MASTER DEGREE IN CYBERSECURITY
UNIVERSITY OF PADUA

User Privacy on Spotify: Predicting Personal Data from Music Preferences

Master Thesis Defense
15th december 2022

Candidate: **Jiancheng Ye**
Supervisor : **Mauro Conti**
Co-supervisor : **Pier Paolo Tricomi**

SPRITZ
Security & Privacy
Research Group

1222 · 2022
800 ANNI

Università
degli Studi
di Padova

DIPARTIMENTO
MATEMATICA

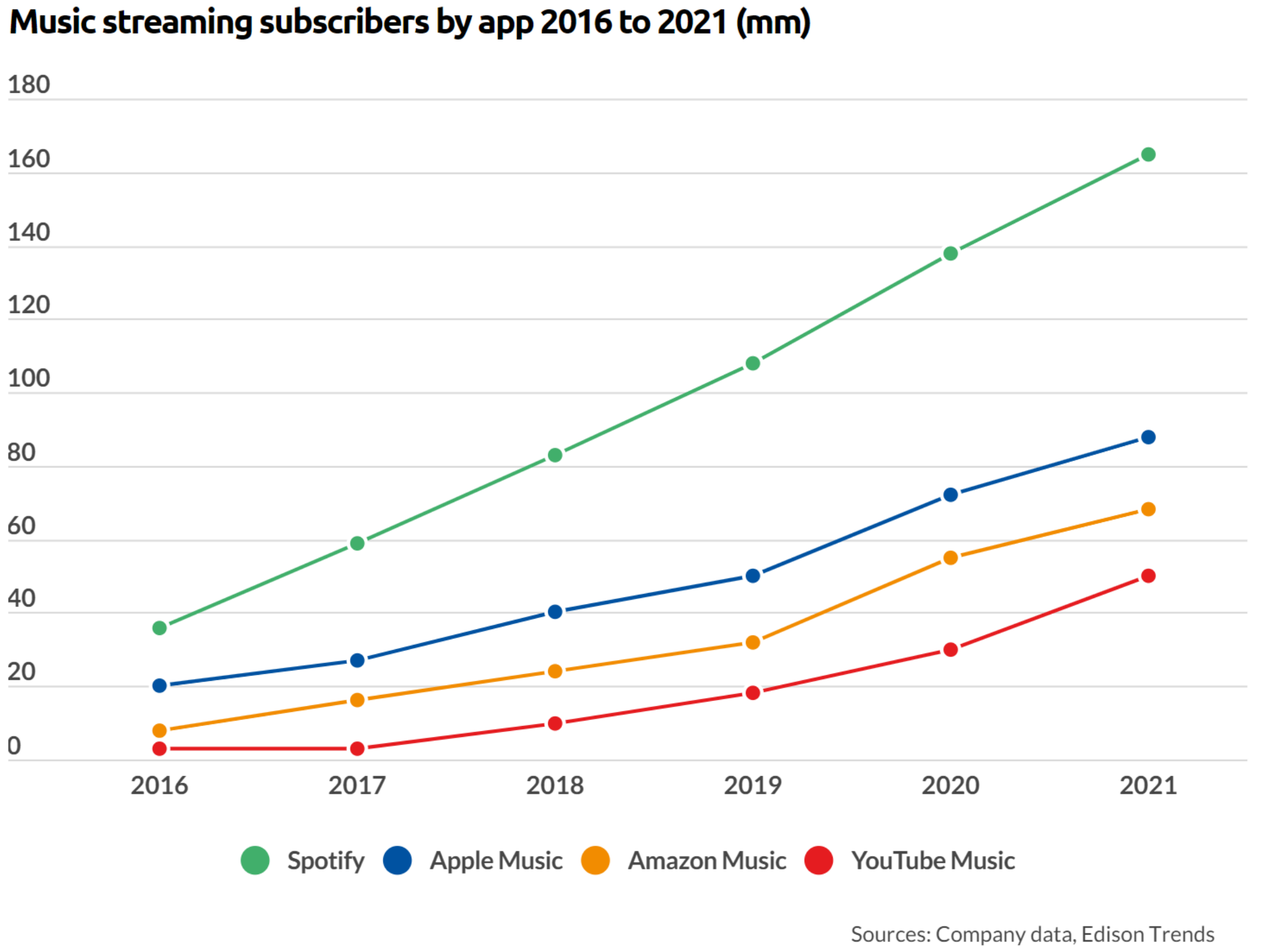
DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

Outline

1. Introduction and motivations
2. Our case study: proposed attack on Spotify
3. Correlations
4. Predictive models and results
5. Discussion and Conclusions

The Impact of Music

- Music is part of our daily life!
- Reveals a great deal about who we are
- People share their playlists and songs on the music platform



Sources: Company data, Edison Trends

The Impact of Music

Is it a good idea to share music interests or will it compromise our privacy?

There is no purposeless data

Everything can be used to infer private information

Many providers let us listen and share the music

Music Streaming Service Panorama

Table 3.1: Music streaming service panorama table

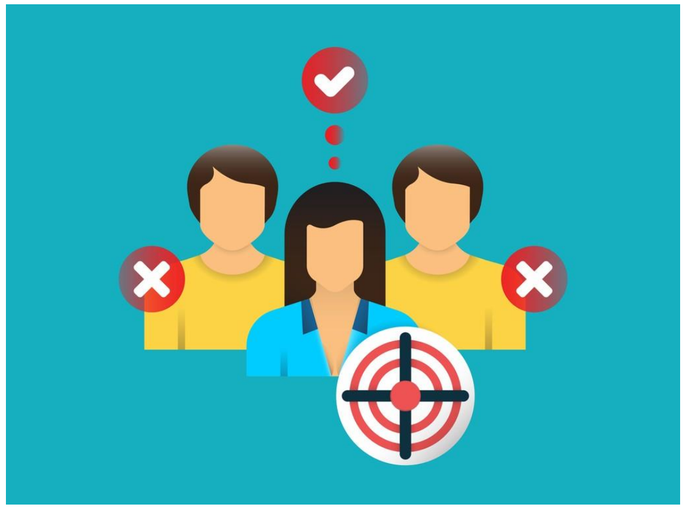
	Release Year	*Monthly Users*	*Paying Users*	*Public playlist Visible*	*Public S song Visible*	*Retrievable Via API*	*Website*
Spotify	2011	422 M	195 M	yes	yes	yes	open.spotify.com
Apple music	2015	88 M	88 M	yes	yes	yes	apple.com/it/apple-music
Tidal	2015	5 M	5 M	yes	yes	yes	tidal.com
Youtube Music	2015	30 M	30 M	yes	yes	yes	music.youtube.com

What If this information
is used in the wrong way?

