Rakamin Academy

ABSOLUTE

Kelompok 7:
1. Fikrie Lazuardi
2. Retno Dwi
3. Yunita Rachmawati
4. Rian Dwi Haryono
5. Wahyuni Salman
6. Retno Harindhi
7. Hidayat Yatul

STAGE 2 (Pre-processing)

Data Cleansing

Handle duplicated data
Tidak terdapat data yang duplicate

Handle outliers
Kami menggunakan zscore < 3 untuk menghilangkan outlier pada semua kolom numerical. Dimana jumlah baris sebelum filter outlier adalah 45211 dan setelah filter outlier jumlah baris menjadi 40211.

```
[ ]  #Kita akan menggunakan Z-score filtering (cenderung lebih konservatif)
     from scipy import stats
     print(f'Jumlah baris sebelum memfilter outlier: {len(dftrain)}')

     filtered_entries = np.array([True] * len(dftrain))

     for col in numericals:
         zscore = abs(stats.zscore(dftrain[col]))
         filtered_entries = (zscore < 3) & filtered_entries

     dftrain = dftrain[filtered_entries]

     print(f'Jumlah baris setelah memfilter outlier: {len(dftrain)}')

     Jumlah baris sebelum memfilter outlier: 45211
     Jumlah baris setelah memfilter outlier: 40211
```

```
[ ]  df_train.duplicated().any()

     False
```

Data Cleansing

Feature Transformation

- Kolom Age

Kami melakukan transformasi kolom age dengan normalisasi karena datanya sudah mendekati distribusi normal.

- Kolom Balance

kita transformasi kolom Balance dengan standarisasi. Karena datanya(highly positive skew).

- Kolom Day

kita transformasi kolom Day dengan normalisasi. Karena sejak awal kolom day tdk memiliki outlier.

- **Kolom Duration**

Kita tranformasi kolom Duration dengan standarisasi. Karena datanya(highly positive skew).

Data Cleansing

Feature Transformation

- **Kolom Campaign**

kita transformasi kolom Campaign dengan standarisasi. Karena datanya(highly positive skew).

- **Kolom Pdays**

kita transformasi kolom PDay dengan normalisasi. Karena kolom pday hanya sedikit memiliki outlier.

- **Kolom Previous**

kita transformasi kolom Pervious dengan normalisasi. Karena kolom pervious hanya sedikit memiliki outlier.

Data Cleansing

Feature Encoding

Strategi yang kita gunakan untuk dataset banking adalah Label Encoding. Kita memilih metode Label Encoding karena menurut kami lebih aman untuk digunakan pada logika machine learning, karena jika menggunakan One Hot Encoding dikhawatirkan pada saat pemilihan algoritma akan menimbulkan hasil 1.0 secara keseluruhan.

Handle Class Imbalance

Kami menggunakan random oversampling + undersampling dengan imblearn. Selain untuk menyeimbangkan data kita gunakan random over dan under sampling agar kita bisa meningkatkan sampel kelas minoritas sama dengan kelas mayoritas lain.

Feature Engineering

Feature Selection (membuang feature yang kurang relevan atau redundan)
Karena distribusi pdays dan previous terlalu tinggi kemungkinan kami akan membuang feature salah satu dari kolom tsb dengan standarisasi dikarenakan datanya (*highly positive skew*).

Feature tambahan
- Feature income nasabah
- Feature email nasabah
- Feature internet banking (aktivitas nasabah)
- Feature debet / kredit nasabah