DER INTERAKTIVE GEHALTS-VERGLEICH – EINE PROFILBASIERTE SCHÄTZUNG VON VERDIENSTEN

Anja Überschaer

↘ **Schlüsselwörter:** Gehaltsrechner — Verdienste nach Berufen — Regression — Verdiensterhebung — Lohnschätzung

ZUSAMMENFASSUNG

Das Statistische Bundesamt entwickelt sich zu einem digitalen, nutzerorientierten und innovativen Informationsdienstleister weiter. Dazu gehören auch neue Produkte, die Informationen nutzungsfreundlich bereitstellen – wie der Gehaltsvergleich. Diese interaktive Webanwendung ermöglicht es den Nutzerinnen und Nutzern, sich auf Basis von individuellen Angaben ein Gehalt schätzen zu lassen. Der Gehaltsrechner ist leicht zu bedienen und legt den Fokus auf individuelle Profile. Dadurch ermöglicht die Anwendung die zielgruppenorientierte und bedarfsgerechte Aufbereitung von Verdienstdaten und steigert somit deren Nutzen.

Der Aufsatz beschreibt den konzeptionellen Aufbau des Gehaltsrechners sowie die Methodik der Schätzung des individuellen Bruttomonatsverdienstes.

↘ ***Keywords:*** *salary calculator — earnings by occupation — regression — earnings survey — wage estimate*

ABSTRACT

The Federal Statistical Office is developing into a digital, user-oriented and innovative information provider. This also includes creating new products which provide information in a user-friendly form, for example the salary calculator for salary comparison. That interactive web application allows users to have a salary estimated on the basis of individual information. The salary calculator is easy to use and focuses on individual profiles. It enables earnings data to be prepared in a target-group-oriented and needs-based manner, thus increasing its usefulness.

This paper describes the conceptual design of the salary calculator and the methodology used to estimate individual gross monthly earnings.

Dr. Anja Überschaer

ist im Referat „Verdienste, Überschuldung“ des Statistischen Bundesamtes tätig. Sie hat im Rahmen eines Doppelmasterprogramms Betriebswirtschaftslehre und Volkswirtschaftslehre studiert. Ihr Tätigkeitsbereich umfasst methodische Fragestellungen rund um die neue Verdienststatistik und sie hat die interaktive Anwendung „Gehaltsvergleich“ entwickelt.

1

Einleitung

Viele Personen interessieren sich für Angaben zu durchschnittlichen Verdiensten. Dabei geht es oft um die Einordnung und Bewertung des eigenen Verdienstes und um Jobbewerbungen, wobei der Beruf, der Ausbildungsabschluss, die Branche und weitere Aspekte eine Rolle spielen. Um dieser Nachfrage gerecht zu werden, ist im Statistischen Bundesamt der interaktive Gehaltsvergleich[1] entwickelt worden. Er richtet sich vorwiegend an Privatpersonen, die beispielsweise aufgrund von bevorstehenden Gehaltsverhandlungen, Bewerbungsgesprächen oder rein aus Interesse eine spezifische Bruttomonatslohnschätzung erhalten möchten.

Die Veröffentlichung durchschnittlicher Verdienste in den Fachveröffentlichungen und in der Datenbank GENESIS-Online des Statistischen Bundesamtes sind für Nutzerinnen und Nutzer, die sehr spezifische und individuelle Informationen benötigen, nicht optimal. So stehen unter anderem Geheimhaltungsvorschriften einer spezifischen Nutzung entgegen, da konkrete Angaben zu Verdiensten nach Berufen, zusätzlich gegliedert nach Bundesländern, Wirtschaftszweigen oder Ausbildungsabschlüssen, nicht veröffentlicht werden dürfen. Zudem wäre eine solche spezifische Aufschlüsselung der Verdienste in Tabellenform nicht nutzungsfreundlich darstellbar. Die Methodik der Regression ermöglicht, dass der Gehaltsrechner Verdienstinformationen, die den Bedürfnissen der Zielgruppe entsprechen, individuell und profilbasiert bereitstellt. Der interaktive Gehaltsvergleich baut durch seine leichte Zugänglichkeit potenzielle Hemmschwellen ab und überzeugt durch die spielerische Aufbereitung der Informationen. Das interaktive Onlinetool ist durch diese Kombination an Eigenschaften – individuelle Informationen, leichte Zugänglichkeit und spielerische Handhabung – auch für Laien attraktiv.

Beim interaktiven Gehaltsvergleich sind verschiedene gehaltsbestimmende Merkmale vorgegeben, zum Beispiel Beruf, Ausbildungsabschluss oder Branche, anhand derer sich der Bruttomonatsverdienst schätzen lässt. Dazu gibt eine Person die für sie passende Ausprägung dieser Merkmale an und erhält eine individuelle Schätzung des Bruttomonatsverdienstes. Der Gehaltsvergleich ermöglicht somit den Nutzerinnen und Nutzern einen schnellen und unkomplizierten Überblick über den (geschätzten) Verdienst, unter der Berücksichtigung von persönlichen und arbeitsplatzspezifischen Merkmalen. Dabei handelt es sich um regressionsbasierte Schätzungen und nicht um Mittelwerte aus der Erhebung an sich, wie sie die Fachveröffentlichungen enthalten. Mit dem Gehaltsvergleich ist es dabei auch möglich, Veränderungen im (geschätzten) Bruttomonatslohn zu beobachten, die entstehen, wenn einzelne verdienstbestimmende Merkmalsausprägungen modifiziert werden.

Der Artikel beschreibt zunächst die Datengrundlage des interaktiven Gehaltsvergleichs (Kapitel 2) und wie die Aufbereitung der Daten erfolgt (Kapitel 3). Kapitel 4 stellt das entwickelte Regressionsmodell vor, während Kapitel 5 die Variablen der Regression ausführlich erläutert. Die Qualität des Modells wird in Kapitel 6 detailliert untersucht. Der Beitrag schließt mit einem Fazit und einem Ausblick auf eine mögliche jährliche Aktualisierung des Gehaltsrechners mit den jeweiligen Daten des Monats April der neuen digitalen Verdiensterhebung.

2

Datengrundlage

Der interaktive Gehaltsvergleich basiert zunächst auf den Daten der Verdienststrukturerhebung, die alle vier Jahre Daten zu den Verdiensten erfasst hat, zuletzt für das Berichtsjahr 2018. An dieser Erhebung nahmen etwa 60 000 repräsentativ ausgewählte Betriebe teil und lieferten unter anderem Angaben zum Bruttomonatslohn von etwa 1 Million abhängig Beschäftigten.[2]

Die Verdienstdaten dieser Erhebung sind untergliedert nach Wirtschaftszweigen und Regionen sowie persönlichen Angaben über die Beschäftigten, wie Geschlecht, Geburtsjahr, die Dauer der Unternehmenszugehörigkeit, Beruf und Ausbildungsabschluss. Zudem wurden Merkmale zum Beschäftigungsverhältnis erhoben, beispielsweise Angaben zu Tarifvertrag, Art der Beschäftigung (befristet oder unbefristet) und der Umfang des

1 Eine Beschreibung der Anwendung und ihrer Methodik befindet sich unter www.destatis.de.

2 Zusätzliche Informationen zur Verdienststrukturerhebung 2018 enthält der Qualitätsbericht (Statistisches Bundesamt, 2018).

Urlaubsanspruchs. Die Verdienststrukturerhebung ermöglicht damit Aussagen über die Verteilung der Verdienste sowie über den Einfluss wichtiger, die individuelle Verdiensthöhe bestimmender Faktoren.

Der Gehaltsrechner nutzt Angaben zu den einzelnen Beschäftigten, zum Beispiel den Verdienst. Diese Angaben stammen aus der Lohnabrechnung der Betriebe, dadurch ist eine hohe Datenqualität gewährleistet. Die auskunftgebenden Betriebe sind gesetzlich verpflichtet, vollständige und korrekte Angaben zu machen. Zudem erfolgt eine Plausibilisierung der Daten durch die Statistischen Ämter der Länder. Auch dies sichert eine sehr genaue und zuverlässige Datenbasis.

Künftig wird der Gehaltsvergleich auf Basis der neuen monatlichen Verdiensterhebung ab 2022 jährlich aktualisiert. Bei der Konzeption des Regressionsmodells wurde daher darauf geachtet, nur Merkmale zu verwenden, die auch in den Daten der neuen Verdiensterhebung vorliegen werden. Dies soll einen möglichst reibungslosen Übergang beim Wechsel der Datengrundlage zur Berechnung des Gehaltsvergleichs sicherstellen.

3

Datenaufbereitung

Aus Gründen der Vergleichbarkeit wurden Teilzeitbeschäftigte bei der Schätzung des Regressionsmodells ausgeschlossen: Deren Bruttomonatsverdienst hängt stark von der Zahl der Arbeitsstunden ab. Branchen mit einem vergleichsweise hohen Anteil an Teilzeitbeschäftigten (die in der Regel einen niedrigeren Bruttomonatsverdienst haben als Vollzeitbeschäftigte) würden die durchschnittlichen Monatsverdienste verzerren. Zudem wäre ein Vergleich der Verdienste über verschiedene Branchen hinweg erschwert, sofern diese einen unterschiedlich hohen Anteil an Teilzeitbeschäftigten haben. Da mehr Frauen als Männer in Teilzeit arbeiten, gilt Gleiches auch für die Verdienste von Männern und Frauen. Auch hier würde die Einbeziehung von Teilzeitbeschäftigten dazu führen, dass die durchschnittlichen Verdienste von Frauen systematisch unterschätzt würden. Das Regressionsmodell basiert daher auf den Verdiensten von etwa 600 000 Personen, die in Vollzeit arbeiten und keine Auszubildenden oder Personen in Altersteilzeit sind.

Im Zuge der Datenaufbereitung wurde auch auf mögliche Ausreißer eingegangen. Nach Aguines und anderen (2013) sind Ausreißer Datenpunkte, die sich stark von anderen Datenpunkten unterscheiden. Den Autoren zufolge kann es sich dabei beispielsweise um Fehler in den Daten handeln (wie Tippfehler) oder um „besondere Fälle“. Letztere unterscheiden sich in ihrer Ausprägung zwar von anderen Fällen, sind aber dennoch korrekt. Eine Entfernung von Ausreißern, bei denen es sich nicht um echte Fehler handelt, hat zur Folge, dass Informationen eines (korrekten) Datenpunkts nicht genutzt werden können. Dies kann dazu führen, dass die Stichprobe die Grundgesamtheit nur noch verzerrt darstellt. Da aufgrund der Plausibilisierung der Daten durch die Statistischen Ämter der Länder falsche Angaben (gerade bei Extremfällen) sehr unwahrscheinlich sind, würde dies lediglich die Varianz künstlich reduzieren. Deshalb werden alle Werte zur Schätzung des Regressionsmodells verwendet.

4

Das Regressionsmodell

Um das Regressionsmodell zu entwickeln, wurde zunächst basierend auf theoretischen Modellen (wie dem Lohnmodell von Mincer, 1974) und mithilfe der vorhandenen Daten Folgendes ermittelt: Welche Merkmale beeinflussen den Verdienst einer Person und wie stark ist ihr Einfluss? Dies ist möglich, da in den Daten sowohl die Verdienste als auch die gehaltsbestimmenden Merkmale hinterlegt sind. Mithilfe des Regressionsmodells, das die ausgewählten Einflussfaktoren und die Größe ihres Einflusses umfasst, können anschließend Schätzungen ausgegeben werden. Die ausgegebenen Zahlenwerte des Gehaltsvergleichs sind folglich Prognosewerte einer Regressionsgleichung. Die Regression wird dabei nicht etwa zur Laufzeit durchgeführt, sondern zuvor bei der Programmierung des Gehaltsvergleichs. Im Gehaltsrechner sind allein die ermittelten Regressionskoeffizienten hinterlegt. Für jede Merkmalsausprägung liegt ein merkmalspezifischer Prognosewert (= Regressionskoeffizient) vor. Die finalen Prognosewerte ergeben sich durch eine Aufsummierung der von der Person individuell ausgewählten merkmalspezifischen Regressionskoeffizient (bei metrischen Variablen erfolgt zunächst die Multiplikation des hinterlegten Prognosewerts mit

der von der Nutzerin oder vom Nutzer angegebenen Zahl) und der Konstanten. Der so ermittelte Wert wird dann der anfragenden Person ausgegeben.

Da es sich bei der Verdienststrukturerhebung um eine Stichprobe mit komplexem Design (zweistufige, geschichtete Stichprobe) handelt, ist dies auch bei der Auswertung zu berücksichtigen. In der vom Statistischen Bundesamt genutzten Statistiksoftware SAS steht hierzu die Prozedur SURVEYREG zur Verfügung (Finke, 2010). Das SURVEYREG-Verfahren führt Regressionsanalysen für Stichprobenerhebungsdaten durch. Das Verfahren kann mit komplexen Stichprobenentwürfen für Umfragen umgehen, einschließlich Entwürfen mit Schichtung, Clustering und ungleicher Gewichtung (SAS Institute Inc., 2013).

5

Variablen der Regression

5.1 Die abhängige Variable Bruttomonatsverdienst

Die abhängige Variable ist der Bruttomonatsverdienst (Gesamtbruttoentgelt gemäß § 1 Absatz 2 Nummer 2c Entgeltbescheinigungsverordnung) abzüglich sonstiger Bezüge wie 13. Monatsgehalt, Urlaubsgeld oder Weihnachtsgeld. Die Entscheidung fiel für die Ausgabe von Monatsverdiensten, da die Verdienste in der Lohnabrechnung der Betriebe in Monatsgehältern vorliegen und auch so erhoben werden. Darüber hinaus handelt es sich bei den monatlichen Bruttoverdiensten um sogenannte laufende Bezüge. Anhand dieser lässt sich nach Abzug von Steuern und Abgaben (die individuell unterschiedlich sind, wie die Beiträge zur Krankenversicherung) das monatlich verfügbare Erwerbseinkommen bestimmen. Somit stellt der Bruttomonatsverdienst eine wichtige Entscheidungsgrundlage dar. Außerdem ist den meisten Personen der Bruttomonatsverdienst aufgrund der gesetzlich vorgeschriebenen Entgeltbescheinigung geläufig. Folglich lassen sich Angaben dazu gut einordnen und leicht vergleichen.

Zudem unterliegen Verdienstdaten einer rechtsschiefen Verteilung (Mincer, 1974). Dies kann durch Logarithmierung korrigiert werden (Fields, 2010), was eine Annäherung an die Normalverteilung ermöglicht. Aus diesem Grund wird in dem Regressionsmodell der logarithmierte Bruttomonatsverdienst als abhängige Variable verwendet:

$$lnY_i = \beta_0 + \sum_{j=1}^{n} \beta_j X_{ij} + e_i$$

Dabei sind:

$ln\ Y_i$ = logarithmierter Bruttomonatsverdienst einer Person i

β_j = Regressionskoeffizient eines Merkmals j

β_0 = Regressionskonstante

X_{ij} = beobachtetes Merkmal j einer Person i

e = Störterm

5.2 Die unabhängigen Variablen

Die nachfolgend beschriebenen Variablen (beziehungsweise Merkmale) wurden in das Regressionsmodell mit aufgenommen, um das Gehalt zu schätzen. Hierbei handelt es sich sowohl um arbeitsplatzspezifische Merkmale (wie den Beruf) als auch um persönliche Merkmale (wie die Ausbildung). Die Auswahl der Merkmale orientiert sich am (bereinigten) Gender Pay Gap (Finke, 2010).

Interessen von Nutzerinnen und Nutzern können sehr unterschiedlich sein. Beispielsweise sucht eine Person Informationen zu einer konkreten Stelle, eine andere möchte jedoch lediglich eine grobe Orientierung über Verdienstmöglichkeiten in einem Beruf. Daher müssen nicht für alle Merkmale zwingend Angaben vorgenommen werden. Während für Beruf, Branche, Ausbildung und Bundesland eine Auswahl zu treffen ist, gibt es bei anderen Merkmalen auch die Option, „keine Angabe“ auszuwählen. In diesem Fall werden für die betroffenen Merkmale Annahmen getroffen. Damit diese Annahmen für möglichst viele Nutzerinnen und Nutzer „passend“ sind, basieren sie auf Durchschnittswerten oder (im Rahmen der Anwendung) „üblichen“ Werten.

Pflichtangaben

Die Pflichtangaben sind immer anzugeben und sind Merkmale, die einen besonders starken Einfluss auf das Ergebnis haben. Diese Merkmale sind nicht übermäßig spezifisch und können auch von Personen angegeben werden, die sich nur grob orientieren möchten.

Beruf

Anhand der Stichwortliste der Bundesagentur für Arbeit, die auf dem Klassifikationsserver der Statistischen Ämter des Bundes und der Länder (klassifikationsserver.de) zur Verfügung steht, werden einzelne Berufsbezeichnungen den Berufsgattungen der Klassifikation der Berufe 2010 zugeordnet. Nähere Informationen zu den Klassifikationen der Berufe sind in Band 1 der Klassifikation der Berufe 2010 (Bundesagentur für Arbeit, 2011) zu finden.

Die Berufsgattung dient als Ausgangsbasis, um die Merkmale Berufsgruppe, Anforderungsniveau und Aufsichtsbeziehungsweise Führungsfunktion abzuleiten. In der Klassifikation der Berufe 2010 wird jeder Berufsgattung eine fünfstellige Zahl zugeordnet (5-Steller). Die ersten drei Stellen des 5-Stellers bezeichnen die Berufsgruppe. Die vierte Stelle gibt Aufschluss darüber, ob es sich um eine Aufsichts- beziehungsweise Führungsfunktion handelt. Umfasst eine berufliche Tätigkeit vorwiegend Aufsichts- oder Führungsaufgaben, wird sie entsprechend als Aufsichts- beziehungsweise Führungsposition definiert. Die letzte Stelle kann die Werte 1 bis 4 annehmen und bezieht sich auf das Anforderungsniveau des Arbeitsplatzes. Das Anforderungsniveau gibt an, wie komplex eine Tätigkeit ist. Es gibt vier Ausprägungen: (1) Helfer- und Anlerntätigkeiten, (2) fachlich ausgerichtete Tätigkeiten, (3) komplexe Spezialistentätigkeiten und (4) hoch komplexe Tätigkeiten.

Beispielsweise gehört die Berufsbezeichnung „Sekretär/in“ der Berufsgattung „Büro- und Sekretariatskräfte (ohne Spezialisierung) – Fachlich ausgerichtete Tätigkeiten“ an (5-Steller 71402). Diese wiederum ist eine Untergruppe der Berufsgruppe „Büro- und Sekretariat“ (3-Steller 714), wird nicht als Aufsicht- beziehungsweise Führungsfunktion eingeordnet (vorletzte Ziffer ungleich 9) und hat das Anforderungsniveau „fachlich ausgerichtete Tätigkeiten“ (Stufe 2).

Der Vorteil einer Aufsplittung des 5-Stellers liegt darin, dass die Nutzerinnen und Nutzer die Bestimmung des Anforderungsniveaus nicht selbst vornehmen müssen. Stattdessen erfolgt dies automatisch über den ausgewählten Beruf und dessen Zuordnung zu einem Anforderungsniveau. Darüber hinaus lässt sich so auch die Anzahl an Dummyvariablen im Regressionsmodell reduzieren.

Branche

In die Regression gehen auch die einzelnen Abteilungen der Branchen ein. Basierend auf der Klassifikation der Wirtschaftszweige, Ausgabe 2008 (WZ 2008) wird zunächst die Oberkategorie (= Abschnitt der WZ 2008) ausgewählt und dann die Unterkategorie (= Abteilung der WZ 2008) bestimmt. Beispielsweise kann als Oberkategorie „Verarbeitendes Gewerbe (Herstellung)“ und als Unterkategorie „Automobilindustrie (Herstellung von Fahrzeugen und Fahrzeugteilen)“ ausgewählt werden. Die Originalbezeichnungen aus der WZ 2008 wurden zum besseren Verständnis leicht angepasst. So wird zum Beispiel statt der Bezeichnung „Herstellung von Kraftwagen und Kraftwagenteilen“ die Formulierung „Automobilindustrie (Herstellung von Fahrzeugen und Fahrzeugteilen)“ verwendet.

Ausbildung

Die Ausbildung einer Person wird mit dem höchsten beruflichen Ausbildungsabschluss erfasst. Dieser umfasst sechs Kategorien: ohne beruflichen Ausbildungsabschluss, Abschluss einer anerkannten Berufsausbildung, Meister (beziehungsweise Techniker- oder gleichwertiger Fachschulabschluss), Bachelor, Master (beziehungsweise Diplom/Magister/Staatsexamen) und Promotion.

Bundesland

Die Nutzerinnen und Nutzer wählen zudem auch das Bundesland aus. Dies ermöglicht die Berücksichtigung von regionalen Verdienstunterschieden.

Wahlangaben

Für die Wahlmerkmale ist keine Angabe erforderlich. Werden diese nicht spezifiziert, wird ein Durchschnittswert hinterlegt.

Alter

Anhand des Alters wird approximativ die Berufserfahrung ermittelt und in die Analyse mit einbezogen. Dazu erfolgt eine näherungsweise Ermittlung über das Alter und die Ausbildung einer Person (Achatz und andere, 2005). Durch diese Form der Berechnung kann es theoretisch vorkommen, dass die Berufserfahrung einen negativen Wert annimmt (potenzielle Berufserfahrung = Alter – durchschnittliche Ausbildungsdauer). Für den Fall, dass das Alter unter der durchschnittlichen Ausbildungsdauer liegt, wird der Wert der Berufserfahrung jedoch auf null gesetzt. Die potenzielle Berufserfahrung wird zudem auch als quadrierter Wert in das Regressionsmodell mit aufgenommen, um den im Zeitverlauf verminderten Einfluss von Berufserfahrung auf den Verdienst abzubilden (Achatz und andere, 2005).

Befristung

Das Regressionsmodell berücksichtigt auch, ob eine Stelle befristet oder unbefristet ist. Gründe für eine Befristung können beispielsweise der zeitlich begrenzte Einsatz im Rahmen einer Projektstelle oder eine Elternzeitvertretung sein.

Anzahl Beschäftigte im Unternehmen

Durch die Aufnahme der Anzahl der Beschäftigten im Unternehmen in das Regressionsmodell wird die Größe des Unternehmens abgebildet. Um den Nutzerinnen und Nutzern die Angabe zu erleichtern, gibt es sechs Kategorien: weniger als 10 Beschäftigte, 10 bis 49 Beschäftigte, 50 bis 249 Beschäftigte, 250 bis 499 Beschäftigte, 500 bis 999 Beschäftigte sowie 1 000 und mehr Beschäftigte.

Tarifbindung

Das Regressionsmodell berücksichtigt auch, ob in dem Betrieb ein Tarifvertrag (Kollektivvertrag oder Firmentarifvertrag) gilt oder nicht.

Dauer der Unternehmenszugehörigkeit

Wie lange eine Person bereits in einem Unternehmen arbeitet, wird über die Dauer der Unternehmenszugehörigkeit erfasst. Da – ähnlich wie bei der Berufserfahrung – der positive Einfluss auf den Verdienst im Zeitverlauf geringer ausfällt, wird auch in diesem Fall der quadrierte Term in das Regressionsmodell mit aufgenommen.

Geschlecht

Das Geschlecht ist ebenfalls als unabhängige Variable im Regressionsmodell enthalten. Allerdings erfolgt aus Gründen der Transparenz keine Auswahl durch die Nutzerinnen und Nutzer. Stattdessen wird das Ergebnis für beide Gruppen ausgegeben. Die Beschränkung auf die binäre Ausgabe (Mann oder Frau, ohne die Angabe von „divers“) ist durch die unzureichende Datenbasis für die Gruppe „divers“ bedingt.

6

Qualität des Modells

Die Güte des Modells wurde zum einen mit klassischen Qualitätskriterien überprüft, wie dem (korrigierten) R^2. Zum anderen wurde der mittlere absolute prozentuale Fehler (mean absolute percentage error, MAPE) herangezogen, um eine Kreuzvalidierung durchzuführen. Des Weiteren werden in der Webanwendung Plausibilisierungsmaßnahmen ergriffen, um Falscheingaben zu unterbinden und eine ausreichende Anzahl an Datensätzen zu gewährleisten.

6.1 R^2 und korrigiertes R^2

Das R^2 (auch Bestimmtheitsmaß genannt) ist ein Gütemaß zur Beurteilung der Qualität eines Regressionsmodells und kann Werte zwischen 0 und 1 annehmen. Es gibt an, wie viel Prozent der Varianz der abhängigen Variable durch das Modell erklärt werden kann. Ein R^2 von 0 bedeutet, dass es keinen Zusammenhang zwischen den unabhängigen Variablen und der abhängigen Variablen gibt, während ein Wert von 1 (dies entspricht 100 %) einem perfekten Zusammenhang entspricht. Da das R^2 immer weiter ansteigt, je mehr Variablen in das

Modell aufgenommen werden, wird gerne das korrigierte R^2 zur Beurteilung herangezogen. Dieses „bestraft" zu komplexe Modelle und ist in der Regel niedriger als das R^2. Das für den Gehaltsvergleich verwendete Regressionsmodell hat ein R^2 von 0,66 und ein korrigiertes R^2 von ebenfalls 0,66. Folglich erklärt das Modell 66 % der Varianz (beziehungsweise Änderungen) der abhängigen Variablen, was für Modelle dieser Art eine gute Leistung ist. Dass das korrigierte R^2 nicht deutlich unter dem R^2 liegt, ist darüber hinaus ein erster Anhaltspunkt dafür, dass das Modell nicht zu viele (unnötige) Variablen umfasst. Weitere Informationen zur Überanpassung enthält der folgende Abschnitt 6.2.

6.2 Kreuzvalidierung

Da das Modell der Vorhersage von Verdiensten dient, ist seine Qualität auch diesbezüglich zu testen. Eine der größten Gefahren bei der Entwicklung von Vorhersagemodellen ist die Überanpassung (Shmueli, 2010). Bei Erstellung eines Modells mit vielen Variablen kann es passieren, dass das Modell übermäßig genau auf die aktuelle Stichprobe eingestellt wird (mit allen stichprobenbedingten Besonderheiten), aber neue Daten deutlich schlechter schätzt (Putka und andere, 2018). Das führt zu einer zu optimistischen Beurteilung der Güte des Modells. Ein überangepasstes Modell kann folglich zwar sehr gut die Daten erklären, mit denen es erstellt wurde, liefert aber bei neuen Daten deutlich schlechtere Schätzungen. Die Qualität eines Modells zur Vorhersage von Ergebnissen wird daher mittels Kreuzvalidierung bestimmt. Im einfachsten Fall unterteilt man die Stichprobe in zwei Teile und nutzt einen Teil der Daten, um das Modell zu erstellen, und den anderen Teil, um mit diesem Modell die Werte für die abhängige Variable zu schätzen. Ist die Schätzqualität in beiden Teilen der Stichprobe gleich gut, dann ist eine Überanpassung unwahrscheinlich und davon auszugehen, dass das Modell die Grundgesamtheit (und nicht nur die Stichprobe) abbildet (John/Roth, 1999).

Daher wurde das Modell zusätzlich noch einmal mit 90 % der Daten geschätzt (Trainingsdatensatz) und mit 10 % der Daten getestet (Validierungsdatensatz). Die Aufteilung erfolgte dabei zufallsbasiert. Zum Vergleich wurde der mittlere absolute prozentuale Fehler herangezogen. Dieses Abweichungsmaß entspricht der mittleren prozentualen Abweichung der geschätzten Werte von den erhobenen Werten (hier: die Bruttomonatsverdienste in Euro). Während im Trainingsdatensatz der mittlere absolute prozentuale Fehler 19,27 % beträgt, liegt er im Validierungsdatensatz bei 19,21 %. Dieser Unterschied ist nicht signifikant, eine Überanpassung ist daher unwahrscheinlich. Folglich kann das Modell zur Schätzung neuer Daten verwendet werden. Um für die Schätzung des finalen Modells nicht auf Informationen verzichten zu müssen, werden dafür alle Daten verwendet. Dies ist möglich, da die Modelle, die mit 90 % beziehungsweise 100 % der Daten geschätzt wurden, nahezu gleich sind und Überanpassung bereits ausgeschlossen wurde.

Darüber hinaus sagt der mittlere absolute prozentuale Fehler aus, dass die vom Modell geschätzten Bruttomonatsverdienste (in Euro) im Mittel etwa 20 % von den erhobenen (= wahren) Bruttomonatsverdiensten abweichen. Dabei ist zu bedenken, dass das Modell den mittleren Bruttomonatsverdienst schätzt. So wie die individuellen erhobenen Verdienste um den Mittelwert der erhobenen Verdienste streuen, streuen die individuellen erhobenen Verdienste auch um den geschätzten Mittelwert. Zudem lassen sich bei einer Schätzung nicht alle Einflussfaktoren perfekt abbilden. Beispielsweise wird erfasst, welchen Abschluss eine Person hat, aber nicht deren Notendurchschnitt. Folglich kann es auch innerhalb dieser Gruppe von Personen mit dem gleichen Abschluss zu Unterschieden im Verdienst kommen. Vereinfacht gesagt wird daher der erhobene Mittelwert (und somit auch der geschätzte Mittelwert) des Bruttomonatsverdienstes für Personen mit einer sehr guten Abschlussnote tendenziell zu niedrig sein, für Personen mit einer sehr schlechten Abschlussnote dagegen eher zu hoch.

Zur Veranschaulichung dieser Differenzen soll ein Beispiel mit dem Beruf-5-Steller „Büro- und Sekretariatskräfte (ohne Spezialisierung) – fachlich ausgerichtete Tätigkeiten" dienen. Für diese Gruppe liegt der Mittelwert der geschätzten Verdienste (3 186 Euro) vergleichsweise nah am Mittelwert der erhobenen Verdienste (3 374 Euro). Die Abweichung beträgt 5,57 %. Der mittlere absolute prozentuale Fehler für diese Gruppe liegt jedoch bei 21,72 %. Da Einzelwerte in der Regel um den Mittelwert streuen, kommt es zu Abweichungen. Folglich weicht die Schätzung des Mittelwerts auch von den individuellen erhobenen Werten ab. Vergleicht man aber die Mittelwerte der Schätzungen mit den Mittelwerten der erhobenen Daten, sind die Unterschiede deutlich

geringer. So liegen die Abweichungen der Mittelwerte (geschätzter und wahrer Bruttomonatsverdienst) für die 500 häufigsten Berufe im Schnitt unter 6,50 %. Daher muss bei der Interpretation der Ergebnisse des Gehaltsvergleichs bedacht werden, dass es sich um Mittelwertschätzungen für das ausgewählte Profil handelt. Die ausgegebenen Werte können und sollen daher nur der Orientierung dienen.

6.3 Plausibilisierung innerhalb der Webanwendung

Maßnahmen zur Qualitätssicherung sind ebenfalls in den interaktiven Gehaltsvergleich integriert. Diese verhindern die Eingabe unplausibler Angaben und die Anzeige von Ergebnissen ohne ausreichende Anzahl an Datensätzen. So darf als Alter nur ein Wert zwischen 16 und 67 Jahren angegeben werden und die Dauer der Unternehmenszugehörigkeit 51 Jahre nicht überschreiten. Darüber hinaus muss die Differenz zwischen Alter und Dauer der Unternehmenszugehörigkeit mindestens 16 Jahre betragen. Des Weiteren werden Ergebnisse nur ausgegeben, wenn in der Kombination Beruf 3-Steller und in der Oberkategorie des gewählten Wirtschaftszweigs mindestens 140 Fälle in den Daten vorliegen. Dies entspricht umgerechnet 30 Fällen für jede Kombination von Wirtschaftszweig 2-Steller (die Gliederungsebene „Abteilungen“ der WZ 2008) und Beruf 3-Steller. Die Einschränkung soll zum einen gewährleisten, dass für die einzelnen Kombinationen dieser sehr stark untergliederten Merkmale genügend Fälle vorliegen. Zum anderen soll sie verhindern, dass versehentlich völlig unsinnige Kombinationen ausgewählt werden können. Sowohl Beruf als auch Wirtschaftszweig haben vergleichsweise hohen Einfluss auf die Höhe des Verdienstes. Doch gerade die Auswahl des korrekten Wirtschaftszweigs ist für Laien nicht ganz einfach. Daher verhindert diese Prüfung auch, dass den Nutzerinnen und Nutzern für falsche Kombinationen von Beruf 3-Steller und Abteilung der WZ 2008 Ergebnisse ausgegeben werden.

7 Fazit und Ausblick

Zielgruppe für die Nutzung des interaktiven Gehaltsvergleichs sind vorwiegend Privatpersonen, die beispielsweise aufgrund von bevorstehenden Gehaltsverhandlungen, Bewerbungsgesprächen oder rein aus Interesse eine spezifische Bruttomonatslohnschätzung erhalten möchten. Den Nutzerinnen und Nutzern soll eine unkomplizierte Schätzung von Bruttomonatsverdiensten für individuell spezifizierte Profile möglich sein. Diese Profile können nach Belieben anhand verschiedener gehaltsbestimmender Merkmale (wie Beruf, Ausbildungsabschluss oder Branche) konfiguriert werden. Verdienstdaten werden somit zielgruppenorientiert aufbereitet und ihr Nutzen dadurch erhöht. Der Gehaltsvergleich trägt folglich zur Ausrichtung des Statistischen Bundesamtes als kundenorientierter und innovativer Informationsdienstleister bei.

Der Anwendung liegt ein Regressionsmodell zugrunde, das eine individuelle Schätzung der Verdienste anhand von Profilen ermöglicht. Dieses Vorgehen verhindert zum einen die Verletzung von Geheimhaltungsvorschriften, die bei einer so spezifischen Darstellung andernfalls unumgänglich wäre. Zum anderen werden Verdienstinformationen auf diese Weise nutzungsfreundlich aufbereitet und leichter zugänglich. Das Verfahren der multiplen Regression erlaubt es, anhand der ermittelten Regressionskoeffizienten neue Profile zu schätzen und den Bruttomonatsverdienst zu prognostizieren. Zur Beurteilung der Qualität des Modells wurde das (korrigierte) R^2 herangezogen. Da es sich um ein Vorhersagemodell handelt, wurde darüber hinaus im Rahmen einer Kreuzvalidierung auch der mittlere absolute prozentuale Fehler (MAPE) betrachtet. Dies stellt sicher, dass keine Überanpassung des Modells auf die zur Schätzung verwendeten Daten vorliegt. Die Qualitätsprüfungen zeigen, dass das Modell geeignet ist, um Schätzungen für neue Daten auszugeben. Dabei ist jedoch zu beachten, dass die geschätzten Verdienste lediglich als Anhaltspunkte zur Orientierung dienen können.

Die Anwendung selbst enthält zudem mehrere Plausibilitätskontrollen, die gravierende Fehleingaben verhindern. Ebenso wird sichergestellt, dass nur dann Verdienstinformationen ausgegeben werden, sofern genügend Daten für das angeforderte Profil vorliegen.

Die Nutzerinnen und Nutzer des interaktiven Gehaltsvergleichs benötigen Verdienstinformationen, um ihre gegenwärtige Situation einzuschätzen. Daher ist eine möglichst aktuelle Datengrundlage erstrebenswert. Die Verwendung der neuen Verdiensterhebung ermöglicht eine solche zeitnahe Aktualisierung der Daten. Im Jahr 2021 wurde einmalig der April erhoben, ab 2022 liefert diese Erhebung monatlich Informationen zu den Verdiensten. Es ist vorgesehen, zur Aktualisierung des Gehaltsrechners jeweils die Aprildaten zu verwenden. Dieser Monat wurde bereits bei der Verdienststrukturerhebung als repräsentativer Berichtsmonat genutzt. Dadurch ist künftig eine Aktualisierung des Rechners bereits im laufenden Jahr möglich. Im Gegensatz zur Verdienststrukturerhebung erfasst die neue Verdiensterhebung darüber hinaus alle Mitarbeiterinnen und Mitarbeiter eines Unternehmens und nicht nur einen Teil davon. Dadurch liegen künftig die Einzeldaten von etwa 7 Millionen Beschäftigten vor. Dieser Zugewinn an Einzeldaten sollte es ermöglichen, für eine größere Anzahl an Beruf-Branchen-Kombinationen Informationen bereitzustellen.

LITERATURVERZEICHNIS

Achatz, Juliane/Gartner, Hermann/Glück, Timea. *Bonus oder Bias? Mechanismen geschlechtsspezifischer Entlohnung*. In: KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie. Jahrgang 57. Ausgabe 3/2005, Seite 466 ff.

Aguines, Herman/Gottfredson, Ryan K./Joo, Harry. *Best practice recommendations for defining, identifying, and handling outliers*. In: Organizational Research Methods. Jahrgang 16. Ausgabe2/2013, Seite 270 ff. DOI: doi.org/10.1177/1094428112470848

Bundesagentur für Arbeit. *Klassifikation der Berufe 2010 – Band 1: Systematischer und alphabetischer Teil mit Erläuterungen*. 2011.

Field, Andy/Miles, Jeremy. *Discovering statistics using SAS*. 2010.

Finke, Claudia. *Verdienstunterschiede zwischen Männern und Frauen 2006*. Herausgeber: Statistisches Bundesamt im Auftrag des Bundesministeriums für Familie, Senioren, Frauen und Jugend. Wiesbaden 2010.

Mincer, Jacob A. *Schooling, experience, and earnings*. New York 1974.

Putka, Dan J./Beatty, Adam S./Reeder, Matthew C. *Modern prediction methods: New perspectives on a common problem*. In: Organizational Research Methods. Jahrgang 21. Ausgabe 3/2018, Seite 689 ff.

SAS Institute Inc. *SAS/STAT® 13.1 User's Guide*. Cary 2013.

Shmueli, Galit. *To explain or to predict?* In: Statistical Science. Jahrgang 25. Ausgabe 3/2010, Seite 289 ff.

Statistisches Bundesamt (Herausgeber). *Verdienststrukturerhebung 2018*. Qualitätsbericht. 2020.

Statistisches Bundesamt (Herausgeber). *Interaktiver Gehaltsvergleich*. Methodenbericht. 2020.

St. John, Caron H. /Roth, Philip L. *The Impact of Cross-Validation Adjustments on Estimates of Effect Size in Business Policy and Strategy Research*. In: Organizational Research Methods. Jahrgang 2. Ausgabe 2/1999, Seite 157 ff. DOI: doi.org/10.1177/109442819922003

Herausgeber
Statistisches Bundesamt (Destatis), Wiesbaden

Schriftleitung
Dr. Daniel Vorgrimler
Redaktion: Ellen Römer

Ihr Kontakt zu uns
www.destatis.de/kontakt

Erscheinungsfolge
zweimonatlich, erschienen im Dezember 2021
Ältere Ausgaben finden Sie unter www.destatis.de sowie in der Statistischen Bibliothek.

Artikelnummer: 1010200-21006-4, ISSN 1619-2907