JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 26, No. 1, pp. 175-182, Jan. 2025

Check for updates

트리기반 머신러닝을 이용한 새로운 동영상 조회수 예측 영향성 인자 분석

김 진 웅[1†] · 서 지 원[2†] · 손 창 호[3*] · 최 승 호[4*]
[1]한성대학교 IT융합공학부 학부과정
[2]한성대학교 컴퓨터공학과 학부과정
[3]육군3사관학교 국방시스템과학과 교수
[4]한성대학교 기초교양학부 조교수

Analysis of Influential Factors for the Prediction of YouTube Views Using Tree-based Machine Learning

Jin-Woong Kim[1‡] · Ji-Won Seo[2‡] · Chang-Ho Son[3*] · Seoung-Ho Choi[4*]

[1]Undergraduate Course, Department of Convergence IT Engineering, Hansung University, Seoul 02876, Korea
[2]Undergraduate Course, Department of Computer Engineering, Hansung University, Seoul 02876, Korea
[3]Professor, Department of System Engineering Korea Army Academy, Gyeongsangbuk-do 38900, Korea
[4]Assistant Professor, College of Liberal Arts, Faculty of Basic Liberal Art, Hansung University, Seoul 02876, Korea

[요 약]

유튜브는 최근 몇 년 간 전 세계에서 가장 인기 있는 미디어 공유 플랫폼으로 자리 잡았다. 유튜브 조회수는 콘텐츠 성공의 중요한 지표로, 정확한 예측이 콘텐츠 전략 개발에 큰 도움이 된다. 그러나 유튜브 조회수를 예측하는데 어떠한 요인이 영향성을 끼치는지에 대해서 알려져 있지 않아 새로운 콘텐츠에 대해서 조회수를 예측하는 것은 어렵다. 새로운 동영상 조회수 예측하는데 미치는 영향 인자를 찾아 도움을 주고자, 우리는 모든 카테고리를 포함하는 유튜브 영상의 메타데이터와 파생 변수를 이용해 새로운 동영상 조회수를 예측하는 프로그램을 제안한다. 우리는 트리 기반 머신러닝 모델 5개를 사용하여 조회수 예측 성능을 검증했고, SHAP 분석을 통해 조회수 예측에 주요 영향을 미치는 특성을 분석했다. 우리는 제안된 방법을 통해 좋아요 수, 동영상 길이, 카테고리 ID 등 여러 특성이 유튜브 조회수 예측에 중요한 영향을 미치는 요인임을 확인했다.

[Abstract]

In recent years, YouTube has become the most popular media sharing platform worldwide. The number of views of YouTube videos is a crucial indicator of content success, and its accurate prediction can significantly help in the development of content strategies. However, the factors influencing YouTube view predictions are not well understood, making it difficult to predict the views of new content. To address this, we propose a method to predict the views of new videos using metadata and derived variables from YouTube videos across all categories. We validated the prediction performance using five tree-based machine learning models and analyzed the key features influencing view prediction through SHapley Additive exPlanations (SHAP) analysis. Through the proposed method, we confirmed that several factors, such as the number of likes, video length, and category ID, have a significant impact on the prediction of YouTube views.

색인어 : 유튜브 조회수 예측, 머신러닝, 트리 기반 모델, SHAP 분석, 인자 분석

Keyword : YouTube View Prediction, Machine Learning, Tree-based Models, SHapley Additive exPlanations (SHAP) Analysis, Factor Analysis

http://dx.doi.org/10.9728/dcs.2025.26.1.175

Received 08 October 2024; **Revised** 06 November 2024
Accepted 11 December 2024

‡ **These authors contributed equally to this work**

***Corresponding Author; Chang-Ho Son, Seoung-Ho Choi**

Tel: ※ 개인정보 표시제한

E-mail: c13981@kaay.ac.kr, jcn99250@naver.com

Ⅰ. 서 론

최근 몇 년 간 유튜브는 전 세계에서 가장 인기 있는 미디어 공유 플랫폼 중 하나로 자리 잡았다. 전 세계의 사용자들은 매일 수십억 개의 비디오를 시청하며, 이는 콘텐츠 제작자와 기업에게 중요한 기회를 제공한다[1]. 유튜브 조회수는 단순한 인기 지표를 넘어, 콘텐츠의 성공을 평가하고 마케팅 전략을 수립하는 데 필수적인 요소이다. 따라서 유튜브 조회수를 정확히 예측하는 것은 콘텐츠 제작자와 기업이 자신의 비디오 콘텐츠에 대한 효과적인 전략을 개발하는 데 큰 도움이 된다[2].

유튜브 조회수 예측은 특히 마케팅과 관련하여 매우 중요한 역할을 한다. 콘텐츠 제작자와 브랜드는 조회수를 기반으로 광고 수익을 창출하고, 콘텐츠의 성과를 분석하여 향후 전략을 수립한다. 높은 조회수는 더 많은 광고 수익으로 이어질 뿐만 아니라, 콘텐츠의 확산에도 긍정적인 영향을 미친다. 또한, 조회수가 많은 콘텐츠는 유튜브의 알고리즘에 의해 더 많이 추천되어, 자연스럽게 더 많은 시청자를 유입시킨다. 이러한 이유로, 정확한 조회수 예측은 광고 캠페인 및 콘텐츠 마케팅 전략의 성공 여부에 큰 영향을 미친다[3].

