Journal of Industrial Information Technology and Application *ISSN(Online): 2586-0852*
Vol.8 No.4

Design of a view prediction system based on YouTube titles using the KernelRidge regression algorithm

Hyeonji Kim[1)] and Yoosoo Oh[2,*)]
[11)] Dept.of Information and Communication Engineering,
Daegu University, Republic of Korea
[2)] School of AI, Daegu University, Republic of Korea

Abstract. The title of a YouTube video is the first thing people see before watching the content. The number of views of a YouTube video differs depending on the level of interest in the title. In this paper, we use a machine learning regression algorithm to analyze the number of views according to the title of Korean YouTube videos. This paper learns YouTube title as the feature value and the number of views as the target value. This paper designs a learning model using the KernelRidge regression algorithm, which achieved the highest performance as a result of analyzing six machine learning regression algorithms (LinearRegression, KNN_Regression, SVR, KernelRidge, DecisionTree Regressor, gradient boosting regressor). The proposed system predicts the number of views through a learned model when a user enters a YouTube title.

Keywords: YouTube, Machine learning, Regression, KernelRidge

1. Introduction

YouTube is a video platform where millions of video creators are active. [1] YouTube has a structure in which the profits of video creators increase as the number of views increases. [2] The composition and length of the title determine high views on YouTube. [1] To achieve elevated views, video creators should use titles matching their videos. However, it is not easy to find a title that reveals the content of the video well and derives people's interest. Therefore, this paper designs a view prediction system according to YouTube titles. YouTube's title data is string data with a nonlinear. In this

* Corresponding author: e-mail yoosoo.oh@daegu.ac.kr
Received: Oct 27, 2024; Accepted: Nov 28, 2024; Published: Dec 31, 2024

paper, to predict the number of views according to the title, encoded using the Countvectorizer algorithm, a frequency-based vectorization technique. Countvectorizer is a method that extracts features from frequently appearing words in each document and vectorizes them. [3] This paper learns the KernelRidge regressor algorithm, a nonlinear regression algorithm. KernelRidge is a nonlinear regression machine learning algorithm that learns the relationship between multiple independent variables and a dependent variable. [4] [5]. The proposed model predicts the number of views when a user enters the YouTube title the user wants to use. Users can select a title based on the expected number of views.

2. Related Research

Haeyeon Park et al. predicted the number of YouTube video views based on deep learning. Hye-yeon Park et al. extracted audio and video information from YouTube videos and learned them along with YouTube metadata (number of subscribers, release period). Based on the training results videos, we confirmed that training audio data, video data, and metadata together is more effective than using only a single piece of information. [2]

William Hoiles et al. analyzed the interactions between YouTube channels and users via YouTube metadata (titles, tags, thumbnails, descriptions). William Hoiles et al. used machine learning regression algorithms through collected YouTube metadata to analyze the impact of metadata on the number of views. William Hoiles et al. confirmed that the Extreme Learning Machine derived the lowest RMSE value for YouTube metadata, and the RandomForest model derived the highest R2-score [1].

Previous related research analyzed the relationship between YouTube and users using various metadata. Additionally, it increases the complexity of the model when various metadata are used. As the complexity of the model increases, more data is required, and the cost of computer computation increases. Therefore, this paper designs a view prediction system using machine learning regression algorithms that achieve high performance even with a simple nonlinear model by using the titles most exposed to users among the YouTube metadata.

3. YouTube View Prediction System

This paper proposes designing a view count prediction system based on YouTube titles using the KernelRidge Regressor algorithm among machine learning regression algorithms. Figure 1 is a diagram of the proposed system.

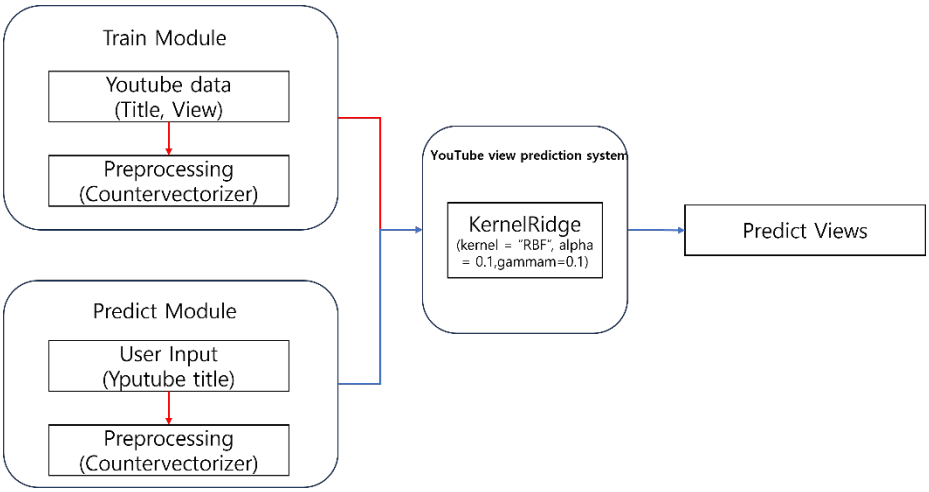

Figure 1. Diagram of the view prediction system according to YouTube title

This paper learns using 1000 YouTube titles and view count data. The proposed system embeds YouTube title data collected using a countervectorizer algorithm. The CountVectorizer algorithm is a frequency-based embedding technique that is intuitive and simple.[3] The KernelRidge Regersor algorithm is used among machine learning regression algorithms to learn embedded data. The KernelRidge Regressor algorithm is a parametric methodology that uses kernel functions to learn to predict parameters.[5][6] The KernelRidge Regressor algorithm is a model that achieves high performance in overfitting and underfitting by regulating the squared weights by imposing a penalty.[6] When the user enters the title data they want to use, the learned model predicts the number of views according to the title entered.

4. Experiments

This paper compares the accuracy after learning through six regression algorithms (LinearRegression, KNN_Regression, SVR, KernelRidge, DecisionTree Regressor, and Gradient Boosting Regressor) to identify the optimal regression machine learning algorithm. This paper compares regression algorithm performance using MSE (Mean Square Error), RMSE (Root Mean Square Error), and R2-Score evaluation indices. As shown in Equation 1, MSE (Mean Square Error) is the squared average error (difference between the predicted value and the actual value).[7] Equation 2 is an RMSE (Root Mean Square Error) formula that takes the root of the MSE value.[7] R2-Score has a value between 0 and 1; the closer it is to 1, the closer it is to show how well the

independent variable represents the dependent variable. The formula for R2-Score is as equation 3.[7]

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i' - y_i)\text{^}$$

Equation 1. MSE(Mean Square Error) equation

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i' - y_i)^2}{n}}$$

Equation 2. RMSE(Root Mean Square Error) equation

$$R^2Score = \frac{SSE}{SST}$$

Equation 3. R2-Score Expression (SSE: sum of squares of the difference between predicted and actual values of the regression model, SSE: sum of squares)

Figures 2, 3, and 4 show the MSE, RMSE, and R2-Score results for six regression machine learning algorithms. As a result of the analysis, the MSE of the KernelRidge model was 0.04 out of the entire MSE range, and the RMSE was 0.08 out of the RMSE as a whole range. Additionally, R2-Score yielded the highest performance at 0.91.

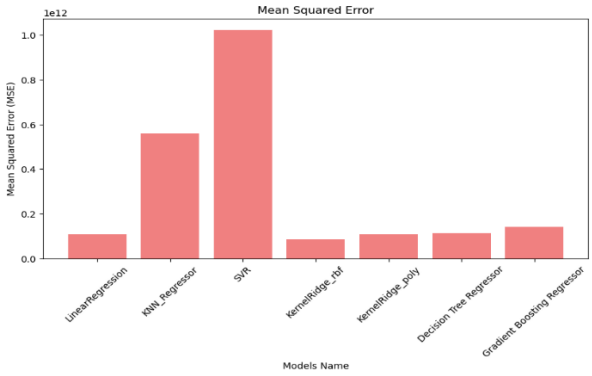

Figure 2. Graph of MSE values for machine learning regression algorithms (LinearRegression, KNN_Regression, SVR, KernelRidge, DecisionTree Regressor, Gradient Boosting Regressor)

