Research paper link : https://arxiv.org/abs/1602.04257

Hospital readmissions are expensive and reflect the inadequacies in healthcare system. In the United States alone, treatment of readmitted diabetic patients exceeds 250 million dollars per year. Early identification of patients facing a high risk of readmission can enable healthcare providers to to conduct additional investigations and possibly prevent future readmissions. This project helps to predict the same using some machine learning algorithms.

Because of high number of missing values, columns weight, payer\_code and medical\_specialty are removed. Since there are adequate number of observations in the dataset, the other missing values need not be imputed and those observations can be deleted.

### Except insulin all other features of medication are removed as they show little to no variation. It has been identified after reading through the research paper that prmary diagnosis is a sufficient indicator of diabetes and hence secondary/addition diagnoses can be removed.

### It has been identified by reading in the research paper that the values that start with 250 are indicators of diabetes

### It is expected that young people are much less prone to diabetes and the data also indicates that most people are within the age groups 60-80. So it made sense to recategorize age groups.

### Converted discharge disposition, admission source and types into 2 levels for ease of analysis. I do not expect them to have much impact in the readmission rate.

### After all the pre-processing, I used logistic regression and random forest to predict the model performance and used bagging and boosting methods to improve accuracy.