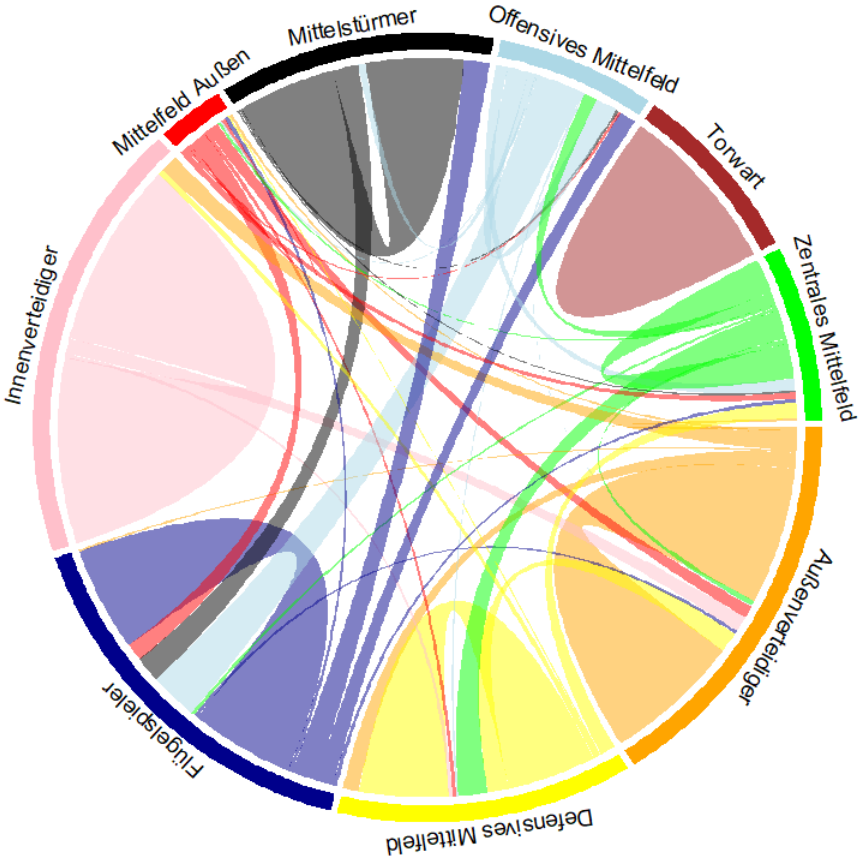
Ludwig-Maximilians-Universität München
Institut für Statistik

Alexander Gerharz

Positionsbezogene Leistungsdatenanalyse von Fußballspielern



Abschlussarbeit zur Erlangung
des akademischen Grades
Master of Science

Datum
24.05.2019

Betreuer
Prof. Dr. Christian Heumann
Dr. Gunther Schauberger

Eigenständigkeitserklärung

Hiermit versichere ich, dass ich die vorliegende Masterarbeit selbstständig verfasst habe. Ich versichere, dass ich keine anderen als die angegebenen Quellen benutzt und alle wörtlich oder sinngemäß aus anderen Werken übernommenen Aussagen als solche gekennzeichnet habe und dass die eingereichte Arbeit weder vollständig noch in wesentlichen Teilen Gegenstand eines anderen Prüfungsverfahrens gewesen ist.

Alexander Gerharz

Abstract

Diese Masterarbeit beschäftigt sich mit den Leistungsdaten und den Positionen von Fußball-Bundesligaspielern und untersucht diese auf Zusammenhänge zwischen den Leistungsdaten und den Positionen. Um dies auszuarbeiten werden ein multinomiales logistisches Regressionsmodell und ein *Random Forest* verwendet und die Effekte der Modelle mithilfe von interpretierbaren Machine Learning Methoden analysiert. Obwohl das Erzeugen der Modelle auf sehr verschiedene Art und Weise funktioniert, kann mit den interpretierbaren Machine Learning Methoden gezeigt werden, dass sich die Effekte in den beiden Modellen sehr ähneln.

Inhaltsverzeichnis

1 Einführung

Viele Sportarten werden in der heutigen Zeit zunehmend quantifiziert und in Datenbanken erfasst. Diese Quantifizierung führt dazu, dass immer mehr statistische Analysen in den verschiedensten Sportarten durchgeführt werden, um die Ereignisse, die in einem Wettkampf stattfinden, besser zu verstehen. In manchen Sportarten wie zum Beispiel dem Baseball, werden schon seit Jahren statistische Analysen durchgeführt, um das Team optimal zusammenzustellen oder das Abschneiden der Teams in einer Saison zu modellieren.

Im Unterschied zum Baseball ist die Analyse von Leistungsdaten im Fußball etwas schwieriger. Während im Baseball ein Spielzug hauptsächlich von den Fähigkeiten zweier Spieler abhängt (dem Ball-werfenden Spieler und dem Ball-schlagenden Spieler), hängt ein Spielzug im Fußball von mehreren angreifenden und verteidigenden Spielern ab und macht somit die Ausgangslage eines einzelnen Spielzuges schon deutlich komplexer. Um diese Spielzüge quantifizierbar zu machen, werden im Fußball mittlerweile allerlei Leistungsdaten erfasst. Angefangen vom Zählen der gespielten Pässe eines Spielers bis hin zum Messen der Höchstgeschwindigkeit und der Laufweite eines Spielers werden immer größere Datengrundlagen geschaffen.

Ein sehr moderner Wert, der häufig für die Analyse einer Spielsituation genutzt wird, ist der "Expected Goals"-Wert. Dieser beschreibt mit welcher Wahrscheinlichkeit in der jeweiligen Spielsituation ein Tor fällt (Nordmann 2016).

Die folgende Arbeit beschäftigt sich damit, wie die erfassten Leistungsdaten zusammenhängen und wie sich diese zwischen den einzelnen Positionen auf dem Fußballfeld unterscheiden. Hierfür werden deskriptive Methoden genutzt, um die gemessenen Leistungsdaten besser zu verstehen und sowohl klassische als auch maschinen-basierte Modellierungen verwendet, um die Zusammenhänge der Leistungsdaten bezüglich ihrer Positionen zu modellieren.

Wichtig ist es aus Modellen Wissen zu generieren. Modellierungen aus dem Bereich des *Machine Learning* haben den Ruf schwer interpretierbar zu sein, weshalb sie oft nur für Prognosen verwendet werden. In den letzten Jahren entwickeln sich jedoch immer mehr Methoden, die versuchen die "Blackbox" einer *Machine Learning*-Methode zu entschlüsseln und die Modelle interpretierbarer zu machen. Aus diesem Bereich des "Interpretierbaren Machine Learnings" werden in dieser Arbeit Methoden verwendet, um das klassische Modell mit dem *Machine Learning*-Modell zu vergleichen und daraus Wissen zu generieren.

Die Modelle sollen die Position einer Beobachtung anhand ihrer Leistungsdaten schätzen. Wenn die Modelle diese Beziehung gut beschreiben, kann anhand der Modelle ausgearbeitet werden, wie sich die Leistungsdaten auf die Positionen auswirken. Im Speziellen soll untersucht werden, ob hohe Werte bestimmter Leistungsdaten für bestimmte Positionen sprechen (z.B. ob eine hohe *Laufweite* eher für einen *Verteidiger* oder einen *Mittelfeldspieler* spricht). Die interessantesten Beziehungen werden dabei durch Methoden des *interpretierbaren Machine Learnings* näher beschrieben und zwischen den beiden Modellen auf Gemeinsamkeiten und Unterschiede untersucht.

Darüber hinaus wird in dieser Arbeit eine Methode vorgestellt, die die Topologie der Daten, die durch die beiden Modelle beschrieben wird, untersuchen soll. Anhand dieser Methode

wird es möglich sein zu überprüfen, welche Positionen bezüglich eines einzelnen Leistungsdatums gegeben der anderen Leistungsdaten im Raum benachbart liegen.

2 Datensatz

2.1 Datengrundlage und Herkunft

Um valide und sinnvolle Analysen durchzuführen wird ein Datensatz mit aktuellen Leistungsdaten aus mehreren Quellen zusammengeführt. Bevor dies geschehen kann, muss jedoch eine Datengrundlage definiert werden.

Im Großen und Ganzen sind alle Spieler, die je in der Bundesliga gespielt haben, Teil der Datengrundlage. Um jedoch nur Spieler in den Datensatz aufzunehmen, die wirklich bezeichnend für die Bundesliga sind, müssen Einschränkungen getroffen werden.

Zum Einen ist es wichtig, dass einzelne ausreißende Spiele die Analysen nicht zu stark verzerren. Daher wird die Datengrundlage auf alle Spieler, die mindestens *4* Spiele (bzw. *360* Spielminuten) in einer Bundesligasaison gespielt haben, reduziert und die Daten werden saisonaggregiert betrachtet.

Zum Anderen bestehen Differenzen zwischen der Bundesliga und den niedrigeren Ligen in Deutschland, bzw. zwischen der Bundesliga und anderen Top-Ligen auf der Welt. Daher wird die Datengrundlage weiterhin auf Spieler reduziert, die mindestens *3* Jahre in der Bundesliga gespielt haben und somit über einen längeren Zeitraum gezeigt haben, dass ihre Fähigkeiten denen eines Bundesligaspielers entsprechen.

Um eine Liste der Namen zu erhalten, wurden die Spielerlisten von Bundesligaprofis auf der Seite *www.weltfussball.de* genutzt ("weltfussball.de" 2018). Anhand dieser Liste wurde nach Leistungsdaten der Bundesligaprofis gesucht. Darüber hinaus konnte überprüft werden, wie viele Saisons ein Spieler in der Bundesliga einem Kader angehörte. Hier sind keine Spielminuten oder gespielte Spiele angegeben, weshalb dieser Filter im Nachhinein gesetzt werden musste.

Auf der Seite *www.sport1.de* waren bis Anfang des Jahres 2019 umfangreiche saisonaggregierte Leistungsdaten aufgelistet ("sport1.de" 2018). Dieser große Umfang an Leistungsdaten existiert jedoch erst seit der Saison 2009/2010. Die Daten vor dieser Saison waren auf nur wenige Leistungsdaten beschränkt. Daher wurde die Datengrundlage auf alle Spieler, die seit 2009/2010 in der Bundesliga gespielt haben ein weiteres mal eingeschränkt.

Die Spielerlisten von *www.weltfussball.de* wurden verwendet, um über die Namen der Spieler die URLs zum Scrapen der Leistungsdaten zu ermitteln. Nicht alle Spieler sind über ihren Namen gefunden worden. Die Teilmenge der gefunden Spieler wurde auf Diskriminierungen bezüglich Herkunft, Position, Alter, Verein und Spielzeit untersucht. Es wurde kein diskriminierendes Muster gefunden, weshalb die Stichprobe als repräsentativ betrachtet wird. Die Leistungsdaten enthielten die Anzahl der Spielminuten, wodurch die Spieler, die weniger als *360* Spielminuten in einer Saison aufgewiesen haben hier gefiltert wurden.

Diese Leistungsdaten bestehen aus:

Leistungsdatum	Reichweite	Gruppierung
Spielminuten	360 - 3060	Spielminuten
Ballkontakte	127 - 3066	Spielbeteiligung Generell
Gespielte Pässe	28 - 2595	Spielbeteiligung Generell
Angekommene Pässe	19 - 2343	Spielbeteiligung Generell
Fehlpässe	9 - 464	Spielbeteiligung Generell
Passquote in %	48 - 94	Spielbeteiligung Generell
Zweikämpfe	0 - 1227	Spielbeteiligung Generell
Zweikampfquote in %	0 - 100	Spielbeteiligung Generell
Laufweite in km	20.84 - 398.67	Körperliche Leistungen
Höchstgeschwindigkeit in km/h	24 - 35	Körperliche Leistungen
Sprints	1 - 1162	Körperliche Leistungen
Tore	0 - 31	Tore
Tore mit dem Fuss	0 - 25	Tore
Kopfballtore	0 - 7	Tore
Elfmetertore	0 - 8	Tore
Verschossene Elfmeter	0 - 3	Tore
Schüsse	0 - 151	Spielbeteiligung Offensiv
Schussvorlagen	0 - 124	Spielbeteiligung Offensiv
Torvorlagen	0 - 20	Spielbeteiligung Offensiv
Abseits	0 - 60	Spielbeteiligung Offensiv
Eigentore	0 - 3	Spielbeteiligung Sonstiges
Fouls	0 - 96	Spielbeteiligung Sonstiges
Gefoult worden	0 - 121	Spielbeteiligung Sonstiges
Gegentore	0 - 70	Torwart
Gehaltene Schüsse	0 - 152	Torwart
Gehaltene Elfmeter	0 - 5	Torwart

Tabelle 1: Übersicht über die Leistungsdaten

Diese Daten sind seit Ende Januar 2019 nicht mehr direkt verfügbar, können jedoch noch über Webarchive gefunden werden. Der Stand der Daten für die Saison 2018/2019 ist der 16.11.2018 (zwischen dem 11. und 12. Spieltag). Im weiteren Teil der Arbeit werden vor allem für die Modellierung die Leistungsdaten auf ihre Spielminuten bezogen, weshalb es unproblematisch ist eine noch laufende Saison hier aufzunehmen.

Um eine detailliertere Information über die gespielte Position eines Profis innerhalb einer Saison zu erhalten, wurden die gespielten Positionen von der Seite *www.transfermarkt.de* an den Datensatz angefügt ("transfermarkt.de" 2018). Dort wurden die gespielten Positionen zusammen mit der Anzahl an Spielen, die die Spieler auf den Positionen in einer bestimmten Saison gespielt haben, erfasst. Diese Positionen erweitern die bisherigen Informationen, die aus *Torwart*, *Verteidiger*, *Mittelfeld* und *Sturm* bestanden, um genauere Angaben. In diesen Daten ist aufgeführt, wie häufig ein Spieler eine bestimmte Position über die Saison bekleidet hat. In Tabelle 2 sind die gespielten Positionen zusammen mit der Anzahl an Beobachtungen, die diese als "Häufigste gespielte Position" aufführen, aufgelistet. Wie zu sehen ist, existiert

Position	Anzahl an Beobachtungen
Torwart	154
Linker Verteidiger	168
Rechter Verteidiger	151
Innenverteidiger	366
Libero	1
Defensives Mittelfeld	246
Linkes Mittelfeld	48
Rechtes Mittelfeld	46
Zentrales Mittelfeld	146
Offensives Mittelfeld	114
Linksaußen	141
Rechtsaußen	132
Hängende Spitze	48
Mittelstürmer	219

Tabelle 2: Anzahl der Beobachtungen pro Position

nur eine Beobachtung mit der Position *Libero*, was bedeutet, dass nur ein Bundesligaprofi in diesem Datensatz über eine komplette Saison hauptsächlich als *Libero* gespielt hat. Daher wurde diese Beobachtung ***nur*** für einen Teil der deskriptiven Analysen der Leistungsdaten verwendet und nicht für die positionsbezogenen Analysen.

Alles in allem enthält der Datensatz, der in den folgenden Analysen untersucht wird, 1980 Beobachtungen. Jede Beobachtung entspricht den saisonaggregierten Leistungsdaten eines Bundesligaspielers von einer bestimmten Saison zwischen 2009/2010 bis 2018/2019. Diese Beobachtungen stammen von 407 verschiedenen Spielern.

2.2 Datenimputationen

Die Leistungsdaten von der Seite *www.sport1.de* weisen ein Leistungsdatum, das keine Ausprägung besitzt (also zum Beispiel einen *Torhüter*, der keinen Schuss abgegeben hat) nicht auf, anstatt dieses mit einer *0* zu erfassen. Daher wurde eine *0*-er Imputation für ausschließlich plausiblen Variablen durchgeführt.

Die meisten Beobachtungen weisen für die ausschließlich für *Torhüter* erfassten Variablen *Gegentore*, *Gehaltene Schüsse* und *Gehaltene Elfmeter* keine Ausprägung auf. Diese wurden für alle Positionen, abgesehen der *Torhüter*, mit einer *0* aufgefüllt.

Im Gegenteil dazu fehlen *Zweikämpfe* ausschließlich bei *Torhütern* und die Anzahl an *Fouls* und wie oft jemand *Gefoult worden* ist bei *Torhütern* und Feldspielern mit sehr wenig Spielminuten. Dies ist sehr plausibel, weshalb diese auch durch *0*en aufgefüllt wurden.

Die Leistungsdaten für *offensive Spielbeteiligungen*, *Tore* und *Eigentore* sind häufig fehlend, aber primär bei defensiven Spielern und *Torhütern*, bzw. im Falle der Eigentore bei offensiven Spielern, weshalb auch diese mit *0*en aufgefüllt wurden.

In den Saisons vor 2011 fehlt häufig die *Laufweite*. Es wird vermutet, dass diese eventuell noch nicht mitgetrackt werden konnte, da eine Laufweite von *0 km* in mindestens *4* Spielen unplausibel ist. Diese Beobachtungen wurden mit `NA`s aufgefüllt.

2.3 Alter

Das Alter der Spieler in einer Saison wurde an ihrem Geburtstag gemessen und gerundet aufgenommen. Da die meisten Spieler während der Saison ihren Geburtstag feiern, musste sich für einen bestimmten Stichtag entschieden werden, an dem das Alter bestimmt wird. Hierfür wurde der 31.12. als Stichtag bestimmt und das Alter an diesem Tag für die gesamte Saison gemessen. Wenn ein Spieler vor diesem Tag Geburtstag hat, ist er die komplette Rückrunde und den Rest der Hinrunde nach seinem Geburtstag bereits ein Jahr älter als zum Start der Saison (also $> 50\%$ der Saison). Wenn ein Spieler erst nach diesem Tag Geburtstag hat, hat er die komplette Hinrunde und die Rückrunde von Start bis zu seinem Geburtstag mit dem Alter gespielt, mit dem er in die Saison gestartet ist (also auch $> 50\%$ der Saison). Daher ergibt die Wahl des 31.12. als Stichtag zur Bestimmung des Alters Sinn.

3 Deskriptive Veranschaulichung der Daten

3.1 Absolute Leistungsdaten

3.1.1 Verteilung der absoluten Leistungsdaten

Die Leistungsdaten sind ganzzahlig und metrisch gegeben. Um die Verteilungen der Leistungsdaten darzustellen, wurden Histogramme verwendet. Diese Histogramme bilden die Verteilung der Leistungsdaten ab. Anhand dieser kann einerseits erkannt werden, ob ein Leistungsdatum einer schiefen Verteilung folgt, und andererseits, ob ein Leistungsdatum mehrgifplig verteilt ist. Da die Leistungsdaten der *Torhüter* sich deutlich von den Leistungsdaten der Feldspieler unterscheiden, werden hier nur die Feldspieler betrachtet (bspw. hat kein *Torhüter* im Datensatz ein *Tor* geschossen oder im *Abseits* gestanden). Für die Variablen, die nur für die *Torhüter* Ausprägungen aufweisen, wurden eigene Histogramme nur mit den *Torhütern* erstellt.

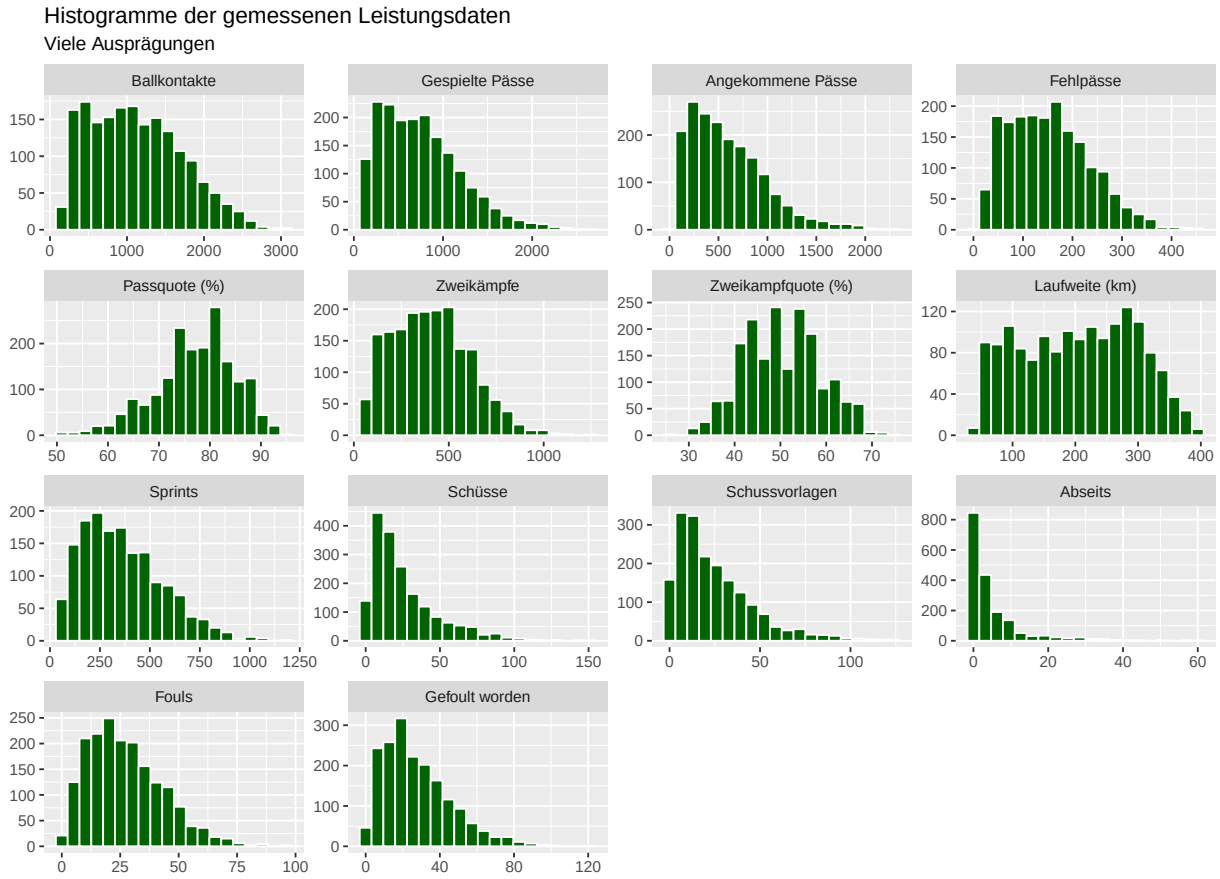


Abbildung 1: Visualisierung der absoluten Leistungsdaten durch Histogramme - Viele Ausprägungen

Für die Leistungsdaten, die eine hohe Anzahl an verschiedenen Ausprägungen aufweisen

(≥ 25), werden Histogramme mit einer festen Anzahl an Balken abgebildet. Dies bedeutet, dass Beobachtungen, die Nahe beieinander liegen in diskrete Klassen aufgeteilt werden und dadurch zusammen abgebildet werden. Würden die Daten nicht zusammengefasst werden, so wäre die Verteilung bei manchen Leistungsdaten nur schwer zu erkennen, da sie viele Ausprägungen aufweisen, die jeweils nur sehr selten (z.B. 1 bis 5 mal für die *Anzahl der gespielten Pässe*) vorkommen. In Abbildung 1 werden alle Beobachtungen, die keine *Torhüter* sind, anhand solcher Histogramme abgebildet. Wie hier zu sehen ist, sind die meisten Leistungsdaten linkssteil verteilt. Die *Passquote* weist hingegen eine rechtssteile Verteilung auf. Die *Laufweite* scheint über den Wertebereich in etwa gleichverteilt zu sein. Die *Zweikampfquote* ähnelt noch am meisten einer Normalverteilung. Alle Leistungsdaten liegen in einem plausiblen Wertebereich.

Die anderen Leistungsdaten, die nur eine sehr niedrige Anzahl an verschiedenen Ausprägungen aufweisen (≤ 25), sind ohne Zusammenfassen für jede Ausprägung gezählt worden und in Histogrammen in Abbildung 2 dargestellt.

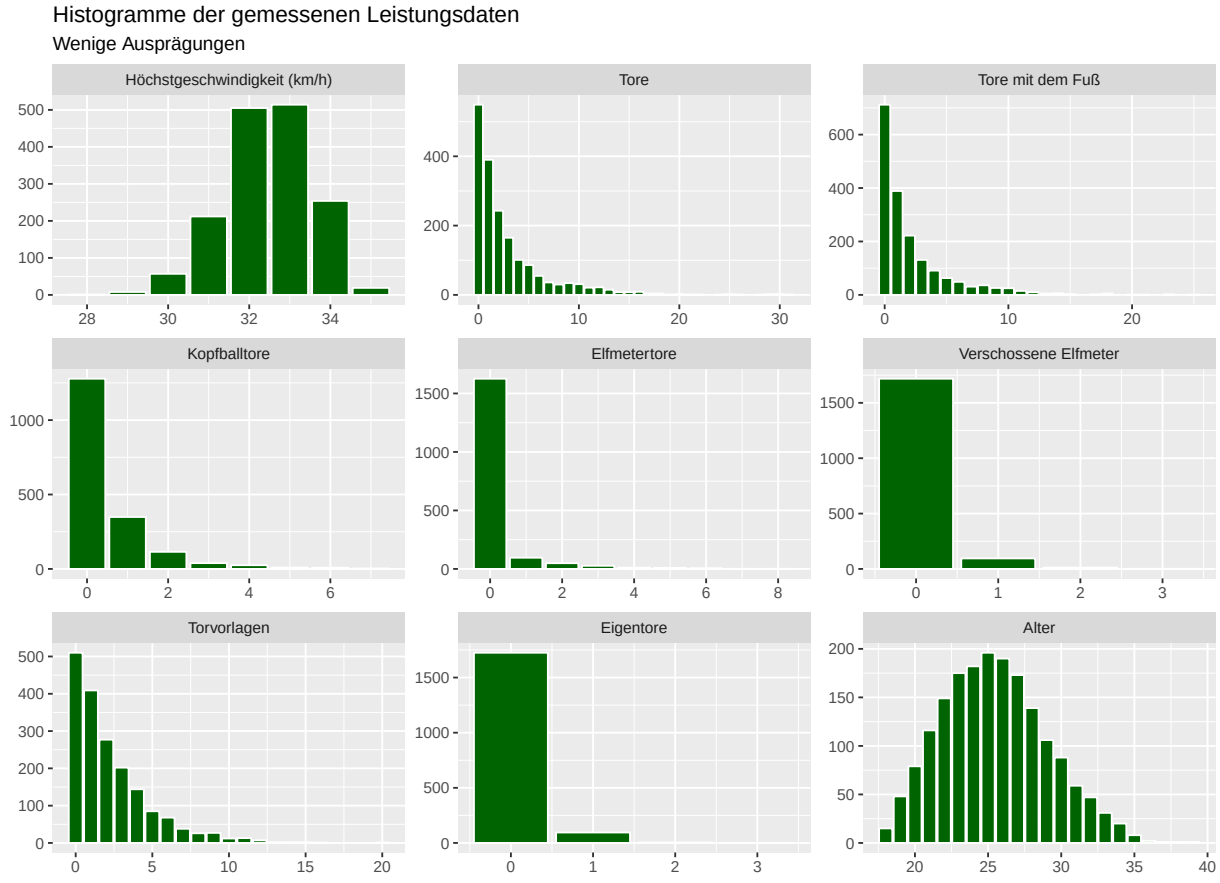


Abbildung 2: Visualisierung der absoluten Leistungsdaten durch Histogramme - Wenige Ausprägungen

Bis auf die *Höchstgeschwindigkeit*, die eine leichte rechtssteile Verteilung aufweist, sind die Leistungsdaten mit wenigen Ausprägungen alle linkssteil.

Um die drei Torhüter-Leistungsdaten nocheinmal genauer zu betrachten, sind diese nur für die *Torhüter* in Abbildung 3 visualisiert.

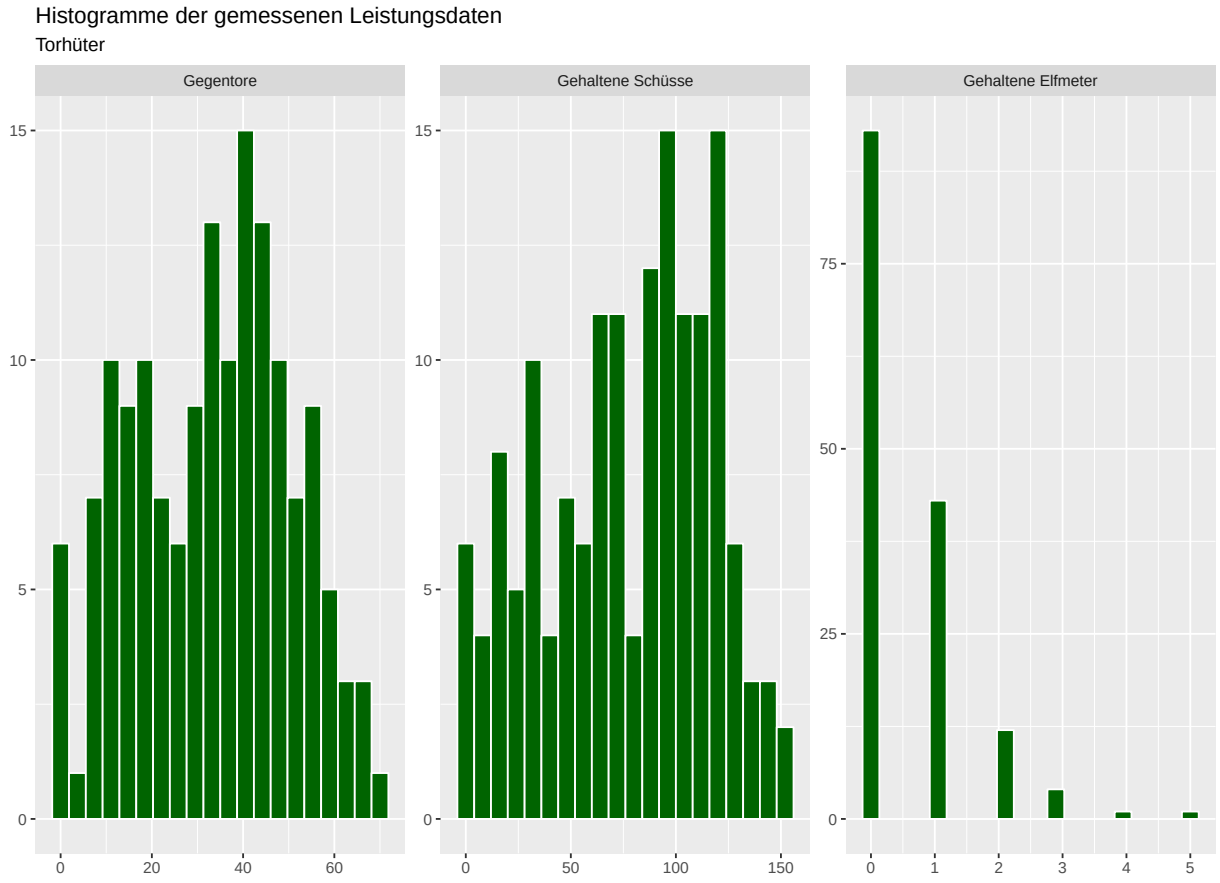


Abbildung 3: Visualisierung der absoluten Leistungsdaten der *Torhüter*

Bis auf eine klare linkssteile Verteilung der *Anzahl an gehaltenen Elfmeter* folgen die anderen beiden Verteilungen keinen eindeutigen Strukturen. Die *Anzahl an Gegentoren* weist eine leicht bimodale Struktur auf, während die *Anzahl an gehaltenen Schüssen* eine leichte rechtssteile Verteilung aufweist.

3.1.2 Zusammenhänge der absoluten Leistungsdaten

Um die Zusammenhänge zwischen den Leistungsdaten zu überprüfen, werden ihre Korrelationen nach Pearson gemessen. Diese Zusammenhänge sind hier als Heatmap dargestellt. Eine rote Kachel steht für eine positive Korrelation zwischen den beiden Leistungsdaten und eine blaue Kachel für eine negative Korrelation. Je höher die Farbsättigung, desto höher die Korrelation nach Pearson.

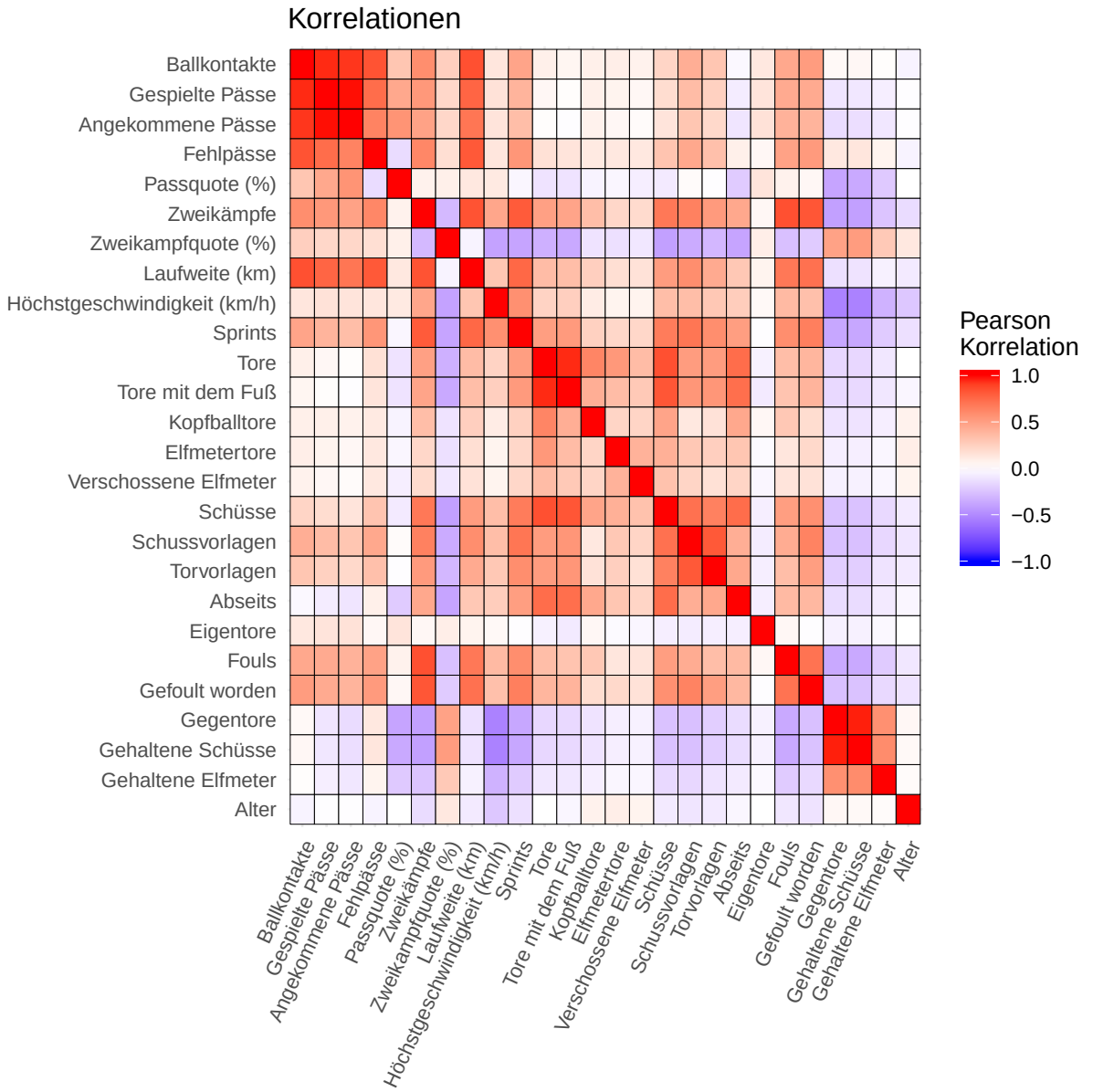


Abbildung 4: Visualisierung der Korrelation nach Pearson der absoluten Leistungsdaten

Auffällig sind die ersten vier aufgeführten Leistungsdaten, also *Ballkontakte*, *Gespielte Pässe*, *Angekommene Pässe* und *Fehlpässe*, da diese alle hoch positiv miteinander korreliert sind. Auch die *Laufweite in km* weist noch eine hohe positive Korrelation mit den vier Leistungsdaten auf. Abgesehen von der *Laufweite* messen all diese Leistungsdaten die generelle Spielbeteiligung der Bundesligaprofis.

Eine weitere Gruppe hoch positiv korrelierter Leistungsdaten sind die *Schüsse*, *Schussvorlagen*, *Torvorlagen* und *Abseitsstellungen*. Diese Leistungsdaten messen die offensive Spielbeteiligung der Bundesligaprofis.

Darüber hinaus weisen die Torhüter-Leistungsdaten *Gegentore* und *Gehaltene Schüsse* eine

sehr hohe positive Korrelation auf, während die Variable *Gehaltene Elfmeter* nur eine leichte positive Korrelation mit den anderen beiden Torhüter-Leistungsdaten aufweist. Dies ist jedoch die höchste mit *0*en befüllte Variable, weshalb hier keine hohe Korrelation mit den anderen Leistungsdaten erwartet werden kann.

Das Alter weist keine hohe Korrelation mit einem der anderen Leistungsdaten auf. Dies bedeutet, dass keine der Leistungsdaten mit steigendem Alter stark linear abfällt, bzw. zunimmt.

3.2 Relative Leistungsdaten

3.2.1 Verteilung der relativen Leistungsdaten

Die bisherigen Visualisierungen weisen die Verteilungen und Zusammenhänge zwischen den absolut gemessenen Leistungsdaten auf. Dies bedeutet, dass Verteilungen und Korrelationen dadurch stark beeinflusst worden sind, wie viel Spielzeit ein Spieler in einer Saison angesammelt hat. Das wirkliche Interesse an diesen Daten steckt aber darin, die Fähigkeiten der einzelnen Spieler zu messen und diese miteinander zu vergleichen. Aus diesem Grund wurden die Leistungsdaten auf ihre Spielzeit bezogen. Da eine Angabe pro Spielminute jedoch schwer zu interpretieren ist, wurden die Leistungsdaten auf ihre Einheit pro 90 Minuten bezogen. Die neuen Leistungsdaten bilden also ab, wie viele *Pässe*, *Zweikämpfe*, *Schüsse*, etc. ein Spieler pro Spiel (exklusive Nachspielzeit) in einer Saison aufweisen konnte.

Die Leistungsdaten *Passquote (in %)*, die *Zweikampfquote (in %)* und die *Höchstgeschwindigkeit (in km/h)* sind bereits relativ, weshalb diese nicht erneut auf ihre Spielzeit bezogen wurden.

Die Verteilungen der relativen Leistungsdaten sieht aus wie folgt:

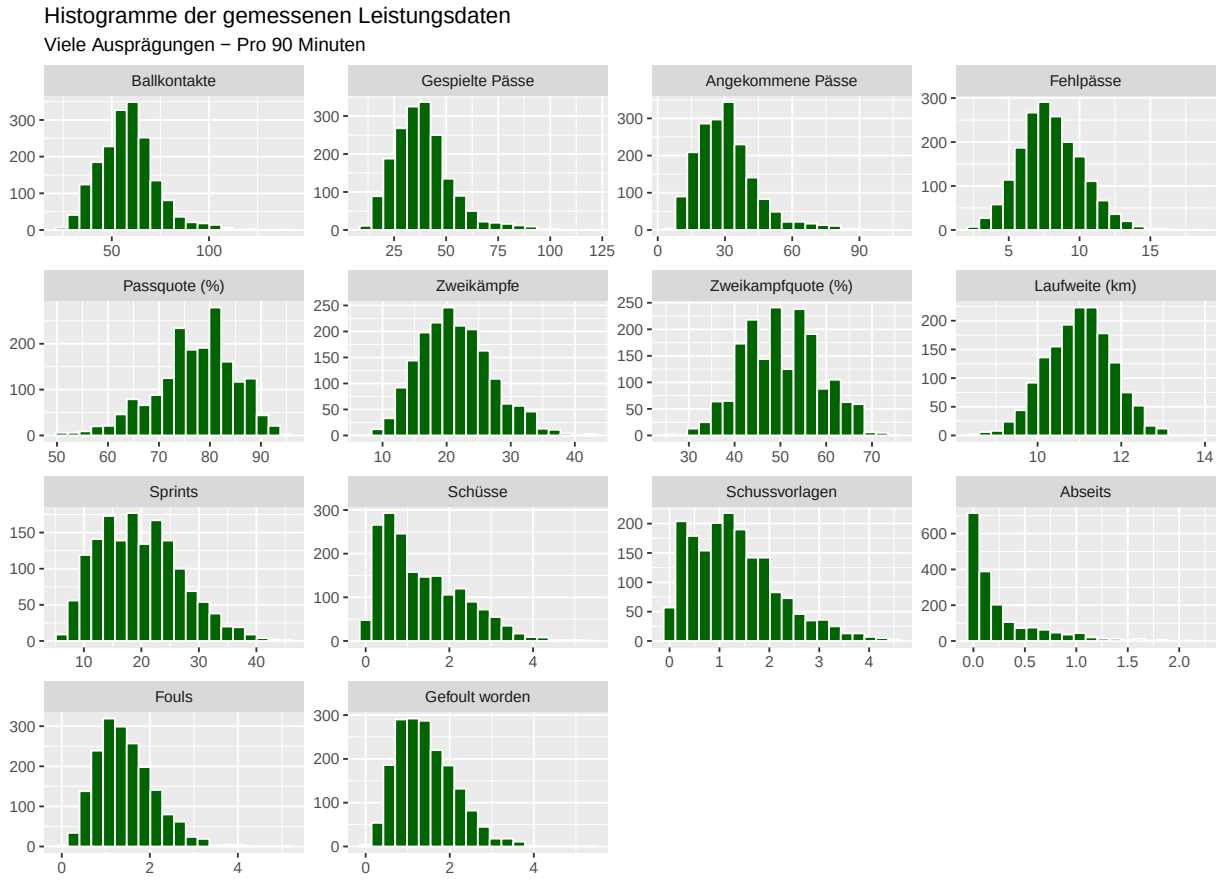


Abbildung 5: Visualisierung der relativen Leistungsdaten durch Histogramme - Viele Ausprägungen - Pro 90 Minuten

Eine große Veränderung ist in der *Laufweite* zu sehen. Diese weist jetzt eine annähernd normalverteilte Form auf. Die weiteren Leistungsdaten ändern die Form der Verteilung nur leicht, der größte Unterschied ist der neue Wertebereich.

Die Leistungsdaten, die zuvor wenige Ausprägungen aufweisen konnten, weisen nun durch das Relativieren eine höhere Anzahl an verschiedenen Ausprägungen auf, wie die Histogramme in Abbildung 6 zeigen.

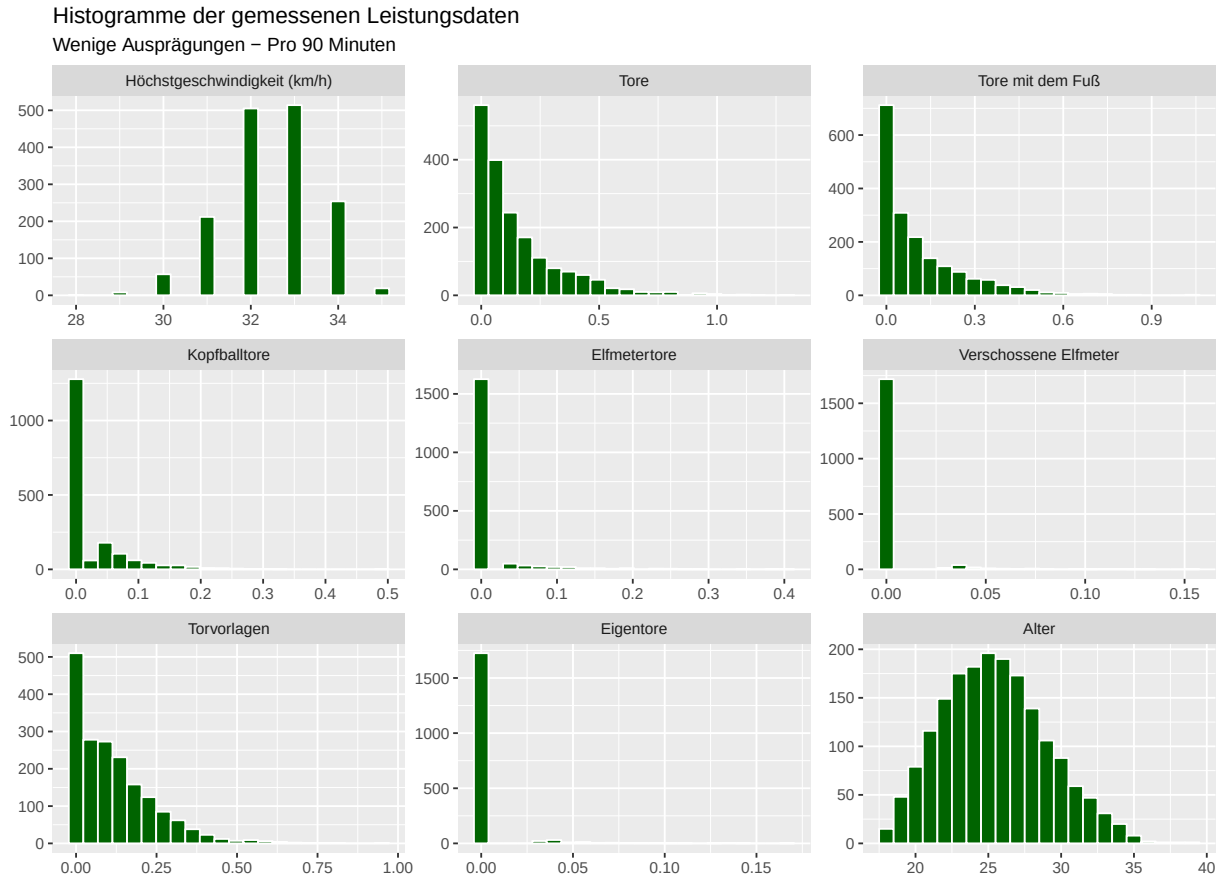


Abbildung 6: Visualisierung der relativen Leistungsdaten durch Histogramme - Wenige Ausprägungen - Pro 90 Minuten

Mit dem die Anzahl der unterschiedlichen Ausprägungen durch das Relativieren steigt, werden abgesehen von der Anzahl der Beobachtungen mit einer *0* als Ausprägung (bspw. weisen Beobachtungen mit *0* erzielten *Kopfballtoren* auch beim Relativieren *0* erzielte *Kopfballtore pro Spiel* auf) die Anzahlen pro Balken im Histogramm kleiner. Visuell entsteht dadurch ein stärkerer Effekt im Vergleich von der Anzahl der *0*en mit den restlichen Ausprägungen. Tendenziell bleiben jedoch alle Verteilungen auch nach Relativieren ihrer Form treu (z.B. linkssteile Verteilungen bleiben nach Relativieren linkssteil).

Auch für die relativen Torhüter-Leistungsdaten wird eine eigenständige Betrachtung durchgeführt, damit ein Überblick über diese Daten gegeben werden kann. Dies ist in Abbildung 7 dargestellt.

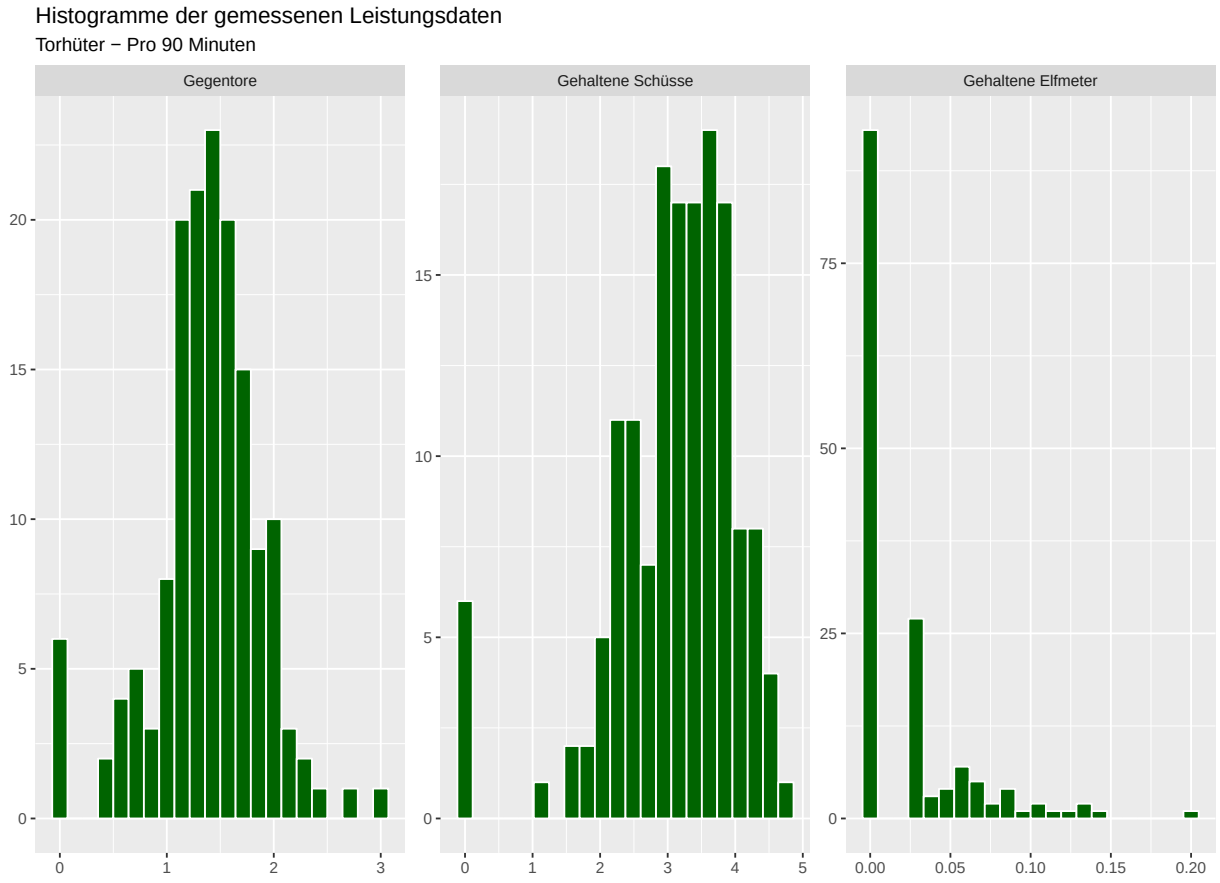


Abbildung 7: Visualisierung der relativen Leistungsdaten der Torhüter - Pro 90 Minuten

3.2.2 Zusammenhänge der relativen Leistungsdaten

An den Korrelationen nach Pearson zwischen den Leistungsdaten finden einige Veränderungen statt, wenn diese nicht absolut, sondern relativ betrachtet werden. Warum dies zu großen Unterschieden führen kann, ist in folgendem Beispiel dargestellt:

- Spieler **A** spielt in 100 Minuten 40 Pässe und 10 Fehlpässe.
- Spieler **B** spielt in 400 Minuten 200 Pässe und 20 Fehlpässe.

Die beiden Variablen *Pässe* und *Fehlpässe* wären in diesem Beispiel positiv korreliert. Werden die Daten jedoch auf ihre Spielminuten bezogen, dann ergibt sich folgende Situation:

- Spieler **A** spielt pro Minute 0.4 Pässe und 0.1 Fehlpass.
- Spieler **B** spielt pro Minute 0.5 Pässe und 0.05 Fehlpässe.

Auf die Spielminuten bezogen ergibt sich für dieses Beispiel eine negative Korrelation zwischen den beiden Variablen *Pässe* und *Fehlpässe*.

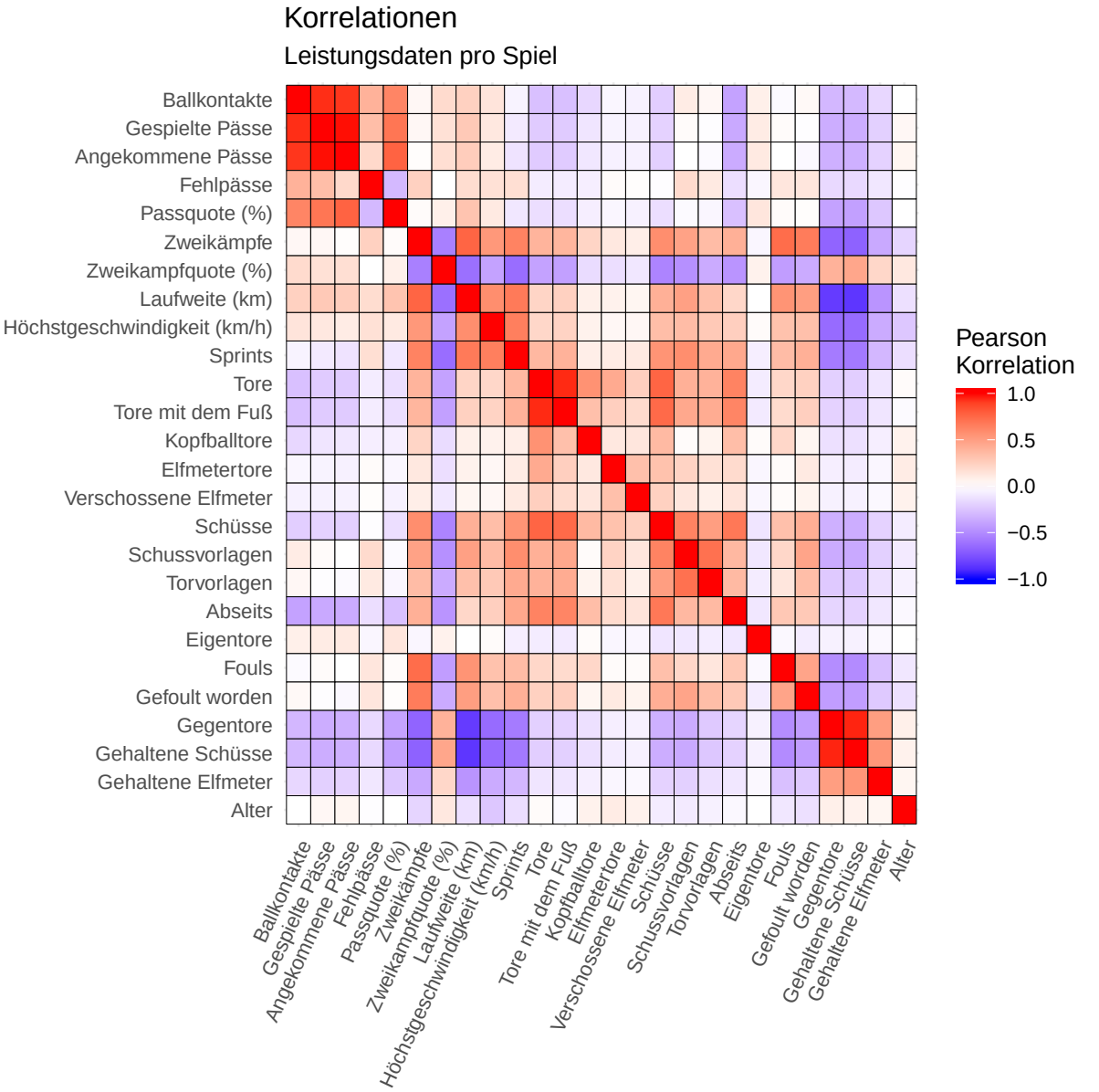


Abbildung 8: Visualisierung der Korrelation nach Pearson der relativen Leistungsdaten

In Abbildung 8 sind die Korrelationen nach relativieren der Leistungsdaten abgebildet. Die hohen positiven Korrelationen, die vorher zwischen der *Laufweite* und den Leistungsdaten, die die Spielbeteiligung beschreiben, gemessen werden konnten, sind nun gegen *0* gesunken.

Die Torhüter-Leistungsdaten *Gegentore* und *Gehaltene Schüsse* haben vorher schwach negative Korrelationen zu den körperlichen Leistungsdaten *Laufweite*, *Höchstgeschwindigkeit* und Anzahl an *Sprints* aufgewiesen. Diese sind durch die relative Betrachtung jedoch deutlich

negativer korreliert.

3.3 Starker Fuss Verteilung auf dem Spielfeld

Wer das ein oder andere Fußballspiel in der Kreisliga verfolgt hat, dem ist bestimmt schon aufgefallen, dass jeder meint er sei ein Experte darin zu wissen, auf welcher Position ein Links-, bzw. ein Rechtsfüßler zu spielen hat. Begründung dafür sind zum Beispiel, dass ein Verteidiger auf dem äußeren Fuß, also ein Rechtsverteidiger auf dem rechten Fuß und ein Linksverteidiger auf dem linken Fuß stark sein muss, um den angreifenden Flügelspieler einfacher am Flanken hindern zu können. Genauso benötigt ein Flügelstürmer einen guten äußeren Fuß, um Flanken zu können. Andere wiederum sind der Meinung, dass der Flügelstürmer einen guten inneren Fuß haben muss, damit er gefährlicher aufs Tor schießen kann.

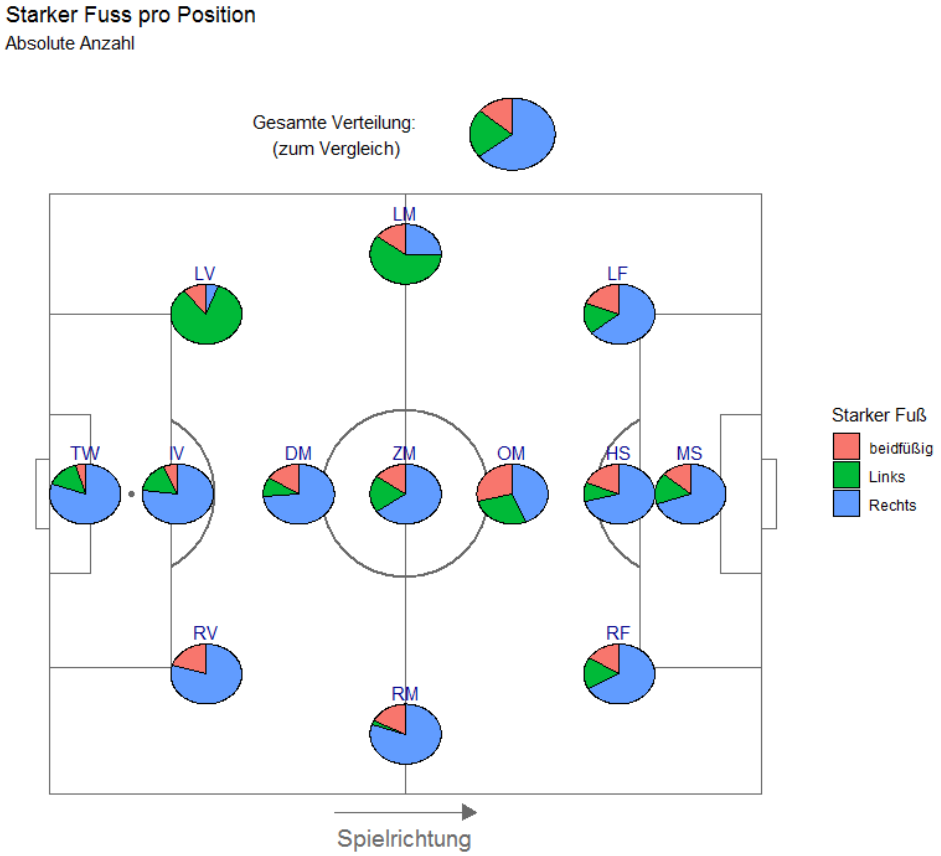


Abbildung 9: Verteilung des starken Fußes auf dem Spielfeld in der Bundesliga

Auf der Homepage von *The Guardian* ist ein Artikel von 2010, der von dieser Thematik handelt (Wilson 2010). In diesem Artikel wird diskutiert, dass immer mehr Flügelspieler auf der 'falschen' Seite spielen. Hier wird ein weiteres Argument gebracht, warum ein Flügelspieler auf der 'falschen' Seite effektiv ist. Ein Verteidiger hat in der Regel einen starken äußeren und schwachen inneren Fuß, damit Flanken verhindert werden können. Ein Flügelspieler

mit einem starken äußeren Fuß tritt also gegen den starken äußeren Fuß des Verteidigers an, wohingegen ein Flügelspieler mit einem stärkeren inneren Fuß gegen den schwächeren inneren Fuß des Verteidigers antritt.

Die Verteilung des starken Fußes der Bundesliga ist in Abbildung 9 dargestellt. Was hier auffällt, ist die Verteilung des starken Fußes bei den Außenverteidigern und den äußeren Mittelfeldpositionen. Während auf der linken Seite vorwiegend Linksfüßler spielen, spielen auf der rechten Seite vorwiegend Rechtsfüßler. Dass auf den Außenverteidigerpositionen also Spieler mit starkem äußeren Fuß eingesetzt werden ist hier deutlich zu erkennen.

Auf der anderen Seite ist zwischen den Verteilungen der linken und rechten Flügelspieler kaum ein Unterschied zu erkennen. Wie auch in der gesamten Verteilung des starken Fußes zu erkennen ist, gibt es ein etwa 65%-iges Übergewicht an Rechtsfüßlern auf beiden Positionen. Das bedeutet, dass viele Rechtsfüßler auf dem entgegengesetzten Flügel spielen, jedoch nur wenige Linksfüßler. Dies könnte jedoch keine taktischen Gründe haben, sondern durch die Verfügbarkeit von Rechts- und Linksfüßlern zu erklären sein. Wenn es nur wenige Linksfüßler gibt, können auch nur weniger taktisch eingesetzt werden.

Eine weitere interessante Erkenntnis ist der weit überdurchschnittlich hohe Anteil an beidfüßigen Spielern und die Unterbesetzung der Rechtsfüßler auf der Position des *offensiven Mittelfelds*. Ein Spieler auf der Position des *offensiven Mittelfelds* hat die Aufgabe Chancen zu kreieren, Abschlüsse zu suchen und die Bälle gezielt zu verteilen. Diese Position ist von einem sehr hohen Anteil an Kreativität geprägt. Im *Journal of Nervous and Mental Disease* wurde 2007 in einem Artikel von Preti und Vellante eine Verbindung zwischen kreativen Künstlern und ihrer starken Hand untersucht (Preti and Vellante 2007). In dieser Studie wurde gemessen, dass der Anteil an nicht-Rechtshändern bei kreativen Menschen größer ist als in ihrer Kontrollgruppe. Wenn sich dies auf Leute übertagen lässt, die nicht-Rechtsfüßler sind, dann könnte dadurch ein natürlicher Zusammenhang gefunden werden, wieso sich die Verteilung des starken Fußes auf der kreativen Position des offensiven Mittelfeldspielers so sehr von der Gesamtpopulation unterscheidet.

Eine weitere Erklärung könnte die Notwendigkeit beider Füße auf dieser Position sein. Ein *offensiver Mittelfeldspieler* muss in der Lage sein den Ball auf beide Flügel zu verteilen. Ist ein Spieler nur rechts- oder linksfüßig, so würde es ihn einen unnatürlichen extra Aufwand kosten den Ball in die "unnatürliche" Richtung nach Außen zu spielen. Gemeint ist damit beispielsweise, dass ein Rechtsfüßler mit der Innenseite seines rechten Fußes den Ball mit Blickrichtung zum gegnerischen Tor einfach auf die linke Seite passen kann. Wenn er einen Pass auf die rechte Seite spielen möchte, muss er entweder die in der Regel unpräzisere äußere Seite des Fußes nutzen oder sich erst umdrehen, um den Pass mit dem rechten Fuß zu spielen. Ist ein Spieler jedoch mit beiden Füßen stark, kann er schnell und präzise den Ball auf beide Seiten des Spielfelds verteilen.

Diese beiden möglichen Erklärungen müssten jedoch erst in einer aufwendigen Studie untersucht werden, um sie zu bestätigen, was in dieser Arbeit jedoch nicht mehr getan wird.

3.4 Auswahl der Leistungsdaten für die Modellierung

Für eine durchdringende Analyse und eine gute Modellierung muss eine Auswahl an Variablen getroffen werden, die dafür relevant sind. Da beispielsweise die Anzahl der *angekommenen Pässe* und die Anzahl der *Fehlpässe* die Gesamtanzahl der *Pässe* ergeben, würde dadurch eine lineare Abhängigkeit der Variablen entstehen, was in einer Modellierung zum Problem der Multikollinearität führt.

Darüber hinaus gibt es Variablen, die sehr hoch miteinander korreliert sind, weswegen es sinnvoll wäre nur eine der beiden in das Modell aufzunehmen, wie zum Beispiel die Anzahl der *geschossenen Tore* und die Anzahl der *mit dem Fuß geschossenen Tore*.

Andere Variablen wiederum weisen die Problematik auf, dass sie nur Momentaufnahmen sind und nicht die Leistung eines Spielers über mehrere Spiele widerspiegeln, wie zum Beispiel die *Höchstgeschwindigkeit*.

Alles in allem wurde das Variablen-Set auf **11** relevante Variablen reduziert (siehe Tabelle 3). Hier sind nicht die absoluten Werte der Leistungsdaten gemeint, sondern die auf ihre Spielminuten bezogenen Werte.

Variable	Beschreibung
Gespielte Pässe	Misst die Spielbeteiligung eines Spielers mit Ball
Angekommene Pässe (in %)	Misst die Qualität der Pässe
Geführte Zweikämpfe	Misst die Spielbeteiligung eines Spielers mit und gegen den Ball
Gewonnene Zweikämpfe (in %)	Misst die Qualität der Zweikämpfe
Begangene Fouls	Misst, wie häufig ein Foul begangen werden muss, um einen Gegner zu stoppen
Gefoult worden	Misst, wie häufig der Gegner foulen muss, um den Spieler zu stoppen
Laufweite	Misst die körperliche Ausdauerleistung eines Spielers
Abseits	Misst die offensive Einsatzbereitschaft eines Spielers
Vorlagen	Misst die Fähigkeit ein Tor vorzubereiten
Geschossene Tore mit dem Fuss	Misst die Fähigkeit Tore zu erzielen
Geschossene Tore mit dem Kopf	Misst die Kopfballstärke eines Spielers

Tabelle 3: Ausgewählte Variablen

Diese Variablen werden für die Modellierungen verwendet. In Abbildung 10 ist die Korrelation der ausgewählten Variablen dargestellt. Die höchste Korrelation nach Pearson besteht zwischen der Anzahl der *geführten Zweikämpfe* und der *Laufweite* mit 0.76. Trotz dieser sehr hohen Korrelation sollen beide Variablen für die Modellierung betrachtet werden, da sie verschiedene Leistungen eines Spielers messen.

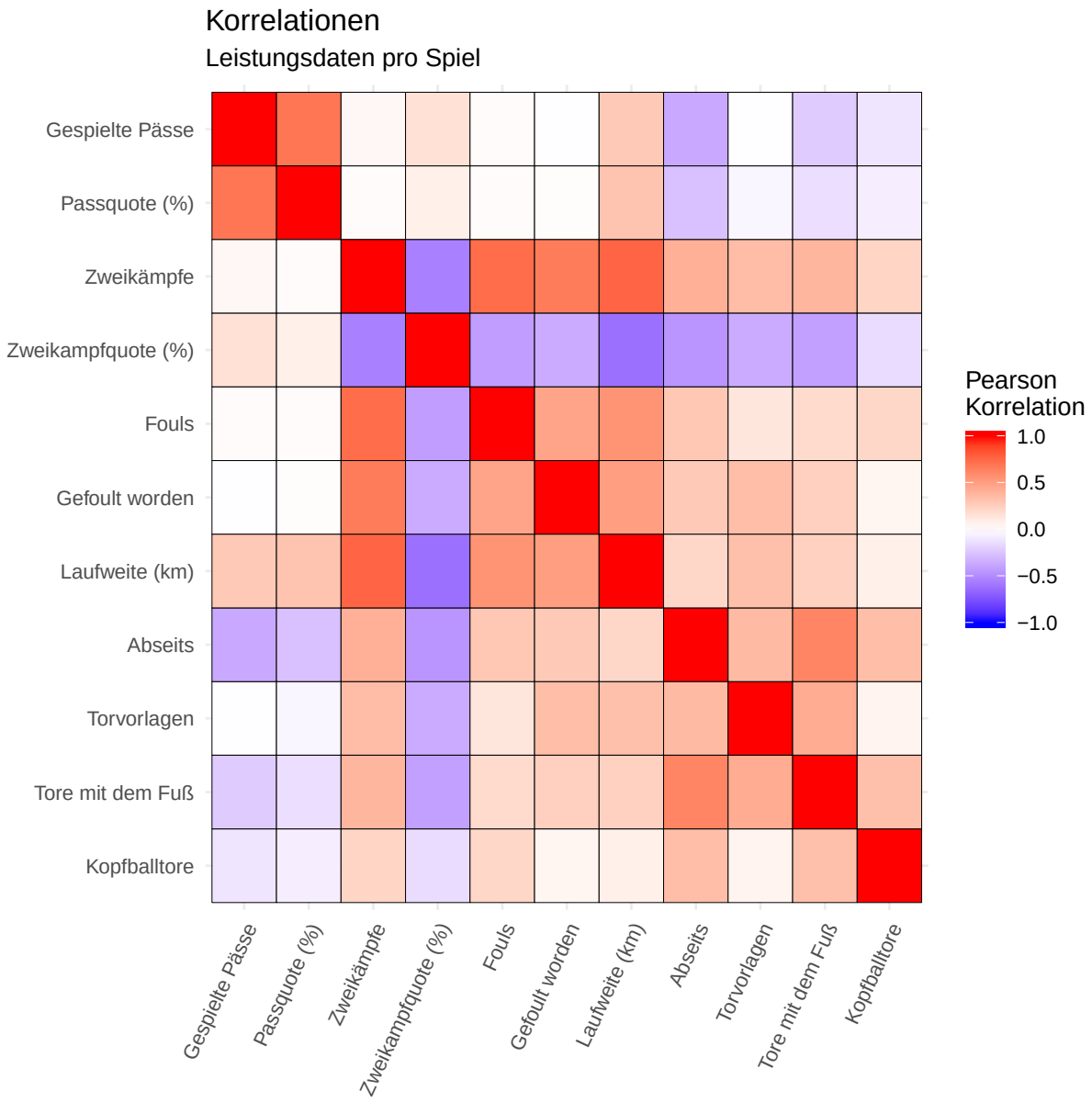


Abbildung 10: Korrelation der ausgewählten Modellvariablen

3.5 Zusammenfassen von Positionen

Wie bereits in Tabelle 2 auf Seite 5 zu sehen ist, sind manche Klassen schwächer besetzt als andere. Dies gilt vor allem für die *äußeren Mittelfeldpositionen* und die *Hängende Spitze*. Um fehlerhafte Analysen durch unterbesetzte Klassen zu vermeiden, werden daher Positionen, die (in etwa) diegleiche Funktion auf dem Spielfeld haben, zusammengefasst.

Um durch dieses Vorgehen die Analysen nicht zu verfälschen wird mit bonferroni-korrigierten t-Tests untersucht, ob sich die Positionen bezüglich ihrer Leistungsdaten signifikant unter-

scheiden. Die Bonferroni-Korrektur ist nötig, da multiple Tests gleichzeitig betrachtet werden (und zwar 1 Test pro Leistungsdatum, das überprüft wurde).

Die Beobachtungen, die dabei überprüft werden, sind jedoch nicht unabhängig, da eine Person sowohl auf dergleichen Position, als auch auf den beiden Positionen, die verglichen werden, mehrere Saisons gespielt haben kann. Daher muss der Datensatz zufällig auf einen Teildatensatz reduziert werden, indem jeder Spieler genau einmal vorkommt. Dies wäre jedoch nur eine Aufnahme für einen einzelnen zufällig gezogenen Teildatensatz und könnte dem Zufall geschuldet Unterschiede aufweisen, die im Gesamtdatensatz jedoch nicht vorhanden sind. Daher werden diese t-Tests 100 mal mit verschiedenen Teildatensätzen wiederholt.

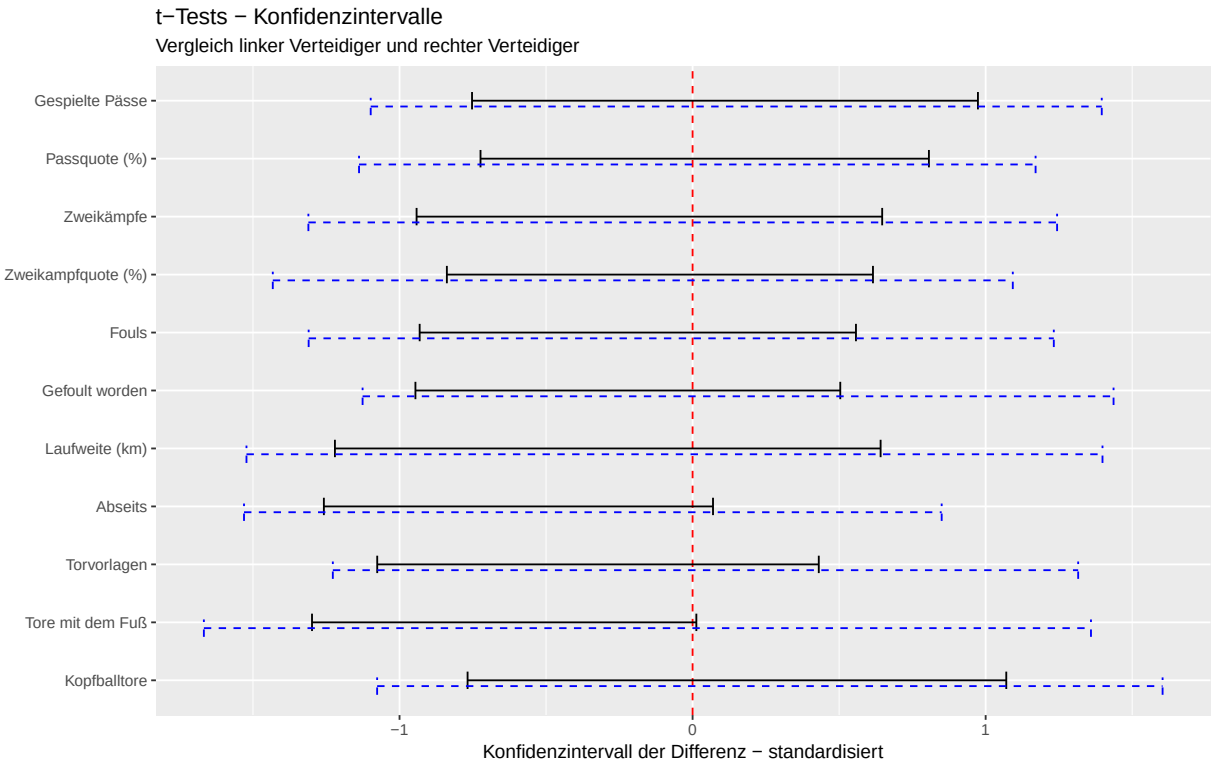


Abbildung 11: Vergleich zwischen *linken* und *rechten Verteidigern*

In Abbildung 11 ist der Vergleich der Leistungsdaten zwischen den linken und rechten Verteidigern abgebildet. Das schwarz eingezeichnete Intervall bildet das Konfidenzintervall der Differenz des jeweiligen Leistungsdatums für **einen einzelnen** beispielhaften Teildatensatz ab. Das blaue Intervall bildet das Minimum und das Maximum der Konfidenzintervalle der Differenz des jeweiligen Leistungsdatums für **alle** Teildatensätze ab. Um die Intervalle miteinander vergleichbar zu machen, wurden sie standardisiert. Der Wert 1 auf der x-Achse bedeutet, dass dieser Punkt eine Standardabweichung von der 0 entfernt ist. Enthält das blaue Intervall die 0, so wird keine signifkante Differenz dieses Leistungsdatums zwischen den beiden betrachteten Gruppen festgestellt. Enthält dieses Intervall die 0 nicht, so wurde eine signifikante Differenz dieses Leistungsdatums zwischen den beiden betrachteten Gruppen festgestellt.

Wie in Abbildung 11 zu sehen ist, unterscheiden sich die linken und rechten Verteidiger nicht

signifikant, weshalb diese beiden Positionen zu einer gemeinsamen *Außenverteidiger*-Position zusammengefasst werden. Wie den weiteren untersuchten Paaren (s. Anhang) zu entnehmen ist, unterscheiden sich auch die *linken* und *rechten Mittelfeldspieler*, die *Links-* und *Rechtsaußen* Spieler und die Spieler auf der *hängenden Spitze* und die *offensiven Mittelfeldspieler* nicht. Die *linken* und *rechten Mittelfeldspieler* werden als *Mittelfeld Außen*-Position zusammengeführt, die *Links-* und *Rechtsaußen* Spieler werden als *Flügelspieler* zusammen gefasst und die als *Hängende Spitze* Spielenden werden zusammen mit den *offensiven Mittelfeldspielern* als gemeinsames *Offensives Mittelfeld* betrachtet.

Das Verwenden der einfachen Bonferroni-Korrektur ist ein sehr konservativer Ansatz. Das bedeutet, dass ein möglicherweise siginifkanter Effekt nicht erkannt werden würde. Das durch 100 Simulationen erzeugte Intervall verbreitert dieses bonferroni-korrigierte Konfidenzintervall noch weiter, was zu einem zu konservativen Intervall führen könnte, das kaum signifikante Effekte erfassen würde. Um zu demonstrieren, dass dies doch geschehen kann, wenn zwei wirklich unterschiedliche Gruppen untersucht werden würden, ist in Abbildung 12 ein Vergleich von *Innenverteidigern* und *offensiven Mittelfeldspielern* aufgeführt, in dem deutlich signifikante Effekte zu erkennen sind. In einem solchen Fall würde das Verbinden dieser beiden Gruppen nicht erlaubt sein.

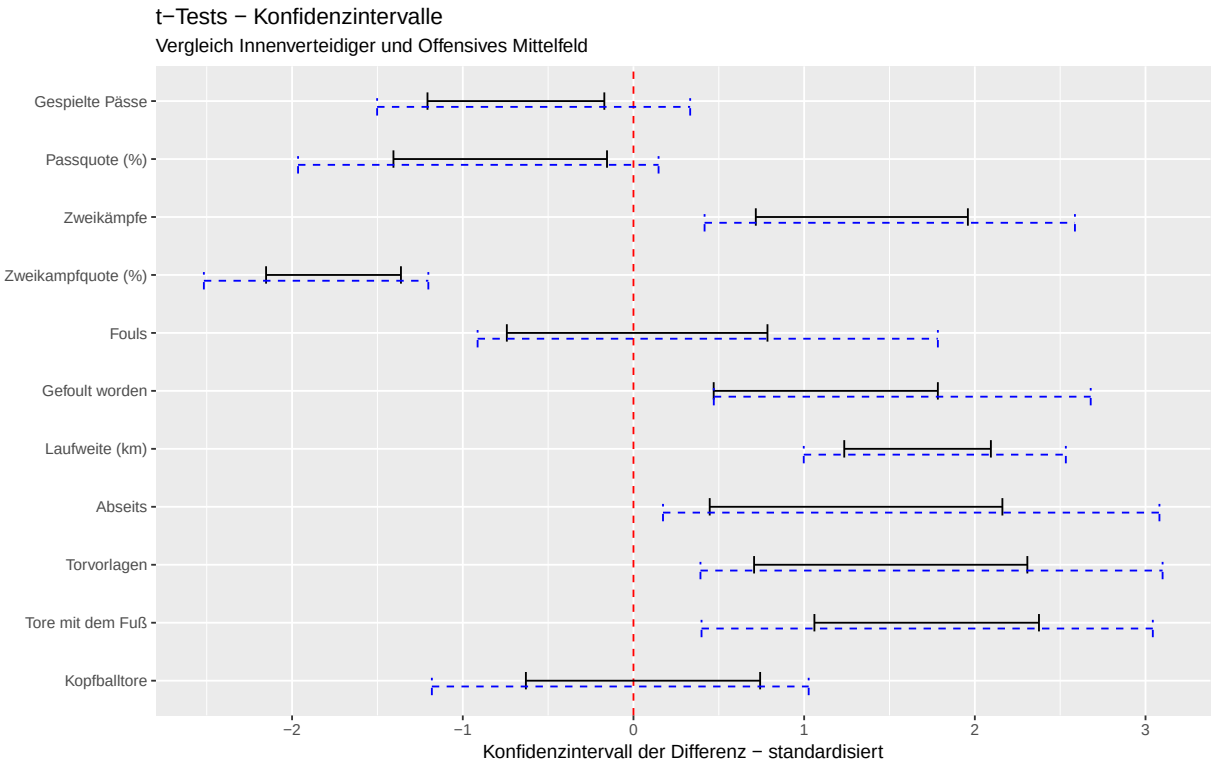


Abbildung 12: Vergleich zwischen *Innenverteidigern* und *offensives Mittelfeldspielern*

Durch diese Klassenzusammenführung ergibt sich eine neue Verteilung der Positionen, die in Tabelle 4 dargestellt ist.

Position	Anzahl an Beobachtungen
Torwart	154
Außenverteidiger	319
Innenverteidiger	366
Libero	1
Defensives Mittelfeld	246
Mittelfeld Außen	94
Zentrales Mittelfeld	146
Offensives Mittelfeld	162
Flügelspieler	273
Mittelstürmer	219

Tabelle 4: Anzahl der Beobachtungen pro Position

3.6 Mittlere Leistungsdaten pro Position

Um einen Eindruck davon zu erhalten, auf welchen Positionen welche Leistungsdaten besonders hoch ausgeprägt sind, werden alle Leistungsdaten bezüglich ihrer Position gemittelt und in Radarplots miteinander verglichen. Um zu vermeiden, dass durch zu viele Klassen die Übersicht verloren geht, wurden die **defensiven Leistungsdaten** für die **defensiven Positionen** und die **offensiven Leistungsdaten** für die **offensiven Positionen** in Abbildung 13 und Abbildung 14 dargestellt. Für diese Visualisierung wurden auch die Variablen verwendet, die nicht für die Modellierung hinzugenommen wurden.

Die in den Radarplots abgebildeten Werte $m_{k,l}*$ werden für die k Positionen und l Variablen mit Formel (1) und Formel (2) ermittelt.

$$m_{k,l} = \frac{1}{N_k} \sum_{i=1}^{N_k} (x_{k,l})_i, \tag{1}$$

wobei N_k die Anzahl der Beobachtungen für Position k darstellt und $(x_{k,l})_i$ den i-ten Wert für Position k und Leistungsdatum l.

Diese Mittelwerte pro Position und Leistungsdatum werden für die Visualisierung mit einem Radarplot auf einen Wertebereich zwischen 0 und 1 skaliert, wobei 0 der natürliche Nullpunkt darstellt und 1 das Maximum der Werte $m_{\bullet,l}$ pro Leistungsdatum. Diese Umskalierung geschieht durch Formel (2).

$$m_{k,l}* = \frac{m_{k,l}}{max(m_{\bullet,l})}, \tag{2}$$

wobei $m_{\bullet,l}$ den Vektor $m_{k,l}$ über alle k für ein festes Leistungsdatum l darstellt.

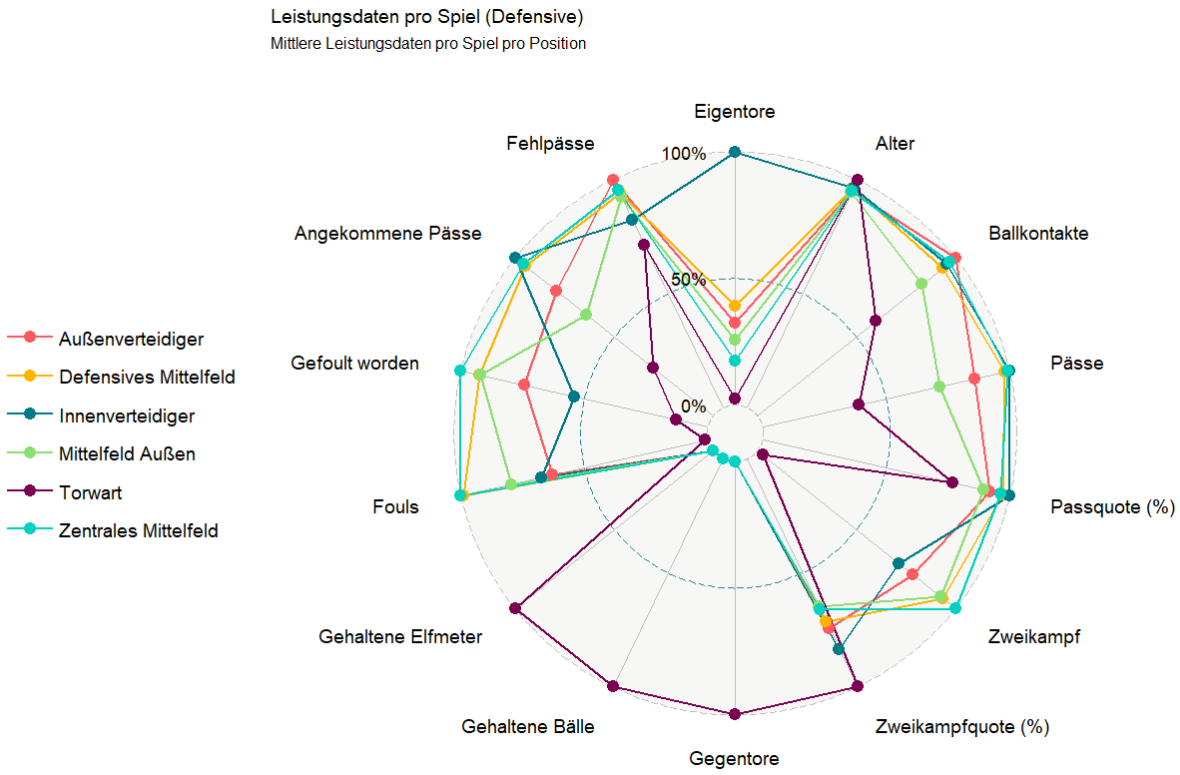


Abbildung 13: Mittlere defensive Leistungsdaten pro Position

In Abbildung 13 fallen sehr schnell die Variablen auf, die nur für *Torhüter* Werte enthalten. In diesen Variablen dominieren natürlich die *Torhüter* im Vergleich zu den anderen Positionen. Darüber hinaus dominieren *Torhüter* in der *prozentualen Anzahl gewonnener Zweikämpfe* und haben ein leicht überdurchschnittlich hohes *Alter* im Vergleich zu den restlichen Positionen. Bei allen anderen Variablen belegen sie deutlich den letzten Platz.

Während die *Außenverteidiger* die meisten *Ballkontakte* und etwas mehr *Fehlpässe* als der Rest aufweisen, weisen die übrigen Spielanteilsvariablen wie die Anzahl *gespielter Pässe*, die Anzahl der *angekommenen Pässe* die Anzahl der *prozentual angekommenen Pässe* und die Anzahl der *prozentual gewonnenen Zweikämpfe* (nach den *Torhütern*) bei den *Innenverteidigern* die höchsten Werte auf. Sehr dominant sind die Zahlen der *Eigentore* bei den *Innenverteidigern*.

Die meisten *Zweikämpfe* und die meisten *Fouls* finden im *zentralen Mittelfeld* statt. Die Spieler dort werden darüber hinaus noch am häufigsten *gefoult*.

