SMARTYOUTUBER: A DATA-DRIVEN ANALYTICAL PLATFORM TO IMPROVE THE SUBSCRIBER GROWTH AAND SUSTAINABILITY USING ARTIFICIAL INTELLIGENCE AND BIG DATA ANALYSIS

Muyang Li[1], Erik Serbicki[2] and Yu Sun[3]

[1]Shanghai Jiaotong University, Shanghai, China, 200240
[2]University of California, Irvine, Irvine, CA 92620
[3]California State Polytechnic University, Pomona, CA, 91768

ABSTRACT

Youtuber is a new type of freelancer, whose economic profit and personal reputation are highly decided by their own popularity on the Internet, which can be reflected directly by the number of subscribers accumulated. In order to develop the management of YouTube channels and get more advertisement benefits, youtubers need to maintain the current subscriber group and appeal to more new followers by making attractive videos. But they lack efficient methods to analyze their video quality and their communication with subscribers so that they can predict their future development and adjust present strategies. In this paper, we applied several machine learning algorithm and models to study the prediction of short and long term future subscriber increase(we call them as growth and sustainability of youtubers) by analyzing youtuber-related information including video content(e.g. topic type, video tags, etc.) and subscriber interaction(e.g. views, likes, comments, etc.). One highest-scoring regression algorithm is proposed to make the out-performing prediction for certain youtubers, and we have proven its rationality and high accuracy in predicting the growth and sustainability of YouTube subscribers with suitable configuration. Apart from establishing algorithms, a relevant website, which offers services for future prediction and improvement suggestions, is created based on the established random forest regression algorithm. This application allows youtubers to completely analyze their current management situation and assists them to increase popularity for both social and economic benefits.

KEYWORDS

Machine Learning, Video Sharing Platform, Artificial Intelligence, Big Data

1. INTRODUCTION

Nowadays, the development of video platforms like YouTube and TikTok create a special internet celebrity culture that everyone, no matter if you are a celebrity or a normal person, can all have the opportunity to exhibit themselves and receive thousands of people's subscriptions. And with more subscriptions, youtubers will get greater economic effect and more popularity. Take YouTube as an example. There are several levels of YouTubers according to their subscribers, which is the most obvious standard for judging YouTubers' popularity. Some YouTubers might obtain a tremendous number of new subscribers within a short time due to one or several supreme hot videos, while some YouTubers have been managing their account for a

David C. Wyld et al. (Eds): ICAITA, CDKP, CBW, SOFT, CMC, Signal, MLT-2023
pp. 61-71, 2023. CS & IT - CSCP 2023 DOI: 10.5121/csit.2023.130605

long period in order to accumulate a huge fan base. The fanbase of YouTubers determines their views of new published videos, their direct economic benefits from videos and relevant advertisement promotion[15].

Managing the YouTube account, keeping high growth of subscribers and maintaining current follower level require multi-dimensional consideration [17][19][20]. There have already been a lot of YouTubers who can be considered as successful, while some also have been experiencing gradual or rapid loss of subscribers. If we study the inner features of the videos of these YouTubers' videos and analyze their secret, it may be helpful for YouTubers to develop further or maintain present success [12][13][14]. We will predict their future change by models and offer constructive suggestions for them. These suggestions would include both changing present possible demerits and strengthening elements and directions that might make videos more attractive.

Many websites and studies have found other methods in predicting the future popularity of YouTubers. Take the website "Social Blade" as an example. It is a service site for data collection and presentation, and their predictions for future growth of subscribers of YouTubers are based on the subscriber growth data and change curves [11]. This type of prediction will focus more on the real-time updating and accumulating subscriber statistics of certain YouTubers, which makes its prediction more intuitive and easily understood. In another different study, they established a lifetime aware regression model to evaluate and predict the influence of video content on the future popularity of YouTube videos. Their model overcomes the problem of lacking enough historical user data and therefore, they don't need to build up new assumptions of network structure. This strategy allows them to quickly predict future growth within short periods like several hours or a day. However, the issue of these studies is that the basis of their prediction only focuses on specific aspects like historical data and lifetime changing content, while the growth of the subscriber group is a result of diverse attributes. This affects their future prediction accuracy for a longer time interval.

