Analyse des Bundesligafußballs anhand von Spielstatistiken

Bachelorthesis

Ludwig-Maximilians-Universität

München

Julian Raith

betreut von
Prof. Dr. Christian Heumann

Januar 2020

Zusammenfassung

Diese Arbeit beschäftigt sich damit, durch welche Attribute eine Fußball-Mannschaft kurzfristigen, aber auch langfristigen Erfolg erzielen kann.
Dazu werden Spielstatistiken des Kickers zu allen Spielen der deutschen Fußball-Bundesliga seit der Saison 2013/14 ausgewertet.
Um dies auszuarbeiten, werden für die Vorhersage einzelner Spiele ein (kumulatives) Logit-Modell und ein Random Forest verwendet. Für das Abschneiden einer Mannschaft in einer ganzen Saison werden lineare Modelle angewandt.
Bei einzelnen Spielen zeigt sich der Einfluss von Torschüssen und der Laufleistung als besonders einflussreich. Für den Erfolg über die ganze Saison sind dagegen spielerische Elemente, wie vor allem der Ballbesitz, entscheidend.

Inhaltsverzeichnis

Kapitel 1

Einleitung und Zielsetzung

18 zu 14 Torschüsse, 47 zu 53 Prozent Ballbesitz, 86 zu 84 Prozent Passquote oder 7 zu 5 Ecken [1]. So lauten einige Spielstatistiken eines Fußballspiels aus Sicht von Team A. Würde man rein intuitiv auf Grundlage derer auf das Spiel und dessen Ergebnis schließen wollen, würde man ein ausgeglichenes und enges Spiel erwarten. Tatsächlich aber gehören diese Statistiken zu einem der berühmtesten, weil deutlichsten, Spiele der jüngeren Fußballgeschichte. Das nach den Zahlen vermeintlich gleichwertige, wenn nicht leicht überlegene, Team A stellt dabei Brasilien dar, die Spielstatistiken beziehen sich auf das Halbfinale der WM 2014 gegen den späteren Weltmeister Deutschland, welches Brasilien mit 1:7 verlor.
Orientiert man nach diesem Spiel, scheinen Statistiken im Fußball zunächst bedeutungslos. Es stellt sich die Frage, wie anhand derer auf den Spielverlauf oder das Ergebnis geschlossen werden soll. Da der Fußball in seiner Komplexität kaum vergleichbar mit anderen Sportarten ist, lässt sich allgemein sagen, dass der Statistik im Fußball klare Grenzen gegeben sind.
„Ballbesitz schießt keine Tore“ lautet eines der bekanntesten Sprichwörter im Fußball. Die Entwicklung des modernen Fußballs scheint diese These zu unterstützen, so nimmt die Geschwindigkeit des Fußballs immer mehr zu. Die Laufleistungen der Spieler haben sich enorm gesteigert, auch die Anzahl und Dauer an intensiven Läufen ist deutlich gestiegen (Moebius (2018)). Dadurch fallen Tore meist nach schnellen Umschaltsituationen, als nach lange Ballbesitzphasen. Doch trotzdem spielt die Statistik im Fußball eine immer größere Rolle, die erhobenen Daten zu Fußballspielen nehmen in ihrer Quantität zu und werden auch immer spezifischer.
Erste Ansätze zur Datenerhebung im Fußball fanden bereits 1950 ohne den weitreichenden Einsatz von Technik statt, indem eine Art Protokoll zu Spielen geführt wurde. Einer der bekanntesten „Datenerheber“ Charles Reep kam zu der Erkenntnis, dass der Großteil der Tore nach drei oder weniger Pässen fallen, was ihn zur Schlussfolgerung kommen ließ, dass der lange Ball in Richtung Tor des Gegners die richtige Spielweise sei, was auch oben genanntes Sprichwort bestätigen würde (Pollard (2002)).
Große Schritte in der Datensammlung konnten durch den Einsatz von Videokameras

[1] https://www.kicker.de/1417879/spieldaten/brasilien-920/deutschland

und der Auswertung des Bildmaterials gemacht werden. Wohingegen dies Anfang der 90er-Jahre noch mit Stift und Papier praktiziert wurde, wird das Filmmaterial mittlerweile in Echtzeit softwareunterstützt analysiert. Mit Hilfe von künstlicher Intelligenz und Sensoren wird versucht die Erfassung von Daten immer weiter zu automatisieren und zu optimieren (Schlipsing u. a. (2013)).

Diese Arbeit untersucht, durch welche Attribute eine Mannschaft kurzfristigen, aber auch langfristigen Erfolg erzielen kann. Dazu werden sowohl der Einfluss von Spielstatistiken auf den Ausgang einzelner Spiele als auch das Abschneiden in einer ganzen Saison untersucht. Dazu werden Spielstatistiken aller Bundesligaspiele seit der Saison 2013/14 ausgewertet.
Nach einem kurzen Überblick über die Geschichte und Identität der Bundesliga in Kapitel 2 wird in Kapitel 3 auf die Beschaffung und Aufbereitung der in der Analyse verwendeten Daten eingegangen. In Kapitel 4 soll eine deskriptive Analyse der Daten erfolgen, wobei insbesondere Beziehungen zwischen einzelnen Spielstatistiken genauer durchleuchtet werden. Kapitel 5 und 6 bilden den Hauptteil der Arbeit und widmen sich der Modellbildung zur Untersuchung des Einflusses von Spielstatistiken auf den Ausgang von einzelnen Spielen und das Abschneiden von Mannschaften in einer ganzen Saison.
In Kapitel 5 werden für die Modellierung sowohl Modelle aus der Familie der Regressionsanalyse für Klassifikationen, als auch gängige nichtparametrische Modelle aus dem Bereich des Machine Learning verwendet. In Kapitel 6 werden die Einflüsse saisonaggregierter Spielstatistiken auf das Abschneiden von Mannschaften in einer Saison mittels linearer Modelle untersucht.

Kapitel 2

Die deutsche Bundesliga

2.1 Geschichte der Bundesliga

Am 28. Juli 1962 wurde die Zusammenführung der bis dahin 5 regionalen Ligen zu einer landesweiten Profiliga ab der Saison 1963/64 beschlossen, wodurch die deutsche Fußball-Bundesliga entstand. Nachdem sich bis Ende der 70er-Jahre noch Borussia Mönchengladbach mit dem FC Bayern München um die Spitzenposition der Liga streiten konnte, nahm ab Beginn der 80er-Jahre die Dominanz der Bayern immer mehr zu und gipfelt in der heutigen Zeit durch den Gewinn von 7 Meisterschaften in Folge.
International war die Bundesliga fast immer in der europäischen Spitze zu finden. Am erfolgreichsten waren dabei die 70er-Jahre, vor allem der dreimalige Gewinn des damalig höchsten internationalen Wettbewerbs, dem Europapokal der Landesmeister (Vorgänger der Champions League), in drei aufeinanderfolgenden Jahren des FC Bayern sorgte dafür.
Die Bundesliga konnte auch in den folgenden beiden Jahrzehnten international durchgängig eine gute Rolle spielen. Nach Gründung der Champions League zur Saison 1992/93 war Borussia Dortmund in der Spielzeit 1996/97 der erste deutsche Sieger dieses Wettbewerbs. Nachdem auch der FC Bayern 2001 den Titel holen konnte, begann gegen Mitte der 2000er-Jahre die erste Schwächephase der Bundesliga im internationalen Vergleich (Havemann (2013)).
Doch mit drei Endspielteilnahmen der Bayern innerhalb von vier Jahren zu Beginn der 2010er-Jahre konnte sich die Bundesliga wieder in der internationalen Spitze etablieren, kämpft in den letzten Jahren jedoch darum, diesen Status zu verteidigen. Die damit verbundenen Schwierigkeiten sind auch der Explosion der Ablösesummen und Gehälter durch den vermehrten Einstieg von Investoren und der rasanten Steigerung der medialen Erlöse vor allem in England geschuldet.
Doch hinter der Premier League und der spanischen La Liga kann sich bereits die Bundesliga als Fußball-Liga mit dem dritthöchsten Umsatz nennen. Mit einem Gesamtumsatz von 3,8 Milliarden Euro (DFL (2018)) liegt sie jedoch deutlich hinter dem Spitzenreiter, der englischen Liga mit 6,5 Milliarden Euro (Deloitte (2018)).

2.2 Wofür steht die Bundesliga?

Während die englische Liga für deren harte Spielweise und schnellen Fußball steht, steht die italienische Liga für deren, auch in der Nationalmannschaft etabliertes, Catenaccio, welches zwar nicht mehr in seiner ursprünglichen Form praktiziert wird, sich aber im Laufe der Jahre als Synonym zum defensiv stabilen Fußball entwickelt hat (Duit (2012)).
Die spanische Liga war in den letzten Jahren stark durch den Einfluss der beiden vermutlich besten Fußballer des letzten Jahrzehnts, Lionel Messi und Cristiano Ronaldo, geprägt, die mit ihren zahlreichen Toren für spannende Meisterschaftskämpfe zwischen deren Teams, dem FC Barcelona und Real Madrid sorgten. Lediglich ein Titelträger der letzten 15 Jahre entsprang nicht diesem Zweikampf, in der Saison 2013/14 konnte Reals Stadtrivale Atletico Madrid die beiden in ihrer Dominanz unterbrechen und die Meisterschaft feiern.
So stellt sich die Frage wofür die Bundesliga steht. Schnell fällt die Alleinherrschaft des FC Bayern mit sieben Meistertiteln in Folge auf, welche jedoch vergleichbar mit der von Juventus Turin in Italien und Paris Saint Germain in Frankreich ist. Zusätzlich steht die Bundesliga seit Jahren für volle Stadien und dem höchsten Zuschauerschnitt aller Fußball-Ligen Europas, weltweit ist sie nach den Zahlen von SportingIntelligence (2017) nach der NFL sogar die Liga, die die zweitmeisten Zuschauer in deren Stadien zieht.
Doch es bleibt die Frage danach, wie sich der deutsche Bundesliga-Fußball vom Fußball der anderen Top-Ligen abhebt, wie etwa der englische durch seine Härte oder der italienische durch die Defensiv- bzw. der spanische durch die Offensivkunst. Um dies zu untersuchen sollen im Folgenden Spielstatistiken der letzten Jahre untersucht werden, um so darauf zu schließen, welche Faktoren in der höchsten deutschen Spielklasse zu kurzfristigem und langfristigem Erfolg führen.

Kapitel 3

Datenbeschaffung und Aufbereitung

3.1 Datenherkunft

Vor Beginn der Analyse soll noch vorgestellt werden, woher die Daten stammen. Für Bundesligaspiele bieten zahlreiche Fußball- und Sportapps Statistiken an, welche beispielsweise die Anzahl an Torschüsse, die Laufleistung oder den Ballbesitz angeben. Recherchiert man, woher diese Daten stammen und wie diese letztendlich durch verschiedene Apps bereitgestellt werden, stößt man relativ schnell auf das Unternehmen Opta.
„Opta sammelt, analysiert und vertreibt umfassende Live-Daten des deutschen Profifußballs zu allen Klubs und Spielern. Mit unseren maßgeschneiderten Services können Sie einzigartige Erlebnisse für Fans schaffen.“, heißt es auf der Homepage von Opta [1].
Opta arbeitet mit verschiedenen Medienunternehmen wie Kicker, Sky oder BILD zusammen, welche die Daten letztendlich nutzergerecht für Fußballfans auf deren Apps, Websites oder Zeitschriften präsentieren dürfen.
„Opta verarbeitet detaillierte Daten von einer Vielzahl an Wettbewerben und Sportarten aus der ganzen Welt. Diese Daten werden in Echtzeit erfasst und anschließend über Feeds an unsere Kunden geliefert.“ [1] Trotz der gleichen Datenquelle sind beispielsweise bei Kicker und Bild Unterschiede in den publizierten Spielstatistiken vorhanden, wohingegen Kicker und Sky nahezu identische Zahlen aufweisen.
Neben den Unterschieden in den Zahlen direkt, werden dem Nutzer auch eine unterschiedliche Auswahl an Daten präsentiert. So gibt beispielsweise Sky im Gegensatz zu Kicker oder BILD keine Information über die Laufleistung eines Teams, jedoch über die Anzahl der gelben und roten Karten. Sowohl die Unterschiede in den Zahlen, als auch in den verschiedenen Spieldaten können mehrere Gründe als Ursache aufweisen. Die bereitgestellten Daten können von den Medienunternehmen unterschiedlich interpretiert werden, was ein bekanntes Problem der Spielstatistiken darstellt.

[1] https://www.optasports.com/de/sport/fußball/deutscher-fußball/

So ist zu sagen, dass es im Allgemeinen keine eindeutige Wahrheit darüber gibt, wer einen Zweikampf gewonnen hat oder in manchen Fällen ob es sich bei einer Aktion um einen Pass, eine Flanke oder einen Torabschluss handelt.
Zwar ist „Die Bundesliga [...] übrigens weltweit die einzige Liga, bei der es einen offiziellen Definitionskatalog für Spieldaten gibt. Dort ist haargenau beschrieben, was eine Ballbesitzphase ist, wie ein Pass definiert wird oder wo ein intensiver Lauf aufhört und ein Sprint anfängt.“ (Herrmann (2015)), doch so werden diese Daten lediglich auf der offiziellen Bundesliga-Homepage in einer nicht befriedigenden Weise bereitgestellt: Zum Einen sind - zumindest zum jetzigen Zeitpunkt - erst Daten ab der Saison 2017/18 verfügbar, des Weiteren sind diese Daten nicht in der gewünschten Form und Vollständigkeit vorhanden, sodass eine weitreichende Analyse möglich wäre, auch wenn die Daten zu gewissen Spielen weitreichender wären als die in dieser Analyse verwendeten Daten.
Über die genauen Gründe der Unterschiede in den Daten, vor allem derer mit Datenquelle Opta, können aber nur Mutmaßungen angestellt werden. Die Uneinheitlichkeit kann auch an verschiedenen Rechten in der Zusammenarbeit der Unternehmen liegen.
Nach Prüfung der verschiedenen Datenquellen wurde sich letztendlich für die Daten des Kicker entschieden. Zum Einen, da die verschiedenen Daten am interessantesten schienen – der Kicker zeigt zwar keine gelben und roten Karten auf der Seite der Spielstatistiken an, jedoch z.B. die Laufleistung – zum Anderen bietet der Kicker eine sehr übersichtliche und strukturierte Benutzeroberfläche, welche die Datenbeschaffung mittels des Web Scrapers erheblich erleichterte.

