The Joy of Machine Learning

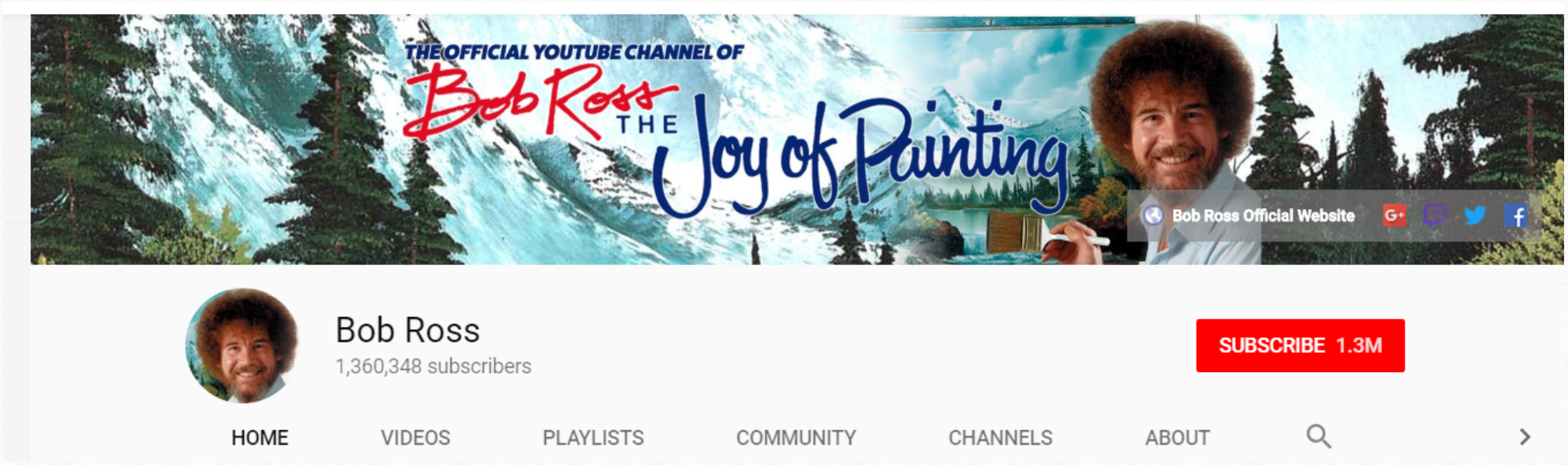
Predicting viewership of Bob Ross' The Joy of Painting episodes on YouTube using painting elements

Background & Client

Bob Ross

You've seen him before. He's the soft spoken guy painting happy clouds, mountains and trees in about twenty-six television minutes, using big housepainting-type brushes and cooing soothing "you can do its" to the audience. His Joy of Painting program is the most recognized, most watched TV art show in history.

Background & Client



Background & Client

- The social media manager for Bob Ross' official website wants to run a competition for Bob Ross 'paint-a-likes'. Contestants ('Bob-testants'?) will film themselves painting in the same style, and submit videos for judging.
- To maximize the potential for visibility of this project, she wants to know what painting topics predict the greatest viewership of existing Bob Ross episodes hosted on YouTube. These elements will be required in contestants' submitted painting videos.

Data Acquisition

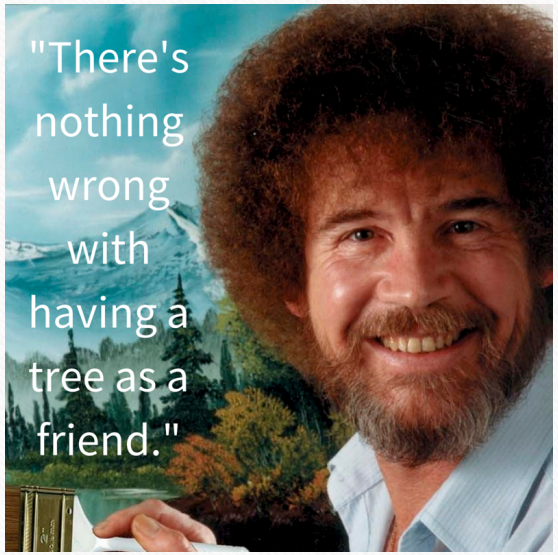
- Fortunately, someone has already sifted through all The Joy of Painting episodes and listed elements present in the final painting
  - These data are hosted on the github for FiveThirtyEight, https://github.com/fivethirtyeight/data/tree/master/bob-ross
- YouTube also provides public access to channel data through its YouTube Data API
  - View count, comment count, like count, etc…

