Early hospital readmission rates of patients with diabetes

Kalyani Asthana, Emily Brunelli, Preethi Seshadri

Introduction

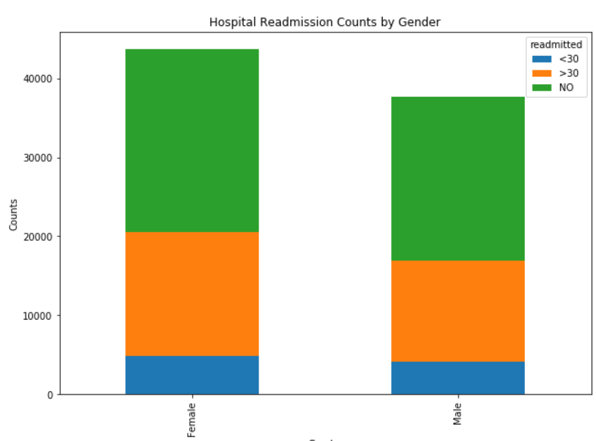
Diabetes is a serious, prevalent issue in the United States -- about 10% of the US population has diabetes and it is the 7th leading cause of death. To our understanding, there have not been many comprehensive studies connecting inpatient care and management with long-term health outcomes. This line of work can be useful for motivating work on proper protocols and prioritizing preventative care. Likewise, in total, early hospital readmissions contribute to 41.3 billion dollars in hospital costs, according to a report from the Agency for Healthcare Research and Quality[3]. Identifying factors that even moderately contribute to readmission rate has the potential to save a lot of money.

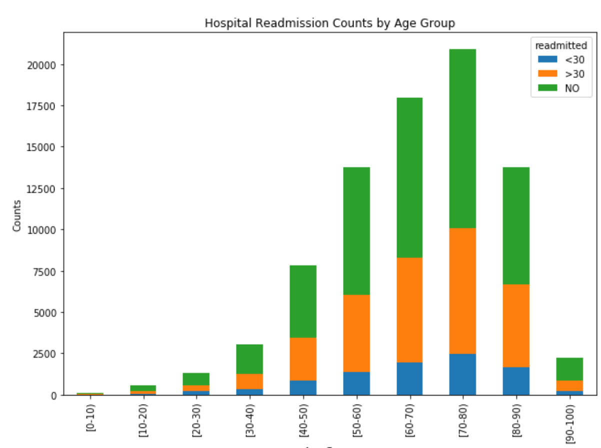
Our dataset (collected from the Health Facts database) consists of roughly 102,000 encounters in which a patient was admitted to a hospitals in the United States, diabetes was entered into the system as a diagnosis, the patient stayed in the hospital between 1 and 14 days, lab tests were performed, and medications were administered. This was collected from 130 hospitals over the span of decade (1999-2008). The original dataset had 55 features, including information about 24 different medications, demographic information about the patient (age, gender, etc.), whether or not there was a change in medication, whether or not the patient was readmitted within <30 days, >30 days, or not at all. (A full overview of original features can be seen by looking at the Strack et. al. paper.)

Our goal with this project is to accurately predict which patients are readmitted to the hospital within 30 days of their initial hospital visit, and better understand which factors contribute to early readmission.

Data

Through initial data exploration, it became clear that there were no obvious differences in hospital readmission across various demographic groups. For example, based on the plots below it is not clear that being a male vs. female patient or being 55 vs. 85 years old makes a noticeable difference in early readmission rates. We also observed similar trends in hospital care related attributes.





The other aspect that stood out was that only a small subset of patients had an early readmission. With imbalanced classes, it is extremely simple to achieve strong accuracy without predicting anything of value. Our initial concern was that the roughly 89-11 split between the majority and minority class would influence our classifiers to predict “no readmission” by default, and this became apparent once we constructed initial models.

There are various approaches to handle class imbalances, such as higher penalties for minority class errors in the loss function. Another approach is to sample data, although oversampling makes exact copies of existing examples which can lead to overfitting and undersampling throws out useful data. We decided to use SMOTE (synthetic minority over-sampling technique) as an attempt to mitigate this issue. As its name suggests, SMOTE generates synthetic data as opposed to duplicating data instances.

In terms of feature selection and data processing, we removed the Encounter ID and Patient Number attributes since they are unique identifiers of the encounter and patient respectively. The weight, payer code, and medical specialties contained a significant proportion of missing values, so we removed those attributes altogether. We also noticed that many of the medication attributes contained nearly all “no” values (indicating that medication was not prescribed). Medication attributes falling into this category were removed, since these would be uninformative for classification. We kept the remaining numerical columns without making any modifications and also treated age as a numerical attribute (converted the age ranges to single values using the midpoint, e.g. [0,10) maps to 5, [10, 20) maps to 15, etc.). We then constructed one-hot encoded representations of the remaining categorical attributes, with a few modifications described in the table in our appendix.

At first, we wanted to group patients into either “<30, >30” or “NO”, i.e. any patients that were readmitted at any point would be in the same category, while patients that were not readmitted had their own category. After looking further into the Strack et. al. paper, what they did was split patients into “<30” and “>30, NO”, because they were only concerned with early readmission. We switched to that grouping, reasoning that patients with late readmission might be more similar to non-readmitted patients, as opposed to early readmitted patients.

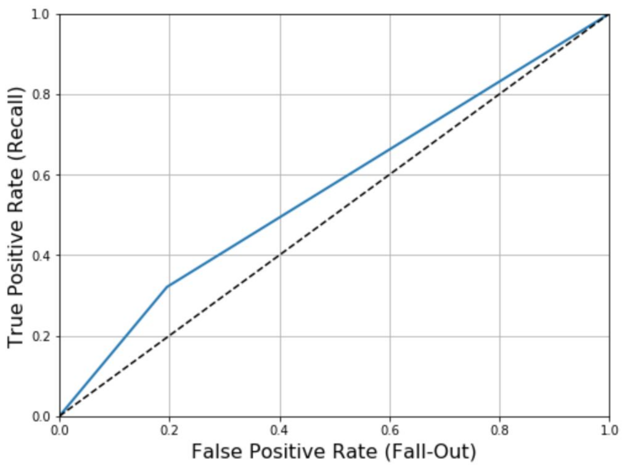
Method and Results

We focused on three groups of models: linear classifiers (logistic regression), ensemble methods (random forests), and neural networks (multi-layer perceptron). Logistic Regression was chosen as a baseline model. Tree-based models such as random forests are typically effective for tabular data, although the data in our case is fairly sparse from one-hot encodings. Random forests are generally well-suited for overfitting because of bagging. Neural networks are more appropriate for capturing complex, nonlinear relationships.

To combine SMOTE with our cross validation pipeline, we ensured to only apply SMOTE to the training folds and not the validation fold. Using SMOTE on the entire training dataset before cross validation can inflate validation performance, since the validation data is also oversampled. We also standardized our numerical features by making them zero mean and unit variance and incorporated this into our cross validation pipeline.

Our initial choice of logistic regression was motivated by the approach in the Strack et al. paper. We tested out L1 and L2 regularization, as well as varied the C parameter. Our best AUROC score was only .557, using L1 regularization and C = 4.64 and was a quick indicator that we needed to move onto more complicated models, as the logistic regression model was severely underfitting.

The next model we tried was a random forest, where we ran cross-validated grid search and found that the best hyperparameters were 100 estimators with depth 6. This gave us an AUROC of 0.579, recall of 0.033, and accuracy of 0.873 which is comparable to the logistic regression performance.

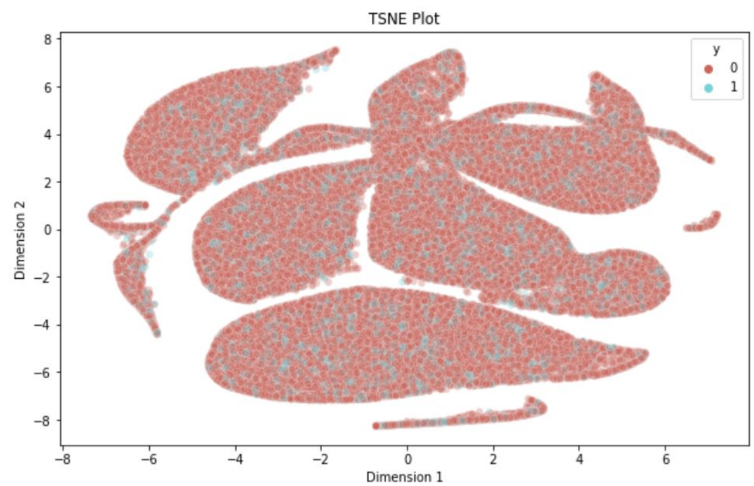


Ultimately, our best model was a neural network, where we also ran a cross-validated grid search and found the best model to be the neural network with 3 hidden layers, with 20, 20 and 15 nodes in each layer respectively. The model was trained using the RELU activation function with a constant learning rate of .015. This model gave us an AUROC of 0.58, a recall of 0.17, a precision of 0.18 and an accuracy of 0.82 on our validation set.

Finally, we chose the above neural network as our best model to make predictions on the test set. We got an AUROC of 0.52, a recall of 0.09, precision of 0.18 and an accuracy of 0.85 on our test set.

Discussion and Future Directions

We were unable to capture a meaningful relationship between our features and early readmission. All of our proposed models were prone to underfitting, since both training and cross validation performance were extremely poor. We also leveraged t-SNE as shown on the right, which performs nonlinear dimensionality reduction and is commonly used to visualize high dimensional data. Ideally we would see some meaningful clustering or groupings based on type of readmission (1 is <30 and 0 is >30/NO in the plot) if the classes were distinct, but instead the data overlaps heavily. This aligns with our findings and indicates that this problem is difficult to solve.



Overall, we were not impressed with our results and believe this project highlighted the challenges of working with real-world, noisy, imbalanced data. In the future, there are a number of other directions we could take with this dataset, including but not limited to:

- Further exploration of resampling techniques: SMOTE did not improve our performance meaningfully, and in some cases led to worse performance. It's possible that undersampling the majority class would have been a better choice for this data. We also did not do an in-depth exploration of the various SMOTE ratios, and instead focused predominantly on a 50-50 split.
- Although specific metrics utilize various decision thresholds (e.g. ROC looks at the tradeoff between TPR vs. FPR for various prediction thresholds), it could be useful to pay more attention to the actual probabilities as opposed to the binary predictions and utilize this information to make more nuanced predictions. In reality, what might be more informative is to have certain patients flagged as "at risk

for readmission" so that steps may be taken to provide better care for them in particular to avoid another hospital visit.

- A more thorough examination of medication usage: Many of the medications in the dataset had extremely low use rates (e.g. 2 or 3 occurrences across the entire dataset) and it was difficult to decide how to properly utilize this information. We think that the addition of a summary statistic regarding medication usage for each patient might be more valuable. Additionally there could be a deeper dive into engineering features that incorporate medical domain knowledge.
- Modifying what is being predicted: We may consider changing the classification to "<30, >30" and "NO" instead, since readmissions as a whole might be a more natural grouping. Additionally, it could be beneficial to narrow the scope of our analysis to a subset of the dataset. For example, the Strack et al. paper focuses on the impact of the HbA1c, so we could reduce our focus to encounters that include this measurement.

Appendix

Description of feature processing for select categorical attributes.

Feature	**Description**
Gender	Male, Female, and Unknown/Invalid (removed Unknown/Invalid because there are only 2 entries)
Admission type	Emergency, Urgent, Elective, Newborn, Trauma, N/A (this groups Not Available, Null, and Not Mapped into one group)
Discharge Disposition	Originally 29 values but only kept Discharged Home and grouped remaining categories into Other (since Discharged Home contains majority of the data)
Admission Source	Originally 21 values but grouped categories that corresponded to Referral into one value, Emergency Room into another, and all remaining values into Other (since previous two groups contains majority of the data)
Diagnosis 1	The primary diagnosis - mapped the first three digits of ICD9 code to diagnosis group (Circulatory, Respiratory, Digestive, Diabetes, Injury, Muskuloskeletal, Genitourinary, Neoplasms, and Other)
Diagnosis 2	The secondary diagnosis (using same mappings as primary)
Diagnosis 3	An additional secondary diagnosis (using same mappings as primary)

References

[1] Bhuvan, Malladihalli S., et al. "Identifying Diabetic Patients with High Risk of Readmission." *ArXiv:1602.04257 [Cs]*, Feb. 2016. *arXiv.org*, http://arxiv.org/abs/1602.04257

[2] Beata Strack, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios, 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. https://doi.org/10.1155/2014/781670

[3] Anika L. Hines, Ph.D., M.P.H., Marguerite L. Barrett, M.S., H. Joanna Jiang, Ph.D., and Claudia A. Steiner, M.D., M.P.H. "Conditions With the Largest Number of Adult Hospital Readmissions by Payer, 2011", April 2014, *Healthcare Cost and Utilization Project,* https://www.hcup-us.ahrq.gov/reports/statbriefs/sb172-Conditions-Readmissions-Payer.pdf

[4] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyera, "SMOTE: Synthetic Minority Over-sampling Technique", arXiv:1106.1813, June 2011, *arXiv.org*, https://arxiv.org/abs/1106.1813

Division of Tasks

Overall our work was split evenly and we worked on this project collaboratively. We have roughly split the tasks as follows:

Kalyani: Worked on the Neural Network model, test set performance
Emily: Worked on logistic regression model, neural net exploration, cross val pipeline used in other models
Preethi: Worked on data processing/feature extraction pipeline, ensemble models

Data exploration, visualization, and project writeup was done by everyone.