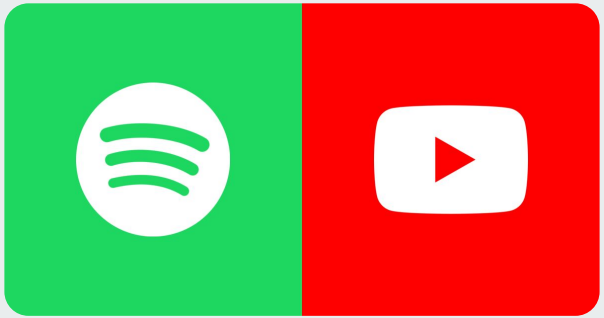
Spotify & Youtube

Performance im Vergleich



Collective Intelligence - SoSe2023

Projekt von Sebastian Braun & Sandra Kiefer

Hochschule **RheinMain**

Inhaltsverzeichnis

- Motivation
- Explorative Datenanalyse (EDA)
- Lösungswege

  → Predict Views/Likes/Kommentare/Streams und Klassifizierung Popularität
- Herausforderungen
- Evaluationsergebnisse
- Diskussion

  → Vergleich zu State-of-the-Art, Eigenschaften der Popularitätsklassen
- Zusammenfassung
- Ausblick

Motivation

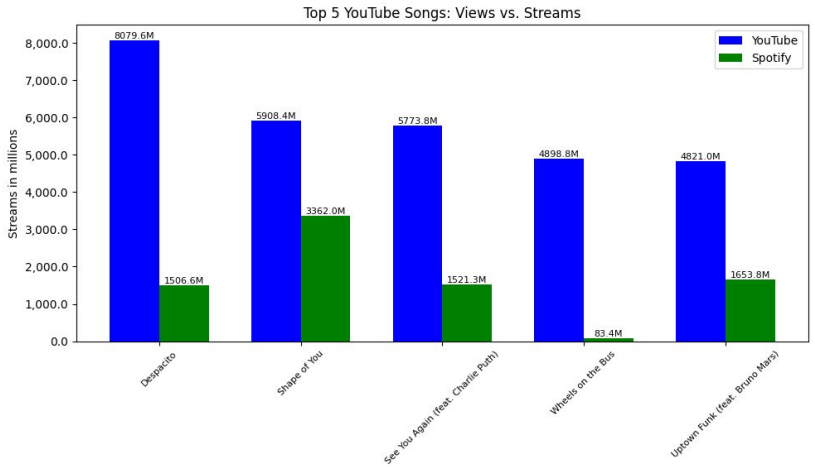
- Performance von Liedern anhand ihrer Eigenschaften vorhersagen
- Objektive Einschätzung der Qualität, Popularität und kommerzieller Erfolgschancen
- Besonders wertvoll für Musikproduzenten und Plattenlabels
- Zielgruppen analysieren
- Gezieltes Marketing und maximieren der Reichweite und Erfolgschancen
- Schwachstellen eines Liedes analysieren

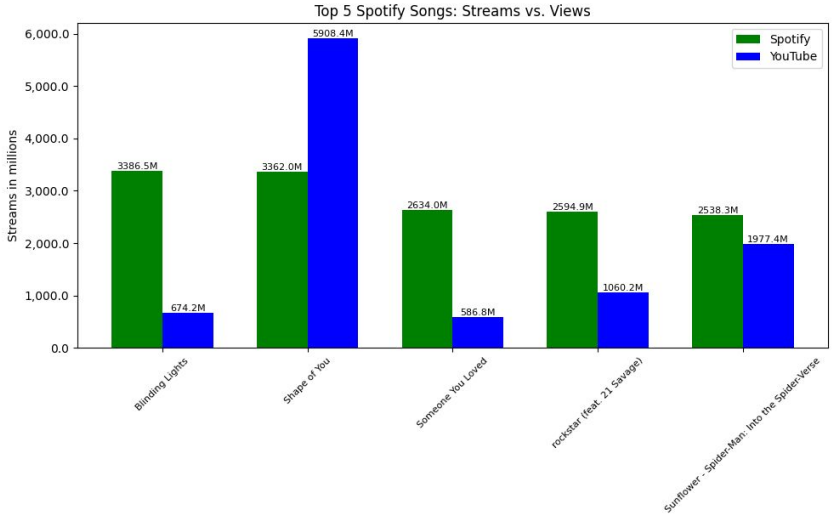
Explorative Datenanalyse

- Datensatz wurde erhoben am 07. Februar 2023
- Umfasst Einträge von 20717 Liedern
- Top 10 Songs verschiedener Künstler
- Aufrufzahlen von YouTube und Spotify
- Eigenschaften der Lieder:

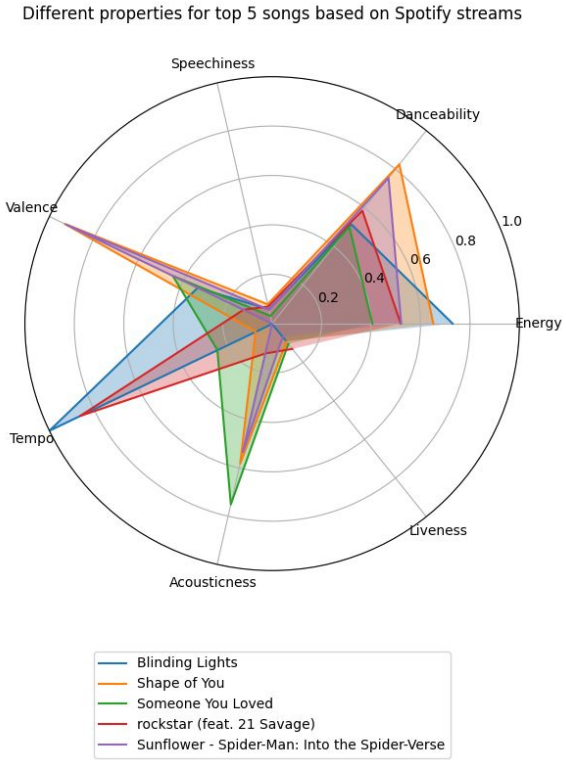
  ⟶ Danceability, Energy, Key, Loudness, Speechiness, Acousticness, Instrumentalness, Liveness, Valence und Tempo

