Who Watches What: forecasting viewership for the top 100 TV networks

Denis Khryashchev[1], Alexandru Papiu[2] Jiamin Xuan[2],
Olivia Dinica[2], Kyle Hubert[2], and Huy Vo[3]

[1] The Graduate Center, City University of New York
365 5th Ave, New York, NY 10016, USA
dkhryashchev@gradcenter.cuny.edu
[2] Simulmedia
11th Floor, 401 Park Ave S, New York, NY 10016, USA
{apapiu,jxuan,odinica,khubert}@simulmedia.com
[3] The City College of New York, City University of New York
160 Convent Ave, New York, NY 10031, USA
hvo@cs.ccny.cuny.edu

Abstract. TV advertising makes up more than one third of total ad spending, and is transacted based on forecast ratings and viewership. Over time, forecast accuracy has decreased due to fragmentation of consumer behavior. Through a comprehensive study we find that an assortment of models combined with an ensemble method leads to better accuracy than any single method. This results in an 11 percent improvement over a naive baseline method, across 100 of the largest networks.

Keywords: TV networks, time series, forecasting, XGboost, Fourier, Facebooks Prophet, seasonal averaging.

1 Introduction

In this paper we examine the properties of 100 largest TV networks in US in terms of aggregated hourly viewership (impressions), and evaluate the performance of several forecasting models against a simple seasonal averaging. Understanding of the behavior of such viewership time series is vital for accurate targeted advertising for even with the rise of digital media, TV advertising spending makes up more than a third of total ad spending. It was estimated by Adweek [1] that the total TV ad spending in the US for 2018 adds up to over 68 billion dollars. The majority of it is bought and sold based on forecast ratings and impressions. Previous research [3] has found that TV forecasts have become less accurate over time due to the fragmentation of audiences and increasing number of networks. Inaccurate forecasts can lead to disruptions in the media planning process and financial losses. Thus, developing models and ways to measure forecasted performance can have a big impact on advertisers' bottom line.

The purchase of television advertisement time is mostly influenced by a television program's predicted performance. Consequently, the prediction and analysis

of television audience sizes has been covered extensively. This analysis has shown that forecasting errors are increasing over time. This trend has been attributed to a series of causes such as the fragmentation of TV audiences due to changing ethnic diversity and increasing education levels in the American population [7]. At the program level, the increased choice of programs and networks for TV viewers has been a large cause of the reduction in forecasting accuracy [7, 3]. Measurements of program quality itself are a high predictor of the error [8].

Yet another level of complexity to viewership forecasting is due to Digital Video Recording (DVR) services that allow viewers to create their own schedules and break the established viewing patterns. Zigmond et. al. [9] discovered that although up to 70% of ads are skipped in the households with DVR, some of the niche ads appear to have a much higher audience retention.

1.1 Related work

Nevertheless, numerous models have been tried to improve these forecasts, mainly at the aggregate level. Most models focus on major networks and only prime-time viewing. J. Arvidsson [2] studied short-term forecasting of on-demand video viewership comparing the performance of a neural network predictor against a simple seasonal averaging with the latter being slightly more accurate. R. Weber [10] reported that Neural Networks and general linear models provided the most accurate short- and long-term forecasts for the viewership data of the 8 major German TV networks with the SMAPE errors ranging from 15% to 28%.

Linear Holt-Winters and ARIMA models were used by R. Neagu [11] for long-term forecasting of Nielsen data with the latter model being less accurate. Pagano et. al. [12] applied autoregressive models (AR, ARX, and STAT) for short-term forecasting of TV ratings in terms of mean viewing time per household per network. The reported normalized RMSE ranged from 0.80 to 0.87.

Meyer et. al. [13] studied how forecast aggregation affects accuracy of predictors on 3 levels: population, segment, and individual. They reported regression models to slightly outperform decision trees and neural networks on all levels of aggregation, and the population level to have the most accurate forecasts.

Nikolopoulos et. al. [14] compared the performance of Multiple Linear Regression, Simple Bivariate Regression, several Neural Networks, and predictors based on the nearest neighbor analysis and human judgment on the Greek TV audience ratings in terms of mean absolute error. Top two models to achieve the highest accuracy of around 9.0 were the models based on 5-nearest neighbors and simple linear regression.

Many recent research works are focused on the effects of exogenous variables on TV viewership due to the overall growth of data collection. Wang et. al. [15] showed the influence of Belgian Pro League soccer games schedule (kickoff time, month, and opponents) on TV viewership and stadium attendance. Gambaro et. al. [16] discovered that news content is a strong predictor of viewership: soft news turn viewers off and vice versa. Belo et. al. [17] concluded that the presence of Time-Shift TV that allows to watch live programs recorded on average has increased TV viewership per household by 4 minutes a day.

1.2 Contributions

Motivated by the increase of the forecasting errors, earlier works predicting TV viewership at the aggregate level, and having the domain expertise of TV advertising at Simulmedia, we make the following contributions in this paper:

- we aggregate set-top box data collected from individual households into hourly viewership time series for the top 100 TV networks in US and determine their periodicity, seasonality, and presence of trends that helps us select better parameters for our predictive models;
- we examine individual performances of 4 forecasting models: seasonal averaging that we use as a baseline predictor, Facebook's prophet, Fourier extrapolation, and XGboost;
- we build an ensemble predictor that reduces the negative effects of overfitting of the 4 individual models. It benefits from the diversity of the models that results in uncorrelated errors between each pair of the models.

2 Viewership data

2.1 Set-top box data aggregation

To accurately measure TV viewing, data scientists at Simulmedia collected viewership data from the set-top boxes of over 5 million US households using different cable providers. These data were weighed and projected to match the national Census measurements using demographic information such as age, gender, income, and presence of children. This census-weighed panel is called SimulPanel. While historically most ratings have been done on the Nielsen panel, we used Simulmedia's panel since the larger sample size allows us to achieve more precise results by minimizing the measurement noise.

The original data were comprised of viewing sequences of individual households at a minute level. We standardized the data in two steps: 1. weighed aggregation of the viewership of all the households at a minute level; 2. averaging the viewership to the hourly level. As a result, we obtained hourly level viewership time series for each of the top 100 networks. One can think of these aggregated time series as the series of the expected counts of the households that will be reached by an ad that was shown at random during that hour. For a given minute m the computations are as follows:

$$x_m = \sum_{i \in H} w_i a_{i,m} \tag{1}$$

where the sum is over the entire household set H, w_i is the household weight, and $a_{i,m}$ is a binary indicator of whether household i watched minute m. The hourly values are acquired from the minute level ones with

$$x_{hour} = \sum_{m \in hour} x_m/60. \tag{2}$$

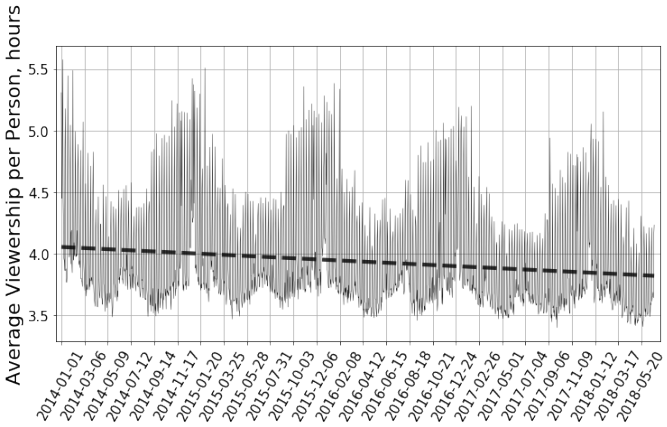

Fig. 1. Long-term trend in average daily viewership

2.2 Periodicity, seasonality, and trend

Aggregated viewership data have a great property: random individual viewing habits are averaged out allowing the global periodic patterns to emerge. Fig. 2 displays the aggregated and normalized viewership data at an hourly level for 100 most viewed TV networks in US as of March, 2018. Visually we can notice that although there are certain differences among the networks, they tend to have a strong hourly and daily periodicity (darker areas where the points overlap).

To measure the strongest periods, we applied Fast Fourier Transform (FFT) [4] with Blackman window [5] to the network viewership time series. Fig. 3 demonstrates the resulting superimposed frequency spectra for the periods within the range of 4 and 744 hours (1 month). The spectra of the top 100 networks strongly overlap and have the largest common magnitudes for the periods of 24, 12, 8, 6, and 168 hours.

TV Viewership has historically been fairly stable in the long term; however, there are certain global trends present. With the advent of digital media and streaming platforms, TV ratings have been undergoing a steady decline for the majority of networks that we forecast. We observed a 5% decrease in individual daily viewership during the period of about 4.5 years as seen in Fig. 1. It is reasonable to consider the trend to serve as a proxy for a decrease in national TV viewership. On the other hand, certain networks might follow different local trends. Overall, these trends have been analyzed in depth by K. Hubert [6].

3 Viewership Forecasting

Based on the domain knowledge, we have selected 5 better performing predictors to forecast aggregate viewership: Baseline predictor that implements simple seasonal averaging, Facebook's Prophet that being an additive regression model

extracts local trends, seasonality, and blends in important days and holidays, Fourier extrapolation that deploys prior rigorous mean averaging, XGBoost that performs gradient tree boosting, and Ensemble model that combines the predictions of the 4 individual models.

3.1 Baseline predictor

Aggregated viewership data are known to have a strong seasonality and to be relatively stable. Therefore, our Baseline model relies on a simple arithmetic averaging of 8 time periods separated by a week (168 hours) from each other. Every t^{th} element in the model's forecast $\hat{y}_{bt}$ is calculated as

$$\hat{y}_{bt} = \frac{1}{8} \sum_{i=1}^{8} x_{t-168i} \tag{3}$$

3.2 Facebook's Prophet

TV viewership follows various periodic patterns that include yearly, weekly, monthly or bimonthly seasonality as we have shown in section 2.2. However, such patterns are interfered with holidays and various local trends.

Taylor and Letham of Facebook introduced the Prophet [18], a decomposable time series model that incorporates a seasonal component, trends, customizable holidays, and an error term:

$$\hat{y}_{pt} = g(t) + s(t) + h(t) + \epsilon_t \tag{4}$$

where $g(t)$ models non-periodic changes in time series. Assuming that our data do not have non-linear saturating trends, we employed linear trend with changepoints

$$g(t) = \left(k + a(t)^T \delta\right) t + \left(m + a(t)^T \gamma\right) \tag{5}$$

where k is the growth rate, δ stands for rate adjustments, m is the offset parameter, $a(t)$ is a vector of binary values with ones corresponding to the locations of certain changepoints, and γ is included to make $g(t)$ continuous.

Seasonal component $s(t)$ is evaluated with standard Fourier series

$$s(t) = \sum_{n=1}^{N} \left[a_n \cos\left(\frac{2\pi n t}{P}\right) + b_n \sin\left(\frac{2\pi n t}{P}\right) \right] \tag{6}$$

where P stands for the period and parameters a_n, b_n are estimated. We extract weekly and yearly seasonality with $P = 7$ and $P = 365.25$ correspondingly.

Holidays and special events term, $h(t)$ contains supplementary regressors initially intended to be used for holidays. However, knowing detailed TV program schedule in advance and assuming that more popular shows are to gain higher viewership than less popular ones, we incorporate both: the information on holidays and program schedules one-hot encoded into $h(t)$.

Last term, ϵ_t represents the errors introduced by any unusual changes not accounted for by the model.

The model as a whole is optimized maximizing a posteriori probability.

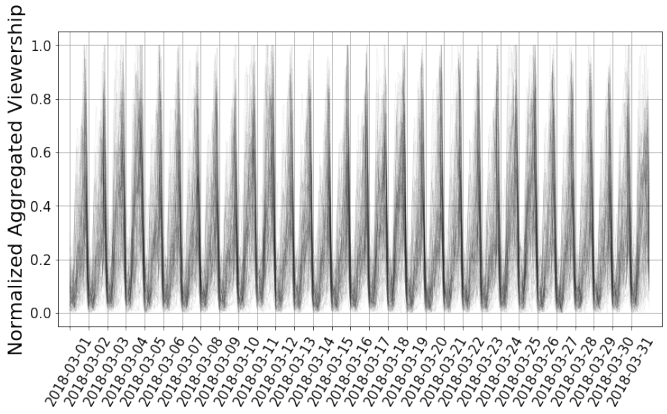

Fig. 2. Normalized aggregated hourly viewership for 100 largest networks during the March of 2018 superimposed

3.3 Fourier Extrapolation

Analyzing Fig. 2 and Fig. 3 we notice a very strong periodic pattern in viewing behavior: on a large scale network viewership has a strong autocorrelation at time lags of 24, 168, and about 1344 hours which correspond to a daily, weekly and bimonthly periodicity in the time series. Strong periodic patterns assume that a Fourier-based extrapolation might be able to capture the repetitions in the signal and efficiently extrapolate it into the future.

However, due to the constant gradual change in the set-top box data: households join and leave the panel with their weights being adjusted; there might appear certain unexpected jumps in the aggregated viewership (not to mention rare hardly predictable events like Super Bowl).

In order to decrease the the negative effects of such viewership jumps, we replace the original historical hourly viewership data with robust location values calculated with Huber's M-estimator [19] which is equivalent to an application of a low-pass filter. We use 7 previous values that are 168 hours apart (weekly seasonality) and minimize the objective function of robust location and scale:

$$\underset{\mu_h,\sigma_h}{\operatorname{argmin}} \sum_{i=1}^{7} \psi\left(\left[\frac{x_{t-168i}-\mu_h}{\sigma_h}\right]^2\right) \tag{7}$$

$$\psi(z) = \min\left(\max(z,-c),c\right) \tag{8}$$

where c is the threshold that limits the range of ψ, μ_h and σ_h stand for the robust estimation of location and scale. In our experiments we found that $c = 1.25$ provides the most accurate seasonal averaging.

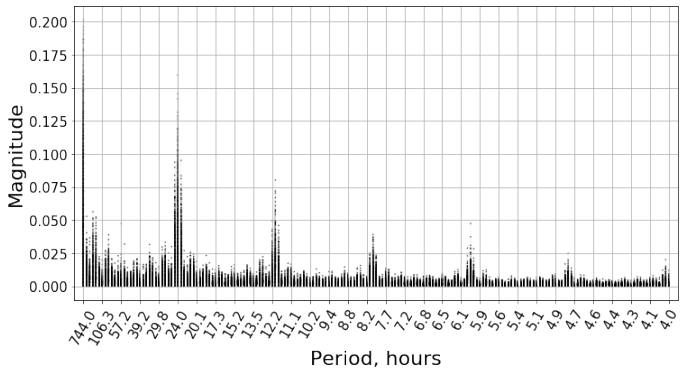

Fig. 3. Spectra of FFT with Blackman window for the viewership of 100 largest networks during the March of 2018 superimposed

After filtering of the original time series we extrapolate it with the use of the standard Fast Fourier Transform [4]. First, the magnitudes are calculated

$$M_k = \sum_{t=1}^{N} x_t e^{-\frac{2\pi i}{N} kt} \tag{9}$$

Then the extrapolation is evaluated with cosine for every t^{th} value

$$\hat{y}_{ft} = \sum_{k=1}^{N} \frac{\Re(M_k)}{N} \cos\left(2\pi\omega_k t + \arg(M_k)\right) \tag{10}$$

where ω_k is the k^{th} frequency corresponding to the magnitude M_k.

Experiments with various windows including Blackman, Hamming, and Parzen [5] as well as zero-padding and detrending did not result in any significant improvement in the overall accuracy of the extrapolation.

3.4 XGBoost

XGBoost [20] is a tree boosting method that incorporates regularization and a 2^{nd} order approximation of the objective function to prevent overfitting and reduce computation time. It allows for the use of an arbitrary objective function.

We combine the viewership data $x = \{x_1, \ldots, x_N\}$, $x_t \in \mathbb{R}$ with m features $X = \{X_1, \ldots, X_N\}, X_t \in \mathbb{R}^m$ that correspond to a specific network to obtain a data set $\mathcal{D} = \{(X_t, x_{t+\tau})\}_1^{N-\tau}$ in which some of the features X_t contain lagged viewership x_t with the maximal lag of τ, and the target is the viewership starting at time period τ. A tree ensemble model is composed of K additive functions

$$\hat{x}_{t+\tau} = \phi(X_t) = \sum_{k=1}^{K} f_k(X_t), f_k \in \mathcal{F} \tag{11}$$

where $\mathcal{F} = \{f(X) = w_{q(X)}\}$ stands for the space of regression trees with f_k representing independent tree structure q with leaf weights w. The functions f_k are found through the regularized minimization of

$$\mathcal{L}(\phi) = \sum_j l\left(\hat{x}_j, x_j\right) + \sum_k \Omega(f_k) \tag{12}$$

where $\Omega(f) = \gamma T + \frac{1}{2}\lambda||w||^2$, T is the number of leaves in a tree, and l denotes a convex differentiable loss function. The gradient boosting is performed through the iterative additive optimization. Denoting the model's prediction of the viewership on the training set at t^{th} time period at step i as $\hat{x}_t^{(i)}$ we minimize

$$\mathcal{L}^{(i)} = \sum_{t=\tau}^{N} l\left[x_t, \hat{x}_t^{(i-1)} + f_i(X_{t-\tau})\right] + \Omega(f_i) \tag{13}$$

Finally, the model's forecast is performed as

$$\hat{y}_x = \phi^*(Y) \tag{14}$$

where Y stands for the features of the testing set and ϕ^* is the trained tree ensemble. The features we used included:

- Lagged viewership at time periods that represent the same Weekday and Hour for different number of weeks in the past based on the seasonality of the series calculated with Fourier Transform;
- Seasonally Averaged Lagged Viewership: averages over the past k weeks for the same NDH (Network, Weekday, Hour). Since some NDHs are relatively volatile, looking at the seasonal average decreases the variance and gives a better prediction for the baseline trend. We also introduced the standard deviation as a feature which prevented the model from overfitting caused by outliers present throughout the 8-week period.
- Program Level Features. Since we were dealing with over 20000 programs, this categorical variable posed many challenges. To solve this problem we looked at the average number of impressions for the program and the network in the past 8 weeks, and used that as a numeric feature. We can think of this as an averaged lagged viewership feature for the programs. For certain future programs we did not have actual program names so we also used features like genre, and whether the program was live or repeated.
- Calendar Features: one-hot encoded important calendar days like Christmas.

Among the model's limitations we identified that while the model did pick certain special programs it did not perform well on extreme outliers (e.g. the Superbowl). It is rather expected of the tree-based models, since the prediction they make is an average of the prior predictions. More specialized models able to automatically detect and remove outliers might provide better accuracy.

Furthermore, the tree-based models are known to be unable to capture trends. As explored earlier, the viewership data seem to have a slowly decreasing trend but this did not pose an issue with the XGBoost model: detrending of the time series before training the XGBoost model did not offer significant improvements.

3.5 Ensemble model

Quite a few researchers have demonstrated that ensemble models generate forecasts more accurate than individual models participating in the ensemble. Wen Shen et al. [21] used an ensemble of 5 clustering techniques for electricity demand forecasting. Taylor et al. [22] combined various modifications of ARMA and GARCH models to improve the accuracy of wind power density forecasts. Kourentzes et al. [23] concluded that an ensemble of neural networks outperforms the best individual neural network model.

In our experiments we noticed that the errors produced by any two individual models have a very weak correlation (see table 1) which could be explained by the diversity of the models. Taking it into account, our ensemble model is in essence a convex combination of the forecasts made by individual models:

$$\hat{y}_e = w_b\hat{y}_b + w_x\hat{y}_x + w_p\hat{y}_p + w_f\hat{y}_f \tag{15}$$

where $w_b + w_x + w_p + w_f = 1$, $w_\cdot > 0$ are the weights assigned to the models.

4 Evaluation

TV viewership data used in real business applications are characterized with a slight processing delay which prohibits running next-day forecasting. Set-top box data become available two weeks after the actual viewership. To take that into account we separate training and testing data sets with a 2-week window.

In order to reduce bias in our models' parameters we performed yearly cross-validation training and testing. We selected 13 testing periods of 30 days from March 1, 2017 to March 1, 2018 each starting in the beginning of the month. The corresponding training sets were constructed from the viewership data collected in a period within 1 year to 2 weeks prior the beginning of each testing period.

TV networks naturally have different sizes of their audiences and total hourly viewership. To accurately measure average performance of our models on the networks of different size, we used Symmetric Mean Absolute Percentage Error (SMAPE). It is a common metric for relative forecasting accuracy evaluation [24][25], and for every model's forecast $\hat{y}_\cdot$ it is defined as

$$SMAPE(\hat{y}_\cdot) = \frac{1}{N}\sum_{t=1}^{N}\frac{2\,|\hat{y}_{\cdot t} - y_t|}{\hat{y}_{\cdot t} + y_t} \tag{16}$$

where y_t stands for actual values observed in the test set and $\hat{y}_{\cdot t}$ is the forecast made by either of the models: $\hat{y}_b$, $\hat{y}_x$, $\hat{y}_p$, $\hat{y}_f$, or $\hat{y}_e$. The metric fits naturally for

	Prophet	**XGBoost**	**Fourier**
Prophet	1.	0.18166874	0.51658387
XGBoost	0.18166874	1.	0.13299841
Fourier	0.51658387	0.13299841	1.

Table 1. Correlation matrix of forecast errors

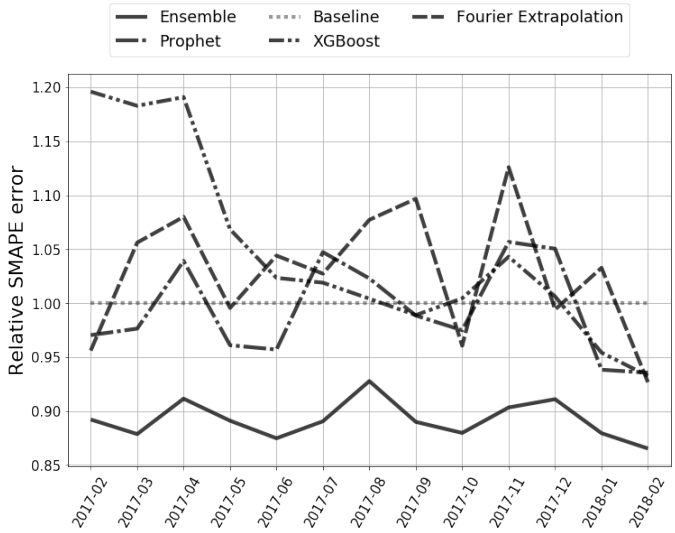

Fig. 4. Mean SMAPE errors normalized with the Baseline

our task because the viewership time series have values $x_t, y_t > 0$ and we restrict our model's forecasts to $\hat{y}_{.t} > 0$.

In table 2 we report the mean SMAPE errors for 100 top networks per testing period per model normalized with the errors made by the Baseline model. The errors are also visualized on Fig. 4.

The results of our experiments indicate that due to the presence of a very dominant seasonal signal that has a period of about 8 weeks, the simple Baseline model is on par with more complex predictors. XGBoost and Fourier Extrapolation produced slightly less accurate forecasts with the mean SMAPE errors of 5% and 3% higher correspondingly. While Prophet performed marginally better being 1% more accurate than the Baseline.

On the other hand, the Ensemble model demonstrated a significant improvement in overall accuracy being about 11% better than the Baseline which can be explained by a very weak correlation between the models' forecasts.

5 Conclusion

As discussed in the introduction, US TV viewing is undergoing a change, as more fragmentation occurs due to consumer choice. This places a focus on having a robust forecast that can handle smaller network feeds or streams with higher variation. We have undergone a study to pull together these methodologies, and have found that ensemble is a powerful way of reducing the overfitting of individual models.

While we have predicted ratings based solely on the viewing behavior exhibited prior to the broadcast, further research should focus on additional ex-

ternalities that may impact movement of viewers through content. Indeed, as found in [6], individual networks have differing trends, this combined with people watching fewer networks consistently means correlative effects may be observed. Another potential direction could include bias reduction in the models' forecasts.

6 Acknowledgment

This work was supported in part by Pitney Bowes 3100041700, Alfred P. Sloan Foundation G-2018-11069, and NSF award 1827505.

References

1. E. Oster: TVs Share of Ad Spend Expected to Continue Its Decline This Year, https://www.adweek.com/agencies/tvs-share-of-ad-spend-expected-to-continue-its-decline-this-year (2018)
2. J. Arvidsson: Forecasting on-demand video viewership ratings using neural networks (2014)
3. P. M. Napoli: The unpredictable audience: An exploratory analysis of forecasting error for new prime-time network television programs. Journal of Advertising, 30(2): pp 53-60 (2001)
4. J. W. Cooley and J. W. Tukey: An algorithm for the machine calculation of complex fourier series. Mathematics of computation, 19(90): pp 297-301 (1965)
5. A. Oppenheim, R. Schafer, and J. Buck: Discrete-time signal processing (1999)
6. K. Hubert: The Hidden Story Behind TVs Ratings Decline, https://www.simulmedia.com/assets/media/Hidden-Story-Behind-TV-Ratings-Decline.pdf (2017)
7. D. B. Hindman and K. Wiegand: The big threes prime time decline. A technological and social context. Journal of Broadcasting & Electronic Media, 52(1): pp 119-135 (2008)
8. S. D. Hunter III, R. Chinta, S. Smith, A. Shamim, and A. Bawazir. Moneyball for tv: A model for forecasting the audience of new dramatic television series. Studies in Media and Communication, 4(2): pp 13-22 (2016)

