Healthcare Analytics Using Texas Hospital Inpatient Discharge Public Use Data File

Sowmiyaa Chakravarthy Vasan
Master of Applied Computing
University of Windsor
Windsor, ON
chakr113@uwindsor.ca

Shailja Gupta
Master of Applied Computing
University of Windsor
Windsor, ON
gupta14p@uwindsor.ca

Disha Amrish Gajjar
Master of Applied Computing
University of Windsor
Windsor, ON
gajjard@uwindsor.ca

***Abstract*—We present a data mining model for healthcare analysis using Texas Hospital Inpatient Discharge Public Use Data File, where are catogorize patients suffering from HIV and drug abuse on the basis of severity of illness using factors such as spec_unit and source of admission.**

I. Introduction

Improving care before treatment is one of the best ways of introducing quality to healthcare. So we present a data mining model for healthcare analysis using Texas Hospital Inpatient Discharge Public Use Data File. In this project, we are Classify patients suffering from HIV and drug abuse on the basis of severity of illness using factors such as spec_unit and source of admission.

II. Motivation

A. Business problem that led to the idea

- There is an observed mismatch between the resources and the demand which is one of the clinical challenges faced today One such resource is doctor's time that needs to be efficiently utilized to improve healthcare
- Clinical efficiency can be improved indirectly by reducing the number of patients in the healthcare that will also result in better population health
- Sometimes patients who doesn't need immediate care are given full access to resources and sometimes patients who are in need of immediate care are delayed
- Healthcare industry has always faced this issue where patients under special category or a condition fall out of care due to their underlying condition adding to their illness or disease
- Hence, clinical efficiency is lost in such failures of retaining care for HIV or drug abuse patients

III. Existing model *Vs* Our model

There are many models available for healthcare analysis. We followed the model/models that follow: Data collection, cleaning, analysis and regression analysis.

- The existing models fall under the general field of "data mining", whereas among the different pathways that are defined under healthcare analytics, our data mining problem falls under "right care" pathway to be provided to the patients
- In the existing models, the focus is solely on Lateness and how Lateness is associated with various factors
- They took variable such as time of appointment , arrival time, late time. However, in our model we have chosen 14 variables

IV. Idea of our model

A. Details of data mining problem

The current objective in healthcare is to have a healthy population over providing better treatment. Improving care before treatment is one of the best ways of introducing quality to healthcare. There is an observed mismatch between the resources and the demand which is one of the clinical challenges faced today. One such resource is doctor's time that needs to be efficiently utilized to improve healthcare. Clinical efficiency can be improved indirectly by reducing the number of patients in the healthcare that will also result in better population health.

Healthcare industry has a varied patient group and it is required by them to service the needs of all kinds of patients. Healthcare for certain groups such as patients with disability, HIV, substance abuse, chronic diseases require special care and extended services. It is important to study and use analytics in this domain to be prepared and to provide the best service without any gap in the process.

In this project, we are considering group of patients suffering from HIV or diagnosed with drug/substance abuse and how analysis of healthcare data can help us point to services that can be improved. In our data mining problem, we focus on patient profiling by grouping patients with HIV and drug abuse condition and they would require special needs and care while being treated for the illness they have been admitted for. Due to their underlying condition, these groups could develop certain side-effects or illness physically or mentally that can be addressed if aware of. This way, it improves clinical efficiency as analysis and study would help in providing right treatment based on their condition and side-effects. This in turn cuts down the patient count for each doctor that increases their efficiency too. Further, admin of healthcare organizations can plan their resources based on the study in order to accommodate such patients as quickly as possible and at an affordable price.

From an admin's view, this project aims to improve healthcare experience for these targeted group (HIV and drug abuse patients) by increasing the special rooms as per demand to provide better care. Also, help in providing pain free services based on their type of illness with pre-existing HIV or drug abuse disorders. Further, as admin it would be a good idea to come up with programs and department that provide counselling for drug abuse patients in order to prepare them towards a healthy living.

Having more information about the patient suffering from HIV and drug abuse order will help the administration of the hospital to improve clinical efficiency. There are different units in hospitals such as Coronary Care Unit, Pediatric Unit, Rehabilitation Unit etc which are recommended to patients depending on their illness. The utmost goal of healthcare industry is to provide care and comfort to the patient while on treatment. On similar lines, as quoted in our data mining problem, patients with HIV or drug abuse condition would require extra care when they come in for other illnesses. Hence, the issue in the healthcare industry to provide better services to specially categorized patients such as HIV or drug abuse patients can be improved by mining into data to predict better services to them. This could also help clinics and admins to understand the resource utilization and predict future allocation based on current utilization trend. It allows clinics to be prepared and to provide better care based on the illness of such patients considering their underlying medical conditions.

Hence, among the different pathways that are defined under healthcare analytics, our data mining problem falls under "right care" pathway to be provided to the patients. Data mining problem also address client's risk for retention in care failure before client falls out of care. Healthcare industry has always faced this issue where patients under special category or a condition fall out of care due to their underlying condition adding to their illness or disease. Hence, clinical efficiency is lost in such failures of retaining care for HIV or drug abuse patients.

B. Data mining process for classification of HIV/drug abuse patients

The data mining problem is worked upon by following the data mining steps. We selected the data for HIV/drug abuse patients at an earlier stage to have a focused dataset to solve our business problem. The data mining problem is to classify a new patient into one of the specialized units in the hospital, that forms our target variable. As we classify patient based on certain parameters, they behave as input variables such as illness, ethnicity, age, type of admission etc. On identifying the required variable for our data mining problem, we selected appropriate data and performed dimension reduction and other visualization techniques to clean, transform and validate the dataset required for classification. Further, we partitioned the data to build a model using training set and to test the accuracy of the model using validation set.

The variables used in the data to classify patients are mostly categorical. Also, the target variable being categorical, the model built in this report is logistic. The additional advantage of logistic regression to easily learn from the training set makes it easy to implement.

V. Steps to construct the model

A. Data Exploration

The THCIC data is cleaned and examined for the following variables that are required for our data mining problem. Based on the identified target variable, required input variables or predicted variables are selected in this part of data mining.

Predictor Variable: Predictor variables are mapped to the target variable through an empirical relationship. They can be categorical, continuous or integer. Predictions can be of three types: decisions, rankings and estimates.

Target Variable: The target variable is the one for which we need the output or classification hence in this case our target variable is SPEC_UNIT_1.

B. Data Identification

We have identified 14 variables that can be used as predictors for our target variable. These predictors can be used to create models that will help us in our predictions. Once we analyse the given data, we need to first identify the data that is useful for us. Hence, we need to reduce our data from 194 columns to 15 columns including the target variable. This reduction in the number of columns is done as the other columns are of no value while classifying the Specialty units, they just increase the volume of data for our business case.Thus, we select those 15 variables and create another sheet of data that does not have any unnecessary columns.

Once we are done with columns, we need to work on reducing the number of rows by identifying the data that we need. Here, we need the data for patients who are 'HIV and drug/alcohol use patients. As we analyse the column PAT_AGE (Patient age) we see that the values 22-26 for that variable specifies these patients. Hence, using this criterion we filter out the data and utilize it for further tasks. Hence after this step our data reduces from 719,371 rows to 52,146 rows. This is done as our classification is based on a subset of patients having a specific condition associated with them. If we include other data, it can lead to wrong predictions as well.

C. Data Cleaning

In this step we must identify the anomalies present in our data. For this we need to check each variable that is involved in our analysis. Hence, we observe that there are not many trash values in each column, but there are blanks and the symbol ['] which need to be filtered out and removed.

This step is needed because having blanks and symbols will cause hindrance in our classification as they add errors to

our analysis. Also, null values specify nothing which means assuming them as zero also is not correct. After we are done with this step for all variables, our number of rows reduce from 52,146 rows to 32529 rows.
Now we try to understand the relationship between predictor and target variables.

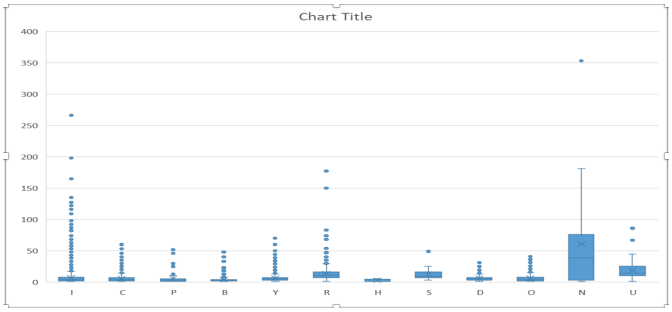

Fig. 1. Relationship between predictor and target variable.

In the above graph (Figure 1) we check the relation between predictor LENGTH_OF_STAY and SPEC_UNIT. As we see the box plot, we can identify the relation between the type of specialty unit and the length of stay. The values are distributed around the median quite uniformly for such a large data set. Even the number of outliers is very less. This shows that using length of stay as a predictor will enhance our analysis. Usage of such variables help us reduce average error in case of our predictions.

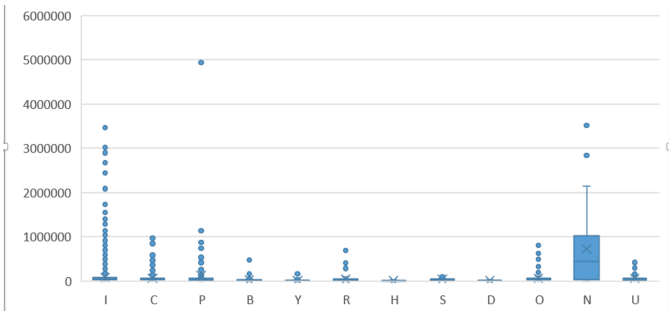

Fig. 2. Box plot - predictor and target variable.

In this graph (Figure 2), we create a box plot to understand the relation between SPEC_UNIT and TOTAL_CHARGES. As we know, that in general cases our charge or cost of treatment depends on the disease. Also, the specialty unit is specific for specific type of disease. Hence, the cost and specialty unit are directly linked to each other. This assumption of ours is proven true by the graph above where the distribution box for the cost is similar for similar specialty units. Even in this case the number of outliers is less. Apart from this let us understand the relation between multiple predictors as well.

The graph (Figure 3) is a stacked bar plot for variables SOURCE_OF_ADMISSION and TYPE_OF_ADMISSION. Similar columns tend to capture similar information which

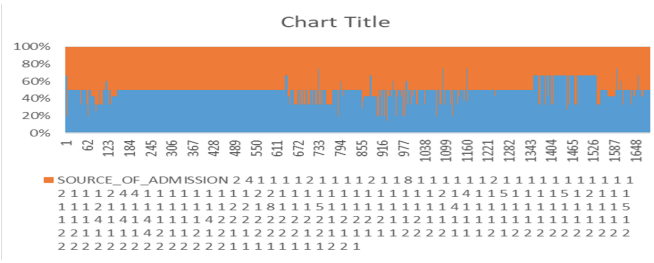

Fig. 3. Stacked bar plot - predictor and target variable.

can lead to multi collinearity. This must be avoided for which we can use just one column instead of both.

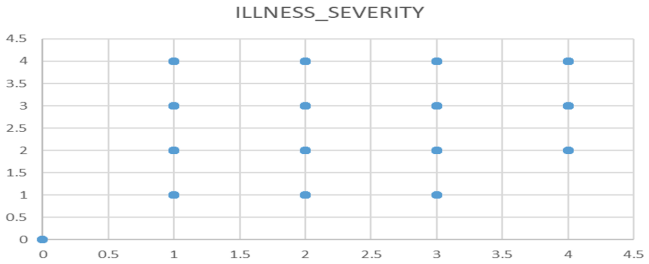

Fig. 4. Illness vs Risk Morality.

It is the similar case with variables ILLNESS_SEVERITY and RISK_MORTALITY. They capture similar information hence we can eliminate either of these. Apart from this we have a lot of categorical variables that need to be transformed to perform further analysis for upcoming reports.

D. Data Import

The primary step for running our analysis is to import the data in to SAS Enterprise Miner. In this case, we need to identify the variables that are predictors for our analysis and the variable that is the target variable. Since we have already identified those in our data exploration step, all we need to do is to set them as they are in this step using the 'Edit variables' option in the SAS Enterprise Miner from the 'File Import' node.
After the node is run successfully, we get the dialog box as shown below which indicates the file has been successfully imported. When we click on results, we can view the summary of the data that has been imported on to the tool.

E. Fixing Issues with Data

As we have seen above, we have already cleaned up the data by removing invalid data (nulls or symbol or any other aberration), after which we imported the files in the tool, now we need to transform it for our use. The data taken into consideration has many limitations on how it can be manipulated and assessed. This is mainly because the data type of each variables differs. The data therefore first needs to be checked if it is numerical or text, continuous, integer or

categorical. This will help us realize what sort of operations need to be performed on them to transform them into a usable format. Say for example, one of the fields is categorical and has strings stored in them. These strings will cause an error if used directly in the logistic regression algorithm as it can only take numeric values. Thus, we need to transform these variables into a form which can be understood by the algorithm, which leads to the creation of dummy variables. These dummy variables are a way of bridging this gap between the data and the model.

Hence, in our data we see that other than 3 interval variables which continuous numeric values. These variables can stay as is since they will be interpreted by the algorithm in a normal manner. Other than these 3, we have all the other 10 variables as categorical. All these variables are necessary for our analysis since they can help us classify our target variable which is again a categorical field. Thus, these variables need to be manipulated in such a way that they can be useful as predictor variables. For this, we will add another node to our diagram which is the 'Transform variables' node. This node takes our clean data as an input and creates dummy variables for all the variables that are categorical. Creating of dummy variables creates n-1 columns for 1 variable which has n categories. This is a transformation of text value to numeric form which is identifiable by the algorithm. Hence, after we run this node, all the 10 variables will have dummy variables created for them. We can click on results to see the output of this node.

F. Creating a Modelset

In the case of supervised algorithm, it means that the algorithm learns through the data and then apply the pattern it understood to the data for which we need to classify our outcomes. The data that we have for this contains 32,529 rows in the table. This data can be split into three parts where one part can be used by the algorithm to learn the pattern of the data, the second part can be used to validate this pattern that the algorithm learned and check if it gives us appropriate results that is with error rate which is tolerable for our use and the third part of the data can be used to test the algorithm. This third dataset is helpful in the case where there are multiple algorithms run and we need to select the best possible model and then run it on this test data set. When an algorithm learns from a training data set, and it applies this model to validation data set it might seem accurate, but this might not necessarily mean that it is correct, hence to learn the characteristic of a data set wholly we need a separate test data. Hence, the three parts of our data will be training data, validation data and test data.

Here in this case, we have divided our data into three parts where 40 percent of the data is training data, 30 percent of data is validation data and the last part of 30 percent of data is test data set on which the model is applied to verify if the algorithm is working accurately. Here what happens is that the algorithm will use the training data set and understand how the target variable varies in accordance with changes in predictor variables. Then it will apply this pattern on our validation data set and check how correctly it is able to classify the records, we can check error percentage for this. After this it again applies this pattern to a separate test data to confirm if the model is working properly.

VI. The Build

The model that we have chosen for our data is Logistic regression. We have chosen this model as our output or target variable is a categorical variable and we are actually classifying our outcomes in this scenario. There are other methods that can do this as well but they have some shortcomings that are overcome using this algorithm, like in k-Nearest Neighbor, the algorithm doesn't learn from the data, it just assigns the variable to nearest cluster. Hence as we see below, I have added a node of Logistic 'Regression'. As we run this, we can generate our results for the data that we are working on.

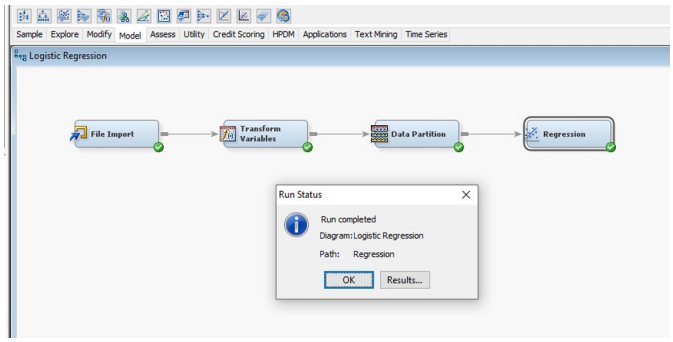

Fig. 5. Tool Run

VII. Assessment and Result

As we click on the results button, we will be able to see the result that are produced by this algorithm. Hence, below we see the output that has been generated. Lift Chart: Lift Chart here shows that the training and validation data set shoe somewhat similar trend in terms of classification. There are places where there is error or deviation, but this seems to be tolerable.

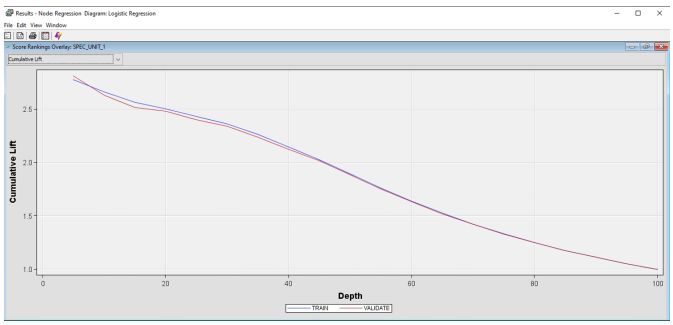

Fig. 6. Result

Fit Statistics: As we look at the fit statistics, we mainly check for the Root Mean Squared Error to check our outcome.

Hence, we see that the value for training data is 0.1795, for validation data it is 0.1814 and for the test data it is 0.1811. Thus, we can say that the model has performed consistently with all the datasets indicating that our result has no issues in terms of data pattern being analysed. Also, the low value of the errors indicates that even the classification has been accurate.

VIII. Conclusion

With this model we can classify patients suffering from HIV or diagnosed with drug/substance abuse and how our analysis of healthcare data can help us point to services that can be improved. The 13 predictor variable can help us understand the requirement of the patients and what kind of Special unit (Target Variable) should be given to them. With the historical data, we generated some test data which can be used by the admin to plan the resource allocation. Model has performed consistently with all datasets indicating that our resutls has no issues in terms of data pattern being analysed. We also obtained low value of errors indicating an accurate classification.