Threats

We may be profiled for targeted advertisement, by surveillance agencies, or in general, become potential victims of malicious activities:

- Cyber-bullying
- Cyber-harassment
- Grooming



Motivations

Predicting Personal Data from Music Preferences on Spotify

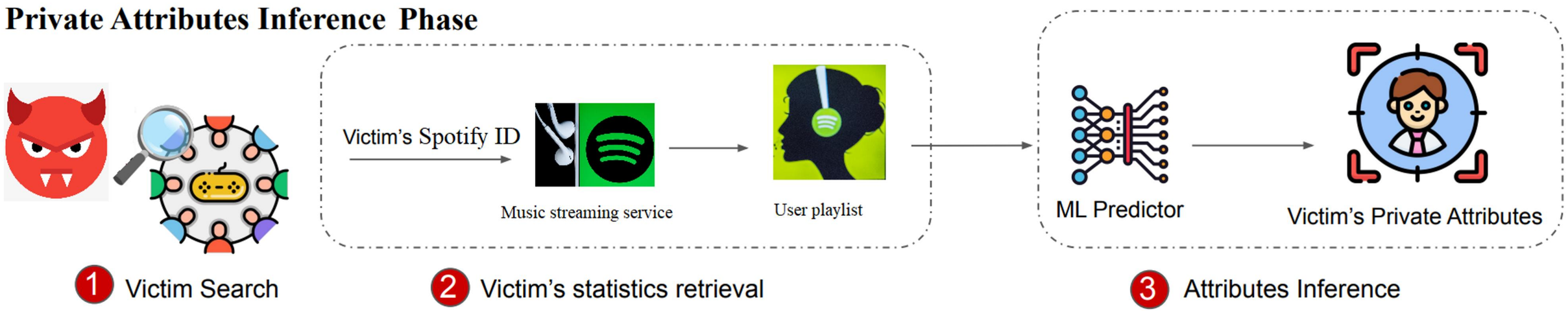
1. Exposing a potentially dangerous privacy threat in an environment rich of publicly available data

2. The success of this attack highlights a crucial privacy threat.

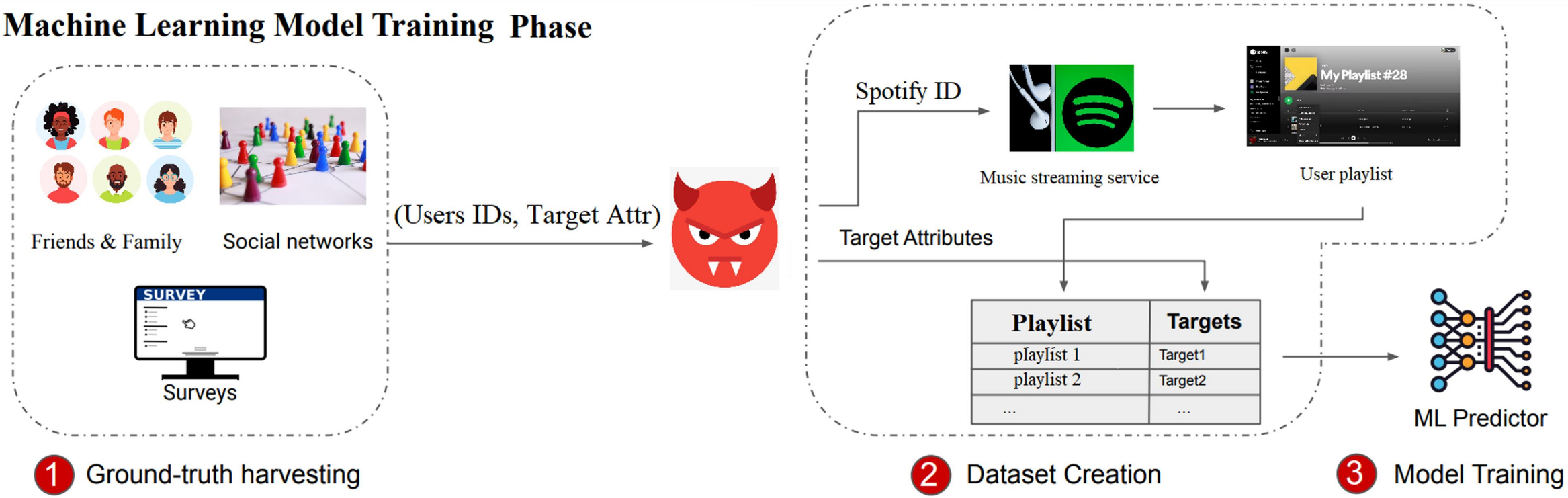
Outline

1. Introduction and motivations
2. **Our case study: proposed attack on Spotify**
3. Correlations
4. Predictive models and results
5. Discussion and Conclusions

