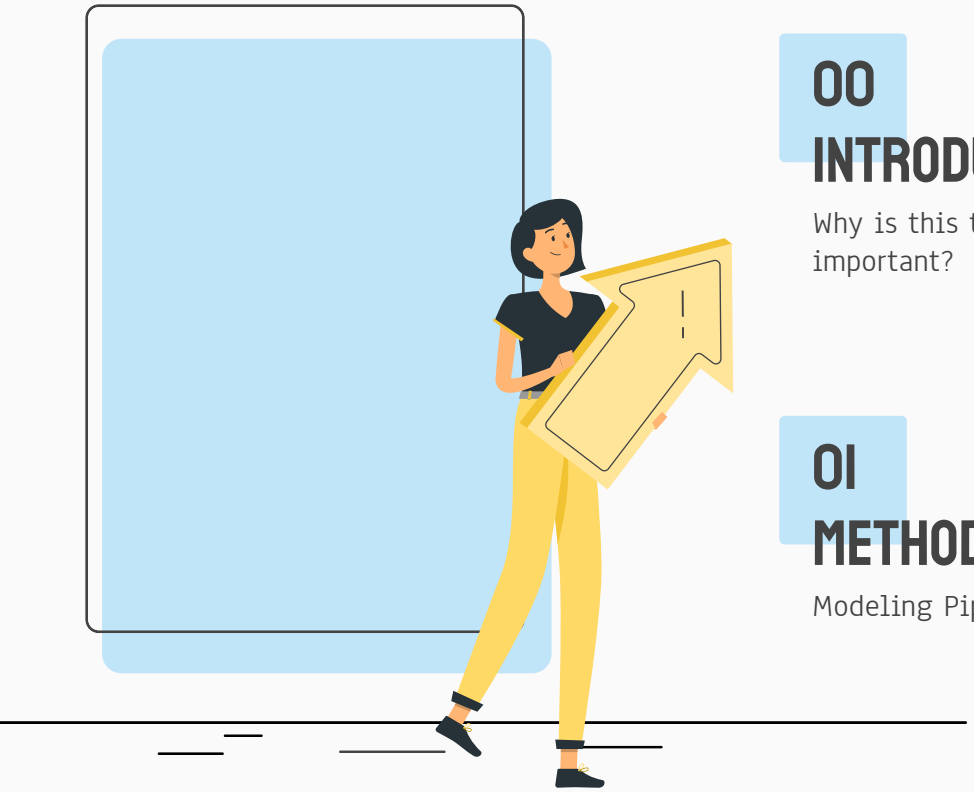
# PREDICTING YOUTUBE VIEW GROWTH WITH RANDOM FOREST

Eustina Kim
Tiffany Feng
Stats 101C Lecture 4



## 00 INTRODUCTION

Why is this this topic important?