3.2 Datenerhebung

Der Kicker stellt Daten seit der Saison 2013/14 immer in gleicher Art und Form zur Verfügung. Die einzelnen Spielstatistiken wurden mit Hilfe des eben genannten Web Scrapers unter Verwendung des R-Paktes „rvest“ extrahiert. Mittels des „Selector-Gadget: point and click CSS selectors“ [2] wurden auf der Kicker Spieldaten-Website die gewollten Daten ausgewählt und in den Webscraper eingefügt.
Mittels der URLs wurden anschließend alle Statistiken einer ganzen Saison extrahiert. Dies wurde für alle Saisons durchgeführt. Problematisch erwies sich dabei die Ermittlung der URLs der Spieldaten-Websites zu allen Spielen einer Saison, da diese nur in wenigen Fällen fortlaufende URLs waren.
Nach zeitaufwendiger Überprüfung und Sammlung der URLs konnten die Daten letztendlich jedoch in der gewünschten Form extrahiert werden.

3.3 Datenaufbereitung

Die extrahierten Daten wurden pro Saison zu Datensätzen zusammengefügt und weitere Variablen zur eindeutigen Identifikation, wie z.B. die Saison, hinzugefügt,

[2] https://selectorgadget.com

um letztendlich einen Datensatz für alle extrahierten Spiele zu erhalten. Der Datensatz enthält für sechs komplette Saisons und den ersten sieben Spieltagen der aktuellen Saison 2019/20 pro Saison 34 Spieltage und pro Spieltag neun Spiele der 18 Bundesliga-Mannschaften. So ergeben sich $34 \times 9 \times 6 + 7 \times 9 = 1899$ Beobachtungen.
Aus diesen Daten wurden weitere Datensätze erzeugt. So wurde neben einem Datensatz, welcher für jede Zeile die Statistiken einer Mannschaft eines Spiels enthält und somit die doppelte Anzahl an Beobachtungen besitzt, ein weiterer Datensatz erzeugt, welcher saisonaggregierte Daten pro Team pro Saison enthält.
Dazu wurde neben den saisonaggregierten Daten aus den extrahierten Spielstatistiken auch eine Tabelle pro Saison erzeugt, und so zu jeder Beobachtung der Tabellenplatz, die erreichten Punkte und die Tordifferenz am Ende der Saison pro Team hinzugefügt.
Der Tabellenplatz wurde wie auch in der Bundesliga zunächst nach der erreichten Punktzahl, dann nach der Tordifferenz und dann nach der Anzahl an geschossenen Tore gebildet. Diese Kriterien waren für die Daten ausreichend um einen eindeutigen Tabellenplatz zu identifizieren. Es ergeben sich somit für sechs vollständige Saisons pro Spielzeit 18 Beobachtungen für die 18 teilnehmenden Teams einer Saison, also insgesamt 108 aggregierte Daten vollständiger Saisons.
Da in den Spieldaten mehrere Variablen wie Laufleistung oder Torschüsse absolute Werte aufweisen, wurde ein weiterer Datensatz gebildet, um den Zusammenhang der Ausprägungen zur reinen Spielzeit zu verhindern.
Bei Betrachtung der Rohdaten, also ohne Bezug zum Gegner, ergeben sich dabei zwei grundlegende Probleme. Zum einen dauert ein Spiel nicht immer gleich lange, zwar wird durch die Nachspielzeit versucht, durch längere Unterbrechungen verlorene Spielzeit, wie etwa durch Verletzungen oder Wechsel, nachzuholen, jedoch liegt dies im Ermessen der Schiedsrichter und hat keine exakten Regeln als Grundlage. Vor allem bei Spielen, welche bereits nach 90 Minuten einen eindeutigen Spielstand zeigen und entschieden sind, hält sich die Nachspielzeit meist in Grenzen.
Zum anderen ist nicht nachvollziehbar, wie viel die wirkliche Netto-Spielzeit, also die gesamte Spielzeit abzüglich aller auch kleinen Unterbrechungen, vor allem durch Einwürfe, Abstöße oder Freistöße, beträgt. In der Bundesliga liegt die durchschnittliche Nettospielzeit nach Ergebungen von CIES Football Observatory (2018) bei etwa 58,5%. Jedoch variiert diese von Spiel zu Spiel stark.
Vor allem die Laufleistung scheint davon stark abhängig zu sein. Bei nur einer Netto-Spielminute länger, in welcher jeder Spieler z.B. 100 Meter laufen würde, würde sich die gesamte Laufleistung pro Team um 1,1 Kilometer erhöhen.
Diese Effekte wurden mittels der Bildung der Differenzen der Spielstatistiken von Heim und Gastteams versucht zu bereinigen. Damit würde eine Spielminute mehr, in der beide Teams gleich viel laufen, keinen Effekt haben, die Differenz aus der Laufleistung beider Teams würde in diesem Fall konstant bleiben.

3.4 Fehlende Werte

Bei genauer Betrachtung der einzelnen Variablen fiel auf, dass es in der Laufleistung der Saison 13/14 bei drei Spielen fehlende Werte bzw. Nuller gab. Auch in anderen Quellen waren keine Werte für diese Spiele auffindbar. Daher ist davon auszugehen, dass eventuell technische Fehler in der Erhebung der Daten auftraten, wodurch diese nicht verfügbar sind.
Um insbesondere für die saisonaggregierten Daten sinnvolle Werte zu gewährleisten, wurden die Nuller durch die durchschnittliche Laufleistung des jeweiligen Teams in dieser Saison ersetzt. Es wurde sich darauf geeinigt, dass eine damit einhergehende Verfälschung der Varianz minimal und für die Analyse unbedenklich ist, weshalb auf eine Korrektur der Varianz verzichtet wurde.

Kapitel 4

Deskriptive Analyse

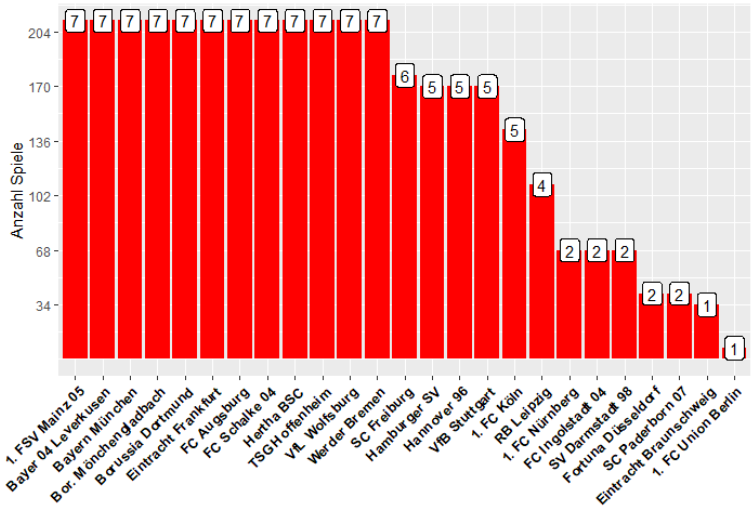
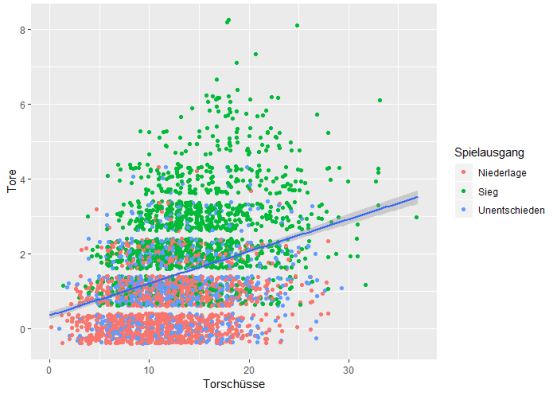


Abbildung 4.1: Anzahl an Spielen und Saisons je Mannschaft in den Daten

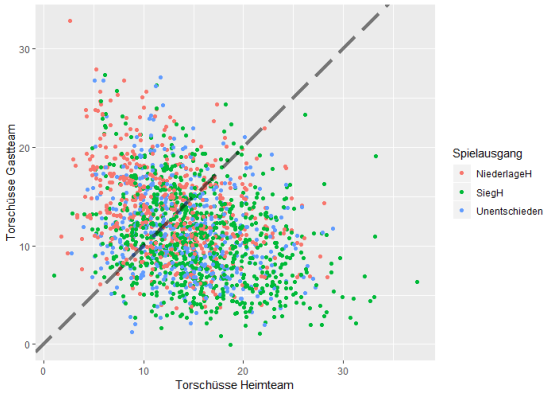
Da eine ausführliche deskriptive Analyse jeder Variable aller Datensätze zu weit führen würde, sollen im Folgenden Zusammenhänge verschiedener Variablen deskriptiv dargestellt werden, bei welchen aus fußballerischer Sicht ein Zusammenhang vermutet werden kann oder welche besonders interessant erscheinen.
Dafür sollen zunächst Variablen aus dem Datensatz der Spieldaten pro Spiel der Bundesliga von der Saison 2013/14 bis 2018/19 und den ersten sieben Spieltagen der Saison 2019/20 betrachtet werden.
Wie in Abbildung 4.1 dargestellt, haben in diesem Zeitraum 25 verschiedene Mannschaften in der Bundesliga gespielt. Zwölf Teams waren dabei in allen sieben Spielzeiten vertreten, Union Berlin spielt erst seit der laufenden Saison 2019/20 in der Bundesliga und hat im Beobachtungszeitraum erst sieben Spiele absolviert und ist daher bei den saisonaggregierten Daten nicht vertreten.

4.1 Torschüsse

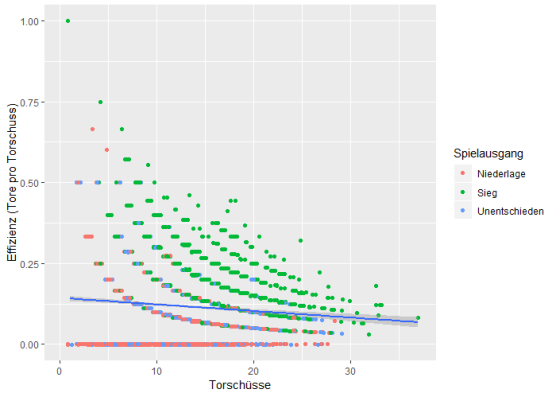
Zunächst soll mit der Anzahl an Torschüssen diejenige Variable betrachtet werden, welche den größten Einfluss auf die Anzahl an Toren, welche selbstredend für eine der später verwendeten Zielvariable, den Ausgang der Spiele, entscheidend ist, vermuten lässt.
Eine bekannte Fußballer-Weisheit besagt: Wer nicht auf das Tor schießt, kann auch kein Tor erzielen. Dieser Aussage kann zunächst anhand der Daten nicht widersprochen werden. Das einzige Spiel im Datensatz, bei welchem eine Mannschaft ohne Torschuss blieb, stellt das Auswärtsspiel von Bremen in München am 8. Spieltag der Saison 2014/15 dar, bei dem die Gäste bei der 6:0 Niederlage bei keinem Versuch aufs Tor auch kein Tor erzielten.
Es ist außerdem keine Beobachtung zu identifizieren, in welcher die Anzahl der Tore die Anzahl der Torschüsse übertrifft, was z.B. aufgrund von Eigentoren durchaus möglich wäre.
Trotzdem stellt sich in diesem Zusammenhang immer die Frage der Effizienz der



(a) Geschätzter linearer Einfluss der Anzahl an Torschüsse auf die Anzahl an Tore in einem Spiel mit Spielausgang



(b) Vergleich der Anzahl an Torschüsse des Heimteams mit Anzahl an Torschüsse des Gastteams in einem Spiel mit Spielausgang aus Sicht des Heimteams



(c) Einfluss der Anzahl an Torschüsse in einem Spiel auf die Effektivität eines Teams, definiert als Tore pro Torschuss, mit Spielausgang

Abbildung 4.2: Torschüsse und Torschusseffizienz

Torschüsse. Dem Hauptstadtklub Hertha Berlin gelang es am 13. Spieltag der Saison 2015/16 aufgrund einer 100-prozentigen Effizienz der Torschüsse bei lediglich einem Torschuss trotzdem drei Punkte einzufahren und das Spiel gegen Hoffenheim 1:0 zu gewinnen.

Die Grafiken zeigen, dass eine erhöhte Anzahl an Torschüssen durchaus den erwartbar starken Einfluss auf die Anzahl an Tore hat. Wird wie in Abbildung 4.2a ein linearer Zusammenhang angenommen, so steigt die erwartete Anzahl an Tore pro zehn Torschüssen um circa eins. Des Weiteren verspricht eine höhere Anzahl an Torschüssen als der Gegner auch eine erhöhte Chance auf den Sieg, wie in Abbildung 4.2b zu sehen ist.
Abbildung 4.2c zeigt, dass die Effizienz der Teams, also die Anzahl der Tore pro Torschuss, mit einer höheren Anzahl an Torschüssen abnimmt. Das lässt sich möglicherweise dadurch erklären, dass Teams mit vielen Torschüssen, welche sich meist aus Spielen mit einer dominanten Mannschaft ergeben, die Konzentration im Abschluss verlieren können und ihre Chancen nicht konsequent nutzen.
Auf der anderen Seite nutzen Teams mit nur wenigen Torschüssen ihre Chancen eventuell konzentrierter.
Zusätzlich kommt eine hohe Anzahl an Torschüssen selbstverständlich erst zu Stande, wenn eine Mannschaft mehr Abschlüsse sucht und auch aus weniger aussichtsreichen Positionen auf das Tor zielt, wodurch sich die Abnahme der Effizienz erklären lässt. Gut zu erkennen ist außerdem der Einfluss der Effizienz auf den Spielausgang. Teams mit einer Effizienz von 0 in einem Spiel, können, selbstredend, maximal ein Unentschieden erzielen.

4.2 Wie kommen Teams zu Torabschlüssen?

So stellt sich die Frage, wie eine Mannschaft zu vielen Torschüssen kommt und somit bei guter Effizienz Tore erzielen kann. Vermuten lässt sich, dass der Ballbesitz oder die Anzahl der Pässe einen hohen Einfluss haben. Wer den Ball lange in den eigenen Reihen hält und viele Pässe spielen kann, versucht sich in der Regel dem Tor anzunähern und den Abschluss zu suchen.

In Abbildung 4.3 zeigt sich ein Trend, dass ein erhöhter Ballbesitz eine erhöhte Torschussanzahl vermuten lässt. Dabei ist auch ein deutlicher Zusammenhang des Ballbesitzes auf die Anzahl an Pässen zu beobachten.

