Final Project

Chenxin Yang, Kevin Chen, Jonathan Martinez

Abstract

Statistical models play a crucial role in analyzing the data and producing quantitative results for future prediction. In this project, we selected 28 influential features from the grant datasets, and tried different models to predict youtube view growth rate. We eventually used a bagging model on the selected features and produced a 1.37725 RMSE on Kaggle.

Introduction

In this project, we aimed to create a supervised statistical learning model using a dataset of 7242 youtube video observations and 260 youtube video feature predictors to predict the video view growth rate in a test data set of 3105 observations. We used statistical learning methods discussed in "*An Introduction to Statistical Learning with Applications in R"* to create our models.

Methodology

a. Prepossessing the data

The raw data we have has different variable types with many missing values, we did the following to clean the data:

First we removed the ID column because it wasn't related. Then we extract the hour data from the video published date attribute, since we assume the publish hour of the day will heavily affect the video view. Next we converted all of the attributes to numeric variables. Then we imputed any missing data with the column means of the attributes. Finally we removed attributes that had zero variance as they will not affect the video view.

b. Feature selection

Firstly, we analyzed the correlation matrix of the data's attributes and identified highly correlated attributes. Then we removed attributes with an absolute correlation of 0.8 or higher, resulting in 172 variables selected.

Next, we performed a random forest model on the selected attributes to help with feature selection. We ranked the importance of features and selected features with importance higher than 10, resulting in 28 variables selected, which are the top 20% of imported variables which can be seen in Figure 1.

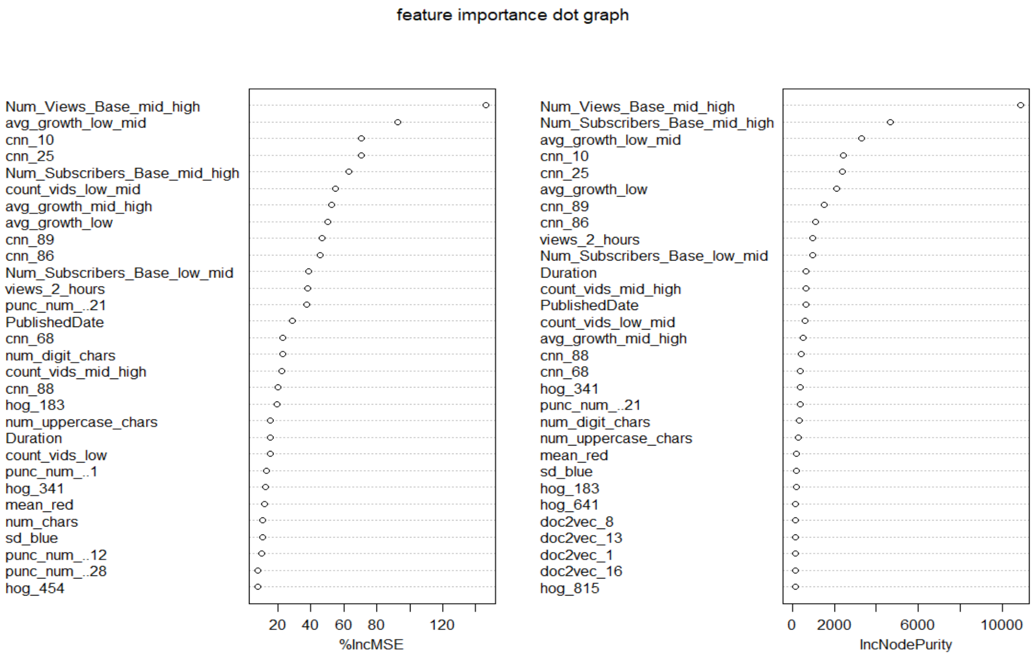

Figure 1

c. Statistical model

Since we were predicting a regression model, we have tried different methods like PCA, SVM, and Random Forest.We found that Random Forest performed better than other regression models that we tried when comparing the kaggle scores, so we used the Random Forest model as our final model. With those important features selected previously, we could fit the model with only important predictors, which would greatly enhance the model performance. To select parameter mtry in the model, we splitted the data into training and validation sets. We then iterated through all possible mtry and trained the model using the training set and stored the rmse of the model on the validation set. We output the graph of mrty parameter vs RMSE and it is shown in Figure 2.

