

Credit Risk Prediction

ID/X Partners - Data Scientist

Presented by
Daviro Yota Nagasan Wahyudi



Daviro Yota Nagasan Wahyudi

About Me

Fresh graduate with a Bachelor's degree in Civil Engineering from Bandung Institute of Technology, graduating with honors (Cum Laude). Successfully completed a Data Science: Machine Learning BootCamp at Rakamin Academy. Proficient in SQL, Python, Data Visualization, Machine Learning. Prepared to excel in data analyst-related roles.

Surabaya, Jawa Timur, Indonesia

yota.dorez@gmail.com

linkedin.com/in/daviroyota/

My Experience

Project Based Internship

Kimia Farma x Rakamin Academy
Apr 2024 – May 2024



Project Based Internship

PT Bank Muamalat Indonesia Tbk
x Rakamin Academy
Mar 2024 – Apr 2024



Quantity Surveyor

At CV. Surabaya Satu
Aug 2022 – Feb 2023



Project Portfolio

Problem Statement

Perusahaan multifinance perlu meningkatkan keakuratan penilaian risiko kredit untuk **mengoptimalkan keputusan bisnis dan mengurangi kerugian**. 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.

Role

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.

Project Portfolio

Goal

Meningkatnya akurasi memprediksi risiko kredit client. Asumsi meningkat 2.5%.

Objective

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

Business Metrics

credit risk prediction accuracy

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.



Data Understanding

Data Overview

- Data terdiri dari 466285 baris dan 75 kolom
- Terdapat 52 kolom numerik dan 23 kolom kategorikal
- Terdapat 40 kolom numerik yang memiliki missing values
- Tidak terdapat duplikat baris
- Beberapa kolom memiliki missing values 100% (akan di drop)

Data Understanding

Dari kolom-kolom data yang ada, beberapa kolom secara definisi yang telah dipahami bisa dirangkum sebagai berikut. Kolom ini mungkin atau tidak akan digunakan sebagai features dalam pengolahan data ini.

1. Distribusi Jumlah Pinjaman (loan_amnt):

- Jumlah pinjaman paling umum: \$10.000. Diikuti beberapa pinjaman \$12.000, \$15.000.
- Jumlah pinjaman yang lebih rendah lebih populer di kalangan pemohon.

2. Suku Bunga (int_rate):

- Suku bunga umum: 12,99%, Diikuti beberapa suku bunga lainnya 10,99%, 15,61%.
- Gambaran rentang suku bunga yang biasanya ditawarkan kepada pemohon.

Data Understanding

3. Distribusi Pendapatan Tahunan (annual_inc):

- Pendapatan tahunan umum: \$60.000. Diikuti pendapatan lainnya \$50.000, \$65.000.
- Mencerminkan profil pendapatan pemohon.

4. Rasio Utang Terhadap Pendapatan (dti):

- Rasio umum: 14,40, 19,20, 12,00.
- Menunjukkan berapa banyak pendapatan yang digunakan untuk membayar utang.

5. Pembayaran Terlambat dalam 2 Tahun Terakhir (delinq_2yrs):

- Sebagian besar pemohon (sekitar 382.954) tidak memiliki pembayaran terlambat.
- Beberapa memiliki hingga 29 pembayaran terlambat.

6. Catatan Publik (pub_rec):

- Sebagian besar pemohon (404.893) tidak memiliki catatan publik. Beberapa memiliki hingga 40 catatan publik.

7. Total Kredit Tersedia (total_rev_hi_lim):

- Batas kredit umum: \$15.000, \$13.500, \$10.000.
- Menunjukkan jumlah kredit yang tersedia bagi pemohon.

8. Pinjaman Tertunggak dalam 12 Bulan Terakhir (collections_12_mths_ex_med):

- Sebagian besar pemohon (sekitar 462.226) tidak memiliki pinjaman tertunggak.

Data Understanding

9. Jangka Waktu Pinjaman (term):

- Sebagian besar pinjaman memiliki jangka waktu 36 bulan, dengan sebagian kecil memiliki jangka waktu 60 bulan.
- Menunjukkan preferensi mayoritas peminjam terhadap jangka waktu pinjaman yang lebih pendek.

10. Peringkat dan Subperingkat Pinjaman (grade dan subgrade):

- Sebagian besar pinjaman memiliki peringkat B dan C, dengan subperingkat B3 dan B4 paling umum.
- Mencerminkan profil risiko yang beragam di antara pemohon pinjaman.

11. Pekerjaan Peminjam (emp_title) dan Lama Bekerja (emp_length):

- Mayoritas peminjam adalah guru, manajer, atau perawat terdaftar, dengan kebanyakan memiliki lebih dari 10 tahun pengalaman kerja.
- Menunjukkan bahwa mayoritas pemohon adalah pekerja dengan pengalaman kerja yang solid.

