Predicting ratings of TV show based on target groups and hosts

Matey Nedyalkov, 221889

DISCOVER YOUR WORLD

Breda University
OF APPLIED SCIENCES

Predicting ratings of TV show based on target groups and hosts.

Matey E. Nedyalkov

Data Science and AI, Breda university of Applied sciences

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

Alican Noyan

January 20, 2023

Abstract

The report starts with introduction, where it is described the whole goal of the project as well what elements are used for predicting the rating. Afterwards, there is an entire section, which is explained how the data was used – it was cleaned, pre-processed, all tables were merged and then visualized. This gave a better start for doing the machine learning algorithm, since there is more data to work on and the information of the data collections is way clearer. Two algorithms were chosen as the most appropriate for solving the use-case. The models were evaluated so it can be reached the highest possible accurate score. The description of the model was followed by explanation of the results whether the algorithm can be useful for the business case or not. Finally, there is ethical part, explaining the policies of the company, especially those related to data processing. The report was finished with conclusion and discussion mentioning the usage of AI and data science not just in this project, but in general. Also, it is asked the question what other features can be useful for the use-cases, so the rating can be predicted.

Table of Contents

Predicting ratings of TV show based on target group and hosts

For block B we, the students of Data Science and Artificial Intelligence in Breda University of Applied Sciences, were hired from the production company Banijay to predict the rating of TV show. We are supposed to deliver machine learning algorithm, which must predict the rating based on the features, we have chosen, and report, where we describe the process of it, results and the ethics. The goal of my project is to find out whether specific target groups and all the hosts can help to predict the ratings of a show. The target groups, which were picked for the project, are the 'boodschapper_20_49' and 'boodschapper_25_54'. To see whether these are the right picked features, the data should be cleaned and pre-processed, to merge all the provided datasets into one big flat file, visualize the findings to get better understanding of the data and for the final – chose the most appropriate machine leaning algorithm for the case. Alongside that, it should be done research on whether the company follows an ethical etiquette and what is its attitude towards the data.

1.1 Introduction to the datasets used.

One of the processes, which had to be done at the beginning, was the EDA – Exploratory Data Analysis. To be able to do this part of the project, tree datasets were provided, which to be worked on – content, rating and Twitter data:

- Content data – it contains information about the shows such as titles, keywords, summary, hosts and at what time they were live.
- Rating data – it includes the ratings (Kdh000), target groups, broadcasts, rating types and time.
- Twitter data – there is the data of the twitter metrics, whether the tweet is original, at what time it was published and id of the person, created the post.

1.1.1 Data cleaning and preparation

Firstly, for each of the datasets was checked whether there are missing values and it ended up this can be noticed in the rating and twitter data. For the latter one should be mentioned that the 'NaN' is in the "referenced_tweets" column, which indicates that the tweet is original. Compared to the other two datasets, in the content dataset there were more things to be done. For instance, there were three columns (text, summary and keywords) in the content data collection, which were in Dutch. To make them more accessible for every user, they were all translated in English using the python library deep translator. Furthermore, the columns, which contain date and time were fixed as the milliseconds in the "start" and "end" columns were removed and combined with the "Date" column as they were renamed to "date_time_start" and "date_time_end" respectively. To make the process easier, the data type was converted to datetime object. Also, the same was done for the rating table with the only difference there were not columns, which shows the period of show, but the at what time it was. Lastly, the IDs were split into “content_id” and “fragment” of a specific show, where the former indicates the show’s ID, while the latter is for a specific part of the show. Later it was found that it was better to keep the fragment to the ID and then separate them after the merge, but this will be explained more in detail further. For the rest of the datasets there is not much to be said beside that in the twitter data the time zone in “created_at” was removed and converted to datetime object. Moreover, the “referenced_tweet” column was filtered to NaN, because as it was mentioned before, it means that this is not a retweet and there is a need only of the original post.

