

Credit Risk Prediction Made Smarter with Machine Learning

VIX Data Scientist ID/X Partners



Created by:

Rofik

Email: rofikn4291@gmail.com

linkedIn: Rofik Rofik

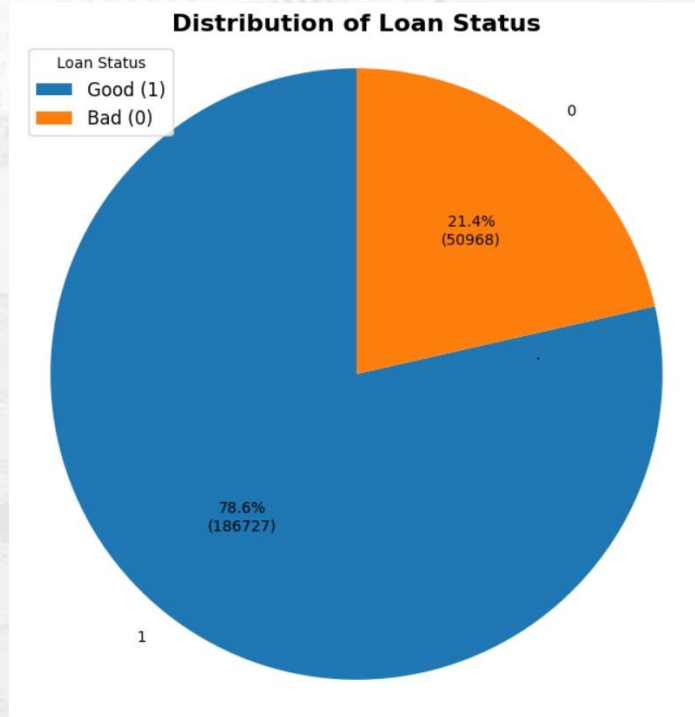
"I'm Rofik, with two years of experience in **Data Science, Data Analysis, and Machine Learning** through learning, research, and hands-on projects. I have experience in **data mining, data analysis, preprocessing, and making predictions**. I'm familiar with tools like **Python, SQL, Pandas, Numpy, Matplotlib, 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."

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?



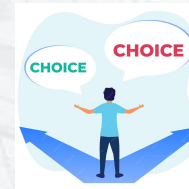
- Bagaimana karakteristik orang yang good loan dan bad loan (gagal bayar)? →
- Bagaimana Machine Learning membantu dalam prediksi credit risk? →
- Apa saja faktor penting dalam keputusan prediksi credit risk? →



