Constructed Youtube Analytics

Group Members: Darren Tsang and Yonatan Khalil

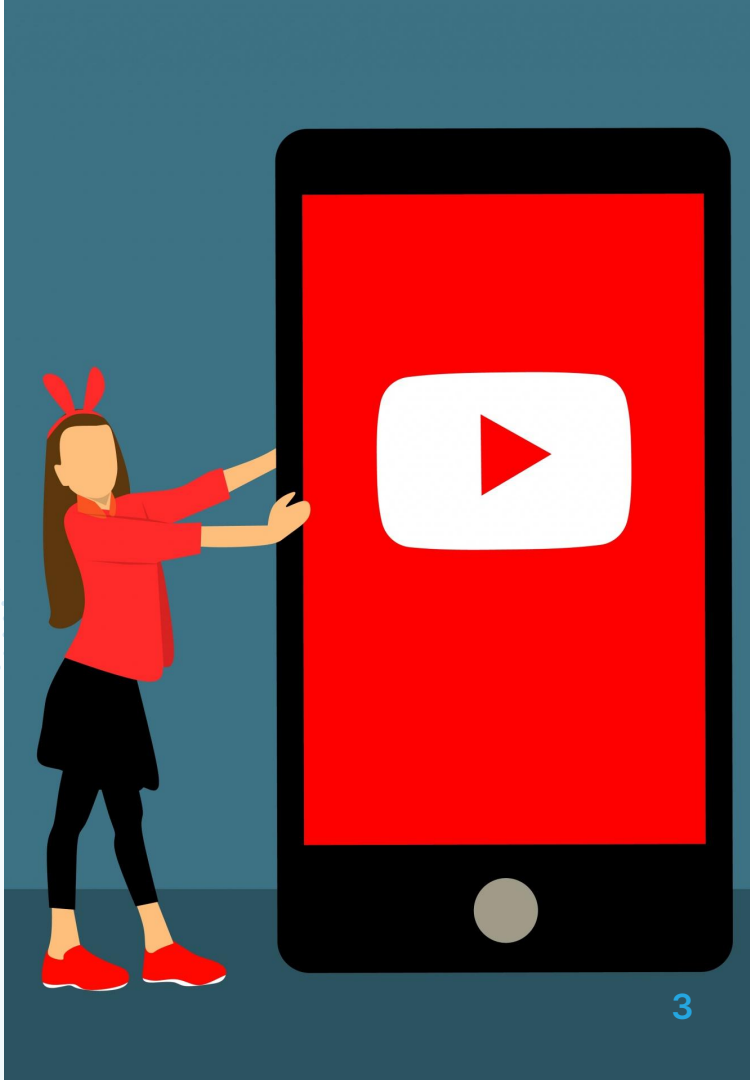
1. Introduction

Just like the first sip of coffee

YouTube

Founded in 2005

Purchased by Google a year later in 2006 and is the main hub for video sharing.



Content Creators

- Rely on Analytics and Projections to improve
- They will often see a plot similar to the one below which approximates their view count in the future.

Our Goal

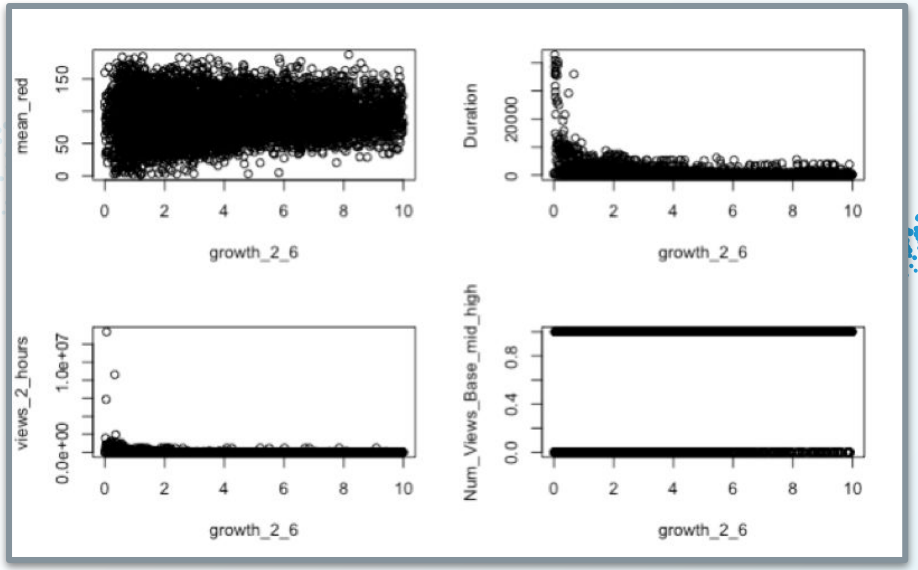
Predict the percentage of change of a video's views between the second and sixth hour

Observations and Predictors

- 7242 Videos
- 258 Predictors (Some Continuous and some Discrete)
- Response Variable: growth_2_6

Exploratory Data Analysis

Visual methods and reading through a few rows of our raw data is an important step in creating a reliable model.



2. Methods

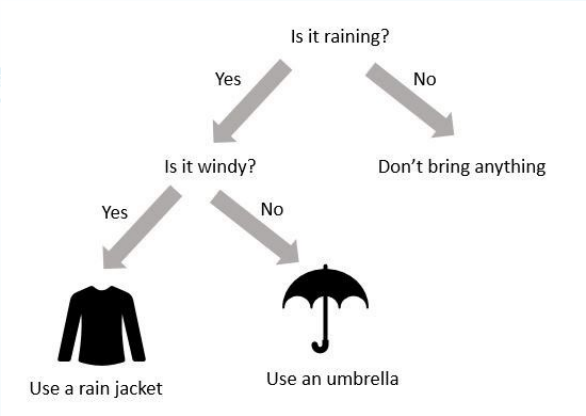
The second sip of coffee

Data Cleaning

- Removed *id* variable
- Used *PublishedDate* to create other variables
  - *month*, *day*, *min_of_day*
- Removed highly correlated variables (< .7)
- Removed columns where all values were 0
  - *eg. min_red*, *min_green*, *min_blue*

Decision Trees

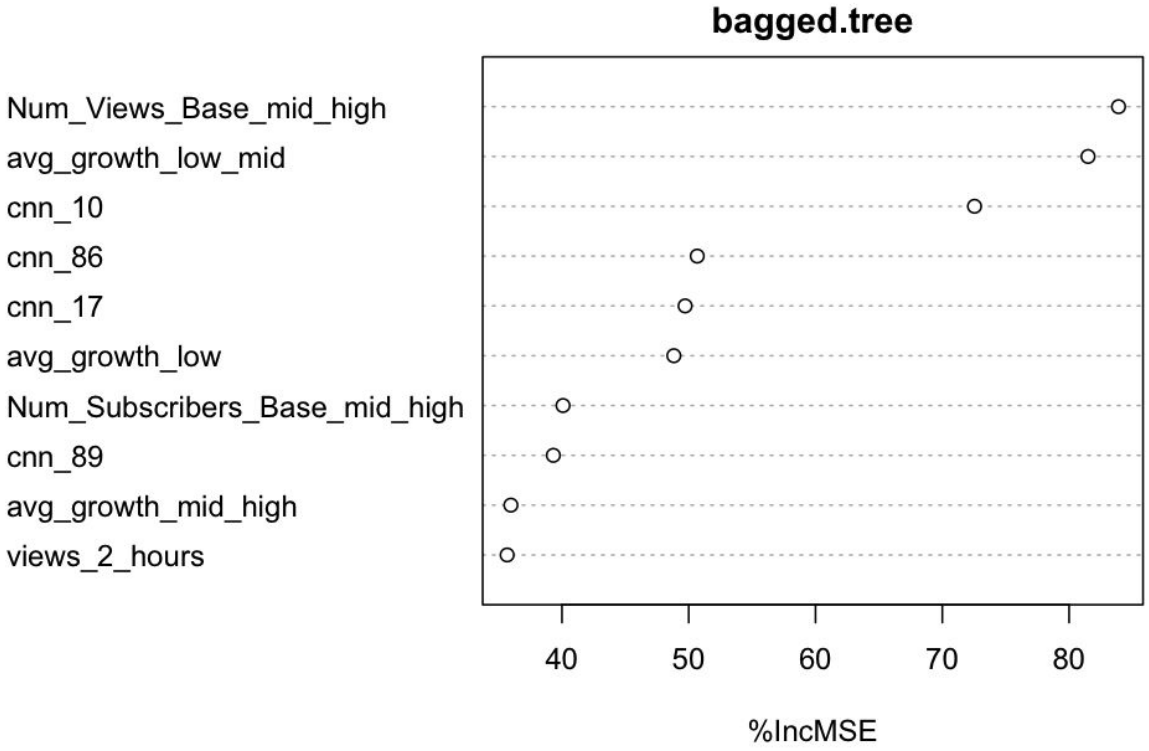
- Relatively simple algorithm that “asks” a question at each node
  - Goes left or right depending on answer
- When you reach a leaf node, you get your response variable



Bagging

- Create many decision trees using a bootstrap sample
- New predictions are run through all decision trees, then the average outcome is taken
- Found that building 500 decision trees was best through trial and error

Important Predictors



3. Conclusions

Shoutout to Coffee Bean and Tea Leaf

1.41472

On the Public Data

1.40174

On the Private Data

Above Every Threshold

Success!

Strengths

- Avoids Multicollinearity and uses an adjustable function
- Simple application of bagging
- Shown to be a good model on both private and public datasets

Chosen vs Best Model

Our Chosen Model

- Kaggle score of 1.40174.
- Airs on the side of caution and simplicity by using a bagging method instead of random forest.

Our Best Model

- Kaggle score of 1.39849.
- Random forest was chosen, but too computationally expensive
- More analysis necessary.

Future Recommendations

- Find the best balance of good predictors and multicollinearity
- Sift through Random Forest (a very promising option)
- Attempt stacked methods, similar to our midterm submission

Overall, given time and alternate methods, there are many other routes we can take to improve our model.

Thanks!

Any questions?