2019 SSC Case Studies in Data Analysis Poster Competition

Case Study #	2
Team Lead (bold) and Members	**Inwook Back** Shu Li
Department and University	Mathematics and Statistics Memorial University of Newfoundland
Supervisor	Dr. Zhaozhi Fan
Project Title	Predicting podcast popularity in iTunes

Introduction

The main focus of this analysis was extracting useful features from the text data. Specifically, keywords and casting names in the raw summary text were extracted using NLP(Natural Language Process). One of the methods used was the Mean (target) Encoding technique, which is widely used in practice to treat high dimensional categorical variables. Linear regression and gradient boosting models were used for modeling.

Objective

The main objective is to predict the number of reviews for each podcast. One important thing is that only a small number of podcasts have large variation in the target time series and the time effect could have limited usefulness. High dimensional categorical features in the data should be carefully handled to minimize possible overfitting issue.

Methods

From the investigation of the target variable, the initial number of reviews are very similar to the last number of reviews among 95% of podcasts. So, we divided podcasts into two categories: one has small variation in time series, and the other has large variation.

For the text data analysis, we used NLP techniques such as “Name Entity Recognition” to extract human names from the summary, and “gloVe word embedding matrix from 6-billion words corpus in Wikipedia” to re-group the subcategory feature. The reason why we did this is because there are too many different categories, where many of them are very similar and can be merged. To extract keyword effects, we tested linear regression model, with other covariates fixed, for all words in the summary column that appeared in more than 25 different podcasts. This is imposed to avoid noise.

The next step was to convert categorical features into continuous features using Mean (target) encoding with 2 phases cross-validation. Methods used for modeling were linear regression, XGBoost, and Random forest. In the hyperparameters tuning in GBM models, we used Bayesian Optimization using cross-validation error.

Results

The important features are “subcategory”, “duration of dates scraped”, “number of new episodes added”, “scores from summary keywords”, “the number of Twitter followers of the artists (the popularity of the artists)”. From the linear regression, the R-squared was 55% for the train dataset and 51% for the validation dataset, which were not very good. The RMSE for the validation set (20% of the train dataset) was 1.32.

Conclusion and Discussion

From the R squared (55%), our model was not very reliable to explain the target variable and should be used with caution. More information or variables need be collected to improve the prediction precision. Our method's strength is that our cross-validation error was very stable in many times of simulations, and robust even when we changed the hyperparameters in the model. This is because we used the mean (target) encoding, which reduces the dimension effectively. Also, we have a small number of features in the model, which increases the interpretability as well.