Wirtschaftliche Untersuchungen,
Berichte und Sachverhalte

IW-Report 35/18

Identifikation von empirischen Unternehmenscharakteristika mittels Machine Learning Verfahren

Gemeinsames Projekt von DATAlovers, Institut der deutschen Wirtschaft und IW Consult
Manuel Fritsch, Dr. Henry Goecke, Andreas Kulpa

Köln, 24.09.2018

INSTITUT DER DEUTSCHEN WIRTSCHAFT

Inhaltsverzeichnis

1 Einleitung

Dieser Projektabschlussbericht fasst die Ergebnisse des gemeinsamen Projektes der DATAlovers AG, dem Institut der deutschen Wirtschaft und der IW Consult zusammen. Das Konsortium hat sich zusammengeschlossen um zu evaluieren, inwieweit Machine Learning Ansätze in Kombination mit den Inhalten von Unternehmensinternetseiten als primäre Informationsquelle für die wissenschaftliche Forschung angewendet werden können. Die grundlegende Aufgabenstellung des Projektes besteht darin zu klären, ob die Kombination aus den neuen Methoden des maschinellen Lernens und den Texten von Internetseiten bei der Identifikation von unternehmerischen Zwillingen wissenschaftlichen Ansprüchen genügt. Hierbei sollen validierte Informationen zu einer vergleichsweise kleinen Zahl an Unternehmen auf die Gesamtheit der Unternehmen übertragen werden. Bei einem Erfolg des Ansatzes würde dies bedeuten, dass mit einer kostengünstigen Methode Ergebnisse für alle deutschen Unternehmen gewonnen werden können. Durch diese „quasi-Vollerhebung“ würden sich viele weitere Anwendungsmöglichkeiten für ein Forschungsinstitut eröffnen.

Die Aufgabenteilung in diesem Projekt gestaltet sich wie folgt: Das Institut der deutschen Wirtschaft und die IW Consult liefern die originären Informationen der Unternehmen und übernehmen die Quantifizierung der Ergebnisse. DATAlovers bringt als originäre Daten die Texte der Internetseiten aller deutschen Unternehmen mit ein, trainiert mit der gesamten Datenmenge einen Algorithmus und bestimmt die Prognosen des Algorithmus.

Allgemein stellt die beschriebene Aufgabe ein Klassifizierungsproblem dar: Mit Hilfe einer großen Datenmenge soll für jedes deutsche Unternehmen entschieden werden, ob es zu einer spezifischen Gruppe gehört oder nicht. Für derartige Fragestellungen bietet sich die Verwendung von Machine Learning Methoden an. Da die Zielgrößen (die jeweiligen Gruppen) bekannt sind, ist die Klasse des überwachten maschinellen Lernens (im Gegensatz zum unüberwachten maschinellen Lernen) anzuwenden. Hierzu gehören beispielsweise die Methoden Logistic Regression, Random Forest, Support Vector Machine oder Decision Tree (vgl. ausführlicher dazu Brownlee, 2016; Provost/Fawcett, 2013).

Diese Ansätze werden vermehrt in der aktuellen Forschung und in der Statistik eingesetzt. Beispielsweise werden die Methoden von Feuerhake/Dumpert (2016) bei der Klassifizierung von Unternehmen in die deutsche Handwerksstatistik verwendet. Dumpert et al. (2016) verwenden diese Ansätze, um Unternehmen in den sogenannten Dritten Unternehmenssektor einzusortieren, bei Finke et al. (2017) erfolgt eine Zuordnung der Müttereigenschaft bei Frauen und Dumpert/Beck (2017) verwenden diese Methoden zur Klassifikation der Staatsangehörigkeit bei Personen. Des Weiteren werden aktuell Machine Learning Algorithmen verwendet, um Datensätze miteinander zu verknüpfen (z. B. Schild et. al., 2017). Damit lässt sich dieses Projekt, von der Methode her, den aktuellen Ansätzen in der amtlichen Statistik zuordnen.

2 Daten und Methoden

2.1 Vorgehensweise

Der neuartige Ansatz dieses Projektes ist die Kombination von zwei Elementen. Auf der einen Seite sind dies Daten aus einer Unternehmensdatenbank der IW Consult. Diese Unternehmensdatenbank deckt nicht die Gesamtheit der deutschen Unternehmen ab, sondern beinhaltet bei den in der Studie betrachteten Unternehmenscharakteristika zwischen 1075 und 1749 Unternehmen. Des Weiteren liegt der Fokus primär auf Industrieunternehmen und industrienahen Dienstleistern. Die in diesem Projekt verwendeten Unternehmensinformationen identifizieren, ob die jeweiligen Unternehmen auslandsaktiv sind, intensive Forschung und Entwicklung betreiben, eine hohe Exporttätigkeit aufweisen oder wie stark ihre Wertschöpfungsketten digitalisiert sind (für eine genauere Erläuterung siehe Kapitel 2.2). Für alle diese vier Charakteristika von Unternehmen wird im Zuge des Forschungsprojektes eine Zuordnung der Grundgesamtheit vorgenommen.

Auf der anderen Seite sammelt DATAlovers die gesamten Inhalte der Internetseiten aller in Deutschland ansässigen Unternehmen (entspricht rund 3 Millionen aktiven Unternehmen).
DATAlovers bekommt für die jeweilige Spezifizierung der Unternehmen aus der Unternehmensdatenbank eine Stichprobe als Trainingsdatensatz von mindestens 800 Unternehmen und trainiert mit diesen Informationen, unter Verwendung der Inhalte der entsprechenden Unternehmensinternetseiten und weiterer Informationen, ein Modell unter der Verwendung eines Machine Learning Algorithmus (für eine genauere Erläuterung der verwendeten Methode siehe Kapitel 2.4). Mit dem auf Basis des Trainingsdatensatzes trainierten Models wird allen anderen deutschen Unternehmen, die nicht im Trainingsdatensatz enthalten sind, die jeweilige Ausprägung der Eigenschaften (Auslandsaktiv Ja/Nein; Forschungsintensiv Ja/Nein; Exportstark Ja/Nein; Grad der Digitalisierung) zugewiesen. Abschließend wird auf Basis der zurückgehaltenen Unternehmensdaten der Unternehmensdatenbank überprüft, wie gut die Klassifikationsergebnisse des Machine Learning Algorithmus sind (siehe Kapitel hierzu 3.2).

Das Ziel der Analyse ist es zu bewerten, ob es mit Hilfe der Inhalte von Internetseiten unter Verwendung von Machine Learning Algorithmen möglich ist, spezifische Charakteristika allen Unternehmen zuzuordnen, ohne dass diese explizit auf einem anderen Wege ermittelt werden müssen. Erreicht diese Methode eine adäquate Güte, würde dies die Möglichkeit eröffnen, „quasi-Vollerhebungen“ von Unternehmen zu generieren.

2.2 Datenbank

Ein Datensatz beinhaltet den **Digitalisierungsgrad** von Unternehmen beziehungsweise deren Geschäftsmodellen. Diese digitalen Geschäftsmodelle wurden auf Grundlage eines Reifegradmodells definiert, welches durch die IW Consult im Rahmen des Projektes für den Zukunftsrat der bayerischen Wirtschaft entwickelt wurde (Lichtblau et al., 2017). Im Rahmen dieses Projektes wurden rund 2.500 Industrieunternehmen und industrienahe Dienstleister danach gefragt, wie stark ihr Geschäftsmodell digitalisiert ist und einen wie großen Anteil ihres Umsatzes sie mit digitalen Produkten erzielen. Das Reifegradmodell unterteilt anhand ihrer Antworten die Unternehmen in die folgenden vier Stufen:

- Stufe 0: Basis computerisiert
- Stufe 1: Unterstützend computerisiert
- Stufe 2: Gestaltend computerisiert
- Stufe 3: Digitalisiert

Die inhaltliche Unterscheidung zwischen den Stufen ist wie folgt:

- **Computerisiert (Stufen 0; 1 und 2)**: Dort werden die Unternehmen zugeordnet, die den Computer nur in Basisfunktionen, unterstützend oder aktiv gestaltend für ihre Geschäftsprozesse nutzen. Bei „unterstützender Computerisierung“ (Stufe 1) dienen Computer und IKT-Systeme eher zur Darstellung und als Informationsquelle. Digital aufbereitete Stammdaten liegen kaum vor. Bei der Stufe 0 gilt dies ebenfalls, jedoch noch in abgeschwächter Form beziehungsweise in weniger Bereichen als für die Stufe 1 nötig. „Gestaltende Computerisierung“ (Stufe 2) bedeutet einen systematischen Einsatz bei der Durchführung von Prozessen. Diese Gruppe ist nicht wirklich digitalisiert, weil ihr das entscheidende Merkmal einer „virtuellen Abbildung der Vernetzung der realen Welt“ fehlt. Daten spielen deshalb keine entscheidende Rolle in den Geschäftsmodellen.

- **Digitalisiert (Stufe 3):** Diese Unternehmen nutzen Daten, IKT-Technologien und das Internet zur „virtuellen Abbildung von Produkten und Prozessen“. Das ist die Voraussetzung für digitale Geschäftsmodelle, deren Rohstoffe Daten sind, die in verpackten Datenmodellen oder Algorithmen neue Wertschöpfung schaffen. Dafür ist die Bereitschaft zum „Teilen von Daten oder Informationen“ über Unternehmensgrenzen hinaus notwendig. In der maximalen Ausprägung können die Systeme „selbstständig und autonom entscheiden“ und sich sogar selbst optimieren. Hierbei tritt der Mensch als Entscheider in den Hintergrund. Dieser höchste Grad der Digitalisierung ist eng verwandt mit dem Konzept Industrie 4.0 von ACATECH und erfordert sehr elaborierte cyber-physische Systeme und den Einsatz künstlicher Intelligenz. Unternehmen mit diesem Reifegrad sind selten und werden auch in absehbarer Zukunft nur in ausgewählten Anwendungsbereichen zu finden sein.

