# Patient Readmission Analytics:

## Team name: Patient Readmission Analytics

## Team members:

Name:Sri Sai Gowtham Nadakudati

Email id: srnada@iu.edu

Name:Bhavani Prasad Rao Ejanthkar

Email id: [bejanthk@iu.edu](mailto:bejanthk@iu.edu)

## Introduction:

Healthcare is one of the costliest and most important among all the industries. The tests that are performed are usually costly and incorrect analyses and medications may lead to detriment in the health of a patient. The objective of this project to create a model that can predict if a patient with diabetes will need readmission or not after primary diagnosis. We will also take into consideration the medications prescribed and predict whether the patient will be readmitted due to diabetes in the future or not.

From the data collected from over 130 hospitals in the US over a period of 10 years, a wide range of data related to patients was collected. We have done some preprocessing to this dataset to perform logistic regression to form a prediction model.

## Pre-processing:

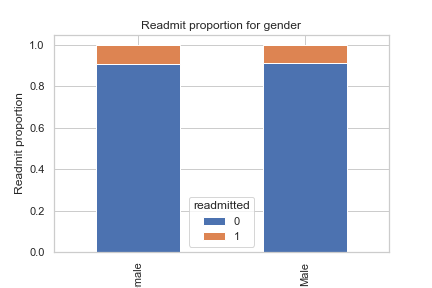
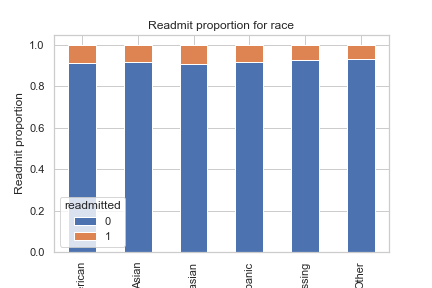
* Features such as patient number, encounter id, payer code etc. were removed from the dataset as they have no logical significance to the prediction of readmission.
* Unary columns were dropped as they do not carry any meaning.
* Weight column was dropped due to extreme sparsity.
* The primary diagnosis column with icd9 values was reclassified to have fewer values based on the type of disease.
* Several categorical values such as admission source and discharge disposition were also reclassified to reduce the number of values for them
* Initially only HbA1c result is added to the preprocessed data to perform regression.
* Class label 'readmitted' is made into binary variable by classifying ‘<30’ as 1 and ‘NO’, ‘>30’ as 0

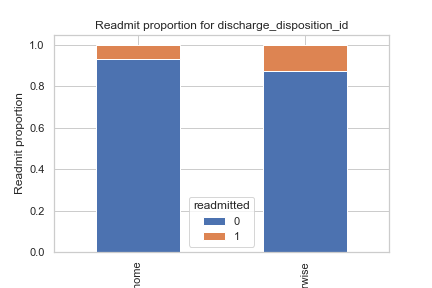
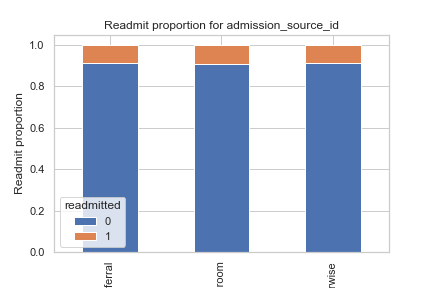
The initial features in the dataset are:

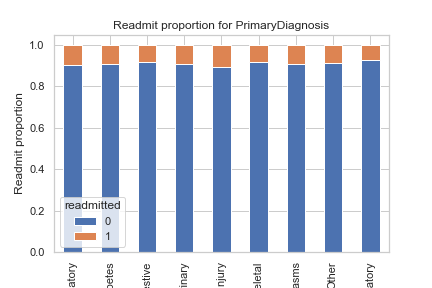
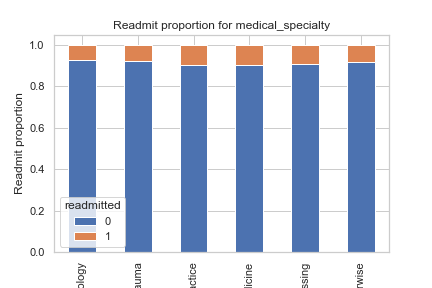
['race', 'gender', 'age', 'discharge\_disposition\_id', 'admission\_source\_id', 'time\_in\_hospital', 'medical\_specialty', 'PrimaryDiagnosis', 'A1Cresult', 'readmitted' ]

## Data exploration:

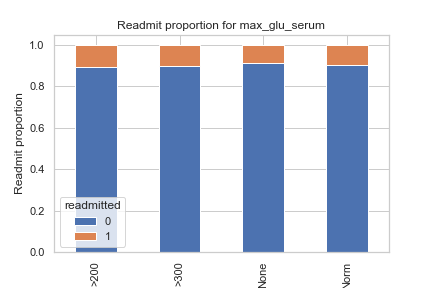
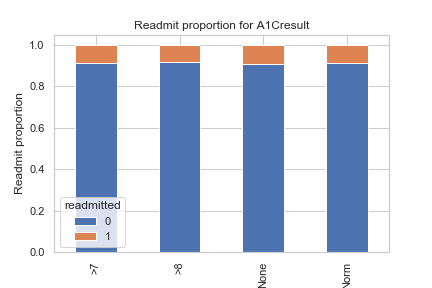
We have seen that readmit observations are about 8.97% and no-readmit observations are about 91.03%. This means that the data is not balanced. We will try to find which variables are ideal for prediction by making plots for all the categorical features.



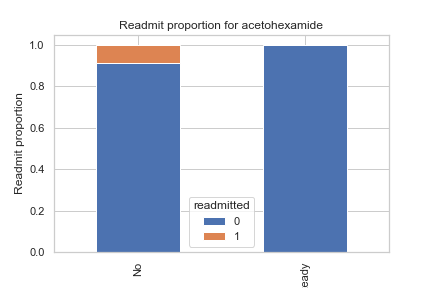
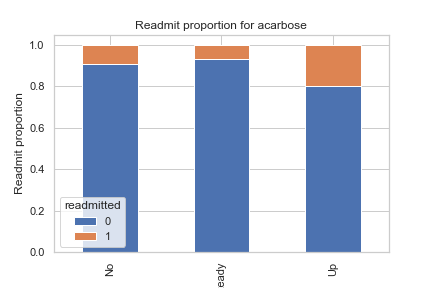


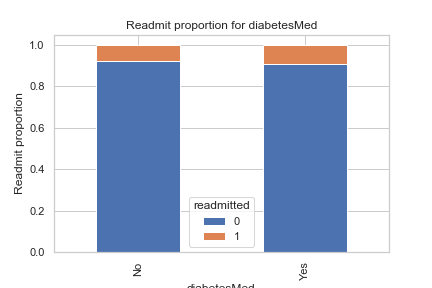
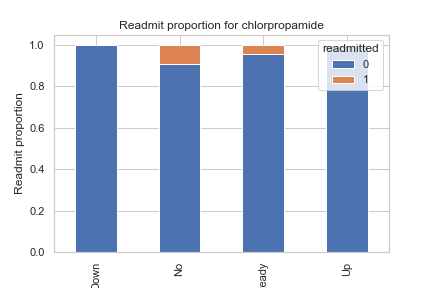


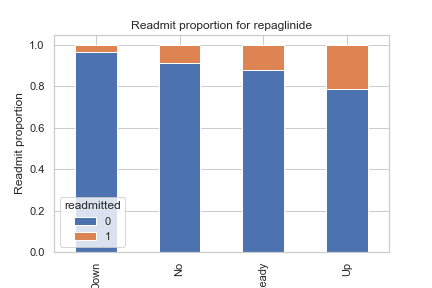
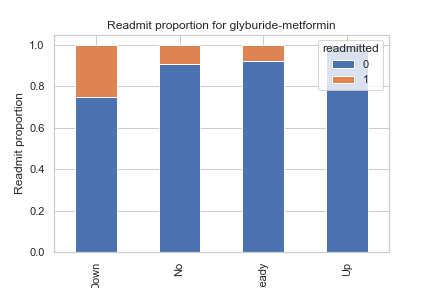
HbA1c result and max\_glu\_serum result



Other medications





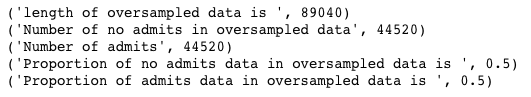


There are several other features we wish to include to improve as many contributing factors as possible. However, as we can see, many of the factors are not very useful. Very few factors such as glyburide-metformin, repaglinide, acarbose etc., seem to be useful.

We tried to include all the features in the model, but the model did not converge, and we were unable to create a regression model. Therefore, we have decided to include very few features initially and progressively check which features are useful in building a model.

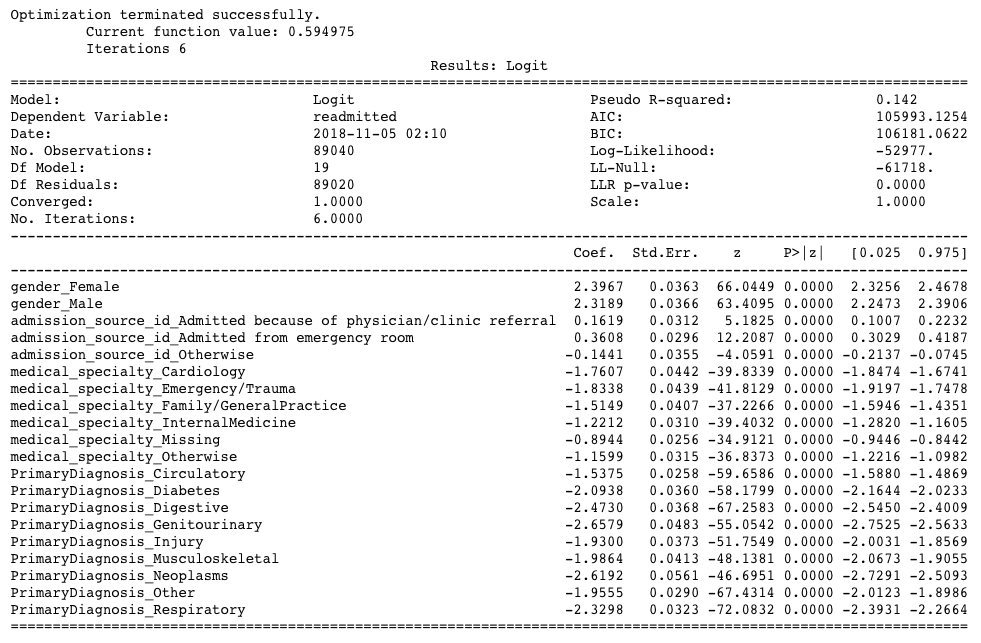
## Oversampling:

Since the data is not balanced, we will perform oversampling using SMOTE. We have split the data into training and test data.



## Recursive feature elimination:

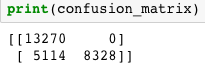
We will create dummies of the feature data and recursively find which values are useful for the model. Using RFE we will create a model to generate most prominent features with small P values and generate our final dataset to create regression model.

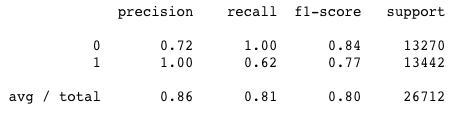


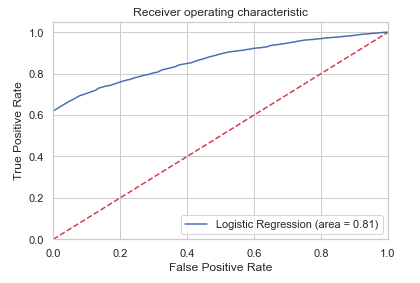
## Logistic Regression Model Fitting:

We have calculated certain parameters using our model. They are as follows:









## Further work:

We need to further include several other features such as medications and test results in the model to improve the prediction. As we increase the features, the accuracy may drop. So, care must be taken not to include such features.

We are currently using Logistic regression, but we also wish to perform the analysis using other methods to find if other models have better prediction rates. We also need to use a hypothesize the best model for our data.