Abstract

SCHERTEL, STACEY LEE. *DATA MINING AND ITS POTENTIAL USE IN TEXTILES: A Spinning Mill. (Under the direction of Dr. George Hodge and Dr. William Oxenham)*

The purpose of this research has been to understand the possible uses of data mining in the Textile Industry, specifically a spinning mill. There is a lot of information published on the theory of data mining, however there is not a lot published on its use in a manufacturing setting. A case study approach was used to help understand how data mining could be used in the manufacturing of textiles. The focus of this research was on a spinning mill in the textiles industry and the processing that is followed with the different data elements available.

Data was collected from a spinning mill operation and then cleansed and merged to create a data warehouse that could be mined using the SAS Enterprise Miner software.

An example ideal data warehouse was created for a spinning mill. In this warehouse the different elements and formats that are needed were listed for each process in the production of a cotton fiber.

Initially a simple data mining process was used however this proved to be ineffective. Due to the successes and failures that were experienced during the research a new data mining process model was created. This model has six major steps, which contains a total of 28 specific activities that may be included in the data mining process model. The proposed model describes how data mining can be implemented in a manufacturing setting.

DATA MINING AND ITS POTENTIAL USE IN TEXTILES: A Spinning Mill

by
Stacey L. Schertel

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Textile Technology Management

Raleigh, North Carolina

2002

APPROVED BY:

_____	_____
Dr. George Hodge Co-Chair	Dr. William Oxenham Co-Chair
_____	_____
Dr. Nancy Cassill Committee Member	Dr. Thomas HoneyCutt Committee Member

To Everyone that believed in me

Biography

Stacey Lee Schertel was born on July 23, 1973 in Woburn, Massachusetts. She grew up in the neighboring town of Burlington, Massachusetts and attended the local high school. She graduated in 1991. After completion of high school she decided to attend the University of Rhode Island studying Textile Marketing. After two years she transferred to North Carolina State University where she earned her Bachelor of Science in Textile and Apparel Management. She graduated Magna Cum Laude in 1995.

L.L. Bean employed her directly following graduation. She worked there for two years and two months. She went through several promotions at L.L. Bean. Upon leaving her final title was Production Engineer. Her goals while at L.L. Bean had shifted and therefore she found herself applying to the Master's program at North Carolina State University. She started her quest for a Master's degree in the fall of 1997. She completed this degree in the December of 1998. Beginning in 1999 the pursuit of a PHD degree began. During both of these degrees, from 1997 to 2001, she held a research assistantship at the College of Textiles, helping Dr. George Hodge and Dr. William Oxenham on a project funded by the National Textile Center.

During the summer of 2001 she married Damon Dowell on Cape Cod where they will take residence upon completion of this degree.

Acknowledgements

I would like to thank my co-advisors, Dr. George Hodge and Dr. William Oxenham, for their guidance and input throughout my research work I would also like to thank Dr. Nancy Cassill because without you I never would have gotten through the A1 paper. I would also like to thank Dr. Thomas Honeycutt for his support and feedback in my research work.

The author would like to thank the company that was willing to participate in the research project. The time of the employees there and their willingness to assist in the research was greatly appreciated. I would also like to thank Dr. Jerry Oglesby at SAS Institute for his assistance in the actual mining expedition.

Finally, and above all else, the author would like to thank all her family members and friends who supported her quest for her PHD degree. Without her husband Damon's constant support and loving words she may never have gotten through this degree. She would especially like to thank her loving mother and father, Linda and Robert, and her sister, Lisa, and her husband, Steve, for their support and encouragement. In addition she would like to thank her "second parents", Paul and Mary, for their words of encouragement and loving advice when things were tough. A special thanks goes to her new in-laws for their understanding, support and words of wisdom in the PHD process. Without all of you, my extended family, this degree would not be possible. Finally, thanks to Abercrombie and Sable for all their loving during this difficult time.

THANKS!!!!!

Table of Contents

List of Figures

List of Tables

1.0 Introduction

In the information technology age that has emerged in the last 10 to 20 years, industries have seen an influx of data that is "generated", but for the most part not fully utilized. In the textile arena this is certainly the case according to the literature available on the subject, very little is written about textiles and data mining. In most mills the data is used for day-to-day evaluation of how a plant is running. Printouts are reviewed to see how the plant is performing, but this data is not used any further.

The purpose of this research is to understand how data mining can be used in Textiles to sift through the massive amounts of data that are being produced each day in this industry to help make more informed decisions. A case study approach will be used in this research, specifically a spinning mill will be evaluated.

As can be seen in the literature review section of this dissertation, there is information based on theory available on this topic. However, in all of the literature that is available there is very little that discusses how data mining can be used in a manufacturing setting. There are examples of how data mining has worked for customer related industries like: fraud detection, insurance, hotels and many others. However, the information that is published does not really explain how the process works. In some cases it describes the process as a black box. The data is inputted and then the outcomes or nuggets of information are outputted from the system. The literature leads individuals to believe that it is as an easy task with no real work involved. The main difference between customer oriented and manufacturing data is that in customer related industries databases have already been established however; the databases still need to be cleaned. In manufacturing environments the process is very different. The gathering of data has to

be determined, production time lags need to be established and the overall database needs to be created. These types of issues will be discussed in this dissertation along with a new process model that has been produced to help manufacturing environments mine their data. Some topics that can be the focus of a data mining project are: what raw material is appropriate for the end product that is being produced, is the equipment efficient, and how is the process performing to create an quality end product.

In Chapter 2 the literature review discusses the basics about data mining. Various definitions of data mining are presented and it is apparent from this review that there is some ambiguity as to the "definition" of data mining. Applications of where data mining has been used are reviewed. The techniques that are present in data mining software packages are discussed. The vendors and their software packages and techniques are compared and the overall process steps are exposed.

Chapter 3 covers objectives of the research. In this chapter the purpose of the research and how the goals of the project will be accomplished are shown. In this chapter there will also be information about the facility that was chosen for the analysis and how it runs. The process used to mine the data, which will be followed for this research, will also be explained. The research hypothesis that is hoped to be proven is laid out to help get a better understanding of what data mining might show from the manufacturing data.

Chapter 4 shows the cleansing process for this case study. In this chapter the different issues that were faced will be shown to help get a feel for what must be accomplished if data mining is to be used. It can be seen in this section that it is not a simple process nor is it a short process. At the beginning of this chapter a glossary of abbreviations and definitions will be given for reference to understand the data involved.

In some cases there will be a full definition while in others the abbreviation will simply be explained.

In Chapters 5 and 6 the actual effort of mining the data will be shown and then discussed. In the first of the two chapters the analysis using the chosen software package, Enterprise Miner – SAS, will be explained. In chapter 6 the results will be discussed and explained. Chapter 7 will introduce the Proposed Data Mining Process Model with explanations on how it is organized. In this section there will be five Phases. Each phase will show what must be completed to move on to the next phase of the process. Finally the conclusions of the research will be stated in Chapter 8 along with suggestions for future studies.

Glossary

"***Classification Tree*** is a decision tree that places categorical variables into classes" [57].

Clustering uses algorithms to help distinguish groups that are similar in nature [57].

Cross selling is determining whether a customer, who has already purchased certain articles, is likely to acquire others (and in what combinations)[22].

Customer lifetime valuation is based on repeat purchases, dollars spent or longevity. From this data mining a company can predict who will become their most valuable customer [22].

Customer segmentation determines what the common characteristics are about their customers. Data mining can help to determine whether they fall under identifiable groups [22].

Data Cleansing is the process of ensuring that all information in datasets are consistently and correctly inputted [59].

Data mining is an analytical tool, a computer software package that is used to sort through data to determine trends, relationships or profiles.

"***Data Visualization*** is the visual interpretation of complex relationships in multidimensional data" [59].

Data Warehouse is a computer based storage system enabling massive quantities of data to be pooled with easy accessibility in a way that is consistent with the organizations' needs [21].

Decision Trees are tree shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Such methods include: Chi Square Automatic Interaction Detection (CHAID), Classification Regression Trees (CART) [59].

Missing Data is data that was missed either because it was not measured, not answered, it was unknown at the time, or it was lost [59].

"***Nearest Neighbor Method*** is a technique that classifies each record in a dataset based on a combination of the classes of the k record most similar to it in a historical dataset" [59].

"***Regression tree*** is a decision tree that predicts values of continuous variables" [59].

2.0 Literature Review

Data mining is an analytical tool, a computer software package that is used to sort through data to determine trends, relationships or profiles [47]. Why should business use this technology? In today's environment of constant changes and new technologies, it is important to understand what the drivers and enablers are of the industry. Drivers are what forces an organization to change while the enablers help the organization to accomplish the actual change. Data mining is not the only driver and enabler present in an organization, it is only one, among many others in today's changing world. Data mining can be considered an enabler because of the need to analyze data to make a company more competitive in the marketplace. The enabler is the means to implement data mining techniques and these are becoming increasingly more technically advanced in new technologies like databases and visualization techniques [8]. The driver of data mining is what comes out of the process of data mining that helps managers make decisions. "Knowing your customers, what they want and need is what tops the priority list for most businesses today" [27] and because of this, data mining is very important for a company to be competitive. The notion of data mining has become more important over the course of time due to the enormous amounts of data that are created in today's business environment. Traditional manual approaches are ruled out because they can no longer perform the analysis in a timely fashion. Humans can deal with three to five dimensions but when more than 10 are used it becomes much more difficult. This is partly why it is important to develop new technologies that can handle this data [14].

Companies have discovered that the ability to derive knowledge from customer related information could help lead to a competitive advantage. There is no shortage of

raw data in today's business environments. What there is a shortage of is how to use this data to be more competitive and in making products more appealing. It is also important for a company to use the data obtained from the process in a timely fashion. For instance, if the manufacturer of your product takes one week from start to finish, it would be imperative to have quality information about this product while it is still being produced. The information is no longer helpful to producing a quality product after it has been completed, shipped and already arrived and is on the shelf at your customer's facilities. If it takes two months for the statistician to find trends about this data, then in most cases it is too late to save that company from possible disaster. However, if a data mining tool can look at the data during the process and show that similar trends have resulted in defective product then this would be valuable information to the producers and possibly save them a lot of money and retain customers' loyalty.

Some of the applications of data mining are in customer retention, customer acquisition, customer lifetime valuation, customer segmentation, cross selling, response modeling and other special niche applications like National Basketball Association scouting. In the textile industry today most of the articles that have been written are showing data mining being used for forecasting, production planning, promotion and distribution. However, for the most part this is only done at the retail level. Very little has been found at the manufacturing end of the textile industry. One example is in the production of mono-filament nylon fibers. In this data mining case the company used classification and regression trees to help determine spin breaks during production [30]. The analysis utilized data that was available from both on and offline measurement and testing. This is similar to what is hoped to be accomplished in the present research,

however a very different production process is followed for a natural fiber versus that of the man-made nylon fibers.

Some of the various definitions of data mining offered are as follows:

- In "Solid State Technology", data mining is defined as “a methodology that finds hidden patterns in large sets of data to help explain the behavior of one or more response variables” [46].

- Jacobs defines data mining “as a process of analyzing data to extract information not offered by the raw data alone” [21].

- Davis says that “data mining uses mathematical algorithms to search for patterns within large volumes of data that are related to business issues. Data mining is a way of discovering hypotheses not verifying them. It also takes human skill to interpret the results accurately. It is thought that data mining is not only a science but also an art” [10].

- DuMouchel says that “a common task of data mining is to search for associations within large databases” [13].

- Hand describes data mining as “a new discipline, lying at the interface of statistics, database technology, pattern recognition, and machine learning, and concerned with the secondary analysis of large databases in order to find previously unsuspected relationships which are of interest or value to the database owners” [19].

- “Data mining gives you the ability to sift through thousands of potential variables to isolate key variables that are highly predictive” [25].

- The SAS organization has defined data mining as “the process of selecting, exploring, and modeling large amounts of data to uncover previously unknown patterns of data for business advantage” [47].

- Alexander says that “data mining is the next step beyond online analytical processing (OLAP). It sifts through data for unknown relationships” [1].

- The University of Birmingham’s Computer Science website has defined data mining as “the nontrivial extraction of implicit, previously unknown, and potentially useful information from data. It uses machine learning, statistical and visualization techniques to discover and present knowledge in a form that are easily comprehensible to humans” [48].

- “Data mining helps to find trends in data that human would never dream of looking for” [7].

- "Data mining is the use of an appropriate set of technologies to exploit patterns of information from massive customer focused databases. However, data mining is not a single technology. Rather, it is a collection of tools that are used to extract information from data. Data mining is not just a technology but it is also a process. It cannot be fully automated as individuals must be active in the process to be sure that the information that is extracted is accurate. In other words the results must be relevant and meaningful to the business [15].

Data mining is therefore a way to analyze the enormous amounts of data that is being generated in today's industries to find trends and relationships that were not previously known to exist. The main reason this new technology works is because of the high-powered computers that are available today at a reasonable price coupled with sophisticated software. Also with the Internet we are able to get more data points than what was ever possible in the past. In the new environment of the world wide web and order processing on line there is an extreme amount of data that can be obtained. While at the same time it can be transmitted from computer to computer with minimal time and effort.

Aspects that must be considered in data mining are:

- Who uses different data and information? (e.g. do different levels of workers from the mill floor to senior management need access to the same information?)
- What format is the data in? (can be obtained from different monitoring systems and can be in different end formats such as: Excel, Word, or only printable formats)

The rest of the chapter reviews the current state of the art of this analytical tool.

2.1 Introduction to Data Mining

Data mining is an analytical tool, usually implemented using a computer software package, that sorts through data to determine trends, relationships, or profiles. Specific data mining tools consist of existing statistical models for traditional statistical analysis such as: clustering [15,36,53,58], linear regression [15,42,53,56], neural networks

[1,15,17,25,36,41,42,46,53,56,58], bayesian networks [25,6], visualization [15,58] and tree based models [1,15,17,25,36,41,42,46,53,56,58]. Many of the applications that are used to data mine are from statistical methods that have been around for years. However, the main difference in today's data mining technology as opposed to old methods is that a high-speed computer is used to find trends and relationships in large amounts of data in a minimal time frame. Without computer and optimized software a traditional statistical analysis of the same data may take an individual month to come up with the same answer, if they ever find the same relationship. The reasoning for this is that data mining can easily find predictive trends and relationships while at the same time they are making a sweep through the data to find hidden patterns all in the same step. This data mining finds hidden trends with minimal time and effort.

A data warehouse is a computer based storage system enabling massive quantities of data to be pooled with easy accessibility in a way that is consistent with the organizations' needs [21]. A data warehouse facilitates the storage and the organization of the data so that it may be readily accessed. Data warehouses became available in the early 1990's. Data warehouses do not sort the data into useful trends, relationships or profiles; it is just a database filled with potential information. It is the use of techniques such as data mining that allows us to sort through the data to research information. In order to acquire the data for the data warehouse, it is important to identify what data is required and where it currently resides. Many times the data that is required will be on different systems and therefore these are all in different formats. This necessitates the use of a cleansing process prior to moving the data into the storage phase. Once the data has been conditioned and correctly stored in the warehouse, it can then be analyzed. It is

the analysis, which reveals trends that enable management to answer business questions based on stored data. In order to develop links between data and/or groups of data according to some specified parameter, it is necessary to use specialized software, of which data mining performs a particular function. According to W.H. Inmon, the creator of the term data warehouse, "a data warehouse is a centralized, integrated repository of information, which means data is cleaned up, merged and redesigned. This can be more or less complicated depending on where all the data is being fed in from one or many different systems" [33].

2.2 Current Signification of Data Mining

Ernst and Young conducted a survey on where data mining is being used and what types of data are being stored in data warehouses in 1999. It was found from the companies that were surveyed, 62% are currently mining data. If the company revenues were less than $5 billion, there was a change in this figure. In companies of this size, 63% of them said they have no plans of using data mining. Conversely, of companies that have revenues greater than $5 billion, 75% of them say they are currently data mining. The top three types of information that are stored in the warehouse are: sales (88%), merchandising (84%) and financial information (72%) [45]. However there is also information stored about planning/forecasting (56%) and operational performance information (54%). These data warehouses are used currently for more broad decision making instead of tactical solutions. Companies look to their data warehouses for query and reporting analysis, executive decision-making, and regional performance reporting and store operations [45].

The META Group also did a survey on data mining. They surveyed, in 1999, any company that currently has data warehouses to see if these companies also plan to purchase data mining tools. Of the companies surveyed, 54% of them said that they do plan to purchase data mining and knowledge discovery tools this year. This is a 20% jump since 1996. However, only 8% of those that were surveyed are currently using data mining software [26]. These numbers seem to be fairly low, the main reason for this is the simple fact that it has not been proven yet as a viable solution. Companies selling data mining tools have not yet established themselves as leaders in the field and therefore companies interested in possibly using these types of products are skeptical. Data mining will not grow until customers are convinced that it will improve business [33].

While there seems to be a disparity in the results of these surveys, there are several possibilities why this may be so. Each survey questioned different companies, the questionnaire may have had very different questions, or the timeframe of the actual survey may have caused enough difference in the mind frames of those companies questioned. However, from the articles there is no obvious justification for the difference in results.

2.3 Requirements of Data Mining

Data mining requires that a database is present, which can also be referred to as a data warehouse, and a software package.

Requirements for data mining are as follows:

- data availability
- efficient access methods
- robustness to data problems
- efficient algorithms
- high performance application server
- flexibility in delivering results [46]

Other requirements of data mining are that the data must also be "cleaned" to help eliminate getting false signals due to outliers [46]. It is important to know what the plans for the data are before trying to set up a data warehouse effectively. If the right data does not exist and the limitations of the data are not known then the software can lead companies and people astray [37]. As Alexander says in his article, "garbage in, garbage out."

2.3.1 How to Condition the Data

In a business environment it has been stated that whomever controls the critical information can then leverage them to create profitability [38]. The idea of collecting this information and using it is not a new one. However, there are new techniques available to do this for instance, data warehouses, data marts and then mining the data to make critical decisions that have been made before without the knowledge from analyzing, massaging and delivery useful information from otherwise unconnected raw data [38]. Examples of how to do this are discussed further in the following paragraph.

The first line of defense to receive quality data is to be certain that at the data entry stage of the process all fields have been inputted completely and correctly. If this is done then it may help to eliminate the likelihood of receiving bad data. This is the easiest stage at which to correct data because it is still fresh in the mind and once this data works its way downstream it will take much more time and effort to correct errors. In some cases it might be important to narrow the range of the data that is being entered and to use integrity constraints. For instance, if a survey has answers of only 1 to 5, an integrity constraint will ensure that no other answer can be entered. As with a spell or grammar check not all of the mistakes are caught. If a word that exists is placed in a sentence and

it is not misspelled, the spell checker will not pick it up and therefore it is not correct but it still exists. This is the same for data entry. An individual can by mistake enter a 2, which is in the 1 to 5 range, although it should really be a 3. This would be considered an error and depending on the magnitude of this error it may or may not be caught. For instance, if looking at a disparity in the results of entering a 2 when it should be a 2.1 this probably would not be caught but if it should be a 2 and a 3 is entered this may show up in the results as an outlier. It is not uncommon for there to be inconsistencies or holes in the data. This often requires the use of some standard default values for the cleansing process. An example of this would be the color "000" on an item that has no color. Because color was not used for the product, it may be blank or hold zeros in some systems. If this color code does not exist in the color description table, then the cleansing process should have instructions such as loading a description of "Unknown Color" or "No Color". It is important to have default mechanisms so that there are no voids in the data even when there is nothing to enter into that particular field. For instance, if there is a blank in the data field then there was no data. For example, perhaps there was no inspection on that particular day. However, if there is a "0" in the field then the value is "0". These are two totally different values and should be looked at this way. The blank should not be assumed to be a "0" value [38].

It is extremely important that the data being collected is quality data. This means the data is accurate, complete and relevant. If this is not the case then the data is not useful to the business. Bad quality data will give false signals. How can a business really be sure the information collected is quality data? There is no way to ensure that the data is quality data because the data that is gathered and collected may not be what is

needed to answer the "question" the company wants to answer. For instance, if I want to know how many students are from India then the nationalities of the students must be contained in the data. No matter what it is still important to cleanse the data or a correct "answer" still will not be found [38]. The data that is stored in the data warehouse needs to be cleansed and de-duplicated. This all needs to be done in a repeatable cycle whether it be hourly, weekly or monthly [39].

2.3.2 Different techniques that can be included in data mining

Han states that present mining models lack human involvement and user control. It is believed that data mining is most effective when the computer searches large databases or counting [18]. However, there are many techniques that are involved in data mining that mix human involvement in a computer environment quite extensively in the world of statistics. Some of the techniques that are incorporated in data mining software packages will be discussed in this section of the dissertation.

2.3.2.1 "Traditional" statistics

Traditional statistics includes such things as: regression, cluster analysis and time series forecasting. Regression is usually used when for building predictive models that forecast some outcome and generate a mathematical algorithm [42]. It can be used when dichotomous (yes/no) variables are being predicted. It is also used to analyze continuous ratio measures such as revenues or profits [42]. Cluster analysis is the process of separating a data set into components that reflect consistent patterns. The clustering of data into sets allows for the data to be put into understandable profiles. This is used typically in marketing to determine who has ordered before or not and are they worth

sending a new catalog. Time series analysis is the comparison of time periods for instance, year to year or month to month.

2.3.2.2 Neural networks

It is claimed that "neural networks are not true data mining" [46], it is only a way of modeling the relationships. Neural networks do not find hidden trends or relationships within the data. The real use of the neural network is to model large amounts of data in a physical model [46]. However, another source says that neural networks are the most common data mining technique. This technique generalizes patterns inside the database for classifications and predictions [5]. Neural networks are used quite extensively in the financial industry because these tend to only model and predict trends [16]. Neural network algorithms quantify only and because of this these are not appropriate to use on qualitative data. It is best to only use neural network algorithms for data that is quantifiable. Neural networks are very good at detecting non-linear relationships and are also good at combining information from many input parameters. Therefore neural networks work well when the predictors are partially redundant [15]. Some consider neural networks to be programs that mimic the brain's ability to learn from its own mistakes [1]. Neural networks are therefore non-linear models that learn through training and they closely resemble biological networks. These work best on small to medium datasets. The main issue with Neural networks is that they are considered to be the black box operation because there is no way to get an explanation of what happened in the process and because of this it inhibits the confidence about the final results.

2.3.2.3 Decision trees

A decision tree is the process of taking the data and constructing a tree that depicts the relationships into classifications. These classifications are typically depicted

by nodes, which signify the data it represents. Decision trees are tree shaped structures, like a family tree, that represent sets of decisions. By completing the different steps in making the decision tree a basic set of rules are generated. Decision trees are useful for several different reasons they are very good at detecting non-linear relationships. They are good at selecting important variables and therefore work well when many of the predictors are irrelevant. Basically decision trees are branching systems that show relationships in the form of a hierarchy, such as an organizational chart [1]. The major advantage is that they are faster and more understandable than neural networks. However, since the methodology uses "if-then" statements to categorize the data it is important for the data to be either categorical or interval. All continuous data needs to be put into categorical data. For instance, if States in the United States are chosen then 50 different columns would be needed that list all the States and only one of these columns would have a yes while the rest would be no's. This type of data can cause problems because of all the extra data that is not needed to get the same answer.

2.3.2.4 Bayesian networks

Bayesian networks are similar to decision trees in that these are also branching structures. However, each circle or node, is a probability of the one that precedes it. In other words one event is conditional on the probability of the event that was before this event. The Bayesian network is based on the Bayesian Inference Theorem. All of the probabilities within this network are conditional. These specify the degree of belief in some proposition based on the assumption that some other propositions are true.

Bayesian networks and decision trees can detect hidden relationships and trends within large amounts of data. These are capable of discovering new relationships and therefore these are true data mining techniques [46].

2.3.2.5 Visualization

Visualization is a technique that is used to make it possible for the analyst to get a deeper understanding of the data. It is often difficult to analyze data when it is in rows and columns so visualization is key to understanding vast amounts of information. Visualization is a way to see the trends and profiles easier. Visualization uses graphical tools to illustrate the data relationships this can range from a simple scatter plot to complex multidimensional representations. In other words a picture is worth a thousand words or in this case data points.

2.4 Where is Data Mining Being Used?

The major acceptance of data mining is in insurance, health care providers, finance, credit card companies, telephone companies, retail and marketing processes [1,15,25,26,47,58]. Data mining is also currently in hotel, catalog, supermarket, rental car, the Internet or E-Commerce, and airline industries [15,25,26,47,58]. These industries use data mining through frequent shopper, flier or rental card clubs (marketing programs) to determine customer's preferences. From this data, they can decide what promotional fliers can be sent to a particular customer. The main reason these industries are using data mining is because they have large quantities of data on customers, products and transactions and need to be able to understand the value within this information [1]. They are looking for behavior patterns, preference patterns, any relationships that are not obvious to humans either because the linkage is obscure or because the volume of data to

sift through is simply too overwhelming [15]. Some of the current applications of data mining are briefly discussed below [22].

