# train.csv

developerCountry

* ~~Menggabungkan nilai “CANNOT IDENTIFY COUNTRY” dan “ADDRESS NOT LISTED IN PLAYSTORE” menjadi “UNKNOWN”~~
* Frequency encoding

countryCode

* Drop missing value
* Frequency encoding

userRatingCount & averageUserRating

* Where userRatingCount = 0 🡪 averageUserRating = 0
* Standarscaler both of them

primaryGenreName

* ~~Mengubah genre sbb~~

~~genre\_mapping = {~~

~~'Travel': 'Travel & Local',~~

~~'Music' : 'Music & Audio',~~

~~'Photography': 'Photo & Video',~~

~~'Reference': 'Books & Reference',~~

~~'Book': 'Books & Reference',~~

~~'Comics' : 'Books & Reference',~~

~~'Social Networking': 'Social',~~

~~'News': 'News & Magazines',~~

~~'Magazines & Newspapers' : 'News & Magazines'~~

~~}~~

* Frequency encoding

downloads

* Mapping nilai yang salah

# Mapping nilai yang salah (akibat konversi tanggal otomatis) ke nilai download asli

download\_fix\_map = {

'01-May': '1 - 5',

'Oct-50': '10 - 50',

'05-Oct': '5 - 10',

'01-Jan': '1 - 10',

'10-Jan': '10 - 100',

'05-Jan': '5 - 10',

'50 - 1': '50 - 100', # ini juga kemungkinan error kebalik, karena normalnya min-max

'500 - 1': '500 - 1000' # idem

}

* Memecah kolom downloads menjadi minDownloads (batas bawah) dan maxDownloads (batas atas)
* Impute missing value pada minDownloads dan maxDownloads dengan median
* Handling outlier: log transform 🡪 winsorizing 🡪 clip nilai minimum 🡪 winsorizing

1. minDownloads
   * standardscaler
2. maxDownloads
   * standardscaler
3. ~~log\_min\_winsor~~
   * ~~standardscaler~~
4. ~~log\_max\_winsor~~
   * ~~standardscaler~~

deviceType

* One hot encoding

hasPrivacyLink

* Drop (korelasi dengan target < 1%)

hasTermsOfServiceLink & hasTermsOfServiceLinkRating

* ~~Gabungkan keduanya, membentuk kolom “terms\_service\_combined” yang kemudian di one-hot encoding ordinal encoding menjadi “terms\_service\_combined\_has\_link\_low” dan “terms\_service\_combined\_no\_link”~~
* ~~Impute missing value dengan -1 dan buat flag missing value untuk membedakan data hasil imputasi dan data asli~~
* Drop kolom hasTermsOfServiceLink
* Ordinal encoding Isi missing value pada hasTermsOfServiceLinkRating dengan -1
  + Missing value diisi dengan “undetermined”
  + Ordinal encoding
    - Low = 0
    - High = 1
    - Undetermined = -1

isCorporateEmailScore

* Buat flag untuk missing value (-1) 🡪 nilai pada isCorporateEmailScore: 0, 99, -1
* Standardscaler

1. isCorporateEmailMissing
   * standardscaler

adSpent

* Rekomendasi: drop karena missing value sangat banyak (81%) dan tidak ada kolom lain yang bisa memprediksi adSpent (baik random sampling maupun iterative imputer tidak berguna)
* Opsi lain (not recommended): Impute missing value dengan random sampling

appAge

* Drop missing value (<1% dari data)
* Standardscaler

appContentBrandSafetyRating

* Drop kolom karena:
  + /Cramér's V sangat rendah bahkan pada data yang tersedia (0.0131)
  + missing value sangat tinggi (88%) 🡪 Risiko memperkenalkan noise jika diimputasi

appDescriptionBrandSafetyRating

* ~~one hot encoding~~ ordinal encoding

mfaRating

* drop kolom mfaRating karena nilainya hampir semuanya low 🡪 ga ada variasi 🡪 ga ada korelasi dengan kolom target

coppaRisk

* Handling imbalanced:

# test.csv

Im gonna experiment mana yang menghasilkan lebih bagus

* Imputasi, scaling, freq encoding menggunakan sebaran data training (X\_train dari train.csv)
* Imputasi, scaling, freq encoding menggunakan sebaran data prediksi sendiri (test.csv). feeling gue bakal bagusan yang ini karena data buat prediksi completely new. Kalo kita impute pake sebaran data yang digunakan untuk training model justru modelnya bisa langsung tau karena udah pernah liat. Ibarat ngerjain ujian soalnya sama persis kayak kisi-kisi

developerCountry

* ~~Menggabungkan nilai “CANNOT IDENTIFY COUNTRY” dan “ADDRESS NOT LISTED IN PLAYSTORE” menjadi “UNKNOWN”~~
* Frequency encoding

