Technical report

Motivation

Due to recent Spotify backlash resulting in a lot of artists leaving, music enjoyers might want to find out if the type of artists they listen to are more likely to boycott the platform due to ethical concerns. Our work could spread awareness of the harmful business model of Spotify and encourage people to stop supporting problematic platforms.

Data collection

We manually collected a list of artists who left Spotify due to boycotting. We used different sources such as news outlets and Instagram. In total, we collected 121 artists who boycotted Spotify. We stored them inside a .csv file listing the artists' names and dates when they announced their departure. Next, we needed to find their corresponding Spotify IDs and append them to the .csv as another column. Since Spotify doesn't provide a reliable tool to fetch an artist's Spotify ID based on their name, we had to manually obtain the IDs by searching for the artists on Spotify and extracting the IDs from their Spotify page URLs.

After collecting all artists' Spotify IDs, we used the Spotify API to fetch artist information as JSON files. The data was retrieved in multiple batches, and we eventually combined and saved all the artists' information into a single JSON file. After having all the artists who boycotted Spotify and their information saved, we proceeded to collect the negative dataset needed for supervised learning i.e. the artists who have not boycotted Spotify. Since obtaining a random sample of Spotify artists is not straightforward, we opted to use a Kaggle dataset that claims to be a large random sample of Spotify artists. The dataset contains 15 000 artists from 2024 including information related to them. Since there is no reliable way to verify whether artists in the dataset have not boycotted Spotify or not, we simply assumed that none of them had boycotted the platform.

Preprocessing

Because a lot of genres were missing from the Spotify API response, we manually collected missing genres using Bandcamp as a primary source. The API response included meaningful fields like followers amount, artist name, popularity (0 to 100 with 100 being the most popular compared to all Spotify artists), a list of genres attributed to the artist, spotify_id

and some urls of images that are shown in spotify. Some fields like external_urls, href and uri were removed from all entries since they were either always null or not useful to our scope. Imputation was not applied here since the amount of missing values was small and only affected the genres, and we manually entered these to the empty genres lists. With a bit of elbow grease and Pandas we joined the manually crafted .csv (fields were name, date-of-leaving, spotify-id) with the Spotify API response .json with the spotify ids of both datasets. We made the Kaggle dataset follow the same structure as our positive dataset by removing unnecessary columns and renaming the others.

The genres variable is a categorical variable with approximately 90 different categories (genres) in the positive dataset. To reduce the number of categories, we selected ten genres that appeared most often in the positive dataset and labeled all other genres as “other”. However, this led to a result where a majority of genres in the negative and positive dataset were identified as “other”. To fix this issue, we converted genre names to more generalized ones to be able to include more genres while keeping the number of variables reasonable, since there are too many obscure genres in the Kaggle dataset.

Lastly, we performed preprocessing before actual machine learning. We removed fields that are not used in the analysis (Name, Id, Date, Image) and added column for the target variable that will either take value 0 or 1 depending on whether an artist has boycotted Spotify or not. Since the genres were categorical, we converted them to numeric format using one-hot encoding.

Exploratory data analysis (EDA)

We explored the visual and numerical amount of different genres compared to one another using Pandas and WordCloud. Genres were also sorted and top 10 genres listed as Spotify-leavers were printed. A bar chart was also created to visualize the difference in the sorted list. Mean and median followers and popularity were printed as numerical values for exploration purposes.

Visualizations

Our main two visualization efforts were the wordcloud of genres and bar charts of genres and artist popularities. To follow the principles of graphical integrity we cleared labeling on the bar charts and toned down design variation as we first thought of more vibrant charts like interactive pie charts.

Learning task

Our task is to predict on given characteristics (follower, popularity, genres) of an artist how likely it is to boycott Spotify at some point. We train our model in a supervised setting, so we have both artists that have boycotted and who have not. Target variable is binary: target = 1 if likely to boycott at some point and target = 0 if not.