하지만 유튜브 조회수 예측은 여러 가지 이유로 어려움을 겪는다. 첫째, 각 비디오는 주제와 형식이 다양하여 조회수에 대한 반응이 상이하다. 예를 들어, 어떤 콘텐츠는 특정 시청자에게 깊이 있는 관심을 받을 수 있지만, 다른 콘텐츠는 예상치 못한 인기를 끌 수 있다. 이러한 예측 불가능성은 조회수 예측에 복잡성을 더한다. 둘째, 유튜브 사용자의 행동 패턴은 매우 다양하고 비예측적이다. 사용자의 선호와 소비 습관은 지속적으로 변화하기 때문에, 어떤 비디오가 예상보다 더 높은 조회수를 기록할 수도, 반대로 낮은 조회수를 기록할 수도 있다. 이러한 유튜브 조회수의 복잡성은 기존의 단순한 통계적 접근 방법으로는 효과적으로 설명하기 어려운 문제이다[4]. 이러한 어려움을 해결하기 위해 우리는 모든 카테고리를 포함한 유튜브 영상의 메타데이터와 파생 변수를 활용한 머신러닝 기반의 유튜브 조회수 예측 프로그램을 제안한다. 제안한 프로그램은 유튜브 새로운 동영상 조회수 예측에 대해 다음과 같은 공헌을 한다.

- 우리는 처음으로 모든 카테고리를 포함한 유튜브 영상 메타데이터와 파생 변수를 활용하여, 새로운 동영상에 대한 조회수 예측에 유용한 프로그램을 제안한다.
- 우리는 새로운 동영상 조회수 예측을 위해 5개의 트리 기반 머신러닝 모델을 개발했다.
- 우리는 SHAP의 Summary 분석을 통해 각 독립 변수가 조회수 예측에 어떻게 기여하는지 분석했다.

나머지 논문은 다음과 같이 구성되어 있다. 2장에서는 본 논문에서 수행한 연구와 관련된 선행 연구들에 관해 기술한다. 3장에서는 제안 방법에 관해 기술한다. 4장에서는 실험 방법에 관해 기술한다. 5장에서는 실험 결과, 6장에서는 결론을 기술하며 논문을 마무리한다.

Ⅱ. 관련 연구

서범근, 그리고 이한준의 연구[5]에서는 유튜브 먹방 콘텐츠의 인기에 영향을 주는 요인들을 식별하여 유튜브 먹방 콘텐츠의 인기를 예측하는 프로그램을 제안했다. 해당 연구에서는 SHAP 분석을 수행하여 조회수 예측 프로그램과 좋아요 수 예측 프로그램에서 각각의 중요 변수를 도출하여 콘텐츠 조회와 좋아요 반응에 대한 선행요인이 다름을 확인했다. 또한 Random Forest, XG Boost, Light GBM 등의 머신러닝 모델을 사용하여 조회수 및 좋아요 수를 예측했다. SHAP 분석 결과 조회수 예측에 가장 큰 영향을 미치는 요인은 구독자 수, 좋아요 수 예측에 가장 큰 영향을 미치는 요인은 크리에이터의 매력도임을 도출했다.

이승규, 오현진, 임형섭, 그리고 정은혜의 연구[6]에서는 유튜브 썸네일을 기반으로 구독자 수 대비 조회수를 예측하는 프로그램을 제안했다. 해당 연구에서는 Google Cloud Vision API와 KoBERT 모델을 활용하여 썸네일의 이미지를 분석하여 파생 변수로 생성했다. 또한 Box-COX 변환을 사용하여 변수 간의 편향을 조정하여 정규성을 확보했으며, Random Forest, Light GBM, 그리고 XG Boost 모델을 활용하여 예측 모델의 성능을 비교했다. SHAP 분석 결과 구독자 수, 영상 길이, 영상 게시 날짜 등의 영상 정보에 대한 변수가 구독자 수 대비 조회수 예측에 중요한 변수라는 결론을 도출했다.

김 슬, 김수인, 박훈, 임수빈, 그리고 정재우의 연구[7]에서는 유튜브 쇼츠 내 뷰티 광고 콘텐츠의 메타 및 시청각 데이터를 수집하여 조회수를 예측하는 프로그램을 제안했다. 해당 연구에서는 영상 게시일, 구독자 수, 광고 표기 방식, 채널 생성 기간 등 총 7개의 메타데이터와 썸네일의 HSV, 유튜버 PrettyScale 값, 목소리 발화 속도 평균값 등 총 9개의 시청각 데이터를 모델의 독립변수로 설정했다. 조회수 예측을 위해 XG Boost 모델을 사용하여 선행 연구와의 성능을 비교했다. SHAP 분석 결과 메타데이터에서는 채널 생성 기간, 광고 직/간접 표시 여부가, 시청각 데이터에서는 유튜버의 외모 매력도와 유튜버 발화 속도의 평균값이 쇼츠 조회수 예측에 주요 변수로 나타났다.

박해연, 이강훈, 장연우, 김형석, 그리고 배성호의 연구[8]에서는 딥러닝 알고리즘을 기반으로 유튜브 먹방 콘텐츠의 조회수를 예측하는 프로그램을 제안했다. 해당 연구에서는 모델의 독립 변수로 영상 데이터, 음성데이터, 메타데이터 등 3가지 성격의 입력 정보를 사용했고, 종속 변수는 유튜브 먹방 콘텐츠의 조회수로 설정했다. 조회수를 예측하기 위해 3개의 LSTM 모델을 이용해 3가지 성격의 입력 정보를 각 모델의

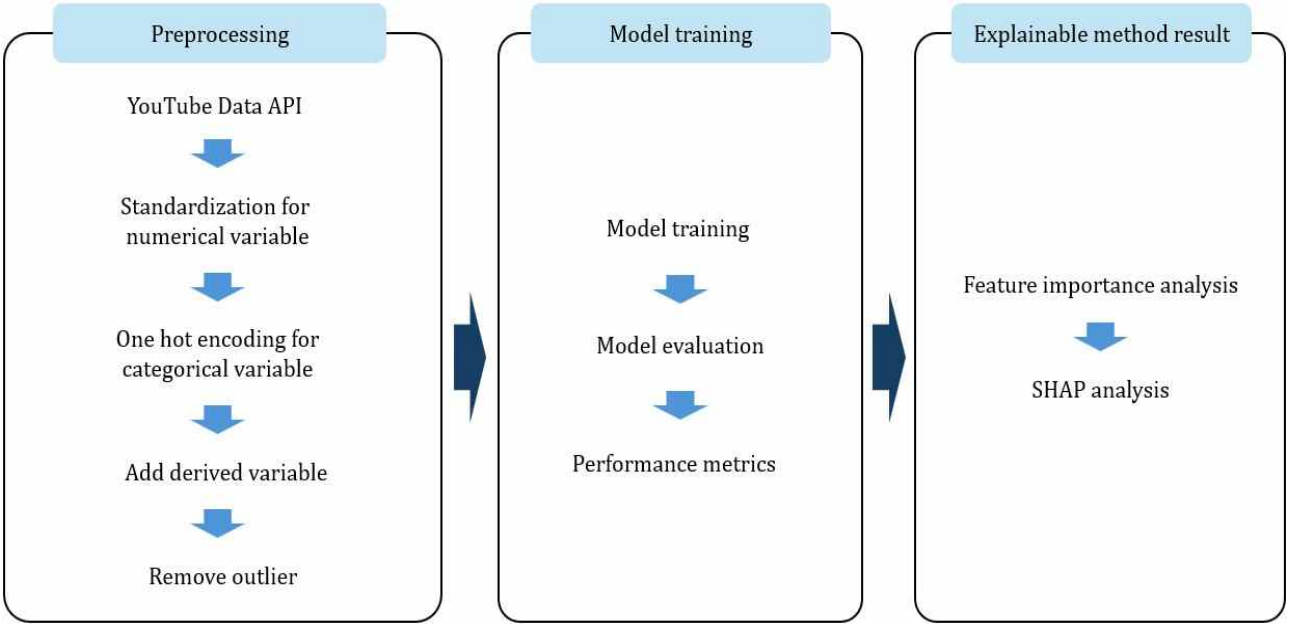

그림 1. 제안된 유튜브 조회수 예측 프로그램의 구성도
Fig. 1. Architecture of the proposed YouTube view prediction program

독립 변수로 사용했다. 이후 각 모델을 융합하여 조회수를 예측했다. 동영상 자체 정보를 통해 동영상의 조회수에 대한 개괄적 예측이 가능하다는 것과 동영상 정보 학습 시 정보의 융합 방법에 따라 모델의 예측 성능이 달라질 수 있다는 결론을 도출했다.

이정선의 연구[9]에서는 고령층의 삶의 만족도를 예측하고, 디지털 정보격차 요인을 탐색하기 위해 트리 기반 앙상블 머신러닝 모델을 사용했다. 연구에서는 Random Forest, XGBoost, LightGBM, 그리고 CatBoost를 사용해 성능을 비교했으며, CatBoost 모델이 가장 우수한 성능을 보였다. 비록 본 연구에서 다루는 유튜브 데이터와 데이터의 성격이 다르지만, 트리 기반 모델들이 다양한 변수 간의 복잡한 상호작용을 잘 포착하고 예측 성능을 높이는 데 효과적이라는 공통점이 있다.

앞서 소개된 연구들은 유튜브 조회수 예측을 위한 다양한 접근법을 제안했다. 연구들은 유튜브 콘텐츠의 인기에 영향을 미치는 요인들을 분석하고, 이를 바탕으로 조회수 및 좋아요 수를 예측하는 프로그램을 개발했다. 특히, SHAP 분석을 통해 두 종속 변수에 영향을 미치는 요인이 다르다는 점을 확인했으며, 썸네일 이미지를 분석하여 구독자 대비 조회수를 예측하는 프로그램도 제안했다. 메타데이터와 시청각 데이터를 결합하여 조회수를 예측하는 접근법이 다수 나타났으며, 딥러닝 기법을 활용한 예측 연구도 진행했다. 이러한 연구들은 머신러닝 및 딥러닝 모델을 적용했지만, 특정 콘텐츠 유형이나 데이터 세트에 초점을 맞춰져 있는 경우가 많았다.

기존 연구와 달리, 우리는 유튜브의 모든 카테고리를 포함하는 데이터셋을 사용하여 조회수를 예측하고, 이 과정에서 각 요인이 조회수 예측에 미치는 영향을 세부적으로 분석한다.

Ⅲ. 제안 방법

우리는 모든 카테고리를 포함한 유튜브 영상 메타데이터와 파생 변수를 활용하여 조회수를 예측하는 프로그램을 제안한다. 이 프로그램은 유튜브 조회수 예측 정확도를 높이기 위해 다양한 변수를 적용하여 조회수 변동을 분석한다. 그림 1은 제안한 프로그램의 구성도를 구체적으로 보여준다.

