PD3結果報告

4EP4-35 鶴瀬 和輝

研究テーマ

YouTube動画の再生数を予測する

背景

- YouTubeは、2019年広告売上高150億ドル、月間アクセス20億人に達するほど、世界で最も普及した動画マーケティングツール

- 再生数を増やすための施策が、日々多くのメディアでノウハウが紹介されている

(動画広告、セミナー、記事)

背景と課題

- 再生数を増やすための施策が、日々多くのメディアでノウハウが紹介されている

各取り組みによりどれだけ視聴回数を伸ばすことが出来るのか、定量データに基づいたノウハウ紹介は少なく、何が効果的か分かっていない

ゴール

YouTuber(動画クリエイター)にとって生命線とされる再生数を伸ばすために、どのような指標を心がける必要があるのかを知り、動画の再生数を予測できるモデルを構築する。

アプローチ・方法

カラム名	説明
Video_id	動画ごとの一意のID
title	動画のタイトル
publishedAt	動画の投稿時間
channelId	チャンネルのID
channelTitle	チャンネルの名前
categoryId	動画カテゴリのID
Collection_date	データレコードの収集日
tags	動画に割り当てられたタグ
likes	高評価数
dislikes	低評価数
Comment_count	コメント数
Thumbnail_link	サムネイルのリンクURL
Comments_disabled	コメントが許可されていない
Rating_disabled	評価が許可されていない
description	動画の説明文
y	再生数

データ概要

- YouTube APIで取得できるメタデータ

- 訓練データ
  レコード数：19720
  列数：16
- テストデータ
  レコード数：29582
  列数：15

アプローチ･方法

- LIghtGBMを使用し、再生数を予測するのに効果的な特徴量をモデルの入力値とした予測モデルを作成し、そのモデルにおける特徴量の重要度を算出する。

具体的にやったこと

- EDA(探索型データ解析)を行い、再生数と相関関係がある特徴量を見つける。

EDA：探索型データ解析は、機械学習タスクの一番最初のフェーズで、データを視覚化する、データのパターンを知ることで、特徴量やターゲットの関係性/相関性を知ること。

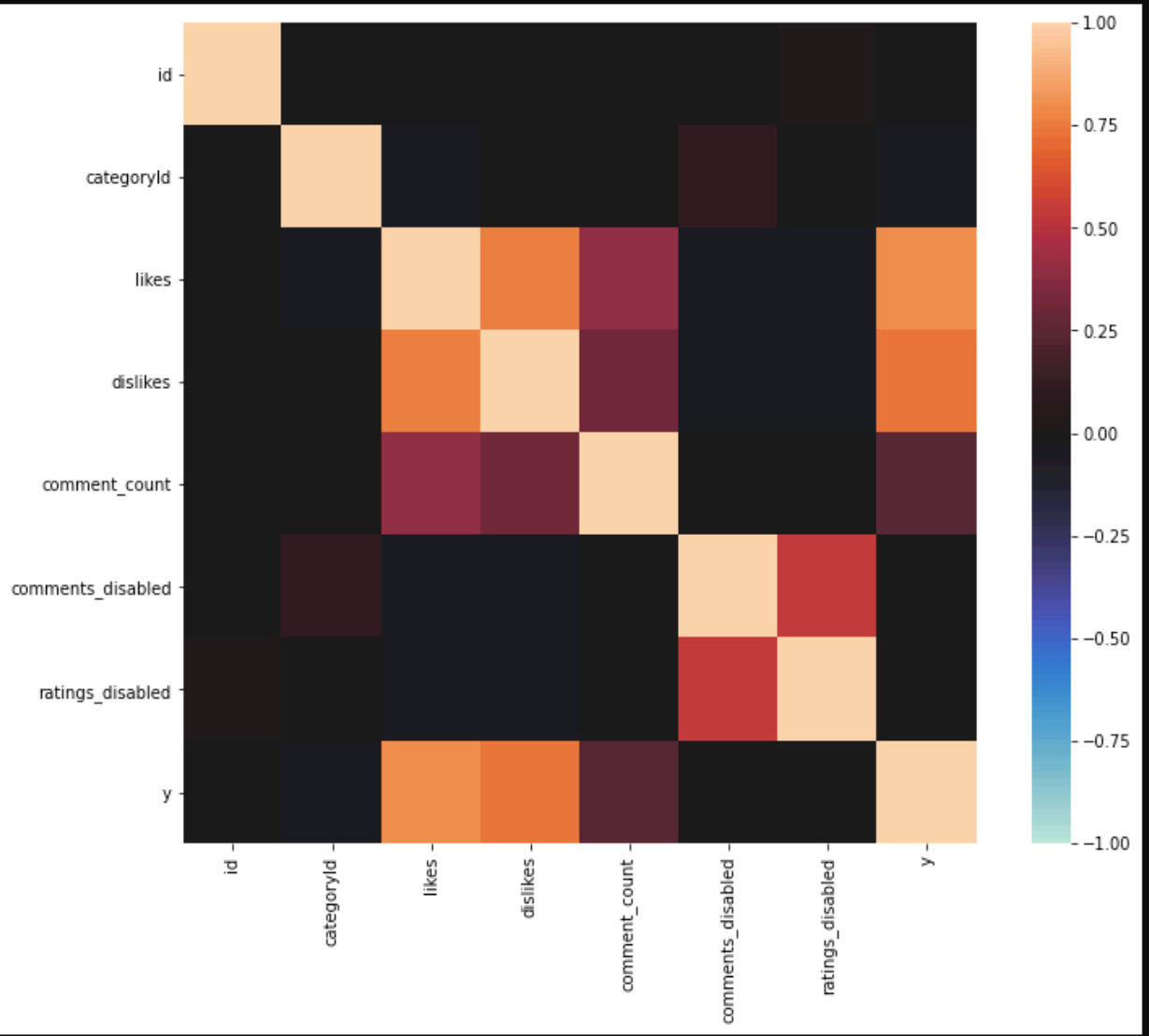
各特徴量の相関関係を表した図

- Pythonの外部ライブラリのPandas、Seabornを使用し、各特徴量の相関関係をヒートマップ図で表示。

- 図より、再生数と相関のある特徴量はLikes、Dislikes、Comment_countであることが分かる。

```
[6]: 1 fig, ax = plt.subplots(figsize=(12, 9))
     2 sns.heatmap(train.corr(), square=True, vmax=1, vmin=-1, center=0)

[6]: <AxesSubplot:>
```



具体的にやったこと

- 効果的な特徴量(Likes、 Dislikes、 Comment_count)
  をLightGBMを使用し、上記の特徴量の予測値を
  出力値とする、それぞれ3つの予測モデルを作成。

予測値を作成する目的として、上記の特徴量には評価、
コメントを行うことが許可されていないデータもあるた
め、そのデータが再生数を予測するモデルの精度に影響
を及ぼすと考えたから。

評価、コメントができないデータの図

- ratings_disabledとcomment_disabledの値がTrueになっているLikes、 Dislikes、 Comment_countを抽出。

- これらの値は0となっており全体の13%にも及ぶ。

[32]:
```
1 d[((d["comments_disabled"] == True) & (d["ratings_disabled"] == True))]
```

[32]:

	likes	dislikes	comment_count	ratings_disabled	comments_disabled
83	0	0	0	True	True
219	0	0	0	True	True
237	0	0	0	True	True
239	0	0	0	True	True
258	0	0	0	True	True
...	...	...	...	...	...
49130	0	0	0	True	True
49133	0	0	0	True	True
49231	0	0	0	True	True
49232	0	0	0	True	True
49272	0	0	0	True	True

2392 rows × 5 columns

効果的な特徴量の予測方法

- LightGBMを使用しテキストデータ(ChannelTitle、descriptionなど)のSVD特徴量、カテゴリIDなどの集計特徴量をモデルの入力とした予測モデルを作成する。

SVD特徴量：今回はTF-IDFで作成したテキスデータの大量の特徴量をPythonの外部ライブラリであるscikit-learnのTrucecatedSVDを使用して次元削除した後の特徴量のこと。

集計特徴量：今回は、カテゴリIDやタグなどのグループごとのコメント数・高評価数・低評価数をそれぞれのコメント数や低評価数で計算した特徴量のこと。

- TF-IDFを使い、channelTitle特徴量に含まれる単語の重要度を数値化した大量の特徴量を生成した図

	00	007004ma1	0094592	009eel	009gabry007	00motivation	00r	00rs00	00throne	01	0120baseball
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...
49297	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
49298	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
49299	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
49300	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
49301	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

49302 rows × 28431 columns

- TrucetedSVDを使用し、大量の特徴量(次元)を20次元に圧縮した図

[10]:

