YOUTUBE ADVIEW PREDICTION USING REGRESSION MODEL

Author-1: **Mr. K.Krishna Reddy** Associate Professor, Department of Computer Science Engineering, Sreyas Institute of Engineering and Technology, Hyderabad, Telangana, India.
krishnareddy@sreyas.ac.in

Author-2: **M.Pranavi Reddy** Department of Computer Science Engineering, Sreyas Institute of Engineering and Technology, Hyderabad, Telangana, India.
Pranavireddymallamgari@gmail.com

Author-3: **Kallem Deepa Rishika Reddy** Department of Computer Science Engineering, Sreyas Institute of Engineering and Technology, Hyderabad, Telangana, India.
reddyrishika3687@gmail.com

Author-4: **Neha Golekar** Department of Computer Science Engineering, Sreyas Institute of Engineering and Technology, Hyderabad, Telangana, India.
nehagolekar12@gmail.com

Author-5: **Akarapu Pradeep Kumar** Department of Computer Science Engineering, Sreyas Institute of Engineering and Technology, Hyderabad, Telangana, India.
pradeepakarapuapk@gmail.com

ABSTRACT

The main goal is to create a machine learning regression that can estimate the number of YouTube adviews based on other parameters. Advertisers on YouTube pay content creators based on how many times their ads are viewed and clicked. They want to estimate the adview based on other metrics like vidid, adviews, published, duration, views, comments, likes etc. CSV files are utilised for training and fitting, and then they are tested to get the best outcomes. As a result, the goal of the project is to train multiple regression models and select the best one for predicting the number of adviews. We validate datasets and packages like Numpy, Pandas, and Sklearn for their form and data type. Also, we visualise and clean the dataset followed by transforming attributes into numerical values. Using different regression algorithms, normalise the data and separate it into training and test sets. The model is saved and used to predict on the test set. To acquire better outcomes, data or information must be improved, filtered, and cleansed before being fed in based on numerous criteria.

INTRODUCTION

Launched in May 2005, YouTube allows billions of people around the world to discover, watch, and share originally-created videos. YouTube allows individuals all around the world to interact, educate, and inspire one another and acts as a distribution platform for original content creators and advertisers, both large and small. The video view count is an important metric for determining a video's popularity or "user engagement," as well as the parameter by which YouTube compensates the content creators. This research aims to forecast how many Ad

views a specific video will receive in order to promote a specific deal or brand. We use a dataset to first train the model. The file train.csv contains around 15000 YouTube videos, which contains metrics and other information Number of views, ad views, likes, dislikes, and comments are among the indicators. Aside from that, the date, duration, and category of the publication are all given. The metric number of ad views which is our target variable for prediction is also available in our csv file. Various plots are used in order to predict the value needed. The data is refined and cleaned before feeding in the algorithms for better results. There are Youtube videos where there is a lot of interaction among the users and the content creators when compared to other social media platforms.Taking this as an advantage, it's an easy place yet affective to market or promote a particular item in this technological world. So in order to predict how many ad views a particular video would get when marketed in a particular published year, we created a Machine Learning project regarding prediction techniques Where a lot of previous data is collected in the form of 2 csv files named train andtest The train dataset has around 15000 records where we divide it into 80 20 in the further or final stages.

OBJECTIVE

The study of popularity of YouTube videos based onmeta-levelfeatures is a challenging problem given the diversity of users, sponsorships and content providers. To define the popularity of YouTube videos, several types of parametric models are utilised, with the view count time series being used to estimate the model parameters.

For instance, ARMA time series models, which are multivariate linear regression models, have been used to predict future video view counts based on previous view count time series.

Specifically, the project finds that youtube sales within a particular range are more popular compared to other users.

Here, we find the best of the regression models and hence try to predict the most approximate and expected values.

LITERATURE SURVEY

The analysis of YouTube video popularity at the meta-level features is a challenging problem given the diversity of users, sponsorships and content providers. To define the popularity of YouTube videos, several types of parametric models are utilised, with the view count time series being used to estimate the model parameters.

For instance, ARMA time series models, which are multivariate linear regression models, have been used to predict future video view counts based on previous view count time series.

Specifically, the project finds that YouTube sales within a particular range are more popular compared to other users.

Here, we find the best of the regression models and hence try to predict the most approximate and expected values.

EXISTING SYSTEM

Regression models like decision tree, linear ,support vector are used to predict the number of adviews and gives the accurate results. This prediction is based

on the metrics such as likes, dislikes, comments etc.
DRAWBACK: By using this models we cannot predict the exact values but we can predict the accurate values

PROPOSED SYSTEM

We simply utilise one regression model, the support vector, which provides better prediction accuracy that is support vector gives less number of errors among other regression models when we test the data for actual predictions and this system helps in predicting the adviews of a particular video which would helps in marketing a particular sale or a brand .after training the data using regression models we test the models by giving some test data from that we predict the actual model that gives less errors for adview predictions . we can also use regression models apart from these to test the data but from those regression models we consider only one model which gives less number of errors.

ADVANTAGE: This aids in the prediction of adviews for a specific video, which aids in the promotion of a specific product or brand.

IMPLEMENTATION

(1) Import the datasets and libraries, then double-check their shape and datatype.
In the first step we need import libraries, dataset and analysed the data by checking its shape, data types.
(2) Other necessary changes include converting characteristics to numerical values.
As we've seen, the data is now in object format.
hence, we have converted the data in float for further process and evaluation and also manipulate time into seconds and date into numeric format and also split the date into year, month and day for further analysis.

- Convert views, likes, dislikes, comment data into numeric using panda.to numeric () with errors="coerce", so that if it is not able to convert to numeric it converts to NULL.
- Converting published date into numeric and splitting it into year, month, day.
- Converting time into seconds' format.
- Converting or labelling the category for faster and easy analysis

(3) Clean the dataset by removing missing values and other things.
And at last, remove the missing values such as null or any other miscellaneous data so that they do not interfere with further process.

- Drop or remove null characters and unnecessary data.
- Rearrange the columns so that it is easy to split while training the data.

(4) Visualise the dataset using plotting using heatmaps and plots.
You may also look at the data distributions for each attribute.
Now for further analysis I have by plotting heatmap and
different plots:

- Year vs Total Ad views

In this plot we can observe plot of total number of ad views in each year and we can observe the increasing trend in each year.

- Year vs view

In this plot we can observe the scatter plot of ad views in each year from 2005 to 2017 and can observe only one video to be above 2000000 and hence we can exclude it before training the data.

- Category vs No. of videos

In this plot we can observe a greater value for category 3 than others categories.

(5) Data should be normalised and In the right ratio, divide the data into training and test sets.

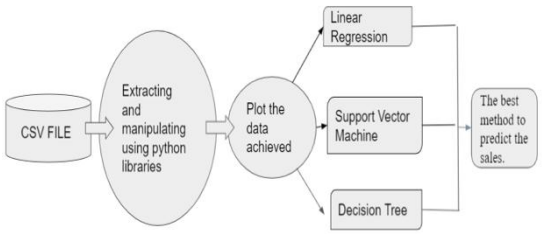
- Separate the data into training and test data.
- Develop a function to calculate mean absolute error, mean square error, and mean square root error.

(6) Use linear regression, Support Vector Regressor and Decision Tree for training and get errors.

- Train the data for each respective model and make a note of errors

ARCHITECTURE

The system architecture is used to expose the process of finding which regression model contains the least errors



CONCLUSION

- This project explores how to use different libraries like numpy, sklearn, pandas and matplot in python.
- Predicting the average sales for a particular XYZ company when it tries to market its product through Youtube is finally achieved

- Automation of these results will reduce much effort and help the sales/marketing team easily predict the scope of customers at the end of the day.

REFERENCES

[1] D. Agarwal, B.-C. Chen, and P. Elango. Spatio-temporal models for estimating click-through rate. Pages 21–30 in Proceedings of the 18th International Conference on the World Wide Web.ACM Press, 2009.

[2] R. Bekkerman, M. Bilenko, and J. Langford. Scaling up machine learning: Parallel and distributed approaches. 2011.

[3] Z. A. Zhu, W. Chen, T. Minka, C. Zhu, and Z. Chen.The application of a revolutionary click model to online advertising.Pages 321–330 in the Proceedings of the third ACM international conference on Web search and data mining. 2010 ACM.

[4] "Computational Programming with Python," Hans Petter Langtangen • SveinLinge

*Available:*https://link.springer.com/book/10.1007/978-3-319-32428-9

[5] "data-analysis-with-Numpy-and-Pandas," says Curtis Miller. Curtis Miller's Hands-On Data Analysis with NumPy and Pandas is now available for download (b-ok.asia)