Predicting Recommendations of Comments on New York Times Articles

Statistics 412 Final Project

Lucy Baden

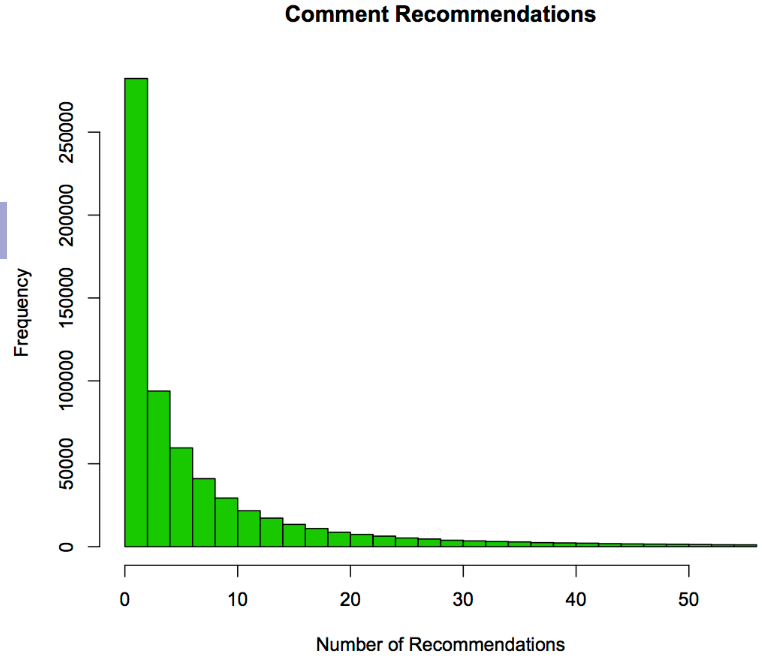
June 7, 2018

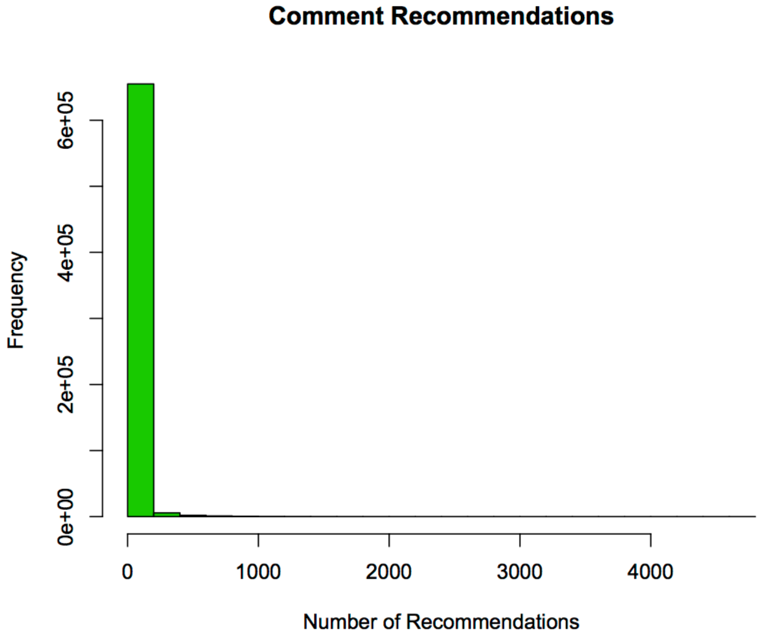
Overview

- Data
- Feature Analysis and Creation
  - Sentiment Analysis
  - Gender
  - Other
- Modeling
  - Gradient-Boosted Machines
  - Random Forests
  - Other
- Conclusion

Data

- Two data files, train_comments and train_articles
- Challenges:
  - □ Many observations (665396)
  - □ Feature selection and creation
  - □ Recommendations is highly skewed

