Book Review

Reviewed by Zhongxian Wang and Ruben Xing*

E-mail: wangj@mail.montclair.edu
E-mail: Xingr@mail.montclair.edu
*Corresponding author

Applied Data Mining for Forecasting Using SAS
by: Tim Rey, Arthur Kordon and Chip Wells
Published 2012
by SAS Institute Inc.
100 SAS Campus Drive, Cary, NC 27513, USA, 324pp
ISSN: 978-1-60764-662-4

Data mining is the science of managing and analysing large datasets and discovering novel patterns. The use of data mining can advance a company's position by creating a sustainable competitive advantage (Han et al., 2012; Wang, 2008). The development of non-linear time series modelling, based on various computational intelligence methods, such as neural networks, support vector machines, and genetic programming have been prompted by the increased demand for forecasting triggered. Scholars continuously work on improving the time series forecasts via various data mining methods in industrial applications. However, the available literature for integrating data mining methods in forecasting is very limited (Corne et al., 2012; Rand and Zilinskas, 2012). The existing books on the market are either focused on forecasting methods or on data mining approaches. In addition, there are very few references that discuss the numerous practical issues of applying forecasting in a business setting.

Applied Data Mining for Forecasting Using SAS is one of the first books on the market that fills this need. This is a terrific book for practitioners since it presents a framework for integrating data mining and time series forecasting and exhibits a methodology for large-scale multivariate industrial forecasting. Readers can 'learn from the kitchen'. The main purpose of this book is to give readers a hands-on tool concerning applying data mining for forecasting different business activities by using some of the most popular software from an industrial perspective in a holistic framework and in a business setting.

In chapter 1, the authors elucidate the business forces that drive the use of data mining for forecasting, list the limitations of classical univariate forecasting, and illustrate what the advantages of integrating data mining and forecasting actually are in the age of big data. Chapter 2 presents a generic work process, akin to Six Sigma methodologies, that helps to integrate the proposed approach into corporate culture. It defines and discusses in detail the process of data mining for forecasting. Defining and optimising work processes is a must in industrial applications since ignoring it may lead to failure. Adopting such a systematic approach is critical in order to solve complex problems and introduce new methods.

Chapter 3 focuses on an enterprise-wide implementation strategy of building hardware, software, and organisational infrastructures that are needed for the successful application of business forecasting. Applying data mining for forecasting in a business requires serious investments in hardware, software and training, but a cultural change must also take place. The importance of integrating the selected options into the existing corporate infrastructure cannot be ignored.

Chapter 4 gives a systematic view of the key technical and non-technical application issues as well as a complete checklist for applying data mining for forecasting. The final success of the application project depends on a good understanding of both the key technical and non-technical issues related to data mining and forecasting. In order to improve model quality and gain the user's trust, a data miner can include economic drivers as model inputs and pay attention to non-technical issues including managing forecasting expectations, handling the politics of forecasting, and avoiding bad practices.

Chapter 5 discusses the main issues related to extracting the necessary knowledge about the data from the experts and organising an effective data collection effort for implementing data mining for forecasting. Identifying the key economic drivers and defining the internal and external data sources and the metadata are important steps. There are the different methods for data extraction and alignment.

Chapter 6 identifies the main data preprocessing steps and emphasises their critical role for high-quality forecasting since 60% to 80% of the data mining work lies in this step. Preparing time series data for forecasting is a critical step in the modelling process since transaction data are not ready to use. Often time, series data might have to be contracted or expanded. All of the data needs to be on the same time frequency and merged. There are various methods for imputing missing data (leading, trailing or embedded) and handling outliers.

Chapter 7 discourses the foundation for the actually doing data mining by providing a practitioner's guide to data mining methods for forecasting. From a practical perspective, it defines the key data mining methods of forecasting, such as similarity analysis, varcluster analysis, principal component analysis, stepwise regression, decision trees, co-integration analysis, and genetic programming.

Chapters 8 through 11, the most important part of the book, offer a practitioner's guide to time series forecasting methods by defining an implementation strategy for successful real-world applications of data mining for forecasting. These chapters present a practitioner's guide of time series forecasting methods that details univariate, multivariate, hierarchical, and non-linear models.

Chapter 8 delivers useful tools for extracting and extrapolating relevant information from the data at hand. Emphasis is placed on developing a user's intuition regarding the techniques presented and on the process of model building associated with time series data. Also, technical details are provided as supplementary resources or references.

Beginning with an ordinary regression model in a time series setting, chapter 9 presents a rational polynomial transfer function framework. Extensions to the regression framework illustrate how dynamic relationships including include lags, shifts, and persistent effects between inputs and the target can be accommodated using the rational polynomial transfer function. Furthermore, accommodating trend and seasonal variation in the transfer function framework, as well as stochastic input variables are provided.

Chapter 10 offers additional modelling topics since many issues arise frequently in applied forecasting. The authors explain and provide solutions to problems associated with a large-scale forecasting scenario. The demonstrated concepts and techniques can be used with datasets of more complicate large-scale forecasting; even the example datasets are small.

Chapter 11 concentrates on non-linear forecasting models including non-linear modelling features, forecasting models based on either neural networks, or support vector machines, or evolutionary computation. Also, both multivariate models and unobserved component models are on the list.

Finally, chapter 12 illustrates the key topics in applying data mining for forecasting on a real business example by detailing an example of data mining with 1,441 variables. First, the authors removed non-informative variables (no data, no variance, and correlated perfectly with other Xs), and then inputted missing data (either by back-casting or within the data). Next, they used rigorous time series-based approaches to reduce the number of Xs through similarity and variable clustering, and finally selected 50 variables for modelling through three different time series methods (similarity, co-integration and cross correlation). The topics covered in chapters 8 to 11 could be used to refine this model even further just in case.

The key features that differentiate this book from other titles on data mining and forecasting are: Integrating data mining and forecasting due to the synergetic benefits in the area of variable reduction and variable selection for building multivariate forecasting models; A broader view of industrial forecasting; Emphasis on practical applications. You should have it.

References

Corne, D., Dhaenens, C. and Jourdan, L. (2012) 'Synergies between operations research and data mining: the emerging use of multi-objective approaches', *European Journal of Operational Research*, Vol. 221, No. 3, pp.469–479.

Han, J., Kamber, M. and Pei, J. (2012) *Data Mining: Concepts and Techniques*, 3rd ed., Morgan Kaufmann Publishers, Burlington, Massachusetts.

Rand, G.K. and Zilinskas, A. (2012) 'Data mining and knowledge discovery via logic-based methods', *Interfaces*, Vol. 42, No. 2, pp.221–223.

Wang, J. (2008) *Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications*, IGI Global, Hershey, Pennsylvania.