

Final Presentation - The Gould-en Rule

Stats 101C Lecture 3

Andy Shen, Ethan Allavarpu

Fall 2020

Section 1

Introduction

Introduction

- With the rise in popularity of YouTube, many people are now making a living off creating YouTube videos
- The more views gained by the video, the more likely it is for that channel to profit
- We are interested in predicting the growth rate in video views between the **second** and **sixth** hour that a YouTube video is published

Section 2

Pre-Processing

- Examined a univariate plot to look for stray points and removed them systematically
 - ▶ Based off personal judgment and inference on the effect of the stray points
- We also remove highly correlated variables as indicated by a heat map
 - ▶ To avoid overfitting based on having too many predictors

Predictor Selection

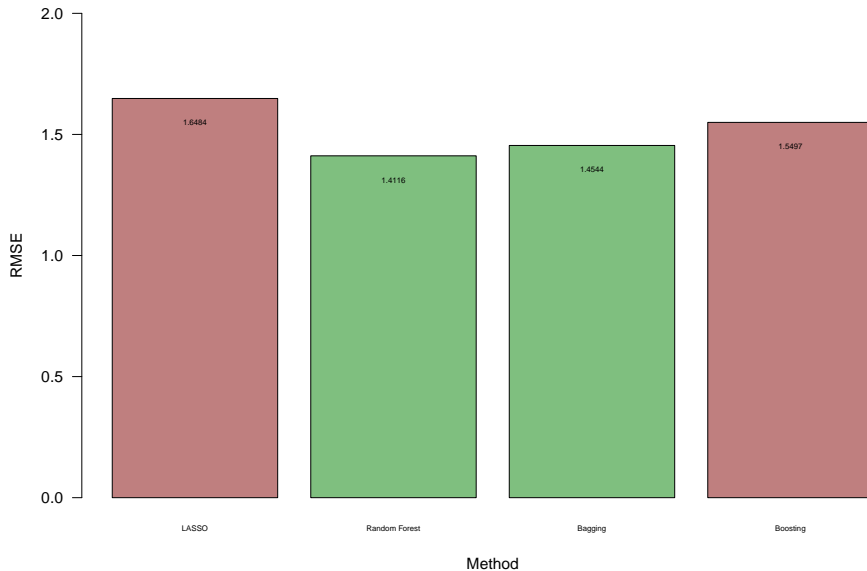
- We use LASSO to select significant predictors
 - ▶ Used to refine predictors from a large subset
 - ▶ LASSO pushes non-significant predictors to zero and keeps the most significant ones
- First fit a LASSO model for a sequence of candidate λ values
- Then select our optimal value of λ as the one that is one standard deviation above the λ value that resulted in the lowest test MSE
- Then extract the predicted coefficients in this LASSO model as our predictors for the candidate model

Section 3

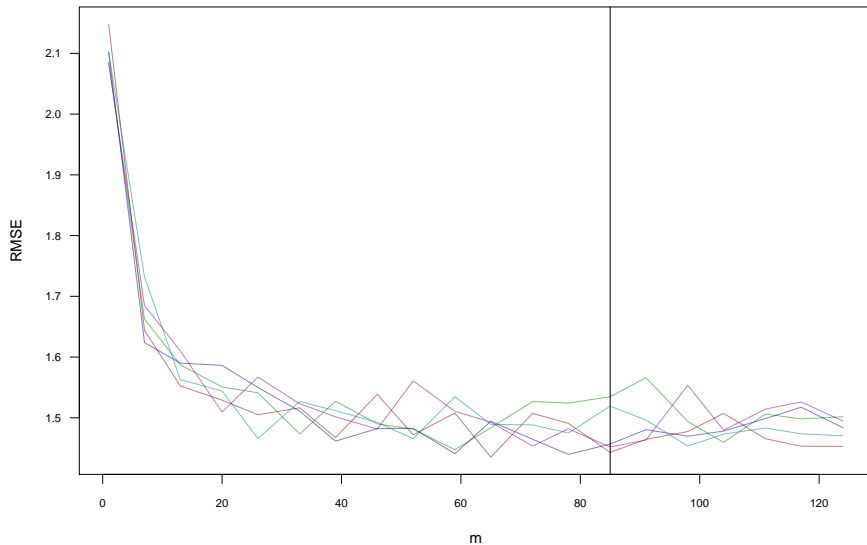
Model Fitting

- Most of our models were fit using bagging or random forest
 - ▶ Only adjusted certain parameters at a time (number of trees, depth, and λ)

RMSE for Different Approaches



Best m for Random Forest



Candidate Model: Random Forest

- After running LASSO to refine predictors, we use this smaller subset to find m for random forest
 - ▶ m is the number of variables the model randomly considers in each node of each decision tree
 - ▶ Find optimal m using 5-fold cross-validation and select m corresponding to the *median* RMSE of the 5 folds
 - ▶ Median is more preferable than mean due to the mean's sensitivity to extreme points
- Once optimal m is selected, fit another random forest model to 80% of the preprocessed training data
 - ▶ Fit this extra model to ensure the model was performing consistently

Section 4

Results

Results

Kaggle scores:

Ⓐ 1.39594

Ⓑ 1.40285

Model (B) only differs from (A) in the sense that $m = p$ as opposed to m equaling the median RMSE. Here p is the number of predictors.

Section 5

Discussion

- We believed our model performed well due to the fact that it works as an ensemble method
 - ▶ Combines multiple individual models to get more accurate responses
- By using cross-validation for our selection of m , we limit the potential effect of a random seed showing us an inaccurately good or bad RMSE