



Duale Hochschule Baden-Württemberg Mannheim

Projektreport

Bundesliga Match Predictions

Studiengang Wirtschaftsinformatik

Studienrichtung Data Science

Verfasser:	Max Bernauer, Philipp Dingfelder, Julius Könning
Matrikelnummer:	5763624, 8687786, 7305370
Firma:	SAP SE, Schaeffler
Modul:	Data Exploration
Kurs:	WWI20DSB
Dozent:	Simon Poll
Bearbeitungszeitraum:	Sommersemester 2022

Ehrenwörtliche Erklärung

Ich versichere hiermit, dass ich die vorliegende Arbeit mit dem Titel “*Bundesliga Match Predictions*” selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Ich versichere zudem, dass die eingereichte elektronische Fassung mit der gedruckten Fassung übereinstimmt.

Ort, Datum

Max Bernauer, Philipp Dingfelder, Julius Könning

Inhaltsverzeichnis

Abbildungsverzeichnis	iii
Tabellenverzeichnis	iv
Abkürzungsverzeichnis	v
1 Zufall und Vorhersehbarkeit im Fußball	1
2 Theoretische Grundlagen	2
2.1 Related Work	2
2.2 Verwendete Technologien und Bibliotheken	2
3 Praktischer Teil	3
3.1 Umsetzung	3
3.2 Ergebnisse und Use Case-Validierung	3
4 Fazit	6
Anhang	

Abbildungsverzeichnis

3.1	Datensatz	3
3.2	Ensemble Learning	4
3.3	Neural Network	4
3.4	Ensemble-Learning	5

Tabellenverzeichnis

Abkürzungsverzeichnis

xxx	Description
-----	-------------

1 Zufall und Vorhersehbarkeit im Fußball

Was macht Fußball so spannend und beliebt? Sepp Herberger hat dazu eine klare Meinung: „Fußball ist deshalb spannend, weil niemand weiß, wie das Spiel ausgeht.“

Dass Sepp Herberger teilweise Recht hat bestätigen immer wieder vermeintliche Underdogs, die den klaren Favoriten in einem Spiel schlagen oder sogar debütieren können. Hierfür gibt es Statistiken, die den Effekt erklären: So fallen 47% aller Tore zufällig, also durch einen Fehler in der Verteidigung oder einen abgefälschten Schuss (.06102018). Außerdem ist jeder dritte Ballkontakt ein unvorhersehbares Ereignis. Es wird also beinahe unmöglich einen perfekten Spielzug mit anschließendem Torerfolg zu planen. (.)

Dennoch gibt es auch Komponenten, die das Gegenteil vermuten lassen. So wissen wir aus Berechnungen, dass die Mannschaft mit dem höheren Marktwert zu 48% gewinnt, das Heimteam zu 45% oder das Team mit der besseren Tabellenplatzierung zu 44%. Inwieweit ist Fußball also etwa durch Machine Learning vorhersehbar oder überwiegt doch die Zufallskomponente? Und ist es tatsächlich möglich Anbieter wie Tipico und Bwin durch ein mathematisches Modell zu schlagen? Diese beiden Fragen versuchen wir im Rahmen der Arbeit zu beantworten.

Das erfolgreiche Vorhersagen von Fußballspielen wäre auch mit einem enormen wirtschaftlichen Nutzen versehen. So kann man durch eine solche Prediction potenziell starken Profit bei etwaigen Wettanbietern erzielen. Andererseits kann man diese Software auch an Wettanbieter verkaufen, damit diese ihre Quoten noch effizienter und genauer berechnen können.

