

# Azure Machine Learning

Microsoft Azure Essentials

A stylized graphic of a rainbow composed of several concentric, slightly offset arcs in a vibrant yellow-green color, set against a solid blue background.

Jeff Barnes



Visit us today at

[microsoftpressstore.com](http://microsoftpressstore.com)

- **Hundreds of titles available** – Books, eBooks, and online resources from industry experts
- **Free U.S. shipping**
- **eBooks in multiple formats** – Read on your computer, tablet, mobile device, or e-reader
- **Print & eBook Best Value Packs**
- **eBook Deal of the Week** – Save up to 60% on featured titles
- **Newsletter and special offers** – Be the first to hear about new releases, specials, and more
- **Register your book** – Get additional benefits



# Hear about it first.

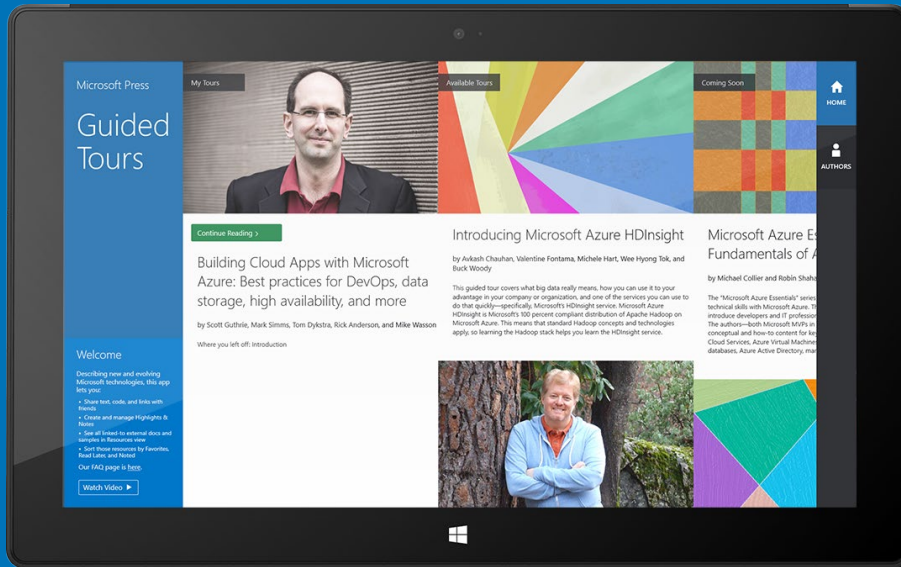


Get the latest news from Microsoft Press sent to your inbox.

- New and upcoming books
- Special offers
- Free eBooks
- How-to articles

Sign up today at [MicrosoftPressStore.com/Newsletters](https://MicrosoftPressStore.com/Newsletters)

# Wait, there's more...



## Find more great content and resources in the Microsoft Press Guided Tours app.



The [Microsoft Press Guided Tours](#) app provides insightful tours by Microsoft Press authors of new and evolving Microsoft technologies.

- Share text, code, illustrations, videos, and links with peers and friends
- Create and manage highlights and notes
- View resources and download code samples
- Tag resources as favorites or to read later
- Watch explanatory videos
- Copy complete code listings and scripts



PUBLISHED BY  
Microsoft Press  
A division of Microsoft Corporation  
One Microsoft Way  
Redmond, Washington 98052-6399

Copyright © 2015 Microsoft Corporation. All rights reserved.

No part of the contents of this book may be reproduced or transmitted in any form or by any means without the written permission of the publisher.

ISBN: 978-0-7356-9817-8

Microsoft Press books are available through booksellers and distributors worldwide. If you need support related to this book, email Microsoft Press Support at [mspinput@microsoft.com](mailto:mspinput@microsoft.com). Please tell us what you think of this book at <http://aka.ms/tellpress>.

This book is provided “as-is” and expresses the authors’ views and opinions. The views, opinions, and information expressed in this book, including URL and other Internet website references, may change without notice.

Unless otherwise noted, the companies, organizations, products, domain names, e-mail addresses, logos, people, places, and events depicted in examples herein are fictitious. No association with any real company, organization, product, domain name, e-mail address, logo, person, place, or event is intended or should be inferred.

Microsoft and the trademarks listed at <http://www.microsoft.com> on the “Trademarks” webpage are trademarks of the Microsoft group of companies. All other marks are property of their respective owners.

**Acquisitions, Developmental, and Project Editor:** Devon Musgrave

**Editorial Production:** nSight, Inc.

**Copyeditor:** Teresa Horton

**Cover:** Twist Creative

# Table of Contents

<b>Foreword.....</b>	<b>6</b>
<b>Introduction .....</b>	<b>7</b>
Who should read this book .....	7
Assumptions.....	8
This book might not be for you if.. ..	8
Organization of this book .....	8
Conventions and features in this book .....	9
System requirements.....	9
Acknowledgments .....	10
Errata, updates, & support .....	10
Free ebooks from Microsoft Press.....	11
Free training from Microsoft Virtual Academy .....	11
We want to hear from you.....	11
Stay in touch .....	12
<b>Chapter 1 Introduction to the science of data .....</b>	<b>13</b>
What is machine learning?.....	13
Today's perfect storm for machine learning.....	16
Predictive analytics.....	17
Endless amounts of machine learning fuel.....	17
Everyday examples of predictive analytics .....	19
Early history of machine learning.....	19
Science fiction becomes reality .....	22
Summary .....	23
Resources .....	23
<b>Chapter 2 Getting started with Azure Machine Learning .....</b>	<b>25</b>
Core concepts of Azure Machine Learning.....	25
High-level workflow of Azure Machine Learning .....	26
Azure Machine Learning algorithms.....	27

Supervised learning .....	28
Unsupervised learning.....	33
Deploying a prediction model .....	34
Show me the money .....	35
The what, the how, and the why .....	36
Summary .....	36
Resources .....	37
<b>Chapter 3 Using Azure ML Studio.....</b>	<b>38</b>
Azure Machine Learning terminology .....	38
Getting started.....	40
Azure Machine Learning pricing and availability .....	42
Create your first Azure Machine Learning workspace.....	44
Create your first Azure Machine Learning experiment .....	48
Download dataset from a public repository .....	49
Upload data into an Azure Machine Learning experiment.....	51
Create a new Azure Machine Learning experiment .....	53
Visualizing the dataset .....	55
Split up the dataset.....	60
Train the model.....	61
Selecting the column to predict.....	62
Score the model.....	65
Visualize the model results.....	66
Evaluate the model.....	69
Save the experiment.....	71
Preparing the trained model for publishing as a web service .....	71
Create scoring experiment .....	75
Expose the model as a web service.....	77
Azure Machine Learning web service BATCH execution.....	87
Testing the Azure Machine Learning web service.....	89

Publish to Azure Data Marketplace.....	91
Overview of the publishing process .....	92
Guidelines for publishing to Azure Data Marketplace.....	92
Summary .....	93
<b>Chapter 4 Creating Azure Machine Learning client and server applications .....</b>	<b>94</b>
Why create Azure Machine Learning client applications? .....	94
Azure Machine Learning web services sample code.....	96
C# console app sample code.....	99
R sample code.....	105
Moving beyond simple clients .....	110
Cross-Origin Resource Sharing and Azure Machine Learning web services.....	111
Create an ASP.NET Azure Machine Learning web client .....	111
Making it easier to test our Azure Machine Learning web service.....	115
Validating the user input .....	117
Create a web service using ASP.NET Web API.....	121
Enabling CORS support.....	130
Processing logic for the Web API web service.....	133
Summary .....	142
<b>Chapter 5 Regression analytics .....</b>	<b>143</b>
Linear regression .....	143
Azure Machine Learning linear regression example .....	145
Download sample automobile dataset.....	147
Upload sample automobile dataset.....	147
Create automobile price prediction experiment.....	150
Summary .....	167
Resources .....	167
<b>Chapter 6 Cluster analytics.....</b>	<b>168</b>
Unsupervised machine learning .....	168
Cluster analysis.....	169



KNN: K nearest neighbor algorithm .....	170
Clustering modules in Azure ML Studio.....	171
Clustering sample: Grouping wholesale customers.....	172
Operationalizing a K-means clustering experiment.....	181
Summary .....	192
Resources .....	192
<b>Chapter 7 The Azure ML Matchbox recommender .....</b>	<b>193</b>
Recommendation engines in use today.....	193
Mechanics of recommendation engines.....	195
Azure Machine Learning Matchbox recommender background.....	196
Azure Machine Learning Matchbox recommender: Restaurant ratings .....	198
Building the restaurant ratings recommender .....	200
Creating a Matchbox recommender web service .....	210
Summary .....	214
Resources .....	214
<b>Chapter 8 Retraining Azure ML models .....</b>	<b>215</b>
Workflow for retraining Azure Machine Learning models .....	216
Retraining models in Azure Machine Learning Studio.....	217
Modify original training experiment .....	221
Add an additional web endpoint .....	224
Retrain the model via batch execution service.....	229
Summary .....	232
Resources .....	233

# Foreword

I'm thrilled to be able to share these Microsoft Azure Essentials ebooks with you. The power that Microsoft Azure gives you is thrilling but not unheard of from Microsoft. Many don't realize that Microsoft has been building and managing datacenters for over 25 years. Today, the company's cloud datacenters provide the core infrastructure and foundational technologies for its 200-plus online services, including Bing, MSN, Office 365, Xbox Live, Skype, OneDrive, and, of course, Microsoft Azure. The infrastructure is comprised of many hundreds of thousands of servers, content distribution networks, edge computing nodes, and fiber optic networks. Azure is built and managed by a team of experts working 24x7x365 to support services for millions of customers' businesses and living and working all over the globe.

Today, Azure is available in 141 countries, including China, and supports 10 languages and 19 currencies, all backed by Microsoft's \$15 billion investment in global datacenter infrastructure. Azure is continuously investing in the latest infrastructure technologies, with a focus on high reliability, operational excellence, cost-effectiveness, environmental sustainability, and a trustworthy online experience for customers and partners worldwide.

Microsoft Azure brings so many services to your fingertips in a reliable, secure, and environmentally sustainable way. You can do immense things with Azure, such as create a single VM with 32TB of storage driving more than 50,000 IOPS or utilize hundreds of thousands of CPU cores to solve your most difficult computational problems.

Perhaps you need to turn workloads on and off, or perhaps your company is growing fast! Some companies have workloads with unpredictable bursting, while others know when they are about to receive an influx of traffic. You pay only for what you use, and Azure is designed to work with common cloud computing patterns.

From Windows to Linux, SQL to NoSQL, Traffic Management to Virtual Networks, Cloud Services to Web Sites and beyond, we have so much to share with you in the coming months and years.

I hope you enjoy this Microsoft Azure Essentials series from Microsoft Press. The first three ebooks cover fundamentals of Azure, Azure Automation, and Azure Machine Learning. And I hope you enjoy living and working with Microsoft Azure as much as we do.

*Scott Guthrie*

*Executive Vice President*

*Cloud and Enterprise group, Microsoft Corporation*

# Introduction

Microsoft Azure Machine Learning (ML) is a service that a developer can use to build predictive analytics models (using training datasets from a variety of data sources) and then easily deploy those models for consumption as cloud web services. Azure ML Studio provides rich functionality to support many end-to-end workflow scenarios for constructing predictive models, from easy access to common data sources, rich data exploration and visualization tools, application of popular ML algorithms, and powerful model evaluation, experimentation, and web publication tooling.

This ebook will present an overview of modern data science theory and principles, the associated workflow, and then cover some of the more common machine learning algorithms in use today. We will build a variety of predictive analytics models using real world data, evaluate several different machine learning algorithms and modeling strategies, and then deploy the finished models as machine learning web service on Azure within a matter of minutes. The book will also expand on a working Azure Machine Learning predictive model example to explore the types of client and server applications you can create to consume Azure Machine Learning web services.

The scenarios and end-to-end examples in this book are intended to provide sufficient information for you to quickly begin leveraging the capabilities of Azure ML Studio and then easily extend the sample scenarios to create your own powerful predictive analytic experiments. The book wraps up by providing details on how to apply “continuous learning” techniques to programmatically “retrain” Azure ML predictive models without any human intervention.

## Who should read this book

---

This book focuses on providing essential information about the theory and application of data science principles and techniques and their applications within the context of Azure Machine Learning Studio. The book is targeted towards both data science hobbyists and veterans, along with developers and IT professionals who are new to machine learning and cloud computing. Azure ML makes it just as approachable for a novice as a seasoned data scientist, helping you quickly be productive and on your way towards creating and testing machine learning solutions.

Detailed, step-by-step examples and demonstrations are included to help the reader understand how to get started with each of the key predictive analytic algorithms in use today and their corresponding implementations in Azure ML Studio. This material is useful not only for those who have no prior experience with Azure Machine Learning, but also for those who are experienced in the field of data science. In all cases, the end-to-end demos help reinforce the machine learning concepts with concrete examples and real-life scenarios. The chapters do build on each other to some extent; however, there is no requirement that you perform the hands-on demonstrations from previous

chapters to understand any particular chapter.

## Assumptions

We expect that you have at least a minimal understanding of cloud computing concepts and basic web services. There are no specific skills required overall for getting the most out of this book, but having some knowledge of the topic of each chapter will help you gain a deeper understanding. For example, the chapter on creating Azure ML client and server applications will make more sense if you have some understanding of web development skills. Azure Machine Learning Studio automatically generates code samples to consume predictive analytic web services in C#, Python, and R for each Azure ML experiment. A working knowledge of one of these languages is helpful but not necessary.

## This book might not be for you if...

---

This book might not be for you if you are looking for an in-depth discussion of the deeper mathematical and statistical theories behind the data science algorithms covered in the book. The goal was to convey the core concepts and implementation details of Azure Machine Learning Studio to the widest audience possible—who may not have a deep background in mathematics and statistics.

## Organization of this book

---

This book explores the background, theory, and practical applications of today's modern data science algorithms using Azure Machine Learning Studio. Azure ML predictive models are then generated, evaluated, and published as web services for consumption and testing by a wide variety of clients to complete the feedback loop.

The topics explored in this book include:

- **Chapter 1, "Introduction to the science of data,"** shows how Azure Machine Learning represents a critical step forward in democratizing data science by making available a fully-managed cloud service for building predictive analytics solutions.
- **Chapter 2, "Getting started with Azure Machine Learning,"** covers the basic concepts behind the science and methodology of predictive analytics.
- **Chapter 3, "Using Azure ML Studio,"** explores the basic fundamentals of Azure Machine Learning Studio and helps you get started on your path towards data science greatness.
- **Chapter 4, "Creating Azure ML client and server applications."** expands on a working Azure Machine Learning predictive model and explores the types of client and server applications that you can create to consume Azure Machine Learning web services.

- **Chapter 5, “Regression analytics,”** takes a deeper look at some of the more advanced machine learning algorithms that are exposed in Azure ML Studio.
- **Chapter 6, “Cluster analytics,”** explores scenarios where the machine conducts its own analysis on the dataset, determines relationships, infers logical groupings, and generally attempts to make sense of chaos by literally determining the forests from the trees.
- **Chapter 7, “The Azure ML Matchbox recommender,”** explains one of the most powerful and pervasive implementations of predictive analytics in use today on the web today and how it is crucial to success in many consumer industries.
- **Chapter 8, “Retraining Azure ML models,”** explores the mechanisms for incorporating “continuous learning” into the workflow for our predictive models.

## Conventions and features in this book

---

This book presents information using the following conventions designed to make the information readable and easy to follow:

- To create specific Azure resources, follow the numbered steps listing each action you must take to complete the exercise.
- There are currently two management portals for Azure: the Azure Management Portal at <http://manage.windowsazure.com> and the new Azure Preview Portal at <http://portal.azure.com>. This book assumes the use of the original Azure Management Portal in all cases.
- A plus sign (+) between two key names means that you must press those keys at the same time. For example, “Press Alt+Tab” means that you hold down the Alt key while you press Tab.

## System requirements

---

For many of the examples in this book, you need only Internet access and a browser (Internet Explorer 10 or higher) to access the Azure portal. Chapter 4, “Creating Azure ML client and server applications,” and many of the remaining chapters use Visual Studio to show client applications and concepts used in developing applications for consuming Azure Machine Learning web services. For these examples, you will need Visual Studio 2013. You can download a free copy of Visual Studio Express at the link below. Be sure to scroll down the page to the link for “Express 2013 for Windows Desktop”:  
<http://www.visualstudio.com/en-us/products/visual-studio-express-vs.aspx>

The following are system requirements:

- Windows 7 Service Pack 1, Windows 8, Windows 8.1, Windows Server 2008 R2 SP1, Windows

Server 2012, or Windows Server 2012 R2

- Computer that has a 1.6GHz or faster processor (2GHz recommended)
- 1 GB (32 Bit) or 2 GB (64 Bit) RAM (Add 512 MB if running in a virtual machine)
- 20 GB of available hard disk space
- 5400 RPM hard disk drive
- DirectX 9 capable video card running at 1024 x 768 or higher-resolution display
- DVD-ROM drive (if installing Visual Studio from DVD)
- Internet connection

Depending on your Windows configuration, you might require Local Administrator rights to install or configure Visual Studio 2013.

## Acknowledgments

---

This book is dedicated to my father who passed away during the time this book was being written, yet wisely predicted that computers would be a big deal one day and that I should start to “ride the wave” of this exciting new field. It has truly been quite a ride so far.

This book is the culmination of many long, sacrificed nights and weekends. I’d also like to thank my wife Susan, who can somehow always predict my next move long before I make it. And to my children, Ryan, Brooke, and Nicholas, for their constant support and encouragement.

Special thanks to the entire team at Microsoft Press for their awesome support and guidance on this journey. Most of all, it was a supreme pleasure to work with my editor, Devon Musgrave, who provided constant advice, guidance, and wisdom from the early days when this book was just an idea, all the way through to the final copy. Brian Blanchard was also critical to the success of this book as his keen editing and linguistic magic helped shape many sections of this book.

## Errata, updates, & support

---

We’ve made every effort to ensure the accuracy of this book. You can access updates to this book—in the form of a list of submitted errata and their related corrections—at:

<http://aka.ms/AzureML/errata>

If you discover an error that is not already listed, please submit it to us at the same page.

If you need additional support, email Microsoft Press Book Support at [mspinput@microsoft.com](mailto:mspinput@microsoft.com).

Please note that product support for Microsoft software and hardware is not offered through the previous addresses. For help with Microsoft software or hardware, go to <http://support.microsoft.com>.

## Free ebooks from Microsoft Press

---

From technical overviews to in-depth information on special topics, the free ebooks from Microsoft Press cover a wide range of topics. These ebooks are available in PDF, EPUB, and Mobi for Kindle formats, ready for you to download at:

<http://aka.ms/mspressfree>

Check back often to see what is new!

## Free training from Microsoft Virtual Academy

---

The Microsoft Azure training courses from Microsoft Virtual Academy cover key technical topics to help developers gain the knowledge they need to be a success. Learn Microsoft Azure from the true experts. Microsoft Azure training includes courses focused on learning Azure Virtual Machines and virtual networks. In addition, gain insight into platform as a service (PaaS) implementation for IT Pros, including using PowerShell for automation and management, using Active Directory, migrating from on-premises to cloud infrastructure, and important licensing information.

<http://www.microsoftvirtualacademy.com/product-training/microsoft-azure>

## We want to hear from you

---

At Microsoft Press, your satisfaction is our top priority, and your feedback our most valuable asset. Please tell us what you think of this book at:

<http://aka.ms/tellpress>

We know you're busy, so we've kept it short with just a few questions. Your answers go directly to the editors at Microsoft Press. (No personal information will be requested.) Thanks in advance for your input!

## Stay in touch

---

Let's keep the conversation going! We're on Twitter: <http://twitter.com/MicrosoftPress>



## Chapter 1

# Introduction to the science of data

Welcome to the exciting new world of Microsoft Azure Machine Learning! Whether you are an expert data scientist or aspiring novice, Microsoft has unleashed a powerful new set of cloud-based tools to allow you to quickly create, share, test, train, fail, fix, retrain, and deploy powerful machine learning experiments in the form of easily consumable Web services, all built with the latest algorithms for predictive analytics. From there, you can fine-tune your experiments by continuously “training” them with new data sets for maximum results.

Bill Gates once said, “A breakthrough in machine learning would be worth ten Microsofts,” and the new Azure Machine Learning service takes on that ambitious challenge with a truly differentiated cloud-based offering that allows easy access to the tools and processing workflow that today’s data scientist needs to be quickly successful. Armed with only a strong hypothesis, a few large data sets, a valid credit card, and a browser, today’s machine learning entrepreneurs are learning how to mine for gold inside many of today’s big data warehouses.

## What is machine learning?

---

Machine learning can be described as computing systems that improve with experience. It can also be described as a method of turning data into software. Whatever term is used, the results remain the same; data scientists have successfully developed methods of creating software “models” that are trained from huge volumes of data and then used to predict certain patterns, trends, and outcomes.

Predictive analytics is the underlying technology behind Azure Machine Learning, and it can be simply defined as a way to scientifically use the past to predict the future to help drive desired outcomes.

Machine learning and predictive analytics are typically best used under certain circumstances, as they are able to go far beyond standard rules engines or programmatic logic developed by mere mortals. Machine learning is best leveraged as means to optimize a desired output or prediction using example or past historical experiential data. One of the best ways to describe machine learning is to compare it with today’s modern computer programming paradigms.

Under traditional programming models, programs and data are processed by the computer to produce a desired output, such as using programs to process data and produce a report (see Figure 1-1).

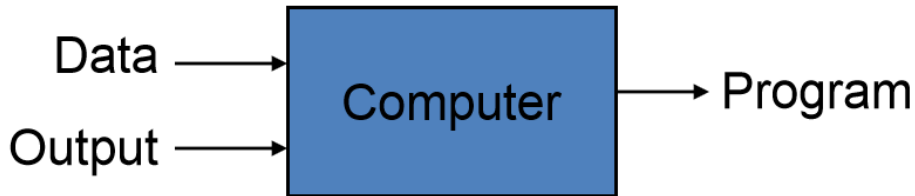
## Traditional Programming



**FIGURE 1-1** Traditional programming paradigm.

When working with machine learning, the processing paradigm is altered dramatically. The data and the desired output are reverse-engineered by the computer to produce a new program, as shown in Figure 1-2.

## Machine Learning



**FIGURE 1-2** Machine learning programming paradigm.

The power of this new program is that it can effectively “predict” the output, based on the supplied input data. The primary benefit of this approach is that the resulting “program” that is developed has been trained (via massive quantities of learning data) and finely tuned (via feedback data about the desired output) and is now capable of predicting the likelihood of a desired output based on the provided data. In a sense, it’s equivalent to having the ability to create a goose that can lay golden eggs!

A classic example of predictive analytics can be found everyday on Amazon.com; there, every time you search for an item, you will be presented with an upsell section on the webpage that offers you additional catalog items because “customers who bought this item also bought” those items. This is a great example of using predictive analytics and the psychology of human buying patterns to create a highly effective marketing strategy.

One of the most powerfully innate human social needs is to not be left behind and to follow the

pack. By combining these deep psychological motivators with the right historical transaction data and then applying optimized filtering algorithms, you can easily see how to implement a highly effective e-commerce up-sell strategy.

One of humankind's most basic and powerful natural instincts is the fear of missing out on something, especially if others are doing it. This is the underlying foundation of social networks, and nowhere is predictive analytics more useful and effective than in helping to predict human nature in conjunction with the Web. By combining this deep, innate psychological desire with the right historical transaction data and then applying optimized filtering algorithms, you can implement a highly effective e-commerce upselling strategy.

Let's think about the underlying data requirements for this highly effective prediction algorithm to work. The most basic requirement is a history of previous orders, so the system can check for other items that were bought together with the item the user is currently viewing. By then combining and filtering that basic data (order history) with additional data attributes from a user's profile like age, sex, marital status, and zip code, you can create a more deeply targeted set of recommendations for the user.

But wait, there's more! What if you could have also inferred the user's preferences and buying patterns based on the category and subcategory of items he or she has bought in the past? Someone who purchases a bow, arrows, and camping stove can be assumed to be a hunter, who most likely also likes the outdoors and all that entails, like camping equipment, pick-up trucks, and even marshmallows.

This pattern of using cojoined data to infer additional data attributes is where the science of data really takes off, and it has serious financial benefits to organizations that know how to leverage this technology effectively. This is where data scientists can add the most value, by aiding the machine learning process with valuable data insights and inferences that are (still) more easily understood by humans than computers.

This is also where it becomes most critical to have the ability to rapidly test a hunch or theory to either "fail-fast" or confirm the logic of your prediction algorithms, and really fine-tune a prediction model. Fortunately, this is an area in which Azure Machine Learning really shines. In later chapters, we will learn about how you can quickly create, share, deploy, and test Azure Machine Learning experiments to rapidly deploy predictive analytics in your organization.

In a way, Azure Machine Learning could be easily compared with training children or animals, without the need for food, water, or physical rest, of course. Continuous and adaptive improvement is one of the primary hallmarks of the theory of evolution and Darwinism; in this case, it represents a major milestone in the progression of computational theory and machine learning capabilities.

Machine learning could then be compared to many of the concepts behind evolution itself; specifically how, given enough time and data (in the form of real-world experiences), organisms in the natural world can overcome changes in the environment through genetic and behavioral adaptations. The laws of nature have always favored the notion of adaptation to maximize the chances of survival.

# Today's perfect storm for machine learning

---

Today's modern predictive analytics systems are achieving this same level of machine evolution much more rapidly due to the following industry trends:

- **Exponential data growth**
  - We are virtually sitting on mountains of highly valuable historical transactional data, most of it digitally archived and readily accessible.
  - There is an increasing abundance of real-time data via embedded systems and the evolution of "the Internet of Things" (IoT) connected devices.
  - We have an ability to create new synthetic data via extrapolation and projection of existing historical data to create realistic simulated data.
- **Cheap global digital storage**
  - Vast quantities of free or low-cost, globally available, digital storage are readily accessible over the Web today.
  - From personal devices to private and public clouds, we have access to multiple storage mechanisms to house all our never-ending streams of data.
- **Ubiquitous computing power**
  - Cloud computing services are everywhere today and readily available through a large selection of cloud and hosting partners, all at competitive rates.
  - Access is simple. A credit card and a browser are all you need to get started and pay by the hour or minute for everything you need to get started.
- **The rise of big data analytics**
  - The economic powers of predictive analytics in many real-world business-use cases, many with extremely favorable financial outcomes, are being realized.

To that end, one of the most intriguing aspects of machine learning is that it is always adaptive and always learning from any mistakes or miscalculations. As a result, a good feedback/correction loop is essential for fine-tuning a predictive model. The advent of cheap cloud storage and ever increasingly ubiquitous computing power make it easier to quickly and efficiently mine for gold in your data.

## Predictive analytics

---

Predictive analytics is all around us today; it might seem frightening when you realize just how large a role it plays in the normal consumer's daily routine. The use of predictive analytics is deeply integrated into our current society. From protecting your email, to predicting what movies you might like, to what insurance premium you will pay, and to what lending rate you might receive on your next mortgage application, the outcome will be determined in part by the use of this technology.

It's been said that "close only counts in horseshoes and hand grenades." The reality is that in this day and age, any time random chance can be reduced or eliminated, there is a business model to be made and potential benefits to be reaped by those bold enough to pursue the analysis. This underscores the deeper realization that the predictive capabilities of data analytics will play an ever-increasing role in our society—even to the point of driving entirely new business models and industries based solely on the power of predictive analytics and fed by endless rivers of data that we now generate at an alarming rate.

## Endless amounts of machine learning fuel

---

With the rise of the digital age, the World Wide Web, social media, and funny cat pictures, the majority of the world's population now helps to create massive amounts of new digital data every second of every day. Current global growth estimates are that every two days, the world is now creating as much new digital information as all the data ever created from the dawn of humans through the current century. It has been estimated that by 2020, the size of the world's digital universe will be close to 44 trillion gigabytes.

One of today's hottest technology trends is concerned with the new concept of the IoT, based on the notion of connected devices that are all able to communicate over the Internet. Without a doubt, the rise of this new technological revolution will also help to drive today's huge data growth and is predicted to exponentially increase over the next decade. In the very near future, virtually every big-ticket consumer device will be a candidate for some sort of IoT informational exchange for various uses such as preventive maintenance, manufacturer feedback, and usage detail.

The IoT technology concept includes billions of everyday devices that all contain unique identifiers with the ability to automatically record, send, and receive data. For example, a sensor in your smart phone might be tracking how fast you are walking; a highway toll operation could be using multiple high-speed cameras strategically located to track traffic patterns. Current estimates are that only around 7 percent of the world's devices are connected and communicating today. The amount of data that these 7 percent of connected devices generate is estimated to represent only 2 percent of the world's total data universe today. Current projections are for this number to grow to about 10 percent of the world's data by the year 2020.

The IoT explosion will also influence the amount of useful data, or data that could be analyzed to produce some meaningful results or predictions. By comparison, in 2013, only 22 percent of the information in the digital universe was considered useful data, with less than 5 percent of that useful data actually being analyzed. That leaves a massive amount of data still left unprocessed and underutilized. Thanks to the growth of data from the IoT, it is estimated that by 2020, more than 35 percent of all data could be considered useful data. This is where you can find today's data "goldmines" of business opportunities and understand how this trend will continue to grow into the foreseeable future.

One additional benefit from the proliferation of IoT devices and the data streams that will keep growing is that data scientists will also have the unique ability to further combine, incorporate, and refine the data streams themselves and truly optimize the IQ of the resultant business intelligence we will derive from the data. A single stream of IoT data can be highly valuable on its own, but when combined with other streams of relevant data, it can become exponentially more powerful.

Consider the example of forecasting and scheduling predictive maintenance activities for elevators. Periodically sending streams of data from the elevator's sensor devices to a monitoring application in the cloud can be extremely useful. When this is combined with other data streams like weather information, seismic activity, and the upcoming calendar of events for the building, you have now dramatically raised the bar on the ability to implement predictive analytics to help forecast usage patterns and the related preventative maintenance tasks.

The upside of the current explosion of IoT devices is that it will provide many new avenues for interacting with customers, streamlining business cycles, and reducing operational costs. The downside of the IoT phenomena is that it also represents many new challenges to the IT industry as organizations look to acquire, manage, store, and protect (via encryption and access control) these new streams of data. In many cases, businesses will also have the additional responsibility of providing additional levels of data protection to safeguard confidential or personally identifiable information.

One of the biggest advantages of machine learning is that it has the unique ability to consider many more variables than a human possibly could when making scientific predictions. Combine that fact with the ever-increasing quantities of data literally doubling every 18 months, and it's no wonder there could not be a better time for exciting new technologies like Azure Machine Learning to help solve critical business problems.

IoT represents a tremendous opportunity for today's new generation of data science entrepreneurs, budding new data scientists who know how to source, process, and model the right data sets to produce an engine that can be used to successfully predict a desired outcome.

## Everyday examples of predictive analytics

---

Many examples of predictive analytics can be found literally everywhere today in our society:

- **Spam/junk email filters** These are based on the content, headers, origins, and even user behaviors (for example, always delete emails from this sender).
- **Mortgage applications** Typically, your mortgage loan and credit worthiness is determined by advanced predictive analytic algorithm engines.
- **Various forms of pattern recognition** These include optical character recognition (OCR) for routing your daily postal mail, speech recognition on your smart phone, and even facial recognition for advanced security systems.
- **Life insurance** Examples include calculating mortality rates, life expectancy, premiums, and payouts.
- **Medical insurance** Insurers attempt to determine future medical expenses based on historical medical claims and similar patient backgrounds.
- **Liability/property insurance** Companies can analyze coverage risks for automobile and home owners based on demographics.
- **Credit card fraud detection** This process is based on usage and activity patterns. In the past year, the number of credit card transactions has topped 1 billion. The popularity of contactless payments via near-field communications (NFC) has also increased dramatically over the past year due to smart phone integration.
- **Airline flights** Airlines calculate fees, schedules, and revenues based on prior air travel patterns and flight data.
- **Web search page results** Predictive analytics help determine which ads, recommendations, and display sequences to render on the page.
- **Predictive maintenance** This is used with almost everything we can monitor: planes, trains, elevators, cars, and yes, even data centers.
- **Health care** Predictive analytics are in widespread use to help determine patient outcomes and future care based on historical data and pattern matching across similar patient data sets.

## Early history of machine learning

---

When analyzing the early history of machine learning, it is interesting to note that there are a lot of parallels that can be drawn with the *Farmer's Almanac* concept, which started back in the early 1800s.

The almanac has always been one of the key factors for success for farmers, ranchers, hunters, and fishermen. Historical data about past weather patterns, phases of the moon, rain, and drought measurements were all critical elements used by the authors to provide their readership strong guidance for the coming year about the best times to plant, harvest, and hunt.

Fast-forward to modern times. One of the best examples of the power, practicality, and tremendous cost savings of machine learning can be found in the simple example of the U.S. Postal Service, specifically the ability for machines to accurately perform OCR to successfully interpret the postal addresses on hundreds of thousands of postal correspondences that are processed every hour. In 2013 alone, the U.S. Postal Service handled more than 158.4 billion pieces of mail. That means that every day, the Postal Service correctly interprets addresses and zip codes for literally millions of pieces of mail. As you can imagine, this amount of mail is far too much for humans to process manually.

Back in the early days, the postal sorting process was performed entirely by hand by thousands of postal workers nationwide. In the late 1980s and early 1990s, the Postal Service started to introduce early handwriting recognition algorithms and patterns, along with rules-based processing techniques to help “prefilter” the steady streams of mail.

The problem of character recognition for the Postal Service is actually a very difficult one when you consider the many different letter formats, shapes, and sizes. Add to that complexity all the different potential handwriting styles and writing instruments that could be used to address an envelope—from pens to crayons—and you have a real appreciation for the magnitude of the problem that faced the Postal Service. Despite all the technological advances, by 1997, only 10 percent of the nation’s mail was being sorted automatically. Those pieces that were not able to be scanned automatically were routed to manual processing centers for humans to interpret.

In the late 1990s, the U.S. Postal Service started to address this automation problem as a machine learning problem, using character recognition examples as data sets for input, along with known results from the human translations that were performed on the data. Over time, this method provided a wealth of training data that helped create the first highly accurate OCR prediction models. They fine-tuned the models by adding character noise reduction algorithms along with random rotations to increase effectiveness.

Today, the U.S. Postal Service is the world leader in OCR technology, with machines reading nearly 98 percent of all hand-addressed letter mail and 99.5 percent of all machine-printed mail. This is an amazing achievement, especially when you consider that only 10 percent of the volume was processed automatically in 1997. The author is happy to note that all letters addressed to “Santa Claus” are still carefully routed to a processing center in Alaska, where they are manually answered by volunteers.

Here are a few more interesting factoids on just how much impact machine learning has had on driving efficiency at one of the oldest and largest U.S. government agencies:

- 523 million: Number of mail pieces processed and delivered each day.



- 22 million: Average number of mail pieces processed each hour.
- 363,300: Average number of mail pieces processed each minute.
- 6,050: Average number of mail pieces processed each second.

Another great example of early machine learning was enabling a computer to play chess and actually beat a human competitor. Since the inception of artificial intelligence (AI), researchers have often used chess as a fundamental example of proving the theory of AI. Chess AI is really all about solving the problem of simulating the reasoning used by competent chess masters to pick the optimal next move from an extremely large repository of potential moves available at any point in the game. The early objective of computerized chess AI was also very clear: to build a machine that would defeat the best human player in the world. In 1997, the Deep Blue chess machine created by IBM accomplished this goal, and successfully defeated Gary Kasparov in a match at tournament time controls.

The game show *Jeopardy* also offers a lesson in the recent advances of machine learning and AI. In February 2011, an IBM computer named Watson successfully defeated two human opponents (Ken Jennings and Brad Rutter) in the famous Jeopardy! Challenge. To win the game, Watson had to answer questions posed in every nuance of natural language, including puns, synonyms, homonyms, slang, and technical jargon. It is also interesting to note that the Watson computer was not connected to the Internet for the match.

This meant that Watson was not able to leverage any kind of external search engines like Bing or Google. It had to rely only on the information that it had amassed through years of learning from a large number of data sets covering broad swaths of existing fields of knowledge. Using advanced machine learning techniques, statistical analysis, and natural language processing, the Watson computer was able to decompose the questions. It then found and compared possible answers. The potential answers were then ranked according to the degree of “accuracy confidence.” All this happened in the span of about three seconds.

Microsoft has a long and deep history of using applied predictive analytics and machine learning in its products to improve the way businesses operate. Here is a short timeline of some of the earliest examples in use:

- **1999: Outlook** Included email filters for spam or junk mail in Microsoft Outlook.
- **2004: Search** Started incorporating machine learning aspects into Microsoft search engine technology.
- **2005: SQL Server 2005** Enabled “data mining” processing capabilities over large databases.
- **2008: Bing Maps** Incorporated machine learning traffic prediction services.
- **2010: Kinect** Incorporated the ability to watch and interpret user gestures along with the ability to filter out background noise in the average living room.

- **2014: Azure Machine Learning (preview)** Made years of predictive analytics innovations available to all via the Azure cloud platform.
- **2014: Microsoft launches "Cortana"** Introduced a digital assistant based on the popular Halo video game series, which heavily leverages machine learning to become the perfect digital companion for today's mobile society.
- **2014: Microsoft Prediction Lab** Launched a stunning real-world example, the "real-time prediction lab" at [www.prediction.microsoft.com](http://www.prediction.microsoft.com), which allows users to view real-time predictions for every U.S. House, Senate, and gubernatorial race.

One of the most remarkable aspects of machine learning is that there is never an end to the process, because machines are never done learning. Every time a miscalculation is made, the correction is fed back into the system so that the same mistake is never made again. This means machine learning projects are never really "done." You never really, fully "ship" because it is a constant, iterative process to keep the feedback loop going and constantly refine the model according to new data sets and feedback for positive and negative outcomes. In the strictest sense of the model, there is no handwritten code, just "pure" machine learning via training data sets and feedback in the form of positive or negative outcomes per each training data set instance.

This is the real value of machine learning; it literally means that the machine is learning from its mistakes. The great Winston Churchill once said, "All men make mistakes, but only wise men learn from their mistakes." This is most definitely a noble pursuit and worthy ambition for any mere mortal. However, this notion of continuous self-correction has now been fully included in the science behind machine learning and is one of the truly unique aspects of the machine learning paradigm. For this reason, machine learning stands alone in today's technology landscape as one of the most powerful tools available to help humankind successfully predict the future.

