# DATA 621 – Business Analytics and Data Mining

Predicting the number of YouTube views using linear regression

Fall 2020 - Group 2 - Final Project

Authors: Avraham Adler, Samantha Deokinanan, Amber Ferger, John Kellogg, Bryan Persaud, Jeff Shamp

# Abstract

- Method
  - Attempted to predict number of views
  - Used criteria including category, number of likes, number of dislikes, and number of comments
  - Engineered features based in ratios of predictors
  - Used linear regression to test for predictive power
- Results
  - Allow ratings and posts which will engender heated discussion.
- Keywords
  - Youtube, linear regression, elastic net, R

# Introduction

- YouTube has changed the future of video entertainment
- YouTube's model connects a user's creativity with a desire for global recognition
- Creators from all over the world gain international prominence using their own equipment and space
- Online video platforms are taking over the entertainment world.
  - 6 out of 10 people already prefer online video platforms or streaming services over live TV
  - Expected in four years—half of viewers under the age of 32 will not pay TV service
- Understanding this world is beneficial to many people and companies

# Literature Review

This problem has been addressed often. For example:

- S. Ouyang, C. Li, and X. Li, “A Peek Into the Future: Predicting the Popularity of Online Videos,” *IEEE Access*, vol. 4, pp. 3026–3033, June 14, 2016, doi: 10.1109/ACCESS.2016.2580911.
- H. Pinto, J. M. Almeida, and M. A. Gonçalves, “Using Early View Patterns to Predict the Popularity of Youtube Videos,” in *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, Rome, Italy, 2013, pp. 365–374, doi: 10.1145/2433396.2433443.
- T. Trzciński and P. Rokita, “Predicting Popularity of Online Videos Using Support Vector Regression,” *Ieee Transactions on Multimedia*, vol. 19, no. 11, Art. no. 11, April 18, 2017, doi: 10.1109/TMM.2017.2695439.
- A. Srinivasan, “Youtube Views Predictor,” December 12, 2017. https://towardsdatascience.com/youtube-views-predictor-9ec573090acb.

# Methodology

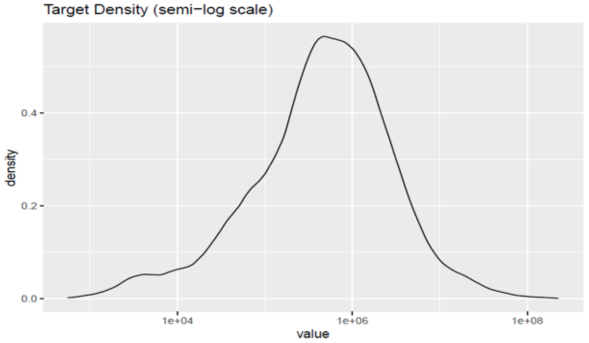
- Data for this project is from Kaggle.
- Data roadblocks:
  - Data appeared to be broken up by country.
  - Actually, country seems to be just location of count capture
  - Most videos had same statistics on same days across countries
  - United States had the most observations so it was selected
- Feature selection
  - Explored relationships between a views and number of likes, dislikes, and comments
  - Used video category and properties as factor features
  - Engineered ratio features
  - Used linear regression to investigate relationship between features and views
- Models compared using RMSE, $R^2$, and MAE

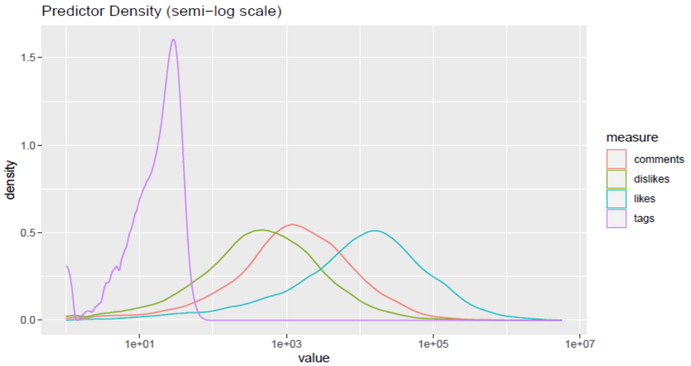
# Data Exploration

- Target Variable: Views
  - On log scale looks Gaussian implying lognormal distribution.

- Numeric Predictors
  - Number of tags is orders of magnitude less than rating or comment variables
  - More likes than comments, More comments then dislikes
  - All three exhibit symmetric Gaussian-like behavior
- Categories
  - Interests Vary by Category
  - In the US
    - 1st is Entertainment
    - 2nd is Music

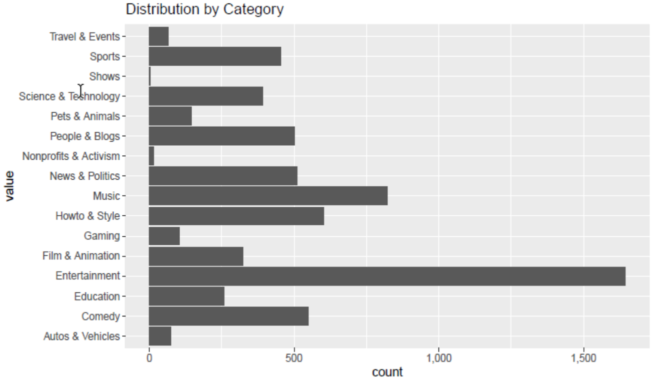
Target Variable




Numeric Predictors

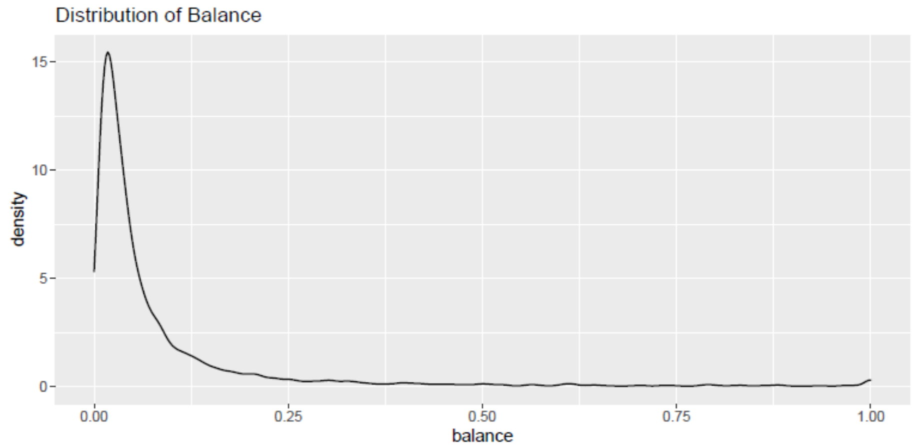
Categories



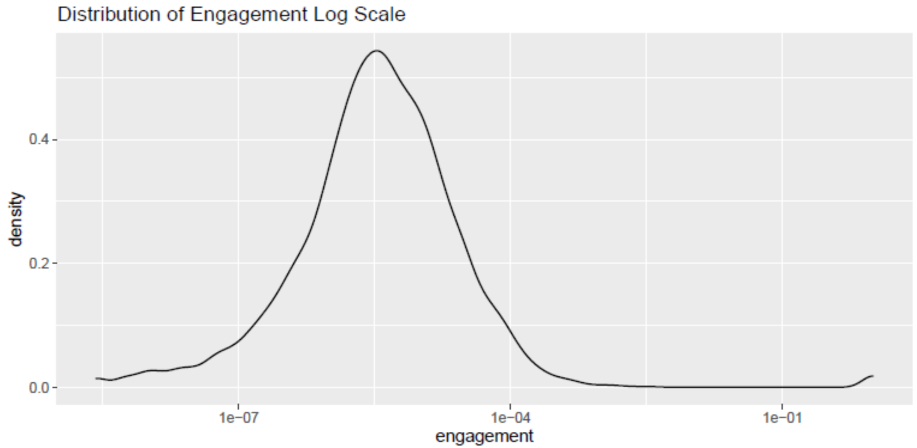
# Feature Selection & Engineering

- Feature Selection—Relationship of views to:
  - Likes, dislikes, comments, and tags
  - Category
  - Disabling comments or ratings
  - Error with the video

- Feature Engineering—Hypothesize that if video is universally loved or panned it will get fewer views
  - Balance: ratio between likes and dislikes
  - Engagement: ratio of comments to sum of likes and dislikes




It’s pretty clear from the distribution of balance that, at least for videos viewed in the US, there is a healthy dose of disagreement!



# Models

Investigated multiple models in linear regression family

- Investigated four models in three families:
  - Model 1: Ordinary Least Squares
    - Equivalent to GLM with Gaussian errors and inditiy link function
  - Model 2a & 2b: GLM with log link function
    - Model 2a: Gaussian Errors
    - Model 2b: Poisson Errors
      - More appropriate for counts in general
  - Model 3: Penalized Regression
    - Elastic Net Model
    - Combines best features of Lasso and Ridge regression
- Details may be found in corresponding paper

# Model Evaluation

- The three models contain both intuitive and counter-intuitive results.
- As expected, videos with more ratings tend to have more views.
- However, as likes outnumber dislikes, that tends to reduce the number of views.
- All the models agree that having a video whose likes and dislikes are close in magnitude increases the propensity for views.

Table 5: Model Performance on Test Set

| Model | RMSE | R2 | MAE |
|---|---|---|---|
| LM | 3,301,299 | 0.675 | 1,068,070 |
| GLM: Gauss+Log | 5,256,922 | 0.175 | 2,186,177 |
| GLM: Poisson+Log | 4,952,845 | 0.267 | 1,838,099 |
| ElasticNet | 3,279,379 | 0.679 | 1,036,441 |

- ElasticNet model performed best
- Clearly the log link is inferior to the identity link for this data set

# Conclusions

- Make sure videos will engender heated discussions
  - Get as many dislikes as likes
- Keep ratings enabled.
  - Allowing users to rate your video drives attention to them
- Investigate potentially disabling comments.
- Stay ***away*** from activism.
- Talk about films or comics.
  - Probably reflects the anime and manga phenomena.
- Talk about cars
- More tags is better

# Future Development

- Investigating more sophisticated algorithms
- Analyzing growth patterns over time
- Finding better sources or methods to include country without overcounting

# Questions?

Thank you for making the time to view our presentation and our findings.