Predictive Analytics Organisationsstrukturen

Serdar Süzen

areto consulting gmbh

Köln

Schlüsselworte
Predictive Analytics, Data Scientist, Klassifikation, Regression, Clustering, Assoziation, Organisationsstrukturen

Einleitung
Mit fortschreitender Technik eröffnen sich im Themenfeld Business Intelligence auch neue Möglichkeiten. Während die in die Vergangenheit gerichteten Reports und Ad-Hoc Anfragen sich mittlerweile Etabliert haben, ermöglicht Predictive Analytics (PA) erstmalig fundierte Prognosen. Diese Prognosen stellen eine neue Qualität der Datenanalyse dar und bieten Unternehmen einen hohen Mehrwert, da sie anstatt reaktionär nun proaktiv handeln können. Dadurch können z.B. Kaufverhalten von Kunden vorhergesagt, Wartungsintervalle für Produktionsanlagen abgeschätzt oder Betrugsfälle aufgedeckt werden. Während PA in den USA bereits seit einigen Jahren für Unternehmen eine wichtige Rolle spielt, wird auch in Deutschland die Relevanz des Themas langsam wahrgenommen.

PA Methoden
Obwohl PA für viele Unternehmen neu ist, handelt es sich um etablierte mathematische und statische Methoden und Algorithmen. Grundsätzlich können die Algorithmen in das Clustering, die Klassifikation, die Regression und die Assoziationsanalyse aufgeteilt werden.

Das Clustering hat das Ziel eine Datenmenge auf Basis vorher festgelegter Variablen zu gruppieren. Dabei werden Objekte die sich „ähnlich“ sind der gleichen Gruppe zugeordnet. Durch das Clustering wird den Daten eine zusätzliche Variable, die Gruppenzugehörigkeit, annotiert. Die fachliche Interpretierbarkeit der Gruppenzugehörigkeit steht dabei nicht im Fokus. Die Ähnlichkeit zwischen Objekten wird anhand eines Distanz- bzw. Ähnlichkeitsmaßes bestimmt. Die Clusteringalgorithmen können ihrer Arbeitsweise nach z.B. in partitionierende (K-Means, EM), hierarchische (agglomerativ, divisiv) oder dichtebasierte Algorithmen und bei ihrer Gruppenzuordnung nach weichen und harten Algorithmen unterschieden werden. Weil es beim Clustering keine Zielvariable gibt, anhand derer entschieden werden kann, ob eine Gruppenzugehörigkeit richtig oder falsch ist, sondern nur wie gut oder schlecht die Aufteilung auf Basis eines Gütemaßes ist, gehört das Clustering zu den Methoden des unüberwachten Lernens.

Im Gegensatz dazu gehören Klassifikationsalgorithmen zu den Methoden des überwachten Lernens. Bei der Klassifikation gibt es eine Zielvariable, welches die Klassenzugehörigkeit, wie z.B. Kreditwürdigkeit oder Betrugsfall, beschreibt, die anhand weiterer beschreibender Variablen vorhergesagt werden sollen. Dabei kann es sich um binäre oder n-äre Ausprägung der Klassen handeln. Um den Klassifikationsalgorithmus auf einen konkreten Kontext anzupassen und die Qualität der Ergebnisse zu überprüfen, werden zwei unterschiedliche Datenmengen benötigt. Die Trainingsmenge beinhaltet die Daten mit den dazu gehörigen Klasseneigenschaften, z.B. Kunden bei denen ein Betrugsverdacht bestätigt wurde. Auf Basis dieser Trainingsmenge lernt der Algorithmus die Muster in den Daten und optimiert die Vorhersage der Klassenzugehörigkeit. Mithilfe einer

Testmenge, die während der Lernphase des Algorithmus zurückgehalten wurde, wird anschließend überprüft, wie gut der Algorithmus auf einer neuen, noch unbekannten Datenmenge arbeitet.

Die Regressionsalgorithmen werden für die Vorhersage genau einer Zielvariable (abhänige Variable) auf Basis einer oder mehrerer beschreibender Variablen (unabhängige Variablen) verwendet. Wie bei Klassifikationsalgorithmen auch, wird bei Regressionsverfahren eine Trainings- und Testmenge mit einer zu vorherzusagenden abhängigen Variable benötigt. Die Regressionsalgorithmen liefern Funktionen, in der die unabhängigen Variablen übergeben werden, um die abhängige Variable vorherzusagen. Die Funktion kann dabei linear oder nichtlinear sein. In diesem Kontext stellt die Zeitreihenanalyse eine spezialform der Regression dar, welche die zeitliche Dimension mit in die Prognose aufnimmt.

Die Assoziationsanalyse hat das Ziel Regelmäßigkeiten in einer Datenmenge zu finden und Regeln abzuleiten und diese auf eine relevante Menge, auf Basis von unterschiedlichen Filtermethoden, zu reduzieren. Ein typischer Anwendungsfall ist dabei die Warenkorbanalyse, bei der die Einkäufe Analysiert werden um z.B. häufige Kombinationen von Produkten zu erkennen. Die Filtermethoden können dabei helfen, um z.B. offensichtliche Regeln wie „Kunden die Cornflakes kaufen, kaufen auch oft Milch“ auszublenden und sich auf ggf. weniger offensichtliche Regeln wie „Kunden die Taschenlampen kaufen, kaufen auch oft Fertigsuppen“ zu konzentrieren und die Abhängigkeit zu untersuchen.

Bei Methoden die trainiert werden müssen besteht die Gefahr des Overfitting. Das Overfitting beschreibt das ein Modell zu sehr an die Daten mit denen es trainiert wurde angepasst ist, sodass das Modell auch das Rauschen in den Daten abbildet, anstelle der Essenz des Zusammenhangs. Dadurch wird zwar meinst eine gute Prognosegenauigkeit für die Datensätze mit denen trainiert wurde erreicht, jedoch leidet darunter meist die Prognosegenauigkeit für zukünftige neue Daten. Was bei einem Modell mit zwei Einflussgrößen leicht erkennbar ist, kann bei hochdimensionalen Daten nur schwer erkannt werden.

Data Scientist

Die unterschiedlichen Methoden gehören unter anderem dabei zum Repertoire des Data Scientist, denjenigen der zwischen Technik/Wissenschaft und Fachlichkeit/Business steht und die Prognosen erstellt. Die Rolle des Data Scientist ist jedoch relativ neu und wird oftmals sehr unterschiedlich definiert. Es stellt sich jedoch heraus, dass der Data Scientist eine breite Palette an Anforderungen erfüllen und unterschiedliche Sichten und Rollen einnehmen muss. Der Data Scientist benötigt eine enge Einbindung bzw. Beziehung zum Business, um zum einen das Geschäftsumfeld mit seinen Prozessen und Beziehungen zu verinnerlichen und zum anderen die analytischen Erkenntnisse und Prognosen dem Business zu verdeutlichen und Maßnahmen und weitere Fragestellungen zu planen. Um dies zu leisten muss der Data Scientist sich in der mathematisch algorithmischen Welt genauso gut auskennen wie mit dem Geschäftsumfeld. Fachliches Know-How, verständliche und prägnante Datenvisualisierung und Kommunikation sind dabei unabdingbar. Daten spielen bei den Prognosen eine zentrale Rollen, sodass der Data Scientist dafür sorgen muss die Daten in optimaler Form bereit stehen. Neben den Prozessen der Datenbereitstellung und Transformation (ETL), muss die Datenqualität aufbereitet werden können, sodass z.B. Ausreißer oder Messfehler entsprechend behandelt, Daten angereichert und in einer auswertbaren Architektur bereitgestellt werden können. Um neue Erkenntnisse zu entdecken, wird zudem ein gewisses Maß an Kreativität vorausgesetzt um Prognose-Prozesse zu erstellen.

Organisationsstrukturen
Um das Thema PA und somit die Data Scientist in das Unternehmen einzuführen, können unterschiedliche Organisationsstrukturen verwendet werden. Die richtige Auswahl der Organisationsstruktur kann dabei im Vornhinein maßgeblich über den Erfolg des Themas PA im Unternehmen entscheiden.

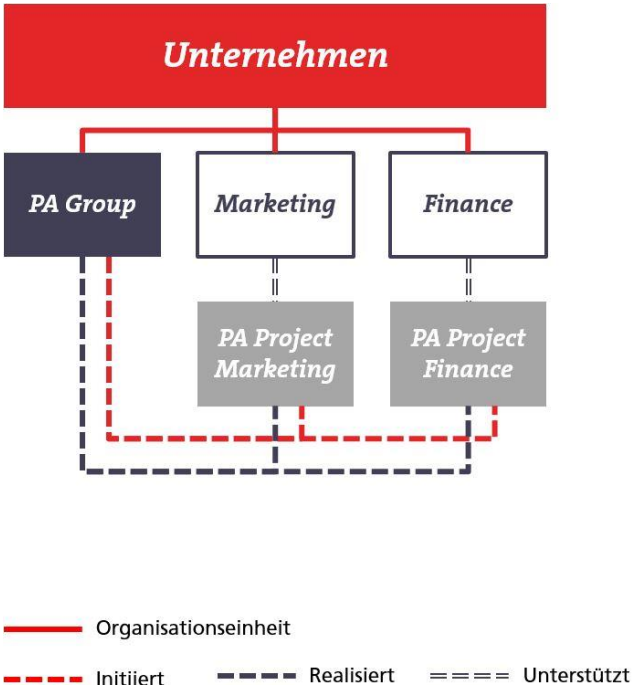

Abb. 1: zentralisierte Organisationsstruktur

Eine zentralisierte Organisationsstruktur, in der das PA Gruppe eine eigene Organisationseinheit bildet, initiiert und realisiert Projekte für andere Abteilungen. Die PA Gruppe legt dabei die unternehmensweite PA-Strategie fest und kann sicherstellen, da die Data Scientist alle der PA Gruppe unterstellt sind, das diese auch verfolgt werden. Da die Data Scientist nicht einer festen Domäne im Unternehmen zugeordnet sind, wird eine fachliche Unterstützung des Fachbereichs benötigt, um zum einen Use Cases zu identifizieren und zum anderen die Daten, Prozesse und Abhängigkeiten zu verstehen.

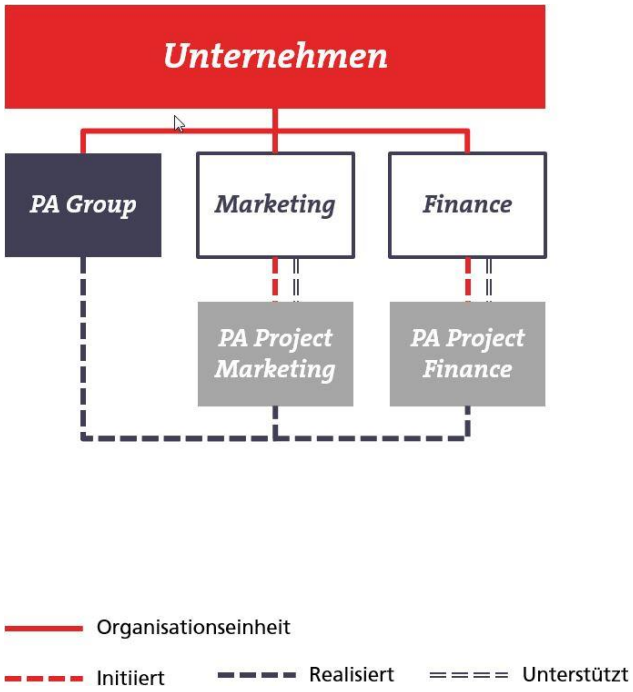

Abb. 2: beratende Organisationsstruktur

Bei einer beratenden Organisationsstruktur existiert eine eigene PA Organisationseinheit, welches auch die strategische Ausrichtung bezüglich PA festlegen kann. Die Data Scientist sind dabei der PA Organisationseinheit unterstellt. Die Projekte werden von unterschiedlichen Abteilungen initiiert und zusammen mit den Data Scientists der PA Organisationseinheit und der Fachlichen Unterstützung der Abteilung durchgeführt. Die für das Projekt gebuchten Data Scientists können dabei nicht nur beratend, sondern auch entwickelnd das Projekt unterstützen.

