Predicting readmissions and length of stay for diabetes patients

Team: Hannah Gonzalez (519, Aditya's cohort), Christopher Snider (519, Aditya's cohort)
Project Mentor TA: Aditya Kashyap

Abstract:

Predicting whether a diabetes patient is likely to be readmitted to the hospital or have a prolonged length of stay could enable the treatment team to intervene early in their care, potentially reducing unnecessary costs and improving the quality of care. We used linear regression and poisson regression with cross validation to predict the length of stay and we used logistic regression, random forest, and gradient boosting classifiers with cross validation to predict readmissions. After doing cross validation and calculating different scores, we concluded that the poisson regression model was the best model of the two to predict the length of stay. While all classification models performed similarly well (AUC 0.67-0.69), we opted to use logistic regression with L1 regularization to evaluate the final test data since it was a more simple model and had greater interpretability.

Github link: https://github.com/isgla/ML_project

Introduction:

MedPac's report to congress showed that $15 billion in spending in 2005 were accounted for readmitting patients to hospitals within 30 days of discharge, many of which were unnecessary. We want to address this problem because properly managing diabetes in patients could reduce the number of readmissions, which could lead to decreasing health care costs, improved quality, and increased patient satisfaction. Secondly, predicting the length of stay in a hospital could help to monitor the quality of patient care, since the length of stay could be a measurement of efficiency and performance of a hospital.

Related Prior Work:

In a previous retrospective analysis of the diabetes inpatient dataset, Strack et. al (2014) found measuring HbA1c during the admission was associated with a lower readmission rate, which may suggest managing diabetes (e.g., measuring HbA1c and adjusting medications) could reduce the risk of readmissions. A systematic review (Robbins et al., 2019) of 83 studies on readmission risk for diabetes patients reported the most commonly-reported risk factors were comorbidities, age, race, and insurance status. We will compare our model's performance with the results Alahmar's et. al. (2018) obtained for predicting the length of stay of patients in the hospital using the same data set. Our models differ in the way we analyzed and preprocessed the data and we will analyze how this affects pour model's performance:

Formal Problem Setup (T, E, P):

T: Our primary outcome of interest is predicting the likelihood of 30-day readmissions for diabetes patients admitted in the hospital. We intend to predict the patient's risk of readmission, represented by $P(readmit^i=1 \mid r^i, g^i, a^i, d^i, t^i, m^i)$, reflecting the patient's race, gender, age, diagnosis, tests, and medications, respectively. Secondarily we will also attempt to predict length of stay, which may serve as an indicator for complications occurring while in the hospital. For patient *i,* we want to predict the length of stay *l,* denoted as $l^{i} = [r^{i}, g^{i}, a^{i}, d^{i}, t^{i}, m^{i}]$.

E: To do this, we will use the Diabetes 130-US hospitals for years 1999-2008 Data Set (see reference section for source). The dataset contains over 100,000 encounters and 47 features. We will create an analytical dataset that includes the population of interest, diabetes patients at risk of readmission (e.g., excludes patients who had died while in the hospital). We will use 70% of the data as our training set and 30% of the data as our test set. We will also randomly select only one encounter per patient in order to not have the same patient showing up in the training/ test sets.

P: We will calculate the AUC and F1 Score for the readmission prediction model and MSE for the length of stay model. We will also consider sensitivity and positive predictive value (at a chosen threshold) since these metrics would be especially relevant for the clinical team. Additionally, for the length of stay model, we will do a cross validation of 5-folds and get the mean score and standard deviation of the score. We will compare the performance of both models in addition to observing the impact of altering the alpha parameter in poisson regression.

Methods:

Preprocessing: We began preprocessing the data by merging several dictionary tables with the main analytic dataset. We had observed there were more than 700 distinct primary diagnoses for patients in the data. Because some of these diagnoses were overly specific, we obtained additional data from a database of diagnosis categories, the HCUP Clinical Classification System (CCS), that we used to merge with the diabetes dataset's ICD-9 codes to categorize the specific codes into related groupings. We evaluated the prevalence of missing data, and noticed the patients' weight and lab results were missing for more than 80% of patients. We created "none" categories for the lab results and used mode imputation for weight (a categorical variable) stratifying by age group (since the overall mode would not be reasonable for patients <10 years old). The mode for weight by age group was obtained from the training data and used for missing data in the validation and test data. One-hot encoding was used to convert categorical features into binary indicator columns. The features created from the training data were used for the validation and test data (so test data was not used for feature selection). We performed standardization using scikit-learn's StandardScaler for the regularized regression models (using the training data as the reference). We excluded patients who were deceased during the hospital admission (since they could not be readmitted). We also randomly selected one observation per patient, so patients with more than one admission would not appear in both the training and test datasets.

Exploratory Data Analysis: We visualized important features of interest against our target outcome, readmission status, using boxplots for continuous and bar charts for categorical data. We created histograms and density plots for the length of stay outcome (see figures 1, 2). We performed chi-square tests across categorical and t-tests across numeric features to identify associations with readmissions.

Modeling: We used linear regression and poisson regression to model length of stay. We used logistic regression (separately with L1 and L2 regularization), random forest, and gradient boosting classifiers with cross validation to predict readmissions. Hyperparameters were tuned using a grid search for the random forest and gradient boosting classifiers. We used scikit-learn's LogisticRegressionCV to perform cross validation and hyperparameter tuning.

Results: We obtained a Test MSE = 7.33 for the length of stay model using linear regression and a Test MSE = 6.88 for the length of stay model using poisson regression. Additionally, we obtained a Training MSE = 6.61 for the length of stay model using linear regression and a training MSE = 6.6 for the length of stay model using poisson regression. For our classification model prediction 30-day readmissions, we obtained an AUC of 0.68 for the random forest classifier, 0.66 for the gradient boosting classifier, 0.67 for logistic regression with L2 regularization, and 0.68 for L1 regularization. Refer to supplemental figures 3.1-3.4 for a graphical representation.

Baseline approaches we compare against:

When predicting the length of stay, Alahmar's et. al. have previously worked with this data set and have approached this problem as a classification of length of stay, but they stated future work included taking a regression approach to predict the length of stay for diabetic patients and suggested using linear regression and the stacked ensemble method to do this. Thus, we decided to start approaching this problem with a linear regression model and continue from there. A prior analysis predicting readmissions among diabetes patients was performed by Alloghani et al (2019), which used several modeling approaches. Their best performing model was a Naïve Bayes classifier, which achieved an AUC of 0.64.

Implementation Details:

In the Poisson Regression for predicting the length of stay, the following alpha parameters were used; alpha is a constant that multiplies the penalty term and thus determines the regularization strength. alpha = 0 is equivalent to unpenalized GLMs. The accuracy is denoted using r2 score metrics from sklearn.
(alpha = 0) 0.23 of r2 score with a standard deviation of 0.01
(alpha = 1.0) 0.23 of r2 score with a standard deviation of 0.01
(alpha = 0.01) 0.23 of r2 score with a standard deviation of 0.01
(alpha = 30) 0.09 of accuracy with a standard deviation of 0.00
(alpha = 100) 0.23 of accuracy with a standard deviation of 0.01

Using scikit-learn's GridsearchCV along with RandomForestClassifier resulted in selecting a max depth of 4, 100 estimators, and using entropy for splitting. For scikit-learn's GradientBoostingClassifier, the best performing model had a max depth of 8, 500 estimators, and a learning rate of 0.01. We used LogisticRegressionCV, also from scikit-learn, with varying regularization penalties and 5-fold cross validation for training the regularized logistic regression models.

Experimental Results:

We compare our models for predicting the length of stay and the 30-Day Readmission below; the metric used for predicting the length of stay was MSE and the metric used for predicting 30-Day Readmissions was AUC. These results are after cross-validation.

Predicting the length of stay

Method Name	Training MSE	Test MSE
Linear Regression	6.61	7.33
Poisson regression with alpha = 1	6.6	6.88

For the length of stay models, we observed that our models predicted the length of stay of a patient quite well in both models. However, we observed that the accuracy of the linear regression model was not better in either the train and test sets because we have one outlier in the predictions. We tried to solve this by applying a linear transformation, thus when using the poisson regression as a model, this one had a better accuracy and decreased the mean squared error in both test and train sets. Please observe figures 4, 5 for a graphical representation of this on the test set. The poisson regression model also resolved the issue of predicting length of stay below zero for some patients. Regarding the question as to if more data should be added to the sets— as we learned in class, the MSE does not depend directly on the number of training samples, thus adding more data will make the MSE to remain approximately the same.

Predicting 30-Day Readmissions

Method Name	Training AUC	Validation AUC
Logistic Regression with L1 Regularization	0.667	0.682
Logistic Regression with L2 Regularization	0.662	0.672
Random Forest Classifier	0.687	0.684
Gradient Boosting Classifier	0.801	0.695

While we used the full training dataset, we found that performance appeared to level off when using around 60% of the training data, or 25,000 cases (see figure 6 for the AUC by sample size). Overall the models predicting 30-readmissions performed similarly well against the held-out validation set, achieving an AUC of 0.672-0.695. We opted to use the logistic regression classifier with L1 regularization for the final model against the test dataset since it was a more simplified model, seemed to have less overfitting (similar training and validation performance as shown above), and its output of odds ratios could be more interpretable to the clinical team. The number of previous inpatient admissions and ED visits in the past year, being discharged to a facility (as opposed to home), and certain CCS diagnosis categories appeared to be the strongest predictors of readmission. Even though we categorized diagnoses, some diagnoses, such as acquired foot deformities, appeared to be overly weighted in the model, which may be due to their rarity leading to complete/quasi-complete separation. The final model (logistic regression with L1 regularization) achieved an AUC of 0.67 on the holdout test set, suggesting the model was not overfitting the training data. At a threshold of 0.51, the sensitivity was at 80% but the positive predictive value was only 10%. Since the outcome was unbalanced in the data, we also calculated the F1 score, which was 0.14.

Conclusions and Future Work

Conclusions:

The model that performed the best after cross validation when predicting the length of stay in the hospital was the Poisson regression. Additionally, we can conclude that the alpha parameter in Poisson Regression does not improve the r2 score, but it can decrease it, thus we observe how modifying this parameter could change the performance of the model.

For the readmission model, we obtained an AUC of 0.67 on the test data, which was slightly better than Alloghani et al (AUC=0.64). However, there was a large difference between the sensitivity and positive predictive value (0.80 vs 0.10 at our chosen threshold of 0.51). Depending on the intensity of the intervention for these patients, the clinical team must weigh the importance of identifying all patients against lowering the false positive rate, which could lead to alert fatigue.

Our analysis was limited to previously collected data. In order to protect the patient's privacy, the data curators categorized some of the features instead of providing raw values that could have been obtained in the clinical setting. In some instances, this likely led to a loss of important information (e.g., when categorizing age into decades or lab results as above/below certain thresholds) that may have reduced the performance of our model. The data was structured with one row per admission, so values that are typically recorded multiple times (e.g., lab results) were aggregated by taking the maximum value). If we were deploying the model in a real-world setting, we may also be interested in the last values recorded prior to discharge or even modeling changes in the values over time.

Future work:

We will improve the accuracy of the length of stay model by trying other linear transformations— we observed how accuracy improved and the error decreased when we used Poisson Regression instead of simple Linear Regression, thus we will try other linear transformations to achieve greater accuracy. We will attempt to improve classification accuracy by resampling to achieve better balance in the outcome of interest as well as for sensitive groups (e.g., race). We believe we could improve performance of both models by adding additional data, especially exact measures for lab results because patients with extreme value may be more likely to have adverse results. We could use internal data (e.g., from the University of Pennsylvania Health System) to test this assumption.

We will constantly monitor the changes in HIPAA law and ensure that our project remains HIPAA compliant. Also, if any of these models were to be deployed, we would frequently do post deployment monitoring to ensure that it is not negatively impacting a certain part of the population. We would also monitor the data quality of the features used by the model to ensure they are being measured consistently.

Ethical Considerations and Broader Impacts:

We considered HIPAA while doing this project. We are aware privacy is a major ethical concern, thus the dataset we used for this analysis was de-identified according to the Safe Harbor method, where direct and indirect identifiers were removed (including names, medical record numbers, dates of service, ages >=90 years, etc.), which ensures confidentiality and integrity to the patients.

The ethical approach followed throughout the project was utilitarian— we considered what was best for most people while also respecting individual rights. We also considered all stakeholders, including patients, physicians, employers, insurance companies, pharmaceutical firms and government, when working on this project.

As to broader impacts; race and age are legally recognized as protected classes, and we acknowledge the use of both as features in our models. This could raise ethical concerns since it could affect a certain part of the population if any of these models were deployed. However, we will follow the notion of fairness of independence, where risk evaluation will be independent of age and race. When evaluating our model's performance within subgroups, the AUC was similar across groups, but white patients had the highest AUC and they were overrepresented in the data (see supplemental table 7). As mentioned future work will entail resampling to ensure groups are represented more equally.

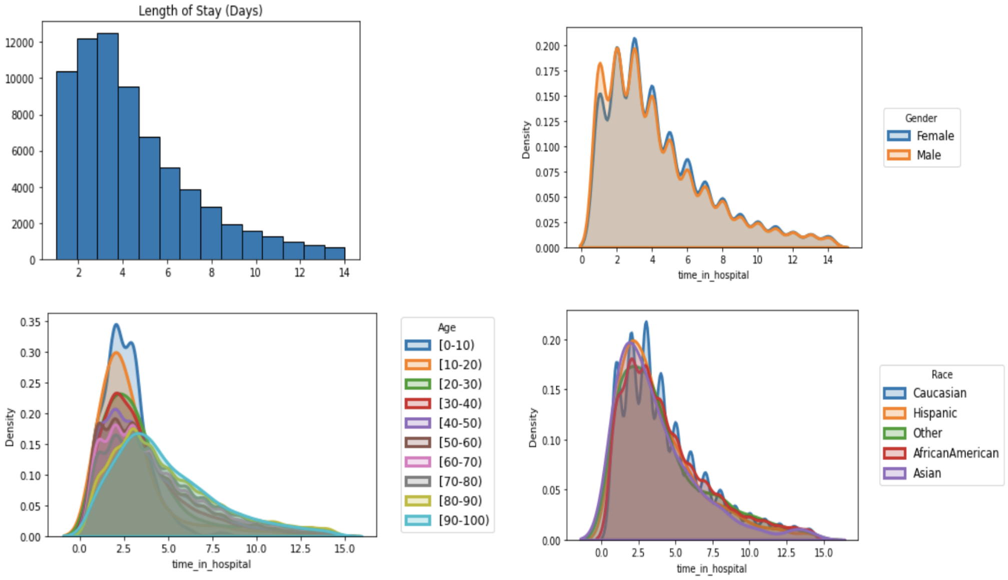
Prior Work / References:

1. Data Source: Diabetes 130-US hospitals for years 1999-2008 Data Set https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008).
2. Additional data source: Clinical Classifications Software (CCS) for ICD-9-CM https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp
3. We analyzed the benefits of aiming to reduce the readmissions in hospitals with this paper: https://www.commonwealthfund.org/publications/newsletter-article/focus-preventing-unnecessary-hospital-readmissions
4. Robbins TD, Lim Choi Keung SN, Sankar S, et al. Risk factors for readmission of inpatients with diabetes: A systematic review. Journal of Diabetes and its Complications 2019;33:398–405. doi:10.1016/j.jdiacomp.2019.01.004
5. 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.
6. A. Alahmar, E. Mohammed and R. Benlamri, "Application of Data Mining Techniques to Predict the Length of Stay of Hospitalized Patients with Diabetes," 2018 4th International Conference on Big Data Innovations and Applications (Innovate-Data), Barcelona, Spain, 2018, pp. 38-43, doi: 10.1109/Innovate-Data.2018.00013.
7. A similar analysis of the diabetes readmission data was performed by Alloghani et al. (2019): Alloghani M, Aljaaf A, Hussain A, *et al.* Implementation of machine learning algorithms to create diabetic patient re-admission profiles. *BMC Med Inform Decis Mak* 2019;19. doi:10.1186/s12911-019-0990-x

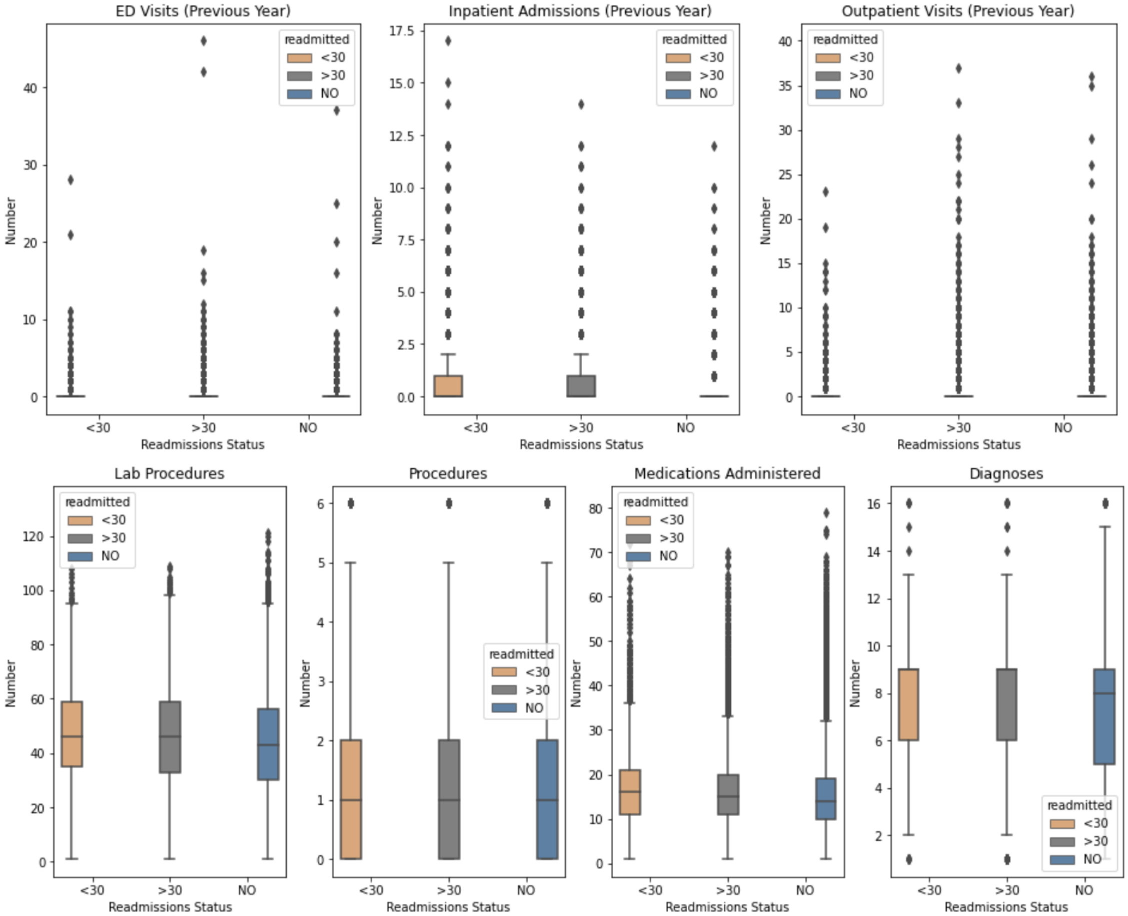
Supplementary material:

Figures and Tables:

1. Analysis for the length of stays (days) in the hospital model

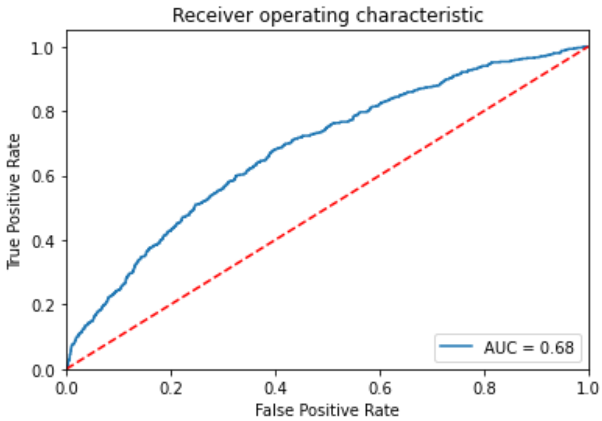


2. Analysis for the 30 day-readmissions model

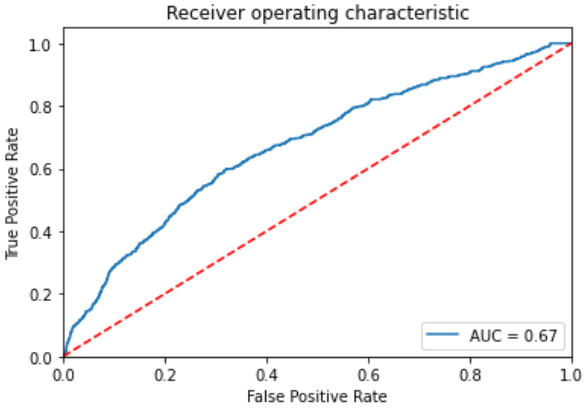


3. Area under the curve (validation dataset)

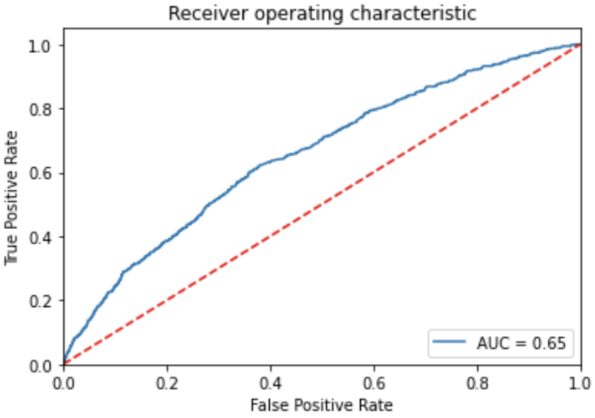
3.1 AUC of Logistic Regression (L1 regularization):



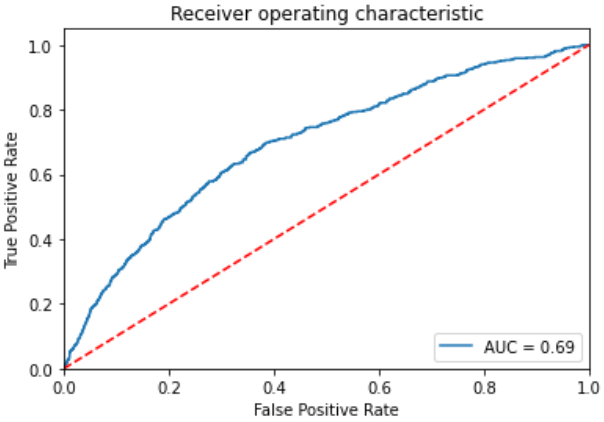
3.2 AUC of Logistic Regression (L2 regularization):



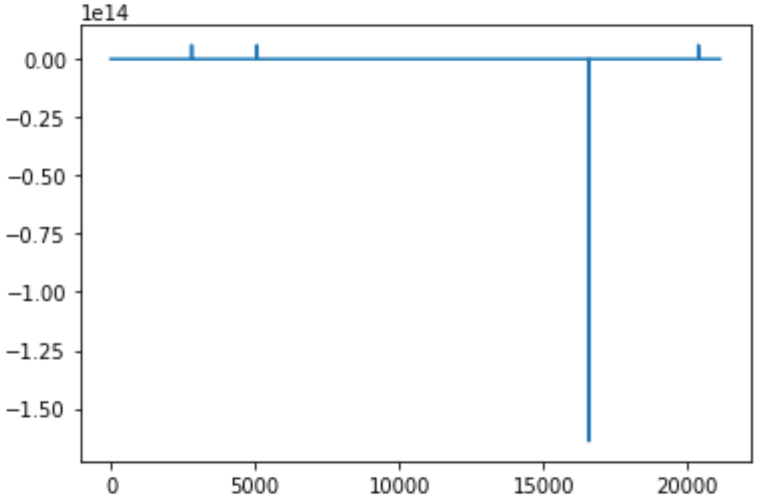
3.3 AUC of random forest classifier:



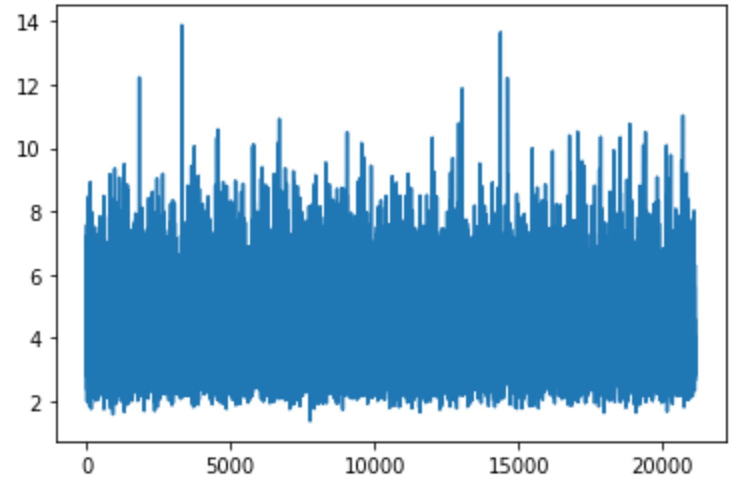
3.4 AUC of gradient boosting classifier:



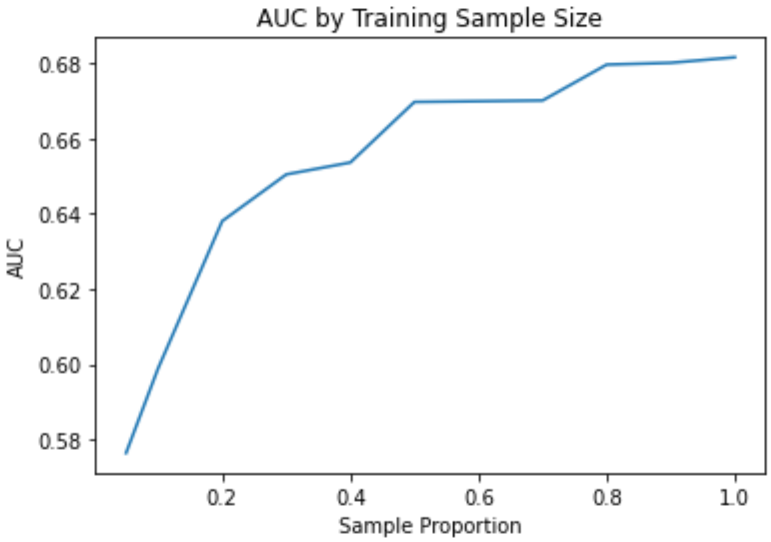
4. Predictions of linear regression model on X test set



5. Predictions of poisson regression model on X test set



6. AUC Score with Varying Training Sample Size



7. Performance across sensitive subgroups

Subgroup	Category	Observed Readmission Rate	Predicted Readmission Rate	Accuracy	AUC	PPV	Recall
race	African American	0.058	0.525	0.500	0.663	0.079	0.716
race	Asian	0.070	0.532	0.487	0.614	0.083	0.636
race	Caucasian	0.078	0.612	0.438	0.675	0.104	0.823
race	Hispanic	0.066	0.487	0.533	0.629	0.089	0.655
race	Other	0.053	0.523	0.504	0.668	0.076	0.750
race	Unknown	0.031	0.430	0.586	0.663	0.055	0.765
gender	Female	0.074	0.599	0.448	0.677	0.101	0.817
gender	Male	0.070	0.574	0.465	0.671	0.095	0.783
gender	Unknown/ Invalid	0.000	0.000	1.000	NaN	NaN	NaN

age	[0-10)	0.020	0.122	0.898	1.000	0.167	1.000
age	[10-20)	0.024	0.297	0.715	0.842	0.061	0.750
age	[20-30)	0.030	0.330	0.682	0.730	0.064	0.700
age	[30-40)	0.062	0.396	0.637	0.760	0.120	0.766
age	[40-50)	0.056	0.390	0.624	0.679	0.090	0.628
age	[50-60)	0.055	0.422	0.603	0.701	0.094	0.725
age	[60-70)	0.077	0.575	0.469	0.678	0.105	0.783
age	[70-80)	0.081	0.705	0.349	0.639	0.096	0.833
age	[80-90)	0.089	0.772	0.296	0.622	0.102	0.884
age	[90-100)	0.077	0.801	0.265	0.575	0.089	0.930