Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2022) 30: 750 – 766

doi:10.55730/1300-0632.3809

Forecasting TV ratings of Turkish television series using a two-level machine learning framework

Busranur Akgul[1], **Tayfun Kucukyilmaz**[2,*]
[1]Applied Data Science Program, TED University, Ankara, Turkey
[2]Computer Engineering Department, Engineering Faculty, TED University, Ankara, Turkey

Received: 30.05.2021 • **Accepted/Published Online:** 24.10.2021 • **Final Version:** 21.03.2022

Abstract: TV rating is a numeric estimate of the popularity of television programs. Forecasting TV ratings is considered an important asset for investment planning of media due to its potential of reducing the risks of future ventures. The aim of this study is to develop a machine learning model capable of efficiently forecasting the TV ratings of Turkish TV series in a practical manner. To this end, two prediction models were proposed for forecasting the TV ratings of television series, facilitating an extensive set of features. A contribution of this study is the inclusion of social media-based features using search trends around television series and exploration of the viability of using these features in place of temporal features.

The study presents an extensive evaluation of the forecast performance of the proposed models. The performance of the proposed models were evaluated using a data collection composed of ratings and various attributes of series and their episodes aired at prime-time Turkish broadcast during 2014 and 2018. In the experiments, a theoretical forecast performance was first established with the inclusion of temporal features. Next, a set of practical models were proposed, replacing temporal attributes with social media-based attributes, relating to internet popularity and visibility of the series. The experiments show that, the proposed models achieve up to 1.65% error rate for the theoretical setting and 7.06% error rate for the practical setting.

Key words: TV rating forecast, TV rating prediction, episode rating forecast, regression, Google trends

1. Introduction

Television broadcasting started as a tape broadcasting service in 1968 and continued with color broadcasting in 1980s in Turkey. The first TV series and their continuations were solely produced by TRT, state channel of Turkey, from 1974 to 1990s. In 1990s, private channels started to appear in the screen and after 2000s, the number of private channels reached up to 400s which led to a competitive environment in the sector. After 2000s, many leading channels distinguish themselves by exporting series from abroad [1].

TV rating is a prime indicator of the popularity of a television program. Traditionally, TV rating is a numeric estimate of the average view percentage and duration of each television program at each household. Collection of TV rating is performed via a special equipment called people-meter, installed at a sample selection of households in different regions around a country, ideally representing different viewer demographics. TV rating measurement was originally intended as a system for giving advertisement decisions in order to match the content of an advertisement with a target audience, based on their socio-economic status [2]. Today, as

*Correspondence: kucukyilmaz@rsm.nl

televised media became the biggest shareholder of telecommunication investments, the importance of TV rating forecast became an indispensable component of the corporate decision making process.

The aim of this study is to evaluate the effectiveness of using machine learning techniques for forecasting TV rating of television series. To this end, we model the TV rating forecast problem as a regression problem, where each episode of a television series is considered an instance and each instance is represented via a set of features extracted from various sources. The features used in this study are categorized into four groups: time-, episode-, series-, and social media-based features. Time-based features represent the past behavior of the television series and are extracted by analyzing the popularity of the past episodes of the particular series. Episode-based features represent the state of a single episode of a particular series, and might change during the lifetime of a show. Series-based features relate to the genre and originality of the series.

The studies in the literature show that although social media data may have correlation with TV ratings, direct application of such data to the TV rating forecast problem requires high maintenance and produces poor results. The study presented here proposes a novel approach for using publicly available Web-based data for the TV forecast problem in an automated fashion. Instead of using text mining techniques or communication network graph analysis in order to infer social media attraction to a particular program, we use a more direct approach: we model the social media popularity of a series as the number of Web searches involving terms related with a series in order to represent how the audience react to and interact with a particular episode. We represent each TV show with a set of keywords that are related with the show and monitor their search patterns for a period of time. Using the fluctuations of the search patterns for these keywords, we aim to infer changes in the popularity of a TV show.

An important caveat of TV rating forecast is that, the forecast has more potential value if the predictions are made before the series has aired. Since using time-based attributes necessitate the availability of past episodes, models containing time-based features would only contribute to a theoretical prediction performance, which provide little practical value in terms of assesing the investment potential. Hence, we also propose alternative models that do not include time-based features in order to assess practical viability of such models. As user interaction with social media tends to start prior to the air time of a series, using social media-based features for estimating the possible popularity or audience reaction to an episode/series also provides a workaround for alleviating the need for past information, which is one of the major practical complications of the TV rating forecast problem.

We tested the proposed models for both theoretical and practical settings using a data collection gathered from seven popular channels in Turkey during the period between 2014 and 2018. Our experiments also showed that low rated and high rated episodes have distinct patterns. As a consequence, we propose a two-level machine learning strategy for more accurate forecast results. In the proposed strategy, first, a binary classification model is fit to each episode in order to estimate whether an episode would be rated high or low, and two separate regression models were trained for low and high rated episodes. Our experiments show that, the proposed technique slightly improve the forecast accuracy of the practical models.

The contributions of this study are as follows:

- We propose the use of Web search trends-based features to assess the popularity of the TV series.
- We examine and evaluate the TV rating forecast problem under two settings, regarding whether the forecast is performed pre- or post-air time of an episode.
- We present a multilevel forecast model, exploiting the difference between high- and low-rated TV series.

- We use an extensive set of features in order to represent the TV series and their episodes, and the propose models achieve improvements compared to similar forecast models in the literature.

The organization of this paper is as follows: In Section 2, previous work on TV rating forecast is examined. Section 3 describes the data collection and its aggregation process, and provides an in-depth look at the features that are used in this study and their taxonomy. Section 4 presents the execution pipeline of the forecast framework, detailing the feature extraction, preprocessing, and model generation processes. In section 5, the experimental framework, baseline and proposed techniques are presented along with the experimental results. We conclude with several final remarks in Section 6.

2. Related work

Since the ratings of a program determine the revenue of a TV show, accurate forecasting of the audience ratings will contribute to many benefits, such as investment planning and time savings for the producers and investors. According to [3], even if the needs for accurate forecasting of TV is clear, it is poorly studied and employed in practice. With the advent of new technologies, especially in data mining and machine learning, many studies apply the emerging techniques to the TV rating forecast problem. After 2000s, most studies focus on individual rating behavior while only a select few focus on social media such as Facebook [4] and Twitter [5, 6] in order to forecast the TV ratings of a program or a channel. An objective of this study is to integrate data collected via the social media to the TV rating forecast problem in a successful manner.

Studies on TV rating forecast can be categorized into two groups based on the type of the data that is used to evaluate TV ratings: individual [7–11] and aggregated data of the all households' people-meter [3, 4, 12–16] values. Meyer & Hyndman's work [7] is unique in the sense that it uses personalized attributes of the users, such as exact viewing times. Even though the individual data provides additional information about behavior of the audiences, such data is generally kept private due to proprietary reasons. Another alternative categorization of the past studies is based on the objective of the forecast. The TV rating forecast can focus on the ratings of a channel [3, 9, 12–14, 16], a program [4, 8, 11, 15], or an individual episode of a program [7, 9].

In the literature, several techniques are employed to the TV rating forecast problem. Classification and regression analysis are among the most popular techniques. Classification algorithms such as neural networks [16], decision trees [7, 16], gradient boosting machine [14], and nested logit [3] are employed in several studies. Ridge regression, step wise regression, linear regression, Bayesian averaging, and genetic algorithms [3, 7, 8, 11–13, 15] are examples of studies using regression for forecasting the TV rating.The study in [9] is one of the few studies that used time series analysis for the TV rating forecast problem. Although various techniques are employed to the problem, the representative features used in past studies mostly remain the same. These features are often extracted from static content such as descriptions of the TV shows or broadcast stream, including day of the week, year, duration, genre, and rerun information for a particular show or episode.

The studies about forecasting the TV ratings based on social media naturally consist of predictors related to specific programs or TV series such as number of posts, likes, shares, and comments for the specified pages. According to [5], forecasting TV ratings based on tweets is not trivial and favorable due to the number of tweets differing between the episodes of a program even if the correlation between the tweets and TV ratings are sometimes very high. [6] also showed that social media ratings based on Twitter is not in direct correlation with actual TV ratings in Turkey due to the fact that users can interact with Twitter feeds before and after the air of a TV show rather than only during the air time. On the other hand, the authors of the study in [4] find out that using Facebook posts has favorable results for forecasting the TV ratings of a program. The

authors used 10 features which are obtained from both social media posts and TV companies. The features in the study include the number of posts, likes, comments and shares of each post in various program fan pages, and the counts of these from the fan page administrators. Even though promising results are reported, the study involve only four popular TV programs. Several other studies also show that the most popular programs show correlation between social media and TV ratings [5, 6, 17].

The authors of the study in [18] contributed to an empirical study for determining the success factors of a never-aired television series. They conducted experiments based on 107 dramatic TV series which are broadcast on the 4 most popular US TV channels. They have used several interesting features, such as whether a TV series being an adaptation and spin-offs of a past series or the past performance of the crew of a series, and the size of the social network of the first episode of a new series; most of which are positively associated with the rating prediction performance. Using generalized least square regression to test their hypotheses, the authors showed that forecasting TV rating of a television series can be achieved with reasonable success.

Multimodal approaches to the forecast problem, are poorly studied in the literature.The study in [13] developed a model forecasting period of 6 hours of channel ratings from an aggregated data using forward step wise regression and presented their results using mean square error as the performance measure. The model facilitates features such as teleplay playback number, type, audience ratings, number of star actors, number of teleplay fans, and TV channel information. Authors also presented a detailed analysis of the relationship between these attributes and the TV rating. The study in [15] proposed using a new set of features for the TV rating forecast problem including type of television, broadcasting time, program name, TV channel, and three main actors of a program using ridge regression. In regard to the range of features, the above study has similarities with our work: we also used an extensive set of features collected and integrated from multiple sources.

The authors in [14] presented an extensive study on TV rating forecast of series using a large set of features and employing various machine learning techniques. Technique-wise, the study provides an invaluable direction for the work presented here. In their study, the authors used TV program characteristics, program performance indicators, promotional support information, and social behavior indicators such as social media ratings as features. They have used several machine learning techniques such as penalized and linear regression, multiple adaptive regression splines, support vector machines, gradient boosting machine, random decision forests, and neural networks. Their results showed that gradient boosting machine models performed best in terms of accuracy and scalability. Recently, the study in [11] also presented a feature-based forecast analysis on the people-meter data of randomly selected 200 households by applying entrophy and regression techniques. In the study, the most popular 6 TV series that broadcast in 2014 were selected and the first 10 episodes of each series were used. There are several differences between the above studies and this work. First, instead of using social media ratings directly, our work employs a more indirect and practical approach for inferring the social media popularity by using a keyword based trend analysis on popular search engines. Second, our study proposes a two-level machine learning algorithm in order to further improve the forecast quality. Third, this study involves a much larger dataset, providing a more factual representation of the TV rating forecast problem. Last, the proposed models used in this study are more generic, and the features are gathered via publicly available channels, allowing an ease of applicability.

3. Dataset and features

A large number of features are extracted from multiple sources in order to describe the state of each episode of the TV series, and are used in our machine learning models. In this section, we first briefly describe the composition of the dataset, then describe the features in detail.

3.1. Dataset

In order to forecast TV ratings, data were collected from three different sources. The data collection contains information of the individual episodes of 100 highest rated TV programs which are aired on Turkish prime-time television between 2014 and 2018. In the collection, information on each episode of a TV series is considered an instance, and there are 5416 such episodes. The response variable, rating data, that has been used in this study entails publicly available, official estimates of audience viewing rates announced by the Turkish Television Audience Measurement Committee (TIAK Inc.). The rating value is considered as a continuous variable ranging from 0.310 to 21.240 (1^{st} Quartile = 3.070, Median = 4.390, Mean = 5.023, 3^{rd} Quartile = 6.250).

Figure 1 presents the rating distribution in the TV ratings data collection used in this study after preprocessing and cleanup. In order to provide a clear view, we have packed the episodes of the TV series into equal-sized bins that represent rating categories. The figure illustrates that, although popular and high rated episodes are relatively few in number, the dataset represents both high- and low-rated episodes successfully.