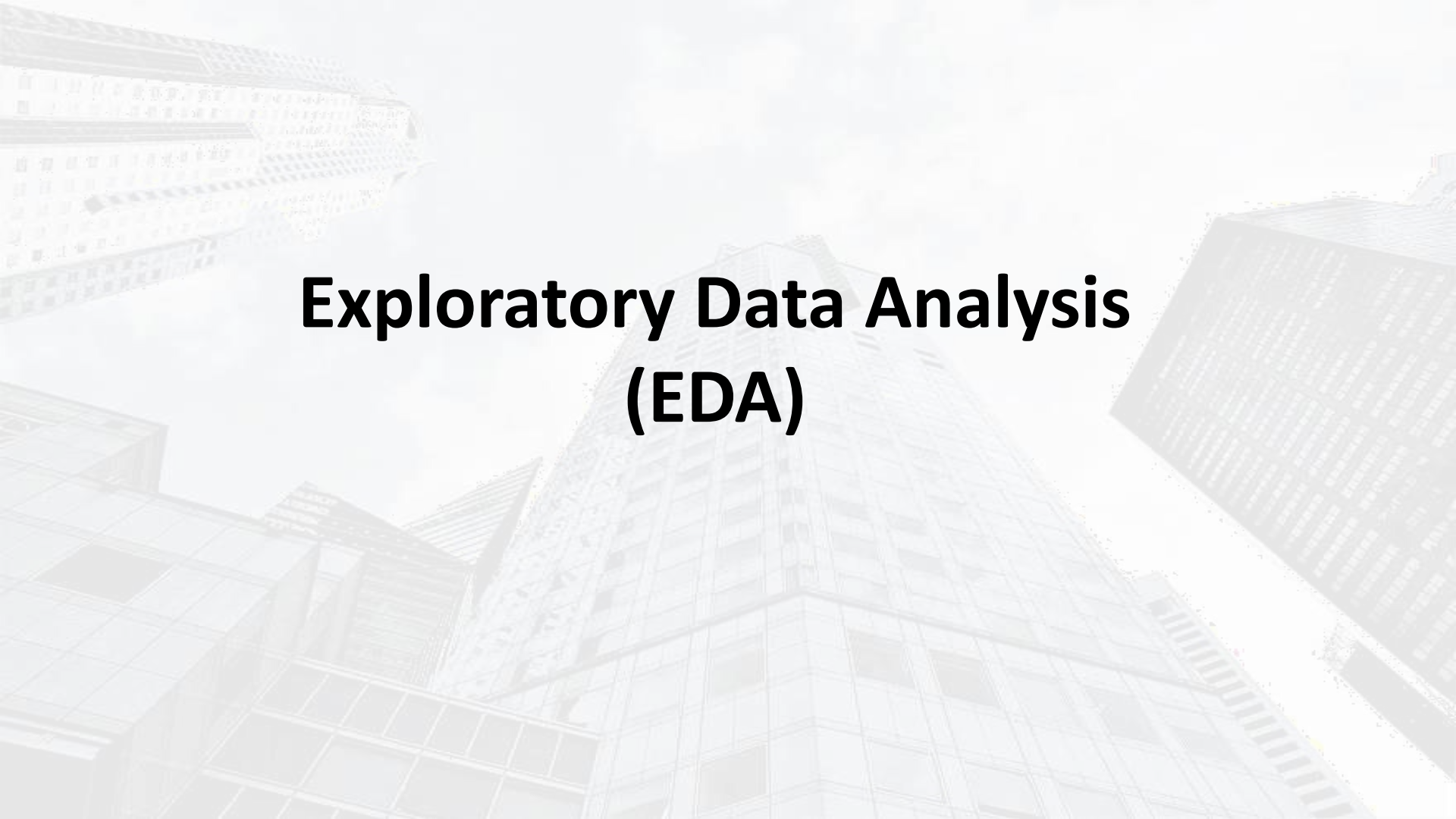
Exploratory
Data Analysis



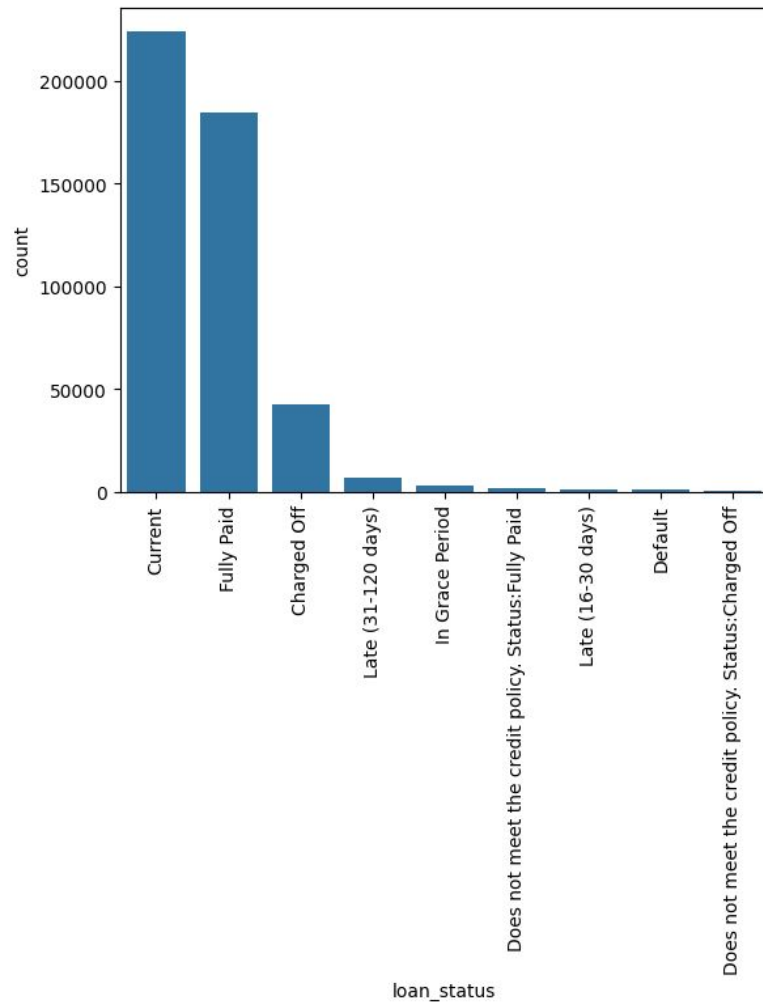
Model
Prediktif



Feature
Importance

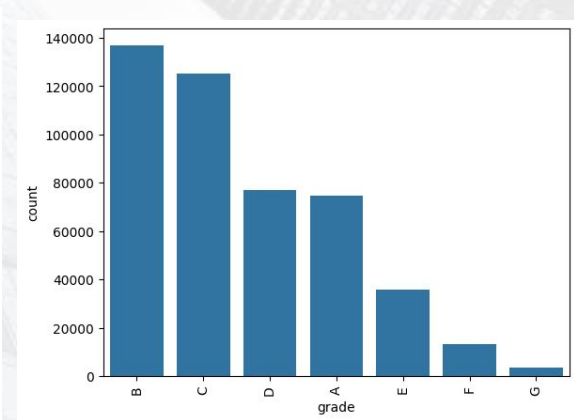
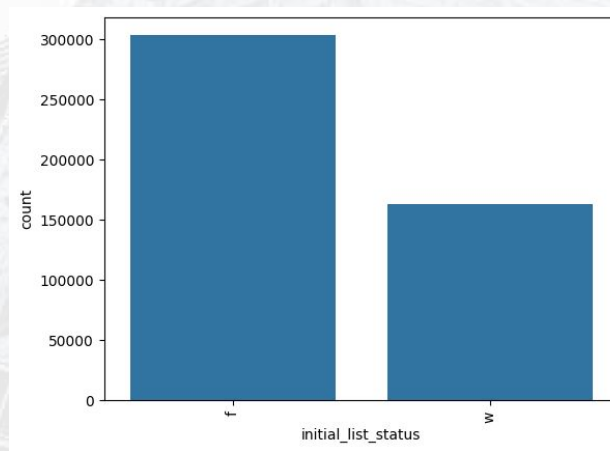
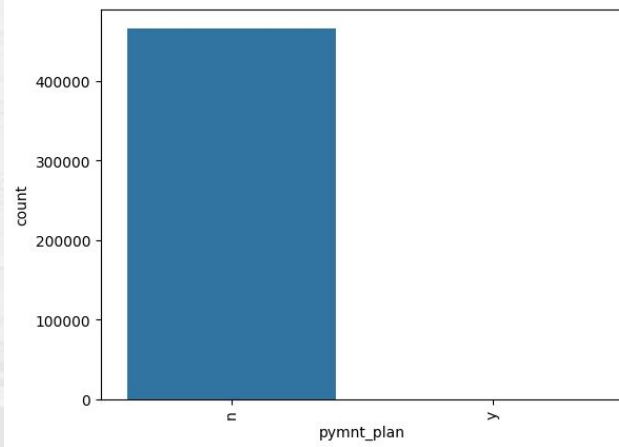
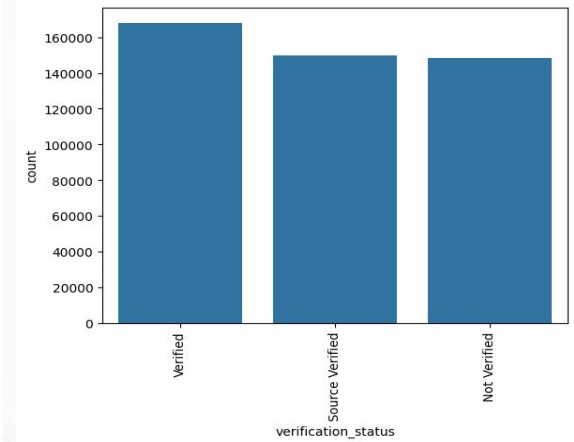
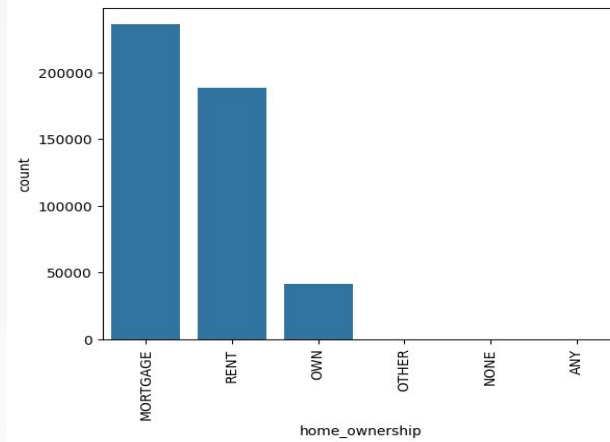
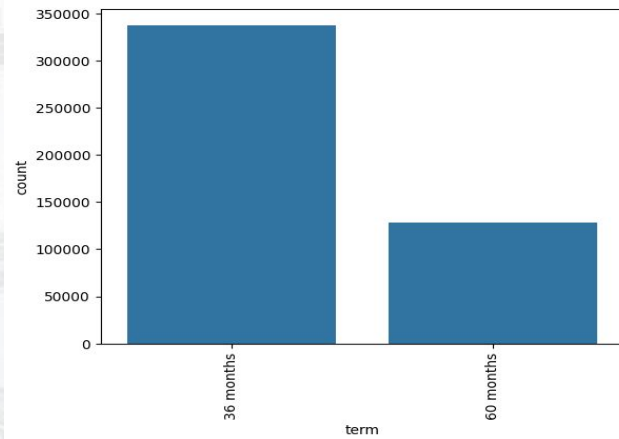


