PREDICTION OF DIABETES READMISSIONS

Aisha Kamara, Data Scientist
October 4th, 2021

PRESENTATION OVERVIEW

01 **Introduction**

Brief background on Diabetes and readmission rates in the US.

02 **Problem Statement**

Problem statement introduction.

03 **Data Cleaning**

Cleaning of the dataset

04 **EDA Findings**

EDA results

05 **Modeling Results**

Models used and results

06 **Conclusion**

Findings and recommendations

INTRODUCTION

3.3 Million

Readmitted Patients

$41 Billion

Hospital Costs

DIABETES

Diabetes is a heterogeneous group of diseases that, through various mechanisms, cause hyperglycemia, commonly known as high blood sugar, a buildup of glucose in a person's bloodstream at dangerously high levels due to a person's body not being able to produce enough insulin in order to regulate glucose within the bloodstream.

PROBLEM STATEMENT

DATASET OVERVIEW

- The data obtained from the USC data repository, represents 100,000+ unique inpatient diabetes medical visits over 10 years (1999–2008) of clinical care at 130 hospitals and integrated delivery networks in the United States.

- The data contains such attributes as patient number, race, gender, age, admission type, time in hospital, medical specialty of admitting physician, number of lab test performed, diagnosis, number of medications, diabetic medications, number of outpatient, inpatient, and emergency visits in the year before the hospitalization, etc.

PROBLEM STATEMENT

Problem 1

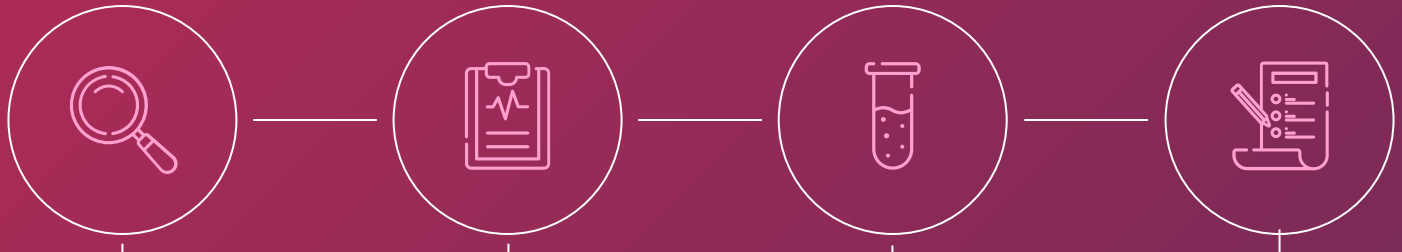
What factors are the strongest indicators of hospital readmission for a diabetic patient?

Problem 2

How well can I predict hospital readmission with "limited" features in this dataset?

DATA CLEANING

CLEANING PROCESS



01

Missing Values

Looked for missing values.

02

Dropping

Dropped features with 90% of their data missing. Features with 2% or less of their data missing, dropped the corresponding rows.

03

Convert Strings

Converted string variables to integers of floats for EDA and modeling purposes

04

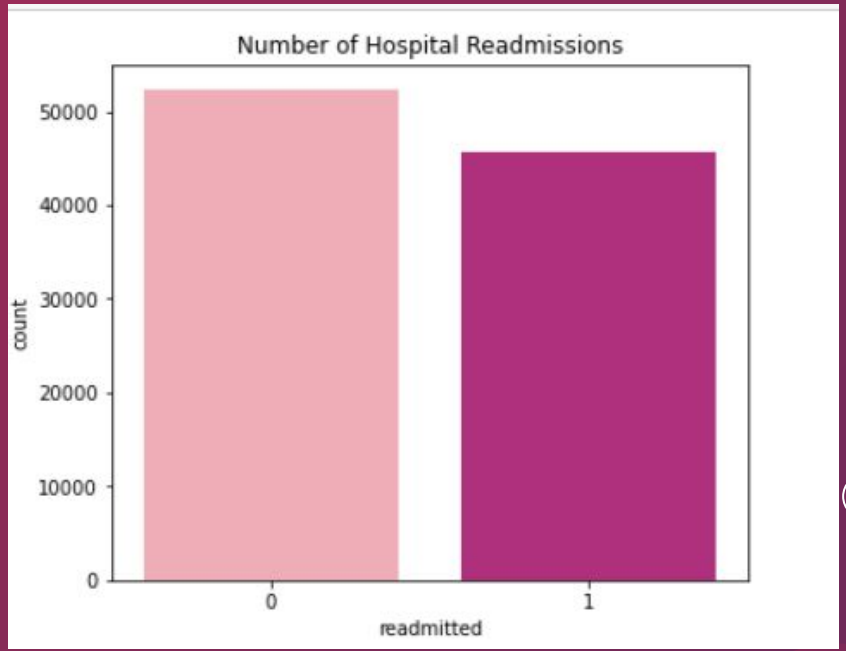
Conclusions

98,052 observations and 48 features remaining for EDA

EDA

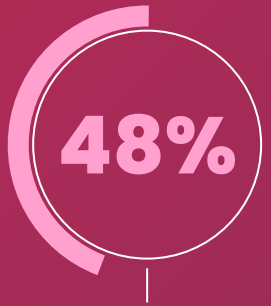
READMITTANCE RATES

- 98,052 Observations
- 47 Features
- 1 Target Feature
- 54% of patient had no record of readmission
- 46% of patients, were readmitted after discharge



Categorical Features

DEMOGRAPHICS



Age

48 % of patients readmitted are between 60-80 years of age.

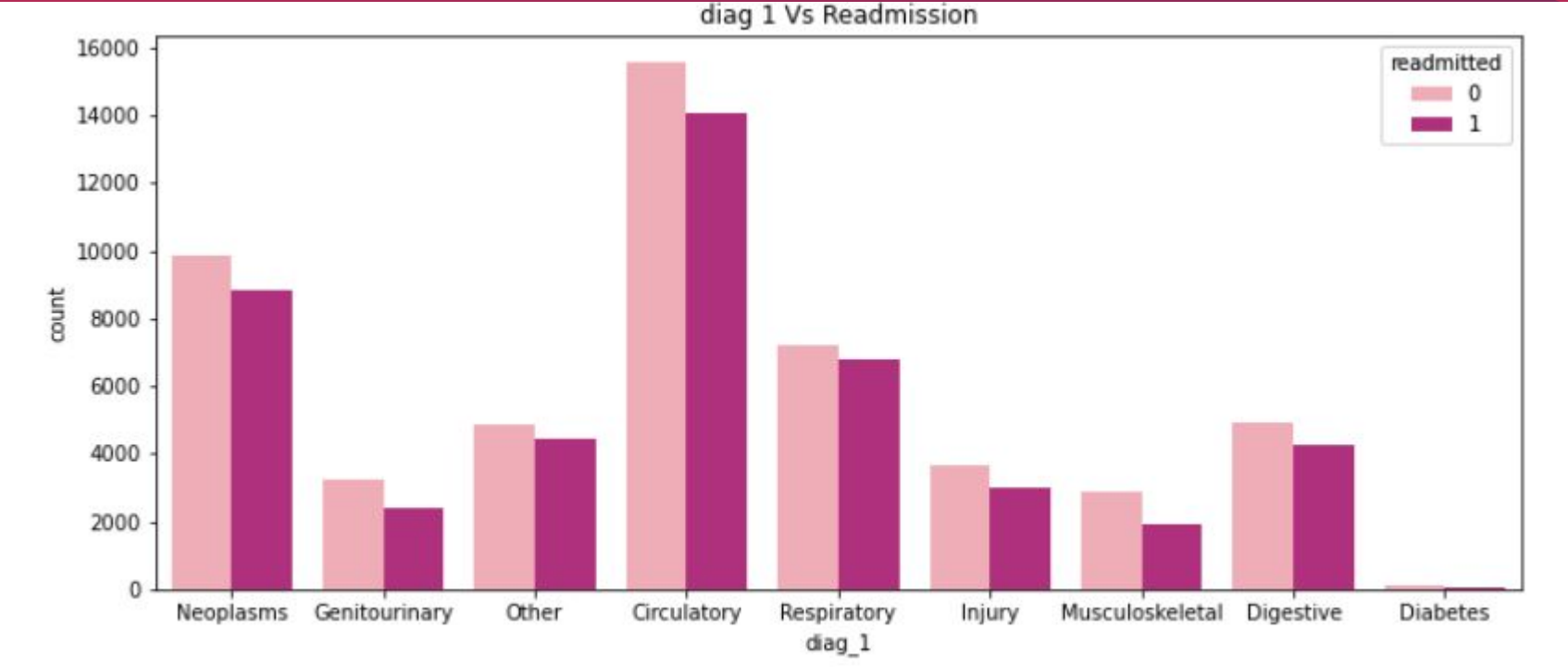
Gender

53 % of readmitted patients are females.

Race

76% of readmitted patients are Caucasian.

DIAGNOSIS TYPE



Continuous Features

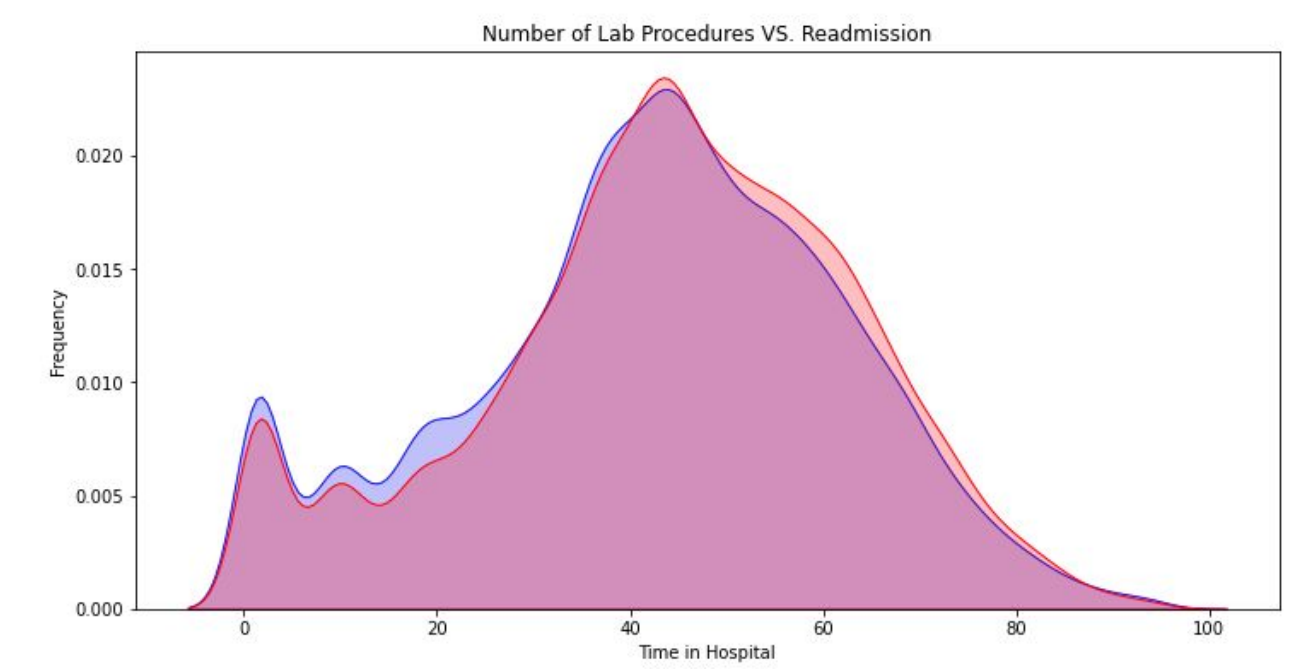
CONTINUOUS FEATURES

- 8 continuous features

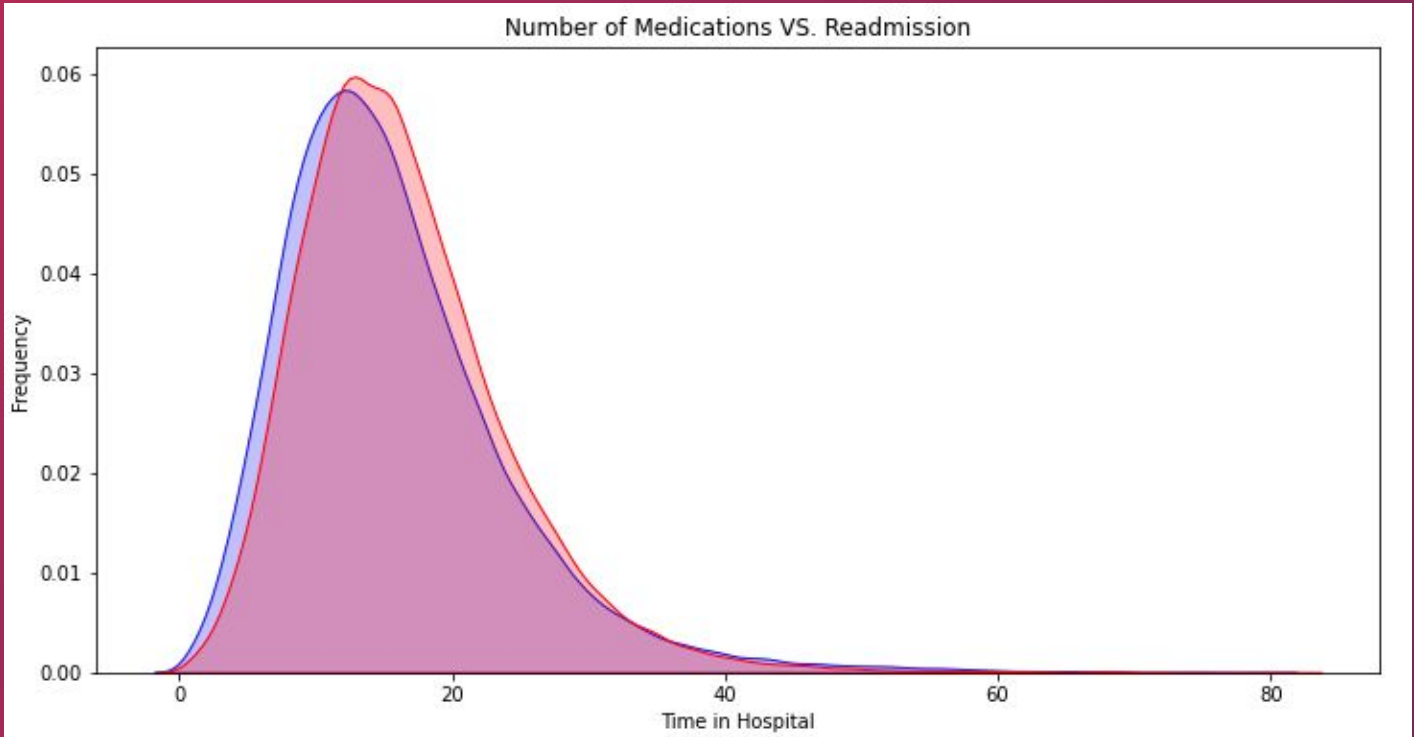
- Correlations
  - Number of impatient (0.21)
  - Number of diagnosis (0.11)

