Banijay Project

Neil Ross Daniel (221270)

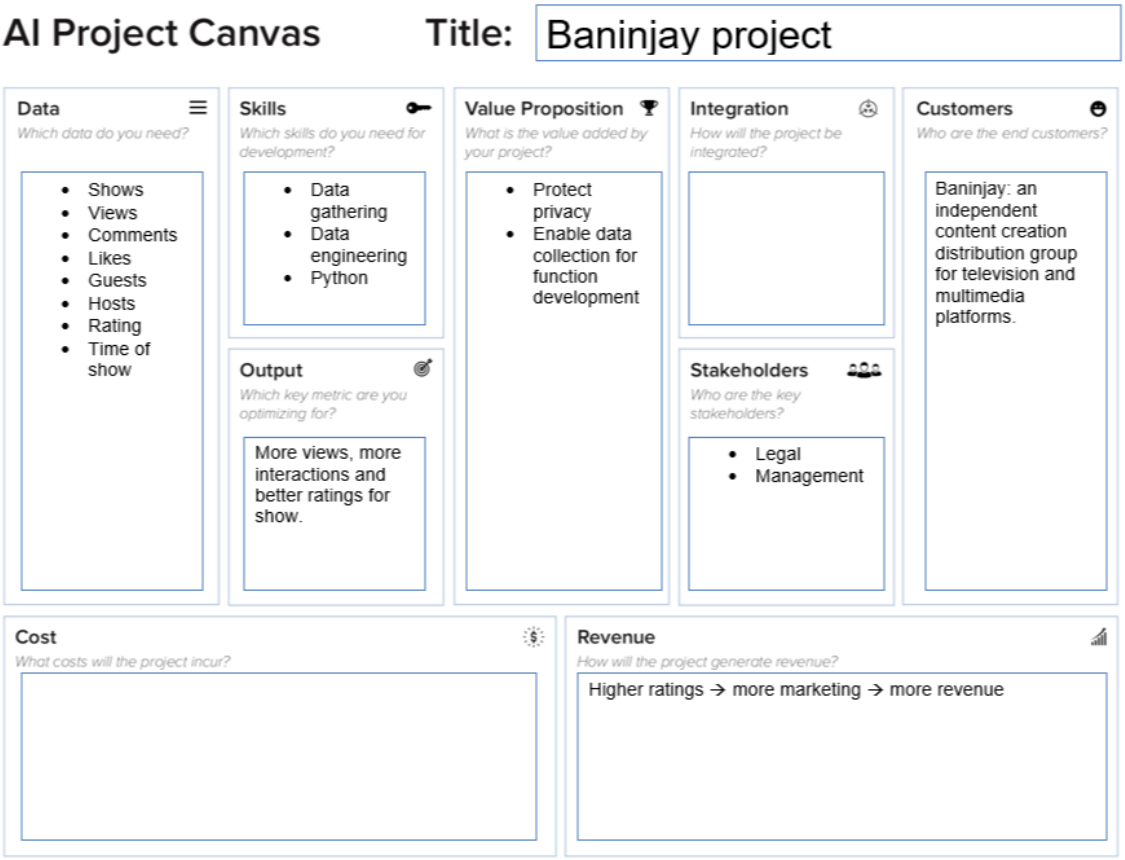
Breda University of Applied Science (BUAS)

Data Science and Artificial Intelligence

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

Mentor: Nitin Bhushan

20th January 2023



Contents

Abstract

Introduction to the business case: This part of the report discusses what the project is about and about Banijay.

Introduction to the datasets used: This part of the report is the Ecplanitory Data Analysation (EDA).

Data cleaning and preperation: This part of the report discusses what we did to clean and prepare the datasets before being able to start the research.

Data exploration: This part of the report discusses what we did to learn more about the datasets that we received.

Introduction to the machine learning models used: This part of the report describes how the machine learning model that was made, why we use the model.

Machine Learning Model: This part of the report describes how the accuracy model was tested.

Ethics: This part of the report describes the ethical parts of this research.

Discussion: This part of the report discusses the results obtained by the visualization, models, gives advice to Banijay.

