ID/X Partners x Rakamin Academy

PREDICTION MODEL

Machine Learning - Prediksi Credit Risk

By Suny Guinesya Ardiansyah

Suny Guinesya Ardiansyah

Data Science Enthusiast

Suny adalah seorang data science enthusiast yang memiliki keterampilan dalam pengolahan dan analisis data. Dengan keahliannya dalam Python, SQL, dan alat BI, Suny mampu mengubah data menjadi informasi yang dapat diandalkan untuk mendukung berbagai keputusan bisnis. Suny terus berupaya mengembangkan pengetahuannya di bidang ini, dengan fokus pada inovasi dan hasil yang berkualitas tinggi.

www.linkedin.com/in/suny-guinesya-ardiansyah-b839761ba

Bandung, Jawa Barat

sannyguinesya@gmail.com

Course and Certification

Scan for Certificate

or access link

https://drive.google.com/drive/u/0/folders/1S4bVl7G61nkJgs_GbIn7gIie4upVuJMa

Project Based Virtual Internship :
Big Data Analytict - Kimia Farma x Rakamin Academy
Awarded a Certificate of Achievement with an average score of 93.32
July 2024

Bootcamp Data Science Rakamin Academy
December 2023 - June 2024

Mini Project Rakamin Academy - Analyzing eCommerce Business Performance with SQL
June 2024

Mini Project Rakamin Academy - Investigate Hotel Business using Data Visualization
July 2024

About Company

id/x partners didirikan pada tahun 2002 oleh ***ex-bankers and management consultants*** yang memiliki pengalaman luas dalam manajemen siklus kredit dan proses, pengembangan skor, dan manajemen kinerja. Pengalaman gabungan kami telah melayani perusahaan-perusahaan di seluruh wilayah Asia dan Australia serta di berbagai industri, khususnya jasa 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 yang terintegrasi untuk membantu klien mengoptimalkan profitabilitas portofolio dan proses bisnis.

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

Project Portofolio

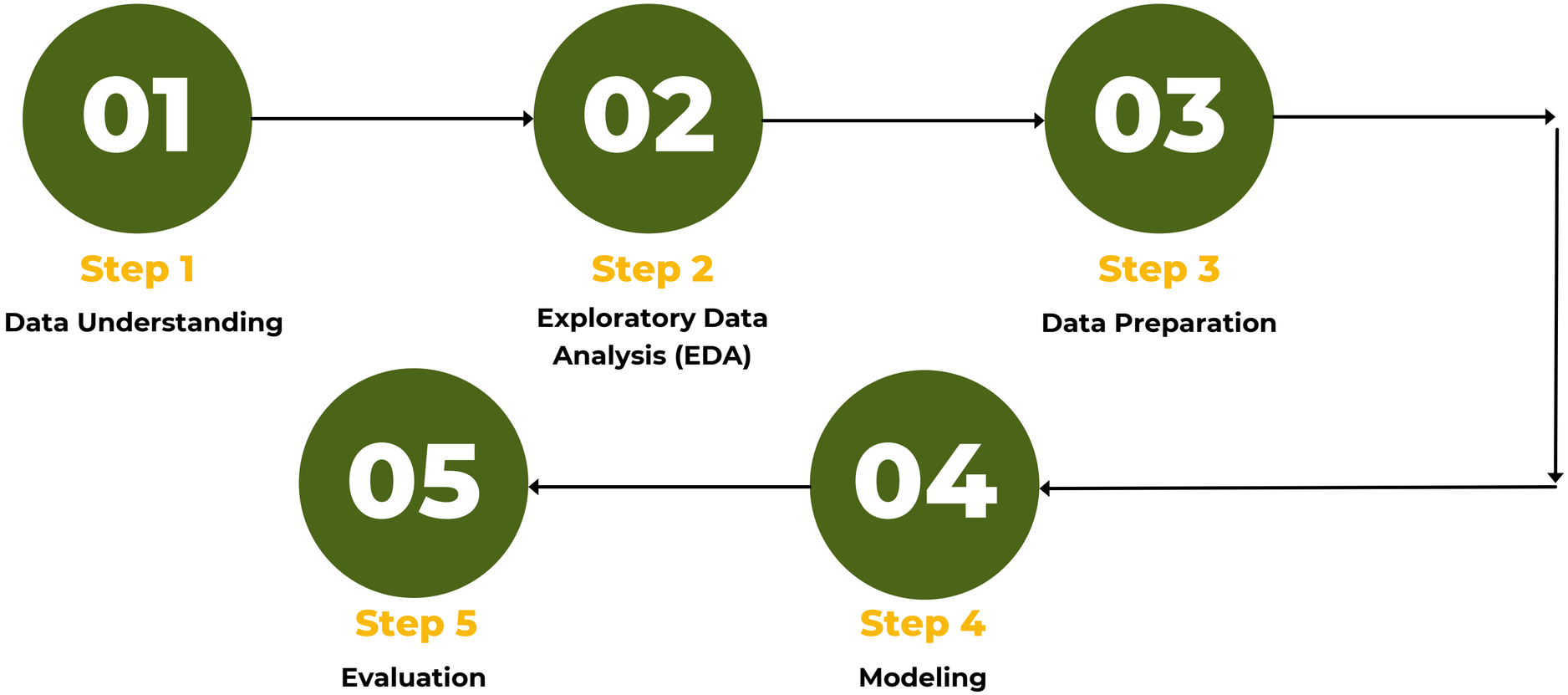
Sebagai **Data Scientist di ID/X Partners**, Anda akan terlibat dalam sebuah proyek dari perusahaan pemberi pinjaman (multifinance), dimana client Anda ingin **meningkatkan keakuratan dalam menilai dan mengelola risiko kredit**, sehingga dapat mengoptimalkan keputusan bisnis mereka dan mengurangi potensi kerugian.

