2006

Comparison of data mining and statistical techniques for classification model

Rochana Lahiri
Louisiana State University and Agricultural and Mechanical College

COMPARISON OF DATA MINING AND STATISTICAL TECHNIQUES FOR CLASSIFICATION MODEL

A Thesis

Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Master of Science

in

The Department of Information Systems & Decision Sciences

by
Rochana Lahiri
B.E., Jadavpur University, India, 1991
December 2006

ACKNOWLEDGEMENTS

It is a moment of great pleasure for me to take this opportunity to express my sincere gratitude to my supervisor, Dr. Helmut Schneider, who took so much interest in my work and went out of his way to help me. I hope that he would oblige me with his valued suggestions and advice in the future too.

I convey my sincere thanks to Dr. Joni Nunnery and Omer Soysal who provided me with valuable inputs regarding my work and helped me all along. I am also grateful to my teachers of the ISDS department for being so cooperative and helpful throughout.

I take the opportunity here to express my deep regards for my late parents who taught me the values of life and who, were they present, would have been very happy at this moment. My very special thanks go to my husband Ramanuj who has been by my side always, been so kind, considerate and understanding and had encouraged me throughout. I also thank Neel, Anindita, Proyag, Shreya, Atri, Rumpa, Bidisha, Anita, Abhijit, Sumita, Amitabha, Udit, Rohit, and Sora for being such nice and supportive friends.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The purpose of this study is to observe the performance of three statistical and data mining classification models viz., logistic regression, decision tree and neural network models for different sample sizes and sampling methods on three sets of data. It is a 3 by 2 by 3 by 8 study where each statistical or data mining method has been employed to build a model for each of 8 different sample sizes and two different sampling methods. The effect of sample size on the overall performance of each model against two sets of test data are observed and compared.

It is seen that for a given dataset, none of the three methods is found to outperform any other and their performances are comparable. This is in contrast to many of the existing studies as cited in the literature review chapter of this thesis. But the absolute value of prediction accuracy varied between the three datasets indicating that the data distribution and data characteristics play a role in the actual prediction accuracy, especially the ratio of the binary values of the dependent variable in the training dataset and the population. The models built with each of the sample size and sampling method for each method were run on two sets of test data to test whether the prediction accuracy was being replicated. It was found that for each of the cases the prediction accuracy was replicated across the test datasets.

1. INTRODUCTION

The management and analysis of information and using existing data for correct prediction of state of nature for use in similar problems in the future has been an important and challenging research area for many years. Information can be analyzed in various ways. Classification of information is an important part of business decision making tasks. Many decision making tasks are instances of classification problem or can be formulated into a classification problem, viz., prediction and forecasting problems, diagnosis or pattern recognition. Classification of information can be done either by statistical method or data mining method.

Data mining (DM) is also popularly known as Knowledge Discovery in Database (KDD). DM, frequently treated as synonymous to KDD, is actually a part of knowledge discovery process and is the process of extracting information including hidden patterns, trends and relationships between variables from a large database in order to make the information understandable and meaningful and then use the information to apply the detected patterns to new subsets of data and make crucial business decisions. The ultimate goal of data mining is prediction – predictive data mining is the most common type of that has the most direct business applications. The process basically consists of three stages: 1) the initial exploration, 2) model building or pattern identification with validation/verification and 3) deployment, i.e., the application of the model to new data in order to generate predictions. Data mining has very intrinsic connection to statistics. Stage (1) involving data cleaning, data transformation and selecting subsets of records use a variety of graphical and statistical methods such as techniques for identifying distributions of variables, reviewing large correlation matrices for coefficients that meet certain thresholds or examining multi-way frequency tables. Multivariate exploratory techniques designed specifically to identify patterns in multivariate or univariate data sets include cluster analysis, factor analysis, discriminant function analysis, multidimensional scaling, log-linear analysis, canonical correlation, stepwise linear and nonlinear (e.g., logit) regression, correspondence analysis, time

series analysis and classification trees. Stage (2) involves considering various models and choosing the best one based on their predictive performance and a variety of techniques to achieve that goal have been developed such as neural network, decision tree, etc. These are often considered the core of 'predictive modeling' techniques and approaches used for these techniques such as regression, discrimination and classification problems usually fall in the area of multivariate statistics, theory of probability, sampling and inference. So, data mining techniques are basically dependent on statistical techniques and combine machine learning algorithms and database management technologies with it and are very suitable for manipulating large number of records, often ranging from few hundred thousands to millions of data instances which are in general highly dimensional and dynamic in nature. The most commonly used techniques in DM based on statistical analysis for predictive modeling, are decision trees and neural network.

Statistical methods alone, on the other hand, might be described as being characterized by the ability to only handle data sets which are small and clean, which permit straightforward answers via intensive analysis of single data sets, which are static, which were sampled in an iid (variables are independent and identically distributed if each has the same probability distribution as the others and all are mutually independent) manner, which were often collected to answer the particular problem being addressed and often which are solely numeric. None of these apply in data mining context.

Literature shows that a variety of statistical methods and heuristics have been used in the past for the classification task. Decision science literature also shows that numerous data mining techniques have been used to classify and predict data; data mining techniques have been used primarily for pattern recognition purposes in large volumes of data. According to literature, statistical and data mining techniques have been used for purposes like bankruptcy prediction (Wilson and Sharda; 1994), educational placement of students (Lin, Huang and Chang; 2004), supporting marketing decisions for target marketing of solo mailings (Levin, Zahavi and Olitsky; 1995) and (Kim and Street; 2004), assessing consumer credit risk (Hand and Henley; 1996) and

customer credit scoring (Hand and Henley; 1997). Different data mining and statistical classification methods have been analyzed for a comparative assessment of classification methods (Kiang; 2004), (Chiang, Zhang and Zhou; 2004) and (Asparoukhov and Krzanowski; 2001). Comparisons have been made between different statistical classification models based on misclassification rates for different data conditions (Finch and Schneider; 2006) and (Meshbane and Morris; 1996).

The objective of this thesis is to draw a comparison between the results obtained on a given set of data when a classification model is built using three different statistical and data mining methods viz., logistic regression, decision tree and neural network models and compare the accuracy and validity of prediction. This thesis also shows the effect of different sample sizes and sampling methods used for the same model and tries to draw a conclusion regarding the influence of sample sizes and sampling methods on classifying data into proper groups.

The datasets used for the analysis for this thesis has been taken from the Louisiana Motor Vehicle Traffic Crash database supplied by the Department of Public Safety and Corrections, Highway Safety Commission of the State of Louisiana.

The data mining classification models used will be Decision Tree model using “Entropy” algorithm for growing the trees and “Standard Error Rule” algorithm for pruning the trees and Neural Network model using multilayer feed forward network (perceptron) architecture with back propagation algorithm. Louisiana Motor Vehicle crash data for two years viz., 2001 and 2002 will be used. The data for year 2001 will be primarily used to build the model whereas the data for year 2002 will be used to test the models. In the original data set, some of the variables are continuous and some are categorical. But each variable involved in the analysis will be converted into categorical variable by defining ranges and assuming certain conditions. A classification model will be built for each of the following dependent variable: 1) Alcohol, 2) Seat Belt usage, 3) Fatality and 4) Single/Multiple vehicle collision. A different set of independent variables will

be used for classifying each of the dependent variables which is the determined as the best variable subset by the statistical methods previously reviewed.

From the Louisiana Motor Vehicle crash data, we have a population of around 20000 observations for each year. For each of the dependent variable, classification models would be developed using the data for year 2000 using 8 different sample sizes viz., 200, 400, 800, 1000, 5000, 10000, 15000 and 20000 and the models would be tested on the data for years 2001 and 2002 to observe the effect of sample sizes on the accuracy of prediction of the dependent variable into the correct group. Also, for the optimum sample size for which best results are obtained, two different methods of sampling viz., random and stratified would be used to observe whether the method of sampling makes any difference in the accuracy of prediction.

By doing the above mentioned analyses, it is expected that we would be able to identify a classification model which works best for the given data and obtain an optimum sample size.

1.1 Contribution of the Research

The literature shows that many studies have been conducted which compares the efficiency of different data mining and statistical methods in classifying data instances into correct groups. A key study in this respect has been done by Kiang (2003) deal with the performance assessment of a few well known classification methods by running the models on synthetic data. The study focuses on the effect of data characteristics on the model performance, where the data characteristics are artificially modified to introduce imperfections like nonlinearity, multicollinearity and unequal covariance. A study by Shavlik, Mooney and Towell (1991) compares the performance of two data mining methods and studies the effect of size of training data on performance and conclude that neural networks can be trained better on small sizes of training data and also that ID3 performed better if the examples are converted to binary representation. Other studies comparing the performance of different data mining or statistical methods have been performed which looked at some or other data characteristics but none of

these studies have looked systematically at the relationship of sample size or the sampling method to the data classification accuracy, especially when the dependent variable is binary and all the predictors are either binary or categorical variables. Some of the studies like the one conducted by Asparoukhov et al. (2001) does perform a comparison of discriminant procedures for binary variables by considering different sets of predictor variables but it does not address the issue of sample size. This study focuses on mainly on the effect of sample sizes and the sampling techniques on the classification accuracy of the three methods viz., logistic regression, decision tree and neural network and look at the performance of each model at different sample sizes for different sampling methods. This study also tries to show that the information content of a dataset is not necessarily dependent only on the size of the dataset. The classification accuracy of a model and its ability to classify independent sets of test data is dependent on the information content of the training dataset that the model is built on, so building a model with a bigger training dataset does not imply better performance.

Also, by running the models for different sample sizes on three different data sets where the ratio of “0” values and “1”values of the dependent variables are quite distinctively different, an effort has been made to study whether there is any difference in the classification accuracies of the three different models depending on this ratio. A similar study was done by Meshbane et al. (1996) where they saw that when the size of one population is much larger than the other, hit-rate is improved by choosing logistic regression model if interest is in classification accuracy of the larger group and choosing predictive discriminant analysis if interest is in classification accuracy of the smaller group. But they have not studied the effect of a hugely disproportionate 0/1 distribution with respect to neural network or decision tree models. This study intends to do so.

Again, unlike any other study, the models built with different sizes of training data have been validated on two different sets of real world test data to verify whether the results are consistent and replicable. The performance of the models on training data alone is not enough to prove the efficacy of the model unless the results are replicable.

Since the kind of study performed for this thesis has never been done before, this study should prove to be a useful contribution towards the knowledge of classification criteria for binary data, especially from the data mining perspective. This study shows that the information content of a training dataset determines the prediction accuracy and that is not dependent on the size of the training data. Also, the distribution of “0”s and “1”s is a factor in determining what method could best classify a given set of data. This study also shows whether the sampling strategy for a particular method and for a particular dataset is important in improving the classification accuracy.

1.2 Organization of the Research

This research is organized into five chapters. In Chapter 2 a review of relevant background literature is discussed which provides the groundwork for the research. In Chapter 3, the methods used for the research is elaborated including the data used, the organization and choice of data variables, conversion of data to suit the research objective and different classification models. Chapter 4 analyzes and discusses the results and performance of the models described in Chapter 3 for various sample sizes followed by a summary and conclusion for the research in Chapter 5.

2. REVIEW OF THE LITERATURE

Data mining and statistical techniques have been used in a large number of areas, especially for business purposes to detect certain patterns in a given population of data. Data mining techniques are very helpful in detecting underlying patterns from large volumes of data.

Data mining technique can be used in bankruptcy prediction as shown by Wilson and Sharda (1994). A major evolution in the studies utilizing financial ratios for bankruptcy prediction was to identify the financial and economic predictors which improve the predictive performance, and two statistical techniques had been used the most: discriminant analysis and logistic regression (Bell, Ribar and Verchio, 1990). Wilson et al., compare the predictive capability of firm bankruptcy using neural networks and classical multivariate discriminant analysis. Discriminant analysis is a statistical technique used to construct classification schemes so as to assign previously unclassified observation to the appropriate group (Eisenbeis and Avery, 1972). But the underlying assumption for the technique is that the discriminating variable has to be jointly distributed according to a multivariate normal distribution. Wilson and Sharda use a number of financial ratios in a multivariate discriminant analysis and contrast it with the predictive capability of neural network which is a data mining methodology to show that neural networks performed significantly better than discriminant analysis to predict firm bankruptcy.

Statistical techniques have been used to predict the correct placement of a student in the appropriate group as shown by Lin, Huang and Chan (2004). Lin et al. have considered five science-educational indicators for each student who is intended to be placed in three reference groups, viz., advanced, regular and remedial, and have compared several discriminant techniques including Fisher's discriminant analysis and kernel-based non-parametric discriminant analysis using five school datasets. Though they have taken care of sampling variation on the resulting error rate by conducting an identical set of analyses on 500 bootstrap samples from School 5 dataset, the study does not show the effect of sample sizes on prediction accuracy. The study

shows that a kernel-based nonparametric procedure performs better than Fisher's discriminant rule.

In the same line, Finch and Schneider (2006) have conducted a study comparing classification accuracy of linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), logistic regression (LR) and classification and regression trees (CART) under a variety of data conditions. Statistical methods for predicting group membership based on a set of measurements have been shown to be very useful in a variety of conditions by Wilson and Handgrave (1995). Decisions regarding admission to various academic programs, entry into treatment regimens and identification of children at risk for academic failure or behavioral problems were often made with the help of statistical prediction techniques such as predictive discriminant analysis (PDA) or logistic regression (Abedi, 1991; Baird, 1975; Remus & Wong, 1982). PDA has two forms – linear (LDA) and quadratic (QDA). LR is an alternative to PDA and it models the odds of being in one group versus the other as a function of the predictor variable. The CART is a truly non-parametric method because there are no assumptions regarding the underlying distribution from which the subjects are drawn. Williams, Lee, Fisher and Dickerman (1999) found that both LR and LDA were better at predicting group membership than CART and that QDA performed worse than the other three. But the issue that had not been addressed was the classification accuracy of any of these procedures when one or more of the predictor variables are categorical instead of continuous. Huberty (1994) recommended using 0 to 1 assignment (dummy coding) and including the variable in the set of predictors when one of the predictors is binary in nature. This approach was supported by earlier work Bryan (1961) and Maxwell (1961). Johnson and Wichern (2002) suggested that LR might be preferable to LDA when one of the variables is of this type. Finch et al. conducted this study using Monte Carlo simulations to compare classification accuracy of LDA, QDA, LR and CART and found that QDA approach had a misclassification rate which was never larger than LDA and LR and in many cases it was lower. When the assumptions of LDA were met, i.e., the data was normally distributed and the

covariance matrices of the groups were equal, LDA. LR and QDA had comparable misclassification rates. However, they saw that CART had higher error rates than the other three. The error rates for LDA and LR went up if the data conditions were not met, while QDA and CART's misclassification rates declined when the covariance matrices were not equal.

Similar study for comparing cross-validated classification accuracies of predictive discriminant analysis and logistic regression classification models under varying data conditions for a two-group classification problem have been done by Meshbane and Morris (1996). Among the methods used for solving two-group classification problems, logistic regression (LR) and predictive discriminant analysis (PDA) are two of the most popular (Yarnold, Hart and Soltysik, 1994). Several studies have compared the classification accuracy of LR and PDA but the results have been inconsistent. Results of three simulation studies (Baron, 1991; Bayne, Beauchamp, Kane and McCabe, 1983; Crawley, 1979) suggest that LR is more accurate than PDA for non-normal data. However, several researchers (Cleary and Angel, 1984; Dey and Astin, 1993; Knoke, 1982; Krzanowski, 1975; Press and Wilson, 1978) found little or no difference in the accuracy of the two techniques using non-normal data. Findings are also inconsistent for degree of group separation. Bayne et al. (1993) found that larger group separation favored PDA while Crawley (1979) found this condition to favor LR. Sample size is yet another data condition yielding inconsistent results. In a simulation study, Harrell and Lee (1985) found that PDA was more accurate than LR for small samples while in a study by Johnson and Seshia (1992) using real data, LR worked better than PDA for small samples. Meshbane et al. (1996) proposed a method whereby separate-group as well as total-sample proportions of correct classifications could be compared for the two models using McNemar's test for contrasting correlated proportions and showed that neither theoretical nor data-based considerations were helpful in predicting which of the models would work better.

In their study, Hand and Henley (1997) conducted a review of different statistical classification methods used for credit scoring i.e., classifying applicants for credit into 'good' and 'bad' risk classes. The authors examined particular problems arising in the credit scoring context and reviewed the statistical methods which have been applied. Hand et al., mention in the study that historically discriminant analysis and linear regression have been most widely used techniques for building score-cards. The first published account of the use of discriminant analysis to produce a scoring system seems to be that of Durand (1941) who showed that the method could produce good predictions of credit replacement. Myers and Forgy (1963) had compared discriminant analysis and regression analysis for credit scoring and Grablowsky and Talley (1981) compared linear discriminant analysis and probit analysis for the same purpose. Orgler (1970) used linear regression analysis in a model for commercial loans and Orgler (1971) used regression analysis to construct score-card for evaluating outstanding loans and found that behavioral characteristics were more predictive of future loan quality than are application characteristics. Wiginton (1980) gave one of the first published accounts of logistic regression applied to credit scoring in comparison to discriminant analysis and concluded that logistic regression gave a superior result. Rosenberg and Gleit (1994) described several applications of neural networks to corporate credit decisions and fraud detection and Davis, Edelman and Gammerman (1992) compared such methods with alternative classifiers. Non-parametric methods, especially nearest neighbor methods, have been explored for credit scoring applications by Chatterjee and Barcun (1970) and Hand (1986). In addition to the mentioned methods, Hand et al., also considered mathematical programming methods, recursive partitioning, expert systems and time varying methods, summarized the various methods in their study, assessed the relative strengths and weaknesses of the methods and have drawn the conclusion that there is no overall 'best' method. What is best depends on the details of the problem: on the data structure, the characteristics used the extent to which it is possible to separate the classes by using those characteristics and the objective of the classification (overall misclassification rate, cost-weighted

misclassification rate, bad risk among those accepted, some measure of profitability, etc.). If the classes are not well separated, then Pr (good risk|characteristic vector) is a flat function, so that the decision separating the classes can not be accurately estimated. In such circumstances, highly flexible methods such as neural networks and nearest neighbor methods are vulnerable to over fitting the design data and considerable smoothing must be used. Nearest neighbor methods are effective with regard to the speed of classification. Neural networks are well suited to situations where there is a poor understanding of the data structure. If there is a good understanding of data structure and the problem, methods which make use of this understanding, such as regression, nearest neighbor and tree-based methods are expected to perform better. The authors infer that in credit scoring, since people have been constructing score-cards on similar data for decades, there is solid understanding and hence, neural networks have not been adopted as a regular production system.

Henley and Hand (1996) have also studied the application of k-nearest-neighbor (k-NN) method, a standard technique in pattern recognition and nonparametric statistics, as a credit scoring techniques for assessing the credit worthiness of consumer loan applicants. The k-NN method is a standard non-parametric technique used for probability density function estimation and classification and was originally proposed by Fix and Hodges (1952) and Cover and Hart (1967). Henley et al. proposed this study to provide a practical classification model that can improve on traditional credit scoring techniques. They proposed an adjusted version of the Euclidean distance metric which attempted to incorporate knowledge of class separation contained in data. To assess the potential of this method, Henley et al., drew a comparison k-NN with linear and logistic regression and decision trees and graphs and showed that the k-NN method with adjusted Euclidean metrics can give slightly improved prediction of consumer credit risk than the traditional techniques, achieving the lowest expected bad risk rate.

It has been observed that most cases that are misclassified by one method can be correctly predicted by other approaches (Tam and Kiang, 1992). A study on comparative analysis of ID3

and neural networks conducted by Dieterrich, Hild and Bakiri (1995) also had similar observations. Breiman (1996) studied the instability of different predictors and concluded that neural networks, classification trees and subset selection in linear regression were unstable while the k-th nearest neighbor method was stable.

A study to compare discriminant procedures for binary variables has been done by Asparoukhov and Krzanowski (2001). Thirteen discriminant procedures were compared by applying them to five real sets of binary data and evaluating their leave-one-out error rates (Lachenbruch and Mickey, 1968). Asparoukhov et al., have also taken into consideration the role of the number of variables in the investigation of classifier effectiveness and have used three versions of each data set containing 'large', 'moderate' and 'small' number of variables and to achieve the later two categories , variable reduction using all-subsets approach based on Kullback's information divergence measure (Hills, 1967) was used. The thirteen classifiers used were Independent binary model (IBM), linear discriminant function (LDF), logistic discrimination (LD), mixed integer programming bases classification (MIP), quadratic discriminant function (QDF), second-order log-linear model (LLM(2)), second-order Bahadur (Bahadur(2)) model, Hill's nearest neighbor estimator (kNN-Hills), adaptive weighted near neighbor estimator, kernel estimator (Kernel), Fourier procedure, multilayer perceptron neural network (MLP) and learning vector quantization neural networks (LVQ). A study by Anderson (1984) shows that under the assumptions of multivariate normal distributions with known parameters and equal covariance matrices in the classes, linear classifiers provide optimal classification. Fisher's (1936) LDF with unbiased estimates in place of unknown parameters maximizes the ratio of the between-sample variance to the within-sample variance. Logistic discrimination, a semi-parametric method avoids the problems of density estimation by assuming a logistic form for the conditional probability (Cox, 1966; Day and Kerridge, 1967; Anderson, 1972). Various nonparametric mathematical programming (MP) – based techniques facilitate a geometric interpretation and a number of studies (Duarte Siva, 1995; Joachimsthaler and Stam,

1988, 1990; Koehler and Erenguc, 1990; Rubin, 1990) have confirmed that MP methods can yield effective classification rules under certain non-normal data conditions, for instance, if the data set is outliers-contaminated or highly skewed. Log-linear models are well-known techniques for analysis of contingency tables and allow the logarithm of the probability of the dependent variable to be estimated as a linear function of main effects and interactions between binary variables (Argesti, 1990). MLP is a popular technique (Ripley, 1994) and the most widely used techniques for the minimization of MLP error criterion is the back-propagation algorithm (Hertz, Krogh and Palmer, 1991). LVQ neural network (Kohonen, 1990) drastically reduces the number of computations at every classification decision. The classification rule is: allocate the given observation to the closest codebook class in terms of Euclidean distance. In their study, Asparoukhov et al. concluded that the traditional statistical classifiers were not well able to cope with small sample binary data but the non-traditional (MLP, LVQ, MIP) classifiers did much better under those circumstances.

Another interesting study for comparison between neural networks and logistic regression for predicting patronage behavior towards web and traditional stores has been done by Chiang, Zhang and Zhou (2006). Different kinds of empirical studies for predicting customer preference for online shopping have been done (Degeratu, Rangaswamy and Wu, 2000; Bellman, Lohse and Johnson, 1999; Kwak, Fox and Zinkhan, 2002). According to Urban and Hauser (1980), these studies are forms of “preference regressions” and they all share the same a priori assumption that the process of consumers’ channel evaluation is linear compensatory, i.e., those models assume that any shortfall in one channel attribute (e.g., immediate possession of a product) can be compensated by enhancements of other channel attributes (e.g., price). Studies show that consumers might judge alternatives based on only one or a few attributes and the process of evaluation might not always be compensatory (Johnson, Meyer and Ghose, 1989; Payne, Bettman and Johnson, 1993). Chiang et al., developed neural network models which are known for their known capability of modeling non-compensatory decision processes and tried to find out whether

non-compensatory choice models using neural network perform better than logit choice models in predicting consumer's channel choice between web and traditional stores. The authors show that for most of the selected products, neural networks significantly outperform logistic regression models in terms of predictive power. Studies by Fadlalla and Lin (2001), Hung, Liang and Liu (1996) and West, Brockett and Golden (1997) also show that in most of the applications where neural networks have been used to model business problems in support of finance and marketing decision-making, neural networks have outperformed traditional compensatory models such as discriminant and regression analysis.

Study has also been done to help make marketing decisions by targeting the right audience for sending promotional materials from among a very large marketing database based on customers' attributes and characteristics by Levin, Zahavi and Olitsky (1995) using a hybrid system called AMOS (Automatic Model Specification). Levin et al. developed AMOS as a fully automatic hybrid system involving traditional statistical and optimization models where a probabilistic approach to model response has been used, which expresses the customer's likelihood of purchase by well defined purchase probabilities. The method used in AMOS to estimate the choice probability (customer's) is a discrete-choice logistic-regression model. Levin et al. tested the AMOS system to show that AMOS targets the mailing better, increasing the return on sales by 5.5%.

In line with the study of Levin et al., Kim and Street (2004) conducted a study for market managers for targeting customers using a data mining approach. Kim et al., used artificial neural networks (Riedmiller, 1994) guided by genetic algorithms (Goldberg, 1989) to develop their predictive model. Genetic algorithms have been known to have superior performance to other search algorithms for data sets with high dimensionality (Kudo and Sklansky, 2000). The key determinants of customer responses were isolated by selecting different subsets of variables using genetic algorithms and those selected variables are used to train different neural networks. The result was a highly accurate predictive model that used only a subset of the original features, thus

simplifying the model and reducing the risk of over-fitting. Kim et al., show that their system maximized the hit rate at fixed target point and also selected a best target point where expected profit from direct mailing was maximized.

Berardi, Patuwo and Hu (2004) presented a principled approach for building and evaluating neural network classification models for decision support system implementation and e-commerce application in their study. The study aimed at understanding how to utilize e-commerce data for Bayesian classification within a neural network framework to yield more accurate and reliable classification decisions and showed that neural networks are ideally suited for noisy data like e-commerce data. In a similar study, Chu and Widjaja (1994) showed that neural networks using a back-propagation based forecasting prototype can be effectively used as a forecasting tool.

A key study with respect to comparative assessment of classification methods has been done by Kiang (2003). In this study Kiang has considered data mining classification techniques viz., neural networks and decision tree models and three statistical methods – linear discriminant analysis (LDA), logistic regression analysis and k-nearest-neighbor (kNN) models, and used synthetic data to perform a controlled experiment in which the data characteristics are systematically altered to introduce imperfections such as nonlinearity, multicollinearity, unequal covariance, etc. The study was performed to investigate how these different classification methods performed when certain assumptions about the data characteristics were violated and Kiang showed that data characteristics considerably impacted the classification performance of the methods. Also, the study conducted by Shavlik, Mooney and Towell (1991) added on in this line by empirically analyzing the effects of three factors on the performance of two AI methods, neural networks and ID3. The three factors considered were size of training data, imperfect training examples and encoding of the desired outputs. Shavlik et al. showed that neural networks performed well with small sizes of training data but they did not emphasize much on the distribution of the data instances. This aspect was looked at by Meshbane et al. (1996) where they

found that when the size of data instances with either a “0” or a “1” is much larger than the other, hit-rate is improved by choosing logistic regression model if interest is in classification accuracy of the larger group and choosing predictive discriminant analysis if interest is in classification accuracy of the smaller group. In a similar line Rendell and Cho (1990) examined the effects of six data characteristics on the performance of two classification methods, ID3 and PLSI (probabilistic learning system). The factors considered in their study include size of training set, number of attributes, scales of attributes, error or noise, class distribution and sampling distribution. The study conducted for this thesis intends to add a new dimension to the finding of these papers by looking at the optimum sample size that is required to train a decision tree, neural network or a logistic regression model and also looks at effect of sampling strategy on the performance of the models. The study also looks at the effect of the ratio of the binary values of the dependent variable in the training data set and how it affects the prediction performance of the three models.

3. METHODS

Three different models have been considered for our research purpose. Two data mining methods viz., decision tree and neural network and one statistical method viz., logistic regression method. The data mining software Insightful Miner version 7.0 has been used for the purpose of building the models. Three sets of analyses have been done using three sets of data. All the three analyses have been done on each of the three datasets for different sample sizes and two different sampling methods viz., simple random sampling and stratified sampling. The data used has been taken from Louisiana Motor Vehicle Traffic Crash database supplied by the Department of Public Safety and Corrections, Highway Safety Commission of the State of Louisiana and from the crash database provided by the Federal government of USA.

3.1 The Data

Louisiana State Government and Federal State Government crash database consists of records of all the recorded accidents and any pertinent data in relation to the accidents. There are six different tables in the Louisiana state database, viz., CRASH_TB, VEHIC_TB, OCCUP_TB, PEDES_TB, TRAIN_TB and TROCC_TB containing the crash details, details of the vehicles involved in the crash, occupant details, details of the pedestrians involved in the crash, details of the train involved in the crash if any and details of train occupants involved in the crash if any, respectively. Each table has a large number of variables.

For the purpose of the analyses for this research, three variables have been chosen as the dependent variables for three different datasets, the details of which are given as follows:

3.1.1 Alcohol Dataset

The first data set shall be referred to as ‘Alcohol’ dataset hereafter and the purpose of analysis for this is to predict correctly whether alcohol is involved in the crash and is a reason for

the crash. The predictor variables used for this analysis have been chosen on a commonsense basis and not on a statistical best-subset basis. For example, to predict whether the blood alcohol test of the driver produced a positive or negative result, predictors like police reported alcohol involvement, hour of the day (alcohol is more likely to be a reason if it is night time), day of the week (more likely during the weekend), injury severity (if alcohol is involved, injury is likely to be more severe, probably fatal), restraint system used (seat belt use not likely if alcohol involvement present), age of the driver (irresponsible driving more likely at teenage), etc. are likely to play a major role. The variables have been converted into categorical variables as this is a requirement for the predictor variables while using decision trees. The list of predictor names used for this analysis along with their descriptions, data types, possible values and conversion rules are given at the Appendix, Table #1.

The data for two years viz., 2001 and 2002 have been considered for the analyses and the models have been built using samples from the data for year 2001. The dependent variable ALC_RES has three possible values, viz., 0, 1 and 2. We are mainly interested with the classes 0 and 1 for ALC_RES. Also, since decision trees can be run for binary variables only, the dataset is cleaned before building the model by removing all records with ALC_RES = 2. There are approximately over 25,000 observations for each of the years after cleaning the datasets. Sample sizes of 200, 400, 800, 1000, 5000, 10000, 15000 and 20000 have been chosen to build the models once using simple random sampling and once using stratified sampling and each model has been validated separately against year 2001 data and year 2002 data. For stratification, driver's age, the DR_AGE variable has been chosen as a stratification variable as age is likely to play a major role in the prediction of alcohol involvement in a crash, to study the ramification on the prediction capability of the models.

When data characteristics is observed, it is seen that the distribution of the dependent variable ALC_RES in the final version of cleaned dataset is more or less uniform with number of instances of "1"s being more than 50% of the number of instances of "0's, both in the year 2001

and year 2002 datasets. This forms the basis of better predictability for data mining models as will be seen later in the models.

3.1.2 Seatbelt Dataset

The second data set shall be referred to as 'Seatbelt' dataset hereafter. The purpose of this set of analyses is to study whether the seat belt usage of the driver can be predicted accurately with the use of a set of variables. As in the first case, a set of predictors have been chosen from the crash database on a common sense basis. Variables like the most severe injury to the driver, the age of the driver, the race of the driver, the extent of damage to the vehicle at the first impact area, presence of alcohol/drugs, sex of the driver, etc, are thought to have a probable influence on the predictability of seatbelt usage. The extent of the importance of the predictors and their predictability is studied in these analyses. The variables have been converted into categorical variables as this is a requirement for the predictor variables while using decision trees. The list of predictor names used for this analysis along with their descriptions, data types, possible values and conversion rules are given at the Appendix, Table #2.

