VON IDG
COMPUTERWOCHE
VOICE OF DIGITAL

Link: https://www.computerwoche.de/a/sechs-stufen-fuer-eine-effiziente-analyse,1899880

Hintergrund Data-Mining-Rezept

Sechs Stufen für eine effiziente Analyse

Datum: 08.07.2009
Autor(en): Christa Manta

Für erfolgreiche BI-Projekte müssen CIOs und Fachabteilung bekanntlich eng zusammenarbeiten und den Prozess gemeinsam anstoßen. Damit das Spiel reibungslos funktioniert und die Analyse zum Erfolg wird, finden sie im Industriestandard CRISP-DM eine Bauanleitung.

Gründe für das Scheitern von **Data-Mining**[1]-Projekten gibt es viele: Mangelhafte **Datenqualität**[2], fehlendes Wissen oder explodierende Kosten etwa oder das Fehlen einer BI-Strategie, von Problemdefinitionen oder konkreten Zielvorgaben für die Analyse. Wichtig für das Gelingen ist, dass alle betroffenen Fachabteilungen in den Data-Mining-Prozess integriert werden und die technischen mit den betriebswirtschaftlichen Kompetenzen verschmelzen. Essenziell ist auch, dass **BI**[3]-Projekte klar definiert werden, ein Anfang und ein Ende haben. Bei der Umsetzung helfen kann CRISP-DM.

CRISP-DM[4] (Cross Industry Standard Process For Data Mining) ist ein Industriestandard, der aus einem Förderprojekt der Europäischen Union von der **Daimler AG**[5], **SPSS**[6], **Teradata**[7] und **OHRA**[8] entwickelt wurde und den Fokus auf die betriebswirtschaftliche Fragestellung richtet. Er ist so etwas wie eine branchenneutrale Bauanleitung für Data-Mining-Projekte, die flexibel genug ist, um individuellen Unterschieden gerecht zu werden und ihren zyklischen Charakter in den Vordergrund rückt. Data Mining ist nämlich ein nicht linearer Prozess, bei dem oft Rücksprünge in die vorherige Phase nötig sind. Die ersten drei Phasen - Projektdefinition, Datensichtung und -aufbereitung - sind dabei oft die aufwändigsten. Sie nehmen bis zu 80 Prozent der gesamten Zeit in Anspruch.

Business Understanding

Am Anfang des Prozesses wird ein Problem definiert oder ein Unternehmensziel festgelegt. Dann werden Kriterien für dessen Erreichen bestimmt. Zum Beispiel kann ein Ziel sein, ein neues Buch im Online-Shop an fünf Prozent der registrierten Kunden zu verkaufen. Daraus leitet sich dann die Frage nach dem Profil der Kunden ab, für die das neue Produkt interessant ist. Für die Problem- oder Zieldefinition ist meistens ein intensiver Austausch zwischen den Fachabteilungen nötig. Eine klare Definition und ein genauer Projektplan helfen dabei, Frustrationen zu vermeiden und später das Projekt zu evaluieren.

Data Understanding

In einem zweiten Schritt werden die Datenquellen ermittelt, die für die Analyse zur Verfügung stehen. Die Datensätze müssen gesichtet und auf ihre Qualität hin beurteilt werden. Das ist ein wichtiger Schritt, denn unvollständige oder fehlerhafte Daten können die Analyse verfälschen.

Data Preparation

Anschließend muss man die relevanten Daten für die Analyse vor- und aufbereiten, also selektieren, bereinigen, integrieren oder formatieren. Wenn nötig werden sie in Abhängigkeit von dem einzusetzenden Algorithmus transformiert, etwa Kunden in Alterskohorten zusammengefasst oder neue Merkmale wie der Gesamtumsatz abgeleitet.

Modelling

In der vierten Prozessphase findet das eigentliche Data Mining statt, nämlich die Bildung eines oder mehrerer geeigneter Modelle. Hierfür muss zunächst die Modelliermethode ausgewählt werden (z.B**. künstliche neuronale Netze**[9], Entscheidungsbäume, **Clusterverfahren**[10], Regression oder Regelinduktion). Anschließend wir ein Test-Design erstellt und damit Trainings- und Testdaten generiert. Schließlich wird das Modell gebaut und mit den richtigen Parametern justiert.

Evaluation

In der Evaluationsphase werden die Ergebnisse der Datenanalyse bewertet. Dafür vergleicht man Data-Mining-Resultate mit den eingangs definierten Erfolgskriterien. Es werden die Modelle ermittelt, die der definierten Problemstellung gerecht werden und akzeptable Daten liefern. In einer Prozessrückschau nimmt man den gesamten DM-Prozess kritisch unter die Lupe. Eventuell muss in einer früheren Phase des Prozesses nachgebessert werden. Zum Schluss werden die nächsten Schritte festgelegt.

Deployment

In der Deployment-Phase wird das Modell auf die aktuellen Daten angewendet. Ein Plan für eine sinnvolle Anwendung der Ergebnisse und die Instandhaltung der Modelle wird erstellt. Dieser Plan muss überwacht und gepflegt werden. Ein Abschlussbereicht oder eine Abschlusspräsentation rundet das Data-Mining-Projekt ab und setzt es für andere, nicht involvierte Personen und Abteilungen in Szene.

Links im Artikel:

[1] **https://www.computerwoche.de/knowledge_center/business_intelligence/1863856/**
[2] **https://www.computerwoche.de/knowledge_center/crm/1895315/**
[3] **https://www.computerwoche.de/schwerpunkt/b/BI.html**
[4] **http://www.crisp-dm.org/index.htm**
[5] **http://de.wikipedia.org/wiki/Daimler_AG**
[6] **http://de.wikipedia.org/wiki/SPSS**
[7] **http://www.teradata.com/t/**
[8] **http://www.ohra.nl/**
[9] **https://www.cio.de/subnet/oracle_bi/890454/index2.html**
[10] **https://www.computerwoche.de/knowledge_center/business_intelligence/1755558/index9.html**
