Predicting podcast popularity in iTunes

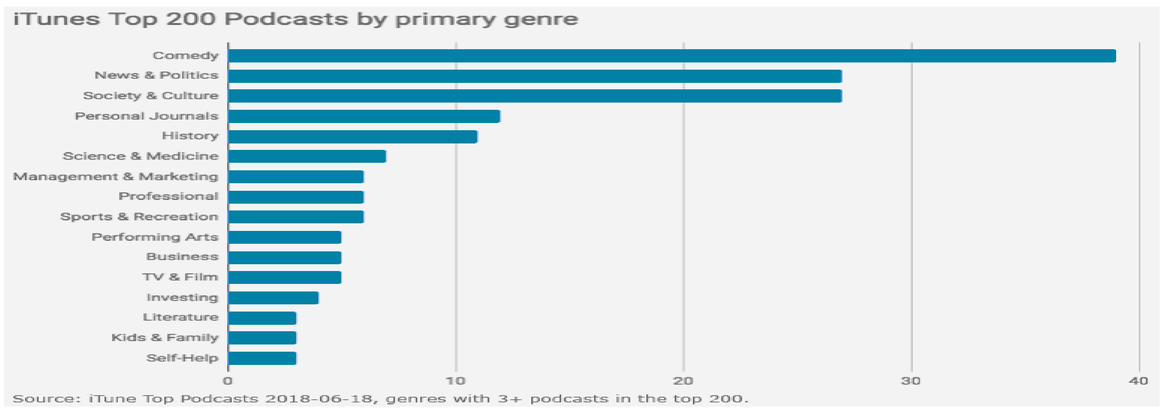
Inwook Back, Shu Li. Mentor: Dr. Zhaozhi Fan

Department of mathematics and statistics, Memorial University

BACKGROUND

- The popularity of podcast have steadily increased

- But podcast shows uneven popularity among genres

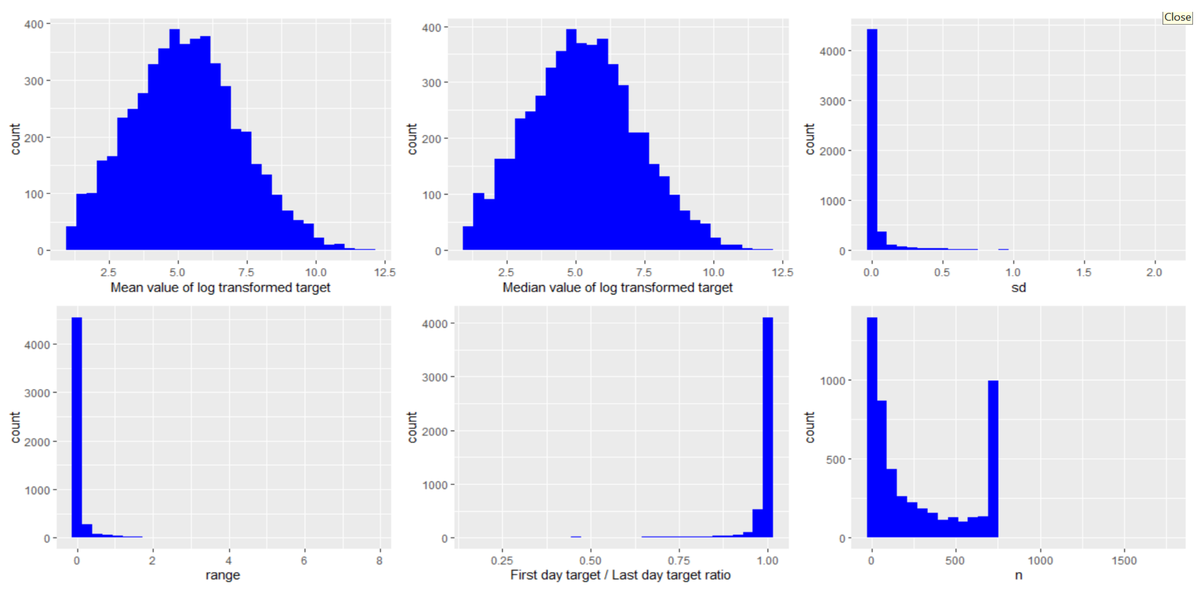


- So, it is meaningful to study which contents are more likely to be popular

PURPOSE OF THIS STUDY

- To investigate various features such as from genres, artists to text data in the summary information of each podcast to predict its popularity
- Popularity could be evaluated using the weighted average of the number of reviews and their rating values. So, our main focus is to predict the number of reviews for podcasts.

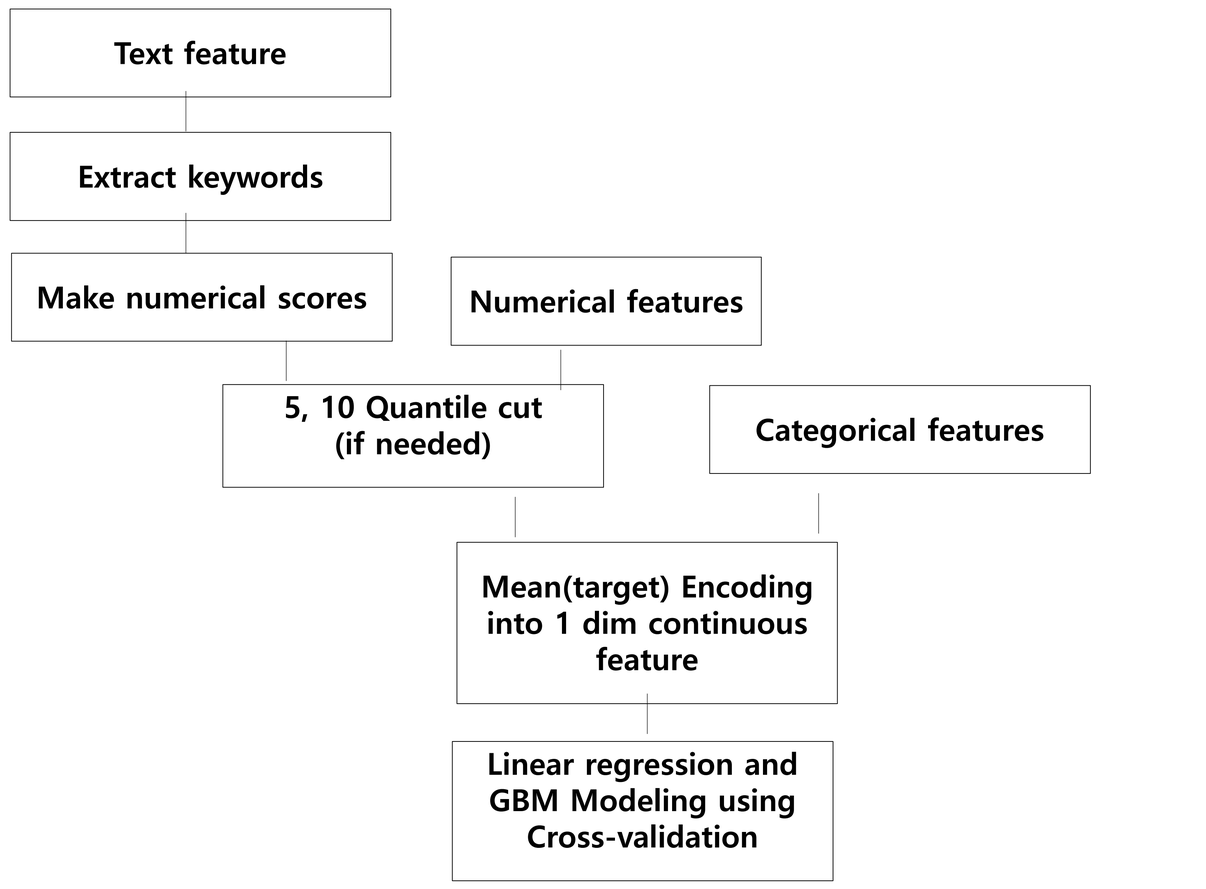
PREVIEW OF THE TARGET VARIABLE



- **Target Variable(Log transformed number of reviews)**
  - ❖ After log-transform, skewness to the right tail was removed effectively.
  - ❖ Most of podcasts(95%) show very small variation in their time series, which means that time effect could be limited.
  - ❖ Our strategy to predict the target variable is to split the scenarios:
    - First : Common titles in train and test set with small variation in time series
    - Second : Common titles in train and test set with large variation in time series
    - Third : New titles only appeared in test set.
  - ❖ For the first case, we used the last day target value in the train set to all prediction, for the second case, we used simple linear model with day count as a offset term, and for the third, we used the fitted value from our final model using other features.

- **Data structure and other variables**
  - Train set size : 1408901 x 10. 5148 titles, Test set size : 462885 x 9. 4087 titles
  - Text feature and Categorical features : Summary, Subcategory, Artist
  - Numerical features : Rating
  - Date : Data scraped date, Podcast released date

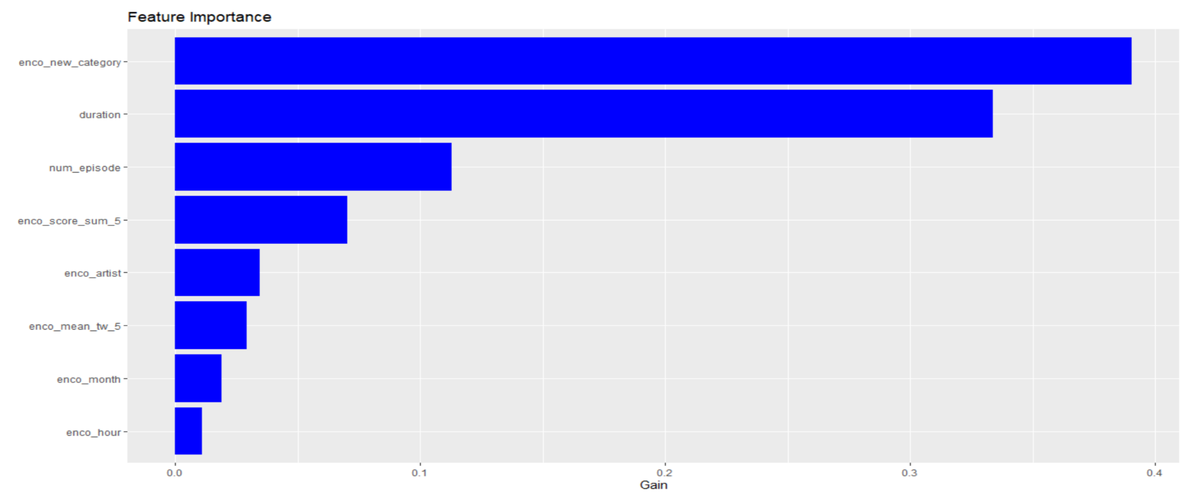
FEATURE ENGINEERING AND MODELING RESULTS



- Text feature : We extracted casting names and keywords from summary column, and then make the scores using these words and name features based on their size of effects.
- To make the subcategory to lower dimension, we leveraged the dependency structure of "gloVe Wiki 6-billion words embedding matrix(external data) using cosine distance of each word vector
- We used quantile cutting for continuous variable to avoid the overfitting, and changed categorical variables into continuous 1-dimension feature using "Mean(target) Encoding"

- **Features that are used in the model**
  - Text related : Keywords and Casting names from the summary
  - Categorical features : Subcategory, Artist(Production), Released hour and month
  - Numeric features : Duration of scraped dates, Number of Twitter followers of casting names(External data), Number of new episodes added

Model features	Estimate	S.E.	P-value
(Intercept)	-10.24	0.41	0.00
Twitter follower	0.36	0.03	0.00
Duration of scraped dates	0.05	0.00	0.00
Number of new episodes added	0.03	0.01	0.00
Subcateogry(numeric score)	0.76	0.02	0.00
Artist(numeric score)	0.43	0.04	0.00
Hour(numeric score)	0.27	0.05	0.00
Month(numeric score)	0.18	0.03	0.00
Keywords(numeric score)	0.79	0.04	0.00
Adjusted R-squared	0.55		



CONCLUSIONS

- Overall, the R-squard was 55% which means this model is not very reliable for prediction
- All variables are significant and their standard errors are also stable. From the feature importance, subcategory, total days posted, number of episodes and keywords effects are so large, so we can mainly use subcategory and keywords in the summary to predict the popularity
- The strength of our method is robustness. 10-kolds CV-error was stable in many times of simulations and the error was also robust when we changed the hyper parameters in the model. This means we effectively removed noises from the features.
- The other merit of our method is that we only used 8 features in the model, and the whole process is reproducible for the new test data