Phase 2: Data Cleaning and Preprocessing

Masheia Dzimba and Peter Mangoro

2025-11-28

Contents

1 Introduction

This document presents Phase 2: Data Cleaning and Preprocessing. In this phase, we clean the response variable, handle missing values, perform feature engineering, and prepare datasets for both Logistic Regression and CART models.

2 Load Original Data

```
Original dataset:  25000  observations,  17  variables
```

3 Clean Response Variable

```
Original readmitted distribution:
```

```
   no   yes
13246 11754
```

```
Binary conversion readmitted distribution:
```

```
    0     1
13246 11754
```

```
      0       1
0.52984 0.47016
```

4 Handle Missing Values

```
Missing values before cleaning:
```

Table 1: Missing Values Before Cleaning

	Variable	Missing_Count	Missing_Percentage
medical_specialty	medical_specialty	12382	49.53
diag_1	diag_1	4	0.02
diag_2	diag_2	42	0.17
diag_3	diag_3	196	0.78

```
--- Dropping rows with missing diag_1 ---
```

```
Rows before:  25000
```

```
Rows after:  24996
```

```
Rows dropped:  4  ( 0.02 %)
```

- We are only going to drop missing data from diag_1 because it only contains 4 missing rows, and since it is the main diagnosis it means we cant assume no medication was given.
- medical_speciality has mor than 12k (49.53%) of missing data and we will not drop it as it valuable data showing that the responsible doctors has no speciality at the time of treatment.
- diag_2 and diag_3 have significant missing data and we are not droping them as this again is important information that shows that patient did recieve the main diagnosis but might not have been treated for 2 or 3 diagnosis.

5 Feature Engineering

Added column to show number of diagnoses distribution:

```
    1     2     3
   21   196 24779
```

Table shows that:

- 24 779 patients had 3 diagnosis

- 196 patients had 2 diagnosis

- 21 patients had 1 diagnosis

Added Medications per day to capture treatment intensity eg:

- 20 medications over 2 days = 10/day (high intensity)

- 20 medications over 10 days = 2/day (lower intensity)

Added Total previous visits with formula:

n_outpatient + n_inpatient + n_emergency

- Sums all previous visit types into one variable.

- Frequent users may have higher readmission risk

6 Feature Selection

Selected variables:

  Categorical: age, medical_specialty, diag_1, change, diabetes_med, glucose_test, A1Ctest

  Numerical: time_in_hospital, n_lab_procedures, n_procedures, n_medications

7 Encode Variables for Models

7.1 Encoding for Logistic Regression (Dummy Encoding)

Binary variables converted:

  change_binary: 11501 'yes' values

  diabetes_med_binary: 19224 'yes' values

```
Age dummy variables created:  6  columns

Medical specialty dummy variables created:  7  columns

Primary diagnosis dummy variables created:  7  columns

Glucose test dummy variables created:  3  columns

A1C test dummy variables created:  3  columns
```

- we are using one hot encoding, so a new column is made for each value, the matching column is set to 1 while the rest are set to 0.
- we do this because Logistic regression expects a binary value
- during model fitting, R will drop one column as the reference category ## Combine Features for Logistic Regression

```
Logistic Regression dataset created:

  Observations:  24996

  Variables:  39

  Response variable: readmitted (binary 0/1)

Reference categories for dummy variables:

  Age: Will use first alphabetically as reference

  Medical specialty: 'Missing' is most frequent (49.53%)

  Primary diagnosis: Most frequent category as reference

  Glucose test: 'no' as reference

  A1C test: 'no' as reference

  Note: model.matrix() creates all levels; will drop reference in model fitting
```

7.2 Encoding for CART (Factor Encoding)

```
Response variable converted to factor:

readmitted_factor
Not_Readmitted     Readmitted
         13244          11752

CART dataset created:
```

```
  Observations:  24996

  Variables:  18

  Response variable: readmitted (factor: Not_Readmitted/Readmitted)

Categorical variables converted to factors:

  - age: ordered factor with  6  levels

  - medical_specialty: factor with  7  levels

  - diag_1: factor with  7  levels

  - change: factor with  2  levels

  - diabetes_med: factor with  2  levels

  - glucose_test: factor with  3  levels

  - A1Ctest: factor with  3  levels
```

8 Summary

This phase successfully cleaned and preprocessed the data:

- **Response variable**: Converted to binary format (0/1)
- **Missing values**: Dropped 4 rows with missing primary diagnosis
- **Feature engineering**: Created 3 derived features
- **Encoding**: Prepared datasets for both model types
- **Final dataset**: 24,996 observations (99.98% retention)