Is there a correlation between the success of a TV show and its Twitter interactions?

Dániel Elek - 223838

DISCOVER YOUR WORLD

Applied Data Science and AI

Monday 19 June 2023

Breda University OF APPLIED SCIENCES

Abstract

This project aims to help Banijay Benelux, a television and production company, by analysing their data and finding a correlation between Twitter interactions and a given show's success. The goal is to provide the company with insights into which shows are successful to make decisions that can yield more profit. I used Python and various libraries such as Pandas, Numpy, Scikit-learn, Matplotlib, and Power BI to manipulate and visualise the data. The project used three datasets provided by Banijay, including information about the content produced, ratings data, and Twitter interactions data. The data was cleaned, combined, and explored through visualisations. Machine learning models were used to find relationships in the data. The project's potential impact is increasing profits for Banijay Benelux through more informed decision-making.

1 Introduction

The use of data science and machine learning in the television and production industry has become increasingly important in recent years. Companies, such as Banijay Benelux, are constantly looking for ways to maximize profits and stay competitive in the market. One way to achieve this goal is through the smart usage of data, which can provide valuable insights into the characteristics and patterns of the industry. In this business case, I aimed to help Banijay Benelux by analysing their data and trying to find a correlation between Twitter metrics and a show's success. The reason being, if I can find a correlation between the two, Banijay will be better equipped to decide which shows are successful, meaning they could better judge which ones to cancel or renew. This would ultimately lead to an increase in profits for the company. To conduct this analysis, I used Python with libraries such as Pandas, Numpy, Scikit-learn, Matplotlib, and Power BI. These powerful tools allowed me to manipulate and visualize the data, providing us with the insights we need to make informed decisions. Overall, this project has the potential to have a significant impact on the company. By providing Banijay Benelux with a better understanding of the correlation between Twitter metrics and a show's success, I can help the company make more profitable decisions, leading to long-term growth and success.

2 The datasets used

I used three datasets provided by Banijay for this project. The first dataset contained information about the content that was produced by the company, including information such as the title, the channel it was broadcast on, and the duration of the show. The second dataset contained ratings data for said content, including information about the viewership and demographics of the audience. The third dataset contained data from Twitter, including tweets about the content. All the datasets were in Dutch.

2.1 Preparing the data

To prepare the data for analysis, I first made two new columns in the content dataset, date_time_start and date_time_end, by combining already existing columns. I then split the id column to create a column for fragments. For the ratings dataset, I only had to combine the date and the time columns into a date_time column. After that, I cleaned both datasets by deleting unnamed columns and sorted them by time. Then I combined the two datasets into one. With the Twitter dataset, I deleted referenced tweets and then combined this dataset with the previous one, creating one final dataset that included all the information from the three sources.

2.2 Exploring the data

Exploring this data is also a crucial step in understanding the characteristics and patterns present in it. One way to do this is through the use of visualisations. By creating visual representations, trends, patterns, and outliers can be easily identified which are not necessarily obvious from just looking at the raw data. In this project, I conducted an exploration of the datasets I was given. I began by examining the basic statistics such as the number of rows and columns, the data types of each column and the distribution of values. I also checked for missing or duplicate data and any other issues that could impact our analysis. After the initial exploration, I created various visualizations to better understand the patterns and relationships within the data. I used histograms, bar charts, and scatter plots to investigate the distribution of values and the relationships between different variables. These visualizations provided a deeper understanding of the data and highlighted areas of interest for further analysis.

3 The machine learning models

Since I was working with huge amounts of data, I used machine learning to help me. To make sure I would get the best possible results, I decided to use two different types of machine learning models: random forest regressor and linear regressor. My reason for this decision is that not every model can capture the correlations in every type of dataset. To make sure that I do not get misled by a model that sees a non-existent relationship in a dataset, the best is always to try out different models and compare their results, both numerically and with visualisations. The precise correlation I was looking for was between “Kdh000” (ratings) and the twitter metrics of likes, retweets, and replies.

Random Forest Regressor is an ensemble method that uses multiple decision trees to make predictions, it works by creating multiple decision trees, each with a random subset of the data, and then averaging the predictions of all the trees to make a final prediction. I used the .score attribute which calculates the determination coefficient of the prediction, also called R squared score. This ranges between 0 and 1 where the closer the score is to 1, the better it is.

