TV POPULARITY SHOW ANALYSIS USING MACHINE LEARNING

HARIHARAN D [1]

Mrs. V. BAKYALASHMI [2]

PG Student [1], Assistant Professor [2],

PG & Research, Department of Computer Applications,

HINDUSTHAN COLLEGE OF ARTS AND SCIENCE

COIMBATORE, INDIA

Abstract: Television is an ever-evolving, multi-billion-dollar business industry. TV shows are more successful Technological society is a huge formula with multiple variables. Art Success is not something that happens, it is studied and Duplicate and apply. Hollywood success is unpredictable, with many movies and sitcoms being hyped and the promise of a hit ended in a box office failure Disappointment. In current research, linguistic exploration describes the relationship with the TV series Appeal to your audience community. Have a decision support system you will need to be able to display reliable and predictable results gives you confidence in investing in new TV series. Of the model presented in this study uses data to explore and determine. This article uses the description Predictive modeling techniques for assessing persistence Television comedy successes: office, big bang theory, Arrested Development, Scrubs, and South Park. The factor is tested for statistical significance for episode ratings Character presence, director, and writer. View these stats Characters are very important to the show itself, the creation and direction of the show ratings, and therefore the success of the show. Use the machine learning-based predictive models such as linear regression, K nearest Neighbors, Stochastic Gradient Descent, Decision Trees and Forests, Neural Networks, Facebook Prophet accurately predict the success of your show. The model represents Fundamentals for Understanding TV Show Success How Producers Can Boost Current TV Success Use or use this data when creating future shows. Deadline Many factors that go into the series, empirical analysis this study shows that there is no one-size-fits-all model for prediction. Rating or success of a television program. But, even if your variables are statistically significant, you can optimize them positively affect the rating.

Keywords: linear regression, K nearest Neighbors, Stochastic Gradient Descent, Decision Trees and Forests, Neural Networks

I. INTRODUCTION

At the time of this writing, TV ratings are often manually projected by teams in the TV channel's forecasting department. These predictions form the basis for where ads and programs appear on the daily television schedule. Regardless of how experienced the employees in this department are, mistakes can happen at work. Workers sometimes advertise at the wrong time of day when creating TV shows. This may be due to misjudgment of slot viewing demographics or TV ratings. This can result in, for example, improperly priced ads or breach of contract, which can lead to financial losses and reduced viewership. Automated processes can help resolve these issues.

The process of deciding when to show an ad is Automate more accurately compared to manual processes. Avoid the wrong ads at the wrong time. That is Potential increase in financial profit for television stations due to reduced labor costs Pricing Your Ads Right and Displaying Better Advertising to a targeted audience. The overall goal of this paper is to develop a prediction module for TV ratings. Automate processes using supervised machine learning Allocate ads to the most appropriate time slots for daily TV programming. Specifically, this project

- Evaluate the predictive accuracy of two or more machine learning models
- Find the best model for predicting TV viewership.
- Based on evaluation and related work previously performed model, choosing the best model,
- Train and tune the best model for predictive accuracy error 1.5%. This is her TV4 mean error in predicting TV ratings at the time.

Only supervised machine learning training models are considered in this work Considerations when using and testing different predictive models to predict TV ratings. More specifically, the related work, testing, and implementation consider only supervised learning models. Only partly due to time constraints

A supervised machine learning training model is tested. Supervised machine learning is a technique that uses acquired knowledge Make predictions from past and current data. So we need people for that handle input and output as well as feedback on accuracy Predict. Unsupervised machine learning works when the training data is unclassified or labeled. Semi-supervised machine learning

uses both supervised and unsupervised learning education. Reinforcement machine learning uses a reward system to train. The training data should be pristine, not labeled data. A study by Indu Kumar et al. [2] Comparison of different monitored machines Learning Algorithms and Their Accuracy in Supervised Learning Predict stock prices. A different algorithm turned out to be better mI found the size of the dataset to be a factor to consider. Who supervises the machine learning algorithm tested in this study is Support Vector Machine. (SVM), Random Forest (RF), K Nearest Neighbors (KNN), Naive Bayes (NB), and Softmax algorithm. Of these tested algorithms, RF was shown to be the best in terms of prediction accuracy for large datasets. To the smaller among the datasets, Naive Bays showed the highest predictive accuracy in stock price inference.

The study uses 12 different technical indicators to Stock price. Reducing the number of indicators resulted in less accurate results when the test is done. In another study by Jiao and Jakubowicz [3], supervised machine learning Logistic Regression (LR), RF, Neural Networks, Gradient Boosted Models Decision trees (GBDT) have been tested and used to predict stock prices.

II. RELATED WORKS

Yu-Hsuan Cheng et al. [4] conduct research to predict viewership Information mechanically collected from the social media platform Face book learning. The machine learning method used was the Back-Propagation Network, Trained in supervised learning. The prediction result is its mean absolute error (MAE) and mean absolute percentage error (MAPE) and according to Lewis' definition of MAPE, the results were promising. according to Lewis's definition of MAPE, if MAPE $\leq$ 20%, it's a good prediction, all their predictions Forecast values accumulated over one week fell under this definition. There was some deviations in certain cases where there was a premiere or finale another program on the same day. The author recommends considering this to reduce such errors.

Another study on machine learning prediction by Sunakshi Magain and others. [5] We investigated the possibility of predicting the popularity of new cars. Company. This study tested several different models for prediction. ANN, LR, RF, and SVM. The author defined the problem as multivariate

I classified it as a regression problem and classified it as supervised learning. Results of this study showed that SVM has the best accuracy (by relatively small margin). Tested model. Furthermore, the authors state the following about future work in this area: They used his SVM and tried to modify it for even better performance. Chongsheng Zhang et al. [6] wrote a study comparing different classifications in 2017 Algorithms and their accuracy. 11 different algorithms were investigated back then, research and everything was cutting edge. Here are the conclusions of the study: GBDT and RF had the highest average prediction accuracy, and SVM had the best prediction accuracy Average prediction accuracy. Regarding training time, GBDT and RF Slow or average at best. In testing and prediction, GBDT was the fastest, but KNN and SVM were the most efficient across execution times. Depending on the algorithm SVM, RF, etc. did not perform well on some of the datasets used during our investigation, GBDT maintained good accuracy on these datasets.

Despite the power of GBDT, this algorithm is still not suitable for e.g. SVMs and HFs. Additionally, the study notes that GBDT is not well covered. Main research and literature. Ramya Akra et al. [7] conducted research on predicting forecasting methods using machine learning. TV shows will be successful. Especially different regression models. To To achieve this, the dataset used included a variety of TV shows, but was not limited to Past TV ratings, authors, directors, titles, broadcast dates, characters and their numbers of row. The success of a TV show is related to ratings, and the higher the rating, the TV ratings were more successful TV shows. The study concluded that there is no unified algorithm for predicting which TV shows are the best. Success. For example, KNN was good at predicting The Office, but no algorithm could predict it. Predict South Park with enough accuracy to show the show's chances of success Predicted.

According to research, the reason for this may be the number Elements that go into the production of a TV show. This study addresses this future work must include factors such as demographics and schedules to ensure accuracy

III. METHODOLOGY

When it comes to sitcoms, it remains the most famous Popular across generations. Office, big bang theory, arrested development, scrubs, south park has It has stood the test of time and remains popular. In this job Use data from these sitcoms in their respective sitcoms Elements that make up a successful sitcom: Title, director, author, first broadcast date, rating, number of copies lines of each character.

A. Description of the problem

In this work, descriptive statistics, visualization, Hypothesis Testing and Predictive Analysis to Understand Differences and influences of various factors on us Dependent variable, episode rating. This process allows Options for two competing hypotheses:
H0:
Same mean or median episode rating By Factors Tested (Director, Writer, and Number) quotes from famous characters)
• HA:
The mean or median episode rating is Elements tested (director, author, number of lines) (voiced by famous characters) Descriptive statistics are an important step in understanding data. For example, when you find a subset of that data does not have a normal distribution, so perform analysis of variance [25]. Instead we Use the Kruskal-Wallis test [26] for nonparametric data. The test is, Median of two or more groups. If the data is not normal Median is a better estimate of the midpoint.