12. Status Kepemilikan Rumah (home_ownership):

- Mayoritas pemohon memiliki rumah dengan hipotek, diikuti oleh yang menyewa.
- Hanya sedikit yang memiliki rumah tanpa hipotek.

13. Status Verifikasi (verification_status):

- Sebagian besar pemohon memiliki status verifikasi atau sumber verifikasi yang sudah diverifikasi.
- Menunjukkan bahwa mayoritas pemohon telah melewati proses verifikasi identitas dan pendapatan.

Data Understanding

14. Tujuan Pinjaman (purpose):

- Sebagian besar pinjaman digunakan untuk konsolidasi utang atau pembayaran kartu kredit, diikuti oleh perbaikan rumah dan tujuan lain seperti pembelian besar.

15. Status Pinjaman (loan_status):

- Sebagian besar pinjaman masih aktif (current) atau sudah lunas, dengan sebagian kecil mengalami gagal bayar atau pembayaran terlambat.

Feature Engineering

Feature Engineering akan dilakukan pada langkah pertama sebelum EDA terkhususkan untuk **Label Defining**. Disebabkan target / label yang berada pada feature '**loan_status**' masih dalam beberapa status, meliputi:

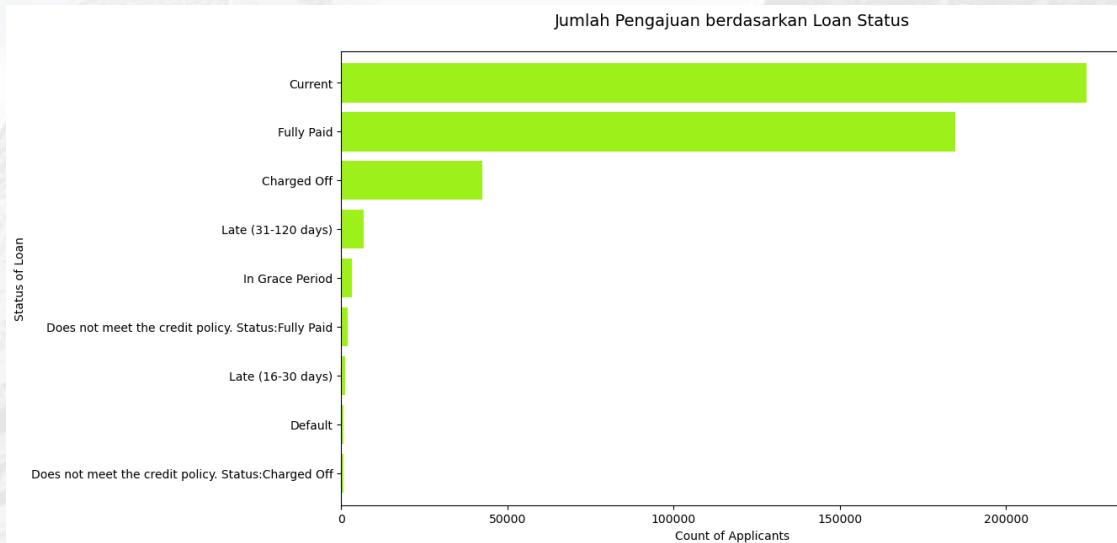
- Current
- Fully Paid
- Charged Off
- Late (16-30 days)
- Late (31-120 days)
- Default
- In Grace Period
- Does not meet the credit policy. Status: Fully Paid
- Does not meet the credit policy. Status: Charged Off

Pada project ini, **pinjaman berisiko buruk** akan diberikan untuk mereka yang memiliki **keterlambatan pembayaran lebih dari 30 Hari**. Status `Current`, `Fully Paid`, `In Grace Period`, `Does not meet the credit policy. Status:Fully Paid`, dan `Late (16-30 days)` yang akan dikategorikan sebagai 'Good risk' dan sisanya sebagai 'Bad risk'. **Sisa Feature Engineering akan dilanjutkan setelah EDA.**