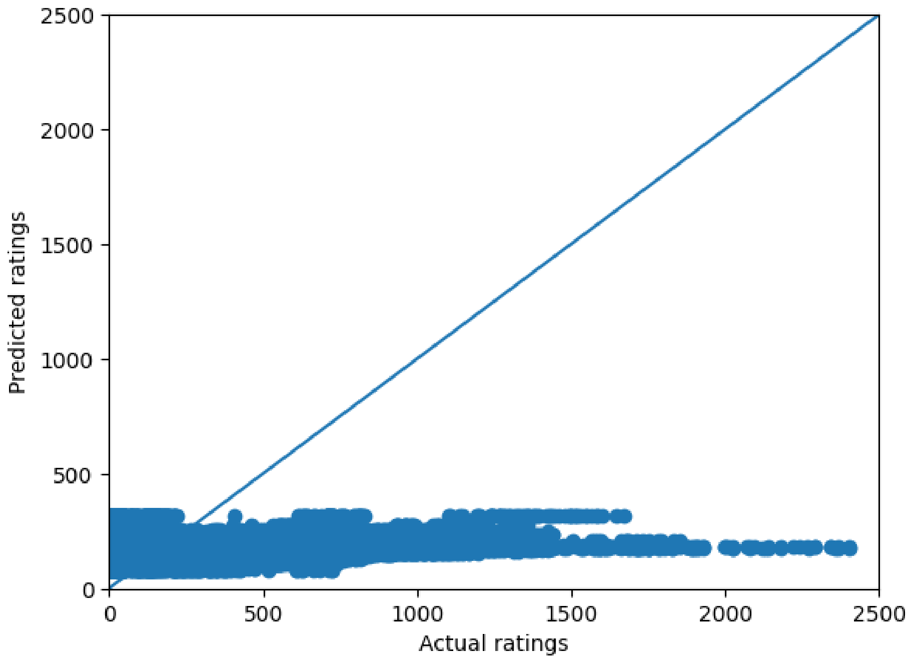
Linear Regression is a powerful and straightforward algorithm that aims to identify the linear relationship between the input features and the output variable. It does this by creating a line of best fit that minimizes the difference between the predicted values and the actual values. This is achieved by finding the line that minimizes the sum of the squared errors. Here, again, the .score attribute calculates the R-squared, which is the correlation coefficient squared.

4 Results

I personally went into this research thinking that there is a correlation, since I am also a Twitter user and I have seen successful shows being discussed there, generating tweets with hundreds of thousands or millions of likes, retweets, and comments. To my surprise it turns out there is none, or at maximum a very weak one.

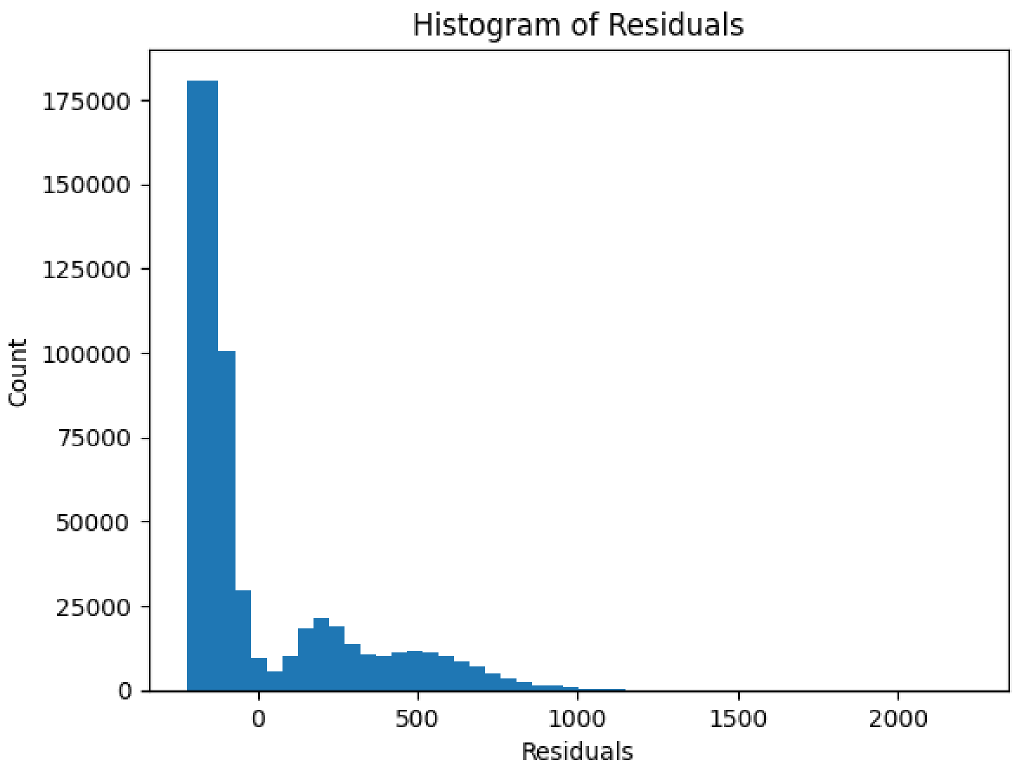
4.1 Random Forest Regressor

With the Random Forest Regressor model, the r2 score I got was 0.0005439969613698636, which is very far from 1, meaning the model does not fit the data at all. The visualizations also supported this.



This scatter plot visualizes the relationship between the predicted values (y_pred) and the actual values (y_test) of the target variable. Each point on the scatter plot represents one observation from the test set, where the x-coordinate is the actual value, and the y-coordinate is the predicted value. If the model is a good fit for the data, the points on the scatter plot would mostly be concentrated along a diagonal line from the bottom left to the top right corner. This would indicate that the predicted values are mostly close to the actual values. If the model would not be a good fit for the data, the points would be scattered all over the plot, with no clear pattern. This would mean that the predicted values are far from the actual values, and that the model is not capturing a relationship between the predictors and the target variable.

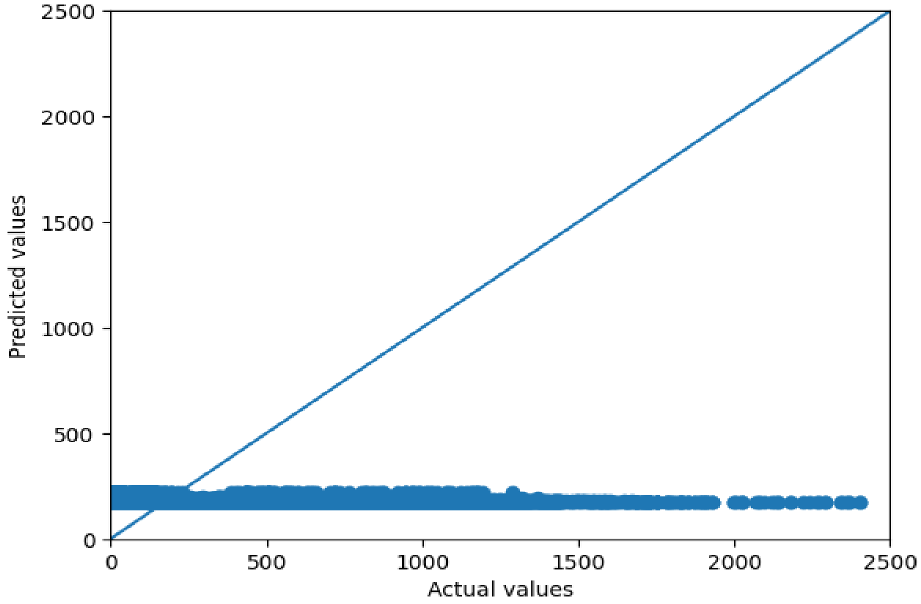
What we can see here is that all of the predicted values are on the bottom, meaning that regardless of what the actual values are, the model predicts every rating to be under around 400, this obviously suggests that the model is not seeing any relationships.



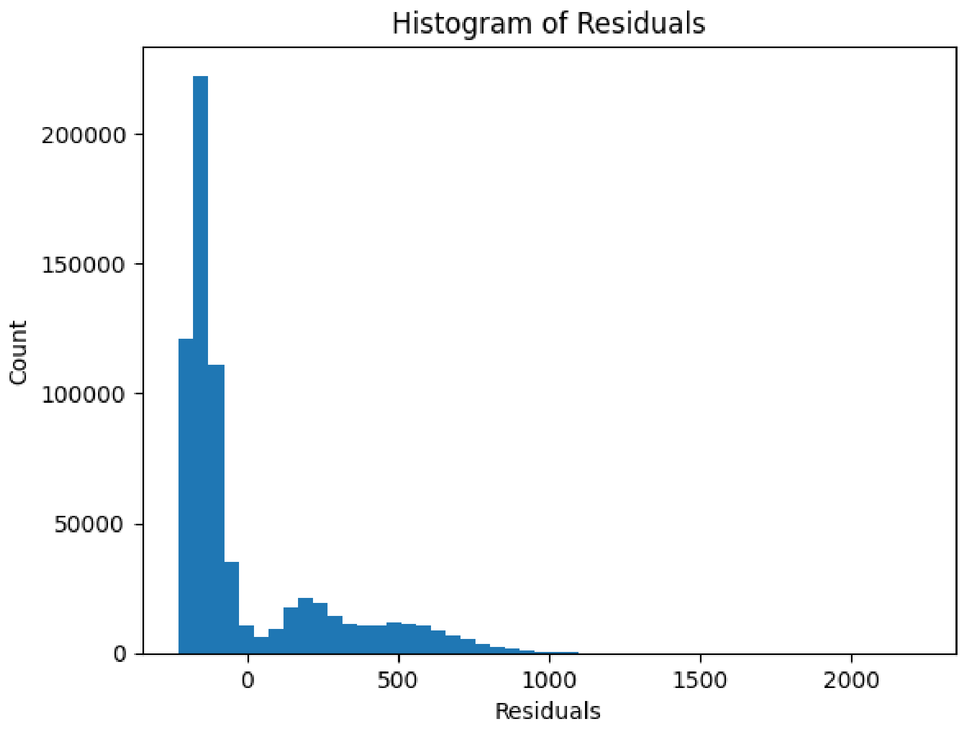
On this histogram the X-axis represents the range of residuals, and the Y-axis represents how many data points fall into each range. The residuals are the differences between the actual values of Y (y_test) and the predicted values of Y (y_pred) from the model. We would expect a good model to have residuals that are randomly distributed around zero. Here we can see that there are many negative residuals very close to zero, and still a lot of positive ones, but these are more spread out. The negative residuals near zero indicate that the model tends to underestimate the target variable. With the positive residuals the greater spread indicates that the model has larger errors when predicting higher values and that these errors vary in magnitude. This is also exactly what we could see on the scatter plot.

4.2 Linear Regression

My Linear Regression model supported this, as it got a score of 4.6079440049751064e-05 which can be written as 0.000046079440049751064. This is so close to zero that it is no better than a random guess.



The scatter plot this time shows something very similar to that of the previous model’s, all of the predicted values are concentrated near the bottom of the graph, meaning that regardless of the real values, it always predicts the ratings to be between around 200 and 300.



The histogram of residuals is also very similar to the random forest regressor's. With lots of negative values close to zero, and still many positive values spread out until around 1100, we can come to the same conclusions as we did with the ensemble model we used previously.

Overall, it was useful to try two different models, as the fact that they both arrived at similar conclusions proves that it is not a specific model's fault that the predictions are not good, just there is simply no correlation between the ratings and the Twitter metrics.

Having found these results, I was left to wonder why there is no correlation as it seemed so obvious to me. I think there are multiple reasons or even combinations of them that could cause this. Maybe people that are more vocal on Twitter are simply a small minority and most people do not actually go to Twitter to talk about their favourite shows. Maybe only Dutch people are the ones who do not talk that much about shows they like on social media. Maybe there is a correlation, just simply not in op1's case. This is a little harder to believe though as the models were very clear about there being no correlation, this would have to be a huge coincidence. Maybe there is something completely different behind all this that I did not even think of. Still, to find an answer to this question more research would have to be done, and that is not the aim of this project.

5 Ethics

5.1 The three elements of an ethical organisational capacity

Professional integrity and accountability of an ethical statistician are essential to maintain trust in statistical analysis and reporting. This includes being honest, adhering to ethical guidelines, ensuring data confidentiality, and properly acknowledging the work of others.

An ethical statistician has a responsibility to other statisticians and statistics practitioners to create a collaborative and supportive environment. This involves sharing knowledge, engaging in open discussions, and providing constructive feedback to promote the advancement of statistical practice.

When working with research team colleagues, an ethical statistician is responsible for supporting integrity and transparency. They need to actively contribute to the team's objectives, communicate statistical findings effectively, and ensure that data and analysis are handled ethically and with proper consideration for the research team's goals.

These obviously apply to the statisticians at Banijay Benelux.

5.2 Privacy notice

I decided to conduct a comprehensive examination of Banijay's privacy notice, as I believe it is important to examine a company's ethics before working for them.

Banijay's privacy notice is a clear document that outlines the company's data collection and usage practices well. It starts by stating the purpose of the notice, which is to inform employees of the types of personal information that the company collects and how it is used. It explains the company's commitment to transparency and compliance with the General Data Protection Regulation (GDPR). A strength of this privacy notice is that it clearly defines who the "data controller" is - Banijay UK - and provides contact information for employees to address any data concerns.

It also provides a list of the types of personal information that the company collects, including name, address, contact details, qualifications, skills, and employment history, which is all data that is completely ethical to collect by an employer. The notice also explains how the company collects this information, including through CVs, forms, and interviews, and notes that in some cases, the company may collect information from third parties, such as references from former employers.

Overall, this privacy notice is rather ethical in my opinion which is also helped by the fact that it is easy to understand. It provides a clear and comprehensive overview of the company's data collection and usage practices and gives employees the information they need to understand how their personal information is being used, which is extremely important.

5.3 Sustainability

The whole of the Banijay group seems to be aware of their importance in making a more sustainable world. Their stance is: "Sustainability and the impact of climate change is central to our industry right now as it should be. Banijay, like its peers, is proactive in reducing its carbon footprint and overall impact on the environment, both corporately and on-set. "(*Sustainability Archives – James T. - We Are Banijay*, n.d.)

They bring up examples of them promoting a more sustainable lifestyle in their shows, as well as donating leftover food. (*Sustainability Archives - Banijay Group - We Are Banijay*, n.d.)

This is obviously the right stance to have on such an important topic, so I find it completely ethical.

5.4 Diversity

I researched Banijay Benelux's stance on diversity, and they strongly stand by it:" Stereotyping, diversity, and inclusion. We are highly aware that the content we produce has a major impact on public opinion and perception. It is important to us to connect with the public and pay attention to stereotyping, diversity, and inclusion. Both in our content and as an employer. That is why we invest - in collaboration with various partners including the Coalitie Imaging in the Media - in a respectful and inclusive setting where equality and diversity are stimulated." (Banijay Benelux, 2022)

Keeping this quote in mind, I found it interesting that during the trip of our programme group to Banijay Benelux' headquarters in Amsterdam, we only met women there throughout the whole day, except for the sports department, where there were only men. I do understand that women are often less represented in positions of more power or more income, but at the same time it would be the best to include both women and men everywhere. Banijay should hire in a way that this problem will be fixed, making sure that both genders are represented equally everywhere.

We cannot lose sight of our goal of achieving equality, so I feel like Banijay could work on this part. My part of this ethics question is also important. I made sure to handle the data with regard to privacy in accordance with GDPR, it did not contain any sensitive or personal information.

6 Conclusion

The project at hand was a large challenge for me with significant potential to increase both profit and customer satisfaction for Banijay. The goal of this endeavour was to find a correlation between Twitter metrics and the success of a TV show, with the hope that this information could be used to inform future decisions about which shows to produce and promote. Although the project ultimately did not yield the desired results in terms of finding a strong correlation between Twitter metrics and a given show’s success, it still provided valuable insights for Banijay. For example, the project showed that since there is not a strong correlation between Twitter metrics and the success of a show, negative comments on Twitter about a show may not be indicative of poor performance. This information can be useful for the company when deciding how much weight to give to social media metrics in their decision-making process. In addition to its practical value, the project also had a positive impact on my personal development. It was a valuable opportunity to work with real company data and make a meaningful contribution to the company's goals, while also improving my skills in Python and Power BI.

Overall, while the project may not have produced the results that were initially hoped for, it still provided valuable insights and experiences that can inform future decisions and projects. It also helped in understanding not to worry too much about negative comments on Twitter as it may not be representative of the majority, but it's important to consider each case separately.

7 References

- *sustainability Archives - James T. - We are Banijay*. (n.d.). Banijay Group - We Are Banijay. https://www.banijay.com/blog/tag/sustainability/
- *sustainability Archives - Banijay Group - We are Banijay*. (n.d.). Banijay Group - We Are Banijay. https://www.banijay.com/blog/tag/sustainability/
- Banijay Benelux. (2022). *Ons verhaal - Banijay Benelux*. https://banijaybenelux.com/ons-verhaal/

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