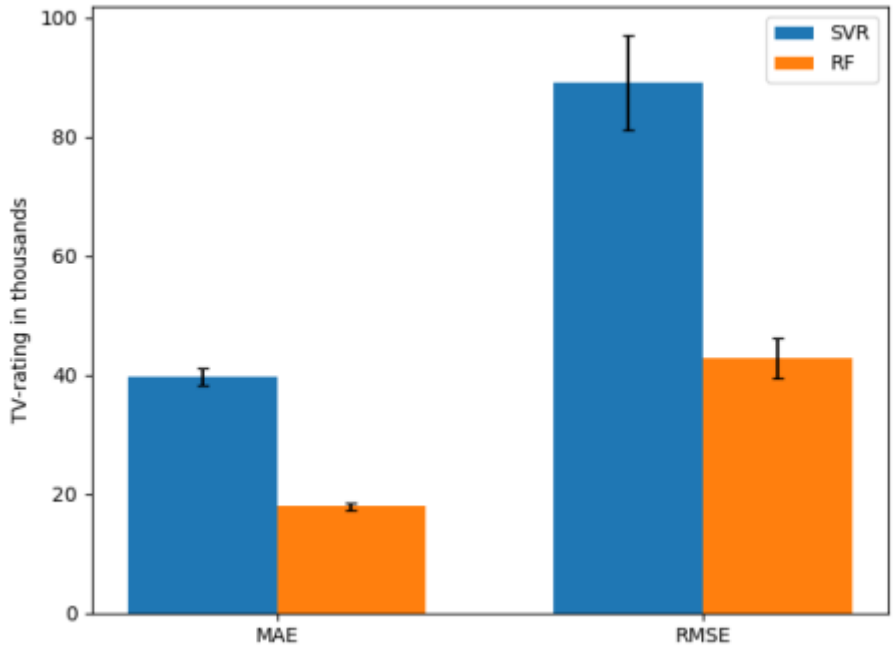

Fig 1 the plot extraction with the correction system

Additionally, visualization helps identify trends within our data. A boxplot is useful for showing the distribution the distribution of the data and how different distributions compare as shown Figure 1. The scatter plot [28] shows the obvious a relationship between two different variables. Sample box a chart visualization of the Office dataset is shown in and Big Bang Theory Pair Plot Visualization Example The dataset is shown in Figure 1. Hypothesis testing [29] is used to test the trend of: determined statistically. Linear regression model [30] indicates a correlation with the presence of characters and episode ratings. The more lines a character speak, the more increased presence in the episode. These models. Helps identify factors that lead to successful episodes. Predictive analytics use data in time series format. Measures how things change over time and is stored in time instructions. Prophet 5 is a way to forecast a time series. Data based on additive models with nonlinear trends

Every year, every week, every day according to seasons and holidays effect. An example of a prophet visualization of seasonal forecasts in Facebook Prophet – Massive Predictions. The South Park dataset is shown in Figure 10. He works best Using Strong Seasonal Effects and Multiple Time Series Season of historical data. Prophet is robust against missing data Trends change and usually handle outliers gracefully. As mentioned above, each episode has the following pertinent elements: Original air dates and ratings used to understand how Episodes occur over time. Exemplary predictions of prophets a visualization of the scrub dataset is shown in Figure 1.

We use machines to predict the outcome of future events learning a predictive regression method based on [31]: Linear regression [30], K nearest neighbors [32], Scholastic Gradient descent [33], decision trees and forests [34], neural network [35]. Machine learning based regression example

A lot of related research shows that the use of supervised machine learning is the most commonly used for prediction. Both SVM and RF use this type of his learning. Among the models that used this type of learning in research, they were most commonly used in his, which showed potential for excellent prediction accuracy. I gave possibilities because most studies had the best or average of in terms of predictive accuracy.

You can narrow down the number of models used and tested. This is because relevant studies have shown with relatively high certainty which models performed well on regression problems. Share exchange value and TV ratings. Television reality shows are increasing day by day with today's generation. There are many ways to find your TV Rating Points (TRP). First, raw data is recorded based on the People's Meter, and views are counted from there. Next, we need to split the entire dataset into clusters based on different channels. Here the dataset consists of channels. Select each channel and count views. Depending on the number of views, rate the channel or view it accordingly. B. If the number of views exceeds 10,000, assign a rating of 10 to the show. If you have new data, add it in the middle of the process and the whole process starts over. We can also update and add new entries in the middle of the process with the help of the proposed algorithm. Rank each TV show with the highest ratings based on the number of views. TRP can be compared between different shows and displayed in bar charts, pie charts and histograms. There are k-means and incremental k-means algorithms for comparing TRPs. A comparison of the two algorithms is very clear in the histogram. This is the easiest way to predict TV show analytics. Inaccurate data can result in erroneous values.

The impact of a character's presence on an episode's rating is calculated by correlating the number of lines spoken by the character with the episode's rating. The boxplot shows the distribution of lines spoken by each character repeated during the run of the series. Linear regression shows that the majority of lines spoken by the cast per episode have no statistically significant effect. However,

her combinations of all characters in the model do. ANOVA is used to test the difference between directors and writers in rating episodes. Some machine learning algorithms use inputs of dialogue spoken by characters, directors, and writers to predict ratings for episodes.

IV. RESULT AND CONCLUSION

Television production is an art, but I suggest it can also be a science. Having a decision support system that predicts success can build confidence in investing in new TV series. The presented model used pieces of historical data to explore and judge sitcom excellence. Not all series are successful, but factors such as character presence, director, and writer have been shown to be accurate predictors of his IMBD rating for a series. In addition, the screenwriter and director, has a stronger impact than the character's presence. Was also not always successful, but the machine learning algorithm was able to predict the score of the sequence more accurately. Office was the most successful in prediction, being able to predict all episodes tested within units of actual ratings, with an R 2 of 0.398. Despite the office's success, South Park fell short of expectations. This work shows that there is no single model for predicting ratings or success. Probably because there are a huge number of factors involved in producing a TV show. The presented model provides a basis for understanding his performance for television shows and how producers can improve the performance for current television shows or use the data to create future shows. Provide. Further analysis should focus on many other factors such as demographics, advertising budgets, web exposure, topics and schedules to fully account for the overall distribution of ratings for shows. However, the variables character presence, director, and writer were statistically significant, so they can be optimized to improve the score**.**

V. REFERENCES

[1] R. Saravanan and P. Sujatha, "A State of Art Techniques on Machine Learning Algorithms: A Perspective of Supervised Learning Approaches in Data Classification," 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2018, pp. 945-949. Cited 27-03- 2020. ISBN: 978-1-5386-2842-3. Available at: https://ieeexplore-ieeeorg.focus.lib.kth.se/document/8663155

[2] I. Kumar, K. Dogra, C. Uterja, P. Yadav, "A Comparative Study of Supervised Machine Learning Algorithms for Stock Market Trend Prediction", 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 2018, Coimbatore, India, pp. 1003-1007. Cited 27-03- 2020. ISBN: 978-1-5386-1974-2. Available at: https://ieeexplore-ieeeorg.focus.lib.kth.se/document/8473214

[3] Y. Jiao and J. Jakubowicz, "Predicting stock movement direction with machine learning: An extensive study on S&P 500 stocks," 2017 IEEE International Conference on Big Data (Big Data), Boston, MA, USA, 2017, pp. 4705-4713. Cited 27-03-2020. ISBN: 978-1-5386-2715-0. Available at: https://ieeexplore-ieeeorg.focus.lib.kth.se/document/8258518/

[4] Y. Cheng, C. Wu, T. Ku, G. Chen, "A Predicting Model of TV Audience Rating Based on Facebook", 2013 International Conference on Social Computing, 2013, Alexandria, VA, USA, pp. 1034-1037. Cited 27-03-2020. ISBN: 978-0-7695-5137-1. Available at: https://ieeexplore-ieee-org.focus.lib.kth.se/document/6693464/

[5] S. Mamgain, S. Kumar, K. M. Nayak and S. Vipsita, "Car Popularity Prediction: A Machine Learning Approach", 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 2018, pp. 1-5. Cited 27-03-2020. ISBN: 978-1-5386-5257-2. Available at: https://ieeexplore-ieee-org.focus.lib.kth.se/document/8697832/

