Predicting Hospital Re-admission in Diabetic Patients

Crystal Chen[1] and Sara Golestaneh[2]

***Abstract*— Hospital re-admissions can be a drain on hospital resources, both in terms of time and money. As such, many in the healthcare industry are interested in reducing readmission rate to reap the numerous benefits. This project studied the factors which had the most significant impact on predicting hospital re-admission in patients with diabetes using a variety of binary classification methods.**

I. INTRODUCTION

Hospital re-admissions can be caused by many factors from inadequate coordination by administrators[1] to insufficient post-discharge care[2]. Regardless of the reason, re-admission can have a negative impact on patients as their health issues may have progressed to a more serious stage compared to their previous visit, putting their life at greater risk. Furthermore, re-admissions increase costs as hospitals must allocate more resources to address cases which has been previously addressed. Lastly, re-admissions place further burdens upon healthcare workers as these re-admissions contribute to overcrowding which many hospitals already face. As such, it is not surprising that there is great interest in identifying factors which contribute to re-admission.

In this project, the objective was to use binary classification to identify significant factors that predict re-admission in patients with diabetes. The data used was from the data set *Diabetes 130 US Hospitals for years 1999 - 2008*. As the title implies, this publicly available dataset involved 130 American hospitals and recorded information regarding diabetic patients and their stay, ranging from their age and race, to their length of stay, the diagnoses and the kinds of medication they were prescribed. Note that as per similar research, only re-admissions under 30 days were considered[3]. Re-admissions over 30 days were not considered to be re-admissions.

II. METHODS

A. Data Cleaning

This dataset contained in total fifty variables in a variety of formats. First, ten highly problematic variables were removed.

- **Weight:** in the original dataset, approximately 97% of patients were not weighed. As such, this variable was dropped as it was unlikely to be informative.
- **Certain medication variables:** in general, there were 23 medication variables which indicated whether the patient was described the medication and/or whether their dosage had changed. However, nine of these features were removed as they fell into one of two categories - either the medication was never prescribed to any patients or the medication was prescribed to fewer than 100 patients. In either case, these features were so biased towards one level that they were not expected to have any explanatory power. The specific medications that were removed were acetohexamide, tolbutamide, troglitazone, examide, citoglipton, glipizide metformin, glimepiride pioglitazone, metformin rosiglitazone, and metformin pioglitazone.

Note that there were other variables which were missing 30-50% of their values (e.g. A1C results and Payer Code). However, these were deemed likely to still have some explanatory power and so, did not warrant removal. Furthermore, A1C results had been shown to be significant for predicting re-admission. As such, only features that were missing over 95% of its data were removed.

Secondly, any patients who had a disposition ID of 11, 13, 14, 19-21 were removed. These patients either passed away or were transferred to a hospice and as such, were unlikely to ever be readmitted[4][5]. Given the target population are patients who have a possibility of being readmitted, these patients were removed as they did not fit that criterion.

Lastly, the remaining variables were processed depending on their type. The remaining numerical variables required no processing because there appeared to be no instances of missing or erroneous values. One notable change is Age which was originally an ordinal variable with patient's age being indicated by a range (e.g. 0-9, 10-19, etc.). This was transformed into a numerical variable by taking the midpoint of each age range as an estimate of patients' age. Converting Age into a numerical variable makes more sense as it is just more common to treat age as a numeric value.

For the remaining categorical variables, any missing/unknown values were grouped into an "Unknown"/"Other" category. Then, they all received the one-hot encoding where each level in a feature becomes its own indicator function. Note however that features such as Diagnosis 1-3 and Medical Specialty contains many levels (e.g. Diagnosis 1 contains 718 levels). As such, if all levels became an indicator function, then overfitting may be an issue in modelling. These particular variables were reduced to only eleven levels where only the top ten most frequent labels were retained and all other labels were relabelled as "Other".

[1]Crystal is a MSc student in Statistics, University of Toronto, Email: chy.chen@mail.utoronto.ca

[2]Sara is a MSc student in Applied Computing, University of Toronto, Email: sara.golestaneh@mail.utoronto.ca

B. Exploratory Data Analysis

The motivation behind the exploratory data analysis (EDA) was to investigate features which were intuitively linked to health; if a patient was experiencing more severe health issues, they may be more likely to be re-admitted. There was also interest in identifying whether certain groups of people were more likely to be readmitted. This analysis was completed through a series of visualizations (i.e. pie charts, stacked bar graphs and box plots) often relating features of interest to rate of re-admission.

C. Balancing out the Data

One of the major problems with this data set that needed to be resolved before modeling, was that the data was highly imbalanced. More specifically, the number of the patients who were readmitted within 30 days was much lower than the number of the patients who were not readmitted (see Fig. 1).

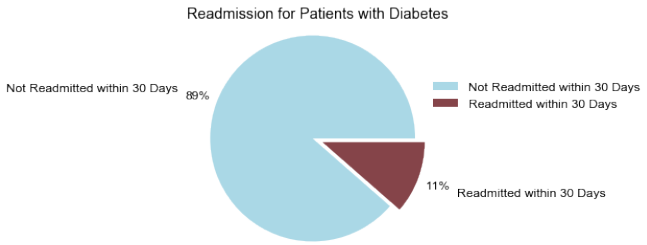

Fig. 1: Proportion of readmitted patients vs. non-readmitted patients

As seen in Fig. 1, readmitted patients only cover up around 11% of all patients. Since the main goal of this project was to correctly classify the patients that were readmitted, not having enough data about this group could pose a problem and could cause models to be highly biased. To address this issue, there are two main solutions:

1) Undersampling the Majority Class: The main way to to do this is by randomly choosing a subset of data in the majority class (also called "Random Undersampling"). This results in a balanced but smaller version of the original data set.

2) Oversampling the Minority Class: There are two main ways to implement oversampling. One way is to simply choose a random subset of the minority class and then duplicate it to populate the dataset and create classes of equal size. The other method is to synthesise a new dataset through SMOTE (Synthetic Minority Oversampling Technique). Random oversampling and undersampling were initially used in this project. However since random oversampling behaved similarly to random undersampling, focus was shifted to comparing random undersampling and SMOTE oversampling method.

D. Modelling

This project focused on applying and comparing three different tree-based models:

1) Random Forest: A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

2) Gradient Boosting: Gradient Boosting builds an additive model in a forward stage-wise fashion and allows for the optimization of arbitrary differentiable loss functions. In each stage, regression trees are fitted on the negative gradient of the binomial or multinomial deviance loss function. Since this is a binary classification, only a single regression tree is induced in each stage.

3) XGBoost: XGBoost (Extreme Gradient Boosting) is similar to normal gradient boosting with a few optimizations. While regular gradient boosting uses the loss function of the base model (e.g. decision tree) as a proxy for minimizing the error of the overall model, XGBoost uses the 2nd order derivative as an approximation. Also, xgboost uses a more regularized model formalization to control over-fitting, which gives it better performance.

Please note that since there were no timestamps associated with the data points in the dataset, they could not be treated as time series data. On the other hand, since multiple rows could belong to one patient visiting the hospital on different days, the data points were not completely independent. For such data, tree-based models are the best, since they do not make any strict assumption about the independence of the data (unlike logistic regression).

E. Cross-Validation

To evaluate the models and compare the results from using oversampling and undersampling, a 5-fold stratified cross validation was used. When doing cross validation along with balancing out techniques, undersampling and oversampling were only applied on the training partitions and the validation partition would remain untouched.

F. Hyper-parameter Tuning

After choosing the best model based on different metrics (especially recall and AUC), 2-fold cross-validation was applied along with a grid search on possible parameters to do the hyper-parameter tuning. The grid search was completed over the following parameters: learning rate: [0.001,0.01,0.1], maximum depth: [1,3,5,7,9] and number of estimators (decision trees): [100,200,...,500]. The metric that was optimized at this stage was AUC.

G. Feature Importance and SHAP values

To interpret the final model, SHAP values were used. SHAP or SHapley Additive exPlanations is a game theoretic approach that is used on tree-based models to compute the contribution of each feature to the predictions. It connects optimal credit allocation with local explanations using the classic Shapley values from game theory and their related extensions [6].

III. Results

A. Exploratory Data Analysis

There are a few important observations to note from the EDA. The main observation is that the dataset is unbalanced in a variety of manners. Recall, there are far fewer readmitted patients (approx. 11.4%) compared to non-readmitted ones (see Fig. 1). This is to be expected as re-admissions are not so high that the proportion of readmitted patients is comparable to non-readmitted patients. This dataset is also unbalanced in terms of race and gender. Approximately 75% of patients are Caucasian with the second largest racial group being African Americans (19%) (see Fig. 8 in Appendix). However, proportion of readmitted patients within each group is relatively constant at around 10-11% (see Fig. 9 in Appendix). Similarly, the population is made up of 53% females and 37% males, with re-admission rates among the two groups being almost the same (11.5% vs. 11.3% respectively) (see Fig. 10 in Appendix).