Tugas Anda adalah **mengembangkan model machine learning** yang dapat **memprediksi risiko kredit (credit risk)** berdasarkan dataset yang disediakan, yang mencakup data pinjaman yang disetujui dan ditolak. Dalam pengembangan modelnya Anda juga perlu melakukan beberapa tahap dimulai dengan Data Understanding, Exploratory Data Analysis (EDA), Data Preparation, Data Modelling, dan Evaluation.

Link Video Penjelasan

https://drive.google.com/file/d/1Yr76JiCvRctNAP8SFBn8F0CDMcaM9usM/view?usp=sharing

Overview



Link Kode Python:

https://github.com/sunyardiansyah/predicyion_model_credit_risk_machine_learning_idx_partners_data_science

Rakamin Academy

Business Understanding

Problem

Adanya keterlambatan pembayaran bahkan nasabah tidak membayar pinjaman

Goals

Membuat Machine Learning untuk mengurangi resiko (Credit Risk)

Objective

Mengidentifikasi fitur-fitur yang kemungkinan nasabah membayar tepat waktu (Good Credit)

Business Metric

Loss Given Default (LGD) : Estimasi kerugian yang akan dihadapi perusahaan jika sebuah pinjaman gagal bayar

Business Understanding

Jumlah Baris dan Kolom

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

Terdapat 75 kolom dan 466.285 Baris

```
[4]: df_loan.head(3)
```

[4]:

	Unnamed: 0	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	grade	sub_grade	emp_title	emp_ler
0	0	1077501	1296599	5000	5000	4975.0	36 months	10.65	162.87	B	B2	NaN	10+ y
1	1	1077430	1314167	2500	2500	2500.0	60 months	15.27	59.83	C	C4	Ryder	< 1
2	2	1077175	1313524	2400	2400	2400.0	36 months	15.96	84.33	C	C5	NaN	10+ y

Dataset yang disajikan berbentuk tabular (tabel) yang memiliki 75 kolom dan 466.285 baris data.

sesuai dengan tujuan (credit risk), fitur loan_status akan menjadi fitur utama pada dataset ini.

Rakamin Academy

Business Understanding

Informasi Awal

Banyak kolom yang tidak memiliki Nilai sama sekali

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

Hapus Fitur 0 non-null

Sudah dipastikan fitur fitur yang tidak memiliki nilai sama sekali tidak akan terpakai, maka semua fitur tersebut dihapus secara langsung.

