Paper 3254-2015

Predicting Readmission of Diabetic Patients using the high performance Support Vector Machine algorithm of SAS® Enterprise Miner™

Hephzibah Munnangi, MS, Dr. Goutam Chakraborty
Oklahoma State University, Stillwater, Ok

ABSTRACT

Diabetes is a chronic condition affecting people of all ages and is prevalent in around 25.8 million people in the U.S. The objective of this research is to predict the probability of a diabetic patient being readmitted. The results from this research will help hospitals design a follow-up protocol to ensure that patients having a higher re-admission probability are doing well to promote a healthy doctor-patient relationship. The data was obtained from the Center for Machine Learning and Intelligent Systems at University of California, Irvine. The data set contains over 100,000 instances and 55 variables such as insulin and length of stay, etc. The data set was split into training and validation to provide an honest assessment of models. Several variable selection techniques such as Stepwise Regression, Forward Regression, LARS, and LASSO were used. Using LARS, prominent factors were identified in determining the patient readmission rate. Numerous predictive models were built: Decision Tree, Logistic Regression, Gradient Boosting, MBR, SVM, and others. The model comparison algorithm in SAS® Enterprise Miner 13.1 identified that the High Performance Support Vector Machine outperformed the other models, having the lowest misclassification rate of 0.363. The selected model has a sensitivity of 49.7% and a specificity of 75.1% in the validation data.

INTRODUCTION

Data related activities in healthcare management consist of data collection, data sharing, and data analytics. Of those three parts, data collection and data sharing have been successfully achieved by implementing various technologies suitable for collecting and sharing health care records. One of the prominent dimensions in health care is to design services and manage resources by understanding patient conditions. Such situations may include not only chronic conditions a patient might be suffering from, but also external factors such as mode of admission into the health care facility, length of stay in the hospital, patient age, gender, ethnicity, and so on. Further foray into analytics has provided tremendous advantages that assist in designing follow-up services specifically catering to chronic illnesses, thereby reducing costs significantly and improving health care services.

Several measures have been taken by the Centers for Medicare & Medicaid Services (CMS) to reduce hospital readmission rates and improve the quality of healthcare in the United States. According to CMS, one of the major issues faced by our healthcare system is avoidable readmissions which cost the Medicare approximately $15 billion per year. On average, Medicare spends double the amount for an episode with one readmission. In order to handle this situation, hospitals are required to publicly report their re-admission rates and those with high readmissions have to pay a re-imbursement penalty as well. Conditions with highest readmission rates were acute myocardial infraction, heart failure, pneumonia and diabetes. Everybody is different and so are their medical needs. A standard umbrella prescription of diet, regimen, and medication may not be suitable for patients with such different conditions. A custom approach for each medical condition is required to identify patients who have higher likelihood of being readmitted [11].

Diabetes is a chronic condition which serves as a precursor for other medical conditions and could trigger other health concerns in future. It has been estimated that about 366 million people worldwide have diabetes, and this number is likely to escalate to 552 million by the year 2030 [2].

Around 25.8 million people in the U.S., which is about 8.3% of the population at present, have diabetes. If the present trend continues, one in every three U.S. adults is likely to have diabetes by the year 2050 [1]. From a monetary consideration, in 2012 alone diabetes has cost the U.S. approximately $245 billion dollars [1] and unchecked diabetes resulted in 23,700 readmissions and cost $251 million [12]. SAS

Enterprise Miner 13.1 is used in this paper to identify patients among various surgical groups who display a higher likelihood of readmission.

LITERATURE REVIEW

Health care organizations continue to develop optimal solutions in chronic care models. Examination of scholarly articles has revealed that out of 39 studies, 32 interventions based on chronic care models have resulted in improvement of at least one outcome measure according to Thomas Bodenheimer, MD, Edward H. Wagner, MD, MPH and Kevin Grumbach, MD [6]. From a financial perspective, studies suggest that out of 27 cases, 18 have confirmed that using chronic care models has reduced health care costs and led to a decrease in the use of health care services [6].

According to the study carried out by Allen M. Glasgow, Jill Weissberg-Benchell, W. Douglas Tynan, Sandra F. Epstein, Chris Driscoll, Jane Turek, EvBeliveau in “Readmissions of Children with Diabetes Mellitus to a Children's Hospital”, it was observed that readmission of diabetic children increased due to missed insulin injections [9].

Another point of view is presented by BeataStrack, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios and John N. Clore in “Impact of HbA1c measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records” where the authors used the multivariate regression method to conclude that greater attention to HbA1c determination may improve outcomes and reduce costs [8].

Using such chronic illness management models to develop personalized interventions in hospitals has yielded better patient outcomes, decreased length of stay, and lessened rates of re-admission in a 30 day period. Jill Koproski, RN, CDE, ZoraydaPretto, MD and Leonid Poretsky, MD, conducted a random feasibility study on the effects of an intervention in hospitalized diabetics by a medical research team and found that it resulted in decreased length of stay in the hospital as well as an overall improvement in the glycemic levels [7].

DATA

This study uses the data obtained from the Center for Machine Learning and Intelligent Systems at University of California, Irvine [4]. It contains clinical records from over 100,000 individual encounters corresponding to over 60,000 distinct patients. The data was collected from over 74 million encounters related to 17 million patients [8]. It was collected over a period of 10 years, from 1999 to 2008, and contains several attributes which correspond to the times of admission and discharge of diabetic patients. These records contain information about various laboratory tests and procedures, diagnosis, and medications that were administered in the duration of the hospital stay. Following are the variables selected for modeling:

Variable Name	Measurement Level	Range of Values
A1Cresult	Nominal	>7, >8, none, norm
admission_source_id	Nominal	1 to 25
admission_type_id2	Nominal	1 to 4
admitted_30days	Binary	0 to 1
age	Nominal	[0-10]... [90-100]
change2	Binary	0 to 1
diabetesmed2	Binary	0 to 1
Diagnosis_1	Nominal	3 to 999
Diagnosis_2	Nominal	5 to 999

Diagnosis_3	Nominal	3 to 999
discharge_disposition_id	Nominal	1 to 28
discharge_disposition_id_2	Nominal	1 to 17
encounter_id	Interval	12522 to 443867222
gender	Char	Male, Female, Other
medical_specialty	Nominal	1 to 23
num_lab_procedures	Interval	1 to 132
num_medications	Interval	1 to 81
num_procedures	Interval	0 to 6
number_diagnoses	Interval	1 to 16
number_emergency	Interval	0 to 76
number_inpatient	Interval	0 to 21
number_outpatient	Interval	0 to 42
patient_nbr	Interval	ID
Race	Nominal	Asian, Caucasian…
Re_admitted	Binary	0 to 1
Readmitted	Nominal	No, <30, >30
time_in_hospital	Interval	1 to 14
Weight	Nominal	[150-170]…[170-200]
weight_new	Interval	165, 175…
24 features for medications	Nominal/Binary	Metformin, Glipizide…… etc.

Table 1: List of variable names, measurement level, description and their range of values

Health Insurance Portability and Accountability Act (HIPPA) requires that any personal information which would identify a patient must be kept confidential. In this research it was ensured that the dataset was HIPPA compliant. In addition, the research conducted did not require any form of consent from the Helsinki Declaration, enabling its exemption from the VCU Institutional Review Board review [8].

The data set “Diabetes 130-US hospitals for years 1999-2008” incorporated the use of International Statistical Classification of Diseases and related health Ppoblems, also known as ICD [10]. ICD is maintained by World Health Organization (WHO) and serves as a “standard diagnostic tool for epidemiology, health management and clinical purposes” [10]. ICD serves as a data dictionary for various kinds of health conditions. Furthermore, to explain the health condition classification system, it provides detailed explanation of the conditions, for instance signs, symptoms, and so on [10]. The dataset used for this study incorporated the majority of the ICD systems ranging from ICD-9001 to ICD-9999 [10].

DATA PREPARATION:

The original dataset had 55 attributes and 101,731 observations. In order to prepare the dataset for modeling, the data was subjected to intense cleaning procedures. The final dataset created has 91,841 observations and 53 attributes as displayed below.