Period	Ensemble	Prophet	XGBoost	Fourier
Mar 2017	0.8922	0.9704	1.1961	0.9561
Apr 2017	0.8787	0.9764	1.1829	1.0561
May 2017	0.9115	1.0392	1.1910	1.0800
Jun 2017	0.8911	0.9610	1.0685	0.9955
Jul 2017	0.8748	0.9570	1.0235	1.0441
Aug 2017	0.8905	1.0473	1.0190	1.0274
Sep 2017	0.9278	1.0230	1.0042	1.0772
Oct 2017	0.8900	0.9886	0.9889	1.0968
Nov 2017	0.8798	0.9749	1.0045	0.9606
Dec 2017	0.9033	1.0567	1.0430	1.1260
Jan 2018	0.9110	1.0505	1.0062	0.9941
Feb 2018	0.8795	0.9383	0.9542	1.0328
Mar 2018	0.8653	0.9356	0.9326	0.9268
Mean, Deviation	0.89, 0.02	0.99, 0.04	1.05, 0.09	1.03, 0.06

Table 2. Mean SMAPE errors normalized with the Baseline

9. D. Zigmond, Y. Interian, S. Lanning, J. Hawkins, R. Mirisola, S. Rowe, Y. Volovich: When viewers control the schedule: Measuring the impact of digital video recording on TV viewership, Key Issues Forums at ARF Audience Measurement Conference (2009)
10. R. Weber: Methods to forecast television viewing patterns for target audiences. Communication Research in Europe and Abroad Challenges of the First Decade, Berlin, De-Gruyter (2002)
11. R. Neagu: Forecasting television viewership. A case study. GE Global Research, 2003GRC039 (2003)
12. R. Pagano, M. Quadrana, P. Cremonesi, S. Bittanti, S. Formwentin, and A. Mosconi: Prediction of tv ratings with dynamic models. ACM Workshop on Recommendation Systems for Television and Online Video, RecSysTV (2015)
13. D. Meyer and R. J. Hyndman: The accuracy of television network rating forecasts. The effects of data aggregation and alternative models, Model Assisted Statistics and Applications, 1(3): pp 147-155 (2005)
14. K. Nikolopoulos, P. Goodwin, A. Patelis, and V. Assimakopoulos: Forecasting with cue information. A comparison of multiple regression with alternative forecasting approaches, European Journal of Operational Research, 180(1): pp 354-368 (2007)
15. C. Wang, D. Goossens, and M. Vandebroek: The impact of the soccer schedule on TV viewership and stadium attendance: evidence from the Belgian Pro League, Journal of Sports Economics, 19 (1): pp 82-112 (2018)
16. M. Gambaro, V. Larcinese, R. Puglisi, J. M. Snyder, Jr.: Is Soft News a Turn-Off? Evidence from Italian TV News Viewership (2017)
17. R. Belo, P. Ferreira, M. de Matos, and F. Reis: The Impact of Time-Shift TV on TV Viewership and on Ad Consumption: Results from Both Natural and Randomized Experiments, A theory of the economics of time, The Economic Journal 81 (324): pp 828-846 (2016)
18. S. J. Taylor and B. Letham: Forecasting at scale, The American Statistician, 72(1): pp 37-45 (2018)
19. P. J. Huber: Robust statistics, International Encyclopedia of Statistical Science, pp 1248-1251, Springer (2011)
20. T. Chen and C. Guestrin: Xgboost. A scalable tree boosting system, In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp 785-794, ACM (2016)
21. W. Shen, V. Babushkin, Z. Aung, and W. L. Woon: An ensemble model for day-ahead electricity demand time series forecasting. In Proceedings of the fourth international conference on Future energy systems, pp 51-62, ACM (2013)
22. J. W. Taylor, P. E. McSharry, R. Buizza, et al: Wind power density forecasting using ensemble predictions and time series models, IEEE Transactions on Energy Conversion, 24(3): pp 775 (2009)
23. N. Kourentzes, D. K. Barrow, and S. F. Crone: Neural network ensemble operators for time series forecasting, Expert Systems with Applications, 41(9): pp 4235-4244 (2014)
24. S. Makridakis and M. Hibon: The m3-competition. Results, conclusions and implications, International journal of forecasting, 16(4): pp 451-476 (2000)
25. S. Makridakis, E. Spiliotis, and V. Assimakopoulos: The m4 competition. Results, findings, conclusion and way forward, International Journal of Forecasting, 34(4): pp 802-808 (2018)