Assessing Machine Learning Models for Predicting Diabetes Hospital Readmission

Tony Ding

Department of Electrical Engineering and Computer Science, MIT

Dr. Aleksander Madry, Dr. Manish Raghavan, Dr. Ashia Wilson

December 13, 2022

Abstract

Diabetes is a common health condition, and those who were diagnosed with diabetes have to monitor their blood sugar levels for an extensive amount of time, usually their whole lives. If not well-managed, individuals may be admitted to healthcare facilities. In addition, hospital readmissions due to diabetes within 30 days may indicate that the patient is especially at risk of medical complications. Numerous predictive models had been built to predict hospital readmission; nevertheless, no previous studies had audited the predictive models for bias regarding hospital readmission. As a result, this paper aims to explore predictive models for the readmission of diabetic patients as well as to determine if there are any potential biases in the decision-making processes of these models. By applying models such as logistic regression, XGBoost, and random forest, we found that model designs for building a predictive program on a diabetes hospital readmissions dataset can indeed introduce a certain amount of bias. It was found that there were more Caucasian patients being readmitted to the hospital than actually needed, and non-Caucasian patients who should be readmitted to the hospital are not predicted by the models to experience readmittance as often compared to Caucasian patients. Furthermore, by comparing the XGBoost and random forest models, the XGBoost models tend to have more equitable predictions while still maintaining comparable predictive ability.

Keywords: diabetes, hospital readmission, machine learning, racial bias

Assessing Machine Learning Models for Predicting Diabetes Hospital Readmission

Introduction and Problem Definition

Diabetes is a chronic health condition where the body does not respond to or does not generate enough insulin in the body, resulting in issues with regulating blood sugar (Centers for Disease Control and Prevention, 2017). There is currently no cure for diabetes, so those diagnosed with diabetes must monitor and manage their blood sugar levels for the rest of their lives. If not controlled well, individuals may be readmitted to a healthcare system due to complications. Readmission within 30 days may indicate that something had gone wrong in the care provided or the patient is especially at medical risk of diabetes. Additionally, readmission is a costly economic burden and affects both the quality of life of patients and the distribution of medical resources. Moreover, the US spends significantly more on the healthcare sector compared to other countries, yet the healthcare outcomes and quality of care are lower than those of other countries that spend less (Lecture_3_Healthcare, 6.3950/6.3952). Despite concerns about the economics of readmission and the prevalence of diabetes, readmission among diabetes patients is not well studied (Shang et al., 2021). Moreover, readmission can be a possible measure of healthcare quality. In summary, readmissions impose significant costs and implications, and creating a model could provide insights to identify patients that are likely to experience readmission within a short period of time.

In this study, we are interested in exploring predictive models for the readmission of diabetic patients and determining if there are any potential biases in the decision-making of the models. We have identified a dataset containing data regarding patients admitted to a hospital for diabetes treatment. Despite previous work developing predictive models for this dataset, we did not find any studies that audited the predictive models for bias. Therefore, we believe that there is value in understanding how the design of the model may present different levels of bias.

Related Work

Multiple models had been built for predicting 30-day readmission for this dataset, but we found that none of them investigated potential bias in their models. In one study, Shang et al. (2021) compared three machine learning algorithms, namely random forest, Naive Bayes, and tree ensemble, and they optimized for the predictive ability of the models by maximizing the average AUC score. In another study, Alturki et al. (2019) designed and reviewed five machine learning algorithms, namely logistic regression, random forest, XGBoost, K-nearest neighbors, and support vector machine. Nevertheless, none of the above studies compared performance metrics across various patient demographics.

When building machine learning models, especially in the healthcare industry, it is important to audit the models and algorithms for bias since inequitable decision-making from models is extremely undesirable. In healthcare, biases can be easily injected into medical records because they are indirect measures of the patient’s actual condition or state. If there are biases in the data, predictive models using this data may propagate these biases if not addressed, potentially resulting in poorer medical care provided to a group of patients (MIT 6.3950/6.3952, 2022).

Last but not least, since the original dataset contained multiple inpatient visits for certain patients, Strack et al. (2014) suggested using only one encounter per patient when analyzing the dataset in order to make sure that the observations are meaningful and statistically significant. In order to achieve such a goal, they suggested converting the criterion for labeling the outcome variable to whether or not patients were readmitted within 30 days, which turns the outcome variable into a binary one (Strack et al., 2014).

Data

The dataset for this project was obtained from the UCI Machine Learning Repository, and the original data was submitted on behalf of the Center for Clinical and Translational Research of Virginia Commonwealth University. This dataset contains the clinical care data from 1998 to 2008 for 130 US hospitals and integrated delivery networks, and it has a total of 55 attributes and 101,766 instances. These attributes contain both demographic information and clinical information. Some example attributes are race, gender, age, time in hospital, number of lab tests performed, HbA1c test result, diagnosis, number of outpatients, number of inpatients, number of medications, emergency visits in the year before the hospitalization, etc. In addition, the original dataset has a total of 3 possible values for the outcome variable *readmitted*. The first possible value is "No", which stands for no readmission for this patient within the 10 years; the second possible value is ">30", which stands for readmission but more than 30 days after the initial admission date; and the last possible value is "<30", which stands for readmission within 30 days of the initial admission date. Moreover, there are a total of 54864 patients labeled as "No"(53.9% of the entire dataset), 35545 patients labeled as ">30"(34.9% of the entire dataset), and 11357 patients labeled as "<30"(11.2% of the entire dataset). One thing worth noticing is that there's no specification in the dataset regarding when exactly were patients readmitted if they were labeled as ">30". Consequently, the timeframe for patients that are labeled as ">30" could theoretically range from 31 days to 10 years. We also plotted the distribution of some demographic variables using pie charts, and you may find them in the appendix as Figure 9 and Figure 10.

Methods

The original dataset has a total of 3 possible results for the outcome variable *readmitted*. However, we are not clear about when exactly patients were readmitted if they were labeled as ">30", which can potentially affect our analysis. Furthermore, as suggested by Strack et al. (2014), it will be much more significant to look at whether or not patients were readmitted within 30 days. Consequently, we transformed the outcome variable *readmitted* to be either readmitted within 30 days(labeled as "1") or not readmitted within 30 days(labeled as "0"). After such transformation, we obtained a total of 11357 patients that were labeled as "1"(11.2% of the entire dataset) and 90409 patients that were labeled as "0"(88.8% of the entire dataset).

Nevertheless, a new issue emerged after the aforementioned transformation on the outcome variable, which is the problem of imbalanced dataset. A visual representation of our current dataset in terms of the outcome variable is shown below in Figure 1. Given the imbalanced condition of our dataset, the models would most likely have poor performances when predicting for the minority outcome group(i.e. patients labeled as "1"). To fill this distribution gap, we decided to use undersampling and oversampling techniques to make sure the two possible labels for the outcome variable are more balanced when used for training models. These methods were all implemented using the functions within "RandomUnderSampler" and "SMOTE", which belong to the Python package "Imbalanced-learn"(imported as imblearn) for undersampling and oversampling respectively.

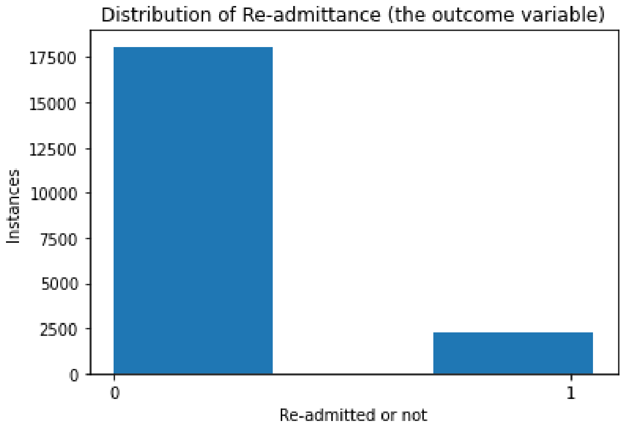


Figure 1 Distribution of the outcome variable *readmitted*

Additional preprocessing was also performed. To be specific, we dropped rows that have more than 30% of the covariates missing and rows with missing data for the outcome variable. Moreover, we dropped columns with more than 45% missing values as well as columns that are irrelevant to our study, such as *patient_nbr*. For missing value imputations, we chose median imputation for continuous features to minimize the effects of outliers, and we chose mode imputation for categorical features. Lastly, we mapped ordinal feature values into fewer categories with larger intervals and performed one-hot encoding for these categorical features.