## 01 METHODOLOGY

Modeling Pipeline

## 02 RESULTS

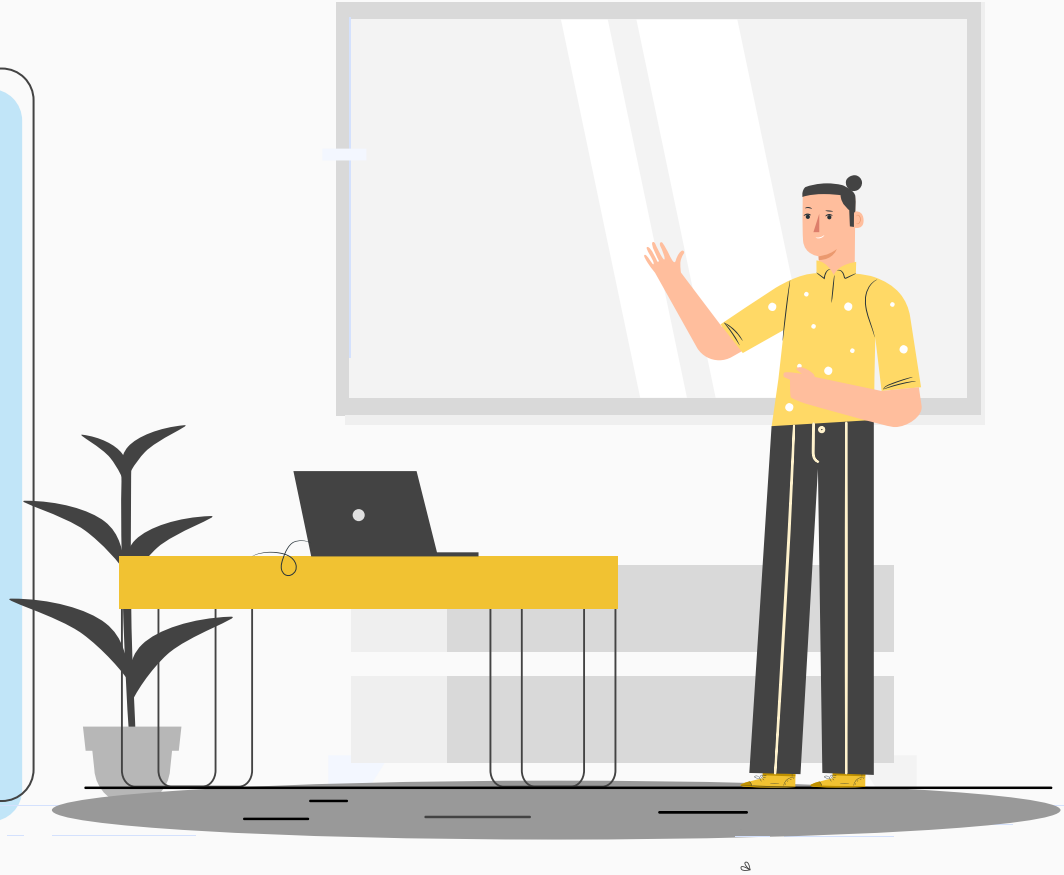
Model Performance

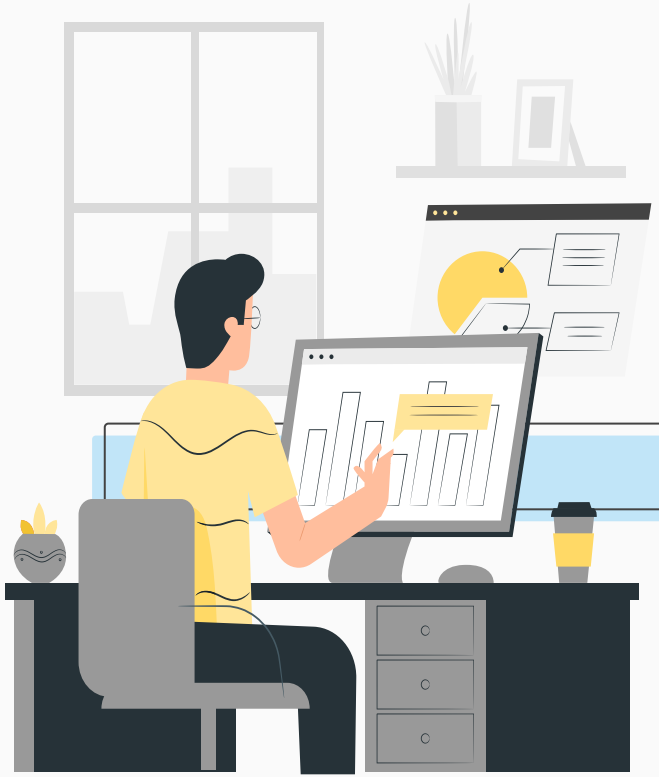
## 03 CONCLUSIONS

Model Strengths and recommendations

# 00.
# INTRODUCTION

Why is this this topic important?





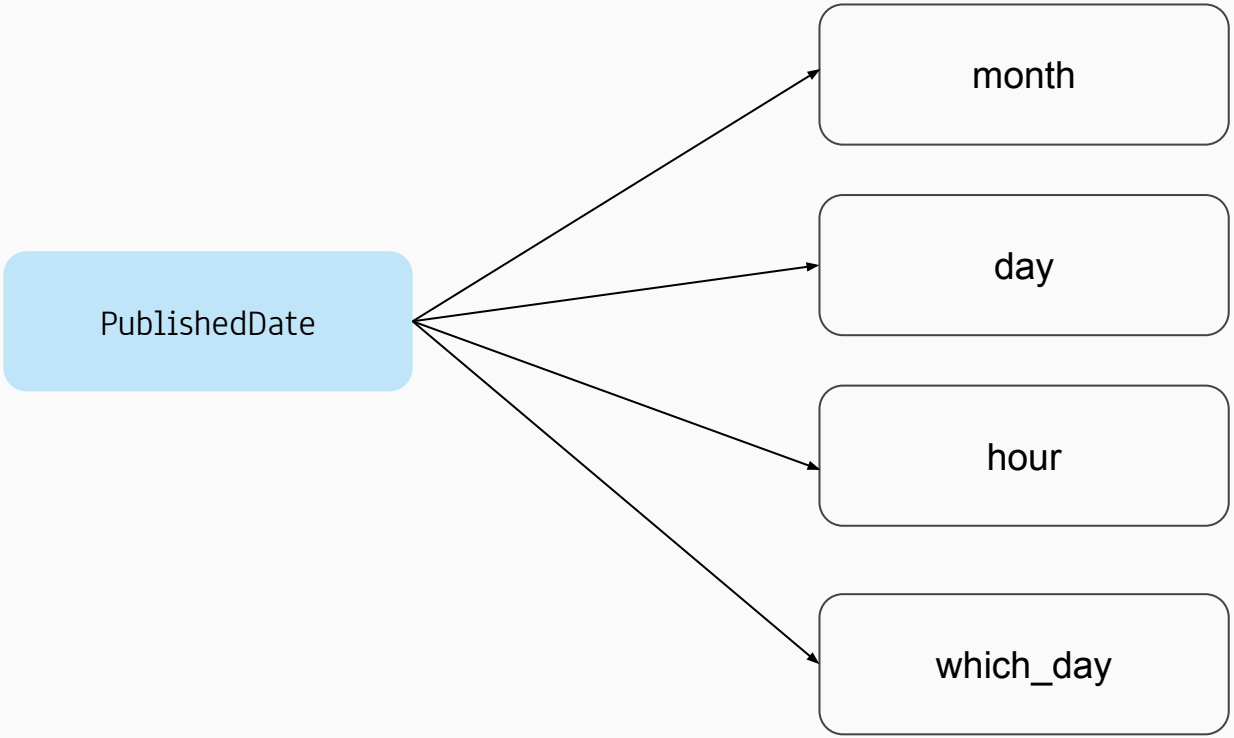
"When people are making the decision to put a piece of content online, they really do truly want to get it in front of the largest audience."

