



**APPLICATION AND MODEL TO CROSS-SELL IPTV SERVICE
UNDER A B2B FRAMEWORK**

By

Claudia Mack

Mémoire présenté en vue de l'obtention du grade de maîtrise ès sciences
(Intelligence d'affaires)

Thesis submitted in partial fulfillment of the requirements for the degree of
Master of Science in Administration (Business Intelligence)

OCTOBER 2015

© Claudia Mack 2015

Résumé

L'industrie des télécommunications est confrontée à une problématique caractérisée par le fait que les consommateurs ont la possibilité d'acheter soit un service par fournisseur ou bien plusieurs services offerts en forfait par un seul fournisseur. Cette étude utilise des techniques de l'intelligence d'affaires, soit l'entreposage de données et l'exploitation de données, afin d'analyser deux problèmes liés au service de télévision sur le protocole Internet (IP) : l'acquisition de nouveaux clients et la vente croisée aux clients existants; le tout dans le contexte de marchés d'affaires où les clients sont des petites entreprises. À partir de cette recherche, deux modèles sont développés. Ceux-ci permettent à une compagnie opérant dans le marché d'affaires d'obtenir des renseignements provenant d'une gamme de bases de données internes ainsi que d'une base externe, fournie par une compagnie de données commerciales, afin de générer une liste de clients potentiels.

La méthodologie utilisée est basée sur une approche d'analyse des bases de données en marketing, qui inclut l'estimation des modèles pour chacune des problématiques, c'est-à-dire celle de la croissance et celle de l'acquisition. Des modèles prédictifs basés sur une analyse de régression logistique furent construits et évalués au niveau de leur pouvoir prédictif et de leur performance pour optimiser les campagnes de la vente croisée et de l'acquisition de clients dans la catégorie microentreprises. Les modèles consacrés à la croissance produisent une liste de clients abonnés soit au service téléphonique ou au service Internet, ayant la plus haute probabilité de s'abonner au service de télévision IP. Les modèles consacrés à l'acquisition produisent une liste de microentreprises n'étant pas clients de la compagnie et ayant la plus haute probabilité de s'abonner au service de télévision IP. Les modèles pourvoient des entreprises en ordre décroissant et groupées en déciles. Les microentreprises dans les déciles supérieurs pourront être contactées pour des fins de marketing direct.

Les modèles obtenus identifient les facteurs ayant une forte influence sur la probabilité de conversion, notamment les profils démographiques des entreprises et leurs comportements de consommation. Ces modèles fournissent un « lift » de 4,88 pour la croissance et de 3,15 pour l'acquisition.

Summary

The telecommunications industry is confronted with the customers' ability to buy their services in a bundle from one single supplier or separately from a number of agents. This research looks at the business intelligence (BI) techniques of data warehousing (DW) and data mining (DM) to analyze two business problems: acquiring new customers and cross-selling to active clients under a business-to-business (B2B) context and pertaining to one single service, Internet protocol television (IPTV). The research develops two final models that will allow a B2B company to obtain information from a range of internal databases and from one leased database, provided by a commercial data supplier, and produce a list of prospects in the small business and microenterprise category.

The methodology in this thesis is based on a marketing database analysis approach, which includes the model assessment for each business problem, namely growth and acquisition. Predictive models based on logistic regression analysis were estimated and evaluated in the context of their performance to optimize cross-selling and acquisition campaigns. Models focusing on growth by cross-selling yielded presently active business phone or business Internet clients with the highest probability to subscribe to the IPTV service. Models focusing on acquisition yielded potential small businesses with no subscription to telecom services supplied by the B2B firm with the highest probability to subscribe to the IPTV service. Each model generates a list of enterprises in descending order and grouped in deciles. The enterprises in the top deciles may be contacted via direct marketing.

The resulting models identified the influential factors in terms of firm demographics and consumer behaviour that cause conversion probabilities. These models provided a lift of 4.88 for growth, and 3.15 for acquisition.

Acknowledgments

I would like to express my gratitude to an exceptional professor and supervisor, Marc Fredette, whose expertise, understanding, guidance and patience added considerably to my graduate experience. I appreciate his vast knowledge and skill, and his assistance in writing this thesis.

I am grateful to the Director and the Senior Manager of the Business Intelligence department for their provision of the data evaluated in this study. Appreciation goes to the ETL development team for their technical assistance throughout my research project, and to every business intelligence specialist for all the instances in which their assistance helped me along the way.

I thank my family for the support they provided me throughout my graduate studies. In particular, I will be forever indebted to my daughter and best friend, Sara, without whose love and encouragement at all times, I would not have concluded this endeavor.

Table of Contents

Résumé.....	ii
Summary.....	iii
Acknowledgments.....	iiv
Table of Contents.....	v
List of Tables.....	vii
List of Figures.....	ix
Chapter 1. Introduction	1
1.1 Objectives.....	2
1.2 Thesis Outline.....	3
Chapter 2. Literature Review.....	4
2.1 Customer Relationship Management	4
2.2 Acquisition and IPTV Adoption	6
2.3 Growth and Cross-Selling	7
2.4 The Telecommunications Sector	7
2.5 The B2B Context	9
Chapter 3. Conceptual Framework.....	13
3.1 Data Warehousing	13
3.1.1 Data Marts.....	15
3.1.2 Accessing Data – OLAP	18
3.2 Data Mining	20
3.2.1 Data Mining Tasks.....	20
3.2.2 Logistic Regression	22
Chapter 4. Modeling Process.....	26
4.1 Methodology	26
4.1.1 Data	29
4.1.2 Data Preprocessing.....	30
4.1.3 Data Cleaning	34
4.1.4 Industry Analysis	35
4.1.5 Model Summary	40
4.2 Cross-Selling Model.....	41
4.2.1 Model A0	43
4.2.1.1 Model Development.....	43
4.2.1.2 ID Variable Manipulation.....	43
4.2.1.3 Model A0 [LA_ID].....	45
Variable Selection	45
Model Estimation	45
4.2.1.4 Model A0 [PK_ID]	48
Variable Selection	48

Model Estimation	48
4.2.2 Model A.....	51
4.2.2.1 Model A [NI].....	53
Variable Selection	54
Model Estimation	54
4.2.2.2 Model A [WI].....	58
Variable Selection	59
Model Estimation	61
4.3 Acquisition Model.....	66
4.3.1 Model B.....	66
4.3.1.1 Variable Selection	66
4.3.1.2 Model Estimation.....	67
Chapter 5. Analysis of Results	70
5.1 Model A	71
5.1.1 Model Scoring	71
5.1.2 Analysis.....	75
5.2 Model B	80
5.2.1 Model Scoring	80
5.3 BI Process Integration.....	81
Chapter 6. Conclusion and Discussion.....	82
6.1 Conclusions.....	82
6.2 Discussion and Future Work.....	84
References	87
Appendix.....	92

List of Tables

Table 4.1	Growth and Acquisition Modeling Overview	27
Table 4.2	Database Structure	32
Table 4.3	Two Analyzing Periods for Cross-Selling Model	33
Table 4.4	One-Digit & Two-Digit Industry Classification	38
Table 4.5	Chi-Square Test of Independence for Industry	39
Table 4.6	Summary of Models	40
Table 4.7	Description of Available Variables after Preprocessing.....	41
Table 4.8	Two A0 Models.....	43
Table 4.9	Description of Available Variables for Model A0	44
Table 4.10	Chi-Square Test of Independence for Model A0 [LA]	45
Table 4.11	Model A0 [LA]	45
Table 4.12	Estimates for Model A0 [LA].....	46
Table 4.13	Chi-Square Test of Independence for Model A0 [PK].....	48
Table 4.14	Model A0 [PK].....	48
Table 4.15	Estimates for Model A0 [PK]	49
Table 4.16	Two Models A	51
Table 4.17	Firmographics for Models A [NI] and A [WI]	51
Table 4.18	Behaviour-Related Variables for Models A [NI] and A [WI]	53
Table 4.19	Chi-Square Test of Independence for Model A [NI]	54
Table 4.20	Stepwise Selection for Model A [NI].....	54
Table 4.21	Model A [NI]	55
Table 4.22	Estimates for Model A [NI].....	56
Table 4.23	Description of New Variables for Model A [WI]	59
Table 4.24	Additional Behaviour-Related Variables for Model A [WI]	60
Table 4.25	Chi-Square Test of Independence for Model A [WI].....	60
Table 4.26	Correlation Matrix for Model A [WI].....	61
Table 4.27	Stepwise Selection for Model A [WI].....	61
Table 4.28	Model A [WI].....	62
Table 4.29	Estimates for Model A [WI]	63
Table 4.30	Description of Variables for Model B	66
Table 4.31	Chi-Square Test of Independence for Model B.....	67

Table 4.32	Model B.....	68
Table 4.33	Estimates for Model B	68
Table 5.1	Comparison of Cross-Selling Models	75
Table 5.2	Odds Ratio Estimates for Model A [NI]	76
Table 5.3	Odds Ratio Estimates for Model A [WI] Part 1	77
Table 5.4	Odds Ratio Estimates for Model A [WI] Part 2.....	78

List of Figures

Figure 2.1	Customer Lifecycle Management	4
Figure 3.1	ETL Process	15
Figure 3.2	Difference between Data Warehouse and Data Mart	16
Figure 3.3	Dependent and Independent Data Marts.....	17
Figure 3.4	Business Intelligence Platform	18
Figure 3.5	Star Schema	19
Figure 3.6	Phases of the CRISP Process for Data Mining.....	21
Figure 3.7	The Logistic Function	24
Figure 4.1	Small Business and Microenterprise Market.....	27
Figure 4.2	Data Process Flow for Prediction Modeling	28
Figure 4.3	Arrangement of Available Data Sources.....	31
Figure 4.4	Industry Code Description and Examples.....	36
Figure 4.5	TV Service by Industry	37
Figure 4.6	Conceptual Cross-Selling Model A [WI]	58
Figure 5.1	Comparative Lift Chart for Models A ₀ [LA] and A ₀ [PK]	72
Figure 5.2	Lift Curve for Model A ₀ [PK].....	72
Figure 5.3	Comparative Lift Chart for Models A [NI] and A [WI].....	73
Figure 5.4	Lift Curve for Model A [WI].....	73
Figure 5.5	ROC Curves for Model A [NI]	74
Figure 5.6	ROC Curves for Model A [WI].....	74
Figure 5.7	Lift Curve for Model B	80
Figure 5.8	ROC Curves for Model B.....	80
Figure 5.9	Project Trees in SAS/EG for Model A.....	81

Chapter 1.

Introduction

For firms offering telecom services in both a business-to-client (B2C) and B2B environments, it is becoming increasingly difficult to acquire new customers as the industry grows more competitive within an oligopolistic market. The same applies when cross-selling new services to current clients for growth purposes. This is especially the case for services providing access to TV, which can operate under a number of technologies: satellite, cable, IPTV, and more recently, over-the-top content (OTT) or live streaming. The latter becoming the rival of all former services as it allows customers to watch their favourite programs, films and live performances and sport events everywhere on practically any Internet-connected device without the involvement of an additional supplier. As a result, telecom firms are susceptible to see their customers cancel their TV service completely.

As a growing number of customers make the move out of cable or satellite, they have the option to either cut the cord or switch to IPTV. Yet, IPTV requires an Internet connection and demands the subscriber to be located within the footprint that supports the IPTV technology, namely the fiber optic network. In Canada, this network has only been introduced across some areas within the two most populated regions, Ontario and Quebec.

The firm that serves as a case study during this research provides telecom services in Canada to business clients in the small business and microenterprise category. In Canada, a business is considered small if it employs between 1 and 99 people, and a microenterprise if it has between 1 to 5 employees. The present research focuses on the business problems of growth and acquisition, and models will be estimated for each one of these strategies. The first model will target current clients

through an appropriate customer growth or development strategy such as cross-selling so that a firm can enhance the value of its present clients. The second model will target potential customers, addressing the customer acquisition process.

There is no previous model devoted to the development or acquisition of clients in this customer category for the selected service that could serve as reference and against which the present models could be compared. However, a 2.07 percent response rate has been set down by the prevailing direct marketing program. The Marketing department has based its targeting process on their instinct by selecting small businesses located within the fiber footprint.

Furthermore, the same company also provides satellite TV to customers located outside the fiber optic network. This feature will be used to compare the needs exerting small business customers to acquire TV as a service regardless of the technology.

The firm claims that IPTV is the first choice for small and micro businesses, who look for innovative solutions to keep their customers entertained and informed. As the firm extends its covered network, it seeks to establish the IPTV service in major cities across six additional locations. Therefore, it is important to have a tool ready to target the prospects with the highest probability to convert within the many new locations to be added in the coming years.

1.1. Objectives

This research is centered on small business and microenterprise customers and has four related goals. First, it seeks to remedy the lack of a previous scheme at the selected telecom company to: (1) cross-sell IPTV to existing clients, and (2) acquire new IPTV customers. Second, it undertakes to forward a reliable list of prospects to adopt IPTV to the Marketing department by developing two models that will yield respectively: (1) the sales propensity index (SPI) associated to active clients, and (2) the SPI associated to potential customers. Third, it attempts to determine the factors that influence small enterprises to take on IPTV. And fourth, it aims to use as much as possible the independent data integration system or data mart, which was an ongoing

endeavor to concentrate relevant data from distinct sources by the IT and BI teams during all stages of this study. Once the data mart is fully completed, further models could readily be implemented.

1.2. Thesis Outline

The thesis is further organized as follows: In Chapter 2, existing literature related to CRM and business intelligence within the B2B market and the telecommunications sector is reviewed. Chapter 3 explains theoretical concepts relevant to business intelligence, in particular the tools to collect data from unstructured sources, and data mining techniques. Chapter 4 presents the data and the adopted methodology to develop and estimate the models. Chapter 5 provides the modeling analysis and evaluates results. In Chapter 6, the research is concluded, and managerial implications and directions for future research are discussed.

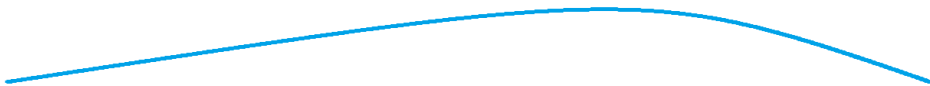
Chapter 2.

Literature Review

2.1. Customer Relationship Management

Customer *acquisition* has become a critical undertaking for both B2C and B2B organizations as marketing has transitioned from a *brand-* to a *customer-*based approach (Blattberg et al., 2008). It has been described as the first of the three customer lifecycle strategies within the customer relationship management (CRM) along with customer *development* and customer *retention* (Kamakura et al., 2005). It has also been defined as the first of four circulatory processes along with customer *retention*, customer *churn*, and customer *win-back* (Kumar & Petersen, 2012). These processes fit into the customer lifecycle management depicted in Figure 2.1 in terms of the objectives, the data, and the tactics the firm can take at each phase.

Figure 2.1 Customer Lifecycle Management
(Adapted from Course 6-105-13: Database Marketing Analysis)



	Acquisition	Onboard	Grow	Retain	Decline
Objectives	Customer acquisition and conversion	Customer engagement	Customer spend lift	Customer recognition & rewarding	Customer reactivation & win-back
Tactics	Limited data support	Potential value scoring & Cross-Partner activation	Potential value scoring & Cross-Selling classification models	Attrition Propensity scoring & Survival analysis models	Attrition propensity scoring & Survival analysis models
Data	Acquisition & conversion models	Attitudinal segments	Understanding share of wallet potential lift	Maintaining spend & reducing churn	Models to identify and target churn

From a marketing perspective, the acquisition process demands greater effort compared with those of growth or retention because it starts by considering the channels taken by the customers to first access the firm, and then elaborates on the promotions that bring them to purchase ([Kamakura et al., 2005](#)). Furthermore, the costs associated to acquisition are typically five times higher than those of retention ([Hung et al., 2006](#)). Nevertheless, empirical evidence suggests that firms can achieve an improved performance by allocating a greater proportion of resources for both acquisition and retention of larger or key, as opposed to smaller, customers.

At the acquisition stage, overlooking which customers would be likely to convert can lead to a waste of resources and potential sales loss. Therefore, a quantitative model, formulated on relevant information about customers, comes as a highly valuable tool to predict who of these customers are likely to adopt a product or service ([Monat, 2011](#)).

At the growth stage, current customers' value can be increased by one of the following three ways: (1) by increasing the use of products or services they already have; (2) by selling them more or higher-margin products or services; or (3) by keeping the customers for a longer period of time ([Rygielski et al., 2002](#)).

Because CRM is concerned with the activities providers undertake in an attempt to develop and maintain successful relation exchanges with their customers, a significant amount of effort has been devoted to the study of relationship marketing over the last two decades. As [Zablah et al. \(2012\)](#) highlight, CRM activities have significant managerial implications because (1) they affect the long-term performance in the marketplace, (2) they reveal the level of profitability by the firm's capacity to build the right type of relationship with the right type of customer, and (3) they indicate that a firm must be able to discriminate between current and prospective customers based on their expected level of long-term profitability.

From the previous discussion, it becomes clear that CRM is a broad topic that requires the support of BI, whose techniques allow companies to become more customer-oriented. BI has emerged over recent years as an extremely powerful approach in extracting meaningful customer information ([Lee & Siau, 2001](#)).

2.2. Acquisition and IPTV Adoption

IPTV is a media technology capable of receiving and retrieving a video stream encoded as a series of Internet protocol packets. Its versatility, which can provide high quality audio and video content, has been the most significant competitive advantage. IPTV service is provided based on high-speed networks and requires the deployment of a broadband convergence network and fiber optic architecture to allow the distribution of television and video signals over the Internet (Shin, 2007).

Fierce competition in the telecommunications sector, especially in the mobile phone, voice over Internet protocol (VoIP), and cable TV, has prompted telecom providers to look into new ways of gaining potential customers by offering the triple play of services, namely phone, data, and video, which cable companies already provide. In North America, IPTV faces a saturated market where cable and satellite television are very strong (Ortiz, 2006).

Since IPTV is relatively a new technology in the minds of customers, two main avenues have been developed in previous research at exploring the adoption of IPTV. Although these studies do not focus on the B2B environment, they provide important insights on the factors explaining the adoption of this service by residential customers. The first avenue seeks understanding through technology acceptance modeling (TAM), which is based on factor analysis from conducting surveys (Choi et al., 2010; Weniger, 2010; Jang & Noh, 2011), while the second avenue obtains results from logistic regression modeling (Atkin et al., 2003; Shin, 2007; Shin & Hwang, 2011).

In using TAM, Weniger (2010) proposes studying the driving forces of users' adoption of IPTV stressing the importance of its perceived qualities, namely content, system security, and interactivity. In using logistic regression, Atkin et al. (2003) found important factors within the demographic information affecting the adoption of IPTV, and introduced variables from surveyed subjects such as the level of sense of humor and quality of life. In using both logistic regression and diffusion models, which included the well-known Bass model, Shin & Hwang (2011) incorporated explanatory variables from surveyed customers such as cost, relative advantage, complexity, and compatibility to assess the factors affecting IPTV diffusion.

2.3. Growth and Cross-Selling

Cross-selling means approaching current customers and encouraging them to increase their engagement with the firm by purchasing one or many additional products or services. It is one of the main tools to strengthen the customer relationship (Kamakura et al., 1991). Understanding and using cross-selling techniques is crucially important for a company because as the customers acquire more products or services from the same provider, the associated switching cost for leaving with a competitor increases (Kamakura et al., 2003). Therefore, cross-selling is considered a strong driver for increasing the number of loyal customers, obtaining higher customer lifetime value, and lowering the customer churn (Akura & Srinivasan, 2005).

Other studies focus on modeling the probability of a successful cross-sale attempt in the financial industry. In an early study by Kamakura et al. (1991) probabilistic predictions are made on whether or not a customer would purchase a particular product based on being the owner of other products. Knott et al. (2002) apply logistic regression and neural networks to predict which product a customer is expected to buy next. The approach is further developed in Li et al. (2005), where also the appropriate time to approach a specific customer is studied by applying a logistic regression and Markov chain Monte Carlo approach. These studies have mainly been made on cross-selling as a method for increasing a company's revenue within the banking services.

2.4. The Telecommunications Sector

Within the subscription services, the telecommunications industry is considered one of the top sectors suffering from customer churn (Jahromi, 2010). At the same time, customers have a long-term relationship with their service provider, and data on their characteristics and their transactional behaviour are stored in the internal databases. For this reason, most of the CRM research in this industry has been focused on retention of individual consumers, i.e. in a B2C setting. Furthermore, as the telecommunications market is often saturated, the pool of available customers is limited, and a service provider has to shift from its acquisition strategy to a retention action plan.

Consequently, academic research using data from telecom providers has focused on retention and churn prediction, particularly for the mobile service (Hung et al., 2006).

Rygielski et al. (2002) highlight that growing competition forces telecommunications providers around the globe to market special pricing programs in order to retain existing customers and attract new ones. As some telecom firms favour untargeted approaches, which rely on mass advertising, many others accumulate detailed records from their customers allowing them to discover similar patterns and thus target clear segments with attractive pricing and feature programs.

Tsai et al. (2010) and Jahromi et al. (2010) have summarized and compared previous work related to modeling customer churn within the telecommunications industry under a B2C framework, which ranges from logistic regression (Kim & Yoon, 2004; Ahn et al., 2006; Seo et al., 2008), classification trees & random forests (Wei & Chiu, 2002; Burez & Van den Poel, 2009), and neural networks (Mozer et al., 2000; Hung et al., 2006; Luo et al., 2007; Tsai & Lu, 2009); to genetic algorithms (Pendharkar, 2009), Markov chains (Burez & van den Poel, 2007) and survival analysis (Burez & Van den Poel, 2008). All of the above stress the importance of factors related to consumers' demographics, purchase behaviour, and satisfaction within the models estimated.

Nonetheless, the mobile telecom market has also been the subject of academic research with respect to growth and acquisition. Cross-selling value-added services (VAS), which are new services to generate more average revenue per user (ARPU), allow firms to diversify their business areas and expand their revenues and profits. Examples of mobile VAS are mobile communication, entertainment, transaction, and information services. Thus, telecom operators can also utilize CRM to find the appropriate prospects for using mobile VAS by understanding their customers' life styles. In Ahn et al. (2011), the use of genetic algorithms to combine several prediction models' outcomes leads to a classification model for cross-selling mobile VAS. They develop three models: a logistic regression, a decision tree, and an artificial neural network, providing weighted averages that were later combined and optimized by a genetic algorithm.

2.5. The B2B Context

Laplaca & Katrichis (2009) compiled academic research addressing marketing on a B2B framework and highlighted the fact that interest in this field, as evidence by the number of publications in specialized journals, does not reflect the relative economic importance of B2B activities. Specifically to the B2B marketing environment, business customers are fewer in number (Jahromi et al., 2014). Moreover, it is much more difficult than in a B2C context to acquire information about industrial customers, both from an internal customer base and from external data sources (Hague & Harrison, 2013).

Debate continues about the best strategies to build an effective B2B database marketing process. However, some essays suggest specific elements that would improve the profiling of business users based on industrial information (Coe, 2001). In order to target leads, it is essential to analyze the current business and market share. This involves acquiring knowledge about the market to identify new or emerging opportunities by carefully researching small niches to narrow and penetrate segments with the highest revenue-generating customer base. Despite the existence of commercially available corporate databases, which provide copious prospects, crucial data are frequently missing. Usually this is because industrial customers have complex buying decision units and processes, and they withhold vital information regarding their buying patterns, decision-making style, and key decision makers. Acquiring the missing information manually would be a time-consuming solution that is bound to add significant costs of acquisition. Therefore, the key question becomes how to feasibly manage the process of acquiring customers under conditions of partial information.

Coe (2004) suggests that characteristics pertaining to *demographics*, or *firmographics* in the B2B context, which serve to organize leads include: company size by number of employees; industry type by the Standard Industrial Classification (SIC) code; company location; company organization (headquarters, branch, affiliate or division); and business age. At the other end, behavioural aspects, which serve to segment current customers include 1) transactional data, 2) response actions, and 3) customer needs. First, transactional data is linked directly to the purchase information, namely the date the transaction took place, the quantity purchased, and the amount spent. Second, response actions track any communications going to and from current

customers, namely any calls received and returned as well as any offers or promotions sent to the client. And third, customer needs represent the most interesting and difficult aspect as it undertakes determining the needs of the business client in terms of the product or service it offers. Since it is not clear who within the company defines its needs, or when these needs change over time, it is essential to conduct regular surveys within the targeted segments to gather important customer information. Since an additional effort in time and cost would be required to collect these data, not all organizations are qualified to undergo regular surveys. Therefore, under austere conditions, behavioural data within a B2B context can be assumed as being “complete” once both transactional data and response actions are part of the customer base.

Although limited research exists in the area of customer acquisition and customer growth under a B2B environment, it has been recently the object of further development with a business intelligence perspective. At the acquisition level, [D’Haen & Van den Poel \(2013\)](#) propose a model integrated to a web application that could run without the need of human interference and provide sales representatives with a list of the leads most likely to respond. The model applies three different modeling methods, namely logistic regression, decision trees, and neural networks. At the growth level, the first research in modeling the factors affecting customers’ satisfaction in a B2B setting go back to [Patterson et al. \(1997\)](#). They argue that the five variables – novelty, purchase decision, decision complexity, stakeholding, and uncertainty – are key factors contributing to the satisfaction of clients. Their conclusions result from modeling pre-purchase buying decision in industrial markets by employing a two-stage longitudinal study of a range of business professional services. Perhaps the most important of their conclusions is that customer satisfaction is strongly and positively associated with repeat purchase intentions in the business professional services.

As seen, previous literature in the B2B context focuses on acquisition and growth separately. The customer’s decision to cross-buy is considered as being a retention strategy. [Bolton et al. \(2008\)](#) perceive it as the current client’s decision to upgrade a service contract with the supplier. A service upgrade is a way of expanding the relationship in which the customer purchases an expanded offering – a higher price, augmented service – instead of repurchasing a low-price service from another supplier.

[Bolton et al. \(2008\)](#) further propose a model where the business client's decision to upgrade is influenced by the decision maker's perceptions of the supplier (relationship or account level variables), the service at a contract level, and the interactions between these two.

The factors affecting cross-buying from the customer's perspective, either positively or negatively, are thus reduced to four basic areas and their interactions: (1) price, (2) service quality, (3) criticality, and (4) satisfaction.

First, price effects are considered in almost every paper that has been written on retention. The concept of dual entitlement suggests that price increases commensurate with cost increases, and these are perceived as fair ([Bolton & Alba, 2006](#)). Analogously, in the cross-buying context, an upgrade or additional service represents more benefits to the business client at a higher cost to the supplier, resulting in a higher price for that extra service at the client's end. The suggested prediction is that cross-buying an additional service is positively related to the current price of the service contract.

Second, when a decision maker experiences poor Quality of Service (QoS), anticipated regret is likely to cause forgone alternatives. Moreover, suppliers delivering consistently a bad service when competition exists and switching costs are low, are forced to exit the marketplace. Attribution theory suggests that people are likely to make causal attributions based on their prior service experiences ([Folkes, 1988](#)). Here, it is important to differentiate between service quality as a set of both customer service and technological issues, i.e. unsuccessful call ratios, supply times for initial connection, response times for operation services, and bill correctness; and the quality of the specific service included in the service agreement. The latter being the main component for the purpose of the present study. In telecommunications, the most important dimensions of quality are availability, reliability, protection, security, and flexibility ([Zhao et al., 2000](#)). On the Internet, QoS is the idea that transmission rates, error rates, and other characteristics can be measured, improved, and, to some extent, guaranteed in advance. In summary, previous research suggest that quality of the service provided has a positive relationship with customer retention.

Third, a small business' likelihood of cross-buying a service is positively related to the decision maker's perception on the criticality of such service to the owned business' operation ([Bolton et al., 2008](#)).

And fourth, customer satisfaction influences the decision to cross-buy in terms of the increase of the business client's commitment to the existing relationship with the supplier ([Narayandas et al., 2004](#)).

In essence, at the contract level, it has been proposed that business clients are influenced by service quality and price. While at the supplier level, business clients are influenced by the decision-maker's perceptions of the buyer-seller relationship – assessments of satisfaction – and criticality, according to its industry-related needs.

Fundamentally, the microenterprise and the telecommunications supplier are engaging in dynamic adaptation ([Dickson, 1992](#)). While business clients engage in a problem-solving process to find the appropriate level of service from the supplier that best meets their needs, service providers over represent the positive features of the service they supply. Consistent with learning theory, this suggests that firms and consumers either in a B2C or B2B context engage in trial-and-error behaviour over time that results in learning.

Another important contribution in the B2B context is that of [Hansen et al. \(2008\)](#), who introduced the corporate reputation, information sharing, distributive fairness, and flexibility to the factors influencing the customer perceived value (CPV). They confirmed that CPV has a key role and serves as a mediator of business activities on relationship outcomes, proving that word of mouth is an indispensable source of credible information for potential customers.

In summary, the current research builds on previous literature to develop one model of service cross-selling and one model of service acquisition under a B2B environment. As previous research has established the importance of considering both acquisition and growth strategies simultaneously, the model is built at a customer level, allowing to explore the underlying relationship between these two phases.

Chapter 3.

Conceptual Framework

Business Intelligence (BI) is a set of concepts, methods, and technologies designed to pursue the goal of turning all the widely separated data in an organization into useful information and knowledge (Hancock, 2006). It is sought to provide decision support and is built over database tools, including data warehousing (DW) and data mining (DM). The functionalities of these tools are complementary and interrelated. At one end, DW refers to the automation of business process to provide easy access to data with a design and approach that is optimized for queries rather than specific transactions; while at the other end, DM being the essential layer of CRM, refers to the analysis of data using a number of techniques.

This chapter is divided into two main sections, each of which presents the theoretical concepts related to BI. The first section describes DW, which describes data warehouses and data marts as well as the tool to access their information. The second section presents DM as a collection of both tasks and techniques, which covers the statistical modeling method used in this study, namely logistic regression.

3.1. Data Warehousing

An enterprise data warehouse is a critical component of a successful CRM strategy (Lee & Siau, 2001). A data warehouse is defined as a subject-oriented, integrated, consistent, non-volatile data repository that provides for the efficient storage, maintenance, and retrieval of historical data showing data evolution over time to support decision-making processes (Golfarelli & Rizzi, 2009).

Data warehouse systems grew in 1990s from the need to retrieve useful information from a huge amount of data accumulated on heterogeneous platforms. The data was made available to decision-makers by formulating queries, allowing them to conduct analyses on relevant information without blocking operational systems. Before their existence, the process of quantifying and evaluating each department's contribution to the global business performance in a large corporation, for example, involved tailor-made queries, retrieving the required data from all the corporate databases, processing calculations to finally produce a report. This approach was demanding in terms of time and resources and usually did not achieve the desired results.

Because data warehouses take advantage of multiple data sources, such as data extracted from production and then stored to enterprise databases, or data from a third party's information systems, it is essential to emphasize its integration and consistency.

Operational data usually cover a short period of time, because most transactions involve the latest data. On the contrary, a data warehouse should enable analyses that cover a few years. For this reason, data warehouses are regularly updated from operational data, which is never deleted, when they are offline. Data warehouses can be essentially viewed as read-only databases to satisfy the users' need for a short analysis query response time.

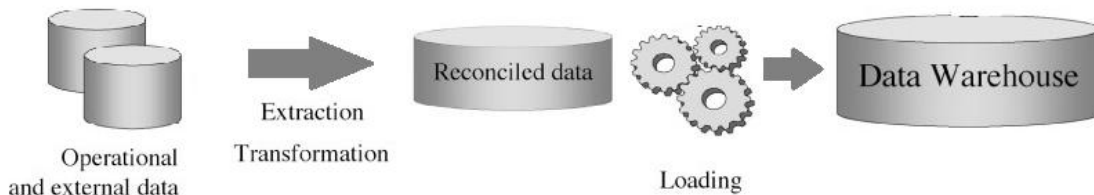
The following is a list of essential requirements for a data warehouse process, i.e. a set of tasks that allow turning operational data into decision-making support information:

- **Accessibility** to users
- **Correctness** and **completeness** with neither lack of information nor duplicity
- **Accuracy** of data as stored values are compliant with real-world values
- **Conciseness** and **consistency** of data as its representation is uniform
- **Flexibility** and **traceability** as data can easily be traced back to its sources
- **Multidimensional** representation for intuitive view of information
- **Freshness** as data is regularly updated

One additional quality feature would refer to the creation of data dictionaries in order to document any type of conflict. Data dictionaries can be created where these relationships can be stored (Golfarelli & Rizzi, 2009).

The process necessary to move data from operational systems, filter it, and load it into the data warehouse or data mart is called *Extraction-Transformation-Loading* (ETL) as depicted in Figure 3.1. A strict quality standard must be ensured from the first phases of the data warehouse project. In general, the quality of a process stands for the way a process meets users' goals. In data warehousing, quality is not only useful for the level of data, but above all for the whole integrated system.

Figure 3.1 ETL Process (Golfarelli & Rizzi, 2009)



Most of the times, data are required for a specific application domain or a specific department within the enterprise, in which case data must be summed up as much as possible into a “smaller data warehouse” called data mart. The architectural difference between a data warehouse and a data mart need to be studied closer.

3.1.1. Data Marts

A data mart is a subset or an aggregation of data stored to a primary data warehouse. It can be seen the simplest form of a data warehouse focusing on a single subject or functional area or company department. Given their single-subject focus, data marts usually draw data from only a few sources, namely internal operational systems, a central data warehouse, or external data.

While data warehouses store and retrieve clean, consistent data effectively from multiple source systems, a data mart is a specialized system often built and controlled by a single department that brings together the data needed for its related applications.

Although the size of a data mart or the decision-support data that it contains can be elaborated, data marts are typically smaller and less complex than data warehouses. Therefore, they are typically easier to build and maintain. Figure 3.2 summarizes the basic differences between a data warehouse and a data mart.

Figure 3.2 Difference between Data Warehouse and Data Mart (Talus, 1998)



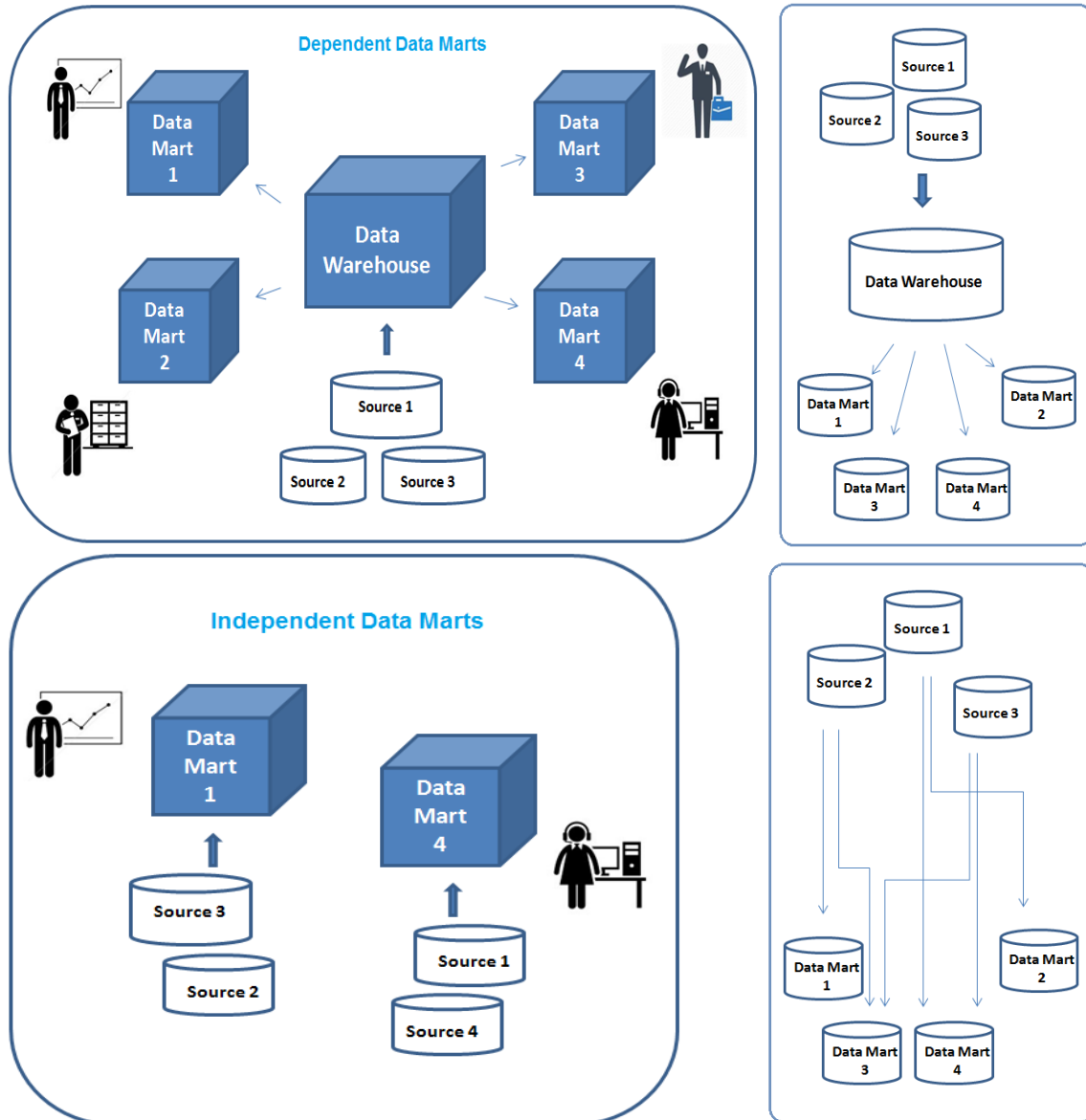
Data marts are often associated with reporting systems and slicing-and-dicing pre-summarized data.

The main difference between independent and dependent data marts is how the data mart is populated; that is, how you get data out of the sources and into the data mart, either from the operational data sources or from the data warehouse, as can be seen in Figure 3.3.

With dependent data marts, this process is somewhat simplified because formatted and summarized – clean – data have already been loaded into the central data warehouse. The ETL process for dependent data marts is mostly a process of identifying the right subset of data relevant to the chosen data mart subject and moving a copy of it, usually in a summarized form.

With independent data marts, one must deal with all aspects of the ETL process, much as done with a central data warehouse. However, the number of sources are expected to be fewer so that the amount of data associated with the data mart is less than the data warehouse, given the focus on a single subject.

Figure 3.3 Dependent vs. Independent Data Marts
 (Adapted from Golfarelli & Rizzi, 2009)



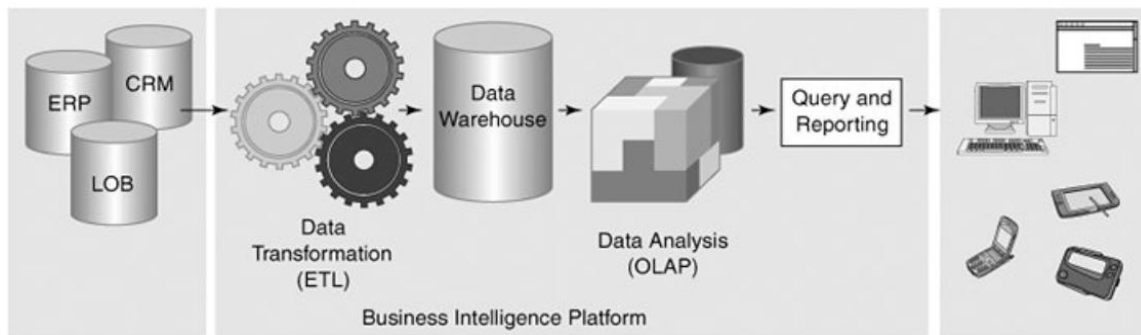
The motivations behind the creation of these two types of data marts are also typically different. Dependent data marts are usually built to achieve improved performance and availability, better control, and lower telecommunication costs resulting from local access of data relevant to a specific department. The creation of independent data marts is often driven by the need to have a solution within a shorter time.

3.1.2. Accessing Data – OLAP

Analysis is the last level common to all data warehouse architecture types. After cleansing, integrating, and transforming data, it is essential to determine how to get the information. End users query data warehouses or data marts to use the stored information as a starting point for additional business intelligence applications, such as what-if analyses and data mining, and to create reports and dashboards.

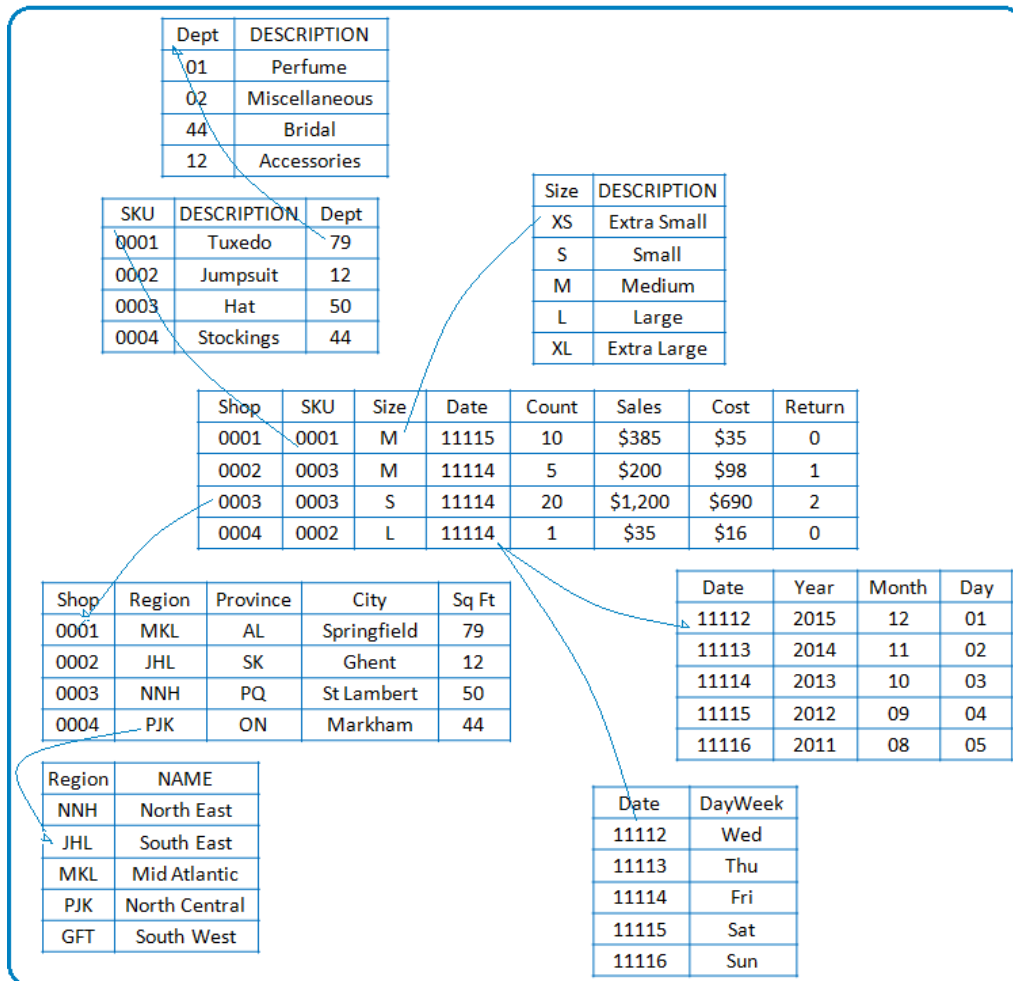
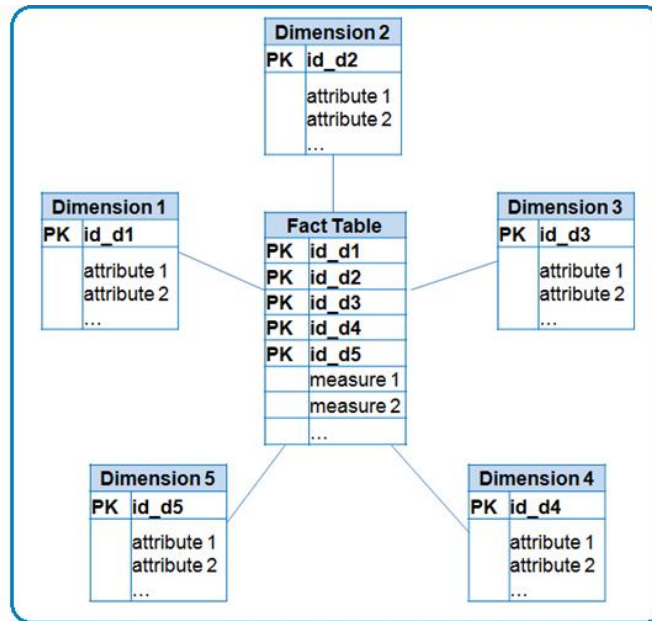
On-line analytical processing (OLAP) is a service providing quick answers to *ad hoc* queries against a data warehouse (Linoff & Berry, 2011). Figure 3.4 shows the business intelligence platform.