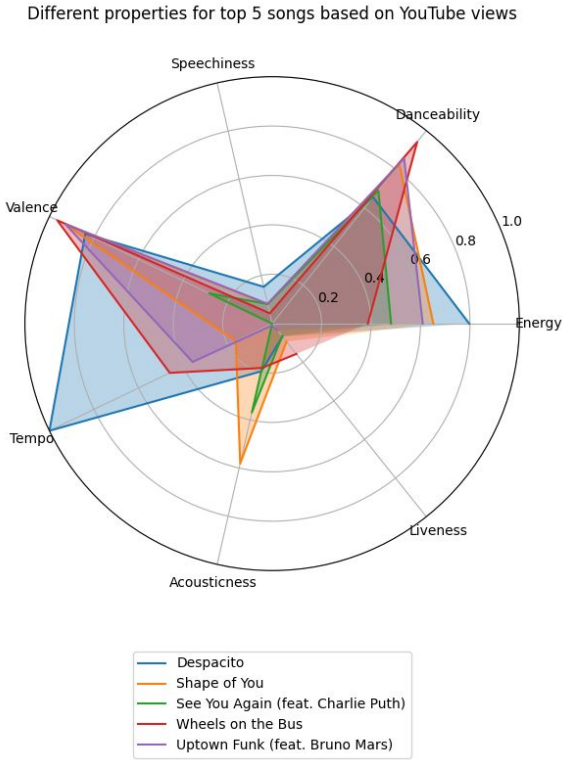
Top 5 YouTube vs. Spotify





Eigenschaften der Top Hits





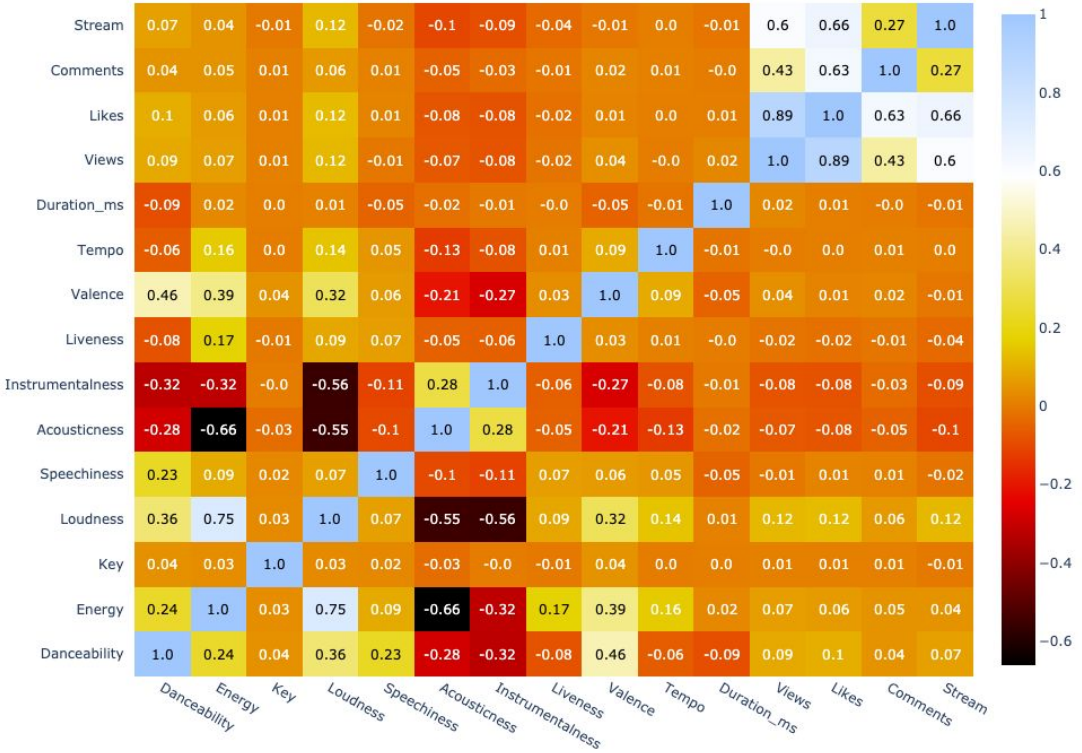
Lösungswege - Predict Views/Likes/Streams ...

Vorhersage	Linear Regression (MSE)	Complex Regression (MSE)
Views	13 845 475 698 058 648	3 186 447 608 515 025 408
Likes	332 489 855 901	51 371 805 025 899
Kommentare	8 512 847 112	30 114 110 934
Streams	29 869 146 780 298 188	167 281 257 616 288 192

→ keine aussagekräftigen Ergebnisse

(da ein Wert in einem sehr großen Wertebereich vorhergesagt werden muss)

Herausforderungen



Lösungswege - Klassifizierung Popularität

Popularitätsklasse	Low Popularity	Moderate Popularity	Good Popularity	High Popularity	Very High Popularity
Popularitätswert (%)	[0-30]	[31-50]	[51-80]	[81-90]	[91-100]

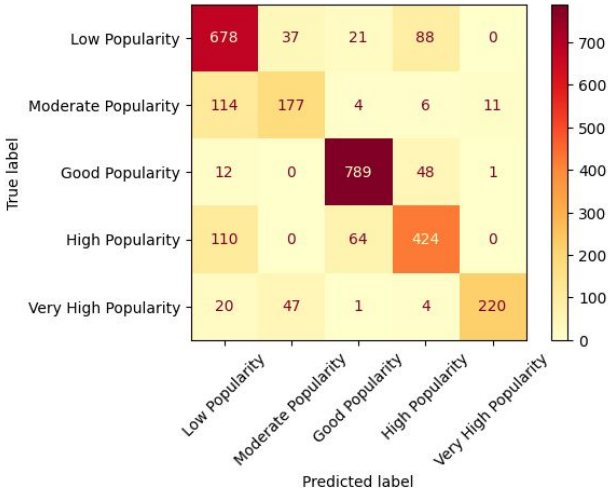
Berechnung des Wertes / der Klasse aus den Views, Likes, Kommentare und Streams

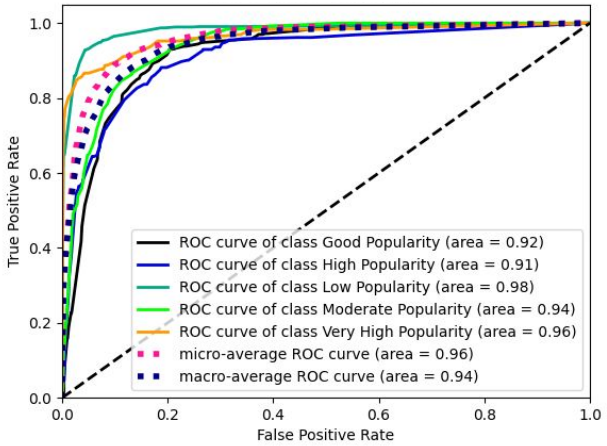
Evaluationsergebnisse - Übersicht

Modell	Accuracy	Precision	Recall
Gradient Boosting	97.71 %	97.35 %	97.35 %
KNN	95.41 %	95.02 %	94.53 %
SVM	95.27 %	94.73 %	94.56 %
Random Forest	94.85 %	95.09 %	92.88 %
Naive-Bayes	88.46 %	87.07 %	86.25 %
Logistic Regression	83.52 %	82.63 %	82.92 %
Decision Tree	79.55 %	79.88 %	75.61 %

Evaluationsergebnisse - schlechteste Klassifizierung

Decison Tree

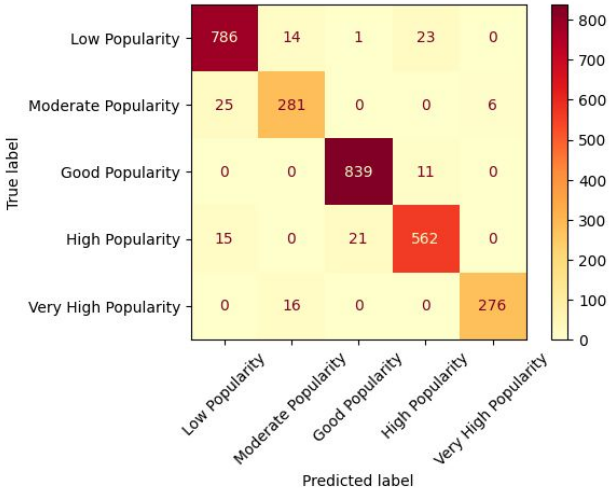


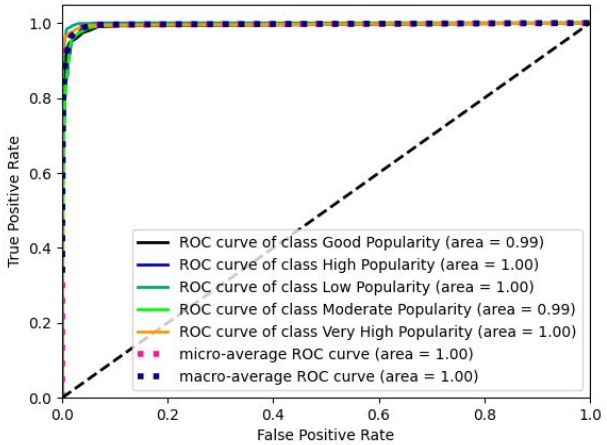


	TP	TN	FP	FN	TPR	TNR	FPR	FNR	Precision	Recall	F1
Low Popularity	678.0	1796.0	256.0	146.0	0.8228	0.8752	0.1248	0.1772	0.7259	0.8228	0.7713
Moderate Popularity	177.0	2480.0	84.0	135.0	0.5673	0.9672	0.0328	0.4327	0.6782	0.5673	0.6178
Good Popularity	789.0	1936.0	90.0	61.0	0.9282	0.9556	0.0444	0.0718	0.8976	0.9282	0.9127
High Popularity	424.0	2132.0	146.0	174.0	0.709	0.9359	0.0641	0.291	0.7439	0.709	0.726
Very High Popularity	220.0	2572.0	12.0	72.0	0.7534	0.9954	0.0046	0.2466	0.9483	0.7534	0.8397

Evaluationsergebnisse - sehr gute Klassifizierung

K-Nearest-Neighbor





	TP	TN	FP	FN	TPR	TNR	FPR	FNR	Precision	Recall	F1
Low Popularity	786.0	2012.0	40.0	38.0	0.9539	0.9805	0.0195	0.0461	0.9516	0.9539	0.9527
Moderate Popularity	281.0	2534.0	30.0	31.0	0.9006	0.9883	0.0117	0.0994	0.9035	0.9006	0.9021
Good Popularity	839.0	2004.0	22.0	11.0	0.9871	0.9891	0.0109	0.0129	0.9744	0.9871	0.9807
High Popularity	562.0	2244.0	34.0	36.0	0.9398	0.9851	0.0149	0.0602	0.943	0.9398	0.9414
Very High Popularity	276.0	2578.0	6.0	16.0	0.9452	0.9977	0.0023	0.0548	0.9787	0.9452	0.9617

Diskussion

Wissenschaftliche Veröffentlichung

“Predicting Music Popularity Using Spotify and YouTube Features”