2.4.1 Customer Retention

It has been determined in different studies that to acquire a new customer can cost a company between 4 and 10 times that of keeping a current customer [15]. This is the driving force behind the loyalty programs that many companies institute. Many loyalty programs reward customers for their continued business. Basically, the premise behind this is that if a customer is happy they will continue to give that company business. If for instance the customer gets no discount or free gift from one company they may be less likely to purchase there if they can get free stuff or discounts from one of the companies competitors [15, 22, 26, 42, 47]. Another dimension of customer retention is cross selling. This will be discussed in more depth later in the dissertation.

2.4.2 Customer Acquisition

Although customer retention is critical to an organization it is also important to acquire new customers. Attracting new customers is what helps the organization grow. The main issue with this is how to acquire customers efficiently. For instance, it is not profitable to send a sale catalog of kids' products to families that do not have children. The costs of acquiring a customer are astronomical so it is a necessity to find a way to obtain customers with as little cost to the company as possible. It is important to profile and segment customers to better understand them. By profiling the customers that are already loyal they can determine what segment of the population that is most likely to buy products from them in the future. From this companies can start to focus their future mailings towards customers that would fit their different types of product categories. It is

essential to market towards customers that would more likely be interested in your products than those that are not in order to acquire a new customer proficiently [15, 22, 26, 42, 47].

2.4.3 Customer Lifetime Valuation

Customer lifetime valuation is based on repeat purchases, dollars spent or longevity [22]. From this data mining, a company can predict who will become their most valuable customer. A company can also use this to help get their customers to purchase at a higher level [42]. Banks use the information that is generated from data mining to determine who their most profitable credit card customers are or possibly their highest risk loan applicants [1]. It can also be used to detect fraud, trend analysis and analyzing profitability. NeoVista Solutions has a knowledge discovery technology that helps retailers leverage store and customer level data to uncover important sales patterns and consumer preferences. The patterns that are determined from this system are then transformed into models to help improve targeting [2].

2.4.4 Customer Segmentation

Customer segmentation determines what characteristics are common to their customers [26]. Data mining can help to decide if they fall into identifiable groups [22]. Data mining is used to help companies better create a more successful marketing strategy. Data mining techniques can be used to identify or target markets, determine what the needs and wants of the customers or businesses in these markets are, and give the marketing strategists valuable information from which intelligent information can be derived. It is a way for companies to help retain customers by using loyalty programs. The greatest benefits are usually achieved by enabling organizations to better understand

individual's wants and needs [4]. Data mining tools do not ensure a successful marketing plan but it does greatly enhance the probability of a successful one [42]. An insurance company in California used IBM's software to mine data about sports car owners. The insurance company was able to extract information about a group of sports car owners that were in the age group of 30 – 50 years old, married and owned two cars. From this data mining expedition, the insurance company decided that this group was not high risk and was able to offer a better rate to their customers that fit this profile. The real importance of this is to identify the not so obvious and offer something their competitors are not [10].

2.4.5 Cross Selling

Cross selling is a further dimension of customer retention. Cross selling is the expansion of products sold to a customer. The additional products are somehow related to other products that should go together. Historical data is used to determine which kinds of products are typically bought in conjunction with each other. If, for example, an individual were to purchase three particular products will they also be likely to purchase additional products X and Y (both) or maybe only X. This is a technique of selling additional products or services to existing customers. For example, if an individual purchases a printer they are probably also likely to purchases paper for that printer. In other words it is a linkage between different products. The customer retention side of this comes into play when the company knows what the customer typically purchases and therefore sends them fliers about additional items that may be discounted when purchasing their regular product. This information is collected on all of the companies

clients in order to data mine it to find a trend in the purchasing behavior of their customers [15, 22, 26, 42, 47].

2.4.6 Response Remodeling

Catalog companies also use a technique similar to data mining to help them decide what customer should or should not receive a particular catalog. If a customer never orders from the summer catalog, then the company should not waste their time and money continuing to send catalogs to that customer when there could be another potential customer who may spend hundreds of dollars if they received the same catalog. This can be referred to as response modeling. Response Modeling can help in identifying a community as a retirement one and then not send any ads out that would be for back to school sales since these would be ineffective [20]. At the same time, they're using data mining tools to help "drill" through a data warehouse to help seek trends and form hypotheses based on the character of the data. This is something that was thought to be an impossible task. It also helps to see if any profiles have changed [20]. Other companies are using this type of software to help establish point of sale trends that may have been overlooked otherwise [6]. Some companies are using point of sales forecasting to help accurately determine sales at the store level [24].

2.4.7 Special Application

IBM has created a program that specifically mines data for NBA (National Basketball Association) scouts. It is called the Advanced Scout. It will find patterns in the massive amounts of statistical data from games and condenses it into useful pieces of information that typical box scores do not show. It helps to establish non-obvious patterns

of the players which helps coaches to learn more about their players and their opponents [3].

2.5 Summary of problems Data Mining Solves

Data mining has been used to overcome a wide range of business issues and problems. Some of the problems include how to:

- Segment customers accurately into groups with similar buying patterns.
- Profile customers most effectively for individual relationship management.
- Dramatically increase response rate from mailings.
- Profile customers to help identify which ones have been the most loyal or are the most likely to respond to certain promotions.
- Understand what motivates customers to leave a company for its competitor.
- Uncover factors affecting purchasing patterns, payments and response rates.
- Predict whether a credit card transaction or an insurance claim will be fraudulent.
- Predict whether credit card customers are likely to transfer their balances to another company within some given timeframe.
- Anticipate customers' future actions, given their history and characteristics.
- Help medical centers and insurance companies better manage costs by determining which combination of procedures will produce the most favorable outcomes [15].

2.6 Areas in Which Data Mining will Produce Good Results

As with any new technology there are always doubts about whether or not it is suitable for any particular circumstance. While no technology will answer every business problem, data mining can be useful for certain things. Data mining is successful at solving problems with these characteristics:

- Have vast amounts of data available,
- The data consists of many variables,
- The data is complex, multivariate, and non-linear,

- Need to predict behavior or outcomes,
- Must find associations and relationships that are not currently understood.

2.7 Who are the Vendors?

IBM: Intelligent Miner, Oracle (Thinking Machines): Darwin, SAS Institute: Enterprise Miner, Angoss International: Knowledge Seeker, and NCR Corporation: TeraMiner Stats are the most recognizable data mining vendors. Other vendors, who are not as recognizable in the literature but they also offer packages for data mining include: DataMind, Pilot Software, Business Objects, NeoVista Solutions, Magnify and Cognos [16].

Table 2.1 shows the major players, the platforms their product functions on and what types of techniques are present in their data mining software products.

Table 2.1: Data Mining Vendors Compared

Vendor: Product Name	**IBM:** Intelligent Miner	**Oracle:** Darwin	**SAS Inst.:** Enterprise Miner	**Angoss Intl.:** Knowledge Seeker	**NCR Corp.:** Teraminer Stats
Platform	AIX 4.1, NVS, AS/400, Windows NT	Windows Windows NT	MacIntosh, Windows NT Unix,	Windows, Windows NT, UNIX	Windows NT, UNIX
Decision Trees	X	X	X	X	
Neural Networks	X	X	X		X
Time Series	X		X		
Prediction	X	X	X	X	
Clustering	X		X		
Associations	X		X		
Bayesian Networks					
Visualization	X	X	X	X	X
Website	[49]	[50]	[47]	[51]	[52]

2.8 Steps in Data Mining

In order for data mining to be successful it is important that it is thought of as a process rather than a set of tools. Therefore it is important that there are steps to follow in the process, and these are as follows:

Step 1: Must have a representative sample of the data that will be explored both statistically and visually. This means that the data that will be collected and then mined must be identified.

Step 2: Apply exploratory statistical and visualization techniques to select and transform the most significant predictive measures. This means that data is sampled to explain visually or numerically any inherent trends or groupings.

Step 3: Model the measures to predict outcomes. At this stage the data is explore both visually and/or numerically for trends and relationships.

Step 4: Test the results to be sure the model is accurate.

Step 5: If needed change the data that is being selected/collected to help focus the model selection process. Changing the data refers to the modification of the data to help focus the model selection process. In some cases it may be apparent that some absent fields in the data set need to be added or deleted.

Always remember that data mining is an iterative process so these steps must be continually used [15].

2.9 What is the Future of Data Mining?

Some believe that the long-term prospects of data mining are truly amazing. There are thoughts about data mining being used on medical research data or on subatomic-particle information. It is thought that maybe computers can reveal new treatments for diseases or new insights on the nature of the universe [11]. In an article by Regalado, the use of data mining to find answers to diseases is starting to take shape. It is also stated that data mining is going to be a core part of drug discovery in the future [32]. At North Carolina State University, Dr. Bruce Weir, is using data mining to look at the

human genome to help reshape how a new medicine is discovered. The hope is to find cures easier and more efficiently. It is felt that even miscellaneous bits of trivia can be turned into power [28].

In an article by Kittler, he states that there must be a better "cleaning" of the data being placed in the data warehouse in order to use it for data mining [25]. He also states that at this time, he does not feel that data mining technology is ready for manufacturing uses. He feels that data mining must have as a starting point derived models that physically relate the association of the various components. There seems to be an opportunity to significantly increase the rate at which volumes of data generated by the manufacturing process can truly be turned into useful information [25].

It is felt that the manufacturing use of data mining is just beginning and, because of this, it is the time to determine how useful data mining will be in the manufacturing of Textiles. This research will outline the current status and future prospects of data mining in textile technology. Some questions to answer would be:

- Is data mining necessary for a textile company to stay competitive (i.e. will it make a company better and more efficient than other textile companies)?
- What role could data mining play in the textile industry?
- How much, if any, is it being used today in this industry?
- How would a textile company get started in incorporating this into their business?

In closing it has been determined through this literature review that data mining, in the sense of finding new trends and relationships that otherwise would not have been found, has not yet be been done in the production of natural fibers in a spinning mill. It has been done by using regressional analysis with statistical methods for instance statistical process control. Because of the competitive edge that data mining will likely

give to companies in the textile industry it cannot be verified for sure that it is not being used presently; however, at this time it has not been documented in a spinning mill for natural fiber production.

3.0 Research Methodology

In this chapter the objectives of the research will be presented. The following subjects will be discussed:

- What were the research objectives
- How will this research be conducted and in what form
- What type of a facility will be analyzed and what does the set up of that company look like

3.1 General Objectives of the Research

The main objective of this research was to explore the data mining application techniques in a textile environment. In this research a case study approach was used to determine the potential benefits afforded by using data mining to analyze factors affecting process and product quality in a spinning mill. Some questions that will be answered are:

- What roles could data mining play in the textile industry?
- How much, if any, is data mining being used today in this industry?
- How would a textile company get started in incorporating this into their business? (How does a company go about it)?
- What is the process of data mining?

The ultimate purpose of this research was to enable the analysis of the plethora of data that is collected from different monitoring and other data input systems. This analysis provides concise information that can in turn be used to determine trends in the spinning process. A prerequisite to success in this project is the development of a standardized and documented process of how to combine the data generated from many online and offline sources, which may be in different formats, into a useable database environment.

3.2 Experimental Procedure

It was decided that this research would be conducted using a case study approach. In this case one company's data will be collected and then analyzed using a data mining software package. The reason for a case study approach is that one can always derive a plan to mine data in a manufacturing environment, but until this has been done there is no way to really fully understand the possibilities of mining data. The reason for only choosing one facility is because of the amount of time needed to clean data and mine it. In the literature there are discussions about companies that take years to gather and clean data in order to mine it [62]. The literature also discusses that it takes 80% of the time to get the data ready for mining [26].

It will be imperative to work through this data mining process in steps. In reviewing the literature it was determined that there should be a procedure to follow for mining data. In this research the steps shown in figure 3.1, that were derived from Thuraisingham's process steps, were followed to help with the organization of this research. The process begins at the bottom of the staircase and works its way upward. What is involved in the research will be discussed in section 3.3.

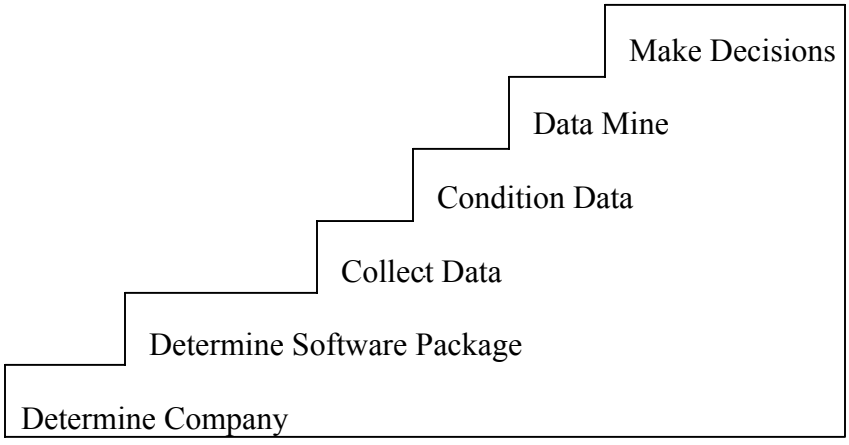


Figure 3.1: Steps to be Completed for this Data Mining Project

3.3 Steps to the Data Mining Process and What These Entail

3.3.1 Determine Company

The first step was determining what type of company to target by looking at different options such as proximity of geographical location; likelihood of their willingness to co-operate; which company already had available a significant source of data? After reviewing what companies were available and willing to co-operate. It was decided that a spinning mill would be the object of the research. At the present time spinning mills are gathering data but it may be possible that the data that is being collected is not useful. It is important to be able to understand and use the data that is being generated by the many different monitoring systems in a mill. In many instances companies are collecting data, but they may not be sure why they are collecting that particular piece of data.

3.3.2 Determine Software Package

The first step of determining the company to study had been completed it was now time to choose the software, step number two. This was not an easy task since from the review of earlier research, it was determined that there are several different mining software packages that could be used for this project. Each of these software packages had there own strengths and weaknesses. Based on the information known about textile companies it was decided to select SAS Enterprise Miner as the software to use. However, the main reasoning behind the choice of SAS was that this software allows for the data to be imported into their software directly from Excel. It was felt that this was an important part of the user friendliness of the SAS software. The reasoning was that all the data that is collected from any company can be downloaded into Excel, which is a fairly well known and used program. This data can then, in turn, be imported into SAS

with a fairly simple macro. The other software package that was also considered was the IBM Intelligent Miner. The main disadvantage of this software was that the user must also learn DB2, their database software, which is very complicated and not user friendly. Both software packages that have been mentioned were available for use from the University and therefore availability of software was not a factor in the decision. However, other software packages that were reviewed were not available and therefore not in the final running.

3.3.3 Collect Data

The next step in this process is to acquire an actual data set that could be mined. However, before this can be accomplished the following must be identified:

- What processes are involved in a spinning plant
- What systems generate the data
- What are the timeframes of the data
- What is the format of the data

See the beginning of Chapter 4 for an explanation of the plant in this research study.

3.3.4 Condition Data

Once the data is collected then it will be conditioned or also sometimes referred to as cleaned. In this case study since there are not earlier works to base the research on the data will be cleaned/conditioned manually. In this approach it will be feasible because only a few months of data will be utilized.

3.3.5 Data Mine

The next step of the process is to take the data that has been collected and cleaned and then import it into the chosen SAS software package and mine the data. At this stage the data will be analyzed using a mining program. A few different analysis tools will be

utilized during the mining, for instance, neural networks, regression models and decision trees.

3.3.6 Make Decisions

The data analysis will be conducted after the data warehouse has been created for the data in the spinning mill. This data will be analyzed using Enterprise Miner Software by SAS that was discussed in Chapter 2. Once this process has been completed then decisions can be made about the usefulness of data mining in a manufacturing environment specifically a spinning mill.

3.4 Set up of the Experiment

The data mining experiment was conducted on a Gateway computer with 500Mhz Pentium III processor. The original datasets were in Excel 1997 by Microsoft before being imported into SAS Enterprise Miner version 8.1. The whole SAS programming system utilizes 814Mb of space on the C drive.

3.5 Possible Outcomes of the Research

The initial intention of this study was to examine relationships that are generally accepted as true in order to verify that the data mining approach was giving valid links. Once this is accomplished it will be used to search for new trends in the data that is currently collected by the spinning mill. Some of the questions of this research are: is it necessary to continue collecting this data? Is all the data important for the company to find the answers to any manufacturing dilemma?

Figure 3.2 gives a conceptual example; this figure is purely what would be considered to be “dream” results, of what might be the outcome of this research. In collecting data there are both inputs and outputs. The data found in Figure 3.2 are a

combination of the two: the fiber information are the inputs while the quality information can be viewed as outputs. From this diagram it is obvious that lots of parameters effect yarn breaks. The real interest is that all of these factors can be actually independent of yarn breaks but are influenced by another parameter which does impact breaks. In effect, do one or more variables directly and indirectly influence this parameter? For example in Figure 3.3 the illustration shows that yarn breaks may be influenced by thick places, thin places, trash sliver, neps, count, etcetera shown by the black lines. However, on closer inspection it can be seen that while micronaire has a direct impact on yarn breaks micronaire also affects thin places, thick places, neps and % CV. Therefore, what may be determined is that in fact micronaire is the key parameter. In essence, it may not be necessary to gather data on thick places, thin places, % CV and neps, because they can be inferred from a measuring micronaire (and if micronaire could be eliminated as a variable then neps may have no effect on yarn breaks).

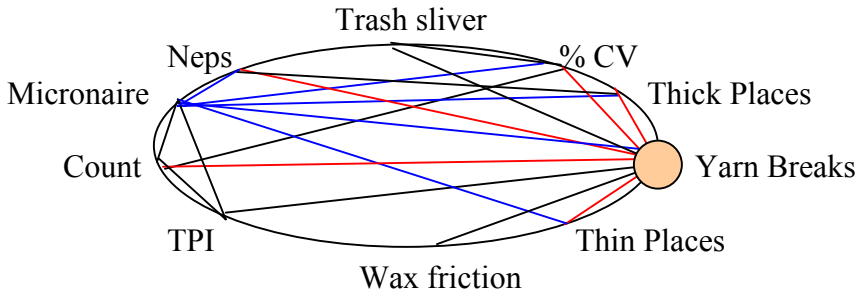


Figure 3.2: Different Data Collected within a Spinning Mill

The following sections discuss the actual case study analysis and results, and draw final conclusions based on the initial research goals.

Glossary and Abbreviations of Terms in Spreadsheets Used in Chapter 4:

AVG: Average

WGT: Weight

S.E.: Single End

Elong: Elongation

Frict: Friction

T.P.I.: Turns per inch

V.L.C.: Variance length curve.

CV: Coeffiecient variance

% : Percentage

HVI: High Volume Instrument data

Laydown Spreadsheet Data Terms:
Layd #: Laydown Number: The incremental number used as an identifier for each 30 bale laydown that is consumed.

Bale #: Is the internally assigned bale number to insure unique file keys.

Bale Count: Number of bales in the laydown.

Layd Pos = Laydown Position: The bale position within the 90-bale laydown.

Mix Cons = Mix Consumption: consumption mix identifier.

Net Weight: Weight of the individual bale.

Mic = Micronaire: Weight per unit length of a fiber.

2.5% SL = 2.5% span length: Length at which only 2.5% of the fiber is longer than that length.

UI = Uniformity Index: Ratio of span length to mean length.

Str = Strength: strength of the fiber in g/Tex.

EL = Elongation: the % of elongation.

Rd = Colorimeter A: The level of grayness in the fiber.

+b = Colorimeter B: The level of yellowness in the fiber.

C-G = Color/Grade: Color grade that a classer would use to classify the fiber.

UV = Ultraviolet reading: This is the ultraviolet reflectance of the fiber.

Card Spreadsheet Data Terms:
Card Number: The value listed is the average of 5 readings from a sample. The date on the report is the date the samples were gathered. The samples are taken from the card mat and the delivered sliver.

Card: Machine number in the plant.

Nep Mat: Number of neps per gram entering the card (within the card mat).

Nep Sliver: Number of neps per gram in the card sliver.

Nep % Elim. = % Elimination: Percent reduction of neps at the card (Mat-Sliver)/100

Dust Mat: Number of dust particles per gram entering the card (within the card mat).

Dust Sliver: Number of dust particles per gram in the card sliver.

Dust % Elim. = % Elimination: Percent reduction of dust particles at the card (Mat-Sliver)/100

Trash Mat: Number of coarse trash particles per gram entering the card (within the card mat).

Trash Sliver: Number of coarse trash particles per gram in the card sliver.

Trash % Elim. = % Elimination: Percent reduction of coarse trash particles at the card (Mat-Sliver)/100

V.F.M. = Visible Foreign Matter: is a calculated value from the dust and trash readings.

V.F.M. Mat: Number of % visible foreign matter per gram entering the card (within the card mat).

V.F.M. Sliver: Number of % visible foreign matter per gram in the card sliver.

V.F.M. % Elim. = % Elimination: Percent reduction of % visible foreign matter at the card (Mat-Sliver)/100

SFC (n): Short fiber content by number.

SFC (w): Short fiber content by weight.

2.5% (n): The upper 2.5 % length of the fibers (only 2.5% of the fibers are longer than this).

Drawing Spreadsheet Data Terms:
MAH: machine

Style: Product (100% cotton) being produced.

Count: Sliver weight (grains/yard).

KLBS: 1000 pound units produced total for shift.

LB/H: Pounds per hour

ACT %: Actual percent efficiency of machine (s).

ST/H: Number of stops per machine hour.

STPH: Mean time stops per hour.

M/D: Stop minutes per doffing or can change.

QS: Quality stops.

A%: Weight variation of the sliver (100yd).

CV%: Weight variation of the sliver (inches).

CVL: Count variation.

TP/KM: Tick places per 1000 meters.

Spinning Spreadsheet Data Terms:
Shift Start: Time of day shift begins (7:00am or 7:00 pm.) it also includes the date.

Yarn Count (Ne): Target yarn count produced on machine.

Efficiency ma%: Actual efficiency of the machine (100 maximum).

Yarn Breaks [1/1000 Rh]: Number of yarn breaks per 1000 rotor hours of production.

Stop Thick Short S [1]: Total number of stops during the shift for short thick places (between 2 and 10cm).

Stop Thick Long L [1]: Total number of stops during the shift for long thick places (greater than 10cm).

Stop Thin Moire M [1]: Total number of moiré stops (clusters of short thick places).

Stop Thin T [1]: Total number of thin places removed.

Yarn Spreadsheet Data Terms:
Date: The date the item was tested.

Count/Mix: Is the count of the item and what mix it is.

Frame: This is the frame the product was produced on.

Average Weight (AVG WGT): this is the average weight of the product.

% CV of Weight: it takes 10 packages to calculate the coefficient variance of the packages.

Skein Break: pounds required to break 120 yard skein of yarn as measured by the Scott J. Tester.

Break Factor: Skein break times the average weight (this normalizes the strength for yarn count).

S.E. Grams (Single End): the single end strength as measured by the Uster Tenso Rapid tester.

S.E. % CV (Single End): Single end variation of yarn strength of 10 packages breaking them 20 times to get the CV of 200 breaks.

RKm Grams/Tex (strength): Normalizing factor for yarn strength g/Tex of the yarn.

% Elongation: Using the Tenso Rapid tester to calculate the % of elongation across 10 packages.

% CV Elongation: Variation of a yarn elongation testing 10 packages and getting 200 readings, 20 per package.

Wax Friction: Friction on metal as measured by the Lawson Hempphill friction tester.

Wax % CV: Variation of the 10 readings.

TPI: Twist in turns per inch.

Hairs >3mm: Average hairs tested from 6 packages using the Zweigle hairiness tester.

Uster % CV: Evenness of the yarn tested on the Uster 3 evenness tester.

% VB: CV of the 10 tests.

V.L.C. 1 yd: 1 yard weight variability as measured by the Uster Tester 3.

V.L.C. 3 yd: 3 yard weight variability as measured by the Uster Tester 3.

V.L.C. 10 yd: 10 yard weight variability as measured by the Uster Tester 3.

V.L.C. 50 yd: 50 yard weight variability as measured by the Uster Tester 3.

Thin Places: Values of the yarns less than 50% level it counts is as a thin place.

Thick Places: More than 50% of the average mass is greater than it counts it as a thick place.

Neps: More than 200% of the average mass is greater than it counts it as a nep.

4.0 Data Quality

Data quality is a challenge for any organization. Quality data must consist of information that a company needs to be able to make informed decisions. There are many studies being conducted about what makes information quality. At the recent 2001 Information Quality Conference at MIT, there were many papers submitted on this subject. Some examples of these are "An Exploratory Investigation Into the Impact of Information Quality Upon the Perceived Value of Information" by Doig, Doherty and Marples [12]. Another paper discussed how to keep on track when undergoing such an initiative. There were many papers that touched upon this topic warranting it discussion in this dissertation. As can be seen in this section of the dissertation data quality, is a major issue in creating a data warehouse in order to accomplish data mining.

