Erklärbare Künstliche Intelligenz in der Wirtschaft: Eine Fallstudie von Einkommensvorhersagen

Nhut Hoa Huynh

Department Informatik, Fakultät Technik und Informatik, Hochschule für Angewandte Wissenschaften Hamburg, Germany
nhuthoa.huynh@haw-hamburg.de

Zusammenfassung. Dieses Paper zielt darauf ab, ein umfassendes Verständnis für die verfügbaren Methoden der erklärbaren künstlichen Intelligenz (XAI) zu erlangen. Um dies zu erreichen, untersucht das Paper den Anwendungsfall der Einkommensprognose in der Wirtschaft, um die spezifischen Funktionalitäten zu erforschen, die verschiedene XAI-Methoden bieten können. Durch die Fokussierung auf die Anwendung von XAI in der Einkommensprognose trägt diese Untersuchung dazu bei, die Transparenz von Modellen der Künstlichen Intelligenz zu erhöhen. Die Ergebnisse dieser Untersuchung liefern Einblicke in das Potenzial von XAI-Techniken zur Förderung verantwortungsvoller KI-Anwendungen in realen Wirtschaftsszenarien.

Schlüsselwörter: Erklärbare Künstliche Intelligenz · XAI · Einkommensvorhersage · Wirtschaft · Transparenz · Interpretierbarkeit · AI-Methoden · Künstliche Intelligenz · Maschinelles Lernen · Ethische Aspekte.

1 Einleitung

1.1 Ziele dieser Arbeit

Das Hauptziel dieser Arbeit ist es, ein besseres Verständnis der verfügbaren Methoden der Erklärbaren Künstlichen Intelligenz (XAI) zu erlangen. Hierzu wird der Anwendungsfall der Einkommensvorhersage in der Wirtschaft untersucht, um herauszufinden, welche spezifischen Funktionen die verschiedenen XAI-Methoden bieten können.

1.2 Gliederung

Die Arbeit ist in die folgenden Kapitel unterteilt: In Kapitel 2 wird das Grundwissen über erklärbare künstliche Intelligenz vorgestellt. Kapitel 3 beschreibt das durchgeführte Experiment. Kapitel 4 enthält das Fazit.

2 Grundlagen der erklärbaren künstlichen Intelligenz

Erklärbare künstliche Intelligenz ist das grundlegende Konzept, das künstliche Intelligenzsysteme transparent und verständlich macht. In der heutigen Zeit, in der sich komplexe Algorithmen und maschinelles Lernen in verschiedenen Bereichen immer mehr durchsetzen, wird die Forderung nach Erklärbarkeit und Nachvollziehbarkeit der von KI-Systemen getroffenen Entscheidungen immer wichtiger. XAI zielt darauf ab, die "Blackbox"- Natur herkömmlicher KI-Modelle zu durchbrechen, indem sie menschenähnliche Erklärungen für ihre Entscheidungen liefert. Dies ermöglicht nicht nur ein tieferes Verständnis der zugrundeliegenden Prozesse, sondern erhöht auch das Vertrauen der Nutzer und erleichtert die Identifizierung potenzieller Fehler oder Vorurteile in den Modellen.
Zu den Grundlagen von XAI gehören Techniken wie die Modellinterpretation, die Verwendung lokaler und globaler Erklärungsmethoden und die Berücksichtigung ethischer Fragen, um eine verantwortungsvolle Anwendung von KI-Technologien zu gewährleisten.
Interpretationsmethoden in der XAI lassen sich aufgrund ihrer Eigenschaften und Anwendungen in verschiedene Typen einteilen.
Erstens gibt es modellagnostische und modellspezifische Methoden. Modellagnostische Verfahren können auf verschiedene Arten von Modellen angewandt werden, was sie für die Interpretation verschiedener KI-Modelle vielseitig einsetzbar macht. Umgekehrt sind modellspezifische Methoden speziell für die Interpretation bestimmter Modelltypen konzipiert und passen ihren Ansatz an die einzigartigen Merkmale dieser Modelle an.
Zweitens lassen sich die Interpretationsmethoden in globale und lokale Erklärungen unterteilen. Lokale Erklärungen zielen darauf ab zu verstehen, wie eine bestimmte Vorhersage durch die Analyse einer bestimmten Instanz oder eines bestimmten Datenpunkts zustande kommt. Im Gegensatz dazu werden bei globalen Erklärungen die allgemeinen Faktoren untersucht, die die Vorhersagen für das gesamte Modell beeinflussen.
Bei den Interpretationsmethoden schließlich kann zwischen modellbasierten und Post-hoc-Ansätzen unterschieden werden. Modellbasierte Methoden sind von Natur aus interpretierbar, wie z. B. lineare Regressionsmodelle, bei denen das Verständnis des Entscheidungsprozesses des Modells direkt und klar ist. Post-hoc-Methoden hingegen versuchen, die Entscheidungsfindung komplexer Modelle zu erklären, und die meisten Interpretationstechniken fallen in diese Kategorie.
Die Einteilung der Interpretationsmethoden hilft den Anwendern, die am besten geeigneten Techniken für ihre spezifischen KI-Modelle und Anwendungsfälle auszuwählen.
Es ist auch von entscheidender Bedeutung, **die Konflikte in der Verwendung von Begriffen in der XAI** zu erkennen, insbesondere zwischen Interpretierbarer KI und Erklärbare KI. Obwohl beide Disziplinen das gemeinsame Ziel verfolgen, die Transparenz von KI-Systemen zu fördern, unterscheiden sie sich oft in ihren Definitionen und Methoden. Der Begriff "Interpretieren" geht über einfache "Erklärungen" hinaus. Es beinhaltet die Fähigkeit, über den Kontext einer Vorhersage hinauszublicken und tiefergehende Verbindungen zu suchen.

Wenn Menschen etwas interpretieren, konstruieren wir Bedeutung und erkennen Zusammenhänge, die über das beobachtete Ereignis und den Kontext hinausgehen.
Der letzte Aspekt, wenn es um XAI geht, ist die **Ethik**. Der ethische Aspekt spielt eine wesentliche Rolle bei der Integration von Transparenz und Verantwortung in Systeme der künstlichen Intelligenz. Die Ethik in der künstlichen Intelligenz befasst sich mit Themen wie Fairness, Vermeidung von Vorurteilen, Datenschutz und Vermeidung von Diskriminierung. Wenn XAI-Systeme nicht ordnungsgemäß entwickelt und überwacht werden, können sie unbeabsichtigte oder sogar schädliche Folgen haben, die sich auf Einzelpersonen oder ganze Gemeinschaften auswirken können. Daher ist es von entscheidender Bedeutung, ethische Richtlinien und Standards sowie Gesetze zu entwickeln, um sicherzustellen, dass XAI-Technologien auf verantwortungsvolle und ethische Weise eingesetzt werden, um das Potenzial von XAI zum Nutzen der Gesellschaft auszuschöpfen.
Durch die Schaffung solider Grundlagen für XAI können wir die potenziellen Vorteile der KI nutzen, ohne die Kontrolle über ihre Entscheidungsprozesse zu verlieren.

3 Experimente

3.1 Datensatz

Für die Durchführung der Experimente wird ein Datensatz gewählt, der häufig in früheren Forschungsarbeiten verwendet wurde, die sich mit Fragen der algorithmischen Fairness und Zuverlässigkeit befassen: der Census-Einkommensdatensatz. Dies wird mit potenziellen Vorurteilen bei Entscheidungen wie dem Zugang zu Krediten und Arbeitsmöglichkeiten in Verbindung gebracht.
Die Census-Einkommensdaten (CI) enthalten Informationen, die zur Vorhersage des Einkommens von Einzelpersonen verwendet werden [1]. Sie umfassen 32.561 Datensätze und 12 Merkmale. Diese Merkmale heißen auf Englisch “age, work class, education, marital status, occupation, relationship, race, sex, capital gain, capital loss, hours per week, country”. Die Klassen sind niedriges Einkommen (weniger oder gleich 50.000 Dollar) oder hohes Einkommen (mehr als 50.000 Dollar).

3.2 Facets Dive

Die Daten werden zunächst mit Facets Dive angezeigt. Die Verwendung von Facets Dive soll dabei helfen, festzustellen, welche der bereitgestellten Informationen für die Einkommensprognose relevant sind.
Das Buch [2] enthält 2 Beispiele:

– Abbildung 1 zeigt, die Einkommenskurve steigt von der Kindheit bis zum Erwachsenenalter an, erreicht einen Spitzenwert und sinkt dann langsam mit dem Alter.

– Abbildung 2 bestätigt nicht nur, dass eine längere Ausbildung zu einem höheren Einkommen führt, sondern zeigt auch deutlich, dass eine längere Ausbildung in Verbindung mit dem Alter zu einem höheren Einkommen führt, was ab dem Alter von 30 Jahren im Bereich der 13- bis 17-jährigen Ausbildung zu beobachten ist.

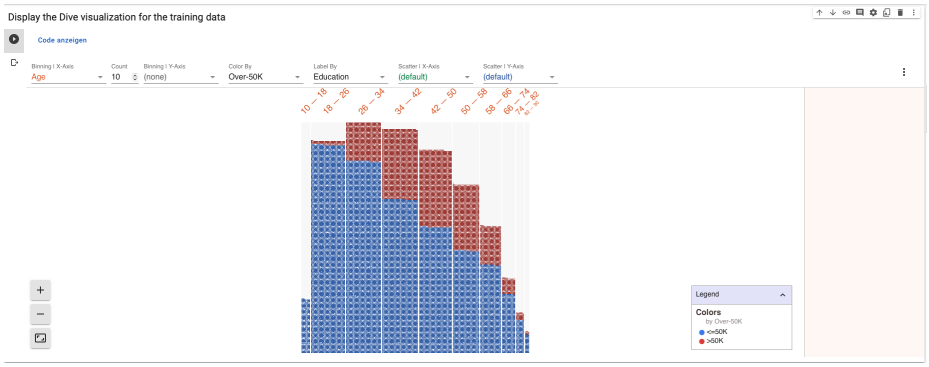

Abb. 1. Einkommen und Alter

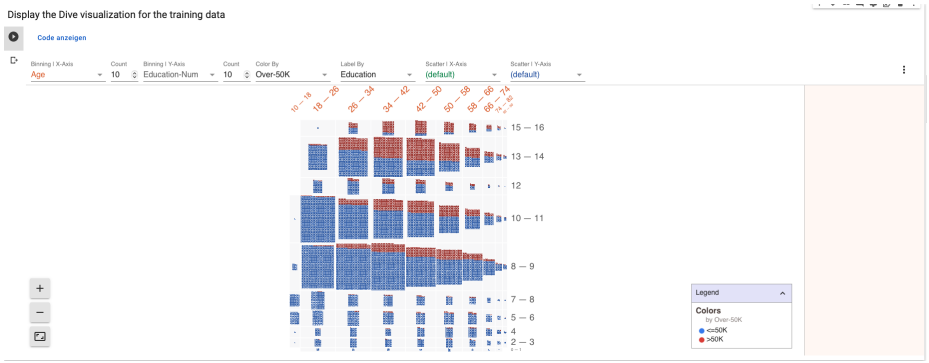

Abb. 2. Einkommen und Ausbildung

Weitere Beispiele sind folgende:

– Abbildung 3 illustriert ein deutliches Einkommensgefälle zwischen Frauen und Männern sowie zwischen verschiedenen Bevölkerungsgruppen.
– Auch beim Familienstand gibt es Einkommensunterschiede, nämlich dass die Verheirateten oft ein höheres Einkommen haben. Dies kann in Abhängigkeit mit dem Alter stehen (Abbildung 4).

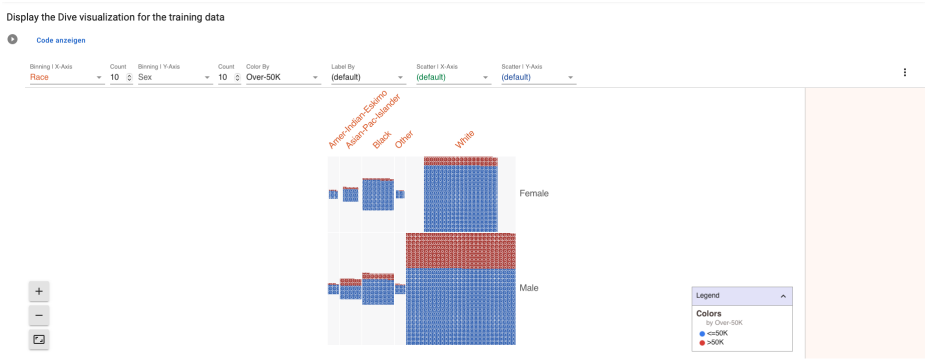

Abb. 3. Einkommen, Geschlecht und Bevölkerungsgruppe

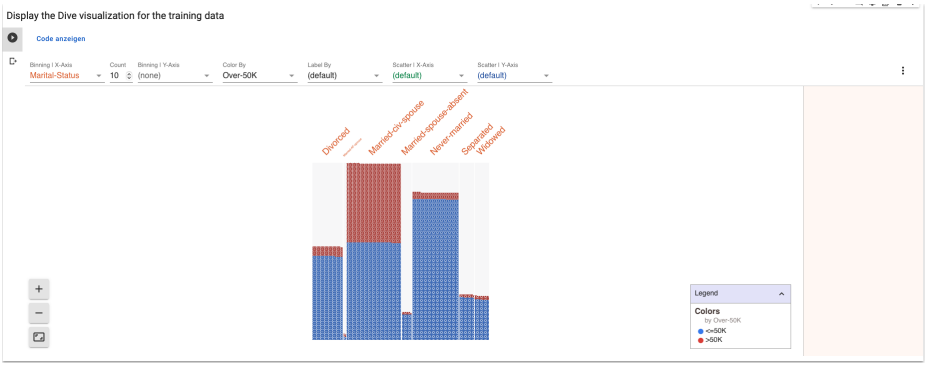

Abb. 4. Einkommen und Familienstand

3.3 What-if Tools

Vor der Durchführung einer benutzerorientierten "What-if"-Untersuchung, hat der Autor des Buches [2] den Datensatz aus ethischen Gründen transformiert, z. B. Race zu Reading, Sex zu Travel Preferences usw. In diesem Paper wird der Datensatz dagegen nicht transformiert. Die beiden Modelle, deren Leistung von WIT analysiert werden soll, nämlich ein linearer Klassifikator und ein tiefes neuronales Netz (DNN), bleiben jedoch gleich.

"What-if" bietet die folgenden drei Werkzeuge:

- Data Point Editor: ein Interface zur Bearbeitung von Datenpunkten und zur Erklärung der Vorhersagen.
- Features: ein Interface zur Visualisierung von Feature-Statistiken
- Performance and Fairness: Werkzeuge zur Messung der Genauigkeit und Fairness von Vorhersagen.

Wir werden den Datensatz und die Leistung der beiden Modelle mit diesen Werkzeugen untersuchen.

Im Data Point Editor können die Merkmale jedes Datenpunkts angezeigt und auch geändert werden, um die ursprüngliche Vorhersage sowie die sich ändernde Vorhersage bei Änderung der Merkmale zu untersuchen. Diese Änderungen veranschaulichen und erklären, wie das Modell "denkt". Außerdem kann man für jeden Datenpunkt die sogenannte "nearest counterfactual" finden und anzeigen. Abbildung 5 zeigt ein Beispiel.

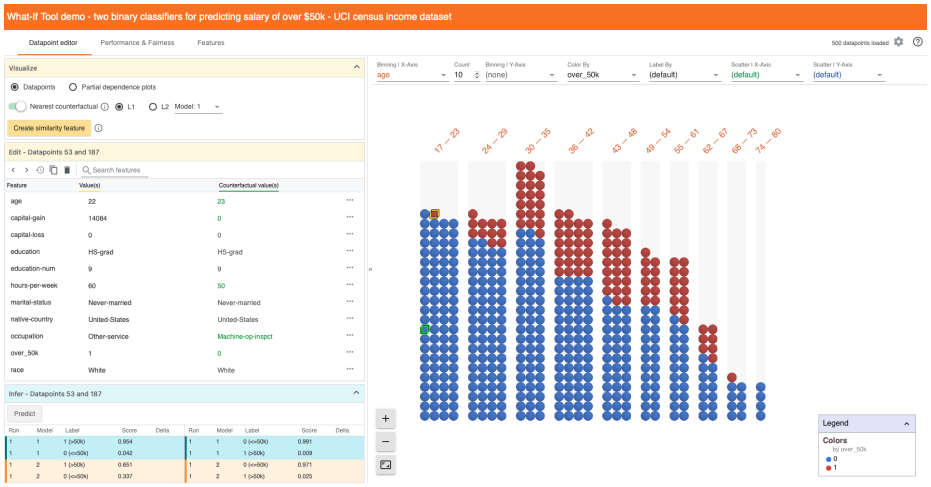

Abb. 5. Ein Datenpunkt und sein sog. nearest counterfactual

Eine weitere Analysemöglichkeit, die sich nicht nur auf die Datenvisualisierung, sondern auch auf die Analyse der Modellleistung bezieht: Man kann außerdem im Data Point Editor sehen, wie sich jeder Datenpunkt im Ergebnis zwischen dem linearen Modell (1) und dem DNN-Modell (2) unterscheidet. Wir müssen nur die Datenpunkte organisieren, indem wir die Streuung der X-Achse auf "inference score 1" und die Streuung der Y-Achse auf "inference score 2" setzen. Punkte außerhalb der Diagonale haben unterschiedliche Ergebnisse zwischen den beiden Modellen (Abbildung 6).

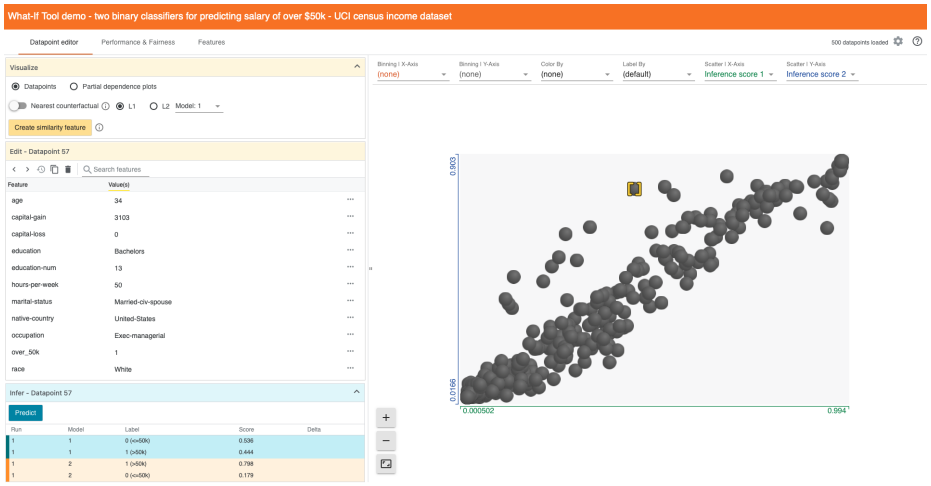

Abb. 6. Ein Datenpunkt, der unterschiedliche Ergebnisse zwischen den beiden Modellen hat

Das zweite Werkzeug von WIT ist “Features”. Im Tab “Features” (Abbildung 7) können die Statistiken der Features gelesen werden. So lassen sich die Merkmale analysieren, indem man sie sortiert, ihre Minimal-, Maximal-, Median- und Mittelwerte sowie die Verteilungsabstände visualisiert oder nach Ungleichmäßigkeiten/fehlenden Datensätzen in ihren Datenverteilungen sucht.

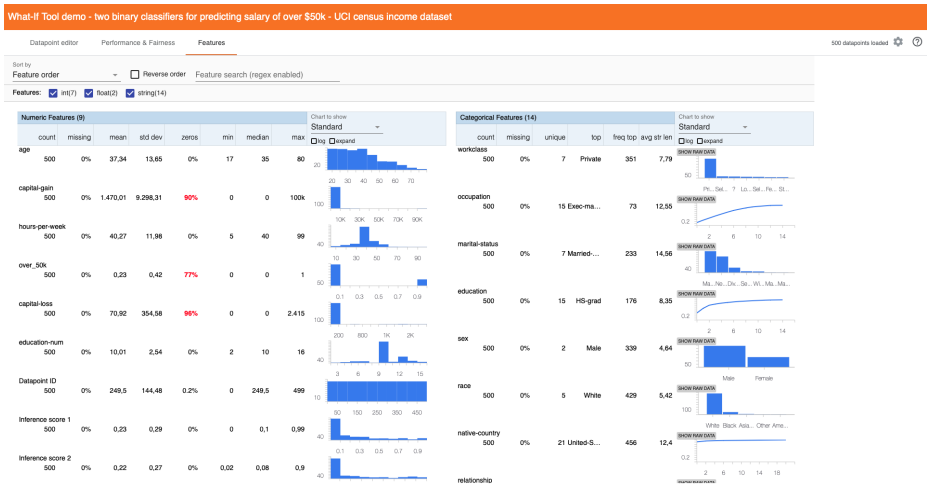

Abb. 7. Features Tab

Zu guter Letzt kann man in “Performance and fairness” (Abbildung 8) die Leistung und Fairness der Modelle anhand von Ground Truth, Cost Ratio, Fairness, ROC-Kurven, Slicing, PR-Kurven und der Konfusionsmatrix der Modelle überprüfen. Die Leistung der Modelle kann zusätzlich allein mit den Änderungen des Thresholds oder zusammen mit den Änderungen des Cost Ratio untersucht werden.

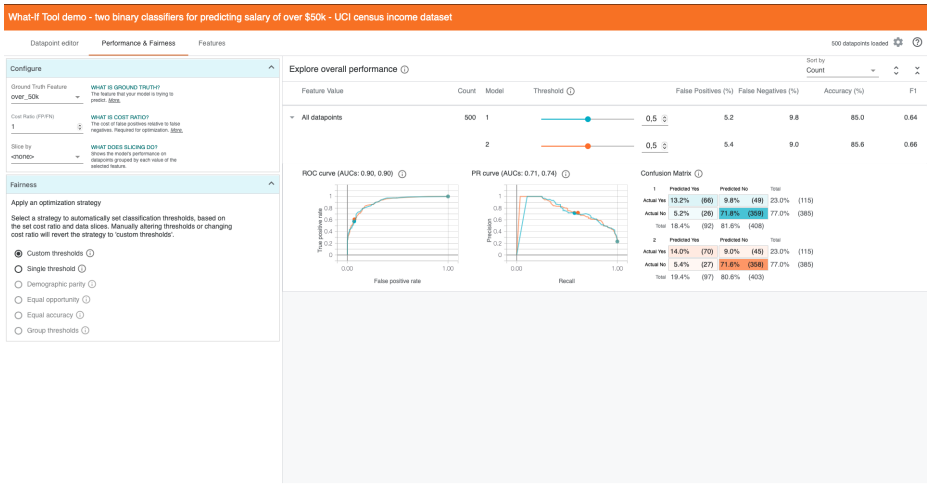

Abb. 8. Performance and fairness Tab

Mit “Performance and fairness” können wir wie in Abbildung 9 den Datensatz auch nach Merkmalen aufteilen, um zu sehen, ob zwei Modelle die gleiche Leistung zwischen den Slices haben. Hier können wir auch den Threshold anpassen, um zu prüfen, ob zum Beispiel ein Modell größere Unterschiede in den Threshold-Werten pro Slice benötigt, um die gewünschte Einschränkung zu erreichen.

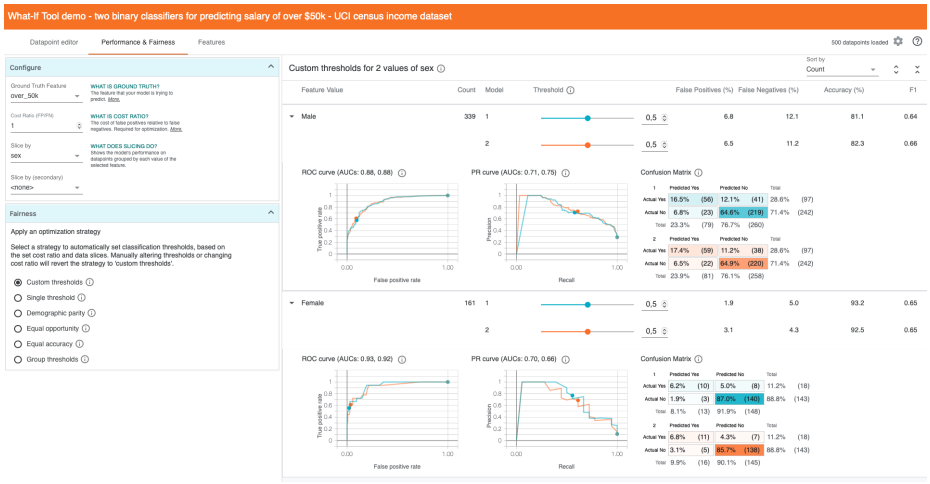

Abb. 9. Slicing

Zusammenfassend lässt sich sagen, dass sowohl der Linearklassifikator als auch das DNN ähnlich gute Leistungen erbracht haben. Das gesamte Beispiel für die Einkommensklassifizierung ist unter https://pair-code.github.io/what-if-tool/demos/uci.html zu finden.

Die nächsten beiden Experimente stellen Techniken vor, die aus dem Buch [3] zusammengefasst sind.

3.4 SHAP

SHAP ist eine modell-agnostische, post-hoc-Methode. Es ist eine lokale Interpretationsmethode, aber man kann sie auf alle Daten anwenden, um ein globales Verständnis zu erhalten. In diesem Paper werden die folgenden Plots erläutert:

- Summary Plot
- Partial Dependence Plot
- Force Plot und Decision Plot (als lokale Interpretationsmethode)

Der Summary Plot ist eine interessante Darstellung, um die Merkmale des Modells zu bewerten, da sie mehr Informationen liefert als die traditionelle Feature Importance:

- Feature Importance: Die Variablen sind in absteigender Reihenfolge ihrer Wichtigkeit sortiert.
- Auswirkung auf die Vorhersage: Die Position auf der horizontalen Achse zeigt an, ob die Datensatzinstanzen mehr oder weniger Einfluss auf die Vorhersage des Modells haben.

– Korrelation: Die Farbe zeigt an, ob es sich um einen hohen oder niedrigen Wert handelt. Die Korrelation eines Merkmals mit der Vorhersage des Modells kann also durch Auswertung seiner Farbe und der Auswirkung auf der horizontalen Achse analysiert werden.

Es wird beispielsweise in Abbildung 10 festgestellt, dass das Alter eine positive Korrelation mit dem Ziel hat, da die Auswirkung auf die Ausgabe mit zunehmendem Wert des Merkmals steigt.

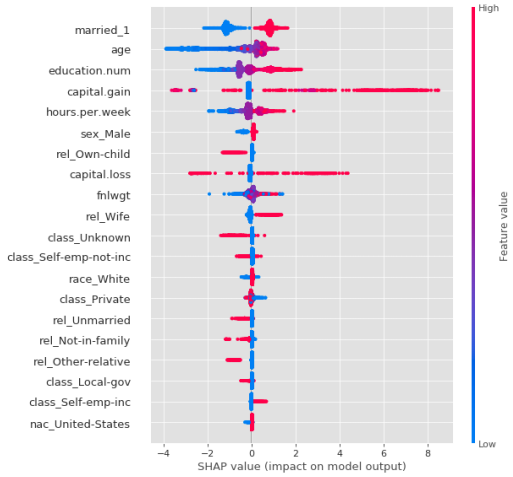

Abb. 10. Summary Plot (SHAP)

Der Partial Dependence Plot (PDP) zeigt den marginalen Effekt, den ein Merkmal individuell auf die Vorhersage hat. Dazu wird das Merkmal modifiziert, ceteris paribus, und die Veränderungen in der durchschnittlichen Vorhersage werden beobachtet. Der PDP kann angeben, ob die Beziehung zwischen dem Merkmal und dem Ergebnis linear, monoton usw ist.

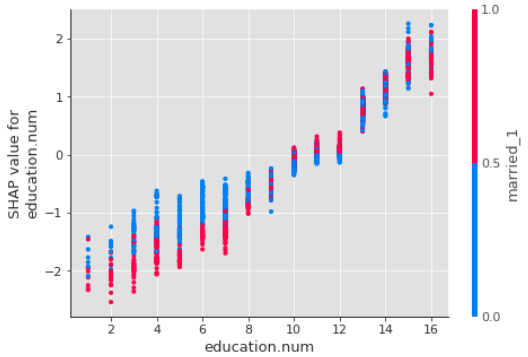

Abb. 11. PDP von Dauer der Ausbildung

Wie in Abbildung 11 dargestellt, besteht es ein linearer Zusammenhang zwischen den Bildungsjahren und der Wahrscheinlichkeit, mehr als 50.000 Dollar zu verdienen. Es wurde auch festgestellt, dass das Merkmal "Familienstand" am stärksten interagiert (das bedeutet, dass verheiratete Personen mit einer hohen Anzahl von Bildungsjahren eher mehr als 50.000 Dollar verdienen).

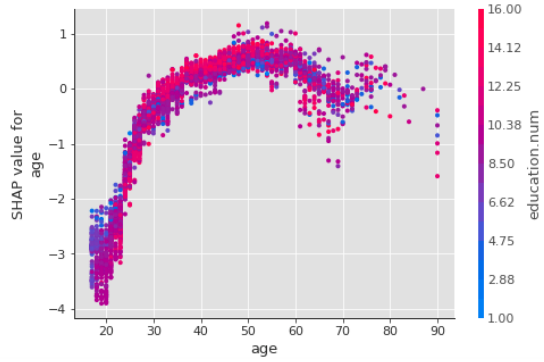

Abb. 12. PDP von Alter

Es scheint in Abbildung 12, dass die Wahrscheinlichkeit, mehr als 50.000 Dollar zu verdienen, in den 50ern am größten ist, und dass das Merkmal, mit dem "Alter" am meisten interagiert, ist "Bildungsjahren".

Der Force Plot (Abbildung 13) zeigt für jedes Merkmal an, welchen Einfluss es auf die Vorhersage hatte. Es sind zwei relevante Werte zu beachten: der "output value" (Modellvorhersage für die Instanz) und der "base value" (durchschnittliche Vorhersage für den gesamten Datensatz). Ein größerer Balken bedeutet eine höhere Auswirkung und die Farbe zeigt an, ob der Merkmalswert die Vorhersage vom "base value" in Richtung 1 (rot) oder 0 (blau) verschoben hat.

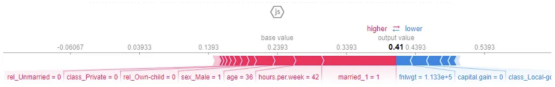

Abb. 13. Force Plot

Der Decision Plot (Abbildung 14) zeigt im Wesentlichen die gleichen Informationen wie das Force Plot. Die graue vertikale Linie ist der "base value" und die rote Linie zeigt an, ob jedes Merkmal den "output value" auf einen höheren oder niedrigeren Wert als die durchschnittliche Vorhersage verschoben hat. Diese Darstellung kann etwas übersichtlicher und intuitiver sein als die vorherige, vor allem, wenn es viele Merkmale zu analysieren gibt.

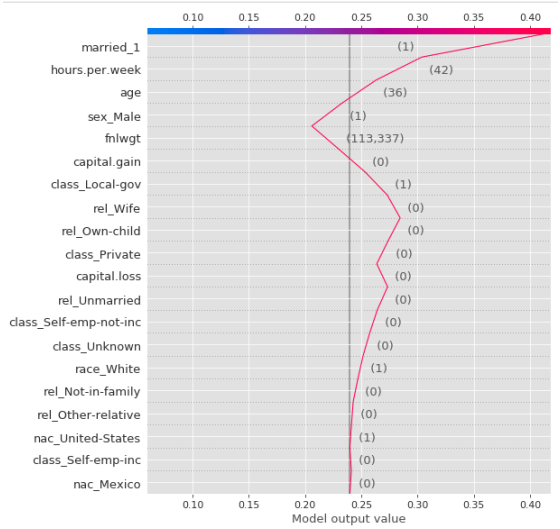

Abb. 14. Decision Plot

3.5 LIME

Ein weiteres bekanntes Werkzeug für lokale Interpretation ist LIME.
LIME analysiert, was mit den Modellvorhersagen geschieht, wenn die Eingabedaten verändert werden. Es erzeugt einen neuen Datensatz mit vertauschten Stichproben und den entsprechenden Vorhersagen des ursprünglichen Modells. Auf diesem synthetischen Satz trainiert LIME interpretierbare Modelle, die dann nach der Nähe der Stichprobeninstanzen zu der interessierenden Instanz gewichtet werden.
Die Erklärung für die Instanz X (z.B. wie in Abbildung 15) ist die des Ersatzmodells, das die Verlustfunktion (Leistungsmaß zwischen der Vorhersage des Ersatzmodells und der Vorhersage des ursprünglichen Modells) minimiert.

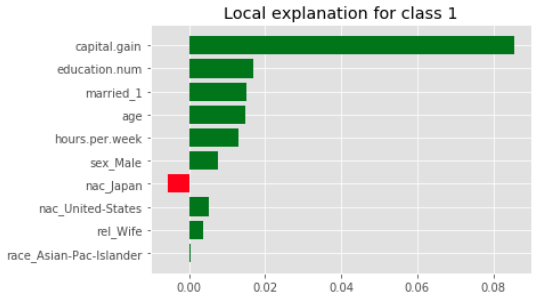

Abb. 15. Relative Wichtigkeit jedes Merkmals bei einem Beispiel

4 Fazit

Zusammengefasst haben die Experimente mit Explainable AI (XAI) unter Verwendung des Census-Einkommensdatensatzes wertvolle Einblicke in die Interpretierbarkeit und Transparenz von KI-Modellen bei der Einkommensprognose geliefert. Facets Dive ermöglichte es uns, die Daten zu visualisieren und relevante Merkmale für die Einkommensprognose zu identifizieren, wobei interessante Muster und Korrelationen aufgedeckt wurden. Die Untersuchung lokaler Interpretationen mit What-if-Tools beleuchtete, wie einzelne Datenpunkte die Modellvorhersagen beeinflussen, und ermöglichte es uns, die Modellleistung und Fairness in verschiedenen Untergruppen zu analysieren. SHAP (SHapley Additive exPlanations) lieferte detaillierte und intuitive Einblicke in die Bedeutung von Merkmalen und partielle Abhängigkeiten, was unser Verständnis des Entscheidungsprozesses des Modells förderte. Und nicht zuletzt erklärte LIME (Local Interpretable Model-agnostic Explanations) effektiv individuelle Vorhersagen, indem es interpretierbare Modelle generierte und die relative Bedeutung jedes Merkmals hervorhob.
Durch diese Forschung haben wir unser Verständnis der verfügbaren XAI-Methoden und ihrer potenziellen Anwendung in realen Szenarien verbessert, insbesondere im Kontext der Einkommensvorhersage im Wirtschaftsbereich. Die aus diesen Experimenten gewonnenen Erkenntnisse tragen zur Entwicklung transparenter und vertrauenswürdiger KI-Modelle bei. Diese interpretierbaren KI-Techniken können dazu beitragen, Fragen der algorithmischen Fairness anzugehen und ethische Überlegungen bei der Anwendung von KI-Systemen zu fördern. Insgesamt unterstreicht die Studie die Bedeutung von XAI, um KI zugänglicher und nachvollziehbarer zu machen und damit letztlich das Vertrauen in KI-Technologien zu stärken.

Literatur

1. Lichman, M., et al.: UCI machine learning repository. Irvine, CA, USA (2013)
2. Rothman, D.: Hands-On Explainable AI (XAI) with Python: Interpret, Visualize, Explain, and Integrate Reliable AI for Fair, Secure, and Trustworthy AI Apps. Packt Publishing Ltd (2020)
3. Molnar, C.: Interpretable Machine Learning. Lulu. com (2020)