유튜브 동영상의 메타데이터를 추출하기 위해 YouTube Data API를 사용하여 제목, 조회수, 좋아요 수, 카테고리 ID 등 동영상의 메타데이터를 수집한다. YouTube Data API는 개발자가 YouTube 플랫폼과 상호작용할 수 있도록 해주는 RESTful API로, 유튜브 동영상, 채널, 재생목록 등의 메타데이터를 검색하고 조작할 수 있으며, 다양한 기능을 통해 YouTube의 데이터를 효과적으로 활용할 수 있도록 지원한다[10].

유튜브 동영상의 메타데이터와 파생 변수를 이용하여 조회수를 예측하기 위해 Random Forest, Extra Trees, CatBoost, LightGBM, 그리고 XGBoost를 포함한 5개 트리 기반 머신러닝 모델을 사용한다. 트리 기반 모델은 구조화된 데이터를 처리하는 데 매우 적합하며, 고차원 데이터셋에서도 우수한 성능을 발휘한다. 특히, 이들 앙상블 기법은 변수 간의 비선형 관계와 상호작용을 효과적으로 포착하는 강점을 지닌다[15],[16]. 이러한 특성 덕분에, 우리는 이 모델들이 유튜브 동영상의 메타데이터와 파생 변수를 분석하는 데 적합하다고 판단했다. 이를 통해 조회수를 예측하는 데 활용하고자 한다.

5개의 트리 기반 머신러닝 모델에 대해 각각 학습을 수행

한 후 가장 성능이 우수한 모델에 대해 SHAP (SHapley Additive exPlanations) 분석과 지니 불순도(Gini impurity) 기반 특성 중요도 분석을 수행한다. 지니 불순도 기반 특성 중요도는 모델이 예측하는 데 있어 각 특성이 얼마나 중요한지 파악하는 데 유용하며, 이를 통해 유튜브 동영상의 메타데이터와 파생 변수 중 조회수 예측에 중요한 요인을 식별한다[11]. SHAP 분석은 각 특징이 모델의 예측에 어떻게 기여하는지를 정량화함으로써 예측 결과에 대한 각 특성의 영향을 명확하게 분석할 수 있도록 도와준다[12]. SHAP 분석을 통해 유튜브 동영상의 메타데이터 중 조회수 예측에 가장 큰 영향을 미치는 요인을 더욱 구체적으로 분석한다.

Ⅳ. 실험 방법

우리는 YouTube Data API를 이용하여 유튜브 동영상의 메타데이터를 수집했다. 무작위로 선택된 400개의 유튜브 동영상에서 각 동영상의 비디오 ID, 제목, 카테고리 ID, 카테고리 이름, 조회수, 좋아요 수, 동영상 길이 등 메타데이터를 추출하고, 수집한 데이터를 50개씩 8개의 CSV 파일로 저장했다.

수집한 동영상의 메타데이터를 바탕으로 동영상이 업로드된 요일 및 해당 요일이 주말인지 여부를 파생 변수로 생성해 모델의 독립 변수로 추가했다. 또한, 동영상 길이에 대해 범주화를 수행하고, IQR 방식을 사용해 이상치를 제거했다. 데이터 전처리 단계에서는 결측치가 포함된 행을 제거하고, 범주형 변수에 대해서는 원 핫 인코딩을 적용했다. 마지막으로, 모델 학습을 위해 범주형 변수를 제외한 나머지 수치형 변수에 Standard Scaler를 적용해 표준화를 진행했다.

이후 학습 데이터를 Train, Validation, Test 데이터셋으로 분할했다. 총 8개의 CSV 파일 중 4개는 Train 데이터로, 2개는 Validation 데이터로, 2개는 Test 데이터로 사용했다. 학습을 위해 모델의 종속 변수는 유튜브 동영상의 조회수로 설정했다.

유튜브 동영상의 메타데이터 및 파생 변수를 이용하여 동영상의 조회수를 예측하기 위해 Random Forest, Extra Trees, CatBoost, LightGBM, 그리고 XGBoost를 포함한 5개의 트리 기반 머신러닝 모델을 사용했다.

Random Forest는 여러 개의 의사결정나무를 종합하여 예측 성능을 향상시키는 방법이다. 각 나무는 부트스트랩 샘플 및 특징 배깅을 활용한 변수의 집합을 통해 생성되며, 모든 트리의 예측을 평균 또는 다수결로 결정하여 최종 예측 결과를 생성한다. 이를 통해 과적합을 방지하고 모델의 일반화 성능을 향상시키는 데에 효과적으로 사용된다[11].

Extra Trees (Extremely Randomized Trees)는 Random Forest와 유사하지만, 더 많은 무작위성을 도입하여 모델을 학습시키는 앙상블 방법이다. Random Forest가 각 노드에서 최적의 분할을 찾는 반면, Extra Trees는 무작위로 선택된 분할 기준을 사용하여 트리를 구성한다. 이로 인해 모델의 편향은 약간 증가할 수 있지만, 분산은 감소하여 과적합을 방지하는데 도움이 된다[13].

CatBoost (Categorical Boosting)는 Yandex에 의해 개발된 Gradient boosting 라이브러리로 범주형 데이터를 효과적으로 처리할 수 있다. CatBoost는 Ordered Boosting 기법을 채택하여 과적합을 방지하고 모델의 일반화 성능을 향상시킨다. 또한, CatBoost는 효율적인 학습과 예측을 위해 최적화된 구현을 제공하며, 대규모 데이터셋 및 복잡한 문제에서도 우수한 성능을 발휘한다[14].