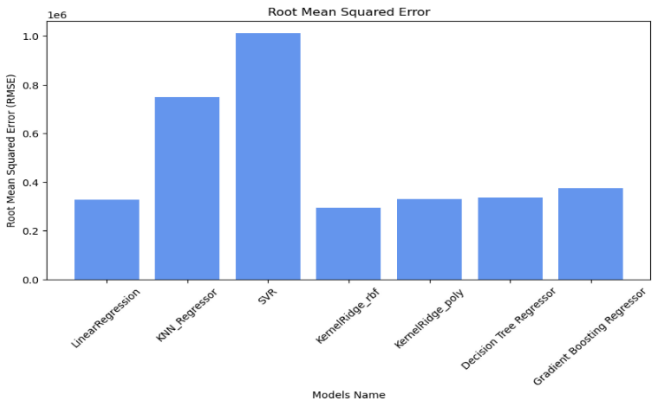

Figure 3. Graph of RMSE values for machine learning regression algorithms (LinearRegression, KNN_Regression, SVR, KernelRidge, DecisionTree Regressor, Gradient Boosting Regressor)

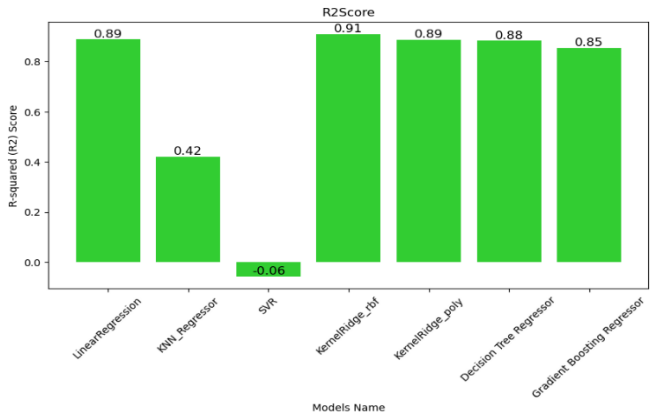

Figure 4. Graph of R2 Score values for machine learning regression algorithms (LinearRegression, KNN_Regression, SVR, KernelRidge, DecisionTree Regressor, Gradient Boosting Regressor)

5. Conclusion

This paper confirmed the algorithm that yields the highest r2Score by analyzing six machine learning regression algorithms. As a result, the KernelRidge algorithm achieved the highest performance with an r2 Score of 0.91. This paper has a simple feature value of one and a small amount of training data of 1000. Therefore, the KernelRidge algorithm and all other algorithms except SVR and KNN Regressor among the models used in the analysis show high r2score performance. In future research, we plan to increase performance by configuring feature values in various and collecting more learning data. This paper designs a view count learning model according to the YouTube

title through KernelRidge and predicts the view count according to the YouTube title the user wants to use through the learned model. The currently proposed system indicates the number of views according to the title. In future research, we plan to design a learning model that includes various metadata by adding the title, the thumbnail image, and the thumbnail content data that affects YouTube views. In addition, future research plans to design a system that predicts the number of views and recommends titles and thumbnails.

References

[1] William Hoiles, “Engagement and Popularity Dynamics of YouTube Videos and Sensitivity to Meta-Data”, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 29, NO. 7, 2017.

[2] Haeyeon Park,Ganghun Lee,Yeonwoo Jang,Hyeongseok Kim,and Sungho Bae. " A New YouTube View Count Prediction Method using DeepAudio-Video Multimodal Learning." Korean Society of Information Scientists and Engineers Academic Presentation Papers 2019.6 (2019): 1902-1904.

[3] Jin Hyeong Jung,and Yong Soo Kim. " Automotive Failure Prediction based on Text Mining of Warranty Data" Reliability application research 20.4 (2020): 357-365.

[4] Seok-hee Han , Jae-hun Jung , Jeong-un Cha ,and Young-gon Kim . " Real time monitoring using Regression algorithm system" Korean Society of Information Scientists and Engineers Academic Presentation Papers 20.1 (2013): 934-936.

[5] Exterkate, Peter, et al. "Nonlinear forecasting with many predictors using kernel ridge regression." International Journal of Forecasting 32.3 (2016): 736-753.

[6] Nguyen, Timothy, Zhourong Chen, and Jaehoon Lee. "Dataset meta-learning from kernel ridge-regression." arXiv preprint arXiv:2011.00050 (2020)

[7] Khan, Mohammad Ayoub, et al. "Performance evaluation of regression models for COVID-19: A statistical and predictive perspective." Ain Shams Engineering Journal 13.2 (2022): 101574.