|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| テーマ番号 | 日程番号を選択テーマ番号を入力 | |  | | |
| プロジェクト テーマ | 和文 | YouTube動画の再生数予測モデルに効果的な指標の解析 | | 指導教員 | 元木 光雄 准教授 |
|  | 英文 | Analysis of Effective Metrics for YouTube Video Views Prediction Model | |  |  |
| プロジェクト メンバー | | 4EP4-35　鶴瀬 和輝 (Kazuki Tsuruse) | | | |

YouTubers are able to make a living by posting videos on YouTube and earning advertising revenue. Although there are many measures and know-hows in the media to increase video views, we don't know what kind of metrics are best for increasing views.  
In this study, we analyze effective indicators for a model to predict the number of YouTube video views as a way to increase the number of YouTube video views.

Prediction, YouTube, LightGBM, Feature importance, YouTuber

# まえがき

近年，世界的なスマートフォンの普及と共に，様々な動画マーケティングツールが世界に広まった．日本における代表的な動画マーケティングツールの一つにYouTubeがある，アルファベット社によると2019年YouTubeは広告売上高150億ドル(約1兆6000億円)で，アルファベットの年間総売上高の約9%を占めており，月間アクセスが20億人ほどであり，他社のサービスと比べても人気があることが分かる[1]．

最近ではYouTubeに動画を投稿し，動画の広告収益を得て生活を担うYouTuberと呼ばれる職業も存在する．この広告収益の値は，動画の再生数や評価の割合によって変化する．よって再生数や高評価の割合が多いほど，収入も増えていくことが分かる．これらの影響から動画の再生数を増やすための施策が，日々多くのメディア(動画広告・セミナー・記事)でノウハウが紹介されている．しかし実際にそれらの取り組みによってどれだけ試聴回数を増やすことができるのか，定量データに基づいたノウハウ紹介は少なく，何が効果的なのかは分かっていない状況である．そこでYouTube動画の再生数を予測し，予測に効果的な指標を解析することによってYouTube動画の再生数を増やすことができる指標を抽出できると考えた．

本研究では，YouTube動画の再生数を増やすための方法として，YouTube動画の再生数予測モデルに効果的な指標の解析を行う．

# 準備

ProbSpaceと呼ばれるデータ分析のプラットフォームに「YouTube動画視聴回数予測コンペティション」が2020年4月〜2020年6月の期間に開催され，450ユーザーが参加した．このコンペティションは，YouTube APIから取得できるメタデータを入力として，動画の視聴回数を予測するアルゴリズムを開発するというコンペティションである．

メタデータの中には，動画コンテンツの質的指標となる高評価数・低評価数・コメント数や，SEOとして重要とされているタイトル名・説明文・タグ・投稿時間，といった情報が含まれている．訓練データのレコード数は19720，カラム数は16．テストデータのレコード数は29582，カラム数は，視聴回数を除いた15である．

本研究では，訓練データとテストデータを結合したデータセットをデータの前処理に使用し，前処理されたデータセットから訓練データのレコード数までを抽出したデータセットをモデルの訓練に使用する．また本予稿では，それぞれのデータの名称を前処理データ，モデル訓練データとする．

# 基本的な解析手法

## 欠損データの補完と特徴量の作成

　まず，本研究で使用する前処理データから探索型データ解析を行い，再生数と相関関係の高い特徴量は，高評価数・低評価数・コメント数であることが分かった．

　次に，前処理データの中で高評価数・低評価数・コメント数の欠損値が全体の13%存在しており，この13%の部分が予測モデルの精度に影響したので，欠損している箇所をLightGBMを使用し，高評価数・低評価数・コメント数それぞれの予測モデルを構築し，高評価数・低評価数・コメント数の予測値で欠損部分の補完を行った．

　特徴量の作成では，コンペティションでも多くの参加者が行っていたカテゴリIDや動画を公開した年ごとの高評価数・低評価数・コメント数の集計特徴量を作成した．集計方法は，トリム平均値，最小値，最大値，中央値などである．また，期間ごと(１日あたり・１ヶ月あたり・一年あたり)の高評価数・低評価数・コメント数の特徴量，高評価数・低評価数・コメント数の実際値と予測値の差分などの特徴量を作成した．またチャンネル名・動画タイトル・説明文などのテキスト特徴量をベクトル化するために、てテキスト特徴呂に対してMeCabによる形態素解析を行った．次にTF-IDF手法により、一万次元を超える単語の重要度の特徴量を作成．最後にLSA(潜在的意味解析)を行い，作成した単語の重要度の特徴量を次元圧縮した．

## LightGBMによる予測モデルの構築

まずモデルの訓練(学習)を行うためにモデル訓練データを用意する．そのために3.1で作成した特徴量を前処理データに追加した後，前処理データをモデル訓練データに変換する．次に，予測モデルの精度にあまり影響しない特徴量をモデル訓練データから削除する．具体的には，次元圧縮後の説明文の特徴量，動画タイトルやチャンネル名などのテキスト特徴量である．最後に，4.1で述べる交差検証法による予測モデルの汎化性能を評価するために，モデル訓練データをモデルを作る訓練データと，モデルを評価するテストデータに分割する．この訓練データとテストデータをLightGBMのモデルの入力値とし，モデル訓練を行うことによって予測モデルの構築ができる．

# 3.3 特徴量の重要度の抽出

　まず3.2で構築した予測モデルから再生数を予測するのに効果的な特徴量の重要度を抽出する．特徴量の重要度とは，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可能であり，ジニ不純度IG(t)を求める計算式を式(1)に示す．

式(1)のIG(t)のtは決定木の任意のノード，cはカテゴリ数，p(i|t) は当該ノードの総数に対するあるカテゴリに属するデータの割合を示す．ジニ不純度の値が0に近いほど分類されており，分類されていなければ値は1に近づく．このジニ不純度を利用した特徴量の重要度I(j)の計算式を式(2)に示す．

I(j)はある特徴量jにおける重要度，F(j)はある特徴量jが分割対象となるノードの集合，Nparent(i)はあるノードiにおけるサンプル数，Gparams(i)はあるノードiにおけるジニ不純度，Nleft\_child(i)・Nright\_child(i)はあるノードiの子ノードのうち，左右のノードにおけるサンプル数，Gleft\_child(i)・Gright\_child(i)はあるノードiの子ノードのうち，左右のノードにおけるジニ不純度である．式(2)から算出される特徴量の重要度の数値が高ければ高いほど，重要度として適切である．本研究では，この特徴量の重要度の計算をLightGBMのfeature\_importanceメソッドによって行い，正規化した特徴量の重要度を抽出する．

# 解析結果の評価

## 評価方法

機械学習を行う際，訓練データでの予測モデルの性能が良いにも関わらず，テストデータでの予測モデルの性能が悪くなってしまう過学習が起こることがある．よって予測モデルの汎化性能を評価するために交差検証法のK-分割交差検証を用いた．K-分割交差検証とは，データをK個に分割してそのうちの一つをテストデータにし，残りのデータを訓練データとして正解率の評価を行う．これをK個のデータすべてが1回ずつテストデータになるようにK回学習を行い，予測モデルの精度の平均をとる手法である．

また，予測モデルの評価指標は，回帰モデルの最も一般的な性能評価指標であるRMSE(二乗平均平方根誤差)とする．RMSEとは，実際値と予測値の誤差の二乗の総和の平均値のルートをとったものである．よってRMSEの値が0に近いほどモデルとしての性能が良いということになる．RMSEの計算式を式(3)に示す．

nはデータの個数，yiは実際値，yˆiは予測値である．

## 評価結果

本研究では，4.1で述べたK-分割交差検証のKの値を8とし，LightGBMによるモデル訓練を8回実施した．それぞれの予測モデルのRMSEを表1に示す．また，モデル訓練を8回実施した中で精度が一番良い時(K=2)の上位10個の特徴量の重要度を表2に示す．

1. 再生数予測モデルのRMSE
2. RMSE of the views prediction model

|  |  |
| --- | --- |
| K回 | RMSE |
| 1 | 0.7555 |
| 2 | 0.6770 |
| 3 | 0.7426 |
| 4 | 0.7178 |
| 5 | 0.7584 |
| 6 | 0.7378 |
| 7 | 0.6901 |
| 8 | 0.7413 |

1. K=2の上位10個の特徴量の重要度
2. Importance of the top 10 features of K=2

|  |  |
| --- | --- |
| 特徴量名 | 特徴量の重要度 |
| year\_min\_diff\_dislikes | 0.2293 |
| dislikes | 0.1337 |
| categoryId\_min\_diff\_dislikes | 0.0979 |
| likes\_by\_year | 0.0547 |
| year\_trim\_mean\_diff\_dislikes | 0.0522 |
| year\_min\_diff\_dislikes | 0.0407 |
| likes\_by\_month | 0.0229 |
| likes\_by\_day | 0.0145 |
| year\_trim\_mean\_diff\_likes | 0.0141 |
| ratings\_disabled\_mean\_ratio\_likes | 0.0127 |

表1より，構築した全ての予測モデルのRMSE値が0.7前後となっており，今回の研究で構築した予測モデルはモデルとしての性能があまり高くないという結果となった．

表2からはlikes\_by\_year・likes\_by\_monthなどの期間ごとの評価数やyear\_trim\_mean\_diff\_dislikesなどのトリム平均値を使った集計特徴量が高く，その中でも低評価に関する特徴量が重要度として効果があることが分かった．

# むすび

本研究では，YouTube APIのメタデータからLightGBMによる再生数を予測するモデルを構築することで，予測モデルから特徴量の重要度を抽出し，再生数を予測するのに効果的な指標を見つけることができた．しかし，モデルとしての性能は高くはないので，今後の課題として，LightGBMのモデルに入力する特徴量を高評価数，低評価数，コメント数などのエンゲージメントに関わる特徴量に厳選する，画像データを利用する．

1. ペイジ・レスキン，ついにYouTubeの広告売上高が明らかに・・・2019年は前年比36%増の150億ドル，

https://www.businessinsider.jp/post-207025，2020年11月19日参照．