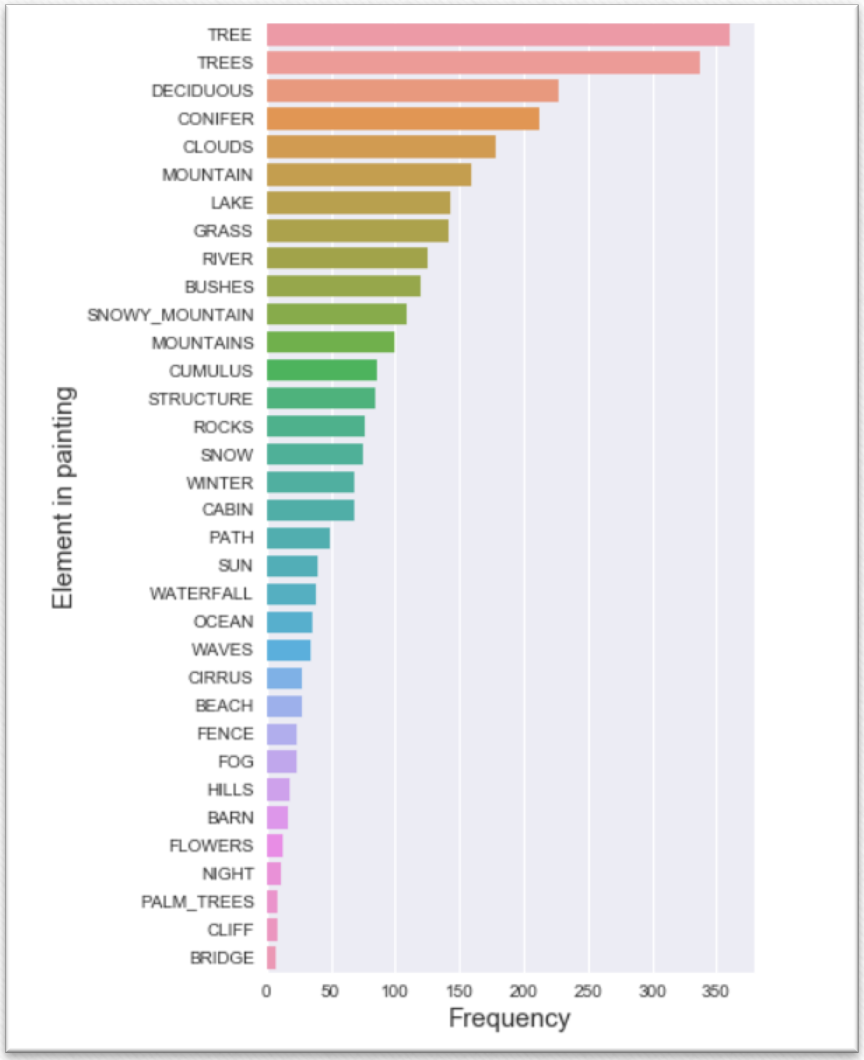
Data cleaning

- Painting elements: provided in a clean, wide csv
- YouTube data:
  - Access YouTube Data API & pull json
  - Read json text for episode title and view count
- Data merge:
  - Join on episode title