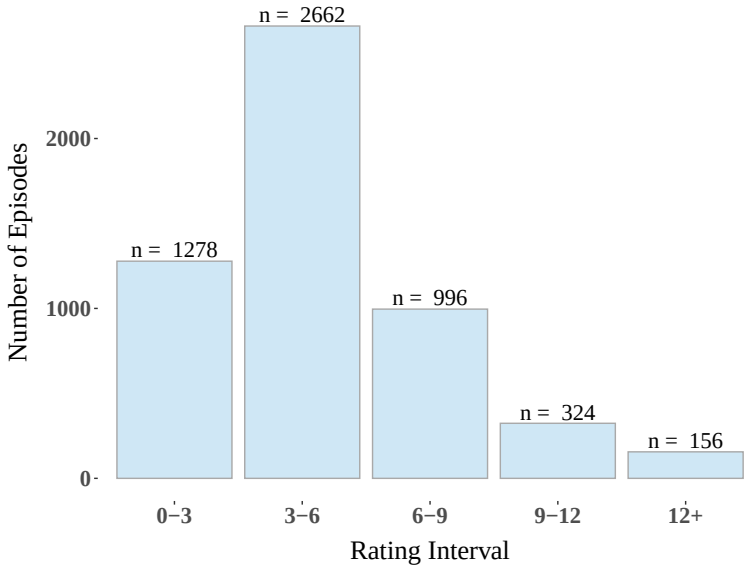

Figure 1. Rating distribution in the TV rating corpus.

Figure 2 summarizes the modular composition of the dataset. In the figure, the accumulated and restructured data is named as the *TV Rating Corpus*. The TV rating corpus contains rating information and features extracted from three separate sources. These sources are labeled as TV ratings, search trends, and series features in the figure.

An important component of the TV rating corpus is the TV ratings. For each of the episodes of a TV series, TV rating of the episode is used as the response variable. Since it is not possible to measure the true television viewing rate, the published TV rating values are used, which are estimated by Turkish

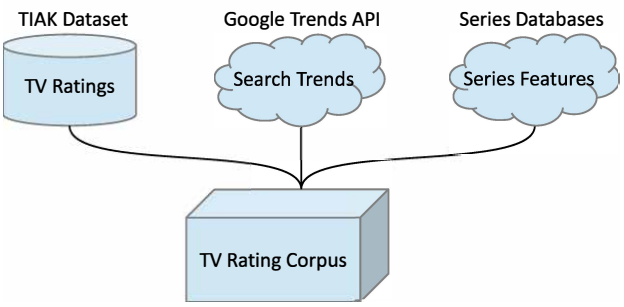

Figure 2. Composition of the TV rating corpus.

Television Audience Measurement Committee (TİAK) [1], by sampling over large populations from different socio-economic backgrounds. According to TİAK, which is a joint-stock company established to organize and supervise television surveillance in Turkey, the sample selection consists of two steps: layering and selection. At layering step, the rating values are collected from sample households while ensuring that the chosen sample is distributed proportionally to represent all socio-economic groups in the country. At the second step, samples are further processed in the collection, respecting the proportions in population sizes in different areas. The sample data and design is maintained in a database, but is also open to change, and is updated every year. During the time of the study, this data was made publicly available by TİAK. The rating component of the TV rating corpus is signified as the "TIAK Dataset" in Figure 2.

Using search trends for representing the attention of the attraction of a TV series is a novel contribution of this study. The motivation behind this decision is that, today, there is no product in the global market that would be unaffected from the online platforms in some way. In order to reflect the impact of online platforms while predicting the ratings of TV series, we have formulated the impact of a TV series on social media as the amount of search traffic it has generated, reflected via search patterns around terms that are related with a series or its episodes. Google Trends API provides an anonymized, categorized, and aggregated time series information on a particular search keyword either globally or in a specific region in a normalized fashion. Normalization is applied in order to prevent regions with the highest search volume from dominating relative search activities. Having the same search trend in more than one region for a given term does not indicate that the search volumes of these regions are the same [2]. In this study, we have collected search trends for both series and players in the series using gtrendsR package [3].

Series features summarize broad information about a TV series and of each episode. In this study, features on TV series and their episodes are gathered from a popular TV series database website [4] extracted at April 18, 2018. The Web pages contain a variety of information on the programs including the short summary of the series, its plot story, genre and categorical information, and its start and end times. For each series in this study, the related Web pages are crawled, scraped, and processed. Gender of the protagonist, information of the originality, and the players of a TV series are other qualifier attributes of TV series that are also used in this study. These additional features are gathered by crawling and parsing two additional internet TV databases [5].

[1] http://tiak.com.tr/tablolar
[2] https://support.google.com/trends/answer/4365533?hl=en
[3] https://cran.r-project.org/web/packages/gtrendsR
[4] https://www.tv8.com.tr/reyting-sonuclari
[5] https://www.diziler.com and https://www.dizisi.info.tr

To improve data quality, preprocessing, data transformation, and discretization techniques are applied [19] to the TV rating corpus. The crawled text is converted into lowercase, spaces between names are transformed to dash (_) characters. Genre and originality features are stored as numeric features, while gender of the protagonist is encoded as a binary one-hot variable. Finally, using temporal information on each TV series, several windowed, time-series based features are extracted.

3.2. Features

In order to provide a clear and organized presentation, the features used in this study are organized into four categories. These categories are: Time-based (TB), Google trends (GT), Series-based (SB), and Episode-based (EB) features. Table 1 provides a view into this categorization, number of features related to each category, representative names, and short descriptions of the features designed and used in this study.

Table 1. Features used in the forecast models.

Type	Feature category	Count	Description
TIME-BASED (TB)	MIN_RATING_ATTR	3	Minimum TV rating of the last 2, 4, and 8 episodes
	MAX_RATING_ATTR	3	Maximum TV rating of the last 2, 4, and 8 episodes
	AVG_RATING_ATTR	3	Average TV rating of the last 2, 4, and 8 episodes
GOOGLE TRENDS (GT)	SEARCH_TREND_RATE	2	Search rate in the last 7 days before air time
SERIES-BASED (SB)	GENRE	12	Genre of the series
	PROTAGONIST GENDER	1	The gender of protagonist
	ORIGINALITY	1	Series being a spin-off or not
EPISODE-BASED (EB)	CHANNEL	1	The TV channel that aired the episode
	YEAR	1	The year that the episode was aired
	DURATION	1	The length of the episode
	DAY	1	Day of the week the episode was aired

Time-based features: These features represent the past TV rating behaviour of a series. In order to simulate a time-series approach via machine learning techniques, we use an online time-windows approach for extracting the time-based features. They are obtained via aggregating the running minimum, maximum, and average TV rating values for the last 2, 4, and 8 episodes of the series. Here, it is important to point out that time-based features require information on past air time, hence an initial allocation of air time for the series. Thus, using these features in a forecast model would have less "practical benefit" as they cannot be used for forecasting potential ratings before air time allocation.