Exploratory Data Analysis

Jumlah Pengajuan berdasarkan Loan Status

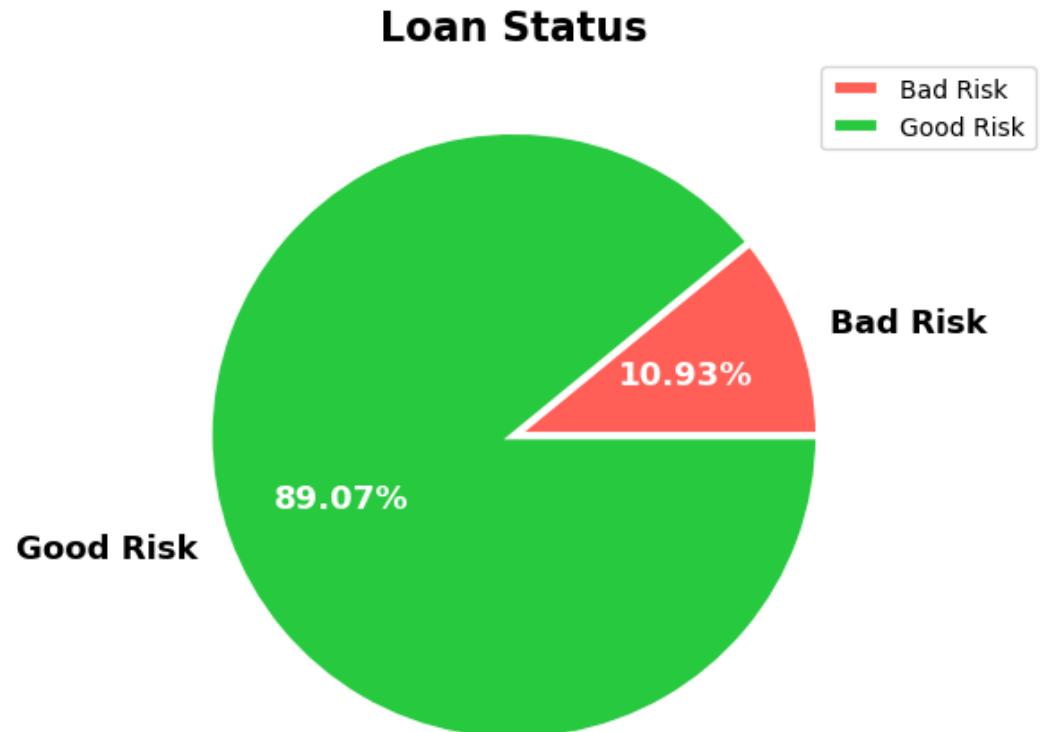
- Status Pinjaman Berlangsung:** Sekitar 48% dari pemohon, yang setara dengan sekitar 224.226 orang, memiliki status pinjaman "Current".
- Status Pinjaman Lunas:** Sekitar 39,6% dari pemohon, atau sekitar 184.739 orang, memiliki status pinjaman "Fully Paid".



Exploratory Data Analysis

Rasio Loan Status

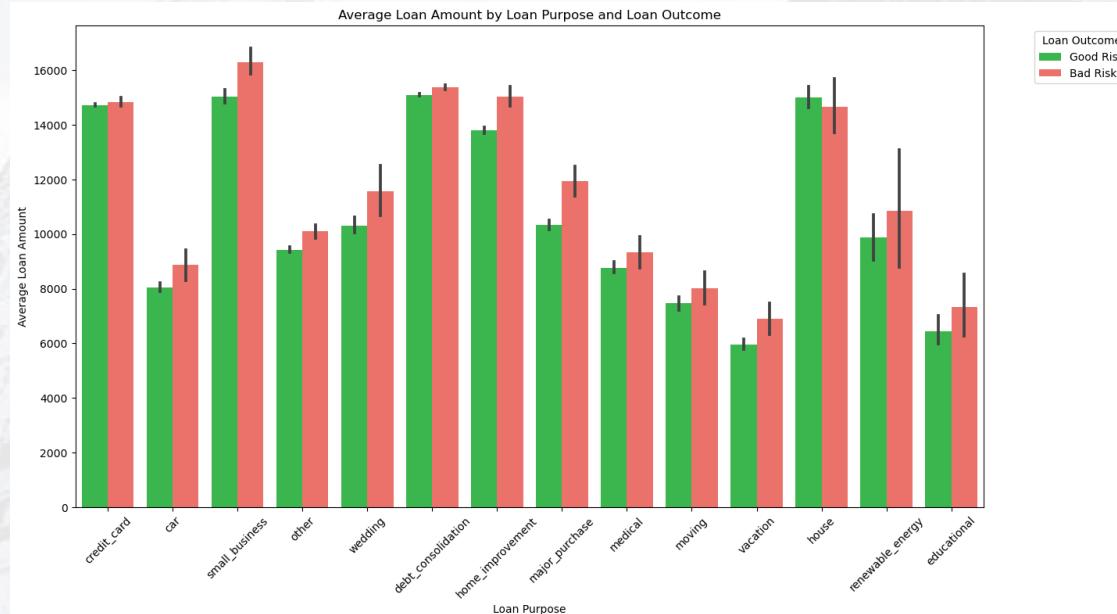
Terdapat Imbalance data target, dengan 'Bad Risk' sebagai minoritas pada 10,93% dibandingkan 'Good Risk' sebesar 89.07%.



Exploratory Data Analysis

Rata-rata Besar Pinjaman vs Tujuan Pinjaman

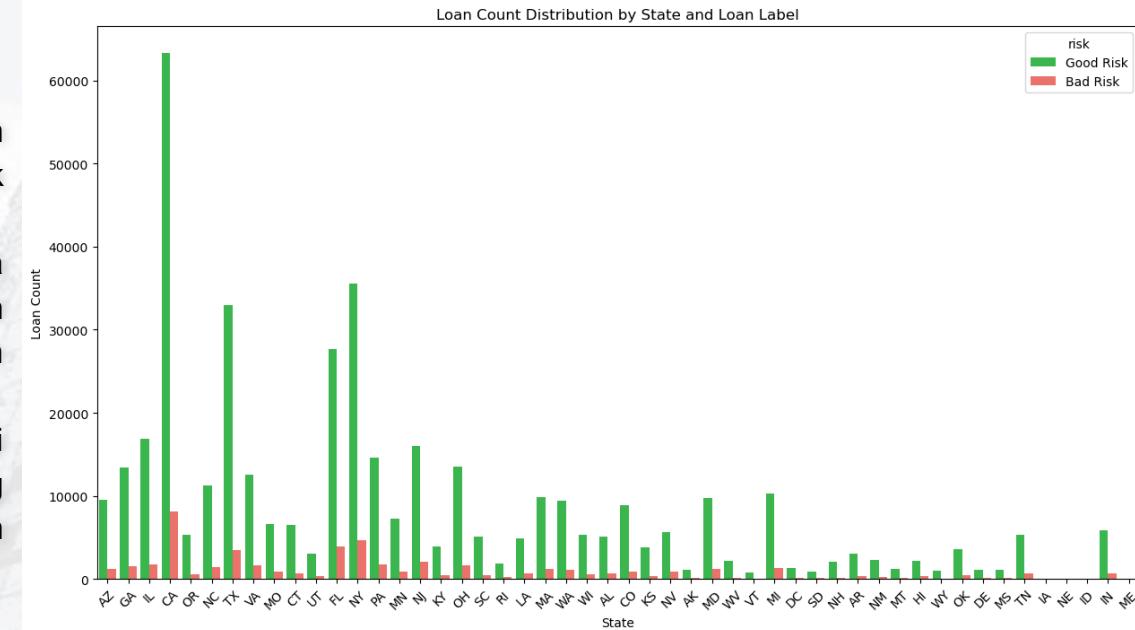
- Usaha Kecil (Small Business):** Menunjukkan rata-rata jumlah pinjaman tertinggi untuk pinjaman Good Risk dan Bad Risk.
- Rumah (House):** Satu-satunya kategori di mana pinjaman Good Risk sedikit lebih tinggi daripada pinjaman Bad Risk.
- Perbaikan Rumah & Pembelian Besar (Home Improvement & Major Purchase):** Menampilkan rata-rata jumlah pinjaman tinggi dengan perbedaan kecil antara pinjaman Good Risk dan Bad Risk.
- Pernikahan & Medis (Wedding & Medical):** Umumnya memiliki rata-rata jumlah pinjaman lebih rendah dibandingkan tujuan lainnya.
- Perbedaan antara Pinjaman Good Risk dan Bad Risk:** Pinjaman Bad Risk biasanya lebih tinggi, menunjukkan risiko yang lebih besar.



Exploratory Data Analysis

Jumlah Pinjaman vs Region

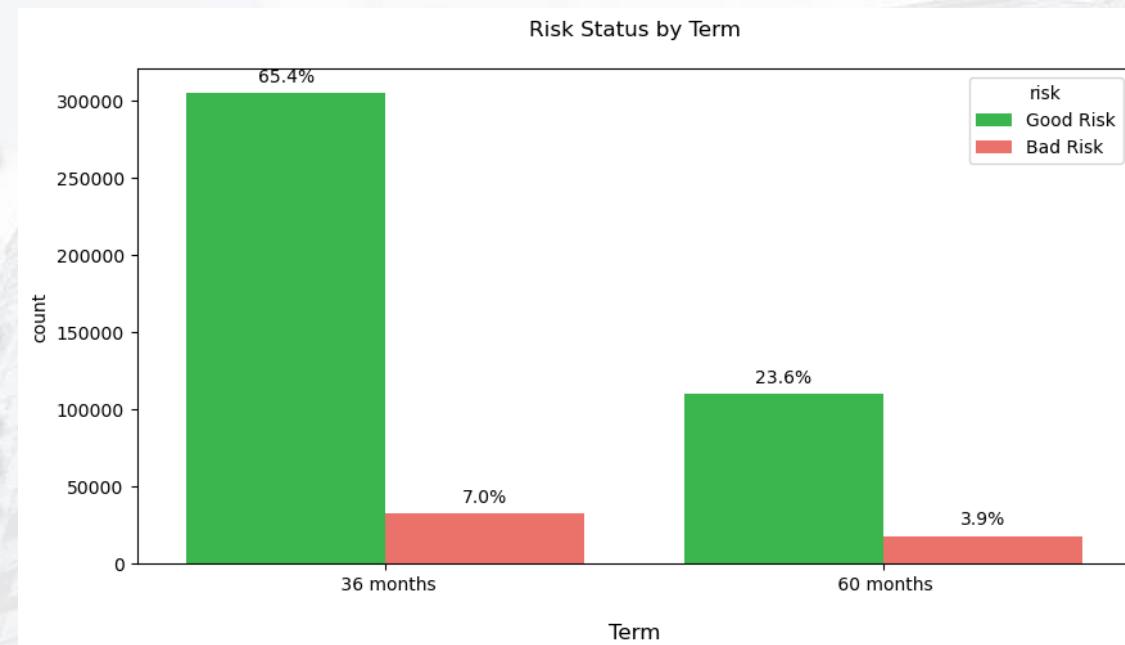
- **California (CA):** Menunjukkan jumlah pinjaman tertinggi, baik pinjaman baik maupun buruk.
- **New York (NY) & Texas (TX):** Kedua negara bagian ini menunjukkan jumlah pinjaman yang tinggi, mengindikasikan pasar pinjaman yang aktif.
- **Florida (FL) & Illinois (IL):** Negara bagian ini juga memiliki jumlah pinjaman yang signifikan, meskipun angkanya lebih rendah dibandingkan CA, NY, dan TX.



Exploratory Data Analysis

Jumlah Pinjaman vs Jangka Waktu Pinjaman

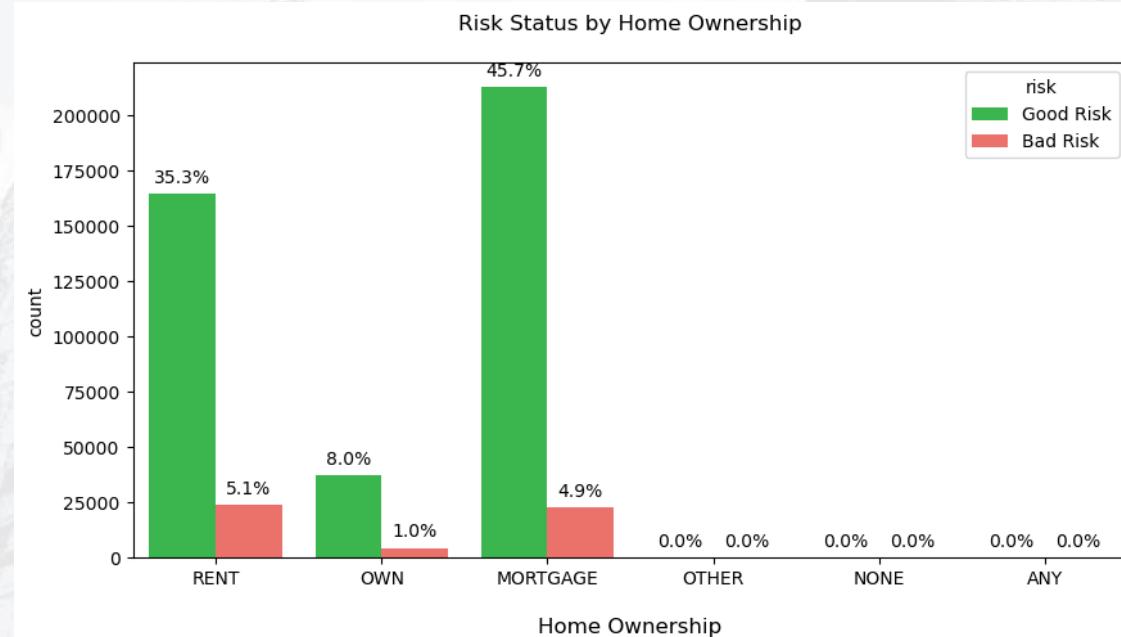
- Jangka waktu pinjaman memberi informasi tentang jumlah pembayaran pinjaman.
- **Hanya ada dua jenis** jangka waktu pinjaman: 36 bulan atau 60 bulan.
- **Mayoritas pinjaman (72,5%)** memiliki jangka waktu yang lebih pendek yaitu **36 bulan**.
- Pinjaman dengan jangka waktu **36 bulan** dua kali lebih mungkin mengalami risiko gagal bayar kredit dibandingkan dengan pinjaman dengan jangka waktu 60 bulan.



