Rakamin Academy X id/x partners

Classifiying Credit Risk : A Systematic Approach

ID/X Partners - Data Scientist

Presented by
Annisa Sekartierra Mulyanto

Yogyakarta

annisasekar220@gmail.com

Annisa Sekartierra Mulyanto

Annisa Sekartierra Mulyanto

<Profile>

Mahasiswa Statistika tahun kedua di Universitas Gadjah Mada. Ambisius dan bertanggung jawab dengan kemampuan manajemen tim dan kepemimpinan yang baik. Berminat pada bidang Data Analysis, Data Science, Data Mining, dan Machine Learning. Terbuka dengan pengalaman baru dengan kemampuan adaptasi yang cepat dan kemauan belajar yang tinggi, Aktif meningkatkan softskill dan hardskill secara akademis maupun non-akademis dengan mengikuti perlombaan, projek, dan pelatihan.

Project Portfolio

Suatu perusahaan finance yang menyediakan jasa layanan pinjaman perlu untuk memahami karakter peminjam agar dapat mengurangi risiko kerugian. Dengan membentuk suatu model prediksi credit risk peminjam tentunya akan dapat menekan angka risiko kerugian tersebut. Model ini dikembangkan dengan harapan agar perusahaan dapat mengetahui karakteristik dari peminjam yang memiliki potensi pinjaman baik dan buruk.

Akan dikembangkan suatu model prediktif dengan metode supervised learning menggunakan data dari Lending Club (2007-2014).

Project explanation video here!

Github Repository link here!

Folder of Project Files here!

1. Data Understanding

Variabel target pada data yang digunakan adalah 'loan_status' yang menunjukkan status peminjaman saat ini. Pada kolom ini terdapat 7 kategori dengan penjelasan sebagai berikut:

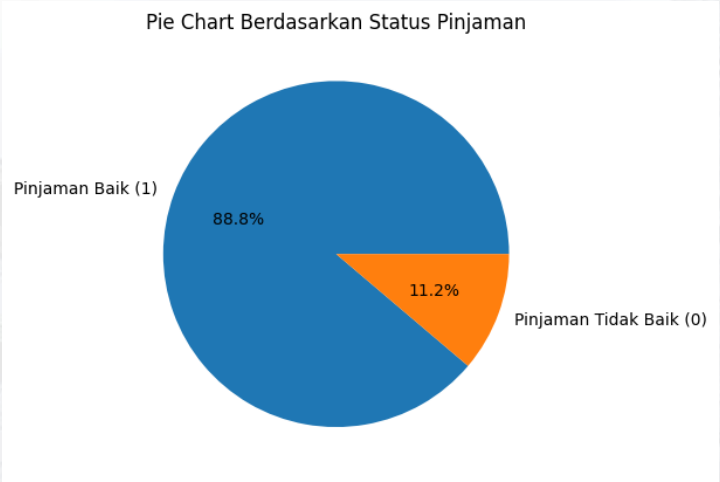
- **Current**: Pinjaman up-to-date pada semua pembayaran yang belum dilunasi.
- **In Grace Period**: Pinjaman sudah jatuh tempo tetapi masih dalam periode tenggang 15 hari.
- **Late (16-30)**: Pinjaman tidak up-to-date selama 16 hingga 30 hari.
- **Late (31-120)**: Pinjaman tidak up-to-date selama 31 hingga 120 hari.
- **Fully paid**: Pinjaman telah dilunasi sepenuhnya, baik pada akhir periode 3 atau 5 tahun atau sebagai hasil dari pelunasan di muka.
- **Default**: Pinjaman tidak up-to-date untuk jangka waktu yang lama.
- **Charged Off**: Pinjaman yang tidak lagi diharapkan untuk menerima pembayaran lebih lanjut.

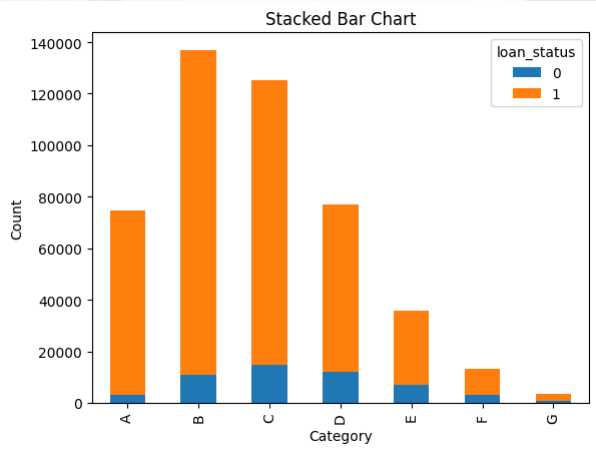
Kami mengidentifikasikan suatu **pinjaman yang baik** (1) apabila status pinjaman saat ini berada pada kategori 'Current', 'In Grace Period', dan 'Fully Paid'.