Deskripsi Kolom

policy_code	466285.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00

Semua baris memiliki nilai police code yang sama.

Kolom lainnya akan di lihat distribusinya pada saat EDA.

Rakamin Academy

Business Understanding

Deskripsi Kolom Kategorikal

	count	unique	top	freq
term	466285	2	36 months	337953
grade	466285	7	B	136929
sub_grade	466285	35	B3	31686
emp_title	438697	205475	Teacher	5399
emp_length	445277	11	10+ years	150049
home_ownership	466285	6	MORTGAGE	235875
verification_status	466285	3	Verified	168055
issue_d	466285	91	Oct-14	38782
loan_status	466285	9	Current	224226
pymnt_plan	466285	2	n	466276
url	466285	466285	https://www.lendingclub.com/browse/loanDetail....	1
desc	125981	124435		234
purpose	466285	14	debt_consolidation	274195
title	466264	63098	Debt consolidation	164075
zip_code	466285	888	945xx	5304
addr_state	466285	50	CA	71450
earliest_cr_line	466256	664	Oct-00	3674
initial_list_status	466285	2	f	303005
last_pymnt_d	465909	98	Jan-16	179620
next_pymnt_d	239071	100	Feb-16	208393
last_credit_pull_d	466243	103	Jan-16	327699
application_type	466285	1	INDIVIDUAL	466285

Tanda Biru

Kolom seharusnya bertipe Numeric

Tanda Hitam

Kolom memiiliki nilai Unik yang terlalu banyak

Tanda Kuning

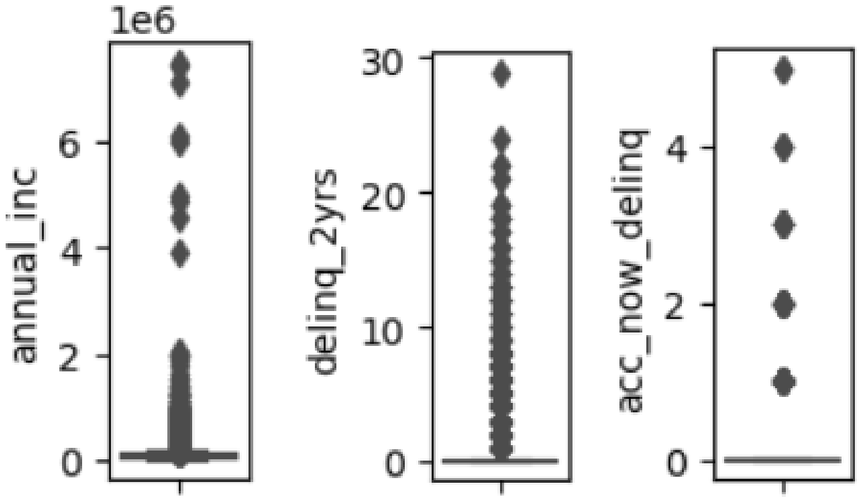
Kolom Bertipe Datetime

Exploratory Data Analysis (EDA)

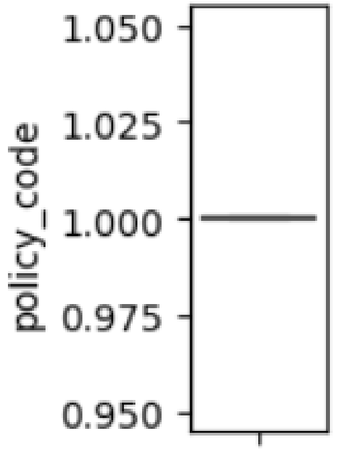
Rakamin Academy

BOXPLOT

Sebanyak 13 Kolom menyerupai 3 contoh di samping, tidak terlihat area datanya, Hal ini berarti banyak sekali outliers yang terdapat pada kolom-kolom tersebut.



Police code hanya memiliki 1 nilai yaitu 1, maka boxplot nya akan menjadi seperti gambar disamping.

