Charlie Xu (cxu2), Stephone Christian (stephone), Christina Pan (capan)
12/15/2017

Beating Diabetes: Predicting Early Diabetes Patient Hospital Readmittance to Help Optimize Patient Care

Project Category: Life Sciences

Introduction

According to the American Society of Diabetes, the cost of care for diabetic and prediabetic patients in the United States is 332 billion USD (Cost of Diabetes, 1). This global epidemic affects over 350 million people, with 3 million people dying each year due to diabetes related complications, predominantly cardiovascular or nephropathic ones (Zhu, 2). By analyzing an extensive diabetic patient dataset accumulated from 130 hospitals in the United States from 1999 to 2008, we wish to reduce the mortality rate of diabetic people by improving patient care, while also reducing the astronomical yearly cost of diabetic patient care.

In order to accomplish this, we will create a model to predict whether a diabetic patient will be readmitted to the hospital in less than 30 days from the last visit, as well as a model that expresses the probability of a diabetic patient being readmitted in less than 30 days from the last visit. Thus, the input to our models will be patient data, and we then use a variety of models (e.g. logistic regression, SVMs, random forests) to output a prediction as to whether the patient will be readmitted early..

Each model is valuable in different ways. The probabilistic model can be used by doctors during a patient visit to help them with decisions such as whether to administer an HbA1 test, a costly procedure, or whether to change the dosage or administer new medications. The doctor could theoretically run the model with different hypothetical features--different medications, different tests--and inform their decision using the the combination of features that reduces the patient's chance of readmission into the hospital. The binary model can be used to infer general patterns in the data and to determine which patient characteristics may lead to early hospital readmission. This may expose weaknesses in the current approach doctors are taking in diabetic patient care, and hopefully pave the path to better health outcomes for diabetic people.

Related Work

Prediction of early readmittance is a topic of major interest, with various studies on this topic focusing on major diseases and procedures using electronic , such as for heart failure (Philbin & DiSalvo) and intestinal surgery (Kiran et al.). However, these studies primarily focus upon manual analysis like Student's t-tests or chi-squared tables (Philbin & DiSalvo; Kiran et al.).

However, there have been a few studies using machine learning on hospital readmission, like heart failure (Shameer et al.). In fact, that model used just a Naive Bayes algorithm to obtain an accuracy of 83%.

There has been some research conducted regarding early readmittance in the case of diabetes, where multivariable logistic regression models were used to assess the impact of HbA1c values on early readmittance. (Beata). While this model did delve thoroughly into the relationship between HbA1c values and early readmittance from a statistical standpoint, particularly with the use of p-values and significance tests, there was not much focus on utilizing the logistic regression model or other machine learning algorithms for predictions. Thus, we think our application of machine learning models to this problem will yield exciting and interesting insights regarding early readmission for diabetes patients.

Tangentially related however widely useful for our applications is a paper published by the University of Waikato on correlation based feature selection for multidimensional data (Hall, M.A). . The dataset we are leveraging for our algorithms has over 55 features, thus techniques on feature selection to diminish noise and pull out signal appeal greatly to our causes.

Dataset and Features

The dataset that we used, which was extracted from the UCI Machine Learning Repository, is a dataset with over 100,000 rows and 55 features extracted from the anonymized electronic patient health records of 150 hospitals from 1999-2008, where each row corresponds to a patient. (Beata, 3) These features can be grouped into 4 categories: Patient Demographic Information, Hospital Metrics and Measurements, and Diabetes Medications. Patient Demographic Information, the smallest category, contains the features of race, gender, and age. Hospital Metrics, the largest category, contains features that pertain to the measurements and metrics that were noted during a patient's hospital visit (e.g. number of procedures, insulin levels, the number of lab procedures a patient went through, the results of an H1A1c test, if taken). Lastly, the Diabetes Medications category contains features that provide information on the patient's usage levels of various diabetes medications (e.g. Glyburide, Metformin).

	race	gender	age	admission_type_id	discharge_disposition_id	admission_source_id	time_in_hospital	medical_specialty	num_lab_procedures	num_procedures	num_medications	number_outpatient	number_emergency	number_inpatient	number_diagnoses	max_glu_serum	A1Cresult	metformin
1	Caucasian	Female	[0-10)	6	25	1	1	Pediatrics-Endocrinology	41	0	1	0	0	0	1	None	None	No
2	Caucasian	Female	[10-20)	1	1	7	3	Missing	59	0	18	0	0	0	9	None	None	No
3	AfricanAmerican	Female	[20-30)	1	1	7	2	Missing	11	5	13	2	0	1	6	None	None	No
4	Caucasian	Male	[30-40)	1	1	7	2	Missing	44	1	16	0	0	0	7	None	None	No
5	Caucasian	Male	[40-50)	1	1	7	1	Missing	51	0	8	0	0	0	5	None	None	No
6	Caucasian	Male	[50-60)	2	1	2	3	Missing	31	6	16	0	0	0	9	None	None	No
7	Caucasian	Male	[60-70)	3	1	2	4	Missing	70	1	21	0	0	0	7	None	None	Steady

We also conducted some preprocessing of our dataset. First, we eliminated features that were extremely sparse (significant majority of feature values are NA), since such columns would have very little predictive value and would cause problems for our

machine learning algorithms. To accomplish this, we identified the most frequently occuring value for each categorical variable and then calculated the percentage of observations associated with that value. Through this method, we determined that 7 columns were too sparse to be useful for our algorithms and were therefore eliminated from our dataset. We also went through non-binary categorical features to identify any values that corresponded to only 1 observation (i.e. only 1 patient experienced/had that particular quality) and then eliminate that observation. Inclusion of such observations might cause our machine learning algorithms to incorrectly determine that the presence of that particular value would perfectly predict the early readmission outcome for that observation. Thus, to accomplish this task, for each non-binary categorical feature, we obtained the number of observations contained within each value of the feature and then eliminated any observations where the given value contained only 1 observation. With this method, we eliminated 8 observations. We did not normalize our dataset, since mostly categorical variables and because the numerical variables were all features like "num_procedures" or "number_diagnoses", the step sizes (and their associated meanings) were already standardized across all numerical features.

Since we want to use the best models after comparing the performances of our different models, we then divided up our dataset into our training, dev, and test sets using a 60/20/20 split. Thus, our training set contained 61,059 observations; our dev set contained 20,351 observations, and our test set contained 20,352 observations.

Methods

As mentioned above, we used a variety of different classification models to assess which one had the greatest predictive value.

1.) Logistic Regression (Binomial)

The binomial logistic regression algorithm works by finding the theta value that maximizes the following log likelihood function:

$$\begin{aligned}\ell(\theta) &= \log L(\theta) \\ &= \sum_{t=1}^{m} y^{(t)} \log h(x^{(t)}) + (1 - y^{(t)}) \log(1 - h(x^{(t)}))\end{aligned}$$

,

where x represents the features matrix, y represents the labels vector, and h(x) is $1/(1 + exp(-\theta^T x)$

2.) Logistic Regression (Multinomial)

The multinomial logistic regression algorithm works essentially the same way as the binomial logistic regression algorithm. However, since multinomial logistic regression is traditionally used to predict the outcome for multi-class variables, the log likelihood function that needs to be maximized is the sum of the log-likelihood functions for each class, as shown below:

$$\begin{aligned}\ell(\theta) &= \sum_{n=1}^{N} \log \left[\frac{\exp(\theta_{y_n}^{\mathrm{T}} \mathbf{x}_n)}{\sum_{c=1}^{C} \exp(\theta_c^{\mathrm{T}} \mathbf{x}_n)} \right] \\ &= \sum_{n=1}^{N} \left[\theta_{y_n}^{\mathrm{T}} \mathbf{x}_n - \log \sum_{c=1}^{C} \exp(\theta_c^{\mathrm{T}} \mathbf{x}_n) \right]\end{aligned}$$

,

where C represents the number of classes within the multi-class outcome variable.

Initially we had wanted to assess multi-class outcome variable, but we ultimately found that one class had too few outcomes to be predicted accurately. Thus, we decided to proceed with predicting a binary variable, but we still used the multinomial logistic regression algorithm to assess its accuracy in our case and to have another model for comparison.

3.) Elastic Net

The elastic net algorithm implements logistic regression, while also incorporating L1 and L2 regularization. Thus, the algorithm finds the beta value that minimizes the following negative likelihood function:

$$\min_{(\beta_0,\beta)\in\mathbb{R}^{p+1}} -\left[\frac{1}{N}\sum_{i=1}^{N} y_i \cdot (\beta_0 + x_i^T\beta) - \log(1 + e^{(\beta_0 + x_i^T\beta)})\right] + \lambda\left[(1-\alpha)||\beta||_2^2/2 + \alpha||\beta||_1\right].$$

,

given that x is the features matrix, y is the labels vector, lambda is the weight for the regularization, and alpha is the parameter for adjusting how much to weight L1 regularization vs L2 regularization.