This section of the dissertation addresses the different procedural and computer issues that had to be dealt with to make the data usable in the chosen software system. It is important to know the different sources of data, which create the final database for a spinning mill. Each step of the mill process will be discussed.

Figure 4.1 shows a process flow diagram about the different data that was collected from this spinning mill. There are different phases in processing baled cotton into a spun yarn and at each phase there are different data collection systems that are being used in the chosen facility. These can either be on line or off line testing systems. As shown in Table 4.1 the plant being used has different systems however Figure 4.1 breaks these into segments of processing for better understanding. At each step there is data that is collected in some format that will be used. However, it can be assumed that this data will also need to be cleaned.

Figure 4.1 shows the different phases of taking a bale of cotton and processing it through the system to create a yarn. What occurs in the process and how long the cotton product stays there are discussed in more detail later in this chapter.

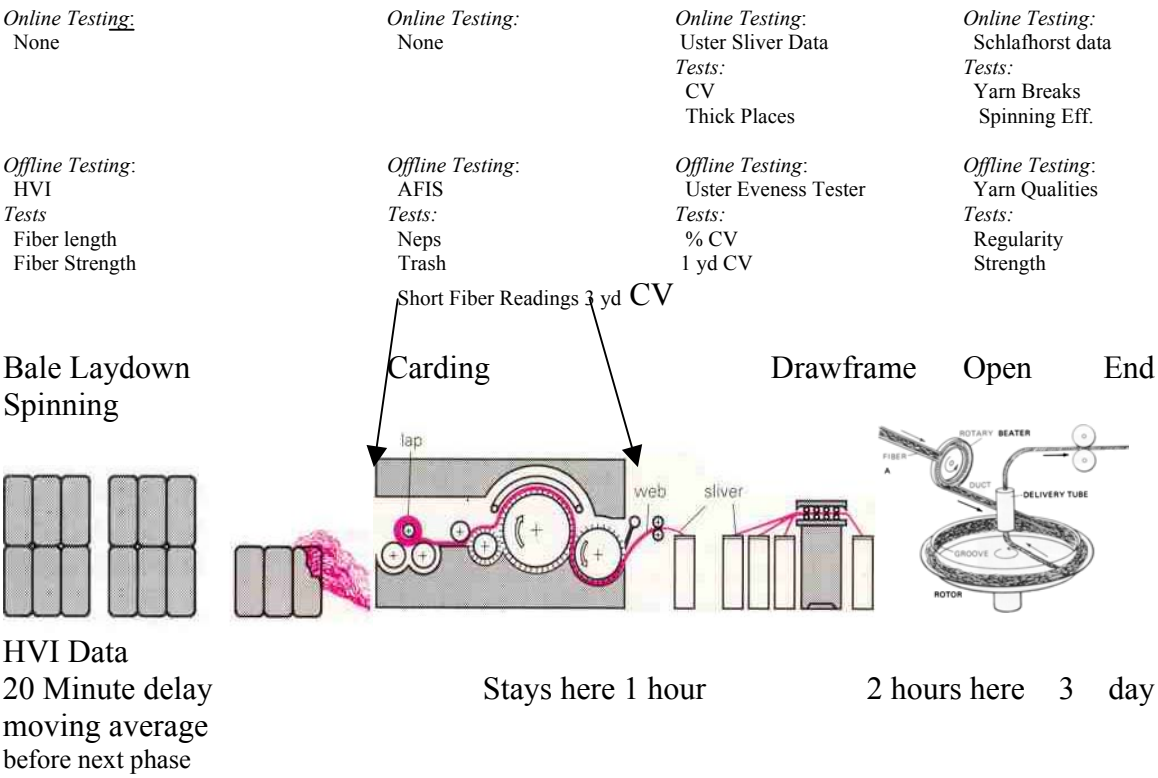


Figure 4.1: Flow of Data obtained through the Spinning Mill

The next step in this process was to get an actual data set that could be mined. On the surface this seems like a relatively easy proposition, but in actuality it was not. The chosen company has different measuring and monitoring systems within their manufacturing plant, and these are shown in Table 4.1. The data on these different systems are saved in different formats and for different lengths of time. Table 4.1 shows what types of data are collected at the plant for each process and what systems these are stored on. It also gives an idea about how long that data is kept in the case study plant.

Table 4.1: Current systems in the chosen plant

System for what process	What is on this system?	Data Systems	How long data is kept
EFS – bale data	Cotton Inc. proprietary HVI	Cotton Inc.	No limit
AFIS – card data	Card sliver data - offline	Excel 2000	No limit
Drawing	Uster sliver data	Uster version 2.6.4	5 Shifts
Spinning	Production information	Schlafhorst - WINCI	No limit
Yarn Quality	Quality information - offline	Excel 2000	1 year

In sections 4.1, 4.2 and 4.3, the different cleansing phases will be reviewed. These are passes through the data to get it ready for data mining. The first phase is determining what is available for each process within spinning to see what types of data are available. Within this phase it is important to create files that will be able to be imported into SAS. A large part of this cleansing process is creating one line of data for each record. The second phase was to examine the data to find a common variable in order to combine all the different data sets. In this case it will be time. However, the biggest issue with time is that this is a manufacturing process and therefore a production lag of time exists between each process. For example, if the fiber is in the bale laydown process then there is no way it can be in carding at the same time. This is true of all things, for instance if a person is at the mall they cannot also be at the hospital. It is the whole premise of a person cannot be in two places at one time. So therefore lags needed to be determined. A lag is the time it takes to get from one process to the next. Figure 4.1 shows the layout of the processes in spinning and at the bottom of this diagram there are times noted of how long it should take for the cotton to move through this process. Phase 2 of the cleansing takes all of the different processes, bale laydown, drawing,

spinning and yarn and transfers the time element to be recognizable by the SAS software. Phase 3 actually deals with the lag times. In this phase the determination of how to merge them is approached.

4.1 Cleansing Process Phase 1

4.1.1 Bale Laydown

The first stage in the processing of cotton fibers is the bale laydown. A bale is a sack or bag of staple fibers, in this case cotton, that are compressed for shipping purposes. Each bale weighs approximately 500 pounds. The USDA classes each bale. This information can be accessed by having the number on the barcode label that is attached to the bale. The data is produced from the HVI test that are done by the USDA and is made available to the company that will be purchasing the bale for their mill. See Appendix A for an example of how cotton is classified this information can also be found at http://ipmwww.ncsu.edu/Production_Guides/Cotton/chptr18.html. While the purpose of this classing is to establish cotton quality, and thus set pricing, the data provided can be used to optimize bale laydown in order to obtain a uniform product line. A standard laydown consists of 30 bales that are placed first in a holding pattern for approximately 36 hours before it will be used. The laydown will take approximately 8 hours to be consumed; this figure is based on a 30-bale laydown at 500 pounds per bale for a total of 15,000 pounds. During this study the participating mill has changed their process to a 90-bale laydown. In this case, the 90-bale laydown, at 500 lbs per bale for a total of 45,000 pounds, will take 3 times that of the 30-bale laydown to be consumed, or one full day (24 hours).

At this stage of the process all of the data is collected offline. Approximately 30% of the bales are retested in this particular company's own testing lab using their own HVI tester. The rest of the time they refer to the USDA data that is available on each bale, this data is then imported into the Cotton Incorporated EFS system. There can be as many as six laydowns present in the mill at one time. The types of information that is gathered about the bale are laydown number, bale number, micronaire, 2.5 % span length, uniformity index, strength, elongation, colorimeter A, colorimeter B, color/grade, ultraviolet reading, laydown position, date consumed, mix consumption and net weight of bales. An example of this can be found in Figure 4.2. The information that is found in this figure comes from a printout form of the data that is available for the laydown data.

MIX 2020 CONSUMPTION BY LAYDOWN PAGE 1

LAYD NO	BALE NO	MIC	2.5% SL	UI	STR	EL	RD	+B	C-G	UV	LAYD POS	DATE CONS	MIX CONS	NET WGT
0003914	785532	48	1.07	.82	26.7	6.3	74.0	8.4	413	70	27	0/10/13	2020	507
	785533	49	1.06	.82	27.8	6.3	73.0	8.7	413	70	29		2020	494
	785578	48	1.06	.82	28.8	6.3	76.0	7.9	411	70	9		2020	491
	807084	47	1.02	.82	25.8	6.3	76.0	8.2	312	73	15		2020	506
	829584	38	1.11	.81	31.0	6.3	80.0	8.1	212	82	30		2020	514
	830824	38	1.09	.82	31.0	6.3	80.0	8.3	212	84	2		2020	512
	831609	49	1.08	.83	29.3	6.3	78.0	6.4	411	60	11		2020	503
	831650	49	1.09	.81	29.0	6.3	79.0	6.2	411	61	21		2020	490
	839521	42	1.14	.83	33.2	6.3	76.0	8.2	312	73	14		2020	513
	870454	40	1.09	.81	27.0	6.3	77.0	9.2	313	85	8		2020	465
	870469	41	1.11	.81	27.2	6.3	79.0	8.5	212	83	28		2020	494
	870471	41	1.11	.81	27.2	6.3	79.0	8.5	212	83	4		2020	508

Figure 4.2: Original Data Set from Bale Laydown (Non Summarized)

The abbreviations with the definitions of what these terms mean for this particular mill can be found at the beginning of this chapter. The actual gathering of data is exported out of the HVI system and downloaded into Excel. During this process the data goes from one line of information to three lines. The reason for this is the format that the data is stored in the EFS data mart. When transferred from one program to another the data will appear differently. The data itself has not changed it just no longer lines up accurately. An example of this can be seen in Table 4.3. This table shows an example of

the data that is averaged. Unlike the previous table that has each individual bale this is the averages of each laydown.

Table 4.2: Bale Laydown Summarized Data Transferred to Excel

		LAYD	BALE	MIC	2.5%	UI	STR	EL	Rd	+b	C-G	UV
MIX	Date Consumed	NO	Count		SL							
2020	07/23/00	3704										
2020	07/23/00	AVG		44	1.08	0.82	28	7	77	8	342	73
2020	07/23/00	COUN	30									

Table 4.2 shows an example of problems with data format that typifies the sort of data conditioning that was necessary. The first is that when taking this data and transferring it into the software to be used it is important to only have one row for the titles of the columns. For instance, laydown number is referred to in this table as LAYD NO however these are two separate rows in the spreadsheet. This needed to be rectified immediately. This was also a problem with other titles and these also had to be changed and put into one row. The next issue is that when transferring this data from EFS to the Excel spreadsheet the one row of data became 3 different rows. The average count, (AVG COUN) is the number of bales in the laydown, went from being a single row to two rows, one of which contained lots of blanks. The information that is needed should be on one row so that there will not seem to be missing data points within the data. This is very important to cleanse out because of the problems this would cause when placed into the software. It would not import properly because it no longer makes any sense to the SAS program. Table 4.3 shows an example of this data cleansed to meet the requirements that have been discussed.

Table 4.3: Cleansed Excel Spreadsheet of Bale Laydown Summarized Data

Date Consumed	LAYD NO	Bales	MIC	2.5% SL	UI	STR	EL	Rd	+b	C-G	UV
07/23/00	3704	30	44	1.08	0.82	28	7	77	8	342	73

Another issue with this data is that there is data on every bale and then there is also summarized data. The data that is in Tables 4.2 and 4.3 are the averaged values of the 30 bales that create one laydown. The problem with using this data is that it does not represent any variability inherent in the different laydowns. Because of this it was determined that it would be important to calculate and include the coefficient variation for some of the columns. The parameters that this was determined for was the micronaire, strength, colorimeter A, colorimeter B and the Color/grade. This information may or may not prove to be useful. However if it is not included in this study there is no way to know if the summarized values are telling anything about the data in the plant. An example of this spreadsheet is in Table 4.4. This is an expansion of Table 4.3, which will now not be used in the final data warehouse. The only difference in this table and the one in SAS is that the first row only occupies one row of data. The reason for this is that only one row can be read into SAS. One row equals one record and therefore if it is in another program in two rows it will not input correctly, SAS can not distinguish between the two. The purpose for the difference in this dissertation is purely the lack of space across the page.

Table 4.4: Bale Laydown Data with Additional CV% columns

LAYD No	MIC	2.5% SL	UI	STR	EL	Rd	+b	C-G	UV	DATE CONS	CV% Mic	CV% Str	CV% Rd	CV% +b	CV% C-G
2000032	42	1.08	0.81	26.8	6.3	74.3	8.1	403	68	12/04/00	7.84	4.30	1.58	5.36	6.86
2000033	42	1.08	0.82	27	6.3	74.1	8.1	405	68	12/04/00	7.76	4.37	1.81	5.45	6.06

During the laydown a machine goes across the top of the bales and takes a little cotton off each bale to move the cotton to the next step of the process. This process is continuous. The process begins by opening the cotton and after this it will move to the cleaning phase and then onto the blending phase. This is referred to as the blow room.

There is no data collected during these processes and therefore they will not be discussed in the following sections.

4.1.2 Carding

The next process that the fiber goes through is carding. This is where the fibers are opened, cleaned, aligned and formed into a continuous, strand called a sliver. At this stage of the process, the only data that is collected is from the AFIS offline testing system. This is only conducted once per month and therefore cannot be used in the mining of this data because there are infrequent data points. The information cannot be combined to determine the production lag time within the plant. There is only one data point per month and to match this with others would be a waste of time. It is merely a snapshot of what is being processed within this mill at any given time. This data seems to only be used in this plant to determine when maintenance of the machine should be conducted. The data that is collected is placed into an Excel spreadsheet once a month and kept for "no limit of time". The type of information that AFIS generates at this phase are the Card number, Mat, Sliver, % Elimination, sliver neps, dust, and trash, these terms are also explained in the glossary at the beginning of this chapter. This information is hand entered monthly by one individual who is in charge of the testing. Once a month the AFIS machine in the company lab generates a report that is then inputted into the report that can be found in Appendix D, which is a customized report for the company that is being used for this project.

4.1.3 Drawing

The drawing process is where the weight per unit length of laps, slivers, slubbings, or rovings are reduced. At this stage of the process there is online testing in

the mill that is being evaluated. The testing is done using the Uster Sliver Data program. There are other systems that can be used, for instance the Rieter Spider Web. In this case the information is only kept for five shifts and the information is on a sheet of paper that does not download onto a system and hence needs to be hand entered if it is to be used for evaluation beyond the original printout. If this paper is misplaced there is no backup of it past the five working shifts. The kinds of information that is available on this printout are the machine number, style of product, count, pounds, pounds per hour, actual efficiency, stops, stop minutes per doffing or can change, quality stops, and weight variation of the sliver in both 100 yards and in inches. These terms are all defined in the glossary and abbreviations section at the beginning of this chapter.

The main issue with this data is that it all has to be hand entered accurately and quickly into an Excel spreadsheet to be used for evaluation. The data is printed out both in the morning and evening shift so there are two readings per day per machine. An example of what this printout looks like can be found in Table 4.5. This is what the printout looks like for the company used in this case study.

Table 4.5: Uster Sliver Data Shift Report

END OF SHIFT MO 11-12- 0 19:00

USTER SLIVERDATA V2.6.4-41 MO 11-12- 0 19:00
STYLE ALL / SINGLE / LAST SHIFT
MONITORED TIME 12H 0M SHIFT 1 MO 11-12- 0 7:00 TILL MO 11-12- 0 19:00

	MACH	STYLE	COUNT	LB	LB/H	ACT%	ST	CD	STM	D	M/D	QS	A%	CV%	CVL	OS	GR
*	1	COTTON	70.00	3505	520*56.2		42		276	96	.4	5	.5	3.6	.1	1	
*	2	COTTON	70.00	3233	495*54.5		35		279	90	.5		.5	2.8	.1	1	
*	3	COTTON	70.00	3099	496*52.1		25		195	85	1.8		.5	3.2	.1	1	
*	4	COTTON	70.00	3798	499*63.5		26		205	104	.6		-.5	3.2	.1	1	
*	5	COTTON	70.00	3790	497*63.5		28		184	104	.8		-1.0	3.1	.1	1	
*	6	COTTON	70.00	3124	498*52.3		28		127	86	2.5		-.5	3.0	.1	1	
S	6	COTTON	70.00	20549	501	57.0	184		1266	565	1.0	5	.0	3.1	.1		

Table 4.6 shows an example of what this data looks like in the Excel spreadsheet. In comparing Tables 4.4 and 4.5 it can be seen that several columns were left out of the final database spreadsheet. For instance in this facility only cotton styles are produced so it is not necessary to have a column that contains only cotton in that field, this is extra data entry that is not needed.

Table 4.6: Sliver Data Entered Into an Excel Spreadsheet

Date	Time	Mach	Count	LB	LB/H	ACT%	ST	CV%
12/8/00	7:00 PM	7	75	3960	565	58.4	53	2.5
12/8/00	7:00 PM	8	75	4241	575	61.5	39	2.5
12/8/00	7:00 PM	9	75	3878	571	56.6	36	2.7
12/8/00	7:00 PM	1	70	2127	519	34.1	31	3.9
12/8/00	7:00 PM	2	70	2881	504	47.6	33	2.9
12/8/00	7:00 PM	3	70	1985	505	32.8	22	3.2
12/8/00	7:00 PM	4	70	2237	499	37.4	25	3.2
12/8/00	7:00 PM	5	70	2301	498	38.5	26	3.2
12/8/00	7:00 PM	6	70	2179	495	36.7	30	3
12/9/00	7:00 AM	7	75	4201	568	61.6	50	2.5
12/9/00	7:00 AM	8	75	4360	572	63.6	39	2.5
12/9/00	7:00 AM	9	75	4094	569	60	55	2.7
12/9/00	7:00 AM	1	70	1231	506	20.3	26	3.8
12/9/00	7:00 AM	2	70	1268	515	20.5	20	3

4.1.4 Open End Spinning

Open End Spinning is a system of spinning based on the concept of introducing twist into the yarn without package rotation by simply rotating the yarn end at a gap or break in the flow of the fibers between the delivery system and the yarn package. The system being used at this time is Schlafhorst –WINCI. At this facility the data is kept indefinitely. The monitoring systems at this time are merely print outs that come off the machine. The mill states that it cannot be saved to a file but rather are simply in paper format that will need to be reentered into a usable file, in this case, the file format will be entered into an Excel file. A representative, Heinzbert Reiners, of the Schlafhorst Company verified this to be true in an email communication. The monitoring system

takes readings continuously throughout the day per machine. However, the data that was supplied by the mill in this case study is the summarized data for the week. This data was all that the plant would provide for this research. This printout gives a per shift summary at 7 a.m. and 7 p.m. for every machine that is available on the monitoring system. However, in this particular printout there is no actual machine specific data. An example of this printout can be found in Figure 4.3.

ust

Schlafhorst
Central Informator Autocorc

Page 1 1/25/01 07:16 AM

Machine: *all* **Grouping:** *NONE*
Time: *10/08/00 07:00 pm - 10/15/00 07:00 am* **Restriction:** *NO*

Shift start	Yarn count [Ne]	Efficiency ma. [%]	Yarn breaks [1/1000Rh]	Stop thick short S [1]	Stop thick long L [1]	Stop thin T [1]	Stop Moire M [1]
10/10/00 07:00 pm	18.0	93.4	238.4	38	44	2	15
10/10/00 07:00 pm	18.0	93.8	215.8	57	8	6	10
10/10/00 07:00 pm	18.0	90.3	274.8	30	18	1	26
10/10/00 07:00 pm	18.0	94.5	280.7	35	26	3	10
10/10/00 07:00 pm	18.0	91.0	274.3	39	17	3	12
10/10/00 07:00 pm	18.0	93.8	183.4	38	19	5	12
10/10/00 07:00 pm	18.0	91.2	221.5	95	109	4	2
10/10/00 07:00 pm	18.0	92.1	303.8	42	80	4	11
10/10/00 07:00 pm	18.0	93.6	150.0	117	109	1	3
10/11/00 07:00 am	18.0	83.2	283.3	39	11	1	14
10/12/00 07:00 pm	18.0	93.0	237.9	44	21	2	12
10/12/00 07:00 pm	18.0	92.4	158.9	137	98	1	4
10/12/00 07:00 pm	18.0	88.9	258.3	104	63	4	4
10/11/00 07:00 am	18.0	93.1	236.5	33	25	2	20
10/11/00 07:00 am	18.0	91.8	254.0	23	16	2	13
10/11/00 07:00 am	18.0	88.1	253.4	32	8	1	28
10/11/00 07:00 am	18.0	90.9	255.4	30	28	0	20
10/11/00 07:00 am	18.0	92.6	166.6	22	16	2	12
10/11/00 07:00 am	18.0	90.8	162.8	118	123	2	0
10/11/00 07:00 am	18.0	86.3	207.5	37	26	3	11
10/11/00 07:00 am	18.0	93.8	247.5	24	11	2	13
10/11/00 07:00 am	18.0	87.0	281.7	28	17	2	4
10/11/00 07:00 am	18.0	87.2	303.7	27	73	2	12
10/11/00 07:00 am	18.0	89.9	203.0	107	74	2	0
10/11/00 07:00 pm	18.0	92.6	299.4	49	6	1	8
10/12/00 07:00 pm	18.0	88.0	214.7	25	17	2	10
10/12/00 07:00 pm	18.0	90.4	321.4	41	95	2	12
10/12/00 07:00 pm	18.0	89.4	326.9	35	19	2	8
10/11/00 07:00 pm	18.0	92.4	281.9	37	43	3	14
10/11/00 07:00 pm	18.0	92.0	258.8	26	6	1	19
10/11/00 07:00 pm	18.0	90.5	304.9	0	0	0	0
10/11/00 07:00 pm	18.0	94.6	255.8	35	22	1	10
10/11/00 07:00 pm	18.0	77.0	238.4	37	28	6	11
10/11/00 07:00 pm	18.0	95.8	321.9	23	5	5	15
10/11/00 07:00 pm	18.0	94.3	181.4	30	9	6	13
10/11/00 07:00 pm	18.0	91.6	315.2	35	6	0	12
10/11/00 07:00 pm	18.0	92.4	308.2	36	83	4	14
10/11/00 07:00 pm	18.0	91.9	166.8	126	100	3	6
10/10/00 07:00 pm	18.0	92.5	256.9	22	16	2	16

Figure 4.3: Schlafhorst Spinning Data Week Summary

The data that is found in Figure 4.3 was then entered into an Excel spreadsheet. Again the machine number is still missing from the data. An example of what this looked like after entering the data can be seen in Table 4.7. The main difference between this dataset and the actual Excel spreadsheet is that each column heading is only one row of information; one record equals one row. For space purposes it is shown here as two rows.

Table 4.7: Schlafhorst Spinning Data Entered Into an Excel Spreadsheet

Shift Start	Yarn Count (Ne)	Efficiency ma. (%)	Yarn Breaks (1/1000 RH)	Stop thick short S	Stop Thick long L	Stop thin T	Stop Moire M
10/03/00 07:00pm	18	85.6	220.8	136	130	0	7
10/03/00 07:00pm	18	94.7	242	42	13	0	13
10/03/00 07:00pm	18	92.7	318.6	51	10	1	12
10/03/00 07:00pm	18	91.9	274.1	76	21	1	13
10/03/00 07:00pm	18	87.1	465.2	35	106	4	18
10/03/00 07:00pm	18	87.5	204	89	61	4	11
10/03/00 07:00pm	18	93.1	213	32	18	4	15
10/03/00 07:00pm	18	88	350.6	34	16	0	15
10/04/00 07:00am	18	91.5	324.9	57	12	1	14
10/05/00 07:00pm	18	95.1	260.2	46	26	0	14
10/04/00 07:00am	18	92	266.9	31	25	0	25
10/05/00 07:00pm	18	92.3	278.9	47	14	6	8
10/05/00 07:00pm	18	85.7	195.7	103	155	1	9
10/04/00 07:00am	18	88.5	310.4	38	14	2	12
10/04/00 07:00am	18	86.5	358.5	22	6	1	11
10/04/00 07:00am	18	89.1	178.7	73	57	1	6
10/04/00 07:00am	18	93.8	269.7	39	30	2	11
10/04/00 07:00am	18	86.8	202.3	123	159	1	10
10/04/00 07:00am	18	92	263.7	47	30	1	16
10/04/00 07:00am	18	92.9	291.7	58	16	0	16
10/04/00 07:00am	18	94.1	201.3	25	6	13	13
10/04/00 07:00am	18	87.7	447	28	82	2	16
10/04/00 07:00am	18	87.8	358.3	20	23	1	6
10/05/00 07:00pm	18	85.5	190	69	60	2	5
10/04/00 07:00pm	18	94.4	321	54	12	2	19
10/05/00 07:00pm	18	90.8	344.8	34	113	2	22

4.1.5 Yarn

At the end of the process a yarn is formed and placed on a package ready for shipment. At this time a random testing is done on the product to obtain final quality

information about the product. This information is hand entered into an Excel spreadsheet that will be kept for one year's time at the case study plant. The type of information that is contained in the spreadsheet are: regularity, strength, elongation, friction, wax percentage, turns per inch, hairiness, thin and thick places. An example of some of the columns in this dataset can be seen in Table 4.8 however, in the actual Excel spreadsheet the column headings again only take up one row not two as shown.

