

Fakultät Informatik

Verknüpfung von Datensätzen über Erwerbsbiographien und Stellenangeboten zum trainieren eines Machine Learning Modells zur Unterstützung bei der Arbeitsvermittlung.

Expose Bachelorarbeit im Studiengang Informatik

vorgelegt von

Robin Feldmann

Matrikelnummer 3538270 - Ca. 9000 Zeichen

© 2024

Dieses Werk einschließlich seiner Teile ist **urheberrechtlich geschützt**. Jede Verwertung außerhalb der engen Grenzen des Urheberrechtsgesetzes ist ohne Zustimmung des Autors unzulässig und strafbar. Das gilt insbesondere für Vervielfältigungen, Übersetzungen, Mikroverfilmungen sowie die Einspeicherung und Verarbeitung in elektronischen Systemen.

Inhaltsverzeichnis

| | |
|--------------------------------|---|
| Ausgangslage | 2 |
| Zielsetzung | 3 |
| Forschungsstand | 3 |
| Methodik | 4 |
| Vorgehensweise | 5 |
| Literaturverzeichnis | 7 |

Ausgangslage

Arbeit ist ein zentraler Bestandteil unseres alltäglichen Lebens und unserer Gesellschaft. Folglich führt Arbeitslosigkeit bei Betroffenen zu gravierenden gesundheitlichen und sozialen Nachteilen. Auch gesamtgesellschaftlich hat Arbeitslosigkeit viele negative Folgen [1]. Die Bundesagentur für Arbeit macht es sich zum Ziel, diese Probleme durch Arbeitsvermittlung zu bekämpfen [2] und schreibt dazu in ihrer Präambel: *Die Arbeitsvermittlung ist ein wichtiger Bestandteil der staatlichen Daseinsvorsorge und gehört zum Kern moderner Dienstleistungen am Arbeitsmarkt.* [3]. Arbeitsvermittlung meint dabei das Zusammenführen von Informationen über Angebot und Nachfrage der Arbeitsmärkte [4].

Dieser Prozess beginnt dabei mit der Erfassung persönlicher Merkmale und Fähigkeiten der Arbeitssuchenden. Im nächsten Schritt werden dem Arbeitssuchenden passend zu seinem Profil Stellenangebote vorgeschlagen [3]. Dieser Schritt wird in der Arbeitsmarktökonomik als Matching Prozess bezeichnet. Der Matching Prozess kann als mathematische Abbildung abstrahiert werden. Die Parameter dieser Abbildung sind Informationen

sowohl über den Arbeitssuchenden als auch über das Stellenangebot. Der Rückgabewert ist die Wahrscheinlichkeit, dass der Arbeitssuchende in der Stelle arbeiten wird [5]. Für das Schätzen so einer Abbildung im Kontext des Arbeitsmarktmatchings eignen sich Methoden des maschinellen Lernens [6].

Zielsetzung

Das Ziel der Bachelorarbeit ist im Zuge eines Praktikums beim Institut für Arbeitsmarkt- und Berufsforschung in Zusammenarbeit mit Fr. Sabrina Mühlbauer zwei Datensätze miteinander zu identifizieren (vgl. **Methodik**). Dann soll anhand dieser miteinander identifizierten Datensätze ein Modell des maschinellen Lernens entwickelt werden. Dieser soll es ermöglichen bei der Arbeitsvermittlung für individuelle Arbeitssuchende anhand von Informationen über deren Lebenslauf und Qualifikation offene Arbeitsstellen daraufhin zu bewerten, wie gut sie zum Arbeitssuchenden passen.

Forschungsstand

Die Arbeit wird im Zuge eines Praktikums beim Institut für Arbeitsmarkt- und Berufsforschung unter der Betreuung von Fr. Sabrina Mühlbauer verfasst. Diese Arbeit soll auf der Veröffentlichung *Machine Learning for Labour Market Matching* aufbauen [6]. In dieser Arbeit wird ein Ansatz beschrieben anhand von statistischen Methoden und Methoden des maschinellen Lernens Arbeitssuchende bei der Auswahl des zukünftigen Berufsfelds zu unterstützen. Hierfür wird ein Datensatz über Berufsbiographien verwendet um anhand der Berufshistorie einer Person Aussagen über das zukünftige Berufsfeld treffen zu können. Hierbei wird die Effektivität von statistischen Methoden mit dem Random Forest und dem K-Nearest-Neighbours Algorithmus verglichen. Fr. Mühlbauer kommt zu dem Ergebnis, dass der Random Forest Algorithmus hierfür am besten geeignet ist [6]. In anderen Veröffentlichungen werden auch sogenannte *Learning to Rank* Ansätze untersucht, um bei Online Jobbörsen Vorschläge für Arbeitssuchende zu bewerten [7]. Es soll für einen Arbeitssuchenden ermittelt werden, in wie fern ein spezifisches

Jobangebot zu diesem passt um so Arbeitsvermittler bei der Auswahl von Jobvorschlägen zu unterstützen. Außerdem soll mit dem XGBoost Algorithmus eine Verbesserung des Random Forest getestet und evaluiert werden.

Methodik

Datengrundlage für die Arbeit ist der Datensatz über Integrierte Erwerbsbiografien (IEB) des Instituts für Arbeitsmarkt und Berufsforschung. Dieser Datensatz enthält Erwerbsverläufe aller Personen die in Deutschland seit 1975 sozialversicherungspflichtig beschäftigt war. Er enthält Informationen über Geburtsjahr und Ausbildung, Leistungsbezug und Beschäftigung [8]. Ein weiterer Datensatz, zu dem zum mir Zeitpunkt dieses Expose noch keine genaueren Informationen erhältlich sind wird historische Daten über offene Stellenangebote beinhalten. In diesem Datensatz wird zu den Stellenangeboten Informationen über deren Erstellungsdatum und Aufslaufdatum, über den Betrieb und in welche Kategorie die Arbeitsstelle fällt, beinhalten. In einem ersten Schritt sollen die Erwerbsbiografien des IEB anhand der Betriebe, der Zeitpunkte und der Art der Beschäftigung mit historischen Stellenangeboten identifiziert werden. Es werden also paare von Datenpunkte erstellt, bei denen sichergestellt ist, dass eine spezifische Arbeitsbiografie eine spezifische Tätigkeit beinhaltet, die vorher als Jobangebot ausgeschrieben war. Diese Kombination aus zusammengehörigen Erwerbsbiografien und Stellenangeboten wird die Grundlage für das Training eines Algorithmus des maschinellen Lernens bilden.

Da als Vorbereitung bereits passende Verknüpfungen zwischen Informationen über Job-suchenden aus dem Datensatz über Erwerbsbiographien und Informationen aus dem Datensatz über Stellenangebote hergestellt wurden, handelt es sich um *überwachtes Lernen*. Das Schätzen von Wahrscheinlichkeiten also von reellwertigen Labels bezeichnet man als Regressionsproblem [9]. Im Laufe der Arbeit sollen verschiedene Algorithmen des überwachten Lernens zum lösen von Regressionsproblemen verglichen und anhand von vorher definierten Metriken bewertet werden. Zum jetzigen Zeitpunkt ist geplant drei solcher Algorithmen zu vergleichen: Support Vector Machine (SVM), Random Forest (RF) und XGBoost (XGB). Diese Liste kann und soll aber im Laufe der Arbeit noch erweitert werden. In der Recherchephase der Bachelorarbeit sollen weitere mögliche Algorithmen anhand des aktuellen Forschungsstands ermittelt werden.

Beim Support Vector Machine Algorithmus werden Datenpunkte in einem Vektorraum dargestellt. Es wird versucht möglichst große Hyperebenen zu finden, welche Klassen von Datenpunkten separieren. Beim Klassifizieren neuer Datenpunkte wird dann überprüft auf welcher Seite der Hyperebene dieser Datenpunkt liegt [10].

Der Random Forest ist ein Algorithmus bei dem der Datensatz zufällig in mehrer Samples zerteilt wird. Auf jedem dieser Samples wird ein Entscheidungsbaum erstellt und bei der Klassifikation wird der Durchschnitt der einzelnen Entscheidungen der Entscheidungsbäume ermittelt [11].

Der XGBoost Algorithmus ist ein *gradient boosting* verfahren. Hierbei werden Modelle iterativ erweitert um ihre jeweiligen Schwächen auszubauen. Da hierbei auch mit Entscheidungsbäumen gearbeitet wird, ist er eine Erweiterung des Random Forest Algorithmus [12].

Zur Evaluation der Modelle wird der Datensatz in einen Trainingsdatensatz und einen Testdatensatz aufgeteilt. Hier sind Aufteilungen von 80% als Trainingsdaten und 20% als Testdaten üblich. Zur Bewertung werden Datenpunkte des Testdatensatzes ausgewertet und die so erhaltenen Werte mit den tatsächlichen Labels verglichen. Diese werden Anhand gebräuchlicher Metriken bewertet. Hierfür üblich sind Mean Squared Error und der R-Score. Mean Squared Error meint hier den durchschnittlichen aufsummierten Quadratischen Fehler. Der R-Score berechnet sich als eins minus der Mean Squared Error geteilt durch die quadratische Standardabweichung [13]. Weitere Metriken zur Bewertung der Modelle sollen im Laufe der Recherchephase ermittelt werden.

Vorgehensweise

Der Prozess eine Bachelorarbeit zu verfassen ist aufwendig. Um dieses große Aufgabe sinnvoll anzugehen wird die Arbeit in kleiner Abschnitte unterteilt. Die Bearbeitung der Bachelorarbeit soll in einer groben Einteilung fünf Schritte umfassen.

In der Recherchephase soll der aktuelle Forschungsstand ermittelt werden. Außerdem sollen Informationen zur aktuellen Methodik gefunden werden. Für diese Phase ist in etwa einen Monat Zeit eingeplant.

In der zweiten Phase werden die Datensätze begutachtet. Zuerst werden die Datensätze explorativ analysiert. Hier soll die Qualität der Datensätze überprüft werden. Auch Informationen über die einzelnen Felder der Daten müssen gesammelt werden. Dann müssen die beiden Datensätze miteinander Identifiziert werden. Hierbei soll herausgefunden werden welche Stellenausschreibung mit welcher Erwerbsbiografie in Zusammenhang steht.

Da die Identifikation nur aufgrund der Betriebe und Zeitpunkte passiert, ist hier besondere Vorsicht geboten. Auch hierbei muss überprüft werden wie gut diese Identifikation funktioniert. Für diese Phase ist auch etwa ein Monat Zeit eingeplant.

Anhand im ersten Schritt recherchierten Methodik sollen nun verschiedene Modelle des maschinellen Lernens anhand der im zweiten Schritt identifizierten Datensätze trainiert werden. Auch müssen verschiedene Parameter der Algorithmen angepasst und optimiert werden. Dafür sind bis zu zwei Monate an Zeit vorgesehen.

Bei der Evaluierung sollen die verschiedenen im dritten Schritt trainierten Modelle verglichen und bewertet werden. Dies geschieht anhand von Metriken, die bei der Recherche der Methodik im ersten Schritt gefunden wurden. Nun sollen Vor- und Nachteile der Modelle anhand dieser Metriken gegenübergestellt werden. Dies sollte ungefähr zwei Wochen Zeit in Anspruch nehmen.

In der letzten Phase sollen die in den vorherigen Schritten gewonnenen Erkenntnisse niedergeschrieben werden. Für das Verfassen der Bachelorarbeit sind in etwa ein Monat an Zeit vorgesehen.

Verläuft alles nach Plan sollte die Bachelorarbeit also in etwa fünf Monaten zu schaffen sein.

Literaturverzeichnis

- [1] F. OSCHMIANSKY, J. BERTHOLD ; BUNDESZENTRALE FÜR POLITISCHE BILDUNG (Hrsg.): *Folgen der Arbeitslosigkeit*. [Online]. <https://www.bpb.de/themen/arbeit/arbeitsmarktpolitik/305686/folgen-der-arbeitslosigkeit>. – Zugriff am: 2023-05-25.
- [2] BUNDESAGENTUR FÜR ARBEIT: *DER WEG ZUR ARBEIT Die Bundesagentur für Arbeit stellt sich vor*. [Online]. https://www.arbeitsagentur.de/datei/dok_ba026145.pdf. – Zugriff am: 2023-05-25.
- [3] OLAF SCHOLZ: *Zeitenwende - Regierungserklärung in der Sondersitzung zum Krieg gegen die Ukraine vor dem Deutschen Bundestag*. [Online]. <https://www.bundesregierung.de/resource/blob/992814/2131062/78d39dda6647d7f835bbe76713d30c31/bundeskanzler-olaf-scholz-reden-zur-zeitenwende-download-bpa-data.pdf>. Version: 2022. – Zugriff am: 2024-08-01.
- [4] F. OSCHMIANSKY ; BUNDESZENTRALE FÜR POLITISCHE BILDUNG (Hrsg.): *Die öffentliche Arbeitsvermittlung*. [Online]. <https://www.bpb.de/themen/arbeit/arbeitsmarktpolitik/317004/die-oeffentliche-arbeitsvermittlung/#node-content-title-6>. – Zugriff am: 2023-05-25.
- [5] WOLFGANG, F.: Der Matching-Prozess. In: *Arbeitsmarktökonomik. Springer-Lehrbuch*. Berlin, Heidelberg : Springer, 1994, Kapitel 6, S. 195ff.
- [6] S. MÜHLBUER, E. WEBER: Machine Learning for Labour Market Matching. In: *Articles on labour market issues* (2022), Nr. 3, S. 5.
- [7] MARIOS KOKKODIS, PANAGIOTIS PAPADIMITRIOU: Hiring Behavior Models for Online Labor Markets. (2015).
- [8] BUNDESMINISTERIUM FÜR ARBEIT UND SOZIALES: *Stichprobe der Integrierten Erwerbsbiografien des IAB*. [Online]. <https://kooperationen.zew.de/dfgflex/links/datensaetze-deutschland/>

stichprobe-der-integrierten-erwerbsbiografien-des-iab-iebs. – Zugriff am: 2023-05-25.

- [9] M. MOHRI, A. ROSTAMIZADEH, A. TALWALKAR: *Foundations of Machine Learning*. London, England : MIT Press, 2018.
- [10] A. MAMMONE, M. TURCHI, N. CHRISTIANINI: Support vector machines. In: *Journal of statistical software* 1 (2009).
- [11] LEO BREIMAN: Random Forests. In: *Machine Learning* 45 (2001), S. 5–35.
- [12] T. CHEN, C. GUESTRIN: XGBoost: A Scalable Tree Boosting System. (2016).
- [13] A. BOTCHKAREV: Performance Metrics in Machine Learning Regression, Forecasting and Prognostics: Properties and Typology. (2018), S. 4.