In order to better understand the significance of content of videos and accumulated subscriber base, it is essential to build up a more comprehensive and multidimensional system to evaluate the relationships between video content, historical view data and YouTuber future popularity [16]. To make analysis clearer and easier, we separate the whole prediction into two parts: growth and sustainability. Growth is a reaction to effects of latest uploaded videos in a short time, which is to help to better understand the reasons behind the increase of subscribers. Therefore, with further analysis and assessment, predictor can provide more detailed and specific suggestions for YouTubers and make them aware of strengths and weaknesses in their videos so that they can modify the content and form of their next videos. This analysis would concentrate more on video tag, types and content of videos as significant features as well as instant user data. Sustainability, however, is built up in the aim of building up long-term prediction, so we preprocess historical subscriber data and frequent content data for a better prediction of future subscriber change trends. And finally, we make a real-time updating service website to provide simple predictions for YouTubers based on our modified regression model and updating big data. The website contents include subscriber growth and sustainability prediction, as well as giving out simple suggestions on making YouTube videos.

Various experiments were conducted to find the most suitable feature matrix, the top-performing models and relevant well-matched configuration so that we can establish a relatively complete and efficient prediction system to evaluate the current management situation of youtubers and preview their short or long-term development [19]. For both the growth of the subscribers and the maintenance of their development, we compared several regression models including linear

regression, polynomial regression, SGD, random forest, SVD, Bayesian ridge, gradient boosting and k-neighbors. By testing and modifying each model to figure out their own best prediction results based on the same training data sets, we compared these models for over 100 trials and calculated their mean accuracy based on the R2 score of models, which explains the degree of connection and similarity between prediction results and testing labels. And finally, the random forest regression offered the optimal prediction for both two parts. Hence, we selected the random forest regression model to further research on. We chose different maximum depth and different numbers and types of decision trees inside the random forest to test and compare their effects. The result is that more maximum depth and more decision trees will enhance the final prediction accuracy on the growth and sustainability of subscribers for specific youtubers. Also, we conducted extra experiments to explore the relevant weights of each feature in the data matrix. We found that the current level of subscribers will not significantly influence the increase of subscribers within the short term for youtubers who have already accumulated a certain number of followers. And the rise of short videos, which are born from TikTok and different from traditional YouTube videos, will obviously influence the increase of subscribers either for short term or long term.

The rest of the paper is organized as follows: I will first introduce the possible challenges that we met during the process of experiments and designing the sample. Then based on the intention to solve these challenges, we proposed relevant solutions and generated them into methods to systematically solve the problems. The following experiment was divided into two parts. The first experiment part focused on the datasets designed for instant growth, while the second part applied other datasets preprocessed with average value of each feature in videos to predict its future sustainability. Also, we introduced some other similar studies and related work, and analyzed their strengths and weaknesses. Finally, conclusion remarks along with future possible improvement work will be described.

2. TECHNICAL CHALLENGES

In order to develop the system, a few technical challenges have been identified as follows.

2.1. The Selection and Preprocessing of YouTube Video Features

In order to find the key to the growth and sustainability of a YouTuber's account, any data related to YouTube channels should be collected and analyzed. It is tough as that data of diverse dimensions must be preprocessed before building up models. From the standpoint of video content, it is complicated for a specific video published by a certain YouTuber to be defined and classified. With rich tags and topics, how to correctly select and filter information of videos requires more work since certain information may have a close relationship with the latest news or fashion that would greatly affect popularity of the videos. But superfluous data would also lead to unessential workload and disturb the final prediction. Meanwhile, it is also necessary to analyze the effective interaction between audience and YouTubers. It can be multiple forms, and how to transform them into a uniform structure for further research needs to be considered. To solve the challenge, we did a huge amount of manual preparation in the early stage of the program to make enough classical samples covering different levels of YouTubers, which also teaches automatic data collecting tools the way of selecting data so that we would be able to obtain larger datasets. We select tags and topics that can be considered as more general representations in videos, and we use one-hot encoding methods to make it structured. All the data related to interaction from the audience would be preprocessed in the base of YouTubers present subscriber number to reach a more rational comparison between YouTubers.

2.2. Selection of Models and Parameters

The real analysis on YouTube data is much more complicated than expected, which means that different values for certain features in datasets for YouTubers would not definitely or obviously decide the final growth scale. Since we separate these YouTubers into different levels, rules and parameters also must be modified to realize better models for different groups of data. Also, same models with different configurations and parameters would have diverse accuracy, which are quite sensitive. To solve this problem, we made several experiments to compare between different models and the same model with different configurations and parameters, so that we would be able to find the best model to predict the growth rate and sustainability of the YouTuber imported.

2.3. The Bias Problem

The third challenge is the bias problem. AI models learn by the dataset that is used when the model is training. However, if there is not enough data the model can be underfitting and predicts the incorrect output. There exist different resources that can provide YouTube Video trends data, however many of them have a limited amount of requests. For example, web scraping websites encounter some issues, some web pages have a limited number of requests that users can perform. To solve the issues, we use APIs that are more flexible when we need to fetch data.

3. SOLUTION

SmartYoutuber is a popularity prediction system based on big data analysis and data update strategy. By applying multi-dimensional evaluation for existing popular YouTubers who are rapidly accumulating a great number of subscribers, we established several data matrices to describe the YouTubers' video quality and their ability to attract new viewers and maintain its original subscriber group. The matrix has a dynamical estimation towards the appealing points of these videos and their relationship with current fashion trends, which ensures their growth and sustainability. With the data matrix, we tested different models to predict the situation of subscribers of YouTubers who are on the rise or decline, in order to help them develop present video business or prevent from popularity loss. To achieve such a series of goals, our system consists of four main components:

- a data processor to web scraping distinct resources
- a multi-dimensional data structure for storing the content features of YouTube videos and subscriber statistics
- evaluate different weights of the features and build up reliable prediction models
- a real-time updated web platform for YouTubers to examine their present state of operation and predict their future trend

3.1. Back-End System

The back-end consists of data processing, data storage and Artificial Intelligence engine. It is written in the Python programming language[6]. For the data processor, we collect data from Twitter, YouTube and Google Trends websites[7][8][9]. We perform web scraping to distinct websites and select the key aspect features that are used for our predictor. After extracting the data for different resources, we store the data in our database. For a multi-dimensional data structure for storing the content features of YouTube videos and subscriber statistics, we use Firebase Firestore. Firebase is a cross-platform SDKs that provides hosting services (e.g.,

analytics, authentication, and storage) for any type of application such as Android, IOS, Unity and Website[4].

To evaluate different weights of the features and build up reliable prediction models, we utilize a Python machine learning library called Scikit-learn and also known as sklearn[5]. Scikit-learn consists of supervised and unsupervised machine learning algorithms and some features are classification, regression and clustering algorithms including support-vector machines, random forests, and k-means. The classification algorithms that are used to evaluate the predictors are Linear, Polynomial, Random Forest, Bayesian Ridge, Stochastic gradient descent (SGD), Gradient Boosting, Support Vector Regression (SVR) and Kneighbors.

3.2. Front-End System

The frontend of the application consists of a website that displays the video details, video analysis, video trends and video recommendation. The website is designed with a bootstrap framework. Bootstrap is a feature-packed front-end toolkit, HTML, CSS, and JavaScript framework, for developing websites[10]. To display the YouTube videos trends and other information the system sends requests to the server which analyzes the video and data to predict the growth of number of subscribers.

