Digital Business University of Applied Sciences

Data Science und Business Analytics

ADS 41 – ADS-04: Tools der Machine Learning

Prof. Dr. Marcel Hebing

Welche Merkmale aus dem vorhandenen Datensatz sind entscheidend für eine präzise Vorhersage von Einkommenskategorien?

Studienarbeit

Eingereicht von: Mareike Lass-Hennemann

Matrikelnummer: 190190

07.06.2024

Inhaltsverzeichnis

Abbildungsverzeichnis

Executive

Im Projekt Vorhersage von Einkommenskategorien werden zwei Machine Learning Modelle entworfen, welche anhand persönlicher Merkmale zukünftige Einkommenskategorien kleiner/gleich oder größer als 50.000 US-Dollar vorhersagen können. Es wird außerdem untersucht, welche Merkmale in welcher Richtung und Stärke entscheidenden Einfluss auf die Vorhersage haben.

Es zeigt sich, dass die wichtigsten Merkmale zur Vorhersage der Familienstand, Kapitalgewinn, der Bildungsweg und verschiedene Berufsgruppen sind. Die Merkmale Familienstand und Berufsgruppe können in Ausprägungen mit positivem und Ausprägungen mit negativem Einfluss aufgeteilt werden. Merkmale wie Herkunftsregion und „race“ haben für die entworfenen Modelle keine Bedeutung.

Der DecisionTree erreicht einen Accuracy Score von 0,86. Mit einem Modell der logistischen Regression können nahezu gleichwertige Ergebnisse (Accuracy: 0,85) erzielt werden.

Die Daten für dieses Projekt sind leider bereits dreißig Jahre alt, weshalb nicht sicher eine Relevanz für aktuelle Verhältnisse abgeleitet werden kann.

Einleitung

Vor allem in Verbindung mit Berufswünschen, Jobbewerbungen und Gehaltsverhandlungen, aber auch im Zusammenhang mit der persönlichen Lebensplanung ist es für viele Menschen interessant zu erfahren, wie ein zukünftiges Gehalt ausfallen könnte oder wie sich dies bei Modifizierung der persönlichen Merkmale verändern könnte. Laut WSI (2021) sind in Deutschland die fünf entscheidenden Faktoren, die die Höhe des Gehalts bestimmen, die folgenden: Anforderungsniveau, Geschlecht, Betriebsgröße, Tarifbindung und Bundesland. Das Handelsblatt (2024) nennt als wichtigste Faktoren die Branche und ebenfalls das Geschlecht und das Bundesland. Weitere Faktoren können lt. WSI (2021) sein: das Berufsfeld, die Berufserfahrung und ob jemand Leitungsverantwortung übernimmt. Für die USA konnte keine einheitliche Nennung solcher Faktoren gefunden werden – verschiedene Internetquellen nennen aber Faktoren wie geographische Region, Bildungsweg, Geschlecht, Alter, Abstammung, Berufsgruppe.

Der vorliegende Datensatz aus den USA enthält nur wenige der vorher genannten deutschen Merkmale, aber einige der amerikanischen: Alter, Beschäftigungsstatus, Bildungsweg, Familienstand, Berufsgruppe, Verwandtschafts-Beziehung, Abstammung, Geschlecht, Kapital-Gewinne, Kapital-Verluste, Wochenarbeitsstunden, Herkunftsland. Kann mit diesen Merkmalen ein Modell entworfen werden, welches präzise Vorhersagen treffen kann? Welche Merkmale sind für die Vorhersage relevant? Sind in den USA andere Merkmale relevant als in Deutschland? Zur Annäherung an diese Fragen werden zunächst Einzelbetrachtungen des Einkommens im Zusammenhang mit den verschiedenen Merkmalen vorgenommen, statistische und später Machine Learning Methoden angewendet. Die entworfenen Modelle können Personen anhand ausgewählter Merkmale mittels Klassifikation den zwei Gehaltskategorien <= oder >50.000 US-Dollar zuordnen, sodass diese eine Einschätzung über ihre Chance auf ein höheres/niedrigeres Gehalt bekommen und zusätzlich erfahren, welche Merkmale diese Chance beeinflussen oder gar verändern können.

