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

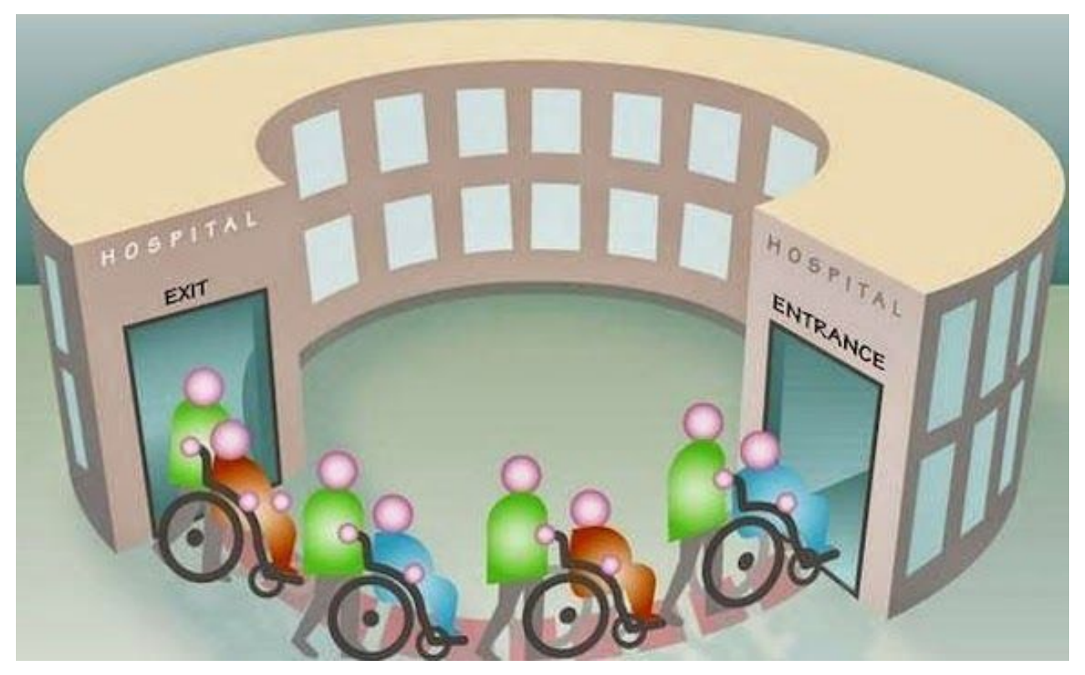
Charlie Xu, Christina Pan, Stephone Christian

{cxu2, capan, stephone}@stanford.edu

world diabetes day
14 November

Motivation

- The US spends $332 billion on diabetic and prediabetic care every year[1]
- Diabetes affects 350 million people with 3 million dying each year due to complications
- Through analyzing diabetic patient data, we hope to help reduce the mortality rate of diabetic people through improving patient care and decrease its cost
- We will analyze data from 130 hospitals in the United States from 1999 to 2008 to create 2 models that will:
  1. Predict whether a diabetic patient will be readmitted to the hospital in less than 30 days (i.e., a binary model)
  2. Predict the probability of readmittance within 30 days for a diabetic patient
- Doctors can use these models during patient visits to guide patient care decisions
- The binary model can be used to infer general patterns in the data and for holistic research on early hospital readmittance



Method

- We decided to pursue a variety of machine learning techniques and assess their performances to find our optimal model

Algorithms Used:

- Binomial Logistic Regression
  - Easy to interpret results
  - Output naturally fulfils our probability model objective
- Multinomial Logistic Regression
  - Same as binomial logistic regression, but simply uses a multinomial function
- Elastic Net
  - Same benefits of logistic regression
  - Also uses both L1 and L2 regularization, which minimizes overfitting
    - L1 regularization also chooses features that contribute the most to the probability of early readmittance
- SVMs
  - Used to potentially infer geometric order from the data
- Random Forests
  - Very flexible and generally highly accurate
  - Resistant to overfitting
  - Can also output probabilities