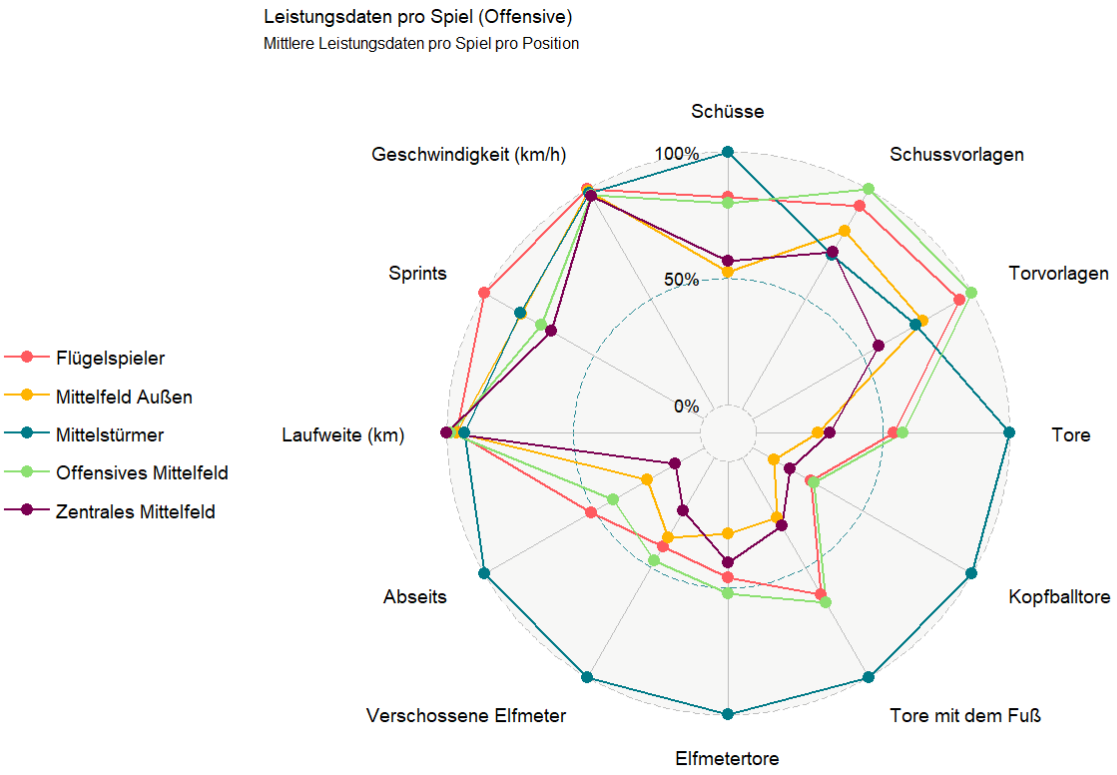


Abbildung 14: Mittlere offensive Leistungsdaten pro Position

In Abbildung 14 fallen die Werte der *Mittelstürmer* sehr schnell auf, da sie in 7 der 12 offensiven Kategorien die höchsten Werte aufweisen, nämlich in der Anzahl der abgegebenen *Schüsse*, der *Tore*, der *Kopfballtore*, der *Tore mit dem Fuß*, der *Elfmetertore*, der *verschossenen Elfmeter* und der *Abseitsstellungen*.

Die meisten *Schussvorlagen* und dann zum Tor führende *Vorlagen* werden von den *offensiven Mittelfeldspielern* abgegeben, dicht gefolgt von den *Flügelspielern*.

Die *Flügelspieler* weisen die höchsten Werte in den schnellen körperlichen Kategorien auf. In der durchschnittlichen Anzahl der *Sprints* führen die *Flügelspieler* deutlich, während sie in der durchschnittlichen *Höchstgeschwindigkeit* nur knapp führen.

Die höchste *Laufweite* weisen die *zentralen Mittelfeldspieler* auf. Diese und die durchschnittliche *Höchstgeschwindigkeit* weisen jedoch zwischen den Positionen die geringsten Differenzen auf.

4 Modellierung der Daten

4.1 Modellierungsziel

Um zu verstehen, wie sich die Leistungsdaten zwischen den verschiedenen Positionen unterscheiden, wird eine Modellierung vorgenommen, in der die Position anhand der Leistungsdaten eines Spielers klassifiziert wird. Mit interpretierbaren Machine Learning Methoden, die auch auf klassische Modelle angewendet werden können, wird dann versucht die Beziehung zwischen den Leistungsdaten und den verschiedenen Positionen zu erarbeiten und somit die Modelle zu vergleichen.

Die Beziehung zwischen den Positionen und den Leistungsdaten wird wie folgt formuliert:

$$\hat{y}_i = f(x_i), \tag{3}$$

mit $\hat{y}_i = PositionSpieler_i$, $x_i = LeistungsdatenSpieler_i$ und $f(\cdot)$ eine Funktion, bzw. ein Modell, das eine Entscheidung darüber fällt, auf welcher Position ein Spieler gegeben seiner Leistungsdaten gespielt hat.

Ziel ist die Interpretation der Beziehung zwischen x und $f(x)$ und zu verstehen, wie die Beobachtungen bezüglich der verschiedenen Positionen im Raum der Leistungsdaten verteilt sind. Darüber hinaus soll die Beziehung zwischen zwei geeigneten, aber verschiedenen, Modellierungen miteinander vergleichbar gemacht werden.

4.2 Methoden

4.2.1 Modellauswahl

Für die Modellierung soll ein Modell aus der Familie der Regressionsanalysen für Klassifikationen und eine gängige Klassifikationsart aus dem Bereich des *Machine Learning* verwendet werden.

Als klassische Regressionsanalyse bietet sich eine multinomiale logistische Regression an. Diese ist eine Erweiterung der binären logistischen Regression und modelliert den Zusammenhang zwischen einem Variablenvektor x und einer kategoriellen Zielgröße y mit $k \geq 2$ Klassen durch einen linearen Prädiktor, der die Chance einer Beobachtung einer Klasse anzugehören im Vergleich zu einer Referenzkategorie schätzt. Da jede Klasse einen Bezug zur Referenzkategorie aufweist, kann für jede Kategorie eine Wahrscheinlichkeit berechnet werden, mit welcher eine Beobachtung zu dieser Klasse gehört. Für diese Modellierung wird sich auf lineare Effekte der Einflussgrößen ohne Interaktionen oder quadratische Effekte beschränkt. Der Hauptgrund dafür ist, dass auch durch lineare Effekte bereits für 8 Klassen (9 Positionen minus eine Referenzkategorie) und 11 Variablen insgesamt 88 Koeffizienten geschätzt werden müssen und der Umfang der Daten nur knapp unterhalb von 2000 Beobachtungen liegt.

Für diese multinomiale logistische Regression wird die Funktion `multinom` aus dem R-Paket `nnet` verwendet. Diese Implementatierung der multinomialen logistischen Regression schätzt

die Regressionskoeffizienten über neuronale Netze. Diese neuronalen Netze updaten iterationsweise die Regressionskeoffizienten, damit der Kleinste-Quadrate-Fehler auf den Trainingsdaten minimiert wird. Dieses Verfahren approximiert den KQ-Schätzer für die Regressionskoeffizienten (Venables and Ripley 2002).

Aus dem Bereich des Machine Learning wird ein (Klassifikations-)*Random Forest* verwendet, wie ihn Breiman vorgeschlagen hat (Breiman 2001). Ein *Random Forest* modelliert die Beziehung zwischen den Einflussgrößen x und der kategoriellen Zielgröße y mit $k \geq 2$ Klassen durch viele möglichst unterschiedliche Bäume mit "guter" Prädiktionsgüte. Jeder Baum klassifiziert eine Beobachtung in eine der Kategorien, wodurch nach Betrachtung aller Bäume eine Gesamtwahrscheinlichkeit für jede Kategorie berechnet werden kann, zu der eine Beobachtung dieser Kategorie angehört.

Für diese Analyse wird die im R-Paket `ranger` verwendete Implementierung von *Random Forests* verwendet. Dies ist eine computational schnelle Implementatierung der *Random Forests*, die jedoch nachgewiesen keine schlechtere Prädiktion vorweist als die originale Implementierung von Breimans *Random Forests* in `R` (Wright and Ziegler 2017).

Die Funktionsweise der beiden Modellierungen wird im weiteren Teil dieser Arbeit an den passenden Stellen näher erläutert und anschließend miteinander verglichen.

4.2.2 Interpretierbares Machine Learning zur Vergleichbarkeit

4.2.2.1 Variable Importance

Während die multinomiale logistische Regression aus dem Bereich der Regressionsanalysen stammt und durch Regressionskoeffizienten leicht zu interpretieren ist, stammen die *Random Forests* aus dem Bereich des *Machine Learnings* und weisen keine leicht zu interpretierbaren Regressionskoeffizienten auf. Um die beiden Modellierungen miteinander vergleichbar zu machen werden interpretierbare *Machine Learning* Methoden angewendet, die auf beide Modelle anwendbar sind.

Eine mit dem *Random Forest* häufig verbundene Methode ist die Berechnung der **Variable Importance**. Die Variable Importance ist ein Maß um die Variablen gemäß ihrer "Wichtigkeit" in der Modellierung einzuordnen. Eine Variable Importance zu messen funktioniert auf verschiedene Arten. Für diese Arbeit wird die von Breiman 2001 vorgeschlagene Idee für das Messen der Variable Importance durch Permutation aufgegriffen.

Diese Implementierung des *Random Forests* beinhaltet eine automatische *Variable Importance* Berechnung nach der Methode "*permutation*". Diese Methode nutzt den Vorschlag von Breiman. Da die beiden Modelle jedoch fair verglichen werden sollen, wird eine eigene *Variable Importance* nach diesem Vorschlag von Breiman berechnet. Um eine Streuung für diese Methode zu erhalten, werden die beiden Modelle 100 mal mit verschiedenen Trainingsdatensätzen gefittet. Aus diesen 100 Wiederholungen wird dann für jede Variable eine Streuung bestimmt.

Um die für die *Random Forests* berechnete *Variable Importance* mit der multinomialen logistischen Regression zu vergleichen, wird die *Variable Importance* wie folgt per Hand berechnet:

1. Ein Trainings- und Testdatensatz wird generiert
2. Das geschätzte Modell $\hat{f}(x)$ wird auf Trainingsdaten gefittet
3. Ein Gütemaß des Modells wird für einen Testdatensatz berechnet
4. Der Testdatensatz wird durch Permutieren einer einzelnen Variable verändert
5. Die Verschlechterung des Gütemaßes wird berechnet
6. Schritt 4. und 5. werden für jede Variable wiederholt

Als Trainings- und Testdatensplitrate wird ein $\frac{2}{3}$ Trainingsdaten- und $\frac{1}{3}$ Testdaten-Split gewählt. Um eine Streuung für die *Variable Importance* zu schätzen, wird dieser Vorgang 100 mal wiederholt. Als Punktschätzer für die *Variable Importance* wird das arithmetische Mittel für jede Variable berechnet, wodurch eine Rangfolge der Variablen bezüglich ihrer Wichtigkeit für das Modell bestimmt wird.

Jeder einzelne Baum eines *Random Forests* könnte als eigenständiges Modell betrachtet werden. Die *Variable Importance*, die durch das Ensemble der einzelnen Bäume generiert wird, ist also bereits aufgrund von vielen Modellen gemittelt. Dadurch werden für die *Random Forest Variable Importance* etwas kleinere Streuungen erwartet. Die Punktschätzung, und damit auch die Rangfolge zwischen beiden Modellen, bleibt jedoch vergleichbar.

4.2.2.2 Partial Dependence Plots

Ein Partial Dependence Plot (kurz PDP) ist eine Visualisierungstechnik, die helfen soll den marginalen Effekt einer bestimmten Variable auf eine Zielgröße über ihren kompletten Wertebereich zu visualisieren.

Die Idee der *partial dependence* ist es den Effekt einer oder mehrerer Variablen in dem Modell durch das Integrieren über die marginale Verteilung der übrigen Kovariablen zu erhalten (Friedman 2001).

Sei x_l die interessierende Variable und $x_{\setminus l}$ die Kovariablen ohne x_l, dann ist

$$\mathbb{E}_{x_{\setminus l}}(\hat{f}(x)) = \int \hat{f}(x_l, x_{\setminus l}) p_{\setminus l}(x_{\setminus l}) dx_{\setminus l} \tag{4}$$

eine Funktion, die die *partial dependence* für x_l bedingt auf $x_{\setminus l}$ abbildet. $p_{\setminus l}(x_{\setminus l})$ ist hier die marginale Wahrscheinlichkeitsfunktion von $x_{\setminus l}$, welche aus den Trainingsdaten ermittelt werden kann (Friedman 2001). Dies funktioniert jedoch nur dann, wenn die Abhängigkeiten zwischen den Kovariablen nicht zu stark sind.

Da $p_{\setminus l}(x_{\setminus l})$ aus den Trainingsdaten ermittelt werden kann, kann (4) zu

$$\overline{f_l}(x_l) = \frac{1}{N} \sum_{i=1}^{n} \hat{f}(x_l, x_{i,\setminus l}) \tag{5}$$

umgeformt werden. Dies bedeutet, dass die *partial dependence* an einem bestimmten Punkt (oder für eine bestimmte Kovariablen-Kombination, falls die *partial dependence* für mehrere Variablen gebildet werden soll) als Durchschnitt über die Prädiktionen des Modells $\hat{f}(x)$ für alle Beobachtungen gebildet wird, wobei x_l einem fixen Wert entspricht.

Algorithmisch wird der Partial Dependence Plot wie folgt für eine bestimmte metrische Variable x_l erzeugt:

1. Definiere Punkte $q = q_1, ...q_r$ innerhalb des Wertebereichs der Variable x_l, an denen der PDP berechnet werden soll
2. Berechne für jede Beobachtung die Prädiktion, mit $x_m = \begin{cases} x_m, & \text{wenn } m \neq l, \\ q_p, & \text{wenn } m = l \end{cases}$
3. Berechne das arithmetische Mittel für die neuen Prädiktionen aller Beobachtungen
4. Wiederhole Schritt 2. und Schritt 3. für alle $p = 1, ..., r$
5. Plotte den gemittelten Verlauf der Prädiktion über den Wertebereich der Variable x_l

Für die Visualisierung in dieser Analyse werden für die Werte q die empirischen Perzentile der jeweiligen Variable x_l betrachtet (also das 0%-Quantil, das 10%-Quantil, das 20%-Quantil...). Da die Prädiktion eine Klassifikation ist, wird für jeden Auswertungspunkt q für jede Beobachtung die Klassenzugehörigkeitswahrscheinlichkeit berechnet, und diese dann für jede Klasse über alle Beobachtungen hinweg separat gemittelt.

Die daraus resultierende Visualisierung der gemittelten Klassenwahrscheinlichkeiten und des Wertebereichs einer Variable gibt den marginalen Effekt dieser Variable auf die verschiedenen Klassenwahrscheinlichkeiten für dieses Modell an. Damit kann festgestellt werden, ob in einem geschätzten Modell $\hat{f}(\cdot)$ eine bestimmte Variable einen linearen, quadratischen oder unstrukturierten Effekt auf die Klassenwahrscheinlichkeit hat.

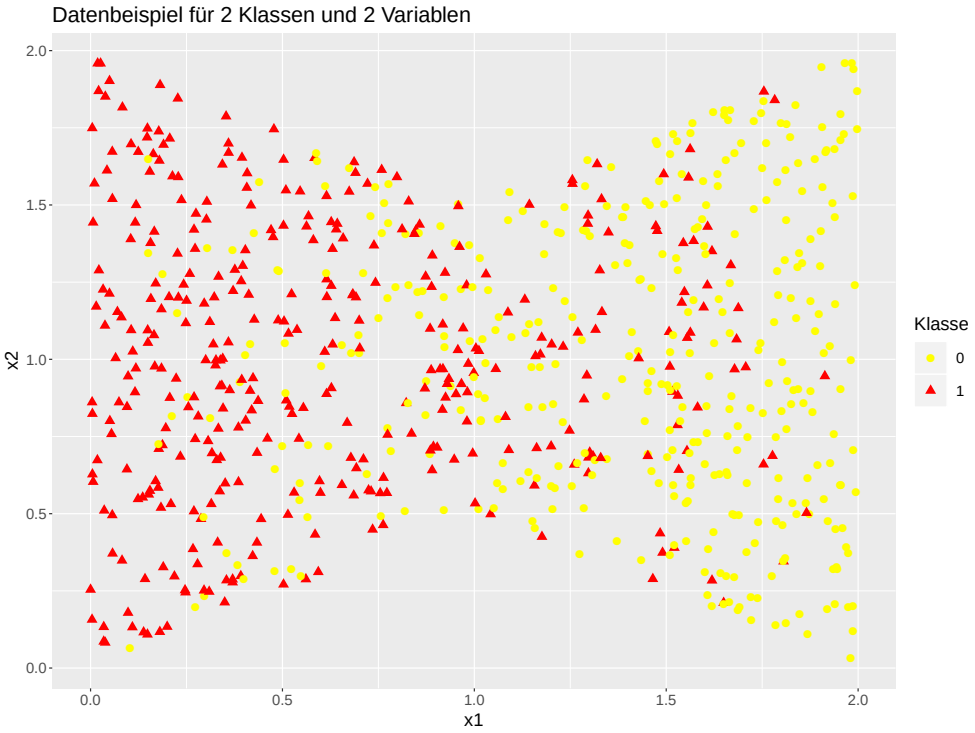


Abbildung 15: Datenbeispiel mit 2 Klassen, 2 Variablen und eindeutigen marginalen Effekten

Um die Interpretation eines Partial Dependence Plots zu erklären, wurde ein Datenbeispiel (Abbildung 15) generiert, das eindeutige marginale Effekte aufweist. Die Wahrscheinlichkeit für Klasse 1 sinkt mit steigendem Wert von **x1**, während **x2** so generiert wurde, dass es unabhängig von $x1$ ist und keinen Einfluss auf die Klasse hat.

Für dieses Datenbeispiel wurden ein multinomiales logistisches Regressionsmodell und ein *Random Forest* trainiert und der Partial Dependence Plot an den Dezilen ausgewertet und veranschaulicht. Der Partial Dependence Plot kann auch an feineren Quantilen ausgewertet werden, bis hin zu allen Werten im Wertebereich der Variable oder sogar künstlich generierten Werten. In Abbildung 16 sind 2 Plots abgebildet, die jeweils aus 2 Sub-Plots bestehen.

Der linke Plot bildet die *Partial Dependence* Kurven für das multinomiale logistische Regressionsmodell ab, wobei der linke Sub-Plot den PDP für **x1** und der rechte Sub-Plot den PDP für **x2** darstellt.

Der rechte Plot bildet die *Partial Dependence* Kurven für den *Random Forest* ab, wobei auch hier der linke Sub-Plot den PDP für **x1** und der rechte Sub-Plot den PDP für **x2** darstellt.

Die gelbe Kurve bildet den Verlauf des marginalen Effekts der jeweiligen Variable auf die Wahrscheinlichkeit für Klasse *0* ab, während die rote Kurve den Verlauf des marginalen Effekts der jeweiligen Variable auf die Wahrscheinlichkeit für Klasse *1* abbildet.

Wie in Abbildung 16 zu sehen ist, erfasst das multinomiale logistische Regressionsmodell den Effekt von **x1** bei Konstanhaltung von **x2** sehr gut. Bei zunehmendem Wert von **x1** erhöht sich die Wahrscheinlichkeit auf Klasse *0* und dementsprechend verringert sich die

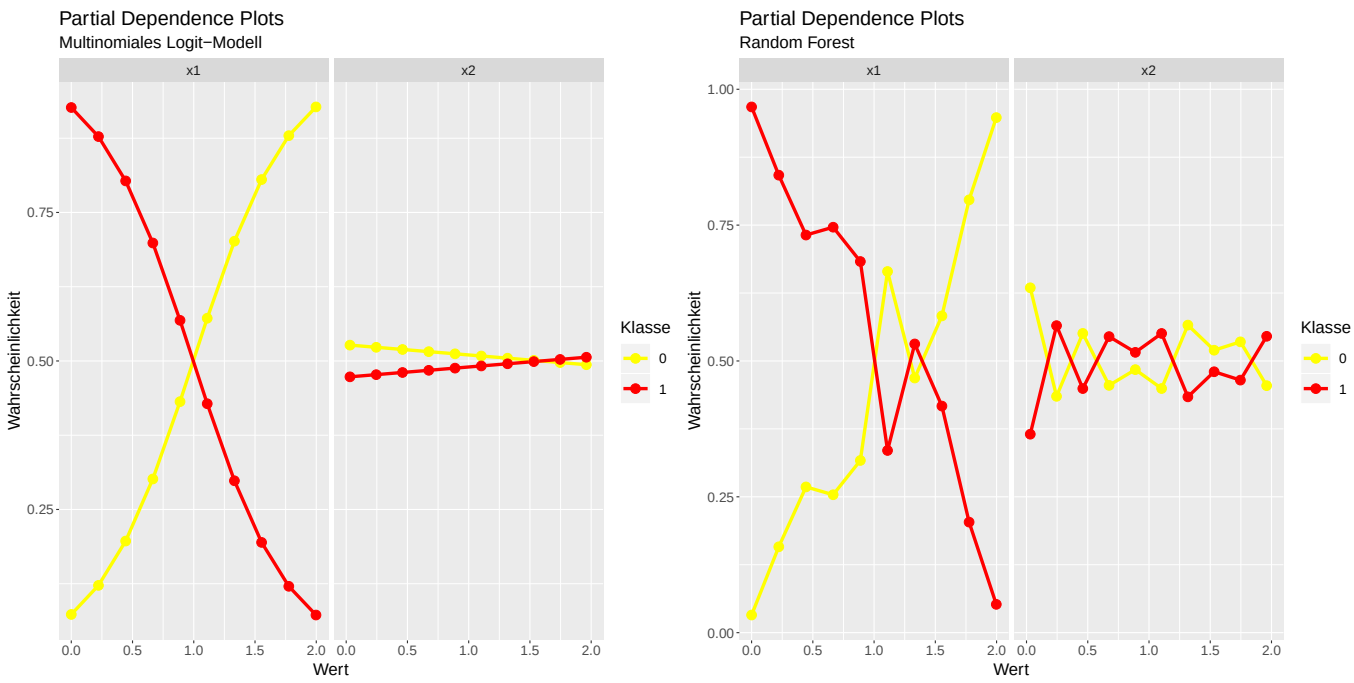


Abbildung 16: Partial Dependence Plots durch multinomiale logistische Regression und *Random Forest*

Wahrscheinlichkeit für Klasse *1*. Der nicht-vorhandene Effekt von **x2** bei Konstanthaltung von **x1** auf die Klasse ist hier zu erahnen, da dieser über den Wertebereich kaum eine Veränderung aufzeigt.

Der *Random Forest* erfasst die zugrunde liegende Logik von **x1** nicht ganz so gut, wie das multinomiale logistische Regressionsmodell. Im mittleren Wertebereich wird ein Knick abgebildet, der im datengenerierenden Prozess nicht vorhanden war. Der steigende Trend der Wahrscheinlichkeit für Klasse *0* bei steigendem **x1** und Konstanthaltung von **x2** kann in dieser Kurve jedoch trotzdem erkannt werden. Der nicht-vorhandene Effekt von **x2** auf die Klasse ist auch hier gut zu erkennen. Es wird ein sehr schwacher Einfluss abgebildet, der beiden Klassen bei Konstanthaltung von **x1** auf dem gesamten Wertebereich von **x2** eine um *0.5* schwankende konstante Wahrscheinlichkeit zuweist.

4.2.2.3 Individual Conditional Expectation Plots

Eine weitere Visualisierungstechnik aus dem Bereich des interpretierbaren Machine Learnings sind die Individual Conditional Expectation Plots (kurz ICE-Plots). Genau wie für den Partial Dependence Plot werden Punkte für jede Variable definiert, an denen für jede Beobachtung eine Prädiktion bei Gleichhalten der anderen Variablen bestimmt wird. Diese werden jedoch an ihren Auswertungspunkten nicht gemittelt, sondern individuell betrachtet. Somit wird ein individueller Verlauf für jede Beobachtung über den Wertebereich einer Variable erzeugt, der zwischen den einzelnen Beobachtungen auf Gemeinsamkeiten und Unterschiede überprüft werden kann.

Der ICE-Plot für eine bestimmte Variable ist also pro Beobachtung i nichts anderes als

$$\hat{f}^{(i)}(x_l) = \mathbb{E}(f(x_l, x_{\setminus l}^{(i)})) \tag{6}$$

ausgewertet an allen relevanten Punkten von x_l (Goldstein et al. 2013). Diese Funktion für jede Beobachtung $i = 1, ..., N$ über den gesamten Wertebereich von x_l ausgewertet, ergibt den ICE-Plot für $\hat{f}(x_l)$. Da für diese Arbeit eine Klassifikation durchgeführt wird, kann für die Wahrscheinlichkeitsprädiktion für jede Klasse ein eigener ICE-Plot erzeugt werden.

Der ICE-Plot wird für eine bestimmte metrische Variable x_l wie folgt erzeugt:

1. Definiere Punkte $q = q_1, ...q_r$ innerhalb des Wertebereichs der Variable x_l, an denen der ICE-Plot berechnet werden soll
2. Berechne für jede Beobachtung die Prädiktion, mit $x_m = \begin{cases} x_m, & \text{wenn } m \neq l, \\ q_p, & \text{wenn } m = l \end{cases}$
3. Wiederhole Schritt 2. für alle $p = 1, ..., r$
4. Plotte die Prädiktion für jede Beobachtung über den Wertebereich der Variable X_l

Da die ICE-Plot Visualisierung einen gemeinsamen Verlauf der Prädiktion über den Wertebereich von X_l darstellen soll, wird auch für diese Visualisierung die Berechnung der Prädiktion an den empirischen Perzentilen der betrachteten Variable x_l ausgewertet (also am 0%-Perzentil, am 1%-Perzentil, ...).

Im Gegensatz zum PDP, der pro Variable für jede Klasse eine gemittelte Kurve angibt, existiert im ICE-Plot für **jede Beobachtung** und **jede Klasse** eine Kurve, was in einem einzelnen Plot zu unerkennbaren Effekten führen würde. Daher wird für jede Klasse ein eigener ICE-Plot erstellt. Daraus resultieren insgesamt 99 verschiedene ICE-Plots (für jede der 9 Klassen und für jede der 11 Variablen).

Der ICE-Plot selbst kann richtungsweisend für Zusammenhänge zwischen den betrachteten Variablen und den betrachteten Klassenwahrscheinlichkeit sein. Möglicherweise nimmt die Wahrscheinlichkeit einer bestimmten Klasse anzugehören über den Wertebereich einer Variable für alle Beobachtungen konstant zu oder ab. In diesem Fall kann von einem monoton steigenden oder fallenden Effekt der Variable auf die Klassenwahrscheinlichkeit gesprochen werden. Häufig passiert es jedoch, dass für manche Beobachtungen die Klassenwahrscheinlichkeit steigt, während sie für andere Beobachtungen fällt. In diesen Fällen kann keine klare Struktur zwischen dem modellierten Zusammenhang zwischen der Variable und der Klasse erkannt werden.

Die in dieser Arbeit weiter behandelten Daten weisen alle verschiedene Verteilungen auf (siehe Kapitel 3.2.1). Weist eine Variable beispielsweise eine bimodale Verteilung mit einer großen Lücke auf, so würde der zugehörige ICE-Plot in diesem Bereich "springen" und einen Verlauf suggerieren, der nicht existiert. Um diese Lücke aufzufangen besteht die Möglichkeit die x-Achse des ICE-Plots nicht aufgrund des Wertebereichs der betrachteten Variable zu

skalieren, sondern auf die empirischen Quantile. Dies entzerrt auch unter anderem eine sehr dichte Datenstelle, in der ein großer Effekt zu sehen ist, welcher jedoch bei quantilsweiser Betrachtung entzerrt betrachtet werden kann.

Da die Kovariablen konstant gehalten werden und den echten Variablen entsprechen, beginnen die ICE-Plots am Minimum des Wertebereichs der betrachteten Variable auf verschiedenen Niveaus (Höhe der Prädiktion bei Transformation von $x_l = min(x_l)$). Um den gemeinsamen Verlauf und nicht die aktuelle Höhe zu betrachten, kann der ICE Plot zentriert werden. Dies bedeutet, dass jede Positionsvorhersage am unteren Rand des Wertebereichs von x_l über den gesamten Verlauf der Kurve subtrahiert wird und somit alle Kurven auf dem Niveau *0* starten. Ein gemeinsames Wachstum oder eine gemeinsame Verringerung der Kurven kann somit einfacher erkannt werden (Goldstein et al. 2013).

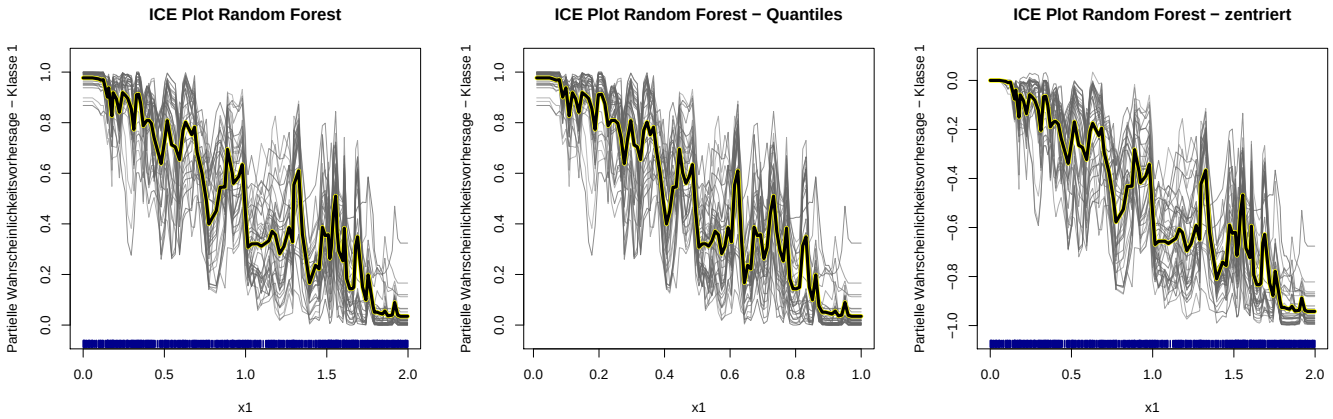


Abbildung 17: ICE Plots für **x1** durch *Random Forest*

In Abbildung 17 sind die drei beschriebenen Typen an ICE-Plots dargestellt. Diese ICE-Plots wurden für das Datenbeispiel aus Abbildung 15 erstellt. Zu sehen ist hier der Zusammenhang von der Variable **x1** und der Wahrscheinlichkeit auf Klasse *1*, so wie ihn der *Random Forest* modelliert. Aus Gründen der Übersichtlichkeit wurde hier zufällig ein Subsample der ursprünglichen Daten generiert.

Der linke Plot beschreibt den Zusammenhang zwischen **x1** und der Wahrscheinlichkeit für Klasse *1* indem für jede Beobachtung an jedem Wert von **x1** eine künstliche Beobachtung erstellt wird (bei Konstanthaltung der übrigen Variablen; in diesem Beispiel **x2**). Alle künstlichen Beobachtungen, die so für eine ursüprünglich zugrunde liegende Beobachtung erstellt wurden, bilden einen grauen Verlauf in dieser Grafik ab. Der gelb-umrandete schwarze Verlauf bildet den für diese Beobachtungen kreierten Partial Dependence Plot ab. Dieser wurde genau wie die ICE-Kurven an den Perzentilen der Verteilung von **x1** ausgewertet. Die Verteilung von **x1** ist durch die blauen Makierungen an der x-Achse zu sehen. Gäbe es zum Beispiel einen Wertebereich von **x1** in dem keine Beobachtungen erfasst wurden, so wäre dies durch die blauen Makierungen gekennzeichnet.

Im Allgemeinen ist eine abfallende Wahrscheinlichkeit für Klasse *1* bei steigendem **x1** erkennbar. Dieser schwankt im Gegensatz zum Partial Dependence Plot für den *Random*

Forest aus Abbildung 16 sehr stark, da er an mehr Punkten ausgewertet wurde und somit kleine Schwankungen eher erfasst werden.

Der mittlere Plot bildet für dieses Datenbeispiel in etwa dengleichen Verlauf ab wie der linke Plot. Der Unterschied ist, dass die x-Achse auf die empirischen Quantile von **x1** skaliert wurde, das heißt, dass bspw. das 20%-Quantil an der Stelle *0.2* liegt und der Median an der Stelle *0.5*. Eine mögliche Sprungstelle würde dadurch überbrückt werden. Wie im späteren Verlauf dieser Arbeit noch gezeigt wird, können Sprungstellen die Interpretation dieser ICE-Kurven schwieriger gestalten, was durch diese quantilsweise Betrachtung behoben werden würde. Der Bereich, in dem viele Datenpunkte liegen (also im oberen und unteren Teil des Wertebereichs von **x1**), wird durch diese Betrachtung etwas entzerrt, während der Bereich, in dem wenige Datenpunkte liegen (im mittleren Teil des Wertebereichs) etwas gestaucht wird.

Der rechte Plot bildet eine zentrierte Art der ICE-Plots ab. Wie im linken Plot zu erkennen ist, beginnen die verschiedenen Verläufe alle auf unterschiedlichen Niveaus (der Auswertungspunkt von **x1** ist zwar dergleiche für alle Beobachtungen, **x2** jedoch nicht!). Diese Niveaus werden im rechten Plot alle zusammengeführt und beginnen damit am gleichen Punkt. Durch die Zentrierung werden nicht nur die absoluten Vorhersagewerte an den verschiedenen Auswertungspunkten vergleichbar gemacht, sondern auch die Verläufe der Vorhersagewerte über den gesamten Wertebereich von **x1**.

An dieser Stelle soll eine Problematik dieser Art Plots erwähnt werden. Werden die Plots an zu wenigen Stellen ausgewertet, so könnten Schwankungen eventuell nicht erkannt werden, da vor und nach einer möglicherweise relevanten Schwankung der Plot ausgewertet wird. Wird aber an sehr vielen Stellen ausgewertet, so werden richtigerweise alle Schwankungen, die durch das Modell prädiktiert werden, abgebildet, jedoch könnte dadurch eine zugrundeliegende Logik, die einen Anwender interessiert, nicht erkannt werden. Vor allem durch Sprungstellen im Wertebereich der betrachteten Variable könnten dadurch Effekte erkannt werden, die dem zugrundeliegenden Zusammenhang zwischen der Variable und der Zielgröße überhaupt nicht entsprechen.

4.2.2.4 Accumulated Local Effect Plots

Eine Schwachstelle der Partial Dependence Plots und der ICE-Plots ist, dass durch die künstliche Datenmanipulation Beobachtungen kreiert werden können, die in der Realität unmögliche Beobachtungen sind. Dies kann für die Leistungsdaten bei Fußballspielern zum Beispiel bedeuten, dass ein Spieler mehr Tore geschossen hat, als er Torschüsse abgegeben hat. Es können auch höchstunwahrscheinliche Beobachtungen auftauchen, zum Beispiel ein Spieler, der pro Spiel nur 10 Pässe spielt, aber davon 9 Fehlpässe sind.

Die Idee ist es die auf x_l bedingten Wahrscheinlichkeiten für die möglichen Kovariablenkombinationen von $x_{\setminus l}$ zu nutzen, damit unwahrscheinliche Kovariablenkombinationen an Gewicht verlieren (Apley 2016).

$$\begin{aligned}\hat{f}_{l,ALE}(x_l) &\equiv \int_{z_{0,l}}^{x_l} \mathbb{E}[f^l(x_l, x_{\setminus l})|x_l = z_l]dz_l - constant \\ &= \int_{z_{0,l}}^{x_l} \int \mathbb{P}_{\setminus l|l}(x_{\setminus l}|z_l) f^l(z_l, z_{\setminus l}) dx_{\setminus l} dz_l - constant,\end{aligned} \tag{7}$$

wobei $f^l(z_l, z_{\setminus l}) = \frac{\partial f(x_l, x_{\setminus l})}{\partial x_l}$ die partielle Ableitung darstellt. Da die partielle Ableitung in der Regel unbekannt ist, wird diese durch Einteilen des Wertebereichs von x_l in Intervalle mit den Intervallgrenzen $\{z_{0,l}, ..., z_{r,l}\}$ diskretisiert. Für diese Intervalle werden finite Differenzen gebildet, wodurch die partielle Ableitung approximiert wird (Scholbeck et al. 2019). Der bedingte Erwartungswert wird anschließend intervallweise durch Monte Carlo Integration geschätzt, wodurch das innere Integral für das jeweilige Intervall, in dem z_l liegt, bestimmt wird. Dadurch entsteht im inneren Integral eine Art Treppenfunktion. Da über all diese intervallmäßigen Erwartungswerte integriert wird, ist die Breite der Intervalle irrelevant. Üblicherweise sollten die Intervalle jedoch entweder äquidistant oder anhand der Quantile der Daten gebildet werden (Apley 2016). Während die äquidistanten Intervalle den Wertebereich in gleich große Bereiche einteilen, hat die quantilsweise Einteilung den Vorteil, dass in Bereichen, in denen viele Beobachtungen vorkommen, feinere Einteilungen gemacht werden. Dadurch können relevante Effekte in kleineren Bereichen genauer erfasst werden.

Die Funktion $\hat{f}_{l,ALE}(X_l)$ ergibt die Kurve für den ALE-Plot für X_l, dessen Höhe durch die abgezogene Konstante bestimmt wird. In der Regel wird die Konstante so gewählt, dass die Kurve "zentriert" ist, was bedeutet, dass die y-Achse der Abweichung vom durchschnittlichen Effekt einer Variable auf die Zielgröße im Modell entspricht.

Der ALE-Plot wird vereinfacht für eine Variable X_l wie folgt erzeugt:

1. Definiere Intervallgrenzen $q = q_0, ...q_r$ innerhalb des Wertebereichs der Variable X_l, zwischen denen der ALE-Plot berechnet werden soll
2. Bestimme für Intervall i einen Teildatensatz S_i mit allen Beobachtungen, für die $q_i \leq X_l < q_{i+1}$ gilt
3. Bestimme für alle Beobachtungen in Teildatensatz S_i die Prädiktion an der unteren und oberen Intervallgrenze i (also $X_l = q_i$ und $X_l = q_{i+1}$) bei Konstanthalten der Kovariablen
4. Bestimme für jede Beobachtung eine lineare Steigung für das Intervall i durch Interpolieren der beiden Prädiktionen an den Intervallgrenzen
5. Berechne eine mittlere Steigung für den Teildatensatz S_i, um eine durchschnittliche Steigung für Intervall i zu erhalten
6. Wiederhole Schritt 2. bis 5. für $i = 0, ..., r - 1$

7. Kumuliere die ermittelten Steigungen für jedes Intervall um eine stetige Kurve zu erhalten

Um die Intervalle klein genug zu machen, damit die Kovariablen ihren Effekt auf den Prädiktionsunterschied der Intervallgrenzen verlieren, aber trotzdem genug Beobachtungen in jedem Intervall zu behalten um stabile Steigungen für die Intervalle zu erhalten, wurde entschieden insgesamt 15 Intervalle mit etwa 100 Beobachtungen pro Intervall für die ALE-Plots zu bilden.

In Abbildung 18 ist für die Datensituation aus Abbildung 15 der Accumulated Local Effects-Plot für **x1** für die Wahrscheinlichkeitsvorhersage für Klasse *1* im *Random Forest* abgebildet. Auf der x-Achse ist der Wertebereich für **x1** zu erkennen. Auf der y-Achse ist die Abweichung der mittleren Prädiktion der Wahrscheinlichkeitsvorhersage für Klasse *1* abgebildet. Wie zu erkennen ist, sinkt die Wahrscheinlichkeitsvorhersage für Klasse *1* mit steigendem **x1**. Auch hier zeigt der Abwärtstrend leichte Schwankungen und ähnelt dem Partial Dependence Plot für **x1**.

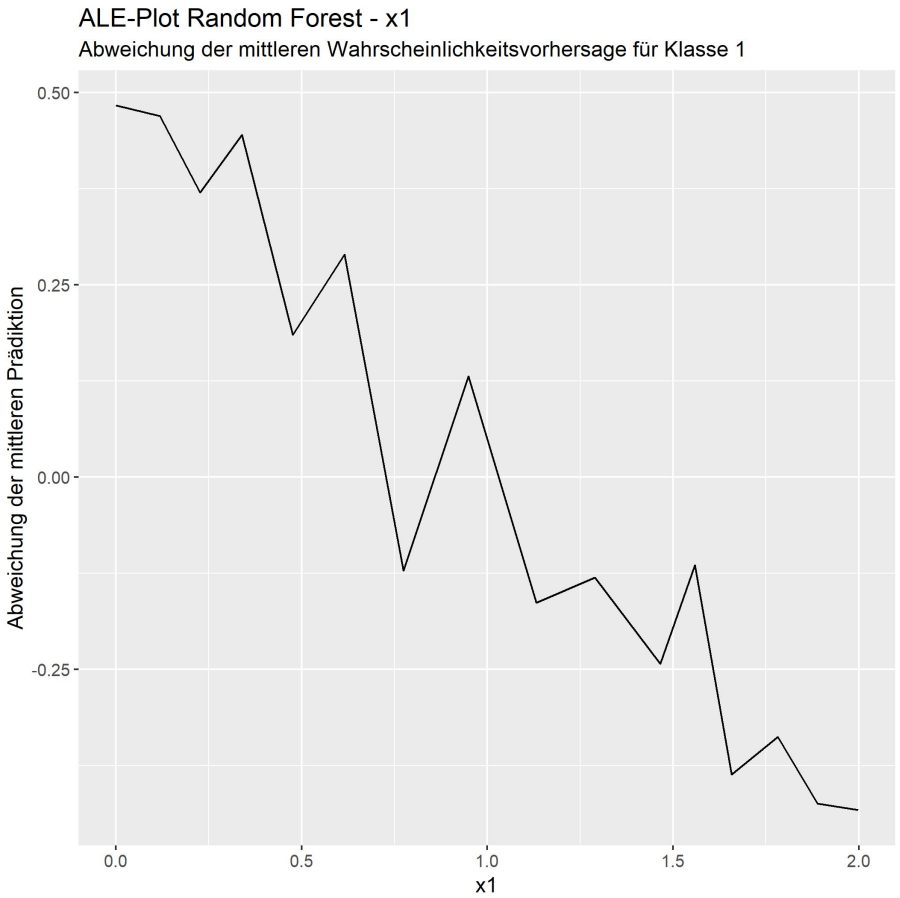


Abbildung 18: ALE-Plot für **x1** durch *Random Forest*

Der Hauptunterschied zwischen diesen beiden Plots ist jedoch die Interpretation der y-Achse. Während beim Partial Dependence Plot die y-Achse als “mittlere Wahrscheinlichkeitsvorher-

sage für Wert x" interpretiert werden kann, gibt der Accumulated Local Effects-Plot die Abweichung der mittleren Wahrscheinlichkeitsvorhersage an. Mit anderen Worten bedeutet ein Wert auf der y-Achse von *0.25*, dass die Wahrscheinlichkeitsvorhersage für Klasse *1* bei diesem Wert x um *0.25* höher ist als die durchschnittliche Wahrscheinlichkeitsvorhersage für Klasse *1*.

Eine wirklich große Diskrepanz kann für Variablen mit starken Abhängigkeiten zu anderen Variablen entstehen. Der Partial Dependence Plot würde einen Punkt x^* mit allen Beobachtungen auswerten, egal wie nah oder weit sie von diesem Punkt entfernt liegen. Dass die Wahrscheinlichkeitsvorhersage dafür durch die Kovariablen stark beeinflusst wird, ist dementsprechend für starke Abhängigkeiten zwischen den Variablen sehr wahrscheinlich. Wenn jedoch nur Beobachtungen betrachtet werden, die Nahe an x^* liegen, so wird dieser Abhängigkeitseffekt der Kovariablen reduziert.

4.2.2.5 Erarbeitung der Topologie der Modelle

Eine interessante Fragestellung, abgesehen von den Beziehungen zwischen den Leistungsdaten und den Positionen, wäre es die Topologie der Daten im mehrdimensionalen Raum näher zu untersuchen. Dafür wird hier eine Methode beschrieben, mit der untersucht wird, welche Klassen in Bezug auf einzelne Leistungsdaten im mehrdimensionalen Raum nebeneinander liegen (also konkret, welche Positionen bezüglich eines Leistungsdatums benachbart sind).

Um die angewendete Methode für diese Untersuchung genauer zu erklären, wird diese im Folgenden anhand von Beispielen dargestellt und erklärt, wie die Resultate zu interpretieren sind.

Angenommen es existiert eine in Abbildung 19 dargestellte Datenlage. In diesem Beispiel existieren 4 verschiedene Klassen und 2 Variablen, durch welche die Klassen perfekt getrennt werden.

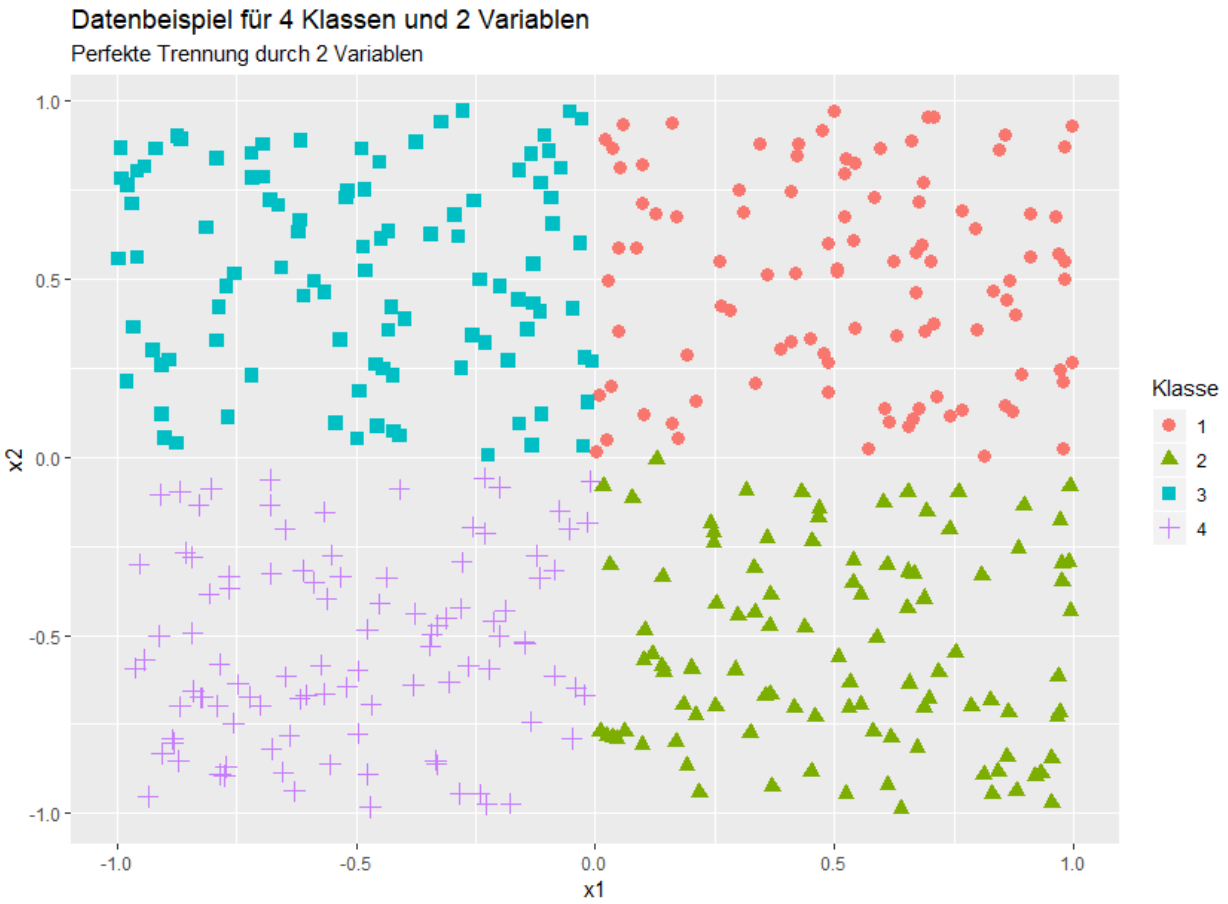


Abbildung 19: Datenbeispiel für perfekt getrennte Klassen

Zwischen jeweils zwei Klassen kann bezüglich jeder Variable ein Bezug formuliert werden, was zu insgesamt $4 * 3 * 2 = 24$ Kombinationen führt (dabei ist die Beziehung "A ist oberhalb von B" und "B ist unterhalb von A" doppelt gezählt). All diese Kombinationen können von unserem Hirn gleichzeitig erfasst und verarbeitet werden, wodurch die Beziehung zwischen den Klassen durch diese Visualisierung schnell erfasst werden kann. Für ein solches Datenbeispiel genügt also die Betrachtung einer 2-dimensionalen Grafik wie dieser, um zu ermitteln, welche Klassen in welcher Beziehung nebeneinander liegen. In der multinomialen logistischen Regression werden Regressionskoeffizienten ermittelt, welche diese Beziehung ausdrücken. Ein **positiver** Koeffizient bedeutet, dass Punkte in der Klasse bezüglich der Punkte in der Referenzkategorie einen **höheren** Wert der betrachteten Variable aufweist, während ein **negativer** Koeffizient ausdrückt, dass Punkte in der Klasse bezüglich der Punkte in der Referenzkategorie einen **niedrigeren** Wert der betrachteten Variable aufweisen. Was jedoch nicht direkt durch die Regressionskoeffizienten ermittelt werden kann, ist die Tatsache, ob zwischen zwei Klassen eine weitere Klasse liegt, oder ob zwei Klassen aufgrund der Kovariablen gar nicht bezüglich einer betrachteten Variable nebeneinander liegen (in Abbildung 19 die diagonal benachbarten Klassen).

Die eigentliche Problematik beginnt jedoch, wenn mehr Klassen auftreten und nicht eindeutig durch 2 Variablen zu trennen sind.

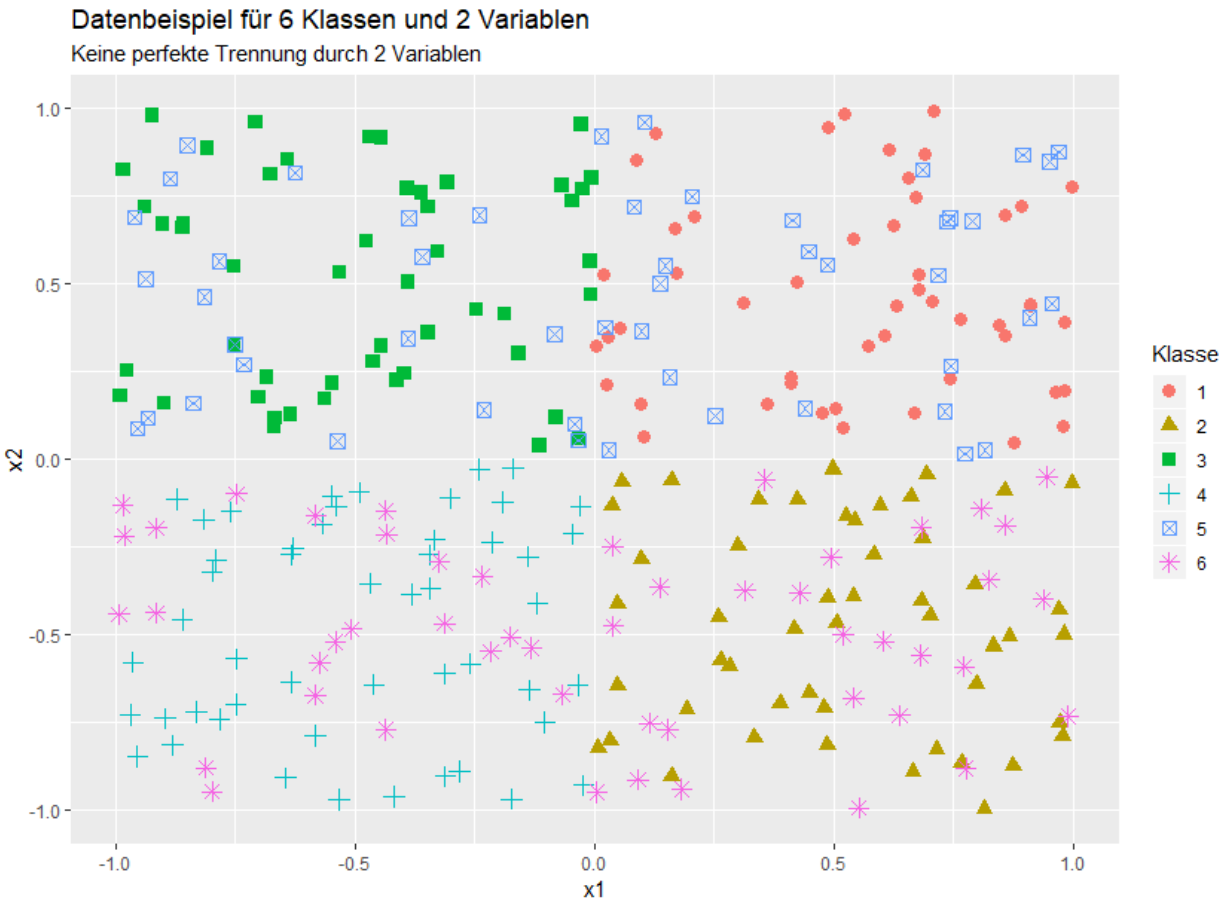


Abbildung 20: Datenbeispiel für nicht perfekt getrennte Klassen

Angenommen es existiert eine in Abbildung 20 dargestellte Datenlage. Wie hier zu erkennen ist, werden die Klassen *1*, *2*, *3* und *4* weiterhin durch die Variablen **x1** und **x2** perfekt voneinander getrennt, existieren zwei weitere Klassen *5* und *6*, welche zwar durch Variable **x2** perfekt voneinander getrennt werden, jedoch zwischen den anderen Klassen liegen. Eine exakte Definition des Nachbarschaftsverhältnisses ist hier erst nach näherer Betrachtung genau anzugeben, da die beiden Variablen nicht reichen die Daten perfekt zu trennen.

Nun wird zusätzlich angenommen, dass eine Variable **x3** existiert, welche die Klassen *5* und *6* perfekt von den Klassen *1*, *2*, *3* und *4* trennt. Eine 3-dimensionale Grafik, in der die Klassen *5* und *6* hinter, bzw. vor den anderen Klassen liegen, wodurch die Daten wieder perfekt getrennt sind, ist mit leichtem Aufwand vorstellbar und es können direkt Nachbarschaftsverhältnisse erfasst werden. Wenn sich die Anzahl der Klassen und die Höhe der Dimensionalität jedoch weiter erhöhen, wird das ganze unvorstellbar (und nur sehr schwer darstellbar).

Ein weiteres Problem ist, dass Klassen oft nicht perfekt trennbar sind und nicht als “einzelne Cluster” im Raum liegen, sondern “punktweise verteilt” sind. Es soll nun eine Methode gefunden werden, die das Nachbarschaftsverhältnis zwischen den verschiedenen Klassen bezüglich der einzelnen Variablen ermittelt.

Das Ziel einer Klassifikation ist es anhand der vorliegenden Daten den Raum in Bereiche

einzuteilen, in denen Beobachtungen Wahrscheinlichkeiten zugewiesen werden können, mit welchen sie den verschiedenen Klassen angehören. Einer neuen Beobachtung kann anhand seiner Lage im Raum eine Klasse zugeordnet werden, die aufgrund der ursprünglichen Daten am wahrscheinlichsten für diese Position ist. Ein sehr gutes Modell teilt den Raum also in "perfekte Bereiche" ein, in denen die verschiedenen Klassen liegen. Es ist somit möglich eine sehr gute Modellierung zu nutzen, um die Lage der einzelnen Klassen im Raum und die Nachbarschaftsverhältnisse zwischen den verschiedenen Klassen zu ermitteln.

Die Methode, die die Nachbarschaftsverhältnisse zwischen den einzelnen Klassen bezüglich einzelner Variablen gegeben der Kovariablen beschreiben soll, geht wie folgt vor:

1. Schätze ein Modell $\hat{f}(\cdot)$
2. Nutze echte (oder für spezielle Betrachtungen simulierte) Daten und merke ihre Prädiktionen $\hat{y}$ durch das Modell
3. Erhöhe/Verringere eine bestimmte Variable geringfügig, während alle anderen Kovariablen gleichgehalten werden
4. Ermittle anhand der manipulierten Daten die neuen Prädiktionen $\hat{y}^*$
5. Betrachte die Wechsel zwischen den einzelnen Klassenvorhersagen

Dieses Vorgehen soll anhand des ersten Datenbeispiels aus Abbildung 19 erläutert werden. In Schritt 1. wird ein Modell (z.B. ein Klassifikationsbaum) anhand der Daten geschätzt und teilt für diese Datensitation den Raum wie in Abbildung 21 ein.

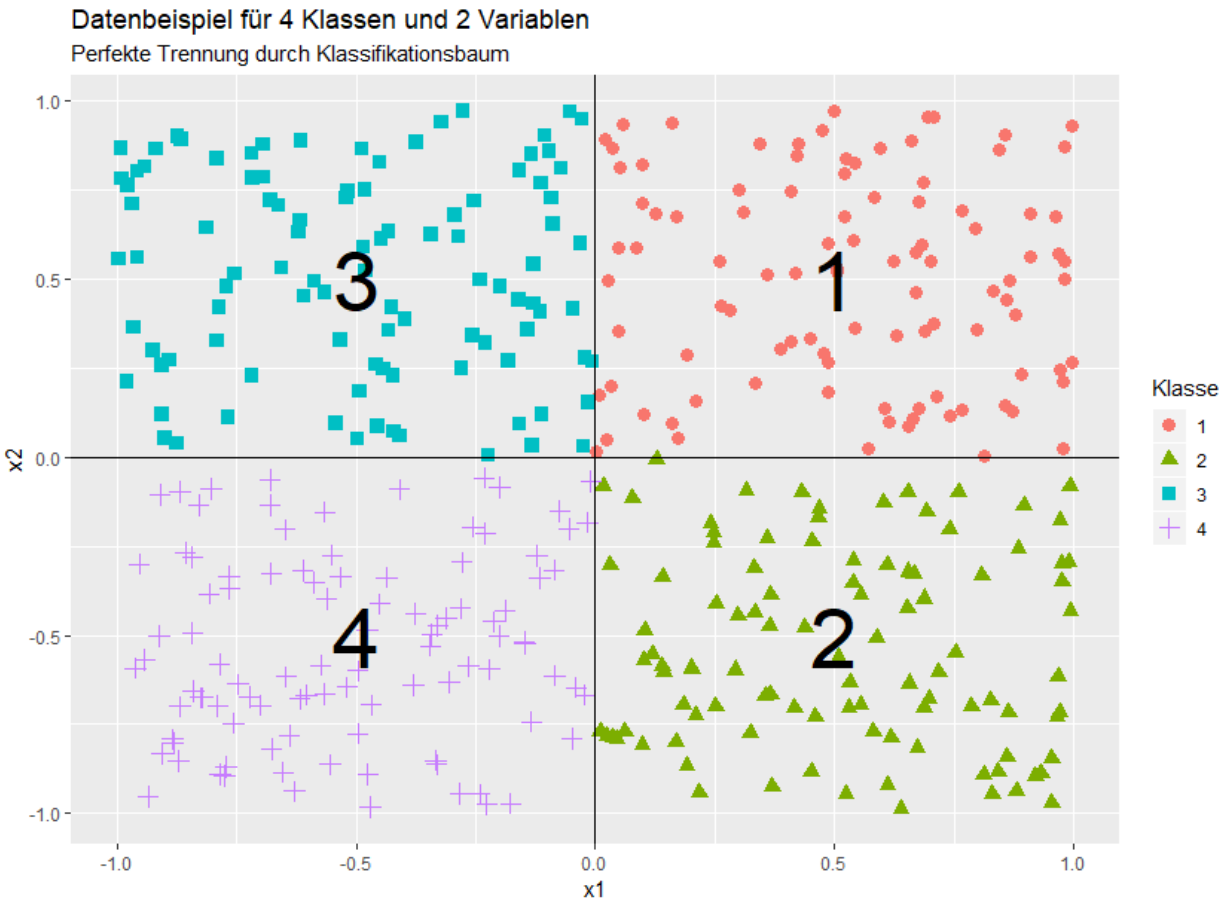


Abbildung 21: Perfekte Trennung durch Klassifikationsbaum

Als erstes soll das Nachbarschaftsverhältnis bezüglich der Variable **x1** betrachtet werden. Um dies zu ermitteln werden die originalen Daten verwendet (ein Trainings- und Testdatensplit ist für diese Methode nicht notwendig) und ihnen wird eine Prädiktion zugewiesen. Für diesen Spezialfall einer perfekten Trennung werden allen Daten als Prädiktion ihre originalen und richtigen Klassen zugewiesen. Nun wird, wie in Schritt 3. beschrieben, der **x1**-Wert jeder Beobachtung leicht erhöht (in diesem Beispiel um 0.1) und wie in Schritt 4. beschrieben die Prädiktion $\hat{y}^*$ für jede Beobachtung ermittelt (siehe Abbildung 22). Wichtig ist, dass durch die Datenmanipulation das Risiko auf unmögliche Datenkonstellationen gering gehalten werden soll, weshalb nur kleine Datenmanipulationen durchgeführt werden sollen.

Die Form der Beobachtung gibt an, welche Prädiktion die Beobachtung vor der Datenmanipulation hatte, und die Farbe, welche Prädiktion eine Beobachtung nach der Datenmanipulation hatte. Wie zu erkennen ist, hat die Vorhersage einiger Beobachtungen aus Klasse *3* zu Klasse *1* gewechselt, während einige Beobachtungen aus Klasse *4* zu Klasse *2* gewechselt haben.

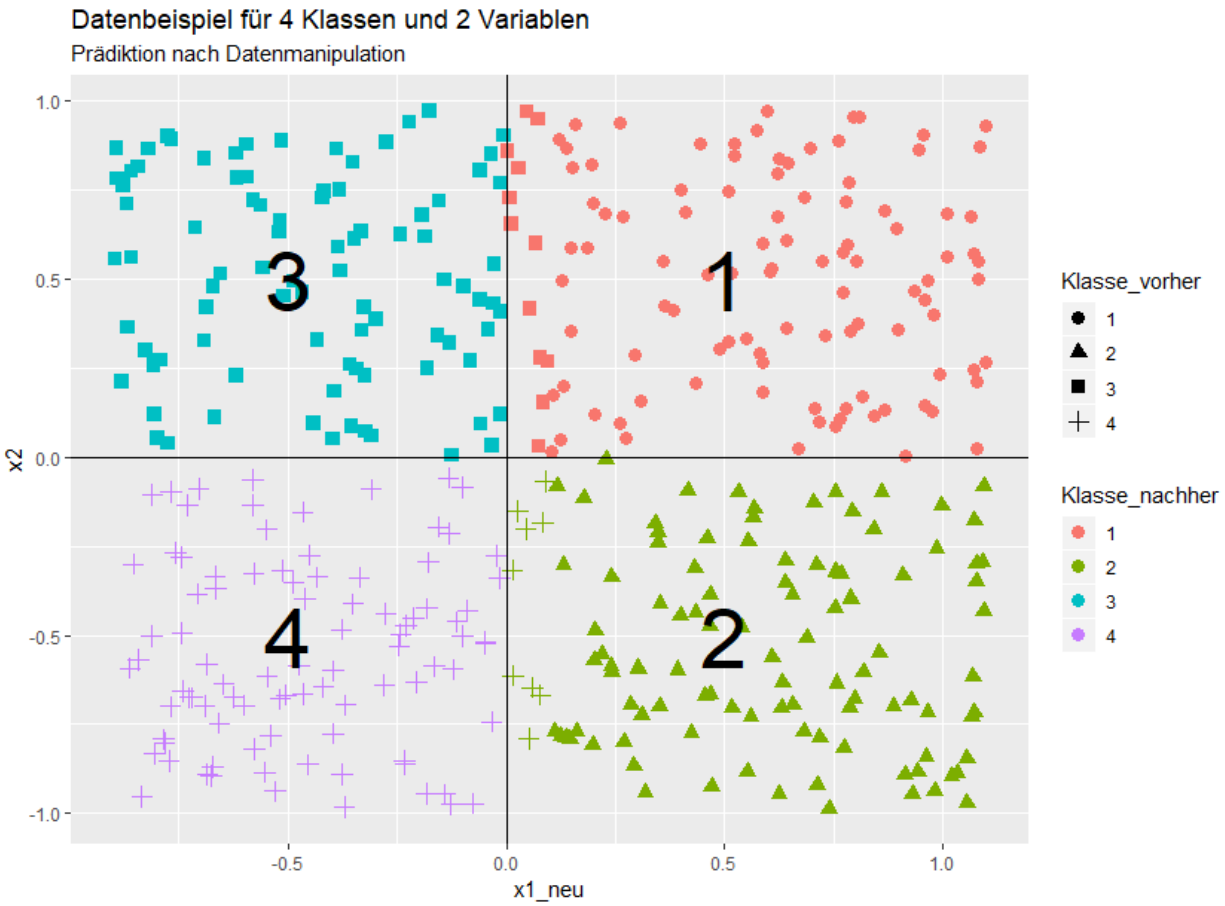


Abbildung 22: Vorhersage nach Datenmanipulation

Die Wechsel der Prädiktionen können anhand einer einfachen 4x4-Matrix erfasst werden (siehe Tabelle 5). Wie an dieser Tabelle abzulesen ist, haben sich nach der Datenmanipulation ein paar Prädiktionswechsel ergeben. Zum Einen haben 12 der Beobachtungen, die vorher in Klasse *3* waren, durch die Datenmanipulation in Klasse *1* gewechselt. Zum Anderen haben 9 der Beobachtungen aus Klasse *4* in Klasse *2* gewechselt. Dies bedeutet, dass bei Konstanthalten der Kovariablen (hier nur **x2**) ein Gebiet mit Klasse *1*, das einen höheren **x1**-Wert aufweist, neben einem Gebiet mit Klasse *3* liegt, und dass ein Gebiet mit Klasse *2*, das einen höheren **x1**-Wert aufweist, neben einem Gebiet mit Klasse *4* liegt. Genau dieses Nachbarschaftsverhältnis ist auch in der Grafik beobachtbar.

Neue Prädiktion in:	1	2	3	4
Original 1	100	0	0	0
Original 2	0	100	0	0
Original 3	12	0	88	0
Original 4	0	9	0	91

Tabelle 5: Prädiktionswechsel der Beobachtungen nach Datenmanipulation

Der Unterschied zum Ermitteln des Nachbarschaftsverhältnisses durch Grafiken ist jedoch, dass diese Methode keinen "dimensionellen Restriktionen" unterliegt und mit beliebig vielen Klassen und beliebig vielen Kovariablen durchgeführt werden kann. Des Weiteren ist der Vorteil gegenüber den Regressionskoeffizienten einer multinomialen logistischen Regression das Verhältnis zwischen allen Klassen gleichzeitig zu ermitteln, während durch die Regressionskoeffizienten nur der Bezug zu einer bestimmten Referenzkategorie bestimmt wird. Darüber hinaus unterliegt diese Methode keiner Restriktion bezüglich der Modellform und kann für alle klassifizierende Modelle angewendet werden.

Ein Nachteil gegenüber der Regressionskoeffizienten einer multinomialen logistischen Regression ist jedoch, dass jede Variable und jede Richtung der Datenmanipulation der Variable einzeln betrachtet werden muss.

Bezüglich der **Interpretation** müssen jedoch einige Dinge beachtet werden. Zum Einen ist es möglich, dass eine Grenze des Modells an einer Stelle liegt, an welcher nur Datenpunkte aus Klasse A, aber keine Datenpunkte aus Klasse B liegen. Dies bedeutet, dass beim Überprüfen einer Richtung (Variable x_m Erhöhen oder Verringern) ein Nachbarschaftsverhältnis festgestellt wird, in die andere Richtung jedoch nicht. Dieses Ergebnis bedeutet **NICHT** bspw. "Ein Gebiet mit Klasse A liegt neben einem Gebiet mit Klasse B mit höherem x_m-Wert, aber kein Gebiet der Klasse B liegt neben einem Gebiet mit Klasse A mit niedrigerem x_m-Wert", sondern "Ein Gebiet mit Klasse A liegt neben einem Gebiet mit Klasse B mit höherem x_m-Wert, aber keine Punkte der Klasse B liegen neben einem Gebiet mit Klasse A mit niedrigerem x_m-Wert". Es könnte zum Beispiel passieren, dass das Modell eine sinnvolle Grenze zieht, dort jedoch eine *unmögliche Datensituation* vorliegt, weshalb dort keine Daten liegen; es könnten allerdings auch einfach keine Beobachtungen dort erhoben worden sein.

Ein weiterer Punkt, der beachtet werden muss, ist das Gesetz der **Transitivität**. Es kann passieren, dass durch punktweise verteilte Gebiete Situationen entstehen, in denen bspw. ein Gebiet der Klasse B an einer Stelle "über" einem Gebiet der Klasse A und an einer anderen Stelle "unter" einem Gebiet der Klasse C liegt, wodurch aber nicht impliziert werden kann, dass ein Gebiet der Klasse A auch "unter" einem Gebiet der Klasse C liegt (vergleiche Abbildung 23). Damit sind die Resultate **nicht** transitiv zu interpretieren.

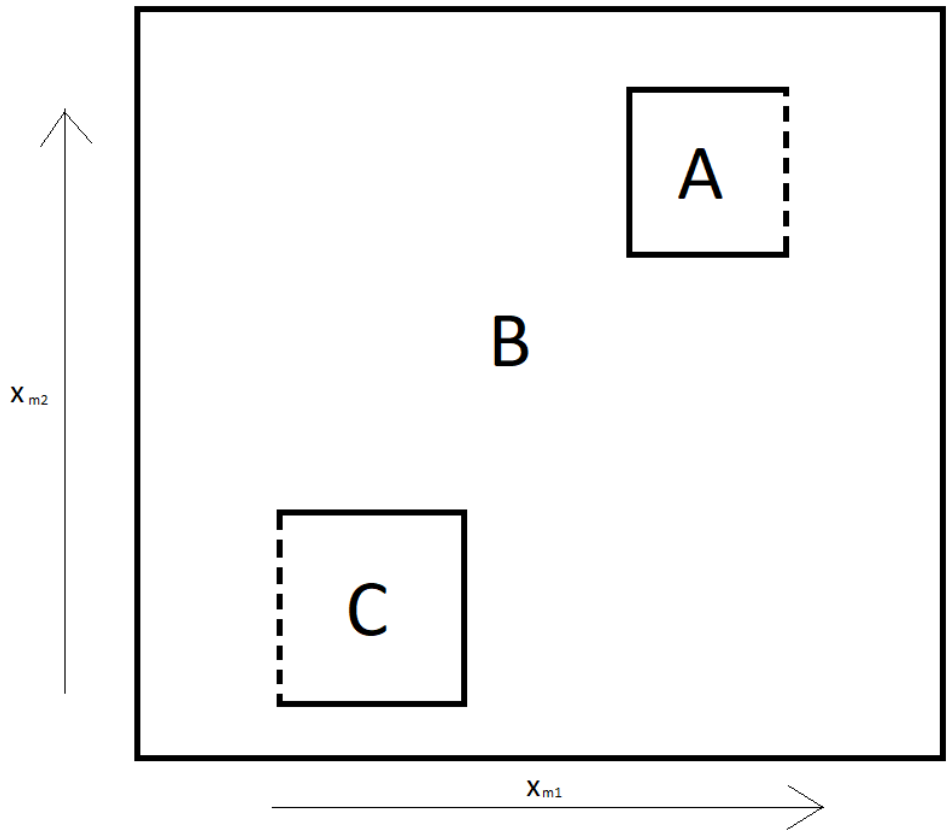


Abbildung 23: Beispiel für Transitivitätsproblem für **m1** $\neq$ **m2**

Alles in allem sind die Ergebnisse also wie folgt zu interpretieren:

1. Wechsel von Klasse A in Klasse B bei Erhöhen von Variable x_m bedeutet, dass Gebiete mit Klasse B existieren, die in Bezug auf x_m oberhalb von Gebieten mit Klasse A liegen (bei Konstanthaltung der anderen Variablen) und umgekehrt existieren Gebiete mit Klasse A, die in Bezug auf x_m unterhalb von Gebieten mit Klasse B liegen!

2. Wechsel von Klasse A in Klasse B bei Verringern von Variable x_m bedeutet, dass Gebiete mit Klasse B existieren, die in Bezug auf x_m unterhalb von Gebieten mit Klasse A liegen (bei Konstanthaltung der anderen Variablen) und umgekehrt existieren Gebiete mit Klasse A, die in Bezug auf x_m oberhalb von Gebieten mit Klasse B liegen!

3. Wenn Punkt 1. und 2. gleichzeitig auftreten, bedeutet es nicht, dass ein Gebiet umschlossen ist, sondern dass an einem Punkt im Raum ein Gebiet mit Klasse A in Bezug auf x_m oberhalb von einem Gebiet mit Klasse B liegt und an einem möglicherweise anderen Punkt ein Gebiet mit Klasse A unterhalb von einem Gebiet mit Klasse B liegt

Die in Tabelle 5 aufgeführte Migrationsmatrix kann in einem Chordgraph visuell dargestellt werden (siehe Abbildung 24). In diesem Graphen ist einerseits zu sehen, wie groß die prädiktierten Klassen vor und nach der Datenmanipulation sind, und andererseits von welcher Klasse in welche andere Klasse die Prädiktion wechselt.

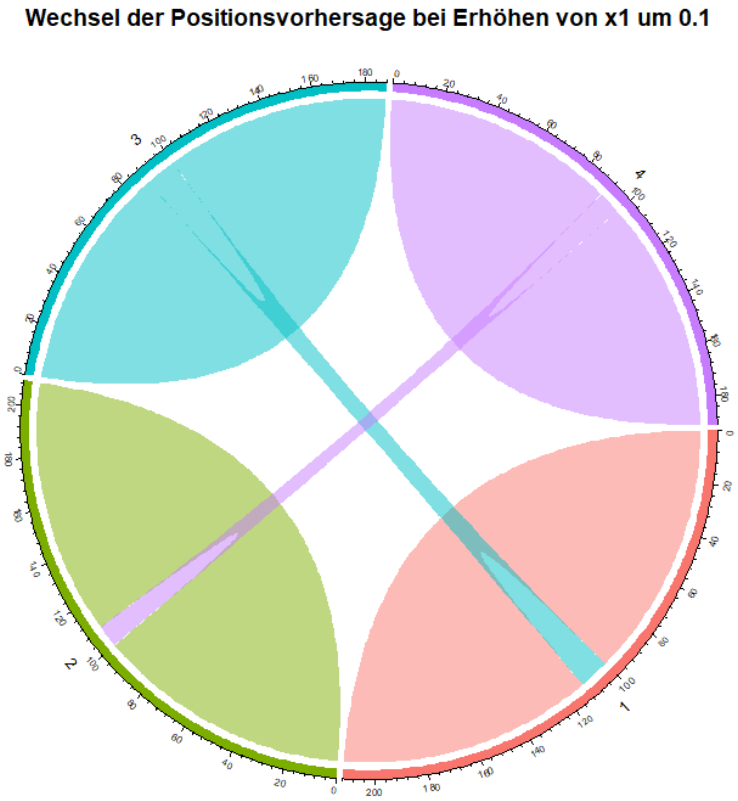


Abbildung 24: Chordgraph als Visualisierung für Migrationsmatrix

Die Zahlensträhle am Rand der verschiedenen Klassen geben die absolute Anzahl an Beobachtungen, die wechseln, bzw. nicht wechseln, an. Jede Klasse wurde so konstruiert, dass sie 100 Beobachtungen enthält. Vor der Datenmanipulation waren genau 100 Beobachtungen durch das Modell richtig prädiktiert. Dies wird an den Zahlenstrählen jeweils von 0 bis 100 angezeigt. In Klasse *3* ist zu erkennen, dass ein Teil der ersten 100 Beobachtungen in Klasse *1* wechselt. Aus Klasse *1* wechselt keine Beobachtung in eine andere Klasse. Folglich wechseln alle 100 Beobachtungen aus Klasse *1* "in sich selbst" und erhalten aus Klasse *3* zusätzliche Beobachtungen.

Durch diese Verbindungen ist also das zu erkennen, was auch in der Migrationsmatrix zu erkennen ist:

- Es existiert ein Gebiet mit Klasse *1*, das bezüglich **x1** oberhalb von einem Gebiet mit Klasse *3* liegt.

- Es existiert ein Gebiet mit Klasse *2*, das bezüglich **x1** oberhalb von einem Gebiet mit Klasse *4* liegt.
- Zwischen den Klassen *1* und *2*, zwischen den Klassen *3* und *4*, zwischen den Klassen *1* und *4* und zwischen den Klassen *2* und *3* können bezüglich **x1** keine benachbarten Gebiete festgestellt werden.

Der letzte Punkt könnte sich bei einem erneuten Überprüfen durch das Verringern von **x1** jedoch ändern (nur nicht für dieses simulierte Datenbeispiel)!

4.3 Modellaufbau in grafischem Kontext

4.3.1 Random Forest

Der Klassifikations-*Random Forest* ist ein Ensemble von verschiedenen Klassifikationsbäumen. Jeder einzelne Baum teilt einen Raum in rechtwinklige Flächen ein, in denen bei einem unbeschnittenen Baum zu 100% eine bestimmte Klasse prognostiziert wird. Um eine neue Beobachtung durch diesen *Random Forest* zu klassifizieren, wird die neue Beobachtung in jeden einzelnen dieser eingeteilten Räume eingesetzt und erhält dadurch eine Klassifikation (Majority-Vote der Bäume oder Wahrscheinlichkeiten für jede Klasse).

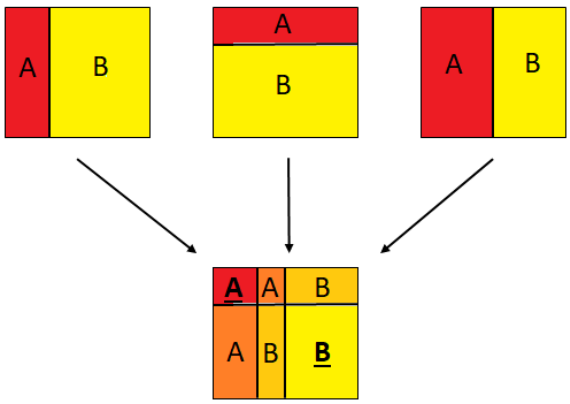


Abbildung 25: Einteilung eines Raums durch Baumstümpfe

In Abbildung 25 ist eine beispielhafte Einteilung eines Raumes durch 3 verschiedene Baumstümpfe abgebildet. Diese 3 Baumstümpfe zusammen teilen den Raum in Bereiche ein, in denen:

1. Zu 100% Klasse A prognostiziert wird, da alle 3 Bäume diesen Bereich Klasse A zuweisen (roter Bereich)

2. Zu 67% Klasse A prognostiziert wird, da 2 der 3 Bäume diesen Bereich Klasse A zuweisen (dunkel-oranger Bereich)

3. Zu 67% Klasse B prognostiziert wird, da 2 der 3 Bäume diesen Bereich Klasse B zuweisen (hell-oranger Bereich)

4. Zu 100% Klasse B prognostiziert wird, da alle 3 Bäume diesen Bereich Klasse B zuweisen (gelber Bereich)

Wie an diesem Beispiel zu sehen ist, wird durch das Übereinanderlegen der Bäume eine rechtwinklige Einteilung im 2-dimensionalen Raum kreiert. Für eine simple Prognose, die die Wahrscheinlichkeiten für eine Klasse nicht beachtet, sondern nur die wahrscheinlichste Klasse betrachtet, können Bereiche zusammengefasst werden. Die roten und dunkel-orangenen Bereiche ergeben einen Bereich, der zu Klasse A gehört, während die hell-orangenen und gelben Bereiche zu einem gemeinsamen Bereich zusammengefasst werden können, der zu Klasse B gehört.

Der unbeschnittene *Random Forest* selbst ist viel komplexer aufgebaut. Anstatt nur einmal die Daten in 2 Bereiche zu trennen, trennt jeder einzelne Baum die Daten solange, bis jede einzelne Beobachtung ein eigenes Gebiet zugewiesen bekommt. All diese Bäume übereinander gelegt ergeben einen weit aus komplexeren Raum (der für mehr als 2 Variablen auch in der Dimension viel komplexer wird), indem viele verschiedene Gebiete mit verschiedenen Vorhersagen liegen.

Wird das hier vorgeschlagene Verfahren zur Erarbeitung der Topologie auf einen *Random Forest* angewendet, so werden Grenzen zwischen den verschiedenen Klassifikationsbereichen gefunden, unabhängig von der Wahrscheinlichkeit, mit der in diesem Bereich klassifiziert wird. Eine solche Grenze kann als direkte Nachbarschaft zweier Klassen bezüglich einer bestimmten Variable interpretiert werden.

Durch diese Einteilung können jedoch leicht eine oder mehrere "Inseln" entstehen, wie in Abbildung 23 angedeutet ist. Wenn also eine kleine Ausreißergruppe dazu beiträgt, dass eine Insel innerhalb einer anderen Klasse entsteht, so kann durch das hier vorgeschlagene Verfahren ein schwaches Nachbarschaftsverhältnis angedeutet werden, wobei die größten Gebiete der beiden Klassen überhaupt nicht aneinander grenzen. Mit "schwachem Nachbarschaftsverhältnis" ist hier gemeint, dass ganz vereinzelt Beobachtungen zwischen den beiden Klassen wechseln, während bei einem "starken Nachbarschaftsverhältnis" viele Beobachtungen zwischen den Klassen wechseln würden, da sie längere gemeinsame Grenzen aufweisen.

4.3.2 Multinomiales Logistisches Regressionsmodell

Der größte Unterschied zwischen der Raumeinteilung durch einen Klassifikations-*Random Forest* und der Raumeinteilung durch ein multinomiales logistisches Regressionsmodell ist das Prinzip der abschnittsweisen Raumeinteilung im Vergleich zu einer stetigen Raumeinteilung. Während der Raum durch den *Random Forest* in Bereiche mit festen Klassifikationswahrscheinlichkeiten eingeteilt wird, welche sich innerhalb eines einzelnen Bereichs nicht

ändert, ändern sich die Klassifikationswahrscheinlichkeiten im multinomialen logistischen Modell stetig im Raum. Ddadurch entstehen "glattere" Übergänge zwischen den verschiedenen Klassifikationsbereichen.

Eine Einteilung in Bereiche ist jedoch trotzdem möglich, da an jedem Punkt im Raum eine bestimmte Klasse als "am wahrscheinlichsten" modelliert wird und somit der gesamte Raum, in dem diese Klasse am wahrscheinlichsten ist, als Bereich für diese Klasse bezeichnet werden kann. Zwischen diesen Bereichen kann das hier vorgeschlagene Verfahren zur Erarbeitung der Topologie angewendet werden und Nachbarschaftsverhältnisse erarbeitet werden.

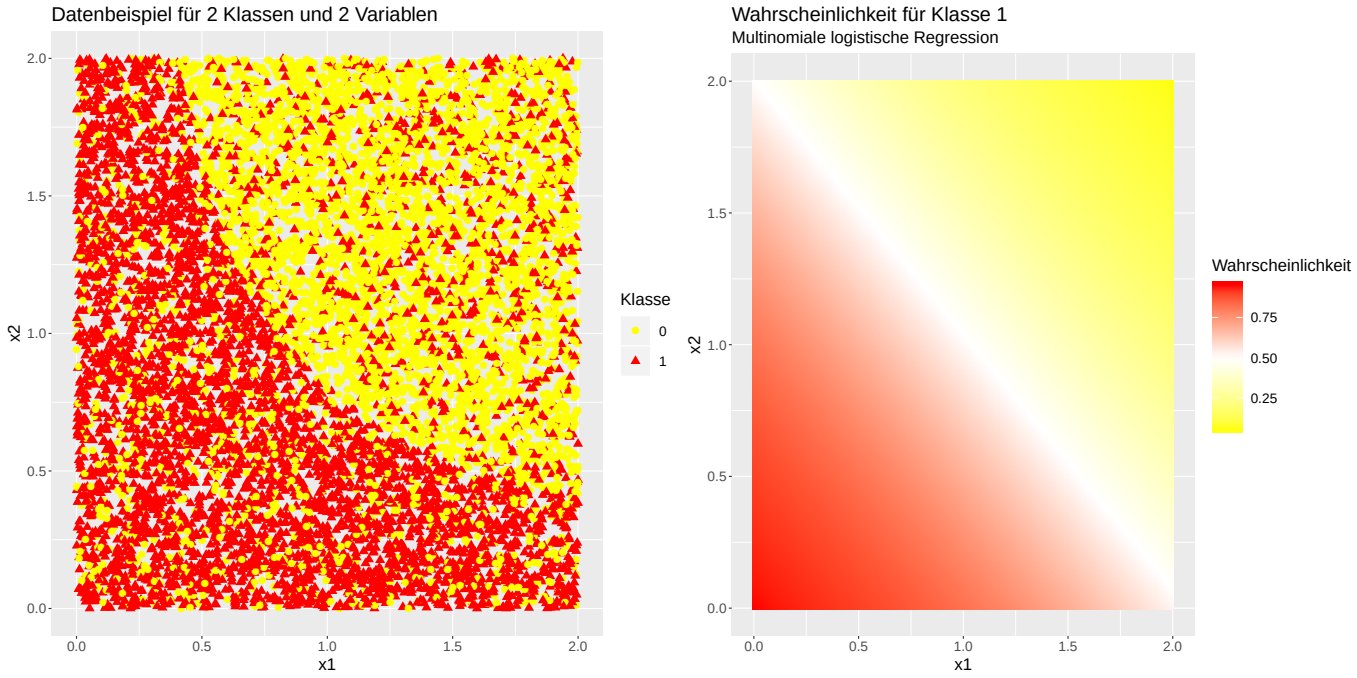


Abbildung 26: Einteilung eines Raums durch Multinomiale logistische Regression

Für Abbildung 26 wurde eine Datengrundlage mit 2 verschiedenen Klassen simuliert. Diese beiden Klassen vermischen sich etwas, können jedoch im 2-dimensionalen Raum deutlich voneinander getrennt werden. Die multinomiale logistische Regression weist jedem Punk im betrachteten Raum eine Wahrscheinlichkeit für Klasse *1* zu, die an einer klar erkennbaren Trenngeraden langsam von über *0.5* auf unter *0.5* wechselt. An dieser Stelle entsteht eine Grenze, die einen Bereich für Klasse *1* von einem Bereich für Klasse *0* trennt.

Für dieses Beispiel würde das vorgeschlagene Verfahren für die Erarbeitung der Topologie ein Nachbarschaftsverhältnis zwischen Klasse *0* und Klasse *1* feststellen, wobei:

- bezüglich **x1** ein Gebiet der Klasse *0* oberhalb eines Gebietes der Klasse *1* liegt bei Konstanthaltung von **x2**
- bezüglich **x2** ein Gebiet der Klasse *0* oberhalb eines Gebietes der Klasse *1* liegt bei Konstanthaltung von **x1**

Ein deutlicher Nachteil der multinomialen logistischen Regression gegenüber der Flexibilität eines *Random Forests* wird jedoch erst bei den Beispielen im folgenden Abschnitt erkannt.

4.3.3 Vergleich zwischen multinomialer logistischer Regression und Random Forest

Für den folgenden Vergleich zwischen der multinomialen logistischen Regression und dem *Random Forest* werden zwei simulierte Datensituationen verglichen (vgl. Abbildung 27).

Die erste Datensituation ähnelt der Datensituation aus Abbildung 26. Hinzu kommt jedoch noch eine dritte Klasse, die in der Nähe des Nullpunkts von **x1** und **x2** vorkommt.

Für die zweite Datensituation wird die dritte Klasse als 2 getrennte Inseln eingeführt, die eine in der Nähe des Nullpunkts und die andere bei hohem **x1** und hohem **x2**. Letztere kommt in ihrem Bereich sogar "rein" vor, das heißt keine Beobachtung einer der anderen beiden Klassen existiert in diesem Bereich.

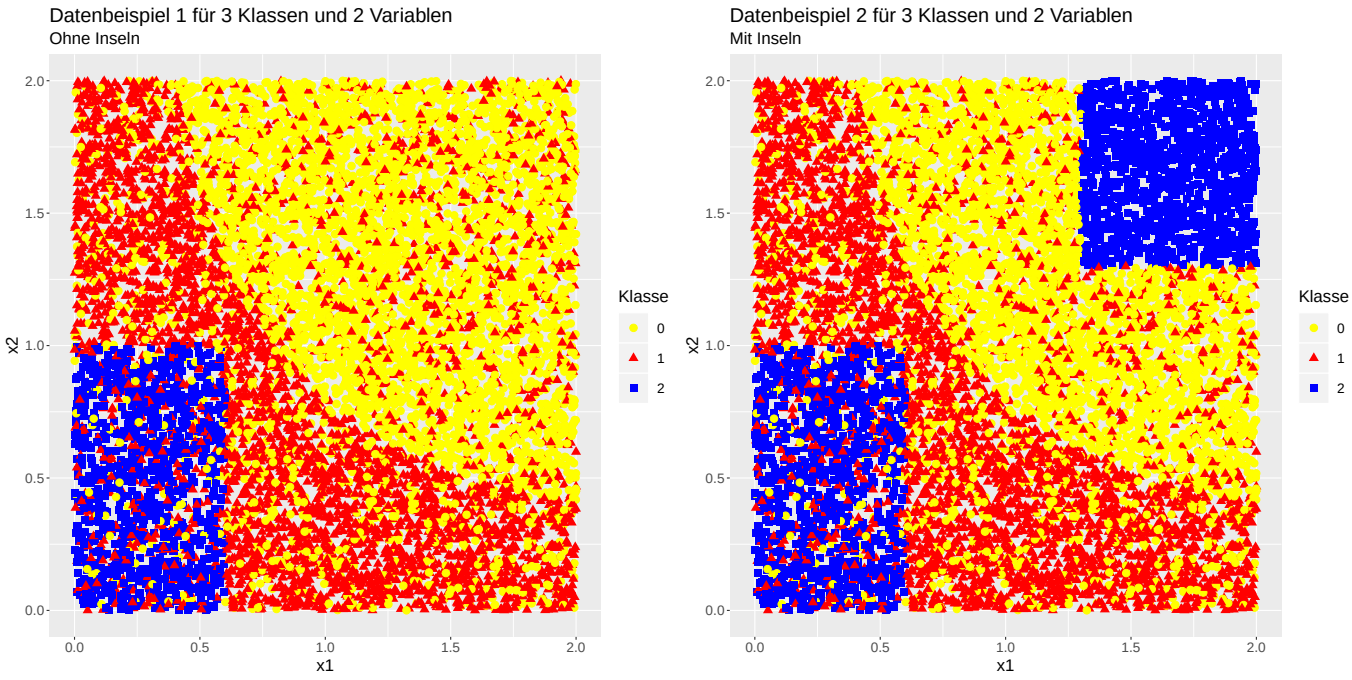


Abbildung 27: Datenbeispiele mit 3 Klassen und 2 Variablen mit und ohne Inseln

Für die beiden Datensituationen wird jeweils ein multinomiales logistisches Regressionsmodell und ein *Random Forest* geschätzt. Künstlich wird nun ein ganz feines Gitter an Punkten genutzt um eine Karte zu erstellen, an welchen Punkten das Modell welche Klasse als am "wahrscheinlichsten" modelliert.

In Abbildung 28 sind 4 verschiedene Raumeinteilungen durch die beiden Modelle zu sehen. In der ersten Zeile ist links die Raumeinteilung für das 1. Datenbeispiel durch das multinomiale logistische Regressionsmodell und rechts die Raumeinteilung für das 1. Datenbeispiel durch

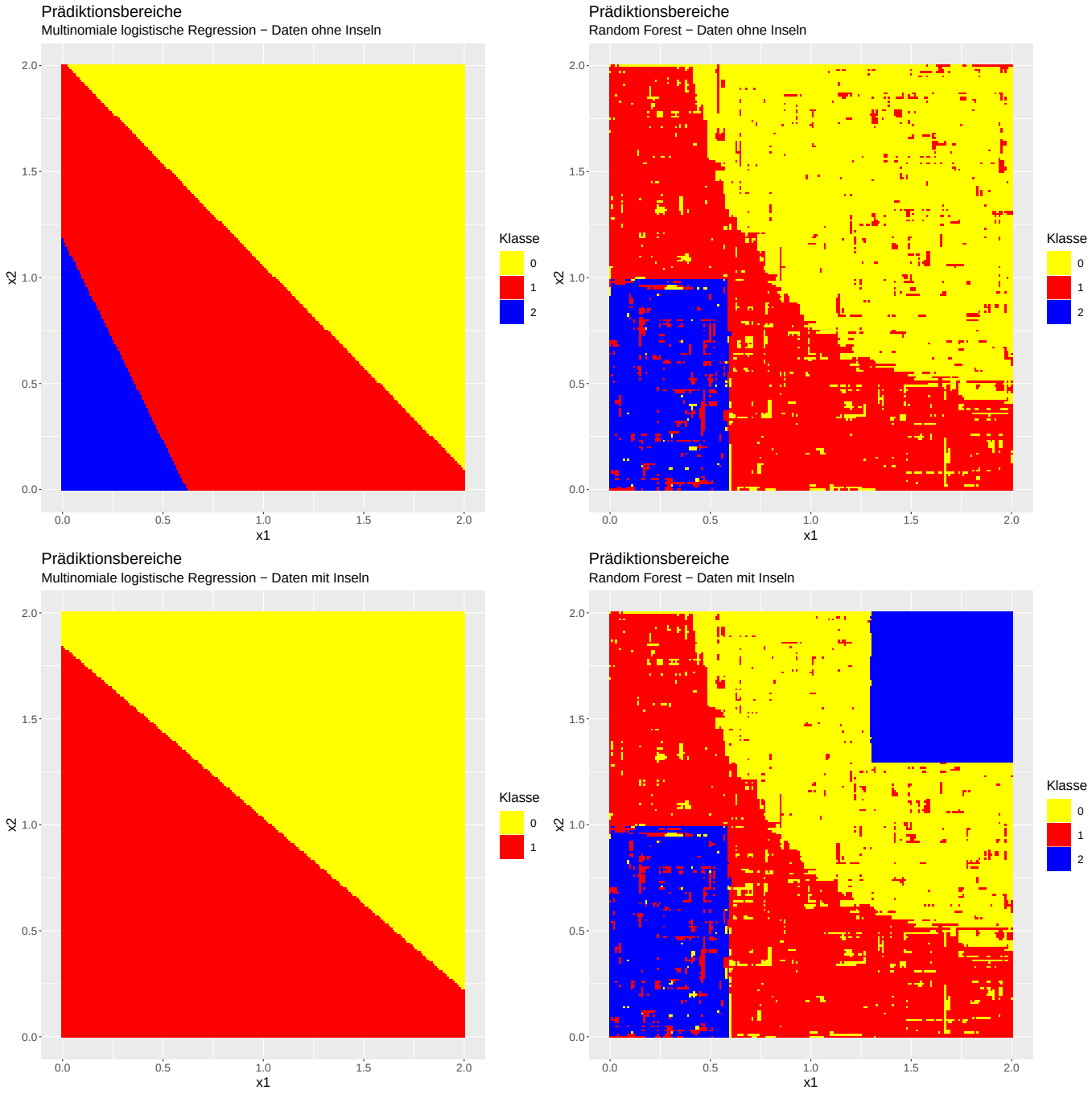


Abbildung 28: Klassifikation durch multinomiales logistisches Regressionsmodell und *Random Forest*

einen *Random Forest* zu sehen. Was schnell auffällt ist, dass der *Random Forest* sowohl die Logik der runden Abtrennung zwischen Klasse *0* und Klasse *1* sehr gut modelliert hat als auch die Logik der linearen Abtrennung zwischen Klasse *1* und Klasse *2*. Das multinomiale logistische Regressionsmodell hat hingegen lineare Abgrenzungen zwischen den Klassen gefunden, die in etwa die Lage der Punkte abbilden.

Ein weiterer Unterschied zwischen den beiden Modellierungen ist die Reinheit der abgetrennten Gebiete. Während durch das multinomiale logistische Regressionsmodell reine Bereiche modelliert werden, in denen jeweils eine Klasse am wahrscheinlichsten vorkommt, modelliert der *Random Forest* viele kleine Inseln innerhalb der großflächigen Bereiche.

Im multinomialen logistischen Regressionsmodell würde sich die Unreinheit dieser Bereiche in den Wahrscheinlichkeiten für die einzelnen Klassen widerspiegeln, was für eine Prädiktion der wahrscheinlichsten Klasse jedoch irrelevant ist.

Ob die Abbildung der Unreinheit im Allgemeinen eine positive oder negative Eigenschaft darstellt, soll an dieser Stelle unbewertet bleiben, da es einerseits die zugrundeliegende Unreinheit widerspiegelt, andererseits jedoch zu einer zu hohen Datenanpassung und damit zu möglichen falschen Prädiktionen führen kann.

Für den hier erbrachten Vorschlag zur Erarbeitung der Topologie der Daten, kann die Unreinheitsmodellierung jedoch zu Problemen führen, da die Hauptlogik, mit der 2 Gebiete voneinander getrennt sind (zum Beispiel Klasse *0* liegt bezüglich **x1** oberhalb von Klasse *1* bei Konstanthaltung von **x2**), unerkannt bleiben kann.

Ein großes Problem der multinomialen logistischen Regression kann im 2. Datenbeispiel erkannt werden. In diesem Beispiel liegt Klasse *2* auf zwei Inseln verteilt im Raum. Der *Random Forest* hat kein Problem die beiden Inseln abzubilden, während in der multinomialen logistischen Regression die Klasse *2* im relevanten Raum überhaupt nicht auftaucht. Die Wahrscheinlichkeit für Klasse *2* ist im gesamten relevanten Raum durch die Wahrscheinlichkeit für Klasse *0* oder Klasse *1* überdeckt.

Mit "relevantem Raum" ist hier der "für die Daten relevante Raum" gemeint. Die Ursprungsdaten lagen in einem Raum $S(\mathbf{x1}, \mathbf{x2})$ mit $0 \leq \mathbf{x1} \leq 2$ und $0 \leq \mathbf{x2} \leq 2$. Auf den gesamten reelen 2-dimensionalen Raum betrachtet, existieren Bereiche, in denen das multinomiale logistische Regressionsmodell Klasse *2* prädiktiert, jedoch ist dies nicht in dem Bereich, in dem Klasse 2 in den Ursprungsdaten auftaucht.

Der "nicht-relevante Raum" ist der Raum, auf den die Modelle nicht trainiert wurden. Da Beobachtungen, die durch diese Modelle prädiktiert werden sollen, eigentlich aus der Grundverteilung stammen sollten, sollten auch keine Probleme durch diese Gebiete entstehen. Es kann jedoch passieren, dass sich in der Grundgesamtheit etwas ändert und die Modelle für Gebiete angewendet werden, auf die sie nicht trainiert wurden. Für manche interpretierbare *Machine Learning*-Methoden werden sogar einige Beobachtungen unvermeidbar in diesen Gebieten erzeugt.

In Abbildung 29 ist zu sehen, dass das multinomiale logistische Regressionsmodell beim "Rauszoomen" des betrachteten Bereichs an einem bestimmten Punkt anfängt Klasse *2* als wahrscheinlichste Klasse zu prädiktieren. Dies liegt jedoch weit außerhalb des Bereichs, in dem Klasse *2* tatsächlich vorkommt. Selbst wenn die beiden Inseln für Klasse *2* außerhalb

des relevanten Raums mit ihren linearen Abgrenzungen weitergeführt werden, so wäre der Raum weit weg von der Stelle, an der Klasse *2* wirklich prädiktiert werden würde.

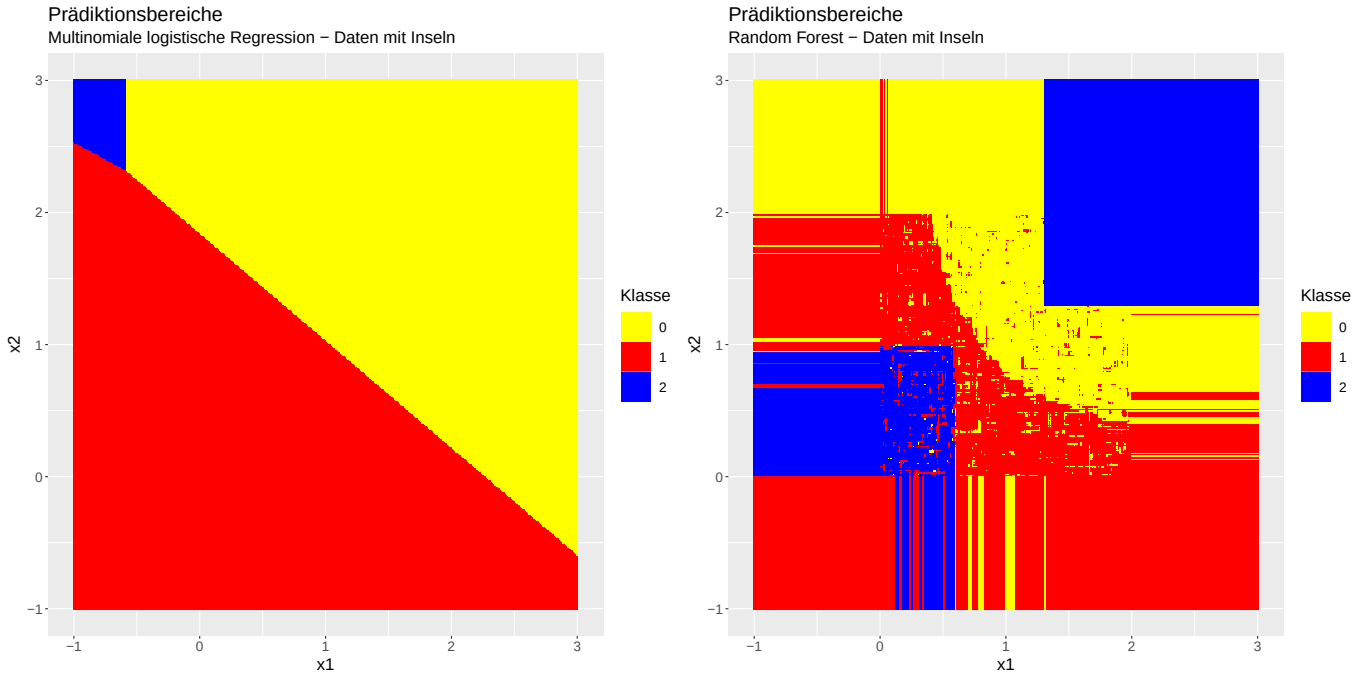


Abbildung 29: Klassifikation durch multinomiales logistisches Regressionsmodell und *Random Forest* außerhalb des relevanten Raumes

Für den *Random Forest* ist zu erkennen, dass der äußerste Punkt innerhalb des relevanten Raumes den Bereich des irrelevanten Raums bestimmt. Wieso dies der Fall ist, kann sogar leicht erklärt werden. Am Rande des Wertebereichs splitten die Bäume des Random Forests einen Raum ab, der in einer Ecke bezüglich 2 Richtungen beschränkt ist und an den Rändern bezüglich 3 Richtungen. Ein Split kann nicht außerhalb des ursprünglichen Wertebereichs erfolgen, weshalb der Rand selbst keine Einschränkung für die Raumaufteilung aufweist. Aufgrund von kleineren Unreinheiten können also außerhalb des relevanten Raumes großflächige Bereiche entstehen, die nichts mit der zugrundeliegenden Logik der Datenverteilung zu tun haben.

Alles in allem sollen trotzdem diese beiden Modelle, die vor allem für das 1. Datenbeispiel verschiedene, aber dennoch gute Grundlogiken, aus den Daten modelliert haben, verwendet werden, um die Lage der Leistungsdaten der Bundesligaspieler im Raum bezüglich ihrer Position zu erarbeiten. Da diese beiden Modelle, wie hier gezeigt, doch sehr unterschiedlich modellieren, sollen Gemeinsamkeiten und Unterschiede für das Erarbeiten der Topologie der Daten ermittelt werden.

5 Ergebnisse

5.1 Hyperparameter Tuning Random Forest

Wie die meisten *Machine Learning* Methoden besitzt auch der *Random Forest* Hyperparameter, die die Prädiktion des Modells verbessern können. Für die Modellierung in dieser Arbeit werden drei dieser Hyperparameter betrachtet:

1. `mtry`: Der `mtry`-Hyperparameter gibt an, wie viele verschiedene Variablen an den einzelnen Splitpunkten in Betracht gezogen werden sollen, um den nächsten Splitpunkt zu bestimmen.
2. `min.node.size`: Der `min.node.size`-Hyperparameter bestimmt, wie viele Beobachtungen in einem Knoten sein sollen, damit der Baum die Daten weiter trennen soll.
3. `ntree`: Der `ntree`-Hyperparameter bestimmt, wie viele Bäume innerhalb des Forests erzeugt werden sollen.

Der `mtry`-Hyperparameter ist in *Random Forests* in den meisten Datensituationen der wichtigste Hyperparameter. Angenommen es wird vermutet, dass nur wenige Einflussgrößen den größten Teil der Daten erklären, dann wäre es wichtig, wenn mindestens einer dieser Einflussgrößen an den Splits ausgewählt werden würde. Für einen solchen Fall wäre ein hoher Wert für `mtry` wichtig. In anderen Datensituationen kann aber auch ein sehr geringer Wert von `mtry` eine gute Modellierung herbeiführen (Liaw and Wiener 2002).

Wie bereits in Abbildung 28 zu sehen ist, bildet ein *Random Forest* aufgrund von kleineren Unreinheiten und Ausreißern diese Unreinheiten mit ab. Dies liegt vor allem daran, dass die Bäume komplett aufgespannt werden. Angenommen es gäbe einen Ausreißer der Klasse *2* in einem Gebiet, in dem sonst nur Klasse *1* vorkommt. Der Ausreißer würde in etwa 62% der Bäume vorkommen (Natur des Bootstrapping), wodurch 62% der Bäume an diesem exakten Punkt Klasse *2* prädiktieren. Da dies mehr als der Hälfte aller Bäume entspricht, prädiktiert der *Random Forest* an exakt dieser Stelle auch für neue Daten Klasse *2*. Das Erhöhen des `min.node.size`-Hyperparameter führt dazu, dass einzelne Ausreißer in einem Terminalknoten mit nahen anderen Beobachtungen vorkommen können. Durch Erhöhen dieses Hyperparameters wird die Gewichtung von Ausreißern reduziert und Unreinheiten somit glättet. Ein zu hoher Wert des `min.node.size`-Hyperparameters würde jedoch dazu führen, dass **echte** Inseln geglättet und nicht mehr erkannt werden würden.

Der `ntree`-Hyperparameter sollte einen gewissen Wert überschreiten, der sich von Datensituation zu Datensituation unterscheidet. Besteht ein *Random Forest* aus zu wenigen Bäumen, könnte es sein, dass die zugrundeliegenden Strukturen noch nicht erkannt worden sind. Aufgrund der Einschränkungen pro Baum (nur wenige Variablen pro Split betrachtet, im Regelfall ungeprunet) könnte es sogar sein, dass ein zu kleiner *Random Forest* schlechter als ein standardmäßiger Klassifikationsbaum modelliert. Sobald eine bestimmte Anzahl an Bäumen gefunden wurde, verbessert sich die Vorhersage jedoch nicht mehr (bzw. nur geringfügig). Die Anzahl der notwenidgen Bäume steigt in der Regel mit der Anzahl der Variablen im *Random Forest* (Liaw and Wiener 2002).

Für den *Random Forest* in dieser Modellierung wurde ein Grid Search für das beste Parametersetting für die folgenden Werte durchgeführt:

Hyperparameter	Minimum	Maximum	Schrittweite
mtry	3	8	1
min.node.size	1	10	1
ntree	200	500	50

Tabelle 6: Hyperparameter-Raum für Grid Search

Als Performance-Maß zum Messen der Güte der einzelnen Hyperparameterkombinationen wurde die *Accuracy* gewählt. Die *Accuracy* misst, welcher Anteil der Beobachtungen durch das Modell richtig klassifiziert wird.

Um dieses Performance-Maß nicht durch Overfitting zu beeinflussen, wurde eine 5-fache Kreuzvalidierung zum Berechnen der *Accuracy* verwendet. Dies bedeutet, dass der Datensatz in 5 Teile eingeteilt wird, auf jeweils 4 dieser Teile das Modell gefittet wird und der 5. Teil, der nicht verwendet wurde, als Testdatensatz verwendet wird. Insgesamt wird dies 5 mal wiederholt, sodass jeder Teildatensatz einmal als Testdatensatz benutzt wird. Aus diesen 5 Moderllierungen ergibt sich eine geschätzte *Gesamtaccuracy*, anhand welcher das beste Parametersetting gefunden werden soll.

In Abbildung 30 sind die Ergebnisse des Hyperparametertunings für den *Random Forest* abgebildet. Jeder Lineplot ist dafür da den Effekt eines bestimmten Hyperparameters auf die *Accuracy* abzubilden, während jede Linie in diesen Lineplots für ein festes Parametersetting der jeweils anderen beiden Hyperparameter steht. Die dicke Linie in der Mitte stellt eine durchschnittliche *Accuracy* pro Parametersetting des betrachteten Hyperparameters dar.

Der obere Lineplot bildet den Effekt des `mtry`-Hyperparameters auf die *Accuracy* ab. Anhand der durchschnittlichen Accuray ist kein eindeutiger Verlauf oder bestes Setting zu erkennen und auch die einzelnen Linien bilden keinen Trend zu einem optimalen Hyperparametersetting ab. Das Maximum liegt bei einem Hyperparametersetting mit einem `mtry` Wert von *6*, was bedeutet, dass an jedem Splitpunkt im *Random Forest* für eine gute *Accuracy* 6 Variablen in Betracht gezogen werden sollen. Es ist jedoch anzumerken, dass das Maximum von 0.684 nur leicht über dem Minimum von 0.666 liegt und somit ein Hyperparametertuning keine wirkliche Verbesserung zu einem ungetuneten *Random Forest* bringt.

Der mittlere Lineplot bildet den Effekt des `min.node.size`-Hyperparameters auf die *Accuracy* ab. Auch dort ist anhand der durchschnittlichen *Accuracy* kein wirklicher Effekt zu erkennen. Das bereits angesprochene Maximum, das mit einem `mtry`-Hyperparameter von *6* gefunden wurde, ist hier bei einem `min.node.size`-Hyperparametersetting von *9* zu finden, was bedeutet, dass die beste *Accuracy* kreuzvalidiert für einen *Random Forest* gefunden wurde, in dem nur bei einer Knotengröße von mindestens 9 Beobachtungen ein weiterer Split durchgeführt wird.

Der untere Lineplot bildet den Effekt des `ntree`-Hyperparameters auf die *Accuracy* ab. Genau wie für die beiden anderen Hyperparameter ist in der durchschnittlichen *Accuracy* kein

wirklich bestes Hyperparametersetting zu erkennen. Das Maximum, das für die beiden anderen Hyperparameter gefunden wurde, liegt bei einem Setting des `ntree`-Hyperparameters von *450*. Dies bedeutet, dass kreuzvalidiert die beste *Accuracy* durch insgesamt 450 Bäume erzeugt wurde.

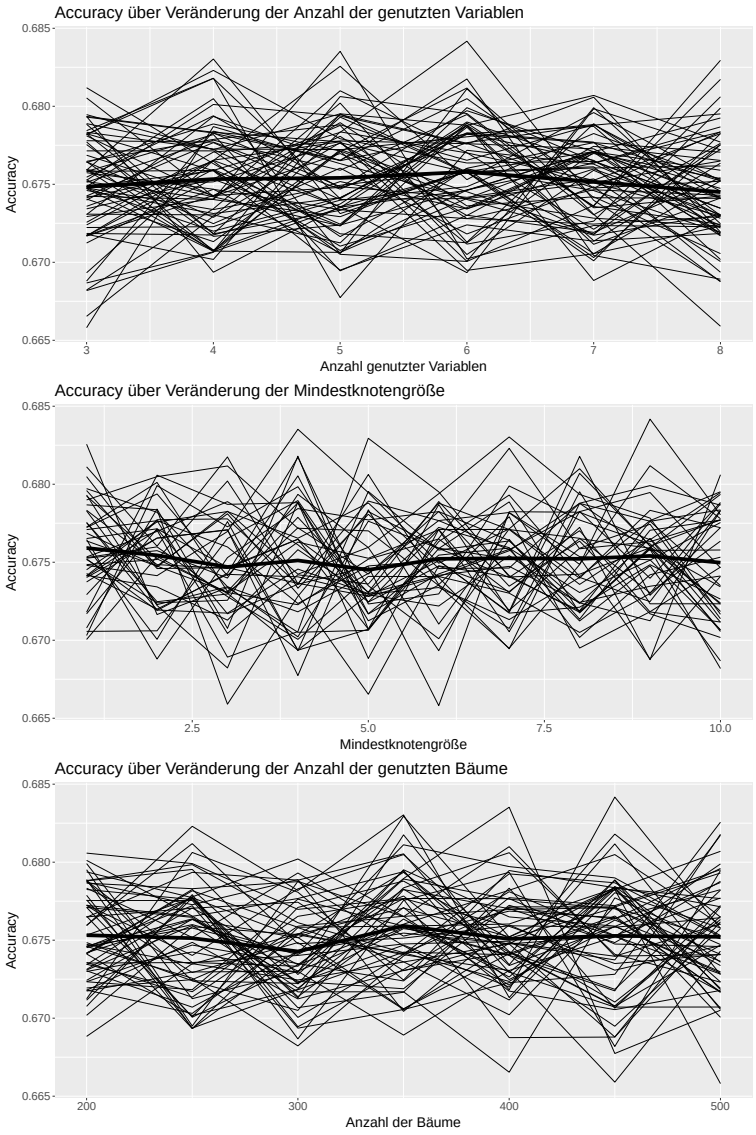


Abbildung 30: Hyperparametertuning für *Random Forest*

Insgesamt scheint das Hyperparametertuning des *Random Forest* bezüglich der *Accuracy* für diesen Datensatz keine großen Unterschiede zu machen. Dem kreuzvalidierten Hyperparametertuning wird jedoch vertraut und das finale Parametersetting von

- $m_{try} = 6$
- $min.node.size = 9$

- $n_{tree} = 450$

gewählt, bei dem eine leichte Verbesserung der *Accuracy* erwartet wird.

5.2 Klassifikationsgüte der Modellierungen

Um zu überprüfen, wie gut die Modelle die Daten einteilen, wird eine 10-fache Kreuzvalidierung auf dem Gesamtdatensatz durchgeführt. Dabei wird der Datensatz in zehn Teile eingeteilt, neun werden zum Erstellen eines Modells verwendet und einer wird verwendet, um eine *Accuracy* zu messen. Dies resultiert in 10 gemessenen *Accuracy*-Werten, die über alle Teile hinweg eine geschätzte Gesamt-*Accuracy* des Modells ergeben. Die Einteilungen des Datensatzes in 10 Teile sind für beide Modellierungen identisch.

	Multinomiale logistische Regression	Random Forest
Accuracy	0.680	0.669

Tabelle 7: Kreuzvalidierte Accuracy für multinomiales logistisches Regressionsmodell und *Random Forest*

In Tabelle 7 ist zu erkennen, dass beide Modelle bei 10-facher Kreuzvalidierung eine ähnliche *Accuracy* aufweisen, wobei das multinomiale logistische Regressionsmodell etwas besser abschneidet. Während des Hyperparameter Tunings wurde jedoch bereits festgestellt, dass die *Accuracy* des *Random Forest* in einem ähnlichen Bereich schwankt, weshalb die beiden Modelle insgesamt als etwa gleich gut bewertet werden.

Darüber hinaus wurde noch überprüft, wie die beiden Modelle für die einzelnen Positionen abschneiden.

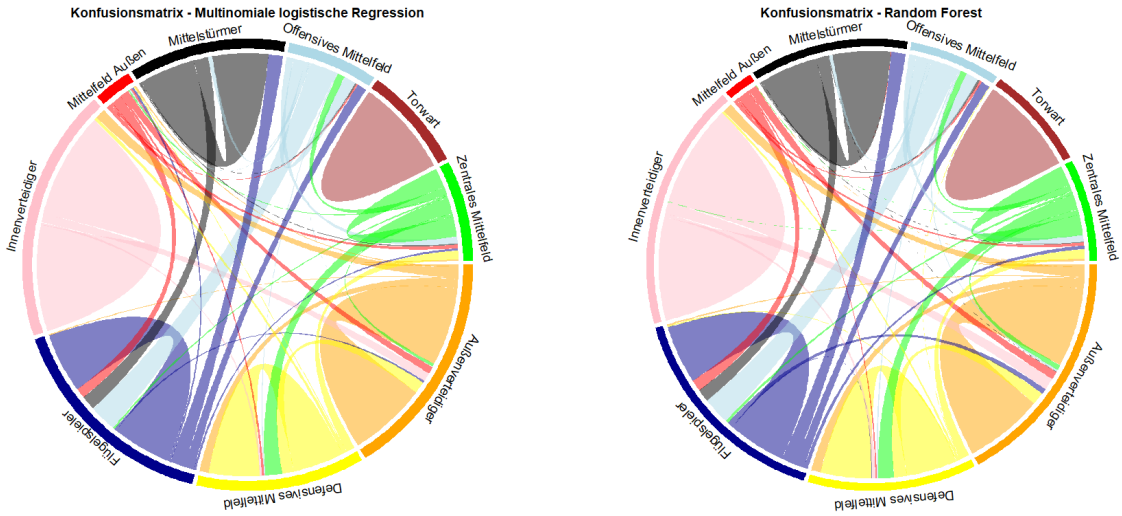
Accuracy für:	Multinomiale logistische Regression	Random Forest
Außenverteidiger	0.77	0.77
Defensives Mittelfeld	0.667	0.657
Flügelspieler	0.66	0.635
Innenverteidiger	0.904	0.869
Mittelfeld Außen	0.07	0
Mittelstürmer	0.764	0.785
Offensives Mittelfeld	0.227	0.234
Torwart	1	1
Zentrales Mittelfeld	0.396	0.396

Tabelle 8: Kreuzvalidierte Accuracy pro Position für multinomiales logistisches Regressionsmodell und *Random Forest*

In Tabelle 8 fällt auf, dass beide Modelle die Positionen in etwa gleich gut modellieren. Manche Positionen sind nahezu perfekt modelliert worden (z.B. die *Torhüter* und die *Innenverteidiger*), wohingegen manche Positionen sehr schlecht modelliert worden sind (z.B. die

äußeren Mittelfeldspieler und die *offensiven Mittelfeldspieler*). Wird die *Accuracy* pro Position mit den Klassengrößen verglichen, fällt auf, dass vor allem kleinere Klassen schlechter modelliert wurden (vgl. Tabelle 4).

Des Weiteren fällt auf, dass die Positionen im Mittelfeld am schlechtesten modelliert werden. Dies könnte dafür sprechen, dass sich die verschiedenen Positionen im Mittelfeld bezüglich der hier ausgewählten Leistungsdaten nicht sehr stark unterscheiden.



(a) Multinomiales logistisches Regressionsmodell (b) *Random Forest*

Abbildung 31: Konfusionsmatritzen als Chordgraphen

Im Gegensatz zu der Visualisierung der Topologie, die in Kapitel 4.2.2.5 anhand von Chordgraphen vorgestellt wurde, werden die Chordgraphen in Abbildung 31 genutzt, um die Fehlklassifikationen der beiden Modellierungen zu vergleichen. Jede eingezeichnete Verbindung, die nicht in der eigenen Klasse endet, steht für einen Anteil der ausgehenden Gruppe, der in die Zielgruppe fehlklassifiziert wurde.

Auf den ersten Blick sehen die beiden Grafiken identisch aus, weisen aber kleine Unterschiede auf (hauptsächlich in der Breite der Anteile). Die in Tabelle 8 aufgeführten auffälligsten Positionen, sind die *Torhüter* und die *äußeren Mittelfeldspieler*. Auch in dieser Grafik unterscheiden sich die beiden Positionen klar von den anderen, da von den Torhütern kein einziger misklassifizierter Teil in eine andere Gruppe ausgeht und von den *äußeren Mittelfeldspielern* kaum eine Verbindung in sich selbst führt (also richtig klassifiziert wurde).

Auf den zweiten Blick fällt auf, dass die Positionen, zwischen denen viel Fehlklassifikation besteht, häufig auch auf dem Spielfeld nebeneinander liegen. Die am schwächsten modellierte Klasse der *äußeren Mittelfeldspieler* wird beispielsweise häufig als *Flügelspieler* (was auf dem Spielfeld direkt vor den *äußeren Mittelfeldspielern* liegt), als *Außenverteidiger* (was direkt dahinter liegt) oder als *zentrale Mittelfeldspieler* (was direkt daneben liegt) klassifiziert. Die Aufgaben, die die Spieler auf diesen Positionen haben, unterscheiden sich nur geringfügig von denen eines *äußeren Mittelfeldspielers*, weshalb die Fehlklassifikationen wahrscheinlich größtenteils in diese Klassen resultieren.

Auch die Position, die am zweit-schlechtesten modelliert wurde, weist diese Zusammenhänge auf. Die *offensiven Mittelfeldspieler* wurden sehr häufig als *Flügelspieler* (was nach außen hin direkt neben den *offensiven Mittelfeldspielern* liegt), als *zentrale Mittelfeldspieler* (was direkt dahinter liegt) oder als *Mittelstürmer* (was direkt davor liegt) klassifiziert.

Es gibt Gütemaße, die die Fehlklassifikationen verschieden gewichten - dies bedeutet beispielsweise, dass eine Fehlklassifikation von Klasse A in Klasse B nicht so schwer gewichtet ist, wie eine Fehlklassifikation von Klasse A in Klasse C - . Wenn die Distanz zwischen zwei Positionen als Gewicht genommen werden würde, könnte es sein, dass die beiden Modelle bezüglich eines solchen Gütemaßes sogar noch besser abschneiden, als die hier gemessene *Accuracy*. Aufgrund der Tatsache, dass im folgenden Teil der Arbeit nur die Prädiktion der wahrscheinlichsten Klasse vorwiegend untersucht wird, ist eine Gütemessung bezüglich eines solchen gewichteten Gütemaßes hier nicht durchgeführt worden.

Alles in allem kann durch diese Betrachtung der Fehlklassifikationen offensichtlich kein großer Unterschied zwischen den beiden grundsätzlich verschiedenen Modellen festgestellt werden.

5.3 Regressionskoeffizienten in multinomialem logistischen Regressionsmodell

Das multinomiale logistische Regressionsmodell hat die Eigenschaft Regressionskoeffizienten zu schätzen, die einzeln interpretiert werden können. Ein positiver Regressionskoeffizient für eine bestimmte Klasse k und eine bestimmte Variable l bedeutet eine höhere Chance auf Klasse k im Vergleich zur Referenzkategorie bei einer höheren Ausprägung der Variable l.

Das Schätzen der Regressionskoeffizienten wurde mit der `multinom`-Funktion aus dem R-Paket `nnet` durchgeführt. Diese schätzt das multinomiale logistische Regressionsmodell mit neuronalen Netzen. Die Konvergenz zu einem Minimum der kleinsten-Quadrate-Approximation ist nach etwa 190 Iterationen eingetreten (+- 10 Iterationen bei der Kreuzvalidierung). Durch diese Modellierung können Punktschätzer und Standardfehler für die Regressionskoeffizienten geschätzt werden. Die geschätzten Regressionskoeffizienten des multinomialen logistischen Regressionsmodells sind in Tabelle 9 abgetragen.

Tabelle 9 zeigt die geschätzten Beziehungen der Referenzkategorie *Zentrales Mittelfeld* mit den anderen Positionen durch die Regressionskoeffizienten des multinomialen logistischen Regressionsmodells auf.

Was bei näherer Betrachtung auffällt, sind die Standardabweichungen der Regressionskoeffizienten für die *Torhüter*, welche oft deutlich größer oder deutlich kleiner als die der restlichen Positionen sind. Bei genauerer Betrachtung der zugrundeliegenden Daten fällt auf, dass für den Teildatensatz der *Torhüter* mehrere Leistungsdaten Null-Vektoren sind. Kein *Torhüter* in diesem Datensatz hat bspw. ein Tor geschossen oder im Abseits gestanden, weshalb die Koeffizienten für die *Torhüter* nicht überinterpretiert werden sollten.

Des Weiteren fallen bei zeilenweiser Betrachtung interessante Zusammenhänge der Leistungsdaten und der Positionen auf. Das *zentrale Mittelfeld* wurde als Referenzkategorie gewählt, da die Position sehr zentral auf dem Fußballplatz steht. Die Reihenfolge der Spalten in

Tabelle 9 wurde auch aufgrund der Platzierung auf einem Fußballfeld festgelegt. Während der *Torhüter* die defensivste Position von allen ist, ist der *Mittelstürmer* die offensivste.

		TW	IV	AV	DM	AM	OM	FS	MS
Gespielte Pässe	coeff.	**-0.08**	**-0.06**	**-0.05**	**-0.02**	**-0.12**	**-0.08**	**-0.13**	**-0.16**
	std.dev.	2.52	0.02	0.02	0.01	0.02	0.02	0.02	0.03
Passquote (in %)	coeff.	**-0.26**	**0.05**	**-0.1**	**0.01**	**-0.09**	**-0.01**	**-0.07**	**-0.13**
	std.dev.	2.67	0.04	0.03	0.03	0.03	0.03	0.03	0.04
Zweikämpfe	coeff.	**-0.42**	**-0.17**	**-0.08**	**-0.1**	**-0.02**	**0.13**	**0.18**	**0.2**
	std.dev.	4.67	0.07	0.05	0.04	0.05	0.05	0.05	0.06
Zweikampfquote (in %)	coeff.	**0.29**	**0.55**	**0.3**	**0.2**	**0.05**	**-0.13**	**-0.08**	**-0.2**
	std.dev.	1.76	0.05	0.04	0.03	0.04	0.04	0.04	0.05
Begangene Fouls	coeff.	**-7.83**	**0.55**	**-1.17**	**0.71**	**-1.24**	**-2.23**	**-2.52**	**-2.25**
	std.dev.	0.27	0.48	0.34	0.27	0.36	0.34	0.34	0.38
Gefoult worden	coeff.	**-2.46**	**-3.35**	**-1.45**	**-0.26**	**-0.56**	**0.21**	**-0.19**	**-0.3**
	std.dev.	0.32	0.45	0.3	0.23	0.31	0.27	0.27	0.34
Laufweite (in km)	coeff.	**-10.21**	**-4.38**	**-2.44**	**-0.22**	**-0.95**	**-0.54**	**-1.2**	**-1.99**
	std.dev.	0.92	0.39	0.27	0.23	0.27	0.23	0.22	0.27
Abseits	coeff.	**6.89**	**-5.92**	**-1.33**	**-6.34**	**4.24**	**5.9**	**6.81**	**8.52**
	std.dev.	0.11	2.52	1.52	1.58	1.17	1.09	1.09	1.14
Vorlagen	coeff.	**-22.02**	**-16.31**	**1.66**	**-5.79**	**3.68**	**3.6**	**4.58**	**2.76**
	std.dev.	0.01	3.34	1.7	1.6	1.6	1.41	1.42	1.7
Tore mit dem Fuß	coeff.	**2.31**	**-10.08**	**-10.31**	**-4.85**	**-3.49**	**3.49**	**2.84**	**5.82**
	std.dev.	0.1	4.16	2.56	1.86	1.93	1.37	1.39	1.56
Kopfballtore	coeff.	**4.65**	**11.36**	**-11.15**	**0.34**	**-8.36**	**8.31**	**4.68**	**16.67**
	std.dev.	0.02	3.02	3.48	2.55	4.05	2.64	2.59	2.76

Tabelle 9: Regressionskoeffizienten des multinomialen logistischen Regressionsmodells mit *Zentralem Mittelfeld* als Referenzkategorie

Was *gespielte Pässe* angeht, so ist dies die wichtigste Aufgabe eines *zentralen Mittelfeldspielers*. Dies spiegelt sich auch in den Regressionskoeffizienten wider, die für alle anderen Kategorien negativ sind. Auffällig ist auch, dass der Regressionskoeffizient negativer wird, je weiter weg ein Spieler von der *zentralen Mittelfeld*-Position spielt (mit den höchsten Beträgen für die *Torhüter* und *Mittelstürmer*).

Die Qualität der Pässe ist jedoch anscheinend für defensivere Positionen höher als für offensive. Den höchsten und positiven Regressionskoeffizienten für die *Passquote* weisen die *Innenverteidigern* auf. Die niedrigsten und negativen Regressionskoeffizienten für die *Passquote* weisen, abgesehen von den *Torhütern*, die *Mittelstürmer* auf.

Die *Anzahl der geführten Zweikämpfe* steigen wohl mit zunehmender offensiver Position des Spielers. Den höchsten und positiven Regressionskoeffizienten weisen die *Mittelstürmer* auf, während den niedrigsten und negativen Regressionskoeffizienten die *Torhüter* gefolgt von den *Innenverteidigern* aufweisen.

Die Qualität der Zweikämpfe ist jedoch anscheinend umgekehrt. Den höchsten und positiven Regressionskoeffizienten für die *Zweikampfquote* weisen die *Innenverteidiger* auf, während die niedrigsten und negativen Regressionskoeffizienten die *Mittelstürmer* aufweisen.

Für die *Anzahl an begangenen Fouls* lässt sich keine klare Struktur auf dem Feld erkennen. Die höchsten positiven Regressionskoeffizienten weisen die *defensiven Mittelfeldspieler* und die *Innenverteidiger* auf, wobei dies die Positionen sind, an denen am häufigsten mit taktischen Fouls gearbeitet wird.

Die am häufigsten *gefoulte* Position ist mit dem einzigen positiven Regressionskoeffizienten die Position des *offensiven Mittelfelds*. Dies deckt sich auf dem Fußballplatz mit der am häufigsten foulenden Position (angreifende gegen verteidigende Mannschaft).

Für die *Laufweite* werden nur negative Regressionskoeffizienten für die verschiedenen Positionen geschätzt, was bedeutet, dass eine hohe *Laufweite* für einen *zentralen Mittelfeldspieler* spricht. Während betragsmäßig die niedrigsten negativen Regressionskoeffizienten für die Mittelfeldpositionen geschätzt werden, sind die negativen Regressionskoeffizienten für die wirklich offensiven und defensiven Positionen betragsmäßig hoch. Dies spricht dafür, dass im Mittelfeld am meisten gelaufen wird.

Was *Abseitsstellungen* und *Vorlagen* angeht, so weisen die wirklich offensiven Positionen einen positiven Regressionskoeffizienten auf, während die defensiven Positionen einen negativen Regressionskoeffizienten aufweisen. Eine Ausnahme bilden die *Außenverteidiger*, die einen positiven Regressionskoeffizienten für *Vorlagen* aufweisen, was dafür spricht, dass sich die *Außenverteidiger* häufig an Angriffen beteiligen. Eine weitere Ausnahme ist der positive Regressionskoeffizient für die *Torhüter* bei der Anzahl an *Abseitsstellungen*. Da dies jedoch wie bereits erwähnt im Datensatz für keinen Torhüter vorgekommen ist, sollte auch dieser Regressionskoeffizient nicht überinterpretiert werden.

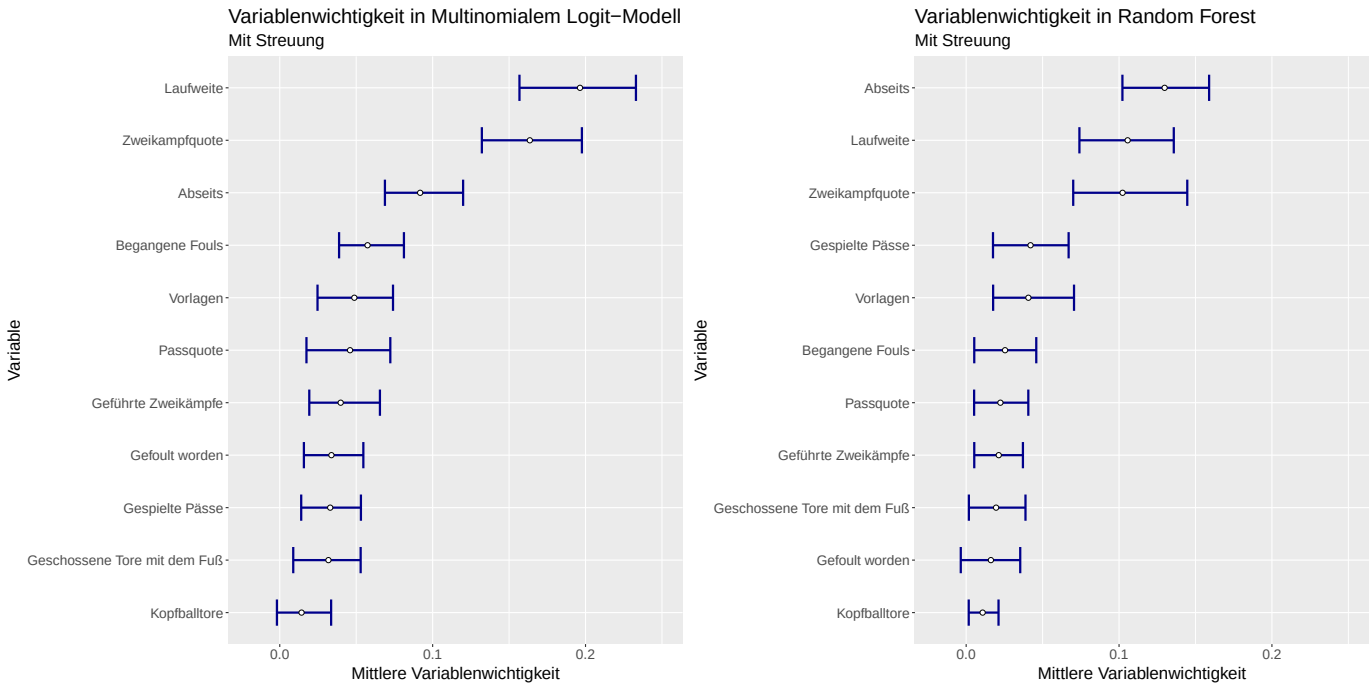
Die höchsten positiven Regressionskoeffizienten für *geschossene Tore* und *Kopfballtore* weisen wie zu erwarten die *Mittelstürmer* auf. Interessant ist hier, dass der zweit-höchste positive Regressionskoeffizient für *Kopfballtore* für die *Innenverteidiger* geschätzt wird. Mit ihrer Größe und Kopfballstärke in der Defensive den Ball aus dem eigenen Strafraum rauszuköpfen und in der Offensive für Gefährlichkeit vor dem gegnerischen Tor zu sorgen ist eine typische Aufgabe für *Innenverteidiger* bei Ecken. Dass diese großen und kopfballstarken Spieler viele *Kopfballtore* erzielen ist also durchaus etwas, das in der Bundesliga beobachtet werden kann.

Alles in allem sind bereits durch die Regressionskoeffizienten Strukturen und Zusammenhänge zwischen den Positionen und den Leistungsdaten zu erkennen. Um diese Zusammenhänge jedoch auch mit den modellierten Zusammenhängen eines *Machine Learning*-Modells vergleichbar zu machen, werden in den folgenden Kapiteln die beiden Modelle mit Methoden verglichen, die für beide Modelle angewendet werden können.

5.4 Variable Importance

Die *Variable Importance* wird wie in Kapitel 4.2.2.1 beschrieben für die beiden Modelle erzeugt. In Abbildung 32 ist die *Variable Importance* für die beiden Modelle dargestellt. Die Rangfolge der Variablen ist durch die Punktschätzer der Variable Importance für die

jeweilige Variable im jeweiligen Modell bestimmt worden. Um Ausreißer nicht zu stark zu gewichten, zeigen die Balken die Streuung durch die Spannweite der gemessenen *Variable Importance*-Werten vom 5%-Quantil bis zum 95%-Quantil an.



(a) Multinomiales logistisches Regressionsmodell (b) *Random Forest*

Abbildung 32: *Variable Importance* für beiden Modelle

In Abbildung 32 ist die *Variable Importance* für das multinomiale logistische Regressionsmodell und den *Random Forest* abgebildet. Die *Variable Importance*-Werte für den *Random Forest* streuen nur unmerklich weniger, als für das multinomiale logistische Regressionsmodell. Darüber hinaus treten die beiden höchsten *Variable Importance*-Werte für die Leistungsdaten in der multinomialen logistisches Regression auf, was bedeutet, dass es für dieses Modell "schlimmer" ist, wenn diese beiden Leistungsdaten permutiert werden, als jede andere Variable für den *Random Forest*. Dies könnte für eine höhere Abhängigkeit der beiden Leistungsdaten mit den anderen Kovariablen sprechen.

Die Reihenfolgen scheinen sich auf den ersten Blick sehr zu unterscheiden. Bei näherer Betrachtung fällt jedoch auf, dass die Reihenfolgen sich sogar ziemlich ähneln.

Die wichtigsten drei Variablen für beide Modelle sind die *Laufweite*, die *Zweikampfquote* und die *Anzahl der Abseitsstellungen*, jedoch in verschiedener Reihenfolge. Dies bedeutet, dass durch das Permutieren dieser drei Leistungsdaten die Prädiktionen der Modelle am schlechtesten werden. Die *Variable Importance* dieser drei Leistungsdaten hebt sich für beide Modelle deutlich von den anderen ab.

Die Ränge dahinter liegen alle ziemlich dicht mit stark überlappenden Streuungsintervallen beieinander. Gemeinsam ist bei beiden Modellen, dass die *Anzahl der Torvorlagen* und die

Anzahl der begangenen Fouls noch eine relativ hohe *Variable Importance* aufweisen, während die *Anzahl der Kopfballtore* und die *Anzahl der geschossenen Tore mit dem Fuß* für beide Modelle eher niedrige Werte der *Variable Importance* aufweisen.

5.5 Partial Dependence Plots

Die Partial Dependence Plots für die beiden Modellierungen sollen Trends für Zusammenhänge zwischen den Leistungsdaten und den Positionen aufzeigen.

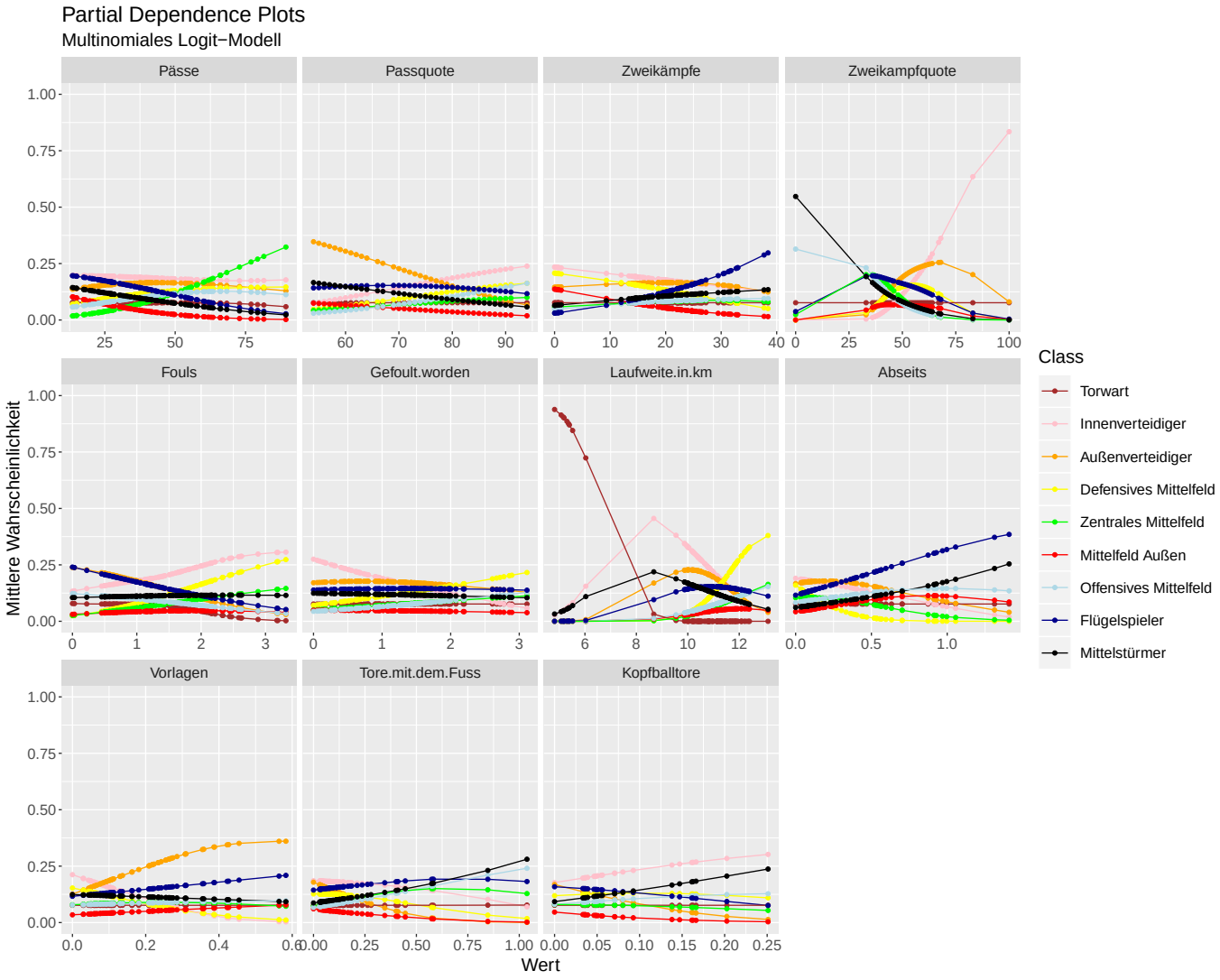


Abbildung 33: Partial Dependence Plot für multinomiales logistisches Regressionsmodell

In Abbildung 33 sind die Partial Dependence Plots für das multinomiale logistische Regressionsmodell aufgeführt. Am auffälligsten sind die Partial Dependence Plots für die *Laufweite* und die *Zweikampfquote*, da dort am meisten Aktivität herrscht. Dies bedeutet, dass der Partial Dependence Plot aufzeigt, dass die mittlere Vorhersage für die verschiedenen Klassen

über den Wertebereich dieser Variablen stark schwankt. Dass diese beiden Variablen wichtig für das Modell sind, wurde bereits in Abbildung 32 gezeigt und bestätigt sich hier.

Ein kleiner Wert der *Laufweite* spricht stark für die Position des *Torhüters* oder des *Innenverteidigers*, während ein hoher Wert für einen *defensiven Mittelfeldspieler* spricht. In Tabelle 9 wurde bezüglich der Laufweite gezeigt, dass alle Positionen im Vergleich zur Referenzkategorie *Zentrales Mittelfeld* negative Koeffizienten aufweisen, also eine hohe *Laufweite* für eine hohe Wahrscheinlichkeit auf *Zentrales Mittelfeld* spricht. In diesem Partial Dependence Plot ist angedeutet, dass die mittlere Wahrscheinlichkeit auf *Zentrales Mittelfeld* am Rand des Wertebereichs anfängt zu steigen, jedoch würde das Maximum der mittleren Wahrscheinlichkeit auf *Zentrales Mittelfeld* außerhalb des Wertebereichs liegen (also in einem Bereich, in dem keine Beobachtungen existieren).

Der Partial Dependence Plot für die *Zweikampfquote* bildet die Regressionskoeffizienten sehr gut ab. Kleine Werte der *Zweikampfquote* sprechen für eine hohe mittlere Wahrscheinlichkeit auf *Mittelstürmer* oder *offensive Mittelfeldspieler*, also für offensivere Positionen. Hohe Werte der *Zweikampfquote* sprechen dagegen für *Innenverteidiger* und *Außenverteidiger*, also eher defensive Positionen. Auch diese Zusammenhänge sind in den Regressionskoeffizienten in Tabelle 9 erkennbar.

Ein Beispiel für eine Variable, die nur wenig Effekt auf die Vorhersagewahrscheinlichkeiten hat, wird im Partial Dependence Plot für die *Anzahl gefoult worden* zu sein dargestellt. Dort ist kaum ein Unterschied in der Prädiktion über den Wertebereich der Variable festzustellen. Auch im Variable Importance Plot (Abbildung 32) war zu erkennen, dass diese Variable für die Prädiktion eher unwichtig ist. Es gibt eine leicht erhöhte Wahrscheinlichkeit auf *Innenverteidiger* bei kleinen Werten für die *Anzahl gefoult worden* zu sein, während hohe Werte für eine leicht erhöhte Wahrscheinlichkeit auf *defensive Mittelfeldspieler* sprechen.

Obwohl die Punktschätzer für die Regressionskoeffizienten in Tabelle 9 für die Variable *Gefoult worden* betragsmäßig viel höher sind als für die *Zweikampfquote*, ergeben sich auf dem gesamten Wertebereich der beiden Variablen für die Variable *Gefoult worden* weniger Schwankungen in den Prädiktionen als für die *Zweikampfquote*. Dies ist etwas, das aus den Regressionskoeffizienten vorher nicht erkannt werden konnte, was in dieser Art des interpretierbaren *Machine Learnings* jedoch sichtbar wird.

Alles in allem bildet der Partial Dependence Plot für das multinomiale logsitische Regressionsmodell die Regressionskoeffizienten ab und lässt darüber hinaus noch andeutungsweise die *Variable Importance* erkennen.

In Abbildung 34 sind die Partial Dependence Plots für den *Random Forest* abgebildet. Die größten Unterschiede, die zu den in Abbildung 33 erzeugten Partial Dependence Plots für die multinomiale logistische Regression auffallen, sind einerseits die kantigere Form der Kurven und andererseits die niedrigeren mittleren Wahrscheinlichkeiten an den Rändern der Wertebereiche. Eine große Gemeinsamkeit ist jedoch, dass die Partial Dependence Plots zu den Variablen, die die höchte *Variable Importance* aufgewiesen haben, auch für den *Random Forest* eine hohe Aktivität aufweisen.

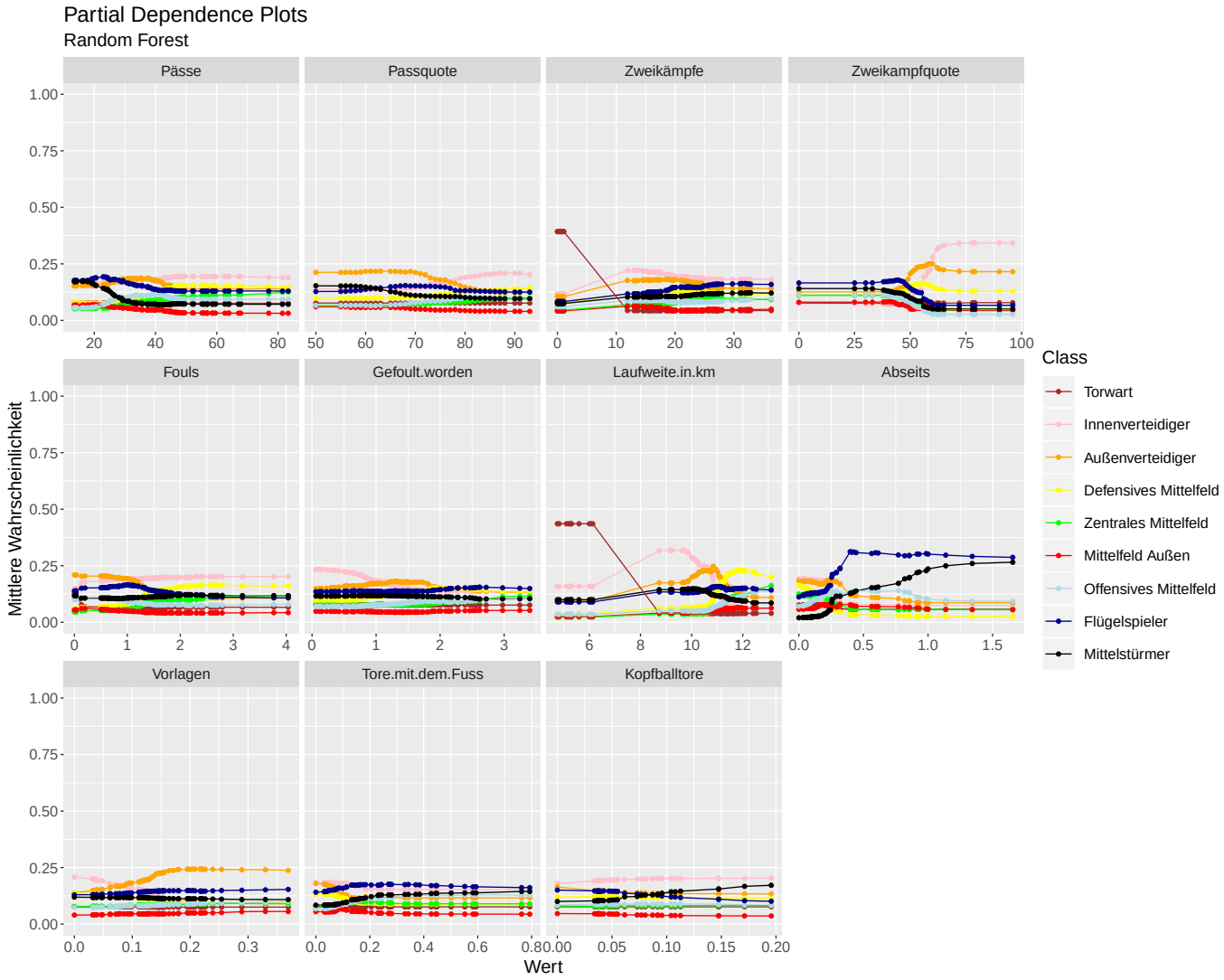


Abbildung 34: Partial Dependence Plot für *Random Forest*

Die *Laufweite* zeigt auch hier für einen geringen Wert eine hohe mittlere Wahrscheinlichkeit für einen *Torhüter*, für eine mittlere *Laufweite* eine hohe mittlere Wahrscheinlichkeit auf einen *Innenverteidiger* und für eine hohe *Laufweite* eine hohe mittlere Wahrscheinlichkeit auf einen *defensiven Mittelfeldspieler*.

Auch der Partial Dependence Plot für die *Zweikampfquote* zeigt einige Gemeinsamkeiten zwischen den Modellierungen auf. Ein hoher Wert der *Zweikampfquote* spricht für eine höhere Wahrscheinlichkeit für *Innenverteidiger* und *Außenverteidiger*. Für einen niedrigeren Wert der *Zweikampfquote* zeigen sich hier jedoch einige größere Unterschiede zwischen den beiden Modellen. Während für niedrige Werte der *Zweikampfquote* die *Mittelstürmer* für beide Modelle die höchste mittlere Wahrscheinlichkeit aufzeigen, ist die Wahrscheinlichkeit im multinomialen logistischen Regressionsmodell ziemlich hoch, im *Random Forest* jedoch nur leicht über den anderen Klassen. Darüber hinaus zeigt das multinomiale logistische Regressionsmodell für niedrige Werte der *Zweikampfquote* auch eine hohe mittlere Wahrscheinlichkeit

für einen *offensiven Mittelfeldspieler*.

Auch für die anderen Variablen ergeben sich viele Ähnlichkeiten, aber auch größere Unterschiede. Der Partial Dependence Plot für die *Anzahl der gespielten Pässe* zeigt zum Beispiel im multinomialen logistischen Regressionsmodell für hohe Werte eine sehr hohe mittlere Wahrscheinlichkeit auf einen *zentralen Mittelfeldspieler*. Der Partial Dependence Plot des *Random Forests* hingegen weist für einen *zentralen Mittelfeldspieler* nur eine recht geringe mittlere Wahrscheinlichkeit für eine hohe *Anzahl an Pässen* auf.

Des Weiteren ist höchst auffällig, dass die mittlere Wahrscheinlichkeit für *Innenverteidiger* und für *Mittelstürmer* bei einem hohen Wert an *Kopfballtoren* zwar für beide Modelle in den Partial Dependence Plots am höchsten ist, jedoch der Wert an sich im multinomialen linearen logistischen Regressionsmodell etwas höher ist als im *Random Forest*. An dieser Stelle sollte aber auffallen, dass es sich bei *Kopfballtoren* um ein eher seltenes Ereignis handelt. Dementsprechend sollte es auch viele Kovariablensettings geben, für die nie eine hohe Anzahl an Kopfballtoren vorgekommen ist. Die mittleren Wahrscheinlichkeiten an diesem Rand mit allen Beobachtungen zu bestimmen, stellt die beiden Modelle also vor das Problem, dass Beobachtungen in Gebieten erzeugt werden, auf welche sie gar nicht trainiert wurden (vgl. Kapitel 4.3.3). Dies könnte in den ICE-Plots auffallen und genau diese Problematik wird in den ALE-Plots versucht zu beheben.

5.6 Individual Conditional Expectation Plots

Insgesamt wurden durch die 11 verschiedenen Variablen, 9 verschiedenen Positionen, 2 verschiedenen Modelle und 3 Arten von ICE-Plots 594 verschiedene ICE-Plots erzeugt. Hier werden jedoch lediglich die wichtigsten und auffälligsten von ihnen interpretiert. Die anderen ICE-Plots sind im elektronischen Anhang zu finden.

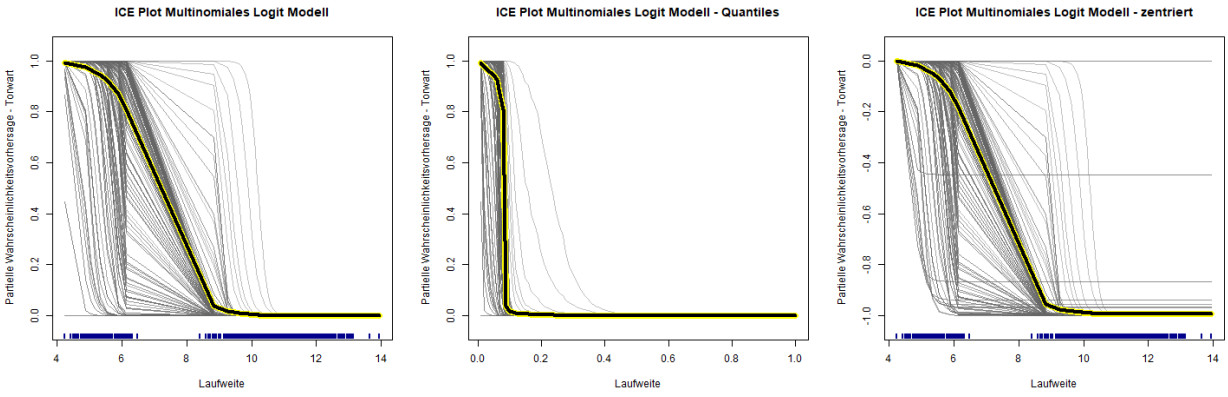


Abbildung 35: ICE-Plot für die Wahrscheinlichkeit auf die *Torhüter*-Position bezüglich der *Laufweite* - multinomiales logistisches Regressionsmodell

Da durch alle Beobachtungen ein viel zu unüberischtlicher Plot entstehen würde, wurden die ICE-Plots auf ein kleines Subsample beschränkt. In diesem Subsample sind alle Beobachtun-

gen der Position, für welche der ICE-Plot die Wahrscheinlichkeiten angibt, und zusätzlich genau so viele Beobachtungen zufällig aus dem restlichen Datensatz ausgewählt.

Bereits bei der *Variable Importance* und in den Partial Dependence Plots wurde die *Laufweite* als wichtige Variable für die Modellierung bestimmt. In Abbildung 35 ist der ICE-Plot für die Wahrscheinlichkeit darauf ein *Torhüter* bezüglich der Laufweite im multinomialen logistischen Regressionsmodell zu sein. Auffällig ist der Wertebereich der *Laufweite*, da dieser eine größere Sprungstelle beinhaltet. Diese Sprungstelle ist in den Dichteplots in Kapitel 3.2.1 nicht zu erkennen, da dort die *Torhüter* nicht miteinbezogen wurden. Durch diese Spieler entsteht jedoch eine stark bimodale Verteilung.

Im linken ICE-Plot in Abbildung 35 ist ersichtlich, dass nach der Sprungstelle die durch das Modell prognostizierte Wahrscheinlichkeit darauf *Torhüter* zu sein für die meisten Beobachtungen äußerst gering ist. Für einen sehr hohen Wert der *Laufweite* ist die Wahrscheinlichkeit darauf *Torhüter* zu sein sogar für **alle** Beobachtungen verschwindend gering. Auch fällt auf, dass für den niedrigsten Wert der *Laufweite* durch die Kovariablenkombinationen der Beobachtungen der gesamte Wertebereich von *0* bis *1* abgedeckt wird, jedoch nahezu alle Kurven bei einer Wahrscheinlichkeit von fast *1* beginnen. Warum dies interessant ist, zeigt sich im Vergleich mit den ICE-Plots des *Random Forests*. Der Quantils-ICE-Plot zeigt diesen Abfall sogar noch extremer auf, da die Lücke der Sprungstelle verschwindet. Im zentrierten ICE-Plot ist sogar zu erkennen, dass die Wahrscheinlichkeit für **alle** Beobachtungen ab dem niedrigsten Wert der *Laufweite* monoton fällt.

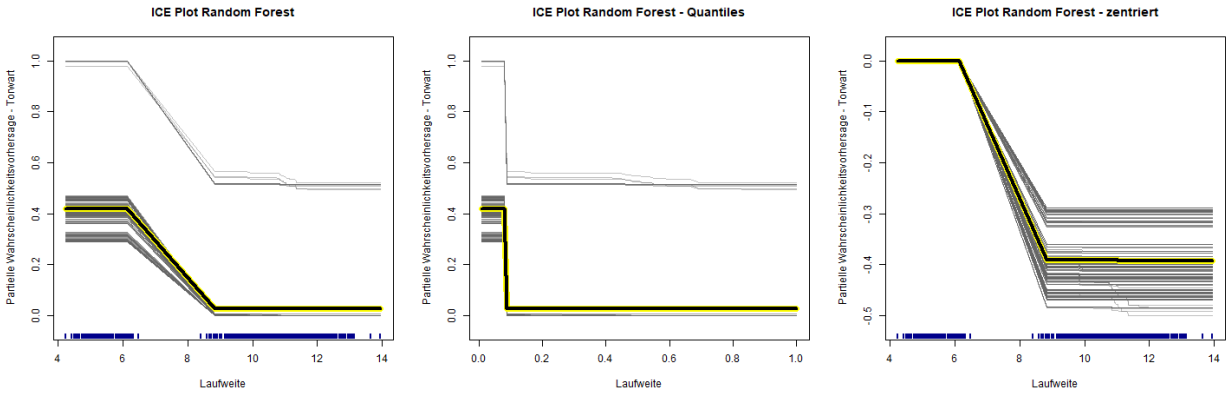


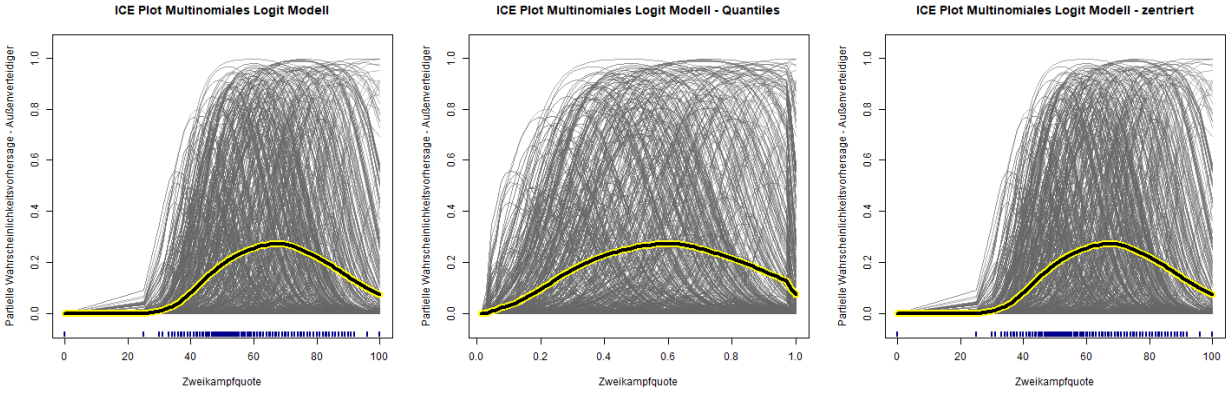
Abbildung 36: ICE-Plot für die Wahrscheinlichkeit auf die *Torhüter*-Position bezüglich der *Laufweite* - *Random Forest*

Der grundsätzliche Trend, den auch die ICE-Plots für das multinomiale logistische Regressionsmodell gezeigt haben, ist auch für den *Random Forest* in Abbildung 36 zu erkennen. Steigt die *Laufweite*, so verringert sich für die Beobachtungen die durch das Modell geschätzte Wahrscheinlichkeit darauf ein *Torhüter* zu sein. Der linke ICE-Plot zeigt dabei einen großen Unterschied zu den ICE-Plots aus Abbildung 35. Während für das multinomiale logistische Regressionsmodell auf dem gesamten Intervall [0,1] der geschätzten Wahrscheinlichkeit Beobachtungen am unteren Rand des Wertebereichs der *Laufweite* beginnen, beginnen die Kurven für den *Random Forest* bei einer Höhe von mindestens *0.3*. Darüber hinaus begin-

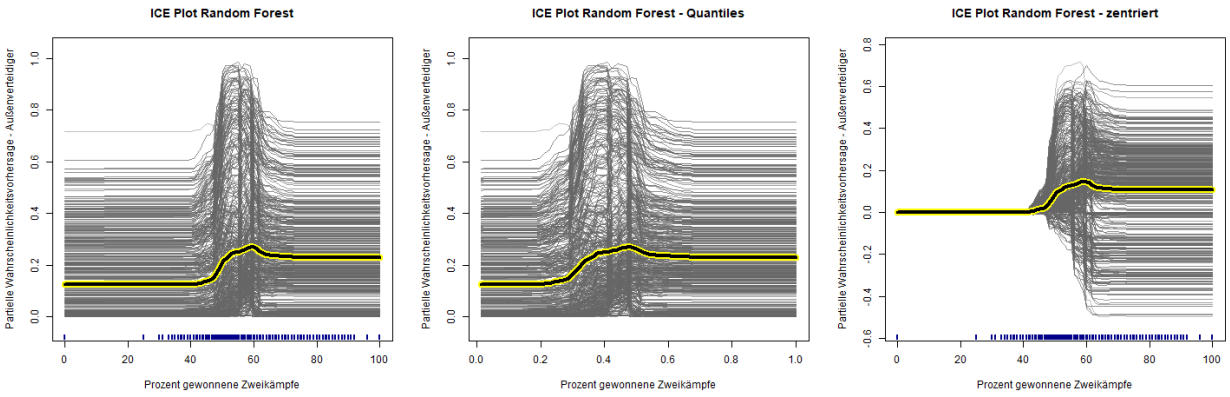
nen die meisten Kurven im Intervall zwischen *0.3* und *0.5* und nicht bei fast *1* wie in den ICE-Plots für das multinomiale logistische Regressionsmodell. Diesbezüglich unterscheiden sich die beiden Modelle also stark voneinander.

Ein weiterer auffälliger Unterschied zwischen den beiden Modellen ist das Verhalten im unteren Teil des Wertebereichs. Während im multinomialen logistischen Regressionsmodell die Wahrscheinlichkeit darauf *Torhüter* zu sein bereits unterhalb der Sprungstelle abnimmt, bleibt sie für den *Random Forest* in diesem Bereich konstant. Dies kann so gedeutet werden, dass der "niedrigste" Split, der für die *Laufweite* in **allen** Bäumen gemacht wird, erst in etwa bei der Sprungstelle beginnt. Der Zusatz "in etwa" ist wichtig, da durch das Bootstrap-Sampling Bäume existieren, in denen der höchste Wert vor der Sprungstelle nicht vorkommt, und daher nicht als Splitkriterium genommen werden kann. Dadurch bestimmt der jeweils nächst höchste Punkt unterhalb der Sprungstelle den Splitpunkt in diesen Bäumen.

Grundsätzlich ist jedoch zu sagen, dass der *Random Forest* die Sprungstelle als eine solche modelliert, obwohl ein Split in einem Baum nicht von der Höhe einer metrischen Variable abhängt, sondern nur von ihrer Einordnung in den Wertebereich.



(a) Multinomiales logistisches Regressionsmodell



(b) *Random Forest*

Abbildung 37: ICE-Plots für die Wahrscheinlichkeit auf die *Außenverteidiger*-Position bezüglich der *Zweikampfquote*

Dieses Verhalten könnte ein großer Indikator dafür sein, warum die *Laufweite* durch die *Variable Importance* so wichtig eingestuft wurde. Mit dem die Wahrscheinlichkeit für einen *Torhüter* nach der Sprungstelle sinkt, steigt sie natürlich für andere Klassen. Eine Variable, die eine Gruppe nahezu "rein" absplitten kann, ist für jede Art der Modellierung äußerst wichtig.

Abgesehen von der *Laufweite* ist auch die *Zweikampfquote* für beide Modelle eine wichtige Variable (vgl. Kapitel 5.4). Ein großer Unterschied zwischen den beiden Modellierungen kann anhand des ICE-Plots für die modellierte Wahrscheinlichkeit für die *Außenverteidiger*-Position in Abbildung 37 erkannt werden.

Die Wahrscheinlichkeitsvorhersage für *Außenverteidiger* liegt für **alle** Beobachtungen im multinomialen logistischen Regressionsmodell für $X_{Zweikampfquote} = 0$ bei nahezu *0%*, unabhängig von der Ausprägung der Kovariablen. Für den *Random Forest* ist die Spannweite der Wahrscheinlichkeitsvorhersagen für die *Außenverteidiger*-Position um einiges größer, wodurch eine klare Abhängigkeit zu den Kovariablen für diesen Punkt festgestellt werden kann.

Auch der weitere Verlauf unterscheidet sich in den beiden Modellen. Wenn eine Kurve bei einem Wert von exakt *0%* beginnt, kann im zentrierten ICE-Plot kein negativer Wert für diese Kurve aufkommen, da die Wahrscheinlichkeit der Kurve nicht unter *0%* fallen kann. Während also für das multinomiale logistische Regressionsmodell im zentrierten ICE-Plot keine (oder nur minimale) Werte unter *0* vorkommen, tauchen im zentrierten ICE-Plot für den *Random Forest* sehr hohe negative Werte auf. Selbst an den Stellen, an denen der Partial Dependence Plot des *Random Forest* einen Anstieg der mittleren Wahrscheinlichkeitsvorhersage für die *Außenverteidiger*-Position aufzeigt, existieren im zentrierten ICE-Plot hohe negative Werte und vor allem auch weiter sinkende Verläufe. Dies impliziert an dieser Stelle eine hohe Abhängigkeit mit den Kovariablen, die die Wahrscheinlichkeitsvorhersage diktieren.

Eine hohe Kovariablenabhängigkeit kann jedoch auch in den ICE-Plots für das multinomiale logistische Regressionsmodell festgestellt werden. Würden Kovariablen überhaupt keinen Einfluss auf die Prädiktion durch eine Variable haben, so wären alle Kurven im ICE-Plot identisch. Die Kurven besitzen der Natur des multinomialen logistischen Regressionsmodells geschuldet eine ähnliche Form. Wie jedoch in den ICE-Plots in Abbildung 37 zu entnehmen ist, bilden die Kurven alle Glockenkurven mit jeweils einem Maximum, an dem die Wahrscheinlichkeitsvorhersage durch das multinomiale logistische Regressionsmodell am höchsten ist. Diese Maxima der einzelnen Kurven liegen aber über eine große Spannweite des Wertebereichs von $X_{Zweikampfquote}$ verteilt, was für eine hohe Ahängigkeit von den Kovariablen spricht.

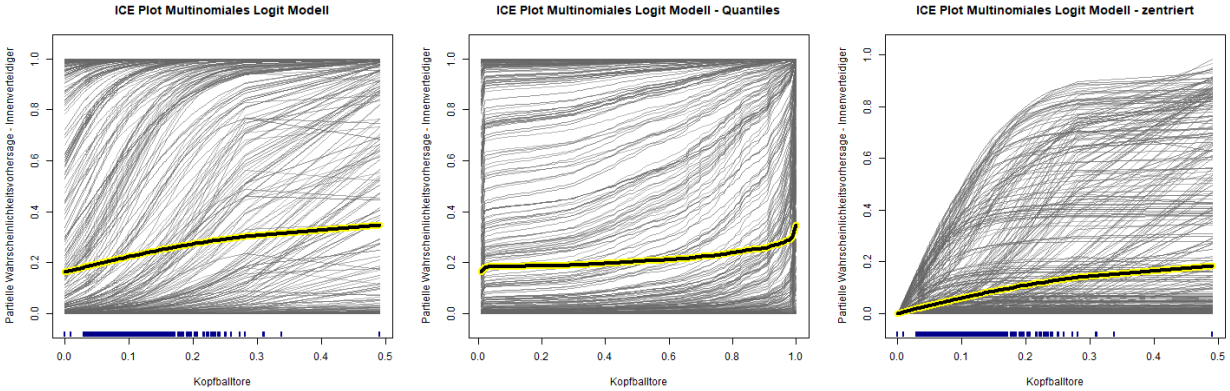
In Abbildung 38 ist im Gegensatz zu den voran gegangenen ICE-Plots mit der *Anzahl an Kopfballtoren* eine Variable abgebildet, die laut der *Variable Importance* für beide Modellierungen eher unwichtig ist (vgl. Kapitel 5.4).

Wie in Abbildung 6 zu sehen ist, ist die Verteilung der Variable $X_{Kopfballtore}$ sehr linkssteil. Daher bietet es sich an den Effekt dieser Variable anhand der Quantile zu beurteilen, damit eine Steigung des ICE-Plots nicht über- oder unterinterpretiert wird. Da die *Innenverteidiger*-Position bereits bei den Regressionskoeffizienten in Tabelle 9 und im Partial Dependence

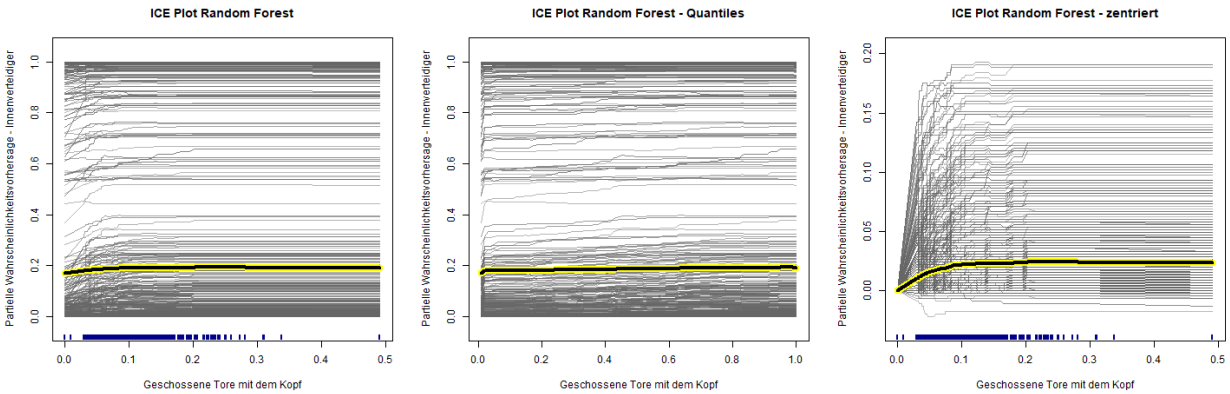
Plot in Abbildung 33 für die *Anzahl an Kopfballtoren* aufgefallen ist, wird diese Position für die ICE-Plots verwendet.

Für das multinomiale logistische Regressionsmodell zeigt der linke ICE-Plot für die meisten Beobachtungen einen sehr schnellen Anstieg der Wahrscheinlichkeitsvorhersage für einen *Innenverteidiger*, während der Quantils-ICE-Plot zwar einen stetigen, aber langsam wachsenden Anstieg für die meisten Beobachtungen abbildet. Es lassen sich Kurven finden, die bei niedrigen Werten von $X_{Kopfballtore}$ noch eine sehr niedrige Wahrscheinlichkeitsvorhersage für die *Innenverteidiger*-Position aufzeigen, für hohe Werte jedoch eine sehr hohe. Im zentrierten ICE-Plot ist dies gut zu erkennen, da eine Kurve, die für hohe Werte von $X_{Kopfballtore}$ fast den Wert *1* erreicht, eine sehr niedrige Wahrscheinlichkeit für niedrige Werte von $X_{Kopfballtore}$ und eine hohe Wahrscheinlichkeit für hohe Werte aufgewiesen hat.

Genau dies ist für den *Random Forest* nicht der Fall. Im Quantils-ICE-Plot ist zu erkennen, dass die meisten Kurven zwar mit ansteigendem Wert von $X_{Kopfballtore}$ wachsen, jedoch nur geringfügig. Der zentrierte ICE-Plot zeigt auch, dass keine Kurve von ihrem Startwert um mehr als *0.2* ansteigt.



(a) Multinomiales logistisches Regressionsmodell



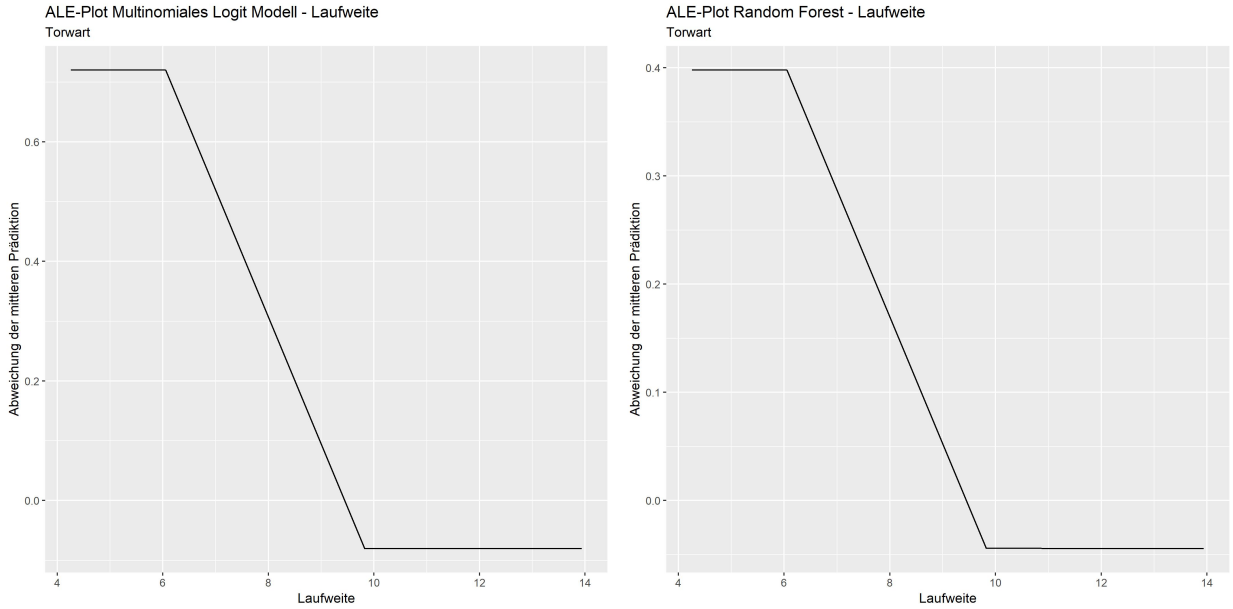
(b) *Random Forest*

Abbildung 38: ICE-Plots für die Wahrscheinlichkeit auf die *Innenverteidiger*-Position bezüglich der *Kopfballtore*

Insgesamt zeigen diese Beispiele für die ICE-Plots und auch die Partial Dependence Plots, dass die einzelnen Leistungsdaten innerhalb der Modelle ähnlich funktionieren (beide wachsen, beide fallen, beide zeigen intervallmäßige An- und Abstiege, ...). Bei genauerer Betrachtung weisen sie jedoch viele Unterschiede auf, wie genau die Variablen die Prädiktionen beeinflussen. Obwohl die beiden Modelle wie in Kapitel 5.2 gezeigt eine ähnliche Modellgüte aufweisen und sogar für die einzelnen Klassen ähnlich gut funktionieren, funktionieren die Modelle intern also sehr verschieden. Anhand des Beispiels in Abbildung 29, in dem dargestellt wurde, wie die beiden Modelle in einem Bereich funktionieren, für den sie nicht trainiert wurden, kann vermutet werden, dass die größten Unterschiede, die in den Partial Dependence Plots und den ICE-Plots gezeigt wurden, dadurch entstehen, dass unwahrscheinliche Variablenkombinationen erzeugt werden. Im nächsten Abschnitt werden ALE-Plots für den Vergleich der beiden Modelle verwendet, die genau gegen dieses Problem vorgehen sollen.

5.7 Accumulated Local Effect Plots

Genauso wie für die ICE-Plots sind für die ALE-Plots mit 9 verschiedenen Klassen, 11 verschiedenen Variablen und 2 verschiedenen Modellen insgesamt 198 Plots entstanden, welche alle in dieser Arbeit zu interpretieren zu viel wäre. Daher wurden auch hier verschiedene ALE-Plots für Variablen und Klassen ausgewählt, die bereits in den Kapiteln zuvor aufgefallen und daher von besonderem Interesse sind.



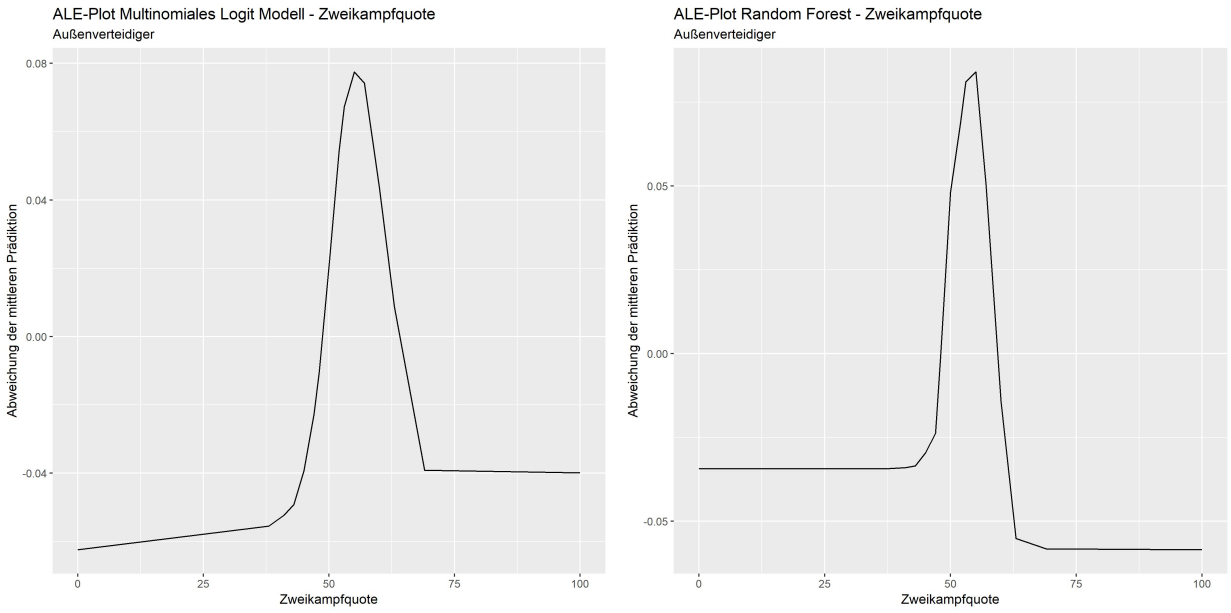
(a) Multinomiales logistisches Regressionsmodell (b) *Random Forest*

Abbildung 39: ALE-Plots für die Wahrscheinlichkeit auf die *Torhüter*-Position bezüglich der *Laufweite*

In Abbildung 39 sind die ALE-Plots für die *Torhüter*-Position im Bezug auf die *Laufweite* dargestellt. In den ICE-Plots in den Abbildungen 35 und 36 war ein klarer Abwärtstrend

mit Sprungstelle zu sehen. Ein großer Unterschied zu den ICE-Plots sollte beim betrachten des ALE-Plots für das multinomiale logistische Regressionsmodell auffallen. Während im dazugehörigen ICE-Plot und auch im Partial Dependence Plot bereits im unteren Teil des Wertebereichs von $X_{Laufweite}$ ein Gefälle zu erkennen war, scheint der ALE-Plot in diesem Bereich konstant.

Der Unterschied zwischen der Wahrscheinlichkeitsvorhersage für einen *Torhüter* im unteren Bereich verglichen mit dem oberen Bereich von $X_{Laufweite}$ ist hinsichtlich des ALE-Plots für den *Random Forest* viel niedriger als für das multinomiale logistische Regressionsmodell. Dieser Effekt ist an der y-Achse leicht zu abzulesen. Im unteren Teil des Wertebereichs der *Laufweite* wird für die Daten, deren wahrer Wert innerhalb dieses Bereichs liegen eine um mehr als *0.7* höhere Wahrscheinlichkeit prädiktiert, als durchschnittlich für alle Beobachtungen, während die Wahrscheinlichkeit im *Random Forest* für diesen Bereich nur um etwa *0.4* höher ist.



(a) Multinomiales logistisches Regressionsmodell (b) *Random Forest*

Abbildung 40: ALE-Plots für die Wahrscheinlichkeit auf die *Außenverteidiger*-Position bezüglich der *Zweikampfquote*

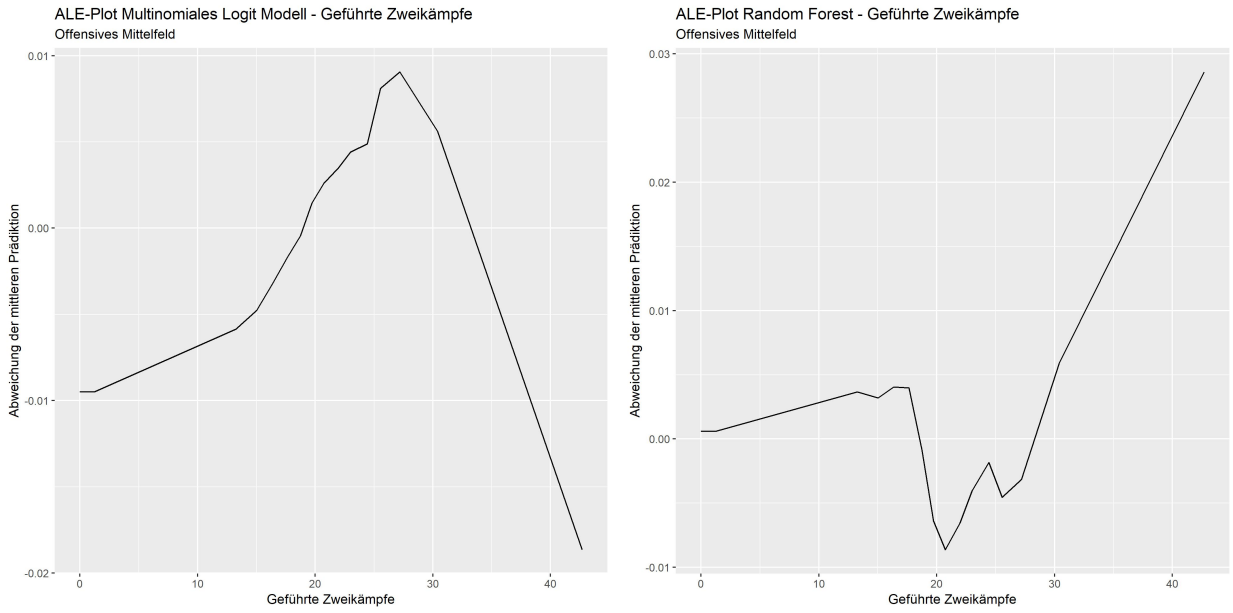
In Abbildung 40 ist der Einfluss der *Zweikampfquote* auf die Wahrscheinlichkeit auf die *Außenverteidiger*-Position in ALE-Plots dargestellt. Diese zeigen für beide Modellierungen sowohl verlaufsmäßige als auch von der Höhe der Abweichung der mittleren Prädiktion sehr große Ähnlichkeiten. Der größte Unterschied ist der Vergleich zwischen dem Niveau im unteren Wertebereich und dem oberen Wertebereich von $X_{Zweikampfquote}$. Im multinomialen logistischen Regressionsmodell hat sowohl der ICE-Plot als auch der Partial Dependence Plot eine sehr niedrige Wahrscheinlichkeitsvorhersage von fast 0% für Beobachtungen mit einem niedrigen Wert der *Zweikampfquote* und eine vergleichsweise hohe bis sehr hohe Wahrscheinlichkeit über den restlichen Wertebereich aufgezeigt. Der ALE-Plot hingegen zeigt nur im mit-

tleren Wertebereich der *Zweikampfquote* eine höhere Abweichung der mittleren Wahrscheinlichkeitsvorhersage an, während er an den beiden Rändern des Wertebereichs sehr niedrige Werte anzeigt.

Der *Random Forest* hingegen hat dieses Verhalten, dass nur im mittleren Wertebereich der *Zweikampfquote* eine hohe Wahrscheinlichkeit auf die *Außenverteidiger*-Position und im unteren und oberen Teil des Wertebereichs eine niedrigere Wahrscheinlichkeit modelliert wird, schon im ICE-Plot angedeutet. Der ICE-Plot zeigt jedoch im oberen Wertebereich der *Zweikampfquote* eine höhere Wahrscheinlichkeitsvorhersage an als im unteren Wertebereich, der ALE-Plot zeigt dies umgekehrt an.

Der Unterschied zwischen dem Effekt, der durch die ICE-Plots in Abbildung 37 angedeutet wird, und dem Effekt, der durch die ALE-Plots in Abbildung 40 gezeigt wird, ist am deutlichsten an den einzelnen Kurven des ICE-Plots für die multinomiale logistische Regression zu erkennen. Während auf einem großen Teil des Wertebereichs der *Zweikampfquote* Maxima der einzelnen Kurven existieren, zeigt der ALE-Plot ganz klar nur im mittleren Teil des Wertebereichs einen deutlichen Anstieg der Wahrscheinlichkeit auf die *Außenverteidiger*-Position an. Dies könnte dafür sprechen, dass die vielen Maxima im restlichen Teil des Wertebereichs durch Beobachtungen entstanden sind, die unwahrscheinliche Variablenkombinationen dort gebildet haben.

Obwohl fast alle ALE-Plots sehr ähnliche Effekte aufweisen, existieren auch ALE-Plots, in denen sich die beiden Modelle sehr stark voneinander unterscheiden.



(a) Multinomiales logistisches Regressionsmodell (b) *Random Forest*

Abbildung 41: ALE-Plots für die Wahrscheinlichkeit auf die *Offensive Mittelfeld*-Position bezüglich der *Anzahl an Zweikämpfen*

In Abbildung 41 sind die ALE-Plots für den Effekt der *Anzahl an geführten Zweikämpfen* auf die Wahrscheinlichkeitsvorhersage auf die *Offensive Mittelfeld*-Position in den beiden Mo-

dellen abgebildet. Während der ALE-Plot in beiden Modellen im unteren Wertebereich von $X_{Zweikämpfe}$ leicht ansteigt, steigt er im multinomialen logistischen Regressionsmodell danach stark an und fällt im oberen Wertebereich wieder stark ab. Im *Random Forest* sinkt der Accumulated Local Effect im mittleren Wertebereich stark und steigt im oberen Wertebereich wieder an.

An dieser Stelle sollte jedoch trotz dieses großen Unterschieds auf die y-Achse hingewiesen werden. Obwohl die *Anzahl an geführten Zweikämpfen* in beiden Modellen eine eher wichtige Variable ist, ändert sich die Wahrscheinlichkeitsvorhersage auf einen *offensiven Mittelfeldspieler* über den gesamten Wertebereich kaum.

Auch bei den weiteren Fällen, in denen sich die ALE-Plots stark voneinander unterscheiden, ist dies zu beobachten. Vor allem ist dies für die *Torhüter* auffällig, jedoch wurde diesbezüglich bereits in Kapitel 5.3 darauf hingewiesen, dass mehrere Variablen für die *Torhüter* ausschließlich aus *0*en bestehen, weshalb diese Effekte nicht überinterpretiert werden sollten.

5.8 Erarbeitung der Topologie der Modelle

Die in Kapitel 4.2.2.5 vorgeschlagene Methode zur Erarbeitung der Topologie der Modelle kann für alle Klassen gleichzeitig erfolgen. Für jede Richtung (Erhöhen und Verringern) werden jedoch 2 verschiedene "Stärken" der Datenmanipulation verwendet, wodurch für 11 Variablen, 2 Richtungen, 2 verschiedene Stärken und die 2 Modelle insgesamt 88 Plots entstehen.

Grundsätzlich soll diese Methode helfen nahe beeinander liegende Klassen zu ermitteln und darüber hinaus herauszufinden, bezüglich welcher Variable diese Nachbarschaft besteht. Dies könnte zum Beispiel helfen folgende Problemstellungen zu lösen:

1. Ein Spieler hat die letzten Jahre mit variierenden, aber immer passenden, Leistungsdaten für seine Position als *Außenverteidiger* in der Bundesliga gespielt. Leider hatte er sich verletzt und ist zusätzlich aufgrund seines Alters nicht mehr in der Lage so viel zu laufen, wie die Jahre zu vor. Auf welcher Position könnte ein *Außenverteidiger*, von dem davon ausgegangen wird, dass seine anderen Fähigkeiten (*Passgenauigkeit*, *Zweikampfquote*, etc.) konstant bleiben, mit einer **niedrigeren Laufleistung** eingesetzt werden?

2. Ein junger Spieler, der bisher die Position eines *zentralen Mittelfeldspielers* bekleidet hat, trainiert in der Sommerpause viele Zweikämpfe, wodurch seine Fähigkeit Zweikämpfe zu gewinnen (**höhere Zweikampfquote**) steigt. Auf welcher Position könnte er durch diese neuen Fähigkeiten flexibel einen verletzten Spieler ersetzen, wenn davon ausgegangen wird, dass seine anderen Leistungen konstant bleiben?

Um die erste der beiden Fragen zu beantworten, muss der Chordgraph für die *Laufweite* angeschaut werden. In Abbildung 42 sind auf der linken Seite zwei Chordgraphen für eine Verringerung der *Laufweite* und auf der rechten Seite für eine Erhöhung der *Laufweite* für das multinomiale logistische Regressionsmodell abgebildet. Die obere Zeile gibt dabei eine

schwache Datenmanipulation (Änderung der Werte um 0.05-Quantile) und die untere Zeile eine etwas stärkere Datenmanipulation (Änderung der Werte um 0.1-Quantile) an.

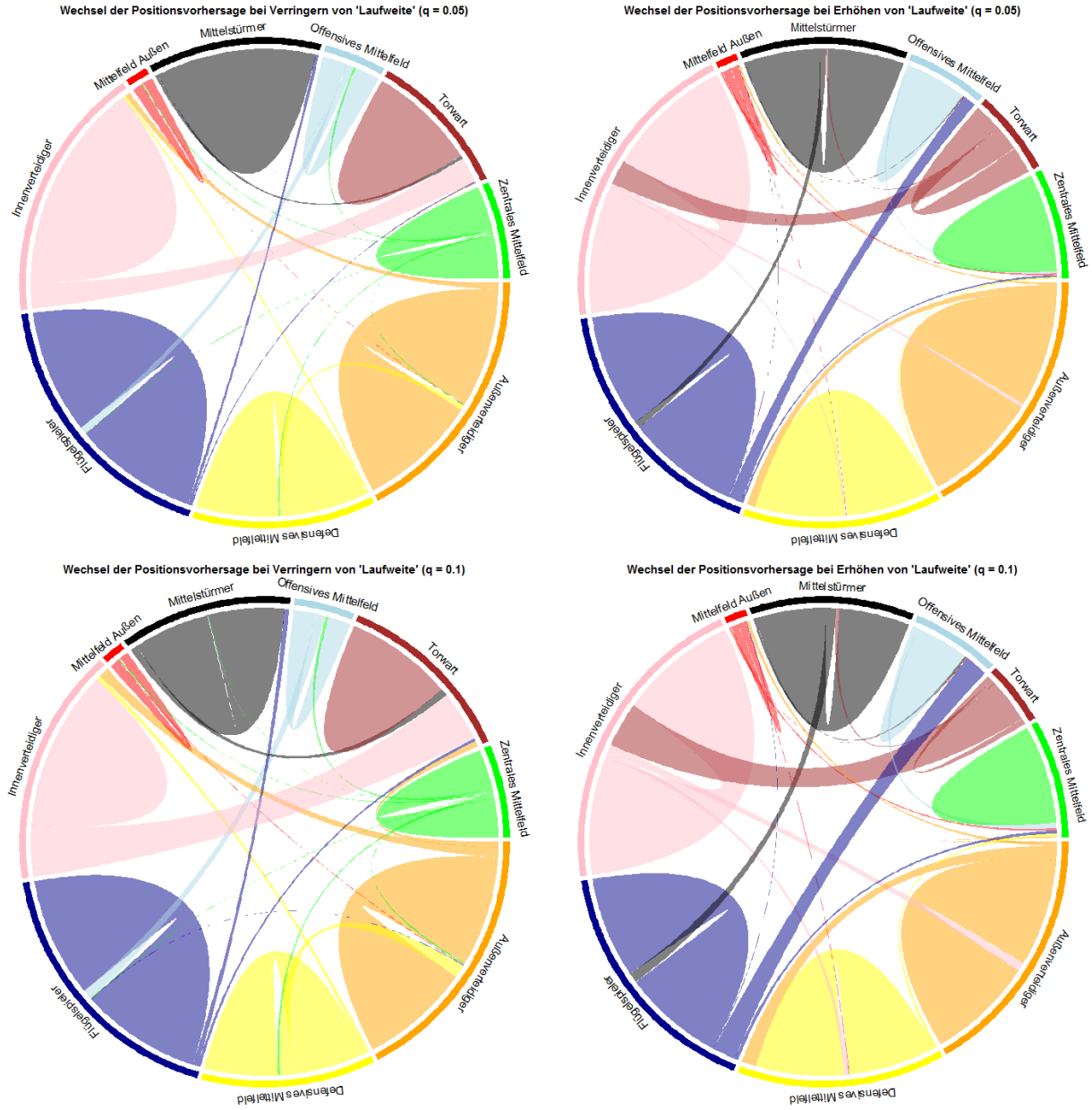


Abbildung 42: Chordgraph für die Nachbarschaftsverhältnisse bezüglich der *Laufweite* im multinomialen logistischen Regressionsmodell

Für die erste Frage sind vor allem die beiden linken Chordgraphen interessant. Ein Spieler, der zuvor *Außenverteidiger* war und auch diese Leistungsdaten passend für die Position erbracht hat und weiter erbringen wird, kann nur eine geringe *Laufweite* für die neue Saison aufs Spielfeld bringen. In dem linken oberen Chordgraphen, ist aufgeführt, dass ein Teil der durch das Modell als *Außenverteidiger* modellierten Beobachtungen bei Konstanthaltung

der Kovariablen und Verringerung der *Laufweite* in die Klasse der *Innenverteidiger* wechseln. Auch bei stärkerer Verringerung der Laufweite, existieren aus der Klasse der *Außenverteidiger* nur Wechsel in die Klasse der *Innenverteidiger*. Dies bedeutet, dass es möglich wäre, den Spieler auf der Position des *Innenverteidigers* auszuprobieren, da er dort gut in die Bundesliga passen könnte.

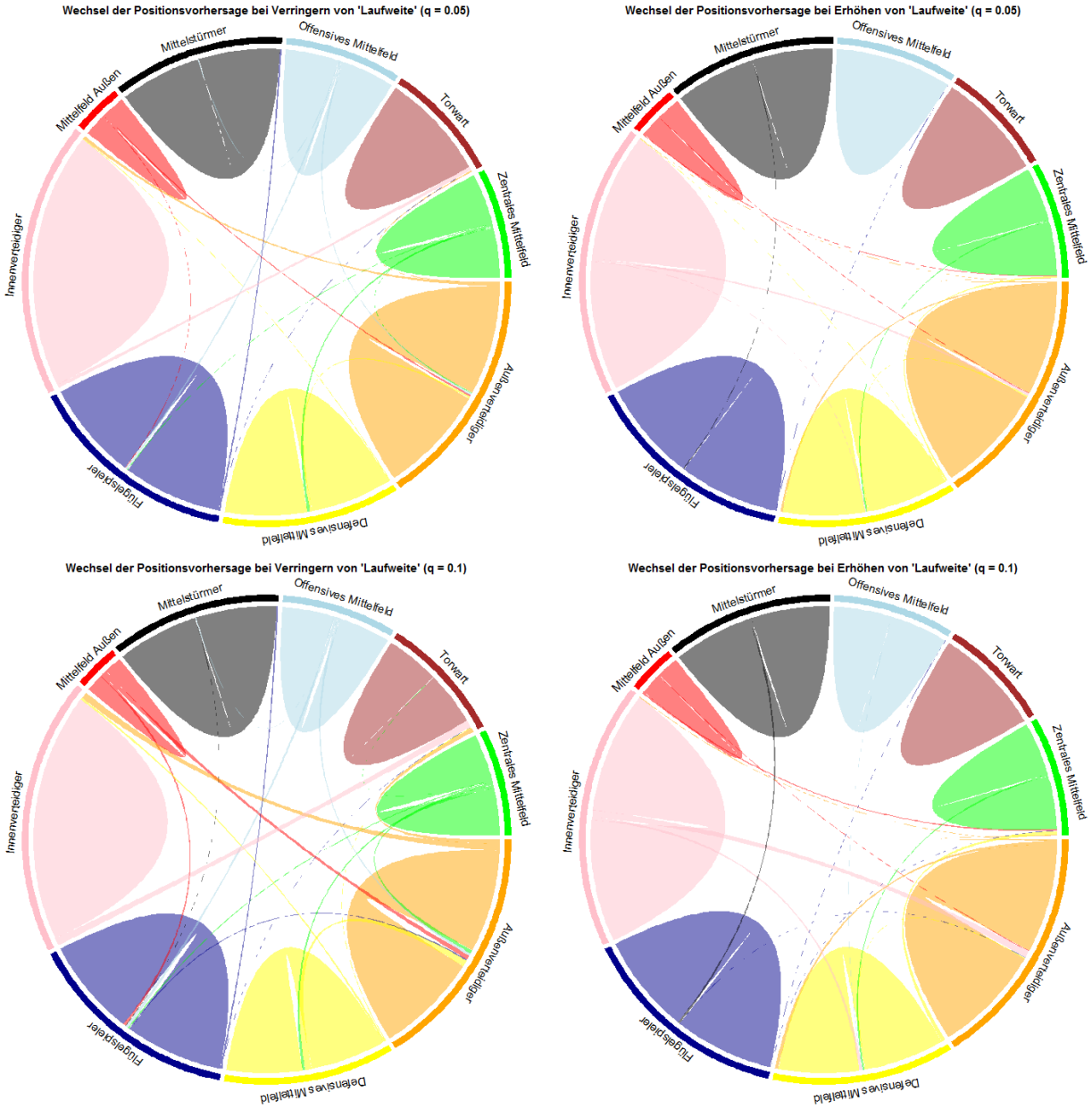


Abbildung 43: Chordgraph für die Nachbarschaftsverhältnisse bezüglich der *Laufweite* im *Random Forest*

Um zu überprüfen, ob durch die Datenmanipulation ein Bereich erreicht wird, in dem sonst keine Beobachtungen existieren, kann als Gegenprobe überprüft werden, ob auch Innen-

verteidiger durch **Erhöhen der Laufweite** in die Klasse der *Außenverteidiger* wechseln. Diese Gegenprobe ist in den beiden rechten Chordgraphen in Abbildung 42 dargestellt. Dort ist zu erkennen, dass es Beobachtungen gibt, die zunächst der Klasse der *Innenverteidiger* zugehören und durch Datenmanipulation in die Klasse der *Außenverteidiger* wechseln.

Hier sollte auffallen, dass dieses Verfahren für schlechte Modelle problematisch werden kann. Angenommen die Punkte sind komplett zufällig im Raum verteilt, dann würde es zwar ein Gebiet geben, indem *Außenverteidiger* vor der Datenmanipulation durch das Modell prognostiziert wird, jedoch könnte es passieren, dass in diesem Gebiet überhaupt keine Beobachtungen für die *Außenverteidiger*-Position existieren (vgl. Modellierungsproblem aus Abbildung 29).

Der *Random Forest* kann genutzt werden, um dieses modellierte Nachbarschaftsverhältnis zu überprüfen. In dem linken unteren Chordgraphen aus Abbildung 43 ist deutlich zu erkennen, dass auch der *Random Forest* einen Teil der als *Außenverteidiger* prognostizierten Beobachtungen bei Verringerung der *Laufweite* als *Innenverteidiger* prognostiziert. Auch in der Gegenprobe zeigt der Chrodgraph auf, dass Beobachtungen von der Prognose als *Innenverteidiger* durch Erhöhen der *Laufweite* als *Außenverteidiger* prognostiziert werden.

Wenn die beiden Modelle durch diese Chordgraphen ohne eine Fragestellung näher miteinander verglichen werden, so fällt sehr schnell das durch das multinomiale logistische Regressionsmodell aufgezeigte Nachbarschaftsverhältnis zwischen den *Torhütern* und *Innenverteidigern* auf, während dieses in den Chordgraphen für den *Random Forest* überhaupt nicht existiert. Dies könnte darauf hinweisen, dass durch die doch grundlegend verschiedene Einteilung des Raumes durch die beiden Modelle Nachbarschaftsverhältnisse aufgezeigt werden könnten, die in Wahrheit nicht existieren, oder die nur durch die Beschaffenheit des Modells existieren. Aus diesem Grund ist diese Methode auch nur dafür geeignet die Topologie in einem Modell und nicht die Topologie der Daten zu bestimmen. Für ein perfekt klassifizierendes Modell wäre dies äquivalent, da die Einteilung des Raumes durch das Modell auch approximativ der Einteilung des Raumes der Daten entspricht. Auch hier ist also zu erkennen, dass diese Methode besser funktioniert, wenn das Modell die vorliegende Datensituation besser klassifiziert. Genauso ratsam ist es diese Nachbarschaftsverhältnisse durch grundlegend verschiedene, aber gut klassifizierende Modelle zu vergleichen.

Wird ein Nachbarschaftsverhältnis von einem Modell erkannt, von einem anderen aber nicht, so empfiehlt es sich die Daten dort genauer anzuschauen. Einerseits könnte dies aufgrund von Interaktionseffekten vorkommen, die trotzdem durch Manipulieren einer einzelnen Variable einen Effekt auf die Prädiktion ausüben. Andererseits kann dies aber auch an der Natur eines Modells liegen, das "gröber" oder "feiner" die Daten einteilt und daher "Unreinheiten" zu diesen Unterschieden führen.

Die zweite Frage kann ebenso durch die potentiellen Nachbarschaftsverhältnisse beantwortet werden. In Abbildung 44 sind die 4 Chordgraphen für die Nachbarschaftsverhältnisse bezüglich der *Zweikampfquote* im multinomialen logistischen Regressionsmodell abgebildet. Für einen Spieler, dessen *Zweikampfquote* sich leicht erhöht und seine anderen Fähigkeiten konstant bleiben, existiert ein modelliertes Gebiet für *zentrale Mittelfeldspieler* unterhalb eines modellierten Gebietes für *defensive Mittelfeldspieler*. Bei ganz genauer Betrachtung existieren auch noch sehr schwache Nachbarschaftsverhältnisse zwischen den *zentralen*

Mittelfeldspielern und den bezüglich der *Zweikampfquote* darüber liegenden *äußeren Mittelfeldspielern* und *Außenverteidigern*. Bei der Gegenprobe durch das Verringern der *Zweikampfquote* bestätigen sich jedoch nur die Nachbarschaftsverhältnisse zu den *defensiven Mittelfeldspielern* und den *äußeren Mittelfeldspielern*.

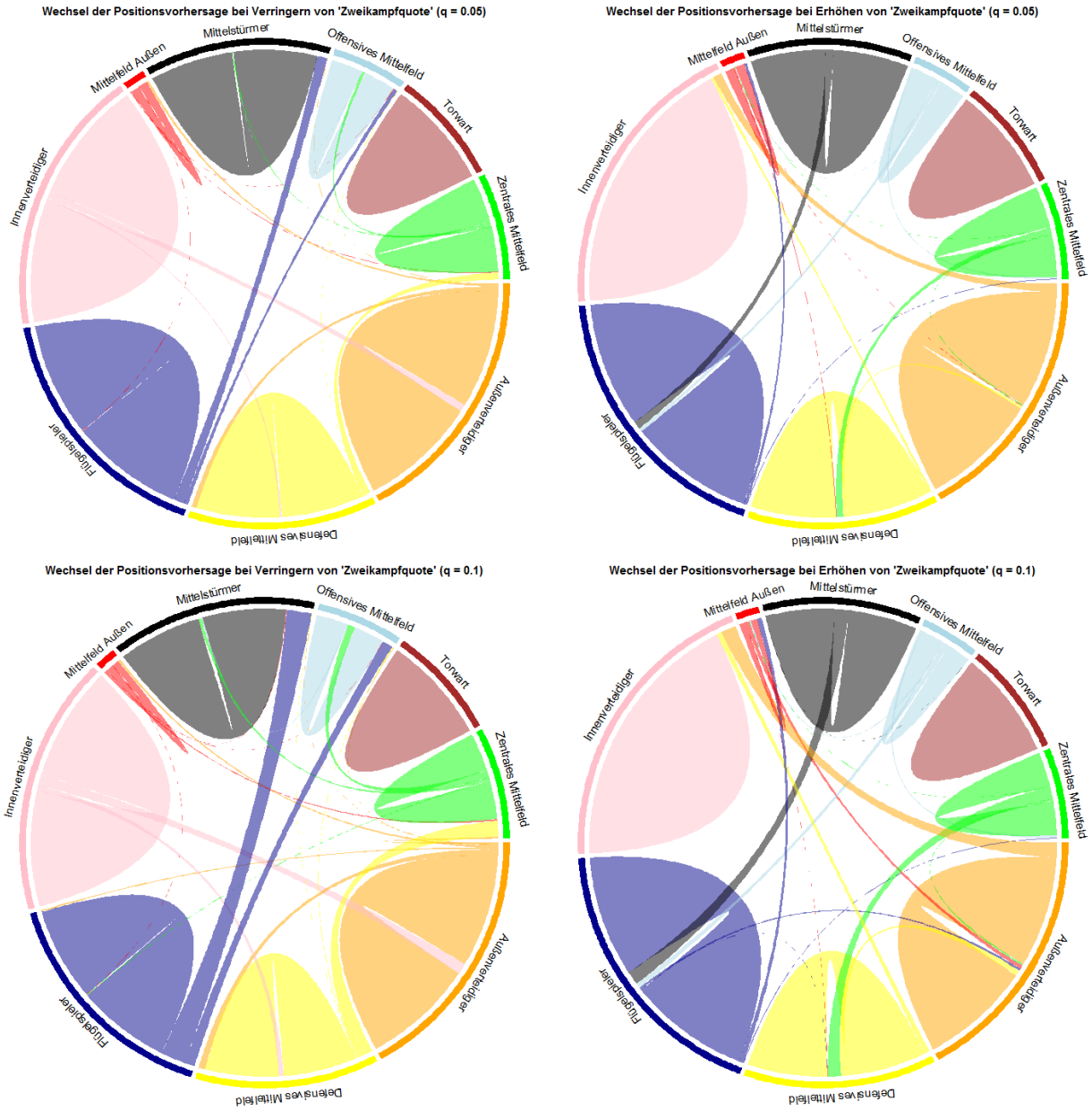


Abbildung 44: Chordgraph für die Nachbarschaftsverhältnisse bezüglich der *Zweikampfquote* im multinomialen logistischen Regressionsmodell

Dies bedeutet, dass es für den Spieler aus Fragestellung 2. möglich sein könnte einen *defensiven Mittelfeldspieler* oder einen *äußeren Mittelfeldspieler* auf Bundesliganiveau zu ersetzen. Das nicht-Bestätigen des Nachbarschaftsverhältns der *zentralen Mittelfeldspieler* und

der *Außenverteidiger* kann so gedeutet werden, dass das Modell zwar 2 benachbarte Gebiete modelliert, an dessen Grenze Beobachtungen im Gebiet der *zentralen Mittelfeldspieler* liegen, jedoch auf der anderen Seite keine Beobachtungen auf der Seite der *Außenverteidiger*. Dort könnte also eine unmögliche oder unwahrscheinliche Datensituation vorliegen.

Um die beiden gefundenen Nachbarschaftsverhältnisse zu bestätigen, können wieder die Chordgraphen für den *Random Forest* betrachtet werden.

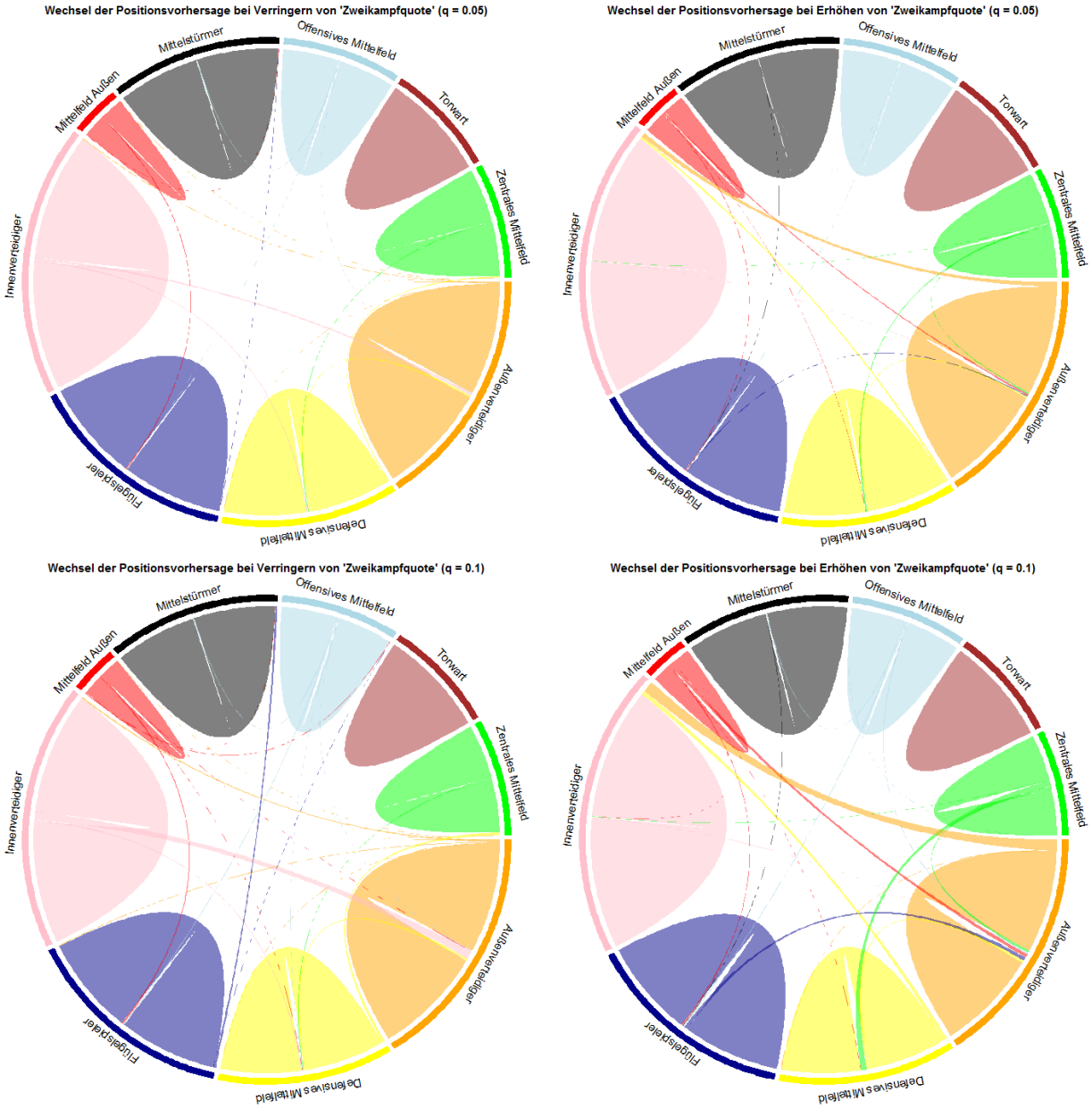


Abbildung 45: Chordgraph für die Nachbarschaftsverhältnisse bezüglich der *Zweikampfquote* im *Random Forest*

In Abbildung 45 wird das Nachbarschaftsverhältnis im *Random Forest* bezüglich der

Zweikampfquote zwischen der Prädiktion als *zentraler Mittelfeldspieler* und der Prädiktion als *defensiver Mittelfeldspieler* durch beide Richtungen der Datenmanipulation bestätigt. Ein Nachbarschaftsverhältnis zwischen dem *zentralen Mittelfeld* und *äußeren Mittelfeld* wird nur gegenläufig angezeigt, sodass ein modelliertes Gebiet mit *zentralen Mittelfeldspielern* oberhalb eines modellierten Gebietes mit *äußeren Mittelfeldspielern* liegt. Ein Nachbarschaftsverhältnis zwischen den *zentralen Mittelfeldspielern* und den *Außenverteidigern* wird, wie im Chordgraph für das multinomiale logistische Regressionsmodell, bei Erhöhen der *Zweikampfquote* in diegleiche Richtung angezeigt, jedoch wird auch dieses hier nicht durch die Gegenprobe bestätigt.

Zusammengefasst ist zu sagen, dass der junge Spieler als guter Ersatz für einen *defensiven Mittelfeldspieler* gelten kann. Das modellierte Gebiet für die *äußeren Mittelfeldspieler* könnte durch Unreinheiten, Ausreißer oder Inseln entstanden sein, da dieses Nachbarschaftsverhältnis durch die beiden Modelle sogar gegenläufig modelliert wird. Da beide Modelle das "einseitige" und sehr schwache Nachbarschaftsverhältnis zu den *Außenverteidigern* modellieren, könnte dies entuell auch durch eine kleine Gruppe Ausreißer, die dort ein Gebiet suggerieren, das nicht dem typischen *zentralen Mittelfeldspieler* entspricht, entstanden worden sein. Ein Einsatz auf dieser Position könnte also funktionieren, ist jedoch mit Vorsicht zu raten.

Praktisch gesehen ergibt sich für diese Frage ein neuer Gedankengang: "Durch hartes Training der Zweikampfquote erhöht sich nicht nur diese, sondern auch die Bereitschaft Zweikämpfe zu führen!" (Theoretisch ausgedrückt: Durch das Erhöhen einer Variable erhöhen sich auch andere). Das Messen der Nachbarschaftsverhältnisse in der Modellierung wurde hier nur bezüglich einzelner Variablen durchgeführt. Existiert jedoch ein hohes Ahängigkeitsverhältnis zu anderen Variablen, so müssten diese gemeinsam erhöht werden. Dadurch würden sich ganz neue Nachbarschaftsverhältnisse ergeben. Dies wird in dieser Arbeit hier nicht weiter überprüft, eröffnet aber einen noch weiteren Anwendungsraum für diese Methode.

6 Fazit

Das Ziel dieser Arbeit war es Zusammenhänge zwischen den Positionen der Bundesligaspieler und ihrer Leistungsdaten mit verschiedenen Methoden zu erkennen und aufzuzeigen. Für die Analysen wurde sowohl ein multinomiales logistisches Regressionsmodell als auch ein *Random Forest* verwendet und anschließend mit Methoden aus dem Bereich des interpretierbaren Machine Learnings verglichen.

Obwohl die beiden Modelle an sich sehr unterschiedlich aufgebaut sind, konnte vor allem durch die Methode der Accumulated Local Effect Plots gezeigt werden, dass sie im Grunde die Daten sehr ähnlich modellieren und sich nur geringfügig voneinander unterscheiden. Anhand dieser Methode konnten Effekte ausgearbeitet werden, die deutlich für eine bestimmte Position sprechen, wie zum Beispiel eine niedrige *Laufweite* für einen *Torhüter*. Die Partial Dependence Plots und die Individual Conditional Expectation Plots haben noch deutlichere Unterschiede zwischen den beiden Modellen aufgezeigt. Im Gegensatz zu den Accumulated Local Effect Plots werden diese Methoden jedoch mit Datenmanipulationen erzeugt, die häufig sehr unwahrscheinliche Datenkombinationen kreieren.

Des Weiteren wurden die Fehlklassifikationen der beiden Modelle betrachtet und entdeckt, dass die meisten Fehlklassifikationen in Positionen geschehen, die auf dem Fußballplatz nebeneinander liegen. Beim Testen der Methode zum Erarbeiten der Topologie der Modelle ist aufgefallen, dass häufig Positionen, die auch auf dem Fußballfeld benachbart sind auch durch die Modelle im Datenraum benachbart modelliert werden. Zusätzlich wurden neue Fragestellungen formuliert, die mit dieser Methode beantwortet werden konnten. Auch für diese Methode ähneln sich die beiden Modellierungen sehr.

Da zwar eindeutige Zusammenhänge zwischen den Positionen und den Leistungsdaten gefunden wurden, diese aber nicht auf ihre Kausalität getestet wurden, bleibt die Frage offen, ob die Position die Leistungsdaten, die von einem Spieler auf dieser Position erzeugt werden, beeinflusst, oder ob ein Spieler mit einer bestimmten Fähigkeit auf einer gewissen Position eingesetzt wird. Dies könnte noch in einer weiterführenden Analyse untersucht werden. Eine weitere Möglichkeit wäre es noch die aufgezeigten Zusammenhänge auf ihre Zeitlosigkeit zu untersuchen. Dabei könnte überprüft werden, ob sich diese Effekte über die Jahre verändern oder konstant bleiben.

Alles in allem wurde in dieser Arbeit gezeigt, dass es eindeutige Zusammenhänge zwischen den Leistungsdaten der Bundesligaspieler und ihrer Positionen gibt, und dass diese mithilfe von verschiedensten Modellen und Methoden ermittelt werden können. Für diese Analysen wurde nach subjektiven Kriterien eine Variablenselektion durchgeführt. Die Ergebnisse wären durch Hinzunahme neuer Leistungsdaten zwar komplexer zu interpretieren; je nach Interesse können für weitere Analysen jedoch Variablen hinzugefügt oder weggelassen werden. Im Gegensatz zu multinomialen logistischen Regressionsmodellen haben *Random Forests* keine Probleme mit Multikollinearität, wodurch vor allem für *Random Forests* eine Hinzunahme von zusätzlichen Variablen für diese Analysen möglich wäre.

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

Abkürzung	Erklärung
AM	Äußere Mittelfeldposition
AV	Außenverteidiger
DM	Defensives Mittelfeld
FS	Flügelspieler
IV	Innenverteidiger
MS	Mittelstürmer
OM	Offensives Mittelfeld
TW	Torwart
ZM	Zentrales Mittelfeld

Quellen

Apley, Daniel W. 2016. “Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models.”

Breiman, Leo. 2001. *Machine Learning* 45 (1). Springer Nature: 5–32. doi:10.1023/a:1010933404324.

Friedman, Jerome H. 2001. *The Annals of Statistics* 29 (5). Institute of Mathematical Statistics: 1189–1232.

Goldstein, Alex, Adam Kapelner, Justin Bleich, and Emil Pitkin. 2013. “Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation.”

Liaw, Andy, and Matthew Wiener. 2002. “Classification and Regression by randomForest.” *RNews* 2: 18–22.

Nordmann, Nils. 2016. “Das angesagteste Statistikmodell im Profifußball.” https://www.welt.de/sport/fussball/article151870094/Das-angesagteste-Statistikmodell-im-Profifussball.html.

Preti, Antonio, and Marcello Vellante. 2007. “Creativity and Psychopathology.” *The Journal of Nervous and Mental Disease* 195 (10). Ovid Technologies (Wolters Kluwer Health): 837–45. doi:10.1097/nmd.0b013e3181568180.

Scholbeck, Christian A., Christoph Molnar, Christian Heumann, Bernd Bischl, and Giuseppe Casalicchio. 2019. “Sampling, Intervention, Prediction, Aggregation: A Generalized Framework for Model Agnostic Interpretations.”

“sport1.de.” 2018. https://www.sport1.de.

“transfermarkt.de.” 2018. https://www.transfermarkt.de.

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with S*. Fourth. New York: Springer. http://www.stats.ox.ac.uk/pub/MASS4.

“weltfussball.de.” 2018. https://www.weltfussball.de/spielerliste/bundesliga-2009-2010/nach-name/1/.

Wilson, Jonathan. 2010. “Why are so many wingers playing on the ’wrong’ wings?” https://www.theguardian.com/sport/blog/2010/mar/24/the-question-inside-out-wingers.

Wright, Marvin N., and Andreas Ziegler. 2017. “ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R.” *Journal of Statistical Software* 77 (1): 1–17. doi:10.18637/jss.v077.i01.

Anhang

Nicht ausgewählte Variablen

Variable	Beschreibung
Alter	Generell wurde sich gegen demografische Variablen entschieden, da Spieler nur an ihren Fähigkeiten gemessen werden sollen
Eigentore	Eigentore passieren einerseits viel zu selten und andererseits geschehen diese oft durch Zufälle, die nicht die Leistung eines Spielers abbilden
Schüsse	Die Anzahl der Schüsse ist höchstkorreliert mit der Anzahl an Toren. Die Anzahl an Toren misst die Fähigkeiten eines Spielers besser als die Anzahl der Schüsse
Schussvorlagen	Die Anzahl der Schussvorlagen ist höchstkorreliert mit der Anzahl an Torvorlagen. Die Anzahl an Torvorlagen misst die Fähigkeiten eines Spielers besser als die Anzahl der Schussvorlagen
Ballkontakte	Die Anzahl der Ballkontakte ist höchstkorreliert mit der Anzahl der gespielten Pässe. Diese messen die Fähigkeiten eins Spielers besser
Angekommene Pässe	Die Anzahl der angekommenen Pässe misst die Fähigkeiten eines Spielers am besten in Kombination mit der Anzahl der gespielten Pässe, was der Passquote entspricht, weshalb diese stattdessen aufgenommen wurde
Fehlpässe	Die Anzahl der Fehlpässe misst die Fähigkeiten eines Spielers am besten in Kombination mit der Anzahl der gespielten Pässe, was der Passquote entspricht, weshalb diese stattdessen aufgenommen wurde
Sprints	Die Anzahl der Sprints ist höchstkorreliert mit der Laufweite, welche die Fähigkeiten eines Spielers besser aufnimmt (konstante Messung gegen Anzahl langer oder kurzer Sprints)
Höchstgeschwindigkeit	Die Höchstgeschwindigkeit bildet nur eine einzelne Momentaufnahme in einem einzigen Spiel ab, was keine konstante Leistungserfassung abbildet
Tore	Die Anzahl der Tore ist höchstkorreliert mit der Anzahl der Tore mit dem Fuss. Die Tore mit dem Fuss wurden aufgenommen, da auch die Kopfballtore aufgenommen werden sollten und diese gemeinsam (abzüglich der Elfmetertore - seltenes Ereignis) die Anzahl der Tore angeben

Elfmetertore	Die Elfmetertore geben nicht wirklich die Fähigkeiten eines Spielers an, da nicht jeder die Chance hat Elfmetertore zu schießen
Verschossene Elfmeter	Da nicht jeder die Chance hat Elfmeter zu schießen hat auch nicht jeder die Chance einen Elfmeter zu verschießen
Gegentore	Diese Variable wurde nur für Torhüter erhoben und ist daher nicht repräsentativ für die Fähigkeiten der einzelnen Spieler
Gehaltene Bälle	Diese Variable wurde nur für Torhüter erhoben und ist daher nicht repräsentativ für die Fähigkeiten der einzelnen Spieler
Gehaltene Elfmeter	Diese Variable wurde nur für Torhüter erhoben und ist daher nicht repräsentativ für die Fähigkeiten der einzelnen Spieler

Tabelle 10: Nicht ausgewählte Variablen

Zusammengefasste Positionen

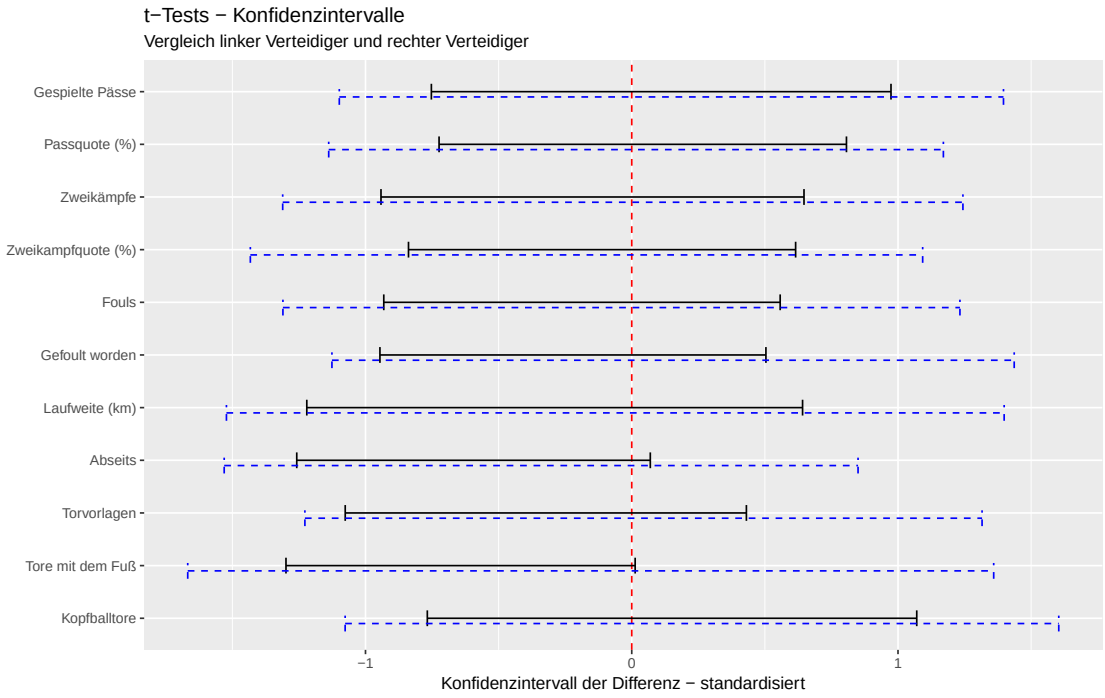


Abbildung 46: Vergleich zwischen *linken* und *rechten Verteidigern*

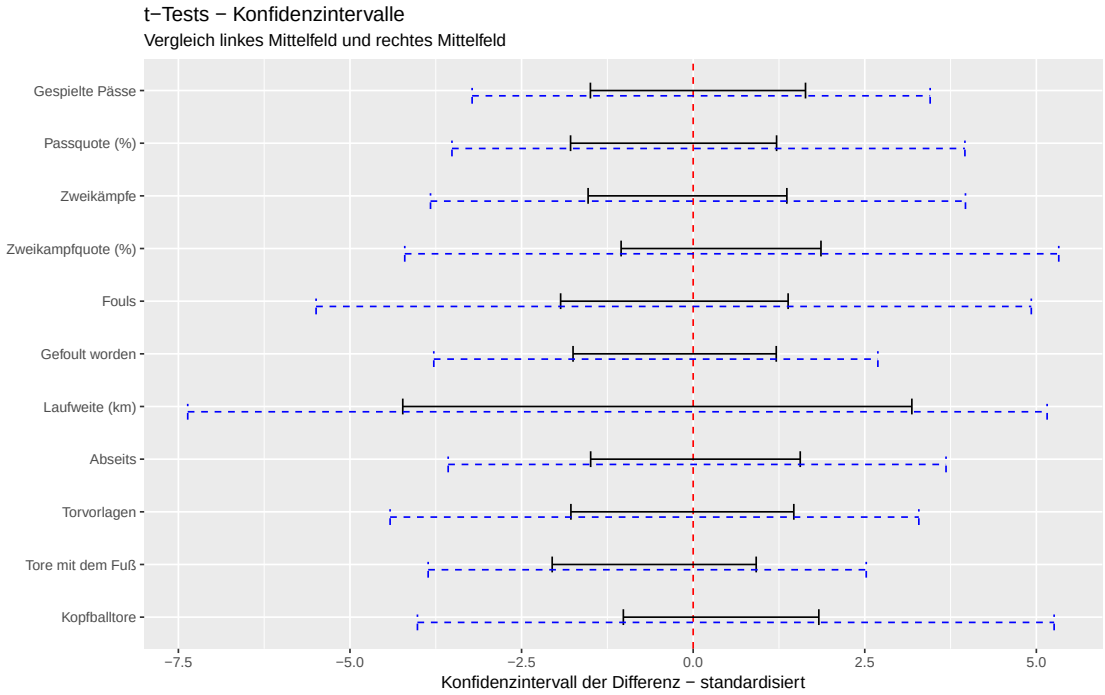


Abbildung 47: Vergleich zwischen *linken* und *rechten Mittelfeldspielern*

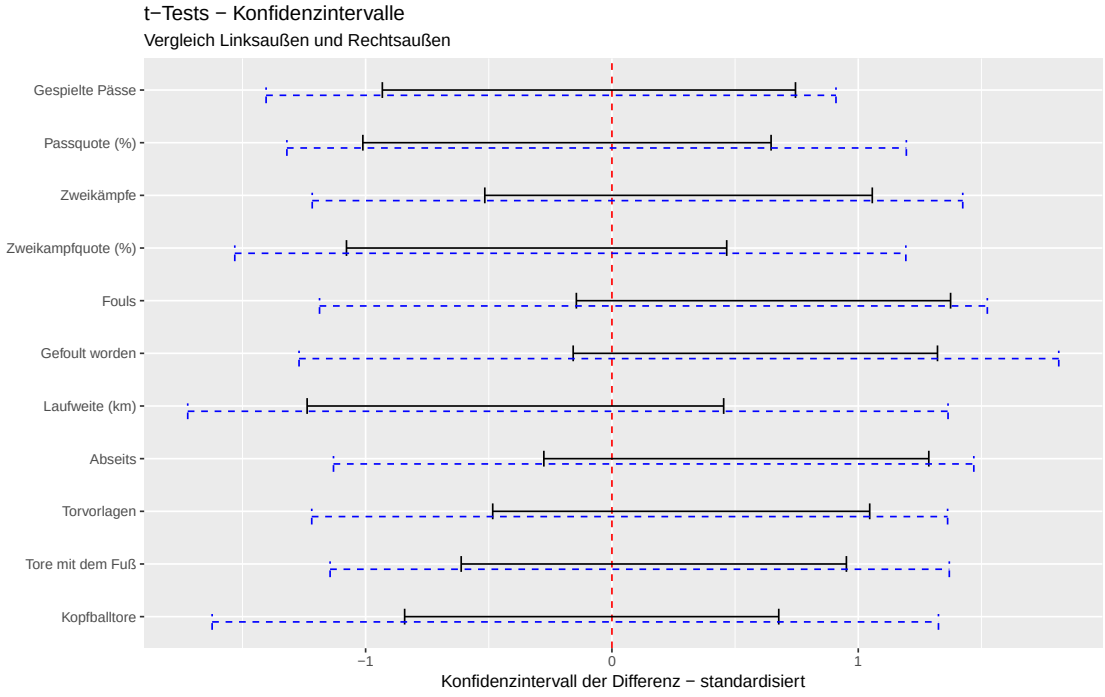


Abbildung 48: Vergleich zwischen *Linksaußen* und *Rechtsaußen*

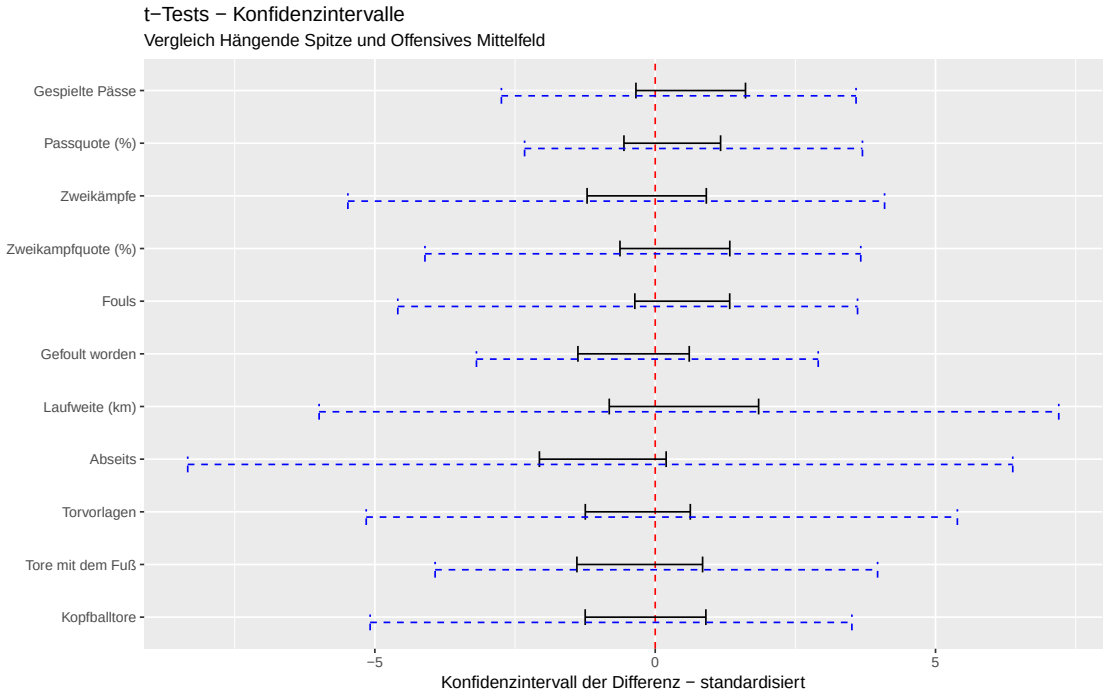


Abbildung 49: Vergleich zwischen *hängender Spitze* und *offensiven Mittelfeld*

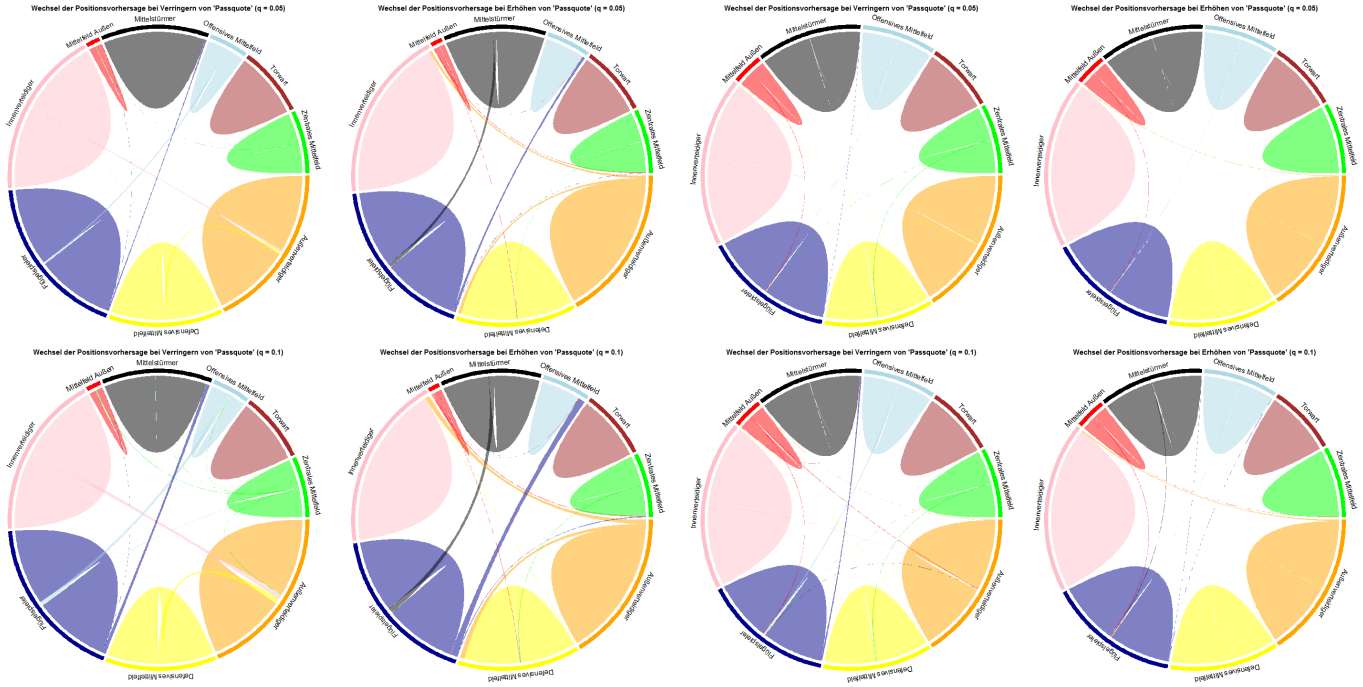
ICE-Plots

Zu viele für gedruckten Anhang - siehe elektronischer Anhang

ALE-Plots

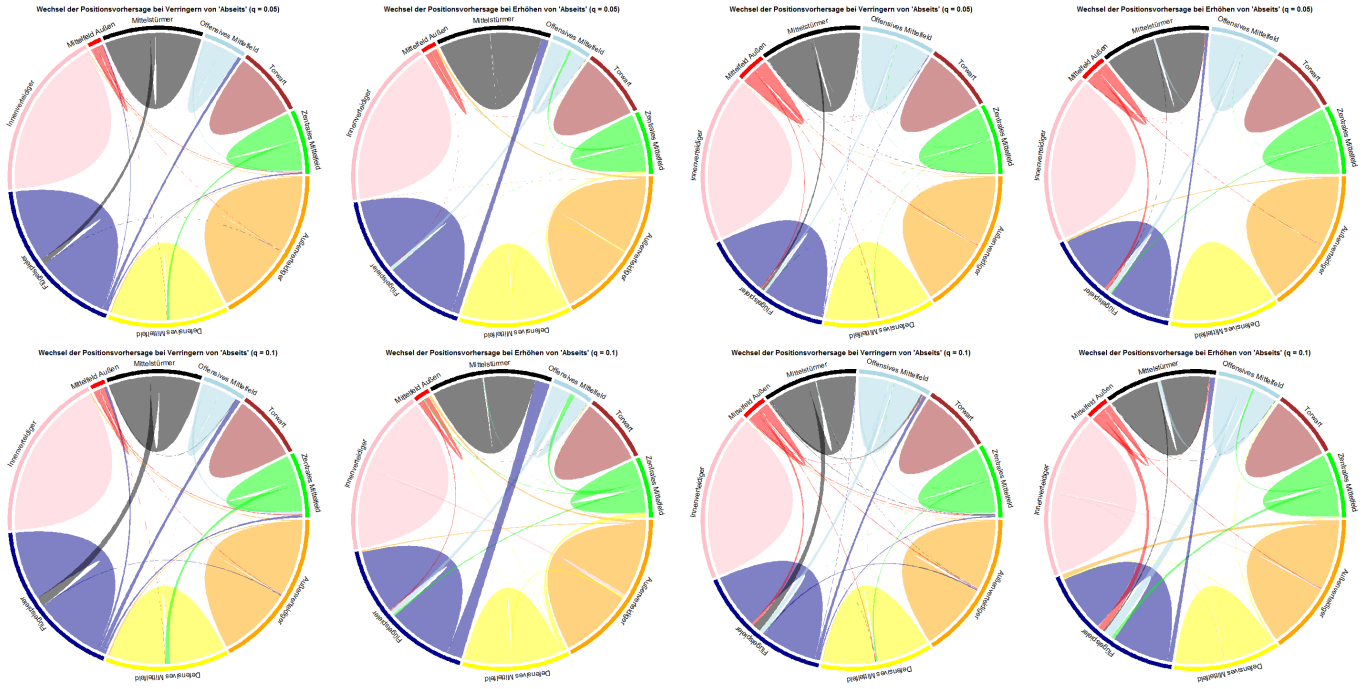
Zu viele für gedruckten Anhang - siehe elektronischer Anhang

Chordgraphen



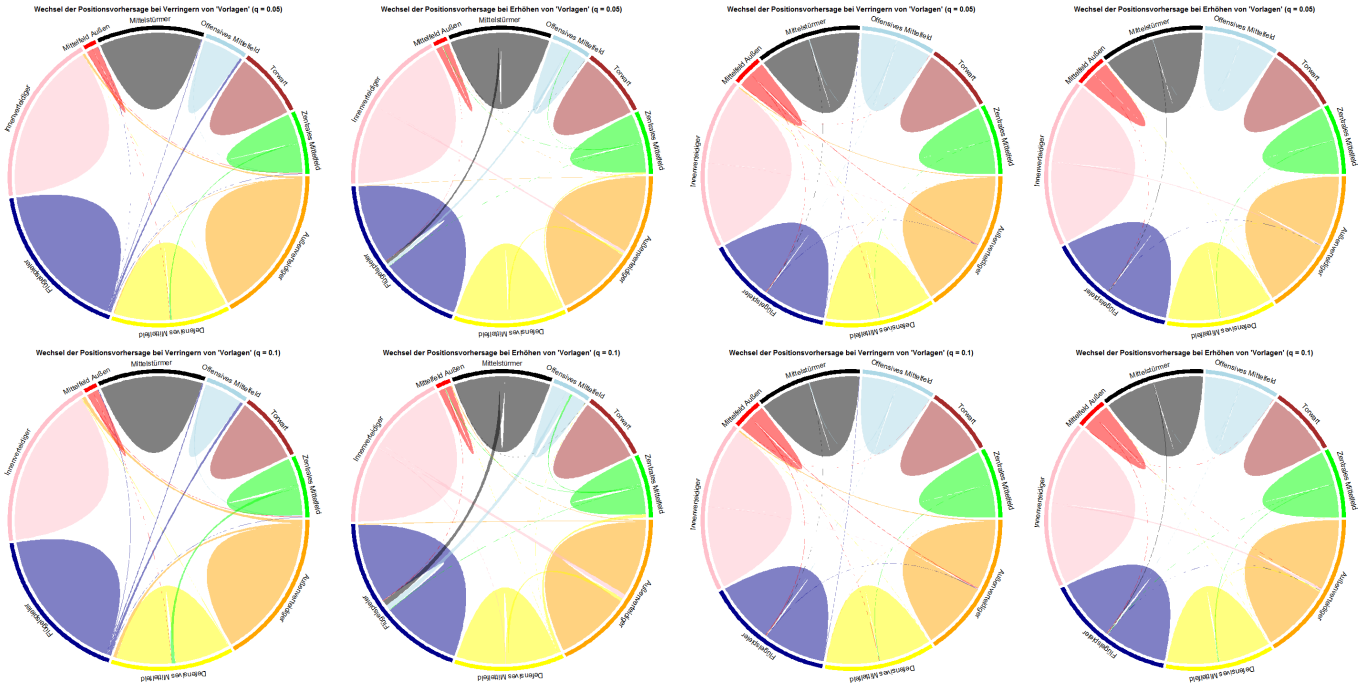
(a) Multinomiales logistisches Regressionsmodell (b) *Random Forest*

Abbildung 50: Chordgraphen für die erarbeiteten Nachbarschaften bezüglich der *Passquote*



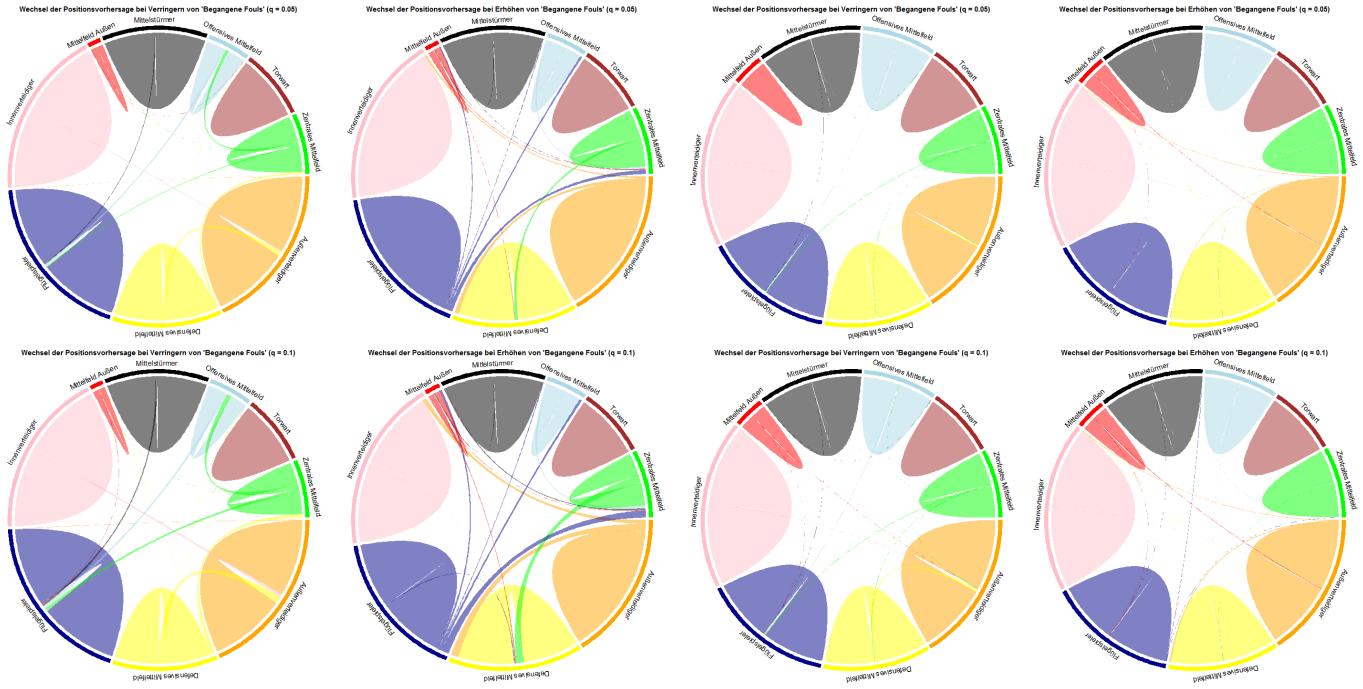
(a) Multinomiales logistisches Regressionsmodell (b) *Random Forest*

Abbildung 51: Chordgraphen für die erarbeiteten Nachbarschaften bezüglich der *Abseitsstellungen*



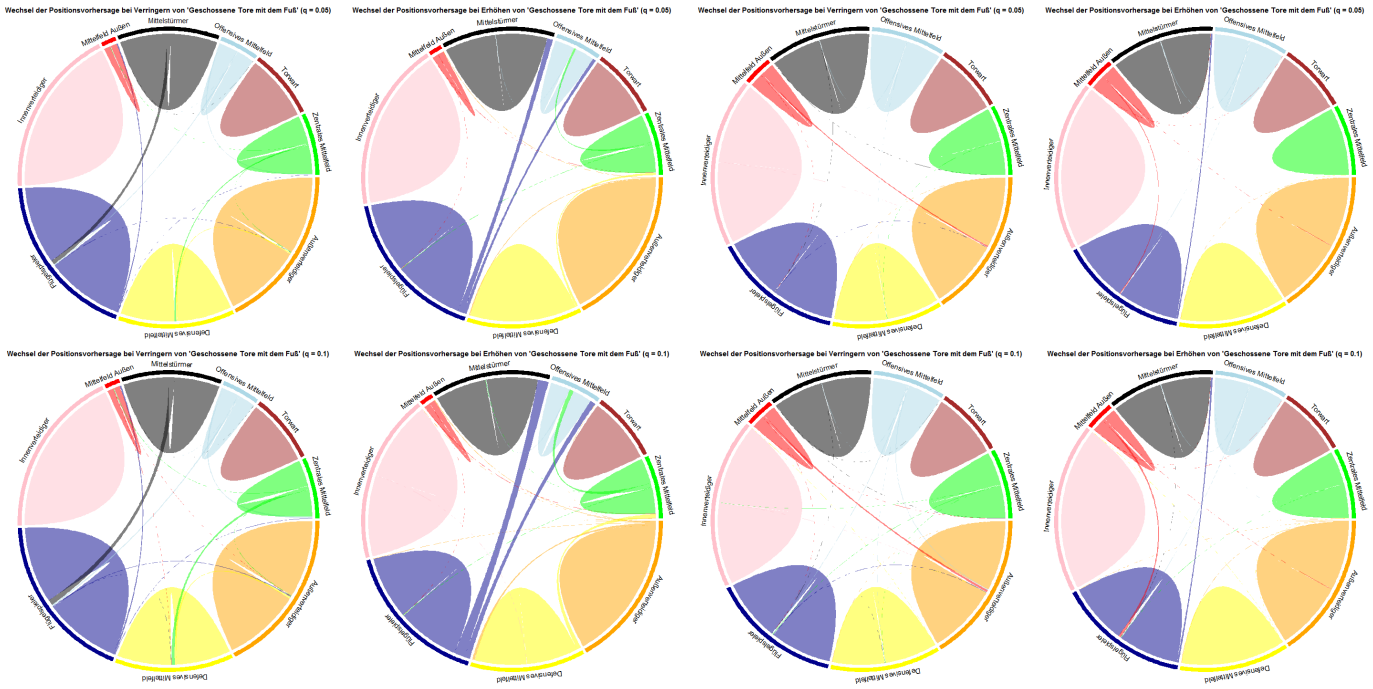
(a) Multinomiales logistisches Regressionsmodell (b) *Random Forest*

Abbildung 52: Chordgraphen für die erarbeiteten Nachbarschaften bezüglich der *Torvorlagen*



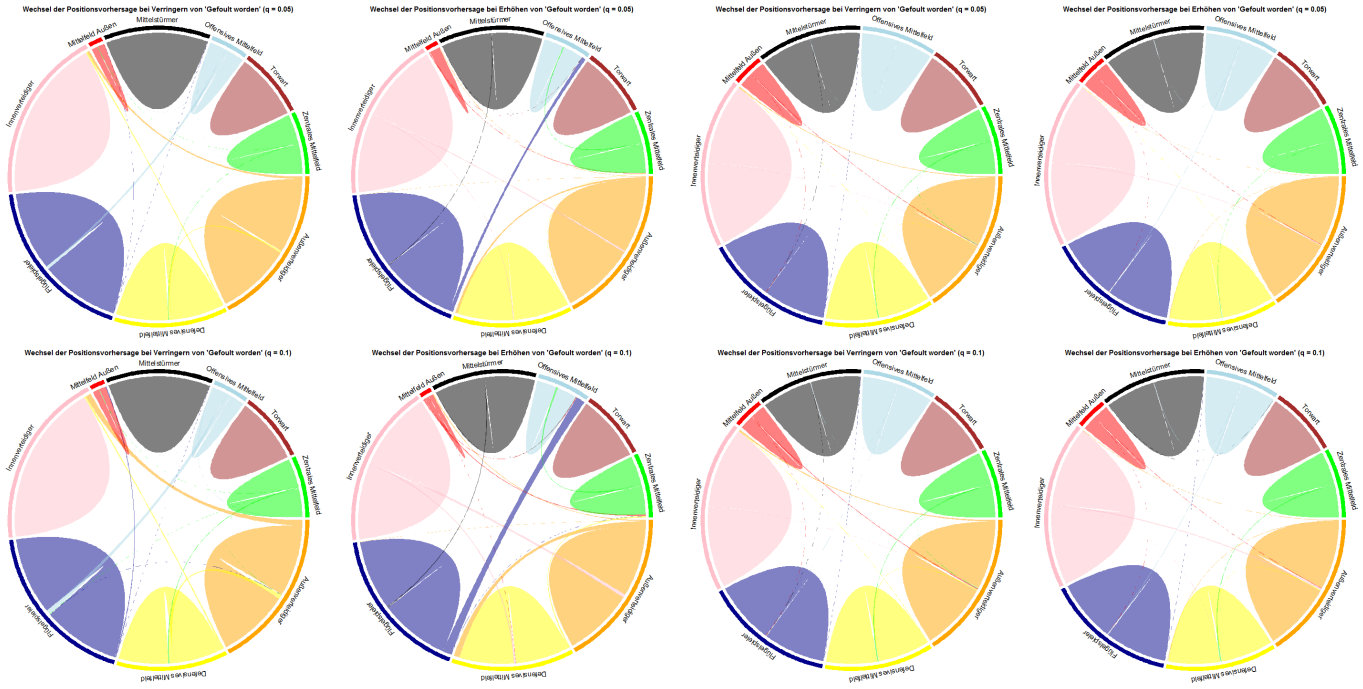
(a) Multinomiales logistisches Regressionsmodell (b) *Random Forest*

Abbildung 53: Chordgraphen für die erarbeiteten Nachbarschaften bezüglich der *Fouls*



(a) Multinomiales logistisches Regressionsmodell (b) *Random Forest*

Abbildung 54: Chordgraphen für die erarbeiteten Nachbarschaften bezüglich der *Tore mit dem Fuss*

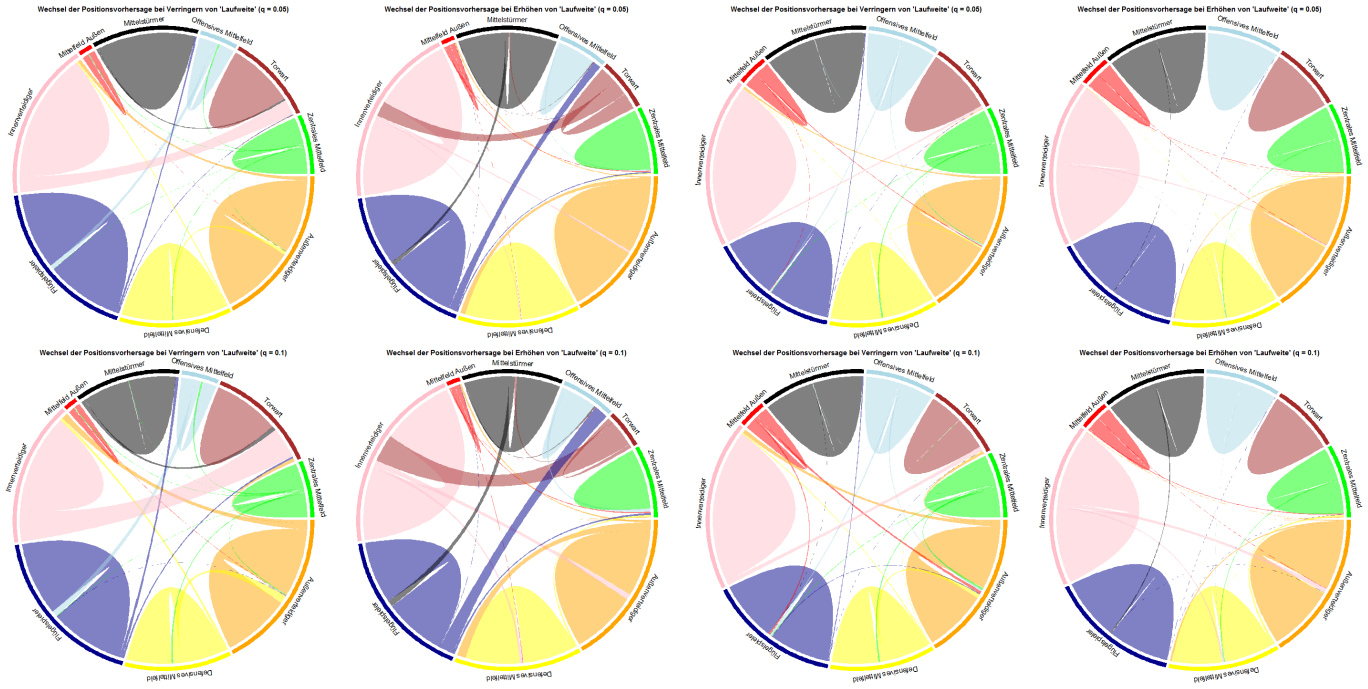


(a) Multinomiales logistisches Regressionsmodell (b) *Random Forest*

Abbildung 55: Chordgraphen für die erarbeiteten Nachbarschaften bezüglich der *Anzahl gefoult zu werden*

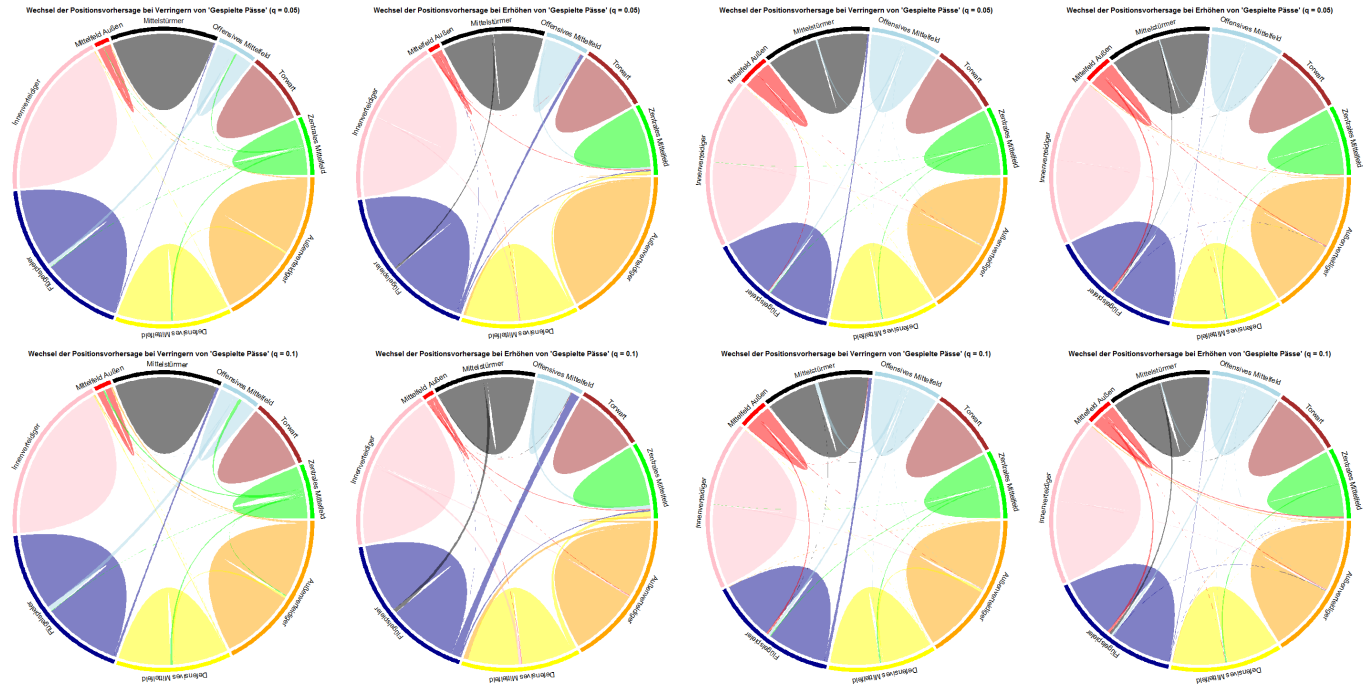
(a) Multinomiales logistisches Regressionsmodell (b) *Random Forest*

Abbildung 56: Chordgraphen für die erarbeiteten Nachbarschaften bezüglich der *Kopfballtore*



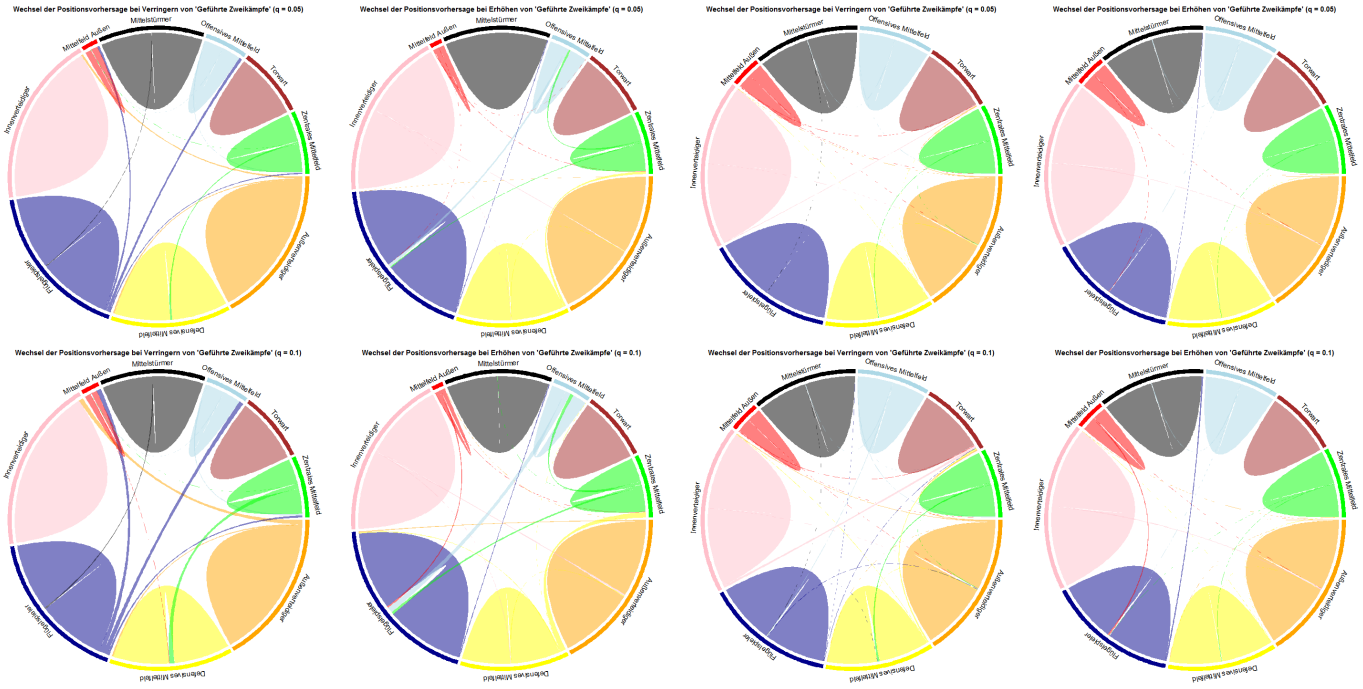
(a) Multinomiales logistisches Regressionsmodell (b) *Random Forest*

Abbildung 57: Chordgraphen für die erarbeiteten Nachbarschaften bezüglich der *Laufweite*



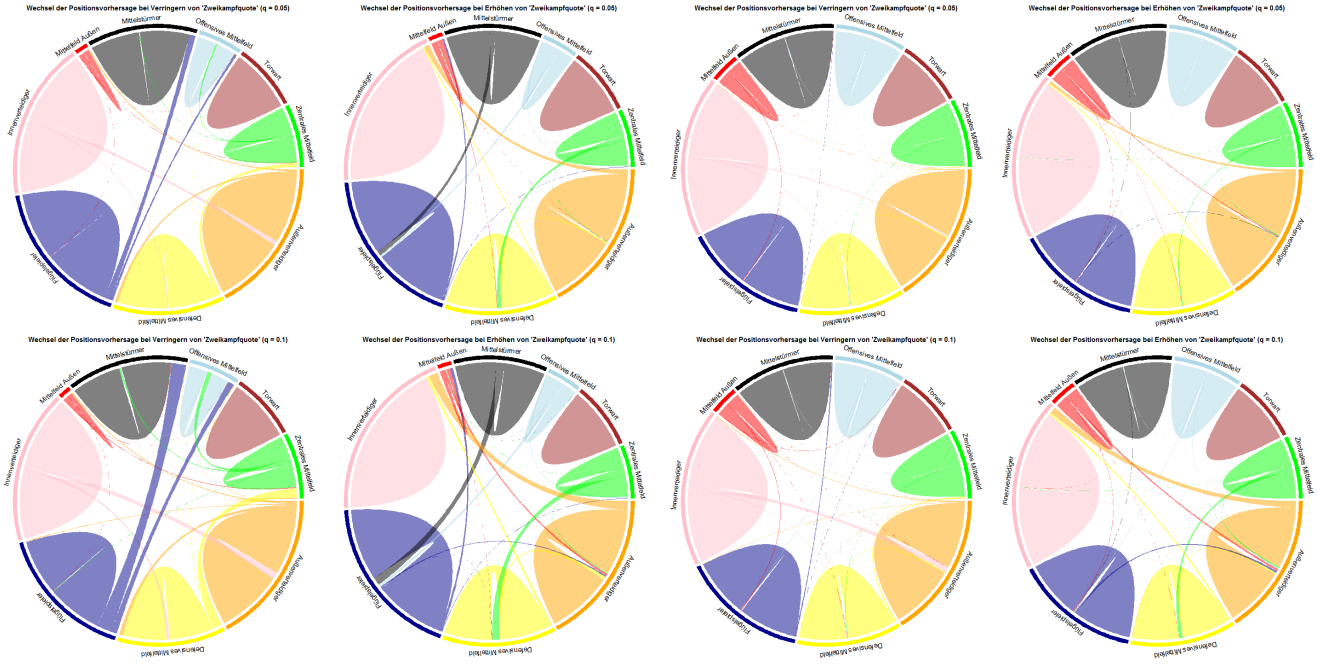
(a) Multinomiales logistisches Regressionsmodell (b) *Random Forest*

Abbildung 58: Chordgraphen für die erarbeiteten Nachbarschaften bezüglich der *Anzahl Pässe*



(a) Multinomiales logistisches Regressionsmodell (b) *Random Forest*

Abbildung 59: Chordgraphen für die erarbeiteten Nachbarschaften bezüglich der *Anzahl Zweikämpfe*



(a) Multinomiales logistisches Regressionsmodell (b) *Random Forest*

Abbildung 60: Chordgraphen für die erarbeiteten Nachbarschaften bezüglich der *Zweikampfquote*