5. Juli 2021

Wochenende Challange

Spielvorhersage der English Premier League mit Maschinellem Lernen

von Guillem Boada

ZUSAMMENFASSUNG

Context. Fußball ist eine der berühmtesten Sportarten weltweit und die English Premier League erreicht ein Publikum von ungefähr 4,7 Billion Zuschauer [*Playing the game: The soft power of sport*, British Council]. Ein Grund für das große Interesse an Sportspielen und Fußball eingeschlossen ist ihrer Komplexität und Unvorhersehbarkeit. Viele Menschen versuchen es täglich, die Ergebnisse von Sportereignissen entweder zum Vergnügen oder sogar für akademische und wirtschaftliche Gründe, z.B. beim Wetten, vorherzusagen. Aufgrund dessen werden jährlich neue Studien veröffentlicht, die die neuesten Vorhersagemethoden im Bereich maschinelles Lernen (ML) anwenden, um diese Vorhersage zu optimieren und immer noch bessere Genauigkeit zu schaffen [Raju et al. 2020, Rahman 2020, Baboota et al. 2018].

Ziele. In dieser Arbeit gehe ich diese Herausforderung an, mit dem Ziel eine Bewertung von sowohl den traditionellen ML Modellen wie etwa random forest, logistic regression (LR), support vector machine, naive bayes, k-nearest neighbours und boosting gradients Klassifikator (GB) als auch von moderneren Methoden wie neuronale Netze (NN) zu machen. Da sind zwei Aufgaben: erstens, eine multiclass Klassifizierung von *home wins*, *draw*, und *home loses*; und zweitens, die genauere Vorhersage von der Toranzahl am Ende jedes Spiels. Da ich Daten der Fußballspiele von August 2010 bis März 2021 habe, ist es mein Ziel, die Spielergebnisse von all den Matches vorherzusagen, die im Jahr 2021 bisher gespielt worden sind. Das sind insgesamt 117 Spiele. Daraus ergibt sich ein Trainingssatz von 3724 Spielen.

Methoden. ***Datenvorverarbeitung***. Zuerst habe ich mich mit den fehlenden Werten befasst und es nach der Faustregel entschieden, die Spiele mit fehlenden Daten zu entfernen, weil sie nicht mehr als 5% der Daten ausmachen. Als Nächstes werden die Spielergebnisse für die multiclass Klassifizierung codiert: *home wins* (+1), *draw* (0), und *home loses* (-1). ***Datentransformationen***. Folgend werden die Features je nach Datentyp transformiert, sodass die Modelle sie als Inputs annehmen können. Nach der Testung von verschiedenen Transformationen, habe ich entschieden, die kategorischen Features durch One-Hot Encoding zu codieren und die numerischen Features durch eine Min-Max-Normalisierung zu skalieren. ***Datentrennung***. Beim Training werden Daten gebraucht, um nicht nur die Parameter von den Modellen zu optimieren, sondern auch die Hyperparameter. Dafür wird der Satz von 3724 Spielen in zwei Teile geteilt: die Matches bis zum Jahr 2019 werden als Trainingssatz und die Matches von dem Jahr 2020 als Validationssatz verwendet. ***Training & Vorhersagen***. Zuallererst wird ein Set von multiclass Klassifizierungsmodelle gesammelt, die in scikit-learn verfügbar sind, z.B. random forest, LR und GB. Hier sind auch binäre Klassifizierungsmodelle möglich, die durch die One-vs-Rest heuristische Methode auch als multiclass Klassifizierungsmodelle angewendet werden können. Mit ihren standardmäßigen Konfigurierungen, oder *out-of-the-box*, werden sie mit dem Trainingssatz trainiert und die resultierenden Genauigkeiten werdenn mit dem Validationssatz berechnet. Davon werden die zwei Modelle ausgewählt, die die höchsten Genauigkeiten ergeben haben. Eine Parametersuche wird über die Hyperparameter von jedem Modell gelaufen, um sie feinzustellen. Diese feingestellten Modelle werden mit dem Trainings- und Validationssatz zusammen trainiert. Dadurch profitiert die Optimierung der Modelle von der maximalen Anzahl an verfügbaren Daten. Zum Schluss werden damit die Spielergebnisse vom Testsatz vorhergesagt. Die Vorhersage von der Toranzahl wird statistisch durchgeführt, indem der Medianwert der Toranzahlen zwischen jedem Paar von Teams in den vorherigen Spielen berechnet wird.

Ergebnisse. Die Genauigkeit der standardmäßigen konfigurierten Modelle mit dem Validationssatz finden sie sich in dem Notebook. Die durchschnittliche Genauigkeit beträgt 57,90% mit einer Standardabweichung von ±6,57%. Die besten zwei Ergebnisse wurden von der LR und dem GB mit einer Genauigkeit von 64,21% bzw. 65,49% erzielt. Danach sind Parametersuchen gelaufen, die spezifisch für jedes Modell sind, weil jedes Modell verschiedene Hyperparameter besitzt. Zum Schluss wurde die Genauigkeit von den parametrisierten Modellen mit dem Validationssatz (LR 66,77%; GB 65,50%) und mit dem Testsatz (LR 63,25%; GB 58,12%) berechnet. In allen Fällen liefert die LR die beste Vorhersage. Aufgrund der verschiedenen Datensätze und Bedingungen von jeder Studie, ist es schwer, Ergebnisse zu vergleichen. Trotzdem liegt meine Genauigkeit nahe an den state-of-the-art Ergebnissen, die im Bereich zwischen 60-70% sich bewegen. Die Vorhersage von Toren war bei 5,13% der Spielen korrekt. Mehr Informationen zu diesem Ergebnis finden sich im Notebook.

Diskussion. Da es drei Labels gibt (*home wins*, *draw*, und *home loses*) liegt das Zufallsniveau während der Klassifizierung bei 33,33%. Deswegen wird behauptet, dass die Modelle Informationen von den Daten gelernt haben, weil die durchschnittliche Genauigkeit von 57,90% sich deutlich über dem Zufallsniveau befindet. Dieser Fakt wird auch durch *explainability* begründet, indem die gelernten Parameter von der LR untersucht wurden und angemessen Features wie z.B. *home_shorts_on_target* und *home_possession* mehr Gewicht haben. Dieses Ergebnis wird in Form einer Wordcloud im Notebook dargestellt.

Schlüsselwörter. Machine Learning, predictive analysis, football match prediction, logistic regression, random forest, support vector machine, neural network, gradient boosting