Google trends features: In order to represent the popularity of a TV series without depending on past episode ratings, we extracted several features reflecting the Google search trends related with the TV series. The Google trends data entail worldwide and region-wide search frequency for keywords on the series in question. Several keywords are extracted from the official Twitter channel of each of the series in the study, common hashtags are extracted, and search patterns of the keywords contained in these hashtags are used via Google Trends API in order to infer the temporal patterns in the search history about the series. Google trends features are extracted using the Google APIs, and consist of average search volume (i.e. number of queries submitted from Turkey region) of queries involving the name of the TV series and its protagonist for 7 days, starting from 7 days before the air time until the date of the air time of an episode.

Note that, using past trend information is practical for predicting the rating of a never-aired TV program. This is motivated by the fact that, although the Web search trends data are also collected using past information similar to time-based features, it is a commonly observed behavior that users often tend to start Web search activity on popular media content even before the air time of a show. Consequently, here we postulate that it is possible to collect, observe, and evaluate the media hype generated by a TV program before air time via analyzing search trends.

Series-based features: These features contain qualifier attributes of TV series such as genre, gender of the protagonist, and originality. The series genre is grouped into 12 categories. We would also like to note that the series can have more than one genre, hence each of the 12 different genres is represented as a separate binary feature in the TV rating corpus. Gender of the protagonist is also binary attribute representing male or female. Inspired by the study of [18], we also included the originality feature, which is used to indicate whether the series is a spin-off from another work such as an adaptation of a novel or a movie.

Episode-Based Features: These features represent episode-level aspects of a series that are expected to change in a weekly basis. To this end, channel, year, duration, and airing day of an episode are considered as features. Within the TV rating corpus, there were seven distinct channels and five distinct year values (2014-2018). Duration is the length of an episode, in minutes, and day is the weekly air date of the episode, which can change during the lifetime of a series.

4. Framework

The procedure used for forecasting TV ratings is presented in Figure 3. The figure summarizes the federation of data from multiple resources, the preprocessing and feature extraction operations that have been applied to the raw data, and the machine learning framework that has been used. The proposed framework consists of three modules that work in a pipelined fashion: data acquisition, preprocessing, and classification/regression modules.

The aim of the data acquisition module is to collect information on prime-time Turkish TV series in an automated fashion and federates separate data sources in order to generate the TV ratings corpus. The detailed descriptions of these sources, and the collected data can be found in Section 3. However, due to lack of information or short air time of several series, some instances within the TV Ratings corpus are not suitable for deducing sound results. In the preprocessing module, series that are aired less than 10 episodes are eliminated from the corpus. Additionally, the information on the series is converted into a feature representation in the preprocessing module: Rating and Google trends data is represented as numeric features while, series features such as the information on genre and protagonist are scraped from the associated Web pages and are represented as one-hot binary features. Last, several numeric, time series-based features are generated in the preprocessing module, and normalization is performed for features that are represented as continuous variables. During preprocessing, no feature selection and/or elimination was performed.

The classification/regression module divides the dataset into two partitions: a training set and a testing set, generate forecast models using the training data, and estimate the forecast performance on the testing data. For the experiments, 10-fold shuffled cross validation that is based on episodes of the series is used for generating the testing and training sets.

For modelling the TV rating forecast problem, we propose two approaches. First, adhering to the literature, we approach the problem as a regression problem with a vector space model where, each episode of a series constitutes to an instance, and the aggregated set of features are used to describe the instances. We then applied

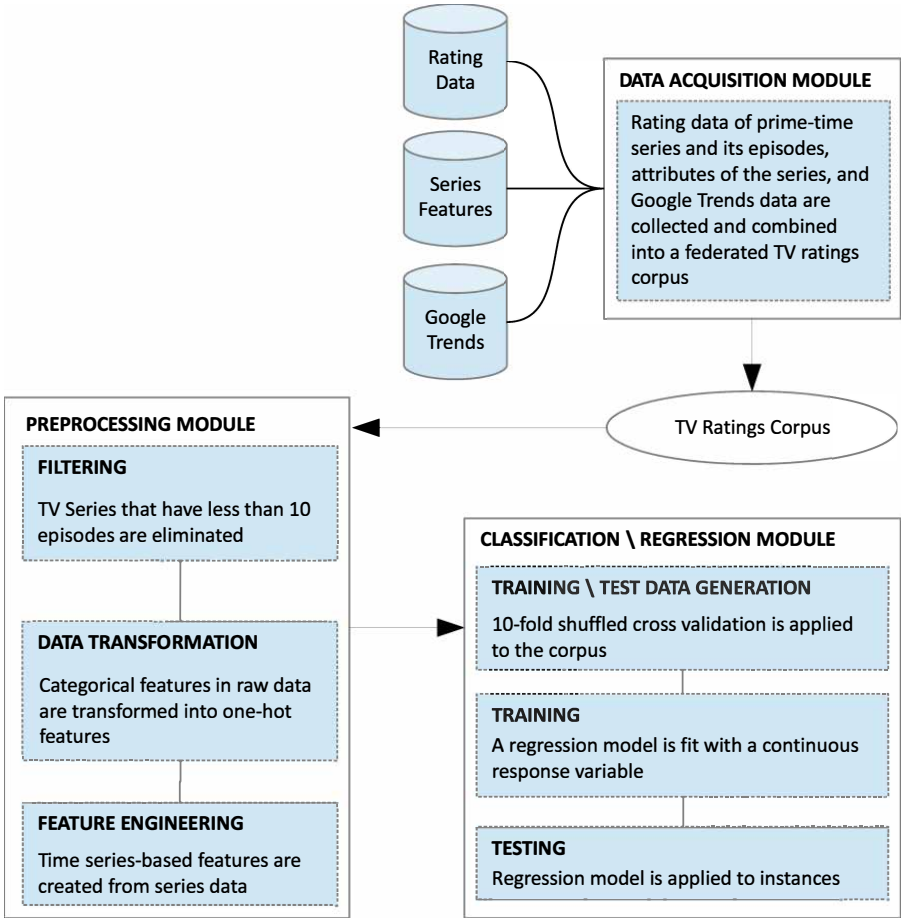

Figure 3. TV rating forecast framework.

a wide range of machine learning algorithms such as multivariate adaptive regression splines (BAGEARTH [20]), forward-backward greedy algorithm (FOBA [21]), Bayesian models (BAYESGLM [22]), rule-based regression (CUBIST [23]), generalized linear models (GLM [24], GLMNB [25]), decision tree regression (M5P [23]), random forests (RF [26]), and support vector machines (SVM [27]) for generating the forecast models. These models are then evaluated over the testing data and the results are reported in Section 5. Table 2 summarizes the parameters of machine learning algorithms used in the experiments.

4.1. TV Rating Prediction Model

We propose two machine learning models for the TV rating forecast problem. The proposed models are unique in the sense that they both use search activity trends in order to represent the state of each episode of a series; an approach applied to the TV rating forecast problem for the first time in the literature. In addition, in an attempt to improve predictive performance, the proposed models are also evaluated while facilitating time-series based features, aggregated in an online fashion with a time-window based mechanic. Here, we describe the two proposed models in detail.

Social media-based regression model [(TB)+EB+SB+GT]: In this model, each instance is represented with all available features presented in Table 1. In this sense, the social media-based regression model

Table 2. Selected parameters for the machine learning algorithms.

Algorithm	Parameter Name	Value
BAGEARTH	number of bootstrap samples (B)	50
BAYESGLM	error Distribution (Family)	Gaussian
CUBIST	# neighbors	1
	# committees	75
FOBA	steps	29
	n.u.	0.5
	λ	$1e^{-5}$
GLM	solver	Coordinate Descent
	θ	$1e^{-10}$
	β	$1e^{-4}$
	α	0.5
GLMNB	link	log
	θ	$1e^{-10}$
RF	number of trees	20
	number of indicators	100
SVM	kernel	Linear
	γ	$2e^{-3}$
	regularization parameter (C)	1

has the most ŕepresentative power out of all models. For a clear presentation, the social media-based regression model is tagged with the name "(TB)+EB+SB+GT" in the experiments notifying the range of features making up the composition of the descriptive attributes in the model. Note that the term "TB" is provided in parenthesis within the name tag, as time-based features will only be used in a theoretical setting and will be included in the name tag, while they will not be used in the experiments involving a practical setting, and will not be included in the name tag.

Two-level forecast model [two-level (TB)+EB+SB+GT]: During the experiments, we observe a consistent behavior in the generated forecast models: the high rated series tend to get overevaluated and low rated series get underevaluated by all regression models. Hence, as a refinement approach, we propose a two-level forecasting framework for the problem. First, a single label binary classification model is trained for each instance, predicting whether the instance is a high- or low-rated episode. In this classification task, the average of the training instances are used as a threshold for differentiating between high and low ratings. Both training and testing sets are divided into two using the outcome of this model. Next, two separate forecast models are generated by using the associated training data where, similar to the classical approach, the forecast problem is modeled as a regression problem. The two generated models, i.e. binary classification model and regression model, are then evaluated over the testing data.

Note that, the former binary classification problem can be considered as a coarser, subproblem of the TV rating regression problem. The motivation behind the two-level forecast model is to approach first the problem at a coarser level with the hope of achieving a higher accuracy than the relatively more challenging, finer problem; next, fit better, more suitable models to each predicted class, at a finer level as a refinement strategy. Here, it is important to note that, this study is not comparing the performance of the binary classification model and the regression model; but propose a two-level strategy that facilitates both for a single forecasting

objective. In this sense, the two-level forecast model uses the same training data for the both coarse and finer version of the problems. But this approach can be justified since the objective of the proposed model is also achieving a rating forecast using the available information base.

For a clear presentation, the two-level model is tagged with the name " two-level (TB)+EB+SB+GT" in the experiments notifying the range of features making up the composition of the descriptive attributes in the model. The term "TB" is handled similar to the social media-based regression model, being used within the tag according to its availability in the experimental setting or not.

5. Experimental results

In this section, we present a comprehensive experimental evaluation of the proposed framework. To this end, we adopted/replicated several forecasting techniques that have been used in the literature. In the experiments, we also consider two separate settings considering the potential value and usefulness of the outcomes: when information on previous episodes is available in the form of time-based features and when information on previous episodes is not available. We ran the experiments on a workstation consisting of two i7 6700HQ 2.6 GHz processors, having 8 cores. The experimental framework has 16 GB of RAM installed.

Note that, forecasting the TV ratings of a series while it is already in the air, and before it starts getting air time are two different problems with different associated practical value. In the former case, it is already possible to infer rough estimates of the TV rating by evaluating the past episode ratings. In this case, predicting the TV rating has little outcome revenue-wise as the series is already allocated air time. We term this setting the "'Theoretical Setting", reflecting its superiority in terms of its high predictive potential and relatively low practical potential. In this setting, past episode information is available for the forecast model. The experimental results on the theoretical setting is presented in Section 5.2. In the latter case, the forecast models can be applied to series even before the allocation of air time; hence can be used to evaluate the monetary utility and benefits of airing a series prior to airtime. In this scenario, it is assumed that prior episode information is unavailable to the generated models. In order to reflect its high potential for generating revenue-wise benefit, we term this setting the "Practical Setting". In this sense, the former setting would provide an upper bound for the possible prediction performance, while the latter setting would serve a more practical outcome. The experimental results on this, more practical setting is presented in Section 5.3. In order to compare our findings, we also adopted two baseline forecasting techniques from the literature. While these techniques are evaluated under the same classification/regression models, they differ in terms of features used to represent an instance. For both settings, the proposed forecast models are compared with the baseline techniques.

5.1. Baseline techniques

As summarized in Section 2, other than channel rating prediction, the TV rating forecast problem focuses on two distinct scenarios in the literature: forecasting the TV rating of a program, or an episode of a program. The widely adopted strategy for both these problems is to approach the problem with a vector space model where, each series or episode is represented with a set of descriptive features. As the response variable for these problems is different, so does the features to represent them. In order to represent the results of these commonly adopted strategies, we used the feature categorizations presented in Table 1 and design two baseline techniques.

