UNIVERSITY OF CALIFORNIA

SANTA CRUZ

Data Mining for Improving Health-Care Resource Deployment

A thesis submitted in partial satisfaction

of the requirements for the degree of

MASTER OF SCIENCE

in

TECHNOLOGY AND INFORMATION MANAGEMENT

by

NANNAN HE

March 2014

The thesis of Nannan He is approved:

_____

Dr. Subhas Desa

_____

Professor Patrick E. Mantey

_____

Professor Brent M. Haddad

_____

Tyrus Miller

Vice Provost and Dean of Graduate Studies

Table of Contents

List of Tables

List of Figures

Abstract

Data Mining for Improving Health-Care Resource Deployment

Nannan He

While the health care industry accounts for a significant large portion of the GDP, the health care system in the US are still relatively inefficient. Before cutting down unnecessary health care expenses, it is important to ensure that individuals who really need medical attention should receive it. For example, if we could predict the hospitalization period (in days) for a potential patient, then we could better predict and distribute health care resources.

In this research, we apply data mining methods and tools to address the problem of predicting future hospitalization periods (in days) for patients from a given set of historical patient data. The data mining techniques that we explored were linear regression, random forest and gradient boosting. For each technique, we used different historical data sets. The combination of data mining techniques and historical datasets enabled us to compare access and choose the combination which provides the best prediction of hospitalization period of a set of patients. Based on the results of our work, the random forest technique provides the best prediction of patient hospitalization.

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my appreciation to all the people who have helped and supported me in this thesis research.

First of all, I would like to express my deep appreciation to my advisor Dr.Subhas Desa, for his guidance, and encouragement through the entire duration of my thesis work. He motivated me to think more deeply about my work. He also made great effort to build the structure and refine every detail of my work. I have learned how to conduct scientific research under rigorous scholarship, which will benefit me through my entire life.

Secondly, I want to thank Tyler Munger for providing me many useful suggestions during the course of my research. He made time to meet with me and help me to build a systematic framework for the research.

I would like to thank my committee members Professors Patrick Mantey and Brent Haddad for reviewing of my thesis and providing valuable and insightful suggestions.

Finally, I heartily thank my close friends and my family for their great support and encouragement during my study.

1 Introduction

1.1 Background and Motivation

According to the World Health Organization (WHO), in 2011 health care was a major component of the US GDP (17.2%), and more than any other nation. Although its health care expenditures were so high, the Commonwealth Fund ranked the United States last in the quality of health care among similar countries, and notes that U.S. care costs the most. In a 2013 Bloomberg ranking of nations with the most efficient health care systems, the United States ranks 46th among the 48 countries included in the study (Bloomberg, 2013).

An important question is why this large expenditure does not provide health care system quality and efficiency. Answering this question would potentially help the healthcare system better deploy financial resources and thus work more efficiently. Studies showed that in year 2006, over $30 billion were spent on unnecessary hospital admissions (Davidson, 2013). This suggests that health care expenditures could be reduced significantly by avoiding or reducing unnecessary hospitalization. To address the hospitalization issue, and plan and manage hospital resources, require better projection of who will need hospital admission, as well as the expected hospitalization period for each patient.

To this end, our research addresses the application and implementation of data mining techniques to predict patient hospitalization periods.

1.2 Research Problem

1.2.1 Research Issue

Data mining approaches solving problem through analysis of massive data. Data mining is defined as the process of discovering patterns in data (Ian H.Witten, 2011), and it turns a dataset into understandable knowledge. As a highly application-driven discipline, data mining has seen great successes in many applications. Applying data mining techniques in medical field is one of such application. Medical data are usually complex, diverse and large in volume. It is hard to find patterns and draw useful information from this data. Data mining techniques provide a means to access the data and generate knowledge. In this research, we address prediction of patient hospitalization periods (in days), based on historical patient records. Systematically applying data mining techniques to perform this prediction is the major research issue.

1.2.2 Research Contribution

The research issues address in this thesis involve implementing data mining techniques for specific medical research problem, which seek to identify which patients will need hospital admission in the third year, and how many days he/she will be in hospital based on previous two years of hospitalization data. In this research we created different prediction models and compared several data mining techniques apply to this problem. The thesis contributions are:

Use of data mining knowledge to complex and massive health care data to perform prediction;

1. A research methodology that integrates data preparation, data mining model building, data mining algorithms comparison, and results evaluation for medical research.

2. The application of research methodology to classification and prediction of individual patient future hospitalization duration.

1.2.3 Data Description

The dataset used was obtained from Heritage Health Prize (Heritage Health Prize , 2012). The dataset processed in this research includes two years of historical patient information. It consists of five tables. They are the the *"Claims"* Table, the *"DaysInHospital_Y2"* Table, the *"DrugCount"* Table, the *"LabCount"* Table and the *"Members"* Table respectively. The overall dataset are summarized in the Table 1 below:

Table 1 Summary of Raw Dataset

Table Name	**Attribute Name**	**Date Type**	**Description**	**Number of Categories**
Members	Member ID	Nominal	Member identifier.	
	Age At First Claim	Ordinal	Age in years at the time of the first claim's date of service computed from the date of birth; Generalized into ten year age intervals.	10
	Sex	Nominal	Biological sex of member: M = Male; F=Female.	3
	Member ID	Nominal	Member identifier.	

Claims	ProviderID	Nominal	Provider identifier	14700
	Vendor	Nominal	Vendor identifier	6388
	PCP	Nominal	Primary care physician pseudonym.	1360
	Year	Nominal	Primary care physician pseudonym.	3
	Specialty	Nominal	Generalized specialty.	13
	PlaceSvc	Nominal	Generalized place of service.	9
	PayDelay	Numeric	Number of days delay between the date of service. Values above 161 days (the 95% percentile) are top-coded as "162+".	
	LengthOfStay	Ordinal	Length of stay, (1-2] weeks;	11
	DSFS	Ordinal	Day since first claim, computed from the first claim for that member for each year, generalized to: [0-1] month,	13
	PrimaryConditionGroup	Nominal	Generalization of primary diagnosis codes	46
	CharlsonIndex	Ordinal	Measure of mortality based on comorbid conditions	6
	ProcedureGroup	Nominal	Broad categories of procedures	18
	SupLOS	Nominal	Indicates if the NULL value for the LengthOfStay variable is due to suppression done during the de-identification process. A value of 1 indicates that suppression was	2

			applied.	
DaysIn Hospital	Member ID	Nominal	Member identifier.	
	Days in Hospital	Numeric	Days the member was hospitalized the next year	16
	ClaimsTruncated	Numeric	Members with truncated claims in the year prior to the main outcome are assigned a value of 1, and 0 otherwise	2
LabCount	Member ID	Nominal	Member identifier.	
	LabCount	Numeric	Count of unique laboratory and pathology tests by DSFS.	7
	Year	Nominal	Primary care physician pseudonym.	
	DSFS	Ordinal	Day since first claim, computed from the first claim for that member for each year, generalized to: [0-1] month,	13
DrugCount	DrugCount	Numeric	Count of unique prescription drugs filled by DSFS.	10
	Member ID	Nominal	Member identifier.	
	Year	Nominal	Primary care physician pseudonym.	
	DSFS	Ordinal	Day since first claim, computed from the first claim for that member for each year, generalized to: [0-1] month,	13

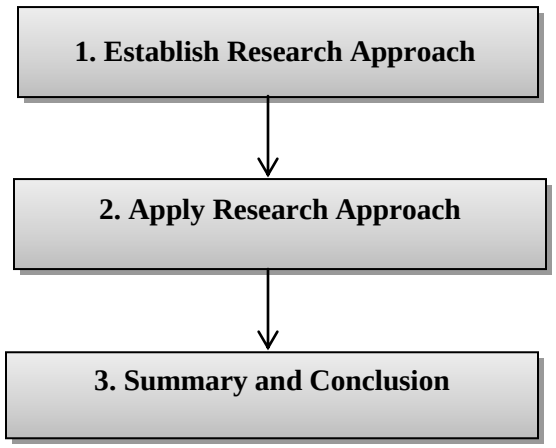
The dataset is very large and its data types vary significantly. In the *"claims"* Table, over one million records are stored and in the *"Member"* Table over 11,300 different

members are stored. For Year2, in the *"DaysInHospital"* table, 76,038 records are stored. And it is noticeable that there are many missing and null values in the table.

1.2.4 Research Methodology

The research work was composed of three phases, sketched below (see Figure 1):

Figure 1 Overview of Research Methodology



1. Establish Research Approach: The research approach is about how to carry out the research in a systematic way, and is the section demonstrating the creativity of the research. It provides the guideline for the entire work. The implementation follows the approach. It includes several aspects, such as which algorithms are applied and how they are applied to data.

2. Apply Research Approach: The implementation section shows the results from following research approach.

3. Summary and Conclusion: In the last section, we draw conclusions from the research work, evaluate the results and propose future work. Although this is

the last part, however, it highlights the value of the research and lays a solid foundation for future work.

1.3 Organization of Work

The thesis is structured as follows. Chapter 2 designs a research approach used in the research and explains how each step in the approach was conducted. Chapter 3 discusses the first three phases of the research approach, which are data preparation, data mining model establishments and data mining algorithm application. Chapter 4 focuses on results evaluation and improvement and is the last phase of the research approach as discussed earlier. Chapter 5 is the final part, summarizing the work and discussing the future work.

2 Theory

In this chapter, we show the development of the approach need to carry out the research. Generally each step in the approach is described at first, and then followed by details of each step. To better explain the third and fourth steps, we have followed by describing the related algorithms and the evaluation methods used.

2.1 Approach

In this section, we will present an overview of how this research was carried out step by step. The approach consisted of four major steps:

1. Data Preprocessing: The objective of data preprocessing was to bring data together and transform the data into the desired format. It includes data cleaning and features generation two parts.

2. Predictive Model Establishment: The aim of building data mining model is to identify the relations between datasets, and then build connections among datasets. In this research, the data mining model helps us mine patterns from previous historical medical data, in order to predict future hospitalization periods. The models' value is to help explain how the data relate.

