Final Project

Implementing a Model to Predict Hospital Readmission from Diabetes Diagnosis

Brianna L. Palmisano, Jianna J. Estevez & Julianna M. Lomonte

St. John's University, The Peter J. Tobin College of Business

BUA 611/3311: Machine Learning for Business

Spring 2024

Summary

- Introduction to the data set and motivation to create machine learning models
- Overview of data including its source and descriptions of the variables
- Pre-processing/munging plan
- Some preliminary results and initial visualizations

Predicting Hospital Readmissions Dataset

- Contains historical, patient data for 10 years
- Contains both categorical and numerical variables
- These variables consist of several measures of diabetes diagnosis (e.g. glucose test or diabetes medication) used to predict hospital readmission

Objectives and Motivation

- **Objective: to showcase whether the diabetes diagnosis measures predict hospital readmission using X models.**
  - **Choosing the best combination of measures of diabetes diagnosis that will result in readmission**
- **Motivation: to alleviate hospital operation costs**
  - **Using 10 years of data will provide us with accurate predictions**
  - **Our goal is to help healthcare businesses allocate resources strategically**

Descriptions of variables

This data consists of both categorical and numerical variables. **The categorical variables were encoded using both regular and one-hot into binary vectors to be included as input features in the model.**

The variables encoded using regular encoding require that their data is given a **hierarchy in terms of how important** that result is to the objective of our model.

In one-hot encoding, each categorical variable is given a column, and the data elements are either hot or cold (0 or 1), a binary vector. One variable is 1 and the rest are 0. This helps us avoid imposing ordinal relationships on categorical variables. Rather, where the **categorical variables don't have a meaningful numerical relationship, they are given one.**

Categorical variables

"medical_specialty" - the specialty of the admitting physician

"age" - age bracket of the patient

"diag_1" - primary diagnosis (Circulatory, Respiratory, Digestive, etc.)

"diag_2" - secondary diagnosis

"diag_3" - additional secondary diagnosis

"glucose_test" - whether the glucose serum came out as high (> 200), normal, or not performed (ordered in importance)

"change" - whether there was a change in the diabetes medication ('yes' or 'no')

"diabetes_med" - whether a diabetes medication was prescribed ('yes' or 'no')

"A1Ctest" - whether the A1C level of the patient came out as high (> 7%), normal, or not performed (ordered in importance)

"readmitted" - if the patient was readmitted at the hospital ('yes' or 'no')

Numerical variables

"time_in_hospital" - days (from 1 to 14)

"n_procedures" - number of procedures performed during the hospital stay

"n_lab_procedures" - number of laboratory procedures performed during the hospital stay

"n_medications" - number of medications administered during the hospital stay

"n_outpatient" - number of outpatient visits in the year before a hospital stay

"n_inpatient" - number of inpatient visits in the year before the hospital stay

"n_emergency" - number of visits to the emergency room in the year before the hospital stay

Preprocessing and Data Munging

medical_specialty
Missing
Other
Missing
Missing
InternalMedicine
...
Missing

```
#dropping bad data
Hospital_df_1=Hospital_df.drop('medical_specialty', axis=1)
Hospital_df_1
```

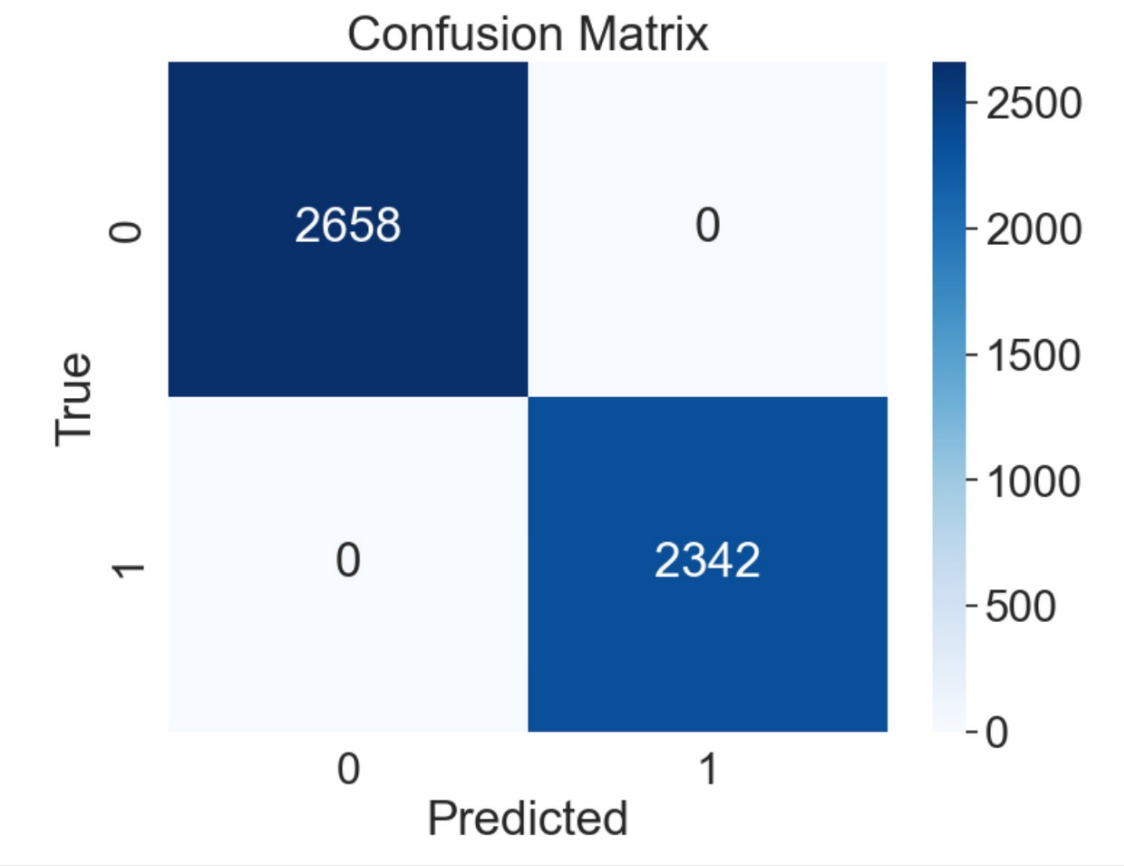
```
#Dropping Missing Values:
Hospital_df_2 = Hospital_df_1 [pd.notna(Hospital_df['diag_1'])]
Hospital_df_3 = Hospital_df_2 [pd.notna(Hospital_df['diag_2'])]
Hospital_df_4 = Hospital_df_3 [pd.notna(Hospital_df['diag_3'])]
Hospital_df_4
```

time_in_hospital	n_lab_procedures	n_procedures	average_time_in_hospital	average_#_of_procedures
8	72	1	4.45332	36.5
3	34	2	4.45332	18.0
5	45	0	4.45332	22.5
2	36	0	4.45332	18.0
1	42	0	4.45332	21.0
...	...	...	...	...
14	77	1	4.45332	39.0
2	66	0	4.45332	33.0
5	12	0	4.45332	6.0
2	61	3	4.45332	32.0
10	37	1	4.45332	19.0

Preprocessing and Data Munging cont.

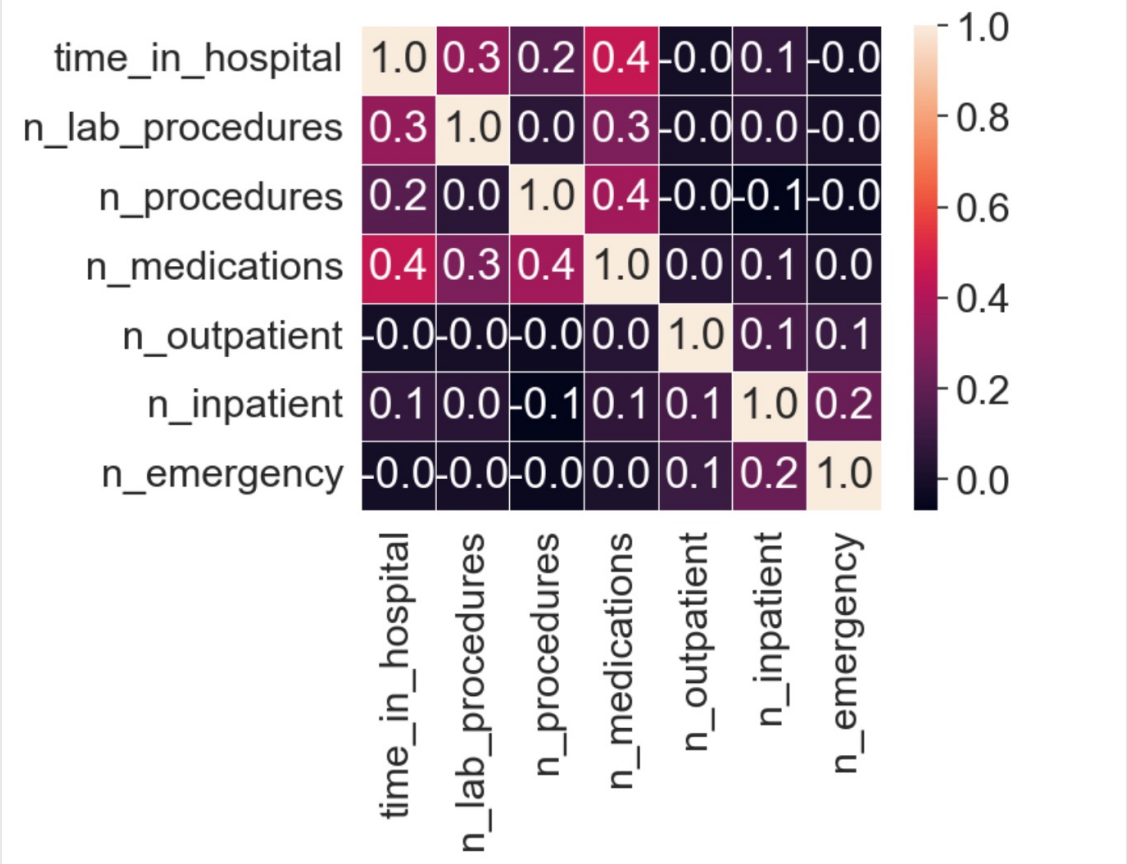
Preliminary results & initial visualizations

Regression 1: Initial Results lead to overfitting



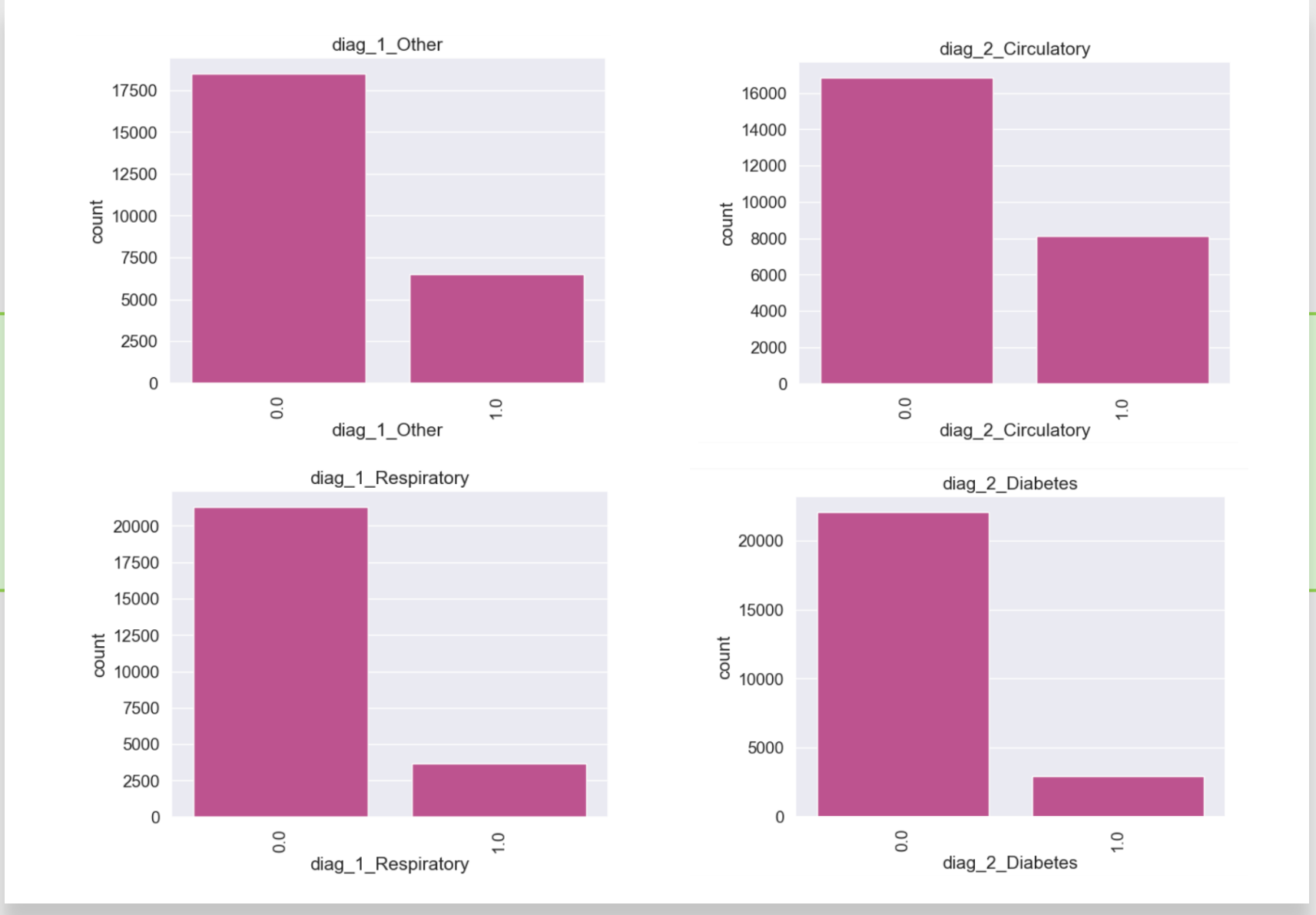

```
Accuracy: 1.00
Precision: 1.00
Recall: 1.00
F1 Score: 1.00
```

Preliminary results & initial visualizations



Correlation Matrix of numerical values

Top Categorical Features



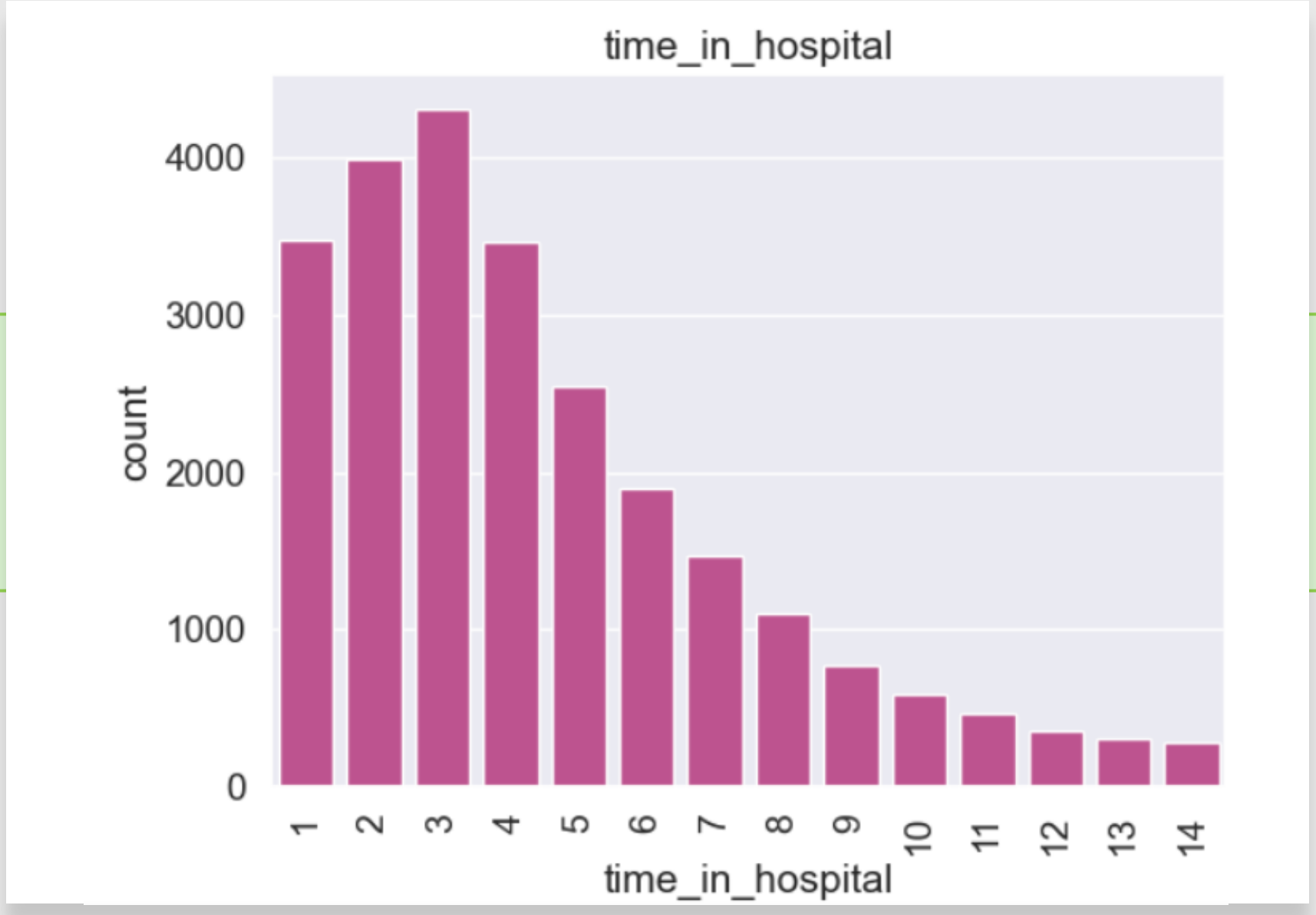
Preliminary results & initial visualizations

When looking at the time patients stay in the hospital (days) compared to other features of importance, more people are leaving after about a 3-day stay than the latter.

This data leads us to predict that the number of days you spend in the hospital has an impact or relationship with readmission rate. The other independent variables also increase in correlation overall, due to this relationship.

As the time variable increases, our other features of importance such as number procedures, lab procedures and medications objectively increases as well.

Top Numerical Feature



Thank you!