Brian_Reppeto_DSC550_Week_6

April 21, 2024

0.0.1 DSC 550 Week :

Activity 6.2

Author: Brian Reppeto 4/19/2024

0.0.2 Milestone 1: Predicting Patient Readmissions for Diabetic Patients

Project Narrative

Hospital readmissions within 30 days post-discharge are not only a significant financial burden on the healthcare system but also often reflect suboptimal patient outcomes and potentially preventable complications. This project addresses the critical challenge of predicting such readmissions among diabetic patients, a group particularly prone to frequent and costly hospitalizations.

The Problem

The specific problem this project targets is the prediction of unplanned readmissions within 30 days among patients diagnosed with diabetes. These readmissions may be due to a variety of factors including inadequate management of diabetes, complications arising from the condition, or insufficient patient education and follow-up care upon discharge. The goal is to provide a predictive tool that can identify at-risk patients before they leave the hospital. This tool will enable healthcare providers to initiate targeted interventions such as personalized discharge planning, enhanced patient education, and tailored follow-up care schedules.

Objectives

The primary objective of this project is to develop a predictive model that uses historical hospital data to forecast the likelihood of a diabetic patient being readmitted within 30 days of discharge. The insights gained from this model will assist healthcare professionals in making informed decisions about patient care strategies and resource allocation.

Data Utilization

The project utilizes data from the Diabetes 130-US hospitals dataset, which comprises information from over 100,000 hospital admissions from 1999 to 2008 and across 130 US hospitals. The dataset includes diverse variables such as patient demographics, admission and discharge statuses, diagnostic codes, number of inpatient visits, and medication details. This rich dataset provides a comprehensive foundation to explore and model the complexities associated with readmissions.

Potential Impact

By accurately predicting readmissions, the model can directly influence the development of personalized medicine approaches and proactive healthcare strategies. Hospitals can use these predictions to reduce readmission rates, thereby decreasing the penalties they face under healthcare regulations like the Hospital Readmissions Reduction Program (HRRP). Moreover, patients benefit from improved healthcare experiences and outcomes, contributing to overall patient satisfaction and health system sustainability.

In conclusion, this project aims to harness the power of machine learning and predictive analytics to tackle a pressing healthcare issue. By doing so, it not only addresses an immediate business need for hospitals but also plays a crucial role in advancing how data-driven strategies can be implemented in clinical settings to enhance patient care.

[82]:
```
# import libraries

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler, LabelEncoder
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import confusion_matrix, classification_report,␣
 ↪accuracy_score
```

[83]:
```
# load the dataset

data = pd.read_csv('diabetic_data.csv')
```

[84]:
```
# data shape

data.shape
```

[84]:
```
(101766, 50)
```

[85]:
```
# head the data

data.head(15)
```

[85]:
```
   encounter_id  patient_nbr             race  gender      age weight  \
0       2278392      8222157        Caucasian  Female   [0-10)      ?
1        149190     55629189        Caucasian  Female  [10-20)      ?
2         64410     86047875  AfricanAmerican  Female  [20-30)      ?
3        500364     82442376        Caucasian    Male  [30-40)      ?
4         16680     42519267        Caucasian    Male  [40-50)      ?
5         35754     82637451        Caucasian    Male  [50-60)      ?
6         55842     84259809        Caucasian    Male  [60-70)      ?
7         63768    114882984        Caucasian    Male  [70-80)      ?
8         12522     48330783        Caucasian  Female  [80-90)      ?
```

```
9          15738     63555939         Caucasian  Female  [90-100)      ?
10         28236     89869032  AfricanAmerican  Female   [40-50)      ?
11         36900     77391171  AfricanAmerican    Male   [60-70)      ?
12         40926     85504905         Caucasian  Female   [40-50)      ?
13         42570     77586282         Caucasian    Male   [80-90)      ?
14         62256     49726791  AfricanAmerican  Female   [60-70)      ?

    admission_type_id  discharge_disposition_id  admission_source_id  \
0                   6                        25                    1
1                   1                         1                    7
2                   1                         1                    7
3                   1                         1                    7
4                   1                         1                    7
5                   2                         1                    2
6                   3                         1                    2
7                   1                         1                    7
8                   2                         1                    4
9                   3                         3                    4
10                  1                         1                    7
11                  2                         1                    4
12                  1                         3                    7
13                  1                         6                    7
14                  3                         1                    2

    time_in_hospital  … citoglipton insulin  glyburide-metformin  \
0                  1  …          No      No                   No
1                  3  …          No      Up                   No
2                  2  …          No      No                   No
3                  2  …          No      Up                   No
4                  1  …          No  Steady                   No
5                  3  …          No  Steady                   No
6                  4  …          No  Steady                   No
7                  5  …          No      No                   No
8                 13  …          No  Steady                   No
9                 12  …          No  Steady                   No
10                 9  …          No  Steady                   No
11                 7  …          No  Steady                   No
12                 7  …          No    Down                   No
13                10  …          No  Steady                   No
14                 1  …          No  Steady                   No

    glipizide-metformin  glimepiride-pioglitazone  metformin-rosiglitazone  \
0                    No                        No                       No
1                    No                        No                       No
2                    No                        No                       No
3                    No                        No                       No
4                    No                        No                       No
```

```
5                     No                         No                        No
6                     No                         No                        No
7                     No                         No                        No
8                     No                         No                        No
9                     No                         No                        No
10                    No                         No                        No
11                    No                         No                        No
12                    No                         No                        No
13                    No                         No                        No
14                    No                         No                        No

    metformin-pioglitazone  change diabetesMed readmitted
0                       No      No          No         NO
1                       No      Ch         Yes        >30
2                       No      No         Yes         NO
3                       No      Ch         Yes         NO
4                       No      Ch         Yes         NO
5                       No      No         Yes        >30
6                       No      Ch         Yes         NO
7                       No      No         Yes        >30
8                       No      Ch         Yes         NO
9                       No      Ch         Yes         NO
10                      No      No         Yes        >30
11                      No      Ch         Yes        <30
12                      No      Ch         Yes        <30
13                      No      No         Yes         NO
14                      No      No         Yes        >30

[15 rows x 50 columns]
```

0.0.3 Data Exploration and Cleaning

[86]:
```python
# explore missing values and clean data

data = data.replace('?', np.nan)  # replace '?' with NaN for clarity
missing_data = data.isnull().sum()/len(data) * 100
print("Percentage of missing data per column:\n", missing_data[missing_data >␣
 ↪0])
```

```
Percentage of missing data per column:
 race                     2.233555
weight                96.858479
payer_code            39.557416
medical_specialty     49.082208
diag_1                 0.020636
diag_2                 0.351787
diag_3                 1.398306
max_glu_serum         94.746772
```

```
A1Cresult              83.277322
dtype: float64
```

[87]:
```
# drop columns with high percentage of missing values and those not relevant␣
↪for the analysis

columns_to_drop = ['weight', 'medical_specialty', 'payer_code']
data.drop(columns=columns_to_drop, inplace=True)
```

[89]:
```
# head the data

data.head(15)
```

[89]:
```
    encounter_id  patient_nbr             race  gender       age  \
0        2278392      8222157        Caucasian  Female    [0-10)
1         149190     55629189        Caucasian  Female   [10-20)
2          64410     86047875  AfricanAmerican  Female   [20-30)
3         500364     82442376        Caucasian    Male   [30-40)
4          16680     42519267        Caucasian    Male   [40-50)
5          35754     82637451        Caucasian    Male   [50-60)
6          55842     84259809        Caucasian    Male   [60-70)
7          63768    114882984        Caucasian    Male   [70-80)
8          12522     48330783        Caucasian  Female   [80-90)
9          15738     63555939        Caucasian  Female  [90-100)
10         28236     89869032  AfricanAmerican  Female   [40-50)
11         36900     77391171  AfricanAmerican    Male   [60-70)
12         40926     85504905        Caucasian  Female   [40-50)
13         42570     77586282        Caucasian    Male   [80-90)
14         62256     49726791  AfricanAmerican  Female   [60-70)

    admission_type_id  discharge_disposition_id  admission_source_id  \
0                   6                        25                    1
1                   1                         1                    7
2                   1                         1                    7
3                   1                         1                    7
4                   1                         1                    7
5                   2                         1                    2
6                   3                         1                    2
7                   1                         1                    7
8                   2                         1                    4
9                   3                         3                    4
10                  1                         1                    7
11                  2                         1                    4
12                  1                         3                    7
13                  1                         6                    7
14                  3                         1                    2
```

	time_in_hospital	num_lab_procedures	…	citoglipton	insulin	\
0	1	41	…	No	No	
1	3	59	…	No	Up	
2	2	11	…	No	No	
3	2	44	…	No	Up	
4	1	51	…	No	Steady	
5	3	31	…	No	Steady	
6	4	70	…	No	Steady	
7	5	73	…	No	No	
8	13	68	…	No	Steady	
9	12	33	…	No	Steady	
10	9	47	…	No	Steady	
11	7	62	…	No	Steady	
12	7	60	…	No	Down	
13	10	55	…	No	Steady	
14	1	49	…	No	Steady	

	glyburide-metformin	glipizide-metformin	glimepiride-pioglitazone	\
0	No	No	No	
1	No	No	No	
2	No	No	No	
3	No	No	No	
4	No	No	No	
5	No	No	No	
6	No	No	No	
7	No	No	No	
8	No	No	No	
9	No	No	No	
10	No	No	No	
11	No	No	No	
12	No	No	No	
13	No	No	No	
14	No	No	No	

	metformin-rosiglitazone	metformin-pioglitazone	change	diabetesMed	\
0	No	No	No	No	
1	No	No	Ch	Yes	
2	No	No	No	Yes	
3	No	No	Ch	Yes	
4	No	No	Ch	Yes	
5	No	No	No	Yes	
6	No	No	Ch	Yes	
7	No	No	No	Yes	
8	No	No	Ch	Yes	
9	No	No	Ch	Yes	
10	No	No	No	Yes	
11	No	No	Ch	Yes	

```
12                        No                      No      Ch            Yes
13                        No                      No      No            Yes
14                        No                      No      No            Yes

   readmitted
0          NO
1         >30
2          NO
3          NO
4          NO
5         >30
6          NO
7         >30
8          NO
9          NO
10        >30
11        <30
12        <30
13         NO
14        >30

[15 rows x 47 columns]
```

[90]:

```
# summary of cleaned data

print(data.info())
```

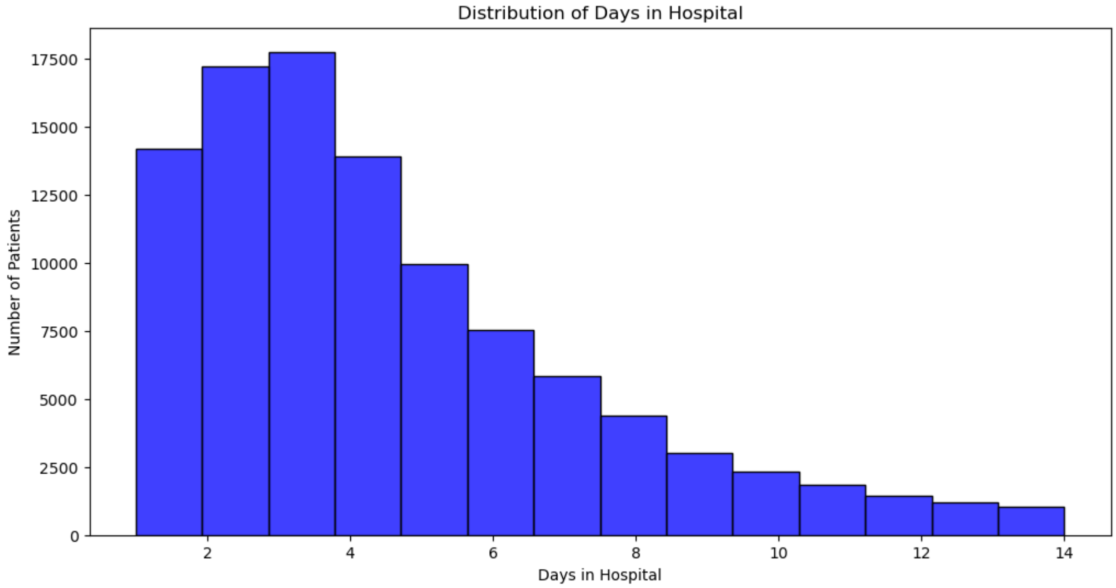
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 101766 entries, 0 to 101765
Data columns (total 47 columns):
 #   Column                    Non-Null Count   Dtype
---  ------                    --------------   -----
 0   encounter_id              101766 non-null  int64
 1   patient_nbr               101766 non-null  int64
 2   race                      99493 non-null   object
 3   gender                    101766 non-null  object
 4   age                       101766 non-null  object
 5   admission_type_id         101766 non-null  int64
 6   discharge_disposition_id  101766 non-null  int64
 7   admission_source_id       101766 non-null  int64
 8   time_in_hospital          101766 non-null  int64
 9   num_lab_procedures        101766 non-null  int64
 10  num_procedures            101766 non-null  int64
 11  num_medications           101766 non-null  int64
 12  number_outpatient         101766 non-null  int64
 13  number_emergency          101766 non-null  int64
 14  number_inpatient          101766 non-null  int64
 15  diag_1                    101745 non-null  object
```

```
 16  diag_2                    101408 non-null  object
 17  diag_3                    100343 non-null  object
 18  number_diagnoses          101766 non-null  int64
 19  max_glu_serum             5346 non-null    object
 20  A1Cresult                 17018 non-null   object
 21  metformin                 101766 non-null  object
 22  repaglinide               101766 non-null  object
 23  nateglinide               101766 non-null  object
 24  chlorpropamide            101766 non-null  object
 25  glimepiride               101766 non-null  object
 26  acetohexamide             101766 non-null  object
 27  glipizide                 101766 non-null  object
 28  glyburide                 101766 non-null  object
 29  tolbutamide               101766 non-null  object
 30  pioglitazone              101766 non-null  object
 31  rosiglitazone             101766 non-null  object
 32  acarbose                  101766 non-null  object
 33  miglitol                  101766 non-null  object
 34  troglitazone              101766 non-null  object
 35  tolazamide                101766 non-null  object
 36  examide                   101766 non-null  object
 37  citoglipton               101766 non-null  object
 38  insulin                   101766 non-null  object
 39  glyburide-metformin       101766 non-null  object
 40  glipizide-metformin       101766 non-null  object
 41  glimepiride-pioglitazone  101766 non-null  object
 42  metformin-rosiglitazone   101766 non-null  object
 43  metformin-pioglitazone    101766 non-null  object
 44  change                    101766 non-null  object
 45  diabetesMed               101766 non-null  object
 46  readmitted                101766 non-null  object
dtypes: int64(13), object(34)
memory usage: 36.5+ MB
None
```

0.0.4 Graphical Analysis

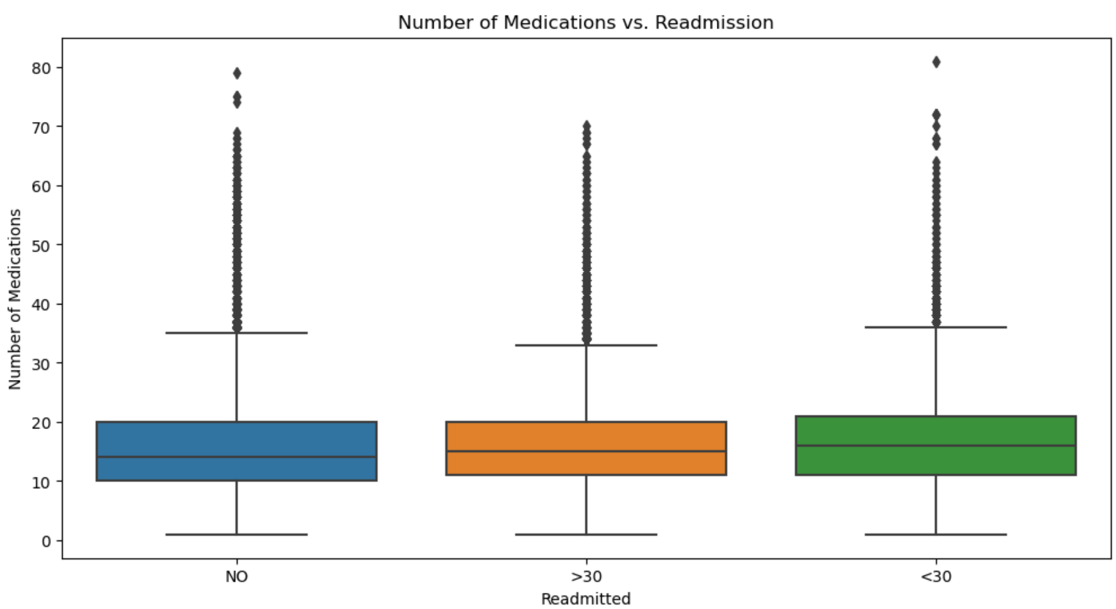
[91]:
```python
# histogram of Distribution of the number of days in hospital

plt.figure(figsize=(12, 6))
sns.histplot(data['time_in_hospital'], bins=14, kde=False, color='blue')
plt.title('Distribution of Days in Hospital')
plt.xlabel('Days in Hospital')
plt.ylabel('Number of Patients')
plt.show()
```



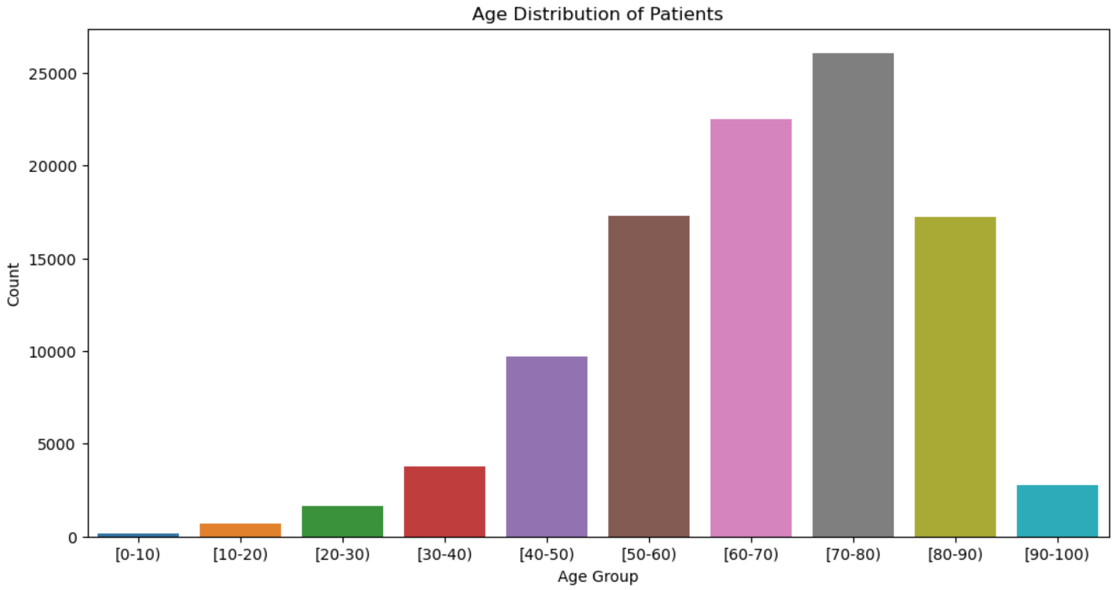
[92]:
```
# number of medications vs. readmissions

plt.figure(figsize=(12, 6))
sns.boxplot(x='readmitted', y='num_medications', data=data)
plt.title('Number of Medications vs. Readmission')
plt.xlabel('Readmitted')
plt.ylabel('Number of Medications')
plt.show()
```



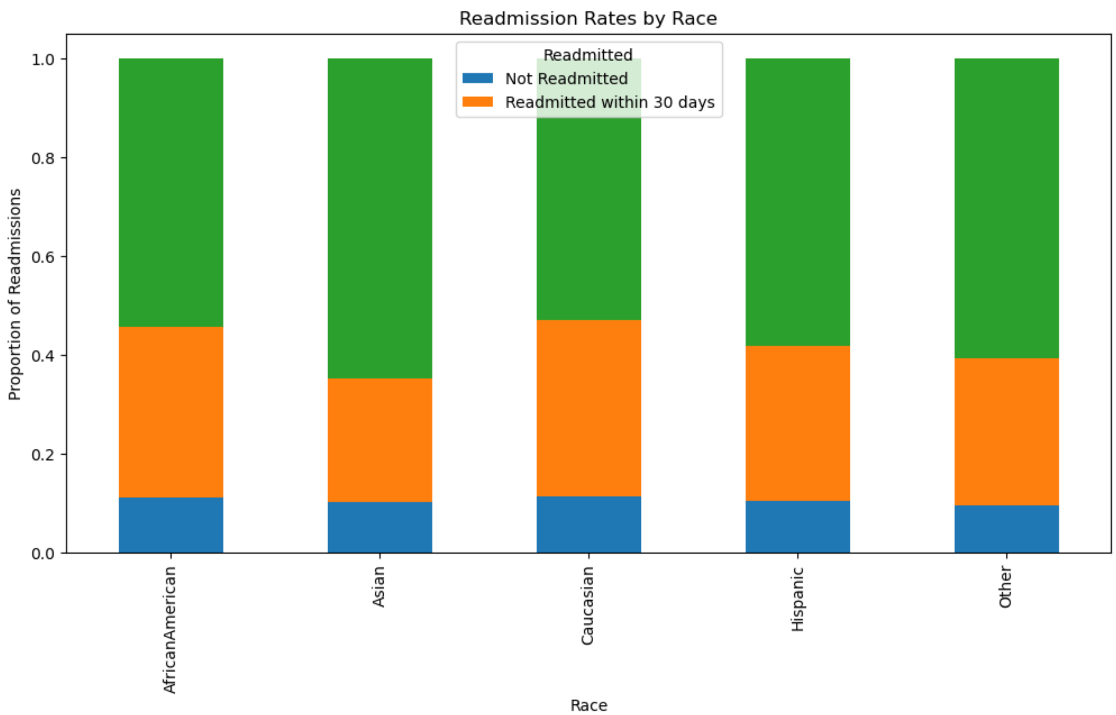
[93]:
```python
# age distribution of the patients

plt.figure(figsize=(12, 6))
sns.countplot(x='age', data=data, order=sorted(data['age'].unique()))
plt.title('Age Distribution of Patients')
plt.xlabel('Age Group')
plt.ylabel('Count')
plt.show()
```



[94]:
```python
# readmission rates by race

readmission_by_race = data.groupby('race')['readmitted'].
    ↪value_counts(normalize=True).unstack().fillna(0)
readmission_by_race.plot(kind='bar', stacked=True, figsize=(12, 6))
plt.title('Readmission Rates by Race')
plt.xlabel('Race')
plt.ylabel('Proportion of Readmissions')
plt.legend(title='Readmitted', labels=['Not Readmitted', 'Readmitted within 30␣
    ↪days'])
plt.show()
```



0.0.5 Analysis of Graphs

Distribution of Days in Hospital:

The histogram shows that the most common duration of hospital stays is between 2 to 4 days. The distribution is right-skewed, indicating that longer stays are less frequent but not uncommon. This suggests that most diabetic patients have relatively short hospital stays, but a subset requires extended care.

Number of Medications vs. Readmission:

From the boxplot comparing the number of medications between readmitted and not readmitted groups, there is a noticeable overlap, but it seems that patients who were readmitted tend to be on slightly more medications. This could imply that patients with more complex medication schedules are at a higher risk of readmission, possibly due to more severe underlying conditions.

Age Distribution of Patients:

The age distribution shows that the majority of the patients fall into the 60-80 age range, with fewer younger patients. This is typical for diabetic cohorts where prevalence increases with age.

Readmission Rates by Race:

The bar chart demonstrates that readmission rates vary somewhat by race. The proportions show that certain racial groups might have higher or lower rates of readmission, which could be important for targeted interventions or understanding disparities in healthcare outcomes.

0.0.6 Conclusion

The graphical analysis provided insights into factors that might influence hospital readmission among diabetic patients. The analysis suggests that duration of hospital stay, complexity of medication regimens, patient age, and race could be significant predictors of readmission risk.

[ ]:
```

```