Rakamin Academy X id/x partners

Prediction Model

ID/X Partners - Data Scientist

Presented by
Nabilah Shamid

Rakamin Academy

Palembang, South Sumatra

nblhshamid08@gmail.com

linkedin.com/in/nabilahshamid/

Nabilah Shamid

Machine Learning & Data Science Enthusiast

An Informatics Engineering Student passionate about technology, data, and machine learning. Experienced in public speaking as an MC and skilled in communication and public relations. Enthusiastic about learning and growth, with a strong foundation in data science, machine learning, and artificial intelligence.

Courses and Certification

Rakamin Academy

Introduction to Microsoft Azure Cloud Services | <link certificate>	**Feb,**	**2025**
Introduction to ML on AWS | <link certificate>	**Jan,**	**2025**
Generative AI with LLM (DeepLearning.AI) | <link certificate>	**Jan,**	**2025**
Generative AI: Introduction & Application | <link certificate>	**Jan,**	**2025**
Build and Deploy ML Solutions in Vertex AI | <link certificate>	**July, 2024**	

About Company

ID/X Partners adalah perusahaan konsultan teknologi yang berbasis pada data dan analitik, yang berfokus pada pengembangan solusi strategis berbasis teknologi untuk klien dari berbagai industri.

Perusahaan ini menawarkan layanan di bidang advanced analytics, machine learning, big data, digital transformation, serta pengembangan solusi berbasis kecerdasan buatan. Dengan pendekatan berbasis data dan pemahaman bisnis yang kuat, ID/X Partners membantu perusahaan dalam mengambil keputusan yang lebih akurat, efisien, dan berdampak jangka panjang.

Kolaborasi lintas disiplin dan penggunaan teknologi mutakhir menjadi ciri utama dalam setiap solusi yang dikembangkan oleh ID/X Partners.

Project Portfolio

Proyek ini bertujuan untuk membangun model machine learning yang mampu memprediksi risiko kredit dari data pinjaman historis. Model ini diharapkan dapat membantu perusahaan multifinance dalam mengambil keputusan yang lebih akurat dan mengurangi potensi kerugian akibat kredit bermasalah.

Dataset yang digunakan berisi informasi pinjaman antara tahun 2007 hingga 2014, dengan berbagai fitur numerik dan kategorikal.

Link code here!

Project explanation video here!

1. Data Understanding

Tahap ini bertujuan untuk memahami karakteristik dan struktur awal dataset historis pinjaman dari tahun 2007-2014 untuk mengidentifikasi kualitas data dan potensi masalah.

Aktivitas Utama

- **Memuat Dataset**: Menggunakan file `loan_data_2007_2014.csv` yang berisi **466.285 baris** dan **75 kolom**.
- **Analisis Statistik Deskriptif**: Melakukan analisis awal untuk memahami distribusi, tipe data, dan ringkasan statistik dari setiap fitur.
- **Identifikasi Kualitas Data**: Melakukan pemeriksaan awal untuk mendeteksi masalah umum seperti nilai yang hilang (*missing values*), duplikasi, dan tipe data yang tidak sesuai.

```
Persentase Missing Values per Kolom:
inq_fi                        100.000000
open_rv_24m                   100.000000
max_bal_bc                    100.000000
all_util                      100.000000
inq_last_12m                  100.000000
annual_inc_joint              100.000000
verification_status_joint     100.000000
dti_joint                     100.000000
total_cu_tl                   100.000000
il_util                       100.000000
mths_since_rcnt_il            100.000000
total_bal_il                  100.000000
open_il_24m                   100.000000
open_il_12m                   100.000000
open_il_6m                    100.000000
open_acc_6m                   100.000000
open_rv_12m                   100.000000
mths_since_last_record         86.566585
mths_since_last_major_derog    78.773926
desc                           72.981975
mths_since_last_delinq         53.690554
next_pymnt_d                   48.728567
tot_cur_bal                    15.071469
```

2. Feature Engineering

Tahap ini bertujuan untuk menciptakan fitur baru yang lebih informatif dan memiliki daya prediksi yang lebih kuat bagi model. Fokusnya adalah mengubah data mentah, khususnya kolom tanggal, menjadi fitur numerik yang lebih bermakna.

Aktivitas Utama

1. **Identifikasi Kolom Tanggal**: Mengidentifikasi kolom `issue_d` (tanggal pinjaman diberikan) dan `earliest_cr_line` (tanggal laporan kredit pertama kali dibuat) sebagai sumber informasi yang potensial.
2. **Transformasi Tipe Data**: Mengubah kedua kolom tersebut dari format teks (`object`) menjadi format `datetime` agar bisa dilakukan kalkulasi.
3. **Membuat Fitur Baru**:
    - Menghitung selisih waktu antara `issue_d` dan `earliest_cr_line` untuk setiap peminjam.
    - Hasil selisih tersebut dikonversi menjadi satuan bulan untuk menciptakan fitur baru bernama **`credit_history_length`**.
4. **Pembersihan Kolom Asli**: Setelah fitur baru berhasil dibuat, kolom-kolom tanggal asli (`issue_d`, `earliest_cr_line`, `last_pymnt_d`, `last_credit_pull_d`) dihapus agar tidak ada informasi redundan.

2. Feature Engineering

Hasil

- Berhasil dibuat sebuah fitur numerik baru, yaitu `credit_history_length`, yang merepresentasikan panjang sejarah kredit seorang peminjam (dalam bulan) pada saat pinjaman diajukan.
- Fitur ini memberikan nilai yang lebih intuitif bagi model, di mana durasi riwayat kredit yang lebih panjang dapat menjadi salah satu indikator penting dalam menilai risiko kredit.

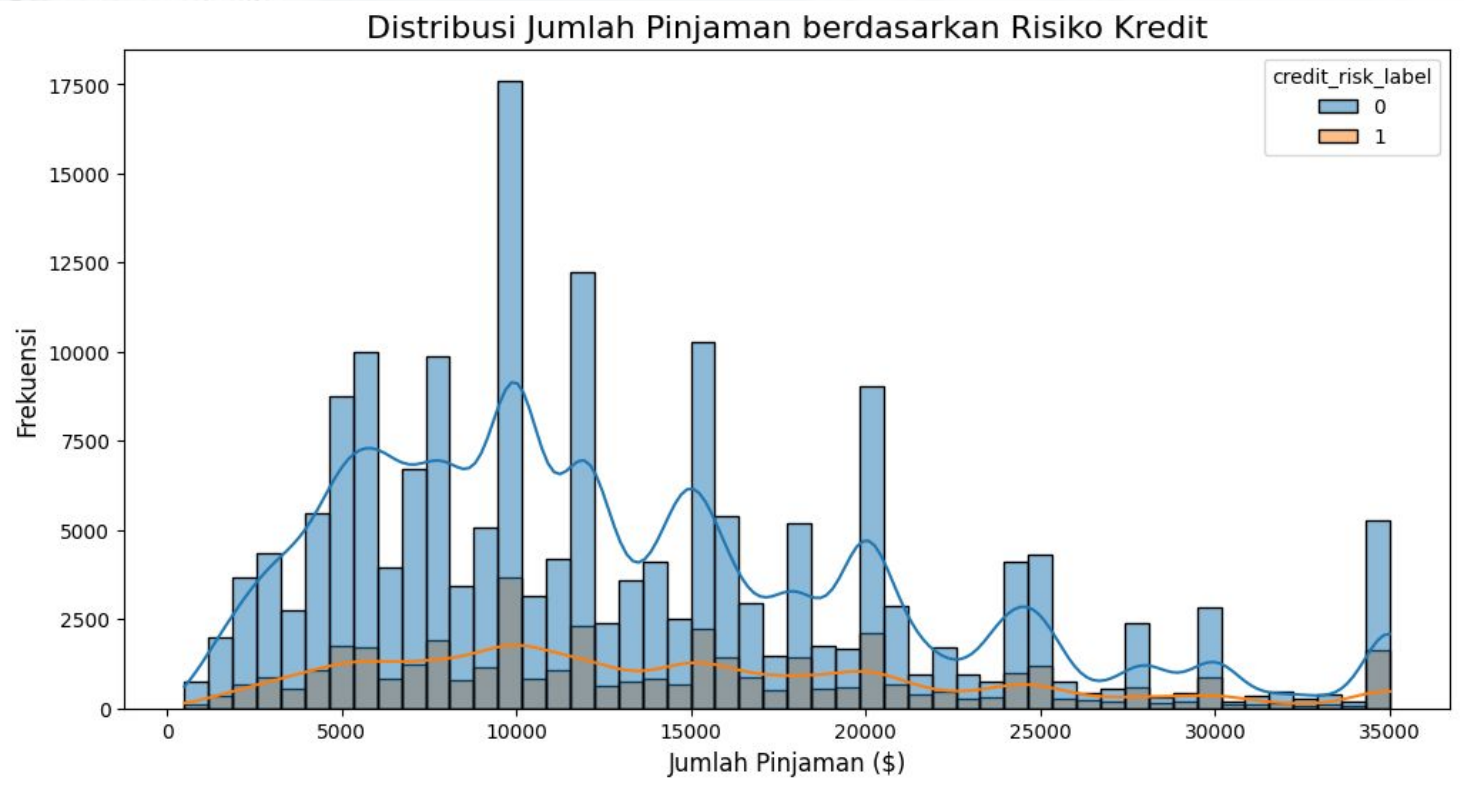
```
Contoh nilai fitur baru 'credit_history_length':
0    327
1    154
2    122
3    192
5     86
Name: credit_history_length, dtype: int64
```

3. Exploratory Data Analysis

Analisis Utama yang Dilakukan

1. **Analisis Variabel Target (`credit_risk_label`)**:
   - Mendefinisikan variabel target biner: **1 (Bad Loan)** untuk pinjaman `Charged Off` dan **0 (Good Loan)** untuk pinjaman `Fully Paid`.
   - Ditemukan bahwa data bersifat **tidak seimbang (*imbalanced*)**, dengan proporsi sekitar **81% Good Loan** dan **19% Bad Loan**.
2. **Analisis Univariat dan Bivariat**:
   - Menganalisis distribusi fitur-fitur individual dan hubungannya dengan risiko kredit.
   - **Grade Pinjaman vs Risiko**: Ditemukan tren yang jelas bahwa semakin rendah grade pinjaman (dari A ke G), semakin tinggi proporsi pinjaman yang gagal bayar.
   - **Grade vs Suku Bunga**: Terbukti ada hubungan kuat di mana grade yang lebih rendah dikenakan suku bunga (`int_rate`) yang lebih tinggi.
3. **Analisis Korelasi**:
   - Membuat *heatmap* korelasi untuk memahami hubungan linear antar fitur numerik.
   - Terlihat korelasi positif antara `int_rate` dan `credit_risk_label`, yang mengonfirmasi bahwa suku bunga yang lebih tinggi berasosiasi dengan risiko yang lebih tinggi.
   - Teridentifikasi adanya korelasi yang sangat kuat (multikolinearitas) antara fitur seperti `loan_amnt`, `funded_amnt`, dan `installment`.

3. Exploratory Data Analysis



4. Data Preparation

Aktivitas Utama

1. **Penghapusan Kolom (Column Removal)**:
   - Kolom yang **tidak relevan** (seperti ID, URL, teks bebas), memiliki **missing values tinggi**, atau berpotensi menyebabkan **kebocoran data (*data leakage*)** dihapus.
   - Penghapusan kolom yang bocor seperti `total_pymnt` dan `recoveries` sangat krusial untuk memastikan model tidak dilatih menggunakan informasi masa depan.
2. **Pembersihan dan Transformasi Fitur**:
   - Membersihkan dan mengubah kolom `term` (jangka waktu) dan `emp_length` (lama kerja) menjadi format numerik yang dapat diolah.
   - Melakukan *feature engineering* pada kolom tanggal untuk membuat fitur baru `credit_history_length`, seperti yang dibahas sebelumnya.
3. **Encoding dan Imputasi**:
   - Mengisi sisa *missing values* pada kolom numerik dengan nilai **median**.
   - Mengubah semua variabel kategorikal yang tersisa menjadi format numerik menggunakan **one-hot encoding**.
4. **Pembagian dan Scaling Data**:
   - Membagi data yang sudah bersih menjadi **80% data latih** dan **20% data uji**.
   - Melakukan standarisasi (*scaling*) pada semua fitur menggunakan `StandardScaler` agar memiliki rentang nilai yang sebanding.

4. Data Preparation

Aktivitas Utama

1. **Penghapusan Kolom (Column Removal)**:
   - Kolom yang **tidak relevan** (seperti ID, URL, teks bebas), memiliki **missing values tinggi**, atau berpotensi menyebabkan **kebocoran data (*data leakage*)** dihapus.
   - Penghapusan kolom yang bocor seperti `total_pymnt` dan `recoveries` sangat krusial untuk memastikan model tidak dilatih menggunakan informasi masa depan.
2. **Pembersihan dan Transformasi Fitur**:
   - Membersihkan dan mengubah kolom `term` (jangka waktu) dan `emp_length` (lama kerja) menjadi format numerik yang dapat diolah.
   - Melakukan *feature engineering* pada kolom tanggal untuk membuat fitur baru `credit_history_length`, seperti yang dibahas sebelumnya.
3. **Encoding dan Imputasi**:
   - Mengisi sisa *missing values* pada kolom numerik dengan nilai **median**.
   - Mengubah semua variabel kategorikal yang tersisa menjadi format numerik menggunakan **one-hot encoding**.
4. **Pembagian dan Scaling Data**:
   - Membagi data yang sudah bersih menjadi **80% data latih** dan **20% data uji**.
   - Melakukan standarisasi (*scaling*) pada semua fitur menggunakan `StandardScaler` agar memiliki rentang nilai yang sebanding.

5. Data Modeling

Tahap ini merupakan tahap membangun dan melatih model *machine learning* yang mampu memprediksi risiko kredit (`credit_risk_label`) berdasarkan data yang telah dipersiapkan pada tahap sebelumnya.

Langkah-Langkah Pemodelan

1. **Pemisahan Fitur (X) dan Target (y)**:
   - Data dibagi menjadi dua bagian: matriks fitur `X` (semua kolom kecuali target) dan vektor target `y` (`credit_risk_label`).
2. **Pembagian Data Latih dan Uji (Train-Test Split)**:
   - Dataset dibagi secara proporsional menjadi **80% data latih** (untuk melatih model) dan **20% data uji** (untuk evaluasi).
   - *Stratifikasi* diterapkan pada variabel target (`stratify=y`) untuk memastikan distribusi "Good Loan" dan "Bad Loan" seimbang di kedua set data.
3. **Scaling Fitur**:
   - Semua fitur numerik distandarisasi menggunakan `StandardScaler`.
   - Tujuannya adalah untuk menyamakan skala nilai antar fitur, yang sangat penting untuk performa model seperti *Logistic Regression*.
4. **Pelatihan Model**:
   - Dua model klasifikasi yang berbeda dilatih menggunakan data latih yang telah di-scaling:
     - **Logistic Regression**: Sebagai model dasar yang wajib digunakan dalam proyek ini.
     - **Random Forest Classifier**: Sebagai model pembanding yang lebih kompleks untuk melihat potensi peningkatan performa.

