WILEY

TIMELY. PRACTICAL. RELIABLE.

Data Mining with Microsoft® SQL Server® 2008

Jamie MacLennan
ZhaoHui Tang
Bogdan Crivat

Table of Contents

Data Mining with Microsoft® SQL Server® 2008

Jamie MacLennan
ZhaoHui Tang
Bogdan Crivat

WILEY

Wiley Publishing, Inc.

Data Mining with Microsoft® SQL Server® 2008

Published by

Wiley Publishing, Inc.

10475 Crosspoint Boulevard

Indianapolis, IN 46256

www.wiley.com

Published by Wiley Publishing, Inc., Indianapolis, Indiana

Published simultaneously in Canada

ISBN: 978-0-470-27774-4

Library of Congress Cataloging-in-Publication Data

MacLennan, Jamie.

Data mining with Microsoft SQL server 2008 / Jamie MacLennan, Bogdan Crivat, ZhaoHui Tang.

p. cm.

Includes index.

ISBN 978-0-470-27774-4 (paper/website)

1. SQL server. 2. Data mining. I. Crivat, Bogdan. II. Tang, Zhaohui. III. Title.

QA76.9.D343M335 2008

005.75′85—dc22

2008035467

To Logan, because he needs it the most.
—Jamie MacLennan
This book is for Cosmin, with great hope that he will someday find math (and data mining) to be fun and interesting.
—Bogdan Crivat

About the Authors

Jamie MacLennan is the principal development manager of SQL Server Analysis Services at Microsoft. In addition to being responsible for the development and delivery of the Data Mining and OLAP technologies for SQL Server, MacLennan is a proud husband and father of four. He has more than 25 patents and patents pending for his work on SQL Server Data Mining. MacLennan has written extensively on the data mining technology in SQL Server, including many articles in *MSDN Magazine*, *SQL Server Magazine*, and postings on `SQLServerDataMining.com` and his blog at `http://blogs.msdn.com/jamiemac`. This is his second edition of *Data Mining with SQL Server*. MacLennan has been a featured and invited speaker at conferences worldwide, including Microsoft TechEd, Microsoft TechEd Europe, SQL PASS, the Knowledge Discovery and Data Mining (KDD) conference, the Americas Conference on Information Systems (AMCIS), and the Data Mining Cup conference.

ZhaoHui Tang is a group program manager at Microsoft adCenter Labs, where he manages a number of research projects related to paid search and content ads. He is the inventor of Microsoft Keyword Services Platform. Prior to adCenter, he spent six years as a lead program manager in the SQL Server Business Intelligence (BI) group, mainly focusing on data mining development. He has written numerous articles for both academic and industrial publications, such as *The VLDB Journal* and *SQL Server Magazine*. He is a frequent speaker at business intelligence conferences. He was also a co-author of the previous edition of this book, *Data Mining with SQL Server 2005*.

Bogdan Crivat is a senior software design engineer in SQL Server Analysis Services at Microsoft, working primarily on the Data Mining platform. Crivat has written various

articles on data mining for *MSDN Magazine* and *Access/VB/SQL Advisor Magazine*, as well as numerous postings on the SQLServerDataMining.com website and on the MSDN Forums. He presented at various Microsoft and data mining professional conferences. Crivat also blogs about SQL Server Data Mining at www.bogdancrivat.net/dm.

Credits

Executive Editor
Robert Elliott

Development Editor
Kevin Shafer

Technical Editors
Raman Iyer; Shuvro Mitra

Production Editor
Dassi Zeidel

Copy Editor
Kathryn Duggan

Editorial Manager
Mary Beth Wakefield

Production Manager
Tim Tate

Vice President and Executive Group Publisher
Richard Swadley

Vice President and Executive Publisher
Joseph B. Wikert

Project Coordinator, Cover
Lynsey Stanford

Proofreader
Publication Services, Inc.