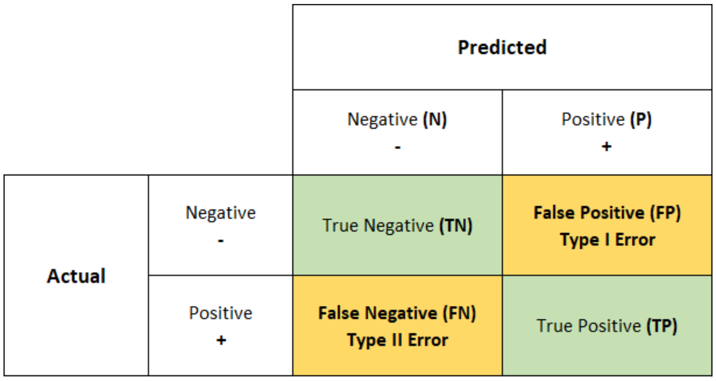
For our modeling part, we implemented a total of three machine learning algorithms, namely multinomial logistic regression, random forest, and XGBoost. Given that our label for this clinical dataset has a binary outcome, we picked logistic regression as our baseline model to gain an overall insight into how this specific dataset responds to machine learning algorithms. To further improve the performance of our logistic regression model, we also applied different penalizations, namely L1 and L2 penalizations, as well as a 5-fold cross-validation to train our logistic regression model. We also tried out different solver functions for logistic regression, such as “liblinear” and “saga”, to see if different solvers would further improve our model. Subsequently, we trained our logistic regression models using the original dataset and tested their performances on the test set. (Note: We did not train our logistic regression models on the oversampled and undersampled training sets because they were merely our baseline model and they also, as we found out and would mention later, performed poorly.)

The second model we deployed was random forest. We tuned the parameter *max_depth* and found out that when *max_depth* = 10, the model was performing well enough. We did not go for a higher value for *max_depth*, despite the slightly higher model accuracy as a result, in order to prevent the potential issue of overfitting. Subsequently, we trained our tuned random forest model separately using the original, the undersampled, and the oversampled dataset; then, we tested these models' performances using our test set.

We implemented our last model using the XGBoost algorithm. We tuned the parameters *learning_rate* and *max_depth* and derived relatively decent model accuracy when *learning_rate* = 0.001 and *max_depth* = 6. Once again, we did not go for a higher value for *max_depth* or a lower value for *learning_rate*, despite the slightly higher model accuracy as a result, in order to prevent the potential issue of overfitting. Subsequently, we trained our tuned XGBoost model separately using the original, the undersampled, and the oversampled dataset; then, we tested these models' performances using our test set.

As for model evaluation, we utilized five evaluation metrics, namely accuracy, precision, recall, F1-score, and AUC score. The models' performances were evaluated by the definitions in Figure 2 below, and our metrics will also be based on that table. To highlight two of the metrics that are more important to this project, precision is the number of true positive predictions divided by all positive predictions(i.e. TP/(TP + FP)), and recall is the number of true positive predictions divided by actual positive cases(i.e. TP/(TP + FN)). Moreover, F1-score is an integrated metric of precision and recall, and it is calculated as

2×(Precision×Recall)/(Precision+Recall). Finally, the AUC score is the area under the receiver operating characteristic curve, which can be obtained using the “roc_auc_score” function.



		Predicted	
		Negative **(N)** **-**	Positive **(P)** **+**
Actual	Negative **-**	True Negative **(TN)**	**False Positive (FP)** **Type I Error**
	Positive **+**	**False Negative (FN)** **Type II Error**	True Positive **(TP)**

Figure 2 The 2×2 table for confusion matrix

We consider accuracy, F1-score, and AUC score as the general metrics to evaluate our models’ performances since these metrics can help us determine the prediction power of each model. On the other hand, precision and recall are considered metrics that can help evaluate bias within a model. Specifically, a low precision is associated with a high false positive rate. Therefore, in our case, a low precision would mean that more patients are readmitted to the hospital than actually needed. On the contrary, a low recall is associated with a high false negative rate, which means that patients who need to be readmitted to the hospital did not get readmittance. Therefore, we would consider a prediction model to be biased toward a certain group of people when that group has a relatively low precision and/or a high recall. With that in mind, we also tested our models separately on different genders and races to see whether each model is biased toward a specific gender or race.