LightGBM (Light Gradient Boosting Machine)은 마이크로소프트가 개발한 Gradient Boosting 기법으로, 대규모 데이터 및 고차원 공간에서 효율적으로 작동하도록 설계되었다. Level-wise 방식을 사용하는 GBM과 XGBoost와 달리 Leaf-wise 방식을 사용하여 최대 손실을 갖는 리프 노드를 지속적으로 분할하여 깊고 비대칭적인 트리를 만든다. 그리고 GOSS (Gradient-based One-Side Sampling)과 EFB (Exclusive Feature Bundling) 기법을 이용하여 학습 속도를 크게 향상한다[15].

XGBoost (Extreme Gradient Boosting)는 Gradient Boosting의 효율성과 성능을 개선하기 위하여 알고리즘을 고도화하고 시스템 디자인을 최적화한다. 정규화, 병렬 처리, 최대 깊이를 지정한 가지치기를 통해 학습 속도와 모델의 일반화 성능을 높인다. 그러나 메모리 사용량이 많고 불균형 데이터에 대해서는 낮은 성능을 보인다는 한계를 지닌다[16].

각 머신러닝 모델에 대해 Grid Search를 이용해 하이퍼파라미터 튜닝을 수행하여 최적의 하이퍼파라미터 값을 도출했다. 이렇게 도출한 최적의 하이퍼파라미터를 사용해 모델 학습을 진행했다.

5개 회귀 모델의 성능을 평가하기 위해 RMSE, MAE, SMAPE, R^2 score 등 4개의 평가지표를 사용했고, 각 평가지표의 수식은 다음과 같다.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{1}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2}$$

$$SMAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{\frac{|y_i| + |\hat{y}_i|}{2}} \times 100 \tag{3}$$

$$R^2\,score = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{4}$$

위 수식에서 n은 데이터의 개수를 의미한다. 수식 (1)은 RMSE (Root Mean Squared Error)를 나타낸다. RMSE는 MSE (Mean Squared Error)의 제곱근으로, 예측값과 실제값 간의 오차를 측정하는 지표이다. MSE는 각 오차의 제곱의 평균을 구한 값으로, 값이 클수록 모델의 예측 정확도가 낮음을 의미한다. RMSE는 MSE와 달리 오차의 단위를 원본 데이터와 동일하게 맞춰 직관적인 해석이 가능하다. RMSE가 작을수록 모델의 예측 성능이 좋다고 평가할 수 있다. 수식 (2)는 MAE (Mean Absolute Error)를 나타낸다. MAE는 실제값과 예측값 사이의 절대 오차의 평균을 의미한다. 항상 0 이상의 값을 가지며 그 값이 클수록 예측값과 실제값 사이의 평균 오차가 크다는 것을 의미한다. 수식 (3)은 SMAPE (Symmetric Mean Absolute Percentage Error)를 나타낸다. SMAPE는 MAPE가 가진 한계를 보완하기 위해 고안된 평가지표로 절대 오차를 절대 실제값과 절대 예측값의 평균으로 나눈 값을 나타낸다. 0%에서 200% 사이의 값을 가지며 그 값이 클수록 예측값과 실제값 사이의 상대적 차이가 크다는 것을 의미한다. 수식 (4)는 R^2 score를 나타낸다 R^2 score는 독립 변수에 의해 설명되는 종속 변수의 변산성(종속 변수 값들이 얼마나 다양하게 변하는지)의 비율을 나타낸다. 보통 0에서 1 사이의 값을 가지며, 그 값이 1에 가까울수록 종속 변수의 변산성이 독립 변수에 의해 잘 설명됨을 의미한다. 한편 그 값이 음수로 나올 경우, 모델이 예측한 값이 종속 변수의 평균보다도 설명력을 갖지 못한다는 뜻으로 잘못된 예측을 하고 있다고 해석할 수 있다.

5개의 모델 학습이 끝난 후 4개의 평가지표를 종합적으로 고려하여 가장 우수한 성능을 보인 모델을 선택하고, 해당 모델을 이용해 지니 불순도 기반으로 특성 중요도를 추출한 후, 특성 중요도를 기준으로 내림차순으로 정렬했다. 이후 상위 15개의 특성 중요도를 시각화했다. 또한 가장 우수한 성능을 보인 모델에 대해 SHAP 분석을 수행한 후 Summary plot과 Force plot을 각각 시각화했다. Summary plot은 각 특성의 SHAP 값을 보여주는 그래프이다. 특성 중요도는 SHAP 값의 절대값 크기로 표현되며, 특성 값의 크기에 따라 점의 색상이 달라지고, 이를 통해 특성 값의 영향을 파악할 수 있다. Force plot은 특정 데이터 포인트에 대한 SHAP 값을 보여주는 그래프이다. 각 특성이 모델의 예측에 어떻게 기여했는지 자세히 보여준다. 각 특성의 SHAP 값은 막대 그래프로 표현되며, 양수 값은 왼쪽, 음수 값은 오른쪽으로 표시된다. 특성 값의 크기에 따라 막대의 색상이 달라지며, 이를 통해 특성 값의 영향을 파악할 수 있다.