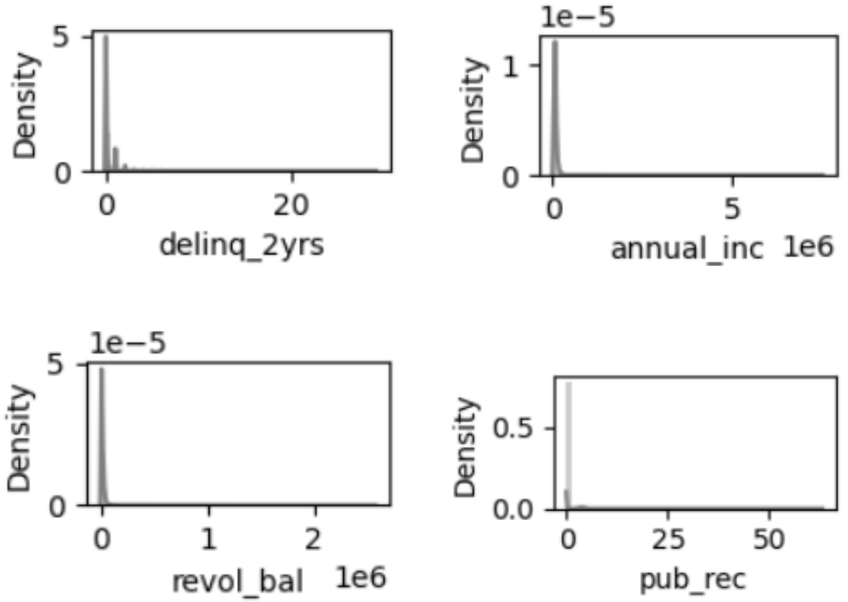


Exploratory Data Analysis (EDA)

Rakamin Academy

DISTRIBUTION PLOT

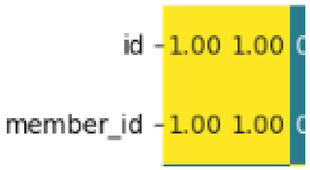
Sebanyak 13 kolom juga memiliki distribusi seperti gambar di samping. Tentunya harus menjadi pertimbangan apakah kolom-kolom tersebut sebaiknya dihapus atau tidak.



Police code hanya memiliki 1 nilai yaitu 1, maka distribution plot terlihat seperti disamping.

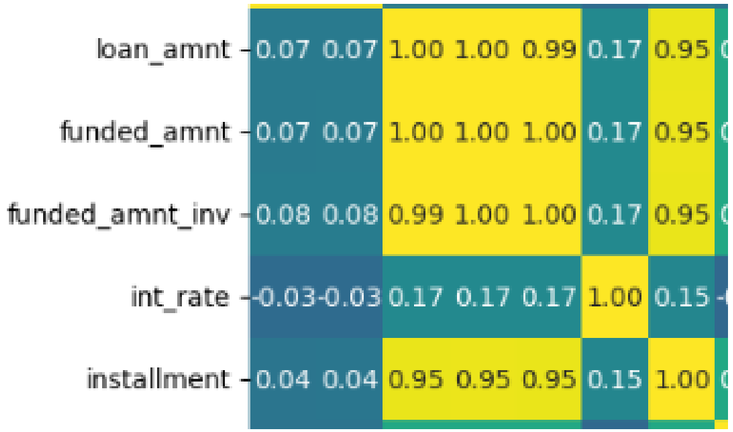
Exploratory Data Analysis (EDA)

