PROJECT BASED INTERNSHIP
RAKAMIN X ID/X PARTNERS

End to End Solution

Study Case: Loan Dataset 2007 - 2014

By: Rashif Dhafin Fairiza

Intro

Pada proyek dari sebuah perusahaan lending company, kita diminta untuk membangun model yang dapat memprediksi credit risk menggunakan dataset yang disediakan oleh company yang terdiri dari data pinjaman yang diterima dan yang ditolak.

Step to Solution

1. Explore Dataset
2. Define Label
3. Feature Engineering
4. Feature Selection
5. Handling Missing Values
6. Feature Scaling & Encoding
7. Machine Learning

Explore Dataset

Explore Dataset

- Dataset terdiri dari 466.285 data dan 74 kolom fitur
- Tidak ada data duplikat pada dataset

Problems with The Dataset?

- Banyaknya kolom fitur pada dataset yang tidak terpakai
- Kolom kolom tersebut terdiri dari kolom identitas, kolom free text, dan kolom yang berisi nilai NULL.
- Terdapat kolom ‘sub_grade’ yang merupakan kolom sub kelas pinjaman yang diajukan.
- Kelas pinjaman sudah diwakili kolom ‘grade’

Solutions?

Penghapusan kolom identitas (‘id’, ‘member_id’, ‘zip_code’), kolom free text (‘url’, ‘desc’), kolom yang berisi nilai NULL (‘annual_inc_joint’, ‘dti_joint’, ‘open_acc_6m’, etc.) dan kolom ‘sub_grade’. Terdapat 23 kolom yang dihapus pada tahap ini.

Define Label

Define Label

Untuk mengetahui credit risk seseorang, kita perlu mengetahui status pinjaman dari orang tersebut. Status pinjaman pada dataset terdapat pada kolom ‘loan_status’.

Define Label

Terdapat 9 status pinjaman, antara lain:

- 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

Define Label

Karena tidak ada status yang menyebut langsung pinjaman baik dan buruk, maka kita perlu mengkategorikan 9 status pinjaman sebelumnya menjadi kategori ‘good loan’ dan ‘bad loan’.

Set Label

- good loan: ‘Current’, ‘Fully Paid’, ‘In Grace Period’, ‘Does not meet the credit policy. Status:Fully Paid’, dan ‘Late (16-30 days)’
- bad loan: ‘Charged Off’, ‘Default’, “Does not meet the credit policy. Status:Charged Off’, dan ‘Late (31-120 days)’

Feature Engineering

Feature Engineering

Dari dataset yang kita miliki, terdapat beberapa kolom penting yang harus diubah agar dapat digunakan pada proses Machine Learning. Kolom-kolom tersebut memuat data integer tetapi berformat string serta kolom dengan format date sehingga kolom tersebut perlu diubah.

Feature Engineering

Terdapat beberapa kolom yang harus dilakukan feature engineering antara lain:

- ‘term’
- ‘emp_length’
- ‘issue_d’
- ‘earliest_cr_line’
- ‘last_pymnt_d’
- ‘next_pymnt_d’
- ‘last_credit_pull_d’

Engineered The Feature

Berikut pengubahan yang dilakukan pada beberapa kolom yang disebutkan sebelumnya:

- Penghilangan beberapa karakter pada string dan konversi data ke integer. Kolom yang diubah: ‘term’ dan ‘emp_length’
- Konversi data dari format tanggal ‘bulan-tahun’ ke selisih waktu yang telah berlalu sejak data yang dimasukkan dalam satuan bulan. Tanggal yang digunakan untuk pengurangan adalah Desember 2017. Kolom yang diubah: ‘issue_d’, ‘earliest_cr_line’, ‘last_pymnt_d’, ‘next_pymnt_d’, dan ‘last_credit_pull_d’

Feature Selection

Feature Selection

Terdapat beberapa kolom pada dataset yang memiliki high cardinality, data kategori tertentu yang sangat dominan, dan kolom yang memiliki nilai korelasi tinggi dengan kolom lain yang bukan label. Kolom-kolom tersebut harus dihapus.

Select the Feature

- Kolom dengan High Cardinality: ‘emp_title’ & ‘title’
- Kolom dengan data kategori tertentu yang sangat dominan: ‘application_type’, ‘pymnt_plan’, dan ‘policy_code’
- Kolom dengan korelasi tinggi dengan kolom selain label (Nilai korelasi > 0.7): 15 kolom (‘funded_amnt’, ‘installment’, ‘total_pymnt’, etc.)

Result

Sampai tahap ini, terdapat 43 kolom yang sudah dihapus sehingga jumlah kolom yang tersisa saat ini adalah 31 kolom.

Handling Missing Values

Handling Missing Values

Pada tahap ini, terdapat dua pendekatan untuk menangani missing values dari tiap kolom antara lain:

1. Menghapus kolom dengan persentase missing values diatas 80%.
2. Mengisi missing values dengan nilai yang relevan dengan kolom.

Drop Column

Terdapat 15 kolom yang memiliki missing values diatas 0% namun hanya 1 kolom yang memiliki missing values diatas 80% yaitu kolom ‘mths_since_last_record’ sehingga kolom tersebut harus dihapus.

Fill Missing Values

14 kolom lain yang memiliki missing values diatas 0% akan diisi dengan nilai 0 kecuali kolom ‘annual_inc’. Kolom ‘annual_inc’ merupakan nilai pendapatan tahunan yang diisikan oleh peminjam saat pendaftaran. Perusahaan lending company pasti tidak menerima orang dengan pendapatan 0 sehingga kita isi kolom ‘annual_inc’ dengan rata-rata nilai pada kolom tersebut.

Result

Dataset ini sudah hampir siap untuk kita olah dengan Machine Learning. Jumlah kolom dataset saat ini adalah 30 kolom.

Feature Scaling & Encoding

Feature Scaling & Encoding

- Kolom fitur dengan data numerikal dilakukan scaling dengan StandardScaler
- Kolom fitur dengan data kategorikal kecuali label dilakukan encoding dengan One Hot Encoding

Result

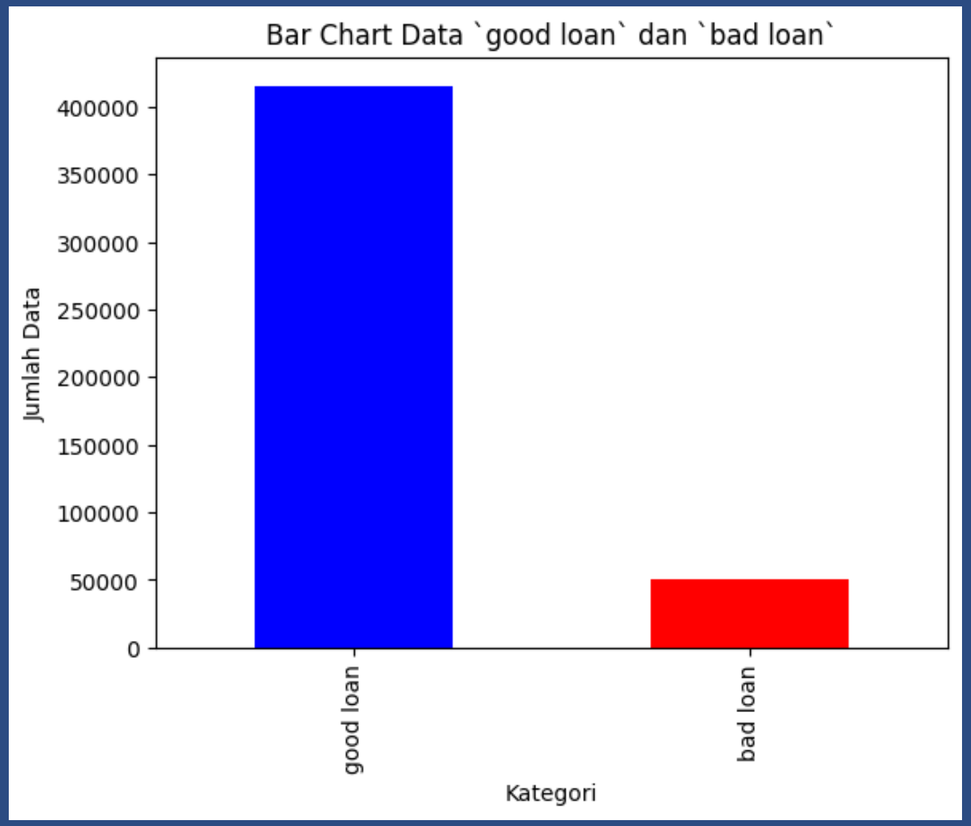
Dataset dengan 99 fitur, 1 label, dan 466.285 data

Machine Learning

Before Machine Learning

Sebelum memprediksi credit risk menggunakan Machine Learning, kita harus pastikan apakah dataset yang kita punya balance atau imbalance.

Is it Balance?



IMBALANCE!!!!

How to Solve It?

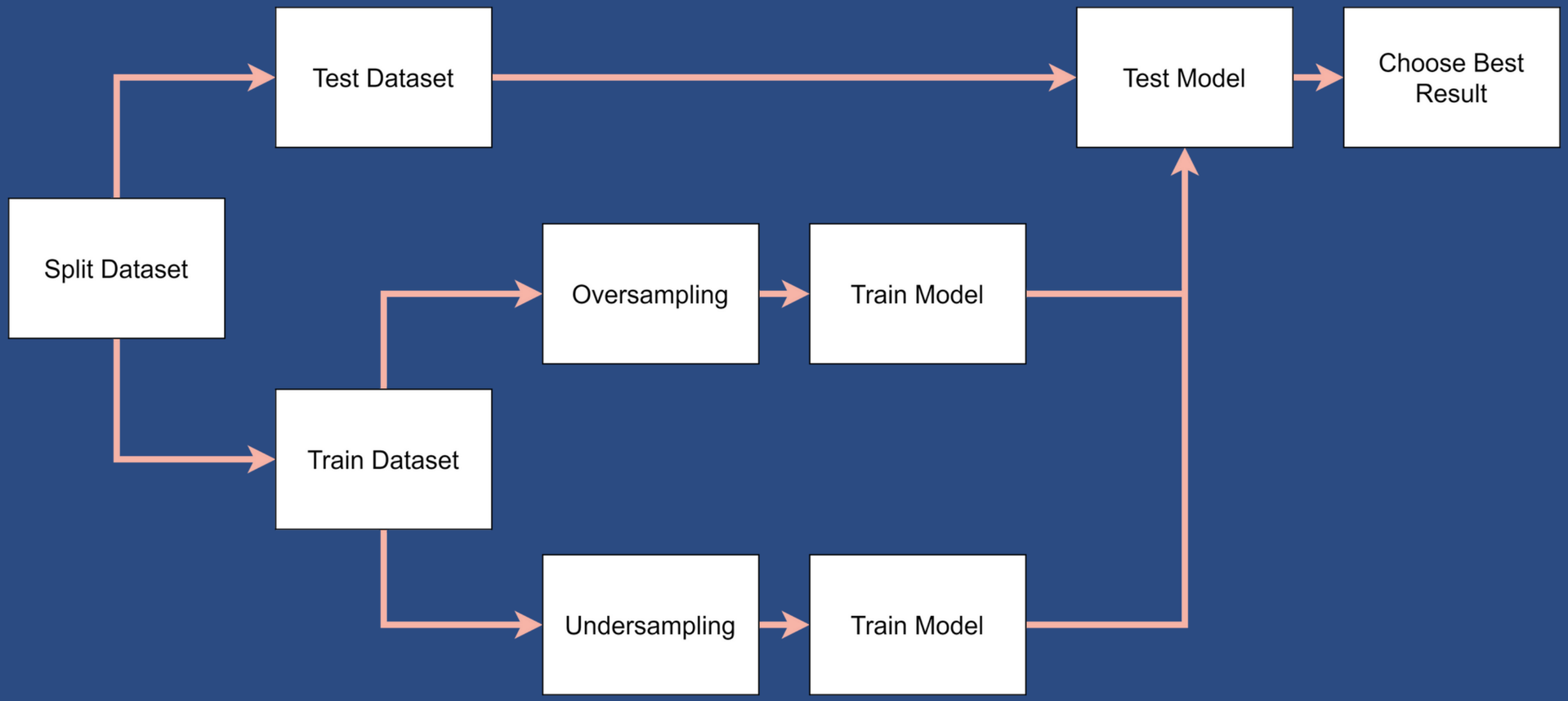
- Oversampling data ‘bad loan’
- Undersampling data ‘good loan’
- Which better? We use it as our prediction scenario
- Untuk hindari data leakage, maka kita lakukan oversampling dan undersampling setelah split dataset menjadi train dataset dan test dataset
- Kita lakukan pada train dataset saja

Machine Learning Model

Karena dataset kita imbalance, maka kita gunakan metode Ensemble Machine Learning. Model yang menggunakan metode Ensemble Machine Learning antara lain:

- Random Forest
- Adaboost
- Gradient Boost

Machine Learning Flow



Split Dataset

- 80% train dataset
- 20% test dataset

Train Dataset

- 373.028 data
- 332.250 data good loan
- 40.778 data bad loan

Test Dataset

- 93.257 data
- 83.067 data good loan
- 10.190 data bad loan

Prediction Scenario 1: Oversampling

Kita lakukan oversampling pada train dataset dan hasilnya sebagai berikut:

- 664.500 data
- 332.250 data good loan
- 332.250 data bad loan

Prediction Scenario 1: Oversampling

Hasil pelatihan model dengan data oversampling dan pengujian model dapat dilihat pada tabel berikut:

Model	Precision	Recall	F1-Score	Accuracy
Random Forest	97%	50%	66%	94%
AdaBoost	42%	72%	53%	86%
Gradient Boosting	42%	73%	53%	86%

Prediction Scenario 2: Undersampling

Kita lakukan undersampling pada train dataset dan hasilnya sebagai berikut:

- 81.556 data
- 40.778 data good loan
- 40.778 data bad loan

Prediction Scenario 2: Undersampling

Hasil pelatihan model dengan data undersampling dan pengujian model dapat dilihat pada tabel berikut:

Model	Precision	Recall	F1-Score	Accuracy
Random Forest	41%	73%	52%	85%
AdaBoost	42%	72%	53%	86%
Gradient Boosting	43%	73%	54%	86%

Compare Result: Scenario 1 vs Scenario 2

Hasil pelatihan model dengan data undersampling dan pengujian model dapat dilihat pada tabel berikut:

Model	Precision	Recall	F1-Score	Accuracy
Random Forest (Scenario 1)	97%	50%	66%	85%
Gradient Boosting (Scenario 2)	43%	73%	54%	86%

Conclusion

- Resampling yang tepat untuk dataset yang diberikan adalah oversampling
- Model yang menghasilkan hasil terbaik adalah Random Forest dengan akurasi 94%

Advice

- Untuk menyempurnakan hasil model Machine Learning, dapat lakukan hyperparameter tuning
- Jika dataset memiliki jumlah data yang banyak dapat menggunakan Deep Learning
- Ubah label menjadi data numerikal seperti 0 atau 1 agar bisa digunakan oleh model Machine Learning yang mengharuskan parameter label berupa data numerik

Thank You