Lastly, we also graphed out the Shapley values and the feature importance plot for the XGBoost model and the random forest model to get an idea of whether or not these two models agree on the most important variables that can help predict readmittance.

Results

The performance of our baseline logistic regression model, despite our efforts, was considered poor as we derived an AUC score of around 0.5(0.5054 for L2-penalized and 0.5032 for L1-penalized). Though the accuracy of this model was high at 0.89, it was solely due to the fact that the model was trying to predict every instance to be a negative label, which also happened to be the dominating label in our datasets, including the test set. As shown below in Figure 3, we graphed out the ROC curves for the logistic regression models with L2-penalization and L1-penalization respectively:

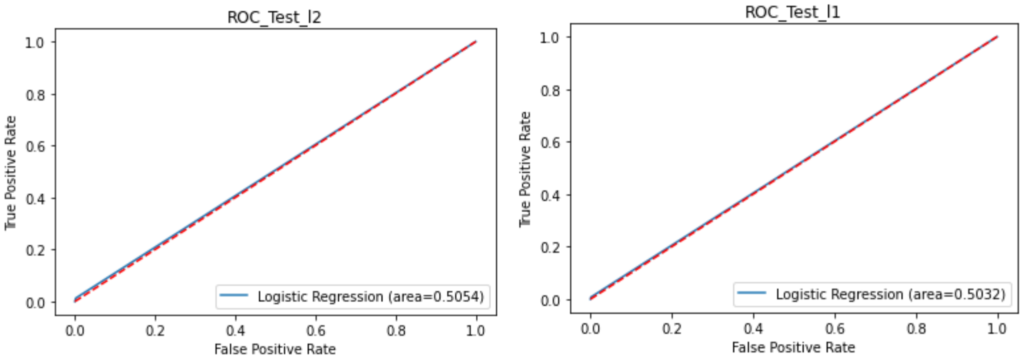


Figure 3 ROC curves for logistic regression models with L2 and L1 penalizations

For the random forest models trained using the original dataset and the undersampled dataset, we achieved AUC scores of 0.651 when evaluated on a test set. However, for the random forest model trained using an oversampled dataset, the AUC score went down to 0.598. The ROC curves are shown below in Figure 4.

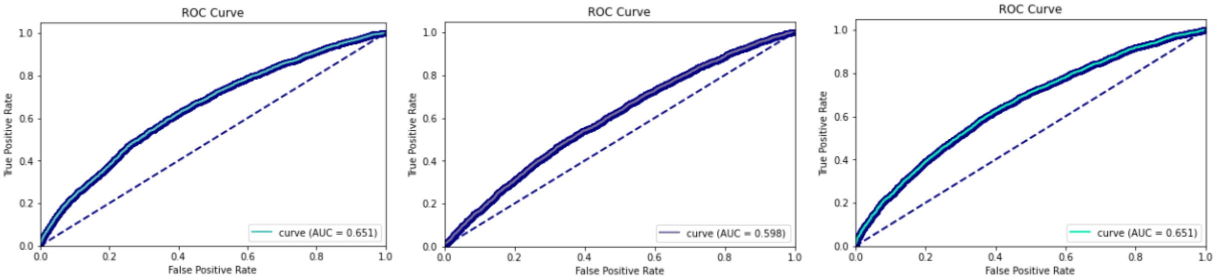


Figure 4 ROC curves for random forest models trained on the original dataset, oversampled dataset, and undersampled dataset respectively from left to right

Lastly, for XGBoost, the highest AUC score, 0.638, was achieved by the model trained on the original dataset, followed by the model trained on the undersampled dataset with an AUC score of 0.636. The model trained on the oversampled dataset had the lowest AUC score of 0.566 among the three XGBoost models.

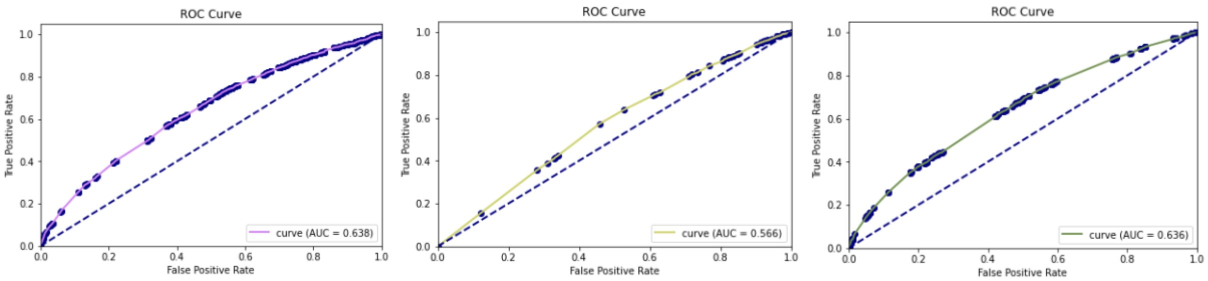


Figure 5 ROC curves for XGBoost models trained on the original dataset, oversampled dataset, and undersampled dataset respectively from left to right

Additionally, as shown below in Figure 6, we plotted the performance metrics of the two models when trained using the original data, oversampling data, and undersampling data using histograms, and we compared the results across different patient characteristics(i.e. different genders and races). Comparing the three plots on the left for Figure 6 to the three on the right (i.e. metrics for evaluating random forest models vs. those for XGBoost), we can see that the random forest models present more observable differences in prediction bias compared to the XGBoost models.