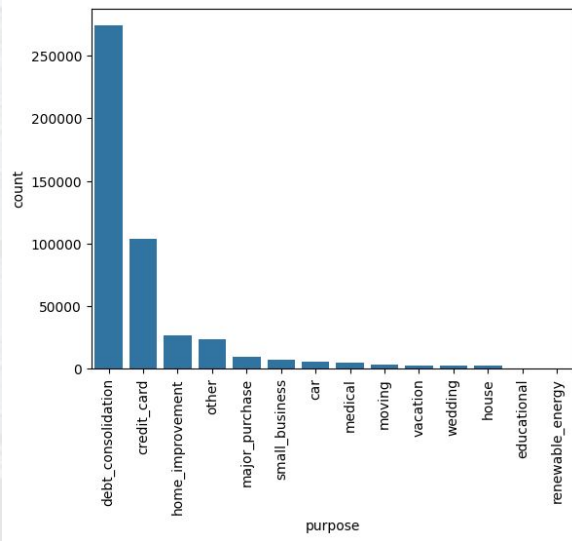
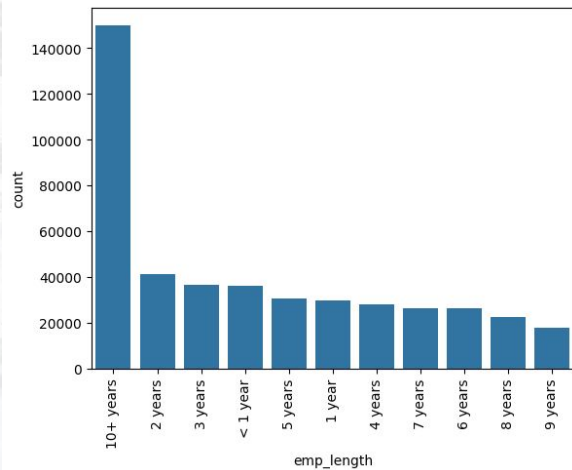
Exploratory Data Analysis (EDA)



- 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.**

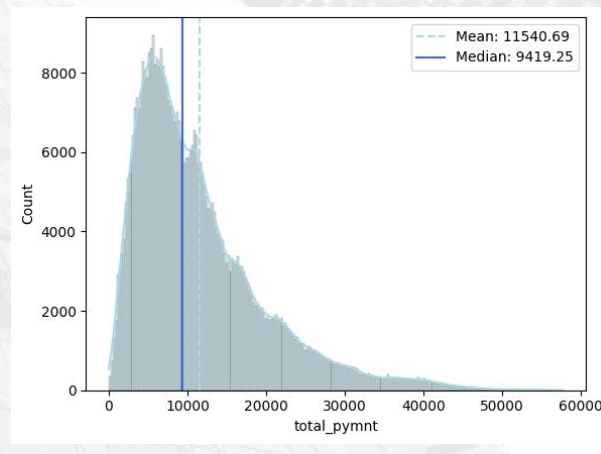
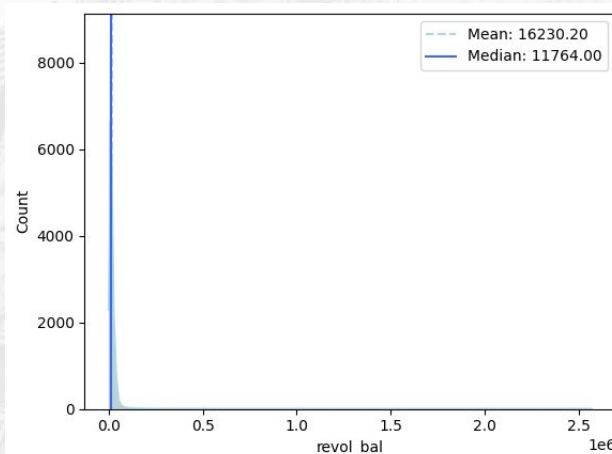
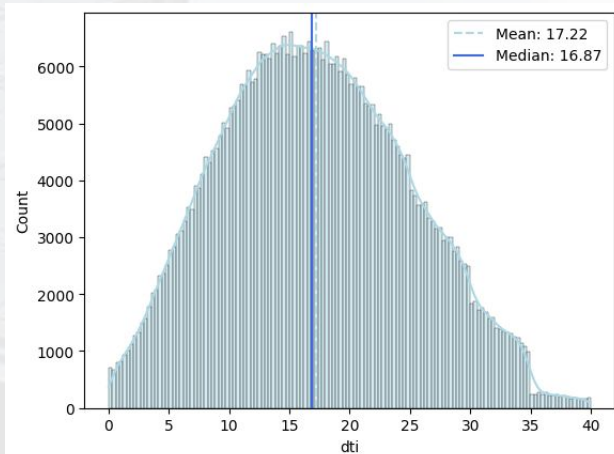
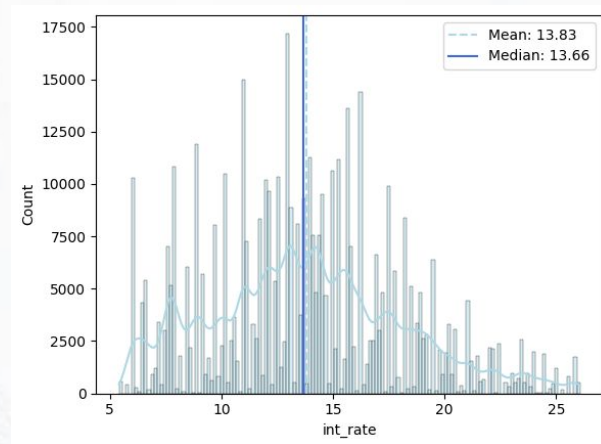
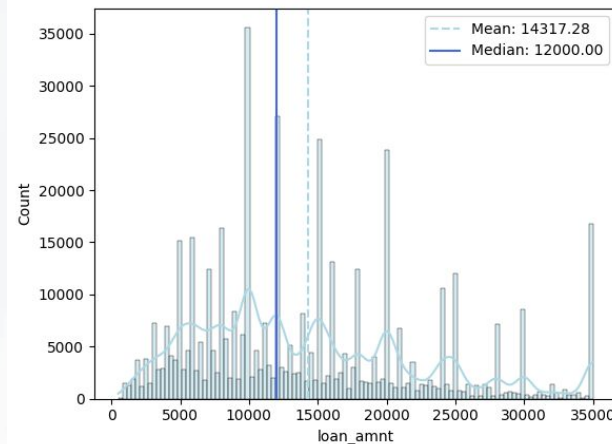
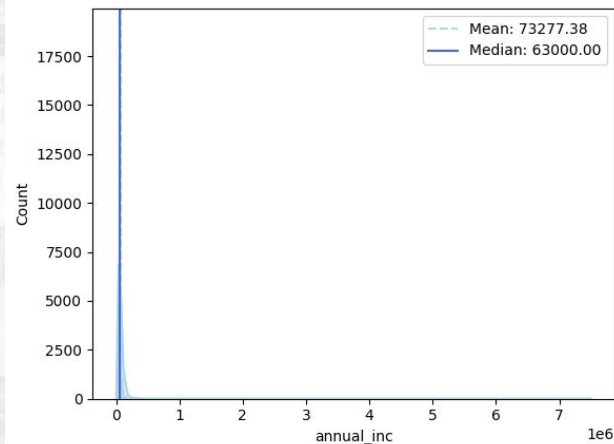
Univariate Analysis