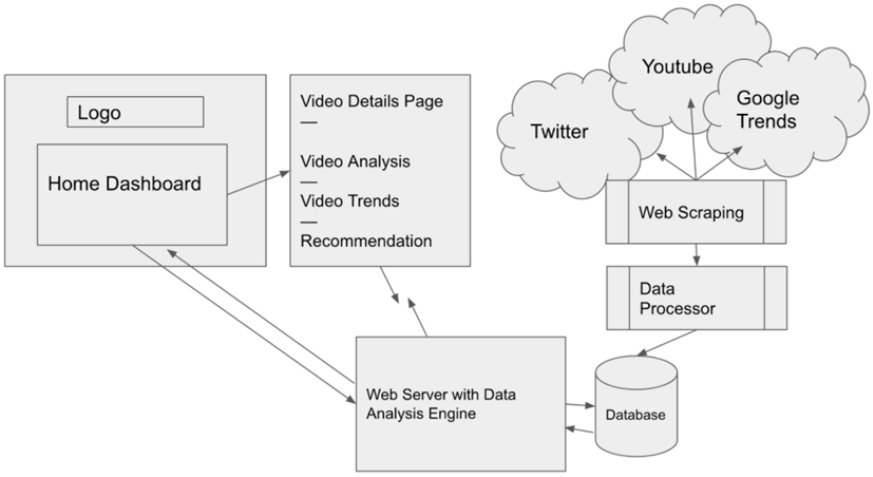

Figure 1. An Overview of the Web-based Data Analytics Video Platform

3.3. Generating Reliable Datasets

In order to generate initial datasets for original training and prediction, we listed a series of possible normal features of YouTube videos including views, comments, likes, tags, topics, types and so on. The targets of datasets are set in a daily growth level for the growth part, while the sustainability is evaluated in a preprocessed way based on the ratio of new average growth over the general number of subscribers. All the data in the growth part is selected and calculated into average values in a range of 7 days, and the data for the sustainability are established from average data in a month or longer. Moreover, we catch the rise of live stream and shorts, which impacts the number of views and increase of subscribers in an unexpected way. Therefore, we treat them as a special feature to analyze the exact influence of short videos and live stream on popularity of YouTubers. After collecting data sets manually, we applied automatic tools to keep accumulating larger amounts of data for further experiments and management of the prediction website we built in step 3.

3.4. Comparison and Selection of Regression Models

With carefully modified and collected datasets, we compared the practical accuracy of different regression models in this specific application scenario. After normalization of the original datasets, we set a suitable training and testing scale of datasets and test the results of models under the same scale to obtain compatible differences. The parameters and configuration of each model were modified and adjusted repeatedly in order to realize the best prediction results for different models, including traditional linear regression models, random forest regression model, Bayesian Ridge regression model, decision tree regression model and so on. Also, to better determine the regression model, we picked up different combinations and diverse weights of features to evaluate their influence on video popularity. Based on the simulation results of these features, we create simulation datasets with adjusting standard presets to test our models.

3.5. Establishment of the Prediction Website and the Application of the Regression Model

We established a service website for YouTubers to predict their future growth and sustainability. This website is mainly based on the automatically collected YouTuber data, which is real-time updating to periodically modifying models. The model we obtained in the second part would be applied here and repeatedly trained. Based on the prediction results, the website should provide simple suggestions for the specific YouTubers about the strengths and weaknesses in their videos.

4. EXPERIMENTS

4.1. Experiment 1

Due to the complexity of content and subscriber data of YouTubers, it is hard to understand the inner relationship between features and figure out how they decide and predict the future growth and sustainability. Hence, in this experiment, we aim to find the best regression algorithm to predict the future daily growth of subscribers based on normalized feature datasets, including video types, current subscriber number, whether the video is short video and number of views, comments and likes in recent videos.

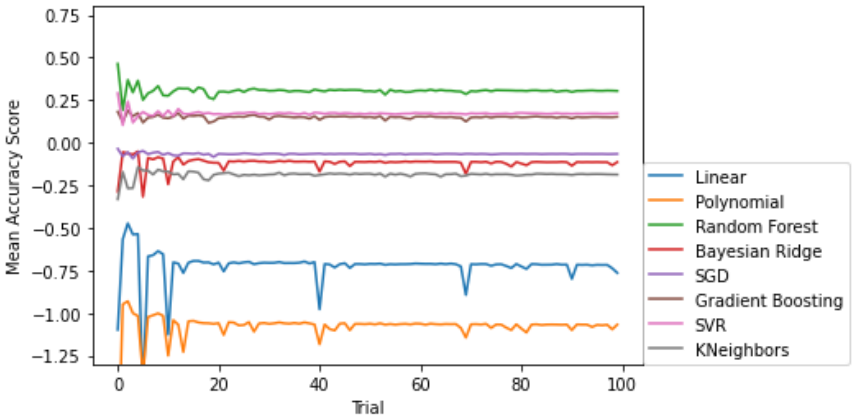

Figure 2. Experiment 1.1 - accuracy of different models

From the result of comparison, there is an obviously distinctive curve of random forest regression. With over 100 trials, the random forest regression consistently outperformed the rest, whose final average accuracy score maintains over 0.35, which is 0.15 greater than the SVR and gradient

boosting. By contrast, polynomial and linear regression exhibit unideal prediction accuracy for this experiment.