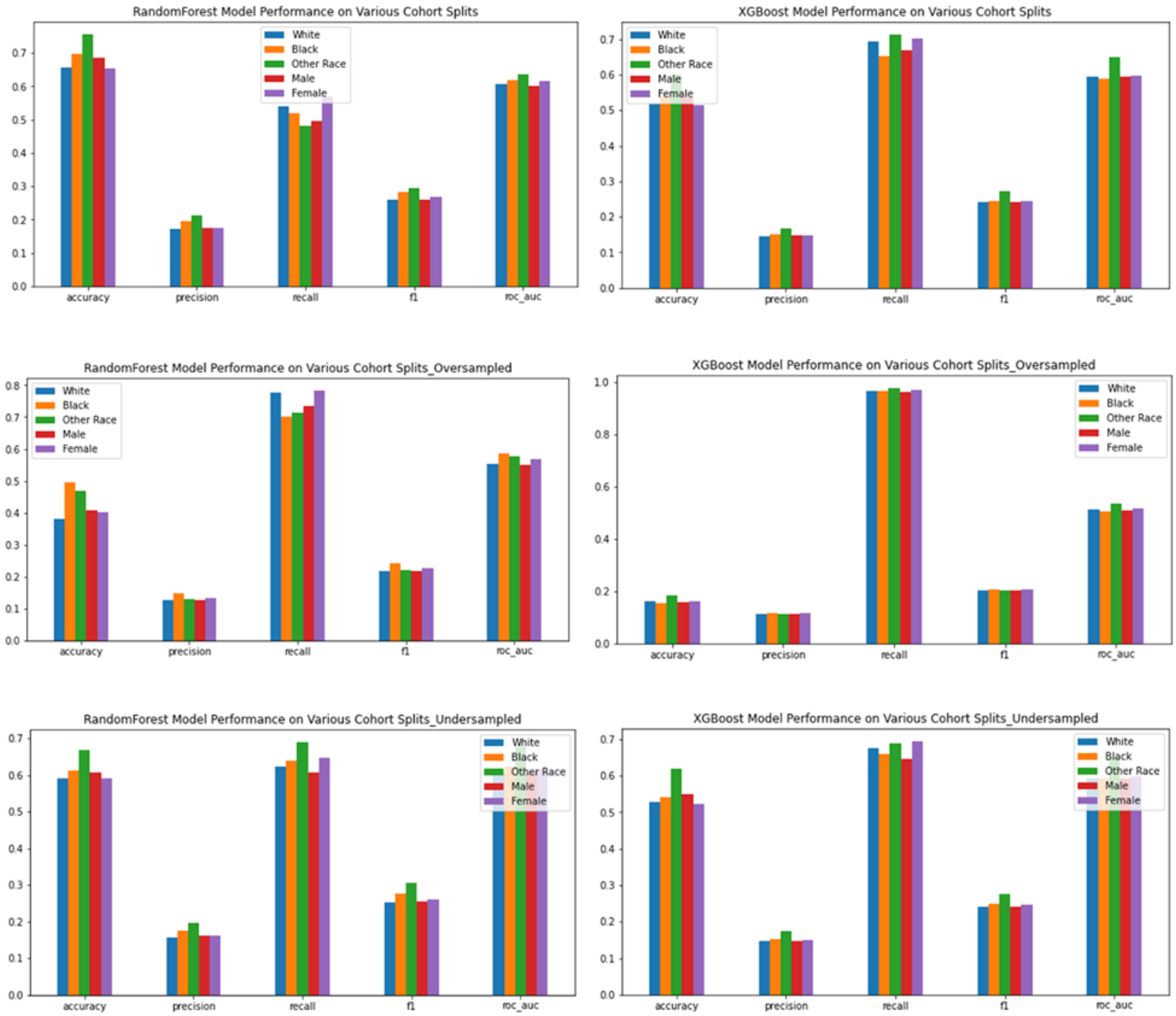


Figure 6 Model evaluation metrics for random forest(left) and XGBoost(right) when trained on the original dataset(top), the oversampled dataset(middle), and the undersampled dataset(bottom)

For the random forest model without oversampling or undersampling, the precision is the lowest for Caucasians, followed by African Americans, and finally other races. The recall is exactly the reverse: Caucasians have the highest recall, followed by African Americans, and finally other races. Based on the definitions of precision and recall mentioned in the methodologies section, we could interpret our findings as there were more Caucasian patients being readmitted to the hospital than actually needed(i.e. there are more false positives for Caucasian patients). When looking at recall, we could see African American patients as well as patients of other races usually have a relatively lower recall than Caucasian patients. This can be interpreted as non-Caucasian patients who should be readmitted to the hospital are not predicted by the model to experience readmittance as often compared to Caucasian patients(i.e. more false negatives for non-Caucasian patients). For this situation, it is more beneficial to err on having higher false positives, rather than higher false negatives, because we would prefer to provide patients with additional resources who may not necessarily need them, rather than missing any patients that would actually need additional medical assistance.

The different combinations of models and preprocessing steps for the training data set propagate varying levels of bias in the model. In general, the XGBoost models have fewer differences in the performance metrics between various patient demographics compared to the random forest models. When comparing the sampling methods for the XGBoost models, the model trained with oversampled dataset presented the most similar metrics across the patient demographics but also had the lowest predictive ability as seen by the lower AUC scores. Both the XGBoost model trained with the original dataset and with the undersampled dataset have superior predictive abilities but metrics evaluating bias differ slightly between different race and gender categories. Comparing these models and sampling methods show the importance of auditing a model for bias as the algorithm choice and modeling design can result in different levels of biases. Consequently, before deploying a model, the tradeoffs between equitable performance and predictive ability should be assessed.

Lastly, because machine learning models like random forest and XGBoost are difficult to interpret, we hope to distinguish the bias using feature importance(Figure 7) for the random forest model and SHAP/Shapley values(Figure 8) for the XGBoost model to understand our models' decision making processes. For random forest, we concluded from the metrics that the model presents predictive bias among races but no bias regarding genders. The feature importance for Caucasians is higher than African Americans, which is in turn higher than other races. On the contrary, though the feature importance for males is higher than that for females, the difference between the two is relatively small. Nevertheless, overall, the differences in feature importance values are small, which makes it hard for us to validate the significance of the bias based on feature importance only. For XGBoost, our conclusion was no bias existence. In alignment with our conclusion, race or gender is not one of the top ten most important features according to Shapley values.

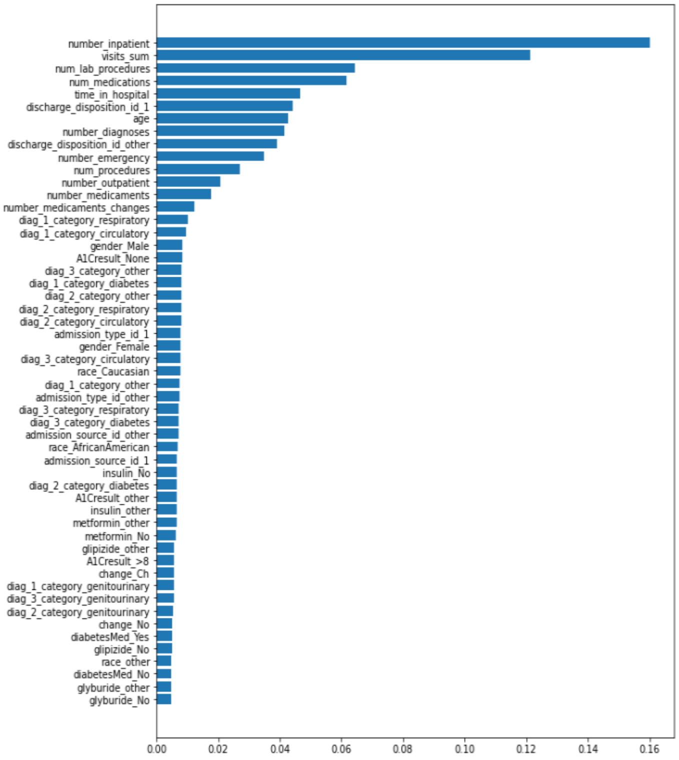


Figure 7 Feature importance for the random forest model trained on the original dataset

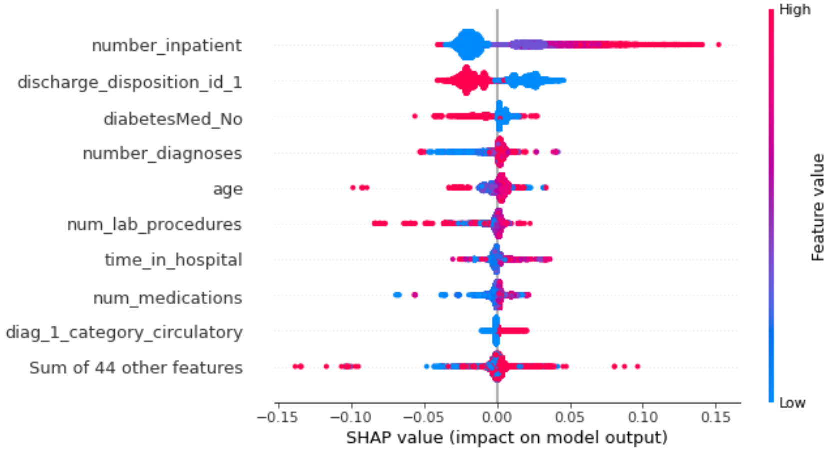


Figure 8 Shapley values for XGboost model trained on the original dataset

Conclusion and Discussion

Through our exploration of model performance metrics and analysis of various machine learning algorithms, we found that model designs for building a predictive program on a diabetes hospital readmissions dataset can introduce a certain amount of bias. Overall, the models presented less-than-desirable predictive performance, and different degrees of bias were observed depending on the model design. Comparing the XGBoost and random forest models, the XGBoost models tend to have more equitable predictions while still maintaining comparable predictive ability. Preprocessing methods, such as oversampling and undersampling, were used to correct data imbalances and deficiencies, and these methods affected not only the predictive capabilities of these models but also the level of bias that propagated in the predictions. Overall, we assessed the bias of decision-making within these models, which has not been done in previous studies that built machine learning models for this dataset.

If a biased model, such as the random forest models in this paper, was deployed, there would be significant implications on the equity of care. For example, a hospital may want to identify whether or not a patient is likely to experience readmission to determine which patients to monitor more closely. If a healthcare system uses the random forest model to identify patients that are most at risk for readmission for diabetes, the model would over-select Caucasians for readmittance, resulting in unfairness as Caucasian patients are more likely to receive medical care and attention than African American patients and other minority patients.

Lastly, there remain a few limitations to our study as well. To be specific, we did not provide statistical proof of the existence of bias since it was hard to validate the statistical significance of the differences between model performance metrics for different populations given the models' generally poor performances on this imbalanced dataset. Hence, we resorted to Shapley values and feature importance tables to buttress our observations and findings. In addition, we removed quite a few instances and features due to missing values, which also meant that we lost a certain amount of prediction power for our models since the values for those features and instances were not completely empty. Future research could focus on improving the model performances by either using superior sampling and imputation techniques or an overall better model. Future papers could also attempt to provide statistical evidence to mathematically prove the existence of bias within the models for populations with certain characteristics.