Painting elements

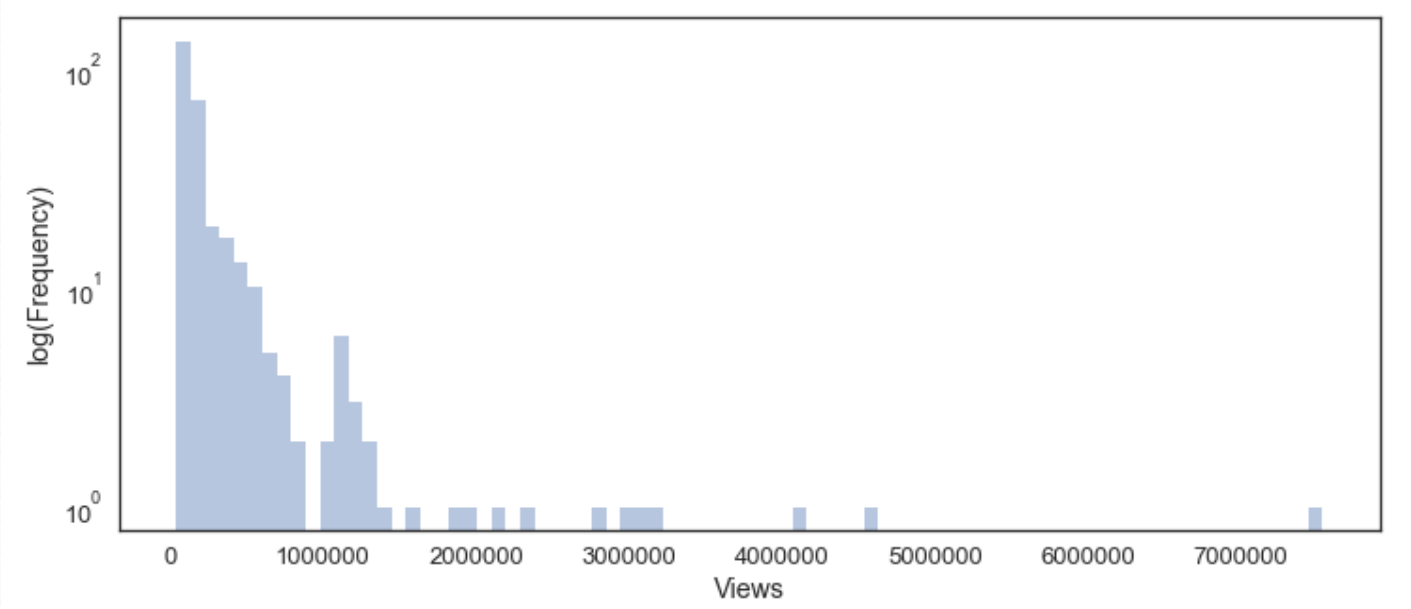
- Most common are trees (singular, multiple, and of different types)
- Then mountain, lake, and grass





Viral Videos

- What is the threshold view count of interest? This histogram shows there is a break in the distribution at about 900,000 views



Initial results

- Decision tree and random forest models are both having a hard time with the low incidence of ‘positives’ (viral videos) in the training data.
  - In most models, none of the positive test cases are predicted positive (i.e. Recall = 0)
  - Best model is a random forest ensemble model with only true one positive predicted

Strategies for improvement

- get more data
- add more features
- fancier feature engineering (e.g. featuretools)
- switching from classification to regression problem (predict # of views)

Strategies for improvement

- ~~get more data~~ (not possible in this problem)
- ~~add more features~~ (not relevant to this problem)
- fancier feature engineering (e.g. featuretools)
- switching from classification to regression problem (predict # of views)

Influential elements

- Top five important features that predict a video going ‘viral’

Element	Importance
clouds	0.119584
fog	0.095371
river	0.095059
lake	0.091317
conifer	0.089633

Best strategy for contest!

- Go for a combination of MOST COMMON elements
  - Trees of any kind (duh)
  - Mountains and clouds

  AND
- Top IMPORTANT elements
  - Clouds & fog, river & lake, conifer tree

Other Information

- Please find relevant code notebooks at
  - https://github.com/mskaerthomason/Springboard-Capstone-2
  - Thank to you my mentor, Everett Wetchler, and Springboard staff
- Say hello on LinkedIn or Twitter:
  - /in/mskaerthomason
  - @thomasonmeg