HEATMAP

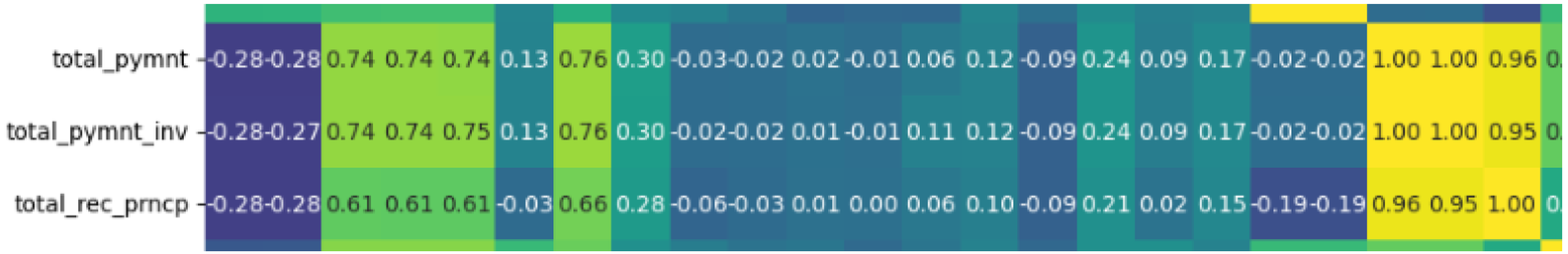


id dan member id akan di hapus karena hanya identitas peminjaman.

Selain int_rate, 4 kolom yang lainnya memiliki hubungan yang sangat tinggi, hal ini tentunya menyebabkan redundan. Maka dari ke empat kolom tersebut 3 kolom akan dihapus.



3 kolom di bawah memiliki hubungan yang sangat tinggi, maka penanganannya sama seperti sebelumnya.



Data Preprocessing

Hapus Kolom

selain kolom yang tidak memiliki nilai, hasil dari EDA saya memutuskan untuk **menghapus 18 kolom** dengan ketentuan sebagai berikut.

- 2 kolom id
- Nilai Missing Values yang lebih dari 48%
- semua baris memiliki nilai berbeda
- Terlalu banyak nilai Unik
- Hanya memiliki 1 nilai
- proporsi di dominasi oleh 1 nilai
- redundan.

Missing Values

	Persentase Missing Values (%)
emp_length	4.51
annual_inc	0.00
title	0.00
delinq_2yrs	0.01
earliest_cr_line	0.01
inq_last_6mths	0.01
open_acc	0.01
pub_rec	0.01
revol_util	0.07
total_acc	0.01
last_pymnt_d	0.08
last_credit_pull_d	0.01
collections_12_mths_ex_med	0.03
acc_now_delinq	0.01
tot_coll_amt	15.07
tot_cur_bal	15.07
total_rev_hi_lim	15.07

- **Menghapus baris** yang memiliki nilai **Missing Values kurang dari 10%**
- **Mengganti nilai** Missing Values **dengan median** untuk yang memiliki **Missing values lebih dari 10%** karena data Skewed

Data Preprocessing

Feature Engineering

- Menghapus Object pada baris nilai kolom **emp_length** dan **term** dan menyisakan angka, kemudian type data di ubah menjadi numeric.
- Memisahkan kolom tanggal dan bulan pada kolom Kolom **issue_d, last_credit_pull_d, last_pymnt_d, dan earliest_cr_line.**
- Menghapus huruf xx pada **zip_code** kemudian diubah type datanya menjadi Numeric.

Membuat Kolom Terget

Kolom **target** diambil dari kolom **loan status** dengan ketentuan sebagai berikut.

- Jika nilai pada loan_status adalah Current, Fully Paid, dan In Grace Period, maka **kredit ridak beresiko** diberi label **0**.
- Selain itu **kredit beresiko** diberi label **1**

Kolom loan_status kemudian di hapus supaya tidak terjadi overfitting.

Data Preprocessing

Outliers

Mengatasi **Outliers** menggunakan **Log Transformation**, karena jika menggunakan IQR menyebabkan terlalu banyak kehilangan data.

Encoding

- Menggunakan **One-Hot Encoding** untuk kolom kategorikal Nominal (**home_ownership, verification_status, purpose, dan initial_list_status**)
- Menggunakan **Label Encoder** untuk kolom kategorikal ordinal (**grade**)

