Microsoft Azure Machine Learning

This book provides you with the skills necessary to get started with Azure Machine Learning to build predictive models as quickly as possible, in a very intuitive way, whether you are completely new to predictive analysis or an existing practitioner.

The book starts by exploring ML Studio, the browser-based development environment, and explores the first step—data exploration and visualization. You will then build different predictive models using both supervised and unsupervised algorithms, including a simple recommender system. The focus then shifts to learning how to deploy a model to production and publishing it as an API.

The book ends with a couple of case studies using all the concepts and skills you have learned throughout the book to solve real-world problems.

Who this book is written for

The book is intended for those who want to learn how to use Azure Machine Learning. Perhaps you already know a bit about Machine Learning, but have never used ML Studio in Azure; or perhaps you are an absolute newbie. In either case, this book will get you up-and-running quickly.

What you will learn from this book

- Learn to use Azure Machine Learning Studio to visualize and pre-process data
- Build models and make predictions using data classification, regression, and clustering algorithms
- Build a basic recommender system
- Deploy your predictive solution as a Web service API
- Integrate R and Python code in your model built with ML Studio
- Explore with more than one case study

Sumit Mund

Explore predictive analytics using step-by-step tutorials and build models to make prediction in a jiffy with a few mouse clicks
In this package, you will find:

- The author biography
- A preview chapter from the book, Chapter 1 'Introduction'
- A synopsis of the book’s content
- More information on Microsoft Azure Machine Learning
About the Author

Sumit Mund is a BI/analytics consultant with about a decade of industry experience. He works in his own company, Mund Consulting Ltd., where he is a director and lead consultant. He is an expert in machine learning, predictive analytics, C#, R, and Python programming; he also has an active interest in Artificial Intelligence. He has extensive experience working with most of Microsoft Data Analytics tools and also on Big Data platforms, such as Hadoop and Spark. He is a Microsoft Certified Solution Expert (MCSE in Business Intelligence).

Sumit regularly engages on social media platforms through his tweets, blogs, and LinkedIn profile, and often gives talks at industry conferences and local user group meetings.
You are reading this probably because you are aware of the importance of machine learning and advanced analytics, such as predictive analytics. While there is an increasing demand for people all over the world who possess these skill sets, there is a real scarcity of data scientists who are skilled enough to deliver applications that involve machine learning and advanced analytics and can create real value from the available data.

The reason for this scarcity is because the field of machine learning and data mining used to be the realm of PhDs and experts in subjects such as math, statistics, and programming combined. It's really difficult to find such unicorns. Again, tasks such as predictive analytics have historically been so difficult that even experts, even if they don't exactly struggle, don't find it easy either. This means that years of experience are needed for newcomers to to get on with it.

In this modern age, predictive analytics is on the verge of being industrialized as it is the key to sustaining and promoting the growth of a business. While the scarcity of "unicorn" data scientists doesn't seem to be ending, organizations are now finding solutions to get over this problem. A leading IT research firm, Gartner, suggests that, in the coming days, a new breed of professionals will emerge, referred to as citizen data scientists. Their emergence may bring about such a change that they may soon outnumber unicorn data scientists by a ratio of 5:1.

You might be wondering now, who are these citizen data scientists and where have they come from? They are existing developers, people from the business analyst community, and, possibly, new graduates as well, who are data-savvy, passionate about advanced analytics, and determined to stretch themselves and go in-depth into data science concepts. They will democratize data science and enable the industrialization of advanced analytics.
All this is happening and will continue to happen because of one reason: the arrival of new tools and platforms that make advanced analytics so easy and present data science as a commodity. While this brings huge opportunities for such vendors, it also bring good news for organizations and professionals who are picking it up. There is no doubt that Azure Machine Learning is a leader in this field and Microsoft offers this to organizations, strategically.

Microsoft's corporate vice president, Joseph Sirosh, who is in charge of Azure Machine Learning, describes Azure Machine Learning, as published in CITEworld: "This is the fastest way to build predictive models and deploy them. Very few tools exist today if you're going to build solution on the cloud and create applications. This way you can build intelligent applications from data, then publish as APIs so you can hook them up very easily from any enterprise application—and even from mobile. We're building it simple enough for a high schooler to be able to use it."

This book is an attempt to extend this vision; driven by simplicity, it sets the mission to develop the necessary skills to get started with Microsoft Azure Machine Learning as quickly as possible. The book assumes no prerequisites other than high school math!

What this book covers

Chapter 1, Introduction, sets the context for the book, and it introduces machine learning, predictive analytics, and Azure ML as a whole. It describes a predictive analytics project through its life cycle.

On your mark: do the background work

Chapter 2, ML Studio Inside Out, explains the ML Studio in detail—the development environment of Azure ML.

Chapter 3, Data Exploration and Visualization, familiarizes you with the concepts related to data exploration and visualizations in the first part of this chapter, and then demonstrates the same using ML Studio.

Chapter 4, Getting Data in and out of ML Studio, describes the different options available for data input and output inside ML Studio.

Chapter 5, Data Preparation, familiarizes you with the different options for data preparation in ML Studio, such as data cleaning, transformation, feature selection, and so on.

Get Set: build and deploy predictive models
Chapter 6, Regression Models, familiarizes you with the different regression algorithms available, and demonstrates the building of different regression models with step-by-step tutorials.

Chapter 7, Classification Models, familiarizes you with the different classification algorithms available and demonstrates the building of different classification models with step-by-step tutorials.

Chapter 8, Clustering, explains clustering and then builds a model using ML Studio and the K-means clustering algorithm.

Chapter 9, A Recommender System, introduces you to the concepts of a recommendation system and also the options available in ML Studio for you to build your own recommender system. It then walks you through building a recommendation system with a simple example.

Chapter 10, Extensibility with R and Python, introduces you to integrating your code in ML Studio using R and Python scripting.

Chapter 11, Publishing a Model as a Web Service, explores how easily you can publish a model in an experiment and make it available as a Web service API for others to consume.

Go: apply your learnings to real-world problems

Chapter 12, Case Study Exercise I, presents a classification problem as a case study exercise.

Chapter 13, Case Study Exercise II, presents a regression problem as a case study exercise.
Welcome to the world of predictive analytics and machine learning! Azure Machine Learning enables you to perform predictive analytics with the application of machine learning. Traditionally, it has been an area for experts. Developing and deploying a predictive modeling solution using machine learning has never been simple and easy, even for experts. Microsoft seems to have taken most of the pain out with this new cloud-based offering that allows you to develop and deploy a predictive solution in the simplest and quickest possible way. Even beginners would find it easy and simple to understand.

This chapter, while setting the context for the rest of the book, will present the related topics from a bird’s eye view.

Introduction to predictive analytics

Predictive analytics is a niche area of analytics that deals with making predictions of unknown events that may or may not be in future. One example of this would be to predict whether a flight will be delayed or not before the flight takes off. You should not misunderstand that predictive analytics only deals with future events. It can be any concerned event, for example, an event where you need to predict whether a given credit card transaction is a fraud or not when the transaction has already taken place. In this case, the event has already taken place. Similarly, if you are given some properties of soil, and you need to predict a certain other chemical property of soil, then you are actually predicting something that is present.
Predictive Analytics leverages tools and techniques from Mathematics, Statistics, Data Mining and Machine Learning plays a very important role in it. In a typical predictive analytics project, you usually go through different stages in an iterative manner, as depicted in the following figure;

**Problem definition and scoping**
In the beginning, you need to understand; what are the business needs and the solutions they are seeking? This may lead you to a solution that lies in predictive analytics. Then, you need to translate the business problem in an analytics problem, for example, the business might be interested in giving a boost to the catalog sales for the existing customers. So, your problem might get translated to predict the number of widgets a customer would buy if you know the demographic information about them, such as their age, gender, income, location, and so on, or the price of an item, given their purchase history of the past several years. While defining the problem, you also need to define the scope of the project; otherwise, it might end up in a never-ending process.

**Data collection**
The solution starts with data collection. In some cases, the data may already be there in enterprise storages or in the cloud, that you just have to utilize and in other cases, you need to collect the data from disparate sources. It may also require you to do some ETL (Extract, Transform, and Load) work as part of data collection.
Data exploration and preparation

After you have all the data you need, you can proceed to understand it fully. You do so by data exploration and visualization. This may also involve some statistical analysis.

Data in the real world is often messy. You should always check the data quality and how it fits for your purpose. You have to deal with missing values, improper data, and so on. Again, data may not be present in the proper format, as you would need it to make predictions. So, you may need some preprocessing to get the data in the desired shape. Often, people call it data wrangling. After this, you can either select or extract the exact features that lead you to the prediction.

Model development

After the data is prepared, you choose the algorithm and build a model to make a prediction. This is where machine learning algorithms come in handy. A subset of the prepared data is taken to train the model and then you can choose to test your model with another set or the rest of the prepared data to evaluate its performance. While evaluating the performance, you can try different algorithms and choose the one that performs the best.

Model deployment

If it is a one-off analysis, you may not bother deploying your trained model. However, often, the prediction made by the model might be used somewhere else. For example, for an e-commerce company, a prediction model might recommend products for a prospective customer visiting the website. In another example, after you have built a model to predict the sales volume for the year, different sales departments across different locations might need to use it to make the forecasts for their region. In such scenarios, you have to deploy your trained model as a web service or in some other type of production, so that others can consume it either by a custom application, Microsoft Excel, or a similar tool.

For most of the practical cases, these phases never remain in isolation and are always worked on in an iterative manner.

This book, with an overview of the different common options available for data exploration and preparation, focuses on model development and deployment. In fact, model development and deployment is the core offering of Azure Machine Learning with the limited options for data exploration and preparation. You can make use of other Azure services, such as HDInsight, Azure SQL Database, and so on, or programming languages outside it for the same.
Machine learning

Samuel Arthur, known to be the father of machine learning, defines it as a field of study that gives computers the ability to learn without being explicitly programmed. To simplify it, machine learning is a scientific discipline that explores the construction and study of algorithms that can learn from data. Such algorithms operate by building a model from example inputs and use that model to make predictions or decisions rather than following strictly static program instructions.

To illustrate, consider that you have a dataset that contains the information about age, education, gender, and annual income of a sufficiently large number of people. Suppose you are interested in predicting someone's income. So, you will build a model by choosing a machine learning algorithm and train the model with the dataset. After you train your model, it can then predict the income of a new person if you provide it with age, education, and gender data. To explain it further, you have not programmed something explicitly, such as if a male's age is greater than 50 and whether he has a master's degree, then he would earn say $100,000 per annum. However, what you did was just choose a generic algorithm and gave it the data, so that it discovers all the relationships between the different variables or features (here, age, gender, and education) with the target variable income. So, the algorithm learned from the data and hence got trained. Now, with the trained algorithm, you can predict someone's income if you know their other variables.

The preceding example is a typical kind of machine learning problem where there exists a target variable or class; here that is income. So, the algorithm learns from the training data or examples and then after being trained, the algorithm predicts for a new case or data point. Such learning is known as the **Supervised Machine Learning**. It works as shown in the following figure:
There is another kind of machine learning where there is no target variable or the concept of training data or examples, so here, the prediction is also of a different kind. Consider the same dataset again that contains data of age, gender, education, and income of a sufficiently large number of people. You have to run a targeted marketing campaign, so you have to divide or group the people into three clusters. In this case as well, you can use a different kind of machine learning generic algorithm on the dataset that would automatically group the people into three groups or clusters. This kind of machine learning is known as \textbf{unsupervised machine learning}.

There is also another kind of machine learning that makes recommendations; remember how Amazon recommends books or Netflix recommends movies—which might surprise you as to how magically they know about a user's choice or taste.

Though machine learning is not limited to these three kinds, for the scope of this book, we would limit it to these three.

Again, the scope of this book and, of course, Azure Machine Learning limits the application of machine learning to just the area of predictive analytics only. You should be aware that machine learning is not limited to this. Machine learning finds it roots in artificial intelligence and powers a variety of applications, some of which you use in everyday life, for example, web search engines, such as Bing or Google are powered by Machine Learning or applications, so also personal digital assistants like Microsoft's Cortana and Apple's Siri. These days, driverless cars are also in the news, which use machine learning. So, such applications are countless.

\section*{Types of machine learning problems}

The following are some of the common kinds of problems solved through machine learning.

\subsection*{Classification}

\textbf{Classification} is the kind of machine learning problem where inputs are divided into two or more classes and the learner produces a model that assigns unknown inputs to one (or multi-label classification) or more of these classes or labels. This is typically handled in a supervised way. Spam detection is an example of classification, where the inputs or examples are e-mail (or other) messages and the classes are "spam" and "not spam" and the model to predict a new e-mail as spam or not are based on example data.
Regression

Regression problems involve predicting a numerical or continuous value for the target variable for the new data given in the dataset with one or more features or dependent variables and associated target values. A simple example can be where you have historical data of the price paid for different properties in your locality for say the last 5 years. Here, the price paid is the target variable and the different attributes of a property, such as the total built-up area; the type of property, such as a flat or semi-detached house; and so on, are different features or variables. A regression problem would be to predict the property price of a new property available in the market for sale.

Clustering

Clustering is an unsupervised learning problem and works on a dataset with no label or class variable. This kind of algorithm takes all of the data and groups them into different clusters say 1, 2, and 3, which were not known previously. The clustering problem is fundamentally different from the classification problem. The classification problem is a supervised learning problem where your class or target variable is known to train a dataset, whereas in clustering, there is no concept of label and training data. It works on all the data, and groups them into different clusters.