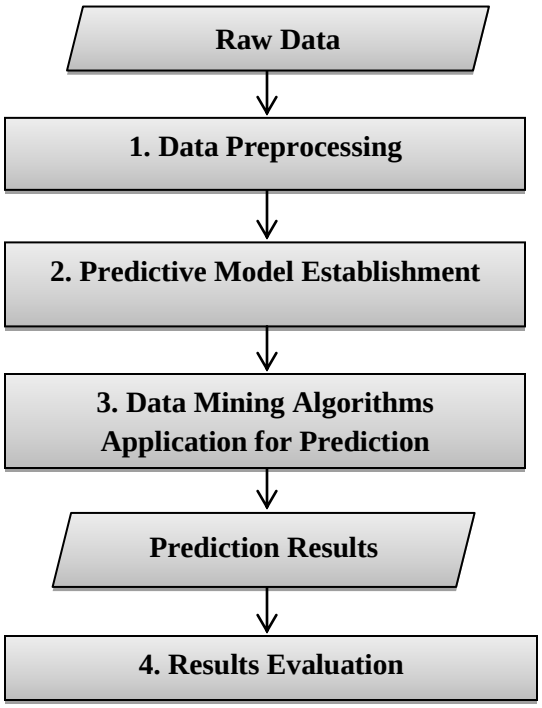
3. Applying Data Mining Algorithms for Prediction Analysis: In this step, we describe the selection of data mining algorithms and then apply them to predictive models. This was the key step in the entire research approach, because the selection of data mining algorithms largely determines the quality

of the results. Within this step, we note that simple ideas often work very well, so we adopted of a “simplicity-first” methodology when analyzing practical datasets.

4. Results Evaluation: The results evaluation and analysis is a process where we inspected the research quality and interpreted the results. It included examination of the performance of each algorithm, the quality of results and determination of the advantage and disadvantage of each algorithm for our research.

The approach is shown as Figure 2. It illustrates how we performed the research approach.

Figure 2 The Overview of Approach of Hospitalization Prediction



2.2 Data Preprocessing

The given raw dataset is problematic. The symptoms include attributes incalculable, measurement units inconsistent and etc. Data preprocessing is an effective way to tackle these problems, helping to improve data quality. The raw dataset is simply tables of record, having little value for data mining. Data preprocessing perform aggregation or decomposition attributes when necessary, to convert the original tables into useful datasets.

Data preprocessing was performed in two parts in our research. One was data cleaning and the other was feature generation. Data cleaning involved filling in missing values, smoothing out noise, identifying outliers in the data (Jiawei Han, 2012). In our research, we focused on the problem of missing values. Missing values bring many problems to data mining work by adding the uncertainty. The common methods to deal with missing values are to ignore the tuple, filling in the missing values manually, using a global constant to fill in missing values, using a measure of central tendency for the attribute to fill in the missing value, using the attribute mean or median for all samples belonging to the same class as the given tuple, using the most probable value to fill in the missing value (Jiawei Han, 2012). Feature generation is the other important data preprocessing task. For exacting useful information in a large dataset, related attributes must be acquired. The related attributes refers to the features. Sometime, features are particular attributes in the dataset, which can be directly used in data mining process. But, more commonly, features are draw by manipulating the dataset. Aggregation and decomposition are

two popular means used to uncover features. However, to what extent the aggregation and decomposition are performed is determined by the requirements of prediction models. Overuse of any of them may results in a feature losing its general trend. In our research, dataset was aggregated to one training dataset and one prediction dataset. The number of features was significantly compressed for data mining purpose.

2.3 Predictive Model Establishment

Predictive model were built in this step. In this dataset, the basic information about members are sex, age and memberId. The claims data, drugcoutn data and labcount data were available for Year1 and Year2. We also have DaysInHospitals(DIH) data for Year2. The goal is to predict Year3 DaysInHospitals data. We call the claims data, drugcount data and labcount data Member Data.

By analyzing the dataset, we saw many possibilities for organizing the dataset. Since our goal was to predict hospitalization period of Year3, we decided that the *"Member"* Table and *"DaysInHospitals"* Table of Year2 should be most useful. But as we also have the *"Member"* Table and *"DaysInHospitals"* Table from Year1, we included it in training dataset to improve the prediction quality. And thus, we could utilize Member Data from Year1 or/and Year2 as training dataset, use Member Data from Year2 or/and Year1 as prediction dataset. Totally, four predictive models were conceived. To simplify remembering the model name, we use year to label model name.

The first Predictive Model is called T1P2 Model, involving Member Data from Year1 and DIH of Year2 as Training Dataset and Member Data from Year2 as Prediction Dataset (see Table 2):

Table 2 Predictive Model 1

Predictive Model 1: T1P2 Model	
Training Dataset	Member Data from Year1 and DIH of Year2
Prediction Dataset	Member Data from Year2
Prediction Goal	DIH of Year3

By adding Member Data from Year2, we got an alternative model, named T1P12 Model (see Table 3):

Table 3 Predictive Model 2

Predictive Model 2: T1P12 Model	
Training Dataset	Member Data from Year1 and DIH of Year2
Prediction Dataset	Member Data from (Year1 and Year2)
Prediction Goal	DIH of Year3

The third model combined two year historical data in training set and used the same prediction set as T1P2 model. As more historical data was added to training dataset, we expected the results will be improved. Similarly, we expanded the prediction set to obtain the third model, T12P2 Model (see Table 4):

Table 4 Predictive Model 3

Predictive Model 3: T12P2 Model	
Training Dataset	Member Data from (Year1 and Year2) and DIH of Year2
Prediction Dataset	Member Data from Year2
Prediction Goal	DIH of Year3

Finally, we came to the fourth model: T12P12 (see Table 5). It used member data from both Year1 and Year2 as training dataset. And it used both member data from Year1 and Year2 to predict DIH data of Year3. This model involved the most data.

Table 5 Predictive Model 4

Predictive Model 4:T12P12 Model	
Training Dataset	Member Data from (Year1 and Year2) and DIH of Year2
Prediction Dataset	Member Data from (Year1 and Year2)
Prediction Goal	DIH of Year3

2.4 Data Mining Algorithms for Prediction

Before we could apply the data mining algorithms to data, we needed to select the appropriate ones. Following our "simplicity-first" methodology, two types of algorithms were adopted. One is regression method, which is common in prediction application. Regression is used to predict the value of a response (dependent) variable from one or more predictor (independent) variables, where the variables are numeric. One typical regression example is linear regression. The other type of algorithm is

classification. Decision trees algorithm is a well-known classification algorithm. However, it has limited ability a facing large amount and complex dataset. Thus, we introduced two improved decision tree algorithm: random forest and gradient boosting.

2.4.1 Linear Regression

Linear regression is a natural technique to consider when we try to establish connection between numeric attributes. Ideally, linear regression is used to identify the relationship between a single predictor value x and related attributes value a_k with respect to a linear distribution. Sometimes, to exaggerate the importance of certain related attributes value than the others, different weights are assigned:

$$x = w_0 + w_1a_1 + w_2a_2 + \ldots + w_ka_k$$

where x is the predictor value; a_1, a_2, …, a_k are the attribute values; and w_0, w_1, …, w_k are weights.

2.4.2 Random Forest

Decision trees (L. Breiman, Classification and regression trees, 1984) are widely used in botany, taxonomy or medical diagnosis. A basic decision tree is a hierarchical set of nodes, starting from a root node, each one containing a decision involving the comparison of an attribute with a given threshold, which then leads to another node or to a leave. The decision tree classification method is computationally simple and easy to understand. The biggest limitation for a basic tree classification is that when data

shows high variance, the prediction accuracy is usually low. To overcome this drawback and maintain advantages, ensemble of trees and letting them vote for the most popular class is recommended. Random forests were introduced by Breiman (Breiman, 2001). Random forests build a randomized decision tree in each iteration of the bagging algorithm, and often produce excellent predictors (Breiman, 2001). Breiman proposed to grow each tree via a random selection (without replacement). Random forest can be built using bagging in tandem with random attribute selection. (Jiawei Han, 2012).

2.4.3 Gradient Boosting

Another case for enhancing robustness of the regression tree is gradient boosting (Friedman, 2001). Boosting is a popular method used to improve model accuracy (Schapire, 2002). It assumes that each model excels at handling certain domains where other models don't perform very well. As each model is built separately, the new model will be influenced by the previously built one, and then improve its performance on the instances that are not well treated by the previous one. Gradient boosting is a flexible data mining method caring model fitting, and it is able to identify the influential attributes in the data mining process.

2.5 Results Evaluation

Last but not least, results evaluation has being valued equally important to the algorithms application process nowadays, because it is the step to test whether the

application is successful. Identifying an efficient method to interpreter the results is not easy. Usually, cross-validation is applied to data mining results evaluations. However, since we had the actual DIH Year3 dataset, we were able to directly compare the prediction results and facts to provide a more immediate representation of the quality of the results. We adopted three commonly used forecasting results evaluation methods: Mean Square Error (MSE), Mean Absolute Deviation (MAD) and Mean Absolute Percentage Error (MAPE).

The mean squared error (MSE) is one of statistics ways to quantify the difference between values implied by an estimator and the true values of the quantity being estimated. MSE measures the average of the squares of the "errors". The error is the difference between predicted value and actual value. The MSE can be related to the variance of the prediction error.

$$MSE = \frac{\sum_{n=1}^{n} E^2}{n}$$

where E is prediction error, where $E = Predict - Actual$.

The Mean Absolute Deviation (MAD) is defined as the average of the absolute deviation over all predictions. MAD can be used to estimate the standard deviation of the random component assuming that the random component is normally distributed.

$$MAD = \frac{\sum_{n=1}^{n} |E|}{n}$$

where E is prediction error, $E = Predict - Actual$.

The mean absolute percentage error (MAPE), also known as mean absolute percentage deviation (MAPD), is a measurement of accuracy of for trend estimation. The absolute value in this calculation is summed for every predicted value in time and divided again by the number of predicted points *n* multiplying by 100, making it a percentage error. It usually expresses accuracy as a percentage.

$$\mathrm{M} = \frac{1}{n}\sum_{n=1}^{n} \left|\frac{E}{A}\right|.$$

where E is prediction error, $A = Actual$.

