テーマ番号	1EP35			
プロジェクトテーマ	和文	YouTube 動画の予測再生数･予測高評価数を取得するシステムの実装	指導教員	元木 光雄 准教授
	英文	Implementation of a system to obtain the predicted number of views and the highest rated number of YouTube videos		
プロジェクトメンバー	4EP4-35　鶴瀬和輝 (Kazuki Tsuruse)			

Abstract　近年,YouTube という動画配信サービスが話題になっており,YouTuber という動画の広告収入を基にして生活を担う職業が誕生している.この動画の広告収入は YouTuber によって,様々であり動画の再生数が高い・高評価が高く取得できる YouTuber の動画が広告収入の単価が高くなることが分かっている.よって YouTuber にとって動画の再生数や高評価数は大切であるため,動画を作成する前に,自身の作成した動画がどの程度の規模になるかを知ることができれば,効率よく再生数,高評価を高く取得できる動画の作成,改善に繋がると思われる.

Keywords　YouTube, predict,One-HotEncoding,LinearRegression,Video

1. テーマの導入

近年,YouTube という動画配信サービスが話題になっており,YouTuber という動画の広告収入を基にして生活を担う職業が誕生している.この動画の広告収入はYouTuberによって,様々であり動画の再生数が高い・高評価が高く取得できる YouTuber の動画が広告収入の単価が高くなることが分かっている.よってYouTuberにとって動画の再生数や高評価数は大切であるため,動画を作成する前に,自身の作成した動画がどの程度の規模になるかを知ることができれば,効率よく再生数,高評価を高く取得できる動画の作成,改善に繋がると思われる.

私がこのテーマを扱う理由として挙げることは,自分も普段から使用している YouTube というサービスの発展に別の角度から貢献したいと考えたからである.また YouTube は視聴者であるユーザーの見たい動画をレコメンドシステムによって提供しているが,クリエイター側に対する提供は少ない.そしてYouTuberとしてまだ収入を得る事が出来ていない底辺YouTuberなどは,この問題が顕著に現れる.よってこのテーマを扱うことによって,まだ自身の動画のジャンルを確立できいないYouTuber達が,どのような動画を作成するば再生数･高評価数を高く取得できる動画に繋がるか知ることが出来る.

またこのテーマに関する現状として,YouTube というコンテンツを消費する側であるユーザー側のサービスやシステムは存在するが,YouTube というサービスを用いて動画を提供する側であるクリエイターのためのサービスやシステムはまだ少ない.よってこのテーマを実施することによってYouTuberにとって,動画を作成する環境を用意することが出来る.

2. 問題の掲示

このテーマを始めるに至った,課題として挙げられるのが,YouTube という動画配信サービスの広告収入を生活の担保にしているYouTuberが,自身の作成した動画が本当に価値ある動画なのか,再生数・高評価数を高く取得できるのかということが動画を YouTube 上に投稿するまで分かる事ができないという事である.

この問題を解決することにより,得られる成果として挙げられるのが,YouTuber の動画作成にかかるコスト(時間,思考,費用)を少なくすることである.

なぜ YouTube の動画の作成にかかるコストが少なくする事が出来るかというと,YouTuber 自身が動画を作成する前に,自身の考えた動画のタイトル,チャンネルの規模(登録者数),動画のカテゴリという情報から,その動画の予測再生数・予測高評価数を所得する事が出来るからである.

予測再生数･予測高評価数を動画を作成する前に取得する事が出来れば,自身の作成する予定の動画が価値ある動画なのか,再生数を稼ぐ事が出来るのかを知る事が可能であり,高い予測再生数・高い予測高評価数を取得できる動画であれば,そのまま動画を作成して投稿する,低い予測再生数・予測高評価数であれば動画を作成するのを検討でき,動画を作成するコストを軽減する事が出来る.

想定するゴールとして,YouTube 動画のタイトル・チャンネルの規模(登録者数)・動画のカテゴリ ID という上右方から,動画の予測再生数・予測高評価数を取得するシステムの実装とする.

3. 方法の掲示

このテーマを始めるに至って,どのように方法で問題の解決に取り組むかというと,YouTube API,kaggle の Trending YouTube Video Statistics データセットから必要なデータを抽出し,機械学習の線形回帰モデルを使用し予測モデルを構築する.

具体的にどのようにするかというと,kaggle のデータセットからは YouTube 動画のタイトル・高評価数を取得し,YouTube API からはチャンネル登録者数を取得する事が可能なので,それらのデータから 3 つの特徴量を作成しそれらのデータを X_train とする.1 つ目の特徴量としてタイトルを TF-IDF を用いて,文章中の単語の重要度を数値化させたデータを作成し,そのデータを特徴量とする.2 つ目の特徴量として動画のカテゴリ ID は,そのままの数値では学習器が学習しにくいので One-Hot エンコーディングによってカテゴリ変数を 0,1 などの変数に変換を行う[(1)].One-Hot エンコーディングをどのように行うかというと,Python の機械学習ライブラリである scikit-learn を用いる.Scikit-leaen には One-Hot エンコーディングをするのに必要な OneHotEncoder という関数があるので,その関数を動画カテゴリ ID のデータに適用させ,ダミー変数化したデータを生成し,そのデータを特徴量とする.3 つ目の特徴量はデータセットから動画の高評価数･低評価数を取得する事が出来るので,高評価数が全体のどれくらいの割合なのかをパーセントで表した数値を特徴量とする.また再生数と高評価数は Y_train というデータの特徴量とする.

上記の X_train と Y_train を線形回帰モデルに組み込み学習させることによって,予測再生数・予測高評価数を取得できるモデルを構築する事が出来る.

,

参考文献

[1] One-Hotエンコーディングなら pandas の get_dummies() を使おう

(https://blog.shikoan.com/pandas-get-dummies/), (参照 2020-1-28).