Scaling

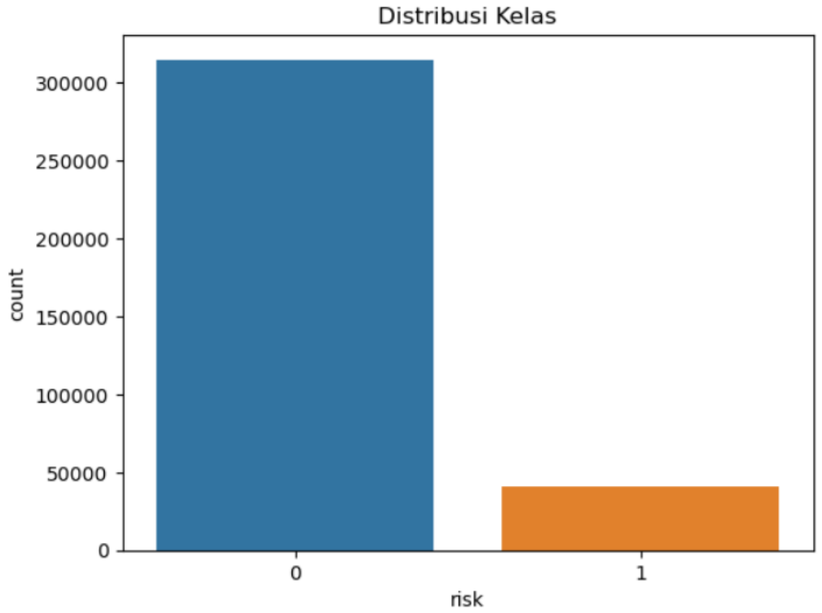
Menggunakan **MinMaxScaler** untuk membuat skala semua kolom sama, tetapi kolom boolean tidak di skala karena sudah bernilai 1 dan 0

Train Test Split

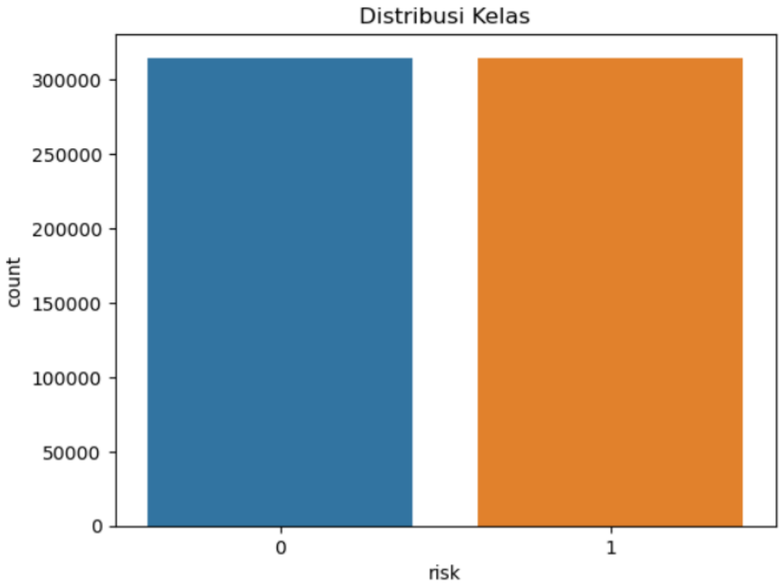
Membagi dataset menjadi data Train dan data Test dengat persentase **70% Train** dan **30% Test**

Data Preprocessing

Imbalanced Class



Penangana **Imbalaced Class** yang dilakukan dengan menggunakan Oversampling **SMOTE.**



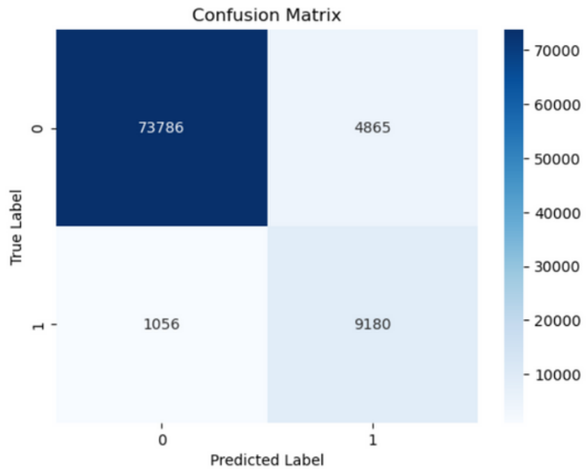
Modeling

```
LogisticRegression
LogisticRegression(random_state=42)
```

Logistic Regression

Melatih Model → **Melakukan Prediksi** → **Evaluasi**

Result




```
Classification Report:
              precision    recall  f1-score   support

           0       0.99      0.94      0.96     78651
           1       0.65      0.90      0.76     10236

    accuracy                           0.93     88887
   macro avg       0.82      0.92      0.86     88887
weighted avg       0.95      0.93      0.94     88887

Accuracy: 0.9334
ROC-AUC Score: 0.9730
```

Dengan nilai **Accuracy 0.93**, berarti model tersebut **memberikan prediksi yang benar sebanyak 93%** dari semua data yang diuji. Nilai **ROC-AUC 0.97** menunjukkan bahwa model ini sangat baik dalam membedakan antara kedua kelas. Ini berarti bahwa dalam 97% dari semua pasangan acak yang terdiri dari satu instance positif dan satu instance negatif. Nilai ini menunjukkan bahwa **model ini secara konsisten dapat membedakan antara kelas-kelas, bahkan ketika threshold berubah.**

Modeling

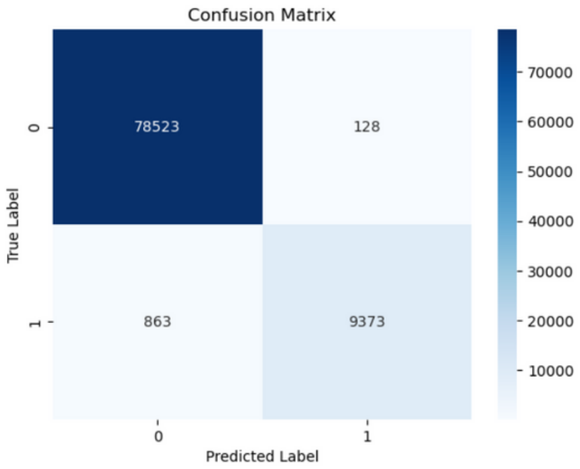
RandomForestClassifier

```
RandomForestClassifier(random_state=42)
```

Random Forest

Melatih Model → **Melakukan Prediksi** → **Evaluasi**

Result



	precision	recall	f1-score	support
0	0.99	1.00	0.99	78651
1	0.99	0.92	0.95	10236
accuracy			0.99	88887
macro avg	0.99	0.96	0.97	88887
weighted avg	0.99	0.99	0.99	88887

Jika melihat pada metric evaluasi, semua hasil sangat mendekati angka 1, bahkan recall berniilai 1, maka hasil dari **model di atas Overfiting**.

Dilakukan sekali lagi dengan pembatasan pembatasan dan melakukan **Cross Validation** guna mengecek apakah memang benar model tersebut Overfiting atau tidak.

Modeling

Random Forest

```
RandomForestClassifier
RandomForestClassifier(random_state=42)
```

Dengan langkah yang sama, tetapi menambahkan batasan batasan sebelum melakukan pelatihan model sebagai berikut

```
model = RandomForestClassifier(
    n_estimators=100,           # Jumlah pohon
    max_depth=10,               # Batasi kedalaman pohon
    min_samples_split=4,        # Jumlah minimum sampel untuk split
    min_samples_leaf=2,         # Jumlah minimum sampel per daun
    max_features='sqrt',        # Pengambilan sampel fitur
    random_state=42             # Untuk reproduktifitas hasil
)
```

