

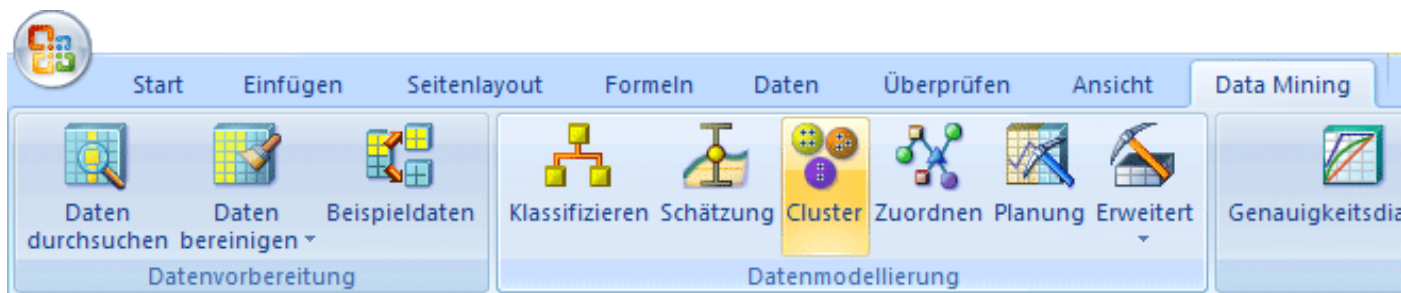
pmmOne

Überblick Data Mining mit Microsoft Excel 2007 und SQL Server

Data Mining, das Finden von Mustern in Daten mit statistischen Methoden, erfreut sich zur Zeit wieder steigender Beliebtheit. Laut den Trendanalysen in Suchmaschinen hat die Nachfrage in den letzten beiden Jahren gewaltig zugelegt und gehört zu den meist nachgefragten Begriffen. Mit einem für Office Anwender kostenlosen Add-in für Excel bietet Microsoft einen einfachen Zugang zu Data Mining für Fachanwender an. Ein Überblick.

Einleitung

Microsoft adressiert das Thema Data Mining rund um die Datenbankplattform SQL Server seit vielen Jahren, der SQL Server unterstützt bereits seit früheren Versionen Data Mining Funktionalitäten, die allerdings wegen der Nichtverfügbarkeit eines benutzerfreundlichen Front-Ends offensichtlich wenig genutzt wurden. Diese Schwäche wurde mit der Office Version 2007 beseitigt, unter dem Microsoft-typisch sperrigen Namen „Microsoft SQL Server 2008 Data Mining-Add-Ins für Microsoft Office 2007“ gibt es seit letztem Jahr ein sehr benutzerfreundliches Addin für Excel. Dieses Add-in ist bei Microsoft zum Download erhältlich und für Anwender mit einer Office 2007 Lizenz sowie Zugang zu einer SQL Server 2005 Datenbank kostenlos. Für die grafische Aufbereitung von z.B. Entscheidungsbäumen wird auch ein Add-in für die Visualisierungssoftware Visio angeboten, die seit einigen Jahren zu Microsoft Office gehört. Ältere Versionen von Excel werden nicht unterstützt, Microsoft will offensichtlich mit dem kostenlosen Zusatzangebot einen weiteren Anreiz für das Upgrade auf Office 2007 in Unternehmen anbieten.



Das Data Mining Ribbon in Microsoft Excel 2007

Obwohl damit der Zugang zu Data Mining auch für Fachanwender mit wenig IT Kenntnissen einfach wurde und das Angebot kostenlos ist, findet die neue Funktionalität bislang weiterhin nur geringes Echo. Der Hauptgrund dafür ist, dass das Add-in nicht bei der Installation von Excel mit installiert wird, sondern getrennt aus dem Internet geladen werden muss und daher den meisten Anwendern schlichtweg unbekannt ist.

Aufbau

Schade eigentlich, denn das Add-in bietet viele grundlegende Data Mining Funktionalitäten unter einer sehr einfach und verständlich aufgebauten Benutzeroberfläche. Der Funktionalitätsumfang adressiert insbesondere Einsteiger ins Thema und unterstützt diese mit vielen Assistenten die Schritt für Schritt durch die Methoden führen. Für einen Großteil der Anwender werden die angebotenen Methoden völlig ausreichen, um einen spürbaren Nutzen bei der Auswertung ihrer Daten zu erreichen.

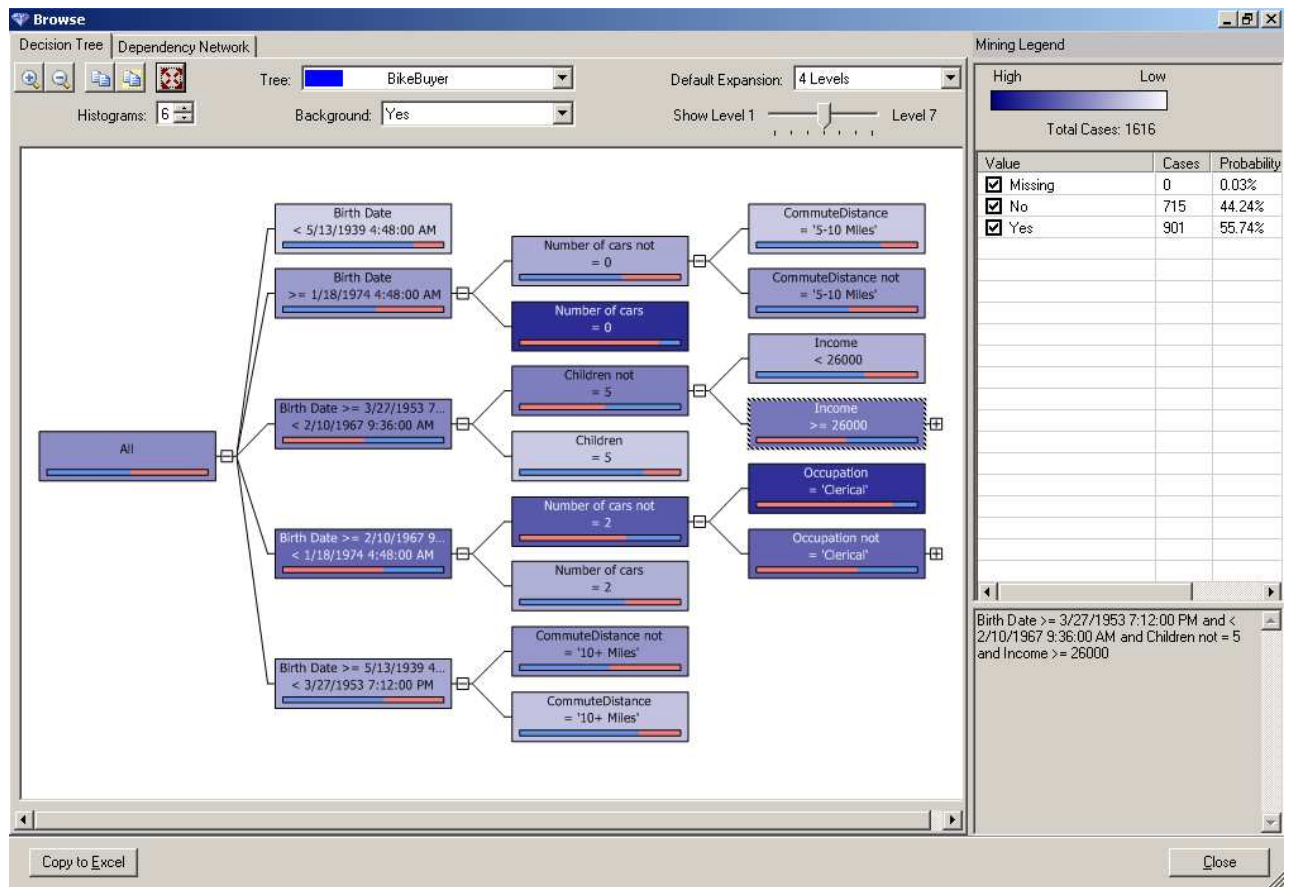


Wizards überall: Insbesondere für Neueinsteiger ins Thema hilfreich

Zum Experimentieren und Überprüfen verschiedener Algorithmen reicht es aus, Testdaten in Excel zu laden und direkt aus Excel Tabellen Analysen mit dem Data Mining Add-in zu erstellen. Das Add-in fordert zwar lizenztechnisch eine aktive Verbindung zu einer SQL Server 2005 Datenbank, die überwiegende Anzahl der Funktionen steht aber auch problemlos mit reinen Excel Dateien zur Verfügung. Nachdem in Excel 2007 die Grenze von 16.000 Zeilen aufgehoben wurde und auch wesentlich größere Datenbestände geladen werden können, kann der Fachanwender auch ohne IT Unterstützung mit Data Mining experimentieren. Dabei erstellte Data Mining Modelle können bei einem späteren Produktivbetrieb dann auf einem SQL Server (Version 2005 SP2 oder 2008 ist erforderlich) gespeichert und angewendet werden.

In Excel werden die Funktionalitäten in zwei Schritten angeboten: Die „Tabellentools“ bestehen aus den beiden „Ribbons“ (so heißen die seit Office 2007 üblichen „Karteikarten“, in denen die Funktionalitäten der Office Produkte gruppiert werden) Analysieren und Entwurf, und bieten Basisfunktionalitäten für Anwender an, die noch nicht mit Data Mining in Berührung gekommen sind. Im Ribbon „Data Mining“ werden dann die „Hardcore“ Data Mining Algorithmen wie Entscheidungsbäume oder Cluster Analysen angeboten. Dort finden sich auch die Funktionen, um Daten für das Data Mining vorzubereiten, beispielsweise um „Datenschmutz“ in den Daten zu

identifizieren und zu bereinigen oder die Daten automatisch in Test- und Trainingsdaten zu unterteilen. Die Funktionen zum Bereinigen der Daten, Identifikation von Ausreißern und erweiterte Szenario und What-if Analysen machen die Installation des Add-ins auch für Anwender interessant, die mit dem eigentlichen Data Mining gar nichts am Hut haben.



Entscheidungsbaum, erstellt über die Funktion "Klassifikation" des Data Mining Add-ins

Sind Data Mining Modelle einmal erstellt und getestet, können diese wahlweise über die grafischen Assistenten des Add-ins angewendet werden oder über hinzugefügte Excel-Formeln wie `dmpredict()`, die sich wie eine einfache Summenformel anwenden lassen, auf eigene Datenbestände angewendet werden.

Ausblick und Bewertung

In der neu releasten Version 2008 des SQL Servers wurden die Data Mining Funktionalitäten ausgeweitet, beispielsweise wurden die Zeitreihenanalysen oder Warenkorbanalysen überarbeitet und verfeinert. Der Umstand, dass Microsoft sich des Themas Data Mining annimmt und einfach zu bedienende Data Mining Funktionalität in Excel integriert, wird das Thema insgesamt voranbringen. Insbesondere die einfache Zugänglichkeit und geringe Einsatzschwelle durch die Integration in Excel und die gute Online-Dokumentation ist hier hervorzuheben. Das Add-in stellt eine starke funktionale Erweiterung des Business Intelligence Angebots von Microsoft dar und wird vielen Anwendern, die Daten in Excel jonglieren, durch die vielen Assistenten auch bei wenig komplexen statistischen Aufgaben hilfreich sein.

In Vorbereitung ist offensichtlich auch ein Thin Client für Data Mining bei Microsoft, der technisch auf den Flash-Konkurrenten „Silverlight“ von Microsoft setzt. Es sollen dabei „Data Mining Services for the Cloud“ angeboten werden, also leicht zugängliche Analysen für Anwender, die sich nicht um die Server Infrastruktur Sorgen machen wollen. Ein erster technischer Preview (siehe Link) dazu ist bereits im Web verfügbar.

Links

Download des Data Mining Add-ins bei Microsoft (Deutsche Version):

<http://www.microsoft.com/downloads/details.aspx?FamilyID=896a493a-2502-4795-94ae-e00632ba6de7&DisplayLang=de>

Einstündiges kostenloses Webseminar zu Data Mining mit Office 2007:

<http://www.pmone.de>

Englische Site mit vielen Informationen, Whitepapers und Beispielen rund um Data Mining mit SQL Server:

<http://www.sqlserverdatamining.com>

Technischer Preview von SQL Server Data Mining Services

<http://www.sqlserverdatamining.com/cloud/>

Rückfragen/Kommentare:

pmOne AG

DEUTSCHLAND

Lindenstraße 12a · D-81545 München
(089) 642499-0

ÖSTERREICH

Pottendorfer Straße 25-27 · A-1120 Wien
(01) 890 28 52-0

SCHWEIZ

Fröschbach 62 · CH-8117 Fällanden
(078) 73807 36

kontakt@pmone.com

www.pmone.de

www.pmone.at

www.pmone.ch