DATA SCIENCE INSTITUTE

WOOLF

ASSIGNMENT

ADVANCED PREDICTIVE MODELING

Python

REPORT

Pillar1-e

Prepared By

William C. Phiri

PG Dip in Data Science

Student No. 4295158639

+353-87-3502102

chenechoz@gmail.com

Athlone, IRELAND

BACKGROUND:

The Data for a Study of Risk Factors Associated with Low Infant Birth Weight. Data were collected at Baystate Medical Center, Springfield, Massachusetts.

DATA CATALOGUE:

Variable	Description	Values/Coding
LOW	Low Birth Weight	0 = Not low, 1 = low
AGE	Age of the Mother	In years (continuous)
LWT	Weight at Last Menstrual Period	In pounds (continuous)
RACE	Race	1 = White, 2 = Black, 3 = Other
SMOKE	Smoking Status During Pregnancy	1 = Yes, 0 = No
PTL	History of Premature Labor	0 = None, 1 = One, 2 = Two, etc. (count)
HT	History of Hypertension	1 = Yes, 0 = No
UI	Presence of Uterine Irritability	1 = Yes, 0 = No
FTV	Number of Physician Visits During First Trimester	0 = None, 1 = One, 2 = Two, etc. (count)

ASSIGNMENT OBJECTIVE:

Consider LOW as dependent variable and remaining variables listed above as independent variables.

QUESTIONS:

1. Import BIRTH WEIGHT data.

2. Cross tabulate dependent variable with each independent variable.

3. Develop a model to predict if birth weight is low or not using the given variables.

4. Generate three classification tables with cut-off values 0.4, 0.3 and 0.55.

5. Calculate sensitivity, specificity and misclassification rate for all three tables above. What is the recommended cut-off value?

6. Obtain ROC curve and report area under curve.

RESULTS:

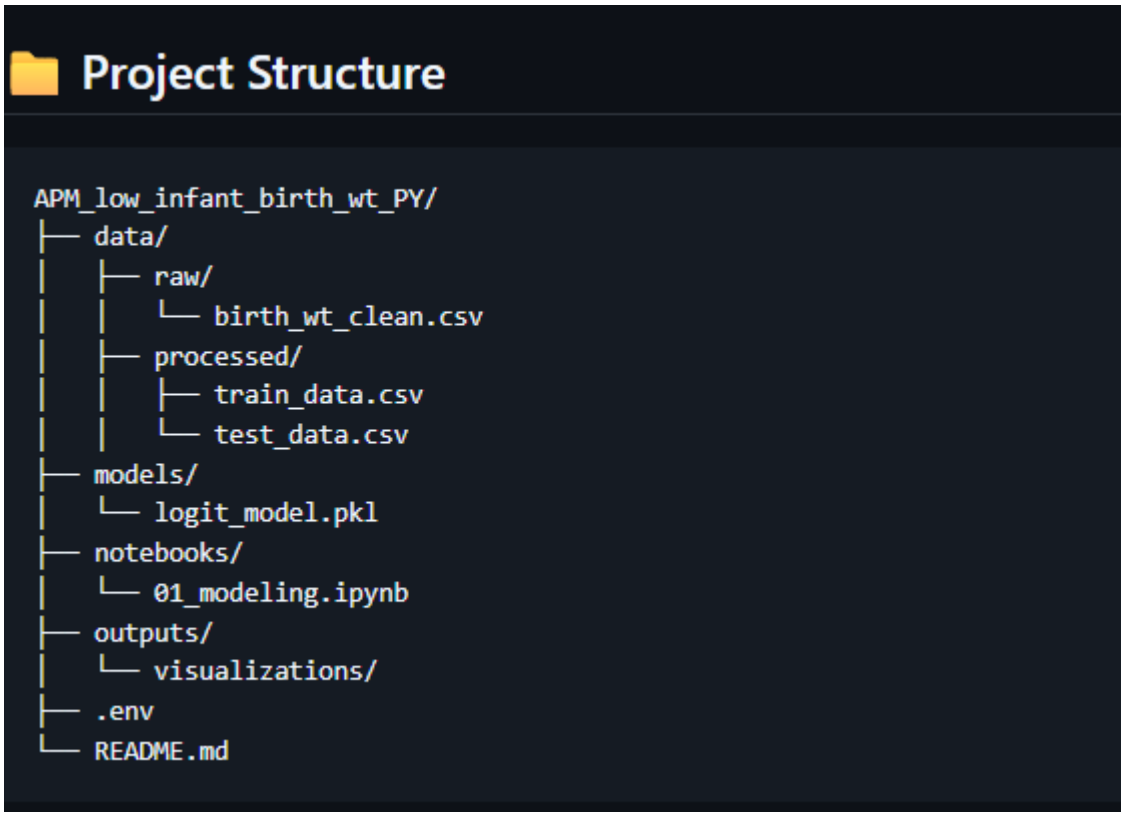
INTRODUCTION

QUESTION 1: Import BIRTH WEIGHT data.