Exploratory Data Analysis

Jumlah Pinjaman vs Kepemilikan Rumah

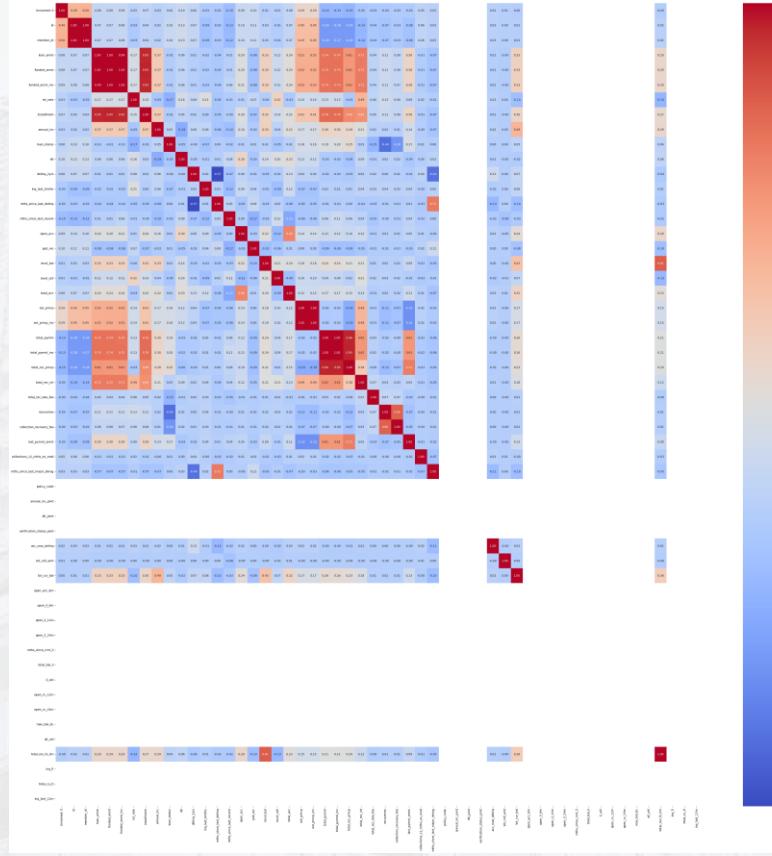
- Kepemilikan Rumah adalah kategori yang disediakan oleh pemohon saat pendaftaran.
- Sebagian besar pemohon memiliki *Mortgage* yang sedang berjalan (51,6%) atau sedang menyewa (40,4%).
- Pemohon dengan **mortgage** atau yang **sedang menyewa** memiliki kemungkinan **risiko gagal bayar kredit yang lebih tinggi**.



Exploratory Data Analysis

Multivariate Analysis: Correlation Heatmap

- Dengan banyaknya jumlah feature (75), tidak bisa ditampilkan secara detail korelasi antar masing-masing feature.
- Terlihat terdapat beberapa yang memiliki warna yang menandakan adanya korelasi yang tinggi.
- Untuk **menghindari redundant feature, maka korelasi di atas 0.7 akan dihapus** salah satu.



Exploratory Data Analysis

Jumlah Missing Value pada Feature

- Terdapat lebih dari 40 Feature yang memiliki value nihil / null.
- Feature dengan >50% Missing value akan didrop, dan sisanya akan diterapkan metode imputasi baik mode, median, atau interpolasi.



Data Preparation

Handling Missing Values

- Semua feature dengan missing values **di atas 50%, akan didrop.**
- Imputasi diterapkan untuk missing values pada feature dengan 50% value.

Redundant Data

- Semua data yang memiliki korelasi tinggi **di atas 0.7** akan didrop salah satunya **untuk menghindari bias** pada model akibat redundansi.

Handling Outlier

- Kalkulasi **Varian (var)** dan **Skewness** masing-masing feature. **Feature yang kurang sesuai akan diterapkan IQR clipping.**
- Sebab hampir >50% data hilang akibat IQR Clipping, sehingga disesuaikan feature tertentu saja, yang menyebabkan tidak ada data yang didrop. (jumlah tetap 466285).

Feature Engineering

Handling Datetime Feature

- 4 feature datetime yang masih categoric telah diubah, earliest_cr_line, last_credit_pull_d, last_pymnt_d, next_pymnt_d.
- **Datetime feature diubah menjadi data numerik “jarak dalam hari” sampai tanggal terakhir pada data (Mar-2016).**

Feature Encode

- Untuk feature kategorikal yang tidak memiliki hierarki atau Boolean feature, akan diterapkan **onehot encoding**.

Split Data

- Data Train dan Test diterapkan split 0.2 random_state 42. **Split diterapkan sebelum data scaling.**
- Train 332250 Data. Test 40778 Data.

