Main Focus:

What attributes drive an athlete to compete at the full IRONMAN distance (140.6 miles)

# IM.csv

This dataset contains 36,249 records and 17 variables from a sample of our 2017 athletes.

|  |  |
| --- | --- |
| Contact.Key | Athlete identifier |
| Resp | 1 = registered for full IRONMAN |
| R2016 | Races ran in 2016 |
| R2015 | Races ran in 2015 |
| R2014 | Races ran in 2014 |
| Age | Age of athlete |
| Finish.Time | Best finish time from historical races |
| Swim.Time | Best swim time from historical races |
| Bike.Time | Best bike time from historical races |
| Run.Time | Best run time from historical races |
| Min\_Year | First year of racing |
| prior\_races | Count of prior races of all time |
| club\_aff | 1 = affiliated with a tri club |
| Finish.Rank | Ranking of best finish time |
| Swim.Rank | Ranking of best swim time |
| Bike.Rank | Ranking of best bike time |
| Run.Rank | Ranking of best run time |

**Focus:** Main focus of the assignment is to use programming language(R) to extract actionable insights that are impactful for Ironman.

Assignment 1:

**Objective:**

The aim of this assignment is to come up with a classification model to find the factors that contribute towards full marathon participation.

**Defining Dataset:**

The dataset has 36249 observations and 17 variables. Targeted variable: Resp(registered for full IRONMAN or not), which have 30949 0’s and only 5300 1’s. Remaining variables are used as an input for prediction of full marathon participation.

**Data Preprocessing:**

* No **missing values** in the data
* **Outlier** treatment

We remove data points with Age < 1 years

and Time Variables (Finish, Run, Bike, Swim) with values = 999999 seconds

We are left with 26782 observations

* Create **buckets** for age(Age2) to identify the particular age groups to target.

Age 30 years and below => prime athletes

Age between 30 and 50 => seasoned athletes

Age above 50 => veteran

* **Remove variables** ‘Contact. Key’ as it **insignificant** to model building and ‘Age’ as we have Age2 variable containing buckets
* Convert necessary variables to **factors** with levels
* Int variables are **scaled** into range [0,1] i.e. normalization. Scaling is done to bring all int variables in the same range.
* Check **class bias**. Class bias is observed as we have 30949 0’s and only 5300 1’s.
* Create **training** and **test** samples

Training data is appropriately created with equal proportions of 0’s and 1’s to solve class bias problem. Remaining data is used for testing. Size of training data is smaller than test data, which is okay, because, we have large number of observations.

Training data has 6626 observations

Test data has 20156observations

**Model Selection:**

We choose Logistic model as we can solve the classification problem as well as get effect of eachinput variable on the output.

**Model Diagnostics:**

Call:

glm(formula = Resp ~ ., family = binomial(link = "logit"), data = trainingData)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.4707 -0.9453 0.0906 0.8941 3.9650

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 0.354044 0.715415 0.495 0.620686

R2016 28.829155 1.540343 18.716 < 2e-16 \*\*\*

R2015 15.031488 1.166331 12.888 < 2e-16 \*\*\*

R2014 10.516842 1.495309 7.033 2.02e-12 \*\*\*

Finish.Time -6.563941 1.589305 -4.130 3.63e-05 \*\*\*

Swim.Time 1.411944 1.078990 1.309 0.190677

Bike.Time 3.489325 1.522611 2.292 0.021925 \*

Run.Time 1.239209 1.549312 0.800 0.423801

Min\_Year2015 0.057307 0.217340 0.264 0.792030

Min\_Year2016 0.021803 0.205334 0.106 0.915435

Min\_Year2017 -0.657871 0.199445 -3.299 0.000972 \*\*\*

prior\_races -63.296664 2.679615 -23.622 < 2e-16 \*\*\*

club\_aff1 0.710454 0.066561 10.674 < 2e-16 \*\*\*

Finish.Rank2 -0.070398 0.161902 -0.435 0.663694

Finish.Rank3 -0.116043 0.214750 -0.540 0.588947

Finish.Rank4 -0.055200 0.263805 -0.209 0.834256

Finish.Rank5 0.051671 0.311255 0.166 0.868150

Finish.Rank6 -0.013172 0.366162 -0.036 0.971304

Finish.Rank7 0.160773 0.433063 0.371 0.710454

Finish.Rank8 0.509762 0.530818 0.960 0.336888

Swim.Rank2 -0.180760 0.150143 -1.204 0.228620

Swim.Rank3 0.075538 0.179499 0.421 0.673881

Swim.Rank4 0.135756 0.205945 0.659 0.509776

Swim.Rank5 0.046473 0.234152 0.198 0.842676

Swim.Rank6 0.048221 0.266001 0.181 0.856148

Swim.Rank7 0.119629 0.311087 0.385 0.700569

Swim.Rank8 0.098388 0.414301 0.237 0.812284

Bike.Rank2 0.103045 0.138471 0.744 0.456779

Bike.Rank3 0.249526 0.169272 1.474 0.140451

Bike.Rank4 0.049879 0.197223 0.253 0.800342

Bike.Rank5 0.153142 0.230042 0.666 0.505595

Bike.Rank6 0.060263 0.266824 0.226 0.821315

Bike.Rank7 -0.344253 0.317744 -1.083 0.278618

Bike.Rank8 -0.324746 0.408354 -0.795 0.426464

Run.Rank2 0.007217 0.138180 0.052 0.958347

Run.Rank3 0.274042 0.178219 1.538 0.124129

Run.Rank4 0.038593 0.217877 0.177 0.859404

Run.Rank5 0.204699 0.260792 0.785 0.432505

Run.Rank6 0.271384 0.311709 0.871 0.383954

Run.Rank7 0.151139 0.377416 0.400 0.688820

Run.Rank8 0.115444 0.489461 0.236 0.813542

Age2seasoned 0.228708 0.080991 2.824 0.004745 \*\*

Age2veteran 0.073557 0.100901 0.729 0.466000

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 9185.6 on 6625 degrees of freedom

Residual deviance: 7572.4 on 6583 degrees of freedom

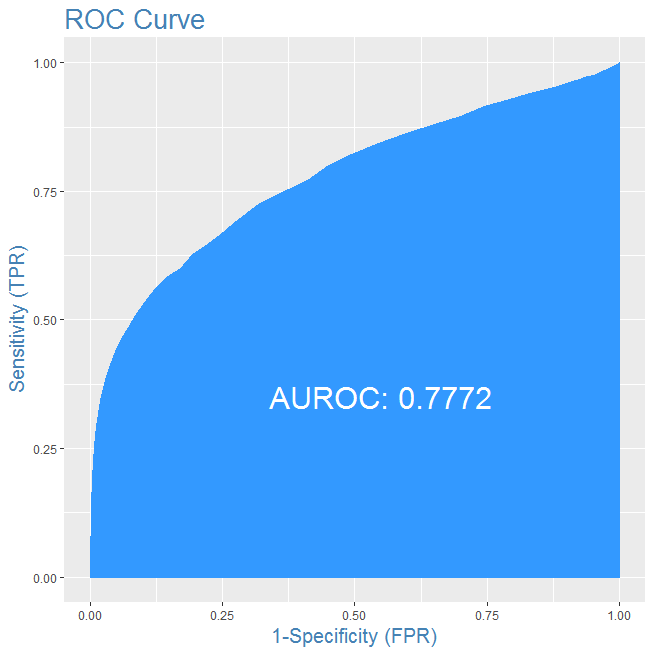
AIC: 7658.4

Number of Fisher Scoring iterations: 4

**Model Accuracy: 94%**

Model Accuracy is percentage match of actual and predicted values. We get a high accuracy on our test data.

**Receiver Operating Characteristics:**



ROC is generated by plotting True Positive Rate against False Positive Rate. The Area under the curve is 77% that represents good predictive ability of our model.

**Variable Importance:**

Overall

R2016 18.71605971

R2015 12.88783634

R2014 7.03322190

Finish.Time 4.13007061

Swim.Time 1.30857864

Bike.Time 2.29167247

Run.Time 0.79984436

Min\_Year2015 0.26367484

Min\_Year2016 0.10618555

Min\_Year2017 3.29851622

prior\_races 23.62155381

club\_aff1 10.67378620

Finish.Rank2 0.43481938

Finish.Rank3 0.54036243

Finish.Rank4 0.20924649

Finish.Rank5 0.16600884

Finish.Rank6 0.03597231

Finish.Rank7 0.37124664

Finish.Rank8 0.96033209

Swim.Rank2 1.20392033

Swim.Rank3 0.42082714

Swim.Rank4 0.65918581

Swim.Rank5 0.19847160

Swim.Rank6 0.18128025

Swim.Rank7 0.38455253

Swim.Rank8 0.23747999

Bike.Rank2 0.74416150

Bike.Rank3 1.47411457

Bike.Rank4 0.25290433

Bike.Rank5 0.66571233

Bike.Rank6 0.22585379

Bike.Rank7 1.08342935

Bike.Rank8 0.79525735

Run.Rank2 0.05222811

Run.Rank3 1.53767129

Run.Rank4 0.17713297

Run.Rank5 0.78491249

Run.Rank6 0.87063363

Run.Rank7 0.40045676

Run.Rank8 0.23585937

Age2seasoned 2.82387099

Age2veteran 0.72900333

**Results:**

Based on p values,coefficients and variable importance we get the top 9 variables contributing towards full marathon participation.

|  |  |  |  |
| --- | --- | --- | --- |
| Serial no. | Variable | Coefficients/Effect | Sign/Direction |
| 1 | prior\_races | -63.296664 | Negative |
| 2 | R2016 | 28.829155 | Positive |
| 3 | R2015 | 15.031488 | Positive |
| 4 | club\_aff1 | 0.710454 | Positive |
| 5 | R2014 | 10.516842 | Positive |
| 6 | Finish.Time | -6.563941 | Negative |
| 7 | Min\_Year2017 | -0.657871 | Negative |
| 8 | Age2seasoned | 0.228708 | Positive |
| 9 | Bike.Time | 3.489325 | Positive |

**Interpretation:**

* prior\_races and Finish.Time have significant(negative) impact on full marathon participation. While, variables R2016, R2015 and R2014 hasvesignificant(positive) impact on full marathon participation. These are the actual predictors and should be our prime focus.
* club\_aff1: When athlete is affiliated to a club, as compared to no affiliation, the odds of taking part in a full marathon will increase by 103.5%
* Min\_Year2017: When first year of racing is 2017 , as compared to when first year of racing is 2014, the odds of taking part in a full marathon will decrease by 48.2%
* Age2seasoned: When athlete lies in the seasoned age group , as compared to when lie in the prime age group, the odds of taking part in a full marathon will increase by 25.69%
* Bike.Time: When best bike time increases by 1 unit(seconds) the odds of taking part in a full marathon will increase by multiple of 32.7 times.

**Recommendations:**

Ideal conditions for an athlete to take part in full marathon

* Less prior races
* Low Best Finish team
* More number of races in the year 2014, 2015 and 2016
* Have club affiliation
* Not have first year of racing as 2017
* Should be in the seasoned age group (30 to 50 years)
* High Best Bike team

**Future Scope:**

I have tried all the different sampling techniques along with Logistic regression. Future scope of the assignment would be to try weighted Logistic Regression.