Oracle Predictive Queries: Der einfache Weg zu neuen Erkenntnissen

Reinhard Mense und Marco Nätlitz, areto consulting gmbh

Mit den in der Version 12c enthaltenen „Predictive Queries" (PQ) erhalten die Anwender die Möglichkeit, explorative Analysen direkt in der Oracle-Datenbank auszuführen, und somit einen einfachen und kostengünstigen Einstieg in die Daten-Analyse. Auf Knopfdruck lassen sich die Daten nach bestimmten Merkmalen untersuchen, um Erkenntnisse über bisher verborgene Chancen und Potenziale zu erhalten. Anhand eines einfachen Beispiels werden die Voraussetzungen und Vorgehensmodelle für den Einsatz von PQ gezeigt sowie Grenzen und Unterschiede zu anderen Methoden und Tools erklärt.

Selten waren sich die führenden Marktforschungs-Unternehmen so einig: Predictive Analytics (PA) wird unisono als der Treiber für mehr Produktivität, bessere Kundenbeziehungen und höhere Umsätze erachtet. PA verheißt effizientere Maschinen-Auslastungen, geringere Ausfälle beziehungsweise Stillstand-Zeiten und eine systematische Identifizierung und Erschließung neuer Marktpotenziale. So ist es kaum verwunderlich, dass nach einer Studie von Pierre Audoin Consultants immerhin 30 Prozent aller Unternehmen Ausgaben für PA-Software in den nächsten beiden Jahren geplant haben. Obschon die Quote hoch erscheint, fragt man sich doch: Warum nur eine Minderheit, warum nur 30 Prozent, was ist mit dem Rest?

PA ist keine Methodik, die sich quasi von allein erschließt und im Handumdrehen oder auf magische Weise revolutionäre Ergebnisse erbringt. Die Analyse sollte vielmehr durch speziell ausgebildete Mitarbeiter vorgenommen werden, die nicht nur über ein tiefgründiges Verständnis der mathematischen Grundlagen verfügen, sondern auch betriebswirtschaftliche Fragestellungen nachvollziehen können. Die Rolle solcher Data Scientists ist ebenso anspruchsvoll wie auch gefragt: Die beiden Autoren T. H. Davenport und D. J. Patil haben diese Aufgabe in den Harvard Business News gar als „the sexiest job of the 21st century bezeichnet." Und dieser hohe Anspruch an das Personal erklärt wiederum zumindest teilweise die bisher geübte Zurückhaltung der Unternehmen bei der Nutzung dieser Technologie. Es fehlen Know-how und eben das geeignete Personal. Viele Betriebe scheuen überdies den hohen Implementierungsaufwand sowie die Komplexität der Anwendungen und befürchten eine zu geringe Datenqualität, wie die Analysten von Pierre Audoin Consultants in ihrer Studie noch herausfanden.

Gerade für jene Unternehmen, die PA bisher eher skeptisch und abwartend betrachtet haben, bieten sich mit den PQ der Version 12*c* der Oracle-Datenbank die Möglichkeit, PA schnell und ohne großen Aufwand zu testen.

Predictive Queries – Ad-hoc-Analysen der Daten

PQ sind ein neues Feature der Datenbank-Version 12*c* und erweitern den Funktionsumfang von SQL. Die Anwendung ist recht einfach: Der Datenbank-Programmierer schreibt ein SQL-Script und wendet es auf die Daten an. Dies geschieht in einem Schritt und erfordert keine tiefgreifenden Kenntnisse der internen Abläufe in der Datenbank. Nach Ausgabe der Ergebnisse wird das durch die Datenbank errechnete Modell wieder gelöscht, was auch als „flüchtige" oder „temporäre" Modellierung bezeichnet wird. Bei der Analyse wird auf die eingebauten Data-Mining-Algorithmen zugegriffen, ohne dass der Anwender jedoch das vollständige Verständnis aller notwendigen Voreinstellungen besitzen oder das sonst übliche Feintuning der Algorithmen vornehmen muss. Dies schränkt die Bandbreite der Nutzung sicherlich ein, erlaubt aber auch dem unerfahrenen Anwender erste Gehversuche mit PA, sodass er sich stärker auf die Daten und deren Ergebnisse als auf komplexe Vorarbeiten und Programmierungen konzentrieren muss.

PQ sind in der Lage, die Daten zu partitionieren und die vorher definierten Analysefunktionen und Queries auf diese Teilmengen anzuwenden. Dies geschieht automatisch, ohne dass es weiterer Eingriffe des Programmierers bedarf. Der Ablauf ist dabei wie folgt: Im Zentrum steht die Prozedur „PREDICTION", die in ein „SELECT"-Statement eingebunden werden kann. Der Prozedur „PREDICTION" kann nun als Parameter übergeben werden, welche Spalte prognostiziert werden soll und welche Spalten zur Modell-Entwicklung herangezogen werden sollen. Gleichzeitig kann mit „PARTITION BY" angegeben werden, welche Spalten eine Partitionierung innerhalb der Daten definieren.

Enthält eine Kunden-Tabelle etwa eine Spalte „Geschlecht" oder „Bundesland", so könnte die Modell-Entwicklung direkt anhand der partitionierten Daten nach Geschlecht beziehungsweise Bundesland erfolgen. Das Ergebnis der Prognose ist das Result-Set des ausgeführten SQL-Statements. Mit dem so erhaltenen Ergebnis lässt sich auf einfache Weise überprüfen, wie aussagekräftig und nutzbar die Daten wirklich sind. Dazu bietet sich folgende Vorgehensweise an: Zur Validierung der Ergebnisse werden historische Daten in zwei unterschiedlichen Zeiträumen herangezogen. Auf die Werte des ersten, früheren Zeitfens-

Beispiel: Oracle Predictive Queries

Wie kann der Preis eines Weins aus Bordeaux ermittelt werden? Die beliebten und häufig teuren Kennerweine werden unter anderem durch die Verkostung von Experten preislich eingestuft. Alternativ müssten die Winzer entlang der französischen Küste über Wetterbedingungen während der Ernte- und Reifezeit befragt werden, denn die kostbarsten Trauben entwickeln einen besonderen Geschmack während heißer und trockener Sommer.

Der Ökonom und Weinliebhaber Prof. Dr. Orley Ashenfelter verließ sich nicht länger auf die menschliche Intuition der Experten und entwickelte ein Predictive-Analytics-Modell, das die Preise von Bordeaux-Wein vorhersagen kann. Dieses Modell (siehe *„http://www.ft.com/intl/cms/s/0/1e9cb152-5824-11dc-8c65-0000779fd2ac.html“*) sorgte für Aufruhr in der Weinbranche, denn plötzlich konnte der Weinhändler die Preise der Weine ermitteln, bevor diese die nötige Reifezeit erreicht hatten (siehe *„https://www.betterment.com/resources/investment-strategy/behavioral-finance-investing-strategy/human-vs-algorithm-investing-a-lesson-from-wine-country“*). Ashenfelter nutzte eine Datenbasis von fünfundzwanzig Bordeaux-Weinen aus den Jahrgängen 1952 bis 1987. Die abhängige Variable – das, was prognostiziert werden soll – ist der typische Preis (P) im Rahmen von Wein-Auktionen aus den Jahren 1990 und 1991.

Pro Jahrgang ermittelte Ashenfelter ebenfalls die unabhängigen Variablen „Durchschnittstemperatur während der Reifezeit“ (DTMP), „Regen während der Ernte“ (REGE) und „Regen im Winter“ (REGW). Der logarithmierte Preis und das Alter der Weine sind in *Abbildung 1* eingetragen. Anhand der eingezeichneten Geraden lässt sich bereits ein linearer Zusammenhang erkennen. Das prädikative Modell von Ashenfelter setzt lineare Regression über die Variablen DTMP, REGE und REGW ein, um die abhängige Variable P, den Preis, zu prognostizieren (siehe *„http://www.liquidasset.com/orley.htm“*).

Mithilfe von Predictive Queries ist das Modell von Ashenfelter schnell nachgebaut. Benötigt wird eine Tabelle „Wein“, die die fünfundzwanzig Jahrgänge inklusive Preis und der Eigenschaften DTMP, REGE und REGW als Spalten enthält. Die Funktion „prediction“ kann nun das lineare Modell für die Vorhersage zukünftiger Weinpreise erstellen *(siehe Listing 1)*.

Die Datenbank erstellt nun automatisch ein Modell, das mithilfe der linearen Regression die abhängige Variable „Preis“ anhand der unabhängigen Variablen „DTMP“, „REGE“ und „REGW“ vorhersagen kann. Werden nun weitere Jahrgänge in die Tabelle eingefügt, für die der Preis bisher unbekannt ist, jedoch „DTMP“, „REGE“ und „REGW“ bekannt sind, dann prognostiziert die oben dargestellte Query die Preise in der Spalte „PREIS_VORHERSAGE“. Das so nachgestellte Modell nach Ashenfelter erreicht eine Prognose-Genauigkeit von circa 83 Prozent gegenüber Testdaten. Zu gleichen Ergebnissen und Preisvorhersagen gelangt in unserem Test auch die Statistikumgebung R, die unter Data Scientist die Plattform der Wahl darstellt (siehe *„https://www.r-project.org“*).

Das Beispiel veranschaulicht die Vor- und Nachteile der Predictive Queries, die ab Version 12*c* in der Datenbank zur Verfügung stehen. Mit der Funktion „prediction“ sind schnell Predictive-Analytics-Modelle gebaut, die erste Ergebnisse und Analysen von Daten erlauben. Sie können ohne zusätzliche Statistik-Software und Detail-Kenntnisse im Bereich der Predictive Analytics eingesetzt werden. Die Modelle sind allerdings nicht parametrisierbar, weshalb sie nicht besser an den Datensatz angepasst werden können. Auch können die Modelle nicht gespeichert werden, denn die Datenbank hält diese nur während der Ausführung der Query im Arbeitsspeicher vor.

```
select
    jahr
  , dtmp
  , rege
  , regw
  , preis
  , prediction(
for preis using dtmp, rege, regw
) over ( ) preis_vorhersage
from
  wein;
```

Listing 1

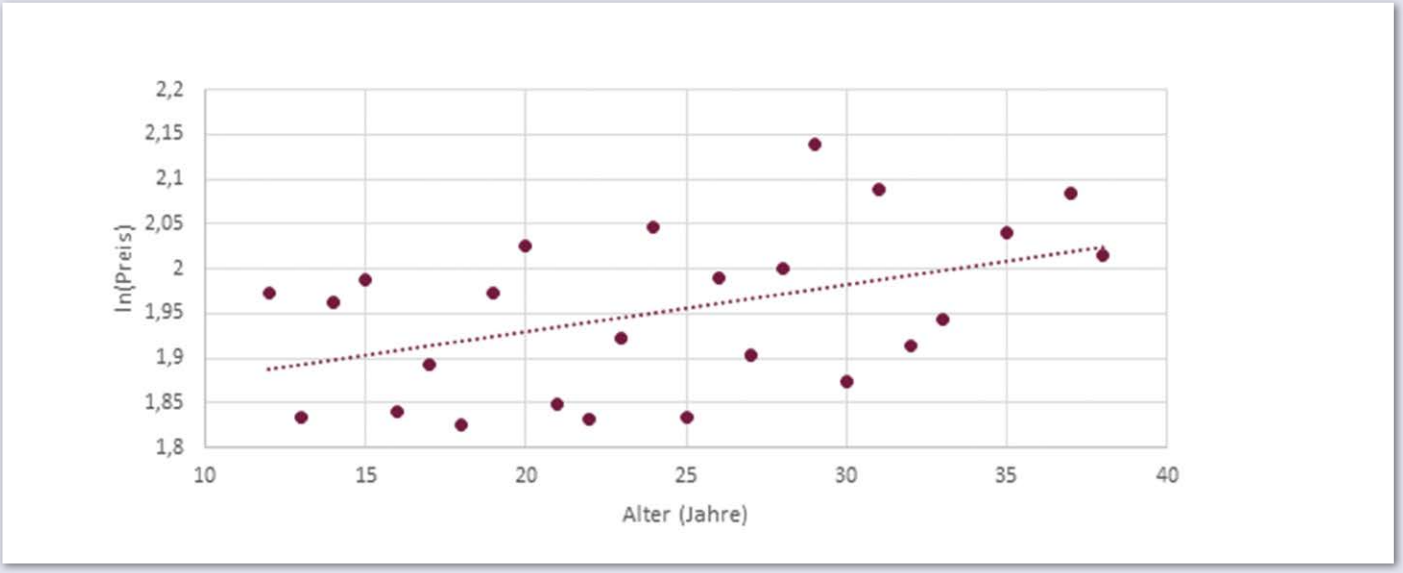

Abbildung 1: Preis (logarithmiert) gegen Alter in Jahren

ters kann der Datenbank-Entwickler zugreifen, die Schätzdaten wie auch die Ist-Daten der nachfolgenden Periode können hingegen nicht eingesehen werden.

Für eben diesen Zeitraum soll nun mithilfe der PQ eine Prognose erstellt und mit den tatsächlich erreichten sowie mit den durch die Experten vorhergesagten Werten verglichen werden. Hier zeigt sich dann, wie groß die Abweichungen sind und welche Erkenntnisse und Mehrwerte für das Unternehmen entstehen könnten.

Am Ende dieses kurzen und einfachen Prozesses stehen sich drei Werteräume gegenüber: Die tatsächlich erreichten Zahlen, die durch die Experten vormals prognostizierten Werte und eben jene Daten, die durch den Einsatz von PQ errechnet wurden. Ordnet man den Abweichungen von den Ist-Werten nun entsprechende Geldwerte zu oder quantifiziert diese anderweitig, so zeigt sich die Wirtschaftlichkeit der angewandten Methoden und somit die fallbezogene Wirksamkeit von PA. Dieses bewährte Vorgehensmodell erfüllt gleich mehrere Funktionen:

- Quantifizierung des Nutzens und Bestimmung des Return on Investment
- Nachweis der Wirksamkeit der PA-Methoden im Vergleich zu konventionellen Verfahren
- Einfache Validierung möglicher Anwendungsfälle
- Schaffen des notwendigen Vertrauens und Akzeptanz von PQ beziehungsweise Predictive Analytics im Allgemeinen

Sollte das Resultat nicht überzeugen, so muss dies nicht das Ende aller Bemühungen sein: Bessere, da auch stärker beeinflussbare Ergebnisse können durch den Einsatz der Advanced Analytics Option (AAO) erzielt werden. Die AAO erweitert die Datenbank um die zwei Komponenten R Enterprise und Data Mining. Oracle stellt eine umfassende Plattform zur Verfügung, die über alle nötigen Werkzeuge verfügt und Unternehmen in die Lage versetzt, maßgeschneiderte Analysen und Modelle im PA-Umfeld zu entwickeln. Denn mit R Enterprise verheiratet Oracle die mächtige und verbreitete Statistikumgebung R mit den bewährten Features der Oracle-Datenbank. Etwaige Analysen können so in R programmiert und durch die Datenbank integriert, parallelisiert sowie skalierbar berechnet werden. Darüber hinaus können mit Oracle Data Mining Workflows erstellt werden, die in der Datenbank integrierte Data-Mining- und PA-Algorithmen auf Daten anwenden. Diese Workflows können komfortabel mit einer grafischen Benutzeroberfläche innerhalb des SQL Developer entwickelt sein. Somit lassen sich mit AAO komplexe und umfassende Probleme im PA-Kontext effizient bewältigen.

Ein weiteres Unterscheidungsmerkmal der beiden Lösungsansätze betrifft das notwendige Fachpersonal. Während die Erprobung der Methoden und damit verbundene erste Schritte auch durch klassische Datenbank-Entwickler durchgeführt werden können, empfiehlt es sich beim Einsatz der AAO, auf Data Scientists oder vergleichbar ausgebildete Mitarbeiter zurückzugreifen. Diese sind nicht nur in der Lage, an den notwendigen Stellschrauben zur Verfeinerung der Ergebnisse zu drehen, sondern bringen auch das notwendige Verständnis für die mathematisch-statistischen Grundlagen und auch für die zu untersuchenden betriebswirtschaftlichen Fragestellungen mit. Spätestens mit dem Einsatz der AAO ist ein systematisches und methodisches Vorgehen unabdingbar. Dazu eignet sich besonders der Standard „Cross Industry Standard Process for Data Mining" (CRISP-DM). Dieses bereits im Jahr 1999 veröffentlichte Rahmenwerk beschreibt den gesamten Ablauf in sechs Phasen:

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

Advanced Analytics – die nächste Stufe benötigt Fachleute

Dieser Artikel betrachtet ausschließlich die Phasen „Data Preparation" und „Modelling", da diese ja durch PQ überwiegend übernommen werden. Im Mittelpunkt der Data Preparation steht die Definition eines finalen Algorithmus für die spätere Modell-Entwicklung. Spätestens zu diesem Zeitpunkt müssen Ausreißer, Messfehler oder fehlende Variablen sowie andere Unzulänglichkeiten ermittelt und bereinigt worden sein. Die mithilfe 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 Modelling-Phase beschreibt die 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 Test-Designs 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 erfordert oder auch zu Änderungen der Teilschritte aus den Vorphasen führen kann.

Fazit

Bereits dieser kurze Abriss zeigt, dass eine systematische und auf Nutzen-Maximierung ausgerichtete Beschäftigung mit dem Thema „Predictive Analytics" keine spontane Angelegenheit sein kann, sondern einen Rahmen benötigt, der bei vielen Firmen nur bedingt gegeben ist. Insofern verstehen sich die PQ als ein guter Einstieg, um „sanft" und ohne hohen Aufwand an das Thema herangeführt zu werden. Die nächsten Schritte erfolgen dann eher mit einem klassischen PA-Tool, mit den richtigen Fragenstellungen und von ausgebildeten Mitarbeitern.

Die weitere Entwicklung von PQ dürfte spannend werden: Welche zusätzlichen Funktionen zur Daten-Analyse und zur Verfeinerung möglicher Ergebnisse wird Oracle zukünftig in den Standardumfang der Datenbank einbauen und wie werden diese Möglichkeiten durch die Anwender angenommen? Der Trend zur Automatisierung von BI- und Analyse-Funktionen wird sich fortsetzen. Inwieweit Predictive Queries das Zeug zu „Business Analytics for the Masses" haben werden, bleibt abzuwarten.

Reinhard Mense
reinhard.mense@areto-consulting.de

Marco Nätlitz
marco.nätlitz@areto-consulting.de