Eine andere Möglichkeit um zu Torschüssen zu kommen stellen Standards dar. So können Freistöße direkt oder indirekt und Ecken nahezu immer indirekt (Ausnahme bilden Ecken, welche durch starken Drall direkt auf das gegnerische Tor geschlagen werden) zu Torschüssen führen. Die Anzahl an Standards soll die Summe aus Ecken und der Variablen Gefoult worden bilden. Auch gegnerisches Abseits hat einen Freistoß zur Folge, dieser ist jedoch immer in der eigenen Hälfte, meist sogar sehr tief in der eigenen Hälfte, auszuführen. Daraus resultieren in der Regel wenige Torschüsse, da diese meist als kurzer Pass ausgeführt werden und nur selten als Flanke Richtung

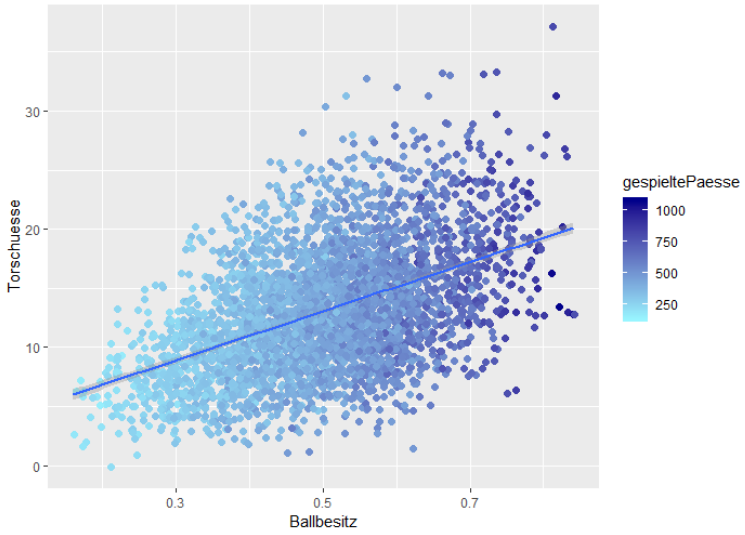
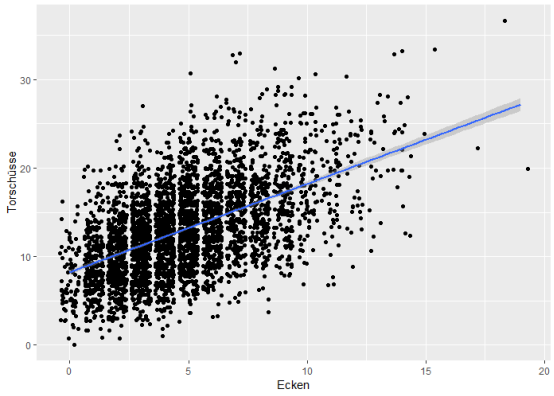
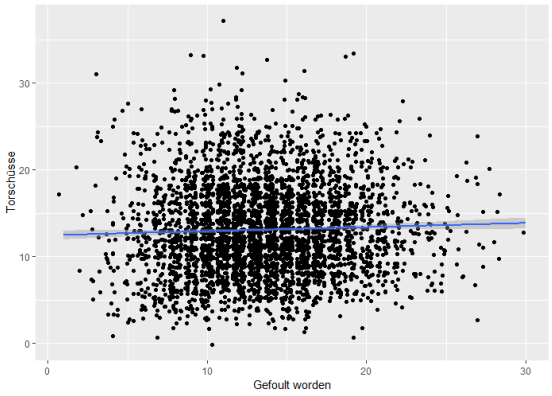


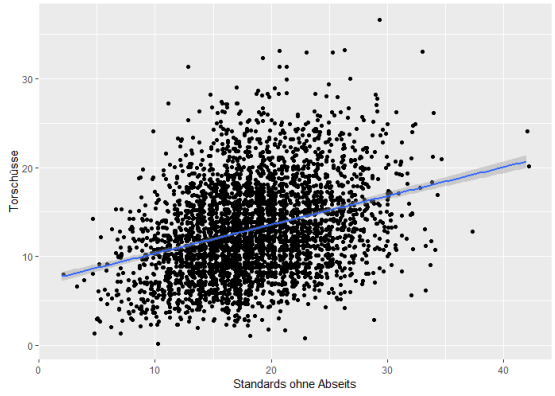
Abbildung 4.3: Einfluss von Ballbesitz auf die Anzahl an Torschüsse in einem Spiel mit Anzahl an gespielten Pässe



(a) Einfluss von Anzahl Ecken auf Anzahl Torschüsse



(b) Einfluss von Anzahl an Gefoult worden, also Freistöße nach gegnerischem Foul, auf Anzahl an Torschüsse



(c) Einfluss von Summe aus Ecken und Gefoult worden auf Anzahl Torschüsse

Abbildung 4.4: Einfluss Standards auf Torschüsse

gegnerisches Tor oder gar als Torschuss ausgeführt werden, weshalb Freistöße nach Abseitsentscheidungen hier nicht beachtet wurden.

Der Einfluss der Gesamtzahl an Standards, also der Summe aus Ecken und Freistößen nach gegnerischem Foul, auf die Anzahl an Torschüsse verwundert zunächst nicht (Abbildung 4.4c). Werden die einzelnen Summanden jedoch separat betrachtet, so überrascht, dass die Anzahl der Freistöße (Abbildung 4.4b) nahezu keinen Einfluss zeigt. Der Einfluss der Summe ergibt sich so fast komplett durch die Anzahl an Ecken (Abbildung 4.4a).
Dies lässt sich möglicherweise dadurch erklären, dass die Anzahl an Freistößen viel mehr durch die Spielweise der Teams bedingt ist. Ein Großteil der Fouls sind vermutlich im Mittelfeld des Spielfeld begangen worden, wodurch sie nur wenige Torschüsse zur Folge haben.
Lediglich Freistöße in Nähe des Strafraums eignen sich für Torschüsse, diese werden jedoch durch das Bewusstsein der damit verbundenen Gefahr versucht zu vermeiden. Durch diese Faktoren ist der sehr geringe Einfluss der Freistöße auf die Torschüsse erklärbar.
Dagegen stellt die Anzahl an Ecken auch einen Indikator für die Offensivstärke einer Mannschaft dar und korreliert auch daher mit der Anzahl an Torschüssen, da eine offensiv starke Mannschaft grundsätzlich viele Ecken und viele Torschüsse zu verzeichnen hat.

4.3 Einfluss von Fouls

Betrachtet man den Einfluss von Gefoult worden, also der Anzahl an Freistößen auf die Anzahl der Torschüsse getrennt nach der Tabellen-Platzierung der Teams genauer, so fällt sofort eine klare Trennung auf. Für eine gute Übersichtlichkeit wurden die Teams nach „gut“ und „schlecht“ getrennt. Gute Teams sind hier Mannschaften, deren durchschnittlicher Tabellenplatz der betrachteten Saisons in den Daten besser als 10 ist, also, welche Teams zur tabellenmäßig besseren Hälfte der Bundesliga zählen.

In Abbildung 4.5 ist zu sehen, dass gute Teams, in diesem Fall vor allem auffällig Bayern München, wenig gefoult werden und, wie zu erwarten, viele Torschüsse abgeben, wohingegen schwächere Teams, wie beispielsweise Ingolstadt, Paderborn oder Braunschweig deutlich häufiger gefoult werden, aber trotz vermeintlich vieler Freistößen wenig Torschüsse zu Stande bringen.
Eine mögliche Erklärung dafür wäre, dass Teams, welche spielerisch limitiert sind und gegen den Abstieg zu kämpfen haben, wie eben genannte Teams, oft zum Mittel des Fouls greifen müssen. Dadurch entwickeln sich auch insgesamt hart geführte Spiele, wodurch auch diese Teams selbst viele Fouls erleiden. Diese Vermutung bewahrheitet sich bei Betrachtung des Einflusses der begangenen Fouls auf die erlittenen Fouls.

Genannte „schlechte“ Teams sind in Abbildung 4.6 in der oberen rechten Ecke

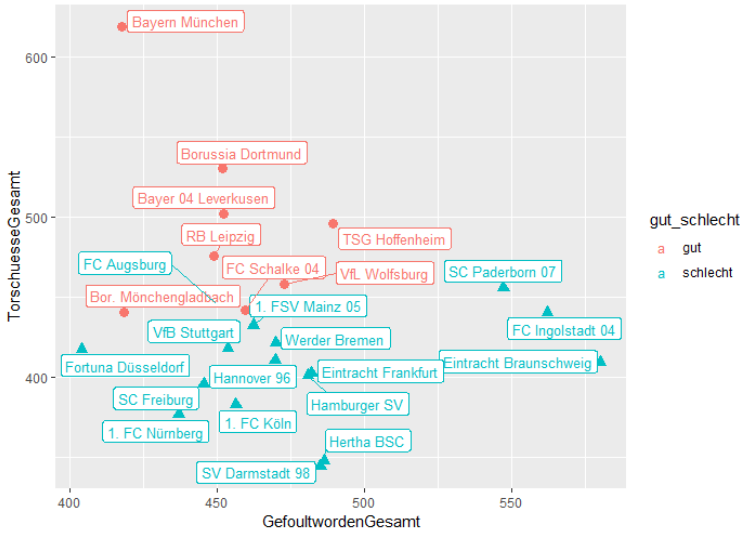


Abbildung 4.5: Einfluss der Gesamtanzahl an Gefoult worden auf Torschüsse in einer Saison gruppiert nach Team und Teams der oberen/unteren Tabellenhälfte (gut/schlecht)

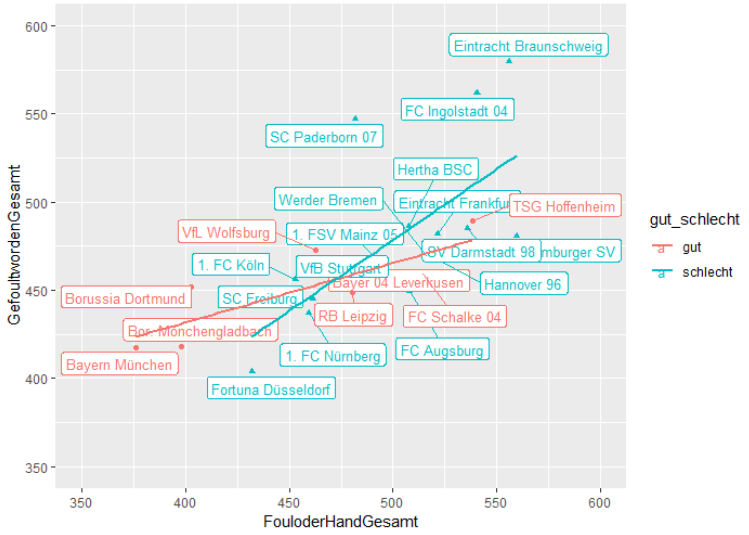
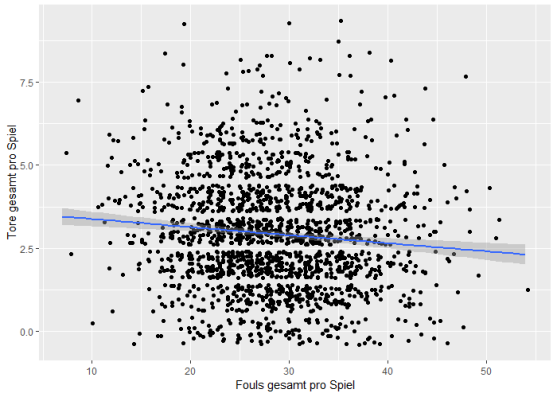


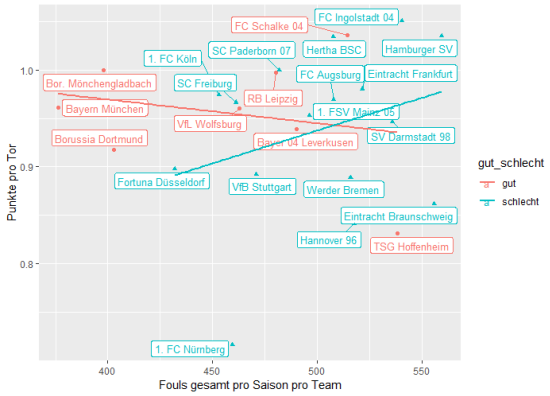
Abbildung 4.6: Einfluss der Gesamtanzahl an begangenen Fouls auf erlittenen Fouls in einer Saison gruppiert nach Team und Teams der oberen/unteren Tabellenhälfte (gut/schlecht)

zu finden also einer hohen Anzahl sowohl an begangenen als auch erlittenen Fouls, Teams wie Bayern, Dortmund oder Gladbach, welche in den letzten Jahren gute Tabellenplätze erreichen konnten, in der linken unteren Ecke, was bedeutet, dass Spiele dieser Teams insgesamt fairer geführt werden.
Als Variable, die in diesem Zusammenhang als latente Variable wirkt, ist zudem die Anzahl der Zweikämpfe zu nennen, zu welcher jedoch keine Zahlen verfügbar sind. Fouls werden bis auf wenige Ausnahmen, wie z.B. Handspiel, in Zweikämpfen begangen, diese geschehen dabei oft auch unabsichtlich. Es ist daher trivial, dass mit steigender Anzahl an Zweikämpfen die Wahrscheinlichkeit für mehr Fouls zunimmt.

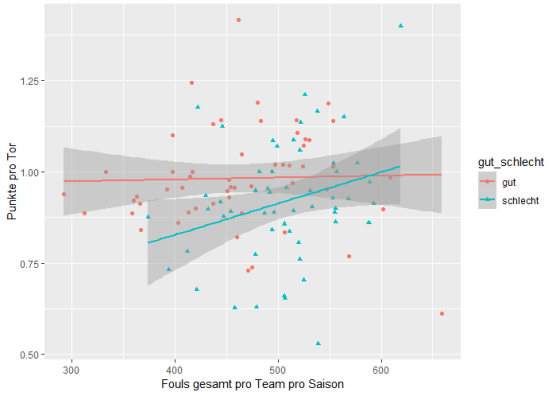
So wurde in der Saison 2015/16 der Aufsteiger Ingolstadt immer wieder für seine „ekelhafte Spielweise“ kritisiert (Weltfussball (2016)). Die Ingolstädter konnten in dieser Saison trotz weniger Tore verhältnismäßig viele Punkte verbuchen, was annehmen lässt, dass eine harte Spielweise zu weniger Toren in einem Spiel und daher mehr erwarteten Punkten pro Tor führt, welcher Einfluss auch in Abbildung 4.7a erkennbar ist.



(a) Einfluss der gesamten Fouls pro Spiel auf die Summe der Tore beider Mannschaften in einem Spiel



(b) Einfluss der durchschnittlichen Gesamtanzahl an Fouls einer Mannschaft pro Saison auf die erreichten Punkte pro Tor gruppiert nach Team



(c) Einfluss der Gesamtanzahl an Fouls einer Mannschaft pro Saison auf die erreichten Punkte pro Tor

Abbildung 4.7: Zusammenhang von Foulspielen und Toren

Bei Spielen mit vielen Fouls leidet oft der Spielfluss und es werden weniger Tore erzielt. So war es für die Ingolstädter trotz insgesamt weniger Tore (33) möglich, viele Punkte (40) einzufahren und letztendlich den Klassenerhalt zu schaffen. Betrachtet man dazu die Anzahl an Punkten pro erzieltem Tor in Abhängigkeit der Anzahl an Fouls, also der harten bzw. weniger harten Spielweise mit vielen Zweikämpfen und Fouls, ist in Abbildung 4.7b und 4.7c zu sehen, dass bei Teams der unteren Tabellenhälfte bei einer Zunahme der Anzahl an Fouls, also einer robusteren Spielweise, ein Tor zu mehr durchschnittlichen Punkten pro Tor führt, was mit dem eben erläuterten stockenden Spielfluss, durch ein foulreicheres Spiel, zu erklären ist.