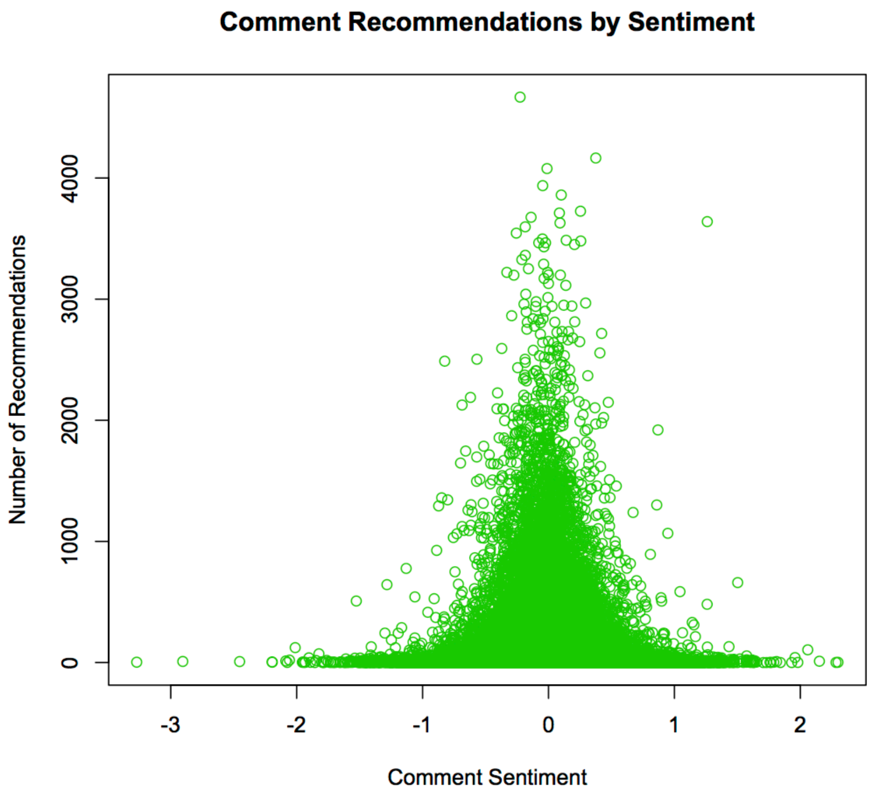


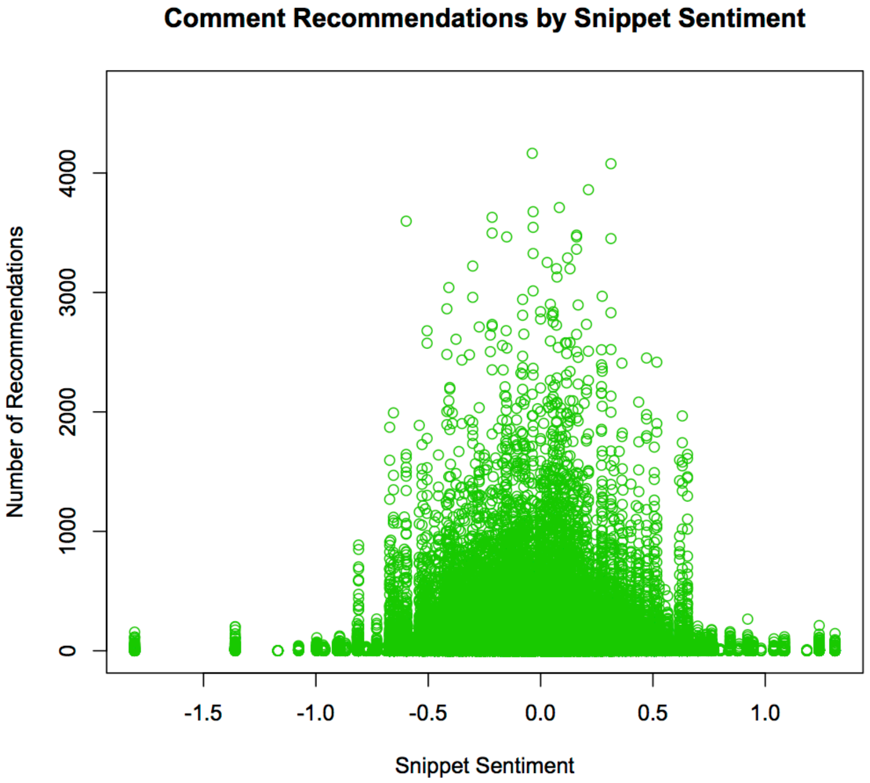


Feature Creation

Sentiment Analysis

- Comment body sentiment and word count
- Article snippet sentiment and word count
- sentimentr package