Role	Measurement Level	Frequency Count
INPUT	BINARY	12
INPUT	INTERVAL	8
INPUT	NOMINAL	25
REJECTED	BINARY	2
REJECTED	INTERVAL	2
REJECTED	NOMINAL	2
REJECTED	UNARY	2

Figure 1 Variable Summary

The nominal variables such as diagnosis, race, medical specialty, and discharge disposition had missing values. These missing values need to be handled prior to being fed into the model as they may introduce instability and bias in models such as regression and support vector machines. In order to avoid imputing missing values in primary, secondary and tertiary diagnosis, the observations that had missing values in them were removed from the analysis using PROC SQL. 'Medical Specialty' variable, containing approximately 50% missing values, pertains to the medical specialty of the physician who attended the patient and prescribed the initial diagnosis. 'Weight' variable has 97% missing values. After careful consideration, weight, payer code, and medical specialty variables have been rejected in model building.

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
num_lab_procedures	INPUT	43.57071	19.70622	93294	0	1	45	132	-0.24695	-0.23371
num_medications	INPUT	16.1344	8.205716	93294	0	1	15	81	1.336144	3.473862
num_procedures	INPUT	1.356239	1.725256	93294	0	0	1	6	1.29748	0.770448
number_diagnoses	INPUT	7.446374	1.930978	93294	0	1	8	16	-0.90207	-0.01788
number_emergency	INPUT	0.198727	0.939338	93294	0	0	0	76	23.52216	1239.687
number_inpatient	INPUT	0.625206	1.249806	93294	0	0	0	21	3.554008	19.70258
number_outpatient	INPUT	0.369413	1.27864	93294	0	0	0	42	8.969296	151.3833
time_in_hospital	INPUT	4.401108	2.963397	93294	0	1	4	14	1.125158	0.846443

Figure 2 Interval variable summary statistics

Variable Name	Role	Number of Levels	Missing	Mode
AlCresult	INPUT	4	0	None
Diagnosis_1	INPUT	688	19	428
Diagnosis_2	INPUT	636	354	276
Diagnosis_3	INPUT	664	1360	250
Gender2	INPUT	2	0	0
acarbose	INPUT	4	0	No
admission_sourceid2	INPUT	7	0	5
admission_type_id2	INPUT	4	0	1
age	INPUT	10	0	[70-80)
change2	INPUT	2	0	0
chlorpropamide	INPUT	4	0	No
diabetesmed2	INPUT	2	0	1
discharge_disposition_id_2	INPUT	10	4227	1
glimepiride	INPUT	4	0	No
glipizide	INPUT	4	0	No
glyburide	INPUT	4	0	No
glyburide_metformin	INPUT	4	0	No
insulin	INPUT	4	0	No
max_glu_serum	INPUT	4	0	None
medical_specialty	INPUT	72	45769	
metformin	INPUT	4	0	No
nateglinide	INPUT	4	0	No
pioglitazone	INPUT	4	0	No
race	INPUT	6	2141	Caucasian
repaglinide	INPUT	4	0	No
rosiglitazone	INPUT	4	0	No
Re_admitted	TARGET	2	0	0

Figure 3 Class Variable Summary Statistics

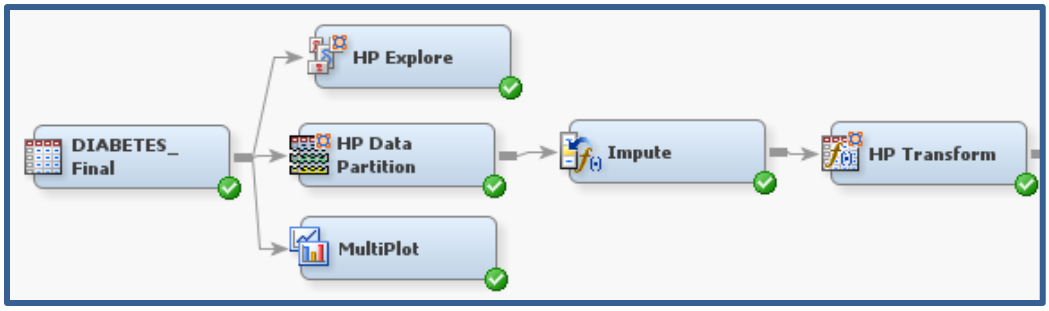
Some variables, for example weight, race, and encounter ID contained anomalies such as questions marks and other special characters. Observations with invalid gender values were deleted. The variables

discharge disposition and race had few percentages of missing values and they were imputed using the Tree imputation method in the Impute node. The nominal variable 'discharge disposition' had 29 levels such as null, unknown, and other etc. These levels were grouped in a more logical way and reduced to 6 levels using SAS code. Similarly, the variables admission type and admission source had 8 and 26 levels. As many levels were similar to each other with subtle differences, they were categorized into 4 and 7 levels respectively. In addition, 11 categorical variables were converted into binary variables for ease of analysis. The data was checked for outliers via box and whisker plots. The observations that were beyond the range of ±3 standard deviation were deemed as outliers and were accordingly removed using SAS code. The target variable of readmission had three levels of '<30' (patient readmitted within 30 days), '>30' (patient was readmitted after 30 days) and 'no' (patient was not readmitted). In this paper, the target variable has been re-coded as a '1' (Patient is readmitted) and '0' (patient is not readmitted) as a binary variable.

The interval variables 'number emergency (number of times a patient was admitted via emergency)' and 'number outpatient (number of times the patient was an outpatient before being admitted)' had high values of skewness and kurtosis. In order to normalize these variables, log transformation was used to reduce the skewness and kurtosis values of these variables.

MODEL BUILDING

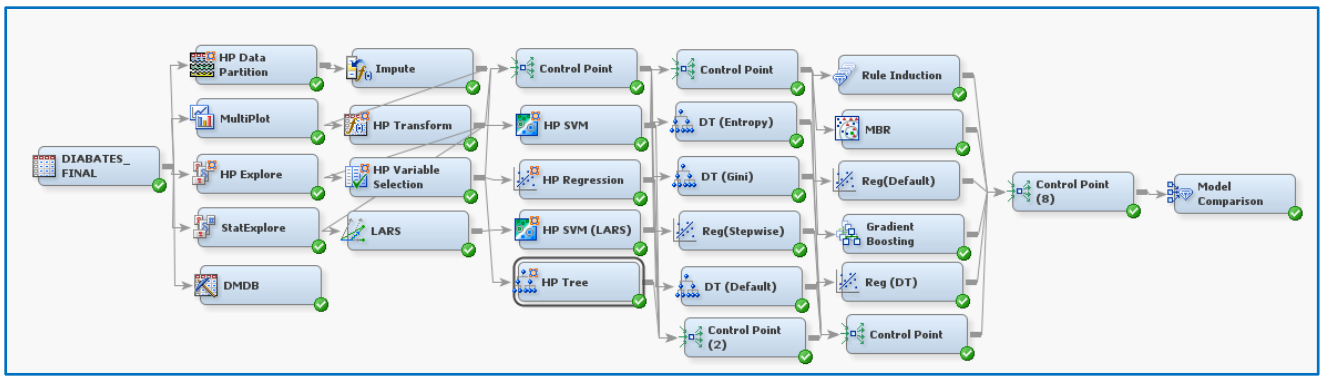
The data was partitioned into train (60%) and validation (40%) datasets to provide an honest model assessment.. As the percent of target variable in each outcome is balanced in this data, there is no need to set prior probabilities.



Display 1 Initial process flow

SAS Enterprise Miner 13.1 provides many new features and nodes. The newer sets of models such as high performance nodes are available in this version and used in the analysis. The primary purpose of this paper is to predict whether or not a patient will be re-admitted and to identify the key factors serving the objective. As the target is a binary variable, the selection criterion used was the validation misclassification rate. For variable selection purposes, the variables selected from different nodes such as the HP variable selection node with variations (LARS, LASSO, and Adaptive LASSO), LARS node, the HP GLM node and the Decision Tree node were used to identify the significant predictors. Among all these variable selection nodes, the variables selected by the LARS model were most instrumental in decreasing the validation misclassification rate.

In order to predict the patient outcome measures, various models were built such as the Decision Tree with variations in the splitting rule target criteria (default, Entropy, Gini, HP Tree), the Regression models with variations in variable selection criteria (default, decision tree, stepwise, HP Tree), the Gradient Boosting with default properties, the Rule Induction node where the binary model used is Tree and the cleanup model used was 'Neural', the MBR model using the RD-Tree method with 16 neighbors, and the Support Vector Machine node with the Linear Kernel Function. The model comparison node has been used to select the model with the least validation misclassification rate. Among all the models, the LARS model (variable selection) used in conjunction with the HP Support Vector Machine had the lowest validation misclassification rate of 0.363.



Display 2 Process Flow Diagram

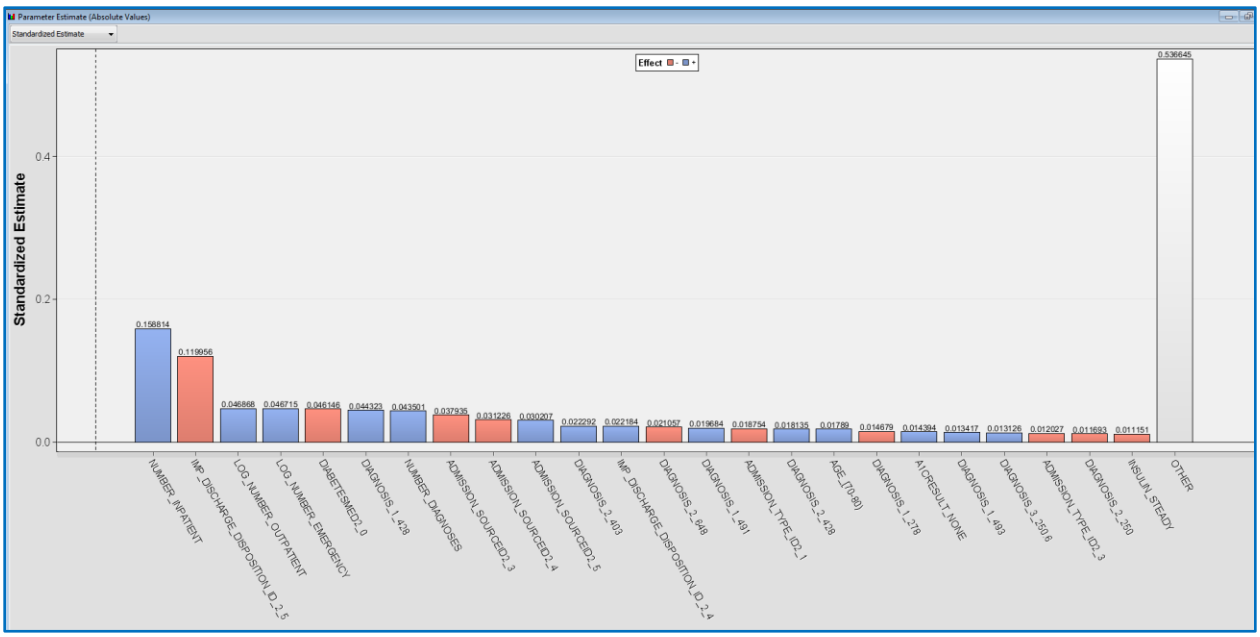
Fit Statistics

Selected Model	Model Description	Target Variable	Selection Criterion: Valid: Misclassification Rate	Train: Misclassification Rate	Train: Roc Index	Valid: Roc Index	Train: Cumulative Lift	Valid: Cumulative Lift
Y	HP SVM (LARS)	Re_admitted	0.366153	0.34407	0.716	0.683	1.642554	1.616555
	HP SVM	Re_admitted	0.366643	0.343145	0.716	0.683	1.640204	1.619492
	Rule Induction	Re_admitted	0.367405	0.370348	0.65	0.651	1.263997	1.259032
	Reg(Stepwise)	Re_admitted	0.367432	0.369803	0.68	0.679	1.615916	1.623017
	HP Tree	Re_admitted	0.367841	0.3714	0.674	0.675	1.617199	1.625038
	DT (Entropy)	Re_admitted	0.370127	0.373795	0.652	0.655	1.53867	1.53555
	DT (Default)	Re_admitted	0.370127	0.373795	0.652	0.655	1.53867	1.53555
	Reg (DT)	Re_admitted	0.370291	0.375338	0.667	0.67	1.61239	1.621548
	DT (Gini)	Re_admitted	0.370536	0.366246	0.667	0.662	1.538504	1.541594
	HP Regression	Re_admitted	0.394654	0.380474	0.669	0.648	1.509364	1.479688
	MBR	Re_admitted	0.428626	0.360722	0.584	0.584	1.306946	1.274643

Output 1 Fit Statistics of Model Comparison Node

EXPLAINING THE BEST MODEL- HP SVM WITH VARIABLE SELECTOR (LARS)

LARS (Least Angle Regression) is used to identify the important predictors for model building. It is a recently developed specialized regression technique that is used for variable selection as well as model fitting and prediction. LARS can handle both numeric and categorical input variables. Using LARS, problems such as violation of hypothesis assumptions and selection bias may be overcome by using information criteria such as AIC and SBC.



Output 2 Parameter Estimate (Absolute Value) LARS

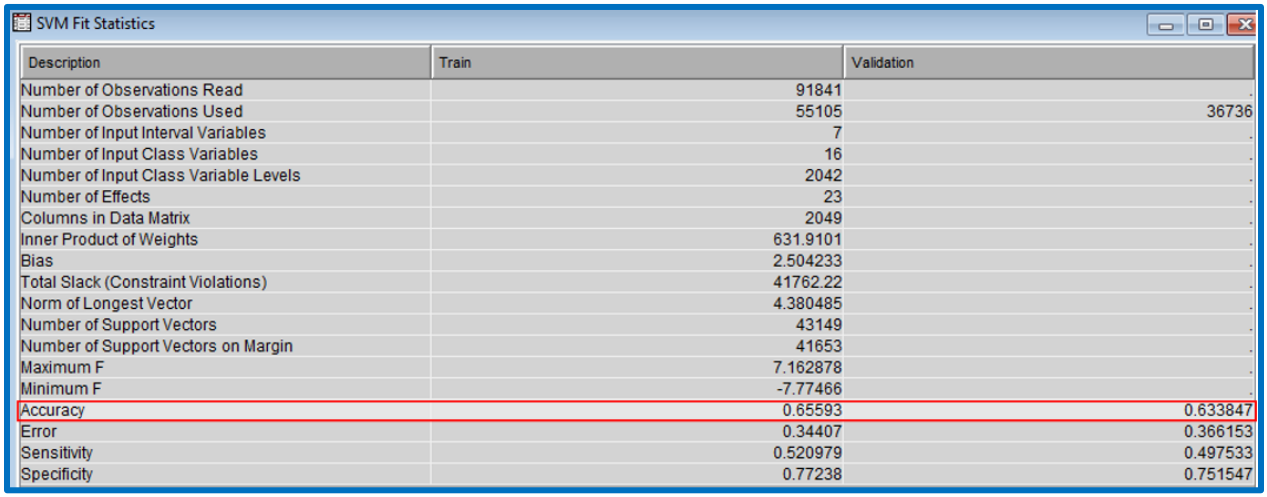
Shown above is the plot of absolute values of the parameter estimates from the LARS model. The variable 'Number_Inpatient' in the Parameter Estimate graph has the highest estimate of 0.158. The important predictors in order of decreasing importance are as follows:

- The number of times a patient has been an inpatient previously has a positive effect on readmission followed by the number of times he/she was an outpatient.
- The number of times a patient was admitted via the emergency room has the next highest positive impact with a likelihood of 4%.
- Diabetic patients suffering from a primary diagnosis of heart failure have a higher chance of readmission.
- Number of diagnoses entered into the system corresponding to each patient is one of the important predictors.
- Diabetic patients who were transferred from another hospital, home health agency, ambulatory surgery center, critical access hospitals, or due to court/law enforcement have reduced chances of being readmitted.
- Patients who were transferred from a skilled nursing facility (SNF) were discovered to have increased chances of readmission.
- In addition, patients suffering from a secondary diagnosis of hypertensive chronic kidney disease demonstrated increased readmission rates.
- Patients who were discharged/ transferred to home with home health service had a positive effect, thus indicating a higher chance of coming back to the hospital.
- Diabetic women with complicated pregnancy or childbirth have displayed a reduced likelihood of being readmitted.
- Patients suffering from a primary diagnosis of chronic bronchitis are more likely to be readmitted; it was also noticed that patients with diabetes mellitus had higher chances of being readmitted.

However, the LARS model is used as a variable selector and how these selected effects are weighed in to the final SVM model is unknown.

Support Vector Machine (SVM) algorithm has been selected as the winner model by the model comparison node since it has the lowest validation misclassification rate of 0.363. It is a supervised machine learning method which is used to solve binary classification problems. Generally, the data in the real world is not linearly separable. SVM maps the data into a higher dimensional space (infinite) which helps in separating the data points more effectively. Inside SVM, for a data-point having n-dimensions (n input variables), a (n-1) hyperplane (a linear combination of the input variables) is generated. This (n-1) dimensional hyperplane is generated using a kernel function which is used to separate the data points in a yes/no, 0/1 or +/- situation. The kernel function used here is the Linear Kernel function

The SVM node takes the training data, selects the weight and bias in such a way that the hyperplane separates the training data points. Once the plane is known, if the plane is on the positive direction, it is termed as positive data; if the plane is on the negative direction, it is termed as negative data. The purpose of SVM is to find the best hyperplane, which maximizes the margin or the boundary vectors (vectors define the boundary of the classes). The best hyperplane provides maximum safety in case the predicted value is incorrect. The optimization problem is to maximize the distance or minimize the width (w) of the hyperplane. Under circumstances where the data is not completely separable, the SVM model imposes a penalty. The idea is to allow penalty for classifications which were wrongly predicted. The following variables were fed into the model: 7 interval variables and 16 class variables (2,042 class variable levels). The number of effects selected by the model is 23.

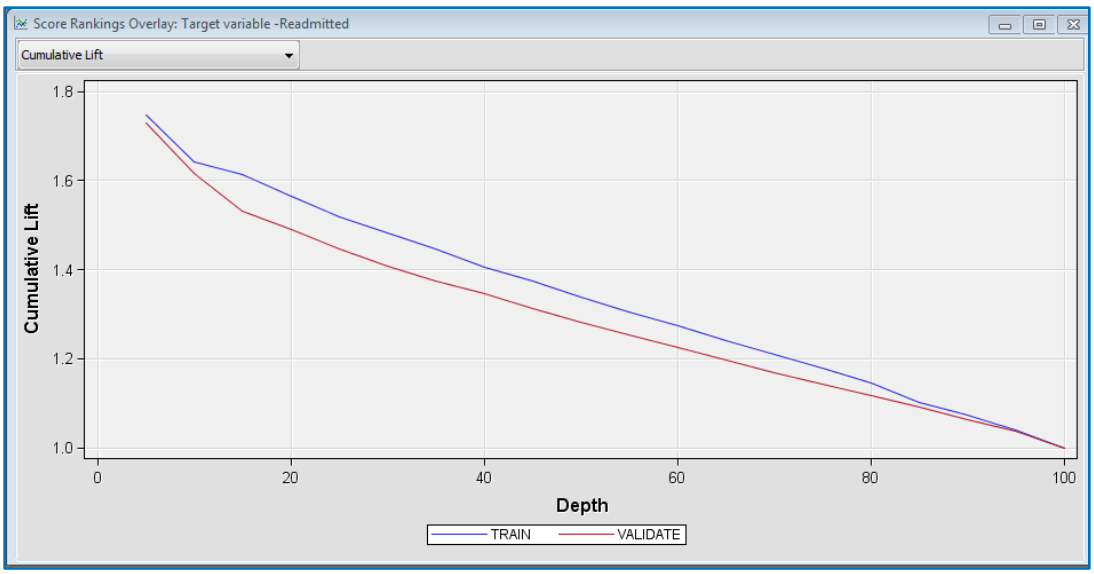


SVM Fit Statistics

Description	Train	Validation
Number of Observations Read	91841	.
Number of Observations Used	55105	36736
Number of Input Interval Variables	7	.
Number of Input Class Variables	16	.
Number of Input Class Variable Levels	2042	.
Number of Effects	23	.
Columns in Data Matrix	2049	.
Inner Product of Weights	631.9101	.
Bias	2.504233	.
Total Slack (Constraint Violations)	41762.22	.
Norm of Longest Vector	4.380485	.
Number of Support Vectors	43149	.
Number of Support Vectors on Margin	41653	.
Maximum F	7.162878	.
Minimum F	-7.77466	.
Accuracy	0.65593	0.633847
Error	0.34407	0.366153
Sensitivity	0.520979	0.497533
Specificity	0.77238	0.751547

Output 3 SVM Fit Statistics

The accuracy obtained by the model is 63.38% for the validation data. With the help of the model and sorting the observations based on probabilities, it can be stated that patients in the top decile have 61.6% higher likelihood of being readmitted.



Output 4 Cumulative Lift SVM

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
Re_admitted	Target variable -Readmitted	_ASE_	Average Squared Error	0.227901	0.231116
Re_admitted	Target variable -Readmitted	_DIV_	Divisor for ASE	110210	73472
Re_admitted	Target variable -Readmitted	_MAX_	Maximum Absolute Error	0.8383	0.825203
Re_admitted	Target variable -Readmitted	_NOBS_	Sum of Frequencies	55105	36736
Re_admitted	Target variable -Readmitted	_RASE_	Root Average Squared E...	0.477389	0.480745
Re_admitted	Target variable -Readmitted	_SSE_	Sum of Squared Errors	25116.92	16980.57
Re_admitted	Target variable -Readmitted	_DISF_	Frequency of Classified ...	55105	36736
Re_admitted	Target variable -Readmitted	_MISC_	Misclassification Rate	0.34407	0.366153
Re_admitted	Target variable -Readmitted	_WRONG_	Number of Wrong Class...	18960	13451

Output 5 Fit statistics SVM

CONCLUSION

The hospital data of in-patients having diabetes as an existing condition in conjunction with other medical illnesses was analyzed to build a predictive model to identify patients who had a higher likelihood of being readmitted. Some of the key factors that drove readmission are the number of times a patient was formerly admitted both as an inpatient and outpatient, primary diagnosis, mode of admission, conditions like heart failure or hypertensive chronic kidney disease and so on. Using the model, it can be stated that patients in the top decile have 61.6% higher likelihood of being readmitted.

DISCLAIMER

The authors do not possess any medical expertise. The results are based purely on the data and are not intended to serve as a medical advice.

REFERENCES

1. http://mitsloanexperts.mit.edu/using-analytics-to-manage-diabetes/
2. http://www.sciencedirect.com/science/article/pii/S0956566313004260
3. http://www.hindawi.com/journals/
4. https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008
5. http://archive.ics.uci.edu/ml/

6. http://jama.jamanetwork.com/article.aspx?articleid=195407
7. http://care.diabetesjournals.org/content/20/10/1553.short
8. http://www.hindawi.com/journals/bmri/2014/781670/
9. http://pediatrics.aappublications.org/content/88/1/98.short
10. http://en.wikipedia.org/wiki/International_Statistical_Classification_of_Diseases_and_Related_Health_Problems
11. https://www.healthcatalyst.com/healthcare-data-warehouse-hospital-readmissions-reduction
12. https://www.fiercehealthfinance.com/story/readmission-lead-413b-additional-hospital-costs/2014-04-20

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Dr. Goutam Chakraborty
Oklahoma State University
Stillwater, OK, 74078
goutam.chakraborty@ okstate.edu

Dr. Goutam Chakraborty is Ralph A. and Peggy A. Brenneman professor of marketing and founder of SAS® and OSU data mining certificate and SAS® and OSU marketing analytics certificate at Oklahoma State University. He has published in many journals such as Journal of Interactive Marketing, Journal of Advertising Research, Journal of Advertising, Journal of Business Research, etc. He has chaired the national conference for direct marketing educators for 2004 and 2005 and co-chaired 2007 data mining conference. He has over 25 years of experience in using SAS® for data analysis. He is also a Business Knowledge Series instructor for SAS®.

Hephzibah Munnangi
Oklahoma State University
Stillwater, OK- 74075
Work Phone: 405-385-4992
Email: hephzibah.munnangi@okstate.edu

Hephzibah Munnangi is a Master's graduate in Management Information Systems from Oklahoma State University. She has a bachelor's degree in electronics and communication engineering. She has two years' experience in using SAS® tools for Database Marketing and Predictive Modeling. She is a Base SAS® 9 certified professional, SAS Certified Statistical Business Analyst Using SAS 9: Regression and Modeling Credential, JMP Software data exploration certified and holds the SAS and OSU Data Mining certification. She is currently working as an Analyst at Ipsos MMA.