3 Research Implementation

In this chapter, we describe the process of research implementation following designed approach. We begin with data preprocessing, and then move to the predictive model establishment. We also elaborate how the algorithms were applied to each model, including analyzing attributes importance, identifying model overfitting and etc. Many tables and figures are presented to facilitate the explanation.

3.1 Data Preprocessing

3.1.1 Data Cleaning

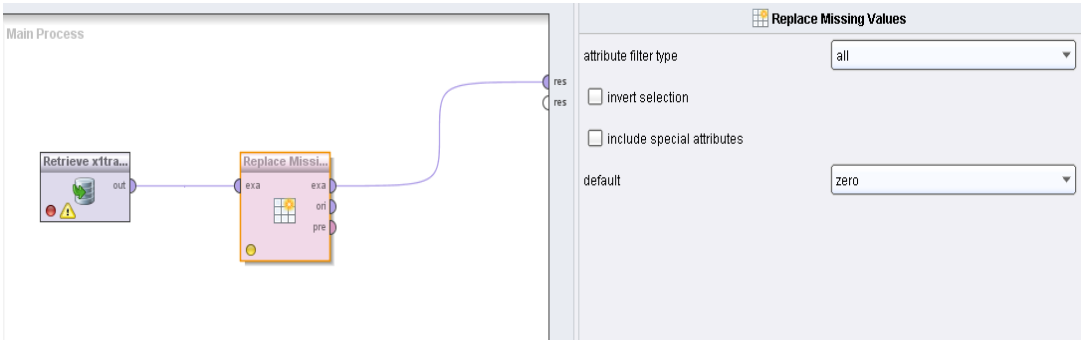
According to Table 1, the raw data contained many different data types. Some columns in the data files contained numerical values, such as *LabCount*, *DrugCount* and *PayDelay*; some had categorical values, such as *AgeAtFirstClaim* and *Year*, and others had binary value, such as *ClaimsTruncated*. There were many columns also having some missing values, such as Sex. We need to preprocess these data firstly. The data cleaning work includes:

1. For Claims table, changed data in "LengthOfStay" to days and used the interval data to represent.
2. For the data in "DSFS" in any table, applied the same methods as above.
3. For data in "PayDelay", replaced descriptions like 4-8 weeks with the average of an interval (6 weeks).

4. For data in “PayDelay”, replaced162+ with 163.
5. For data in CharlsonIndex, replaced original data with its upper bound, such as using 2 to replace “1-2”.
6. In LabCount table , replaced 10+ with 11.
7. For DrugCount, replaced 7+ with 8.

For missing values, they were replaced by value zero (see Figure 3) in RapidMiner.

Figure 3 Filling in Missing Values



3.1.2 Feature Generation

Overlooking the given dataset, each member had one or more claims in dataset. And for each claim, over 30 attributes were recorded. For particular attributes, it contained many categories for each attributes. Directly performing data mining on such dataset is futile. The most imperative step is to reorganize the dataset and decompose it. All the information of *Claims*, *Labcount*, *Drugcount* for each patient in a particular year must be extracted. The following features were created for each member (see Table 6). The total number of generated feature was 32.

Table 6 Generated Feature Summary

Features for Each Patient
Numer of ClaimsTruncated
Number of DaysInHospital
Number of each primary care physicians
Number of each vendors
Number of each providers
Number of each specialties
Number of each placessvc
Number of each primary condition groups
Number of each ProcedureGroup
Number of times each specialty shows up
Number of times each placesvc shows up
Number of times each primary condition group shows up
Average, max, min of PayDealy
Average, max, min of LengthOfStay
Average, max, min of DSFS
Average, max, min of CharlsonIndex
Average, max, min, sum of DrugCount
Average, max, min, sum of LabCount

The benefit for reorganization the dataset was reflected in the second step. It helped different model to access to corresponding years of data more easily.

3.2 Predictive Model Establishment

As we discussed in the Chapter 2, four predictive models to predict hospitalization period of Year3 were established in the research.

1. *Predictive Model 1:T1P2 Model*

Training Set: *Member Data from Y1 and DIH Data from Y2*

Prediction Set: *Member Data from Y2*

2. *Predictive Model 2:T1P12 Model*

Training Set: *Member Data from Y1 and DIH Data from Y2*

Prediction Set: *Member Data from (Y1 + Y2)*

3. *Predictive Model 3:T12P2 Model*

Training Set: *Member Data from (Y1 + Y2) and DIH Data from Y2*

Prediction Set: *Member Data from Y2*

4. *Predictive Model 4:T12P12 Model*

Training Set: *Member Data from (Y1 + Y2) and DIH Data from Y2*

Prediction Set: *Member Data from (Y1 + Y2)*

For each of them we generated different dataset. Totally, four data sheets were created:

1. *Training Dataset: Member Data from Y1 and DIH Data from Y2*
2. *Training Dataset: Member Data from (Y1 + Y2) and DIH Data from Y2*
3. *Prediction Dataset: Member Data from Y2 for prediction DIH Year3*
4. *Prediction Dataset: Member Data from (Y1 + Y2) for prediction DIH Year3*

In the above dataset, table 1 and 3 were simply draw from reorganized dataset with respect to corresponding year. For dataset 2 and 4, a new dataset aggregating member data from Year1 and Year2 was generated by re-computing all the features.

Considering different categories existing in each different attributes, the total involved attributes number for prediction was 112. The format of member datasets is bellowing (see Table 7). The four datasets contained all the bellowing attributes. The major difference for training dataset and prediction dataset was that DIH data in prediction dataset was empty.

Table 7 Member Data Attributes Summary

Attribute Name	Type
Memberid	Numeric
ClaimsTruncated	Numeric
DaysInHospital	Numeric
num_ProviderID	Nominal
num_Vendor	Nominal
num_PCP	Nominal
num_Specialty	Nominal
num_PlaceSvc	Nominal

num_PrimaryConditionGroup	Nominal
num_ProcedureGroup	Nominal
sp_ane	Nominal
sp_obs	Nominal
sp_dia	Nominal
sp_eme	Nominal
sp_gen	Nominal
sp_int	Nominal
sp_lab	Nominal
sp_oth	Nominal
sp_pat	Nominal
sp_ped	Nominal
sp_reh	Nominal
sp_sur	Nominal
ps_amb	Nominal
ps_hom	Nominal
ps_ind	Nominal
ps_inp	Nominal
ps_off	Nominal
ps_oth	Nominal
ps_out	Nominal
ps_urg	Nominal
pcg_ami	Nominal
pcg_app	Nominal
pcg_art	Nominal
pcg_cancra	Nominal
pcg_cancrb	Nominal
pcg_cancrm	Nominal
pcg_cat	Nominal
pcg_chf	Nominal
pcg_cop	Nominal

pcg_fla	Nominal
pcg_fxd	Nominal
pcg_gib	Nominal
pcg_gio	Nominal
pcg_gynec1	Nominal
pcg_gyneca	Nominal
pcg_heart2	Nominal
pcg_heart4	Nominal
pcg_hem	Nominal
pcg_hip	Nominal
pcg_inf	Nominal
pcg_liv	Nominal
pcg_metab1	Nominal
pcg_metab3	Nominal
pcg_mis	Nominal
pcg_miscl1	Nominal
pcg_miscl5	Nominal
pcg_msc	Nominal
pcg_neu	Nominal
pcg_oda	Nominal
pcg_peri	Nominal
pcg_perv	Nominal
pcg_pnc	Nominal
pcg_pne	Nominal
pcg_prg	Nominal
pcg_ren1	Nominal
pcg_ren2	Nominal
pcg_ren3	Nominal
pcg_res	Nominal
pcg_roa	Nominal
pcg_sei	Nominal

pcg_sep	Nominal
pcg_skn	Nominal
pcg_str	Nominal
pcg_tra	Nominal
pcg_uti	Nominal
pg_ane	Nominal
pg_em	Nominal
pg_med	Nominal
pg_pl	Nominal
pg_rad	Nominal
pg_sas	Nominal
pg_scs	Nominal
pg_sds	Nominal
pg_seo	Nominal
pg_sgs	Nominal
pg_sis	Nominal
pg_smcd	Nominal
pg_sms	Nominal
pg_sns	Nominal
pg_so	Nominal
pg_srs	Nominal
pg_sus	Nominal
PayDelay_Max	Nominal
PayDelay_min	Nominal
PayDelay_avg	Nominal
LengthOfStay_Max	Nominal
LengthOfStay_min	Nominal
LengthOfStay_avg	Nominal
DSFS_Max	Nominal
DSFS_min	Nominal
DSFS_avg	Nominal

CharlsonIndex_Max	Nominal
CharlsonIndex_min	Nominal
CharlsonIndex_avg	Nominal
drugcount_max	Nominal
drugcount_min	Nominal
drugcount_avg	Nominal
drugcount_sum	Nominal
labcount_max	Nominal
labcount_min	Nominal
labcount_avg	Nominal
labcount_sum	Nominal

3.3 Data Mining Algorithms Application for Prediction

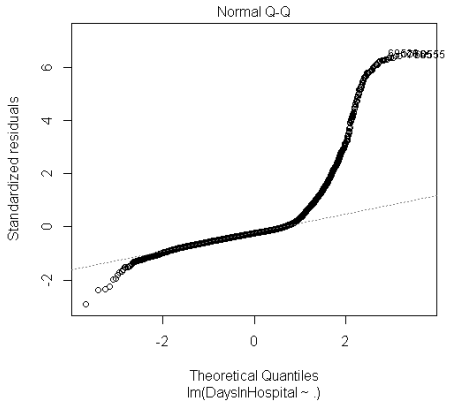
3.3.1 Linear Regression Algorithm for Prediction

Liner regression is an approach to modeling the relationship between a scalar dependent variable y and one or more explanatory variables denoted x. For linear regression, the algorithm is building relationship between target value (DIH for Year3) and other attributes.