So, to put it simply, if you have a dataset and a class/label or target variable as a categorical variable, and you have to predict the target variable for a new dataset based on the given dataset (example), then this is a classification problem. If you are just given a dataset with no label or target variable and you just have to group them into \( n \) clusters, then it's a clustering case.

Common machine learning techniques/

algorithms

The following are some of the very popular machine learning algorithms:

Linear regression

Linear regression is probably the most popular and classic statistical technique used for regression problems to make prediction for a continuous value from one or more variables or features. This algorithm uses a linear function and it optimizes the coefficients that fit best to the training data. If you have only one variable, then you may think of this model as a straight line that best fits the data. For more features, this algorithm optimizes best hyperplane that fits the training data.
Logistic regression

Logistic regression is a statistical technique used for classification problems. It models the relationship between a dependent variable or a class label and independent variables (features) and then makes a prediction of a categorical dependent variable or a class label. You may think of this algorithm as a linear regression for a classification problem.

Decision tree-based ensemble models

A decision tree is a set of questions or decisions and their possible consequences arranged in a hierarchical fission. While the plain decision tree is not very powerful, an assembly of trees with the averaged out results can be very effective. These are ensemble models and differ by how the decision is sampled or chosen. Random forest or decision forest and boosted decision tree are two very popular and powerful algorithms. Decision tree-based algorithms can be used for both classification and regression problems.

Neural networks and deep learning

Neural networks algorithms are inspired by how a human brain works. It builds a network of computation units, neurons, or nodes. In a typical network, there are three layers of nodes: first, the input layer, the middle layer or hidden layer, and in the end, the output layers. Neural networks algorithms can be used for both classification and regression problems.

A special kind of neural networks algorithms where there are more than three layers along with the input and output layers and more than one hidden layers are known as Deep learning algorithms. These are getting increasingly popular these days because of remarkable results.

Though Azure Machine Learning is capable of deep learning (convolutional neural network—a flavor of the deep learning model as of writing of this book), the book does not include it.

Introduction to Azure Machine Learning

Microsoft Azure Machine Learning or in short Azure ML is a complete cloud service. It is accessible through the browser Internet Explorer (IE) 10 or its later versions. This means that you don't need to buy any hardware or software and don't need to worry about deployment and maintenance.
So, it's a fully managed cloud service that enables analysts, data scientists, and developers to build, test, and deploy predictive analytics into their applications or in a standalone analysis. It turns machine learning into a service in the easiest possible way and lets you build a model visually through drag and drop. Azure ML helps you to gain insight even of massive datasets, bringing all the benefits of the cloud by integrating other big data that processes an Azure service such as HDInsight (Hadoop) to machine learning.

Azure ML is powered by a decent set of machine learning algorithms. Microsoft claims that these are state-of-the-art algorithms coming from Microsoft Research and some of these actually power flagship products, such as Bing search, Xbox, Cortana, and so on.

**ML Studio**
Azure Machine Learning Studio or in short ML Studio is the development environment for Azure ML. It's totally browser-based and hence is accessible from a modern browser, such as IE 10 or its later versions. It also provides a collaborative environment where you can share your work with others.

ML Studio provides a visual workspace to build, test, and iterate on a predictive model easily and interactively. You create a workspace and create experiments inside it. You can consider making an experiment inside ML Studio as a project where you drag and drop datasets and analysis modules onto an interactive canvas, connecting them together to form a predictive model. Usually, you iterate your model's design, edit the experiment, save a copy if desired, and run it again. When you're ready, you can publish your experiment as a web service, so that it can be accessed by others or other applications.

When your requirement can't be met visually by dragging and dropping modules, ML Studio allows you to extend your experiment by writing code in either R or Python scripting. It also provides you a module that allows you to play with data using SQL queries.
Summary

You just finished the first chapter, which not only introduces you to predictive analytics, machine learning, and Azure ML, but also sets the context for the rest of the book. You started by exploring predictive analytics and learned about the different stages for a typical predictive analytics task. You then moved on to a high-level understanding of machine learning by gaining some knowledge about it. You also learned about the common type of problems solved through machine learning and some of the popular algorithms. After that, you got a very high-level overview of Azure ML and ML Studio.

The next chapter is all about ML Studio. It introduces you to the development environment of Azure ML with an overview of the different components of ML Studio.
Where to buy this book


Click here for ordering and shipping details.