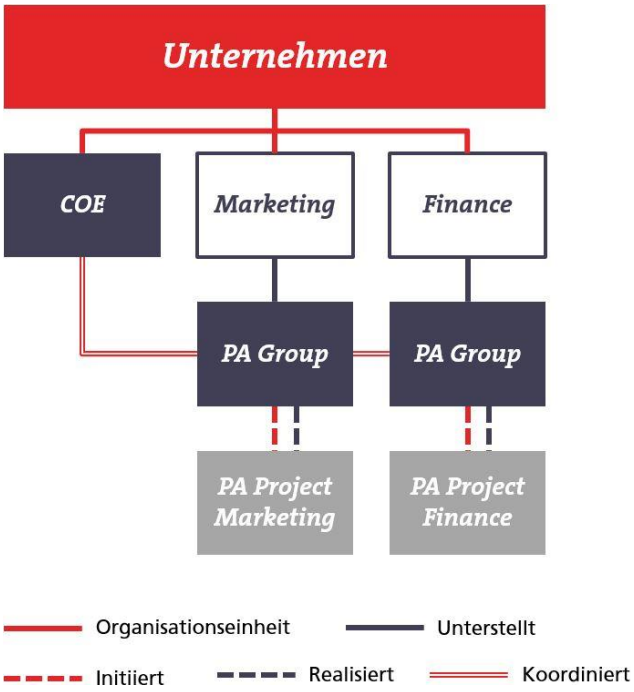

Abb. 3: Center of Excellence Organisationsstruktur

Das Center of Excellence (COE) beschreibt eine Organisationsstruktur den Organisationseinheiten eigene PA Gruppen unterstellt sind, die durch ein COE koordiniert werden. Das COE legt die unternehmensweite PA Strategie fest und stellt sicher, dass die jeweiligen PA Gruppen sich danach richten. Die COE bildet zudem eine Community und sorgt für regen Erfahrungs- und Wissensaustausch. Die Data Scientists in den Gruppen stehen dabei in engem Kontakt zur Abteilung und kennen sich dementsprechend in der Domäne fachlich aus, weshalb sie potenziale erkennen und die Abteilungen durch sinnvolle Prognosen unterstützen können. Aus diesem Grund können die PA Gruppen PA Projekte selbst initiieren.