OUTPUT: Refer to GitHub repository at LINK HERE for the model and the associated notebooks.

NOTE: For educational purposes the dataset BIRTH WEIGHT.csv was uploaded to a PostgreSQL database Neon.Tech to facilitate enhancing a knowledge gap of being able to query data directly from a database and incorporating this into the work structure to try and simulate real world work-flows.

Project Repository Structure:



QUESTION 2: Cross tabulate dependent variable with each independent variable.

(Continuous variables were left out)

Summary Table: Crosstabulation of Predictors vs Low Birth Weight

1. Race

Race	0 (Normal)	1 (Low)	Total
1	78	18	96
2	23	17	40
3	34	19	53

Interpretation:

- Race **1 (White)** has the highest count of normal births.
- Race **2 (Black)** has a **high proportion** of low birth weight (17/40 = 42.5%).
- Race **3 (Other)** also shows significant risk (~36%).

Insight: Race 2 may be more associated with risk of low birth weight in this dataset.

2. Smoke (Smoking during pregnancy)

Smoke	0 (Normal)	1 (Low)	Total
0	83	30	113
1	26	24	50

Interpretation:

- **Smokers (1)** have a higher rate of low birth weight (24/50 = 48%).
- **Non-smokers (0)** have a lower rate (30/113 ≈ 26.5%).

Insight: Smoking is strongly associated with increased risk of low birth weight.

3. **PTL** (History of Premature Labor)

PTL	0 (Normal)	1 (Low)	Total
0	92	34	126
1	8	10	18
2	4	5	9
3	1	2	3

Interpretation:

- **PTL = 0 (No history)** has the majority of births, but 34 were low.
- As PTL increases (e.g., 2 or 3), the proportion of low birth weights increases.
    - PTL=2 → 5/9 were low (≈56%)
    - PTL=3 → 2/3 were low (≈67%)

Insight: Past premature labor may be predictive of low birth weight.

4. **HT** (History of Hypertension)

PTL	0 (Normal)	1 (Low)	Total
0	101	44	145
1	8	10	18

Interpretation:

- Hypertension history (HT = 1) shows high risk:
    - 10/18 = 55% of these births are low birth weight
- Without hypertension (HT = 0), only 44/145 ≈ 30% were low.

Insight: Maternal hypertension is a significant risk factor.

5. **UI** (Uterine Irritability)

UI	0 (Normal)	1 (Low)	Total
0	96	35	131
1	13	19	32

Interpretation:

- Uterine irritability (UI = 1) increases the chance of low birth weight:
    - 19/32 ≈ 59% of these cases are low
- No irritability (UI = 0) has a lower rate (~27%).

Insight: Uterine irritability is another red flag.

6. **FTV** (First Trimester Visits)

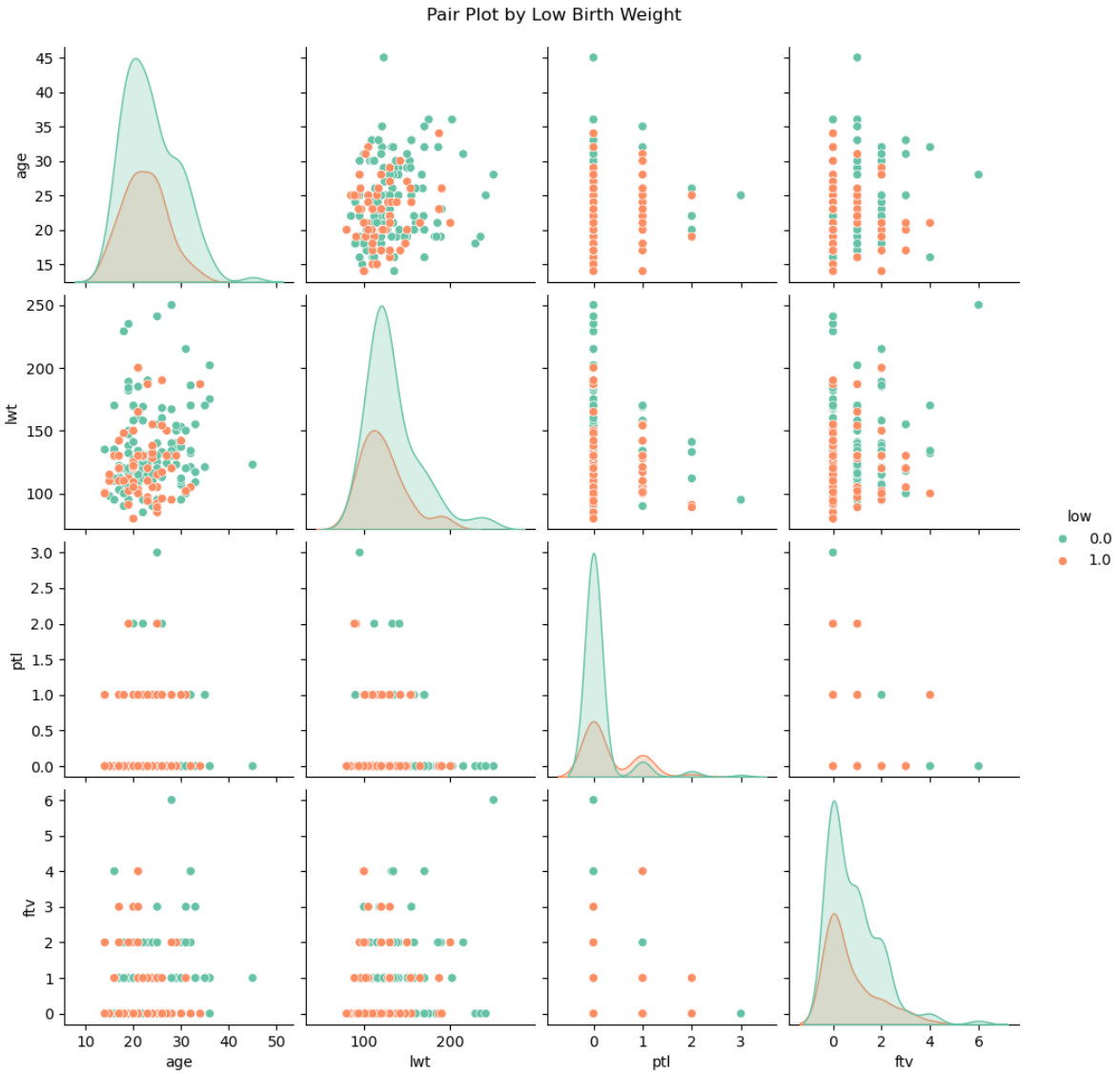
FTV	0 (Normal)	1 (Low)	Total
0	38	14	52
1	42	18	60
2	19	7	26
3	9	15	24

Interpretation:

- Low FTV (0–1 visits) still sees high low birth weight
- FTV=3 has 15/24 low births = 62.5%

Insight: Not clearly linear — could be confounded by high-risk mothers visiting more frequently.

Data Visualisation:



QUESTION 3: Develop a model to predict if birth weight is low or not using the given variables.

```
Confusion Matrix:
[[26  0]
 [ 9  3]]

Classification Report:
              precision    recall  f1-score   support

         0.0       0.74      1.00      0.85        26
         1.0       1.00      0.25      0.40        12

    accuracy                           0.76        38
   macro avg       0.87      0.62      0.63        38
weighted avg       0.82      0.76      0.71        38

  Feature  Coefficient
3   smoke     0.422931
2    race     0.404149
5      ht     0.390333
6      ui     0.285877
1     lwt    -0.219737
4     ptl     0.186786
7     ftv    -0.127036
0     age     0.032151
```

Considering positive coefficients increases the likelihood of the outcome of low birth weight, in this instance we can conclude the following about our variables;

Variable	Coefficient	Interpretation
smoke	***0.423***	Smoking increases the risk of low birth weight.
race	***0.404***	Being non-white (race=2 or 3) increases the risk.
ht	***0.39***	History of hypertension increases the risk.
ui	***0.286***	Uterine irritability increases the risk.
ptl	***0.187***	Previous premature labor increases the risk.
age	***0.032***	Slight increase in risk with age (small effect).
lwt	*−0.220*	Higher maternal weight reduces the risk.
ftv	*−0.127*	More prenatal visits reduce the risk.

Features like ***smoke, ht, ui***, and ***race*** have the strongest positive influence on predicting low birth weight. On the other hand features ***like lwt (last weight)*** and ***ftv (prenatal care visits)*** are protective.

QUESTION 4/5: Generate three classification tables with cut-off values 0.4, 0.3 and 0.55. Then Calculate sensitivity, specificity and misclassification rate for all three tables above. What is the recommended cut-off value?

```
Threshold: 0.4
----------------------------
Confusion Matrix:
[[TN: 25, FP: 1]
 [FN: 7, TP: 5]]
Sensitivity (Recall for 1): 0.42
Specificity (Recall for 0): 0.96
Misclassification Rate:     0.21

Threshold: 0.3
----------------------------
Confusion Matrix:
[[TN: 19, FP: 7]
 [FN: 6, TP: 6]]
Sensitivity (Recall for 1): 0.50
Specificity (Recall for 0): 0.73
Misclassification Rate:     0.34

Threshold: 0.55
----------------------------
Confusion Matrix:
[[TN: 26, FP: 0]
 [FN: 11, TP: 1]]
Sensitivity (Recall for 1): 0.08
Specificity (Recall for 0): 1.00
Misclassification Rate:     0.29
```

Recommendation

Threshold = 0.4 appears ss the best compromise:

- You get **decent sensitivity** (42%)—meaning you're identifying nearly half of true low birth weight cases.
- **High specificity** (96%)—avoiding too many false positives.
- **Lowest misclassification rate** (21%).

Why not 0.3?

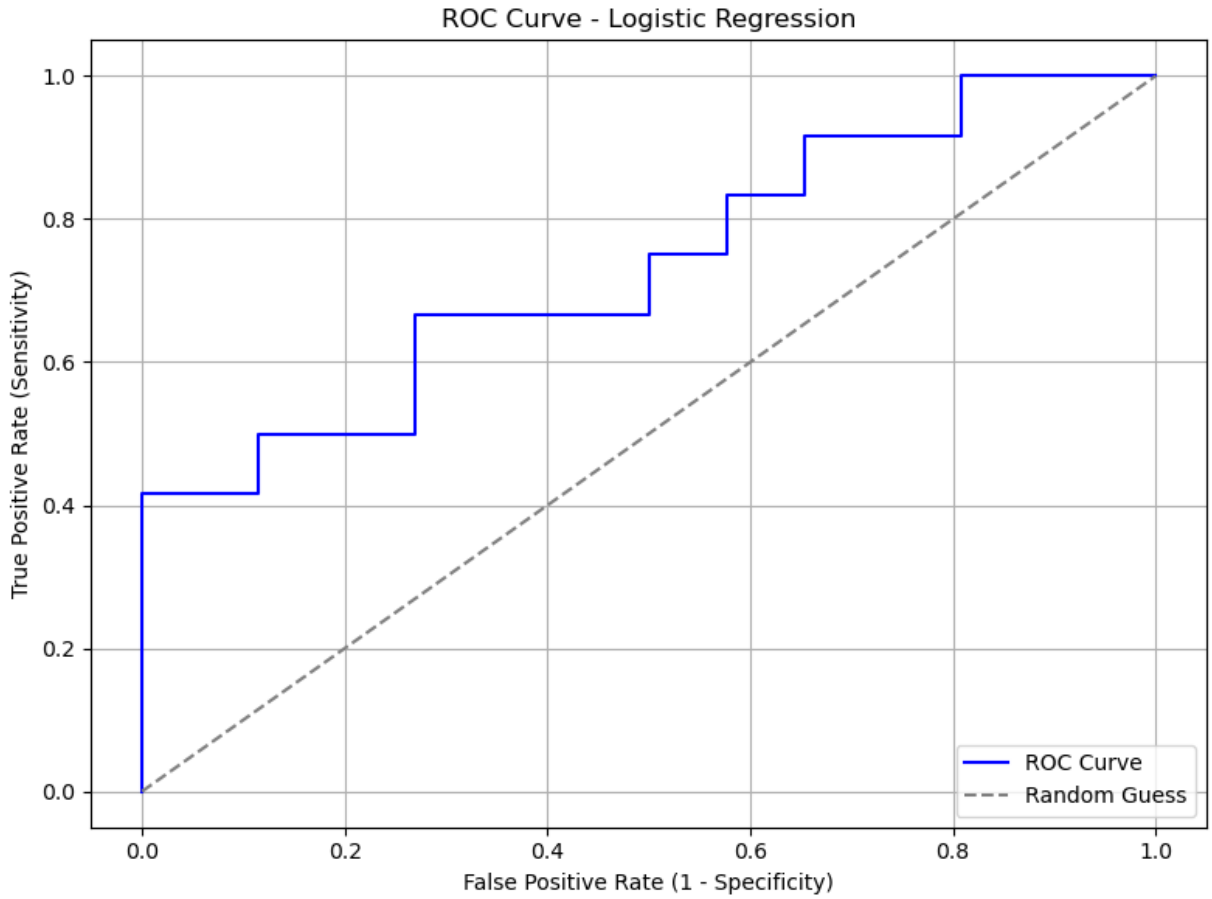
Although it has the **best sensitivity**, it also has:

- **High false positives** (specificity drops to 73%)
- **Worst misclassification rate** (34%)

Why not 0.55?

It completely **fails to detect low birth weight** (sensitivity = 8%), even though it never misclassifies non-low cases (specificity = 100%).

QUESTION 6: Obtain ROC curve and report area under curve.



From the plot:

- The ROC curve is consistently above the diagonal (random guess) line.
- This shows the model has some predictive power and performs better than random chance.

Area Under Curve(AUC) = 0.73

The model has **fair discrimination** — it's reasonably good at distinguishing between **low birth weight** and **normal weight** infants.