countryCode

* Impute missing value dengan modus (mungkin “GLOBAL” modusnya, cari tau dulu modus di data training apa )
  + Iya global. Ini dicari manual karena udah telanjur bikin modelnya tanpa fungsi menghitung modus pada train set dan lama lagi kalo harus train ulang. But pretty sure emang hasilnya global karena 82% dari kolom ini bernilai global
* Frequency encoding

userRatingCount & averageUserRating

* Where userRatingCount = 0 🡪 averaeUserRating = 0
* Standarscaler both of them

primaryGenreName

* ~~Mengubah genre sbb~~

~~genre\_mapping = {~~

~~'Travel': 'Travel & Local',~~

~~'Music' : 'Music & Audio',~~

~~'Photography': 'Photo & Video',~~

~~'Reference': 'Books & Reference',~~

~~'Book': 'Books & Reference',~~

~~'Comics' : 'Books & Reference',~~

~~'Social Networking': 'Social',~~

~~'News': 'News & Magazines',~~

~~'Magazines & Newspapers' : 'News & Magazines'~~

~~}~~

* Frequency encoding

downloads

* Mapping nilai yang salah

# Mapping nilai yang salah (akibat konversi tanggal otomatis) ke nilai download asli

download\_fix\_map = {

'01-May': '1 - 5',

'Oct-50': '10 - 50',

'05-Oct': '5 - 10',

'01-Jan': '1 - 10',

'10-Jan': '10 - 100',

'05-Jan': '5 - 10',

'50 - 1': '50 - 100', # ini juga kemungkinan error kebalik, karena normalnya min-max

'500 - 1': '500 - 1000' # idem

}

* Memecah kolom downloads menjadi minDownloads (batas bawah) dan maxDownloads (batas atas)
* Impute missing value pada minDownloads dan maxDownloads dengan median
* Handling outlier: log transform 🡪 winsorizing 🡪 clip nilai minimum 🡪 winsorizing

1. minDownloads
   * standardscaler
2. maxDownloads
   * standardscaler
3. log\_min\_winsor
   * standardscaler
4. log\_max\_winsor
   * standardscaler

deviceType

* One hot encoding

hasPrivacyLink

* Drop (korelasi dengan target < 1%)

hasTermsOfServiceLink & hasTermsOfServiceLinkRating

* Gabungkan keduanya, membentuk kolom “terms\_service\_combined” yang kemudian di ~~one-hot encoding~~  ordinal encoding menjadi “terms\_service\_combined\_has\_link\_low” dan “terms\_service\_combined\_no\_link”

isCorporateEmailScore

* Buat flag untuk missing value (-1) 🡪 nilai pada isCorporateEmailScore: 0, 99, -1
* Standardscaler

1. isCorporateEmailMissing
   * standardscaler

adSpent

* Rekomendasi: drop karena missing value sangat banyak (81%) dan tidak ada kolom lain yang bisa memprediksi adSpent (baik random sampling maupun iterative imputer tidak berguna)
* Opsi lain (not recommended): Impute missing value dengan random sampling

appAge

* Ubah negative value menjadi 0
* Impute missing value dengan median (gunakan sebaran dari train)
* Standardscaler

appContentBrandSafetyRating

* Drop kolom karena:
  + /Cramér's V sangat rendah bahkan pada data yang tersedia (0.0131)
  + missing value sangat tinggi (88%) 🡪 Risiko memperkenalkan noise jika diimputasi

appDescriptionBrandSafetyRating

* ~~one hot encoding~~ ordinal encoding

mfaRating

* drop kolom mfaRating karena nilainya hampir semuanya low 🡪 ga ada variasi 🡪 ga ada korelasi dengan kolom target

coppaRisk

* Handling imbalanced:

Rubrik komentar

|  |
| --- |
| Drop kolom log\_min\_winsor dan log\_max\_winsor pada both train.csv dan test.csv  Pada test.csv, tambahin manually kolom terms\_service\_combined\_has\_link\_nan yang diisi dengan nilai False  Drop kolom “isCorporateEmail\_missing” pada both train.csv dan test.csv  Drop kolom “isCorporateEmailScore” pada both train.csv dan test.csv  Drop kolom “appAge” pada both train.csv dan test.csv |

1. Games
2. Entertainment
3. Music & Audio
4. Music
5. Photography
6. Photo & Video
7. Art & Design
8. Graphics & Design
9. Video Players & Editors
10. Lifestyle
11. Shopping
12. Personalization
13. Education
14. Productivity
15. Books & Reference
16. Reference
17. Book
18. Comics
19. Business
20. Finance
21. Tools
22. Utilities
23. Travel & Local
24. Travel
25. Social
26. Social Networking
27. Communication
28. News & Magazines
29. News
30. Medical
31. Health & Fitness
32. Beauty
33. Food & Drink
34. Sports
35. Auto & Vehicle
36. House & Home
37. Weather
38. Events
39. Dating
40. Stickers
41. Navigation