MAST5955: Predictive Modelling

Assessment 2

Name: Sagari Muraliegaran

Contents

Data Summary

Numerical Summary

Table 1 shows the summary statistics for numeric variables and data types for all variables in the dataset and categorical variables. (Appendix 1)

Variable	Min	1st Quartile	Median	Mean	3rd Quartile	Max	Class
Medal	0	0	0	0.3627	1	1	Numeric
Previous Medal	-	-	-	-	-	-	Character
Top 10	-	-	-	-	-	-	Character
Country	-	-	-	-	-	-	Character
Main Sponsor	-	-	-	-	-	-	Character
Age	19	21	23	23.53	26	28	Numeric
Hours training	14	15.4	17	16.97	18.5	20	Numeric

Table 2 shows the summary of categorical variables, showing their categorical and corresponding counts in the dataset. (Appendix 2)

Variable	Categories	Counts
Previous Medal	No, yes	No:1313 | Yes: 1036
Top 10	No, yes	No: 1637 | Yes: 712
Country	A,B,C	A: 816 |B: 716|C: 772
Main Sponsor	A,N,P,U	A:763 | N: 693 | P: 446 | U: 447

Table 1 summarises the dataset's numerical and categorical variables. Key numeric variables including minimum, quartiles, medians, means and maximums. For example, Medal, a binary variable, has a mean of 0.3627, indicating that 36.27% of athletes received a medal. The age spans from 19 to 28, with a median of 23 and a mean of 23.53, indicating a balanced distribution. Training lasts 14 to 20 hours, with a median of 17 and a mean of 16.97, indicating a symmetric distribution. Categorical variables, such as Previous Medal, Top 10, Country and Main Sponsor, are distinguished by their data types, as numeric statistics do not apply.

Table 2 displays the frequency distribution of categorical variables. In terms of prior medals, 1036 competitors won one, whereas 1313 did not. Top 10 implies that 1637 athletes are not in the top 10, whereas 712 are. The country categorises athletes into three groups: A (816), B (716), and C(772). The main sponsor list reveals that A sponsors 763 athletes, N sponsors 693, P sponsors 446 and U sponsors 447.

Table 1 summarises numerical distributions, whereas Table 2 focuses on the composition of categorical variables, giving a comprehensive understanding of the dataset.

Graphical Summary

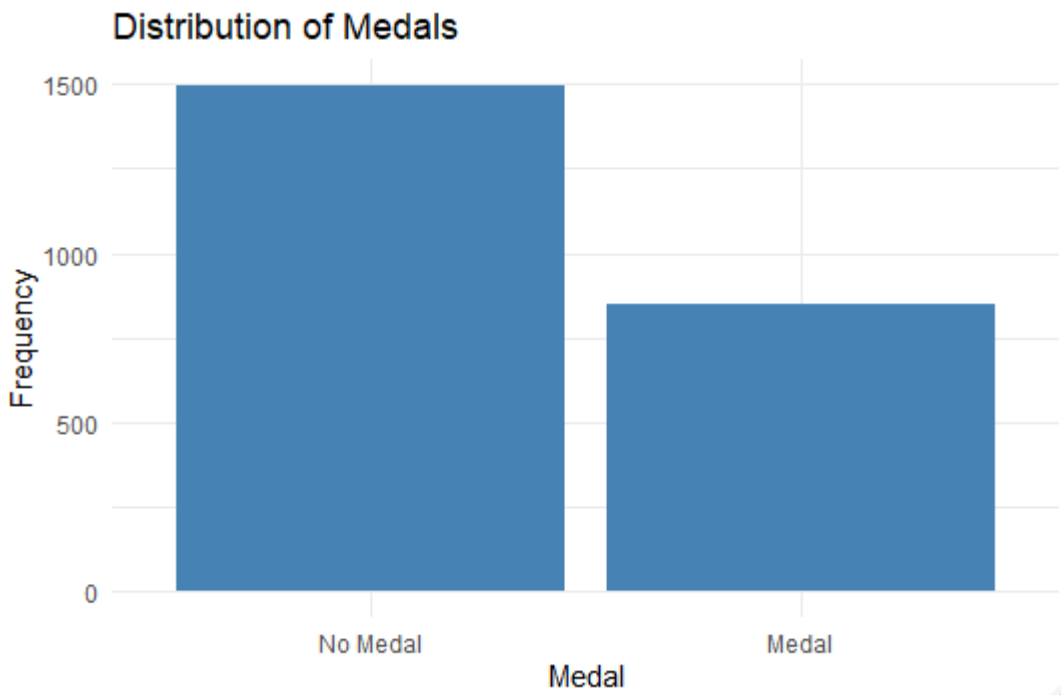


Figure 1 shows a bar chart of the distribution of athletes who won a medal ("Medal") versus those who did not ("No Medal"). (Appendix 3)

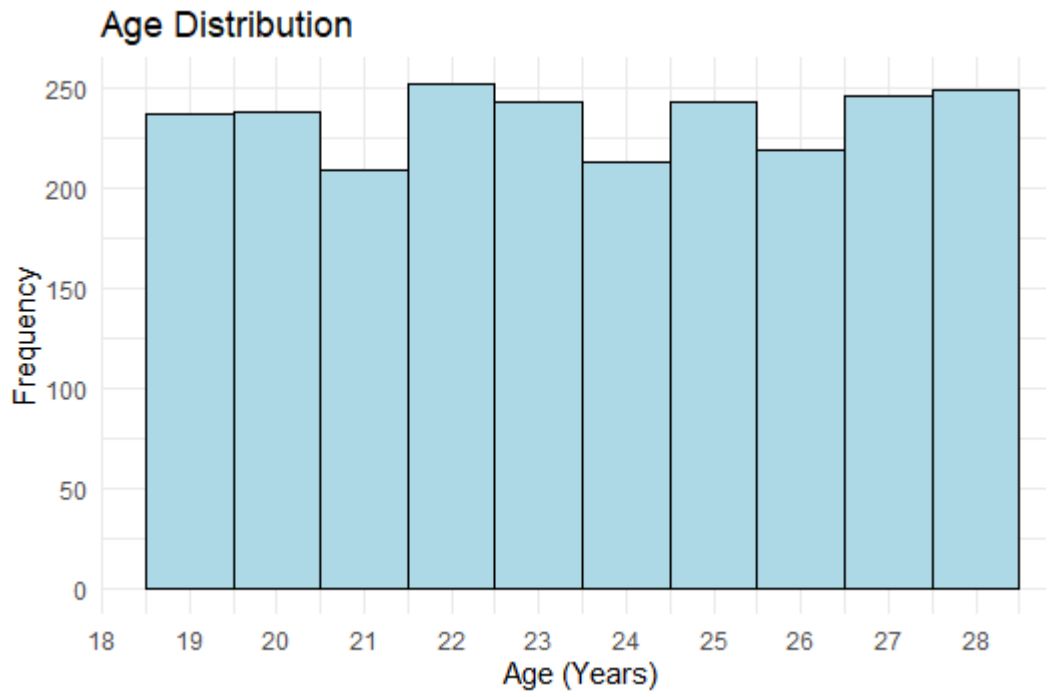


Figure 2 shows a histogram representing the age distribution of athletes. (Appendix 4)

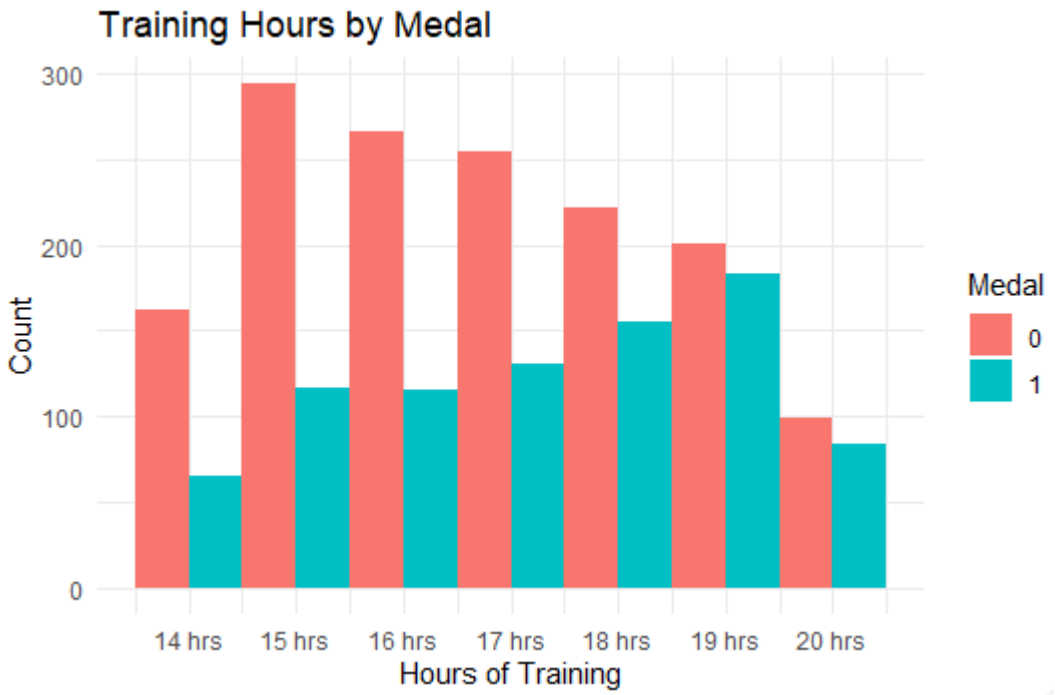


Figure 3 shows a bar chart with the distribution of weekly training hours for athletes who won a medal (1) vs those who did not (0). (Appendix 5)

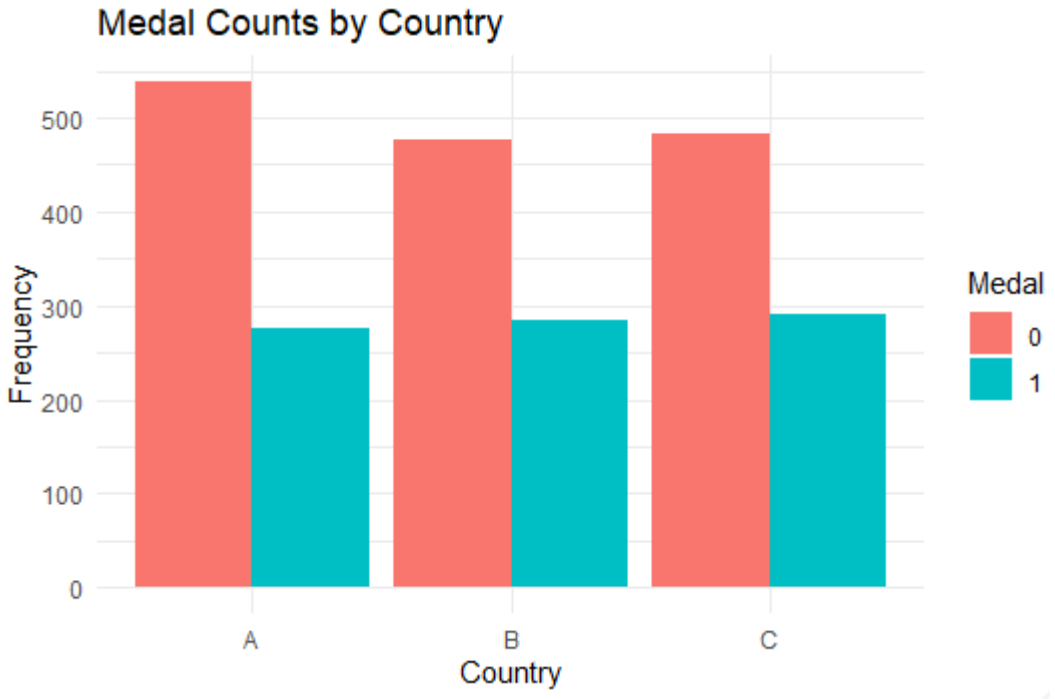


Figure 4 is a bar chart that shows the distribution of medal winners (1) and non-winners (0) in countries A, B and C. (Appendix 6)

Figures 1-4 exhibit graphs that illustrate the distribution of medals, athlete ages, training hours and medal counts by country. Figure 1 shows that more athletes did not win a medal than those who did, indicating an overall imbalance in medal-winning results. Figure 2 shows the age distribution of athletes which ranges from 19 to 28 years old, with a little peak around age 22, indicating a higher frequency of athletes at that age. Figure 3 shows training hours, revealing that athletes who train 15 to 17 hours per week are the most prevalent, while gold winners are more fairly spread, indicating that training hours do not ensure success. Finally, Figure 4 compares medal winners and non-winners in countries A, B, and C. Country A has the most athletes overall, but the proportion of athletes is somewhat greater in countries B and C. These graphs offer an overview of the major elements driving medal results and illustrate patterns in the dataset.

Model Selection and Interpretation

The logistic regression model's primary purpose is to predict an athlete's chances of winning a medal using a collection of factors. These predictors include Previous Medal (if the athlete has won a medal in the past), Top 10 (whether the athlete is rated in the top 10), Country (the athlete's home country), Main Sponsor, Age and Hours Training (the number of hours training per week). This research assists in identifying the most relevant aspects that contribute to medal performance.

Full Logistic Regression Model

The initial model includes all predictors:

$$\text{Logit}(P) = \beta0 + \beta1(\text{Previous Medal}) + \beta2(\text{Top 10}) + \beta3(\text{Country}) + \beta4(\text{Main Sponsor}) + \beta5(\text{Age}) + \beta6(\text{Hours Training})$$

The model was built using stepwise regression, which ensured simplicity and accuracy by eliminating variables depending on their contribution, as evaluated by the Akaike Information Criterion (AIC). A low AIC indicates a better balance of complexity and fit. Starting with the entire model, non-significant variables such as Country and Main Sponsor were excluded due to their high p-values and little influence. The resulting model is cost-effective, prioritises essential predictors, and enhances interpretability while lowering the danger of overfitting.

Table 3 is the Stepwise regression process showing the removal of non-specific predictors (e.g. Main Sponsor and Country) and the corresponding changes in AIC. The final model includes Previous Medal, Top 10, Age and Hours Training, indicating an improved fit. (Appendix 7)

Step	**Model**	**AIC**
1	Full model (all predictors)	2966.63
2	Removed Main Sponsor	2962.36
3	Removed Country	2960.95
Final	Previous Medal, Top 10, Age, Hours Training	2960.95

The stepwise regression process improved the logistic regression model's fit while lowering complexity by removing non-significant factors. The original model, which included Previous Medal, Top 10, Country, Main Sponsor, Age and Hours Training, had an AIC of 2966.63. The first phase was removing the Main Sponsor, which reduced the AIC to 2962.36. In the following step, Country was removed decreasing the AIC to 2960.95. The remaining predictors (Previous Medal, Top 10, Age and Hours Training) were kept since they greatly improved the model's fit. The final model, with an AIC of 2960.95, shows enhanced efficiency by preserving just significant predictors. By deleting non-significant variables, the

model becomes more interpretable and focuses on the main determinants impacting medal achievement.

Final Model

Logit(P) = -2.79218 + 0.48818(Previous Medal) + 0.51578(Top10) – 0.05419(Age) + 0.17683 (Hours Training)

Table 4 shows the final model's coefficients (Appendix 8)

Variable	Estimate	Std. Error	Z Value	Pr(>|z|)
Intercept	-2.79218	0.57184	-4.883	1.05e-06
Previous Medal (Yes)	0.48818	0.08852	5.515	3.49e-08
Top 10 (Yes)	0.51578	0.09404	5.485	4.14e-08
Age	-0.05419	0.01529	-3.544	0.000394
Hours Training	0.17683	0.02536	6.995	1.87e-12

Table 5 shows the metrics to assess model performance (Appendix 8)

Metric	Value	Degrees of Freedom
Null Deviance	3077.0	2348
Residual Deviance	2946.6	2339
AIC	2966.6	N/A

The final logistic regression model predicts the likelihood of winning a medal based on the 4 relevant predictors: previous medal tip 10, age, and training hours. The model equation indicates that winning a prior medal (0.48818) and ranking in the top 10 (0.51578) considerably improve the chance of earning a medal, with very significant p-values ($p < 0.001$). Younger athletes are somewhat more likely to win ($p = 0.000394$) since age has a minor negative influence (-0.05419). Training hours (0.17683) are a significant predictor of victory ($p = 1.87e-12$).

Model performance measurements help to confirm its efficacy. The Null deviation (3077.0) reflects the deviation of a model without predictors, but the Residual Deviance (2946.6) shows a decrease, demonstrating that the additional predictors enhance model fit. The AIC (2966.6) demonstrates a decent balance between model complexity and prediction accuracy, demonstrating its effectiveness. Overall, the modified model efficiently identifies critical parameters impacting medal-winning chances, hence enhancing interpretability while preserving good predictive ability.

Model Validation

Residual analysis

Residual analysis assesses the model's fit by spotting trends, outliers or deviations from assumptions. A Q-Q plot compares observed residuals to the theoretical normal distribution to assess model adequacy and identify potential problems.

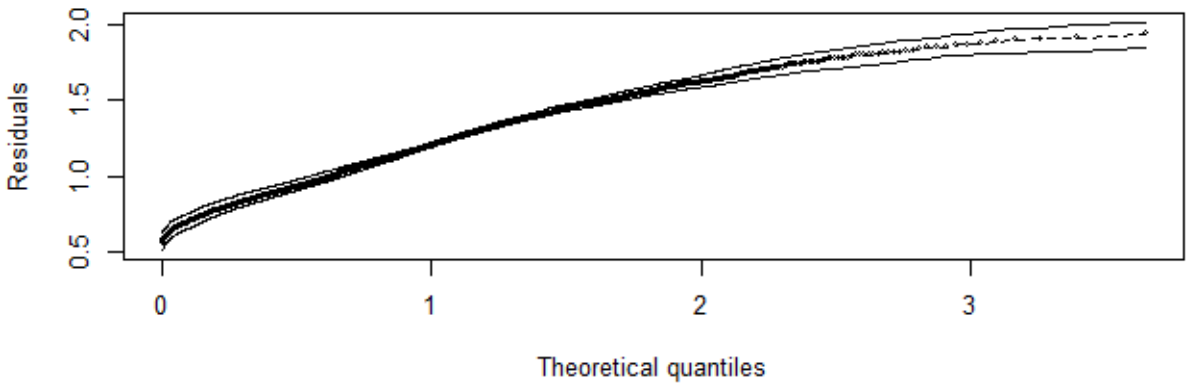


Figure 3 shows a Q-Q plot of residuals showing good overall fit as points align with the diagonal line, with minor deviations at the tails indicating discrepancies in extreme predictions. (Appendix 9)

The residual analysis examines the logistic regression model's goodness of fit. The Q-Q plot compares the residuals' theoretical and actual quantiles. The points in the centre area nearly follow the diagonal reference line, indicating that the residuals are roughly normally distributed and that the majority of the data fits well. However, deviations from the line at the tails suggest that the model's fit for extreme observations may be compromised, maybe die to outliers or limits in collecting unusual occurrences. While the model worked well for the majority of the data, the tail deviations may require further diagnostics to assure robustness.

Model Prediction

Case 1

This section shows the expected chances of winning a medal for athletes aged 19 to 28 years old from Country B who have never won a medal, are not ranked in the top 10, have Main Sponsor N and train 17 hours per week.

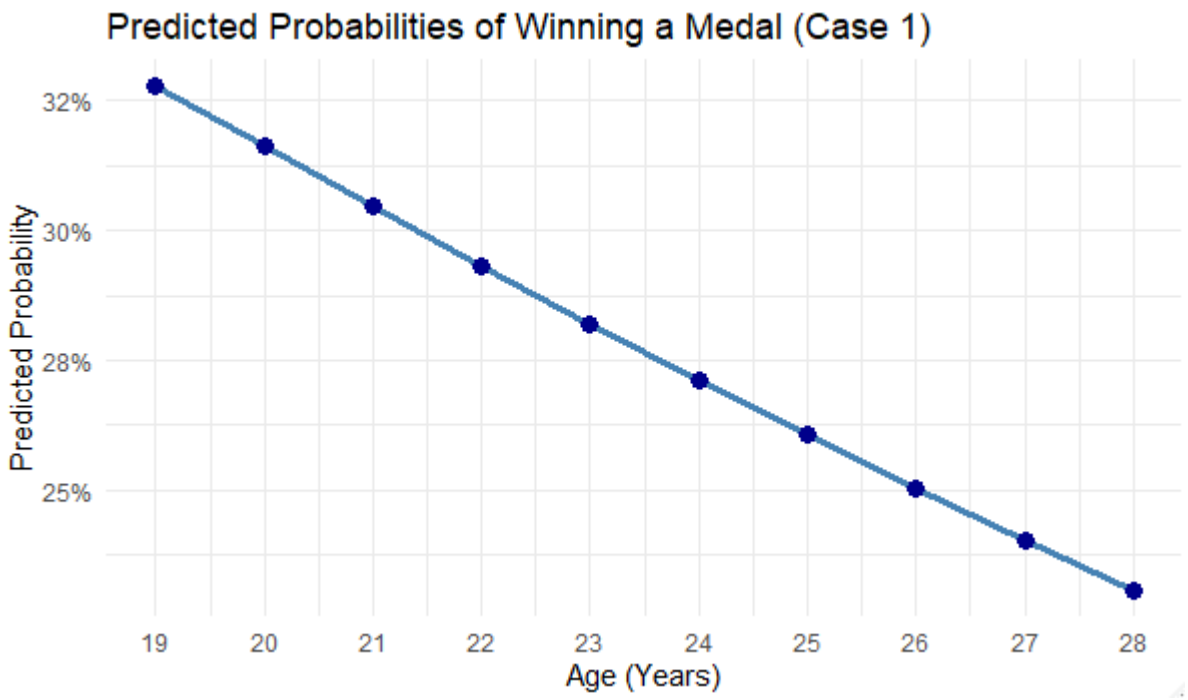


Figure 4 shows predicted probabilities of winning a medal for athletes aged 19 to 28 years (Case 1) – (Appendix 10 and Appendix 11)

Athletes meeting the parameters in Case 1 have a lower chance of earning a medal with age. Athletes aged 19 had the highest likelihood of winning (32.8%), which subsequently decreased to 23.1% by age 28. This trend shows that age has a negative influence on medal-winning chances, which is consistent with the findings of the final regression model, in which age was found to be a significant predictor with a negative effect on the probability of success.

Case 2

This section shows the predicted probabilities of winning a medal for athletes who meet the following conditions: Previous Medal (Yes), Top 10 (Yes), Main Sponsor (U), aged 23 years, and training between 14 and 20 hours per week. – (Appendix 12 and Appendix 13)

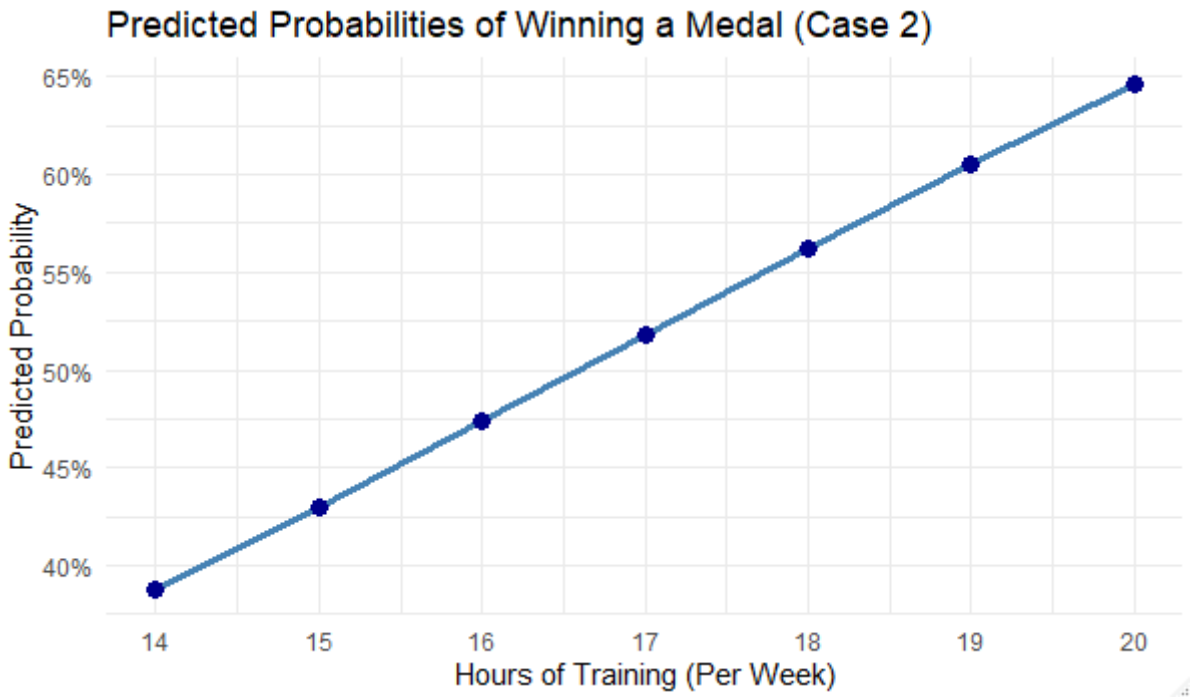


Figure 5 shows predicted probabilities of winning a medal for athletes training 14 to 20 hour per week (Case 2)

The likelihood of winning a medal grows as the amount of training hours increases. Athletes who train 14 hours per week have a 38.7% probability of earning a medal, which increases to 64.6% if they train 20 hours per week. This trend demonstrates the large positive impact of increased training hours on medal-winning chances, which is consistent with the model's conclusions that training is a strong predictor of success.

Conclusion

This report used logistic regression to investigate the factors that influence an athlete's chances of winning a medal. The final model uses Previous Medal, Top 10, Age and Training Hours as significant predictors, with training hours having the most beneficial influence. Model evaluation metrics, such as residual deviance and AIC, verified a decent fit while slight deviations in residual analysis indicated possible areas for improvement. The Case 1 and Case 2 predictions showed how younger athletes and those who practice for longer hours are more likely to succeed. While the model is useful, further enhancements might include more predictors or evaluating non-linear relationships to enhance accuracy and generalisability.

Appendix

Appendix 1

This is the R code Numerical Summary (Table 1 in the report).

```r
library(readxl) #For reading Excel files
library(ggplot2) #For visualisation


athletes_medal <- read_excel("Data Analytics/MAST5955- Predictive Modelling/CW/athletes_medal.xlsx")

#Summary statistics for numeric variables
summary(athletes_medal)

```

This was the output code which is in Table 1

```
> summary(athletes_medal)
     medal        previous_medal        top10             country
 Min.   :0.0000   Length:2349        Length:2349        Length:2349
 1st Qu.:0.0000   Class :character   Class :character   Class :character
 Median :0.0000   Mode  :character   Mode  :character   Mode  :character
 Mean   :0.3627
 3rd Qu.:1.0000
 Max.   :1.0000
 main_sponsor            age        hours_training
 Length:2349        Min.   :19.00   Min.   :14.00
 Class :character   1st Qu.:21.00   1st Qu.:15.40
 Mode  :character   Median :23.00   Median :17.00
                    Mean   :23.53   Mean   :16.97
                    3rd Qu.:26.00   3rd Qu.:18.50
                    Max.   :28.00   Max.   :20.00
>
```

Appendix 2

This is the R code for the Numerical Summary (Table 2 in the Report)

```r
# Count unique levels for each categorical variable
cat_vars <- c("previous_medal", "top10", "country", "main_sponsor")
for (var in cat_vars) {
  print(paste("Summary for", var, ":"))
  print(table(athletes_medal[[var]]))
}
```

This was the output code which is in Table 2

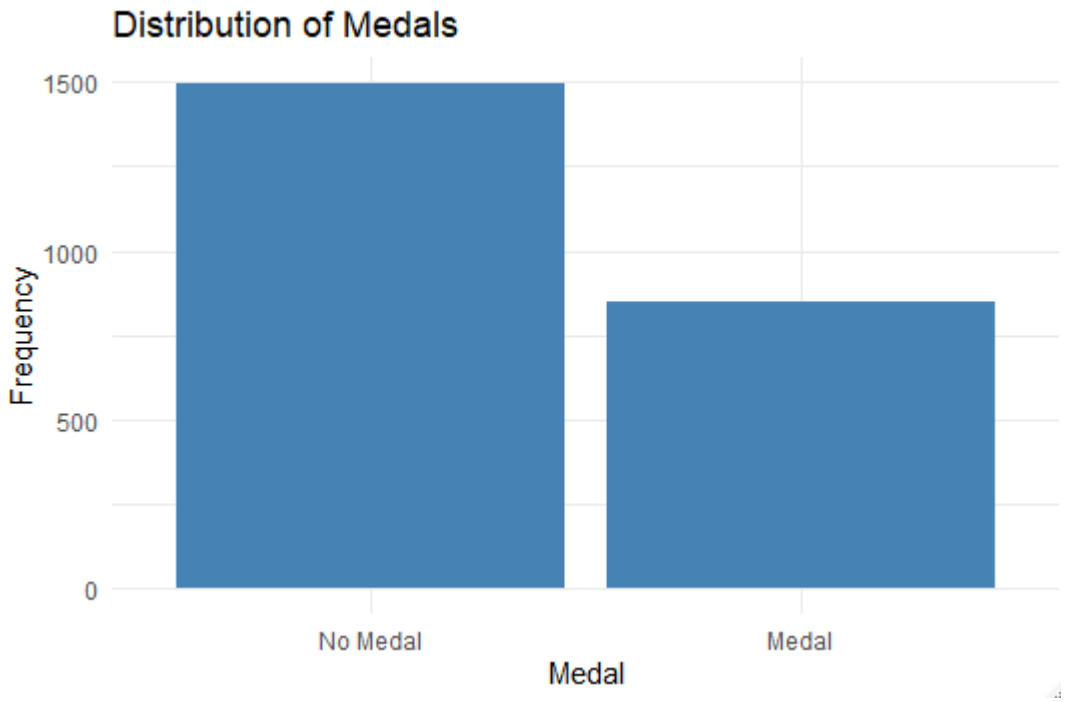
```
[1] "Summary for previous_medal :"

  no  yes
1313 1036
[1] "Summary for top10 :"

  no  yes
1637  712
[1] "Summary for country :"

  A   B   C
816 761 772
[1] "Summary for main_sponsor :"

  A   N   P   U
763 693 446 447
```

Appendix 3

This is the R code for Figure 1 in Graphical Summary

```
#Visualisation: Distribution of response variable (medal)
ggplot(athletes_medal, aes(x = factor(medal, labels = c("No Medal", "Medal")))) +
  geom_bar(fill = "steelblue") +
  theme_minimal() +
  labs(title = "Distribution of Medals", x = "Medal", y = "Frequency")

```

This is the output of this code

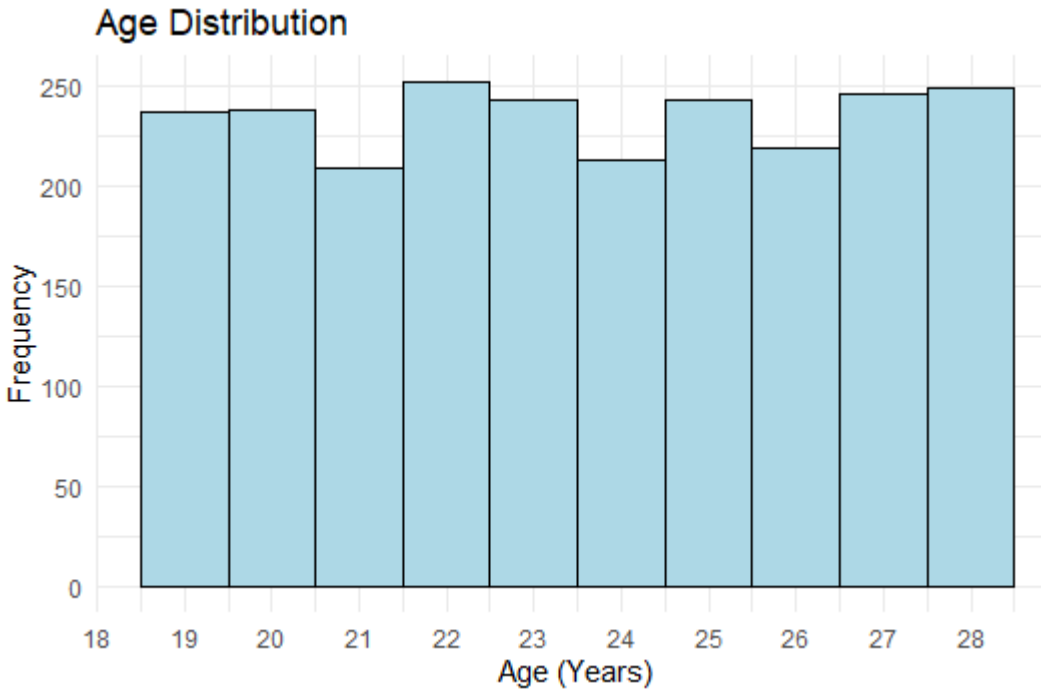


Appendix 4

This is the R code for Figure 2 in Graphical Summary:

```r
#Visuation: Age distribution
ggplot(athletes_medal, aes(x = age)) +
  geom_histogram(fill = "lightblue", bins = 10, col = "black") +
  scale_x_continuous(breaks = seq(18,28, by = 1)) +
  theme_minimal() +
  labs(title = "Age Distribution", x = "Age (Years)", y = "Frequency")

```

This is the output of the R code:

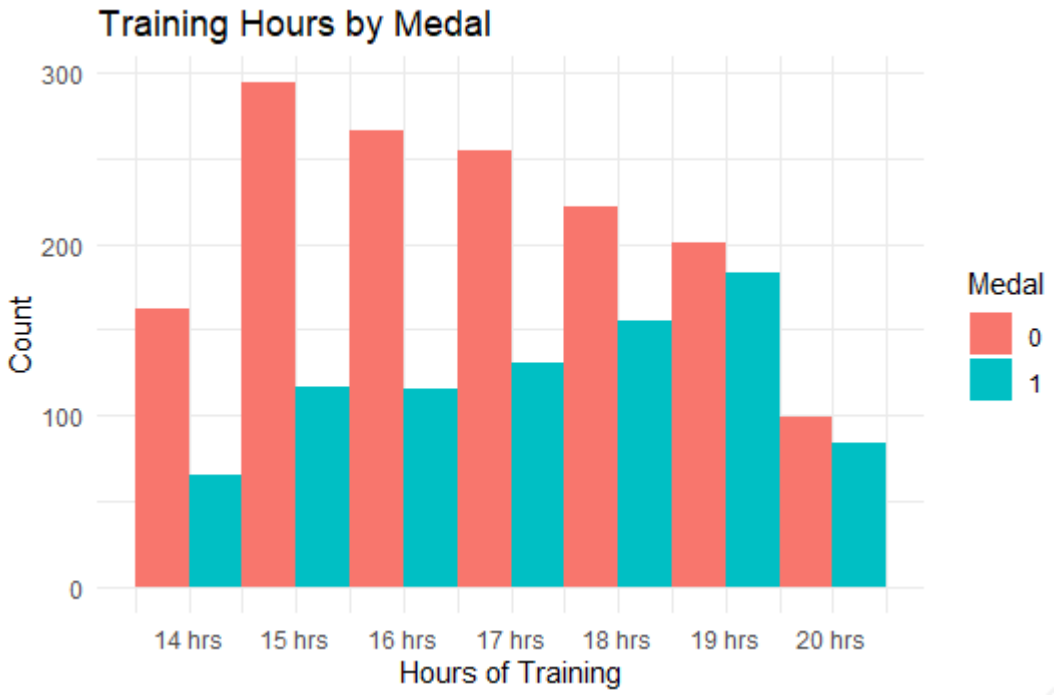


Appendix 5

This is the R code for Figure 3 in Graphical Summary

```r
#Visualisation: Training hours vs Medal
ggplot(athletes_medal, aes(x = hours_training, fill = factor(medal))) +
  geom_histogram(position = "dodge", binwidth = 1) +
  scale_x_continuous(
    breaks = seq(14,20, by = 1),
    labels = paste(seq(14, 20, by = 1), "hrs")
  ) +
  theme_minimal()+
  labs(title = "Training Hours by Medal", x = "Hours of Training", y = "Count", fill = "Medal")

```

This is the output of the R code:

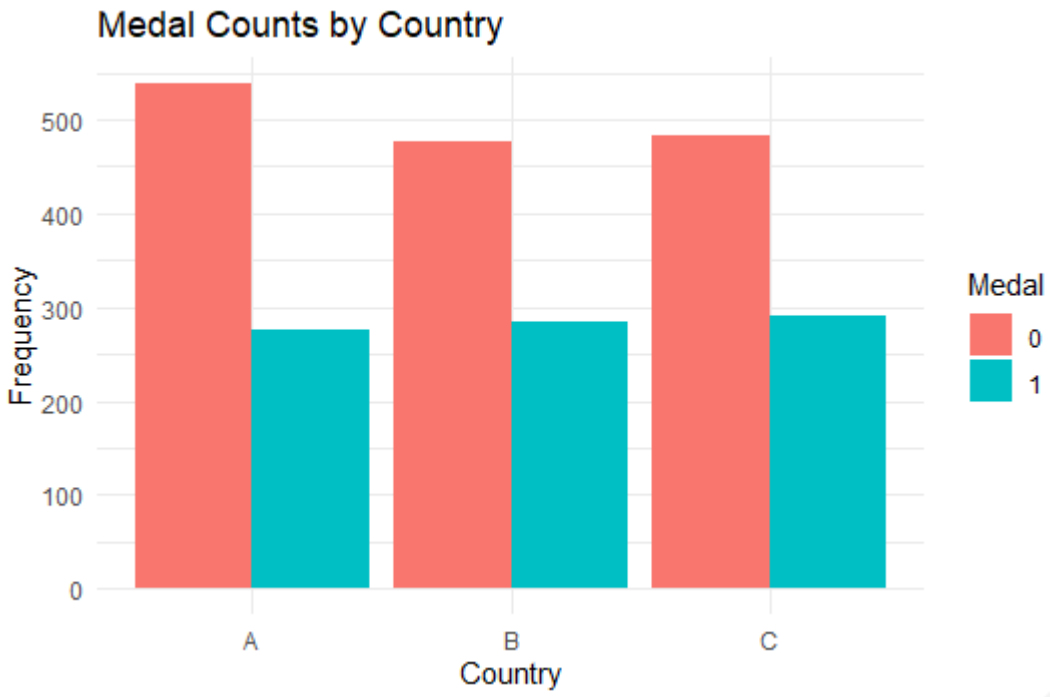


Appendix 6

This is the R code for Figure 4 in Graphical Summary

```r
#Visualisation: Medal counts by Country
ggplot(athletes_medal, aes(x = country, fill = factor(medal))) +
  geom_bar(position = "dodge") +
  scale_y_continuous(
    breaks = seq(0, 500, by = 100), #Define specific intervals for the y- axis
    labels = seq(0, 500, by = 100), #Ensure clear labeling
  )+
  theme_minimal() +
  labs(title = "Medal Counts by Country", x = "Country", y = "Frequency", fill = "Medal")

```

This is the output of the R code:



Appendix 7

This is the R code for the first regression model with all the predictors:

```
#Model Building
##Fit logistic regression model
full_model <- glm(medal ~ previous_medal + top10 + country + main_sponsor + age + hours_training,
                  family = binomial, data = athletes_medal)
model_step_aic <- step(full_model, direction = "both")
summary(model)
```

This is the output of the R code:

```
Start:  AIC=2966.63
medal ~ previous_medal + top10 + country + main_sponsor + age +
    hours_training

                 Df Deviance    AIC
- main_sponsor    3   2948.4 2962.4
- country         2   2949.3 2965.3
<none>                2946.6 2966.6
- age             1   2959.2 2977.2
- top10           1   2976.1 2994.1
- previous_medal  1   2977.2 2995.2
- hours_training  1   2997.3 3015.3

Step:  AIC=2962.36
medal ~ previous_medal + top10 + country + age + hours_training

                 Df Deviance    AIC
- country         2   2950.9 2960.9
<none>                2948.4 2962.4
+ main_sponsor    3   2946.6 2966.6
- age             1   2961.1 2973.1
- top10           1   2977.7 2989.7
- previous_medal  1   2978.8 2990.8
- hours_training  1   2999.0 3011.0


Step:  AIC=2960.95
medal ~ previous_medal + top10 + age + hours_training

                 Df Deviance    AIC
<none>                2950.9 2960.9
+ country         2   2948.4 2962.4
+ main_sponsor    3   2949.3 2965.3
- age             1   2963.6 2971.6
- top10           1   2980.9 2988.9
- previous_medal  1   2982.0 2990.0
- hours_training  1   3000.9 3008.9
> summary(model)
```

```
Call:
glm(formula = medal ~ previous_medal + top10 + country + main_sponsor +
    age + hours_training, family = binomial, data = athletes_medal)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -2.79218    0.57184  -4.883 1.05e-06 ***
previous_medalyes  0.48818    0.08852   5.515 3.49e-08 ***
top10yes           0.51281    0.09426   5.440 5.32e-08 ***
countryB           0.14966    0.10855   1.379 0.167971
countryC           0.15744    0.10782   1.460 0.144237
main_sponsorN     -0.08846    0.11237  -0.787 0.431177
main_sponsorP     -0.08534    0.12800  -0.667 0.504939
main_sponsorU      0.05853    0.12674   0.462 0.644205
age               -0.05420    0.01531  -3.540 0.000401 ***
hours_training     0.17863    0.02536   7.044 1.87e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3077.0  on 2348  degrees of freedom
Residual deviance: 2946.6  on 2339  degrees of freedom
AIC: 2966.6

Number of Fisher Scoring iterations: 4
```

Appendix 8

```
57  # Final logistic regression model
58  final_model <- glm(medal ~ previous_medal + top10 + age + hours_training,
59                     family = binomial, data = athletes_medal)
60  summary(final_model)
61
```

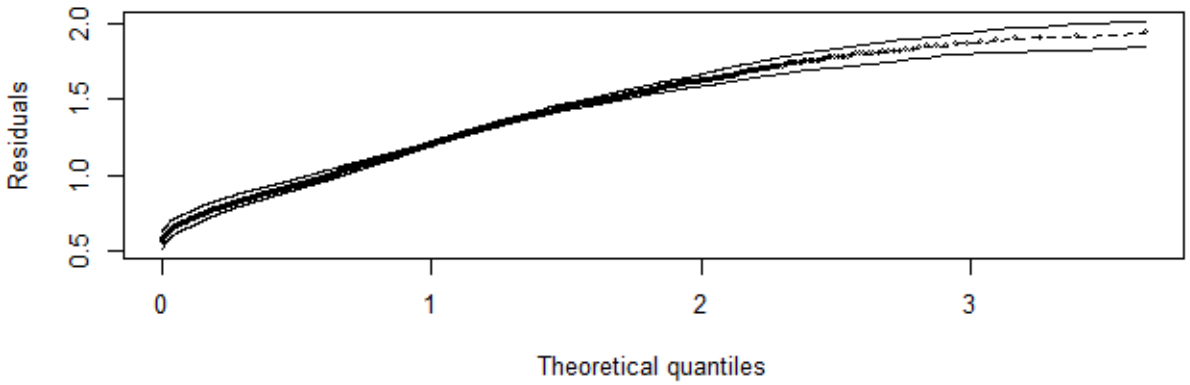
```
Step:  AIC=2960.95
medal ~ previous_medal + top10 + age + hours_training

                Df Deviance    AIC
<none>               2950.9 2960.9
+ country        2   2948.4 2962.4
+ main_sponsor   3   2949.3 2965.3
- age            1   2963.6 2971.6
- top10          1   2980.9 2988.9
- previous_medal 1   2982.0 2990.0
- hours_training 1   3000.9 3008.9
> summary(model)
```

Appendix 9

This is the r code for the residual analysis

```r
#Model validation
install.packages("hnp") #Install if not already installed
library(hnp)
hnp(final_model)
```

This is the output for residual Analysis (Figure 3)



Appendix 10

This is the R code for the Predictions of Case 1:

```r
#New Data for prediction
#Case 1: Athletes aged 19-28, country "B", no previous medal, not in top 10, main sponser "N", training 17 hours/week
case1 <- data.frame(
  previous_medal = factor("no", levels = c("no","yes")),
  top10 = factor("no", levels = c("no", "yes")),
  country = factor("B",levels = c("A","B","C")),
  main_sponsor = factor("N", levels = c("A", "N", "P", "U")),
  age = seq(19,28, by = 1),#Range of ages 19-28
  hours_training = 17 #Average training hours
  )

#Predict probability for Case 1
case1$predicted_probability <- predict(final_model, newdata = case1, type = "response")

print("Predicted Probabilties for Case 1:")
print(case1)

```

This is the output of prediction case 1

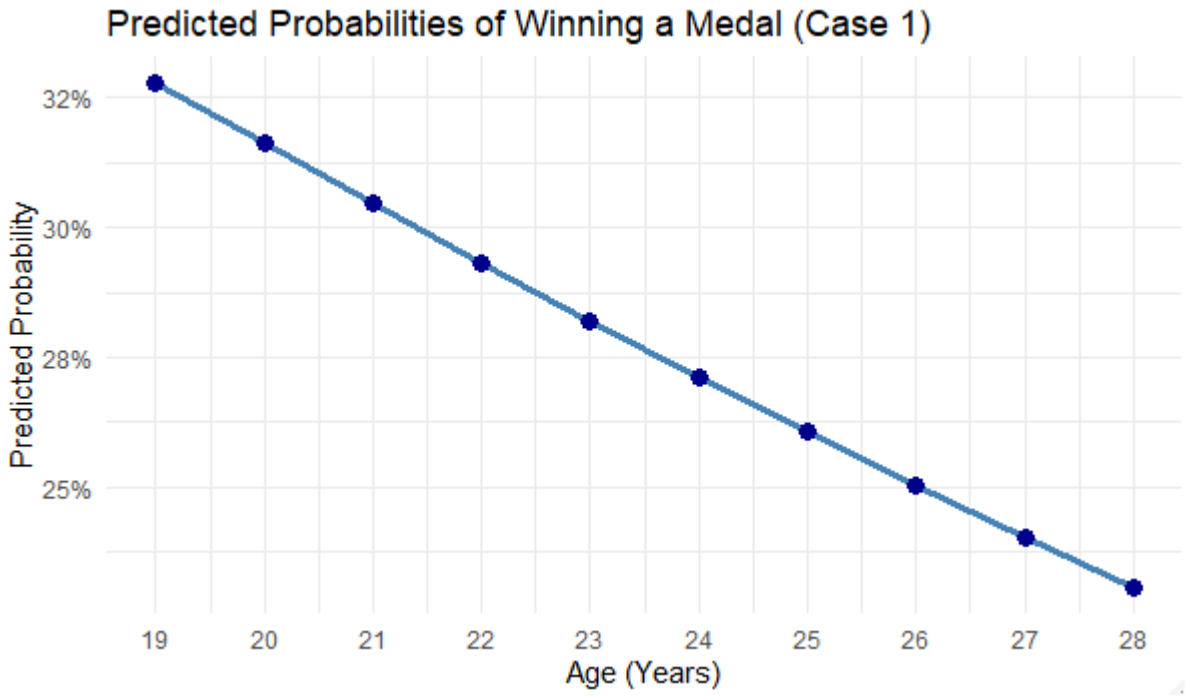
```
> #Predict probability for Case 1
> case1$predicted_probability <- predict(final_model, newdata = case1, type = "response")
>
> print("Predicted Probabilties for Case 1:")
[1] "Predicted Probabilties for Case 1:"
> print(case1)
   previous_medal top10 country main_sponsor age hours_training
1              no    no       B            N  19              17
2              no    no       B            N  20              17
3              no    no       B            N  21              17
4              no    no       B            N  22              17
5              no    no       B            N  23              17
6              no    no       B            N  24              17
7              no    no       B            N  25              17
8              no    no       B            N  26              17
9              no    no       B            N  27              17
10             no    no       B            N  28              17
   predicted_probability
1              0.3279523
2              0.3161227
3              0.3045264
4              0.2931731
5              0.2820715
6              0.2712289
7              0.2606517
8              0.2503454
9              0.2403141
10             0.2305612
```

Appendix 11

This is the R code for Prediction Case 1 with the use of Appendix 10 data:

```
#Plot predicted probabilities for Case 1
library(ggplot2)
ggplot(case1, aes(x = age, y = predicted_probability)) +
  geom_line(col = "steelblue", size = 1.2) + #Line plot for predicted probability
  geom_point(col = "darkblue", size = 3) + #Points for each age
  theme_minimal() +
  labs(
    title = "Predicted Probabilities of Winning a Medal (Case 1)",
    x = "Age (Years)",
    y = "Predicted Probability"
  ) +
  scale_x_continuous(breaks = seq(19,28,1)) + #Ensure all ages are shown
  scale_y_continuous(labels = scales::percent_format(accuracy = 1)) # show percentages

```

This is the output of prediction case 1 which is represented in figure 4 with the data in Appendix 10.



Appendix 12

This is the R code for Prediction Case 2

```
#Case 2: Athletes training 14-20 hours/week, country "A", previous medal, top 10, main sponsor "U", age 23
case2 <- data.frame(
  previous_medal = factor("yes", levels = c("no", "yes")),
  top10 = factor("yes", levels = c("no","yes")),
  country = factor("yes",levels = c("A","B","C")),
  main_sponsor = factor("U", levels = c("A","N","P","U")),
  age = 23,
  hours_training = seq(14, 20, by = 1) #Range of training hours
)

#Predict probabilities for Case 2
case2$predicted_probability <- predict(final_model, newdata = case2, type = "response")

print("Predicted Probabilities for Case 2")
print(case2)
```

This is the output of Prediction Case 2 (the data):

```
> print("Predicted Probabilities for Case 2")
[1] "Predicted Probabilities for Case 2"
> print(case2)
  previous_medal top10 country main_sponsor age hours_training predicted_probability
1            yes   yes    <NA>            U  23             14             0.3874458
2            yes   yes    <NA>            U  23             15             0.4301857
3            yes   yes    <NA>            U  23             16             0.4739922
4            yes   yes    <NA>            U  23             17             0.5182026
5            yes   yes    <NA>            U  23             18             0.5621299
6            yes   yes    <NA>            U  23             19             0.6051043
7            yes   yes    <NA>            U  23             20             0.6465131
> #Plot predicted probabilities for Case 2
```

Appendix 13

The is the R code for Prediction Case 2 which uses Appendix 12 data:

```
#Plot predicted probabilities for Case 2
library(ggplot2)

ggplot(case2, aes(x = hours_training, y = predicted_probability)) +
  geom_line(col = "steelblue", size = 1.2) + # Line plot for predicted probabilities
  geom_point(col = "darkblue", size = 3) + # Point for each training hour
  theme_minimal() +
  labs(
    title = "Predicted Probabilities of Winning a Medal (Case 2)",
    x = "Hours of Training (Per Week)",
    y = "Predicted Probability"
  ) +
  scale_x_continuous(breaks = seq(14,20,1)) + # Ensure all training hours are shown
  scale_y_continuous(labels = scales::percent_format(accuracy = 1)) # Show percentages
```

The output is presented as figure 5: