

Predictive Analytics with Microsoft Azure Machine Learning

Build and Deploy Actionable
Solutions in Minutes



Roger Barga
Valentine Fontama
Wee Hyong Tok

Apress®

Predictive Analytics with Microsoft Azure Machine Learning: Build and Deploy Actionable Solutions in Minutes

Copyright © 2014 by Roger Barga, Valentine Fontama, and Wee Hyong Tok

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.

ISBN-13 (pbk): 978-1-4842-0446-7

ISBN-13 (electronic): 978-1-4842-0445-0

Trademarked names, logos, and images may appear in this book. Rather than use a trademark symbol with every occurrence of a trademarked name, logo, or image we use the names, logos, and images only in an editorial fashion and to the benefit of the trademark owner, with no intention of infringement of the trademark.

The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Managing Director: Welmoed Spahr

Lead Editor: James DeWolf

Development Editor: Douglas Pundick

Technical Reviewers: Jacob Spoelstra and Hang Zhang

Editorial Board: Steve Anglin, Mark Beckner, Gary Cornell, Louise Corrigan, James DeWolf,

Jonathan Gennick, Robert Hutchinson, Michelle Lowman, James Markham,

Matthew Moodie, Jeff Olson, Jeffrey Pepper, Douglas Pundick, Ben Renow-Clarke,

Dominic Shakeshaft, Gwenan Spearing, Matt Wade, Steve Weiss

Coordinating Editor: Kevin Walter

Copy Editor: Mary Behr

Compositor: SPi Global

Indexer: SPi Global

Artist: SPi Global

Cover Designer: Anna Ishchenko

Distributed to the book trade worldwide by Springer Science+Business Media New York, 233 Spring Street, 6th Floor, New York, NY 10013. Phone 1-800-SPRINGER, fax (201) 348-4505, e-mail orders-ny@springer-sbm.com, or visit www.springeronline.com. Apress Media, LLC is a California LLC and the sole member (owner) is Springer Science + Business Media Finance Inc (SSBM Finance Inc). SSBM Finance Inc is a **Delaware** corporation.

For information on translations, please e-mail rights@apress.com, or visit www.apress.com.

Apress and friends of ED books may be purchased in bulk for academic, corporate, or promotional use. eBook versions and licenses are also available for most titles. For more information, reference our Special Bulk Sales-eBook Licensing web page at www.apress.com/bulk-sales.

Any source code or other supplementary materials referenced by the author in this text is available to readers at www.apress.com. For detailed information about how to locate your book's source code, go to www.apress.com/source-code/.

Contents at a Glance

About the Authors	xi
Acknowledgments	xiii
Foreword	xv
Introduction	xix
■ Part 1: Introducing Data Science and Microsoft Azure Machine Learning	1
■ Chapter 1: Introduction to Data Science.....	3
■ Chapter 2: Introducing Microsoft Azure Machine Learning.....	21
■ Chapter 3: Integration with R	43
■ Part 2: Statistical and Machine Learning Algorithms...	65
■ Chapter 4: Introduction to Statistical and Machine Learning Algorithms.....	67
■ Part 3: Practical Applications	85
■ Chapter 5: Building Customer Propensity Models.....	87
■ Chapter 6: Building Churn Models.....	107
■ Chapter 7: Customer Segmentation Models	129
■ Chapter 8: Building Predictive Maintenance Models.....	143
Index	163

Contents

- About the Authors..... xi**
- Acknowledgments xiii**
- Foreword xv**
- Introduction xix**

- Part 1: Introducing Data Science and Microsoft Azure
Machine Learning 1**
- Chapter 1: Introduction to Data Science..... 3**
 - What Is Data Science? 3
 - Analytics Spectrum 4
 - Descriptive Analysis..... 5
 - Diagnostic Analysis..... 5
 - Predictive Analysis..... 5
 - Prescriptive Analysis 6
 - Why Does It Matter and Why Now? 7
 - Data as a Competitive Asset 7
 - Increased Customer Demand 8
 - Increased Awareness of Data Mining Technologies..... 8
 - Access to More Data..... 8
 - Faster and Cheaper Processing Power..... 9
 - The Data Science Process 11