- Outliers
  - Number of Lab Procedures
  - Number of Medication

Number of Lab Procedures



Number of Medications



MODELING

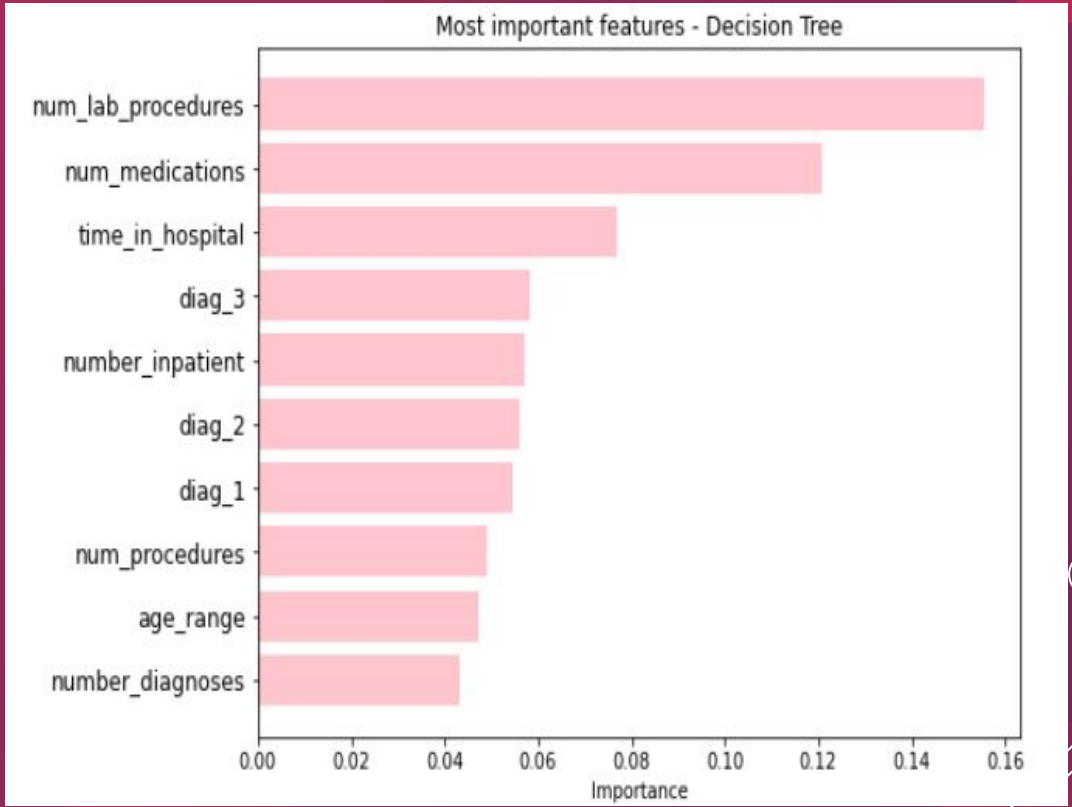
Logistic Regression, KNN, Random Forest Classifier

BASIC MODEL RESULTS

- Random Forest Classifier
- **Training Score:** 0.9999
- **Testing Score:** 0.6207
- **Accuracy:** 0.6194
- **Recall Score:** 0.5134

MOST IMPORTANT FEATURES

- Decision Tree Classifier
- Most important feature option
- Limited to 10 features out of the 50 in the initial dataset
- Limited 3 classification models to just these 10 features.



MOST IMPORTANT FEATURES

RESULTS

- Logistic Regression model
- **Training Score:** 0.6090
- **Testing Score:** 0.6093
- **Accuracy:** 0.6083
- Lowest recall score

- KNN Model
- **Recall Score:** 0.4990

PRINCIPAL COMPONENT ANALYSIS

RESULTS

- Overall scores were lower than the scores computed by the limited features variation of each model.
- Logistic Regression function, sensitivity score of 0.5414

GRID SEARCH FINE TUNING

RESULTS

- Produced the best scores
- Random Forest model
- **Training Score:** 0.9999
- **Testing Score:** 0.6218
- **Accuracy:** 0.6181
- **Recall Score:** 0.5151

CONCLUSION

CONCLUSION

- Ten major features are found to have high impact on diabetes patient readmission.
- Although not the best scores, still beneficial for medical practitioners to pay attention to these features
- Using Grid search for each of our classification models produced the best accuracy and sensitivity scores.
- The best model, Random Forest Classifier, provided an accuracy score of 0.62 and a sensitivity score of 0.52.
- Attempt to use other models moving forward.
- Use more current data.
- Including new data such as family history may be helpful in increasing primary diagnosis rates and effectively decrease readmission rates.
- Given the current COVID pandemic, I'd be interested in exploring how these past 2 years have affected readmission rates.

THANKS

Do you have any questions?

akamara@xyz.com
+1 240 333 4456
xyzconsulting.com