In der verwendeten Unternehmensdatenbank sind für rund 1.100 Unternehmen zusätzlich Ergebnisse zu Indikatoren hinterlegt, die sich in mehreren Studien als Erfolgsfaktoren für deutsche Unternehmen erwiesen haben. Konkret handelt es dabei um dichotome (Erfolgsfaktor

liegt vor oder Erfolgsfaktor liegt nicht vor) Informationen zu **Exportaktivität**, **Auslandproduktionsstätten** und zu **Forschungsaufwendungen**.

2.3 Daten zum Trainieren des Modells unter Verwendung eines Machine Learning Algorithmus

Im Kern sind zum Training des Modells Daten aus den folgenden Bereichen eingeflossen:

Tabelle 2-1: Daten zum Training des Modells

Bereich	Beispieldaten	Datenmenge
Handelsregisterinformationen	Wirtschaftsschlüssel, Alter des Unternehmens	1 GB
Stammdaten	Ort, Name	600 MB
Strukturinformationen	Mitarbeiter-/Umsatzklassen	50 MB
Informationen aus dem DIGITAL INDEX		50 MB
Website und Social Media Content	Als Rohtextinformation	17 TB
Extrahierte Informationen aus der Website des Unternehmens	Verlinkungen, Bilder, Meta-Informationen	25 GB
Extrahierte Informationen aus den Social Media Profilen des Unternehmens	Follower, Likes, Posting-Frequenzen	2 GB

Quelle: Eigene Darstellung

Die Daten werden sowohl strukturiert als auch in unstrukturierter Form den Lernalgorithmen bereitgestellt.

2.4 Angewendete Methode zur Identifikation der Zwillingsunternehmen

Im Kern kam eine Best-Practice Kombination der Verarbeitung der Textinformationen durch ein neuronales Netz zur Vektorisierung der Textelemente sowie einer anschließenden Zuordnung mittels eines Random Forest Algorithmus zum Einsatz. Für die Vektorisierung der Volltexte wurde das von Google entwickelte Word2Vec Modell genutzt. Als Basisdaten für den Random Forest Algorithmus kamen die oben beschriebenen kategoriellen Daten (wie Branchen), numerische Daten (wie DIGITAL INDEX) und Merkmalsvektoren aus dem vorhergehenden Natural Language Processing (NLP) zum Einsatz.

Die Performance des Modells wurde mittels stratifizierter Kreuzvalidierung gemessen und optimiert. Hierbei wird in jedem der Kreuzvalidierungsschritte jeweils eine Teilmenge der Daten beiseitegelegt und ein auf Basis der restlichen Daten generiertes Modell klassifiziert. Das Modell, welches im Durchschnitt der Kreuzvalidierung am besten performt, wird anschließend zur Klassifikation aller Unternehmen genutzt.

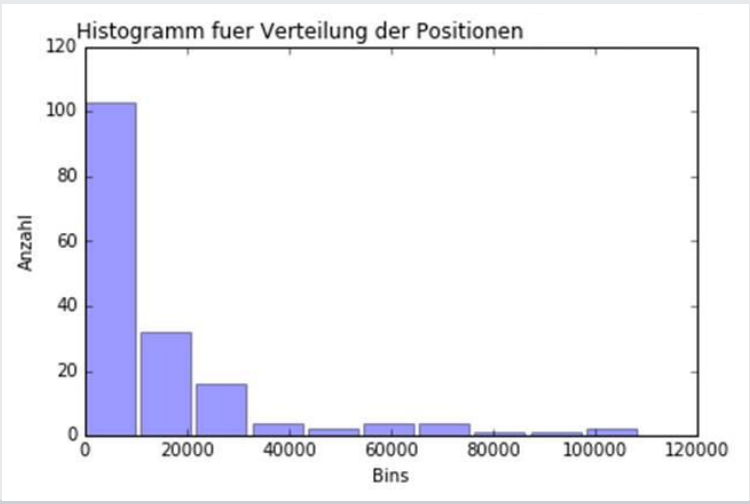
Alternative Machine Learning Verfahren abseits des Random Forest Modells wurden für diese Auswertung nicht überprüft. Der Random Forest Algorithmus zur Modellbestimmung hat sich in vergangenen Projekten bisher beständig als das Modell mit der stabilsten und besten Vorhersagegüte herausgestellt.

3 Ergebnisse

3.1 Deskriptive Statistik

Einen ersten Einblick in die Ergebnisse des Ansatzes liefert die Verteilung der prognostizierten Zuordnungen in der Grundgesamtheit. Nach einer zehnfachen Kreuzvalidierung ergeben sich die unten dargestellten Verteilungen. Hierbei zeigen die Histogramme und Verteilungen, in welche Wahrscheinlichkeitsgruppe der Algorithmus die einzelnen Unternehmen zugeordnet hat. Bei der Exportinformation finden sich gut 100 Unternehmen des Trainingsdatensatzes in der Gruppe der rund 10.000 Unternehmen mit der höchsten Wahrscheinlichkeit für eine starke Exportaktivität wieder (Tabelle 3-1). Insgesamt befinden sich 80 Prozent der zugeordneten Unternehmen unter den 5 Prozent der Grundgesamtheit, die nach dem Algorithmus die höchste Wahrscheinlichkeit haben, die Ausprägung „hohe Exportaktivität“ zu besitzen. Über 95 Prozent der zugeordneten Unternehmen befinden sich unter den Top 25 Prozent der Grundgesamtheit.

Tabelle 3-1: Deskriptive Statistik: Exportaktivität

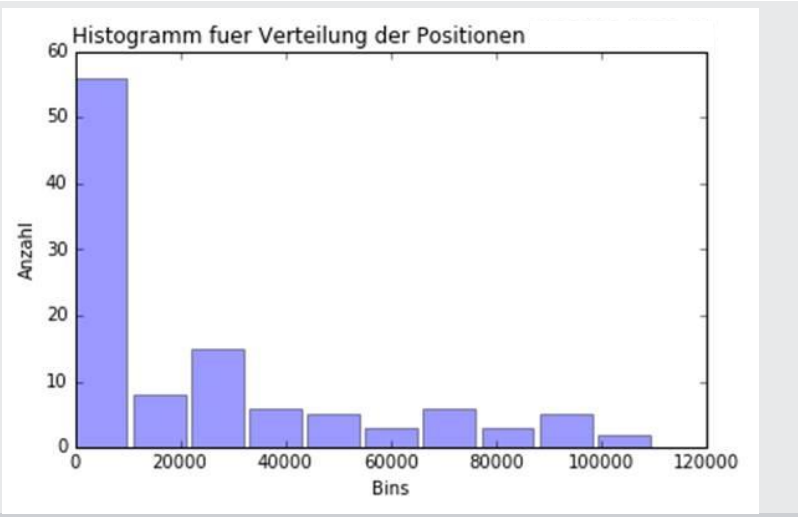


Unter Top …	Anteil in %
5 Prozent:	80.0
10 Prozent:	87.5
15 Prozent:	91.3
20 Prozent:	93.5
25 Prozent:	95.7

Quelle: Eigene Berechnungen

Bei der Auslandsaktivität und der Forschungs- und Entwicklungsaktivität sehen die Ergebnisse ähnlich aus, allerdings etwas schlechter als bei der Exportaktivität. Knapp 66 beziehungsweise knapp 63 Prozent der zugeordneten Unternehmen finden sich in den 5 Prozent aller Unternehmen wieder, bei denen die Wahrscheinlichkeit für eine starke Auslandsaktivität beziehungsweise Forschungs- und Entwicklungsaktivität am höchsten ist. Jeweils gut 90 Prozent der zugeordneten Unternehmen befinden sich unter den Top 25 Prozent der Grundgesamtheit.

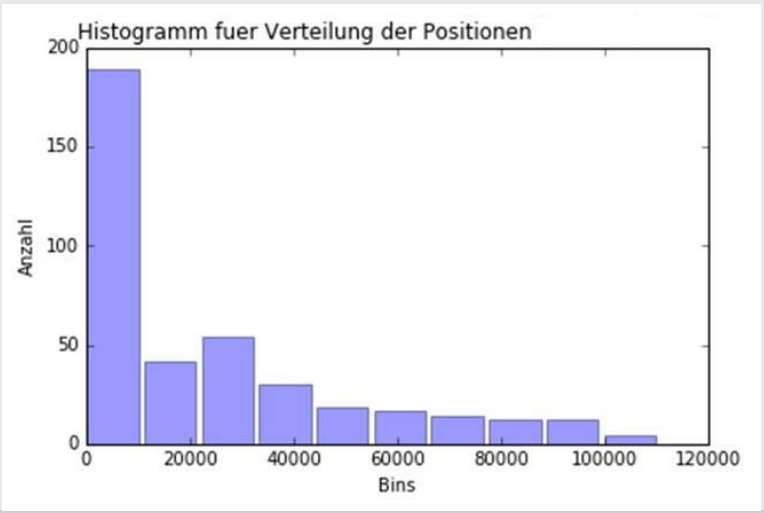
Tabelle 3-2: Deskriptive Statistik: Auslandsaktivität



Unter Top …	Anteil in %
5 Prozent:	65.8
10 Prozent:	75.8
15 Prozent:	85.0
20 Prozent:	90.8
25 Prozent:	94.2

Quelle: Eigene Berechnungen

Tabelle 3-3: Deskriptive Statistik: Forschungs- und Entwicklungsaktivität

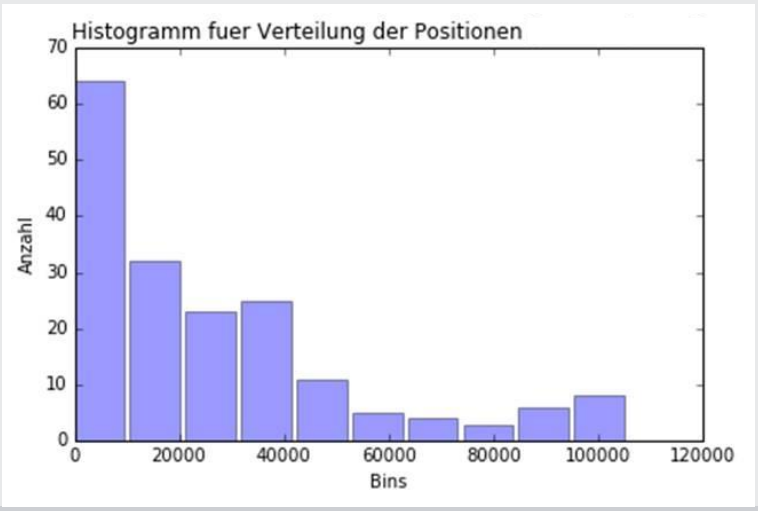


Unter Top …	Anteil in %
5 Prozent:	62.7
10 Prozent:	76.7
15 Prozent:	86.7
20 Prozent:	90.5
25 Prozent:	92.8

Quelle: Eigene Berechnungen

Eine weitere Verschlechterung ergibt sich bei den deskriptiven Ergebnissen zum Digitalisierungsgrad der Unternehmen. Prinzipiell ist diese Aufgabe jedoch auch schwieriger, da nicht nur zwischen der Zugehörigkeit oder der nicht Zugehörigkeit einer Gruppe vom Algorithmus unterschieden werden muss, sondern es vier separate Gruppen gibt. In der Tabelle 3-4 ist beispielhaft analog zu den vorherigen Tabellen die deskriptive Statistik für die Zuordnung der Unternehmen in die höchste Digitalisierungsstufe dargestellt. Gut 40 Prozent der der höchsten Digitalisierungsstufe zugeordneten Unternehmen befinden sich unter den 5 Prozent der Grundgesamtheit, die nach dem Algorithmus die höchste Wahrscheinlichkeit haben, die Ausprägung „höchste Digitalisierungsstufe“ zu besitzen.

Tabelle 3-4: Deskriptive Statistik: Digitalisierung von Unternehmen (höchste Stufe)



Unter Top …	Anteil in %
5 Prozent:	41.7
10 Prozent:	56.6
15 Prozent:	60.9
20 Prozent:	67.5
25 Prozent:	70.2

Quelle: Eigene Berechnungen

3.2 Validierung der Prognosegüte des Machine Learning Algorithmus

Die Generierung und Validierung der Güte der Prognose des Machine Learning Algorithmus bei der beschriebenen Aufgabe erfolgt in vier aufeinanderfolgenden Schritten:

1. Aufteilung der Daten aus der Unternehmensdatenbank in einen Trainings- und einen Kontrolldatensatz
2. Trainieren des Machine Learning Algorithmus auf die Inhalte der Unternehmensinternetseiten unter Verwendung des Trainingsdatensatzes
3. Zuordnung der jeweiligen Charakteristika für alle Unternehmen unter Verwendung des angelernten Machine Learning Algorithmus
4. Quantifizierung der Güte über den Abgleich der durch den Algorithmus zugeordneten Charakteristika für die Unternehmen aus dem Kontrolldatensatz

Im ersten Schritt werden die in der Unternehmensdatenbank vorhandenen Daten in einen Trainings- und einen Kontrolldatensatz geteilt. Der Kontrolldatensatz steht zum Trainieren des Algorithmus nicht zur Verfügung. Damit wird gewährleistet, dass die Quantifizierung der Güte des Machine Learning Algorithmus auf Daten basiert, die dieser nicht bereits kennt. Dieser Schritt ist nötig, um ein hartes Maß an die Güte der Prognosen zu legen und die Ergebnisse an unbekannten Daten zu testen (Christen, 2012, 34). Damit entgeht man zudem von vornherein der möglichen Kritik an einer mangelnden Unabhängigkeit der Daten für den Fall, dass der Algorithmus sehr gute Modellprognosen liefern sollte (Dumpert et al., 2016).

Im zweiten Schritt werden nur die Daten des Trainingsdatensatzes verwendet. Aus diesem Datensatz wird die Zuordnung der Unternehmen zu den jeweiligen Charakteristika genutzt, um den Machine Learning Algorithmus zu trainieren und allgemeingültige, übertragbare Muster für

die Grundgesamtheit zu identifizieren. Mit Hilfe dieser identifizierten Muster werden die Charakteristika allen Unternehmen zugeordnet (Schritt 3). Aus dieser per Algorithmus klassifizierten Grundgesamtheit werden die Unternehmen des Kontrolldatensatzes betrachtet. Abschließend werden für diese Unternehmen die zugeordneten Charakteristika aus dem Machine Learning Algorithmus mit den tatsächlichen Charakteristika aus der Unternehmensdatenbank verglichen und so die Güte des Machine Learning Algorithmus bestimmt.

Die Prognosegüte wird mit Hilfe von Konfusionsmatrizen ausgewertet. Tabelle 3-5 zeigt allgemein den Aufbau einer Konfusionsmatrix. Diese stellt bei einer Zuordnung zu lediglich einer Ausprägung die folgenden vier möglichen Fälle dar:

- *a*: Anzahl der Unternehmen, bei denen der Algorithmus eine Zuordnung zuweist (Modellprognose positiv), die auch in den Daten das Charakteristikum aufweisen (Daten positiv)
- *b*: Anzahl der Unternehmen, bei denen der Algorithmus eine Zuordnung zuweist (Modellprognose positiv), die in den Daten nicht das Charakteristikum aufweisen (Daten negativ)
- *c*: Anzahl der Unternehmen, bei denen der Algorithmus eine Zuordnung nicht zuweist (Modellprognose negativ), die jedoch in den Daten das Charakteristikum aufweisen (Daten positiv)
- *d*: Anzahl der Unternehmen, bei denen der Algorithmus eine Zuordnung nicht zuweist (Modellprognose negativ), die auch in den Daten das Charakteristikum nicht aufweisen (Daten nicht)

Die von dem Algorithmus korrekt vorhergesagten Fälle sind damit die Zellen *a* und *d*. Hieraus kann die jeweilige Modellgüte berechnet werden.

Tabelle 3-5: Konfusionsmatrix allgemein

		Daten			
		Positiv	Negativ		
Modell-prognose	Positiv	a	b	Richtig positiv:	a/(a+b)
	Negativ	c	d	Richtig negativ:	d/(c+d)
				Modellgüte = (a+d)/(a+b+c+d)	

Quelle: Eigene Berechnungen

3.3 Ergebnisse der Validierung

Aus allen beschriebenen Datensätzen ergibt sich jeweils eine Konfusionsmatrix, aus der die Güte der Prognose des Algorithmus abgelesen werden kann. Bei der Zuordnung zu lediglich einer Ausprägung (Exportaktivität; Auslandsaktivität; Forschungs- und Entwicklungsaktivität) lässt sich exakt die oben dargestellte Konfusionsmatrix bauen. Die Ergebnisse für die Exportaktivität zeigen, dass 186 Unternehmen des Kontrolldatensatzes vom Algorithmus als exportintensiv klassifiziert werden, die tatsächlich zu den exportintensiven Unternehmen gezählt werden können (Tabelle 3-6), während 23 Unternehmen fälschlicherweise dieser Gruppe zugeordnet werden. Hieraus ergibt sich eine Zuordnung „richtig positiv“ von 89 Prozent. Auf der anderen Seite werden 47 Unternehmen vom Algorithmus als nicht exportstark identifiziert, die tatsächlich zu den nicht exportstarken Unternehmen gehören, während 39 exportaktive Unternehmen falsch zugeordnet werden. Dies stellt eine Trefferquote von 55 Prozent dar. Insgesamt erzielt der Algorithmus bei der Zuordnung der Unternehmen im Hinblick auf deren Exportintensität eine Modellgüte von 79 Prozent.

Tabelle 3-6: Konfusionsmatrix: Exportaktivität

		Daten			
		Positiv	Negativ		
Modell-prognose	Positiv	186	23	Richtig positiv:	0,89
	Negativ	39	47	Richtig negativ:	0,55
				Modellgüte: 0,79	

Quelle: Eigene Berechnungen

Bei der Zuordnung zur Auslandsaktivität zeigen sich sehr ähnliche Ergebnisse (Tabelle 3-7). 217 Unternehmen des Kontrolldatensatzes werden vom Algorithmus korrekterweise den auslandsaktiven Unternehmen zugeordnet (Quote von 84 Prozent). 14 Unternehmen werden korrekterweise den nicht auslandsaktiven Unternehmen zugeordnet (Quote von 35 Prozent). Insgesamt erzielt der Algorithmus bei der Zuordnung der Unternehmen im Hinblick auf deren Auslandsaktivität eine Modellgüte von 78 Prozent.

Tabelle 3-7: Konfusionsmatrix: Auslandsaktivität

		Daten			
		Positiv	Negativ		
Modell-prognose	Positiv	217	41	Richtig positiv:	0,84
	Negativ	26	14	Richtig negativ:	0,35
				Modellgüte: 0,78	

Quelle: Eigene Berechnungen

Weniger gut ist der Algorithmus bei der Prognose der Forschungs- und Entwicklungsaktivität von Unternehmen (Tabelle 3-8). Von den Unternehmen aus dem Kontrolldatensatz werden zwar 103 korrekterweise als forschungs- und entwicklungsintensiv klassifiziert (Quote von 87 Prozent), allerdings werden nur 66 Unternehmen richtigerweise als nicht forschungs- und entwicklungsintensiv vom Algorithmus erkannt (Quote von 34 Prozent). Insgesamt hat der Algorithmus nur eine Güte von 54 Prozent und ist damit nur minimal besser als ein Münzwurf.

Tabelle 3-8: Konfusionsmatrix: Forschungs- und Entwicklungsaktivität

		Daten			
		Positiv	Negativ		
Modell-prognose	Positiv	103	16	Richtig positiv:	0,87
	Negativ	126	66	Richtig negativ:	0,34
				Modellgüte: 0,54	

Quelle: Eigene Berechnungen

Bei dem Datensatz zum Grad der Digitalisierung war den Unternehmen nicht wie bei den anderen Datensätzen ein binäres Charakteristikum zuzuordnen, sondern eine Einteilung in vier verschiedene Stufen vorzunehmen. Dies stellt für den Algorithmus eine wesentlich komplexere Aufgabe dar, als bei den bisher verwendeten Datensätzen. Der Referenzpunkt des Zufalls ist nun nicht mehr eine Trefferquote von 50 Prozent, sondern bei vier Stufen liegt dieser Wert bei 25 Prozent. Die Ergebnisse zeigen, dass der Algorithmus am besten bei der korrekten Zuordnung der Stufe 0 (geringste Computerisierung) und der Stufe 3 (Digitalisiert) ist (57 respektive 43 Prozent Trefferquote). Die korrekte Zuordnung zu der Stufe 1 und der Stufe 2 ist wesentlich geringer (29 respektive 27 Prozent Trefferquote). Damit ergibt sich insgesamt eine Prognosegüte, die mit 36 Prozent zwar besser als der Zufall, jedoch weit von einem sehr guten Ergebnis entfernt ist. Inwiefern das Vorgehen bei Variablen mit mehr als 2 Ausprägungen angepasst oder bewertet werden muss, sollte durch die Durchführung zukünftiger Projekte genauer betrachtet werden.

Tabelle 3-9: Konfusionsmatrix: Digitalisierungsgrad

		Daten				
		Stufe 0	Stufe 1	Stufe 2	Stufe 3	Richtig positiv:
Modell-prognose	Stufe 0	40	14	7	9	0,57
	Stufe 1	52	42	24	28	0,29
	Stufe 2	24	21	30	36	0,27
	Stufe 3	15	14	26	41	0,43
						Modellgüte: 0,36

Quelle: Eigene Berechnungen

3.4 Einordung der Ergebnisse

Bleibt die Frage nach der Bewertung der gemessenen Güte. Aufgrund der Anwendung des Projektes auf deutsche Unternehmen erfolgt die Einordung der Ergebnisse im Vergleich zu Anwendungen von maschinellem Lernen in der deutschen Statistik.

Bei der Klassifizierung von Unternehmen in der deutschen Handwerksstatistik sprechen Feuerhake/Dumpert, 2016) bei einer Fehlklassifizierungsrate von knapp 6 Prozent (einer Modellgüte von 94 Prozent) von einem „akzeptablen Wert“. Bei einer Zuordnung von Unternehmen in den sogenannten Dritten Unternehmenssektor kommen Dumpert et. al. (2016) auf eine Modellgüte von etwa 86 Prozent und ziehen das Fazit, dass dieses Ergebnis in der üblichen Größenordnung der Modellgüte von 64 bis 97 Prozent liegt. Finke et al. (2017) erzielen bei der Zuordnung der Müttereigenschaft eine Modellgüte von etwa 74 Prozent und verwenden diese Ergebnisse, um anschließend weitere Berechnungen beim gender pay gap durchzuführen. Die Klassifikation der Staatsangehörigkeit resultiert nach Dumpert/Beck (2017) in einer Modellgüte von 82 Prozent. In Schild et al. (2017) werden mit Machine Learning Algorithmen Datensätze miteinander verknüpft. Dort ergibt sich eine Modellgüte von knapp 96 Prozent.

Mit jeweils knapp 80 Prozent Modellgüte in den Variablen Exportaktivität und Auslandsaktivität befinden sich die Ergebnisse dieses Projektes in der Größenordnung, die in der Literatur Akzeptanz finden. Nichtsdestotrotz besteht noch wesentliches Verbesserungspotenzial, da die Präzision der Algorithmen bei anderen Anwendungsfällen oft über 90 Prozent liegt.

Die Zuordnung einer starken Forschungs- und Entwicklungsaktivität stellt sich mit einer Modellgüte von 54 Prozent nur unwesentlich besser als der Zufall dar. Herauszufinden, worin dieser starke Abfall der Performance begründet ist, wäre eine Aufgabe für zukünftige Forschung. Eine thesenartige Begründung wäre, dass Begriffe mit Bezug zu Aktivitäten mit dem oder im Ausland (beispielsweise über ausländische Städtenamen) vom Algorithmus klarer als Muster identifizierbar sind, als Begriffe, die eine starke Forschungs- und Entwicklungsaktivität implizieren können.

Bleiben abschließend die Ergebnisse der Zuordnung des Digitalisierungsgrades. Hierbei liegt die Modellgüte des Algorithmus bei 36 Prozent. Dies ist mit Abstand der niedrigste Wert, jedoch war bei dieser Aufgabe auch eine Zuordnung in vier ordinal skalierte Klassen und nicht nur in zwei Gruppen zu vollziehen. Die Güte ist zwar signifikant besser als eine Zufallszuordnung, liegt aber weit unterhalb der für die Wissenschaft nötigen Werte. Im Detail zeigt sich, dass insbesondere die Einordnung in die Subgruppe 1 und 2 dem Algorithmus Schwierigkeiten bereitet. Deren Digitalisierungsgrad ergibt sich durch die Addition der Ausprägung verschiedener Digitalisierungsmerkmale. Die Zuordnung zu den extremen Gruppen Gruppe 0 und Gruppe 3 gelingt hingegen etwas besser. Es steht zu überprüfen, ob eine Zuordnung von verschiedenen über Grenzwerte gebildeten Klassen durch einen Machine Learning Algorithmus grundsätzlich sinnvoll ist, oder hier durch die potentiell hohe Menge an betrachteten Fällen im Grenzbereich für die Einzelinterpretation der Ergebnisse ein solches Verfahren für wissenschaftliche Untersuchungen weniger geeignet ist.

4 Fazit und Ableitungen

Insgesamt reichen die Ergebnisse dieses Projektes von ernüchternd bis ermutigend: Die Zuordnung über Machine Learning Algorithmen unter Verwendung der Texte von Internetseiten scheint für komplexe und vielstufige Zuordnungen wie dem Digitalisierungsgrad (noch) nicht geeignet. Einfachere Zuordnungen in nur zwei unterschiedliche Klassen gelingen hingegen teilweise schon recht gut. An dieser Stelle würde es sich lohnen, in weitere Arbeit zu investieren, um die Güte der Modelle respektive der Algorithmen weiter zu verbessern.

Maschinelles Lernen stellt eine vielversprechende Möglichkeit zur Beantwortung verschiedener Fragestellungen auf Basis unstrukturierter Daten dar. Die hohen Aktivitäten dieser Big Data Anwendung auch in den Bereichen Wirtschaftswissenschaften, Statistik und Forschung implizieren, dass hier große Potentiale bestehen. Gleichwohl schwankt die Güte der Ergebnisse je nach Fragestellung, Datenumfang und Qualität sowie der verwendeten Methode.

Allgemein hat sich der Ansatz bewährt, bei der Bearbeitung einer Fragestellung mittels Machine Learning die Ergebnisse mit einem Kontrolldatensatz bewerten zu können. Da die Generierung solcher Datensätze potentiell mit einem hohen Ressourcenaufwand verbunden ist, sollte ex ante eine möglichst gute Einschätzung gewonnen werden, bei welchen Fragestellungen und Datensätzen solche Analysen eine hohe Erfolgswahrscheinlichkeit besitzen und bei welchen weniger. Es sollte vorab möglichst effektiv abgeschätzt werden, welche Projekte sich mit Hilfe dieser Methodik bearbeiten lassen und realistischer Weise zu einem zufriedenstellenden Ergebnis führen könnten. Das Verfahren ist für kommende Aufgabenstellungen weiterhin als Option mit hohem Potential zur Beantwortung spezifischer Fragestellungen zu bewerten, unterliegt jedoch in der Anwendbarkeit verschiedenen Restriktionen.

Tabellenverzeichnis

Literaturverzeichnis

Brownlee, Jason (2016), Master Machine Learning Algorithms. Discover How They Work and Implement Them From Scratch, o. O.

Christen, Peter (2012), Data Matching. Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection, Berlin

Dumpert, Florian / Beck, Martin (2017), Einsatz von Machine-Learning-Verfahren in amtlichen Unternehmensstatistiken, in: AStA Wirtschafts- und Sozialstatistisches Archiv, Bd. 11, Nr. 2, S. 83–106

Dumpert, Florian / von Eschwege Katja / Beck, Martin (2016), Einsatz von Support Vector Machines bei der Sektorzuordnung von Unternehmen, in: WISTA Wirtschaft und Statistik, 2016, Nr. 1, S. 87–97

Feuerhake, Jörg / Dumpert, Florian (2016), Erkennung nicht relevanter Unternehmen in den Handwerksstatistiken, in: WISTA Wirtschaft und Statistik, 2016, Nr. 2, S. 79–94

Finke, Claudia / Dumpert, Florian / Beck, Martin (2017), Verdienstunterschiede zwischen Männern und Frauen. Eine Ursachenanalyse auf Grundlage der Verdienststrukturerhebung 2014, in: WISTA Wirtschaft und Statistik, 2016, Nr. 2, S. 43–62

Lichtblau, Karl et al. (2017), Ökonomische Aspekte der Digitalisierung, in: vbw – Vereinigung der Bayerischen Wirtschaft e.V. (Hrsg.), Neue Wertschöpfung durch Digitalisierung, München, S. 45–117.

Provost, Foster / Fawcett, Tom (2013), Data Science for Business. What You Need to Know About Data Mining and Data-Analytic Thinking, Sebastopol, CA

Schild, Christopher-Johannes / Schultz, Simone / Wieser, Franco (2017), Linking Deutsche Bundesbank Company Data using Machine-Learning-Based Classification. Technical Report 2017-1, o. O.