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

Outliers

- 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

Candidate Models

Random Forest

Bagging