Predicting Risk of Re-hospitalization for Congestive Heart Failure Patients

Jayshree Agarwal

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Masters of Science

University of Washington

2012

Reading Committee:

Dr. Senjuti Basu Roy, Chair

Dr. Ankur Teredesai

Program Authorized to Offer Degree:
Institute of Technology - Tacoma

University of Washington

Abstract

Predicting Risk of Re-hospitalization for
Congestive Heart Failure Patients

Jayshree Agarwal

Chair of the Supervisory Committee:
Assistant Professor Dr. Senjuti Basu Roy
Institute of Technology

Congestive Heart Failure (CHF) is one of the leading causes of hospitalization, and studies show that many of these admissions are readmissions. Identifying patients who are at a greater risk of hospitalization, can guide implementation of appropriate plans to prevent these readmissions. In the field of medical sciences, prediction of such outcomes is a challenging task since it involves integration of various variables associated with patients, such as patients' socio-demographic factors, health conditions, health care utilization and factors related to health care providers. This work aims at analyzing the problem and building an effective predictive model to identify patients who are at a greater risk of future hospital admissions. We propose several classification algorithms to that end. The precursory step to the actual model building process is the information extraction phase; this step seems to be prohibitively challenging due to the prevalence of noise in the dataset, heterogeneity and diverse nature of the sources, and sparsity to name a few. Our initial results are encouraging, as we significantly outperform the existing predictive model proposed by the researchers at Yale University. Our solutions are empirically evaluated by using a health care data set provided by Multicare Health System (MHS).

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I am grateful to many people without whose help and contributions this work could not have been completed. First and foremost I would like to express my deepest gratitude to my advisor and committee chair Dr. Senjuti Basu Roy who has continuously supported, encouraged and guided me with her knowledge, ideas and suggestions. Without her persistent help, this work would not have been possible. My appreciation extends to Dr. Ankur Teredesai who has always been there for me and guided me through this journey with his encouragement and invaluable advice. I would also like to thank David K Hazel, Si-Chi Chin, Kiyana Zolfaghar and Mehrdad Rohani for their continuous support and invaluable inputs.

I would also like to express my deepest appreciation to Dr. Lester Reed, senior vice president of Multicare Health System and Dr. Paul Amoroso for providing us with this wonderful opportunity and the required financial support. I express my gratitude to data architects at Multicare, Yoshi Williams and Eric Johnson for their invaluable inputs and help in enhancing my domain knowledge.

I am indebted to my parents and family for their love and support and making me the person I am today. Lastly but not the least, I am deeply thankful to my husband Amit Agarwal who has always been there for me and without whom this journey would not have been possible.

DEDICATION

To my parents and my husband Amit

Chapter 1

INTRODUCTION

The hospital readmission rate has progressively increased over the past few years especially for heart failure patients. This work is focused on developing predictive models that calculate the probability of a given patient discharged with Congestive Heart Failure (CHF) being readmitted to the hospital. In this introductory chapter we discuss the impact of unplanned hospitalizations and illustrate the need for predicting the risk of readmission for heart failure patients. We also provide the formal definition of the problem we are trying to solve, highlight the challenges faced in the process, and summarize our contributions.

1.1 Motivation

Congestive Heart Failure (CHF) is one of the leading causes of hospitalization, especially for adults older than 65 years of age [1, 19]. The prevalence and incidence of CHF have considerably increased over the past few years [19]. Studies show that many of these hospitalizations are readmissions and CHF is one of the primary reasons behind multiple hospitalizations within a short time-span [16, 24, 6]. All cause 30 day readmission rate for patients with CHF has increased by 11% between 1992 and 2001 [14]. These readmissions act as a substantial contributor to the rising health care costs[10]. Based on the 2005 data of Medicare beneficiaries, it has been estimated that 12.5% of Medicare admission due to CHF were followed by readmission within 15 days, accounting for about $590 million in healthcare costs[14]. Furthermore, the readmission rate is used as a screening tool for monitoring the quality of service and efficiency of care provided by the health care providers [4]. Center for Medicare and Medicaid Services (CMS) has started using the 30 day all cause HF readmission rate as a publicly reported efficiency measure.

Readmission rates are attributable to the quality of care provided by the hospitals [3]. Readmission can result from a variety of reasons, including early discharge of patients,

improper discharge planning and poor care transitions. Studies have shown that targeted interventions before or after discharge can reduce the probability of getting readmitted, especially in elderly patients, and decrease the overall medical costs [22, 21, 7]. Proper pre-discharge planning [23] and post discharge plans like home based follow up [17] and patient education [15, 12], can considerably reduce the readmission rates and improve the health outcome of patients. Therefore, many such readmissions can be prevented if the interventions are fully established and the type and quality of care provided to patients is improved.

Identifying patients at high risk of readmission can guide health care providers to develop programs to improve the quality of care and institute targeted interventions, thus reducing the readmission rate and the cost incurred in these re-hospitalizations. This can also facilitate proper resource utilization by the hospitals.

While actionable insights [13, 11] could be gathered by analyzing and mining the data generated by health-care industries, there are several non-trivial challenges involved in the process. First and foremost, the generated data is *complex* and *extremely voluminous*. Next, one has to understand several *factors* that lead to multiple readmissions. Finally, it is important to devise *solution techniques* that are effective and robust in handling the domain specific complex data. While the overall problem has been identified as an extremely important problem in the health-care domain, unfortunately, not many solutions are known [14, 18, 20, 2] at present that seem to be effective. To that end, this thesis formally studies the risk-of-readmission prediction problem in the context of CHF.

1.2 Problem Definition

Formally, we study the problem of building a predictive model that can identify the patients with CHF who are most likely to get readmitted. To achieve the overall task, we study the following two sub-problems:

1. Identify the factors that influence readmission of patients discharged with CHF.

2. Build a predictive model such that for a given patient discharged with CHF, the model is able to provide the likelihood of the patient getting readmitted within 30 days or

60 days of discharge.

1.3 Challenges

We highlight the challenges involved in solving the aforementioned problem. The first and foremost challenge is to interpret and identify the pertinent *factors* (attributes or variables) and data values present in the high dimensional medical dataset, that attributes to readmission of CHF patients. This task is challenging as the data is closely coupled with a medical domain. The next challenge is the task of *data cleaning and preparation*. On one hand, the real world medical dataset is noisy, inconsistent, and may consist of significant amount of missing values; on the other hand, an effective and accurate predictive model assumes the presence of balanced and clean dataset. Even before the actual model building tasks, we need to preprocess the data effectively and make it suitable for predictive modeling. Finally, one has to have an in-depth understanding of different predictive modeling techniques, so that effective algorithmic solutions could be devised to solve the prediction task. Note that, as we discuss later on, many such prediction techniques are required the use of labeled dataset; thus, we further preprocess the data and augment it with appropriate class labels.

1.4 Contribution

Our primary contributions could be summarized as follows:

- We study an important and ongoing problem in health-care domain in collaboration with Multicare Health-care System (MHS), a leading health care provider in the state of Washington. We formalize the problem of predicting risk of readmission of the CHF patients. We study the attributes and factors pertinent to cause of readmission for CHF patients.

- We propose effective predictive modeling techniques and devise novel solutions to solve the problem.

- We perform an extensive experimental study using a complex real world dataset (provided by MHS) that demonstrate the effectiveness of our proposed method. Our

proposed method outperforms the existing model proposed by the researchers at Yale University.

The rest of the thesis is organized as follows: in chapter 2, we introduce the various techniques that we employ towards solving this problem and the evaluation metrics that are used. Chapter 3 discusses the basic approaches and methods employed to address the different challenges mentioned earlier. In chapter 4, we explain the experimental setup along with the dataset description, and describe at length the experiments performed and the attribute sets used in the experiments. Chapter 5 summarizes the experimental results. Finally in chapter 6, we outline future work and conclude the thesis.

Chapter 2

TECHNICAL BACKGROUND

There are various data mining techniques we can leverage on for solving the given problem. The complex non linear relationships among the variables related to hospital readmission for CHF patients, can be modeled by using both supervised and unsupervised learning techniques. In the former the model is constructed by analysing a training set made up of database tuples and their associated class labels. This is in contrast to unsupervised learning technique where the information regarding the class labels is not known. Classification is one of the supervised learning techniques where the model is trained with the help of labeled dataset and the performance of the model is evaluated on test set which is independent of the training set. If the performance is satisfactory then it is used to classify new data. In order to address the problem of predicting incidence of hospital readmission we will be exploiting the supervised classification technique. It is possible to apply this technique because the information regarding the readmission of patients is available and the dataset can be augmented with the appropriate class label information. There are various classification methods proposed by researchers. We are choosing three such techniques which are some of the most popular and powerful techniques used to model and predict medical outcomes. In the next section we describe the chosen techniques and we also discuss the evaluation metrics used for assessing the accuracy of the predictive model.