Table 4.8: Yarn Dataset Example

DATE	*COUNT/ MIX*	FRAME	AVG WGT	% CV of WGT	SKEIN BREAK	BREAK FACTR
10/17/00	20kt 220	8	20.24	1.33	73.67	1491.12
10/17/00	20kt 220	9	20.36	1.88	74.37	1514.23
10/17/00	20kt 220	10	20.47	0.42	69.16	1415.79
10/17/00	20kt 220	11	20.16	1.77	76.43	1540.76
10/18/00	20kt 220	12	19.95	1.18	79.11	1578.16
10/24/00	18kt 220	8	20.02	1.04	70.94	1420.26
10/24/00	20kt 220	9	20.08	1.19	72.22	1450.17
10/24/00	20kt 220	10	20.18	1.24	71.74	1447.79
10/24/00	20kt 220	11	20.18	0.84	75.68	1527.12
10/24/00	20kt 220	12	20.28	1.54	68.69	1393.00

The main issue with this spreadsheet is that it takes up to three rows to get the title in small width column. This needed to be changed in order to have it entered into a SAS program correctly; one row must equal one column for importation. There is only one reading per week per frame on this type of data per count makeup. In the case study that is being conducted only one type of count mix, 20/20, is being examined. This number refers to the count of the yarn and what mix it is.

4.2 Cleansing Phase 2

Once the first phase of cleansing the data was completed it was determined that it would now need to be transferred into a SAS dataset. In order to do this it became very

apparent that the data, that was already in four separate Excel spreadsheets, would need to be merged. In order to do this it was important to find a common ground. When looking at the plant set up and the different lags from starting in a laydown to a finished yarn it was determined that the date and time would be the key to this particular data. Once this was realized it was then time to look at each dataset to see how time was recorded at each step. The following section will discuss the different time issues and what codes had to be written for each set of data.

4.2.1 Laydown

In the Excel spreadsheet, which was imported into a SAS dataset, there was a column titled "date consumed". This date was read into the system as year/month/date because this is how it was downloaded from the HVI dataset on EFS. However, SAS recognizes a date by the month/date/year. Figure 4.4 shows the code that needed to be written in order for the date to appear correctly in the SAS dataset. First a new library is created to store the data set into SAS. Then the data is imported using the import wizard and it replaces the Excel spreadsheet with a SAS database. A new column is also created called xdate this reformats the date consumed data from the original dataset to make it represented as month/date/year so it can be understood in SAS.

```
libname stacey 'd:\jerry';

                      PROC IMPORT OUT= WORK.laydown1
                           DATAFILE= "d:\jerry\LAYDOWN1.xls"
                           DBMS=EXCEL2000 REPLACE;
                           GETNAMES=YES;
                      RUN;
                      data stacey.laydown1;
  set work.laydown1;
                            format
   xdate_cons date7.
                            ;
```

Figure 4.4: SAS Code for Converting Date for Laydown Data

4.2.2 Carding

This dataset was not dealt with for date issues because it will not be used in the merged dataset and, therefore, the date is not a problem at this time.

4.2.3 Drawing

In this dataset the date and time were inputted using the import wizard. When the date is taken off the machine and put into an Excel spreadsheet it is inputted into one row as date/month/year. Adding another complication is that another cell contains the time of the reading. This will either be 7 a.m. or 7 p.m. as the company gets a morning and then a night reading or one per shift. When this is transferred into SAS two new columns are created, one for the date and another to correspond with the time. This code creates a true time and date variable. This can be seen in Figure 4.5. The first half of the code imports from the "jerry" file and then it stores the data in the created library. The data is imported again from an Excel spreadsheet to create a SAS data set. In this code the date and time are taken apart to create two separate columns of information. Then this information is reformatted to bring the information into a SAS understandable format of mmddyy. The time, in minutes or military time, is also calculated within this code and a separate column is created for this as well.

```
libname stacey 'd:\jerry';
PROC IMPORT OUT= WORK.drawing1
      DATAFILE= "d:\jerry\DRAWING1.xls"
      DBMS=EXCEL2000 REPLACE;
      GETNAMES=YES;
      RUN;
data stacey.drawing1;
  set work.drawing1;
  format
    xtime time5.
     xdate datetime13.
;
```

Figure 4.5: SAS Code for Converting Date for Drawing Data

4.2.4 Open End Spinning

At first glance, this dataset seemed to be transferring from the Excel spreadsheet to the SAS dataset without any problems. However, once it was compared to the Drawing dataset it was determined that it also had problems because it would need to be merged with the other data. In this dataset the date entered correctly but the concern became that the time in this set was being transferred as either 7:00 a.m. or 7:00 p.m. Military time was not being utilized as with the other dataset. Therefore, this column also needed to be adapted for SAS. This was done using the code in Figure 4.6. The beginning of the code does the same thing as with the two previously discussed codes. Again two new columns are created in the SAS data set one for the date and one for time. The time needed to be converted from 7 a.m. and 7 p.m. to military time. This is done in the code with the indexing of shift starts. If x1 is equal to zero then the time is 07:00 and if the x2 variable is zero the 19:00 is entered into the time column. The data is also put into the SAS format of mmddyy.

```
libname Stacey 'd:\jerry';

PROC IMPORT OUT= WORK.schlaf1
            DATAFILE= "d:\jerry\SCHLAF1.xls"
            DBMS=EXCEL2000 REPLACE;
     GETNAMES=YES;
RUN;

data stacey.schlaf1;
  set work.schlaf1;
  format
    xdate datetime13.
        time time5.
  ;
  x1=index(shift_start,'07:00am');
  x2=index(shift_start,'07:00pm');
  if x1 ne 0 then do;
    time='07:00't;
        x3=input(substr(shift_start,1,x1-1),mmddyy8.);
        xdate=dhms(x3,hour(time),0,0);
  end;
  if x2 ne 0 then do;
```

```
                time='19:00't;
                    x3=input(substr(shift_start,1,x2-1),mmddyy8.);
                    xdate=dhms(x3,hour(time),0,0);
              end;
              drop x1 x2 x3;
            run;
```

Figure 4.6: SAS Code for Converting Date for Schlafhorst Data

4.2.5 Yarn Data

In this dataset there were no major complications with the date as it was entered in a SAS recognizable format of month/date/year. However, a code needed to be written to put it into a data variable format instead of as a character string. A data variable format is one that is in numbers while the character string only sees the numbers as characters in the alphabet. This can be seen in Figure 4.7. The beginning of this code creates the library and then imports the data from Excel using the import wizard.

```
Libname Stacey 'd:\jerry';

PROC IMPORT OUT= WORK.yarn20220
            DATAFILE= "d:\jerry\yarn20220_only.xls"
            DBMS=EXCEL2000 REPLACE;
     GETNAMES=YES;
RUN;
data stacey.yarn20220;
    set work.yarn20220;
    format
       xdate date7.
;
```

Figure 4.7: SAS Code for Converting Date for Yarn Data

4.3 Cleansing Phase 3

Once all the dates and time were accurate it was time to determine the lag times and how best to merge the 4 different SAS databases to create one SAS database.

4.3.1 Bale Laydown

The laydown data helps to secure the rest of the data, meaning if the start of a laydown can be determined then hopefully the rest of the data can be lagged off of this

starting point. The issue with this is that it is assumed that when the data is recorded is actually being utilized which can be referred to as brought into the production process. It has been determined that the "Data Consumed" column cannot possibly be when it is actually utilized. It can be better understood as when the bale laydown is brought into the production plant. After careful analysis of the data it could be seen that physically there is no way to consume upwards of 6 bale laydowns in one day. As discussed earlier in section 4.1.1 a 30-bale laydown takes 8 hours to consume proving that only 3 laydowns of 30 bales or one laydown of 90 bales can be consumed in a 24-hour period. A visit to the plant was taken in order to verify what this column of information really meant. Upon closer investigation it was firmed up that this date is when the laydown is formed, meaning when the laydown is transferred onto a truck as a laydown. This does not mean it was actually placed into the process and used that day as originally defined. Therefore, in creating a database at this time it must be assumed that the "date consumed" column is that the bale laydown has been selected and that it will be in the process in the next few days. This is a bit risky when trying to link data across an entire process that begins with a somewhat unstable start. In this case, it will be used to see how it will work out.

4.3.2 Carding

At this stage of the cleansing process carding data has been removed due to the lack of actual data. This has been discussed earlier in this chapter. However, at this point it would be important to point out that if this data was collected on a daily basis instead of once a month it could be used to link the data across all processes.

4.3.3 Drawing

This dataset as discussed earlier is collected twice a day across all machines.

Each machine is recorded with the average for that particular shift. It might be more helpful to get every hour of data in order to better utilize time lags since it takes approximately one hour and 20 minutes to get from laydown to drawing. This is calculated by the fact that there is a 20-minute delay in the mixing chamber and then carding takes approximately one hour to accomplish. This is all assuming that there is no backup built into the process. See Appendix C to see a layout of the system with lags for the plant utilized in this case study.

4.3.4 Spinning

This dataset is also collected twice a day; however, only the summary data has been collected. There is no way from the data that has been provided for this study to determine what machine has done what. Therefore, it can be concluded that without knowing this it would be impossible to determine which machine is running the worst. However, other information may be obtained from the summary data. Also, because of the time it takes to produce a yarn a three-day moving average of laydowns will need to be used to determine a new variable to link with the rest of the data. The reason for this is that a can of sliver takes 48 hours to process. With the 48 hours plus the time waiting to process a 3-day moving average is being assumed. The 3-day period prior to and including the date of the yarn is the best approach. An example would be a yarn produced on the 22nd of March, this would be considered an average of the laydowns produced on March 20, 21 and 22.

5.0 Combining of Dataset to be Used in Enterprise Miner by SAS

At this point all of the data that has been collected is now cleansed. However, these datasets are still separate databases. In Chapter 4 it was discussed how and what needed to be cleansed but not what was necessary to be done to create a final data mart. In this chapter the codes that were used on the datasets to create what will be discussed along with some of the results that have been generated by using the final data mart along with different targeted variables.

5.1 Code for Laydown Dataset

The laydown dataset was regenerated using the code that is in Figure 5.1. This code allowed the datasets to first be sorted into numerical order by laydown number. The main reason for this statement is for making sure that laydown numbers are in increasing order so as to not create the wrong "xjoin" variable later in the code. Once this has been done a new variable is created called "xjoin" this variable is imperative in order to later join the other datasets. The "xjoin" variable is a newly created date variable; it allows for one laydown per day. The dataset itself began on December 8, after they switched to 90 bale laydowns, and one laydown can be processed per day therefore it was assumed that the data will start on the 8th of December then each additional laydown will be in process the next day. This code also allows for the Christmas shutdown and is coded to start with January 2nd for the determining the date of the laydown after the holidays. This is down by taking the number that is calculated and subtracting 13 so that it may begin at one all over again.

The purpose of generating the "xjoin" variable is to create common ground between the datasets that are to follow in the process. Without this variable there would be no way to join the other datasets.

```
proc sort data=stacey.laydown1;
  by layd_no;
run;

data stacey.laydown1;
    set stacey.laydown1;
    format
       xjoin date7.
  ;
xjoin=.;
 if _n_ le 13 then xjoin=intnx('day','08dec00'd,_n_);
 if _n_ gt 13 then xjoin=intnx('day','02jan01'd, (_n_-13));
run;
```

Figure 5.1: Laydown Code for New Data Variable

5.2 Code for Drawing Dataset

The drawing dataset is reworked to be able to combine it with the laydown dataset that was recreated. In the new drawing code the data is sorted by the date, and only by date; time is eliminated, and then by machine per day for each column in the dataset. The reason that the time of am and pm is eliminated is because there is nothing to link these with versus the other processes. The mean is then calculated for different outputs to create one per time frame. Once this is accomplished the code continues on to create a lagged dataset. It was determined by the lag times calculated for the plant which can be see in Appendix B and C that a 2 day moving average of this data needed to be created. The 2 day moving average is generated using the proc expand command in SAS. The moving average must be created using the reverse moving average because the dataset needed to be reversed this is accomplished by taking the series and using n-1 instead. The convert statement within this code simply renames the variable that is now the

moving average. Next by using the keep command the original variables are dropped and only lagged variables remain to be used. Once this part of the code has been run then the laydown data is joined together with the drawing dataset using the merge command. If the date exists in both datasets then it is a match if not the extras are dropped from the dataset. See Figure 5.2 for more specifics.

```
data work.drawing_day;
  set stacey.drawing1;
  xjoin=datepart(xdate);
run;
proc sort data=work.drawing_day;
  by mach xjoin;
run;

proc means data=work.drawing_day noprint;
  by mach xjoin;
  format
    xjoin date7.
  ;
  var lb lb_h act_ st cv_;
  output
    out=work.drawing_summary
    mean=
  ;
run;

proc expand data=work.drawing_summary out=work.drawing_averages method=none;
  by mach;
  id xjoin;
  convert lb=lb_movavg2 / transform=(reverse movave 2 reverse);
  convert lb_h=lbhr_movavg2 / transform=(reverse movave 2 reverse);
  convert act_=act_movavg2 / transform=(reverse movave 2 reverse);
  convert st=st_movavg2 / transform=(reverse movave 2 reverse);
  convert cv_=cv_movavg2 / transform=(reverse movave 2 reverse);
run;
data stacey.drawing_lagged;
  set work.drawing_averages;
  keep
    lb_movavg2
          lbhr_movavg2
          act_movavg2
          st_movavg2
```

```
        cv_movavg2
        xjoin
        mach
  ;
run;
proc sort data=stacey.laydown1;
  by xjoin;
run;

proc sort data=stacey.drawing_lagged;
  by xjoin mach;
run;

data stacey.lay_draw;
  merge
    stacey.laydown1(in=in1)
        stacey.drawing_lagged(in=in2)
  ;
  by xjoin;
  if in1 and in2;
run;
```

Figure 5.2: Code for Drawing Regeneration

5.3 Code for Spinning Data

The spinning data is from the Schlafhorst system and this data also needs to be combined with the laydown and drawing data. In order to join this set of data with the others a code also had to be written for it. In this code the data is rolled up into daily averages after it has been sorted to be sure that the data is also in order from latest to most current date. Again in this dataset an "xjoin" variable is created so that merging can be accomplished. This is a very similar process as to what was done for the drawing data. The main difference becomes the averaging number in this set a three-day moving average is created instead of a two day like in drawing. This code uses the proc expand, convert and reverse commands as well. The reverse command allows us to get the averages of the data from 1,2 and 3 instead of 10, 9 and 8. The old data is dropped and the lagged data is saved in its place. Once this has been accomplished these datasets can

be merged into one using the "xjoin" variable. In this case a command is used to be sure that if a date exists in the laydown and drawing datasets it must also be present in the Schlahorst – spinning dataset.

```
data work.schlaf1;
  set stacey.schlaf1;
  if yarn_count__ne_=20;
  format
    xjoin date7.
  ;
  xjoin=datepart(xdate);
run;
proc sort data=work.schlaf1;
  by xjoin;
run;
proc means data=work.schlaf1 noprint;
  by xjoin;
  var efficiency_ma_____ -- stop_moire_m;
  output
    out=work.schlaf_averages
          mean=
  ;
run;
proc sort data=work.schlaf_averages;
  by xjoin;
proc expand data=work.schlaf_averages out=work.schlaf_lag method=none;
  convert efficiency_ma_____=efficiency_ma_lag3 / transform=(reverse movave 3
reverse);
  convert yarn_breaks__1_1000_rh_=yarn_breaks__1_100_lag3 / transform=(reverse
movave 3 reverse);
  convert stop_thick_short_s=stop_thick_short_s_lag3 / transform=(reverse movave 3
reverse);
  convert stop_thick_long_l=stop_thick_long_l_lag3 / transform=(reverse movave 3
reverse);
  convert stop_thin_t=stop_thin_t_lag3 / transform=(reverse movave 3 reverse);
  convert stop_moire_m=stop_moire_m_lag3 / transform=(reverse movave 3 reverse);
run;

data stacey.schlaf_lagged;
  set work.schlaf_lag;
  keep
    efficiency_ma_lag3 -- stop_moire_m_lag3 xjoin
  ;
run;
```

```
proc sort data=stacey.lay_draw_yarn;
  by xjoin;
run;
proc sort data=stacey.schlaf_lagged;
  by xjoin;
run;
data stacey.lay_draw_yarn_schlaf;
  merge
    stacey.lay_draw_yarn(in=in1)
          stacey.schlaf_lagged(in=in2)
  ;
  by xjoin;
  if in1 and in2;
run;
```

Figure 5.3: Code for Schlafhorst – Spinning Data

5.4 Code for Yarn Data

The yarn data is sorted by date just like all of the other datasets. After this has been accomplished an "xjoin" variable is created. Since this dataset only has one output per week new variables have to be created. Every week the one row of data is used to build a representative dataset. For instance, if the row of data is read on the 12th of December it may then be copied for data on the 9, 10 and 11th of the month as well. This part of the code can be seen in Figure 5.4. Again using the same code as in drawing and spinning the lags are created for this data as well as the three-day moving averages. Once the new variables are saved then these are joined with the laydown, drawing and spinning data that was already combined.

```
if xdate='12dec00'd then do;
  xjoin='09dec00'd; output;
  xjoin='10dec00'd; output;
  xjoin='11dec00'd; output;
end;
```

Figure 5.4: Building Rows of Yarn Data

Now that one data mart has been created the analysis using SAS Enterprise Miner can be done. The final format of the data mart is a spreadsheet that has rows and

columns. Each record represents one row and each column represents an attribute of that particular record. The data itself is by both product and by process. The data has repeated rows within it because of the production lag times. A laydown is averaged from 90 bales of data this in turn is then compared to the next process which is drawing in this case and that has readings twice a day. In Table 5.1 an example of what the data consists of is shown. This table includes all of the following information: Letter of the column it represents, name of the attribute, units it is collected in, format of the data (real, integer, date/time), and frequency that it is collected. Definitions of the "Name" column can be found in the glossary at the beginning of Chapter 4.

Table 5.1: Final Format of the Data Collected

Column #	Name	Units	Format	Frequency
A	LAYD_No	Number	Integer	1/day
B	MIC	Weight/unit length	Real	1/day
C	__5__SL	Length/Number	Real	1/day
D	_UI	Ratio	Real	1/day
E	_STR	g/Tex	Real	1/day
F	_EL	Percentage	Real	1/day
G	__Rd	Number	Real	1/day
H	__b	Number	Real	1/day
I	C_G	Number	Real	1/day
J	UV	Number	Real	1/day
K	DATE_CONS	Date	Date/time	1/day
L	CV_Mic	Weight/unit length	Real	1/day
M	CV__Strength	Length/Number	Real	1/day
N	CV_Rd	Number	Real	1/day
O	CV__b	Number	Real	1/day
P	CV_C_G	Number	Real	1/day
Q	xdate_cons	Date	Date/time	1/day
R	xjoin	Date	Date/time	1/day
S	Mach	Number	Integer	2/day
T	lb_movavg	Lbs	Real	2/day
U	lbhr_movavg	Lbs/hour	Real	2/day
V	act_movavg	Percentage	Real	2/day
W	st_movavg	Number	Real	2/day
X	cv_movavg	Number	Real	2/day
Y	avg_wgt	Lbs	Real	1/week

Table 5.1 (continued): Final Format of the Data Collected

Z	__cv_ofwgt	Number	Real	1/week
AA	skein_break	Lbs	Real	1/week
AB	break_factr	Lbs	Real	1/week
AC	s_e_grams	g/Tex	Real	1/week
AD	s_e____c_v	g/Tex	Real	1/week
AE	rkm_grams__tex	g/Tex	Real	1/week
AF	__elong	Percentage	Real	1/week
AG	__cv_elong	Percentage	Real	1/week
AH	wax_frict	Number	Real	1/week
AI	wax__cv	Number	Real	1/week
AJ	t_p_l	Number	Real	1/week
AK	hairs__3_mm	Number	Real	1/week
AL	uster__c_v	Number	Real	1/week
AM	uster__v_b	Number	Real	1/week
AN	vlc_1_yd	g/Tex	Real	1/week
AO	vlc_3_yds	g/Tex	Real	1/week
AP	vlc_10_yds	g/Tex	Real	1/week
AQ	vlc_50_yds	g/Tex	Real	1/week
AR	thin_places	Number	Integer	1/week
AS	thick_places	Number	Integer	1/week
AT	neps___200	Number	Integer	1/week
AU	cone_dia	Number	Real	1/week
AV	cone_wgt	Number	Real	1/week
AW	cone_dens	Number	Real	1/week
AX	wax_std	Number	Real	1/week
AY	efficiency_ma	Percentage	Real	1/week
AZ	yarn_breaks__1_100	Number	Integer	1/week
BA	stop_thick_short_s	Number	Integer	1/week
BB	stop_thick_long_l	Number	Integer	1/week
BC	stop_thin_t	Number	Integer	1/week
BD	stop_moire_m	Number	Integer	1/week

6.0 Analysis of Data Mining Using Enterprise Miner

In the first part of this chapter an explanation of how the Enterprise Miner software package functions is explained in order to get a better understanding of what has occurred in this research. This part of the research can now begin since the data collection and data preparation has been completed. In the second half of the chapter the results from the data mining will be reviewed.

6.1 SAS Enterprise Miner

Figure 6.1 shows a completed SAS Enterprise Miner diagram showing each labeled node. Before the diagram can be drawn a new project needs to set up. Creating a new project is very similar to creating a new Microsoft Word document. Once the new project is established a library needs to be assigned in order to have access to the dataset that will be used in SAS. Then nodes can be dragged onto the screen. Once this is accomplished each node can be opened to change the defaults that are already set up in SAS Enterprise Miner.

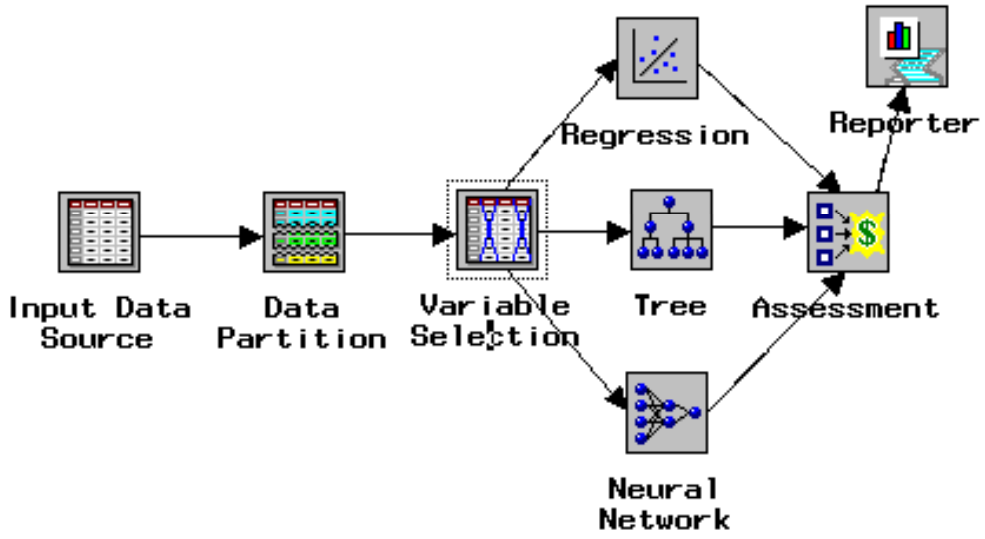


Figure 6.1: SAS Enterprise Miner Diagram for the Research Conducted

6.1.1 Input Data Source Node

The first step in the process is to apply the dataset to the input data source. This node is how the data that has been cleansed and conditioned for data mining is inputted into the process. Once the data is attached to the node, which is done by simply finding the path to that database. The model role of the data can now be defined. At this stage items may be rejected that are not useful to the mining or if the data is redundant. For instance, in the input data source that is being used in this research some of the data columns are rejected because these have been transposed into another column. The date-consumed column was used to create a new data column with the lag times so this column is not useful in the actual mining of this data so it is rejected. The target variable is identified in column one. This is done by right clicking the mouse and selecting set model role. This is also how a variable is rejected. Once this has been done changes can be saved to the Input Data source and the node is closed. The final data mart that was created from this research was 56 variables and 181 records. The data has duplications because of the time lags associated with production data. As stated earlier a laydown takes one day to be totally consumed so it will have to be duplicated to data in drawing that corresponds for that day. Also in this case there is no missing data because it is duplicated to handle the one to many relationships that exist in this data set.

