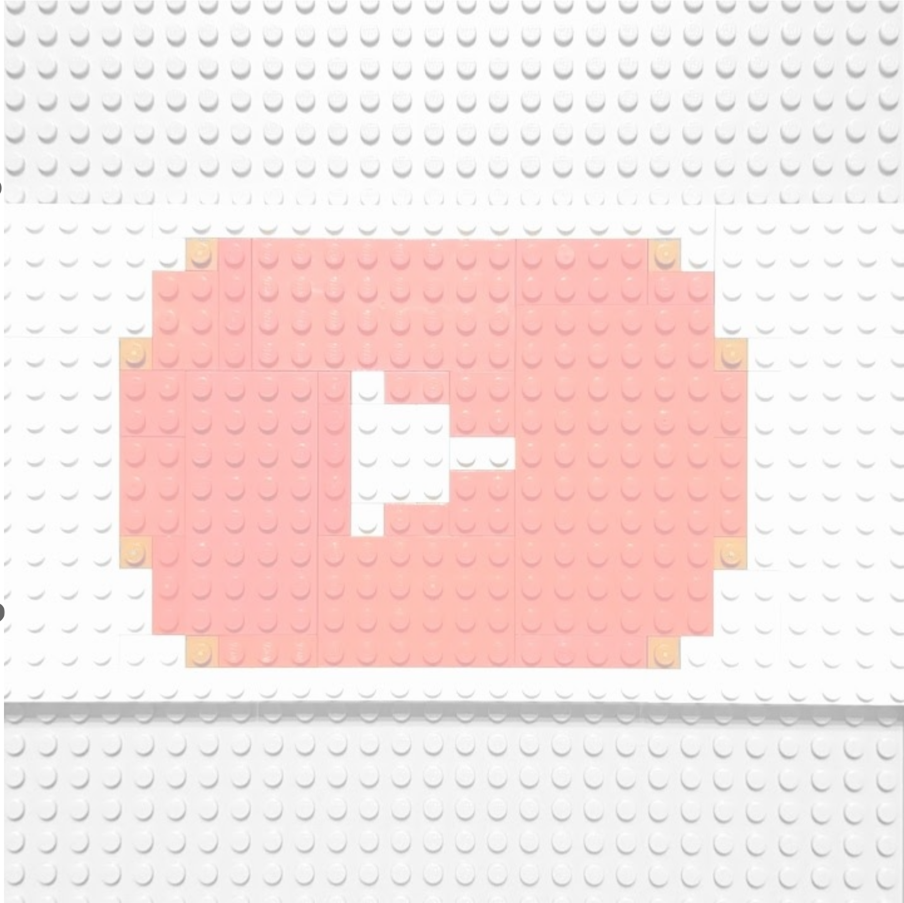
Building the Optimal Youtube Tech Channel