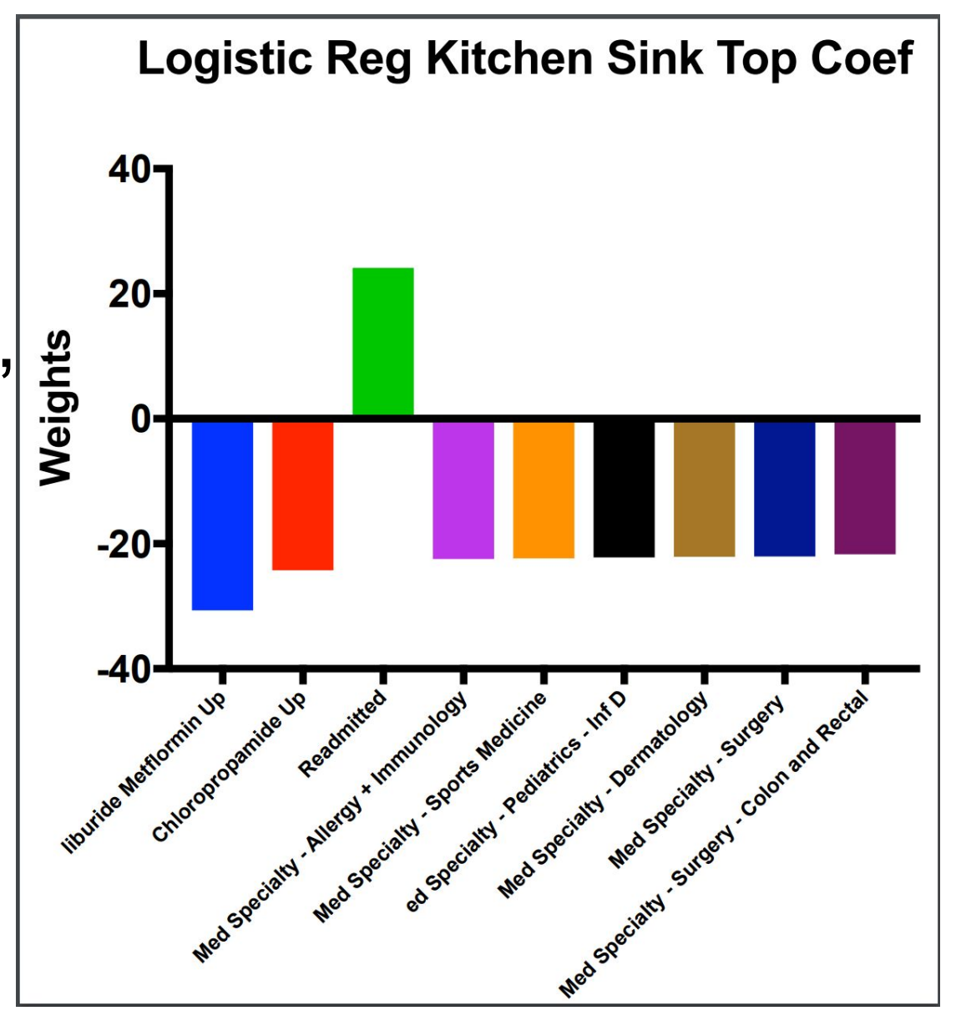
Analysis and Results

Preprocessing:

- When running our experiments, we discovered 7 columns that provided little predictive value
  - They were too sparsely populated and were eliminated from the dataset

We looked at the coefficients for our logistic regression models for some preliminary inference analysis. Overall, we found:

- Early readmittance is negatively correlated with seeing doctors with the specialties of Immunology, Sports Medicine, Pediatrics, Surgery, Colon & Rectal Surgery
- Early readmittance is highly correlated with overall readmittance
- Early readmittance is highly negatively correlated with the increased dosage of Chloropropamide & Gliburide Metaflormin



