PD3中間報告会

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。

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
describe	動画の説明文
y	再生数

アプローチ・方法

使用するデータ：
Probspace YouTube動画視聴回数予測
コンペティションのデータ

レコード数：19720
列数：17

YouTube動画の再生数を予測する

アプローチ・方法

EDA(探索型データ解析)を行い、再生数を予測するのに効果的な特徴量を見つける。

以下の特徴量を用意

- dislikes(低評価数)、likes(高評価数)comment_count(コメント数)の四則演算等,log,二乗などのaggregation(集計)特徴量
- dislikes、likes、comment_countの予測数
- テキスト(channelTitle,CategoryIdなど)のaggregation特徴量
- テキストのtf-idf -> svd、doc2vec、tf-idf+t-sne
- Target Encoding
- テキストの中にkeywordあり、なしのbinary特徴量

アプローチ・方法

評価やコメントをすることが許可されていないものに関しては値が０。

[19]:
```
1 train[['likes', 'dislikes', 'comment_count', 'y']][(train['comments_disabled'] == True) & (train['ratings_disabled'] == True)]
```

[19]:

	likes	dislikes	comment_count	y
83	0	0	0	15056
219	0	0	0	658
237	0	0	0	29185
239	0	0	0	8013
258	0	0	0	5259
...	...	...	...	...
19634	0	0	0	6377
19651	0	0	0	90468
19657	0	0	0	4027999
19699	0	0	0	90541
19701	0	0	0	461454

960 rows × 4 columns

アプローチ・方法

・モデリングでは、LightGBMという勾配ブースティング
フレームワークを用いる

なぜLightGBMを使うのか？
・カテゴリ変数に対して特別な処理を自動的に実行してくれる
ので、One-Hotエンコーディングの手間を無くせる
・既存のデータセットを極力加工せずに利用するという観点で、
特徴量エンジニアリングの負担を軽減してくれる

進捗状況

特徴量について

〇・データ収集。Probspace YouTube動画視聴回数予測コンペ

〇・収集したデータをEDA(探索型データ解析)を行う

〇・特徴量エンジニアリング

- numerical data(dislikes, likes, comment_count)のaggregation特徴量を作成
- text dataはtf-idf特徴量をsvd特徴量に変換、そしてaggregation特徴量を作成
- Target Encoding

モデルについて

×・シングルモデルでは、LightGBM、XGBoost、CatBoostを試す

×・各々のモデルのパラメータ調整

今後の進め方について

- BERTを使ってテキストデータ(title, description, chanmelTitle, tags)の特徴徴出
- 進捗状況でできていなかったモデリングを行う