Perbedaan dengan akurasi data testing juga tidak jauh pula. Ini menunjukkan bahwa **Random Forest** tidak overfitting dan berjalan dengan **performa baik**.

Dengan nilai **0.98**, berarti model dapat memprediksi model dengan benar sebanyak **98%** dari semua data yang diuji.

Data Modeling

Training
Model

Prediction

Evaluation

LightGBM

Akurasi LightGBM pada testing: 0.994013461621435

Performa pada model LightGBM berada di akurasi 0.99 atau mendekati 1. Oleh karena itu perlu dilakukan pemeriksaan apakah model ini terjadi overfitting atau tidak, maka dilakukan evaluasi dengan cross-validation.

Akurasi LightGBM (CV): 0.9752294965690871

Setelah dilakukan evaluasi dengan cross-validation, akurasi berada di angka 0.97. Hal ini menunjukkan bahwa rata-rata model memiliki akurasi 97% ketika diuji dengan data yang berbeda dalam proses cross validation. Ini menunjukkan model memiliki performa yang konsisten dan dapat bekerja dengan baik.

```
c:\Users\ASUS A456UR\anaconda3\envs\tf_env\lib\site-packag
warnings.warn(
===== LightGBM =====
Accuracy: 0.9940
Classification Report:
      precision    recall  f1-score   support

      0       0.99       1.00       1.00      82423
      1       1.00       0.95       0.97      10285

 accuracy          0.99          0.99      92708
 macro avg         1.00          0.97          0.98      92708
 weighted avg       0.99          0.99          0.99      92708

Confusion Matrix:
[[82403   20]
 [  535 9750]]
```

Data Training

Data Modeling

Training
Model

Prediction

Evaluation

CatBoost

```
==== CatBoost ====
Accuracy: 0.9933
Classification Report:
      precision    recall  f1-score   support

     0       0.99       1.00       1.00      82423
     1       1.00       0.94       0.97      10285

...
[ 601  9684]]
```

Data Training

Akurasi CatBoost pada testing: 0.9933339086163007

Performa pada model **CatBoost** berada di akurasi **0.99** atau mendekati 1. Oleh karena itu perlu dilakukan pemeriksaan apakah model ini terjadi overfitting atau tidak, maka dilakukan evaluasi dengan **cross-validation**.

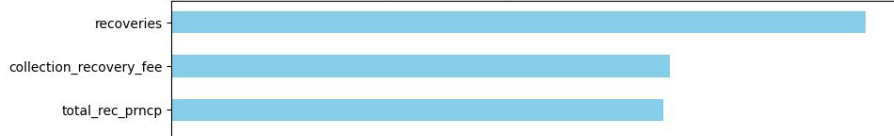
Akurasi CatBoost (CV): 0.9839343418408253

Setelah dilakukan evaluasi dengan **cross-validation**, akurasi berada di angka **0.98**. Hal ini menunjukkan bahwa rata-rata model memiliki akurasi **98%** ketika diuji dengan data yang berbeda dalam proses cross validation. Ini menunjukkan model memiliki performa yang **konsisten** dan dapat **bekerja dengan baik**.

Feature Importance

Random Forest

Top 10 Feature Importance (Random Forest)



LightGBM

Top 10 Feature Importance (LightGBM)



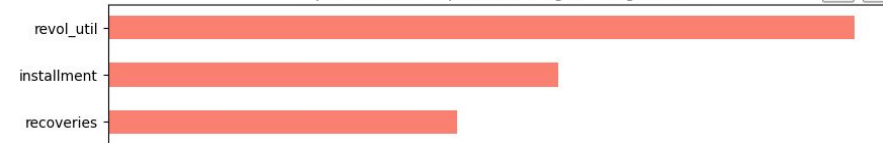
CatBoost

Top 10 Feature Importance (CatBoost)



Logistic Regression

Top 10 Feature Importance (Logistic Regression)



Dari keempat model tersebut memiliki fitur importance yang berbeda-beda, tetapi ada pula yang sama seperti LightGBM dan CatBoost. Fitur-fitur tersebut dapat menjadi **pertimbangan yang optimal untuk menentukan keputusan apakah pengajuan pinjaman disetujui atau tidak.**

7. Conclusion

Berdasarkan analisis risiko kredit yang dilakukan, **model Random Forest, LightGBM dan CatBoost** menunjukkan performa terbaik dalam memprediksi risiko kredit berdasarkan **F1-Score**. **Feature importance** mengidentifikasi bahwa faktor utama seperti **recoveries, total_rec_prncp, dan revol_util** memiliki pengaruh signifikan dalam klasifikasi risiko. Implementasi model machine learning ini dapat **meningkatkan akurasi deteksi risiko**, membantu **mengurangi potensi kredit macet**, serta mendukung pengambilan keputusan yang lebih **efektif dan berbasis data** dalam evaluasi pinjaman.

Thank You



Rakamin
Academy



id/x

partners