Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einführung

Die vorliegende Simulationsstudie befasst sich mit dem Thema: Personal im Unternehmen halten. Dabei geht es um gewisse Muster in der Fluktuation des Unternehmens „Pay Solutions GmbH“. Die Pay Solutions GmbH zeichnet sich durch maßgeschneiderte Compliance-Lösungen in einer sich ständig wandelnden Geschäftsumgebung aus. Ihre Mission ist es, Unternehmen dabei zu unterstützen, die komplexen Anforderungen und Regulierungen im Finanzbereich zu verstehen und zu erfüllen, um einen reibungslosen Geschäftsbetrieb sicherzustellen.

Die Fluktuation kann in drei Arten gegliedert werden: natürliche Fluktuation, unternehmensinterne Fluktuation oder unternehmensexterne Fluktuation. Eine natürliche Fluktuation beschreibt altersbedingte Gründe wie Vorruhestand, Rente oder Altersteilzeit. Personalwechsel innerhalb des Unternehmens, beispielsweise durch Beförderungen, fallen in die Rubrik unternehmensinterne Fluktuation. Wohingegen eine aktive Kündigung von Seiten der angestellten Person als unternehmensexterne Fluktuation angesehen wird [1].
Vorliegende Simulationsstudie betrachtet lediglich die unternehmensexterne Fluktuation.

Um die Simulationsstudie umzusetzen werden zunächst in Kapitel 2 Variablen für die Grundgesamtheit der Studie erstellt. Anschließend wird eine Stichprobe dieser in Kapitel 3 näher analysiert. Abschließend wird in Kapitel 4 die optimale Modellflexibilität untersucht.

2 Erzeugung einer Grundgesamtheit

Das folgende Kapital beinhaltet die Erzeugung der Grundgesamtheit mit N = 1.000.000 Elementen, welche die Variablen zum Thema 5 „Personal im Unternehmen halten“ enthält. Zunächst werden sechs potenziell erklärenden Variablen Alter, Betriebszugehörigkeit, Bildungsabschluss, Gehalt, Zeit seit Gehaltserhöhung und Work-Life-Balance erstellt. Auf Basis dieser wird im Anschluss die Zielvariable Abwanderung generiert.

2.1 Alter

Die Verteilung der Variable Alter basiert auf einer Statistik des ifo Instituts bei welcher die Altersstruktur in Unternehmen in Deutschland nach Branchen und Anzahl der Beschäftigten untersucht wurde [2]. Da das durchschnittliche Renteneintrittsalter bei ca. 65 Jahren [3] und der oft frühestmögliche Berufseinstieg im Alter von 16 Jahren stattfindet [4], wurden diese beiden Werte als Altersgrenzen für die Simulationsstudie festgelegt. Zudem liegt das Durchschnittsalter der Beschäftigten in Baden-Württemberg zwischen 40,9 und 43,2 Jahren, weshalb das Durchschnittsalter von 43 Jahren für die Altersverteilung angenommen wurde [5]. Für eine Verteilung welche in einem bestimmten Intervall liegen soll, bietet sich die Beta-Verteilung an. Diese kann Verteilungen in einem beidseitig beschränkten Wertebereich simulieren. Standardgemäß nimmt diese aber nur Werte im Intervall [0;1] an, um das Intervall von [16;65] zu simulieren gilt es die Beta-Verteilung mit Hilfe der folgenden Formel zu transformieren:

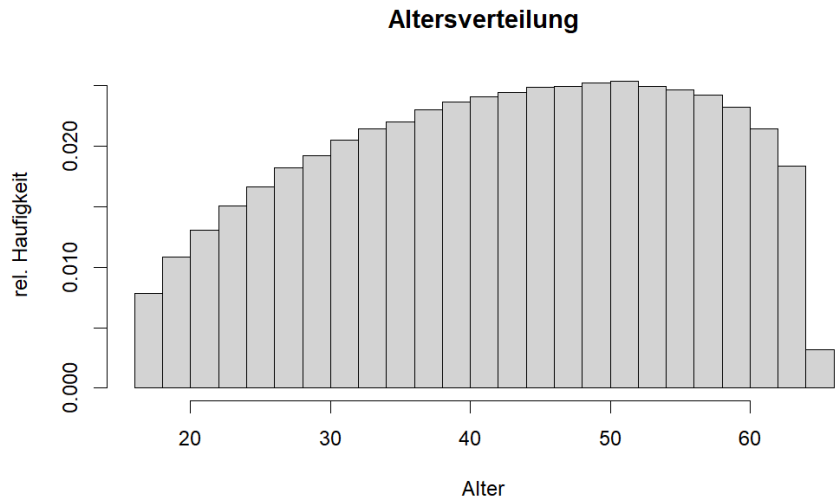
$$y_i = a + b * x_i$$

Zusätzlich gilt für die Beta-Verteilung $E(X) = \frac{p}{p+q}$ und $VAR(X) = \frac{pq}{(p+q+1)(p+q)^2}$.

Dadurch erhält man für den Parameter a den Wert 16 und für den Parameter b den Wert 49. Um den Erwartungswert von 43 Jahren abzubilden, müssen noch die Parameter p und q bestimmt werden. Dabei folgt p = (27/22) * q und q = 1,2. Der Parameter q kann selbst definiert werden. Er regelt die Varianz der Werteverteilung zum Mittelwert. Je höher der Parameter q gewählt wird, desto näher ist die Werteverteilung am Mittelwert und damit die Varianz geringer. Je kleiner q gewählt wird, desto höher die Varianz der Verteilung. Parameter p hingegen ergibt sich, wenn $\frac{p}{p+q}$ in der obigen Gleichung für x_i eingesetzt wird.

In der Programmiersprache R wurde in RStudio diese Beta-Verteilung mit der rbeta() Funktion umgesetzt und die erhaltenen Altersdaten noch auf die nächste ganze Zahl gerundet.

Folgende Abbildungen zeigen die Beta-Verteilung für die Variable Alter und ihre Eigenschaften.




```
   vars       n  mean    sd median trimmed   mad min max range  skew kurtosis   se
X1    1 1000000 43.01 12.72     44   43.32 14.83  16  65    49 -0.17    -1.02 0.01
```

Abbildung 1: Altersverteilung durch Beta-Verteilung und ihre Eigenschaften

Durch die simulierte Altersverteilung wird folgende Altersstruktur erreicht, welche in Tabelle 1 gezeigt wird. Diese deckt sich gut mit der zuvor erwähnten Statistik des ifo Instituts [2].

Altersverteilung	Anzahl gesamt	Häufigkeit
16 bis 25 Jahre	109.772	11,0 %
26 bis 35 Jahre	197.480	19,7 %
36 bis 45 Jahre	236.937	23,7 %
46 bis 55 Jahre	250.512	25,1 %
56 bis 65 Jahre	205.299	20,5 %

Tabelle 1: Altersverteilung

2.2 Betriebszugehörigkeit

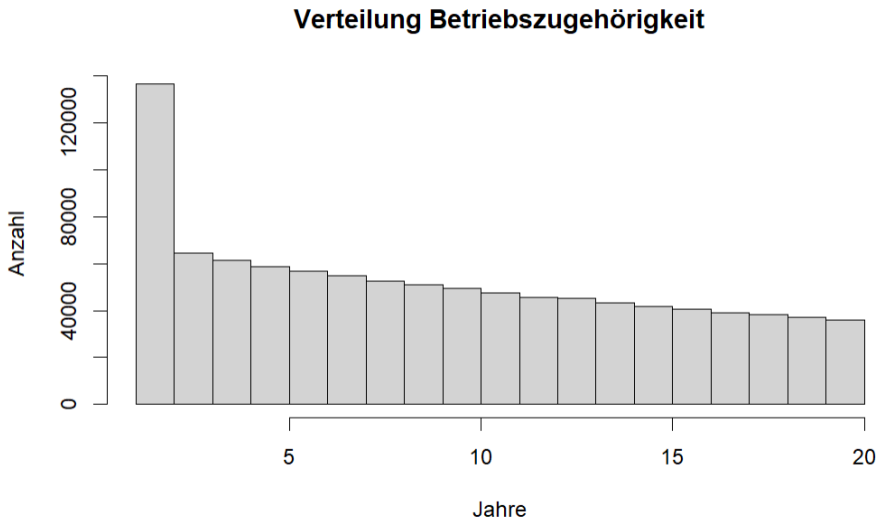
Neben dem Alter wird in dieser Simulationsstudie die Betriebszugehörigkeit einer Person simuliert. Die Betriebszugehörigkeit gibt an wie lange eine Person bereits beim aktuellen Arbeitgeber beschäftigt ist. Dies kann als ein wichtiger Indikator für die Stabilität der Beschäftigung angesehen werden und sich darüber hinaus auf die Zufriedenheit der Beschäftigten auswirken. Laut dem Statistischen Bundesamt waren im Jahr 2022 knapp 43% der Erwerbstätigen seit mindestens zehn Jahren bei ihrem Arbeitgeber beschäftigt [6].

Die Simulation der Variable Betriebszugehörigkeit beläuft sich im Wertebereich [1;20] wobei eins = erstes Jahr im Unternehmen, zwei = zweites Jahr im Unternehmen, … und 20 = zwanzigstes Jahr und länger im Unternehmen beschreibt. Demnach ist die Variable abhängig von dem zuvor generierten Alter, da eine beschäftigte Person frühstens mit 16 Jahren im Unternehmen starten konnte. Somit erhalten Personen die 16 Jahre alt sind auch nur die Betriebszugehörigkeit = 1 und sind somit das erste Jahr im Unternehmen. Wenn das Alter größer als 16 Jahre ist, so wird die maximale Zugehörigkeitsdauer durch die Gleichung „Alter – 16 + 1“ ausgerechnet und in dieser

möglichen Spanne der Zugehörigkeit ein zufälliger Wert vergeben. Wenn eine Person 18 Jahre alt ist kann sie demnach die Werte: eins, zwei oder drei bei der Variable Betriebszugehörigkeit erhalten. Da der Wertebereich bis maximal 20 beschränkt ist, ist die Spanne der Zugehörigkeit auch auf den Wert 20 beschränkt.

In RStudio wurde diese Logik durch die Funktion „berechneBetriebsZugehoerigkeit“ umgesetzt welche die Zugehörigkeitsspanne mit der min()-Funktion berechnet und die Zufallszahl innerhalb dieser Spanne mit der sample()-Funktion vergibt.

Schlussendlich veranschaulicht Abbildung 2 die Verteilung der Variable Betriebszugehörigkeit. Diese Verteilung hat einen Mittelwert von 9,35 Jahren welcher die zuvor genannten Informationen des Statistischen Bundesamtes gut widerspiegelt. Der Wert = 1 (erstes Jahr im Unternehmen) überragt die anderen Werte deutlich, da dieser für jedes Alter vergeben wird bzw. vergeben werden kann und die restlichen Werte nur eingeschränkt vergeben werden können.



	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
X1	1	1000000	9.35	5.72	9	9.13	7.41	1	20	19	0.24	-1.14	0.01

Abbildung 2: Verteilung Betriebszugehörigkeit und ihre Eigenschaften

2.3 Bildungsabschluss

Die dritte Variable Bildungsabschluss, beschreibt in dieser Simulationsstudie den höchsten Bildungsabschluss einer Person. Hierbei werden vier Ausprägungen unterschieden welche in folgender Tabelle dargestellt sind.

Ausprägung	Beschreibung / akademischer Grad
1	Lehre / Ausbildung
2	Bachelor
3	Master
4	Doktor

Tabelle 2: Ausprägungen der Variable Bildungsabschluss und Bedeutung

Diese Variable ist ebenfalls vom Alter der Person abhängig, da es für die hier verwendeten Bildungsabschlüsse unterschiedliche Altersangaben gibt, wann dieser Abschluss durchschnittlich erreicht wird. So beträgt das Durchschnittsalter mit welchem ein Bachelor Abschluss erreicht wird 23,8 Jahre und ein Master Abschluss wird durchschnittlich mit 26,4 Jahren absolviert [7]. Wohingegen ein Doktortitel im Schnitt mit 30 Jahren erreicht wird [8]. Für die Verteilung des Bildungsabschlusses auf die 1.000.000 Personen wurde die Verteilung der Bevölkerung in Deutschland nach beruflichem Bildungsabschluss im Jahr 2022 herangezogen [9]. Da in dieser Statistik allerdings auch andere Ausprägungen als die in dieser Simulationsstudie verwendeten betrachtet wurden, diente diese Auswertung als Orientierung für die Verteilung der Bildungsabschlüsse.

Dadurch das die Variable Bildungsabschluss vom Alter abhängt, werden Personen im Alter von 16 bis einschließlich 19 lediglich dem Bildungsabschluss = 1 zugeteilt. Im Alter von 20 bis einschließlich 25 ist es möglich den Bildungsabschluss = 1 oder 2 zu erhalten. Diese Verteilung wird unter Verwendung der logistischen Funktion berechnet. Jüngere Personen haben eine höhere Wahrscheinlichkeit Bildungsabschluss = 1 zu erhalten, während ältere Personen eher Abschluss = 2 erhalten. Ab dem Alter von 26 Jahren ist es möglich jeden Bildungsabschluss zugeteilt zu bekommen. Hierbei wurde sich an der zuvor erwähnten Statistik orientiert und die Wahrscheinlichkeiten für die entsprechenden Bildungsabschlüsse wie in Tabelle 3 dargestellt, verteilt.

