REINFORCEMENT OF BUSINESS INTELLIGENCE APPLICATIONS IN SRI LANKAN LIFE INSURANCE INDUSTRY

Nayomi Kanchana Wickramasekara

(169141J)

Degree of Master of Business Administration in Information Technology

Department of Computer Science and Engineering

University of Moratuwa

Sri Lanka

December 2017

REINFORCEMENT OF BUSINESS INTELLIGENCE APPLICATIONS IN SRI LANKAN LIFE INSURANCE INDUSTRY

Nayomi Kanchana Wickramasekara

(169141J)

The dissertation was submitted to the Department of Computer Science and Engineering of the University of Moratuwa in partial fulfilment of the requirement for the Degree of Master of Business Administration in Information Technology.

Department of Computer Science and Engineering

University of Moratuwa

Sri Lanka

December 2017

DECLARATION

I declare that this is my own work and this thesis does not incorporate without acknowledgement of any material previously submitted for a Degree or Diploma in any other University or institute of higher studies and does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant the University of Moratuwa the non-exclusive right to reproduce and distribute my thesis/dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

………………………………. ………………………

N K Wickramasekara Date

The above candidate has carried out research for the Masters' thesis under my supervision.

……………………………….. ………………….

Dr A S Perera Date

COPYRIGHT STATEMENT

I hereby grant the University of Moratuwa the right to archive and to make available my thesis or dissertation in whole or part in the University Libraries in all forms of media, subject to the provisions of the current copyright act of Sri Lanka. I retain all proprietary rights, such as patent rights. I also retain the right to use this thesis or dissertation in future works (articles or books etc) in whole or part.

Date

ABSTRACT

Business Intelligence is not a newer technology. Instead, it's an integrated solution for businesses, where business requirements are the key factors that drive technology innovation.

Nowadays Business Intelligence in financial organizations has been implemented and operated mainly to support decision making using knowledge as a strategic factor. Business Intelligence takes a vital role in insurance domain especially in life insurance sector where BI help firms in gaining business advantage mainly in decision making.

In the life insurance industry, using classification techniques on customer and product databases seems to be very effective. One of the best applications where classification can be used in the life insurance industry is for the regularity of life insurance policyholders for instalment payment depending on their behavioural attributes. That is deciding whether a life insurance policyholder is regular or irregular in premium payments by considering his or her behavioural attributes such as their demographic, social, cultural and economic data.

So in order to achieve the objective of this research, which is reinforcing business intelligence applications in Sri Lankan life insurance industry, primary data of 400 life insurance policyholders have been collected from different life insurance companies in Sri Lanka, considering the regularity of policyholders' premium payments. Five different classification techniques such as Naïve Bayes, Multi-Layer Perceptron, IBK, PART and SMO, which have been identified as most significant in classifying regularity of policyholders' premium payments, have been applied on primary data, in order to decide whether life insurance policyholder is regular or irregular in premium payments. Finally, those five classification techniques have been evaluated using evaluation techniques in order to come up with the best BI model in classifying regularity of policyholders' premium payments for Sri Lankan life insurance industry.

Key words: Business Intelligence, Naïve Bayes, Multi-Layer Perceptron, IBk, PART, SMO

ACKNOWLEDGEMENT

First of all, my deepest gratitude expressed to Dr A.S. Perera, my supervisor for this research, for letting me to take this research project and guiding me from the beginning. Sir, your valuable comments and encouragements drive me to a successful endpoint of this research.

I appreciate the guidance provided by Dr Chandana Gamage, Ms Jeeva Padmini, who gave their valuable advice, criticisms to see the best from this research.

Finally, it is a pleasure to thank my parents, my sister, colleagues and all supporters who helped me to succeed in this research with their valuable support and advice.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

A/L - Advanced Level

AUC - Area Under the Curve

BI - Business Intelligence

CRM - Customer Relationship Management

FP - False Positive

HNB - Hatton National Bank

IBk - Instance Bases learning with parameter k

ID3 - Iterative Dichotomiser 3

IT - Information Technology

KNN - K-nearest neighbours

MLP - Multi-Layer Perceptron

O/L - Ordinary Level

PART - Projective Adaptive Resonance Theory

QP- Quadratic Programming

ROC - Receiver Operating Characteristics

SMO- Sequential Minimal Optimization

SVM- Support Vector Machine

TP - True Positive

WEKA - Waikato Environment for Knowledge Analysis

1. INTRODUCTION

1.1 Background

The service sector of the world economy has grown substantially since World War II. The worldwide insurance industry has had an average annual growth rate of over 10 per cent since 1950. During the mid-1980s, the international life insurance industry grew at an average annual rate greater than 25 per cent.

Life insurance provides individuals and the economy as a whole with a number of important financial services. It is an instrument to manage income risk by providing coverage against income loss from death, as well as an investment vehicle for long-term savings. Much empirical research has shown the overwhelming positive relationship between insurance industry development and economic growth. A United Nation's report has acknowledged that a sound national life insurance and reinsurance market are an essential characteristic of economic growth. In recent decades, the life insurance sector has grown in economic importance since it forms an essential component of the global financial market.

Sri Lankan life insurance industry is still a growing market with relatively low penetration levels where awareness of the importance of life insurance is also lower than in more developed markets. But as the population becomes increasingly wealthier, delivery channels such as bancassurance and online insurance are growing in popularity and show major potential for development. Sri Lankan life insurance market is extremely competitive with 15 players including three international companies, where only four large players currently dominate the market which features the Oligopoly market structure (Insurance Regularity Commission of Sri Lanka, 2013).

The Insurance Board of Sri Lanka has been established for the purpose of development, supervision and regulation of the insurance industry and to ensure that insurance business in Sri Lanka is carried out with integrity and in a professional and prudent manner, with a view to safeguarding the interests of the policy-holders and potential policyholders. From time to time several amendments have been incorporated into the principal enactment and subordinate legislation made to strengthen the insurance industry. The Board is also in the process of introducing the Risk-Based Capital Model for insurance companies under the Act (Insurance Regularity Commission of Sri Lanka, 2013).

A Business Intelligence System is a key component of a company's IT framework. It is the component that enables business users to report on, analyze and optimize business operations to reduce costs and increase revenues. Most companies use this component for strategic and tactical decision making where the decision-making cycle may span a time period of several weeks or months. Competitive pressures, however, are forcing companies to react faster to changing business conditions and customer requirements. As a result, there is now a need to use Business Intelligence to help drive and optimize business operations on a daily basis, and, in some cases, even for intraday decision making.

The concept of Business Intelligence was brought up by Gartner Group in 1996. Business intelligence is an umbrella term that includes the applications, infrastructure and tools, and best practices that enable access to and analysis of information to improve and optimize decisions and performance (Wikipedia, 2017).

BI technologies provide historical, current and predictive views of business operations. Common functions of business intelligence technologies include reporting, online analytical processing, analytics, data mining, process mining, complex event processing, business performance management, benchmarking, text mining, predictive analytics and prescriptive analytics. BI technologies can handle large amounts of structured and sometimes unstructured data to help identify, develop and otherwise create new strategic business opportunities. They aim to allow for the easy interpretation of these big data. Identifying new opportunities and implementing an effective strategy based on insights can provide businesses with a competitive market advantage and long-term stability.

Generally, these systems will illustrate business intelligence in the areas of customer profiling, customer support, market research, market segmentation, product profitability, statistical analysis, and inventory and distribution analysis to name a few.

IT systems support three main types of application processing: business transaction processing, BI processing and collaborative processing. Business transaction processing drives day-to-day business operations and supports business activities such as order entry, inventory control, shipping, billing and so forth. BI processing reports and analyzes business transaction processing, and provides information about how well this processing is meeting business requirements. Business users employ the output produced by BI applications to optimize business transaction processing to more closely match business goals and requirements. This optimization process involves discussions between business experts about possible ways of improving business processes. The interaction between these business users is enabled by collaborative processing.

The key challenge of Business Intelligence is how to use technologies to creatively address major business issues and achieve operation effectiveness where BI technologies could enable the development of cutting-edge business models and achieves great business impact. A successful BI application could significantly help the transformation of the insurance selling approach from product-oriented to customer needs-oriented and create tremendous market/customer value.

1.2 Motivation

The major problem identified in Sri Lankan Life Insurance Industry is the intense competition between insurers for existence as well as for new entrants to the market though there are steady regulations in the insurance market. A weak relationship among insurers, reinsures & brokers too have become a problem in Sri Lankan life insurance sector. To survive & remain viable in the life insurance sector, it is necessary to identify new ways to increase their own market shares.

Furthermore, customer expectations are becoming complex nowadays and they are expecting high personalization for their services. So the ultimate results have lead to a new economic paradigm centred on the customer and this new paradigm will induce many pressures on insurers such as pressure on capital, pressure on sales volume etc.

When companies implement new technologies, the main goal comes as innovation where some technologies may present innovation in the way of price, while other innovations may assist in user-friendliness or a boost in productivity. So the emergence of new technologies such as Business Intelligence, Data mining etc in global arena will become a challenge to Sri Lankan life insurance sector in finding innovative ways of how to use these new technologies in order to increase user-friendliness, productivity or their product prices.

1.3 Research Scope

Business intelligence has long been cited for the benefits it can deliver to organizations. Like any new technology it requires capital investment, but in the long-term, the benefits far outweigh the expenditure and can help to move the business forward into the era of real-time decision-making ability.

By utilizing operational data as part of the BI arsenal, organizations are able to gain insight into near real-time data about key performance parameters that affect order volumes, inventory levels, employee productivity etc, thus allowing for more effective decision-making. However, while the advantages of operational BI may be numerous and clear, proper implementation of such a solution is complex and poses many challenges to the organization.

One aspect towards creating and implementing successful operational BI is to conduct a proper business analysis. The traditional gap between business and IT needs to be closed for such an initiative, as IT needs to work closely with business to understand and identify data and business requirements. The business also needs to align with IT to understand and realize the benefits of this to the organization and any sort of technological implementation.

Another challenge is the need to draw data from a wider range of sources than traditional BI, which makes the need for the coveted "single view" of the organization even more vital to avoid duplication and wasted processing time.

This research will be mainly focused on Sri Lankan Life Insurance Industry where the main objective will be to propose suitable business intelligence applications that are applicable to Sri Lankan life insurance industry which will reinforce the Sri Lankan life insurance industry.

1.4 Problem Statement

Traditional business intelligence has its usefulness worth as a tool for providing information on strategic planning and high-level decision-making. There is an increasing need for BI to be extended across the wider enterprise to incorporate operational data and allow for accurate operational decision-making. Most organizations now employ some form of BI, and this is no longer a source of significant competitive advantage. Operational BI is emerging as a tool to once again create a real competitive edge for organizations, empowering executives, line managers and other business professionals across the enterprise by leveraging the power of information at the operations level.

So this research will mainly be looking to how Business Intelligence applications will impact to the reinforcement of Sri Lankan Life Insurance Industry such as how operational BI helps frontline workers and operational managers to access relevant information more quickly, improving efficiency where operational BI can also reduce costs, improve proactive decision-making and provide a significant competitive advantage.

1.5 Research Objectives

Following are the ultimate goals of this research:

- Identify business intelligence applications used in the global insurance industry
- Identify the use of business intelligence in Sri Lankan life insurance industry
- Propose a suitable business intelligence application that is applicable to Sri Lankan life insurance industry

In order to achieve the above objectives, first of all, it is required to study global insurance industry trends. After achieving the main objective of this research which is proposing a suitable business intelligence application that is applicable to Sri Lankan life insurance industry, it is needed to evaluate that model to Sri Lankan Life Insurance Industry as well.

2. LITERATURE REVIEW

2.1 Global Life Insurance Industry

Insurance is protecting against uncertainties. It is a protection against financial loss arising from the happening of an unexpected event. Insurance companies collect a premium to provide security for the above-mentioned purpose.

Insurance may be described as a mechanism for transferring risks so that the losses suffered by a few members of a group is borne by the contributions (or premiums) of the many. The first function of insurance is to transfer risk by replacing uncertainty with certainty. The uncertainty as to whether a loss may occur and if so how much it will cost is replaced by paying a known fixed amount in advance that is the premium. The second function is to establish a common fund from which losses will be paid. The third important function is to have a method of providing a fair contribution (premium) to be paid by all those who are members of the fund (the insured). The members' (the insured’s') contribution ought to be fair in relation to the degree of risk and the value of such risk that the member brings into or exposes the fund.

Life insurance is an appropriate financial tool for managing and mitigating the financial risk associated with untimely death. The risk is inseparable from life and nobody is exempt from it. Obviously, some people are exposed to greater risks than others. To a greater or lesser extent the risk to life and property due to natural perils, such as flood, storm & tempest and earthquake, and manmade perils such as theft and those arising from the negligence of others as well as ourselves, are some of the more common that is constantly with us. However, Life Insurance decisions are often complex. The choice of a life insurance product for a consumer is now a problem of plenty.

Life insurance market of transition economies had experienced a rapid growth over the last decade, indicating the increased importance of this sector as a financial intermediary. A key decision the individuals or families take is whether to buy life insurance or not. The reason behind considering such a decision is to protect against possible loss of income. Life insurance provides individuals and the economy as a whole with a number of important financial services. In the face of escalating urbanization, the mobility of the population and formalization of economic relationships between individuals, families and communities, life insurance has taken increasing significance as a way for individuals and families to manage income risk. Also, life insurance products encourage long-term savings and the re-investment of substantial sums in private and public sector projects.

However, with the deregulation of the financial services market, greater competition is entering the life insurance market. The growing popularity of term life insurance, which is less dependent on its price structure on the detailed medical histories and other information that is used in the life insurance industry, has further placed pressure on the life insurance industry to develop more efficient and cost-effective system and methods for distribution. This effect has only intensified as the number of term life insurance carriers who use modem communications methods such as the Internet to attract and interact with potential purchasers of insurance increases.

The steady deregulation of the life insurance market, the emergence of new technologies, increasing competition from existing and new entrants, are all resulting in a new economic paradigm centred on the customer. The new paradigm will induce many pressures on insurers. Some of the more important ones will be:

- Pressure on capital
- Pressure on volumes
- Pressure on the use of Information Technology

2.2 Sri Lankan Life Insurance Industry

The Life Insurance Industry in Sri Lanka is growing rapidly and it has become more and more important to keep pace with the growth of the industry through technological advancements and innovative ideas to market the organization to the masses. Portfolio of products offered by life insurance providers has diversified, over the years, attracting more customers than ever. Accumulation of operational data inevitably follows from this growth in the industry.

To survive & remain viable from intense competition in the life insurance sector, it is necessary to identify new ways to increase market share. Further Customer expectations are becoming higher and higher and they expect high personalization, online access to account information, online need analysis, online premium payments & policy administration and online claims initiation. High policy administration cost is another issue faced by insurance companies. A weak relationship among insurers, reinsurers & brokers is another problem. Global trends in the insurance sector are unmatchable with local services. If local companies do not identify these trends in incorrect time they have to face a lot of business difficulties. There exists an increasing need to convert their data into a corporate asset in order to stay ahead and gain a competitive advantage.

2.3 Determinants of Life Insurance Demand

Redzuan, H. (2011) confirms that factors of an economic and financial nature strongly stimulate the demand for life insurance. Higher education, employment by someone else, income, number of dependents and better perception about insurance firms have improved the chances of taking life insurance. However, Redzuan says that demographic and social factors such as age, level of education and economic factors such as life insurance price and inflation have had a negative relationship with the odds of taking life insurance.

Sliwinski et al (2013) have performed a study on determination of life insurance demand in Poland. The study has confirmed that factors of an economic and financial nature strongly stimulate the demand for life insurance. However, education level and social benefits have had a negative impact on life insurance demand.

Sarkodie et al (2015) have looked into determinants of life insurance demand of Ghana. They say that income, higher education, number of dependents, employment by someone else and better perception about insurance firms have improved the chances of taking life insurance. Age, however, has had a negative relationship with the odds of taking life insurance. The number of dependents has been statistically significant at 1%. Age and type of employment both have been significant at 5% while income and education level have been significant at 10%.

Çelik and Kayali (2009) have found a positive relationship between income and odds of taking insurance. Moreover, Çelik and Kayali say that higher education has negatively influenced the odds of taking life insurance.

Rahman et al (2017) have selected a set of attributes that would represent a life insurance policy. They have included demographic information such as age, occupation, gender and policy details such as agent, the term of the policy, sum assured, premium etc. They have constructed another attribute to represent the number of other policies a policyholder may have with the company.

2.4 Business Intelligence

Business Intelligence Systems are referred to as an integrated set of tools, technologies and programme products that are used to collect, integrate, analyze and make data available. These systems are to support decision-making on all management levels. They differ from traditional Management Information Systems by a wider subject range, multivariate analyses of semi-structured data that come from different sources and their multidimensional presentation. The BI systems contribute to optimizing business processes and resources, maximizing profits and improving proactive decision-making.

Business intelligence and knowledge discovery are the most common academic disciplines for data mining. There are several types of data mining models such as Association, Classification, Clustering, Forecasting, Regression, Sequence Discovery, Visualization etc.

Classification is a process of finding a model (or a function) that describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. The derived model will be based on the analysis of a set of training data (data objects whose class label is known). Naïve Bayes, IBK, SMO, Multilayer Perceptron and PART are some examples of Classification Techniques.

- Naive Bayes Classifier

  The Naïve Bayes classifier applies to learn tasks where each instance $\boldsymbol{x}$ is described by a conjunction of attribute values and where the target function $\boldsymbol{f}(\boldsymbol{x})$ can take any value from some finite set $\boldsymbol{V}$. When to use moderate or large training set available, attributes that describe instances are conditionally independent given classification.

- Multi-Layer Perceptron

  Multi-Layer Perceptron is a feed-forward neural network with one or more layers between the input and output layer. Feedforward means that data flows in one direction from input to the output layer (forward). This type of network is trained with the back propagation learning algorithm.

- Instance Bases learning with parameter k

  Instance Bases learning with parameter k is a K-nearest neighbour classifier. IBk's KNN parameter specifies the number of nearest neighbours to use when classifying a test instance and the outcome is determined by majority vote. An appropriate value for K can be selected based on cross-validation. The distance is calculated using distance measures such as Euclidean distance, Minkowski distance or Mahalanobis distance.

- Projective Adaptive Resonance Theory

  The input for Projective Adaptive Resonance Theory algorithm is the vigilance and distance parameters. The basic architecture of PART is similar to ART neural networks which have been shown to be very effective in self-organizing stable recognition codes in real time in response to arbitrary sequences of input patterns.

- Sequential Minimal Optimization

  Sequential Minimal Optimization solves the Support Vector Machine Quadratic Programming problem by decomposing it into QP sub-problems and solving the smallest possible optimization problem, involving two Lagrange multipliers, at each step.

Model Evaluation is an integral part of the model development process. It helps to find the best model that represents data and how well the chosen model will work in the future. Evaluating model performance with the data used for training is not acceptable in data mining because it can easily generate over-optimistic and over fitted models.

- Cross-Validation

  When only a limited amount of data is available, to achieve an unbiased estimate of the model performance, use k-fold cross-validation. In k-fold cross-validation, it divides the data into k subsets of equal size. Then build models k times, each time leaving out one of the subsets from training and use it as the test set.

- Confusion Matrix

  A confusion matrix shows the number of correct and incorrect predictions made by the classification model compared to the actual outcomes (target value) in the data. The matrix is NxN, where N is the number of target values (classes).

  Accuracy - the proportion of the total number of predictions that were correct

  Positive Predictive Value or Precision - the proportion of positive cases that were correctly identified

  Negative Predictive Value - the proportion of negative cases that were correctly identified

  Sensitivity or Recall - the proportion of actual positive cases which are correctly identified

  Specificity - the proportion of actual negative cases which are correctly identified

- True Positive

  A true positive test result is one that detects the condition when the condition is present.

- False Positive

  A false positive test result is one that detects the condition when the condition is absent.

- Receiver Operating Characteristic

  Receiver Operating Characteristic graph is a technique for visualizing, organizing and selecting classifiers based on their performance. In a ROC curve, the true positive rate (sensitivity) is plotted in function of the false positive rate (1- specificity) for different cut off points.

- Area Under the Curve

  Area Under the Receiver Operating Characteristic curve is often used as a measure of the quality of the classification models. A random classifier has an area under the curve of 0.5, while AUC for a perfect classifier is equal to 1. In practice, most of the classification models have an AUC between 0.5 and 1.

2.5 Business Intelligence and Global Life Insurance Industry

Life insurance industry needs direct contacts with their customers in order to provide services to satisfy their customers' where they will continue in buying life insurance policies, and/or to renew their life insurance products. This is the main source of revenue for the life industry. Due to its specialized nature, life insurance industry needs up-to-date information in order to modify their products and services to attract potential customers. The best source of data is a market survey, where results may provide information about customers' needs, their purchase intentions, service requirements etc.

Some of the tactical issues faced by life insurance sector nowadays are understanding customer retention patterns, better understanding their claim patterns and identifying types of policyholders who are at more risk etc. These problems could affect the amount of premium and subsequently profitability in the industry where Business intelligence can be used as a solution by using a variety of BI techniques.

Mosley (2012) has discussed the application of correlation, clustering, and association in the insurance industry by analyzing social media such as Twitter posts. He says the results of these analyses could help to identify keywords and concepts in the social media data and can facilitate the application of this information by insurers. Further, he says as insurers analyze this information and apply the results of the analysis in relevant areas, they will be able to proactively address potential market and customer issues more effectively.

Umamaheswari & Janakiraman (2014) have summarized the role of data mining in the Insurance Industry as in Table 2.1. Umamaheswari & Janakiraman have analyzed patterns under four data mining techniques such as clustering, classification & prediction, association and summarization.

Mohapatra & Tiwari (2009) says that data mining techniques combined with business intelligence could be used within holistic frameworks which include customer relationship management, human resource management, claim management, channel management and asset management for life insurance industry where it has been found that these frameworks help life insurance decision making process by making it faster and effective. Table 2.2 is a summarization of areas in the life insurance business process where business intelligence could be best used for decision making.

Devale & Kulkarni (2012) have introduced different exhibits for discovering knowledge in the form of association rules, clustering, classification and correlation suitable for data characteristics in the life insurance sector. Further, they say that an understanding of probability and statistical distributions is necessary to absorb and evaluate acquiring new customers, retaining existing customers, performing sophisticated classification and correlation between policy designing and policy selection.

Table 2.1: Role of data mining in the Insurance Industry

Data Mining Technique	**Pattern**
Clustering	Customer having similar characteristics
	Analysis of customer attrition in the insurance sector
	Policy most likely to be used, most unlikely to be used
	Segments related to policy
Classification & Prediction	Predicting consumer behaviour
	Predicting the likelihood of success of policies
	Classifying the historical customer records
	Prediction of what type of policy most likely to be retained, most likely to be left
	Predicting insurance product behaviour and attitude
	Predicting the performance progress of segments throughout the performance period
	Prediction to find what factors will attract new avenues in the Insurance sector
	Classify trends of movements through the organization for successful/unsuccessful customer historical records
Association	Discovery of such association that promotes the business technique
Summarization	Provides summary information
	Various multidimensional summary reports
	Statistical summary information

Note: Reprinted from "Role of Data mining in Insurance Industry" by K. Umamaheswari & S. Janakiraman. 2014, An international journal of advanced computer technology, 3(6), 961-966.

Devale & Kulkarni says clustering technique could be used to acquire new customers in which the first cluster specifies the group of customers holding life security policy while the second group holds customer for Tax benefit policy and the third group is for those customers holding policy for investment. For an example what they suggest is that when an agent approaches a particular customer, the agent could enter the demographic data of that customer in terms of age, occupation, income and education. Then each individual factor could be compared with the means of each cluster and the difference will be calculated. After comparing each different for each group, the closest cluster could be finalized which has the least difference.

According to Devale & Kulkarni, classification too can be used to targeting life insurance customers or designing new life insurance products. What they say is that normally classes can be created according to policy term, premium mode and premium amount based on age, income and occupation where policy term can be decided according to age and occupation while premium mode and the premium amount can be according to income and occupation. So particular class for the particular customer could be created where policy term, premium mode and the premium amount can be mentioned in it in terms of percentage.

Adding more to above mentioned techniques, Devale & Kulkarni suggest correlation to identify the relation between policy designing and selection factors where to assign numbers to different policy designing and selection factors such as 1 for life security, 2 for investment and 3 for tax benefit etc. and then while analyzing previous policies suggesting respective numbers both for policy designing and selection factors. Using those information life insurance companies would be able to find increasing or decreasing trends between two factors.

Xu et al (2005) have demonstrated how innovative BI infrastructure and application could effectively address major business challenges in China life insurance industry and achieve operational excellence and business impact. It is an IT infrastructure of a newly established for a life insurance brokerage company, which provides professional life insurance consulting service to mass affluent customers. Employees in different departments of the company have been able to enjoy a various level of BI applications based on their roles/responsibilities. That is a set of powerful BI applications have been built based on this infrastructure and multiple well-established databases, to provide the services to BI users such as Life insurance consulting service for insurance consultants, CRM services for Business Development and Customer Service Department, Sales management services for Sales Management Department, Market/product analysis services for Research Department, Management and decision supporting services for Management etc as in Figure 2.1 and Figure 2.2.

Table 2.2: BI applications in Life Insurance business process

Business process	**BI application**
Customer relationship management	Customer Profitability
	Customer Lifetime Value
	Customer Segmentation
	Attrition Analysis
	Affinity Analysis
	Target Marketing
	Campaign Analysis
	Cross-Selling
Channel management	Agent and Sales Force Deployment
	Agent Development and Relationship Management
	E-Business Development
Actuarial	Risk Modeling
	Re-insurance
	Profitability Analysis
Underwriting and Policy Management	Premium Analysis
	Loss Analysis
Claims Management	Claims Analysis
	Fraud Detection
	Claims Estimation
Finance and Asset Management	Budgeting
	Asset Liability Management
	Financial Ratio’s Analysis
	Profitability Analysis
	Web Reporting and Analysis
Human Resources	Human Resource Reports/ Analytics
	Manpower Allocation
	HR Portal
	Training and Succession Planning
Corporate Management	Dashboard Reporting
	Statutory Reporting
	Customer Information Services

Note: Reprinted from "Using Business Intelligence for Automating Business Processes in Insurance" by S. Mohapatra & M. Tiwari. 2009, IJACT: International Journal of Advancements in Computing Technology, 1(2), 92-98.

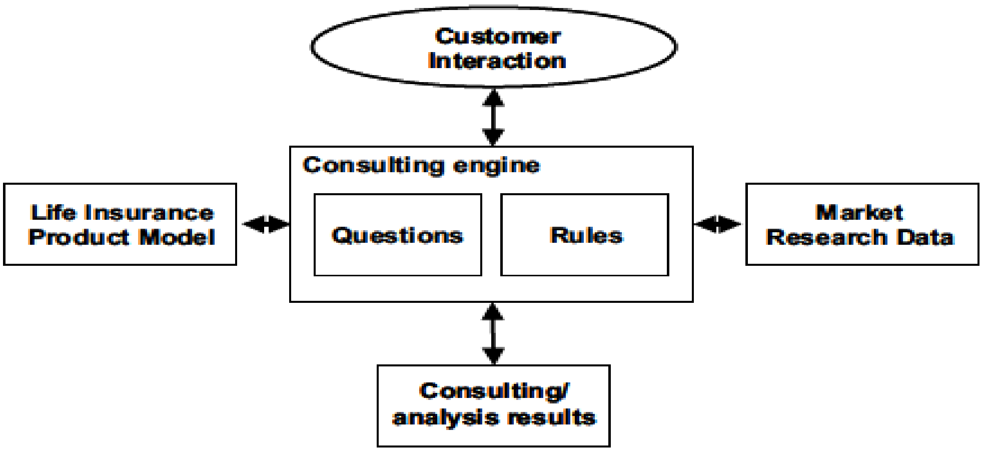

Figure 2.1: A Consulting Engine for Life Insurance Market. Reprinted from "Business intelligence-a case study in the life insurance industry" by Z. Xu, M. Zhang & X. Jiang, 2005. In *e-Business Engineering, 2005. ICEBE 2005. IEEE International Conference on* (pp. 129-132). IEEE.

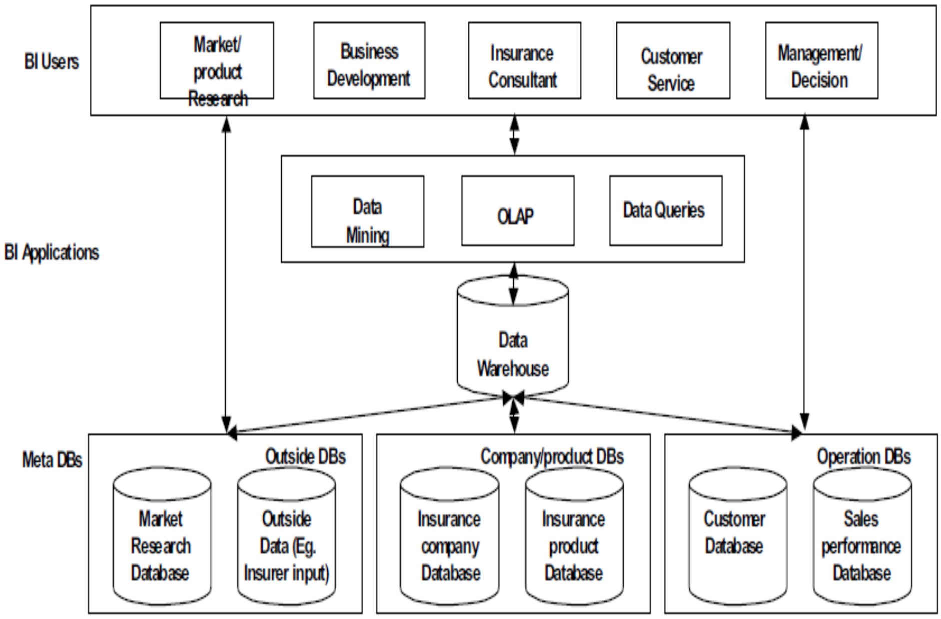

Figure 2.2: A Comprehensive BI Application Infrastructure for a Life Insurance Company. Reprinted from "Business intelligence-a case study in the life insurance industry" by Z. Xu, M. Zhang & X. Jiang, 2005. In *e-Business Engineering, 2005. ICEBE 2005. IEEE International Conference on* (pp. 129-132). IEEE.

Liao et al (2009) have investigated functionalities that best fit the life insurance consumers’ needs and wants for life insurance products by extracting specific knowledge patterns and rules from consumers and their demand chain. They have used a priori algorithm and clustering analysis as methodologies for knowledge extraction and results have illustrated as market segments and demand chain analysis on life insurance market in Taiwan in order to propose suggestions and solutions to the life insurance firms for new life insurance product development and marketing.

Shyng et al (2007) have addressed the effect of attributes on the combination values of decisions that insurance companies make to satisfy customers’ needs using Rough Set Theory. Their approach has redefined the value set of attributes through expert knowledge by reducing the independent dataset and reclassifying it. The results have demonstrated that the redefined combination values of attributes could contribute to the precision of decisions in insurance marketing. They have used a hit test that incorporates 50 validated sample data into the decision rule so that the hit rate has reached 100%. Consequently, they believe that the effects of attributes on combination values could be fully applied in research into insurance marketing.

Balaji & Srivatsa (2012) have used decision tree approach which is a widely used method since it is efficient and can deal with both continuous and categorical variables and generates understandable rules. Balaji & Srivatsa have applied ID3 decision tree algorithm on a life insurance data set for predicting customer preferences towards the life insurance policy preferences under the product type based on the spilt attribute. The adoption of the supervised learning technique for prediction analysis has used the demographic attributes of the customer. The decision tree approach implemented by Balaji & Srivatsa clearly delineates the customer segment based on the spilt attribute and contributes to retaining the profitable customers. So what they finally suggest is that the segment of life insurance customers resulted based on spilt attributed could be utilized for cross selling and up selling of life insurance products.

Rahman et al (2017) have applied different classification techniques on the data provided by a life insurance company in Bangladesh to classify customers as regular or irregular based on their given attributes in order to predict the class label for future customers. That is to determine whether a customer is regular or irregular in premium instalment payment. The initial data set has had about 282,282 policyholders where they have pruned it to 10,000 policyholders for faster processing of data. They have selected 10 attributes (policy term, age, sex, occupation, urban-rural, marital status, sum-assured, division, premium payment mode and regularity) from the original data set using attribute selection techniques to properly classify the data. Rahman et al have introduced a new attribute named ‘Regularity’ which consists of two values such as regular/irregular where regularity has been determined based on the starting date of a policy, the last payment dates of every policy. Classification techniques such as RIPPER, Naïve Bayes, IBK, SMO, Multilayer Perceptron and PART have been applied on data and a comparative analysis for the performance of the classifiers have been done to find the best-suited classifier for the acquired dataset. They have used “WEKA” for implementing classification techniques and tenfold cross-validation technique to validate the dataset. From the results, it has been evident that using PART classification gives the maximum correctly classified instances where the performance of the PART algorithm has been significant for the dataset compared to others.

2.6 Business Intelligence and Sri Lankan Life Insurance Industry

In Sri Lankan context, insurance companies are reluctant to use the off-the-shelf software because they are very costly and companies are not much convinced of what Business Intelligence can do for them. This opens up many opportunities and challenges for BI researchers to convince these companies of the commercial viability of BI efforts.

Customer attrition is an increasingly pressing issue faced by many Sri Lankan insurance providers today. Retaining customers who purchase life insurance policies has become an even bigger challenge since the policy duration spans for more than twenty years. Therefore Sri Lankan insurance companies are eager to reduce these attrition rates in the customer-base by analyzing operational data.

Goonetilleke & Caldera (2013) have analyzed customer attrition by classifying all policyholders who are likely to terminate their policies using classification techniques such as Decision trees and Neural Networks. Models generated have been evaluated using ROC curves and AUC values. Using these models, Goonetilleke & Caldera have suggested that customers who are at high risk of attrition can then be targeted for promotions to reduce the rate of attrition.

3. METHODOLOGY

3.1 Overview of Chapter

The methodology includes the process that has been used to collect information and data such as interviews, surveys and other research techniques where this research is in support of making business decisions for strategic planning in Sri Lankan life insurance industry by reinforcing business intelligence applications.

This chapter will deliver vital feedback using the methods used to collect data, the samples gathered from the population and to a final conclusion on business intelligence applications in Sri Lankan life insurance industry.

3.2 Introduction to Methodology

Analyzing data of life insurance companies gives an important insight on how the customers are reacting to the offered insurance policies by the companies. This information can be used to predict the behaviour of future policyholders. Life insurance companies maintain a large database on their customers and policy-related information. Business Intelligence techniques applied with proper preprocessing of data prove to be very efficient in extracting hidden information from data stored by life insurance companies.

Life Insurance policyholders often have to pay a monthly or yearly premium against their policy where it comes as the matter of regularity of a customer. Insurance companies are interested in the regular customers because it increases their chance of profit as a company. Again customers who are regular are benefited from the insurance policies they bought because it ensures their chances of successful insurance claims.

On the other hand, the reason for people to get interested in becoming a life insurance policyholder is the demand for life insurance. There are attributes which have been identified as determinants of life insurance demand such as sex, age, marital status, dependents, occupation, living environment etc. If a person is getting interested in becoming a policyholder he should possess a high demand for life insurance. If he has a high demand there is a high probability for him to become a regular customer by paying premium payments regularly.

Therefore, in order to proceed with the regularity of premium payments of a customer, determinants of life insurance demand have been considered. So based on the literature study, 10 attributes have been selected such as gender, age, civil status, dependents, occupation, living environment, sum-assured, policy term and premium payment mode as determinants of life insurance demand that have been figured out to be effective in determining the regularity of a customer.

Brief descriptions of these selected attributes are as follows:

(a) Gender of policyholder

(b) Age of the policyholder when he starts the policy

(c) Civil status of the policyholder whether single or married

(d) Dependants - the number of children

(e) Occupation category of the policyholder when s/he starts the policy such as agriculture, business, government sector, private sector etc

(f) Living environment refers to the policyholders dwelling status whether s/he is living in an urban area, semi-urban area or a rural area

(g) Sum-assured refers to a pre-decided amount which the insurer promises to pay the nominee in case of the policyholder’s death

(h) Policy term which refers to the time span through which the policyholder will pay his or her premiums

(i) Premium payment mode marks the breakdown of the payment of the policyholder’s total premium

(j) Regularity marks the policyholder as “regular” or “irregular” based on whether s/he is paying the premiums on time

So one of the best Business Intelligence applications where it can be applied in life insurance industry is to classify the life insurance policyholders as regular or irregular. According to the literature study done, in order to classify the life insurance customers as regular or irregular, classification algorithms have been proved to be very useful.

So the aim of this research which is reinforcing business intelligence applications in Sri Lankan life insurance industry could be achieved via classifying the life insurance policyholders based on their behavioural attributes using Business Intelligence so that it can predict the class label for future life insurance customers. Since classification techniques have been approved as one of the best approaches to look on regularity of life insurance policyholders according to the literature study, five different classification techniques such as Naïve Bayes, Multi-Layer Perceptron, IBk, PART and SMO have been used in this research and their performances have been compared to find the best-suited classifier for the acquired dataset.

3.3 Data Gathering

The target population for this research is life insurance policyholders in Sri Lankan Life Insurance companies. Since the population is large and the nature of the study is about premium payment regularity of life insurance policyholders, the most appropriate techniques are the quantitative methods.

According to studies carried out, secondary data have been used in previous researches. Therefore in order to gather data for the research, first it was searched for secondary life insurance data from Ceylinco Life Insurance. But due to the confidentiality of data maintained by Ceylinco Life Insurance, it was unable to collect the secondary life insurance data from them. Therefore it was decided to collect primary life insurance data.

Therefore the proposed research was carried out for a selected sample of life insurance policyholders where a questionnaire was used as the data collection method since interviews were not the most suitable method due to time constraints. The questionnaire was distributed among life insurance policyholders from three life insurance companies in Sri Lanka, Ceylinco Life Insurance, HNB Assurance and Union Assurance.

3.4 Population and Sample

Stratification is an efficient research sampling design, that provides more information with a given sample size. Therefore, *disproportionate stratified random sampling* has been adopted for the present study.

According to Krejcie & Morgan (1970), as the population size increases, the sample size increases at a diminishing rate and remains relatively constant at slightly more than 380. The relationship between sample size and the total population is illustrated in Figure 3.1.

The target population for this research is life insurance policyholders in Sri Lankan Life Insurance companies which are more than two million. Therefore the sample size of 380 is considered accurate for the present study where a sample of 400 was the targeted sample size.

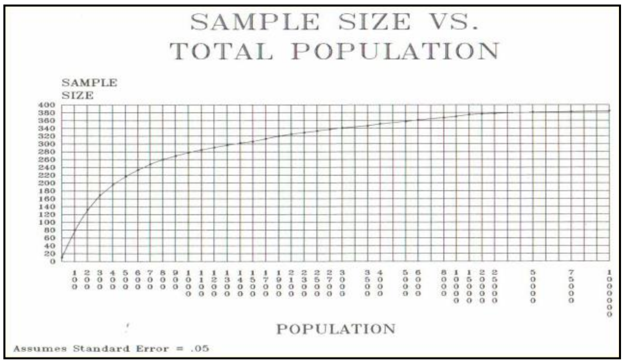

Figure 3.1: Sample Size vs. Total Population. Reprinted from "Determining sample size for research activities" by R. V. Krejcie & D. W. Morgan, 1970. *Educational and psychological measurement*, *30*(3), 607-610.

3.5 Designing the Questionnaire

The approach for obtaining data for this study was collecting primary data via a questionnaire. The questionnaire was distributed among 400 life insurance policyholders from three life insurance companies in Sri Lanka, Ceylinco Life Insurance, HNB Assurance and Union Assurance. The questionnaire comprises with eleven questions in order to collect certain attributes and behavioural factors of life insurance policyholder such as policy term, age, sex, occupation, marital status, premium mode, premium amounts and regularity of premium payments. The relevant questionnaire is in Appendix A for reference.

3.6 Methodology

Research approach mainly contains two steps - "Data Preprocessing" and "Classification Technique Implementation". In order to preprocess data, it is very much required to analyze and understand data properly. Pruning data is needed so that the attributes could have maximum effect on specifying a customer as regular or irregular.

In "Classification Technique Implementation" phase, different classification techniques have been applied which have been identified via literature survey as most effective models to apply on insurance data to find out the most effective classifier for Sri Lankan life insurance data. These will be discussed in detail in the following subsections.

3.6.1 Data preprocessing

Selected data have been transformed into different forms that are appropriate for applications of identified classification techniques. Since this research classification is based on two classes: regular or irregular, in order to collect information regarding regularity of policyholders premium payments, whether payments are regular or irregular, the questionnaire had to be designed in following way since it will not be user friendly to ask the question from customer directly whether their payments are regular or irregular.

11. Do you pay premium regularly?

Always on time □
Most of the time □
Sometimes □
Occasionally on time □
Never on time □

In order to extract the relevant data that is needed for the research proceedings, the above answer of policyholders had to be mapped with whether premium payments are regular or irregular. So the answers of "Always on time" and "Most of the time" were mapped with regular class and the other three answers "Sometimes", "Occasionally on time" and "Never on time" were mapped with irregular class.

3.6.2 Classification technique implementation

A number of classification techniques such as Naïve Bayes, Multi-Layer Perceptron, IBK, PART and SMO have been used and compared in this research so that the best classification model could be determined.

3.6.3 WEKA for data mining

In order to carry out the classification process, WEKA was selected to assist machine learning algorithms since it has all the functions required to analyze the data set by performing the machine learning algorithms.

There were several reasons for selecting that WEKA software application. Those are,

1. An open source software tool
2. Inbuilt functions for generating rules, making predictions, data conversions, etc
3. Data preparation and data selection algorithms are included
4. Easily accessible online tutorials, guides and user manual
5. Consisted of understandable, user-friendly interfaces

4. DATA ANALYSIS

4.1 General Overview of Data

Table 4.1 gives an overall description of the primary data set. The dataset comprises eleven attributes where they have been categorized into four main sections such as Demographics, Social, Cultural and Economical.

Category	***Attribute Name***	***Possible Values***
Demographics	Gender	Male, Female
	Age	10-20, 20-30, 30-40, 40-50, 50-60
Social	Civil Status	Married, Single
	Number of Children	None, One, Two, Three, More than Three
	Occupation	Agriculture, Business, Government Sector, Private Sector, None
Cultural	Level of Education	Primary, O/L, A/L, Diploma, Graduated, Post Graduated
	Living Environment	Urban, Semi Urban, Rural
Economical	Policy Term	1-10, 10-20, 20-30, 30-40, 40-50
	Sum Assured	< 100000, 100000 - 250000, 250000 - 500000, 500000 - 750000, 750000 - 1000000, 1000000 <
	Premium Payment Mode	Single Payment, Yearly, Biannually, Quarterly, Monthly
	Frequency of Payment	Always on time, Most of the time, Sometimes, Occasionally on time, Never on time

Table 4.1: Data Source

4.2 Statistical Analysis

This section discusses some attributes of life insurance policyholders that were figured out to be very effective in determining the regularity of life insurance premium payments. Therefore, this illustration shows some specific areas that were highly considered within this research process to acquire hidden knowledge. This analysis helped to get the preliminary knowledge of the data source which was very much useful for the next step of the research process. The data source consists of 400 life insurance policyholders.

4.2.1 Demographics

According to Figure 4.1, three out of five represents male in the collected data set which might give a simple insight of which the most of the life insurance policyholders are to be males than females.

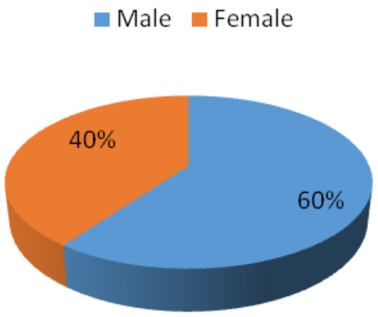

Figure 4.1: Variation over Gender

The Figure 4.2shows a trend of increase in buying life insurance policies while people are getting aged from the year group of 10-20 to age group 30-40 and then a decrease up until age group 50-60. So the highest number of buying life insurance policies seems to happen in between the age of 30-40 where an increase of responsibilities might be the reason.

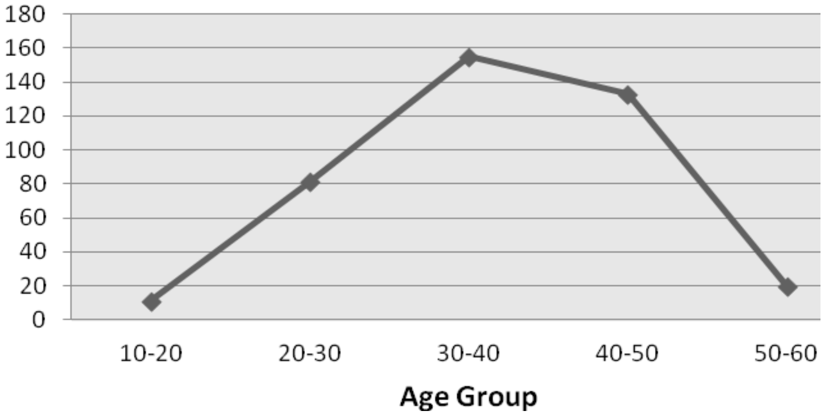

Figure 4.2: Variation over age groups

4.2.2 Social

Figure 4.3 depicts that ninety-three per cent of data source represents married people which might give a simple insight where most of the life insurance policyholders are married people because they tend to buy life insurance policies than single people due to the higher responsibilities they have had to carry on.

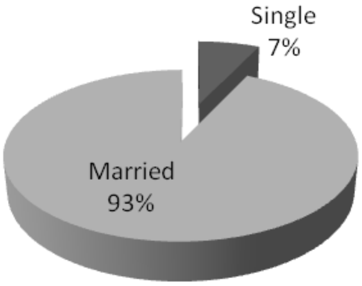

Figure 4.3: Variation over Civil Status

Dataset according to Figure 4.4 shows that most of the policyholders have only one child and it is 50% of the dataset. 25% have two children where more than 10% don't have children. A few policyholders have more than 2 children as well.

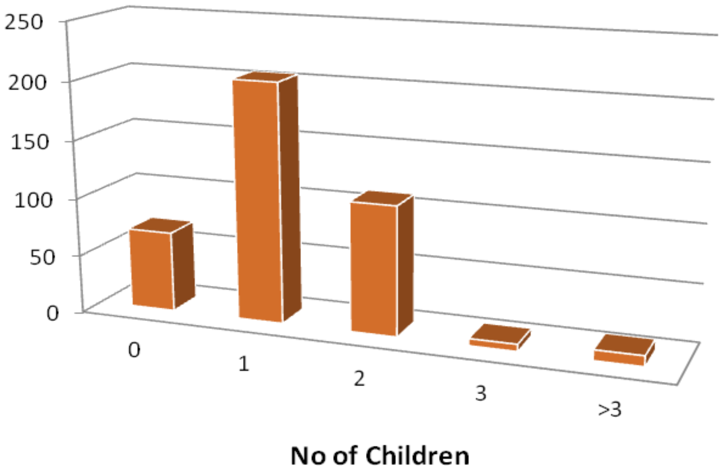

Figure 4.4: Variation over the number of children

According to Figure 4.5, the collected dataset shows that more than 50% of policyholders have been occupied in the private sector where only 13% have their career in government sector. Nearly 20% of life insurance policyholders are businessmen where more than 10% are in the agriculture field.

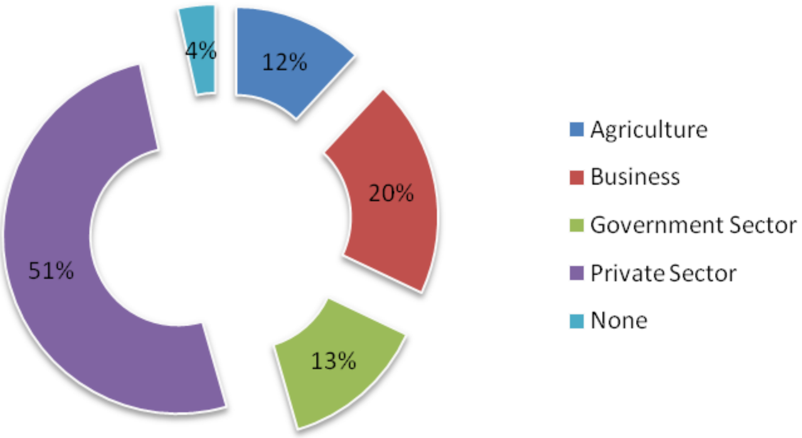

Figure 4.5: Variation over Occupation

4.2.3 Cultural

According to the data source in Figure 4.6, higher numbers of policyholders have been graduated and it is nearly 40% of the dataset. More than 25% have Advanced Level educational Qualification where nearly 10% have got the Ordinary Level educational qualification. There are few policyholders who have got an only primary level education as well.

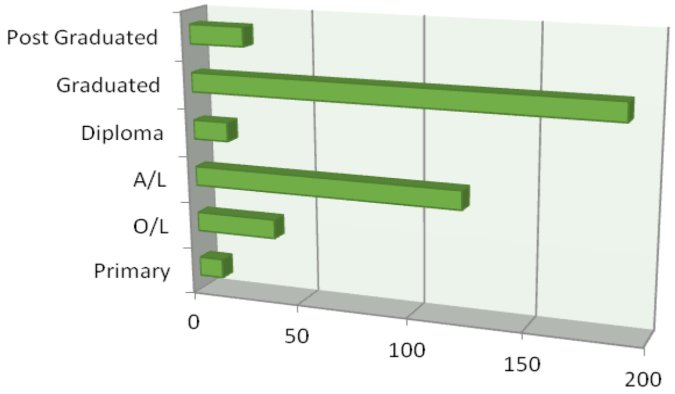

Figure 4.6: Variation over Educational qualifications

The data source in Figure 4.7 illustrates that most of the life insurance policyholders live in a semi-urban environment where it is nearly 50% of the dataset. Nearly 25% of them live in urban areas where a few policyholders live in a rural environment as well.

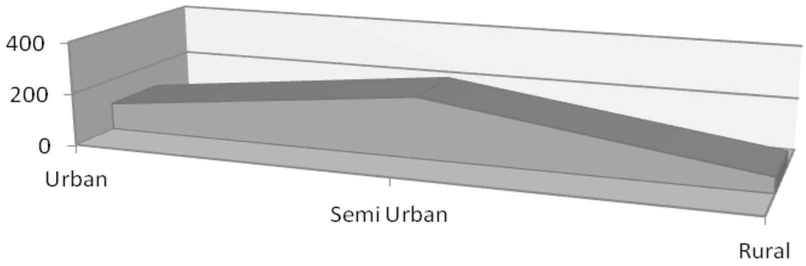

Figure 4.7: Variation over Living environments

4.2.4 Economical

According to Figure 4.8, a higher number of life insurance policyholders in the data source have tended to buy policies of insurance term for 10-20 years where more than 25% of policyholders have bought policies with life duration of 1-10 years. The trend of buying policies for life term more than 20 years shows a slight decrease up until 50 years.

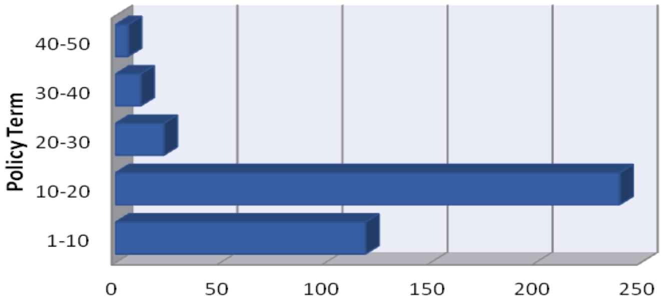

Figure 4.8: Variation over Policy Term

The data source in Figure 4.9 illustrates the most numbers of policyholders have got insured for less than Rs. 100,000. More than 10% have insured for the rupee value in between 100,000-250,000. Only a very few people have bought life insurance for more than Rs.750,000.

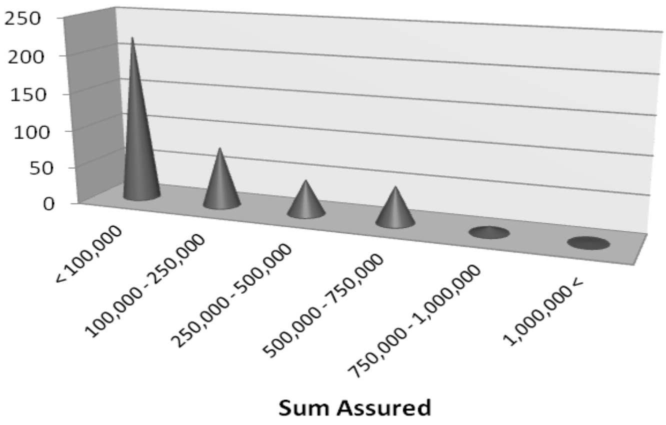

Figure 4.9: Variation over Sum Assured

According to Figure 4.10, higher numbers of policyholders have tended to pay premiums monthly and it is 75% of the dataset. Some policyholders have paid premiums in single payments where one few people have to pay annually and biannually.

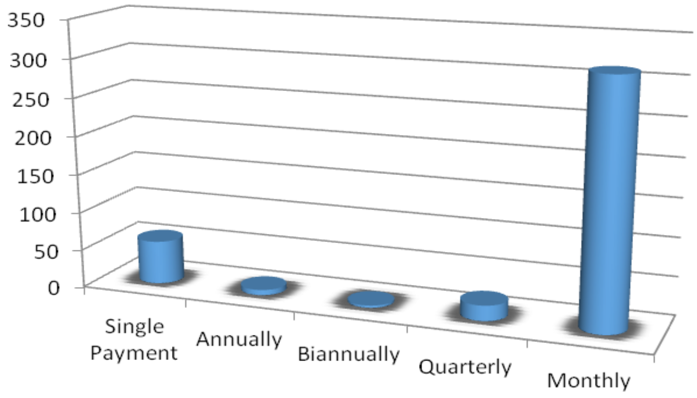

Figure 4.10: Variation over Premium Payment Methods

According to Figure 4.11, fifty-three per cent of data source illustrates that life insurance policyholders pay premiums regularly where forty-seven per cent pays premiums irregularly. Policyholders who pay premiums regularly include who pay premiums always on time and the policyholders who pay most of the time on time. The policyholders who are irregular in premium payments are policyholders who pay premiums sometimes on time, occasionally on time and never on time.

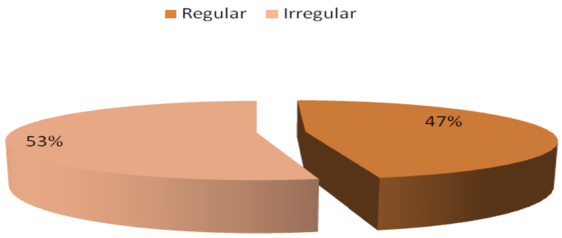

Figure 4.11: Variation over Premium Payment Regularity

4.3 Statistical Tests

Chi-Square Test of Independence

The Chi-Square Test of Independence determines whether there is a statistical independence or association between categorical variables. Table 4.2 gives the Pearson Chi-Square values of attributes with an association to the Regularity of premium payment attribute for collected sample data. Pearson Chi-Square significance value is the probability of observing the sample outcome if attributes are independent of the entire population. Significance level has been chosen as 0.05for this analysis where it is said that the association between two variables is statistically significant if Significance (2-sided) < 0.05. Results of Chi-Square Tests are at Appendix B for reference.

According to Table 4.2, three attributes: policy term, sum assured and premium payment mode have Pearson Chi-Square values less than chosen significance level α = 0.05, where we can conclude that there are associations between policy term and regularity of premium payments, sum assured and regularity of premium payments as well as premium payment mode and regularity of premium payments. All other attributes have Pearson Chi-Square values greater than significance level α = 0.05, where we can conclude that there is not enough evidence to suggest an association between regularity of premium payments and other attributes such as gender, age, civil status, level of education, living environment, number of children and occupation. Although these attributes were selected as effective in determining the regularity of premium payments according to literature study, the reason for not having associations here in between regularity of premium payments and other attributes might be the small sample data set of 400 records which have been used for this research.

Attribute	**Pearson Chi-Square**
Gender	0.287
Age	0.904
Civil Status	0.225
Number of Children	0.488
Occupation	0.664
Level of Education	0.802
Living Environment	0.722
Policy Term	0.000
Sum Assured	0.001
Premium Payment Mode	0.000

Table 4.2: The Chi-Square test of independence (association with Regularity of premium payments)

4.4 Evaluation

Since the dataset used for this research consists of a limited number of records, cross-validation has been used as the evaluation technique. In this case, 10-fold cross-validation has been used where it involves running the learning algorithm 10 times to build and evaluate 10 classifiers. Evaluation using classification has been used to validate the result models. The evaluation results of classifiers are in Appendix C for reference.

4.3.1 Comparison of models

A model to be comparatively better, it has to have the highest ROC Area, highest TP Rate and lowest FP Rate. Table 4.3 gives an overall evaluation for five models, Naive Bayes, Multi-Layer Perceptron, IBk, PART and SMO. According to Figure 4.13, Naive Bayes model performs better than all other models since it has the highest ROC Area, as well as Figure 4.12, illustrates highest TP Rate and lowest FP Rate for Naive Bayes than all other models. When comparing three evaluation techniques, IBk has shown a poor performance among other models while Multi-Layer Perceptron and PART goes on average.

Model	**TP Rate**	**FP Rate**	**ROC Area**
Naive Bayes	0.668	0.357	0.69
Multi Layer Perceptron	0.613	0.392	0.67
IBk	0.623	0.395	0.633
PART	0.638	0.372	0.675
SMO	0.645	0.384	0.63

Table 4.3: Comparison of models using classification evaluations

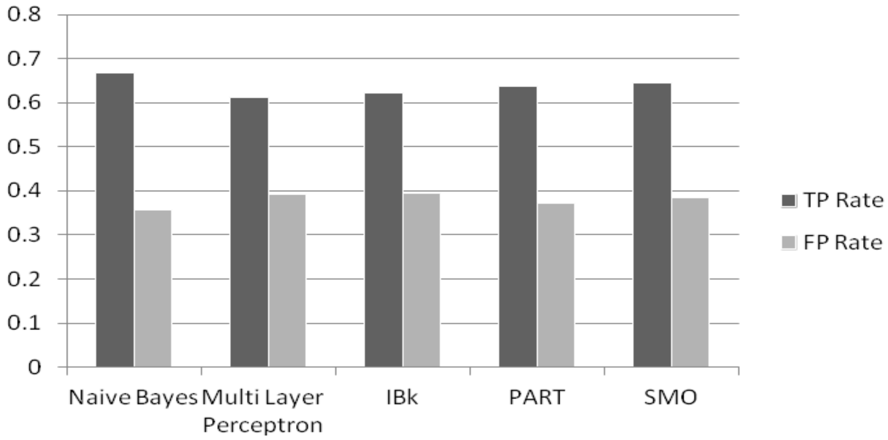

Figure 4.12: Model comparison in TP and FP rates

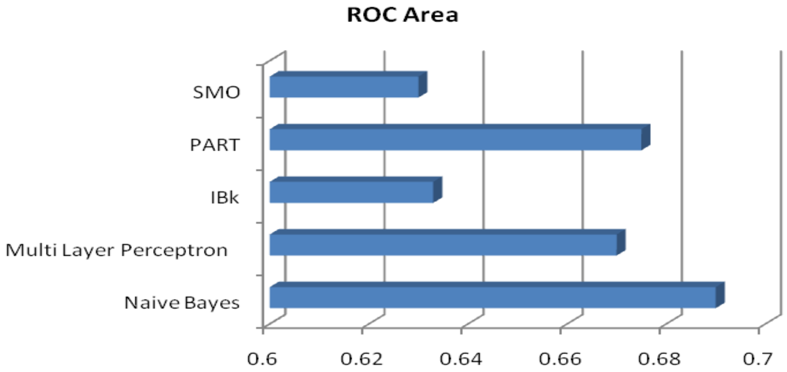

Figure 4.13: Model comparison in ROC Area

Table 4.4 represents model evaluation results that have been by Rahman et al (2017) for premium payment regularity for a life insurance company in Bangladesh where they have used a secondary data set of 10,000 policyholders. They used six different classification techniques: RIPPER, Naïve Bayes, IBk, SMO, Multilayer Perceptron and PART where the performance of PART algorithm had been significant on their dataset compared to other models. Naive Bayes have had a high FP-rate where both Naive Bayes and IBk have been less efficient on correctly classified instances. SMO and Multilayer Perceptron had worked moderately.

When comparing model evaluation results of current work with Rahman et al (2017) work, according to Table 4.3 and Table 4.4, Rahmans' work has a higher TP rate and a lower FP rate than current work. This might be due to the higher number of records used by Rahman where their work has used approximately 10,000 records from secondary data while current research contains only 400 records from primary data. Other than to that, PART has given the best performance for Rahmans' work while Naive Bayes has performed better for current work. Reason for this might be samples used for researches where Rahman has done this for Bangladesh while current research is for Sri Lanka.

Model	**TP Rate**	**FP Rate**
Naive Bayes	0.70	0.30
Multi-Layer Perceptron	0.71	0.28
IBk	0.72	0.28
PART	0.73	0.27
SMO	0.71	0.28
RIPPER	0.72	0.27

Table 4.4: Model evaluation done by Rahman et al (2017)

4.5 Summary

The first part of this chapter has discussed a statistical analysis of data source which is the initial step of the research process while mapping those data with knowledge gain through the literature study. Data source have been analyzed using different types of statistical charts with various categories. Attributes have categorized into five main sections such as demographics, social, cultural and economical. Some attributes have given significant variations while some haven't. It's worth to carry out a statistical analysis to get a general idea about the data set before other proceedings of the research.

The second part of this chapter has discussed the evaluation of classification models in order to come up with the best suitable model for classification of premium payment regularity. According to evaluation results, Naive Bayes model has performed better than all other models because it had the highest ROC Area, highest TP Rate as well as lowest FP Rate than other four models: Multi-Layer Perceptron, IBk, PART and SMO.

5. RECOMMENDATION & CONCLUSION

The key challenge of Business Intelligence is how to use technologies to creatively address major business issues and achieve operational effectiveness. In life insurance industry, BI can help firms to gain business advantage mainly to support in decision making where BI could enable the development of cutting-edge business models and achieve a great business impact in life insurance industry. Life insurance companies need to know essentials in decision making where BI techniques could help in order to compete in the market of life insurance such as in acquiring new customers, retaining existing customers, policy designing, policy selection etc.

In order to achieve the research objective of the research which is reinforcing business intelligence applications in Sri Lankan life insurance industry, the approach which has used here was applying some BI techniques on life insurance customer data to identify a better BI model that is applicable to Sri Lankan life insurance industry. Since classification techniques have been proved to be very useful in classifying customers according to their attributes, main focus of the research was to apply different classification techniques on life insurance data from Sri Lankan life insurance companies where normally classes can be created according to policy term, premium mode, premium amount, age, income, occupation and premium payment regularity.

According to literature study done, it was identified that certain attributes and behavioural factors of life insurance policyholder such as policy term, age, sex, occupation, marital status, premium mode and premium amounts have been effective in determining the regularity of policyholders in instalment payments. Therefore it was decided to use classification techniques on life insurance customer data to classify a non-regular life insurance policyholder from a regular life insurance policyholder in order to come up with a classifier that could effectively determine the regularity of policyholders in instalment payments in Sri Lankan life insurance industry.

Since it was unable to collect secondary life insurance data from Ceylinco Life Insurance due to confidentiality maintained by the company itself, a questionnaire was designed to collect life insurance data from life insurance policyholders and a sample of 400 has been collected from three life insurance companies in Sri Lanka, Ceylinco Life Insurance, HNB Assurance and Union Assurance. These collected life insurance data certain attributes and behavioural factors of life insurance policyholder such as policy term, age, sex, occupation, marital status, premium mode, premium amounts and regularity of premium payments.

Five different classification techniques such as Naïve Bayes, Multi-Layer Perception, IBK, PART and SMO have been used in order to classify the customers as regular or irregular based on their given attributes so that it can predict the class label for future customers. A comparative analysis of the performance of these classifiers also have been done using evaluation criteria such as ROC curves and five different models gave different results.

From the results achieved, the best performance was achieved by the Naïve Bayes classifier. Naive Bayes classifier is a good model for any domain and for any size of a dataset. It might give more reliable results for larger ones but for moderate ones also it gives good results than other models which too were occurred in this research. Naive Bayes classifier gave the best results than the other models in this research where we can come to a conclusion that Naive Bayes classifier could effectively determine the regularity of policyholders in instalment payments in Sri Lankan life insurance industry.

By applying these BI classification techniques, life insurance companies could fully exploit data about customers' premium payment patterns and behaviours as well as to help reduce fraud in premium payments and enhance risk management. For example, insurance companies could identify behavioural attributes of policyholders who pay premiums regularly such as policy term, age, sex, occupation, marital status, premium mode and premium amount where they could get used for marketing and advertising campaigns. Not only that at the time of purchase of a new life policy, insurance companies will be able to identify whether this new customer will be a regular customer or irregular by entering his/her behavioural attributes such as policy term, age, sex, occupation, marital status, premium mode, premium amount etc.

Another instance is policyholders who pay premiums regularly are much more likely to renew or buy new policies while policyholders who are irregular in premium payments are less likely to renew their policies. So by offering some discounts to policyholders who are irregular in premium payments considering above mentioned their behavioural attributes, insurance firms can add some value to themselves and thereby will be able to increases customer loyalty. On the other hand, insurance firms can suggest new life insurance products to policyholders who pay premiums regularly considering their behavioural attributes such as policy term, age, sex, occupation, marital status, premium mode, premium amount.

Because of the limited scope of the research, only four hundred records from policyholders were able to collect. So as a future work, it might be able to give better results from a bigger data set.

There are various methods that can be suggested as classifier improvement strategies. This research has approached the regularity problem as a classification task. It is also possible to explore the dataset to find any strong associations between the data. An association analysis that satisfies a minimum support and a threshold will be useful for the insurance provider to act on these rules to promote to the customer.

REFERENCES

Albashrawi, M. (2016). Detecting Financial Fraud Using Data Mining Techniques: A Decade Review from 2004 to 2015. *Journal of Data Science*, *14*(3), 553-569.

Balaji, S., & Srivatsa, S. K. (2012). Decision Tree induction based classification for mining Life Insurance Databases. *Int J Comput Sci Inf Technol Secur (IJCSITS)*, *2*, 699-703.

Çelik, S., & Kayali, M. M. (2009). Determinants of demand for life insurance in European countries. *Problems and perspectives in management*, *7*(3), 32-37.

Cerny, P. A., & Proximity, M. A. (2001). Data mining and neural networks from a commercial perspective. In *ORSNZ Conference Twenty Naught One* (pp. 1-10).

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly*, *36*(4).

Devale, A. B., & Kulkarni, R. V. (2012). Applications of data mining techniques in life insurance. *International Journal of Data Mining & Knowledge Management Process*, *2*(4), 31-40.

Goonetilleke, T. O., & Caldera, H. A. (2013). Mining life insurance data for customer attrition analysis. *Journal of Industrial and Intelligent Information*, *1*(1).

Hedgebeth, D. (2007). Data-driven decision making for the enterprise: an overview of business intelligence applications. *Vine*, *37*(4), 414-420.

Data mining techniques. In IBM. Retrieved November 18, 2017, from https://www.ibm.com/developerworks/library/ba-data-mining-techniques/index.html.

Insurance Regularity Commission of Sri Lanka. Retrieved November 18, 2017, from http://www.ibsl.gov.lk/insurance-companies.html.

Klumpes, P. J. (2004). Performance benchmarking in financial services: Evidence from the UK life insurance industry. *The Journal of Business*, *77*(2), 257-273.

Krejcie, R. V., & Morgan, D. W. (1970). Determining sample size for research activities. *Educational and psychological measurement*, *30*(3), 607-610.

Liao, S. H., Chen, Y. N., & Tseng, Y. Y. (2009). Mining demand chain knowledge of life insurance market for new product development. *Expert Systems with Applications*, *36*(5), 9422-9437.

Mehregan, S., & Samizadeh, R. (2012). Customer Retention Based on the Number of Purchase: A Data Mining Approach. *International Journal of Management and Business Research*, *2*(1), 41-50.

Mohapatra, S., & Tiwari, M. (2009). Using Business Intelligence for Automating Business Processes in Insurance. *IJACT: International Journal of Advancements in Computing Technology*, *1*(2), 92-98.

Mosley Jr, R. C. (2012). Social media analytics: Data mining applied to insurance Twitter posts. In *Casualty Actuarial Society E-Forum* (Vol. 2, p. 1).

Ngai, E. W., Xiu, L., & Chau, D. C. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert systems with applications*, *36*(2), 2592-2602.

Olszak, C. M., Ziemba, E., & Koohang, A. (2006). Business Intelligence Systems in the Holistic Infrastructure Development Supporting Decision-Making in Organisations. *Interdisciplinary Journal of Information, Knowledge & Management*, *1*.

Poleto, T., de Carvalho, V. D. H., & Costa, A. P. C. S. (2015, May). The roles of big data in the decision-support process: an empirical investigation. In *International Conference on Decision Support System Technology* (pp. 10-21). Springer, Cham.

Rahman, M. S., Arefin, K. Z., Masud, S., Sultana, S., & Rahman, R. M. (2017). Analyzing Life Insurance Data with Different Classification Techniques for Customers' Behavior Analysis. In *Advanced Topics in Intelligent Information and Database Systems* (pp. 15-25). Springer International Publishing.

Redzuan, H. (2011). Analysis Of The Demand For Life Insurance And Family Takaful. *Universiti Teknologi Mara Doctor of Philosophy*.

Sarkodie, E. E., & Yusif, H. M. (2015). Determinants of Life Insurance Demand, Consumer Perspective-A Case Study of Ayeduase-Kumasi Community, Ghana. *Business and Economics Journal*, *6*(3), 1.

Shollo, A. (2011). Using business intelligence in IT governance decision making. Governance and Sustainability in Information Systems. Managing the Transfer and Diffusion of IT, 3-15.

Shyng, J. Y., Wang, F. K., Tzeng, G. H., & Wu, K. S. (2007). Rough set theory in analyzing the attributes of combination values for the insurance market. *Expert Systems with Applications*, *32*(1), 56-64.

Sliwinski, A., Michalski, T., & Roszkiewicz, M. (2013). Demand for life insurance—An empirical analysis in the case of Poland. *The Geneva Papers on Risk and Insurance-Issues and Practice*, *38*(1), 62-87.

Umamaheswari, K., & Janakiraman, S. (2014). Role of Data mining in Insurance Industry. *An international journal of advanced computer technology*, *3*(6), 961-966.

Wikipedia. (2017). Business Intelligence. Retrieved November 18, 2017, from https://en.wikipedia.org/wiki/Business_intelligence.

Xu, Z., Zhang, M., & Jiang, X. (2005, October). Business intelligence-a case study in life insurance industry. In *e-Business Engineering, 2005. ICEBE 2005. IEEE International Conference on* (pp. 129-132). IEEE.

APPENDIX A: QUESTIONNAIRE INSTRUMENT

The Regularity of Life Insurance policy premium payments

1. What is your Gender?

Male □ Female □

2. What is your Civil Status?

Single □ Married □

2. b How many children do you have (Please answer if relevant)?

None □

One □

Two □

Three □

More than three □

3. What is your educational level?

Primary □

O/L □

A/L □

Diploma □

Graduated □

Post Graduated □

4. Are you living in an urban area or a rural area?

Urban □ Semi Urban □ Rural □

5. Do you have a life insurance policy?

Yes □ No □

Please answer if relevant,

6. What was your age when you first obtained this policy?

10-20 □

20-30 □

30-40 □

40-50 □

50-60 □

7. What was your occupation when you first obtained this policy?

Agriculture □

Business □

Government Sector □

Private Sector □

None □

8. What is the validity period of the policy (years)?

1-10 □

10-20 □

20-30 □

30-40 □

40-50 □

9. What is the pre-decided amount which the insurer promises to pay the beneficiary in case of the policyholder's death (Rs.)?

< 100,000	□
100,000 - 250,000	□
250,000 - 500,000	□
500,000 - 750,000	□
750,000 - 1,000,000	□
1,000,000 <	□

10. What is the payment mode of paying premiums for this policy?

Single Payment	□
Yearly	□
Biannually	□
Quarterly	□
Monthly	□

11. Do you pay premium regularly?

Always on time	□
Most of the time	□
Sometimes	□
Occasionally on time	□
Never on time	□

APPENDIX B: CHI-SQUARE TESTS OF INDEPENDENCE

GENDER

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	1.131[a]	1	.287		
Continuity Correction[b]	.924	1	.336		
Likelihood Ratio	1.131	1	.288		
Fisher's Exact Test				.307	.168
N of Valid Cases	400				

AGE

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	1.037[a]	4	.904
Likelihood Ratio	1.036	4	.904
N of Valid Cases	400		

MARITAL STATUS

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	1.473[a]	1	.225		
Continuity Correction[b]	1.035	1	.309		
Likelihood Ratio	1.498	1	.221		
Fisher's Exact Test				.245	.155
N of Valid Cases	400				

NO OF DEPENDANTS

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	3.431[a]	4	.488
Likelihood Ratio	3.457	4	.484
N of Valid Cases	400		

OCCUPATION

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	2.390[a]	4	.664
Likelihood Ratio	2.388	4	.665
N of Valid Cases	400		

EDUCATIONAL LEVEL

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	2.332[a]	5	.802
Likelihood Ratio	2.337	5	.801
N of Valid Cases	400		

LIVING ENVIRONMENT

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	.652[a]	2	.722
Likelihood Ratio	.652	2	.722
N of Valid Cases	400		

POLICY TERM

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	24.916[a]	4	.000
Likelihood Ratio	25.566	4	.000
N of Valid Cases	400		

SUM_ASSURED

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	20.608[a]	5	.001
Likelihood Ratio	21.318	5	.001
N of Valid Cases	400		

PAYMENT_MODE

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	72.754[a]	4	.000
Likelihood Ratio	85.304	4	.000
N of Valid Cases	400		

APPENDIX C: EVALUATION OF CLASSIFIERS

SMO

```
=== Summary ===

Correctly Classified Instances        258              64.5    %
Incorrectly Classified Instances       142              35.5    %

=== Detailed Accuracy By Class ===

          TP Rate   FP Rate   Precision   Recall    F-Measure   ROC Area     Class
          0.854     0.594      0.621      0.854     0.719       0.63         Irregular
          0.406     0.146      0.71       0.406     0.517       0.63         Regular
Avg.      0.645     0.384      0.663      0.645     0.625       0.63

=== Confusion Matrix ===

   a   b   <-- classified as
 182  31 |   a = Irregular
 111  76 |   b = Regular
```

PART

```
=== Summary ===

Correctly Classified Instances        255              63.75   %
Incorrectly Classified Instances       145              36.25   %

=== Detailed Accuracy By Class ===

          TP Rate   FP Rate   Precision   Recall    F-Measure   ROC Area     Class
          0.709     0.444     0.645       0.709     0.676       0.675        Irregular
          0.556     0.291     0.627       0.556     0.589       0.675        Regular
Avg.      0.638     0.372     0.637       0.638     0.635       0.675

=== Confusion Matrix ===

   a   b   <-- classified as
 151  62 |   a = Irregular
  83 104 |   b = Regular
```

Naive Bayes Classifier

```
=== Summary ===

Correctly Classified Instances         267               66.75   %
Incorrectly Classified Instances       133               33.25   %

=== Detailed Accuracy By Class ===

         TP Rate   FP Rate   Precision   Recall   F-Measure   ROC Area    Class
         0.845     0.535      0.643      0.845    0.73        0.69        Irregular
         0.465     0.155      0.725      0.465    0.567       0.69        Regular
Avg.     0.668     0.357      0.681      0.668    0.654       0.69

=== Confusion Matrix ===

   a   b   <-- classified as
 180  33 |   a = Irregular
 100  87 |   b = Regular
```

Multi-Layer Perceptron

```
=== Summary ===

Correctly Classified Instances         245               61.25   %
Incorrectly Classified Instances       155               38.75   %

=== Detailed Accuracy By Class ===

         TP Rate   FP Rate   Precision   Recall   F-Measure   ROC Area    Class
         0.643     0.422      0.634      0.643    0.639       0.67        Irregular
         0.578     0.357      0.587      0.578    0.582       0.67        Regular
Avg.     0.613     0.392      0.612      0.613    0.612       0.67

=== Confusion Matrix ===

   a   b   <-- classified as
 137  76 |   a = Irregular
  79 108 |   b = Regular
```

IBk

```
=== Summary ===

Correctly Classified Instances         249               62.25   %
Incorrectly Classified Instances       151               37.75   %

=== Detailed Accuracy By Class ===

          TP Rate   FP Rate   Precision   Recall    F-Measure   ROC Area      Class
          0.746     0.519      0.621      0.746     0.678        0.633        Irregular
          0.481     0.254      0.625      0.481     0.544        0.633        Regular
Avg.      0.623     0.395      0.623      0.623     0.615        0.633

=== Confusion Matrix ===

   a   b   <-- classified as
 159  54 |   a = Irregular
  97  90 |   b = Regular
```