

Brian_Reppeto_DSC550_Week_8

May 5, 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.

```
[153]: # 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
```

```
[154]: # load the dataset

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

```
[155]: # data shape

data.shape
```

```
[155]: (101766, 50)
```

```
[156]: # head the data

data.head(15)
```

```
[156]:
```

	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

```
[157]: # 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
```

```
[158]: # 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)
```

```
[159]: # head the data
```

```
data.head(15)
```

```
[159]:
```

	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]

[160]: *# 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

```

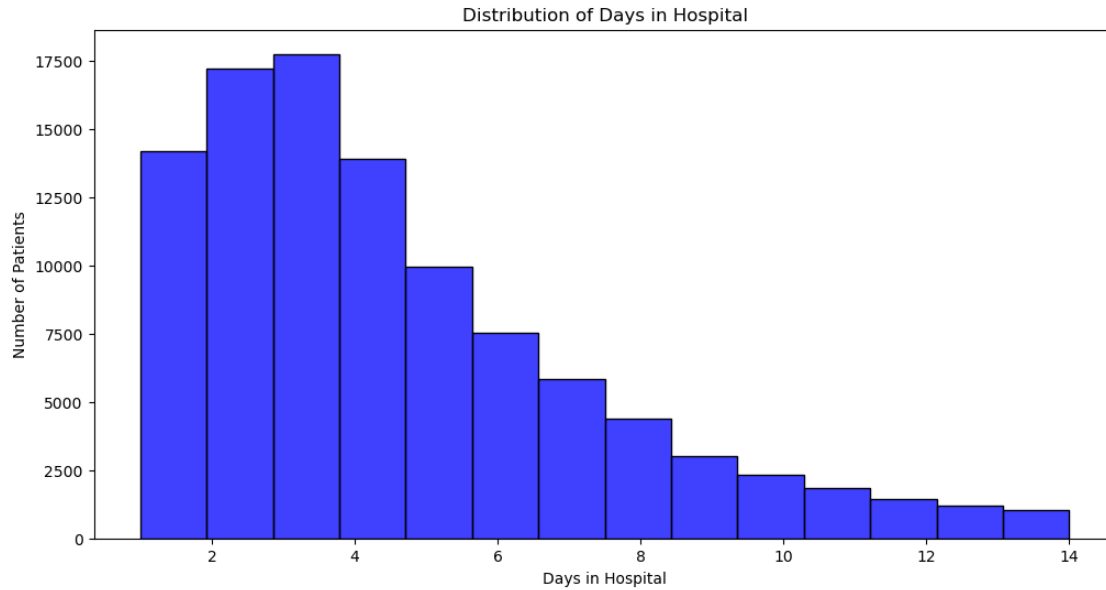
0.0.4 Graphical Analysis

```

[161]: # histogram of Distribution of the number of days in hospital

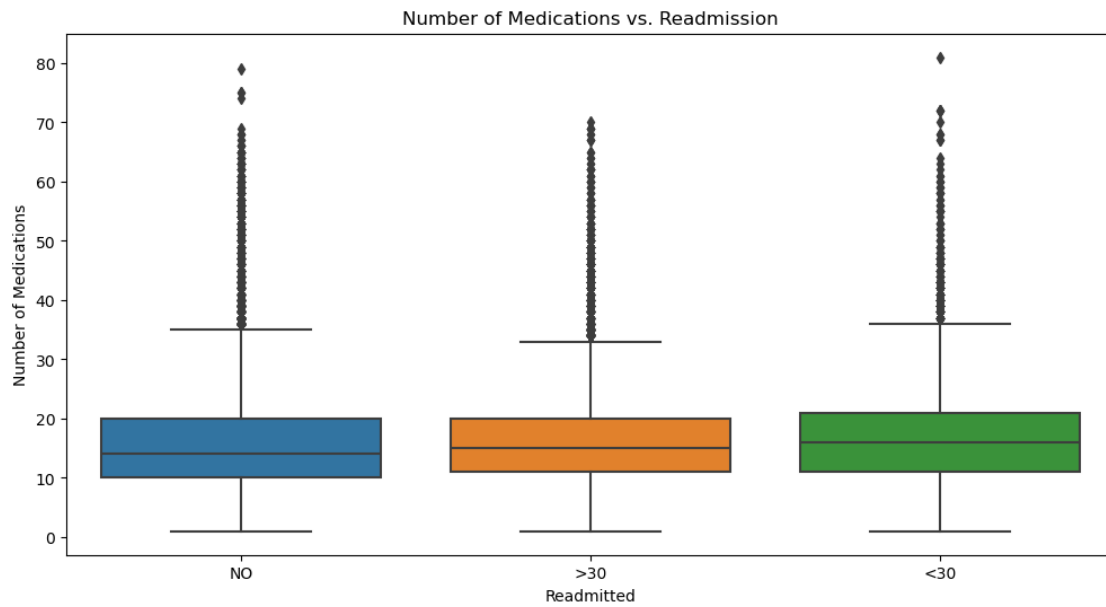
plt.figure(figsize=(12, 6))
sns.histplot(data['time_in_hospital'], bins=14, kde=False, color='blue')
plt.title('Distribution of Days in Hospital')
plt.xlabel('Days in Hospital')
plt.ylabel('Number of Patients')
plt.show()

```

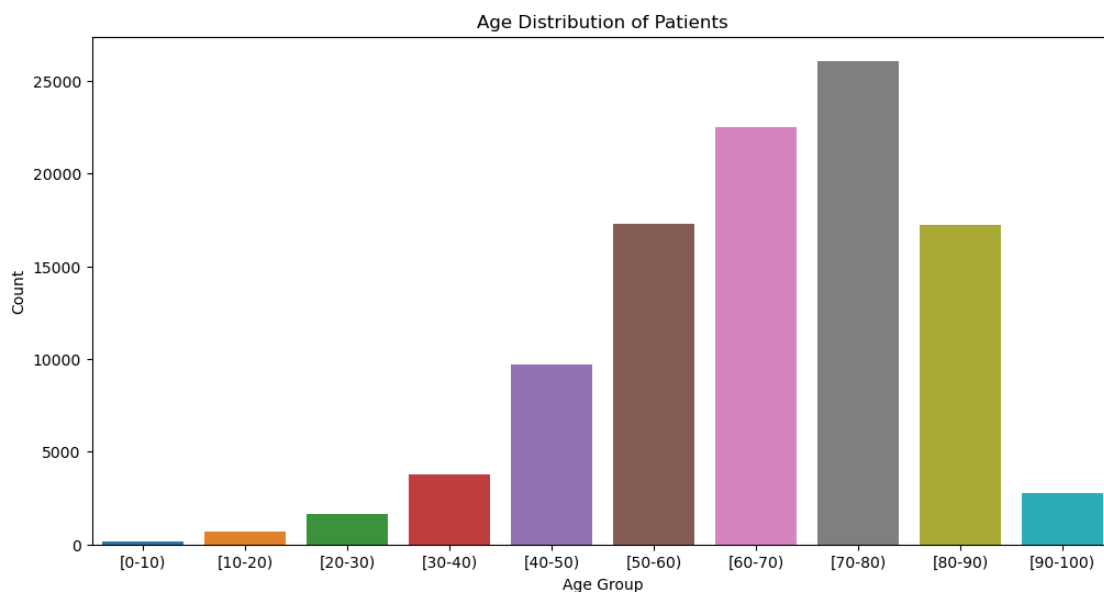
```
[162]: # number of medications vs. readmissions
```

```
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()
```