4.) Naive Bayes

This method uses the classic Naive Bayes algorithm, which assumes that all of the features are independent of one another. To predict y, we use:

$$P(y \mid x_1, \ldots, x_n) \propto P(y) \prod_{i=1}^{n} P(x_i \mid y)$$

$$\Downarrow$$

$$\hat{y} = \arg\max_y P(y) \prod_{i=1}^{n} P(x_i \mid y),$$

5.) Random Forests

The random forests algorithm is an ensemble method which constructs decision trees based on a bootstrapped sample from the dataset. The candidate features for each split point in the decision tree are generated using a bootstrapped set of the features of the overall dataset, and the bootstrapped set is then used to determine the optimal feature to be used for the decision rule. This process generates an individual decision tree classifier, which, due to the use of bootstrapping, is a model that has low bias, but high variance. Thus, by aggregating predictions across all trees, the random forests algorithm can reduce the overall variance of the model, while still keeping bias relatively low. Furthermore, in our case, by determining what percentage of decision trees voted for the "1" class, we can also find the probability of predicting "1" associated with each observation.

5.) Support Vector Machines

Support Vector machines (SVMs) are a type of powerful classifier capable of drawing a hyperplane through multidimensional data. The distance a point is from said hyperplane informs the confidence of the prediction/classification of the data. The motivation for using an SVM for early readmission prediction on the dev set was to see whether there existed some geometric order to the data. I ran the SVM using the fitcsvm MatLab package. I trained said model on the train set and determined its accuracy on both the train and dev set, the results of which I reported in the table below.

6.) Neural Networks

Neural Networks use several layers of hidden functions whose output become the input of the next layer. With these layers, neural networks can do automatic feature selection. For our implementation of the neural network, we used the sigmoid function (below) as the activation function as well as cross entropy loss (below), given that the output was also a sigmoid function.

$$f(x) = \frac{1}{1 + e^{-(x)}}$$

$$-\frac{1}{N} \sum_{i=1}^{N} y_i log(h_\theta(x_i)) + (1 - y_i) log(1 - h_\theta(x_i))$$

Results and Discussions.

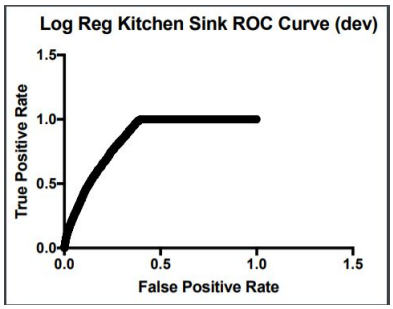
As stated before, we also ran experiments using our models, primarily around error analysis. To accomplish this, we first established 0-1 loss to be the standard metric by which we would assess accuracy. We then generated a naive classifier, which would predict the most frequently occurring label (in our case, 0) and then established the 0-1 loss gained from that naive classifier as a baseline for comparing the errors of our other models. We then found the 0-1 loss for each of our models on the training dataset, the results of which are documented below:

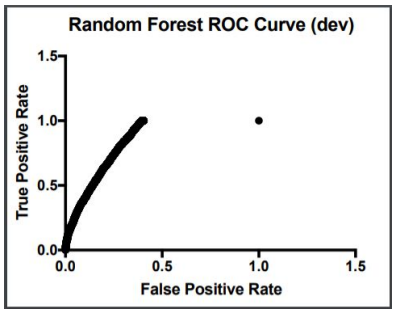
Model Used	Train 0-1 Loss	Train Accuracy	Dev 0-1 Loss	Dev Accuracy
Naive Classifier	0.1102794	0.8897206	0.113748	0.8875731
Naive Bayes	0.110991	0.889009	0.112230	0.8864352
Binomial Log Reg (kitchen sink)	0.109051	0.890949	0.1134532	0.8865468
Binomial Log Reg (selected features based on significance in kitchen sink model)	0.1101812	0.8898188	0.1138463	0.8861537
Multinomial Log Reg (kitchen sink)	0.1091493	0.8908507	0.1134532	0.8865468
Elastic Net	0.1101484	0.8898516	0.113748	0.886252

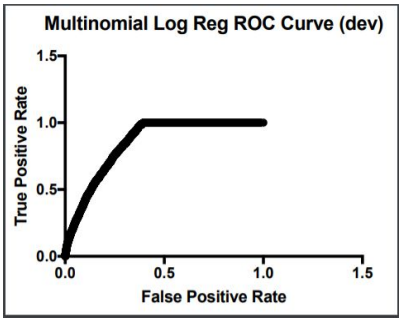
Random Forests	0.03513283	0.9648672	0.1118809	0.8881191
SVM (no Kernel)	.1095	.8905	.1110	.8890
SVM (3rd order polynomial Kernel)	.1095	.8905	.1110	.8890
SVM (Gaussian Kernel)	.1050	.8950	.1110	.8890
Neural Network	0.110991	0.889009	0.112304	0.8875731

As can be seen from our table, for both the train set and the dev set, the logistic regression models (both binomial and multinomial) achieved marginal improvement over the baseline error, while the elastic net model surprisingly performed slightly worse than the regular logistic regression models. The naive bayes model performed about the same, while the SVM models also did about the same compared to the baseline, regardless of the kernel function. However, the random forests model did very well on the train set, and although the same exemplary performance was not reflected when applied to the dev set, the random forests model still achieved the highest accuracy out of all the other models.

