**Predicting Diabetic Patient Readmissions**

**Introduction**

The objective of this project is to develop machine learning models that will predict whether diabetic hospital patients will be readmitted within 30 days. For a bit of context, the Affordable Care Act created the Hospital Readmission Reduction Program to improve the quality of healthcare for Americans by tying hospital payments to patient readmission rates. In other words, this program seeks to incentivize hospitals to provide high quality healthcare by financially penalizing hospitals with higher readmission rates.

This project endeavors to aid in achieving the goals set forth by the Hospital Readmission Reduction Program by positively impacting two key stakeholders in the American healthcare system:

1. **Patients:** Correctly identifying patients who are likely to be readmitted and will enable hospitals to take preventative measures so that patients will receive the best possible healthcare and avoid adverse health consequences
2. **Hospitals:** Tying lower readmission rates to higher hospital revenues will foster efficiency from a financial perspective, allowing hospitals to reallocate a larger pool of resources to important expenses such as personal protective equipment

In short, lower readmission rates lead to better health for patients and better fiscal health for hospitals.

**Dataset**

The dataset being analyzed in this project was procured from over 10 years (1999-2008) of clinical care in 130 hospitals for the purposes of analyzing the impact of HbA1c (blood sugar level) on hospital readmissions. It is suitable for this project of developing a predictive model for hospital readmissions because it contains over 50 features with each instance representing patient and hospital outcomes (whether they were readmitted to the hospital).

To qualify as an instance in this dataset, each observation has met the following criteria:

1. It is an inpatient encounter (a hospital admission)
2. It is a diabetic encounter, that is, one during which any kind of diabetes was entered to the system as a diagnosis
3. The length of stay was at least 1 day and at most 14 days
4. Laboratory tests were performed during the encounter
5. Medications were administered during the encounter

**Citation:** 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.

**Process**

The process for this project entails the training of Decision Tree, Logistic Regression, Random Forest and XGBoost algorithms. The models will then be evaluated on the AUC (Area under (ROC) Curve), which serves as a measure of model performance across all possible classification thresholds. After plotting and determining each model's AUC, feature importances will be plotted for each model, whenever possible, to discern which factors are most likely to contribute towards a patient's readmission.

Once each of the algorithms are evaluated at a base level (default parameters), we will try to optimize each algorithm through hyperparameter and other machine learning techniques (bootstrap aggregating and stacking) to see if more robust models can be created.

The basic outline of what is described above can be summarized as follows:

1. Data Processing
2. Training Base Machine Learning Algorithms
3. Evaluating Base Machine Learning Algorithms (AUC / Feature Importance)
4. Hyperparameter Tuning with GridSearchCV + Other Machine Learning Techniques (Bagging / Stacking)
5. Evaluating Optimized Machine Learning Algorithms (AUC / Feature Importance)

**Data Processing**

Before diving into training our base machine learning models, it is crucial to examine the data and process them so they are suitable for machine learning models. In terms of first steps, features must be processed to address a number of issues that must be resolved before training machine learning models. When observing this particular dataset, the two prevalent issues are missingness and an abundance of categorical features.

**Missingness**

With regards to missingness, features with missing values must be imputed or dropped (if most of the data is missing or if the variance is low) as done below:



**Numeric and Categorical Features**

Once the issue of missingness is resolved, the next step is to transform numeric and categorical features. One thing that is clear when taking a closer look at this dataset is that few key categorical features such as A1C result (which measures recent blood sugar levels) are actually quantifiable features that can be measured. Thus, as a part of the data processing, proxy values have been mapped to categories for such features as shown below:



Once numeric features and categorical features are properly delineated, numeric features are scaled and categorical features are encoded so that all data are numeric in line with what is suitable for training machine learning models.

**Imbalanced Dataset**

One issue that is noticed when examining the output variable is that the dataset is heavily imbalanced towards the majority class (not readmitted). If we do not address this issue, our models will have poor predictive power for the minority (readmitted) class.

To remedy this problem, we can implement the Synthetic Minority Oversampling Technique (SMOTE) to upsample the minority class in our training data set.

**Base Models**

**Optimized Models**

Hyperparameter Tuning

Bagging

Stacking

**Conclusion**

After evaluating the machine learning models that we have trained in this project, we were able to achieve a maximum AUC of 0.63 with the following models:

1. Logistic Regression Classifier
2. Tuned Logistic Regression Classifier
3. Bagged Logistic Regression Classifier
4. Tuned Random Forest Classifier
5. Stacked Classifier

In terms of preemptive actions that can be taken to reduce patient readmissions, leveraging a few of features deemed as important by our highest performing machine learning models trained in this project can help:

* Number of Inpatient visits: patient intervention teams should pay close attention to frequent visitors to the hospital and appropriate log the reason for each visit to address the root cause of admission
* Diabetes Medication: more medications could indicate the potential for more complications - as such key barometers of diabetic health such as a1C (blood sugar levels) should be tracked
* Number of Diagnoses: diabetes in conjunction with other health conditions can lead to severely adverse complications - hospitals should ensure that all health conditions are addressed
* Age: patient intervention teams should maintain regular and routine communication with older patients to anticipate any reason for readmission

All things being said, thank you very much for taking the time to go through my analysis of diabetic readmissions. In this notebook we learned how to implement the Decision Tree, Logistic Regression, Random Forest, and XGBoost algorithms. Additionally, we tried to improve the performance of our models through hyperparameter tuning as well as useful machine learning techniques such as bootstrap aggregation and stacking. For models where applicable and interpretable, we leveraged feature importance to discern which aspects of a patient encounter could serve as a flag for hospitals to take action to prevent a patient from being readmitted.

**Future Work**

In terms of future work, the feature engineering space would be one lever that can be used to improve model performance. The way to best do so, in our view is to conduct some qualitative research to strengthen domain expertise in areas such as medications and other health conditions that could have a big impact on diabetic patients.

Being able to better contextualize all of these features with domain knowledge, would no doubt, lend towards a better patient experience!