Homework Solution - Regression

Team:

1. Muhammad Iqbal
2. A Nahda La Roiba
3. Ilham Maulana
4. Clara Natalie S
5. Eka Apriyani
6. R. Rani Indah Salamah
7. Sekar Ayu Larasati
8. Firstandy Edgar Dhafa

Submission:

1. Docs : https://docs.google.com/document/d/1FyfeirgEGmlkf4vE7zQNjhYe_0VFJ5aY/edit
2. PPT: https://docs.google.com/presentation/d/1nb2uOeTjPFwTEnydfEChtREYeKc2SBCr27SWo0DVI2s/edit#slide=id.g23d8b19c754_8_7
3. Notebook: https://colab.research.google.com/drive/1K9oCJKxVhiR0O-vv9TT8LCW7PUgoCt5D#scrollTo=cTRNgga18mR5

Simple Exploratory Data Analysis

DESCRIPTIVE STATISTICS

- Dalam descriptive statistic, kami mencoba memanggil kolom, jumlah row data beserta tipe datanya menggunakan df.info.
- Pada dataset youtube_statistics.xlsx terdapat 36791 baris dan 18 kolom. Dataset tersebut memiliki 4 jenis tipe data, yaitu object, int64, bool, dan datetime64[ns].
- Tipe data pada kolom trending_date dan publish_time harusnya menggunakan tipe data datetime64[ns] karena menunjukkan data tanggal dan data waktu.
- Serta pada kolom description terdapat missing value yaitu sebanyak 45 baris.
- Pada kolom likes, comment_count, dan desc_len nilai rata-rata hitungnya (mean) lebih besar bila dibandingkan dengan nilai kuartil tengah (median), hal tersebut menunjukkan bahwa grafik distribusi frekuensi kolom-kolom tersebut cenderung miring (skew) ke kanan. Sedangkan pada kolom views dan len_title niai mean lebih kecil, yang menunjukkan bahwa grafik distribusi frekuensinya miring (skew) ke kiri.
- Pada kolom comment_count dan desc_len standar deviasinya sangat tinggi, yang menunjukkan bahwa nilainya sangatlah beragam, hal tersebut juga ditandai dengan nilai minimal dan maksimalnya yang berbeda jauh.
- Mayoritas kolom memiliki jumlah unique yang banyak, kecuali kolom dengan tipe boolean yaitu comments_disabled, ratings_disabled, dan video_error_removed

Dari data deskriptif ini, kemungkinan yang bertipe data non numerik berisi kalimat seperti title, channel_title, dan description tidak akan dipakai dalam modelling, sedangkan untuk data object

lain seperti trending_date dan publish_time akan dipertimbangkan untuk dipakai dalam feature extraction.

UNIVARIATE ANALYSIS (NUMERICAL)

Fitur 'views', 'likes', 'dislikes', 'comment_count'

- Grafik boxplot pada fitur 'views', 'likes', 'dislikes', 'comment_count' menunjukkan adanya banyak outlier, yang dapat memengaruhi perhitungan statistik seperti mean dan standar deviasi, serta mempengaruhi interpretasi keseluruhan data.
- Grafik boxplot yang tidak menunjukkan adanya "kotak" atau sebaran data yang kompak di sekitar median dapat menjadi indikator bahwa terdapat variasi ekstrem dalam jumlah views, likes, dislikes, dan comment pada video
- Sisanya memiliki outliers yang tidak begitu banyak sampai mengganggu distribusi data terpusat.

Fitur ‘No_tags’

- Mayoritas video cenderung menggunakan jumlah tag yang relatif sedang hingga cukup banyak. Namun, penting untuk mencatat bahwa terdapat nilai-nilai outliers di atas nilai maksimum, yang berarti ada sejumlah video yang menggunakan jumlah tag yang jauh lebih tinggi dari rata-rata. Distribusi dari data ini menunjukkan kecenderungan untuk mengikuti distribusi normal, tetapi dengan sedikit pergeseran ke kanan (positively skewed).

Fitur ‘desc_len’

- Sebagian besar pembuat video cenderung memilih untuk memberikan deskripsi yang relatif seimbang dan terperinci, tetapi tidak terlalu panjang. Namun, patut dicatat bahwa ada beberapa nilai-nilai outliers yang berada di atas nilai maksimumnya, yang menunjukkan adanya sejumlah video dengan deskripsi yang sangat panjang. Grafik distribusi menunjukkan adanya ekor yang panjang ke kanan, dimulai dari angka 2000 kata ke atas. Pre-processing mungkin termasuk mengatasi nilai-nilai ekstrem atau outlier yang terdapat pada ekor distribusi.

Fitur ‘len_title’

- Sebagian besar judul video memiliki panjang kata yang relatif serupa dalam rentang ini. Nilai maksimum pada 100 kata menunjukkan adanya batasan dalam panjang judul video, mungkin sebagai hasil dari aturan platform atau preferensi konten. Tidak adanya outlier dalam distribusi juga menunjukkan bahwa sebagian besar judul video cenderung memiliki panjang kata yang masuk akal dan tidak ekstrem. Grafik distribusi pada fitur len_title menunjukkan kecenderungan condong ke arah kiri. Namun, menariknya, ekor distribusi mulai terbatas pada rentang 0 hingga sekitar 100 kata. Hal ini bisa menunjukkan adanya pembatasan panjang kata dalam judul video.

Fitur ‘comment_disabled’ dan ‘ratings_disabled’

- Mayoritas video memungkinkan interaksi melalui komentar, dan like & dislike, sementara sebagian kecil video memiliki fitur komentar yang dinonaktifkan.

Fitur ‘video_error_or_removed’

- Sebagian besar video tetap aktif dan dapat diakses, sedangkan sejumlah sangat kecil video mengalami kendala seperti error teknis atau penghapusan dari platform.

Hampir semua kolom di atas mengalami data yang imbalance

MULTIVARIATE ANALYSIS (NUMERICAL)

Berdasarkan korelasi heatmap, kebanyakan korelasi dari fitur tersebut adalah lemah atau sangat lemah (korelasi negatif), yaitu fitur no_tags, desc_len, dan len_title. Sedangkan terdapat 4 fitur yang berkorelasi positif dan kuat (>= 0,5), yaitu views, likes, dislikes, dan comment_count.

Tindak lanjut :

- Fitur positif : Fitur-fitur yang memiliki korelasi positif dan kuat yaitu views, likes, dislikes, dan comment_count tetap dipertahankan untuk dimasukkan untuk pemodelan ML.
- Fitur negatif :
- 'category_id' → Meskipun memiliki korelasi sangat lemah terhadap semua fitur yang lain, namun harus mempertimbangkan untuk memasukkannya ke dalam model, karena kategori video dapat memengaruhi jumlah views.
- 'no_tags' → Fitur harus dipertimbangkan juga, karena jumlah tags yang digunakan dalam video mungkin masih memiliki pengaruh pada popularitasnya.
- Pada 'No_tags', 'desc_len', 'len_title' karena fitur tersebut adalah panjang kata, jadi dapat mengubah fitur-fitur ini menjadi kategori berdasarkan rentang nilai tertentu, misal 'pendek', 'sedang', 'panjang'.

Data Preprocessing

Pada data preprocessing, kami melakukan:

1. Drop missing values

   Terdapat *missing values* pada kolom description yaitu sebanyak 45 baris atau sebesar 0.12% dari total data. Kemudian baris yang terdapat *missing values* tersebut akan kami hapus.

2. Adjust data type

   Pada tahap ini kami mengubah tipe data pada kolom trending_date dan publish_time menjadi tipe datetime64[ns] karena menunjukkan data tanggal dan waktu. Serta pada kolom dengan tipe boolean seperti comments_disabled, ratings_disabled, dan video_error_or_removed kami rubah menjadi tipe data int64.

3. Drop duplicate values

   Pada dataset terdapat data duplicate sejumlah 4228 data atau 11.51% dari total data (36746). Langkah selanjutnya data duplicate tersebut kami hapus.

4. Remove irrelevant column

   Pada tahap ini kami menghapus kolom title, channel_title, tags, dan description. Penghapusan tersebut dikarenakan pada kolom hanya terdapat data teks saja dan bukan termasuk bentuk kategori.

Feature Engineering

Dalam feature engineering, kita melakukan:

1. Feature Extraction
    - Mengekstrak trending_date dan publish_date untuk mendapatkan tanggal.
    - Menambah feature day_difference dari hasil pengurangan trending_date dikurangi publish_date.
    - Menambah feature day_part dari publish_time:hour
    - Menambah feature publish_isweekend dari feature publish_date
2. Drop Features, mendrop feature yang tidak digunakan pada model.
3. Remove Outliers, meremove data yang termasuk dalam outlier.
4. Normalization, melakukan normalisasi dengan menggunakan metode log transformation dan Z-Score agar rentang nilai tiap variabel dan skala data menjadi seragam, data terdistribusi normal dengan nilai mean 0 dan std 1.