Hasil di Slide selanjutnya

Modeling

Random Forest

```
RandomForestClassifier(random_state=42)
```

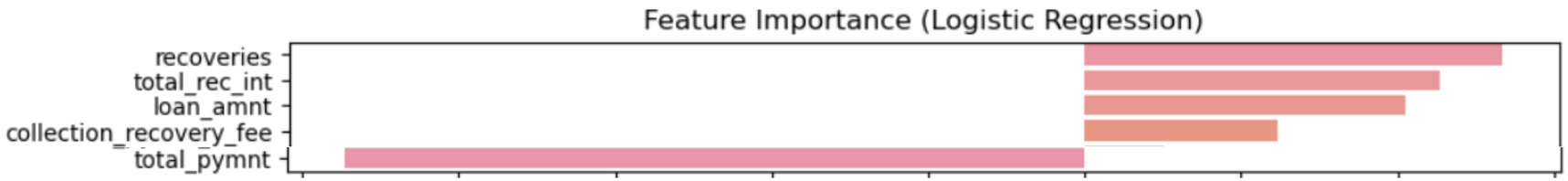
Result

```
Akurasi: 0.9663280344707325
Classification Report:
              precision    recall  f1-score   support

           0       0.98      0.98      0.98     78651
           1       0.84      0.88      0.86     10236

    accuracy                           0.97     88887
   macro avg       0.91      0.93      0.92     88887
weighted avg       0.97      0.97      0.97     88887

Confusion Matrix:
[[76931  1720]
 [ 1273  8963]]
Cross-Validation Scores: [0.93959332 0.97418966 0.9770707  0.97463412 0.97673736]
Mean CV Accuracy: 0.9684450300009523
```

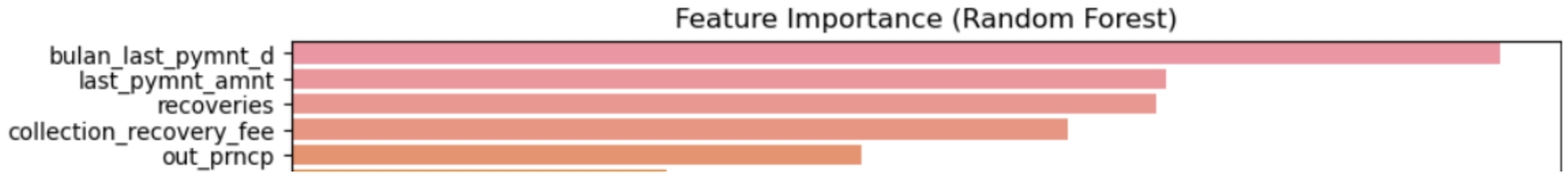
Terlihat nilai menurun dari yang sebelumnya. Nilai **accuracy 0.97** artinya **97% persen Model dapat memprediksi dengan tepat.** Apakah ini Overfitting? Setelah dilakukan **Cross-Validation**, di dapat **Mean CV Accuracy sebesar 0.97** hal ini menunjukkan bahwa rata-rata model memiliki akurasi 97% ketika diuji dengan data yang berbeda dalam proses cross-validation. Ini menunjukkan **model ini memiliki performa yang konsisten dan kemungkinan besar akan bekerja dengan baik pada data yang belum pernah dilihat sebelumnya.**

Feature Importance

Logistic Regression



Random Forest



Feature Importance dari ke dua Model memiliki urutaan berbeda, tetapi ada featur importance yang sama. Pertimbangan untuk mengoptimalkan feature-feature tersebut sangatlah **penting untuk keputusan apakah pengajuan pinjaman akan di setujui atau tidak,**

Kesimpulan

Model yang dibuat memiliki nilai yang tinggi, tetapi tidak overfitting karena telah dilakukan cross validation. Tentunya **model ini dapat di implementasikan terhadap data baru** dan kemungkinan besar akan sesuai dengan apa yang diharapkan yaitu akan terdeteksinya pola credit yang beresiko. Maka **dengan adanya prediksi ini, perusahaan dapat untuk tidak menyetujui nasabah yang terdeteksi sebagai resiko.** Dengan ini **kerugian dapat diminimalisir.**

Link Report

Code Lengkap

https://github.com/sunyardiansyah/predicyion_model_credit_risk_machine_learning_idx_partners_data_science

Video Penjelasan

https://drive.google.com/file/d/1Yr76JiCvRctNAP8SFBn8F0CDMcaM9usM/view?usp=sharing

Rakamin Academy

id/x partners

THANK YOU

For your nice attention

email :
sannyguinesya@gmail.com

Adress :
Bandung, Jawa Barat

Linkedin :
www.linkedin.com/in/suny-guinesya-ardiansyah-b839761ba