6.1.2 Data Partition Node

The main purpose of this node is to separate the data into different groups. For instance, if lots of data is available then the it should be separated into training, validation and testing data. In this case study there were not thousands of records so it was decided to simply have 80% of the data training and 20% validation with no testing data. This decision is really dependent on the size of the warehouse. Once the node is in place an

arrow will be connected from the first node to the second. The arrow is accomplished by placing the mouse on the first node, clicking the left mouse button and then drag to the second node and release the mouse and the arrow will appear.

6.1.3 Variable Selection Node

The purpose of this node is to enable the evaluation of how important the input variables are when predicting or classifying the target variable. The target variable is what is used to create a data mining success story. For instance, in trying to detect whether fraud will occur or not the data must show which cases were fraudulent so that it may be used to predict if others will be as well. This node will remove variable based on their R-Square and Chi square values, basically it is looking for large percentages of missing values. In this research project this node has been included even though most of the conditioning was done prior to the data being entered into SAS. In this case the variables are not missing because of the work that was ahead of time and the lags that are created to be able to use the data points that remained. However, it is a good idea to always put this node in even when it might not be needed.

6.1.4 Regression Node

Two Crows defines modeling as an important function of data mining. "Modeling produces a model that can be either descriptive or predictive. A descriptive model helps in understanding the processes or behavior while a predictive model is an equation or set of rules that makes it possible to predict what is an unseen or an unmeasured value." [57] Modeling nodes are the choices within the SAS software package that allow for the different techniques available in data mining to be looked at these were discussed in the literature review. The regression node is a modeling node that enables the user to fit both a linear and a logistic regression model. Logistic

regression is a way to predict a binary variable such as a yes/no answer. [57] In this node there are no limitations to what variable is chosen for the target. It can be continuous, ordinal or binary. This node supports all of the following methods, these are chosen by using the selection tab after it has opened the regression node:

- Forward
- Backward
- Stepwise

Forward regression is a way to add variables to the model. The candidates are added when they are systematically significant to the target variable until none of the remaining effects meet the entry significance level or until the stop criterion is met.

Backward regression works in the opposite manner of forward regression. The backward regression begins with all the candidate effects in the model and it then systematically removes the ones that are not significant to the target variable. This method is not recommended when the target variable is binary or ordinal and there are many candidate effects or many levels to be dealt with in producing the model.

Stepwise begins the process with no candidate effects in the model and it then systematically adds the effects that are significantly associated to the target variable. Even though an effect has been added stepwise allows for the removal if it is really not significant. Stepwise uses cutoffs to both add and remove variables to the model. Stepwise is a combination of forward and backward regression.

As with all the other nodes so far in order to use it, it needs to be dragged onto the open desktop and then right click the mouse to open it. Once it has been open then the defaults may be changed, however, it is not necessary. At the same time it must be connected to the previous nodes for it to run.

6.1.5 Decision Tree Node

The decision tree node is also a modeling node. It allows for the fitting of a tree and it uses decision tree algorithms like CHAID (Chi-Squared Automatic Interaction Detection) and CART (Classification and Regression Trees). As with the other nodes to use it, it will be dragged onto the workspace and then be connected to the previous node. In this case that would be the variable selection node. The tree node can also be directly attached to the input data source node as well.

6.1.6 Neural Network Node

The Neural Network node also models the data. The purpose of this node is to construct, train and then validate using multiplayer feed forward neural networks. As discussed in the literature review neural networks are used to physically model large amounts of data, with many columns and possibly billions of records. In the case of this research there were many columns but only a few, 180, actual records so neural networks were unable to produce any results.

6.1.7 Assessment Node

This node is used to run the models and assess each modeling technique in one place. The assessment node shows charts such as: profit, return on investment, diagnostic charts and threshold based charts. In our case these are not very informative charts because this information is not present in the data that is involved on a spinning shop floor. However, it is a good practice to always include this node in a model.

6.1.8 Reporter Node

The reporter node is a way to generate an html report that contains all of the process flow analysis. Each report gives headers, an image of the process flow that was used like in Figure 6.1 and a sub report of each of the nodes contained in the diagram. This is a great way to remember what was chosen in each node especially if changes to

the defaults that are contained in each node are made. An example of the report that is generated from this node can be seen in Appendix E.

6.2 Results

The results that will be shown in this section of the dissertation are all based on the modeling of using the yarn breaks as the target variable. In this research it was decided that yarn breaks was a good variable to analyze versus the data that was collected. Each node that has already been explained in sections 6.1.1 through 6.1.8 will now show examples of the outputs of the data mining example in this portion of the dissertation.

6.2.1 Input Data Source

The data mart that has been discussed in Chapters 4 and 5 will be inputted into this node to have the data available to use in SAS Enterprise Miner software package. The data set that has been entered is shown in Table 6.1. The information that is provided in this table are the variables that are in the dataset, their model role, measurement type, format and what type of variable and the label for the variable.

Table 6.1: Input Data Source Variables

Name	Model Role	Measurement	Type	Format	Variable Label
LAYD_NO	rejected	interval	num	BEST12.	LAYD No
MIC	input	ordinal	num	BEST12.	MIC
__5__SL	input	binary	num	BEST12.	2#5% SL
_UI	input	binary	num	BEST12.	UI
_STR	input	ordinal	num	BEST12.	STR
_EL	rejected	unary	num	BEST12.	EL
__RD	input	ordinal	num	BEST12.	Rd
__B	input	binary	num	BEST12.	+b
C_G	input	ordinal	num	BEST12.	C-G
UV	input	ordinal	num	BEST12.	UV
DATE_CONS	rejected	nominal	char	$7.	DATE CONS
CV_MIC	input	interval	num	BEST12.	CV%Mic
CV__STRENGTH	input	interval	num	BEST12.	CV% Strength
CV_RD	input	interval	num	BEST12.	CV%Rd
CV__B	input	interval	num	BEST12.	CV%+b
CV_C_G	input	interval	num	BEST12.	CV%C-G

Table 6.1 (continued): Input Data Source Variables

Name	Model Role	Measurement	Type	Format	Variable Label
XDATE_CONS	rejected	ordinal	date	DATE7.	
XJOIN	rejected	interval	date	DATE7.	
MACH	input	ordinal	num	BEST12.	Mach
LB_MOVAVG2	input	interval	num	BEST12.	LB
LBHR_MOVAVG2	input	interval	num	BEST12.	LB/H
ACT_MOVAVG2	input	interval	num	BEST12.	ACT%
ST_MOVAVG2	input	interval	num	BEST12.	ST
CV_MOVAVG2	input	interval	num	BEST12.	CV%
AVG_WGT_LAG3	input	interval	num	BEST12.	AVG WGT
CVOFWGTLAG3	input	interval	num	BEST12.	% CV of WGT
SKEIN_BREAK	input	interval	num	BEST12.	SKEIN BREAK
BREAK_FACTR	input	interval	num	BEST12.	BREAK FACTR
S_E_GRAMS	input	interval	num	BEST12.	S#E# GRAMS
S_E____C_V	input	interval	num	BEST12.	S#E# % C#V#
RKM_GRAMSTEX	input	interval	num	BEST12.	RKmGRAMS/TEX
ELONG_LAG3	input	interval	num	BEST12.	% ELONG
CV_ELONG	input	interval	num	BEST12.	% CV ELONG
WAX_FRICT	input	interval	num	BEST12.	WAX FRICT
WAX__CV	input	interval	num	BEST12.	WAX %CV
T_P_I	input	interval	num	BEST12.	T#P#I#
HAIRS__3_MM	input	interval	num	BEST12.	HAIRS >3 mm
USTER__C_V	input	interval	num	BEST12.	USTER% C#V#
USTER__V_B	input	interval	num	BEST12.	USTER% V#B#
VLC_1_YD	input	interval	num	BEST12.	VLC 1-YD
VLC_3_YDS	input	interval	num	BEST12.	VLC 3-YDS
VLC_10_YDS	input	interval	num	BEST12.	VLC 10-YDS
VLC_50_YDS	input	interval	num	BEST12.	VLC 50-YDS
THIN_PLACES	input	interval	num	BEST12.	THIN PLACES
THICK_PLACES	input	interval	num	BEST12.	THICK PLACES
NEPS___200	input	interval	num	BEST12.	NEPS (+200%)
CONE_DIA	input	interval	num	BEST12.	CONE DIA#
CONE_WGT	input	interval	num	BEST12.	CONE WGT
CONE_DENS	input	interval	num	BEST12.	CONE DENS
WAX_STD	input	interval	num	BEST12.	WAX STD
EFFICIENCYMa	input	interval	num	BEST12.	Eff. ma# (%)
YARN_BREAKS__1_100	target	interval	num	BEST12.	Yarn Breaks (1/1000 RH)
STOP_THICK_SHORT_S	input	interval	num	BEST12.	Stop thick short
STOP_THICK_LONG_L	input	interval	num	BEST12.	StopThicklong
STOP_THIN_T	input	interval	num	BEST12.	Stop thin
STOP_MOIRE_M	input	interval	num	BEST12.	Stop Moire

Once the data is entered into SAS, the program gives a breakdown of these variables. For instance, the minimum number, maximum number, mean, deviation and %

missing are all in this breakdown. A small sample of the variables in this data set can be seen in Table 6.2.

Table 6.2: Breakdown of Variables

Name	Min	Max	Mean	StdDev	Missing %
LAYD_NO	2000032.00	2000052.00	2000041.85	0.000	0
CV_MIC	6.09	7.85	6.73	0.740	0
CV__STRENGTH	3.91	6.24	5.25	0.712	0
CV_RD	1.45	1.87	1.63	0.116	0
CV__B	5.03	5.55	5.29	0.141	0
CV_C_G	3.45	7.27	5.59	1.012	0
XJOIN	14953.00	14985.00	14967.05	0.000	0
LB_MOVAVG2	944.00	4868.00	3418.49	665.760	0
LBHR_MOVAVG2	246.75	573.75	524.54	40.957	0

6.2.2 Data Partition

The data partition node takes all of the variables from the input data source node and then separates these out to separate data sets. Basically in this research 80% of the data was allocated to the training data set and 20% to the validation data set. In this node there really is not any output to see.

6.2.3 Variable Selection

In this node the variables are selected. The main reason for rejection of the variable in the data set that is being used in this research is because of a low R squared. Some examples from the SAS outputs can be seen in Table 6.3.

Table 6.3: Variable Selection Examples

Name	Role	Rejection Reason
MIC	input	
__5__SL	rejected	Low R2 w/ target
STOP_THICK_SHORT	input	
STOP_THICK_LONG	input	
CONE_DENS	rejected	Low R2 w/ target
HAIRS__3_MM_LAG3	rejected	Low R2 w/ target

6.2.4 Regression

The regression node does a basic statistical analysis. Figure 6.2 shows a distribution of the yarn breaks across the sample. It allows us to see in a visual way where the data falls on the scale. The regression node also determines the t-value, F-

Figure 6.2: Percentages of Yarn Breaks Allocated

value, standard error and degrees of freedom for the variables that have been selected in the variable selection node.

6.2.5 Decision Tree

The decision tree node creates a tree diagram based on a series of simple rules. Each rule will assign an observation to a segment/node of the tree based on the value of the input. The leaves on the tree are determined by the F test and the split allows for minimum of one observation to fall into each leaf. The original node is only to have two branches sprouting out of it. The depth of the tree is only allowed to be six. From these guidelines Figure 6.3 was created again using the yarn breaks as the target variable. Inside the node the tree diagram that has been created can be viewed. However, in this version of SAS Enterprise Miner the whole tree diagram cannot be extracted to Word.

Also within the node there is information about the complexity of the tree, it allows for the viewing of the split balance. Figure 6.4 shows an example of a tree diagram for the yarn breaks example. As can be seen in this Figure the tree diagram is very similar in layout as an organizational chart.

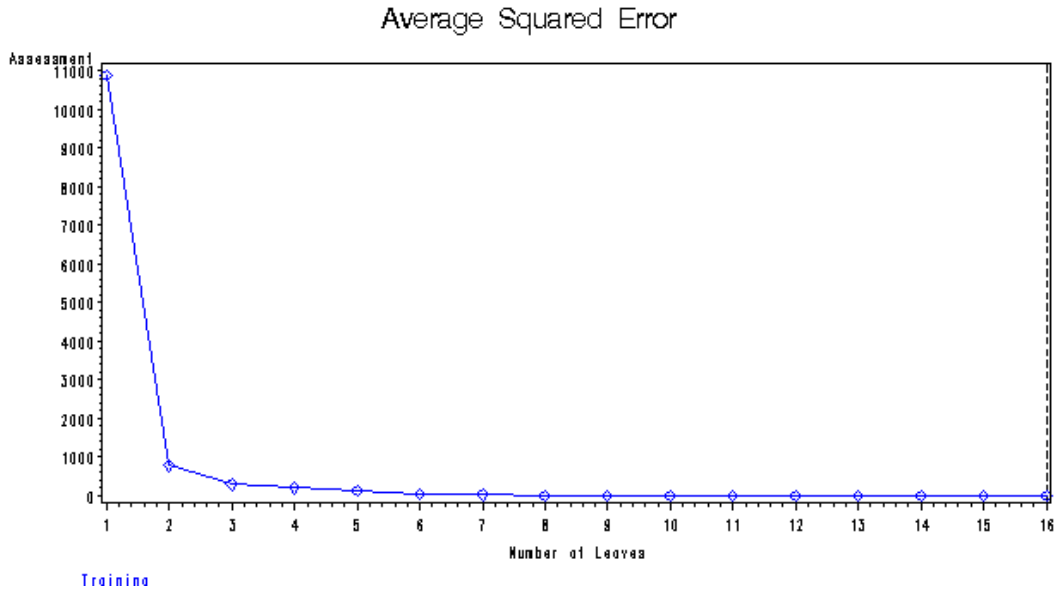


Figure 6.3: Decision Tree Model Assessment Plot

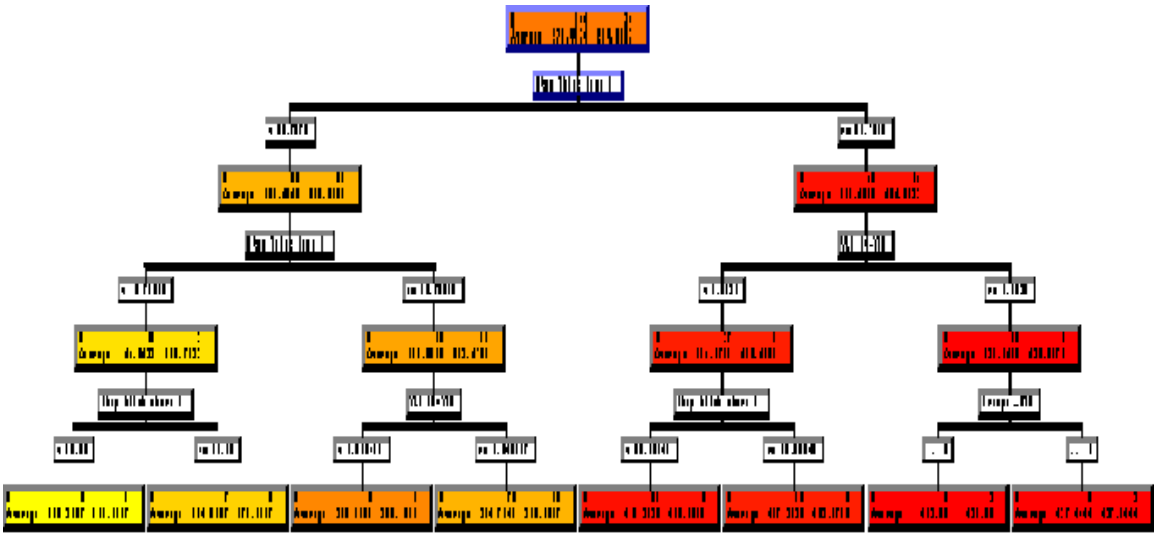


Figure 6.4: Decision Tree Model

6.2.6 Neural Networks

The Neural Network node was not very useful with analyzing the data set that was prepared for this research. The dataset was much too small to be used. In Figure 6.5 the results of the neural network node can be seen. From looking at the results it can be seen that there was an error for this function. Therefore, it has been verified from this research that a neural network is only useful for larger databases.

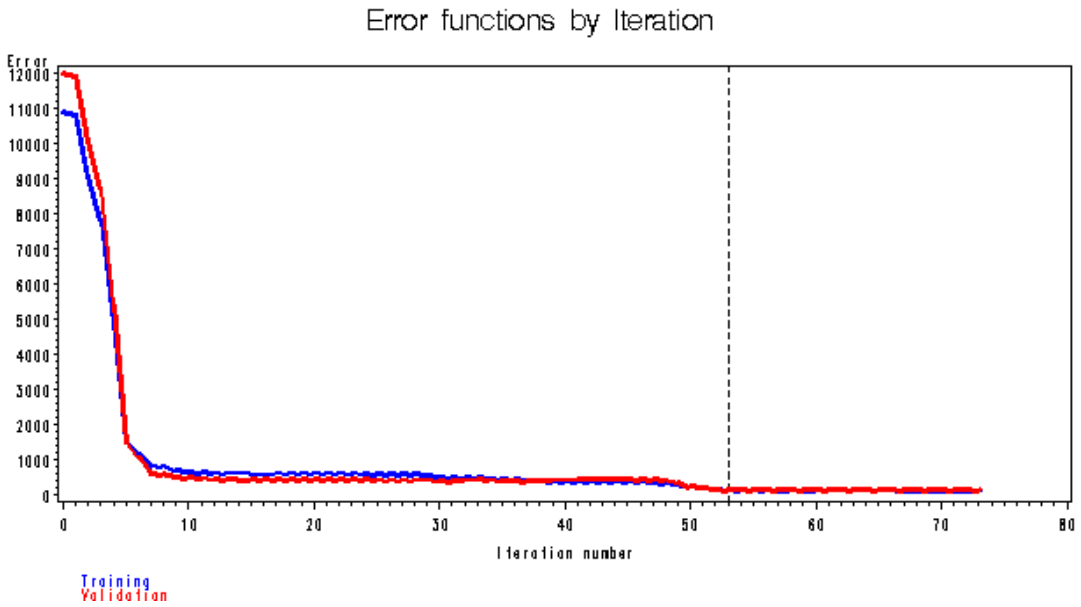


Figure 6.5: Neural Network Optimization Plot Results

6.2.7 Final Comments

The different models that were tried were the regression, neural net and decision trees. Each of these modeling techniques was attempted using different target variables. Some of the variables that were set as target variables are as follows: Micronaire, skein break, Micronaire CV, Break Factor, and yarn breaks. Even though this can be thought of as a cause and effect analysis the Micronaire and Micronaire CV were tried just to be sure that this was true. The target should be an effect and this was proven in the analysis of trying the Micronaire variables.

Visualization techniques are used for each modeling method. From the original diagram that is drawn, example Figure 6.1, to the different graphs that are present for each node and modeling technique. There are also visualization techniques involved in the final layout of the results that are presented in the final report that is created; this can be seen in Appendix E. The information that is gathered is not in just tabular format but when possible graphics are presented to make understanding all the results more effective.

In this research the results did not prove to be very effective. Overall the data seemed to be almost perfect in significance factors, which really was not the case, see Appendix E. The main contribution to this is believed to be due to the fact that each process was collected for different time frames and because of this a lot of summary data needed to be used. It was also difficult to create lags that were completely accurate to the hour because some of the processes did not have the time available for this research. As with any new idea there are learning curves and in this case it is believed that some of this could have been prevented if a model of how to go about mining data for a manufacturing facility already existed prior to this endeavor. Since a model does not exist one has been created to help others through this grueling process in the future, see chapter 7.

In the case of this project a time basis is used to match the different information, however it is nearly impossible. The reason for this is that there is a lot of guessing involved in knowing when the actual product would have moved from one machine process to the next. Even if time lags are established properly there could still be discrepancies. Mainly this is because of the lack of data that is collected. In the case of

this data there is not enough data points to be able to match across processes. Also there is too much summarization of the data. Another issue with the data is the actual storage of it.

7.0 Data Mining Process Model

In the course of reviewing the literature for this research project it became very apparent that a process needed to be established for future projects to be successful, initially a six-step process was developed which can be seen in Figure 7.1.

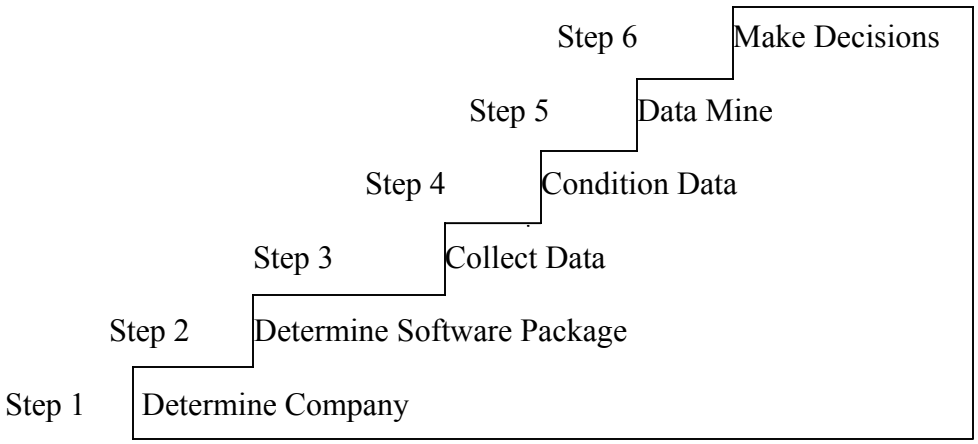


Figure 7.1: Initial Six-Step Process for a Data Mining Project

Once the project was underway it was determined that the initial six-step process was a good starting point but in practice it was not nearly enough for any company to work with to craft a successful data mining effort. Also during this time it was established that some of the initial steps were not correct, these will be discussed in more detail during this chapter and can be seen in Figure 7.2 this is the revised six-step process that will be used to expand the new data mining process model. The initial six-step process was derived from what seemed to be the most logical way to attempt to gather what was needed to mine their data. It was initially thought that if the first step could not be completed then the staircase to the next step in the process could not be attempted. It was determined that in fact that was not the way the data mining process worked but in fact it had more parallels to the childhood game of Chutes and Ladders than it did to a staircase. By using this analogy to explain the process makes it more understandable to a

larger audience. In working through this research it became very apparent that the process was not a lot of steps that needed to be completed to move on to the rest but in actuality if a step is completed inaccurately then they would have to slide back, or beyond, to a previous step hence the chute. In the data mining process model the chute will be represented by "If Then" statements that if not answered positively would cause the slide to an earlier step. At the same time ladders are represented by connecting arrows that allow for the skipping of steps in the process because they are already completed or a positive answer to the question occurs.

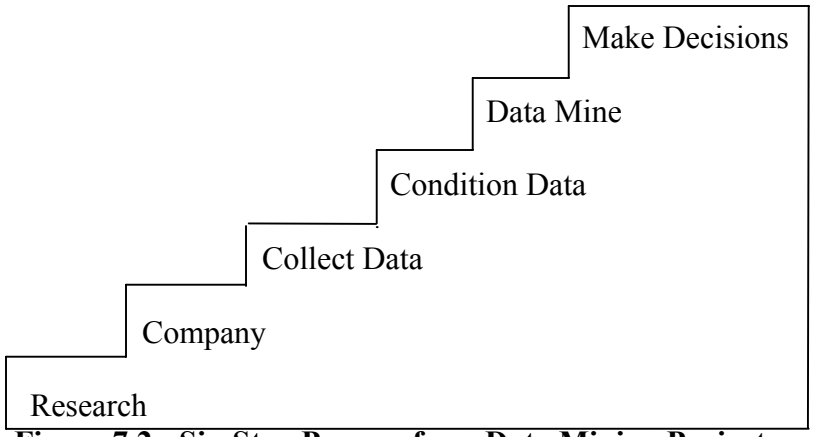


Figure 7.2: Six-Step Process for a Data Mining Project

The six-step process was used during this project to begin the gathering of data in order to answer the questions of whether or not data mining can be used in a manufacturing environment but more specifically in a Textile spinning facility. In this chapter the new data mining process model will be discussed. The entire model that will be discussed in this chapter can be seen in Figure 7.3.

7.1 Overview of the Model

The main purpose of this new model is to help anyone that wants to data mine be able to mine their data. The process itself is done by following the rules of flow-charting to be able to determine step by step what needs to be done in order to meet the end goal

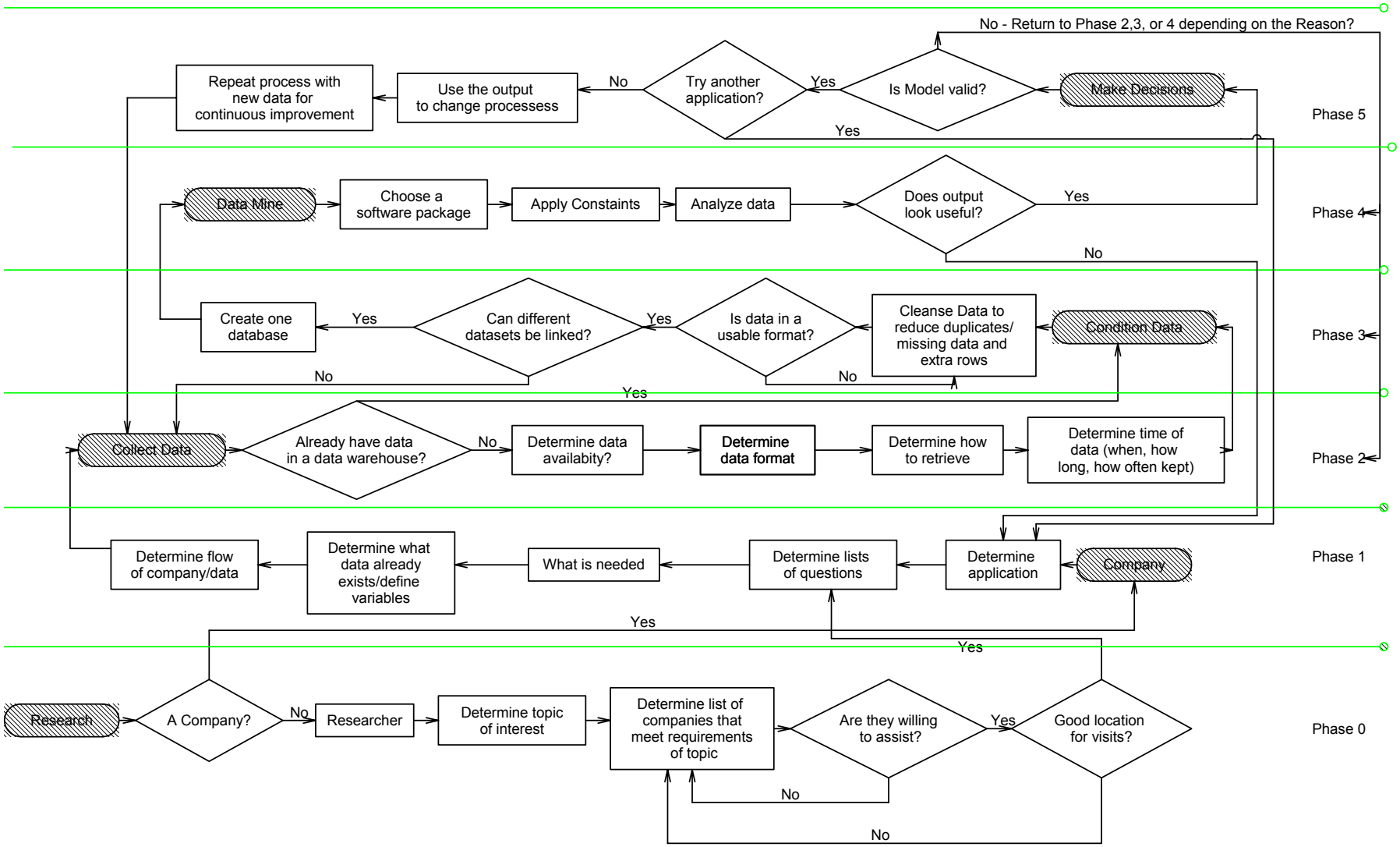


Figure 7.3: Schertel Data Mining Process Model

of the project. The purpose of this model is to minimize the learning curve for companies using it because the process is already laid out for them.

The actual layout of the process model is that it starts at the left bottom corner of the chart and works its way to the top left corner. Each step that was in Figure 7.2 represents a row or phase of the process. For example the Research step is represented in Phase 0. The process begins with Phase 0 because this phase is only necessary for a researcher. The data mining process really begins in Phase 1. Each step from Figure 7.2 can be denoted by an oval with violet diagonal lines running through it, refer to Figure 7.4 for this representation.

Figure 7.4: Step Representation

Diamonds that have two arrows coming out of them and linking to another step represents the "If-Then" statements. On each of these diamond diagrams there are two separate lines that exit the box these help determine where to go from the box depending on the answer to the question within the box. This is what was referred to earlier as the "Chutes and ladders" analogy. A chute would cause a backwards step in the process while a ladder would help movement forward and in some cases skip steps in the process. An example of this representation can be seen in Figure 7.5.

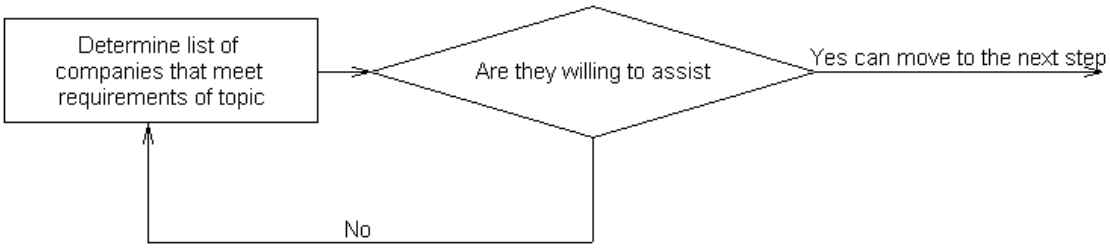


Figure 7.5: Flow Chart Representation

Another type of diagram is a rectangle this represents the smaller sub steps of each major stair of the staircase. In order to move up the staircase the work in the major stair must be completed first. The question that can be raised is what is really involved in each step? The data mining process model begins to break down the issues that must be resolved to more effectively mine the data. Figure 7.6 also shows how the rectangle is represented.

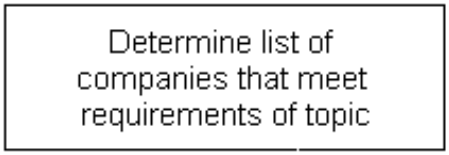


Figure 7.6: Sub Step Rectangle Representation

The arrows that connect the different diagram representations help guide the individual through the process depending on the output of each step. These can also be seen in Figure 7.5. Without these arrows there would be no clear cut path of what step should be worked on next. These add order to the whole process model.

On the end of each row there is also text that helps to verify what phase of the process is currently being worked on. At the same time there is also a line between the phases that helps distinguish what information falls in which category more easily. Each Step in the revised six-step process represents a separate phase in the model. Each lower phase in the process must be completed in order to move up to the next Phase. For instance, it is not possible to be on Phase 3 without completing Phases 1 and 2.

7.2 Data Mining Process Model Phase explanations

7.2.1 Phase 0: Research

Phase 0 is for the sole purpose of a researcher that does not belong to a specified company or project. This phase has been included in the diagram for future projects like

the one that has been discussed in this research. If a company, the ladder part of the "If Then statement" is used to move to Phase 1. The first sub step of the phase after determining that a company is not involved but that a researcher is involved is to determine the topic of interest that could prove useful if investigated. Usually this is done by seeing what has been done in the past or simply by reviewing where there might be a need. After the topic has been decided it is time to decide a list of what companies meet the requirements of the chosen topic. This could be as simple as in this project of deciding first that the topic would be a spinning mill and then the second to list all known textile companies. After this has been completed it is time to move to the next sub step, which is finding out whether or not any of the companies on the list would be willing to assist in the project at hand. At this stage there may be many different choices or just one. This all depends on what type of research is planned. In some fields companies may determine that their data is too confidential to help with the research. The last main question to deal with in this phase is whether or not the company's proximity to the location is adequate or not. If a company is chosen that is more than a days drive away it may be hard to answer questions that can only be dealt with in person. This should be considered especially if on a tight budget. For instance, if the data does not really make any sense then it is easier to go investigate where and how it is generated than to have to rely on hearsay or another person's description, as they say a picture is worth a thousand words. Once these sub steps are completed it is time to move onto Phase 1.

In Phase 0 the process was generated from the literature research for this project. The literature helped to determine what had been done in this field already. The questions that are in this section were not changed because of the project. In other words

in this Phase there were no surprises from what was expected to what was determined and the literature was enough to work through this phase accurately.

7.2.2 Phase 1: Company

In the company process step there are two ways to get to this phase. Either the individual has completed Phase 0 or Phase 0 was skipped because they are a company and there is no reason to do Phase 0. In any case, once Phase 1 begins there are two separate beginning paths if this is not a research project but rather a company then the company takes the ladder directly to this Phase 1. At this point they will settle on what type of application they want to use the data mining for in their business operations. This can be done through brainstorming sessions. If a researcher there is no need to redo the first subset as this was decided in Phase 0 therefore the researcher should skip to the second sub step in the Phase. At this sub step the company joins in to follow the same path with the researcher. At this time the list of questions that should be asked will be listed and the discussion about what data needs to be collected in order to answer the questions at hand. Once this has been decided it will be imperative to decipher what already exists for data variables. At this time it will be helpful to know what the variables are, how these are collected and what the definitions are of each one. It can be tricky to get the exact definition at this point but it will be helpful to know as much as possible up front in the process. At this stage the more information that is known the better off the process will be in later phases. Once this sub step has been completed it will be important to establish how the data flows through the process. This helps to resolve how the actual database will be set up.

Phase 1 was mostly derived from the literature like in Phase 0. The literature helps in understanding the importance of background work it also establishes the type of information that is needed to data mine. In this phase the information here is still fairly basic and obvious to even a first time data miner.

7.2.3 Phase 2: Collect Data

Phase 2 is one of the most critical in the process. In this stage it is important to lay the groundwork for what will happen in later phases. The first sub step is to conclude whether or not a data warehouse exists in a company. The main question to ask is does a data warehouse exist for the application that it will be used for? If there is a warehouse it may not be updated, contain the right information or be in a format that will work. If indeed it does exist then the company may want to bypass the rest of this step and move onto Phase 3, Condition data. Otherwise continue the process model to the next sub step, which is to determine what is available for data. There may be many systems that contain different information that needs to be joined together in order to meet the goals of the application. At this time it is also essential to find the format of the data. This could be a paper format that came from a system that only prints, it could be a flat file, ASCII file or could be on a disk. This sub set is very important because if the data cannot be accessed then the application will fail. At some point it will be imperative to determine how and where the data is generated and how best to access it. If the data comes from a summarized database then the information that is needed could already be sifted out from the original data set. Therefore, the "gold nuggets" of information may never found because they are no longer present in the database. Because of what was just stated the next sub set seems almost self-explanatory. This sub step is to determine how the data

should be retrieved. Is it something that is collected by the minute, hour, half-day, day or by the month?

The data needs to be useful for the intended application. If the data generated is not done often enough or maybe too often then problems may arise. The main issue with it being done too often is how long it will take to actually run the algorithms in the data mining packages. Also under this sub step the data must be considered because sometimes it is retrieved from different programs that will need to be able to dump into one common program for consistency. Phase 2 was determined both from the literature and from lessons learned in this project. The main issue with this project is the problem with the gathering of data across many different operating systems that then collect data at different time intervals. See Table 7.1 for a comparison of the different systems and how the data is collected for a spinning mill as an example of what a company may be up against. As can be seen in this table there are discrepancies when the data is collected. In other words there is no matching of time. When trying to link data it is important to be able to tie with a commonality. In most data this would be some type of time element.

Table 7.1: Chart of Current Data Systems and Collection of Data

SYSTEM FOR WHAT PROCESS	WHAT IS ON THIS SYSTEM?	DATA SYSTEMS	HOW OFTEN IS DATA COLLECTED
EFS	Cotton Inc. proprietary HVI	Cotton Inc.	Per day
Drawing	Uster sliver data	Uster version 2.6.4	Per Shift/Per Frame
Spinning	Production information (modem)	Schlafhorst -WINCI	Per Shift
Yarn Quality	Quality information - offline	Excel 2000	Min. once/week

Some companies only store their data for a small amount of time. What if for instance a new database is created using old data then there needs to be a common time thread. This issue involves more production type data than customer data. Customer data is not really time sensitive except for knowing how old that information really is. The production data however is time sensitive in that if comparing old data it is essential to have data that is for the same time period elements. In Table 7.1 there are four different process systems with separate time elements of how long the data is stored on computer. An example of data time element issues is that if out of these four systems only two have data from two months ago and the other two systems have only kept data for one shift or one week then the longest matching of the old data is whichever has the least amount of data is the amount that must be used. Otherwise there would be a lot of missing data values within the newly created database not allowing data mining to work properly and thoroughly. Once this type of information is verified and understood then one can move on to the next step, which is the actual conditioning of the data.

7.2.4 Phase 3: Condition Data

Conditioning the data for the use in data mining is the most time consuming and difficult task that is involved in the process. Many companies think that they have data that is in great shape and can be mined almost immediately. This is for the most part never the case, even with companies that have decent data it can still take almost 60% of the time.

In a recent conference, M2001 Data Mining at SAS Institute in Cary, North Carolina [59], Dick De Veaux indicated that cleansing the data is more of a project than originally anticipated. In earlier literature [1, 14, 24] it was stated that the

cleansing/conditioning of the data took anywhere from 60% to 95% of the time to actually preparing the data. However De Veaux, said that the only time it is 60% of the effort is when one already has a data warehouse established. He said in all actuality most cases are upwards of 95%. For example, he is currently working with a plane wing manufacturer that they have been cleansing and reworking their data for one year already and it still cannot be considered useful. Jerry Oglesby, Jim Georges, and Bob Lucas of SAS Institute and Herb Edelstein of Two Crows Corporation all have confirmed that it is more time consuming than originally published [62].

The main reason for this cleansing/conditioning is that in most cases the data is in such disarray. There will be data that is missing. For instance, if a telecommunications company has a customer that purchases service from the company they most likely fill out a service agreement. Maybe in filling out their application to be granted service they leave a question blank. The big question ends up being what does that company do with the blank space? Do they average other lines of data that are close in properties to get what they estimate should go there or maybe they call the person back and reenter the correct data. In many cases this will be infeasible. If the data is not cleansed properly and kept cleansed then it becomes very difficult to evaluate the data. In data mining systems if there is missing data then it is thrown out of the mix when analyzing the data unless a smoothing type algorithm is utilized. In a case that was done at SAS Institute approximately 60% of the rows had at least one data element missing. Other cleansing issues involve the transfer of the data into one system. In some cases the data will not import correctly. As mentioned earlier in the dissertation sometimes data points that should only be on one row become 3 rows. In this case there are not only too many rows

used to equal one data set but then there also becomes the issue of lots of empty/missing data points. In this case one will have three times as many blanks than if it imported correctly.

Another issue is that of having extra data. For instance, catalog companies send out catalogs in order to create sales. If a catalog company sends one to a Ms. Diane Floor and one to Diane Frances Floor then they have duplicate entries of the same individual. These same phenomena can also occur with addresses. Let's just say that Diane gets a catalog at 2323 Bowing Street and then Frank Smith also gets one at 2323 Bowing Street maybe possibly only one catalog needed to be sent because it is one address with two people sharing an apartment. Is it still necessary to send out two catalogs? In some cases it may just be a waste of money and resources.

During the cleansing/conditioning of the data the data linkages need to be established. For instance, in manufacturing the linkages need to be across the production process. Usually this will be done by time lags. In customer data it will be linked by a customer number variable. After the links have been determined one database can be created that contains all the information needed to mine the data. The mining can only be done from one database that contains all information in a usable format.

7.2.5 Phase 4: Data Mine

If all phases up until this point are done accurately the mining itself should not be too difficult. However, before any mining can occur a software package needs to be selected that will work best for the application and situation of the company. For instance, if a company is using DB2 software from IBM then they should use the IBM Intelligent Miner package because these are a one to one type application. The software

package that was used for this project was SAS Enterprise Miner. Enterprise Miner allows the importing of data directly from other programs such as: Excel as long as the data is in a recognizable format. Table 7.2 (which for clarity is a repeat of Table 2.1) shows the key data mining players and what these have to offer along with the platforms that these work on. Companies need to review price, features and database (are they relational 1 to 1) compatibility before making the decision of which one to use.

Table 7.2: Data Mining Vendors Compared

Vendor: Product Name	**IBM:** Intelligent Miner	**Oracle** (Thinking Machine): Darwin	**SAS Institute:** Enterprise Miner	**Angoss Intl.:** Knowledge Seeker	**NCR Corp.:** Teraminer Stats
Platform	AIX 4.1, NVS, AS/400, Windows NT	Windows, Windows NT	MacIntosh, Windows	Windows, Windows NT, UNIX	Windows NT, UNIX
Algorithms (linear, regressions and/or clustering	X	X	X	X	X
Decision Trees				X	
Neural Networks	X	X	X		X
Bayesian (learning) Networks		X			
Visualization	X	X			X
Website	[46]	[47]	[44]	[48]	[49]

Once the package has been decided upon it will then be time to proceed with learning how the system works and how best to utilize the features. A data mining program has many different statistical tools within the package and therefore it is

important to run many different scenarios to be sure the one that is chosen is the best representation of the data being processed.

After the actual analysis is complete it will then be time to review the results. Is the information, nuggets of gold, really useful? In the literature [7, 28] there was an example of an insurance agency that learned trends about some niche market of sports car drivers and were able to use this information to sell a better insurance package to those individuals. However, there is the well-known market basket example of beer and diapers. In this data mining expedition it was determined that a shopper purchasing diapers on a Thursday evening also has a high percentage of purchasing beer at the same time. The output of this data-mining scenario was not necessarily a gold nugget of information because there has not been any determination of what to do with it. The question is does the grocery store put beer and diapers next to each other on the shelf? Maybe advertising an item at a lower price and one at a higher price in order to make more money on the other? It is information like this that can be seen as useless to someone that is mining his or her data. If this is the type of information that is being generated from the analysis it may be important to return (slide down the chute) to Phase 1. If the output is gold nuggets of information then one may continue in Phase 4. A gold nugget is something that has been found within the data that would not have been found without it. However, it also is categorized as something that is useful to the company that finds it. It needs to be able to be acted upon.

7.2.6 Phase 5: Make Decisions

The last step in the process is to make the decisions from what has been generated from mining the data in order to answer the questions that were raised that caused the use

of data mining in the first place. From this output of information it can be determined whether this is something that is a one-time change or whether it would be a continuous improvement item. For instance, if it is a process change in that a machine is causing a lot of downtime and not producing product up to specifications then that particular machine could be either retired and replaced or maintenance could be done on the machine to bring it back to specifications. This could be used for a one-time look at how the machinery is performing or it could be continuously watched to be sure that the problems do not arise again in the future. In reviewing the output from the data mining it should be reviewed to help change the process in the future. It is there to help improve what is already being done in the facility so that the process is more efficient than it was prior to the data mining. No company can be considered perfect and because of this the output needs to be used to change the way things are being done currently. After the decisions have been made the company must return and work their way back through the process by collecting more data and continuously watching and using the new and cleansed data to make more improvements. At this stage it will be important to sift through what was collected in the process that was just completed and be sure that whatever is being collected is useful. In some cases there may not be enough collected and in others there may be fields that are not necessary to get the same answers. Also if the process works for this company it would be smart to start the process over at Phase 1 in which they decided on an application. If it can work for one problem it should work for others. From this entire process there should be a recommendation about what types of data should be collected, how and in what format.

7.3 Ideal Data Warehouse

In this section the ideal warehouse data points will be shown by the process within a spinning mill. Beginning with Laydown data and ending with the yarn data. All of the data that are mentioned in these sections should be collected for a minimum of one year a twice a day intervals. All of the abbreviations and explanation of terms that are to follow can be found in Chapter 4.

7.3.1 Bale Data

Bale data should be collected on each bale that will be included in a laydown. The reason for this information is so that it may be used to get the CV on each data point that is collected on the Laydown data if so desired. Data points that should collected are as follows:

Name	Format
♦ Bale number	Integer
♦ Micronaire	Real
♦ Uniformity Index	Real
♦ Strength	Real
♦ Elongation	Real
♦ Colorimeter Rd	Real
♦ Colorimeter B	Real
♦ Color/Grade	Real
♦ Ultraviolet	Real
♦ Laydown Number (included in)	Integer

7.3.2 Laydown Data

Laydown data should be collected for each laydown that has entered the plant. The most important information about this data is the date it starts processing in the laydown. It will also be important to get the actual time frame of when it starts to be processed and when the next laydown begins the process. The following list shows the data that should be collected.

Name	Format
♦ Laydown number	Integer
♦ Micronaire	Real
♦ Uniformity Index	Real
♦ Strength	Real
♦ Elongation	Real
♦ Colorimeter Rd	Real
♦ Colorimeter B	Real
♦ Color/Grade	Real
♦ Ultraviolet	Real
♦ Data Entered process	Military Time/date (mmddyy)
♦ Date Completed process	Military Time/date (mmddyy)
♦ CV of Micronaire	Real
♦ CV of Strength	Real
♦ CV of Rd	Real
♦ CV of B	Real
♦ CV of Color/Grade	Real
♦ CV of Elongation	Real

7.3.3 Carding Data

In this case study the carding data was not used because of the lack of data. In the case study facility the carding data was only collected once per month and was therefore not useable in this research. However, in future studies it would be beneficial to have this data available. The only way that this can occur is that it must be collected more frequently. The data that should be collected on the carding process in the list that follows. All of this information should be categorized by the date and machine in which it has been collected on.

Name	Format
♦ Date and Time of Reading	Date(mmddyy) /Military time
♦ Card Number (machine number)	integer
♦ Neps for Mat, Sliver and Percent Elimination	Real
♦ Dust for Mat, Sliver and Percent Elimination	Real
♦ Trash for Mat, Sliver and Percent Elimination	Real
♦ Visible Foreigh Matter for Mat, Sliver and Percent Elimination	Real
♦ SFC(n), SFC(w) and 2.5%(n) for both Mat and Sliver data	Real

7.3.4 Drawing Data

The drawing data should be collected by machine, date and time. In this case study this was done however, the data that was collected is only for two separate times one being 7 a.m. and the other 7 p.m., which can be thought of as the end of the shift. There is no data collected in between these two hours. The following bulleted list is what should be collected at this process.

Name	Format
♦ Date	Date/time
♦ Time	Date/time
♦ Machine Number	Integer
♦ Count (sliver weight)	Real
♦ Pounds per Shift	Real
♦ Pounds per Hour	Real
♦ ACT% (actual efficiency)	Real
♦ Stops per Machine	Real
♦ Coefficient Variation Percentage	Real

7.3.5 Spinning Data

The spinning data should not only be collected by machine but also by spindle number. The data that is collected should not be data that has been summarized per shift but constant running data. Again this data should be collected and stored for more than the summary of the shift, as with the drawing data there should be data throughout the shift. The data that should be collected for this process is as follows:

Name	Format
♦ Date	Date/time
♦ Time	Date/time
♦ Machine Number	Integer
♦ Spindle Number	Integer
♦ Yarn Count	Real
♦ Efficiency of Machine	Real
♦ Yarn Breaks	Real
♦ Stops Thick Short	Real

- Stops Thick Long — Real
- Stops Thin — Real
- Stops for Moire — Real
- Rotor Speed — Real)
- Twist — Real) can be derived from each other
- Delivery Speed — Real)

7.3.6 Yarn Data

The yarn data was being collected for each machine at least once per week. In creating a new dart mart it will be important to get tests of at least one package per machine per day. This data is very important because it can tell someone the actual quality of the yarn that has been produced. Some things that should be in this spreadsheet are as follows:

Name	**Format**
Date	Date/time
Count	Real
Frame	Integer
Average Weight	Real
CV of Weight	Real
Skein Break	Real
Break Factor	Real
Single End Grams	Real
Single End Percent CV	Real
Percent Elongation	Real
Percent CV Elongation	Real
Wax Friction	Real
Wax Percent CV	Real
Twists per Inch	Real
Hairs greater than 3mm	Real
Uster Percent CV	Real
Yard Weight Variability for 1 Yard, 3 Yard, 10 Yard and 50 Yard Tests	Real
Thick Places	Real
Thin Places	Real
Cone Diameter	Real
Cone Weight	Real
Cone Density	Real

As with the case study that has been discussed in this dissertation all of the processes and their data elements need to be combined to create one data mart. This should be done similar to how it was done in this research but with more data per day as discussed in Section 7.3. It will be important to collect at least one year's worth of data to really get results that are based on a lot of inputted data.

8.0 Conclusions and Recommendations for Future Studies

8.1 Conclusions

The research in this study focused on the use of data mining in a textile spinning mill and whether or not it would prove to be useful in that area. The methodology involved a single case study approach because of the time burden involved in mining data. In this case study data was gathered from a plant that was willing to participate in the research and that met the initial guidelines when choosing one. The data was collected from the spinning process and the machines involved. Many of the questions that were proposed and discussed throughout the research will be recapped in this section.

1. Who uses different data and information? In this case study the main users of the data that is collected is those working in the plant, running the machines. However, the vice president of manufacturing also gets copies of the Excel spreadsheets each week to monitor the efficiency of the plant and the quality of the product. The information that is gathered and entered is not everything that generated from the machines it is simply a summary of the data.
2. What is the format is of the data? The data format has been discussed in this dissertation in Chapter 4 and 5. Figure 4.1 gives a layout of the plant and what each monitoring system is involved in each phase of the production process. Appendix C shows another layout of this plant. Table 4.1 lists the different systems that are involved at each stage of the process with how long this data is kept. Table 5.1 lists the actual format of each attribute that was collected for this project. Chapter 7 discusses the Ideal Data Warehouse that has been proposed from this research.

3. In order to acquire the data for the data warehouse it is important to identify what data is required and where it currently resides. In this instance it was determined that all of the data that is available was needed for the data warehouse in order to better understand and mine the production process. Table 4.1 shows the data systems. Chapter 4 reviewed where the actual data for this case study was acquired from and it explains each stage of the process. Chapter 7 gives what is believed to be the Ideal Data Warehouse for spinning data.
4. Is Data mining useful for a textile company to stay competitive (i.e. will it make a company better and more efficient than other textile companies)? This study really brought to the forefront that although there is a lot of data available in a plant there really is not an easy way to access it. It was determined that if a data warehouse was created and there was more data points the plant could use data mining on a regular basis to help make the process more effective. At this time the data that is available to examine the past is minimal and not detailed enough for each process to really get a good look at the entire process as one entity. For example the timing of the data collection was an issue in that there was no consistency of each process collection time. Some of the data was collected twice a day, some once a week, and others only once a day. Since the data was on different systems it was very labor intensive to enter the data into one system for it to be mined. The plant that was examined in this case study felt that their data was fairly complete. This is a common misconception when data mining is applied.

5. What role could data mining play in the textile industry? It could play an important role in being able to analyze the plethora of data over longer periods of time because the data will be available in one database. Currently most plants still have printouts of the data to see how a plant is running and from this it is difficult to compare past production runs to what is in the plant now. The data mining could offer more detailed analysis if the data is collected and stored properly.
6. How much, if any, is it being used today in this industry? As discussed in the literature review data mining is hardly being used in the textile industry. There were only a few examples and most of these were in the apparel and consumer end of the business. For instance it is being used for forecasting and marketing to the customer [6, 24]. The other cases that were found were in the fiber production process of a man made fiber [29] and the classification of wool fibers using neural networks [40].
7. How would a textile company get started in incorporating this into their business? The data mining process model that was proposed in this research would be a great starting point for any textile company to introduce data mining to the company.
8. What is the process of data mining? The data mining process model again shows the steps that are involved in this process.

This study showed what is really involved in mining data. From this experience a new proposed data mining process model was developed. The model contains 6 major

steps with 28 sub steps. This model was explained in Chapter 7 and the model itself can be seen in Figure 7.3.

8.2 Future Studies

There is very little information about data mining and its use in a manufacturing environment in today's literature. Future research should address this informational need. This section presents several recommendations for future work based on the present research contributions. Some of the ideas are merely extensions to the study that was already done while others may be more in depth. Hopefully, these ideas will spark someone to continue studying data mining and its use in textiles.

It would also be interesting to expand the research to include a monthly, yearly or crop change comparison worth of data. It would also be worthwhile to get permission from a mill to spend a month in the plant collecting the data using the process model in chapter 7. From this research one could have a better knowledge about what really needs to be collected and how to go about doing this effectively. It was difficult in this research to accomplish this because there was no process to follow so there was no way to know that it wasn't being done effectively to get great results. Also because one is at the mercy of what the plant has available for data it was difficult going into the process knowing whether or not it would prove to be useful results from the data. In the end it was determined that more data points that could be linked or as referred to in this dissertation as lagged across the process would have giving better results. However, from the start there was no way to be sure since in past research individuals have found that summarizing results was all right to do because a plant runs roughly the same types of product through the process.

In the future it would be intriguing to do this study again with more data (at least one years worth) and also to get in the plant and be a part of collecting the data. If one is allowed to be in the plant monitoring the collection of data it will be easier to understand where the data has been derived from. Also they will be more likely to get permission if the person doing the research is collecting the data.

It would also be interesting to take the process model and attempt data mining in other areas of textiles. For instance, in this case we looked at the process of cotton spinning but what about other types of fibers like man-made, silk or wool. It would also be interesting to see how it would work for other processes in the supply chain like the dyeing and finishing plants, weaving, knitting or apparel manufacturing.

9.0 References

[1] Alexander, S. "Users find tangible rewards digging into data mines." InfoWorld, July 7, 1997.

[2] Anonymous 1. "RTMS and NeoVista extend opportunity for retailers to reach customers partnership combines individualized marketing with predictive modeling." Business Wire, New York, June 7, 1999.

[3] Anonymous 2. "Data mining and the art of coaching… How IBM advanced scout put the spotlight on NBA sixth man award winner Darrell Armstrong." Business Wire, New York, May 14, 1999.

[4] Anahory, S. "Effective data mining." Hospitality, London, May 1999, p27.

[5] Berry, M. "Data Mining Techniques." John Wiley and Sons, Inc, N.Y., New York, 1997.

[6] Bonner, S. "Forecasting Fashion." Apparel Industry Magazine, March 1996, p 32-34.

[7] Bransten, L. "Technology – Power tools – Looking for patterns: Data mining enables companies to better manage the reams of statistics they collect; the goal: spot the unexpected." Wall Street Journal – eastern edition, New York, June 21, 1999.

[8] Cabena, P. Hadjinian, P. Stadler, R. Verhees, J. Zanasi, A. "Discovering Data Mining: From Concept to Implementation." Prentice Hall, Upper Saddle River, New Jersey, 1998.

[9] Cser, L.; Korhonen, A.S.; Gulyas, J.; Mantyla, P.; Simula, O.; Reiss, G.; Ruha, P. "Data Mining and State Monitoring in Hot Rolling." Proceedings of the Second International Conference on Intelligent Processing and Manufacturing of Materials IPMM '99, 1999, Vol. 1, p 529-536.

[10] Davis, B. "Data Mining Transformed." InformationWeek, 09/06/99, Issue 751, p86.

[11] DeJesus, E. "Turn Computers loose on your data, and you don't know what they'll come up with – that's the whole point." Byte, Oct 99.

[12] Doig, G. Doherty, N. and Marples, C. "An Exploratory Investigation Into the Impact of Information Quality Upon the Perceived Value of Information". Proceedings of the Sixth International Conference on Information Quality, November 2001, p138-152.

[13] DuMouchel, W. "Bayesian Data Mining in Large Frequency Tables, With an Application to the FDA Spontaneous...", American Statistician, Aug 99, Vol. 53 Issue 3, p 177.

[14] Fayyad, U. "The digital physics of data mining." Communications of the Association for Computing Machinery, New York, NY, March 2001.

[15] Flanagan, T. "Mining for a Competitive Advantage in Your Data Warehouse." ATG's Data Warehousing Technology Guide Series, Applied Technology Group, Natick, Massachusetts, 1997.

[16] Groth, R. "Data Mining: A hands on approach for business professionals." Prentice Hall, Upper Saddle River, New Jersey, 1998.

[17] Hall Jr., O. "Mining the store." The Journal of Business Strategy, Boston, Massachusetts, Mar/Apr 2001.

[18] Han, J. Lakshmanan, L. and Ng, R. "Constraint-Based, Multidimensional Data Mining." IEEE, August 99, p 2.

[19] Hand, D.J. "Data Mining: Statistics and More?" The American Statistician, 1998, 52, p 112-118.

[20] Hill, S. "I see a tall, dark man in Wal-Mart looking for wide leg jeans….." Apparel Industry Magazine, August 1998, p 78-82.

[21] Jacobs, P. "Data mining: What general managers need to know." Harvard Management Update, Oct 99, Vol. 4, Issue 10, p 8.

[22] Johnstone, K. "Components and applications for data warehousing." Imaging service bureau news. Westport, May/June 1999, p 10-11.

[23] Kahn, B. Katz-Haas, R, and Strong, D. "Organizational Realism Meets Information Quality Idealism: The Challenges of Keeping an Information Quality Initiative Going." Proceedings of the Sixth International Conference on Information Quality, November 2001, p. 20-32.

[24] Kilarski, D. "Data Mining: Improving on Intuition." Consumer Goods, May/June 1998, Vol. 7, No. 3. p 28-29.

[25] Kittler, R. Wang, W. "The Emerging Role for Data Mining." Solid State Technology, Nov 99, Vol. 42, Issue 11, p 45.

[26] Lach, J. "Data Mining Digs In." American Demographics, July 99, Vol. 21, Issue 7, p 38.

[27] LaMonica, M. "Know Your Customers." InfoWorld, July 7, 1997.

[28] Maciag, G. "Who's mining the agency?" National Underwriter, Erlanger, January 15, 2001.

[29] Mastrangelo, C.M. Porter J.M. "Data Mining in a Chemical Process Application." IEEE International Conference on Systems, Man and Cybernetics, 1998, Vol. 3, p 2917-2921.

[30] Menezes, J. "Database makers deepen mining capabilities." Computing Canada, 07/30/99, Vol. 25, Issue 29, p19.

[31] Mieno, F. Sato, T. Shibuya, Y. Odagiri, K. Tsuda, H. Take, R.. "Yield Improvement Using Data Mining System." IEEE International Symposium on Semiconductor Manufacturing Conference Proceedings, 1999, p 391-394.

[32] Regalado, A. "Mining the Genome." Technology Review, Sep/Oct 99, p 56.

[33] Restivo, K. "The drill on data mining." Computer Dealer News, Willowdale. April 9, 1999, Vol. 15, Issue 14, p 29-30.

[34] Rogoski, R. "SAS partners on workgroup software." Triangle Business Journal, 10/22/99, Vol. 15, Issue 7, p 31.

[35] Rosenthal, A. Wood, D. Hughes, E. and Prochnow, M. "Data Quality in the Small: Providing Consumer Information", Proceedings of the Sixth International Conference on Information Quality, November 2001, p 153-161.

[36] Rubenking, N. "Hidden Messages." PC Magazine, May 22, 2001, Vol. 20, No. 10, p 86-88.

[37] Russo, A. Vanecko, J. "Taking the Mystery Out of Mining and Modeling." Direct Marketing, Aug 99, Vol. 62, Issue 4, p 42.

[38] Ryan, J. "A Practical Guide to Achieving Enterprise Data Quality." ATG's Data Warehousing Technology Guide Series, Applied Technology Group, Natick, Massachusetts, 1997.

[39] Secker, M. "Do you understand your customer?" Telecommunications, Dedham, Massachusetts, March 2001.

[40] She, F.H. Kong, L.X. Nahavandi, S. and Kousani, A.Z. "Woolnet: A Hybrid Artificial Neural Network Wool Fibre Identifier." 2001 Textile Institute Conference Proceedings, 2001.

[41] Thuraisingham, B. "A Primer for Understanding and Applying Data Mining." IT Pro, Jan/Feb. 2000, p 28-31.

[42] Vanecko, J. Russo, A. "Taking the Mystery Out Of Mining and Modeling." Direct Marketing, Sept 99, Vol. 62 Issue 5, p 52.

[43] Weir, B. "Mining the Human Genome." Data Mining Technology Conference Proceedings, October 2000.

[44] Whiting, R. "Data Mining for the Masses." InformationWeek, 09/20/99, Issue 753, p 26.

[45] "Exclusive Research: Data mining." Chain Store Age, Oct 99, Part 2 of 2, Vol. 75, Issue 10, p 42.

[46] "Data mining in brief." Solid State Technology, Nov 99, Vol. 42, Issue 11, p 48.

[47] http://www.sas.com SAS website

[48] http://www.cs.bham.as.uk Introduction to Data Mining

[49] http://www.ibm.com IBM website

[50] http://www.oracle.com Oracle website

[51] http://www.angoss.com Angoss International website

[52] http://www.ncr.com NCR Corporation website

[53] http://www-pub.cise.ufl.edu/~ddd/cap6635/Fall-97/Short-paper/10.htm "Data Mining." Visited 4/4/01.

[54] http://www.math.mcmaster.ca/peter/sora/case_studies_00/data_mining.html "Data Mining – Case Study." Visited 4/4/01.

[55] http://www.anderson.ucla.edu "Data Mining: What is Data Mining?." Visited 4/3/01.

[56] http://wwwmaths.anu.edu.au/~johnm/dm/dmpaper.html "Data Mining from a Statistical Perspective." Visited 4/4/01.

[57] http://www.twocrows.com/glossary.html "Two Crows Data Mining Glossary." Visited 4/3/01.

[58] http://www.rpi.edu/~arunmk/dm1.html "Data Mining." Visited 4/4/01.

[59] http://www3.shore.net/~kht/text/wp9501/wp9501.shtm "An Overview of Data Mining at Dun and Bradstreet." Visited 4/3/01.

[60] http://www.twocrows.com/whitep.htm "Scalable Data Mining." Small, R. and Edelstein, H. Visited 4/4/01.

[61] http://www.consumergoods.com "Kraft Data Mining Transforms Marketing and Margins." Consumer Goods Technology, September 2000, Vol 9, No. 7.

[62] M2001 Conference Proceeding and Lectures. October 1-5, 2001, Cary, North Carolina.

Appendices

Appendix A: Cotton Classification

Quality Parameter	Columns	**Description**
Gin Code Number	1-5	Consists of five digits, the first two digits are the classing office and the last three identify the gin.
Gin Bale Number	6-12	A seven digit bale number is assigned by the gin. At this time a barcode identification tag is placed with the sample.
Date Classed	13-18	This is the data that the bale is classed at the classing office.
Module, Trailer, or Single Bale	19	This one digit code identifies the sample as a single bale = 0, module = 1 or a trailer = 2.
Module/Trailer Number	20-24	A five digit number to identify the module/trailor number assigned by the gin.
Bales in Module/Trailer	25-26	A two digit number that tells the number of bales averaged in the module/trailor to determine the value.
Producer Account	27-29	This space is reserved for the USDA.
Color Grade	30-31	A grade is given by the classer to identify the color of the cotton. (yellowness and whiteness)
Fiber Length (32nd)	32-33	The HVI systems measures the length of the fiber in 100ths of an inch.
Mike (Micronaire)	34-35	Fiber fineness is measured using an airflow instrument in the HVI system.
Strength	37-40	The fiber strength is tested and measured in grams/Tex of the force required to break the fiber.
Leaf Grade	41	This is the amount of plant leaf left in the cotton after ginning and is determined on a 1 (best) -7 basis by the classes.

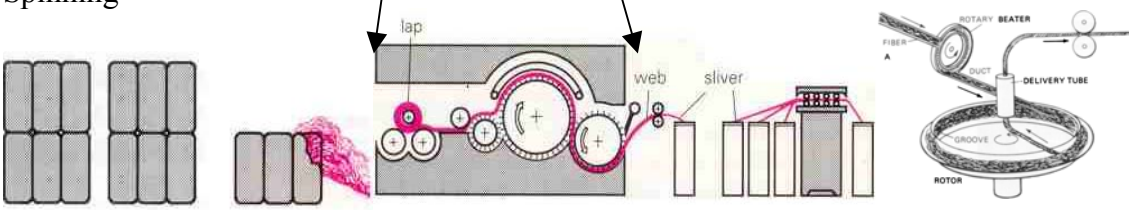
Quality Parameter	Columns	Description
Extraneous Matter	42-43	This is any debris left in the cotton (bark, grass, oil, dust, etc.) these are determined and noted by the classer.
Remarks	44-45	These columns are left for the classer to identify anything special about the cotton.
HVI Color Code	47-48	Two digits to report the color grade.
Color Quadrant	49	This determines the color differences within a color grade for more precision.
HVI Rd	50-51	Indicates how light or dark the sample is.
HVI + b	52-54	Indicates how yellow a sample is.
HVI Trash Percent Surface	56-57	A video scanner helps to determine the trash particles on the surface of the sample.
Fiber Length (100th)	58-62	Gives the length of the fiber in 100^{th} of lengths.
Length Uniformity Percent	63-64	Is a two digit number that identifies the degree of uniformity in the sample.
Upland or Pima	65	Tells whether the sample is an Upland or Pima cotton. (Upland = 1, Pima = 2)
Record Type	66	Tells what kind of record it is. (Original = 0, Review = 1, Rework = 2, Duplicate = 3, and Correction = 4)
CCC Loan Premiums and Discounts	67-71	A five digit code to tell what the CCC loan and discount points are for the Upland cotton.

This table has been derived from the following source
http://ipmwww.ncsu.edu/Production_Guides/Cotton/chptr18.html
By: Keith Edmisten
Crop Science Extension Specialist -- Cotton

Appendix B: Spinning Mill Layout with Lag Times

Online Testing: None	*Online Testing:* None	*Online Testing*: Uster Sliver Data *Tests:* CV Thick Places	*Online Testing:* Schlafhorst data *Tests:* Yarn Breaks Spinning Eff.
Offline Testing: HVI *Tests* Fiber length Fiber Strength	*Offline Testing*: AFIS *Tests:* Neps Trash Short Fiber Readings 3 yd CV	*Offline Testing*: Uster Eveness Tester *Tests:* % CV 1 yd CV	*Offline Testing*: Yarn Qualities *Tests:* Regularity Strength

Bale Laydown Carding Drawframe Open End Spinning



HVI Data
20 Minute delay moving average before next phase

Stays here 1 hour

2 hours here

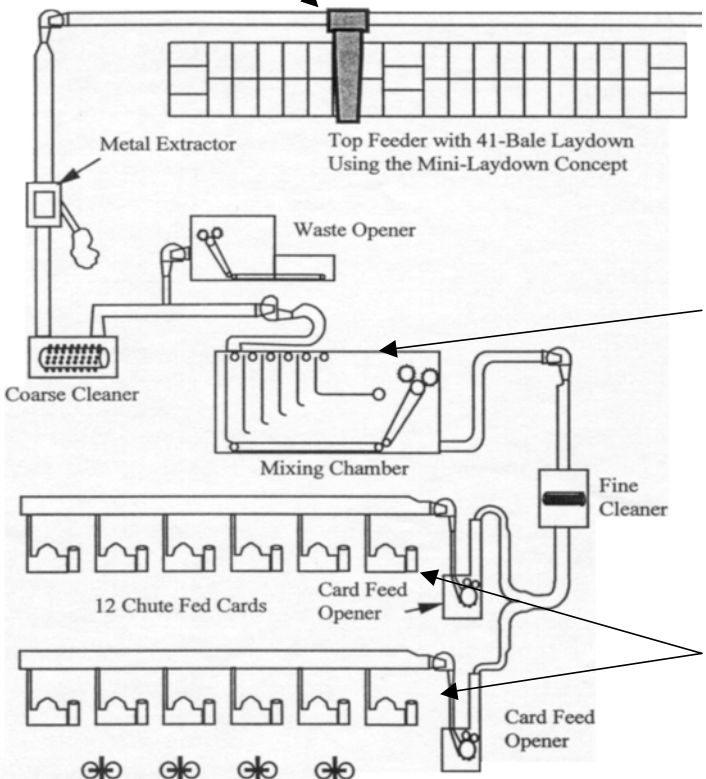
3 day

Appendix C: Another View of Lags

Oct. 17th would be in laydown

Offline HVI Data

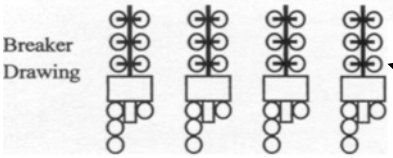
HVI Data is tested prior to bale laydown. Ex. tested On Oct. 16th.



20 Minute delay occurs here

Offline AFIS

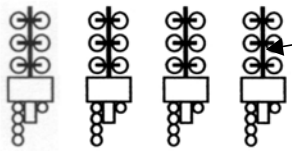
Stays in carding for 1 hour



Online Uster Sliver

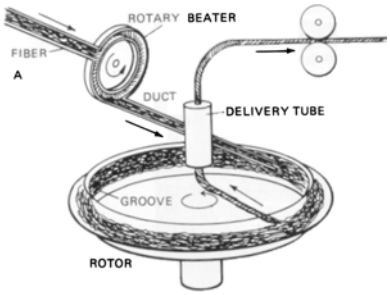
Stays in drawing for 2

Offline Uster Evenness Tester



Online Schlafhorst

Offline Random testing



The can of sliver may remain for approximately 2 Days because of this a 3 day moving average is used. Average of Oct. 17, 18 and 19th.

Appendix D: Spreadsheet of AFIS Data Collected

					SLIVER NEPS				% ELIMINATION					DUST			TRASH			V.F.M.		
					LAST 4 CHECKS				LAST 4 CHECKS													
CARD	MAT	SLIVER	%ELIM		9-6	8-9	7-10	6-12	9-6	8-9	7-10	6-12	CARD	MAT	SLIVER	ELIM	MAT	SLIVER	ELIM	MAT	SLIVER	ELIM
1	469	134	71.43	1	131	147	133	99	70.29	70.78	75.00	81.77	1	215	103	52.09	54	16	70.37	1.15	0.27	76.52
2	501	78	84.43	2	139	140	147	114	65.85	74.50	66.97	79.72	2	242	33	86.36	60	7	88.33	1.30	0.08	93.85
3	498	62	87.55	3	23	49	41	23	94.84	90.26	90.99	95.93	3	261	84	67.82	76	14	81.58	1.70	0.18	89.41
4	417	86	79.38	4	95	121	143	86	75.07	71.93	69.11	83.56	4	268	59	77.99	59	10	83.05	1.35	0.12	91.11
5	403	61	84.86	5	63	99	143	86	86.74	78.80	69.31	83.59	5	223	55	75.34	53	11	79.25	1.52	0.15	90.13
6				6	58	91	77	61	87.76	79.78	84.54	89.97	6									
7	486	91	81.28	7	55	93	59	67	87.72	77.48	88.08	88.22	7	224	52	76.79	57	11	80.70	1.27	0.18	85.83
8	600	88	85.33	8	65	117	111	82	82.62	73.71	77.02	84.29	8	224	54	75.89	61	15	75.41	1.28	0.16	87.50
9	433	56	87.07	9	87	85	99	102	79.38	82.29	78.80	82.17	9	235	45	80.85	60	9	85.00	1.15	0.12	89.57
10	495	80	83.84	10	54	64	97	57	87.14	85.05	77.28	89.66	10	205	91	55.61	45	11	75.56	1.12	0.21	81.25
11	503	113	77.53	11	103	131	148	89	75.48	73.43	69.04	81.98	11	267	37	86.14	66	5	92.42	1.75	0.12	93.14
12	481	101	79.00	12	28	101	148	181	93.99	78.91	71.26	71.41	12	161	31	80.75	41	9	78.05	1.12	0.09	91.96
13	502	104	79.28	13	123	143	182	107	71.53	67.94	62.00	78.77	13	185	41	77.84	53	5	90.57	0.95	0.10	89.47
14	529	25	95.27	14	25	50	55	49	94.58	90.14	87.41	90.67	14	249	27	89.16	66	4	93.94	1.91	0.06	96.86
15	509	41	91.94	15	39	118	45	75	90.93	75.21	89.75	84.21	15	220	33	85.00	60	7	88.33	1.37	0.11	91.97
16	487	38	92.20	16	41	72	105	48	89.80	78.95	75.41	90.93	16	225	40	82.22	45	10	77.78	1.17	0.10	91.45
17	449	77	82.85	17	116	96	105	103	77.03	73.98	77.66	75.88	17	206	52	74.76	41	14	65.85	0.82	0.25	69.51
18	453	59	86.98	18	79	91	99	101	84.92	76.84	78.66	77.25	18	159	65	59.12	37	9	75.68	0.74	0.13	82.43
19	413	90	78.21	19	119	122	123	132	76.25	74.85	71.59	71.05	19	184	50	72.83	35	23	34.29	0.67	0.37	44.78
20	499	69	86.17	20	35	109	116	125	93.62	74.47	77.95	72.35	20	203	61	69.95	49	11	77.55	1.42	0.17	88.03
21	452	97	78.54	21	83	96	93	99	80.47	78.62	80.54	78.38	21	212	95	55.19	47	21	55.32	0.95	0.35	63.16
22	358	115	67.88	22	117	205	132	181	74.00	33.66	70.60	60.31	22	165	43	73.94	43	20	53.49	1.15	0.32	72.17
23	445	31	93.03	23	22	39	218	112	94.94	91.08	48.58	74.94	23	202	54	73.27	55	7	87.27	1.14	0.15	86.84
24	530	134	74.72	24	75	109	117	93	85.27	75.40	69.45	78.91	24	165	41	75.15	45	9	80.00	1.08	0.14	87.04

Appendix D Continued:

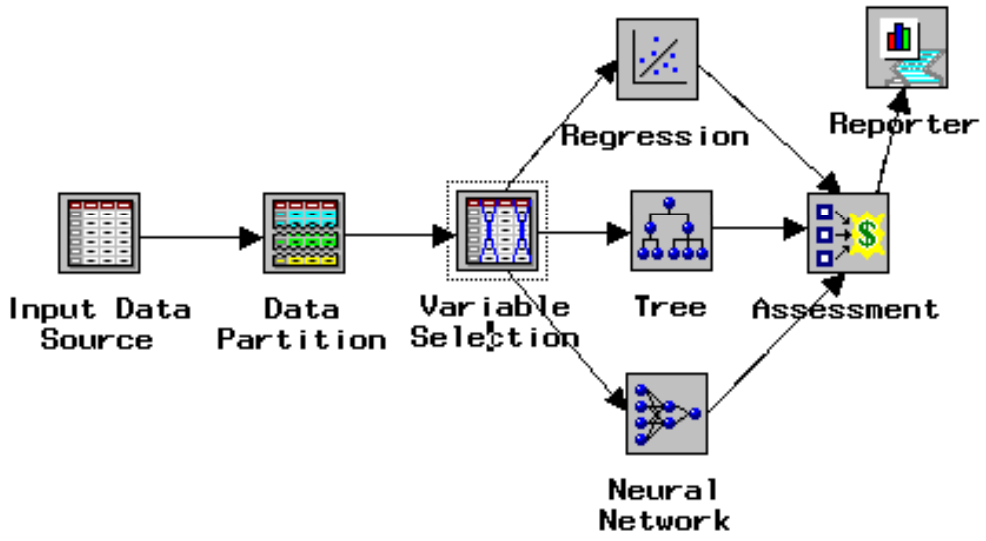
	MAT				SLIVER		
CARD	SFC(n)	SFC(w)	2.5%(n)	CARD	SFC(n)	SFC(w)	2.5%(n)
1	43.7	15.4	1.33	1	38.0	14.2	1.35
2	40.5	13.8	1.34	2	34.4	12.5	1.37
3	39.8	13.2	1.37	3	37.9	14.1	1.39
4	43.3	14.7	1.33	4	33.2	12.0	1.39
5	41.1	13.9	1.35	5	35.7	13.0	1.37
6				6			
7	38.0	12.4	1.36	7	36.8	13.3	1.41
8	36.8	13.3	1.41	8	34.8	12.7	1.38
9	41.5	14.5	1.35	9	37.1	13.0	1.37
10	41.6	13.8	1.35	10	39.8	15.0	1.36
11	39.8	13.4	1.35	11	37.7	13.9	1.37
12	38.8	13.0	1.36	12	36.6	13.2	1.38
13	39.8	13.4	1.34	13	38.4	14.1	1.36
14	39.4	13.4	1.33	14	40.3	15.6	1.36
15	42.2	14.6	1.34	15	35.4	13.3	1.37
16	40.9	13.7	1.35	16	36.2	13.1	1.39
17	35.0	11.7	1.33	17	34.3	12.5	1.35
18	36.0	11.7	1.35	18	36.7	13.5	1.34
19	37.2	12.6	1.33	19	35.2	12.6	1.36
20	36.9	12.3	1.35	20	34.5	12.4	1.36
21	38.0	12.8	1.32	21	34.2	12.0	1.38
22	38.4	13.1	1.35	22	32.5	11.7	1.38
23	34.8	11.1	1.35	23	33.2	12.0	1.39
24	37.7	12.7	1.35	24	34.8	13.0	1.37

Appendix E: Reporter Node Output Example

SAS Enterprise Miner

User : sasjzo
Date : 06FEB2002:18:43:16
Notes:

"EM Workspace" :



STACEY.LAY_DRAW_YARN_SCHLAF

Input Data Settings:

Source Data: STACEY.LAY_DRAW_YARN_SCHLAF (180 rows, 56 columns)
Output: EMDATA.VIEW_YFV
Description: STACEY.LAY_DRAW_YARN_SCHLAF
Role: RAW
Metadata Sample: EMPROJ.SMP_VIQP (180 rows)

- All variables
- Interval Variables

- Class Variables
- Notes: not available

Data Partition

- **Partition Settings**

  Method: SIMPLE RANDOM
  Partition percentages: Training: 80%, Validation: 20%, Test: 0%

- Output
- Log
- Training Code
- Notes: not available

Variable Selection

- Results
- **Settings:**

  Manual selections
  No manual selections were made.

  Target associations
  Selection criterion: R-square
  Squared correlation<0.005
  Stepwise R2 improvement<0.0005
  Ignore 2-way interactions
  Do not bin interval variables (AOV16)
  Use only grouped class variables
  Cutoff:0.5

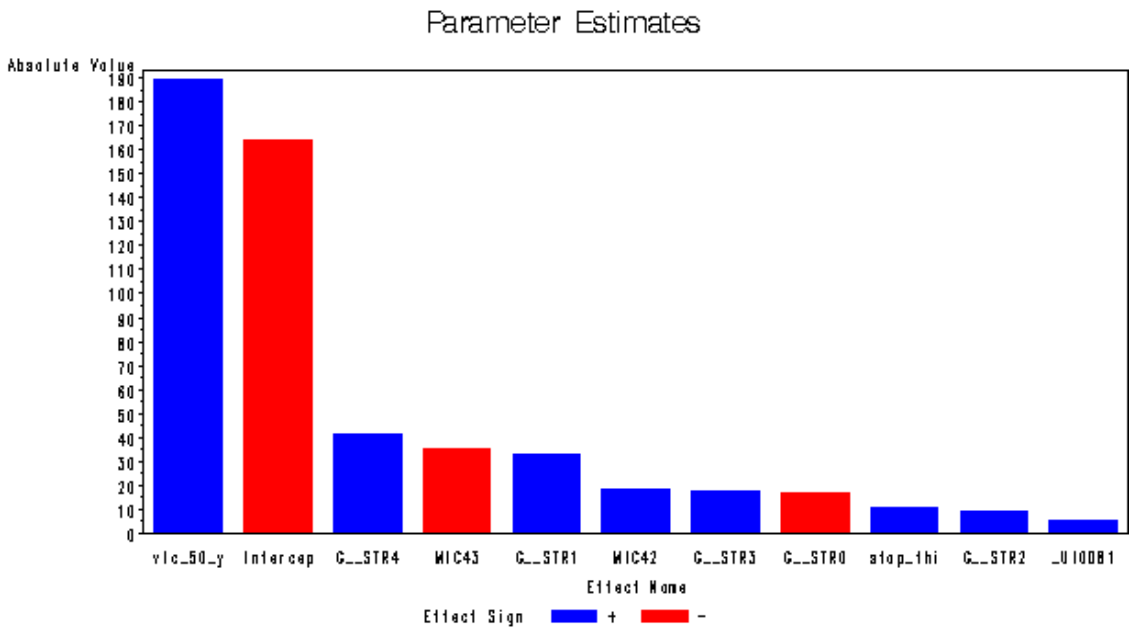
  Missing values
  Reject variables with more than 50% missing values

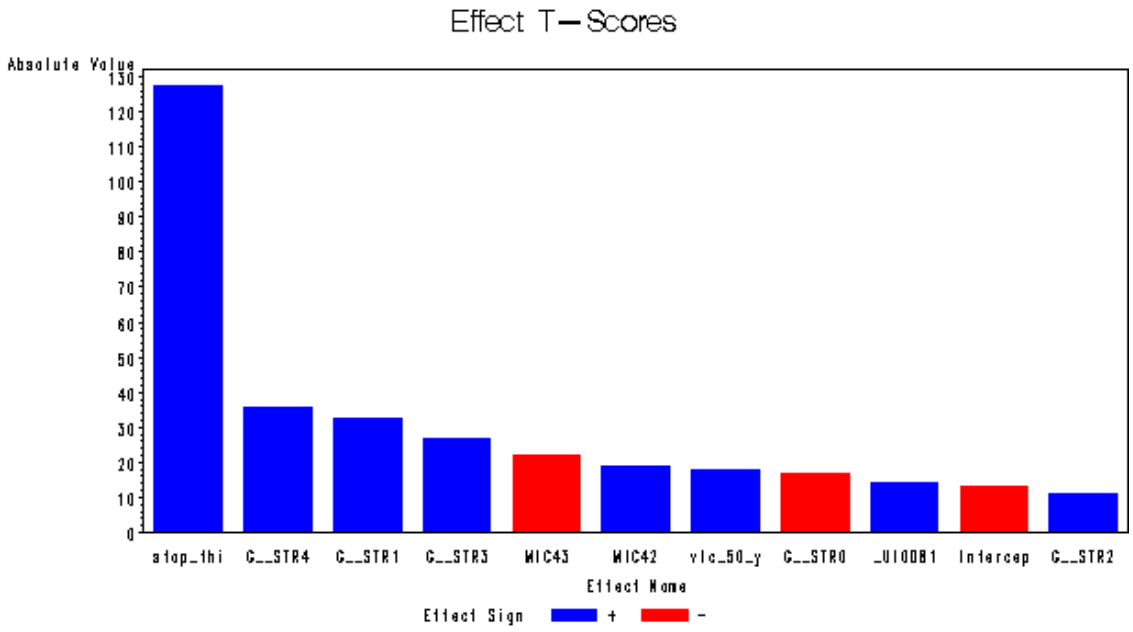
  Hierarchies
  Do not reject variables in hierarchies

- Variables
- Output
- Log
- Training Code
- Score Code
- Notes: not available

Regression

- **Parameters:**
  - Estimates Table





- **Fit Statistics**

Fit Statistic	Training	Validation	Test
Akaike's Information Criterion	296.76228929	.	.
Average Squared Error	6.6471125976	5.8556405602	.
Average Error Function	6.6471125976	5.8556405602	.
Degrees of Freedom for Error	132	.	.
Model Degrees of Freedom	12	.	.
Total Degrees of Freedom	144	.	.
Divisor for ASE	144	36	.
Error Function	957.18421405	210.80306017	.
Final Prediction Error	7.8556785244	.	.
Maximum Absolute Error	4.9231186603	4.9231186603	.
Mean Square Error	7.251395561	5.8556405602	.
Sum of Frequencies	144	36	.
Number of Estimate Weights	12	.	.
Root Average Sum of Squares	2.5781994875	2.4198430859	.
Root Final Prediction Error	2.8027983382	.	.
Root Mean Squared Error	2.6928415403	2.4198430859	.
Schwarz's Bayesian Criterion	332.40004889	.	.
Sum of Squared Errors	957.18421405	210.80306017	.
Sum of Case Weights Times Freq	144	36	.

- Target Information:

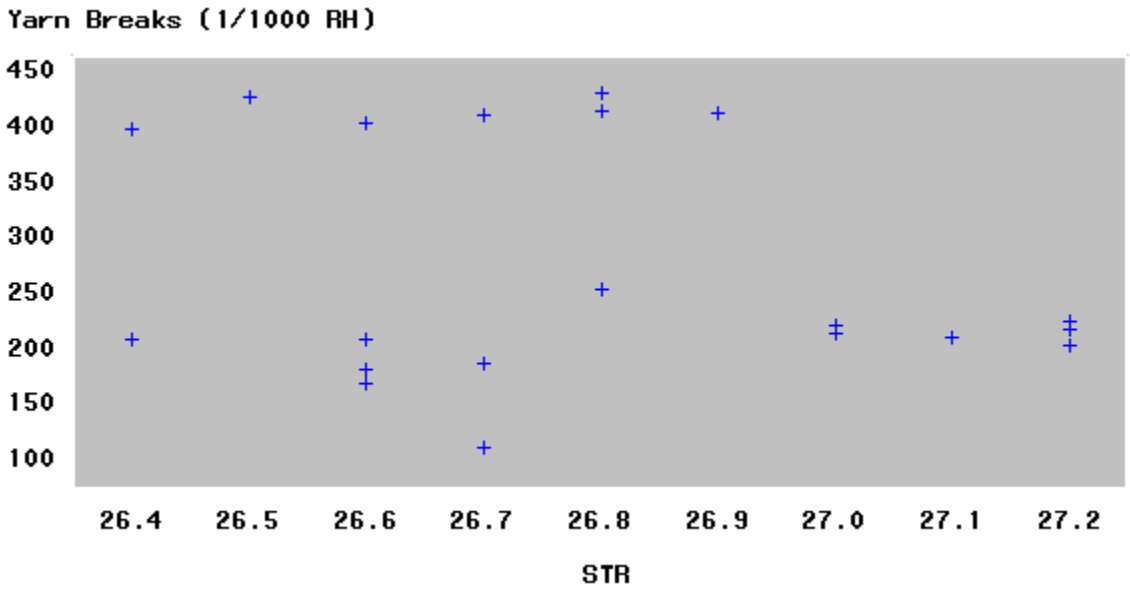
  Name: YARN_BREAKS__1_100_LAG3
  Label: Yarn Breaks (1/1000 RH)
  Measurement: interval

No profile information defined

- **Regression Settings:**

  Regression type: LINEAR
  Selection method: None
  Optimization technique: DEFAULT

- Output



The DMREG Procedure

Data Set: EMDATA.DMDB3XEK
Response Variable: yarn_breaks__1_100_lag3 Yarn Breaks (1/1000 RH)
Number of Observations: 144
Error Distribution: Normal
Link Function: Identity

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr>F
Model	11	1565327	142302	19624.2	<.0001
Error	132	957.184214	7.251396	.	.
Corrected Total	143	1566285	.	.	.

Model Fitting Information

R-square	0.9994	Adj R-sq	0.9993
AIC	296.7623	BIC	300.9276
SBC	332.4000	C(p)	12.0000

Type III Analysis of Effects

Effect	DF	Type III SS	F Value	Pr > F
G__STR	5	58575.1213	1615.554	<.0001

_UI	1	1471.7002	202.9541	<.0001
stop_thick_short_s_l	1	34024.7673	4692.168	<.0001
stop_thick_long_l_la	1	24947.3547	3440.352	<.0001
MIC	2	3574.8584	246.4945	<.0001
vlc_50_yds_lag3	1	2243.5272	309.3925	<.0001

Analysis of Parameter Estimates

Parameter		DF	Estimate	Standard Error	t Value	Pr>|t|
Intercept		1	-164.2	12.4253	-13.21	<.0001
G__STR	0	1	-16.5020	0.9938	-16.60	<.0001
G__STR	1	1	32.7241	1.0116	32.35	<.0001
G__STR	2	1	9.5708	0.8797	10.88	<.0001
G__STR	3	1	17.6225	0.6582	26.77	<.0001
G__STR	4	1	41.5965	1.1700	35.55	<.0001
_UI		1	5.6469	0.3964	14.25	<.0001
stop_thick_short_s_l		1	3.5282	0.0515	68.50	<.0001

The DMREG Procedure

Analysis of Parameter Estimates

Parameter		DF	Estimate	Standard Error	t Value	Pr>|t|
stop_thick_long_l_la		1	7.4398	0.1268	58.65	<.0001
MIC	42	1	18.2401	0.9656	18.89	<.0001
MIC	43	1	-35.0150	1.5906	-22.01	<.0001
vlc_50_yds_lag3		1	189.1	10.7529	17.59	<.0001

- Log
- Training Code
- Score Code

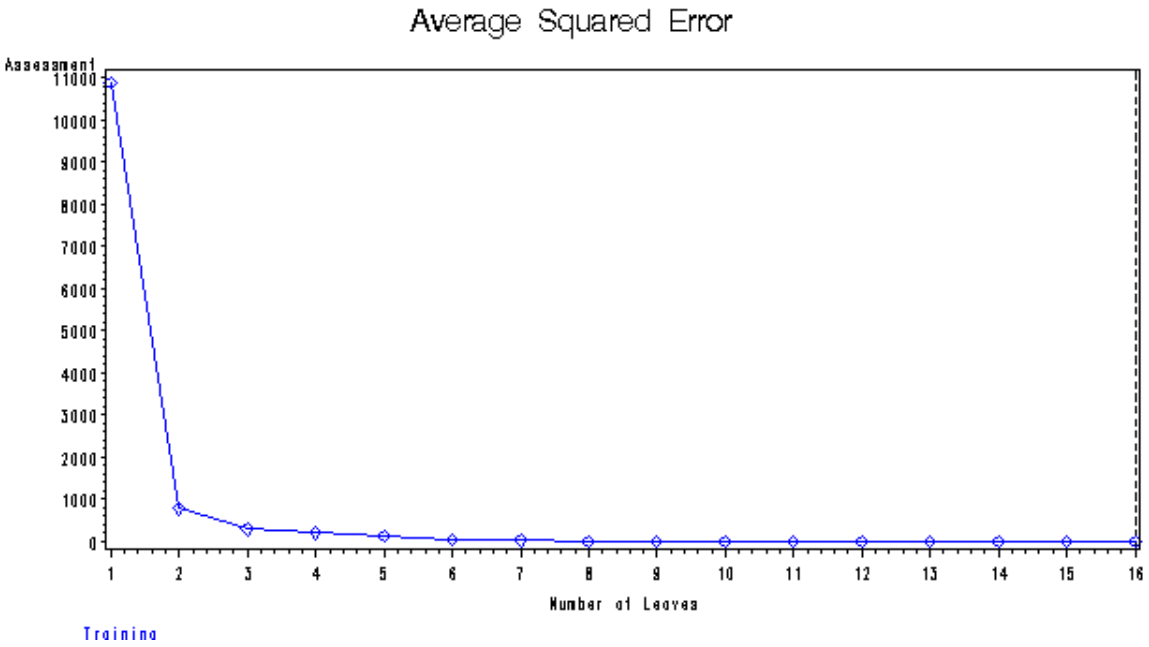
Model assessment settings

Train data set is not selected for assessment.
Validation data set is selected for assessment.
Test data set is not selected for assessment.
Scored data set: 5000 observations are saved for interactive model assessment.

- Notes: not available

Tree

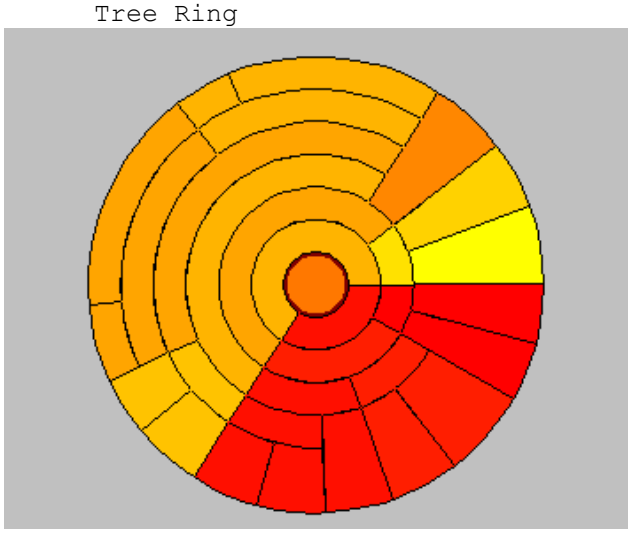
Model assessment plot:



Fit Statistic	Training	Validation	Test
Average Squared Error	1.160	0.9654	.
Sum of Squared Errors	167.077	34.7543	.
Root Average Squared Error	1.077	0.9825	.
Maximum Absolute Error	3.305	3.3049	.
Divisor for ASE	144.000	36.0000	.
Total Degrees of Freedom	144.000	.	.
Number of Estimated Weights	16.000	.	.
Sum of Frequencies	144.000	36.0000	.
Sum Case Weights * Frequencies	144.000	36.0000	.

LEAF ID	N	V N	ROOT ASE	V ROOT ASE	AVERAGE	V AVERAGE
8	8	1	0.00000	0.00000	118.317	118.317
9	7	2	0.00000	0.00000	174.517	174.517
10	8	1	0.00000	0.00000	260.188	260.188
28	22	5	0.40087	0.41595	215.415	215.501
29	6	3	0.00000	0.00000	209.750	209.750
30	23	4	2.66656	2.91072	227.078	229.822
31	8	1	0.00000	0.00000	220.240	220.240
24	7	2	0.00000	0.00000	193.600	193.600
25	6	3	0.00000	0.00000	187.456	187.456
26	7	2	0.00000	0.00000	417.771	417.771
27	7	2	0.00000	0.00000	419.129	419.129
19	7	2	0.00000	0.00000	421.158	421.158
20	7	2	0.00000	0.00000	403.673	403.673
21	9	0	0.00000	.	410.250	.
14	6	3	0.00000	0.00000	433.650	433.650
15	6	3	0.00000	0.00000	437.444	437.444

Tree Ring



- English rules
- Sequence
- Matrix

Target information
Name: YARN_BREAKS__1_100_LAG3
Label: Yarn Breaks (1/1000 RH)
Measurement: interval

Tree settings

No profile information defined
Splitting criterion: F Test
Significance Level: 0.2
Minimum number of observations in a leaf: 1
Observations required for a split search: 2
Maximum number of branches from a node: 2
Maximum depth of tree: 6
Splitting rules saved in each node: 5
Surrogate rules saved in each node: 0
Treat missing as an acceptable value
Model assessment measure: Average Squared Error
Subtree: Best assessment value
Observations sufficient for split search: 144
Maximum tries in an exhaustive split search: 5000
P-value adjustment: KASS DEPTH
Apply KASS BEFORE choosing number of branches

- Log

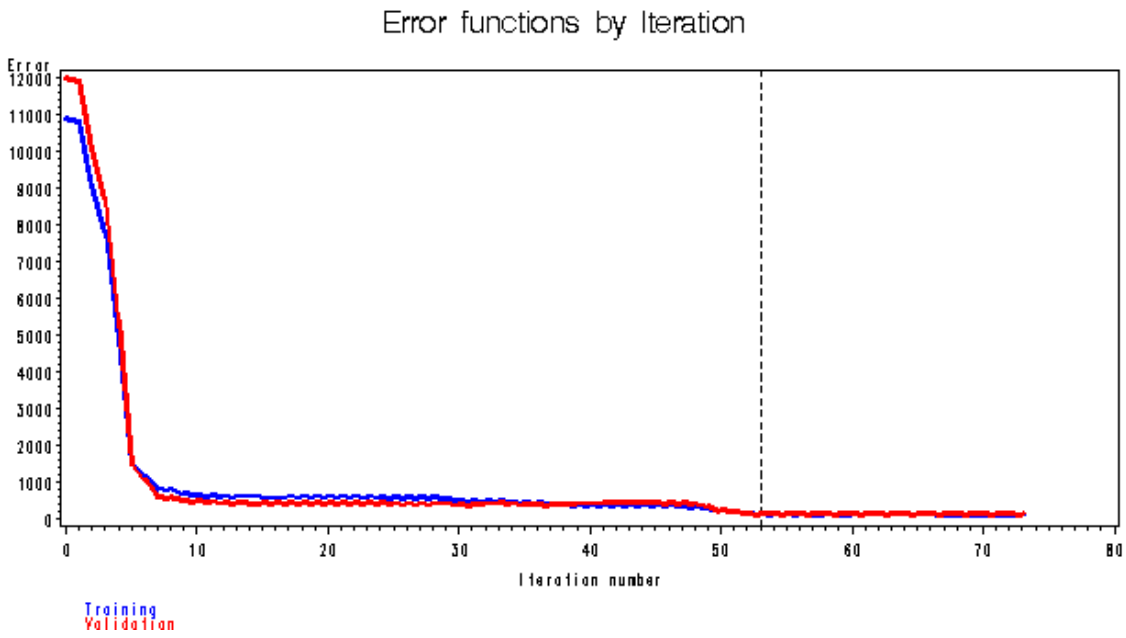
- Score Code

Model assessment settings
Train data set is not selected for assessment.
Validation data set is selected for assessment.
Test data set is not selected for assessment.
Scored data set: 5000 observations are saved for interactive model assessment.

- Notes: not available

Neural Network

Optimization plot:



Fit Statistic	Training	Validation	Test
[TARGET=YARN_BREAKS__1_100_LAG3]	.	.	.
Average Error	126.69	147.42	.
Average Squared Error	126.69	147.42	.
Sum of Squared Errors	18243.41	5306.95	.
Root Average Squared Error	11.26	12.14	.
Root Final Prediction Error	12.41	.	.
Root Mean Squared Error	11.85	12.14	.
Error Function	18243.41	5306.95	.
Mean Squared Error	140.33	147.42	.
Maximum Absolute Error	26.85	26.85	.
Final Prediction Error	153.98	.	.
Divisor for ASE	144.00	36.00	.

Model Degrees of Freedom	14.00	.	.
Degrees of Freedom for Error	130.00	.	.
Total Degrees of Freedom	144.00	.	.
Sum of Frequencies	144.00	36.00	.
Sum Case Weights * Frequencies	144.00	36.00	.
Akaike's Information Criterion	725.21	.	.
Schwarz's Baysian Criterion	766.79	.	.

- Network settings

No profile information defined

- Variables
- Output
- Log
- Training Code
- Score Code

Model assessment settings
Train data set is not selected for assessment.
Validation data set is selected for assessment.
Test data set is not selected for assessment.
Scored data set: 5000 observations are saved for interactive model assessment.

Assessment

End Report