Rangkuman pengerjaan Stage 2 Final Project

Data Cleansing

Pada proses ini, kami berusaha untuk membersihkan data dari missing value, duplicate, outliers, dan imbalance. Tujuannya agar data siap untuk diolah dan bisa lebih mudah dimengerti oleh Machine Learning.

a. Handle missing value

Dari 16 fitur yang ada, terdapat 4 fitur yang memiliki missing value. Fitur-fitur tersebut ialah **duration, campaign, pdays, previous, y.** Setelah melakukan analisis terhadap dataset bank, kami memutuskan untuk menghapus baris data yang kosong karena persentase baris kosong tersebut kurang dari 1%. Dengan menghapus baris-baris yang kosong, kami dapat mempertahankan sebagian besar informasi yang ada dalam dataset tanpa mengorbankan jumlah sampel yang signifikan. Pendekatan ini memungkinkan kami untuk melanjutkan analisis dengan dataset yang lebih lengkap dan representatif, yang diharapkan dapat menghasilkan hasil yang lebih akurat dan reliable dalam konteks analisis yang dilakukan terkait dengan data bank tersebut.

	feature	missing_value
0	age	0
1	job	0
2	marital	0
3	education	0
4	default	0
5	balance	0
6	housing	0
7	loan	0
8	contact	0
9	day	0
10	month	0
11	duration	346
12	campaign	382
13	pdays	217
14	previous	268
15	poutcome	0
16	y	308

b. Handle duplicated data

Dalam mengelola dataset bank, kami telah melakukan penghapusan data duplikat dengan menggunakan subset kolom 'age', 'job', 'balance', 'loan', 'campaign', dan 'y'. Dengan melakukan penghapusan ini, kami memastikan bahwa setiap entri dalam dataset hanya mewakili satu observasi unik yang tidak memiliki duplikat dengan kombinasi nilai pada kolom-kolom yang disebutkan. Setelah dilakukan drop duplicated dan handle missing value, data yang semula berjumlah **45663** menjadi **41285.**