Conclusion: This part of the report concludes, summerizes everything that was previously mentioned in the report.

Introduction to the business case

In this block we where tasked to improve a business process using digital processes.

We received this project from the Banijay Group, an independent content creation and distribution Group for Television and other media platforms (Banijay, 2022). The goal of the Banijay group is to create, produce and deliver high-quality, multi-genre content and increase its intellectual property though creative entrepreneurship and business acumen. In this project the client (Banijay) has approached us as the data scientist with an aim to analyse their current TV viewership metrics and present a plan to improve their data usage(with multiple sources) and ultimately, help them understand what drives the shows popularity (BUAS, 2022).

Introduction to the datasets used

Data cleaning and preparation

We got 3 different datasets: Content data, Ratings data, and Twitter data.

The Content dataset was given as a pkl. To clean and prepare the Content data, the data was checked if it had any missing values, checked if there was any duplicate values in the data set, deleted the milliseconds part of the date_time column, the *date_time* column was split into *date_time_start* and *date_time_end*, the *date_time_start* and the *date_time_end* was changed into datetime format and finally the *id* column was split into *show_id* and *fragment* columns.

The ratings data was given as a CSV. To clean and prepare the Ratings data, the data was checked if it had any missing value, checked if there were any duplicate values in the data set, combined the *date* column and *time* column to create a *date_time* column and the *date_time* column was changed to datetime format.

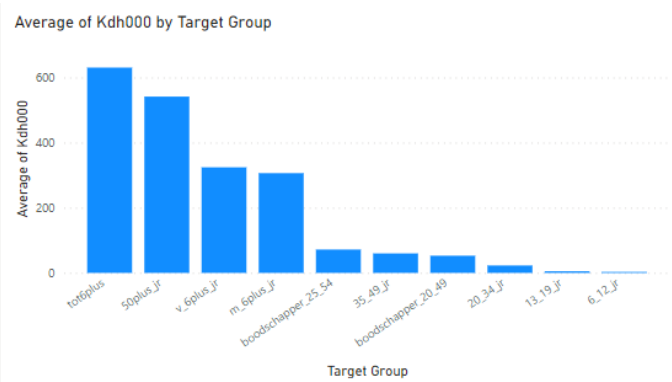
The twitter dataset was given as a json. To clean the and prepare twitter data, the *created_at* column was converted into a datetime format, the tweets were filtered out to only keep the tweets that are not referenced to another tweet (only keeping the columns with *NaN* values).

Data exploration

For the data exploration Power Bi was used, and the analysis focused on the *total* ratings type using *Kdh000* as and indicator on how well a show is performing.

We started b looking at the target audience of the shows, understanding the audience's preferences are and how they differ from each other. This was done by looking at the average Kdh000 and the average Kdh000 per target group.

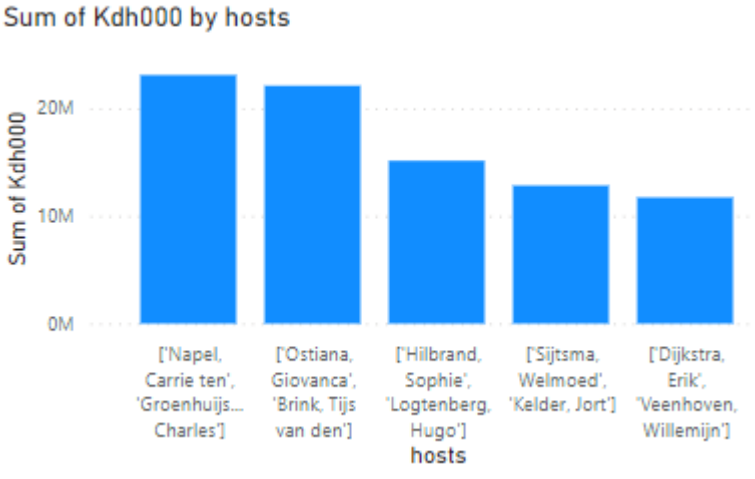
(1)



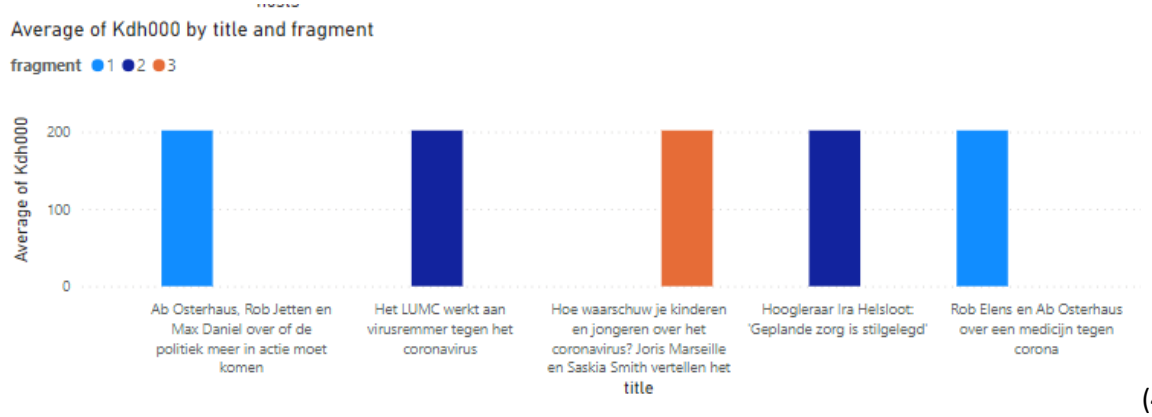
(2)

Here you can see that the average Kdh000 is 201.80 . The most highly rated target groups are *tot6plus*, *50plus_jr* and *v_6plus_jr.*

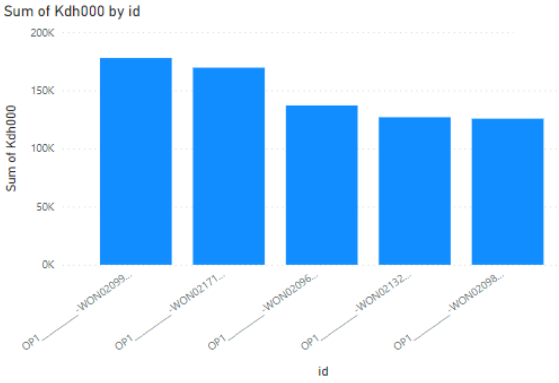
Then we analysed the content to see what kind of content was most liked by the audience. This was done by first checking the most favourite hosts (3), then visualizing the 5 most highly rated show and the 5 most highly rated fragments(4). Finally we visualized the 5 best fragment according to the ratings (5)and made a word cloud of the keywords used in the content of the said fragments(6).



(3)



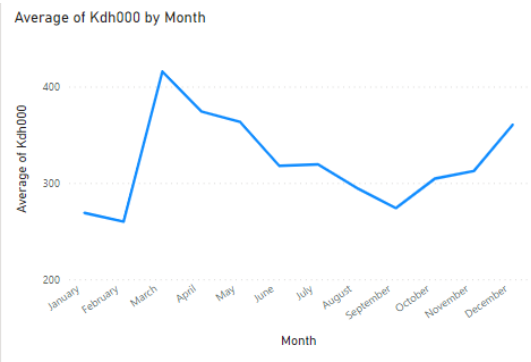
(4)



(5)

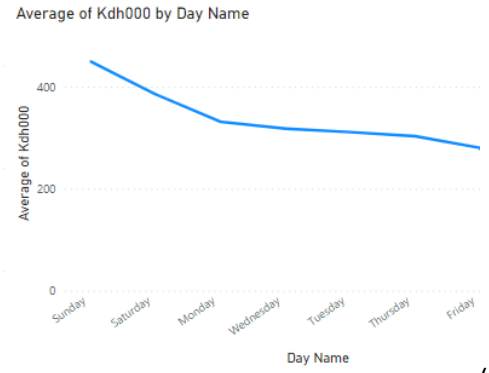
(6)

Finally, we looked at the trends in the ratings data and tried to understand how the ratings changed over time. This was done by first checking the average rating for all shows per month (7)and then average ratings for all shows per day of the week(8). Then we checked the average ratings for all the target groups per day of the week(9).



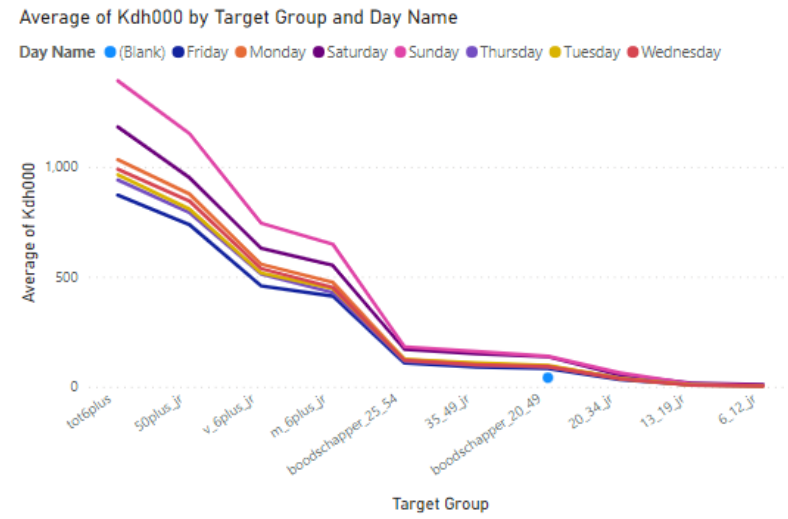
(7)

This trend show that the ratings slowly decline in the beginning of the year, have a massive increase during March and April, gradually decrease till October then slowly increase again until December.



(8)

This trend shows that the ratings are higher during Sunday and Saturday then decrease gradually. This is the case because during the weekdays people have to work/ go to school/ are busy not many people will watch the shows. During the weekends on the other hand people are free and watch more shows, so the ratings are higher on Saturday and Sunday.



(9)

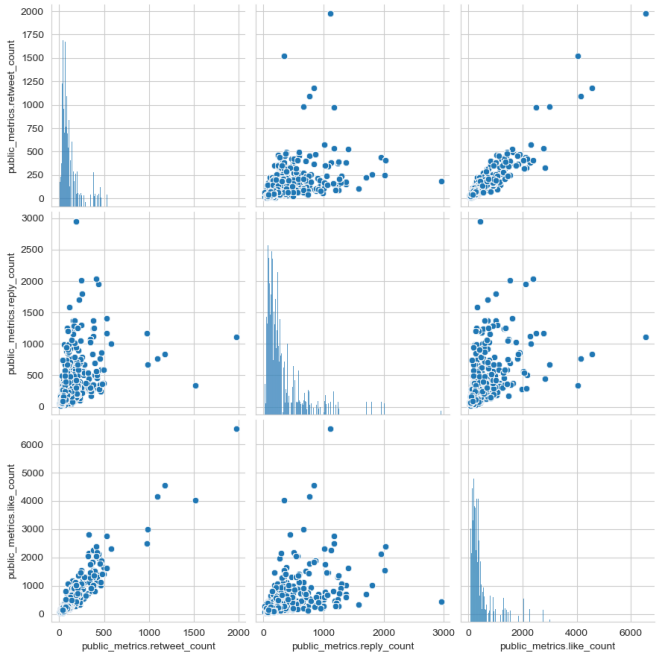
As shown above the most highly rated target groups are *tot6plus*, *50plus_jr* and *v_6plus_jr*. It also shows that Saturday and Sunday are the days where the groups watch the shows the most and so rate it higher.

Introduction to the machine learning models

Machine learning models are algorithms that can identify patterns/ make prediction on future data. Today, companies need to use predictive modelling to maximize profits. Banijay for example has given the project to analyse their current TV viewership metrics and present a plan to improve their data usage and help them understand what drives the shows popularity.

Machine learning models can be classified into supervised and unsupervised learning. The main difference is that supervised algorithms need labelled input and output training data. Unsupervised learning on the other hand can process unlabelled datasets. (Shin, 2022). A supervised machine learning algorithm was the best option to choose from in the research as the datasets were labelled.

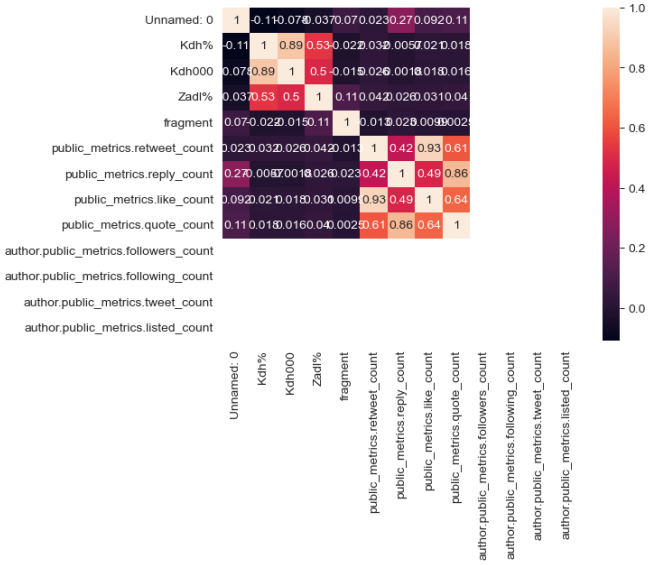
The research question that I came up with is if we could predict the ratings (Kdh000) of a show based on the twitter metrics. Twitter metrics include the number of retweets, the number of replies and the amount of likes a tweet has. This means that the ratings (Kdh000) are the target variable, and the twitter metrics are the features (number of retweets, the number of replies and the amount of likes a tweet has).



(10)

Before starting with the research, I had to clean the data, prepare the data and explore the data.

To clean and prepare the data I checked if there were any missing values and I checked the datatypes of all the columns. To prepare the data I checked the relationships between the numerical variables for the twitter metrics using a pair plot(10).



I realised that the more likes on the tweet the less likes and reply. To check if there were any other correlations between columns , I visualized the correlation between the columns using a heatmap(11).

(11)

Now That I had finished cleaning, preparing and exploring the data, I could start with making the machine learning model.

To make the machine learning model, I defined my X (twitter metrics: retweets, replies, likes) and my Y (ratings: Kdh000). Then I split my data into training and test data, I made 80% of my data training data and the other 20% test data. I then built a linear regression class.

Now that I have split my data and built a linear regression class, I built my machine learning model using the linear regression class and the training data.

Machine learning model

Now that I had built my model, I wanted to check how accurate my model was. To evaluate the performance of the model, I used coefficient determination (R^2 score). In the coefficient determination, the closer the value/ score is the better the model. I first checked the score of the training data.

```
Train R^2:  0.0010843639436458608
```

The coefficient of determination on the training date is just greater than 0.001. The model is not good, it would be better if the score was closer to 1. Then I checked the score of the model on the test data.

```
Test R^2:  0.00086007601664539 47
```

The coefficient of determination of the training set is lesser than 0.001. The model score for the model on the training data is also not high. The performance of the model on the training data is close to the performance of the test data. This means that there is no overfitting problem.

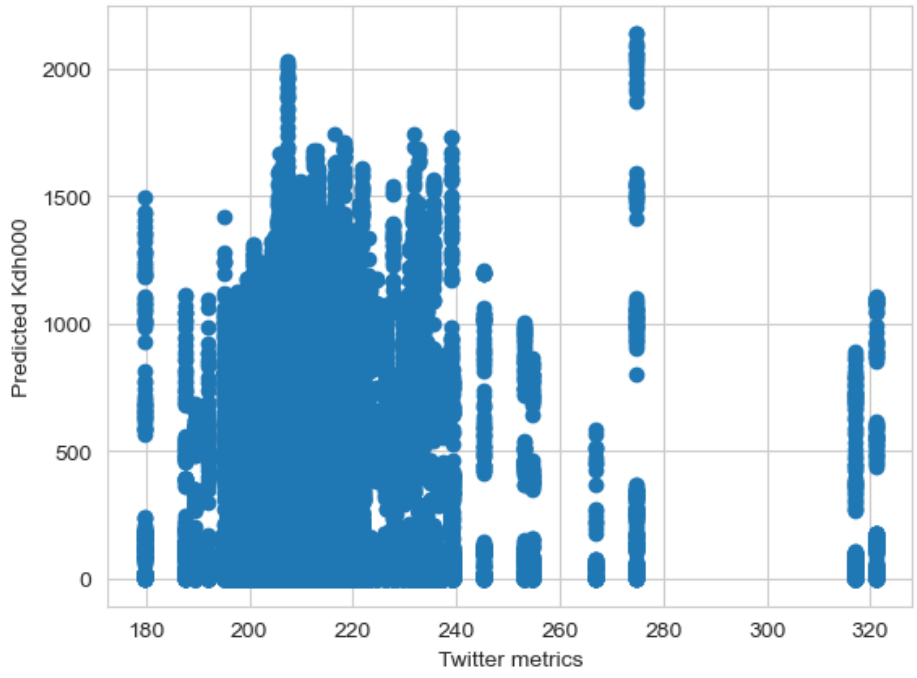
As the coefficient determination score was low for bot the training model and the test model, I decided to look into another metric: the mean squared error.

```
Mean squarer error: 90201.59390516437
```

Using the mean squared error, I can calculate the standard deviation by calculating the square root of the mean squared error.

```
Standard deviation: 300.335801903743
```

To visualize the predicted model in a better way, I visualized the model using a scatter plot(12).



(12)

This scatterplot of the model reflects the low R^2 score of the training model and the test model, as the points are not scattered in a very linear manner.

Ethics

The ethical strategies of a company result in a company having competitive advantage ahead of its competitors. So it is important to look at the norms and ethical principles of a company and the people working in it. To look at the ethical aspects of a company, you look at the three elements that are vital for ethical organizational capacity: Ethical company, Ethical process and tools and finally Ethical people (employees and clients).

Ethical company

The company can be considered ethical if it follows policies (like the GDPR) and behaviour towards their employees and their shareholders. Banijay for example is ethical according to this, as their code of conduct states that the employee's welfare is their priority (Banijay)(13). The code of conduct in Banijay also states that Banijay want you to feel safe and comfortable in your work environment and the way Banijay work with other companies (Banijay). Banijay is also very transparent on the type of data that the company will collect, who will have access to the data, why the company collects the data (Banijay, 2022).

One part of the company that might be seen as unethical is the diversity of the team. Especially that is portrayed on the website, in the website the majority of people portrayed are male while most of the employees during the Banijay visit were female (Banijay, 2022). This can be improved by simply portraying more female employees in the website. Another part of the diversity is that most of the employee's are Dutch, a way to fix this is to employ more foreign employees. Foreign people have another way of thinking and knowledge so by employing them you bring in a unique insight on problems.

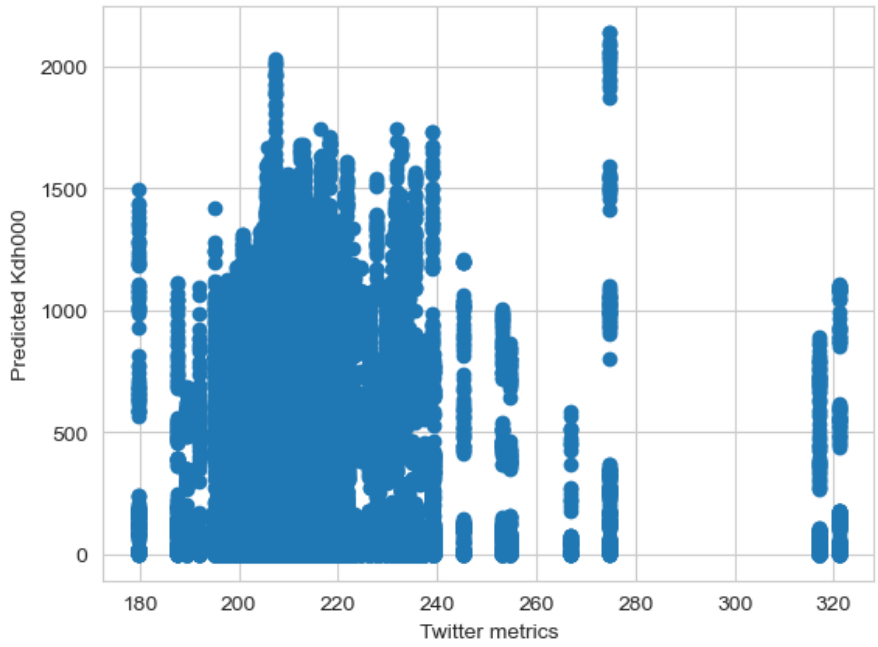
Ethical process and tools

A process can be classified as ethical if ethics is incorporated in the development process. The process and tools that Banijay use is also ethical as they collect their data is ethical. They follow the GDPR rules when collecting data as they ask consent to the data that they collect. For example, if you look at the data from twitter, all of the data that is collected is publicly available information that Banijay uses to learn what their audience thinks about their shows. This way Banijay follows the GDPR rules to make sure the data is collected and used in an appropriate manner. Banijay is also very ethical in the sense of how they protect the data because Banijay has internal policies to make sure the data is not lost or misused (Banijay, 2022).

Ethical people

This part is about the ethical behaviour of the clients and professionals towards their customers, society, owners, environment, suppliers. Ethical people also focus on the awareness of ethics, , morals and acting as responsible professionals. The people in Banijay are ethical as well as they have a an ethical behaviour towards their customers, clients, owners, society, environment and suppliers. This is proven again by their code of conduct where it says that Banijay values the employees welfare (13),(14). The code of conduct in Banijay also states that Banijay want you to feel safe and comfortable in your work environment and the way Banijay work with other companies (Banijay).

Discussion



As shown above in the scatter plot, the predicted ratings are not very linear. This means that as the twitter metrics increases the ratings(Kdh000) decreases. As for the linear regression machine learning model, we can predict the ratings of a show based on the twitter metrics, but the accuracy of this model is low.

The average ratings of all the shows comes at 201.80 , the most highly rated target groups are *tot6plus*, *50plus_jr* and *v_6plus_jr*. This makes sense as the people that are older than 6 watch television more, so they have higher ratings.

The ratings are most high during March and April and the most high during Saturday's and Sundays. This is because most people have more free time during the weekends, so the relax by watching shows.

According to the models and visualizations, the advice I would give Banijay is that if they want more ratings you should air the shows during the weekends. The shows will have even more ratings if the show is aired during March and April. The content of the show will also make a difference as the

target group with the highest ratings is *tot6plus,* so Banijay have to air more shows that have their interests.

Conclusion

To come back to my research question, could we predict the ratings (Kdh000) of a show based on the twitter metrics? The answer is yes, but as mentioned previously in the report my model is not very accurate as it has a very low coefficient determination score. For future analysis the coefficient determination score will have to be improved for a more accurate model. This coefficient determination score (R2 score) can be improved by adding more independent variables. This is because the R2 score thinks that every independent variable in the model helps explain the dependant variables and its variations. The model can also be improved by tuning its hyperparameter with validation sets.

If we look at the ethical strategies, principles and norms of Banijay, Banijay is very ethical. Banijay, has rules and policies in place for the way they behave, the way they work (collecting and protecting the data).The only thing that could be improved in ethical side is the diversity of the company. Banijay could employ more foreign employees to provide a new perspective and insights on problems. Overall Banijay is a very ethical company.

Reference

Banijay. (2022, November 7). *Our People*. Retrieved from Banijay: https://www.banijay.com/our-people/

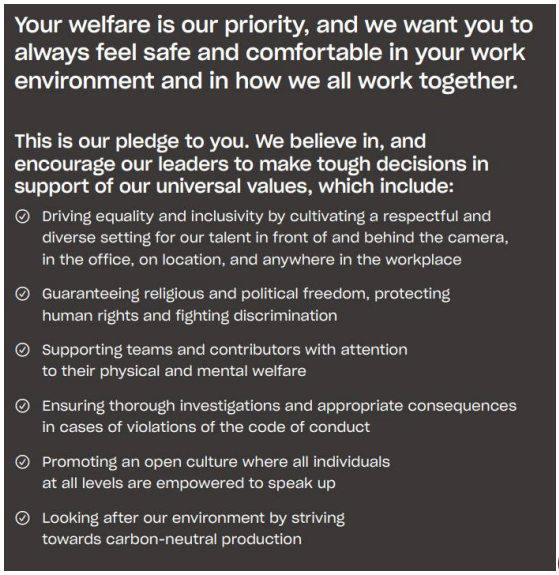
Banijay. (2022, December 7). *Our Story*. Retrieved from Banijay: https://www.banijay.com/our-story/

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

Banijay. (n.d.). *Banijay Code of Conduct.* Retrieved from Banijay: https://www.banijay.com/wp-content/uploads/2022/12/Banijay-Code-of-Conduct-v15.pdf

BUAS. (2022). *Block B - Data Understanding and Preparation*. Retrieved from ADAI: https://adsai.buas.nl/Year1/BlockB/

Shin, T. (. (2022, November 10). *Machine learning models explained in 6 minutes*. Retrieved from Medium: https://towardsdatascience.com/all-machine-learning-models-explained-in-6-minutes-9fe30ff6776a



(13)

(14)

When handling personal data, we follow these privacy principles:

- **Lawfulness, Fairness and Transparency:** We always process personal data in this way and inform individuals of this and their rights through a clear and detailed privacy policy.
- **Purpose Limitation:** We only collect personal data for specified, explicit and legitimate purposes and do not process it further in a manner incompatible with those purposes. We only process personal data for the purposes indicated in the privacy policy given to the individual.
- **Accuracy:** Personal data should be accurate and, where necessary, kept up to date. We delete or update incorrect or out-of-date data right away.
- **Storage Limitation:** We don't keep data in a way that someone could be identified from it for any longer than necessary for the purpose for which it was processed.
- **Data Minimisation:** We make sure that personal data is adequate, relevant and limited to what is necessary in relation to the purpose for which it is collected.
- **Integrity and Confidentiality:** We use appropriate technical and organisational measures to protect personal data against unauthorised or unlawful processing and accidental loss, destruction or damage.
- **Privacy by Design and by Default:** When developing products and services, we consider the protection of personal data from the design phase. Measures are implemented to ensure that, by default, only personal data that is necessary for the purpose of the processing is processed.
- **Accountability:** We are responsible for and we must all be able to demonstrate compliance with these principles.

If you're not sure what is permissible, ask your legal department. Data breaches can expose the Company to penalties and harm our reputation.

Banijay

Respecting intellectual property

We're home to some of the world's top scripted and unscripted brands and multi-platform titles.

Our IP mustn't be infringed or used and distributed without the right permission. Using "pirated" or illegally obtained intellectual property also isn't allowed: in fact, copying or using materials beyond applicable legal boundaries without the owner's consent is theft.

Keeping track of information

Any information has to be recorded and reported in a fair, timely, full and accurate way, following good business practices, applicable accounting standards and local laws. All documents, files, records and reports that you acquire or create while working for us are our property. Please only remove originals or copies from your office if it's to do with work (and return them when required).

Using communication tools

Your phone, e-mail, Internet and other communication facilities and appliances are our property for business purposes only. You can use these within reason, so long as it doesn't interfere with your work. We may access and review data held, following applicable legislation and best practices. Any evidence of wrongful use may lead to disciplinary action. Always be mindful that email, voicemail messages and internet usage are potentially subject to interception and may be disclosed during litigation or an investigation.

All social media is public and should therefore be used in a way that's transparent, truthful and sensible, and that won't cloud Banijay's reputation.

(16)