Predicting the Growth Rate of Youtube Videos

Catherine Jennifer, Anshuman Mahalley, Priya Jain

STATS 101C Team Fire Breathing Rubber Ducks

December 2020

Introduction

- Youtube is the 2nd most visited site in the world, with almost 2 billion users.[1]
- Youtube's revenue came to $15.1 billion for the full year 2019, up 36% from 2018.[1]
- In this project, we would like to **predict the growth rate of a Youtube video between the second and sixth hour since it was uploaded.**

[1] "YouTube Revenue and Usage Statistics (2020)." Business of Apps, 17 Nov. 2020, www.businessofapps.com/data/youtube-statistics/.https://www.businessofapps.com/data/youtube-statistics/

Preprocessing

- Extracted data regarding the hour of day on which videos were published from the PublishedDate column; replaced with the new PublishedHour column.

  8/5/20 7:27 ⇒ 7

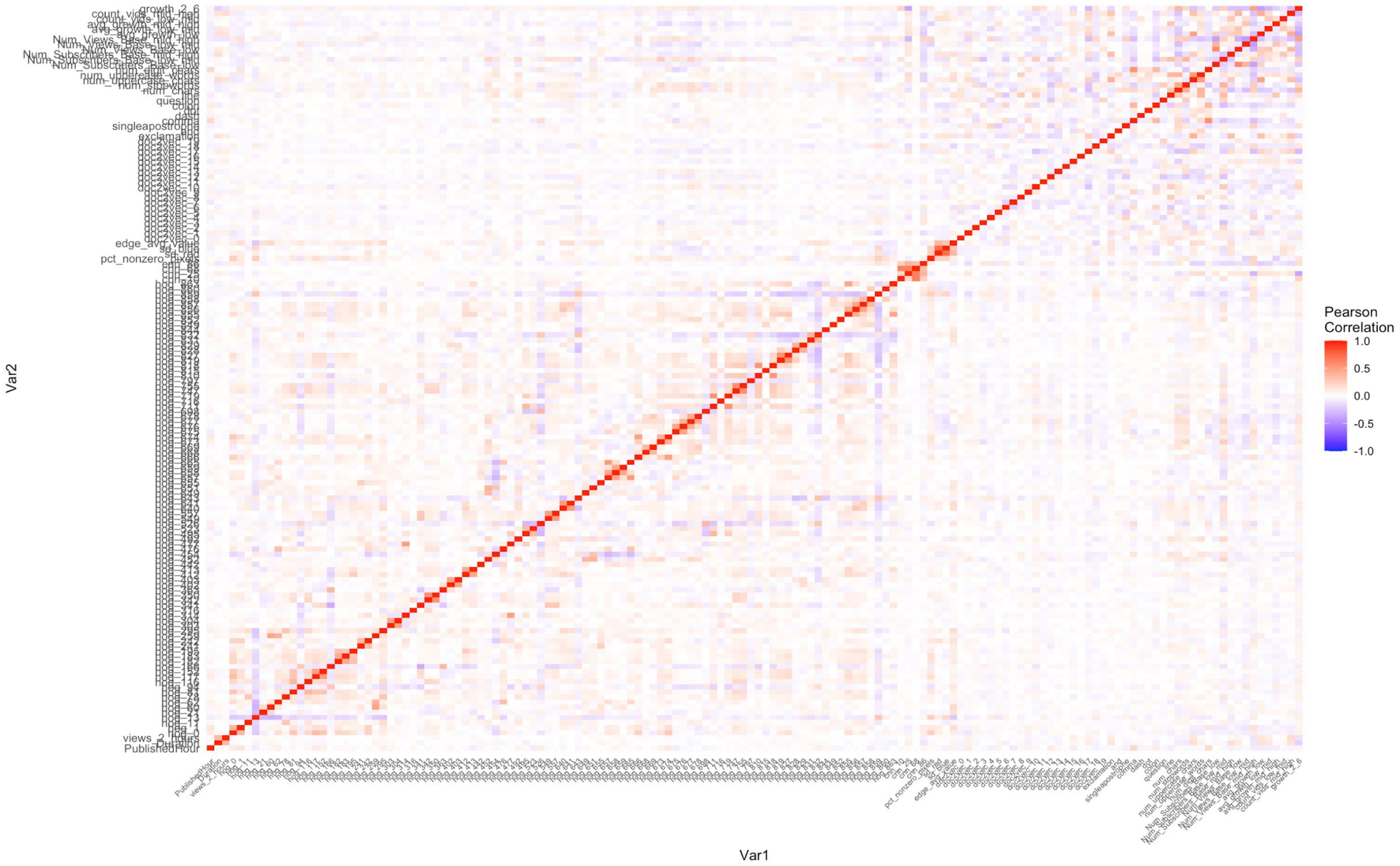
- Removed id and renamed several predictors from punctuation marks into letters.

  “punc_num_,” ⇒ “punc_num_comma”

- Removed variables with standard deviation of 0.

Preprocessing

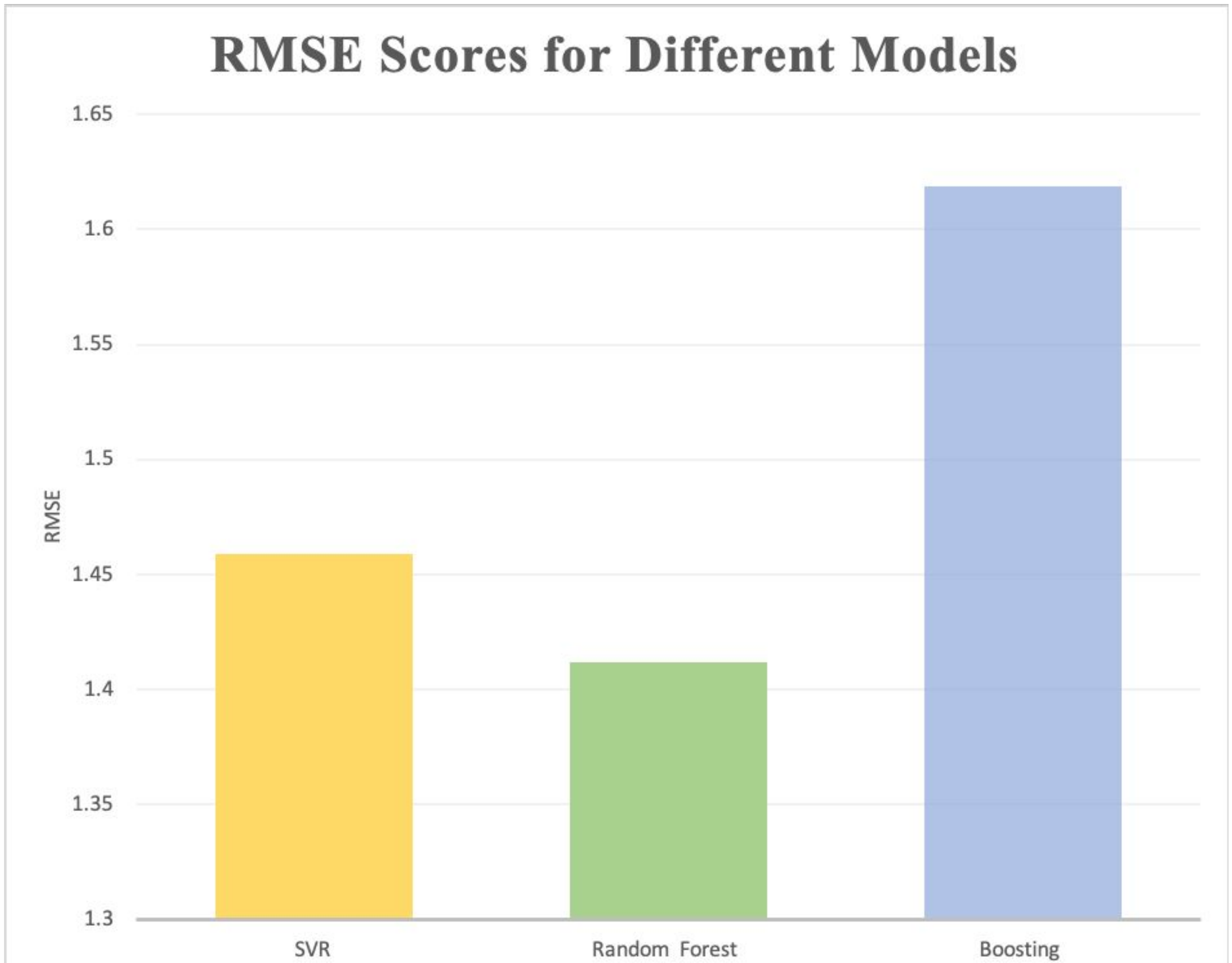
- Removed highly correlated variables (> 0.75)



