Rakamin Academy

Homework Machine Learning Preparation

Kelompok 2 :

- Wisnu Pri Hartono
- Muhammad Zulfarhan
- Arman Lukman
- Farki Mahbubi
- Radithya Arif Pambudi
- Surya Praviarti
- Raihan Damar

Teknis Pengerjaan

1. Pekerjaan dilakukan secara berkelompok, sesuai kelompok Final Project
2. Masing-masing anggota kelompok tetap perlu submit ke LMS (jadi bukan perwakilan)
3. File yang perlu dikumpulkan:
   - File jupyter notebook (.ipynb) yang berisi source code.
   - File slides (.pdf) simple slides presentasi yang berisi rangkuman dari apa saja yang telah dilakukan.
4. Upload hasil pengerjaanmu melalui LMS.
   - Masukkan semua file ke dalam 1 file dengan format ZIP.
   - Nama File: ML Preparation - .zip

Rangkuman

1. Descriptive Statistics

Menggunakan function info dan describe pada dataset terdapat beberapa temuan yaitu :

- Terdapat kolom dengan tipe data kurang sesuai, atau nama kolom dan isinya kurang sesuai yaitu kolom id dengan tipe data int dan kolom exclusive dengan tipe data int.
- Terdapat kolom yang memiliki nilai kosong yaitu kolom category, rating, number_of_reviews, love, price, dan value_price.
- Terdapat kolom yang memiliki nilai summary agak aneh yaitu pada kolom id dan exclusive yang sebaiknya menampilkan nilai count, unique, top, dan freq.

2. Univariate Analysis

Menggunakan visualisasi untuk melihat distribusi masing-masing kolom (feature maupun target) terdapat temuan sebagai berikut :

- Fitur id dan rating mengalami distribusi data dengan nilai skewed negatif. Sedangkan fitur number_of_reviews, love, price, dan value_price distribusi data dengan nilai skewed positif. Dan yang menarik ialah untuk fitur exclusive distribusi data bimodal.
- Tiap fitur di kelompok data numerical ada outlier kecuali exclusive.
- Pada fitur di kelompok data category memiliki category yang terlalu banyak.
- Hal yang perlu di folllow up pada saat preprocessing data: Mengubah fitur id dan exclusive menjadi categorical, mengubah data yang terdistribusi kekanan atau kekiri menjadi distribusi normal.

Rangkuman

3. Multivariate Analysis

Melakukan multivariate analysis (seperti correlation heatmap dan category plots, sesuai yang diajarkan di kelas). Ditemukan beberapa temuan, yaitu :

- feature yang paling relevan dan harus dipertahankan adalah id, rating, exclusive, brand, dan category.
- Terdapat Korelasi antar-feature yaitu korelasi antara love dengan number_of_reviews dan price dengan value_price yang mana nilai korelasinya diatas 0.7 sehingga perlu di drop fiturnya karena kemungkinan fitur tersebut redundan.

4. Data Cleansing

Dilakukan pembersihan data, sesuai yang diajarkan di kelas, yaitu :

- Dari setiap kolom data yang kosong, tampaknya memiliki jumlah yang kurang dari 10% dari data keseluruhan (terbilang sedikit). Tim kami memilih data kosong pada number_of_reviews, price dan value_price di drop, sedangkan sisanya disi dengan nilai modus/mean.
- Terlihat pada plot sebelumnya distribusi setiap kolom memiliki skew atau tidak memiliki distribusi yang normal, maka tim kami memilih metode IQR untuk digunakan pada handling outliers.
- Untuk efisiensikan model maka dilakukan Log Transformation pada dataset untuk MERUBAH bentuk sebaran data menjadi mendekati normal.
- Melakukan One-hot Encoding untuk mengubah feature categorical menjadi numeric dengan menjadikan masing-masing nilai unik feature tersendiri pada kolom Brand dan Category .
- Distribusi nilai unik pada target sangat timpang antara 0 dengan 1, oleh karena itu tim kami melakukan resampling dengan metode SMOTE.

Rangkuman

5. Feature Engineering

Dilakukan pembersihan data, sesuai yang diajarkan di kelas, yaitu :

- Feature selection (membuang feature yang kurang relevan atau redundan). Saat dibuat plot heatmap, fitur love berkorelasi kuat dengan number_of_reviews dan value_price kuat dengan price. Tim kami memutuskan untuk membuang fitur love & value_price.
- Feature extraction (membuat feature baru dari feature yang sudah ada). Fitur yang dapat diekstraksikan dari fitur yang ada misalnya fitur ratio dari number_of_reviews dengan rating.
- Feature tambahan (selain yang sudah tersedia di dataset) yang mungkin akan sangat membantu membuat performansi model semakin bagus yaitu :
    - Fitur Age of Product
    - Fitur Waktu (Hari, Bulan atau Tahun)
    - Fitur Diskon dari Price
    - Fitur Ketersediaan Produk di toko Online atau Tidak
    - Fitur jumlah produk terjual
    - Fitur keanggotaan
    - Fitur rating layanan pelanggan
    - Fitur ketersediaan produk kustom
    - Fitur asal produk
    - Fitur tanggal peluncuran

Rakamin Academy

Terima Kasih !