Product Exclusive Classification

Memprediksi apakah suatu produk eksklusif atau tidak berdasarkan fitur yang tersedia

Agi Rahmawandi_Batch 57

Descriptive Statistics

Memahami karakteristik dan struktur data dari Product_Exclusive_Classification.csv, didapat kesimpulan sebagai berikut :

A. Type data dan nama kolom sudah sesuai

- - int64: id, exclusive
- - float64: rating, number_of_reviews, love, price, value_price
- - object: brand, category

B. kolom yang memiliki nilai kosong adalah category,rating, number_of_reviews,love,price,value_price

C. Terdapat beberapa kolom yang memiliki summary yang dirasa kurang tepat.

- - kolom love mempunyai nilai`max 1.300.000` ke nilai `mean 17.563` maupun ke nilai `median 5.500` yang terlampau jauh,kemungkinan ada outlier
- - kolom number_of_reviews nilai `max 19000 `ke nilai`mean 303.57`maupun ke nilai `median 56` yang terlampau jauh,kemungkinan ada outlier
- - begitupun dengan kolom price dan price_value jarak nilai `max 549` ke median ataupun ke mean nya terlihat jauh, kemungkinan juga ada outlier

Pendekatan Grafis

Melihat bentuk distribusi data dari masing masing kolom :

- Kolom rating, data berkumpul di rentang angka 3 sampai 5, karena ada nilai minimal 0, perlu di pertimbangkan apakah dihapus ataukan digunakan.
- Kolom number_of review, Love,Price,value_price terlihat memiliki pola yang sama yakni bertumpuk dinilai kecil/ skewed positif, serta memiliki outlier yang cukup jauh. Perlu dilakukan scaling dan handling outlier saat data pre-processing
- Lalu pada kolom exclusive angka 0 (‘tidak’) lebih banyak dari angka 1 (‘ya’) datanya imbalace, perlu penangan data imbalance dalam pre-processing.
- Lalu pada kolom berupa kategory, yakni kolom ‘brand’ dan ‘category’, kategorinya terlalu banyak perlu dilakukan encoding

Multivarite

Dari korelasi antar kolom kita bisa melihat bahwa :

- Kolom yang paling relevan dan dipertahankan ialah kolom price senilai -18 yang berkolerasi negatif.
- Value_price redundan dengan kolom price, makan akan didrop atau dibuat feature baru dari dua feature ini..
- Kolom love dan number_of_reviews redundan dan akan diambil salah satu.
- Meski memilki nilai korelasi cukup besar kolom ‘id’ akan di drop karena kolom id merupakan Identifier, bukan fitur.

Data Cleaning

Melakukan data cleaning :

- **Mising value**, mengisi nilai kosong dengan memakai mode, dan hanya nilai price yang diberi nilai rata2.
- **Dulplicated data** , tidak ada nilai duplikat.
- **Outlier** , Karena data seperti **price_value, price, love, dan number_of_reviews** tampak sangat skewed dan memiliki outlier ekstrem, maka digunakanlah **IQR** untuk mengilangkan outlier.
- **Transformation ,** Melakukan transformasi menggunakan **logarithmic transformation** pada fitur yang terdistribusi skewed, dan mempunyai nilai outlier ekstrim.
- **Encoding,** karenanya jenis kategorinya begitu banyak maka digunakan **Target Encoding**
- **Class imbalance** Tager Exclusive bersifat imbalance **74% Tidak exclusive** dan **26% exclusive**. penanganan imbalance dilakukan menggunakan **SMOTE** (Synthetic Minority Over-sampling Technique) untuk menyeimbangkan jumlah sampel

Feature Engineering

Feature Selection :

- Kolom **price dan value_price** nilainya cukup signifikan berkorelasi negatif terhadap target, namun redundan 0.99 antar keduanya artinya informasi yang dibawa keduanya hampir sama,bisa di hapus salah satu seprti **value_price** setelah datanya diekstak menjadi feature baru.
- **brand_encoded dan category_encoded** meski memiliki nilai yang cukup signifikan 0.81 terhadap target, ini merupakan predikor yang penting, jadi tidak akan dihapus.
- Feature **id** akan dihilangkan meski memilki nilai yang cukup signifikan terhadap target, karena id adalah identifier unik, tidak bermakna prediktif .

Feature extraction :

Membuat feature baru untuk menghitung tinggkat discount yang didapat dari feature price dan value_price, di dapat kolom **‘discount_pct’ untuk melihat presentase, dan kolom ‘price_gap’ untuk melihat nilai dari selisih price dan value price.**

Feature Engineering

Feature tambahan

Berikut feature tambahan yang mungkin bisa membantu dalam performansi model lebih baik :

- **brand_tier,** membuat kategori luxury,premium,common
- **product_age**, umur produk setelah diluncurkan
- **limited_edition**, apakah produk termasuk limited edition atau bukan
- **seasonal_product**, produk yang direalease saat event tertentu (hari raya, natal, tahun baru,dsb)

b