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 45663 entries, 0 to 45662
Data columns (total 17 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   age        45663 non-null  int64
 1   job        45663 non-null  object
 2   marital    45663 non-null  object
 3   education  45663 non-null  object
 4   default    45663 non-null  object
 5   balance    45663 non-null  int64
 6   housing    45663 non-null  object
 7   loan       45663 non-null  object
 8   contact    45663 non-null  object
 9   day        45663 non-null  int64
 10  month      45663 non-null  object
 11  duration   45317 non-null  float64
 12  campaign   45281 non-null  float64
 13  pdays      45446 non-null  float64
 14  previous   45395 non-null  float64
 15  poutcome   45663 non-null  object
 16  y          45355 non-null  object
dtypes: float64(4), int64(3), object(10)
memory usage: 5.9+ MB
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 41285 entries, 0 to 45210
Data columns (total 17 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   age        41285 non-null  int64
 1   job        41285 non-null  object
 2   marital    41285 non-null  object
 3   education  41285 non-null  object
 4   default    41285 non-null  object
 5   balance    41285 non-null  int64
 6   housing    41285 non-null  object
 7   loan       41285 non-null  object
 8   contact    41285 non-null  object
 9   day        41285 non-null  int64
 10  month      41285 non-null  object
 11  duration   41285 non-null  float64
 12  campaign   41285 non-null  float64
 13  pdays      41285 non-null  float64
 14  previous   41285 non-null  float64
 15  poutcome   41285 non-null  object
 16  y          41285 non-null  object
dtypes: float64(4), int64(3), object(10)
memory usage: 5.7+ MB
```

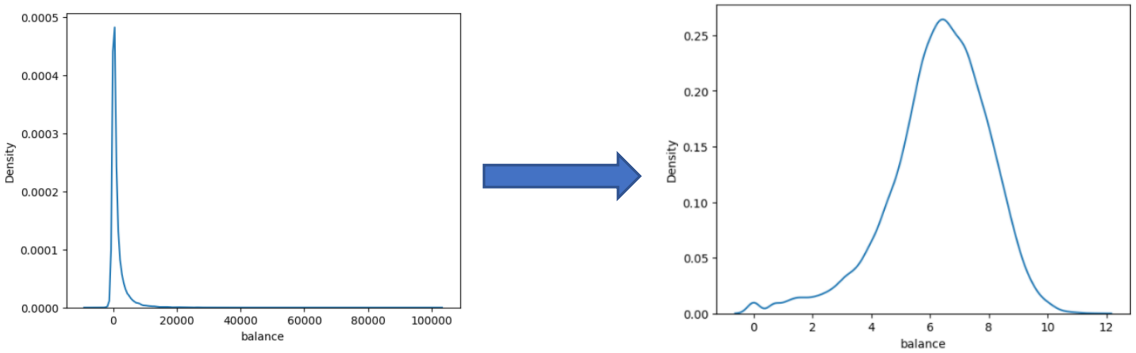
c. Handle outliers

Berdasarkan analisis dataset Bank tersebut, kami tidak perlu menghapus outlier karena nilai-nilai yang ekstrem masih masuk akal atau dapat dijelaskan secara beralasan. Hal ini menunjukkan bahwa dataset tersebut mengandung variasi yang wajar dan tidak ada observasi yang secara signifikan melenceng dari pola umum yang terlihat. Oleh karena itu, outlier-outlier yang ada dalam dataset bank dapat dianggap sebagai bagian yang sah dari variasi data dan tidak perlu dihapus dalam proses analisis.

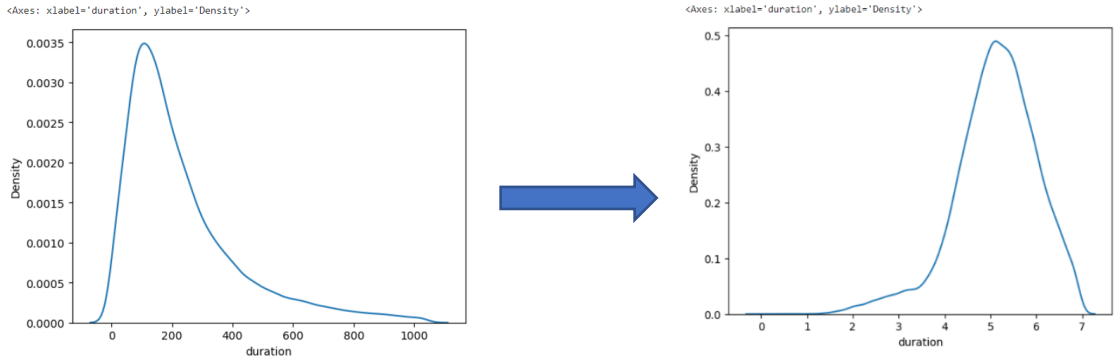
d. Feature Transformation

Pada kesempatan kali ini, kami melakukan log transformation. Log Transformation digunakan pada data yang right-skewed. Distribusi hasil transformasi akan mendekati distribusi normal, seperti pada gambar berikut:

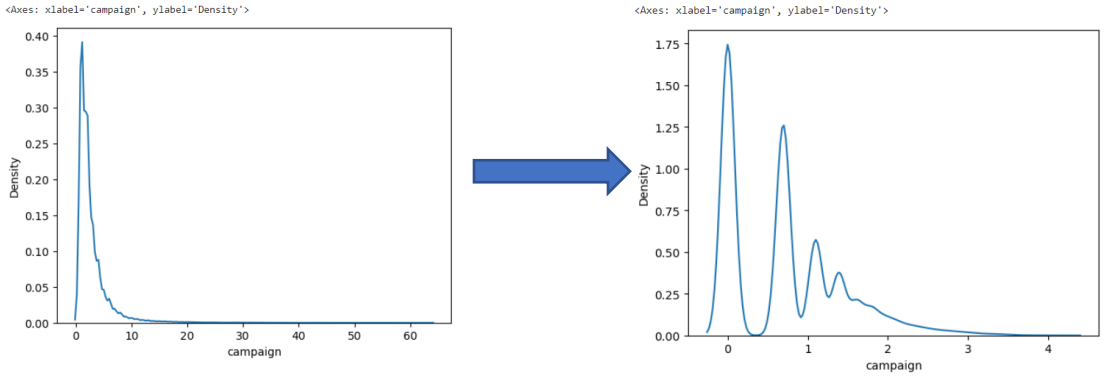