4.4 Spielstil

Nun soll der Spielstil der Teams genauer betrachtet werden: Darunter sind Variablen zu verstehen, welche Indikatoren für die spielerische Stärke eines Team darstellen. Betrachtet man dazu diejenigen Variablen, welche für Ballbesitz und Pässe stehen, so sind sehr hohe Korrelationen festzustellen. Insbesondere ergeben sich manche Variablen aus Linearkombinationen anderer, wie beispielsweise die Passquote aus dem Quotienten aus angekommenen und gespielten Pässen und die Fehlpässe aus der Differenz zwischen gespielten und angekommenen Pässen. Außerdem korreliert der Ballbesitz sehr stark mit gespielten Pässen und angekommenen Pässen, wie in Abbildung 4.8 zu sehen ist.

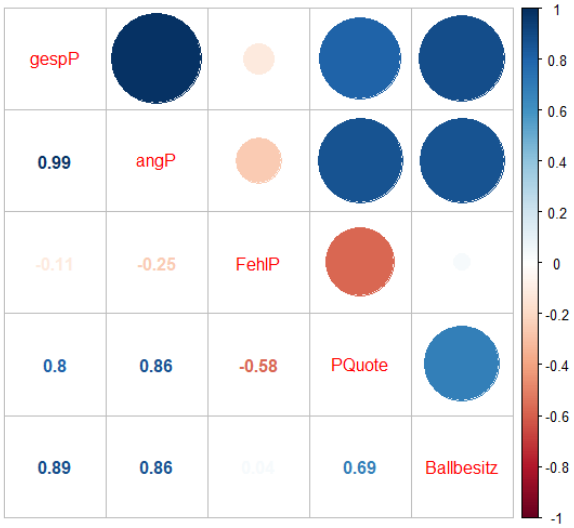


Abbildung 4.8: Korrelationsplot der Variablen, welche den Ballbesitz und sämtliche Passwerte darstellen

Um eine sinnvolle Auswahl aus den Pass- und Ballbesitz-Variablen zu treffen, müssen dazu inhaltliche Überlegungen getroffen werden. Der Ballbesitz sollte vor Allem als Kennzahl der Dominanz im Vergleich zum anderen Team beibehalten werden, dafür ist dadurch etwa die Anzahl an gespielten Pässen hinfällig, da diese, wie erwähnt, eine sehr hohe Korrelation durch einen nahezu linearen Zusammenhang zum Ballbesitz aufzeigt.
Auch die Betrachtung der Fehlpässe scheint nicht ausreichend, da diese in Bezug auf den Ballbesitz an Aussagekraft verlieren: Beispielsweise kann eine Mannschaft, welche in einem Spiel 1000 Pässe spielt und davon lediglich 100 Fehlpässe spielt, die gleiche Anzahl an Fehlpässen aufweisen, wie eine Mannschaft, die von 200 Pässen 100 Fehlpässe spielt. So steht die Anzahl an Fehlpässen nicht wirklich für die Qualität des Spiels einer Mannschaft, wohingegen die Passquote, welche in Beispiel 1 90% und in Beispiel 2 50% betragen würde, einen deutlichen Indikator dafür darstellt, und so geeignet ist, die Qualität des Ballbesitzes einer Mannschaft zu messen. Trotzdem muss die in Abbildung 4.8 erkennbare Korrelation der beiden Größen

weiter beachtet werden, welche auch dadurch zu Stande kommt, dass eine Mannschaft bei hoher Ballbesitz-Quote sicherer wird und prozentual weniger Fehlpässe spielt, bzw. eine hohe Passquote zu weniger Ballverlusten und damit verbundenen Ballbesitzphasen des Gegners führt.

4.5 Laufleistung und Heimvorteil

Zu vermuten wäre, dass Mannschaften, welche viel Ballbesitz haben und dadurch auch viele Pässe spielen können, den Gegner laufen lassen und selbst weniger Laufarbeit verrichten müssen.

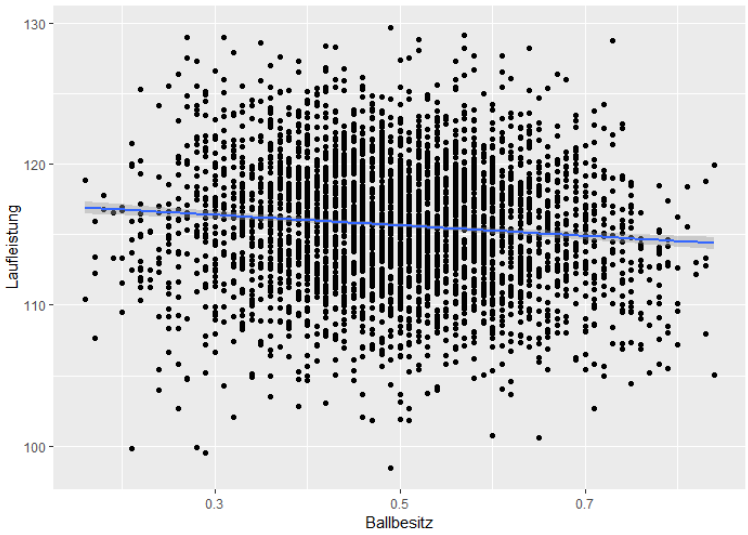


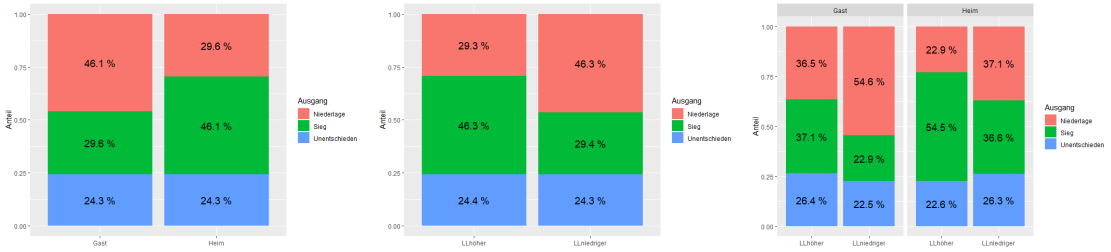
Abbildung 4.9: Einfluss von Ballbesitz auf Laufleistung

Dieser Effekt kann in Abbildung 4.9 beobachtet werden, der Einfluss ist jedoch begrenzt, lediglich bei sehr hohen Ballbesitzquoten sind hohe Laufleistungen nicht zu erkennen, weil diese in derartigen Spielen eventuell für das dominante Team auch gar nicht nötig sind. Plottet man in Abbildung 4.10 die Laufleistung der Heimmannschaft gegen die der Gastmannschaft ist ein klarer Trend erkennbar.

Punkte unterhalb der Winkelhalbierenden, also Spiele bei denen die Heimmannschaft die Gastmannschaft in der Laufleistung übertrifft werden sehr oft von der Heimmannschaft gewonnen. Wird dies genauer betrachtet und nach Spielen unterschieden, in welchen die Laufleistung aus Sicht des betrachteten Teams höher bzw. niedriger als die des Gegners ist, lässt sich dieser Effekt auch deutlich quantifizieren. Insgesamt werden 46.3% der Spiele von einem Team gewonnen, das eine höhere Strecke als das gegnerische Team zurückgelegt hat (Abbildung 4.11b).
So ergibt sich dabei ein Chancenverhältnis von $\frac{46.3\%}{29.4\%} = 1.57$ bei der Chance auf Sieg von Teams deren Laufleistung höher ist im Vergleich zu Teams deren Laufleistung niedriger ist.
Der Vorteil, den man sich dadurch erarbeitet, ist in etwa mit dem des Heimvorteils

Abbildung 4.10: Laufleistung Heim/Gast nach Spielausgang

zu vergleichen, da insgesamt 46.1% der Heimspiele gewonnen werden (Abbildung 4.11a). Das Chancenverhältnis auf Sieg beträgt $\frac{46.1\%}{29.6\%} = 1.56$ im Verhältnis von Heim- zu Gastteams.
Kombiniert man diese beiden Variablen, so bilden sich vier Gruppen (Abbildung 4.11c).



(a) Spielausgang in Prozent getrennt nach Heim/Gast

(b) Spielausgang in Prozent getrennt nach Laufleistung höher/niedriger

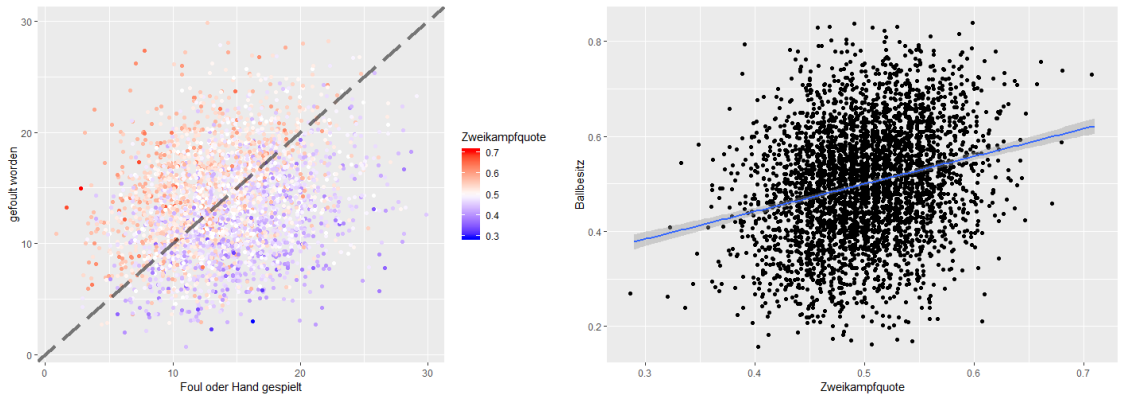
(c) Spielausgang in Prozent getrennt nach Laufleistung höher/niedriger und Heim/Gast

Abbildung 4.11: Spielausgang nach Laufleistung und Heim/Gast

Ist die Laufleistung von Heimteams höher, so steigert sich die Sieg-Quote auf 54,5%, es ergibt sich ein Chancenverhältnis von $\frac{54.5\%}{36.6\%} = 1.49$, also die Chance auf Sieg bei Heimteams, deren Laufleistung größer ist, im Verhältnis zur Chance auf Sieg von Heimteams, deren Laufleistung geringer ist.
Dasselbe auf Auswärtsteams angewandt, ergibt sich ein Chancenverhältnis von $\frac{37.1\%}{22.6\%} = 1.64$ für Gastteams mit höherer Laufleistung als der Gegner im Verhältnis zu Gastteams mit niedrigerer Laufleistung auf Sieg. Der Effekt ist damit bei den Gastteams sogar noch stärker zu beobachten.

4.6 Zweikämpfe

Neben der Laufleistung gilt auch die Zweikampfquote als ein weiterer Indikator für die physische Stärke einer Mannschaft. Beide Variablen beziehen sich auch auf die Intensität des Spiels einer Mannschaft.
Als latente Variable kann hier der Fitnesszustand der einzelnen Spieler und damit der gesamten Mannschaft gesehen werden. Ist eine Mannschaft austrainiert und fit, sind eine höhere Laufleistung als auch eine bessere Zweikampfführung möglich.
Die Zweikampfquote stellt einen entscheidenden Einfluss auf die Anzahl bzw. Differenz an begangenen und erlittenen Fouls dar.



(a) Zusammenhang zwischen begangenen und erlittenen Fouls mit Zweikampfquote

(b) Zusammenhang zwischen Zweikampfquote und Ballbesitz

Abbildung 4.12: Grafiken zu Zweikampfquote

Ein Team mit hoher Zweikampfquote begeht zahlenmäßig und auch verglichen zum Gegner weniger Fouls und erleidet dafür gleichzeitig mehr Fouls verglichen zum Gegner wie in Abbildung 4.12a zu sehen ist. Diese Teams befinden sich oberhalb der Winkelhalbierenden. Sehr hohe bzw. niedrige Zweikampfquoten ergeben sich fast nur, wenn die Differenz an erlittenen und begangenen Fouls hoch ist. Dies ist auch zu erwarten, da ein Zweikampf oft auch durch ein, teilweise auch nicht absichtliches, Foulspiel verloren wird.
Außerdem führt eine hohe Zweikampfquote auch zu einem höheren Ballbesitz wie Abbildung 4.12b zeigt. Werden Zweikämpfe verloren, so hat dies auch meist einen Ballverlust und damit den Wechsel des Ballbesitzes zur Folge.

Kapitel 5

Welche Variablen beeinflussen den Ausgang einzelner Spiele?

5.1 Modellierungsziel

Im Folgenden soll sowohl unter Verwendung von parametrischen als auch nichtparametrischen Modellen untersucht werden, welche Variablen, also welche Spieldaten, den Ausgang von Spielen der deutschen Bundesliga in den letzten Jahren am entscheidendsten beeinflusst haben, um so herauszustellen, welche Spielweise in der Bundesliga zum Erfolg führt.
Außerdem soll ermittelt werden, bei welchem Anteil der Spiele anhand der Spieldaten der richtige Spielausgang (Sieg/Unentschieden/Niederlage bzw. Sieg/kein Sieg) konstruiert werden kann.
Wie in Kapitel 3.3 erläutert, sollen dazu die Differenzen der Spieldaten im Vergleich zum Gegner pro Spiel betrachtet werden, um eine bessere Vergleichbarkeit der einzelnen Beobachtungen zu gewährleisten.
Für alle Modelle werden gleiche Einflussvariablen gewählt, um diese dann in deren Performance gegenüberzustellen. Als Zielvariable wird der Ausgang des Spiels aus Sicht des betroffenen Teams gesehen, auf das genaue Spielergebnis wird keine Rücksicht genommen, was bedeutet, dass es mit Sieg, Unentschieden und Niederlage drei Stufen der kategorialen (geordneten) Zielvariable gibt.
Außerdem werden einzelne Modelle zusätzlich mit der binären Zielvariable Sieg/kein Sieg gerechnet.

5.2 Variablenauswahl

Für die Modelle sollen möglichst alle erhobenen Spielstatistiken verwendet werden, um deren Einfluss zu quantifizieren und zu vergleichen und so die einflussreichsten Variablen zu finden.
Eingeschränkt werden die Passvariablen, wie in Kapitel 4.4 erläutert. Es wird nur der prozentuale Anteil des Ballbesitzes und die Passquote verwendet, um die Quantität und Qualität des Ballbesitzes zu messen.

