Preface

According to the technology consulting group Gartner, most emerging technologies go through what they term the "hype cycle." This is a way of contrasting the amount of hyperbole or hype versus the productivity that is engendered by the emerging technology. The hype cycle has three main phases: *peak of inflated expectation*, *trough of disillusionment*, and *plateau of productivity*. The third phase refers to the mature and value-generating phase of any technology. The hype cycle for predictive analytics (at the time of this writing) indicates that it is in this mature phase.

Does this imply that the field has stopped growing or has reached a saturation point? Not at all. On the contrary, this discipline has grown beyond the scope of its initial applications in marketing and has advanced to applications in technology, Internet-based fields, health care, government, finance, and manufacturing. Therefore, whereas many early books on data mining and predictive analytics may have focused on either the theory of data mining or marketing-related applications, this book will aim to demonstrate a much wider set of use cases for this exciting area and introduce the reader to a host of different applications and implementations.

We have run out of adjectives and superlatives to describe the growth trends of data. Simply put, the technology revolution has brought about the need to process, store, analyze, and comprehend large volumes of diverse data in meaningful ways. The scale of data volume and variety places new demands on organizations to quickly uncover hidden trends and patterns. This is where data mining techniques have become essential. They are increasingly finding their way into the everyday activities of many business and government functions, whether in identifying which customers are likely to take their business elsewhere, or mapping flu pandemic using social media signals.

Data mining is a class of techniques that traces its roots to applied statistics and computer science. The process of data mining includes many steps: framing the problem, understanding the data, preparing data, applying the right techniques to build models, interpreting the results, and building processes to

deploy the models. This book aims to provide a comprehensive overview of data mining techniques to uncover patterns and predict outcomes.

So what exactly does the book cover? Very broadly, it covers many important techniques that focus on predictive analytics, which is the science of converting future uncertainties to meaningful probabilities, and the much broader area of data mining (a slightly well-worn term). Data mining also includes what is called descriptive analytics. A little more than a third of this book focuses on the descriptive side of data mining and the rest focuses on the predictive side of data mining. The most common data mining tasks employed today are covered: classification, regression, association, and cluster analysis along with few allied techniques such as anomaly detection, text mining, and time series forecasting. This book is meant to introduce an interested reader to these exciting areas and provides a motivated reader enough technical depth to implement these technologies in their own business.

WHY THIS BOOK?

The objective of this book is twofold: to help clarify the basic concepts behind many data mining techniques in an easy-to-follow manner, and to prepare anyone with a basic grasp of mathematics to implement these techniques in their business without the need to write any lines of programming code. While there are many commercial data mining tools available to implement algorithms and develop applications, the approach to solving a data mining problem is similar. We wanted to pick a fully functional, open source, graphical user interface (GUI)-based data mining tool so readers can follow the concepts and in parallel implement data mining algorithms. RapidMiner, a leading data mining and predictive analytics platform, fit the bill and thus we use it as a companion tool to implement the data mining algorithms introduced in every chapter. The best part of this tool is that it is also open source, which means *learning* data mining with this tool is virtually free of cost other than the time you invest.

WHO CAN USE THIS BOOK?

The content and practical use cases described in this book are geared towards business and analytics professionals who use data in everyday work settings. The reader of the book will get a comprehensive understanding of different data mining techniques that can be used for prediction and for discovering patterns, be prepared to select the right technique for a given data problem, and will be able to create a general purpose analytics process.

We have tried to follow a logical process to describe this body of knowledge. Our focus has been on introducing about 20 or so key algorithms that are in widespread use today. We present these algorithms in following framework:

1. A high-level practical use case for each algorithm.
2. An explanation of how the algorithm works in plain language. Many algorithms have a strong foundation in statistics and/or computer science. In our descriptions, we have tried to strike a balance between being academically rigorous and being accessible to a wider audience who don't necessarily have a mathematics background.
3. A detailed review of using RapidMiner to implement the algorithm, by describing the commonly used setup options. If possible, we expand the use case introduced at the beginning of the section to demonstrate the process by following a set format: we describe a problem, outline the objectives, apply the algorithm described in the chapter, interpret the results, and deploy the model. Finally, this book is neither a RapidMiner user manual nor a simple cookbook, although a recipe format is adopted for applications.

Analysts, finance, marketing, and business professionals, or anyone who analyzes data, most likely will use these advanced analytics techniques in their job either now or in the near future. For business executives who are one step removed from the actual analysis of data, it is important to know what is possible and not possible with these advanced techniques so they can ask the right questions and set proper expectations. While basic spreadsheet analyses and traditional slicing and dicing of data through standard business intelligence tools will continue to form the foundations of data exploration in business, especially for past data, data mining and predictive analytics are necessary to establish the full edifice of data analytics in business. Commercial data mining and predictive analytics software tools facilitate this by offering simple GUIs and by focusing on applications instead of on the inner workings of the algorithms. Our key motivation is to enable the spread of predictive analytics and data mining to a wider audience by providing both conceptual framework and a practical "how-to" guide in implementing essential algorithms. We hope that this book will help with this objective.

Vijay Kotu
Bala Deshpande