- Fitur balance



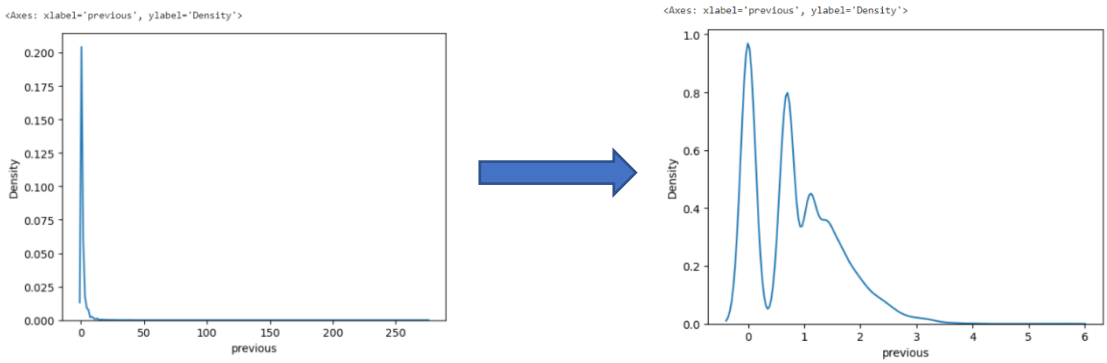
- Fitur duration



- Fitur campaign



- Fitur previous



e. Feature Encoding

Feature Encoding adalah proses mengubah feature categorical menjadi feature numeric. Kami mengubah beberapa fitur yang semula categorical menjadi numerical.

- Label Encoding

  Kami mengubah fitur marital, education, housing, loan, month yang semula category menjadi numeric.

  Before:

	age	job	marital	education	default	balance	housing	loan
0	58	management	married	tertiary	no	2143	yes	no
1	44	technician	single	secondary	no	29	yes	no
2	33	entrepreneur	married	secondary	no	2	yes	yes
3	47	blue-collar	married	unknown	no	1506	yes	no
4	33	unknown	single	unknown	no	1	no	no

  After:

	age	job	marital	education	default	balance	housing	loan
0	58	management	1	3	no	2143	1	0
1	44	technician	0	2	no	29	1	0
2	33	entrepreneur	1	2	no	2	1	1
3	47	blue-collar	1	0	no	1506	1	0
4	33	unknown	0	0	no	1	0	0

- One hot encoding
  Melalui teknik ini, kami memecah fitur job menjadi fitur tersendiri.

job_entrepreneur	job_housemaid	job_management	job_retired	job_self-employed	job_services	job_student	job_technician	job_unemployed	job_unknown
0	0	1	0	0	0	0	0	0	0
0	0	0	0	0	0	0	1	0	0
1	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1

f. Handle class imbalance
   - Melakukan pemisahan feature menjadi dua yaitu kategorikal dan numerical.
   - Menghitung matriks korelasi antara fitur menggunakan fungsi '.corr()'
   - Melakukan visualisasi data dengan heatmap untuk melihat korelasi
   - Melakukan pemisahan data (split) menjadi data train dan data set
   - Melakukan class imbalance terhadapat fitur target 'y' menggunakan oversampling (SMOTE) dan undersampling

Feature Engineering

a. Feature selection

Feature selection dilakukan untuk menentukan fitur dari fitur sebelumnya atau menghapus fitur yang kurang relevan yang akan digunakan untuk modelling. Pada *feature selection* ini dilakukan analisis heatmap dan pairplot untuk mengetahui apakah terdapat korelasi yang kuat dari data yang akan menyebabkan multikoloniearitas dan menjadikan model tidak optimal.

Pada feature selection ini dilakukan untuk menentukan feature terbaik yang akan digunakan untuk modelling, didapat 3 kesimpulan:

- 1. Korelasi: fitur `marital` memiliki korelasi cenderung kuat dengan `age` yang diduga akan menyebabkan multikoliearitas sehingga harus dibuang salah satu. Pada kasus ini fitur `marital` yang dibuang karena dari lini bisnis fitur `age` sudah merepresentasikan tingkat nasabah deposito
