Team CTS Stats 101C Final Project

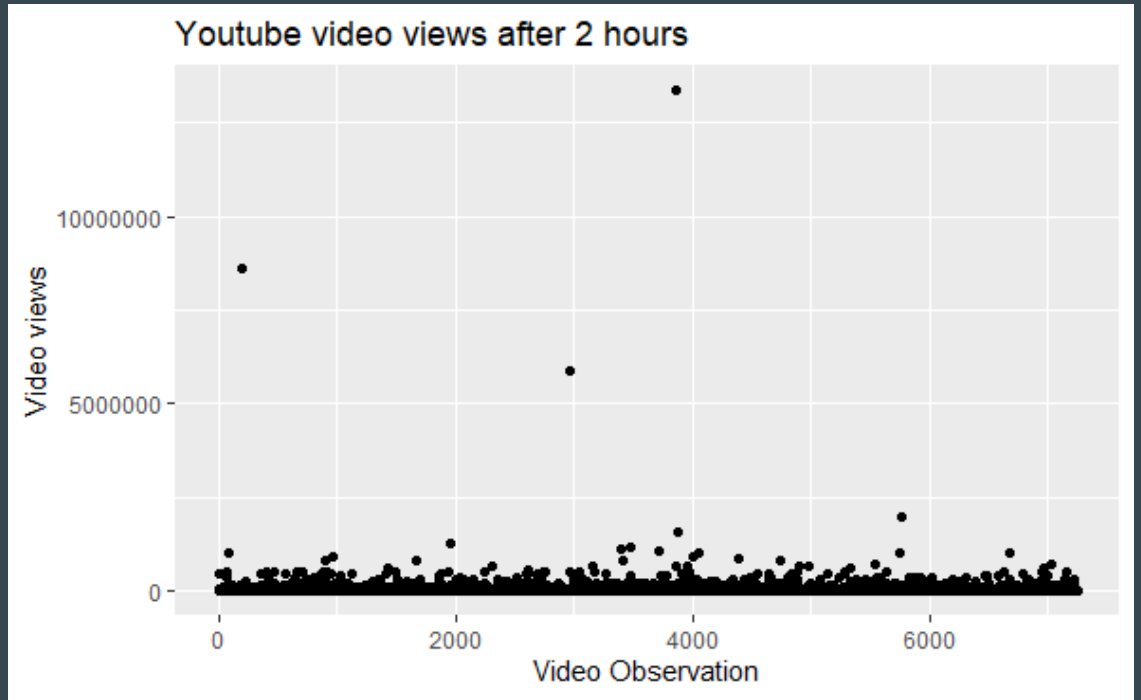
Chenxin Yang, Jonathan Martinez, Kevin Chen

Introduction

Can you predict youtube video view growth?

Introduction

Can you predict youtube video view growth?



Methodology- Data Cleaning

	PublishedDate <chr>
1	4/17/2020 10:38
2	8/31/2020 9:56
3	8/16/2020 12:15
4	8/22/2020 9:00
5	8/22/2020 14:18
6	7/24/2020 21:16

Removed the ID column

Extract the hour data from the video published date attribute

	PublishedDate <dbl>
1	10
2	9
3	12
4	9
5	14
6	21

Converted all of the attributes to numeric variables.

Replaced any missing data with the column means of the attributes.

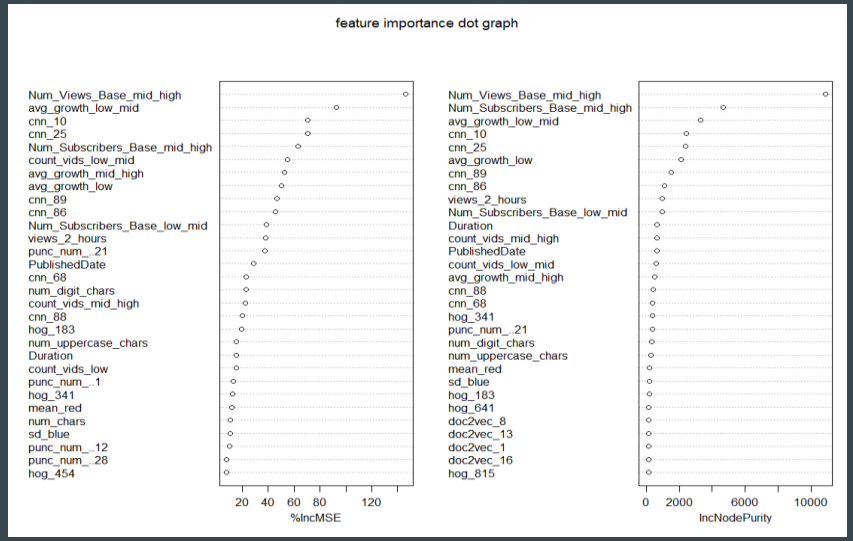
Removed attributes that had zero variance as they will not affect the video view.

Methodology- Feature selection

Analyzed the correlation matrix & Removed attributes with an absolute correlation of 0.8 or higher

Methodology- Feature selection

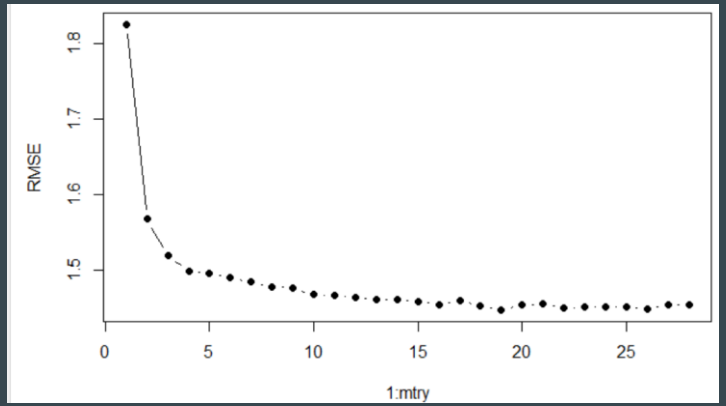
Performed bagging on the processed data ranked the importance of features and selected top 20% of most important features.



Methodology - Model selection

Tried different methods like PCA, SVM, and Random Forest. Random Forest works best.

To select parameter mtry in the model, we split 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 validation set.



Methodology - Model tuning

mtry=28 will generate the best kaggle score. When mtry=28, the random forest model becomes the bagging model.

Results

Kaggle score of 1.37725 with 40% of the test data

Kaggle score of 1.39752 with 60% of the test data

The evaluation metric used is RMSE.

Model 4 Kaggle score: 1.42100

Conclusion

Highlight & Recommendations

- Two layers of variable selection -> extracted the important predictors (260 -> 28)
- Our ranks did not change -> good balance between bias and variance