Predicting Hospital Readmissions of Diabetic Patients

Final project report
DSI Huda Saeed
GitHub

1. Introduction

Motivation

Diabetes is a common disease that increases the risk of potentially life-threatening heart problems. Hence, to optimize diabetes care, it is useful to have a model that predicts whether a diabetic patient will be readmitted to the hospital in less than 30 days (which indicates the patient has health issues) as the physician would be able to reevaluate the patient's treatment. Furthermore, it would be in the hospital's interest to minimize their visits and thus costs, which such a model would facilitate in its attempt to preemptively identify the patients at risk of readmission and alert physicians to take action so they need not return.

Cerner Corporation Dataset

The dataset comes from Health Facts, a data warehouse that gathers data from hospitals across the USA, by Cerner Corporation, a health IT company [1]. Healthcare facilities that use the Cerner Electronic Health Record System can volunteer to share their data which I then found on the UCI Machine Learning Repository [1]. This dataset represents ten years (1999-2008) of diabetic care at 130 American hospitals [2]. Each row represents a hospital visit of a patient who was diagnosed with diabetes, who stayed in the hospital for 1 to 14 days, and for whom laboratory tests were performed and medications were administered [2].

Previous Work

A health data scientist, Andrew Long, developed a predictive model with this dataset with the same aim [3]. His best model was a gradient-boosting classifier which caught 58% of the readmissions and is about 1.5 times better than a model that outputs random predictions [3]. Additionally, a study published in *BMC Medical Informatics and Decision Making* found the Random Forest Classifier to be the best performer with an AUC of 0.64 [4].

2. Exploratory Data Analysis

Features, Target Variable, and Data Cleaning

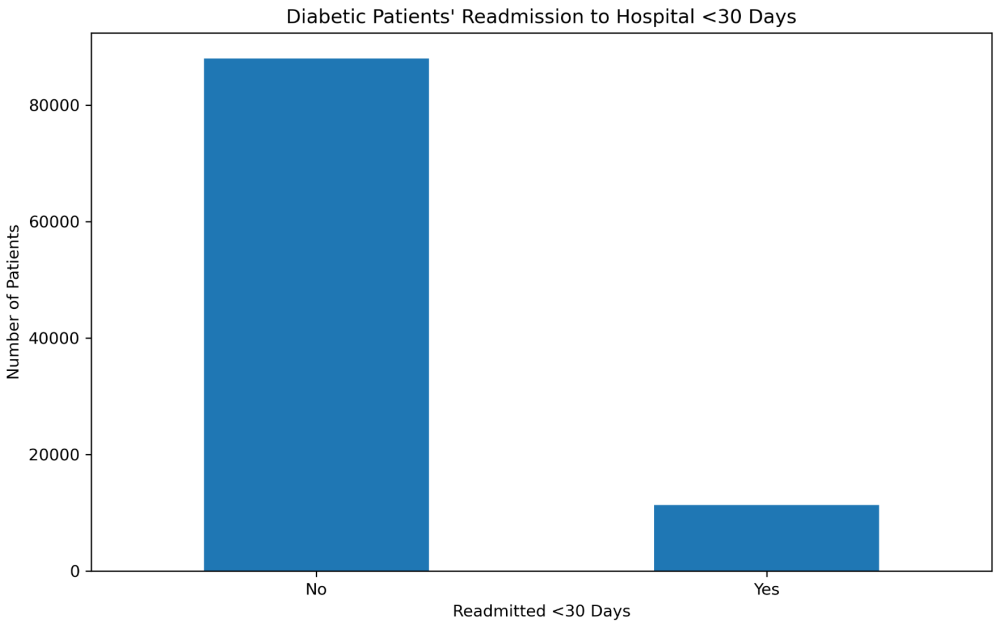
The dataset has 101,766 rows and 49 features which include demographic information, medical administrative information (for example, the type of specialist who treated the patient), and information about various medications. The binary target variable was created based on the readmission variable, where 1 is assigned to those readmitted <30

days and 0 to everyone else. Since each row represents a hospital visit, some patients repeat, so this dataset has a group structure. I removed patients who passed away or were discharged to hospices (which take terminally ill patients) as they cannot be readmitted. Ultimately, the cleaned dataset had 99,340 rows.

Data Exploration

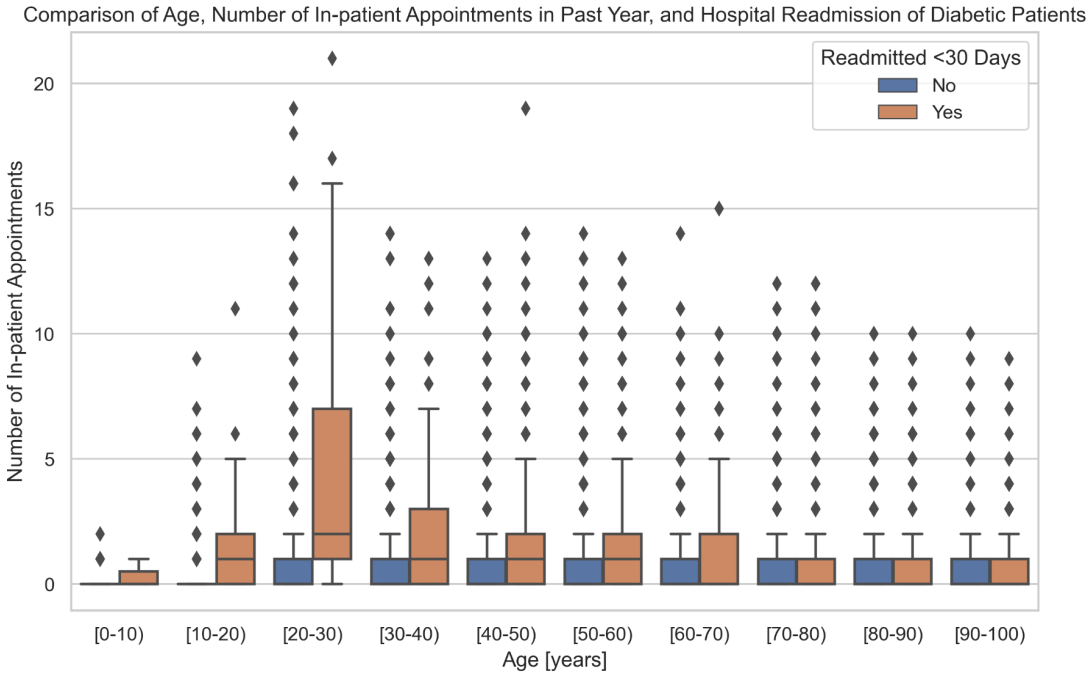
Through .value_counts and .describe I investigated every feature and replaced categories with 'NA' or replaced 'NA' with 'No test' based on the dataset's codebook.

I found that the target was imbalanced, with 11.39% of the observations readmitted in less than 30 days.

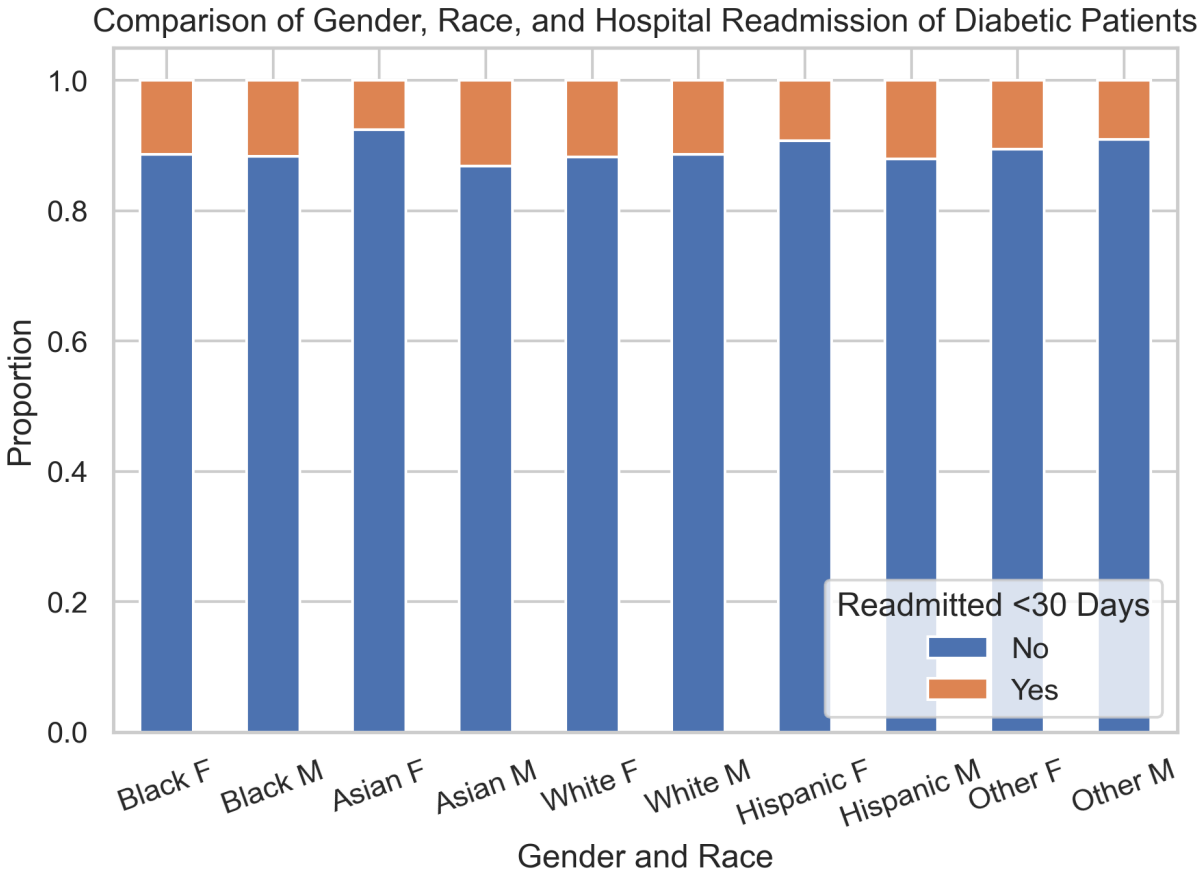


<Figure 1. The readmission target is imbalanced >

I also found that 20-30-year-olds who were readmitted had a higher number of past in-patient appointments than 20-30-year-olds who were not readmitted; this indicated a high usage of hospital resources so perhaps additional preventative measures are needed for this population.



<Figure 2. The readmitted and non-readmitted subgroups for each age group except 20-30-year-olds have a comparable number of in-patient appointments in the past year>



<Figure 3. The readmitted proportion of each race is similar across men and women except for Asians>

Lastly, I found that the fraction of diabetic Asian men who were readmitted is double the fraction of diabetic Asian women who were readmitted, which demonstrates the importance of these demographics for this problem.

Missingness

None of the quantitative features had missing values. 98.97% of the observations had missing values, mostly due to weight. The features with missing values are displayed in Table 1 with their percent of missing values; these were all resolved with a “missing” category.

Feature	Percentage of Missing Values (%)
race	2.25
admission_type_id	9.89
admission_source_id	6.74
discharge_disposition_id	3.72
weight	96.85
payer_code	39.66
medical_specialty	48.95
diag_1	0.02
diag_2	0.36
diag_3	1.43

<Table 1. Percentage of missing values for each feature in the data>

3. Methods

Dataset Split

Due to the group structure and the imbalanced target, StratifiedGroupKFold was used. Since the entire dataset was too computationally expensive to train on the algorithms, a sample was retrieved through the first fold of StratifiedGroupKFold out of 7 folds, which had 14,022 observations. A larger sample size was not chosen due to computational expense. The test set was then created through the first fold of StratifiedGroupKFold with 6 folds on the sample. The remaining 5 folds were split into 4 folds through StratifiedGroupKFold into training and validation sets to achieve an approximate 6:2:2 ratio between the training, validation, and test sets. For the XGBoostClassifier, the validation set was used as the evaluation set with early stopping rounds set to 10.

Preprocessing

OneHotEncoder() was used to preprocess categorical features, OrdinalEncoder() was used to preprocess ordinal categorical features, and StandardScaler() was used to preprocess quantitative features. One of the medications, troglitazone, was dropped as a feature since it was withdrawn from medical use and is thus irrelevant to deployment [5]. Preprocessing ultimately resulted in 167 features.

Evaluation metric

The f1-score was chosen as an evaluation metric due to the imbalanced nature of the target. Beta was determined as 1 since I am interested in identifying all true positives to capture all patients at risk of readmission so physicians have a chance to improve their treatment (recall) in addition to minimizing false positives due to the cost in resources and time given to the reevaluation of treatment of a patient who received a positive prediction (precision). Hence, a balance between recall and precision is desired which aligns with a beta value of 1.

Machine learning pipeline

Logistic Regression, Support Vector Classifier, Random Forest Classifier, and XGBoostClassifier were trained for this analysis. Parameters were tuned on each algorithm to optimize its performance (see Table 2).

Algorithm	Parameter grid
Logistic Regression	Penalty: None, l1, l2 Class weight: {0:0.4,1:0.6}, {0:0.3,1:0.7}, {0:0.2,1:0.8} C: 0.01, 0.1, 1, 10, 100
Random Forest Classifier	Max depth: None,1, 2, 3, 4, 5, 6, 7, 8, 9, 10 Max features: None, 0.25, 0.5, 0.75, 1.0 Class weight: balanced, balanced_subsample
Support Vector Classifier	Gamma: 0.001, 0.01, 0.1, 1, 10, 100, 1000 C: 0.1, 1, 10
XGBoost Classifier	Alpha: 0, 0.01, 0.1, 1, 10, 100 Max depth: 1, 3, 10, 30, 100

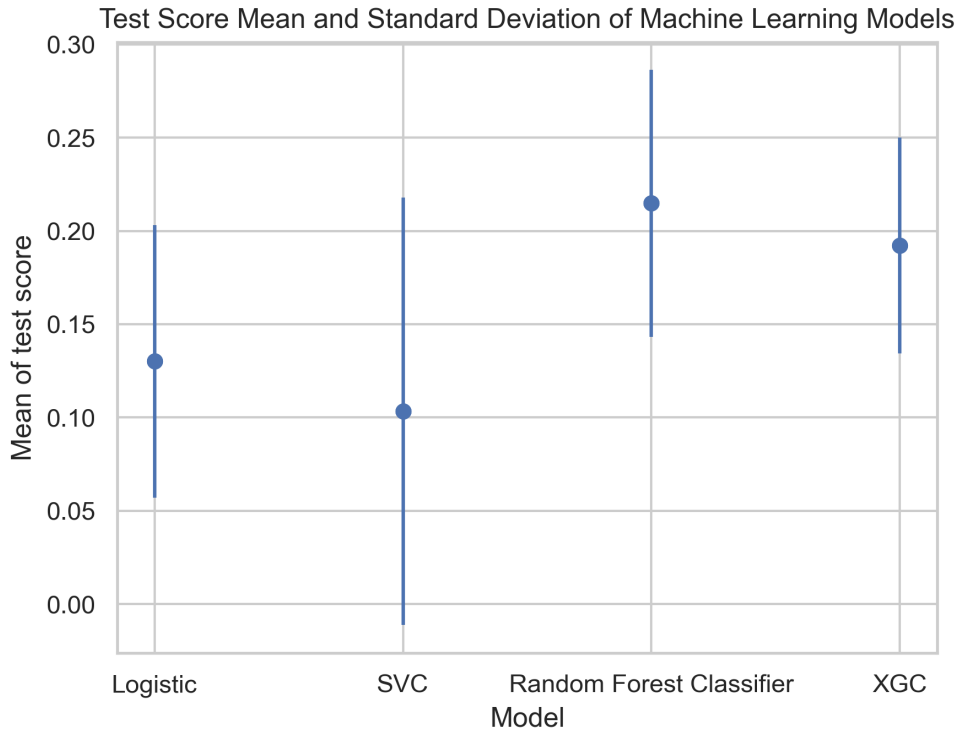
<Table 2. Tuned parameter values for each machine learning algorithm>

To address uncertainties from splitting and non-deterministic methods, the sample is split as described earlier over five random states and then preprocessed. Each algorithm is fit on the training set for each parameter combination with fixed random states to ensure reproducibility. Additionally, XGBoostClassifier has a custom f1 evaluation metric used during the early stopping rounds. Train, validation, and test scores are computed and saved. Lastly, the test sets, models, and predictions are saved to investigate global and local feature importance.

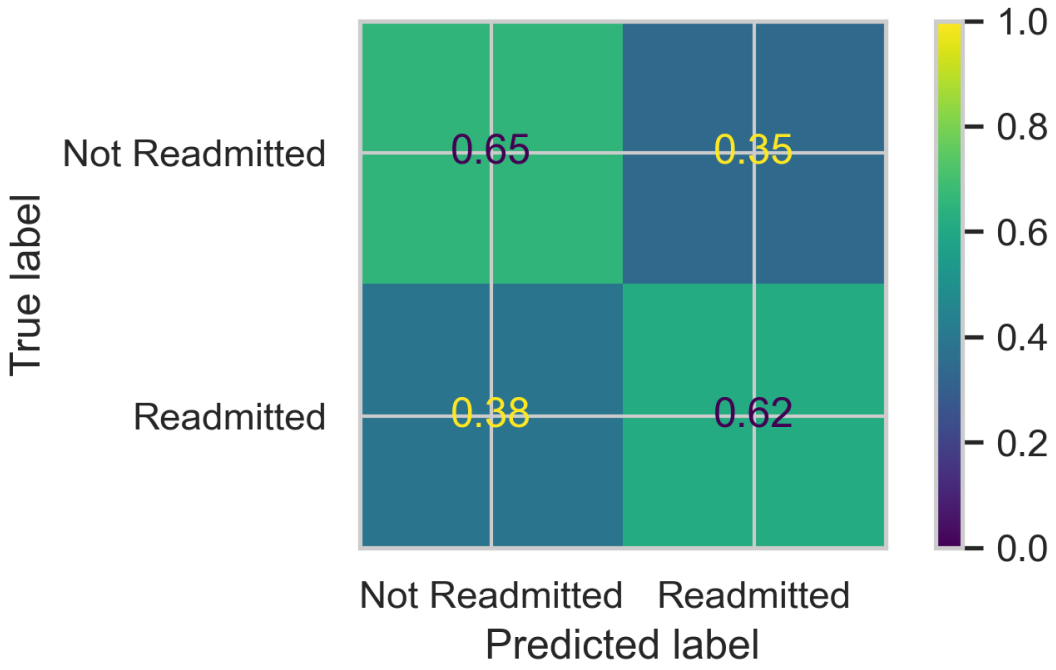
4. Results

	Mean Test Score	Standard Deviation	(Mean test score-baseline)/ Standard deviation
Logistic Regression	0.130	0.073	-0.923
Random Forest Classifier	0.216	0.072	0.254
Support Vector Classifier	0.103	0.115	-0.824
XGBoost Classifier	0.194	0.058	-0.059

<Table 3. Test score mean, standard deviation, and comparison to baseline for each algorithm>

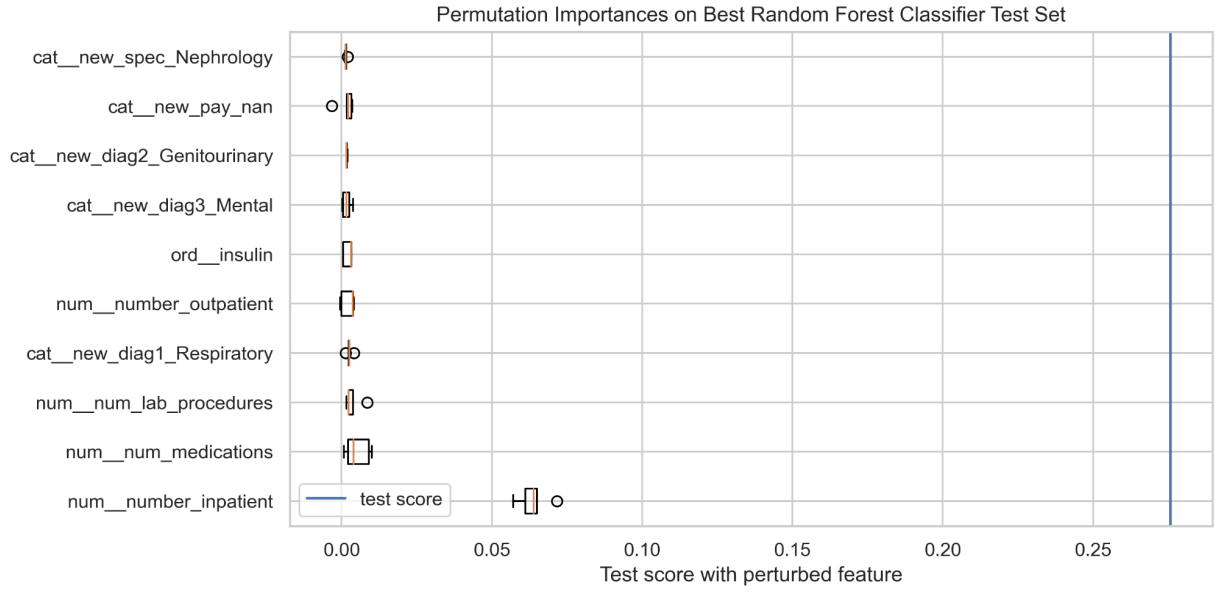


<Figure 4. Test score mean and uncertainty for each algorithm due to splitting and non-deterministic methods>

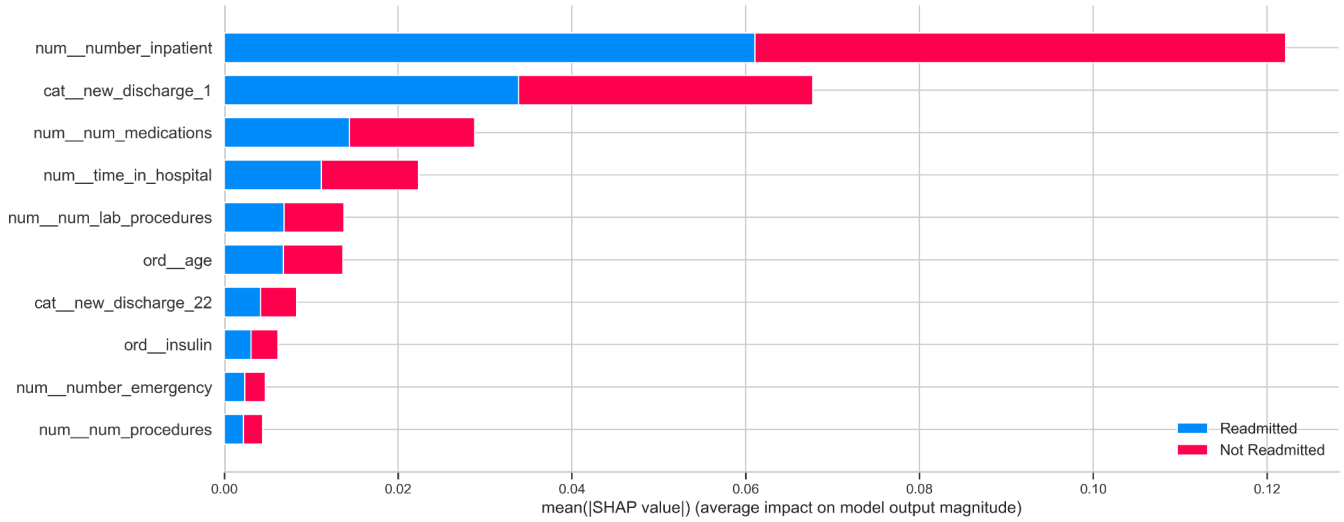


<Figure 5. Confusion matrix with best Random Forest Classifier test set>

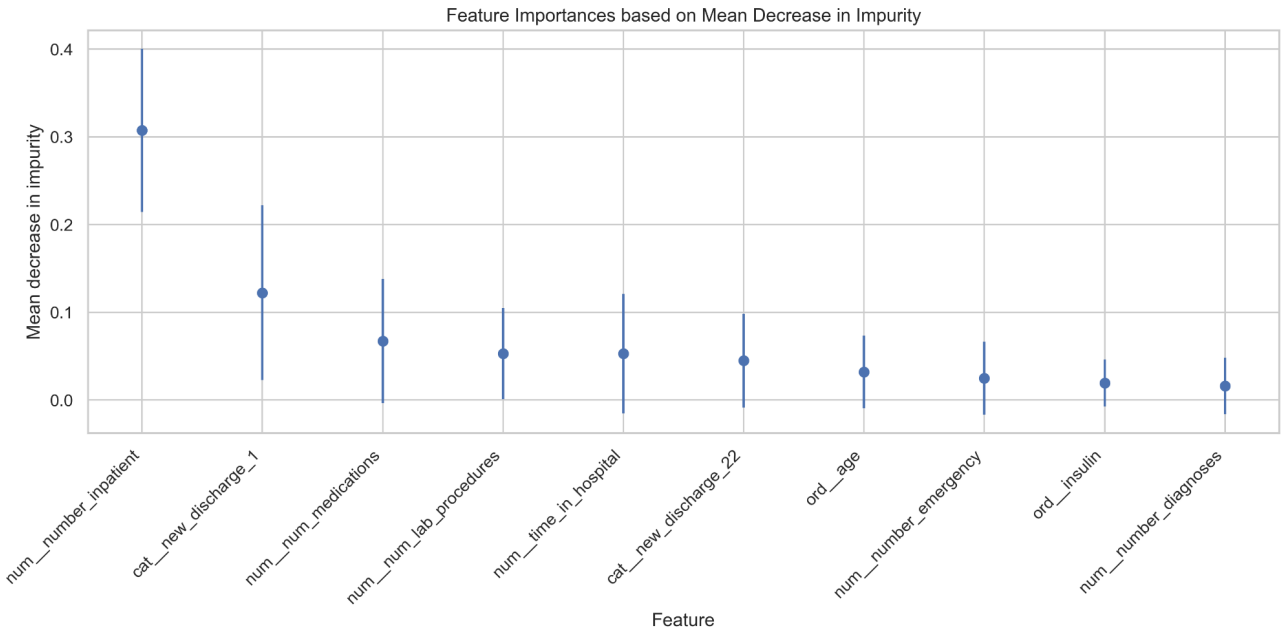
The baseline f1 score is 0.198, which was calculated with the true labels of the sample and all positive predictions. The Support Vector Classifier had the lowest mean f1 score, followed by Logistic Regression, and XGBoostClassifier. The Random Forest Classifier was the only algorithm that on average had a higher f1 score than the baseline, by 0.254 standard deviations, and is thus the best model. From Figure 5 we can see that it captured 62% of the readmissions and 65% of those who were not readmitted in the test set; hence, it performs moderately well as it captures the majority of each group.



<Figure 6. Permutation importances on best Random Forest Classifier test set>



<Figure 7. Global SHAP feature importance on best Random Forest Classifier test set>

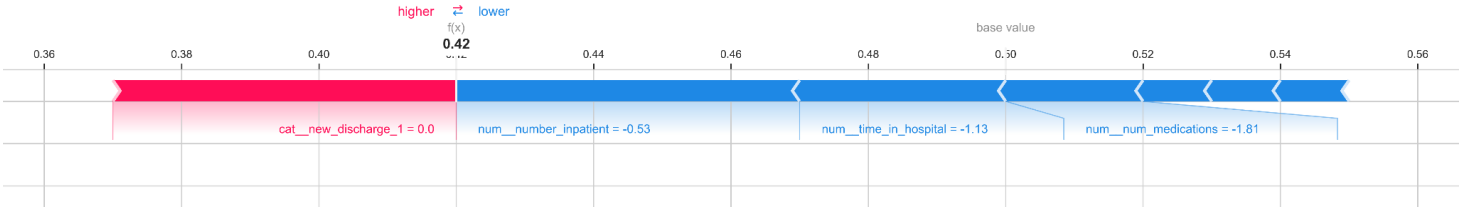


<Figure 8. Feature importance based on mean decrease in impurity>

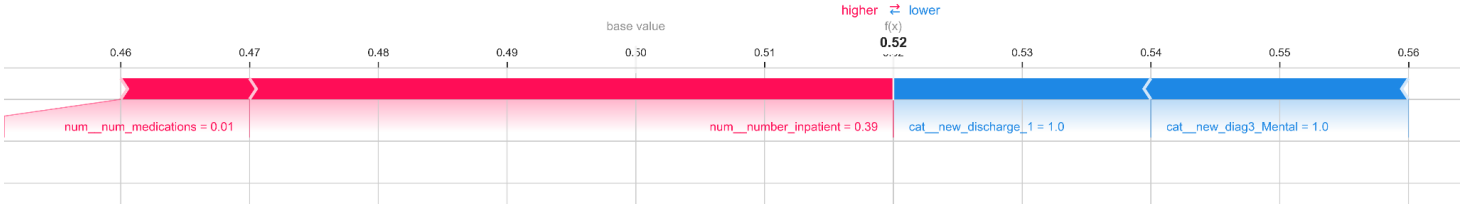
Permutation importance, global SHAP feature importance, and mean decrease in impurity were used to calculate global feature importance. Permutation importance found whether the diabetic patient's physician was a nephrologist to be the most important feature to predict readmission. This is logical as diabetes is linked with kidney problems that would require a nephrologist, which may be indicative of a more serious problem that would require readmission in under a month [6]. Global SHAP and mean decrease in impurity both found the number of inpatient appointments in the past year, whether the patient was discharged to home, and the number of medications to be the top three important features. This is logical since if a patient had a higher number of appointments or medications, or if they were not discharged home, they may have more issues to warrant a readmission. Overall, the change in insulin medication, and the number of medications, labs, and inpatient appointments were in the top 10 most important features across all three metrics.

I found the diagnoses identified in permutation importance surprising: First, whether a patient's secondary diagnosis was a genitourinary disease (of the genitals and urinary organs) was identified which I discovered to be consistent with the literature as diabetes has serious effects on the genitourinary system (urinary tract infections, fungal infections, and more) [7]. Second, whether a patient's primary diagnosis was a respiratory disease was also identified which I discovered to also be consistent with the literature (diabetics have increased oxidative stress which affects their lung tissue) [8]. Lastly, whether a patient's tertiary diagnosis was a mental illness was also identified to help predict

readmission. Perhaps these diagnoses in addition to the diabetic diagnosis increase health complications and the chance of readmission.



<Figure 9. SHAP local value for index 0>



<Figure 10. SHAP local value for index 1981>

For index 0, not discharged home pushes the probability for readmission up as this indicates the patient is not well, while low time in hospital, number of inpatient appointments in the past year, and number of medications push the probability down for a correct prediction of no readmission. These features are consistent with the global feature results.

For index 1981, the high number of inpatient appointments in the past year and medications push the probability of readmission up while discharged to home and a tertiary mental illness diagnosis pushes the probability down for a correct prediction of readmission. Strangely, the mental health diagnosis pushes the readmission probability down; perhaps mental disorders do not warrant hospital readmissions as much as physical disorders or diabetic patients have effective treatment plans for this disease type [9] [10].

5. Outlook

The best model was the Random Forest Classifier with a validation f1-score of 0.292, a max depth of 6, a class weight of balanced, and max features of 0.25.

To improve the predictive power of this model, other algorithms could be employed, such as a gradient-boosting classifier which was used successfully for this problem as

mentioned earlier. Furthermore, with greater computing power and time, the use of a greater portion of the dataset to train the algorithms could also result in a better-performing model. Additional data about the patients' socio-economic background, family history of illness, lifestyle habits, and past diagnoses could also provide the model with stronger predictive power.

6. References

[1] Rizvi, Ali, et al. "Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records." *BioMed Research International*, vol. 2014, 2014, p. 781670. Hindawi Publishing Corporation, doi:10.1155/2014/781670

[2] Clore,John, Cios,Krzysztof, DeShazo,Jon, and Strack,Beata. (2014). Diabetes 130-US hospitals for years 1999-2008. UCI Machine Learning Repository. https://doi.org/10.24432/C5230J.

[3] Long, Andrew. "Using Machine Learning to Predict Hospital Readmission for Patients with Diabetes with Scikit-Learn." *Medium*, Towards Data Science, 30 Jan. 2020, towardsdatascience.com/predicting-hospital-readmission-for-patients-with-diabetes-using-scikit-learn-a2e359b15f0.

[4] Shang Y, Jiang K, Wang L, Zhang Z, Zhou S, Liu Y, Dong J, Wu H. The 30-days hospital readmission risk in diabetic patients: predictive modeling with machine learning classifiers. BMC Med Inform Decis Mak. 2021 Jul 30;21(Suppl 2):57. doi: 10.1186/s12911-021-01423-y. PMID: 34330267; PMCID: PMC8323261.

[5] LiverTox: Clinical and Research Information on Drug-Induced Liver Injury [Internet]. Bethesda (MD): National Institute of Diabetes and Digestive and Kidney Diseases; 2012-. Troglitazone. [Updated 2018 Jun 6]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK548142/

[6] "Diabetes and Chronic Kidney Disease." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 30 Dec. 2022, www.cdc.gov/diabetes/managing/diabetes-kidney-disease.html#:~:text=Each%20kidney%20is%20made%20up,which%20can%20damage%20kidneys%20too.

[7] "Diabetes and Its Impact on Your Urinary and Sexual Health." *Diabetes and Its Impact on Your Urinary and Sexual Health - Urology Care Foundation*, www.urologyhealth.org/healthy-living/urologyhealth-extra/magazine-archives/spring-2017/diabetes-and-its-impact-on-your-urinary-and-sexual-health. Accessed 9 Dec. 2023.

[8] "Lung Diseases Associated with Diabetes." *Das Diabetesinformationsportal*,

www.diabinfo.de/en/living-with-diabetes/complications/lungs.html#:~:text=Diabetes%20is%20accompanied%20by%20ongoing,cancer%20cells%20in%20the%20lungs. Accessed 9 Dec. 2023.

[9] “Mental Illness and the Family: Is Hospitalization Necessary?” *Mental Health America*, mhanational.org/is-hospitalization-necessary. Accessed 9 Dec. 2023.

[10] “Diabetes and Mental Health.” *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 15 May 2023, www.cdc.gov/diabetes/managing/mental-health.html#:~:text=People%20with%20diabetes%20are%202,often%20gets%20worse%2C%20not%20better.