Sedangkan kategori yang lainnya akan diidentifikasikan sebagai **pinjaman yang tidak baik** (0)

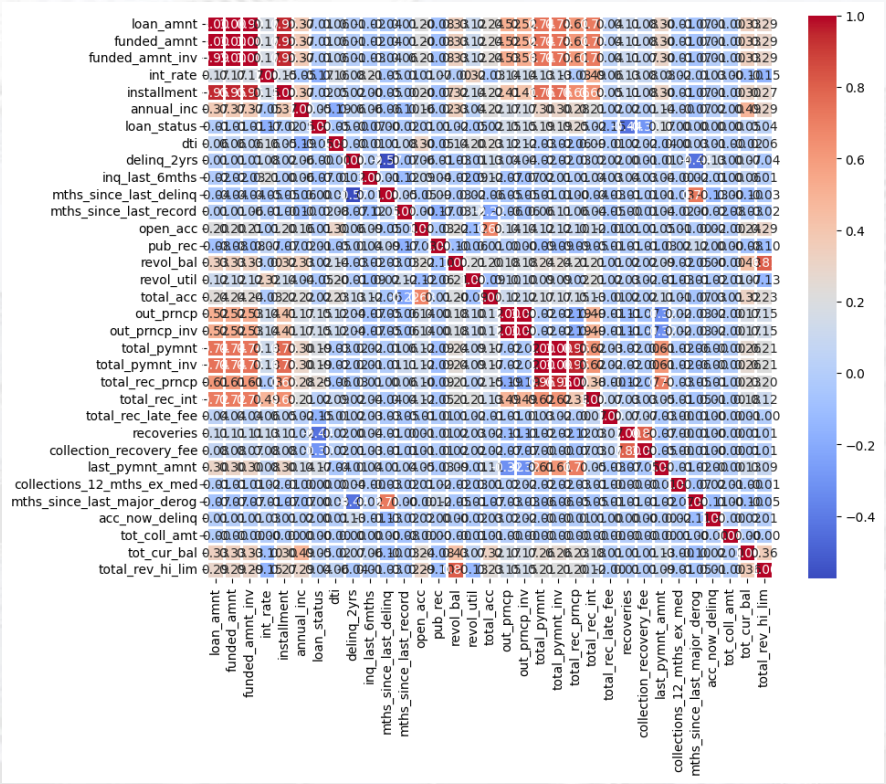
2. Exploratory Data Analysis

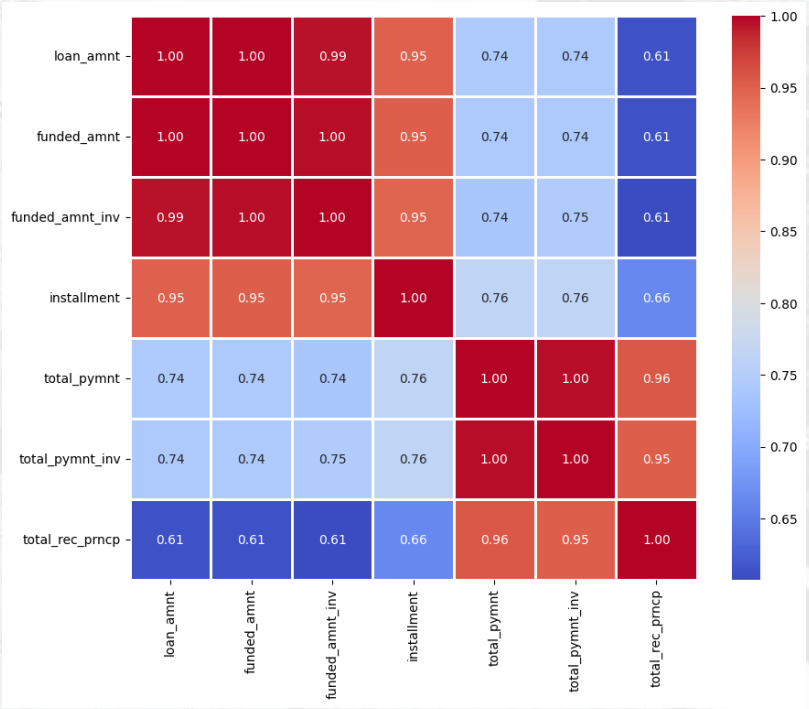
Didapatkan line insight dari tahap ini yang dijelaskan lebih lanjut sebagai berikut:



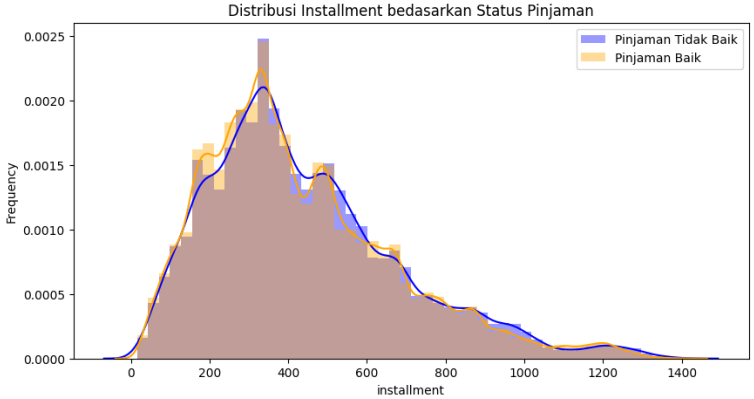


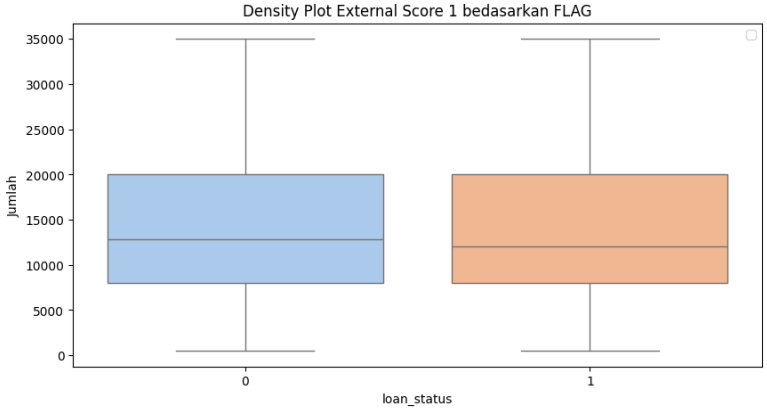
2. Exploratory Data Analysis

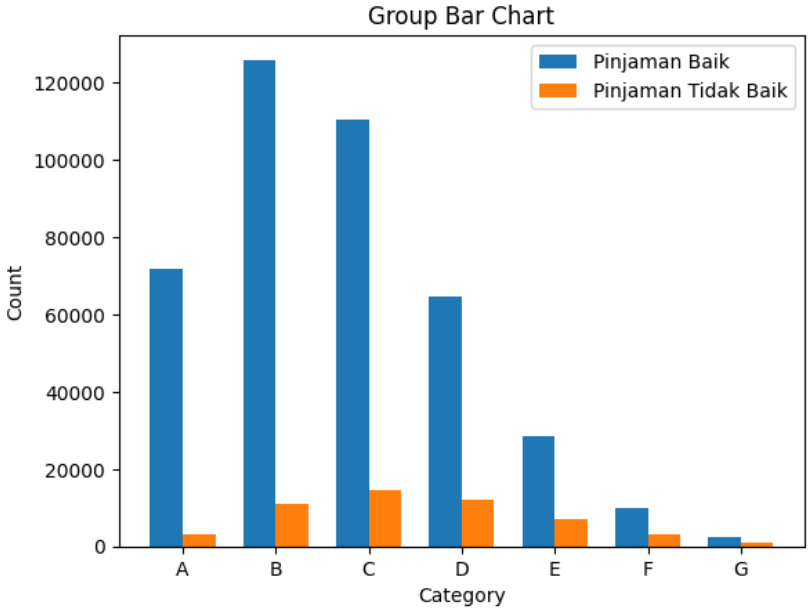




2. Exploratory Data Analysis







3. Feature Engineering

Pada tahap ini, dilakukan **ekstraksi fitur** baru menggunakan informasi dari kolom yang menyimpan data dalam format tanggal.

- **month_issue_since_crline** menunjukkan waktu dalam satuan bulan yang menunjukkan jangka waktu dari waktu peminjam membuka rekening kredit pertamanya (earliest_cr_line) sampai pendanaan pinjaman terbaru (issue_d).
- **month_last_pymnt_since_issue** menunjukkan jangka waktu dalam satuan bulan dari waktu pertama kali pinjaman terkini didanai sampai saat peminjam melakukan pembayaran terakhir untuk pinjaman terkini.

Selain ekstraksi fitur dilakukan pula **penghapusan fitur** yang dirasa tidak akan memberikan kontribusi yang baik pada analisis berikutnya. Beberapa diantara fitur tersebut adalah id, member_id, url, dan title. Dihapuskan pula beberapa kolom kategorik yang memiliki banyak kategori dan tak memungkinkan untuk dilakukan rekategorisasi, yaitu ‘emp_title’. Terakhir, dihapuskan pula fitur kategori ‘application_type’ karena hanya memiliki satu kategori.

4. Data Preparation

Bagian ini dibagi menjadi **dua tahap**, tahap awal dan tahap akhir. **Tahap awal** dilakukan sebelum EDA untuk membersihkan data dari kesalahan-kesalahan sederhana meliputi ketidakkonsistenan, duplikasi, dan kesalahan format.

Tahap kedua dilakukan setelah melakukan data splitting dimana proses preprocessing ini menggunakan pipeline untuk mencegah adanya data leakage. Proses yang dilakukan pada tahap ini meliputi pengisian missing values, normalisasi, dan encoding.

Missing value dilakukan imputasi dengan IterativeImputer() untuk variabel numerik dan SimpleImputer() metode 'most_frequent' untuk data ketegorik. Metode normalisasi yang dilakukan adalah MinMaxScaler() untuk seluruh variabel numerik. Lalu dilakukan pula encoding data kategorik menggunakan OrdinalEncoder() untuk fitur berskala ordinal dan LabelEncoder() untuk fitur berskala nominal.

5. Data Modeling

Menggunakan Pipeline, telah dicoba 10 algoritma klasifikasi berbeda dan digunakan metrik evaluasi **F1 Score** untuk memilih model terbaik. Metrik ini dipilih karena dapat merepresentasikan performa untuk imbalance data. Hasil pemodelan menggunakan pipeline mengeluarkan performa sebagai berikut.

	Model	Balanced Accuracy	F1 Score	ROC AUC
0	LogisticRegression	0.870519	0.984000	0.944706
1	LogisticRegressionCV	0.898221	0.987160	0.966794
2	GradientBoostingClassifier	0.906183	0.988341	0.967906
3	HistGradientBoostingClassifier	0.915520	0.989314	0.970526
4	CatBoostClassifier	0.917042	0.989515	0.971599
5	XGBClassifier	0.921413	0.989926	0.973195
6	BernoulliNB	0.731684	0.890018	0.817419
7	GaussianNB	0.560797	0.948405	0.560897
8	MultinomialNB	0.619961	0.912043	0.731173
9	KNeighborsClassifier	0.531134	0.938886	0.633634

Bedasarkan metrik **F1 Score**, dipilih model **XG Boost Classifier** sebagai model terbaik.

6. Evaluation

Secara lengkap performa dari model terbaik adalah sebagai berikut:

	precision	recall	f1-score	support
0	0.99	0.84	0.91	9982
1	0.98	1.00	0.99	80606
accuracy			0.98	90588
macro avg	0.99	0.92	0.95	90588
weighted avg	0.98	0.98	0.98	90588

Balance Aquracy = 0.9214131960119586
F1 score = 0.9899256873459503
ROC AUC = 0.9731949455773735

F1-score bernilai **98,99%**. Artinya, model `best` secara keseluruhan memiliki keseimbangan **sangat baik** antara presisi dan recall, di mana 98,99% dari prediksi positif adalah benar, dan model dapat mengidentifikasi 98,99% dari total instance yang sebenarnya positif.

7. Conclusion

Beberapa langkah dalam peningkatan performa model klasifikasi dapat dilakukan dengan cara berikut:

- Melakukan penanganan data imbalance dalam data splitting
- Melakukan cross validation untuk menenjukan hyperparameter
- Melakukan imputasi missing value dengan algoritma machine learning terutama untuk variabel kategorik.

Dengan model ini (atau model yang sudah ditingkatkan) Lending Club dapat menggunakannya untuk **menekan risiko kerugian akibat ketidaktepatan pemberian pinjaman kepada customer**. Model yang telah dikembangkan menunjukkan **performa yang luar biasa** menggunakan algoritma XG Boost Classifier. Dengan mengimplementasikan ini, Lending Club dapat mengurangi kemungkinan memberikan pinjaman ke calon peminjam yang dirasa tidak akan membayar pinjaman secara penuh. Dampaknya, Lending Club akan dapat secara tepat sasaran memberikan pinjaman ke calon peminjam yang layak dengan karakteristik tertentu.

Thank You

Rakamin Academy X id/x partners