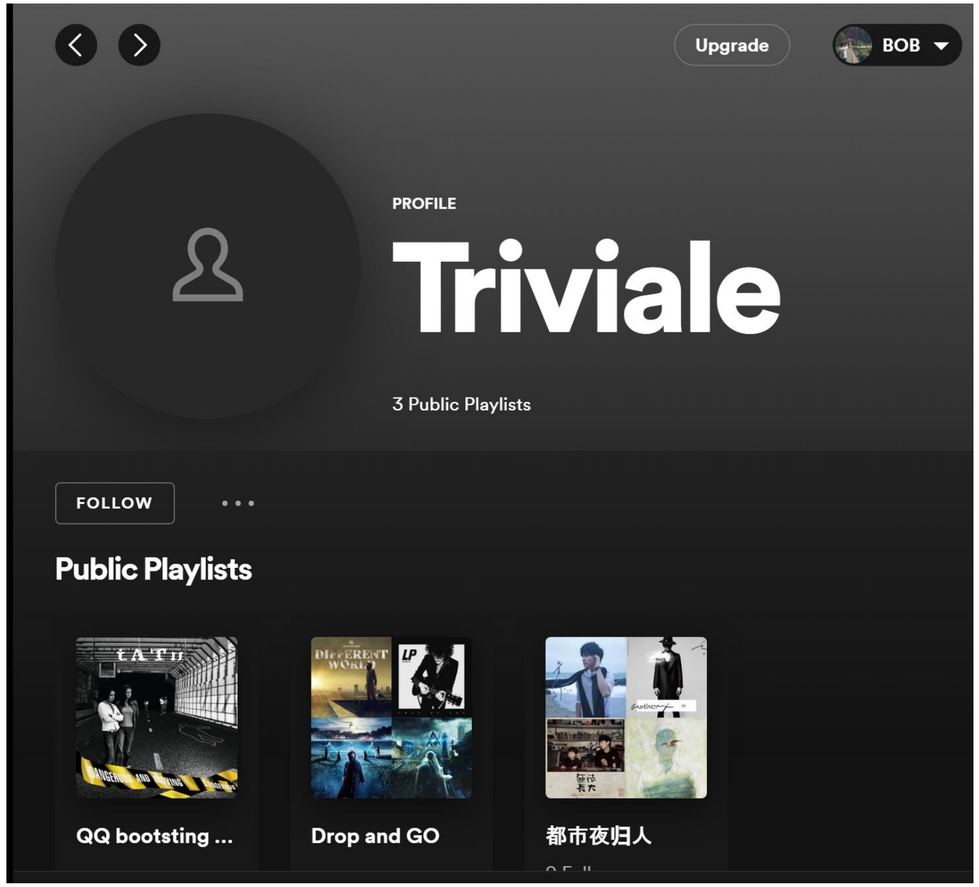
Proposed Attack

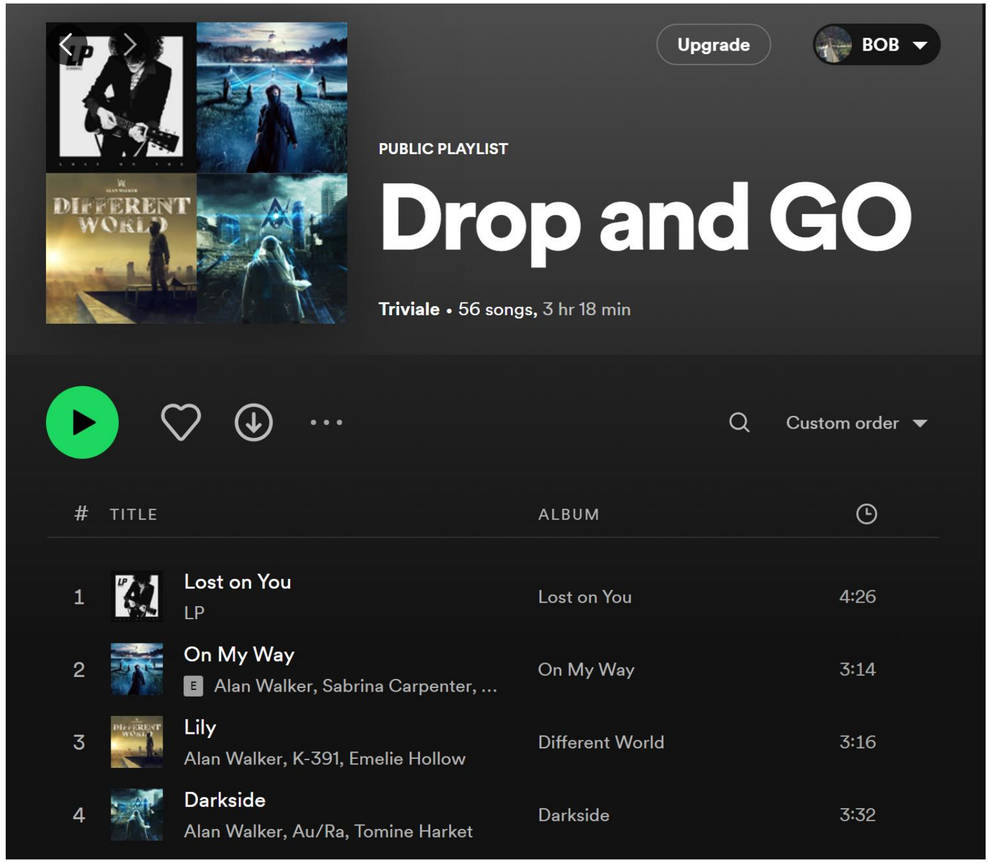


Proposed Attack



Our Case Study: Spotify





Data Collection

Private data and Spotify ID collected through an online *survey*

↓

Music preference data collected through *API Spotify*

1222·2022 800 ANNI

UNIVERSITÀ DEGLI STUDI DI PADOVA

SPRITZ SECURITY & PRIVACY RESEARCH GROUP

Spotify Survey

Dear participant,

Welcome to this short survey!
We are a research group from the University of Padua, Italy, and we are conducting a study on Spotify.

This survey should take approximately **3-4 minutes** to complete. We ask you to provide information about yourself, your Spotify usage, and your personality.

We will store the data collected through this survey anonymously on our servers.

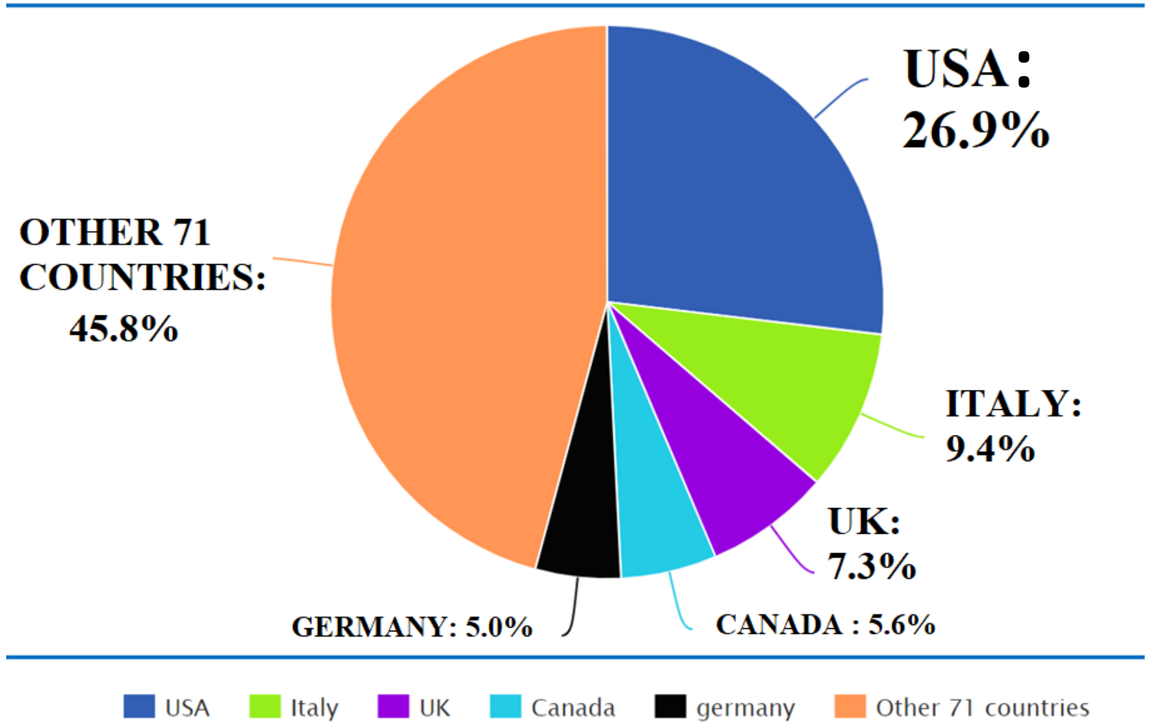
If you have any questions feel free to email jiancheng.ye@studenti.unipd.it

Please forward this link to anyone else who may be interested.

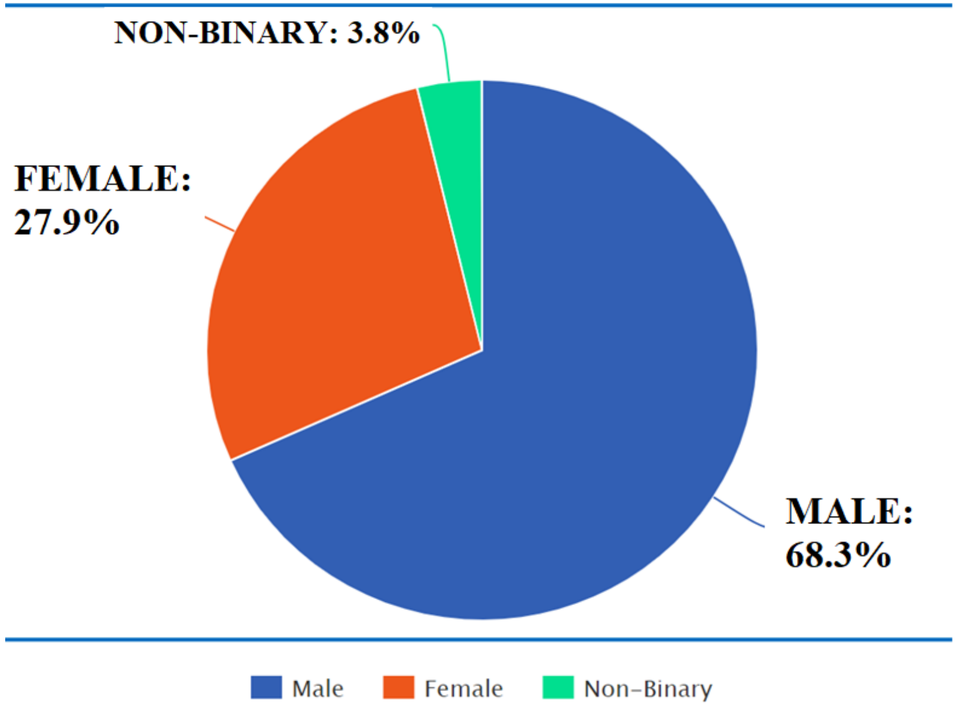
Thanks for your collaboration!
The SPRITZ Group

