Predicting Diabetes Hospital Readmissions

***Abstract*—This paper implements a classification model for predicting hospital readmissions of diabetic patients. It also analyses the possible biases present in our dataset and model.**

I. Project Proposal

A. Introduction

Classifying at-risk hospital patients is an already well-researched task [1]. We aim to specify this approach to diabetic patients using a logistic regression model to predict whether a diabetic patient will or will not be readmitted to hospital.

The number of cases of diabetes has risen drastically from 108 million in 1980 to 422 million in 2014. With diabetes, comes the increased likelihood of major life altering, or even life-threatening, illnesses such as blindness, kidney failure, and heart attacks. In 2019, an estimated 1.5 million deaths were caused by diabetes [2]

With a functioning classification model, we will be able to triage patients more accurately based upon their predicted classification of readmittance. At a hospital level, this will reduce the wastage of medical facilities and supplies, which are generally in short supply and have high usage and maintenance costs. At a personal level, we will be able to alleviate the various physical and mental effects of frequent hospital admissions as well as to ideally reduce the mortality rate of diabetic patients.

Diabetes has two distinct variants – type 1 and type 2 – with the latter being caused by obesity which is becoming increasingly prevalent in developed countries. In type 1 diabetes, the pancreas fails to produce enough insulin for the body whilst in type 2 diabetes the body is unable to effectively use the insulin it produces. This fundamental difference results in different symptoms and severities, so it is important for our model to differentiate between the two.

B. Dataset

Our dataset [3] contains 10 years (1999-2008) of hospital admission information, from 130 US hospitals, regarding 100,000 patients with diabetes. It supplies relevant attributes such as patient identification, demographics, diagnoses, inpatient statistics, and medical administration.

However, one of the limitations of this dataset is its inability to differentiate between type 1 and type 2 cases which may result in a reduction in our model's performance due to the large differences between the two variants as discussed earlier. There may be other features that correlate with the diabetes type, such that our machine learning algorithm will be able to implicitly differentiate between them, but this will not have as strong an influence on our model as an explicit binary feature.

The data is now over a decade old, which may also have an influence over our predictive capacity in a current real-world implementation of our model. For example, we may find that a specific medication is no longer administered; what our model may have previously used as a strong predictor of readmission, is now no longer appropriate.

C. Bias

The dataset contains demographic information such as race, gender, and age which need to be respected when training our model as possible demographic biases may occur. We do not investigate the fairness of gender and age as it is expected that these features *will* contribute to the likelihood of someone being readmitted. For example, it is obvious that older people will have a higher chance of being readmitted due to a higher vulnerability to a larger number of diseases.

Without respecting the possible biases surrounding race, we may unfairly misclassify individuals that would result in insufficient treatment. For example, there may be an inherent structural bias in the dataset that affects medication distribution based on race. Possible explanations for these biases may be the result of systematic racism in the United States medical system [4] or from the collection and selection process of our dataset.

D. Methodology

First, we clean, bin, encode, normalise, and reduce any elementary bias in our dataset so that it is applicable for our given task and model.

We choose to implement a logistic regression algorithm and must be aware of its inherent assumptions. It assumes independent observations, a binary dependent variable, and a large sample size, which are all satisfied by our dataset. However, it also assumes linearity between the independent variables and log odds, as well as little or no multicollinearity between independent variables. We do not explicitly check either. Disregarding these assumptions is likely to lead to reductions in our model's performance.

We then balance the dataset with respect to our protected race attributes and apply the same logistic regression model in an attempt to identify any historical bias in our dataset relating to this demographic. We compare the performance and bias of this model against our previous model.

Now that we have observed this bias, we implement a fair machine learning solution to mitigate it and analyse the results. Our final model should remain appropriate to the task and ideally bias free.

Our analysis is performed using python. We use pandas and numpy for general data and mathematical manipulation, scikit-learn for implementing our classification model, and matplotlib and seaborn for graphically representing our results.

II. Data Analysis

A. Dataset Cleaning

We begin by extracting the ground truth label from the 'readmitted' feature which differentiates between no readmission, readmission within 30 days, and readmission outside of 30 days. We choose readmission within 30 days to represent our positive label, since the longer the time before readmission, the more likely that it is due to an unrelated illness rather than the patient's treatment in hospital.

We then mitigate some elementary bias within our dataset. The 'encounter id' and 'patient nbr' features need to be removed as they provide a direct correlation with our label. We also remove duplicate patient encounters to ensure that our observations are statistically independent, which is a required assumption for logistic regression. Finally, we drop observations that result in either 'death' or 'discharge to hospice' since it is evident that these cases will not be readmitted to hospital.

We also remove almost all medication data since the distribution of useful information – values that signify medication was used – is extremely sparse and contributes very little to our model. Based on the number of missing entries for each feature, we choose to drop columns that have 25% or more missing values and subsequently drop all remaining rows with any missing data.

Based on the previous efforts made by Strack et al. [3] and using 'id_mappings.csv' as a reference, we bin many features with an excessive number of unique values. The exact binning process is shown in 'Create Mapping Dicts.py'.

Our models require strictly numerical data as inputs. To satisfy this, we encode our categorical data using 'one hot encoding', creating many binary dummy variables for each unique value of each feature. This is superior to 'label encoding', which simply converts string values to a numeric counterpart since it does not create a false correlation between values in categorical features.

Although not strictly necessary for logistic regression, we also normalise our (non-label) data to between 0 and 1 since it increases the speed in which logistic regression converges and allows for comparable features.

B. Demographic Analysis

After splitting our dataset into its demographic race groups (African American, Asian, Caucasian, Hispanic, and Other), we perform some simple analysis on each group against the available features[1].

The first obvious check for bias is to assess the proportion of each group that are readmitted. We find that there is a ~36% higher chance of being readmitted if you are Caucasian than if you satisfy the race group 'Other'. It is worth noting that the group Other is severely underrepresented in our dataset, only having 960 cases against the 42684 regarding Caucasians. As a result, the mean of the readmitted feature for Other is much more likely to be skewed by a few outliers.

We find that the mean values for the number of lab procedures, number of medications, and number of outpatient visits for Caucasian patients are higher than those of the other groups which is likely the cause of the higher readmittance rate. It may also be the results of an inherent bias within the medical system or our dataset.

It is also worth noting that the mean values of the time in hospital, number of lab procedures, number of medications, number of outpatient visits, and number of emergencies is significantly lower for the Asian demographic. We also find that the variance for almost all the features is significantly smaller than that of the other groups in each respective feature. This narrower distribution about the mean implies a stronger likelihood of bias than that of the Caucasian group discussed earlier. However, as with the Other group, the Asian demographic is severely underrepresented which is likely responsible for the differences observed.

We also evaluate the 3 most common values for each categorical feature for each race group. Since most of these features were binned into fewer unique values, there is less deviation between the groups. However, we do notice one oddity regarding the African American demographic; the most common age (50-60) is much lower than that of all other groups. This is surprising as we expect more old people to be admitted to hospital, as shown by the other groups. However, this higher frequency of younger people is not reflected in the chance of being readmitted which lies just below the highest rate of the Caucasian demographic.

III. Conventional Implementation

A. Chosen Algorithm and Implementation

We choose to implement logistic regression as our classification algorithm. It is not too complex so efficient to train and does not take too long to optimise through hyperparameter tuning. Logistic regression also assigns coefficients to features which we can use to interpret the importance of each variable on predicting hospital readmissions.

We first split our dataset into 75% training data and 25% testing data. We notice that our data has an imbalance of approximately 1:8 positive to negative labels. Due to this imbalance, if a classifier were to predict only negative labels, it would achieve an accuracy of approximately 90% in our test dataset which is not functional since the positive labels are more valuable in our task.

Hence, we aim to increase the task-specific performance of our classification model using hyperparameter tuning. The scoring functions available are not appropriate for evaluating the imbalanced data so we create a custom scoring metric (balanced accuracy) that weights the true positive rate (TPR) and the true negative rate (TNR) equally.

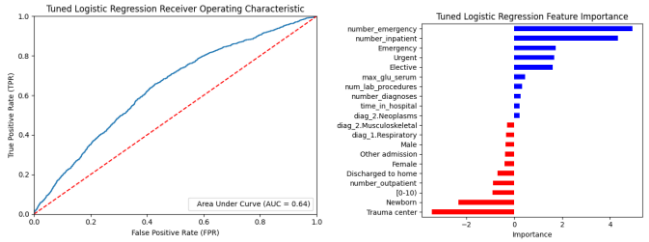
B. Algorithm Performance Evaluation

	precision	recall	f1-score	support
0.0	0.93	0.65	0.77	12745
1.0	0.14	0.55	0.22	1281
accuracy			0.64	14026

We achieve an overall accuracy of 64%. Although not amazing, it is similar to other logistic regression algorithms evaluated on this dataset [5] so we can affirm its consistency. Hence, it is safe to assume that the poor performance is more likely because of an inherent lack of correlation with the dependent variable within the dataset itself.

[1] All results are available through running the main implementation code

From the recall scores, we can see that our model struggle to correctly predict positive labels more than negative labels, this is most likely still due to the imbalance, although the extent of which we have reduced by hyperparameter tuning.



We can also use a receiver operating characteristic (ROC) curve to compare the TPR against the FPR at various thresholds. The blue line shows the ROC curve of our model whilst the red line represents a random classifier. We can see that our model performs better than random predictions which is quantified through use of the area under the curve (AUC) – a value of 0.5 is associated with a random classifier.

We also calculate the 10 best features for predicting the positive and negative labels respectively from our logistic regression models coefficients. This graphic shows that the number of emergency and inpatient visits are the most important predictors for a positive label as expected.

C. *Balanced Implementation and Evaluation*

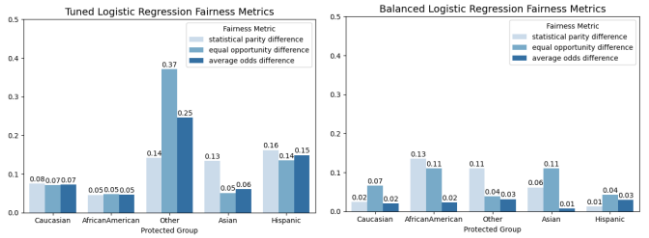
We also create a model from training and testing sets that have a balanced distribution of our protected race groups. As a result, we would expect to find no performance bias if our dataset contained no historical bias. For sake of brevity, the performance plots are not shown but are available in the given folder.

We notice very little change in precision from our first model, however we notice an unexpected increase in both recall scores. During this process, we end up losing a lot of data to match the count of the smallest race group. This is responsible for most fluctuations in our model's performance

Another option would be to generate synthetic data points – using SMOTE for example – to balance the groups, but this would introduce its own bias since all the new cases are generated from old ones so cannot introduce much variance.

D. *Bias Observations*

We use 3 fairness metrics to observe bias in our model: statistical parity, equal opportunity, and average odds. The bias of our two models is shown below where a result of 0 indicates there is no bias in the protected group.



We observe that our unbalanced model discriminates most severely towards the race group Other. This is most likely due to its under-representation in our dataset. When balancing our data to ensure equal representation of minority groups we can see that the overall bias is reduced. However, there is still bias prevalent across all race groups which we can conclude is due to historical bias in our dataset.

IV. Fair Machine Learning Implementation

A. *Fairness Mitigation Algorithm Implementation*

We implement the pre-processing bias mitigation algorithm called reweighing. It reweights each sample in our training set, with respect to each protected group to ensure fairness before classification.

It is preferred over in-processing and post-processing algorithms since it more effectively reduces bias without significant reductions on performance. In fact, we find that the performance of our reweighed model is practically identical to our first model.

B. *Fairness Mitigation Algorithm Performance Evaluation*

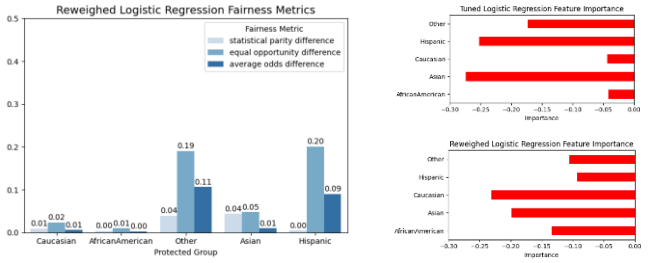
	precision	recall	f1-score	support	
0.0	0.93	0.65	0.77	9684	Caucasian
1.0	0.14	0.55	0.22	992	
	precision	recall	f1-score	support	
0.0	0.92	0.72	0.81	96	Asian
1.0	0.25	0.60	0.35	15	

We note that the Caucasian majority group performs almost identically to our dataset as a whole. As the most prevalent demographic and being hardly biased in our first model, we expect a similar performance as the reweighing would have a lesser effect.

If we had fully mitigated the bias, we would expect to see the same performance in the minority as the majority group. By comparing all metrics, we find that our model performs better on the Asian group. This shows that there is still some algorithmic bias prevalent towards the minority group which is supported by the fairness plots below.

It is also likely that this bias is due to or amplified by the size of the support of the Asian group since each prediction has a higher weighting on the performance metrics.

C. *Fairness Mitigation Algorithm Bias Observations*



We observe from the first plot that bias is mitigated in all protected groups compared against our unbiased algorithm. However, the previously most biased groups still remain significantly more biased after reweighing. It is interesting to note that statistical parity difference is reduced much more than the other two metrics. This is probably because statistical parity is the only metric not dependent on the original label.

From comparing the second plots, we observe that our models weighting on the three minority groups – Other, Hispanic, and Asian – is also reduced. This implies a reduction in bias however it comes at the cost of increasing the importance and therefore bias of our two previously least biased majority groups – Caucasian and African American.

[1] J. Miles, J. Turner, R. Jacques, J. Williams, S. Mason, “Using machine-learning risk prediction models to triage the acuity of undifferentiated patients entering the emergency care system: a systematic review”

[2] World Health Organisaton (WHO), <https://www.who.int/news-room/fact-sheets/detail/diabetes>

[3] B. Strack, J. P. DeShazo, C. Gennings, J. L. Olmo, S. Ventura, K. J. Cios, J. N. Clore, “Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records”

[4] J. Feagin, Z. Bennefield, “Systematic racism and U.S. health care”

[5] Y. Hu, M. Sokolova, “Explainable multi-class classification of medical data”