Kework K. Kalustian[1,2]

[1]Max-Planck-Institut für empirische Ästhetik (Frankfurt am Main),

[2]Goethe-Universität (Frankfurt am Main)

Spotify-Streaming der täglichen deutschen Top 200-Charts vor und während der COVID-19-Pandemie: Exploration und Vorhersage von Streaminghäufigkeiten

Die COVID-19-Pandemie stellt ein Ereignis dar, das weitreichende Konsequenzen nach sich zieht. Etliche negative Konsequenzen sind zu beobachten. Hierbei kann aber ebenso festgestellt werden, dass Menschen während der bundesweiten Kontaktbeschränkungen anscheinend vermehrt musikalische Angebote nutzen; sowohl medial über verschiedene Streaming Service-Anbieter (Apple Music, Spotify, YouTube etc.) als auch praktisch (z.B. kollektives Singen auf Balkonen, virtuelles Musizieren über YouTube).

Da dieses Phänomen des vermehrten Musikhörens und des Musizierens im Zusammenhang mit der COVID-19-Pandemie als Bewältigungsstrategie (coping) interpretiert werden kann, um mit den sozialen Einschränkungen besser umgehen zu können, lässt sich daher fragen: Wenn während des Zeitraums, in dem in Deutschland bundesweite Kontakteinschränkungen (inklusive Lockerungen) gegolten haben (11.03.20–29.06.20), die mittleren Streaminghäufigkeiten der täglichen deutschen Top 200-Spotify-Charts tatsächlich erheblich höher als vor der COVID-19-Pandemie waren, inwiefern ist dann das tatsächliche Streamingverhalten (operationalisiert als tägliche Streaminghäufigkeiten) mit akzeptabler Genauigkeit, basierend auf Daten zu den täglichen mittleren Streaminghäufigkeiten vor der COVID-19-Pandemie, vorhersagbar?

Um die Forschungsfrage beantworten zu können, wurden durch Web-Scraping von der Spotify-Internetseite u.a. Chartplatzierungen, Tracktitel und Streaminghäufigkeiten für den Zeitraum vom 01.01.2019–29.06.2020 gesammelt und für die nachfolgenden Datenanalysen bereinigt und aufbereitet (sämtliche Datenbeschaffungen, -aufbereitungen und -analysen erfolgten mittels R). Anschließend wurde der gesamte Datensatz heuristisch in zwei Zeiträume aufgeteilt, um die notwendigen Referenzwerte zu erhalten: Der „Vor-Pandemie-Zeitraum“ dauert vom 01.01.2019–10.03.20; der (untersuchte) „Pandemie-Zeitraum“ vom 11.03.20–29.06.20. Das Entscheidungskriterium für diese Trennung ist, dass am 11.03.20 der dritte COVID-19-Todesfall in Deutschland gemeldet wurde und die WHO die COVID-19-Krise offiziell als Pandemie einstufte. Der 29.06.20 wurde als Enddatum für den Untersuchungszeitrum gewählt, da die bundesweiten Kontaktbeschränkungen bis zu diesem Datum einheitlich gegolten haben.

Nachdem die Unterdatensätze für beide Zeiträume nach Wochentagen gruppiert und tägliche Median-Streaminghäufigkeiten für die jeweiligen Wochentage ermittelt wurden, ist ein signifikant starker Unterschiedseffekt nach Cohen (1992) in den zentralen Tendenzen der Median-Streaminghäufigkeiten der nach Wochentagen gruppierten Zeiträume (wöchentlicher $M_{\text{Vor-Pandemie-Zeitraum}}$ = 90871.4; wöchentliche $SD_{\text{Vor-Pandemie-Zeitraum}}$ = 8882.15 sowie wöchentlicher M_{Pandemie} = 96156.6; wöchentlicher SD_{Pandemie} = 8005.66) bei einem einseitigen t-Test (α = 5%) – entsprechend der Hypothese, dass während der Pandemie höhere Streaminghäufigkeiten beobachtet werden können – feststellbar ($t_{(6)} = 10.03$; $p < .001$; $r = .971$). Aufbauend auf diesem erwarteten Befund erfolgte die Analyse der Zeitreihen der beiden Zeiträume. Konkret: Die beiden Unterdatensätze wurden so aufgeteilt, dass der Testdatensatz den postulierten Pandemie-Zeitraum darstellt (etwa 20 % des gesamten Datensatzes) und der Vor-Pandemie-Zeitraum den Trainingsdatensatz (etwa 80 %). Die Trainingsdatensatz-Varianz wurde durch eine Transformation (natürlicher Logarithmus) stabilisiert, während durch eine anschließende Differenzierung saisonale Trends reduziert wurden. Um zu überprüfen, wie viele Zeiteinheiten im stationierten Trainingssatz verstreichen müssen, bis sich dieselbe Periode (bzw. dasselbe Intervall) wiederholt, wurden durch die Methode der gewöhnlichen kleinsten Quadrate (OLS) Schätzungen der sogenannten Lags (Zeitverschiebungen bzw. Zeitintervall) ermittelt. Es konnten 7 Lags ermittelt werden, was insofern Sinn ergibt, als wöchentliche Trends (7 Tage) in einem kombinierten Streu- und Liniendiagramm visuell bei dem originalen (nicht-stationierten) Trainingsdatensatz zu beobachten sind (s. Abbildungen 2.1 und 2.2 im Poster). Gemäß den Ergebnissen von KPSS-Tests und des erweiterten Dickey-Fuller-Tests konnte schließlich nachgewiesen werden, dass der saisonal-differenzierte und stabilisierte Trainingsdatensatz keinen Hinweis darauf liefert, eine zugrundeliegende Einheitswurzel des saisonal-differenzierten und stabilisierten Trainingsdatensatzes anzunehmen. Das heißt, der vorbereitete Trainingsdatensatz entspricht einem stationären Prozess mit Trend.

Vor diesem Hintergrund kommen klassische lineare Modelle nicht infrage, weshalb sowohl mittels Random Forest (RF)-Regression als auch mit einer kNN-Regression die täglichen tatsächlichen

Median-Streaminghäufigkeiten des Pandemie-Zeitraums vorhergesagt wurden. Die beiden Algorithmen wurden deshalb gewählt, weil sie eine vergleichsweise große Teststärke bei gleichzeitiger Robustheit aufweisen. Für die Vorhersage wurde der Trainingsdatensatz mit den 7 Lags in einen zweidimensionalen Euklidischen Raum mit einem zusätzlichen Lag eingebettet, um eine für die Algorithmus lernbare Matrix zu schaffen, in der die 7 Lags als Prädiktoren für den zusätzlichen Lag fungieren. Nach erfolgter Hyperparameteroptimierung mittels 10-facher Kreuzvalidierung wurde schließlich der RF-Algorithmus mit 500 Regressionsbäumen trainiert, sodass Punktschätzungen der tatsächlichen Median-Streaminghäufigkeiten erfolgen konnten (Vorhersagehorizont von 111 Tagen). Der kNN-Algorithmus wurde mit einem Kappa-Wert von 15 trainiert, der ebenfalls durch eine 10-fache Kreuzvalidierung ermittelt wurde. Nachdem die Retransformierung der Punktschätzungen erfolgte, wurden die Fehlermetriken (Accuracy) des RF-Modells ermittelt (MAE_{RF} = 5151.36; $RMSE_{RF}$ = 6633.3; $MAPE_{RF}$ = 5.29) und überprüft, um wie viel Prozent das trainierte Modell genauere Punktschätzungen leistet als ein naives Benchmark-Modell (saisonal-naiv): Die RF-Punktschätzungen sind um 49 % genauer ($UMBRAE_{RF\ :\ sainsonal\text{-}naiv}$ = .51); außerdem sind die zentralen Tendenzen der RF-Punkt-schätzungen (M_{RF} = 103930; SD_{RF} = 3750.82) und die der tatsächlichen Werte ($M_{Tatsächlich}$ = 96295.6; $SD_{Tatsächlich}$ = 9315.41) *nicht* signifikant verschieden (zweiseitiger t-Test: $t_{(110)}$ = .66; p = .51). Die Fehlermetriken des kNN-Modells sind dem RF-Modell ähnlich: MAE_{kNN} = 5314.8; $RMSE_{kNN}$ = 6926.32; $MAPE_{kNN}$ = 5.6. Wobei die kNN-Punktschätzungen um 1 % weniger genau sind als das RF-Modell in Bezug auf das saisonal-naive Benchmark-Modell ($UMBRAE_{kNN\ :\ sainsonal\text{-}naiv}$ = .52). Allerdings zeigt hier ein zweiseitiger t-Test, dass die zentralen Tendenzen des kNN-Modells erheblich von den tatsächlichen Werten (s.o.) abweichen (M_{kNN} = 98803.3; SD_{kNN} = 9427.39), sodass dieser Unterschied als mittelstarker Effekt eingestuft werden kann (zweiseitiger t-Test: $t_{(110)}$ = -4.1; p < .001; r = .36) .

