Ratings of a show based on the hosts

Max Meiners
214936

DISCOVER YOUR WORLD

ADS&AI

Saturday 20 January 2022

Breda University OF APPLIED SCIENCES

The prediction of the ratings of a show based on the hosts of that show

Max W. Meiners

Data Science & AI, Breda University of Applied Sciences

FAI1.P2-01 Project 1B ADS&AI 2022-23

Bram Heijligers & Zhanna Kozlova

January 20, 2023

Abstract

This report presents an in-depth analysis of the content, ratings, and Twitter data of ‘’Op1’’, a popular television show produced by Banijay. The goal of this report is to use machine learning models to predict the ratings of a show based on the content of the show by analysing the data we were given.

The report begins by providing an introduction to what exact data we were given and how I handled the data. Afterward, I will talk about how the merging of the datasets got into play. Furthermore, I will be talking about the different Machine Learning models that I used and the scores of those models.

The report concludes with an overview of the key findings and suggestions for how the show can continue to improve and grow its audience using the insights from the machine learning models. It will also be useful for Banijay to understand the audience’s preferences, and to make data-driven decisions to improve its programming.

Index

1 Introduction

In this report, I will be utilizing several datasets that contain lots of information related to the television program called Op 1, from NPO1. The data includes details about the program's broadcast date and time, station details, the target audience, and various metrics related to its popularity and engagement. The datasets have been collected from a reliable source and have been pre-processed to ensure the accuracy and completeness of the information. The datasets include a mix of numerical and categorical variables. It did contain some missing values, but I cleaned the datasets sufficiently so that there were no more missing values in there. I will dive deeper into cleaning datasets in the EDA section of this report.
The data includes information about the program's target audience, such as the age group of the viewers, which can be used to analyze which age groups are most likely to watch the program. Additionally, the data includes various metrics related to the program's popularity and engagement on Social Media, such as the number of tweets, replies, likes, and quotes. This information can be used to understand the program's reach and influence on Social Media. The data also includes information about the program's hosts, which can be used to analyze the popularity of the hosts and the impact they have on the program's success. Additionally, the data includes the length of the program, which can be used to analyze whether the length of the program has any correlation with its popularity. Furthermore, the data includes the type of ratings, which can be used to analyze the popularity of the program based on the type of ratings.
In this report, I will use various techniques to explore the dataset and uncover patterns, trends, and insights that can help us better understand the performance of the program. The goal is to provide actionable insights that can be used to guide future decisions regarding the program's content, broadcast schedule, and target audience. The analysis of this dataset will provide valuable information that can be used to improve the program's performance and reach a wider audience.

2 Data preparation

For this project, we were asked to create a machine-learning model that predicts the ratings of a show based on the content of the show, and the Twitter metrics. To get to this point, we were given three different datasets:
Content data. This data contains columns like the date when a show aired, the hosts from that show, the length, and the summary of that show.

1. Ratings data. This data contains columns like the time the show started, the station, the target group, the ratings type, and the Kdh000.
2. Twitter data. This data contains columns like the language of the tweet, the id of the author, at what time the tweet was created, and the username of the author.
3. For the data preparation, I will only be talking about the content- and the ratings data. The Twitter data will be talked about in the EDA section of this report.

The content- (figure one) and the ratings data set (figure two) were provided by Banijay. Because the two datasets were already cleaned by Banijay, I only had to drop the ''NaN'' values in the datasets and convert the columns named ''date'' and ''Datum'' from an Object to a DateTime. In figure one, you can see the ''id'' column, which contains the id of the show and the fragment of that show. I split this column, making two new columns named ''show_id'' and ''fragment''. To make it even clearer, I combined created two new columns named ''date_time_start'' and ''date_time_end'', where I combined the date with the start and the end time of the show. After this was done, I could work on merging the two datasets.

2.1 Merging the datasets

Merging the two datasets was very difficult. It took me several days to create a code that worked. The code took a surprising one hour to run. With the code I created I, unfortunately, encountered the problem that some columns were deleted, so I had to come up with a different one before I could continue with the EDA. With the help of some peers, I used a new code, which only ran for around eight minutes, and managed to merge the two datasets correctly.
Figure three shows what the merged data set looks like after I did the merge. With the merge, it created ''NaN'' values, so I had to drop them. After I did that, I could finally move on to the EDA.

	date	hosts	id	length	start	end	title	summary	keywords
0	01-02-2021	[Hilbrand, Sophie, Logtenberg, Hugo]	OP1_____-WON02197428_01_segment	00:21:21	22:20:11:10	22:41:32:12	Gerard Smetsers, Annelie Jager en Károly Illy ...	De basisscholen mogen weer open, maar dat bete...	[directeur, gesloten, basisscholen]
1	01-02-2021	[Hilbrand, Sophie, Logtenberg, Hugo]	OP1_____-WON02197428_02_segment	00:11:04	22:41:36:17	22:52:41:03	IC-arts Hugo Touw wil versoepeling van de coro...	Intensivisten pleiten voor versoepeling van de...	[accepteren, waarom, coronamaatregelen]

Figure one

	Datum	Time	Program	Station	Target Group	Broadcast Type	Ratings Type	Kdh%	Kdh000	Zadl%
0	2020-01-06	22:18:00	op1	npo1	tot6plus	live/prerecorded uitzendingen	uitzenddag	9.676969	1546.863437	30.881672
1	2020-01-06	22:18:00	op1	npo1	tot6plus	live/prerecorded uitzendingen	uitgesteld	1.484044	237.224411	24.692727

Figure two

Figure three

3 Exploratory Data Analysis

3.1 Average Kdh000 per target group

To be able to understand the target audience better, I created a bar chart to visualize the average Kdh000 per target group. This chart can be seen in figure four. The x-axis represents the target group and the y-axis represents the Kdh000. The chart shows that the target group named ‘’tot6plus’’, which is all the ratings of the target group combined has the highest average ‘’Kdh000’’, followed by 50 plus, ‘’v_6plus_jr’’ and ‘’m_6plus_jr’’. ‘’v_6plus_jr’’ and ‘’m_6plus_jr’’ is all of the people, 6 years and up, divided into men and women.

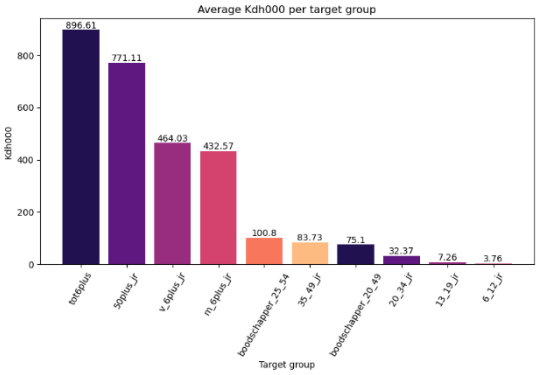

Figure four

For this visual, and the coming ones, I decided to add annotations to the bars, so it would be visible what the average Kdh000 for each target group is.

3.2 Top five most highly rated hosts

To get to know what the five most highly rated, I created another bar chart, which can be seen in figure five, where the x-axis represents the name of the hosts and the y-axis represents the average Kdh000. As you can see, ‘’Carrie ten Napel’’, ‘’Charles Groenhuijsen’’ and ‘’Welmoed Sjitsma’’ are the most highly rated, with an average Kdh000 of 452.41. ‘’Erik Dijkstra’’, ‘’Willemijn Veenhoven’’, ‘’Sophie Hilbrand’’, ‘’Hugo Logtenberg’’, and ‘’Tim Hofman’’ close off the top five with an average Kdh000 of 406.3.

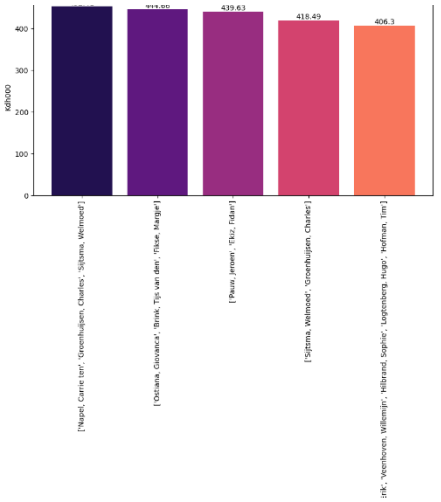

Figure five

3.3 Top 5 most highly rated shows

A bar chart was created to be able to visualize the top 5 most highly rated shows of Op1. This bar chart can be seen when looking at figure six. The x-axis represents the show name, also known as the id, and the y-axis represents the average Kdh000. The average Kdh000 goes from 684.21 to 609.39.

