Von der Ideenfindung zum erfolgreichen PA-Projekt

Viele Unternehmen tun sich schwer, erste Projekte im Umfeld von Predictive Analytics (PA) in Angriff zu nehmen und zu erproben. Zumeist fehlt ebenso erfahrenes Personal wie auch ein geeignetes Umsetzungsmodell. Bei der Realisierung von Projekten kann auf ein bewährtes Rahmenwerk zurückgegriffen werden, das bereits in den Jahren 1996 bis 1999 im Rahmen eines EU-Projektes entstanden ist: Die Unternehmen SPSS, Teradata, Daimler und OHRA entwickelten den Standard CRISP-DM (Cross Industry Standard Process for Data Mining) der sich auch für die Durchführung von PA-Projekten bestens eignet.

Das Rahmenwerk beschreibt den gesamten Ablauf in sechs Phasen, die mit ihren jeweiligen Aufgaben und Ereignissen sowie zu beachtenden Kriterien nachfolgend beschrieben werden:

- Business Understanding
- Data Understanding
- Data Preparation
- Modelling
- Evaluation
- Deployment

siehe Abbildung 1

Business Understanding hat insbesondere die Beschreibung der Projektziele und Anforderungen zum Inhalt und soll eine detaillierte Vorgabe der strategischen und betriebswirtschaftlichen Ziele in den Kontext von PA erzeugen. Dabei ist es zwingend erforderlich, die Aufgabenstellung so konkret wie möglich zu formulieren. Ein Beispiel: Der Absatz in einem definierten Segment soll um 10 Prozent gesteigert werden. Dabei sollen die Profile und Kaufgewohnheiten der Konsumenten in den vergangenen 12 Monaten untersucht werden, um Rückschlüsse zu erhalten, mit welcher Zielansprache und welchen (Zusatz-) Produkten eine solche Umsatzsteigerung erreicht werden kann.

Diese Phase ist extrem wichtig für den erfolgreichen Verlauf des gesamten Projektes und des möglichen Geschäftserfolges: Fehler, die hier begangen werden, wirken sich auf alle nachfolgenden Schritte aus. Im schlimmsten Falle liefern sie dann „richtige Ergebnisse auf nicht gestellte Fragen".

Business Understanding kann in folgende Teilprozesse oder –aufgaben untergliedert werden:

- Bestimmung der Geschäftsziele mit Zieldefinitionen aus Sicht von Management und Fachbereichen, Projektplan und Festlegen der Erfolgskriterien und Kennzahlen
- Betrachtung des Geschäftsumfeldes mit den Ressourcen (Personal, Daten, Hardware, Software, etc.) und den Rahmenbedingungen inklusive der Schaffung eines Risikoplanes sowie eines Glossars zum einheitlichen Verständnis wichtiger Begriffe und Kennzahlen
- Bestimmung der Ziele des PA-Einsatzes und Festlegung der Testszenarien
- Festlegung der einzusetzenden Methoden und Tools und Prüfung auf ihre Eignung im Kontext des Projektes

Innerhalb von **Data Understanding** erfolgt die Sammlung und Bewertung der zur Verfügung stehenden Daten vor allem in Hinblick auf die Datenqualität. In diesem Abschnitt soll geprüft werden, ob die Daten in ausreichender Weise gepflegt und aktualisiert wurden, so dass sie für die Analysen einsetzbar sind und nicht von vornherein verfälschende Ergebnisse vorprogrammiert sind.

- Sammlung der Daten und Prüfung, ob sie in Bezug auf Menge, Aktualität und Qualität für das angestrebte Ziel ausreichend sind und problemlos integriert werden können
- Beschreibung der Daten hinsichtlich Datentyp, Felder, Attribute, Formate und weiterer Merkmale sowie optionale Beschaffung weiterer (externer) Daten
- Durchführung der explorativen Datenanalyse, wobei ein erster

Von der Ideenfindung zum erfolgreichen PA-Projekt

Überblick der vorliegenden Daten erreicht wird. Dabei wird insbesondere nach Abhängigkeiten, Auffälligkeiten, Verteilungen, Ausreißern oder fehlenden Daten gesucht.

- Die Daten werden mit Hilfe von Histogrammen, Box-Plots oder anderen Darstellungsformen visualisiert

Besonders wichtig: Datenqualität

Datenqualität ist eine grundlegende Voraussetzung für ein erfolgreiches PA-Projekt und kann zu keinem Zeitpunkt kompensiert werden. Aus diesem Grunde sind mögliche Unzulänglichkeiten bereits im Vorfeld zu bestimmen und zu eliminieren. Messfehler oder Ausreißer haben im Kontext von PA eine völlig andere Bedeutung und einen anderen Stellenwert als bei klassischen Data Warehouse Anwendungen.
Die in dieser Phase gewonnenen Erkenntnisse über die Datenqualität können eventuell Rückkopplungen zur ersten Phase Business Understanding erzeugen und eine Anpassung der anfänglichen Ziele und Fragestellungen erfordern. Mitunter mündet dieser Prozess auch in der Schlussfolgerung, dass die vorhandenen Daten für das zu erzielende Ergebnis völlig ungeeignet sind.
Im Mittelpunkt der Phase Data Preparation steht die Definition eines finalen Algorithmus für die spätere Modellierung. Spätestens zu diesem Zeitpunkt müssen Ausreißer, Messefehler oder fehlende Variablen sowie andere Unzulänglichkeiten ermittelt und bereinigt worden sein. Die mit Hilfe des eingesetzten Algorithmus gewonnenen Datensätze werden nach zuvor festgelegten Kriterien gefiltert, aggregiert und transformiert. Gegebenenfalls müssen die Daten noch durch weitere Informationen angereichert werden, um weitere Aspekte in die Analyse einfließen zu lassen.
Die Phase **Modelling** beschreibt die Phase der Auswahl einer geeigneten PA-Methode zur Untersuchung der Aufgabenstellung. Auch hier sind verschiedene Teilprozesse zu durchlaufen bis ein Modell entsteht, das zunächst mit Test- und Trainingsdaten und anschließend mit Daten aus den produktiven Systemen gespeist wird.

- Entwickeln eines Testdesigns mit Definition der Qualitätskennzahlen und Aufteilungen der Daten in eine geeignete Test- und Trainingsmenge
- Auswahl und Erprobung der Algorithmen und deren Dokumentation
- Erstellung des Modells und einer Dokumentation mit Begründung der Auswahl für diesen Typus

Schlussendlich erfolgt der Test der Modelle, der unter Umständen die Neuauswahl eines besser geeigneten Algorithmus erfordern oder auch zu Änderungen der Teilschritte aus den Vorphasen führen kann.
Im Prozessschritt **Evaluation** werden die Resultate der Datenanalyse betrachtet und mit den im Business Understanding formulierten Erfolgskriterien abgeglichen. Vorrangiges Ziel ist eine Übereinstimmung mit den festgelegten Geschäftsstrategien und mit dem angestrebten Nutzen. Es können sich auch neue Erkenntnisse ergeben, die zwar im Grunde nicht angestrebt wurden, gleichwohl aber einen noch höheren Nutzen für das Unternehmen darstellen können und somit konsequent verfolgt werden sollten.
Dies erfordert gegebenenfalls eine Anpassung einzelner Schritte und Maßnahmen in den Vorphasen und eine Neudefinition der Ziele, der Daten sowie der Algorithmen. Der gleiche Effekt tritt ein, sollten (Teil-) Ergebnisse unbrauchbar sein oder sollte die wesentliche Fragestellung nicht ausreichend oder korrekt beantwortet werden.
Deployment schlussendlich umfasst die Präsentation der erzielten Ergebnisse und die (regelmäßige) Anwendung des Modells mit den definierten Daten. Hierbei ist zu beachten, dass ein Modell nur für den Zeitpunkt der Fertigstellung und der danach unmittelbar erfolgten Anwendung die ausreichende Qualität besitzt. Veränderungen der Kenngrößen und der Rahmenbedingungen wie auch der zu untersuchenden Fragestellungen erfordern die Erstellung eines neuen oder modifizierten Modells. Dies sollte in einem Plan für eine regelmäßige Überprüfung und Anpassung des Modells berücksichtigt werden.

Von der Ideenfindung zum erfolgreichen PA-Projekt

Abbildung 1: