Video Channel Presentation

Predicting Growth Rate of Youtube Views

Team: Dada’s Lambda
Member: Qinyi Chen, Yue Wu, Hanyue Zhang

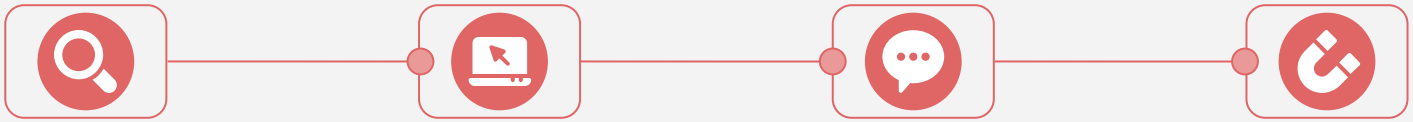
01. Introduction

How to make the views of videos grow faster?

02. Data Pre-processing

How to improve the explanatory power?

Data Pre-processing



Data Dimension

260 variables and 7242 observations

Cross-validation

randomly split 30% of the data into a validation dataset

“PublishedDate”

split in to “month” and “date”

Combine Levels

merge 12 dummy variables to 4 categorical variables

Pre-processing: Data Transformation

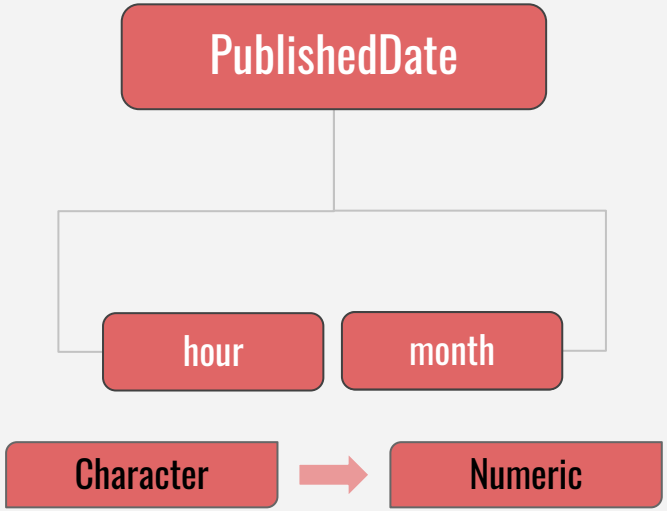
"PublisedDate": indicates when the video was published on YouTube, including the date and time

$$Formula1: "month" = Month + \frac{Date}{days\ of\ this\ month}$$

$$Ex1: December\ 12^{th} = 12 + \frac{12}{31} = 12.39$$

$$Formula2: "hour" = Hour + \frac{Minute}{60}$$

$$Ex2: 3:20 = 3 + \frac{20}{60} = 3.33$$



Pre-processing: Data Transformation

Combine Levels

Avg growth low	Avg growth low mid	Avg growth mid high
1	0	0
0	1	0
0	0	1
0	0	0

Avg growth
1
2
3
4

- 4 Channel Features (3 variables for each):
  - Num_Subscribers
  - Num_Views
  - Avg_Growth
  - Count_Vids
- 4 levels of each feature
  - Low
  - Between low and medium
  - Between medium and high
  - High

03. Feature Selection

Which predictors are significantly important?

Feature Selection: Low-variance

If the variable's variance is very low or close to zero, it indicates that this variable cannot provide significant contributions to the prediction of the model.

Set 100 non-zero observations of each variable as our simple-baseline approach to remove features.

Remove all variables that contain no more than 100 non-zero observations.

As a result, we remove 31 variables.

Example:

	cnn_0	punc_num_..2
1	0	0
4	0	0
6	0	0
7	0	0
9	0	0
10	0	0
11	0	0
12	0	0
14	0	0
15	0	0
16	0	0
17	0	0
18	0	0
20	0	0

Feature Selection: Lasso

Lasso is one of the shrinkage methods that force some non-significant variables' coefficients towards 0.

Remove variable, which has a 0 coefficient, as our second baseline approach to further feature removal.

We remove 107 variables.

Example:

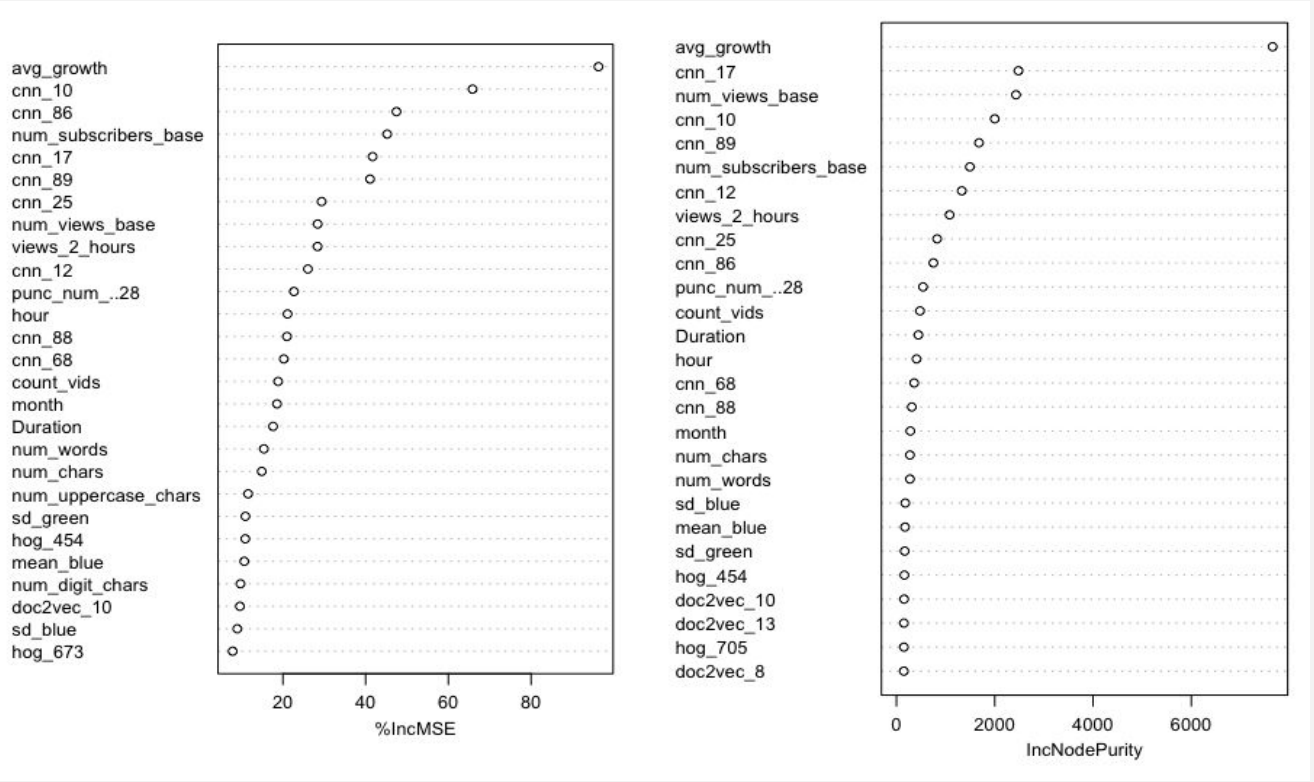
	s0
count_vids2	7.467583e-01
count_vids3	-5.532219e-02
count_vids4	-8.782446e-02
avg_growth2	-2.272451e-01
avg_growth3	2.770488e-01
avg_growth4	2.349401e+00
Num_Views_Base2	-2.982982e-01
Num_Views_Base3	1.971340e-01
Num_Views_Base4	.
Num_Subscribers2	9.866527e-01
Num_Subscribers3	.
Num_Subscribers4	-1.826909e-01
hour	5.297481e-04
Duration	-5.251272e-05
views_2_hours	-2.078658e-08

Feature Selection: Random Forest

Since the dataset is high dimensional and it includes a large number of correlated predictors, the Random Forest method going through all possible splits on a smaller predictor subset size m is usually helpful.

We used a typical m = p / 3 = 37 to construct a regression Random Forest model.

Feature Selection: Random Forest — Variable Importance



The variable's importance of the Random Forest model gave us some potential insight into the variables predicting power.

We keep 26 variables as the final predictors.

04. Statistical Model

Which model has the best predictive capability?

Statistical Model

Method	Validation	Training
Random Forest	**1.468522**	**0.592646**
Bagging	1.470501	0.587117
Lasso	1.678724	1.653955
Ridge	1.678905	1.654272
Linear Regression	1.679488	1.653883
Boosting	1.595208	1.453578

We apply six methods, Multivariate Linear Regression, Lasso, Ridge, Bagging, Random Forest, Boosting, to the training dataset, and compute the root mean squared error on both training dataset and validation dataset.

This table shows the comparison of the RMSE estimation from these six methods.

Results

1.39081 — Kaggle Public Leaderboard

1.41226 — Kaggle Private Leaderboard

1.468522 — Validation (30%)

05. Conclusion

What is the strength of our model?

Conclusion

- Strength
  - Data Transformation (significantly contribute to the prediction of the model)
  - Model Selection (comparison all six methods to compute the best one)

Video Channel Presentation

100.000

Thank you for watching