Figure 3.4 Business Intelligence Platform (Hancock, 2006)



The term on-line analytic processing refers to technology that allows users of multidimensional databases to generate views, which are descriptive or comparative summaries of data and other analytic queries. Despite its name, analyses referred as OLAP do not need to be performed on-line or in real-time. The term applies to analyses on multidimensional databases or a “cube” that may contain dynamically updated information from queries referencing various types of data. An OLAP cube can be integrated into corporate database systems to allow analysts and managers to monitor the performance of the business or the market. The final result of an OLAP cube can be very simple and in the form of frequency tables, descriptive statistics, cross-tabulations, removal of outliers, and other forms of cleaning the data. OLAP is mainly performed by the well-known star schema. Figure 3.5 depicts a star schema design and database structure, which is comprised of two main components, the fact and the dimensions tables.

Figure 3.5 Star Schema, Design (Benhima et al., 2013) and Structure (Linoff & Berry, 2011)



OLAP enables users to easily and selectively extract data from different points of view. These *views* are derived from tables, and they store aggregate data used for typical OLAP queries, thus minimizing execution costs related to joining operations between large tables.

3.2. Data Mining

Data mining has a fundamental and critical role to play in CRM as it enables the transformation of customer data into useful information and knowledge (Rygielski et al., 2002). Data mining is about the successful exploitation of large amount of data that uses a number of statistical algorithms to find patterns and correlations in the data and report models back to the user (Linoff & Berry, 2011; Lee & Siau, 2001). As opposed to traditional hypothesis testing, which is designed to verify *a priori* hypotheses about relations between variables, exploratory data analysis in data mining is used to identify systematic relations between variables when *a priori* expectations regarding the nature of those relations exist. In a typical data mining process, many variables are taken into account and compared, using a variety of techniques in the search for systematic patterns. All of which help identify valuable customers, predict future behaviour, and make proactive and knowledge-based decisions (Rygielski et al., 2002).

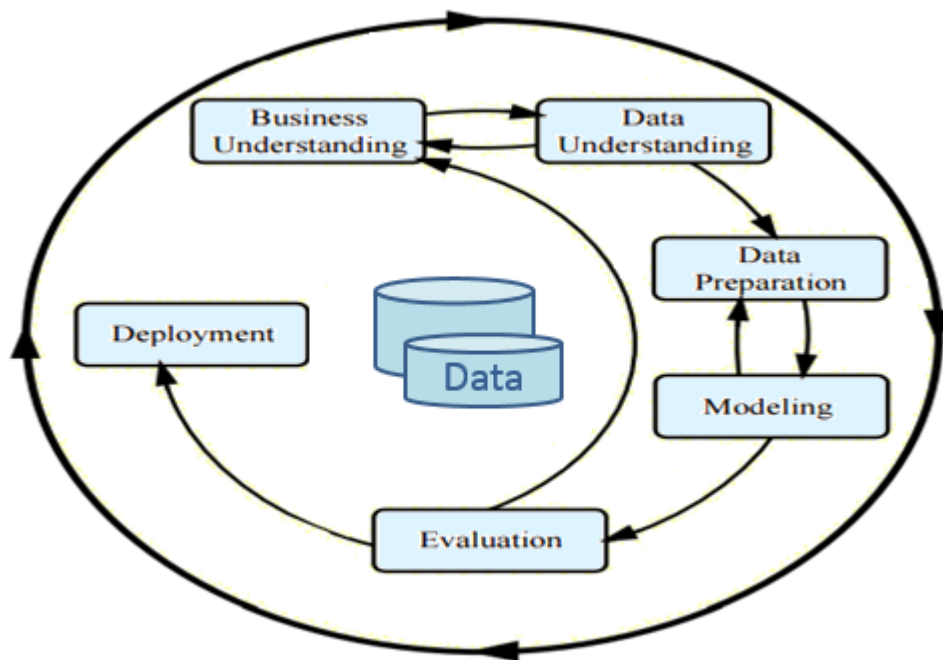
The initial value of a data warehouse or data mart comes from automating existing processes, such as reporting. But the biggest returns are the improved access to data that can spur innovation and creativity. Although data mining techniques can operate on any kind of unprocessed or even unstructured information, they can also be applied to the data views and summaries generated in a warehouse environment to provide in-depth and multidimensional knowledge. In other words, data mining techniques are considered to represent either a different analytic approach, or as an analytic extension of OLAP.

3.2.1. Data Mining Tasks

In the business environment, complex data mining projects may require the coordinated efforts of various experts or departments throughout an entire organization.

In the data mining literature, various general frameworks have been proposed to serve as blueprints for how to organize the process of gathering data, analyzing data, disseminating results, implementing results, and monitoring improvements. One such model, Cross-Industry Standard Process (CRISP) for data mining was proposed in the mid-1990s by a European consortium of companies to serve as a non-proprietary standard process model for data mining. This approach postulates the sequence of steps shown in Figure 3.6.

Figure 3.6 Phases of the CRISP-DM Process for Data Mining (Adapted from Wirth, 2000)



Another framework of this kind is the approach proposed by SAS Institute called SEMMA, which focuses more on the technical activities typically involved in a data mining project.

Sample → Explore → Modify → Model → Assess

This model is concerned with the process of how to integrate data mining methodology into an organization, how to convert data into information, how to involve important stakeholders, and how to disseminate the information in a form that can easily be converted into resources for strategic decision making. The term Predictive Data

Mining is usually applied to identify data mining projects with the goal to identify a statistical model that can be used to predict some response of interest. For example, a credit card company may want to engage in predictive data mining, to derive a (trained) model or set of models that can quickly identify transactions which have a high probability of being fraudulent. Other types of data mining projects may be more exploratory in nature, in which case drill-down descriptive and exploratory methods would be applied. Data reduction is another possible objective for data mining, e.g. to aggregate or amalgamate the information in very large data sets into useful and manageable parts.

Lee & Siau (2001) highlight the importance of statistics as an indispensable component in data selection, sampling, and evaluation of extracted knowledge within data mining. Statistics also deals with the data cleaning process by offering the techniques to detect outliers, to smooth data, and to choose imputation methods due to missing data. Techniques in clustering and designing of experiments, as well as descriptive and predictive modeling processes come all from the statistics sphere to assist the data mining massive data latitude.

The choice of a statistical model for generating cross-buying and acquisition predictions relies on the dichotomous feature of the dependent variable (Stevens, 2009), which could include: logistic regression, neural networks and classification trees. However, there are a few practical considerations that might determine the choice of model *a priori*. Logistic regression is particularly easy to implement, widely available, applicable when only one service or product will be purchased next, and can become a straightforward application to score new data once the model is fitted (Knott et al., 2002).

3.2.2. Logistic Regression

Following a long tradition of theoretical and empirical work to predict customer acquisition behaviour (Bose & Chen, 2009; Gupta et al., 2006), the customer's decision to adopt the IPTV service will be represented by a binary choice model, i.e. logistic regression. Logistic regression is a model used for prediction of the probability of occurrence of an event by fitting data to a logistic curve. It makes use of several predictor variables, which may be either numerical or categorical. Logistic regression is

used extensively in marketing applications such as predicting customers' propensity to purchase or cease a subscription (Jahromi et al., 2014; Knott et al., 2002; Hansotia & Wang, 1997).

As explained in Larocque (2013), prospects' probability of response is modeled in logistic regression as a conditional probability p of the dependent variable Y taking the value of 1 given X ,

$$p = E(Y = 1|X) = (P(Y = 1|X))$$

where X is a set of n variables X_1, X_2, \dots, X_n affecting the acquisition of a customer,

$$p = P(Y = 1|X_1, X_2, \dots, X_n)$$

The logistic regression model is a particular case of the generalized linear models, and as such, p can be re-expressed in terms of a link function. Therefore, by letting p be a linear function of X , the possible probability values $[0,1]$ can be 'linked' to the possible linear values $[-\infty, \infty]$. A common choice is to transform p into a logistic (or *logit*) function given by

$$\text{Logit}(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n$$

or

$$\frac{p}{1-p} = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n}$$

and

$$p = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n)}}$$

which represents the probability of acquiring a customer or the sales probability index (SPI) given the variables X_1, X_2, \dots, X_n .

Once the function is estimated, it can be used to predict future values of the dependent variable, having gathered the values of the independent variables. However, as stated in Allison (2012), the best measure to interpret the outcomes of the logistic regression is the odds ratio, which is the number of times success occurred compared to the number of times failure occurred.

$$\frac{p}{1-p} = \frac{P(Y = 1|X)}{P(Y = 0|X)} = \frac{P(Y = 1|X)}{1 - P(Y = 1|X)}$$

In order to calculate the odds ratio for a particular independent variable, for example X_1 ,

$$\frac{P(Y = 1|X_1)}{1 - P(Y = 1|X_1)} = e^{\beta_0 + \beta_1 X_1} = e^{\beta_0} e^{\beta_1 X_1}$$

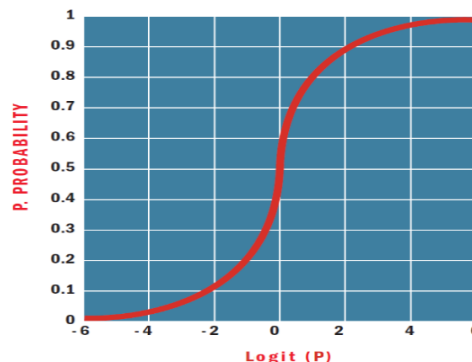
and when X_1 increases by 1,

$$\frac{P(Y=1|X_1+1)}{1-P(Y=1|X_1+1)} = e^{\beta_0} e^{\beta_1(X_1+1)} = e^{\beta_0} e^{\beta_1 X_1} \underbrace{e^{\beta_1}}$$

the odds ratio is multiplied by e^{β_1} . This is valid for all X_n .

The *logit* transformation is illustrated in Figure 3.7. As seen, if the probability is 0.5, the odds are even and the *logit* is zero. Negative *logits* represent probabilities below 0.5, and positive *logits* correspond to probabilities above 0.5.

Figure 3.7 The Logistic Function (Monat, 2011)



The central matter in a logistic regression analysis is the determination of the coefficients (βs) as the best estimates under a significance criteria. This is accomplished by analyzing historical data that is associated to the independent or explanatory variables to the dependent or probability of conversion variable.

At all points, the emphasis is on interpreting the values obtained for the coefficients and the odds ratios.

Chapter 4.

Modeling Process

In Chapter 3, data warehousing concepts and data mining techniques relevant to the present research were reviewed. Chapter 4 presents the method to design the models and the data, including the process to access the available databases and create the analysis data sets.

In this section, the pursued methodology occurs as a natural path following the former concepts, namely the use of logistic regression as a data mining tool to obtain the probability of a customer to adopt the IPTV service. The data preparation and statistical analyses were conducted within a SAS, SQL and Teradata environment.

4.1. Methodology

As stated in Chapter 2, the research focuses on both (a) the business growth and (b) the customer acquisition. In order to achieve (a), this study proposes Model A, which will address the customer development process by cross-selling IPTV within the current business phone and Internet database. In order to achieve (b), the study further proposes Model B, which will address the customer acquisition process by seeking new leads from a commercially available database.

Both models must provide a sales propensity index (SPI), yielding a list of prospects grouped by deciles and sorted in descending order. Table 4.1 presents the basic modeling difference in variable selection.

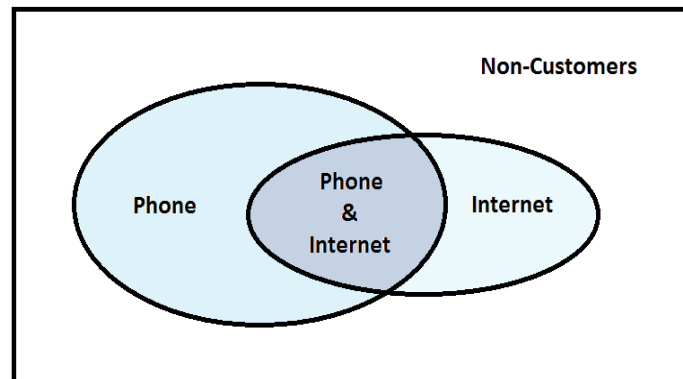
Table 4.1 Growth and Acquisition Modeling Overview

	Model A Cross Selling to Clients	Model B Acquisition of New Clients
Internal Data Behaviour Variables	Both consumption and response to direct mailing are available.	First time client acquisition is not based on consumption.
External Data Firmographics	Available information describing demographics of business clients – B2B format.	Available information describing demographics of business clients – B2B format.

In the proposed methodology, only those customers who did not initially subscribe to all three services (phone, Internet and IPTV) in a bundle at the moment of subscribing with the firm were examined for Model A.

All active customers defined as small or micro businesses were considered for analysis. These customers were subscribed to a) wired phone, b) Internet or c) wired phone and Internet. Figure 4.1 displays the current customer base within the market.

Figure 4.1 Small Business and Microenterprise Market



Ease of use and robustness of results have made logistic regression a popular binary predictor among marketing academics as well as the first choice for customer cross-selling and acquisition modeling. Therefore, in this research, logistic regression is used as the sole technique against which the performance of the marketing response rate will be compared.

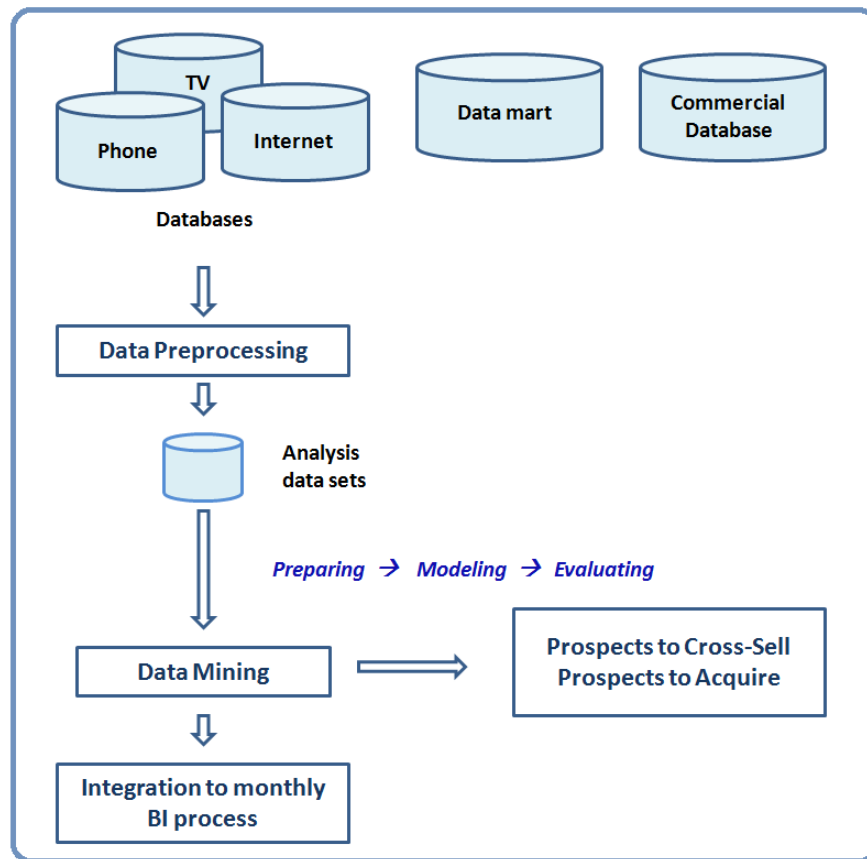
The target variable is a binary variable that takes the values “0” or “1”. As an example, it could take values for the month of June 2014 as follows:

“1”: The customer subscribes to IPTV during the month of June 2014.

“0”: The customer does not have an active IPTV service at the end of June 2014.

Once the analysis data sets are created, the stages involved in developing each one of the models will follow the CRISP process described in Chapter 3.

Figure 4.2. Data Process Flow for Prediction Modeling



Stage 1: Preparing – Exploring, identifying the most relevant variables for regression model via correlation and chi-square testing, transforming and selecting variables.

Stage 2: Modeling – Model building, estimation and validation. Considering various models and choosing the best one, based on their predicting performance, i.e. explained variability of target variable and stability of results.

Stage 3: Evaluating – Model scoring and deployment. Use the model selected as best in the previous stage and applying it to new data in order to generate predictions or estimates of the expected outcome.

Figure 4.2 presents the proposed data flow across the system to implement the predictive models. Both operational and data mart bases have been put together in the initial data sets. These would be treated to produce a single analysis data set with condensed information suitable for the modeling phase, also referred to as the calibration data set.

4.1.1. Data

Access to operational and data mart libraries related to the subscription of micro business clients was provided by a major telecommunications firm in Canada. The raw data used to build the cross-selling predictive models resulted from the preprocessing and cleaning process to create two data sets: the first over a period of six months (from March to August 2014) and the second over a period of ten months (from March to December 2014). The raw data used to build the acquisition predictive model resulted from the preprocessing and cleaning process to create only one data set over the period of ten months (from March to December 2014). This data set will be described in §4.3.1.

Two available data sources were identified: a) the firm's internal databases, containing behavioural information about current clients, and b) an external, commercially available database with corporate data of businesses located in Canada.

Each internal service-related database has its own client identifier. As a result, any client subscribed to more than one service will have exactly as many identifiers as to the number of services he is subscribed. Consequently, it is essential to detect this kind of customers and assign them a single identification number. This consolidation process is accomplished by matching all clients with the prospects listed in the external database

by their company name and their business address. Thus, in addition to creating a single observation per client with the internal information, the firmographic details supplied by the external database are also collected. The final and main identifier is the primary key generated by the data mart, which takes the external or third party identification number as a single identifier of a company.

4.1.2. Data Preprocessing

The present section pertains to the gathering of the internal data in order to create the analysis data sets for Models A and B. This was achieved by collecting the total number of customers subscribed to phone and Internet through the data mart and some operational databases.

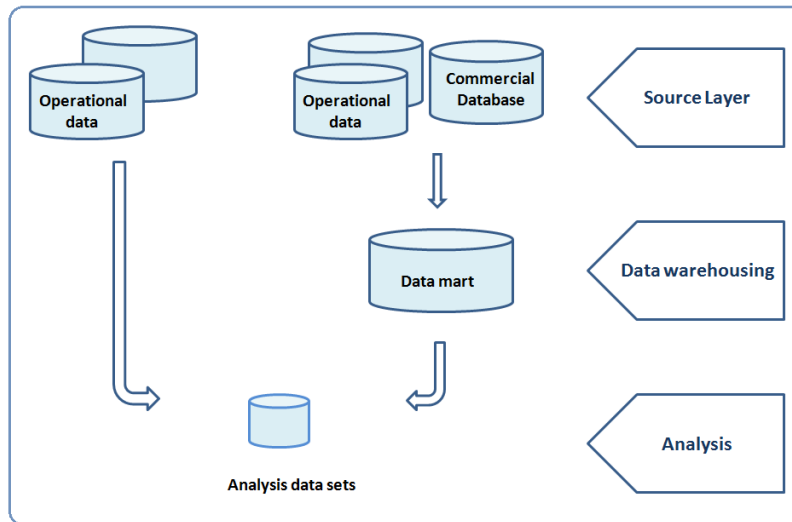
The first step required familiarizing with the available information sources. As data from a number of operational databases were not yet incorporated into the data mart, a multidimensional approach was not yet possible. As a result, obtaining the required data was performed with regular SQL queries, by creating tables from several joined databases.

These table joining process was possible by means of three identifiers: 1) each service legacy account identification (one for phone, one for Internet) that combined will be referenced as legacy account (LA) identification; 2) the data mart primary key (PK) identification; and 3) the commercially available database identification, which will be called Third Party (TP) identification, to keep its anonymity.

- 1) LA – Combination of phone and Internet account numbers
- 2) PK – Primary key from data mart
- 3) TP – Third party identification number

The disposition of the available data is depicted in Figure 4.3. By creating tables referenced with a combination of the three above mentioned identifiers, it was possible to collect different variables associated with clients.

Figure 4.3 Arrangement of Available Data Sources



The available databases concerned wired phone, Internet, satellite TV and IPTV. Matching of variables from internal and external data sources allowed obtaining the maximum percentage of information from all client firms. For each active service, i.e. phone and Internet, the following databases were accessed:

- Orders
- Activations
- Billing
- Customer Service
- Marketing Communications

The main databases are presented with the following, proposed notation: M_name for an operational database, and V_name for a view table from data mart as shown on Table 4.2.

As explained in §4.1.1, the legacy account identifiers associated with each service were not consolidated. Therefore, a single customer may have a different phone, Internet and TV service identifier within each database. The data mart was created by matching the client's internal identifiers with the external third-party data by its business name and address.

Table 4.2 Database Structure

	Database
Operational	M_phone_activation
	M_internet_activation
	M_phone_billing
	M_internet_billing
	M_iptv_orders
	M_satellite_orders
	M_consumption
	M_customer_service
Data Mart	V_phone_accounts
	V_internet_accounts
	V_profile_codes
	V_service_codes
	V_promo_tool
	V_external_database

After the data mart matching process, it was found that a single PK identifier could have one or more associated phone and Internet account identifiers. Analogously, a single phone and Internet account identifiers could have one or more associated PK identifiers. This resulted in having duplicate clients and represented an issue to be solved at a further stage of the data mart implementation.

For the purpose of the present research, a solution to remove duplicate clients was devised through the arrangement of observations by the mix of current accounts for one single client, as it will be explained in § 4.2.1.2.

Understanding the limitations of the data due to lack of defined and consistent coding and discrepancy in data entry is essential. For example, there were five variables pertaining to the location or region coming from three different operational databases plus the external database. These variables have different names and different values for a unique client. In order to obtain one and only one variable with the location of the client, a decision was taken to choose the variable with the lowest number of missing values and then impute the missing values left with the values found within the other three variables. If at the end of this process any missing values still remained, they were imputed based on the postal code.

In order to provide an initial model with the first six months of establishing the data mart, Model A0 was proposed. Four months later, during which major modifications to the data mart were applied, a further Model A was developed with the resulting data set on ten months.

It is worth noting the following four data features of the present study. First, the data mart was under development and was not yet completed by the initial six-month modeling process. It was an ongoing process that challenged the task of collecting the required variables due to their changing position or removal from referenced tables on a weekly basis. For this reason, approximately 30% of the complete customer base was used for the six-month data model. Second, during the following four months the data mart was further developed and was set to provide a broader customer base for the ten-month data model. Third, direct mail promotions were sent to clients throughout the two modeling intervals. And fourth, a number of customers discarded and re-subscribed to the Internet service, only that this time by adding the IPTV service and thus generating a different service identifier. Provided that an Internet connection is a necessary requirement to acquire IPTV, attention is restricted to the initial acquisition spell in the modeling process. Therefore, the results for the model addressing growth may not accurately reflect the behaviour of all customers but only the IPTV subscribers who did not sign up to IPTV at the beginning of their tenure as a customer, i.e. new subscribers to the service trio (phone, Internet and IPTV) or duo (Internet and IPTV). Nevertheless, the modeling environment could be enlarged to examine these groups in more detail once the data mart is completed, becoming an area of future research.

Table 4.3 depicts the time frames for both models A0 and A. Essentially, the second ten-month period made less use of the operational data sources, as more variables were incorporated into the data mart views.

Table 4.3 Two Analyzing Periods for Cross-Selling Model

Model A0	Model A
<i>Six-month data</i>	<i>Ten-month data</i>
<i>March to August, 2014</i>	<i>March to December, 2014</i>
<i>30% of client database</i>	<i>Total client database</i>

The resulting preprocessed and unbalanced six-month data set was comprised of **33,161** observations, from which 129 or 0.4% had IPTV (CrossSell = 1) and 33,032 or 99.6% did not have IPTV (CrossSell = 0).

The resulting preprocessed and unbalanced ten-month data set was comprised of **110,756** observations, from which 1260 or 1.14% had IPTV (CrossSell = 1), and 109,496 or 98.86% did not have IPTV (CrossSell = 0). A smaller data set was reserved for modeling acquisition with Model B. This data set was comprised of **6,576** observations, from which 76 or 1.17% had IPTV (Acquired = 1), and 6,500 or 98.84% did not have IPTV (Acquired = 0).

4.1.3. Data Cleaning

The first step in cleaning the data was to check and reject all outliers and extreme values on continuous variables; for example, extreme negative values for dollar amounts. Cleaning inconsistent data resulting from matching customers from operational databases was performed as follows. Given the ratio of IPTV converted clients on the total base was approximately 1.14%, a good way to balance the analysis data set was to remove all observations with the target variable set to '0', which presented a large number of missing values on external variables related to demographic information. In other words, all variables from the phone and Internet sources with missing values for non-converted clients were removed. In order to prevent the modeling results from becoming biased as a result of proceeding in this way, all clients who converted during each one of the observed months and thus having the target variable set to '1', were verified to confirm they did not show a high proportion of missing values. In the event these observations presented a missing value, they were imputed following the procedure as explained in the following paragraph. Nevertheless, it is important to note the risk of bias due to imputing missing values, which was performed on observations which ha the target variable set to either '0' or '1'.

Imputation was performed on behavioural binary variables of the form 'Y' or 'N' where the value was obvious; for example, an extra option on a current service that was missing would be imputed as 'N'. Similarly, missing region codes were collected from the available databases to consolidate the information coming from different sources.

Thus, the 33,161 preprocessed observations from the six-month period were reduced for cross selling to 25,131 observations. This data set was composed of 129 (0.5%) IPTV cross-buyers (*CrossSell* = 1) and 25,002 (99.5%) IPTV non-buyers (*CrossSell* = 0).

Similarly, the 110,756 preprocessed observations from the ten-month period were reduced for cross selling to 87,208 observations. This data set was composed of 1260 (1.4%) IPTV cross-buyers (*CrossSell* = 1) and 85,948 (98.55%) IPTV non-buyers (*CrossSell* = 0).

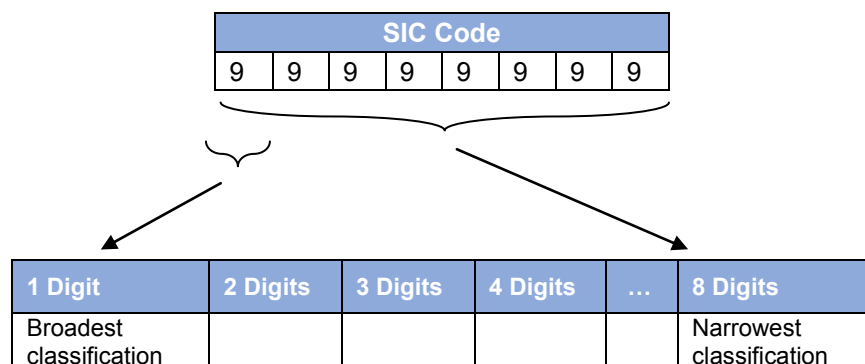
It is important to understand the design of the data that will be used for all models in this study. The data sets contain one observation on each subject so that each subject's information is independent of that of another subject. Therefore, assumption of independence of data is made as a requirement of statistical techniques.

4.1.4. Industry Analysis

The Standard Industrial Classification (SIC) code is composed of eight digits, ranging from 0-9 and thus providing approximately a total of 19,300 different categories. Too many categories in a variable while estimating a logistic regression may create an undesirable output. Therefore, an industry analysis was executed to reduce the number of categories of this variable.

From the eight digits of the SIC code, the first one indicates the broadest classification. The SIC description and some examples are displayed in Figure 4.4.

Figure 4.4 Industry Code Description and Examples



1 Digit	Description
0	Agriculture, forestry and fishing
1	Construction
2	Manufacturing – Food, textile, apparel, paper, printing, chemicals and petroleum
3	Manufacturing – Rubber, leather, stone, metal, machinery, electrical and measurement inst.
4	Transportation, communications, electric, gas & sanitary services
5	Retail Trade
6	Finance, Insurance & real estate
7	Services – Hotels, personal services, business services, automotive repair, misc. repair, motion pictures, amusement and recreation
8	Services – Health, legal, educational, social, museums & art galleries, membership organizations, miscellaneous services
9	Public administration

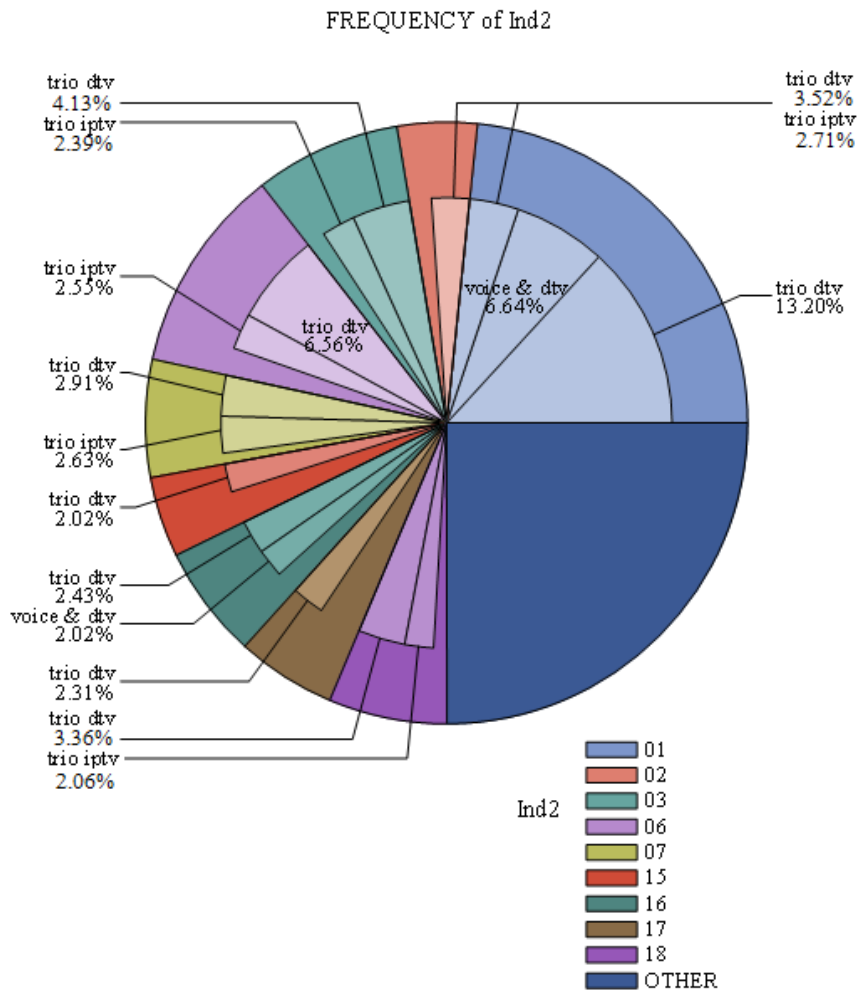
4 Digit	Description
9 3 5 3	Public administration Local government service industries General administrative services Taxation administration, local
2 6 5 4	Manufacturing Paper and allied products Paperboard containers and boxes Manufactures sanitary food containers

8 Digit	Description
8 2 4 3 9903	Services Educational services Correspondence schools and vocational schools Data processing schools Engaged in software training, computer
3 9 3 1 0043	Manufacturing Miscellaneous manufacturing industries Musical instruments Manufactures musical instruments Manufactures ukuleles and parts

In order to determine the categories that could be combined, the approach was based on the industries requiring a TV service. In consequence, a classification was developed according to the significance of each industry towards the use of the TV service in general. The process involved accounting industries to create new combinations of categories, using one or two digits of the SIC code.

The proportion of industries within the data set of the ten-month period is depicted in Figure 4.5. The advantage of discovering the industries adopting TV in any of the two available technologies, satellite and IPTV, allowed the creation of industry categories with a suitable number of digits and avoided the use of industries not requiring a TV service to operate.

Figure 4.5 TV Service by Industry



By grouping industries in each category and testing the significance of the industry variable towards the use of TV, it was possible to obtain the final combination for modeling purposes. For example, while the frequency of the *Services* industry, in particular the *Accommodation & food services* is high, the *Agriculture, Manufacturing,* and *Public administration* showed to be negligible; thus allowing to combine the latter in one category.

Three variables were tested: *Industry01* (one-digit classification with 5 categories), *Industry02* (two-digit classification with 17 categories), and *Industry02x* (two-digit classification with 24 categories). These are depicted on Table 4.4. For confidentiality reasons, the order or number associated to each category is not displayed.

Table 4.5 shows these industry variables with their respective chi-square and p-values. As seen, the use of two digits considerably increases the significance of the SIC code (p-value<.0001). But since *Industry02x* displays a higher chi-square value, this is the variable, which was simply named *Industry*, used for the models on the ten-month interval.

The models developed during the six-month interval used a variation of variable *Industry02x* that split one category in two, which resulted in a two-digit classification with 25 categories. This variable was also named *Industry*.

Table 4.4 One-Digit & Two-Digit Industry Classification

1 Digit	<i>Industry01</i>
0,1,2,3,4,9	Agriculture, construction, manufacturing, transportation, and public administration
5	Retail Trade
6	Finance, Insurance & real estate
7	Services – Hotels, personal services, business services, automotive repair, misc. repair, motion pictures, amusement and recreation
8	Services – Health, legal, educational, social, museums & art galleries, membership organizations, miscellaneous services

2 Digits	Industry02	2 Digits	Industry02x
01-19, 49	Construction, mining, agriculture, sanitary	01-19,91-99	Construction, mining, agriculture, public admin.
20-39	Manufacturing	20-39	Manufacturing
40-47	Transportation	40-49	Transportation and communications
48, 79, 84	Arts, entertainment, recreation, and hotels	50	Wholesale trade, durable
50, 51	Wholesale trade	51	Wholesale trade, non-durable
52-54, 56-59	Retail trade	52	Retail trade – hardware
55	Retail trade – Auto dealers	53	Retail trade – department
58	Retail trade – Eating and drinking places	54	Retail trade – grocery
60-64,66-69,91-99	Finance, insurance and public admin.	55	Retail trade – auto dealers
65	Real estate	56	Retail trade – apparel
70	Hotels and lodging places	57	Retail trade – furniture
72,73,76,78,81,88,89	Services – Personal, business, legal	58	Retail trade – eating and drinking places
75	Automotive repair and garages	59	Retail trade – miscellaneous
80, 83	Health care and social assistance	60-64, 66-69	Finance and insurance
82	Educational services	65	Real estate
86	Membership organizations	70, 79, 84	Arts, entertainment, recreation, and hotels
87	Professional services	72	Personal services
		73	Business services
		75	Automotive repair and garages
		76	Repair and miscellaneous services
		78, 89	Motion pictures and miscellaneous services
		81	Legal services
		82	Educational services
		80, 83	Health care and social assistance

Table 4.5 Chi-Square Test of Independence for Industry

<i>Variable</i>	<i>Chi-square</i>	<i>p-value</i>
<i>Industry01</i>	11.3340	<.0231
<i>Industry02</i>	49.8861	<.0001
<i>Industry02x</i>	57.8565	<.0001

4.1.5. Model Summary

A model can be obtained using either an estimation and a validation data sets, or a cross-validation process on the same data set to prevent over fitting. In this study, all models are estimated on the data extracted for the six- or ten-month intervals and scored on a new data set extracted from the subsequent month. In other words, the data set spanning from March to August was scored on data for September, and the data set spanning from March to December was scored on data for January 2015. The following sections will describe in detail the five models, which are summarized on Table 4.6. Because the six-month data provided a reduced number of observations with a positive value for the target variable (< 0.002%), Model B was only developed with the ten-month data.

Table 4.6 Summary of Models

Stage		Data		Referenced by	Name	Interactions
Growth	Model A	6 months	Model A0	Legacy Account ID	Model A0 [LA]	None
				Data mart ID	Model A0 [PK]	None
		10 months	Model A	Data mart ID	Model A [NI]	None
				Data mart ID	Model A [WI]	Two-way
Acquisition	Model B	10 months	Model B	Data mart ID	Model B	None

4.2. Cross-Selling Model

Estimation of the cross-selling model was developed over two intervals in the interest of the telecommunications provider to issue an initial model by the end of September 2014. This first model, Model A0, was estimated on the analysis data set created with active micro business clients between March and August 2014; and a second model, Model A, was estimated on the analysis data set created with active micro business clients between March and December 2014. Customers subscribed to either a business wired phone or a business Internet service, including customers subscribing to IPTV on a monthly basis during these two periods were selected to build these two analysis data sets

Almost the totality of the services provided in Canada are shared by the two main locations, 68.8% Ontario and 31.2% Quebec. Table 4.7 provides the description of the available variables that will be used for both Models A0 and A.

Table 4.7 Description of Available Variables after Preprocessing

Dependent variable		
<i>CrossSell</i>	<i>Binary</i>	<i>Equal to "1" if the current client adopted IPTV during an observed month, and "0" otherwise.</i>
Internal Independent Variables		
<i>PhoneDate</i>	<i>Date</i>	<i>Date of subscription of wired phone service.</i>
<i>IntDate</i>	<i>Date</i>	<i>Date of subscription of Internet service.</i>
<i>PhoneRev</i>	<i>Numerical</i>	<i>Monthly fee paid by client for wired phone service.</i>
<i>IntRev</i>	<i>Numerical</i>	<i>Monthly fee paid by client for Internet service.</i>
<i>Hosting</i>	<i>Binary</i>	<i>Equal to "1" if the client purchased a Hosting service, and "0" otherwise.</i>
<i>iProtect</i>	<i>Binary</i>	<i>Equal to "1" if the client purchased an Internet protection service, and "0" otherwise.</i>
<i>iSecurity</i>	<i>Binary</i>	<i>Equal to "1" if the client purchased an Internet security service, and "0" otherwise.</i>
<i>Bundle</i>	<i>Binary</i>	<i>Equal to "1" if the client holds services in a bundle, and "0" otherwise.</i>

<i>DMail</i>	<i>Categorical</i>	<i>Title of direct mailing sent to client.</i>
<i>DMailDate</i>	<i>Date</i>	<i>Date of direct mailing sent to client.</i>
<i>ClientCalls</i>	<i>Numerical</i>	<i>The number of calls the client has made to customer service.</i>
External Independent Variables		
<i>Industry</i>	<i>Categorical</i>	<i>SIC, eight-digit code of prospect firm's industry.</i>
<i>Employees</i>	<i>Categorical</i>	<i>Number of employees of client.</i>
<i>Location</i>	<i>Categorical</i>	<i>Geographic location of prospect firm by location.</i>
<i>Language</i>	<i>Categorical</i>	<i>Working language of client, i.e. English, French, or bilingual.</i>
<i>Organization</i>	<i>Categorical</i>	<i>Organizational structure of client, i.e. headquarters, branch, division, affiliate, and the like.</i>
<i>FRI</i>	<i>Categorical</i>	<i>Financial Risk Index assigned to the client. Indicates risk of business default within 12 months. 1= Lowest risk; 5= Highest risk. Zero if prospect has an unfavourable history.</i>
<i>CCI</i>	<i>Categorical</i>	<i>Commercial Credit Index assigned to the client. Indicates risk of delinquent payments over the next 12 months. 1= Lowest risk; 5= Highest risk. Zero if prospect has an unfavourable history.</i>

4.2.1. Model A0

A first set of models was developed to cross-sell the IPTV service in order to fulfill a SPI mandate for the Telecommunications supplier. This model used a six-month data, from March to August 2014.

The main goal was to provide a first glance of the use of an analytical approach to obtain a general customer profile by means of a list of current business clients with the highest likelihood to acquire the service.

4.2.1.1 Model Development

A logistic regression was sought to predict purchase likelihood for the IPTV service based uniquely on the available variables. In order to keep simplicity, variables were kept without any interaction, and the industry code was reduced to the minimum one-digit code, the broadest classification.

Two arrangements of clients were created in order to compare the performance of the data providing from the data mart and those coming directly from the operational databases as shown on Table 4.8. The approach was then to compare the observations with a main account number with those with a primary key designation.

Table 4.8 Two A0 Models

Model A0 [LA]	Model A0 [PK]
<i>Industry</i>	<i>Industry</i>
<i>AccountMix_LA</i>	<i>AccountMix_PK</i>
<i>Promotion</i>	<i>Promotion</i>
<i>Bundle</i>	<i>Bundle</i>

4.2.1.2 ID Variable Manipulation

The two new variables pertaining to the mix of accounts to which each client was subscribed, namely AccountMix_LA and AccountMix_PK, were created. Each one gathering the data of a particular client: (1) from the referenced legacy account, and (2)

from the data mart matching process. The first approach resulted in a list of customers classified by their phone and Internet accounts, or Legacy Accounts (LA) identifier. This identification is defined as AccountMix_LA. The second approach resulted in a list of clients by the primary key created due to the data mart, defined as PK. For these two approaches, the available variables are described on Table 4.9.

Table 4.9 Description of Available Variables for Model A0

Dependent variable		
<i>CrossSellAcquired</i>	<i>Binary</i>	<i>Equal to "1" if the current client adopted IPTV, and "0" otherwise.</i>
Independent variables		
<i>Industry</i>	<i>Categorical</i>	<i>SIC, two-digit code of client's industry.</i>
<i>Duration</i>	<i>Categorical</i>	<i>Duration of relationship with client in years, based on the first acquired service, i.e. wired phone or Internet.</i>
<i>AccountMix_PK</i>	<i>Categorical</i>	<i>The combination of all service accounts to which the client is currently subscribed, based on primary key.</i>
<i>AccountMix_LA</i>	<i>Categorical</i>	<i>The combination of all service accounts to which the client is currently subscribed, based on phone and Internet accounts.</i>
<i>ServiceMix</i>	<i>Categorical</i>	<i>The combination of all services to which the client is currently subscribed.</i>
<i>Fee</i>	<i>Categorical</i>	<i>Average monthly fee paid by client for all current services.</i>
<i>Promotion</i>	<i>Binary</i>	<i>Equal to "1" if an IPTV direct mail promotion was sent to the client during the previous month, and "0" otherwise.</i>

4.2.1.3 Model A0 [LA]

Variable Selection

Chi-square test of independence was conducted for all independent variables in relation to the dependent variable *CrossSell*. The null hypothesis for each test is that the two variables are independent or not related. The alternative hypothesis is that the two variables are dependent or related. Table 4.10 shows their chi-square values and corresponding p-values.

Table 4.10 Chi-Square Test of Independence for Model A0 [LA]

<i>Variable</i>	<i>Chi-square</i>	<i>p-value</i>
<i>Industry</i>	11.3340	<.0231
<i>AccountMix (LA)</i>	57.8565	<.0001
<i>Bundle</i>	84.0346	<.0001
<i>Promotion</i>	167.6506	<.0001

Model Estimation

Table 4.11 displays the model fit statistics further to the logistic regression.

Table 4.11 Model A0 [LA]

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1739.374	1517.247
SC	1747.244	1745.493
-2 Log L	1737.374	1459.247

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	278.1270	28	<.0001
Score	570.4707	28	<.0001
Wald	326.9351	28	<.0001

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Industry	24	34.3919	0.0779
AccMix_LA	2	38.3823	<.0001

Type 3 Analysis of Effects			
Effect	DF	Wald	
		Chi-Square	Pr > ChiSq
Bundle	1	75.7209	<.0001
Promotion	1	154.4684	<.0001

Model A0 [LA] is estimated using the following dependent variable.

$$\text{Logit}[P(\text{CrossSell} = 1 | \text{Independent Variables})]$$

And the values of the independent variables are denoted by the values on the Estimates column of Table 4.12.

Table 4.12 Estimates for Model A0 [LA]

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-4.4609	0.2187	415.9488	<.0001
Industry	01	1	0.5183	0.2624	3.9023	0.0482
Industry	02	1	2.2792	0.6756	11.3811	0.0007
Industry	03	1	-0.2254	0.5796	0.1512	0.6974
Industry	04	1	-0.6648	0.3863	2.9613	0.0853
Industry	06	1	0.3189	0.5943	0.2879	0.5915
Industry	07	1	-0.1118	0.2953	0.1434	0.7049
Industry	08	1	-0.3008	0.5023	0.3587	0.5492
Industry	09	1	-0.0795	0.2879	0.0764	0.7823
Industry	10	1	0.2407	0.4157	0.3352	0.5626
Industry	11	1	-0.6687	0.9761	0.4693	0.4933
Industry	12	1	1.2863	0.7411	3.0123	0.0826
Industry	13	1	-0.1625	0.5052	0.1035	0.7477
Industry	14	1	-1.3415	0.9860	1.8511	0.1737
Industry	15	1	0.2451	0.5740	0.1823	0.6694
Industry	16	1	-0.9994	0.9720	1.0573	0.3038
Industry	17	1	-0.1138	0.3886	0.0858	0.7696
Industry	18	1	0.2068	0.3342	0.3829	0.5360
Industry	19	1	-0.5001	0.3302	2.2948	0.1298
Industry	20	1	-0.1603	0.4576	0.1227	0.7261
Industry	21	1	-0.2972	0.6993	0.1806	0.6708
Industry	22	1	0.5487	0.3911	1.9681	0.1606
Industry	23	1	0.2777	0.3909	0.5048	0.4774
Industry	24	1	-0.4175	0.4143	1.0156	0.3136
Industry	25	1	0.5346	0.3801	1.9775	0.1597
AccMix_LA	1_0	1	-0.6239	0.2049	9.2671	0.0023
AccMix_LA	M_1	1	0.5787	0.1912	9.1663	0.0025
Bundle		1	0.9624	0.1106	75.7209	<.0001
Promotion		1	2.3023	0.1852	154.4684	<.0001

As we can see from the estimates on Table 4.12, 6 out of 29 independent variables and variable indicators are significant at a p-value of 5% or better ($<.0001 < p\text{-value} < 0.05$) once the other variables are in the model, with many of the indicators for the industry being statistically non-significant. Thus, the model suggests the following insights:

- A client who has a business classified as Industry 01 or Industry 02 has a positive effect on cross-buying likelihood against the reference category, Industry 05.
- A small business receiving a promotion over the previous month has a positive effect on cross-buying likelihood against a business not receiving any promotion.
- A business client who has bundled services has a positive effect on cross-buying likelihood against clients not having their services in a bundle.
- A small business client who has multiple phone accounts and one Internet account has a positive effect on cross-buying likelihood against a client who has one land phone account and one Internet account.
- Finally, it suggests that a client who has only one phone account and no Internet account has a negative effect on cross-buying likelihood a client who has one land phone account and one Internet account.

All of the above when the rest of the variables are kept equal in the model.

4.2.1.4 Model A0 [PK]

Variable Selection

Chi-square test of independence was conducted for all independent variables in relation to the dependent variable *CrossSell*. The null hypothesis for each test is that the two variables are independent or not related. The alternative hypothesis is that the two variables are dependent or related. Table 4.13 shows their chi-square values and corresponding p-values.

Table 4.13 Chi-Square Test of Independence for Model A0 [PK]

<i>Variable</i>	<i>Chi-square</i>	<i>p-value</i>
<i>Industry</i>	11.3340	<.0231
<i>AccountMix (PK)</i>	49.8861	<.0001
<i>Bundle</i>	84.0346	<.0001
<i>Promotion</i>	167.6506	<.0001

Model Estimation

Table 4.14 displays the model fit statistics further to the logistic regression.

Table 4.14 Model A0 [PK]

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1619.524	1390.804
SC	1627.656	1634.760
-2 Log L	1617.524	1330.804

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	286.7203	29	<.0001
Score	597.8464	29	<.0001
Wald	333.1671	29	<.0001

Type 3 Analysis of Effects			
Effect	DF	Wald	
		Chi-Square	Pr > ChiSq
Industry	24	44.2360	0.0102
AccMix_PK	2	35.7682	<.0001
Bundle	1	77.0197	<.0001
Promotion	1	150.3503	<.0001

Model A0 [PK] is estimated using the following dependent variable.

$$\text{Logit}[P(\text{CrossSell} = 1 | \text{Independent Variables})]$$

And the values of the independent variables are denoted by the values on the Estimates column of Table 4.15.

Table 4.15 Estimates for Model A0 [PK]

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald	
					Chi-Square	Pr > ChiSq
Intercept		1	-5.0306	0.2065	384.3391	<.0001
Industry	01	1	1.1573	0.2529	8.2848	0.0040
Industry	02	1	3.0003	0.6000	17.4614	<.0001
Industry	03	1	0.3940	17.5308	0.0005	0.9821
Industry	04	1	0.0145	17.5256	0.0000	0.9993
Industry	06	1	1.0172	17.5316	0.0034	0.9537
Industry	07	1	0.6053	17.5238	0.0012	0.9724
Industry	08	1	-0.2969	17.5351	0.0003	0.9865
Industry	09	1	0.2833	17.5245	0.0003	0.9871
Industry	10	1	0.0326	17.5351	0.0000	0.9985
Industry	11	1	-0.0150	17.5485	0.0000	0.9993
Industry	12	1	1.9640	17.5366	0.0125	0.9108
Industry	13	1	0.4978	17.5285	0.0008	0.9773
Industry	14	1	-0.6580	17.5490	0.0014	0.9701
Industry	15	1	1.0108	17.5307	0.0033	0.9540
Industry	16	1	-0.2940	17.5483	0.0003	0.9866
Industry	17	1	-0.0432	17.5284	0.0000	0.9980
Industry	18	1	0.7160	17.5248	0.0017	0.9674
Industry	19	1	0.0493	17.5247	0.0000	0.9978
Industry	20	1	0.1921	17.5287	0.0001	0.9913
Industry	21	1	0.3927	17.5352	0.0005	0.9821
Industry	22	1	1.3423	17.5256	0.0059	0.9389
Industry	23	1	0.7826	17.5264	0.0020	0.9644
Industry	24	1	0.1807	17.5271	0.0001	0.9918
Industry	25	1	1.3434	17.5253	0.0059	0.9389
AccMix_PK	M_1	1	0.9001	0.2239	16.1651	<.0001
AccMix_PK	1_0	1	-0.9559	0.1614	35.0658	<.0001
Bundle		1	0.9917	0.1130	77.0197	<.0001
Promotion		1	2.3677	0.1931	150.3503	<.0001

As we can see from the estimates on Table 4.15, 6 out of 29 independent variables and variable indicators are significant at a p-value of 5% or better ($<.0001 < p\text{-value} < 0.05$) once the other variables are in the model, with many of the indicators for the industry being statistically non-significant. Thus, the model suggests the following insights:

- A client who has a business classified as Industry 01 or Industry 02 has a positive effect on cross-buying likelihood against the reference category, Industry 05.
- A small business receiving a promotion over the previous month has a positive effect on cross-buying likelihood against a client not receiving any promotion.
- A business client who has bundled services has a positive effect on cross-buying likelihood against clients not having their services in a bundle.
- A small business client who has multiple phone accounts and one Internet account has a positive effect on cross-buying likelihood against a client who has one land phone account and one Internet account.
- Finally, it suggests that a client who has only one phone account and no Internet account has a negative effect on cross-buying likelihood a client who has one land phone account and one Internet account.

All of the above when the rest of the variables are kept equal in the model.

4.2.2. Model A

The second modeling period seeks deeper knowledge regarding the main influential factors explaining the variability of the probability of current small business clients to acquire the IPTV service. Two models were developed over the ten-month data collected from March to December 2014. Table 4.16 presents both A models. The first one, A [NI], uses the available variables without their interactions; while the second, A [WI], includes transformed variables and adds two-way variable interactions.

Table 4.16 Two Models A

Model A [NI]	Model A [WI]
<i>Supplier perception</i>	<i>Customer perception & experiences</i>
<i>No variable interactions</i>	<i>Two-way variable interactions</i>

Table 4.17 provides some of the essential firmographics of the analysis data set. A wide variety of small businesses is represented in the sample. It is important to note that while the variables pertaining to the financial risk and commercial credit indexes, FRI and CCI, are obviously ordered, the difference between the various levels is not consistent. Because it is believed that the distances between these levels are not equal, the variables were kept categorical during the modeling process.

Table 4.17 Firmographics for Models A [NI] and A [WI]

Characteristic	Data Composition (%)
Number of employees	
1 – 5	62.5
6 – 20	29.2
21 – 50	6.2
More than 50	2.2
Location	
Ontario	66.7
Quebec	33.3
Language of business	
English	68.3
French	27.4
Bilingual	4.3

Financial Risk Index

0	0.2
1	0.4
2	12.2
3	31.3
4	54.5
5	1.3

Commercial Credit Index

0	0.2
1	8.9
2	31.3
3	17.7
4	22.1
5	19.8

Organizational structure

<i>Single location</i>	97.1
<i>Headquarters</i>	2.9
<i>Branch</i>	0.1

Business Type (Industry)

<i>Health care, social assistance</i>	13.5
<i>Business services</i>	10.8
<i>Construction</i>	8.7
<i>Retail trade – eating and drinking places</i>	7.4
<i>Manufacturing</i>	7.4
<i>Personal services</i>	6.8
<i>Retail trade – miscellaneous</i>	5.0
<i>Automotive repair and garages</i>	4.2
<i>Wholesale trade, durable</i>	4.1
<i>Motion picture and misc. Services</i>	3.4
<i>Real estate</i>	3.4
<i>Transportation and communications</i>	3.4
<i>Retail trade – grocery</i>	3.0
<i>Arts, entertainment, recreation, and hotels</i>	2.7
<i>Legal services</i>	2.7
<i>Wholesale trade, non-durable</i>	2.3
<i>Finance and insurance</i>	2.2
<i>Retail trade – apparel</i>	2.0
<i>Retail trade – auto dealers</i>	1.7
<i>Retail trade – furniture</i>	1.6
<i>Retail trade – miscellaneous</i>	1.6
<i>Educational services</i>	1.1
<i>Retail trade – hardware</i>	0.9
<i>Retail trade – department</i>	0.4

4.2.2.1 Model A [NI]

Table 4.18 provides the composition of the ten-month period analysis data set in terms of the essential behaviour-related variables, i.e. the duration as a client with the supplier and the number of phone lines and Internet services adopted. The account mix shows that the majority of the business clients only have one phone line service.

Table 4.18 Behaviour-Related Variables for Models A [NI] and A [WI]

Characteristic	Data Composition (%)
Duration with the company in years	
1 – 4	21.0
5 – 10	21.3
10 – 15	17.1
More than 15	40.6
Account Mix	
Many phone lines & One Internet service	8.2
One phone line & One Internet service	39.1
One phone line & Zero Internet service	52.7
Hosting Add-In	
No	94.9
Yes	5.1
Email Add-In	
No	91.5
Yes	8.5
iProtection Add-In	
No	81.6
Yes	18.4
iSecurity Add-In	
No	96.7
Yes	3.3
Bundle	
No	95.9
Yes	4.1
Promotion Sent	
No	81.7
Yes	18.3

Variable Selection

In order to select the appropriate variables for Model A [N1], chi-square independent testing was performed for each of the variables with respect to the dependent variable. Table 4.19 shows the chi-square values for each relevant variable.

Table 4.19 Chi-Square Test of Independence for Model A [N1]

<i>Variable</i>	<i>Chi-square</i>	<i>p-value</i>
<i>Industry</i>	357.5354	<.0001
<i>Employees</i>	10.6343	0.0139
<i>Location</i>	47.2945	<.0001
<i>Language</i>	43.7955	<.0001
<i>CCI</i>	34.2687	<.0001
<i>Duration</i>	58.4668	<.0001
<i>AccountMix</i>	40.5101	<.0001
<i>iProtect</i>	563.0661	<.0001
<i>Hosting</i>	15.6792	<.0001
<i>Fee</i>	52.5072	<.0001
<i>Bundle</i>	354.8068	<.0001
<i>Promotion</i>	82.7267	<.0001

Model Estimation

Table 4.20 Stepwise Selection for Model A [N1]

Summary of Stepwise Selection							
Step	Effect		DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
	Entered	Removed					
1	iProtect		1	1	563.0661		<.0001
2	Industry		23	2	292.2492		<.0001
3	Promotion		1	3	129.2878		<.0001
4	Bundle		1	4	82.0194		<.0001
5	Province		1	5	58.5716		<.0001
6	Duration		3	6	25.5264		<.0001
7	Hosting		1	7	15.8784		<.0001
8	CCI		5	8	23.7963		0.0002
9	Employees		3	9	12.6360		0.0055
10	AccountMix		2	10	6.8361		0.0328

A logistic regression with the predictor variables indicating a significant chi-square test of independence in relation to the target variable, *CrossSell*, was performed with a stepwise selection. The summary of this selection is presented on Table 4.20, while the model fit statistics are displayed on Table 4.21. It is important to note that no two-way interactions were introduced in the model.

Table 4.21 Model A [N]

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	13181.411	12250.128
SC	13190.788	12643.922
-2 Log L	13179.411	12166.128

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1013.2833	41	<.0001
Score	1302.5929	41	<.0001
Wald	1119.2985	41	<.0001

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Bundle	1	72.6567	<.0001
AccountMix	2	6.8087	0.0332
Duration	3	25.1132	<.0001
Industry	23	263.2449	<.0001
Location	1	57.8705	<.0001
Employees	3	9.4118	0.0243
CCI	5	32.1013	<.0001
Hosting	1	15.6832	<.0001
iProtect	1	258.7060	<.0001
Promotion	1	149.7099	<.0001

Model A [NI] is estimated using the following dependent variable.

$$\text{Logit}[P(\text{CrossSell} = 1 | \text{Independent Variables})]$$

And the values of the independent variables are denoted by the values on the Estimates column of Table 4.22.

Table 4.22 Estimates for Model A [NI]

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-5.7480	0.5658	103.2013	<.0001
Bundle		1	0.7764	0.0911	72.6567	<.0001
AccountMix	M_1	1	-0.2298	0.1290	3.1757	0.0747
AccountMix	1_0	1	0.0939	0.0672	1.9497	0.1626
Duration	a	1	-0.0208	0.0832	0.0622	0.8031
Duration	b	1	0.1468	0.0817	3.2260	0.0725
Duration	c	1	0.3609	0.0827	19.0627	<.0001
Industry	01	1	1.2707	0.2364	28.8899	<.0001
Industry	02	1	0.3031	0.2906	1.0880	0.2969
Industry	03	1	0.2182	0.2485	0.7705	0.3800
Industry	04	1	0.2773	0.3700	0.5620	0.4535
Industry	06	1	0.1805	0.2442	0.5463	0.4598
Industry	07	1	0.2883	0.2509	1.3209	0.2504
Industry	08	1	-0.2834	0.3190	0.7891	0.3744
Industry	09	1	0.3586	0.3799	0.8911	0.3452
Industry	10	1	-0.3043	0.6248	0.2372	0.6262
Industry	11	1	0.1523	0.2873	0.2811	0.5960
Industry	12	1	0.4500	0.2950	2.3269	0.1272
Industry	13	1	-0.3444	0.3377	1.0400	0.3078
Industry	14	1	-0.0587	0.3520	0.0278	0.8675
Industry	15	1	-0.0853	0.2737	0.0970	0.7555
Industry	16	1	0.4397	0.2518	3.0491	0.0808
Industry	17	1	0.0496	0.2473	0.0402	0.8412
Industry	18	1	0.9135	0.2485	13.5177	0.0002
Industry	19	1	0.1274	0.3391	0.1410	0.7073
Industry	20	1	0.3439	0.2892	1.4140	0.2344
Industry	21	1	-0.4199	0.3232	1.6876	0.1939
Industry	22	1	0.2293	0.2784	0.6784	0.4102
Industry	23	1	0.1607	0.2729	0.3469	0.5559
Industry	24	1	0.2419	0.2946	0.6743	0.4116
Location		1	0.5318	0.0699	57.8705	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Employees	a	1	0.4361	0.2481	3.0885	0.0788
Employees	b	1	0.3734	0.2480	2.2665	0.1322
Employees	c	1	0.0476	0.2755	0.0298	0.8629
CCI	1	1	-0.1908	0.4743	0.1619	0.6874
CCI	2	1	-0.1386	0.4659	0.0885	0.7661
CCI	3	1	-0.4131	0.4688	0.7765	0.3782
CCI	4	1	-0.3356	0.4683	0.5135	0.4736
CCI	5	1	-0.6212	0.4711	1.7388	0.1873
Hosting		1	0.4399	0.1111	15.6832	<.0001
iProtect		1	1.1130	0.0692	258.7060	<.0001
Promotion		1	0.8190	0.0669	149.7099	<.0001

As we can see from the estimates on Table 4.22, 8 out of 41 independent variables and variable indicators are significant at a p-value of 5% or better ($<.0001 < p\text{-value} < 0.05$). Some indicators in the *Industry* and *Duration*, and all those related to the number of *Employees*, *Account combination* and the commercial credit index (*CCI*) are found statistically non-significant ($p\text{-value} > 0.05$) once the other variables are in the model.

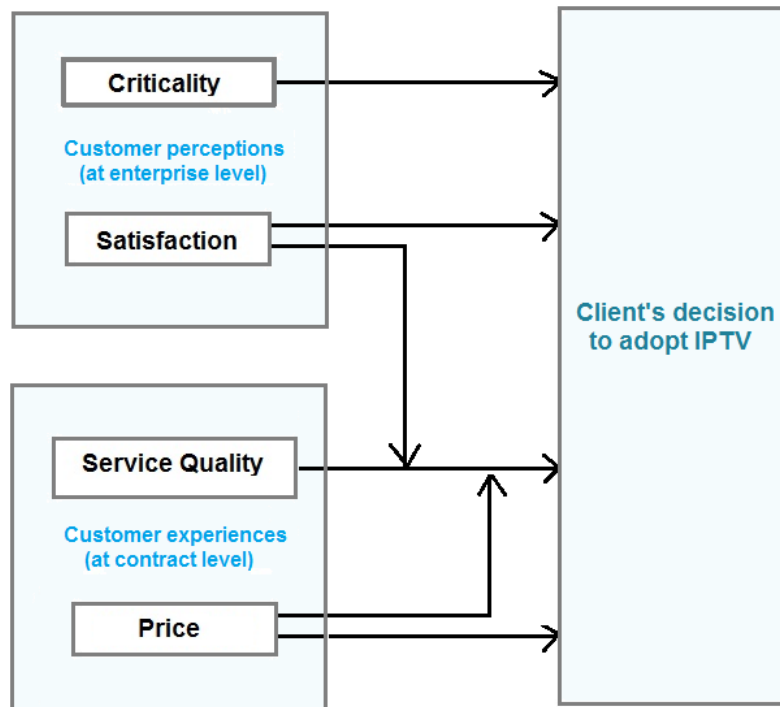
Analysis of the significant variables will be discussed in terms of the odds estimates in Chapter 5.

4.2.2.2 Model A [w1]

This section will explain in detail the basis of the second model A [w1], which was developed on the grounds of previous research, isolating the factors that uniquely affect the client's cross-buying decision. From the small business' perspective, the decision to adopt an extra service, IPTV, is the process of solving the problem of market matching and congruence.

As stated in Chapter 2, Bolton et al. (2008) identified four main components affecting the decision to cross-buy: criticality, satisfaction, service quality, and price. Therefore, new variables were created according to the conceptual structure depicted in Figure 4.6.

Figure 4.6 Conceptual Cross-Selling Model A [w1] (Adapted from Bolton et al., 2008)



The only variable already available relates to the need or criticality of having the IPTV service for the business' operation. This is obtained with the variable *Industry*.

Variable Selection

New variables were created. First, in order to obtain the price, the quantitative representation of the variable pertaining to the monthly fee was retrieved, and thus transformed into a logarithmic numeric value, *LogPrice*.

Second, the service quality in terms of the current service was represented by the variable *QoS*, which sums up the Internet protection and Internet security add-ins to the account of the client, if any. It was zero if none of the features were present, one if only one of these was added, and two if both were opted in.

And third, unable to conduct a survey targeting the same business clients taken in the analysis data set, satisfaction was gathered as the inverse of the number of calls made at the customer service. Simplicity was chosen by considering that a business client is satisfied when it does not call customer service at all, or calls up to 1 time. Therefore, the variable *Satisfied* was set to “0” if the number of calls was greater or equal to two, and “1” otherwise.

The description of the new variables and the composition within the analysis data set are displayed on Tables 4.23 and 4.24 respectively. For confidentiality reasons, the composition of the price is not presented.

Table 4.23 Description of New Variables for Model A [w1]

<i>Independent variables</i>		
<i>Price</i>	<i>Numerical</i>	<i>Average monthly fee paid by client for all current services.</i>
<i>LogPrice</i>	<i>Numerical</i>	<i>Logarithmic of current price paid by client for all current services.</i>
<i>QoS</i>	<i>Numerical</i>	<i>Quality of service measured by the number of additional features, i.e. Internet protection and/or Internet security acquired by the client.</i>
<i>ClientCalls</i>	<i>Numerical</i>	<i>The number of calls the client has made to customer service.</i>
<i>Satisfied</i>	<i>Binary</i>	<i>Equal to “1” if count of calls made by the client is less than 2, and “0” otherwise.</i>

Table 4.24 Additional Behaviour-Related Variables for Model A [WI]

<i>Characteristic</i>	<i>Data Composition (%)</i>
Quality of Service (QoS)	
1	78.8
2	20.7
3	0.5
Client Calls	
0	84.7
1	8.9
2	2.9
3	1.4
4	0.8
5	0.2
6	0.4
7	0.1
8	0.2
9	0.4
Satisfied	
<i>Yes</i>	93.6
<i>No</i>	6.4

In order to select the appropriate variables, chi-square independent testing was performed for each one of the variables with respect to the dependent variable, CrossSell, as seen on Table 4.25. The correlation between the quantitative variables can be viewed in the correlation matrix of Table 4.26.

Table 4.25 Chi-Square Test of Independence for Model A [WI]

<i>Variable</i>	<i>Chi-square</i>	<i>p-value</i>
<i>QoS</i>	484.3905	<.0001
<i>ClientCalls</i>	3258.4067	<.0001
<i>Satisfied</i>	1340.6741	<.0001

Table 4.26 Correlation Matrix for Model A [w]

Pearson Correlation Coefficients, N = 87208 Prob > r under H0: Rho=0			
	Price	Logprice	CrossSell
Price	1.00000	0.83055 <.0001	-0.00087 0.7962
Logprice	0.83055 <.0001	1.00000	0.01092 0.0013
CrossSell	-0.00087 0.7962	0.01092 0.0013	1.00000

Model Estimation

A logistic regression with the predictor variables indicating a significant chi-square test of independence in relation to the target variable, *CrossSell*, was performed with a stepwise selection. The summary of this selection is presented in Table 4.27, while the model fit statistics are displayed on Table 4.28. It is important to note that all possible pair interactions were introduced in the model.

Table 4.27 Stepwise Selection for Model A [w]

Summary of Stepwise Selection						
Step	Effect		DF	Number In	Score Chi-Square	Pr > ChiSq
	Entered	Removed				
1	Satisfied		1	1	1340.6741	<.0001
2	QoS		1	2	327.1159	<.0001
3	Industry		23	3	305.1752	<.0001
4	Promotion		1	4	135.9708	<.0001
5	Bundle		1	5	103.5347	<.0001
6	Location		1	6	62.4520	<.0001
7	QoS*Satisfied		1	7	34.2848	<.0001
8	Duration		3	8	21.2159	<.0001
9	AccountMix*Duration		6	9	92.3242	<.0001
10	Bundle*Duration		3	10	39.4638	<.0001
11	CCI		5	11	18.1170	0.0028
12	Employees		3	12	12.3434	0.0063
13	Promotion*Employees		3	13	27.1715	<.0001
14	AccountMix*Employees		6	14	30.3265	<.0001
15	Bundle*AccountMix		2	15	44.1739	<.0001
16	LogPrice		1	16	9.9073	0.0016
17	LogPrice*QoS		1	17	68.5472	<.0001

The stepwise selection found significant the two-way interactions between the following variables: the accounts mix (*AccountMix*) and the duration of the client-supplier

relationship (*Duration*), the existence of bundled services (*Bundle*) and the duration of the client-supplier relationship (*Duration*), the micro business having received a promotional offer (*Promotion*) and the number of employees (*Employees*), the existence of bundled services (*Bundle*) and the accounts mix (*AccountMix*), the service quality (*QoS*) and paid price (*LogPrice*), and finally the service quality (*QoS*) and customer satisfaction (*Satisfied*). The meaning of these interactions is discussed at the analysis and results stage, Chapter 5.

Table 4.28 Model A [w]

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	13181.411	11353.309
SC	13190.788	11944.000
-2 Log L	13179.411	11227.309

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1952.1025	62	<.0001
Score	3053.3108	62	<.0001
Wald	2045.1235	62	<.0001

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Satisfied	1	206.1886	<.0001
QoS	1	77.0593	<.0001
Industry	23	263.0922	<.0001
Promotion	1	0.0462	0.8298
Bundle	1	10.6893	0.0011
Location	1	50.7554	<.0001
QoS*Satisfied	1	24.9378	<.0001
Duration	3	34.4617	<.0001
AccountMix*Duration	6	84.2632	<.0001
Bundle*Duration	3	75.5298	<.0001
CCI	5	23.8204	0.0002
Employees	3	2.1393	0.5440
Promotion*Employees	3	23.7475	<.0001
AccountMix*Employees	6	11.5836	0.0719
Bundle*AccountMix	2	38.5550	<.0001
LogPrice	1	77.7568	<.0001
QoS*LogPrice	1	67.6690	<.0001

Model A [wI] is estimated using the following dependent variable.

$$\text{Logit}[P(\text{CrossSell} = 1 | \text{Independent Variables})]$$

And the values of the independent variables are denoted by the values on the Estimates column of Table 4.29. The full equation is provided in the Appendix.

Table 4.29 Model A [wI] – Cross-Sell Estimates

Analysis of Maximum Likelihood Estimates							
Parameter			D F	Estimate	Standard Error	Wald Chi- Square	Pr > Chi Sq
Intercept			1	-10.0574	0.9273	117.6398	<.0001
Industry	01		1	1.4048	0.2404	34.1358	<.0001
Industry	02		1	0.3604	0.2945	1.4978	0.2210
Industry	03		1	0.4281	0.2517	2.8925	0.0890
Industry	04		1	0.2978	0.3747	0.6316	0.4268
Industry	06		1	0.2400	0.2470	0.9441	0.3312
Industry	07		1	0.4079	0.2543	2.5735	0.1087
Industry	08		1	-0.1677	0.3226	0.2703	0.6031
Industry	09		1	0.4904	0.3839	1.6320	0.2014
Industry	10		1	-0.3474	0.6436	0.2914	0.5893
Industry	11		1	0.3125	0.2925	1.1415	0.2853
Industry	12		1	0.5809	0.2987	3.7832	0.0518
Industry	13		1	-0.0528	0.3436	0.0236	0.8779
Industry	14		1	0.1238	0.3553	0.1215	0.7275
Industry	15		1	0.0175	0.2771	0.0040	0.9496
Industry	16		1	0.6260	0.2562	5.9704	0.0145
Industry	17		1	0.1347	0.2504	0.2893	0.5906
Industry	18		1	1.0891	0.2520	18.6843	<.0001
Industry	19		1	0.3838	0.3433	1.2495	0.2636
Industry	20		1	0.2514	0.2926	0.7383	0.3902
Industry	21		1	-0.3914	0.3272	1.4317	0.2315
Industry	22		1	0.2883	0.2819	1.0460	0.3064
Industry	23		1	0.2752	0.2760	0.9943	0.3187
Industry	24		1	0.3668	0.2986	1.5086	0.2193
Promotion			1	-0.1187	0.5525	0.0462	0.8298
Duration	a		1	-0.6781	0.1441	22.1324	<.0001
Duration	b		1	-0.0154	0.1102	0.0195	0.8889
Duration	c		1	0.2000	0.1142	3.0664	0.0799
Employees	a		1	0.4007	0.3027	1.7528	0.1855
Employees	b		1	0.3410	0.3028	1.2682	0.2601
Employees	c		1	0.2462	0.3647	0.4557	0.4996

Analysis of Maximum Likelihood Estimates							
Parameter			D F	Estimate	Standard Error	Wald Chi- Square	Pr > Chi Sq
CCI	1		1	-0.1039	0.4810	0.0466	0.8290
CCI	2		1	-0.0417	0.4726	0.0078	0.9296
CCI	3		1	-0.2791	0.4755	0.3444	0.5573
CCI	4		1	-0.2070	0.4752	0.1898	0.6631
CCI	5		1	-0.4701	0.4781	0.9670	0.3254
Bundle			1	0.7229	0.2211	10.6893	0.0011
Location			1	0.5035	0.0707	50.7554	<.0001
AccountMix*Duration	M_1	a	1	0.3129	0.3886	0.6482	0.4208
AccountMix*Duration	M_1	b	1	-0.1015	0.3114	0.1063	0.7444
AccountMix*Duration	M_1	c	1	0.3000	0.3436	0.7621	0.3827
AccountMix*Duration	1_0	a	1	1.5783	0.1815	75.6363	<.0001
AccountMix*Duration	1_0	b	1	0.5137	0.1724	8.8842	0.0029
AccountMix*Duration	1_0	c	1	0.2890	0.1757	2.7069	0.0999
Bundle*Duration		a	1	-1.8520	0.2513	54.3229	<.0001
Bundle*Duration		b	1	-0.3354	0.2552	1.7276	0.1887
Bundle*Duration		c	1	0.0989	0.2806	0.1242	0.7246
Bundle*AccountMix		M_1	1	-0.4467	0.6341	0.4962	0.4812
Bundle*AccountMix		1_0	1	1.2595	0.2116	35.4126	<.0001
Promotion*Employees	a		1	1.1333	0.5585	4.1178	0.0424
Promotion*Employees	b		1	0.9110	0.5655	2.5950	0.1072
Promotion*Employees	c		1	-0.7253	0.6902	1.1044	0.2933
AccountMix*Employee	M_1	a	1	-0.1392	0.2809	0.2457	0.6201
AccountMix*Employee	M_1	b	1	-0.5841	0.2841	4.2274	0.0398
AccountMix*Employee	M_1	c	1	0.1974	0.3840	0.2643	0.6072
AccountMix*Employee	1_0	a	1	-0.3025	0.1334	5.1448	0.0233
AccountMix*Employee	1_0	b	1	-0.2944	0.1477	3.9719	0.0463
AccountMix*Employee	1_0	c	1	-0.3169	0.3294	0.9257	0.3360
QoS			1	3.8446	0.4380	77.0593	<.0001
LogPrice			1	1.1061	0.1254	77.7568	<.0001
QoS*LogPrice			1	-0.6697	0.0814	67.6690	<.0001
Satisfied			1	-2.7595	0.1922	206.1886	<.0001
QoS*Satisfied			1	0.6197	0.1241	24.9378	<.0001

As we can see from the estimates on Table 4.29, 19 out of 62 independent variables and variable indicators are significant at a p-value of 5% or better ($<.0001 < p\text{-value} < 0.05$), with some indicators in the *Industry*, and all those related to the number of *Employees*, *Duration*, *Promotion*, *AccountMix* and the commercial credit index (*CCI*) are

found statistically non-significant ($p\text{-value} > 0.05$) once the other variables are in the model.

Attention is particularly drawn to the presence of two-way interactions and their individual components due to the existent collinearity, which is the result of having two or more correlated predictor variables in a regression model.

Indeed, one must be careful in analyzing the effect of each interaction because when we say that a single interaction increases by one, it does not imply that the main effects do, and vice versa.

The two-way variable interactions are difficult to interpret, but in an effort to understand the underlying elements contributing to the probability in adopting IPTV, the following two interactions are worth mentioning.

1. The model may suggest that the strength of the relationship between a client's likelihood of cross-buying and the perceived service quality on the focal contract is weaker when the price of the contract is low than when it is high.
2. The model may suggest that the strength of the relationship between a client's likelihood of cross-buying and the perceived service quality on the focal contract is stronger when the decision maker's overall satisfaction with the supplier is high than when it is low.

It is important to note that the interaction of variables may diminish the significance of the individual variables comprising the interaction. However, these are not removed from the model as they provide global significance to the interaction.

Analysis of the significant variables and their two-way interactions will be discussed in terms of the odds estimates in Chapter 5.

4.3. Acquisition Model

4.3.1. Model B

Model B targets new prospects not currently subscribing to any service. A logistic regression was used to predict purchase likelihood for the IPTV service.

Because the proportion of completely new clients was very small within the total customer base, all observations with target variable set to '0' and having missing values were removed from the modeling process to reduce the number of clients without the IPTV service. It is important to note the risk of bias due to the imputation of missing values within the observations having the target variable set to '1'. The final analysis data set was comprised of 6,576 observations, from which 76 were new IPTV clients.

4.3.1.1 Variable Selection

In order to select the appropriate variables, chi-square independent testing was performed for each one of the variables described on Table 4.30 with respect to the dependent variable *Acquired*, as seen on Table 4.31.

Table 4.30 Description of Variables for Model B

Dependent variable		
<i>Acquired</i>	<i>Binary</i>	<i>Equal to "1" if the prospect firm was newly acquired, and "0" otherwise.</i>
External Independent Variables		
<i>Industry</i>	<i>Categorical</i>	<i>SIC, two-digit code of prospect firm's industry.</i>
<i>Employees</i>	<i>Categorical</i>	<i>Number of employees of client.</i>
<i>Location</i>	<i>Categorical</i>	<i>Geographic location of prospect firm.</i>
<i>Language</i>	<i>Categorical</i>	<i>Working language of client, i.e. English, French, or bilingual.</i>
<i>CCI</i>	<i>Categorical</i>	<i>Commercial Credit Index assigned to the client.</i>
<i>FRI</i>	<i>Categorical</i>	<i>Financial Risk Index assigned to the client.</i>
<i>Organization</i>	<i>Categorical</i>	<i>Organizational structure of client, i.e. headquarters, branch, division, affiliate, and the like.</i>
<i>Promotion</i>	<i>Binary</i>	<i>Equal to "1" if an IPTV promotion was sent to the prospect during the previous month, and "0" otherwise.</i>

Table 4.31 Chi-Square Test of Independence for Model B

<i>Variable</i>	<i>Chi-square</i>	<i>p-value</i>
<i>Industry</i>	357.5354	<.0001
<i>Employees</i>	10.6343	0.0139
<i>Location</i>	47.2945	<.0001
<i>Language</i>	43.7955	<.0001
<i>CCI</i>	34.2687	<.0001
<i>FRI</i>	58.4668	<.0001
<i>Organization</i>	354.8068	<.0001
<i>Promotion</i>	82.7267	<.0001

4.3.1.2 Model Estimation

The logistic regression presented concerns regarding the convergence criterion when considering all variables. In logistic regression, it has been admitted that with small to medium-sized data sets, it may not be possible to obtain the maximum likelihood estimates. These situations occur if the responses and non-responses can be perfectly separated by a single factor or by a non-trivial linear combination of factors (Albert & Anderson, 1984), thus presenting either a complete or a quasi-complete separation of data points. One suggestion is to determine which variable is causing separation problems and omit it from the model. Heinze and Schemper (2002) proved that the penalized maximum likelihood estimation method, proposed by Firth (1993) to reduce bias in logistic regression in small samples always provides finite estimates of parameters under complete or quasi-complete separation. Therefore, having tried the STEPWISE, the BACKWARD and the FORWARD selections in PROC LOGISTIC, and obtaining no satisfactory results with any combination of variables, these selections were replaced by the FIRTH option to use Firth's method.

Variables were suppressed from the modeling process one at a time until the convergence criterion was satisfied. The final model could only hold the *Industry* variable. Table 4.32 presents the model fit statistics.

Table 4.32 Model B

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	802.451	799.447
SC	809.243	962.435
-2 Log L	800.451	751.447

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	49.0049	23	0.0012
Score	67.9418	23	<.0001
Wald	61.3951	23	0.0004

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Industry	23	52.4204	0.0004

Model B is estimated using the following dependent variable.

$$\text{Logit}[P(\text{Acquire} = 1 | \text{Independent Variables})]$$

And the values of the independent variables are denoted by the values on the Estimates column of Table 4.33.

Table 4.33 Estimates for Model B

Analysis of Penalized Maximum Likelihood Estimates						
Parameter		D F	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-4.5662	0.1768	667.2990	<.0001
Industry	01	1	1.6523	0.2660	38.5734	<.0001
Industry	02	1	0.3735	0.6369	0.3439	0.5576
Industry	03	1	-0.8137	0.6327	1.6542	0.1984
Industry	04	1	-0.5152	1.3781	0.1398	0.7085
Industry	06	1	-0.0934	0.3745	0.0622	0.8031
Industry	07	1	-0.0533	0.4872	0.0120	0.9128
Industry	08	1	-1.5844	1.3698	1.3379	0.2474
Industry	09	1	-0.1434	1.3837	0.0107	0.9175

Analysis of Penalized Maximum Likelihood Estimates						
Parameter		D F	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Industry	10	1	0.3035	1.3941	0.0474	0.8277
Industry	11	1	-0.3920	0.8059	0.2366	0.6267
Industry	12	1	0.7595	0.6398	1.4092	0.2352
Industry	13	1	0.0193	0.8082	0.0006	0.9809
Industry	14	1	0.2849	0.8102	0.1236	0.7251
Industry	15	1	-0.7857	0.8045	0.9538	0.3287
Industry	16	1	0.3343	0.4179	0.6399	0.4237
Industry	17	1	0.1530	0.3749	0.1666	0.6832
Industry	18	1	0.4337	0.4888	0.7871	0.3750
Industry	19	1	-0.7271	1.3756	0.2794	0.5971
Industry	20	1	0.9198	0.4914	3.5043	0.0612
Industry	21	1	-0.3826	0.8060	0.2253	0.6350
Industry	22	1	-0.4815	0.8055	0.3574	0.5500
Industry	23	1	0.2563	0.5456	0.2207	0.6385
Industry	24	1	-0.0621	0.8077	0.0059	0.9387

As we can see from the estimates on Table 4.33, only one out of the 24 *Industry* indicators is significant at a p-value of 5% or better ($p\text{-value} < .0001$) once the other indicators are in the model, and the rest of the indicators are statistically non-significant.

This suggests that a client who has a business classified as Industry 01 has a positive and increasing effect on IPTV acquisition likelihood against the reference category, Industry 05.

Chapter 5.

Analysis of Results

This chapter will present the predictive performance of both models A and B, growth and acquisition, by means of testing or scoring on new data sets. The results will be compared in terms of the following methods.

There are at least four ways to evaluate predictive accuracy in the context of the present research. The first is to compare one model versus another by means of a lift chart, which sketches the relationship between a predicted ordering of customers in terms of their likelihood to buy. The lift is a measure of the effectiveness of a predictive model, calculated as the ratio between the results obtained with and without the predictive model. In other words, it provides the performance of the model based on its power to select appropriate customers compared to choosing customers at random.

A second way to evaluate predictive accuracy is by means of the area under the receiver operating characteristic (ROC) curve, also known as the AUC. The AUC is a common metric to measure the predictive capability of a model by depicting the sensitivity against the specificity of the model. The sensitivity measures the fraction of converted clients that the test correctly identifies as positive. The specificity measures the fraction of non-converted clients that the test correctly identifies as negative. Therefore, the AUC can vary from 0.5, which depicts a random or worthless model, to 1, which depicts a perfect model.

A third way to evaluate a model can be made by comparing to a benchmark. The standard measure against which the growth models will be compared relates to the marketing campaigns to develop presently active clients.

And finally, while often overlooked, a fourth way to evaluate a model can be in terms of the insights it generates by interpreting the odds ratios. It is possible to have more confidence in extrapolating a model to some future time period if it is believed that the underlying parameters make sense.

The scoring results will be presented by means of comparative lift curves and/or AUCs. And the insights from the estimated odds ratios generated for each significant factor will be further discussed for the selected models.

5.1. Model A

This section presents the comparative performance of the models built for growth by cross-selling IPTV. For the models developed during the six-month interval, Models A0 [LA] and A0 [PK], only the lift curves on scoring data are discussed. While for the models developed during the ten-month interval, Models A [NI] and A [WI], the lift curves, the AUCs on scoring data, and the insights from the estimated odds ratios are discussed.

5.1.1. Model Scoring

Models A0 [LA] and A0 [PK] were scored on data from the month of September 2014. These represent the simplest approach at obtaining an initial model due to the incorporation of a reduced number of variables and the absence of any type of variable interactions or variable transformations, while being built under a tight timeframe. The main goal was to compare the performance of the models by accessing the data directly from the operational databases, based on the legacy account identification, against the use of the data mart, based on the primary key identification. Accordingly, the models reflect the difference resulting from implementing a data warehousing environment against the absence of such a structure.

All in all, both models outperform the 2.07 benchmark lift specified by the previous marketing mailing campaign, channeled to 66.85% of the phone and Internet client base.

The lift curves of Models A0 [LA] and A0 [PK] are compared in Figure 5.1, and the lift for the second model is isolated in Figure 5.2. As seen, the model based on the PK identifier provides a higher lift of 2.55 compared to the 2.15 lift of the model based on the LA identifier, thus giving confidence to the structure and use of the data mart.

Figure 5.1 Comparative Lift Chart for Models A0 [LA] and A0 [PK]

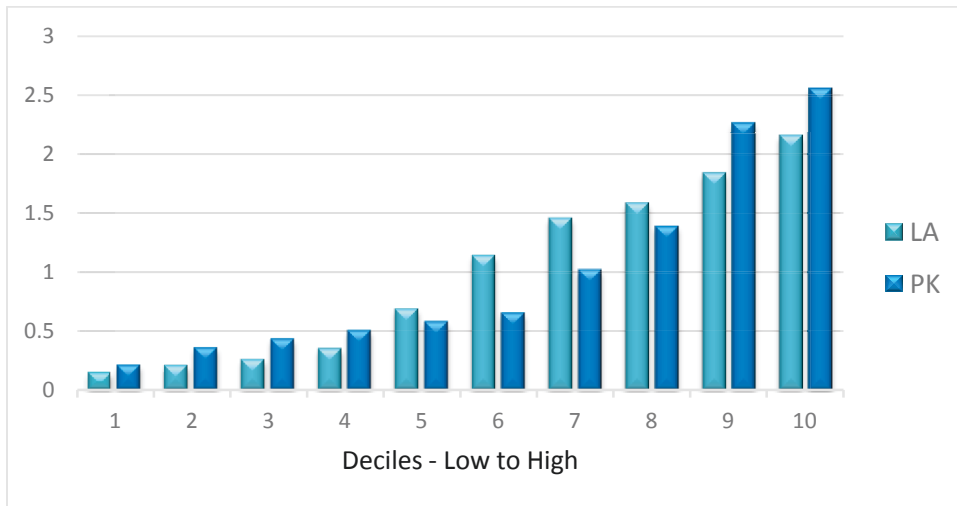
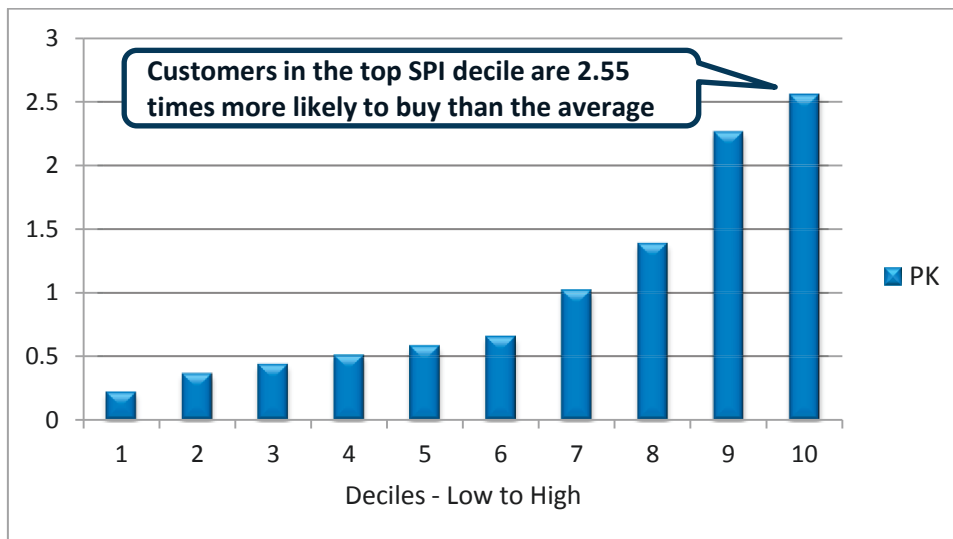


Figure 5.2 Lift Curve for Model A0 [PK]



The lift curves of Models A [NI] and A [WI] were scored on data from the month of January 2015. Their lift curves are compared in Figure 5.3. Figure 5.4 isolates the model that incorporated two-way interactions. This model provides a higher performance with a lift of 4.88 compared to the 4.13 lift of the first model.

Figure 5.3 Comparative Lift Chart for Models A [NI] vs. A [WI]

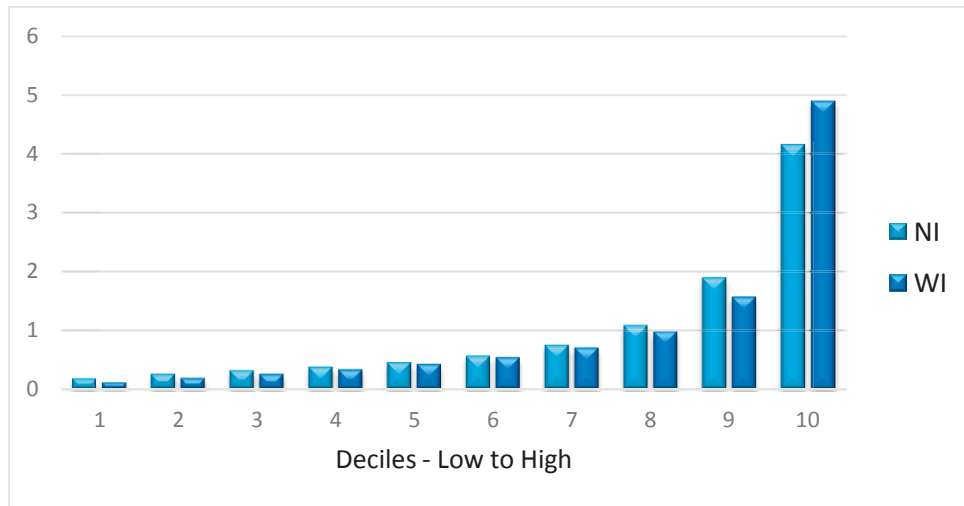
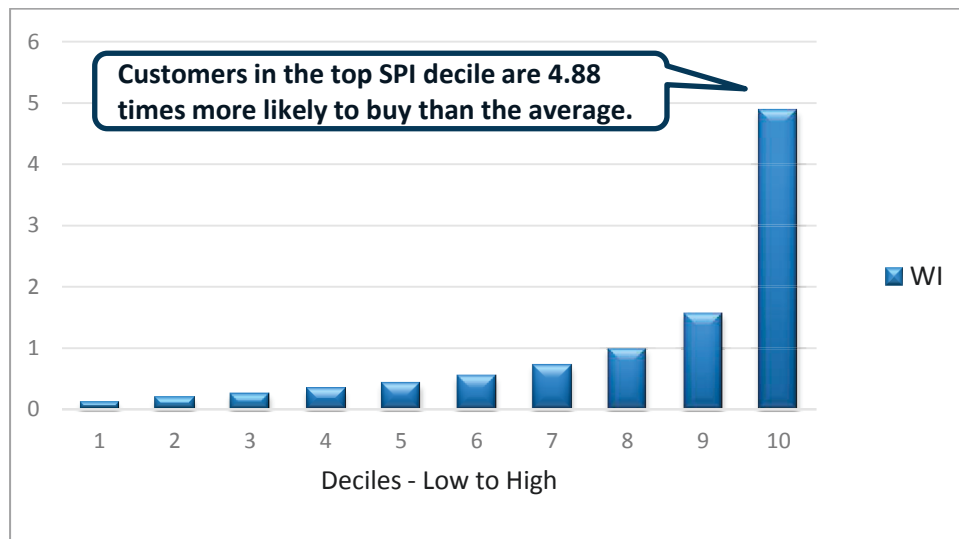


Figure 5.4 Lift Curve for Model A [WI]



The AUC curves during the estimation and scoring processes for both models are presented in Figures 5.5 and 5.6, respectively. The red line depicts a random model with 0.5 area. It can be seen that the scoring process provides a smaller area for both models. But the difference is smaller for Model A [W].

Figure 5.5 ROC Curves for Model A [NI]

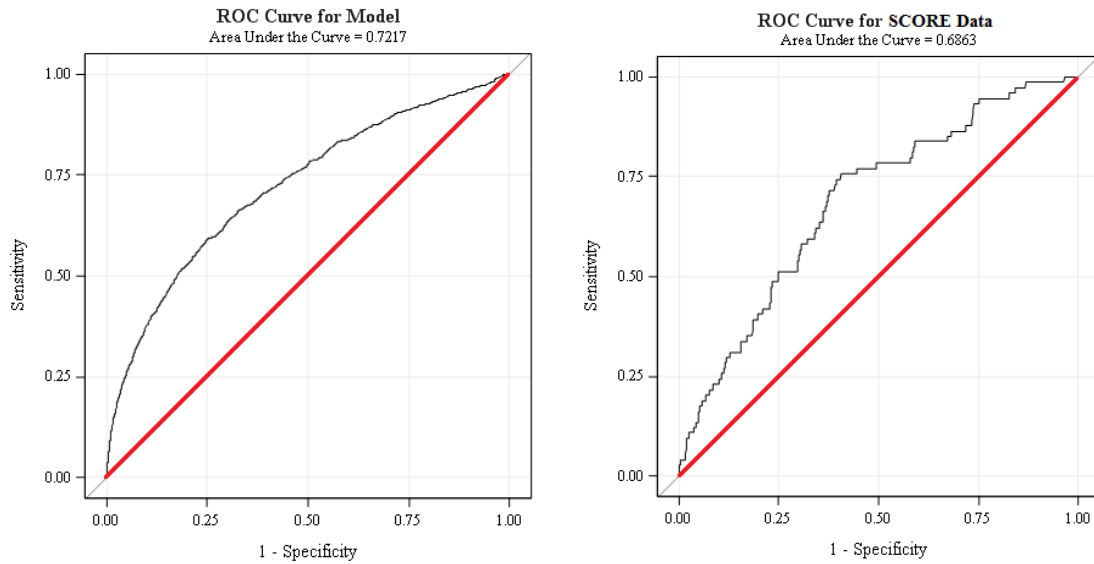
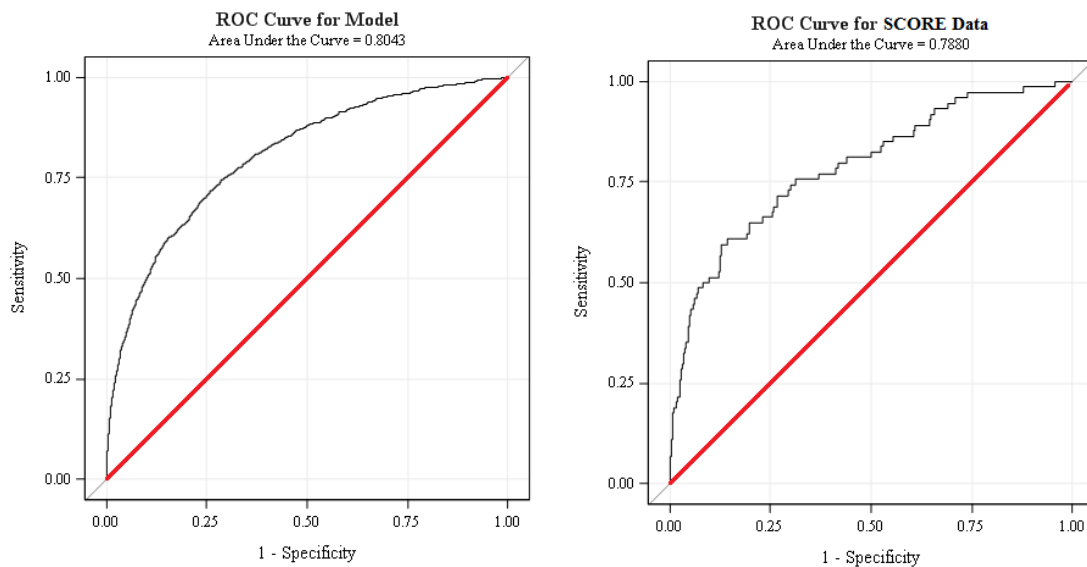


Figure 5.6 ROC Curves for Model A [W]



Comparison of the models estimated over the ten-month period is presented in Table 5.1 as they considered a larger representativeness of the data collected, which includes seasonal changes over time. The performance of the models is compared by means of the fit statistics for the scored data set. The table lists the variables entered, the Akaike's information criteria (AIC), the area under the curve (AUC), and the lifts obtained.

Table 5.1 Comparison of Cross-Selling Models

	Model A [NI] No variable interactions	Model A [WI] With variable interactions
AIC	12250.128	11353.309
AUC / ROC	0.7217 / Score 0.6863	0.8043 / Score 0.7880
Lift Top Decile	4.13	4.88
Variables	<i>Industry</i> <i>Employees</i> <i>Location</i> <i>CCI</i> <i>Duration</i> <i>AccountMix</i> <i>Bundle</i> <i>Promotion</i> <i>iProtect</i> <i>Hosting</i>	<i>Industry</i> <i>Employees</i> <i>Location</i> <i>CCI</i> <i>Duration</i> <i>AccountMix</i> <i>Bundle</i> <i>Promotion</i> <i>QoS</i> <i>LogPrice</i> <i>Satisfied</i> <i>QoS * LogPrice</i> <i>QoS * Satisfied</i> <i>AccountMix * Duration</i> <i>AccountMix * Employees</i> <i>Bundle * Duration</i> <i>Promotion * Employees</i>

5.1.2. Analysis

The logistic regression generates probabilities that serve as scores for each customer. A perennial issue in database marketing is how far down the list to target. The answer depends on whether the probabilities generated by the statistical model can be interpreted in absolute terms.

PROC LOGISTIC computes odds ratio estimates only for variables not involved in interactions or nested terms because when a variable is involved in an interaction, there is no single odds ratio estimate assigned. Rather, the odds ratio for the variable

depends on the level or levels of the interacting variables. Therefore, a list of the odds ratios is computed only for the statistically significant variables for both models.

Table 5.2 Odds Ratio Estimates for Model A [NI]

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Industry 01 vs 05	3.563	2.242	5.664
Industry 18 vs 05	2.493	1.532	4.057
Duration c vs d	1.435	1.220	1.687
Location	1.702	1.484	1.952
Bundle	2.174	1.818	2.598
Hosting	1.553	1.249	1.930
iProtect	3.044	2.658	3.486
Promotion	2.268	1.989	2.586

The following insights are gained from the odds ratios of Model A [NI], presented on Table 5.2. It is important to note that the table includes only the variables that are statistically significant at a 5% or better ($<.0001 < p\text{-value} < 0.05$) once the other variables are in the model.

With respect to the industry, it can be seen that when all other variables are kept equal, the clients having a business classified as Industry 01 have their odds ratio multiplied by 3.563 against clients having a business classified as Industry 05. This would represent an increase in odds by 256.3%. Similarly, clients having a business classified as Industry 18 have their odds ratio multiplied by 2.493 against clients having a business classified as Industry 05, which represents an increase in odds by 149.3%.

In terms of duration on being a client with the telecom firm, customers who started a relationship with the supplier identified as category “c” have their odds ratio multiplied by 1.435 against clients who started a relationship with the supplier identified as category “d”, which represents an increase in odds by 43.5%.

Regarding the geographical location, clients having a business in Location 1 have their odds ratio multiplied by 1.702 against clients having a business in Location 0, which represents an increase in odds by 70.2%.

With respect to bundled services, clients granted a bundle have their odds ratio multiplied by 2.174 against clients who do not have this benefit, which represents an increase in odds by 117.4%.

In terms of additional features, clients opting for hosting or Internet protect features have their odds ratio multiplied by 1.553 and 3.044 respectively against clients who do not have these options, which represents an increase in odds by 55.3% and 204.4% respectively.

And finally, clients receiving a direct mail with an IPTV promotion have their odds ratio multiplied by 2.268 against clients who do not receive such promotion, which represents an increase in odds by 126.8%.

Table 5.3 Odds Ratio Estimates for Model A [wI] – First Part

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Industry 01 vs 05	4.075	2.609	6.728
Industry 16 vs 05	1.870	1.156	3.171
Industry 18 vs 05	2.972	1.854	5.003
Location	1.654	1.442	1.903

The following insights are gained from the odds ratios of Model A [wI], presented on Table 5.3. It is important to note that the table includes only the variables that are statistically significant at a 5% or better ($.0001 < p\text{-value} < 0.05$) once the other variables are in the model.

With regard to the industry, it can be seen that when all other variables are kept equal, the clients having a business classified as Industry 01 have their odds ratio multiplied by 4.075 against clients having a business classified within Industry 05. This would represent an increase in odds by 307.5%. Similarly, clients having a business classified as Industry 16 have their odds ratio multiplied by 1.870 against clients having a business classified within Industry 05, which represents an increase in odds by 87.0%. And finally, clients having a business classified as Industry 18 have their odds ratio

multiplied by 2.972 against clients having a business classified as Industry 05, which represents an increase in odds by 197.2%.

In terms of the geographical location, clients having a business in Location 1 have their odds ratio multiplied by 1.654 against clients having a business in Location 0, which represents an increase in odds by 65.4%.

In order to obtain the odds ratio estimates of simple effects within the two-way interactions when a categorical variable interacts with a continuous or binary variable, the ODDS RATIO statement of the PROC LOGISTIC was used, as this SAS statement produces the odds ratios at the mean of the interacting covariate. The estimates depicted on Table 5.4 shows the odds ratios for the two-way interactions between QoS (three categories) and *LogPrice*, and QoS (three categories) and *Satisfied*.

Table 5.4 Odds Ratio Estimates for Model A [W1] – Second Part

Odds Ratio Estimates and Wald Confidence Intervals			
Label	Estimate	95% Confidence Limits	
Logprice at QoS =3	0.405	0.309	0.531
Logprice at QoS =2	0.792	0.693	0.905
Logprice at QoS =1	1.547	1.376	1.739
Satisfied at QoS=1	0.118	0.099	0.139
Satisfied at QoS=2	0.219	0.182	0.263
Satisfied at QoS =3	0.405	0.273	0.604

[Relevant rows of the regression estimates on Table 4.29]

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > Chi Sq
QoS		1	3.8446	0.4380	77.0593	<.0001
LogPrice		1	1.1061	0.1254	77.7568	<.0001
Satisfied		1	-2.7595	0.1922	206.188	<.0001
QoS*LogPrice		1	-0.6697	0.0814	67.6690	<.0001
QoS*Satisfied		1	0.6197	0.1241	24.9378	<.0001

By recalling the relevant rows of the regression estimates on Table 4.29, the model suggests that a client who pays a higher price for the focal ongoing service has a

positive and increasing effect on cross-buying likelihood. It suggests that a client who is satisfied with the focal contract has a negative effect on cross-buying likelihood. And it also suggests that a client who perceives a high quality of service has a positive and increasing effect on the cross-buying likelihood. The above when all other variables are kept equal.

From the odds ratios of the main variables with two-way interactions shown on Table 5.4, the following insights are gained. The effect is particularly strong for the service quality variable, QoS. Decreasing the perception the client has over the service quality of the focal services, significantly enhances the odds that the customer will purchase IPTV (odds ratio increases from 0.405 to 1.547) as the price increases.

Similarly, increasing the perception that the client has over the quality of the focal services, significantly enhances the odds that the customer will purchase IPTV (odds ratio increases from 0.118 to 0.405) as the client satisfaction increases.

Finally, with respect to the interaction of two categorical variables, the ODDSRATIO statement would produce the odds ratios at each level of a particular interacting covariate, which would make their interpretation extremely complex. Consequently, the odds ratios for the variables *Employees* (four categories), *Duration* (four categories) and *AccountMix* (three categories), which interact in a two-way mode between themselves as well as with respect to the binary variables *Promotion* and *Bundle*, are out of the scope of interpretation in the present study.

The face validity of these relationships enhance the general confidence in the predictive validity of the model.

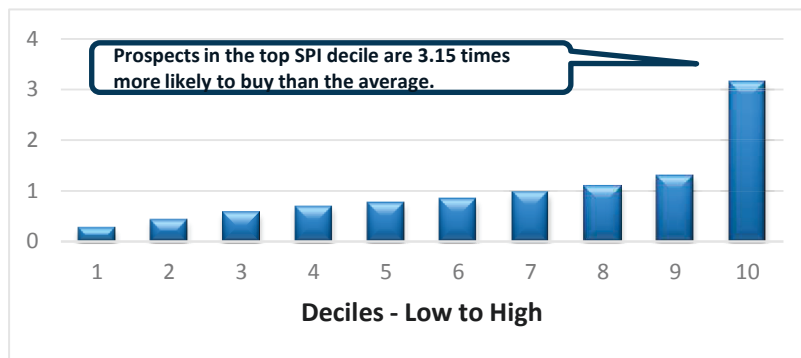
5.2. Model B

This section presents the performance of the model built for acquiring new customers to adopt IPTV. As only one model was developed during the ten-month term, the lift curve and AUC on scoring data are presented.

5.2.1. Model Scoring

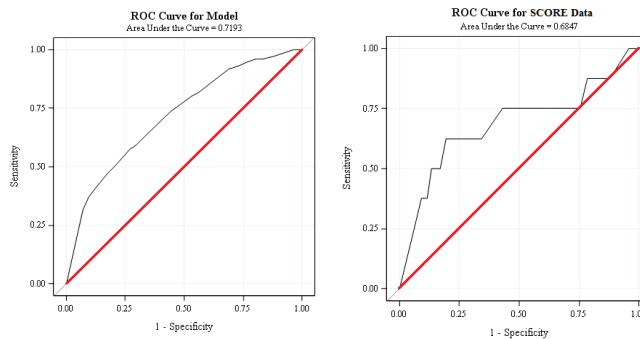
The model addressing acquisition of new IPTV customers was scored on data from the month of January 2015. The lift with prospect opportunities reached the value of 3.15 as shown in Figure 5.7.

Figure 5.7 Lift Curve for Model B



Both areas under the curve for Model B, at the estimation and scoring phases, are shown graphically in Figure 5.8. As it can be seen, the ROC for estimating the model displays a soft curve, while the scoring curve has a step like form.

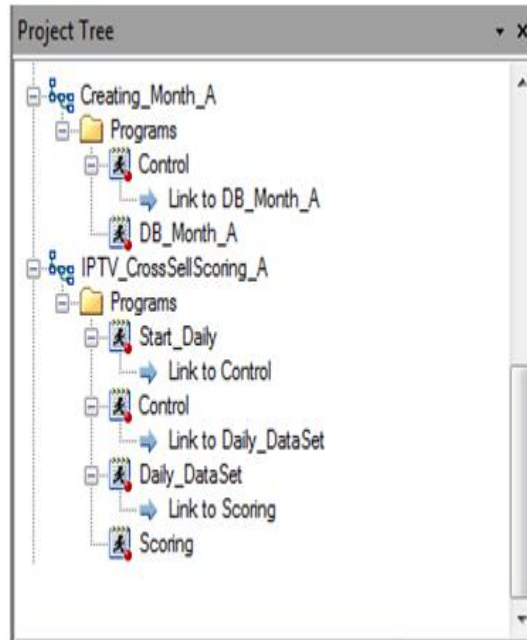
Figure 5.8 ROC Curves Model B



5.3. BI Process Integration

The process to run the propensity index on a monthly basis is depicted in Figure 5.9. The trees for Model A include creating the data sets for scoring and applying the model to obtain the list of clients with their cross-selling propensity index.

Figure 5.9 Project Trees in SAS/EG for Model A



The model was saved with the INMODE option so that it can score the customer data base on a monthly basis.

Chapter 6.

Conclusion and Discussion

6.1. Conclusions

This chapter discusses the findings emerging from the statistical analysis presented in the previous chapters.

The best performing model on growth was Model A [W]. The model considers factors associated with the client's current services, from the customer's perspective as in a utility function. It examines the specific factors that firms may use to enhance the market matching process. The perception of service quality, price, and satisfaction are considered as single variables and coupled variable interactions. The interaction of categorical variables in the complex B2B service environment proves to be helpful and provides findings suggesting that the small firm's assessment of service quality will be positively biased for satisfied decision makers and negatively biased for those who are dissatisfied.

At the supplier's level, it is proposed that microenterprises are influenced by current service quality and price associated to the currently subscribed services. At the customer's level, it is proposed that the small businesses are influenced by decision-maker perceptions of the buyer-seller relationship, especially assessments of satisfaction and quality of service (QoS) of currently subscribed services (phone and/or Internet) and criticality of the small business's TV service needs, which may be considered as industry-related, as discussed in §4.2.2.2.

Furthermore, the model takes into account the location, the size, the duration as a client, the number of accounts, the use of bundled services, and the susceptibility to promotional offers.

Cross-selling is a valuable technique used in sales to increase order size and to transform single-product buyers into multi-product or multi-service clients. It has evolved into a strategy for customer development or growth within the context of CRM, directed toward increasing the firm's share of the customer spending, enlarging the ambit of the relationship with the customer, as well as gaining customer retention.

At the other end, being the first step and foundation of the whole CRM process, customer acquisition is considered as involving more costs than developing existing customers. Indeed, as the meager results for Model B indicate, in order to obtain significant insights regarding customers' choices, a further study would be required. This could be accomplished via a survey, in which the specific needs or criticality as well as the levels of satisfaction with the current supplier may be collected. Moreover, obtaining data from competitors could be a substantial improvement to provide a better suited model with the purpose to cross-sell and to acquire new subscribers. Information such as the competitors' prices and the additional features they provide would increase the number of appealing explanatory variables.

Even if information technology is necessary, it is not sufficient for effective implementation of growth and acquisition strategies. Several organizational barriers might prevent the successful implementation of such strategies. To mention one, the traditional organizational structure along product or service lines may create a "silo" mentality within the departments. Thus, managers of one product or service do not feel responsible for other products or services without realizing that they are all serving the same customers. Database segregation arises, and managers on one side of the corporation ignore the relationship between their clients and other units of the firm. This inevitably prevents any use of acquisition analysis and thus a collaborative cross-selling strategy. As explained at the beginning of this research, what would be the basics of CRM is the client-centric approach. Therefore, this avenue starts with a thorough analysis of the customer's needs.

The analytical tool for cross-selling and acquiring customers are only instruments that facilitate the implementation of growth and acquisition strategies. But three major technological components of the firm must be in place prior to effectively implementing

such strategies. First, managers must have already a comprehensive customer database with detailed activities of each customer in order to apply analytical tools. Second, if a data warehouse does not exist at the corporate level, a department's data mart should at least be in operation to provide managers with a full view of each customer across all points of contact. In this way, service representatives could access relevant information to identify cross-selling opportunities while communication is under way. And third, managers must consider customer database management as a tool not only for operational and functional capabilities, but also as a strategic competence. Therefore, data analysts should be allowed to study the accumulated data. This analysis should be viewed as a fundamental task in CRM research, in addition to the well-known capability to report key performance indicators (KPIs). Thinking "outside the box" is only feasible if analytic tools are adopted to get insights and produce useful models. Otherwise, the firm will be left behind, treating vital data to generate isolated reports.

The problems arising due to an ineffective organization of unstructured data during the creation of a data mart could be resolved by changing the process of gathering the data. A suggestion to create a consolidated database would contemplate offering clients to opt in for bundled services. This is a convenient approach to collect data for all services subscribed by a single and unique client. And it is the easiest way to build a reliable customer data base from the very start, at the data entry level. At the same time, offering bundled services has the advantage of serving the customer with up-to-date features, including special prices and quality of service improvements related to the network, i.e. security, protection, and service hours.

6.2. Discussion and Further Work

Based on the model analysis performed with the limited data, it can be concluded that there are several internal and external factors contributing to the phased adoption of IPTV by small businesses and microenterprises. These include: the duration as a client, number of accounts, add-in features on current services, price, satisfaction and service quality perceptions, as well as susceptibility to promotional offers, all at the internal level; and business industry, size, location, and credit and financial indexes, at the external level.

The limitations of the present research include the weak proportion of converted clients against non-converted customers for both cross-selling and acquisition models. The data sets on both intervals contain non-independent data within the covered months because of the existence of the same client, which may occur up to the maximum number of months in each data set, i.e. six or ten times. The limitation of having missing values at the external database on firm demographics also weakens the validity of the models since these values forced the procedure to either remove the observations from the data sets, or impute the values only for the non-converted clients, i.e. with the target variable set to “0”. This would add a bias to the obtained results.

The implications of the present study are explained next. Suppliers compete on the basis of quality as well as price, and customers are better served by effective competition. According to the best cross-selling model developed in this study, A [WI], the “serve, and then sell” philosophy holds and must be maintained for the cross-selling strategy to work.

The following research questions, however, remain to be answered.

1. How would loyalty factor in the likelihood of cross-selling to current clients and acquiring new customers?
2. How would data from the competition, including penetration rates, pricing, and featured options factor in the likelihood of cross-selling to current clients and acquiring new customers?

The following are suggestions to obtain information – variables – pertaining to loyalty with the aid of a survey:

- Price sensitivity through questioning the commitment to the supplier if an alternative supplier offered a 10% price discount. Customers lowering their commitment levels if there was a lower price available from another supplier are considered less loyal.

- Tracking customers (longitudinal modeling) acquired through low price strategies, such as via a price promotion, to confirm they would be less loyal due to having adopted the service under promotional circumstances, and study expected tenure and churn rate.
- Request customers to identify the “best firm” in the telecommunications industry and whether it was their current supplier. Usually customers indicating that a firm other than their current supplier was “best” had been acquired primarily through promotions, even though the customer preferred another supplier.
- Customer expectations are usually formed relative to competitive alternatives among suppliers. Therefore, a competitor’s actions, claims, and promises can directly influence a customer’s expectations, and as a result, the customer’s decision making on cross-buying or adopting a service for the first time can be affected.

The above are suggestions for further analysis in the modeling and strategic roles of CRM research.

References

- Ahn, H., Han, S. & Lee, Y. (2006). Customer Churn Analysis: Churn Determinants and Mediation Effects of Partial Defection in the Korean Mobile Telecommunications Service Industry. *Telecommunications Policy*, 38, pp. 552-568.
- Ahn, H., Ahn, J.J., Oh, K.J. & Kim, D.H. (2011). Facilitating Cross-Selling in a Mobile Telecom Market to Develop Customer Classification Model based on Hybrid Data Mining Techniques. *Expert Systems with Applications*, 38(5), pp. 5005-5012.
- Akura, M. & Srinivasan, K. (2005). Research Note: Customer Intimacy and Cross-Selling Strategy. *The Journal of Media Economics*, 51(6), pp. 1007-1012.
- Albert, A. & Anderson, J.A. (1984). On the Existence of Maximum Likelihood Estimates in Logistic Regression Models. *Biometrika*, 71, pp. 1-10.
- Allison, P.D. (2012). Logistic Regression Using SAS: Theory and Application. SAS Institute. Second Edition.
- Atkin, D.J., Neuendorf, K. & Jeffres L.W. (2003). Predictors of Audience Interest in Adopting Digital Television. *Management Science*, 16(3), pp. 159-173.
- Benhima, M, Reilly, J.P., Naamane, Z., Kharbat, M, Kabbah, M.I. & Esqalli, O. (2013). Design and Implementation of a Telco Business Intelligence Solution using eTOM, SID and Business Metrics: Focus on Data Mart and Application on Order-to-Payment End to End Process. *IJCSI International Journal of Computer Science Issues*; Vol. 10, Issue 3, No. 1, May 2013, pp. 331-355.
- Blattberg, R.C., Kim, P., Kim B.D. & Neslin, S.A. (2008). Database Marketing: Analyzing and Managing Customers. London, Springer Verlag.
- Bolton, L. & Alba, J. (2006). Price Fairness: Good and Service Differences and the role of Vendor Costs. *Journal of Consumer Research*, 33(2), pp. 258-265.
- Bolton, R.N., Lemon, K.N. and Verhoef, P.C. (2008). Expanding Business-to-Business Customer Relationships: Modeling the Customer's Upgrade Decision. *Journal of Marketing*, 72, January 2008, pp. 46-64.
- Bose, I., & Chen, X. (2009). Quantitative Models for Direct Marketing: A Review from Systems Perspective. *European Journal of Operational Research*, 195, pp.1-16.

- Burez, J. & Van den Poel, D. (2007). CRM at a Pay-TV Company: Using Analytical Models to Reduce Customer Attrition by Targeted Marketing for Subscription Services. *Expert Systems with Applications*, 32(2), pp.277-288.
- Burez, J. & Van den Poel, D. (2008). Separating Financial from Commercial Customer Churn: A Modeling Step towards Resolving the Conflict between the Sales and Credit Department. *Expert Systems with Applications*, 35(1), pp. 497-514.
- Burez, J. & Van den Poel, D. (2009). Handling Class Imbalance in Customer Churn Prediction. *Expert Systems with Applications*, 36, pp.4626-4636.
- Choi, H., Kim, Y. & Kim, J. (2010). An Acceptance Model for an Internet protocol Television Service in Korea with Prior Experience as a Moderator. *The Service Industries Journal*, 30(11), pp. 1883-1901.
- Coe, J.M. (2001). A Road Map for B-to-B Database Marketing. *Target Marketing*, June 2001; 24, 6; ABI/INFORM Complete, pp. 65-72.
- Coe, J.M. (2004). A Road Map for B-to-B Database Marketing. *Journal of Interactive Marketing*, Spring 2004; 18(2); pp. 62-74.
- D'Haen, J. & Van den Poel, D. (2013). Model-Supported Business-to-Business Prospect Prediction Based on an Interactive Customer Acquisition Framework. *Industrial Marketing Management*, 42(4), pp. 544-551.
- Dickson, P. (1992). Toward a General Theory of Competitive Rationality. *Journal of Marketing*, 56 (January); pp. 69-86.
- Firth, D. (1993). Bias Reduction of Maximum Likelihood Estimates. *Biometrika*, 80, pp. 27-38.
- Folkes, V.S. (1988). Recent Attribution Research in Consumer Behavior: A Review and New Directions. *Journal of Consumer Research*, 14 (March), pp. 548-565.
- Golfarelli, M & Rizzi, S. (2009). Data Warehouse Design. Modern Principles and Methodologies. McGraw-Hill/Osborne. Ch. 1 and 8.
- Gupta, S., Hanssens, D., Hardie, B., Kahn, W., Kumar, V. & Lin, N. (2006). Modeling Customer Lifetime Value. *Journal of Service Research*, 9, pp. 139-155.
- Hague, P. & Harrison, M. (2013). Market Segmentation in B2B Markets. www.b2binternational.com/publications/b2b-segmentation-research.
- Hancock, J.C. & Toren, R. (2006). *Practical Business Intelligence with SQL Server 2005*. Addison-Wesley Professional.
- Hansen, H., Samuelsen, B.M. & Silseth, P.R. (2008). Customer Perceived Value in B-t-B Service Relationships: Investigating the Importance of Corporate Reputation. *Industrial Marketing Management*, 37, pp. 206-217.

- Hansotia, B.J. & Wang, P. (1997). Analytical Challenges in Customer Acquisition. *Journal of Direct Marketing*, 11(2), pp. 7-19.
- Heinze, G. & Schemper, M. (2002). A Solution to the Problem of Separation in Logistic Regression. *Statistics in Medicine* 21, pp. 2409-2419.
- Hung, S.Y., Yen, D.C. & Wang, H.Y. (2006). Applying Data Mining to Telecom Churn Management. *Expert Systems and Applications*. 31(3), pp. 515-524.
- Jahromi, A.T., Sepheri, M.M., Teimourpour, B. & Choobdar, S. (2010). Modeling Customer Churn in a Non-Contractual Setting: The Case of Telecommunications Service Providers. *Journal of Strategic Marketing*, 18(7), pp. 587-598.
- Jahromi, A.T., Stakhovych, S. & Ewing, M. (2014). Managing B2B Customer Churn, Retention and Profitability. *Industrial Marketing Management*, Elsevier, <http://dx.doi.org/10.1016/j.indmarman.2014.06.016>, pp. 1-11.
- Jang, H.Y. & Noh, M.J. (2011). Customer Acceptance of IPTV Service Quality. *International Journal of Information Management*, Elsevier, 31, pp. 582-592.
- Kamakura, W.A., Ramaswami, S., and Srivastava R. (1991). Applying Latent Trait Analysis in the Evaluation of Prospects for Cross-Selling of Financial Services. *International Journal of Research in Marketing*, 8, pp. 329-349.
- Kamakura, W.A., Wedel, M., de Rosa, F., and Mazzon, J.A. (2003). Cross-selling through Database Marketing: A Mixed Data Factor Analyzer for Data Augmentation and Prediction. *International Journal of Research in Marketing*, 20, pp. 45-65.
- Kamakura, W., Ansari, A., Bodapati, A, Fader, P., Iyengar, R., Naik, P., Nesli, S, Sun, B., Verhoef, P.C., Wedel, M., Wilcox, R. (2005). Choice Models and Customer Relationship Management. *Marketing Letters*, 16:3/4, pp. 279-291.
- Kim, H. & Yoon, C. (2004). Determinants of Subscriber Churn and Customer Loyalty in Korean Mobile Telephony. *Telecommunications Policy*, 2, pp. 751-765.
- Knott, A., Hayes & A., Neslin, S.A. (2002). Next-Product-To-Buy Models for Cross-Selling Applications. *Journal of Interactive Marketing*, 16(3), pp. 59-75.
- Kumar, V. & Andrew Petersen, J. (2012). *Statistical Methods in Customer Relationship Management*. John Wiley & Sons, Ltd., Ch. 3.
- LaPlaca, P.J. & Katrichis, J.M. (2009). Relative Presence of Business-to-Business Research in the Marketing Literature. *Journal of Business-to-Business Marketing*, 16, pp. 1-22.
- Larocque, D. (2013). Notes de cours: Analyse multidimensionnelle appliqué (6-602-07). HEC Montreal, version 2013.

- Lee, S.J. & Siau, K. (2001). A Review of Data Mining Techniques. *Industrial Management & Data Systems*, 101/1, pp. 41-46.
- Linoff, G.S. & Berry, M.J.A. (2011). *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*. John Wiley & Sons. Third Edition.
- Luo, B., Shao, P. & Liu, D. (2007). Evaluation of Three Discrete Methods on Customer Churn Model based on Neural Networks and Decision Tree in PHSS. *Proceeding of the 1st Inter Symp Data, Privacy and E-Commerce*, Washington DC, pp. 95-97.
- Monat J.P. (2011). Industrial Sales Lead Conversion Modeling. *Marketing Intelligence & Planning*, 29, pp. 178-194.
- Narayandas, D. & Rangan, K.V. (2004). Building and Sustaining Buyer-Seller Relationships in Mature Industrial Markets. *Journal of Marketing*, 68 (July), pp. 63-77.
- Ortiz, S. (2006). Phone Companies Get into the TV Business. *IEEE Computer Society*, (October) pp. 12-15.
- Patterson, P.G., Johnson, L.W., Spreng, R.A. (1997). Modeling the Determinants of Customer Satisfaction for Business-to-Business Professional Services. *Academy of Marketing Science Journal*, 25(1), pp. 4-17.
- Pendharkar, P. (2009). Genetic Algorithm based Neural Network Approaches for Predicting churn in Cellular Wireless Networks Service. *Expert Systems with Applications*, 36, pp. 6714-6720.
- Rygielski, C., Wang, J.C., Yen, D.C. (2002). Data Mining Techniques for Customer Relationship Management. *Technology in Society*, 24, pp. 483-502.
- Seo, D., Ranganathan, C. & Babad, Y. (2008). Two-Level Model of Customer Retention in the US Mobile Telecommunications Service Market. *Telecommunications Policy*, 32, pp. 182-196.
- Shin, D.H. (2007). Socio-Technical Analysis of IPTV: A Case Study of Korean IPTV. *info*, 9(1), pp. 65-79.
- Shin, D.H. & Hwang, Y.S. (2011). Examining the Factors Affecting the Rate of IPTV Diffusion: Empirical Study on Korean IPTV. *Journal of Media Economics*, 24, pp. 174-200.
- Stevens, J.P. (2009). *Applied Multivariate Statistics for the Social Sciences*. Fifth Edition. University of Cincinnati.
- Talus Company. *A Practical Guide to Building Data Marts*. (1998) Wiley, public computer books. ftp://ftp.wiley.com/public/computer_books/updates/guide.pdf.
- Tsai, C.F. & Lu, Y.H. (2009). Customer Churn Prediction by Hybrid Neural Networks. *Expert Systems with Applications*, 36(10), pp. 12547-12553.

- Tsai, C.F. & Lu, Y.H. (2010). Data Mining Techniques in Customer Churn Prediction. *Recent Patents on Computer Science*, 3, pp. 28-32.
- Wei, C. & Chiu I. (2002). Turning Telecommunications Call Details to Churn Prediction: A Data Mining Approach. *Expert Systems with Applications*, 23, pp. 103-112.
- Weniger, S. (2010). User Adoption of IPTV: A Research Model. *23rd Bled eConference eTrust: Implications for the Individual, Enterprises and Society*, June 20-23, 2010, pp. 154-165.
- Zablah, A.R., Bellenger, D.N., Straub, D.W., Johnston, W.J. (2012). Performance Implications of CRM Technology Use: A Multilevel Field Study of Business Customers and Their Providers in the Telecommunications Industry. *Information Systems Research*, 23(2), pp. 418-435.
- Zhao, W., Olshefski, D. and Schulzrinne, H. (2000). Internet Quality of Service: an Overview. *Technical Report CUCS-003-00, Department of Computer Science, Columbia University*, February 2000, pp. 1-11.

Appendix

Final Cross-Selling Equation – Model A [wI]

$$\begin{aligned} \text{Logit} = & -10.0574 + 1.4048 \text{Ind}_{01} + 0.3604 \text{Ind}_{02} + 0.4281 \text{Ind}_{03} + 0.2978 \text{Ind}_{04} \\ & + 0.400 \text{Ind}_{06} + 0.4079 \text{Ind}_{07} - 0.1677 \text{Ind}_{08} + 0.4904 \text{Ind}_{09} - 0.3474 \text{Ind}_{10} \\ & + 0.3125 \text{Ind}_{11} + 0.5809 \text{Ind}_{12} - 0.0528 \text{Ind}_{13} + 0.1238 \text{Ind}_{14} + 0.0175 \text{Ind}_{15} \\ & + 0.6260 \text{Ind}_{16} + 0.1347 \text{Ind}_{17} + 1.0891 \text{Ind}_{18} + 0.3838 \text{Ind}_{19} + 0.2514 \text{Ind}_{20} \\ & - 0.3914 \text{Ind}_{21} + 0.2883 \text{Ind}_{22} + 0.2752 \text{Ind}_{23} + 0.3668 \text{Ind}_{24} + 0.1039 \text{CCI}_1 \\ & - 0.0417 \text{CCI}_2 - 0.2791 \text{CCI}_3 - 0.2070 \text{CCI}_4 - 0.4701 \text{CCI}_5 - 0.1187 \text{Promotion} \\ & + 0.7229 \text{Bundle} - 2.7595 \text{Satisfied} + 3.8446 \text{QoS} + 0.5035 \text{Location} \\ & 97 + 1.1061 \text{LogPrice} - 0.6697 \text{LogPrice} * \text{QoS} + 0.6123 \text{Satisfied} * \text{QoS} \\ & + 0.4007 \text{Employees}_{01} + 0.3410 \text{Employees}_{02} + 0.2462 \text{Employees}_{03} \\ & - 0.6781 \text{Duration}_{01} - 0.0154 \text{Duration}_{02} + 0.2000 \text{Duration}_{03} \\ & + 0.3129 \text{AccountMix}_{01} * \text{Duration}_{01} - 0.1015 \text{AccountMix}_{01} * \text{Duration}_{02} \\ & + 0.3000 \text{AccountMix}_{01} * \text{Duration}_{03} + 1.5783 \text{AccountMix}_{02} * \text{Duration}_{01} \\ & + 0.5137 \text{AccountMix}_{02} * \text{Duration}_{02} + 0.2890 \text{AccountMix}_{02} * \text{Duration}_{03} \\ & - 0.1392 \text{AccountMix}_{01} * \text{Employees}_{01} - 0.5841 \text{AccountMix}_{01} * \text{Employees}_{02} \\ & + 0.1974 \text{AccountMix}_{01} * \text{Employees}_{03} - 0.3025 \text{AccountMix}_{02} * \text{Employees}_{01} \\ & - 0.2944 \text{AccountMix}_{02} * \text{Employees}_{02} - 0.3169 \text{AccountMix}_{02} * \text{Employees}_{03} \\ & - 0.4467 \text{Bundle} * \text{AccountMix}_{01} + 1.2595 \text{Bundle} * \text{AccountMix}_{02} \\ & - 1.1333 \text{Promotion} * \text{Employees}_{01} + 0.9110 \text{Promotion} * \text{Employees}_{02} \\ & - 0.7253 \text{Promotion} * \text{Employees}_{03} - 1.8520 \text{Bundle} * \text{Duration}_{01} \\ & - 0.3354 \text{Bundle} * \text{Duration}_{02} + 0.0989 \text{Bundle} * \text{Duration}_{03} \end{aligned}$$