Professional School
ISO 9001 : 2000 Organisation

PREDICT YOUTUBE VIDEO ENGAGEMENT

Problem Statement, Variable Description & Deliverables

YouTube Video Views Prediction – Objective & Deliverables

Problem Statement:

EngageMark Analytics is struggling to help content creators and brands understand how their videos will perform. They want to predict things like how many views, likes, and new subscribers a video will get, but it's hard to figure out because audience behavior is so unpredictable. With so much data to analyze, it's challenging to find clear answers. The company needs a simple and reliable way to predict video performance and figure out what makes people engage with videos, so they can give better advice to their clients and help them succeed.

Data Description:

The dataset contains information about YouTube video performance and viewer engagement. Each row represents the engagement data for a specific video on a particular day. Below are the key data fields and their descriptions:

Video Information

- **video_id**: A unique identifier for each video.
- **day**: The date on which the engagement data was recorded (format: YYYY-MM-DD).

Viewer Engagement Metrics

- **views**: Total number of times the video was viewed on that day.
- **redViews**: Number of views where users interacted with premium content (e.g., YouTube Red).
- **comments**: Total number of comments posted on the video.
- **likes**: Total number of likes the video received.
- **dislikes**: Total number of dislikes the video received.
- **shares**: Number of times the video was shared with others.

Playlist Metrics

- **videosAddedToPlaylists**: Number of times the video was added to playlists.
- **videosRemovedFromPlaylists**: Number of times the video was removed from playlists.

Watch Time Metrics

- **estimatedMinutesWatched**: Total estimated minutes spent watching the video.
- **estimatedRedMinutesWatched**: Estimated watch time for premium content interactions.

- **averageViewDuration**: The average length of time viewers spent watching the video (in seconds).
- **averageViewPercentage**: The percentage of the video watched on average.

Annotations and Cards Metrics

- **annotationClickThroughRate**: Percentage of viewers who clicked on video annotations.
- **annotationCloseRate**: Percentage of viewers who closed annotations.
- **annotationImpressions**: Number of times annotations were shown to viewers.
- **annotationClickableImpressions**: Number of times clickable annotations were shown.
- **annotationClosableImpressions**: Number of times annotations with close options were shown.
- **annotationClicks**: Total number of clicks on annotations.
- **annotationCloses**: Total number of times annotations were closed.
- **cardClickRate**: Percentage of viewers who clicked on video cards.
- **cardTeaserClickRate**: Percentage of viewers who clicked on teaser cards.
- **cardImpressions**: Total number of times video cards were shown.
- **cardTeaserImpressions**: Total number of times teaser cards were shown.
- **cardClicks**: Total number of clicks on video cards.
- **cardTeaserClicks**: Total number of clicks on teaser cards.

Subscriber Metrics

- **subscribersGained**: Number of subscribers gained on the day due to the video.
- **subscribersLost**: Number of subscribers lost on the day due to the video.

Deliverables:

- **Understand the data variables properly**. Check the variable description to understand the data properly.
- **Clean the data:** Clean the data, that is, fill the missing values (if any), treat the outliers (or odd values), etc. Ensure each variable's data is as per the nature of the variable (e.g. – Date field should contain only date values – can extract year, month and day of the week, and numeric column should be formatted as numeric, etc.).
- **Conduct EDA (Exploratory Data Analysis) on the cleaned Data:** Summarize, explore the data and then decide your strategy. Make note of any important assumptions that you make.
- **Uni-variate and Bi-variate Analysis:** Check the distribution of independent variables and

also compare them with the dependent variable.

- **Feature Engineering:** Create new meaningful features based on the existing features by applying some aggregation functions on them.
- **Hypothesis Testing:** Hypothesis testing in statistics is a way for you to test the results of a survey or experiment to see if you have meaningful results. You should give a brief summary of the data and a summary of the results of your statistical test. In the discussion, you can discuss whether your initial hypothesis was supported or refuted.
- **Identify the most important variables (or data parameters) that affect the final decision:** Identify the impact of each variable on the final result graphically (correlation / scatter plots, regression plots, etc.). Keep those variables that affect the final outcome.
- **Develop and Validate Samples:** Divide samples into 2 parts: Development Sample (70%) & Validation Sample (30%). Build your analysis model using the Development Sample, and validate it on the validation sample and then predict on the test sample.
- **Model Building:** Analyze the dependent variable and decide which technique out of regression or classification to use and hence build the model.
- **Improving model accuracy:** We know that machine learning algorithms are driven by parameters. These parameters majorly influence the outcome of the learning process. So, find the optimum value for each parameter to improve the accuracy of the model and repeat this process with a number of well performing models.
- **Model Comparison:** Comparing each model with other similar models and then choose that model which gives highest accuracy. But it is not necessary that higher accuracy models always perform better (for unseen data points). So, to find the right accuracy of the model, you must use a cross validation technique before finalizing the model.