- 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

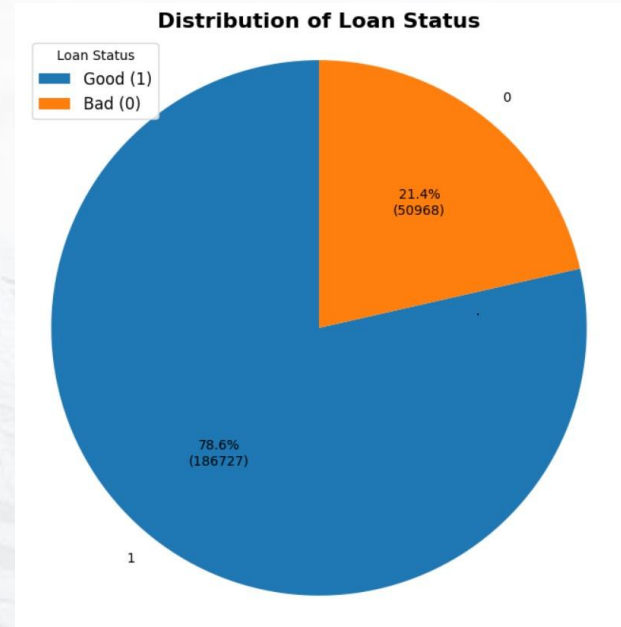


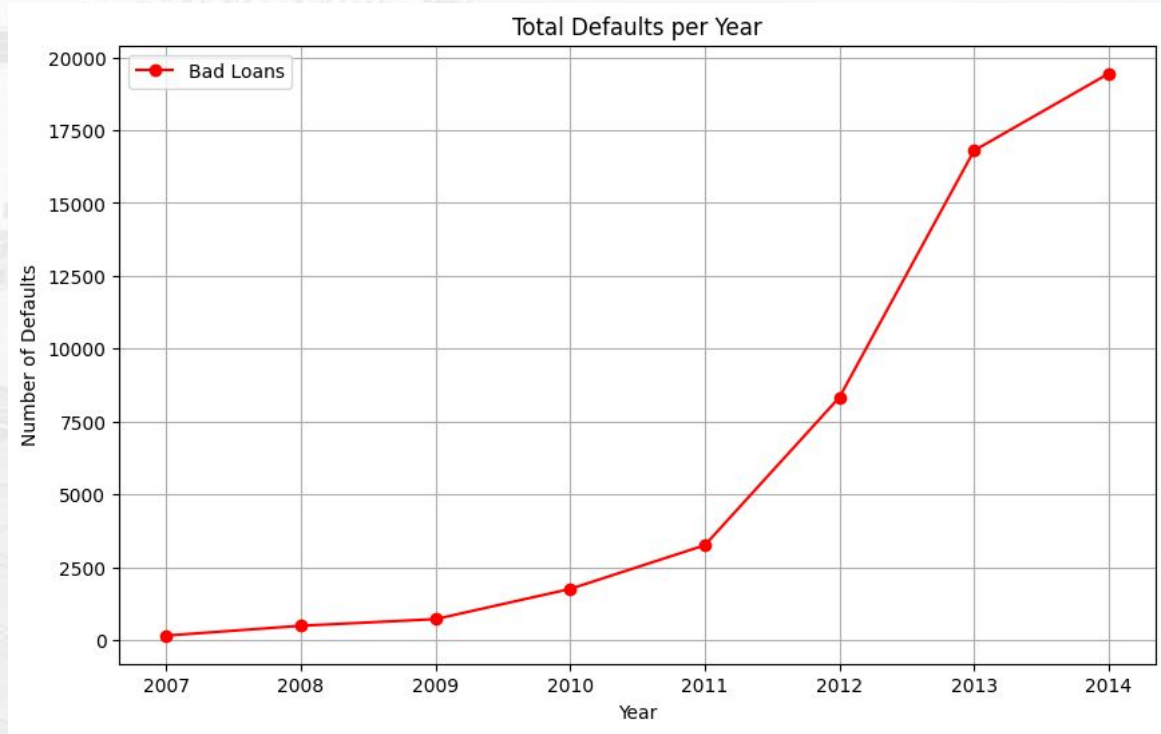
- Beberapa peminjam memiliki total pendapatan tahunan yang sangat tinggi, meskipun rata-rata annual income adalah 73.227.
- Jumlah pinjaman bervariasi, dengan rata-rata sebesar 14.317.
- Suku bunga (int rate) juga bervariasi, dengan rata-rata 13,83%.
- DTI (Debt-to-Income ratio) (Rasio hutang dan pendapatan 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

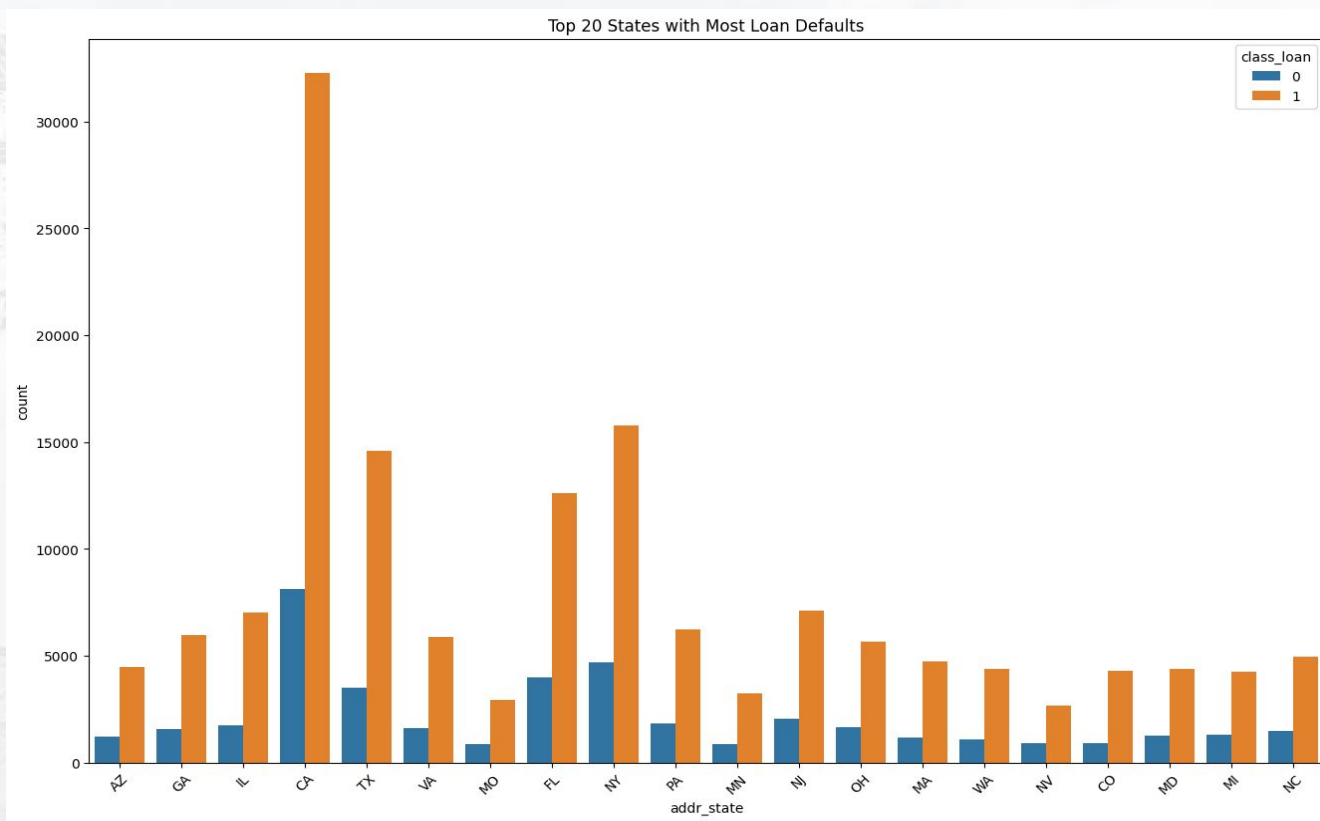
1. Fully Paid: Pinjaman sudah dilunasi sepenuhnya.
 2. Current: Pinjaman sedang berjalan dan pembayaran tepat waktu.
 3. In Grace Period: Pinjaman dalam periode tenggang setelah jatuh tempo.
 4. Charged Off: Pinjaman dianggap gagal bayar dan tak bisa dilunasi.
 5. Default: Peminjam gagal bayar sesuai 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.
 9. 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 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.**



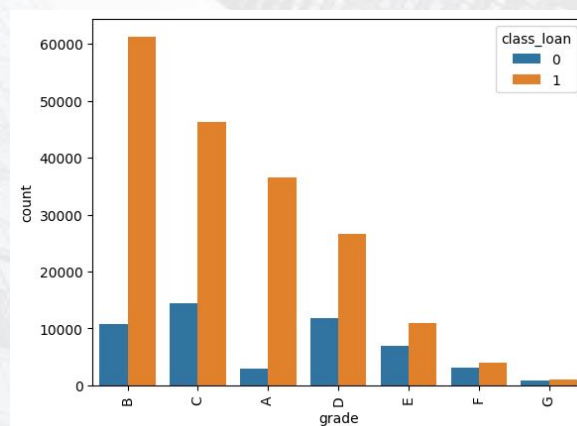
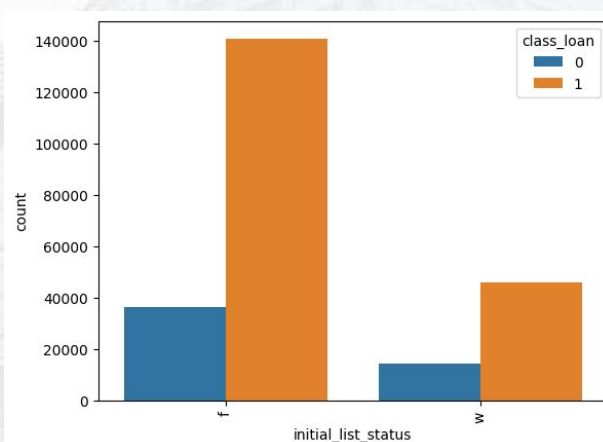
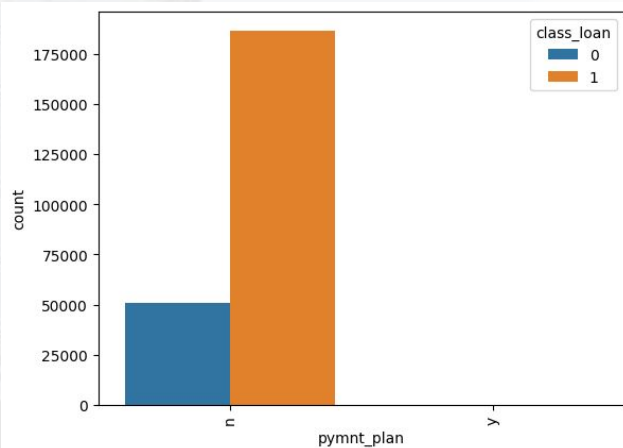
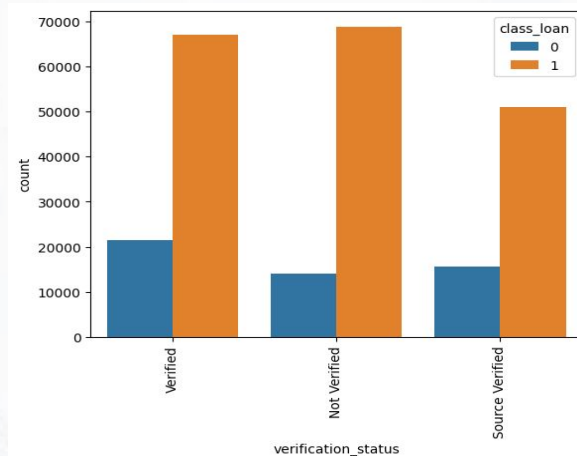
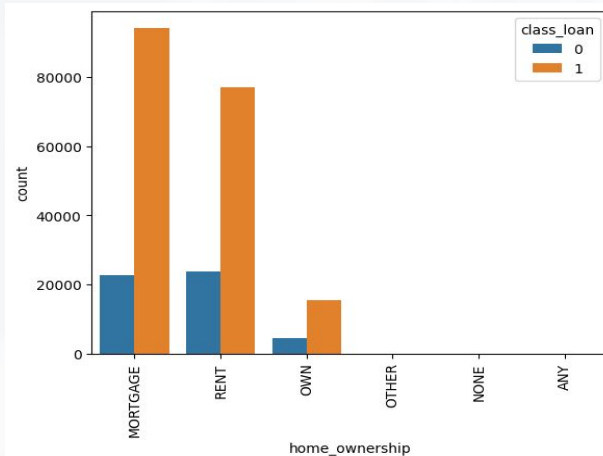
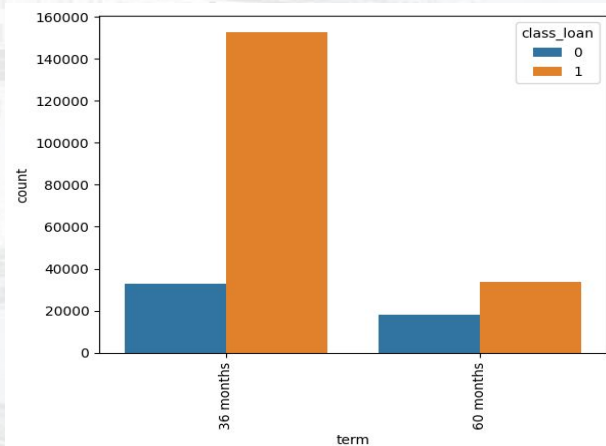


Terlihat bahwa **jumlah gagal bayar di perusahaan meningkat setiap tahunnya**. Hal ini menunjukkan bahwa seiring berjalannya waktu, perusahaan semakin kesulitan dalam membedakan peminjam yang layak diberikan pinjaman dan yang berisiko gagal bayar.

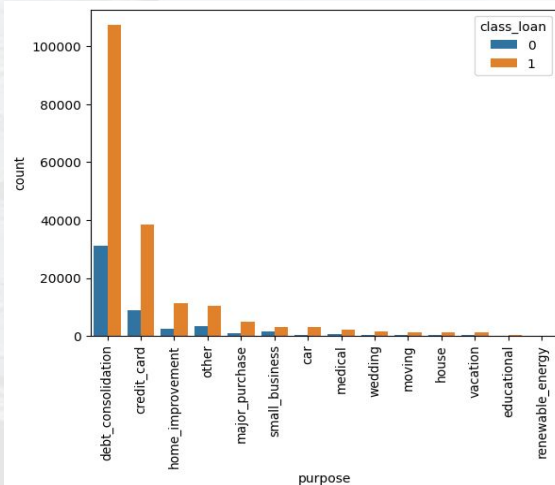
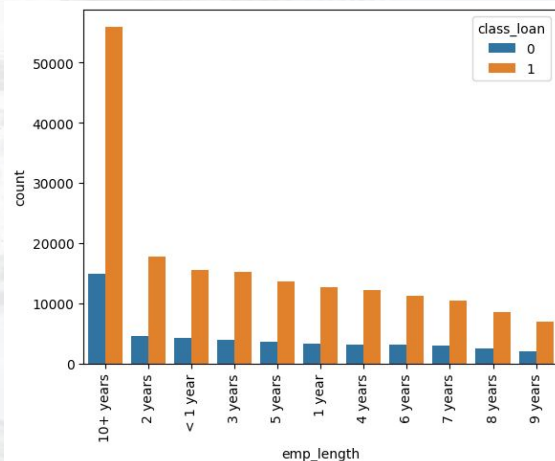


Orang-orang di negara bagian dengan kode CA, TX, FL, dan NY banyak mengajukan pinjaman ke perusahaan, yang menguntungkan perusahaan. Namun, meskipun demikian, tetap ada sebagian yang gagal bayar.

Bivariate Analysis

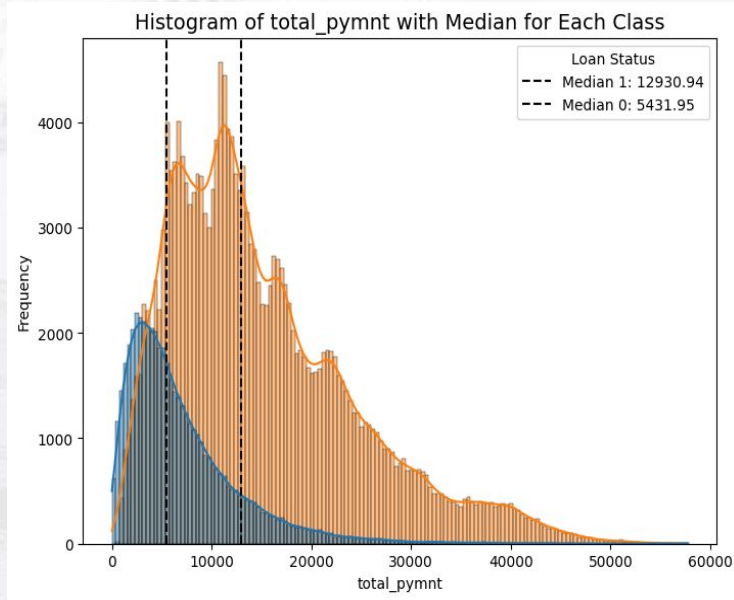


Bivariate Analysis



- 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.

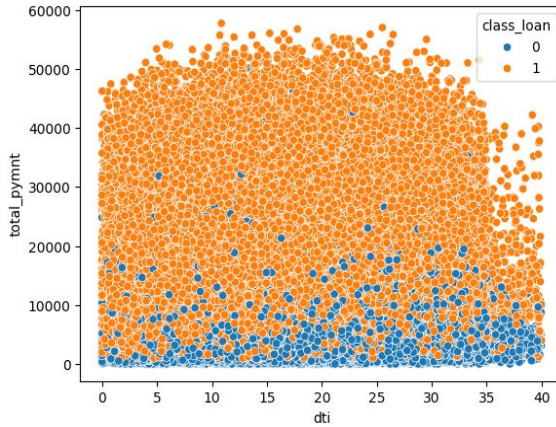
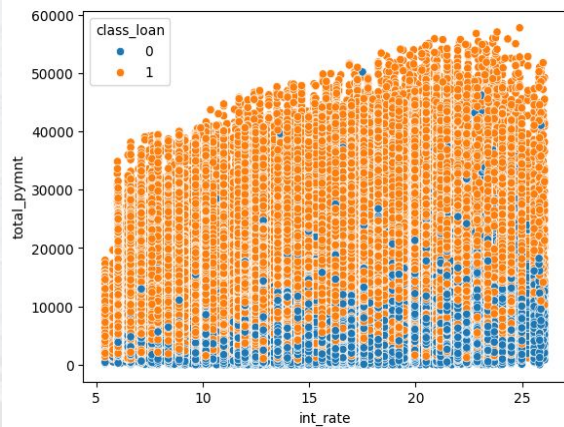
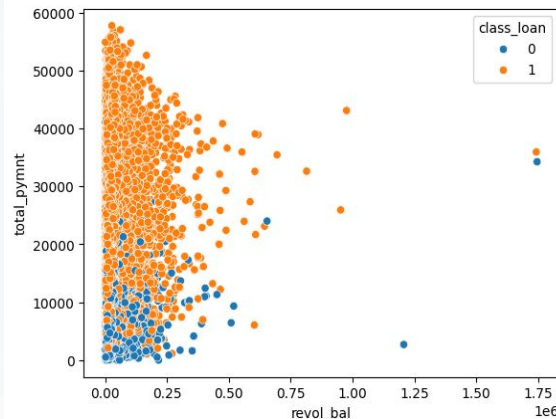
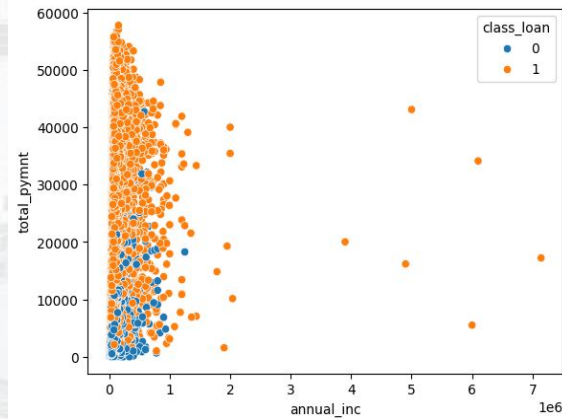
Bivariate Analysis