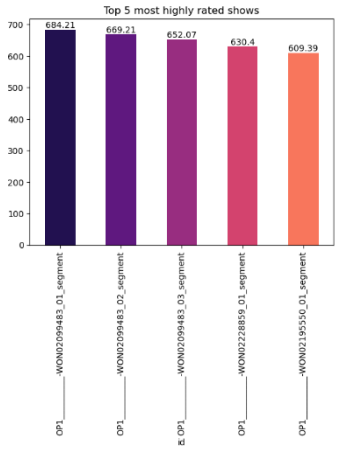

Figure six

3.4 Word cloud

I created a word cloud of the keywords of the shows. This word cloud can be seen when looking at figure seven. Some words that appear are ‘’maatregelen’’, ‘’coronavirus’’, ‘’gasten’’, and ‘’overstromingen’’. These are some of the topics that were talked about on shows.

Figure seven

3.5 Average rating for all shows, per month and day of the week

For both visuals, I created a line chart to best visualize it. Starting with the average rating for all shows, per month, which can be seen in figure eight. It clearly shows that the average rating in the month of April was the highest with 336.44 Kdh000. When looking at the other end, it shows that the average rating was at its lowest in the month of September with 231.46 Kdh000.

Moving on to the average rating for all shows, per day of the week, which can be seen in figure nine. For this visual, I also created a line chart. The chart shows that the average rating is at its lowest on Friday with an average Kdh000 of 265.01, where it is at 430.95 Kdh000 on Sunday. This is the highest average rating of the week. It clearly shows that the average ratings rise when the shows air on the weekend.

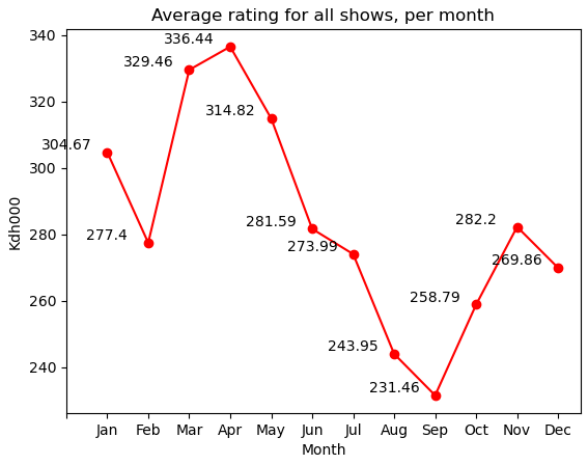

Figure eight

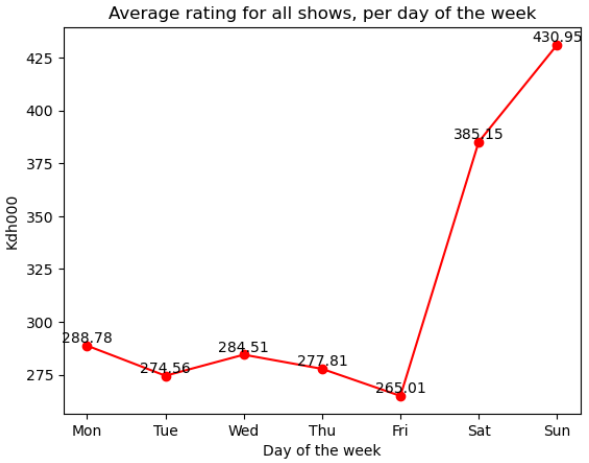

Figure nine

3.6 Average rating for all shows, per target group

When looking at figure ten, you can see that a line chart was created to best visualise the average rating for all shows for the target group ''tot6plus'', per day of the week. What can be seen here is that the average rating on Friday is at its lowest with a Kdh000 of 829.53, and it is at its highest on Sunday with a Kdh000 of 1335.87.