References

Alturki, L., Aloraini, K., Aldughayshim, A., & Albahli, S. (2019, November). Predictors of readmissions and length of stay for diabetes related patients. In *2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)* (pp. 1-8). IEEE.

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

Centers for Disease Control and Prevention. National diabetes statistics report, 2017. Centers for Disease Control and Prevention website. www.cdc.gov/diabetes/pdfs/data/statistics/national-diabetes-statistics-report.pdf. Updated July, 18 2017. Accessed December 11, 2022.

MIT 6.3950/6.3952 (September 2022) Lecture_3_Healthcare, Slide 24 and Slide 28

Rubin DJ. Hospital readmission of patients with diabetes. Curr Diab Rep. 2015 Apr;15(4):17. doi: 10.1007/s11892-015-0584-7. *Corrected and republished in: Curr Diab* Rep. 2018 Mar 13;18(4):21. PMID: 25712258.

Shang, Y., Jiang, K., Wang, L., Zhang, Z., Zhou, S., Liu, Y., Dong, J., & Wu, H. (2021). The 30-days hospital readmission risk in diabetic patients: predictive modeling with machine learning classifiers. *BMC medical informatics and decision making*, *21*(Suppl 2), 57. https://doi.org/10.1186/s12911-021-01423-y

Suresh, A. (2021, June 22). *What is a confusion matrix?* Medium. Retrieved December 12, 2022, from https://medium.com/analytics-vidhya/what-is-a-confusion-matrix-d1c0f8feda5

Appendix

Here's the Google Colab link to all of my code for this paper.

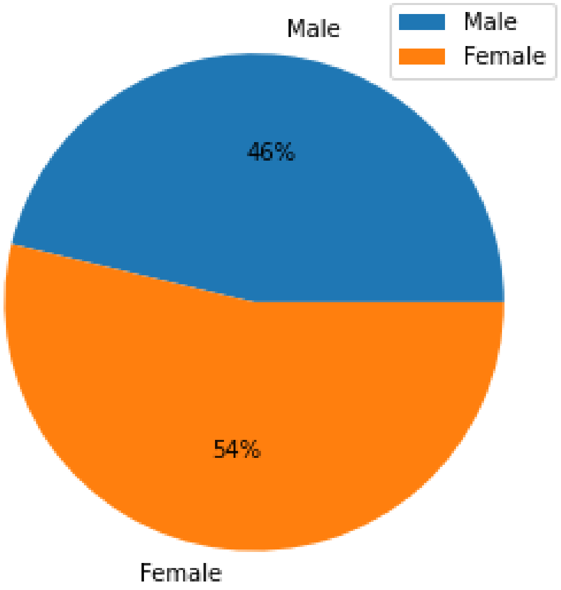


Figure 9 Gender distribution

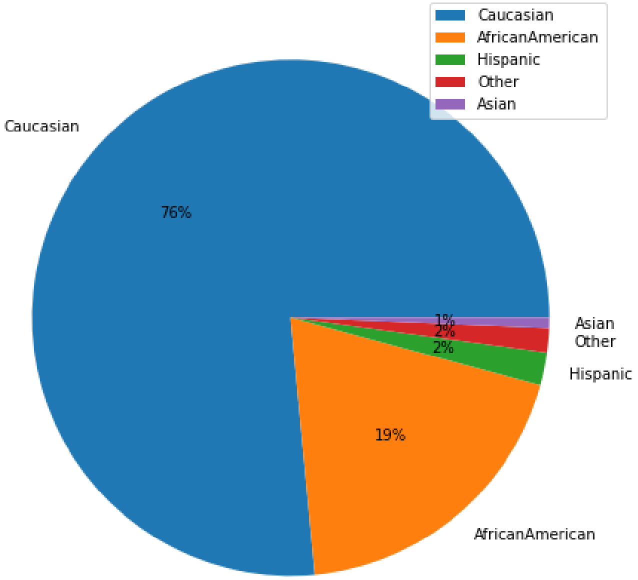


Figure 10 Race distribution