Die Trainings- und Testdaten stammen aus dem Jahr 1994. 1995 lag das durchschnittliche Jahreseinkommen in den USA bei 27.845 US-Dollar (Bureau of Labor Statistics, 1996), in Deutschland zum Vergleich bei 23.984 Euro (Institut Arbeit und Qualifikation der Universität Duisburg-Essen, 2024).

Daten und Methoden

Die Daten zu diesem Projekt wurden freundlicherweise von Prof. Dr. Hebing zur Verfügung gestellt. Sie stammen aus der Datenbank des amerikanischen Census Bureau von 1994 und wurden ursprünglich von Ronny Kohavi und Barry Becker veröffentlicht. Es stehen bereits getrennte Datensätze für Training und Tests zur Verfügung. Diese enthalten 32 537 (train) bzw. 16 276 (test) Fälle nach Entfernung von Duplikaten (24 bzw. 5) mit jeweils 15 Merkmalen. Bereits beim Einlesen der Daten wurde die Großschreibung der Ausprägungen in Kleinschreibung überführt, zudem wurden alle Leerzeichen entfernt. Als abhängige Variable wird das binäre, kategoriale Merkmal „income“ identifiziert. Es liegen die Werte als string in den zwei Ausprägungen „<= 50K“ und „>50K“ vor. Diese werden zur besseren Interpretierbarkeit umgewandelt in 0 und 1. Es liegen 14 unabhängige Variablen vor, davon numerisch: „age”, „final_weight”, „education_num”, „capital_gain”, „capital_loss”, „hours_per_week” und kategorial: „workclass“, „education“, „marital_status“, „occupation“, „relationship“, „race“, „sex“, „native_country“.

Sowohl um der Projektfrage gerecht zu werden, aber auch um Overfitting zu vermeiden, die Trainingszeit zu reduzieren und das Modell leichter interpretieren zu können, wurden die Features gleich zu Beginn sorgfältig ausgewählt und reduziert (Nguyen & Zeigermann, 2021).

Die Variable „final_weight“ wird von vornherein aus der Modellierung ausgeschlossen, da diese der Gewichtung der Fälle bei Anwendung von verallgemeinernden statistischen Methoden dient und in diesem Projekt somit keine Aussagekraft hat. Die Variable „education“ wird ebenfalls zu Beginn ausgeschlossen, da die Variable „education_num“ dieselben Ausprägungen numerisch darstellt. Ausgeschlossen werden außerdem „workclass“ (dt. Beschäftigungsstatus, hier wird wenig Aussagekraft bei 70% Ausprägung „private“ und zusätzlich vielen fehlenden Werten vermutet, der Autorin ist nicht ganz klar, was „private“ als Beschäftigungsstatus bedeuten soll?) und „relationship“ (überschneidet sich mit „marital_status“; zudem unvollständig und nicht eindeutig, da immer nur ein Beziehungsattribut vorhanden ist, Personen ja aber in Bezug auf verschiedene Personen unterschiedliche Status haben können). Nach Auswertung der logistischen Regression in statsmodels werden außerdem noch „race“ und „native_country“ als nicht signifikante Merkmale identifiziert – beide Merkmale liegen zudem recht konzentriert mit 85% bzw.

90% der Fälle hauptsächlich in einer Ausprägung vor („white“, „united-states“). Es wurde überlegt auch Kapitalgewinn und -verlust auszuschließen, da hier ebenfalls 91% bzw. 95% der Fälle keine Gewinne/Verluste haben, also die Verteilung sehr ungleich ist, aber die Variablen sind sowohl in statsmodels als auch in den Machine Learning Modellen signifikant und bringen eine Verbesserung der Modellleistung.

Zur Modellierung werden nun die folgenden Variablen verwendet:

- „age“: Alter der befragten Personen; Ausprägungen zwischen 17 bis 90Jahre;
- „education_num“: Schul-/Ausbildungsstatus in numerischer Entsprechung von „education“; Ausprägungen z.B. Highschool-Graduate, College-Graduate, Bachelors, Masters, usw.
- „capital_gain“: Kapitalgewinn einer Person; Ausprägungen zwischen 0 bis 100K US-Dollar
- „capital_loss“: Kapitalverlust einer Person; Ausprägungen zwischen 0 bis 4356 US-Dollar
- „hours_per_week“: Stunden die eine Person pro Woche arbeitet; Ausprägungen zwischen 1 bis 99 Stunden
- „marital_status“: Familienstand; Ausprägungen z.B. Married-civ-spouse, Never-married, Divorced usw.
- „sex“: Geschlecht; Ausprägungen female, male
- „occupation“: Berufsgruppe, in der eine Person arbeitet; fehlende Werte werden in „other-services“ umgewandelt; die Gruppen ' Armed-Forces', ' Protective-serv', ' Priv-house-serv' werden in der Gruppe ' protect-priv-serv' zusammengefasst; weitere Ausprägungen z.B. Prof-specialty, Craft-repair, Exec-managerial, Adm-clerical, Sales, Machine-op-inspct

Auf die numerischen Variablen wird im Rahmen einer Pipeline zur Anreicherung der Daten PolynomialFeatures angewendet sowie eine Standardskalierung vorgenommen. Die kategorialen Variablen werden mit dem OneHotEncoder in numerische umgewandelt. Eine polynomiale Anreicherung auch der kategorialen Daten bringt keine Verbesserung der Leistung der Modelle.

Um einen Überblick über die Zusammenhänge zu bekommen, wird sowohl ein Report mit SweetVIZ erstellt und ausgewertet, als auch die Variable income in Zusammenhang gestellt

mit jeder einzelnen unabhängigen Variable. Nachfolgend wird mithilfe der LogitRegression in statsmodels die Richtung und Stärke des Einflusses der unabhängigen Variablen in verschiedenen Kombinationen auf die Variable „income“ untersucht. Hieraus ergibt sich der Ausschluss von zwei weiteren Merkmalen (race, native_country s.o.) aus der Modellierung der Machine Learning Modelle. Abschließend werden entsprechend der Erfordernisse des Datensatzes zwei Machine Learning Modelle trainiert: ein LogisticRegressor und ein DecisionTreeClassifier – beide gehören zum supervised learning im Bereich der Klassifikation – die Anwendung von Klassifikationsmodellen ergibt sich aus dem vorhandenen gelabelten Datensatz mit einer binär-kategorialen Zielvariablen (Geron, 2023). Auf Basis des LogisticRegressors werden die Koeffizienten und die Richtung und Stärke ihres Einflusses auf die Zielvariable untersucht. Auf Basis des DecisionTree kann zusätzlich die Wichtigkeit (Importance) der einzelnen Features für das Modell berechnet und ausgegeben werden – dies bietet den Vorteil, dass auch nicht-lineare Beziehungen zwischen Features und Zielvariable berücksichtigt werden (Nguyen & Zeigermann, 2021).

Ergebnisse

Aus der EDA über die Trainingsdaten ergibt sich, dass etwa 76% der Fälle in der Kategorie <=50K sowie 24% der Fälle in der Kategorie >50K liegen. Chancen (Wahrscheinlichkeit größer/gleich 0.4) auf ein hohes Einkommen haben lt. der Einzelbetrachtungen der verschiedenen Merkmale z.B. Personen, die eine der folgenden Merkmalsausprägungen erfüllen: Wochenarbeitszeit 50 bis 59 Stunden, Abschluss Bachelor oder Master oder Doctorates oder Prof-school, Berufsgruppe Führungskräfte oder Fachkräfte, Familienstand verheiratet, Kapitalgewinn über 10.000 US-Dollar, Kapitalverlust zwischen 1750 bis 3000 US-Dollar. Zudem können sich ein höheres Alter (ab 40Jahre), ein längerer Bildungsweg (mindestens 8Jahre), männliches Geschlecht und bestimmte Herkunftsländer (Indien, Iran, Taiwan, Frankreich) positiv auswirken (im Sinne einer Wahrscheinlichkeit größer 0.3 für hohe Einkommen).

Mit einer LogitRegression in statsmodels lässt sich nur ein Pseudo R-squared von maximal 0.411 erzielen, was auf eine geringe Erklärungsfähigkeit des Modells hinweist. Der Familienstand und einige Berufe haben die größten Koeffizienten und weisen somit den größten Einfluss auf das Einkommen auf, während Kapitalgewinn und -verlust die kleinsten Koeffizienten und somit nur sehr geringen Einfluss auf das Einkommen haben.

Es werden die Machine Learning Algorithmen Logistic Regressor und Decision Tree mit den Daten trainiert. Der Logistic Regressor erreicht einen Accuracy Score von 0.85 mit wenig Overfitting auf die Testdaten. Der Classification Report weist für niedriges Einkommen einen Precision Score von 88% aus – es werden also 88% der Fälle richtig als niedriges Einkommen vorhergesagt. Gleichzeitig findet das Modell 93% der niedrigen Einkommen (Recall-Score). Für hohe Einkommen sagt das Modell 72% der Fälle richtig voraus, findet aber nur 59% aller hohen Einkommen.

Die größten positiven Koeffizienten haben die Merkmale und Ausprägungen von Kapitalgewinn, Familienstand verheiratet und Kapitalgewinn*Wochenarbeitszeit. Familienstand unverheiratet und die Berufsgruppe Landwirtschaft/Fischerei haben die „größten“ Koeffizienten im negativen Bereich. Eine detaillierte Auflistung ist in Abbildung 1 zu sehen:

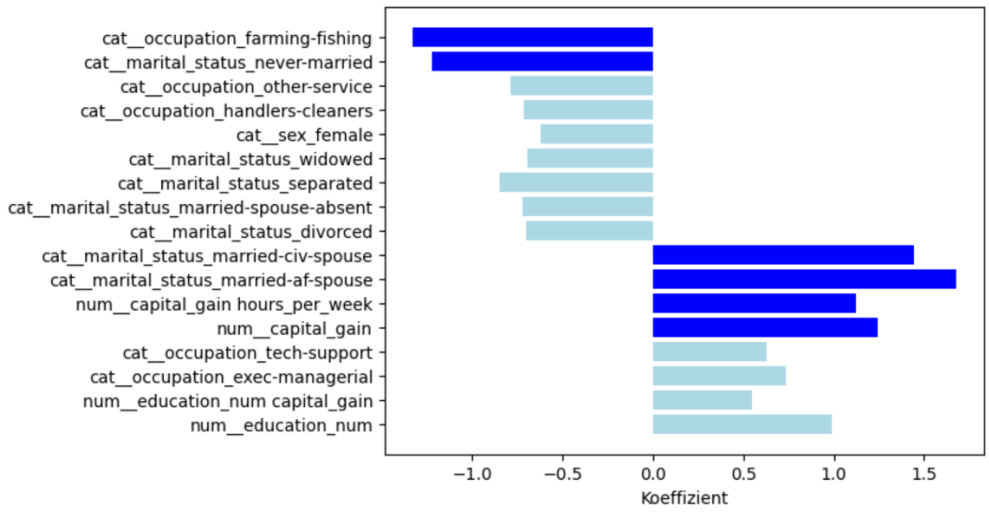

Abbildung 1: Koeffizienten des Models LogisticRegressor

Durch die Betrachtung der Koeffizienten innerhalb der Merkmale Familienstand und Berufsgruppe wird jeweils eine Aufsplittung deutlich in Ausprägungen mit positivem Koeffizienten und Ausprägungen mit negativem Koeffizienten.

Der Decision Tree erreicht nach Hyperparameter-Tuning einen Accuracy Score auf den Testdatensatz von 0.857 mit wenig Overfitting auf die Trainingsdaten. Der Classification Report ergibt für niedriges Einkommen eine Rate von 88% richtiger Vorhersagen (Precision). Das Modell findet insgesamt 95% der niedrigen Einkommen (Recall). Für hohe

Einkommen liegt das Modell in 77% der Fälle richtig, findet aber insgesamt nur 57% der hohen Einkommen.

Die wichtigsten Merkmale und Ausprägungen für das Modell sind Familienstand verheiratet, Kapitalgewinn, Bildungsweg, Alter* Bildungsweg, Kapitalverlust, Bildungsweg*Wochenarbeitszeit.

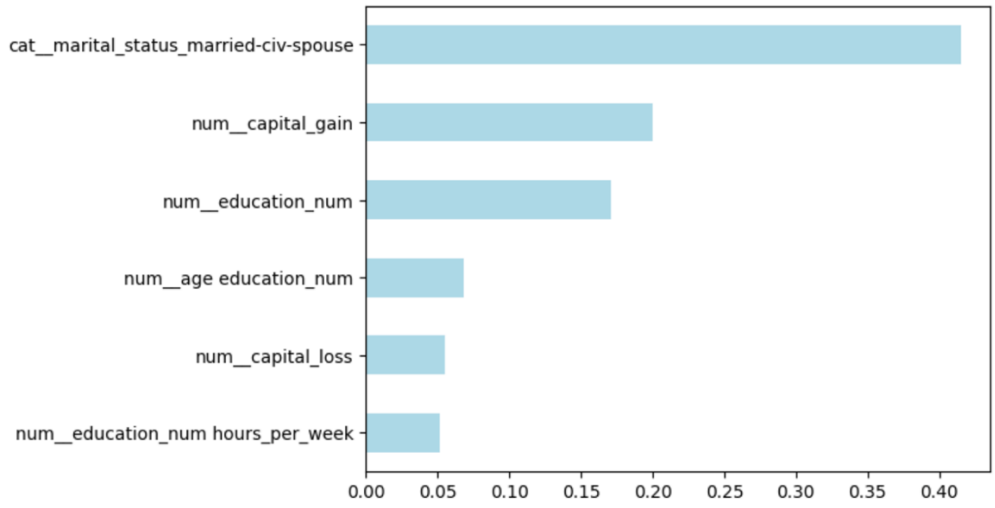

Abbildung 2: Feature Importances des Modells DecisionTreeClassifier

Diskussion und Handlungsempfehlung

Anhand der gegebenen Merkmale kann eine recht präzise Vorhersage zumindest für die niedrigen Einkommen getroffen werden - niedrige Einkommen werden mit einer Präzision von 88% vorhergesagt, hohe jedoch nur mit einer Präzision von 72 bis 77%. Die F1-Scores sind für beide Modelle sehr ähnlich. Möchte man noch präzisere Vorhersagen insbesondere für die hohen Einkommen erreichen, könnte in einem weiteren Schritt ein RandomForestClassifier probiert werden – durch das Training verschiedenster Bäume und die Mittelung ihrer Ergebnisse können ggf. genauere Gesamt-Ergebnisse bei geringerem Overfitting erreicht werden (Harrison, 2021).

Als entscheidend werden die Merkmale Familienstand, Kapitalgewinn, Bildungsweg und verschiedene Berufsgruppen identifiziert. Mithilfe des LogisticRegressor können zudem Berufsgruppen und Familienstände mit positivem sowie negativem Einfluss auf die Einkommenshöhe unterschieden werden. Die Ergebnisse über positive und negative

Einflüsse verschiedener Berufsgruppen und Familienstände sind wohl keine Überraschung, sondern bestätigen allgemeine Annahmen über Einkommenshöhen in der Gesellschaft.

Dass aber die Merkmale Familienstand und Kapitalgewinn sich grundsätzlich als die wichtigsten für die Vorhersage herausstellen, ist unerwartet, da diese laut aktuellen Aussagen weder für den deutschen noch den amerikanischen Arbeitsmarkt tatsächlich relevant für die Einschätzung des Gehalts sind. Für den amerikanischen Arbeitsmarkt scheint heute zumindest weiterhin der Bildungsweg relevant zu sein. Der Bildungsweg könnte bzgl. des deutschen Arbeitsmarktes in eine Richtung sortiert werden, wie die Merkmale Anforderungsniveau und Leitungsverantwortung aus den aktuellen Empfehlungen – da beide einen längeren Bildungsweg voraussetzen.

Wenn auch unerwartet, könnte dennoch für das Merkmal Familienstand mit Hinblick auf Steuervorteile und das Merkmal Bildungsweg mit Hinblick auf komplexere und somit besser bezahlte Jobs eine kausale Begründung für höhere Einkommen hergeleitet werden. Diese Einschätzung wird bzgl. des Bildungsweges zusätzlich durch die Richtung und Stärke der Koeffizienten des LogisticRegressor bekräftigt: für die vermutlich länger ausgebildeten Berufsgruppen Führungskräfte und technischer Support liegen positive Koeffizienten vor, während die Berufsgruppen mit vermutlich einfacherem Bildungsweg Reinigungskräfte, Servicekräfte und Landwirtschaft/Fischerei negative Koeffizienten bilden.

Ein kausaler Zusammenhang in der Richtung „… begründet hohes Einkommen“ kann jedoch nicht eindeutig für Kapitalgewinn und -verlust hergeleitet werden, zumal Kapital in beide Richtungen – Zuwachs und Verlust – deutlich die Chance auf ein hohes Einkommen erhöht. Ein Erklärungsansatz könnte darin liegen, die Kausalität umzudrehen: nicht der Kapitaleinsatz begründet das hohe Einkommen, sondern vermutlich sind Menschen mit einem hohen Einkommen eher bereit, Kapital grundsätzlich einzusetzen.

Abschließend bleibt zu sagen, dass das Modell heute nicht ohne Überarbeitung eingesetzt werden sollte, da die zugrundeliegenden Daten dreißig Jahre alt sind. Beispielsweise lag das aktuelle Durchschnittsgehalt in den USA im Jahr 2022 mehr als doppelt so hoch bei 61.185 US-Dollar (in Deutschland zum Vergleich bei 54.997 Euro) (IMK, 2024), sodass mindestens die Einkommenskategorien angepasst werden sollten. Die Recherche deutet zudem darauf hin, dass möglicherweise heute andere Merkmale zur Einkommensvorhersage relevant sein könnten, als damals und ggf. auch in den USA andere

Merkmale relevant sind, als in Deutschland. Hier wäre eine Erhebung alter und neuer Merkmale in beiden Ländern und ein Vergleich sowohl der Wichtigkeit alter gegen neue Merkmale in Machine Learning Modellen als auch Deutschland gegen die USA interessant.

Literaturverzeichnis / Quellen

Bureau of Labor Statistics (1996). „AVERAGE ANNUAL PAY LEVELS IN METROPOLITAN AREAS, 1995". Abgerufen am 20.05.2024 von https://www.bls.gov/news.release/history/anpay2_110896.txt#:~:text=The%20average%20annual%20pay%20level,on%20September%2025%2C%201996

Geron, A. (2023). Praxiseinstieg Machine Learning mit Scikit-Learn, Keras und TensorFlow (3.Auflage). Heidelberg: O'Reilly.

Handelsblatt (2024). „So hoch ist das Durchschnittsgehalt hierzulande". Abgerufen am 20.05.2024 von https://www.handelsblatt.com/unternehmen/einkommen-in-deutschland-so-hoch-ist-das-durchschnittsgehalt-hierzulande/26628226.html

Harrison, M. (2021). „Machine Learning Die Referenz" (1.Auflage). Heidelberg: O'Reilly.

IMK (2024). „Vergleich Deutschland-USA: Situation hierzulande bei 10 von 15 wichtigen ökonomischen und sozialen Kenngrößen besser". Abgerufen am 20.05.2024 von https://www.boeckler.de/pdf/pm_imk_2024_01_31.pdf

Institut Arbeit und Qualifikation der Universität Duisburg-Essen (2024). „Entwicklung der durchschnittlichen Löhne/Gehälter 1995-2023". Abgerufen am 20.05.2024 von https://www.sozialpolitik-aktuell.de/files/sozialpolitik-aktuell/_Politikfelder/Einkommen-Armut/Datensammlung/PDF-Dateien/tabIII1.pdf

Nguyen, C. N. & Zeigermann, O. (2021). Machine Learning kurz & gut (2.Auflage). Heidelberg: O'Reilly.

WSI (2021). „Diese fünf Faktoren bestimmen ihr Gehalt“. Abgerufen am 20.05.2024 von https://www.boeckler.de/pdf/pm_ta_2021_03_30.pdf