Learning approach

Our first choice of model was logistic regression. We chose this model because of its simplicity and interpretability of results. However, the small size of the positive dataset quickly found out to be problematic. Our small sample size of positives (n≈100) restrictd the number of independent variables we can include in the logistic regression model. Although our final preprocessed data could have fitted with the restrictions of logistic regression, we instead decided to use random forest.

Aside from the small size of the positive sample, a major issue with the data was the class imbalance. The ratio of artists who have boycotted Spotify to those who have not was highly skewed. To tackle this issue, we decided to use synthetic minority oversampling technique (SMOTE). We will next describe in more detail how actual learning actually took place and what the results were.

First and foremost, we randomly shuffled our data. Then we split the data to 80% training set, 10% validation set and 10% test set. We used stratified split in each step to ensure the preserving of the original class ratios. After obtaining the splits, we utilized SMOTE to handle class imbalances in the training set. Then we started the actual training part. We performed simple hyperparameter tuning by training Random Forest classifiers with different numbers of trees (100, 200, 300) and selecting the model that achieved the highest F1 score on the minority class in the validation set. The best number of trees turned out to be 100 with a f1 score of 0.57 in the validation set. The confusion matrix of the validation set showed that the model predicted the majority class (artists not left) perfectly, correctly labeling all 1503 employees and making no errors. In contrast, the minority class (artists who left) was predicted less accurately: out of 10 employees who actually left, only 4 were correctly identified, while 6 were missed.

The results on the test set were even more disappointing. While the model still correctly predicted the majority class (artists not left) in most cases, it struggled significantly with the

minority class (artists who left). Out of 11 artists who actually left, only 1 was correctly identified, and 10 were not. This led to a very low F1 score of 0.13 for the minority class, highlighting that the model fails to generalize well to unseen data when it comes to predicting artists who left, despite the overall accuracy appearing high due to the imbalance in class distribution.

Communication of results

Communication of results was a presentation in a classroom using a slideshow. The slideshow presentation was crafted and trained with the help of the course material's Patrick Dang's elevator pitch tutorial. We succeeded in keeping the presentation within the 3-minute mark. Our contact details were given at the end of the slideshow if the target audience would have wanted to learn more afterwards.

Data privacy and ethical considerations

We acknowledge that people not willing to leave an unethical platform might be harassed by those who did. We acknowledge that the alternative platforms we propose might become unethical as well in the future. Although the project is not going to be worked on in the near future, the GitHub issues pages stay open for any missed data privacy and ethical considerations.

Added value

Two main added values: Awareness and proof of concept. More people are now aware of the ethical and usability issues of Spotify. They can now apply this knowledge to other platforms as well. As a proof of concept we have trained a model with some appropriate data to check whether or not an artist is leaving Spotify. For further development this work can be used as an example, just like any other proof of concept.

Mini-project canvas reflection

Our motivation stayed the same throughout the project and we believe our initial intention is still valid. Data collection also worked out as planned in the mini-project canvas with the exception of using X/Twitter. Data collection worked out as initially planned with a few exceptions to manually scraping some missing information like genres to certain artists.If we could have foreseen the workload needed for this project and course we would have probably just been happy with the GitHub repository and the slides as final deliverables

rather than aiming for a static web page using interactive visualizations. Since the learning task and learning approach gave us some leeway in the project's mini-project phase went according to plan with some minor adjustment. Many of the actual implementations were figured on the go and are previously described in this technical report. Communication of results indeed did not end up being a website but a slideshow and our GitHub repository. The README is written in a blog post format and can be easily converted to a such if given enough time to fiddle around with static site generator tech. Data privacy and ethical considerations did not change throughout the project. Added value stayed the same although the magnitude of the added value can vary depending on the perspective. The possible future steps could involve figuring out solutions to the unbalanced data sets of positives and negatives and deploying a non-technical interface for an end-user to try out the model we have trained.

According to the lecturer, the essence of the course is to ask the right questions and communicate properly. We believe we have succeeded in these goals on a needed level.