For this dataset also, data for two years viz., 2001 and 2002 have been considered for the analyses and the models have been built using samples from the data for year 2001. The dependent variable DR_PROTSYS_CD has three possible values, viz., 0, 1 and 2. We are mainly interested with the classes 0 and 1 for DR_PROTSYS_CD. Also, decision trees can be run for binary variables only. So, the dataset is cleaned before building the model by removing all records with DR_PROTSYS_CD = 2. After cleaning, the dataset for 2001 has approximately 20,000 observations and there are around 27,000 observations for year 2002. Sample sizes of 200, 400, 800, 1000, 2000, 5000, 10000, 15000 and 20000 have been chosen to build the models once using simple random sampling and once using stratified sampling and each model has been validated separately against year 2001 data and year 2002 data. For stratification, driver's age, viz. the DR_AGE variable has been chosen as a stratification variable as age is likely to play a

major role in the prediction of seatbelt usage in a crash, assuming that teenagers are more likely to disobey the seatbelt rule.

The distribution of the dependent variable DR_PROTSYS_CD in the final cleaned version of the datasets for both years 2001 and year 2002 is very much skewed with the number of instances of "0"s being only about 6-8% of the number of instances of "1"s. This may pose a problem for the classification of data with the data mining models.

3.1.3 Fatality Dataset

The third dataset would be termed as 'Fatality' as the motive of the analyses is to study whether a set of predictors are able to predict correctly whether an accident is fatal or non-fatal. As before variables such as alcohol involvement in the crash, previous violations of the driver, number of occupants wearing a seatbelt in the crash, number of vehicles involved in the crash, etc are assumed to be likely to have a correlation to the dependent variable and are considered as the predictor variables. The importance of the predictor variables in classifying the dependent variables and the accuracy of prediction is studied in the analyses. The variables have been converted into categorical variables as this is a requirement for the predictor variables while using decision trees. The list of predictor names used for this analysis along with their descriptions, data types, possible values and conversion rules are given at the Appendix, Table #3.

Fatality is denoted by the variable SEVERITY_CD which is used to designate the most severe injury in the crash. Code "A" is for a fatal crash, "B" for incapacitating/severe, "C" for non-incapacitating/moderate, "D" for possible/complaint and "E" for no injury. Since we are interested in the capability of the independent variable in predicting a fatal crash correctly, records with SEVERITY_CD of "A", "B" or "C" only have been chosen from the datasets of two years viz., 2001 and 2002 for the analyses and the models have been built using samples from the data for year 2001. There are approximately over 13,000 observations for each of the years with a SEVERITY_CD of "A", "B" or "C". The codes "A" and "B" have been grouped into

group “1” and “C” into group “0”, since we assume that an incapacitating or severe injury is as good as a fatal injury and it is just a matter of chance that the driver or passenger survived instead of getting killed. Since the population is 13,000, sample sizes of 200, 400, 800, 1000, 2000, 5000 and 10000 have been chosen to build the models once using simple random sampling and once using stratified sampling and each model has been validated separately against year 2001 data and year 2002 data. For stratification, alcohol i.e., the EST_ALCOHOL variable has been chosen as a stratification variable as alcohol is assumed to be likely to play a major role in a fatal crash. The choice of the stratification variable is also ratified by the results of the models with random sample where alcohol involvement is seen to be the most important variable in predicting the fatality of the crash.

The distribution of the dependent variable SEVERITY_CD in the final cleaned version of the datasets for both years 2001 and year 2002 is very much skewed with the number of instances of “0”s being only about 7% of the number of instances of “1”s. This may pose a problem for the classification of data with the data mining models

3.2 Decision Tree

Decision trees are powerful and popular tools for classification and prediction. They are attractive due to the fact that in contrast to other machine learning techniques such as neural networks, they represent rules that human beings can understand. Decision tree is a classifier in the form of a tree structure (as shown in fig 3.1) where each node is either a **leaf node**, indicating the value of the target attribute or class of the examples, or a **decision node**, specifying some test to be carried out on a single attribute-value, with one branch and sub-tree for each possible outcome of the test. A decision tree can be used to classify an example by starting at the root of the tree and moving through it until a leaf node is reached, which provides the classification of the instance.

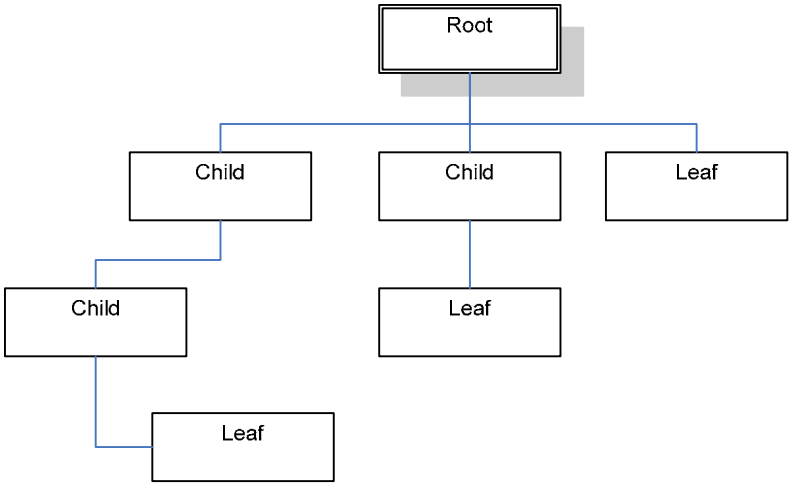

Figure 3.2.1 A Decision Tree

Decision trees represent a set of decisions. These decisions generate rules for classification of a dataset using the statistical criterion: entropy, information gain, Gini index, chi-square test, measurement error, classification rate, etc. There are two stages, tree construction and post-pruning, and five tree algorithms are in common use, viz., CART, CHAID, ID3, C4.5 and C5.0. Most algorithms that have been developed for learning decision trees are variations on a core algorithm that employs a top-down, greedy search through the space of possible decision trees.

The algorithm used for building the models for this thesis is CART i.e., Classification and Regression Tree. In this algorithm, the condition of split is Information Gain and involves the measurement of how much information one can win by choosing a certain variable when deciding upon the variable on the basis of which to split the tree. The measurement of information used is Entropy (in bits). The dependent variable has been converted into a binary variable and the independent variables have been converted into categorical variables and a binary split is done.

For measuring entropy the following assumptions are made:

- S is a sample of training instances
- P_p is the proportion of positive instances in S
- P_n is the proportion of negative instances in S

Entropy measures the impurity of S and is given as $\text{Entropy}(S) = – P_p \log P_p – P_n \log P_n$.

Entropy(S) is the expected number of bits needed to encode class (p or n) of a randomly drawn member of S under the optimal, shortest length-code because information theory states that optimal length code assigns $-\log_2 P$ bits to message having probability P. So, expected number of bits to encode p or n of a random member of S: $P_p (- \log P_p) + P_n (- \log P_n)$. The information gain Gain(S, A) is the expected reduction in entropy due to sorting on A and is given as:

$\text{Gain}(S, A) = \text{Entropy}(S) - \Sigma_{v \text{ in values}(A)} |S_v| / |S| \text{ Entropy}(S_v)$, where S_v is the set of training instances remaining from S after restricting to those for which attribute A has value v. So, when a branching of a decision tree occurs, the choice of the variable by which the split is made is base upon the condition of maximum information gain, i.e., the variable enabling the maximum information gain is chosen as the splitting variable. This process is repeated at each node until the leaf nodes are obtained.

A decision tree can be grown until every node is pure, i.e., the leaf nodes can be divided no further and the members within each leaf node belong to only one class. A maximal classification tree gives 100% accuracy on training data but it is a result of over fitting and would give poor prediction on test data. Tree complexity is a function of the number of leaves, the number of splits and the depth of the tree. A well-fitted tree has low bias and low variance. To avoid over fitting a tree needs to be right sized by either forward-stopping or stunting the growth or growing the tree to its full length and then pruning it back. For the analyses done for this research, the tree is grown and then pruned back using standard error rule. The error rate of an entire tree is the percentage of the records that are misclassified and the standard error rate pruning denotes the cutting off of weak branches, the ones with high misclassification rate which is measured on validation data (a separate set of data from the training data). Pruning the full tree increases the overall error rate for the training set, but the reduced tree will generally provide better predictive power for the test data.

3.3 Neural Network

A neural network is a software (or hardware) simulation of a biological brain (sometimes called Artificial Neural Network or 'ANN'). The purpose of a neural network is to learn to recognize patterns in a given data set. In the human brain, a typical neuron collects signals from others through a host of fine structures called *dendrites*. The neuron sends out spikes of electrical activity through a long thin strand known as an *axon*, which splits into thousands of branches. At the end of each branch, a structure called a *synapse* converts the activity from the axon into electrical signals that inhibit or excite activity in the connected neurons. When a neuron receives excitatory input that is sufficiently large compared with its inhibitory input, it sends a spike of electrical activity down its axon. Learning occurs by changing the effectiveness of the synapses so that the influence of one neuron on another changes.

These neural networks may be built by typically programming in a computer to emulate the essential features of neurons and their interconnections. However, because the knowledge of neurons is incomplete and computing power is limited, the models are necessarily gross idealizations of real networks of neurons. An important application of neural network is pattern recognition which can be implemented using a feed-forward neural network that has been trained accordingly. During training the network is trained to associate outputs with input patterns. When the network is used, it identifies the input pattern and tries to output the associated output pattern. The power of neural network comes to life when a pattern that has no output associated with it, is given as an input. In this case, the network gives the output that corresponds to a taught input pattern that is least different from the given input pattern.

Neural networks are capable of modeling extremely complex, typically non-linear functions. Each neuron has a certain number of inputs, each of which has a weight assigned to it. The weight is an indication of the importance of the incoming signal for that input. These weighted inputs are added together and if they exceed a pre-set threshold value, the neuron fires. The input value

received from a neuron is calculated by summing the weighted input values from its input links. An activation function takes the neuron input value and produces a value which becomes the output value for the neuron and is passes to other neurons in the network. This is called multilayer perceptron (MLP). The number of parameters in a MLP with one hidden layer with *h* neurons and *k* inputs is $h(k+1) + h + 1 = h(k+2) + 1$. By adjusting the weights on the connections between layers, the perceptron output can be "trained" to match a desired output. Weights are determined by adding an error correction value to the old weight. The amount of correction is determined by multiplying the difference between the actual output (x[j]) and target (t[j]) values by a learning rate constant C. If the input node output (a[j]) is a 1, that connection weight is adjusted, and if it sends 0, it has no bearing on the output and subsequently, there is no need for adjustment. The process can be represented as:

$W_{ij(new)} = W_{ij(old)} + C(t_j - x_j)a_i$, where C = learning rate. The training procedure is repeated until the network performance no longer improves.

For the analyses done for this thesis, a MLP neural network is employed, which is a feed-forward neural network using resilient propagation utilizing sigmoid activation functions. The number of iterations that the software runs has been configured to 50. Another task was to select the number of hidden layers and the number of nodes in each layer. Many studies have reported (Jain and Nag, 1997) no improvement of neural network performance with more than one hidden layer. It was confirmed in several trail sessions during an evaluation that compared the performance of each network with one or two layers for the analyses done here, a slightly improved performance was observed with two hidden layers. So, for this research, a MLP network with two layers has been considered. Also, though a large number of hidden nodes may increase training performance, but at the expense of generalization and computation cost. Here, the performance was experimented with a number of hidden nodes and ten nodes in a layer were chosen. The initial weights selected by the software are random and the final weights are the best weights obtained by error reduction at a convergence tolerance of 0.0001. The learning rate is set

at 0.001 and the weight decay at 10. The percent of sample data that the software uses to validate the model is set at 10 with 2 hidden layers and 10 nodes per hidden layer. Thus the activation function is a double sigmoid function as shown below:

$F(sum_j) = w_1/(1+ \exp(sum_j)) + w_2/(1 + \exp(sum_j))$, where sum_j is the scalar product of an input vector and weights to the node j either at a hidden layer or at the output layer and w_1 and w_2 are the initial weights.

3.4 Logistic Regression

A logistic regression model is used when the dependent variable is a categorical variable as in this case and the predictor variables may be continuous or categorical. This is semi-parametric model where there are no multivariate normality and equal dispersion assumptions required for the data. A logistic function of the following form is used:

$Y = 1 / (1 + e^y)$, $y = a + \Sigma_{i=1,n} b_i X_i$, where X_i represents the set of individual variables, b_i is the coefficient of the i^{th} variable, and Y is the probability of a favorable outcome. The outcome Y is a Bernoulli random variable.

4. RESULTS AND DISCUSSION

The results of the analyses performed on the three different datasets for the three different models are given as following:

4.1 Alcohol Dataset Analysis with Decision Tree

When the decision tree model was built using the Alcohol dataset using year 2001 crash data for different sample sizes and the sampling method used was simple random sampling, the analyses showed that the most important variable in classifying the variable ALC_RES into the correct class is DRINKING for all the sample sizes. The next important variables in terms of predicting ALC_RES differed when the sample sizes were different. The prediction rates also varied according to the sample size.

To test the effect of sample size on the results, a variation was also performed. When a sample size of 400 was chosen, the same sample was reproduced three times to make it a sample of 1200 and the decision tree model was run for this 1200 instances. The purpose was to study whether the sample size alone affected the results or was it the information contained within the sample. If the classification accuracy is governed by the sample size, the sample of 1200 would give a better result though the information content of the 1200 sample is same as that of the 400 sample.

Table 4.1.1 shows the summary of the results along with the importance of variables in predicting the dependent variable for the training data while Table 4.1.2 and Table 4.1.3 show the summary of results for different sample sizes when the models built for each sample size was applied to test the validity of prediction for the whole dataset for years 2001 and 2002 respectively. The graphs plotting the overall % agreement against the sample sizes for the training data and test data for years 2001 and 2002 are shown in Figure 4.1.2, Figure 4.1.2 and Figure

4.1.3 respectively. If the classification agreement % for the "1" and "0" values of ALC_RES is observed it is seen that they are comparable, given the ratio of "0" to "1" is less than 2:1.

Table 4.1.1 Decision Tree result on training Alcohol data (random sampling)

Sample Size	% Agree			Important Predictors in order of Relative Importance
	0	1	Overall	
200	90.0	78.6	86.0	drinking, hour
400	89.2	87.9	88.8	drinking, age, rest_use, body_typ
800	93.1	85.4	90.4	drinking, m_harm, age, rest_use, hour, body_typ
1000	92.2	81.4	88.4	drinking, hour, age, rest_use, body_typ
1200 (400*3)	95.4	90.0	93.5	drinking, age, rest_use, hour, body_typ, violchg1, ve_forms, inj_sev, day_week
5000	93.1	78.7	88.1	drinking, hour, age, rest_use, m_harm, body_typ
10000	93.5	77.3	87.8	drinking, hour, ve_forms, rest_use, age, body_typ, m_harm
15000	93.5	77.2	87.8	drinking, hour, age, m_harm, rest_use, ve_forms, body_typ, sex
20000	92.7	79.0	87.9	drinking, hour, m_harm, age, rest_use, body_typ

Table 4.1.2 Decision Tree result on year 2001 Alcohol data (random sampling)

Sample Size	% Agree		
	0	1	Overall
200	90.6	72.9	84.1
400	88.3	78.0	84.5
800	91.2	78.7	86.6
1000	89.8	80.2	86.3
1200 (400*3)	91.7	74.5	85.2
5000	93.0	77.5	87.2
10000	93.3	77.5	87.5
15000	93.4	76.8	87.3
20000	92.6	78.6	87.4

Table 4.1.3 Decision Tree result on year 2002 Alcohol data (random sampling)

Sample Size	% Agree with test data		
	0	1	Overall
200	91.1	74.7	85.0
400	88.7	79.0	85.1
800	91.6	79.8	87.1
1000	90.4	81.6	87.1
1200 (400*3)	91.1	73.5	84.6
5000	93.4	79.0	88.0
10000	93.7	79.5	88.3
15000	93.9	78.5	88.1
20000	93.2	79.8	88.1

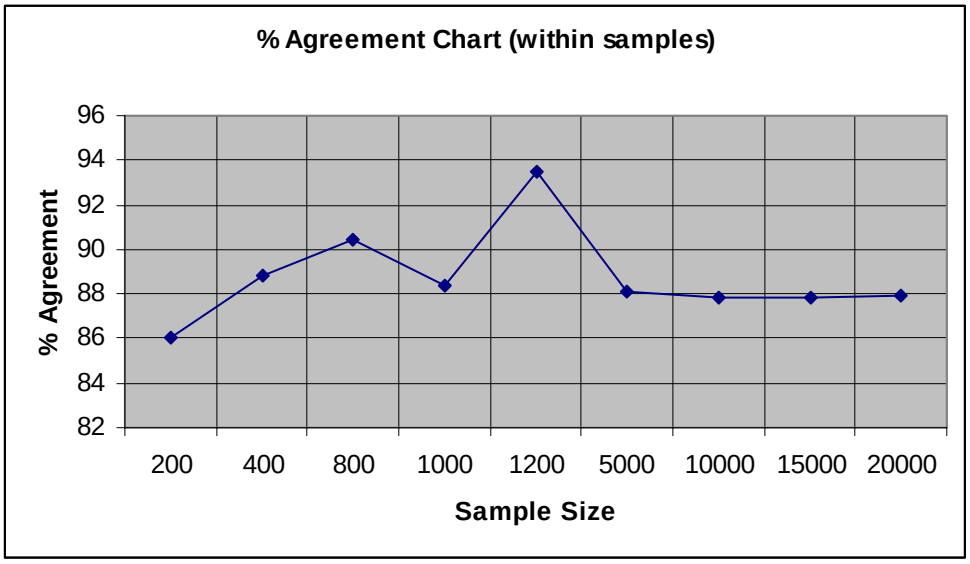

Figure 4.1.1 Decision Tree result on training Alcohol data (random sampling)

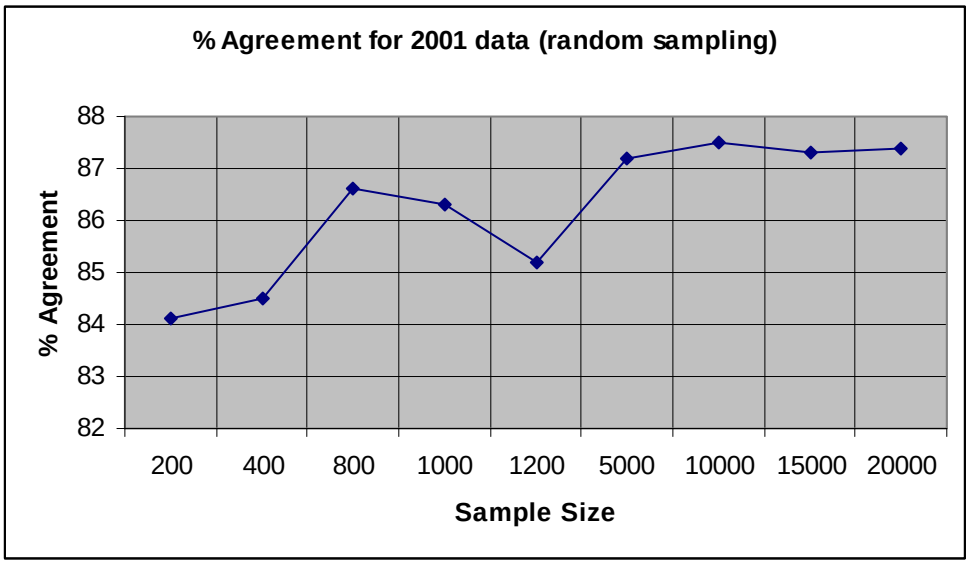

Figure 4.1.2 Decision Tree result on year 2001 Alcohol data (random sampling)

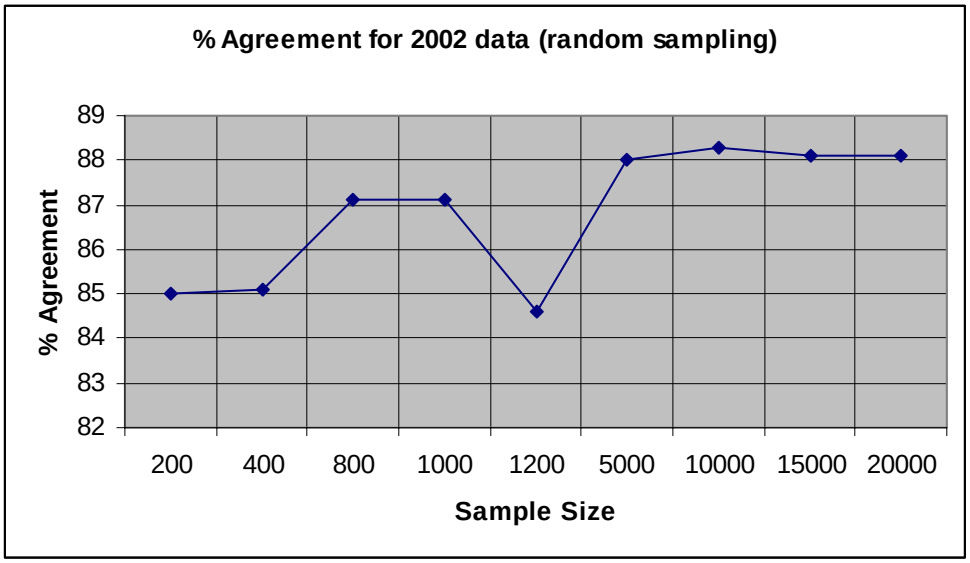

Figure 4.1.3 Decision Tree result on year 2002 Alcohol data (random sampling)

Thus it is seen that the overall prediction classification accuracy for the training data was higher than that for the test data for both the years for all sample sizes. For the sample size of 1200 (400 sample size repeated three times), it is observed that the prediction accuracy shoots up to 94% for the training data giving an impression that increasing the sample size gives a better classification accuracy. But if the graphs for the test data results are observed, for both the test datasets, it is seen that the classification accuracy falls for the sample size 1200. Thus, it shows that the impression that was obtained by observing the training data results is false. The actual information contained in a sample influences the classification accuracy of a decision tree model. The information contained in the sample of size 1200 was the same as that in the sample of size 400.

If the result for sample size 1200 is ignored, it is seen that classification accuracy reached a plateau at the sample size of 1000 for training data and not much could be gained in terms of prediction accuracy by increasing the sample size over 1000. The overall classification accuracy for the training data at the sample size of 1000 was around 88%. But when the test results are observed, it is seen that a plateau is reached at the sample size of 5000, where the classification

accuracy was around 87% for 2001 data and 88% for 2002 data and increasing the sample size beyond 5000 did not help in predicting the test data more accurately. By running the model for the full datasets for both the years 2001 and 2002, it was observed that the classification accuracy was replicated.

When the decision tree models were built by using a stratified sampling method, stratifying by the driver's age variable, DRINKING was found to be the most important variable in classifying the dependent variable, as in the case of random sampling. The next best predictor varied according to the sample sizes and the prediction accuracies also varied according to the sample sizes. Table 4.1.4 shows the summary of the results along with the importance of variables in predicting the dependent variable for the training data while Table 4.1.5 and Table 4.1.6 show the summary of results for different sample sizes when the models built for each sample size was applied to test the validity of prediction for the whole dataset for years 2001 and 2002 respectively. The graphs plotting the overall % agreement against the sample sizes for the training data and test data for years 2001 and 2002 are shown in Figure 4.1.4, Figure 4.1.5 and Figure 4.1.6 respectively.

Table 4.1.4 Decision Tree result on training Alcohol data (stratified sampling)

Sample Size	% Agree			Important Predictors in order of Relative Importance
	0	1	Overall	
200	93.4	74.6	87.5	drinking, inj_sev, sex, rest_use, hour, body_typ, ve_forms
400	94.2	78.6	88.8	drinking, hour, rest_use
800	96.0	77.6	89.2	drinking, hour, ve_forms, age, body_typ, sex, rest_use
1000	93.8	80.4	88.6	drinking, hour, age, ve_forms, inj_sev
5000	92.5	78.3	87.2	drinking, hour, age, m_harm, rest_use, inj_sev, body_typ, ve_forms
10000	93.2	79.1	88.0	drinking, hour, age, rest_use, ve_forms, body_typ, inj_sev
15000	93.2	77.9	87.5	drinking, hour, m_harm, age, rest_use, body_typ
20000	93.0	77.7	87.3	drinking, hour, rest_use, age, body_typ, inj_sev, m_harm

Table 4.1.5 Decision Tree result on year 2001 Alcohol data (stratified sampling)

Sample Size	% Agree		
	0	1	Overall
200	90.9	76.2	85.4
400	92.8	75.4	86.3
800	93.6	74.2	86.4
1000	90.5	78.7	86.1
5000	92.5	78.3	87.2
10000	92.5	79.1	87.5
15000	92.7	77.8	87.2
20000	92.7	77.6	87.1

Table 4.1.6 Decision Tree result on year 2002 Alcohol data (stratified sampling)

Sample Size	% Agree with test data		
	0	1	Overall
200	91.5	77.9	86.4
400	93.5	77.2	87.3
800	94.2	76.3	87.5
1000	90.8	80.8	87.0
5000	93.2	79.9	88.2
10000	92.9	80.7	88.3
15000	93.2	79.3	88.0
20000	93.2	79.1	87.9

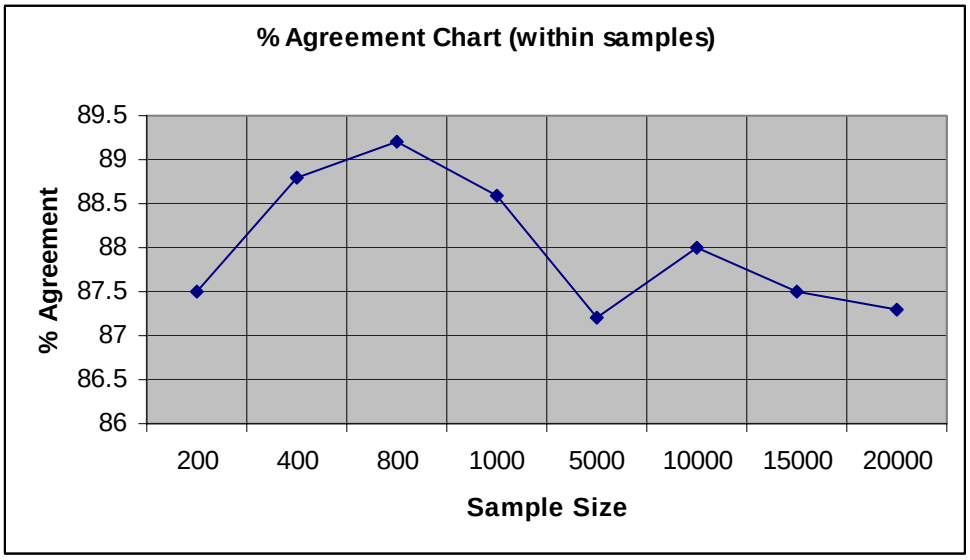

Figure 4.1.4 Decision Tree result on training Alcohol data (stratified sampling)

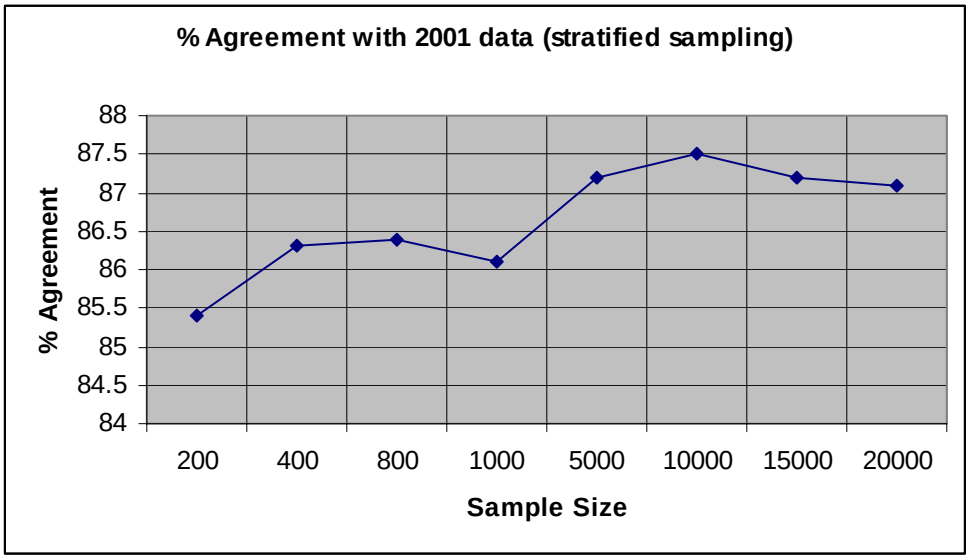

Figure 4.1.5 Decision Tree result on year 20012 Alcohol data (strat. sampling)

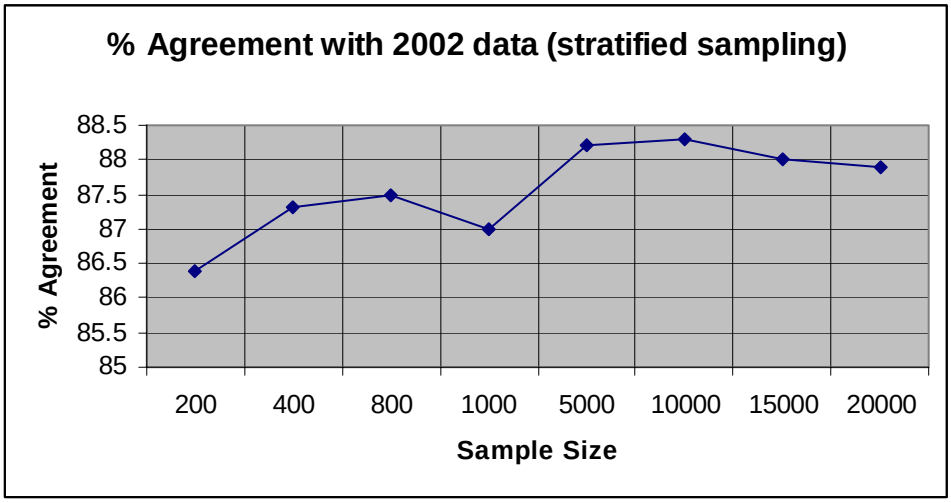

Figure 4.1.6 Decision Tree result on year 2002 Alcohol data (strat. sampling)

The results in case of stratified sampling show an interesting variation. The training data as well as the test data graphs show two-humped curves where the prediction accuracy reached a maximum of around 89% for training data and then faltered off. For test data, the classification accuracy reached a maximum value at the sample size of 5000 as in the case of random sampling method and did not improve any further by increasing the sample size. For 2001 data the

prediction accuracy at a sample size of 5000 was around 87% while that for 2002 data, it was 88%. Thus, it is seen that, even if the sampling method is stratified, the prediction accuracy is consistently replicated over different test datasets.

4.2 Alcohol Dataset Analysis with Logistic Regression

As in the case of decision trees, when the logistic regression analysis was performed using the Alcohol dataset for year 2001 crash data with different sample sizes and the sampling method used was simple random sampling, the analyses showed that the single most important variable in classifying the variable ALC_RES into the correct class is DRINKING for all the sample sizes. The next important variables in terms of predicting ALC_RES differed when the sample sizes were different. The prediction rates also varied according to the sample size. Table 4.2.1 shows the summary of the results along with the importance of variables in predicting the dependent variable for the training data while Table 4.2.2 and Table 4.2.3 show the summary of results for different sample sizes when the models built for each sample size was applied to test the validity of prediction for the whole dataset for years 2001 and 2002 respectively. The graphs plotting the overall % agreement against the sample sizes for the training data and test data for years 2001 and 2002 are shown in Figure 4.2.2, Figure 4.2.2 and Figure 4.2.3 respectively.

If the classification agreement % for the "1" and "0" values of ALC_RES is observed it is seen that they are comparable, given the ratio of "0" to "1" is less than 2:1.

Table 4.2.1 Logistic Regression result on training Alcohol data (random sampling)

Sample Size	% Agree			Important Predictors in order of Relative Importance
	0	1	Overall	
200	93.1	81.4	89.0	drinking, age, hour, rest_use, inj_sev, violchg1, body_typ, ve_forms, sex, day_week, m_harm
400	93.8	85.0	90.8	drinking, hour, age, rest_use, inj_sev, m_harm, body_typ, violchg1, sex, ve_forms, day_week,
800	93.1	82.1	89.2	drinking, age, hour, rest_use, ve_forms, body_typ, sex, violchg1, inj_sev, m_harm, day_week,
1000	93.8	78.9	88.6	drinking, hour, age, inj_sev, body_typ, rest_use, violchg1, m_harm, ve_forms, sex, day_week,

(table cont.)

5000	93.1	79.1	88.2	drinking, hour, age, rest_use, inj_sev, body_typ, sex, violchg1, m_harm, ve_forms, day_week,
10000	93.0	78.3	87.9	drinking, hour, age, body_typ, rest_use, inj_sev, sex, m_harm, violchg1, ve_forms, day_week,
15000	93.2	78.7	88.1	drinking, hour, age, body_typ, rest_use, inj_sev, sex, ve_forms, violchg1, m_harm, day_week,
20000	93.1	79.0	88.2	drinking, hour, age, body_typ, rest_use, inj_sev, sex, ve_forms, violchg1, m_harm, day_week,

Table 4.2.2 Logistic Regression result on year 2001 Alcohol data (random sampling)

	% Agree		
Sample Size	**0**	**1**	**Overall**
200	89.8	77.5	85.3
400	91.4	78.9	86.8
800	91.9	79.1	87.2
1000	92.5	78.0	87.1
5000	93.1	78.3	87.6
10000	93.0	78.6	87.7
15000	93.2	78.5	87.8
20000	93.1	78.6	87.8

Table 4.2.3 Logistic Regression result on year 2002 Alcohol data (random sampling)

	% Agree with test data		
Sample Size	**0**	**1**	**Overall**
200	89.7	78.4	85.5
400	91.9	80.3	87.5
800	92.2	80.0	87.6
1000	92.9	79.6	87.9
5000	93.7	79.6	88.4
10000	93.5	80.1	88.5
15000	93.7	79.8	88.5
20000	93.6	80.1	88.5

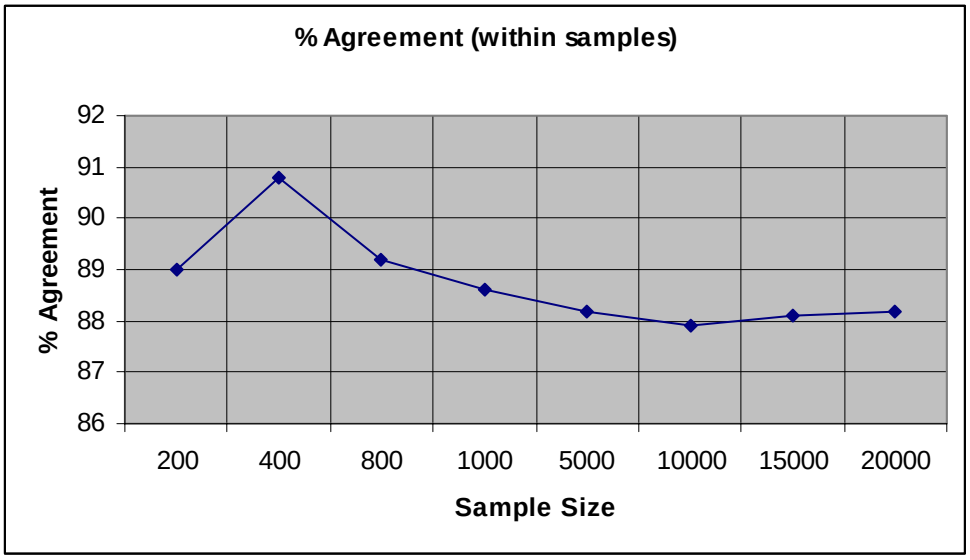

Figure 4.2.1 Logistic Regression result on training Alcohol data (random sampling)

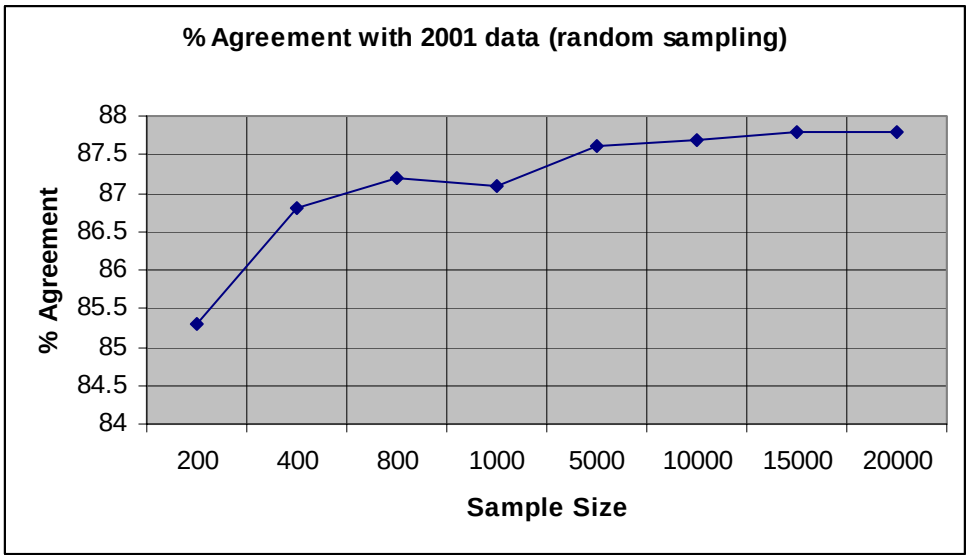

Figure 4.2.2 Logistic Regression result on year 2001 Alcohol data (random sampling)

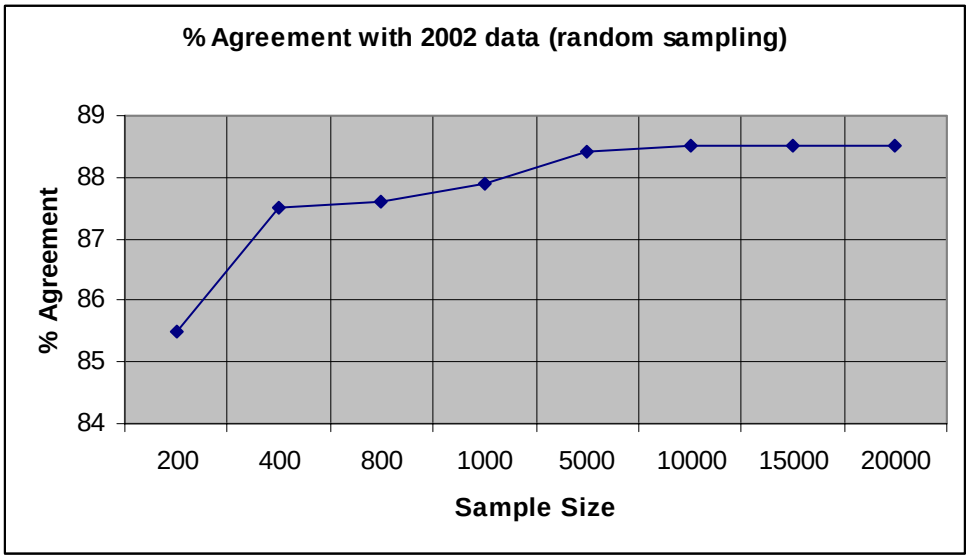

Figure 4.2.3 Logistic Regression result on year 2002 Alcohol data (random sampling)

It is seen that the overall prediction classification accuracy for the training data was almost the same as that for the test data for both the years for all sample sizes. The classification accuracy reached a plateau at the sample size of 5000 for training data and not much could be gained in terms of prediction accuracy by increasing the sample size over 5000. The overall classification accuracy for the training data at the sample size of 5000 was around 88%. When the test results are observed, it is seen that a plateau was reached at the sample size of 5000 as with the training data, where the classification accuracy was around 87.6% for 2001 data and 88.6 % for 2002 data. Increasing the sample size beyond 5000 did not help in predicting the test data more accurately. By running the model for the full datasets for both the years 2001 and 2002, it was observed that the classification accuracy was replicated. The classification accuracy % for the logistic regression model was almost the same as that of the decision tree model when simple random sampling method was used.

When the logistic regression analyses were performed on samples drawn from the year 2001 Alcohol dataset using stratified sampling method, stratifying by the driver's age, DR_AGE variable, as in the case of random sampling, DRINKING was found to be the single most

important variable in classifying the ALC_RES. The next best predictor varied according to the sample sizes and the prediction accuracies also varied according to the sample sizes. Table 4.2.4 shows the summary of the results along with the importance of variables in predicting the dependent variable for the training data while Table 4.2.5 and Table 4.2.6 show the summary of results for different sample sizes when the models built for each sample size was applied to test the validity of prediction for the whole dataset for years 2001 and 2002 respectively. The graphs plotting the overall % agreement against the sample sizes for the training data and test data for years 2001 and 2002 are shown in Figure 4.2.4, Figure 4.2.5 and Figure 4.2.6 respectively.

Table 4.2.4 Regression result on training Alcohol data (stratified sampling)

Sample Size	% Agree			Important Predictors in order of Relative Importance
	0	1	Overall	
200	93.0	80.6	88.5	drinking, hour, rest_use, violchg1, age, body_typ, sex, m_harm, day_week, inj_sev, ve_forms
400	84.1	94.9	91.0	drinking, hour, inj_sev, age, rest_use, violchg1, body_typ, day_week, m_harm, ve_forms, sex
800	92.4	81.9	88.6	drinking, hour, rest_use, age, body_typ, m_harm, violchg1, inj_sev, sex, day_week, ve_forms
1000	76.8	91.1	85.8	drinking, hour, age, inj_sev, violchg1, body_typ, sex, ve_forms, rest_use, m_harm, day_week,
5000	92.9	80.9	88.5	drinking, hour, age, body_typ, rest_use, inj_sev, sex, violchg1, m_harm, ve_forms, day_week,
10000	93.2	79.1	88.0	drinking, hour, age, body_typ, rest_use, inj_sev, sex, ve_forms, violchg1, m_harm, day_week,
15000	93.0	80.0	88.2	drinking, hour, age, rest_use, body_typ, inj_sev, sex, ve_forms, violchg1, m_harm, day_week,
20000	92.8	79.7	88.0	drinking, hour, age, rest_use, body_typ, inj_sev, sex, violchg1, ve_forms, m_harm, day_week,

Table 4.2.5 Logistic Regression result on year 2001 Alcohol data (stratified sampling)

Sample Size	% Agree		
	0	1	Overall
200	88.4	75.8	83.7
400	89.7	81.1	86.5
800	92.1	78.2	86.9
1000	81.8	79.3	87.2
5000	92.5	79.4	87.7
10000	92.8	79.2	87.8
15000	92.7	79.3	87.8
20000	92.7	79.5	87.8

Table 4.2.6 Logistic Regression result on year 2002 Alcohol data (stratified sampling)

Sample Size	% Agree with test data		
	0	1	Overall
200	88.7	77.5	84.5
400	90.0	82.6	87.2
800	92.2	79.1	87.3
1000	92.3	80.7	87.9
5000	93.1	80.9	88.5
10000	93.1	80.6	88.4
15000	93.2	80.8	88.5
20000	93.0	80.8	88.4

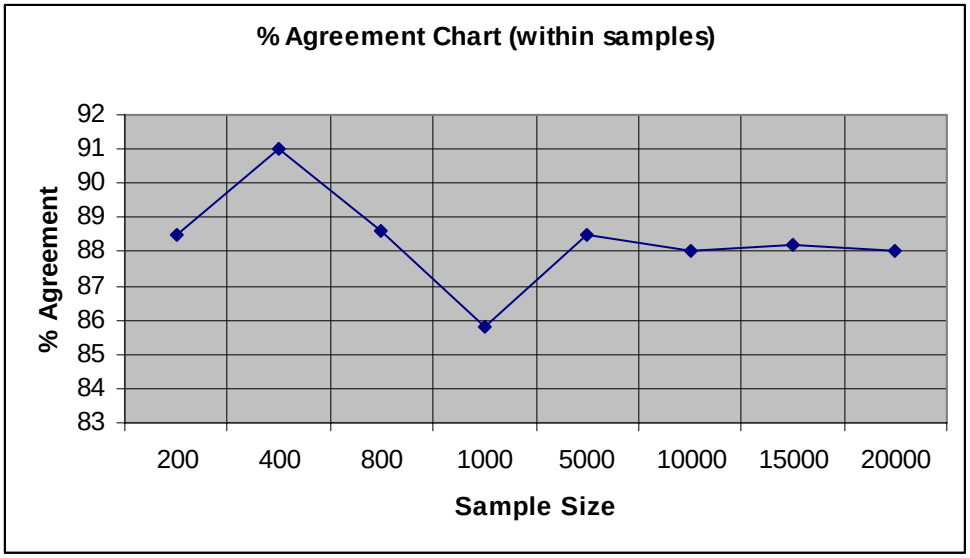

Figure 4.2.4 Logistic Regression result on training Alcohol data (stratified sampling)

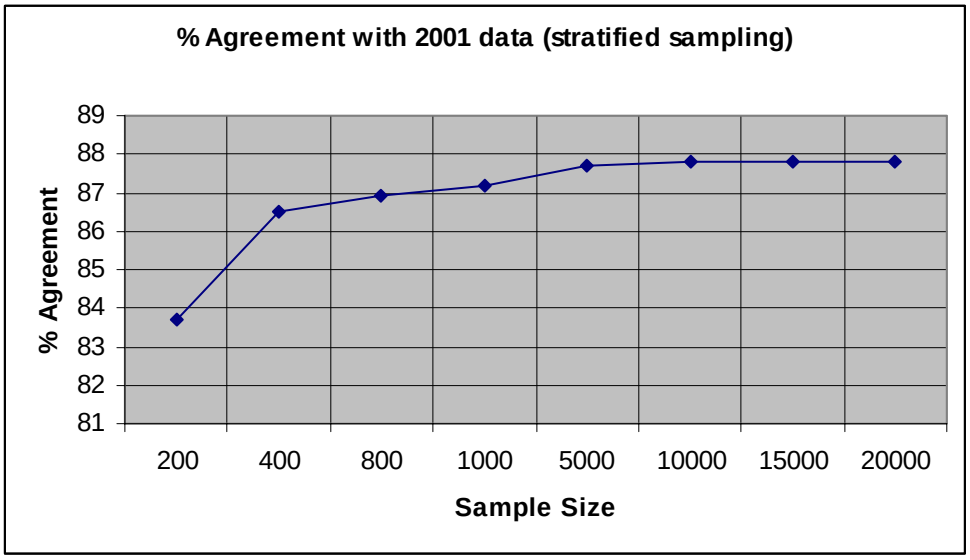

Figure 4.2.5 Logistic Regression result on year 2001 Alcohol data (strat. sampling)

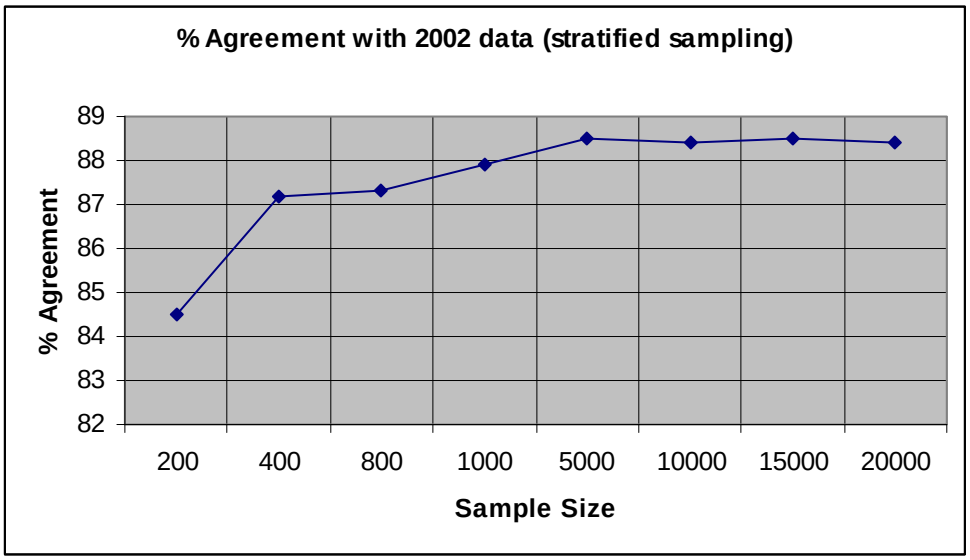

Figure 4.2.6 Logistic Regression result on year 2002 Alcohol data (strat. sampling)

As in the case of random sampling, the training data as well as the test data graphs in case of stratified sampling show that the prediction accuracy do not appreciate after the sample size is increased beyond 5000. For test data, the classification accuracy reaches a maximum value at the sample size of 400 after which it falls and reaches a steady value of around 88% at the sample size of 5000. For 2001 data the prediction accuracy at a sample size of 5000 is around 87.8%

while that for 2002 data, it is 88.5%. So, even if the sampling method is stratified, the prediction accuracy is consistently replicated over different test datasets. Also, the prediction accuracy does not vary by any appreciable amount even if the sampling method is different.

4.3 Alcohol Dataset Analysis with Neural Network

When neural network model was built using the Alcohol dataset using year 2001 crash data for different sample sizes and the sampling method used was simple random sampling, the analyses showed that the prediction accuracy varied according to sample size. Table 4.3.1 shows the summary of the results listing the classification accuracy for different sample sizes for the training data while Table 4.3.2 and Table 4.3.3 show the summary of results for different sample sizes when the models built for each sample size was applied to test the validity of prediction for the whole dataset for years 2001 and 2002 respectively. The graphs plotting the overall % agreement against the sample sizes for training data and test data for years 2001 and 2002 are shown in Figure 4.3.2, Figure 4.3.2 and Figure 4.3.3 respectively.

If the classification agreement % for the "1" and "0" values of ALC_RES is observed it is seen that they are comparable, given the ratio of "0" to "1" is less than 2:1.

Table 4.3.1 Neural Network result on training Alcohol data (random sampling)

	% Agree		
Sample Size	**0**	**1**	**Overall**
200	88.5	85.7	87.5
400	93.1	87.9	91.2
800	92.3	86.4	90.2
1000	88.3	87.4	88.0
5000	91.1	82.9	88.2
10000	91.4	81.4	87.9
15000	91.7	81.3	88.1
20000	91.1	82.5	88.1

Table 4.3.2 Neural Network result on year 2001 Alcohol data (random sampling)

	% Agree		
Sample Size	**0**	**1**	**Overall**
200	90.0	79.4	86.1
400	89.6	82.3	86.9

(table cont.)

800	91.0	80.6	87.1
1000	85.9	86.1	85.9
5000	91.1	82.0	87.7
10000	91.4	81.7	87.8
15000	91.6	81.3	87.8
20000	91.1	82.3	87.8

Table 4.3.3 Neural Network result on year 2002 Alcohol data (random sampling)

Sample Size	% Agree with test data		
	0	1	Overall
200	90.0	79.8	86.2
400	90.1	83.1	87.4
800	91.3	81.7	87.7
1000	86.7	86.7	86.7
5000	91.3	82.9	88.1
10000	91.6	82.6	88.2
15000	91.9	82.3	88.3
20000	91.4	83.0	88.2

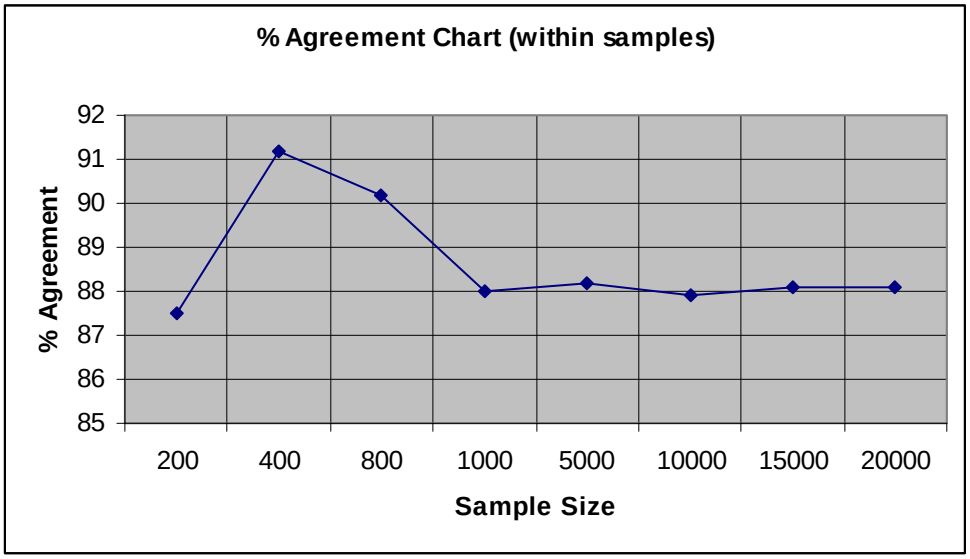

Figure 4.3.1 Neural Network result on training Alcohol data (random sampling)

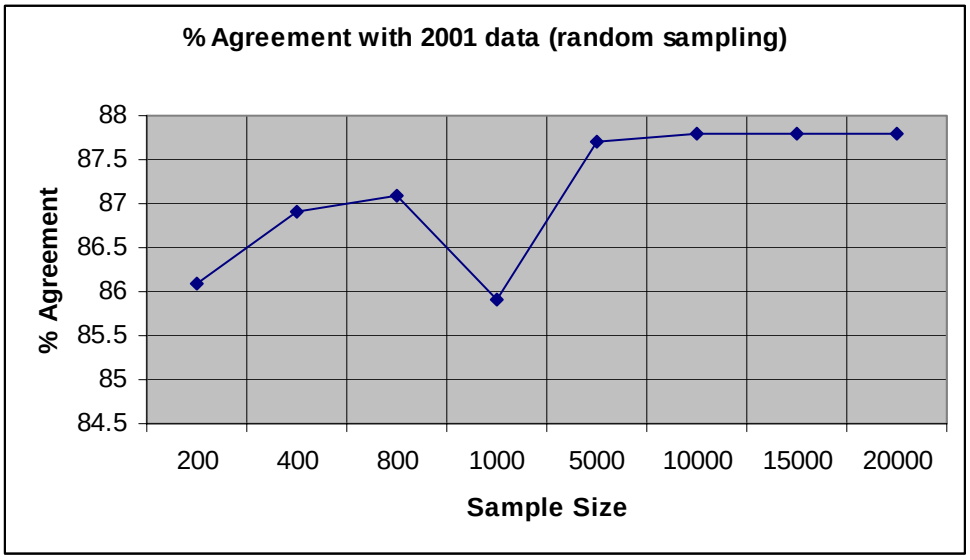

Figure 4.3.2 Neural Network result on year 2001 Alcohol data (random sampling)

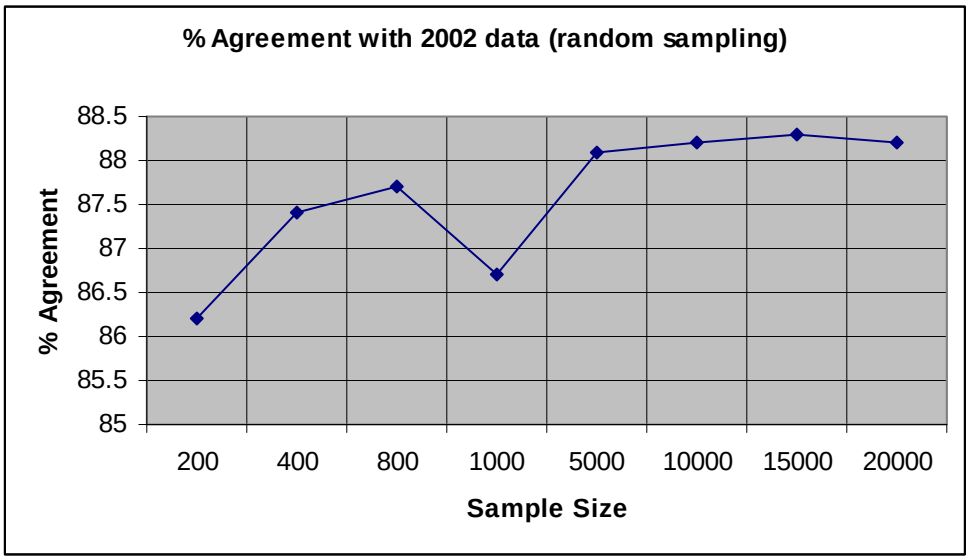

Figure 4.3.3 Neural Network result on year 2002 Alcohol data (random sampling)

It is seen that the overall prediction classification accuracy for the training data was a bit higher than that for the test data for both the years for all sample sizes. The classification accuracy reached a plateau at the sample size of 1000 at about 88% for training data while a maximum accuracy of 91% was obtained at a sample size of 400. Not much could be gained in terms of prediction accuracy by increasing the sample size over 1000. But when the test results

are observed, it is seen that a plateau is reached at the sample size of 5000, where the classification accuracy was around 87.8% for 2001 data and 88.2% for 2002 data and increasing the sample size beyond 5000 did not help in predicting the test data more accurately. By running the model for the full datasets for both the years 2001 and 2002, it was observed that the classification accuracy was replicated. It was also evident that the neural network method was not any more or less efficient in classifying ALC_RES correctly than the decision tree or logistic regression models when the samples were drawn by simple random sampling.

When the neural network models were built by using a stratified sampling method, stratifying by the driver's age variable, the prediction accuracies varied according to the sample sizes. Table 4.3.4 shows the summary of the results listing prediction accuracies for different sample sizes for the training data while Table 4.3.5 and Table 4.3.6 show the summary of results for different sample sizes when the models built for each sample size was applied to test the validity of prediction for the whole datasets for years 2001 and 2002 respectively. The graphs plotting the overall % agreement against the sample sizes for the training data and test data for years 2001 and 2002 are shown in Figure 4.3.4, Figure 4.3.5 and Figure 4.3.6 respectively.

Table 4.3.4 Neural Network result on training Alcohol data (stratified sampling)

	% Agree		
Sample Size	**0**	**1**	**Overall**
200	91.1	93.5	92.0
400	89.3	82.8	86.8
800	92.0	85.3	89.5
1000	90.7	84.3	88.4
5000	91.8	82.5	88.4
10000	92.5	80.6	88.1
15000	90.6	82.5	87.6
20000	91.5	81.5	87.8

Table 4.3.5 Neural Network result on year 2001 Alcohol data (stratified sampling)

	% Agree		
Sample Size	**0**	**1**	**Overall**
200	84.4	86.1	85.0
400	87.8	83.0	86.0
800	91.7	80.3	87.5
1000	90.6	80.1	86.7

(table cont.)

5000	91.1	82.2	87.8
10000	92.7	79.6	87.8
15000	90.7	82.8	87.8
20000	91.5	81.5	87.8

Table 4.3.6 Neural Network result on year 2002 Alcohol data (stratified sampling)

Sample Size	% Agree with test data		
	0	1	Overall
200	84.9	87.2	85.8
400	88.0	83.0	86.1
800	92.2	81.6	88.2
1000	91.0	81.0	87.2
5000	91.4	83.0	88.2
10000	93.0	80.8	88.4
15000	90.9	83.5	88.1
20000	91.6	82.6	88.2

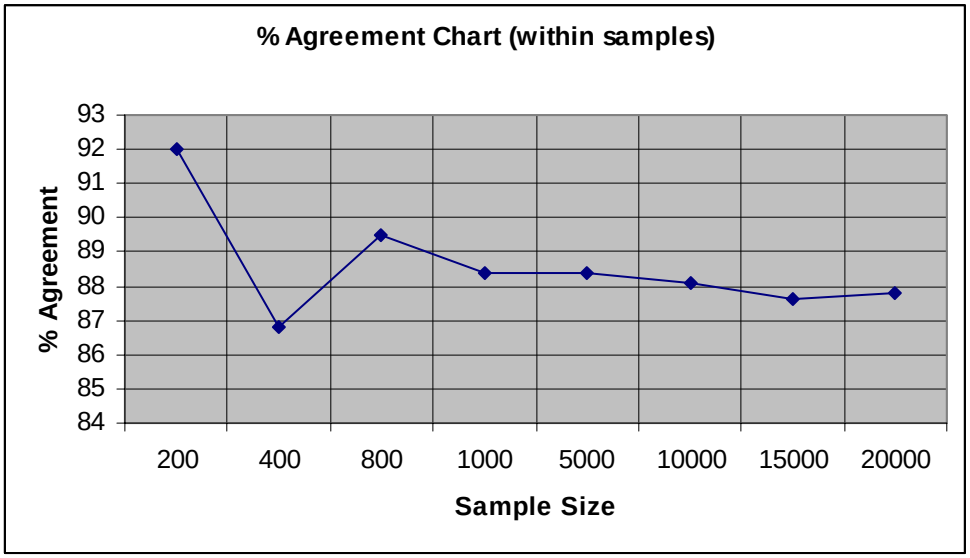

Figure 4.3.4 Neural Network result on training Alcohol data (stratified sampling)

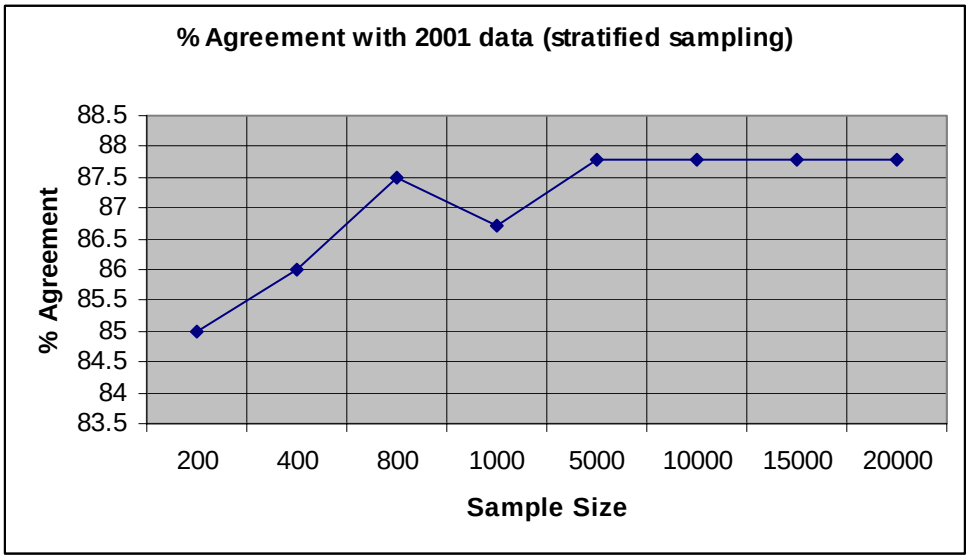

Figure 4.3.5 Neural Network result on year 2001 Alcohol data (stratified sampling)

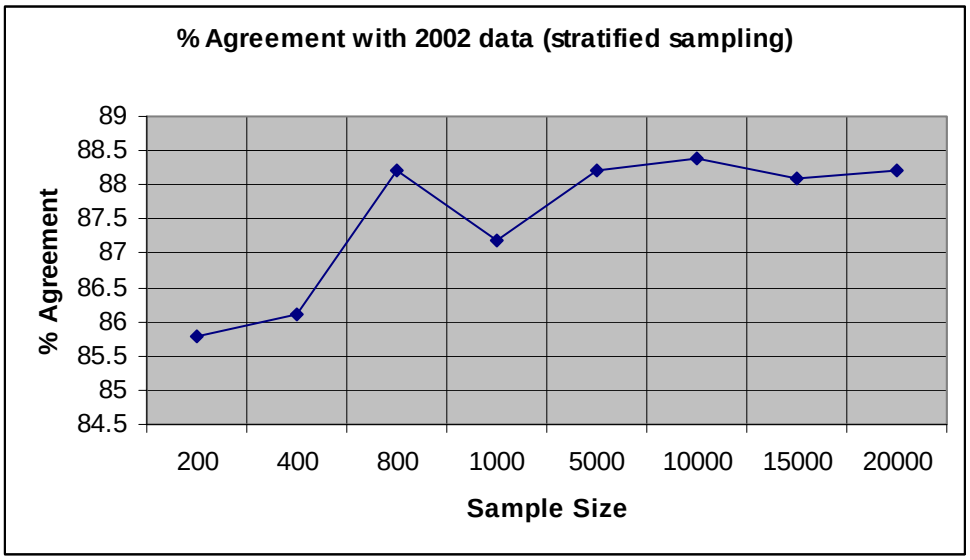

Figure 4.3.6 Neural Network result on year 2002 Alcohol data (stratified sampling)

The results in case of stratified sampling show that the overall prediction classification accuracy for the training data was slightly higher for smaller sample sizes than that for the test data for both the years. For larger sample sizes both the training data and test data had comparable prediction accuracy rates. The classification accuracy reached a plateau at the sample size of 1000 at around 88% for training data. Not much could be gained in terms of prediction

accuracy by increasing the sample size over 1000. But when the test results are observed, it is seen that a plateau is reached at the sample size of 5000, where the classification accuracy was around 87.8% for 2001 data and 88.2% for 2002 data and increasing the sample size beyond 5000 did not help in predicting the test data more accurately. This exactly coincides with the results obtained when neural network model was run and the method of sampling was random sampling. The results show that the model performed consistently for the full datasets for both the years 2001 and 2002 and the classification accuracy was replicated. It was also evident that the neural network method was not any more or less efficient in classifying ALC_RES correctly than the decision tree or logistic regression models when the method of sampling was stratified sampling.