Series based model [SB]: In this model, each series is represented with a set of descriptive features reflecting the state of each series. In order to replicate and evaluate the performance of series-based models used in the literature [17, 18], we have used series-based features (SB) as the descriptive features in the model.

This baseline is tagged with the name "SB", reflecting the set of features used in the model, and will be referred with this tag in the upcoming experiments.

Episode-based model [SB+EB]: In this approach, each episode of a series is considered as a separate instance and represented accordingly. In the episode-based model, each episode is represented with a set of descriptive features reflecting the state of each episode. In order to replicate and evaluate the performance of episode-based models [3, 9, 11, 12, 15], we have used both episode-based (EB) and series-based (SB) features as the descriptive features in the model. In order to improve clarity of the presentation, this baseline is tagged with the name "SB+EB", reflecting the set of features used in the model, and will be referred with this tag in the upcoming experiments.

We would also like to point out that, since the above mentioned past studies are evaluated on a variety of datasets, and using a wide variety of features, a direct comparison of the result accuracy may be misleading. Hence, we have adopted the aforementioned baselines by replicating the adopted feature representations in these studies, using the features that are available in the presented Turkish prime time TV rating corpus.

Table 3 presents the result performance of the baseline models. In the table, each technique is represented as columns. For each technique, execution time, root mean square error (RMSE), error relative to the mean (RMSE%), and accompanying p-value is presented. The rows of the table summarize the results of different machine learning algorithms applied to the forecast problem for the two baseline techniques. The results of the experiments reveal that, both series- and episode-based models can successfully forecast TV rating of an episode of a series. The RMSE values indicate that, while series attributes contain information that can be used to infer the TV rating, episode-based attributes can substantially improve the result performance as the episode-based models perform consistently better. The p-values in the table also indicate that all models perform within an acceptable level of significance, indicating that TV rating forecast is a practically viable problem, and application of machine learning models would present fruitful results.

Table 3. Performance of baseline forecast techniques.

	SB				SB + EB			
Model	Time (s)	RMSE	RMSE %	p-value	Time (s)	RMSE	RMSE %	p-value
BAGEARTH	364.68	3.10	14.58	3e-72	1246.41	2.58	12.15	9e-73
BAYESGLM	3.30	3.15	14.85	7e-71	4.24	2.60	12.26	9e-72
CUBIST	**171.34**	**1.79**	**8.45**	**7e-79**	**265.98**	**1.63**	**7.66**	**6e-81**
FOBA	5.66	3.15	14.85	7e-71	14.88	2.60	12.26	9e-72
GLM	2.21	3.15	14.85	7e-71	14.88	2.60	12.26	9e-72
GLMNB	51.74	3.15	14.85	1e-70	72.77	250	11.79	3e-70
M5P	1.68	2.54	11.98	1e-72	1.43	1.79	8.43	6e-79
RF	1966.50	1.85	8.71	5e-78	3653.84	**1.57**	**7.40**	1e-80
SVM	85.36	3.10	14.58	2e-82	165.19	2.52	11.88	1e-81

5.2. Experiments on theoretical setting

The purpose of this experiment is to evaluate the result performance of the proposed techniques. In this setting, the TV rating values of the past episodes are used as a part of the descriptive features. Hence, the model is capable of forecasting TV ratings for a series that is already allocated air time; however, the model cannot be used for evaluating the potential success of a TV series before air time and hence has limited practical use.

Table 4 presents the result of the forecast performance of both social media-based regression (TB + EB + SB + GT) and two-level forecast models (two-level TB + EB + SB + GT).

Table 4. Proposed TV rating forecast framework performance on the theoretical setting.

	TB + EB + SB + GT				two-level TB + EB + SB + GT			
Model	Time (s)	RMSE	RMSE %	p-value	Time (s)	RMSE	RMSE %	p-value
BAGEARTH	**146.47**	**0.35**	**1.65**	**1e-99**	**226.76**	**0.35**	**1.65**	**2e-96**
BAYESGLM	**3.83**	**0.35**	**1.65**	**2e-99**	**6.35**	**0.35**	**1.65**	**7e-96**
CUBIST	273.36	0.37	1.74	4e-101	390.06	0.37	1.74	1e-97
FOBA	**12.44**	**0.35**	**1.65**	**1e-99**	**27.95**	**0.35**	**1.65**	**7e-96**
GLM	**2.23**	**0.35**	**1.65**	**2e-99**	**3.48**	**0.35**	**1.65**	**7e-96**
GLMNB	**36.39**	**0.35**	**1.65**	**1e-99**	**48.77**	**0.35**	**1.65**	**6e-96**
M5P	**0.94**	**0.35**	**1.65**	**2e-99**	**1.74**	**0.35**	**1.65**	**8e-96**
RF	3139.66	0.38	1.79	4e-100	6758.29	0.38	1.79	1e-96
SVM	169.09	0.36	1.70	2e-100	255.63	0.36	1.70	6e-96

The rows of Table 4 indicate the machine learning algorithms used in the study. The columns of the table are divided into two partitions, representing the proposed social media-based and two-level regression models respectively. For each model, the computation time, RMSE, RMSE %, and p-value of the significance test (i.e. two-tailed t-test) is presented. Bold entries in the table notifies the best performing results, in terms of either prediction performance or computational time. The results show that, both models perform consistently better than the baseline models presented in Table 3. Compared to the baseline models, the RMSE is reduced to 0.35, which constitutes to a 6.8% improvement. For both proposed models, the significance tests show similar results, notifying the significance on the results of all models.

Comparison of the two proposed models reveal that, in terms of result quality, there is no significant difference between the two models. Additionally, the two-level model consistently requires larger computation times in all algorithms, as the model generates a classification and a regression model. The inability of the two-level regression model to improve the result performance can be attributed to one of the two possible causes. Either, generating a two-level model does not improve the proposed model, or both algorithms perform close to an optimal solution within the possible room for improvement.

5.3. Experiments on practical setting

Although experiments under the theoretical setting present successful results, the practical utility of the generated models are limited; as they cannot be used prior to air time. In an attempt to evaluate the performance with the absence of temporal data, we design a second experimental setting. Note that compared to the baseline techniques, the models in this experiment also utilize search trends data relating to the popularity of a series. Table 5 summarizes the results of the experiments.

The table is organized similar to Table 4: the rows of the table are divided into two partitions, representing the experimental results of social media-based and two-level regression models, respectively. The results show that, compared to the results of the theoretical setting, in the absence of time-series based features, the prediction performance is consistently lower for both regression models. However, the results also indicate that, TV rating of a series can be predicted quite successfully by the social media-based regression model, with error rates ranging

Table 5. Proposed TV rating forecast framework performance on the practical setting.

	EB + SB + GT				two-level EB + SB + GT			
Model	Time (s)	RMSE	RMSE %	p-value	Time (s)	RMSE	RMSE %	p-value
BAGEARTH	1456.82	2.65	12.49	5e-66	802.00	1.97	9.27	1e-82
BAYESGLM	4.79	2.62	12.35	2e-67	4.93	2.02	9.51	3e-81
CUBIST	**277.80**	**1.57**	**7.40**	**8e-77**	**292.58**	**1.50**	**7.06**	**1e-80**
FOBA	13.72	2.62	12.35	2e-67	15.08	2.02	9.51	3e-81
GLM	2.64	2.62	12.35	2e-67	3.42	2.02	9.51	3e-81
GLMNB	62.68	2.51	11.83	1e-65	45.65	1.92	9.04	3e-80
M5P	1.82	1.78	8.39	1e-75	1.34	1.79	8.43	2e-81
RF	**3425.94**	**1.60**	**7.54**	**2e-75**	**4583.28**	**1.56**	**7.34**	**7e-81**
SVM	912.30	2.51	11.83	6e-77	212.02	2.07	9.75	4e-85

between 7.40% and 12.49%. The experiments conducted using two-level regression model further improve the forecast performance of the proposed models up to 7.06% RMSE, which indicates a 0.34% improvement over the model that does not contain a binary classification state (i.e., social media-based regression model). As a consequence of using an increased number of attributes and training multiple machine learning models, the results also indicate a relative increase in computation costs as indicated by the computational time.

Figure 4 presents the receiver operator characteristic curve (ROC) for the binary classification classes of low- and high-rated episodes. The area under the curve (AUC) for both classes are 0.96 and 0.95, respectively indicating a high accuracy of the generated models for prediction of both classes. Although the experiments on the two-level prediction model show that, the proposed model improves the prediction accuracy of the class representing the low-rated episodes relatively more than the class representing the high-rated episodes, the difference, as also indicated by the ROC curve, is not significant. Also note that, as ROC curve does not depend on the distribution of data, the figure also indicates that the generated models are not effected by the distributional imbalances of the data and each class is represented sufficiently in the models.

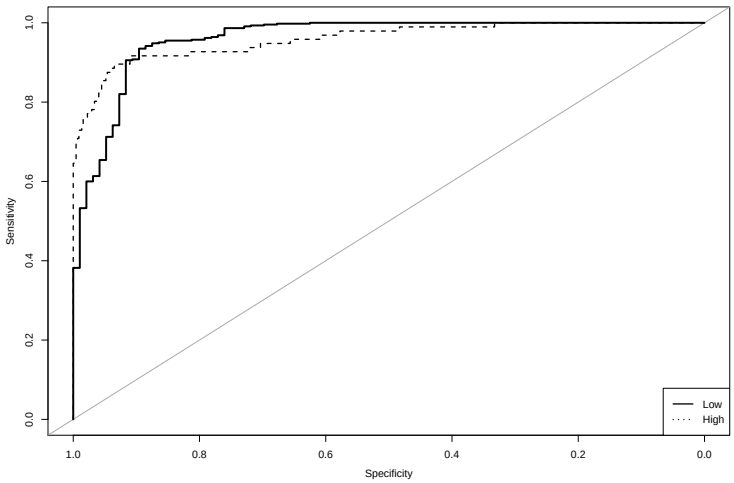

Figure 4. ROC curve for binary classification in two-level forecast models.

An important conclusion of the practical experiments is that, the two-level regression model consistently improved the forecast performance in this setting. Combined with the results achieved using the theoretical setting, the results show that, using time-series based attributes along with other descriptive features in the theoretical setting indeed produce results that is close to the possible room for improvement in the forecast problem, and that the inability of the two-level regression model for improving the result performance in the theoretical setting is possibly due to achieving close-to-optimal prediction performance in the theoretical setting.

The results indicate that, it is possible to forecast TV ratings of a TV series with an acceptable margin of error using an ensemble of features. The results also show that, user Web search trends can be integrated into the forecast problem, improving the forecast performance successfully. We would also like to point out the fact that, user Web search trends data are completely anonymous, publicly available, and easily accessible through public APIs, allowing similar research to be conducted for other human-centered forecast problems.

6. Conclusion

In this study, two social media-based forecast models are proposed for predicting the ratings for prime-time Turkish TV series, using an extensive set of descriptive attributes involving time, episode, series, and search trend-based features. The generated models were then evaluated under two settings. First, theoretical forecast performance is established in a setting where, the models facilitate past information via a set of temporal features. Second, practical forecast performance is evaluated in a setting where, the models do not contain temporal features. In the latter models, search trends acted as a proxy to assess potential popularity of a series. During the experiments, distinct behavior of high and low rated series is identified and a multilevel model is proposed. In the model, first a binary classifier is trained for predicting whether a series has high or low rating, then for each of these classes two separate regression models are fit, as a finer refinement strategy. Prediction performance of the proposed models is also evaluated and 1.65% and 7.06% error rates are achieved for theoretical and practical settings, respectively.

An important contribution of this study was integrating search trends to the TV rating forecast problem. In this work, we hypothesize that the search trends around a TV series would provide valuable information regarding its popularity and potential success. As search activity around a series tend to start a long time before the airing of a series, collecting/using search trends in order to describe the series before air time can be a viable and practical strategy for predicting TV ratings for never-aired TV shows. Our experiments also supported this claim and showed that, in terms of performance, search trends analysis could indeed replace temporal data in the forecast models successfully. We would also like to point out the fact that, in terms of both computational complexity and practicality, using search trends provides an easy to use, publicly available proxy for estimating audience reaction compared to other forms of social media data. Another contribution of this study is using a binary classification model for categorization of series into two classes: high- and low-rated series, and generate class-specific regression models for the TV rating forecast. Our results show that, the proposed model slightly improve the forecast performance with a marginal increase in computational cost.