For greater insight into the performance of our models, we found their ROC curves and AUC values. The ROC curves of our 3 most accurate models (binomial logistic regression, random forests, and multinomial logistic regression) from which we could yield an ROC curve are below:







As can be seen from the ROC curves, both the binomial logistic regression and multinomial logistic regression curve are fairly comparable to each other. Moreover, the random forests ROC curve looks fairly similar to the other models, albeit with a slightly steeper ascent to the curve. (The gap in the model is because the thresholds stopped iterating at that point and then jumped to set the threshold at 0.). Thus, as with the accuracy metrics, the models appear fairly similar to each other, based on their ROC curves.

Finally, to better understand the types of errors that our classifiers were making, we generated a confusion matrix for each model based on its performance on the dev set. The confusion matrices for our 3 most accurate models (binomial logistic regression, random forests, and multinomial logistic regression) are included below. Also note that we did not select the Neural Net, despite its higher accuracy, because its confusion matrix showed that it was simply predicting all 0s):

LRKS	Actual 0	Actual 1
Pred 0	17931	2135
Pred 1	131	153

RF	Actual 0	Actual 1
Pred 0	17869	2071
Pred 1	194	217

LRMKS	Actual 0	Actual 1
Pred 0	17931	2134
Pred 1	131	154

As can be seen from the confusion matrices, despite their accuracy numbers, unfortunately our models do not perform terribly well when predicting early readmittances. The LRMKS and LRKS models perform identically with regards to predicting instances when a patient is not readmitted early, and they perform nearly identically well with regards to predicting instances with a patient is readmitted early (LRMKS does slightly better). However, the Random Forests model is of interest. While it does perform worse than the other two models in terms of predicting instances when a patient is not readmitted early, it performs better than the other models when predicting instances when a patient is readmitted early. Thus, since we are interested in predicting early readmittances, the Random Forests model is more useful to us.

Based on our analysis of its overall objective accuracy, improvement over the baseline in accuracy, ROC curve, and confusion matrix, we determined that the best model was the random forests model. Thus, we then ran our model upon our test set, which yielded a test accuracy of 0.8909636.

In addition, we also looked at the weights given from our logistic regression model. From this model, the most significant covariates were an increase in dosages of Chloropropamide and Glyburide Metflormin as well as seeing doctors with the specialities of allergy & immunology, sports medicine, infections, dermatology, surgery, and colon and rectal surgery.

For highly negative weights associated the increase of Chloropropamide and Glyburide Metflormin, we surmise that the increase in these dosages suggests that patient's situation for which they have been admitted is not severe. This is because both Chloropropamide and Glyburide Metflormin can only be used to treat cases of diabetes that are not severe (WebMD, 2017). For instance, these drugs cannot treat diabetic ketoacidosis, a symptom of severe shortage of blood sugar, a severe condition associated with diabetes.

Not every medical specialty appears to have a strong negative correlation; in fact, only specialties that either incorporate direct treatment of diabetes and/or symptoms of diabetes or treat cases that are not affected by diabetes are included. For instance, on the cases directly involving diabetes, diabetics often encounter skin complications (American Diabetes Association, 2017). On the other hand, sports medicine usually involves physical injuries such as broken ACLs, so diabetes would not factor in to the equation, especially since those who would get injured in the first place would most likely have their blood sugar levels in check since they exercise enough to obtain the injury in the first place. This hints treatment for these types of cases are more effective. Therefore, from this analysis, we can see that more work should be done researching cases where diabetes and another issue intermix for unusual results.

Conclusion/Future Work

Overall, we took a comprehensive approach to our investigation of our diabetes dataset. Given our dual objectives of developing the most accurate predictive model and developing a model that would provide doctors with greater information on a patient's chances of early readmittance, we conducted both an analysis of the coefficients from the logistic regression models that we developed as well as ran an extensive generation and selection process to find our most accurate predictive model.

As mentioned above, our group found analysis of the feature weights from the logistic regression models to be fairly interesting. These weights provide us with some information on what makes a patient more or less likely to be readmitted early. Among said features, medical specialty, and the prescription of certain drugs seem to be negatively correlated to early readmission, which - via other background information - hinted at how the usage of certain prescriptions can signal the severity of a case and/or illness and also what medical specialties are most effective.

Our highest-performing algorithms were our binomial logistic regression model, our multinomial logistic regression model, and our random forests model, although overall our prediction algorithms did not perform as strongly as we had expected. We believe that the class imbalance within the readmitted_early variable of our dataset biased our algorithms in such a way as to predict non-early readmittances with much greater frequency and confidence. This belief is supported by our confusion matrix investigation. Thus, if we had more time and resources, we would have adjusted our algorithms to more heavily penalize false negatives in any future investigations. Furthermore, in the future, we would also be interested in running predictions on other aspects of the dataset (e.g. predicting A1C values).

Contributions

Charlie - I cleaned and preprocessed the dataset to eliminate irrelevant information and make the data usable for our machine learning algorithms, and wrote the Dataset and feature section for the paper. Furthermore, I defined the error metric that we would use as well as defined and wrote the code for our baseline for error analysis. I also helped define the goals and potential applications for our projects.I worked on the binomial logistic regression, multinomial logistic regression, elastic net, and random forests model,

Stephone - I helped clarify the goals of our project, wrote the code for the SVM models, and helped in the interpretation of our results. I helped design the poster and wrote fractions of this very final writeup. I also graphed our results and ROC curves.

Christina - I found the data set, wrote the code for the Naive Bayes model as well as the Neural Network. I also wrote code that converted the categorical data to one hot versions. I also helped interpret our results, design the poster, and write the final writeup.

References

Breiman, Leo. "Random Forests." */Www.stat.berkeley.edu/~Breiman/*, Jan. 2001, www.stat.berkeley.edu/~breiman/randomforest2001.pdf.

"Glyburide-Metformin Oral : Uses, Side Effects, Interactions, Pictures, Warnings & Dosing." *WebMD*, WebMD, www.webmd.com/drugs/2/drug-19823/glyburide-metformin-oral/details.

Hall, M.A. (2000). Correlation-based feature selection of discrete and numeric class machine learning. (Working paper 00/08). Hamilton, New Zealand: University of Waikato, Department of Computer Science.

Hospital Readmissions Reduction Program. (n.d.). Retrieved December 13, 2017, from https://www.medicare.gov/hospitalcompare/readmission-reduction-program.html

Kiran, R. P., Delaney, C. P., Senagore, A. J., Steel, M., Garafalo, T., & Fazio, V. W. (2004). Outcomes and prediction of hospital readmission after intestinal surgery. *Journal of the American College of Surgeons*, *198*(6), 877-883.

Philbin, E. F., & DiSalvo, T. G. (1999). Prediction of hospital readmission for heart failure: development of a simple risk score based on administrative data. *Journal of the American College of Cardiology*, *33*(6), 1560-1566.

Shameer, K., Johnson, K. W., Yahi, A., Miotto, R., Li, L. I., Ricks, D., ... & Moskovitz, A. (2017). Predictive modeling of hospital readmission rates using electronic medical record-wide machine learning: a case-study using Mount Sinai Heart Failure Cohort. In *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2017* (pp. 276-287).

"Skin Complications." *American Diabetes Association*, www.diabetes.org/living-with-diabetes/complications/skin-complications.html?referrer=https%3A%2F%2Fwww.google.com%2F.

"The Staggering Cost of Diabetes." *American Diabetes Association*, 2017, www.diabetes.org/diabetes-basics/statistics/infographics/adv-staggering-cost-of-diabetes.html

Zhu, Meiying et al. "Mortality Rates and the Causes of Death Related to Diabetes Mellitus in Shanghai Songjiang District: An 11-Year Retrospective Analysis of Death Certificates." *BMC Endocrine Disorders* 15 (2015): 45. *PMC*. Web. 21 Nov. 2017.

Code Libraries Used: Keras (keras.ai), Numpy, Pandas, MatLab basic packages (fitcsmv, loss), tidyverse, caret, caretEnsemble, nnet, ROCR, glmnet

References for our Dataset

Beata Strack, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios, and John N. Clore, "Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records," BioMed Research International, vol. 2014, Article ID 781670, 11 pages, 2014.

" Diabetes 130-US Hospitals for Years 1999-2008 Data Set ." *UCI Machine Learning Repository: Diabetes Data Set*, UCI Center for Machine Learning and Intelligent Systems, 3 May 2014, archive.ics.uci.edu/ml/datasets/diabetes 130-us hospitals for years 1999-2008.