Außerdem wird die Anzahl der erzielten Tore und die Anzahl der Gegentore nicht betrachtet, da diese offensichtlich einen sehr hohen Einfluss auf den kategorialen Ausgang haben. Vielmehr soll betrachtet werden, durch welche Spielweise ein Team Tore erzielt und so ein Spiel gewinnen kann, wodurch die Anzahl der Tore eher eine Zielvariable als eine Einflussvariable darstellen würde.

	Variable	Ausprägungen
1	Torschussdiff	-31 bis +31
2	Laufleistungdiff	-14.99 bis +14.99
3	Passquotediff	-38.00 bis +38.00
4	BallbesitzBdiff	-68.00 bis + 68.00
5	Zweikampfdiff	-42.00 bis +42.00
6	Fouldiff	-19.00 bis +19.00
7	Gefoultdiff	-20.00 bis +20.00
8	Eckendiff	-19.00 bis +19.00
9	Abseitsdiff	-10.00 bis +10.00
10	HeimGast	Heim/Gast
11	Ausgang	Sieg/Unentschieden/Niederlage
12	Ausgang binär	Sieg/kein Sieg

Tabelle 5.1: Einfluss- und Zielvariablen mit deren Ausprägungen bzw. Range

Die Beziehung zwischen den Spieldaten und dem Spielausgang wird wie folgt formuliert:

$$E(Y_i|x_i) = f(x_i) \tag{5.1}$$

mit $Y_i = Spielausgang$ und $x_i = Differenzen\ der\ Spieldaten\ aus\ Sicht\ des\ beobachteten\ Teams$, und $f(\cdot)$ eine Funktion, bzw. ein Modell, das anhand der in Tabelle 5.1 aufgelisteten Spielstatistiken eine Entscheidung über den Spielausgang fällen soll.

5.3 Modellauswahl

Für die Modellierung sollen Modelle aus der Familie der Regressionsanalyse für Klassifikationen und gängige nichtparametrische Modelle aus dem Bereich des Machine Learning verwendet werden.
Als klassische Regressionsanalyse bietet sich ein kumulatives Logit-Modell an, welches die ordinale Struktur der Daten ausnutzt. Bei der Reduzierung auf eine binäre Zielvariable wird ein klassisches Logit-Modell verwendet. Aus dem Bereich der nichtparametrischen Verfahren zur Klassifizierung wird ein Entscheidungsbaum (CART) und ein Random Forest verwendet, durch welchen die Variable Importance betrachtet werden kann.

5.4 Kumulatives Logit Modell

5.4.1 Theorie

Die folgenden Aufzeichnungen beziehen sich auf Fahrmeir u. a. (2007):
Die Zielvariable $Y_i \in \{Niederlage, Unentschieden, Sieg\}$ ist kategorial und ordinalskaliert. Zusätzlich liegen Kovariablen x_i vor, die nicht von der Kategorie abhängen. Die Zielvariable Y_i wird mit einer latenten Variable U_i durch

$$Y_i = r \iff \theta_{r-1} < U_i < \theta_r, \qquad r = 1, ..., q \tag{5.2}$$

verknüpft, mit $-\infty = \theta_0 < \theta_1 < ... < \theta_c = \infty$. Für U_i gelte ein lineares Modell

$$U_i = -x_i\beta + \epsilon_i \tag{5.3}$$

mit Verteilungsfunktion F. Dann gilt für Y_i das kumulative Modell

$$P(Y_i \leq r|x_i) = F(\theta_r + x_i'\beta), \qquad r = 1, ..., c \tag{5.4}$$

bzw. äquivalent dazu

$$P(Y_i = r|x_i) = F(\theta_r + x_i'\beta) - F(\theta_{r-1} + x_i'\beta). \tag{5.5}$$

Die Wahl von F entscheidet über das resultierende Modell, F ist in den verwendeten Modellen die logistische Verteilungsfunktion, man erhält das kumulative Logit-Modell:

$$F(x) = \frac{\exp(x)}{1 + \exp(x)} \qquad \longrightarrow \qquad P(Y \leq r|x_i) = \frac{\exp(\theta_r + x_i'\beta)}{1 + \exp(\theta_r + x_i'\beta)} \tag{5.6}$$

oder äquivalent dazu:

$$\log \frac{P(Y_i \leq r|x_i)}{P(Y_i > r|x_i)} = \theta_r + x_i'\beta. \tag{5.7}$$

Besondere Eigenschaft des Modells ist, dass die kumulativen Chancen proportional über alle Kategorien sind, da das Verhältnis von zwei durch x_i bzw. $\tilde{x}_i$ charakterisierte Subpopulationen gegeben ist durch

$$\frac{P(Y_i \leq r|x_i)/P(Y_i > r|x_i)}{P(Y_i \leq r|\tilde{x}_i)/P(Y_i > r|\tilde{x}_i)} \tag{5.8}$$

und damit unabhängig von Kategorie r ist.

5.4.2 Praxis

Zunächst soll ein kumulatives Modell mit Hilfe der polr-Funktion (Proportional Odds Logistic Regression) aus dem „MASS“-Paket auf allen in Tabelle 5.1 vorgestellten Variablen gefittet werden. Um die Ergebnisse besser interpretieren zu können, werden die Variablen in Prozentangaben zu ganzen Zahlen transformiert. Aufgrund

	Variable	Value	exp(-value)
1	Torschussdiff	0.0874	0.9163
2	Laufleistungdiff	0.1590	0.8530
3	Passquotediff	0.0110	0.9890
4	BallbesitzBdiff	-0.0055	1.0055
5	Zweikampfdiff	0.0295	0.9709
6	Fouldiff	-0.0359	1.0366
7	Gefoultdiff	-0.0443	1.0453
8	Eckendiff	-0.0596	1.0614
9	Abseitsdiff	-0.0180	1.0181
10	HeimGastHeim	0.3361	0.7145
	Intercepts		exp(value)
1	Niederlage | Unentschieden	-0.4272	0.6523
2	Unentschieden | Sieg	0.7633	2.1453

Tabelle 5.2: $\beta-$Koeffizienten des linearen Prädiktors je Variable in Value-Spalte und exponierter negativer Wert in exp(-value)-Spalte zur Interpretation des kumulativen Logit Modells und Intercepts

der Rundung der Prozentangaben der Daten des Kickers auf ganze Zahlen ergeben sich bei den Differenzen nur gerade Zahlen.

Die Intercepts lassen sich dabei durch die Verwendung der Differenzdaten durchaus sinnvoll interpretieren. Die Chance auf Niederlage im Verhältnis zu einer höheren Kategorie, also zu einem Unentschieden oder einem Sieg, liegt bei 0.652, falls alle Variablen den Wert 0 aufweisen, also ausgeglichene Daten in allen Spielstatistiken vorliegen und die binäre Variable Heim/Gast den Referenz-Wert Gast annimmt.
Die Chance auf Unentschieden oder niedriger im Verhältnis zu höheren Kategorien, also Niederlage und Unentschieden im Verhältnis zu Sieg liegt bei 2.145, falls alle Differenzen den Wert 0 annehmen und ein Gastteam betrachtet wird.
Die $\beta-$Koeffizienten müssen aufgrund der Modellform der polr-Funktion, welche sich durch

$$logitP(Y \leq k|x) = \alpha_k - \eta \tag{5.9}$$

darstellt, negativ interpretiert werden. Dabei steht α_k für die kategorienspezifischen Interceps und η für den linearen Prädiktor ohne Intercept.
Für die Interpretation werden also die exponierten negativen $\beta-$Koeffizienten betrachtet. Die Interpretation lautet dann wir folgt:

- Die Chance auf einen Spielausgang der Kategorie i oder niedriger im Verhältnis zu einer höheren Kategorie
  - fällt mit dem Zuwachs einer bestimmten Variable um 1 um den multiplikativen Faktor $exp(-\beta_k)$, falls $exp(-\beta_k) < 1$.

– steigt mit dem Zuwachs einer bestimmten Variable um 1 um den multiplikativen Faktor $exp(-\beta_k)$, falls $exp(-\beta_k) > 1$.

Als auffällige Variablen ergeben sich dabei die Torschüsse, die Laufleistung und die Heim/Gast Variable, welche allesamt einen multiplikativen Chancenfaktor kleiner 1 besitzen, wodurch sich die Chance auf eine niedrige Ausgangskategorie verringert und im Umkehrschluss auf eine höhere Kategorie erhöht.
Wider Erwarten spielt dagegen etwa der Ballbesitz kaum eine Rolle, was den Schluss zulassen würde, dass das physische Element Laufleistung wesentlich mehr zum Spielausgang einzelner Spiele beiträgt als das spielerische Element Ballbesitz, dessen Einfluss sogar in die andere Richtung zeigt.
Um das Modell zu veranschaulichen, soll auf beispielhaften Daten die Wahrscheinlichkeit je Ausgangskategorie vorhergesagt werden. So ergeben sich für ein Heim-Team, welches im Vergleich zum Gegner 10 Torschüsse mehr abgegeben hat und 10km mehr gelaufen ist und ansonsten in allen Statistiken keine Unterschiede aufweist, die Wahrscheinlichkeiten in Tabelle 5.3. Ein Sieg ist so sehr wahrscheinlich, wohingegen ein Heimteam, welches in allen anderen Statistiken, bis auf Laufleistung und Torschüsse, um 10 überlegen ist (bzw. bei Gefoultwordendiff -10 für realistische Werte), sogar eine höhere Wahrscheinlichkeit für eine Niederlage als einen Sieg aufweist.

	Niederlage	Unentschieden	Sieg
1	3.82%	7.73%	88.45%
2	39.59%	28.72%	31.69%

Tabelle 5.3: Wahrscheinlichkeiten pro Ausgangskategorie des kumulativen Modells auf:
Heimteam, Torschussdiff = 10, Laufleistungsdiff = 10, alle anderen 0 (1)
Heimteam, Torschussdiff = 0, Laufleistungsdiff = 0, Gefoultwordendiff = -10 alle anderen 10 (2)

Um zu sehen, inwieweit anhand der Spielstatistiken auf den Ausgang von Bundesligaspielen geschlossen werden kann, soll ein Train-Test-Split vorgenommen werden, wobei der Train-Datensatz 80 Prozent der Daten beinhaltet und der Test-Datensatz, auf welchem das Modell mit Hilfe der Accuracy evaluiert werden soll, 20 Prozent.
Nach Fitten des Modells auf den Trainingsdaten soll nun mit Hilfe der vorausgesagten Wahrscheinlichkeiten für jede der drei Ausgangskategorien pro Beobachtung entschieden werden, welcher Spielausgang am wahrscheinlichsten ist, indem einfach das Maximum der drei Werte genommen wird, was dann als prognostizierter Wert gesehen wird. Werden die prognostizierten Werte mit den wahren Werten verglichen, so ergibt sich in Tabelle 5.4 eine 9-Felder-Tafel (Confusion Matrix).
Auffällig dabei ist, dass das Modell kein Unentschieden vorhersagt. Sieht man sich dazu die vorhergesagten Wahrscheinlichkeiten an, aufgrund derer das Modell die Entscheidung der Einordnung in die Zielkategorie trifft, so ergibt sich bei den vorhergesagten Wahrscheinlichkeiten für ein Unentschieden ein Maximum von knapp unter 30%, was es bei beliebiger Verteilung der Wahrscheinlichkeiten der anderen beiden Zielkategorien Sieg und Niederlage unmöglich macht, ein Zeilenmaximum zu

erreichen.
Als Accuracy = $\frac{richtig\,eingeordnete\,Spiele}{Gesamtanzahl\,Spiele}$ ergibt sich ein Wert von 53.3%. Das bedeutet, dass der Spielausgang von 53.3% der Spiele des Testdatensatzes anhand der Spielstatistiken in die richtige Zielkategorie eingeordnet werden kann und diese in der 9-Felder-Tafel in Tabelle 5.4 auf der Diagonalen liegen.
Um auch ein Unentschieden vorherzusagen, wurden zwei Lösungsversuche verfolgt:

	Reference		
Predicton	Niederlage	Unentschieden	Sieg
Niederlage	200	88	84
Unentschieden	0	0	0
Sieg	87	96	205

Tabelle 5.4: Confusion Matrix der Vorhersage auf den Test-Daten mit kumulativem Logit-Modell mit drei Ausgangskategorien

Eine mögliche Variante, um mit großer Sicherheit auch Unentschieden zu predicten, ergibt sich, indem der vorhergesagte Ausgang nicht durch die maximale Wahrscheinlichkeit bestimmt wird, sondern die drei Kategorien mit den vorhergesagten Wahrscheinlichkeiten pro Ausgangs-Kategorie ausgewählt werden.
Dabei werden viele Unentschieden predicted, jedoch haben nur etwa 25% davon auch Unentschieden als wahren Wert. Insgesamt verschlechtert sich die Accuracy deutlich auf knapp über 40%.
Ein weiterer Lösungsansatz basiert auf der Vermutung, dass kein Unentschieden vorhergesagt wird, da durch das Vorkommen der negativen Beobachtungen in den Daten aufgrund von Heim und Gast-Teams, die jeweils den Gegner in Partien bildeten, auf jeweils negativen Daten bei einem Unentschieden die gleiche Zielkategorie steht. Um diese Vermutung zu testen, wurde dasselbe Modell jeweils nur auf Heim bzw. Gastdaten gefittet und evaluiert, um keine negative Daten zu erhalten. Doch auch hier wurden keine Unentschieden vorhergesagt.
Insgesamt bleibt festzuhalten, dass der Anteil an Unentschieden in den Daten mit etwa einem Viertel gering ist. Außerdem ist das Unentschieden die mittlere Zielkategorie. So ergibt es auch inhaltlich durchaus Sinn, dass es sehr schwierig ist, Unentschieden vorherzusagen, wenn sich die Zielkategorie durch nur ein Tor mehr in die eine oder andere Richtung verschieben kann.

5.5 Logit Modell

5.5.1 Theorie

Folgende Aufzeichnungen beziehen sich erneut auf Fahrmeir u. a. (2007):
Die binären Zielvariablen $y_i \in \{keinSieg, Sieg\}$ sind 0/1-kodiert und bei gegebenen Kovariablen $x_{i1}, ..., x_{ik}$ (bedingt) unabhängig.

Die Wahrscheinlichkeit $\pi_i = P(y_i = 1|x_{i1}, ..., x_{ik})$ und der lineare Prädiktor

$$\eta_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} = x_i'\beta \tag{5.10}$$

sind durch eine Responsevariable $h(\eta_i) \in [0, 1]$ miteinander verknüpft:

$$\pi_i = h(\eta_i). \tag{5.11}$$

Wird als Linkfunktion der Logit-Link verwendet, so ergibt sich das Logit-Modell:

$$\pi = \frac{\exp(\eta)}{1 + \exp(\eta)} \qquad \Longleftrightarrow \qquad \log \frac{\pi}{1 - \pi} = \eta. \tag{5.12}$$

Mit dem linearen Prädiktor

$$\eta_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} = x_i'\beta \tag{5.13}$$

gilt für die Chance (odds)

$$\frac{\pi_i}{1 - \pi_i} = \frac{P(y_i = 1|x_i)}{P(y_i = 0|x_i)} \tag{5.14}$$

das multiplikative Modell

$$\frac{P(y_i = 1|x_i)}{P(y_i = 0|x_i)} = \exp(\beta_0) \cdot \exp(x_{i1}\beta_1) \cdot ... \cdot \exp(x_{ik}\beta_k) \tag{5.15}$$

Wird z.B. x_{i1} um 1 auf $x_{i1} + 1$ erhöht, so gilt für das Verhältnis der Chancen

$$\frac{P(y_i = 1|x_{i1}, ...)}{P(y_i = 0|x_{i1}, ...)} \Big/ \frac{P(y_i = 1|x_{i1} + 1, ...)}{P(y_i = 0|x_{i1} + 1, ...)} = \exp(\beta_1). \tag{5.16}$$

Zur Interpretation der βs ergibt sich:
$\beta_k > 0$: Chance $P(y_i = 1)/P(y_i = 0)$ wird größer,
$\beta_k < 0$: Chance $P(y_i = 1)/P(y_i = 0)$ wird kleiner,
$\beta_k = 0$: Chance $P(y_i = 1)/P(y_i = 0)$ bleibt gleich.

5.5.2 Praxis

Wie in Absatz 5.4.2 zu sehen ist, konnte ein kumulatives Modell mit den drei Spielausgangskategorien kein Unentschieden vorhersagen, da die geschätzten Wahrscheinlichkeiten zu niedrig waren. Daher soll im Folgenden ein Modell mit binärem Output geschätzt werden, das Unentschieden als alleinige Kategorie ausschließt.

Es ergeben sich die beiden Kategorien kein Sieg(0) und Sieg(1), es wird also die Wahrscheinlichkeit für Sieg modelliert.
Die Richtung der Einflüsse der $\beta-$s ist für alle Variablen gleich der Richtung im kumulativen Modell, sie zeigen also die gleiche Richtung auf den Einfluss für eine höhere Zielkategorie an.

	Variable	Value	exp(value)
1	Intercept	-0.7498	0.4725
2	Torschussdiff	0.0890	1.0930
3	Laufleistungdiff	0.1706	1.1861
4	Passquotediff	0.0135	1.0136
5	BallbesitzBdiff	-0.0059	0.9941
6	Zweikampfdiff	0.0264	1.0268
7	Fouldiff	-0.0354	0.9652
8	Gefoultdiff	-0.0395	0.9613
9	Eckendiff	-0.0631	0.9388
10	Abseitsdiff	-0.0285	0.9719
11	HeimGastHeim	0.2921	1.3393

Tabelle 5.5: $\beta-$Koeffizienten des linearen Prädiktors je Variable in Value-Spalte und exponierter Wert in exp(value)-Spalte zur Interpretation des Logit-Modells mit Intercept

Einen hohen Einfluss zeigen dabei erneut die Torschuss- und Laufleistungsdifferenz und die binäre Heim/Gast-Variable auf. So erhöht sich beispielsweise die Chance auf Sieg um den multiplikativen Faktor 1.19, falls sich die Laufleistungsdifferenz um 1 zu Gunsten des betrachteten Teams erhöht, vorausgesetzt alle anderen Einflüsse bleiben gleich.
Als Output im Logit-Modell wird eine Wahrscheinlichkeit für das Eintreten des Ereignisses 1, also Sieg, modelliert. Um auch auf dieses Modell zu evaluieren sollen Einser vorausgesagt werden, wenn das Modell eine Wahrscheinlichkeit größer von 0.5 vorhersagt und vice versa.
Betrachtet man dazu erneut die Confusion-Matrix in Tabelle 5.6, welche sich hier

	Reference	
Predicton	kein Sieg	Sieg
kein Sieg	391	166
Sieg	80	123

Tabelle 5.6: Confusion Matrix der Vorhersage auf den Test-Daten mit Logit-Modell mit binärer Ausgangskategorie

durch eine Vier-Felder-Tafel ergibt, so ergibt sich eine Accuracy von 67,6%. Auffällig dabei ist, dass das Modell lediglich in einem Viertel der Fälle eine 1, also einen Sieg vorhersagt, wohingegen 37,8% der Daten einen Sieg als Output haben. Verschiebt man die Grenze, ab welcher Wahrscheinlichkeit ein Sieg vorausgesagt werden soll, so kann, abhängig von Train-Test-Split, eine leicht höhere Accuracy erreicht werden, was in Abbildung 5.1 zu sehen ist.

Dabei gibt die Sensitivität an, wie viel Prozent der vorhergesagten Einser wahr sind, stellt sich also durch $\frac{TruePositive}{TruePositive+TrueNegative}$ dar. Diese ist besonders hoch, wenn der Cut-Off hoch gewählt wird, so werden nur bei hohen Wahrscheinlichkeiten für einen Sieg auch ein Sieg vorhergesagt, jedoch nimmt dadurch selbstverständlich auch

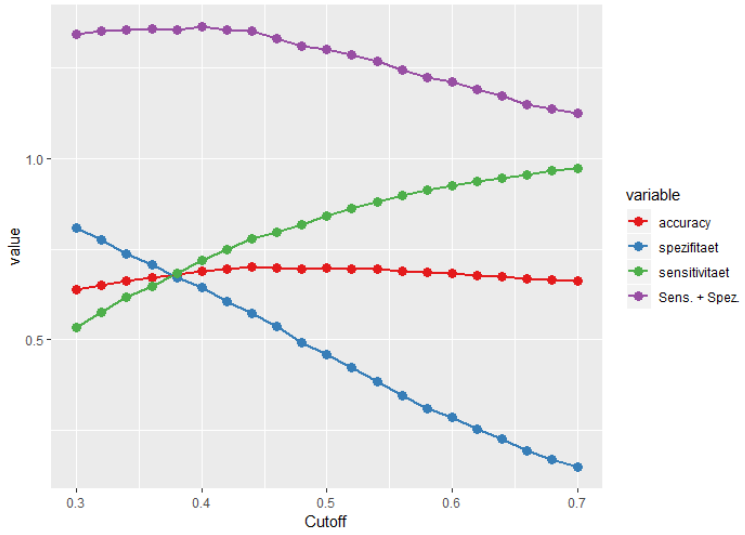


Abbildung 5.1: Accuracy, Sensitivität und Spezifität bei verschiedenen Cut-Off

die Anzahl an Sieg-Vorhersagen ab.
Die Spezifität stellt den Anteil der richtig vorhergesagten Nuller dar, was sich durch $\frac{TrueNegative}{TrueNegative+FalsePositive}$ darstellen lässt. Der Verlauf ergibt sich entgegengesetzt zur Sensitivität.

5.6 CART und Random Forest

Als Vergleich zu oben verwendeten klassischen Regressionsanalysen sollen nun Verfahren aus dem Bereich des Machine Learnings verwendet werden. Zunächst soll ein einfacher Klassifikations- und Regressionsbaum mit Hilfe der R-Funktion „rpart“ aus dem gleichnamigen R-Paket verwendet werden und anschließend ein Random Forest mit Hyperparametertuning aus dem R-Paket „random forest“ um die Klassifikationsgüte zu steigern. Ziel ist es erneut, diejenigen Variablen zu finden, welche die Klassifikation entscheidend beeinflussen.

5.6.1 Theorie

Folgende Aufzeichnungen beziehen sich auf Breiman (2017) und Hastie und Tibshirani (2011):
Man beobachtet Merkmalsvektoren $x_1, ..., x_n$ der Objekte $a_1, ..., a_n$. Ziel ist es auf eine der kategorialen Variablen $Y_1, ..., Y_n$ zu schließen, wodurch sich ein Klassifikationsproblem ergibt. Die populärste Methode dafür sind Classification and Regression Trees (CART). Bei Entscheidungs- und Klassifizierungsbäumen wird der p-dimensionale Merkmalsraum durch rekursives Partitionieren sukzessive in Teilmengen des Merkmalsraums aufgeteilt. Diese Teilmengen des Merkmalsraums sollen so gebildet werden, dass sie bezüglich der Zielvariable möglichst homogen sind.
Grundprinzip CART:

- Betrachte in jedem Schritt binäre Splits, d.h. in jedem Schritt wird eine (bereits gebildete) Teilmenge weiter in genau zwei Teile aufgeteilt.
- Betrachte in jedem Split genau eine Variable, die einen neuen Split bestimmt.
- Das Resultat ist eine disjunkte Zerlegung des Merkmalsraums, die in einer Baumstruktur dargestellt werden kann.

So enthält der Baum zum Start die gesamte Population, davon ausgehend wird in jedem Schritt die bestmögliche Aufteilung bezüglich Y gesucht. Um zu erkennen, welche Variable die höchste Aussagekraft besitzt und anhand welcher der nächste Split vollzogen wird, werden Maße verwendet, die die „Unreinheit“ eines Knotens A_q quantifizieren. Dazu wird in dieser Analyse der Gini-Index verwendet (siehe (Breiman u. a., 1984)). Es wird diejenige Einflussgröße x_i gewählt, für die das Maß minimal ist. Dies wird wiederholt, bis alle Knoten vollkommen homogen sind oder ein Stoppkriterium erfüllt ist. Dazu können verschiedene Stoppkriterien betrachtet werden:

- minimale Anzahl an Beobachtungen pro Blatt
- minimale Verbesserung der Zuordnung im nächsten Schritt
- maximale Anzahl an Baumlevels

Es ergibt sich ein leicht verständliches und interpretierbares Modell, das sich graphisch darstellen lässt, jedoch große Instabilität der Bäume bei Änderungen in den Daten aufweist.
Um ein stabileres Modell zu erhalten, wird außerdem ein Random Forest verwendet, welcher eine Weiterentwicklung des Bagging darstellt, kurz für Bootstrap Aggregating, ein Verfahren, das verschiedene Regressions- bzw. Klassifizierungsmodelle mittelt und daraus ein gleich gewichtetes Modell erstellt (Breiman (1996).

- Ziehe B Bootstrap Stichproben (mit Zurücklegen) aus den vorhandenen Daten und fitte B einzelne Modelle
- Vorhersage für neue Beobachtungen x^*: Aggregiere die Ergebnisse aus den B Modellen für x^*, z.B. durch averaging (metrische Daten) oder majority vote (kategoriale Daten)

So werden beim Random Forest statt einem (CART) gleich mehrere Tress für gegebene Trainingsdaten betrachtet. Um Kollinearität zu vermeiden, wird eine Zufallskomponente eingeführt:

- Konstruiere Baum anhand einer Bootstrap Stichprobe aus dem originalen Datensatz.
- Fitte einen Baum, wobei aus den existierenden M Kovariablen bei jedem Split nur m << M zufällig ausgewählte Merkmale zur Auswahl herangezogen werden.

- Jeder Baum wird so groß wie möglich gezüchtet.

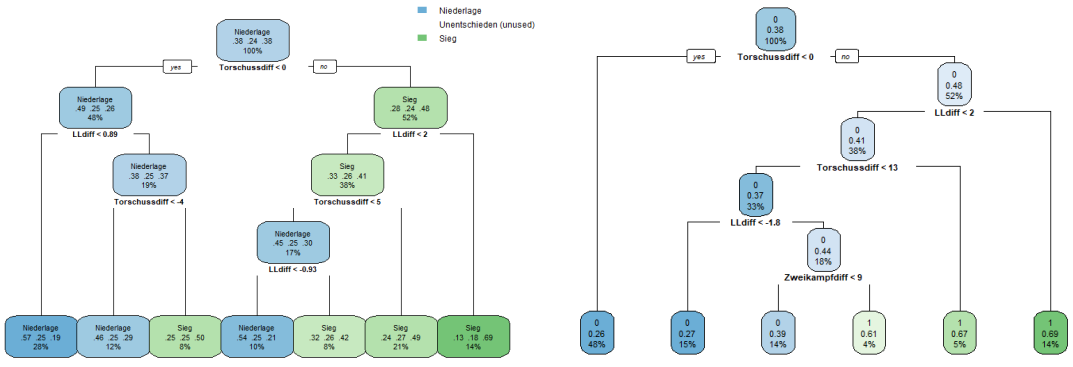
Die Klassifikation erfolgt dann in diesem Modell zu kategorialen Daten anhand des häufigsten Wertes.

5.6.2 Praxis

Es wurde erneut der Trainings- und Testdatensatz verwendet, welche sich in der Verteilung der Zielvariable kaum unterscheiden. Für das Erstellen des CART können dem control-Argument der Funktion *rpart* mittels *rpart.control* verschiedene Werte übergeben werden:

- minsplit: Minimale Anzahl an Beobachtungen, die in einem Knoten vorhanden sein müssen, damit ein Split zustandekommt.
- minbucket: Minimale Anzahl an Beobachtungen in einem Endknoten.
- cp: Komplexitätsparameter. Es werden nur Splits vorgenommen, welche die/den allgemeine(n) Unreinheit/Mangel, um den faktor cp reduzieren.
- maxnodes: Legt die maximale Tiefe des endgültigen Baumes fest, wobei der Stammknoten als Tiefe 0 gezählt wird.

Zunächst wird ein Baum mit den Default-Parametern erstellt (siehe dazu RDocumentation [1]). So ergibt sich ein Baum, welcher lediglich die Variablen Torschussdifferenz und Laufleistung für die Verästelung des Baumes verwendet. Diese Variablen minimieren das Unreinheitsmaß und scheinen so am einflussreichsten für die Entscheidung des CART.