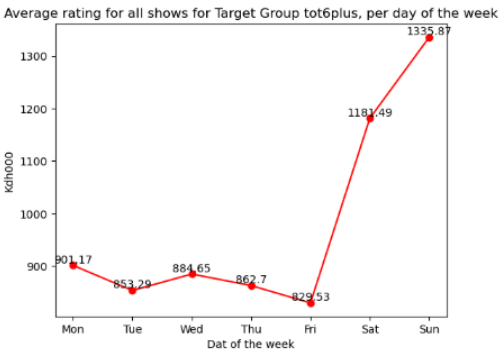

Figure ten

3.7 Social Media Analysis

After the visualizing part, I moved on to analyzing Social Media. The Twitter data we achieved from Twitter. After reading the dataset, I started pre-processing the data. Because there were some ''NaN'' values in the dataset, I had to drop them. I also dropped the columns that were not needed. In the end, I was left with the columns ''created_at,'', ''public_metrics_retweet_count'', ''public_metrics_like_count'', ''public_metrics_like_count'', ''public_metrics_quote_count'', and ''author.public_metrics.followers_count''. I then moved on to merging this dataset with the merged content and ratings dataset. To get that to work, I had to rename the column ''created_at'' to ''Datum'', because I will be merging the datasets on ''Datum''. ''created_at'' also had the time and date combined, so I moved on to splitting them, while only keeping the date. Afterwards, I converted the column to a DateTime object and sorted the data frame by ''Datum''. I then moved on to grouping the data frame by ''Datum'' and summing all of the ''metrics'' columns, instead of ''author.public_metrics.followers_count'', here I took the average. I continued looking at the merged dataset, to see if the column named ''Datum'' was also a DateTime object. After this was done I finally merged the two datasets.

4 Machine Learning models

For this project, I ended up using two different Machine Learning models to predict the Kdh000 metric; a Linear Regression and a Decision Tree. I decided to go for these supervised learning models because they can learn over time. Both models were trained using a dataset that included dummy columns for ''hosts'', ''Datum'', and ''Target Group'' as independent variables and the ''Kdh000'' as the dependent variable. I used the ''train_test_split'' function to split the columns of the x-axis. One in the X_train and one in the X_test. In X_train I have 237029 rows and 684 columns, and in the X_test I have 59258 rows and 684 columns. As you can see, the number of columns stayed the same, but the number of rows did not. 80% got in X_train, and 20% got in X_test.

One of the pros of a Linear Regression model is that it's simple and easy to understand, but that it's sensitive to outliers, which can greatly affect the model's performance. One of the pros of a Decision Tree is that it can handle both categorical and numerical variables, but it may not always result in the most accurate predictions.

4.1 Describing the models

The Linear Regression model, which can be seen in figure eleven, is a supervised learning model used to predict a continuous outcome variable based on one or more predictor variables. The relationship between the variables is represented by a linear equation, and the goal is to find the fitting line through the data points. The model was able to predictions for the Kdh000 metric based on the values of the independent variables, and the results were visualised using a scatter plot.

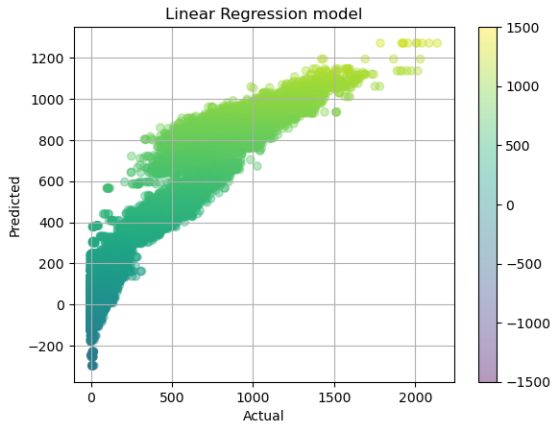

Figure eleven

best-
make

The Decision Tree Regression model, which can be seen in twelve, is a type of supervised learning model that is used both classification and regression tasks. In this case, I used regression task. It is a flowchart-like tree structure, as you see in figure thirteen, where an internal node represents a feature, the branch represents a decision rule, and each leaf represents the outcome. The Decision Tree Regression was trained on the same dataset as the Linear Regression model. The model was able to make predictions for the Kdh000 metric based on the values of the independent variables, and the results were visualised using both a decision tree diagram and a scatter plot.