2.1 Logistic Regression

Logistic regression is one of the most popular techniques used for predicting dichotomous variable. It is widely used to solve binary classification problem but the application of logistic regression is also extended to the cases where the dependent variable is multinomial in nature. In logistic regression the probability of an outcome is modeled as a function of

independent predictor variables using the following logistic function

$$log[\frac{p}{1-p}] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i. \tag{2.1}$$

where p is the probability of the outcome of interest, β_0 is the intercept term, $\beta_1 ... \beta_i$ are the β coefficients associated with each variables, $X_1 ... X_i$ are the predictor variables, and i is the unique subscript denoting each variable [9]. Logistic regression model uses maximum likelihood estimation for calculating the coefficients. It is an iterative process that converges when the coefficient values, that maximizes the likelihood function is obtained [9]. In order to infer the importance of the predictor variables, the β coefficients can easily he converted into the corresponding odds ratios by raising e to the power of coefficient. For example, if β_i is 0.75 we obtain the odd ratio by raising exponent constant to the power of β_i which is approximately 2.12. This means that the probability of our outcome variable being 1 is 2.12 times as the value of X_i is increased by 1 unit. Once the coefficients are learnt by the model the probability of the outcome of new data can be obtained by the following equation

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_i X_i)}} \tag{2.2}$$

2.2 Naive Bayesian Classifier

Bayesian classifier is a statistical classifier that predicts the class membership probability, in other words the probability of a given tuple belonging to a particular class. This classifier assumes that the effect of one attribute value on a given class is independent of the values of the other attributes. This assumption is called as *class conditional independence* which greatly simplifies the learning process [8]. The basic principle of this classifier is the *Bayes theorem.*

2.2.1 Bayes Theorem

Let $\mathbf{X}$ be data tuple that is described by measurements made on a set of n attributes. Let H be the hypothesis that the data tuple belongs to a specified class C. $P(H|\mathbf{X})$ is the *posterior probability*, the probability that the hypothesis H holds given the value of tuple as $\mathbf{X}$. $P(H)$ is the *prior probability* of H, the probability of hypothesis H being true independent of

the value of $\mathbf{X}$. Similarly $P(\mathbf{X}|H)$ is the *posterior probability* of $\mathbf{X}$ conditioned on H and $P(\mathbf{X})$ is the *prior probability* of $\mathbf{X}$. According to Bayes Theorem the posterior probability, $P(H|\mathbf{X})$ can be calculated from the following equation:

$$P(H|\mathbf{X}) = \frac{P(\mathbf{X}|H)P(H)}{P(\mathbf{X})} \tag{2.3}$$

2.2.2 Classification

Let the number of classes be m, $C_1, C_2, \ldots, C_m$ and $\mathbf{X} = (x_1, x_2, , x_n)$ be the n dimensional attribute vector for which the prediction has to be done. The naive Bayesian classifier predicts that $\mathbf{X}$ belongs to the class having the highest posterior probability. In other words the classifier predicts that the tuple $\mathbf{X}$ belongs to the class C_i only if the below condition is satisfied.

$P(C_i|\mathbf{X}) > P(C_j|\mathbf{X})$ for $1 <= j <= m, j \neq i$ [8]

The class C_i for which $P(C_i|\mathbf{X})$ is maximized is called the *maximum posteriori hypothesis* [8]. From Bayes Theorem:

$$P(C_i|\mathbf{X}) = \frac{P((X)|C_i)P(C_i)}{P(\mathbf{X})} \tag{2.4}$$

Since $P(\mathbf{X})$ is identical for all the classes so it can be ignored, so only $P((X)|C_i)P(C_i)$ needs to be maximized. For high dimensional data the estimation of $P((X)|C_i)$ from the given set of training tuples is computationally expensive. To make the computation easy, the naive assumption of class-conditional independence is made which presumes that the attribute values are conditionally independent of one another [8]. Hence,

$$P(\mathbf{X}|C_i) = \prod_{k=1}^{n} P(x_k|C_i) \tag{2.5}$$

With this assumption naive Bayes classifier simplifies the learning process. In various domains, the performance of naive Bayes classifier is comparable to other sophisticated classifiers like decision tree and neural network classifiers.

2.3 Support Vector Machine

Support vector machine is a method for classification of both linear and nonlinear data. It is based on the following idea. It maps the original training data into a higher dimensional

feature space, using nonlinear mapping and finds an optimal separating hyperplane that separates the two classes [8]. SVM searches for a hyperplane that has the maximum distance to the closest points in the training set termed as *support vectors*. This plane is also called as *maximum marginal hyperplane* (MMH) which gives the maximum separation between the classes [8]. Let the data set D consists of set of points $\mathbf{X_i}$ where $i = 1, 2, \ldots, N$ and each point is associated with two class identified by label $y_i \in \{+1, -1\}$. If we consider the data belonging to two classes are linearly separable then the separating hyperplane (MMH) can be written as

$$\mathbf{W}.\mathbf{X} + b = 0 \tag{2.6}$$

where $\mathbf{W}$ is the weight vector, $\mathbf{W} = w_1, w_2, \ldots, w_n$, n is the number of attributes and b is a scalar [8]. The MMH can be rewritten as the decision boundary

$$d(\mathbf{X^T}) = \sum_{i=1}^{l} y_i \alpha_i \mathbf{X_i} \mathbf{X^T} + b_0 \tag{2.7}$$

where y_i is the class label of the support vector $\mathbf{X_i}$, $\mathbf{X^T}$ is the test tuple, α_i and b_0 are numeric parameters determined by solving the quadratic optimization problem and l is the number of support vectors [8]. The classification of the test tuple is performed by computing the sign of the right hand side in the above equation. If the data is not linearly separable then each input point $\mathbf{X}$ is mapped to another point $\mathbf{Z} = \phi(\mathbf{X})$ of a higher dimensional space [8]. The decision hyperplane in the new space is represented as

$$d(\mathbf{Z}) = \mathbf{W}.\mathbf{Z} + b \tag{2.8}$$

In order to reduce the computation complexity the dot product of two points $\mathbf{X_i}$ and $\mathbf{X_j}$, $\phi(\mathbf{X_i}).\phi(\mathbf{X_j})$ is replaced by a kernel function $K(\mathbf{X_i}, \mathbf{X_j})$. With this replacement the equation can be rewritten as

$$d(\mathbf{X^T}) = \sum_{i=1}^{l} y_i \alpha_i K(\mathbf{X_i} \mathbf{X^T}) + b_0 \tag{2.9}$$

where y_i is the class label of the support vector $\mathbf{X_i}$, $\mathbf{X^T}$ is the test tuple, α_i and b_0 are solution of the quadratic programming problem and l is the number of support vectors. There are different kernel functions possible. It can be linear, polynomial, gaussian radial

basis function or sigmoid. SVM always finds a global solution and is less prone to over fitting [8].

2.4 Model Evaluation

Once a classification model is built we need to assess how accurate our classifier is in predicting the class label of tuples. The model is trained on training dataset and the accuracy of the learned model is measured on a test set consisting of class labeled tuples that were not used in training the model. There are various metrics for evaluating the predictive accuracy of the classifier and different methodologies are employed to calculate these metrics.

2.4.1 Evaluation Metrics

1. Accuracy

   Accuracy of a classifier on a given test set is determined be the percentage of test set tuples that are correctly classified by the classifier [8].

$$Accuracy = \frac{No.ofTruePositive + No.ofTrueNegative}{No.ofTestTuples} \tag{2.10}$$

2. Precision

   Precision can be considered as a measure of exactness which gives the percentage of tuples that are predicted as positive are actually such [8]. A precision of 1 means that all the tuples that are predicted by the classifier as positive belong to the positive class but it does not tell us how many positive tuples are mislabeled by the classifier.

$$Precision = \frac{No.ofTruePositive}{No.ofTruePositive + No.ofFalsePositive} \tag{2.11}$$

3. Recall

   Recall is a measure of completeness which gives the percentage of positive tuples that are labeled as such [8]. A recall score of 1 means that every tuple belonging to the positive class is labeled as positive but it does not tell us how many tuples were

incorrectly labeled as positive.

$$Recall = \frac{No.ofTruePositive}{No.ofTruePositive + No.ofFalseNegative} \tag{2.12}$$

4. F-Measure

   There is an inverse relationship between precision and recall where it is possible to increase one at the expense of reducing the other. F measure combines them into a single measure. It is the harmonic mean of precision and recall.

$$F = \frac{2 * precision * recall}{precision + recall} \tag{2.13}$$

5. Receiver operating characteristic curve

   This visual tool is used for comparing the performance of two or more classification models. The ROC curve for a given model shows the trade-off between the true positive rate and false positive rate [8]. True positive rate is the proportion of positive tuples that are correctly labeled by the model and false positive rate is the proportion of negative tuples that are incorrectly labeled as positive by the model. The area under the ROC curve (AUC) is a measure of the accuracy of the model. The closer the area is to 0.5, less accurate the model is. AUC of 0.5 indicates the model performance is no better than by chance. AUC greater than 0.5 indicates that the model has more discriminative ability than just random guessing. A model with perfect accuracy will have an area of 1.0.

2.4.2 k-fold Cross Validation

This technique is widely used for obtaining reliable classifier accuracy because of its low bias and variance. While performing k-fold cross validation the experimental data set is divided equally into k subsets. Out of them, k-1 subsets are used as training data set and the rest is used as test data set. This is repeated k times allowing each tuple to be used for training the same number of times and once for testing [8]. Finally, in order to evaluate the predictive model efficiency the average of the results from the k iterations is used.

Chapter 3

APPROACH

We formulate the problem of predicting the risk of readmission for CHF patients as a binary classification problem, and leverage various available classification methodologies. Critical factors influencing early recurrent admissions are identified through the help of domain experts and review of related studies. These factors are used as predictor variables in the model. As we are dealing with a real world heterogeneous and noisy dataset provided by MHS, the data needs to be thoroughly analysed and integrated before the actual modeling task can be performed. Several experiments are conducted to evaluate the performance of the chosen classification methods in predicting the likelihood of readmission for CHF patients. Figure 3.1 describes the non-trivial steps involved in building the predictive model. It consists of 4 major phases: data understanding, data preprocessing, modeling, and evaluation. The next few sections discuss each of them in detail and summarize the non-trivial challenges encountered in the process of solving the problem.

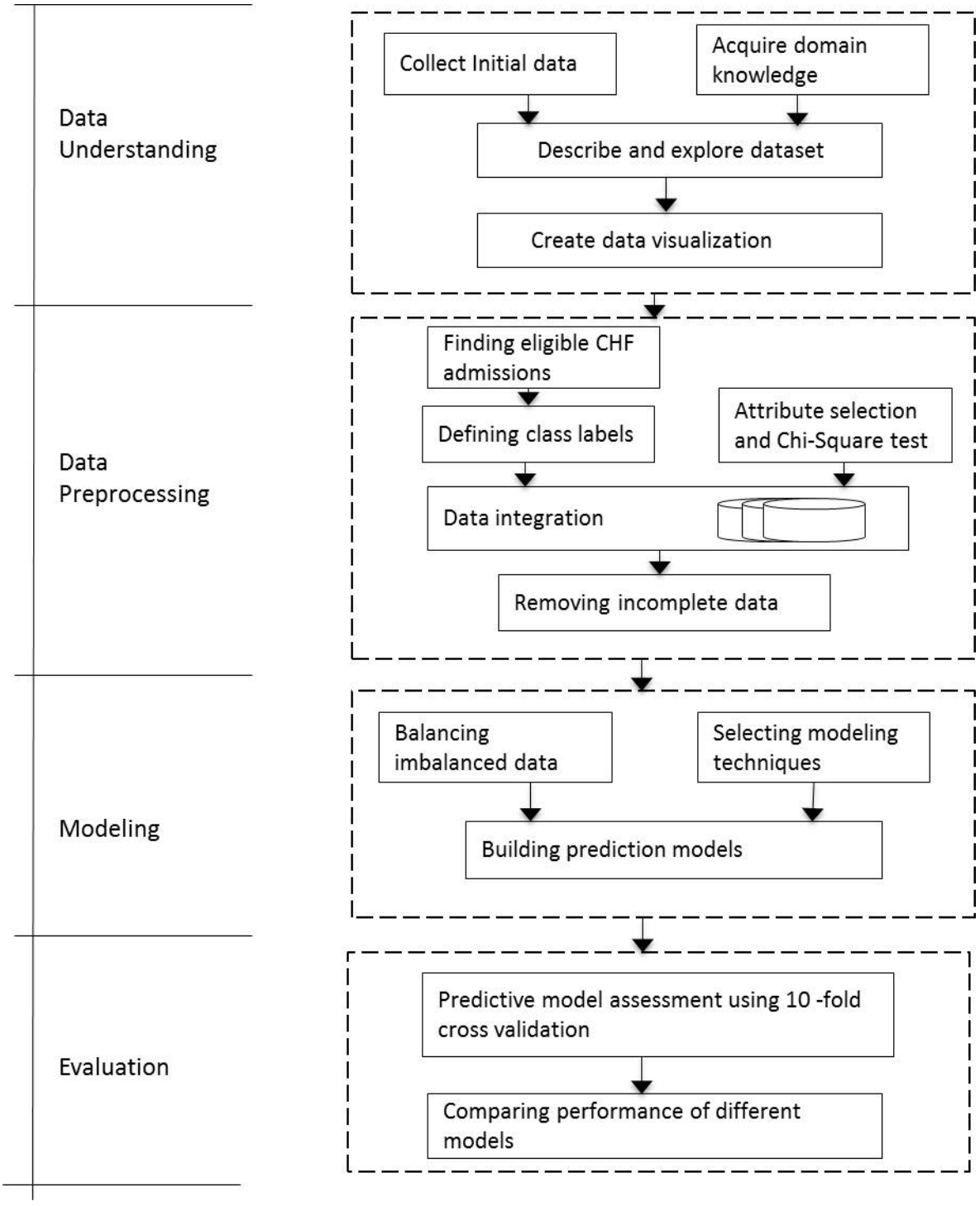

Figure 3.1: The different steps involved in the process

3.1 Data Understanding

This phase involves exploring the raw data in order to gain initial insights and discover interesting actionable patterns. Getting familiar with the data is extremely important because knowledge of the data is useful in data preprocessing, a prerequisite for building a predictive model. Real world clinical data is noisy and heterogeneous in nature, and in order to get the data ready for modeling, it is necessary to closely examine the attributes and their values. Gaining insights into the data, such as the type of attributes present, the kind of data values for each attribute and their distribution, help with subsequent analysis. With the help of valuable inputs from the experts in MHS, we acquired the requisite domain knowledge to explore the dataset. The dataset has information on patient socio-demographic characteristics, marital status, ethnicity, diagnosis, discharge disposition and other cost related factors pertaining to a particular hospital admission. Detailed description of the attributes is present in the next chapter. In addition various data visualizations are also created which help in viewing data through graphical means. These visualizations have been very useful in understanding the correlation between the attributes and their relation with readmission. It further helped in identifying outliers and inconsistencies in the data.

3.2 Data Preprocessing

Once we have a good understanding of the data we need to prepare it for modeling. Data preprocessing is a precursory step to the actual modeling and helps in constructing the final homogeneous data set suitable for training predictive models. This phase poses several challenges due to the presence of heterogeneity in the data and prevalence of missing values and inconsistencies. The various steps involved in completing this task are elaborated in the subsequent sections.

3.2.1 Finding eligible CHF admissions

Admissions of patients with discharge diagnosis of CHF are identified as the potential index CHF hospitalization to be used for training. In this work we only consider patients with a discharge diagnosis of the International Classification of Diseases, 9th Revision, Clinical

Modification Codes (ICD-9 CM) related to CHF, listed in Table 3.1, as either primary or secondary diagnosis. Our entity of observation is each CHF hospitalization. Once all the admissions related to CHF have been identified, we exclude the hospitalizations encountered with in-hospital deaths in our analysis because we are more concerned about predicting readmissions. We also consider all inter hospital transfers as readmissions.

Table 3.1: The ICD-9 CM codes for CHF

ICD-9 CM codes	Description
402.01	Malignant hypertensive heart disease with heart failure
402.11	Benign hypertensive heart disease with heart failure
402.91	Unspecified hypertensive heart disease with heart failure
404.01	Malignant hypertensive heart and kidney disease with heart failure and with chronic kidney disease stage I through stage IV, or unspecified
404.03	Malignant hypertensive heart and kidney disease with heart failure and chronic kidney disease stage V or end stage renal disease
404.11	Benign hypertensive heart and kidney disease with heart failure and with chronic kidney disease stage I through stage IV, or unspecified
404.13	Benign hypertensive heart and kidney disease with heart failure and chronic kidney disease stage V or end stage renal disease
404.91	Unspecified hypertensive heart and kidney disease with heart failure and with chronic kidney disease stage I through stage IV, or unspecified
404.93	Unspecified hypertensive heart and kidney disease with heart failure and chronic kidney disease stage V or end stage renal disease
428.XX	Heart Failure codes

3.2.2 Class label definition

Once we have identified the CHF hospitalizations to be used for training, the next step is to define class labels for these instances which is one of the foremost requirements of classification process.

The outcomes that we are trying to predict here are 30-day readmissions and 60-day readmissions. 30-day or 60-day readmission is defined as presence of at least 1 hospitalization within 30 or 60 days (respectively) of discharge after an index hospitalization due to CHF. We chose a 30 day timeframe because it is a clinically meaningful timeframe for hospitals and medical communities to take action to reduce the probability of readmission [14]. Further, 30 day readmission rate is also used by CMS as a potential efficiency measure of hospitals [14]. The readmissions considered here are all-cause readmission. This is because from a patients point of view any kind of hospitalization is a key concern [14]. Secondly, from a financial point of view all readmissions irrespective of the cause, incur cost. Finally, even if a patient is readmitted for ailments unrelated to CHF, it may still be strongly correlated with the quality of care received by the patient in the previous admission due to CHF. The class label is defined as a dichotomous variable which is evaluated based on whether a patient is readmitted within 30 days or 60 days of discharge. It takes one of two values, 0 or 1, which corresponds to no readmission and occurrence of readmission respectively. Each eligible CHF admission is assigned a readmit or no readmit label based on the presence of next admission within 30 days or 60 days from the discharge date. We train our classifier models to predict the test samples and evaluate the performance based on the prediction.

3.2.3 Attribute selection

Hospital readmission is a complex phenomenon governed by multiple variables because there is no universal factor that can be used to predict the risk of readmission. One of our major challenges is to determine the subset of factors that has a significant impact on readmission of patients from the myriad of attributes present in the data set. Based on input from domain experts and literature review, we assume that the primary factors related to readmission are patients' socio-demographic variables, health parameters, discharge disposition and other

cost related factors like length of stay. Socio-demographic variables include patient's age, gender, marital status and ethnicity. Other medical conditions in addition to CHF, such as severity of illness, ejection fraction value and blood pressure are some of the health parameters taken into consideration. Discharge destination of patients tells us whether a patient is taken care of by a home healthcare worker or extended nursing facility after discharge. In our research we did not find much existing literature or previous work on the given problem. We came across a previous work of researchers at Yale University who tried to solve a similar problem by developing predictive models [14]. In our initial set of experiments, we have considered the set of variables that they included in building their predictive model. We have also conducted a comprehensive set of experiments with other additional factors. Further, we employed the *chi-square* test to discover correlation between the aforementioned variables and the outcome we are trying to predict.

3.2.4 Data integration and removal of missing values

After identifying the predictor variables, it is necessary to combine them into a homogeneous form amenable to modeling. All the predictor variables are spread across multiple tables with primary and foreign key relationships. Using SQL queries, all the selected predictor variables are retrieved from different tables and combined into a single coherent data set which is used for modeling. The dataset provided to us is a real world, noisy clinical data with high prevalence of missing values, outliers and incomplete data. It is necessary to clean this dataset and resolve the inconsistencies because dirty data can result in unreliable output. From the unified dataset obtained, missing values and outliers are removed so that the data is suitable for training the predictive models.

3.3 Modeling

This phase consists of the actual model building task which involves selecting and applying various data mining prediction techniques. We leverage various classification techniques to build a model that can predict if a patient discharged with diagnosis of CHF will get readmitted within 30 days and 60 days of discharge. A comprehensive set of experiments are performed with several classification models, namely logistic regression, naive Bayes

classifier and support vector machine with RBF kernel, and the performance of these models is compared to determine their efficacy for the given dataset. These classification techniques are selected because they are some of the most popular and powerful techniques for solving binary classification problems.

It is observed that the number of patients not getting readmitted highly outnumbers the patients getting readmitted, which makes the dataset very imbalanced. This poses a challenge, since most classifiers are designed to optimize accuracy irrespective of the class distribution, and a classifier built with such high skewed class distribution would inevitably predict the majority class far more frequently than the alternative class. They tend to give higher weightage to the majority class and ignore the less frequently occurring class. This is not an acceptable situation for our problem because our class of interest is the minority class, consisting of patients who have higher chances of getting readmitted. In order to improve the classification accuracy of the class-imbalanced data we use various sampling techniques, including oversampling and undersampling. These techniques alter the class distribution of the training data so that both the classes are well represented. Oversampling works by resampling the rare class tuples so that the resulting training set has equal number of tuples for each class [8]. On the other hand, undersampling works by decreasing the number of tuples belonging to the majority class by randomly eliminating tuples until they are equal to the other class [8]. Among other evaluation metrics, we also consider recall and AUC, as they provide insight into the performance of classifiers on skewed dataset.

3.4 Evaluation

The models developed are evaluated in order to assess the quality of prediction of various classification models. In this phase, we determine the effectiveness of our models with the help of various evaluation metrics mentioned in Section 2.4.1. In our experimental set up, some part of the data set is used for training the model and the remaining is utilized for testing. We use 10-fold cross validation method for evaluation. This is a recommended validation method due to its relatively low bias and variance. The metrics - precision, recall, F measure and AUC are the criteria used to compare the performance of different models, since use of these metrics is common in applications with skewed data sets and

where prediction of one class is substantially important than the other. Once we establish that the model's accuracy is acceptable, it can be used to predict the risk of readmission for new patients.

Chapter 4

EXPERIMENTS

We conduct a comprehensive set of experiments using the data set obtained from MHS. In this section we provide a description of the data set, along with the different attributes chosen and the various experiments performed

All our experiments are conducted using the framework provided by MHS. We have used R-studio and R version 2.15.1 for developing the models. The database used is SQL server 2008.

4.1 Dataset description

The dataset used to develop our readmission prediction model is provided by the Multicare Health System. The dataset consists of CHF hospitalization for patients discharged since 2009. It provides information of 8975 patients who had been admitted with a diagnosis of CHF and the number of discharges generated by these patients during 2009-2012 is 17136. Out of these 3073 admissions have been due to CHF as the primary diagnosis and remaining 14063 admissions had CHF as secondary diagnosis. There are 1077 in-hospital deaths that are not included in building the model. The data provided is in a relational database form. It is spread across multiple base relations with primary and foreign key relationships defined among them. Various supporting views are provided to get a complete understandig of the patients related to heart failure. There are tables that provide the detailed information of a patient's history and related socio-demographic factors. The key attributes of these tables are age, gender, race, marital status and education. There are other relation tables consisting of set of attributes, describing the complete primary and secondary diagnosis (ICD-9 codes) of the patients, ejection fraction value, blood pressure and other clinical data pertaining to their treatment. Information about the discharge disposition of the patients, including the medication prescribed at the time of discharge and follow-up plan can also be

obtained from these tables. Further administrative information of patients like the length of stay, mode of payment, and insurance plan can also be acquired from these tables. After removing the tuples with incomplete information the remaining number of instances on which the model is built is 4481 for predicting the risk of readmission for 30 days and 4182 for 60 days. Ejection fraction value is the attribute with the highest number of missing values. The percentage of missing data points for ejection fraction value is approximately 70%. Out of the instances used for modeling, 27% of the patients are readmitted within 30 days and 38% of patients are readmitted within 60 days. This shows that the dataset is highly imbalanced because the number of patients who encounters readmissions is far less than the patients not re-hospitalized.

4.2 Attribute sets

We generated four different sets of features and conducted our experiments using all the feature sets. The first set of features which we call as **Baseline** consists of the predictor variables used by researchers in Yale University who have tried to solve the similar problem [14]. The variables are shown in Table 4.1. It consists of two demographic variables namely age and gender and other variables related to the diagnosis information derived from the primary and secondary diagnosis of the patient. In order to group the diagnosis ICD-9 CM codes into variables, the CMS hierarchical condition categories (CCs) are used [14]. These variables are binary in nature and a CC is specified as present if it is coded in any of the primary or secondary diagnosis of the considered index admissions. Another feature set named as **New** consisting of other socio-demographic, clinical and administrative data is generated based on literature review and inputs from domain experts. List of all the features present in this set is found in Table 4.2. The third set **All** consists of all the attributes present in the previous two sets.

The fourth set **Correlated** comprises of the attributes that strongly imply hospital admission within 30 days and 60 days based on the available data. These features are selected by conducting chi-square test and using the significance level of 0.05 and are shown in Table 4.2 and Table 4.4. Significant clinical predictors include blood pressure, ejection fraction value, the diagnosis information. Number of secondary diagnosis appears to be a

significant discriminator. Other clinical variables like length of stay, severity of illness and risk of mortality contribute to the predictive power of the model. Discharge destination and status are also positively correlated with readmission. Of the demographic variables age, sex are not found have influence on readmission.

We develop our models using all the different predictor variable sets and compare the performance of these models.

Table 4.1: The features in attribute set *Baseline*

Variable	CC codes	Type	Mean/No. of Domain Values
Age		Numerical	70.98
Gender		Categorical	2 (M,F)
Congestive heart failure	CC 80	Categorical	2 (0,1)
Acute coronary syndrome	CC 81, 82	Categorical	2 (0,1)
Arrhythmias	CC 92,93	Categorical	2 (0,1)
Cardio-respiratory failure and shock	CC 79	Categorical	2 (0,1)
Valvular and rheumatic heart disease	CC 86	Categorical	2 (0,1)
Vascular or circulatory disease	CC 104-106	Categorical	2 (0,1)
Chronic atherosclerosis	CC 83,84	Categorical	2 (0,1)
Other and unspecified heart disease	CC 94	Categorical	2 (0,1)
Hemiplegia, paraplegia, paralysis, functional disability	CC 67-69,100-102,177,178	Categorical	2 (0,1)
Stroke	CC 95,96	Categorical	2 (0,1)
Renal failure	CC 131	Categorical	2 (0,1)
COPD	CC 108	Categorical	2 (0,1)
Diabetes and DM complications	CC 15-20,119,120	Categorical	2 (0,1)
Disorders of fluid/electrolyte/acid-base	CC 22,23	Categorical	2 (0,1)

continued on the next page

Variable	CC codes	Type	Mean/No. of Domain Values
Other urinary tract disorders	CC 136	Categorical	2 (0,1)
Decubitus ulcer or chronic skin ulcer	CC 148,149	Categorical	2 (0,1)
Other gastrointestinal disorders	CC 36	Categorical	2 (0,1)
Peptic ulcer, hemorrhage, other specified gastrointestinal disorders	CC 34	Categorical	2 (0,1)
Severe hematological disorders	CC 44	Categorical	2 (0,1)
Nephritis	CC 132	Categorical	2 (0,1)
Dementia and senility	CC 49,50	Categorical	2 (0,1)
Metastatic cancer and acute leukemia	CC 7	Categorical	2 (0,1)
Cancer	CC 8-12	Categorical	2 (0,1)
Liver and biliary disease	CC 25-30	Categorical	2 (0,1)
End-stage renal disease or dialysis	CC 129,130	Categorical	2 (0,1)
Asthma	CC 110	Categorical	2 (0,1)
Iron deficiency and other/unspecified anemias and blood disease	CC 47	Categorical	2 (0,1)
Pneumonia	CC 111-113	Categorical	2 (0,1)
Drug/alcohol abuse/dependence/psychosis	CC 51-53	Categorical	2 (0,1)
Major pysch disorders	CC 54-56	Categorical	2 (0,1)
Depression	CC 58	Categorical	2 (0,1)
Other psychiatric disorders	CC 60	Categorical	2 (0,1)
Fibrosis of lung and other chronic lung disorders	CC 109	Categorical	2 (0,1)
Protein-calorie malnutrition	CC 21	Categorical	2 (0,1)

Table 4.2: The features in attribute set *New*

Variable	CC codes	Type	Mean/No. of Domain Values
Marital status	A patient's marriage status	Categorical	9
Ethnic group	Patient's ethnic background	Categorical	9
Discharge follow-up category	Description of the different follow up days	Categorical	7
Avg. Blood Pressure Category	Blood pressure category of the average of the last 3 blood pressure taken just prior to Discharge	Categorical	9
IsHFPrimary	Flag indicates whether Heart Failure is Primary or secondary diagnosis	Categorical	2 (Y,N)
Secondary Diagnosis Count	No of Secondary diagnosis	Numerical	18
Admit source	Category value corresponding to the admission source	Categorical	6
Admit type	Category value corresponding to the admission type for patient	Categorical	4
Discharge destination	Category value corresponding to the discharge destination for patient	Categorical	70
Discharge status	Category value corresponding to discharge disposition for patient	Categorical	15

continued on the next page

Variable	CC codes	Type	Mean/No. of Domain Values
Discharge APRDRG Severity of illness	Severity of illness- 4(most severe), 1(least severe)	Categorical	4
Discharge APRDRG Risk of mortality	Risk of mortality- 4(highest risk), 1(lowest risk)	Categorical	4
Length of stay	Number of days between admit date and discharge date	Numerical	5.175
Ejection fraction value	Ejection fraction value measured	Numerical	55.26

Table 4.3: The list of correlated attributes to readmission for 30 days timeframe

Discharge Follow-up Category	Disorders of fluid/electrolyte/acid-base
Avg. Blood Pressure Category	Decubitus ulcer or chronic skin ulcer
Secondary ICD9Diagnosis Count	Peptic ulcer, hemorrhage, other specified gastrointestinal disorders
Discharge Destination	Severe hematological disorders
Discharge Status	Nephritis
Discharge APRDRG Severity of illness	Dementia and senility
Discharge APRDRG Risk of Mortality	Metastatic cancer and acute leukemia
Marital Status	Cancer
Ejection Fraction Value	Liver and biliary disease
Length of Stay	End-stage renal disease or dialysis
Acute coronary syndrome	Iron deficiency and other/unspecified anemias
Cardio-respiratory failure and shock	Pneumonia
Vascular or circulatory disease	Drug/alcohol abuse/dependence/psychosis
Chronic atherosclerosis	Depression
Hemiplegia, paraplegia, paralysis, functional disability	Fibrosis of lung and other chronic lung disorders
Stroke	Protein-calorie malnutrition
Renal Failure	Diabetes and DM complications
COPD	

Table 4.4: The list of correlated attributes to readmission for 60 days timeframe

Discharge Follow-up Category	Hemiplegia, paraplegia, paralysis, functional disability
Avg. Blood Pressure Category	Renal Failure
IsHFPrimary	COPD
Secondary ICD9Diagnosis Count	Diabetes and DM complications
Admit Source	Disorders of fluid/electrolyte/acid-base
Admit Type	Decubitus ulcer or chronic skin ulcer
Discharge Destination	Peptic ulcer, hemorrhage, other specified gastrointestinal disorders
Discharge Status	Severe hematological disorders
Discharge APRDRG Severity of illness	Nephritis
Discharge APRDRG Risk of Mortality	Dementia and senility
Length of Stay	Cancer
Ethnic Group	End-stage renal disease or dialysis
Ejection Fraction Value	Iron deficiency and other/unspecified anemias
Acute coronary syndrome	Depression
Arrhythmias	Fibrosis of lung and other chronic lung disorders
Cardio-respiratory failure and shock	Protein-calorie malnutrition
Vascular or circulatory disease	Chronic atherosclerosis

4.3 Experiments

We perform a comprehensive set of experiments with different attribute sets and various class balancing techniques. Both over sampling and under sampling techniques are employed to achieve class balance. The list of all the experiments are listed in Table 4.5. These same set of experiments are conducted for each classification method and for both 30 days and 60 days time frame.

Table 4.5: The list of experiments performed

Attribute Set	Sampling Technique	Experiment Name
Baseline	No Sampling	BaselineNoBalancing
	Undersampling	BaselineUnder
	Oversampling	BaselineOver
New	No Sampling	NewNoBalancing
	Undersampling	NewUnder
	Oversampling	NewOver
All	No Sampling	AllNoBalancing
	Undersampling	AllUnder
	Oversampling	AllOver
Correlated	No Sampling	CorrelatedNoBalancing
	Undersampling	CorrelatedUnder
	Oversampling	CorrelatedOver

Chapter 5

RESULTS

In this section we describe the experimental results for each individual classifier. In particular, we discuss how different classifiers perform with different attribute sets and class balancing techniques. We compare various empirically obtained evaluation metrics such as precision, recall, F measure, accuracy and AUC, corresponding to classifiers for each attribute set and sampling technique, and discuss the general trends observed in the empirical results.

5.1 Logistic Regression

Let us first examine the performance of the logistic regression classification technique with different attribute sets and sampling techniques. All of the experiments mentioned in Section 4.5 for predicting the risk of readmission within 30 days and 60 days of discharge, are conducted. We employ the 10-fold cross validation method for evaluating performance of the model by comparing the actual and the predicted probability of readmission of patients. The results for 30 days can be seen in Table 5.1 and the results for 60 days are shown in Table 5.2. As can be seen from Figures 5.1 and 5.3 we observe that the AUC of the model increases considerably when using attributes derived from the Chi square test as compared to the *baseline* attributes. This illustrates that the predictive power of the model is enhanced by using predictor variables with high statistical significance and high correlation with the likelihood of readmission within 30 days and 60 days. These results also indicate that the class balancing techniques do not significantly help in improving AUC. On the other hand, Figures 5.2 and 5.4 show that in absence of class balancing, the recall is very low. This indicates that the model classifies most of the test tuples in the negative class which corresponds to the patients who are not likely to get readmitted within the given time frame, because of the presence of high skewness in the original dataset. Use of class balancing

techniques results in a significant increase in the recall value of the models, in turn enabling the predictive model to correctly classify more patients with a higher likelihood of being readmitted. However, these results do not indicate either of the sampling techniques; viz. oversampling and undersampling, to be significantly better than the other, with both having comparable performance. Logistic regression did a better job in predicting the readmission risk for 30 days interval as compared to 60 days of discharge. This can be concluded from the area under the ROC curve obtained in both the cases.

Table 5.1: The 10-fold cross validation results of Logistic regression for 30 days

Experiment	Precision	Recall	F-measure	Accuracy	AUC
BaselineNoBalancing	0.5517	0.0045	0.0090	0.7676	0.5936
BaselineUnder	0.2888	0.5441	0.3773	0.5823	0.5920
BaselineOver	0.2902	0.5385	0.3771	0.5862	0.5931
NewNoBalancing	0.7782	0.1216	0.2103	0.7789	0.6213
NewUnder	0.2935	0.6211	0.3986	0.5461	0.6153
NewOver	0.3018	0.6044	0.4026	0.5655	0.6291
AllNoBalancing	0.6405	0.1386	0.2279	0.7725	0.6365
AllUnder	0.3072	0.6123	0.4092	0.5717	0.6252
AllOver	0.3182	0.5802	0.4110	0.5973	0.6325
CorrelatedNoBalancing	0.7269	0.1486	0.2468	0.7533	0.6410
CorrelatedUnder	0.3624	0.5385	0.4333	0.6170	0.6343
CorrelatedOver	0.3608	0.5033	0.4203	0.6225	0.6353

Table 5.2: The 10-fold cross validation results of Logistic regression for 60 days

Experiment	Precision	Recall	F-measure	Accuracy	AUC
BaselineNoBalancing	0.5241	0.0344	0.0646	0.6803	0.5889
BaselineUnder	0.3834	0.5380	0.4477	0.5743	0.5890
BaselineOver	0.3868	0.5374	0.4498	0.5783	0.5890
NewNoBalancing	0.6050	0.1649	0.2591	0.6416	0.5834
NewUnder	0.4367	0.4883	0.4611	0.5660	0.5793
NewOver	0.4518	0.4990	0.4742	0.5794	0.5926
AllNoBalancing	0.5689	0.2517	0.3490	0.6430	0.6203
AllUnder	0.4593	0.5393	0.4960	0.5835	0.6137
AllOver	0.4766	0.5311	0.5023	0.6	0.6220
CorrelatedNoBalancing	0.5668	0.2455	0.3427	0.6421	0.6260
CorrelatedUnder	0.4706	0.5459	0.5055	0.5942	0.6258
CorrelatedOver	0.4744	0.5321	0.5016	0.5983	0.6271

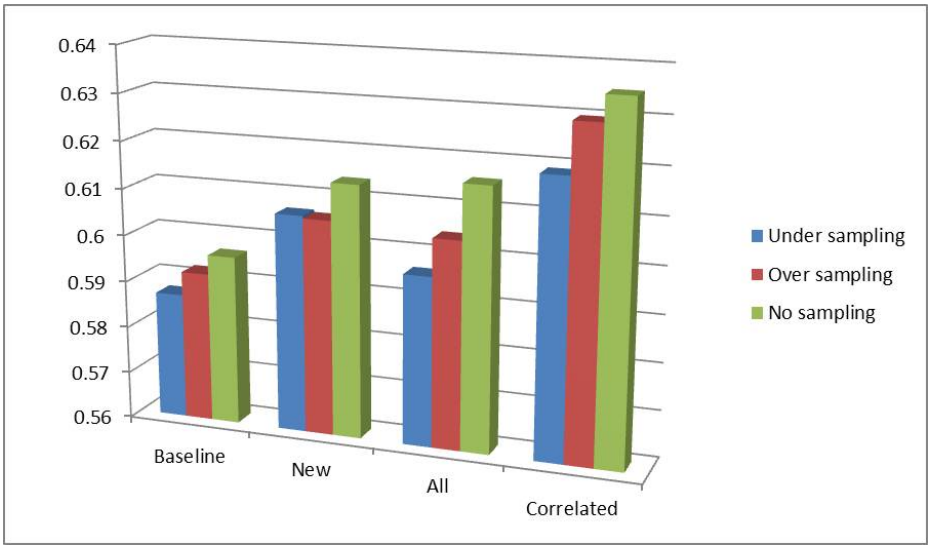

Figure 5.1: The AUC of logistic regression for 30 days

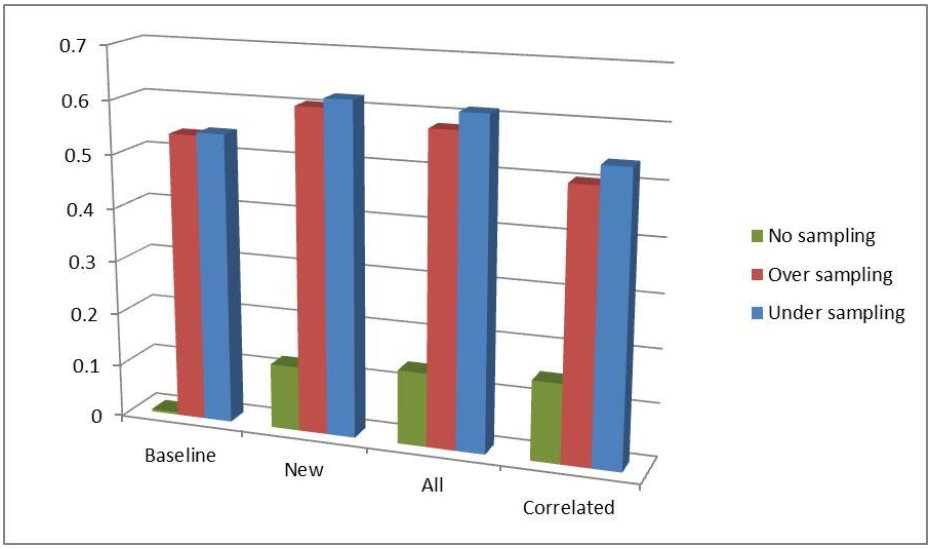

Figure 5.2: The Recall of logistic regression for 30 days

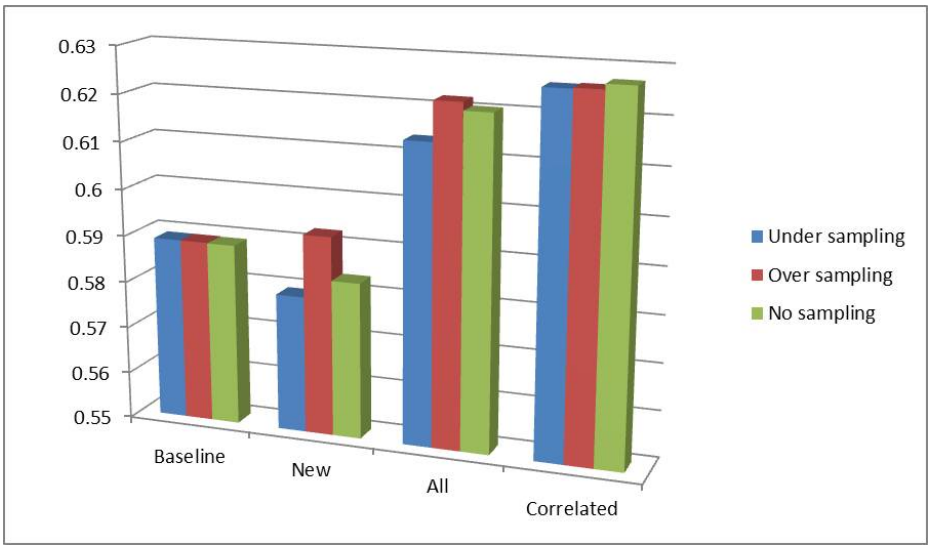

Figure 5.3: The AUC of logistic regression for 60 days

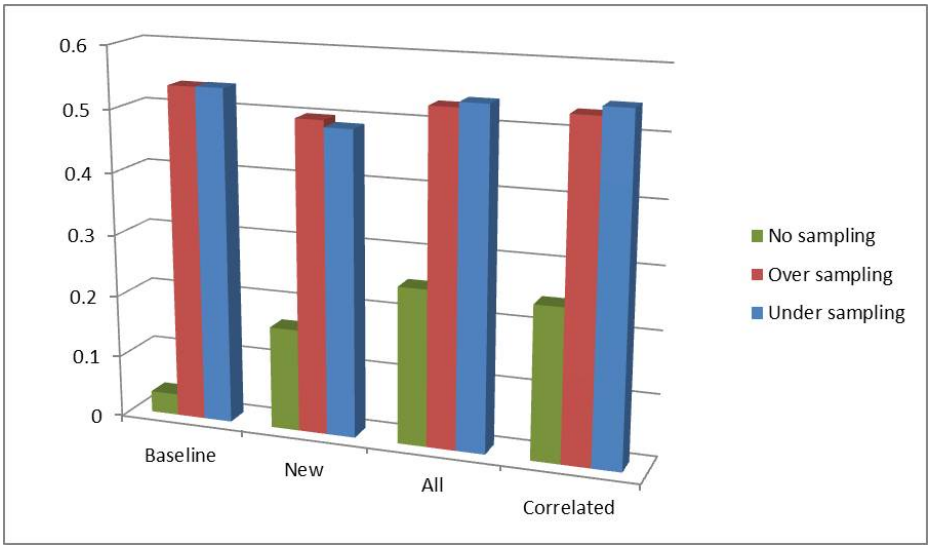

Figure 5.4: The Recall of logistic regression for 60 days

5.2 Naive Bayes Classifier

The set of experiments performed for the logistic regression classification technique are repeated with the Naive Bayes classifier and results shown in Table 5.3 for 30 days and Table 5.4 for 60 days. We do not find any statistically significant difference in discrimination between logistic regression and naive Bayes classifier because the results obtained for both the classifiers are quite similar. Firstly, the AUC is highest when modeling with the attribute set *correlated*; 64% compared to a value of 60% when using the baseline attributes. As with logistic regression, the recall improves with class balancing as can be seen in Figures 5.6 and 5.8. The recall is very high when oversampling is employed for predicting the risk of readmission of patients within 60 days of discharge. In general, no class balancing results in high precision value but low recall value.

Table 5.3: The 10-fold cross validation results of Naive Bayes for 30 days

Experiment	Precision	Recall	F-measure	Accuracy	AUC
BaselineNoBalancing	0.4489	0.2163	0.2923	0.7293	0.6054
BaselineUnder	0.3358	0.5332	0.4114	0.5902	0.5990
BaselineOver	0.3358	0.4600	0.3858	0.6121	0.5836
NewNoBalancing	0.4169	0.3295	0.3229	0.6853	0.6427
NewUnder	0.3473	0.3786	0.1991	0.6789	0.6075
NewOver	0.3325	0.3582	0.2922	0.6034	0.5613
AllNoBalancing	0.4299	0.3308	0.3229	0.6900	0.6387
AllUnder	0.3646	0.4028	0.3172	0.6165	0.6263
AllOver	0.3447	0.4205	0.3577	0.6060	0.5877
CorrelatedNoBalancing	0.4038	0.3282	0.3228	0.6807	0.6467
CorrelatedUnder	0.3611	0.3912	0.3172	0.6177	0.6303
CorrelatedOver	0.3471	0.4427	0.3577	0.6065	0.5998

Table 5.4: The 10-fold cross validation results of Naive Bayes for 60 days

Experiment	Precision	Recall	F-measure	Accuracy	AUC
BaselineNoBalancing	0.4899	0.2902	0.3629	0.6115	0.5908
BaselineUnder	0.4467	0.5326	0.4859	0.5678	0.5880
BaselineOver	0.4457	0.5224	0.4806	0.5619	0.5794
NewNoBalancing	0.4910	0.4916	0.4224	0.5590	0.5947
NewUnder	0.4910	0.4916	0.4224	0.5590	0.5947
NewOver	0.4130	0.7967	0.5379	0.4727	0.5913
AllNoBalancing	0.4734	0.5558	0.4826	0.5703	0.6170
AllUnder	0.4977	0.4953	0.4501	0.5901	0.6161
AllOver	0.4244	0.7684	0.5338	0.4902	0.6147
CorrelatedNoBalancing	0.4705	0.5577	0.4826	0.5688	0.6165
CorrelatedUnder	0.4999	0.4986	0.4501	0.5895	0.6166
CorrelatedOver	0.4184	0.7734	0.5338	0.4860	0.6143

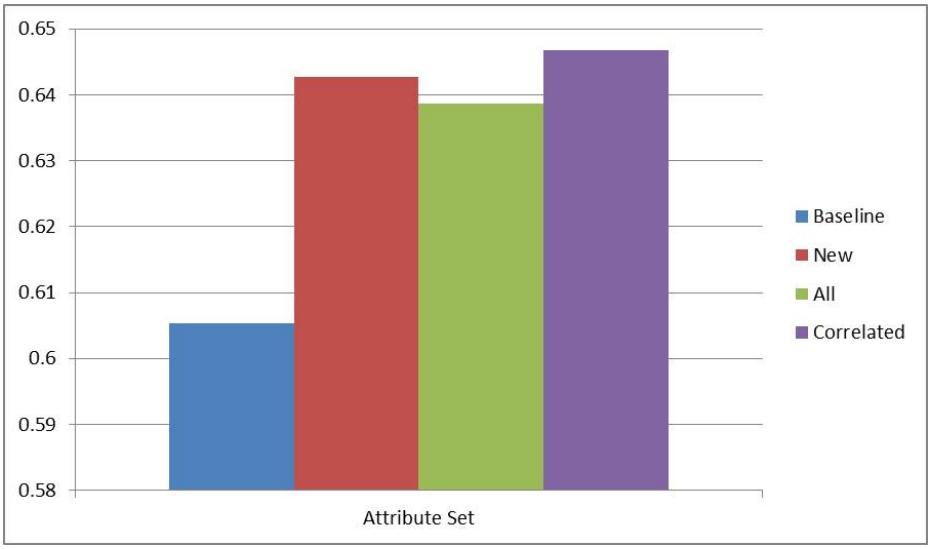

Figure 5.5: The AUC of Naive Bayes classifier for 30 days

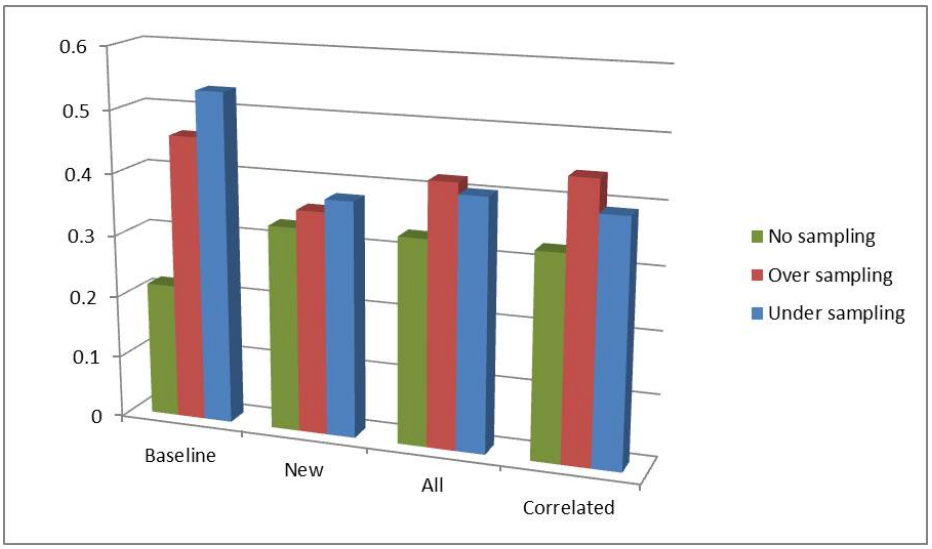

Figure 5.6: The Recall of Naive Bayes classifier for 30 days

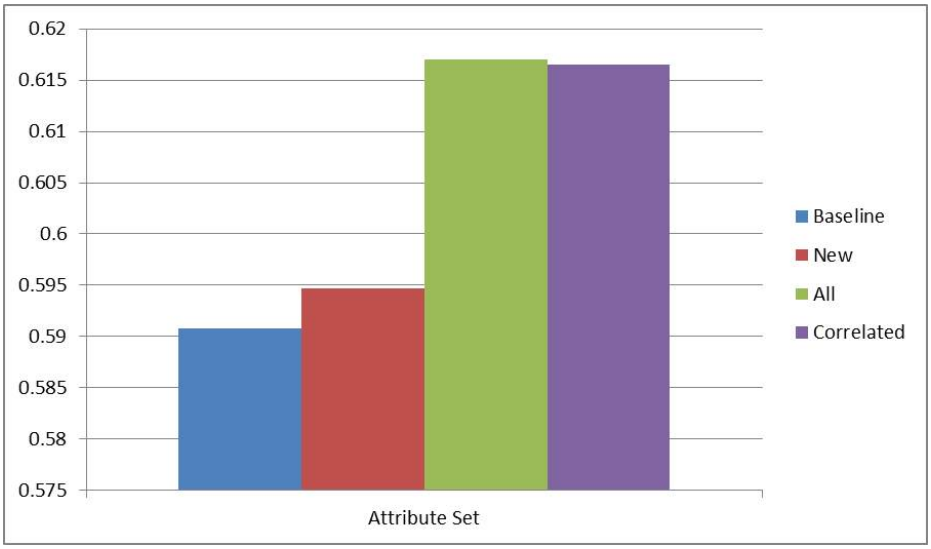

Figure 5.7: The AUC of Naive Bayes classifier for 60 days

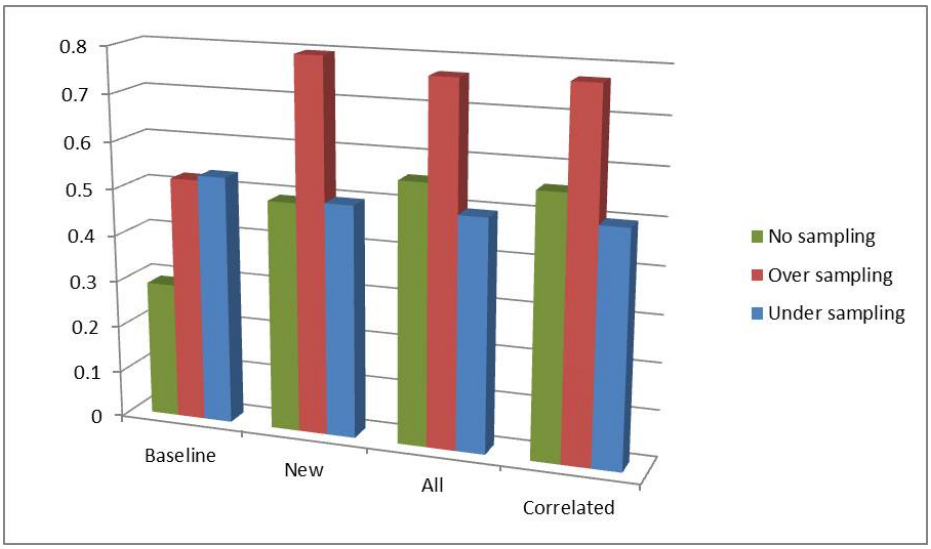

Figure 5.8: The Recall of Naive Bayes classifier for 60 days

5.3 Support Vector Machine (SVM)

Similar set of experiments are also conducted for SVM with radial basis function (RBF) kernel. In general, it is seen that the performance of the model is improved when *correlated* attribute set is used as compared to *baseline* attribute set. Like the other models this also gives better results when used for predicting the risk of readmission for 30 days as compared to 60 days. Class balancing techniques significantly improves the recall measure thus increasing the ability of the model to accurately classify patients who are at a higher risk of getting readmitted.

Table 5.5: The 10-fold cross validation results of Support Vector Machine for 30 days

Experiment	Precision	Recall	F-measure	Accuracy	AUC
BaselineNoBalancing	0.25	0.0016	0.0032	0.7272	0.5607
BaselineUnder	0.3289	0.5509	0.412	0.5723	0.5754
BaselineOver	0.3366	0.4540	0.3866	0.6082	0.5849
NewNoBalancing	0.7546	0.1338	0.2273	0.7526	0.5854
NewUnder	0.3411	0.5238	0.4132	0.5955	0.6056
NewOver	0.3411	0.5238	0.4132	0.5955	0.6056
AllNoBalancing	0.7692	0.1395	0.2362	0.7546	0.6281
AllUnder	0.3636	0.5566	0.4377	0.6111	0.6357
AllOver	0.3849	0.4515	0.4155	0.6546	0.6323
CorrelatedNoBalancing	0.7623	0.1395	0.2359	0.7542	0.6224
CorrelatedUnder	0.3558	0.5524	0.4329	0.6064	0.6358
CorrelatedOver	0.3929	0.4597	0.4237	0.6600	0.6290

Table 5.6: The 10-fold cross validation results of Support Vector Machine for 60 days

Experiment	Precision	Recall	F-measure	Accuracy	AUC
BaselineNoBalancing	0.5750	0.0988	0.1686	0.6296	0.5837
BaselineUnder	0.4382	0.5582	0.4910	0.5600	0.5907
BaselineOver	0.4503	0.4940	0.5784	0.4711	0.5818
NewNoBalancing	0.6410	0.1629	0.2599	0.6471	0.5817
NewUnder	0.4487	0.5261	0.4843	0.5741	0.5896
NewOver	0.4506	0.4682	0.4592	0.5809	0.5831
AllNoBalancing	0.6697	0.2309	0.3435	0.6644	0.6225
AllUnder	0.4751	0.5765	0.5209	0.5969	0.6267
AllOver	0.4896	0.4909	0.4902	0.6119	0.6234
CorrelatedNoBalancing	0.6353	0.2083	0.3137	0.6535	0.6159
CorrelatedUnder	0.4699	0.5764	0.5178	0.5918	0.6272
CorrelatedOver	0.4896	0.4505	0.4661	0.6076	0.6172

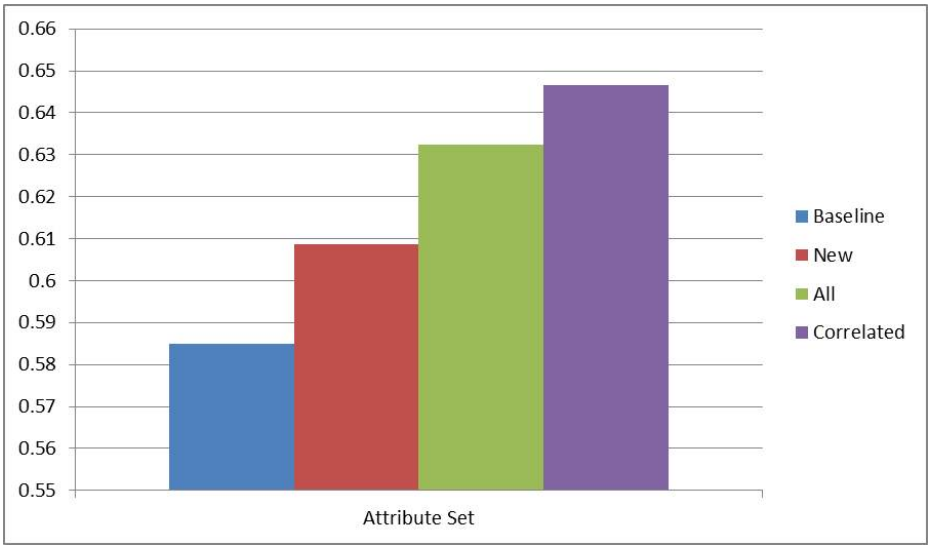

Figure 5.9: The AUC of Support Vector Machine for 30 days

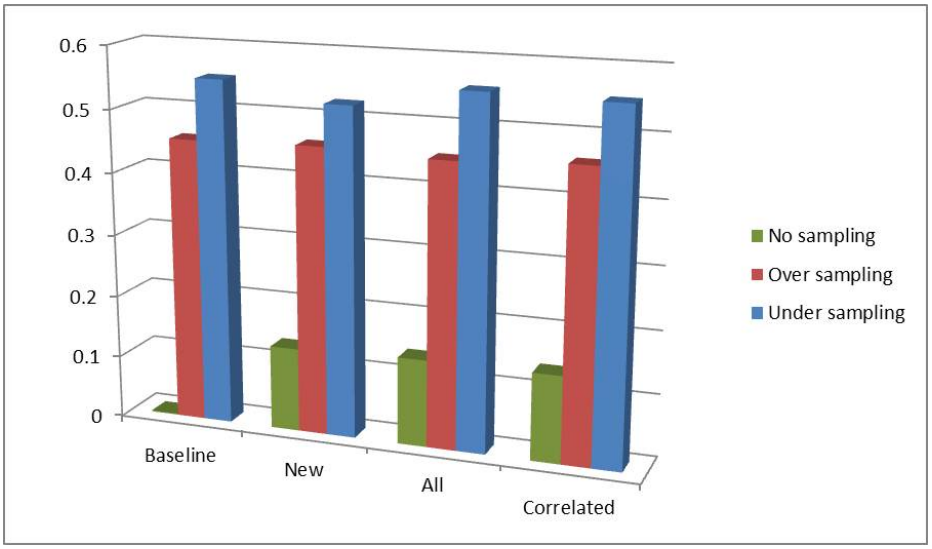

Figure 5.10: The Recall of Support Vector Machine for 30 days

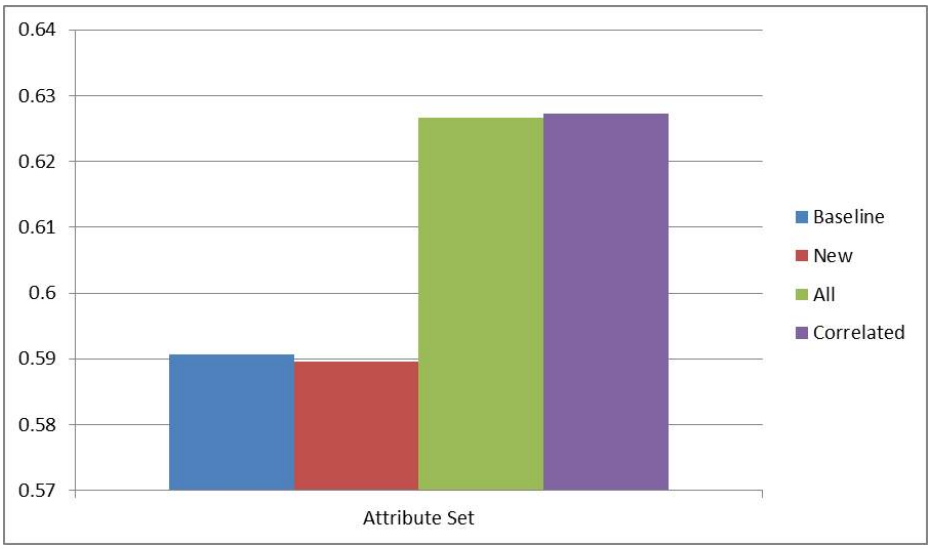

Figure 5.11: The AUC of Support Vector Machine for 60 days

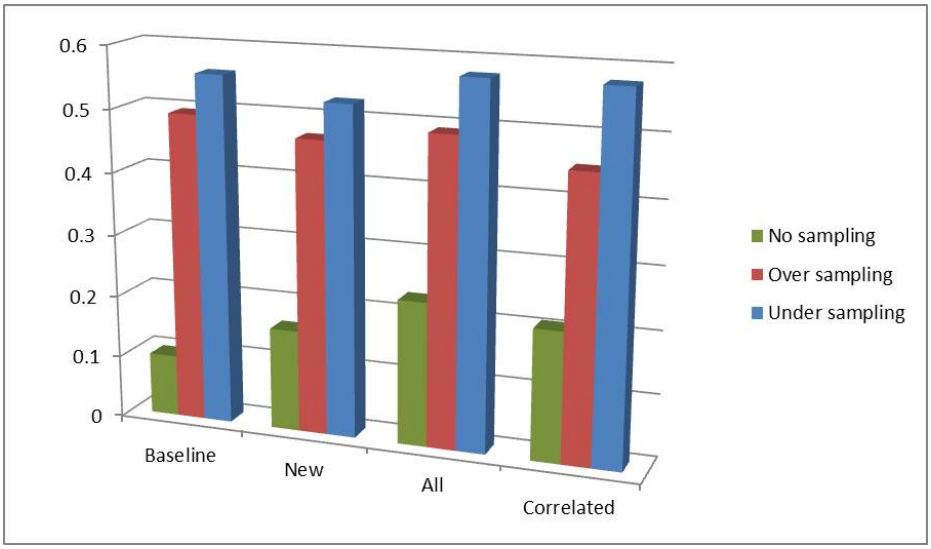

Figure 5.12: The Recall of Support Vector Machine for 60 days

5.4 Summary

It is observed that all the above models do not have a statistically significant difference in their performance. In other words, there is no significant advantage of one algorithm over the other except that logistic regression is a much simpler classification technique to use as compared to others. We were able to obtain a higher AUC value by using the *correlated* attribute set. All the models do an excellent job in predicting patients who were not hospitalized resulting in higher value of precision but lower recall and F measure, when no class balancing is performed. The reason behind this is the skewness in the class distribution (i.e., there are more patients not hospitalized than the ones encountering readmissions). Both oversampling and undersampling techniques improves the recall and F measure value but the improvement is at the cost of decrease in precision value. The sampling techniques do not substantially affect the area under the ROC curve.

Chapter 6

CONCLUSION AND FUTURE WORK

In this work, we have analyzed a large, high dimensional clinical dataset provided by MHS towards identifying the risk factors related to readmission of patients discharged with diagnosis of CHF, within 30 days and 60 days of discharge. The problem is formalized as a binary classification problem and different prediction models are developed and validated. We have compared the results of three different classification algorithms namely, logistic regression, naive Bayes and support vector machine with different attribute sets. The experimental results show that the models developed using correlated attribute set, obtained by chi-square test has the best performance in terms of area under the ROC curve. All the three classification algorithms have a comparable performance, however logistic regression and naive Bayes performs a little better than support vector machine on the current dataset. The recall has a significant improvement by employing class balancing techniques including oversampling and undersampling. The most important application of our model lies in identifying patients who are at a greater risk of getting readmitted within 30 days or 60 days of discharge. This model can be used by physicians to monitor the quality of inpatient care and reduce the readmission rate.

Through our research we have identified several potential areas of further work. Firstly, this work is not an exhaustive investigation of all known classification techniques. Other classification methodologies such as neural network, ensemble methods, can be explored. Neural networks is a well-known technique used for classification problem and has been successfully applied towards development of effective prediction models. Ensemble techniques are used to create improved composite classification models and have typically proven to be more accurate than base classifiers. Secondly, we have not covered all predictor variables in the dataset in developing the prediction models. It is interesting to perform a more exhaustive exploration of additional features in the dataset and study their relevance towards

predicting the risk of readmission. Thirdly, use of more sophisticated feature selection technique is required. We have employed a simple feature selection technique using chi-square test which gives the statistical significance of the variables. With more advanced feature selection techniques better attribute sets can be identified which may improve the performance of predictive model. Also, the current work restricts itself to developing prediction models for readmission risk within two fixed time frames: 30 days and 60 days from the date a patient is discharged. Instead it would be interesting to consider models aimed at predicting the expected number of days after discharge, within which a patient is likely to get readmitted. Finally, the efficacy of the developed predictive models can be truly judged only in the real world. Hence it is necessary to deploy the model in the field so that the actual performance of the model can be evaluated.

BIBLIOGRAPHY

[1] Kirkwood F. Adams, Gregg C. Fonarow, Charles L. Emerman, Thierry H. LeJemtel, Maria Rosa Costanzo, William T. Abraham, Robert L. Berkowitz, Marie Galvao, and Darlene P. Horton. Characteristics and outcomes of patients hospitalized for heart failure in the united states: Rationale, design, and preliminary observations from the first 100,000 cases in the acute decompensated heart failure national registry (ADHERE). *American Heart Journal*, 149(2):209–216, February 2005.

[2] G.F. Anderson and E.P. Steinberg. Predicting hospital readmissions in the medicare population. *Inquiry*, pages 251–258, 1985.

[3] C. M. Ashton, D. H. Kuykendall, M. L. Johnson, N. P. Wray, and L. Wu. The association between the quality of inpatient care and early readmission. *Annals of Internal Medicine*, 122:415415, 1995.

[4] Uri Balla, Stephen Malnick, and Ami Schattner. Early readmissions to the department of medicine as a screening tool for monitoring quality of care problems. *Medicine*, 87(5):294–300, September 2008.

[5] Alex Bottle, Paul Aylin, and Azeem Majeed. Identifying patients at high risk of emergency hospital admissions: a logistic regression analysis. *JRSM*, 99(8):406–414, August 2006.

[6] MD Chin and MD Goldman. Correlates of early hospital readmission or death in patients with congestive heart failure. *The American Journal of Cardiology*, 79(12):1640–1644, June 1997.

[7] Parry C Coleman EA. The care transitions intervention: Results of a randomized controlled trial. *Archives of Internal Medicine*, 166(17):1822–1828, September 2006.

[8] J. Han and M. Kamber. *Data mining: concepts and techniques*. Morgan Kaufmann, 2006.

[9] W. Hosmer David and L. Stanley. Applied logistic regression. *John Wiley&amp*, 2000.

[10] Stephen F. Jencks, Mark V. Williams, and Eric A. Coleman. Rehospitalizations among patients in the medicare fee-for-service program. *New England Journal of Medicine*, 360(14):1418–1428, 2009.

[11] H. Kaur and S. K. Wasan. Empirical study on applications of data mining techniques in healthcare. *Journal of Computer Science*, 2(2):194200, 2006.

[12] Todd M. Koelling, Monica L. Johnson, Robert J. Cody, and Keith D. Aaronson. Discharge education improves clinical outcomes in patients with chronic heart failure. *Circulation*, 111(2):179–185, January 2005.

[13] H. C. Koh and G. Tan. Data mining applications in healthcare. *Journal of Healthcare Information ManagementVol*, 19(2):65, 2011.

[14] H. M. Krumholz, S. L. T. Normand, P. S. Keenan, Z. Q. Lin, E. E. Drye, K. R. Bhat, Y. F. Wang, J. S. Ross, J. D. Schuur, and B. D. Stauffer. *Hospital 30-day heart failure readmission measure methodology. Report prepared for the Centers for Medicare & Medicaid Services.*

[15] Harlan M Krumholz, Joan Amatruda, Grace L Smith, Jennifer A Mattera, Sarah A Roumanis, Martha J Radford, Paula Crombie, and Viola Vaccarino. Randomized trial of an education and support intervention to preventreadmission of patients with heart failure. *Journal of the American College of Cardiology*, 39(1):83–89, January 2002.

[16] Parent EM Krumholz HM. REadmission after hospitalization for congestive heart failure among medicare beneficiaries. *Archives of Internal Medicine*, 157(1):99–04, January 1997.

[17] Brooten D Naylor MD. Comprehensive discharge planning and home follow-up of hospitalized elders: A randomized clinical trial. *JAMA: The Journal of the American Medical Association*, 281(7):613–620, February 1999.

[18] K.J. Ottenbacher, P.M. Smith, S.B. Illig, R.T. Linn, R.C. Fiedler, and C.V. Granger. Comparison of logistic regression and neural networks to predict rehospitalization in patients with stroke. *Journal of clinical epidemiology*, 54(11):1159–1165, 2001.

[19] William W. Parmley. Pathophysiology and current therapy of congestive heart failure. *Journal of the American College of Cardiology*, 13(4):771–785, March 1989.

[20] Edward F Philbin and Thomas G DiSalvo. Prediction of hospital readmission for heart failure: development of a simple risk score based on administrative data. *Journal of the American College of Cardiology*, 33(6):1560–1566, 1999.

[21] Wright SM Phillips CO. Comprehensive discharge planning with postdischarge support for older patients with congestive heart failure: A meta-analysis. *JAMA: The Journal of the American Medical Association*, 291(11):1358–1367, March 2004.

[22] Michael W. Rich, Valerie Beckham, Carol Wittenberg, Charles L. Leven, Kenneth E. Freedland, and Robert M. Carney. A multidisciplinary intervention to prevent the readmission of elderly patients with congestive heart failure. *New England Journal of Medicine*, 333(18):1190–1195, 1995.

[23] Joanne Kraenzle Schneider, Susan Hornberger, Jane Booker, Alyce Davis, and Randy Kralicek. A medication discharge planning program measuring the effect on readmissions. *Clinical Nursing Research*, 2(1):41–53, February 1993.

[24] J. M. Vinson, M. W. Rich, J. C. Sperry, A. S. Shah, and T. McNamara. Early readmission of elderly patients with congestive heart failure. *Journal of the American Geriatrics Society*, 38(12):1290, 1990.