Chapter 1: Why Industry Needs Data Mining For Forecasting

1.1 Overview

In today's economic environment there is ample opportunity to leverage the numerous sources of time series data that are readily available to the savvy decision maker. This time series data can be used for business gain if the data is converted first to information and then to knowledge—knowing what to make when for whom, knowing when resource costs (raw material, logistics, labor, and so on) are changing or what the drivers of demand are and when they will be changing. All this knowledge leads to advantages to the bottom line for the decision maker when times series trends are captured in an appropriate mathematical form. The question becomes how and when to do so. Data mining processes, methods and technology oriented to transactional type data (data that does not have a time series framework) have grown immensely in the last quarter century. Many of the references listed in the bibliography (Fayyad et al. 1996, Cabena et al. 1998, Berry 2000, Pyle 2003, Duling and Thompson 2005, Rey and Kalos 2005, Kurgan and Musilek 2006, Han et al. 2012) speak to the many methods and processes aimed at building prediction models on data that does not have a time series framework. There is significant value in the interdisciplinary notion of data mining for forecasting when used to solve time series problems. The intention of this book is to describe how to get the most value out of the host of available time series data by using data mining techniques specifically oriented to data collected over time. Previous authors have written about various aspects of data mining for time series, but not in a holistic framework: Antunes, Oliveira (2006), Laxman, Sastry (2006), Mitsa (2010), Duling, Lee (2008), and Lee, Schubert (2011).

In this introductory chapter, we help build the case for using data mining for forecasting and using forecasting as a competitive advantage. We cover the explosion of available economic time series data, the basic background on forecasting, and the limitations of classical univariate forecasting (from a business perspective). We also define what a time series database is and what data mining for forecasting is all about, and lastly describe what the advantages of integrating data mining and forecasting actually are.

1.2 Forecasting Capabilities as a Competitive Advantage

Information Technology (IT) Systems for collecting and managing transactional data, such as SAP and others, have opened the door for businesses to understand their detailed historical transaction data for revenue, volume, price, costs and often times even the whole product income statement. Twenty-five years ago IT managers worried about storage limitations and thus would design "out of the system" any useful historical detail for forecasting purposes. With the decline of the cost of storage in recent years, architectural designs have in fact included saving various prorated levels of detail over time so that companies can fully take advantage of this wealth of information. IT infrastructures were initially put in place simply to manage the transactions. Today, these architectures should also accommodate leveraging this history for business gain by looking at it from an advanced analytics view point. Various authors have discussed this framework in detail (Chattratichat et al. 1999, Mundy et al. 2008, Pletcher et al. 2005, Duling et al. 2008).

Large corporations generally have many internal processes and functions that support businesses—all of which can leverage quality forecasts for business gain. This is beyond the typical supply chain need for having the right product at the right time for the right customer in the right amount. Some companies have moved to a lean pull replenishment framework in their supply chains. This lean approach does not preclude the use of high-quality forecasting processes, methods, and technology.

In addition to those who analyze the supply chain, many other organizations in a corporation can use high-quality forecasts. Finance groups generally control the planning process for corporations and deliver the numbers that the company plans against and reports to Wall Street. Strategy groups are always in need for medium- to long-range forecasts for strategic planning. Executive sales and operations planning (ESOP) demand medium-range forecasts for resource and asset planning. Marketing and sales organizations always need short- to medium-range forecasts for planning purposes. New business development (NBD) incorporates medium- to long-range forecasts in the NPV (net present value) process for evaluating new business opportunities. Business managers themselves rely heavily on short- and medium-term forecasts for their own businesses data but also need to know about the market. Since every penny saved goes straight to a company's bottom line, it behooves a company's purchasing organization to develop and support high-quality forecasts for raw material, logistics, materials and supplies, and service costs.

Differentiating a planning process from a forecasting process is important. Companies do in fact need to have a plan to follow. Business leaders do in fact have to be responsible for the plan. But claiming that this plan is in fact a forecast can be disastrous. Plans are what we "feel we can do" while forecasts are mathematical estimates of what is most likely. These are *not* the same; but both should be maintained. In fact, the accuracy of both should be maintained over a long period of time. When reported to Wall Street, accuracy in the actual forecast is more important than precision. Being closer to the wrong number does not help.

Given that so many groups within an organization have similar forecasting needs, why not move towards a "one number" framework for the whole company? If finance, strategy, marketing and sales, business ESOP, NBD, supply chain and purchasing are not using the same numbers, tremendous waste can result. This waste can take the form of rework or mismanagement if an organization is not totally aligned with the same numbers. Such cross-organizational alignment requires a more centralized approach that can deliver forecasts that are balanced with input from the business and financial planning parts of the corporation. Chase (2009) presents this corporate framework for centralized forecasting in his book called *Demand Driven Forecasting*.

1.3 The Explosion of Available Time Series Data

Over the last 15 years, there has been an explosion in the amount of time series-based data available to businesses. To name a few, Global Insights, Euromonitor, CMAI, Bloomberg, Nielsen, Moody's Economy.com, Economagic—not to mention government sources such as www.census.gov, www.statistics.gov.uk/statbase, www.statistics.gov.uk/hub/regional-statistics, IQSS database, research.stlouisfed.org, imf.org, stat.wto.org, www2.lib.udel.edu, and sunsite.berkeley.edu. All provide some sort of time series data—that is, data collected over time inclusive of a time stamp. Many of these services are available for a fee, but some are free. Global Insights (www.ihs.com) contains over 30,000,000 time series. It has been the authors' collective experience that this richness of available time series data is not the same worldwide.

This wealth of additional time series information actually changes how a company should approach the time series forecasting problem in that new processes, methods, and technology are necessary to determine which of the potentially thousands of useful time series variables should be considered in the exogenous or multivariate in an X forecasting problem (Rey 2009). Business managers do not have the time to scan and plot all of these series for use in decision making. Statistical inference is a reduction process and data mining techniques used for forecasting can aid in the reduction process.

In order to provide some structure to data concerning various product lines consumed in an economy, there has long been a code structure used to represent an economies market. Various government and private sources provide this data in a time series format. This code structure is called NAICS (*North American Industry Classification System*) in North America (www.census.gov/naics)**.** Various sources provide historical data in this classification system, but some also produce forecasts (Global Insights). For global product histories, an international system was recently deployed (ICIS—International Code Industry System). This system is at a higher level than the NAICS codes. For reference, there are cross-walk tables between the two (www.naics.com/). Both of these systems, among others, provide potential Y variables for a corporation's market forecasting endeavors. In some cases, depending on the level of detail being considered, these same sources may even be considered Xs.

Many of these sources offer databases for historical time series data but do not offer forecasts themselves. Other services, such as Global Insights and CMAI, do in fact offer forecasts. In both of these cases though, the forecasts are developed based on an econometric engine versus simply supplying individual forecasts. There are many advantages to having these forecasts and leveraging them for business gain. How to do so by leveraging both data mining and forecasting techniques will be discussed in the remainder of this book.

1.4 Some Background on Forecasting

A couple of important distinctions about time series models are important at this point. First, the one thing that differentiates time series data from transaction data is that the time series data contains a time stamp (day, month, year.) Second, time series data is actually related to "itself" over time. This is called serial correlation. If simple regression or correlation techniques are used to try and relate one time series variable to another, without regard to serial correlation, the business person can be misled. Therefore, rigorous statistical handling of this serial correlation is important. Third, there are two main classes of statistical forecasting approaches detailed in this book. First there are univariate forecasting approaches. In this case, only the variable to be forecast (the Y or dependent variable) is considered in the modeling exercise. Historical trends, cycles, and the seasonality of the Y itself are the only structures considered when building the univariate forecast model. In the second approach, where a multitude of time series data sources as well as the use of data mining techniques come in, various Xs or independent (exogenous) variables are used to help forecast the Y or dependent variable of interest. This approach is considered multivariate in the X or exogenous variable forecast model building. Building models for forecasting is all about finding mathematical relationships between Ys and Xs. Data mining techniques for forecasting become all but mandatory when 100s or even 1000s of Xs are considered in a particular forecasting problem.

For reference purposes, short-range forecasts are defined as one to three years, medium-range forecasts are defined as three to five years, and long-term forecasts are defined as greater than five years. Generally, the authors agree that anything greater than 10 years should be considered a scenario rather than a forecast. More often than not, in business modeling, quarterly forecasts are being developed. Quarterly data is the frequency that the historical data is stored and forecast by the vast majority of external data service providers. High-frequency forecasting might also be of interest even in finance where data can be collected by the hour or minute.

1.5 The Limitations of Classical Univariate Forecasting

Thanks to new transaction system software, businesses are experiencing a new richness of internal data, but, as detailed above, they can also purchase services to gain access to other databases that reside outside the company. As mentioned earlier, when building forecasts using internal transaction Y data only, the forecasting problem is generally called a univariate forecasting model. Essentially, the transaction data history is used to define what was experienced in the past in the form of trends, cycles, and seasonality to then forecast the future. Though these forecasts are often very useful and can be quite accurate in the short run, there are two things that they cannot do as well as the multivariate in X forecasts: They cannot provide any information about the "drivers" of the forecasts. Business managers always want to know what variables drive the series they are trying to forecast. Univariate forecasts do not even consider these drivers. Secondly, when using these drivers, the multivariate in X or exogenous models can often forecast further in time, with accuracy, then the univariate forecasting models.

The 2008–09 economic recession was evidence of a situation where the use of proper Xs in a multivariate in X "leading indicator" framework would have given some companies more warning of the dilemma ahead. Services like ECRI (Economic Cycle Research Institute) provided reasonable warning of the downturn some three to nine months ahead of time. Univariate forecasts were not able to capture these phenomena as well as multivariate in X forecasts.

The external databases introduced above not only offer the Ys that businesses are trying to model (like that in NAICS or ISIC databases), but also provide potential Xs (hypothesized drivers) for the multivariate in X forecasting problem. Ellis (2005) in "Ahead of the Curve" does a nice job of laying out the structure to use for determining what X variables to consider in a multivariate in X forecasting problem. Ellis provides a thought process that, when complemented with the data mining for forecasting process proposed herein, will help the business forecaster do a better job of both identifying key drivers and building useful forecasting models.

Forecasting is needed not only to predict accurate values for price, demand, costs, and so on, but it is also needed to predict when changes in economic activity will occur. Achuthan and Banerji—in their *Beating the Business Cycle* (2004) and Banerji in his complementary paper in 1999—present a compelling approach for determining which potential Xs to consider as leading indicators in forecasting models. Evans et al. (2002), as well as www.nber.org and www.conference-board.org, have developed frameworks for indicating large turns in economic activity for large regional economies as well as for specific industries. In doing so, they have identified key drivers as well. In the end, much of this work shows that, if we study them over a long enough time frame, we can see that many of the structural relations between Ys and Xs do not actually change. This fact offers solace to the business decision maker and forecaster willing to learn how to use data mining techniques for forecasting in order to mine the time series relationships in the data.

1.6 What is a Time Series Database?

Many large companies have decided to include external data, such as that found in Global Insights, as part of their overall data architecture. Small internal computer systems are built to automatically move data from the external source to an internal database. This practice, accompanied with tools like the SAS® Data Surveyor for SAP (which is used to extract internal transaction data from SAP), enables both the external Y and X data to be brought alongside the internal Y and X data. Often the internal Y data is still in transactional form that, once properly processed, can be converted to time series type data. With the proper time stamps in the data sets, technology such as Oracle, Sequel, Microsoft Access or SAS itself can be used to build a time series database from this internal transactional data and the external time series data. This database would now have the proper time stamp and Y and X data all in one place. This time series database is now the starting point for the data mining for forecasting multivariate in X modeling process.

1.7 What is Data Mining for Forecasting?

Various authors have defined the difference between “data mining” and classical statistical inference (Hand 1998, Glymour et al. 1997, and Kantardzic 2011, among others). In a classical statistical framework, the scientific method (Cohen 1934) drives the approach. First, there is a particular research objective sought after. These objectives are often driven by first principles or the physics of the problem. This objective is then specified in the form of a hypothesis; from there a particular statistical “model” is proposed, which then is reflected in a particular experimental design. These experimental designs make the ensuing analysis much easier in that the Xs are orthogonal to one another, which leads to a perfect separation of the effects therein. So the data is then collected, the model is fit and all previously specified hypotheses are tested using specific statistical approaches. In this way, very clean and specific cause-and-effect models can be built.

In contrast, in many business settings a set of “data” often contains many Ys and Xs, but there was no particular modeling objective or hypothesis in mind when the data was being collected in the first place. This lack of an original objective often leads to the data having multi-collinearity—that is, the Xs are actually related to one another. This makes building cause-and-effect models much more difficult. Data mining practitioners will mine this type of data in the sense that various statistical and machine learning methods are applied to the data looking for specific Xs that might predict the Y with a certain level of accuracy. Data mining on transactional data is then the process of determining what set of Xs best predicts the Ys. This is quite different than classical statistical inference using the scientific method. Building adequate prediction models does not necessarily mean that an adequate cause-and-effect model was built, again, due to the multi-collinearity problem.

When considering time series data, a similar framework can be understood. The scientific method in time series problems is driven by the economics or physics of the problem. Various structural forms can be hypothesized. Often there is a small and limited set of Xs that are then used to build multivariate in X times series forecasting models or small sets of linear models that are solved as a set of simultaneous equations. Data mining for forecasting is a similar process to the transaction data mining process. That is, given a set of Ys and Xs in a time series database, the goal is to find out what Xs do the best job of forecasting the Ys. In an industrial setting, unlike traditional data mining, a data set is not normally available for doing this data mining for forecasting exercise. There are particular approaches that in some sense follow the scientific method discussed earlier. The main difference here will be that time series data cannot be laid out in a “designed experiment” fashion. This book goes into much detail about the process, methods, and technology for building these multivariate in X time series models while taking care to find the drivers of the problem at hand.

With regard to process (previously discussed), various authors have reported on the process for data mining transactional data. A paper by Azevedo and Santos (2008) compared the KDD process, SAS Institute’s SEMMA (Sample, Explore, Modify, Model, Assess) process and the CRISP data mining process. Rey and Kalos (2005) review the Data Mining and Modeling process used at The Dow Chemical Company. A common theme in all of these processes is that there are many Xs, and therefore some methodology is necessary to reduce the number of Xs provided as input to the particular modeling method of choice. This reduction is often referred to as variable or feature selection. Many researchers have studied and proposed numerous approaches for variable selection on transaction data (Koller 1996, Guyon 2003). One of the main concentrations of this book will be on an evolving area of research in variable selection for time series type data.

At a high level, the data mining process for forecasting starts with understanding the strategic objectives of the business leadership sponsoring the project. This is often secured via a written charter that documents key objectives, scope, ownership, decisions, value, deliverables, timing and costs. Understanding the system under study with the aid of the business subject matter experts provides the proper environment for focusing on and solving the right problem. Determining from here what data helps describe the system previously defined can take some time. In the end, it has been shown that the most time-consuming step in any data mining prediction or forecasting problem is the data processing step where data is defined, extracted, cleaned, harmonized and prepared for modeling. In the case of time series data, there is often a need to harmonize the data to the same time frequency as the forecasting problem at hand. Then there is often a need to treat missing data properly. This may be in the form of forecasting forward, backcasting or simply filling in missing data points with various algorithms. Often the time series database has hundreds if not thousands of hypothesized Xs in it. So, just as in data mining for transactional data, a specific feature or variable selection step is needed. This book will cover the traditional transactional feature selection approaches, adapted to time series data, as well as introduce various new time series specific variable reduction and variable selection approaches. Next, various forms of time series models are developed; but, just as in the data mining case for transaction data, there are some specific methods used to guard against overfitting, which helps provide a robust final model. One such method is dividing the data into three parts: model, hold out, and out of sample. This is analogous to training, validating, and testing data sets in the transaction data mining space. Various statistical measures are then used to choose the final model. Once the model is chosen, it is deployed using various technologies.

This discussion shows how and why it is important that the subject matter experts' knowledge of a company's market dynamics is captured in a form that institutionalizes this knowledge. This institutionalization actually surfaces through the use of mathematics, specifically statistics, machine learning and econometrics. When done, the ensuing equations become intellectual property (IP) that can be leveraged across the company. This is true even if the data sources are in fact public, since how the data is used to capture the IP in the form of mathematical models is in fact proprietary.

The core content of the book is designed to help the reader understand in detail the process described in the previous paragraphs. This will be done in the context of various SAS technologies, including SAS® Enterprise Guide®, SAS Forecast Server and various SAS/ETS® time series procedures like PROC EXPAND, PROC TIMESERIES, PROC ARIMA, PROC SIMILARITY, PROC Xll/12, as well as the SAS® Enterprise Miner™ time series data mining nodes, and others.

1.8 Advantages of Integrating Data Mining and Forecasting

The reason for integrating data mining and forecasting is simply to provide the highest-quality forecasts possible. Business leaders now have a unique advantage in that they have easy access to thousands of Xs, and the knowledge about a process and technology that enables data mining on time series data. With the tools now available through various SAS technologies, the business leader can create the best explanatory (cause and effect) forecasting model possible, and this can be accomplished in an expedient and cost efficient manner.

Now that models of this type are easier to build, they then can be used in other applications, including scenario analysis, optimization problems, and simulation problems (linear systems of equations as well as non-linear system dynamics). All in all, the business decision maker is now prepared to make better decisions with these advanced analytics forecasting processes, methods and technologies.

1.9 Remaining Chapters

The next chapter defines and discusses in detail the process of data mining for forecasting. In Chapter 3, details are given about how to set up an infrastructure for data mining for forecasting. Chapter 4 covers issues with data dining for forecasting applications. This then leads to data collection in Chapter 5 and data preparation in Chapter 6, which has an entire chapter dedicated to the topic since 60–80% of the work lies in this step. Chapter 7 discusses the foundation for the actually doing data mining by providing a practitioner's guide to data mining methods for forecasting. Chapters 8 through 11 present a practitioner's guide to time series forecasting methods. Chapter 12 finishes the book by walking through an example of data mining for forecasting from start to finish.