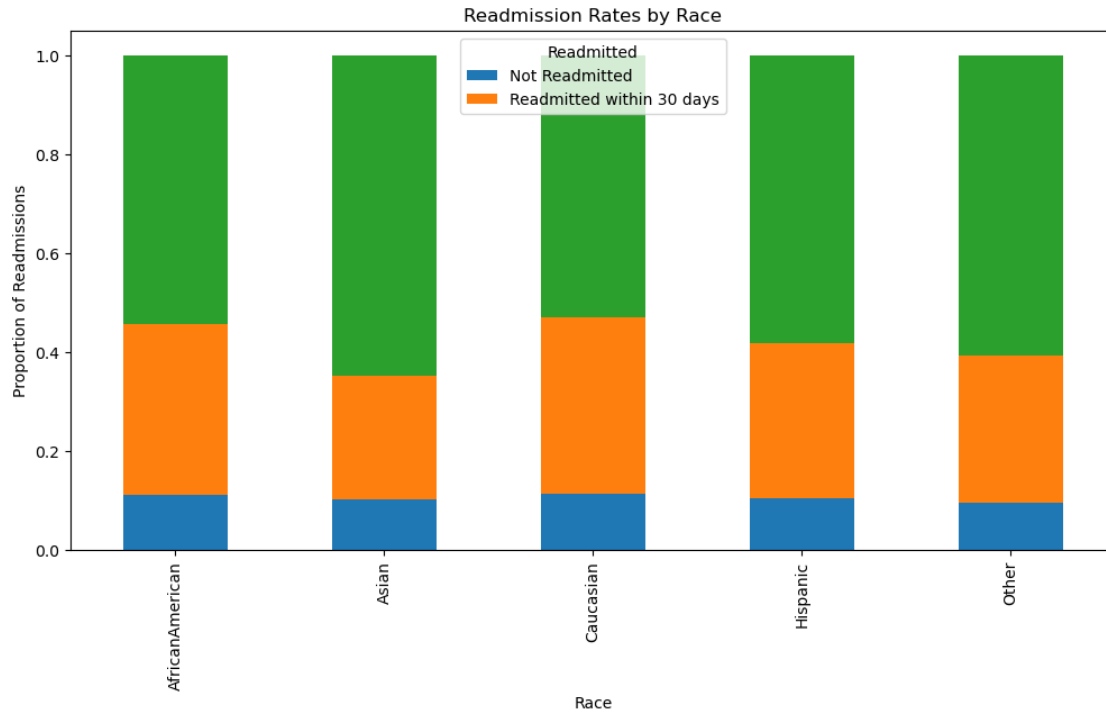
```
[163]: # age distribution of the patients
```

```
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()
```



```
[164]: # 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.

0.0.7 DSC 550 Week :

Activity 8.2 Term Project Milestone 2

Author: Brian Reppeto 5/2/2024

```
[165]: # 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

[166]: # drop columns with high percentage of missing values and those not relevant
    # for the analysis
    # Weight Medical Spec and Payer code were dropped above

columns_to_drop = ['encounter_id', 'patient_nbr']
data.drop(columns=columns_to_drop, inplace=True)
```

0.0.8 Dropped Features

Encounter ID and Patient Number: These are identifiers for hospital stays and patients, respectively. They are unique to each record and do not provide predictive value for readmission.

Weight: Since the missing data for weight is over 95%, it is not useful due to the lack of sufficient data.

Payer Code & Medical Specialty: Have a high proportion of missing values, relevance to readmission might be limited compared to clinical features.

```
[167]: # drop patients that can not be readmitted
    # the discharge disposition is expired or hospice and not readmittable

ids_to_remove = [11, 13, 14, 19, 20, 21]
data = data[~data['discharge_disposition_id'].isin(ids_to_remove)]

[168]: # verify the filtering
```

```
print(data['discharge_disposition_id'].unique())
```

```
[25  1  3  6  2  5  7 10  4 18  8 12 16 17 22 23  9 15 24 28 27]
```

```
[169]: # head the data
```

```
data.head(15)
```

```
[169]:
```

	race	gender	age	admission_type_id	\
0	Caucasian	Female	[0-10)	6	
1	Caucasian	Female	[10-20)	1	
2	AfricanAmerican	Female	[20-30)	1	
3	Caucasian	Male	[30-40)	1	
4	Caucasian	Male	[40-50)	1	
5	Caucasian	Male	[50-60)	2	
6	Caucasian	Male	[60-70)	3	
7	Caucasian	Male	[70-80)	1	
8	Caucasian	Female	[80-90)	2	
9	Caucasian	Female	[90-100)	3	
10	AfricanAmerican	Female	[40-50)	1	
11	AfricanAmerican	Male	[60-70)	2	
12	Caucasian	Female	[40-50)	1	
13	Caucasian	Male	[80-90)	1	
14	AfricanAmerican	Female	[60-70)	3	

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

	num_lab_procedures	num_procedures	num_medications	...	citoglipton	\
0	41	0	1	...	No	
1	59	0	18	...	No	
2	11	5	13	...	No	
3	44	1	16	...	No	
4	51	0	8	...	No	

5	31	6	16 ...	No
6	70	1	21 ...	No
7	73	0	12 ...	No
8	68	2	28 ...	No
9	33	3	18 ...	No
10	47	2	17 ...	No
11	62	0	11 ...	No
12	60	0	15 ...	No
13	55	1	31 ...	No
14	49	5	2 ...	No

	insulin	glyburide-metformin	glipizide-metformin	glimepiride-pioglitazone	\
0	No	No	No	No	
1	Up	No	No	No	
2	No	No	No	No	
3	Up	No	No	No	
4	Steady	No	No	No	
5	Steady	No	No	No	
6	Steady	No	No	No	
7	No	No	No	No	
8	Steady	No	No	No	
9	Steady	No	No	No	
10	Steady	No	No	No	
11	Steady	No	No	No	
12	Down	No	No	No	
13	Steady	No	No	No	
14	Steady	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 45 columns]

```

[171]: # missing values for 'race' with the most common category

data['race'] = data['race'].fillna(data['race'].mode()[0])

# for numerical columns fill missing values with the median

for col in data.select_dtypes(include=['int64', 'float64']).columns:
    data[col] = data[col].fillna(data[col].median())

```

```

[172]: # tail the data

data.tail(15)

```

```

[172]:
      race  gender  age  admission_type_id  \
101751  Caucasian  Male  [70-80)           3
101752    Other  Female  [40-50)           3
101753    Other   Male  [40-50)           1
101754  Caucasian  Female  [70-80)           1
101755    Other  Female  [40-50)           1
101756    Other  Female  [60-70)           1
101757  Caucasian  Female  [70-80)           1
101758  Caucasian  Female  [80-90)           1
101759  Caucasian   Male  [80-90)           1
101760 AfricanAmerican  Female  [60-70)           1
101761 AfricanAmerican   Male  [70-80)           1
101762 AfricanAmerican  Female  [80-90)           1
101763    Caucasian   Male  [70-80)           1
101764    Caucasian  Female  [80-90)           2
101765    Caucasian   Male  [70-80)           1

```

	discharge_disposition_id	admission_source_id	time_in_hospital	\
101751	6	1	13	
101752	1	1	3	
101753	1	7	13	
101754	1	7	9	
101755	1	7	14	
101756	1	7	2	
101757	1	7	5	
101758	1	7	5	
101759	1	7	1	
101760	1	7	6	
101761	3	7	3	
101762	4	5	5	
101763	1	7	1	
101764	3	7	10	
101765	1	7	6	

	num_lab_procedures	num_procedures	num_medications	...	citoglipton	\
101751	77	6	65	...	No	
101752	13	1	5	...	No	
101753	51	2	13	...	No	
101754	50	2	33	...	No	
101755	73	6	26	...	No	
101756	46	6	17	...	No	
101757	21	1	16	...	No	
101758	76	1	22	...	No	
101759	1	0	15	...	No	
101760	45	1	25	...	No	
101761	51	0	16	...	No	
101762	33	3	18	...	No	
101763	53	0	9	...	No	
101764	45	2	21	...	No	
101765	13	3	3	...	No	

	insulin	glyburide-metformin	glipizide-metformin	\
101751	Up	No	No	
101752	Steady	No	No	
101753	Down	No	No	
101754	Steady	No	No	
101755	Up	No	No	
101756	Steady	No	No	
101757	Steady	No	No	
101758	Up	No	No	
101759	Up	No	No	
101760	Down	No	No	
101761	Down	No	No	
101762	Steady	No	No	

101763	Down	No	No
101764	Up	No	No
101765	No	No	No

	glimepiride-pioglitazone	metformin-rosiglitazone	\
101751	No	No	No
101752	No	No	No
101753	No	No	No
101754	No	No	No
101755	No	No	No
101756	No	No	No
101757	No	No	No
101758	No	No	No
101759	No	No	No
101760	No	No	No
101761	No	No	No
101762	No	No	No
101763	No	No	No
101764	No	No	No
101765	No	No	No

	metformin-pioglitazone	change	diabetesMed	readmitted
101751	No	Ch	Yes	NO
101752	No	Ch	Yes	NO
101753	No	Ch	Yes	NO
101754	No	Ch	Yes	>30
101755	No	Ch	Yes	>30
101756	No	No	Yes	>30
101757	No	No	Yes	NO
101758	No	Ch	Yes	NO
101759	No	Ch	Yes	NO
101760	No	Ch	Yes	>30
101761	No	Ch	Yes	>30
101762	No	No	Yes	NO
101763	No	Ch	Yes	NO
101764	No	Ch	Yes	NO
101765	No	No	No	NO

[15 rows x 45 columns]

Explanation: The 'age' feature is converted from ranges to a more usable numerical format representing the mid-point of each range.