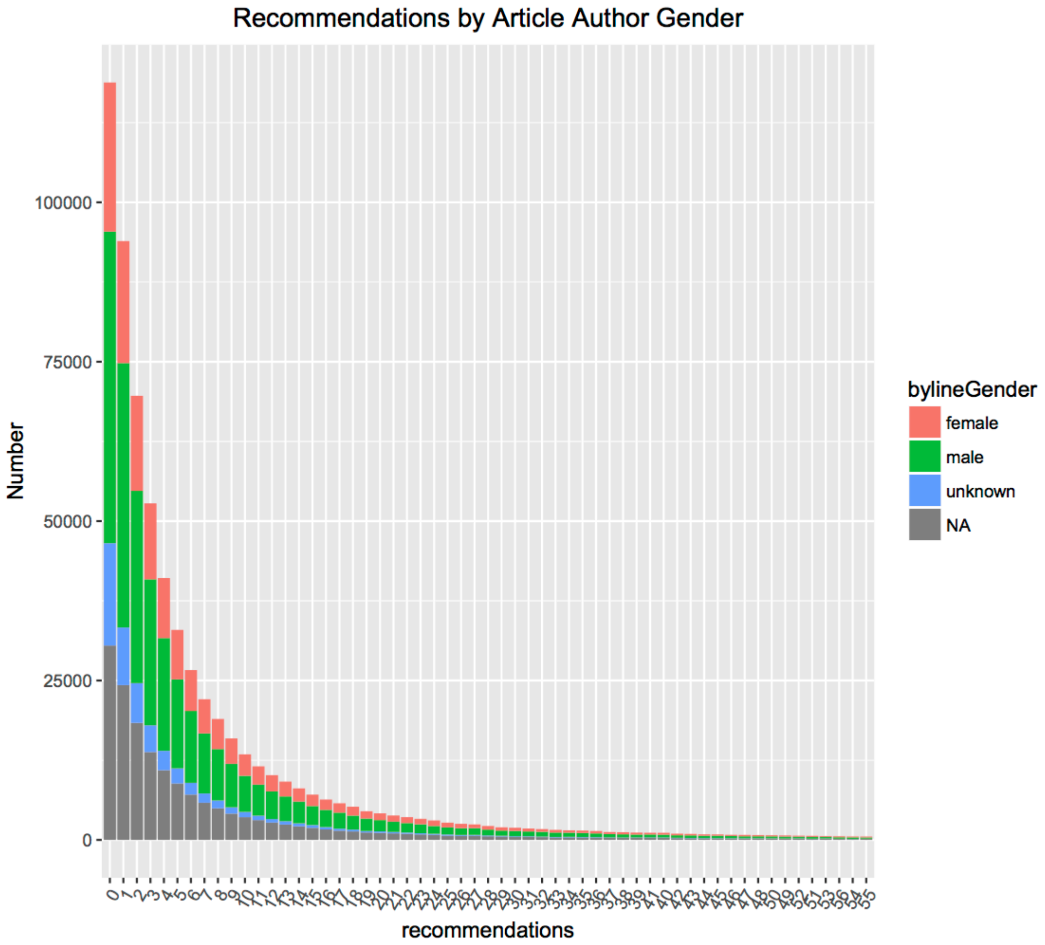

Comment Recommendations by Sentiment


Comment Recommendations by Snippet Sentiment

Feature Creation

Commenter and Author Gender

- Took commenter first name from userDisplayName
- Took author first name from byline
- gender package
    - Probability that the name is male/female
    - Social Security data
    - Non-name usernames were categorized as ‘unknown’
- Significant difference in recommendations mean between gender categories



Feature Creation

Other Features

- Keywords: topKeyword, topKeywordBin
  - Created a list of all keywords
  - Found which keyword for each article appears the most frequently on the list: topKeyword
  - Top five keyword categories: topKeywordBin
- Bins for categorical variables: typeOfMaterial, newDesk, and sectionName
- Approval time
- Default pic

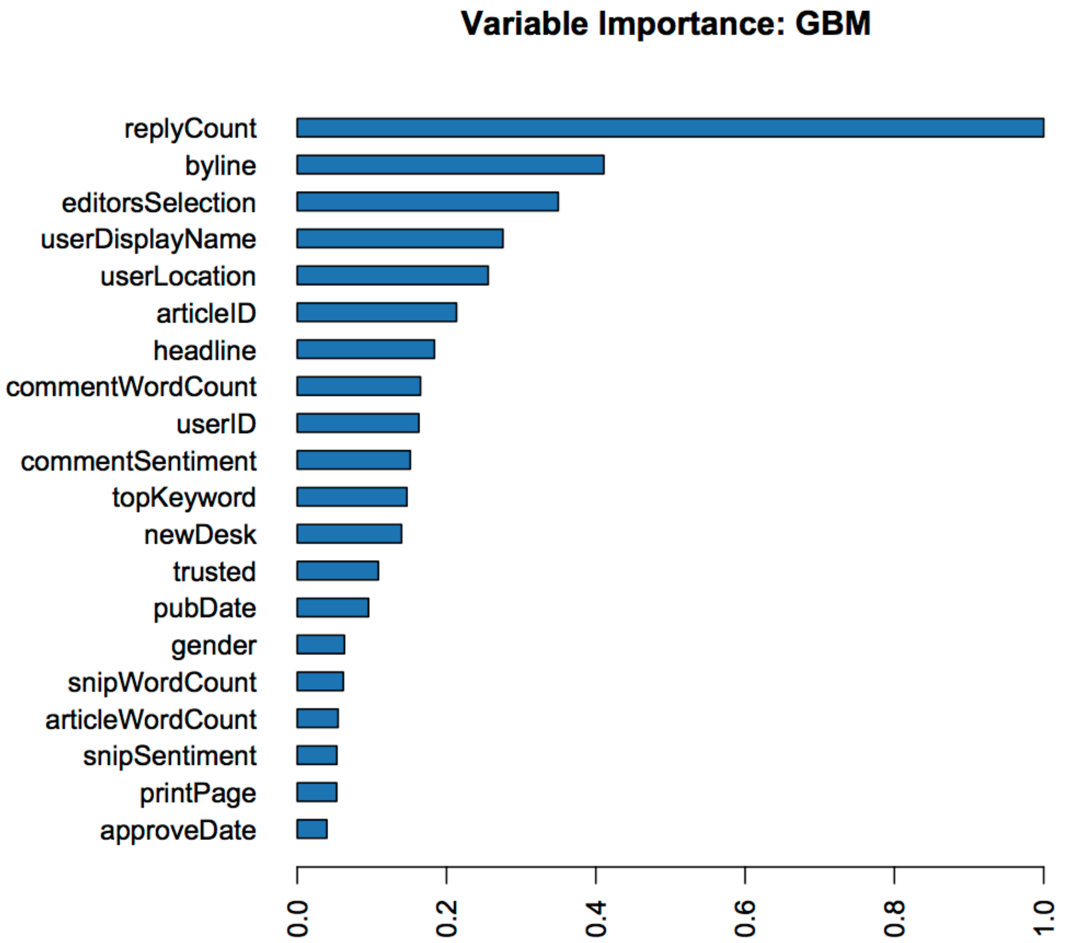
Modeling

- Trained models on 75% of the data
- Validation set for model tuning
- Feature selection and parameter optimization
- Random grid search
- Best models re-fitted on all data and submitted to Kaggle

Modeling

Gradient Boosted Machine

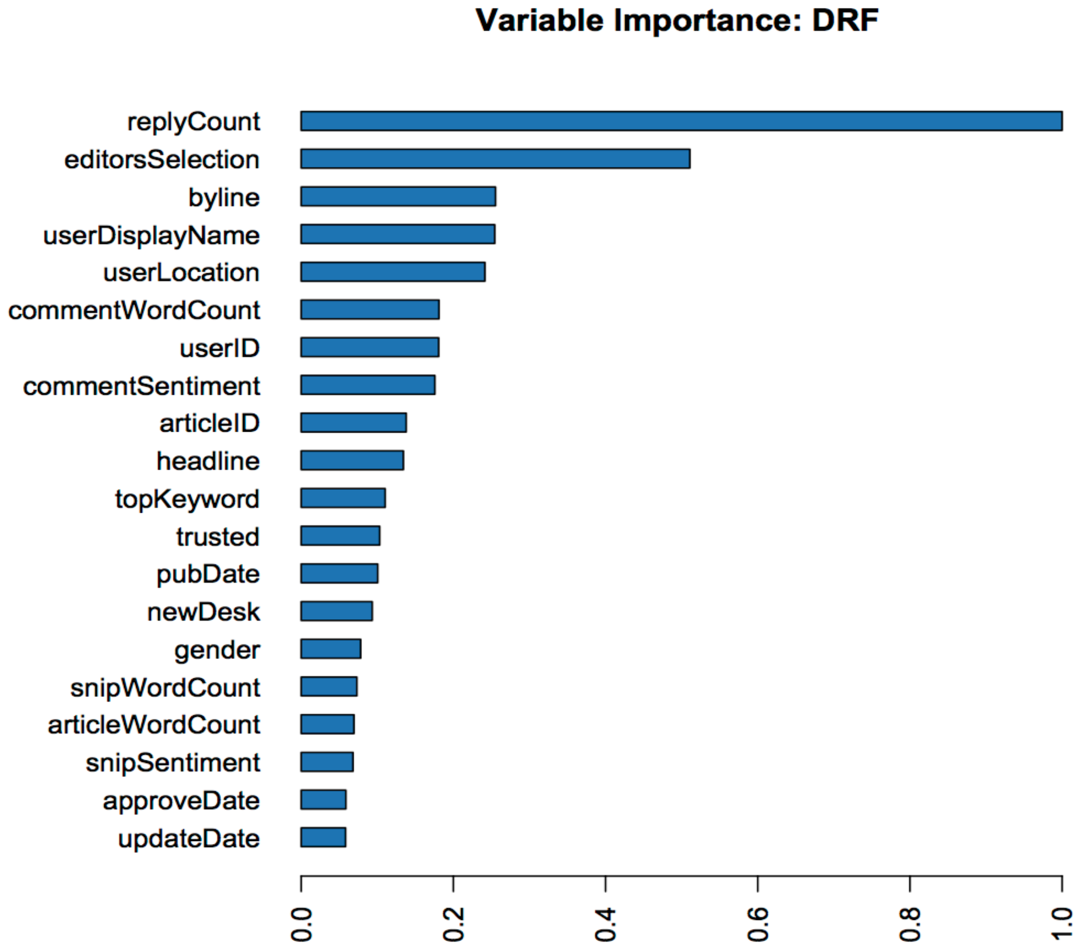
- Best GBM had an MAE of 16.01 on the validation set, 16.9 on Kaggle test set
  - 200 trees, max depth = 13, learn rate = 0.05, min_rows = 3



Modeling

Random Forest

- Best random forest had an MAE of 16.5 on the validation set
  - 50 trees, max depth = 20, min_rows = 1
- In general, random forests did not perform as well as GBMs



Modeling

Other Models

- Neural networks
  - Slower, required reduced variable set, worse MAE
- GLMs
  - H2O, caret, MASS, and base R
  - Significantly worse performance

Conclusion

- H2O gradient boosted machines performed best followed by random forests
- Best model MAE:
  - 16.0 on the validation set
  - 16.6 on the public Kaggle leaderboard
  - 16.9 on the private Kaggle leaderboard
- Sentiment analysis and gender features were relatively important in the models, while binned categorical variables were among the least important
- Best models were about 0.8 better than random numbers on the Kaggle test data

Questions