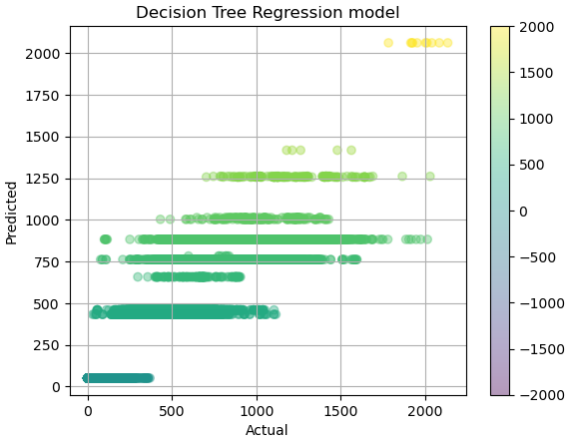

figure
for
it for a
can

node
model

Figure twelve, Decision Tree Regression visualised in

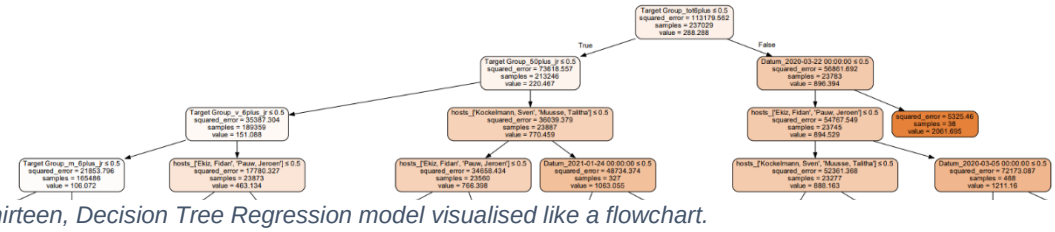

Figure thirteen, Decision Tree Regression model visualised like a flowchart.

4.2 The results of the models

Comparing the results of the two models shows that the Linear Regression model performed slightly better than the Decision Tree Regression model in predicting the Kdh000 metric. However, both models were able to make accurate predictions.

The scores for the Linear Regression model:

- R2: .93
- Mean Squared Error (MSE): 7696.97
- Root Mean Squared Error (RMSE): 87.73

The R-squared value of .93 indicates that the model explains 93% of the variance in the target variable. The higher the number, the better the model is performing. The MSE and RMSE values indicate that the model has a relatively low level of error when predicting the target variable. The lower the MSE and RMSE, the better the fit.

The scores for the Decision Tree model:

- R2: .89
- Mean Squared Error (MSE): 13072.48
- Root Mean Squared Error (RMSE): 114.33

The R-squared value of .89 indicates that the model explains 89% of the variance in the target variable, which is slightly lower than the Linear Regression model. The MSE and RMSE values also indicate that it has a relatively low level of error when predicting the target variable.

4.3 Hyper tuning the model

Because I found no way to hyper-tune my Linear Regression model, I chose to hyper-tune my Decision Tree Regression model. First, I split the the data into training and validation sets. Afterwards, I defined a list of possible hyperparameter values, where ‘’max_depth’’ goes from two to ten, and ‘’min_samples_leaf from one to ten. Furthermore, a grid search was performed by looping through all possible combinations of ‘’max_depth’’ and ‘’min_samples_leaf’’. For each combination, a Decision Tree Regressor model is created using the current hyperparameters, then fit to the training data, and evaluated on the validation data using the MSE metric. After each iteration, the validation MSE, ‘’max_depth’’ and ‘’min_samples_leaf’’ are printed so that the user can see the performance of the model for each set of hyperparameters.

5 Ethical aspects

Banijay works responsibly and in accordance with GDPR standards. They work in a small team and frequently confer with one another. They are extremely open about what they intend to do with the data they have gathered. They also offer client information and their privacy policies on their website. If you do not want to share data with the company, Banijay offers the option to have your data removed.

