Phase 4: Logistic Regression Model

Masheia Dzimba and Peter Mangoro

2025-12-02

Contents

1 Introduction

This document presents Phase 4: Logistic Regression Model. We build a logistic regression model to predict 30-day hospital readmissions, interpret coefficients and odds ratios, and evaluate model performance.

2 Load Data

```
Dataset:  24996  observations,  39  variables
```

3 Train/Test Split

```
Training set:  17498  observations (70%)

Testing set:  7498  observations (30%)

Readmission rate - Training:  47 %

Readmission rate - Testing:  47.05 %
```

4 Build Model

```r
# Build model formula
predictor_vars <- setdiff(colnames(data_train), "readmitted")
formula_str <- paste("readmitted ~", paste(predictor_vars, collapse = " + "))
formula_obj <- as.formula(formula_str)

# Fit the model
model_logistic <- glm(formula_obj,
                      data = data_train,
                      family = binomial(link = "logit"))
```

5 Logistic Regression Equation

5.1 Mathematical Formulation

The logistic regression model uses the logistic function to model the probability of readmission:

$$P(\text{Readmitted} = 1 \mid X) = \frac{e^{\beta_0 + \sum_{i=1}^{p} \beta_i X_i}}{1 + e^{\beta_0 + \sum_{i=1}^{p} \beta_i X_i}}$$

Or equivalently,

$$\text{logit}(P) = \ln\left(\frac{P}{1-P}\right) = \beta_0 + \sum_{i=1}^{p} \beta_i X_i$$

Where:

- $P(\text{Readmitted} = 1 \mid X)$ = Probability of readmission given predictor variables
- β_0 = Intercept (baseline log-odds)
- $\beta_1, \beta_2, ..., \beta_p$ = Coefficients for each predictor variable
- $X_1, X_2, ..., X_p$ = Predictor variables

5.2 Fitted Model Equation

$$\begin{aligned}\text{logit}(P) = &-2.575 + 0.4305 \cdot X_{\text{n\_diagnoses}} + 0.3944 \cdot X_{\text{n\_inpatient}} \\ &+ 0.3161 \cdot X_{\text{medspec\_medical\_specialtyEmergency.Trauma}} + 0.3057 \cdot X_{\text{age\_age70.80}} \\ &+ 0.2941 \cdot X_{\text{medspec\_medical\_specialtyMissing}} + ... \text{ (other predictors)}\end{aligned}$$

Where:

- P = Probability of readmission
- $\beta_0 = -2.575$ (intercept)
- $\beta_{\text{n\_diagnoses}} = 0.4305$
- $\beta_{\text{n\_inpatient}} = 0.3944$
- $\beta_{\text{medspec\_medical\_specialtyEmergency.Trauma}} = 0.3161$
- $\beta_{\text{age\_age70.80}} = 0.3057$
- $\beta_{\text{medspec\_medical\_specialtyMissing}} = 0.2941$

5.3 Example Calculation

For a patient with:

- n diagnoses = 3
- n inpatient = 2
- medspec medical specialtyEmergency.Trauma = 0

Logit calculation:

$$\begin{aligned}\text{logit}(P) &= -2.575 + 0.4305 \cdot 3 + 0.3944 \cdot 2 + 0.3161 \cdot 0 \\ &= -0.4947\end{aligned}$$

Probability calculation:

$$P = \frac{e^{-0.4947}}{1 + e^{-0.4947}} = 0.3788$$

Interpretation: This patient has a 37.88% probability of readmission.

5.4 Coefficient Interpretation

Coefficients (β):

- **Sign**: Positive coefficients increase the log-odds (and probability) of readmission; negative coefficients decrease it
- **Magnitude**: Larger absolute values indicate stronger effects

Example: If $\beta_{\text{n\_inpatient}} = 0.38$:

- A one-unit increase in previous inpatient visits increases the log-odds of readmission by 0.38
- This corresponds to an odds ratio of $e^{0.38} = 1.46$ (46% increase in odds)

5.5 Gradient (Rate of Change)

The **gradient** represents how quickly the probability changes with respect to each predictor:

$$\frac{\partial P}{\partial X_i} = \beta_i \cdot P(1 - P)$$

Key Points:

- The gradient is **not constant** - it depends on the current probability P
- Maximum gradient occurs when $P = 0.5$ (steepest part of the S-curve)
- The gradient is smaller when P is close to 0 or 1 (flatter parts of the curve)

6 Model Summary

```

Call:
glm(formula = formula_obj, family = binomial(link = "logit"), 
    data = data_train)

Coefficients: (6 not defined because of singularities)
                                                  Estimate Std. Error z value
(Intercept)                                     -2.5749951  0.5259409  -4.896
time_in_hospital                                 0.0109356  0.0087986   1.243
n_lab_procedures                                 0.0016485  0.0009307   1.771
n_procedures                                    -0.0401598  0.0106444  -3.773
n_medications                                    0.0030947  0.0029733   1.041
n_outpatient                                     0.1055523  0.0155366   6.794
n_inpatient                                      0.3943691  0.0173317  22.754
n_emergency                                      0.2283614  0.0305429   7.477
n_diagnoses                                      0.4304785  0.1664149   2.587
medications_per_day                             -0.0052265  0.0061653  -0.848
total_previous_visits                                   NA         NA      NA
change_binary                                    0.0389718  0.0375622   1.038
diabetes_med_binary                              0.2178359  0.0436772   4.987
```

```
age_age40.50                                     0.0571183 0.1059320   0.539
age_age50.60                                     0.1148449 0.1005245   1.142
age_age60.70                                     0.2026348 0.0983698   2.060
age_age70.80                                     0.3056551 0.0972168   3.144
age_age80.90                                     0.2831670 0.0991876   2.855
age_age90.100                                           NA        NA      NA
medspec_medical_specialtyCardiology              0.2826373 0.1006864   2.807
medspec_medical_specialtyEmergency.Trauma        0.3160638 0.0956412   3.305
medspec_medical_specialtyFamily.GeneralPractice  0.2659748 0.0953258   2.790
medspec_medical_specialtyInternalMedicine        0.1077578 0.0868588   1.241
medspec_medical_specialtyMissing                 0.2940955 0.0782374   3.759
medspec_medical_specialtyOther                   0.1347454 0.0891943   1.511
medspec_medical_specialtySurgery                        NA        NA      NA
diag1_diag_1circulatory                          0.0483365 0.0514149   0.940
diag1_diag_1diabetes                             0.1917932 0.0738409   2.597
diag1_diag_1digestive                           -0.0197057 0.0670315  -0.294
diag1_diag_1injury                              -0.2339514 0.0753557  -3.105
diag1_diag_1musculoskeletal                     -0.1613060 0.0865966  -1.863
diag1_diag_1other                               -0.1584385 0.0520793  -3.042
diag1_diag_1respiratory                                 NA        NA      NA
glucose_glucose_testhigh                         0.0404400 0.1363745   0.297
glucose_glucose_testno                           0.0015331 0.0978307   0.016
glucose_glucose_testnormal                              NA        NA      NA
a1c_a1ctesthigh                                  0.1737157 0.0847070   2.051
a1c_a1ctestno                                    0.1897011 0.0739780   2.564
a1c_a1ctestnormal                                       NA        NA      NA
                                                Pr(>|z|)
(Intercept)                                     9.78e-07 ***
time_in_hospital                                0.213912
n_lab_procedures                                0.076510 .
n_procedures                                    0.000161 ***
n_medications                                   0.297959
n_outpatient                                    1.09e-11 ***
n_inpatient                                      < 2e-16 ***
n_emergency                                     7.62e-14 ***
n_diagnoses                                     0.009688 **
medications_per_day                             0.396584
total_previous_visits                                 NA
change_binary                                   0.299490
diabetes_med_binary                             6.12e-07 ***
age_age40.50                                    0.589751
age_age50.60                                    0.253264
age_age60.70                                    0.039405 *
age_age70.80                                    0.001666 **
age_age80.90                                    0.004306 **
age_age90.100                                         NA
medspec_medical_specialtyCardiology             0.004999 **
medspec_medical_specialtyEmergency.Trauma       0.000951 ***
medspec_medical_specialtyFamily.GeneralPractice 0.005268 **
medspec_medical_specialtyInternalMedicine       0.214750
medspec_medical_specialtyMissing                0.000171 ***
medspec_medical_specialtyOther                  0.130866
medspec_medical_specialtySurgery                      NA
diag1_diag_1circulatory                         0.347153
```

```
diag1_diag_1diabetes                           0.009394 **
diag1_diag_1digestive                          0.768776
diag1_diag_1injury                             0.001905 **
diag1_diag_1musculoskeletal                    0.062501 .
diag1_diag_1other                              0.002348 **
diag1_diag_1respiratory                              NA
glucose_glucose_testhigh                       0.766820
glucose_glucose_testno                         0.987497
glucose_glucose_testnormal                           NA
a1c_a1ctesthigh                                0.040288 *
a1c_a1ctestno                                  0.010339 *
a1c_a1ctestnormal                                    NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 24194  on 17497  degrees of freedom
Residual deviance: 22913  on 17465  degrees of freedom
AIC: 22979

Number of Fisher Scoring iterations: 4
```

7 Regression Output

Table 1: Top 5 Most Significant Variables: Coefficients and Statistics

Variable	Coefficient	Std_Error	Z_Value	P_Value
n_outpatient	0.1056	0.0155	6.7938	0
n_inpatient	0.3944	0.0173	22.7542	0
n_emergency	0.2284	0.0305	7.4767	0
(Intercept)	-2.5750	0.5259	-4.8960	0
diabetes_med_binary	0.2178	0.0437	4.9874	0

Table 2: Top 5 Most Significant Variables: Odds Ratios and Confidence Intervals

Variable	Odds_Ratio	CI_Lower	CI_Upper	P_Value
n_outpatient	1.1113	1.0780	1.1457	0
n_inpatient	1.4834	1.4339	1.5347	0
n_emergency	1.2565	1.1835	1.3341	0
(Intercept)	0.0762	0.0272	0.2135	0
diabetes_med_binary	1.2434	1.1414	1.3545	0

Table 3: Additional Significant Variables (p < 0.05)

Variable	Coefficient	Odds_Ratio	P_Value
n_procedures	-0.0402	0.9606	0.0002
medspec_medical_specialtyMissing	0.2941	1.3419	0.0002
medspec_medical_specialtyEmergency.Trauma	0.3161	1.3717	0.0010
age_age70.80	0.3057	1.3575	0.0017
diag1_diag_1injury	-0.2340	0.7914	0.0019
diag1_diag_1other	-0.1584	0.8535	0.0023
age_age80.90	0.2832	1.3273	0.0043
medspec_medical_specialtyCardiology	0.2826	1.3266	0.0050
medspec_medical_specialtyFamily.GeneralPractice	0.2660	1.3047	0.0053
diag1_diag_1diabetes	0.1918	1.2114	0.0094

8 Hypothesis Testing

Hypothesis Testing

For each variable in the model:

H_0: $\beta_{\text{variable}} = 0$ (variable has no effect on readmission)

H_1: $\beta_{\text{variable}} \neq 0$ (variable affects readmission)

Table 4: Top 5 Most Significant Variables

Variable	Coefficient	P_Value	Odds_Ratio
n_outpatient	0.1056	0e+00	1.1113
n_inpatient	0.3944	0e+00	1.4834
n_emergency	0.2284	0e+00	1.2565
(Intercept)	-2.5750	1e-06	0.0762
diabetes_med_binary	0.2178	1e-06	1.2434

9 R-squared Interpretation

McFadden's Pseudo R^2: 0.053

Interpretation: The model explains approximately 5.3 % of the variance in readmission status.

10 Model Evaluation

Table 5: Confusion Matrix

	0	1
0	3136	2101
1	834	1427

Table 6: Model Performance Metrics

Metric	Value	Percentage
Accuracy	0.61	60.86
Precision	0.63	63.11
Recall (Sensitivity)	0.40	40.45
Specificity	0.79	78.99
F1-Score	0.49	49.30

11 ROC Curve

```
Area Under the Curve (AUC):  0.6469
```

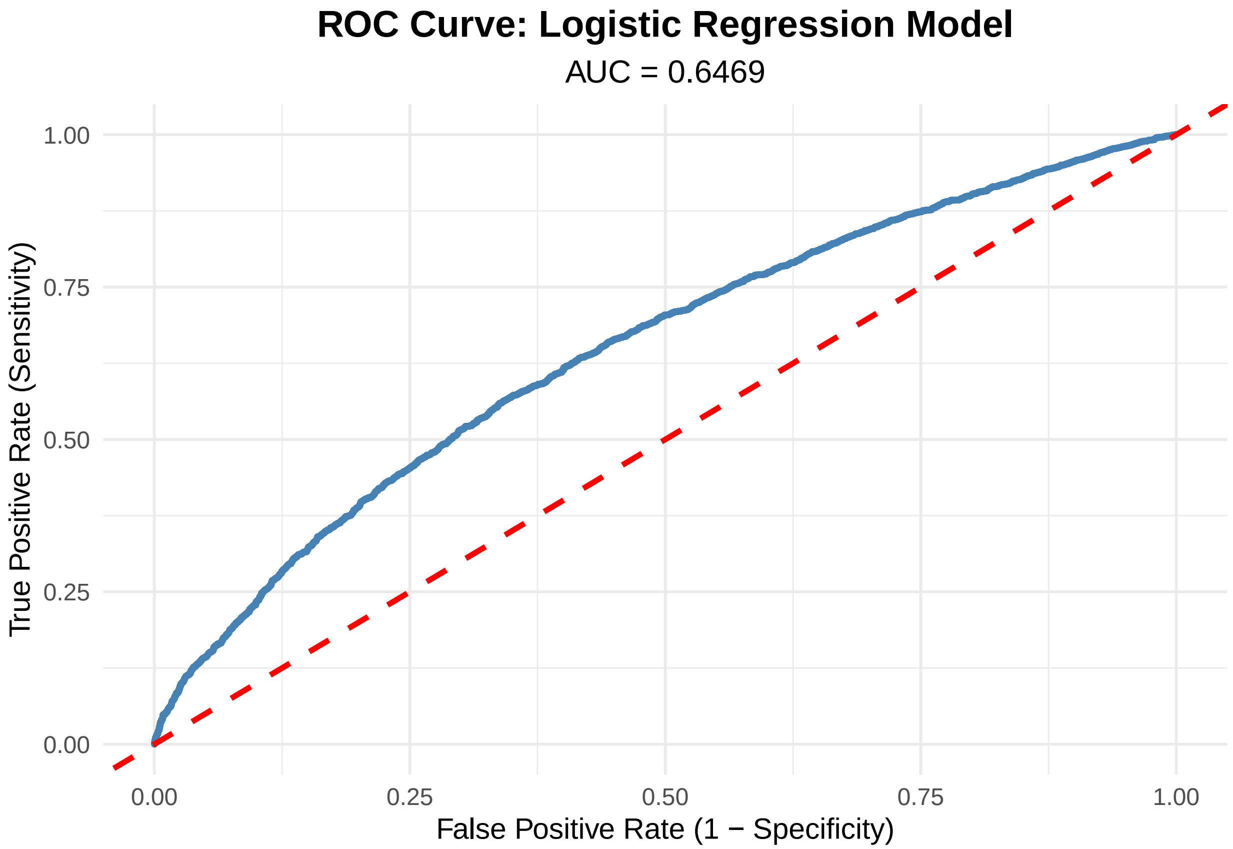

Figure 1: ROC Curve: Logistic Regression Model

12 Odds Ratios Interpretation

```
Top Variables with Highest Odds Ratios (Risk Factors):
```

Table 7: Top 5 Risk Factors (Highest Odds Ratios)

Variable	Odds_Ratio	CI_Lower	CI_Upper	P_Value
n_diagnoses	1.5380	1.1099	2.1311	0.0097
n_inpatient	1.4834	1.4339	1.5347	0.0000
medspec_medical_specialtyEmergency.Trauma	1.3717	1.1372	1.6545	0.0010
age_age70.80	1.3575	1.1220	1.6425	0.0017
medspec_medical_specialtyMissing	1.3419	1.1511	1.5643	0.0002

Table 8: Additional Risk Factors

Variable	Odds_Ratio	CI_Lower	CI_Upper	P_Value
age_age80.90	1.3273	1.0928	1.6122	0.0043
medspec_medical_specialtyCardiology	1.3266	1.0890	1.6160	0.0050
medspec_medical_specialtyFamily.GeneralPractice	1.3047	1.0824	1.5727	0.0053
n_emergency	1.2565	1.1835	1.3341	0.0000
diabetes_med_binary	1.2434	1.1414	1.3545	0.0000

13 Summary

- **Accuracy**: 60.86%
- **AUC**: 0.647
- **Pseudo R^2**: 5.3%
- **19 significant variables** identified ($p < 0.05$)
- **Top predictor**: n diagnoses (OR: 1.54)