4.4 Seatbelt Dataset Analysis with Decision Tree

When the decision tree model was built using the Seatbelt dataset using year 2001 crash data for different sample sizes and the sampling method used was simple random sampling, the analyses showed that the most important variable used for classifying the dependent variable DR_PROTSYS_CD (driver's protection system) into the correct class was DR_EJEC_CD (code for ejection of driver) for all the sample sizes except sample size 400. The next important variables in terms of predicting DR_PROTSYS_CD differed when the sample sizes were different. The prediction rates also varied according to the sample size. Table 4.4.1 shows the summary of the results along with the importance of variables in predicting the dependent variable for the training data while Table 4.4.2 and Table 4.4.3 show the summary of results for different sample sizes when the models built for each sample size was applied to test the validity of prediction for the whole dataset for years 2001 and 2002 respectively. The graphs plotting the overall % agreement against the sample sizes for the training data and test data for years 2001 and 2002 are shown in Figure 4.4.2, Figure 4.4.2 and Figure 4.4.3 respectively.

If the classification agreement % for the "1" and "0" values of DR_PROTSYS_CD is observed, it is seen that there is a great anomaly in the classification accuracy of "0" and "1" which can be attributed to the fact that the ratio of "0" to "1" is less than 1:7.

Table 4.4.1 Decision Tree result on training Seatbelt data (random sampling)

Sample Size	% Agree			Important Predictors in order of Relative Importance
	0	1	Overall	
200	30.8	99.5	95.0	dr_ejec_cd
400	12.0	99.5	93.9	num_veh, damage_ext1_cd, veh_type_cd
800	31.8	99.3	95.6	dr_ejec_cd, dr_inj_cd, dr_age, veh_type_cd, num_veh, dr_sex
1000	32.4	99.0	94.3	dr_ejec_cd, dr_inj_cd, dr_age, veh_type_cd, num_veh, dr_sex
2000	21.8	99.5	94.8	dr_ejec_cd, veh_type_cd, dr_inj_cd, dr_airbag_cd
5000	24.0	99.8	95.6	dr_ejec_cd, num_veh, est_alcohol, veh_type_cd, dr_age, damage_ext1_cd, dr_airbag_cd
10000	17.2	99.9	95.2	dr_ejec_cd, num_veh, severity_cd, dr_a_d_pres_cd
15000	24.9	99.5	95.3	dr_ejec_cd, dr_inj_cd, est_alcohol, veh_type_cd, dr_airbag_cd, dr_age
20000	20.8	99.8	95.4	dr_ejec_cd, est_alcohol, dr_inj_cd, num_veh, veh_type_cd, dr_airbag_cd, dr_age, damage_ext1_cd, dr_race

Table 4.4.2 Decision Tree result on year 2001 Seatbelt data (random sampling)

Sample Size	% Agree		
	0	1	Overall
200	17.4	99.8	95.1
400	5.4	99.0	93.7
800	20.8	99.0	94.6
1000	24.3	98.6	94.4
2000	21.2	99.6	95.1
5000	19.1	99.7	95.2
10000	17.2	99.9	95.2
15000	24.8	99.4	95.2
20000	20.7	99.7	95.3

Table 4.4.3 Decision Tree result on year 2002 Seatbelt data (random sampling)

Sample Size	% Agree with test data		
	0	1	Overall
200	12.5	99.7	92.6
400	4.5	98.8	91.2
800	15.7	98.6	91.9
1000	18.4	98.2	91.8
2000	15.1	99.3	92.4

(table cont.)

5000	13.0	99.6	92.6
10000	12.1	99.9	92.8
15000	18.0	99.3	92.7
20000	13.6	99.6	92.6

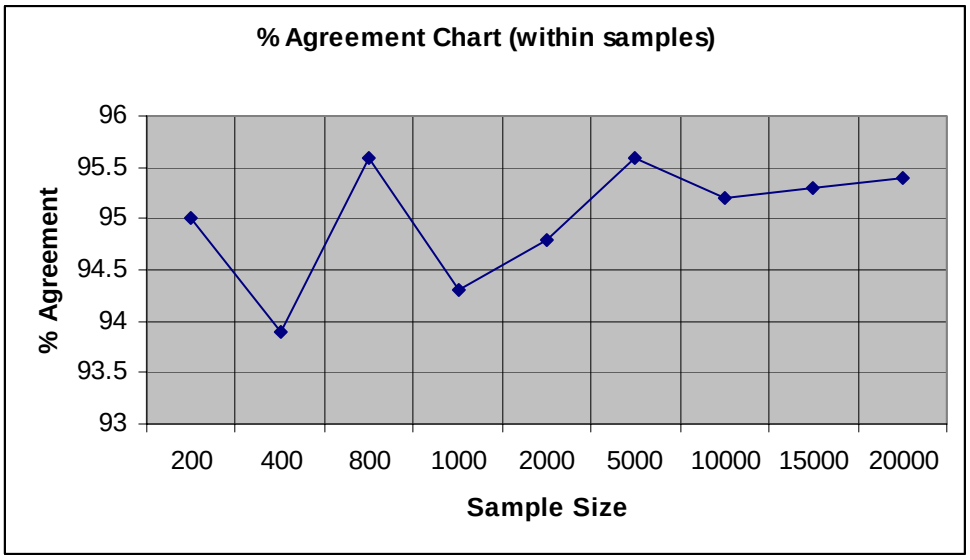

Figure 4.4.1 Decision Tree result on training Seatbelt data (random sampling)

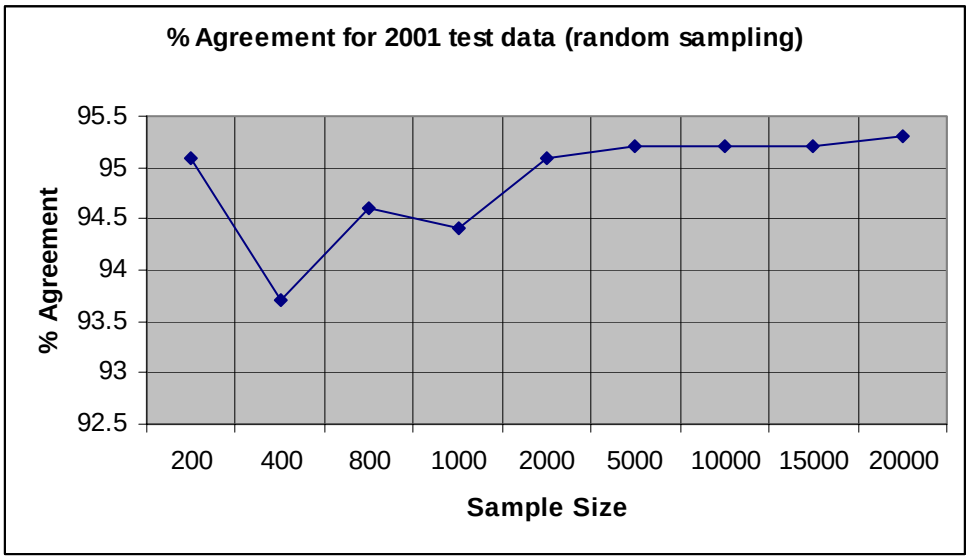

Figure 4.4.2 Decision Tree result on year 2001 Seatbelt data (random sampling)

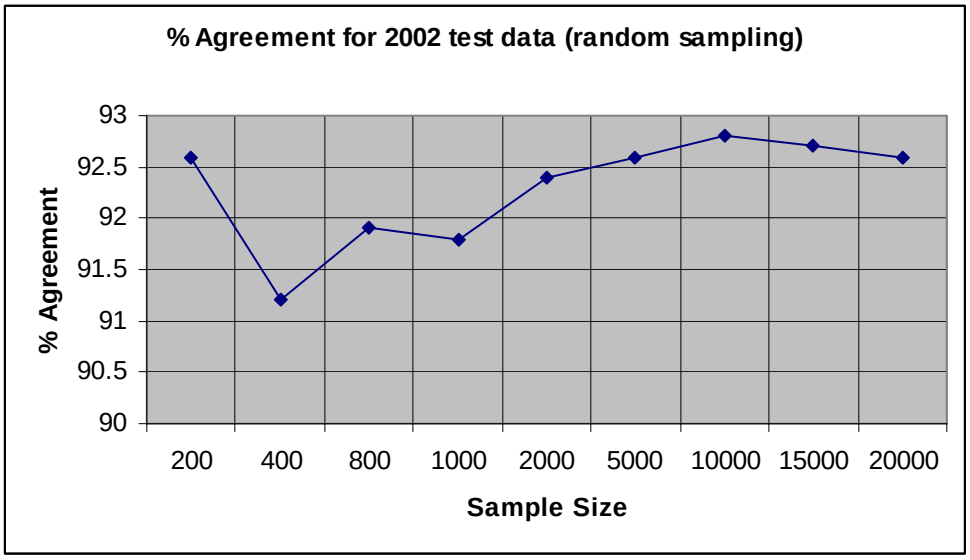

Figure 4.4.3 Decision Tree result on year 2002 Seatbelt data (random sampling)

It is seen that the overall prediction classification accuracy for the training data was higher than that for the test data for the year 2002 data but was almost the same for year 2001 data. The models were built using samples from year 2001 data. The classification accuracy did not improve appreciably when sample size was increased beyond 5000 for the training data. The overall classification accuracy for the training data at the sample size of 5000 was around 96%. When the test results are observed, it is seen that a plateau was reached at the sample size of 5000, where the classification accuracy was around 95.2% for 2001 data and 92.6% for 2002 data and increasing the sample size beyond 5000 did not help in predicting the test data more accurately. By running the model for the full datasets for both the years 2001 and 2002 for the seatbelt data, the classification accuracy was seen to be not exactly replicated. While the validation of the models on the test data which was the same as the population from which the samples were drawn (year 2001) performed well and equivalent to that of the training data, the validation of the models on a completely new dataset (year 2002 data) did not perform as well.

When the decision tree models were built by using a stratified sampling method, stratifying by the driver's age variable, similar to the case of random sampling, DR_EJEC_CD was found to

be the most important variable in classifying the dependent variable for all sample sizes except for sample size of 200. The next best predictor varied according to the sample sizes and the prediction accuracies also varied according to the sample sizes. Table 4.4.4 shows the summary of the results along with the importance of variables in predicting the dependent variable for the training data while Table 4.1.5 and Table 4.4.6 show the summary of results for different sample sizes when the models built for each sample size was applied to test the validity of prediction for the whole dataset for years 2001 and 2002 respectively. The graphs plotting the overall % agreement against the sample sizes for the training data and test data for years 2001 and 2002 are shown in Figure 4.4.4, Figure 4.4.5 and Figure 4.4.6 respectively.

Table 4.4.4 Decision Tree result on training Seatbelt data (stratified sampling)

ample Size	% Agree			Important Predictors in order of Relative Importance
	0	1	Overall	
200	33.3	98.4	95.5	dr_age, dr_airbag_cd, veh_type_cd, num_veh, severity_cd
400	26.9	100.0	95.2	dr_ejec_cd
800	17.0	100.0	95.1	dr_ejec_cd
1000	26.4	99.9	96.0	dr_ejec_cd
2000	32.2	99.5	95.6	dr_ejec_cd, est_alcohol, damage_ext1_cd, dr_sex, severity_cd, dr_race, veh_type_cd
5000	23.6	99.7	95.6	dr_ejec_cd, est_alcohol, dr_inj_cd, dr_airbag_cd, veh_typ_cd
10000	22.1	99.8	95.4	dr_ejec_cd, num_veh, dr_airbag_cd, dr_age, est_alcohol, dr_inj_cd, severity_cd, veh_type_cd
15000	22.9	99.7	95.4	dr_ejec_cd, est_alcohol, dr_airbag_cd, dr_inj_cd, num_veh, veh_type_cd, severity_cd, damage_ext1_cd, dr_sex, dr_race, dr_age
20000	17.9	99.9	95.3	dr_ejec_cd, est_alcohol, veh_type_cd, dr_inj_cd, dr_airbag_cd, dr_age

Table 4.4.5 Decision Tree result on year 2001 Seatbelt data (stratified sampling)

Sample Size	% Agree		
	0	1	Overall
200	2.5	97.7	92.3
400	16.3	99.9	95.2
800	16.3	99.9	95.2
1000	17.4	99.8	95.1
2000	19.5	99.8	95.6
5000	20.5	99.7	95.2

(table cont.)

10000	19.7	99.7	95.2
15000	19.9	99.8	95.6
20000	17.9	99.9	95.3

Table 4.4.6 Decision Tree result on year 2002 Seatbelt data (stratified sampling)

Sample Size	% Agree with test data		
	0	1	Overall
200	2.0	97.6	89.9
400	11.5	99.9	92.7
800	11.5	99.9	92.7
1000	12.5	99.7	92.6
2000	17.2	98.6	92.0
5000	14.2	99.5	92.6
10000	13.0	99.6	92.6
15000	14.4	99.5	92.6
20000	12.1	99.9	92.8

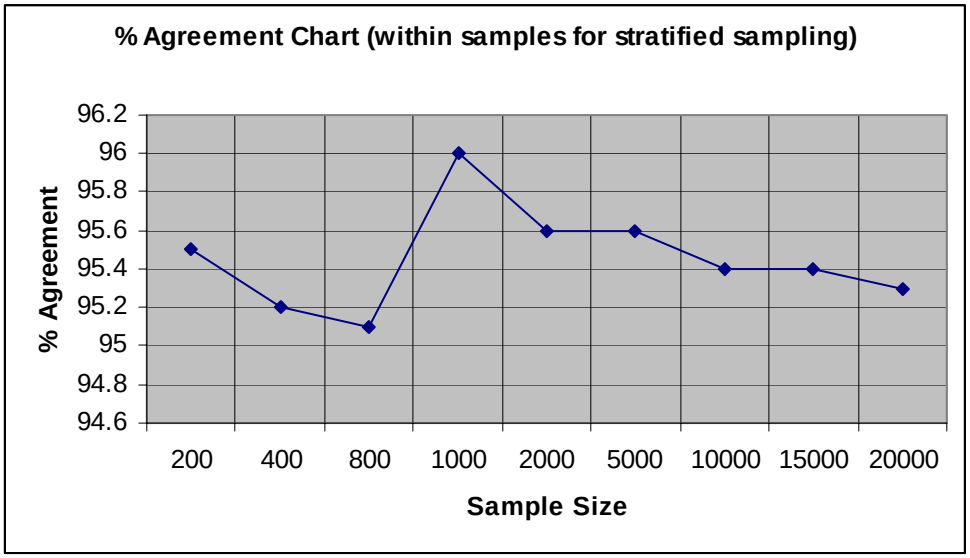

Figure 4.4.4 Decision Tree result on training Seatbelt data (stratified sampling)

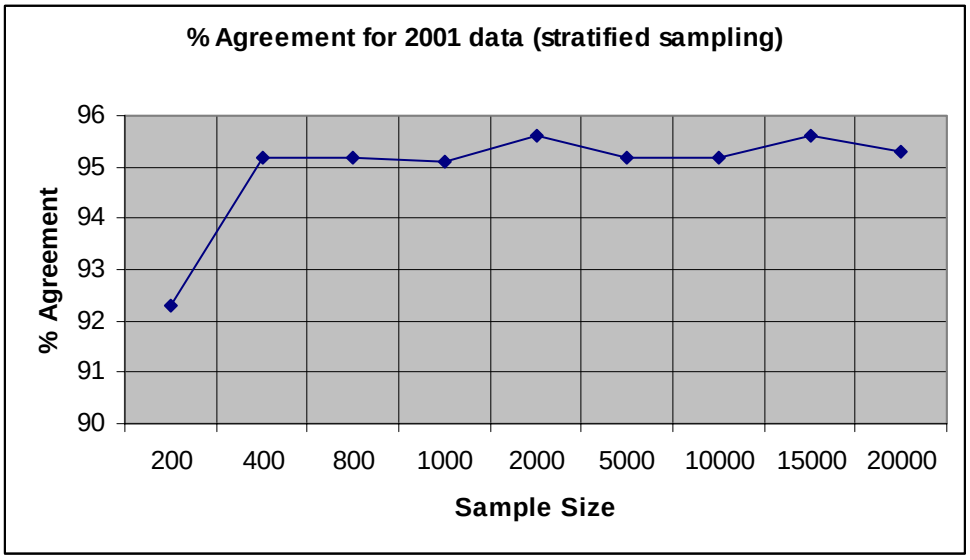

Figure 4.4.5 Tree result on year 2001 Seatbelt data (strat. sampling)

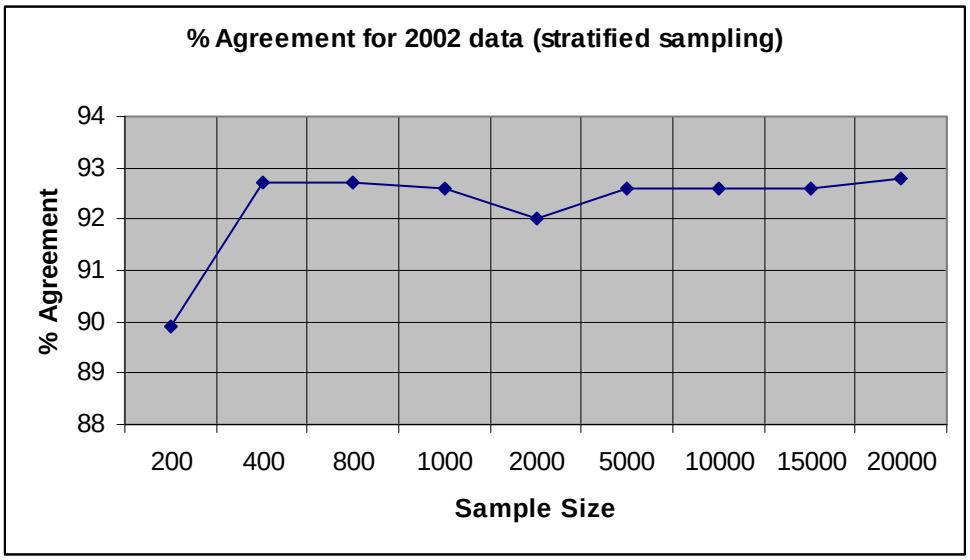

Figure 4.4.6 Decision Tree result on year 2002 Seatbelt data (strat. sampling)

The results in case of stratified sampling show that for training data the prediction accuracy did not improve if the sample size was increased beyond 2000. Actually the classification accuracy for the training data varied within a very tight range starting from 95.1% to 96% for different sample sizes. For test data, the classification accuracy reached a maximum value at the sample size of 400 for both the datasets and the prediction accuracy did not improve by any

means by increasing the sample size beyond 400. This is indeed an interesting observation. But the prediction accuracy in case of 2001 data was higher at above 95% than that of 2002 data which was below 93%, both at the sample size of 400. This may be attributed to the fact that the samples were drawn from 2001 data and 2002 data was an entirely new set of data. So, it is seen that, if the sampling method is stratified, for decision tree models for the Seatbelt datasets, a much lower sample size is required and the prediction accuracy is not consistently replicated over different test datasets. This is in contrast with what was observed for the Alcohol dataset and the characteristic of data may be responsible for that.

Since we see that the prediction rate for the 0 values of the dependent variable is very low for decision trees (for both the sampling methods), and this could be attributed to the very high ratio of 1:0 in the population, another set of analysis is done with the dataset from the 2001 population where only the records with value of injury codes (SEVERITY_CD) "A" and "B" are chosen. The classification of the variable SEVERITY_CD in the original dataset is done as "A" = 1, "B" = 0 and 2 = others and all records with a value of SEVERITY_CD = 2 are removed. This leaves out a much reduced dataset with only 395 records which has a much less skewed distribution of "0"s and "1"s for the dependent variable. The ratio of 0:1 for DR_PROTSYS_CD in this reduced dataset is around 1:3. When the decision tree model is built with the whole reduced 2001 dataset and the model is tested on 2002 data (which is also reduced as only records with SEVERITY_CD = "A" or "B" are retained), the results obtained are shown in the table 4.4.7.

Table 4.4.7 Decision Tree results on modified Seatbelt training and test data

	% Agree with test data		
Data	**0**	**1**	**Overall**
Year 2001 training data	63.7	96.7	89.1
Year 2002 test data	50.0	97.3	86.0

Thus, it is seen that though the overall prediction rate is somewhat lower than that obtained for the original Seatbelt data, the prediction accuracy of "0" values improve a great deal over the original data. This may be attributed to the more even distribution of 0 and 1 values of the

dependent variable. This might be an indicator that decision tree model is not very accurate in predicting correctly the value of the dependent variable which has a very low occurrence in the population.

4.5 Seatbelt Dataset Analysis with Logistic Regression

As in the case of decision trees, when the logistic regression analysis was performed using the Seatbelt dataset for year 2001 crash data with different sample sizes and the sampling method used was simple random sampling, the analyses showed that there was no consistency among the variables that were identified as the most important variable in classifying the dependent variable DR_PROTSYS_CD into the correct class for different sample sizes. Table 4.5.1 shows the summary of the results along with the importance of variables in predicting the dependent variable for the training data while Table 4.5.2 and Table 4.5.3 show the summary of results for different sample sizes when the models built for each sample size was applied to test the validity of prediction for the whole dataset for years 2001 and 2002 respectively. The graphs plotting the overall % agreement against the sample sizes for the training data and test data for years 2001 and 2002 are shown in Figure 4.5.2, Figure 4.5.2 and Figure 4.5.3 respectively.

If the classification agreement % for the "1" and "0" values of DR_PROTSYS_CD is observed, there is a huge difference in the prediction accuracy of "1"s and that of "0"s and as the sample size increases, the difference between prediction accuracy of "1"s and that of "0"s increase. This can be attributed to the fact that the ratio of "0" to "1" in the population is less than 1:7.

Table 4.5.1 Logistic Regression result on training Seatbelt data (random sampling)

Sample Size	% Agree			Important Predictors in order of Relative Importance
	0	1	Overall	
200	100.0	100.0	100.0	dr_age, dr_sex, num_veh, damage_ext1_cd, dr_ejec_cd, dr_inj_cd, dr_airbag_cd, severity_cd, veh_type_cd, est_alcohol, dr_race, dr_a_d_pres_cd
400	34.8	99.4	95.0	num_veh, veh_type_cd, dr_sex, damage_ext1_cd, severity_cd, dr_inj_cd, dr_age, dr_a_d_pres_cd, dr_airbag_cd, dr_race, est_alcohol, dr_ejec_cd

(table cont.)

800	24.0	100.0	97.3	num_veh, veh_type_cd, dr_sex, dr_age, dr_airbag_cd, damage_ext1_cd, dr_a_d_pres_cd, dr_ejec_cd, est_alcohol, dr_inj_cd, severity_cd, dr_race
1000	25.5	99.6	95.2	veh_type_cd, dr_ejec_cd, dr_inj_cd , severity_cd, dr_a_d_pres_cd, dr_sex, damage_ext1_cd, dr_airbag_cd, dr_age, dr_race est_alcohol, num_veh,
2000	25.8	99.8	95.9	dr_ejec_cd, dr_inj_cd, veh_type_cd, dr_age, severity_cd, num_veh, damage_ext1_cd, dr_airbag_cd, est_alcohol, dr_sex, dr_a_d_pres_cd, dr_race,
5000	18.2	99.9	94.8	veh_type_cd, dr_ejec_cd, dr_inj_cd, num_veh, damage_ext1_cd, dr_age, dr_sex, est_alcohol, dr_airbag_cd, dr_race, severity_cd, dr_a_d_pres_cd
10000	18.4	99.8	95.4	dr_ejec_cd, dr_inj_cd, veh_type_cd, dr_age, severity_cd, dr_sex, num_veh, dr_airbag_cd, est_alcohol, dr_race, damage_ext1_cd, dr_a_d_pres_cd
15000	16.1	99.9	95.1	dr_ejec_cd, veh_type_cd, dr_inj_cd, dr_age, est_alcohol, dr_airbag_cd, num_veh, dr_sex, severity_cd, dr_race, damage_ext1_cd, dr_a_d_pres_cd
20000	16.7	99.9	95.1	dr_ejec_cd, veh_type_cd, dr_inj_cd, num_veh, dr_age, est_alcohol, dr_airbag_cd, dr_sex, severity_cd, dr_race, damage_ext1_cd, dr_a_d_pres_cd

Table 4.5.2 Logistic Regression result on year 2001 Seatbelt data (random sampling)

	% Agree		
Sample Size	**0**	**1**	**Overall**
200	21.8	95.4	91.1
400	23.8	97.4	93.2
800	15.8	99.6	94.8
1000	17.0	99.6	94.8
2000	17.6	99.8	95.1
5000	16.7	99.9	95.1
10000	16.6	99.9	95.1
15000	16.3	99.9	95.1
20000	16.6	99.9	95.1

Table 4.5.3 Logistic Regression result on year 2002 Seatbelt data (random sampling)

	% Agree with test data		
Sample Size	**0**	**1**	**Overall**
200	17.6	94.6	88.3
400	19.3	96.8	90.5
800	11.6	99.4	92.2
1000	12.6	99.5	92.4
2000	12.9	99.8	92.7
5000	12.1	99.8	92.6
10000	12.4	99.9	92.7

(table cont.)

15000	12.2	99.9	92.7
20000	12.1	99.9	92.7

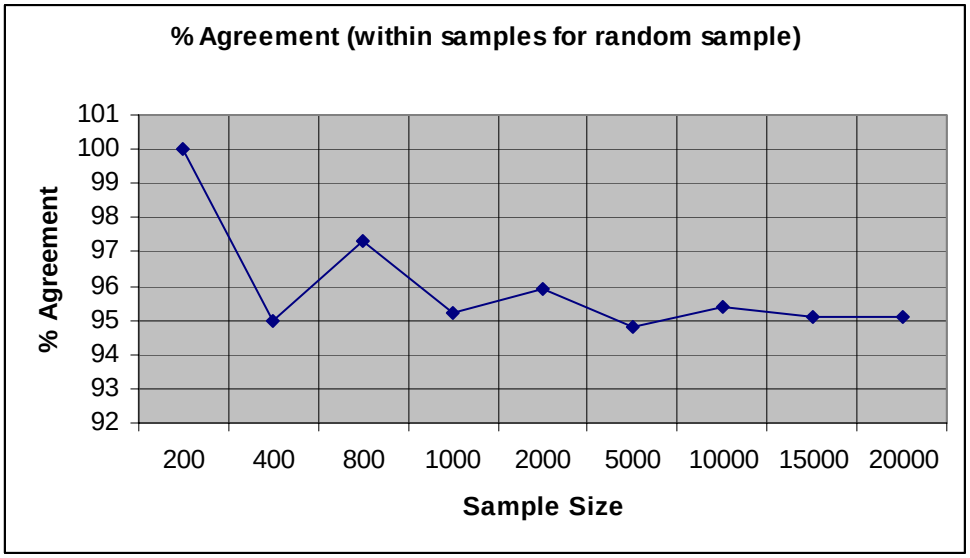

Figure 4.5.1 Logistic Regression result on training Seatbelt data (random sampling)

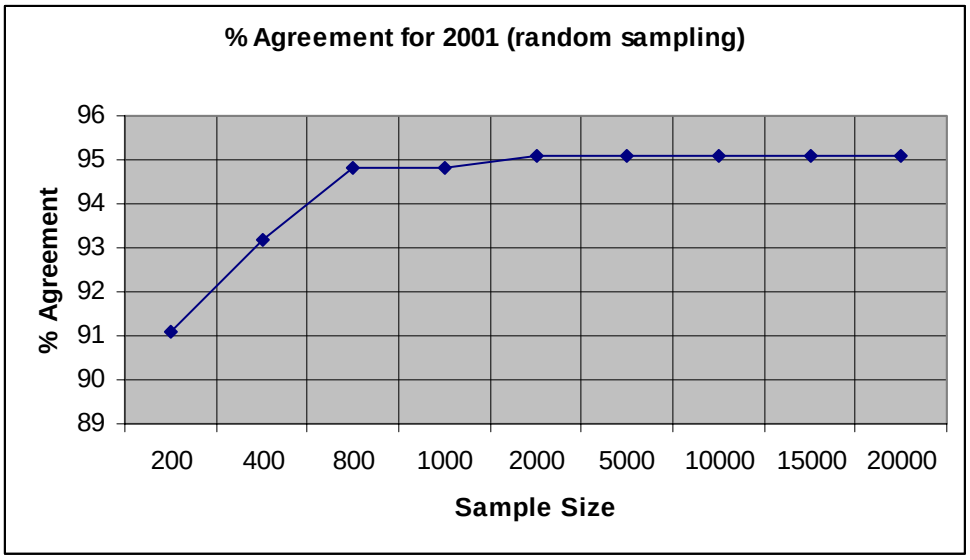

Figure 4.5.2 Logistic Regression result on year 2001 Seatbelt data (random sampling)

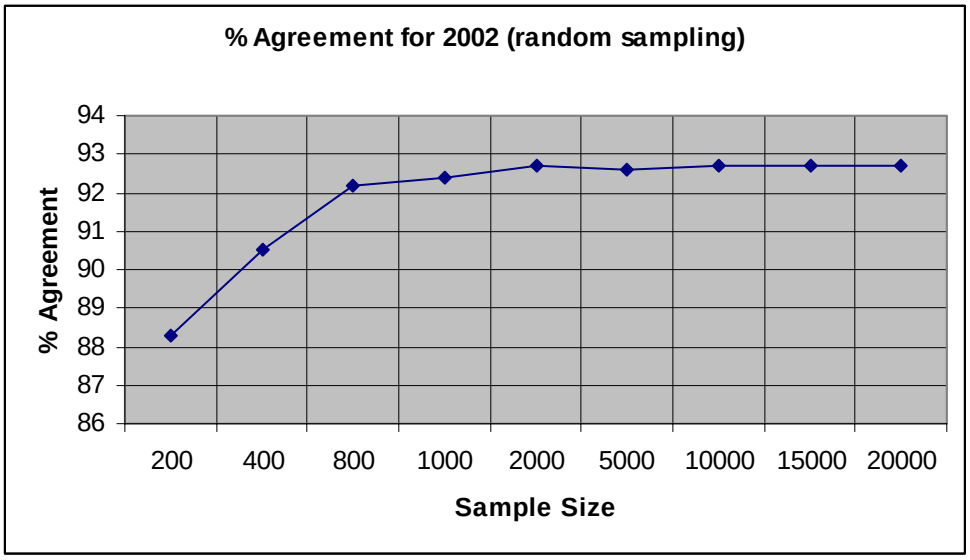

Figure 4.5.3 Logistic Regression result on year 2002 Seatbelt data (random sampling)

It is seen that the overall prediction classification accuracy for the training data was almost the same as that for year 2001 test data for higher sample sizes while it was higher than that of year 2002 test data for all sample sizes. The classification accuracy attained sort of a stable value at the sample size of 1000 for training data and not much could be gained in terms of prediction accuracy by increasing the sample size over 1000. The overall classification accuracy for the training data at the sample size of 5000 was around 95%. When the test results are observed, it is seen that a plateau was reached at the sample size of 800 for both sets of test data, where the classification accuracy was around 95% for 2001 data and a little less than 93 % for 2002 data. This may be attributed to the fact that the samples were drawn from 2001 data and 2002 data was an entirely new set of data. So, it is seen that, if the sampling method is random, for logistic regression models for the Seatbelt datasets, a much lower sample size is required and the prediction accuracy is not consistently replicated over different test datasets. This is in contrast with what was observed for the logistic regression model for Alcohol dataset and the characteristic of data may be responsible for that.

