Vorhersagen verschiedener Formel 1 Events

Fiete Scheel

6. Januar 2026

Inhaltsverzeichnis

1 Einleitung

1.1 Code und Daten

Der Code der in diesem Projekt verwendet wurde ist unter https://github.com/mcfiet/dl-project zu finden.

1.2 Motivation

Die Formel 1 ist ein datengetriebener Sport, in dem Performance von vielen Faktoren abhängt: Fahrer, Team, Strecke, Startposition und Rennverlauf. Durch die Verfügbarkeit strukturierter Telemetrie- und Ergebnisdaten ist es möglich, solche Einflüsse quantitativ zu modellieren. Vorhersagen sind für Fans, Teams und Medien interessant, weil sie Vergleiche über Saisons hinweg ermöglichen und Entscheidungen (z.B. Strategieeinschätzungen) stützen können.

1.3 Problemstellung

Die zentrale Frage dieser Arbeit ist, ob sich aus vor dem Rennen verfügbaren Informationen eine zuverlässige Vorhersage über Rennergebnisse ableiten lässt. Konkret wird ein Klassifikationsproblem betrachtet: Für jeden Fahrer soll vorhergesagt werden, ob er in einem Grand Prix Punkte erzielt.

Im nächsten Schritt soll ein Regressionsproblem betrachtet werden: So soll dann über die Rennzeit die Position in einem Grand Prix vorhergesagt werden. Damit soll später ein Podium, dann der Sieger usw. vorhergesagt werden. Die Herausforderung liegt in der Heterogenität der Daten (kategorische und numerische Merkmale), der saisonalen Dynamik sowie in der begrenzten Menge an Beispielen pro Saison.

1.4 Ziel dieser Arbeit

Ziel ist der Aufbau eines reproduzierbaren Datensatzes aus FastF1-Daten sowie die Entwicklung und der Vergleich geeigneter Machine-Learning-Modelle für die Punktevorhersage. Dazu werden Merkmale wie Startposition, Qualifying-Performance, Saisonschnitt der Punkte und Streckentypen abgeleitet. Die Modelle werden auf Trainings-, Validierungs- und Testdaten evaluiert und mit geeigneten Metriken (u.a. F1-Score und Balanced Accuracy) verglichen. Abschließend werden die Ergebnisse interpretiert und Limitationen sowie mögliche Erweiterungen diskutiert.

2 Stand der Forschung

In der Forschung gibt es bisher wenige Arbeiten, die sich mit der Vorhersage von Rennergebnissen in der Formel 1 beschäftigen. Es gibt einige private Projekte und Blogs, die sich mit diesem Thema auseinandersetzen (vgl. Mehta, 2025), jedoch fehlt es an wissenschaftlichen Veröffentlichungen, die systematisch verschiedene Ansätze vergleichen und evaluieren. Einige Lösungen nutzen jedoch vergleichbare Datensätze und Merkmale, wie sie in dieser Arbeit verwendet werden. Oft wurde das Problem

mit einem Random Forest Classifier (vgl. Roger, 2024) oder Gradient Boosting Modellen (vgl. Antaya, Mariana, 2025) angegangen, da diese Modelle gut mit tabellarischen Daten umgehen können und weniger anfällig für Overfitting sind.

3 Methodik

3.1 Datengrundlage

Als Datenquelle dient das Python-Paket FastF1, das offizielle F1-Ergebnis- und Qualifyingdaten bereitstellt. Für jede Saison und jedes Rennen werden die Qualifying- und Rennsessions geladen und pro Fahrer ein Datensatz erzeugt. Jede Zeile entspricht damit einem Fahrer in einem Grand Prix. Verwendet werden ausschließlich vor dem Rennen verfügbare Informationen, um Datenleckage zu vermeiden. Die Merkmale umfassen sowohl kategorische Identifikatoren als auch numerische Leistungsindikatoren:

- **Kategorisch:** Fahrer (`driver_id`), Team (`constructor_id`) und Strecke (`circuit_id`).
- **Numerisch:** Startposition aus dem Qualifying (`grid_position`), Qualifying-Deltas zum Sessionsbesten und zum Teamkollegen (`quali_delta`, `quali_tm_delta`), kumulierte Saisonpunkte von Fahrer und Team vor dem Rennen (`season_pts_driver`, `season_pts_team`), gleitender Durchschnitt der letzten drei Rennen (`last_3_avg`) sowie einfache Kontextvariablen für Stadtkurse und Regenrennen (`is_street_circuit`, `is_wet`).

Das Ziellabel ist `points_scored` und markiert, ob ein Fahrer im Rennen Punkte erzielt hat (binäre Klassifikation).

3.2 Modellarchitekturen

Die Daten sind tabellarisch mit einer Mischung aus kategorischen und numerischen Merkmalen. Daher werden mehrere Modellfamilien verglichen: eine logistische Regression als lineare Baseline, ein Random-Forest-Modell als nichtlineares Ensemble sowie Gradient-Boosting-Modelle (XGBoost bzw. LightGBM), die sich in ähnlichen Aufgaben als robust erwiesen haben. Die Boosting-Modelle dienen als Hauptvergleich, da sie nichtlineare Effekte und Interaktionen zwischen Merkmalen effizient abbilden können.

3.3 Trainingssetup

Die Datensätze werden nach Saison getrennt, um zeitliche Leckage zu vermeiden. Im vorliegenden Setup werden die Saisons 2015–2023 als Training, 2024 als Validierung und 2025 als Test verwendet (insgesamt 3740/479/479 Beispiele). Kategorische Variablen werden per One-Hot-Encoding kodiert, numerische Merkmale werden imputiert (Median) und skaliert. Für unausgewogene Klassen wird in den Baumverfahren eine balancierte Gewichtung verwendet. Die Modellselektion erfolgt über die Validierungsdaten, wobei ein Entscheidungsschwellenwert auf dem Validierungs-*F1*-Score

abgestimmt und anschließend auf dem Testdatensatz evaluiert wird. Als Metriken werden u.a. *F1*-Score und *Balanced Accuracy* berichtet.

4 Experimente

4.1 Zusammenfassung

5 Ergebnisse und Diskussion

5.1 Ergebnisse und Diskussionen

6 Fazit und Ausblick

6.1 Fazit und Ausblick

Kritik an der Arbeit

Learnings

Ausblick

Literatur

Antaya, Mariana. (2025). *2025_f1_predictions* [GitHub repository]. Verfügbar 6. Januar 2026 unter https://github.com/mar-antaya/2025_f1_predictions

Mehta, N. (2025). *How I built an F1 race prediction app as my first machine learning project* [Medium blog post]. Verfügbar 6. Januar 2026 unter https://medium.com/@nityachintan/how-i-built-an-f1-race-prediction-app-as-my-first-machine-learning-project-7e2e9cc89826

Roger, W. (2024). *Formula 1 Race Prediction* [Kaggle notebook]. Verfügbar 6. Januar 2026 unter https://www.kaggle.com/code/yanrogerweng/formula-1-race-prediction