Predictor Selection

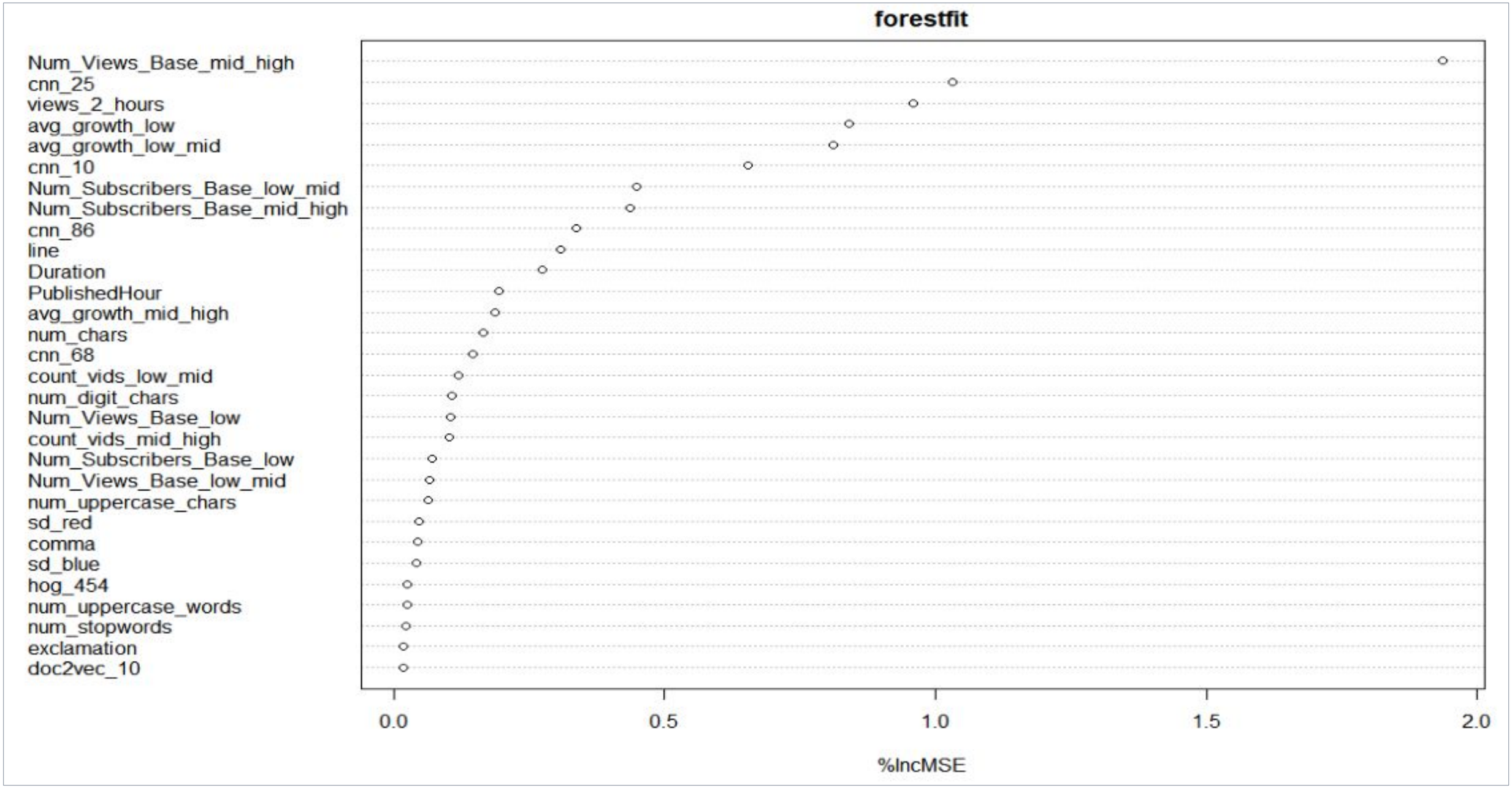
- Applied LASSO regression to select predictors for candidate models (coefficients not equal to zero)
- Considered Support Vector Regression (SVR) as our candidate model.
    - Acknowledges presence of non-linearity in the data and allows us to fit the error within a certain threshold.

Model Selection



Predictor Selection

- Random Forest Variable importance plot:



Final Model Results

- Random forest fit with the final selected predictors:

```
Call:
 randomForest(formula = growth_2_6 ~ ., data = draft_cor1, mtry = recommended.mtry2,
    tunegrid = tunegrid2, importance = TRUE)
                Type of random forest: regression
                      Number of trees: 500
No. of variables tried at each split: 9

          Mean of squared residuals: 2.158884
                    % Var explained: 68.8
```

- The RMSE of our final model is 1.41172 according to the Kaggle public leaderboard, and 1.42997 according to the private leaderboard.

Conclusion

- Looking at the top 10 predictors, cnn* is included thrice which indicates the **importance of thumbnail features** in a Youtube video's virality.
- **Video Duration:** The COVID-19 pandemic has resulted in increase of short-form video viewing.
    - 10 minute videos are conducive to Youtube's recommendation algorithm.[2]

[2] "Creators Are Making Longer Videos to Cater to the YouTube Algorithm." Digiday, 24 Mar. 2020, digiday.com/future-of-tv/creators-making-longer-videos-cater-youtube-algorithm/.

Model Limitations

- Use Principal Components Regression to understand the predictors hog_* and cnn_*.
- Perform transformations on binary variables such as Num_Subscribers_*, Num_Views_* to create new features
- Fit model with most important predictors

Thank you!