Final Project - The Gould-en Rule

Stats 101C Lecture 3

Andy Shen, Ethan Allavarpu

Fall 2020

Introduction and Data Cleaning

The purpose of this regression analysis was to predict the growth percentage of a newly uploaded YouTube video during the 2nd through 6th hour of its publication (`grow_2_6`). In this analysis, we employ a variety of regression techniques to a variety of attributes that make up a YouTube video.

We cleaned the data by plotting each predictor variable against `grow_2_6` and examined each univariate plot for possible associations between predictor variables and the growth percentage. We noticed that there existed many outliers or stray points that did not belong in the plot. However, instead of removing these outliers, we left in the model, believing that there may be an underlying relationship explained by the outliers. However, we do remove highly correlated variables as indicated by a correlation matrix heat map. All predictor pairs with a correlation coefficient over 0.9 are removed.

We manipulated the `PublishedDate` variable by converting it from a date and time into the total number of seconds elapsed from 1 January, 2020. We also combine the binary variables of number of subscribers, views, growth, and video count and convert them into a factor woth four levels.

Predictor Selection

In order to refine our subset of predictors, we use the LASSO to select significant predictors. We first fit a LASSO model for a sequence of candidate λ values. From there, we 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. From these parameters, we extract the predicted coefficients in this LASSO model as our predictors for the candidate model.

We use LASSO to refine our predictors because this technique shrinks coefficient estimates to 0 and keeps the most important ones. As such, we are only interested in the predictors with nonzero coefficients. LASSO is unrelated to the our candidate model (TBD/bagging), and it was only used as a technique to refine the large number of predictors in the data set.

Code Appendix