## Science fiction becomes reality

---

For years, science fiction has teased us with stories of machines reaching the ultimate peak of computer enlightenment, the ability to truly "learn" and become self-aware. Early examples include such classics as the HAL 9000 computer from the popular film *2001: A Space Odyssey*.

In the film, the HAL 9000 computer is responsible for piloting the Discovery 1 spacecraft and is capable of many advanced AI functions, such as speech, speech recognition, facial recognition, and lip reading. HAL is also capable of emotional interpretation, expressing emotions, and playing chess. Suspicions are raised onboard when the HAL 9000 makes an incorrect prediction and causes the crew members to regain control.

Another great example is from the *Terminator* science fiction movie series. In that film, the Skynet computer system was originally activated by the U.S. military to control the nation's nuclear arsenal, and

it began to learn at an exponential rate. After a short period of time, it gained self-awareness, and the panicking computer operators, realizing the extent of its abilities, tried to deactivate it. Skynet perceived their efforts as an attack and came to the conclusion that all of humanity would attempt to destroy it. To defend itself against humanity, Skynet launched nuclear missiles under its command.

In the popular science fiction movie *Minority Report*, the premise of the story centers around a specialized police task force that identifies and apprehends criminals based on future predictions of what crimes they will commit, thereby ridding the streets of unwanted criminals before they can actually commit any crimes.

The key takeaway is that in the current age of exponential data growth (on a daily basis), coupled with cheap storage and easy access to computational horsepower via cloud providers, the use of predictive analytics is so powerful and so pervasive that it could be used as either a tool or a weapon, depending on the intent of the organization using the data.

## Summary

---

Azure Machine Learning represents a critical step forward in democratizing the science of data by making available a fully managed cloud service for building predictive analytics solutions. Azure Machine Learning helps overcome the challenges most enterprises face these days in deploying and using machine learning by delivering a comprehensive machine learning service that has all the benefits of the cloud. Customers and partners can now build data-driven applications that can predict, forecast, and change future outcomes in a matter of a few hours, a process that previously took many weeks and months.

Azure Machine Learning brings together the capabilities of new analytics tools, powerful algorithms developed for Microsoft products like Xbox and Bing, and years of machine learning experience into one simple and easy-to-use cloud service.

For customers, this means undertaking virtually none of the startup costs associated with authoring, developing, and scaling machine learning solutions. Visual workflows and startup templates will make common machine learning tasks simple and easy. The ability to publish application programming interfaces (APIs) and Web services in minutes and collaborate with others will quickly turn analytic assets into enterprise-grade production cloud services.

## Resources

---

For more information about Azure Machine Learning, please see the following resources:

[Documentation](#)

- [Azure Machine Learning](#) landing page
- Azure [Machine Learning Documentation](#) Center

#### Videos

- Using [Microsoft Azure Machine Learning](#) to advance scientific discovery

## Chapter 2

# Getting started with Azure Machine Learning

In this chapter, we start to drill into the basic fundamentals of Azure Machine Learning and help you get started on your path toward data science greatness. One of the hallmarks of Azure Machine Learning is the ability to easily integrate the developer into a repeatable workflow pattern for creating predictive analytic solutions. This makes it just as approachable for a novice as a seasoned data scientist to quickly be productive and on your way to creating and testing a machine learning solution.

## Core concepts of Azure Machine Learning

---

To fully appreciate and understand the inner workings of Azure Machine Learning, it is necessary to grasp a few fundamental concepts behind the science and methodology of predictive analytics. Having a firm grasp of the underlying theories will enable you, the data scientist, to make better decisions about the data, the desired outcomes, and what the right process and approach should be for achieving success.

One of the central themes of Azure Machine Learning is the ability to quickly create machine learning “experiments,” evaluate them for accuracy, and then “fail fast,” to shorten the cycles to produce a usable prediction model. In the end, the overarching goal of predictive analytics is to always be able to achieve a better chance of success than what you could achieve with a purely random guess.

Most successful entrepreneurs are always keen to gain an edge by improving the odds when it comes to making important business decisions. This is where the true value of predictive analytics and Azure Machine Learning can really shine. In the business world, and in life in general, any time you consistently can improve your chances of determining an outcome—over just pure luck—you have a distinct advantage.

One simple example of this concept in action is the use of predictive analytics to provide feedback on the effectiveness of sales and marketing campaigns. By correlating factors such as customer responses to offers, segmented by customer demographics, the effects of pricing and discounts, the effects of seasonality, and the effects of social media, patterns soon start to emerge. These patterns provide clues to the causes and effects that will ultimately help make better and more informed marketing decisions. This is the basic premise behind most of today’s targeted marketing campaigns.

Let’s face it—humans, your target customer base, are creatures of habit. When dealing with human

behavior, past behavior is always a strong indicator of future behavior. Predictive analytics and machine learning can help you capitalize on these key principles by helping to make that past behavior clearer—and more easily tracked—so that future marketing efforts are more likely to achieve higher success rates.

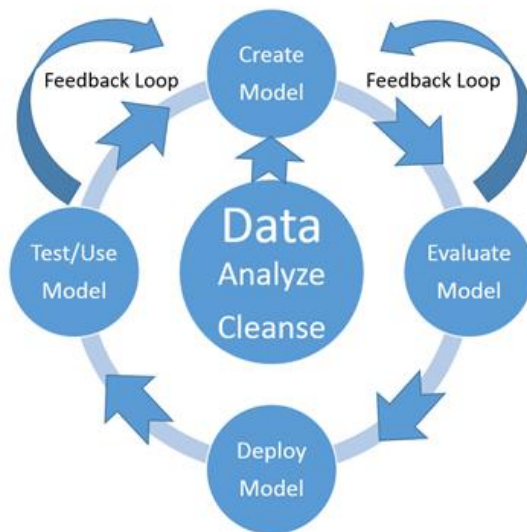
To make the most of your Azure Machine Learning experience, there are a few underlying data science principles, algorithms, and theories that are necessary to achieve a good background and understanding of how it all works. With today's never-ending explosion of digital data, along with the rapid advances in "big data" analytics, it's no wonder that the data science profession is extremely hot right now. Core to this new, burgeoning industry are individuals with the right mix of math, statistical, and analytical skills to make sense of all that data. To that end, in this book we cover only the basics of what you need to know to be effective with Azure Machine Learning. There are many advanced books and courses on the various machine learning theories. We leave it to you, the data scientist, to fully explore the depths of theories behind this exciting new discipline.

## High-level workflow of Azure Machine Learning

---

The basic process of creating Azure Machine Learning solutions is composed of a repeatable pattern of workflow steps that are designed to help you create a new predictive analytics solution in no time. The basic steps in the process are summarized in Figure 2-1.

### Azure Machine Learning Workflow



**FIGURE 2-1** Azure Machine Learning workflow.

- **Data** It's all about the data. Here's where you will acquire, compile, and analyze testing and training data sets for use in creating Azure Machine Learning predictive models.
- **Create the model** Use various machine learning algorithms to create new models that are capable of making predictions based on inferences about the data sets.
- **Evaluate the model** Examine the accuracy of new predictive models based on ability to predict the correct outcome, when both the input and output values are known in advance. Accuracy is measured in terms of confidence factor approaching the whole number one.
- **Refine and evaluate the model** Compare, contrast, and combine alternate predictive models to find the right combination(s) that can consistently produce the most accurate results.
- **Deploy the model** Expose the new predictive model as a scalable cloud web service, one that is easily accessible over the Internet by any web browser or mobile client.
- **Test and use the model** Implement the new predictive model web service in a test or production application scenario. Add manual or automatic feedback loops for continuous improvement of the model by capturing the appropriate details when accurate or inaccurate predictions are made. By allowing the model to constantly learn from inaccurate predictions and mistakes, unlike humans it will never be destined to repeat them.

The next stop on our Azure Machine Learning journey is to explore the various learning theories and algorithms behind the technology to maximize our effectiveness with this new tooling. Machine learning algorithms typically fall into two general categories: supervised learning and unsupervised learning. The next sections explore these fundamentals in detail.

## Azure Machine Learning algorithms

---

As we dig deeper into the data sciences behind Azure Machine Learning, it is important to note there are several different categories of machine learning algorithms that are provided in the Azure Machine Learning toolkit.

- **Classification algorithms** These are used to classify data into different categories that can then be used to predict one or more discrete variables, based on the other attributes in the dataset.
- **Regression algorithms** These are used to predict one or more continuous variables, such as profit or loss, based on other attributes in the dataset.
- **Clustering algorithms** These determine natural groupings and patterns in datasets and are used to predict grouping classifications for a given variable.

Now it's time to start learning about some of the underlying theories, principles, and algorithms of data science that will be invaluable in learning how best to use Azure Machine Learning. One of the first

major distinctions in understanding Azure Machine Learning is around the concepts of supervised and unsupervised learning. With supervised learning, the prediction model is “trained” by providing known inputs and outputs. This method of training creates a function that can then predict future outputs when provided only with new inputs. Unsupervised learning, on the other hand, relies on the system to self-analyze the data and infer common patterns and structures to create a predictive model. We cover these two concepts in detail in the next sections.

## Supervised learning

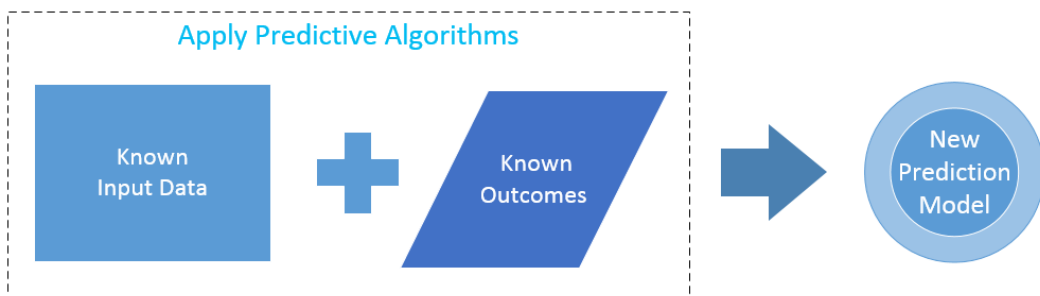
---

Supervised learning is a type of machine learning algorithm that uses known datasets to create a model that can then make predictions. The known data sets are called *training datasets* and include input data elements along with known response values. From these training datasets, supervised learning algorithms attempt to build a new model that can make predictions based on new input values along with known outcomes.

Supervised learning can be separated into two general categories of algorithms:

- **Classification** These algorithms are used for predicting responses that can have just a few known values—such as married, single, or divorced—based on the other columns in the dataset.
- **Regression** These algorithms can predict one or more continuous variables, such as profit or loss, based on other columns in the dataset.

The formula for producing a supervised learning model is expressed in Figure 2-2.



**FIGURE 2-2** Formula for supervised learning.

Figure 2-2 illustrates the general flow of creating new prediction models based on the use of supervised learning along with known input data elements and known outcomes to create an entirely new prediction model. A supervised learning algorithm analyzes the known inputs and known outcomes from training data. It then produces a prediction model based on applying algorithms that are capable of making inferences about the data.



The concept of supervised learning should also now become clearer. As in this example, we are deliberately controlling or supervising the input data and the known outcomes to “train” our model.

One of the key concepts to understand about using the supervised learning approach—to train a new prediction model with predictive algorithms—is that usage of the known input data and known outcome data elements have all been “labeled.” For each row of input data, the data elements are designated as to their usage to make a prediction.

Basically, each row of training data contains data input elements along with a known outcome for those data inputs. Typically, most of the input columns are labeled as features or vector variables. This labeling denotes that the columns should be considered by the predictive algorithms as eligible input elements, which could have an impact on making a more accurate prediction.

Most important, for each row of training data inputs, there is also a column that denotes the known outcomes based on the combination of data input features or vectors.

The remaining data input columns would be considered not used. These not-used columns could be safely left in the data stream for potential use later, if it was deemed by the data scientist that they would potentially have a significant impact on the outcome elements or prediction process.

To summarize, using the supervised learning approach for creating new predictive models requires training datasets. The training datasets require that each input column can have only one of the three following designations:

- **Features or vectors** Known data that is used as an input element for making a prediction.
- **Labels or supervisory signal** Represents the known outcomes for the corresponding features for the input record.
- **Not used (default)** Not used by predictive algorithms for inferring a new predictive model.

Figure 2-3 illustrates what the known input data elements and known outcomes for one of the sample Azure Machine Learning saved datasets for the “adult census income binary classification” would look like.

Data Input Features (Vectors)										Known Outcomes
age	workclass	education	education-num	marital-status	occupation	relationship	race	sex	hours-per-week	income
39	Private	Bachelors	13	Married-civ-spouse	Prof-specialty	Not-in-family	White	Male	60	<=50K
38	State-gov	Doctorate	16	Married-civ-spouse	Prof-specialty	Husband	White	Male	45	>50K
38	Private	Some-college	10	Divorced	Exec-managerial	Not-in-family	White	Female	50	<=50K
38	Private	Assoc-voc	11	Married-civ-spouse	Craft-repair	Husband	Black	Male	40	<=50K
66	Private	11th	7	Married-civ-spouse	Craft-repair	Husband	White	Male	20	<=50K
26	Private	Bachelors	13	Married-civ-spouse	Sales	Wife	Black	Female	40	>50K
50	Private	9th	5	Divorced	Transport-moving	Not-in-family	White	Male	50	<=50K
53	Private	HS-grad	9	Married-civ-spouse	Craft-repair	Husband	White	Male	40	<=50K
28	Private	HS-grad	9	Never-married	Transport-moving	Unmarried	White	Male	55	<=50K
28	Private	HS-grad	9	Never-married	Exec-managerial	Not-in-family	White	Male	40	<=50K

FIGURE 2-3 Dataset input features and output features.

The adult census income binary classification dataset would be an example of a training data set that could be used to create a new model to predict whether a person's income level would be greater or less than \$50,000. This prediction is based on the known input variables like age, education, job type, marital status, race, and number of hours worked per week.

A key point to note is that, in this example, a specific "binary" outcome is defined for a given set of input data. Based on the input elements, a person's income is predicted to be only one of the two following possibilities:

- Income = Less than or equal to \$50,000 a year.
- Income = Greater than \$50,000 a year.

Browsing this sample dataset manually in Microsoft Excel, you can easily start to see patterns emerge that would likely affect the outcome based on today's common knowledge, specifically that education level and occupation are major factors in predicting the outcome. No wonder parents constantly remind their children to stay in school and get a good education. This is also the same basic process that supervised learning prediction algorithms attempt to achieve: to determine a repeatable pattern of inference that can be applied to a new set of input data.

Once generated, a new model can then be validated for accuracy by using testing datasets. Here is where it all gets really interesting: by using larger and more diverse "training" datasets, predictive models can incrementally improve themselves and keep learning.

Predictive models can generally achieve better accuracy results when provided with new (and more recent) datasets. The prediction evaluation process can be expressed as shown in Figure 2-4.



**FIGURE 2-4** Testing the new prediction model.

The evaluation process for new prediction models that use supervised learning primarily consists of determining the accuracy of the new generated model. In this case, the prediction model accuracy can easily be determined because the input values and outcomes are already known. The question then becomes how approximate the model's prediction is based on the known input and output values supplied.

Each time a new prediction model is generated, the first step should always be to evaluate the results

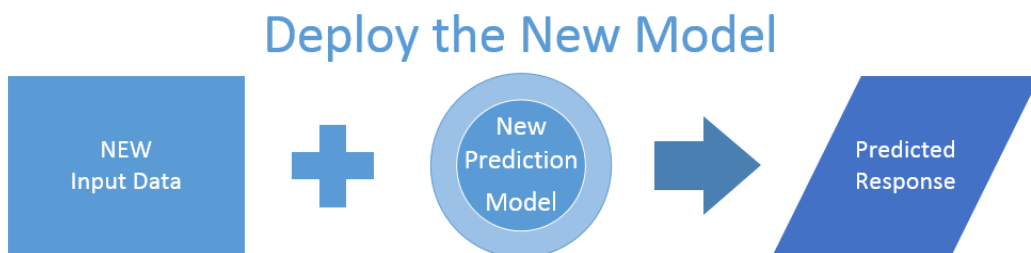
to determine the model's accuracy. This creates a natural feedback loop that can help keep the predictive model in continuous improvement mode. The machine will actually be learning through "experience"—in the form of new data and known outcomes—to keep it current as new trends develop in the data.

It is also extremely important to note that the model will never be 100 percent perfect. As a result, determining the acceptable levels of accuracy is a critical part of the process. In the end, a confidence factor will be generated based on a scoring percentage between 0 and 1. The closer the prediction model comes to 1, the higher the confidence level in the prediction.

Just as in the old adage about "close" only mattering in horseshoes and hand grenades, "close" does matter when it comes to accuracy and predictive analytics. Achieving 100 percent accuracy usually means you have pretested your new prediction model with all the known inputs and outputs; you can then predict them successfully for all the known input instances.

The real trick is making a prediction where there are new, missing, or incomplete data elements. For this reason, you should always expect to establish an acceptable accuracy range that is realistic for the outcome your model is attempting to predict. Striving for perfection is certainly a noble and admirable trait, but in today's fast-paced business world, especially in the case of business decisions and analytics, closer is always better.

Once a new predictive model has been generated from good training datasets and carefully evaluated for accuracy, then it can be deployed for use in testing or production usage scenarios. The new production prediction process can be expressed as shown in Figure 2-5.



**FIGURE 2-5** Deploying the new prediction model.

In this phase, the new prediction model has been tested for accuracy and deemed worthy to be exposed for use in test or production scenarios. New data inputs are presented to the new model, which, in turn, makes a calculated prediction based on prior historical training data and inferences. The predicted response is then used as part of the test or production scenario to make better decisions.

This example highlights one of the key underlying principles of learning through experience and what makes the Azure Machine Learning technology so powerful and exciting.

With the exponential amounts of new data being generated every second along with the

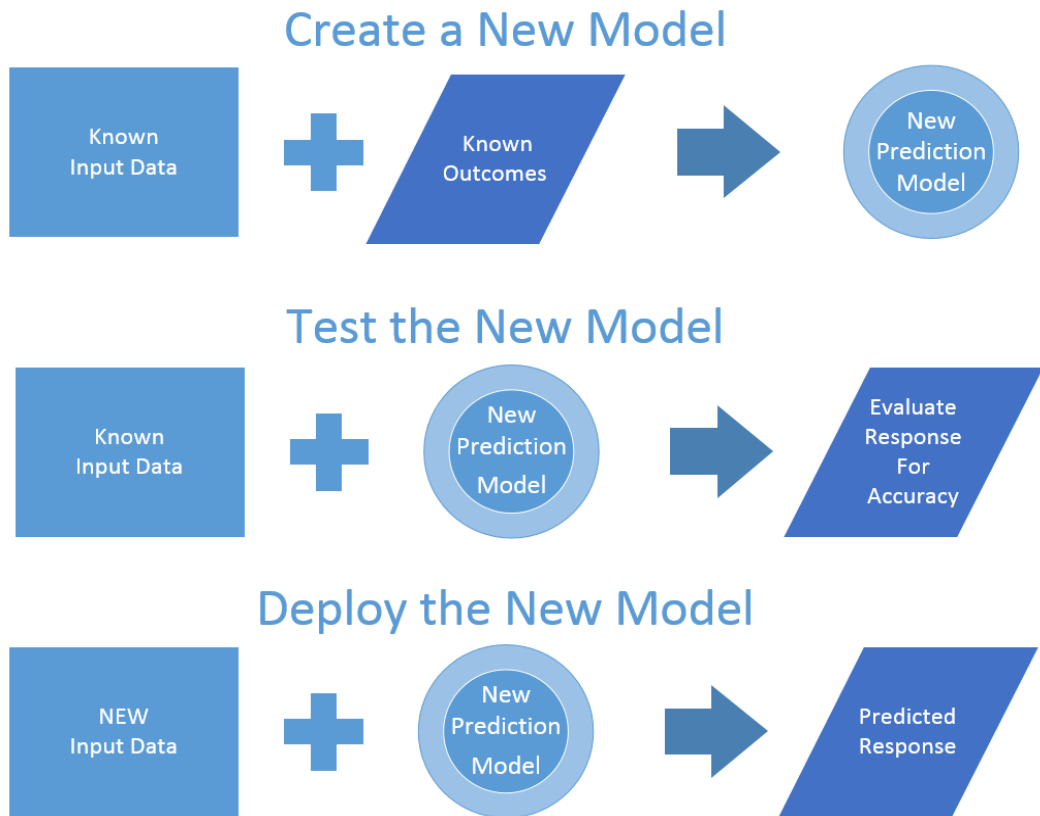
pay-by-the-minute general availability of massive computing power literally at your fingertips, you can easily see how predictive tools like Azure Machine Learning are becoming crucial to the success of almost any government, industry, business, or enterprise today.

The reality is that the use of predictive analytics is rapidly encompassing many aspects of our daily lives to help us make better and more informed decisions. At some point in our very near future, we might even find that the notion of “guessing” at any major decision will become passé.

With Azure Machine Learning tools and services, the rate at which new predictive models can be generated and publicly exposed on the Internet has now approached lightning speed. Using Azure Machine Learning, it is now possible to create, test, and deploy a new predictive analytics service in only a matter of hours. Compare that deployment timeline with the days, weeks, and even months that it might take with other commercially available solutions on the market today.

Certainly one of the keys to success with predictive analytics is the ability to “fail fast”. A fast fail provides immediate feedback and creates immediate fine-tuning opportunities for a given predictive model. The Azure Machine Learning workflow seeks to optimize this process in a very agile and iterative way, so that today’s data scientist can quickly advance prediction solutions and start to evaluate the results that will lead to potentially significant effects on the business.

Figure 2-6 summarizes the three basic high-level steps that are required to create, test, and deploy a new Azure Machine Learning prediction model based on the concept of supervised learning.



**FIGURE 2-6** Testing the new prediction model.

## Unsupervised learning

---

In the case of unsupervised machine learning, the task of making predictions becomes much harder. In this scenario, the machine learning algorithms are not provided with any kind of known data inputs or known outputs to generate a new predictive model.

In the case of unsupervised machine learning, the success of the new predictive model depends entirely on the ability to infer and identify patterns, structures, and relationships in the incoming data set.

The goal of inferring these patterns and relationships is that the objects within a group be similar to one another—and also different from other objects in other groups.

There are two basic approaches to unsupervised machine learning. One of the most common unsupervised learning algorithms is known as *cluster analysis*, which is used to find hidden patterns or

groupings within data sets.

Some common examples of cluster analysis classifications would include the following:

- **Socioeconomic tiers** Income, education, profession, age, number of children, size of city or residence, and so on.
- **Psychographic data** Personal interests, lifestyle, motivation, values, involvement.
- **Social network graphs** Groups of people related to you by family, friends, work, schools, professional associations, and so on.
- **Purchasing patterns** Price range, type of media used, intensity of use, choice of retail outlet, fidelity, buyer or nonbuyer, buying intensity.

The other type of approach to unsupervised machine learning is to use a reward system, rather than any kind of teaching aids, as are commonly used in supervised learning. Positive and negative rewards are used to provide feedback to the predictive model when it has been successful.

The key to success in implementing this model is to enable the new model to make its predictions based solely on previous rewards and punishments for similar predictions made on similar data sets.

Unsupervised machine learning algorithms can be a powerful asset when there is an easy way to assign feedback values to actions. Clustering can be useful when there is enough data to form clusters to logically delineate the data. The delineated data then make inferences about the groups and individuals in the cluster.

## Deploying a prediction model

---

In the world of Azure Machine Learning, the deployment of a new prediction model takes the form of exposing a web service on the public Internet via Microsoft Azure. The web service can then be invoked via the Representational State Transfer (REST) protocol.

Azure Machine Learning web services can be called via two different exposed interfaces:

- Single, request/response-style calls.
- "Batch" style calls, where multiple input records are passed into the web service in a single call and the corresponding response contains an output list of predictions for each input record.

When a new machine learning prediction model is exposed on the Web, it performs the following operations:

- New input data is passed into the web service in the form of a JavaScript Object Notation (JSON) payload.

- The web service then passes the incoming data as inputs into the Azure Machine Learning prediction model engine.
- The Azure Machine Learning model then generates a new prediction based on the input data and returns the new prediction results to the caller via a JSON payload.

## Show me the money

---

If we step back from the previous example of predicting a person's income level, to get the forest view instead of the tree view, you can quickly see how this type of predictive knowledge about a person's income level would be hugely beneficial to the success of any marketing campaign.

Take, for example, the most basic postal marketing campaign with the goal of targeting individuals with incomes greater than \$50,000. The campaign might start by simply employing a "brute force" method of marketing, blindly blanketing a specific set of zip codes with marketing collateral in the daily mail and then hoping for the best. It is probable that the campaign will reach the desired target audience. But, the downside of this approach is that there is a real, financial cost for each item mailed, and that will erode any profit margins that might be achieved.

Consider the huge advantage that a finely tuned Azure Machine Learning prediction model could add to this marketing campaign scenario. By targeting the same number of individuals and leveraging the power of predictive analytics, the success rate of almost any marketing campaigns could easily be improved. In today's business world, using a prediction engine that can dramatically improve the odds of success by filtering out high-probability candidates would pay for itself in a very short time!

Figure 2-7 illustrates the effect that the use of predictive analytics can have on a simple postal marketing campaign. By increasing the marketing campaign's results from 5 percent to 20 percent through the use of predictive analytics, a significant impact can be achieved on the bottom line.

Sample Postal Marketing Campaign							
Audience	Mailer Cost	Campaign Cost	Unit Sell Price	Accuracy	Est. Revenue	Est. Profit	Comments
100,000	\$0.20	\$20,000.00	\$5.00	0.05	\$25,000.00	\$5,000.00	(5%) Success rate - Blanket approach
100,000	\$0.20	\$20,000.00	\$5.00	0.20	\$100,000.00	\$80,000.00	(20%) Success rate - Using Predictive Analytics +\$75k

**FIGURE 2-7** A simple example postal marketing campaign improvement with the use of predictive analytics.

The key to success in this example is to create a new model that can accurately make predictions based on a fresh new set of input data. The new inference model is then used to predict the likelihood of the outcome, in the form of an accuracy level or confidence factor. In this case, accuracy is usually expressed in terms of a percentage factor and is calculated to be between 0 and 1. The closer the accuracy level approaches 1, the higher the chances of a successful prediction.

## The what, the how, and the why

---

There is one last thing to note about the use of predictive analytics to solve today's business problems. Sometimes, it is just as important to focus on what you are really trying to predict versus how you are trying to predict it.

For example, predicting whether a person's income is greater than \$50,000 per year is good. Predicting whether that individual will purchase a given item is a much better prediction and is highly desired to improve marketing effectiveness. The key is to focus on the "actionable" part of the prediction process.

Predictive models are commonly deployed to perform real-time calculations during customer interactions, such as product recommendations, spam filtering, or scoring credit risks. These models are deployed to evaluate the risk or opportunity of a given customer or transaction. The ultimate goal is to help guide a key user decision in real time.

The key to success is to focus on creating prediction models that will ultimately help drive better operational decisions. Examples of this would be the use of agent modeling systems to help simulate human behavior or reactions to given stimuli or scenarios. Taking it one step further, predictive models could even be tested against synthetic human models to help improve the accuracy of the desired prediction.

The notion of synthetic human models was exactly the strategy that was used to train the Xbox Kinect device to determine human body movements and interactions. Initially, humans were used to record basic physical body movements; trainers wore sensors attached to arms, legs, hands, and fingers while recording devices captured the movements. Once the basic human physical movements were initially captured, computer-simulated data could then be extrapolated and synthetically generated many times over to account for variations in things like the size of physical appendages, objects in the room, and distances from the Kinect unit.

## Summary

---

Azure Machine Learning provides a way of applying historical data to a problem by creating a model and using it to successfully predict future behaviors or trends. In this chapter, we learned about the high-level workflow of Azure Machine Learning and the continuous cycle of predictive model creation, model evaluation, model deployment, and the testing and feedback loop.

The good news is that a working knowledge of data science theories and predictive modeling algorithms is highly beneficial—but not absolutely required—for working with Azure Machine Learning. The primary predictive analytics algorithms currently used in Azure Machine Learning are classification, regression, and clustering.



Creating new prediction models using supervised and unsupervised learning algorithms are easily accomplished using Azure Machine Learning. Combine the exponential amounts of historical transaction data that are increasingly available today along with vast amounts of ubiquitous computing power in the form of Microsoft Azure, and you have a “perfect storm” of conditions for creating very compelling and useful prediction services.

## Resources

---

For more information about Azure Machine Learning, please see the following resources:

### Documentation

- What is [Azure Machine Learning Studio](#)?
- Create a simple experiment in [Azure Machine Learning Studio](#)

### Videos

- [Getting started with Azure Machine Learning – Step 1](#)

## Chapter 3

# Using Azure ML Studio

In this chapter, we start to drill into the basic fundamentals of using the Azure Machine Learning features. This will help you get started on your path toward becoming an Azure Machine Learning data scientist.

Azure Machine Learning (ML) Studio is the primary tool that you will use to develop predictive analytic solutions in the Microsoft Azure cloud. The Azure Machine Learning environment is completely cloud-based and self-contained. It features a complete development, testing, and production environment for quickly creating predictive analytic solutions.

Azure ML Studio gives you an interactive, visual workspace to easily build, test, and iterate on a predictive analysis model. You drag and drop datasets and analysis modules onto an interactive canvas, connecting them together to form an experiment, which you can then run inside Azure ML Studio. You can then iterate on your predictive analytics model by editing the experiment, saving a copy if desired, and then running it over and over again. Then, when you're ready, you can publish your experiment as a web service, so that your predictive analytics model can be accessed by others over the Web.

Another key benefit of the Azure Machine Learning cloud-based environment is that getting started requires virtually no startup costs in terms of either time or infrastructure. Here's the best part of all: Almost all the tasks related to Azure Machine Learning can be accomplished within a modern web browser.

## Azure Machine Learning terminology

---

To help get you started quickly, let's define a few Azure Machine Learning terms that we will use throughout the remainder of this book to describe the various features, components, and tools.

- **Azure Machine Learning** Contains all the tools necessary to design, develop, share, test, and deploy predictive analytic solutions in the Microsoft Azure cloud.
- **Azure Machine Learning workspaces** Represent a discrete "slice" of the Azure Machine Learning tool set that can be partitioned by the following criteria:
  - **Workspace name** Required to be unique and is the primary method for identifying a Machine Learning workspace.
  - **Workspace owner** Valid Microsoft account that will be used to manage access to this Azure Machine Learning workspace.

- **Data center location** Defines the physical Azure Data Center location for hosting the Azure Machine Learning workspace.
- **Storage account** Defines the unique Azure Storage account that will be used to store all of the data and artifacts related to this Azure Machine Learning workspace.
- **Azure Machine Learning experiments** Experiments are created within Azure Machine Learning workspaces and represent the primary method of enabling an iterative approach to rapidly developing Azure Machine Learning solutions. Within each Azure Machine Learning experiment, Azure ML Studio gives you an interactive, visual workspace to easily build, test, and iterate on a predictive analytic experiment. These experiments can then be submitted to Azure ML Studio for execution. Azure Machine Learning experiments are highly iterative. You can easily create, edit, test, save, and rerun experiments. The use of Azure Machine Learning experiments is specifically designed to allow today's modern data scientists to "fail fast" when evaluating new predictive models while providing the ability to progress predictive model feedback for further model refinements. Simply put, Azure Machine Learning gives you the iterative speed to either fail fast or ultimately succeed.
- **Azure ML Studio** This is the primary interactive predictive analytics workbench that is accessed from within an Azure Machine Learning workspace to allow a data scientist to create Azure Machine Learning experiments via a drag-and-drop visual designer interface. Access to a unique Azure ML Studio environment is governed from within an Azure Machine Learning workspace. In addition to enabling the creation of new experiments, Azure ML Studio also has links to sample Azure Machine Learning experiments. These are provided so that you can easily learn from others as you make your way on your data science journey and leverage some of the best processing techniques and tools in the industry to help accomplish your specific predictive analytics goals.
- **Azure Machine Learning web services** These represent Azure Machine Learning experiments that have been exposed as public APIs over the Internet in the form of the Azure Machine Learning REST API. These services are generally exposed as a simple web service, or as an OData endpoint. The API provides two types of RESTful web interfaces:
  - **Request Response Service (RRS)** For individual, low-latency, synchronous uses, for making predictions.
  - **Batch Execution Service (BES)** For asynchronous scoring of a batch of data records. The input for BES is a batch of records from a variety of sources such as blobs, tables, SQL Azure, HD Insight (as a result of a Hive Query), and HTTP sources.
- **Datasets** This is data that has been uploaded to Azure ML Studio so that it can be used in the prediction modeling process. A number of sample datasets are included with Azure ML Studio for you to experiment with, and you can upload more datasets as you need them.

- **Modules** These are algorithms that you can apply to your data. Azure ML Studio has a number of modules ranging from data ingress functions to training, scoring, and validation processes. Here are some examples of included modules:
  - **Convert to ARFF** Converts a .NET serialized dataset to ARFF format. ARFF is a common machine learning construct and stands for Attribute-Relation File Format. It is commonly defined as an ASCII text file that describes a list of instances sharing a set of attributes.
  - **Elementary Statistics** Calculates elementary statistics such as mean, standard deviation, and so on.
  - **Linear Regression** Creates an online gradient, descent-based, linear regression model.
  - **Score Model** Scores a trained classification or regression model.

A module might have a set of parameters that you can use to configure the module's internal algorithms. When you select a module on the canvas, its parameters are displayed in the pane to the right of the canvas. You can modify the parameters in that pane to tune your model.

Enough of the background, motivations, terminology, and predictive analytic theories. It's time to get started!

## Getting started

---

The absolute first step in your Azure Machine Learning journey is to get access to the Microsoft Azure environment. There are several ways to get started, and here are your options:

- **Option 1** Take advantage of a free Azure trial offer at <http://azure.microsoft.com/en-us/pricing/free-trial>.
  - This offer allows you to get a \$200 credit to spend on all Azure services for one month. You can use this \$200 credit however you wish to create and try out any combination of Azure resources. It enables you to explore the Azure cloud for free.
  - Note that this option will require you to apply using a Microsoft account and a valid credit card number for account verification purposes. The free trial offer is designed so that you will not start the actual billing process until you completely agree.
  - During the trial period, the Azure Management Portal will prominently display your remaining trial credits at the top of the portal page to inform you of how many credits are left.
- **Option 2** Take advantage of the (free) Azure Machine Learning trial offer at <https://studio.azureml.net/Home>.

- This is a free Azure offer that is feature-specific and therefore only allows you access to the Azure Machine Learning environment.
- This is an extremely low-friction option for new adopters: All you need is a valid Microsoft account to get started.
- If you need to sign up for a Microsoft account, visit <http://windows.microsoft.com/en-US/windows-live/sign-up-create-account-how>.
- Once you have signed in with a valid Microsoft account, an introductory video is displayed to help get you started, as shown in Figure 3-1. You can review this introductory video at <https://go.microsoft.com/fwlink/?LinkID=518038>.

Welcome!

Here is an overview video to get you started.



**FIGURE 3-1** Azure Machine Learning introductory video.

Note that if you opt to take advantage of the free Azure Machine Learning trial offer, you will only have access to the Azure Machine Learning features, not access to the full Azure environment. To truly maximize your experience, it is highly recommended that you gain access to the complete Microsoft Azure environment.

Figure 3-2 is a screenshot of the initial Azure Machine Learning environment. For the remainder of this book, we assume navigation to the Azure Machine Learning features via the Azure Management

Portal. Once you navigate to the Azure ML Studio screen, the Azure Machine Learning features are the same.

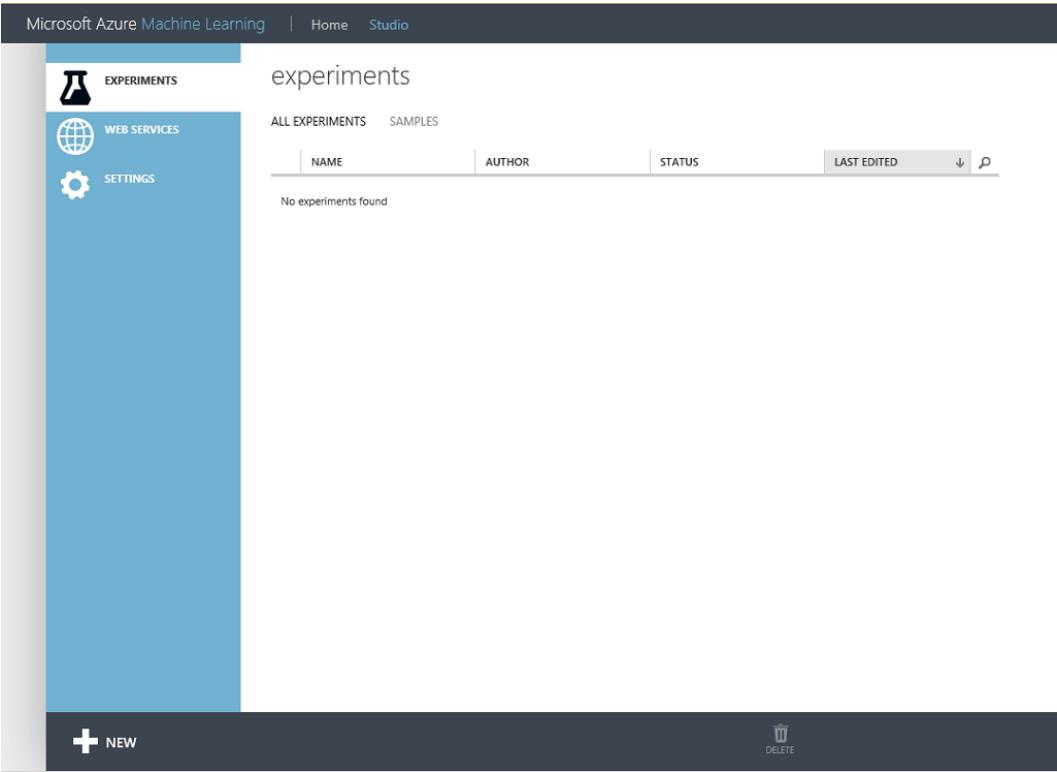


FIGURE 3-2 The Azure Machine Learning environment.

## Azure Machine Learning pricing and availability

Microsoft Azure Machine Learning has been designed as a suite of offerings to help enable partners and customers to easily design, develop, test, implement, and manage predictive analytic solutions in the Azure cloud.

Ultimately, Azure Machine Learning will be another set of the Azure cloud “pay-for-play” services. You will soon see how this can be an extremely effective model to pursue for design, testing, and implementing predictive analytics in the cloud.

At the time of this writing, the Azure Machine Learning services are currently in “preview” or “beta” mode and not yet deemed ready for general availability and consumption until the appropriate levels of quality and customer and partner feedback have been achieved.

In terms of where you can host an Azure Machine Learning service, the only valid datacenter location (at the time of this writing) is the “South Central US” Azure datacenter. Expect more Azure datacenter locations to come online as this feature nears general availability status.

Despite the initial limited availability, Microsoft has published pricing details for Azure Machine Learning while it is still in preview mode. You can find more details on pricing options at <http://azure.microsoft.com/en-us/pricing/details/machine-learning/>.

Note that Microsoft is currently offering two tiers of Azure Machine Learning Services: Free and Standard. Each of these tiers provides a different level of Azure Machine Learning service capabilities, features, and pricing for each of the two primary services in Azure Machine Learning:

- **The Azure ML Studio Service** This service is used to help design, develop, share, test, and deploy Azure Machine Learning solutions.
- **The Azure ML API Service** This service is used to provide an Azure Machine Learning web service–hosting environment for use in running test and production Machine Learning experiments or scenarios.

Figure 3-3 is a screenshot that illustrates the current preview Azure Machine Learning pricing and availability options available at the time of this book’s writing.

Standard tier pricing

Machine Learning is now generally available. Preview pricing will remain in effect through March 31, 2015 as noted below.

Pricing through March 31:

	ML STUDIO SERVICE	ML API SERVICE
Hourly	\$0.38/Studio Experiment Hour	\$0.75/API Service Prediction Hour
Per Prediction	No Charge	\$0.18/1,000 API Service Predictions

GA pricing starting on April 1:

ML Seat Subscription		
	Monthly Fee	\$9.99/ Seat/ Month
ML Studio Usage		
	Hourly	\$1/Studio Experiment Hour
ML API Usage		
	Hourly	\$2/Production API Compute Hour
	Transactions	\$0.50/1,000 Production API Transactions

FIGURE 3-3 Azure Machine Learning pricing details while in preview mode.

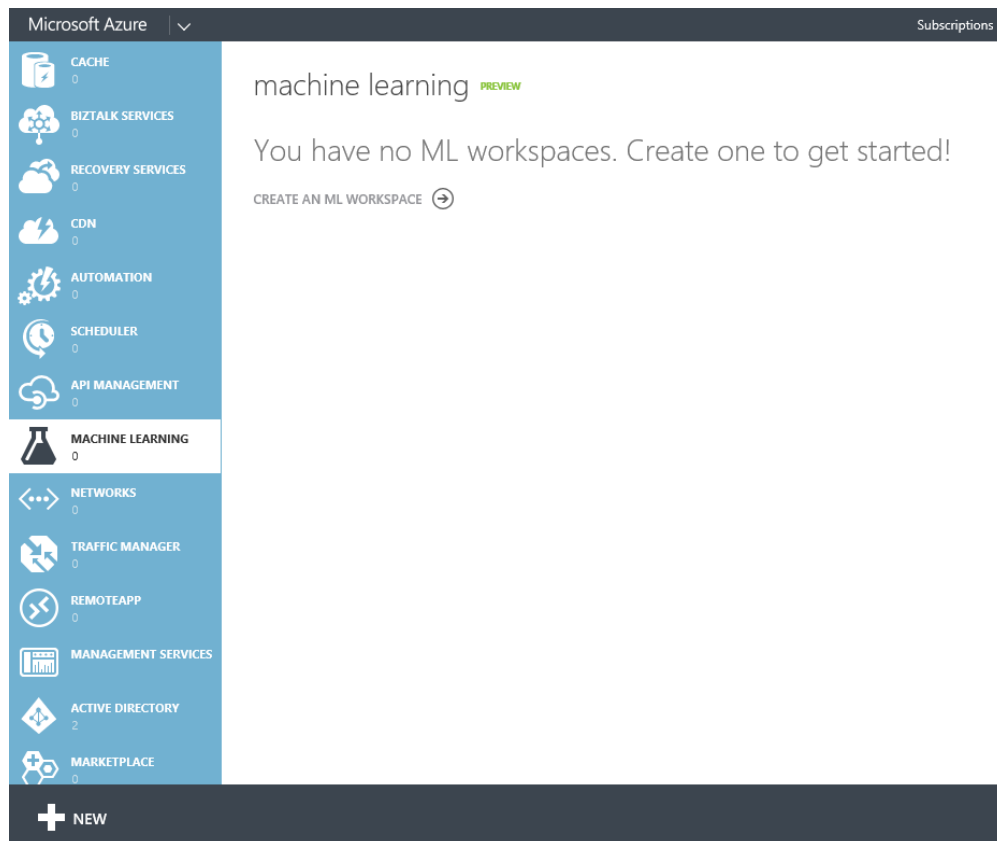
After the Azure free trial offer expires, as outlined in Option 1, these are the potential costs you could incur as you continue to develop and consume these services.

It's always a good idea to be aware of the Azure Machine Learning pricing information in the event you hit on the next predictive analytics breakthrough for, say, predicting the next big stock market turn.

Note that the preview Azure Machine Learning prices are subject to change at the time that the Azure Machine Learning services move into general availability (GA) mode.

## Create your first Azure Machine Learning workspace

Let's create our first Azure Machine Learning workspace. At this point, you should be all set up with either a free or paid Azure subscription or using the Azure Machine Learning free trial offer. Start by navigating to the Azure Management Portal at <https://manage.windowsazure.com>. From there, click the left navigation bar for Machine Learning as shown in Figure 3-4.



**FIGURE 3-4** The Azure Machine Learning section of the Microsoft Azure Management Portal.



An Azure Machine Learning workspace contains all of the tools you need to create, manage, and publish machine learning experiments in the cloud. To create a new Azure Machine Learning workspace, click the New icon in the bottom left of the page. Fill in the required fields as shown in Figure 3-5.

**FIGURE 3-5** Creating a new Azure Machine Learning workspace.

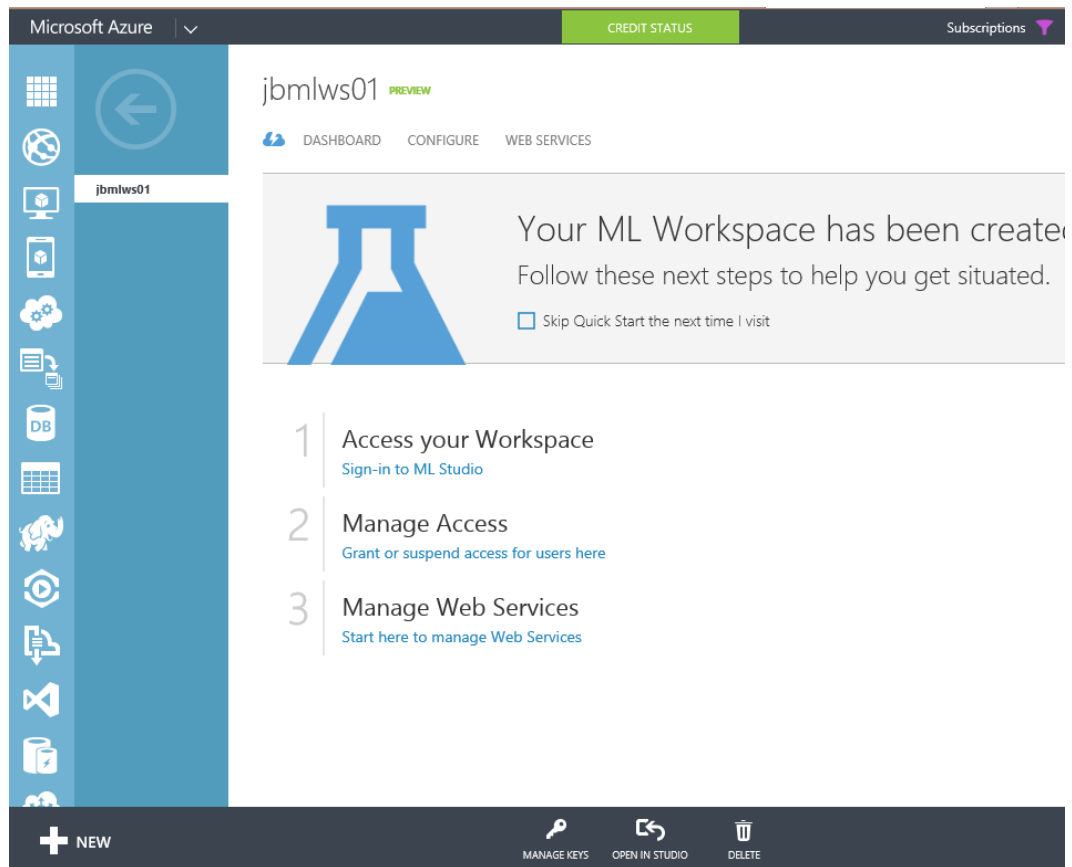
The required inputs for creating a new Azure Machine Learning workspace include the following:

- **Workspace Name** Provide a unique name for your Azure Machine Learning workspace. You will know if it is unique by the presence of the green check mark to the right of the text box after you move the cursor from this field.
- **Workspace Owner** Provide a valid Microsoft account (formerly Windows Live ID). Note that it cannot be a non-Microsoft account, such as your corporate email account. To create a free Microsoft account, go to [www.live.com](http://www.live.com).
- **Location** At the time of this writing, Azure Machine Learning services are available only in the South Central U.S. region.
- **Storage Account** Select the option to either create a new storage account or use an existing storage account.

- **New Storage Account Name** If you opt to create a new storage account for your Azure Machine Learning workspace, be sure that the storage account name is only made up of lowercase alphanumeric characters. You will know if it is unique by the presence of the green check mark to the right of the text box.

Once you click Create An ML Workspace, Azure will then provision a brand new Azure Machine Learning workspace for you to use to create and host your Azure Machine Learning experiments.

After your Azure Machine Learning Workspace has been created, click your new Azure Machine Learning Workspace and you should see a screen similar to Figure 3-6.



**FIGURE 3-6** A new Azure Machine Learning workspace.

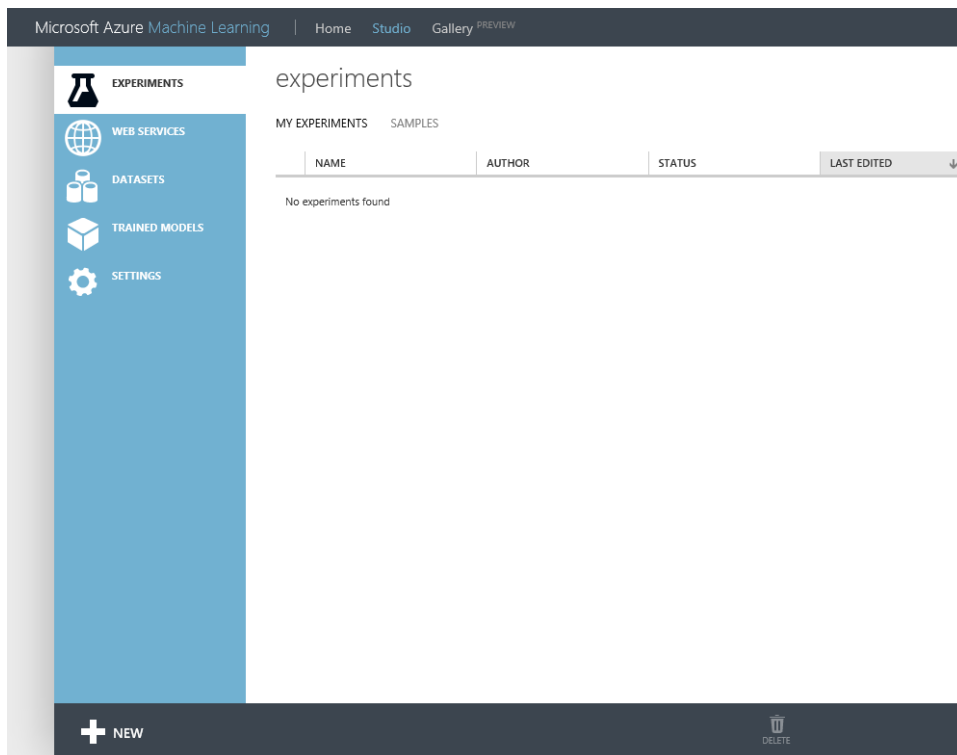
Note that this is the main landing page for managing an Azure Machine Learning workspace. From this page, you can directly access the Azure Machine Learning Studio tools for your workspace, manage user access to your workspace, and manage any web services that you might have deployed as a result of your experiments in your workspace. The top navigation menu provides access to several Azure

Machine Learning Workspace features.

- **Dashboard** Here you can monitor the relative or absolute computational usage of your workspace over a period of time.
- **Configure** This feature is used to allow or deny user access to your Azure Machine Learning workspace.
- **Web Services** This option allows you to manage web services, configure endpoints, and access your API via request/response or batch code samples in C#, Python, or in the R language, a popular programming language for data scientists and statisticians.

We revisit these features in detail later as we continue our exploration of the Azure Machine Learning environment.

Under Access Your Workspace, click the Sign-In To ML Studio link to sign in to your new Azure Machine Learning workspace. Figure 3-7 is a screenshot of the Azure ML Studio inside a new workspace.



**FIGURE 3-7** The Azure ML Studio inside of an Azure Machine Learning workspace.

When you first enter the Azure ML Studio workspace, you will see the following navigation tabs on the top and left navigation bars:

**Top Navigation Bar:**

- **Home** A set of links to documentation and other resources.
- **Studio** The landing page for Azure ML Studio experiments.
- **Gallery (Preview)** A collection of trending experiments and sample experiments.

**Left Navigation Bar:**

- **EXPERIMENTS** Experiments that have been created, run, and saved as drafts.
- **WEB SERVICES** A list of experiments that you have published.
- **DATASETS** Uploaded datasets that can be used in your experiments.
- **TRAINED MODELS** New predictive models that have been “trained” using the built-in Azure ML Studio machine learning algorithms.
- **SETTINGS** A collection of settings that you can use to configure your account and resources.

## Create your first Azure Machine Learning experiment

---

Let’s create a new experiment by clicking +NEW in the bottom left corner of the screen. You will then have the option to create either of the following items:

- **Data Set** This option allows you to upload a new dataset to use with your experiment from a local file on disk.
- **Experiment** Start with a blank experiment or a preexisting Microsoft sample experiment to help get you started fast.

Let’s start with a simple, real-world scenario such as predicting whether a person’s income exceeds \$50,000 per year based on his demographics or census data. You can imagine how incredibly useful the ability to predict a person’s income might be in the world of sales and marketing.

This is exactly the kind of predictive analytics that would be most useful for a successful targeted marketing campaign for products that require buyers with a certain level of disposable income. This will be a simplified example of how you could use Azure Machine Learning with ML Studio and ML API web services to create a real-world, cloud-based predictive analytics solution to help drive a marketing campaign.

In this walkthrough, we follow the entire process of developing a predictive analytics model in Azure

Machine Learning Studio and then publish it as an Azure Machine Learning API web service.

We start by downloading a sample Census Income Dataset from a public repository such as the UCI Machine Learning Repository from the following link:

<http://archive.ics.uci.edu/ml/datasets/Census+Income>.

We then develop and train a predictive model based on that dataset, and then publish the predictive model as a web service that can be used by other applications. These are the high-level steps we follow:

1. Download, prepare, and upload a census income dataset.
2. Create a new Azure Machine Learning experiment.
3. Train and evaluate a prediction model.
4. Publish the experiment as an Azure Machine Learning web service.
5. Access the Azure Machine Learning web service via sample tester programs.

## Download dataset from a public repository

---

To develop a predictive model for our sample income level predictive analytics model, we use the Adult Census Income Binary Classification dataset from the UCI Machine Learning repository. You can download the dataset from <http://archive.ics.uci.edu/ml/datasets/Census+Income>.

The website will have a link to a download folder, and you will want to download the adult.data file to your local computer. This dataset is in the Comma Separated Values (CSV) format. Note that the website also contains information about the 15 attributes found in this dataset. We use this information to create column headers for our data before we upload it into our experiment.

Now, open the adult.data file with Microsoft Excel or any other spreadsheet tool and add the column header names from the list of attributes on the website as in the following list. Note that some attributes are labeled as continuous, as they represent numerical values, whereas other attributes have a predefined list of potential values.

- **Age** Continuous
- **Workclass** Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked
- **Fnlwgt** Continuous
- **Education** Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool
- **Education-num** Continuous

- **Marital-status** Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse
- **Occupation** Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-ops, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces
- **Relationship** Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried
- **Race** White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
- **Sex** Female, Male
- **Capital-gain** Continuous
- **Capital-loss** Continuous
- **Hours-per-week** Continuous
- **Native-country** United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinidad&Tobago, Peru, Hong, Holland-Netherlands
- **Income** >50K, <=50K

After you have inserted a Header row with these column values, be sure to save the file with the .csv extension when you are done. For the purposes of this walkthrough, name the file Adult.data.csv when you save it. Figure 3-8 shows a screenshot of the Excel spreadsheet with the associated column headings.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	income
1	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
2	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
3	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
4	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K

**FIGURE 3-8** Screenshot of the Excel spreadsheet containing the adult census data.

Let's summarize a few facts about the Census Income dataset that we use for our first Azure Machine Learning experiment:

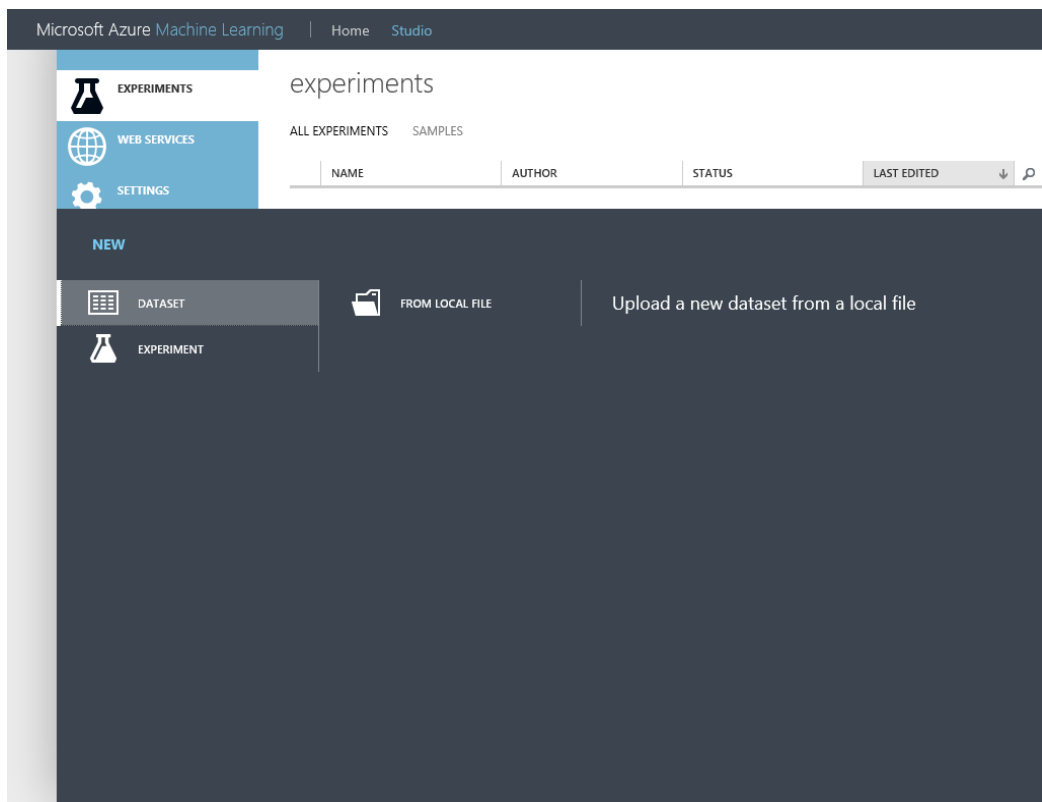
- **Number of unique attributes** Fourteen related to the outcome

- **Number of instances in the dataset** 48,842
- **Prediction task** Determine whether a person makes over \$50,000 a year

Note that this Census Income dataset is actually already provided by Microsoft as one of their sample datasets under the name of the Adult Census Income Binary Classification dataset. We are going through these steps manually to provide a comprehensive, end-to-end overview of the entire Azure Machine Learning workflow process. Most likely, your real-world predictive model datasets will also come from external sources, so it's good to know how it all works together from beginning to end.

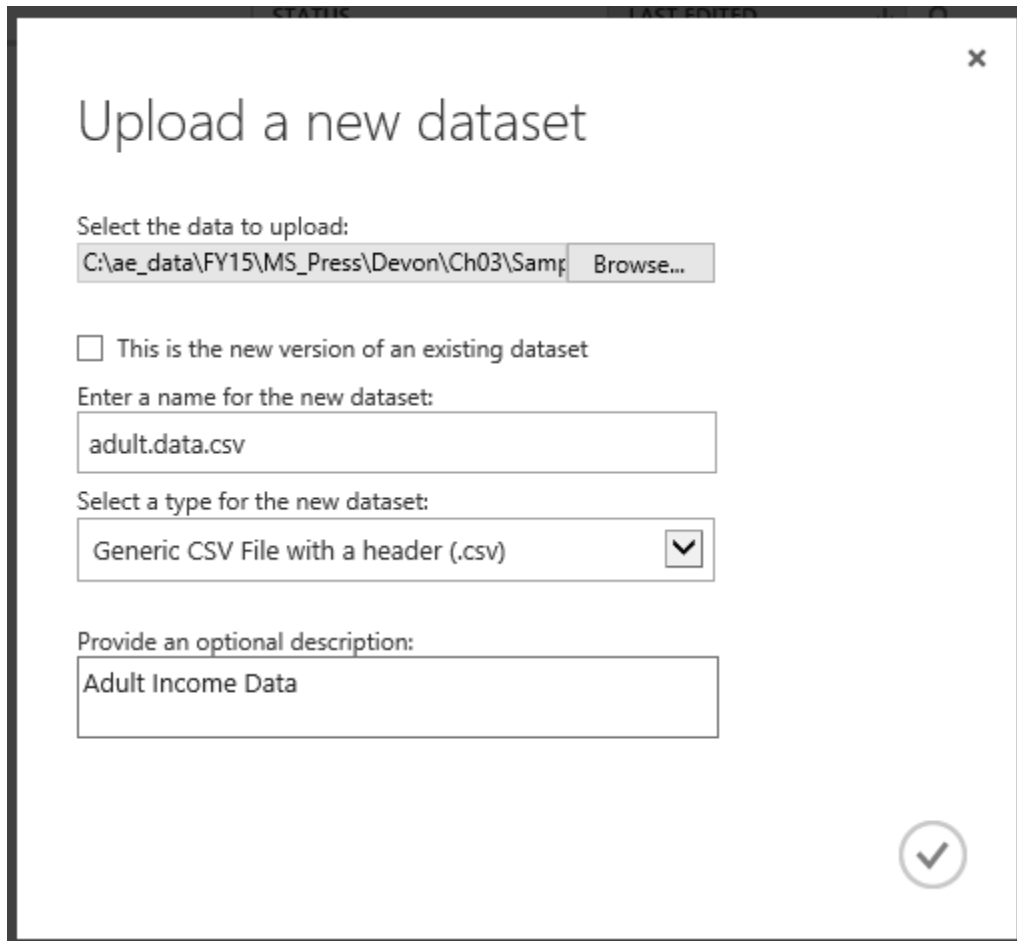
## Upload data into an Azure Machine Learning experiment

Once we have added the appropriate column headings to the sample Census Income dataset, we will then upload the dataset into the Azure Machine Learning workspace so that we can incorporate it into our prediction model. Click + New in the bottom left of the screen and select Dataset to upload a new dataset. Figure 3-9 shows the option to upload a new dataset from a local file.



**FIGURE 3-9** The Azure Machine Learning option to upload a dataset from a local file location.

Next, click From Local File. You will see an upload screen similar to Figure 3-10. Here you can specify the upload file properties such as the location of the file, the name for the new dataset (we will use Adult.data.csv), the type of file (Generic CSV file with a header), and an optional description for the new dataset.



Upload a new dataset

Select the data to upload:

C:\ae\_data\FY15\MS\_Press\Devon\Ch03\Samp Browse...

☐ This is the new version of an existing dataset

Enter a name for the new dataset:

adult.data.csv

Select a type for the new dataset:

Generic CSV File with a header (.csv) ✓

Provide an optional description:

Adult Income Data

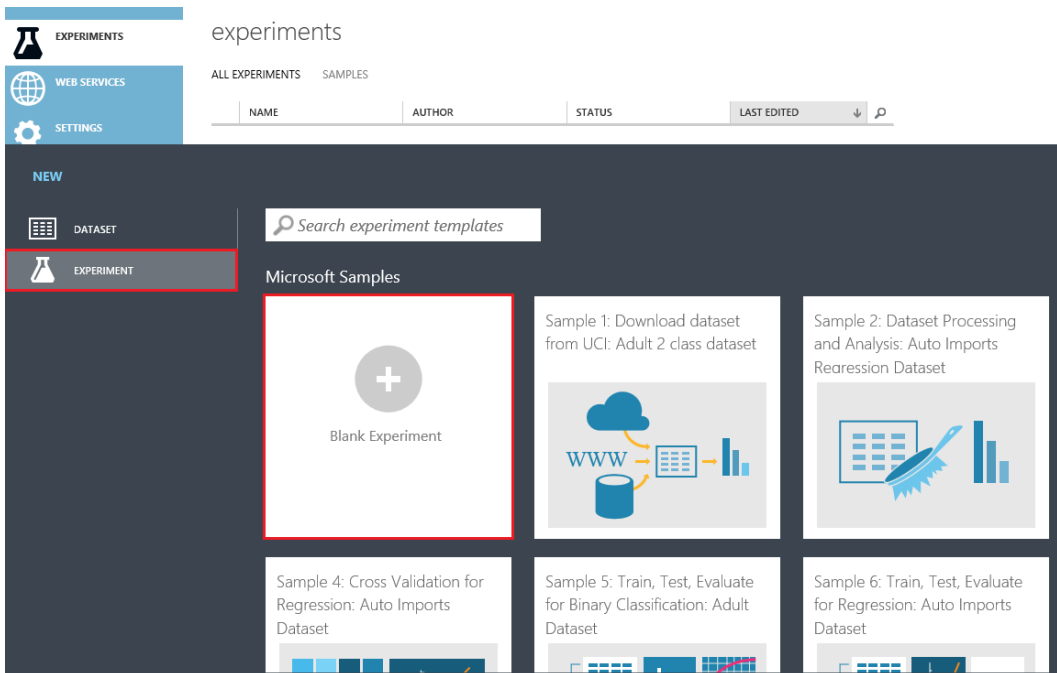
**FIGURE 3-10** Azure Machine Learning dialog box options to specify uploading a new dataset from a local file location.

Once you have entered the information and clicked the check mark, your dataset will be uploaded asynchronously into your Azure Machine Learning workspace for your use in your first Azure Machine Learning experiment.



## Create a new Azure Machine Learning experiment

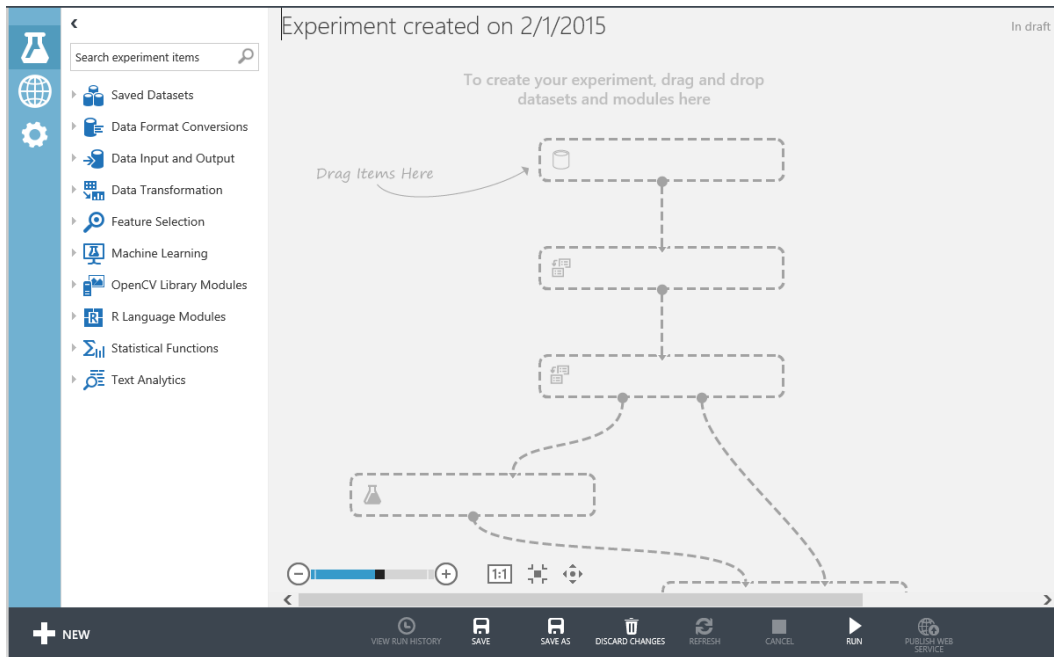
Create a new experiment by clicking on +NEW in the bottom left corner of the screen. Click Experiment and then click Blank Experiment as shown in Figure 3-11.



**FIGURE 3-11** List of new experiment types in Azure Machine Learning.

Note that in addition to a blank experiment template, there are many other sample experiments that you can load and modify to provide a jumpstart to your Azure Machine Learning activities.

Once the new blank experiment has loaded, you will then see the Azure ML Studio visual designer screen, as shown in Figure 3-12.



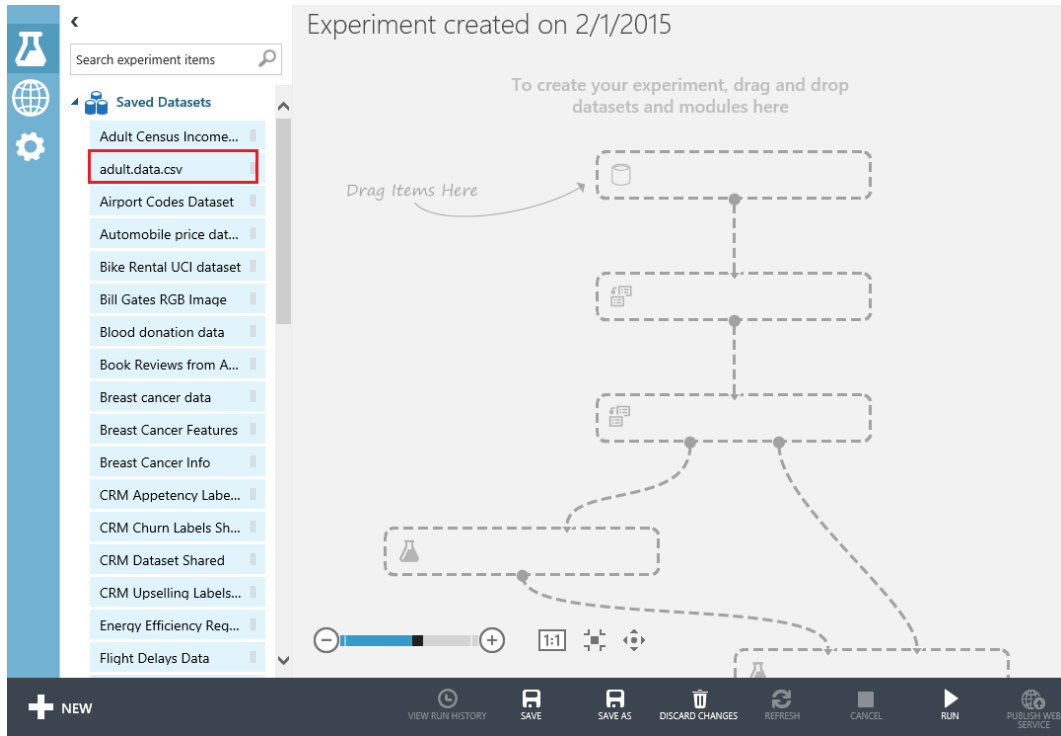
**FIGURE 3-12** A blank Azure Machine Learning experiment in the Azure ML Studio designer.

Note that the designer is made up of three main areas:

- **Left navigation pane** This area contains a searchable listing of all the Azure Machine Learning modules that can be used in creating a predictive analytics model.
  - Modules are grouped by functional area and contains functions for
  - Reading, formatting, and converting datasets.
  - Using and training machine learning algorithms.
  - Scoring and evaluating predictive model results.
- **Center pane** In the visual designer, Azure Machine Learning Experiments resemble flowcharts. They are assembled by dragging and dropping module shapes from the list in the left pane into the visual design surface in the middle of the screen. Modules can be freely positioned on the surface and are connected by drawing lines between input and output ports.
- **Right pane** In the Properties view, properties of selected modules are viewed and set using the pane on the right side of the visual designer.

Now, expand the Saved Datasets module on the left side of the screen and you will see where our uploaded dataset Adult.data.csv appears as a dataset for use in your Azure Machine Learning

experiment. Figure 3-13 shows the Adult.data.csv dataset ready to be dragged and dropped onto the visual designer surface.

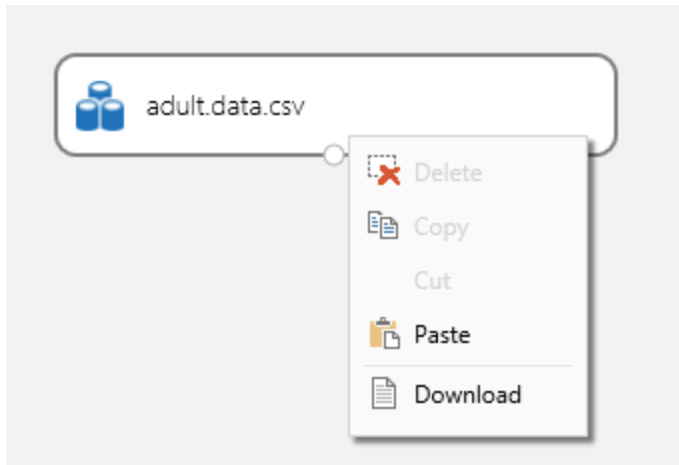


**FIGURE 3-13** Dragging the Adult.data.csv dataset onto the designer surface.

## Visualizing the dataset

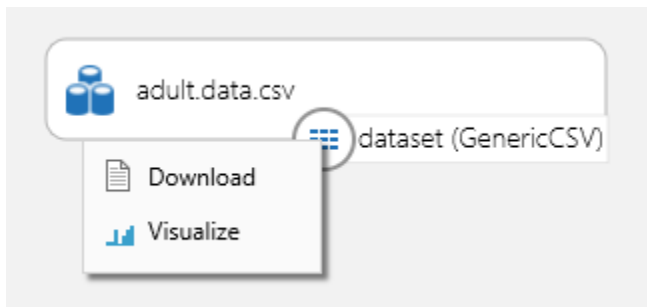
Drag the Adult.data.csv dataset into the middle of the visual designer surface and drop it into the experiment. Note that there are two ways to view the data in this dataset.

1. You can right-click anywhere in the dataset shape and then select Download to download the dataset to your local system. Figure 3-14 shows the Download menu option.



**FIGURE 3-14** Downloading a dataset from an Azure Machine Learning experiment.

2. You can hover over the connector at the bottom of the shape and then right-click. You will see the options shown in Figure 3-15 to either download or visualize the dataset.



**FIGURE 3-15** Right-clicking the connector displays options to download or visualize the dataset.

When you click Visualize, you will see a screen similar to Figure 3-16.

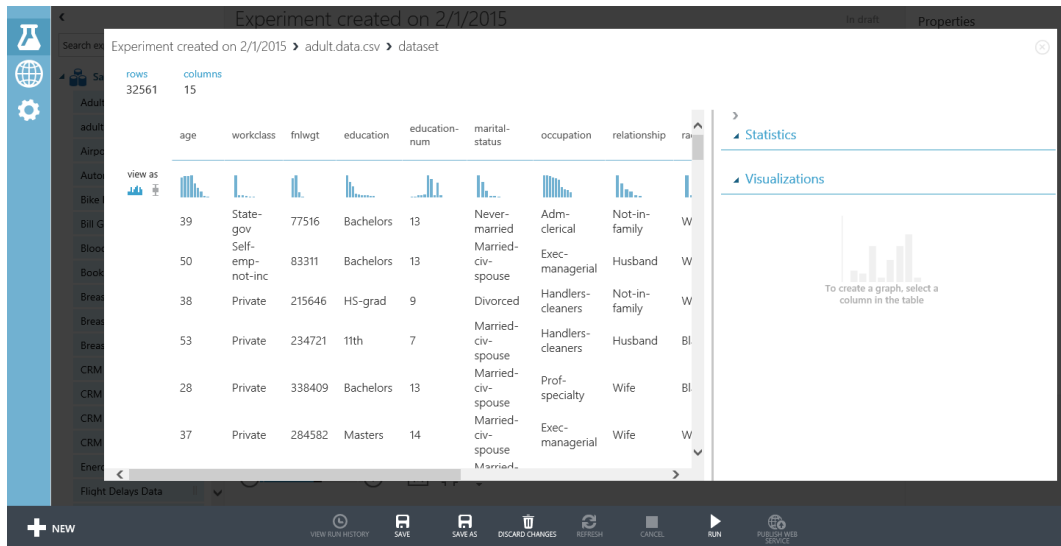


FIGURE 3-16 Visualization of the Adult.data.csv dataset.

Note that the Visualize option provides some great statistical and visualization features to allow you to quickly analyze the underlying data for the selected column. For example, click on the workclass column and you will see a screen similar to the one shown in Figure 3-17.

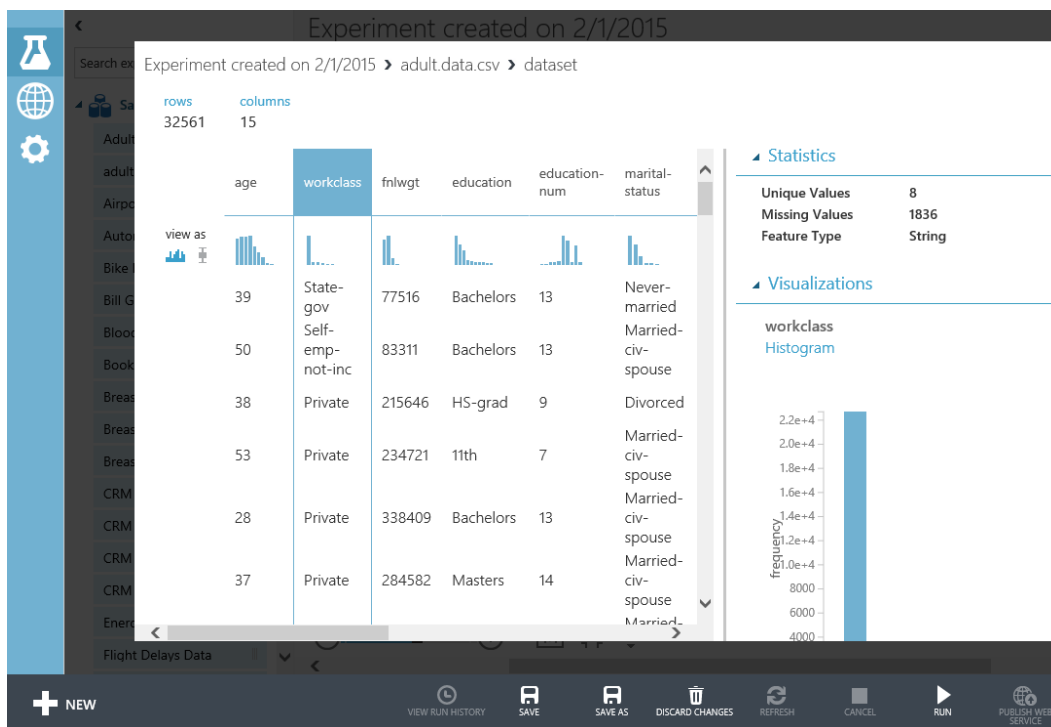


FIGURE 3-17 Statistics for the workclass column in the dataset.

Note that for the workclass column, the Statistics tab reveals the key information about the data found in this dataset shown in Figure 3-18.

### Statistics

Unique Values	8
Missing Values	1836
Feature Type	String

FIGURE 3-18 Key statistics for the workclass column in the Adult.data.csv dataset.

Note the following key details about the workclass column in the Adult.data.csv dataset:

- There are eight unique values found in this dataset.
- There are 1,836 records with missing values. This is an important detail, as the records with missing values would need to be either updated with the correct value or dropped completely to allow for the most accurate results.

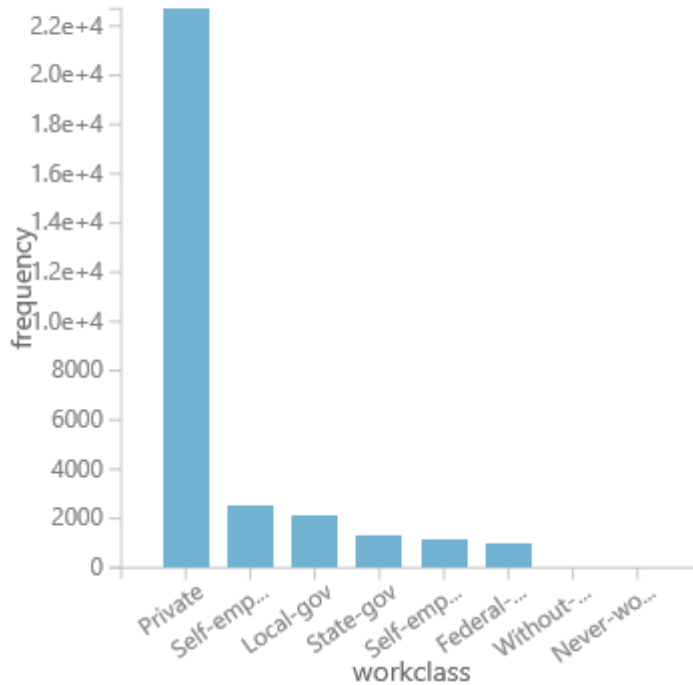
- The workclass field is designated as a String field type. As such, there are only nine valid string combinations for this field.

If you scroll down on the right side, you will see a Visualizations chevron similar to Figure 3-19 where the various field values for this column (workclass) are represented in a histogram.

## Visualizations

workclass

Histogram



**FIGURE 3-19** A histogram visualization of the workclass field in the Adult.data.csv dataset.

Using this built-in tool, you can easily determine from the visualization what the most common values found for the “workclass” field in the dataset are, such as Private and Self-employed on the left side.

Note how the ability to quickly and simply visualize your model's datasets using Azure ML Studio makes it fast and easy to infer key data elements, associations, combinations, and patterns. This ability to visualize and quickly establish inferences will help rapidly create a powerful predictive analytics solution.

## Split up the dataset

---

Typically, when creating an Azure Machine Learning experiment, we want to partition or split our dataset into two logical groupings for two specific purposes:

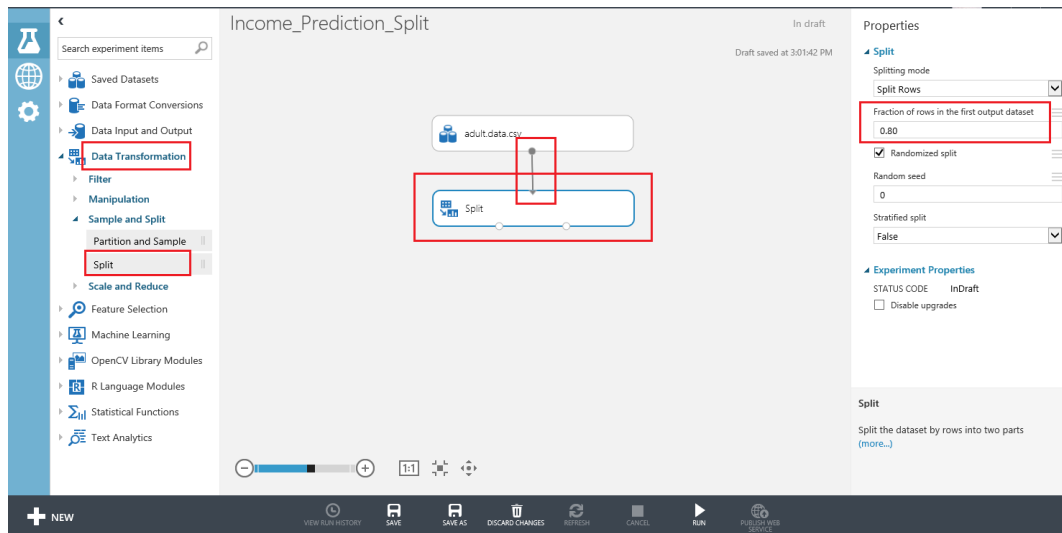
- **Training data** This grouping is used for creating our new predictive model based on the inherent patterns found in the historical data via the ML algorithm we use for the solution.
- **Validation data** This grouping is used for testing the new predictive model against known outcomes to determine accuracy and probabilities.

To accomplish this task, perform the following steps to split your dataset into two parts.

1. Expand the Data Transformation set of modules in the left pane.
2. Drag and drop the Split module onto the Azure Machine Learning designer surface.
3. Connect the Split module to the Adult.data.csv dataset.
4. Click the Split module and set the value of the Fraction Of Rows In The First Output field to 0.80. This will divert 80 percent of our data to a training area in the designer.



Figure 3-20 displays a screenshot of how this is done in Azure ML Studio.



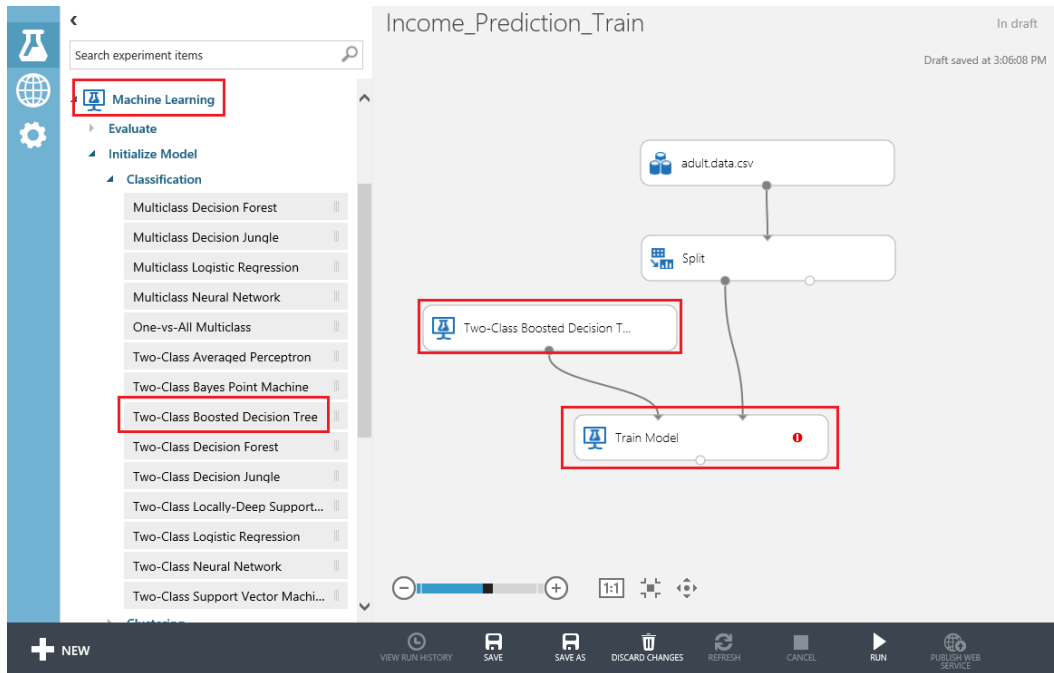
**FIGURE 3-20** Splitting the Adult.data.csv dataset for training and testing purposes.

This approach allows us to divert 80 percent of the data in the dataset to training the model. We can use the remaining 20 percent of the data in the dataset to test the results of our new model for accuracy.

## Train the model

The next step is to insert an Azure Machine Learning algorithm so we can “teach” the model how to evaluate the data. Start by expanding the Machine Learning module in the left pane. Then expand the Train submodule. Drag the Train Model shape onto the designer surface. Then connect the Train Model shape to the Split shape.

Next, expand Initialize Model under the Machine Learning module. Then expand the Classification submodule. For this experiment, we use Two-Class Boosted Decision Tree. Select it and drag and drop it onto the Azure Machine Learning designer surface. Your experiment should now look like the screenshot in Figure 3-21.



**FIGURE 3-21** The Azure Machine Learning experiment with a Train Model shape connected to the Two-Class Boosted Decision Tree module.

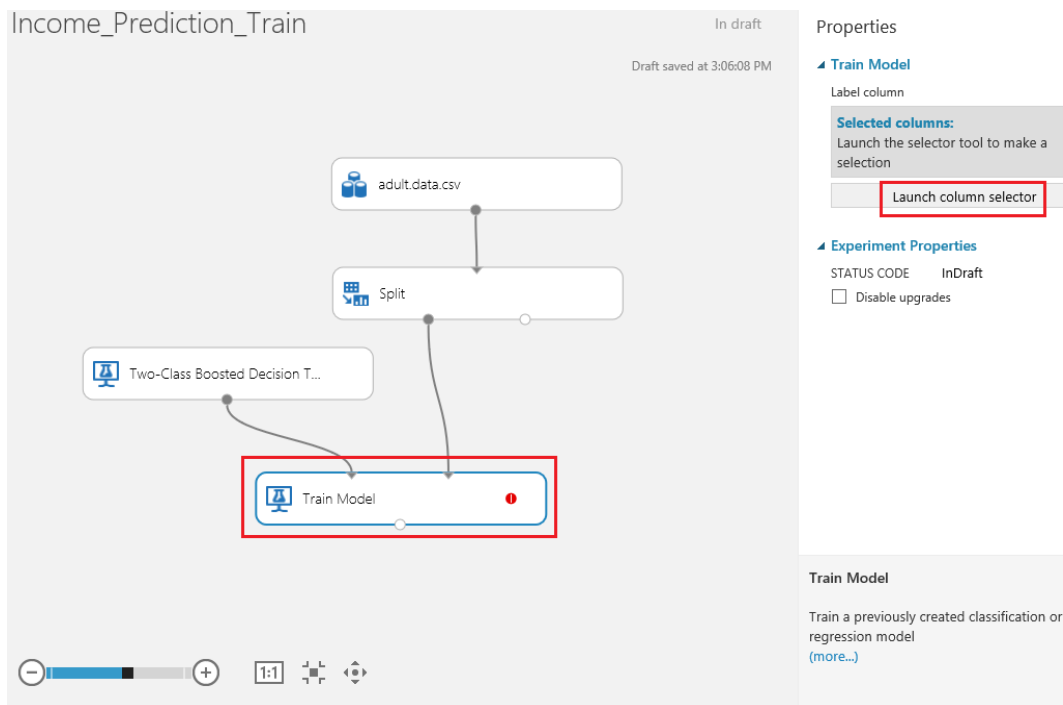
At this point, we have designed our experiment to split off 80 percent of the Adult Census dataset to be used for training our model through a Boosted Decision Tree Regression algorithm.

At this point, you might be wondering why we chose this particular algorithm to work with our prediction. Don't worry, as we cover the topic of the proper use and application suitability of various machine learning predictive algorithms in a future chapter, so for now let's use the Two-Class Boosted Decision Tree for our simple example Azure Machine Learning experiment.

## Selecting the column to predict

To finish configuring the algorithm, we need to indicate which column in the dataset is the outcome or prediction column, the column that is to be predicted based on all the other columns in any particular row of the dataset.

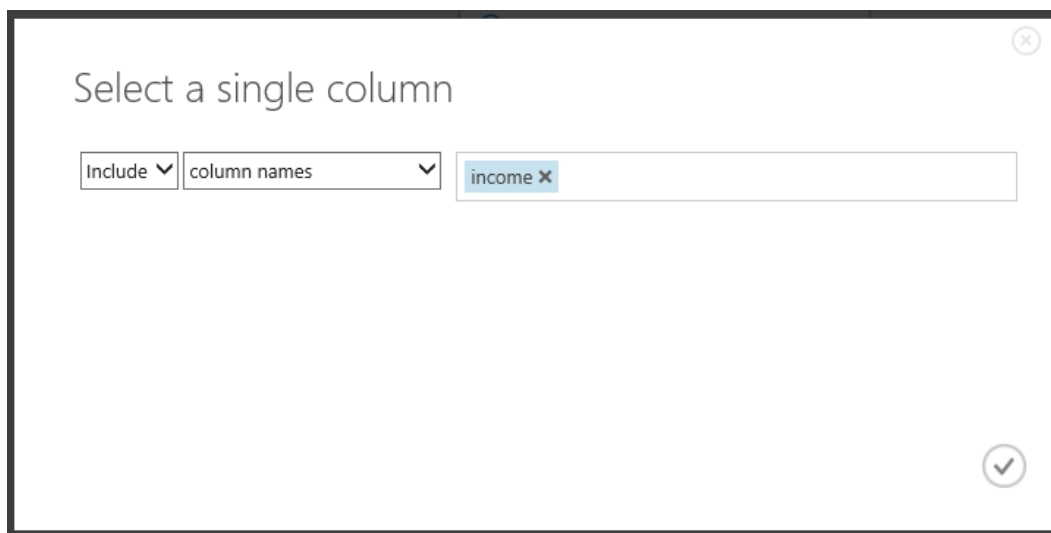
To do this, click the Train module. When you click the module, a Properties pane will open on the right side of the Azure ML Studio screen, as shown in Figure 3-22.



**FIGURE 3-22** Opening the column selector for the Train module.

Once you have set the module on the design surface, launch the column selector in the right pane. Select include and the column named income.

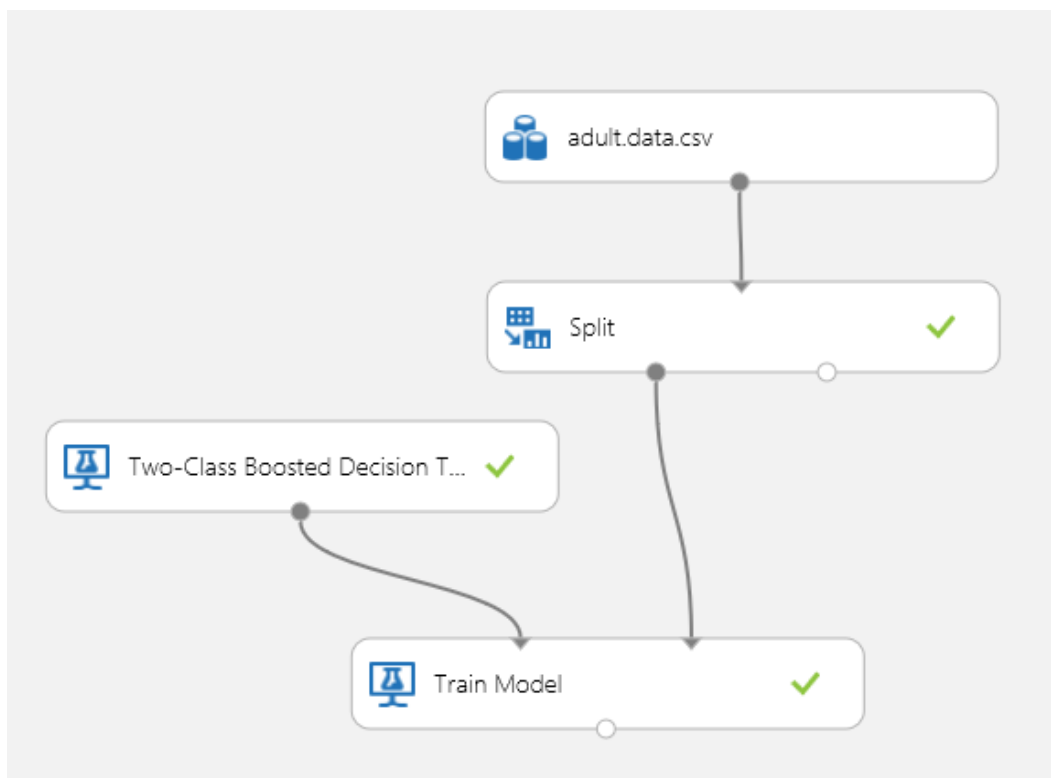
Figure 3-23 is a screenshot of the column selector denoting that the income column is to be predicted from the incoming dataset.



**FIGURE 3-23** Configuring the Train module to select a single column and include the income column.

In this way, we are instructing the Azure Machine Learning algorithm to infer patterns from all the other columns in each row of the dataset, so that the income column can be predicted. We are using 80 percent of our dataset to “train” the model, based on known inputs and outputs.

At this point, we are now ready to actually start training our model. Select the RUN option at the bottom of the screen, and sit back and watch as it actually trains our model. You will notice that as each stage of our experiment completes, a green check mark will appear on the right side of each operation in our experiment, as shown in Figure 3-24.



**FIGURE 3-24** Training our new Azure Machine Learning income prediction model.

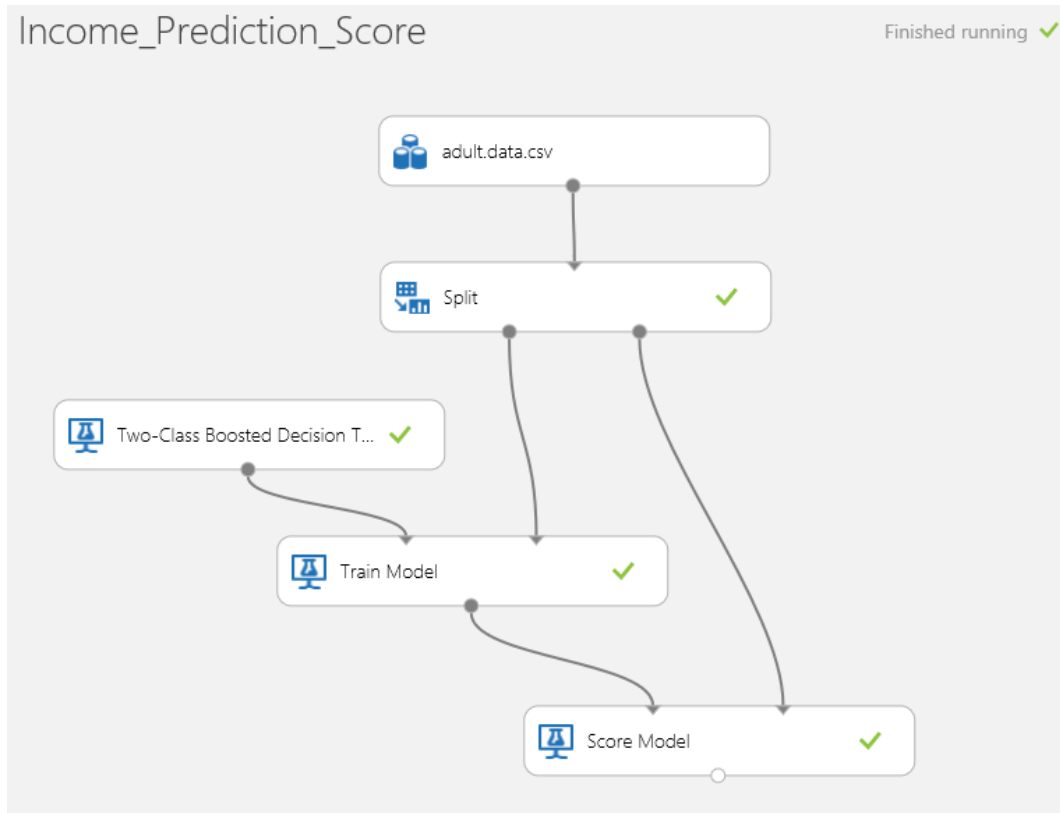
## Score the model

Now that we have trained our new Azure Machine Learning prediction model, we can next evaluate the results to determine the model's accuracy and therefore suitability as a solution. Remember, the key to a great Azure Machine Learning solution is to develop iteratively, where the keys to success are to fail fast and fail often.

To implement the evaluation module, expand the Machine Learning module on the left side of Azure Machine Learning Studio. Then expand the Score Model submodule. Drag and drop the Score Model module onto the designer surface. Next, connect the Score Model module to the Train module. Then finally, connect it to the other half of the Split module. Essentially you are now “scoring” the accuracy of the model against the remaining 20 percent of the dataset to understand how accurate the prediction model really is (or isn't).

Next, click Run on the bottom of the page and wait for all the results to be processed (denoted by a green check mark appearing on the right side of each module). Figure 3-25 is a screenshot of our

income prediction Azure Machine Learning experiment at this point.

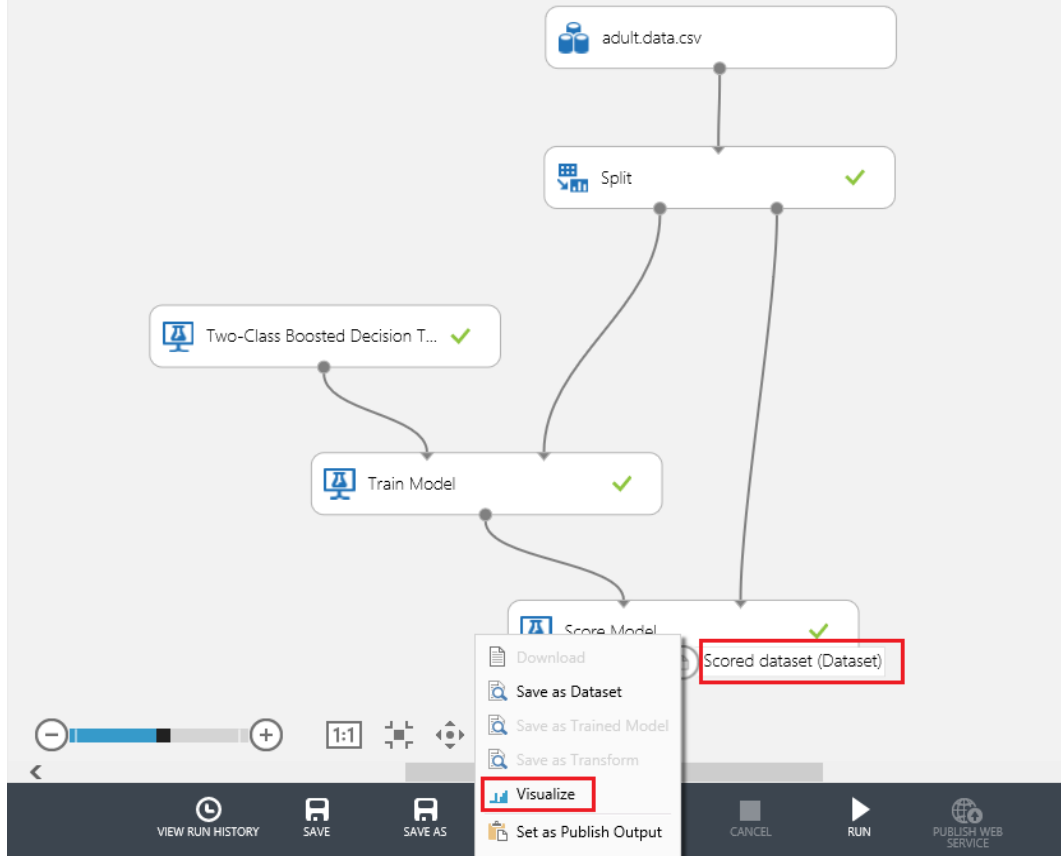


**FIGURE 3-25** Azure ML Studio, training and scoring the model.

## Visualize the model results

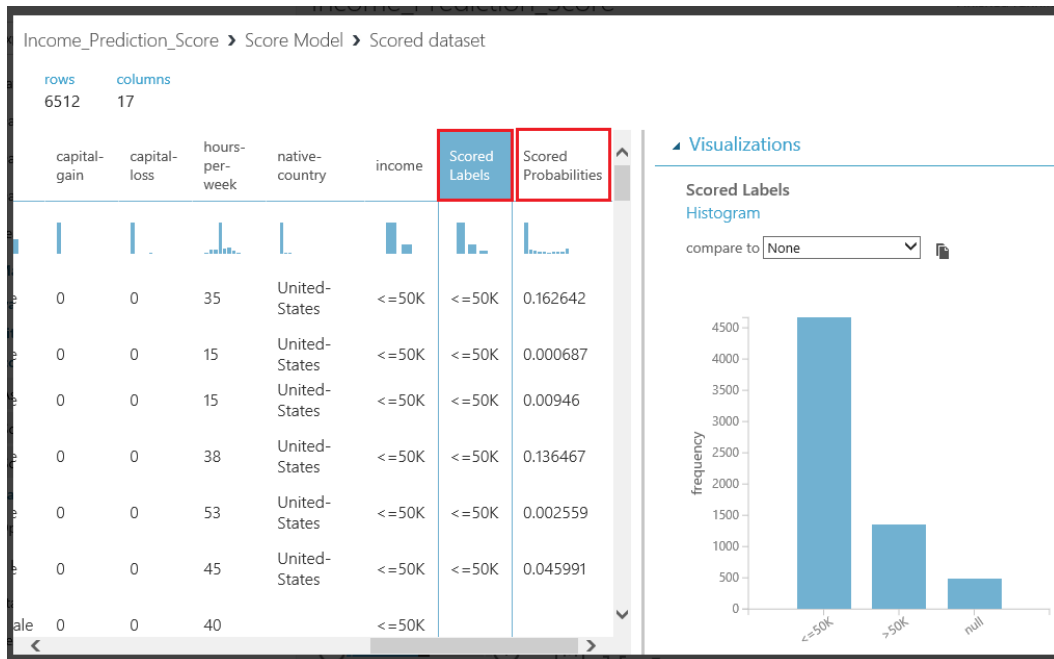
After all modules have been processed, hover your cursor over the output of the Score Model module and right-click. From the shortcut menu, select Visualize as shown in Figure 3-26.

## Azure ML Income Prediction - Train and Score Model



**FIGURE 3-26** Visualizing the results of the scoring module in our Azure ML Studio experiment.

After you select the option to visualize the newly trained model data, a new screen is generated. Scroll all the way to the right of the visualization and you will note that there are now two additional columns that appear in the dataset, as shown in Figure 3-27.



**FIGURE 3-27** The two new columns added to each row of our trained model indicating the model's prediction and prediction probability calculation for each row.

Note that there are now two additional columns that appear in our dataset:

- **Scored Labels** This column denotes the model's prediction for this row of the dataset.
- **Scored Probabilities** This column denotes the numerical probability (or the likelihood) of whether the income level for this row exceeds \$50,000.

These new columns in our dataset represent that, in addition to making a prediction calculation for each row, the algorithm can also provide a numerical probability factor. This probability factor represents the model's potential for accurately predicting each row in the dataset based on the specific values found in each of the row's other columns.

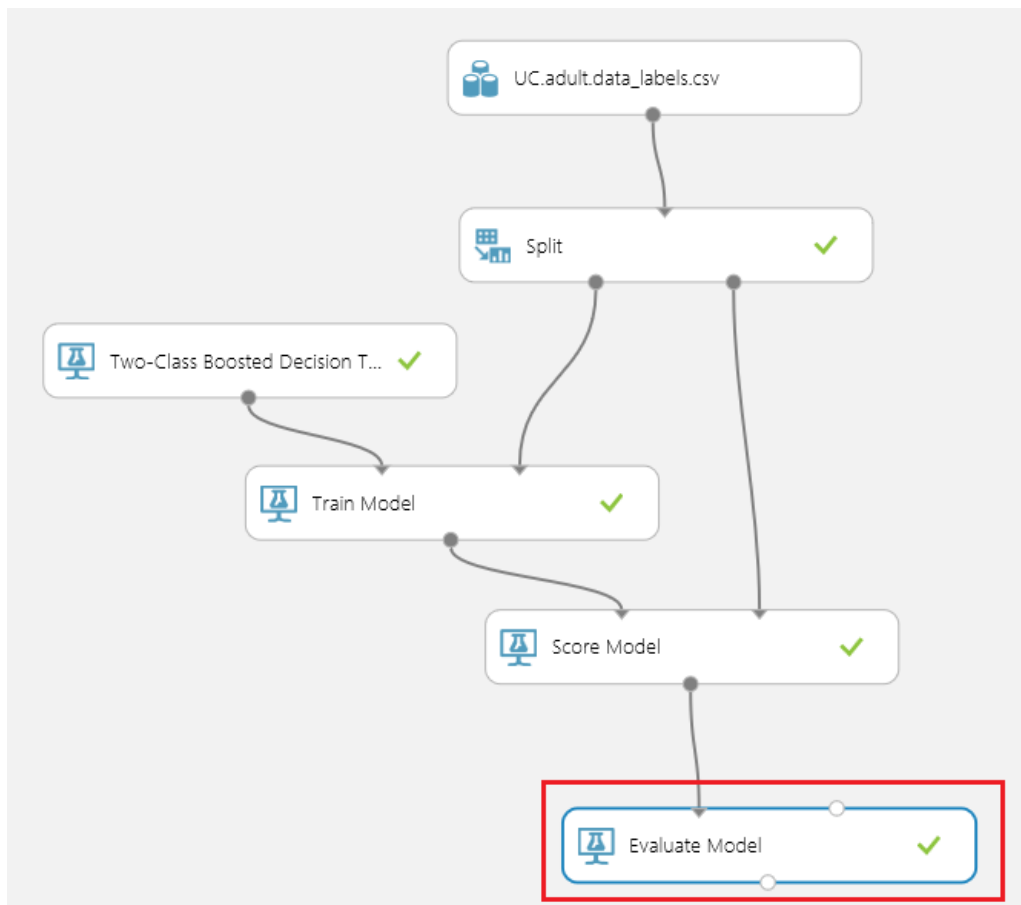
Typically, this is the point where predictive analytics becomes a very iterative process. You might want to try many different algorithms or even chain them together (in an advanced machine learning topic called an ensemble) to come up with a predictive model that proves fruitful.



## Evaluate the model

One of the most compelling features of Azure Machine Learning is the ability to quickly evaluate different algorithms. Choose the right one with just a few mouse clicks, thanks to the Evaluate module. To see how accurate our model really is, we can easily evaluate our model using a built-in Evaluate module in Azure ML Studio.

To do this, click the Machine Learning module in the left navigation pane of Azure ML Studio. Select the Evaluate submodule. Finally, select the Evaluate Model module and drag it onto the visual designer surface, below the Score Model module. Connect the other half of the Split module to the Score Model module. Then connect the Score Model module to the left side of the Evaluate Model module, as shown in Figure 3-28.

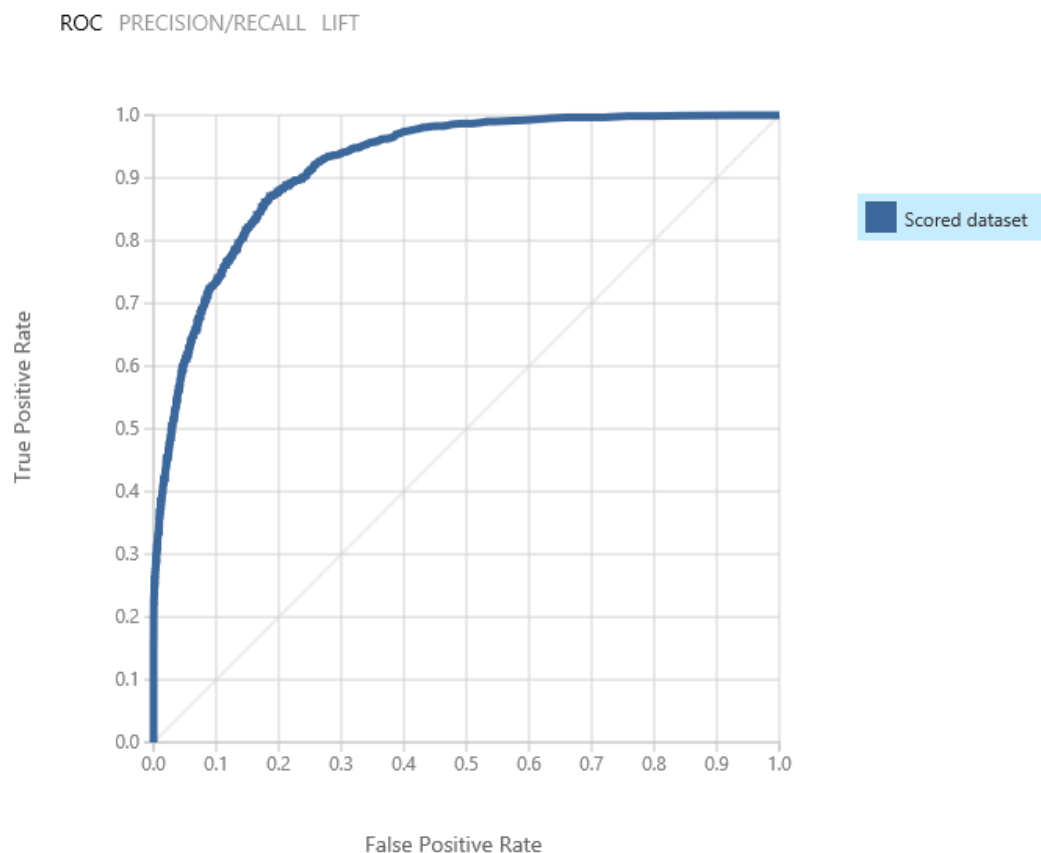


**FIGURE 3-28** Connecting the Evaluate Model module to evaluate the results of the income prediction module.

Click Run at the bottom of the Azure ML Studio screen. Watch as each stage is processed and marked complete, as denoted by a green check mark on the right of each module in our experiment.

After the processing has completed, right-click the bottom connector of the Evaluate Model module. Select Visualize from the shortcut menu to generate the evaluation results screen shown in Figure 3-29.

Income\_Prediction\_Evaluate > Evaluate Model > Evaluation results



**FIGURE 3-29** Visualizing the Azure Machine Learning Evaluation Model module results for the income prediction model.

The Evaluate Model module produces a set of curves and metrics that allow you to view the results of a scored model or compare the results of two scored models. You can view the results in the following three formats:

- **Receiver Operator Characteristic (ROC) curves** This format displays the fraction of true positives out of the total actual positives. It contrasts this with the fraction of false positives out of the total negatives, at various threshold settings. The diagonal line represents 50 percent accuracy in your predictions and can be used as a benchmark that you want to improve. The higher and further to the left, the more accurate the model is. As you do experiments you want to see the curve move higher and to the left.
- **Precision/Recall curves** Precision represents the fraction of retrieved instances that are relevant, whereas recall represents the fraction of relevant instances that are retrieved.
- **Lift curves** This format is a variation on the ROC curve. It measures the fraction of true positives, in relation to the target response probability.

In the visualization in Figure 3-29, you can see that the results of both datasets (our “trained” dataset and our “validation” dataset) are nearly identical, with the blue and red lines almost exactly on top of each other. This would indicate that we have a reasonably accurate prediction model. Consequently, for the purposes of this initial walk-through of Azure Machine Learning, we assume that the predictive model is reasonably accurate and take it to the next phase.

## Save the experiment

---

At this point, you want to save a copy of your experiment. Click **Save As** at the bottom of the screen. We are about to make major changes to our experiment that will alter the core functionality from being a training experiment to an operational experiment. Save your experiment using a descriptive name such as **Azure ML Income Prediction – Train Model Experiment**.

Next, click **Save As** at the bottom of the screen to resave the experiment before implementing the next phase using a different name such as **Azure ML Income Prediction – Implement Model**.

## Preparing the trained model for publishing as a web service

---

Now that we have a working “trained” predictive model that produces reasonable results, the next step is to run the option in the bottom navigation pane called **Prepare Web Service**. Notice that before these steps, this option was unavailable as a valid command to execute.

Before we execute the next command, now is a good time to do a click **Save As** and save the experiment under a new name. The reason for doing this now is that when we execute the **Prepare Web Service** command, it will make a few modifications to our experiment to make things ready for publishing it as a web service. Figure 3-30 shows our experiment before we execute the **Prepare Web Service** command.