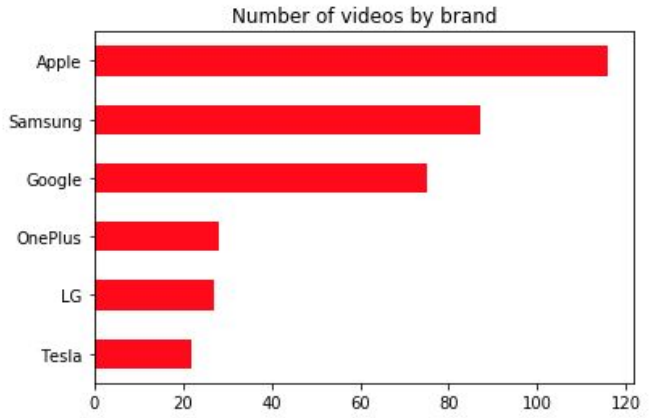
A guide by Jack Tann & Carter Bouley

Youtube Video Building Blocks

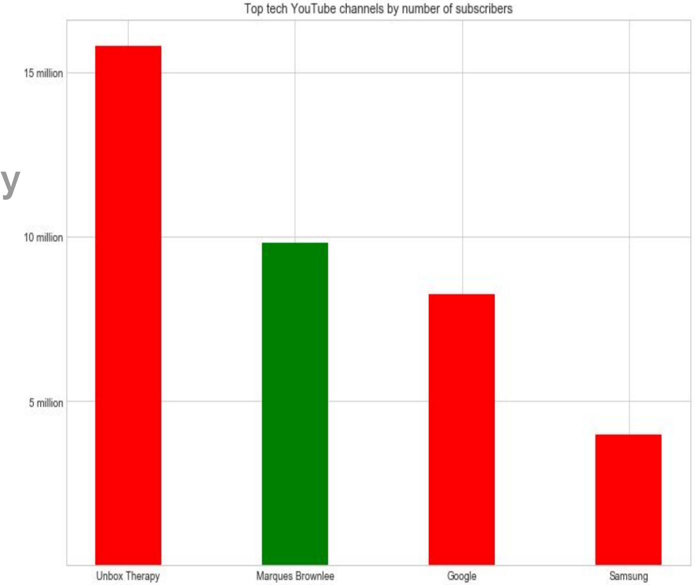
- How long should the video be?
- Which brands should you discuss?
- How many tags is too many tags?
- How often should you post?
- When should you post?
- Should you encourage comments?



Marques Brownlee



- All videos are tech related
- Consistent title formatting
- Sample size of over 1,000
- Upward trajectory in popularity
- Diversity in brand coverage
- Over 1.5bn views
- 9.8M subscribers



Methodology

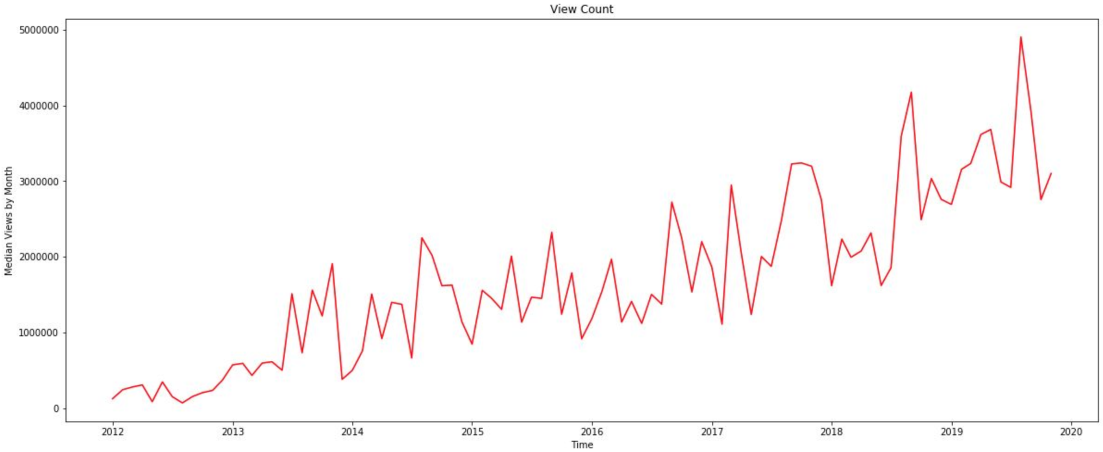
- **Extract**
  We navigated the YouTube API system to extract the video parameters we wanted

- **Transform**
  Converted our data into numerical and categorical formats which could be passed into a Machine Learner, for example:
  - Extracting brand names from titles
  - Extracting seasons from upload date

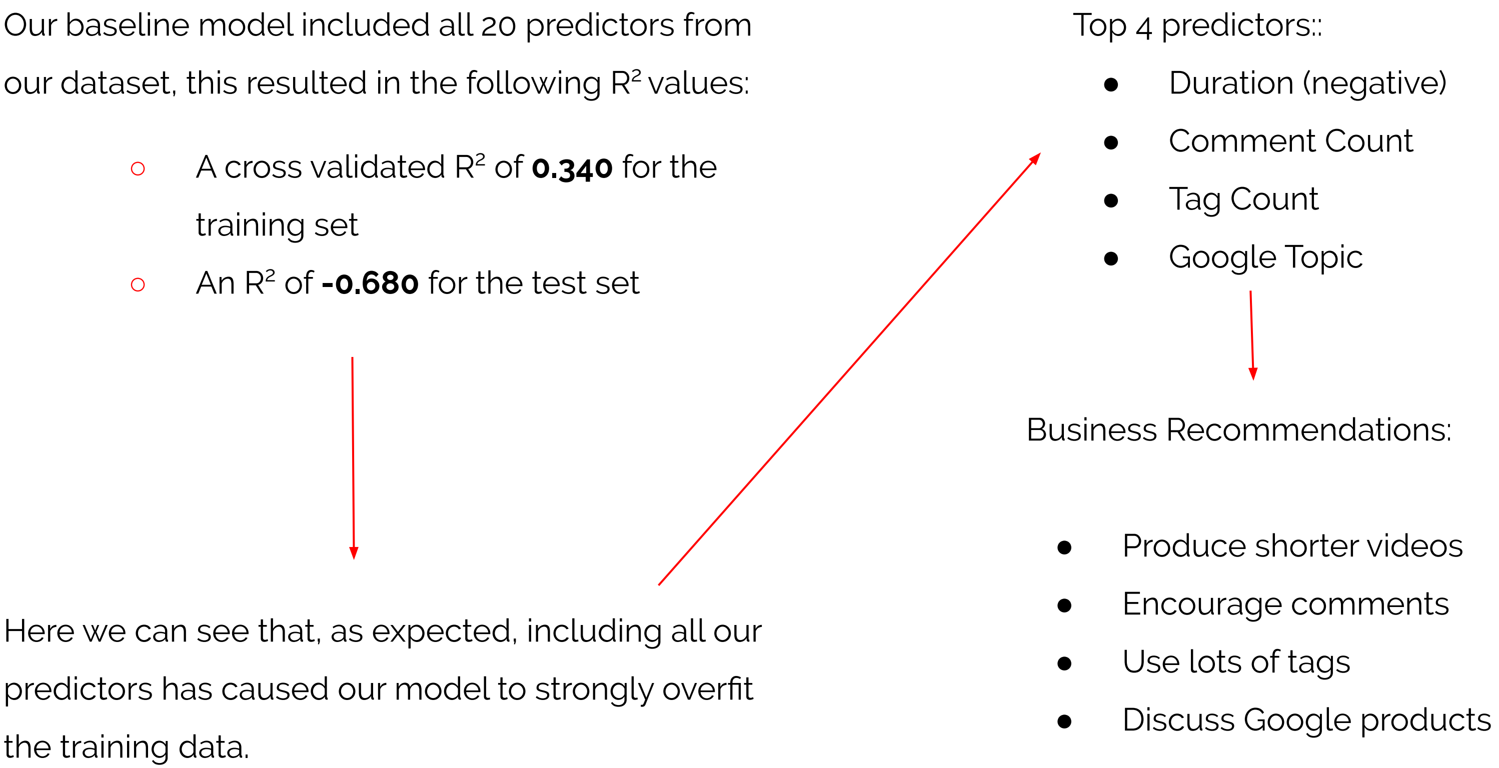
- **Regression Model**
  Optimized our model based on features and strength of relationship to predict view count.

- **Time Series Analysis**

  Predicting future views based on historical data



Baseline Model



Our baseline model included all 20 predictors from our dataset, this resulted in the following R^2 values:

- A cross validated R^2 of **0.340** for the training set
- An R^2 of **-0.680** for the test set

Here we can see that, as expected, including all our predictors has caused our model to strongly overfit the training data.

Top 4 predictors::

- Duration (negative)
- Comment Count
- Tag Count
- Google Topic

Business Recommendations:

- Produce shorter videos
- Encourage comments
- Use lots of tags
- Discuss Google products

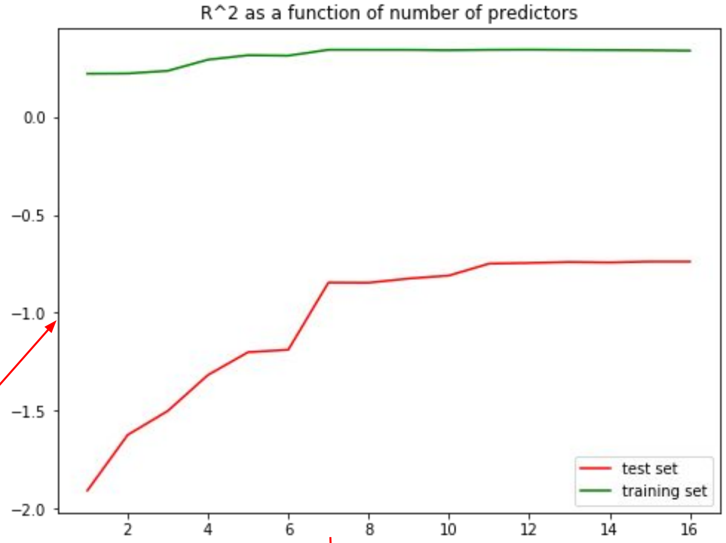
First Iteration

To deal with the overfitting from our baseline model we performed a series of feature selection methods to remove any redundant predictors. This included:

- Inspecting model coefficients
- Setting a minimum variance threshold
- Recursive feature elimination

After removing 11 predictors from our baseline model, we obtained the following R^2 values:

- A cross validated R^2 of **0.340** for the training set
- An R^2 of **-0.750** for the test set



Here we have observed that reducing the model complexity by a factor of 2 has marginally worsened the model fit for the test set.

Second Iteration

To account for possible interactions between the predictors we added every possible first order combination of the remaining predictors from our reduced feature space to our first iteration model. In order to mitigate the extra complexity arising from these new predictors, we applied L1 and L2 regularisation independently and compared their effectiveness at improving the model fit for the test set.

Using L1 regularisation, we obtained the following R^2 values:

- A cross validated R^2 of **-0.661** for the training set
- An R^2 of **-10.7** for the test set

Using L2 regularisation, we obtained the following R^2 values:

- A cross validated R^2 of **0.405** for the training set
- An R^2 of **0.430** for the test set

Here it is clear that L2 regularisation performed considerably better than L1 regularisation.

Final Model

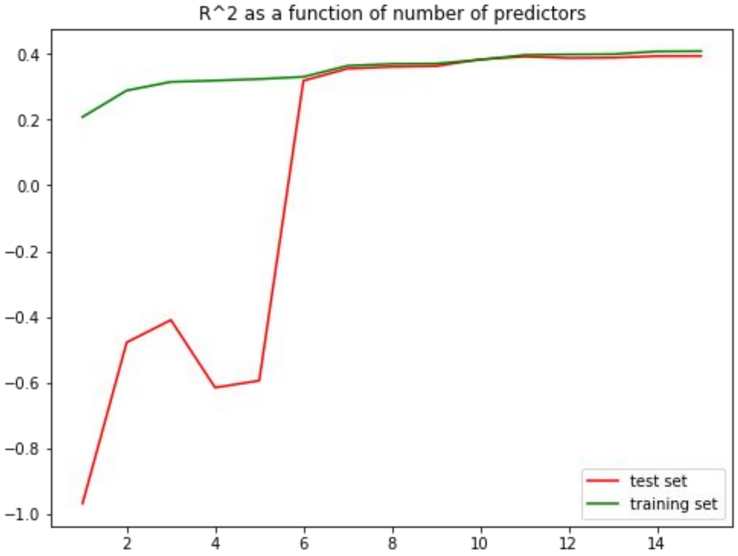
To combat the 3 fold increase in complexity arising from considering interaction effects, we applied further recursive feature elimination to remove a further 9 predictors from our model.