According to the previous experiment, we select random forest regression as the ideal model for further research. Based on the characteristics of random forest regression, we modify its maximum depth and the number of decision trees(estimators) inside the random forest to observe relevant changes in the final prediction accuracy in the same level of data sets and trials.

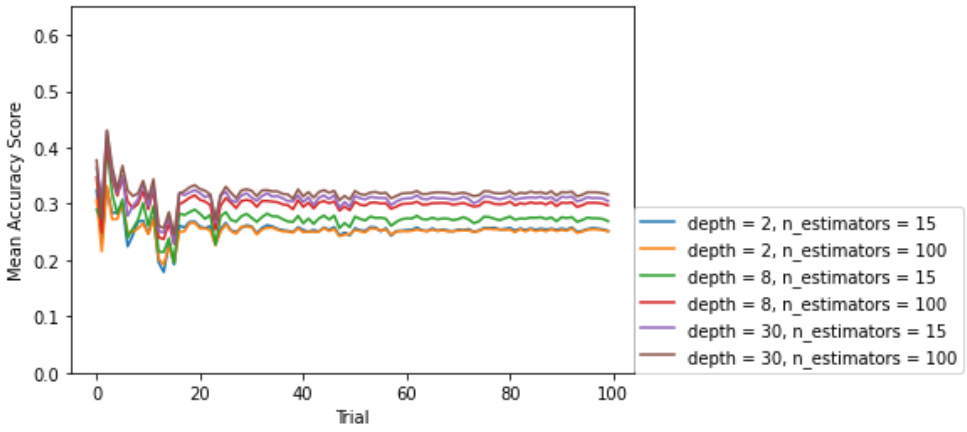

Figure 3. Experiment 1.2 - Random Forest Regression with different parameters

From the figure above, with greater depth and higher number of estimators, this regression model exhibits better prediction results, which is over 30% higher than regression with low depth and a smaller number of estimators. This helps to improve the performance of our final model to predict the future development of youtubers.

In order to find specific weights of different features on the prediction results, we tested the effect of two features, which are current subscriber number and whether videos are shorts, so that we can observe the significance of these features in the evaluation.

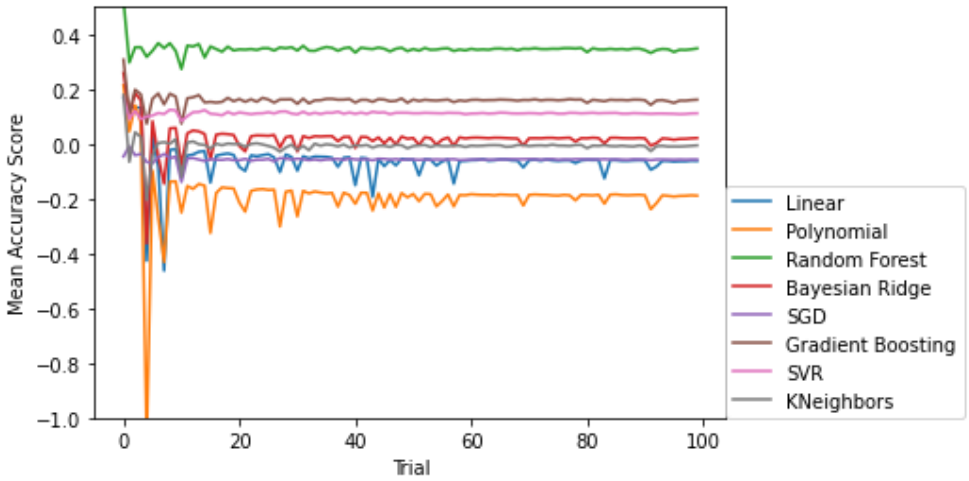

Figure 4. Experiment 1.3 - accuracy of different models without current subscriber number