One of the most important processes during the whole work with the data was the merging between all the table. Firstly, it had to be matched the content and rating data through “content_id” column as it started with the creating of a separated data frame used as a look up table, which contains only the unique values of “date_time” and “date”. In this way, the IDs would be matched with their respective day and hour. When it was done, the new look up table was merged with the rating data by “content_id” and subsequently the result was combined with the content dataset again on the same column as well. As I mentioned previously, there was an issue with the ‘content_id’ and the fragment of the show – if the fragment was split from the ID before the merge and then the process is done – the fragments are not correctly matched with the hours. The issue was resolved when the fragment was divided from it after the linking the tables. Afterwards, this new big table should be connected to the twitter data. Since we need only the twitter metrics, the data was categorised by ‘created_at’ column, as it were taken only the numerical data (the twitter metrics) and the date was extracted from it as the hours, minutes and second were removed. This means that every day contains the average values of all the metrics for their respective date. In this case, it was calculated the total engagement rate for each day. Then, the finished table was merged to the content and rating data and that is how it was created the final flat file.

1.1.2 Data exploration

The data exploration process started with checking the size of the tables and their content. It ended up that among of them, the rating dataset was significantly bigger. Then, when the first merged table was created (with the content and ratings data), there were implemented visualizations of the data. Firstly, the target group data was categorised with the groupby() function as it was taken the mean of the rating (Kdh000) for each target group as it can be seen in Figure 1

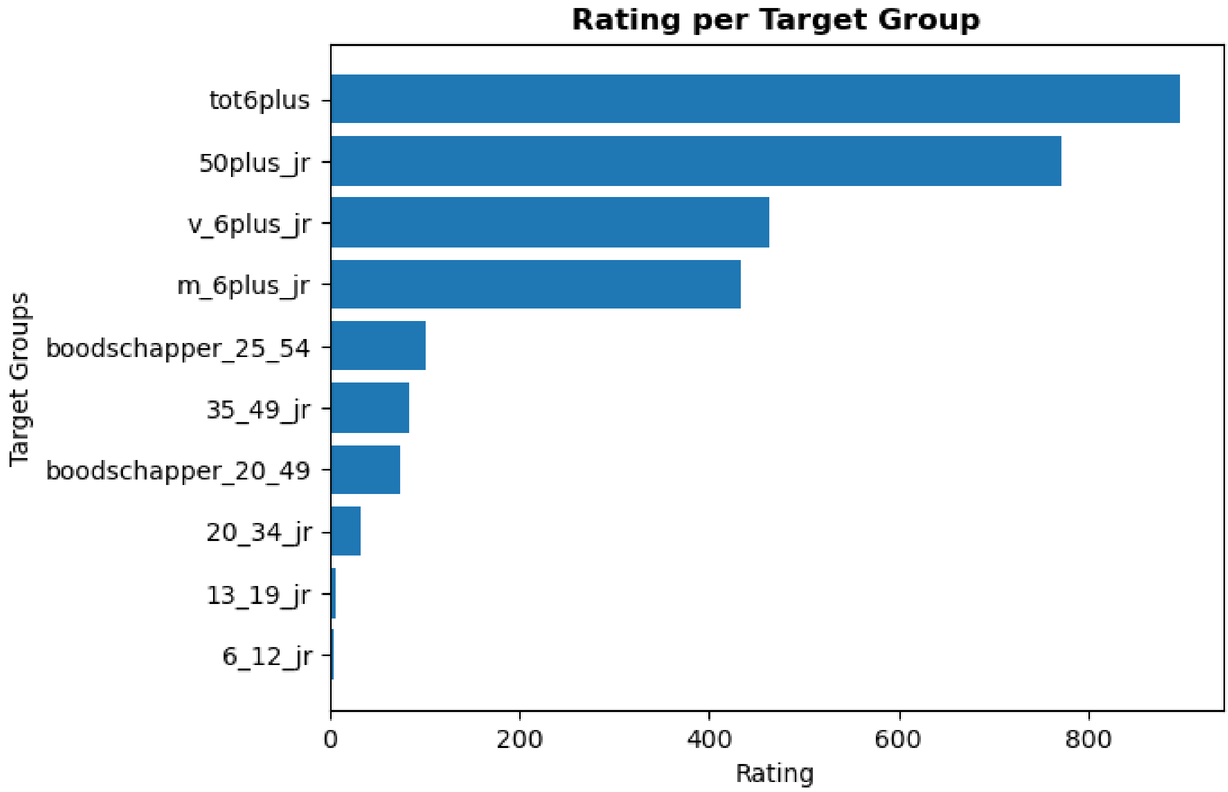

Figure 1

It can be noticed that the group with the highest average rating is the ‘tot6plus’, while ‘6_12_jr’ is with the lowest. Also, the first four category have higher mean than the rest.

Afterwards, it was checked the top five hosts and as it can be seen on the graph the group of Carrie ten Napel, Charles Groenhuijsen and Welmoed Sijtsma is leading.

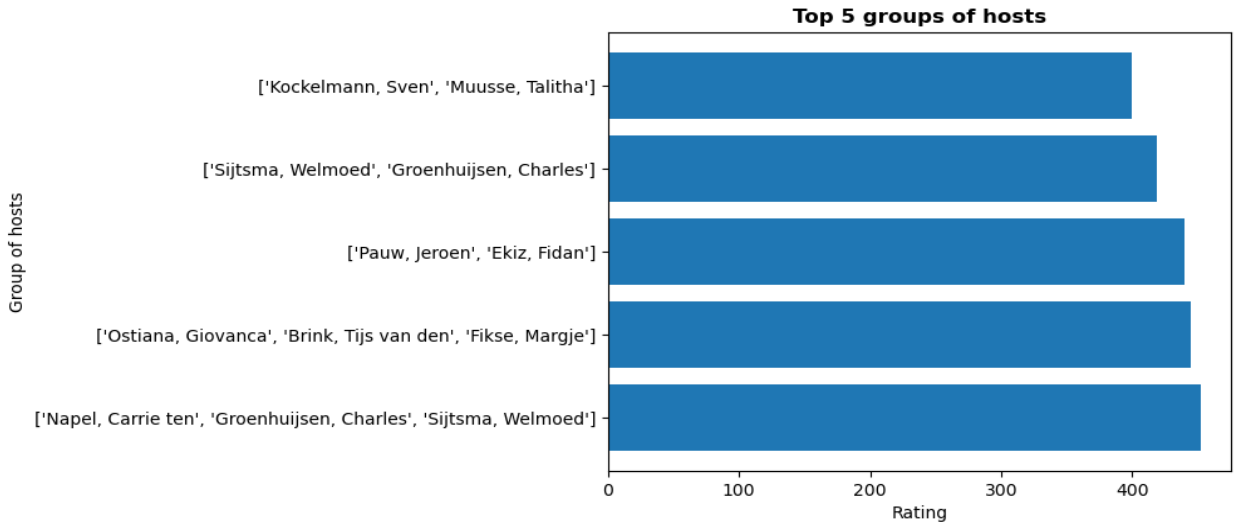

Figure 2

Furthermore, I was supposed to check the top five rated fragments of shows. It is intriguing to be noted that the top three fragments are from the same show, while the rest are from completely different.