```
[173]: # convert 'age' to a numerical average
# the 'age' feature is converted from
# ranges to a more usable numerical format representing the mid-point of each
↳ range.
```

```
data['age'] = data['age'].apply(lambda x: (int(x.split('-')[0][1:]) + int(x.
↳split('-')[1][:-1])) // 2)
```

```
[174]: # tail the data
```

```
data.tail(15)
```

```
[174]:
```

	race	gender	age	admission_type_id	\
101751	Caucasian	Male	75	3	
101752	Other	Female	45	3	
101753	Other	Male	45	1	
101754	Caucasian	Female	75	1	
101755	Other	Female	45	1	
101756	Other	Female	65	1	
101757	Caucasian	Female	75	1	
101758	Caucasian	Female	85	1	
101759	Caucasian	Male	85	1	
101760	AfricanAmerican	Female	65	1	
101761	AfricanAmerican	Male	75	1	
101762	AfricanAmerican	Female	85	1	
101763	Caucasian	Male	75	1	
101764	Caucasian	Female	85	2	
101765	Caucasian	Male	75	1	

	discharge_disposition_id	admission_source_id	time_in_hospital	\
101751	6	1	13	
101752	1	1	3	
101753	1	7	13	
101754	1	7	9	
101755	1	7	14	
101756	1	7	2	
101757	1	7	5	
101758	1	7	5	
101759	1	7	1	
101760	1	7	6	
101761	3	7	3	
101762	4	5	5	
101763	1	7	1	
101764	3	7	10	
101765	1	7	6	

	num_lab_procedures	num_procedures	num_medications	...	citoglipton	\
101751	77	6	65	...	No	
101752	13	1	5	...	No	
101753	51	2	13	...	No	
101754	50	2	33	...	No	

101755	73	6	26 ...	No
101756	46	6	17 ...	No
101757	21	1	16 ...	No
101758	76	1	22 ...	No
101759	1	0	15 ...	No
101760	45	1	25 ...	No
101761	51	0	16 ...	No
101762	33	3	18 ...	No
101763	53	0	9 ...	No
101764	45	2	21 ...	No
101765	13	3	3 ...	No

	insulin	glyburide-metformin	glipizide-metformin	\
101751	Up	No	No	
101752	Steady	No	No	
101753	Down	No	No	
101754	Steady	No	No	
101755	Up	No	No	
101756	Steady	No	No	
101757	Steady	No	No	
101758	Up	No	No	
101759	Up	No	No	
101760	Down	No	No	
101761	Down	No	No	
101762	Steady	No	No	
101763	Down	No	No	
101764	Up	No	No	
101765	No	No	No	

	glimepiride-pioglitazone	metformin-rosiglitazone	\
101751	No	No	
101752	No	No	
101753	No	No	
101754	No	No	
101755	No	No	
101756	No	No	
101757	No	No	
101758	No	No	
101759	No	No	
101760	No	No	
101761	No	No	
101762	No	No	
101763	No	No	
101764	No	No	
101765	No	No	

metformin-pioglitazone change diabetesMed readmitted

101751	No	Ch	Yes	NO
101752	No	Ch	Yes	NO
101753	No	Ch	Yes	NO
101754	No	Ch	Yes	>30
101755	No	Ch	Yes	>30
101756	No	No	Yes	>30
101757	No	No	Yes	NO
101758	No	Ch	Yes	NO
101759	No	Ch	Yes	NO
101760	No	Ch	Yes	>30
101761	No	Ch	Yes	>30
101762	No	No	Yes	NO
101763	No	Ch	Yes	NO
101764	No	Ch	Yes	NO
101765	No	No	No	NO

[15 rows x 45 columns]

```
[175]: # label readmission target as binary for whether readmission occurred within 30
        ↪ days
```

```
data['readmitted'] = data['readmitted'].apply(lambda x: 1 if x == '<30' else 0)
```

```
[176]: # tail the data
```

```
data.tail(15)
```

```
[176]:
```

	race	gender	age	admission_type_id	\
101751	Caucasian	Male	75	3	
101752	Other	Female	45	3	
101753	Other	Male	45	1	
101754	Caucasian	Female	75	1	
101755	Other	Female	45	1	
101756	Other	Female	65	1	
101757	Caucasian	Female	75	1	
101758	Caucasian	Female	85	1	
101759	Caucasian	Male	85	1	
101760	AfricanAmerican	Female	65	1	
101761	AfricanAmerican	Male	75	1	
101762	AfricanAmerican	Female	85	1	
101763	Caucasian	Male	75	1	
101764	Caucasian	Female	85	2	
101765	Caucasian	Male	75	1	

	discharge_disposition_id	admission_source_id	time_in_hospital	\
101751	6	1	13	
101752	1	1	3	

101753	1	7	13
101754	1	7	9
101755	1	7	14
101756	1	7	2
101757	1	7	5
101758	1	7	5
101759	1	7	1
101760	1	7	6
101761	3	7	3
101762	4	5	5
101763	1	7	1
101764	3	7	10
101765	1	7	6

	num_lab_procedures	num_procedures	num_medications	...	citoglipton	\
101751	77	6	65	...	No	
101752	13	1	5	...	No	
101753	51	2	13	...	No	
101754	50	2	33	...	No	
101755	73	6	26	...	No	
101756	46	6	17	...	No	
101757	21	1	16	...	No	
101758	76	1	22	...	No	
101759	1	0	15	...	No	
101760	45	1	25	...	No	
101761	51	0	16	...	No	
101762	33	3	18	...	No	
101763	53	0	9	...	No	
101764	45	2	21	...	No	
101765	13	3	3	...	No	

	insulin	glyburide-metformin	glipizide-metformin	\
101751	Up	No	No	
101752	Steady	No	No	
101753	Down	No	No	
101754	Steady	No	No	
101755	Up	No	No	
101756	Steady	No	No	
101757	Steady	No	No	
101758	Up	No	No	
101759	Up	No	No	
101760	Down	No	No	
101761	Down	No	No	
101762	Steady	No	No	
101763	Down	No	No	
101764	Up	No	No	
101765	No	No	No	

	glimepiride-pioglitazone	metformin-rosiglitazone	\
101751	No	No	
101752	No	No	
101753	No	No	
101754	No	No	
101755	No	No	
101756	No	No	
101757	No	No	
101758	No	No	
101759	No	No	
101760	No	No	
101761	No	No	
101762	No	No	
101763	No	No	
101764	No	No	
101765	No	No	

	metformin-pioglitazone	change	diabetesMed	readmitted
101751	No	Ch	Yes	0
101752	No	Ch	Yes	0
101753	No	Ch	Yes	0
101754	No	Ch	Yes	0
101755	No	Ch	Yes	0
101756	No	No	Yes	0
101757	No	No	Yes	0
101758	No	Ch	Yes	0
101759	No	Ch	Yes	0
101760	No	Ch	Yes	0
101761	No	Ch	Yes	0
101762	No	No	Yes	0
101763	No	Ch	Yes	0
101764	No	Ch	Yes	0
101765	No	No	No	0

[15 rows x 45 columns]

```
[177]: # find all column names
```

```
print(data.columns)
```

```
Index(['race', 'gender', 'age', 'admission_type_id',
      'discharge_disposition_id', 'admission_source_id', 'time_in_hospital',
      'num_lab_procedures', 'num_procedures', 'num_medications',
      'number_outpatient', 'number_emergency', 'number_inpatient', 'diag_1',
      'diag_2', 'diag_3', 'number_diagnoses', 'max_glu_serum', 'A1Cresult',
      'metformin', 'repaglinide', 'nateglinide', 'chlorpropamide',
      'glimepiride', 'acetoexamide', 'glipizide', 'glyburide', 'tolbutamide',
```

```

'pioglitazone', 'rosiglitazone', 'acarbose', 'miglitol', 'troglitazone',
'tolazamide', 'examide', 'citoglipton', 'insulin',
'glyburide-metformin', 'glipizide-metformin',
'glimepiride-pioglitazone', 'metformin-rosiglitazone',
'metformin-pioglitazone', 'change', 'diabetesMed', 'readmitted'],
dtype='object')

```

```
[178]: # number columns
```

```

num_col=['time_in_hospital', 'num_lab_procedures', 'num_procedures', 'num_medications', 'number_outpatient',
         'number_emergency', 'number_inpatient', 'number_diagnoses']

```

```
[179]: # categorical columns
```

```

cat_col=['race', 'gender', 'max_glu_serum', 'A1Cresult', 'metformin',
         ↪ 'repaglinide', 'nateglinide',
         'chlorpropamide', 'glimepiride', 'acetoexamide', 'glipizide',
         ↪ 'glyburide', 'tolbutamide',
         'pioglitazone', 'rosiglitazone', 'acarbose', 'miglitol',
         ↪ 'troglitazone', 'tolazamide',
         'glimepiride-pioglitazone', 'metformin-rosiglitazone',
         ↪ 'metformin-pioglitazone', 'change',
         'diabetesMed']

```