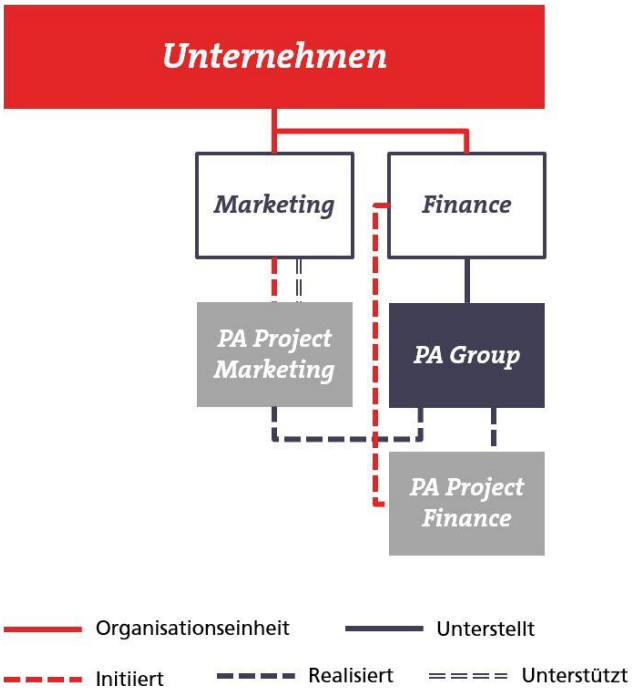

Abb. 4: funktional zentrierte Organisationsstruktur

Bei der funktional zentrierten Organisationsstruktur wird die PA-Gruppe der Organisationseinheit untergeordnet, die sich hauptsächlich mit dem Thema befasst. Die Organisationseinheit gibt dabei die PA-Strategie vor und initiiert die PA-Projekte. Da die Data Scientist sich hauptsächlich in der Domäne einer Organisationseinheit befinden, wird das fachliche Know How über Daten, Prozesse usw. angehäuft, sodass die PA Gruppe eigenständig, ohne fachliche Unterstützung durch die übergeordnete Organisationseinheit, Projekte durchführen kann. Anders sieht es aus, wenn die PA-Gruppe beauftragt wird, ein Projekt für eine andere Organisationseinheit bzw. Abteilung durchzuführen. Da sich die Data Scientist in anderen Domänen nicht auskennen würden, wäre die fachliche Unterstützung durch die beauftragende Organisationseinheit bzw. Abteilung und eine enge Kommunikation, erfolgsentscheidend.

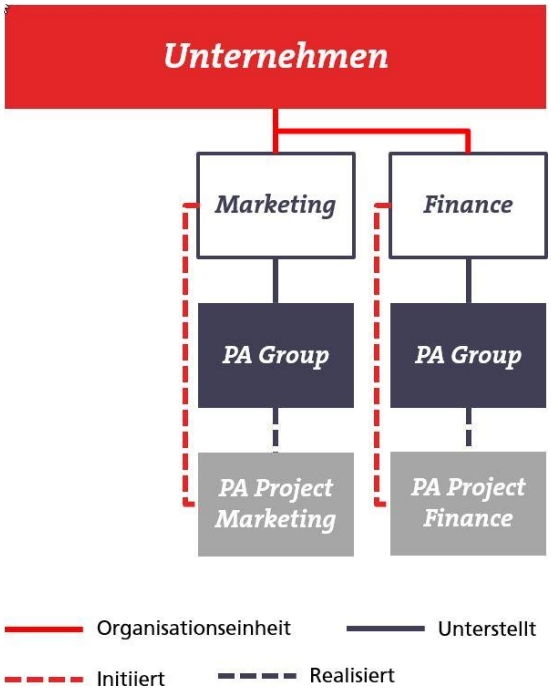

Abb. 5: dezentrale Organisationsstruktur

Bei der vollständig dezentralisierten Organisationsstruktur werden mehrere unabhängige und geschlossene PA-Gruppen ins Leben gerufen, die jeweils unterschiedlichen Organisationseinheiten unterstellt sind. Jede Organisationseinheit bzw. Abteilung definiert dabei eine individuelle, an seine Domäne angepasste Strategie und entsprechende Ziele. Die PA-Gruppe führt nur Projekte für die Ihnen übergeordnete Organisationseinheit bzw. Abteilung durch. Aufgrund der Spezialisierung können Projekte ohne fachliche Unterstützung durchgeführt werden.

Den Unternehmen stehen bei der organisatorischen Integration des Themas unterschiedliche Möglichkeiten zur Auswahl. Eine klare Struktur hat sich jedoch, aufgrund weniger Erfahrungswerte nicht eindeutig durchgesetzt. Vielmehr muss dabei der Reifegrad des Themas, die Teamgröße, verfügbare Ressourcen, strategische Ausrichtung usw. bewertet werden, um eine passende Struktur zu wählen. Während zentralere Strukturen z.B. eine Gesamtstrategie für das Unternehmen definieren und Erfahrungs- und Wissensaustausch fördern, haben dezentralere Strukturen den Vorteil geringeren Overhead zu verursachen, Spezialisierung in einer Domäne zu ermöglichen und abteilungsspezifische Strategien und Ziele zu verfolgen.

Kontaktadresse:
Serdar Süzen
areto consulting gmbh
Schanzenstr. 6-20
51063 Köln

Telefon: 0221 6695750
Fax: +49 (0) 12-345 6788
E-Mail serdar.suezen@areto-consulting.de
Internet: www.areto-consulting.de