2 Theoretische Grundlagen

2.1 Related Work

Nach Recherche von Vergleichsprojekten wurden einige Ideen aufgegriffen. Beim Preprocessing wurden einige relevante Features hinzugefügt. Begonnen wurde hierbei mit Elo und Angriffs- und Verteidigungsstärke (**Accso.14102021**). Als Vergleichsprojekt wurde ein mathematischer Ansatz gefunden. Dieser ist mit R programmiert und schafft laut Angaben der Quelle 64% accuracy (**Doan.15.3.2019**). Für die grundlegende Idee des Business Use Case wurde eine weitere Quelle herangezogen (**Hartley.05102022**).

2.2 Verwendete Technologien und Bibliotheken

Für die Datenvorbereitung werden klassische Python-Module wie Pandas, Sklearn und Matplotlib genutzt.

Bei der Modellimplementierung wurde vorerst ein Decision Tree Classifier genutzt. Dieser stammt aus dem Sklearn-Modul tree und erreichte keine nennenswerten Ergebnisse. Nach Ausprobieren dieses Classifiers wurde sich für das Ensemble-Learning entschieden. Dabei nutzt man mehrere Algorithmen um sie zeitgleich zu vergleichen und potenziell zu vereinen. Die Algorithmen wie beispielsweise Logistic Regression oder Gaussian Naive Bayes stammen auch von Sklearn. Final wurde ein Neural Network implementiert. Dazu wurde Keras genutzt. Aus Keras wurde außerdem eine Random Parameter Search zur weiteren Optimierung angewandt. Durch Matplotlib wurden die Ergebnisse visualisiert. Außerdem wurde eine Poisson-Verteilung genutzt um einen mathematischen Ansatz zu implementieren.

3 Praktischer Teil

3.1 Umsetzung

Die gewählte Datenbasis enthält alle Bundesliga-Spiele seit der Saison 2005/2006 mit 65+ Features. Während des Preprocessings wurde die Anzahl der Features durch das Löschen nicht relevanter Features deutlich reduziert. Diese Features wurden dann harmonisiert und mit einer Korrelationsmatrix auf ihre Aussagekraft geprüft. Zur Verbesserung des Datensatzes wurden die aktuelle Form der Mannschaften (Elo) sowie Angriffs- und Verteidigungswerte errechnet. Um die Aussagekraft weiter zu erhöhen wurden außerdem Features eingeführt, die einen direkten Vergleich der Mannschaften ermöglichen. Diese Features sind bspw. Differenzen der Elo und Angriffs- sowie Verteidigungswerte oder die Anzahl der Punkte und Tore, die eine Mannschaft während der letzten Spiele erzielt hat.

Date	HomeTeam	AwayTeam	FTHG	FTAG	FTR	HTHG	HTAG	HTR	HS	AS	HST	AST	HF	AF	HC	AC	HY	AY	HR
11.08.2006	Bayern Munich	Dortmund	2	0	H	1	0	H	12	16	9	7	14	20	4	4	1	2	0
12.08.2006	Hamburg	Bielefeld	1	1	D	0	1	A	26	10	19	5	29	22	11	2	3	2	0
12.08.2006	Leverkusen	Aachen	3	0	H	2	0	H	20	8	11	5	22	8	5	1	3	0	1
12.08.2006	Mainz	Bochum	2	1	H	1	0	H	11	19	5	9	17	15	1	8	2	3	0
12.08.2006	M'gladbach	Cottbus	2	0	H	0	0	D	13	11	8	7	11	24	3	5	0	4	0
12.08.2006	Schalke 04	Ein Frankfurt	1	1	D	1	0	H	24	8	13	5	17	19	7	5	1	2	0
12.08.2006	Stuttgart	Nurnberg	0	3	A	0	2	A	10	11	8	8	12	24	5	6	1	2	0
13.08.2006	Hannover	Werder Bremen	2	4	A	1	1	D	14	15	6	8	18	14	2	11	1	3	0
13.08.2006	Wolfsburg	Hertha	0	0	D	0	0	D	26	10	15	4	30	18	10	5	4	1	0
18.08.2006	Nurnberg	M'gladbach	1	0	H	1	0	H	11	11	7	6	19	20	2	5	2	3	0
19.08.2006	Aachen	Schalke 04	0	1	A	0	0	D	12	11	7	7	33	16	5	4	5	2	0
19.08.2006	Cottbus	Hamburg	2	2	D	0	1	A	17	15	10	12	17	17	8	8	1	2	0
19.08.2006	Dortmund	Mainz	1	1	D	0	0	D	17	6	10	4	18	30	7	3	2	1	0
19.08.2006	Ein Frankfurt	Wolfsburg	0	0	D	0	0	D	19	8	9	5	24	31	10	1	5	2	1
19.08.2006	Hertha	Hannover	4	0	H	2	0	H	11	5	9	1	11	18	2	2	1	2	0
19.08.2006	Werder Bremen	Leverkusen	2	1	H	1	1	D	19	21	10	9	20	18	8	13	3	2	0
20.08.2006	Bielefeld	Stuttgart	2	3	A	0	1	A	15	13	9	9	16	25	4	7	1	1	0
20.08.2006	Bochum	Bayern Munich	1	2	A	0	1	A	6	21	4	13	23	25	2	4	1	2	0
25.08.2006	Schalke 04	Werder Bremen	2	0	H	1	0	H	12	8	9	6	23	22	7	10	3	3	0

Abbildung 3.1: Datensatz.

Dadurch, dass die Zielvariablen bereits im Datensatz vorliegen handelt es sich um ein Supervised Learning-Problem. Dieses wurde wie im Theorieteil erwähnt mit drei verschiedenen Methoden angegangen. Dem Ensemble Learning, einem Neural Network sowie einem mathematischen Ansatz.

3.2 Ergebnisse und Use Case-Validierung

Angefangen beim Ensemble Learning erkennt man verschiedene Accuracys je Algorithmus.

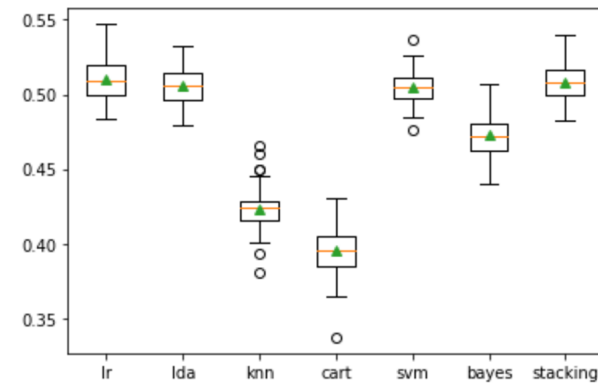


Abbildung 3.2: Ensemble Learning.

Hierbei sind die Logistic Regression und Linear Discriminant Analysis am Besten. Das Stacking Modell welches alle Modelle vereint schneidet auch gut ab. Alle drei liegen bei einer Accuracy von etwas über 50%. Bei Betrachtung des Neural Networks fällt auf, dass die Accuracy schon nach wenigen Epochen auf etwas unter 50% konvergiert.

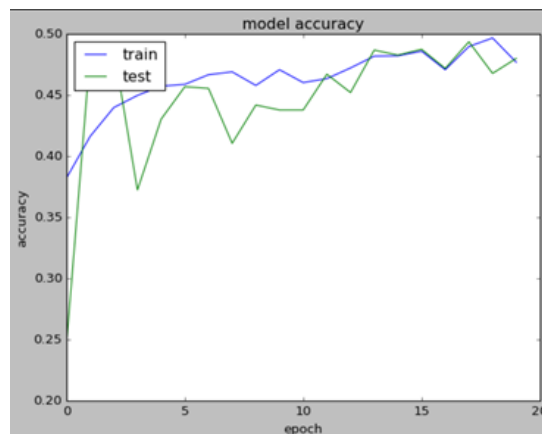


Abbildung 3.3: Neural Network.

Nach einer Random Parameter Search konnte diese auf 53% angehoben werden. Zuletzt haben wir mit der Poisson-Verteilung des mathematischen Ansatzes eine Accuracy von etwa 50 Prozent erreicht. Die Angaben der Quelle konnten somit nicht validiert werden.

Um zu testen, wie gut das Modell unter Realbedingungen funktioniert wurden im Verlauf der Saison 2021/22 Wetten simuliert. Für die Simulation wurden die Predictions des Modells mit den Predictions von vier verschiedenen Wettanbietern verglichen. Um die Predictions der Wettanbieter zu erhalten wurde zuerst der Durchschnitt der Odds der Wettanbieter für jedes Ereignis (Sieg Heimteam, Unentschieden, Sieg Auswärtsteam) ermittelt.

Dieser Durchschnitt wurde dann in eine Prozentzahl umgewandelt und nach dem Maximum gefiltert um eine Prediction für jeden Spieltag zu erhalten. Um auf die Predictions zu setzen mussten diese einen gewissen Schwellwert (XX% Sicherheit) überschreiten. Bei Überschreitung des Schwellwerts wurde je nach Höhe der % Zahl ein bestimmter Betrag gewettet. (Grafik wie viel ab welcher % Zahl gewettet wird)

Auf die gesamte Saison 2021/22 betrachtet ergibt sich folgende Grafik.



Abbildung 3.4: Betting.

Es lässt sich erkennen, dass das Modell am Ende der Saison zwar mehr Gewinn gemacht hat, die Wettanbieter aber im Laufe der Saison besser performen. Auf die Bedeutung der Grafik in Bezug auf den Business Use Case wird im Fazit genauer eingegangen.

4 Fazit

Abschließend lässt sich feststellen, dass sich die Accuracy des Neural Networks im Verlauf des Projekts Schritt für Schritt durch das Optimieren der Datenbasis und des Modells selber erhöhen ließ. Die größte Erhöhung dabei kam durch das Fokussieren auf den direkten Unterschied der Werte der Mannschaften sowie die Random Parameter Search des Modells. Final konnte eine 53% Accuracy erreicht werden.

Wie in der Grafik in Abschnitt drei von Kapitel 3 gezeigt war es möglich mit Hilfe des Modells profitabel zu wetten. Vergleicht man die Profit Kurve des Modells jedoch mit der der Wettanbieter fällt auf, dass diese im Verlauf der Saison Intervalle mit höheren Gewinnen aufweisen, ihre Predictions also besser funktionieren. Das lässt sich darauf zurückführen, dass das Modell mit maximal 65% Sicherheit das Ergebnis eines Spiels vorhersagt. Die Modelle der Wettanbieter weisen eine durchschnittlich höhere Sicherheit in Bezug auf die Ergebnisse der Spiele aus, was dazu führt, dass diese in der Simulation höhere Beträge einsetzen können und insgesamt auch mehr Wetten eingehen als unser Modell. Um also langfristig die Wettanbieter gewinnbringend zu schlagen müsste das Modell weiterverbessert werden um eine höhere Accuracy zu erhalten. Mögliche Ansätze hierfür sind: [Noch einfügen]

Zusammenfassend lassen sich die Ergebnisse des Projekts folgendermaßen darstellen: Das Projekt war relativ erfolgreich. Durch die 53 prozentige Accuracy sowie das halbwegs profitable Wetten konnte gezeigt werden, dass sich Fußballergebnisse zumindest in Ansätzen algorithmisch mit Hilfe von Modellen beschreiben lassen. Damit ist der Use Case zum Großteil erfüllt. Es ist jedoch kritisch anzumerken, dass auch sehr viele nicht genau spezifizierbare Zufallsvariablen einen Einfluss auf den Ausgang eines Spiels haben, sodass sich nie mit 100%er Wahrscheinlichkeit der genaue Ausgang eines Fußballspiels vorhersagen lässt.