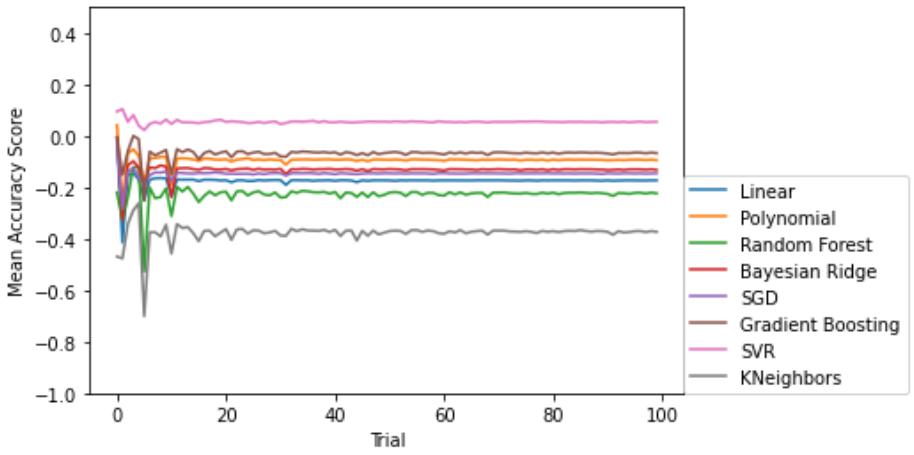

Figure 5. Experiment 1.4 -accuracy of different models without checking if shorts

By removing different features, we can figure out that the influence of current subscriber level on the final prediction is relevantly little when removed, which means that the instant growth of YouTubers does not rely on their present follower number. Attention that this is a conclusion when we compare different youtubers who have accumulated a certain number of subscribers. It might not be suitable for small youtubers. However, the influence of shorts is greater. Without checking if the videos are shorts or not, the final prediction is greatly affected, and accuracy becomes much lower for all the models. This reminds of the rise of shorts in most video platforms. The appearance of shorts tremendously increases people to watch more videos and subscribe to more channels than before.

4.2. Experiment 2

In order to analyze the sustainability of a specific YouTuber, we select YouTuber datasets with quite different features scales. We want to make the average output of YouTuber more clear and understand the inner characteristics of their videos over a rather long period. Then it would help us to build up a more suitable regression model to predict the future subscriber situation within a certain long time. We applied procedures similar to the previous experiments to test different regression models and modified their configuration to improve the chosen model.

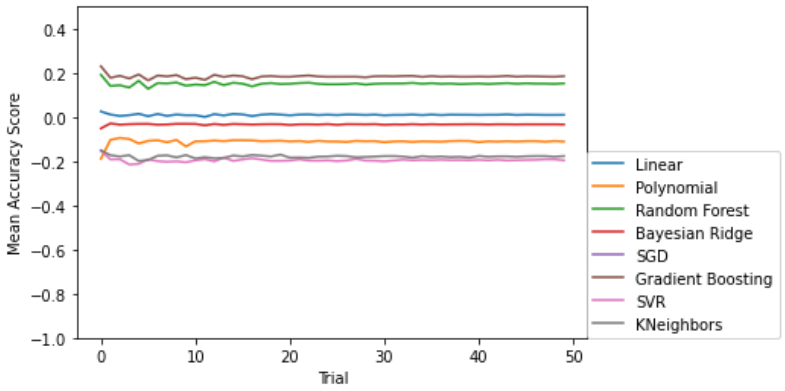

Figure 6. Experiment 2.1 Mean accuracy of different models for sustainability

From the figure we obtained, we could find out that the random forest still applies to the prediction of sustainability. Meanwhile, gradient boosting regression also performs well in this experiment to give even better prediction results.

Then we selected gradient boosting regression for further study. We modified its max depth of the model with different n_estimators to observe the changes of prediction accuracy.

