Rakamin Academy X id/x partners

Credit Risk Prediction

ID/X Partners - Data Scientist

Presented by
Rifqi Mufiddin

<Rifqi Mufiddin>

I am a fresh graduate of Informatics Engineering with a strong passion for Data Science. I refined my skills through bootcamps and internships, culminating in certification as an Associate Data Scientist from BNSP. Eager to apply my knowledge in real-world scenarios, I am poised to contribute innovative data solutions and insights that advance corporate goals.

<Surabaya>

<rifqimufiddin07@gmail.com>

<Github Profile>

<Linkedin Profile>

About Company

ID/X Partners (PT IDX Consulting) adalah perusahaan konsultan yang didirikan pada tahun 2002. Mereka telah melayani perusahaan di Asia, Australia, dan berbagai industri, termasuk layanan keuangan, telekomunikasi, manufaktur, dan ritel. Spesialisasi utama **ID/X Partners** adalah dalam memanfaatkan solusi data analytics and decisioning (DAD), yang digabungkan dengan manajemen risiko dan disiplin pemasaran terintegrasi. Mereka membantu klien untuk mengoptimalkan profitabilitas portofolio dan proses bisnis mereka.

ID/X Partners menawarkan layanan konsultasi yang komprehensif dan solusi teknologi, menjadikannya sebagai penyedia layanan lengkap dalam industri tersebut. Dengan fokus pada inovasi, analisis data, dan pengambilan keputusan, mereka membantu klien mereka untuk mencapai tujuan bisnis mereka secara efisien dan efektif.

Project Portfolio

Latar Belakang: Perusahaan pemberi pinjaman (multifinance) bekerja sama dengan ID/X Partners ingin meningkatkan keakuratan dalam menilai dan mengelola risiko kredit. Hal ini diharapkan dapat mengoptimalkan keputusan bisnis mereka dan mengurangi potensi kerugian.

Data Yang Tersedia: Dataset yang disediakan mencakup data pinjaman yang disetujui dan ditolak, termasuk atribut seperti profil peminjam, riwayat pembayaran, dan performa kredit.

Problem Statement: Tujuan proyek ini adalah mengembangkan model machine learning yang dapat memprediksi risiko kredit (credit risk) berdasarkan dataset yang tersedia. Model ini diharapkan dapat memberikan rekomendasi yang akurat dalam menilai kelayakan peminjam dan mengurangi kemungkinan terjadinya kerugian bagi perusahaan pemberi pinjaman.

Link code here!

1. Data Understanding

Features	Description
id	Unique identification for each loan.
funded_amnt	The amount of loan requested by the customer.
term	The loan in months.
int_rate	The annual interest rate given on the loan.
installment	The monthly payment amount that must be paid by the customer.
loan_status	The current status of the loan (Target).
etc	Other features.

- Data terdiri dari 466285 baris dan 75 kolom
- Terdapat 52 kolom numerik dan 22 kolom kategorikal
- Terdapat 40 kolom numerik yang memiliki missing values
- Tidak terdapat duplikat baris
- Beberapa kolom memiliki missing values 100% (akan di drop)
- Dataset yang digunakan: Loan Data(Data Dictionary)

1. Data Understanding

1. **Loan Amount Distribution Based on Loan Amount (loan_amnt):** A 10,000 loan is the most common, followed by loans of 12,000 and 15,000. This indicates that lower loan amounts tend to be more popular among applicants.
2. **Interest Rate (int_rate):** The most common interest rate is around **12.99%**, followed by **10.99%** and **15.61%**. This provides an overview of the range of interest rates typically offered to applicants.
3. **Annual Income Distribution (annual_inc):** An annual income of **60,000** is the most common, followed by **50,000** and **65,000**. This shows the common income profile of applicants.
4. **Debt-to-Income Ratio (dti):** The most common values for the debt-to-income ratio are around **14.40**, **19.20**, and **12.00**. This can provide information on how much of the applicant's income is allocated to debt payments.
5. **Number of Late Payments in the Last 2 Years (delinq_2yrs):** The majority of applicants (around **382,954**) have no late payments in the last 2 years, but some have up to **29** late payments.
6. **Number of Public Records (pub_rec):** The majority of applicants (around **404,893**) have no public records, but some have up to **40** public records.
7. **Total Available Credit (total_rev_hi_lim):** The most common highest credit limit is **15,000**, **13,500**, and **10,000**. This can provide an insight into the amount of credit available to applicants.
8. **Number of Loans Delinquent in the Last 12 Months (collections_12_mths_ex_med):** The majority of applicants (around **462,226**) have no loans delinquent in the last 12 months.

1. Data Understanding

1. **Loan Term (term):**
    - The majority of loans have a term of **36 months**, with a small portion having a term of **60 months**. This indicates the preference of the majority of borrowers for shorter loan terms.
2. **Loan Grade and Subgrade:**
    - The majority of loans have a grade of **B** and **C**, with subgrades **B3** and **B4** being the most common. This may reflect a diverse risk profile among loan applicants.
3. **Borrower's Occupation (emp_title) and Employment Length (emp_length):**
    - The majority of borrowers are **teachers**, **managers**, or **registered nurses**, with most having **more than 10 years** of work experience. This indicates that the majority of applicants are workers with solid work experience.
4. **Home Ownership Status (home_ownership):**
    - The majority of applicants own homes with a **mortgage**, followed by those who **rent**. Only a small number own homes **outright** without a mortgage.
5. **Verification Status (verification_status):**
    - Most applicants have a **verified** or **source verified** verification status. This indicates that the majority of applicants have passed the identity and income verification process.
6. **Loan Purpose (purpose):**
    - The majority of loans are used for **debt consolidation** or **credit card payment**, followed by **home improvement** and other purposes such as **major purchases**.
7. **Loan Status (loan_status):**
    - The majority of loans are still **active (current)** or have been **fully paid**, with a small portion being **charged off** or having **late payments**.

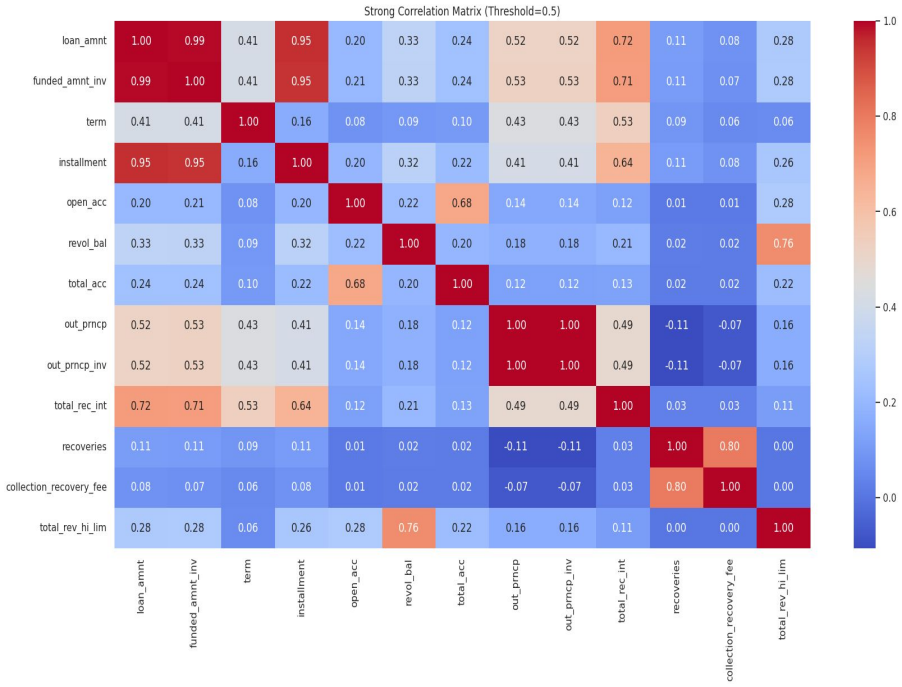
2. Feature Engineering

Create Target

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

3. Exploratory Data Analysis

What are the strong correlations between different features in the dataset?



- **loan_amnt** dan **funded_amnt_inv** (0.99) → jumlah pinjaman yang diminta hampir selalu sama dengan jumlah yang didanai.
- **loan_amnt** dan **installment** (0.95) → jumlah pinjaman yang lebih tinggi umumnya berhubungan dengan cicilan yang lebih tinggi.
- **out_prncp** dan **out_prncp_inv** memiliki korelasi sempurna (1.00) → sisa pokok pinjaman dan sisa pokok pinjaman yang diinvestasikan adalah sama.

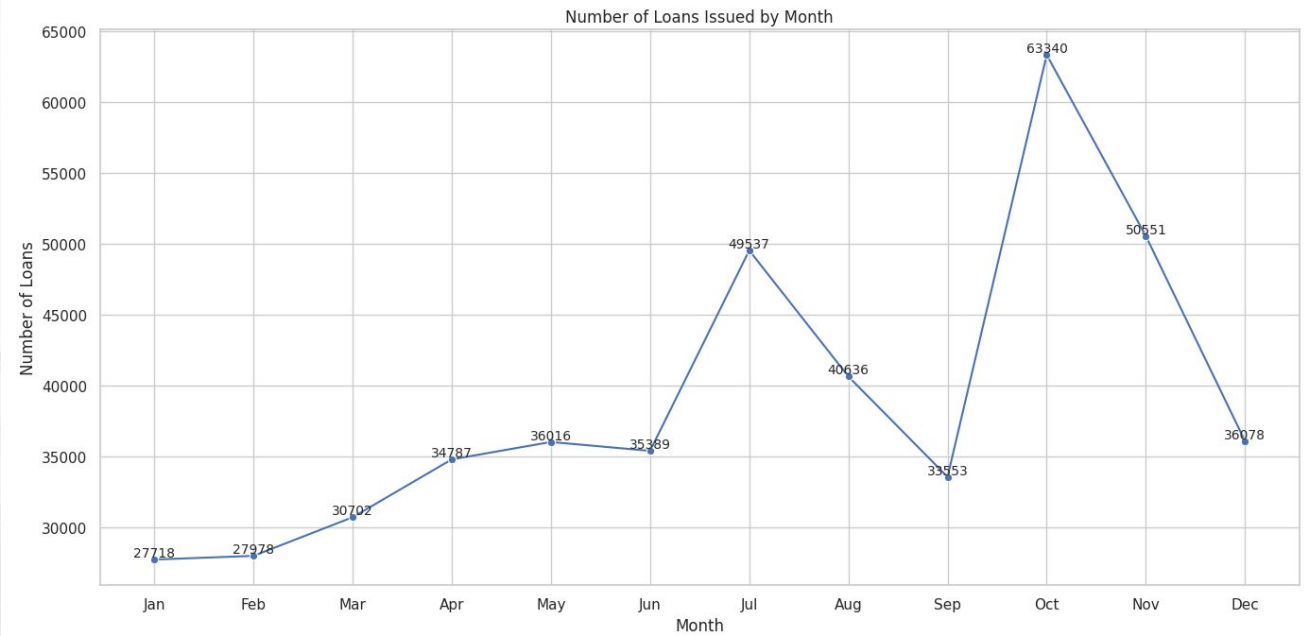
3. Exploratory Data Analysis

What are the strong correlations between different features in the dataset?

- **total_rec_int** dengan **loan_amnt** (0.72), **funded_amnt_inv** (0.71), dan **installment** (0.64) → total jumlah bunga yang diterima berhubungan erat dengan besarnya pinjaman dan cicilan.
- **revol_bal** dan **total_rev_hi_lim** (0.76) → saldo kredit bergulir berhubungan erat dengan batas kredit total yang tinggi.
- **out_prncp** dan **total_rec_int** (0.49) → sisa pokok pinjaman masih memiliki hubungan dengan total bunga yang diterima.
- **open_acc** dan **total_acc** (0.68) → jumlah akun kredit yang terbuka berkaitan erat dengan jumlah akun kredit total.
- **recoveries** dan **collection_recovery_fee** (0.80) → biaya pemulihan koleksi berhubungan erat dengan jumlah pemulihan yang dilakukan.

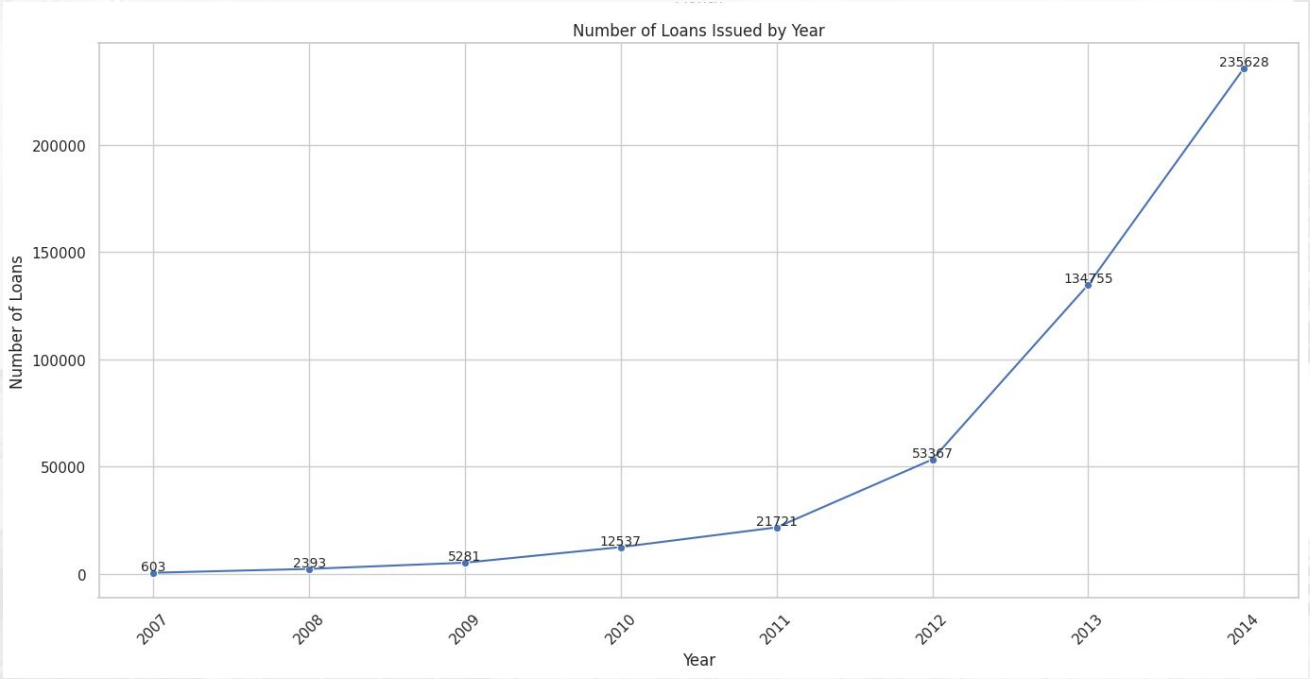
3. Exploratory Data Analysis

What Are the Monthly and Yearly Trends in Loan Issuance?



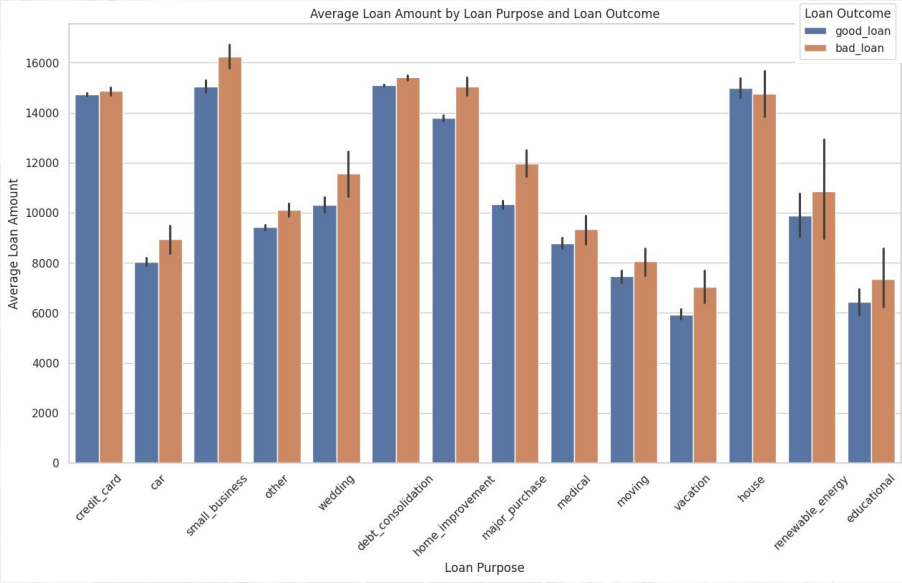
3. Exploratory Data Analysis

What Are the Monthly and Yearly Trends in Loan Issuance?



3. Exploratory Data Analysis

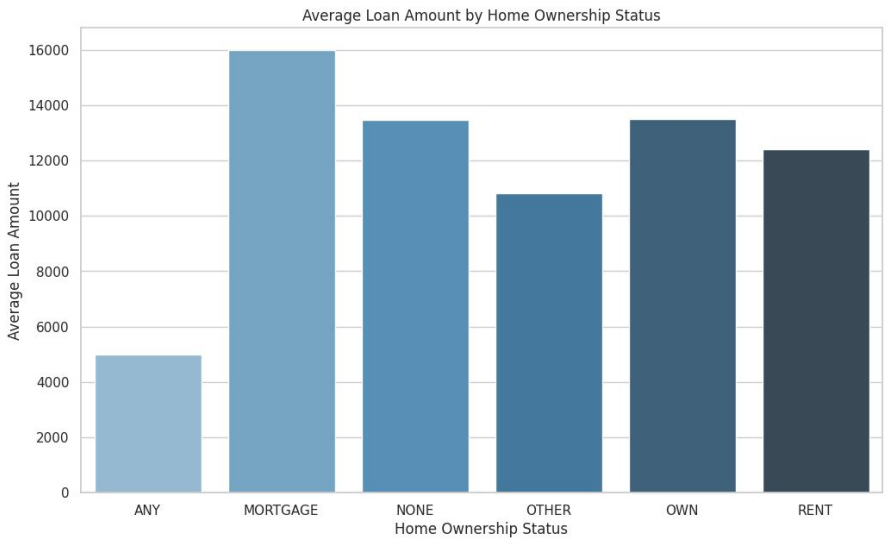
Is there a difference in the average loan amount between good loans and bad loans based on loan purposes?



- **Small Business**: Rata-rata pinjaman tertinggi untuk pinjaman baik dan buruk.
- **House**: Satu-satunya pinjaman baik yang sedikit lebih tinggi dari pinjaman buruk.
- **Home Improvement & Major Purchase**: Rata-rata pinjaman tinggi dengan perbedaan kecil antara pinjaman baik dan buruk.
- **Wedding & Medical**: Rata-rata pinjaman lebih rendah dibanding tujuan lain.
- **Perbedaan Pinjaman Baik dan Buruk**: Pinjaman buruk umumnya lebih tinggi, menunjukkan risiko lebih tinggi.

3. Exploratory Data Analysis

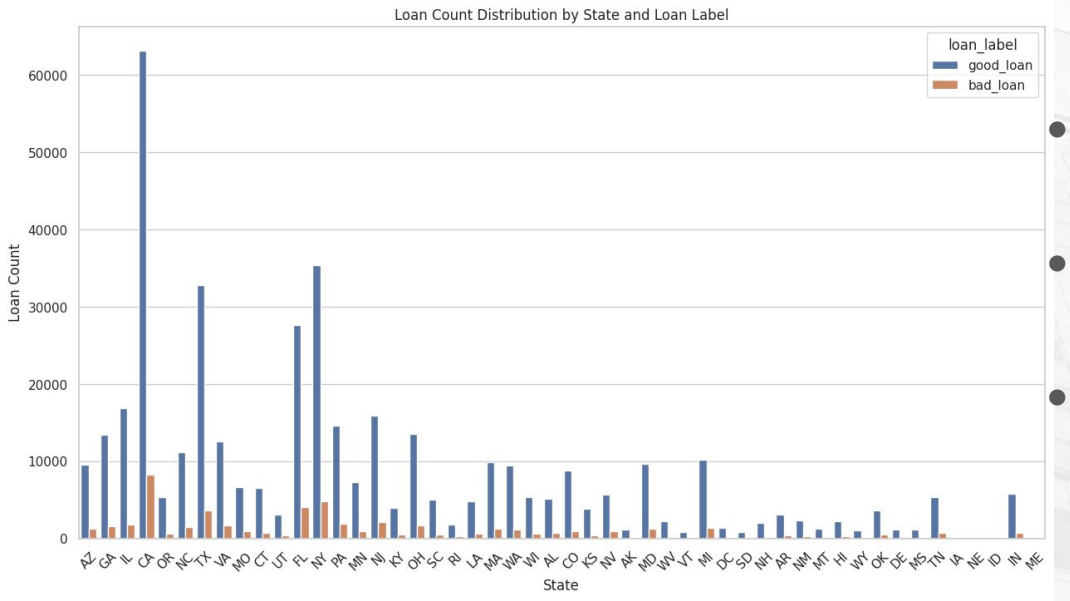
Does home ownership status affect the approved loan amount?



- **Mortgage**: Rata-rata pinjaman tertinggi.
- **Own & Rent**: Rata-rata pinjaman cukup tinggi, sedikit di bawah mortgage.
- **None**: Rata-rata pinjaman masih tinggi meski tanpa kepemilikan rumah.
- **Any**: Rata-rata pinjaman terendah.
- **Status Kepemilikan Rumah**: Hipotek menunjukkan pinjaman jumlah besar. Perbedaan jumlah pinjaman ini penting untuk penilaian kredit dan risiko.

3. Exploratory Data Analysis

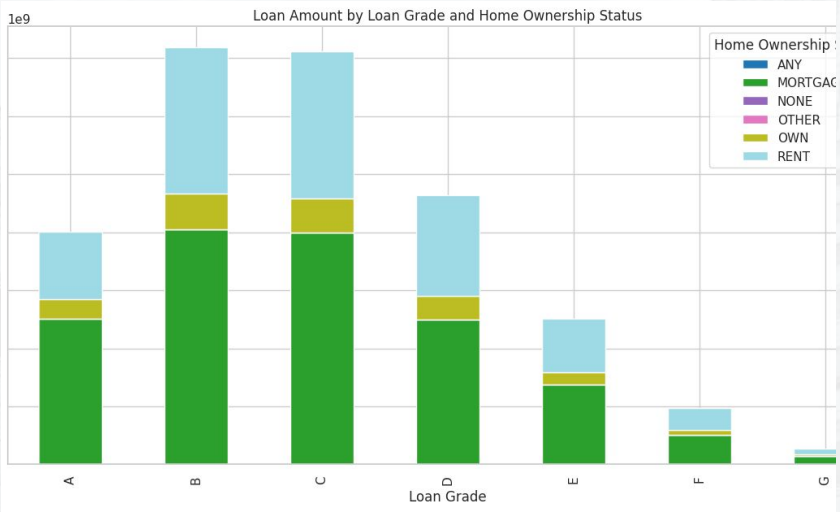
Is there a significant difference in the number of loans issued based on the region (addr_state) and loan label (loan_label)?



- **California (CA)**: Jumlah pinjaman tertinggi untuk pinjaman baik dan buruk.
- **New York (NY) & Texas (TX)**: Jumlah pinjaman tinggi, menunjukkan pasar pinjaman aktif.
- **Florida (FL) & Illinois (IL)**: Jumlah pinjaman signifikan, namun lebih rendah dari CA, NY, dan TX.

3. Exploratory Data Analysis

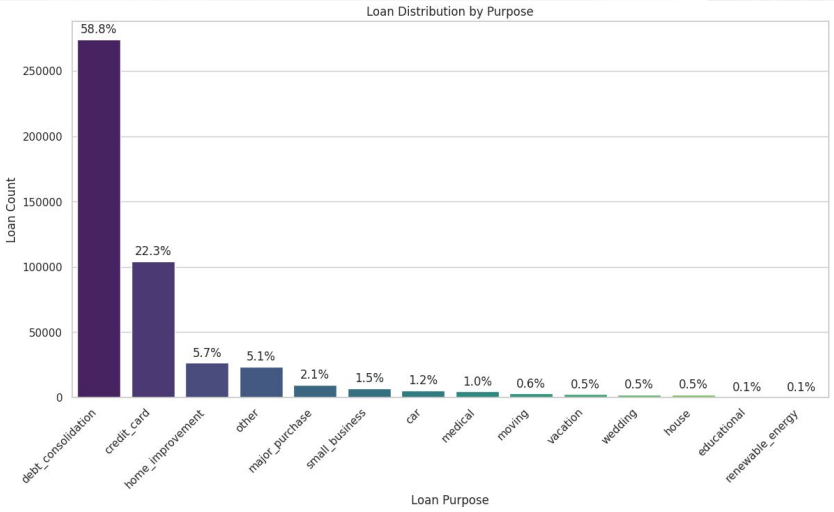
How is the loan amount distributed by loan grade and home ownership status?



- **Grade B dan C**: Jumlah pinjaman tertinggi, menandakan risiko menengah paling umum.
- **Grade A**: Jumlah pinjaman signifikan, tetapi lebih rendah dari Grade B dan C.
- **Grade D hingga G**: Jumlah pinjaman lebih rendah, mencerminkan risiko lebih tinggi dan penawaran lebih ketat.
- **Mortgage**: Jumlah pinjaman tertinggi di semua kelas, terutama pada Grade B dan C.
- **Rent dan Own**: Jumlah pinjaman signifikan tetapi lebih rendah dibanding peminjam dengan hipotek.
- **None, Other, dan Any**: Jumlah pinjaman lebih sedikit, menunjukkan status kepemilikan rumah ini kurang diprioritaskan.

3. Exploratory Data Analysis

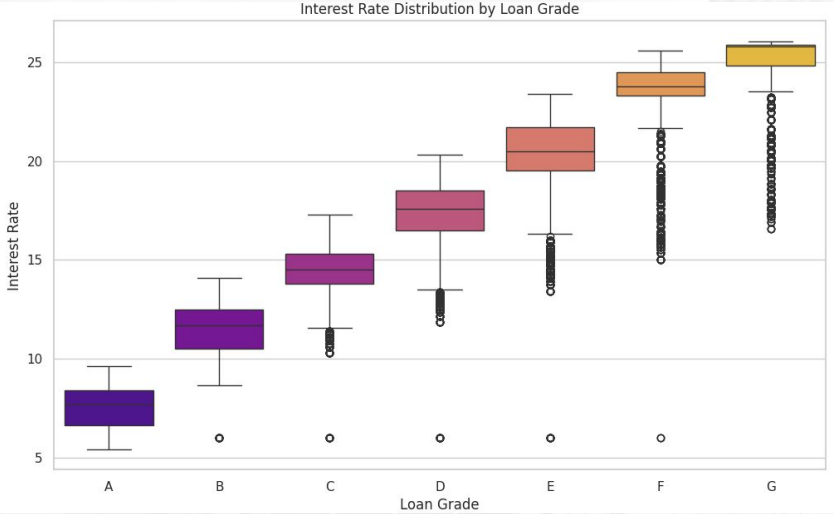
What is the most common purpose for loan applications?



- **Tujuan Populer**: Konsolidasi utang (58.8%) dan kartu kredit (22.3%) paling umum.
- **Tujuan Lain**: Home improvement, major purchase, dan small business memiliki porsi lebih kecil.
- **Variasi Tujuan**: Pendidikan dan energi terbarukan sangat jarang diajukan.

3. Exploratory Data Analysis

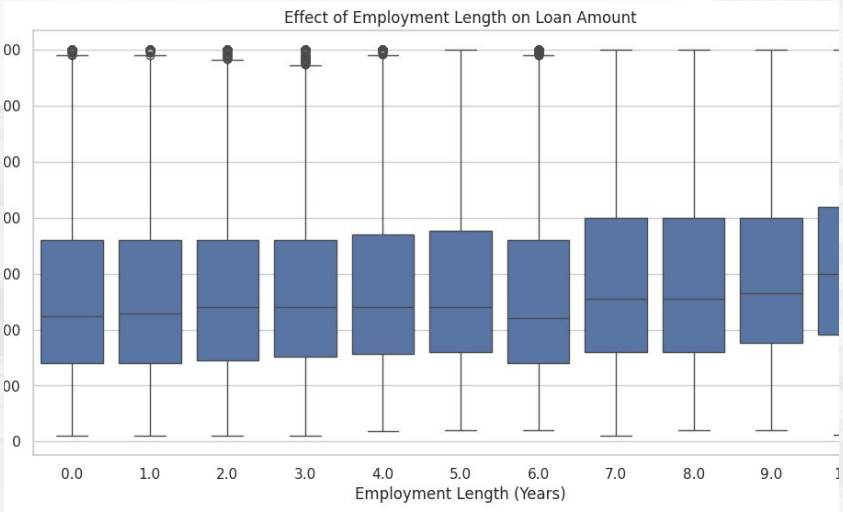
What is the interest rate distribution by loan grade?



- **Hubungan Suku Bunga dan Kelas**: Kelas pinjaman lebih rendah (A ke G) memiliki suku bunga lebih tinggi.
- **Rentang Suku Bunga**: Rentang lebih luas pada kelas pinjaman lebih rendah.
- **Risiko dan Biaya**: Peminjam risiko tinggi dikenakan suku bunga lebih tinggi.

3. Exploratory Data Analysis

Does employment length (emp_length) affect the approved loan amount?



- **Distribusi Konsisten**: Median pinjaman berkisar antara 10.000 hingga 15.000 USD di seluruh lama bekerja, tanpa perbedaan signifikan.
- **Rentang Variasi Luas**: Banyak nilai pencilan menunjukkan faktor lain mempengaruhi jumlah pinjaman.
- **Tidak Ada Tren Jelas**: Lama bekerja tidak menentukan jumlah pinjaman.

4. Data Preparation

Handle Missing Values	Penghapusan Kolom TIdak Relevan	Konversi Data
- Drop kolom 100% hilang - Imputasi nilai (median dan modus)	Menghapus kolom tidak informatif atau berpotensi bocor.	-'emp_length' dan 'term' -Kolom tanggal ke datetime.

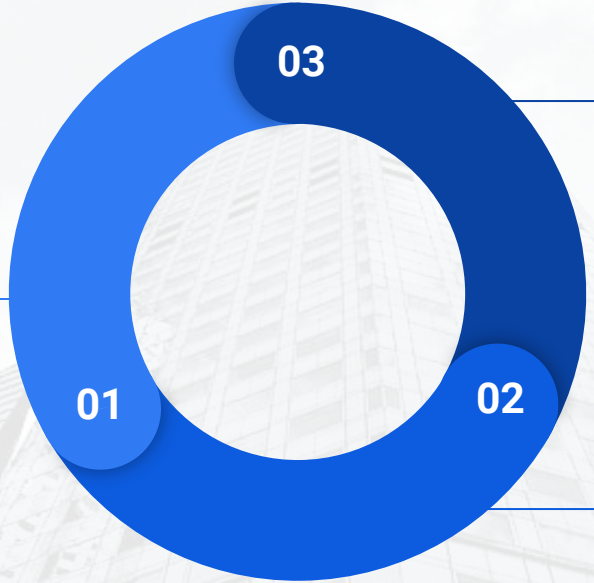
Penanganan Outliers	Encoding Variabel Kategorikal	Penskalaan Data
IQR Clipping pada beberapa kolom.	One-Hot dan Label Encoding.	- StandardScaler dan MinMaxScaler.

5. Data Modeling

Split Data	Pemodelan	Resampling
Stratified train-test split (80-20)	• Logistic Regression • Decision Tree • Random Forest	• Manual Oversampling • SMOTE

Hyperparameter Tuning	Dimensionality Reduction	Other Resampling
• RandomizedSearch CV • Cross-validation (recall)	• PCA	• ADASYN • NearMiss

6. Evaluation



Classification Report

ROC AUC Score

Confusion Matrix

6. Evaluation(No Sampling)

Diutamakan FN & Recall “0” untuk meminimalkan kesalahan prediksi “bad loan”

Model	Recall	ROC AUC	FN	Accuracy
Logistic Regresision	55%	91%	4646	94%
Decision Tree	79%	87%	2201	95%
Random Forest	72%	95%	2902	97%

6. Evaluation(Manual Sampling)

Diutamakan FN & Recall “0” untuk meminimalkan kesalahan prediksi “bad loan”

Model	Recall	ROC AUC	FN	Accuracy
Logistic Regresision	80%	92%	2080	89%
Decision Tree	77%	87%	2353	95%
Random Forest	75%	95%	2614	97%

```
def manual_resample(X_train, y_train, random_state=42):
    X_train_combined = pd.concat([X_train, y_train], axis=1)
    majority_class = X_train_combined[X_train_combined['loan_label'] == 1]
    minority_class = X_train_combined[X_train_combined['loan_label'] == 0]
    minority_oversampled = resample(minority_class, replace=True, n_samples=len(majority_class), random_state=random_state)
    X_train_resampled = pd.concat([majority_class, minority_oversampled])
    y_train_resampled = X_train_resampled['loan_label']
    X_train_resampled = X_train_resampled.drop(columns=['loan_label'])
    return X_train_resampled, y_train_resampled
```

6. Evaluation(SMOTE)

Diutamakan FN & Recall “0” untuk meminimalkan kesalahan prediksi “bad loan”

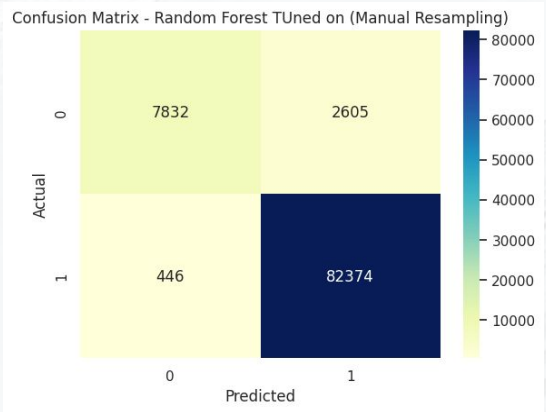
Model	Recall	ROC AUC	FN	Accuracy
Logistic Regresision	79%	91%	2188	88%
Decision Tree	77%	79%	2201	82%
Random Forest	75%	94%	2632	96%

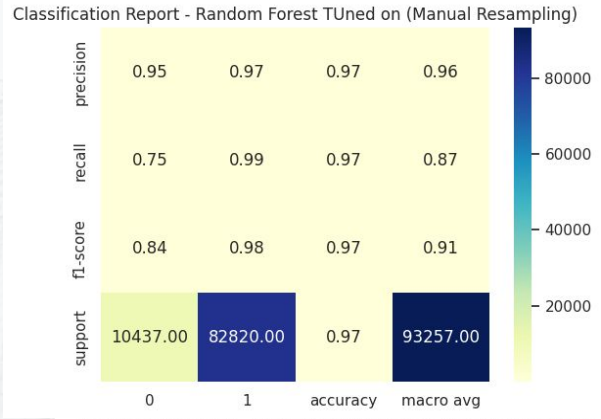
6. Evaluation(Tuning Random Forest)

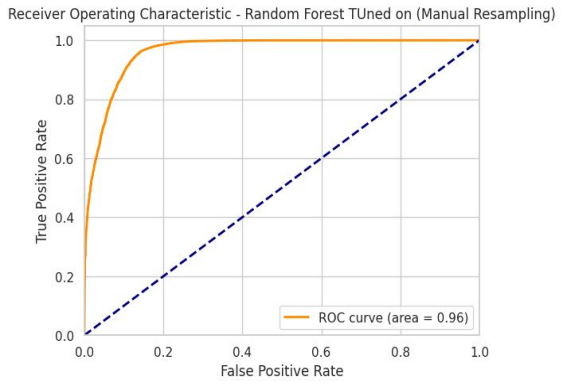
Tuning dilakukan hanya untuk model RF dikarenakan model RF memiliki performa terbaik. Menggunakan **'RandomizedSearchCV'** dan didapatkan paramater terbaik yaitu **'n_estimators'**=300, **'max_depth'**=30, **'max_features'**=sqrt dengan rata-rata skor **0.994**

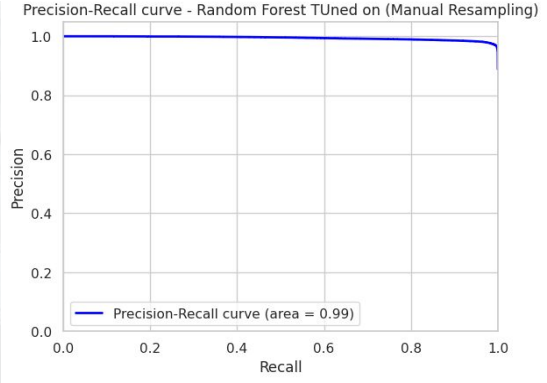
Model	Recall	ROC AUC	FN	Accuracy
No Sampling	72%	95%	2955	97%
Manual Sampling	75%	96%	2605	97%
SMOTE	74%	95%	2694	96%
ADASYN	74%	95%	2750	96%
NearMiss	90%	81%	1004	44%
PCA	75%	93%	2611	94%

6. Evaluation(Tuning RF - Manual Sampling)



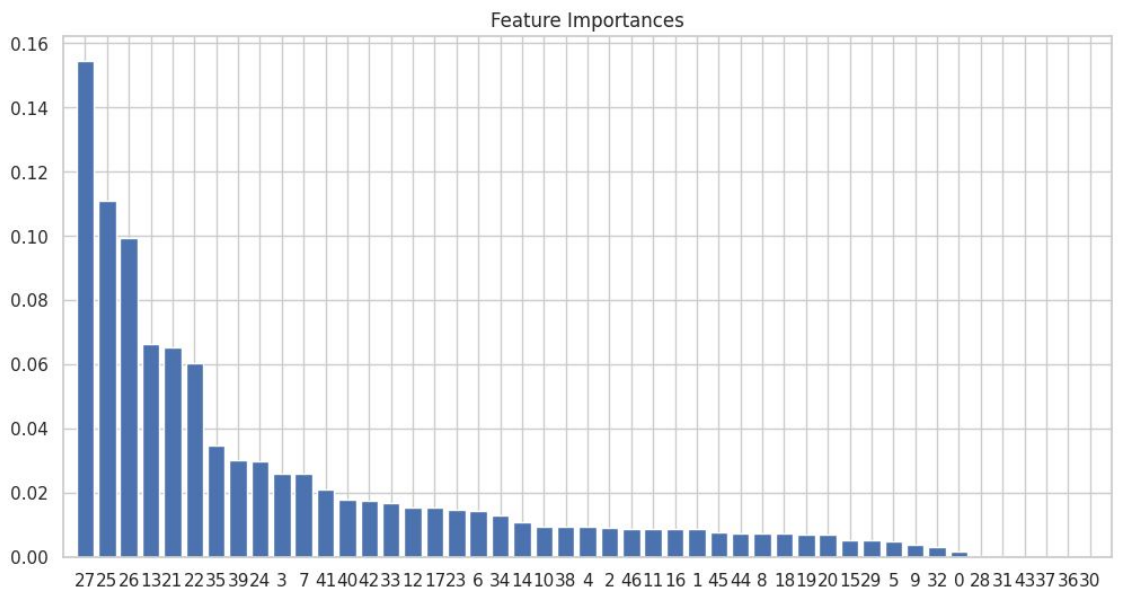






6. Evaluation(Feature Importances)

10 Fitur Penting: ['last_pymnt_amnt', 'recoveries', 'collection_recovery_fee', 'inq_last_6mths', 'out_prncp', 'out_prncp_inv', 'home_ownership_MORTGAGE', 'home_ownership_RENT', 'total_rec_late_fee', 'int_rate']



7. Conclusion(Business Insights)

Karakteristik Debitur Berisiko:

- Debitur dengan riwayat kredit buruk, pendapatan rendah, dan rasio hutang terhadap pendapatan tinggi cenderung memiliki risiko tinggi menunggak pembayaran.
- Faktor-faktor lain seperti pekerjaan, status pernikahan, dan jumlah tanggungan juga signifikan dalam penilaian risiko kredit.
- Loan Amount dan Funded Amount: Jumlah pinjaman yang diminta hampir selalu sama dengan jumlah yang didanai, menunjukkan bahwa hampir semua permintaan pinjaman dipenuhi.
- Installment dan Loan Amount: Jumlah pinjaman yang lebih tinggi berhubungan dengan cicilan yang lebih tinggi, yang memengaruhi kemampuan debitur untuk membayar kembali.
- Out Prncp dan Out Prncp Inv: Sisa pokok pinjaman dan sisa pokok pinjaman yang diinvestasikan sama, menunjukkan konsistensi dalam pelaporan data.

7. Conclusion(Actionable Insights)

- **Pengetatan Kriteria Penilaian Kredit:** Memperketat kriteria berdasarkan faktor risiko yang teridentifikasi seperti riwayat kredit, pendapatan, dan rasio hutang terhadap pendapatan.

- **Peningkatan Monitoring dan Pengawasan:** Monitoring ketat untuk pinjaman yang diberikan kepada debitur berisiko tinggi, termasuk program mitigasi risiko yang lebih agresif.

- **Edukasi dan Konsultasi Keuangan:** Memberikan edukasi mengenai pentingnya riwayat kredit yang baik dan manajemen keuangan, serta menyediakan layanan konsultasi keuangan.

- **Penyesuaian Produk Kredit:** Mengembangkan produk kredit yang lebih fleksibel sesuai dengan kebutuhan debitur berisiko tinggi, serta menyediakan opsi restrukturisasi pinjaman.

7. Conclusion(EDA)

Distribusi Fitur dan Korelasi:

- **Total Rec Int dengan Loan Amount**: Total bunga yang diterima berhubungan erat dengan besarnya pinjaman dan cicilan.
- **Revol Bal dan Total Rev Hi Lim**: Saldo kredit bergulir berhubungan erat dengan batas kredit total yang tinggi.
- **Open Acc dan Total Acc**: Jumlah akun kredit terbuka berkaitan erat dengan jumlah akun kredit total.

Tujuan Pinjaman:

- **Konsolidasi Utang dan Kartu Kredit**: Tujuan pinjaman paling umum, menunjukkan kebutuhan besar untuk pengelolaan hutang.
- **Mortgage dan Status Kepemilikan Rumah**: Pinjaman terbesar diberikan kepada pemilik rumah dengan hipotek, menunjukkan prioritas dan risiko yang lebih rendah.

Kelas Pinjaman dan Suku Bunga:

- **Kelas Pinjaman Lebih Rendah (A ke G)**: Kelas pinjaman yang lebih rendah memiliki suku bunga lebih tinggi, mencerminkan risiko yang lebih besar.

7. Conclusion(Teknis)

Model Terbaik:

- **Random Forest**: Dengan hyperparameter tuning, model ini menunjukkan performa terbaik dalam hal recall dan akurasi, mampu menangani data yang tidak seimbang dengan baik.

Evaluasi Model:

- Model dievaluasi menggunakan berbagai metrik seperti precision, recall, f1-score, dan ROC AUC. Teknik visualisasi seperti confusion matrix, ROC curve, dan precision-recall curve digunakan untuk memahami performa model secara mendalam.

7. Conclusion(Rekomendasi Bisnis)

- **Implementasi Model:** Mengimplementasikan model prediktif dalam sistem penilaian kredit untuk meningkatkan akurasi dan efisiensi.

- **Strategi Mitigasi Risiko:** Mengembangkan strategi mitigasi risiko yang lebih baik berdasarkan karakteristik debitur yang berisiko tinggi.

- **Peningkatan Layanan Pelanggan:** Memberikan layanan konsultasi keuangan untuk membantu debitur mengelola hutang mereka dengan lebih baik.

Thank You