	0	1	2	3	4	5	
0	-1.74256423e-18	-3.81323150e-18	-1.25109462e-18	4.44693285e-19	9.95385054e-19	-1.53437903e-18	1.71560
1	3.23333834e-09	4.28460738e-03	2.76656988e-01	1.37113123e-15	-4.80227371e-04	7.73523344e-07	-3.30685
2	-2.37528385e-18	-5.16734484e-18	-1.68296775e-18	6.56619586e-19	1.47375708e-18	-2.26345088e-18	2.65552
3	-2.00887545e-17	-4.02019822e-17	3.10331706e-18	-5.75261794e-19	4.76878313e-18	-4.48477603e-18	-5.22152
4	1.61881162e-18	3.02018759e-18	-2.64154474e-18	1.79683778e-18	1.05699816e-18	3.45875622e-19	3.65413
...	...	...	...	...	...	...	
49297	6.54758833e-15	2.95936599e-08	4.12667072e-06	-6.06840880e-16	-1.45260832e-08	1.55817180e-10	5.75839
49298	-4.39955024e-19	-9.73885618e-19	-3.21430330e-19	1.01069120e-19	2.64775830e-19	-4.05584066e-19	4.661421
49299	5.79496417e-11	5.45360901e-05	6.82882556e-05	3.05796241e-18	-2.97383744e-07	1.30490186e-07	-1.25462
49300	9.80235266e-10	1.56082380e-03	8.15688624e-04	1.41238253e-14	6.24838253e-01	-2.26123277e-06	-5.34938
49301	4.12311341e-16	8.22855026e-16	-1.04415797e-16	-2.72346025e-17	2.26174670e-16	-1.01360506e-16	-2.47248

49302 rows × 20 columns

提案手法

- （高評価数、低評価数、コメント数）の集計特徴量、　一定の期間ごとの（高評価数、低評価数、コメント数）特徴量などをモデルの入力とした、それぞれ3つの再生数の予測モデルを作成する。
- 上記の特徴量全てをモデルの入力とした再生数の予測モデルを作成する。
- それぞれのモデルのFeature Importanceを算出し、再生数を予測するために効果的な特徴量の類似点、指標を見つける。
- データの検証方法としては交差検証法、モデルの評価方法はRMSE(平均二乗偏差)を採用。

Feature Importanceについて

- Feature Importanceとは？

特徴量の重要度のことであり、その特徴量の分類がターゲットの分類にどれぐらい寄与しているかを測る指標である。
重要度の計算はジニ不純度をもとにして計算ができる。

- LightGBMやXGBoostなどの勾配ブースティングフレームワークには、モデルから特徴量の重要度を算出できるfeature_importanceメソッドがあります。

- 本研究では、LIghtGBMを使用しています。

ジニ不純度の計算について

•ジニ不純度の定義
G(k)：あるノードkにおける不純度
n：ターゲットラベルの数
p(i)：あるノードkにおけるターゲットラベルiの頻度

$$G(k) = \sum_{i=1}^{n} p(i) \times (1 - p(i))$$

あるノードにおいて完全にサンプルが分類されている場合は、
ジニ不純度の値は0になる。分類されていなければ値は1に近づく。

ジニ不純度を利用した重要度の計算

- ある特徴量jにおける重要度の定義
  出力された 数値が高ければ重要度が高い。

$$I(j) = \sum_{i=1}^{n \epsilon F(j)} (N_{parent(i)} \times G_{params(i)}) - (N_{leftchild(i)} \times G_{leftchild(i)} + N_{rightchild(i)} \times G_{rightchild(i)})$$

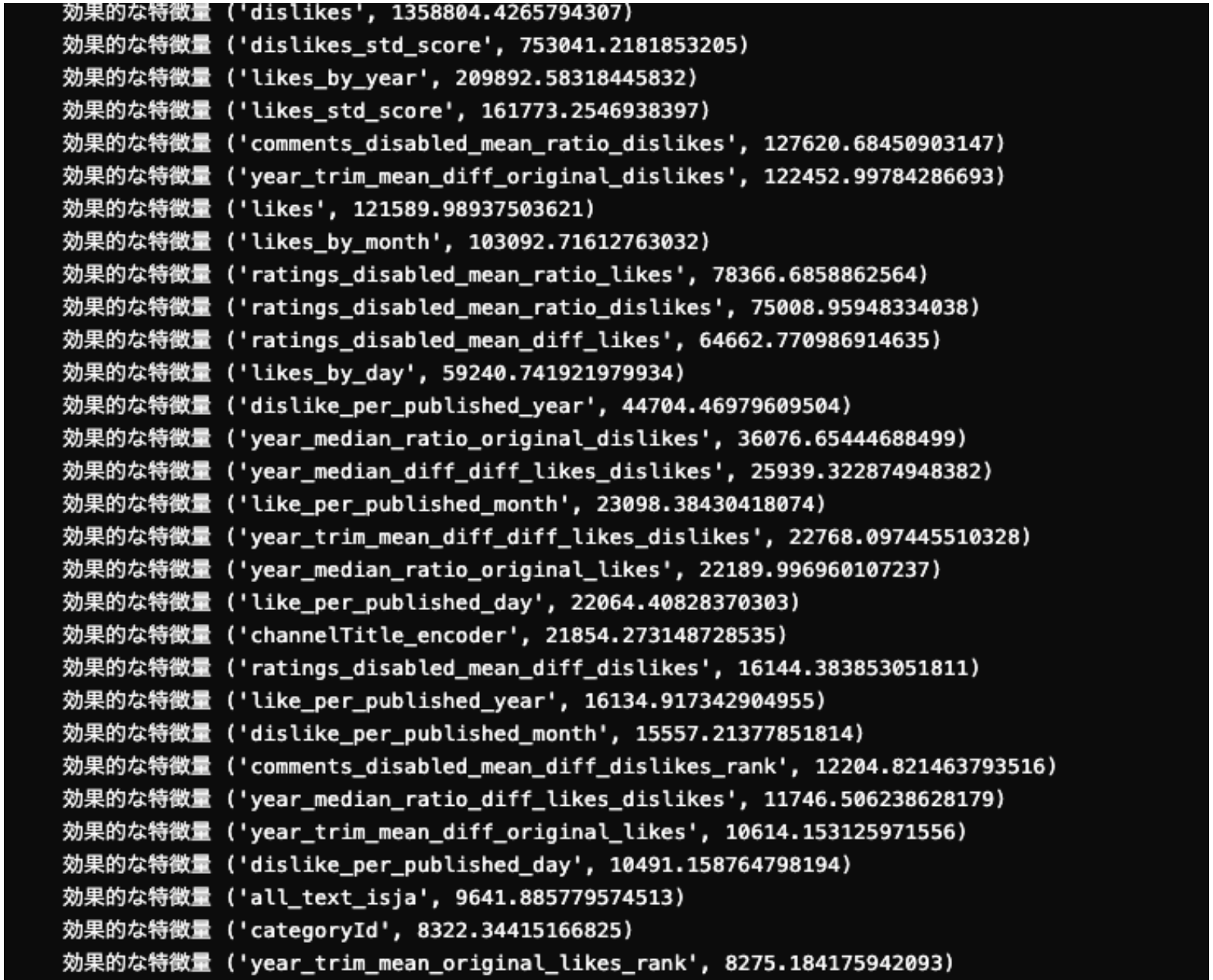
I(j)：ある特徴量jにおける重要度
F(j)：ある特徴量jが分割対象となるノードの集合
$N_{parent(i)}$：あるノードiにおけるサンプル数
$N_{left\_child(i)}$：あるノードiの子ノードのうち、左側のノードのサンプル数
$N_{right\_child(i)}$：あるノードの子ノードのうち、右側のノードのサンプル数
$G_{params(i)}$：あるノードiにおけるジニ不純度
$G_{left\_child(i)}$：あるノードiの子ノードのうち、左側のノードにおけるジニ不純度
$G_{right\_child(i)}$：あるノードiの子ノードのうち、右側のノードにおけるジニ不純度

算出した特徴量の重要度の図

- Likes,Dislikes,Comment_countの集計特徴量が重要度として高いことが分かる。

- publish_day,publish_year,by_yearなどの期間ごとの評価数、コメント数も特徴量の重要度として高いことが分かる。

- RMSEの値は0.718…となった。



まとめ

- YouTube動画の再生数は、高評価やコメントなどのエンゲージメントに影響する。

- エンゲージメント全体における低評価の割合も動画の再生数に影響する。

- 出来たこと
  再生数を予測するための予測モデルの構築。

- 出来てないこと
  決定木の可視化、指標の選定。

カラム名	説明
video_id	動画ごとの一意のID
title	動画のタイトル
publishedAt	動画の投稿時間
channelId	チャンネルのID
channelTitle	チャンネルの名前
categoryId	動画カテゴリのID
collection_date	データレコードの収集日
tags	動画に割り当てられたタグ
likes	高評価数
dislikes	低評価数
comment_count	コメント数
thumbnail_link	サムネイルのリンクURL
comments_disabled	コメントが許可されていない
rating_disabled	評価が許可されていない
description	動画の説明文
y	再生数