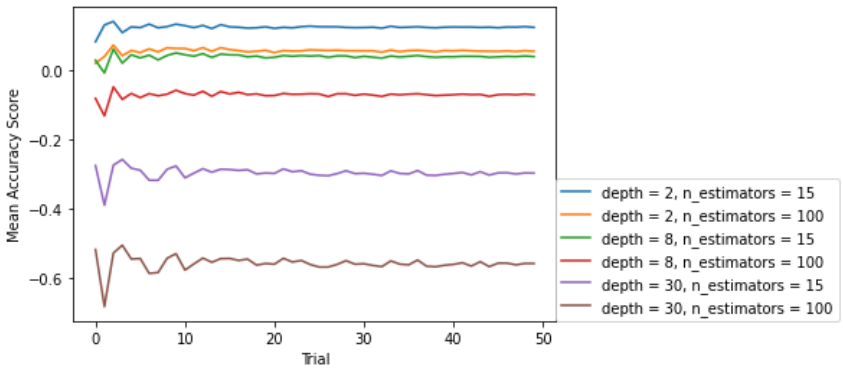

Figure 7. Experiment 2.2 Gradient boosting regression with different parameters

This experiment shows that less depths and more n_estimators would make the result more close to the real results.

5. RELATED WORK

Henrique, J. et al propose two models to predict the future popularity of web content and they selected YouTube as the classical example to research on[1]. The basis of their research is the historica user data accumulated by early popularity measures. The goal of their research is to design and evaluate the possibility of advertisements and recommendation systems, which finally helps to rapidly increase subscriber growth by adjusting these systems. Their model would focus more on the data of users rather than the exact content of videos, and the models would fit the videos with tremendously high popularity in the early stage well. Combination of both user information and video features is not covered in this paper, and their analysis concentrates more on the popularity of certain videos rather than YouTubers or YouTube channels. Our model would inspect more dimensions of YouTubers, trying to include all the possible influence on popularity to help YouTubers to manage their accounts instead of just raising the popularity of certain videos.

Changsha, Z. et al predict YouTube video popularity in a different way avoiding using historical data before prediction[2]. They tried to make their prediction running much faster and suit a more complicated network structure, so they proposed a new lifetime aware regression model to predict future popularity based on recent periodic popularity data. Their prediction is established in a more recent future and short intervals are covered. We would like to investigate the prediction for both the recent future situation and the long-term increase of subscribers of YouTubers. So we would focus more dimensions and cover more information to build up our regression model.

Barjasteh, I. et al present a study of how to analyze, measure and compare aspects of YouTube trending videos[3]. In their study they collected and monitored related statistics of 8,000 YouTube videos over nine months and inspected the profile of users who upload trending videos.

Their results show that Trending videos and their channels and YouTube content that has not been labeled as trending have distinct statistical attributes when compared to each other and reveal a highly asymmetric directional relationship among different categories of trending videos. Similarly to Barjasteh, I. et research, we are monitoring and analyzing the YouTube data to observe the key aspects that influence YouTube video trends. However, our research proposes an Artificial Intelligence engine to predict the growth of future subscribers.

6. CONCLUSIONS

To conclude, our main goal of the experiments is to predict the future growth of subscriber number of youtubers and the long-term sustainable development of youtubers. In order to realize the prediction, we tested different regression models to trace the weights of diverse features of videos and youtuber data and compared their performance in this application of predicting close future increase of subscribers and how they maintain the current subscriber level. We figured out that the random forest regression is the outperforming one to achieve the best average accuracy in over 100 trials. Also, based on the discovery of the comparison of models, we adjust the depth and estimator number to monitor the optimization of the random forest regression model in prediction. We found that more depth and more estimators will improve the final accuracy. Therefore, a random forest regression with high depth and more estimators will predict the future growth and sustainability of subscriber number well. With this model and such a suitable configuration, we can simply guess youtuber's development to some degree and give specific youtubers suggestions on their strengths and weaknesses to help them improve their video quality and the management of their YouTube channels.

Despite the observation and prediction results we have already obtained, there are still some limitations in my method. First, the accuracy of my model remains to be raised. There remain a lot of much more mature and optimized models to optimize the accuracy of the experiments. Higher levels of machine learning algorithms allow more sample features to be accessed and put into evaluation. Secondly, the data sets are still not wide enough and big enough. The inner characteristics that would affect the future growth and sustainability of subscribers need to be included and processed to make the prediction more correct. Finally, short-term growth of subscribers may be influenced by sudden unexpected events or political effects, which may deactivate the prediction.

Future improvement can focus on the inner connection between social activity and video quality. Channel popularity can be associated with social current fashion to make predictions keep in pace with the latest news, which may help to establish more accurate predictions and to accumulate followers in an efficient way.

REFERENCES

[1] Pinto, Henrique, Jussara M. Almeida, and Marcos A. Gonçalves. "Using early view patterns to predict the popularity of YouTube videos." Proceedings of the sixth ACM international conference on Web search and data mining. 2013.

[2] Ma, Changsha, Zhisheng Yan, and Chang Wen Chen. "LARM: A lifetime aware regression model for predicting YouTube video popularity." Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. 2017.

[3] Barjasteh, Iman, Ying Liu, and Hayder Radha. "Trending videos: Measurement and analysis." arXiv preprint arXiv:1409.7733 (2014).

[4] Chatterjee, Nilanjan, et al. "Real-time communication application based on android using Google firebase." Int. J. Adv. Res. Comput. Sci. Manag. Stud 6.4 (2018).

[5] Buitinck, Lars, et al. "API design for machine learning software: experiences from the scikit-learn project." arXiv preprint arXiv:1309.0238 (2013).

[6] vanRossum, Guido. "Python reference manual." Department of Computer Science [CS] R 9525 (1995).
[7] Mandloi, Lokesh, and Ruchi Patel. "Twitter sentiments analysis using machine learninig methods." 2020 International Conference for Emerging Technology (INCET). IEEE, 2020.
[8] Szilagyi, Istvan-Szilard, et al. "Google trends for pain search terms in the world's most populated regions before and after the first recorded COVID-19 Case: infodemiological study." Journal of medical Internet research 23.4 (2021): e27214.
[9] Kready, Joseph, et al. "YouTube data collection using parallel processing." 2020 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW). IEEE, 2020.
[10] Wehrens, Ron, Hein Putter, and Lutgarde MC Buydens. "The bootstrap: a tutorial." Chemometrics and intelligent laboratory systems 54.1 (2000): 35-52.
[11] Hu, Yuheng, Lydia Manikonda, and Subbarao Kambhampati. "What we instagram: A first analysis of instagram photo content and user types." Eighth International AAAI conference on weblogs and social media. 2014.
[12] Thelwall, Mike, Pardeep Sud, and Farida Vis. "Commenting on YouTube videos: From Guatemalan rock to el big bang." Journal of the American Society for Information Science and Technology 63.3 (2012): 616-629.
[13] Elvers, Todd, and Padmini Srinivasan. "What's trending?: mining topical trends in ugc systems with YouTube as a case study." Proceedings of the Eleventh International Workshop on Multimedia Data Mining. ACM. Vol. 4. 2011.
[14] Cheng, Xu, Cameron Dale, and Jiangchuan Liu. "Statistics and social network of YouTube videos." 2008 16th Interntional Workshop on Quality of Service. IEEE, 2008.
[15] Ding, Yuan, et al. "Broadcast yourself: understanding YouTube uploaders." Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference. 2011.
[16] Susarla, Anjana, Jeong-Ha Oh, and Yong Tan. "Social networks and the diffusion of user-generated content: Evidence from YouTube." Information systems research 23.1 (2012): 23-41.
[17] Tan, Zhiyi, et al. "A novel time series approach for predicting the long-term popularity of online videos." IEEE Transactions on Broadcasting 62.2 (2016): 436-445.
[18] Ma, Changsha, Zhisheng Yan, and Chang Wen Chen. "Forecasting initial popularity of just-uploaded user-generated videos." 2016 IEEE International Conference on Image Processing (ICIP). IEEE, 2016.
[19] Yu, Honglin, LexingXie, and Scott Sanner. "The lifecyle of a YouTube video: Phases, content and popularity." Proceedings of the international AAAI conference on web and social media. Vol. 9. No. 1. 2015.
[20] Figueiredo, Flavio, et al. "On the dynamics of social media popularity: A YouTube case study." ACM Transactions on Internet Technology (TOIT) 14.4 (2014): 1-23.