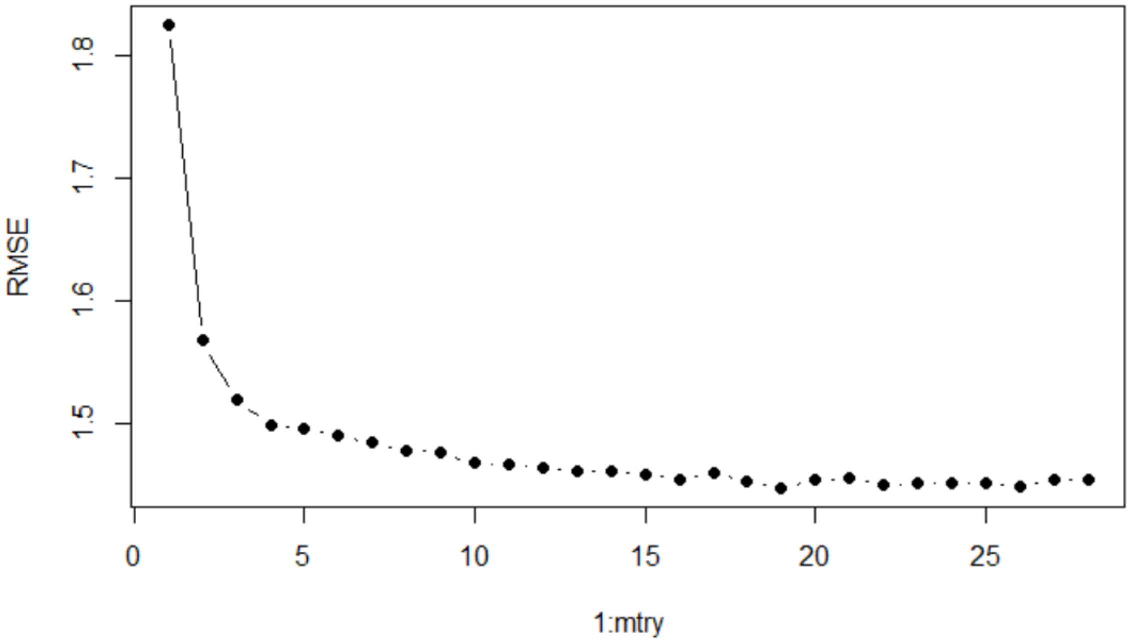

Figure 2

We found mtry = 19, 22, 26, 28 contribute to lowest rmse in the validation set. After trying those parameters and comparing their relative score on kaggle, we found that mtry=28 will generate the best kaggle score. When mtry=28, the random forest model becomes the bagging model.

Results

Using our final bagging model, we obtained a Kaggle score of 1.37725 with 40% of the test data. We were able to have a better prediction score than Model 4, which has a score of 1.42100.

General Conclusions & Evaluations

We believe that one of the reasons our model works well was because of our data prepossessing process. We chose our predictors using both correlation matrix and random forest for feature selection. This process helps us to mainly focus on the important predictors and pruning the model. In addition, the bagging method reduced the amount of variance by averaging the trees and improved the accuracy for our random forest model. The fact that our model did not change its standing at all after using the 60% of testing data shows that creating a simple model using the most important predictors, instead of a complicated one with some less important predictors, is better for prediction in this case because of lower variance.

Our model does have space for improvements. For instance, we could try different ways of imputing the missing data. Possible methods include Hot deck imputation, regression imputation, interpolation, and extrapolation. Another way to possibly improve our model is to transform some variables with the appropriate function and see if they result in high correlation with our target variable.