Survey Population

750 valid answers from users from different countries

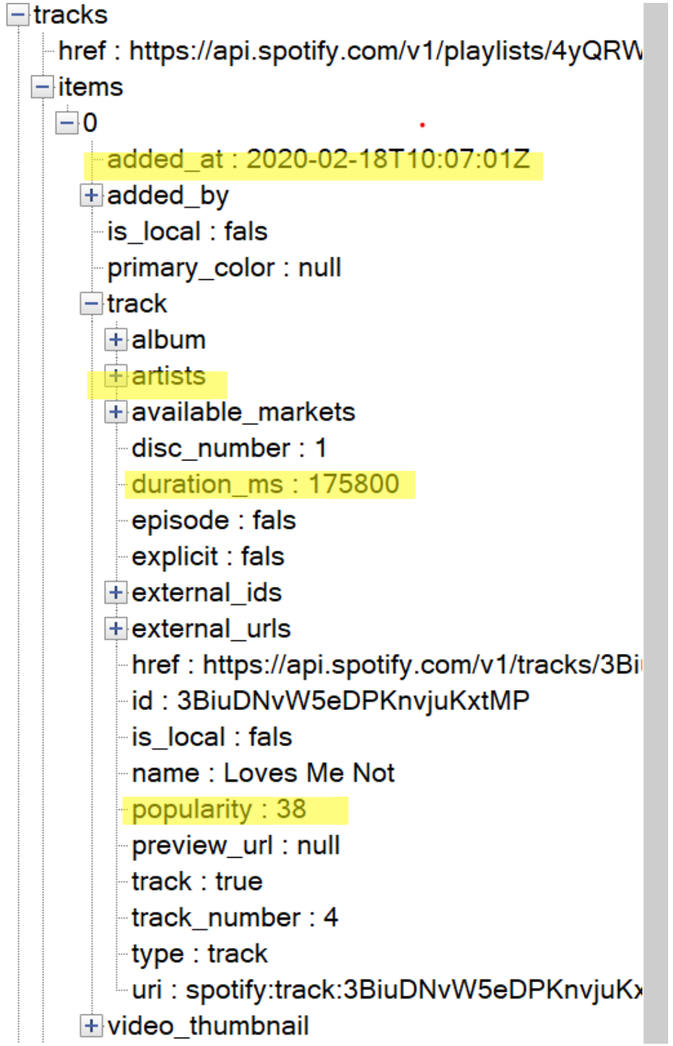


Gender



The age ranges from 13 to 55, with the majority between 14 and 36.

Spotfy Data: Song



Each track (song) in the JSON file is made by the following :

• Artist: contains information about the artists e.g. unique Spotify ID, names, and links to their Spotify pages.

• Time: indicates the date when the user added the song in the playlist;

• Duration: indicates the time duration of the song expressed in milliseconds

• Popularity: a number between 0-100 which indicates the popularity of the song in the Spotify market

Playlist Features

Content features

– **prevalent song language**

– average years publication

– number of songs

– number of followers

– ratio male artist

– ratio female artist

Audio features

features relative to the audio statistical given by directly by Spotify API:

– **danceability:** describes the suitability of a track for dancing in terms of its tempo and rhythm stability. A value of 0.0 is least danceable and 1.0 is most danceable;

– instrumentalness: indicates whether a track contains no vocals;

– tempo: which indicates the overall estimated tempo of a track in beats per minute (BPM);

Datasets

Analysis done at {
- Playlist level (8777×87)
- Song level (402.999×65)

In total, data about 8777 playlists and 402,999 songs

Main Target Features

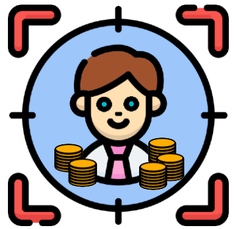
Main targets:

1. Gender
2. Age
3. Occupation
4. Economic
5. Personality → { Extroversion, Agreeableness, Conscientiousness, Neuroticism, Openness }

Gender, Age, and Personality

➜to target precise categories of user

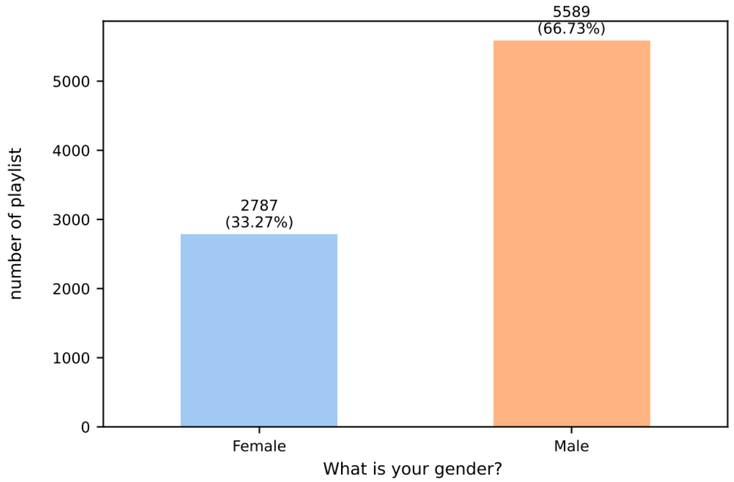
Occupation and Economic



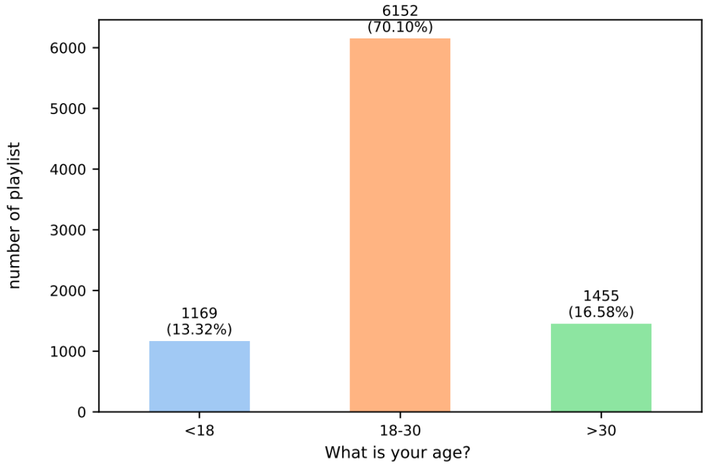
➜to find best victims to advertisement or phishing

Main Target Features Analysis

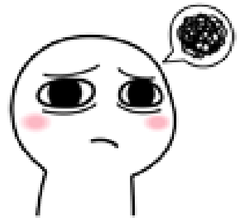
Distribution of **gender** at playlist level



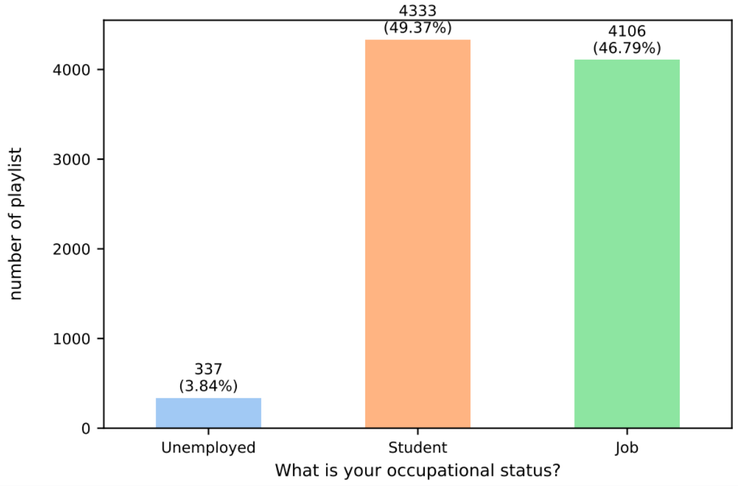
Distribution of **age** at playlist level



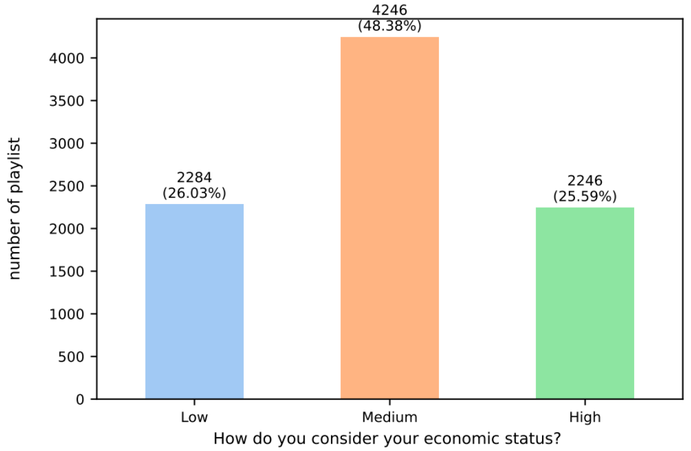
PROBLEM:
Class imbalance



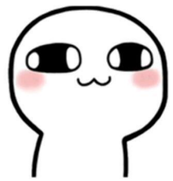
Distribution of **occupation** at playlist level



Distribution of **economic** at playlist level



SOLUTION:
Data oversampling



Outline

1. Introduction and motivations
2. Our case study: proposed attack on Spotify
3. **Correlations**
4. Predictive models and results
5. Discussion and Conclusions

Correlation Metrics

Employed correlation metrics

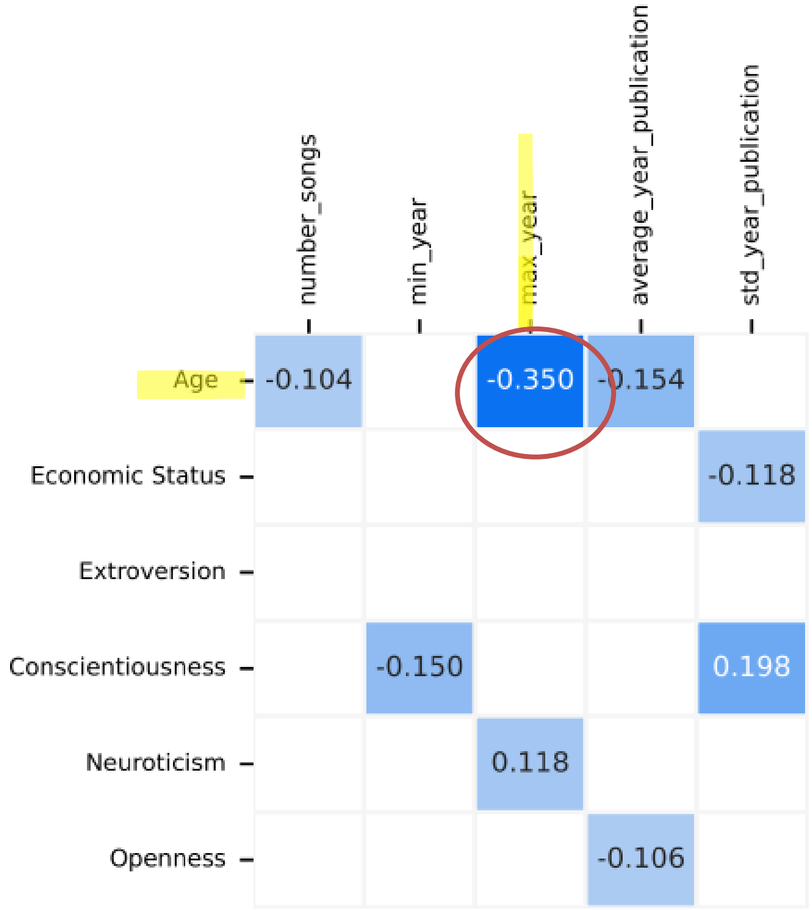
- **Numerical** vs **numerical (I.e. age with number of song)** :

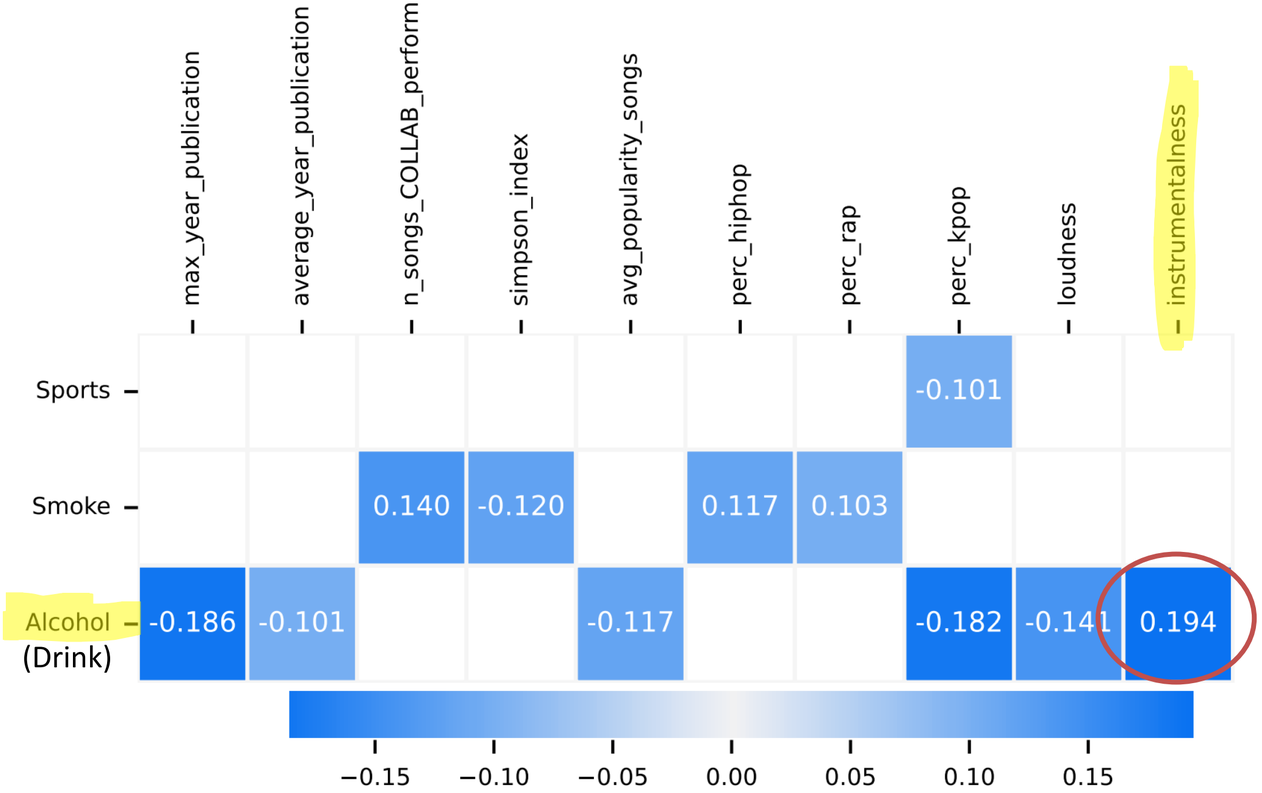
  Spearman's **ρ**$\in [-1, +\ 1]$

- Categorical **nominal target (I.e. gender, occupation)** :

  Logistic regression and count how many features are correlated significance level
  - gender: 37
  - occupation: 39

Spearman’s ρ


number_songs
min_year
max_year
average_year_publication
std_year_publication
Age
Economic Status
Extroversion
Conscientiousness
Neuroticism
Openness
-0.104
-0.350
-0.154
-0.118
-0.150
0.198
0.118
-0.106


max_year_publication
average_year_publication
n_songs_COLLAB_perform
simpson_index
avg_popularity_songs
perc_hiphop
perc_rap
perc_kpop
loudness
instrumentalness
Sports
Smoke
Alcohol
(Drink)
-0.101
0.140
-0.120
0.117
0.103
-0.186
-0.101
-0.117
-0.182
-0.141
0.194
−0.15
−0.10
−0.05
0.00
0.05
0.10
0.15

Outline

1. Introduction and motivations
2. Our case study: proposed attack on Spotify
3. Correlations
4. **Predictive models and results**
5. Discussion and Conclusions

Predictive Models

Experimental settings:

- 5 models tested: Logistic Regression, SVM, Ridged Classifier, Decision Tree and Random Forest.
- Balance
    - limiter parameter
    - User division
    - Oversampling
- 2 approaches : Using all features and using best features
- Macro F1-score as performance metric
- Experiement repeated 5 times

Results: Song Level

target feature	dummy stratified		svm		logistic regression		decision tree		ridge classifier		random forest	
	mean	std dev	mean	std dev	mean	std dev	mean	std dev	mean	std dev	mean	std dev
age	26.90%	0.0028	35.78%	0.039	32.12%	0.0096	37.16%	0.0097	31.55%	0.010	37.72%	0.007
gender	48.36%	0.0086	57.61%	0.0309	63.51%	0.0195	62.61%	0.0244	62.99%	0.0180	**63.88%**	0.02130
economic	31.83%	0.0055	30.67%	0.0124	31.78%	0.018	33.57%	0.0097	31.66%	0.0183	33.86%	0.0107
occupation	28.47%	0.0052	32.17%	0.0259	34.48%	0.019	32.86%	0.0117	38.04%	0.0180	33.86%	0.0107
marital	28.47%	0.0068	32.14%	0.0110	33.14%	0.011	33.00%	0.007	31.87%	0.0138	33.39%	0.0107
sport	47.78%	0.0101	45.86%	0.0570	52.09%	0.012	48.99%	0.0236	50.31%	0.0134	51.49%	0.0145
smoke	45.11%	0.0141	50.23%	0.0272	53.52%	0.015	51.91%	0.009	50.53%	0.0196	52.47%	0.0233
drink	49.69%	0.0032	50.42%	0.0256	53.42%	0.012	51.47%	0.030	53.28%	0.0146	53.39%	0.0222
country	21.30%	0.0052	25.02%	0.0052	23.74%	0.0052	22.42%	0.0198	23.17%	0.002	25.44%	0.002
livewith	41.95%	0.007	45.02%	0.037	48.02%	0.0126	47.30%	0.0168	43.39%	0.009	48.50%	0.005
agreeableness	29.91%	0.0135	30.23%	0.014	32.77%	0.027	33.50%	0.0038	31.99%	0.0247	33.74%	0.010
extroversion	31.97%	0.003	29.51%	0.008	**36.22%**	0.010	32.12%	0.0277	34.86%	0.009	35.18%	0.012
consciousness	30.27%	0.009	31.70%	0.023	34.90%	0.010	33.77%	0.0147	33.63%	0.022	34.77%	0.0055
neuroticism	32.77%	0.002	30.82%	0.018	35.75%	0.014	35.37%	0.006	35.58%	0.0153	36.93%	0.0184
openness	29.35%	0.002	29.69%	0.024	31.89%	0.006	35.14%	0.0317	31.67%	0.0086	35.86%	0.0206

Target = Increasement less than 10% vs dummy

Target = Increasement more than 10% vs dummy

Results: Playlist Level

Target feature	best limiter	features used	dummy	svm	LR	DT	RI	RF
Age	8	all features	28.2%	38.3%	39.9%	41.8%	38.8%	43,5%
Economic	12	all features	29.3%	36.3%	35.4%	30.5%	35.5%	32.3%
Gender	4	best features	47.6%	59.9%	68,1%	61,1%	68,2%	71,7%
Occupation	6	best features	30.77%	40.3%	32.2%	34.1%	31.9%	33.0%
Country	6	all features	17.7%	25.4%	26.4%	21.0%	23.4%	30.4%
Sport	6	best features	45.4%	54.7%	46.2%	49.3%	45.5%	46.8%
Smoke	6	best features	45.1%	52.0%	55.2%	55.7%	52.2%	60.2%
Drink	4	best features	53.0%	54.9%	59.3%	54.8%	59.8%	62.8%
Marital	8	all features	27.4%	36.3%	41.0%	41.7%	40.3%	45.1%
Livewith	8	all features	43.4%	51.9%	49.8%	48.3%	49.7%	54.5%
Agreeableness	8	best features	28.5%	31.1%	32.9%	30.7%	33.2%	34.7%
Conscientiousness	6	best features	33.4%	37.3%	35.9%	33.5%	33.6%	37.9%
Extroversion	8	best features	30.4%	30.5%	35.9%	35.3%	35.4%	37.8%
Neuroticism	10	best features	32.1%	34.0%	41.1%	35.2%	40.4%	41.0%
Openness	10	best features	30.9%	35.9%	32.4%	33.7%	32.6%	36.5%

Discussion

- **Applicability to other music streaming service:** Assure to find music preference features that correlate with private information (i.e. music genres, favorite artists, audio analysis, etc.).

- **Possible countermeasures**: enhance privacy policies

- **Limitations**:
  - Actually tested only on Spotify
  - Number of participants (750 participants)
  - Lack of literature (in particular for security of music streaming services)

Conclusions & Future Research

Recap:

- Promising results but need to be improved
- Can inspire and will be likely become a wider research area
- Applicability to other music streaming service could be a real threat someday

Future research should probably focus on:

- Improving models
- Using more playlists or songs for prediction
- Testing on more music platforms

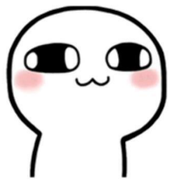
Thank you for your attention!

Q&A time

Backup A: Survey

Divided into 3 sections:

- **General Information**, where private data was gathered
- **Spotify use experience Information**, to assess user experiences with the Spotify
  + attentions and coherence checks
- **Personality Questions**, big five personality test in 10 points



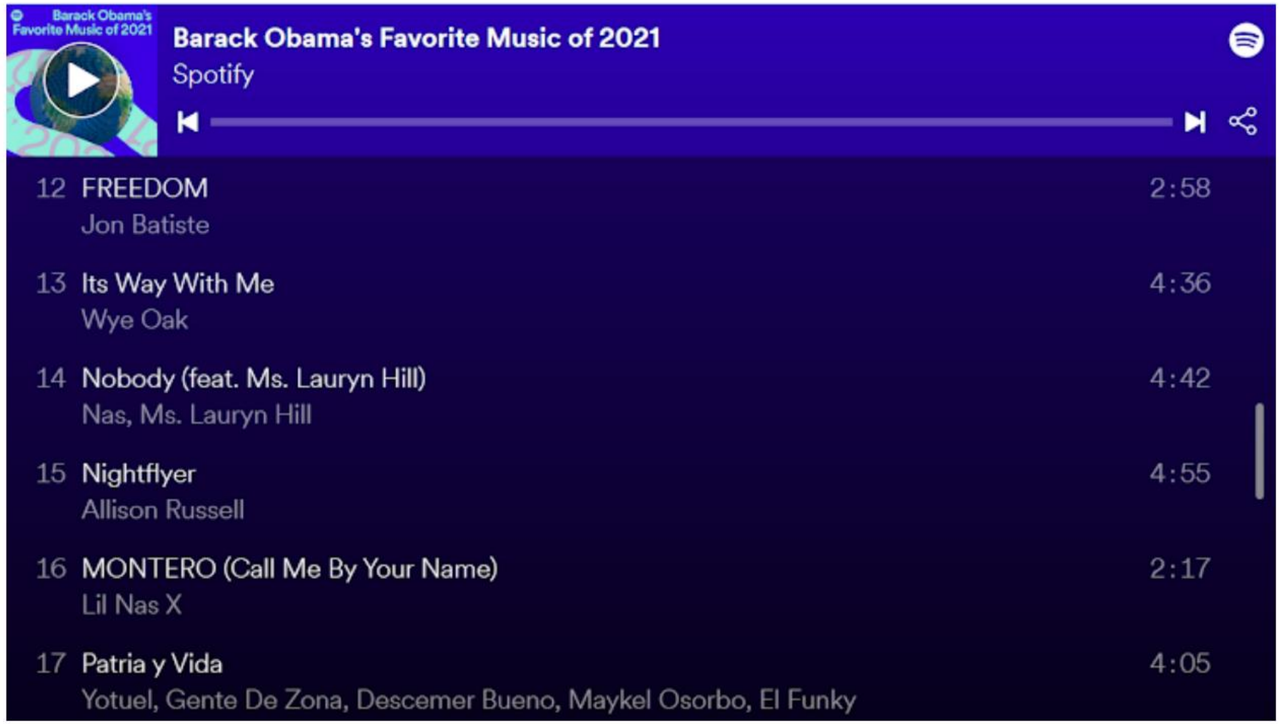
Are people willing to answer surveys providing private information? YES

Backup A: Survey

Example of question for attention check:

Given the image BELOW, which of the following songs is preferred by Barack Obama?

- Without Me - Eminem
- Freedom - Jon Batiste
- My heart will go on - Céline Dion
- Oskar Schuster - Les Sablons



Backup A: Survey

Personality questions:

24. I see myself as someone who...

... is reserved: *

	1	2	3	4	5	
Disagree Strongly	○	○	○	○	○	Agree Strongly

... is generally trusting: *

	1	2	3	4	5	
Disagree Strongly	○	○	○	○	○	Agree Strongly

... tends to be lazy: *

	1	2	3	4	5	
Disagree Strongly	○	○	○	○	○	Agree Strongly

... is relaxed, handles stress well: *

	1	2	3	4	5	
Disagree Strongly	○	○	○	○	○	Agree Strongly

... has few artistic interests: *

	1	2	3	4	5	
Disagree Strongly	○	○	○	○	○	Agree Strongly

... is outgoing, sociable: *

	1	2	3	4	5	
Disagree Strongly	○	○	○	○	○	Agree Strongly

... tends to find fault with others: *

	1	2	3	4	5	
Disagree Strongly	○	○	○	○	○	Agree Strongly

... does a thorough job: *

	1	2	3	4	5	
Disagree Strongly	○	○	○	○	○	Agree Strongly

... gets nervous easily: *

	1	2	3	4	5	
Disagree Strongly	○	○	○	○	○	Agree Strongly

...has an active imagination: *

	1	2	3	4	5	
Disagree Strongly	○	○	○	○	○	Agree Strongly

Backup B: Target Features

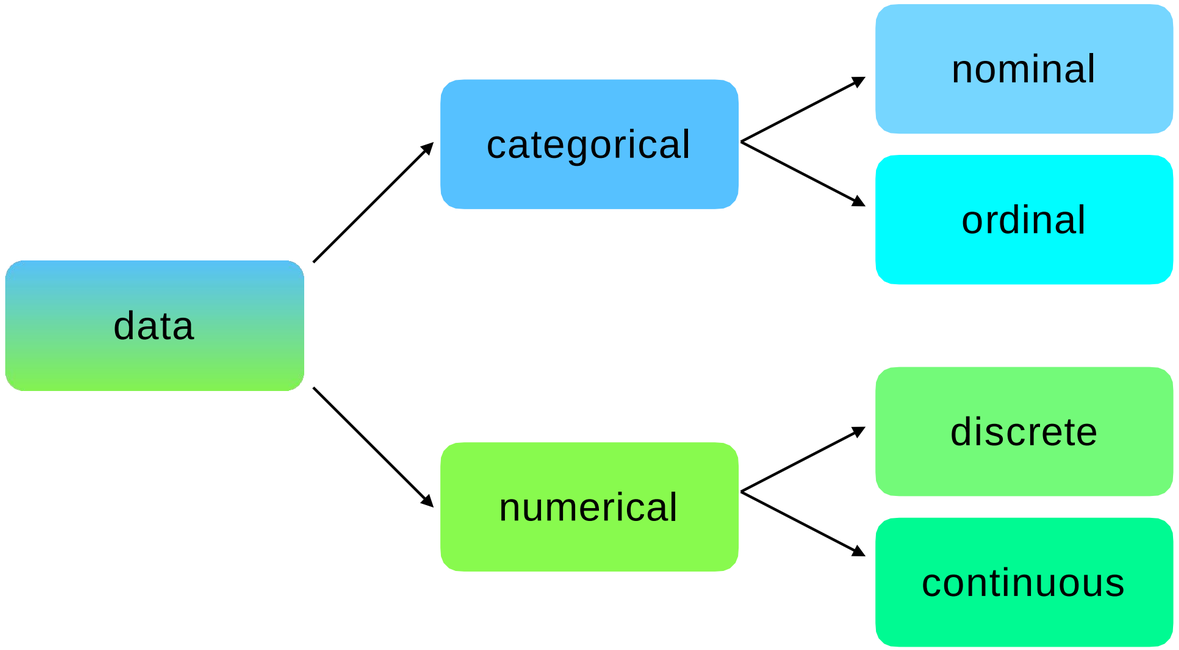
Main targets:

1. Gender
2. Age
3. Occupation
4. Economic
5. Personality

Secondary targets:

1. Marital state
2. Drink
3. Smoke
4. Live with
5. Sport
6. Country

Backup C: Data Types



‘yellow’, ‘blue’, ‘green’

‘low’ < ‘mid’ < ‘high’

1, 2, 3

1.14 , 2.033, 3.0

Backup D: Correlation Tables

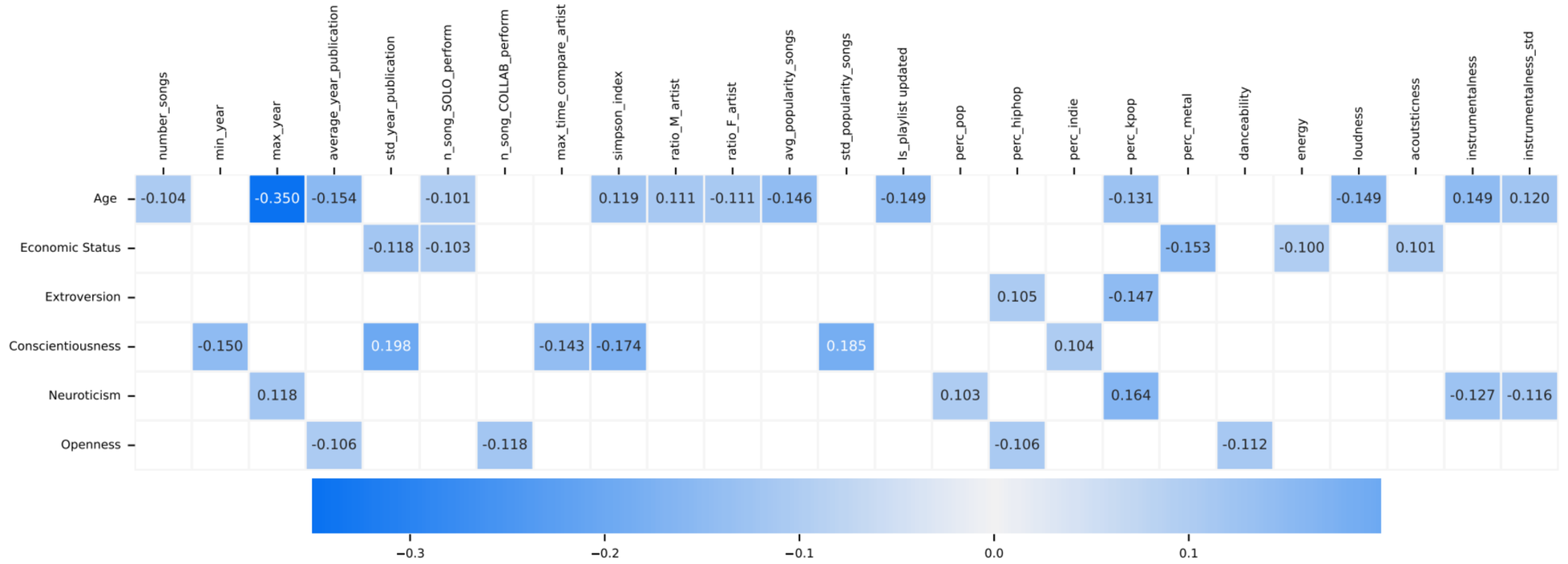


Figure 6.1: Significant (p-value ≤ 0.01) Spearman's correlation indices computed for age, economic, personality and dataset's numericial features at playlist level

Backup D: Correlation Tables

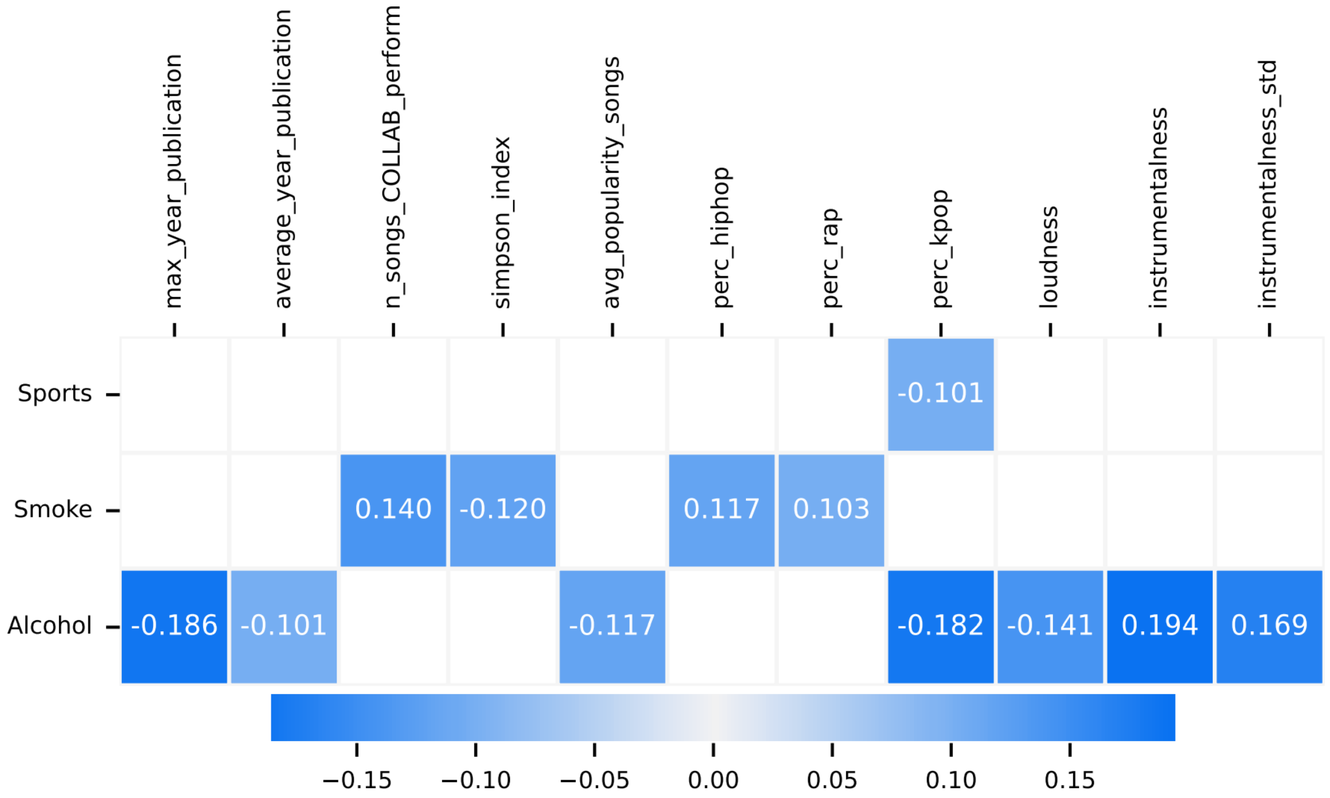


Figure 6.2: Significant (p-value ≤ 0.01) Spearman's correlation indices computed for sport, smoke, drink and dataset's numericial features at playlist level

Backup D: Correlation Tables

Table 6.1: Significant Correlations at different p-values for nominal target

Target	a<0.01	a<0.005	a<0.001
Marital status	31	29	26
Gender	37	36	33
Occupation	39	37	36
Livewith	12	12	7
Country	40	38	34

Backup E: F1 score

		actual		
		positive	**negative**	
predicted	**positive**	True Positive (TP)	False Positive (FP)	$precision = \frac{TP}{TP + FP}$
	negative	False Negative (FN)	True Negatives (TN)	
		$recall = \frac{TP}{TP + FN}$		

$$F1 - score = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Backup F: P-values

The p-value (or probability value), tells you how likely it is that your data could have occurred under the null hypothesis.

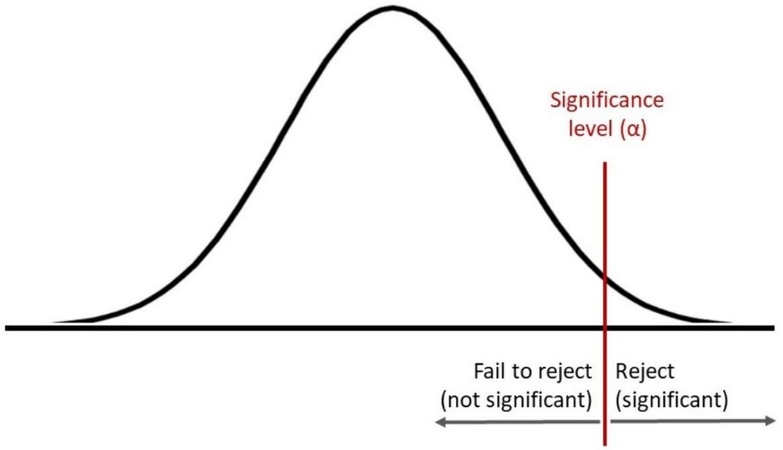
A smaller p-value means that there is stronger evidence to reject the null hypothesis.

In our case, when evaluating correlations, the null hypothesis is the independence of the two tested variables.

Backup F: P-values

eg. if p-value = 0.01 ➡the 1% of times we would see a test statistics that is at least extreme as the one we are considering if the two variables were indeed independent (i.e. null hypothesis true).

Hypothesis testing:
(right tail)



Backup G: Spearman's index

6.1.1 Spearman's ρ

Spearman's test is a non-parametric test to measure correlations between numerical or ordinal variables. It captures the monotonic relationship between data and returns a value $\rho \in [-1, 1]$, where 0 implies that the two variables are actually independent and the sign highlights the direction of the correlation.