Estimating Views

YouTube is a video-sharing website owned by Google. It allow users to upload, view, like, dislike, comment, and report videos. The videos are categorized and also have tags that are related to the video's content. There are billions of videos uploaded and many have up to billions of views. Therefore, in this problem, we will focus only on data from users in the US and we would like to understand the factors that influence the amount of views per video.

To derive insights and answer these questions, we take a look at a dataset containing the top trending YouTube videos in the US throughout several months. Our data has a total of 15 columns and 40003 observations, split across a training set and a test set. Each observation corresponds to a different video.

Training data: youtube_train.csv
Test data: youtube_test.csv

Here is a detailed description of the variables:

- **video_id**: A number that uniquely identifies the video
- **title**: The video's title
- **views**: The amount of views the video has
- **likes**: The amount of likes the video has
- **dislikes**: The amount of dislikes the video has

- **comment_count**: The amount of comments the video has
- **logviews**: The natural logarithm of the **views** variable.
- **loglikes**: The natural logarithm of the **likes** variable.
- **logdislikes**: The natural logarithm of the **dislikes** variable.
- **logcomments**: The natural logarithm of the **comment_count** variable.
- **category_id**: A number that uniquely identifies the video's category
- **category**: The title of the video's category
- **tags**: The amount of tags the video has
- **publish_month**: The month that the video was published (1-12)
- **trending_month**: The month that the video trended (1-12)

Problem 1 - Exploratory Data Analysis

0.0/5.0 points (graded)
Load youtube_train.csv into a data frame called train.

How many rows are in the training dataset?

Answer: 30002

What is the average amount of likes per video in the training dataset?

Answer: 75461.28

What is the category of the video with most views in the training set?

- [x] Music ✔
- [ ] Sports
- [ ] Comedy
- [ ] Entertainment
- [ ] Shows
- [ ] Film and Animation

Which category has the least amount of dislikes in the training set?

- [ ] Education
- [ ] Pets and Animals
- [x] Shows ✔
- [ ] People and Blogs
- [ ] Comedy
- [ ] Science and Technology
- [ ] Sports

In the training set, out of the videos with at least 1 million likes, how many have at least 100,000 comments?

Answer: 184

Submit You have used 0 of 2 attempts

Answers are displayed within the problem

Problem 2.1 - Simple Linear Regression

0.0/3.0 points (graded)

For the rest of this problem, we will be working with log(views), log(likes), log(dislikes), and log(comment_count), which helps us manage the outliers with excessively large amounts of views, likes, dislikes, and comments. The values of log(views), log(likes), log(dislikes), and log(comment_count) are found in the columns logviews, loglikes, logdislikes and logcomments, respectively.

What is the value of log(views) that our baseline model predicts?

Answer: 13.37152

What is the correlation between log(views) and log(dislikes) in the training set?

Answer: 0.8722925

Choose the most reasonable answer from the following statements:

- ○ Higher log of dislikes are associeted with higher log of views, likely because the popular videos often have many dislikes. ✔
- ○ High log of dislikes are associated with less log of views.
- ○ There is no association between log of dislikes and log of views.

Submit You have used 0 of 2 attempts

ℹ Answers are displayed within the problem

Problem 2.2 - Simple Linear Regression (cont'd)

0.0/4.0 points (graded)

Create a linear model that predicts log(views) using log(dislikes).

What is the coefficient of log(dislikes)?

Answer: 0.786930

Load youtube_test.csv into a data frame called test.

What is the R2 on the test set?

Answer: 0.7519569

Submit You have used 0 of 2 attempts

Problem 3 - Adding More Variables

0.0/6.0 points (graded)

As good practice, it is always helpful to first check for multicollinearity before running larger models.

Examine the correlation between the following variables:

logdislikes, loglikes, logcomments, tags, publish_month, and trending_month

Which of the following pairs of variables have correlation with magnitude above 0.8? Select all that apply.

- ☐ logdislikes, loglikes ✔
- ☐ logcomments, logdislikes ✔
- ☐ tags, logcomments
- ☐ trending_month, tags
- ☐ publish_month, trending_month ✔
- ☐ logcomments, loglikes ✔

Create a linear model that predicts log(views) using the following variables:

logdislikes, tags, and trending_month.

We have excluded loglikes, logcomments, and publish_month due to concerns about multicollinearity.

What is the value of the intercept?

Answer: 8.361

What is the R2 on the test set?

Answer: 0.7530081

Submit You have used 0 of 2 attempts

Answers are displayed within the problem

Problem 4 - Interpreting Linear Regression

0.0/3.0 points (graded)

Using the model from Problem 3, which of the following variables are significant at a level of 0.001 (p-value below 0.001)? Select all that apply.

- ☐ logdislikes ✔
- ☐ loglikes
- ☐ tags ✔
- ☐ logviews
- ☐ trending_month ✔

Using the model from Problem 3, how would you interpret the coefficient of tags?

- ○ All else being equal, an increase in tags is associated with a 1.655e-04 increase in log(views). ✔
- ○ All else being equal, an increase in tags is associated with a 1.655e-04 decrease in log(views).

Using the simple model from Problem 2, if the amount of dislikes is 1000, how many views does the model predict the video has?

- ○ 795.2351
- ○ 3182852
- ○ 1906.064
- ○ 928263.8 ✔

Submit You have used 0 of 2 attempts

ℹ Answers are displayed within the problem

Problem 5 - CART and Random Forest

0.0/10.0 points (graded)

In addition to the linear regression model, we can also train a regression tree. Use the same variable as used in the simple model, logdislikes. Train a regression tree with cp = 0.05.

Looking at the plot of the tree, how many different predicted values are there?

Answer: 4

What is the R2 of this model on the test set?

Answer: 0.65701

The out-of-sample R2 does not appear to be very good under regression trees, compared to a linear regression model. We could potentially improve it via cross validation.

Set seed to 100, run a 10-fold cross-validated cart model, with cp ranging from 0.0001 to 0.005 in increments of 0.0001. What is the optimal cp value on this grid?

Answer: 0.0001

What is the R2 of this new model on the test set?

Answer: 0.7528402

Create a random forest model that predicts log(views) using the same variable as the CART model, with nodesize = 200 and ntree = 50. Set the random seed to 100.

What is the R2 of this new model on the test set?

Answer: 0.7513246

Submit You have used 0 of 2 attempts

Answers are displayed within the problem