When the logistic regression analyses were performed on samples drawn from the year 2001 Seatbelt dataset using stratified sampling method, stratifying by the driver's age, DR_AGE variable, the importance of the predictor variables in classifying the dependent variable DR_PROTSYS_CD and the classification agreements for different sample sizes, as in the case of random sampling, no single variable was found to be the most important variable in classifying DR_PROTSYS_CD for all sample sizes. The prediction accuracies also varied with models for different sample sizes. Table 4.5.4 shows the summary of the results along with the importance of variables in predicting the dependent variable for the training data while Table 4.5.5 and Table 4.5.6 show the summary of results for different sample sizes when the models built for each sample size was applied to test the validity of prediction for the whole dataset for years 2001 and 2002 respectively. The graphs plotting the overall % agreement against the sample sizes for the training data and test data for years 2001 and 2002 are shown in Figure 4.5.4, Figure 4.5.5 and Figure 4.5.6 respectively.

Table 4.5.4 Logistic Regression result on training Seatbelt data (stratified sampling)

Sample Size	% Agree			Important Predictors in order of Relative Importance
	0	1	Overall	
200	36.4	100.0	95.9	veh_type_cd, dr_race, severity_cd, dr_age, dr_a_d_pres_cd, dr_airbag_cd, num_veh, damage_ext1_cd, dr_ejec_cd, dr_sex, dr_inj_cd , est_alcohol
400	15.8	100.0	95.4	dr_age, severity_cd, dr_airbag_cd, damage_ext1_cd, dr_inj_cd, veh_type_cd, dr_race, dr_a_d_pres_cd, dr_sex, est_alcohol, dr_ejec_cd, num_veh,
800	14.3	100.0	94.0	dr_age, num_veh, veh_type_cd, severity_cd, dr_airbag_cd, damage_ext1_cd, dr_sex, dr_a_d_pres_cd, est_alcohol, dr_race, dr_inj_cd, dr_ejec_cd
1000	21.6	100.0	95.4	dr_age, est_alcohol, num_veh, dr_airbag_cd, dr_a_d_pres_cd, dr_inj_cd , damage_ext1_cd, dr_sex, dr_race, veh_type_cd, dr_ejec_cd, severity_cd
2000	18.2	99.9	95.2	dr_inj_cd, veh_type, num_veh, dr_airbag_cd, dr_age, damage_ext1_cd, severity_cd, dr_sex, dr_a_d_pres_cd, dr_ejec_cd, cd, est_alcohol, dr_race
5000	13.6	100.0	95.1	dr_ejec_cd, dr_airbag_cd, dr_inj_cd, dr_age, veh_type_cd, num_veh, dr_sex, est_alcohol, severity_cd, damage_ext1_cd, dr_a_d_pres_cd dr_race

(table cont.)

10000	16.7	99.8	95.3	dr_ejec_cd, veh_type_cd, dr_inj_cd, , num_veh, est_alcohol, dr_dr_age, dr_race, severity_cd, dr_airbag_cd, dr_sex damage_ext1_cd, dr_a_d_pres_cd
15000	16.0	99.9	95.0	dr_ejec_cd, veh_type_cd, dr_inj_cd, dr_age, num_veh, est_alcohol, dr_airbag_cd, severity_cd, dr_race, dr_sex, damage_ext1_cd, dr_a_d_pres_cd
20000	16.5	99.9	95.1	dr_ejec_cd, veh_type_cd, dr_inj_cd, dr_age, num_veh, est_alcohol, dr_airbag_cd, dr_sex, severity_cd, dr_race, damage_ext1_cd, dr_a_d_pres_cd

Table 4.5.5 Logistic Regression result on year 2002 Seatbelt data (stratified sampling)

	% Agree		
Sample Size	**0**	**1**	**Overall**
200	36.4	100.0	95.9
400	18.9	98.6	94.0
800	16.7	99.6	94.8
1000	16.9	99.8	95.0
2000	18.2	99.9	95.2
5000	16.4	99.9	95.1
10000	15.9	99.9	95.1
15000	16.5	99.9	95.1
20000	16.6	99.9	95.1

Table 4.5.6 Logistic Regression result on year 2001 Seatbelt data (stratified sampling)

	% Agree with test data		
Sample Size	**0**	**1**	**Overall**
200	15.3	97.4	90.7
400	14.7	98.3	91.4
800	12.3	99.4	92.3
1000	12.1	99.8	92.7
2000	11.0	99.9	92.6
5000	12.1	99.9	92.7
10000	11.5	99.9	92.7
15000	12.3	99.9	92.7
20000	12.1	99.9	92.7

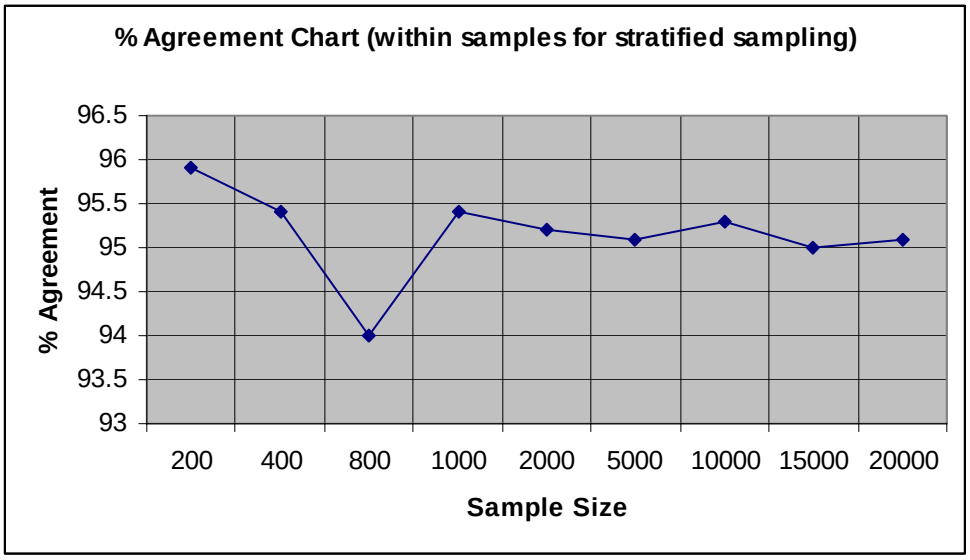

Figure 4.5.4 Logistic Regression result on training Seatbelt data (stratified sampling)

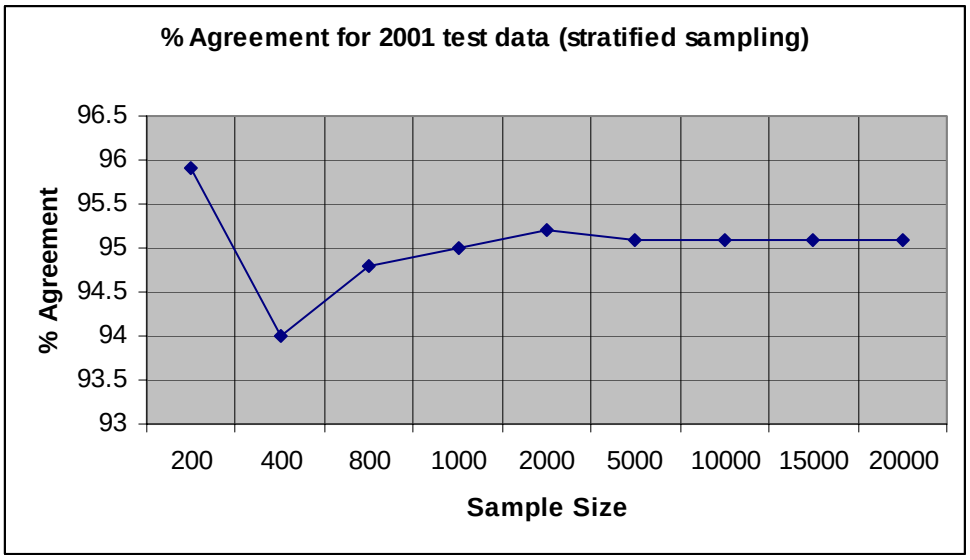

Figure 4.5.5 Logistic Regression result on year 2001 Seatbelt data (strat. sampling)

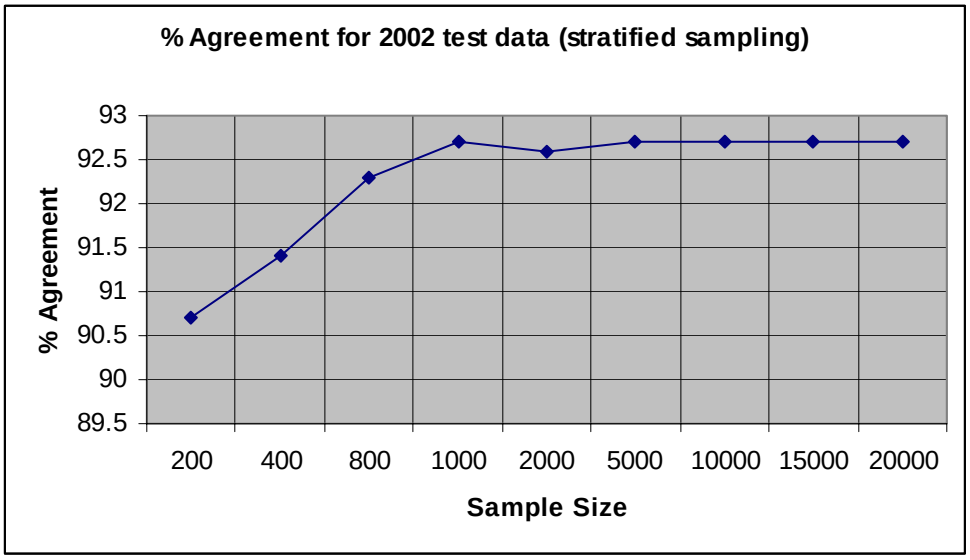

Figure 4.5.6 Logistic Regression result on year 2002 Seatbelt data (strat. sampling)

In this case of logistic regression model with Seatbelt dataset, when the sampling method is stratified, the training data as well as the test data graphs show that the prediction accuracy does not appreciate after the sample size is increased beyond 1000. For training data the classification accuracy does not vary greatly with sample sizes and stays within a range of 94% to 96% and reaches a stable value of a little above 95% at sample size of 1000. For year 2001 test data, the classification accuracy is also a little above 95% at a sample size of 1000 and stays at the same value for greater sample sizes. For 2002 test data the classification accuracy is a little below 93% at a sample size of 1000 and the accuracy does not improve with bigger samples. This is in line with the results obtained with random sampling method and the prediction accuracy does not vary if the sampling method is different for both sets of test data.

Since we see that the prediction rate for the 0 values of the dependent variable is very low with logistic regression method also (for both the sampling methods), and this could be attributed to the very high ratio of 1:0 in the population, as in the case of decision tree, another set of analysis is done with the dataset from the 2001 population where only the records with value of injury codes (SEVERITY_CD) “A” and “B” are chosen. The classification of the variable

SEVERITY_CD in the original dataset is done as “A” = 1, “B” = 0 and 2 = others and all records with a value of SEVERITY_CD = 2 are removed. This leaves a much reduced dataset with only 395 records which has a much less skewed distribution of “0”s and “1”s for the dependent variable. The ratio of 0:1 for DR_PROTSYS_CD in this reduced dataset is around 1:3. When the logistic regression model is built with the whole reduced 2001 dataset and the model is tested on 2002 data (which is also reduced as only records with SEVERITY_CD = “A” or “B” are retained), the results obtained are shown in the table 4.5.7.

Table 4.5.7 Logistic Regression results on modified Seatbelt training and test data

Data	% Agree with test data		
	0	1	Overall
Year 2001 training data	63.9	96.2	88.4
Year 2002 test data	50.0	97.7	86.6

Thus, it is seen that though the overall prediction rate is somewhat lower than that obtained for the original Seatbelt data, the prediction accuracy of “0” values improve a great deal over that with the original data. This may be attributed to the more even distribution of 0 and 1 values of the dependent variable. This might be an indicator that logistic regression model is not very accurate in predicting correctly the value of the dependent variable which has a very low occurrence in the population.

4.6 Seatbelt Dataset Analysis with Neural Network

When neural network model was built using the Seatbelt dataset using year 2001 data for different sample sizes, the sampling method being random sampling, the analyses showed that the prediction accuracy varied according to sample size. Table 4.6.1 shows the summary of the results listing the classification accuracy for different sample sizes for the training data while Table 4.6.2 and Table 4.6.3 show the summary of results for different sample sizes when the models built for each sample size was applied to test the validity of prediction for the whole dataset for years 2001 and 2002 respectively. The graphs plotting the overall % agreement against

the sample sizes for training data and test data for years 2001 and 2002 are shown in Figure 4.6.2, Figure 4.6.2 and Figure 4.6.3 respectively.

If the classification agreement % for the "1" and "0" values of DR_PROTSYS_CD in the training data as well as the test data is observed, it is seen that except for the sample size of 20,000 which is almost equal to the population size, the prediction accuracy of "1"s are 100% for all other sample sizes and that of "0"s are 0%. This can be attributed to the fact that the ratio of "0" to "1" in the population is less than 1:7. Hence when the neural network is taught to read patterns from the training data, most of the time it is trained to predict a "1" irrespective of the predictor values, so it predicts a "1" every time for DR_PROTSYS_CD and fails to predict the "0"s. So, it is successful in predicting "1"s correctly 100% of the time and "0" correctly 0% of the time.

Table 4.6.1 Neural Network result on training Seatbelt data (random sampling)

	% Agree		
Sample Size	**0**	**1**	**Overall**
200	0.0	100.0	91.9
400	0.0	100.0	93.1
800	0.0	100.0	94.7
1000	0.0	100.0	94.2
2000	0.0	100.0	93.7
5000	0.0	100.0	93.8
10000	0.0	100.0	94.1
15000	0.0	100.0	94.2
20000	18.9	99.7	95.1

Table 4.6.2 Neural Network result on year 2001 Seatbelt data (random sampling)

	% Agree		
Sample Size	**0**	**1**	**Overall**
200	0.0	100.0	94.2
400	0.0	100.0	94.2
800	0.0	100.0	94.2
1000	0.0	100.0	94.2
2000	0.0	100.0	94.2
5000	0.0	100.0	94.2
10000	0.0	100.0	94.2
15000	0.0	100.0	94.2
20000	14.7	99.6	92.7

Table 4.6.3 Neural Network result on year 2002 Seatbelt data (random sampling)

Sample Size	% Agree with test data		
	0	1	Overall
200	0.0	100.0	91.8
400	0.0	100.0	91.8
800	0.0	100.0	91.8
1000	0.0	100.0	91.8
2000	0.0	100.0	91.8
5000	0.0	100.0	91.8
10000	0.0	100.0	91.8
15000	0.0	100.0	91.8
20000	14.7	99.6	92.7

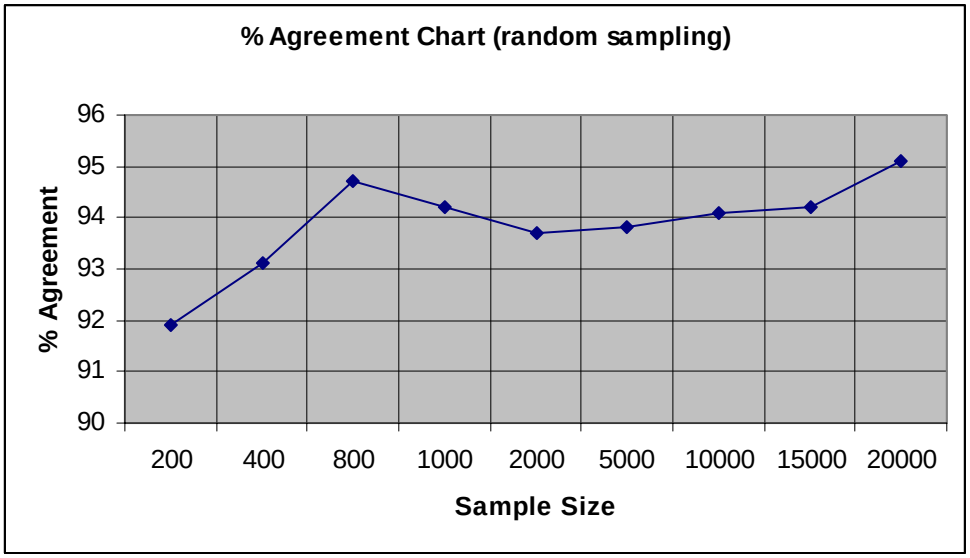

Figure 4.6.1 Neural Network result on training Seatbelt data (random sampling)

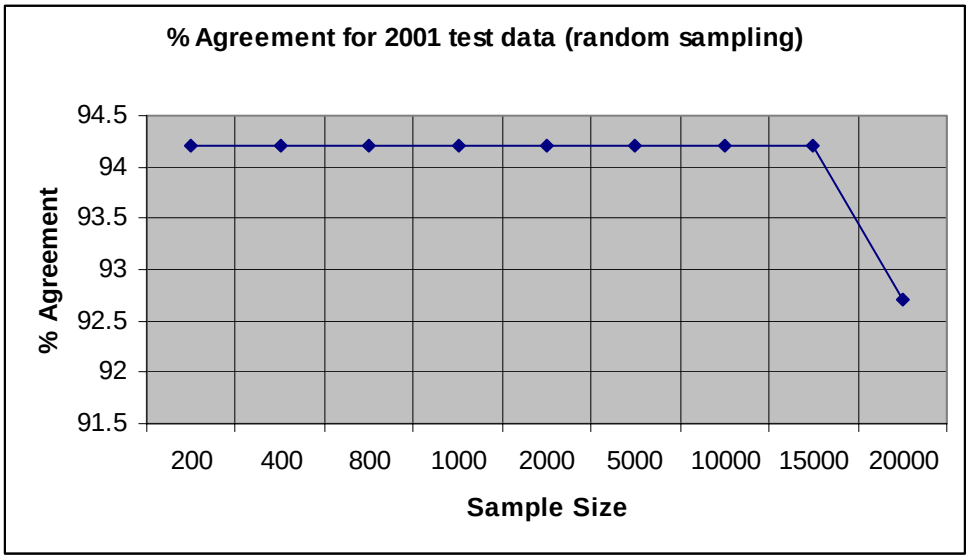

Figure 4.6.2 Neural Network result on year 2001 Seatbelt data (random sampling)

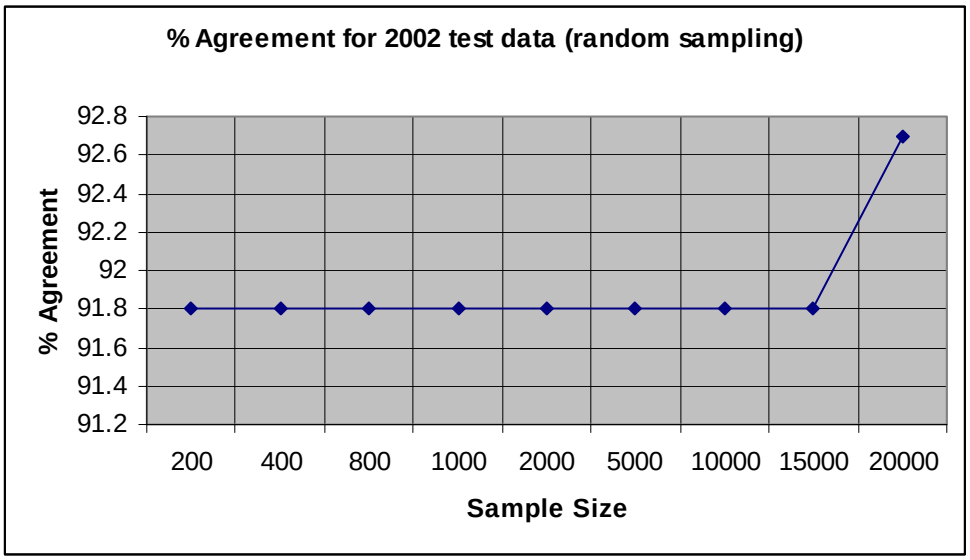

Figure 4.6.3 Neural Network result on year 2002 Seatbelt data (random sampling)

As the model predicts a "1" 100% of the time for DR_PROTSYS_CD, and fails to predict any of the "0" values correctly, excepting for a sample size of 20,000, the overall prediction rate for the test data for both years 2001 and 2002 remain same over the sample sizes 200 through 10,000. For sample size of 20,000, the overall classification rate falls below the constant rate for year 2001 data and rises for year 2002 data, though the individual classification % for "0"s and

"1"s are the same for both the years. This might be due to the difference in actual numbers of "0" and "1" values of DR_PROTSYS_CD in the two datasets. This is also reflected in the classification accuracy graph for the training data. If the training results are observed, it is seen that for different sample sizes the overall prediction accuracies are different though the prediction accuracy for "1" is always 100% and that for "0" is always 0%. This is due to the difference in the actual numbers of "0"s and "1"s in different sample sizes.

When the neural network models were built by using a stratified sampling method, stratifying by the driver's age variable, the prediction accuracies varied according to the sample sizes. Table 4.6.4 shows the summary of the results listing prediction accuracies for different sample sizes for the training data while Table 4.6.5 and Table 4.6.6 show the summary of results for different sample sizes when the models built for each sample size was applied to test the validity of prediction for the whole datasets for years 2001 and 2002 respectively. The graphs plotting the overall % agreement against the sample sizes for the training data and test data for years 2001 and 2002 are shown in Figure 4.6.4, Figure 4.6.5 and Figure 4.6.6 respectively.

As in the case of neural network model with random sampling from the year 2001 Seatbelt dataset, it is seen that except for the sample size of 800, 15000 and 20000, the prediction accuracy of DR_PROTSYS_CD for value "1" is 100% for all other sample sizes and that of value "0" is 0% and can be attributed to the fact that the ratio of "0" to "1" in the population is less than 1:7 which makes it difficult for the neural network model to correctly predict the "0" values of DR_PROTSYS_CD and it always predicts a "1" value irrespective of the values of the predictors. So, it is successful in predicting "1"s correctly 100% of the time and "0" correctly 0% of the time.

Table 4.6.4 Neural Network result on training Seatbelt data (strat. sampling)

Sample Size	% Agree		
	0	1	Overall
200	0.0	100.0	95.3
400	0.0	100.0	94.2
800	24.4	99.7	94.8
1000	0.0	100.0	92.8
2000	0.0	100.0	94.2

(table cont.)

5000	0.0	100.0	94.2
10000	0.0	100.0	94.2
15000	17.2	99.9	95.2
20000	16.4	99.9	95.0

Table 4.6.5 Results Neural Network result on year 2001 Seatbelt data (strat. sampling)

	% Agree		
Sample Size	**0**	**1**	**Overall**
200	0.0	100.0	94.2
400	0.0	100.0	94.2
800	16.1	99.8	94.9
1000	0.0	100.0	94.2
2000	0.0	100.0	94.2
5000	0.0	100.0	94.2
10000	0.0	100.0	94.2
15000	17.2	99.8	95.1
20000	12.4	99.9	92.7

Table 4.6.6 Neural Network result on year 2002 Seatbelt data (strat. sampling)

	% Agree with test data		
Sample Size	**0**	**1**	**Overall**
200	0.0	100.0	91.8
400	0.0	100.0	91.8
800	12.1	99.7	92.6
1000	0.0	100.0	91.8
2000	0.0	100.0	91.8
5000	0.0	100.0	91.8
10000	0.0	100.0	91.8
15000	12.8	99.8	92.7
20000	12.4	99.9	92.7

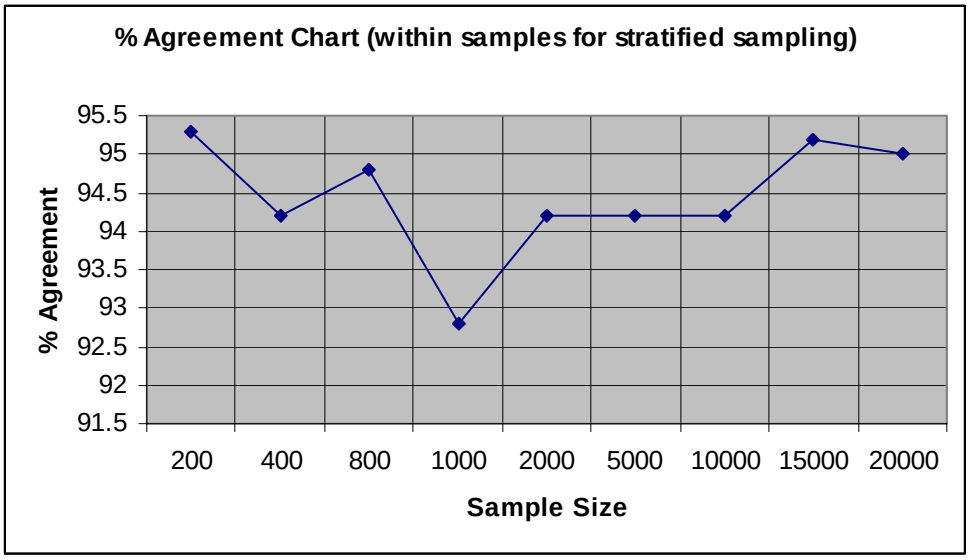

Figure 4.6.4 Neural Network result on training Seatbelt data (strat. sampling)

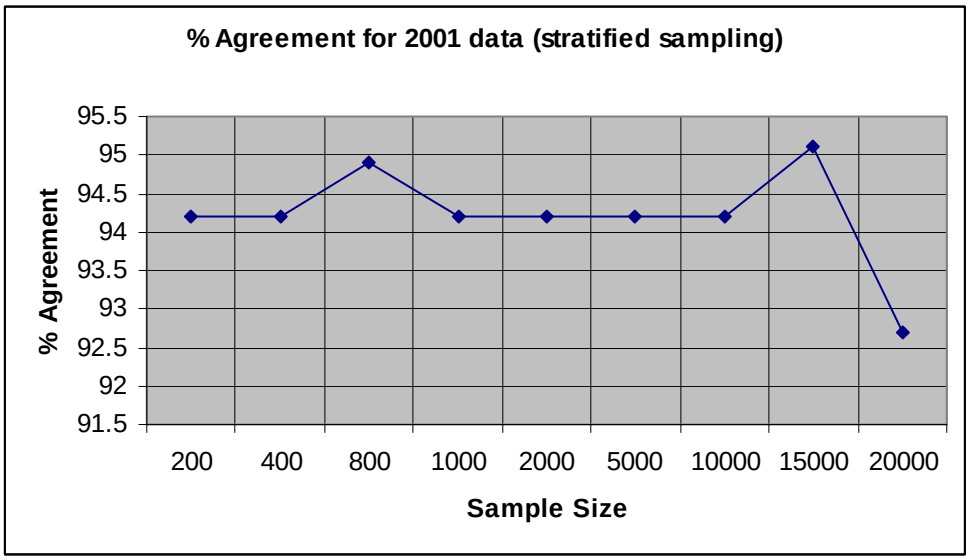

Figure 4.6.5 Neural Network result on year 2001 Seatbelt data (strat. sampling)

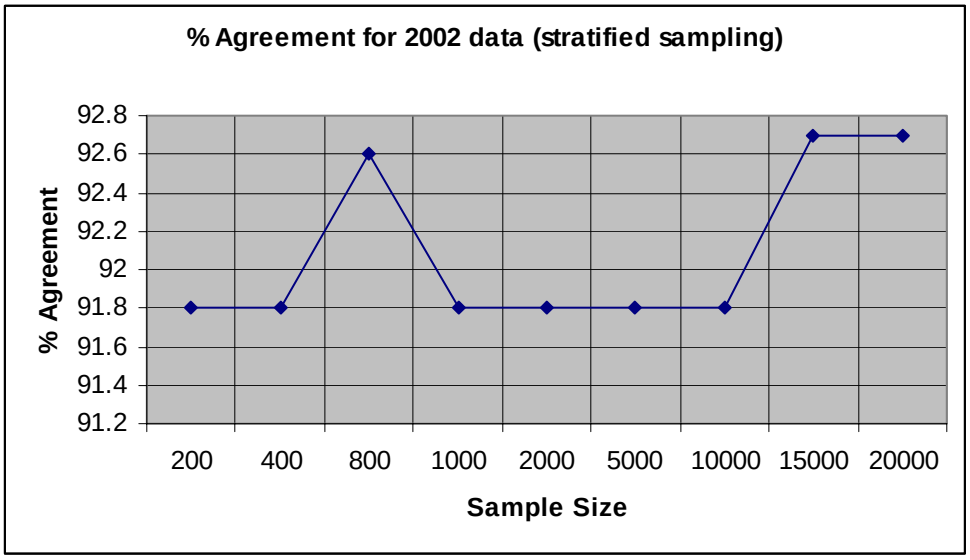

Figure 4.6.6 Neural Network result on year 2002 Seatbelt data (strat. sampling)

As the model predicts a "1" 100% of the time for DR_PROTSYS_CD, and fails to predict any of the "0" values correctly, excepting for a sample size of 800, 15000 and 20000, the overall prediction rate for the test data for both years 2001 and 2002 remain same over other sample sizes. For sample sizes of 800 and 15000, the overall classification rates for both 2001 data and 2002 data rise above the constant rate, while for sample size of 20000, the overall classification rate falls below the constant rate for year 2001 data and rises for year 2002 data, though the individual classification % for "0"s and "1"s are the same for both the years. This might be due to the difference in actual numbers of "0" and "1" values of DR_PROTSYS_CD in the two datasets. This is also reflected in the classification accuracy graph for the training data. If the training results are observed, it is seen that for different sample sizes the overall prediction accuracies are different though the prediction accuracy for "1" is always 100% and that for "0" is always 0%. This is due to the difference in the actual numbers of "0"s and "1"s in different sample sizes.

So, it can be inferred here that, when the distribution of the dependent variable is skewed, a neural network is not able to predict correctly the value which has a low occurrence, irrespective

of the method of sampling. It tends to classify the dependent variable for every instance into the class that predominates in the population.

Since we see that the prediction rate for the 0 values of the dependent variable is very low with neural network method also (for both the sampling methods), and this could be attributed to the very high ratio of 1:0 in the population, as in the case of decision tree and logistic regression, another set of analysis is done with the dataset from the 2001 population where only the records with value of injury codes (SEVERITY_CD) “A” and “B” are chosen. The classification of the variable SEVERITY_CD in the original dataset is done as “A” = 1, “B” = 0 and 2 = others and all records with a value of SEVERITY_CD = 2 are removed. This leaves a much reduced dataset with only 395 records which has a much less skewed distribution of “0”s and “1”s for the dependent variable. The ratio of 0:1 for DR_PROTSYS_CD in this reduced dataset is around 1:3. When the neural network model is built with the whole reduced 2001 dataset and the model is tested on 2002 data (which is also reduced as only records with SEVERITY_CD = “A” or “B” are retained), the results obtained are shown in the table 4.6.7.

Table 4.6.7 Neural Network results on modified Seatbelt training and test data

Data	% Agree with test data		
	0	**1**	**Overall**
Year 2001 training data	62.7	93.9	86.4
Year 2002 test data	50.0	95.8	85.2

Thus, it is seen that though the overall prediction rate is somewhat lower than that obtained for the original Seatbelt data, the prediction accuracy of “0” values improve a great deal over that with the original data. This may again be attributed to the more even distribution of 0 and 1 values of the dependent variable. This is an indicator that neural network model is not very accurate in predicting correctly the value of the dependent variable which has a very low occurrence in the population.

This shows that none of the three models can predict correctly the value of the variable which has a very low occurrence in the population.

4.7 Fatality Dataset Analysis with Decision Tree

When the decision tree model was built using the Fatality dataset using year 2001 data for different sample sizes and the sampling method used was simple random sampling, the analyses showed that the most important variable used for classifying the dependent variable SEVERITY_CD (most severe injury in the crash) into the correct class was EST_ALCOHOL (alcohol involvement in the crash) for all the sample sizes. The next important variables in terms of predicting SEVERITY_CD differed when the sample sizes were different. The prediction rates also varied according to the sample size. Table 4.7.1 shows the summary of the results along with the importance of variables in predicting the dependent variable for the training data while Table 4.7.2 and Table 4.7.3 show the summary of results for different sample sizes when the models built for each sample size was applied to test the validity of prediction for the whole dataset for years 2001 and 2002 respectively. The graphs plotting the overall % agreement against the sample sizes for the training data and test data for years 2001 and 2002 are shown in Figure 4.7.2, Figure 4.7.2 and Figure 4.7.3 respectively.

If the classification agreement % for the "1" and "0" values of SEVERITY_CD is observed, it is seen that there is a huge difference in the classification accuracy of "0" and "1" which increases with the sample size. This can be attributed to the fact that the dependent variable distribution in the population is highly skewed and the ratio of value "1" to value "0" for SEVERITY_CD is less than 1:4 in 2001 data and 2002 data. Also, since the population is around 13000 records, models were run for sample sizes of 200, 400, 800, 1000, 5000 and 10000. However, when the model for 10000 was run, a decision tree could not be built as there was no split on a dependent variable that could reduce the prediction error by more than the specified minimum deviance. This might also be attributed to the distribution of data.

Table 4.7.1 Decision Tree result on training Fatality data (random sampling)

Sample Size	% Agree			Important Predictors in order of Relative Importance
	0	1	Overall	
200	98.8	22.9	85.5	est_alcohol, num_occ

(table cont.)

400	98.8	8.3	82.5	num_occ_no_seatb, est_alcohol, aggressive
800	98.8	8.1	81.9	est_alcohol, num_occ_no_seatb, num_occ num_veh, aggressive, trk_bus_inv, violation, man_coll_cd
1000	99.1	4.8	79.6	est_alcohol, trk_bus_inv
2000	99.8	2.1	81.6	est_alcohol, trk_bus_inv
5000	99.8	2.4	81.8	est_alcohol, trk_bus_inv, man_coll_cd, num_occ_no_seatb

Table 4.7.2 Decision Tree result on year 2001 Fatality data (random sampling)

Sample Size	% Agree		
	0	1	Overall
200	94.4	11.9	79.5
400	98.2	5.8	81.5
800	97.8	6.9	81.4
1000	98.7	3.1	81.4
2000	99.6	1.7	81.9
5000	99.6	1.9	81.9

Table 4.7.3 Decision Tree result on year 2002 Fatality data (random sampling)

Sample Size	% Agree with test data		
	0	1	Overall
200	94.3	12.6	79.4
400	98.3	5.3	81.3
800	98.1	7.1	81.5
1000	98.7	2.6	81.2
2000	99.8	1.5	81.9
5000	99.7	1.9	81.9

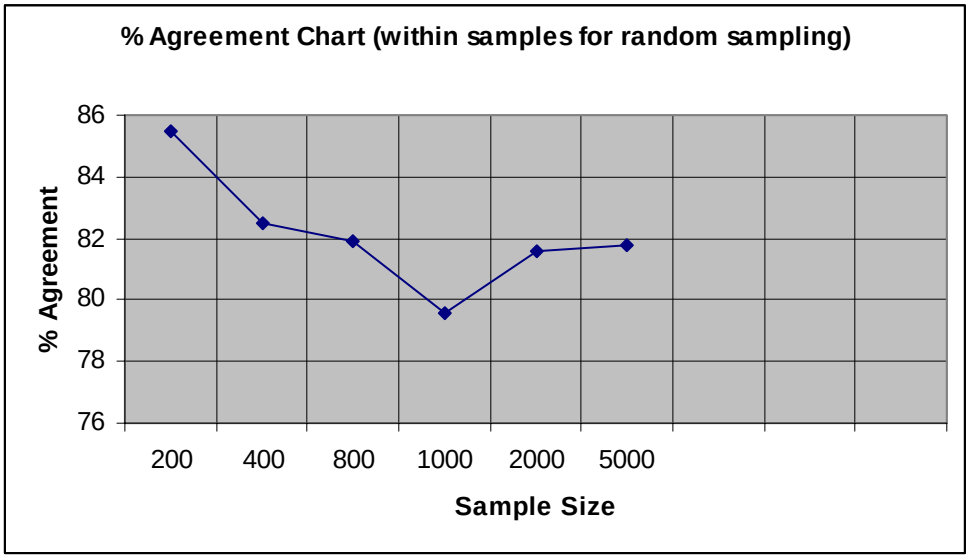

Figure 4.7.1 Decision Tree result on training Fatality data (random sampling)

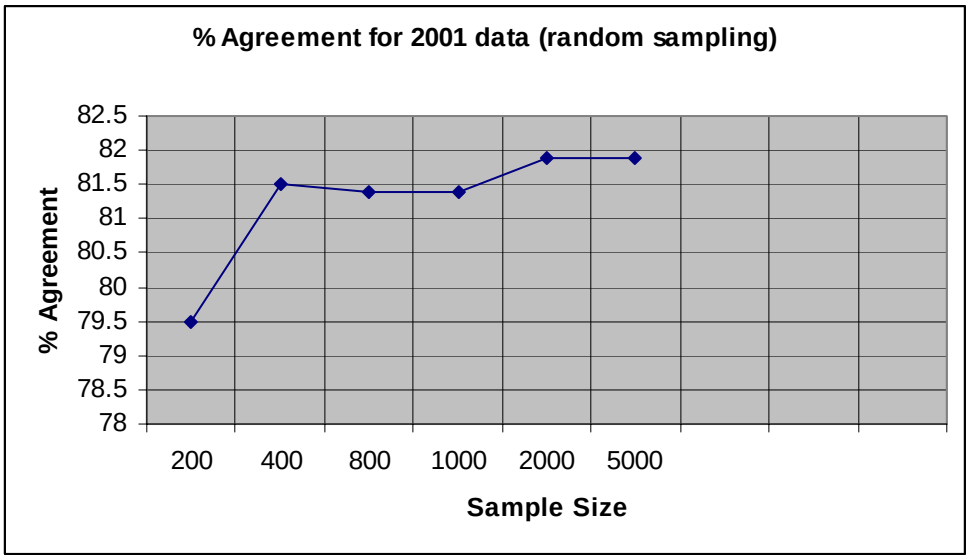

Figure 4.7.2 Decision Tree result on year 2001 Fatality data (random sampling)

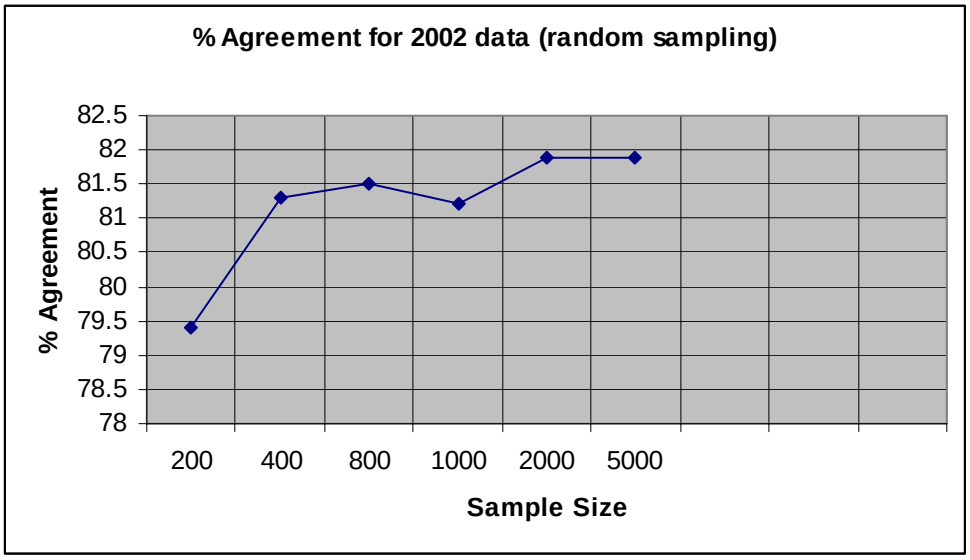

Figure 4.7.3 Decision Tree result on year 2002 Fatality data (random sampling)

It is seen that the overall prediction classification accuracy for test data for both the years was higher than that for the training data for smaller sample sizes. While the classification agreement of the training data was in a range of 80% to 86%, the classification accuracy of the both the test datasets were in a tighter range of 79.5% to 82%. For both the test data sets, classification accuracy did not improve much when the sample size was increased beyond 400. The results were replicated for both the datasets.

When the decision tree models were built by using a stratified sampling method, stratifying by the variable EST_ALCOHOL or the alcohol involvement in the crash, EST_ALCOHOL was found to be the most important variable for predicting the dependent variable SEVERITY_CD for all sample sizes except 200. Table 4.7.4 shows the summary of the results along with the importance of variables in predicting the dependent variable for the training data while Table 4.7.5 and Table 4.7.6 show the summary of results for different sample sizes when the models built for each sample size was applied to test the validity of prediction for the whole dataset for years 2001 and 2002 respectively. The graphs plotting the overall % agreement against the

sample sizes for the training data and test data for years 2001 and 2002 are shown in Figure 4.7.4, Figure 4.7.5 and Figure 4.7.6 respectively.

In line with the results of the decision tree model for Fatality dataset with random sampling, if the classification agreement % for the "1" and "0" values of SEVERITY_CD is observed, it is seen that there is a huge difference in the classification accuracy of "0" and "1" which increases with sample size. Again, this can be attributed to the fact that the dependent variable distribution in the population is highly skewed and the ratio of value "1" to value "0" for SEVERITY_CD is around 22:100 in 2001 data and 6:100 in 2002 data. Also, since the population is around 13000 records, models were run for sample sizes of 200, 400, 800, 1000, 5000 and 10000. In case of stratified sampling, the decision tree model for 10000 could be constructed as opposed to the case for random sampling.

Table 4.7.4 Decision Tree result on training Fatality data (strat. sampling)

Sample Size	% Agree			Important Predictors in order of Relative Importance
	0	1	Overall	
200	98.8	8.3	82.5	trk_bus_inv, man_coll_cd
400	98.5	10.7	82.0	est_alcohol, aggressive, num_occ_no_seatb, num_veh, trk_bus
800	98.8	8.7	82.0	est_alcohol, num_veh, trk_bus_inv, num_occ_no_seatb, violation, man_coll_cd
1000	99.8	1.7	82.6	est_alcohol, man_coll_cd, violation, num_veh
2000	99.2	4.9	83.9	est_alcohol, num_occ_no_seatb, aggressive, man_coll_cd, violation, num_veh
5000	99.9	1.2	82.2	est_alcohol, trk_bus, man_coll_cd, num_occ_no_seatb
10000	99.7	2.1	81.9	est_alcohol, num_occ_no_seatb, trk_bus_inv, man_coll_cd, violation

Table 4.7.5 Decision Tree result on year 2001 Fatality data (strat. sampling)

Sample Size	% Agree		
	0	1	Overall
200	98.9	3.2	81.6
400	97.9	4.4	81.0
800	97.3	3.4	80.3
1000	99.4	1.2	81.7
2000	98.7	4.4	81.6

(table cont.)

5000	99.8	0.9	81.9
10000	99.6	1.9	81.9

Table 4.7.6 Decision Tree result on year 2002 Fatality data (strat. sampling)

Sample Size	% Agree with test data		
	0	1	Overall
200	99.1	2.9	81.5
400	98.4	3.8	81.1
800	97.4	3.0	80.2
1000	99.4	1.1	81.5
2000	98.7	3.6	81.4
5000	99.8	0.7	81.8
10000	99.7	1.9	81.9

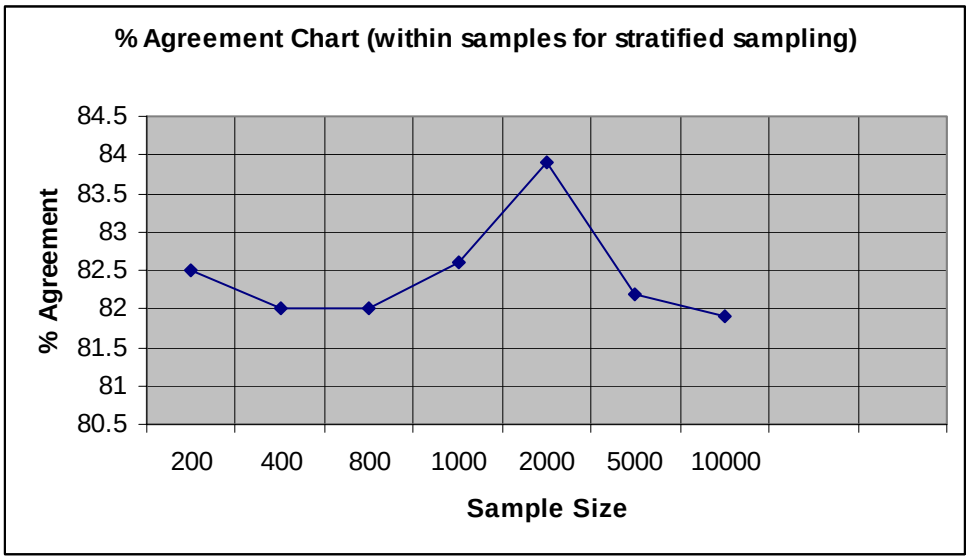

Figure 4.7.4 Decision Tree result on training Fatality data (strat. sampling)

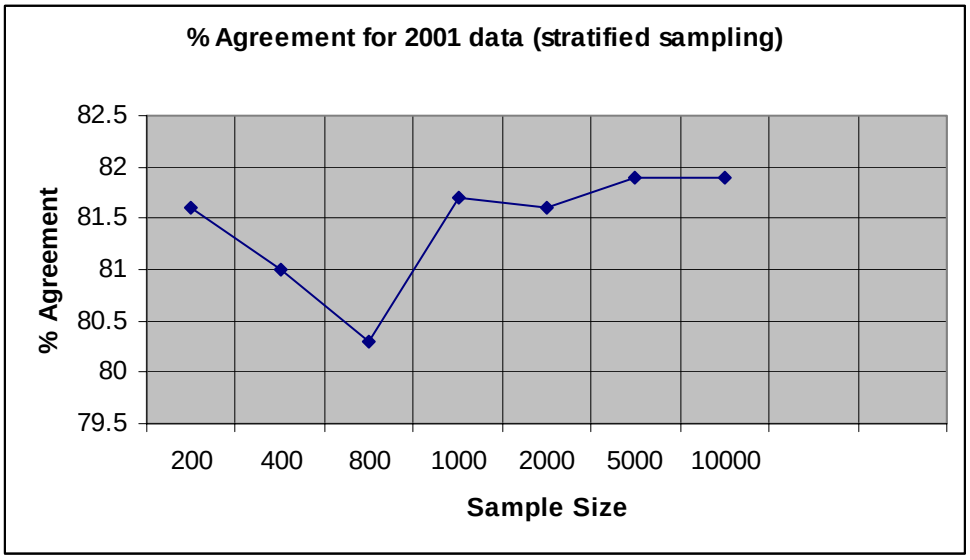

Figure 4.7.5 Decision Tree result on year 2001 Fatality data (strat. sampling)

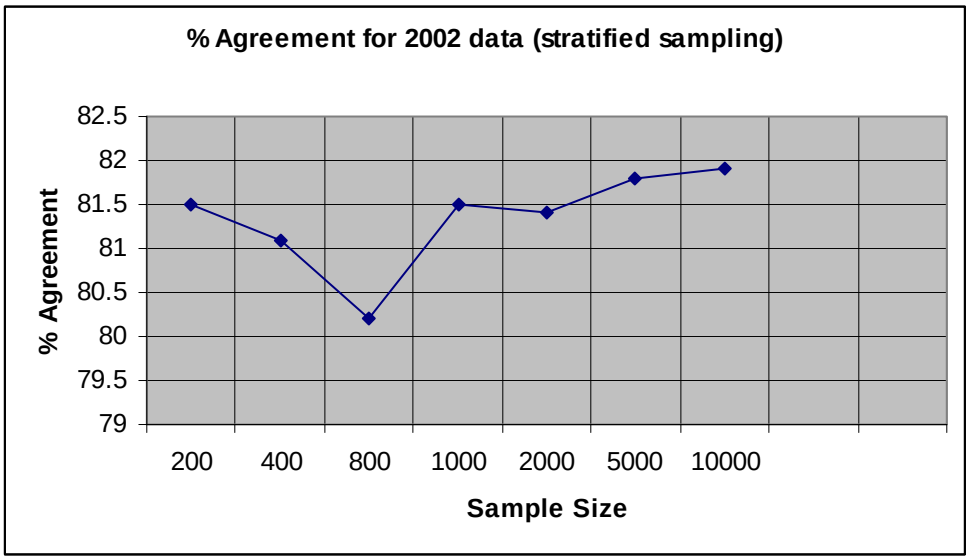

Figure 4.7.6 Decision Tree result on year 2002 Fatality data (strat. sampling)

The results in case of stratified sampling show that the prediction accuracy for training data was a bit higher than that for test data for both datasets. Actually the classification accuracy for the training data varied within a very tight range between 82% and 84% while that for test data for 2001 and 2002 was between 80% and 82%. The classification agreement for both the training

data and the test data showed no appreciable variation with respect to sample size and it was replicated for both the datasets.

Since we see that the prediction rate for the 1 values of the dependent variable is very low with decision tree method for the Fatality dataset also (for both the sampling methods), and this could be attributed to the very high ratio of 0:1 in the population, another set of analysis is done with the dataset from the 2001 population where only the records with value of injury codes (SEVERITY_CD) "A" and "B" are chosen. The classification of the variable SEVERITY_CD in the original dataset is done as "A" = 1, "B" = 0 and 2 = others and all records with a value of SEVERITY_CD = 2 are removed. This leaves a much reduced dataset with only 2419 records which has a much less skewed distribution of "0"s and "1"s for the dependent variable. The ratio of 1:0 for SEVERITY_CD in this reduced dataset is less than 1:2. When the decision tree model is built with three sample sizes of 1000, 2000 and the whole set chosen from the reduced 2001 dataset and the model is tested on the complete reduced 2001 and 2002 datasets (which is also reduced as only records with SEVERITY_CD = "A" or "B" are retained), the results obtained are shown in the table 4.7.7.

Table 4.7.7 Decision Tree results on modified Fatality training and test data

Data	Model with Sample Size	% Agree with test data		
		0	1	Overall
Year 2001 training data	1000	83.2	54.3	73.1
	2000	79.6	47.0	68.0
	Whole (2419)	83.5	44.8	69.7
Year 2001 test data	1000	80.3	49.6	69.4
	2000	79.6	46.8	68.0
	Whole (2419)	83.5	44.8	69.7
Year 2002 test data	1000	79.2	42.9	67.1
	2000	78.1	44.7	67.1
	Whole (2419)	82.3	41.7	68.8

It is seen that though overall prediction rate is much lower than that obtained for the original Fatality data but the prediction accuracy of "1" values improve a great deal over that with the original data. This can again be attributed to the more even distribution of 0 and 1 values of the

dependent variable. This again shows that decision tree model is not very accurate in predicting correctly the value of the dependent variable which has a very low occurrence in the population. The overall fall in the accuracy prediction may be due to the choice of predictors. An optimal choice of predictors would definitely increase the classification accuracy level.

4.8 Fatality Dataset Analysis with Logistic Regression

As in the case of decision trees, when the logistic regression analysis was performed using the Fatality dataset for year 2001 data with different sample sizes and the sampling method used was simple random sampling, the analyses showed that except for sample sizes of 200 and 400, EST_ALCOHOL was identified as the most important variable in classifying the dependent variable SEVERITY_CD into the correct class for all other sample sizes. The prediction rates varied according to the sample size. Table 4.8.1 shows the summary of the results along with the importance of variables in predicting the dependent variable for the training data while Table 4.8.2 and Table 4.8.3 show the summary of results for different sample sizes when the models built for each sample size was applied to test the validity of prediction for the whole dataset for years 2001 and 2002 respectively. The graphs plotting the overall % agreement against the sample sizes for the training data and test data for years 2001 and 2002 are shown in Figure 4.8.2, Figure 4.8.2 and Figure 4.8.3 respectively.

If the classification agreement % for the “1” and “0” values of SEVERITY_CD is observed, there is a huge difference in the prediction accuracy of “1”s and that of “0”s and as the sample size increases, the difference between prediction accuracy of “1”s and that of “0”s increase. This can be attributed to the fact that the ratio of “1” to “0” in the population is less than 1:5. Also, since the population is around 13000 records, models were run for sample sizes of 200, 400, 800, 1000, 5000 and 10000.

Table 4.8.1 Logistic Regression result on training Fatality data (random sampling)

Sample Size	% Agree			Important Predictors in order of Relative Importance
	0	1	Overall	
200	96.4	8.6	81.0	violation, est_alcohol, aggressive, num_veh, man_coll_cd, num_occ_no_seatb, num_occ, trk_bus_inv
400	100.0	0.0	85.8	num_veh, num_occ, aggressive, trk_bus_inv, violation, man_coll_cd, num_occ_no_seatb, est_alcohol
800	99.7	3.6	83.1	est_alcohol, trk_bus_inv, num_veh, aggressive, num_occ, man_coll_cd, num_occ_no_seatb, violation
1000	100.0	2.5	84.7	est_alcohol, trk_bus_inv, aggressive, man_coll_cd, num_occ, num_occ_no_seatb, violation, num_veh
2000	99.6	1.1	82.5	est_alcohol, num_occ_no_seatb, num_veh, man_coll_cd, trk_bus_inv, aggressive, num_occ, violation
5000	99.8	0.3	82.1	est_alcohol, trk_bus_inv, num_occ_no_seatb, num_veh, aggressive, man_coll_cd, num_occ, violation
10000	99.9	0.4	82.2	est_alcohol, trk_bus_inv, num_occ_no_seatb, num_veh, aggressive, num_occ, man_coll_cd, violation

Table 4.8.2 Logistic Regression result on year 2001 Fatality data (random sampling)

Sample Size	% Agree		
	0	1	Overall
200	96.8	5.1	80.2
400	99.9	0.3	81.9
800	99.1	2.7	81.6
1000	99.7	1.3	81.9
2000	99.6	1.2	81.8
5000	99.9	0.5	81.9
10000	99.9	0.4	81.9

Table 4.8.3 Logistic Regression result on year 2002 Fatality data (random sampling)

Sample Size	% Agree with test data		
	0	1	Overall
200	97.3	4.2	80.3
400	99.9	0.1	81.7
800	99.4	3.0	81.8
1000	99.9	1.1	81.9
2000	99.7	1.5	81.8
5000	99.9	0.5	81.8
10000	100.0	0.4	81.8

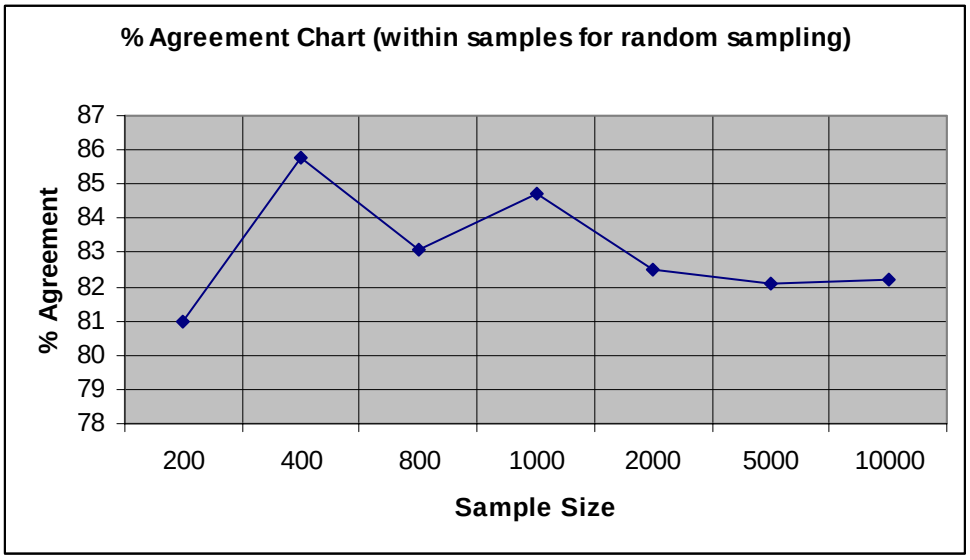

Figure 4.8.1 Logistic Regression result on training Fatality data (random sampling)

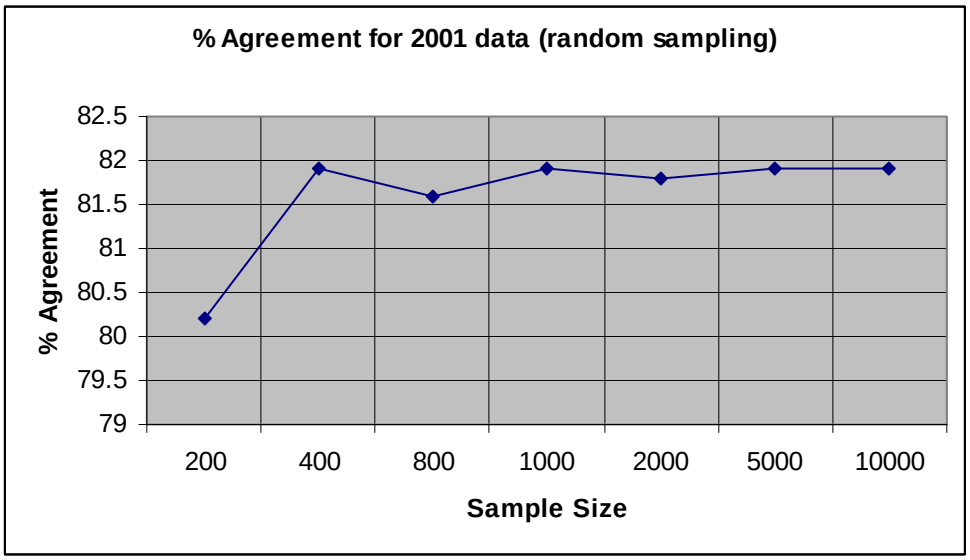

Figure 4.8.2 Logistic Regression result on year 2001 Fatality data (random sampling)

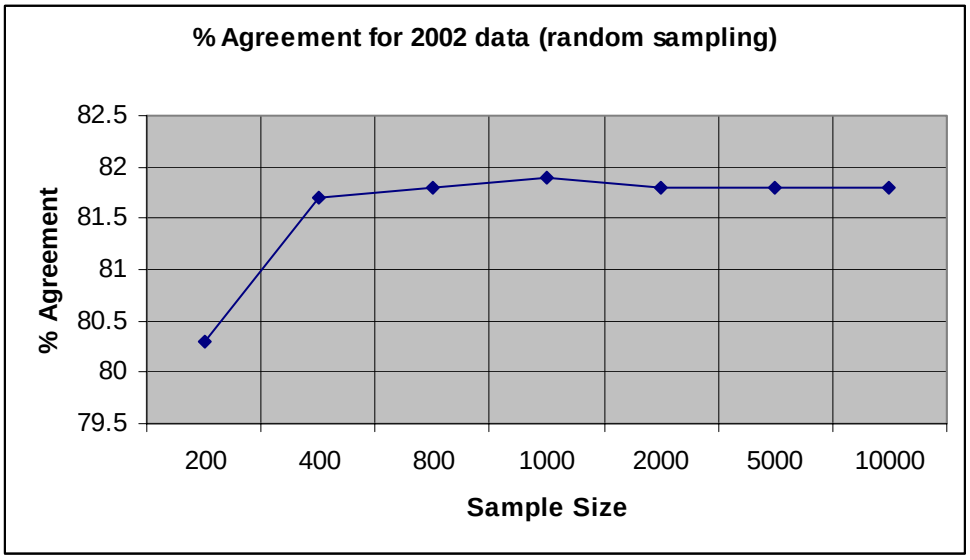

Figure 4.8.3 Logistic Regression result on year 2002 Fatality data (random sampling)

It is seen that the overall prediction classification accuracy for the training data was a bit higher than that of year 2001 and 2002 test data for all sample sizes. The classification accuracy attained a highest value at a sample size of 400 for training data and the accuracy fell when the sample size was increased. For both 2001 and 2002 test datasets, the classification accuracy reached a stable value of a little less than 82% at the sample size of 400 and not much could be gained in terms of prediction accuracy by increasing the sample size over 400. So, it is seen that, if the sampling method is random, for logistic regression models for the Fatality datasets, a much lower sample size is required and the prediction accuracy is consistently replicated over different test datasets.

When the logistic regression analyses were performed on samples drawn from the year 2001 Fatality dataset using stratified sampling method, stratifying by alcohol involvement in the crash, EST_ALCOHOL variable, the importance of the predictor variables in classifying the dependent variable SEVERITY_CD and the classification agreements for different sample sizes were similar to the case of logistic regression model with random sampling. Except for the sample size of 200, EST_ALCOHOL was found to be the most important variable in classifying

SEVERITY_CD for all other sample sizes. The prediction accuracies varied with models for different sample sizes. Table 4.8.4 shows the summary of the results along with the importance of variables in predicting the dependent variable for the training data while Table 4.8.5 and Table 4.8.6 show the summary of results for different sample sizes when the models built for each sample size was applied to test the validity of prediction for the whole dataset for years 2001 and 2002 respectively. The graphs plotting the overall % agreement against the sample sizes for the training data and test data for years 2001 and 2002 are shown in Figure 4.8.4, Figure 4.8.5 and Figure 4.8.6 respectively.

Table 4.8.4 Logistic Regression result on training Fatality data (strat. sampling)

Sample Size	% Agree			Important Predictors in order of Relative Importance
	0	1	Overall	
200	98.2	13.9	83.0	num_occ, est_alcohol, num_veh, trk_bus_inv, aggressive, num_occ_no_seatb, man_coll_cd, violation
400	100.0	0.0	86.2	est_alcohol, num_veh, man_coll_cd, num_occ, aggressive, num_occ_no_seatb, trk_bus_inv, violation
800	99.8	0.0	80.0	est_alcohol, man_coll_cd, num_occ_no_seatb, num_veh, aggressive, num_occ, violation, trk_bus_inv
1000	99.6	0.0	83.5	est_alcohol, aggressive, trk_bus_inv, violation, num_veh, man_coll_cd, num_occ_no_seatb, num_occ
2000	99.6	1.6	81.7	est_alcohol, num_veh, trk_bus_inv, num_occ, num_occ_no_seatb, man_coll_cd, violation, aggressive
5000	99.9	0.1	81.9	est_alcohol, trk_bus_inv, num_veh, num_occ_no_seatb, aggressive, num_occ, man_coll_cd, violation
10000	99.4	0.4	82.0	est_alcohol, trk_bus_inv, num_occ_no_seatb, num_veh, num_occ, man_coll_cd, aggressive, violation

Table 4.8.5 Logistic Regression result on year 2001 Fatality data (strat. sampling)

Sample Size	% Agree		
	0	1	Overall
200	97.8	6.5	81.2
400	100.0	0.0	81.9
800	99.7	0.4	81.7
1000	99.8	0.4	81.8
2000	99.6	1.9	81.9
5000	100.0	0.1	81.9
10000	99.8	0.5	81.9

Table 4.8.6 Logistic Regression result on year 2002 Fatality data (strat. sampling)

Sample Size	% Agree with test data		
	0	1	Overall
200	97.9	6.0	81.2
400	100.0	0.0	81.8
800	99.8	0.7	81.7
1000	99.9	0.5	81.8
2000	99.7	1.6	81.8
5000	100.0	0.1	81.8
10000	99.9	0.4	81.8

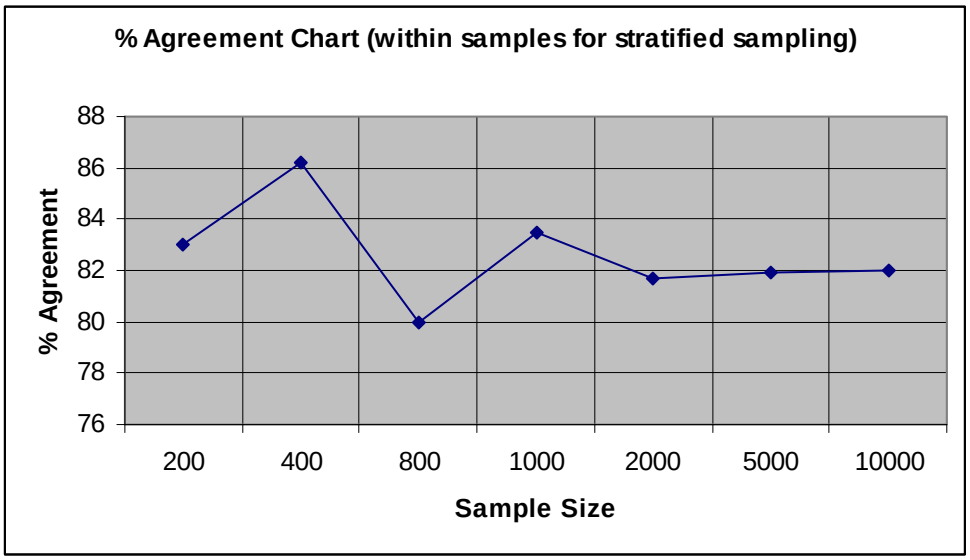

Figure 4.8.4 Logistic Regression result on training Fatality data (strat. sampling)

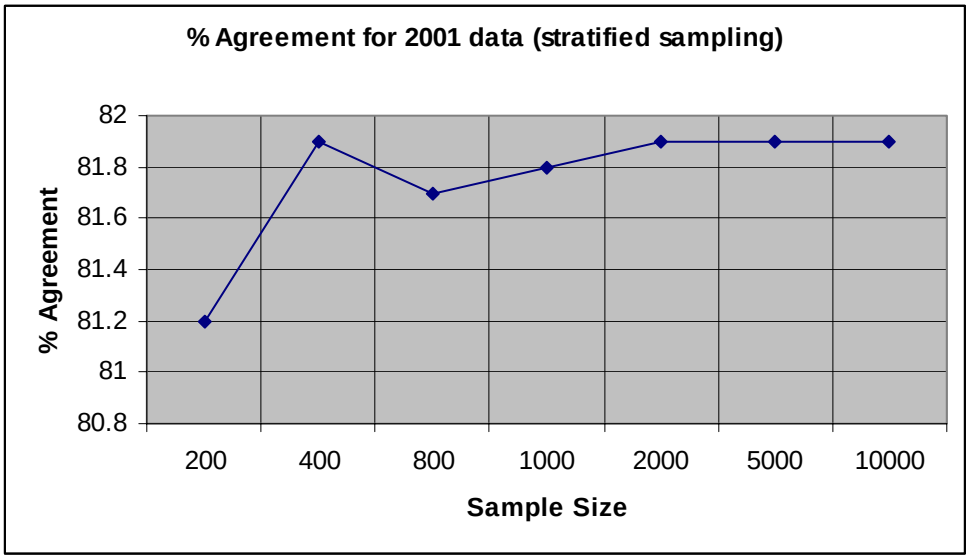

Figure 4.8.5 Logistic Regression result on year 2002 Fatality data (strat. sampling)

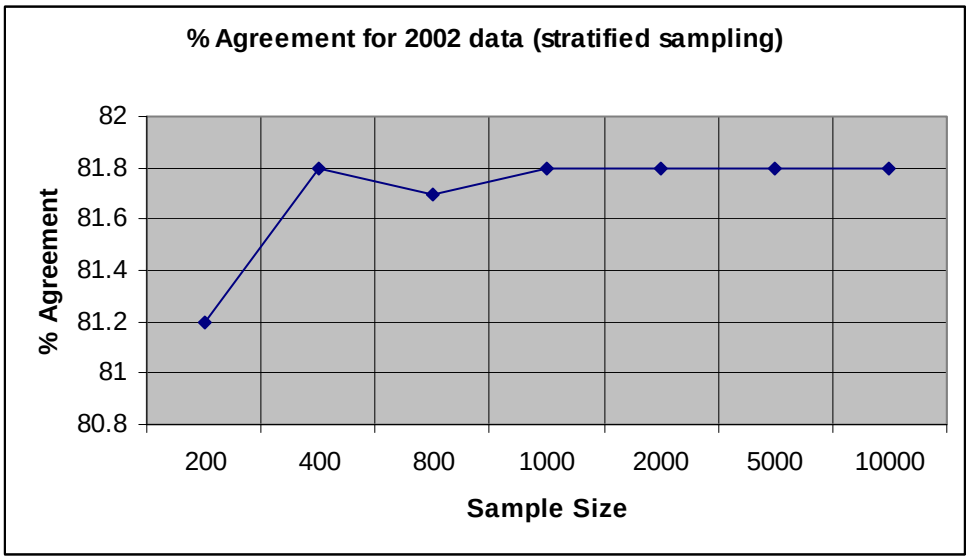

Figure 4.8.6 Logistic Regression result on year 2002 Fatality data (strat. sampling)

In this case of logistic regression model with Fatality dataset, when the sampling method is stratified, the prediction accuracy in case of training data varies between 80% and 86% and does not improve much after the sample size is increased beyond 1000. For both sets of test data, the prediction accuracies for all sample sizes are in a very tight range, between 81.2% and 81.8%. So, the classification accuracy hardly varies with sample size, and after sample size of 800, the

classification accuracy reaches a stable value. This is in line with the results obtained with random sampling method and the prediction accuracy does not vary if the sampling method is different for both sets of test data.

Since we see that the prediction rate for the 1 values of the dependent variable is very low with logistic regression method too for the Fatality dataset (for both the sampling methods), and this could be attributed to the very high ratio of 0:1 in the population, as before, another set of analysis is done with the dataset from the 2001 population where only the records with value of injury codes (SEVERITY_CD) "A" and "B" are chosen. The classification of the variable SEVERITY_CD in the original dataset is done as "A" = 1, "B" = 0 and 2 = others and all records with a value of SEVERITY_CD = 2 are removed. This leaves a much reduced dataset with only 2419 records which has a much less skewed distribution of "0"s and "1"s for the dependent variable. The ratio of 1:0 for SEVERITY_CD in this reduced dataset is less than 1:2. When the logistic regression model is built with three sample sizes of 1000, 2000 and the whole set chosen from the reduced 2001 dataset and the model is tested on the complete reduced 2001 and 2002 datasets (which is also reduced as only records with SEVERITY_CD = "A" or "B" are retained), the results obtained are shown in the table 4.8.7.

Table 4.8.7 Logistic Regression results on modified Fatality training and test data

Data	Model with Sample Size	% Agree with test data		
		0	1	Overall
Year 2001 training data	1000	89.9	26.5	68.1
	2000	83.2	39.0	67.6
	Whole (2419)	82.6	41.2	67.9
Year 2001 test data	1000	90.7	24.9	67.3
	2000	83.3	39.8	67.8
	Whole (2419)	82.6	41.2	67.9
Year 2002 test data	1000	90.0	25.9	69.3
	2000	82.6	40.2	68.6
	Whole (2419)	82.4	41.3	68.8

It is seen that though overall prediction rate is considerably lower than that obtained for the original Fatality data but the prediction accuracy of "1" values improve a greatly over that with

the original data. This can again be attributed to the more even distribution of 0 and 1 values of the dependent variable. This again shows that logistic regression model is not very accurate too in predicting correctly the value of the dependent variable which has a very low occurrence in the population. The overall fall in the accuracy prediction may again be attributed to the choice of predictors. An optimal choice of predictors would definitely increase the classification accuracy level.

4.9 Fatality Dataset Analysis with Neural Network

When neural network model was built using the Fatality dataset using year 2001 data for different sample sizes, the sampling method being random sampling, the analyses showed that the prediction accuracy varied according to sample size. Table 4.9.1 shows the summary of the results listing the classification accuracy for different sample sizes for the training data while Table 4.9.2 and Table 4.9.3 show the summary of results for different sample sizes when the models built for each sample size was applied to test the validity of prediction for the whole dataset for years 2001 and 2002 respectively. The graphs plotting the overall % agreement against the sample sizes for training data and test data for years 2001 and 2002 are shown in Figure 4.9.2, Figure 4.9.2 and Figure 4.9.3 respectively.

If the classification agreement % for the “1” and “0” values of SEVERITY_CD in the training data as well as the test data is observed, it is seen that for all sample sizes, the prediction accuracy of “0”s are 100% for all other sample sizes and that of “1”s are 0%. This can be attributed to the fact that the ratio of “1” to “0” in the population is less than 1:5. Hence when the neural network is taught to read patterns from the training data, most of the time it is trained to predict a “0” irrespective of the predictor values, so it predicts a “1” every time for SEVERITY_CD and fails to predict the “1”s. So, it is successful in predicting “0”s correctly 100% of the time and “1” correctly 0% of the time. As with the other models, the neural network

model was also run for sample sizes of 200, 400, 800, 1000, 5000 and 10000 since the population is around 13000 records.

Table 4.9.1 Neural Network result on training Fatality data (random sampling)

	% Agree		
Sample Size	**0**	**1**	**Overall**
200	100.0	0.0	80.0
400	100.0	0.0	83.2
800	100.0	0.0	84.1
1000	100.0	0.0	82.8
2000	100.0	0.0	82.5
5000	100.0	0.0	82.3
10000	100.0	0.0	82.0

Table 4.9.2 Neural Network result on year 2001 Fatality data (random sampling)

	% Agree		
Sample Size	**0**	**1**	**Overall**
200	100.0	0.0	81.9
400	100.0	0.0	81.9
800	100.0	0.0	81.9
1000	100.0	0.0	81.9
2000	100.0	0.0	81.9
5000	100.0	0.0	81.9
10000	100.0	0.0	81.9

Table 4.9.3 Neural Network result on year 2002 Fatality data (random sampling)

	% Agree with test data		
Sample Size	**0**	**1**	**Overall**
200	100.0	0.0	81.8
400	100.0	0.0	81.8
800	100.0	0.0	81.8
1000	100.0	0.0	81.8
2000	100.0	0.0	81.8
5000	100.0	0.0	81.8
10000	100.0	0.0	81.8

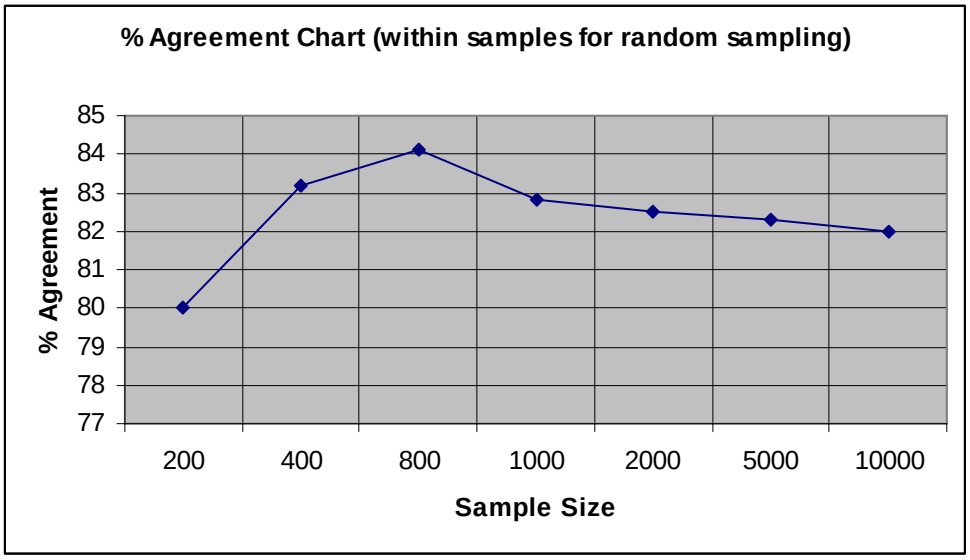

Figure 4.9.1 Neural Network result on training Fatality data (random sampling)

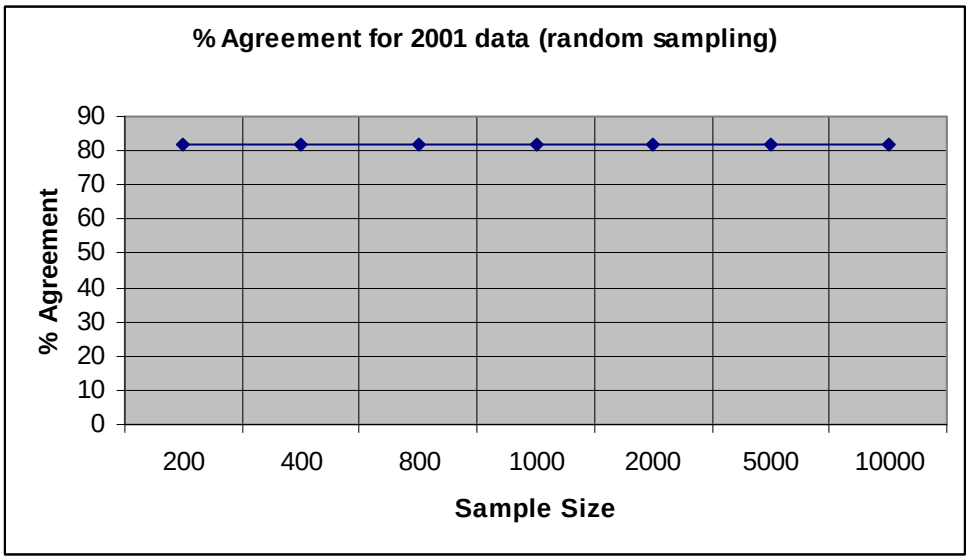

Figure 4.9.2 Neural Network result on year 2001 Fatality data (random sampling)

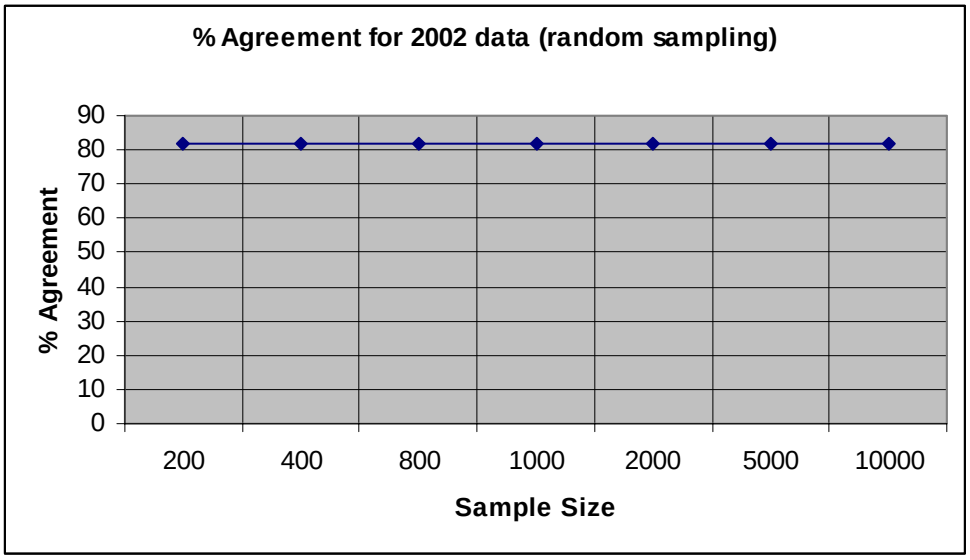

Figure 4.9.3 Neural Network result on year 2002 Fatality data (random sampling)

As the model predicts a "0" 100% of the time for SEVERITY_CD, and fails to predict any of the "1" values correctly, the overall prediction rate for the test data for both years 2001 and 2002 remain same over the sample sizes 200 through 10,000. If the training results are observed, it is seen that for different sample sizes the overall prediction accuracies are different though the prediction accuracy for "1" is always 100% and that for "0" is always 0%. This might be due to the difference in actual numbers of "0" and "1" values of SEVERITY_CD in different sample sizes.

When the neural network models were built by using a stratified sampling method, stratifying by the alcohol involvement in the crash, EST_ALCOHOL variable, the prediction accuracies varied according to the sample sizes. Table 4.9.4 shows the summary of the results listing prediction accuracies for different sample sizes for the training data while Table 4.9.5 and Table 4.9.6 show the summary of results for different sample sizes when the models built for each sample size was applied to test the validity of prediction for the whole datasets for years 2001 and 2002 respectively. The graphs plotting the overall % agreement against the sample sizes for the

training data and test data for years 2001 and 2002 are shown in Figure 4.9.4, Figure 4.9.5 and Figure 4.9.6 respectively.

It is very interesting to note that, similar to the neural network model with stratified sampling for Seatbelt dataset, in the case of neural network model with stratified sampling with Fatality dataset also, it is seen that except for the sample size of 800, the prediction accuracy of SEVERITY_CD for value "0" is 100% for all other sample sizes and that of value "1" is 0% and can be attributed to the fact that the ratio of "0" to "1" in the population is less than 1:5 which was the situation in case of Seatbelt dataset too. This makes it difficult for the neural network model to correctly predict the "1" values of SEVERITY_CD and it always predicts a "0" value irrespective of the values of the predictors. So, it is successful in predicting "0"s correctly 100% of the time and "1" correctly 0% of the time.

Table 4.9.4 Network result on training Fatality data (strat. sampling)

	% Agree		
Sample Size	**0**	**1**	**Overall**
200	100.0	0.0	81.0
400	100.0	0.0	82.0
800	99.5	0.6	78.6
1000	100.0	0.0	82.2
2000	100.0	0.0	82.0
5000	100.0	0.0	82.5
10000	100.0	0.0	81.9

Table 4.9.5 Neural Network result on year 2001 Fatality data (strat. sampling)

	% Agree		
Sample Size	**0**	**1**	**Overall**
200	100.0	0.0	81.9
400	100.0	0.0	81.9
800	99.7	0.5	81.8
1000	100.0	0.0	81.9
2000	100.0	0.0	81.9
5000	100.0	0.0	81.9
10000	100.0	0.0	81.9

Table 4.9.6 Neural Network result on year 2002 Fatality data (strat. sampling)

Sample Size	% Agree with test data		
	0	1	Overall
200	100.0	0.0	81.8
400	100.0	0.0	81.8
800	99.8	0.9	81.7
1000	100.0	0.0	81.8
2000	100.0	0.0	81.8
5000	100.0	0.0	81.8
10000	100.0	0.0	81.8

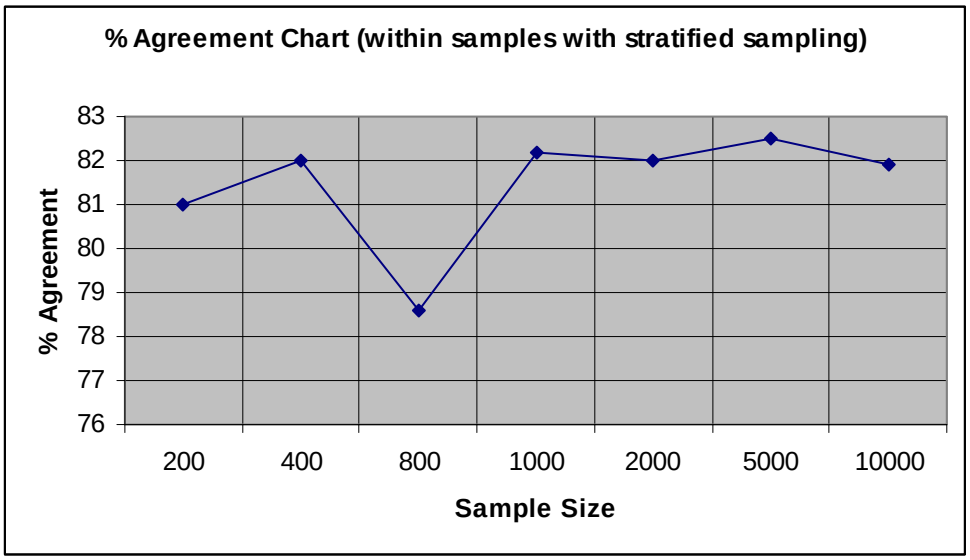

Figure 4.9.4 Neural Network result on training Fatality data (strat. sampling)

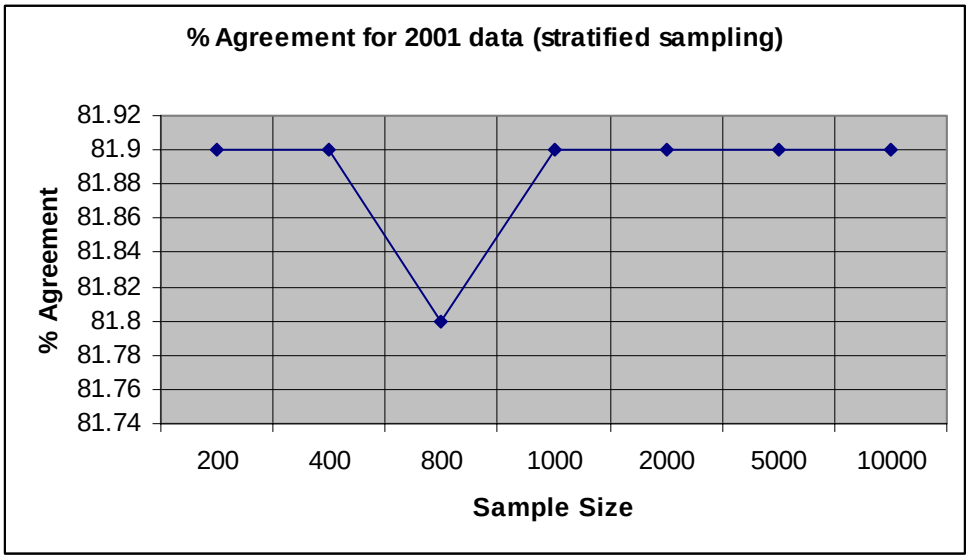

Figure 4.9.5 Neural Network result on year 2001 Fatality data (strat. sampling)

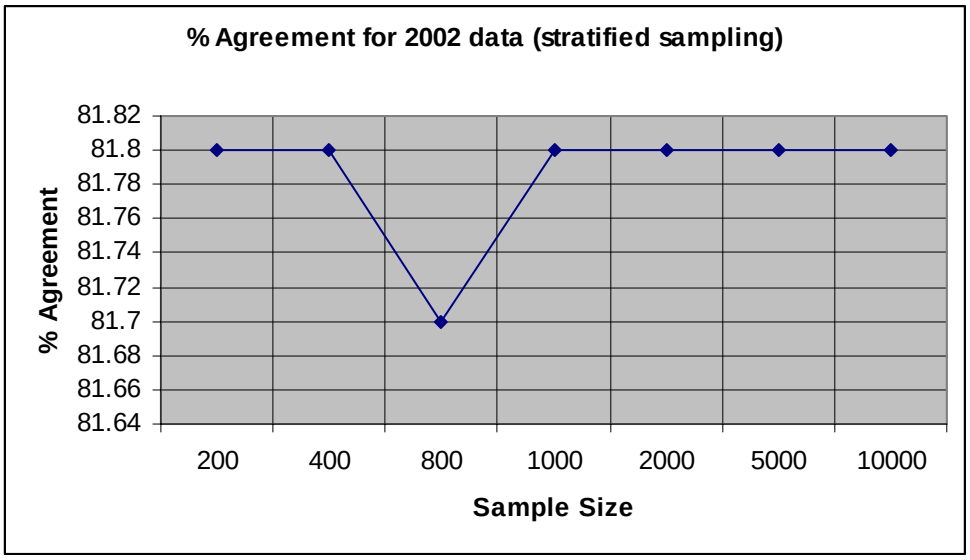

Figure 4.9.6 Neural Network result on year 2002 Fatality data (strat. sampling)

As the model predicts a "0" 100% of the time for SEVERITY_CD, and fails to predict any of the "1" values correctly, excepting for a sample size of 800, the overall prediction rate for the test data for both years 2001 and 2002 remain same over other sample sizes. For sample size 800, the overall classification rates for both 2001 data and 2002 data fall marginally by 0.1% below the

constant rate. If the training results are observed, it is seen that for different sample sizes the overall prediction accuracies are different though the prediction accuracy for "0" is always 100% and that for "1" is always 0%. This is due to the difference in the actual numbers of "0"s and "1"s in different sample sizes.

Since we see that the prediction rate for the 1 values of the dependent variable is very low with neural network method too for the Fatality dataset (for both the sampling methods), and this could be attributed to the very high ratio of 0:1 in the population, as before, another set of analysis is done with the dataset from the 2001 population where only the records with value of injury codes (SEVERITY_CD) "A" and "B" are chosen. The classification of the variable SEVERITY_CD in the original dataset is done as "A" = 1, "B" = 0 and 2 = others and all records with a value of SEVERITY_CD = 2 are removed. This leaves a much reduced dataset with only 2419 records which has a much less skewed distribution of "0"s and "1"s for the dependent variable. The ratio of 1:0 for SEVERITY_CD in this reduced dataset is less than 1:2. When the neural network model is built with three sample sizes of 1000, 2000 and the whole set chosen from the reduced 2001 dataset and the model is tested on the complete reduced 2001 and 2002 datasets (which is also reduced as only records with SEVERITY_CD = "A" or "B" are retained), the results obtained are shown in the table 4.9.7.

Table 4.9.7 Neural Network results on modified Fatality training and test data

Data	Model with Sample Size	% Agree with test data		
		0	1	Overall
Year 2001 training data	1000	75.0	59.1	69.2
	2000	77.0	52.9	68.4
	Whole (2419)	75.7	55.5	68.5
Year 2001 test data	1000	74.6	56.2	68.0
	2000	76.7	53.8	68.6
	Whole (2419)	75.7	55.5	68.5
Year 2002 test data	1000	73.8	53.5	67.1
	2000	75.6	52.7	68.0
	Whole (2419)	74.4	53.4	67.5

It is seen that though overall prediction rate is considerably lower than that obtained for the original Fatality data with neural network but the prediction accuracy of “1” values improve greatly over that with the original data, though the prediction accuracy of “0”s fall. This phenomenon can also be attributed to the more even distribution of 0 and 1 values of the dependent variable. This again shows that neural network model is not very accurate too in predicting correctly the value of the dependent variable which has a very low occurrence in the population. The overall fall in the accuracy prediction may again be attributed to the choice of predictors. An optimal choice of predictors would increase the classification accuracy level.

5. CONCLUSION

In the study conducted for the purpose of this thesis, the main objective was to compare the performance of three statistical and data mining classification models viz., logistic regression, decision tree and neural network models for different sample sizes and sampling methods on three sets of data. The data distributions in the three sets were very different.

By looking at the results obtained for the Alcohol dataset, it can be concluded that if the distribution of the dependent variable is not skewed, the classification accuracy for all the methods are consistent and it cannot be said that one method classifies the dependent variable significantly better than another. Moreover, when the models were applied to the test datasets, it is seen that a stable value of classification accuracy was reached at a sample size of 5000. The classification accuracy could not be improved by increasing the sample size. Also, the sampling method did not have any significant effect on the classification accuracy. This is clearly indicated in the figure 5.1 where the classification accuracies for all the models when applied to year 2002 test data, for different sample sizes and sampling methods, have been plotted against the sample sizes.

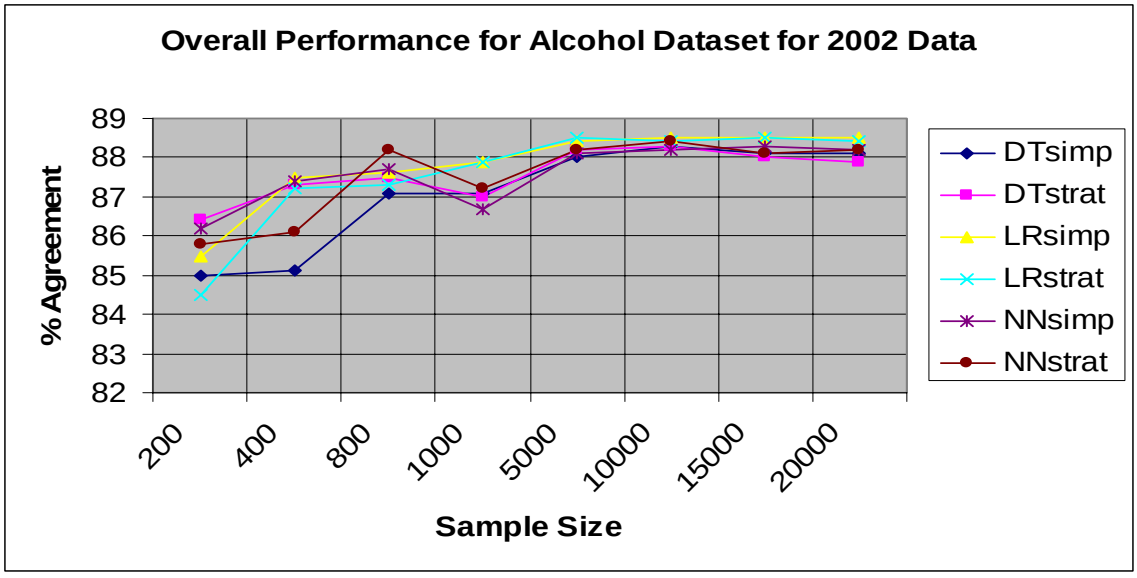

Figure 5.1 Performance graphs of all the models for year 2002 Alcohol dataset

It was also seen that the information contained in the sample rather than the sample size was responsible for the classification accuracy of a model. This was demonstrated when the sample of 400 for Alcohol dataset was reproduced thrice to make a sample of 1200 and the decision tree model was built with this sample. When the model was tested for the test datasets, it was seen that the performance was lower than the sample of 1000 and the prediction accuracy was the same as that for the training data for sample of 400.

For the Seatbelt dataset, when the overall performance is compared for all the tree models with different sample sizes and different sampling methods, it is seen that the overall classification accuracy for all the three methods were the same and varied between a very narrow range. Also, for all the methods, the sample size at which maximum classification accuracy was attained was seen to be 1000. Increasing sample sizes beyond 1000 did not help in classifying any better. This is illustrated in figure 5.2 where the classification accuracies obtained when each of the models for different sample sizes and sampling methods was tested on 2002 test dataset are plotted against the sample sizes.

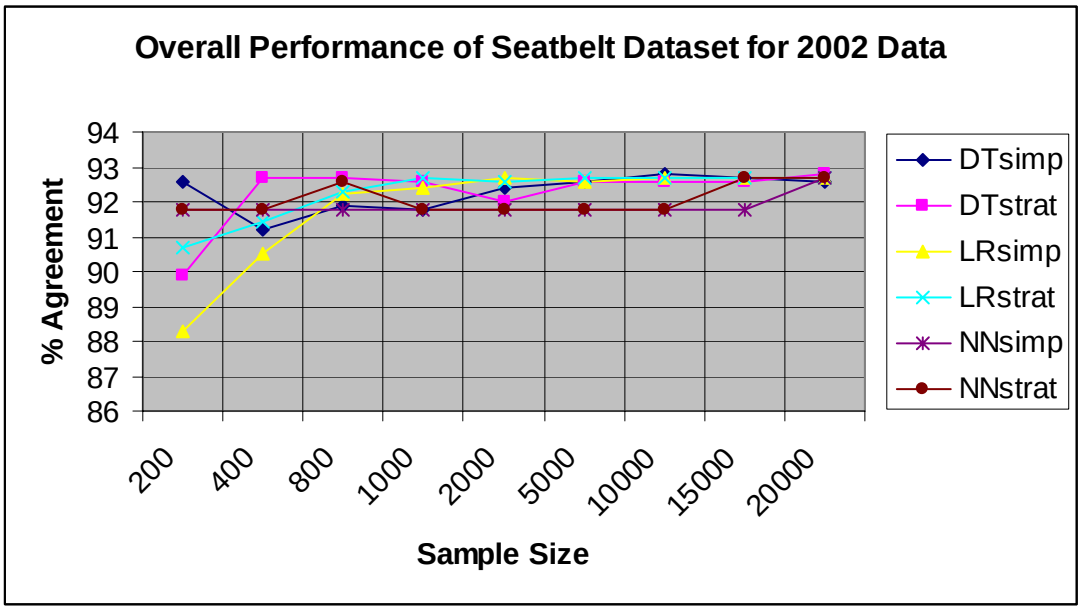

Figure 5.2 Performance graphs of all the models for year 2002 Seatbelt dataset

But, when the neural network model results were examined in details for the Seatbelt dataset, it was seen that the prediction accuracy for a "0" value of the dependent variable was 0 as compared to that for a "1" value of the dependent variable which was 100%. This could be the result of the skewed distribution of the dependent variable in the dataset.

The overall performance pattern of the models for the Fatality dataset was also very similar to that of the Seatbelt dataset, though the absolute values of the classification accuracy were much lower. For all the methods, the sample size at which stable classification accuracy was attained was seen to be 1000. Increasing sample sizes beyond 1000 did not help in improving the classification accuracy. This is illustrated in figure 5.3 where the classification accuracies obtained when each of the models for different sample sizes and sampling methods was tested on 2002 test dataset are plotted against the sample sizes. Also, as in the case of Fatality dataset, when the neural network model results were examined in details, it was seen that the prediction accuracy for a "1" value of the dependent variable was 0 as compared to that for a "0" value of the dependent variable which was 100%. This again could be the result of the skewed distribution of the dependent variable in the dataset.

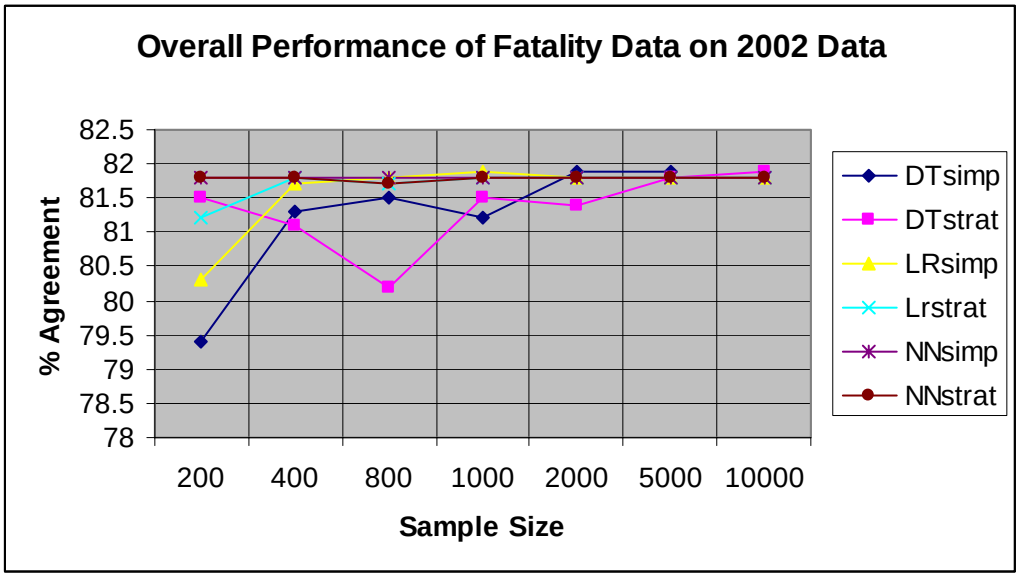

Figure 5.3 Performance graphs of all the models for year 2002 Fatality dataset

So it can be concluded from the results of this study that a very large training dataset is not required to train a decision tree or a neural network model or even for logistic regression models to obtain fairly high classification accuracy. The information content of a training dataset which affects the training process of a model and the classification accuracy is not governed by the size of the dataset. In all of the three datasets that the models were fitted to, the overall performance of the models reached a steady value at the sample size of 1000, irrespective of the total population size from which the samples were taken which was around 25,000 in case of Alcohol dataset, 27,000 for the Seatbelt dataset and 13,000 in case of Fatality dataset. This was seen to be true for all the three methods, irrespective of the data distribution or data quality. This is an important discovery, especially in the context of data mining, because data mining had evolved to deal with very large volumes of data and it takes lots of time to train data mining models, especially neural networks. So, if the classification accuracy is found to not to be dependent on the size of training dataset and a relatively small dataset of 1000 instances is optimum to train a neural network, it would mean a huge savings in terms of time and computational resources.

Moreover, the study also shows that the sampling method has not affected whatsoever, the classification accuracy of the models. But, the ration of the “1” values and “0” values of the dependent variable seems to play an important role in the individual classification accuracies of “0”s and “1”s for all the three models, especially neural networks, though it does not affect the overall accuracy as a whole. The neural network is seen to fail to predict a single “0” correctly when the ratio of “1” to “0” is very high in the training dataset and the target population. This is also proven by the experiment done where the Seatbelt and Fatality datasets are tweaked to make the distribution of 0 and 1 for the dependent variable more even. The prediction accuracy of the value with low occurrence in the population was seen to improve appreciable. But the overall prediction accuracy deteriorated and this might be attributed to the fact that the predictors were not good enough, though this could not be verified within the scope of the study. It can be generally said that, if the accuracy of prediction of one group is more important than the other

(like in the case of a financial institution in deciding whether to grant or deny credit to a customer, it may decide that it is more important to classify a customer correctly into good credit group than the bad credit customers), it would be unwise to train any data mining model, especially a neural network model if the 0/1 ratio in the training dataset and the target dataset is abnormal. Though this point has been addressed by Meshbane et al. (1996) in the context of logistic regression and predictive discriminant analysis, no one had studied before this, the effect with neural network or decision trees. The overall classification accuracy of all the three methods were very much comparable and no one method over performed any other and this was true for all the three datasets, which agrees to the results of some previous studies and contradicts some.

There are some limitations of this study. The Louisiana motor vehicles crash dataset contained a huge number of variables, out of which only a few were chosen based on common-sense, the factors that would normally be believed to affect the value of the dependent variable, in all the three datasets. A better approach would have been to select the subset of predictors that "best" explain, in a statistical sense, the dependent variable. Unless there is a prior knowledge based on either theoretical or practical grounds about the set of predictors explaining the given phenomenon, the specification process is usually an extensive amount of trial and error process with numerous subsets of predictors. At least theoretically, all possible combinations of predictors must be considered and evaluated using any of the methods discussed in literature for model selection. Since this was not possible due to resource and time constraints, it is not possible to say definitely that the results obtained were not because of a wrong choice of predictor set. The difference in the prediction accuracy values for the Seatbelt and Fatality datasets, though both the datasets apparently seemed to have similar distribution of the dependent variables, could be attributed to the relationship of the dependent variables with the independent variables. This could be a scope of future study.

BIBLIOGRAPHY

Agresti, A. (1990). Categorical Data Analysis, Wiley, New York.

Anderson, J.A. (1972). "Separate sample logic discrimination", Biometrica, 59, 19-35.

Anderson, T. W. (1984). An Introduction to Multivariate Statistical Analysis, 2nd Edition, Wiley, New York.

Asparoukhov, O. K. and Krzanowski, W.J. (2001). "A comparison of discriminant procedures for binary variables", *Computational Statistics and Data Analysis*, 38, 139-160.

Baird, L.L. (1975). "Comparative prediction of first year graduate and professional school Grades in six fields", *Educational and Psychological Measurement*, 35, 941-946.

Baron, A. E. (1991). "Misclassification among methods used for multiple group discrimination – The effects of distributional properties", *Statistics in Medicine*, 10, 757-766.

Bayne, C. K., Beacuchump, J. J., Kane, V. E. and McCabe, G. P. (1983). "Assessment of Fisher and logistic linear and quadratic discrimination models", *Computational Statistics and Data Analysis*, 1, 257-273.

Bedi, J. (1991). "Predicting graduate academic success from undergraduate academic performance: a canonical correlation study", *Educational and Psychological Measurement*, 51, 151-160.

Bellman, S., Lohse, G. L. and Johnson, E.J. (1999). "Predictors of online buying behavior", *Communication of the ACM*, 42, 32-38.

Berardi, V. L., Patuwo, B. E. and Hu, M. Y. (2004). "A principled approach for building and evaluating neural network classification models", *Decision Support Systems*, 38, 233-246.

Breiman, L (1996). "Bagging predictors", *Machine Learning*, 24, 123-140.

Bryan, J. G., (1961). *Scientific Report No. 2: Calibration of qualitative and quantitative variables for use in multiple-group discriminant analysis,* Hartford, CT: The Travelers Insurance Company.

Chatterjee, S. and Barcun, S. (1970). "A nonparametric approach to credit screening", *Journal of American Statistical Association*, 65, 150-154.

Chiang, W. K., Zhang, D. and Zhou, L. (2006). "Predicting and explaining patronage behavior toward web and traditional stored using neural networks: a comparative analysis with logistic regression", *Decision Support Systems*, 41, 514-531.

Chu, C. -H. and Widjaja, D. (1994). "Neural network system for forecasting method selection", *Decision Support Systems"*, 12, 13-24.

Cleary, P. D. and Angel, R. (1984). "The analysis of relationships involving dichotomous dependent variables", *Journal of Health and Social Behavior*, 25, 334-348.

Cover, T.M. and Hart, P.E. (1967). “Nearest neighbor pattern classification”, *IEEE Trans. Inform. Theory*, 13, 21-27.

Cox, D. R. (1966). “Some procedures connected with the logistic qualitative response curve”. In: *David, F. N. (Ed.), Research Papers in Statistics: Festschrift for J. Neyman*, Wiley, London, 55-71.

Crawley, D. R. (1979). “Logistic discriminant analysis as an alternative to Fisher’s linear discriminant function. *New Zealand Statistics*, 14, 21-25.

Davis, R. H., Edelman, D. B. and Gammerman, A. J. (1992). “Machine learning algorithms for credit card applications”, *IMA Journal of Mathematics Applied in Business and Industry*, 4, 43-51.

Day, N.E. and Kerridge, D. F. (1967). “A general maximum likelihood discriminant”, *Biometrics*, 23, 313-323.

Degeratu, A. M., Rangaswamy, A. and Wu, J. (2000). “Consumer choice behavior in online and traditional supermarkets: the effects of brand name, price and other search attributes”, *International Journal of Research in Marketing*, 17, 55-78.

Dey, E. L. and Astin, A. W. (1993). “Statistical alternatives for studying college student retention: A comparative analysis of logit, probit and linear regression”, *Research in Higher Education*, 34, 569-581.

Dierrerich, T.G., Hild. H. and Bakiri, G. (1995). “A comparison of ID3 and back propagation for English text-to-speech mapping”, *Machine Learning*, 18, 51-80.

Duarte Silva, A.P. (1995). ” Minimizing classification costs in two-group classification analysis”, *Unpublished PhD. Dissertation*, The University of Georgia.

Durand, D. (1941). *Risk Elements in Consumer Installment Financing*, New York: National Bureau of economic Research.

Eisenbeis, R. and Avery, R. (1972). “Discriminant Analysis and Classification Procedures”, Lexington Books, Lexington MA.

Fadlalla, A. and Lin, C. -H. (2001). “An analysis of the applications of neural networks in finance”, *Interfaces*, 31, 112-122.

Finch, W.H. and Schneider, M.K.(2006). “Misclassification rates for four methods of group classification”, *Educational and Psychological Measurement*, 66, 240-257.

Fisher, R. A. (1936). “The use of multiple measurements in taxonomic problems”, *Annals of Eugenics*, 7, 179-188.

Fix, E. and Hodges, J. (1952). “Discriminatory analysis, nonparametric discrimination: consistency properties”, *Report 4, Project 21-49-004*, US Air Force School of Aviation Medicine, Randolph Field.

Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley, New York, MA.

Grablowsky, B. J. and Talley, W.K. (1981). "Probit and discriminant function for classifying credit applicants: a comparison", *J. Econ. Bus.*, 33, 254-261.

Hand, D. J. (1986) "New instruments for identifying good and bad credit risks: a feasibility study", *Report*, Trustee Savings Bank, London.

Hand, D. J. and Henley, W. E. (1997). "Statistical classification methods in consumer credit scoring: A review", *Journal of the Royal Statistical Society*, 160, 523-541.

Harrell, F. E. Jr. and Lee, K. L. (1985). " A comparison of the discrimination of discriminant analysis and logistic regression under multivariate normality", *In P. K. Sen (Ed), Biostatictics: Statistics in biomedical, public health and environmental sciences*, 333-343.

Henley, W. E. and Hand, D.J. (1996). "A k-nearest-neighbor classifier for assessing consumer credit risk", *The Statistician*, 45, 77-95.

Hertz, J., Krogh, A. and Palmer, R. G. (1991). *Introduction to the Theory of Neural Computation*, Addison-Wesley, Redwood City.

Hills, M. (1967). "Discrimination and allocation with discrete data", *Applied Statistics*, 16, 237-250.

Huberty, C. J. (1994). *Applied Discriminant Analysis.* New York: John Wiley.

Hung, S., Liang, T. and Liu, V. W. (1996). "Integrating arbitrage pricing theory and artificial neural networks to support portfolio management", *Decision Support Systems*, 18, 301-316.

Jain, B. A., and Nag, B. N. (1997). "Performance evaluation of neural network decision models", *Journal of Management Information Systems*, 14, 201-216.

Johnson, F. I., Meyer, R. J. and Ghose, S. (1989). "When choice models fail: contemporary models in negatively correlated environments", *Journal of Marketing Research*, 26, 255-270.

Johnson, R.A. and Wichderm, D.W. (2002). *Applied Multivariate Statistical Analysis*, Prentice Hall, Upper Saddle River, NJ.

Joachimsthaler, E. A. and Stam. A. (1988). "Four approaches to the classification problem in discriminant analysis: an experimental study", *Decision Science*, 19, 322-333.

Joachimsthaler, E. A. and Stam. A. (1990). "Mathematical programming approach for the classification problem in two-group discriminant analysis", *Multivariate Behavioral Research*, 25, 427-454.

Johnston, B. and Seshia, S. S. (1992). "Discriminant analysis when all variables are ordered", *Statistics in Medicine*, 11, 1023-1032.

Kiang, M. (2003). "A comparative assessment of classification methods", *Decision Support Systems*, 35, 441-454.

Kim, Y. -S. and Nick Street, W. (2004). “An intelligent system for customer targeting: a data mining approach”, *Decision Support Systems*, 37, 215-228.

Knoke, J. D. (1982). “Discriminant analysis with discrete and continuous variables”, *Biometrics*, 38, 191-200.

Koehler, G. J. and Erenguc, S. S. (1990). “Minimizing misclassification in linear discriminant analysis”, *Decision Science*, 21, 63-85.

Kohonen, T. (1990). “The self-organizing map”, *Proceedings of the IEEE 78*, 1990, 1464-1480.

Krzanowski, W. J. (1975). “Discrimination and classification using both binary and continuous variables”, *Journal of the American Statistical Association*, 70, 782-790.

Kudo, M. and Sklansky, J. (2000). “Comparison of algorithms that select features for pattern classifiers”, *Pattern Recognition*, 33, 25-41.

Kwak, H., Fox, R.J. and Zinkhan, G. M. (2002). “What products can be successfully promoted and sold via the Internet?”, *Journal of Advertising Research*, 42, 23-38.

Lachenbruch, P. A. and Mickey, M. R. (1968). “Estimation of error rates in discriminant analysis”, *Technometrics*, 10, 1-11.

Levin, N., Zahavi, J. and Olitsky, M. (1995). “AMOS – A probability-driven, customer-oriented decision support system for target marketing of solo mailings”, *European Journal of Operational Research*, 87, 708-721.

Lin, M., Huang, S. and Chang, Y. (2004). “Kernel-based discriminant technique for educational placement”, *Journal of Educational and Behavioral Statistics*, 29, 219-240.

Maxwell, S.E. (1961). “Canonical variate analysis when the variates are dichotomous”, *Educational and Psychological Measurement*, 21, 259-271.

Meshbane, A. and Morris, J. D. (1996). “Predictive discriminant analysis versus logistic regression in two-group classification problems”, *American Educational Research Association annual meeting*, New York.

Myers, J. H. and Forgy, E. W. (1963). “The development of numerical credit evaluation systems”, *Journal of American Statistical Association*, 58, 799-806.

Orgler, Y. E. (1970). “A credit scoring model for commercial loans”, *Journal of Money Credit Banking*, Nov., 435-445.

Payne, J. W., Bettman, J. R. and Johnson, E. J. (1993). *The Adaptive Decision Maker*, Cambridge University Press, New York.

Press, S. J. and Wilson, S. (1978). “Choosing between logistic regression and discriminant analysis”, *Journal of the American Statistical Association*, 73, 699-705.

Remus, W. and Wong, C. (1982). “An evaluation of five models for the admission decision”, *College Student Journal*, 16, 53-59.

Rendell, L. and Cho, H. (1990). “Empirical learning as a function of concept character”, *Machine Learning*, 5, 267-298.

Riedmiller, M. (1994). “Advanced supervised learning in multi-layer perceptrons – from back propagation to adaptive learning algorithms”, *International Journal of Computer Standards and Interfaces*, 16, 265-278.

Ripley, B. (1994). “Neural networks and related methods for classification”, *Journal of the Royal Statistical. Society: Series B*, 56, 409-456.

Rosenberg, E. and Gleit, A. (1994). “Quantitative methods in credit management: a survey”, *Operations Research*, 42, 589-613.

Rubin, P. A. (1990). “Heuristic solution procedures for a mixed-integer programming discriminant model”, *Managerial Decision Econometrics*, 11, 255-266.

Scott, E. (1978). “On the financial applications of discriminant analysis: comment”, *The Journal of Financial and Quantitative Analysis*, 13, 201-210.

Shavlik, J. W., Mooney, R.J. and Towell, G.G. (1991). “Symbolic and neural learning algorithms: an experimental comparison”, *Machine Learning*, 6, 111-144.

Urban, G. L. and Hauser, J. R. (1980). *Design and Marketing of New Products, first edition*, Prentice-Hall, Englewood Cliffs, NJ.

West, P., Brockett, P. L. and Golden, L. L. (1997). “A comparative analysis of neural networks and statistical methods for predicting consumer choice”, *Marketing Science*, 16, 370-391.

Wiginton, J. C. (1980). “A note on the comparison of logit and discriminant models of consumer credit behavior”, *Journal of Financial Quantitative Analysis*, 15, 757-770.

Williams, C.J., Lee, S.S., Fisher, R.A. and Dickerman. L.H. (1999). “A comparison of statistical methods for parental screening for down syndrome”, *Applied Statistical Methods in Business and Industry*, 15, 186-195.

Wilson, R.L. and Hardgrave, B. C. (1995). “Predicting graduate student success in an MBA program: regression versus classification”, *Educational and Psychological Measurement*, 55, 186-195.

Wilson, R. L. and Sharda, R. (1994). “Bankruptcy prediction using neural networks”, *Decision Support Systems*, 11, 545-557.

Yarnold, P. R., Hart, L. A. and Soltysik, R. C. (1994). “Optimizing the classification performance of logistic regression and Fisher’s discriminant analysis”, *Educational and Psychological Measurement*, 54, 73-85.

APPENDIX: DATA DEFINITIONS

1. Alcohol Dataset Data Definitions and Classification criteria

Data Name	Data Description	Data Type	Data Values	Classes	Range
ALC_RES	alcohol involvement in crash	char(2)	0 to 94 denoting blood alcohol level of 0.00% to 0.94% in increments of 0.01, 95-test refused, 96-none given, 97-AC test performed, result unknown, 99-unknown	0, 1	0 - <=95 & = 0, 1 - >0 & < 95, 2 - > 95
DRINKING	police reported alcohol involvement	char(1)	0-alcohol not involved, 1-alcohol involved, 8-not reported, 9-unknown (police reported)	1, 2, 3	1 - 0, 2 - 1, 3 - 8,9
HOUR	hour of crash	smallint	1 through 24	1, 2, 3, 4	1 - <=4, 2 - {5,7}, 3 - {18,20}, 4 - other
DAY_WEEK	day of week of the crash	char(1)	1-Mon, 2-Tues, 3-Wed, 4-Thurs, 5-Fri, 6-Sat, 7-Sun	1, 2	1 - {5,7}, 2 - (1,4}
VE_FORMS	number of vehicles, including trains, involved in this crash; must be at least one	smallint	sum of the number of road vehicles from the VEHIC_TB table involved in this crash, plus the number of trains, from the TRAIN_TB table, involved in this crash	1, others	1, others
INJ_SEV	injury severity	char(1)	0-no injury, 1-possible injury, 2-non-incapacitating evident injury, 3-incapacitationg injury, 4-fatal injury, 5-injured, severity unknown, 6-died prior to accident, 9-unknown	1, 2, 3, 4	1 - 4, 2 - {1, 3}, 3 - 5,6,9, 4 - 0
REST_USE	restraint system used	char(2)	0-15 denoting different kinds of restraints, 99-unknown	1, 2, 3	1 - 0, 2 - 3, 3 - 1,2,4-99
AGE	age of the driver at the time of crash	smallint		1, 2, 3, 4, 5, 6	1 - <=17, 2 - [18, 20], 3 - [21, 44], 4 - [45, 64], 5 - >=65, 6 - others
BODY_TYP	vehicle body type	char(2)	0-97 denoting different vehicle body types like convertible, 2-door sedan, 3-door hatchback, 4-door sedan, minivan, truck, etc., 99-unknown	0, 1, 2, 3, 4	0 - others, 1 - {1-9}, 2 - {20 - 22, 28 - 41, 45 - 49}, 3 - {80 - 89}, 4 - {12, 24 - 25, 50 - 59}

(table cont.)

SEX	sex of driver	char(1)	1-male, 2-female, 9-unknown	1, 2, 3	1 - 1, 2 - 2, 3 - 9
VIOLCHG1	previous violations charged against the driver	char(2)	0-98 denoting different types of violations, 99-unknown	0, 1	0 - others, 1 - {11 - 16, 18 - 19, 01 - 09, 99}
M_HARM	most harmful event	char(2)	1-49 denoting different kinds of harmful events, like overturn, fire/explosion, immersion, gas inhalation, fall from vehicle, injured in vehicle, other non-collision, pedestrian, etc., 99-unknown	0, 1	0 - others, 1 - {01, 18, 23 - 24, 26 - 38, 42 - 43, 99}

2. Seatbelt Dataset Data Definitions and Classification criteria

Data Name	Data Description	Data Type	Data Values	Classes	Range
SEVERITY_CD	most severe injury in crash	char(1)	A-fatal, B-incapacitating/severe, C-non-incapacitating/moderate, D-possible/complaint, E-no injury	0, 1, 2	0 = E, 1 = A/B, 2 = C/D
NUM_VEH	number of vehicles, including trains, involved in this crash; must be at least one	smallint	sum of the number of road vehicles from the VEHIC_TB table involved in this crash, plus the number of trains, from the TRAIN_TB table, involved in this crash	1 , others	1, others
DAMAGE_EXT1_CD	code for extent of damage to vehicle at first impact area	char(1)	A-none, B-very minor, C-minor, D-moderate, E-moderate/severe, G-severe, H-very severe, I-unknown, -not reported	0, 1, 2	0 = A/B/C, 1 = D/E/G/H, 2 = I/
DR_A_D_PRES_CD	code for presence of alcohol and/or drugs for driver	char(1)	A-neither alcohol or drugs present, B-yes(alcohol present), C-yes(drugs present), D-yes(alcohol and drugs present), E-not reported, F-unknown	0, 1, 2	0 = A, 1 = B/C/D, 2 = E/F
DR_AGE	driver's age at time of crash	smallint		1,2,3,4,5, 6	1=(<=17), 2=(18,20), 3=(21,44), 4=(45,64), 5=(>=65), 6=others

(table cont.)

DR_AIRBAG_CD	code for airbag usage	char(1)	A-deployed, B-not deployed, C-not deployed/switched off, D-not applicable, E-unknown	0,1,2	0=B/C, 1=A, 2=D/E
EST_ALCOHOL	alcohol involvement in the crash	char(1)	N - no, Y - yes	0,1	0 = N, 1 = Y
DR_EJEC_CD	code for ejection of driver	char(1)	A-not ejected, B-totally ejected, C-partially ejected, D-unknown	0,1,2	0=A, 1=B/C, 2=D
DR_INJ_CD	code for injury to driver	char(1)	A-fatal, B-incapacitating/severe, C-non-incapacitating/moderate, D-possible/complaint, E-no injury	0,1,2	0=E, 1=A/B, 2=C/D
DR_PROTSYS_CD	code for driver protection system used	char(1)	A-none used, B-shoulder belt used only, C-lap belt used only, D-shoulder and lap belt used, E-child safety seat improperly used, F-child safety seat used, G-helmets used, H-restraint use unknown	0,1,2	0=A,1=B/C/D, 2=E/F/G/H
DR_RACE	race of driver	char(1)	W-white, B-black, I-Indian, O-other, -not specified	W,B,I,O	W,B,I,O
DR_SEX	sex of driver	char(1)	M-male, F-female	M,F	M,F
VEH_TYPE_CD	code type for vehicle	char(1)	A-passenger car, B-light truck/pickup, C-van, D - A, B or C/trailer, E-motorcycle, F-pedal cycle, G-off-road vehicle, H-emergency vehicle, I-school bus, J-other bus, K-motor home, L-single unit truck, M-truck with trailers, N-farm equipment, O-other, -not reported	0,1,2	0=A/C/D, 1=B, 2=others

3. Fatality Dataset Data Definitions and Classification criteria

Data Name	Data Description	Data Type	Data Values	Classes	Range
SEVERITY_CD	most severe injury in crash	char(1)	A-fatal, B-incapacitating/severe, C-non-incapacitating/moderate, D-possible/complaint, E-no injury	0, 1	0 = others, 1 = A/B

(table cont.)

NUM_VEH	number of vehicles, including trains, involved in this crash; must be at least one	smallint	sum of the number of road vehicles from the VEHIC_TB table involved in this crash, plus the number of trains, from the TRAIN_TB table, involved in this crash	1 , 2	1 = 1, 2 = others
MAN_COLL_CD	code for the manner of collision of the crash	char(1)	A-non-coll w/motor veh, B-rear end, C-head on, D-right angle, E-left turn-angle, F-left turn-opp direction, G-left turn-same direction, H-right turn-angle, I-right turn-opp direction, J-sidesw-same direction, K-sidesw-opp direction, L-other	0, 1, 2	0 = A/B/L, 1 = C/D/E/F/G, 2 = H/I/J/K
TRK_BUS_INV	at least one Uniform Truck Bus supplement completed	char(1)	N - no, Y - yes	0, 1	0 = N, 1 = Y
EST_ALCOHOL	alcohol involvement in the crash	char(1)	N - no, Y - yes	0, 1	0 = N, 1 = Y
AGGRESSIVE	aggressive driving	char(1)	0-non-aggressive, 2-careless operation, 3-failure to yield, 4-disregarded traffic control, 5-following too closely, 6-over safe speed limit, 7-over stated speed limit, 8-cut-in/improper pass	0, 1	0 = 0, 1 = others
VIOLATION	vehicle violation at the time of crash	char(1)	A-exceeding stated speed limit, B-exceeding safe speed limit, C-failure to yield, D-following too closely, E-driving left of center, F-cutting in/improper passing, G-failure to signal, H-made wide right turn, I-cut corner on left turn, J-turned from wrong lane, K-other improper turning, L-disregarded traffic control, M-improper starting, N-improper parking, O-failed to set out flags/flares, P-failed to dim headlights, Q-vehicle condition, R-driver condition, S-careless operation, T-unknown violation, U-no violation, V-other	0,1	0 = Q,R,T,U,V, 1 = others
NUM_OCC	number of occupants including driver in the vehicle	smallint	sum of the count of number of occupants in the vehicle plus 1 (for driver)	1, other	1 = 1, 2 = others

(table cont.)

NUM_OCC_NO_SEATB	number of occupant of the vehicle with no seatbelt	smallint		1, other	1 = 1, 2 = others

VITA

Rochana Lahiri studied electrical engineering at Jadavpur University at Calcutta, India, from 1987 to 1991, and earned a Bachelor of Engineering degree in 1991. She entered the energy industry with Durgapur Projects Limited, Durgapur, India, as an electrical engineer and worked there from 1993 to 1996. Later on she changed from power industry to information technology industry where she worked on different operating systems like mainframe, Windows and UNIX. Before she moved to the United States, she had been working for HCL Perot Systems, NOIDA, in India from 1998 to 2004.

In 2004, Ms Lahiri entered the master's program in the Department of Information Systems and Decision Sciences at the E. J. Ourso College of Business at Louisiana State University, Louisiana. She earned the degree of Master of Science in Information Systems & Decision Sciences in December, 2006.