Indexer
Ted Laux

Cover Image
© Darren Greenwood/Design Pics/ Corbis

Acknowledgments

First of all we would like to acknowledge the help from our data mining team members and other colleagues in the Microsoft SQL Server Business Intelligence (BI) organization. In addition to creating the best data mining package on the planet, most of them gave up some of their free time to review the text and sample code. Direct thanks go to Shuvro Mitra, Raman Iyer, Dana Cristofor, Jeanine Nelson-Takaki, and Niketan Pansare for helping review our text to ensure that it makes sense and that our samples work. Thanks also to the rest of the data mining team, including Donald Farmer, Tatyana Yakushev, Yimin Wu, Fernando Godinez Delgado, Gang Xiao, Liu Tang, and Bo Simmons for building such a great product. In addition, we would like to thank the SQL BI management of Kamal Hathi and Tom Casey for supporting data mining in SQL Server.

SQL Server 2008 Data Mining (including the Data Mining Add-Ins) is a product jointly developed by the SQL Server Analysis Services team and other teams inside Microsoft. We would like to thank colleagues from Excel —notably Rob Collie, Howie Dickerman, and Dan Battagin, whose valuable input into the design of the Data Mining Add-Ins guaranteed their success. Also thanks to those in the Machine Learning and Applied Statistics (MLAS) Group, headed by Research Manager David Heckerman, who continue to advise us on deep algorithmic issues in our product. We would like to thank David Heckerman, Jesper Lind, Alexei Bocharov, Chris Meek, Bo Thiesson, and Max Chickering for their contributions.

We would like to give special thanks to Kevin Shafer for his close editing of our text, which has greatly improved the quality of this manuscript. Also thanks to Wiley Publications acquisitions editor Bob Elliot for his support and patience.

Special thanks from Jamie to his wife, April, who yet again supported him through the ups and downs of authoring a book, particularly during painful rewrites and recaptures of screen shots, while taking care of our kids and the world around me. Elalu, honey.

Bogdan would like to thank his wife, Irinel, for supporting him, reviewing his chapters, and some really helpful hints for capturing screen shots.

Foreword

The world is absolutely exploding with digitally born data. Financial transactions, online advertising analytics, consumer preference information, and the results of scientific discovery mean tremendous volumes of data exist in both structured and unstructured stores today. And it is growing faster than ever before, fueled by both technology and a new generation of people adopting and integrating technology into all aspects of their lives.

Business intelligence practitioners struggle to make sense of the data in their charge to help their businesses operate with better understanding of what is influencing results. Trends are evolving and changing more quickly than ever before. It is no longer enough to look at historical data to just determine what happened. Aided by data mining, you can more readily understand why something happened. It can make the difference in whether history —good or bad — repeats itself. Because trends change at such great speed today, automated analysis and sophisticated algorithms for identifying trends, finding outliers, and predicting future courses quickly can be the difference between winning and just competing. Data mining provides the means to make sense of tremendous volumes of data by automating the processes of categorizing and clustering common elements, identifying trends and anomalies in the data, and predicting what will happen given those factors.

I have had the pleasure to work alongside (and learn directly from) Jamie MacLennan and Bogdan Crivat. They are passionate about the difference that technology can make in our lives, and committed to putting the tools necessary to make sense of the expanding world of data into everyone's hands. In this book, they share their passions with you, clearly explaining data mining concepts, and how to apply

them in common situations using the very algorithms and tools they authored themselves as part of Microsoft SQL Server. This book provides an opportunity for you to learn straight from the source, too. I am sure you will discover that this text is a valuable resource.

Tom Casey
General Manager, SQL Server Business Intelligence
Microsoft Corporation

Introduction

Microsoft SQL Server 2008 is the third version of SQL Server that ships with included data mining technology. Since it was introduced in SQL Server 2000, data mining has become a key feature of the larger product. Data mining has grown from an isolated part of SQL Server Analysis Services with two algorithms, to an intrinsic part of the SQL Server Business Intelligence (BI) platform that is fully integrated with OLAP, Integration Services, and Reporting Services. Other Microsoft applications (such as Microsoft Dynamix CRM and Microsoft Performance Point Server) seamlessly integrate SQL Server Data Mining to accentuate their functionality with predictive power.

SQL Server Data Mining has become the most widely deployed data mining server in the industry, with many third-party software and consulting companies building on, specializing, and extending the platform. Enterprise, small and medium business, and even academic and scientific users have all adopted or switched to SQL Server Data Mining because of its scalability, availability, extensive functionality, and ease of use.

This book serves as a guide to SQL Server Data Mining, explaining how it works, providing detailed technical and practical discussions of the SQL Server Data Mining technology, and demonstrating why you should deploy and use SQL Server Data Mining for yourself.

How This Book Is Organized

This book is written to provide you with the knowledge necessary to implement successful data mining solutions using SQL Server, by introducing the overall space,

familiarizing you with the tools, giving depth and breadth on the Microsoft data mining algorithms, and then providing details on various ways to implement data mining solutions.

The book starts with introductory chapters that outline the tools, technologies, and ideas you need to leverage SQL Server Data Mining. Then each of the SQL Server data mining algorithms is described in detail in its own chapter. The subsequent chapters describe how you can integrate SQL Server Data Mining into other parts of the SQL Server BI suite. The latter part of the book deals with architecture and programming issues, and gives examples of some data mining implementation scenarios.

Following is a brief description of the chapters:

- **Chapter 1: Introduction to Data Mining** —This chapter introduces not only the book, but also the technology. It contains a detailed definition of what exactly is meant by the term *data mining*, and discusses what kinds of problems are addressed by this technology.
- **Chapter 2: Applied Data Mining Using Office 2007** —This chapter provides an overview of the Table Analysis Tools for Office 2007 add-in, which is a rich set of tools for Excel that are usable by any information worker. This chapter explains how and why you use these tools, and provides guidance on how to get the best results.
- **Chapter 3: Data Mining Concepts and DMX** —This chapter is critical to understanding the SQL Server Data Mining platform. It explains the underlying concepts of how you should think about a data mining problem, as well as providing a learn-by-example framework for Data Mining Extensions (DMX) to SQL.
- **Chapter 4: Using SQL Server Data Mining** —This chapter introduces you to building data mining solutions using Business Intelligence Development Studio (BI Dev

Studio). In addition to a basic overview, it provides a wide range of tips and tricks that can make the difference between a successful project and a failed one. This chapter also covers using SQL Server Management Studio to access and secure data mining objects. In addition, it tells you how you can expose your data mining models through SQL Server Reporting Services.

- **Chapter 5: Implementing a Data Mining Process Using Office 2007** —This chapter explores the remaining tools in the Data Mining Add-ins for Office 2007. As described in this chapter, these tools provide more functionality than BI Dev Studio and SQL Server Management Studio alone, but they also have limitations that prevent them from exposing the full functionality of SQL Server Data Mining. In any case, this chapter will allow you to best take advantage of the Microsoft Office tools for data mining.
- **Chapters 6–12: the algorithm chapters** —Each of these chapters is devoted to one or more of the algorithms included with SQL Server Data Mining. In each of the chapters, you will find a basic description of the algorithm, followed by usage scenarios that will help you understand how, when, and where you apply each algorithm. Each chapter describes how you create, train, interpret, and apply models using the specified algorithms. The chapters wrap up with a deeper technical dive into how the algorithms work.
- **Chapter 13: Mining OLAP Cubes** —This chapter provides a brief introduction to Online Analytical Processing (OLAP) and the OLAP functionality of SQL Server Analysis Services. The chapter examines how and when you perform data mining on OLAP cubes. It also includes details on how to implement popular OLAP mining scenarios.

- **Chapter 14: Data Mining with SQL Server Integration Services** —This chapter introduces SQL Server Integration Services (SSIS) and describes its various components. It then details the tasks and transformations that you use to implement data mining solutions in your data integration packages. This chapter also describes how to use the text mining components to prepare unstructured data for data mining scenarios.
- **Chapter 15: SQL Server Data Mining Architecture** —This is the first chapter that moves away from tools and concepts and starts to delve into the programming and administration aspects of SQL Server Data Mining. This chapter discusses the architecture of a server-based data mining system, including the XML for Analysis (XMLA) protocol that underlies all client-server communication. The chapter also describes the administration of a data mining server, including server properties that are important for SQL Server Data Mining and data mining security roles.
- **Chapter 16: Programming SQL Server Data Mining** —This chapter details the programming interfaces for SQL Server Data Mining, and includes several examples of the programmatic creation, training, and application of data mining objects.
- **Chapter 17: Extending SQL Server Data Mining** —This chapter describes how you can extend SQL Server Data Mining with your own functionality. It shows you how to create stored procedures for adding operations to DMX. It also describes how you can implement your own data mining algorithms to plug into SQL Server Data Mining and exploit its features. Additionally, this chapter describes how you can write your own data mining visualizations to display patterns in either the supplied algorithms or your own algorithm

implementations, and embed them in BI Dev Studio and SQL Server Management Studio.

- **Chapter 18: Implementing a Web Cross-Selling Application** —This chapter walks you through a common data mining scenario—implementing a recommendation engine and integrating it into a retail website. It includes sample queries and code to get you started.
- **Chapter 19: Conclusion and Additional Resources** —In addition to wrapping up the book, this chapter provides a list of valuable links where you can find additional information and help with your data mining projects. It also includes references to some other reading materials that you can refer to if you want to learn more about data mining.

This book also includes two helpful appendixes:

- **Appendix A: Data Sets** —This appendix contains a brief description of the various data sets used in this book.
- **Appendix B: Supported Functions** —This appendix provides, for your reference, a list of all the supported DMX functions. It also contains lists of all Visual Basic for Applications (VBA) and Excel functions that you can call from DMX. It also describes some supplemental stored procedures provided by the authors to assist with the sample queries presented throughout the text.

Who Should Read This Book

This book is primarily designed for the SQL Server user who is curious about data mining. A working knowledge of SQL will be greatly beneficial in understanding DMX and the DMX queries sprinkled throughout the book. However, non–SQL users can still benefit from the Office 2007 and the algorithm chapters. Readers who are interested in

programming SQL Server Data Mining should understand .NET and the C# languages to apply the relevant chapters.

For those of you who have read the previous edition of this book, *Data Mining with SQL Server 2005* (Indianapolis: Wiley, 2005), welcome back! In this text, you will find comprehensive material on the new functionality of Microsoft SQL Server 2008 Data Mining plus new examples for most algorithm and scenarios described in the text.

Conventions

To help you get the most from the text and keep track of what's happening, a number of conventions are used throughout the book.

> **Note**
>
> Notes and other information that is supplemental to the current discussion are offset and placed in italics like this.

Within the main text, the following conventions are used:

- Important words or terms are *italicized* when they are first introduced in the text.
- Combination keyboard strokes are shown like this: Ctrl+A.
- Filenames, URLs, and code within the text are differentiated from the rest of the text with a special font, as shown in this example: `persistence.properties`
- Blocks (or snippets) of code are shown two different ways:

```
In code examples, new and important code is highlighted with
a gray background.
```

```
The gray highlighting is not used for code that's less important in the
present context, or has been shown before.
```

Tools You Will Need

In order to get the most benefit from this book, you will need access to the SQL Server 2008 Analysis Services software. SQL Server 2008 Analysis Services is included with the Standard, Enterprise, and Developer editions of Microsoft SQL Server 2008. Time-based evaluation versions are available for download at http://www.microsoft.com/sql. To follow along with Chapters 2.1 and 5.1, you will also need Microsoft Office 2007 and SQL Server 2008 Data Mining Add-Ins for Office 2007. Evaluation versions of Microsoft Office 2007 are available at www.microsoft.com/office, and the free download of the Data Mining Add-Ins is available at www.microsoft.com/sql/dm.

You'll also want to have the AdventureWorksDW2008 database installed. Instructions for accessing this database can be found in the ReadMe file on this book's website.

What's on the Website

Most chapters in this book have supplemental materials that you can download from www.wiley.com/go/data_mining_SQL_2008. As appropriate for the chapter, the site contains SQL Server database backups, SQL Server Analysis Services database backups, project files, DMX query files, and/or source code. Each chapter directory contains a `readme` file that describes how to use the downloads for that chapter.

This book will launch you into the world of SQL Server Data Mining. After you absorb all the information contained within, you will be well on your way to adding predictive and descriptive analytics to your daily life. With its powerful development environment and APIs, Microsoft SQL Server Data Mining can change how you and every user in your organization view and interact with data. Take the leap and discover the hidden sweets locked away in the data you