Implementierung von Boosting-Algorithmen zur Modellierung der Qualifikationswahrscheinlichkeit im Eurovision Song Contest (ESC)

Ein Projekt von Karl und Joris Parthier

14. Januar 2025

Projektüberblick

Welchen Einfluss hat das Geschlecht des Künstlers?
Wie beeinflusst die Reihenfolge des Auftritts die Wahrscheinlichkeit der Qualifikation?

Diese Fragen und weitere widmeten wir uns in unserem "Jugend forscht"-Projekt. So entwickelten wir verschiedene Gradient Boosted Trees, das sind maschinell optimierte Algorithmen, die basierend auf den Fehlern vorheriger Modelle versuchen, sich sequenziell selbst beizubringen, diese zu erkennen und zu korrigieren.

Im Rahmen des Projekts entwickelten wir verschiedene Merkmale, wie beispielsweise die Messung der sprachlichen Distanz zwischen Sprachen, die Analyse der Songtexte mithilfe des Flesch–Kincaid Readability Tests, sowie die musikalische Zerlegung eines Musikstücks unter Verwendung der Python-Bibliothek Librosa. Ziel war es, zu untersuchen, wie diese Merkmale die Wahrscheinlichkeit einer Qualifikation beeinflussen.

Inhaltsverzeichnis

1 Fachliche Kurzfassung

Der Eurovision Song Contest ist einer der ältesten jährlichen internationalen Musikwettbewerben der Welt. Trotz der umfänglichen Analyse des Eurovision Song Contest in unzähligen Recherche-Feldern ist bisher nur wenig über die Qualifikationsrunden bekannt. In unserem Projekt "Boosting-Algorithmen zur Modellierung der Qualifikationswahrscheinlichkeit im ESC", knüpfen wir daher an verschiedenen Recherche Ideen an und modellieren die Qualifikationswahrscheinlichkeiten der verschiedenen Teilnehmer des ESC (Eurovision Song Contest) basierend auf den Teilnehmerbeiträgen im Zeitraum von 2018 bis 2024. Dabei trainierten wir unterschiedliche Konfigurationen von LightGBM-Klassifikatoren auf Performancefaktoren, externen Faktoren, Wettbewerbsfaktoren und weiteren Einflussfaktoren, beschränkten aber die Merkmalsaufzeichnungen auf 6 Tage vor dem Finale, um den prädiktiven Character beizubehalten. Unsere Ergebnisse zeigen, dabei eine deutliche Verbesserung gegenüber einem Random Classifier und treffen dabei konstant bessere Vorhersagen, als die Quoten von Buchmachern. Zusätzlich nehmen wir eine Auswertung der Merkmalsbedeutung vor und versuchen dabei unsere Ergebnisse in einem Fazit zu bündeln.

2 Motivation und Fragestellung

Im Jahr 2022 erzählten uns österreichische Bekannte von ihrer Enttäuschung, dass ihr Teilnehmer die Qualifikation nicht geschafft hatte. Dies nahmen wir als Anlass uns grundsätzlich über die Wahrscheinlichkeit der Qualifikation Gedanken zu machen.

Wir entwickelten daraus 4 Hypothesen:

- *H1: Wettbewerbsfaktoren wie die Startreihenfolge haben einen signifikanten Einfluss auf die Qualifikationswahrscheinlichkeit.*
- *H2: Externe Faktoren, wie die multidimensionale Nähe und Affinität zwischen Ländern, beeinflussen die Qualifikationswahrscheinlichkeit signifikant.*
- *H3: Performance-Merkmale eines Beitrags, wie die Komplexität der Lyrik sowie andere musikalische Merkmale sind ebenfalls signifikant für die Qualifikation.*
- *H4: Gradient-Boosted-Tree-Klassifikatoren können in Kombination mit den analysierten Merkmalsgruppen bessere Qualifikationswahrscheinlichkeiten vorhersagen als die Quoten von Buchmachern.*

3 Hintergrund und theoretische Grundlagen

Bereits aus der Literatur geht hervor, dass das Voting beim Eurovision Song Contest (ESC) durch sogenannte "Voting Blocks"beeinflusst wird. Diese Ländergemeinschaften vergeben regelmäßig hohe Punkte aneinander, basierend auf nicht-wettbewerbsrelevanten Faktoren Gatherer 2006. Darauf aufbauend stellten Ginsburgh und Noury 2008 für den Zeitraum von 1956 bis 2003 fest, dass sprachliche und kulturelle Gemeinsamkeiten das Wahlverhalten im ESC signifikant beeinflussen. Ergänzend dazu identifizierten Spierdijk und Vellekoop 2009 die Bedeutung der geographischen Nähe als weiteren entscheidenden Faktor im Zeitraum von 1975 bis 2003. Antipov und Pokryshevskaya 2017 belegten, dass die Reihenfolge des Auftritts ebenfalls das Voting-Bias der Jury und der Televoter beeinflusst. Ein weiterer Einflussfaktor ist die Soft Power, wie Aðalheiðardóttir 2022 herausfand. Besonders für kleinere Länder und sogenannte "große schwache Länder"— also Staaten, die nach internationalem Ansehen streben und deren internationale Angelegenheiten stark davon abhängen — spielt Soft Power eine wichtige Rolle Magnúsdóttir und Thorhallsson 2011. Aydin 2022 untersuchte den Einfluss von Texten (Lyrics) und anderen musikalischen Eigenschaften auf den ESC mithilfe von Machine-Learning-Methoden, darunter Random Forests und neuronale Netzwerke, basierend auf ECHO-Nest-Daten. Die Modelle, die nur auf Audio- oder nur auf Textdaten trainiert wurden, schnitten besser ab als ein zufälliger Klassifikator. Millner u. a. 2015 analysierten zusätzlich Faktoren, wie die mögliche Bevorzugung des Vorjahres Siegers (Host-Effekt) und bewerteten die Bedeutung der Anzahl, sowie des Geschlechts der Künstler. Dabei zeigten sich Geschlechtsunterschiede und eine besondere Relevanz für Transgender-Künstler. Auch die Sprache des Beitrags beeinflusste das Abstimmungsverhalten, wobei insbesondere Französisch als nachteilig identifiziert wurde. Wir knüpfen in dem Sinne an die vorausgehende Literatur an, indem wir bereits bekannte Einflussfaktoren ergänzen und auf die Qualifkationsrunde beziehen.

4 Vorgehensweise

Ferner konnten wir auf Grundlage von Literatur und eigenen Ideen mögliche relevante Feature in interne und externe Einflussfaktoren unterteilen. Intern definiert sich hier als wettbewerbspezifische Faktoren, wobei externe Faktoren generalisierte Aussagen über den Teilnehmer treffen. Im Folgenden werden die zu untersuchenden Merkmale, Definitionen und Merkmalgruppe zusammengefasst und erläutert.

5 Interne Faktoren

5.1 Audioanalytische Merkmale

Variable	Definition	Erfassungsmethode
bpm	Beats per Minute (BPM) quantifiziert die Anzahl der Taktschläge pro Minute in einer musikalischen Komposition	Kontinuierliche Messung mittels Python-Bibliothek librosa
onset_strength	Quantifiziert die Intensität des Einsetzens musikalischer Ereignisse (z.B. Noten oder Schläge) im Audiosignal	ZTM DW MD[1]
rms_energy	Repräsentiert die durchschnittliche Signalintensität des Audiosignals	ZTM, DW sowie MD
spectral_centroid	Charakterisiert den Frequenzschwerpunkt eines Klangs	ZTM, DW sowie MD
spectral_rolloff	Indikator für die spektrale Energieverteilung	ZTM, DW sowie MD
mfcc	Mel-Frequency Cepstral Coefficients repräsentieren die charakteristischen Frequenzmerkmale eines Audiosignals	ZTM, DW sowie MD
chroma	Repräsentation der Tonhöhenverteilung basierend auf den zwölf Halbtönen der chromatischen Skala	ZTM, DW sowie MD
key	Tonart basierend auf Librosa-Chromabändern-Analyse	Integer (0-11)

5.2 Textanalytische Merkmale

Variable	Definition	Erfassungsmethode
Average_Syllables_Word[2]	Durchschnittliche Silbenanzahl pro Wort in den Songtexten	Nur für englischsprachige oder dominant englischsprachige Titel
Percentage_Unique_Words	Prozentualer Anteil distinktiver Wörter im Verhältnis zur Gesamtwortanzahl	Nur für englischsprachige oder dominant englischsprachige Titel
Percentage_Difficult_Words[3]	Prozentualer Anteil als "schwierig"klassifizierter Wörter gemäß Dale-Chall Readability Formula	Nur für englischsprachige oder dominant englischsprachige Titel

5.3 Sprachliche Kategorisierung

Variable	Definition	Kodierung
Language_English	Englisch als dominante Sprache	Binär (1/0)
Language_French	Französisch als dominante Sprache	Binär (1/0)
Language_Italian	Italienisch als dominante Sprache	Binär (1/0)
Language_Portuguese	Portugiesisch als dominante Sprache	Binär (1/0)
Language_Spanish	Spanisch als dominante Sprache	Binär (1/0)
Language_Other	Residualkategorie für nicht spezifizierte dominante Sprachen	Binär (1/0)

5.4 Künstler spezifische Merkmale

Variable	Definition	Kodierung
gender_Female	Weibliche Interpretin	Binär (1/0)
gender_Male	Männlicher Interpret	Binär (1/0)
gender_Mix	Geschlechtergemischte Darbietung	Binär (1/0)
gender_other_gender	Alternative Geschlechtsidentifikation	Binär (1/0)

6 Externe Faktoren

6.1 Brand Finance Indikatoren (GSPI)

Variable	Definition	Erfassungsmethode
Gspi[4]	Global Soft Power Index - misst die Fähigkeit eines Landes, internationale Ziele zu erreichen	Erfassung des Punktwerts, des globalen Rangs sowie der jährlichen Veränderung in Punkten sowie Prozent
Fam	Familiarity - misst den Bekanntheitsgrad eines Landes	Erfassung des Punktwerts, des globalen Rangs sowie der jährlichen Veränderung in Punkten sowie Prozent

Reput	Reputation - evaluiert das internationale Ansehen eines Landes	Erfassung des Punktwerts, des globalen Rangs sowie der jährlichen Veränderung in Punkten sowie Prozent
Influence	Quantifiziert den aktiven und passiven Einfluss eines Landes	Erfassung des Punktwerts, des globalen Rangs sowie der jährlichen Veränderung in Punkten sowie Prozent

6.2 Hofstede Kulturdimensionen

Variable	Definition	Erfassungsmethode
PDI	Power Distance Index[5] - misst die Akzeptanz ungleicher Machtverteilung	Die Erfassung des Punktewertes sowie der Differenz zum durchschnittlichen ESC-Teilnehmer des jeweiligen Jahres erfolgt basierend auf den Werten der Hofstede Insights Website
IDV	misst, ob eine Gesellschaft individuelle Freiheit oder Kollektivismus betont.	Die Erfassung des Punktewertes sowie der Differenz zum durchschnittlichen ESC-Teilnehmer des jeweiligen Jahres erfolgt basierend auf den Werten der Hofstede Insights Website
MAS	Maskulinität vs. Femininität – erfasst, inwieweit eine Gesellschaft Werte wie Durchsetzungsvermögen, Erfolg und Wettbewerb im Vergleich zu femininen Werten wie Fürsorge, Kooperation und Lebensqualität betont	Die Erfassung des Punktewertes sowie der Differenz zum durchschnittlichen ESC-Teilnehmer des jeweiligen Jahres erfolgt basierend auf den Werten der Hofstede Insights Website

UAI	Uncertainty Avoidance Index - quantifiziert den Umgang mit Unsicherheit	Die Erfassung des Punktewertes sowie der Differenz zum durchschnittlichen ESC-Teilnehmer des jeweiligen Jahres erfolgt basierend auf den Werten der Hofstede Insights Website
LTO	Long-Term vs. Short-Term Orientation (LTO) – misst die zeitliche Ausrichtung gesellschaftlicher Werte, wobei langfristige Orientierung auf Zukunftsplanung, Ausdauer und Sparsamkeit abzielt, während kurzfristige Orientierung Traditionen, soziale Verpflichtungen und schnelle Ergebnisse in den Vordergrund stellt	Die Erfassung des Punktewertes sowie der Differenz zum durchschnittlichen ESC-Teilnehmer des jeweiligen Jahres erfolgt basierend auf den Werten der Hofstede Insights Website
IVR	Indulgence vs. Restraint -erfasst die Toleranz gegenüber Genuss, wobei indulgente Gesellschaften Freude fördern und restriktive Gesellschaften Zurückhaltung betonen	Die Erfassung des Punktewertes sowie der Differenz zum durchschnittlichen ESC-Teilnehmer des jeweiligen Jahres erfolgt basierend auf den Werten der Hofstede Insights Website

6.3 Wettbewerbsspezifische Faktoren

Variable	Definition	Erfassungsmethode
xd_prior	Gewinnwahrscheinlichkeit basierend auf Betfair Exchange Odds[6]	5 Zeitintervalle:[7] 6d, 10d, 25d, 35d, 45d
elo_score	Historische Finalperformance basierend auf dem Elo-Rating-System	Kontinuierlicher Wert
elo_score2	Historische Televoting-Performance basierend auf dem Elo-Rating-System	Kontinuierlicher Wert
qual_odds_xd	Qualifikationswahrscheinlichkeit basierend auf Betsson-Quotierungen[8]	4 Zeitintervalle: [9]5d, 9d, 24d, 34d

percentile_rank	Percentile im jeweiligen Semifinale	Normalisiert nach Teilnehmeranzahl

6.4 Andere Externe Faktoren

Variable	Definition	Erfassungsmethode
ldn	Normalisierte Levenshtein-Distanz	ZTM für alle Teilnehmerländer
distance_km	Geographische Distanz zwischen Hauptstädten	Aggregierte Distanzmessung
border_proportion	Anteil gemeinsamer Landesgrenzen zu Teilnehmerländern	Proportion (0-1)

Anmerkungen

[1]Die ZTM (Maße der zentralen Tendenz) umfassen das arithmetische Mittel (Variablensuffix: mean), den Median (Variablensuffix: median), die Standardabweichung (Variablensuffix: std-dev), die Schiefe (Variablensuffix: skewness) und die Kurtosis (Variablensuffix: kurtosis). Zusätzlich wird der DW-Test (Durbin-Watson-Test) zur Messung der Autokorrelation herangezogen, wobei der Variablensuffix durbin-watson verwendet wird. MD=Mahalanobis-Distanz siehe Karl Parthier 2025

[2]Kaggle 2025

[3]Sher 2025

[4]BrandFinace 2025

[5]Insights 2025

[6]Eurovisionworld 2025

[7]bezogen auf Tage vor dem Finale

[8]Eurovisionworld 2025

[9]bezogen auf den Tag vor dem 2. Semifinal(adjustiert um einen Tag)

7 Auswertung der Variablen durch Gradient-Boosted-Classifier

Um die Auswertung der Variablen und deren Wichtigkeit vorzunehmen, gehen wir zunächst auf die verwendeten Methodiken und die mathematischen Grudlagen der Modelle ein.

Decision Tree[1]

Ein Decision Tree ist eine Methode des maschinellen Lernens, die Datenpunkte kontinuierlich anhand bestimmter Parameter aufteilt (splittet). Dabei wird der Prädiktorraum (predictor space) in diskrete, nicht überlappende Regionen segmentiert. Jede Beobachtung, die in eine dieser Regionen fällt, erhält dieselbe Vorhersage wie alle anderen Punkte innerhalb dieser Region. Für Regressionsprobleme ist diese Vorhersage der Durchschnittswert der Zielvariablen in der Region, während bei Klassifikationsproblemen die häufigste Klasse (Mode) als Vorhersage gewählt wird. Die Wahrscheinlichkeit einer Klasse wird dabei als Verhältnis der Anzahl der Punkte dieser Klasse zur Gesamtanzahl der Punkte in der Region berechnet.

Ein Decision-Tree-Algorithmus gehört zur Kategorie CART (Classification and Regression Trees). Typischerweise verwendet dieser Algorithmus ein greedy, top-down-Verfahren, bei dem bei jedem Split der beste lokale Split berechnet wird, ohne Rücksicht darauf, ob dies später zu einer global optimalen Lösung führt. Das Konzept der rekursiven Partitionierung (recursive partitioning) bedeutet, dass der Datenraum in Teilbereiche aufgeteilt wird, die wiederum weiter unterteilt werden, bis ein festgelegtes Kriterium, wie z. B. die Homogenität der Zielvariablen, erfüllt ist.

Die Verlustfunktion (Loss Function), die ein baumbasierter Algorithmus minimieren möchte, ist oft eine Variante von Cross-Entropy oder Gini-Impurity. In unserem Fall, bei einem binären Klassifikationsproblem (qualifiziert = 1, nicht qualifiziert = 0), verwenden wir die Binary Cross-Entropy (BCE). Diese lautet:

$$\text{BCE} = -\frac{1}{n}\sum \left[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)\right] \tag{1.1}$$

Ein Split wird in Abhängigkeit von einem Feature j und einem Schwellenwert t definiert als:

$$S(j,t) = \{\{X \mid X_j < t\}, \{X \mid X_j \geq t\}\} \tag{1.2}$$

Dieser Split teilt die Parent-Region R_p in zwei Child-Regionen R_1 und R_2, basierend darauf, ob die Werte des Features kleiner oder größer (gleich) dem Schwellenwert t sind.

Die optimale Split-Bedingung maximiert die Differenz zwischen dem Verlust der Parent-Region und der Summe der Verluste der Child-Regionen:

$$\max\left(L(R_p) - [L(R_1) + L(R_2)]\right) \tag{1.3}$$

Dabei wird nach Splits gesucht, die die Verluste der Child-Regionen minimieren, um die Gesamtverbesserung im Vergleich zur Parent-Region zu maximieren.

[1] siehe Townshend 2025

8 Regularisierung von Trees

Die Regularisierung von Entscheidungsbäumen ist eng mit dem Bias-Variance-Tradeoff verknüpft. Dieser besagt, dass eine schlechte Abstimmung der Hyperparameter des Modells zu einem unausgewogenen Verhältnis zwischen Bias und Varianz führt:

Bias tritt auf, wenn das Modell zu stark vereinfacht wird und die Daten zu sehr generalisiert. In diesem Fall werden die zugrunde liegenden Beziehungen zwischen den Merkmalen und den Zielvariablen nicht korrekt abgebildet.

Varianz bedeutet, dass das Modell zu komplex ist. Es passt sich den Trainingsdaten zu stark an, was zu Überanpassung führt und dazu, dass das Modell schlecht auf neue Datenpunkte generalisiert.

Um dies zu vermeiden, wird Hyperparameter-Tuning durchgeführt. Dabei wird mithilfe von Cross-Validation (Kreuzvalidierung) die Leistung verschiedener Hyperparameter-Kombinationen bewertet. Cross-Validation ist ein Verfahren, bei dem die Daten in Teilmengen (Folds) aufgeteilt werden:

Die Testergebnisse aus den verschiedenen Folds werden dann in Form des arithmetischen Mittels zusammengefasst. Diese Technik hat den Vorteil, dass das Modell fairer bewertet wird, da es nicht auf demselben Dataset trainiert und evaluiert wird. Ein zusätzlicher Vorteil ist, dass im Vergleich zu einem einfachen Train-Test-Split der Informationsverlust minimiert wird. Die Hyperparameter, die wir genutzt haben, sind die folgenden:

- **n_estimators:** Dieser Parameter ist boostingspezifisch und beschreibt die Anzahl der verwendeten Weak Learners.
- **learning_rate:** Dieser boostingspezifische Parameter skaliert die Gewichtung der Vorhersagen des Baumes auf die Residuen.
- **max_depth:** Dieser Parameter beeinflusst die Tiefe der Baumexpansion, also wie viele Nodes ein Baum maximal haben kann.
- **min_split_gain:** Mindestzugewinn im Fehlermaß, damit ein Split gerechtfertigt ist.
- **min_child_samples:** Dieser Parameter steuert beim Baumaufbau, wie viele Beobachtungen in einem Leaf Node vorhanden sein müssen, damit weiteres Splitten gerechtfertigt ist.
- **bagging_fraction:** Ein Parameter, der die Varianz reduziert, indem zufällige Teilmengen des Trainingsdatensatzes für das Training ausgewählt werden.
- **feature_fraction:** Ein Parameter, der angibt, welcher Anteil der Merkmale für das Training verwendet wird, um die Varianz weiter zu reduzieren.

9 Gradient Boosting

Beim Boosting [2] wird zunächst ein sogenannter weak learner initialisiert. Dabei handelt es sich um ein Modell, das eine hohe Fehlerrate aufweist und viele Residuen hinterlässt. Im Gradient Boosting werden dann baumbasierte Klassifikatoren (tree-based classifiers) auf diese Residuen gefittet. Das kombinierte Modell verbessert sich dadurch, hinterlässt aber weiterhin Residuen. Auf diese werden sequentiell weitere Bäume gefittet.

Der Gradient-Boosting-Algorithmus kann in drei Schritten zusammengefasst werden:

Baue einen schwachen Prädiktor (weak predictor). Passe einen neuen Entscheidungsbaum an die Residuen an, um die Fehler vorherzusagen. Addiere die gewichteten Vorhersagen des Baums, skaliert mit einer Lernrate (learning rate). Wiederhole den Prozess, bis der Fehler ausreichend reduziert ist.

Gradient Boosting beschreibt somit ein additives Modell, bei dem Entscheidungsbäume sequenziell angepasst werden, um die Fehler vorheriger Bäume zu reduzieren. Der Prozess lässt sich mathematisch wie folgt darstellen:

$$\hat{Y} = \hat{F}_k(X_1, \ldots, X_m) \tag{2.1}$$

Dabei gilt:

- $\hat{Y}$ = Vorhersage der Zielvariable (Response).
- $X_1, \ldots, X_m$ = Prädiktorvariablen (Features).
- $\hat{F}_k$ = k-ter schwacher Lerner.

Die Residuen werden wie folgt berechnet:

$$h_k(X_1, \ldots, X_m) = Y - \hat{F}_k(X_1, \ldots, X_m) \tag{2.2}$$

$$\hat{h_k}(X_1, \ldots, X_m) = \hat{F}_{K+1}(h_k(X_1, \ldots, X_m)) \tag{2.3}$$

Dabei ist:

- h_k = Residuen im k-ten Schritt.; $\hat{h_k}$ ist der Schätzer des k-ten Residuals.
- Y = wahrer Wert der Zielvariable.

Im nächsten Schritt gilt:

$$h_{k+1}(X_1, \ldots, X_m) = h_k(X_1, \ldots, X_m) - \hat{F}_{k+1}(h_k(X_1, \ldots, X_m)) \tag{2.4}$$

Das lässt sich verallgemeinern:

[2] Pyrcz 2025 auch Hastie, Tibshirani und Friedman 2009

$$\hat{Y} = \hat{F}(X_1, \ldots, X_m) = \sum_{k=1}^{K} \hat{F}_k(X_1, \ldots, X_m) \tag{2.5}$$

Gradient Boosting beschreibt somit ein additives Modell, bei dem Entscheidungsbäume sequenziell angepasst werden, um die Fehler vorheriger Bäume zu reduzieren. Dieser Modelltyp wurde zunächst auf unser komplettes Featureset sowie der abhängigen Variable "qualified" trainiert. Wir evaluieren das Modell basierend auf einer Randomized Gridsearch mit verschiedenen Hyperparameterkonfigurationen und ließen uns schließlich den Feature-Importance-Plot für dieses Modell ausgeben.

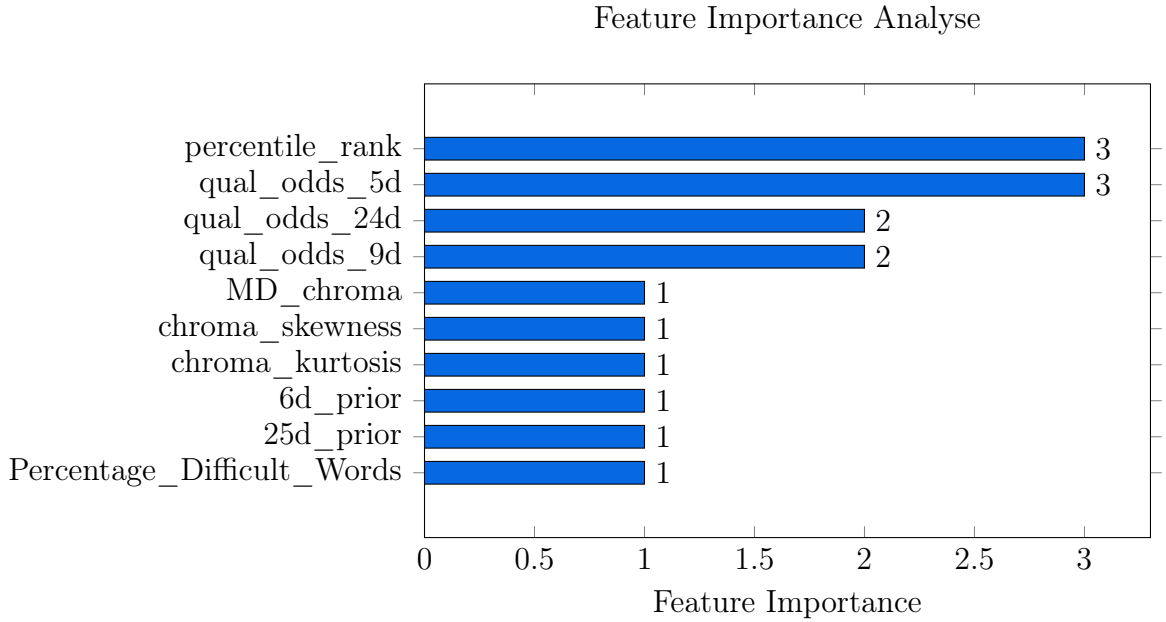

Abbildung 1: Feature-Importance-Diagramm: Darstellung der 10 wichtigsten Merkmale basierend auf der Feature-Importance-Plot

10 Hyperparameter-Tuning

Um die Anzahl der Merkmale weiter zu reduzieren, wendeten wir zunächst die Randomized Grid Search auf alle potenziellen Merkmale an, übernahmen die besten Hyperparameter-Einstellungen und entfernten anschließend Merkmale schrittweise basierend auf dem Feature-Importance-Plot ("one at a time"). Dieser Prozess wurde so lange fortgesetzt, bis sich die Ergebnisse auf Basis einer stratified Cross-Validation nicht weiter verbesserten. Dabei bestand das Cross-Validation-Set aus den Jahren 2018 bis 2023, während wir 2024 als Out-of-Sample-Daten für weitere Tests reservierten. Dieses finale Subset der Merkmale wurde anschließend durch zusätzliches Hyperparameter-Tuning verfeinert.

Das beste Feature-Subset bestand aus den folgenden Merkmalen:

- 'percentile_rank', 'qual_odds_5d', 'qual_odds_24d', 'qual_odds_9d', 'MD_chroma', 'chroma_skewness', 'chroma_kurtosis'

Die optimalen Hyperparameter waren:

- $n\_estimators = 340$
- $min\_split\_gain = 2.2$
- $min\_child\_samples = 3$
- $max\_depth = 6$
- $learning\_rate = 0.5$
- $feature\_fraction = 0.7$
- $bagging\_fraction = 0.96$

Dabei verwendeten wir folgenden Code für die Bestimmung der Hyperparameter auf der Basis von GRID-Search CV bzw. Randomized Grid Search CV:

Die Kreuzvalidierungsmetrik ist definiert als:

$$cv(k) = \frac{1}{k} \sum_{i=1}^{k} BCE_i \tag{2.6}$$

wobei BCE_i die Binär-Kreuzentropie für den i-ten Fold ist und k die Gesamtanzahl der Folds darstellt.

Python-Code

Unten ist die Python-Implementierung der Grid-Search mit LightGBM:

```
from sklearn.model_selection import GridSearchCV,
   RandomizedSearchCV

def grid_search(params, random=False):
    model = lgb.LGBMClassifier(objective='binary')

    if random:
        grid = RandomizedSearchCV(model, params, cv=kfold, n_iter
           =10000, n_jobs=-1, scoring='neg_log_loss')
    else:
        grid = GridSearchCV(model, params, cv=kfold, n_jobs=-1,
           scoring='neg_log_loss')


    grid.fit(X_esc, y_esc)
```

```
    best_params = grid.best_params_
    print("Beste Parameter:", best_params)
    best_score = grid.best_score_
    print("Trainingsscore: {:.3f}".format(best_score))


params = {
    'n_estimators': [5, 10, 15, 20, 40, 60, 80, 100, 120, 140, 160,
        180, 200, 220, 240, 260, 280, 300, 320, 340, 360, 380,
       400],
    'learning_rate': [0.01, 0.05, 0.1, 0.12, 0.14, 0.16, 0.18, 0.2,
        0.25, 0.3, 0.4, 0.42, 0.44, 0.46, 0.48, 0.5, 0.55, 0.6,
       0.65, 0.7],
    'max_depth': [1, 2, 3, 4, 5, 6, 7],
    'min_split_gain': [0, 0.05, 0.1, 0.15, 0.5, 1, 2, 2.2 , 2.4,
       2.6, 2.8, 3],
    'min_child_samples': [2 ,3, 4, 5, 6, 7, 8, 9, 10, 12, 14, 16,
       18, 20, 30, 40],
    'bagging_fraction': [0.6,  0.7, 0.8, 0.9,0.92, 0.94, 0.96,
       0.98, 1],
    'colsample_bytree':[0.2, 0.3, 0.4, 0.5, 0.7, 0.8, 0.9, 0.92,
       0.94, 0.96, 0.98, 1],
    'feature_fraction': [0.6, 0.7, 0.8, 0.9, 0.92, 0.94, 0.96,
       0.98, 1]
}
grid_search(params, random=True)
```

Listing 1: Gitter-Suche mit LightGBM

Um die Performance dieser Methodik im Vergleich zu den Quoten der Buchmacher zu bewerten, setzten wir das zuvor beschriebene Verfahren auf verschiedene Konfigurationen des Cross-Validation-Sets an, wobei wir die ausgewählten Features konstant hielten. Wir begannen das Training mit den Jahren 2018 und 2019 und führten einen Out-of-Sample-Test mit den Daten von 2021 durch. (Das Jahr 2020 wurde aufgrund der COVID-19-Pandemie ausgeschlossen.) In jeder Iteration erweiterten wir das Cross-Validation-Set um ein weiteres Jahr und verschoben den Out-of-Sample-Test um ein Jahr nach vorn. Dabei wurden jeweils neue Hyperparameter-Einstellungen verwendet.
Die Evaluation dieser Methodik wird im Folgenden veranschaulicht.

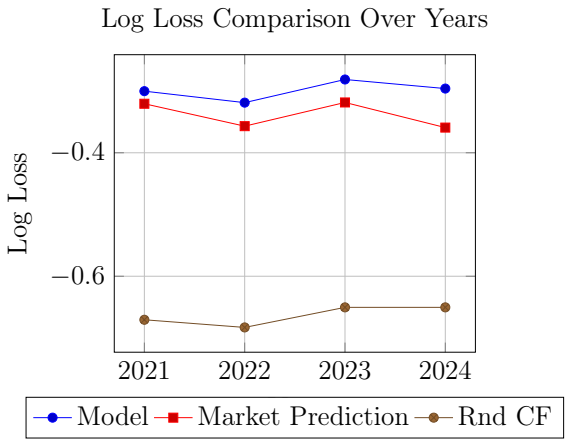

Abbildung 2: Log Loss Comparison

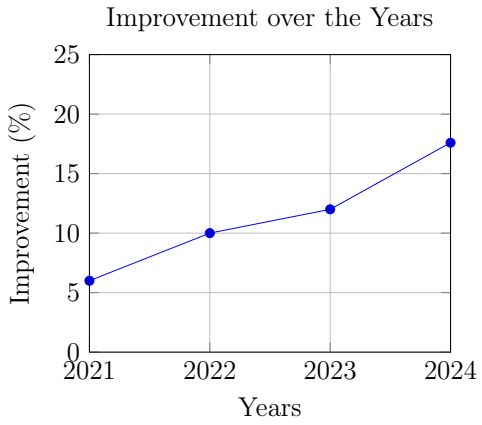

Abbildung 3: Improvement Percentage

11 Ergebnisse:

Die Hypothesen wurden anhand von Feature-Importance-Plots und Modellauswahl analysiert. Die Ergebnisse zeigen, dass die Startreihenfolge sowohl im allgemeinen Feature-Importance-Plot als auch im finalen Feature-Selected-Modell prominent vertreten war. Diese Beobachtung bestätigt die Ergebnisse von Antipov et al. (2017), Bruine de Bruin (2005) sowie Clerides und Stengos (2006). Externe Faktoren, wie Soft Power und verwandte Kategorien, waren in den Modellen nicht signifikant nachweisbar. Dennoch schließt dies die Ergebnisse von Ádalheiðardóttir (2023) nicht aus, die betont, dass Softpower-arme Länder dem Eurovision Song Contest mehr Bedeutung beimessen. Unsere Analyse zeigt jedoch, dass Soft Power keinen direkten Einfluss auf die Qualifikationswahrscheinlichkeit hat.

Obwohl Long-Term Orientation im Feature-Importance-Plot enthalten war, wurde es nicht in das finale Modell aufgenommen. Dies widerspricht Noury und Ginsburgh (2005), die den Einfluss kultureller Dimensionen wie Power Distance und Uncertainty Avoidance in linearen Modellen auf einem Signifikanzniveau von 5 % nachweisen konnten. Unser Modell zeigte hier keinen klaren Effekt. Die normale Levenshtein-Distanz, die sprachliche Ähnlichkeit misst, war nur in Form der Schiefe relevant und wurde nicht in das finale Modell aufgenommen. Dies steht im Gegensatz zu Ginsburgh und Noury (2005), die einen signifikanten Einfluss auf einem 1 %-Niveau nachwiesen.

Performance-Merkmale wie die Anzahl der Künstler und das Geschlecht waren keine signifikanten Prädiktoren. Diese Ergebnisse widersprechen Stephanie Günter (2015), die in linearen Modellen eine hohe Signifikanz bis zu einem 1% -Niveau für die Variable Transgenderfeststellte. Ebenso war der Sprachendummy, speziell für Französisch, nicht relevant. Im Gegensatz zu Aydin(2022) identifizierten wir mehrere signifikante musikalische und lyrische Prädiktoren. Besonders hervorgehoben wurden die Chromascheiefe, Chromakurtosis und die Mahalanobis-Distanz aller Chromamerkmale, die sich im Feature-Selected-Modell als signifikant erwiesen. Die lyrische Analyse, insbesondere der Anteil

schwieriger Wörter, war auf dem Feature-Importance-Plot sichtbar, wurde aber nicht ins finale Modell aufgenommen.

Anders als Andre Bos (2019) fanden wir keinen signifikanten Einfluss von Stimmen aus den Vorjahren in Form einer Lagged Vote auf die Qualifikationsrunde, weder für das ELO-Ranking auf Grundlage der Finalreihenfolge noch für das auf den Halbfinaldaten basierende ELO-System.

Wir konnten hingegen die Signifikanz der Wettquoten im Zusammenhang mit der Qualifikationswahrscheinlichkeit nachweisen. Damit knüpfen wir im weiteren Sinne an die Ergebnisse von Kumpulainen et al. (2020) an, die mittels Sentimentanalyse von Tweets eine starke Korrelation mit dem Abstimmungsverhalten belegten. In diesem Kontext lässt sich argumentieren, dass Quoten im weiteren Sinne als Maß für Sentimentanalysen verstanden werden können, da Buchmacher die kollektiven Einschätzungen und Präferenzen des Publikums effektiv in einer Zahl bündeln.Im Bezug auf den Modellvergleich mit den Quoten der Wettanbieter gelingt es uns, mit der angewandten Methodik jedes Jahr präzise Vorhersagen zu treffen. Erwähnenswert ist dabei, dass die Verbesserung gegenüber den Quoten eine gewisse Korrelation mit der Anzahl der Trainingspunkte im Cross-Validation-Datensatz aufweist (siehe Abbildungen 2 und 3).

12 Ergebnisdiskussion

Einige Diskrepanzen zwischen den Literaturergebnissen und unseren Resultaten lassen sich auf methodologische Unterschiede zurückführen. Im Gegensatz zu linearen Modellen aus früheren Studien verwenden wir Entscheidungsbaumverfahren, welche fundamental anders funktionieren. Zudem könnte der Fokus auf die Qualifikationsrunde, die sich in Teilnehmeranzahl, Voting-System und Allokationsziehung stark vom Finale unterscheidet, zu Abweichungen führen. Die mangelnde Relevanz von Affinitätsmaßen könnte durch Aggregationseffekte erklärt werden, da diese auf einer Gesamtebene statt auf bilateralen Länderpaaren analysiert wurden, was potenzielle Ausgleichseffekte zwischen positiver und negativer Affinität hervorrufen könnte.

In Bezug auf die Analyse der „lagged votes“ lässt sich festhalten, dass sich die Teilnehmer in den Jahren, insbesondere im Semifinale, kontinuierlich ändern. Dies verhindert, dass über die Zeit hinweg stabile Affinitäten zwischen den Ländern aufgebaut werden. Infolgedessen konnten keine langfristigen Affinitäten zwischen den Ländern im Rahmen der untersuchten Jahre identifiziert werden, was die Interpretation der Ergebnisse beeinflusste.

13 Fazit

Diese Arbeit hat verschiedene zuvor untersuchte Faktoren zusammengeführt, in einer gemeinsamen Analyse betrachtet und die Anzahl der berücksichtigten Faktoren erweitert. Dabei konnte gezeigt werden, dass die Modellierung der Qualifikationsrunde möglich ist, wobei das musikalische Merkmal der Tonhöhe(Chroma) eine zentrale Rolle spielt. Die Voting-Bias wurde in unserem Modell ausschließlich durch den Wettbewerbsfaktor der Startreihenfolge sichtbar, während Affinitätsmaße zwischen Ländern keine Relevanz zeigten und externe Faktoren somit insgesamt als nicht bedeutend eingestuft wurden.

Insgesamt konnten wir durch unseren Gradient-Boosted-Classifier die Vorhersagewahrscheinlichkeit über die Quoten von Buchmachern konstant verbessern.

Dennoch weist das Projekt auch einige Limitationen auf. Erstens konzentrierte sich die Analyse ausschließlich auf die Qualifikationsrunde, wodurch die Frage offenbleibt, inwieweit Merkmale, die für die Qualifikationsrunde irrelevant sind, im Finale an Bedeutung gewinnen könnten oder ob relevante Variablen ihre Bedeutung beibehalten. Zweitens stellt die Modellauswahl eine Einschränkung dar: Aufgrund der unterschiedlichen Hyperparameter und der Anzahl der berücksichtigten Merkmale konnten nicht alle möglichen Modellkonfigurationen bewertet werden. Dies könnte potenziell zu einer verzerrten Evaluation führen, die auf einem einzigen Modell basiert.

Weitere interessante Faktoren, die in diesem Projekt nicht berücksichtigt wurden, jedoch potenziell relevant sein könnten, umfassen die Popularität und Beliebtheit eines Künstlers sowie seines Songs. Diese Merkmale wurden bewusst ausgelassen, da es schwierig ist, auf entsprechende Variablen und deren zeitliche Entwicklung zuzugreifen.

Referenzen

Antipov, Evgeny A. und Elena B. Pokryshevskaya (2017). „Order effects in the results of song contests: Evidence from the Eurovision and the New Wave“. In: *Judgment and Decision Making* 12.4, S. 415–419.

Aydin, R. (2022). „Predicting Eurovision scores based only on lyrics and audio features“. Magisterarb. Tilburg University. URL: https://arno.uvt.nl/show.cgi?fid=160779.

Aðalheiðardóttir, Melgar (2022). „Small states and large weak states in the Eurovision Song Contest: Gaining soft power and asserting identity“. Master's thesis. University of Iceland. URL: https://skemman.is/bitstream/1946/43067/4/Small%20States%20and%20Large%20Weak%20States%20in%20Eurovision%20-%20MA%20thesis%20Hera%20Melgar%202022.pdf.

BrandFinace (2025). *Global Soft Power Index*. https://brandirectory.com/softpower. Accessed: 2025-01-10.

Eurovisionworld (2025). *Odds*. https://eurovisionworld.com/odds/eurovision. Accessed: 2025-01-10.

Gatherer, Derek (2006). „Comparison of Eurovision Song Contest simulation with actual results reveals shifting patterns of collusive voting alliances“. In: *Journal of Artificial Societies and Social Simulation*.

Ginsburgh, Victor und Abdul G. Noury (2008). „The Eurovision Song Contest. Is voting political or cultural?“ In: *European Journal of Political Economy*, S. 41–52.

Hastie, Trevor, Robert Tibshirani und Jerome Friedman (2009). „Boosting and additive trees“. In: *The elements of statistical learning: Data mining, inference, and prediction*. 2. Aufl. Springer, S. 337–384.

Insights, Hofstede (2025). *Country Comparison*. https://www.hofstede-insights.com/country-comparison/. Accessed: 2025-01-10.

Kaggle (2025). *Lyrics of every Eurovision song in the history*. https://www.kaggle.com/datasets/minitree/eurovision-song-lyrics. Accessed: 2025-01-10.

Karl Parthier, Joris Parthier (2025). *Implementierung-Boosting-Algorithmen-ESC*. Accessed: 2025-01-14. URL: https://github.com/joris-parthier/Implementierung-Boosting-Algorithmen-ESC.

Magnúsdóttir, Gunnhildur Lily und Baldur Thorhallsson (2011). „The Nordic States and Agenda-Setting in the European Union: How Do Small States Score?“ In: *Stjórnmál og stjórnsýsla* 7.1, S. 203–224.

Millner, Ralf u. a. (2015). *Fair oder Foul? Punktevergabe und Platzierung beim Eurovision Song Contest*. Jena Contributions to Economic Research 2015/2. Ernst-Abbe-Hochschule Jena.

Pyrcz, Michael (2025). *Machine Learning: Gradient Boosting*. https://www.youtube.com/watch?v=___T8_ixIwc. Accessed: 2025-01-10.

Sher, Amir (2025). *The New Dale-Chall Familiar Words List*. https://www.kaggle.com/datasets/amirsher/the-new-dalechall-familiar-words-list. Accessed: 2025-01-10.

Spierdijk, Laura und Michel Vellekoop (2009). „The structure of bias in peer voting systems: lessons from the Eurovision Song Contest“. In: *Empirical Economics*, S. 403–425.

Townshend, Raphael (2025). *Lecture 10 - Decision Trees and Ensemble Methods*. https://www.youtube.com/watch?v=wr9gUr-eWdA. Accessed: 2025-01-10.