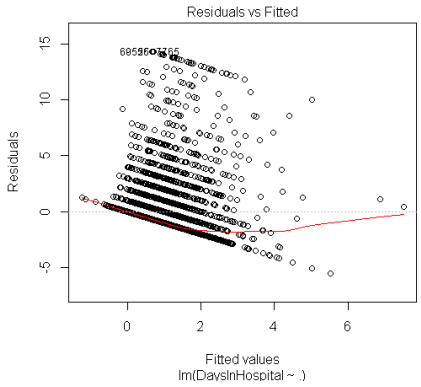
(1) Applying on Predictive Model 1: T1P2 Model

The algorithm was applied in R, which is open source data processing software to process large dataset. The model parameters were represented as below (see Figure 4):

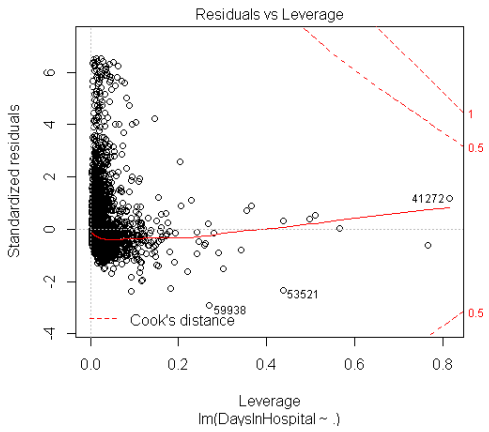
Figure 4 Linear Model Plot for Data Mining Model 1



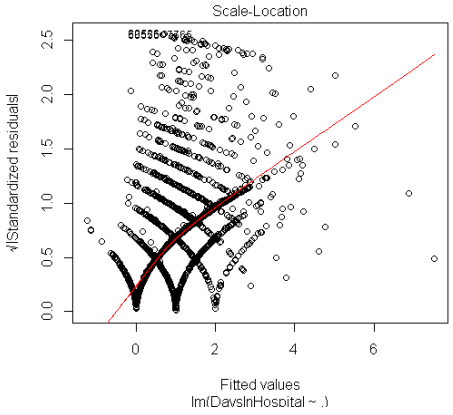
(a) Q-Q plot



(b) Residuals vs. Fitted



(c) Residuals vs. Leverage



(d) Scale-Location

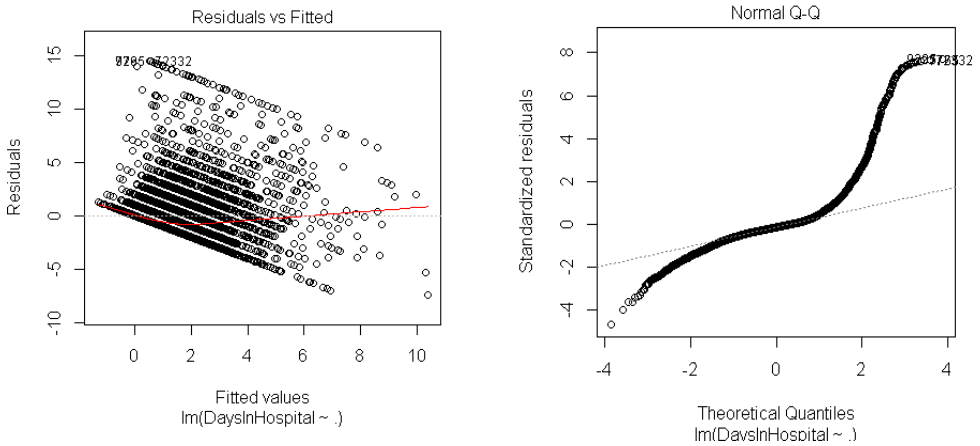
The first normal Q–Q plot ("Q" stands for quantile) is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other. It is a powerful visualization tool allowing the user to view whether there is a shift from one distribution to another. The first half of the residuals fit the theoretical residuals; however, the second was derivate from it. The

second plot residuals vs. fits plot is a scatter plot of residuals on the y axis and fitted values (estimated responses) on the x axis. The plot is used to detect non-linearity, unequal error variances, and outliers. We can see the residuals distribute unbalanced around 0-line, indicating that the assumption that the relationship was linear is not quite fit. And most residuals were above the 0-line, suggesting the variances of the error terms were not equal. And the point at the right end might be an outlier, which stood out from large cluster. The third plot residuals vs. leverage gives the labeled points that we may want to investigate as possibly having undue influence on the regression relationship. The fourth plot scale-location plot is similar to the residuals versus fitted values plot, but it uses the square root of the standardized residuals, which also suggested that we have outliers in the dataset.

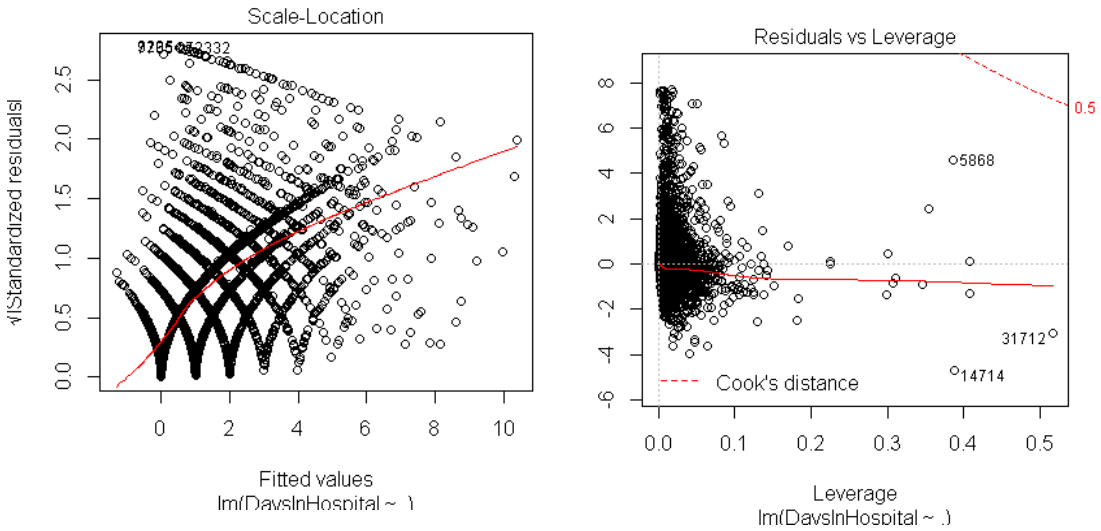
(2) Applying on Predictive Model 3: T12P2 Model

The model parameters are represented as below (see):

Figure 5 Linear Model Plot for Data Mining Model 3



(a) Residuals vs. Fitted (b) Q-Q plot



(c) Scale-Location (d) Residuals vs. Leverage

From the above figure, we can see that Model 3 shared the similar characters with Model 1. The major difference was in residuals vs. leverage plot. It suggested that some new outliers occurs and rang of standardized residuals became larger and the leverage decreased.

3.3.2 Random Forest Algorithm for Prediction

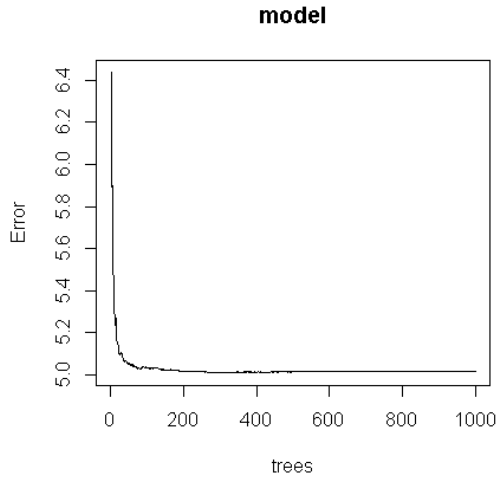
Random forests are one of the popular ensemble methods, mainly used to increase overall accuracy by learning and combining a series of individual (base) classifier models. In the algorithm setting, n-tree was set to 1000. It indicated the number of trees to grow. This should not be set to too small a number, to ensure that every input row gets predicted at least a few times. M-try was set to 3. It is the number of variables randomly sampled as candidates at each split. Note that the default values were different for classification (sqrt(p) where p is number of variables in x) and regression (p/3). Node-size is the Minimum size of terminal nodes. Setting this

number larger caused smaller trees to grow (and thus take less time). Parameter of importance is True, which will assess the importance of predictors.

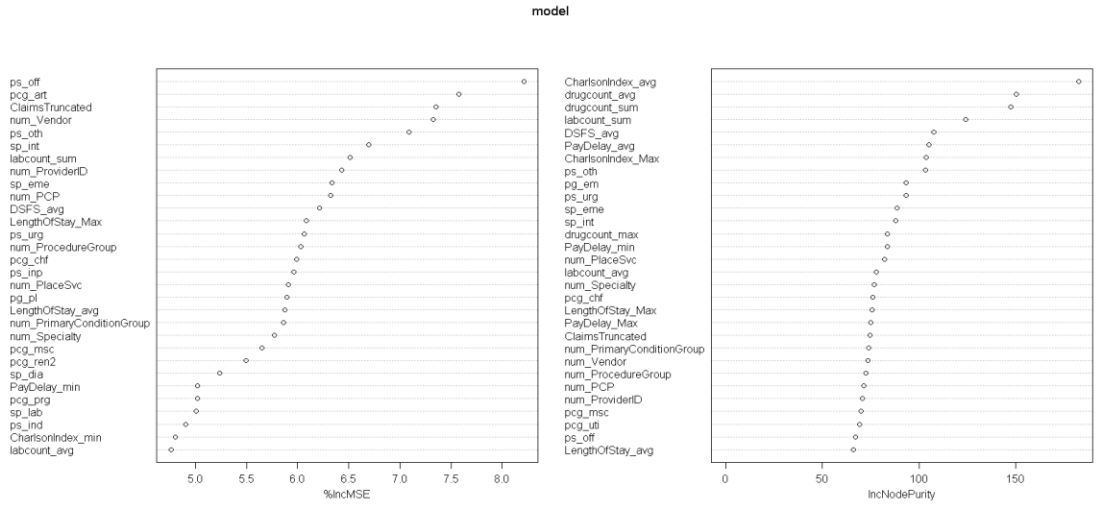
(1) Applying on Predictive Model 1: T1P2 Model

Model 1 used first year historical data as training set and also only used second year as prediction set. The model parameter can be seen from Figure 6:

Figure 6 Random Forest Plot for Data Mining Model 1



(a) Model Plot



(b) Variable Importance Plot

```
Error in MDSplot(model) : model does not contain a proximity matrix
> getTree(model,3, labelVar=TRUE)
    left daughter right daughter                split var split point status
1               2              3                   pg_eso       0.500     -3
2               4              5              drugcount_sum      14.500     -3
3               6              7 num_PrimaryConditionGroup      14.500     -3
4               8              9 num_PrimaryConditionGroup      11.500     -3
5              10             11                   pcg_roa       0.500     -3
6              12             13           CharlsonIndex_min       1.000     -3
7               0              0                        <NA>       0.000     -1
8              14             15                 pcg_misc15       2.500     -3
9              16             17                 pcg_cancrm       2.500     -3
10             18             19              drugcount_sum      91.500     -3
11             20             21                 pcg_heart2       2.500     -3
12             22             23          LengthOfStay_avg       1.250     -3
13              0              0                        <NA>       0.000     -1
14             24             25           ClaimsTruncated       0.500     -3
15              0              0                        <NA>       0.000     -1
16             26             27           ClaimsTruncated       0.500     -3
17              0              0                        <NA>       0.000     -1
18             28             29                   pcg_liv       0.500     -3
19              0              0                        <NA>       0.000     -1
20             30             31                   sp_lab       2.500     -3
21              0              0                        <NA>       0.000     -1
22             32             33                   ps_amb       4.500     -3
23              0              0                        <NA>       0.000     -1
24             34             35          LengthOfStay_Max       4.500     -3
25              0              0                        <NA>       0.000     -1
26             36             37                   sp_reh       4.500     -3
27             38             39                   pg_ads       2.500     -3
28             40             41                 pcg_heart4       0.500     -3
29              0              0                        <NA>       0.000     -1
30             42             43                   pcg_ami       1.500     -3
31             44             45           CharlsonIndex_avg       1.830     -3
32             46             47                 pcg_cancrb       2.500     -3
33              0              0                        <NA>       0.000     -1
34             48             49           CharlsonIndex_Max       1.000     -3
35              0              0                        <NA>       0.000     -1
36             50             51                 pcg_cancrb       4.500     -3
37              0              0                        <NA>       0.000     -1
38             52             53              labcount_max       6.500     -3
39              0              0                        <NA>       0.000     -1
40             54             55                   pcg_msc       0.500     -3
41             56             57              labcount_min      10.000     -3
42             58             59                   ps_urg       7.500     -3
43             60             61                 pcg_cancrb       1.500     -3
136           166            167                   num_PCP       3.500     -3
137           168            169                    ps_out       0.500     -3
138           170            171                 pcg_cancrb       0.500     -3
139           172            173                    sp_dia       2.500     -3
140             0              0                        <NA>       0.000     -1
141             0              0                        <NA>       0.000     -1
142           174            175       num_ProcedureGroup      10.500     -3
143             0              0                        <NA>       0.000     -1
144           176            177              drugcount_min       2.500     -3
145             0              0                        <NA>       0.000     -1
146           178            179                   pcg_fxd       0.500     -3
147             0              0                        <NA>       0.000     -1
148             0              0                        <NA>       0.000     -1
149             0              0                        <NA>       0.000     -1
150             0              0                        <NA>       0.000     -1
151             0              0                        <NA>       0.000     -1
152           180            181          PayDelay_Max      84.500     -3
153             0              0                        <NA>       0.000     -1
154           182            183                   pcg_tra       5.000     -3
155             0              0                        <NA>       0.000     -1
156           184            185                   pcg_perv       1.500     -3
157           186            187                    sp_lab       9.000     -3
158           188            189                    pg_srs       0.500     -3
159           190            191                    sp_ane       0.500     -3
160             0              0                        <NA>       0.000     -1
161           192            193           CharlsonIndex_Max       1.000     -3
162           194            195                    sp_dia       1.500     -3
163           196            197                 pcg_misc15       0.500     -3
164             0              0                        <NA>       0.000     -1
165             0              0                        <NA>       0.000     -1
166           198            199       num_ProcedureGroup       1.500     -3
167             0              0                        <NA>       0.000     -1
168             0              0                        <NA>       0.000     -1
169             0              0                        <NA>       0.000     -1
170           200            201                   ps_urg       1.500     -3
171             0              0                        <NA>       0.000     -1
172           202            203                   pcg_res       0.500     -3
173             0              0                        <NA>       0.000     -1
174             0              0                        <NA>       0.000     -1
175             0              0                        <NA>       0.000     -1
176             0              0                        <NA>       0.000     -1
177             0              0                        <NA>       0.000     -1
178             0              0                        <NA>       0.000     -1
179             0              0                        <NA>       0.000     -1
180             0              0                        <NA>       0.000     -1
116             0              0                        <NA>       0.000     -1
117             0              0                        <NA>       0.000     -1
118             0              0                        <NA>       0.000     -1
119             0              0                        <NA>       0.000     -1
120             0              0                        <NA>       0.000     -1
121             0              0                        <NA>       0.000     -1
122             0              0                        <NA>       0.000     -1
123             0              0                        <NA>       0.000     -1
124             0              0                        <NA>       0.000     -1
125             0              0                        <NA>       0.000     -1
126             0              0                        <NA>       0.000     -1
127             0              0                        <NA>       0.000     -1
128             0              0                        <NA>       0.000     -1
129             0              0                        <NA>       0.000     -1
130             0              0                        <NA>       0.000     -1
131             0              0                        <NA>       0.000     -1
132             0              0                        <NA>       0.000     -1
133             0              0                        <NA>       0.000     -1
134             0              0                        <NA>       0.000     -1
135             0              0                        <NA>       0.000     -1
136             0              0                        <NA>       0.000     -1
137             0              0                        <NA>       0.000     -1
138             0              0                        <NA>       0.000     -1
139             0              0                        <NA>       0.000     -1
140             0              0                        <NA>       0.000     -1
141             0              0                        <NA>       0.000     -1
142             0              0                        <NA>       0.000     -1
143             0              0                        <NA>       0.000     -1
144             0              0                        <NA>       0.000     -1
145             0              0                        <NA>       0.000     -1
146             0              0                        <NA>       0.000     -1
147             0              0                        <NA>       0.000     -1
148             0              0                        <NA>       0.000     -1
149             0              0                        <NA>       0.000     -1
150             0              0                        <NA>       0.000     -1
151             0              0                        <NA>       0.000     -1
152             0              0                        <NA>       0.000     -1
153             0              0                        <NA>       0.000     -1
154             0              0                        <NA>       0.000     -1
155             0              0                        <NA>       0.000     -1
156             0              0                        <NA>       0.000     -1
157             0              0                        <NA>       0.000     -1
158             0              0                        <NA>       0.000     -1
159             0              0                        <NA>       0.000     -1
160             0              0                        <NA>       0.000     -1
161             0              0                        <NA>       0.000     -1
162             0              0                        <NA>       0.000     -1
163             0              0                        <NA>       0.000     -1
164             0              0                        <NA>       0.000     -1
165             0              0                        <NA>       0.000     -1
166             0              0                        <NA>       0.000     -1
167             0              0                        <NA>       0.000     -1
168             0              0                        <NA>       0.000     -1
169             0              0                        <NA>       0.000     -1
170             0              0                        <NA>       0.000     -1
171             0              0                        <NA>       0.000     -1
172             0              0                        <NA>       0.000     -1
173             0              0                        <NA>       0.000     -1
174             0              0                        <NA>       0.000     -1
175             0              0                        <NA>       0.000     -1
176             0              0                        <NA>       0.000     -1
177             0              0                        <NA>       0.000     -1
178             0              0                        <NA>       0.000     -1
179             0              0                        <NA>       0.000     -1
180             0              0                        <NA>       0.000     -1
181             0              0                        <NA>       0.000     -1
182             0              0                        <NA>       0.000     -1
183             0              0                        <NA>       0.000     -1
184             0              0                        <NA>       0.000     -1
185             0              0                        <NA>       0.000     -1
186             0              0                        <NA>       0.000     -1
187             0              0                        <NA>       0.000     -1
188             0              0                        <NA>       0.000     -1
189             0              0                        <NA>       0.000     -1
190             0              0                        <NA>       0.000     -1
191             0              0                        <NA>       0.000     -1
192             0              0                        <NA>       0.000     -1
193             0              0                        <NA>       0.000     -1
194             0              0                        <NA>       0.000     -1
195             0              0                        <NA>       0.000     -1
196             0              0                        <NA>       0.000     -1
197             0              0                        <NA>       0.000     -1
198             0              0                        <NA>       0.000     -1
199             0              0                        <NA>       0.000     -1
200             0              0                        <NA>       0.000     -1
201             0              0                        <NA>       0.000     -1
202             0              0                        <NA>       0.000     -1
203             0              0                        <NA>       0.000     -1
```

c) Random Forest Plot

In the random forest of Model 1, totally, 203 trees were generated. The top three influential factors were *pc_off*, *pcg_art* and *claimsTruncated*. And for the purity of tree nodes, the attributes *claimsTruncated*, *drugcount_avg* and *drugcount_sum* were ranked as first three.

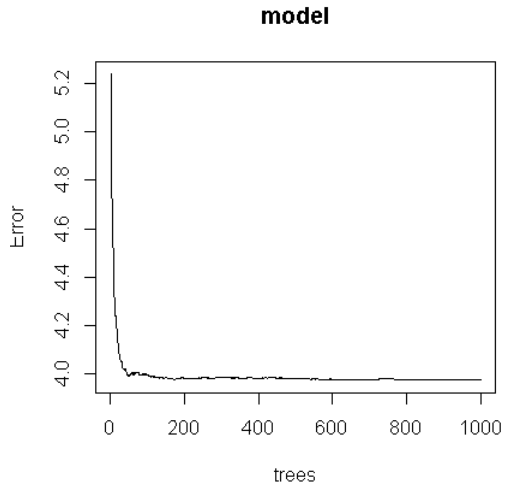
(2) Applying on Predictive Model 2: T1P12 Model

Due to having the same training set, the description of Model 2 was the same as the Model 1.

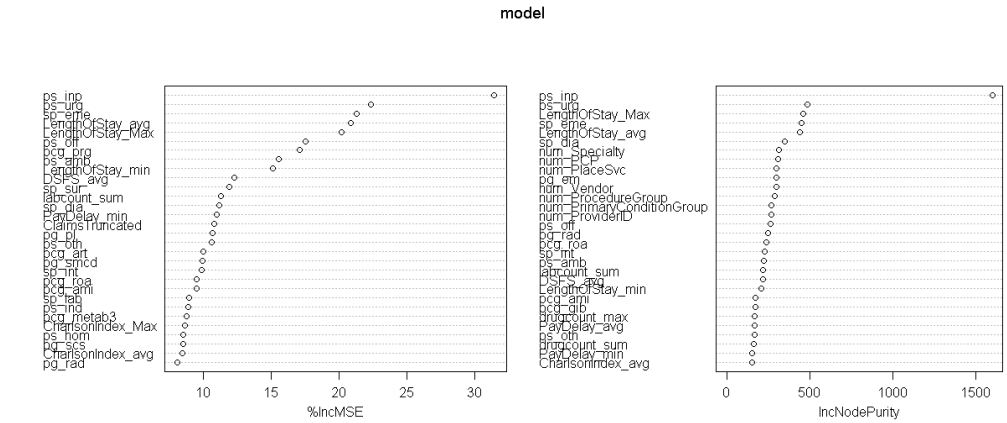
(3) Applying on Predictive Model 3: T12P2 Model

The training set of Model 3 incorporated two years of historical data and kept the same prediction set. The model description was displayed as below (see Figure 7).

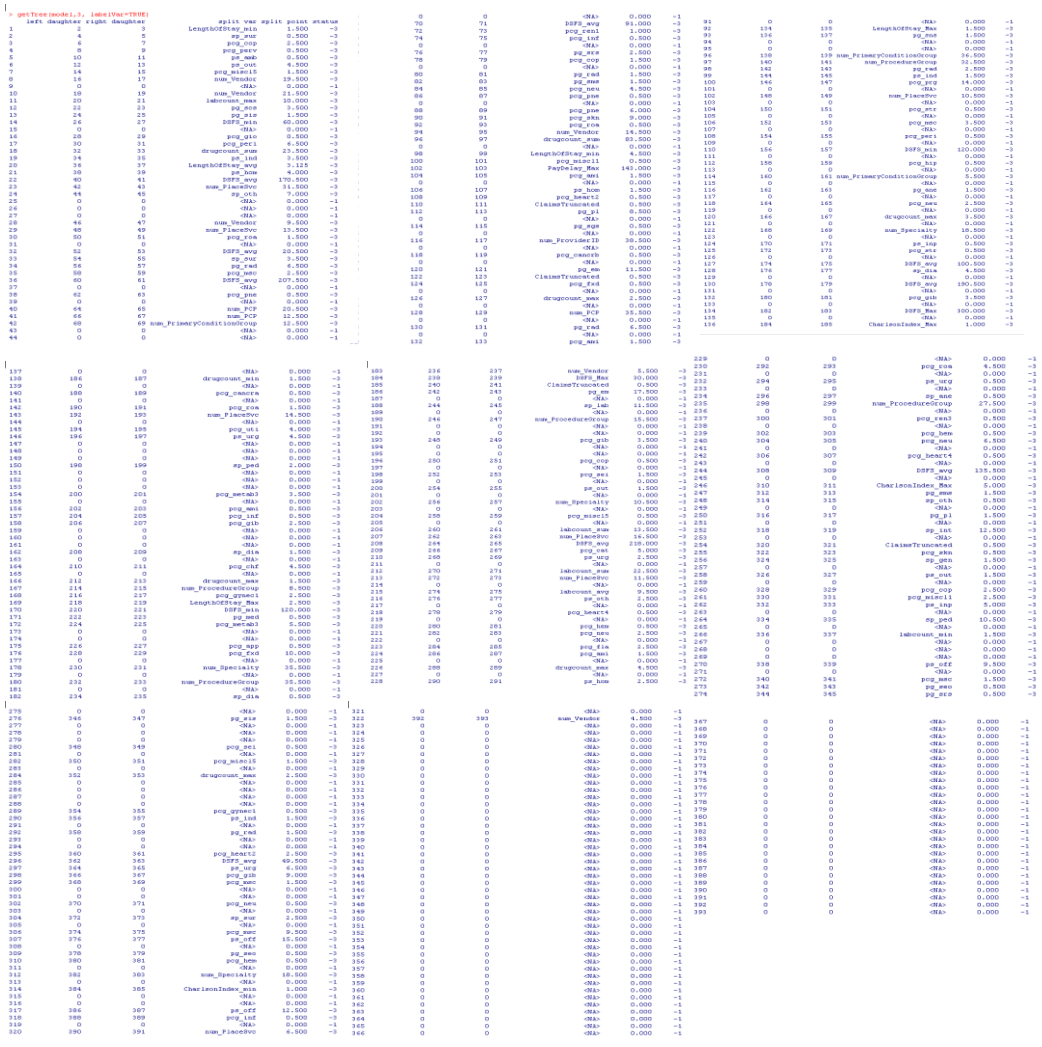
Figure 7 Random Forest Plot for Data Mining Model 3



(a) Model Plot



(b) Variable Importance Plot



(c) Random Forest Plot

In the random forest Model 2, 393 trees were grown. It increased over 100 trees in this model. The top three influential factors were *ps_ing*, *ps_urg* and *sp_eme*. And for the purity of tree nodes, the attributes *ps_ing*, *ps_urg* and *LengthOfStay_Max* were ranked as the first three.

(4) Applying on Predictive Model 4: T12P12 Model

Since the training dataset of Model 4 was the same as the Model 3, the model description was the same also.

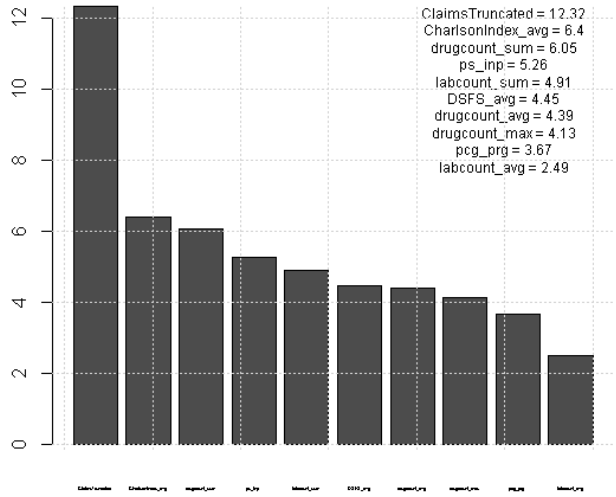
3.3.3 Gradient Boosting

For gradient boosting, parameters setting influence the results performance. The shrinkage is a parameter applied to each tree in the expansion, which was set to 0.05. It is also known as the learning rate or step-size reduction. Distribution is a very important parameter. It is either a character string specifying the name of the distribution to use or a list with a component name specifying the distribution and any additional parameters needed. We specify our distribution model as “gaussian”. N-tree is the total number of trees to fit. This is equivalent to the number of iterations and the number of basic functions in the additive expansion. N-tree was set to 500. Interaction-depth is the maximum depth of variable interactions, which was set to 4. N-minobsinnode is minimum number of observations in the trees terminal nodes and it was set to 50.

(1) Applying on Predictive Model 1: T1P2 Model

Figure 8 told us which factors affect the prediction most. *ClaimsTruncated* gained the high scores as 12.32. And its score doubled the score of second factor, *ChrisonIndex*, 6.4. Then, the score of *Drugcounr_sum*, *ps_ing*, *Labcount_sum*, *DSFS_avg*, *drugcount_avg* and *drugcount_max* were all over 4.

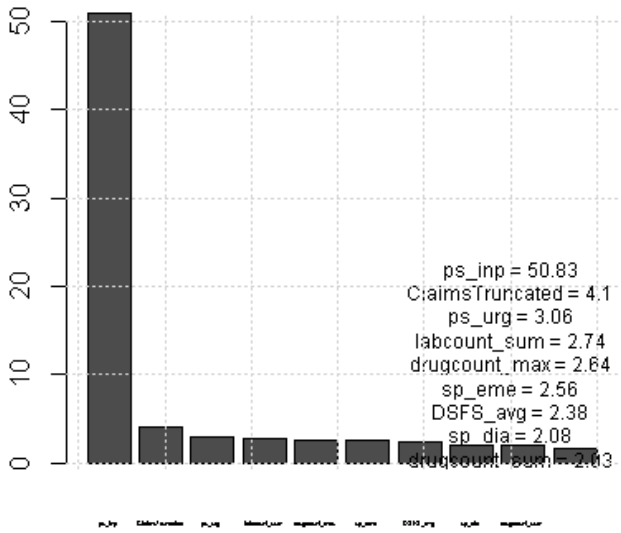
Figure 8 Model 1 Importance Attribute Plot



(2) Applying on Predictive Model 3: T12P2 Model

The gradient boosting Model 3 was quite different from Model 1. In Model 1, besides *ClaimsTruncated*, other factors gained similar scores. However, in Model 3, the differences among each factor were significant (see Figure 9). The most influential factor is *ps_ing* as 50.83 score. The rest of the factors only gained the score around 3.

Figure 9 Mode3 1 Importance Attribute Plot



4 Results Evaluation

In this chapter, we evaluated the results generated from last chapter and analyzed the evaluation results. The evaluation was conducted from two aspects, one was from numeric revaluation, and the other was from classic forecasting results evaluation. Results evaluation is a means to measure which algorithm fits the data characters better and be more excel at revealing the knowledge of the dataset. Thus, it is an important component of the entire research.

4.1 Numeric Evaluation

Before applying the evaluation methods, we provided a general overview of the results. First of all, we compared the range of prediction results for *DayInHospital* for Year3 (see Table 8). The Max, Min and Ave in the table refer to the maximum number, minimum number and average number in each *DayInHospital* prediction results for Year3. Meanwhile, the corresponding value for the actual DIH data of Year2 and Year3 is presented in table 9 in purpose of comparison.

Table 8 Prediction Results General Comparison (unit: days)

	Linear Regression			**Random Forest**			**Gradient Boosting**		
	Max	**Min**	**Ave**	**Max**	**Min**	**Ave**	**Max**	**Min**	**Ave**
Predictive T1P2 Model	8.9	-1.26	0.07	3.5	0	0.07	6.2	-0.56	0.5
Predictive T1P12 Model				4.1	0	0.15			
Predictive T12P2 Model	8.9	-1.26	0.07	4.8	0	0.06	10	-0.76	0.4

Predictive T12P12 Model				5.9	0	0.14			

Table 9 Data summary of DIH data of Year2 and Year3 (unit: days)

	Max	**Min**	**Ave**
DIH data of Year2	15	0	0.47
DIH data of Year3	15	0	0.44

Then, we plotted the actual *DayInHospital* data for Year2 and Year3, and prediction results from four models, in order to compare them visually (see figure 10).

Figure 10 Actual data and Prediction results plots comparison

From the above tables 8 and Figure 10, we can see that:

1. Gradient boosting prediction results had the most similar properties to the actual data.

Its average value and maximum values were closest to the actual data, although it was not very identical seen from the distribution plot. The random forest algorithm failed to predict the large value, but it was the only one algorithm provided no negative values and applied to all the four models.

2. Linear regression also is able to predict large value in the results.

Referring to the linear regression equation, x value was the DIH data for Year3 and a_k is attributes from Member Data and DIH data from past years. Weights were calculated and applied automatically in prediction process. Thus, we can see that each algorithm had its own merits and drawback. Following, we used mathematic evaluation methods to provide a more detail analysis of the results.

4.2 Forecasting Results Evaluation

In Chapter 2, we briefly discussed that we will use Mean Square Error (MSE), Mean Absolute Deviation (MAD) and Mean Absolute Percentage Error (MAPE) to estimate the results. When we looked at the results, we found many negative values and zero values existing. If we directly apply the three statistic evaluation method, the

evaluation results could be inaccurate. Thus, we must modify the evaluation methods to suit our dataset. We examined each method and came up a way to modify it.

For Mean Square Error (MSE), the original expression is $MSE = \frac{\sum_{n=1}^{n}(E)^2}{n}$. We changed it to $MSE = \frac{\sum_{n=1}^{n}(E+1)^2}{n}$, because many results values were between -1 and 1, if square them, the number is too small to calculate and the evaluation will not show big different.

For Mean Absolute Deviation (MAD), the original expression is $MAD = \frac{\sum_{n=1}^{n}|E|}{n}$. Because the absolute values will eliminate the effect of negative values, we did not change this expression.

For Mean Absolute Percentage Error (MAPE), the original expression is $M = \frac{1}{n}\sum_{n=1}^{n}\left|\frac{E}{A}\right|$. Although the concept of MAPE sounds very simple and convincing, it has one major drawback. If there are zero values, there will be a division by zero. In our case, as we noticed that in the actual DIH data of Year3, most of them were value zero. This will cause calculation problem because of the division. Thus, to deal with this problem, we added 1 both to numerator and denominator. The modified expression was $M = \frac{1}{n}\sum_{n=1}^{n}\left|\frac{E+1}{A+1}\right|$.

The evaluation results were summarized as below (see Table 10):

Table 10 Results Evaluation Summary

		MSE	MAD	MAPE
Linear Regression	*Predictive T1P2 Model*	2.849	0.490	86.511
	Predictive T12P2 Model	2.849	0.490	86.512
Random Forest	*Predictive T1P2 Model*	2.820	0.486	86.409
	Predictive T1P12 Model	3.077	0.551	93.954
	Predictive T12P2 Model	2.823	0.486	86.245
	Predictive T12P12 Model	3.054	0.542	92.815
Gradient Boosting	*Predictive T1P2 Model*	3.669	0.792	124.855
	Predictive T12P2 Model	3.623	0.723	115.411

The largest MSE was gained by gradient boosting for model T1P2 as 3.669; while the smallest MSE was gained by random forest for model T1P2, 2.820. The largest MAD was also gained by gradient boosting for model T1P2, 0.792; while the smallest MAD was gained by random forest for model T1P2 and model T12P2, 0.486. For MAPE, the largest value was also gained by gradient boosting for model T1P2, 124.855; the smallest value was gained by random forest for model T12P2, 86.245.

From the above data, we concluded that:

1. Random forest generally provided the best results.

It performed best in model T1P2 and model T12P2. We deemed that the random forest was the most accurate algorithm to handle the large dataset. The reason is that random forests algorithm is not as sensitive as others to the number of attributes selected for consideration at each split. And the accuracy of a random forest depends on the strength of the individual classifiers and a measure of the dependence between them as well. Moreover, only random forest was implemented by all four models, whereas the other two algorithms only implemented in the first and third model. The reason could be that, in model T1P12 and model T12P12, the prediction set involved two year historical member data and may demand too much space for process.

2. The worst model prediction result is from Gradient Boosting, especially for model T1P2.

The reason is that, in model T1P12 and model T12P12, the prediction set involved two year historical member data and may demand too much space for process.

3. The average error percentage of the experiments is high.

The primary possible reason is the missing value and zero value take large portion of the original dataset. They affected the accuracy of the results value. The second reason is due into the nature of medical data. Because we were lack of deep understand of attributes in the dataset, we were not able to identify the most relevant attributes. Thus, when too many attributes involved in the data mining, some less relevant attributes may affect the model accuracy, and therefore, affect the results.

5 Conclusions and Future Work

In this research, we used data mining techniques to address prediction of hospitalization period of patients. The contributions of the research are: first, we demonstrated the importance of building appropriate predictive models, which is neglected in the work of many. Most data mining research focuses on algorithm instead of data modeling. However, in our research, we explored several feasible predictive models and used the results from different models to find the most accurate algorithm; second, we applied three data mining techniques: linear regression, random forest and gradient boosting, in our research. Each algorithm provided different accuracy with each model and reflected the inherent properties of the algorithm. The conclusion from our research is that the random forest techniques are the best techniques of prediction patient hospitalization periods with this dataset.

The historical dataset we used had 112 attributes (e.g. Memberid, num_ProviderID). Some of those are may be relatively unimportant to the prediction of hospitalization period (e.g. Memberid). In the future, we might explore the segmentation of the attributes into different classes, and then only use the more important attributes in our prediction techniques. In addition, we could explore the addition of other (new) attributes such as sex and age.

Bibliography

Bloomberg. (2013). *Bloomberg Visual Data: Most efficient health care: Countries.* Retrieved from http://www.bloomberg.com/visual-data/best-and-worst/most-efficient-health-care-countries

Breiman, L. (2001). Random forests. *Machine Learning*.

Davidson, K. A. (2013). *The Most Efficient Health Care Systems In The World.* The Huffington Post.

Dursun Delen, G. W. (2004). Predicting breast cancer survivability: a comparison of three data mining methods. *Artificial Intelligence in Medicine*.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statictics*, 1189-1232.

Heritage Health Prize. (n.d.). Retrieved from http://www.heritagehealthprize.com/c/hhp

Ian H.Witten, E. F. (2011). *Data Mining:Practical Machine Learning Tools and Techiniques.* Morgan Kaufmann.

Jiawei Han, M. K. (2012). *Data Mining:Concepts and Techiniques.* Morgan Kaufmann Publishers is an imprint of Elsevier.

Katharina Morik, P. B. (1999). Combining statistical learning with a knowledge-based approach - a case study in intensive care monitoring.

L. Breiman, J. F. (1984). *Classification and regression trees.* Wadsworth, Belmont.

Mestrom, W. (2011). *My milestone 1 solution to the Heritage Health Prize.*

Perry, M. J. (2009). *Almost 4 Out of 10 Uninsured Americans Live in Households Making More Than $50,000 Per Year*. Retrieved from http://mjperry.blogspot.com/2009/09/383-of-uninsured-americans-live-in.html#sthash.R7qwAWRV.dpu.

Perry, M. J. (n.d.). *Almost 4 Out of 10 Uninsured Americans Live in Households Making More Than $50,000 Per Year*. Retrieved from http://mjperry.blogspot.com/2009/09/383-of-uninsured-americans-live-in.html#sthash.R7qwAWRV.dpu.

Schapire, R. (2002). The boosting approach to machine learning – an overview. *MSRI Workshop on Nonlinear Estimation and Classification.* Springer.

Appendices

A. Feature Generation (Use one-year-history model as example):

A.1 Extract claims_per_member

SELECT year, Memberid,

Count(ProviderID) AS num_ProviderID, Count(Vendor) AS num_Vendor, Count(PCP) AS num_PCP, Count(Specialty) AS num_Specialty, Count(PlaceSvc) AS num_PlaceSvc, Count(PrimaryConditionGroup) AS num_PrimaryConditionGroup,Count(ProcedureGroup) AS num_ProcedureGroup,

sum(IIF(Specialty = 'Anesthesiology', 1, 0)) AS sp_ane, sum(IIF(Specialty = 'Obstertrics and Gynecology', 1, 0)) AS sp_obs, sum(IIF(Specialty = 'Diagnostic Imaging', 1, 0)) AS sp_dia, sum(IIF(Specialty = 'Emergency', 1, 0)) AS sp_eme, sum(IIF(Specialty = 'General Practice', 1, 0)) AS sp_gen, sum(IIF(Specialty = 'Internal', 1, 0)) AS sp_int, sum(IIF(Specialty = 'Laboratory', 1, 0)) AS sp_lab, sum(IIF(Specialty = 'Other', 1, 0)) AS sp_oth, sum(IIF(Specialty = 'Pathology', 1, 0)) AS sp_pat, sum(IIF(Specialty = 'Pediatrics', 1, 0)) AS sp_ped, sum(IIF(Specialty = 'Rehabilitation', 1, 0)) AS sp_reh, sum(IIF(Specialty = 'Surgery', 1, 0)) AS sp_sur,

sum(IIF(PlaceSvc = 'Ambulance', 1, 0)) AS ps_amb, sum(IIF(PlaceSvc = 'Home', 1, 0)) AS ps_hom, sum(IIF(PlaceSvc = 'Independent Lab', 1, 0)) AS ps_ind, sum(IIF(PlaceSvc = 'Inpatient Hospital', 1, 0)) AS ps_inp, sum(IIF(PlaceSvc = 'Office', 1, 0)) AS ps_off, sum(IIF(PlaceSvc = 'Other', 1, 0)) AS ps_oth, sum(IIF(PlaceSvc = 'Outpatient Hospital', 1, 0)) AS ps_out, sum(IIF(PlaceSvc = 'Urgent Care', 1, 0)) AS ps_urg,

sum(IIF(PrimaryConditionGroup = 'AMI', 1, 0)) AS pcg_ami, sum(IIF(PrimaryConditionGroup = 'APPCHOL', 1, 0)) AS pcg_app, sum(IIF(PrimaryConditionGroup = 'ARTHSPIN', 1, 0)) AS pcg_art,

```
sum(IIF(PrimaryConditionGroup = 'CANCRA', 1, 0)) AS pcg_cancra,
sum(IIF(PrimaryConditionGroup = 'CANCRB', 1, 0)) AS pcg_cancrb,
sum(IIF(PrimaryConditionGroup = 'CANCRM', 1, 0)) AS pcg_cancrm,
sum(IIF(PrimaryConditionGroup = 'CATAST', 1, 0)) AS pcg_cat,
sum(IIF(PrimaryConditionGroup = 'CHF', 1, 0)) AS pcg_chf,
sum(IIF(PrimaryConditionGroup = 'COPD', 1, 0)) AS pcg_cop,
sum(IIF(PrimaryConditionGroup = 'FLaELEC', 1, 0)) AS pcg_fla,
sum(IIF(PrimaryConditionGroup = 'FXDISLC', 1, 0)) AS pcg_fxd,
sum(IIF(PrimaryConditionGroup = 'GIBLEED', 1, 0)) AS pcg_gib,
sum(IIF(PrimaryConditionGroup = 'GIOBSENT', 1, 0)) AS pcg_gio,
sum(IIF(PrimaryConditionGroup = 'GYNEC1', 1, 0)) AS pcg_gynec1,
sum(IIF(PrimaryConditionGroup = 'GYNECA', 1, 0)) AS pcg_gyneca,
sum(IIF(PrimaryConditionGroup = 'HEART2', 1, 0)) AS pcg_heart2,
sum(IIF(PrimaryConditionGroup = 'HEART4', 1, 0)) AS pcg_heart4,
sum(IIF(PrimaryConditionGroup = 'HEMTOL', 1, 0)) AS pcg_hem,
sum(IIF(PrimaryConditionGroup = 'HIPFX', 1, 0)) AS pcg_hip,
sum(IIF(PrimaryConditionGroup = 'INFEC4', 1, 0)) AS pcg_inf,
sum(IIF(PrimaryConditionGroup = 'LIVERDZ', 1, 0)) AS pcg_liv,
sum(IIF(PrimaryConditionGroup = 'METAB1', 1, 0)) AS pcg_metab1,
sum(IIF(PrimaryConditionGroup = 'METAB3', 1, 0)) AS pcg_metab3,
sum(IIF(PrimaryConditionGroup = 'MISCHRT', 1, 0)) AS pcg_mis,
sum(IIF(PrimaryConditionGroup = 'MISCL1', 1, 0)) AS pcg_miscl1,
sum(IIF(PrimaryConditionGroup = 'MISCL5', 1, 0)) AS pcg_miscl5,
sum(IIF(PrimaryConditionGroup = 'MSC2a3', 1, 0)) AS pcg_msc,
sum(IIF(PrimaryConditionGroup = 'NEUMENT', 1, 0)) AS pcg_neu,
sum(IIF(PrimaryConditionGroup = 'ODaBNCA', 1, 0)) AS pcg_oda,
sum(IIF(PrimaryConditionGroup = 'PERINTL', 1, 0)) AS pcg_peri,
sum(IIF(PrimaryConditionGroup = 'PERVALV', 1, 0)) AS pcg_perv,
sum(IIF(PrimaryConditionGroup = 'PNCRDZ', 1, 0)) AS pcg_pnc,
sum(IIF(PrimaryConditionGroup = 'PNEUM', 1, 0)) AS pcg_pne,
sum(IIF(PrimaryConditionGroup = 'PRGNCY', 1, 0)) AS pcg_prg,
sum(IIF(PrimaryConditionGroup = 'RENAL1', 1, 0)) AS pcg_ren1,
```

```
sum(IIF(PrimaryConditionGroup = 'RENAL2', 1, 0)) AS pcg_ren2,
sum(IIF(PrimaryConditionGroup = 'RENAL3', 1, 0)) AS pcg_ren3,
sum(IIF(PrimaryConditionGroup = 'RESPR4', 1, 0)) AS pcg_res,
sum(IIF(PrimaryConditionGroup = 'ROAMI', 1, 0)) AS pcg_roa,
sum(IIF(PrimaryConditionGroup = 'SEIZURE', 1, 0)) AS pcg_sei,
sum(IIF(PrimaryConditionGroup = 'SEPSIS', 1, 0)) AS pcg_sep,
sum(IIF(PrimaryConditionGroup = 'SKNAUT', 1, 0)) AS pcg_skn,
sum(IIF(PrimaryConditionGroup = 'STROKE', 1, 0)) AS pcg_str,
sum(IIF(PrimaryConditionGroup = 'TRAUMA', 1, 0)) AS pcg_tra,
sum(IIF(PrimaryConditionGroup = 'UTI', 1, 0)) AS pcg_uti,

sum(IIF(ProcedureGroup = 'ANES', 1, 0)) AS pg_ane, sum(IIF(ProcedureGroup = 'EM', 1, 0)) AS pg_em, sum(IIF(ProcedureGroup = 'MED', 1, 0)) AS pg_med, sum(IIF(ProcedureGroup = 'PL', 1, 0)) AS pg_pl, sum(IIF(ProcedureGroup = 'RAD', 1, 0)) AS pg_rad, sum(IIF(ProcedureGroup = 'SAS', 1, 0)) AS pg_sas, sum(IIF(ProcedureGroup = 'SCS', 1, 0)) AS pg_scs, sum(IIF(ProcedureGroup = 'SDS', 1, 0)) AS pg_sds, sum(IIF(ProcedureGroup = 'SEOA', 1, 0)) AS pg_seo, sum(IIF(ProcedureGroup = 'SGS', 1, 0)) AS pg_sgs, sum(IIF(ProcedureGroup = 'SIS', 1, 0)) AS pg_sis, sum(IIF(ProcedureGroup = 'SMCD', 1, 0)) AS pg_smcd, sum(IIF(ProcedureGroup = 'SMS', 1, 0)) AS pg_sms, sum(IIF(ProcedureGroup = 'SNS', 1, 0)) AS pg_sns, sum(IIF(ProcedureGroup = 'SO', 1, 0)) AS pg_so, sum(IIF(ProcedureGroup = 'SRS', 1, 0)) AS pg_srs, sum(IIF(ProcedureGroup = 'SUS', 1, 0)) AS pg_sus,

max(PayDelay) AS PayDelay_Max, min(PayDelay) AS PayDelay_min, avg(PayDelay) AS PayDelay_avg, max(LengthOfStay) AS LengthOfStay_Max, min(LengthOfStay) AS LengthOfStay_min, avg(LengthOfStay) AS LengthOfStay_avg, max(DSFS) AS DSFS_Max, min(DSFS) AS DSFS_min, avg(DSFS) AS DSFS_avg, max(CharlsonIndex) AS CharlsonIndex_Max, min(CharlsonIndex) AS CharlsonIndex_min, avg(CharlsonIndex) AS CharlsonIndex_avg

Into claims_per_member

FROM Claims
```

group by year, Memberid;

A.2 Extract drugcount_per_member

SELECT MemberID AS MemberID_dc, Year AS Year_dc, Max(drugcount) AS drugcount_max, Min(drugcount) AS drugcount_min, Avg(drugcount) AS drugcount_avg, sum(drugcount) as drugcount_sum

INTO DrugCount_summary

FROM DrugCount

GROUP BY MemberID, year;

A.3 Extract labcount_per_member

SELECT MemberID AS MemberID_dc, Year AS Year_dc, Max(labcount) AS labcount_max, Min(labcount) AS labcount_min, Avg(labcount) AS labcount_avg, sum(labcount) as labcount_sum

INTO LabCount_summary

FROM LabCount

GROUP BY MemberID, year;

A.4 Form training model table

SELECT a.*, b.* INTO ClaimsFromY1

FROM Y2Target AS a LEFT JOIN claims_per_member AS b ON (a.year=b.year) AND (a.memberid=b.memberid);

claims from Y2:

SELECT a.*, b.* INTO ClaimsanddcFromY1

FROM claimsfromY1 AS a LEFT JOIN drugcount_summary AS b ON (a.a_memberid=b.memberid_dc) AND (a.a_year=b.year_dc);

SELECT a.*, b.* INTO ClaimsanddcandlcFromY1

FROM ClaimsanddcFromY1 AS a LEFT JOIN labcount_summary AS b ON (a.a_memberid=b.memberid_dc) AND (a.a_year=b.year_dc);

A.5 Form prediction table

SELECT a.*, b.* INTO ClaimsFromY2

FROM Y3Target AS a LEFT JOIN claims_per_member AS b ON (a.memberid=b.memberid) AND (a.year=b.year);

SELECT a.*, b.* INTO ClaimsanddcFromY2

FROM claimsfromY2 AS a LEFT JOIN drugcount_summary AS b ON (a.a_memberid=b.memberid_dc) AND (a.a_year=b.year_dc);

SELECT a.*, b.* INTO ClaimsanddcandlcFromY2

FROM ClaimsanddcFromY2 AS a LEFT JOIN labcount_summary AS b ON (a.a_memberid=b.memberid_dc) AND (a.a_year=b.year_dc);