Den Ergebnissen der Fehlermetriken folgend, ist festzuhalten, dass die in Anschlag gebrachte Lernstrategie für den RF-Algorithmus als auch für den kNN-Algorithmus (Lag-Einbettungen) als Vorhersage-Modelle zufriedenstellend ausfällt, da sowohl etwa der *MAPE*-Wert deutlich unter 10 liegt, was nach Lewis (1982) als sehr akkurate Vorhersage gilt, als auch eine Kreuzvalidierung über einen t-Test keinen signifikanten Unterschied in den zentralen Tendenzen der tatsächlichen und vorhergesagten RF-Werte ergab. Dass hierbei das kNN-Modell insgesamt jedoch schlechter abschneidet, kann vermutlich daran liegen, dass der kNN-Algorithmus tendenziell dazu neigt, eher überanzupassen, sodass während der Hyperparameteroptimierung zwar bessere Ergebnisse erzielt, aber ungesehene Validierungs- oder gar Testdaten entsprechend schlechter geschätzt werden. Diese Vermutung deckt sich auch mit dem Befund, dass das kNN-Modell während der Kreuzvalidierung zur Identifikation des optimalen Hyperparameters (κ) weniger Fehler aufweist ($RMSE_{kNN\text{-}CV}$ = 0.070) als das RF-Modell bei seiner Hyperparameteroptimierung ($RMSE_{RF\text{-}CV}$ = 0.074).

Die Frage nach der inhaltlichen Relevanz einer derartigen Vorhersage-Strategie kann z.B. so beantwortet werden, dass unter Anwendung dieser Strategie auch multivariate Modelle denkbar sind (z.B. unter Einbeziehung der Chart-Positionen und SongIDs sowie z.B. unter Einbeziehung der Sentimentqualitäten von Song-Texten), die dann als Grundlage für datumssensitive und personalisierte Musikempfehlungssysteme dienen können, wodurch eine datengetriebene Anwendung von Erkenntnissen zum Musiknutzungsverhalten und zum Moodmanagement anvisiert ist.

Schlüsselwörter: Musik und COVID-19, Musiknutzungsverhalten, Random Forest- und kNN-Regression Spotify-Streaming, Zeitreihen

Wichtigste Quellen:

Bergmeir, C, & Benítez, J.B. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences*. 191, 192–213.

Breiman, L. (2001). Random Forests. *Machine Learning. 45*, 5–32. DOI: 10.1023/A:1010933404324

Efron, B. & Tibshirani, R. (1986). Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Statistical Science, 1*(1), 54–75. DOI:10.1214/ss/1177013815.

Gareth, J., Witten, D., Hastie, T., Tibshirani, R. (2013). *An Introduction to Statistical Learning. With Applications in R. 112*. New York: Springer.

Hyndman, R.J., & Athanasopoulos, G. (2018). *Forecasting: principles and practice*, OTexts: Melbourne, Australia. OTexts.com/fpp2. Zugriff: 01.08.20.

Wyner, A. J., Olson, M, Bleich, J. (2017). Explaining the success of adaboost and random forests as interpolating classifiers, *The Journal of Machine Learning Research. 18*, 1558–1590.

Juslin, P.N., & Sloboda, J.A. (Hrsg.). (2010). Series in affective science. Handbook of music and emotion: Theory, research, applications. Oxford University Press.