



# Leveraging Machine Learning for Improved Credit Risk Prediction

VIX Data Scientist ID/X Partners



Created by: Rofik

Email: rofikn4291@gmail.com

linkedln: Rofik Rofik

"I'm Rofik. I have experience in data mining, data analysis, preprocessing, and making predictions using data. I'm familiar with tools like Python, SQL, Pandas, Numpy, Matplotlib, Scikit-learn, BigQuery, and have worked on creating dashboards using Google Data Studio and Tableau. I'm eager to continue learning and developing my skills in the data field. I am hardworking and a team player, always open to collaboration and new challenges."



# **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**

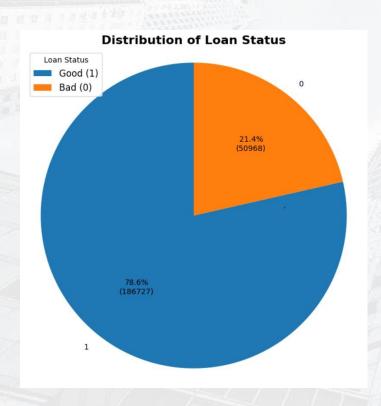
id/x partners didirikan pada tahun 2002 oleh mantan bankir dan konsultan manajemen yang memiliki pengalaman luas dalam manajemen siklus dan proses kredit, pengembangan skoring, dan manajemen kinerja. Pengalaman gabungan kami telah melayani korporasi di seluruh wilayah Asia dan Australia serta di berbagai 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.



# **Case Study**



Di tengah meningkatnya jumlah calon nasabah, sebuah perusahaan pemberi pinjaman menghadapi tantangan besar dalam memprediksi risiko kredit dengan cepat dan tepat untuk menjaga stabilitas operasional.

Hanya 78.6% nasabah yang berhasil mengembalikan pinjamannya. Evaluasi risiko kredit manual terbukti lambat dan kurang akurat, sehingga meningkatkan angka gagal bayar dan risiko kerugian finansial.

Tanpa penerapan model prediksi risiko kredit yang efektif, perusahaan berisiko mengalami kerugian lebih besar, kehilangan kepercayaan pemangku kepentingan, dan kesulitan bersaing di pasar yang semakin kompetitif.

Bagaimana mengidentifikasi calon nasabah yang berisiko gagal bayar?

Bagaimana membuat model otomatisasi dalam penilaian kredit?



# **Business Question**

- Bagaimana karakteristik orang yang good loan dan bad loan
   (gagal bayar)?
- Bagaimana Machine Learning membantu dalam prediksi credit risk?
- Apa saja faktor penting dalam keputusan prediksi credit risk?



Exploratory
Data Analysis



Model Prediktif



Feature Importance

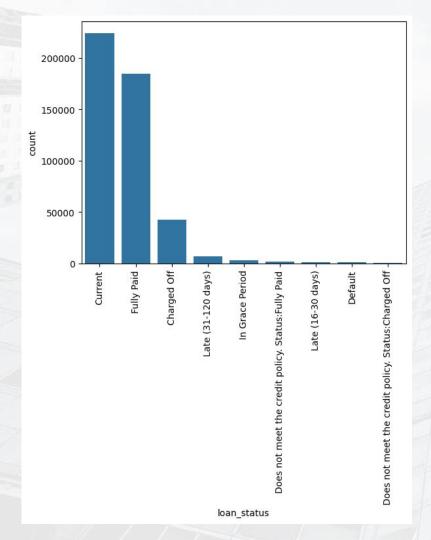
### **Dataset**

Fitur	Deskripsi
term	Jumlah bulan dalam pinjaman, bisa 36 atau 60 bulan.
home_ownership	Status tempat tinggal peminjam (misal: menyewa, memiliki rumah, dll).
verification_status	Status verifikasi pendapatan peminjam (terverifikasi atau tidak).
loan_status	Status pinjaman saat ini (misalnya, dilunasi, tertunggak, dll).
pymnt_plan	Perencanaan pembayaran pinjaman yang telah disepakati.
purpose	Tujuan atau kategori dari pinjaman (misalnya, untuk usaha, pendidikan, dll).
initial_list_status	Status awal saat pinjaman terdaftar (misalnya, publik atau tidak).
annual_inc	Penghasilan tahunan peminjam.
loan_amnt	Jumlah total pinjaman yang diajukan.
int_rate	Tingkat bunga yang dikenakan pada pinjaman.
dti	Rasio utang terhadap pendapatan, menghitung kemampuan peminjam untuk membayar kembali pinjaman.
revol_bal	Saldo total dari kredit bergulir yang dimiliki peminjam.
total_pymnt	Jumlah total pembayaran yang telah diterima untuk pinjaman.
delinq_2yrs	Jumlah kejadian keterlambatan pembayaran lebih dari 30 hari dalam 2 tahun terakhir.
inq_last_6mths	Jumlah pemeriksaan kredit dalam 6 bulan terakhir.
open_acc	Jumlah akun kredit terbuka yang dimiliki peminjam
pub_rec	Jumlah catatan publik yang dapat menurunkan skor kredit (misal, kebangkrutan).
revol_util	Rasio penggunaan kredit bergulir, mengukur seberapa banyak kredit yang digunakan oleh peminjam.
total_acc	Jumlah total akun kredit yang dimiliki peminjam.
out_prncp	Sisa pokok pinjaman yang belum dibayar untuk total pinjaman yang didanai.
total_rec_prncp	Jumlah pokok pinjaman yang telah diterima hingga saat ini.
total_rec_int	Jumlah bunga yang telah diterima hingga saat ini.
total_rec_late_fee	Jumlah biaya keterlambatan yang telah diterima hingga saat ini.
recoveries	Menunjukkan jika ada pembayaran yang telah dipulihkan pada pinjaman yang bermasalah.
collection_recovery_fee	Biaya yang terkait dengan pemulihan pembayaran pinjaman yang terlewat atau koleksi.
last_pymnt_amnt	Jumlah total pembayaran terakhir yang diterima dari peminjam.
collections_12_mths_ex_med	Jumlah koleksi (utang yang belum dibayar) dalam 12 bulan terakhir, kecuali koleksi medis.
acc_now_delinq	Jumlah akun yang sedang mengalami keterlambatan pembayaran saat ini.
issue_d	Tanggal saat pinjaman diterbitkan (didanai).
last_pymnt_d	Tanggal pembayaran terakhir yang diterima.

### **Dataset**

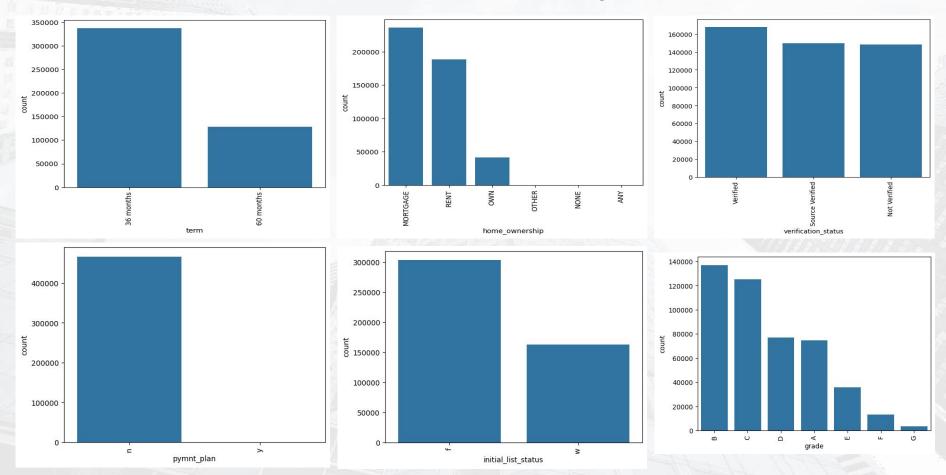
Fitur	Deskripsi	
earliest_cr_line	Bulan saat akun kredit pertama kali dibuka.	
last_credit_pull_d	Tanggal terakhir kali peminjam diperiksa kreditnya.	
grade	Kategori rating pinjaman (misalnya, A, B, C, dll) yang menunjukkan kualitas pinjaman.	
emp_length	Lama bekerja peminjam, diukur dalam tahun.	
addr_state	Negara bagian tempat peminjam tinggal.	

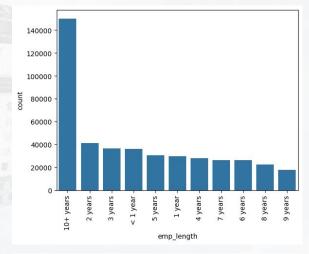
# **Exploratory Data Analysis** (EDA)

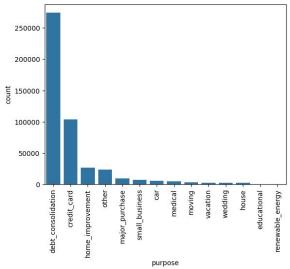


- Sebagian besar status pinjaman di perusahaan menunjukkan kondisi yang positif, dengan mayoritas nasabah masih aktif mengangsur (current) dan banyak yang telah melunasi pinjamannya.
- Namun, adanya pinjaman dalam status charged off mengindikasikan potensi kerugian yang perlu dikelola dengan strategi mitigasi risiko yang lebih baik.

# **Univariate Analysis**

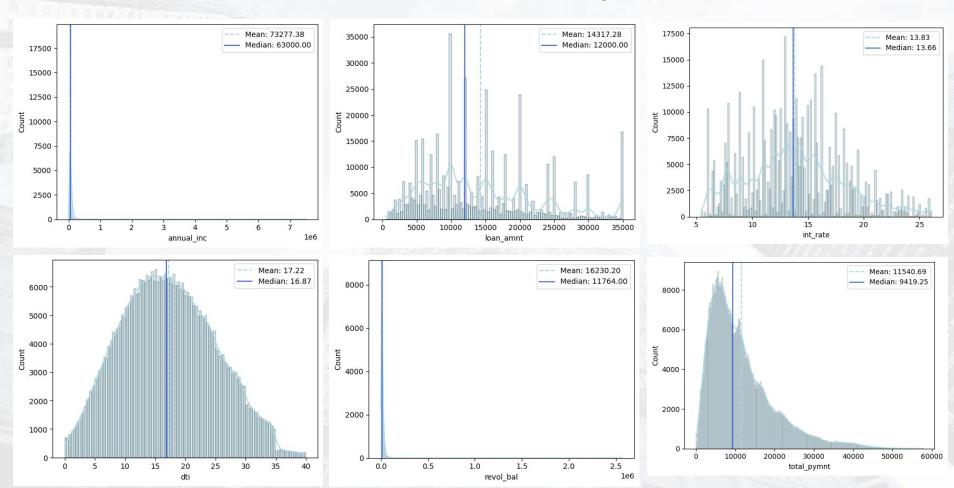






- Mayoritas peminjam memilih periode pinjaman 36 bulan.
- Sebagian besar sedang mengangsur tempat tinggal atau menyewa, dan hanya sedikit yang sudah memiliki rumah sendiri.
- Sebagian besar nasabah tidak berencana menunggak pembayaran.
- Sebagian besar pinjaman telah didanai oleh beberapa investor, dengan mayoritas peminjam berada di grade B dan C, (mencerminkan risiko kredit sedang).
- Peminjam umumnya sudah bekerja lebih dari 10 tahun, dengan status pinjaman yang didominasi oleh verifikasi yang valid.
- Tujuan utama peminjaman adalah untuk melunasi hutang sebelumnya.

# **Univariate Analysis**

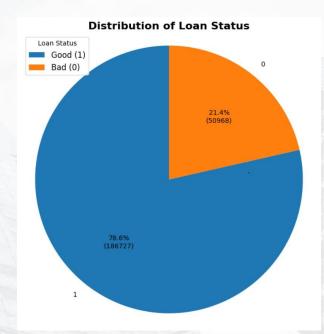


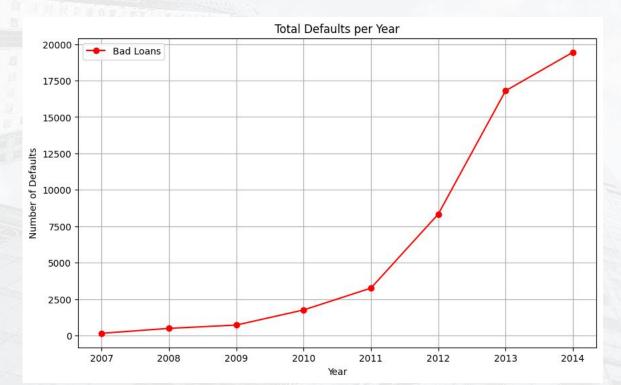
- Beberapa peminjam memiliki total pendapatan tahunan yang sangat tinggi, meskipun rata-rata annual income adalah 73.227.
- Jumlah pinjaman bervariasi, dengan rata-rata sebesar 14.317.
- Suku bunga (int rate) juga bervariasi, dengan rata-rata 13,83%.
- DTI (Debt-to-Income ratio) (Rasio hutang dan pendapatan bulanan) peminjam juga beragam, dengan rata-rata 17,22%.
- Revolving balance (saldo hutang yang belum terbayar) cenderung rendah, dengan rata-rata 16.230,20.
- Total pembayaran (total pymnt) menunjukkan distribusi yang condong positif, dengan beberapa peminjam yang telah melakukan pembayaran utang dalam jumlah besar.

# Pelabelan

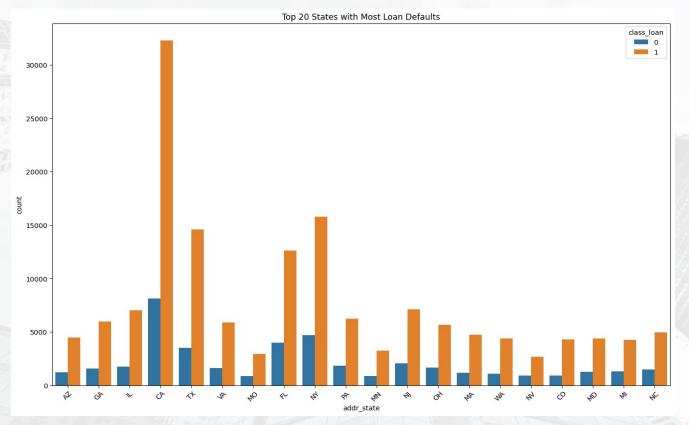
- 1. Fully Paid: Pinjaman sudah dilunasi sepenuhnya.
- 2. Current: Pinjaman sedang berjalan dan pembayaran tepat waktu.
- 3. In Grace Period: Pinjaman dalam periode tenggang setelah jatuh tempo.
- 4. Charged Off: Pinjaman dianggap gagal bayar dan tak bisa dilunasi.
- 5. Default: Peminjam gagal bayar sesuai ketentuan.
- 6. Late (31-120 days): Pinjaman terlambat 31-120 hari, berisiko tinggi.
- 7. Late (16-30 days): Pinjaman terlambat 16-30 hari, masih dalam peringatan.
- 8. Does not meet the credit policy. Status: Fully Paid: Pinjaman lunas tapi tidak memenuhi kebijakan kredit.
- Does not meet the credit policy. Status: Charged Off: Pinjaman gagal bayar dan tidak memenuhi kebijakan kredit.
  - Status Fully Paid dan Does not meet the credit policy. Status: Fully Paid, adalah status yang dapat dipastikan good loan pada peminjam. -> Dilakukan pelabelan 1 pada status ini
- dan status Charged Off, Default, Late (31-120 days) dan Does not meet the credit policy.
   Status:Charged Off, adalah status loan yang dapat dipastikan bad loan. -> dilakukan pelabelan 0 pada status-status ini.

Dalam analisis dan pembuatan model prediksi risiko kredit, akan lebih akurat apabila hanya menggunakan data yang sudah dipastikan berlabel good loan dan bad loan. Hasil pelabelan menunjukkan bahwa persentase good loan mencapai 78,6%, sementara bad loan sebesar 21,4%. Hal ini mengindikasikan bahwa perusahaan masih belum cukup efektif dalam memprediksi peminjam yang berisiko gagal bayar, maupun yang berhasil membayar. Persentase bad loan yang mencapai 21,4% menunjukkan bahwa potensi kerugian perusahaan masih cukup besar.



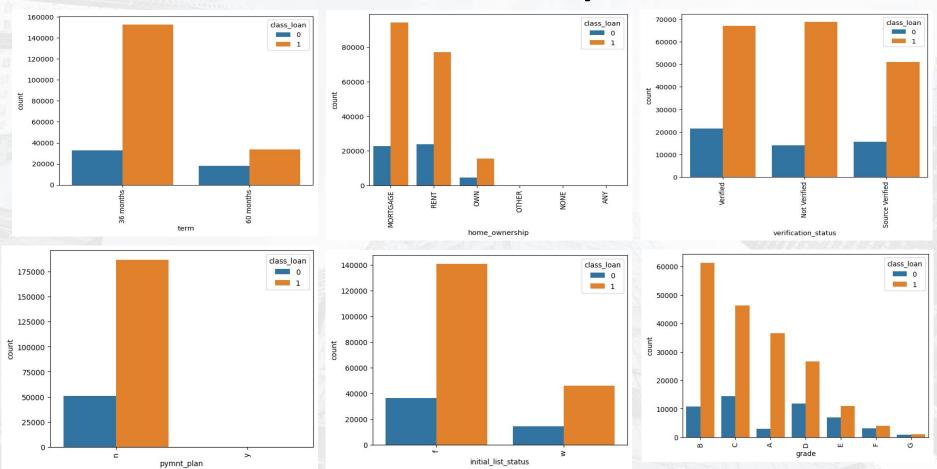


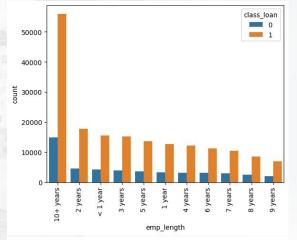
Terlihat bahwa jumlah gagal di perusahaan bayar meningkat setiap tahunnya. Hal ini menunjukkan bahwa seiring berjalannya waktu, perusahaan semakin kesulitan dalam membedakan peminjam layak diberikan yang pinjaman dan yang berisiko gagal bayar.

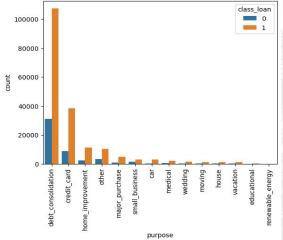


Orang-orang di negara bagian dengan kode CA, TX, FL, dan NY banyak mengajukan pinjaman ke perusahaan, yang menguntungkan perusahaan. Namun, meskipun demikian, tetap ada sebagian yang gagal bayar.

# **Bivariate Analysis**



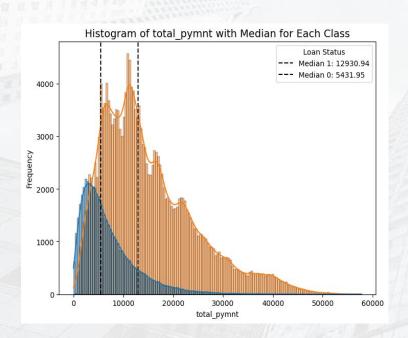




### **Bivariate Analysis**

- Sebagian besar peminjam memilih jangka pembayaran 36 bulan, tetapi kalangan ini juga mendominasi kasus gagal bayar.
- Mayoritas peminjam yang gagal bayar adalah mereka yang menyewa atau mengangsur tempat tinggal.
- Meski status mereka sudah terverifikasi dan didanai oleh beberapa investor, tingkat gagal bayar tetap tinggi.
- Gagal bayar paling banyak terjadi pada peminjam di grade B, C, dan D.
- Meskipun banyak dari mereka memiliki pengalaman kerja lebih dari 10 tahun dan tujuan pinjamannya untuk melunasi hutang sebelumnya, gagal bayar tetap sering terjadi di kalangan tersebut.

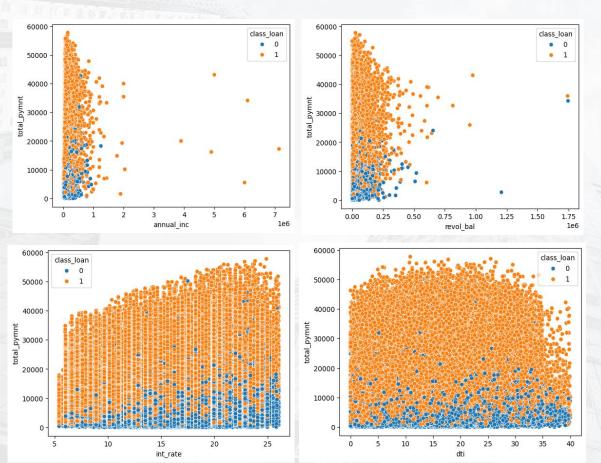
# **Bivariate Analysis**



Peminjam yang gagal bayar cenderung memiliki total pembayaran (total payment) yang jauh lebih rendah dibandingkan dengan mereka yang berhasil melunasi pinjamannya. Median total pembayaran peminjam gagal bayar hanya sebesar 5.431,95, jauh lebih rendah dibandingkan dengan median peminjam yang melunasi pinjamannya, yaitu 12.930,94.

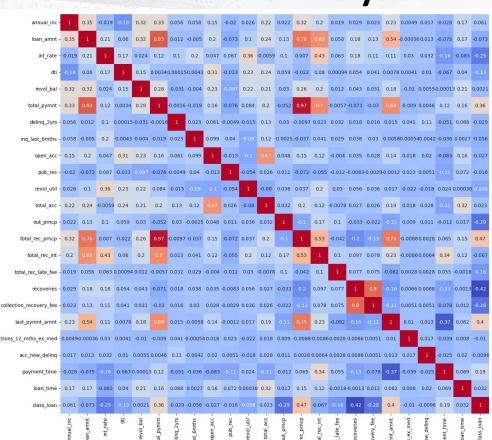
Ini menunjukkan bahwa peminjam yang gagal bayar umumnya berhenti membayar di awal atau pertengahan masa angsuran.

### **Multivariate Analysis**



- Peminjam yang gagal bayar cenderung memiliki total pembayaran rendah, berhenti membayar di awal atau pertengahan masa pinjaman, dan memiliki saldo utang yang lebih besar.
- Gagal bayar terjadi pada total pembayaran yang rendah, meskipun suku bunga dan terhadap rasio hutang pendapatan bervariasi, menunjukkan bahwa faktor-faktor ini tidak cukup untuk menjamin bahwa peminjam akan menjadi good loan.

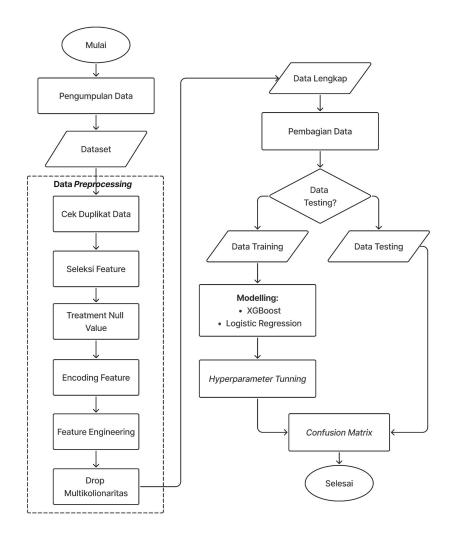
### **Multivariate Analysis**



Beberapa fitur yang memiliki korelasi kuat dengan kelasnya,

- Positif: total\_rec\_prncp, last\_pymnt\_amnt, total\_pymnt (semakin besar pokok pinjaman yang telah dibayar, jumlah pembayaran terakhir dan jumlah pembayaran yang telah dibayar, maka mereka lebih cenderung good loan dan tidak akan gagal bayar)
- Negatif: recoveries, out\_prncp, int\_rate (Semakin besar sisa pokok pinjaman yang belum dibayar, jumlah pemulihan dari gagal bayar, dan biaya pemulihan koleksi, maka peminjam lebih cenderung termasuk kategori bad loan atau berisiko gagal bayar).





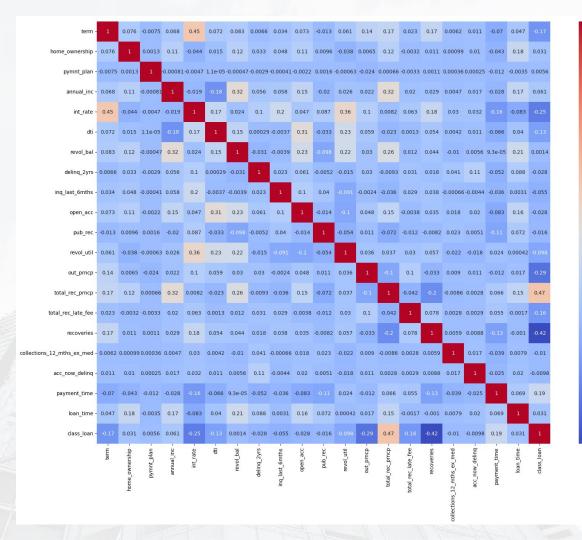
- Tidak ada data duplikat
- Menghapus baris yang mengandung null value (karena jumlah data cukup banyak). Sisanya 227835 baris data yang akan digunakan
- Encoding fitur term, home ownership, grade, pymnt plan.
- Ekstraksi fitur baru.
- Menghapus fitur yang berkorelasi tinggi dengan fitur lainnya, yaitu: verification\_status, loan\_status, purpose, initial\_list\_status, grade, total\_rec\_int, last\_pymnt\_amnt, collection\_recovery\_fee, total\_acc, total\_pymnt, loan\_amnt.
- Pembagian data training dan data testing dengan presentase 80%:20%.
- Pemodelan dengan XGBoost dan Logistic Regression
- Tunning Parameter dengan GridSerachCV.

Selain membuat fitur baru berupa class loan untuk kelas good loan dan bad loan, dibuat fitur baru berupa:

- Payment Time (Durasi Pembayaran):
- Ini mengukur berapa lama waktu yang dibutuhkan peminjam untuk membayar setelah pinjaman diberikan. Mencari selisih antara tahun dan bulan antara tanggal pembayaran terakhir dan tanggal pemberian pinjaman
   Payment Time =(Tahun Pembayaran Terakhir-Tahun Pemberian Pinjaman)
   ×12+(Bulan Pembayaran Terakhir-Bulan Pemberian Pinjaman)
- Loan Time (Durasi Pinjaman):
- Ini mengukur waktu yang telah berlalu sejak peminjam membuka akun kreditnya hingga saat kredit terakhir diambil atau diperiksa oleh perusahaan. mencari selisih antara tahun dan bulan antara tanggal pengambilan kredit terakhir dan tanggal pembukaan akun kredit.

Loan Time=(Tahun Pengambilan Kredit Terakhir-Tahun Pembukaan Akun) ×12+(Bulan Pengambilan Kredit Terakhir-Bulan Pembukaan Akun)

Fitur yang memiliki korelasi lebih dari 0.5 dengan fitur lainnya akan di drop, karena hanya akan menambah redundansi dalam model. Hal ini dapat menyebabkan multikolinearitas, yang dapat mempengaruhi kualitas model dan mengurangi interpretabilitas.





# Metrik yang difokuskan:

### Accuracy:

Akurasi mengukur seberapa banyak prediksi yang benar dibandingkan dengan total prediksi yang dibuat oleh model.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Akurasi memberikan gambaran umum tentang seberapa baik model dalam memprediksi apakah seseorang akan melunasi pinjaman atau tidak

### **Precision:**

Precision mengukur seberapa akurat prediksi model dalam mengidentifikasi bad loan yang benar-benar gagal bayar.

$$Precision = \frac{TP}{TP + FP}$$

Precision fokus pada mengurangi jumlah false positive, yaitu orang yang diprediksi akan membayar, tetapi pada kenyataannya gagal bayar.



### Hasil sebelum di tuning:

### **XGBoost:**

Accuracy (Test Set): 0.971 Accuracy (Train Set): 0.978 Precision (Test Set): 0.976 Precision (Train Set): 0.979 Recall (Test Set): 0.988 Recall (Train Set): 0.992 F1-Score (Test Set): 0.982 F1-Score (Train Set): 0.986 ROC AUC (Test Set): 0.993 ROC AUC (Train Set): 0.997 Recall (Crossval Train): 0.993 Recall (Crossval Test): 0.988

### **Logistic Regression:**

Accuracy (Test Set): 0.948 Accuracy (Train Set): 0.949 Precision (Test Set): 0.954 Precision (Train Set): 0.953 Recall (Test Set): 0.982 Recall (Train Set): 0.983 F1-Score (Test Set): 0.968 F1-Score (Train Set): 0.968 ROC AUC (Test Set): 0.973 ROC AUC (Train Set): 0.973 Recall (Crossval Train): 0.983 Recall (Crossval Test): 0.984

# Parameter yang di tuning:

### **XGBoost:**

- n\_estimators: Jumlah pohon dalam model, 100-150 pohon moderat.
- max\_depth: Kedalaman pohon, 6-7 untuk menghindari overfitting.
- learning\_rate: Tingkat pembelajaran,
   0.05-0.1 agar model tidak belajar terlalu cepat.
- min\_child\_weight: Minimum sampel per node, 5-10 untuk mencegah overfitting.
- subsample: Persentase data yang digunakan, 80% untuk generalisasi.
- colsample\_bytree: Persentase fitur yang dipilih, 80% untuk mengurangi overfitting.
- gamma: Batas penurunan loss function,
   0-0.1 untuk stabilitas pohon.

### **Logistic Regression:**

- C: Kekuatan regularisasi, diuji pada 0.01, 1, dan 100 untuk mengontrol overfitting.
- solver: Algoritma yang digunakan untuk optimasi, menggunakan 'liblinear' untuk model kecil.
- penalty: Jenis penalti yang digunakan, 'l2' untuk regularisasi yang lebih stabil.
- max\_iter: Jumlah iterasi maksimum, diset 100 untuk memastikan konvergensi model.

# Hasil setelah di tuning:

### **XGBoost:**

```
Accuracy (Test Set): 0.971
Accuracy (Train Set): 0.974
Precision (Test Set): 0.974
Precision (Train Set): 0.976
Recall (Test Set): 0.989
Recall (Train Set): 0.992
F1-Score (Test Set): 0.982
F1-Score (Train Set): 0.984
ROC AUC (Test Set): 0.993
ROC AUC (Train Set): 0.995
Recall (Crossval Train): 0.992
Recall (Crossval Test): 0.989
```

### **Logistic Regression:**

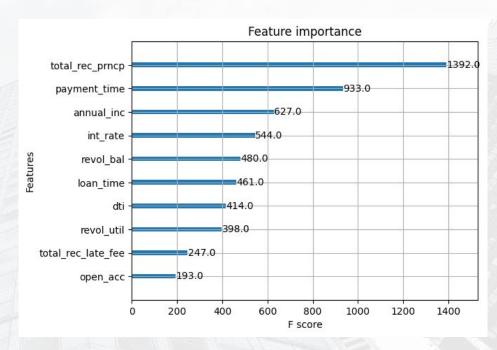
```
Accuracy (Test Set): 0.944
Accuracy (Train Set): 0.944
Precision (Test Set): 0.947
Precision (Train Set): 0.947
Recall (Test Set): 0.985
Recall (Train Set): 0.985
F1-Score (Test Set): 0.965
F1-Score (Train Set): 0.966
ROC AUC (Test Set): 0.970
ROC AUC (Train Set): 0.969
Recall (Crossval Train): 0.984
Recall (Crossval Test): 0.982
```

Dilakukan tunning parameter untuk memastikan model menggunakan parameter parameter yang tepat dan terbaik. Hasil menunjukkan bahwa model masih berperforma baik. Dengan accuracy dan precision yang tinggi pada XGBoost.



XGBoost mampu memberikan prediksi yang akurat dengan jumlah peminjam yang diprediksi sebagai good loan, namun sebenarnya termasuk dalam bad loan, tetap rendah.

# Top 10 fitur yang penting dalam memprediksi risiko kredit.



# Rekomendasi

- 1. Identifikasi Risiko Berdasarkan Total Pembayaran Tertunda
- Data menunjukkan peminjam yang gagal bayar memiliki total pembayaran jauh lebih rendah, menunjukkan bahwa mereka berhenti membayar lebih awal.
- Action: Buat sistem pemantauan untuk mendeteksi pembayaran yang tertunda pada tahap awal angsuran. Kirim pengingat atau tawarkan opsi restrukturisasi jika peminjam menunjukkan tanda keterlambatan pembayaran dalam 2–3 angsuran pertama.
- 2. Tingkatkan Validasi untuk Peminjam di Grade Risiko Tinggi (B, C, D)
- Grade B, C, dan D menunjukkan tingkat gagal bayar yang tinggi meskipun telah melalui proses verifikasi.
- Action: Tingkatkan kriteria evaluasi kredit untuk grade ini, seperti mempertimbangkan rasio DTI, saldo utang (revol\_bal), dan jumlah pinjaman sebelumnya dalam keputusan pemberian pinjaman.

# Rekomendasi

- 3. Segmentasi Berdasarkan Tujuan Pinjaman
  - Sebagian besar peminjam menggunakan pinjaman untuk melunasi hutang sebelumnya, tetapi kelompok ini juga mendominasi gagal bayar.
- Action: Tambahkan analisis kemampuan pembayaran tambahan pada kelompok ini, dengan menilai jumlah hutang yang ingin dilunasi terhadap pendapatan tahunan dan pengeluaran mereka.
- 4. Penguatan Kebijakan untuk Wilayah Risiko Tinggi (CA, TX, FL, NY)
  - Wilayah ini memiliki jumlah pengajuan tinggi namun tingkat gagal bayar signifikan.
  - Action: Implementasikan kebijakan yang lebih ketat di wilayah ini, seperti menaikkan ambang batas minimum pendapatan atau menurunkan jumlah maksimum pinjaman untuk peminjam berisiko.
- 5. Implementasi Model Prediksi XGBoost untuk menilai resiko kredit.
  - Update secara berkala untuk menyesuaikan dengan perubahan pola data dan kondisi ekonomi, guna memastikan akurasi dan relevansi prediksi.

### Simulasi Bisnis

- Tanpa Model
- Berdasarkan data historis:
- 21.4% dari total pinjaman gagal bayar.
- Jika total portofolio pinjaman adalah Rp100 miliar, maka kerugian mencapai:
   21.4% × Rp100miliar= Rp21,4miliar
- Dengan Model (XGBoost) Kinerja Model:
- Accuracy (Test Set): 97.1%
   Artinya, model mampu memprediksi dengan benar 97.1% dari total data.
- Precision (Test Set): 97.4%
   Dari semua pinjaman yang diprediksi good loan, 97.4% benar-benar layak bayar.
- Kerugian dengan Model:
   Model salah memprediksi 2.6% (100% 97.4%) pinjaman bad loan sebagai good loan.
- Jika total portofolio pinjaman adalah Rp100 miliar, kerugian akan menjadi:
   2.6%×Rp100miliar=Rp2,6miliar

### **Simulasi Bisnis**

### 3. Keuntungan Menggunakan Model:

### Pengurangan Kerugian:

- Tanpa model, kerugian adalah Rp21,4 miliar.
- Dengan model, kerugian turun menjadi Rp2,6 miliar.
- Penghematan sebesar:
   Rp21,4miliar-Rp2,6miliar=Rp18,8miliar

### Efisiensi Keuangan:

Perusahaan berhasil mengurangi kerugian sebesar 87.85% dengan menggunakan model.

# Thank You!

Rakamin