During this study, we also observe the lack of benchmark datasets for the TV rating forecast problem. We sympathize with the fact that, collecting and publicizing data that contain demographic characteristics of users or users' viewing habits may be problematic due to privacy concerns. However, there are already studies in the literature using such data, but with little effort for replication of the produced results. In this study, although we attempted to collect a comprehensive data from several publicly available resources, we cannot produce a comparative analysis due to the lack of such reproducible resources. An important part of future

research should be preparation and dissemination of even larger control and baseline data collections for the rating forecast problem involving a wider range of user and item attributes.

There are several other future directions this work can point to. The work presented in this paper verifies that social media-related analysis can positively effect TV rating forecast performance: more personalized analysis involving user habits and preferences, involving user demography prediction can also be integrated into TV rating forecast problem, or even during the data collection and estimation process. Using information collected from multiple social media platforms, and generating automated models for fusion of such information bases in a seamless fashion can be crucial for future research considering the scarcity of rating forecast data. Additionally, we firmly believe that this study shows that using search trends in an automated fashion in forecast problems is an adequate replacement for using temporal data, and other forecast related problems can facilitate such features in the future.

Acknowledgment

Tayfun Küçükyılmaz conceived the idea, Büşranur Akgül performed the experiments, all authors interpreted the results andF prepared the manuscript.

References

[1] Senturk R. Televizyon dizilerinin kesfi, turk televizyon dizileri efsanesi ve gercekler. Istanbul, TR: Kaknüs İletişim, 2018 (in Turkish).

[2] Tekelioglu O. Televizyon halleri: dizi dizi Turkiye. Istanbul, TR: Habitus Yayincilik, 2017 (in Turkish).

[3] Danaher P, Dagger T. Using a nested logit model to forecast television ratings. International Journal of Forecasting 2012; 28 (3): 607–622. doi:10.1016/j.ijforecast.2012.02.008

[4] Cheng YH, Wu CM, Ku T, Chen GD. A predicting model of TV audience rating based on the Facebook. In: 2013 International Conference on Social Computing; Alexandria, VA, USA; 2013. pp. 1034–1037. doi: 10.1109/SocialCom.2013.167

[5] Sommerdijk B, Sanders E, van den Bosch A. Can tweets predict tv ratings? In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16); Portoroz, Slovenia; 2016. pp. 2965–2970.

[6] Krık AM, Domaç A. Sosyal medya üzerinden televizyon reyting olcumlerinin analizi: Twitter ornegi. Akademik Sosyal Araştırmalar Dergisi 2016; 1 (5): 414–430 (in Turkish).

[7] Meyer D, Hyndman RJ. The accuracy of television network rating forecasts: The effects of data aggregation and alternative models. Model Assisted Statistics and Applications 2005; 1 (3): 147–155.

[8] Huang HL, Lee HC, Shu LS, Lai SC, Tsai TM et al. Predicting television ratings and its application to Taiwan cable TV channels. In: 2nd International Symposium on Computer, Communication, Control and Automation. Atlantis Press; 2015. pp. 189–193.doi:10.2991/3ca-13.2013.48

[9] Pagano R, Quadrana M, Cremonesi P, Bittanti S, Formwentin S et al. Prediction of tv ratings with dynamic models. In: ACM Workshop on Recommendation Systems for Television and Online Video (RecSysTV'15); Vienna, Austria; 2015.

[10] Poslodova A. Recommendation system for next generation of smart TV. MS, Griffith University, Queensland, Australia, 2017.

[11] Song L, Shi Y, Tso GKF, Lo HP (2021). Forecasting week-to-week television ratings using reduced-form and structural dynamic models. International Journal of Forecasting 2021; 37 (1): 302–321.

[12] Danaher PJ, Dagger TS, Smith MS. Forecasting television ratings. International Journal of Forecasting 2011; 27 (4): 1215–1240. doi:10.1016/j.ijforecast.2010.08.002

[13] Nan M, Patrick W, Qin H, Wenjia L, Ying Z et al. Prediction of television audience rating based on Fuzzy Cognitive Maps with Forward Stepwise Regression. International Journal of Pattern Recognition and Artificial Intelligence. 2017; 31 (7): 1–13. doi: 10.1142/S0218001417500203

[14] Sereday S, Cui J. Using machine learning to predict future tv ratings. Neilsen Journal of Measurement. 2017; 17 (1): 3–12.

[15] Nan M, Sicheng Z, Zhen S, Xiuping W, Yun Z. An improved ridge regression algorithm and its application in predicting TV ratings. Multimedia Tools and Applications 2019; 78(1): 525–536. doi: 10.1007/s11042-017-5250-4

[16] Wang L. Forecast Model of TV Show Rating Based on Convolutional Neural Network. Complexity, 2021. doi:10.1155/2021/6694538

[17] Nixon L. Predicting your future audience's popular topics to optimize TV content marketing success. In: Proceedings of the 2nd International Workshop on AI for Smart TV Content Production, Access and Delivery. Seattle, WA, USA; 2020.

[18] Hunter III SD, Chinta R, Smith S, Shamim A, Bawazir A. Moneyball for tv: A model for forecasting the audience of new dramatic television series. Studies in Media and Communication 2016; 4 (2): 13–22.

[19] Han JPJ, Kamber M. Data mining: concepts and techniques. Waltham, MA, USA: Elsevier, 2011.

[20] Friedman J, Hastie T, Tibshirani R. The elements of statistical learning. New York, NY, USA: Springer series in statistics, 2001.

[21] Zhang T. Adaptive forward-backward greedy algorithm for sparse learning with linear models. Advances in Neural Information Processing Systems 2009; 1 (1): 1921–1928.

[22] Gelman A, Jakulin A, Pittau MG, Su YS. A weakly informative default prior distribution for logistic and other regression models. The Annals of Applied Statistics 2008; 2 (4): 1360–1383.

[23] Quinlan JR. Induction of decision trees. Machine Learning 1986; 1 (1): 81–106.

[24] Nelder J,Wedderburn R. Generalized linear models. Journal of the Royal Statistical Society 1972; 135 (3): 370–384. doi:10.2307/2344614

[25] Venables WN, Ripley BD. Modern Applied Statistics with S-Plus. New York, NY, USA: Springer-Verlag, 2002.

[26] Breiman L. Random forests. Machine Learning 2001; 45 (1): 5–32.

[27] Platt J. Sequential minimal optimization: A fast algorithm for training support vector machines. Microsoft 1998.