This resulted in the following R² values:

- ○ A cross validated R² of **0.330** for the training set
- ○ An R² of **0.319** for the test set

Top 4 predictors:

- Comment count
- Duration (negative)
- Review topic / duration interaction (negative)
- Comment count / like dislike ratio interaction



Final Business Recommendations:

- Produce short videos that are predominantly review based
- Encourage comments, whilst maintaining an impartial view when comparing tech brands

Limitations

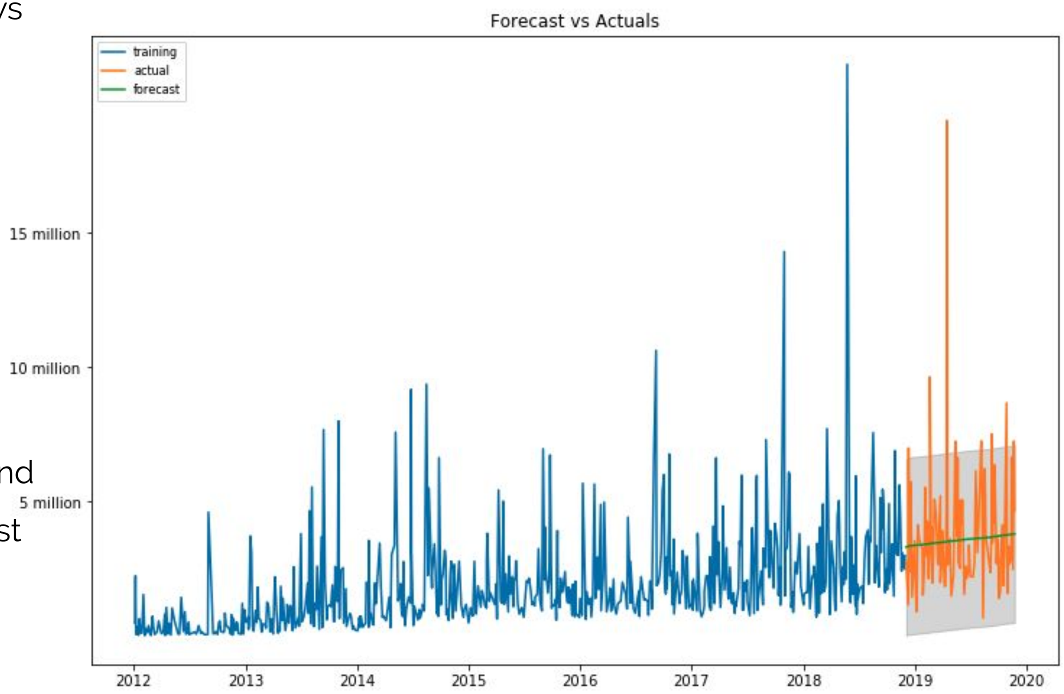
- Video quality is hard to quantify using meta-attributes
- Virality is unpredictable
- Not accounting for popularity growth over time
- Heavily skewed distribution of view count
- Limited numerical variables

Time Series Analysis

Time Series Analysis

- We subsequently ran an ARIMA model on the change in views across each video.
- This model predicted that future video views were likely to land within the shaded confidence interval 95% of the time.
- This model has a MAPE of 0.506
  - This implies 49.4% accuracy in predicting the next 100 videos view count (orange)

- A future model combining our regression and time series model would likely yield the best results.


Forecast vs Actuals
training
actual
forecast
15 million
10 million
5 million