Modeling

Modelling (membandingkan 6 algoritma)

- Evaluate Model: linear

RMSE (test): 0.32
RMSE (train): 0.32
r2 (test): 0.81
r2 (train): 0.81
r2 (cross-val test): 0.80
r2 (cross-val train): 0.81

- Evaluate Model: ridge

RMSE (test): 0.32
RMSE (train): 0.32
r2 (test): 0.81
r2 (train): 0.81
r2 (cross-val test): 0.80
r2 (cross-val train): 0.81

- Evaluate Model: lasso

RMSE (test): 0.73
RMSE (train): 0.73
r2 (test): 0.045
r2 (train): 0.045
r2 (cross-val test): 0.002
r2 (cross-val train): 0.045

- Evaluate Model: xgb

RMSE (test): 0.22

RMSE (train): 0.17
r2 (test): 0.90
r2 (train): 0.94
r2 (cross-val test): 0.86
r2 (cross-val train): 0.94

- Evaluate Model: rf

RMSE (test): 0.241
RMSE (train): 0.0903
r2 (test): 0.89
r2 (train): 0.98
r2 (cross-val test): 0.86
r2 (cross-val train): 0.98

- Evaluate Model: ANN

RMSE (test): 0.32
RMSE (train): 0.20
r2 (test): 0.89
r2 (train): 0.95
r2 (cross-val test): 0.92
r2 (cross-val train): 0.94

Dari hasil di atas :

- Nilai hasil **cross-validation (CV)** menjadi yang lebih kami pertimbangkan, karena itu mencerminkan keadaan performa model yang lebih **nyata** ketimbang tanpa CV.
- nilai **RMSE** untuk ***RF*** dan ***XGB*** relatif **paling kecil**, mengindikasikan nilai error juga lebih kecil ketimbang algoritma lainnya.
- Nilai **R2** untuk ***linear*** dan ***ridge*** **lebih fit** karena jarak nilai train dan test nya relatif sama, namun algoritma ***RF*** dan ***XGB*** memiliki **akurasi** yang relatif **lebih besar** untuk train dan test nya.

Dalam tahap ini, kami akhirnya memutuskan untuk memilih RF dan XGboost untuk selanjutnya dilakukan tuning, karena memberikan performa R2 yang cukup baik ketimbang algoritma lainnya dan layak disandingkan dengan algoritma ANN yang memiliki nilai paling fit dan performa yang paling tinggi.

- Tuned Random Forest:

r2 (cross-val test): 0.8583153646874264
r2 (cross-val train): 0.9459570483809715

- Tuned Xgboost

r2 (cross-val test): 0.8732655106561074
r2 (cross-val train): 0.9819369424406489

Secara umum, hasil tuning dari XGB dan RF tidak memberikan dampak signifikan pada performa R2. ANN tetap memiliki R^2 yang paling tinggi pada data pengujian (cross-val test), yaitu 0.926. Rentang nilai R^2 yang berdekatan antara cross-val test dan cross-val train pada ANN mengonfirmasi efisiensi model ANN dibandingkan dengan Tuned XGBoost Regressor dan Tuned Random Forest.

Summary

- Kesimpulannya, bahwa algoritma ANN menjadi algoritma dengan performa yang terbaik untuk memprediksi jumlah views, berdasarkan tingkat error yang kecil, nilai performa yang besar dan model yang bisa menggeneralisasi setiap faktor yang mewakili.
- Rekomendasi bisnis yang bisa dilakukan adalah mempertimbangkan faktor-faktor seperti waktu upload video untuk marketing produk seperti waktu peak time (menyesuaikan waktu di setiap negara) saat weekdays. Karena jam diluar peak time saat weekday kurang begitu efektif jadi tidak disarankan upload video di waktu ini. Namun untuk weekend, tidak ada batasan khusus karena secara umum banyak orang menonton youtube kapan saja saat weekend khususnya video yang baru dipublish.

Appendix

1. EDA : Laras, Eka, Ilham
2. Prepro : Rani, Dhafa
3. Modelling : La Roiba, Clara, Iqbal
4. Docs : Together

Kesulitan :
- modelling
- memilih parameter yang tepat untuk tuning
- menunggu tuning selesai
- feature extraction dan selection