—CHAD HURLEY, CO-FOUNDER OF YOUTUBE

# 01.
# METHODOLOGY

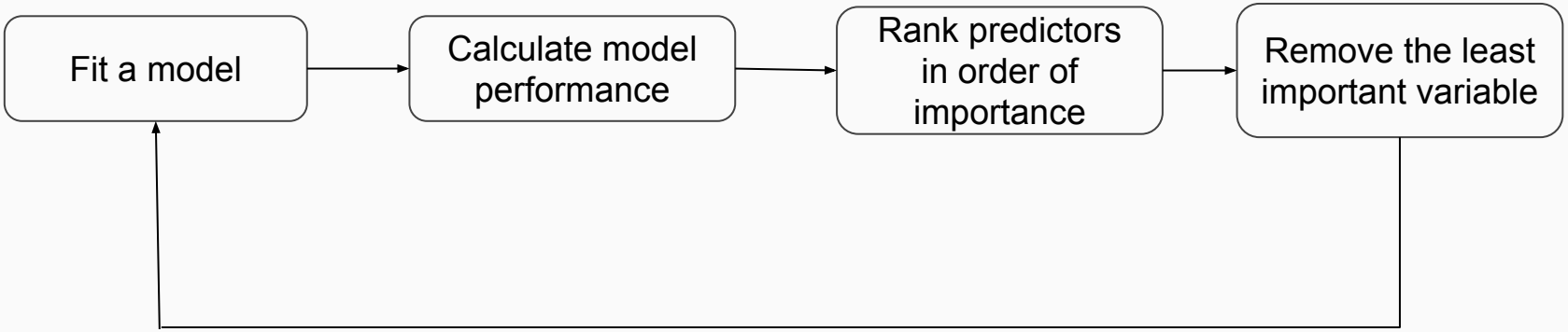
Modeling Pipeline

# PRE-PROCESSING: DATA PREPARATION



# PRE-PROCESSING: VARIABLE SELECTION

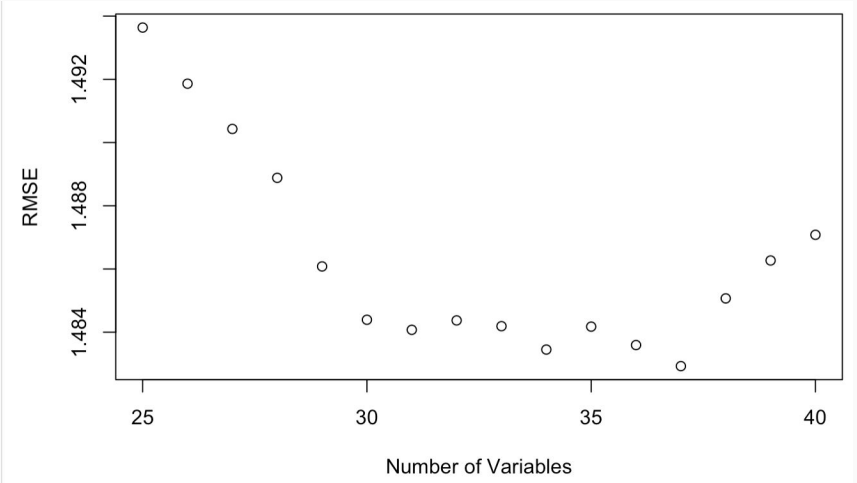
- Random Forest-Recursive Feature Algorithm (RF-RFE)




- Result: Returns a subset of predictors that gives the best model performance

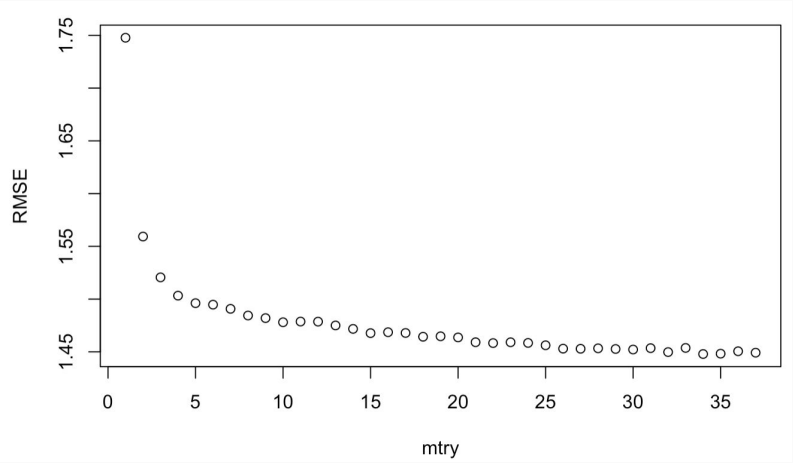
# PRE-PROCESSING

- Advantages of RF-RFE
  - Recommends a subset - less burden on the user!
  - More flexible - not limited to linear regression
  - Reduce the effect of correlations (Gregorutti et al. 16)
- Result on our data:
  - 37 predictors
    - 14 thumbnail image features
    - 9 video title features
    - 9 channel features
    - 4 other features
  - RMSE = 1.483



# FINAL MODEL

- Final Model:
  - Random Forest with 37 predictors, mtry = 37, ntree = 500
  - 10-fold cross validation
  - technically bagging because # of predictors = mtry
  - RMSE = 1.445015



# 02.
# RESULTS

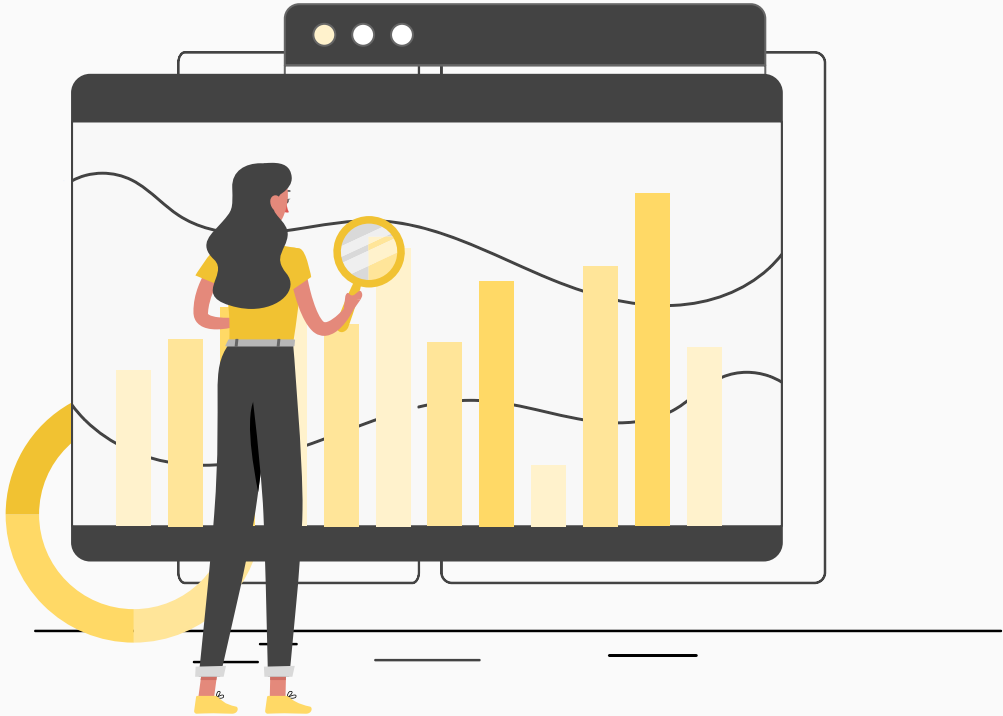
Model Performance

# RESULTS

- Scores on Kaggle:
  - Public = 1.37229
  - Private = 1.39725
  - 1.81886% increase in the score ⟶ Model performance is consistent!

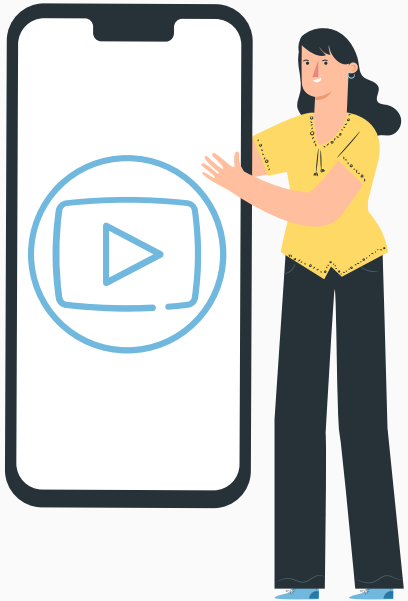
# 03.
# CONCLUSIONS

Model Strengths and recommendations



# CONCLUSIONS

- Advantages of random forest:
  - Works well with large, complex datasets.
  - Non-parametric.
  - Robust to outliers.
- Why our model works:
  - parameter tuning
  - cross validation
- Recommendations for improving model performance:
  - More data pre-processing of variables.



# THANK YOU!

Eustina Kim
Tiffany Feng
Team Chi Beta Gamma
STATS 101C, Lecture 4