Digital Business University of Applied Sciences

Studiengang: Data Science and Business Analytics (B. Sc.)

Dozent: Marcel Hebing

Hauptteil: 10 Seiten (laut Absprache mit M. Hebing am 09.11.2022)

Studienarbeit ADS-04 - Machine Learning

Human Resources Analytics

Foto: www.pexels.com

Eingereicht von: Conny Brintzinger

Matrikelnummer: 190044

Datum: 06.11.202

Abstract

Der Fachkräftemangel zwingt Unternehmen, ihre Recruiting-Strategien zu überdenken und neue Wege im Werben um neue Mitarbeiter zu beschreiten. Das hier betrachtete, im Bereich Big Data und Data Science tätige Unternehmen bietet erfolgreich Kurse und Ausbildungen an und richtet sich an Interessenten, die potenziell auch Kandidaten zur Übernahme ins Unternehmen sein könnten, da sie durch ihre Teilnahme an den Kursen bereits Data Science Expertise belegt haben und eine positive Wahrnehmung des Unternehmens vorausgesetzt werden kann. Bei der Einschreibung zu den Kursen werden für jeden Kandidaten persönliche Daten zu Demografie, Ausbildung und Berufserfahrung erhoben. Diese Datenbasis liefert wichtige Erkenntnisse, die das Unternehmen zur Rekrutierung neuer Mitarbeiter aus dem Kursteilnehmer-Pool nutzen möchte. Zur Vorhersage der Jobwechsel-Wahrscheinlichkeit wurden unterschiedliche Modelle evaluiert. Das beste Model liefert eine Accuracy von 78%, Precision von 86% und Recall von 79% für Vorhersagen, ob ein Kandidat einen Job-Wechsel anstrebt oder nicht. Der City Development Index (CDI) ist dabei der wichtigste Teiber für die Entscheidung pro oder contra Jobwechsel. Beim Recruiting sollte mit diesen Vorzügen aktiv geworben werden, sofern das Unternehmen einen Standort mit hohem CDI bieten kann.

Inhaltsverzeichnis

Anhang: Jupyter Notebooks

Notebook 1: EDA_HR_Analysis.ipynb

Notebook 2: ML_HR_Analysis.ipynb

Abbildungsverzeichnis

1. Einleitung

Es herrscht ein Mangel an IT-Expertinnen und Experten in Deutschland. Laut einer repräsentativen Bitkom-Befragung von mehr als 850 Unternehmen ist die Zahl freier Stellen für IT-Fachkräfte im Jahr 2021 um 12% auf 96.000 vakante Stellen gestiegen. Etwa zwei Drittel der befragten Unternehmen erwarten für die Zukunft eine Verschärfung des Fachkräftemangels (Paulsen & Holdampf-Wendel, 2022). Besonders gefragt sind IT-Security-Fachkräfte, Softwareentwickler, IT-Architekten sowie Data Scientists und Big Data Experten (Schmole, kein Datum). Im Schnitt dauert es heute 180 Tage, eine offene IT-Stelle zu besetzen (Kranz, 2022). Viele Stellen bleiben unbesetzt. Zurückzuführen ist der steigende Bedarf an IT-Fachkräften vor allem auf die Digitalisierung. Digitaler Wandel und Transformation sind branchenübergreifende Entwicklungen, denen sich kein Unternehmen verschließen kann (Kranz, 2022).

Umso wichtiger erscheint ein effektives Recruiting. Klassische Methoden der Personalgewinnung, wie Stellenanzeigen, Personalberatung oder Headhunter können Einstellungskosten in Höhe von 30 bis 50% des Jahresgehaltes der zu besetzenden Stelle verursachen (stellenanzeigen.de, 2022) und führen oft nicht zum gewünschten Erfolg. Insofern gilt es, neue Wege der Personalgewinnung zu beschreiten.

Das hier betrachtete, im Bereich Big Data und Data Science tätige Unternehmen bietet erfolgreich Kurse und Ausbildungen an und richtet sich somit an Interessenten, die potenziell Kandidaten zur Übernahme ins Unternehmen sein könnten, da sie durch ihre Teilnahme an den Kursen bereits Data Science Expertise und eine positive Wahrnehmung des Unternehmens belegt haben. Bei der Einschreibung zu den Kursen werden für jeden Kandidaten persönliche Daten zu Demografie, Ausbildung und Berufserfahrung erhoben. Diese Datenbasis kann als wichtige Ressource zum Recruiting neuer Mitarbeiter aus dem Kursteilnehmer-Pool angesehen werden.

Forschungsfragen:

1. Kann die Wahrscheinlichkeit, dass ein Kandidat einen Jobwechsel anstrebt, prognostiziert werden?

2. Welche Faktoren beeinflussen die Entscheidung potenzieller Bewerber am stärksten?

2. Daten und Methoden

Daten-Grundlage:

Link: https://www.kaggle.com/datasets/arashnic/hr-analytics-job-change-of-data-scientists

Datum des Herunterladens: 12.07.2022

Beteiligte Personen: Möbius (Eigentümer)

Lizenz: CC0, Public Domain

Ursprüngliche Quelle: https://datahack.analyticsvidhya.com/contest/janatahack-hr-analytics/True/#ProblemStatement

Methodik der Sammlung: Änderung der Daten durch Augmentation

Die Daten wurden aufgeteilt in Trainingsdaten (aug_train.csv) und Testdaten (aug_test.csv) zur Verfügung gestellt. Die Zielgröße ist in der ersten Spalte des Trainingsdatensatzes hinterlegt. Für die Testdaten hingegen wurden die Zielgrößen in einer zusätzlichen Datei (sample_submission.csv) gespeichert. Die Zielgröße (Target) ist ein binärer Wert von 1 oder 0. Alle Zielgrößen in der Datei sample_submission.csv weisen jedoch den Wert 0.5 als Zielgröße auf. Insofern ist der Test-Datensatz nicht für die Analyse geeignet und wurde verworfen. Für die Analyse wurde ausschließlich die aug_test.csv-Datei (19.158 Beobachtungen) verwendet und nach dem Zufallsprinzip in 80% Trainings- und 20% Testdaten aufgeteilt:

Trainingsdaten (df_train): 14368 Zeilen und 14 Spalten

Testdaten (df_test): 4790 Zeilen und 14 Spalten

Eine Auflistung aller Spalten und deren Bedeutung kann den Notebooks oder der Readme-Datei entnommen werden.

Fehlende Daten / Kodierungsprobleme:

Eine erste Analyse der Daten erfolgte via Pandas Profiling Report und ergab insgesamt 7.2% fehlende Werte im Datensatz, im Einzelnen:

gender: 23,7% (3410) fehlende Werte

enrolled_university: 2,0% (290) fehlende Werte

education_level: 2,4% (342) fehlende Werte

major_discipline: 14,7% (2105) fehlende Werte

experience: 0,2% (46) fehlende Werte

company_size: 31,1% (4475) fehlende Werte

company_type: 32,2% (4628) fehlende Werte

last_new_job: 2,2% (312) fehlende Werte

Darüber hinaus ist die Spalte *city* durch eine hohe Kardinalität gekennzeichnet.

Der Datensatz ist imbalanced und weist eine starke Verzerrung zugunsten der Kandidaten, die keinen Jobwechsel anstreben, auf:

Target 0: 75,1% entspricht 10.797 Kandidaten, die keinen Jobwechsel anstreben

Target 1: 24,9% entspricht 3.571 Kandidaten, die offen für einen Jobwechsel sind

Daten-Transformation:

Um Machine Learning Ansätze etablieren zu können, müssen die Daten so vorbereitet und transformiert werden, dass sie keine fehlenden und ausschließlich numerische Werte bzw. Zahlen zwischen 0 und 1 enthalten. Für numerische und kategoriale Daten stehen jeweils unterschiedliche Möglichkeiten zur Transformation zur Verfügung. Zum Preprocessing der Trainingsdaten wurden jeweils unterschiedliche Ansätze getestet:

<u>Notebook 1 (EDA_HR_Analysis.ipynb)</u>

Es wurde ein Preprocessing-Workflow definiert, der jede Spalte des Datensatzes entsprechend des Datentyps sowie Art und Verteilung der Datenmerkmale behandelt. Fehlende numerische Werte wurden mit dem jeweiligen Durchschnitt ersetzt, während fehlende kategoriale Daten mit dem am häufigsten vorkommenden Wert aufgefüllt wurden. Zur Kodierung der kategorialen Merkmale wurden Labelencoder eingesetzt, die jeder Kategorie einen Zahlenwert zuweisen, also beispielsweise für die fünf Kategorien in *enrolled_university* die numerischen Werte 0, 1, 2, 3 und 4.

Der vorliegende Datensatz ist imbalanced hinsichtlich der Target-Größe, was zur Verzerrung der Analyseergebnisse führen kann. Es wurden Elemente für die Minderheitenklasse synthetisch mittels Oversampling erzeugt, um den Datensatz auszubalancieren.

Notebook 2 (ML_HR_Analysis.ipynb)

Im zweiten Notebook wurden die Preprocessing- Schritte in Form von Pipelines für Maschinelles Lernen organisiert und der Arbeitsprozess damit automatisiert. Es wurden drei Preprocessing-Pipelines entwickelt, die unterschiedliche Methoden der Daten-Transformation repräsentieren. Anschließend wurde die Performance der Pipelines in Verbindung mit unterschiedlichen ML-Modellen, die im folgenden Abschnitt näher erläutert werden, evaluiert.

Transformation Pipeline 1 beinhaltet den SimpleImputer (numeric and categorical values), StandardScaler (numeric values) und OneHotEncoder (categorical values).

Transformation Pipeline 2 ist ähnlich wie Pipeline 1 aufgebaut, enthält jedoch statt dem StandardScaler einen MinMaxScaler zur Transformation der numerischen Werte

Transformation Pipeline 3 beinhaltet wie Pipeline 2 den SimpleImputer und MinMaxScaler. Beim Encoding der kategorialen Daten wurde jedoch in zwei Gruppen unterschieden. Für die ordinal skalierten Werte wurde der OrdinalEncoder eingesetzt, um die Bedeutung deren Sortierung zu erhalten. Alle andren kategorialen Daten wurden mittels OneHotEncoder transformiert. Die Spalten *enrollee_id* und *company_type* (mehr als 20% fehlende Werte) wurden gelöscht.

Anschließend wurde ein Oversampling in allen drei transformierten Datensätzen etabliert, um die Datensätze auszubalancieren und eine gleiche Verteilung von Kandidaten, die offen für eine Jobwechsel sind und solchen, die keinen Wechsel anstreben, zu erreichen.

Nach Abschluss des Preprocessing liegen demnach zur Analyse drei unterschiedlich vorbereitete Datensätze vor, jeweils als Version mit oder ohne Oversampling,

Analyse-Methoden und Modelle:

Nachdem die vorliegenden Daten bereinigt und transformiert vorlagen, wurden zur Beantwortung der Forschungsfrage geeignete Methoden und Modelle ausgewählt und angewandt.

In Notebook 1 wurde eine Vorhersage mittels GridSearch / KNeighborsClassifier angestoßen, um die Wahrscheinlichkeit, dass ein Kandidat einen Jobwechsel anstrebt, vorauszusagen. Darüber hinaus wurden Decision Trees und Logistic Regression genutzt, um den Einfluss der einzelnen Variablen auf die Entscheidung zu visualisieren.

Die in Notebook 2 transformierten Datensätze wurden Decision Tree, Logistic Regression und Random Forest Classifier-Modellen zugeführt und unterschiedlich tiefe Modelle (1-8 Layer) getestet, um den jeweiligen Score (Accuracy) zu bestimmen und damit die Güte der Modelle im Zusammenspiel mit den unterschiedlich transformierten Daten zu evaluieren. Via GridSearch wurde anhand ausgewählter Preprocessing Pipelines jeweils eine Confusion Matrix erstellt, um die Ergebnisse detaillierter in Hinblick auf Recall und Precision bewerten zu können. Im folgenden Abschnitt sollen die Resultate dargestellt und verglichen werden.

3. Ergebnisse

Zunächst werden die Ergebnisse aus Notebook 1 dargestellt:

Logit Regression

Um der Frage nachzugehen, welche Faktoren den stärksten Einfluss auf die Entscheidung eines Kandidaten haben, ob er einen Jobwechsel anstrebt oder nicht, wurden mehrere Logit-Regression Modelle getestet, wobei Model 4 mit dem höchsten Pseudo R-squared von 0.151 am besten abgeschnitten hat. Der Wert besagt, dass nur 15.1% der Varianz der Standardvariablen durch das Modell erklärt werden, was ein relativ geringer Wert ist. Liegen die p-Werte einer Variable unter 0,05, ist sie als relevanter Einflussfaktor einzustufen. Besonders niedrige p-Werte und damit hohen Einfluss haben die Variablen *experience* (unterschiedlich je nach Jahren Erfahrung). Die Variable *last_new_job* ist hochrelevant, wenn der Kandidat noch keinen Job hatte, also gerade den Einstieg ins Berufsleben plant. Zusätzlich werden *experience*, *enrolled_university*, *education_level* und *city_development_index* als hoch signifikant eingestuft. Der *company_type* scheint kaum relevant, während die *company_size* signifikanten Einfluss hat.

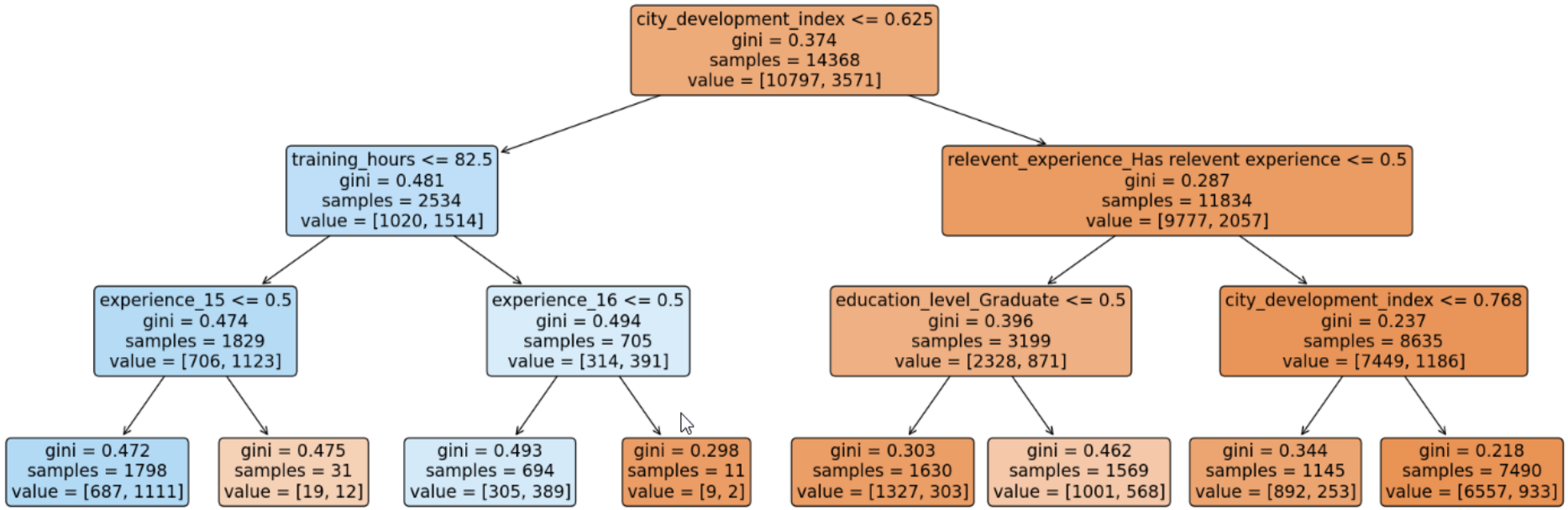

Abbildung 1: Decision Tree Plot mit einer Tiefe=3

Decision Tree

Um den Weg der Entscheidungsfindung graphisch darzustellen, wurde ein Decision Tree erstellt, welcher in *Abbildung 1* zu sehen ist. Das hierarchische, baumartig aufgebaute Diagramm stellt einen gerichteten Entscheidungsweg dar. Ausgehend vom initialen Wurzelknoten schließen sich drei Entscheidungsebenen an. Jeder Entscheidungsknoten repräsentiert eine binäre ja/nein-Entscheidung- also ja, der Kandidat ist offen für einen Job-Wechsel oder nein, der Kandidat strebt keinen Jobwechsel an. Wie im Diagramm zu sehen, stellt der *city_development_index* des Kandidaten den Wurzelknoten dar und somit den größten Einflussfaktor. Liegt der Index unter 0.625, sind 37.4% der Kandidaten bereit für einen Jobwechsel. Von diesen Kandidaten wünschen 48.1% einen Jobwechsel, wenn sie 82.5 oder mehr Trainingsstunden absolviert haben. Auf der dritten Ebene spielt die Berufserfahrung eine Rolle. Dem rechten Pfad folgend, wird ersichtlich, dass auch die relevante Erfahrung einen großen Einfluss hat. Kandidaten, die bisher keine relevante Erfahrung sammeln konnten und deren *city_development_index* unter 0.768 liegt, sind ebenfalls an einem Jobwechsel interessiert.

Feature Importance

Eine Evaluation der Feature Importance der Variablen bestätigt die zentrale Bedeutung des City-Development-Indexes der Kandidaten. Mit deutlichem Abstand folgen die Merkmale company_size und relevant_experience.

Voraussagen mittels KNeighborsClassifier / Grid Search

Die Predictions des KNN-Classifiers lieferten eine Accuracy von 78%, eine Precision von 86% und einen Recall in Höhe von 79%. Für die hier behandelte Frage ist der Recall das aussagekräftigste Maß, um die Qualität der Prognosen zu beschreiben. Es sollen möglichst viele der Kandidaten identifizierte werden, die offen für einen Stellenwechsel sind. Wenn einige Kandidaten als wechselwillig eingestuft werden, die es nicht sind, ist das ein vergleichsweise kleines Problem und besser, als potenzielle neue Mitarbeiter für das Unternehmen nicht zu finden. Die Ergebnisse wurden als Confusion Matrix dargestellt, welche in *Abbildung 2* zu sehen ist. Es wurden 2007 Kandidaten richtig erkannt, die tatsächlich keinen Jobwechsel anstreben (true negative) und 2471, die einen Jobwechsel anstreben (true positive). Weitere 870 Kandidaten, die tatsächlich nicht wechseln möchten, wurden als

jobsuchend eingestuft (false positive). Darüber hinaus wurden 405 Kandidaten, die wechseln möchten, nicht als jobsuchend erkannt (false negative).

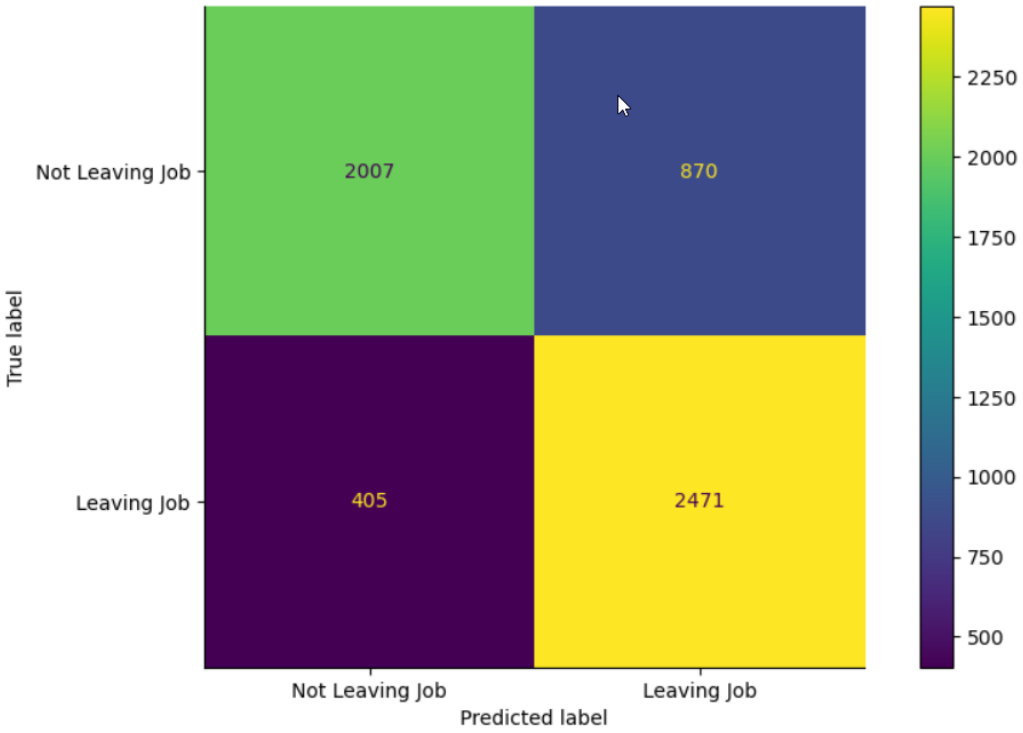

Abbildung 2: Confusion Matrix aus Notebook 1 (Grid Search / KNeighborsClassifier)

In Notebook 2 wurde mit verschiedenen Preprocessing Pipelines gearbeitet und deren Performance anhand des Scorings unterschiedlicher Prediction-Modelle bewertet, mit folgenden Ergebnissen:

Oversampling Test

Anhand unterschiedlich tiefer Logistic Regression- und Decision Tree Classifier-Modelle wurden die Score-Werte der Pipelines 1 bis 3 jeweils mit und ohne Oversampling verglichen. Durch das Oversampling konnte hier keine Verbesserung der Werte erzielt werden. Den höchsten Accuracy-Score erzielte der Decision Tree Classifier mit Tiefe 4 und Pipeline 3 (Score: 0.7933).

Model / Preprocessing Pipeline Kombinationen

Die vorbereiteten Daten aus Pipeline 1, 2 und 3 (ohne Oversampling) wurden nun an drei unterschiedliche Vorhersage-Modelle (DecisionTreeClassifier, RandomForestClassifier,

LogisticRegression) übergeben und für eine Tiefe von 1 bis 8 Layers getestet. Die besten Ergebnisse erzielten folgende Kombinationen:

Pipeline: 3 DecisionTree, i=6: Training Score 0.8035 vs Test Score 0.7891

Pipeline: 2 DecisionTree, i=7: Training Score 0.8050 vs Test Score 0.7852

Traings- und Testscore liegen nah beieinander, was für ein gutes Model ohne Overfitting spricht.

Confusion Matrix

Auf Grundlage der Ergebnisse wurde zu den beiden besten Ergebnissen sowie zusätzlich zu Pipeline 3 mit Oversampling jeweils eine Confusion Matrix erstellt, um die Ergebnisse detaillierter bewerten zu können. Es zeigt sich, dass Pipeline 3 mit Oversampling die beiden anderen Pipelines in Hinblick auf den Recall outperformt.

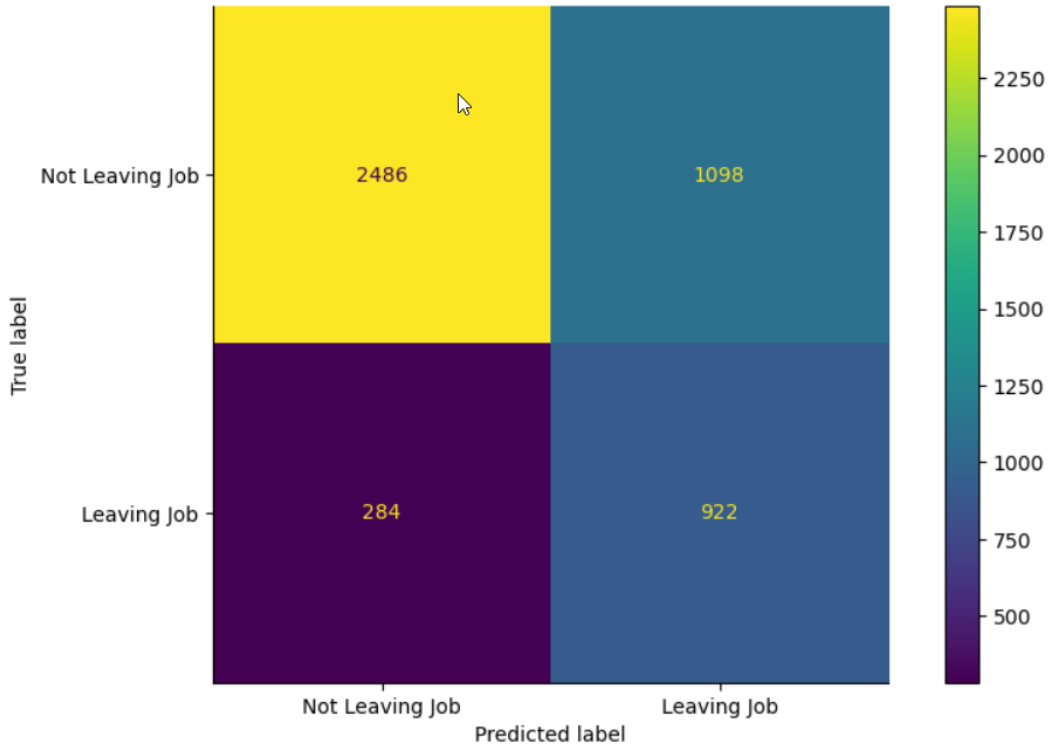

Abbildung 3: Confusion Matrix aus Notebook 2 (Grid Search / DecisionTreeClassifier mit den transformierten Daten der Preprocessing Pipeline 3 inklusive Oversampling)

Wie in *Abbildung 3* zu erkennen, liegt die Accuracy zwar nur bei 71% und damit im Vergleich zu den beiden anderen Modellen (75 und 77%) niedriger, erreicht aber deutlich bessere Werte für Recall und True Positive-Rate. Dieser Effekt ist vermutlich dem Oversampling zuzuschreiben, da dadurch beide Target-Größen gleiche Gewichtung erhalten.

4. Diskussion

Es wurden verschiedene Herangehensweisen- sowohl bezüglich des Preprocessings der Daten als auch der Modelle zur Berechnung der Wahrscheinlichkeit, dass ein Kandidat einen Jobwechsel anstrebt, getestet. Es konnten viele gute Lösungen mit hohen Accuracy-Werten gefunden werden. Entscheidend ist jedoch in erster Linie die True-Positive-Rate, denn es sollen möglichst viele Jobsuchende und Kandidaten, die einem Jobwechsel aufgeschlossen gegenüberstehen, gefunden werden. Für diese Aufgabe hat sich der KNN-Classifier mit Grid Search, der unter Punkt 1.8 in Notebook1 zu finden ist, als am besten geeignet erwiesen. Von denen, die einen Jobwechsel planen, werden 2471 gefunden und nur 405 fehlerhaft gelabelt. Das kann als sehr gutes Ergebnis angesehen werden, denn insbesondere, wenn es um die Vorhersage menschlichen Verhaltens geht, stoßen Modelle oft an Grenzen. Menschen treffen ihre Entscheidungen auf Basis vielfältiger Beweggründe, die sehr individuell sind und oft nicht im Modell erfasst werden können.

Darüber hinaus wurde als Hauptgrund für einen Jobwechsel ein niedriger City Development Index (CDI) von unter 0.625 ermittelt. Der Index umfasst wichtige Merkmale zum Lebensstandard, wie Infrastruktur, Bildung, Wasserqualität und Gesundheitswesen. Wenn das Unternehmen mit einem Standort mit hohem CDI werben kann, sollte darauf ein Schwerpunkt gelegt und beim Recruiting mit diesen Vorzügen aktiv geworben werden. Darüber hinaus bieten Berufseinsteiger ohne relevante Erfahrung eine interessante Zielgruppe. Aufgrund der Teilnahme an den angebotenen Kursen ist von einer positiven Wahrnehmung des Unternehmens auszugehen, sodass ein Angebot zum Direkteinstieg lohnend erscheint. Im Bereich 0 bis 6 Jahre Erfahrung ist die Motivation, den Job zu wechseln, augenscheinlich höher als bei Kandidaten mit längerer Erfahrung. Die erhobenen Daten bieten vielfältige Möglichkeiten der zielgenauen Ansprache von Kandidaten und versprechen hohe Erfolgsraten beim Anwerben neuer Mitarbeiter. Daher wird empfohlen, die Analyse der Kursteilnehmer-Daten und die daraus resultierenden Predictions zur Jobwechsel-Motivation in die Recruiting-Strategie des Unternehmens aufzunehmen. Auf diese Weise lassen sich Kosten für die externe Mitarbeitersuche einsparen und Stellen schneller besetzen. Zudem werden dadurch Kosten, die durch das Kursangebot entstehen, teilweise kompensiert. Und zuletzt werden Mitarbeiter geworben, die durch die Kursteilnahme bereits mit dem Unternehmen, deren Arbeitsweise und Technologien vertraut sind, was lange Einarbeitungszeiten spart.

Literaturverzeichnis und Quellen

Kranz, J.-D. (02. August 2022). it-talents.de. Abgerufen am 02. November 2022 von https://it-talents.de/it-news/der-it-arbeitsmarkt-2022-fakten-entwicklungen-hypothesen/

Paulsen, N., & Holdampf-Wendel, A. (03. Januar 2022). bitkom.org. Abgerufen am 04. November 2022 von https://www.bitkom.org/Presse/Presseinformation/IT-Fachkraefteluecke-wird-groesser

Schmole, M. (kein Datum). get-in-it.de. Abgerufen am 04. November 2022 von https://www.get-in-it.de/magazin/arbeitswelt/it-arbeitsmarkt/so-sieht-der-it-arbeitsmarkt-aus

stellenanzeigen.de. (07. Juli 2022). Abgerufen am 04. November 2022 von https://www.stellenanzeigen.de/arbeitgeber/wecruit/was-kostet-eine-neueinstellung/