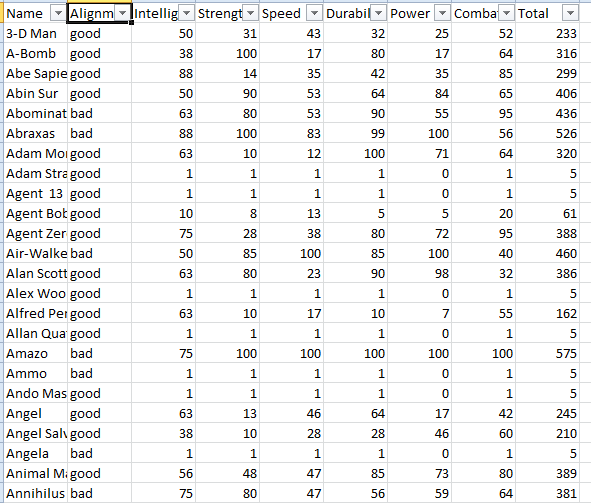
I used a dataset of 597 Marvel superheroes showing their Intelligence, Strength, Speed, Durability, Power and Combat in integer values. I downloaded the csv file from kaggle.com. The data fields includes: Name, Intelligence, Strength, Speed, Durability, Power and Combat and Total as shown below.



The alignment column in this dataset says whether the particular Marvel character is good or bad. Now, I am going to feed this dataset into logistic regression model and look if the Alignment field is dependent on the remaining fields i.e I am going to check whether Intelligence, Strength, Speed, Durability, Power and Combat values of a Marvel character can predict if the character is good or bad.

I removed few columns like Name and total which are not needed for this calculation.

Let me load this file into R studio,

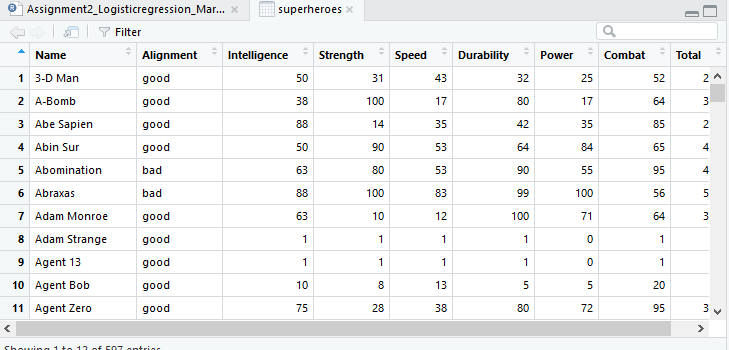
#Load the data

superheroes <- read.csv("C:/python/charcters\_stats.csv")

names(superheroes)

[1] "Name" "Alignment" "Intelligence" "Strength" "Speed"

[6] "Durability" "Power" "Combat" "Total"



#Split dataset into training and testing by count

nrow(superheroes)

[1] 597

I am going to consider 75% of the data (1 to 447 records) as training dataset and remaining 150 records as testing dataset.

limit <- floor((nrow(superheroes)/4)\*3)

training <- superheroes[1:limit, ]

testing <- superheroes[(limit+1):nrow(superheroes), ]

summary(superheroes)

Name Alignment Intelligence Strength Speed

Nova : 2 bad :165 Min. : 1.00 Min. : 1.00 Min. : 1.00

3-D Man : 1 good:432 1st Qu.: 1.00 1st Qu.: 1.00 1st Qu.: 1.00

A-Bomb : 1 Median : 50.00 Median : 10.00 Median : 23.00

Abe Sapien : 1 Mean : 44.21 Mean : 28.77 Mean : 26.96

Abin Sur : 1 3rd Qu.: 75.00 3rd Qu.: 53.00 3rd Qu.: 38.00

Abomination: 1 Max. :113.00 Max. :100.00 Max. :100.00

(Other) :590

Durability Power Combat Total

Min. : 1.00 Min. : 0.00 Min. : 1.00 Min. : 5.0

1st Qu.: 1.00 1st Qu.: 0.00 1st Qu.: 1.00 1st Qu.: 5.0

Median : 32.00 Median : 37.00 Median : 48.00 Median :254.0

Mean : 41.23 Mean : 39.93 Mean : 42.79 Mean :223.9

3rd Qu.: 80.00 3rd Qu.: 67.00 3rd Qu.: 70.00 3rd Qu.:349.0

Max. :120.00 Max. :100.00 Max. :101.00 Max. :581.0

#Fit a logistic regression model using training data

marvel\_model<-glm(Alignment~Intelligence+Strength+Speed+Durability+Power+Combat, data=training, family=binomial)

summary(marvel\_model)

Call:

glm(formula = Alignment ~ Intelligence + Strength + Speed + Durability +

Power + Combat, family = binomial, data = training)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.9418 -1.2537 0.7053 0.7938 1.3820

Coefficients:

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

(Intercept) 1.273473 0.191348 6.655 2.83e-11 \*\*\*

Intelligence -0.013711 0.005929 -2.313 0.0207 \*

Strength -0.013912 0.005441 -2.557 0.0106 \*

Speed 0.011172 0.006743 1.657 0.0976 .

Durability -0.000306 0.005742 -0.053 0.9575

Power 0.001946 0.005474 0.356 0.7222

Combat 0.007754 0.005509 1.408 0.1593

---

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 529.80 on 446 degrees of freedom

Residual deviance: 510.88 on 440 degrees of freedom

AIC: 524.88

Number of Fisher Scoring iterations: 4

After fitting the model with the training dataset, I see some good p-values for Intelligence, Strength and Speed. Let me use this Model to predict the testing dataset and create the confusion matrix

#Use the fitted model to do the predictions in the testing data

model\_pred\_prob<-predict(marvel\_model, testing, type="response")

model\_pred\_direction<-rep("bad", 150)

model\_pred\_direction[model\_pred\_prob>0.5]<-"good"

#Create a confusion matrix and compute misclassification rate

table(testing$Alignment, model\_pred\_direction)

mean(model\_pred\_direction!=testing$Alignment)

model\_pred\_direction

bad good

bad 2 38

good 5 105

[1] 0.2866667

Looks like a decent result and has predicted good characters very well and the mean is also 0.3 which is not bad.

**My inferences on this Logistic Regression Model are as follows:**

1. I have used Logistic regression Model for this dataset over any other model like linear because I am trying to predict a binary field i.e. whether a particular Marvel character is good or bad.
2. Parameters like Intelligence, Strength and Speed had pretty good p-values when compared to the other independent variables.
3. The Model predicted accurately 105 good characters out of 107.