PERSÖNLICH

„Data Mining: Mein Mittel der Wahl für fast alle Analyse-Lagen“ Ein Praxisbericht von Josef Schmid

Seit der Version 4 – damals als Clementine – setze ich die Software SPSS Modeler fast täglich und immer noch mit wachsender Begeisterung ein. IBM SPSS Modeler, die führende Data Mining-Lösung, ist ideal unter anderem für die Optimierung von Marketingkampagnen, Neukundengewinnung, Kundenbindung, Cross- und Upselling, Risiko-Minimierung und Betrugserkennung.

Umfassende Datenumformungsmöglichkeiten

Datenumformungen und die Zusammenführung verschiedenster Datenquellen werden immer wichtiger für Advanced Analytics: Praktisch nie können Datenquellen einfach so verwendet werden, wie sie vorhanden sind. Ebenso ist das Kreieren neuer Felder oftmals einer der wichtigsten Erfolgsfaktoren für ein gutes Modell. Gerade in diesem Bereich ist SPSS Modeler extrem mächtig. Ich kann mich nicht erinnern, dass eine nötige Umformung nicht irgendwie doch möglich war. String-, Datums-, Offsetfunktionen – Modeler bietet so ziemlich alles, was das Herz begehrt. Dies hat mir schon sehr oft aus der „Patsche“ geholfen.

Zeitreihen – ein extrem mächtiger Algorithmus, der aber Anforderungen an die Daten stellt

Die automatische Erstellung von Zeitreihen, d.h. die Erstellung von Forecasts, ist ein ziemlich einzigartiges Feature von Modeler. Ich kenne diverse Kundinnen und Kunden, die diesen Algorithmus mit viel Erfolg in der Praxis einsetzen und erstaunlich genaue Voraussagen erhalten. Modeler kann sehr schnell für eine grosse Anzahl von Zeitreihen individuelle Vorhersagemodelle automatisch erstellen; dabei sind 1000 Modelle in einigen Sekunden eigentlich kein Problem.

Für Zeitreihen müssen die Daten jedoch in einer ganz bestimmten Art und Weise aufbereitet sein: Die Zeitpunkte müssen in Zeilen, die Fälle (Produkte, Kunden …) sollten als Spalten vorliegen. Die Umformung von Daten in ein solches Format bewältigt Modeler mit „links“ – der Restructure Node kombiniert mit einem Aggregate Node erledigt dies im Handumdrehen.

Aber: Mit dem Essen kommt der Appetit. In einem konkreten Projekt sollten plötzlich nicht mehr nur 1000 Modelle erstellt werden, sondern es ging um mehrere Millionen von Modellen. Dies ist eigentlich nicht unbedingt ein Problem – aber in diesem konkreten Fall sollte der ganze Prozess innerhalb einer Datenbank durchgeführt werden, was Modeler ja mittels SQL-Pushback vorbildlich unterstützt. Und hier setzen normalerweise die Schwierigkeiten ein: Wie oben beschrieben sollten die Fälle in Spalten vorliegen, wir benötigten also mehrere Millionen Spalten. Gewisse Datenbanken unterstützen jedoch nur 1024 Spalten – also nur einen Bruchteil der benötigten Anzahl!

Modeler Scripting – der Schritt zur Vollautomatisierung

Guter Rat ist normalerweise teuer – aber mit Modeler lässt sich auch hier recht schnell eine Lösung finden: Da manche Datenbanken nur 1024 Spalten unterstützen, und dies war hier der Fall, muss der Modellierungsprozess in mehrere tausend Teilschritte unterteilt und wiederholt werden. Hier kommt nun eine Funktion von Modeler zum Tragen, die nach meiner Erfahrung nur wenigen Anwendern bekannt ist: Mittels Modeler Scripting lässt sich die Unterteilung des Prozesses in Teilschritte automatisieren.

Die implementierte Lösung schliesslich sieht so aus, dass auf Modeler Server ein Scheduler einen Modeler Prozess anstösst, der automatisch die Daten aus der Datenbank liest, in Teilschritte unterteilt, automatisch für jeden einzelnen der mehreren Millionen Datensätze ein individuelles Forecasting Modell erstellt und

schliesslich diese Modelle benützt, um konkrete Voraussagen in die Datenbank zu schreiben.

„Das Schweizer Taschenmesser" der Daten-Analyse

Modeler überrascht immer wieder mit der Vielfalt seiner Fähigkeiten und bietet individuelle Lösungen, die ich ursprünglich nicht für möglich hielt… Manche unterschätzen – verleitet durch die Einfachheit der Oberfläche – die enorme Vielfalt der Funktionen unter der „Motorhaube". Modeler ist wirklich eine Art „Schweizer Taschenmesser" der Analyse – Unmögliches wird möglich und man kann es für fast alles einsetzen! Und – last but not least – die Voraussagegenauigkeit der Modelle lässt ebenso immer wieder staunen. ●

Josef Schmid, Managing Partner Dynelytics AG
arbeitet seit den Anfängen des Data Mining intensiv und voller Begeisterung mit den Möglichkeiten, die diese Art der Datenanalyse mit selbstlernenden Algorithmen bietet. Seine langjährige Erfahrung in Kundenprojekten in allen Branchen und mit den unterschiedlichsten Fragestellungen erlauben es ihm, kreative und effiziente Lösungsansätze für praktisch jedes Datenanalyse-Problem zu finden. Und für Josef Schmid liegt es auf der Hand, dass der Bedarf und die Einsatzfelder an Data Mining mit den zunehmenden technischen Entwicklungen im Bereich Big Data noch zunehmen werden.
j.schmid@dynelytics.com,
+41 (0) 44 266 90 30