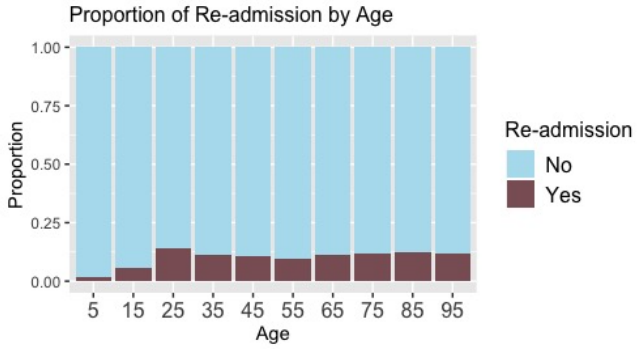

Fig. 2: Proportion of readmitted patients vs. non-readmitted patients for each age group

A majority of patients were also in the 50-89 age range (see Fig. 11 in Appendix), which is unsurprising given that older people are more likely to experience health problems and so, visit the hospital more. Unexpectedly however, re-admission rates peak in the 20-29 age group and remain relatively stable for patients aged 30-99 (see Fig. 2). One would expect patients who are more likely to experience more severe health issues at a higher frequency (i.e. older patients) to be readmitted more than younger, healthier patients. Keeping in mind that this dataset involves diabetic patients, one potential explanation is that patients aged 20-29 are likely beginning to live independently and are at an age where they are still somewhat inexperienced with managing their illness. Thus, they may be readmitted more often compared to older patients who have more experience and even younger patients who are cared for by their parents.

It also appears that particular diagnoses are prone to higher rates of re-admission compared to others. The proportion of readmitted patients under and over 30 days vs. non-readmitted patients for the first, second and third diagnoses can be observed in Fig. 3, 4 and 5 respectively.

In the first diagnosis, the top three outcomes with the highest re-admission rates are, in decreasing order, heart failure (ID: 428), chronic bronchitis (ID: 491) and acute myocardial infarction (ID: 410). In the second diagnosis, the top three outcomes are hypertension renal disease (ID: 403), disorders of fluid, electrolytes and acid-balance (ID: 276) and once again, heart failure. In the third diagnosis, hypertension renal disease and fluid, electrolytes and acid-balance disorders were once again among the top three outcomes with the highest re-admission along with chronic renal failure (ID: 585). Based on these observations, heart failure, hypertension renal disease and disorders of fluid, electrolytes and acid-balance are conditions that doctors should possibly pay closer attention to as they are not only some of the most frequent diagnoses and have the highest re-admission rates, but also appear multiple times in different diagnoses. As an aside, the "Other" outcome in all three diagnoses have high re-admission rates, but recall that any outcome that was not one of the ten most frequent diagnoses were collapsed into this category. As such, the category contains too wide range of diagnoses to make any concrete observations about it.

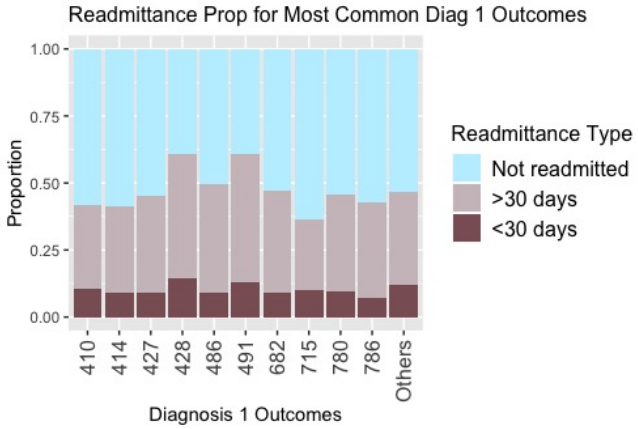

Fig. 3: Proportion of readmitted patients over and under 30 days vs. non-readmitted patients for top 10 most common outcomes of first diagnosis

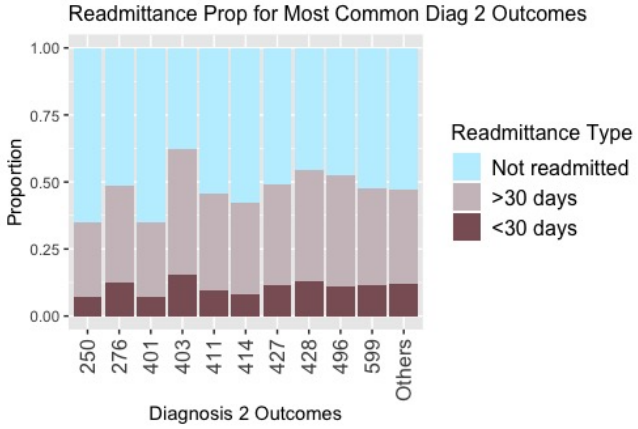

Fig. 4: Proportion of readmitted patients over and under 30 days vs. non-readmitted patients for top 10 most common outcomes of second diagnosis

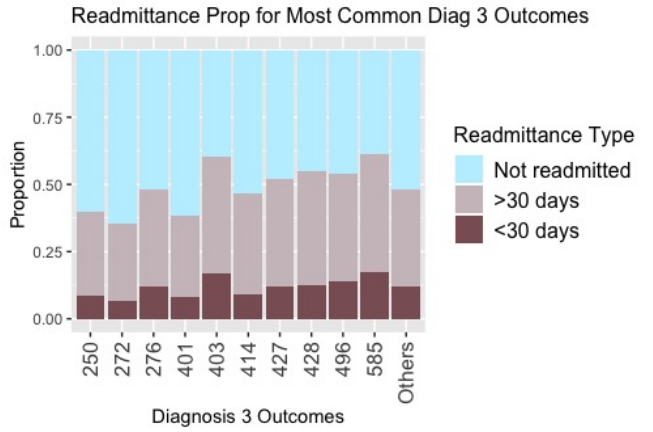

Fig. 5: Proportion of readmitted patients over and under 30 days vs. non-readmitted patients for top 10 most common outcomes of third diagnosis

Lastly, the distributions of the number of medications that patients regularly take as well as the total number of procedures they undergo (whether be it lab tests or other procedures) did not seem to differ between patients who were readmitted and patients who were not readmitted. This indicates that these features likely will not play a significant role in predicting whether or not a patient is readmitted.

To reproduce the results of this section, see the EDA Rmarkdown file.

B. Model Results Using Random Undersampling and SMOTE

After evaluating the models using 5-fold cross-validation, it was observed that SMOTE resulted in higher accuracy in general. The difference in the accuracies can be seen in Table I evaluated using cross-validation (Also see Fig. 12 and Fig. 13 in Appendix).

Model Name	Random Undersample Mean accuracy	SMOTE Mean accuracy
Random Forest	0.615	0.735
Gradient Boosting	0.651	0.884
XGBoost	0.652	0.884

TABLE I: Random Undersampling vs SMOTE Results Using Cross-validation

Although it seems like SMOTE results in higher accuracy, other metrics such as the recall of the positive class is more important for the purposes of this project. Table II shows the difference in accuracy, recall for the positive class and AUC metrics between the two methods when models were evaluated on the test set. Unfortunately, SMOTE resulted in much lower recall and AUC compared to random undersampling. In other words, when SMOTE was used to balance the dataset, models became more conservative and labeled as few positive cases as possible, resulting in lower recall but higher accuracy. As mentioned, since both higher AUC and higher accuracy along with good recalls for positive class are more important for this project, XGBoost with random undersampling was chosen as producing the best results (since the AUC is a little bit higher compared to other models).

Model Name	Random Undersample Recall	Acc	AUC	SMOTE Recall	Acc	AUC
Random Forest	0.63	0.61	0.615	0.34	0.74	0.565
Gradient Boosting	0.59	0.65	0.620	0.03	0.89	0.510
XGBoost	0.59	0.65	0.623	0.03	0.89	0.509

TABLE II: Random Undersampling vs SMOTE: AUC and Recall of Readmitted patients evaluated on Test set

The best parameters resulted from the grid search were: learning rate: 0.1, maximum depth:3 and number of estimators: 100, which is by coincidence the same default parameters that we used to train our model.

Fig. 6 shows the confusion matrix of the best model. As we can see the main reason behind the low accuracy is the high number of false positives.

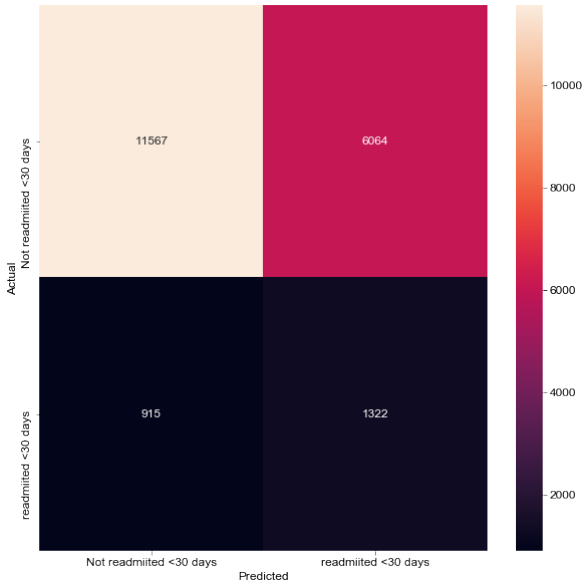

Fig. 6: XGBoost Confusion Matrix on the Test set

To reproduce the results of this section, run the Undersample notebook. For SMOTE results, run the Oversample notebook.

C. SHAP Feature Importance Results

Fig. 7 shows the top 20 features based on the shap values.

As shown in the graph above, some of the features such as the number of inpatient visits in the year preceding the encounter or being discharged to another rehab facility have a positive affect on being readmitted within 30 days. In other words if the patient has had a higher number of inpatient visits or if they were transferred to another rehab facility, there is higher probability that they will be readmitted within 30 days. Features such as being discharged to home,

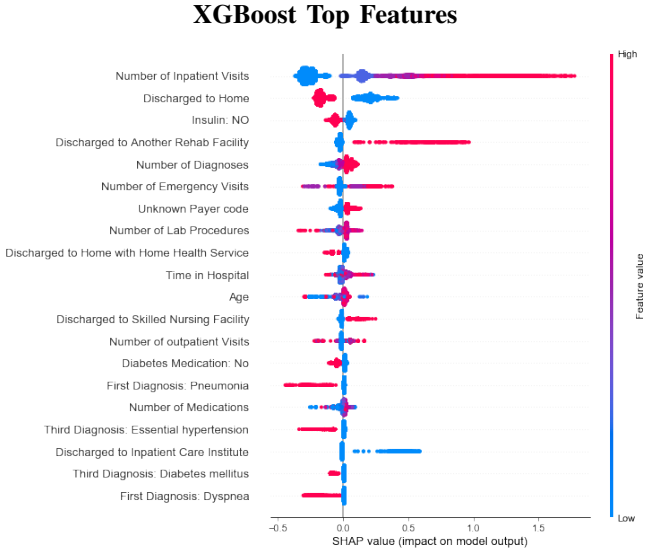

Fig. 7: Top 20 Features of the XGBoost Model Based on SHAP values. Every patient has one dot on each row. The x position of the dot is the impact of that feature on the model's prediction for the patient, and the color of the dot represents the value of that feature for the patient. Dots that don't fit on the row pile up to show density. Since the XGBoost model has a logistic loss the x-axis has units of log-odds (Tree SHAP explains the change in the margin output of the model)

no insulin and no diabetes medication have a negative effect on the prediction, meaning if the patient is discharged to home or if they were not prescribed any insulin (or any diabetes medication), there is higher chance that they will not get readmitted within 30 days. Furthermore, the higher the number of diagnoses, the higher the chance of being readmitted within 30 days (based on the model). Older patients and patients that spent a longer time in the hospital are more likely to be readmitted within 30 days. Two final interesting observations are that if the patient is diagnosed with pneumonia or dyspnea (a breathing problem) as the first diagnosis or essential hypertension or diabetes mellitus as the third diagnosis, there will be a lower probability of getting readmitted within 30 days.

To reproduce the results of this section, see the Undersample notebook, SHAP Values section.

IV. Discussion Conclusion

Overall, this project revealed both surprising and unsurprising results. Through EDA, it was observed that features such as gender and race likely do not play a major role in predicting re-admission given the similar rates between groups. Similar conclusions were made regarding number of medications that patients take as well as the total number of procedures they undergo during their visit. It was surprising to see that while the majority of patients were over 50, 20-29 year olds experienced the highest re-admission rate. This could possibly be due to the new-found independence that patients in this age group experience and the lack of experience they have with managing diabetes. Lastly, it was noted that diagnoses such as hypertension renal disease, heart failure and disorders of fluid, electrolytes and acid balance were common conditions that had high re-admission rates.

However, the results of modelling notes that the most significant factors which predict re-admission are a combination of both the patient's patterns of behaviour (e.g. number of emergency visits and number of inpatient visits) and particular outcomes of their visit (e.g. discharged to home/another rehab/nursing facility) and the condition of their wellness which is somewhat indicated by whether they were prescribed insulin or any diabetes medication or not. Also, the models attributes higher rate of re-admission to older ages which is not necessary true.

V. Future Considerations

There are a few final considerations to be noted for this project. Firstly, as with all data analysis, preprocessing is key. In this project, the number of categories in diagnosis and medical specialty were reduced by keeping only the top ten most frequent diagnoses/specialties and re-categorizing all others as "Other". While this is a reasonable way to reduce the number of features, this method is limited to only interpretations regarding specific diseases. Alternatively, medical classes could have been used to reduce the number of categories. For example, the diagnoses could be classified into their ICD 9 classes (e.g. diseases of circulatory system, nervous system, respiratory system, etc.). This would have allowed for more general interpretations of the results. Secondly, only one type of oversampling and undersampling was investigated. Many other oversampling methods (e.g. SVMSMOTE) or undersampling methods (e.g. Condensed Nearest Neighbours) or even a combination of those could potentially result in higher AUC or recalls. Lastly, calibrating the models could potentially produce better results.

References

[1] MEDPAC, "Report to congress: Promoting greater efficiency in medicare," 2007.

[2] A. Hernandez, M. Greiner, G. Fonarow, B. Hammill, P. Heidenreich, and C. Yancy, "Relationship between early physician follow-up and 30-day readmission among medicare beneficiaries hospitalized for heart failure.," *Journal of the American Medical Association*, vol. 303(17), pp. 1716–1722, 2010.

[3] H. Felix, "Why do patients keep coming back? results of a readmitted patient survey," *Social Work in Health Care*, vol. 54(1), pp. 1–15, 2015.

[4] ""Diabetes 130-US hospitals for years 1999-2008 - dataset by uci," data.world, 28-feb-2018. [Online]. Available: https://data.world/uci/diabetes-130-us-hospitals-for-years-1999-2008. [Accessed: 24-feb-2020].."

[5] A. Long, "Using machine learning to predict hospital readmission for patients with diabetes with scikit-learn." Medium, 30-Jan-2020. [Online]. Available: https://towardsdatascience.com/predicting-hospital-readmission-for-patients-with\-diabetes-using-scikit-learn-a2e359b15f0. [Accessed: 24-Feb-2020].

[6] "Slundberg, "slundberg/shap," GitHub, 21-feb-2020. [online]. Available: https://github.com/slundberg/shap#citations. [Accessed: 24-feb-2020]."

VI. Appendix

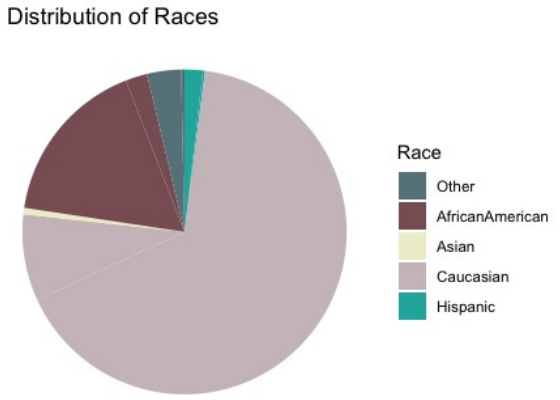

Fig. 8: Distribution of races

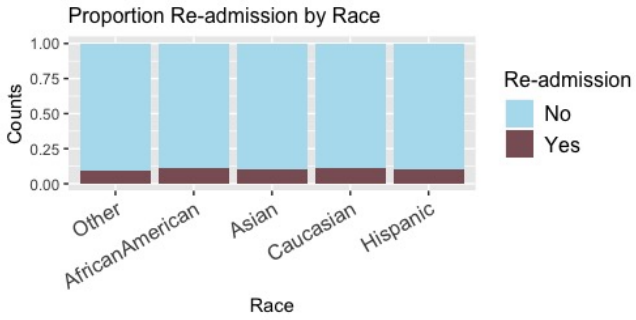

Fig. 9: Proportion of readmitted patients in each race

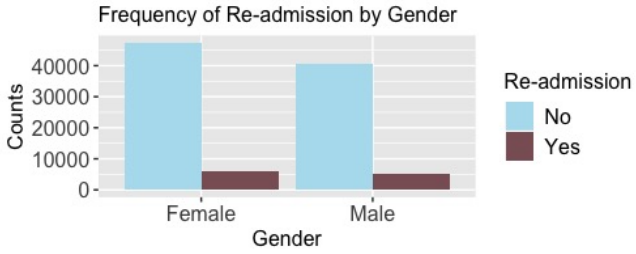

Fig. 10: Frequency of readmitted patients vs. non-readmitted patients for various genders

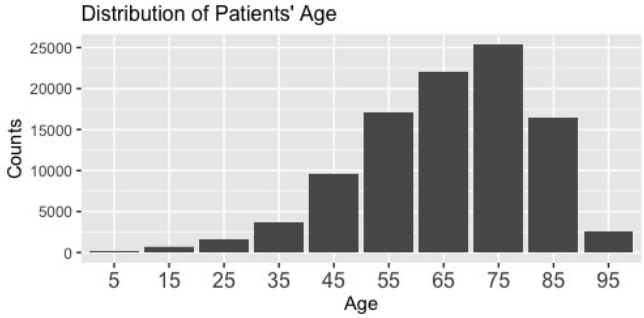

Fig. 11: Number of patients in each age group

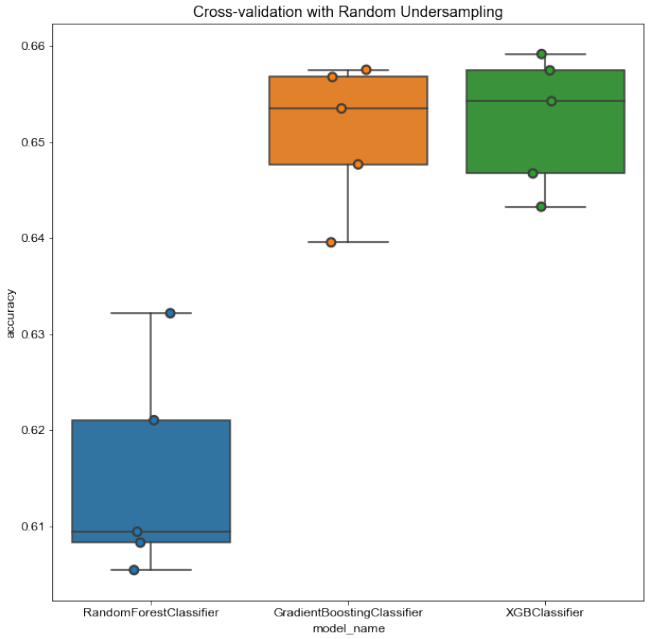

Fig. 12: Boxplot of accuracies of three models using Random undersampling and 5-fold cross-validation

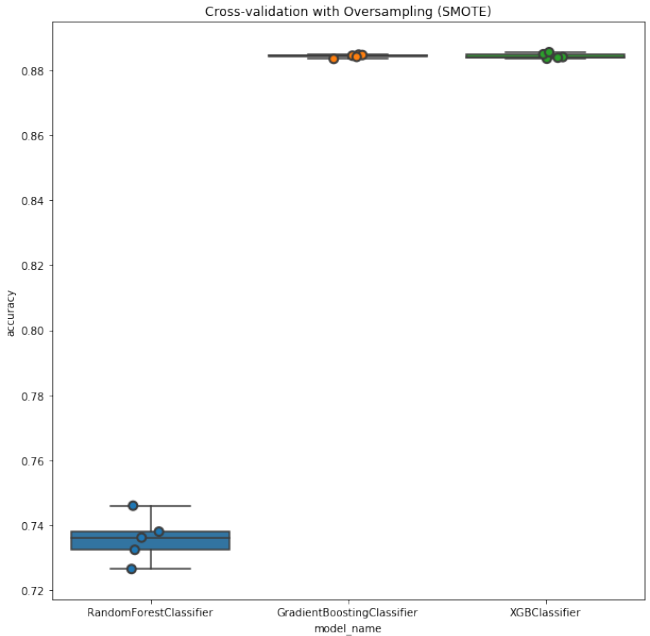

Fig. 13: Boxplot of accuracies of three models using SMOTE (oversampling) and 5-fold cross-validation