V. 실험 결과

유튜브 동영상 카테고리의 인기 정도를 파악하기 위해 유튜브 동영상 카테고리별 조회수의 평균과 표준편차를 분석했다. 분석 결과는 표 1에 나와 있다. 표를 살펴보면 코미디, 여행 및 행사, 뉴스 및 정치, 영화 또는 애니메이션 카테고리의 조회수 평균이 다른 카테고리들에 비해 높게 나타났다. 이는 이들 카테고리가 시청자에게 평균적으로 더 큰 흥미를 유발하고 있음을 시사한다.

표 1. 유튜브 동영상 카테고리별 조회수의 평균 및 표준편차.
Table 1. Mean and standard deviation of YouTube video views by category

Category Name	Mean (SD)
Comedy	991,674 (510,324)
Entertainment	635,138 (474,158)
Film and Animation	874,087 (765,968)
Gaming	341,408 (102,391)
Howto and Style	179,161 (36,475)
Music	578,095 (570,275)
News and Politics	899,684 (628,252)
People and Blogs	422,962 (280,770)
Pets and Animals	324,916 (217,342)
Science and Technology	475,680 (313,297)
Sports	758,000 (543,830)
Travel and Events	971,767 (594,254)

표 2는 유튜브 동영상 조회수 예측에 사용한 5개의 트리 기반 머신러닝 모델의 성능을 비교하여 보여준다. 성능은 테스트 데이터를 기준으로 평가했다. 4개의 평가지표를 종합적으로 고려했을 때, CatBoost 모델이 가장 우수한 성능을 보였으며, R^2 score가 0.833으로 가장 높게 나타났다. 이는 CatBoost 모델이 유튜브 동영상 조회수 예측에 가장 적합하고, 종속변수의 변동성을 효과적으로 설명한다는 것을 의미한다.

유튜브 동영상 조회수 예측에서 가장 우수한 성능을 보인 CatBoost 모델을 이용해 지니 불순도 기반 특성 중요도를 추출하고 그림 2에 시각화했다. 그 결과, 좋아요 수, 동영상 길이, 카테고리 ID, 월요일 업로드 여부 등의 특성이 유튜브 동영상 조회수 예측에 가장 큰 영향을 미치는 요인으로 나타났다.

유튜브 동영상의 메타데이터 및 파생 변수가 유튜브 동영상 조회수 예측에 어떻게 기여했는지 분석하기 위해, 가장 우수한 성능을 보인 CatBoost 모델에 대해 SHAP 분석을 수행한 후, Summary plot과 Force plot을 각각 그림 3과 그림 4에 시각화했다. Summary plot에서는 SHAP 값 기준으로 상위 10개의 특성에 대한 결과만 시각화했다. 그림을 살펴보면 좋아요 수, 동영상 길이, 카테고리 ID 특성이 유튜브 동영상 조회수 예측에 가장 크게 기여한 것으로 나타났다.

그림 4는 CatBoost 모델의 SHAP 값을 시각화한 Force plot을 보여준다. 동영상 길이 특성은 유튜브 동영상 조회수 예측값을 증가시키는 방향으로 기여했으며, 좋아요 수와 카테고리 이름이 '사람과 블로그'인 경우는 유튜브 동영상 조회수 예측값을 감소시키는 방향으로 기여했다.

표 2. 회귀 모델 성능 비교 결과.
Table 2. Evaluating regression model performance results

Regression Models	Data type	RMSE	MAE	SMAPE	R² score
Random Forest [11]	Validation	0.121	0.085	20.516	0.617
	Test	0.106	0.080	18.675	0.785
Extra Trees [12]	Validation	0.122	0.091	21.574	0.613
	Test	0.106	0.077	18.728	0.784
CatBoost [13]	Validation	0.116	0.083	20.236	0.646
	Test	**0.093**	**0.065**	**17.458**	**0.833**
LightGBM [14]	Validation	0.136	0.093	26.139	0.519
	Test	0.107	0.075	19.372	0.780
XGBoost [15]	Validation	0.118	0.086	22.448	0.634
	Test	0.098	0.070	18.879	0.816

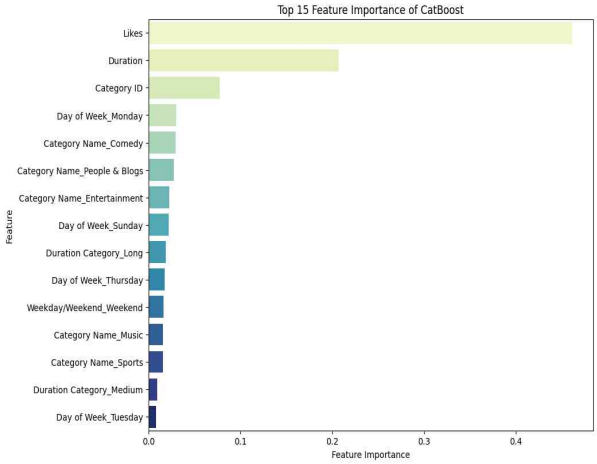

그림 2. CatBoost 모델의 상위 15개 특성 중요도를 보여주는 그래프
Fig. 2. Visualizing top 15 feature importances of CatBoost model

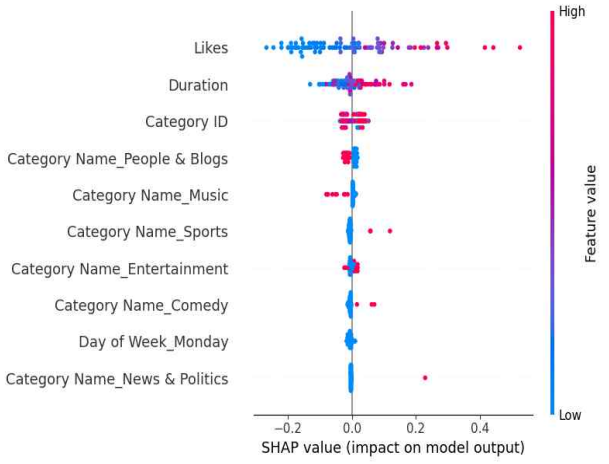

그림 3. CatBoost 모델의 SHAP 분석을 통한 특성 중요도 시각화
Fig. 3. Visualizing feature importance of CatBoost model with Summary plot

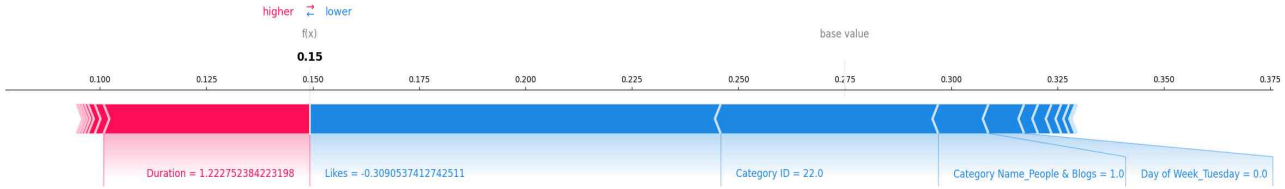

그림 4. CatBoost 모델의 SHAP 값을 시각화한 Force plot
Fig. 4. Visualizing SHAP values of CatBoost model with Force plot

Ⅵ. 결 론

우리는 모든 카테고리를 포함하는 유튜브 동영상 메타데이터와 파생 변수를 활용하여 조회수를 예측하고, 예측 과정에서 중요한 요인들을 분석했다. YouTube Data API를 통해 유튜브 동영상의 제목, 조회수, 좋아요 수, 카테고리 ID 등 다양한 메타데이터를 추출하고, 예측 성능 향상에 기여할 수 있는 파생 변수를 추가했다. 이러한 특성들은 예측 모델 학습에 중요한 기여를 했으며, 이를 통해 유튜브 조회수를 예측하는 효과적인 방법을 제시했다.

유튜브 동영상의 카테고리가 조회수에 미치는 영향을 분석하기 위해, 각 카테고리별 조회수의 평균과 표준편차를 시각화했다. 그 결과, 코미디, 여행 및 행사, 뉴스와 정치 카테고리의 조회수 평균이 다른 카테고리보다 상대적으로 높게 나타났다. 이는 이들 카테고리가 평균적으로 높은 조회수를 기록하며 시청자들의 높은 관심을 반영하고 있음을 나타낸다.

유튜브 동영상 조회수 예측을 위해 동영상의 메타데이터 및 파생 변수를 사용하여 5개의 트리 기반 머신러닝 모델을 학습하고, 4개의 평가지표를 활용하여 성능을 평가했다. 그 결과, CatBoost 모델이 다른 모델들과 비교하여 가장 우수한 성능을 나타냈으며, R^2 score는 0.833으로 나타났다. CatBoost는 Gradient Boosting Algorithm을 기반으로 순차적으로 오류를 보정하고 예측 성능을 향상하는 방식으로 작동한다. 특히, CatBoost는 데이터 순서나 분포에 민감하지 않으며, 범주형 변수를 자동으로 처리해 중요한 특성들을 효과적으로 반영할 수 있었다. 이러한 CatBoost의 특성 덕분에 유튜브 동영상 조회수 예측에서 뛰어난 성능을 발휘했고, 조회수 예측 변산성의 주요 요인들을 효과적으로 설명할 수 있었다[14].

유튜브 동영상의 메타데이터 및 파생 변수가 유튜브 동영상 조회수 예측에 미치는 영향을 분석하기 위해 CatBoost 모델을 활용해 지니 불순도를 기반으로 특성 중요도를 추출했다. 상위 15개의 특성을 시각화한 결과, 좋아요 수, 동영상 길이, 카테고리 ID 특성이 유튜브 동영상 조회수 예측에 가장 큰 영향을 미치는 특성으로 나타났다. 이어서, SHAP 분석을 통해 좋아요 수, 동영상 길이, 카테고리 ID가 주요 특성임을 다시 확인했으며, 특정 데이터 포인트에 대해 Force plot을 시각화한 결과 카테고리가 '사람과 블로그'일 경우와 좋아요 수가 낮은 경우에는 유튜브 동영상 조회수 예측을 낮추는 방향으로, 반대로 동영상의 길이가 긴 경우에는 조회수 예측을 높이는 방향으로 기여하는 경향을 발견했다.

이 연구는 유튜브 동영상 조회수 예측을 위한 새로운 접근법을 제시하며, 기존 연구들과의 차별화되는 중요한 점들을 가지고 있다. 기존 연구들은 주로 특정 콘텐츠 유형이나 데이터셋에 집중하여 조회수 예측을 수행했으나, 본 연구는 유튜브의 모든 카테고리를 포함한 메타데이터와 파생 변수를 활용하여 더 포괄적이고 일반화 가능한 예측 모델을 구축했다. 이로 인해 기존 연구들과의 직접적인 성능 비교는 어렵지만, 다양한 콘텐츠 유형에 적용 가능한 예측 모델을 제시하고 각 카테고리가 조회수에 미치는 영향을 세부적으로 분석할 수 있었다. 이러한 차별화된 접근은 콘텐츠 제작자와 마케팅 전문가에게 유용한 인사이트를 제공하며, 기존 연구들과 비교했을 때 보다 넓은 범위에서의 예측 정확도를 높이는 데 기여할 수 있다.