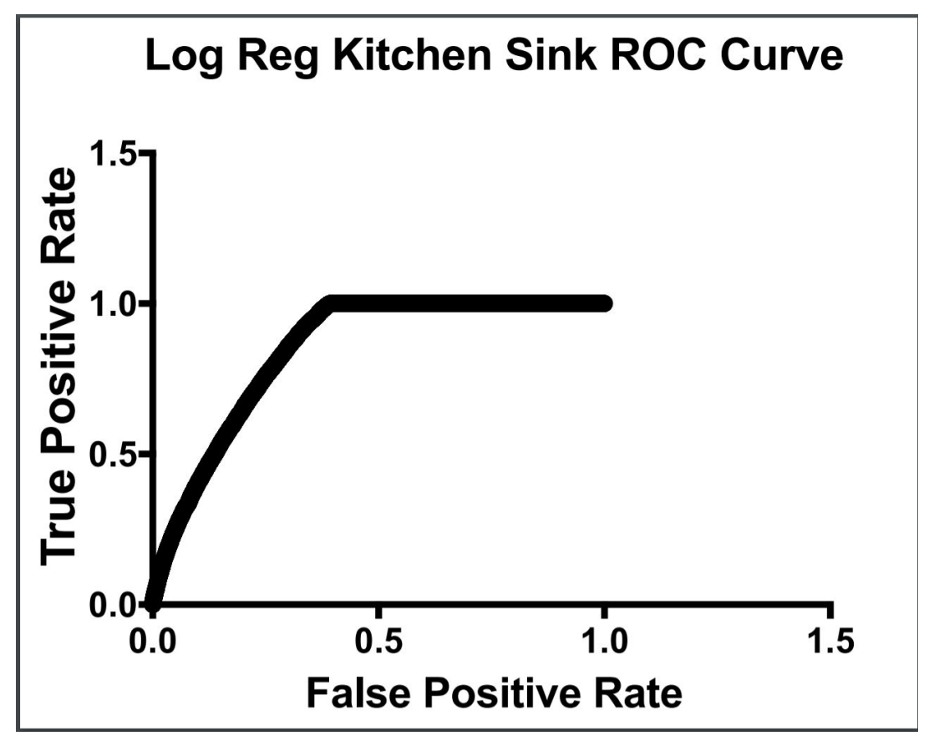
Analysis of Models

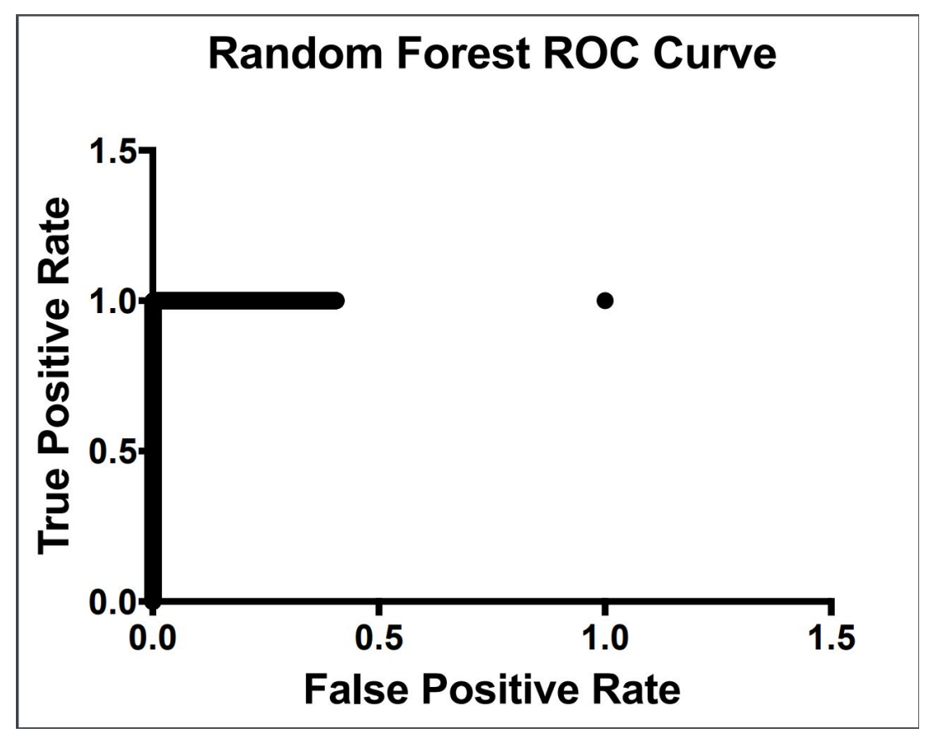
- Established 0 -1 loss as standard metric by which we would assess accuracy
- Use 0-1 loss of naive classifier, which would predict the most frequently occurring label (in our case, 0) as our baseline for comparing accuracies

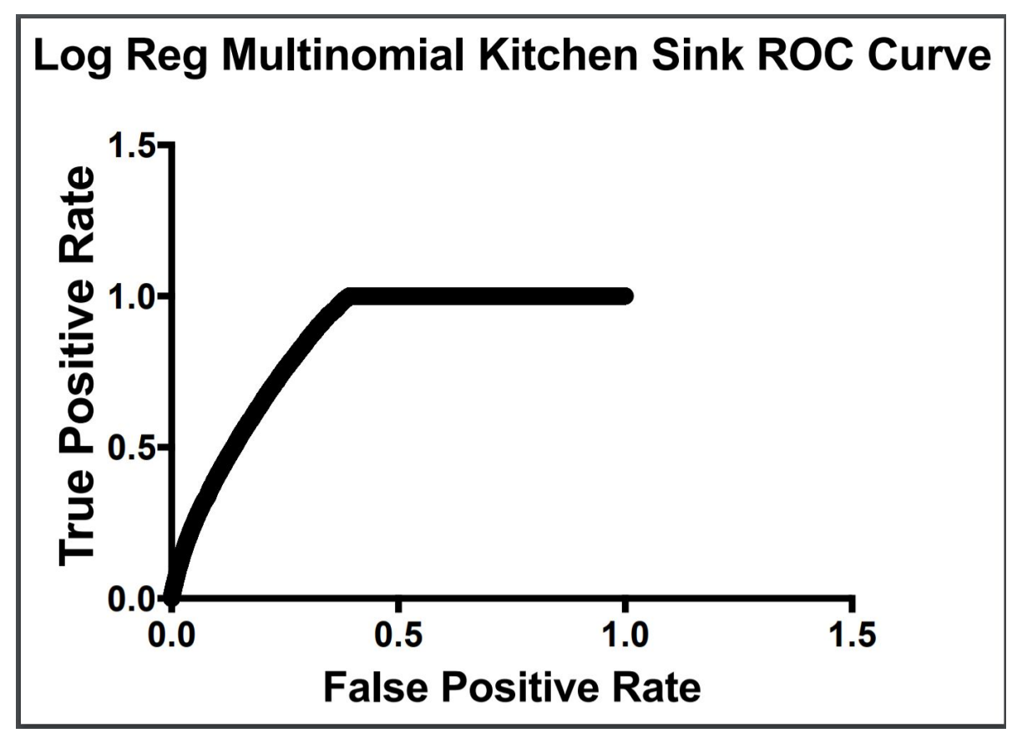
The 0-1 losses and accuracies of our models on the training set are shown below

Model Used	Model Description	Train 0-1 Loss	Train Accuracy
Naive Classifier	Predict most common label (in our case, 0)	0.1104794	0.8895206
Binomial Log Reg	Used all features (kitchen sink)	0.1102688	0.8897312
Binomial Log Reg	Used features determined to be significant in kitchen sink model	0.1107321	0.8892679
Multinomial Log Reg	Used all features	0.1102969	0.8897031
Elastic Net	Log Reg with both L1 and L2 regularization; Used all features	0.1034323	0.8965677
Random Forests	Use of bagging and boosting to generate set of decision trees	0.01532954	0.98467046
SVM	SMO solver, no Kernel	.1123	.8877
SVM	SMO solver, 3rd order polynomial Kernel	.1618	.8382
SVM	SMO solver, gaussian Kernel	.1123	.8877

For greater insight into their performance, we found our models' ROC curves and AUC values. The ROC curves of our 3 most accurate models are below:







Finally, to better understand the types of errors that our classifiers were making, we generated a confusion matrix for each model. The confusion matrices for our 3 most accurate models are included below:

LRB,KS	Act. Vals	
Pred. Vals	0	1
0	63191	7681
1	174	189

Rand. Forests	Act. Vals	
Pred. Vals	0	1
0	63365	1393
1	0	6477

LRM, KS	Act. Vals	
Pred. Vals	0	1
0	63188	7680
1	177	190

After our analyses of our models, we determined that our Random Forests model was the best model to use, given its high accuracy, excellent ROC curve, and its highly favorable confusion matrix values. Thus, we then ran our model on the test set, which yielded a test accuracy of 0.8877752

Conclusions

Weight analysis:

- Consulting doctors with particular specialties (e.g., pediatrics, sports medicine, dermatology, colon & rectal surgery) are very negatively correlated with readmittance
  - For diabetics, either the doctors in their specialties frequently deal with cases strongly related to diabetes or cases that are orthogonal to diabetes
- Having been readmitted is a strong hint towards future readmittance
  - Hints that if a case's root cause has been resolved, there is low likelihood of readmittance
- Drug prescriptions indicate the relative severity of the problem
  - E.g., an upwards dosage of chlorpropamide has a very negative weight; since chlorpropamide only deals with minor diabetic issues, this would signal the the case is most likely not very severe
- Analysis of correlation trends can be very valuable to both doctors and hospitals
  - These models can help doctors make patient care decisions
    - E.g., the strong correlation between past readmittance and early readmittance could help doctors be particularly vigilant on readmitted patients
  - Models for prediction can help hospitals with scheduling as well as help insurers more effectively gauge risk

Future Areas for Investigation

- On top of our existing models, we aim to create a classifier featuring a neural network
  - With this neural network, we will experiment with different activation functions
  - In addition, we will do additional analysis on the results of this models
- Once these models are created, we also aim to do some inference analysis into covariates that were determined to be significant

Acknowledgements

Thank you to Lucio Dery, Professor Boneh and Professor Ng for their guidance and support!

Citations

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

2.) 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.

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.