```
[180]: # head num_col
```

```
data[num_col].head()
```

```
[180]:
```

	time_in_hospital	num_lab_procedures	num_procedures	num_medications	\
0	1	41	0	1	
1	3	59	0	18	
2	2	11	5	13	
3	2	44	1	16	
4	1	51	0	8	

	number_outpatient	number_emergency	number_inpatient	number_diagnoses
0	0	0	0	1
1	0	0	0	9
2	2	0	1	6
3	0	0	0	7
4	0	0	0	5

```
[181]: # head cat_col
```

```
data[cat_col].head()
```

```
[181]:
```

	race	gender	max_glu_serum	A1Cresult	metformin	repaglinide	\
0	Caucasian	Female	NaN	NaN	No	No	
1	Caucasian	Female	NaN	NaN	No	No	
2	AfricanAmerican	Female	NaN	NaN	No	No	
3	Caucasian	Male	NaN	NaN	No	No	
4	Caucasian	Male	NaN	NaN	No	No	

	nateglinide	chlorpropamide	glimepiride	acetoexamide	...	rosiglitazone	\
0	No	No	No	No	...	No	
1	No	No	No	No	...	No	
2	No	No	No	No	...	No	
3	No	No	No	No	...	No	
4	No	No	No	No	...	No	

	acarbose	miglitol	troglitazone	tolazamide	glimepiride-pioglitazone	\
0	No	No	No	No	No	
1	No	No	No	No	No	
2	No	No	No	No	No	
3	No	No	No	No	No	
4	No	No	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 rows x 24 columns]

Explanation: The new feature `meds_x_time` is created to capture the interaction between the number of medications a patient is on and their length of stay in the hospital. This could be a predictor of complexity and readmission risk.

```
[182]: # interaction between number of medications and time in hospital

data['meds_x_time'] = data['num_medications'] * data['time_in_hospital']
```

Explanation: Missing values are filled based on the type of data in each column. Categorical variables are filled with the most frequent value (mode), while numerical variables use the average value (mean).

```
[183]: # fill missing values with the mode for categorical and mean for numerical
        ↪ columns

for column in data.columns:
    if data[column].dtype == 'object':
        data[column].fillna(data[column].mode()[0], inplace=True)
```



```

else:
    data[column].fillna(data[column].mean(), inplace=True)

```

Explanation: Categorical variables such as ‘race’ and ‘gender’ are transformed into a format that can be provided to machine learning models, which require numerical input.

```

[184]: # create dummy variables for categorical features

data = pd.get_dummies(data, drop_first=True) # drop_first to avoid dummy_
↪variable trap

```

```

[185]: data.head()

```

```

[185]:   age  admission_type_id  discharge_disposition_id  admission_source_id  \
0     5                 6                      25                1
1    15                 1                      1                7
2    25                 1                      1                7
3    35                 1                      1                7
4    45                 1                      1                7

      time_in_hospital  num_lab_procedures  num_procedures  num_medications  \
0                   1                   41                0                1
1                   3                   59                0               18
2                   2                   11                5               13
3                   2                   44                1               16
4                   1                   51                0                8

      number_outpatient  number_emergency  ...  insulin_Up  \
0                   0                   0  ...    False
1                   0                   0  ...     True
2                   2                   0  ...    False
3                   0                   0  ...     True
4                   0                   0  ...    False

      glyburide-metformin_No  glyburide-metformin_Steady  glyburide-metformin_Up  \
0                   True                False                False
1                   True                False                False
2                   True                False                False
3                   True                False                False
4                   True                False                False

      glipizide-metformin_Steady  glimepiride-pioglitazone_Steady  \
0                   False                False
1                   False                False
2                   False                False
3                   False                False
4                   False                False

```

	metformin-rosiglitazone_Steady	metformin-pioglitazone_Steady	change_No	\
0	False	False	True	
1	False	False	False	
2	False	False	True	
3	False	False	False	
4	False	False	False	

	diabetesMed_Yes
0	False
1	True
2	True
3	True
4	True

[5 rows x 2319 columns]

Explanation: This step focuses on reducing the dataset to include only the features that are most likely to influence the readmission outcome. I will be looking at this in step 3.

[]: