Rakamin Academy X id/x partners

Prediction Modeling

ID/X Partners - Data Scientist

Presented by

Alfriando C Vean

Sleman, D.I. Yogyakarta

alfriandocv@gmail.com

Alfriando C Vean

Alfriando C Vean

Jobseeker

Saya adalah lulusan dari Universitas Mercu Buana Yogyakarta jurusan Informatika pada tahun 2024. Didorong oleh ketertarikan pada bidang data, saya memutuskan untuk mendaftar pada *Data Science Bootcamp* **yang diselenggarakan oleh Rakamin Academy. Saya ingin menimba ilmu dan pengalaman sebanyak-banyaknya selama mengikuti** *Bootcamp* **dan** *Project Based Internship* **di Rakamin Academy. Fokus saya adalah menjadi seorang** *Data Scientist*, **di mana saya mengombinasikan ilmu** *programming* **yang saya dapat di perguruan tinggi dengan kemampuan mengolah data yang didapat di Rakamin Academy.**

Project Portfolio

Perusahaan multifinance perlu meningkatkan keakuratan penilaian risiko kredit untuk mengoptimalkan keputusan bisnis dan mengurangi kerugian. Kami mengembangkan model machine learning menggunakan data pinjaman dari Lending Club (2007-2014) untuk memprediksi risiko kredit, dengan fokus pada metrik bisnis seperti kerugian dan margin keuntungan bersih. Analisis data ini bertujuan untuk mengidentifikasi pola yang mengindikasikan pinjaman berpotensi buruk atau berisiko, tanpa asumsi yang kuat, untuk mendukung pengambilan keputusan investasi.

Project explanation video here!

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.

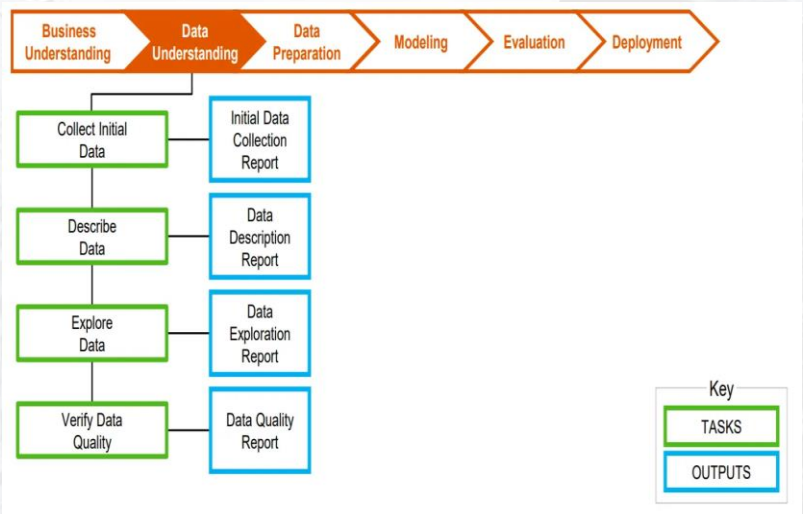
id/x partners menyediakan layanan konsultasi yang mengkhususkan diri dalam memanfaatkan solusi analitik data dan pengambilan keputusan (DAD) yang dikombinasikan dengan disiplin manajemen risiko dan pemasaran terintegrasi untuk membantu klien mengoptimalkan profitabilitas portofolio dan proses bisnis.

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

1. Data Understanding

Data Understanding adalah tahap kedua dalam proses CRISP-DM (Cross-Industry Standard Process for Data Mining) yang fokus pada pengumpulan dan penilaian kualitas data. Tahap ini melibatkan empat tugas utama:

1. **Mengumpulkan Data Awal**: Mengidentifikasi data yang tersedia, metode pengambilan, dan masalah yang mungkin dihadapi.
2. **Mendeskripsikan Data**: Memeriksa properti data yang diperoleh, termasuk format, kuantitas, dan isi dari setiap tabel atau dataset.
3. **Menjelajahi Data**: Menggunakan pertanyaan ilmu data untuk mendapatkan wawasan awal melalui kueri, visualisasi, dan laporan ringkasan.
4. **Memverifikasi Kualitas Data**: Memastikan data cukup bersih dan relevan untuk analisis yang akan dilakukan



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	int64
5	funded_amnt_inv	466285 non-null	float64
6	term	466285 non-null	object
7	int_rate	466285 non-null	float64
8	installment	466285 non-null	float64
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
...			
73	total_cu_tl	0 non-null	float64
74	inq_last_12m	0 non-null	float64

Data asli memiliki 75 kolom dan 466285 baris.

	index	Total Null	Percentage Null
0	open_rv_24m	466285	100.000000
1	inq_fi	466285	100.000000
2	open_rv_12m	466285	100.000000
3	il_util	466285	100.000000
4	mths_since_rcnt_il	466285	100.000000
5	total_bal_il	466285	100.000000
6	open_il_24m	466285	100.000000
7	open_il_12m	466285	100.000000
8	open_il_6m	466285	100.000000
9	open_acc_6m	466285	100.000000
10	dti_joint	466285	100.000000
11	annual_inc_joint	466285	100.000000
12	max_bal_bc	466285	100.000000
13	all_util	466285	100.000000
14	inq_last_12m	466285	100.000000
15	total_cu_tl	466285	100.000000
16	verification_status_joint	466285	100.000000
17	mths_since_last_record	403647	86.566585
18	mths_since_last_major_derog	367311	78.773926
19	desc	340304	72.981975
20	mths_since_last_delinq	250351	53.690554
21	next_pymnt_d	227214	48.728567
22	total_rev_hi_lim	70276	15.071469
23	tot_cur_bal	70276	15.071469
24	tot_coll_amt	70276	15.071469
25	emp_title	27588	5.916553
26	emp_length	21008	4.505399
27	last_pymnt_d	376	0.080637
28	revol_util	340	0.072917
29	collections_12_mths_ex_med	145	0.031097
30	last_credit_pull_d	42	0.009007
31	acc_now_delinq	29	0.006219
32	delinq_2yrs	29	0.006219
33	inq_last_6mths	29	0.006219
34	earliest_cr_line	29	0.006219
35	open_acc	29	0.006219
36	pub_rec	29	0.006219
37	total_acc	29	0.006219
38	title	21	0.004504
39	annual_inc	4	0.000858

Data tidak memiliki baris yang duplikat, namun terdapat 39 kolom yang memiliki nilai *null* dengan jumlah dan persentase beragam mulai dari yang sangat sedikit sampai mendominasi 100%

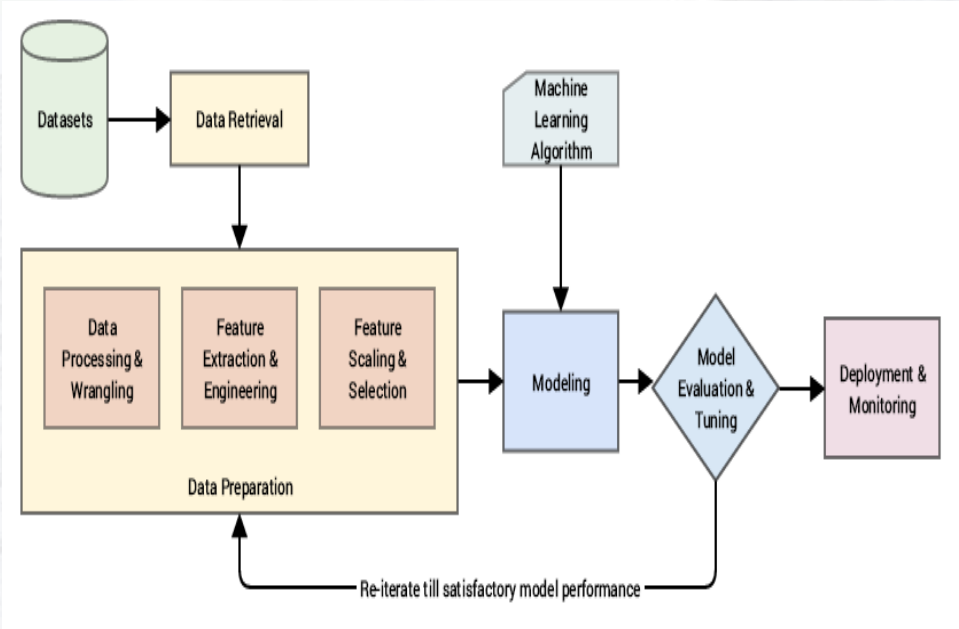
loan_status	
Current	224226
Fully Paid	184739
Charged Off	42475
Late (31-120 days)	6900
In Grace Period	3146
Does not meet the credit policy. Status:Fully Paid	1988
Late (16-30 days)	1218
Default	832
Does not meet the credit policy. Status:Charged Off	761

Data ini memiliki target dari kolom *loan_status* yang mewakili status dari pemohon pinjaman.

2. Feature Engineering

Feature Engineering dalam ilmu data adalah proses kreatif yang melibatkan pemilihan, transformasi, dan pembuatan fitur baru dari data mentah untuk meningkatkan kinerja model pembelajaran mesin. Ini termasuk:

1. **Seleksi Fitur**: Memilih fitur yang paling relevan dengan masalah yang sedang dihadapi.
2. **Transformasi Fitur**: Mengubah skala atau distribusi fitur untuk meningkatkan interpretasi oleh model.
3. **Penciptaan Fitur**: Menggabungkan atau memodifikasi fitur untuk menghasilkan informasi yang lebih berguna.
4. **Ekstraksi Fitur**: Mengidentifikasi dan mengekstrak informasi penting dari data mentah.



1. *Handling Null Value*
- *Null* >40% → *drop*
- *Null* <40% dan fitur bertipe numerik → median
- *Null* <40% dan fitur bertipe kategorikal → modus.

2. *Handling High Correlation*
Korelasi >0,7 → *drop*.

3. *Handling Unrelevant Features*
Korelasi <0,2 → *drop*.

4. *Transformation*
Metode → *Yeo-Johnson*

5. *Feature Encoding*
'grade', 'verification_status', 'home_ownership', 'initial_list_status' → *One-Hot Encoding*

6. *Datetime Features*
Aggregate dan ubah ke numerikal

7. *Handling Target*
'Does not meet the credit policy. Status:Fully Paid' → 'Fully Paid',
'Does not meet the credit policy. Status:Charged Off' → 'Charged Off'

'Fully Paid', 'Current', 'In Grace Period' → Good (1)
Others → Bad (0)

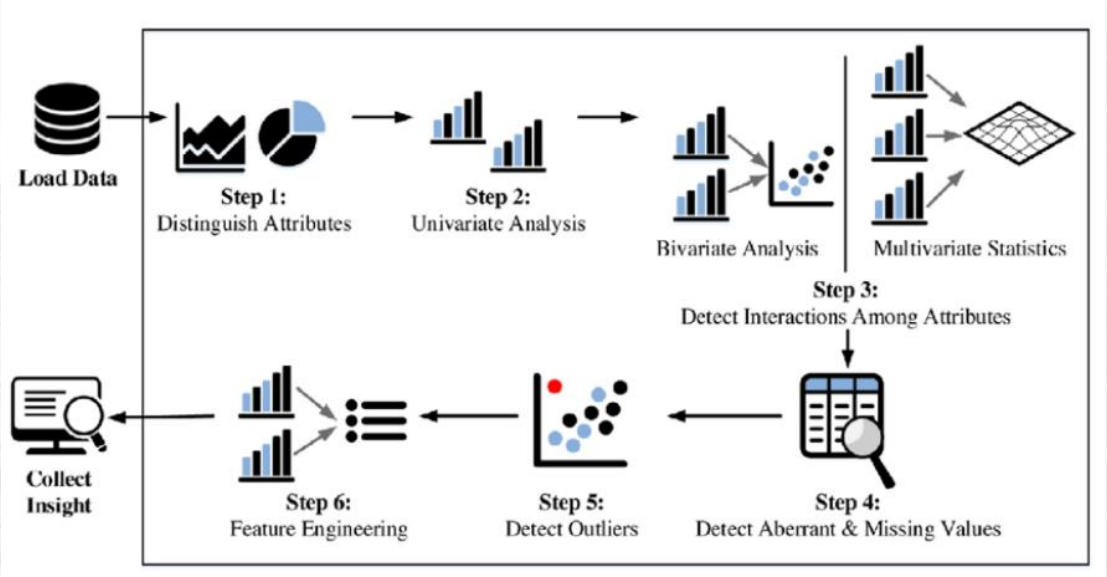
```
Total Rows : 466285
Total Features : 40
```

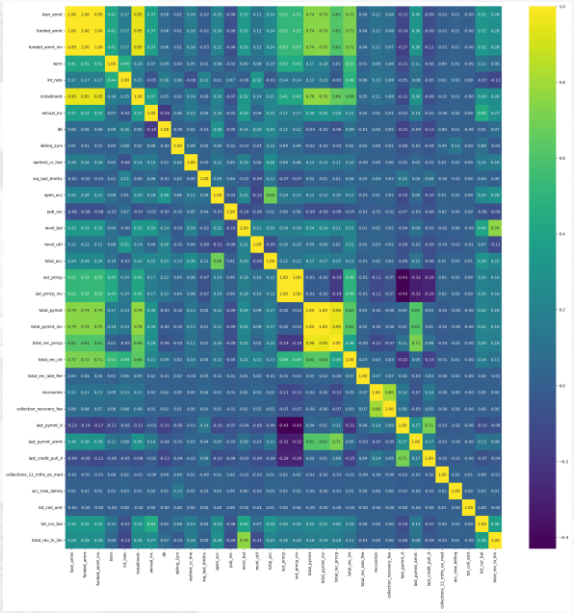
```
float64    21
int64      19
```

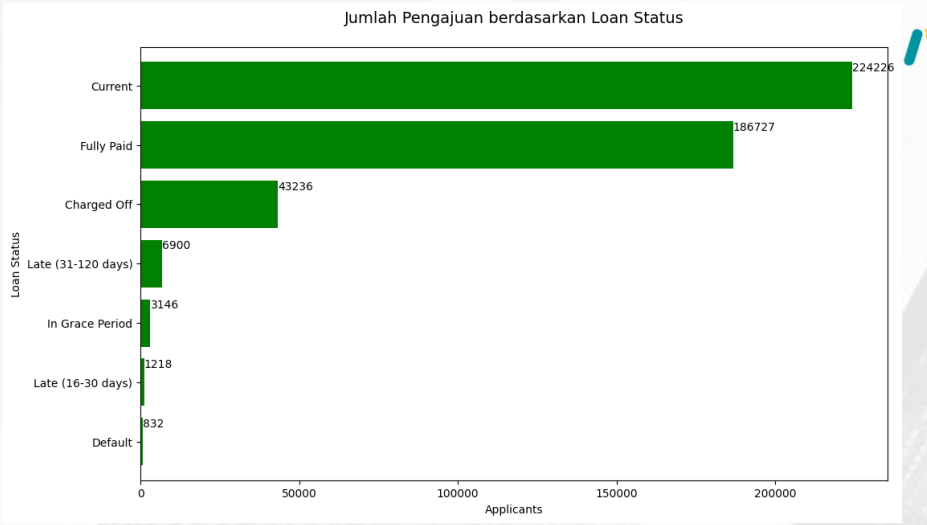
3. Exploratory Data Analysis

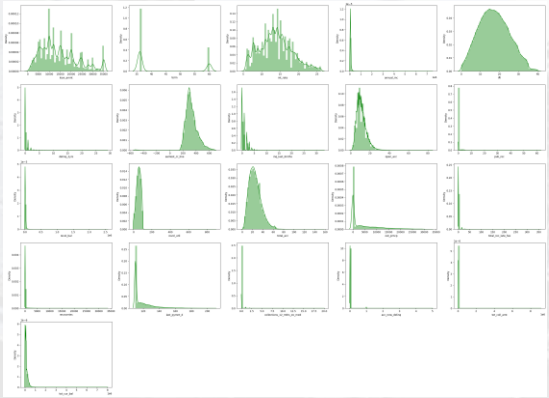
Exploratory Data Analysis (EDA) dalam ilmu data adalah proses analisis awal data untuk memahami karakteristik utama, menemukan pola, anomali, dan hubungan antar variabel. Proses EDA biasanya meliputi:

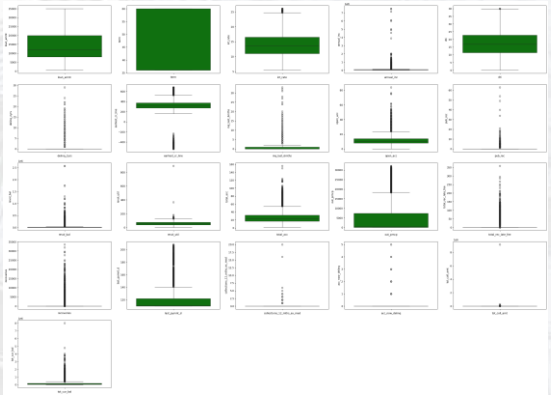
1. **Gambaran Umum Dataset**: Memahami jumlah observasi, jenis fitur, dan data yang hilang.
2. **Statistik Deskriptif**: Meringkas data numerik melalui ukuran tendensi sentral dan dispersi.
3. **Visualisasi Data**: Menggunakan grafik dan diagram untuk menggambarkan distribusi dan hubungan data.
4. **Evaluasi Kualitas Data**: Memeriksa kebersihan dan konsistensi data.

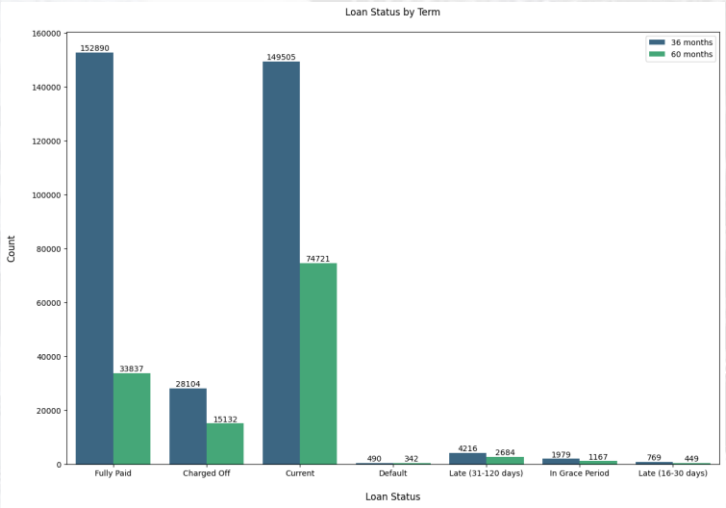





Jumlah Pengajuan berdasarkan Loan Status
Loan Status
Applicants
Current
Fully Paid
Charged Off
Late (31-120 days)
In Grace Period
Late (16-30 days)
Default
224226
186727
43236
6900
3146
1218
832
0
50000
100000
150000
200000






Loan Status by Term
Count
Loan Status
36 months
60 months
Fully Paid
Charged Off
Current
Default
Late (31-120 days)
In Grace Period
Late (16-30 days)
152890
33837
28104
15132
149505
74721
490
342
4216
2684
1979
1167
769
449

4. Data Preparation

1. *Features & Target Split*

```
X = df_final.drop(labels=['binary_loan_status'],axis=1)
y = df_final[['binary_loan_status']]
```

2. *Train & Test Split*

```
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size = 0.2, stratify=y, random_state = 42)
```

3. *Handling Imbalance*

Metode → SMOTE

```
X_train, y_train = over_sampling.SMOTE(
    sampling_strategy=1).fit_resample(X_train, y_train)
```

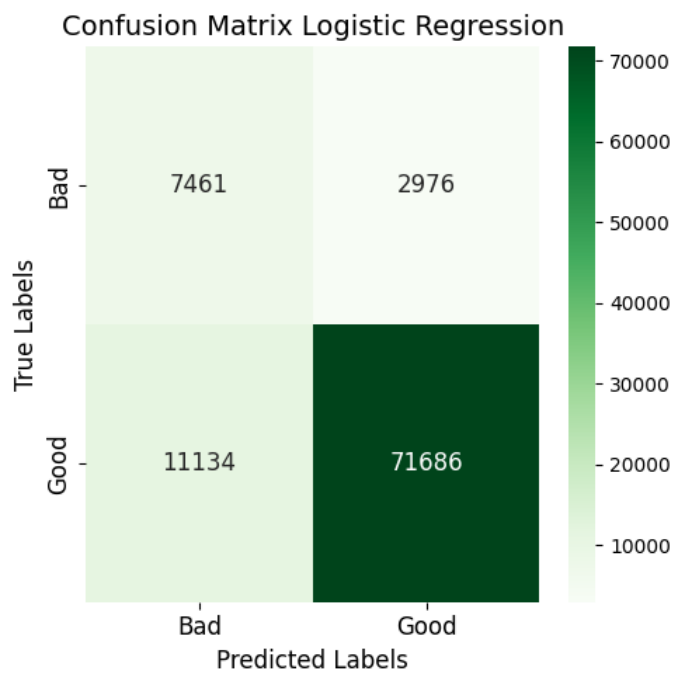
```
binary_loan_status
0                   331279
1                   331279
```

5. Data Modeling

Terdapat 6 jenis model yang diuji, dengan hasil sebagai berikut:

Model	ROC-AUC		Recall	
	Train	Test	Train	Test
Logistic Regression	0,88	0,88	0,86	0,86
Decision Tree	1,00	0,78	1,00	0,42
Random Forest	1,00	0,92	1,00	0,58
LightGBM	0,99	0,84	0,99	0,32
Ada Boost	0,97	0,90	0,91	0,54
XGBoost	0,99	0,82	0,99	0,17

6. Evaluation & Conclusion



Dari perbandingan beberapa model, diperoleh model terbaik yaitu:

Logistic Regression

Dapat disimpulkan bahwa jika menggunakan model ini:

- Dapat mengenali 71686 pinjaman yang sebenarnya baik (good).
- Dapat menghindari 7461 pinjaman yang sebenarnya berisiko (bad).
- Dapat menekan kesalahan prediksi pinjaman berisiko sebagai pinjaman baik menjadi 2976.
- Dan Dapat menekan kesalahan prediksi pinjaman baik sebagai pinjaman berisiko menjadi 11134.

Thank You

GitHub