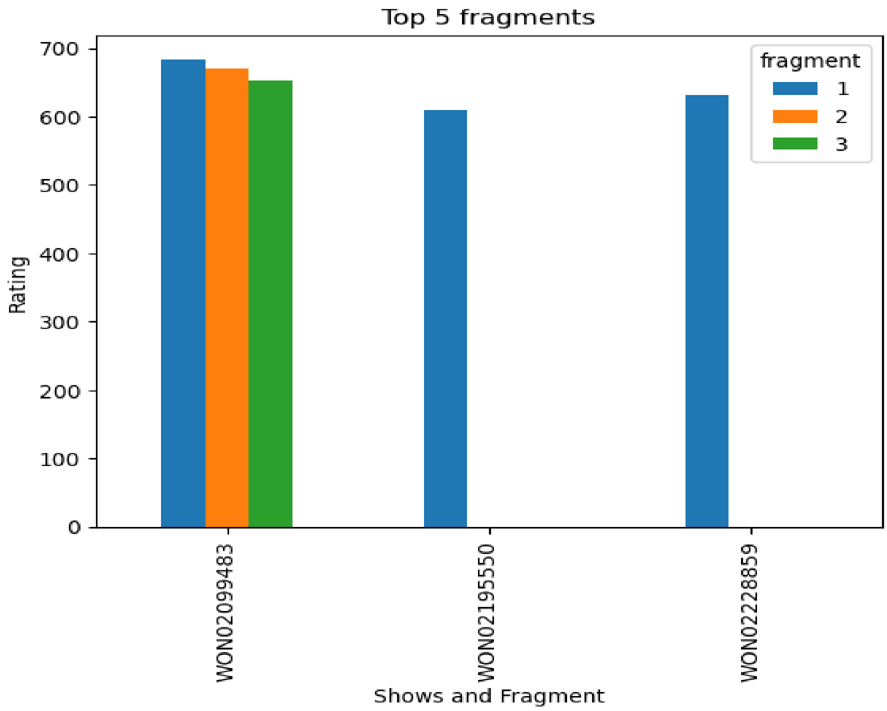

Figure 3

Then, it was created a word cloud of the top five keywords (Figure 4) based on the above-mentioned visualisation with the help of the word cloud library.

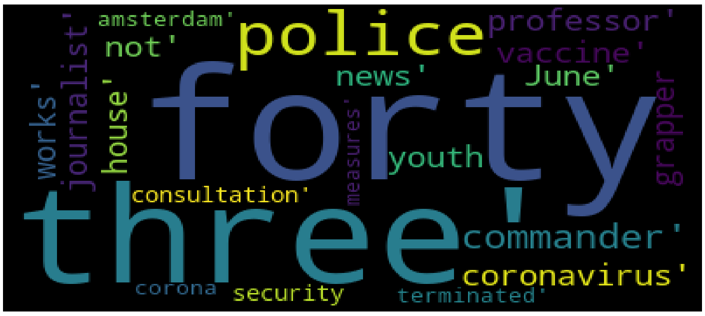

Figure 4

Then it was supposed to check whether the rating is increasing or decreasing throughout the year or the weak. To make this possible, it was taken the average number of ratings for each month and date.

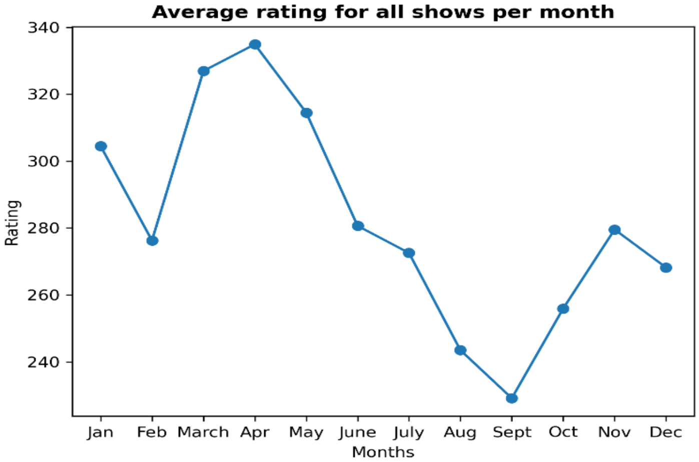

Figure 5

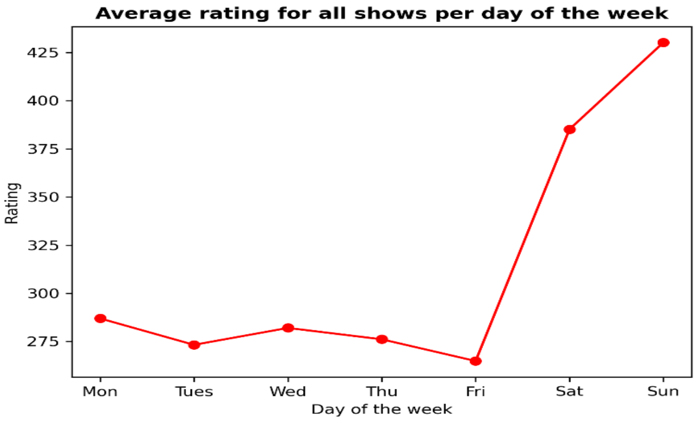

Figure 6

At the beginning of the year, there is a small fall to February, but then it went up significantly until April. Furthermore, this is followed by a significant decrease by September and then rise steadily till December. Moreover, the case during the week is completely different. It can be noticed that there is a slight reduce from the beginning of the week to Friday. However, there is a skyrocket in the weekend, which can make the conclusion that television is mostly watched at that time. Also, the target group ('tot6plus') also was visualised, but it has the same line trend, so there is no need to be visualized.

Then, the dataset with the twitter metrics and engagement rate, it was checked which shows have the highest engagement rate. It can be noticed that the highest total engagement rate is around 0.25%, which is significantly higher than the others. Also, the last two are with almost equal values.

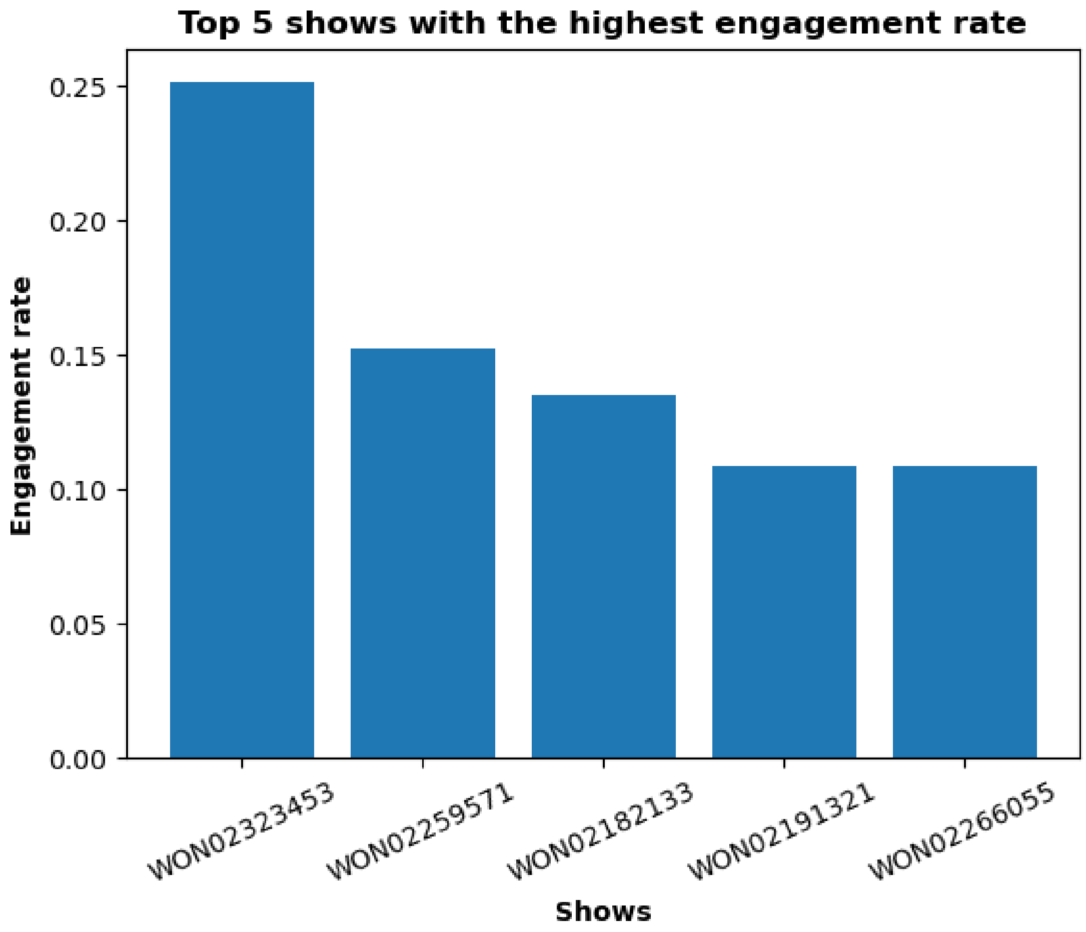

Figure 7

Since the previous goal of the project was to be created a linear regression, which would be able to help to predict rating, it was created a scatter plot and the correlation coefficient of the data was checked. It ended up the data does not look linear, and the correlation coefficient was almost 0 – 0.066169. That influenced the change of the problem statement and choosing more efficient features.

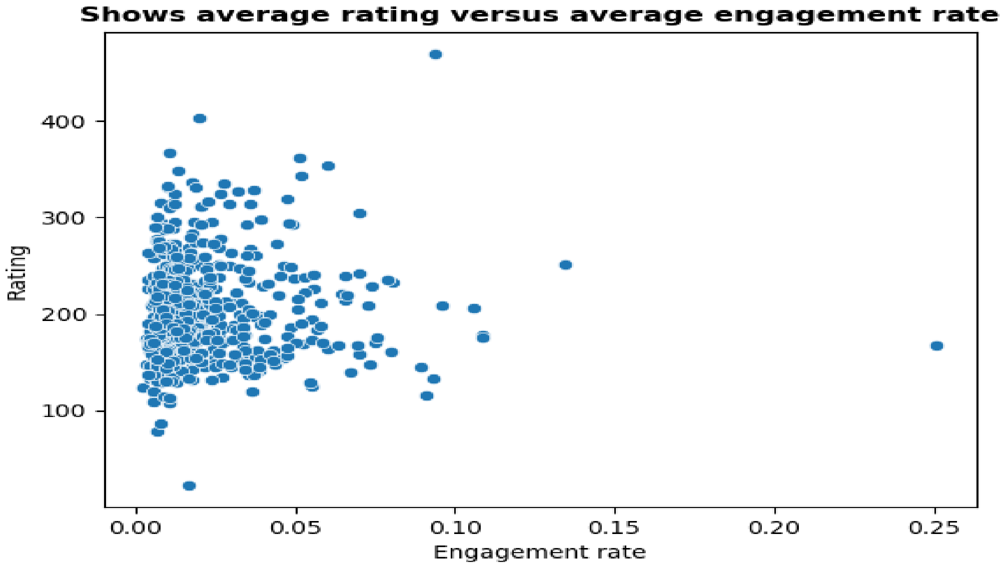
Shows average rating versus average engagement rate
Rating
400
300
200
100
0.00
0.05
0.10
0.15
0.20
0.25
Engagement rate

Figure 8

1.2 Introduction to the machine learning models used.

For the implementation of the machine learning were used a few sklearn libraries. Apparently, it looks more reasonable to be used regression models such as linear or decision tree, because we are working with continuous data. However, to be sure everything is fine, both can be implemented so afterwards can be make comparison between the models results. Evaluation models will be implemented to check whether the performance will be better.

For a reminder, the features, which were used for the model were two target groups - “boodschapper_25_54”, “boodschapper_20_49” and the hosts. I have decided to focus on these specific target groups, because both represent the people, who take care of the household and with it comes the assumption they will spend money on products from advertisements. The ‘hosts’ column was chosen to see whether they can influence on the change of the rating.

1.2.1 Description of the model

Since all target groups contain non-numerical values, it had to be made dummies variables i.e., the string values were converted into integers – 0s and 1s and new column for each target group was created. The features and the target value were picked, it was time to separate them on training and test data as the former is 75% of the whole data, while the latter is the rest of it – 25%. Furthermore, it was implemented cross-validation with the idea to evaluate the performance of the model, which meant that the data would be separated into n number of portions and in my case would be five.

On the other hand, linear regression did not seem to be enough and ended up there is another machine learning algorithm, which can be used for this use-case – decision tree. It is an algorithm, which breaks down the data into smaller pieces and at the same time a decision tree develops. There are two types of it – classification and regression and in the case of the project, the latter seems to be the most appropriate one. The decision tree was set with maximum depth of eight, which is basically the depth of the tree. Maximum depth was chosen based on a hyperparameter tuning using validation test – there were not big difference between the scores, so I decided to go for eight. If the argument is None, the nodes will expand until the leaves are full or contain less min_sample_split. Also, min_sample_split was used to split an internal node.

1.3 Machine learning results

Since the models were trained and tested, it is time to show the results. Firstly, it will be about the linear regression results. The accuracy   of the model was calculated and it ended up being 10%, which is low. Furthermore, the mean square error was inspected. Before presenting the value, the mean square error shows the distance between a data point to the regression line. In the case of the linear regression, it is 319.66. Also, the cross-validation has the similar accuracy as the standard regression. Since it was chosen to separate the data into 5 portions, there are five outputs. Afterwards, the mean of them it was found, and it is exactly 10%. Standard deviation is 0.00153 and quantile is 0.10007 and 0.104.

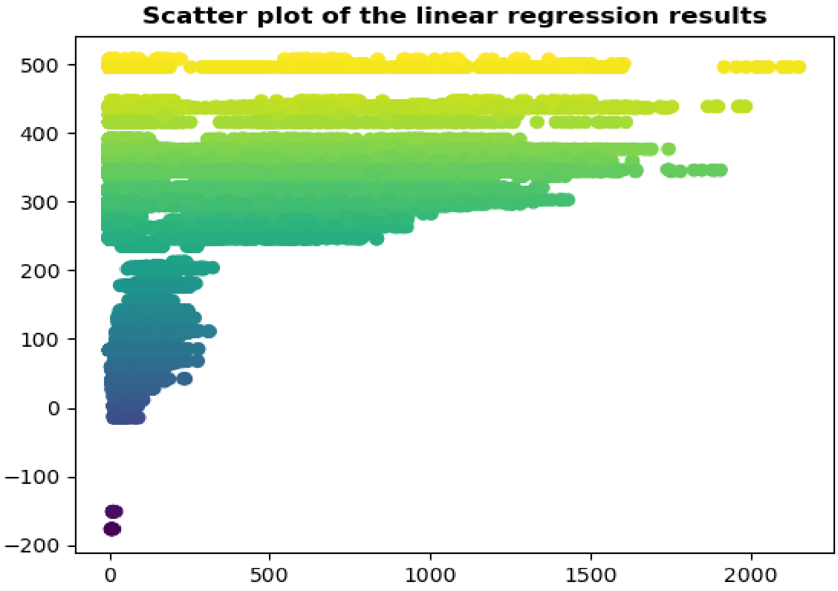

Figure 9

On the other hand, the decision tree regression model ended up being not as accurate as the previously described one. The accuracy score of this model is significantly lower – 0.034 (3%). This basically means that this model cannot be useful for the analysis. Moreover, the mean square error is way higher as it is 331.57.

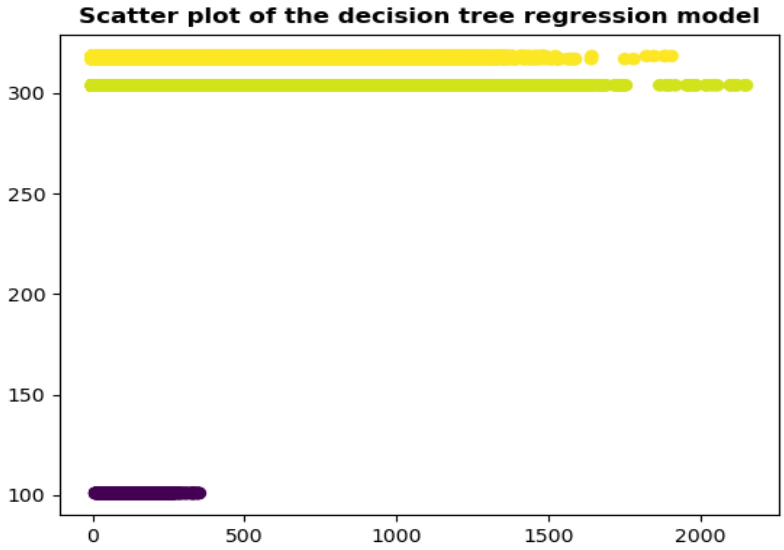

Figure 10

With of the results being explained, it can be reached the conclusion that between both models, linear regression is a little more appropriate than the decision tree regression, even though the score of the former is 10%. Also, based on the score – yes, it can be predicted the rating with the hosts and the target groups ‘boodschapper_20_49’ and ‘boodschapper_25_54’, but it will not be as accurate as if other features were chosen.

1.4 Ethical aspects

During the project, a significant focus was placed on ensuring that the company, Banijay Benelux, implemented ethical policies. Research was conducted to determine the company's adherence to ethical practices. An examination of the company's privacy policy on their website revealed that one of their subsidiaries, EndemolShine, adheres to laws and regulations for protecting data, including the Personal Data Protection Act and the GDPR. The company is transparent in its processing of personal data and only uses such information with customer's consent. Additionally, the use of cookies ensures that the information collected is anonymous.

Furthermore, the company has provided an email address for customers to ask questions about the usage of their data and has a designated data protection officer responsible for ensuring compliance with data protection laws. Regarding the use of machine learning in the project, the employees were responsible for implementing the algorithm to make predictions based on features.

In terms of the company's treatment of its employees, our experience as students from Breda University of Applied Sciences was positive. We were provided guidance by a data scientist and were warmly welcomed during our visit to the company. As a member of the team, I found myself to be competent in both programming and AI, as well as the ethical usage of Artificial Intelligence. Additionally, the team ensured that the data was kept safe and in compliance with the GDPR framework.

Overall, Banijai Benelux has good policies I place for the use and protection of data. The only suggestion for improvement would be to increase diversity within the data science team, specifically in terms of gender, to bring different perspectives to the table. However, no major issues have been identified and the company is doing an excellent job its data policies.

1.5 Conclusion and discussion

When the results of the model were ready, I was certain that this model is not good enough. I have the assumption that if other features were chosen the performance might be even better. ‘Boodschapper_20_49’, ‘boodschapper_25_54’ and hosts do not look like a great combination, so it can start the conclusion what features would be good for predicting the rating of TV show. It might be other target groups, it might be somethings else, to be found out, it must be done another research.

Also, the whole project it shows what advantages gives the usage of machine learning and AI not just in this project, but in general. The data and well-made algorithm can help us take decision on improving in different spheres. Here the results are supposed to check whether Banijay must focus more on the mentioned in the report target groups or in other features from the data. AI can be used in different ways to improve performances in different ways. For instance, it can be used in the sport to analyses the athletes performance and many other ways, which can improve our daily live.

References

Banijay Benelux. (2021, May 11). *Privacy Policy* . https://banijaybenelux.com/privacy-policy/

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

BUAS Appendix 1

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