향후 연구에서는 더 많은 변수와 데이터를 포함하여 예측 모델의 성능을 향상하고, 다른 소셜 미디어 플랫폼에 대한 유사한 분석을 수행하여 보다 포괄적인 이해를 도모할 수 있을 것이다. 또한, 특정 카테고리의 동영상에 대한 조회수 예측의 변동성을 줄이기 위해 특성 공학(feature engineering) 기법을 추가로 적용할 필요가 있다.

감사의 글

이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2022R1F1A1062959).

참고문헌

[1] IGAWorks. YouTube App Analysis Report [Internet]. Available: https://www.mobileindex.com/insight-report.

[2] Linda Window. The Impact of YouTube Views on Content Creators [Internet]. Available: https://lindawindow.com/the-impact-of-youtube-views-on-content-creators/.

[3] Keyword Search. Mastering YouTube Advertising: Key Strategies [Internet]. Available: https://www.keywordsearch.com/blog/mastering-youtube-advertising-key-strategies.

[4] V. Gupta, A. Diwan, C. Chadha, A. Khanna, and D. Gupta, "Machine Learning Enabled Models for YouTube Ranking Mechanism and Views Prediction," arXiv:2211.11528, November 2022. https://doi.org/10.48550/arXiv.2211.11528

[5] B. Seo and H. Lee, "A Machine Learning-Based Popularity Prediction Model for YouTube Mukbang Content," *Journal of Internet Computing and Services*, Vol. 24, No. 6, pp. 49-55, December 2023.

[6] S. G. Lee, H. J. Oh, H. S. Lim, and E. H. Choung, "Predictive Views per Subscriber Counts Model Based on Youtube Thumbnails Using Machine Learning Approach," in *Proceedings of the 2024 Conference of the Korea Contents Association*, Busan, pp. 5-6, May 2024.

[7] S. Kim, S. I. Kim, H. Park, S. B. Lim, and J. W. Jung, "Predicting the Number of Views for YouTube Shorts - Focusing on Beauty Advertising Contents," in *Proceedings of the 2024 Summer Conference of the Korean Society of Computer Information*, Jeju, pp. 131-132, July 2024.

[8] H. Y. Park, G. H. Lee, Y. W. Jang, H. S. Kim, and S. H. Bae, "A New YouTube View Count Prediction Method Using Deep Audio-Video Multimodal Learning," in *Proceedings of the 2019 Korea Computer Congress (KCC 2019)*, Jeju, pp. 1902-1904, June 2019.

[9] J.-S. Lee, "Exploration of Digital Divide Factors Affecting Life Satisfaction in Older People: Using Tree-Based Ensemble Machine Learning and SHAP," *Journal of Digital Contents Society*, Vol. 25, No. 7, pp. 1847-1860, July 2024. https://doi.org/10.9728/dcs.2024.25.7.1847

[10] The Startup Founder. All You Need to Know About the YouTube Data API [Internet]. Available: https://www.thestartupfounder.com/all-you-need-to-know-about-the-youtube

-data-api/.

[11] L. Breiman, “Random Forests,” *Machine Learning*, Vol. 45, No. 1, pp. 5-32, October 2001. https://doi.org/10.1023/A:1010933404324

[12] S. M. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS ’17)*, Long Beach: CA, pp. 4768-4777, December 2017. https://doi.org/10.48550/arXiv.1705.07874

[13] P. Geurts, D. Ernst, and L. Wehenkel, “Extremely Randomized Trees,” *Machine Learning*, Vol. 63, No. 1, pp. 3-42, April 2006. https://doi.org/10.1007/s10994-006-6226-1

[14] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, “CatBoost: Unbiased Boosting with Categorical Features,” in *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS ’18)*, Montréal, Canada, pp. 6639-6649, December 2018. https://doi.org/10.48550/arXiv.1706.09516

[15] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, ... and T.-Y. Liu, “LightGBM: A Highly Efficient Gradient Boosting Decision Tree,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS ’17)*, Long Beach: CA, pp. 3149-3157, December 2017.

[16] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD ’16)*, San Francisco: CA, pp. 785-794, August 2016. https://doi.org/10.1145/2939672.2939785

김진웅(Jin-Woong Kim)

2019년 : 한성대학교 IT융합공학부 학부과정

※관심분야 : 의료 빅데이터, 의료 인공지능

서지원(Ji-Won Seo)

2020년 : 한성대학교 컴퓨터공학과 학부과정

※관심분야 : 인공지능, 생성형 AI, 빅데이터

손창호(Chang-Ho Son)

2002년 : 육군사관학교 기계공학과 공학사

2006년 : North Carolina State University 산업공학 석사

2012년 : 서울대학교 산업공학 박사

2012년～현 재: 육군3사관학교 국방시스템과학과 교수 및 산학협력단장

※관심분야 : 기술경영, 빅데이터분석, 인공지능

최승호(Seoung-Ho Choi)

2018년 : 한성대학교 전자정보공학과(공학사)

2020년 : 한성대학교 전자정보공학과(공학석사)

2023년～현 재: 한성대학교 기초교양학부 조교수

※관심분야 : 딥러닝