Feature Transformation

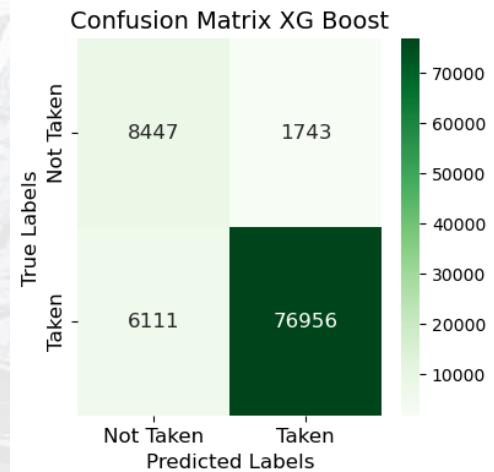
- Transformasi menggunakan **Standar scaler dan Min Max Scaler**. Min Max Scaler diterapkan untuk data-data yang memiliki standar deviasi yang relatif tinggi dibandingkan mean nya. Sisanya diterapkan standard scaler.

Handle Class Imbalance

- Oversampling menggunakan **SMOTE**.

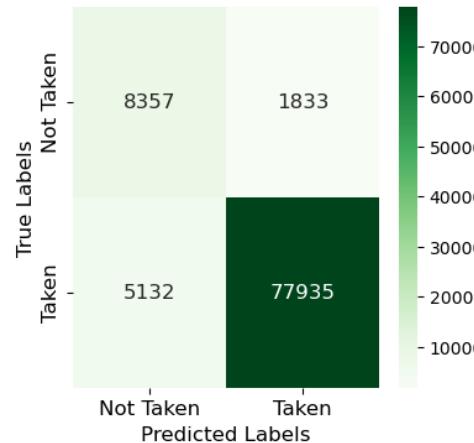
Data Modeling

No	Model	Accuracy	Precision	Recall	F1 Score	AUC (Test)	AUC (Train)
1	Logistic Regression	0.8727	0.9576	0.8968	0.9262	0.8647	0.9327
2	Random Forest	0.8408	0.9711	0.8464	0.9045	0.9055	1.0000
3	Decision Tree	0.3210	0.9743	0.2442	0.3905	0.5958	1.0000
4	AdaBoost	0.8366	0.9686	0.8440	0.9020	0.9147	0.9932
5	XGBoost	0.9158	0.9779	0.9264	0.9514	0.9550	0.9976



Data Modeling

Confusion Matrix XG Boost with Best Parameters



Hyperparameter Tuning

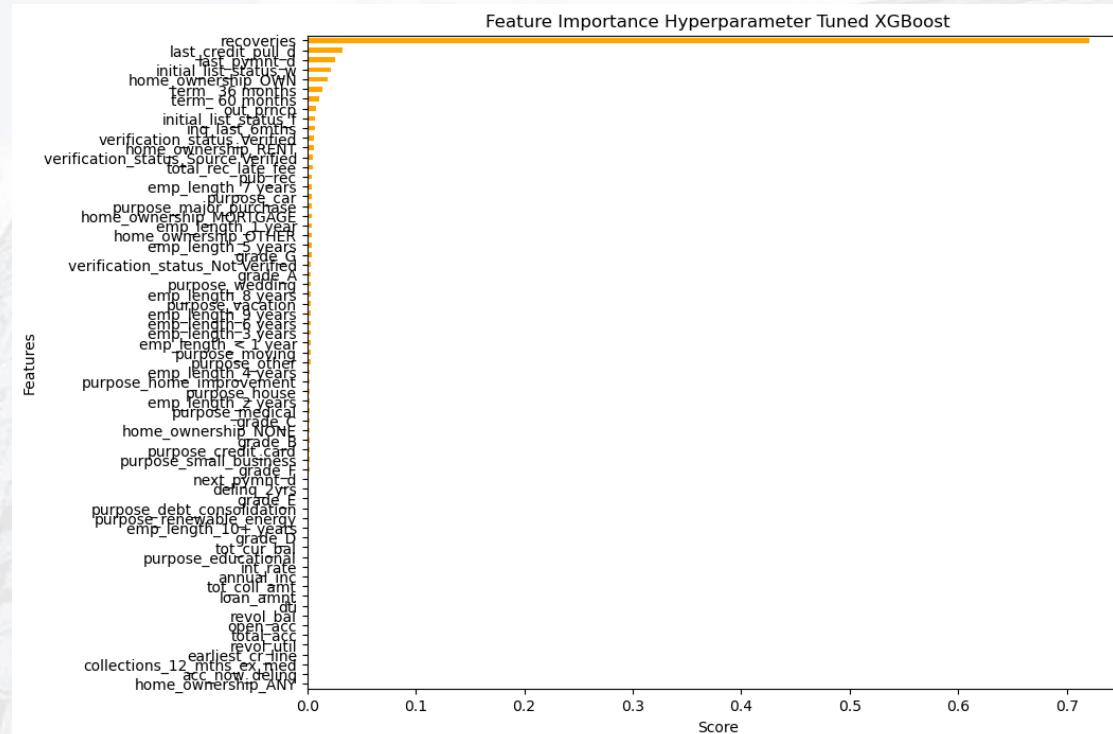
- Menggunakan RandomSearch Cross Validation untuk model terbaik, XGBoost.
- Didapat parameter {'subsample': 0.6, 'n_estimators': 300, 'max_depth': 15, 'learning_rate': 0.05, 'colsample_bytree': 1.0}

No	Model	Accuracy	Precision	Recall	F1 Score	AUC (Test)	AUC (Train)
1	Logistic Regression	0.8727	0.9576	0.8968	0.9262	0.8647	0.9327
2	Random Forest	0.8408	0.9711	0.8464	0.9045	0.9055	1.0000
3	Decision Tree	0.3210	0.9743	0.2442	0.3905	0.5958	1.0000
4	AdaBoost	0.8366	0.9686	0.8440	0.9020	0.9147	0.9932
5	XGBoost	0.9158	0.9779	0.9264	0.9514	0.9550	0.9976
6	XGBoost Tuned	0.93	0.98	0.94	0.96	0.96	1.00

Evaluation

Feature Importance

1. Recoveries: feature yang menunjukkan apakah rencana pembayaran telah diterapkan untuk pinjaman
2. Last_credit_pull_d: tanggal menarik credit untuk pinjaman terakhir.
3. Last_pymnt_d: tanggal pembayaran terakhir
4. Initial_list_status_w: Status pencatatan pinjaman utuh (whole) atau tidak
5. Home_ownership_OWN: status kepemilikan rumah, memiliki (own) atau tidak.



Conclusion

Model Terbaik

XGBoost. Dengan hyperparameter tuning, model ini menunjukkan performa terbaik dalam hal akurasi. Menjawab goal untuk meningkatkan akurasi pemilihan risiko peminjam, **dari 89.07% meningkat menjadi 93%** tingkat keakurasiannya menentukan risiko.

Dengan menggunakan model, bisa ditargetkan hanya mereka yang memiliki good risk saja yang diberikan izin untuk mendapatkan pinjaman. Hal ini memiliki tingkat keakurasiannya 93% dan **memiliki cost lebih rendah akibat tidak semua nasabah perlu dikonsiderasi atau ditawarkan pinjaman.**

Conclusion

Insights Bisnis dari Feature Importance

1. Rencana Pembayaran Pinjaman (Recoveries)

1. **Insight Bisnis:** Pinjaman yang memerlukan rencana pembayaran memiliki risiko gagal bayar yang lebih tinggi.
2. **Rekomendasi:** Terapkan strategi intervensi awal untuk pinjaman yang masuk dalam rencana pembayaran untuk mengurangi tingkat gagal bayar.

2. Pemantauan Kredit (Last_credit_pull_d)

1. **Insight Bisnis:** Pemantauan kredit berkelanjutan membantu mengidentifikasi perubahan dalam profil risiko peminjam.
2. **Rekomendasi:** Tingkatkan frekuensi penarikan kredit untuk peminjam berisiko tinggi guna memastikan penilaian risiko yang terkini.

3. Kepatuhan Pembayaran (Last_pymnt_d)

1. **Insight Bisnis:** Pembayaran tepat waktu adalah indikator penting dari kesehatan keuangan peminjam.
2. **Rekomendasi:** Buat sistem peringatan otomatis untuk pembayaran terlambat dan inisiasi kontak dengan peminjam untuk menawarkan dukungan atau opsi restrukturisasi.

Conclusion

4. Status Pencatatan Pinjaman (Initial_list_status_w)

1. **Insight Bisnis:** Status pencatatan awal dapat memengaruhi kinerja pinjaman.
2. **Rekomendasi:** Bandingkan kinerja pinjaman utuh dan fraksional untuk mengidentifikasi tren dan menyesuaikan strategi pemberian pinjaman sesuai dengan itu.

5. Kepemilikan Rumah (Home_ownership_NOW)

1. **Insight Bisnis:** Kepemilikan rumah adalah indikator kuat dari stabilitas keuangan.
2. **Rekomendasi:** Berikan pertimbangan yang lebih baik kepada pemilik rumah dalam model risiko kredit, tetapi juga perhitungkan volatilitas potensial dari penyewa dan yang memiliki hipotek.

Thank You



Rakamin
Academy



id/x partners

Project GitHub [here!](#)

Project explanation video [here!](#)