|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| テーマ番号 | 1EP15 | |  | | |
| 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) | | | |

YouTube advertising revenue increases as the percentage of views and high ratings increases. Various ideas for improving the number of views and high ratings have been proposed in literature, but there is little quantitative analysis of what is effective. Therefore, our goal is to "identify what metrics are effective in predicting the number of views for YouTube videos by analyzing them. The validation method is to build a model for predicting the number of views using the gradient boosting framework and calculate the importance of features.

Prediction, YouTube, LightGBM, Feature importance, Machine Learning

# まえがき

近年，世界的なスマートフォンの普及とモバイル通信の高速化に伴い，様々な動画サービスが世界に広まった．代表的な動画サービスの一つにYouTubeがある．

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と呼ばれる職業が存在する．この広告収益は，動画の再生数や評価の割合によって変化し，再生数や高評価の割合が多いほど，収入も増えていく．これらの影響から，動画の再生数を増やすためのアイデアが，日々多くのメディア（動画広告，セミナー，記事）で紹介されている．

しかし，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．

そこで，本研究では，「YouTube動画の再生数予測に効果的な指標を解析することによって，再生数を増やすことができる指標を抽出できる」という仮説を立て，その検証を行う．

# 使用するデータ

データ分析のプラットフォームであるProbSpaceにおいて「YouTube動画視聴回数コンペティション」が2020年4月から2020年6月の期間に開催された．本コンペでは，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（再生数）を予測するアルゴリズムを開発する．提供されるデータには，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，視聴回数や，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．訓練データのデータ数は19,720である．テストデータのデータ数は29,582であり視聴回数を除いてある．本研究では，このコンペティションで提供されたデータを用いる．

# 基本的な解析手法

機械学習手法の一つである決定木の勾配ブースティングフレームワークにより，学習器を作成したのち，算出される重要度が高い特徴量を再生数予測モデルに効果的な指標として考える．

本節では，欠損データの補完と特徴量の作成，LightGBMによる予測モデルの構築，特徴量ごとの重要度を抽出する方法について述べる．

## 欠損データの補完

入力データには，高評価数，低評価数，コメント数の欠損値が全体の13%存在している．この部分が予測モデルの精度に影響していると考えられる．よって，欠損している箇所に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

まず，数値で表現される指標に対する特徴量の作成について述べる．特徴量はいくつかの指標を集計することで作成する．具体的に用いる指標は，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，コメント数であり，集計方法にはトリム平均値，最小値，最大値，中央値などを用いる．また，期間ごと（１日あたり，１ヶ月あたり，1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．

次に，テキストで表現される指標に対する特徴量の作成について述べる．まず，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，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

3.2で作成した特徴量，訓練データをLightGBMの入力データとして，再生数予測モデルを構築する．具体的に入力データ数は19,720，特徴量が616種類である．

しかし，入力データだけでは，モデルの訓練を行うのに不十分なので，LightGBMのハイパーパラメータを調整する．具体的には，モデルの訓練実行前にハイパーパラメータを調整することで，モデルの性能向上や過学習の抑制，効率の良い学習などが設定できる．

## 特徴量ごとの重要度の抽出

3.3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り，ジニ不純度を基にして計算できる．決定木の任意のノードのジニ不純度*IG*(*t*)は，  
で求められる．ただし，はカテゴリ数を表し， は決定木のノード総数に対して，カテゴリに属するデータの割合を表す[1]．ジニ不純度は0から1をとり，小さいほど対応する特徴量が重要となる．このジニ不純度を利用した特徴量の重要度を算出するまでの計算式を以下に示す[2]．

は，各ノードにおけるジニ不純度の減少を計算しており，は，あるノードiにおけるサンプル数，，は，あるノードの子ノードのうちの左右のノードにおけるサンプル数，はあるノードにおけるジニ不純度，，は，あるノードの子ノードのうち，左右のノードにおけるジニ不純度を表す．は，を利用し，分母が全ノードにおける不純度の減少の総和であり，分子が対象の特徴量による分岐ノードでのジニ不純度の減少の総和から求められる．

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，正規化した特徴量ごとの重要度を抽出する．

# 実データへの適用

本節では，実データを用いて，予測モデルを評価する方法，評価結果，特徴量の重要度について述べる．

## 予測モデルの評価

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．そこで，予測モデルの汎化性能を実験的に評価する

### 評価方法

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値である．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．はデータの個数，は実際値， は予測値を表すRMSEの計算式を以下の式に示す[3]．

### 評価結果

本研究では，4.1.1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を図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1に示す．

1. 予測モデルのRMSE
2. RMSE of predictive model
3. K=2における上位10個の特徴量の重要度
4. The importance of the top 10 features at 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 |

図1より，構築した全ての予測モデルのRMSE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．

## 特徴量の重要度

表1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．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．そして，画像データなどの利用がある．

1. 近藤悠介，教育データ分析入門2 （第4回），https://www.f.waseda.jp/yusukekondo/IEDA19/IEDA2\_05.html，2020年12月10日参照．
2. mlcourse.ai，Topic5．Ensembles and random forest Part3．Feature Importance，  
   https://mlcourse.ai/articles/topic5-part3-feature-importance/，2020年12月18日参照．
3. biostatistics，回帰分析の評価指標，https://stats.biopapyrus.jp/glm/lm-evaluation.html，2020年12月20日参照．