Predicting Customer Satisfaction from Dental Implants Perception Data

PREDICTING CUSTOMER SATISFACTION FROM DENTAL IMPLANTS PERCEPTION DATA

BY

OMNYA ELMASSAD, B.Sc.

A THESIS

SUBMITTED TO THE DEPARTMENT OF MATHEMATICS AND STATISTICS

AND THE SCHOOL OF GRADUATE STUDIES

OF MCMASTER UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

Master of Science (2013) McMaster University

(Statistics) Hamilton, Ontario, Canada

TITLE: Predicting Customer Satisfaction from Dental Implants Perception Data

AUTHOR: Omnya Elmassad

B.Sc., Statistics and Computer Science

University of Khartoum, Khartoum, Sudan

SUPERVISOR: Dr. Román Viveros-Aguilera

NUMBER OF PAGES: xi, 88

Abstract

In recent years, measuring customer satisfaction has become one of the key concerns of market research studies. One of the basic features of leading companies is their success in fulfilling their customers' demands. For that reason, companies attempt to find out what essential factors dominate their customers' purchasing habits.

Millennium Research Group (MRG) - a global authority on medical technology market intelligence - uses a web-based survey tool to collect information about customers' level of satisfaction. One of their surveys is designed to gather information about the practitioner's level of satisfaction on different brands of dental implants. The Dental Implants dataset obtained from the survey tool has thirty-four attributes, and practitioners were asked to rank or specify their level of satisfaction by assigning a score to each attribute.

The basic question asked by the company was whether the attributes were useful to make customer behavior predictions. The aim of this study is to assess the reliability and accuracy of these measures and to build a model for future predictions, then, determine the attributes that are most influential

in the practitioners' purchasing decisions. Classification and regression trees (CART) and Partial least squares regression (PLSR) are the two statistical approaches used in this study to build a prediction model for the Dental Implants dataset.

The prediction models generated, using both of the techniques, have relatively small prediction powers; which may be perceived as an indication of deficiency in the dataset. However, getting a small prediction power is generally expected in market research studies. The research then attempts to find ways to improve the power of these models to get more accurate results. The model generated by CART analysis tends to have better prediction power and is more suitable for future predictions. Although PLSR provides extremely small prediction power, it helps finding out the most important attributes that influence the practitioners' purchasing decisions. Improvements in prediction are sought by restricting the cases in the data to subsets that show better alignment between predictors and customer purchasing behaviour.

Acknowledgements

Writing this thesis would have not been possible without the help and support of all those incredible individuals around me.

First and foremost, I wish to thank my supervisor Dr. Román Viveros-Aguilera for his constant guidance, patience and limitless support. I am so grateful for the knowledge I gained from him which goes beyond the scope of this research. To him again my gratitude with deep respect.

I would also like to acknowledge Millennium Research Group for providing me with the data necessary for this research. Especial thanks are due to my supervisor and mentor, Michael Walker and the Product Development and Improvement team whose assistance proved to be invaluable; both at professional and personal levels.

I owe deepest gratitude to my friends and relatives for their ongoing help and encouragement; in particular Thuraya Elnaeim, Fayad El-Shaikh and Omnia Osman. As well I am greatly indebted to my parents, my brothers, and sister for their undying love and support throughout this journey.

Thanks and gratitude in the beginning and the end are to Allah for giving

me the strength and support to pursue my studies and reach the finish line.

Contents

List of Figures

Chapter 1

Market Research

The leading companies in the market are the ones which know how to deliver high quality products and services. Many companies that attempt to compete in the marketplace resort to consultants to help them improve their performances. These consultants are specialized in product life-cycle and product improvement methodologies. They help companies in addressing the various quality-related issues to their products and services. In addition to consultants, companies also rely on analysts who study and conduct research on consumers' behavior and consumers' willingness to purchase products and services these companies offer (Hayes, 2008).

Accurate and thorough information about prospective and existing customers, the competitors, and the industry in general help solve marketing challenges most businesses likely face. Market research assesses the overall market by surveying customers' likes and dislikes. It allows the company to

learn customers' preferences and buying habits. This could be achieved by mining sets of quantitative data to uncover patterns and correlations that enable more effective marketing.

The studies market researchers conduct help companies to determine whether they have been able to satisfy their customers or not. They also provide companies with crucial information about the various factors that affect their business. Additionally, these studies help decision makers in formulating plans, taking necessary measures and evaluating the business performance.

1.1 Customer's Satisfaction

Attracting and catching customer's attention is a critical attribute of a successful marketing campaign. Defining the target market through determining the customers the business is helping, and how the product will solve customers' problems will undoubtedly have a positive impact on increasing the company's profits. Markets must put the needs and interests of the customer first. Measuring customer satisfaction is a concept many companies apply and has increasingly become a critical element of business strategy. In a competitive marketplace, where companies compete for customers, there is a need to understand how to retain existing customers and how to better attract new ones (Hayes, 2008).

Customers perceptions and attitudes are used to assess the quality of

products and services the company provides. Gathering this information about customer satisfaction could be done using different tools and methodologies. These tools must accurately measure these perceptions and attitudes (Hayes, 2008).

There are many tools and methods companies might use to measure customer satisfaction. The direct methods are based on personally contacting the customers to get their feedback. This could be achieved in many ways. Companies could get customer feedback through third party agencies, in-house call centers or through face-to-face conversation or meeting. Another approach is to distribute surveys and questionnaires among customers to collect information (Hayes, 2008).

There are indirect methods that companies sometimes rely on for assessing customer satisfaction such as customer complaints and customer loyalty. Customer complaints are the issues and problems reported by the customer. If the number of complaints the company gets in a specific period of time is high, this implies that customers are not happy. On the other hand, a smaller number of complaints means the company is performing well, and there is a high level of customer satisfaction (Hayes, 2008).

Customer loyalty is also a measure of customer satisfaction. A customer is considered loyal if they repurchase the product or services from a company on a regular basis. These loyal customers are the satisfied ones, and this indirectly measures customer satisfaction (Hayes, 2008).

1.2 The Case Study

The Millennium Research Group (MRG), a Decision Resources Group company located in Toronto is a global authority in medical technology market intelligence that focuses on the medical technology industry. MRG developed and maintained a survey-based tool called Perception Pulse. This tool allows clients to assess the dynamics of customer habits behavior and loyalty in competitive medical technology markets (Millennium Research Group, 2013).

Dental Implants Perception Data (DIPD) is one of the surveys distributed to general dentists, oral and maxillofacial surgeons, periodontists, prosthodontists and other dental specialists. These groups of respondents have been selected from 13 countries in North America, Europe and Asia. The survey has 43 different questions, ranging from basic information about the respondent, the brand he or she uses, and whether a brand or product has met or exceeded his or her expectations to a set of attributes to evaluate the product's specification. The main objective of DIPD dataset is to determine the attributes that are most influential in the practitioners' purchasing decisions (Millennium Research Group, 2013).

1.3 Research Problem and Methodology

Different statistical approaches have been previous used, on data collected through the surveys to measure overall customer satisfaction. Previous research was conducted using the regular multiple linear regression to build a model to predict the response variables for future data points. Others used factor and principal components analysis to assess the relationship among the predictors attributes. Although these methods are good enough for predicting and determining the most crucial factors, they do not reveal enough knowledge about what motivates the customers' purchasing decisions and trends in customer behaviour. Lacking this knowledge might not help the company discover insights about the customers' attitudes and perceptions in purchasing certain types of brands.

The aim of this research is to use more advanced statistical techniques to assess the reliability of the collected dataset and then find the best prediction model that takes into account the correlations between the explanatory variables in DIPD dataset.

Two approaches have been studied and applied on the dataset. These techniques include Classification and Regression Trees (CART) and Partial Least Square Regression (PLSR). Both methods provide models used to predict a response variable given the values of a set of predictors. These prediction models help in determining the attributes with the biggest impact on the customer's purchasing habits.

Note that the objective is not to evaluate the statistical methods as such, as these methods have been assessed in many different ways in the past. Rather, the objective is to apply them to evaluate the predictive power of the DIPD.

1.4 Organization of the Thesis

The thesis is organized in six chapters as follows.

Chapter one is an introductory chapter. It states the problem of the research and the methodology followed. It gives an introduction about market research analysis and a brief background about the company that owns the data.

Chapter two contains more detailed information about the dataset, the industry and the firms it belongs to. It gives descriptions of the response and explanatory variables from marketing prescriptive and the methodology used in collecting the data.

Chapter three has the primarily analysis of the study focusing on individual and pairs of variables. It contains the summary tables and the descriptive analysis of the dataset.

Chapter four is about the first formal approach used to examine the dataset, which is the Classification and the Regression Tress (CART). It starts by giving an introduction about the method and its mathematical model and then applying the approach on the DIPD dataset to generate

prediction models.

Chapter five describes the second formal approach in this study which is the Partial Least Squares Regression and its application to the DIPD dataset to generate prediction models for customer satisfaction.

Chapter Six includes a summary about the prediction models that are generated using the two statistical approaches and the conclusion of the study.

Chapter 2

Dental Implants Perception Data

2.1 Dental Implants

A dental implant is a "plastic or metal anchor that is inserted into a jawbone to provide permanent support for a crown, fixed bridge, or a denture when the bone itself would provide insufficient support" (Mosby, 2009).

Dental implants are surgically implanted in the jawbone while the patient is under anesthesia. The surgery is a time consuming procedure especially when there are many implants to be placed. The dental implant procedure involves two steps. The first step is a surgery to place the dental implant by drilling a hole in the jawbone, and then the implant is placed into that hole. The second is to uncover the implant after a healing period of about three

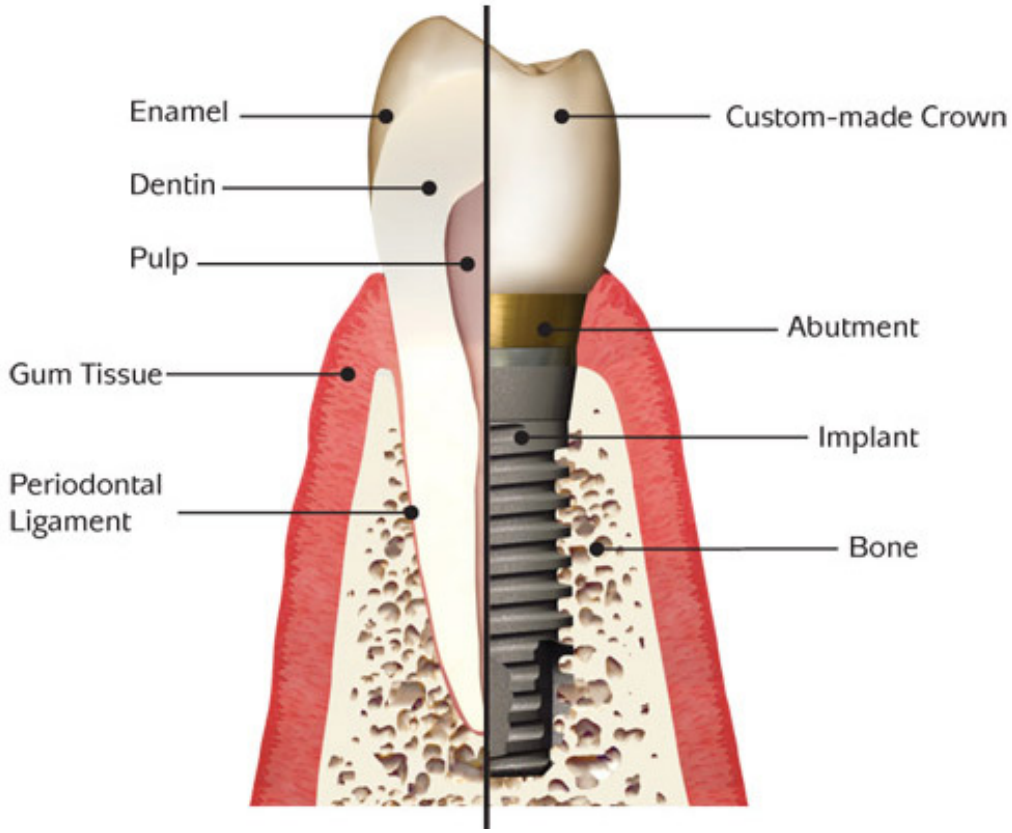

Figure 2.1: The Different Parts of Dental Implants.

to six months to expose the implant and then attach the crown (American Academy of Periodontology, 2012).

According to the American Academy of Periodontology, there are two main types of dental implants:

Endosteal implants: "These are surgically implanted directly into the jawbone. Once the surrounding gum tissue has healed, a second surgery is needed to connect a post to the original implant. Finally, an artificial tooth (or teeth) is attached to the post individually, or grouped on a bridge or denture" (American Academy of Periodontology, 2012).

Subperiosteal implants: "These consist of metal frames that are fitted onto the jawbone just below the gum tissue. As the gums heal, the frame becomes fixed to the jawbone. Posts, which are attached to the frame, protrude through the gums. As with endosteal implants, artificial teeth are then mounted to the posts" (American Academy of Periodontology, 2012).

The MRG survey-based tool collects information about different implant brands used by practitioners. The survey asks the respondents about their decision-making preferences to make a dental implant purchase. This is achieved by asking them to rate a set of attributes related to the company or the brand, the product's quality and services the respondent receives.

The data obtained in DIPD are gathered from different countries including Canada, China, France, Germany, Italy, Japan, Korea, South Netherlands, Spain, Sweden, Switzerland, United Kingdom and United States. The survey targets six main groups of respondents:

General practitioner / Dentist: A primary care provider for patients in all age groups responsible for the diagnosis, treatment, and overall coordination of services to meet oral health needs of patients.

Endodontist: A dentist specializing in root canal issues and the tooth pulp and tissues surrounding the root of a tooth (Mosby, 2009).

Oral and maxillofacial surgeon: A dentist specializing in the diagnosis and surgical, adjunctive treatment of diseases, injuries and defects involving the hard and soft tissues of the mouth, jaws, face, skull and associated structures (Mosby, 2009).

Orthodontist: A dentist specializing in the realignment of dental displacement, and the neuromuscular and skeletal abnormalities of the orofacial structures (Mosby, 2009).

Periodontist: A dentist specializing in diseases of the gums and the other supporting tissues surrounding the teeth (Mosby, 2009).

Prosthodontist: A dentist specializing in the restoration and replacement of missing or deficient teeth to maintain the oral function, comfort, appearance and health of patients (Mosby, 2009).

2.2 Loyalty Matrix

Loyalty Matrix is a set of measurements used to assess customer loyalty in the marketing fields. Marketing companies create them to ensure that customers purchase more frequently and exclusively at their company. These

loyalty measurements have been shown to be a good indicator of company's financial growth. Companies need to make sure they retain and increase their number of customers in order to succeed in their businesses. The objective of Loyalty Matrix is to measure the customers' endorsement of, and approval of a company. This could be achieved by assessing the likelihood that the customer will keep his/her positive behavior towards the company in the future (Hayes, 2008).

The Loyalty Matrix usually has attributes that can be used in assessing characteristics such as:

Satisfaction: Overall impression the product or the company left on a customer.

Advocacy: Likelihood that the customer will be an advocate for the company. This includes measurements such as the possibility to recommend to a friend, choose again and communicate a positive behavior.

Repurchase Intention: Likelihood that the customer will increase their purchase frequency of that product and the possibility of them remaining with the company.

Perceived Value: Ratio of benefits received from companies or products relative to cost. This attribute reflects the customer's evaluation of what is fair for the perceived cost of a product.

2.3 Responses and Explanatory Variables

The DIPD dataset has 5 groups of variables. The first one consists of 12 variables concentrating on *general information* about the respondent and the period of time in which the data were obtained.

The second group is the Loyalty Matrix which includes the *response variables*. In this part, the respondents rate attributes on a scale from 0 to 10, to indicate their level of satisfaction for the experience of using this brand of implants. The attributes the Loyalty Matrix has are *Overall Satisfaction* which relates to the user's experience with a company and the extent to which his/her needs have been met, *Advocacy* which captures the user's willingness to recommend a company to colleagues, *Perceived Value* which relates to the cost paid for that product, and *Repurchase Intention* which expresses the likelihood a user will continue using a company's products in the future.

The third group of variables contains six *company attributes*. The respondents give a rate on a scale from 0 to 100, to indicate their level of satisfaction with the company's performance. The first attribute relates to the company and brand image. The second attribute refers to new technologies, which include rating the continuous stream of innovation and the new technologies the company uses. The third attribute is called the Product Range, which asks about the breadth and variety of product. The fourth

attribute is the Supporting Evidence, and this is used to assess the performance of the company product and the based research evidence in support of this product. The fifth attribute asks to assess the net purchase price of the product. The last attribute is the Limited Backorders which indicates the predictable availability of the product.

The fourth group of variables in DIPD contains another six attributes used to assess the *product quality*. These attributes are rated on a scale from 0 to 100. This group contains Success Rates, Ease of Use and Simplicity, Accelerated Treatment which asks about the ability for accelerated patient treatment, Product Familiarity, Endorsement by Key Opinion Leaders and Restorative Preference.

The fifth and last group contains five *service attributes* also rated on a scale from 0 to 100 to indicate the level of satisfaction with the services that a particular company provides. These attributes include: Quality of Sales Representative, Customer Service and Technical Support, Education and Training Programs, Ongoing Support for Practice Growth and Helpful Website and On-line Capabilities.

2.4 Training and Testing Data

Separating data into training and testing sets is a fundamental part in statistical analysis particularly on evaluating the performance of prediction models. Most of the data is used for training, and a smaller portion of it is

used for testing. The training set is used to build the model. The testing one is used to measure the model's performance in predicting the response variable from a new set of observations. In this analysis, 70% of the data is taken as a training set, the rest is used for testing.

2.5 The Scope and Objectives of the Study

Many companies need to figure out the derived importance of various perception attributes which influence loyalty behavior and purchasing decisions of their customers. One way to get that piece of information is to let purchasers rank these attributes in order of importance, thereby enabling these companies to focus on those attributes that have the most impact on their customers purchasing decisions. MRG develops the DIPD survey tool in order to assist firms in assessing their performances and to better focus their sales, marketing, and product development.

The respondents to this survey were asked to rate each attribute on a scale of (0-100) or (0-10), where 0 represents "thoroughly dissatisfied" and 100/10 represents "thoroughly satisfied". The data set has four response variables and 30 explanatory variables, see Table 1 for details. However, 17 of the explanatory variables are used as the predictors in this study.

The main objectives of this study are:

- Assess the DIPD data regarding its reliability to predict response variables (Customer Satisfaction, Advocacy, Perceived Value and Repurchase Intention) on the basis of the information collected on explanatory variables.
- Identify such prediction models if they exist, by fitting them to training data.
- Test the prediction models on test data.
- Prior to addressing these issues, conduct some preliminary descriptive analysis to help in the understanding of the data and uncover some of the underlying features, and perform data cleaning as needed.

Table 2.1: Description of all the variables in the Dental Implant Perception Data.

Response Variables			
Name	*Type*	*Scale*	*Description*
Satisfaction	Numerical	[0,10]	Respondent's score for this manufacturer on "overall satisfaction"
Advocacy	Numerical	[0,10]	Respondent's score for this manufacturer on "likelihood to recommend" (advocacy)
Perceived.Value	Numerical	[0,10]	Respondent's score for this manufacturer on "perceived value"
Repurchase.Intention	Numerical	[0,10]	Respondent's score for this manufacturer on "repurchase intention"
Predictors			
Name	*Type*	*Scale*	*Description*
Survey.ID	Numerical	(0, inf)	Unique identifier for each survey
Period	Categorical		Year and quarter of the survey
Country	Categorical		Country of respondent
Manufacturer	Categorical		Manufacturer that is being scored in this row/record
Familiarity	Categorical		Familiarity with manufacturer being scored in this row (currently used, formerly used, never used familiar, never used not familiar)
Specialty	Categorical		Respondent's dental specialization
Practice.Type	Categorical		Respondent's dental practice type
Years.In.Specialty	Numerical	[1, inf)	Number of years the respondent has been in practice
Area	Categorical		Geographic region of the US in which the respondent practices
Num.of.implants	Numerical	[0, inf)	Number of dental implants placed by the respondent in the past quarter
Num.of.crowns	Numerical	[0, inf)	Number of crowns placed by the respondent in the past quarter

Num.of.abutments	Numerical	[0, inf)	Number of abutments placed by the respondent in the past quarter
Num.of.bone grafts	Numerical	[0, inf)	Number of bone grafts placed by the respondent in the past quarter
Per.Success.Rate	Numerical	[0, 100]	Success rates the respondent has had with procedures using this manufacturer's products
Company.Image..Brand	Numerical	[0, 100]	Respondent's score for this manufacturer on this attribute
New.Technologies	Numerical	[0, 100]	Respondent's score for this manufacturer on this attribute
Product.Range	Numerical	[0, 100]	Respondent's score for this manufacturer on this attribute
Supporting.Evidence	Numerical	[0, 100]	Respondent's score for this manufacturer on this attribute
Net.Purchase.Price	Numerical	[0, 100]	Respondent's score for this manufacturer on this attribute
Limited.Backorders	Numerical	[0, 100]	Respondent's score for this manufacturer on this attribute
Success.Rates	Numerical	[0, 100]	Respondent's score for this manufacturer on this attribute
Ease.of.Use	Numerical	[0, 100]	Respondent's score for this manufacturer on this attribute
Accelerated.Treatment	Numerical	[0, 100]	Respondent's score for this manufacturer on this attribute

Product.Familiarity	Numerical	[0, 100]	Respondent's score for this manufacturer on this attribute
Restorative.Preference	Numerical	[0, 100]	Respondent's score for this manufacturer on this attribute
Endorsement.by.KOLs	Numerical	[0, 100]	Respondent's score for this manufacturer on this attribute. KOL stands for "Key Opinion Leaders" in the filed
Sales.Reps	Numerical	[0, 100]	Respondent's score for this manufacturer on this attribute
Customer.Service	Numerical	[0, 100]	Respondent's score for this manufacturer on this attribute
Training.Programs	Numerical	[0, 100]	Respondent's score for this manufacturer on this attribute
Practice.Support	Numerical	[0, 100]	Respondent's score for this manufacturer on this attribute
Helpful.Web.site	Numerical	[0, 100]	Respondent's score for this manufacturer on this attribute

Chapter 3

Univariate and Association Descriptions

Raw data is difficult to investigate especially with a large number of cases or observations. Univariate analysis and association assessments are used to summarize and describe the basic features and highlights of the data. This chapter provides a summary about each response and explanatory variable, their distributions and explores their associations.

3.1 Cleaning the Data and Handling of Missing Values

Missing data has always been a problem when it comes to conducting statistical analysis. These missing values might result in degrading the representativeness of the sample. This might lead to distort inferences about the population. In the DIPD dataset, missing values occur when respondents submit the survey without answering all the questions and leave the analyst to deal with no record values. Another problem respondents might incur in, is entering incorrect values while filling up the survey. Such a problem might lead to misinterpretation of the analysis results. The DIPD survey does contain some incomplete and inconsistent data.

One of the first steps in statistical analysis is cleaning the datasets. Data cleaning routines attempt to impute missing values, detect outliers and correct inconsistencies in the data.

As a start, this study performed a validity check across the responses and the predictors in the DIPD data set to ensure that all the entered values are within the correct ranges.

Data collected by the DIPD survey tool covers three years, from 2010 to 2012. The data set was collected from different types of practitioners with some missing values. To avoid any kind of distortion in the analysis and since it is a large dataset, all missing and incorrect values have been excluded. This routine led to a clean dataset that has 11,504 cases. The

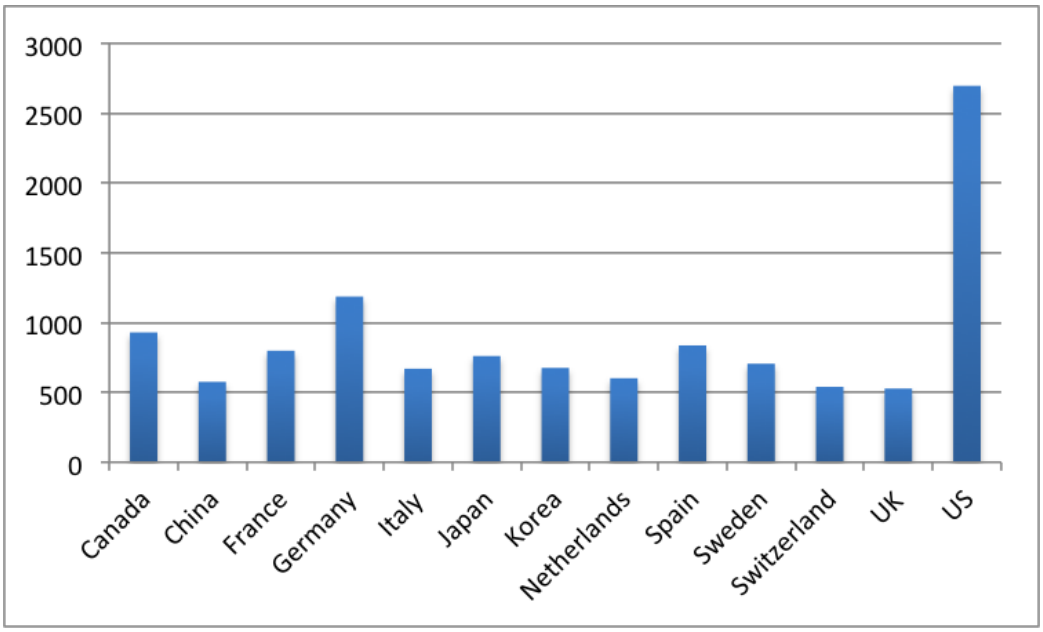

Figure 3.1: Participation Counts per Country.

surveys have been distributed in 13 countries. The main contributions are United States, Germany and Canada with percentages equal to 23%, 10% and 8%, respectively. See Figure 3.1 for a complete account of the participation counts across the countries surveyed.

3.2 Distribution of Responses and Predictors

This section provides quantitative descriptions of all the variables in the DIPD dataset regarding their distributions. Simple summaries and graphs for each response across the different countries are presented in order to give an overall idea of how these variables look like and what they represent.

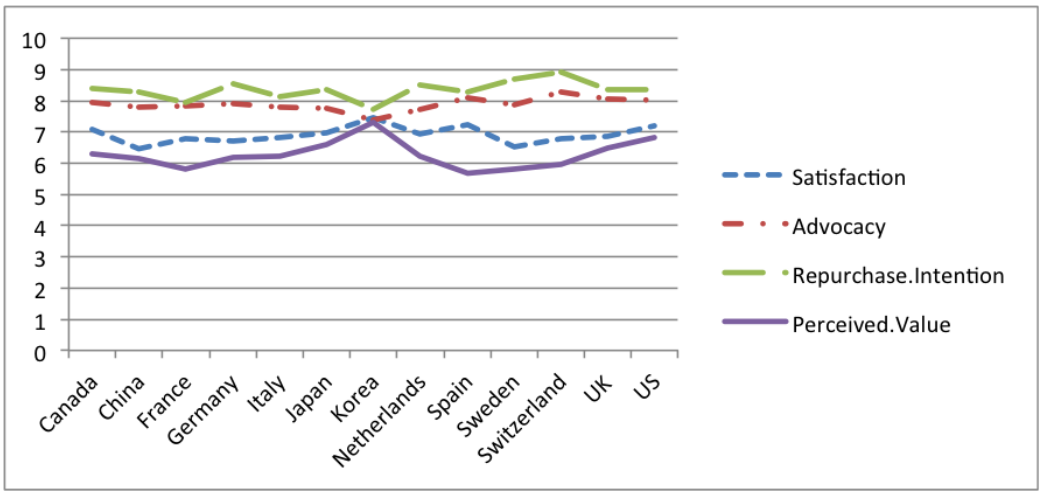

Figure 3.2: Means of the Response Variables per Country.

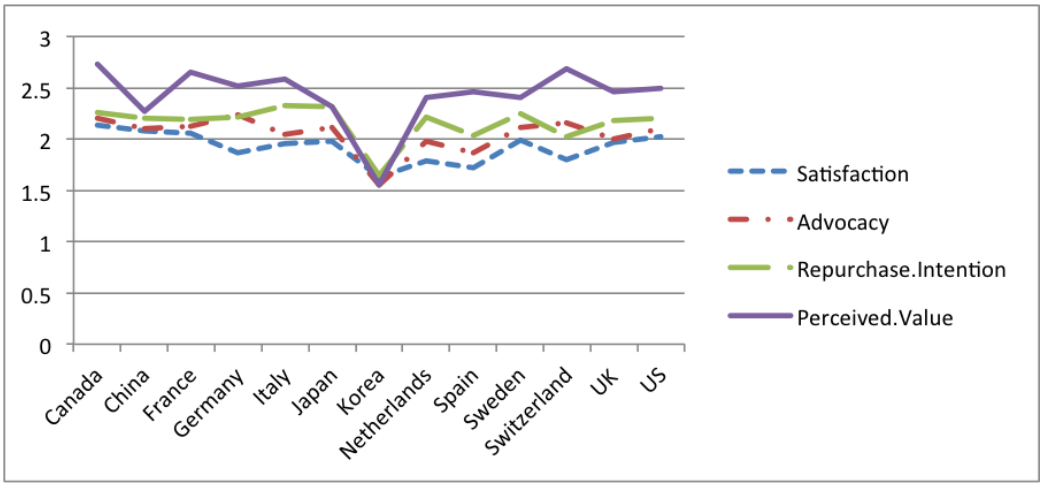

Figure 3.3: Standard Deviations of the Response Variables.

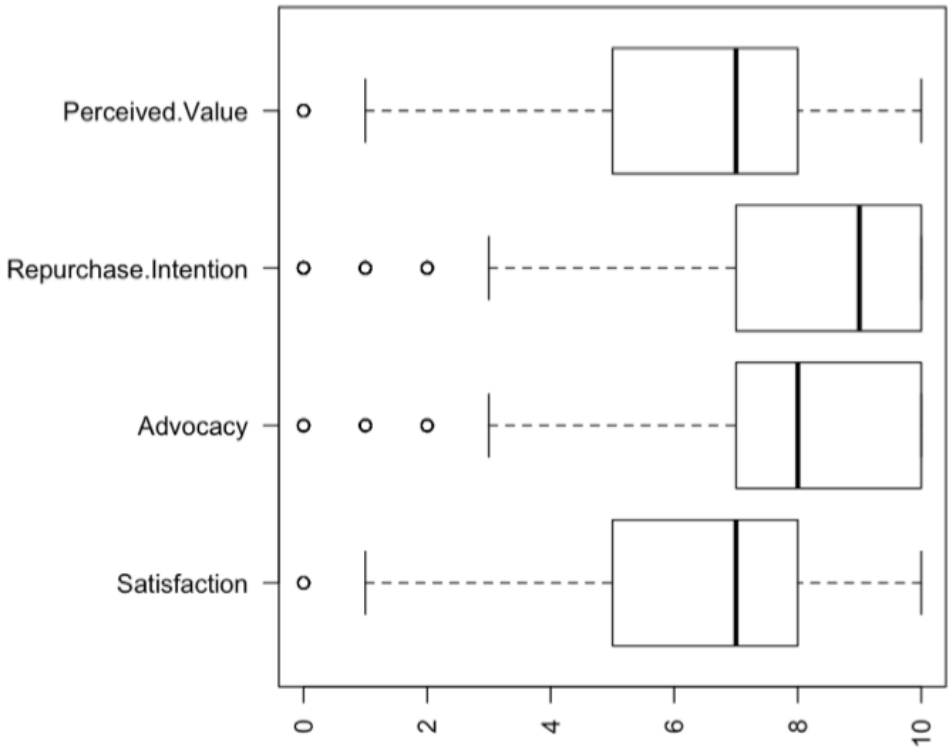

Figure 3.4: Boxplots of the Response Variables.

Means and medians have been calculated to figure out where the data points tend to fall. The standard deviations measure the variation in the response and give an idea of how the data points are spread out around their means. Each of the produced figures has four lines corresponding to the each response variables.

Figure 3.2 shows that all the responses' means are somewhat similar across the countries except for Perceived.Value.

It shows that Korea has the highest mean of Perceived.Value (mean = 7.299) while Spain tends to have the lowest one (mean = 5.677). Both Advocacy and Repurchase.Intention tend to have the highest means. All the mean values fall within 7-9 score points. On the other hand, Perceived.Value has the smallest mean values across all the countries with the means ranging from 5 up to 8.

Figure 3.3 shows the standard deviation (SD) in each response across the different countries. The SD values are relatively small and are similar from country to country except for Perceived.Value. The SD values range between 1.5 and 3 overall. Note that Korea exhibits the smallest SD values, and they are always identical across all the response variables.

Figure 3.4 gives an idea of the distribution of each response variable. All the responses tend to have a negatively skewed distribution which indicates that a large number of respondents gave higher scores when they filled in the survey.

Looking at the means' plots of the explanatory attributes, shown in Figure

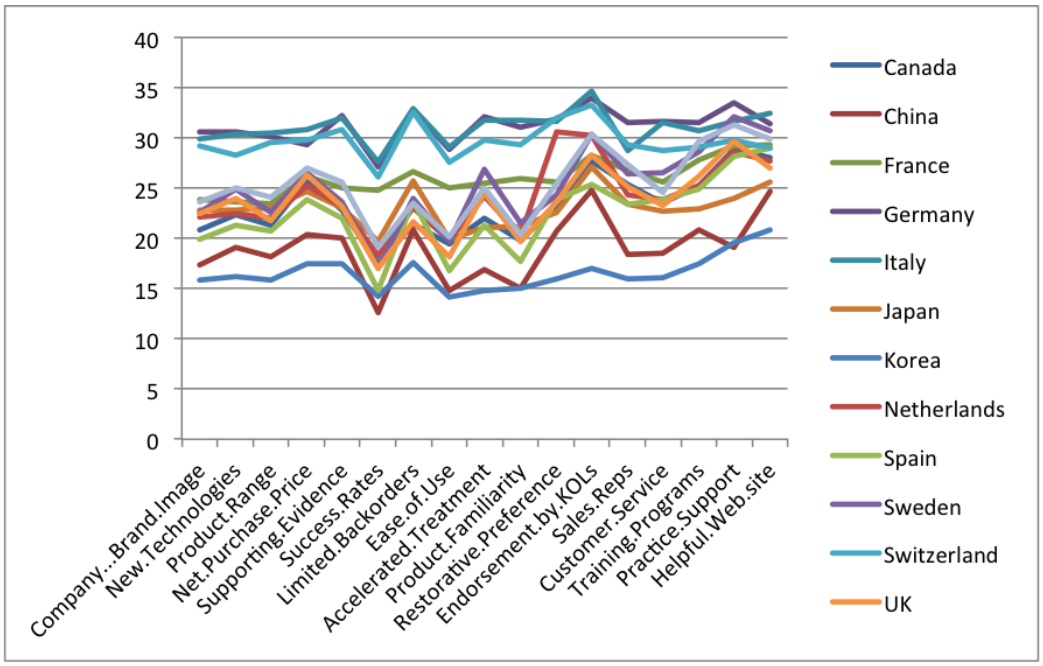

Figure 3.5: Means of the Explanatory Variables.

3.5, all the countries tend to have the same patterns. Net.Purchase.Price and Practice.Support turn out to have the smallest scores compared to other attributes while attributes such as Product.Familiarity and Success.Rate were given the highest ranks. The means range from 40 to 90. Countries such as Italy, Switzerland and Germany tend to have small means, others such as China and Korea have the highest mean values.

A standard deviation plot is used to display the variability in the explanatory attributes. Figure 3.6 shows the shifts in variation, in particular the standard deviations are widely spread out. Although the means plot shows that Italy, Switzerland and Germany have small mean values, it turns

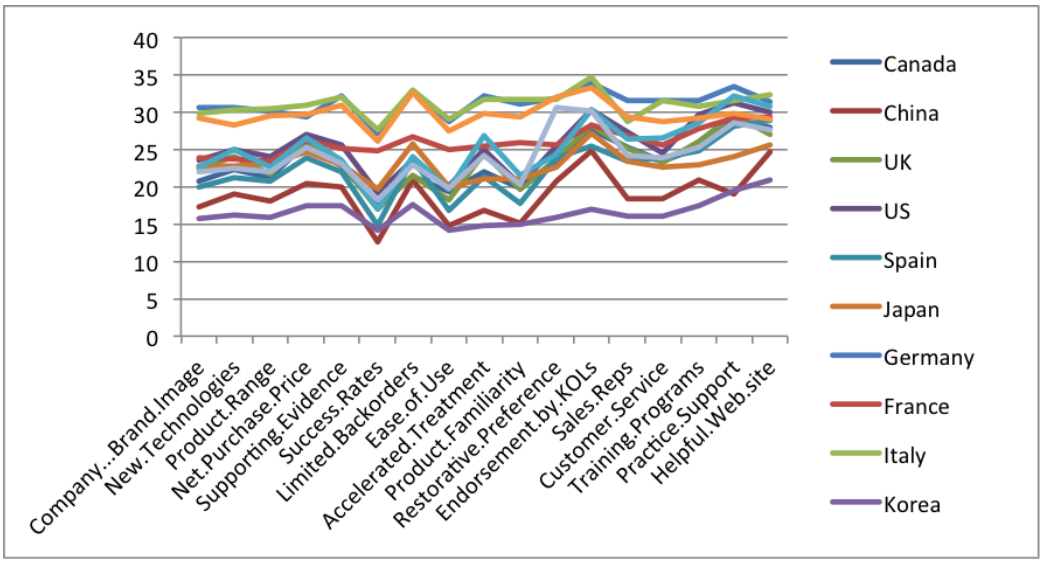

Figure 3.6: Standard Deviations of the Explanatory Variables.

out that these countries have the highest standard deviation values ranging from 24 up to 35. Such a problem might arise when outliers exist in the dataset.

Boxplots provide a graphical summary of key features of a distribution such as the center, the spread of the middle of the data and also help in detecting outliers.

The boxplots shown in Figure 3.7 confirm that most of the predictors have similar asymmetric distributions with a long tail on the left which indicates that the distributions are negatively skewed. Also, the scores given by the respondents tend to cluster toward the upper end of the scale (high scores) while fewer scores occur toward the lower end.

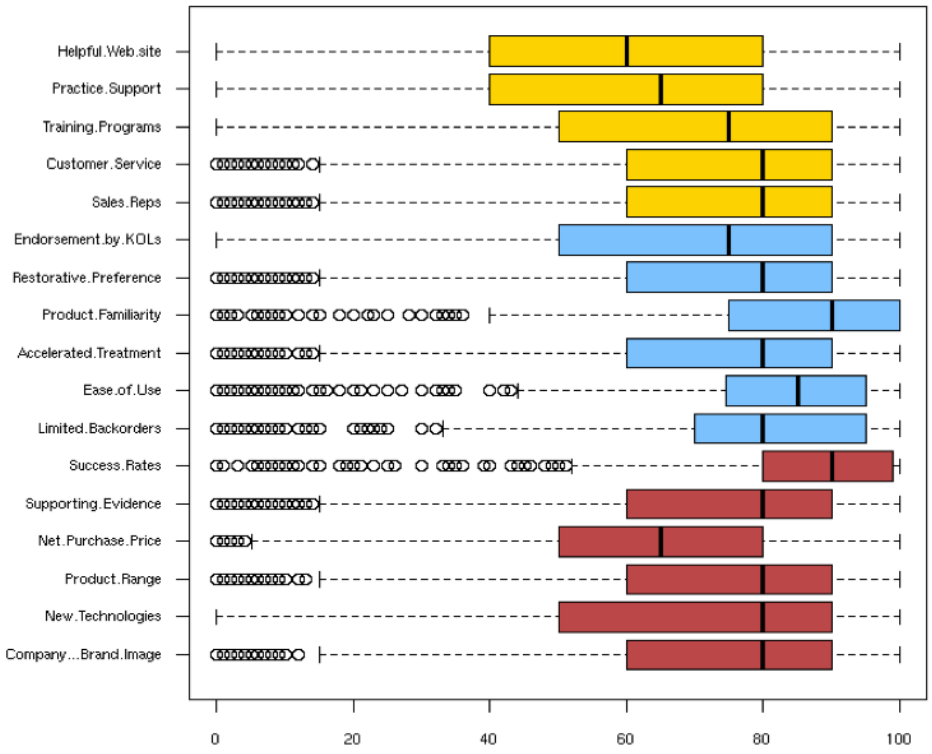

Figure 3.7: Boxplot of the Explanatory Variables.

The boxplots show that Helpful.Web.sites, Practice.Support, Endorsement.by.KOLs and New.technologies are the only predictors that do not have outlying scores. On the other hand, attributes such as Success.Rate, Product.Familiarity and Ease.of.Use have extreme outlying scores.

Another approach to view these attributes is by the group they belong to. Company attributes - red boxes - show similarity in their distributions except for Success.Rate and Net.Purchase.Price attributes. Product quality attributes - blue boxes - is the group with the most attributes with lots of outliers. Services attributes - yellow boxes - show some similarity in their distribution with fewer outliers.

The distributions of all the excluded respondents have been examined as well. It turns out these respondents have similar distributions to the remaining respondents. Thus, we expect that excluding such cases will not affect the results of our analyses. Figure 3.8 and 3.9 display the distributions of these respondents.

3.3 Associations

The correlation coefficients have been calculated in order to measure the degree of association between the predictors and the loyalty matrix attributes. Table 3.1 displays the correlations between the predictors and the response variables. Note that all the correlations are positive. The small values of the coefficients are an indication of a weak relationship between the

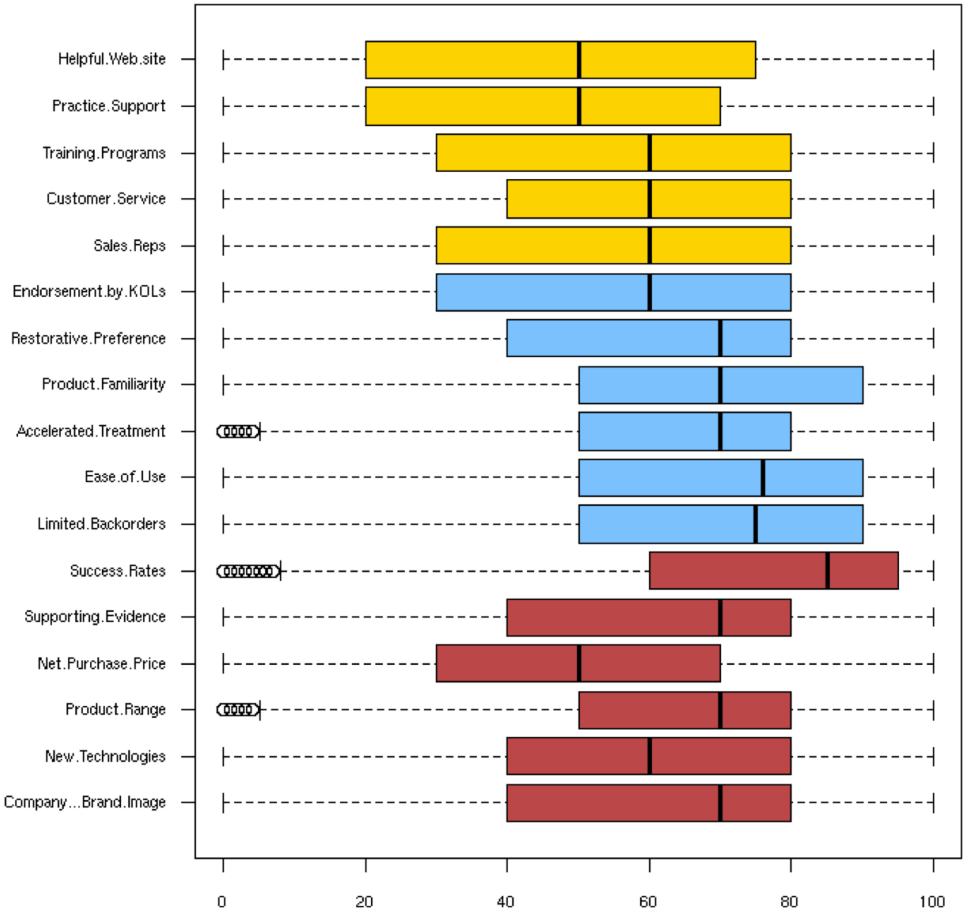

Figure 3.8: Boxplot of the Explanatory Variables for the Cases with Missing Values.

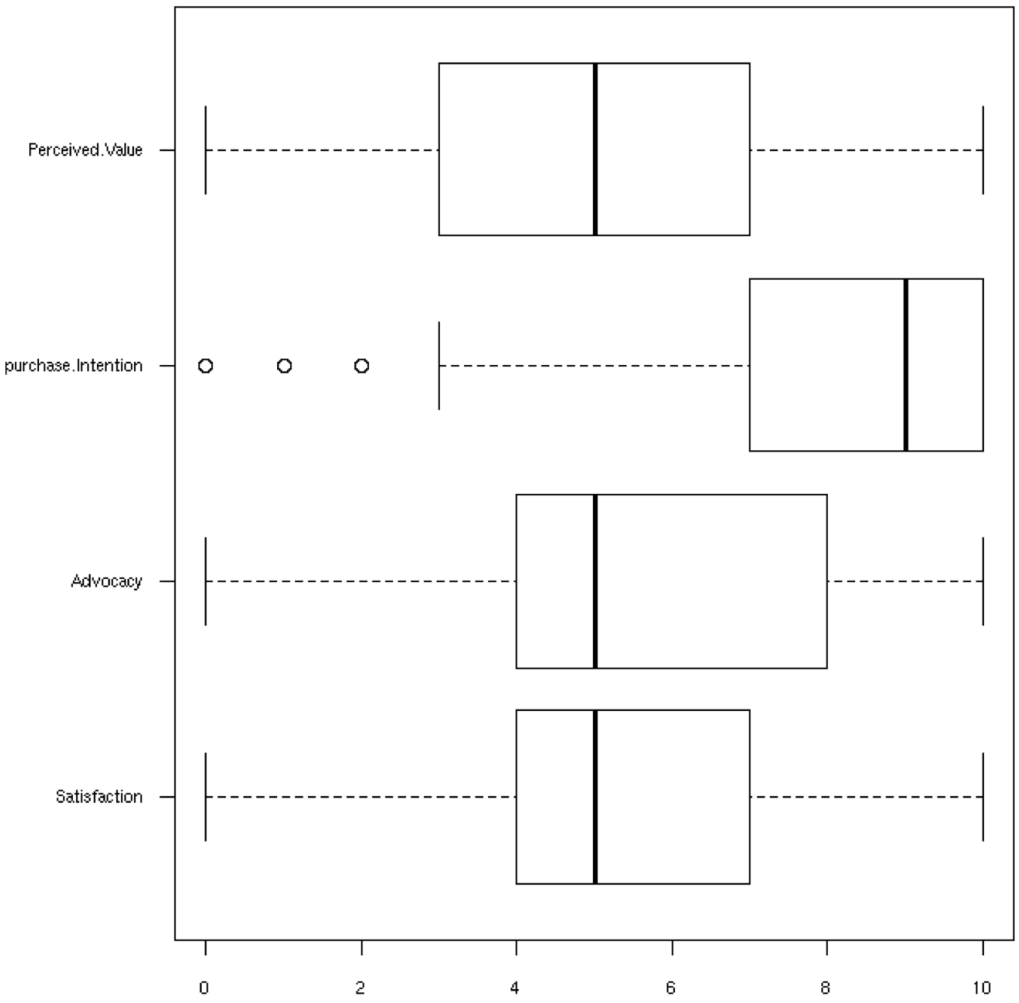

Figure 3.9: Boxplot of the Response Variables for the Cases with Missing Values.

responses and the predictors.

The strongest association is between Perceived.Value and Net.Purchase.Price with a correlation coefficient of 0.4386. Sales.Reps, Company...Brand.Image, New.Technologies and Customer.Service predictors have the next highest correlation coefficients with the response variables (Satisfaction, Advocacy and Repurchase.Intention). However, all these correlations are weak. This is the first indication of a low predictive power for the developed prediction models.

The associations between the response variables themselves have been calculated, and the correlation coefficients are displayed in Table 3.2. The results show that the response variables are positively correlated. The correlations are weak except for Repurchase.Intention and Advocacy whose correlation is moderate (0.6336).

3.4 Test of Multicollinearity

Multicollinearity arises when two or more of the predictors in a regression model are exactly or nearly linearly dependent. This typically causes problems when fitting regression models (Liao and Valliant, 2012).

When the correlation between two independent variables is equal to 1 or -1, it is said that there is perfect multicollinearity among variables and

Table 3.1: Correlation Between Predictors and Response Variables.

Variable	Satisfaction	Advocacy	Repurchase.Intention	Perceived.Value
Company...Brand.Image	0.2699	0.356	0.3029	0.1372
New.Technologies	0.2573	0.3122	0.2539	0.1300
Product.Range	0.2148	0.2887	0.2420	0.1087
Net.Purchase.Price	0.2284	0.2241	0.2043	0.4386
Supporting.Evidence	0.2213	0.2936	0.2381	0.0849
Success.Rates	0.1863	0.2521	0.2518	0.0861
Limited.Backorders	0.1590	0.2187	0.2106	0.1099
Ease.of.Use	0.2131	0.2867	0.2762	0.1817
Accelerated.Treatment	0.1944	0.2337	0.2036	0.1510
Product.Familiarity	0.1847	0.2623	0.2762	0.1146
Restorative.Preference	0.2278	0.2873	0.2428	0.1504
Endorsement.by.KOLs	0.1797	0.2327	0.1706	0.0772
Sales.Reps	0.3272	0.3580	0.3045	0.1791
Customer.Service	0.2820	0.3341	0.2839	0.1931
Training.Programs	0.2440	0.2902	0.2548	0.1168
Practice.Support	0.2795	0.2669	0.2130	0.1998
Helpful.Web.site	0.2148	0.2170	0.1650	0.1880

Table 3.2: The Correlation Matrix for the Response Variables.

Variable	Satisfaction	Advocacy	Repurchase.Intention	Perceived.Value
Satisfaction	1	0.5776	0.4162	0.4136
Advocacy	0.5776	1	0.6336	0.4070
Repurchase.Intention	0.4162	0.6336	1	0.3776
Perceived.Value	0.4136	0.4070	0.3776	1

the high correlation among some of them suggests that data-based multicollinearity exists (Simon, 2013).

The predictions of a response variable from a set of predictors will be accurate even with the existence of multicollinearity. The strong relationship between the variables has no effect on how well the regression model predicts the response. Additionally, the calculated R^2 from the model will be the same as the one produced by a regression model in the absence of multicollinearity. If however, the aim is to understand how the various predictors impact the response, then multicollinearity causes a problem. This is because multicollinearity among the predictors leads to unreliable p-values and therefore incorrect understanding of the impact of the predictors on the developed model (Simon, 2013).

High correlation coefficients do not necessarily imply multicollinearity. Multicollinearity is a multivariate problem, thus the simple bivariate correlation matrix is not capable of detecting multicollinearity. The problem with multicollinearity is not only that two predictors are highly correlated, but that one predictor is nearly dependent on others.

To assess multicollinearity, the relationship between the predictors and how well each predictor is predicted from the others should be investigated. This means to examine the R^2 of each predictor regressed on the other variables. The greater the linear dependence among the predictor and the other predictors, the larger the R^2 (Simon, 2013).

A variance inflation factor (VIF) measures how much the variance of

the estimated coefficients is inflated when multicollinearity exists. The VIF is given by the formula VIF $= \frac{1}{1-R^2}$ where the quantity $\mathbf{1 - R^2}$ gives an estimation of the proportion of variance in the independent variable which is not explained by its relationship with the other predictors. High values of VIF indicate a high degree of multicollinearity. Most of the researches regards a VIF of a value of 10 as a cut-off point of serious multicollinearity. However, a few authors suggest that if any of the VIF of the predictors exceed 5, then multicollinearity should be investigated (Simon, 2013).

The HH package in R (Heiberger, 2012) has a VIF function that gives a vector of the VIF values for each predictor.

Table 3.3 shows the VIFs for the DIPD predictors split by country. The values greater than 5 are bolded.

The table shows that Italy respondents tend to have high correlated predictors. With VIF equals 6.716, 5.069 and 5.525, multicollinearity has been detected. Note however that these values are much smaller than the standard cut-off point of 10, countries such Canada, US, UK, France and Sweden have small VIFs, hence indicating the absence of linear relationship among their predictors.

Table 3.3: The VIF Values for DIPD Dataset by Country.

Predictors	**CA**	**CH**	**UK**	**US**	**SP**	**JA**	**GE**	**FR**	**IT**	**KO**	**SW**	**Swiss**	**NE**
Company...Brand.Image	3.031	3.110	3.014	2.943	3.165	3.466	3.945	3.013	4.783	3.089	2.767	4.221	2.81
New.Technologies	3.290	4.507	3.100	3.872	4.245	4.597	**5.460**	3.192	**6.716**	3.838	2.603	4.916	3.101
Product.Range	3.548	3.177	3.160	3.422	3.086	3.392	4.827	3.447	4.357	3.179	2.961	4.925	3.613
Net.Purchase.Price	1.442	2.164	1.448	1.525	1.286	1.574	1.602	1.432	2.115	2.555	1.447	1.892	1.415
Supporting.Evidence	2.936	3.200	2.854	2.929	2.900	3.629	4.053	3.005	**5.096**	3.024	2.811	4.215	2.581
Success.Rates	2.648	2.206	2.408	2.575	1.845	2.699	3.338	3.344	3.467	2.546	2.327	3.985	2.288
Limited.Backorders	2.523	2.197	2.426	2.351	1.524	1.856	3.565	2.448	3.842	2.903	2.033	4.708	2.391
Ease.of.Use	3.066	3.98	2.415	3.1904	2.409	3.050	4.634	2.857	**5.525**	3.102	3.165	**5.013**	2.703
Accelerated.Treatment	1.825	4.403	1.686	1.856	1.638	3.108	2.582	2.129	3.402	3.832	1.632	2.858	1.871
Product.Familiarity	2.738	2.828	2.336	2.906	2.170	2.423	4.333	3.432	4.548	3.746	2.493	4.281	3.07
Restorative.Preference	2.034	1.892	1.566	1.930	1.599	2.269	2.573	2.837	3.014	3.774	2.013	2.739	1.654
Endorsement.by.KOLs	1.709	1.512	1.643	1.814	1.762	1.902	2.345	2.218	2.445	2.607	1.745	3.037	1.967
Sales.Reps	2.428	3.313	2.483	2.334	2.375	2.819	2.756	2.152	3.021	3.558	2.62	3.341	2.775
Customer.Service	2.537	4.362	3.272	2.908	2.381	2.925	3.982	2.625	3.796	4.605	3.154	4.594	3.13
Training.Programs	2.597	2.660	2.510	2.847	2.329	3.180	3.108	2.222	4.122	4.024	2.471	3.608	2.281
Practice.Support	2.318	3.265	2.091	2.733	2.091	2.933	2.217	2.205	4.284	3.904	1.891	2.154	1.891
Helpful.Web.site	1.727	2.350	1.901	1.958	1.600	2.249	2.002	1.859	2.113	2.832	1.62	1.96	1.613

Chapter 4

Regression Trees to Predict Perception

Two of the most popular techniques used in data mining as prediction models are the classification and the regression trees (CART). These non-parametric statistical approaches can be a good choice for the purpose of getting fairly accurate results and when the dataset has large number of observations and variables. These approaches are extremely resistant to outliers and could be used when the analysis aims to identify the important variables in the dataset. (Steinberg and Colla, 1995)

Classification and regression trees are very common. These two techniques build trees in which each node represents a choice between a number of alternatives and each leaf denotes as a classification. Classification trees are used to obtain a prediction model for a response variable Y that takes a

finite number of unordered values (classes), from a set of predictors X. The method partitions the joint range of X into k disjoint sets $A_1, A_2, .., A_k$, such that the predicted value of Y is j if X belongs to A_j where k is the number of classes that Y has and $j = 1, 2, .., k$. Regression trees develop a prediction model of a regression type that is used to predict a response variable Y which takes on continuous values also based on a set of predictors X (Venables and Ripley, 2002).

The prediction models are developed by recursively partitioning the data and attaching a simple prediction model to each node. Suppose Y is a response variable to be predicted using matrix X that has N predictors $x_1, x_2, ..., x_N$. The prediction model for Y is obtained by growing a binary tree where each of the leaves (terminal nodes) of the tree represents a cell of the partition. For each node within this tree, a test to one of the predictors in X is applied. The result then is used to go into one of two possible directions of that particular sub-tree. This recursive operation ends by reaching a leaf node where Y is being predicted (Venables and Ripley, 2002).

The CART algorithm which was developed by Breiman et al. (1984), is the most popular of several algorithms developed to extract these prediction trees. From a user perspective this CART method could be summarized in three steps:

- Step 1: Build the tree using a recursive partitioning technique to select predictors and split the data points after the tree is identified.

- Step 2: The Pruning procedure. The result of this step is a nested subset of trees starting from the fully-extended tree and continuing the process until only one node of the tree remains.
- Step 3: Select the optimal tree that best fits the data.

4.1 Classification and Regression Trees Algorithm

There are two key ideas underlying classification and regression trees. The first is the recursive partitioning of the set of the predictors. The second is the pruning using a validation data set.

The partitioning method starts by finding one binary partition that maximizes the information about Y, this step gives a root and two child nodes. At each child node the same procedure is repeated by partitioning the set identified up to that point that would give the maximum information about Y and minimize the total impurity of its two child nodes. CART stops growing the tree when further splits give less than a minimal amount of extra information, or when it would result in nodes containing less than a specified percentage of the total data (Venables and Ripley, 2002).

4.1.1 Partitioning Algorithm

- Step 1: Start with a single node containing all the points y in Y. Calculate m_c and SS, where $m_c = \frac{1}{c}\sum_{i=1}^{c} y_i$ the prediction for leaf c and $SS = \sum_{c \in leaves(T)} \sum_{i \in C} (y_i - m_c)^2$.

- Step 2: If all the points in the node have the same value for all the predictors, stop. Otherwise, examine all the binary splits of all the predictors and choose the one which reduces SS as much as possible. If the largest decrease in SS is less than some threshold, or one of the resulting nodes contains less than q points, stop. Otherwise, take that split and create two new nodes.

- Step 3: In each new node, go back to Step 1 (Regression Trees, 2006).

4.2 Complexity Parameter and Pruning

Most of the time, the partitioning algorithm results in producing a extremely large tree that is likely to be over-fitting the data. The idea behind the pruning is to avoid over-fitting and build a tree that fits the data well.

CART algorithm prunes the tree T using a two-stage algorithm called cost complexity pruning. This algorithm, which was introduced by Breiman et al. (1984), starts by generating a sequence of alternative pruned trees as a first step, then a tree selection procedure is carried out to obtain the final model that best fits the data and would fit also well other data generated

from the same process (Venables and Ripley, 2002).

According to Breiman et al. (1984), a pruned sub-tree T^* is obtained by pruning off a branch T_t from a tree T where $T^* = T - T_t$.

For any sub-tree $T < T_{max}$, the complexity for that particular tree is defined as T_{size}, the number of terminal nodes in T. The cost complexity measure $R_\alpha(T)$ is defined as

$$R_\alpha(T) = R(T) + \alpha T_{size}$$

where T_{max} stands for a fully-expanded tree, T stands for a pruned sub-tree and $\alpha \geq 0$ is a real number called the complexity parameter. $R(T)$ represents the misclassification cost of T on the training data. Thus, the cost complexity measure is formed by the combination of the misclassifcation cost and a cost penalty for the tree complexity.

The next step of cost complexity pruning is to find the pruned sub-tree which minimizes $R_\alpha(T)$ for each value of α (Breiman et al., 1984).

4.3 Application to Prediction Perception Variables in the DIPD Survey

It is of our interest to determine which variables or attributes influence the decision of practitioners when they purchase dental implants. Classification and regression trees are non-parametric statistical approaches that could be

used to investigate such a problem.

The R package "**rpart**" (Therneau, Atkinson and Ripley, 2012) was implemented by Terry Therneau and Beth Atkinson (Therneau and Atkinson, 1997) to perform the CART analysis.

The `rpart` function was designed to build trees for two types of responses. The type of the response is specified using the argument `method` which takes either "anova" that leads to extract a regression tree for a numeric response, or "class" that leads to build a classification tree with a categorical response. Below is a part of the R code which uses `rpart` to fit a prediction model for Satisfaction.

```
> Im_S = rpart( Satisfaction~Company...Brand.Image + New.Technologies
+ Product.Range + Supporting.Evidence + Net.PurchasePrice + Limited.Backorders
+ Success.Rates + Ease.of.Use + Accelerated.Treatment + Product.Familiarity
+ Restorative.Preference + Endorsement.by.KOLs + Sales.Reps + Customer.Service
+ Training.Programs + Practice.Support + Helpful.Web.site , data = Data , method = "anova")
```

In 2009, a more specific package was introduced, "**rpartOrdinal**" (Archer, 2010) that has alternative splitting functions for fitting a classification tree when interest lies in predicting an ordinal response.

The use of the `rpartOrdinal` function is almost the same as the `rpart` function. It assumes that a set of numerical scores are assigned to the ordered categories of the response.

A few years later, a new R package "**rpartScore**" (Galimberti, Soffritti and Maso, 2012) was introduced to solve the unexpected results that arise

when using "**rpartOrdinal**" to conduct the CART analysis on ordinal response (Galimberti, Soffritti and Maso, 2012).

The response variables in the DIPD data are actually ranks. The implant practitioners are asked to give scores to determine their level of satisfactions with each response variable. Treating these ranks as numerical variables to apply regression trees and then extract a predictive model might lead to a exceptionally small prediction power and hence a small probability of correct classifications. On the other hand, treating these ranks as categorical variables and ignoring the fact that these classes or categories follow a certain order might lead to inaccurate results. For these reasons, the "**rpartScore**" is used to perform the CART analysis on the DIPD and hence treat the responses as ordinal variables.

rpartScore package assigns a set of increasing scores $s_1 < s_2 < ... < s_M$ to the ordered categories of the response variable Y. The misclassifications costs that **rpartScore** uses to build the prediction trees can be denoted by considering suitable transformation of the absolute differences between pairs of scores. The generalized Gini impurity function is calculated using the following formula

$$i_{GG}(t) = \sum \sum C(s_k|s_l)p(s_k|t)p(s_l|t)$$

Where $p(s_m|t)$ be the proportion of units in node t that belongs to the mth category of Y for $m = 1, ..., M$.

$C(s_k|s_l)$ represents the misclassification cost of assigning category s_k to unit belongs to category s_l. $C(s_k|s_l) = |s_k - s_l|$.

As a first step, a tree is built up using the `rpartScore` function. The predictive tree is constructed for each of the response variables using 17 predictors. The R code needed is

```
> Im_S.S = rpartScore ( Satisfaction~Company...Brand.Image + New.Technologies
+ Product.Range + Supporting.Evidence + Net.PurchasePrice + Limited.Backorders
+ Success.Rates + Ease.of.Use + Accelerated.Treatment + Product.Familiarity
+ Restorative.Preference + Endorsement.by.KOLs + Sales.Reps + Customer.Service
+ Training.Programs + Practice.Support + Helpful.Web.site , data = Data )
```

The `rpartScore` grows a tree that has not yet been pruned to a final size. Two other functions `printcp` and `plotcp` are used to extract information stored in the `rpartScore` object to help in deciding the value of the complexity parameter `cp` that should be used to prune the fitted trees. For instance,

```
> printcp(Im_P.S)
rpartScore(formula = Perceived.Value ~ Company...Brand.Image +
New.Technologies + Product.Range + Supporting.Evidence +
Net.Purchase.Price + Limited.Backorders + Success.Rates +
Ease.of.Use + Accelerated.Treatment + Product.Familiarity +
Restorative.Preference + Endorsement.by.KOLs + Sales.Reps +
Customer.Service + Training.Programs + Practice.Support +
Helpful.Web.site, data = Data)

Variables actually used in tree construction:
```

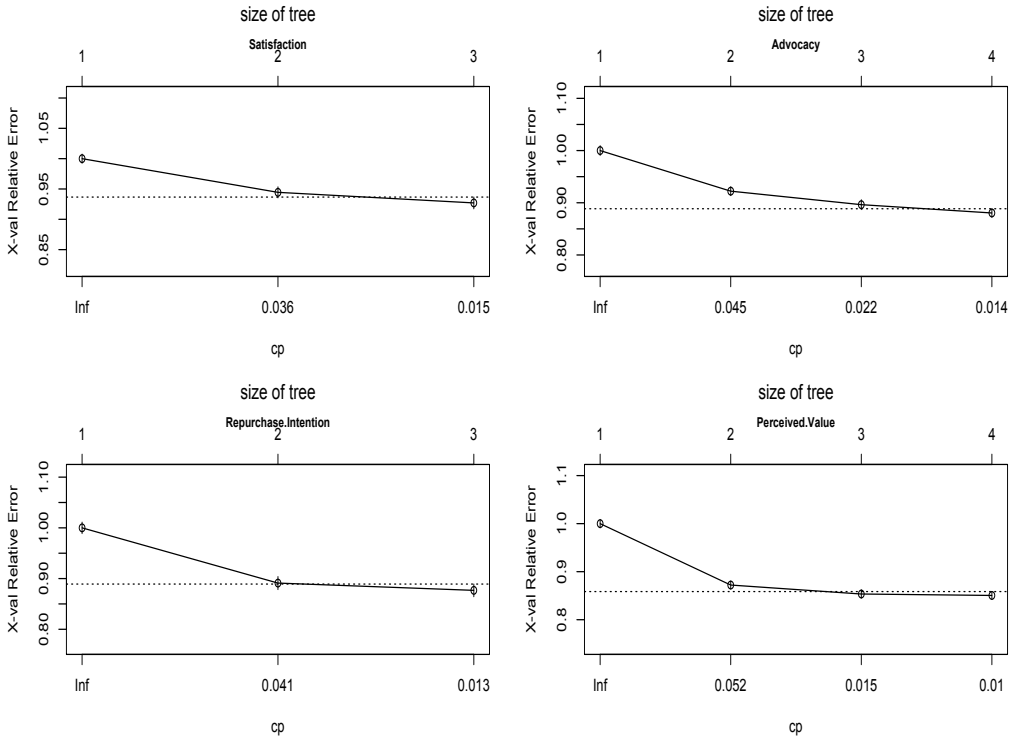

Figure 4.1: CP Plots of the Response Variables.

```
[1] Net.Purchase.Price
Root node error: 16711/8048 = 2.0764
n=8048 (1 observation deleted due to missingness)
CP       rel
1 0.127820      0   1.00000 1.00000 0.0085622
2 0.021423      1   0.87218 0.87218 0.0083053
3 0.010113      2   0.85076 0.85351 0.0086088
4 0.010000      3   0.84064 0.85052 0.0080761
```

The value of `cp` is chosen in a way that minimizes the `xerror`. This

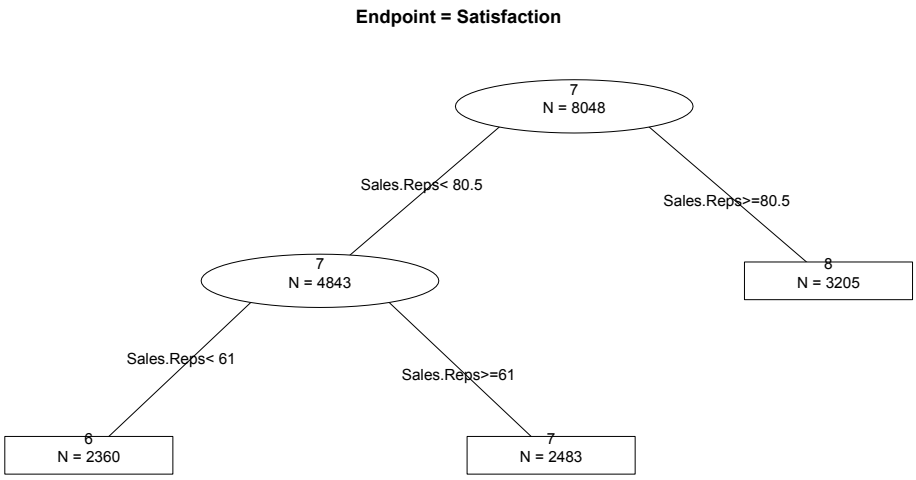

Figure 4.2: The Pruned Tree for Satisfaction.

value could also be chosen by examining the `cp plot` and taking the leftmost pruning point with the value below the dashed-line. After determining the `cp` value for each response variable, the fitted trees could be pruned to produce optimal trees that fit the data well (Galimberti, Soffritti and Maso, 2012).

Looking at the extracted tree for each response variable, we find that one or two out of the 17 predictors are used to build the prediction model. According to the analysis these predictors influence the purchasing behavior of the practitioners.

The final prediction model for Satisfaction relies only on one predictor which is Sales.Rep. Based on the extracted model, if the respondent assigns a value that is greater than 80 to Sales.Reps this is an indication of his high

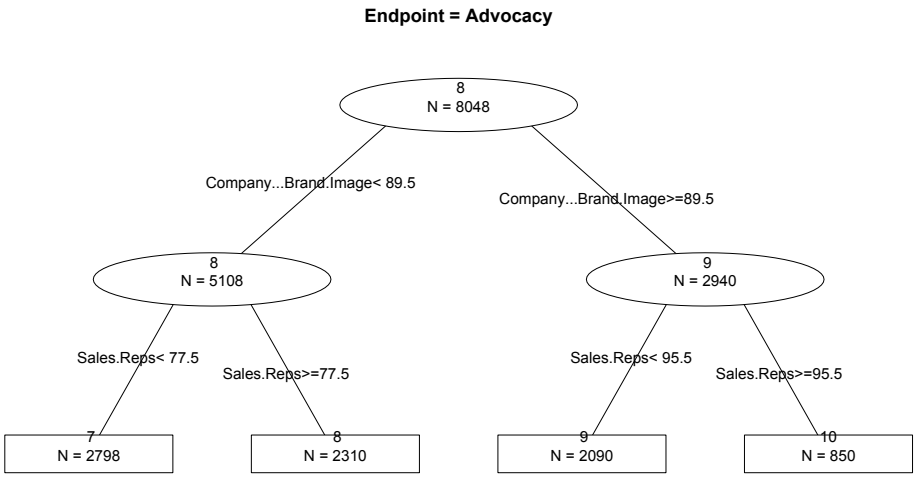

Figure 4.3: The Pruned Tree for Advocacy.

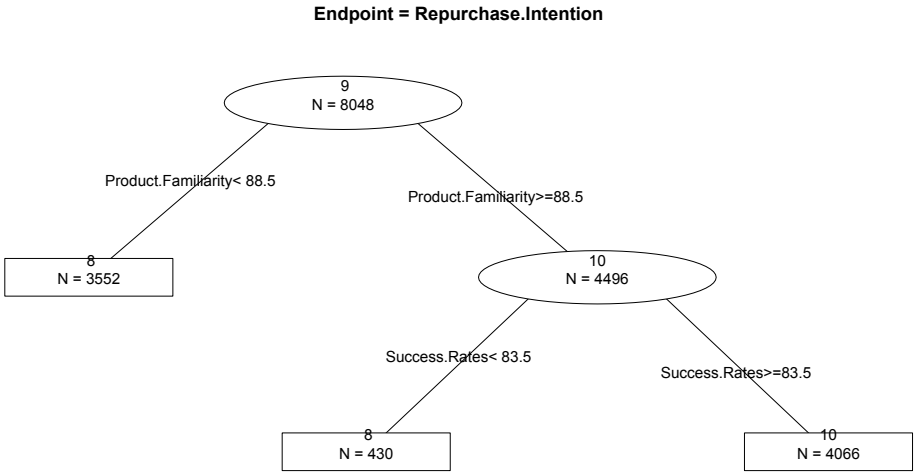

Figure 4.4: The Pruned Tree for Repurchase.Intention.

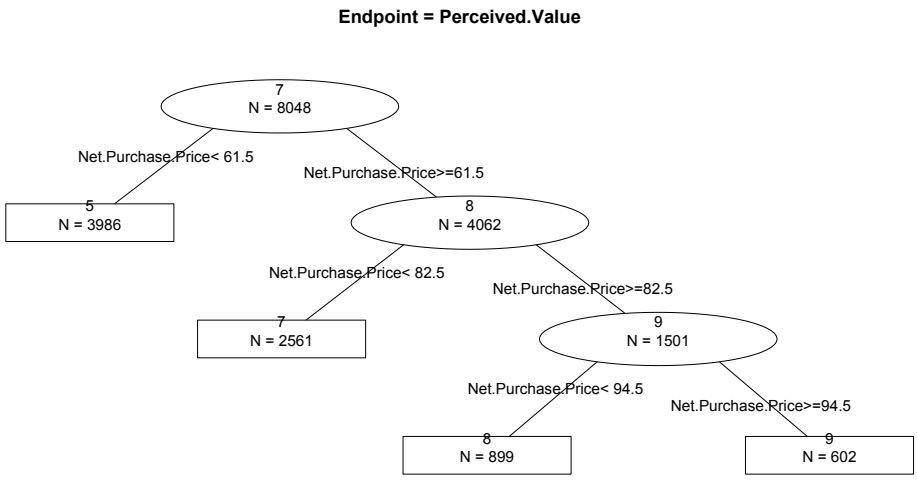

Figure 4.5: The Pruned Tree for Perceived.Value.

Satisfaction and he gives an 8 as a rank for that brand of dental implant. On the other hand, assigning values that are smaller than 61 to Sales.Reps give a smaller rank value to Satisfaction which is equal to 6.

For Advocacy, the final prediction model relies on Company...Brand.Image and Sales.Rep Predictors. High values assigned to Company...Brand.Image is an indication of a high advocate. When Company...Brand.Image is greater than 89 and Sales.Reps is greater than 95 this gives an advocacy score that is equal to 10. Small ranks of Advocacy are given by practitioner who assigns a value that is smaller than 77 to Sales.Reps.

For Repurchase.Intention, the final prediction model relies on Success.Rates

Table 4.1: Contingency Table for the Observed and the Predicted Scores.

Observed Scores											
Predicted Scores	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Satisfaction											
6	34	31	47	86	139	759	249	356	362	211	98
7	6	4	11	31	87	567	219	461	635	309	126
8	3	4	12	17	51	576	205	389	748	790	429
Advocacy											
7	50	36	37	67	60	345	245	334	392	202	280
8	10	10	17	27	38	318	219	476	761	505	702
9	6	3	3	7	13	74	71	127	357	379	498
10	6	6	3	3	2	50	22	52	164	218	857
Perceived.Value											
5	204	113	198	343	347	1013	388	485	421	213	236
7	19	7	28	61	137	482	281	504	617	338	248
9	11	8	9	11	15	147	61	148	233	278	448
Repurchase.Intention											
8	63	43	63	55	61	557	277	480	608	490	1337
10	19	10	10	13	13	216	90	207	374	466	2600

and Product.Familiarity. The Repurchase.Intention of a practitioner for repurchasing a specific dental implant is equal to 10 when this practitioner assigns a value that is greater than 88 to Product.Familiarity and a value greater than 83 to Success.Rate attributes.

For Perceived.Value, the final model relies only on Net.Purchase.Price. Perceived.Value gets small ranks (equal to 5) when practitioner assigns values smaller than 61 to Net.Purchase.Price.

In order to assess the predictive performance of these prediction trees the contingency table for the observed and the predicted scores is examined. From Table 4.1, the percentages of hitting the right classification are calculated.

Repurchase.Intention prediction model shows the highest percentage of getting the right ranks, which is equal to 40%. Satisfaction, Advocacy and Perceived.Value show remarkably lower percentages; 18%, 29% and 22% respectively.

In order to assess the association between the observed and predicted values, the Kendall's Tau-b coefficients are calculated from the produced contingency tables. Kendall's Tau-b coefficient is usually used to determine whether two variables that lie on an ordinal scale with possible ties are correlated. The estimator of Kendall's Tau-b coefficient takes into account the number of concordant and discordant pairs of observations in the two ordinal variables along with the number of the tied pairs. By tied pairs we mean the pairs of observations that have equal values of X and equal values of Y. The coefficient ranges between -1 and 1. The formula to get these coefficients is

$$(P - Q)/\sqrt{(N^2 - \sum r_i^2)(N^2 - \sum c_j^2)}$$

where:

P is the number of pairs of observations that place in the same order.

Q is the number of pairs of observations that place in the opposite order.

r_i and c_j are the total counts of row i and column j in the contingency table.

The coefficients tend to be small and range between 0.26 and 0.36. These small values indicates that associations are weak. The results agree with the

previously noticed weak associations between response and predictor variables.

Although the response variables actually have 10 classes (ranks), the prediction trees do not capture all the values for the response and they totally ignore the cases that have small ranks (0-5). This might happen due to the few number of respondents who have assigned these values as ranks for their level of satisfaction. By examining the training dataset, we found that this category of responses (0-5) represents small percentages of the data. For Satisfaction, Advocacy, Perceived.Value and Repurchase.Intention, these percentages are 7.04%, 5.14%, 19.33% and 4.42%, respectively.

To capture these cases we decided to classify the values for each response variable. This leads to increase the probability of hitting the right classification by decreasing the number of the predicted classes. The range for the response variables is divided into three categories, "Low" which contains the small ranks (6 and below), "Medium" which contains ranks (7 and 8) and "High" which captures the (9 and 10) ranks.

These three categories are treated as ordinal numbers and the CART analysis is applied again to improve the prediction performance of the fitted trees.

For Satisfaction, it appears that small ranks for New.Technologies (below 65) reduces the Satisfaction score given by the respondent.

As for the prediction model for Advocacy, it appears that only Company...Brand.Image affects the model.

Table 4.2: Kendall's tau-b coefficients Before and After Grouping the Response Variables.

Response Variable	**10-Scale**	**3-Scale**
Satisfaction	0.261	0.562
Advocacy	0.343	0.579
Perceived.Value	0.360	0.461
Repurchase.Intention	0.310	0.436

For Repurchase.Intention, the tree shows that Product.Familiarity and Company...Brand.Image have the most impact on the prediction model. A score higher than 88 for Product.Familiarity leads to high scores for Repurchase.Intention. Moreover, even if Product.Familiarity gets lower scores but Company...Brand.Image has a high score that is greater than 88 this leads to high scores for Repurchase.Intention.

The prediction model for Perceived.Value has not changed. However the new model captures the small values for the response variable. Scores below 61 for Net.PurchasePrice lead to lower scores for Perceived.Value.

Although this grouping has a slight effect on the trees' structure for the responses, it improves the predictive performance of the prediction models for Satisfaction, Advocacy, Repurchase.Intention and Perceived.Value to reach 79%, 66%, 65% and 73% respectively. The associations between the observed and the predicted classes are examined. Decreasing the number of the predicted classes leads to increases in the correlation coefficient for kendalls tau-b and these values range between .4 and .6. Table 4.2 has the coefficient values before and after categorizing the range of the response variables.

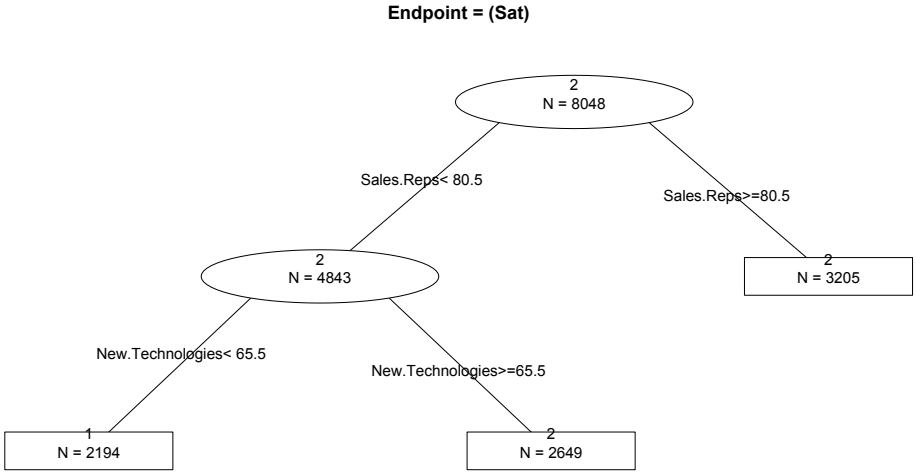

Figure 4.6: The Optimal Tree to Predict Satisfaction. 1=Low, 2=Medium, 3= High.

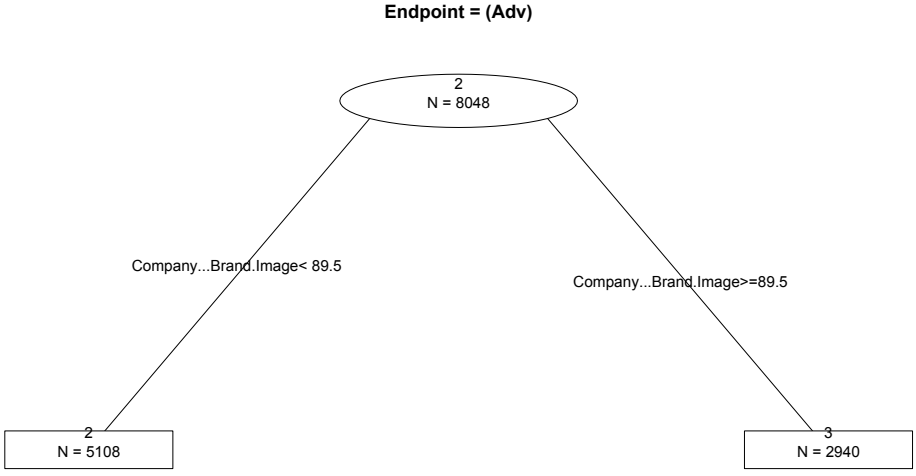

Figure 4.7: The Optimal Tree to Predict Advocacy. 1=Low, 2=Medium, 3= High.

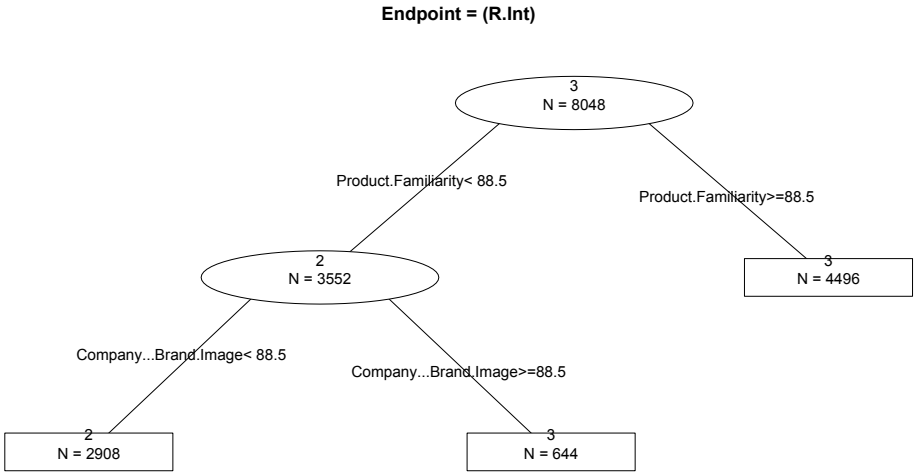

Figure 4.8: The Optimal Tree to Predict Repurchase.Intention. 1=Low, 2=Medium, 3= High.

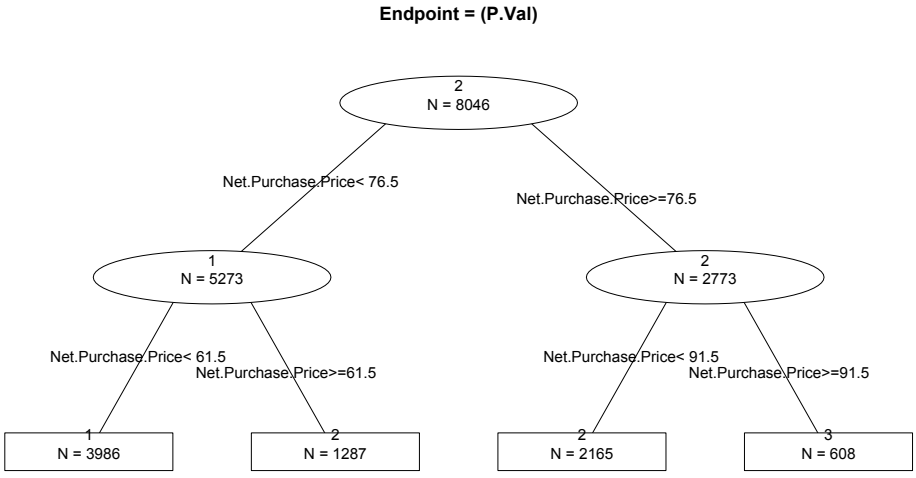

Figure 4.9: The Optimal Tree to Predict Perceived.Value. 1=Low, 2=Medium, 3= High.

Table 4.3: The Percentage of Hitting the Right Classifications for the Training and Testing Sets Before Grouping the Response Variables.

Response Variable	**Training Dataset**	**Testing Dataset**
Satisfaction	18%	19%
Advocacy	29%	31%
Perceived.Value	22%	20%
Repurchase.Intention	40%	39%

Table 4.4: The Percentage of Hitting the Right Classifications for the Training and Testing Sets After Grouping the Response Variables.

Response Variable	**Training Dataset**	**Testing Dataset**
Satisfaction	79%	80%
Advocacy	66%	64%
Perceived.Value	65%	64%
Repurchase.Intention	73%	75%

4.4 Model Validation

The fitting presented in the previous section were done after training portion of the data. The pruning process is aimed at producing a tree that not only fits the training data well but also that results in a model that predicts well in new data obtained by the same sampling process. In this regard, an additional form of process validation is to assess the performance of the models chosen on fresh data. There is where the other portion of the data in the split discussed earlier, the testing data is useful. Recall that the training data contained 8,052 cases (70% of the data) while the testing data contains 3,452 cases (30% of the data).

The performance of the prediction models built using CART analysis is

measured in terms of their success rates, which is the percentage of correctly classifying the response variables in the dataset.

Tables 4.1 and 4.2 display the success rates for both the training and testing datasets before and after grouping the response variables. Comparing the error rates for the training and testing datasets we realize that the differences in the rates are insignificant and this means that the prediction model which is developed using the training dataset is valid and could be used for future data points. However, the success rates for the generated model after grouping the scores are larger. This shows that grouping the value of response variables in classes improves prediction and increases its success rate, thus the model that is fitted using these data leads to more accurate results.

Chapter 5

Prediction Through Partial Least Squares (PLS) Modelling

5.1 PLS Components

5.1.1 PLS Regression

Modeling one or several dependent variables by means of a set of predictors has been one of the most common problems in data-analytical studies. Partial Least Squares Regression (PLSR) is a statistical method for modeling linear relations between two data matrices X and Y. This approach was first developed by Herman Wold in 1975 for the modeling of datasets in terms of chains of matrix blocks, called path models. In 1980, the model for two blocks X and Y was modified by Svante Wold and Harald Martens to

better suit the data from science and technology. The PLSR model is developed from a training set of K cases with N X-variables denoted by x_n where $n = 1, 2, ..., N$ and M Y-variables Y_m where $m = 1, 2, ..., M$. This regression model has the ability to model data with collinear variables in both X and Y matrices (Wold, Sjstrm, and Eriksson, 2001).

The first step in PLSR is to center the data matrices X and Y by subtracting their averages and then dividing by their standard deviations, resulting in X_0 and Y_0, respectively. This guarantees that each of the given variables has the same weight and prior importance when performing the analysis. The PLSR model then works to find a few underlying variables, called latent variables *(LV*'s). The number of these *LV*'s is unknown in advance and PLSR has specific algorithms to help in estimating this number (Wold, Sjstrm, and Eriksson, 2001).

A set of A orthogonal factors X-scores and Y-scores are extracted one by one from the original variables to form the two matrices $T = [t_1, t_2, ..., t_A]$ and $U = [u_1, u_2, ..., u_A]$, respectively. The a-th PLSR factors t_a and u_a are the weighted sums of the centered variables $t_a = X_0 w_a$ and $u_a = Y_0 q_a$. The T matrix is multiplied by the loading matrix P, where $X = TP' + E_X$. These loadings are calculated so that the X-residuals E_x are small. The U is also multiplied by the weight matrix C where $Y = UC' + E_y$ so that the residual matrix for the response Y, E_y is minimized (Wold, Sjstrm, and Eriksson, 2001).

5.1.2 Interpretation of the PLSR Model

PLSR forms a set of new X variables t_a where $a = 1, 2, ..., A$, as linear combinations of the original X's, and then uses these new t's as predictors of Y. The scores t and u contain information about the variables and their similarities and dissimilarities in the developed model. The weight matrix W is computed so that it maximizes the covariance between the response matrix Y and the T scores matrix. These weights provide information about how the variables combine to form the quantitative relation between X and Y and which of the x variables are significant (large w_a-values), and which ones provide the same information (similar w_a-values) (Wold, Sjstrm, and Eriksson, 2001).

The PLSR model produces residuals for both Y ans X matrices. These residuals are used to assess the amount of information in the data that PLSR fails to capture, in other words the part of the data that are not explained by the model PLSR generated. Large values for the residuals indicate that the model is poor, and the dataset possibly contains outlier data points. The normal probability plot of the residuals of the response variable Y helps identify these outliers.

5.2 PLSR Algorithms

Several PLSR algorithms have been developed to compute the PLSR components. The "**pls**" package (Mevik, Wehrens and Liland, 2011) in R

provides implementations for three of the PLSR algorithms: the nonlinear iterative partial least squares (NIPALS) algorithm, the kernel algorithm and the SIMPLS algorithm. These three algorithms differ in the computation time required to estimate the model components as well as the numerical accuracy for these components.

It has been shown in the literature that the kernel and NIPALS algorithms produce the same results. However, the first one has the shortest computation time. The SIMPLS algorithm produces the same t scores as the other two algorithms for single-response models, and slightly different results for multi-response models (Mevik and Wehrens, 2007).

The NIPALS algorithm is the standard algorithm for computing the PLSR components using the "**pls**" package. It starts with the centered form of the X and Y matrices, and proceeds as follows (Wold, Sjstrm and Eriksson, 2001):

1. Select a starting vector for u, usually one of the columns of Y. With a vector y, $u = y$.

2. Calculate the X-weights, w: $w = X^{'}u/u^{'}u$, norm w to $||w|| = 1.0$

3. Calculate X-scores, t: $t = Xw$

4. Calculate Y-weights, c: $c = Y^{'}t/t^{'}t$

5. Calculate an updated set of Y-scores, u: $u = Yc/c^{'}c$

6. Convergence is tested on the change in t, i.e., $||t_{old} - t_{new}||/||t_{new}|| < \varepsilon$, where ε is "small". If convergence has not been reached, return to step 2, otherwise continue with step 7, and then step 1. If there is only one y-variable, the procedure converges in a single iteration, and one proceeds directly step 7.

7. Remove the present component from X and Y and use these deflated matrices as X and Y in the next component. Here, the deflation of Y is optional; the results are equivalent whether Y is deflated or not $p = X't/(t't)$ $X = X - tp'$, $Y = Y - tc'$

8. Continue with the next component (back to step 1) until cross-validation indicates that there is no more significant information in X about Y.

This algorithm is applicable for both multivariate and univariate Y-variable matrix.

5.3 Selecting the Number of PLS Components

Although using a large number of components results in a good fitting model for the current observed data set, it might sometimes lead to over-fitting (getting a well fitting model with a low predictive power). Hence, it is always useful to choose a number of components that reduce the expected error when predicting the response variable from future observations. By

choosing a number of components, we mean retaining only significant components that explain more than 5% of the original variance in the response variable (Carrascal, Galvn and Gordo, 2009).

Cross-validation (CV) is a general statistical method used for choosing the number of components in PLSR. This method avoids over-fitting the data by not using the same dataset to fit a model and to estimate the prediction error. CV divides the data set into a number of groups K, also called segments which are randomly selected. Out of the K segments, a single segment is retained as the validation data for testing the fitted model, and the remaining $k-1$ segments are used as training data. The cross-validation process is then repeated K times, with each of the K segments used exactly once as the validation data. The K results from this iterative process are averaged to produce a single estimation (Mevik, and Wehrens, 2007).

It is of interest to find out which predictors have the most influence on predicting the response variable. Large coefficients show high importance of the particular predictors in modeling the response variable Y, while large loading values indicate that these predictors are essential in modeling X (Wold, Sjstrm, and Eriksson, 2001). A good measure of importance that takes into account both coefficients and loadings are the variable importance for the projection *(VIP)*. This measurement is based on the weighted sums of squares *(SS)* and it is given by

$$VIP_{jA} = \sqrt{n \sum_{a=1}^{A} w_{ja}^{*2} SS_a(Y) / \sum_{a=1}^{A} SS_a(Y)}$$

where $j = 1, ..., N$ and A is the number of selected components in the fitted model (Chong and Jun, 2005).

$$SS_a = b_a^2 * \sum_{k=1}^{K} t_{a_k}$$

where b_a and and t_a are the regression coefficient vector and the score vector of the ath component, respectively (Chong and Jun, 2005).

5.4 Application to Prediction of Perception Variables in the DIPD Survey

In assessing the reliability of the DIPD data, this study aims to find a prediction model with a reasonable predictive power - by fitting to a training data - to help in predicting the loyalty matrix for future respondents. The "**pls**" package in R has been used to fit the model. Before the PLSR component is extracted, the DIPD data is randomly divided into training and testing sets. About $\frac{2}{3}$ of the data are used for training and $\frac{1}{3}$ of it is used for testing purposes. A separate model for each of the response variables is fitted using 17 predictors.

```
> Data.plsr.S=plsr(Satisfaction~Company...Brand.Image+New.Technologies+
Product.Range+Supporting.Evidence+Net.Purchase.Price+Limited.Backorders+
Success.Rates+Ease.of.Use+Accelerated.Treatment+Product.Familiarity+
Restorative.Preference+Endorsement.by.KOLs+ Sales.Reps+Customer.Service+
Training.Programs+ Practice.Support+Helpful.Web.site , data=Data, validation="CV")
```

Table 5.1: The Percentage of the Variance Explained by Each Components for the Training Dataset for Each Response Variable.

Response	Comp 1	Comp 2	Comp 3	Comp 4	Comp 5	Comp 6	Comp 7	Comp 8	Comp 9	Comp 10
Satisfaction	52.67	11.16	4.38	6.15	4.85	3.44	3.91	2.83	3.99	2.698
Advocacy	50.10	10.05	3.71	6.58	4.39	4.01	3.71	4.30	2.30	3.38
Repurchase.Intention	50.15	7.63	4.56	6.03	4.87	3.63	2.38	3.15	2.82	5.14
Perceived.Value	46.39	6.07	4.14	8.08	5.09	4.64	3.32	4.34	2.69	1.91

The model was built using the *(CV)* cross-validation method which divides the data into segments - the default is 10 segments - which are randomly selected. To assess the optimal number of components for producing the model, a validation plot is used. The "**pls**" package produced the validation set using the RMSEP (Root mean squared error of prediction). The optimum number of components selected that minimizes RMSEP. Figure 5.1 shows a measure of prediction performance using RMSEP function against the number of components.

Table 5.1 shows the percentage of the variance explained by each component. The first five components explain most of the variation in the data.

One way to assess the predictive power of the model is to examine the goodness of fit R^2 of the prediction model which is given by

$$R^2 = 1 - \frac{SS_{Error}}{SS_{Total}}$$

This quantity indicates how well the model explains the variability in the data and is an indication of its potential to predict new observations. Figure 5.2 displays R^2 values for each number of component for each response variable. These values seem to be low, which leads to the conclusion that

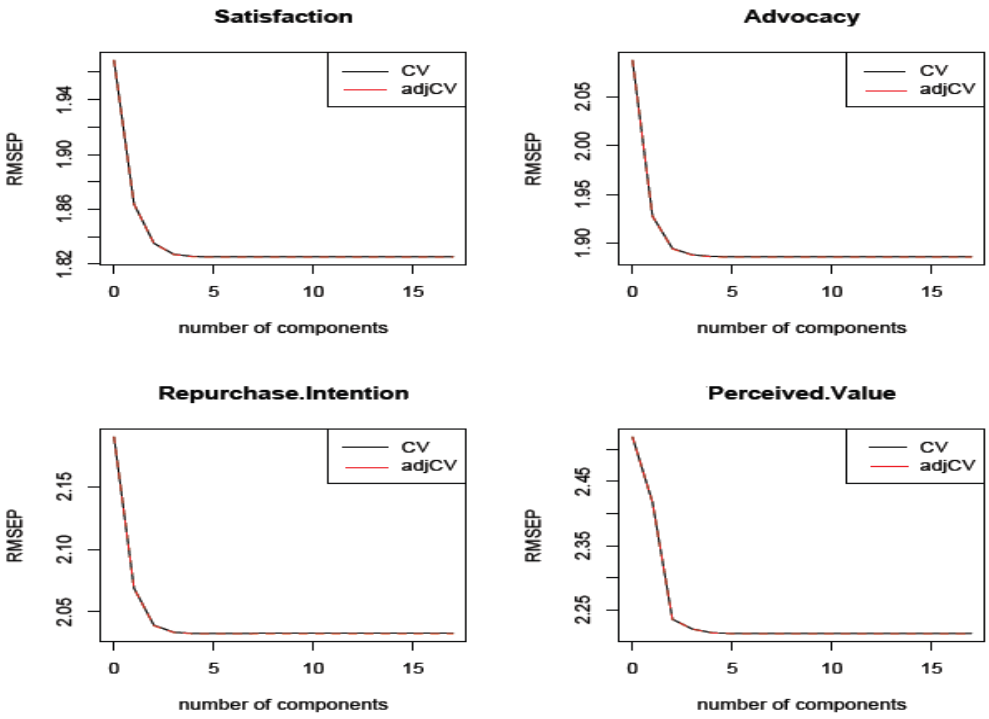

Figure 5.1: Cross-Validation Plot for the Training Data Set.

the prediction model is poor and it will not result in generating accurate predicted values. However, Donald R. Lehmann, in one of his papers states "Much of the research on consumer behavior has resulted in R^2 in the .05 and .10 range. As such, it had indicated little individual-level predictive power but significant relationships among variables such as income and TV viewing time. These variables are usually measured on a 5-8 point scale. Hence it is not at all surprising that people with incomes between 5 and 10 thousand dollars vary considerably in the amount of time they spend watching TV." (Lehmann, 1975). According to Lehmann, these low R^2 values reflect the high level of variation in the respondents' scores which is expected in such studies. Although the analysis results in low R^2's, significant effects that have practical importance might be present (Lehmann, 1975).

The prediction plot that shows the predicted versus the measured values has also been examined in order to assess the goodness of fit, see Figure 5.3. The data points are spread all over the plot which provides another indication of the weakness of the produced model. All of these results indicate that the resulting models are not effective for prediction.

5.5 Partitioned Partial Least Squares (PPLSR)

Analysts resort to PLS regression to build a prediction model that captures the variability in both the responses and the predictors. When the

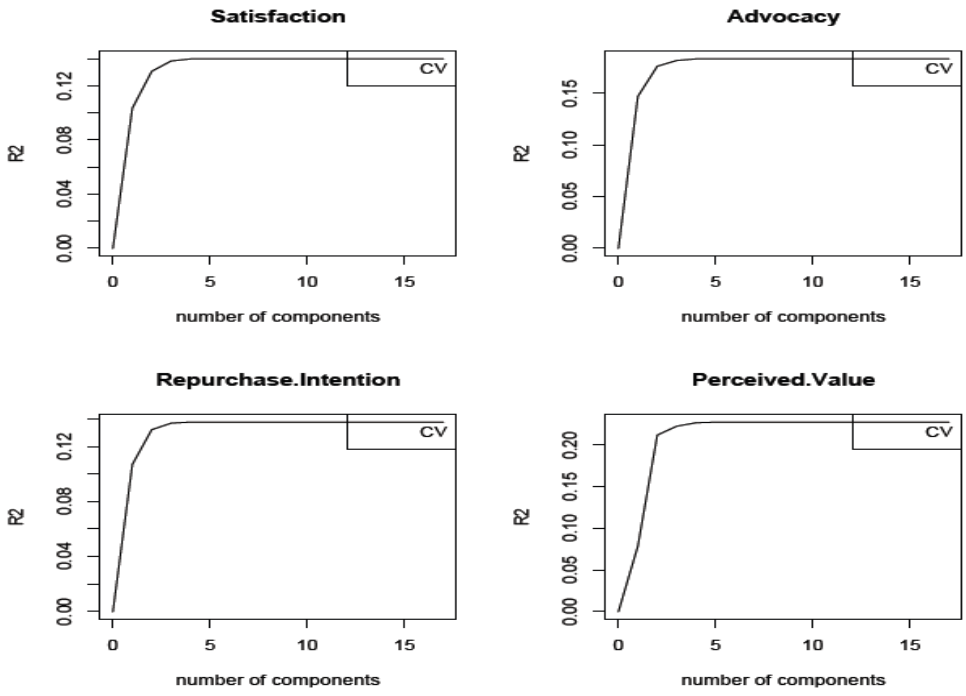

Figure 5.2: R^2 Plots for the Training Data Set for Each Response Variable.

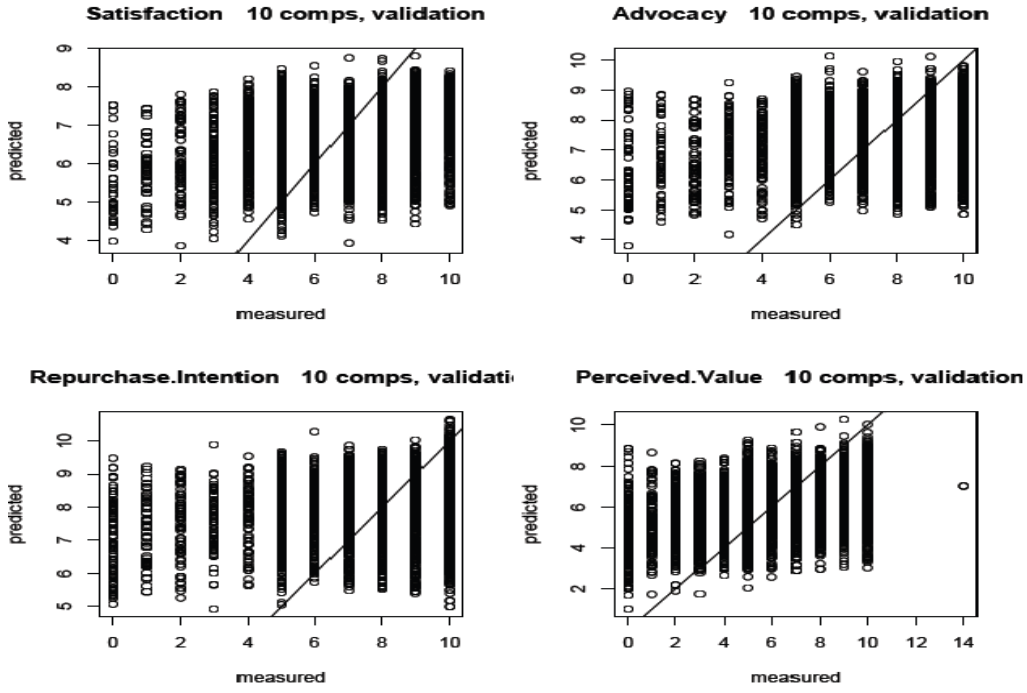

Figure 5.3: Measured VS. Predicted Values Plot for Response Variables.

dataset contains a large number of predictors the variability among the variables might dominate the model and the extracted components represent and emphasize the variance among the predictors at the expense of the prediction of the responses (Andersen and Runger, 2011). This would lead to produce a prediction model with low predictive power which would not be suitable for future predictions. Furthermore, the loadings and coefficients related to the generated components become difficult to interpret. In order to solve these issues some modifications have been applied to the regular PLS algorithm. One of the proposed algorithms tries to determine the most relevant predictors to predict the response and then use them to generate a PLS model. Using just part of the predictors could lead to a significant improvement in the predictive power of the PLS model.

The Partitioned Partial Least Squares Regression (PPLSR) algorithm was proposed by Andersen and Runger (2011) where it was used for analyzing data from a pharmaceutical batch fermentation process. In cases where the PLSR model does not predict well, the PPLSR algorithm is implemented to improve the performance of the generated PLSR prediction model.

The main idea in the PPLSR algorithm is to split the dataset into subgroups that only have part of the predictors. Basicaly, the predictive power for each subgroup is examined by building a PLSR model for each particular group. Once the subgroup with the highest predictive power is identified, only the selected predictors within that subgroup are included to generate the final PLSR model. Then this model is generated using the most relevant

predictors, the ones that are best suited for the prediction of the response variable (Andersen and Runger, 2011).

5.5.1 The Improved PPLSR Algorithm

The DIPD data was collected from respondents from different countries and different specialization groups. In addition to selecting predictors, the modified PPLSR algorithm searches for the group of respondents that provides the most accurate scores for the predictors to generate a more reliable prediction model.

The algorithm starts by selecting a subset A of the categorical predictors in the DIPD data set. The data points in the DIPD set are then divided into several categories based on the values taken by the predictors in A - cases that share the same values for the categorical predictors in A are placed in the same category. In order to determine which group of the cases has the highest prediction power, for each category, the prediction power for all possible subsets of the original predictors set is calculated.

Enhanced/Improved PPLSR algorithm:

1. Select a subset A of the categorical predictors from matrix X.

2. Categorize the original DIPD data set based on the predictors in A.

3. For each category:

(a) Select a subgroup of the original predictors in matrix X that contains at least 30% of the predictors.

(b) For each subgroup, extract a maximum of C PLS components while using the same response variable for each subgroup.

(c) For each generated model, calculate the R^2 quantity using exactly three PLS components.

(d) Repeat steps 3.a to 3.d for all possible subgroups of the original predictor set, keeping track of the calculated R^2's.

4. Select the category and the subgroup pair with the highest R^2.

5. For the chosen subgroups, build the PLS prediction model using only the predictors which belong in that group.

5.5.2 Application to the DIPD Data Set

In this study, three categorical predictors: Practice.Type, Specialty and Country have been chosen to represent the matrix A. Practice.Type is a categorical predictor with 2 levels: Surgical and Surgical-Prosthetic, while Specialty has 7 levels: General Practitioner, Endodontist, Prosthodontist, Orthodontist, Oral and Maxillofacial surgeon, Periodontist and Other Specialty. The subset that contains the data points that belong to the Surgical-Prosthetic category and are either Oral and Maxillofacial surgeon or Periodontist shows a considerable improvement in R^2. Taking Advocacy as an

example, fitting the PLSR model for all the data points gives an R^2 equal to 0.187 while choosing a subset of the data based on the Practice.Type and Specialty gives an R^2 equal to 0.336.

From the chosen subset the data point has been divided into smaller subsets based on the country of the practitioners. Fitting the PLSR model for some of these smaller subsets leads to higher R^2 values and therefore better prediction models. For Satisfaction, it turns out that the subset that includes practitioners from Canada, United Kingdom, Japan, South Korea, Germany, Italy, Sweden and Switzerland gives the highest R^2. For Advocacy, practitioners that belong in Canada, United Kingdom, Japan, South Korea, Germany, Italy, Sweden and Netherlands give the highest R^2. For Repurchase.Intention, the subset with practitioners from US, France, Spain and Korea has the highest R^2. For Perceived.Value, the subset with practitioners from Canada, China, United Kingdom, Spain, France, South Korea, Italy, Switzerland and Netherlands has the highest R^2.

Table 5.2 shows the improvement in R^2 after excluding some set from the DIPD data set. The process of getting the final data set with the highest R^2 values includes three steps. The first column has R^2 values for the all the cases in the DIPD data set. Step 1 excludes all respondents who are not Surgical-Prosthetic category and are not either Oral and Maxillofacial surgeon or Periodontist. Step 2 divides the chosen data points into smaller subsets based on the country of practitioners and then chooses the subgroups with higher R^2 values. Step 3 performs the PPLSR algorithm and chooses

Response	DIPD Data Set	Step 1	Step 2	Step 3
Satisfaction	0.1422	0.2787	0.3464	0.3543
Advocacy	0.1872	0.3359	0.4411	0.4486
Repurchase.Intention	0.1407	0.2294	0.2837	0.2984
Perceived.Value	0.2256	0.2651	0.3013	0.3013

Table 5.2: R^2 Values from the PPLSR Fits.

the most relevant predictors that lead to produce a model with the higher R^2. The results show some improvement on R^2 over the standard PLSR model. However, the final subsets that give the highest R^2 are extremely small and they represent 7% of the original dataset.

Looking at the RMSEP plots from Figure 5.4 we can see that the RMSEP values stop decreasing after three components. As a result, the new prediction model is built using just three PLSR components. Although some improvement appears in R^2 for the generated model after excluding some of the data points, the predicted vs measured values figures show remarkably modest performance.

Figure 5.6 examines the residuals' normality. The normal probability plot should produce an almost straight line if the data comes from a normal distribution. The QQ plots for Satisfaction and Perceived.Value show that the residuals are well-behaved. Although the QQ plots for Advocacy and Repurchase.Intention show small departures from the straight line, they are still considered normal.

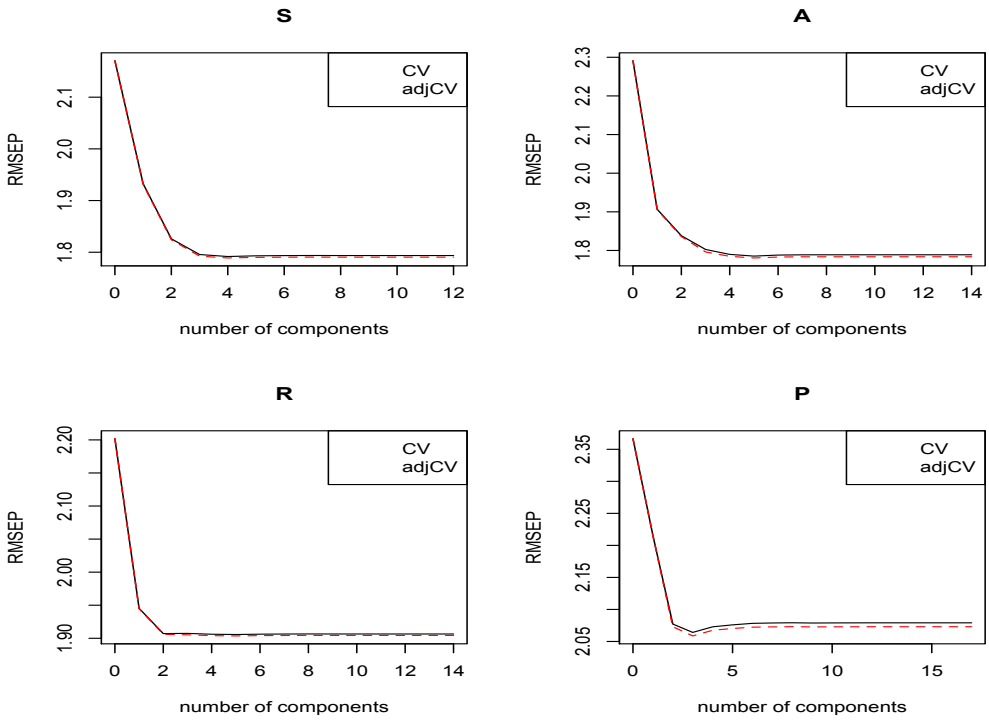

Figure 5.4: RMSEP Plots for the New Training Data Sets. S, A, R and P are the new responses after excluding some data points using the PPLSR algorithm.

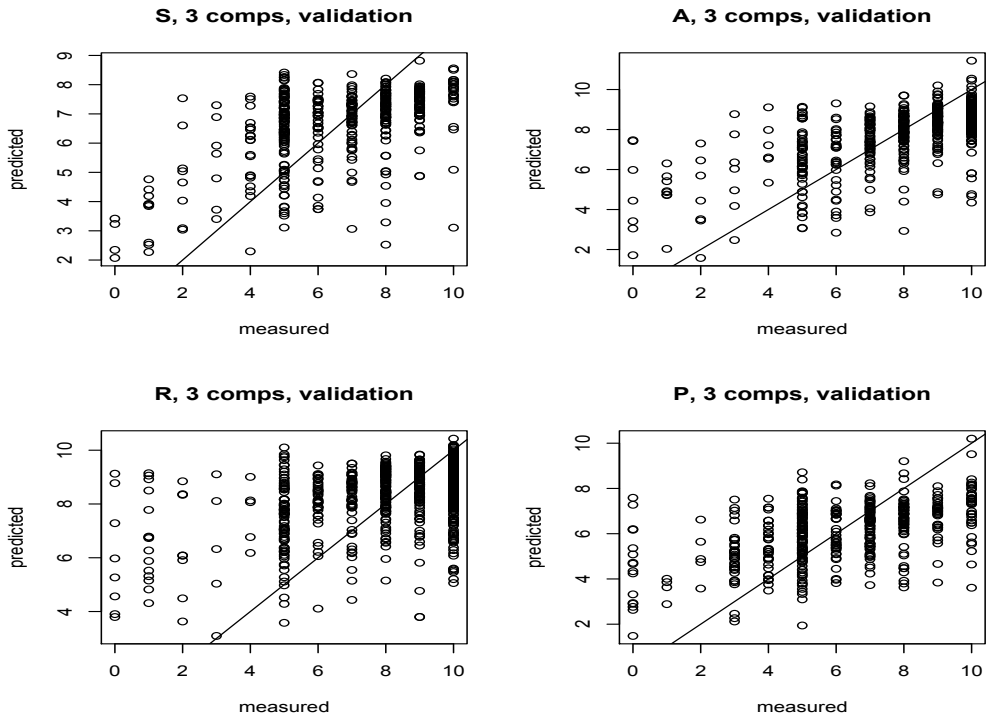

Figure 5.5: Measured VS. Predicted Values Plot for the New Response Variables. S: Satisfaction, A: Advocacy, R: Repurchase.Intention and P: Perceived.Value.

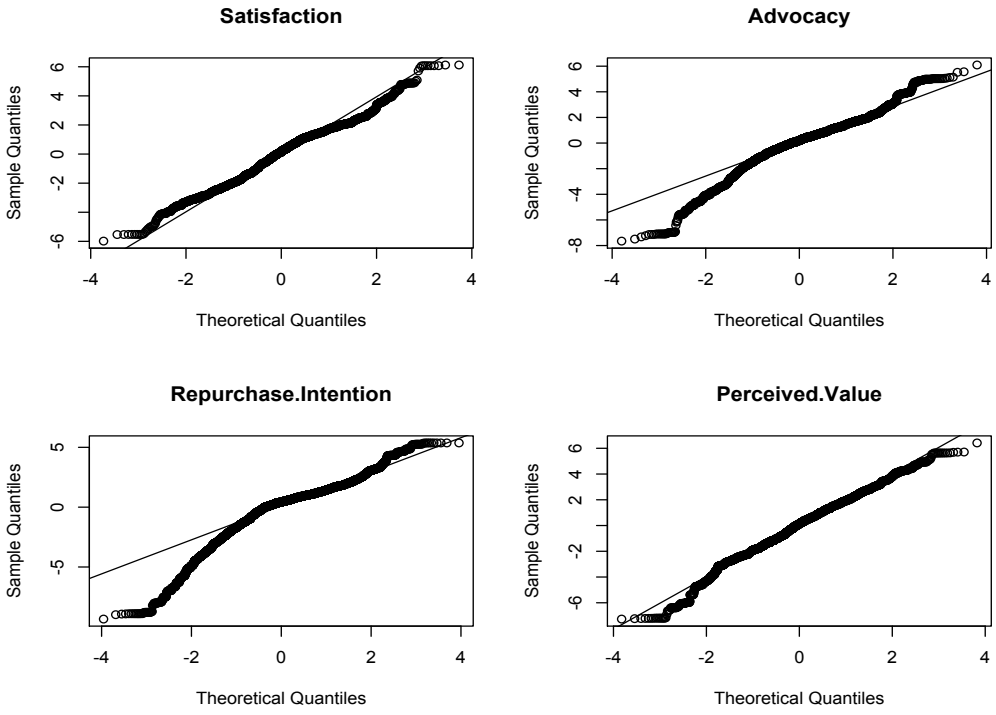

Figure 5.6: Normal QQ Plots for Residuals for Each Response Variable.

5.6 The Importance of the Predictors

The PLS coefficients could be used to get an idea of the impact of each predictor in predicting the response variables. Table 5.3 illustrates the coefficients of the predictors for predicting each response variable. Looking at the signs of the coefficients we can tell which attributes increases the response scores and which decreases its scores. For example, high scores for Limited.Backorders and Endorsement.by.KOLs have a negative impact on the customer's loyalty matrix and assigns small scores for each of the responses. For Perceived.Value, Accelerated.Treatment has a negative coefficient which means that high values for that attribute leads to a decrease in the response's score. On the other hand, Sales.Reps has the highest positive coefficients for Satisfaction, Advocacy and Repurchase.Intention and this indicate that the increases in Sales.Reps's scores lead to increases in these responses.

The Variable Importance in Projection (VIP) quantity measures the importance of each variable used to build a PLS model. All the predictors with a VIP value close to or greater than 1 can be considered essential. The predictors with VIP value significantly less than 1 are less influential. Table 5.4 shows that Sales.Reps have the most impact on predicting Satisfaction, Advocacy and Repurchase.Intention while Net.Purchase.Price has the most impact on Perceived.Value. On the other hand, a predictor such as Ease.of.Use (small VIP values) is less important to build the prediction model for the loyalty matrix.

Table 5.3: Predictor Coefficients for the PLS Models Fitted to each Loyalty Matrix Response Variable.

Predictors	**Satisfaction**	**Advocacy**	**Repurchase Intention**	**Perceived Value**
Company...Brand.Image	0.016	0.020	0.017	0.001
New.Technologies	-0.003	-0.002	-	0.002
Product.Range	-	-	0.001	-0.003
Net.Purchase.Price	0.007	0.001	0.007	0.047
Supporting.Evidence	-0.001	0.008	0.004	0.009
Success.Rates	-	-0.001	0.012	-0.015
Limited.Backorders	-0.013	-0.033	-0.010	-0.006
Ease.of.Use	-0.001	0.006	0.008	0.007
Accelerated.Treatment	0.001	0.009	-0.004	-0.021
Product.Familiarity	-0.002	0.0125	0.002	-0.005
Restorative.Preference	-	-	-	0.003
Endorsement.by.KOLs	-0.001	-0.003	-0.003	-0.001
Sales.Reps	0.039	0.036	0.014	0.008
Customer.Service	-	0.002	0.009	0.005
Training.Programs	-0.007	-0.001	-0.001	0.005
Practice.Support	0.010	0.008	0.004	0.008
Helpful.Web.site	-	-	-	-0.005

Table 5.4: The VIP Values of the Predictors for each Response Variable.

Predictors	**Satisfaction**	**Advocacy**	**Repurchase Intention**	**Perceived Value**
Company...Brand.Image	0.908	0.964	1.141	0.738
New.Technologies	0.841	0.877	-	0.796
Product.Range	-	-	0.886	0.707
Net.Purchase.Price	0.675	0.655	0.909	2.334
Supporting.Evidence	0.857	0.933	0.959	0.866
Success.Rates	-	0.471	0.775	0.819
Limited.Backorders	0.845	1.107	0.883	0.768
Ease.of.Use	0.612	0.670	0.730	0.519
Accelerated.Treatment	0.707	0.782	0.7058	1.268
Product.Familiarity	0.629	0.700	0.864	0.671
Restorative.Preference	-	-	-	0.803
Endorsement.by.KOLs	0.799	0.874	1.174	1.018
Sales.Reps	1.764	1.556	1.308	0.815
Customer.Service	-	1.151	1.107	0.727
Training.Programs	1.111	1.159	1.120	0.828
Practice.Support	1.515	1.463	1.195	1.099
Helpful.Web.site	-	-	-	0.869

5.7 Model Evaluation

The aim of model evaluation is assessing its performance in predicting the response variable for new data points. A proper model should have comparable performance on new data points as on the training set. As mentioned earlier, the data set was divided randomly into training and testing sets. The first set is used to fit the model. The second one is used to assess the performance of the fitted model. In addition to examining the MSE values for the training and testing data, the differences between the variance of the data sets and their MSEs are also examined. A perfect prediction model is the one with MSE value which is small compared to the variance of the data that are used to fit that model. Table 5.6 shows the variance for each of the responses along with the MSE value for its prediction model. Although the MSE values are less than the variances, which indicates that the generated model produces some improvement, these values are high and lead to the conclusion that the model is only moderately successful.

Although excluding some of the data points and some of the predictors shows an improvement in R^2, the reduction in the MSE values before and after the exclusion is not significant. Furthermore, the MSE values for both the training and testing data sets are almost the same and this indicates that the generated models have about the right number of components.

Table 5.5: Comparing the Variance and MSE for the Training Data

	Training Data Set		**Testing Data Set**	
Response Variable	**Variance**	**MSE**	**Variance**	**MSE**
Satisfaction	3.87	3.32	3.79	3.24
Advocacy	4.35	3.54	4.25	3.57
Perceived.Value	6.34	4.91	6.03	4.68
Repurchase.Intention	4.79	4.11	4.84	4.26

Table 5.6: Comparing the Variance and MSE for the Training Data After Excluding Some Data Points Using PPLSR Algorithm.

	Training Data Set		**Testing Data Set**	
Response Variable	**Variance**	**MSE**	**Variance**	**MSE**
Satisfaction	4.71	3.06	3.77	2.93
Advocacy	5.23	2.92	3.75	2.86
Perceived.Value	5.58	3.89	5.63	4.72
Repurchase.Intention	4.84	3.46	3.61	3.81

Chapter 6

Conclusions and Future Work

6.1 Summary of Results

- The early basic correlation analysis revealed weak sample correlations between prediction attributes and response variables. In the case of Satisfaction, Advocacy and Repurchase.Intention, the most promising predictors were Sales.Reps and Company...Brand.Image with sample correlations in the range 0.30 to 0.36. For Perceived.Value, the highest sample correlation ($r = 0.44$) was achieved with predictor Net.Purchase.Price. These findings insinuated limited potential among the attributes to predict the response variables. Of course, there is always the possibility of higher prediction power when used in combination, for instance, in an appropriate regression model.

- Classifications and regression trees (CARTs) were used to develop empirical tree models for predicting response values from attributes. The methods are appealing because: (i) they are nonparametric, (ii) they take into account nonlinearities, and (c) they have been very well studied, providing capabilities to both fit and validate the models. Our response variables are ordinal and were treated as such in fitting CARTs. First, trees were fitted to each response variable keeping the original scale 0 to 10 (11 response values). The correct classification rates achieved were in the range 44% to 57%. These are, of course, modest classification rates. Second, the 11 response values were categorized into 3 groups: Low, Medium, High. The correct classification rates obtained were from 60% to 82%. Note that, as might be expected, the predictors most highly correlated with the responses were the leading predictors driving the fitted trees (Figures 4.2-4.9).

- Partial least squares regression was used to develop prediction models for the response variables from the set of predictors. Additionally, it was used to determine the most influential attributes on predicting the response variables. PLSR is usually used to construct prediction models when the predictors are many and highly collinear. Since multicollinearity was detected in parts of the DIPD dataset, PLSR was chosen to avoid possible problems these correlated predictors might cause. This method treats the responses as numerical variables. First, all the cases in the DIPD dataset are used to fit the prediction models.

These models tend to have small R^2 that ranges between 0.1 to 0.2 and high MSEs which range from 3.32 to 4.91. Second, an algorithm was implemented to look for subsets in the DIPD that would give the highest R^2 and improve the predictive power of the generated models. This algorithm led to prediction models with higher R^2 range from 0.2 to 0.4 and MSEs range from 2.92 to 3.89. The VIP values are then calculated to determine the predictors that have the most impact on the responses. Table 5.4 has the VIP values for the predictors. Predictors with VIP values greater than 1 are the ones most highly correlated with the responses.

- The CART approaches produce a better model with a higher predictive power compared to the PLS regression model. However, the PLS prediction model helps get a clearer picture of which attributes are useful for future predictions. Both techniques show that there is some deficiency in the data set. The MSE values for the generated model are pretty large which lead to the conclusion companies can not fully rely on them because they are poor. Although CART analysis produces prediction models with reasonable MSE values, it only uses 1 or 2 out of 17 attributes for prediction purposes. Performing PLS regression and applying the PPLSR algorithm shows that some data that come from specific countries are less reliable. For some of the responses, the subset of the DIPD data set that belongs to US produces models with a remarkably small predictive power. Moreover, the Practice.Type and

Specialty of the respondents play a significant role in deciding how reliable the data are.

- Similar studies have been conducted on consumers behavior's using logistic and regular multiple regression and most of them produced prediction models with small R^2 values. According to Lehmann (1975), these small values of R^2 are expected. In his paper he states that most of the studies record R^2 that range from 0.05 to 0.1. Although the R^2 values produced by PLSR are small, they are not smaller than 0.1.

6.2 Future Work

CART and PLSR lead to generate prediction models with a moderate prediction power. However, there has been a growing interest in using modern data mining techniques in studying customer's behaviour and customer's satisfaction. Decision trees and neural networks are more advanced techniques that could be applied on the DIPD dataset to generate more accurate prediction models. On the other hand, some work could be done in order to identify in systematic way the subsets that would result in better models. This analysis uses only three predictors to identify these groups. For future studies, other attributes such as years in specialty and number of implant surgeries done by each practitioner could be used to identify these subsets.

Bibliography

[24] American Academy of Periodontology (2012). Dental Implants Perio.org. (n.d.). *Perio.org*. Retrieved January 21, 2013, from *http://www.perio.org/consumer/dental-implants*.

[24] Andersen, S., and Runger, G. (2011). Partitioned Partial Least Squares Regression with Application to Batch Fermentation Process. *Journal of Chemometrics, 25*(4), 159-168.

[24] Archer, K. (2010). rpartOrdinal: An R Package for Deriving a Classification Tree for Predicting an Ordinal Response. *Journal of Statistical Software, 34*(7), 1-17.

[24] Breiman, L. (1984). Classification and Regression Trees. Belmont, Calif.: Wadsworth International Group.

[24] Carrascal, L., Galvn, I., and Gordo, O. (2009). Partial least squares regression as an alternative to current regression methods used in ecology. *Oikos, 118*(5), 681-690.

[24] Chong, I., and Jun, C. (2005). Performance of some variable selection methods when multicollinearity is present. *Chemometrics and Intelligent Laboratory Systems , 78* , 103112.

[24] Galimberti, G., Soffritti, G., and Maso, M. D. (2012). Classification Trees for Ordinal Responses in R: The rpartScore Package. *Journal of Statistical Software, 47*(10), 1-25.

[24] Hayes, B. E.. Measuring customer satisfaction and loyalty: survey design, use, and statistical analysis methods. 3rd ed. Milwaukee, Wis.: ASQ Quality Press, 2008. Print.

[24] Heiberger, R. M. (2012). HH: Statistical Analysis and Data Display: Heiberger and Holland. R package version 2.3-17. *http://CRAN.R-project.org/package=HH.*

[24] Lehmann, D. R. (1975). Validity and Goodness of Fit in Data Analysis. *Advances in Consumer Research, 02*, 741-750.

[24] Liao, D., and Valliant, R. (2012). Variance inflation factors in the analysis of complex survey data . *Statistics Canada, Catalogue No. 12-001-X*, 38(1), 53-62.

[24] Mevik, B., and Wehrens, R. (2007). The pls Package: Principal Component and Partial Least Squares Regression in R. *Journal of Statistical Software*, 18(2), 1-24.

[24] Mevik, B., Wehrens, R., and Liland, K.(2011). pls: Partial Least Squares and Principal Component regression. R package version 2.3-0. *http://CRAN.R-project.org/package=pls.*

[24] Millennium Research Group, Inc (MRG). (n.d.). Vital medical technology insight and analytics - Millennium Research Group, Inc (MRG). Retrieved March 27, 2013, from *http://mrg.net/About-Us.aspx*

[24] Mosby, I. (2009). *Mosby's Medical Dictionary* (8th ed.). St. Louis, MO: Mosby.

[24] Piccarreta, R. (2008). Classication Trees for Ordinal Variable. *Computational Statistics, 23*, 407-427.

[24] R Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, *http://www.R-project.org/.*

[24] Regression Trees (2006). (n.d.). 36-350: Data Mining. Retrieved January 10, 2012, from *www.stat.cmu.edu/ cshalizi/350-2006/lecture-10.pdf*

[24] Therneau, TM., and Atkinson, EJ. (1997). Introduction to Recursive Partitioning Using the rpart Routine. *Technical Report 61, Section of Biostatistics, Mayo Clinic, Rochester. http://www.mayo.edu/hsr/techrpt/61.pdf.*

[24] Therneau, TM., and Atkinson, EJ. and Ripley, B. (2012). rpart:

Recursive Partitioning. R package version 4.10. *http://CRAN.R-project.org/package=rpart.*

[24] Simon, L. (n.d.). Multicollinearity. *STAT 501 Regression Methods*. Retrieved January 21, 2013. *http://online.stat.psu.edu.*

[24] Steinberg, D., and Colla, P. (1995). CART: Tree-Structured Nonparametric Data Analysis.

[24] Venables, W. N., and Ripley, B. D. (2002). Modern applied statistics with S (4th ed.). New York: Springer.

[24] Wold, S., Sjstrm, M., and Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems, 58*(2), 109-130.