6. Evaluation

Hasil Perbandingan Model

Setelah dilakukan perbaikan pada masalah *data leakage*, kedua model memberikan hasil performa yang realistis sebagai berikut:

Model	Accuracy	Precision (Bad Loan)	Recall (Bad Loan)	ROC-AUC
Logistic Regression	82.59%	97.35%	5.14%	71.85%
Random Forest	82.51%	94.13%	6.17%	75.98%

6. Evaluation

Evaluation bertujuan untuk mengukur dan membandingkan performa model *Logistic Regression* dan *Random Forest* secara objektif menggunakan 20% data uji yang belum pernah dilihat sebelumnya untuk menentukan model mana yang terbaik.

Metrik Evaluasi

Performa model diukur berdasarkan metrik klasifikasi utama berikut:

- **Accuracy**: Persentase total prediksi yang benar.
- **Precision (Bad Loan)**: Dari semua yang diprediksi "Bad Loan", seberapa banyak yang benar. Penting untuk mengurangi risiko salah menolak nasabah potensial.
- **Recall (Bad Loan)**: Dari semua "Bad Loan" yang sebenarnya, seberapa banyak yang berhasil diidentifikasi. Penting untuk meminimalkan kerugian.
- **ROC-AUC**: Kemampuan model secara keseluruhan dalam membedakan antara "Good Loan" dan "Bad Loan".

7. Conclusion

Proyek ini berhasil mengembangkan model *machine learning* untuk memprediksi risiko kredit (credit risk) bagi sebuah perusahaan *multifinance.* Setelah melalui tahapan *Data Understanding, EDA, Data Preparation,* dan *Modelling*, dua model yaitu **Logistic Regression** dan **Random Forest** telah dilatih dan dievaluasi.

Model Pilihan

Berdasarkan hasil evaluasi, **Random Forest Classifier direkomendasikan sebagai model pilihan**. Meskipun memiliki akurasi yang serupa dengan Logistic Regression, Random Forest menunjukkan performa keseluruhan yang lebih unggul, terutama pada metrik **ROC-AUC (75.98%)**, yang menandakan kemampuan lebih baik dalam membedakan antara pinjaman baik dan buruk.

Temuan Utama dan Tantangan

- **Tantangan *Data Leakage***: Tantangan terbesar dalam proyek ini adalah adanya **kebocoran data (*data leakage*)** yang awalnya menghasilkan akurasi tidak realistis (99%). Masalah ini berhasil diatasi dengan menghapus fitur-fitur yang membocorkan informasi masa depan (seperti total_pymnt dan recoveries), sehingga evaluasi model menjadi lebih valid.
- ***Recall* Rendah Akibat Data Tidak Seimbang**: Kelemahan utama dari model akhir adalah **nilai *recall* yang rendah** untuk kelas "Bad Loan". Hal ini disebabkan oleh sifat data yang tidak seimbang (81% "Good Loan" vs 19% "Bad Loan"), yang membuat model kesulitan mengidentifikasi semua pinjaman berisiko.

7. Conclusion

Rekomendasi dan Langkah Selanjutnya

1. **Untuk Implementasi Saat Ini**: Model **Random Forest** dapat digunakan sebagai alat bantu awal untuk tim penilai kredit, dengan catatan bahwa model ini lebih andal dalam mengonfirmasi pinjaman yang baik daripada mendeteksi pinjaman yang buruk.
2. **Untuk Pengembangan Lanjutan**: Sangat disarankan untuk menerapkan teknik penanganan data tidak seimbang, seperti **SMOTE (*oversampling*)**, pada data latih. Tujuannya adalah untuk meningkatkan *recall*, sehingga kemampuan model dalam mendeteksi pinjaman yang berisiko gagal bayar dapat ditingkatkan, yang pada akhirnya akan mengurangi potensi kerugian perusahaan.

Thank You

Rakamin Academy X id/x partners