Bildungsabschluss	Wahrscheinlichkeit
1 = Lehre / Ausbildung	60%
2 = Bachelor	25%
3 = Master	10%
4 = Doktor	5%

Tabelle 3: Wahrscheinlichkeitsverteilung für Vergabe der Bildungsabschlüsse ab 26 Jahren

In RStudio wurde dies durch die Funktion „berechneBildungsabschluss“ umgesetzt, welche mittels der sample()-Funktion die entsprechende Verteilung der Variable Bildungsabschluss in Abhängigkeit vom Alter und den Wahrscheinlichkeiten für die möglichen Bildungsabschlüsse verteilt hat.

Abbildung 3 zeigt die Verteilung der Bildungsabschlüsse aller Personen mit einem Balkendiagramm.

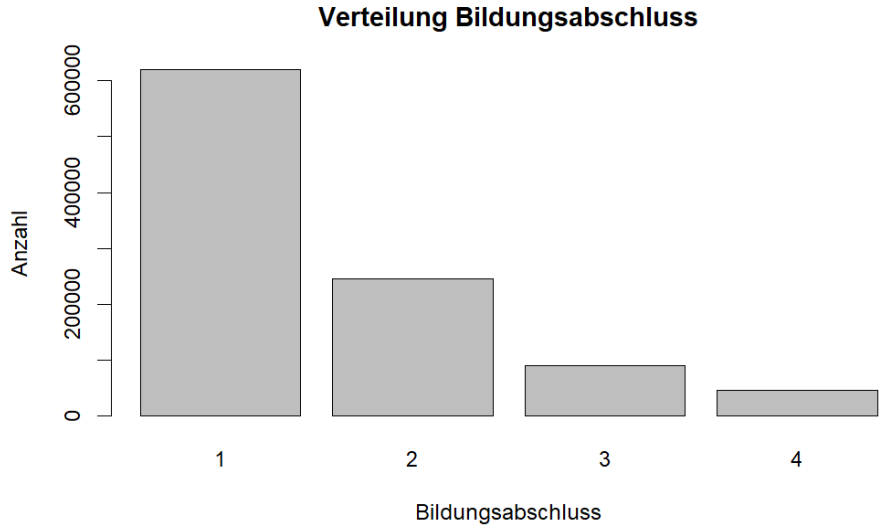


Abbildung 3: Verteilung Bildungsabschluss

2.4 Gehalt

Die Gehälter welche in der Simulationsstudie auf die 1.000.000 Personen verteilt werden, spiegeln die vierte Variable Gehalt wider. Diese Variable ist abhängig von dem zuvor vergebenen Bildungsabschluss. Dadurch werden Gehaltsschwankungen simuliert welche zwischen den Bildungsabschlüssen auftreten. Hierzu hat das Statistische Bundesamt eine Statistik für die Bruttomonatsverdienste für Vollzeitbeschäftigte nach Ausbildungsabschluss veröffentlicht [10]. Diese Statistik dient als Orientierung für die Verteilung des Gehalts in der Simulationsstudie. Die nachfolgend simulierten Werte entsprechen dem Bruttojahresgehalt der jeweiligen Person. Hierbei ist keine nähere Unterscheidung zwischen geleisteten Arbeitsstunden, Vollzeit- oder Teilzeitbeschäftigung oder ähnlichem. Für die Verteilung wird die Beta-Verteilung (bereits im Kapitel 2.1 Alter beschrieben) verwendet. Da ein höherer Bildungsabschluss in der Regel auch ein höheres Gehalt bedeutet, wird für jeden Bildungsabschluss eine eigene Beta-Verteilung erstellt. Die für die Verteilung und die Transformation genutzten Parameter können je nach Bildungsabschluss ausfolgender Tabelle entnommen werden.

Bildung	**Lehre / Ausbildung**	**Bachelor**	**Master**	**Doktor**
Maximum	80.000	90.000	110.000	140.000
E(x)	45.000	55.000	74.000	105.000
a (=Minimum)	25.000	37.000	48.000	60.000
b	55.000	53.000	62.000	80.000
q (frei wählbar)	3	3	3	3
p	4/7	18/35	13/18	9/7
Formel	**a + b * rbeta(sim, p, q)**			

Tabelle 4: Parameter der Beta-Verteilung für den jeweiligen Bildungsabschluss

In RStudio wurde für jeden Bildungsabschluss mit der rbeta()-Funktion die Bruttojahresgehaltsverteilung simuliert. Im Folgenden werden diese Verteilungen grafisch durch Histogramme dargestellt. Zudem zeigt Abbildung 5 die Gesamtverteilung der Variable Gehalt.

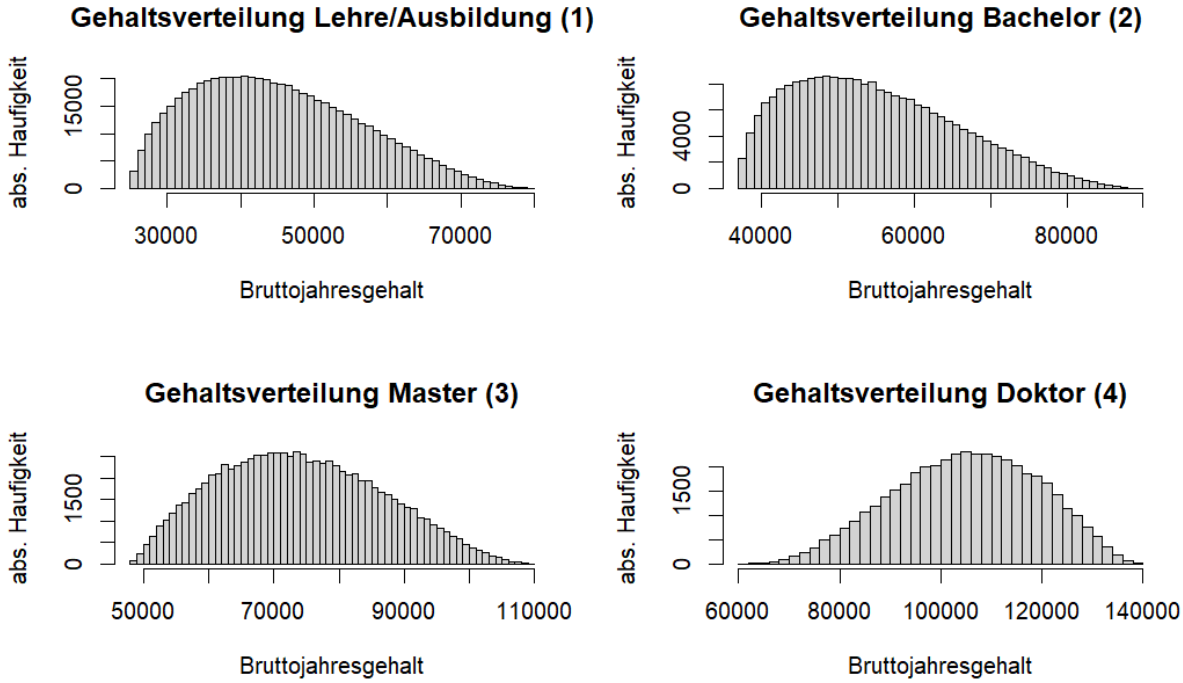
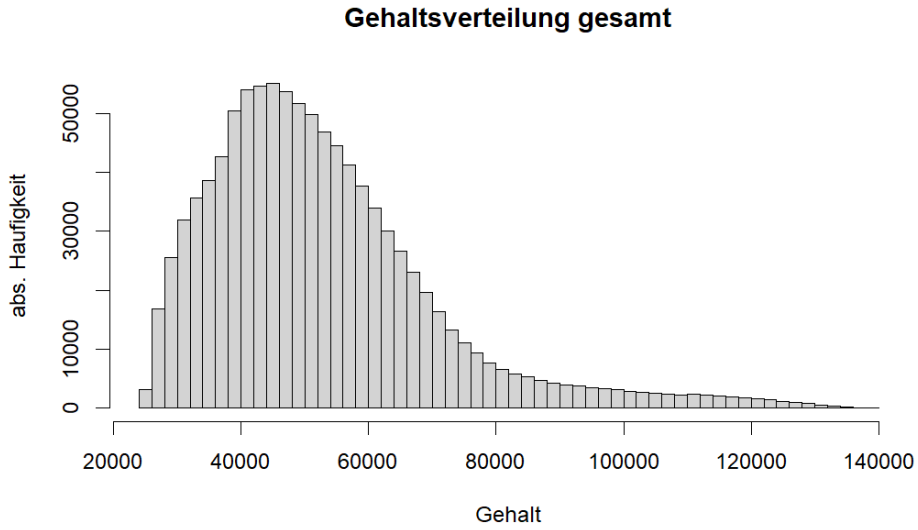


Abbildung 4: Gehaltsverteilung nach Bildungsabschluss



	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
X1	1	1000000	52740.91	18106	49455	50535.05	15046.91	25009	139283	114274	1.38	2.56	18.11

Abbildung 5: Gehaltsverteilung gesamt und ihre Eigenschaften

2.5 Zeit seit Gehaltserhöhung

Die fünfte Variable der Simulationsstudie ist Zeit seit Gehaltserhöhung. Diese Variable gibt an, wie viele Jahre seit der letzten Gehaltserhöhung vergangen sind.

Aus mehreren Quellen kann diesbezüglich entnommen werden, dass es sinnvoll ist, etwa alle ein bis zwei Jahre [11] oder alle 18 bis 24 Monate nach einer Gehaltsanpassung beim Arbeitgeber zu fragen [12].

Mit diesen Informationen soll eine Verteilung simuliert werden, welche einen Durchschnitt von ca. 1,6 bis 1,8 Jahre als Ergebnis hat. Demnach wurde der Wertebereich für die Variable Zeit seit Gehaltserhöhung auf null bis vier Jahre definiert. Hierbei beschreiben die Zahlen folgendes:

- 0 = letzte Gehaltserhöhung im aktuellen Jahr
- 1 = letzte Gehaltserhöhung vor einem Jahr
- 2 = letzte Gehaltserhöhung vor zwei Jahren
- 3 = letzte Gehaltserhöhung vor drei Jahren
- 4 = letzte Gehaltserhöhung vor vier Jahren

Um die Variable zu generieren muss die Betriebszugehörigkeit der jeweiligen Person beachtet werden, sodass es nicht möglich ist, dass eine Person vor drei Jahren die letzte Gehaltserhöhung erhalten hat, obwohl sie erst seit diesem Jahr im Unternehmen ist.

Dementsprechend ist die Variable Zeit seit Gehaltserhöhung von der Variable Betriebszugehörigkeit abhängig.

Im ersten Schritt wird mittels der Poisson-Verteilung eine ganze Zufallszahl generiert welche kleiner oder gleich vier ist, da der Wert vier die obere Grenze für den Wertebereich der Variable Zeit seit Gehaltserhöhung ist.

Die Poisson-Verteilung wird insbesondere dafür eingesetzt, seltene Zähl-Ereignisse zu modellieren und hat die Wahrscheinlichkeitsfunktion:

$$f_x(x) = e^{-\lambda} \frac{\lambda^x}{x!}$$

Zudem gilt:

$$E(x) = \lambda$$

$$VAR(x) = \lambda$$

Im zweiten Schritt wird überprüft, ob der berechnete Poisson-Wert (null bis vier) kleiner oder gleich als die „Betriebszugehörigkeit – 1“ ist.
Ist dies der Fall, so wird die Funktion mit diesem Wert beendet.
Ist dies nicht der Fall, so wird erneut ein Poisson-Wert generiert, bis ein geeigneter und logischer Wert gefunden wurde, welcher kleiner oder gleich der „Betriebszugehörigkeit – 1“ ist.

In beiden Fällen wird der berechnete Poisson-Wert als Variablenwert für die Variable Zeit seit Gehaltserhöhung gesetzt. Da diese Berechnung von der Variable Betriebszugehörigkeit abhängt, erhält jede Betriebszugehörigkeit eine geeignete Variable Zeit seit Gehaltserhöhung.

In RStudio wird diese Logik durch die Funktion „berechne_zeit“ umgesetzt. Mit Hilfe der Schleifen repeat{} und while{} wird sichergestellt dass ein geeigneter Wert ausgewählt wird. Mit der rpois()-Funktion werden die Werte generiert.

Um einen Durchschnittswert im Bereich 1,6 bis 1,8 Jahre bei der Verteilung zu erhalten wurde deshalb der Erwartungswert = 2,2 Jahre gesetzt. Damit erhält die Verteilung einen Durchschnittswert von 1,66 Jahren bei der Verteilung der Variable Zeit seit Gehaltserhöhung.

Durchschnittswert und Erwartungswert unterscheiden sich hier, da Werte nur unter bestimmten Einschränkungen vergeben werden können.

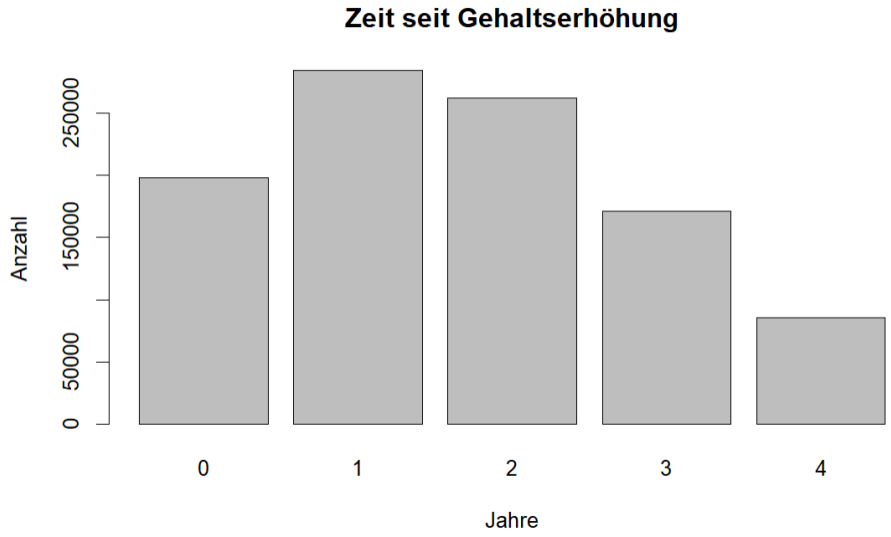


Abbildung 6: Verteilung Zeit seit Gehaltserhöhung

2.6 Work-Life-Balance

Die sechste potenziell erklärende Variable ist die Variable Work-Life-Balance. Diese spiegelt eine subjektive Einschätzung des Mitarbeiters / der Mitarbeiterin wider, inwieweit er oder sie das Berufsleben mit dem Privatleben vereinbaren kann. Dabei kann die Variable die Werte 0 = schlecht, 1 = ausgeglichen und 2 = gut annehmen. In der Referenzquelle hat Deutschland mit einer Work-Life-Balance von acht aus zehn Punkten sehr positiv abgeschlossen [13]. Um die Variable auf drei Ausprägungen auszuweiten werden deshalb die Wahrscheinlichkeiten 9% für den Wert 0 = schlecht, 15% für den Wert 1 = ausgeglichen und 76% für den Wert 2 = gut angenommen.

Diese Zuordnung und Verteilung lässt sich in RStudio mit der sample()-Funktion umsetzen.

Das nachfolgende Balkendiagramm veranschaulicht die Gesamtverteilung der Variable Work-Life-Balance.

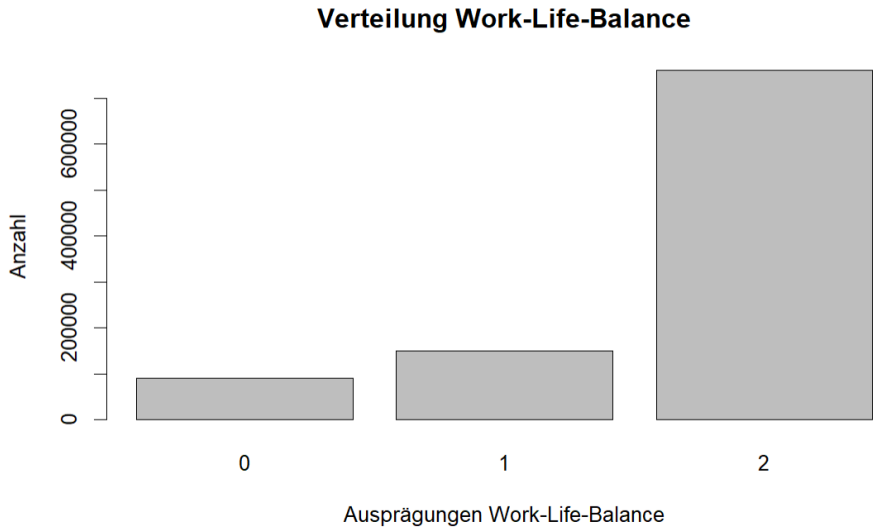


Abbildung 7: Verteilung Work-Life-Balance

2.7 Simulation der Zielvariable

In diesem Unterkapitel wird die Simulation der Zielvariable näher erläutert. Die Zielvariable gibt an, ob eine Person in der Grundgesamtheit von dem Unternehmen abwandert (= kündigt) oder nicht abwandert. Infolgedessen wird die Zielvariable Abwanderung genannt und kann die Werte 0 und 1 annehmen. Dabei steht die Zahl 1 dafür, dass eine Person von dem Unternehmen abwandert und die Zahl 0 dafür, dass eine Person nicht von dem Unternehmen abwandert und somit dem Unternehmen erhalten bleibt.

Die Zielvariable Abwanderung ist von vier der sechs generierten Variablen funktional abhängig, wovon eine Variable ordinal oder nominal skaliert ist. In der vorliegenden Simulationsstudie ist die Zielvariable Abwanderung von folgenden Variablen funktional abhängig:

- Alter (metrisch)
- Betriebszugehörigkeit (metrisch)
- Gehalt (metrisch)
- Bildungsabschluss (ordinal)

Aufgrund der ordinalen Skalierung der Variable Bildungsabschluss ist es notwendig Dummy-Variablen zu erstellen um der funktionalen Abhängigkeit gerecht zu werden damit zu einem späteren Zeitpunkt das logistische Regressionsmodell angewendet werden kann.

Dummy-Variablen sind künstliche Variablen die verwendet werden, um qualitative Daten zu repräsentieren und werden erstellt, um jede Kategorie als separate Binärvariable zu repräsentieren.

Die Variable Bildungsabschluss hat insgesamt vier mögliche Ausprägungen (1 = Lehre / Ausbildung, 2 = Bachelor, 3 = Master, 4 = Doktor). Aus diesem Grund ist es nötig dafür vier Dummy-Variablen zu erstellen.

Die Wahrscheinlichkeit p_i wird für jede Person in der Grundgesamtheit mithilfe der nachfolgenden Formel (= Sigmoid-Funktion) berechnet:

$$p_i := \frac{1}{1 + e^{-(\beta_0+\beta_1*a_i+\beta_2*b_i+\beta_3*c_i+\beta_4*d_i+\beta_5*e_i+\beta_6*f_i+\beta_7*g_i)}}$$

Hierbei repräsentiert a_i die Variable Alter, b_i ist die Betriebszugehörigkeit, c_i ist das Gehalt und d_i, e_i, f_i und g_i sind jeweils die Dummy-Variablen für den Bildungsabschluss.

Um die Wahrscheinlichkeiten berechnen zu können, müssen zuvor noch die β-Werte für die einzelnen erklärenden Variablen definiert werden:

- $\beta_0 = 1$
- $\beta_1 = -0{,}15$
- $\beta_2 = -0{,}20$
- $\beta_3 = 0{,}00015$
- $\beta_4 = -2$
- $\beta_5 = -1$
- $\beta_6 = 4$
- $\beta_7 = 4$

Bei der Wahl der β-Werte sind die absoluten Werte welche eine Variable annehmen kann von entscheidender Bedeutung, da diese mit dem entsprechendem β-Wert multipliziert werden. So kann beispielsweise die Variable Gehalt absolute Werte von 25.000 bis einschließlich 140.000 annehmen, wohingegen die Variable Betriebszugehörigkeit lediglich Werte von 1 bis einschließlich 20 annehmen kann. Hätten beide Variablen nun die gleichen β-Werte, so wäre der Einfluss der Variable Gehalt auf das Ergebnis der Wahrscheinlichkeitsberechnung überdimensional größer als der Einfluss der Variable Betriebszugehörigkeit. Die Entscheidung ob eine Person abwandern würde oder nicht wäre demnach größtenteils von der Variable Gehalt beeinflusst.

Bei der Wahl der β-Werte wurde darauf geachtet, dass die Variablen Gehalt (zugehörig: β_3) und die beiden Dummy-Variablen für den Bildungsabschluss = 3 (Master, zugehörig: β_6) und Bildungsabschluss = 4 (Doktor, zugehörig: β_7) eine stärkere Gewichtung auf die Wahrscheinlichkeit einnehmen als die anderen Variablen und damit die Richtung des Exponenten beeinflussen. Dies zeigt sich unteranderem dadurch, dass die β-Werte der Dummy-Variablen für den Bildungsabschluss = 3 und 4 höher gesetzt wurden als die β-Werte der Dummy-Variablen für den Bildungsabschluss = 1 und 2.

Dadurch soll für die spätere Betrachtung und Analyse der Daten sichergestellt werden, dass Personen mit einem höheren Bildungsabschluss bzw. Gehalt eher kündigen, als Personen mit einem geringeren Gehalt bzw. Bildungsabschluss.

Außerdem wurden die β-Werte entsprechend so gewählt, dass der Wertebereich des Exponenten in der Sigmoid-Funktion überwiegend zwischen -5 und 5 liegt. Diese Verteilung wird durch folgendes Histogramm in Abbildung 8 veranschaulicht. Die Ausreiser auf der rechten Seite des Histogramms bei den x-Werten größer als 5 sind durch die Variable Gehalt und deren großen Wertebereich zu erklären. Die Ausreiser kommen in diesem Fall durch die hohen Gehälter zustande und können nicht durch die anderen β-Werte bzw. Werte der Variablen ausgeglichen werden.

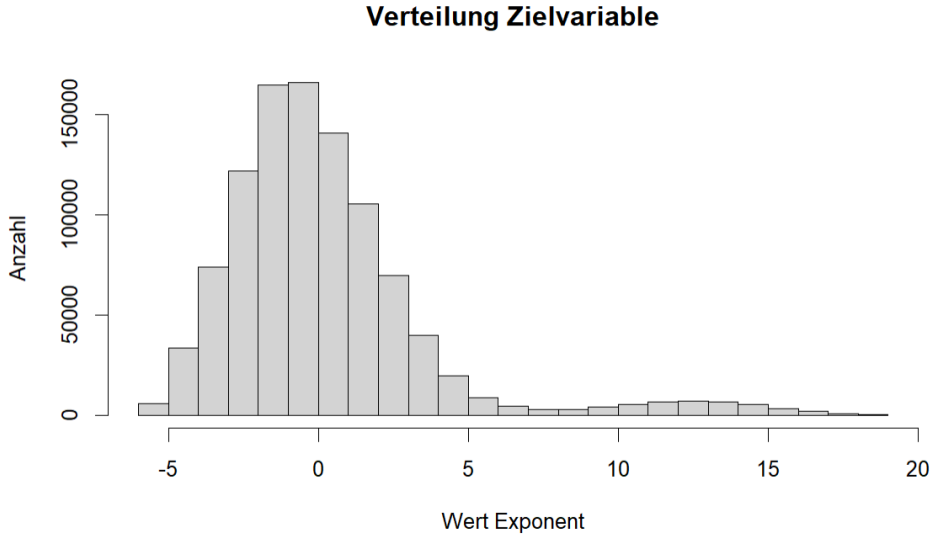


Abbildung 8: Verteilung Zielvariable

Zudem ist bei der Wahl der β-Werte das zustande kommende Rauschen ($\overline{VAR(\varepsilon)}$) der Daten von Bedeutung. Ein zu großes Rauschen impliziert keine Abhängigkeit der Zielvariable von den β-Werten wodurch zum späteren Zeitpunkt kein geeignetes Vorhersagemodell erstellt werden kann. Ein zu kleines Rauschen würde allerdings bedeuten, dass das trainieren eines perfekten Modells möglich wäre, da die Daten dann kaum verrauscht sind. Deshalb ist es wichtig ein mittleres Rauschen zu erzeugen um die späteren Aufgaben welche auf der Grundgesamtheit basieren nicht negativ zu beeinflussen. Aus diesem Grund wurde ein Rauschen im Bereich von 0,07 bis 0,12 angestrebt. Dieses Rauschen entspricht später dem minimalen MSE eines trainierten Modells.

Die beiden Variablen Zeit seit Gehaltserhöhung und Work-Life-Balance werden in der Berechnung und somit in der Generierung der Zielvariable Abwanderung nicht beachtet. Somit haben diese keinen direkten Einfluss auf die Abwanderung einer Person.

In folgendem Histogramm wird die Verteilung der berechneten Wahrscheinlichkeit / Ergebnis der Sigmoid-Funktion deutlich.

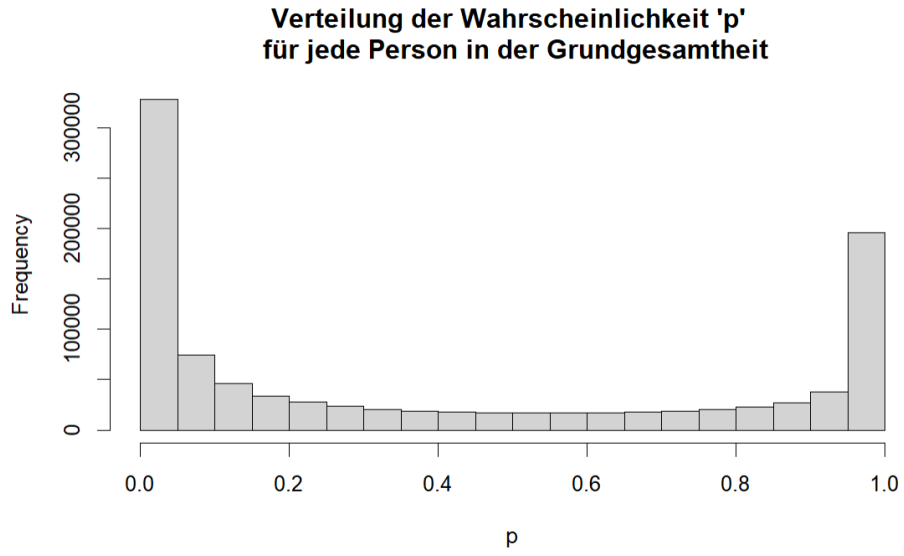


Abbildung 9: Verteilung p für die Grundgesamtheit

Im Anschluss werden normalverteilte Zufallszahlen u_i zwischen 0 und 1 mit der runif()-Funktion in RStudio erstellt. Die zuvor berechnete Wahrscheinlichkeit p_i wird nun mit der Zufallszahl u_i verglichen. Wenn die Zufallszahl u_i kleiner oder gleich der Wahrscheinlichkeit p_i ist, wird eine 1 vergeben, falls nicht wird eine 0 vergeben. Dies ist gleichzeitig die Simulation der Zielvariable Abwanderung.

Nachfolgendes Balkendiagramm zeigt die Verteilung der Zielvariable Abwanderung. Die Ausprägung = 1 bedeutet, dass eine Person von dem Unternehmen abwandert. Dem entsprechen ca. 40%. Die Ausprägung = 0 bedeutet, dass eine Person von dem Unternehmen nicht abwandert. Dem entsprechen ca. 60%.

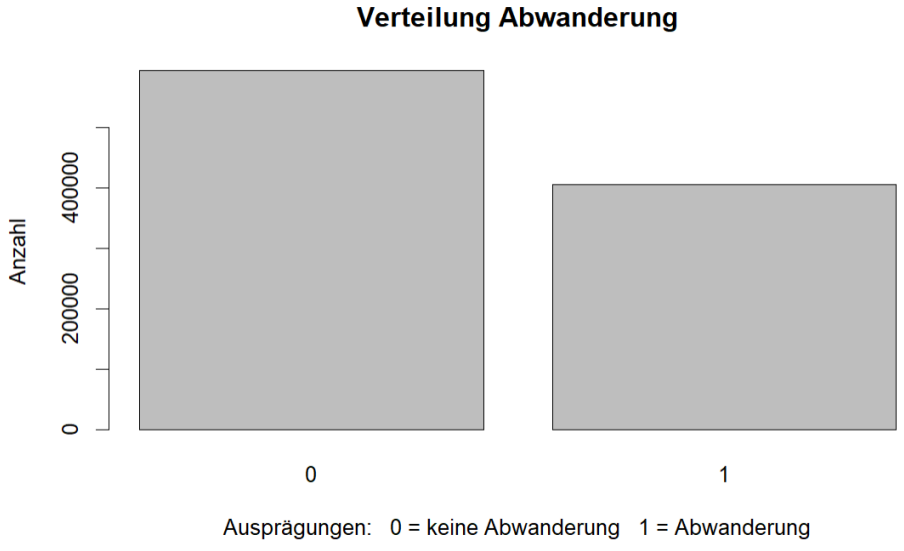


Abbildung 10: Verteilung Abwanderung

Letztlich wird die Stichprobe der Grundgesamtheit im Umfang von N = 10.000 Elementen für den Data Scientist erstellt. Die Stichprobe wird zufällig aus der Grundgesamtheit entnommen.

In RStudio wird hierzu dieser Code ausgeführt:

df_dataScientist = daten[sample(nrow(daten), 10000, replace = FALSE),]

Das Attribut „replace = FALSE“ stellt sicher, dass jeder Datensatz nur einmal in der Stichprobe erscheint.

3 Simulation der Perspektive des Data Scientist

Der Personalchef der Pay Solutions GmbH hat mit der Zeit die Erfahrung gemacht, dass es innerhalb des Unternehmens gewisse Muster bei der Fluktuation gibt. Dies möchte der Personalchef näher untersuchen um Mitarbeiter zukünftig besser im Unternehmen halten zu können. Hierzu beauftragt er einen Data Science Freelancer, welcher die Fluktuation analysieren soll. Der Personalchef übergibt dem Data Scientist eine Stichprobe mit N = 10.000 Datensätzen welche Informationen über das Alter, die Betriebszugehörigkeit, den Bildungsabschluss, das Gehalt, die Zeit seit der letzten Gehaltserhöhung, die Work-Life-Balance und die Aussage ob die Person vom Unternehmen abgewandert ist oder nicht beinhaltet.

Ein beispielhafter Auszug aus dieser Stichprobe zeigt nachfolgend Tabelle 5.

Alter	Betriebs-zugehörigkeit	Bildungs-abschluss	Gehalt	Zeit seit Gehaltserhöhung	Work-Life-Balance	Ab-wanderung
35	10	1	34.513	1	2	0
47	11	1	33.814	2	2	0
40	17	2	59.193	4	2	0
49	15	2	40.464	2	2	0
19	1	1	34.585	0	2	0
42	16	4	112.146	1	0	1

Tabelle 5: Auszug Datenstichprobe des Data Scientist

Wovon eine Abwanderung letztlich abhängt weiß der Personalchef nicht, weshalb der Data Scientist basierend auf dem für ihn zur Verfügung gestellten Datensatz ein Modell entwickeln soll mit welchem zukünftig vorhergesagt werden kann, ob eine Person vom Unternehmen abwandert oder nicht.

Um die Aufgabe des Personalchefs der Pay Solutions GmbH zu lösen, geht der Data Scientist nach dem CRISP-DM-Modell vor. Dieses Modell ist eine bewährte Methode zur Anleitung einer solchen Datenanalyse und steht für „Cross-Industry Standard Process for Data-Mining“ [14]. Das Modell besteht aus verschiedenen Phasen welche in nachfolgender Abbildung 11 dargestellt sind und im weiteren Verlauf dieses Kapitels erläutert werden.

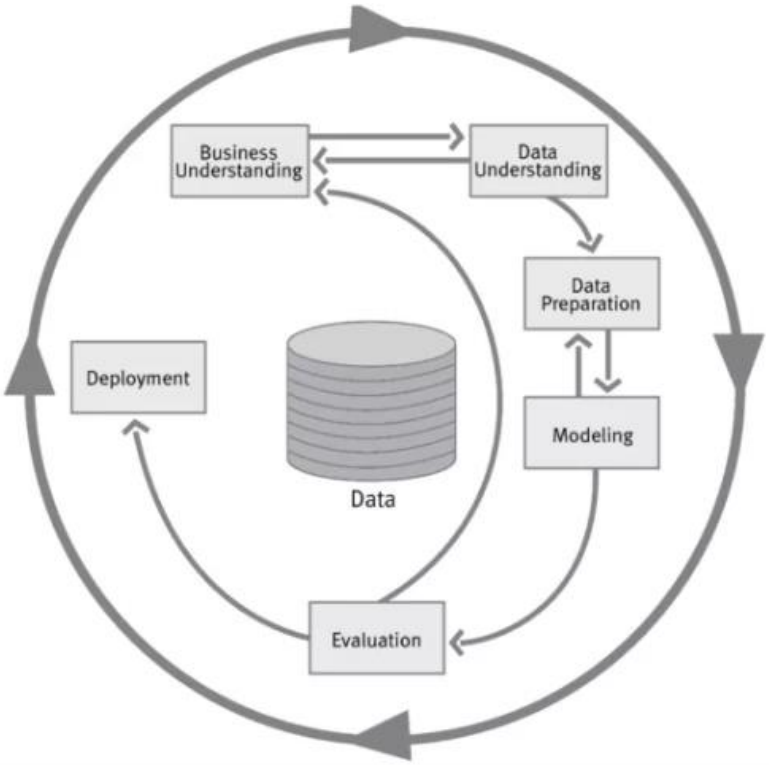


Abbildung 11: CRISP-DM-Modell [15]

3.1 Business Understanding

Die erste Phase des Modells beschreibt das „Business Understanding“. Dabei soll die betriebswirtschaftliche Problemstellung so präzise wie möglich beschrieben werden. Diese Problemstellung zeigt die konkreten Anforderungen an die Datenanalyse auf und bildet die Grundlage für alle weiteren Schritte und Phasen des Modells [15].

Dem Data Scientist wird schnell klar, dass der Personalchef der Pay Solutions GmbH etwas gegen die Fluktuation des Unternehmens tun möchte. Denn eine zu hohe Fluktuation bedeutet, dass regelmäßig wertvolles Fachwissen und Kompetenzen dem Unternehmen verloren gehen. Durch die Neubesetzung der Stellen fallen zudem hohe Kosten an. Nicht zu vergessen, dass eine hohe Fluktuation negative Effekte auf die Unternehmenskultur haben kann und potenzielle neue Mitarbeiter von dem Unternehmen abschrecken könnten [16].

Das Modell welches der Data Scientist entwickelt, kann dazu genutzt werden potenziell abwanderungsgefährdete Mitarbeiter systemseitig zu identifizieren, sodass der Personalchef die jeweiligen Führungskräfte rechtzeitig darauf aufmerksam machen kann und diese etwas gegen eine potenzielle Abwanderung des Mitarbeiters unternehmen können.

3.2 Data Understanding

In der Phase des „Data Understanding“ verschafft man sich einen Überblick über die vorliegenden Daten und versucht bestehende Zusammenhänge zu identifizieren. Außerdem werden die Daten bewertet, inwieweit sie für die Analyse ausreichend sind oder ob es Datenlücken gibt [15].

Mit Hilfe der Funktion summary() kann in RStudio eine gute Übersicht über die vorliegenden Daten generiert werden. So zeigt sie für einen Datensatz mehrere statistische Informationen für jede Spalte auf, wie beispielsweise das Minimum und Maximum oder auch den Median und den Mittelwert. Eine beispielhafte Ausgabe der summary()-Funktion ist in Abbildung 12 zu sehen.

```
     Alter       Betriebszugehoerigkeit Bildungsabschluss      Gehalt       Zeit_seit_Gehaltserhoehung Work_Life_Balance  Abwanderung    
 Min.   :16.00   Min.   : 1.000         Min.   :1.000     Min.   : 25083   Min.   :0.000              Min.   :0.000     Min.   :0.0000  
 1st Qu.:33.00   1st Qu.: 4.000         1st Qu.:1.000     1st Qu.: 40366   1st Qu.:1.000              1st Qu.:2.000     1st Qu.:0.0000  
 Median :44.00   Median : 9.000         Median :1.000     Median : 49459   Median :2.000              Median :2.000     Median :0.0000  
 Mean   :43.07   Mean   : 9.295         Mean   :1.571     Mean   : 52892   Mean   :1.635              Mean   :1.682     Mean   :0.4117  
 3rd Qu.:54.00   3rd Qu.:14.000         3rd Qu.:2.000     3rd Qu.: 61109   3rd Qu.:2.000              3rd Qu.:2.000     3rd Qu.:1.0000  
 Max.   :65.00   Max.   :20.000         Max.   :4.000     Max.   :137978   Max.   :4.000              Max.   :2.000     Max.   :1.0000  
```

Abbildung 12: Anwendung der summary()-Funktion auf den Datensatz

```
'data.frame':   10000 obs. of  7 variables:
 $ Alter                    : num  35 47 40 49 19 57 58 54 61 24 ...
 $ Betriebszugehoerigkeit   : num  10 11 17 15 1 6 12 16 11 2 ...
 $ Bildungsabschluss        : num  1 1 2 2 1 1 1 1 3 1 ...
 $ Gehalt                   : num  34513 33814 59193 40464 34585 ...
 $ Zeit_seit_Gehaltserhoehung: int  1 2 4 2 0 4 2 2 2 1 ...
 $ Work_Life_Balance        : num  2 2 2 2 2 2 2 2 2 2 ...
 $ Abwanderung              : num  0 0 0 0 0 0 0 0 1 1 ...
```

Abbildung 13: Anwendung der str()-Funktion auf den Datensatz

In Kombination mit der str()-Funktion (siehe Abbildung 13) welche die Struktur eines Datensatzes beschreibt, lassen sich erste Informationen und Ableitungen für den Data Scientist treffen.

So kann daraus entnommen werden, dass es sich bei den Variablen Alter, Betriebszugehörigkeit, Gehalt und Zeit seit Gehaltserhöhung sehr wahrscheinlich um metrische Variablen handeln muss. Eine nominale oder ordinale Skalierung werden hingegen die Variablen Bildungsabschluss, Work-Life-Balance und Abwanderung haben, wobei die Variable Abwanderung angibt, ob eine Person abwandert oder nicht, da diese nur die Werte 0 oder 1 annehmen kann.

Diese Variable wird direkt vom Data Scientist näher betrachtet und durch folgendes Balkendiagramm (Abbildung 14) lässt sich erahnen, dass der Wert 1 dafürsteht, dass eine Person vom Unternehmen abgewandert ist. Wohingegen der Wert 2 aussagt, dass eine Person noch Teil des Unternehmens ist. Mit Hilfe der sum()-Funktion kann die Anzahl der beiden Ausprägungen bestimmt werden. So sind von den 10.000 Personen, 4117 abgewandert und 5883 Personen nicht abgewandert.

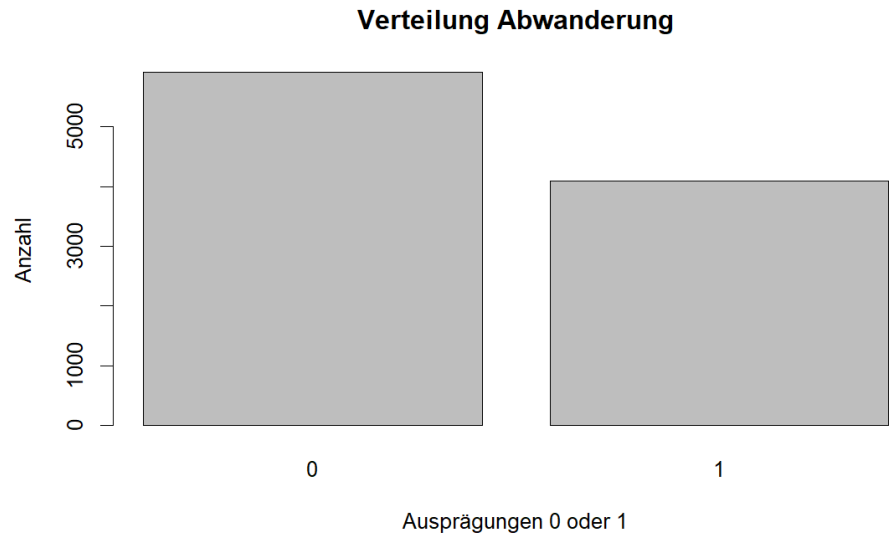


Abbildung 14: Verteilung Abwanderung Data Scientist

Um die Variable Alter besser untersuchen zu können, wird ein Histogramm für die Altersverteilung erstellt.

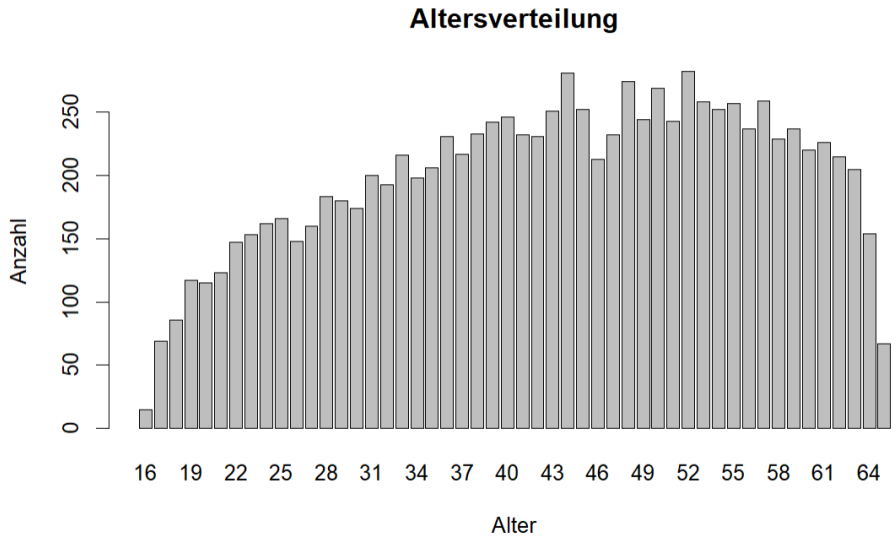


Abbildung 15: Altersverteilung Data Scientist

Da diese Verteilung dem Data Scientist nicht ausreicht, erstellt er ein Boxplot und teilt die Alterswerte in Altersgruppen ein und veranschaulicht diese mittels einem Balkendiagramm. Zusätzlich berechnet er mit der mean()-Funktion den Mittelwert der Variable Alter.

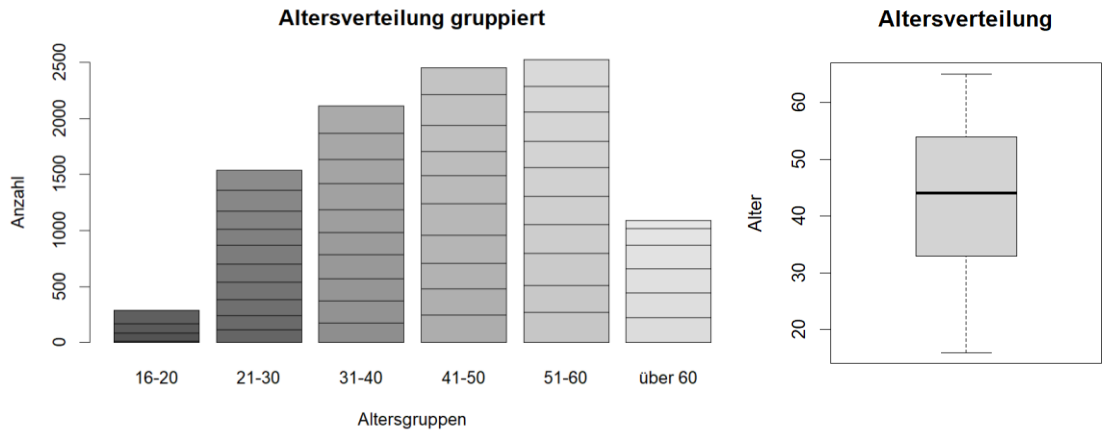


Abbildung 16: Altersverteilung gruppiert und Boxplot Data Scientist

Dadurch lässt sich erkennen, dass das durchschnittliche Alter 43,07 Jahre beträgt und der Großteil der Altersverteilung zwischen 30 und 60 Jahren liegt.

Auch die Betriebszugehörigkeit veranschaulicht der Data Scientist mittels einem Histogramm und der mean()-Funktion.

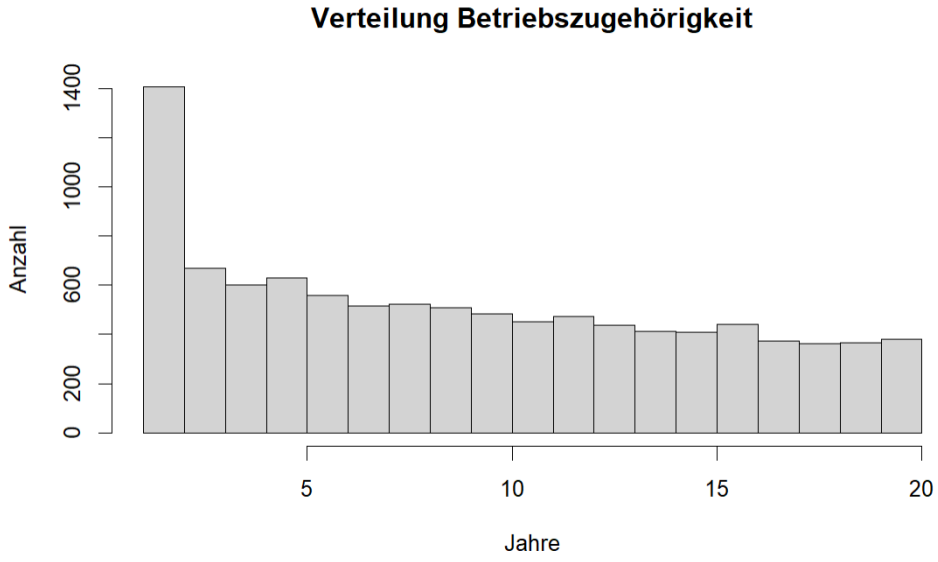


Abbildung 17: Verteilung Betriebszugehörigkeit Data Scientist

Dabei stellt er fest, dass sich ein Großteil der Personen im ersten Jahr im Unternehmen befinden und das die sonstige Betriebszugehörigkeit relativ ausgeglichen ist aber desto größer die Werte werden immer mehr abnimmt. Die durchschnittliche Betriebszugehörigkeit der Pay Solutions GmbH beträgt 9,3 Jahre.

Bei der Untersuchung der Variable Bildungsabschluss wird deutlich, dass diese ordinal skaliert ist. Die Ausprägung 1 bedeutet, dass der höchste Bildungsabschluss einer Person die Lehre / Ausbildung ist. Ausprägung 2 beschreibt einen Bachelor, 3 einen

Master und die Ausprägung 4 einen Doktor als höchsten Bildungsabschluss. Die Verteilung wird durch ein Balken- und Kreisdiagramm deutlich.

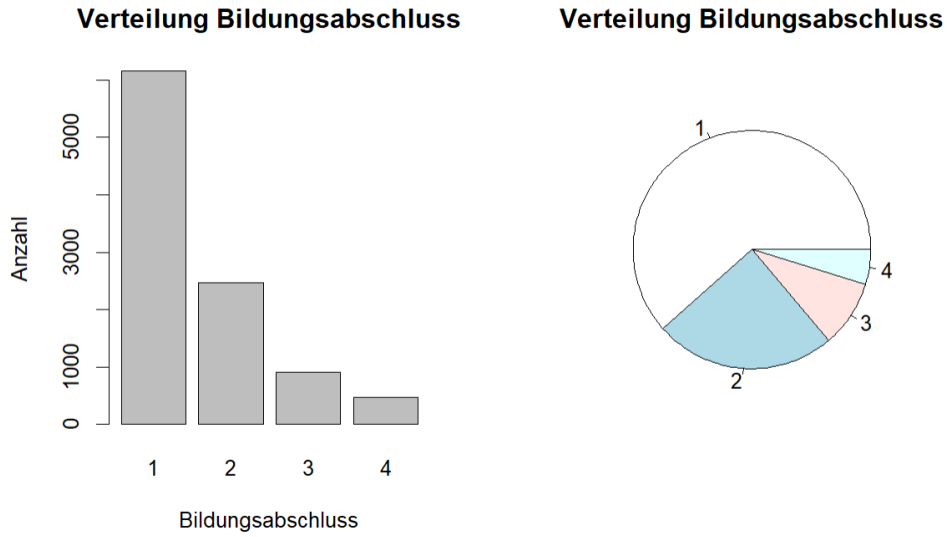


Abbildung 18: Verteilung Bildungsabschluss Data Scientist

Die Verteilung der Variable Gehalt wird mit einem Histogramm veranschaulicht, welches zeigt, dass die Mehrheit ein Bruttojahresgehalt von ca. 30.000€ bis 60.000€ bezieht. Der Boxplot stärkt diese Annahme. Zudem kann mit der mean()-Funktion der das Durchschnittsgehalt von 52.891,85€ berechnet werden.

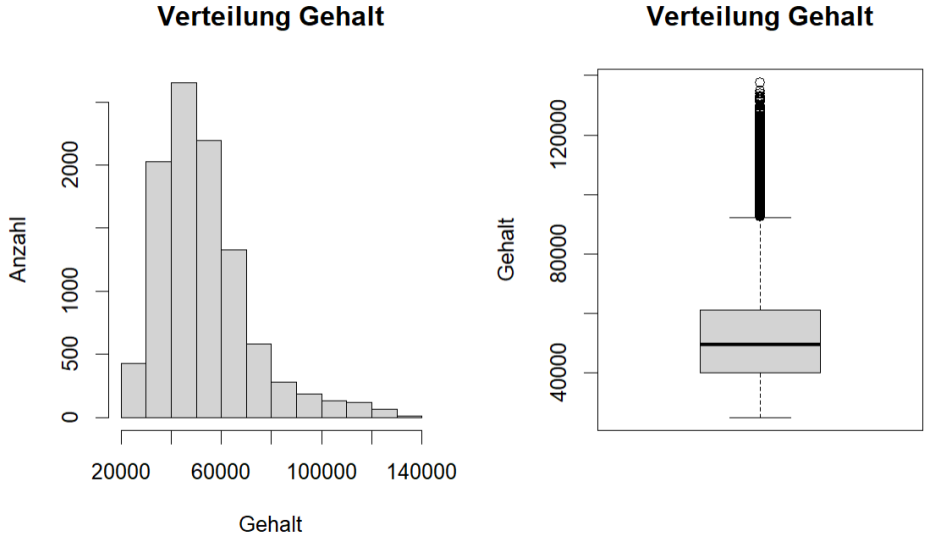


Abbildung 19: Verteilung Gehalt Data Scientist

Da der Data Scientist die Vermutung hat, dass das Gehalt auch in gewissem Maße mit dem Bildungsabschluss zusammenhängt, wendet er die aggregate()-Funktion an um auf Basis des Bildungsabschlusses das Durchschnittsgehalt zu berechnen. Dies wird durch folgendes Balkendiagramm deutlich. Dieses bestätigt seine Annahme. Es lässt sich erkennen, dass das Durchschnittsgehalt mit einem höheren Bildungsabschluss zunimmt.

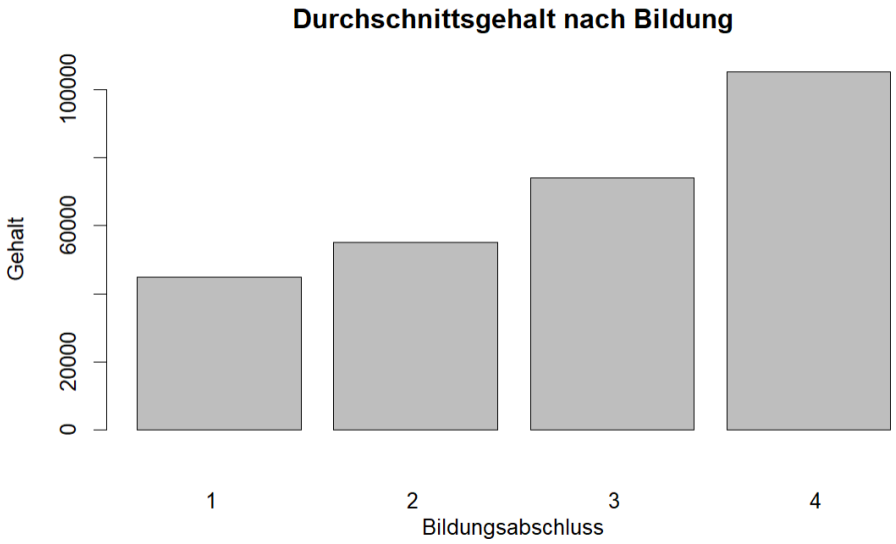


Abbildung 20: Durchschnittsgehalt nach Bildung Data Scientist

Bei der Untersuchung der Variable Zeit seit Gehaltserhöhung zeigt sich, dass im Durchschnitt ca. 1,64 Jahre vergehen bis eine Person eine Gehaltserhöhung erhält. Die Verteilung veranschaulicht der Data Scientist mit einem Balkendiagramm.

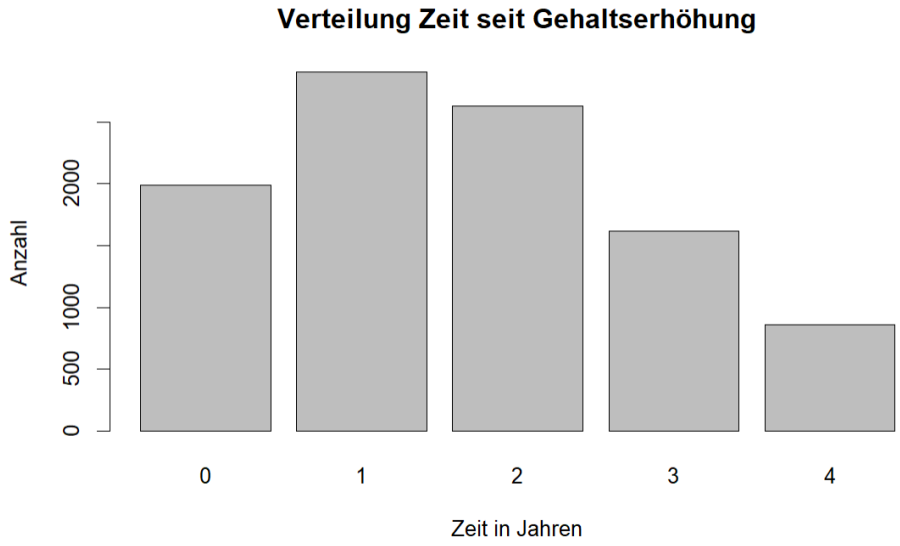


Abbildung 21: Verteilung Zeit seit Gehaltserhöhung Data Scientist

Die Verteilung der Variable Work-Life-Balance nimmt lediglich Werte von 0 bis 2 an, bei welcher 0 für eine schlechte, 1 für eine ausgeglichene und 2 für eine gute Work-Life-Balance steht. Mittels eines Kreisdiagramms lässt sich diese gut veranschaulichen.

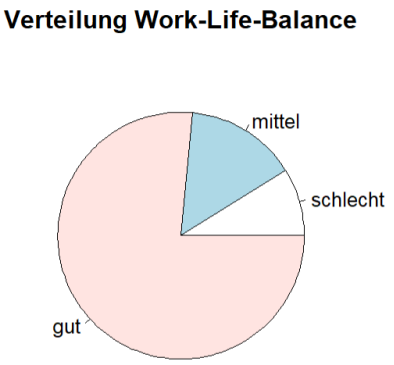


Abbildung 22: Verteilung Work-Life-Balance Data Scientist

Nachdem der Data Scientist alle Variablen näher untersucht hat, möchte er seine Annahme, dass das Gehalt mit dem Bildungsabschluss zusammenhängt nochmals untersuchen. Näher kann er mittels der Korrelation generell lineare Zusammenhänge zwischen den verschiedenen Variablen erkennen. Die Korrelation ist ein Maß um die stärke einer statistischen Beziehung von zwei Variablen zueinander festzustellen. Dabei kann die Korrelation Werte von -1 bis 1 annehmen. Der Wert 1 beschreibt, dass eine Variable positiv mit einer anderen korreliert. Ein Beispiel dafür wäre, dass das Gehalt ansteigt, je höher der Bildungsabschluss ist. Der Wert -1 beschreibt hingegen eine negative Korrelation zwischen zwei Variablen. Ein Beispiel hierfür wäre, dass mit zunehmendem Alter, die Lebenserwartung der Person sinkt. Dabei ist die Korrelation immer ungerichtet, bedeutet, dass sie keine Informationen darüber enthält ob eine Variable eine andere bedingt [17].

Um eine Korrelationsmatrix von den vorliegenden Daten zu erstellen nutzt der Data Scientist die Funktionen cor() und corrplot() in RStudio.

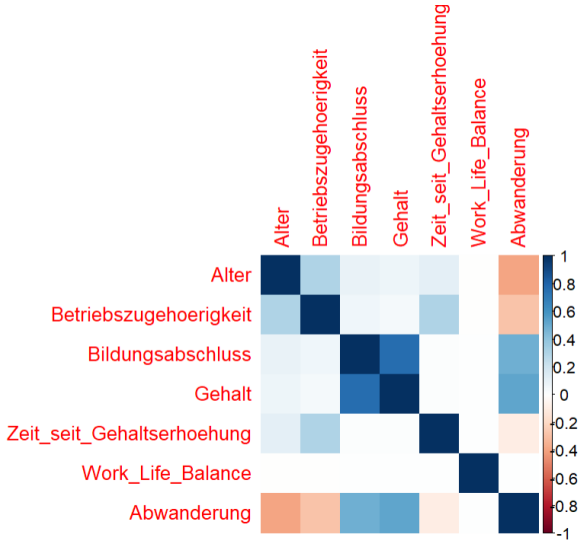


Abbildung 23: Korrelationsmatrix Data Scientist

Dabei lässt sich feststellen, dass die Variable Gehalt tatsächlich positiv mit der Variable Bildungsabschluss korreliert. Was für den Personalchef aber noch deutlich wichtiger ist, ist die Tatsache, dass die Variable Gehalt und Bildungsabschluss ebenso positiv mit der Variable Abwanderung korrelieren. Außerdem korrelieren die beiden Variablen Alter und Betriebszugehörigkeit negativ mit der Variable Abwanderung. Es kann abgeleitet werden, dass Personen mit einem höheren Gehalt und Bildungsabschluss eher abwandern als Personen mit einem geringeren Gehalt und Bildungsabschluss. Zudem wandern eher jüngere Personen und Personen mit einer kürzeren Betriebszugehörigkeit ab.

3.3 Data Preparation

In der Phase der „Data Preparation“ werden die Daten transformiert und bereinigt um sie für die Anwendung des Modells vorzubereiten [15]. Der Data Scientist muss also Vorbereitungen treffen um im Anschluss das logistische Regressionsmodell anwenden zu können.

Hierfür müssen zuerst die kategorialen Variablen welche mehr als zwei Ausprägungen haben in Dummy-Variablen umgewandelt werden. Dies betrifft die Variablen Bildungsabschluss und Work-Life-Balance. Durch die Dummy-Variablen können qualitative Daten in quantitative Daten umgewandelt werden (nähere Erläuterung Kapitel 2.7).

Die Dummy-Variablen erstellt der Data Scientist mittels der Library „psych“ welche die Funktion dummy.code() bereitstellt um Dummy-Variablen zu erstellen.

Eine weitere Vorbereitung die der Data Scientist trifft bevor er das logistische Regressionsmodell anwendet, ist die Aufteilung des Datensatzes in Trainings- und Testdaten. Hierbei verwendet er eine Aufteilung von 80% Trainingsdaten und 20% Testdaten. Die Trainingsdaten werden im Nachhinein dafür verwendet, um die Modelle mit diesen Daten zu trainieren. Mit Hilfe der Testdaten können die Modelle im Anschluss an das Training validiert werden.

Zur Aufteilung in Trainings- und Testdaten verwendet der Data Scientist die sample()-Funktion.

```
index = sample(1:nrow(df_DS_erw), 0.8 * nrow(df_DS_erw))

train_data = df_DS_erw[index, ]

test_data = df_DS_erw[-index, ]
```

3.4 Modelling

Die Phase „Modelling“ beschäftigt sich damit eine Modellierungstechnik auszuwählen mit welcher das Modell oder die Modelle erstellt werden. Anschließend werden die Modelle mit den Trainingsdaten trainiert und überprüft [15].

Der Data Scientist wendet die logistische Regression an um zu überprüfen ob es einen Zusammenhang zwischen der Variable Abwanderung und den anderen Variablen gibt. Hierbei geht er nach dem Prinzip der „Backward-Selection“ vor. Dies ist ein Selektionsverfahren bei welchem zunächst alle Variablen in einem Modell berücksichtigt werden. Nach und nach werden dann nicht signifikante Variablen aus dem Modell entfernt, bis das Beste gefunden wird [18].

Um die logistische Regression in RStudio umzusetzen, nutzt der Data Scientist die glm()-Funktion. Das erste Modell wird mit allen Variablen erstellt.

model1 =
glm(Abwanderung ~ Alter + Betriebszugehoerigkeit + Gehalt +
Zeit_seit_Gehaltserhoehung + Dummy_Bildung2 + Dummy_Bildung3 +
Dummy_Bildung4 + Dummy_WLB1 + Dummy_WLB0,
data = train_data, family = binomial())

Dabei ist es wichtig das Argument „family = binomial()“ zu setzen. Über das Argument wird festgelegt, welche Verteilung die logistische Regression anwendet. Da der Data Scientist ein Modell für die Variable Abwanderung, welche nur die Werte 0 und 1 annehmen kann erstellt, wird die „family=binomial()“ gesetzt um sicherzugehen, dass das Modell auf eine binomiale Verteilung abzielt.

```
Call:
glm(formula = Abwanderung ~ Alter + Betriebszugehoerigkeit +
    Gehalt + Zeit_seit_Gehaltserhoehung + Dummy_Bildung2 + Dummy_Bildung3 +
    Dummy_Bildung4 + Dummy_WLB1 + Dummy_WLB0, family = binomial(),
    data = train_data)

Coefficients:
                            Estimate    Std. Error z value      Pr(>|z|)
(Intercept)                -1.139526423   0.194146191  -5.869 0.00000000437 ***
Alter                      -0.146077812   0.004337399 -33.679       < 2e-16 ***
Betriebszugehoerigkeit     -0.201583777   0.009122087 -22.098       < 2e-16 ***
Gehalt                      0.000150133   0.000004838  31.031       < 2e-16 ***
Zeit_seit_Gehaltserhoehung -0.022278698   0.035274404  -0.632         0.528
Dummy_Bildung2              1.078767602   0.090957763  11.860       < 2e-16 ***
Dummy_Bildung3              6.213853573   0.439673497  14.133       < 2e-16 ***
Dummy_Bildung4             15.879961278 238.196633540   0.067         0.947
Dummy_WLB1                  0.042106392   0.112072333   0.376         0.707
Dummy_WLB0                  0.080236780   0.139973095   0.573         0.566
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Abbildung 24: Ergebnis Model 1

Abbildung 24 zeigt die Zusammenfassung des ersten Modells mit allen Variablen. Um das Modell zu überprüfen können die p-Werte (letzte Spalte „Pr(>|z|)“) in Betracht gezogen werden. Diese geben Informationen darüber, ob die geschätzten Koeffizienten der Variablen statistisch signifikant sind. Die p-Werte helfen Hypothesen zu testen, ob es eine tatsächliche Beziehung zwischen den erklärenden Variablen und der zu erklärenden Zielvariable (Abwanderung) gibt. Ein kleiner p-Wert, typischerweise $p<0,05$ weist darauf hin, dass es starke Evidenz gegen die Nullhypothese (keine Beziehung der Variable zur Zielvariable) gibt. Dadurch kann diese verworfen werden und es kann angenommen werden, dass es bei dieser Variable eine signifikante Beziehung zur Zielvariable gibt. Demnach werden die Variablen Zeit seit

Gehaltserhöhung, Dummy Bildung4 und die beiden Dummy-Variablen der Work-Life-Balance als nicht signifikant angesehen. Da der Data Scientist nach der Backward-Selection vorgeht, entfernt er zunächst die Variable mit dem höchsten p-Wert. In diesem Fall wird die Variable Dummy_Bildung4 entfernt und die weiteren Modelle mit den übrigen Variablen zu trainieren.

Zusätzlich kann das Modell auf Kollinearität überprüft werden. Die Kollinearität beschreibt den Fall, dass zwei oder mehr Variablen in einem Zusammenhang stehen und eine lineare Abhängigkeit besteht. Eine Kollinearität kann zu einer Instabilität des Modells führen weshalb mittels der Funktion vif() diese überprüft wird. VIF steht hierbei für den Variance Inflation Factor.

```
          Alter     Betriebszugehoerigkeit                    Gehalt Zeit_seit_Gehaltserhoehung
       1.681139                   1.393092                  1.866625                   1.118834
Dummy_Bildung2            Dummy_Bildung3           Dummy_Bildung4                 Dummy_WLB1
       1.197257                   1.052574                  1.000000                   1.019205
     Dummy_WLB0
       1.020539
```

Abbildung 25: Kollinearität Modell 1

Wie in Abbildung 25 zu sehen, weist keine Variable eine Kollinearität größer 5 (größer als 5 oder 10 deutet auf Multikollinearität hin [19]) vor, weshalb keine lineare Abhängigkeit zwischen den Variablen besteht. Somit muss zusätzlich zu der aufgrund der nicht Signifikanz ausgeschlossenen Variable Dummy_Bildung4 keine weitere aus dem Modell entfernt werden.

Der Data Scientist wendet nun solange die Backward-Selection an, also entfernt nicht signifikante Variablen und überprüft die Kollinearität, bis das perfekte logistische Regressionsmodell gefunden wurde. Bei diesem sind alle Variablen hoch signifikant und haben p-Werte von kleiner 0,05 wodurch angenommen werden kann, dass es eine signifikante Beziehung von diesen erklärenden Variablen zu der zu erklärenden Zielvariable gibt.

```
Call:
glm(formula = Abwanderung ~ Alter + Betriebszugehoerigkeit +
    Gehalt + Dummy_Bildung2 + Dummy_Bildung3, family = binomial(),
    data = train_data)

Coefficients:
                      Estimate   Std. Error z value Pr(>|z|)
(Intercept)         -1.509135856  0.177821182  -8.487   <2e-16 ***
Alter               -0.146096335  0.004324959 -33.780   <2e-16 ***
Betriebszugehoerigkeit -0.202663715  0.008672339 -23.369   <2e-16 ***
Gehalt               0.000158421  0.000004615  34.328   <2e-16 ***
Dummy_Bildung2       0.976991171  0.088901030  10.990   <2e-16 ***
Dummy_Bildung3       6.059135247  0.438684155  13.812   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Abbildung 26: Das beste logistische Regressionsmodel nach der Backward-Selection

Durch die Backward-Selection kann das beste Modell gefunden werden, welches in Abbildung 26 dargestellt ist. Es bleiben die Variablen Alter, Betriebszugehörigkeit, Gehalt, Dummy-Variable Bildung 2 und Dummy-Variable Bildung 3 (= Bildungsabschluss) übrig.

Dies bedeutet, dass diese vier erklärenden Variablen eine tatsächliche signifikante Beziehung zu der zu erklärenden Zielvariable Abwanderung haben.

Zusätzlich zeigt Abbildung 26 nähere Eigenschaften des Modells in der Spalte „Estimate“. Diese Werte beschreiben die β-Werte des Modells mit welchen die Variablen multipliziert werden. Diese Werte werden in folgender Tabelle 6 mit den zuvor festgelegten β-Werten, welche für die Simulation der Zielvariablen nötig waren, gegenübergestellt.

Variable	β-Wert des Modells	β-Wert real
Intercept	-1,51	1
Alter	-0,15	-0,15
Betriebszugehörigkeit	-0,20	-0,20
Gehalt	0,00016	0,00015
Dummy_Bildung2	0,98	-1
Dummy_Bildung3	6,06	4

Tabelle 6: Gegenüberstellung β-Werte

Hierbei lässt sich feststellen, dass lediglich der Intercept und die Dummy-Variable Dummy_Bildung2 ein falsches Vorzeichen im Modell erhalten, wobei der Betragswert der beiden Variablen nahezu identisch zu den realen β-Werten ist. Die Differenz der Vorzeichen gleicht allerdings die Dummy-Variable Dummy_Bildung3 aus, da diese einen β-Wert von 6,06 im Modell vorweist obwohl der reale β-Wert 4 ist. Die übrigen β-Werte von den Variablen Alter, Betriebszugehörigkeit und Gehalt stimmen im Modell mit den realen β-Werten überein. Die β-Werte des Modells decken die zuvor definierten β-Werte gut ab.

Die vorhergesagte Modellwahrscheinlichkeit für eine Abwanderung einer Person kann mit Hilfe einer S-Kurve veranschaulicht werden.

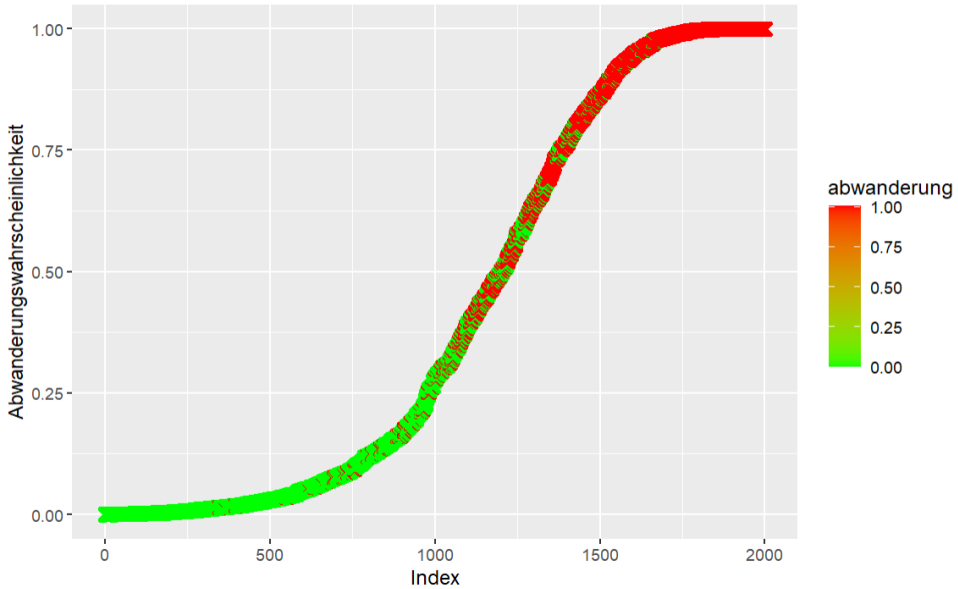


Abbildung 27: Modellwahrscheinlichkeit für eine Abwanderung Data Scientist

3.5 Evaluation

In der fünften Phase „Evaluation“ wird das trainierte Modell bewertet. Hierbei wird untersucht ob die Modellqualität ausreicht um der Zielsetzung der Datenanalyse gerecht zu werden [15]. Diese Phase ist zudem die abschließende Phase der Betrachtung innerhalb dieser Simulationsstudie.

Eine gängige Metrik zur Bewertung eines logistischen Regressionsmodell ist der Mean Squared Error (MSE). Der MSE beschreibt die durchschnittliche quadratische Abweichung und berechnet für jeden Datenpunkt die quadratische Differenz zwischen dem vorhergesagten und dem tatsächlichen Wert. Dabei bedeutet ein niedriger MSE-Wert, dass die Vorhersagen des Modells im Durchschnitt nah an den tatsächlichen Werten liegen.

Der Data Scientist erreicht bei seinem besten Modell (Abbildung 26) einen MSE von 0,0790 sprich ~0,08. Zum Vergleich: das Rauschen welches bei der Erstellung der Zielvariable entstand lag bei 0,0804 sprich ~0,08 was wiederum dem geringsten MSE des Modells entsprechen sollte. Demnach bildet der MSE des Modells das Rauschen der Daten sehr gut ab.

Neben dem MSE kann eine Konfusionsmatrix zur Bewertung des Modells herangezogen werden. Eine Konfusionsmatrix ist eine Tabelle welche die Anzahl der richtig und falsch klassifizierten Ausprägungen darstellt. Dabei hat die Konfusionsmatrix vier Einträge:

- True Positive (TP): Ausprägungen die korrekt als positiv (1) klassifiziert wurden
- True Negative (TN): Ausprägungen die korrekt als negativ (0) klassifiziert wurden
- False Positive (FP): Ausprägungen die fälschlicherweise als positiv (1) klassifiziert wurden
- False Negative (FN): Ausprägungen die fälschlicherweise als negativ (0) klassifiziert wurden

In RStudio kann eine solche Konfusionsmatrix mittels der confusionsMatrix()-Funktion aus der library „caret“ erstellt werden. Die Konfusionsmatrix liefert eine tabellarische Darstellung der Leistung eines Klassifikationsmodell wie in Tabelle 7 ersichtlich wird. Zuvor wurden mittels der ifelse()-Funktion und einem Cutoff-Wert von 0,5 die Vorhersagen für eine Abwanderung auf 1 und 0 verteilt. Bei einer Vorhersage von größer oder gleich 0,5 wurde der Wert 1 vergeben, andernfalls der Wert 0. Der Cutoff-Wert entscheidet demnach ab welchem Schwellenwert eine Vorhersage als positiv oder negativ betrachtet wird. Bei der Festlegung des Cutoff-Werts kann die ROC-Kurve (Receiver Operating Characteristics) Hilfe leisten, diese wird im weiteren Verlauf erläutert.

Konfusionsmatrix	**Person ist tatsächlich nicht abgewandert (False)**	**Person ist tatsächlich abgewandert (True)**
Modell sagt keine Abwanderung vorher (False)	1063	102
Modell sagt eine Abwanderung vorher (True)	117	718

Tabelle 7: Konfusionsmatrix

Zusätzlich zur Konfusionsmatrix werden noch verschiedene Qualitätsmetriken für das Modell berechnet. Unteranderem die:

- Accuracy (Genauigkeit): gibt an wie viele der Vorhersagen insgesamt korrekt waren
- Precision (Präzision): gibt an wie viele der positiv vorhergesagten Ausprägungen tatsächlich positiv sind
- Sensitivity (Sensitivität) gibt an wie viele der tatsächlich positiven Ausprägungen korrekt als positiv vorhergesagt wurden
- Specificity (Spezifität): gibt an wie viele der tatsächlich negativen Ausprägungen korrekt als negativ vorhergesagt wurden

Das Modell erreicht eine Accuracy = 0,8905, eine Precision = 0,8599, eine Sensitivity = 0,9008 und eine Specificity = 0,8756.

Außerdem kann der ROC die Leistung und Qualität eines Modells grafisch abbilden (Abbildung 28). Die Kurve bietet einen Überblick über die Beziehung zwischen Sensitivity und Specificity und gibt so eine grafische Übersicht über den Verlauf der verschiedenen Klassifikationsschwellen wodurch eine passende Schwelle bzw. Cutoff-Wert festgelegt werden kann. Bei der Wahl des Cutoff-Werts können verschiedene Vorgehen gewählt werden. Beispielsweise ein ausgewogenes Verhältnis zwischen Sensitivity und Specificity oder je nach Anwendungsfall mehr Wert auf die Minimierung von „False Positive“-Ausprägungen oder „False Negative“-Ausprägungen gelegt werden. Mit Hilfe des Bayes-Klassifikator p = 0,5 kann gezeigt werden, dass ein Cutoff-Wert von 0,5 zur kleinstmöglichen Fehlerrate führt, weshalb dieser Wert auch in der vorliegenden Simulationsstudie angewendet wurde [20]. Die Fläche unterhalb dieser ROC-Kurve wird als Area Under the Curve (AUC) bezeichnet. Der AUC-Wert liegt zwischen 0 und 1 und kann als Maß für die Modellgüte verwendet werden. Für das erstellte Modell des Data Scientist liegt der AUC bei 0,9588.

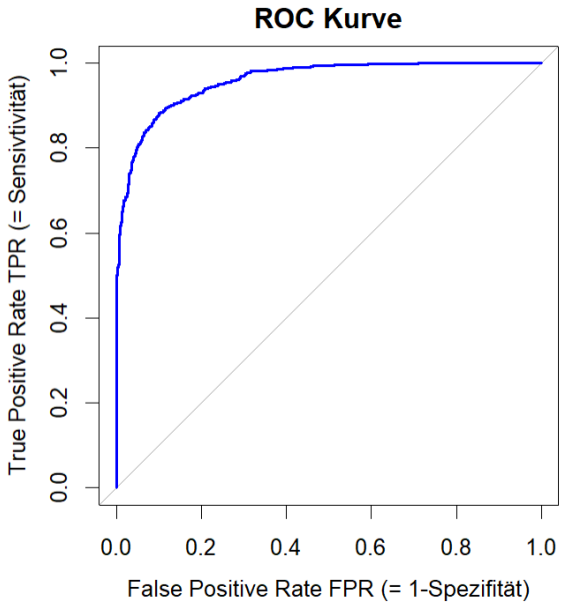


Abbildung 28: ROC-Kurve

Durch die in diesem Unterkapitel 3.5 Evaluation vorgestellten Entscheidungskriterien kann der Data Scientist unterschiedliche Modelle bewerten und Vergleiche zu weiteren Modellen ziehen.

4 Analysen zur optimalen Modellflexibilität

Folgendes Kapitel beschäftigt sich mit der Anwendung des KNN-Algorithmus auf den vorliegenden Datensatz des Data Scientist mit N = 10.000 Elementen aus Kapitel 3. KNN bedeutet „k-Nearest Neighbors“ und ist ein weit verbreiteter Algorithmus im Bereich des maschinellen Lernens und wird vor allem bei Klassifikations- und Regressionsaufgaben eingesetzt.

Für die Klassifikation einer neuen Eingabe (hier: entsprechender Datensatz zu einer Person) sucht der KNN-Algorithmus die „k“ Trainingspunkte, die dieser Eingabe am nächsten sind. Die Mehrheit der Klassenzugehörigkeiten (hier: Ausprägung 1 oder 0) dieser k-nächsten Nachbarn bestimmen die Vorhersage für die Eingabe. Da der Algorithmus anfällig gegenüber Ausreißern ist, muss das optimale k festgelegt werden.

Für den vorliegenden Fall der Simulationsstudie wird der Datensatz mit verschiedenen k-Werten analysiert, wobei sich k durch $k \in \{1, 2, \dots, 10, 20, \dots, 90, 100, 200, 300, 400\}$ definiert.
Bevor KNN trainiert werden kann, muss jedoch sichergestellt werden das ausschließlich metrische Variablen verwendet werden. Dies ist von Bedeutung, da der KNN-Algorithmus Abstände zwischen Datenpunkten berechnet um die Nähe zu bestimmen. Da nicht-metrische Variablen keine messbaren Abstände vorweisen, kann dies zu unsinnigen Schlussfolgerungen führen. Deshalb nutzt der Data Scientist lediglich die metrischen Variablen (Alter, Betriebszugehörigkeit, Gehalt und Zeit seit Gehaltserhöhung) des ihm vorliegenden Datensatzes. Um die Skalen der metrischen Variablen besser vergleichen zu können werden diese vor Anwendung von KNN noch skaliert um sicherzustellen, dass alle Variablen einen vergleichbaren Einfluss haben.

Bei der Anwendung von KNN für die oben definierten k-Werte kann der entsprechende Trainingsdaten-MSE und Testdaten-MSE berechnet werden. Dieser Verlauf wird in nachfolgender Abbildung 29 dargestellt. Die Abbildung veranschaulicht die Entwicklung der Testdaten-MSEs in blau und der Trainingsdaten-MSEs in rot. Daraus lässt sich erkennen, dass sich die beiden MSE-Werte für die k-Werte größer gleich 100 kaum unterscheiden. Doch je kleiner der k-Wert ab k = 100 wird, desto mehr unterscheiden sich die beiden MSE-Werte. Dies wird vor allem bei k = 30 und k = 20 deutlich. Dort strebt der Testdaten-MSE gegen unendlich und der Trainingsdaten-MSE gegen 0. Ableitend lässt sich feststellen, dass ein zu kleiner k-Wert anfälliger für das Rauschen oder die Ausreißer des Modells ist. Gleichzeitig kann ein zu klein gewähltes k zu Overfitting führen, wodurch das Modell möglicherweise zu stark auf den Datensatz reagiert. Das Overfitting wird in Abbildung 29 durch den Verlauf des Trainingsdaten-MSE deutlich, welcher bei minimalen k-Werten gegen 0 strebt und somit eine „perfekte Vorhersage“ bzw. ein „perfektes Modell“ beschreibt. Wiederum kann ein zu großes k zu Underfitting führen was eine zu starke Generalisierung des Modells aussagt welches dadurch beispielsweise Datenmuster und -strukturen übersieht da es durch die große Anzahl an k-Werten, die für die Klassifizierung verglichen werden, eine generalisierte Vorhersage trifft.

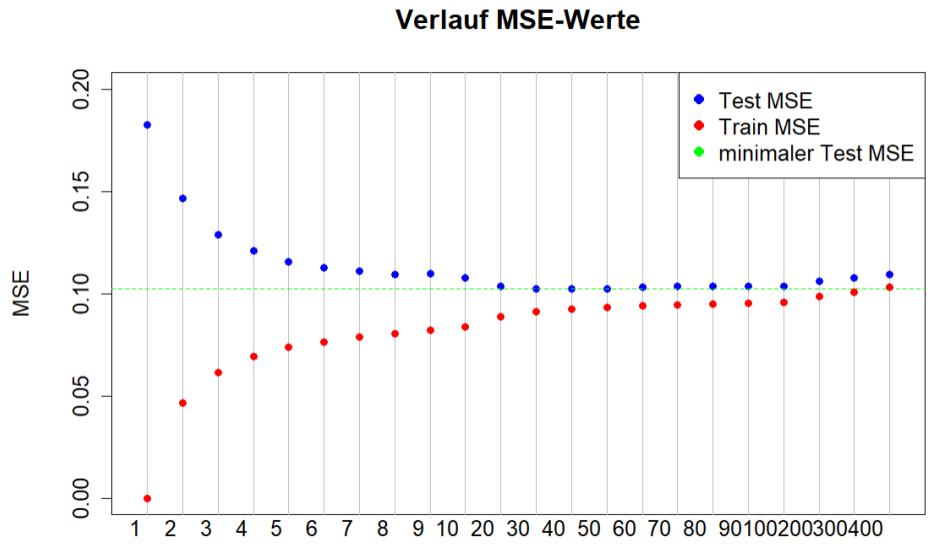


Abbildung 29: Verlauf MSE-Werte in Abhängigkeit zu k

Um das optimale k festzulegen werden die Testdaten-MSEs für die verschiedenen k-Werte verglichen und der minimale Testdaten-MSE gesucht. Für die vorliegenden Daten wird der minimale Testdaten-MSE = 0,1024 bei einem k-Wert von k = 50 erreicht. Dieser Wert wird zudem in Abbildung 29 durch die gestrichelte grüne Linie repräsentiert.

Letztlich kann der minimale MSE des KNN-Algorithmus mit dem minimalen MSE des logistischen Regressionsmodells und dem Rauschen der Daten verglichen werden.

- **Rauschen = 0,0804**
- **Logistisches Regressionsmodell = 0,790**
- **KNN (k_{opt} = 50) = 0,1024**

Es lässt sich feststellen, dass der minimale MSE der logistischen Regression von ~0,08 näher an dem Rauschen der Daten ~0,08 liegt als der minimale MSE des KNN-Algorithmus ~0,10. Die logistische Regression erzielt auf Basis der für den Data Scientist verfügbaren Daten ein besseres Ergebnis bzw. besser passende Vorhersagen als der KNN-Algorithmus.

Der Data Scientist kann dem Personalchef der Pay Solutions GmbH demnach das logistische Regressionsmodell übergeben mit welchem der Personalchef zukünftig abwanderungsgefährdete Personen identifizieren kann um somit die Fluktuation im Unternehmen nachhaltig zu senken.

Literaturverzeichnis

[1] Onlyfy, Fluktuation – Definition und Arten, [online] https://onlyfy.com/de/glossar/fluktuation/#welche-arten-der-mitarbeiterfluktuation-gibt-es? [abgerufen am 24.01.2024].

[2] Statista Altersstruktur in Unternehmen, Altersstruktur in Unternehmen in Deutschland nach Branchen und Anzahl der Beschäftigten im 2. Quartal 2023, [online] https://de.statista.com/statistik/daten/studie/1404501/umfrage/altersstruktur-in-unternehmen-nach-branchen-und-groesse/ [abgerufen am 04.01.2024]

[3] Demografie Portal, Renteneintrittsalter, [online] https://www.demografie-portal.de/DE/Fakten/renteneintrittsalter.html [abgerufen am 04.01.24]

[4] Bundesinstitut für Berufsbildung, Datenreport, [online] https://www.bibb.de/datenreport/de/2019/101256.php [abgerufen am 04.01.24]

[5] Statistisches Landesamt Baden-Württemberg, Mehr als ein Drittel der Beschäftigten ist fünfzig Jahre und älter, [online] https://www.statistik-bw.de/Presse/Pressemitteilungen/2021102 [abgerufen am 04.01.24]

[6] Statistisches Bundesamt, Dauer der Beschäftigung beim aktuellen Arbeitgeber, [online] https://www.destatis.de/DE/Themen/Arbeit/Arbeitsmarkt/Qualitaet-Arbeit/Dimension-4/dauer-beschaeftigung-aktuell-Arbeitgeber.html [abgerufen am 04.01.23]

[7] Deutsche Hochschulwerbung, Infografik Absolventen, [online] https://www.hochschulwerbung.de/wp-content/uploads/2017/04/Infografik_Absolventen.pdf [abgerufen am 04.01.24]

[8] Spiegel, Frauen promovieren schneller – und seltener, [online] https://www.spiegel.de/panorama/bildung/frauen-promovieren-schneller-und-seltener-a-8889b231-c1c1-42f1-b39d-2141981c9b90 [abgerufen am 04.01.24]

[9] Statista Verteilung beruflicher Bildungsabschluss, Verteilung der Bevölkerung in Deutschland nach beruflichem Bildungsabschluss im Jahr 2022, [online] https://de.statista.com/statistik/daten/studie/3276/umfrage/bevoelkerung-nach-beruflichem-bildungsabschluss/ [abgerufen am 04.01.24]

[10] Destatis, Gehaltsvergleich 2022: Neben dem Beruf ist der Bildungsabschluss entscheidend, [online] https://www.destatis.de/DE/Presse/Pressemitteilungen/2023/05/PD23_200_62.html [abgerufen am 05.01.24]

[11] Workwise, Gehaltserhöhung - Wie viel und wann?, [online] https://www.workwise.io/karriereguide/karriere/gehaltserhoehung [abgerufen am 05.01.24]

[12] Academics, Gehaltserhöhung fordern, [online] https://www.academics.de/ratgeber/gehaltserhoehung-gehaltsverhandlung-tipps#subnav_gruende_fuer_eine_gehaltserhoehung_wann_ist_sie_angemessen [abgerufen am 05.01.24]

[13] Statista Leben und Arbeiten, Länder mit der besten Work-Life Balance, [online] https://de.statista.com/infografik/13073/laender-mit-der-besten-work-life-balance/ [abgerufen am 05.01.24]

[14] IBM, CRISP-DM-Übersicht, [online] https://www.ibm.com/docs/de/spss-modeler/saas?topic=dm-crisp-help-overview [abgerufen am 11.01.24]

[15] Datasolut, CRISP-DM, [online] https://datasolut.com/crisp-dm-standard [abgerufen am 11.01.24]

[16] Haufe Online Redaktion, Mitarbeiterfluktuation mit Onboarding senken, [online] https://www.haufe.de/personal/hr-management/fluktuation-wechselbereitschaft-der-arbeitnehmer-steigt_80_193940.html#:~:text=Eine%20hohe%20Fluktuation%20ist%20f%C3%BCr,Vakanz%20und%20Neubesetzung%20der%20Stelle. [abgerufen am 11.01.24]

[17] Statista Lexikon, Definition Korrelation, [online] https://de.statista.com/statistik/lexikon/definition/77/korrelation/ [abgerufen am 12.01.24]

[18] simplilearn, What is Backward Elimination Technique In Machine Learning, [online] https://www.simplilearn.com/what-is-backward-elimination-technique-in-machine-learning-article [abgerufen am 16.01.24]

[19] StatistikGuru, Multiple lineare Regression, [online] https://statistikguru.de/spss/multiple-lineare-regression/voraussetzung-multikollinearitaet.html [abgerufen am 16.01.24]

[20] Statistical Learning, Theoretische Ergänzungen von Prof. Schürle, Kapitel: Logistische Regression, S. 62