(a) Entscheidungs- und Klassifikationsbaum mit allen drei Ausgangskategorien

(b) Entscheidungs- und Klassifikationsbaum mit binärer Zielvariable Sieg(1)/kein Sieg(0)

Abbildung 5.2: CARTs

[1] https://www.rdocumentation.org/packages/rpart/versions/4.1-15/topics/rpart.control

Dieser Baum ist nun einfach zu interpretieren, indem den verschiedenen Verästelungen gefolgt wird. Folgt man in Abbildung 5.2a immer der rechten Abzweigung, ergibt sich für ein Team, das eine Torschussdifferenz höher 0 und eine Laufleistungsdifferenz größer 2 aufweist, eine Wahrscheinlichkeit von 69% für Sieg, 18% für Unentschieden und 13% für eine Niederlage. Der Anteil der Beobachtungen in diesem Endblatt beträgt dabei 14%.
Würde man das Komplexitätsparameter sukzessive heruntersetzen, so wird als nächste Variable die Zweikampfdifferenz in den Baum integriert.
Auch bei diesem Modell taucht das „Problem“ der Nicht-Vorhersage von Unentschieden wieder auf, es ergibt sich als unused category, was bedeutet, dass keiner der verschiedenen Baumpfade zur Entscheidung auf Unentschieden leitet, da die Wahrscheinlichkeit für Unentschieden in den Endblättern stets niedriger als eine der anderen beiden Zielkategorien ist.
Auch bei diesem Verfahren kann eine Confusion-Matrix erzeugt und eine Accuracy

	Reference		
Predicton	Niederlage	Unentschieden	Sieg
Niederlage	200	82	84
Unentschieden	0	0	0
Sieg	87	102	205

Tabelle 5.7: Confusion Matrix der Vorhersage auf den Test-Daten mit CART mit drei Ausgangskategorien

berechnet werden, welche sich mit 53,3% etwa im selben Bereich bewegt wie die des kumulativen Modells, auch die Einordnung in die drei Ausgangskategorien ist sehr ähnlich.
Wird die Zielvariable erneut binärisiert, und zwischen Sieg/kein Sieg unterschieden, so ergibt sich in Abbildung 5.2b ein Baum mit ähnlichen Entscheidungskriterien. Bei Verwendung des default-Werts des Komplexitätsparameter von 0.01 wird hier bei Fitten des Modells auf dem Trainingsdatensatz bereits die Zweikampfquote mitaufgenommen. Evaluiert auf den Test-Datensatz ergibt sich eine Accuracy von 65.3%. Es werden mit 22.5% sogar noch weniger Siege vorhergesagt als im Logit-Modell.

	Reference	
Predicton	kein Sieg	Sieg
kein Sieg	398	191
Sieg	73	98

Tabelle 5.8: Confusion Matrix der Vorhersage auf den Test-Daten mit CART mit binärer Ausgangskategorie

Um die Instabilität gegenüber kleinen Änderungen in den Daten eines Entscheidungsbaumes zu reduzieren, werden nun Random Forests verwendet, welche sich zusätzlich auch sehr gut eignen, die Wichtigkeit der verschiedenen Einflussvariablen

auf die Modellbildung zu messen.
Um das Modell zu evaluieren, wird Kreuzvalidierung angewandt, wobei der Trainingsdatensatz in eine bestimmte Anzahl k an gleich großen Teilen, genannt „folds“, gesplittet wird. Das Modell wird dann auf $\frac{k-1}{k}$tel der Daten trainiert und auf $\frac{1}{k}$tel der Daten evaluiert. Das wird für jeden Split der Daten wiederholt. Am Ende wird die Accuracy über alle k Splits gemittelt.
Die Anzahl der Splits wurde auf 5 festgelegt, wodurch je 20% der Daten zum Testen benutzt werden. Diese Entscheidung für ein relativ kleines k beruht auch auf dem nicht übermäßig großen Datenumfang, um nicht zu kleine folds zu erhalten.
Wird mit diesem Setup ein Random Forest trainiert und auf die Accuracy getestet, so ergeben sich für verschiedene voreingestellte Werte von *mtry* ein Maximum von 52.6% Accuracy für $mtry = 2$, *mtry* steht dabei für die Anzahl an Variablen, die in jedem Split als Kandidaten nach dem Zufallsprinzip ausgewählt werden.
Nun soll anhand von Hyperparametertuning versucht werden, diese Accuracy weiter zu steigern. Dazu wird zunächst der optimale Wert an *mtry* gesucht, indem wie bei allen folgenden Parametern auch verschiedene Werte bei Festhaltung der anderen (bereits optimierten) Werte in R getestet werden. Der *mtry*-Wert wird hier beispielsweise über alle möglichen Werte getestet, welche sich bei 10 Variablen von 1-10 ergeben, wobei sich ein optimaler Wert von 3 herausstellt. Als maximale Tiefe des endgültigen Baumes (*maxnodes*) ergibt sich ein Wert von 33. Als Mindestgröße der Endknoten (*nodesize*) sorgt ein Wert von 14 für die optimale Accuracy. Für den Parameter *ntree*, der Anzahl an gezüchteten Bäumen, wird 250 als bester Wert ermittelt.
Nun kann unter Verwendung der gewählten Parametern ein Random Forest gefittet

	Reference		
Predicton	Niederlage	Unentschieden	Sieg
Niederlage	201	86	88
Unentschieden	0	0	0
Sieg	86	98	201

Tabelle 5.9: Confusion Matrix der Vorhersage auf den Test-Daten mit Random Forest mit drei Ausgangskategorien

werden. Die Accuracy kann dadurch auf einen Wert von 57.0% gesteigert werden. Das Modell auf den Testdatensatz angewandt, führt zu einer Accuracy von 52.9%, wobei wie in den vorherigen Modellen keine Unentschieden predictet werden, wie Tabelle 5.9 zeigt.
Im Vergleich dazu ergab sich für den Random Forest mit default-Parametern, also ohne Hyperparamtertuning, eine Accuracy von 0.493, wobei dieses Modell in Summe sogar 41 Unentschieden predictet, davon jedoch nur 10 auch wirklich diese Ausgangskategorie besitzen, wodurch erneut der Schluss gezogen werden kann, dass eine Vorhersage von Unentschieden nicht befriedigend dargestellt werden kann.
Von vorrangigem Interesse ist jedoch weiterhin der Einfluss der verschiedenen Spielstatistiken auf die Modellbildung, welche durch die Variable Importance dargestellt

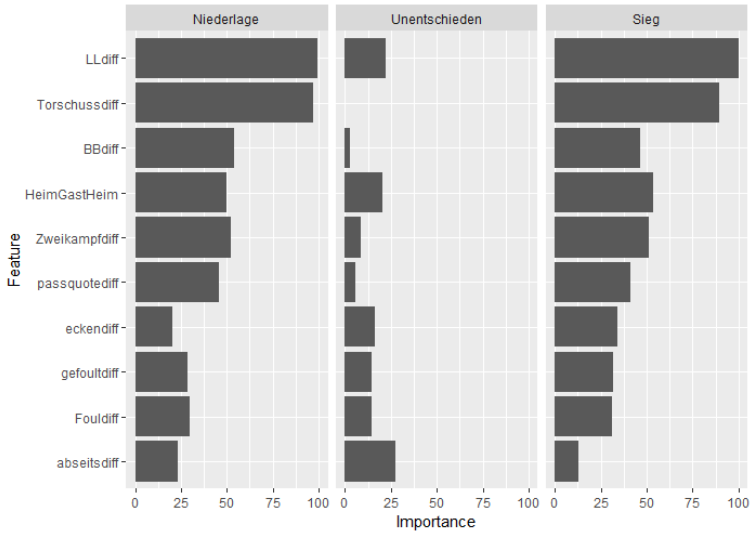


Abbildung 5.3: Variable Importance in % je Variable je Ausgangskategorie des Random Forest

werden soll. Für die hier verwendete kategoriale Zielgröße erhält man kategorienspezifische Variable Importances, welche in Abbildung 5.3 dargestellt sind. Diese werden durch die Summe der Fehlerabnahme bei Split nach einer Variable berechnet (für genauere Beschreibung siehe Strobl u. a. (2008)).
Die relative Wichtigkeit ergibt sich durch die Wichtigkeit einer Variable geteilt durch die Wichtigkeit der Variable mit dem höchsten Wichtigkeits-Wert, wodurch sie auf 0 bis 1 normiert ist.
Es ist zu erkennen, dass erneut die Differenz in der Laufleistung und die Torschussdifferenz deutlich als einflussreichste Variablen zu erkennen sind. Für die Kategorie Unentschieden hat keine Variable eine Importance von höher als 30%, was jedoch aufgrund der Nicht-Vorhersage von Unentschieden auch nicht sinnvoll interpretierbar ist. Die Importance unterscheidet sich in ihrem Einfluss auf Sieg und Niederlage in der Reihenfolge der restlichen Variablen, wobei jedoch keine deutlichen Unterschiede in den Zahlen erkennbar sind. Nach den beiden wichtigsten Variablen folgen die Variablen Heim/Gast, Ballbesitzdifferenz und die Differenz der Zweikampfquote.

Kapitel 6

Welche Variablen beeinflussen das Abschneiden eines Teams in der kompletten Saison?

Im Folgenden sollen Modelle gebildet werden, welche das Abschneiden eines Teams über die komplette Saison hinweg vorhersagen und anhand derer erneut auf die Wichtigkeit von einzelnen Variablen geschlossen werden soll, um zu evaluieren, welche Wichtigkeit verschiedene Spielstatistiken in Bezug auf den Erfolg eines Teams über die ganze Saison besitzen. Dazu werden saisonaggregierte Einflussvariablen verwendet, welche alle Spiele einer Mannschaft einer Saison zusammenfassen. Um den Erfolg einer Mannschaft zu messen, könnten verschiedene Zielvariablen verwendet werden. Möglich sind die geschossenen und kassierten Tore bzw. die Tordifferenz einer Mannschaft über die ganze Saison, der Tabellenplatz oder die Anzahl an erreichten Punkten. Es wurde sich für die Anzahl an Punkten einer Mannschaft pro Saison entschieden, da diese im Gegensatz zu den Toren, welche nicht abbilden können, ob eine Mannschaft auch Spiele gewinnen konnte, eine klare Aussagekraft über den Erfolg einer Mannschaft besitzen. Im Gegensatz zum Tabellenplatz, welcher keine Aussagekraft über den Abstand zweier Tabellennachbarn gibt - diese können punktgleich sein, ebenso aber viele Punkte auseinanderliegen - und damit lediglich eine Ordinalskala besitzt, können diese auf einer Verhältnisskala eingetragen werden.
Außerdem sollen verschiedene Scores gebildet werden, welche die Stärke einer Mannschaft in verschiedenen Bereichen abbilden, um auch auf diesen ein Modell zu fitten und zu evaluieren, welche Merkmale am erfolgversprechendsten sind. Dabei werden folgende Kategorien gebildet:

Kategorie	Variablen
Offensiv	Tore, Torschüsse
Defensiv	Gegentore, gegnerische Torschüsse
Physik	Laufleistung, Zweikampf
Spielerisch	Ballbesitz, Passquote

Tabelle 6.1: Scores-Kategorien mit zugehörigen Spielstatistiken

6.1 Theorie

Die Daten sind gegeben durch $(y_i, x_{i1}, ..., x_{ik}), i = 1, ..., n$ zu einer metrischen Variable y und metrisch oder binär kodierten kategorialen Regressoren $x_1, ..., x_k$.
Als Modellform ergibt sich:

$$y_i = \beta_0 + \beta_1 x_{ik} + ... + \beta_k x_{ik} + \epsilon_i, \qquad i = 1, ..., n. \tag{6.1}$$

Die Fehler $\epsilon_1, ..., \epsilon_n$ sind unabhängig und identisch verteilt (i.i.d.) mit

$$E(\epsilon_i) = 0, \qquad Var(\epsilon_i) = \sigma^2. \tag{6.2}$$

Die geschätzte lineare Funktion

$$\hat{f}(x_1, ..., x_k) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_k x_k \tag{6.3}$$

kann als Schätzung $\hat{E}(y|x_1, ..., x_k)$ für den bedingten Erwartungswert von y bei gegebenen Kovariablen $x_1, ..., x_k$ angesehen und damit zur Prognose von y verwendet werden. Diese wird wieder mit $\hat{y}$ bezeichnet.

(Fahrmeir u. a. (2007))

6.2 Praxis

6.2.1 Saisonaggregierte Daten

Zunächst wurden die Spielstatistiken für bessere Interpretierbarkeit nach der allgemein zulässigen Form $X' = a + bX$ transformiert. Sie wurden durch 34 geteilt, um so die durchschnittlichen Daten pro Spiel zu erhalten, außerdem wurden die Prozent-Daten zusätzlich wie in vorherigen Analysen auf ganze Zahlen transformiert, sodass z.B. die Steigerung der Variable Ballbesitz um 1, die Steigerung des durchschnittlichen Ballbesitzes eines Teams einer ganzen Saison um einen Prozentpunkt bedeutet. Werden die Koeffizienten des Modells mit allen aggregierten Spielstatistiken als Einflussgrößen betrachtet, so ergeben sich deutliche Unterschiede zu den Modellen für die einzelnen Spiele.

Die Richtung bzw. das Vorzeichen der Koeffizienten überrascht dabei vor allem bei der Variable Laufleistung. Auch den Einfluss der Ecken würde man positiv vermuten. Bei den Variablen Fouls/Gefoult worden/Abseits ist dagegen keine Richtung eindeutig erwartbar, die Einflüsse sind nicht signifikant. Die Einflüsse der anderen Variablen zeigen die erwartete Richtung auf. Der Ballbesitz, welcher in den Modellen zu den einzelnen Spielen noch kaum Einfluss zeigt, ergibt sich als deutlich signifikantester Wert.
Der Intercept lässt sich in diesem Modell nicht sinnvoll interpretieren, da Werte von Null für die durchschnittlichen Spielstatistiken in fast allen Kategorien nicht realisierbar sind.

	Variable	Value	p-Wert
1	Intercept	-96.1834	0.1826
2	TorschüsseGesamt	1.7683	0.0228
3	LaufleistungGesamt	-0.1138	0.7970
4	PassquoteGesamt	0.0599	0.8809
5	BallbesitzGesamt	1.2208	0.0009
6	ZweikampfquoteGesamt	1.5830	0.0401
7	FoulGesamt	0.0419	0.9571
8	GefoultwordenGesamt	-0.5668	0.4730
9	EckenGesamt	-2.2030	0.1867
10	AbseitsGesamt	2.6853	0.1846

Tabelle 6.2: Koeffizienten des linearen Modells je Variable in Value-Spalte und dazugehörigem p-Wert

Ausgehend vom Intercept kann das Modell bei sehr niedrigen Werten negative Werte für die Anzahl an Punkten prognostizieren. Nimmt man die Minima bzw. Maxima der einzelnen durchschnittlichen Spielstatistiken je nach Richtung des Einflusses (bei negativem Einfluss das Maximum und andersherum), so ergibt sich ein prognostizierter Punktewert von circa 0 bei Kombination der ungünstigsten Werte, welche in den Daten vorhanden sind. Bei günstigster Kombination ergibt sich ein prognostizierter Punktewert von knapp unter 100, was bei 34 Spielen eine durchschnittliche Punkteausbeute von 2.94 ergibt, welche theoretisch erreichbar ist. So liegen selbst die ungünstigsten Kombinationen im Bereich des Erreichbare, ein Problem durch Vorhersage negativer Werte ergibt sich so nicht. Die wirklich prognostizierten Werte liegen zwischen 21 und 84 Punkte. Die Werte lassen sich beispielhaft am Torschuss folgendermaßen erläutern: Ein Torschuss mehr im Durchschnitt pro Spiel steigert die erwartete Punktzahl in einer Saison um 1,77.
Um eine Vergleichbarkeit der Güte des Modells mit folgenden linearen Modellen zu gewährleisten, wird der R^2-Wert von 0.67 festgehalten.
Der große Unterschied zu den Modellen für einzelne Spiele besteht in den beiden Variablen Ballbesitz und Laufleistung. Wohingegen die Laufleistung neben der Anzahl an Torschüssen die einflussreichste Spielstatistik für den Ausgang einzelner Spiele darstellt und der Ballbesitz kaum einen Einfluss zeigt, ist dies in den Modellen für die Anzahl an Punkten pro Saison konträr. Eine mögliche Erklärung zeigt sich bei genauerer Betrachtung der beiden Variablen.
In Abbildung 6.1 ist deutlich zu erkennen, dass eine Trennung der Daten nach guten und schlechten Teams in y-Richtung erkennbar ist, in x-Richtung dagegen nicht. Dies bedeutet, dass über die Saison erfolgreiche Teams meist mehr Ballbesitz haben, wohingegen die durchschnittliche Laufleistung darauf keinen Einfluss zu haben scheint.
Werden die Teams mit sehr hohen Ballbesitzwerten betrachtet, ist darunter vor allem Bayern und Dortmund vertreten. Diese Teams sind auch diejenigen, deren Talent bzw. Fähigkeit am höchsten einzuschätzen ist, wodurch es diesen Teams leichter

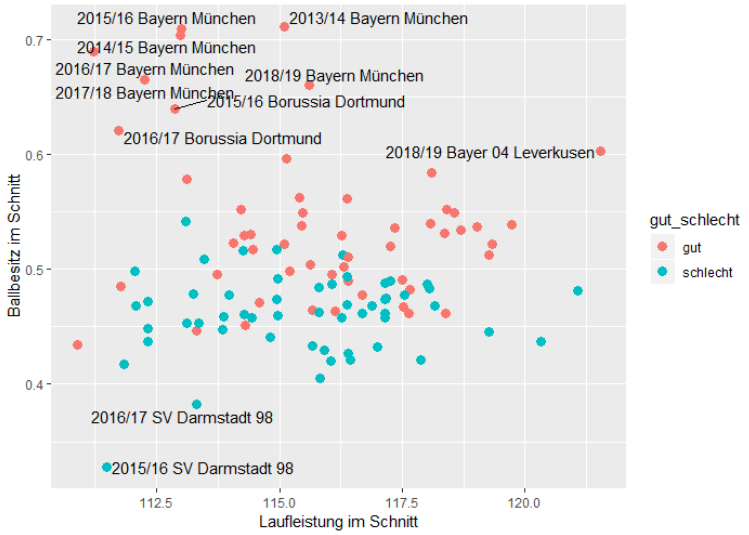


Abbildung 6.1: Laufleistung und Ballbesitz im Schnitt pro Saison pro Team gruppiert nach Teams der oberen/unteren Tabellenhälfte (gut/schlecht)

fällt, den Ball in den eigenen Reihen zu halten und dies auch über die ganze Saison zu praktizieren. Das Talent des Teams kann hier als latente Variable gesehen werden. Anhand dieses Modells soll außerdem herausgefunden werden, welche Teams mehr bzw. weniger Punkte erreicht haben, als das Modell vorhersagt. Man könnte so interpretieren, welche Teams in der Lage waren, mehr Punkte zu erreichen als ihre Spielstatistiken über die Saison hinweg aussagen würden. Gründe dafür können weitere latente Variablen wie eine gute Mentalität, gute Mannschaftsführung des Trainers oder Ähnliches ein.

So konnte beispielsweise Schalke 04 in der Saison 2017/18, als das Team Vize-

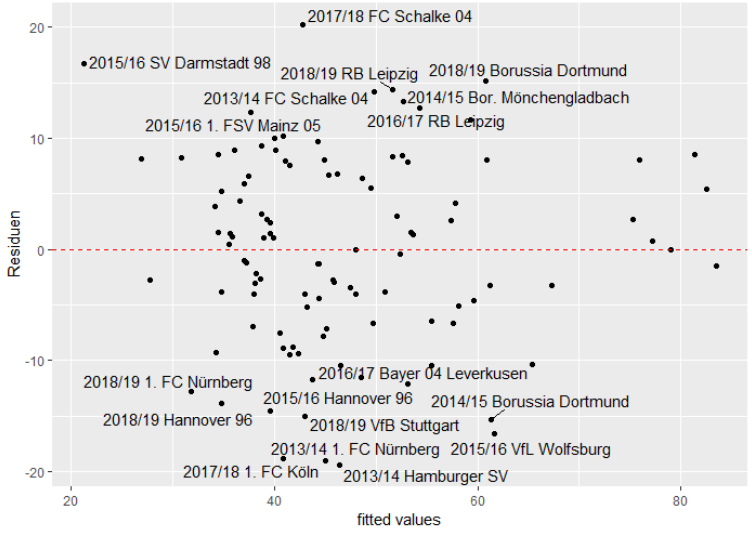


Abbildung 6.2: Fitted Values gegen Residuen des linearen Modells der saiosonaggregierten Daten

Meister wurde, circa 20 Punkte mehr erreichen, als das Modell anhand der saisonaggregierten Daten vorhersagt. Man könnte interpretieren, dass deren Punktanzahl also nicht deren „statistisches Können“ widerspiegelte. Die Platzierungen in der Vor- und Folgesaison mit 10 und 14 würden diese These in gewisser Weise unterstützen.

6.2.2 Score-Daten

Um verschiedene Scores zu bilden, welche durch ihre gleichen Skalen für eine gute Vergleichbarkeit sorgen sollen, wurden Ränge gebildet. Dabei wurden pro saisonaggregierter Spielstatistik alle Beobachtungen geordnet und absteigend Ränge vergeben Das heißt, dass beispielsweise das Team mit der höchsten durchschnittlichen Ballbesitzquote in einer Saison aller 108 Beobachtungen den Wert eins zugewiesen bekommt, das Team mit der niedrigsten den Wert 108. Bei gleichen Werten wurde der Durchschnittsrang vergeben. Die Scores aus Tabelle 6.1 wurden dann jeweils mit den ausgewählten Spielstatistiken nach dieser Form gebildet:

$$Score = \frac{1}{k} \sum_{Spielstatistik_i=1}^{k} rank(Spielstatistik_i). \tag{6.4}$$

Es werden zwei Modelle gebildet: Zum einen soll der Einfluss von Offensiv- und Defensiv-Score, zum Anderen der Einfluss von spielerischem und physischem Score verglichen werden.

Variable	Value	p-Wert
Intercept	76.1245	$2 \cdot 10^{-16}$
Offensiv-Score	-0.2568	$2 \cdot 10^{-16}$
Defensiv-Score	-0.2802	$2 \cdot 10^{-16}$

(a) Koeffizienten Modelle mit Offensiv- und Defensiv-Score

Variable	Value	p-Wert
Intercept	67.2008	$2 \cdot 10^{-16}$
Spielerisch-Score	-0.3563	$1.15 \cdot 10^{-14}$
Physisch-Score	-0.0169	0.784

(b) Koeffizienten Modelle mit Physisch- und Spielerisch-Score

Abbildung 6.3: Koeffizienten zu den Modellen mit Score-Daten

Bei ersterem Modell ergeben sich sehr ähnliche Einflüsse von Offensive und Defensive. Bei einer Verschlechterung des Offensiv-Scored um einen Platz (also der Steigerung der Variable um 1) ergibt sich eine Minderung der erwarteten Punktzahl in der Saison um 0.26, beim Defensiv-Score um 0.28, welches dadurch einen leicht höheren Einfluss besitzt.
Der Anteil der Varianz, der durch das Modell erklärt werden kann, liegt bei $R^2 = 0.8$. Dieser vergleichsweise hohe Wert liegt sicher an der Mitaufnahme der geschossenen und kassierten Tore, welche für bisherige Modelle nicht betrachtet wurden. Interpretieren lässt sich also, dass die Defensive ein wenig einflussreicher auf die Anzahl an erreichten Punkte zu sein scheint als die Offensive.
Das zweite Modell zeigt dagegen deutliche Unterschiede der beiden Scores. Das spielerische Score hat im Gegensatz zum physischen Score einen hochsignifikanten Einfluss, was sich durch den in Kapitel 6.2.1 beobachteten deutlichen Einfluss des

Ballbesitzes erklären lässt. Es ergibt sich ein R^2-Wert von 0.5. Das Modell könnte außerdem maximal eine Punktzahl von 66.84 (beide Scores mit Wert 1) vorhersagen und ist damit nicht für die Vorhersage geeignet und dient lediglich dem Vergleich des Einflusses der beiden Scores.

Kapitel 7

Fazit

Bei der Analyse des Einflusses der Spielstatistiken auf einzelne Spiele ergaben sich die Laufleistung und die Torschüsse als einflussreichste Variablen. Bei der Modellierung des Spielausgangs wurde dabei deutlich, dass kein Modell ein Unentschieden als Ausgangskategorie vorhersagt, was jedoch auch inhaltlich anhand der niedrigen Wahrscheinlichkeiten und dem Dasein als mittlere Kategorie, von welcher aus ein Sprung zu Sieg oder Niederlage lediglich ein weiteres Tor bedarf, erklärbar ist.
Bei der Modellierung der erreichten Punkte einer Mannschaft pro Saison anhand von saisonaggregierten Daten zeigte sich dagegen der Ballbesitz als einflussreichste Variable, dessen Effekt bei den Modellen zu den einzelnen Spielen noch gering war. Beim Vergleich der Offensiv- und Defensiv-Scores waren kaum Unterschiede im Einfluss erkennbar, wohingegen spielerische Elemente, vor allem durch den hohen Ballbesitzeinfluss, über die Saison gesehen einen deutlich stärkeren Einfluss als physische Elemente haben.
Als mögliche Erklärung für den hohen Einfluss des Ballbesitzes bei den saisonaggregierten Daten wurde vermutet, dass sich physische Daten über die gesamte Saison hinweg ausgleichen, wohingegen das dominante Auftreten spielerisch starker Teams über die Saison hinweg konstant ist, was auch eine Folge des Talents der einzelnen Spieler und damit einer ganzen Mannschaft ist.
Es scheint also möglich einzelne Spiele durch Kampf und gute physische Leistungen für sich zu entscheiden. Für den Erfolg über die gesamte Saison, reichen diese Attribute jedoch nicht aus.
Daraus ergibt sich auch die Herausforderung für Trainer, deren Spieler für einzelne Spiele zu motivieren, um gute physische Werte zu zeigen, was zweifelsfrei auch eine Sache der Fitness ist. Gleichzeitig scheint es über die Saison hinweg jedoch wichtig zu sein, die spielerische Stärke der Mannschaft zu verbessern, um auch langfristig erfolgreich zu bleiben.

Literaturverzeichnis

[Breiman 1996] Breiman, Leo: Bagging predictors. In: *Machine learning* 24 (1996), Nr. 2, S. 123–140

[Breiman 2017] Breiman, Leo: *Classification and regression trees*. Routledge, 2017

[Breiman u. a. 1984] Breiman, Leo ; Friedman, Jerome ; Olshen, Richard ; Stone, Charles: Classification and regression trees. Wadsworth Int. In: *Group* 37 (1984), Nr. 15, S. 237–251

[Deloitte 2018] Deloitte: *Annual Review of Football*. 2018. – URL https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/sports-business-group

[DFL 2018] DFL: *Report 2018: German professional football generates revenue in excess of 4 billion for the first time – 14 Bundesliga clubs with revenue over 100 million*. 2018. – URL https://www.dfl.de/en/news/2018-dfl-report-german-professional-football-generates

[Duit 2012] Duit, Nino: *Taktische Revolutionen (3) - Rappan, Herrera und der Catenaccio*. 2012. – URL https://abseits.at/in-depth/taktik-theorie/taktische-revolutionen-3-rappan-herrer

[Fahrmeir u. a. 2007] Fahrmeir, Ludwig ; Kneib, Thomas ; Lang, Stefan ; Marx, Brian: *Regression*. Springer, 2007

[Hastie und Tibshirani 2011] Hastie, Trevor ; Tibshirani, Robert: Statistical learning. In: *Learning* 2 (2011), Nr. 08

[Havemann 2013] Havemann, Nils: *Samstags um halb vier: Die Geschichte der Fussballbundesliga*. Siedler Verlag, 2013

[Herrmann 2015] Herrmann, Karol: *Tracking: Ein Blick hinter die Kulissen der Datenerhebung*. DFL. 2015. – URL https://www.bundesliga.com/de/bundesliga/news/inside-trackingdaten-ein-blick-hint

[Moebius 2018] Moebius, Karsten: Vom Rasenschach zum Tempo-Fussball. In: *MDR Wissen* (2018)

[Observatory 2018] Observatory, CIES F.: *Effective playing time in 37 European competitions*. 2018. – URL https://football-observatory.com/IMG/sites/b5wp/2018/242/en/

[Pollard 2002] Pollard, Richard: Charles Reep (1904-2002): pioneer of notational and performance analysis in football. In: *Journal of Sports Sciences* 20 (2002), Nr. 10, S. 853–855

[Schlipsing u. a. 2013] Schlipsing, Marc ; Salmen, Jan ; Igel, Christian: Echtzeit-Videoanalyse im Fußball. In: *KI-Kuenstliche Intelligenz* 27 (2013), Nr. 3, S. 235–240

[SportingIntelligence 2017] SportingIntelligence: *GLOBAL ATTENDANCES - Best attended domestic sports leagues in the world*. 2017. – URL http://www.sportingintelligence.com/finance-biz/business-intelligence/global-atte

[Strobl u. a. 2008] Strobl, Carolin ; Boulesteix, Anne-Laure ; Kneib, Thomas ; Augustin, Thomas ; Zeileis, Achim: Conditional variable importance for random forests. In: *BMC bioinformatics* 9 (2008), Nr. 1, S. 307

[Weltfussball 2016] Weltfussball, Karol: *HSV Ekelhafte Ingolstaedter Spielweise*. 2016. – URL https://www.weltfussball.de/news/_n2104666_/hsv-ekelhafte-ingolstaedter-spielweise

Tabellenverzeichnis

Abbildungsverzeichnis

Eigenständigkeitserklärung

Ich versichere, dass ich die vorgelegte Bachelorarbeit eigenständig und ohne fremde Hilfe verfasst, keine anderen als die angegebenen Quellen verwendet und die den benutzten Quellen entnommenen Passagen als solche kenntlich gemacht habe. Diese Bachelorarbeit ist in keinem anderen Kurs in dieser oder einer ähnlichen Form vorgelegt worden.

Name, Vorname: Raith, Julian

München, den 24.01.2019

Unterschrift: