SAP TAP DAY Challenge

Team 2 member: Bing Wu

Team Idea Overview

Objective: Forecast how popular will a Youtube video be.

Product use case: When a user open the product, there will be a few forms for the user to fill about the relevant information which are listed below. After the user clicks submit, the user will get a prediction of the estimated number of views based on the information the user input and potential recommendation for the video.

The user input we choose:

- Title
- Category
- Tags
- Comments disabled
- Ratings disabled

Architecture of my Contribution

1. Define challenge objective and formulate the challenge to a data science problem.
   a. I need to use the USvideo data to predict the number of views a video will have with the user input information
2. Load the data and check the basic information about the dataset such as the type of variables and number of data points
3. Perform feature engineering such as create new variables
   a. Calculate the length of the title, the number of tags
   b. Create a new dataframe to be used in prediction modeling
4. Build a random forest regression model to predict the number of views a video will have

Demo of contribution

Let’s have look at the notebook

What You Would Do Next?

1. I would try to acquire more data such as how long the video is, how many videos does this users created before, how many followers this user has, how long since the user start the channel.
2. I would redefine the the way we measure popularity using the possibility of being popular or not.
    a. Popular definition: if the average number of views in a day is great than the median of all number of views, we define this video as popular.
3. I would update the category_id variable to the real category name because that makes more sense for human to understand
4. I would use text analysis on tag, and description to explore the what kind of effect do the context of these variabel have on the prediction
5. I would try to explore data in different countries instead of focusing on US data
6. I would try to build more models such as linear regression, decision tree, and Boosting
7. Rewrite the notebook which can be used for developers and designers later
8. Explore more about the relationship between X variables and Y variable, giving recommendation to the Youtuber

Thank you!

Bing Wu