- 2. Random Forest: cara ini digunakan untuk menentukan fitur terbaik untuk modelling sehingga didapatkan fitur terbaik yaitu `balance`, `age`, `day`, `month`, `education`, `housing`, `loan`, `default`
- 3. Feature `pdays` juga harus dihapus dikarenakan memiliki nilai negatif sehingga akan mengganggu modelling

b. Feature extraction
   Feature extraction dilakukan untuk membuat fitur baru dari fitur yang sudah ada. Berikut hasil pekerjaan kami:
   - Rasio Balance Terhadap Durasi: Rasio saldo terhadap durasi memberikan perbandingan antara seberapa besar saldo rekening pelanggan dibandingkan dengan durasi kontak yang dilakukan. Jika rasio ini tinggi, dapat menunjukkan bahwa pelanggan dengan saldo yang tinggi cenderung lebih mungkin memberikan tanggapan atau merespons kampanye.
   - Usia Tersegmentasi: Segmentasi usia ini dapat membantu meningkatkan interpretasi model dan juga memungkinkan model untuk mengenali pola atau tren yang spesifik untuk kelompok usia tertentu.
   - Kontak per Hari : dapat mengidentifikasi berapa banyak kontak yang optimal dilakukan dalam satu hari untuk mencapai tujuan tertentu tanpa menyebabkan kelelahan atau kejenuhan pada pelanggan.
   - Kesimpulan Peningkatan Interaksi Sebelumnya : Dengan tingkat interaksi yang lebih tinggi, ada kemungkinan bahwa pelanggan akan memberikan respons yang lebih baik terhadap kampanye pemasaran saat ini. Mereka mungkin lebih cenderung untuk berlangganan, membeli, atau melakukan tindakan yang diinginkan oleh kampanye tersebut.

c. Tuliskan minimal 4 feature tambahan
   - Address
     Kita dapat mengkategorikan nasabah berdasarkan alamat atau domisili. Dengan begitu, kita dapat melihat pola perilaku pembelian, apakah di suatu kota kecenderungan membeli nasabah tinggi/rendah.
   - Credit score
     Fitur ini memberikan informasi apakah nasabah lancar atau tidak dalam membayar hutang. Dengan begitu, kita dapat mengetahui apakah nasabah berpotensi atau tidak membuka deposito berjangka.Rata-rata saldo tahunan
     Fitur ini berguna untuk mengetahui rata-rata saldo tahunan nasabah. Dengan begitu, kita dapat memetakan apakah nasabah masuk dalam kategori rata-rata low, medium, atau high. Jika masuk kategori medium/high, besar kemungkinan nasabah berpotensi untuk membuka deposito berjangka.
   - Children
     Fitur anak ini berguna untuk mengetahui apakah nasabah sudah memiliki anak atau belum.
     Jumlah anak dapat memengaruhi keputusan seseorang untuk membeli sesuatu, termasuk deposito berjangka. Nasabah yang sudah memiliki anak tentu perlu mengatur keuangannya dengan bijak. Oleh karena itu fitur ini bisa menjadi faktor penting dalam memprediksi konversi.

GIT

https://github.com/alfiansetiawan13/Stage-2_Group-3_Finest-Professionals