Forecasting iTunes podcast popularity out-of-sample: using in-sample validation for longitudinal and matching methods.

Melissa Van Bussel[1], Dylan Spicker[2]. Faculty Mentor: Dr. Song Cai[1].

1: School of Mathematics and Statistics, Carleton University, 2: Department of Statistics and Actuarial Science, University of Waterloo.

BACKGROUND

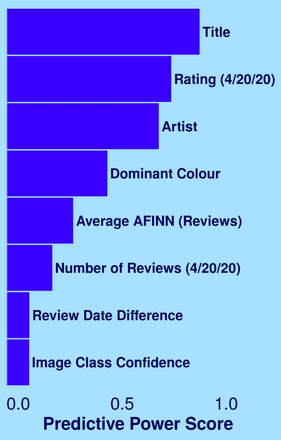
Apple Podcasts supports streaming of audio and video podcasts for free through the Podcasts app for iOS or through iTunes for Mac and Windows computers. Although Apple does not directly produce podcasts, Apple Podcasts users can subscribe to podcast episodes from a virtually unlimited number of podcast producers from all over the world. Listeners can express their opinions of a particular podcast by leaving a "rating" (5-star system where a higher number of stars means a better rating) or a "review" (text commentary). Apple Podcasts data provide the opportunity to gain invaluable insight into viral culture and trends in media.

OBJECTIVES

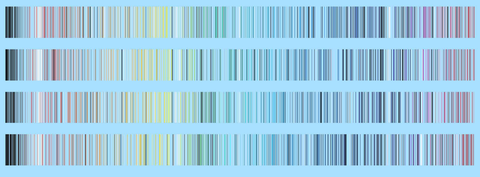
For this case study, we were provided with metadata on 6202 unique audio podcasts available through Apple Podcasts. These data were collected through automated hourly scraping performed from November 1st to December 1st, 2019 for the training data and from January 10th to 20th, 2020 for the unlabelled dataset. We aimed to use the provided metadata along with newly engineered covariates to determine predictors of podcast popularity, both in terms of their mean rating value and the number of reviews. We sought to discover temporal trends and predict the number of reviews for the podcasts in the unlabelled dataset.

POPULARITY

We created several new covariates to determine which characteristics of podcasts predict their popularity. We did this by scraping the webpage of each podcast and engineering new variables. The Predictive Power Score[1] is plotted (right) for each of the covariates, where higher scores represent a stronger ability to predict the number of reviews (note that this ordering was effectively the same in predicting rating value). Aside from podcast title and artist, we found that the current rating value (as of April 20th, 2020) as well as the dominant colour in the cover art of a podcast were strong predictors of popularity. Information about the 3 most recent reviews of a podcast (such as the average sentiment score as evaluated by the AFINN lexicon, and the number of days that passed between the reviews) also had some predictive power. The YOLO Object Detection System[2] was used to classify the objects in the cover art, but this did not prove to be useful.

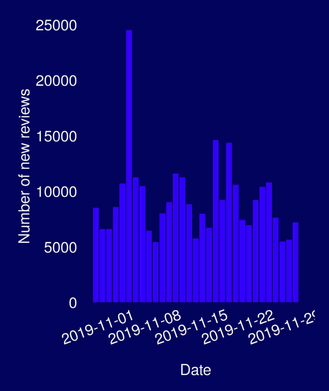


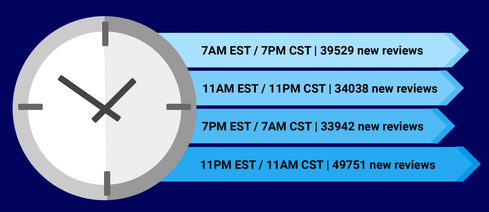
We also performed a colour analysis of the dominant colours in the cover art of each of the podcasts in the training data. The colour spectra below are grouped by quartiles based on the maximum number of reviews for each podcast. We found that the distribution of colours for popular podcasts is more even, indicating that more "unique" colours are used in these images. Many high-hue reds are used by popular podcasts.



TRENDS

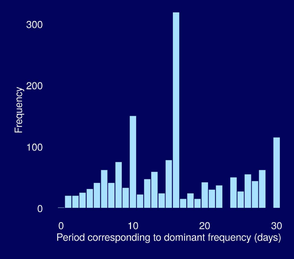
To determine whether the trends in popularity (in terms of number of reviews) change over the timespan of the dataset, we plotted the number of new reviews during each day across all podcasts in the training data (right). We observed a weekly trend, where weekends tend to receive fewer new reviews than weekdays (the first day in the plot is a Friday). We suspect this is because many people listen to podcasts at work or while commuting.





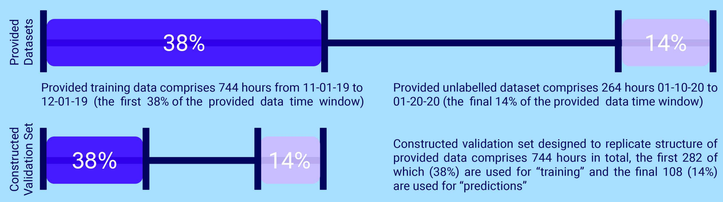
We then repeated a similar process to explore daily trends. Examining the number of new reviews grouped by hour, we found peaks at 7am and 11pm both in EST (America) and CST (China), suggesting that many people listen to podcasts during their morning commute or right before bed.

We performed spectral analysis on each of the time series in the training data. The periodogram for each series was used to extract the dominant frequency. The analysis revealed that periods of 15 days were most common, followed by 11, 30, and 7 days. We noticed that some of the podcasts had periods that matched with their episode release schedule.



PREDICTIONS

In order to perform model selection, we constructed a validation set. This simulated the characteristics of the provided data, using the set of podcasts in the training set not included in the unlabelled dataset. Models were fit based on data falling in the first ~38% of the observed window. The MAE for the models was computed using observations in the final ~13% of the validation set window. The remaining data (~49%) were not considered. MAE results on the validation set are shown in the table.



Noting that "title" was a highly predictive variable, we split the data into individual time series, with one for each podcast. The predictions were categorized into two cases:

1. Test podcast present in training data
2. Test podcast not present in training data

The selected model involved an out-of-sample forecast for the podcasts in case #1 based on mean interarrival times over the training set. Podcasts in case #2 were matched to podcasts from case #1, selecting the most similar *as of April 20th* (the day we scraped the current number of reviews from iTunes). The mean of the out-of-sample predictions of the matched observations were then used.

Model	MAE
Interarrival Time + Matching	**43.62**
Linear Mixed Model	59.42
Poisson Generalized Linear Mixed Model	80.32
MLPs for Temporal Hierarchies	96.22
ELMs for Temporal Hierarchies	99.18

CONCLUSIONS

We find the number of reviews to be a more meaningful measure of podcast popularity. A high number of reviews tends to mean a high rating value; however, knowing that a podcast has a high rating value does not provide reliable information about the number of reviews. Since leaving a review takes considerable effort, the number of reviews is a strong measure of listener engagement and thus popularity.

We have engineered a set of interesting covariates with high predictive power, whose usefulness extends beyond the scope of this case study. Our innovative approach combined techniques of sentiment analysis, computer vision, image analysis, missing data imputation, time series forecasting, and longitudinal data analysis. Our main limitations were related to the structure of the data. Since there were no overlapping dates between the training and unlabelled data, we had to perform out-of-sample forecasting with a forecast horizon that was larger than the length of the original time series, compounding the challenge of model validation. Many values were missing, and our covariates were mostly time-invariant which meant they were not helpful during imputation.

REFERENCES

1. 8080 Labs. *A Python Implementation of the Predictive Power Score (PPS)*. GitHub repository, https://www.github.com/8080labs/ppscore. 2020.
2. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. *You Only Look Once: Unified, Real-time Object Detection.* arXiv preprint arXiv:1506.02640, 2015.