Predicting Diabetes Patient Hospital Readmission

Pavel Zimin, PhD
Mentor: Ramkumar Hariharan, PhD

Problem Statement

- Hospital readmission is a highly preventable cause for high healthcare costs
- The ability to predict hospital readmission will help prioritize patients that will benefit from hospital discharge follow up programs

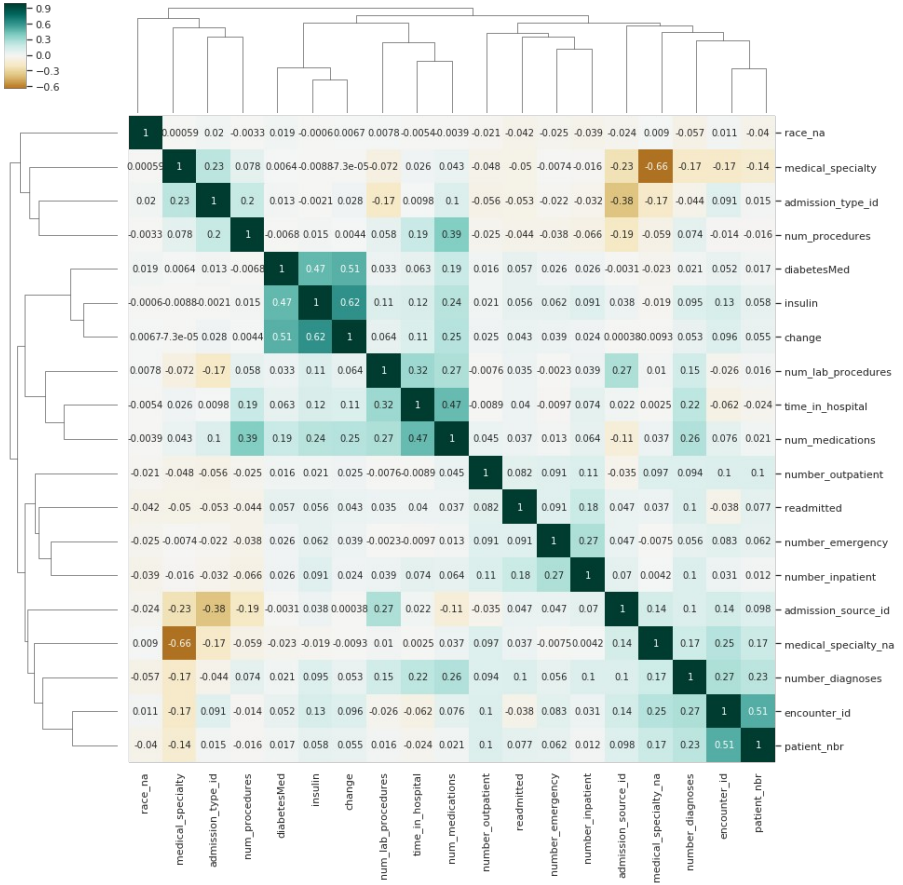
Business use cases:

- The outcome of this analysis will be helpful to the hospital healthcare teams with prioritizing patient support program
- This analysis will benefit patients who will receive improved health care, decreased chances of readmission while incurring smaller cost

Data Wrangling Steps

- Hospital readmission data were downloaded from UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008)
- The dataset contains 101,766 observations of unique hospital encounters with 50 variables: 13 columns of integer type, 37 columns of object type
- Medical diagnosis codes with their hierarchical groupings were downloaded from a GitHub repository (https://github.com/sirrice/icd9.git) and merged with the readmission data set
- For each variable with missing values a separate column was created with values indicating the missing values
- Each categorical variable was encoded with integer values, the code was saved in a dictionary
- No outliers were removed

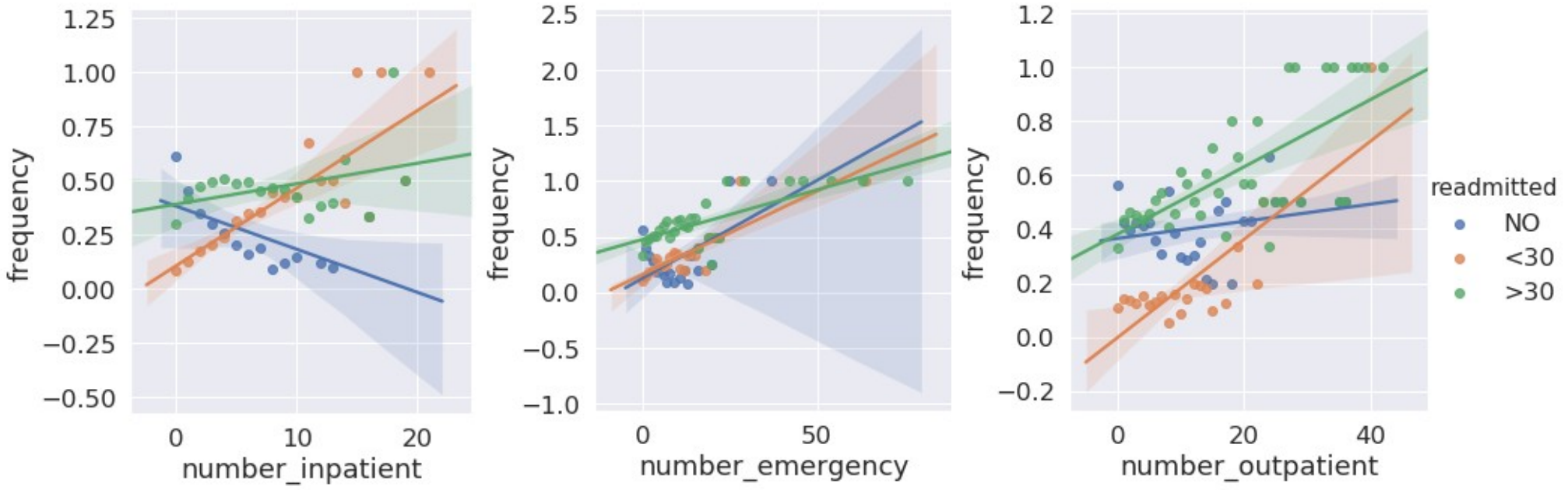
Correlation Clustermap



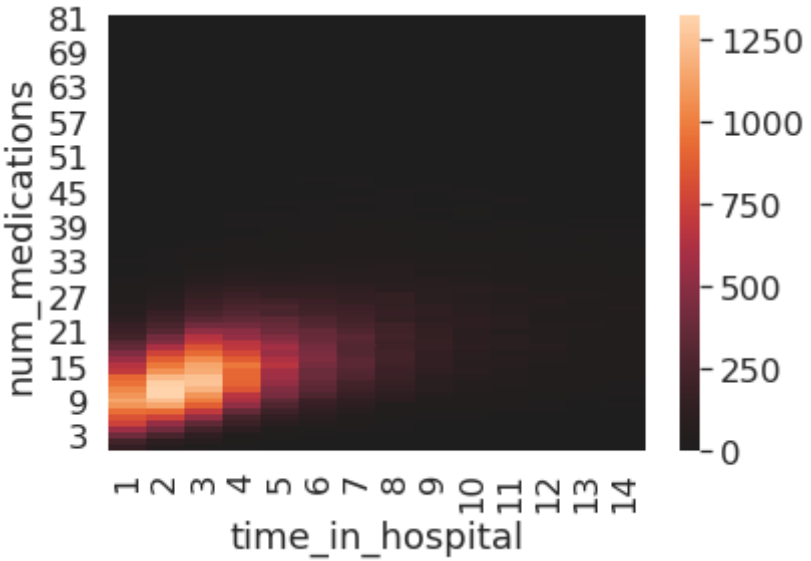
- moderate to low correlation between the readmitted variable and other variables
- variables showing largest correlations are:

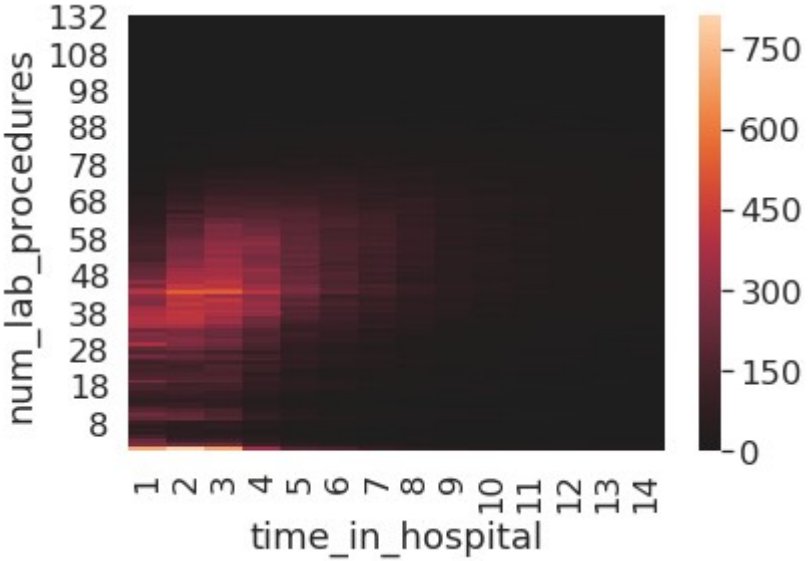
1) number_inpatient (the number of inpatient visits in the year preceding the encounter)
2) number_emergency (the number of emergency visits in the year preceding the encounter)
3) number_outpatient (the number of outpatient visits in the year preceding the encounter)

Frequencies of Selected Variables

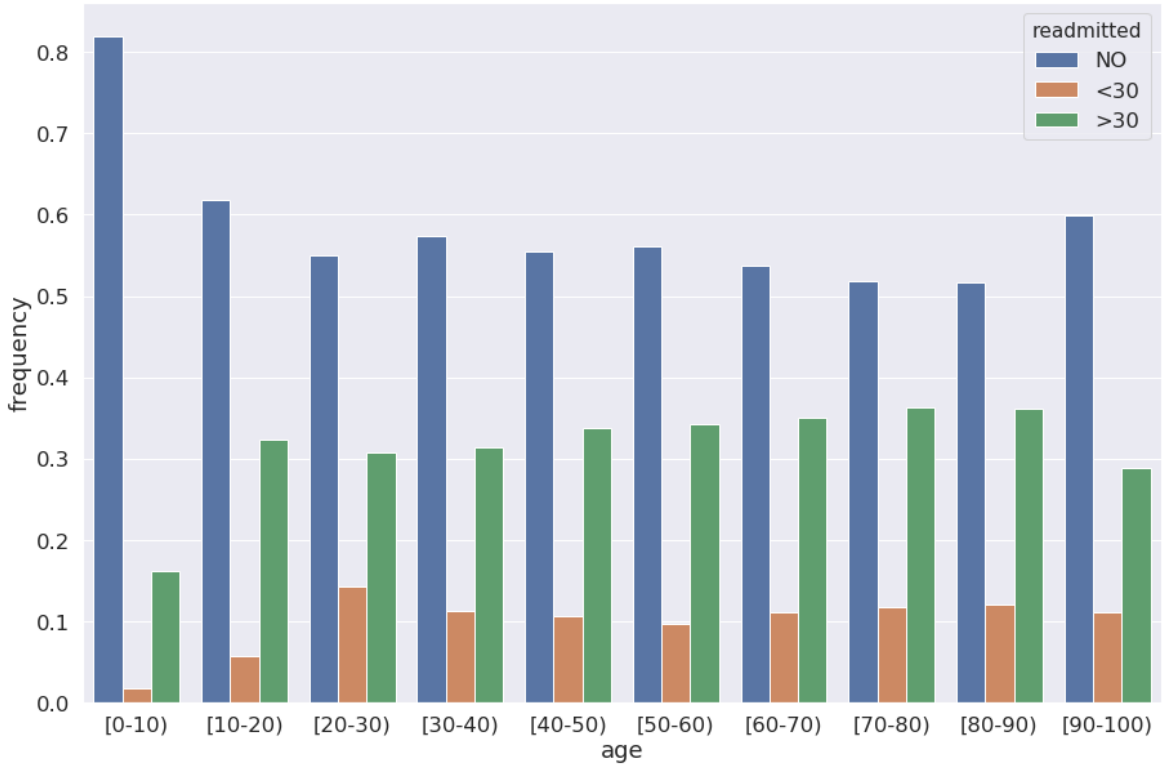


Sanity Check of the Data Set

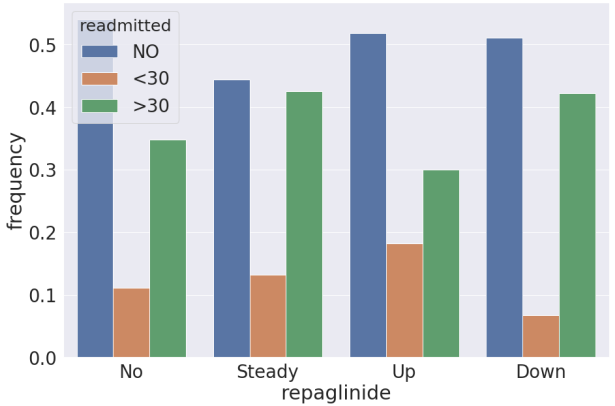


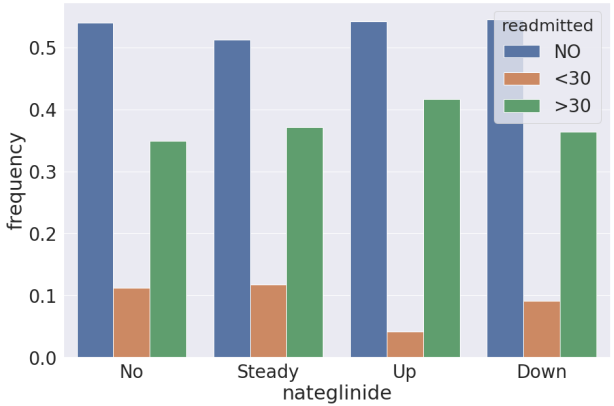


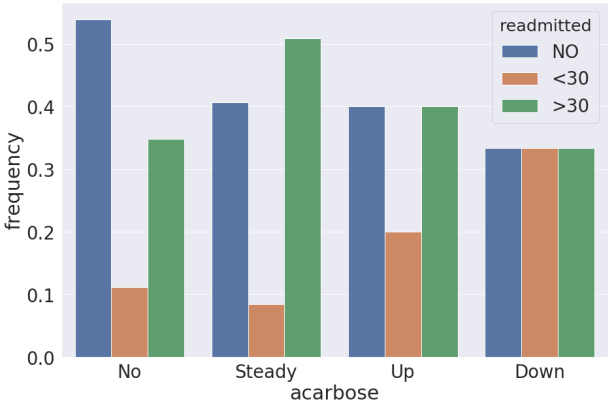
Age Group Frequencies



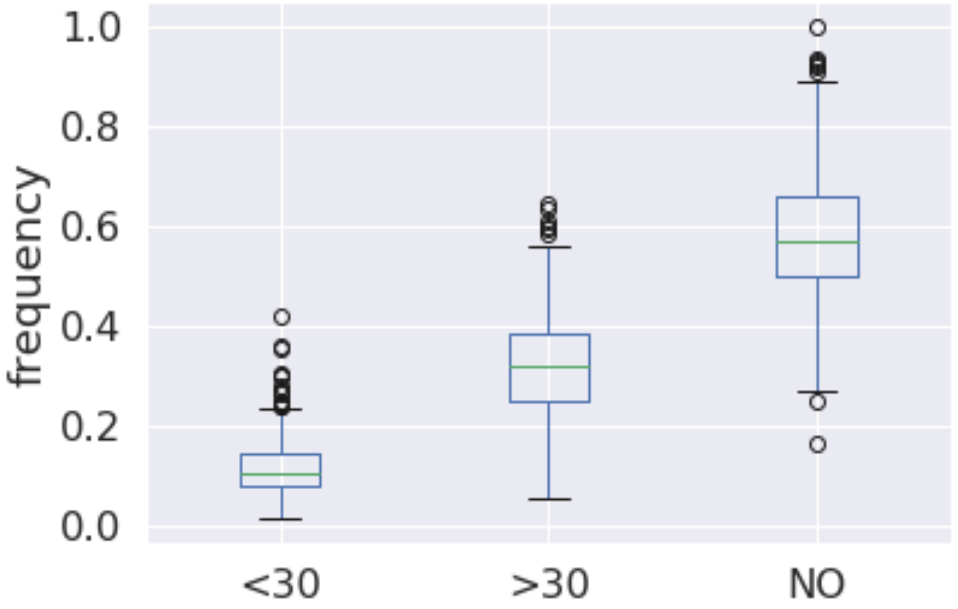
Frequencies of medication dosage changes







Primary Diagnosis Frequency Distributions



The top primary diagnoses in the group with readmission within 30 days:

1) encounter for other and unspecified procedures and aftercare,
2) diabetes with renal manifestations,
3) peritonitis and retroperitoneal infections.

Conclusions

- There is moderate to low correlation between the readmitted variable and other variables.
- The variables showing the largest correlations are: number_inpatient, number_emergency, and number_outpatient.
- The most dramatic changes in the frequencies of medication changes were for the following medications for treating diabetes: repaglinide, nateglinide, and acarbose.
- Distribution of the primary diagnoses shows that for some primary diagnoses the frequency of readmission within 30 days is much higher than the median frequency for that group.
- The top primary diagnoses in the group with readmission within 30 days are
    1) encounter for other and unspecified procedures and aftercare
    2) diabetes with renal manifestations
    3) peritonitis and retroperitoneal infections