[6] C. Zhang, C. Liu, X. Zhang, G. Almpanidis, "An up-to-date comparison of stateof-the-art classification algorithms", Expert Systems with Applications, Vol. 82, 2017, pp. 128-150. Cited 27-03-2020. ISSN: 0957-4174. Available at: https://www.sciencedirect.com/science/article/abs/pii/S0957417417302397

[7] R. Akula, I. Garibay. "Forecasting the Success of Television Series using Machine Learning", IEEE SoutheastCon 2019, Huntsville, AL, USA, 2019. Cited 27- 03-2020. Available at: https://www.researchgate.net/publication/332530593_Forecasting_the_Success_ of_Television_Series_using_Machine_Learning

[8] Scikit-learn. 1.4. Support Vector Machines. Cited 03-24-2020. Available from: https://scikit-learn.org/stable/modules/svm.html

[9] S. Ray, "A Quick Review of Machine Learning Algorithms," 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), Faridabad, India, 2019, pp. 35-39, Cited 05-05-2020. ISBN: 978-1- 7281-0211-5. Available at: https://ieeexplore.ieee.org/document/8862451

[10] A. Géron, "Hands-On Machine Learning with Scikit-Learn & TensorFlow", 1 ed. O'Reilly Media, Inc, USA, 2017-03-24, pp 154-155. ISBN: 9781491962299.

[11] Scikit-learn. sklearn.svm.SVR. Cited 05-05-2020. Available from: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html

[12] A. Géron. "Hands-On Machine Learning with Scikit-Learn & TensorFlow". 1 ed. Sebastopol, CA, USA, O'Reilly Media, Inc. 2017-03-24. pp 181 ISBN: 9781491962299.

[13] L. Breiman. "Random Forests", Machine Learning, Vol. 45, 2001, pp. 5–32. Cited 27-04-2020 ISSN: 0885-6125 (Print) 1573-0565 (Online). Available at: https://doi.org/10.1023/A:1010933404324

[14] Juan Huo, Tingting Shi and Jing Chang, "Comparison of Random Forest and SVM for electrical short-term load forecast with different data sources," 2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS), Beijing, 2016, pp. 1077-1080. Cited 29-04-2020. ISBN: 978-1-4673- 9904-3 Available at: https://ieeexplore.ieee.org/document/7883252 "Time cost"

[15] Scikit-learn. 3.2.4.3.2. sklearn.ensemble.RandomForestRegressor. Cited 05- 05-2020. Available at: https://scikitlearn. org/ stable /modules/generated /sklearn. ensemble.Random ForestRegressor. ht ml

[16] A. Géron. "Hands-On Machine Learning with Scikit-Learn & TensorFlow". 1 ed. Sebastopol, CA, USA, O'Reilly Media, Inc. 2017-03-24. pp 37-40 ISBN: 9781491962299.

[17] A. Géron. “Hands-On Machine Learning with Scikit-Learn & TensorFlow”. 1 ed. Sebastopol, CA, USA, O'Reilly Media, Inc. 2017-03-24. pp 62-64 ISBN: 9781491962299.
[18] A. Géron. “Hands-On Machine Learning with Scikit-Learn & TensorFlow”. 1 ed. Sebastopol, CA, USA, O'Reilly Media, Inc. 2017-03-24. pp 39-40 ISBN: 9781491962299.
[19] Scikit-learn. Scikit-Learn Machine Learning in Python. Cited 20-04-2020. Available at: https://scikit-learn.org/stable/
[20] Scikit-learn. sklearn.grid_search.RandomizedSearchCV. Cited 20-04-2020. Available at: https://scikitlearn.org/0.16/modules/generated/sklearn.grid_search.RandomizedSearchCV.ht ml
[21] Scikit-learn. Comparing randomized search and grid search for hyperparameter estimation. Cited 20-04-2020. Available at: https://scikitlearn.org/stable/auto_examples/model_selection/plot_randomized_search.html
[22] Panda. pandas. Cited 20-04-2020. Available at: https://pandas.pydata.org/
[23] NumPy developers. NumPy. Cited 20-04-2020. Available at: https://numpy.org/