We also operated responsibly by protecting the privacy of those who shared their data, resulting in anonymized data. We used an API to scrape Twitter data, and we did it ethically because it was done in accordance with Twitter's terms of service. The only thing that may be considered unethical is that we have user ids, which could be used to locate an account with much effort.
Who is in charge of AI? We are personally liable for AI's faults in the project. We created the model, thus if something is wrong with it, we should repair it.

Banijay is open about how they intend to use the information they collect. They also respect people's privacy by stating that the data they collect will be kept private and will not be shared. Banijay supplied us the data, but because everything is anonymous, we didn't acquire any data from people; the only data we have are Twitter account ids, which won't go you very far. Banijay cannot keep the comments that accounts submit under Banijay's posts secret because they are public.
There are numerous approaches to strengthen ethical (AI) organizational capabilities, such as being transparent, which involves being honest in how you proceed with AI. This guarantees that we notify people about the data we are gathering and the benefits it will provide them. Another alternative is to have control over biases; to ensure AI is not racist or discriminating towards a person, datasets and algorithms must be improved and updated.
Banijay's staff is incredibly diverse. There is no difference between males and women. There are many women in the organization, but they are evenly dispersed in terms of roles. It's worth noting that Banijay's whole data team is made up of women. Banijay has no public policy on whether to hire more males or women.

6 Conclusion

We, students from BUas, were asked by Banijay to come up with Machine Learning models to improve show ratings. To do this, we have received several datasets from the company. To make a Machine Learning model, I used various techniques to explore the datasets and uncover patterns between the sets. This provided valuable information. Cleaning and preparing the data I began with, and then moved on to merging the datasets.
Afterward, I looked at the Twitter metrics data we received. Here I dropped the ‘’NaN’’ values, cleaned and prepared the dataset, and then merged it with the previously merged dataset.
After visualizing the datasets, the Machine Learning models came into play. I created a Linear Regression model and a Decision Tree Regressor model. It turned out that the Linear Regression model had an R2 score of 93%, whereas the Decision Tree Regressor model had a score of 89%. I would suggest Banijay to go with a Linear Regression model.
The ethical standards are of utmost importance at Banijay, they are following GDPR standards and they are committed to promoting ethical and responsible business practices throughout their organization.

7 References

Art. 4 GDPR – Definitions - General Data Protection Regulation (GDPR). (2018, March 29). General Data Protection Regulation (GDPR). https://gdpr-info.eu/art-4-gdpr/

Banijay Benelux. (2021, May 11). *Privacyverklaring Kandidaten- en Deelnemersgegevens - Banijay Benelux*. http://defigners.wpengine.com/privacyverklaring-kandidaten-en-deelnemersgegevens/

Privacy Notice - Banijay Group - We are Banijay. (2022, July 20). Banijay Group - We Are Banijay. https://www.banijay.com/privacy-notice/

Banijayrights.com. (n.d.). Retrieved January 20, 2023, from https://www.banijayrights.com/privacy-policy

Banijayrights.com. (n.d.). Retrieved January 20, 2023, from https://www.banijayrights.com/modern-slavery-policy

Lawton, G., & Wigmore, I. (2023, January 18). *AI ethics (AI code of ethics)*. WhatIs.com. https://www.techtarget.com/whatis/definition/AI-code-of-ethics

Life at Banijay. (2022, December 16). Banijay Group - We Are Banijay. https://www.banijay.com/life-at-banijay/

Banijay Benelux. (2021, May 11). *Privacyverklaring Kandidaten- en Deelnemersgegevens*. http://defigners.wpengine.com/privacyverklaring-kandidaten-en-deelnemersgegevens/

Science, O.-. O. D. (2022, January 4). *Best Ethical AI Research for 2021 - ODSC - Open Data Science*. Medium. https://odsc.medium.com/best-ethical-ai-research-for-2021-195ebe25ae90

Games

Leisure & Events

Tourism

Media

Data Science & AI

Hotel

Logistics

Built Environment

Facility

Mgr. Hopmansstraat 2
4817 JS Breda

P.O. Box 3917
4800 DX Breda
The Netherlands

PHONE
+31 76 533 22 03
E-MAIL
communications@buas.nl
WEBSITE
www.BUas.nl

DISCOVER YOUR WORLD

Breda University
OF APPLIED SCIENCES