Kelompok: Market Insider
Stage: 2
Mentor: Kevin
Pukul/ Tanggal: 20:00/ 13 Desember 2023

Pembagian tugas di stage ini:

1. Achmad Hilman Shadiqin - Data Analyst
2. Riyan Maula - Data Analyst
3. Nabilah Astiarini - Data Analyst
4. Andreawan Sofian - Data Scientist
5. Figo Akmal Munir - Data Scientist
6. Dzakwan Darussalam - Data Scientist

Poin Pembahasan:

Sesi mentoring ini membahas mengenai tahap Data Pre-Processing yang telah didiskusikan oleh tim sebagai berikut:

1. Handling Missing Values

Pada bagian ini dilakukan identifikasi apakah ada missing values atau tidak pada dataset dan melakukan handling jika terdapat missing values.

2. Handling Duplicates Values

Melakukan identifikasi apakah terdapat duplicate values atau tidak, jika ada maka akan dilakukan handling.

3. Handling Outliers

Melakukan handling terhadap outliers yang sebelumnya sudah kita ketahui pada tahap EDA.

4. Feature Extraction

Membuat feature Baru dari feature yang ada, dan mengekstrak feature penting dari data yang sudah ada.

5. Feature Encoding

Pada tahapan ini dilakukan perubahan pada feature dari feature categorical menjadi feature numeric.

6. Feature Transformation

Mengubah feature kedalam bentuk yang lebih mudah dipahami oleh model.

7. Feature Selection

Pada tahapan ini dilakukan analisa data yang bertujuan untuk melihat feature yang memiliki pengaruh paling optimal.

8. Imbalance Handling

Pada tahap ini dilakukan handling terhadap class imbalance pada data target.

Kelompok: Market Insider
Stage: 2
Mentor: Kevin
Pukul/ Tanggal: 20:00/ 13 Desember 2023

Hasil Diskusi:

1. Handling Missing Values, Duplicates Values & Outliers

- Untuk menangani missing values, data duplicates dan outlier, telah dilakukan dengan menghapus datanya, karena persentase data dibawah 10%.
- Ketiga tahap ini sudah dilakukan dengan baik dan tidak ada notes/comment untuk tahap ini.

2. Feature Extraction

- Membuat feature baru berdasarkan status hubungan (Relationship Status):

Values Absurd dan YOLO diasumsikan masuk kedalam kategori not in relationship. Suggest dari mentor untuk menghapus kedua values tersebut karena sulit untuk dideskripsikan dan jumlah data tersebut masih dibawah 10%.

- Membuat feature baru berdasarkan jumlah keluarga (Family size):

Terdapat comment untuk jumlah anggota keluarga jika statusnya in relationship maka family size 2 dan jika not in relationship family size belum tentu bernilai 2, pertimbangkan untuk status lain seperti widow dimana status tersebut termasuk kedalam kategori not in relationship namun belum tentu tidak memiliki anak, maka harus di pastikan kembali dan merubah logicnya.

- Membuat feature baru (Costumer Lifespan, Year, Total Purchases, Total Spending, Total Offers, Spending Purchases Ratio, Deal Purchases Ratio, At least One Campaign, More One Campaign, Primary Needs & RFM Score) dan mengubah value pada beberapa feature (Education & Marital Status) sudah dilakukan dengan baik dan tidak ada notes/comment.

3. Feature Encoding

- Pada Feature Education melakukan feature encoding menggunakan Ordinal Encoder. Disarankan juga melakukan One Hot Encoding dikarenakan values bersifat ordinal.

4. Feature Transformation

- Menilai bentuk distribusi berdasarkan skewness & kurtosis untuk penentuan standarisasi. Selain hasil dalam bentuk table, disarankan untuk menambah plotnya.

Kelompok: Market Insider
Stage: 1
Mentor: Kevin
Pukul/ Tanggal: 20:00/ 29 November 2023

Rakamin Academy

Hasil Diskusi:

5. Feature Selection

- Pada tahap ini sudah dilakukan Anova untuk feature numerikal – kategorikal. Disarankan untuk menambah Chi-square untuk feature kategorikal-kategorikal agar bisa dipisahkan antara fitur kategorikal dan numerical.

6. Imbalance Handling

- Sudah dilakukan dengan baik menggunakan SMOTE.

Notes tambahan:

- Untuk data splitting bisa dilakukan di awal menjadi data training dan data testing sebelum memulai tahapan lainnya. Lalu data training ditindaklanjuti dengan pre-processing seperti handling outlier, feature transformation dll. Data testing tidak perlu diproses.
- Di tahap selanjutnya, bisa dicek kembali mengenai feature dengan korelasi yang tinggi terhadap response.
- Pertimbangkan feature baru yang merupakan gabungan dari beberapa feature. Contoh: jika feature RFM terpilih, maka feature Recency, Total Purchase, dan Total Spending perlu dihapus/tidak digunakan agar tidak redundant/double. Atau jika feature Total Children (gabungan dari Total Kidhome dan Total Teenhome) terpilih, maka feature Total Kidhome dan Total Teenhome dihapus.

Kelompok: Market Insider
Stage: 1
Mentor: Kevin
Pukul/ Tanggal: 20:00/ 29 November 2023

Tindak Lanjut:

Setelah mendapatkan feedback dari mentor, berikut adalah perbaikan dan tambahan pada beberapa sub-tahap di tahap Data Pre-Processing:

Sebelum pre-processing, telah dilakukan data splitting menjadi data training dan data testing. Lalu dilanjutkan dengan sub-tahapan Pre-Processing.

1. Feature Extraction

- Memperbaiki logic untuk feature Family Size untuk mengetahui jumlah anggota keluarga.

2. Feature Transformation

- Menambahkan plot distribusi untuk memperlihatkan sebaran data secara visual.

3. Feature Selection

- Menambahkan uji Chi-Square pada feature kategorikal-kategorikal dan mengecek redundant analysis.