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

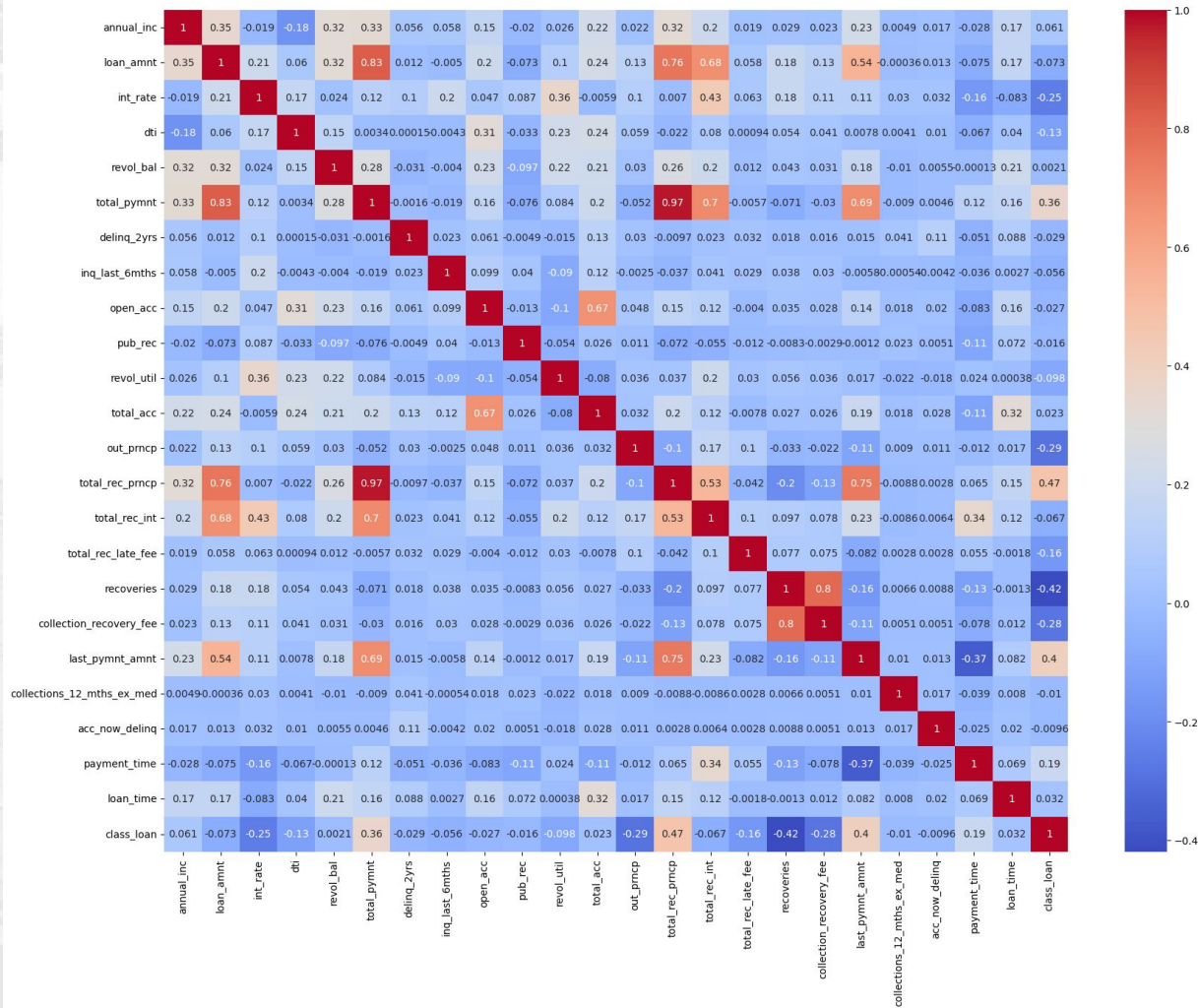


- Peminjam yang gagal bayar cenderung memiliki total pembayaran rendah, berhenti membayar di awal atau pertengahan masa pinjaman, dan memiliki saldo utang yang lebih besar.
- Gagal bayar terjadi pada total pembayaran yang rendah, meskipun suku bunga dan rasio hutang terhadap pendapatan bervariasi, menunjukkan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

Multivariate Analysis

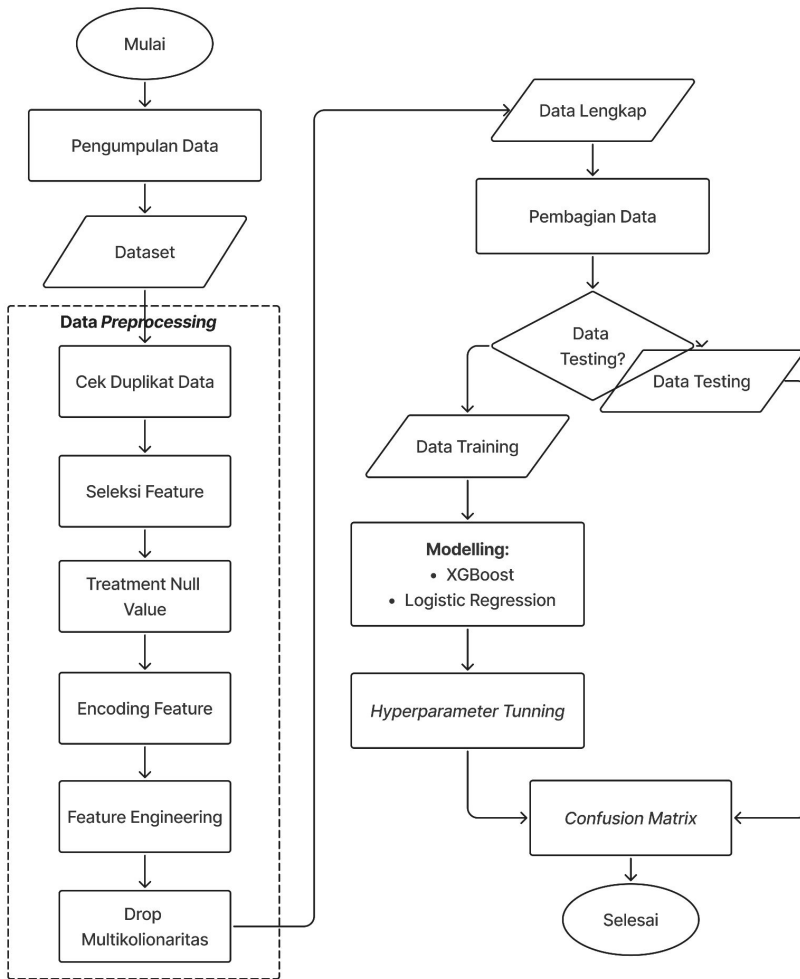
Beberapa fitur yang memiliki korelasi kuat dengan kelasnya,

- Positif: total_rec_prncp, last_pymnt_amnt, total_pymnt
- Negatif: recoveries, out_prncp, int_rate.





Data Preprocessing



- Tidak ada data duplikat
- Hanya menggunakan fitur yang cukup relevan dengan penilaian risiko kredit yaitu berjumlah 32.
- Menghapus baris yang mengandung null value (karena jumlah data cukup banyak).
- 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
- Tuning Parameter dengan GridSerachCV.

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

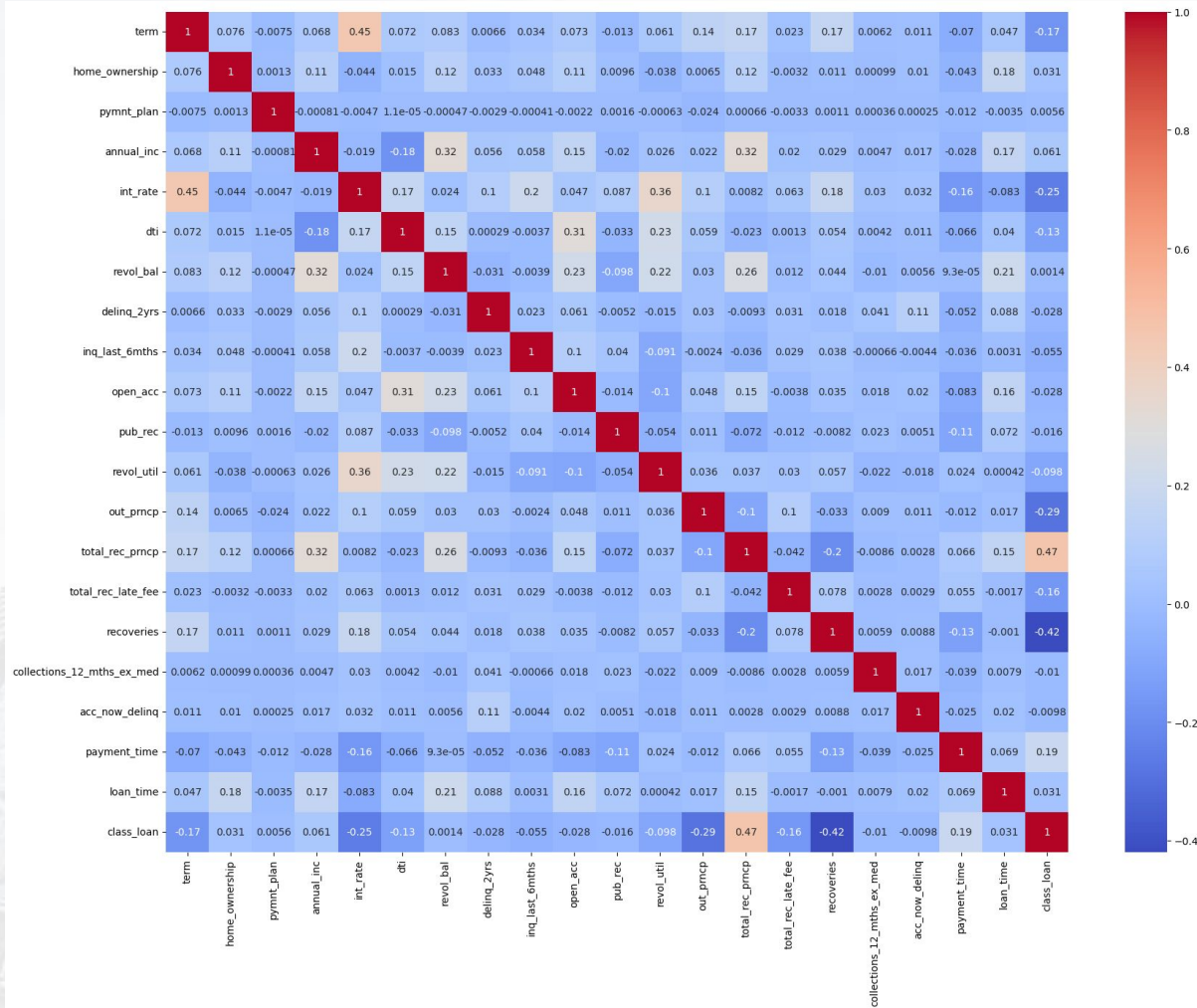
- Payment Time (Durasi Pembayaran):
- Ini mengukur berapa lama waktu yang dibutuhkan peminjam untuk membayar setelah pinjaman diberikan. Mencari selisih antara tahun dan bulan antara tanggal pembayaran terakhir dan tanggal pemberian pinjaman

$$\text{Payment Time} = (\text{Tahun Pembayaran Terakhir} - \text{Tahun Pemberian Pinjaman}) \times 12 + (\text{Bulan Pembayaran Terakhir} - \text{Bulan Pemberian Pinjaman})$$

- 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.

$$\text{Loan Time} = (\text{Tahun Pengambilan Kredit Terakhir} - \text{Tahun Pembukaan Akun}) \times 12 + (\text{Bulan Pengambilan Kredit Terakhir} - \text{Bulan Pembukaan Akun})$$

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.





Modelling

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.

A low-angle, upward-looking perspective of several modern skyscrapers. The buildings are rendered in a light, desaturated gray tone, creating a sense of height and architectural scale. The sky is filled with soft, white clouds, and the overall composition is clean and minimalist.

Evaluasi

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:

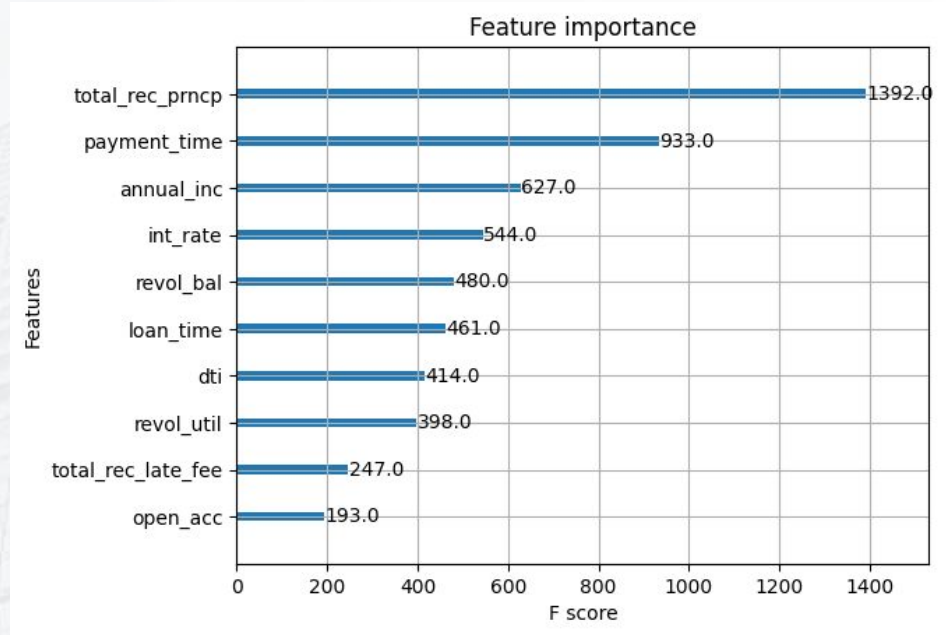
```
Accuracy (Test Set): 0.944
Accuracy (Train Set): 0.944
Precision (Test Set): 0.947
Precision (Train Set): 0.947
Recall (Test Set): 0.985
Recall (Train Set): 0.985
F1-Score (Test Set): 0.965
F1-Score (Train Set): 0.966
ROC AUC (Test Set): 0.970
ROC AUC (Train Set): 0.969
Recall (Crossval Train): 0.984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

Dilakukan tuning 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.

Top 10 fitur yang penting dalam memprediksi risiko kredit.



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



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.

Rekomendasi

1. Implementasi Model XGBoost
Gunakan model XGBoost sebagai alat prediksi risiko untuk menyaring peminjam berisiko tinggi sebelum persetujuan pinjaman, terutama di grade B, C, dan D.
2. Fokus pada Fitur Utama
Prioritaskan `total_rec_prncp`, `last_pymnt_amnt`, dan `total_pymnt` sebagai indikator utama dalam evaluasi risiko kredit dan pemantauan pembayaran.
3. Update Model Secara Berkala
Lakukan pelatihan ulang model secara periodik untuk mengakomodasi perubahan karakteristik data atau tren baru, sehingga akurasi prediksi tetap optimal.

Thank You!