Common Data Science Techniques	14
Classification Algorithms	14
Clustering Algorithms	15
Regression Algorithms.....	16
Simulation	17
Content Analysis.....	17
Recommendation Engines.....	18
Cutting Edge of Data Science.....	18
The Rise of Ensemble Models	18
Summary.....	20
Bibliography	20
■ Chapter 2: Introducing Microsoft Azure Machine Learning.....	21
Hello, Machine Learning Studio!	21
Components of an Experiment.....	22
Five Easy Steps to Creating an Experiment.....	23
Step 1: Get Data.....	24
Step 2: Preprocess Data	26
Step 3: Define Features	29
Step 4: Choose and Apply Machine Learning Algorithms	31
Step 5: Predict Over New Data	33
Deploying Your Model in Production.....	35
Deploying Your Model into Staging.....	36
Testing the Web Service	39
Moving Your Model from Staging into Production	39
Accessing the Azure Machine Learning Web Service.....	40
Summary.....	42

- **Chapter 3: Integration with R** **43**
 - R in a Nutshell 43
 - Building and Deploying Your First R Script..... 45
 - Using R for Data Preprocessing..... 50
 - Using a Script Bundle (Zip)..... 54
 - Building and Deploying a Decision Tree Using R 58
 - Summary..... 64

- **Part 2: Statistical and Machine Learning Algorithms...** **65**
- **Chapter 4: Introduction to Statistical and Machine Learning Algorithms** **67**
 - Regression Algorithms 67
 - Linear Regression..... 68
 - Neural Networks..... 70
 - Decision Trees 72
 - Boosted Decision Trees..... 73
 - Classification Algorithms..... 74
 - Support Vector Machines..... 76
 - Bayes Point Machines 78
 - Clustering Algorithms 79
 - Summary..... 83

- **Part 3: Practical Applications.....** **85**
- **Chapter 5: Building Customer Propensity Models.....** **87**
 - The Business Problem..... 87
 - Data Acquisition and Preparation 88
 - Loading Data from Your Local File System 88
 - Loading Data from Other Sources 89
 - Data Analysis 91

Training the Model.....	99
Model Testing and Validation.....	101
Model Performance.....	102
Summary.....	106
■ Chapter 6: Building Churn Models.....	107
Churn Models in a Nutshell	107
Building and Deploying a Customer Churn Model.....	109
Preparing and Understanding Data	109
Data Preprocessing and Feature Selection	114
Classification Model for Predicting Churn	121
Evaluating the Performance of the Customer Churn Models.....	125
Summary.....	127
■ Chapter 7: Customer Segmentation Models	129
Customer Segmentation Models in a Nutshell	129
Building and Deploying Your First K-Means Clustering Model	130
Feature Hashing	133
Identifying the Right Features	134
Properties of K-Means Clustering.....	135
Customer Segmentation of Wholesale Customers	138
Loading the Data from the UCI Machine Learning Repository	138
Using K-Means Clustering for Wholesale Customer Segmentation.....	139
Cluster Assignment for New Data.....	141
Summary.....	142

■ **Chapter 8: Building Predictive Maintenance Models..... 143**

 Overview 143

 The Business Problem..... 145

 Data Acquisition and Preparation 145

 The Dataset 145

 Data Loading..... 146

 Data Analysis 149

 Training the Model..... 152

 Model Testing and Validation..... 154

 Model Performance 155

 Model Deployment 158

 Publishing Your Model into Staging 158

 Moving Your Model from Staging into Production 160

 Summary..... 161

Index..... 163

About the Authors



Roger Barga is a General Manager and Director of Development at Amazon Web Services. Prior to joining Amazon, Roger was Group Program Manager for the Cloud Machine Learning group in the Cloud & Enterprise division at Microsoft, where his team was responsible for product management of the Azure Machine Learning service. Roger joined Microsoft in 1997 as a Researcher in the Database Group of Microsoft Research, where he directed both systems research and product development efforts in database, workflow, and stream processing systems. He has developed ideas from basic research, through proof of concept prototypes, to incubation efforts in product groups. Prior to joining Microsoft, Roger was a Research Scientist in the Machine Learning Group at the Pacific Northwest National Laboratory where he built and deployed machine learning-based solutions. Roger is also an Affiliate Professor at the University of Washington, where he is a lecturer in the Data Science and Machine Learning programs.

Roger holds a PhD in Computer Science, a M.Sc. in Computer Science with an emphasis on Machine Learning, and a B.Sc. in Mathematics and Computing Science. He has published over 90 peer-reviewed technical papers and book chapters, collaborated with 214 co-authors from 1991 to 2013, with over 700 citations by 1,084 authors.



Valentine Fontama is a Principal Data Scientist in the Data and Decision Sciences Group (DDSG) at Microsoft, where he leads external consulting engagements that deliver world-class Advanced Analytics solutions to Microsoft's customers. Val has over 18 years of experience in data science and business. Following a PhD in Artificial Neural Networks, he applied data mining in the environmental science and credit industries. Before Microsoft, Val was a New Technology Consultant at Equifax in London where he pioneered the application of data mining to risk assessment and marketing in the consumer credit industry. He is currently an Affiliate Professor of Data Science at the University of Washington.

In his prior role at Microsoft, Val was a Senior Product Marketing Manager responsible for big data and predictive analytics in cloud and enterprise marketing. In this role, he led product management for Microsoft Azure Machine Learning; HDInsight, the first

Hadoop service from Microsoft; Parallel Data Warehouse, Microsoft's first data warehouse appliance; and three releases of Fast Track Data Warehouse. He also played a key role in defining Microsoft's strategy and positioning for in-memory computing.

Val holds an M.B.A. in Strategic Management and Marketing from Wharton Business School, a Ph.D. in Neural Networks, a M.Sc. in Computing, and a B.Sc. in Mathematics and Electronics (with First Class Honors). He co-authored the book *Introducing Microsoft Azure HDInsight*, and has published 11 academic papers with 152 citations by over 227 authors.



Wee-Hyong Tok is a Senior Program Manager on the SQL Server team at Microsoft. Wee-Hyong brings over 12 years of database systems experience (with more than six years of data platform experience in industry and six years of academic experience).

Prior to pursuing his PhD, Wee-Hyong was a System Analyst at a large telecommunication company in Singapore, working on marketing decision support systems. Following his PhD in Data Streaming Systems from the National University of Singapore, he joined Microsoft and worked on the SQL Server team. Over the past six years, Wee-Hyong gained extensive experience working with distributed engineering teams from Asia and US, and was responsible for shaping the SSIS Server, bringing it from concept to release in SQL Server 2012. More recently, Wee-Hyong was part of the Azure Data Factory team, a service for orchestrating and managing data transformation and movement.

Wee Hyong holds a Ph.D. in Data Streaming Systems, a M.Sc. in Computing, and a B.Sc. (First Class Honors) in Computer Science, from the National University of Singapore. He has published 21 peer reviewed academic papers and journals. He is a co-author of two books, *Introducing Microsoft Azure HDInsight* and *Microsoft SQL Server 2012 Integration Services*.

Acknowledgments

I would like to express my gratitude to the many people in the CloudML team at Microsoft who saw us through this book; to all those who provided support, read, offered comments, and assisted in the editing, and proofreading. I wish to thank my coauthors, Val and Wee-Hyong, for their drive and perseverance which was key to completing this book, and to our publisher Apress, especially Kevin Walter and James T. DeWolf, for making this all possible. Above all I want to thank my wife, Terre, and my daughters Amelie and Jolie, who supported and encouraged me in spite of all the time it took me away from them.

—Roger Barga

I would like to thank my co-authors, Roger and Wee-Hyong, for their deep collaboration on this project. Special thanks to my wife, Veronica, and loving kids, Engu, Chembe, and Nayah, for their support and encouragement.

—Valentine Fontama

I would like to thank my coauthors, Roger and Val, for working together to shape the content of the book. I deeply appreciate the reviews by the team of data scientists from the CCloudML team. I'd also like to thank the Apress team who worked with us from concept to shipping. And I'd like to thank Juliet, Nathaniel, Siak-Eng, and Hwee-Tiang for their love, support, and patience.

—Wee-Hyong

Foreword

Few people appreciate the enormous potential of machine learning (ML) in enterprise applications. I was lucky enough to get a taste of its potential benefits just a few months into my first job. It was 1995 and credit card issuers were beginning to adopt neural network models to detect credit card fraud in real time. When a credit card is used, transaction data from the point of sale system is sent to the card issuer's credit authorization system where a neural network scores for the probability of fraud. If the probability is high, the transaction is declined in real time. I was a scientist building such models and one of my first model deliveries was for a South American bank. When the model was deployed, the bank identified over a million dollars of previously undetected fraud on the very first day. This was a big eye-opener. In the years since, I have seen ML deliver huge value in diverse applications such as demand forecasting, failure and anomaly detection, ad targeting, online recommendations, and virtual assistants like Cortana. By embedding ML into their enterprise systems, organizations can improve customer experience, reduce the risk of systemic failures, grow revenue, and realize significant cost savings.

However, building ML systems is slow, time-consuming, and error prone. Even though we are able to analyze very large data sets these days and deploy at very high transaction rates, the following bottlenecks remain:

- ML system development requires deep expertise. Even though the core principles of ML are now accessible to a wider audience, talented data scientists are as hard to hire today as they were two decades ago.
- Practitioners are forced to use a variety of tools to collect, clean, merge, and analyze data. These tools have a steep learning curve and are not integrated. Commercial ML software is expensive to deploy and maintain.
- Building and verifying models requires considerable experimentation. Data scientists often find themselves limited by compute and storage because they need to run a large number of experiments that generate considerable new data.
- Software tools do not support scalable experimentation or methods for organizing experiment runs. The act of collaborating with a team on experiments and sharing derived variables, scripts, etc. is manual and ad-hoc without tools support. Evaluating and debugging statistical models remains a challenge.

Data scientists work around these limitations by writing custom programs and by doing undifferentiated heavy lifting as they perform their ML experiments. But it gets harder in the deployment phase. Deploying ML models in a mission-critical business process such as real-time fraud prevention or ad targeting requires sophisticated engineering. The following needs must be met:

- Typically, ML models that have been developed offline now have to be reimplemented in a language such as C++, C#, or Java.
- The transaction data pipelines have to be plumbed. Data transformations and variables used in the offline models have to be recoded and compiled.
- These reimplementations inevitably introduce bugs, requiring verification that the models work as originally designed.
- A custom container for the model has to be built, with appropriate monitors, metrics, and logging.
- Advanced deployments require A/B testing frameworks to evaluate alternative models side-by-side. One needs mechanisms to switch models in or out, preferably without recompiling and deploying the entire application.
- One has to validate that the candidate production model works as originally designed through statistical tests.

The automated decisions made by the system and the business outcomes have to be logged for refining the ML models and for monitoring.

The service has to be designed for high availability, disaster recovery, and geo-proximity to end points.

When the service has to be scaled to meet higher transaction rates and/or low latency, more work is required to provision new hardware, deploy the service to new machines, and scale out.

All of these are time-consuming and engineering-intensive steps, expensive in terms of both infrastructure and manpower. The end-to-end engineering and maintenance of a production ML application requires a highly skilled team that few organizations can build and sustain.

Microsoft Azure ML was designed to solve these problems.

- It's a fully managed cloud service with no software to install, no hardware to manage, no OS versions or development environments to grapple with.
- Armed with nothing but a browser, data scientists can log on to Azure and start developing ML models from any location, from any device. They can host a practically unlimited number of files on Azure storage.

- ML Studio, an integrated development environment for ML, lets you set up experiments as simple data flow graphs, with an easy-to-use drag, drop, and connect paradigm. Data scientists can avoid programming for a large number of common tasks, allowing them to focus on experiment design and iteration.
- Many sample experiments are provided to make it easy to get started.
- A collection of best of breed algorithms developed by Microsoft Research is built in, as is support for custom R code. Over 350 open source R packages can be used securely within Azure ML.
- Data flow graphs can have several parallel paths that automatically run in parallel, allowing scientists to execute complex experiments and make side-by-side comparisons without the usual computational constraints.
- Experiments are readily sharable, so others can pick up on your work and continue where you left off.

Azure ML also makes it simple to create production deployments at scale in the cloud. Pre-trained ML models can be incorporated into a scoring workflow and, with a few clicks, a new cloud-hosted REST API can be created. This REST API has been engineered to respond with low latency. No reimplementing or porting is required—a key benefit over traditional data analytics software. Data from anywhere on the Internet (laptops, web sites, mobile devices, wearables, and connected machines) can be sent to the newly created API to get back predictions. For example, a data scientist can create a fraud detection API that takes transaction information as input and returns a low/medium/high risk indicator as output. Such an API would then be “live” on the cloud, ready to accept calls from any software that a developer chooses to call it from. The API backend scales elastically, so that when transaction rates spike, the Azure ML service can automatically handle the load. There are virtually no limits on the number of ML APIs that a data scientist can create and deploy—and all this without any dependency on engineering. For engineering and IT, it becomes simple to integrate a new ML model using those REST APIs, and testing multiple models side-by-side before deployment becomes easy, allowing dramatically better agility at low cost. Azure provides mechanisms to scale and manage APIs in production, including mechanisms to measure availability, latency, and performance. Building robust, highly available, reliable ML systems and managing the production deployment is therefore dramatically faster, cheaper, and easier for the enterprise, with huge business benefits.

We believe Azure ML is a game changer. It makes the incredible potential of ML accessible both to startups and large enterprises. Startups are now able to use the same capabilities that were previously available to only the most sophisticated businesses. Larger enterprises are able to unleash the latent value in their big data to generate significantly more revenue and efficiencies. Above all, the speed of iteration and experimentation that is now possible will allow for rapid innovation and pave the way for intelligence in cloud-connected devices all around us.

■ FOREWORD

When I started my career in 1995, it took a large organization to build and deploy credit card fraud detection systems. With tools like Azure ML and the power of the cloud, a single talented data scientist can accomplish the same feat. The authors of this book, who have long experience with data science, have designed it to help you get started on this wonderful journey with Azure ML.

—Joseph Sirosh
Corporate Vice President, Machine Learning, Microsoft Corporation.

Introduction

Data science and machine learning are in high demand, as customers are increasingly looking for ways to glean insights from their data. More customers now realize that business intelligence is not enough as the volume, speed, and complexity of data now defy traditional analytics tools. While business intelligence addresses descriptive and diagnostic analysis, data science unlocks new opportunities through predictive and prescriptive analysis.

This book provides an overview of data science and an in-depth view of Microsoft Azure Machine Learning, the latest predictive analytics service from the company. The book provides a structured approach to data science and practical guidance for solving real-world business problems such as buyer propensity modeling, customer churn analysis, predictive maintenance, and product recommendation. The simplicity of this new service from Microsoft will help to take data science and machine learning to a much broader audience than existing products in this space. Learn how you can quickly build and deploy sophisticated predictive models as machine learning web services with the new Azure Machine Learning service from Microsoft.

Who Should Read this Book?

This book is for budding data scientists, business analysts, BI professionals, and developers.

The reader needs to have basic skills in statistics and data analysis. That said, they do not need to be data scientists or have deep data mining skills to benefit from this book.

What You Will Learn

This book will provide the following:

- A deep background in data science, and how to solve a business data science problem using a structured approach and best practices
- How to use Microsoft Azure Machine Learning service to effectively build and deploy predictive models as machine learning web services
- Practical examples that show how to solve typical predictive analytics problems such as propensity modeling, churn analysis, and product recommendation.

At the end of the book, you will have gained essential skills in basic data science, the data mining process, and a clear understanding of the new Microsoft Azure Machine Learning service. You'll also have the frameworks for solving practical business problems with machine learning.