- selbstständige Musik-Feature-Extraction
- bestes Modell = Random Forest mit 79,6% Accuracy

Artikel

“Effect of Feature selection on the Accuracy of Music Popularity Classification Using Machine Learning Algorithms”

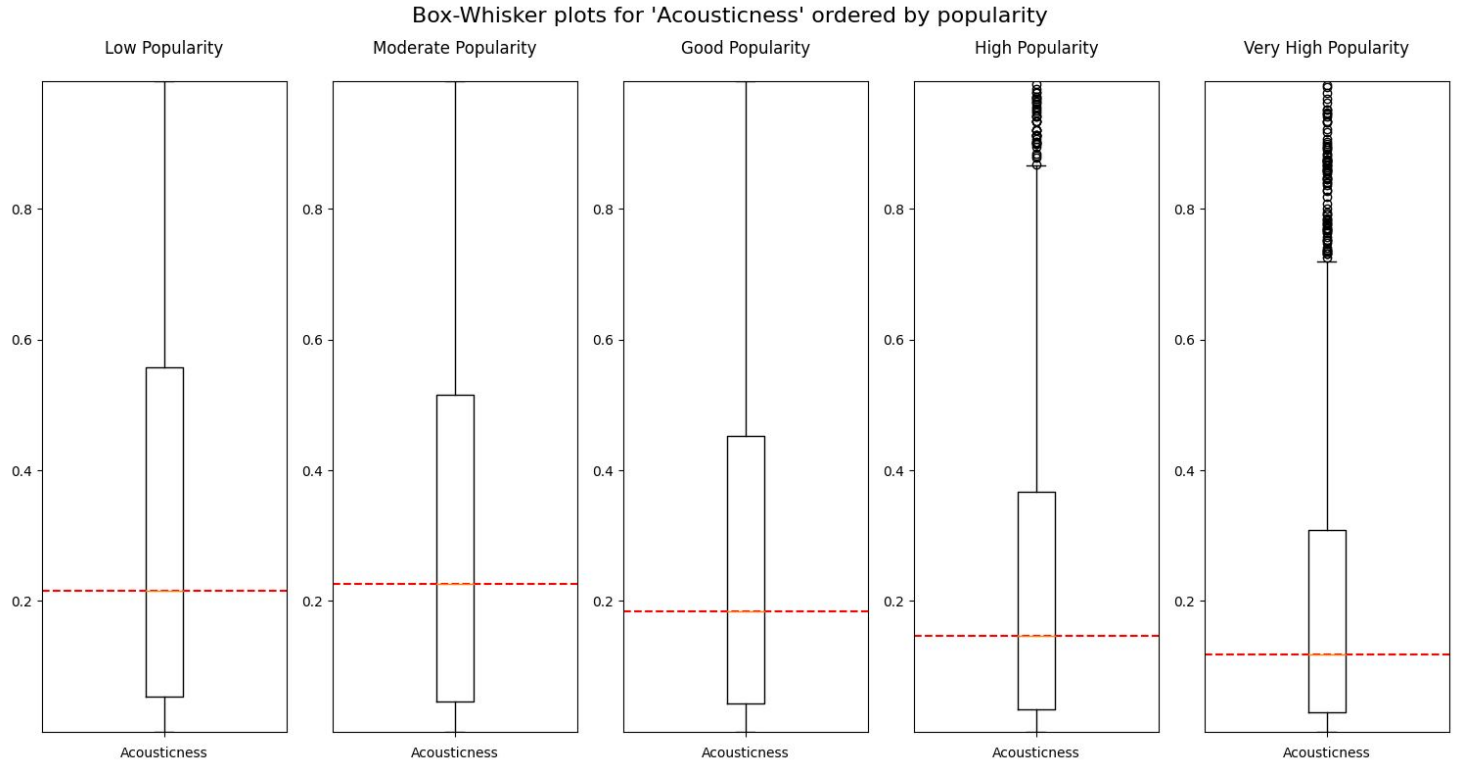
- Einbeziehung von Social-Media-Merkmalen
- Modell mit 95,15% Accuracy
- Optimierung der Laufzeit

Python Notebook auf Kaggle

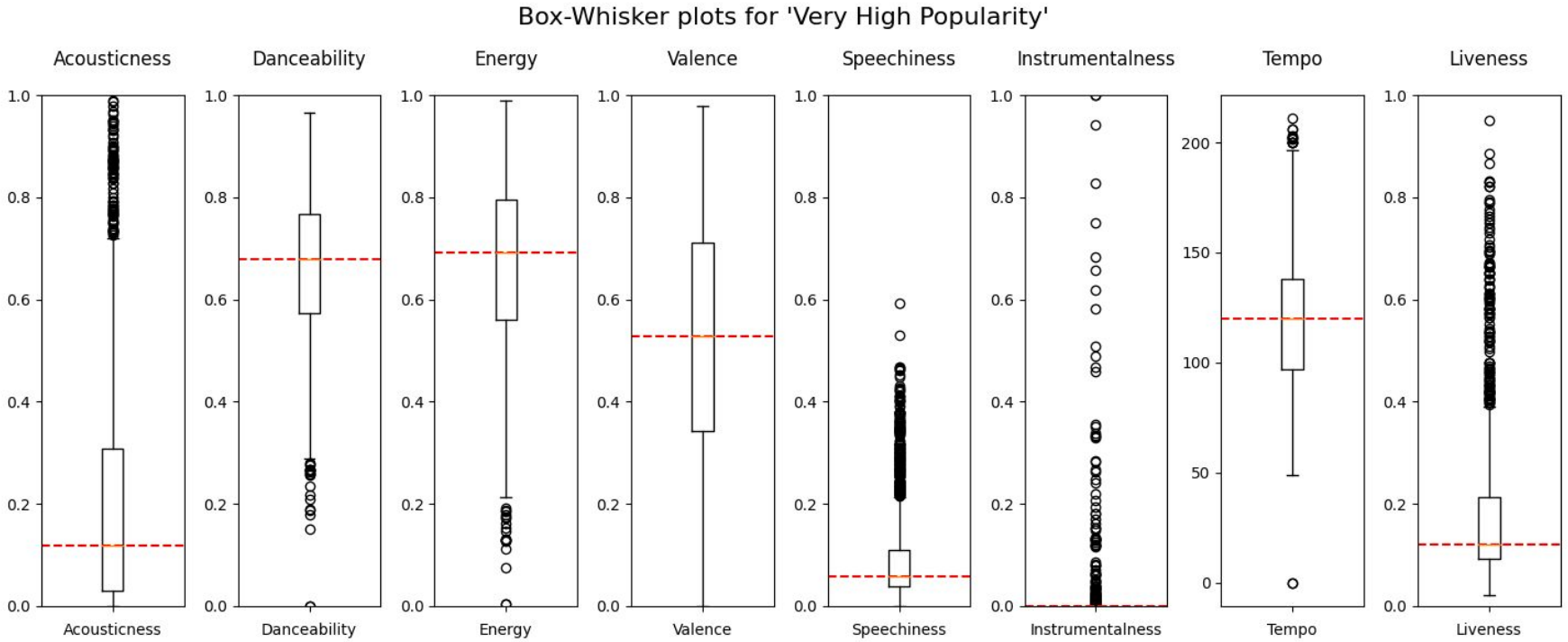
“Youtube/Spotify EDA and Simple Regression”

- Fokus auf die Visualisierung des Datensatzes
- gescheitertes Prediciton Modell mit Linear Regression

Eigenschaften der Popularitätsklassen im Vergleich



Eigenschaften der höchsten Popularität



Zusammenfassung

- Erfolgreiche Entwicklung eines Modells zur Vorhersage der Performance
- Erste Baseline zur Vorhersage genauer Zahlen erfolglos
- Zweite Baseline zur Klassifizierung on Popularitätsklassen
- Zweite Baseline mit verschiedenen Algorithmen getestet und verbessert
- Bestes Ergebnis für Gradient Boosting mit 97% Accuracy, Precision & Recall
- Einfluss der Eigenschaften anhand Box Whisker Plots ablesen
- Eigenschaften populärer Lieder: geringer akustischer Anteil, sehr tanzbar, energetisch und fröhlich, Gleichgewicht zwischen Gesang und Melodie, Geschwindigkeit um 120 BPM

Ausblick

- weitere Unterteilung in noch mehr Popularitätsklassen

  → um genauere (kleinschrittigere) Vorhersagen/Klassifizierungen zu treffen

- eigene Popularitätsklassen für Views, Likes, Kommentare und Streams

  → um Aussagen über nur einen Parameter tätigen zu können

- Hinzunahme weiterer Features in den Datensatz

  → um genauere Klassifizierung der Lieder zu ermöglichen

  → Betrachtung der Korrelationen ist dabei wichtig