

FIFA球員違約金之預測模型

INTRODUCTION TO DATA MINING






目錄




- 動機
- 分析
 - 資料處理
 - 繪圖
 - 迴歸模型
- 結論



動機

想知道什麼原因會影響球員違約金，我覺得原因可能會是年齡、薪資等...因此在Kaggle上找到FIFA的球員資料，利用線性迴歸模型找出其關聯性，並試著預測違約金的金額。



程式碼-資料處理

有關金額的欄位資料皆為文字型態且包含€和計量單位(M、K)，統一將其單位換成K轉換成數值型態。

由於違約金(Release_Clause)為要預測的項目，刪除其值為NA的資料。其他則用平均數填補NA值

```
df$Work_Rate<- sub("/.*", "", df$Work_Rate)
df$Wage <- gsub("[€KM]", "", df$Wage)
df$Wage <- as.numeric(df$Wage)

df$Release_Clause <- gsub("[€KM]", "", df$Release_Clause)
df$Release_Clause <- as.numeric(df$Release_Clause)
df$Release_Clause <- ifelse(grepl("M", fifa_data$Release_Clause),
df$Release_Clause * 1000, df$Release_Clause)
df$Release_Clause <- as.numeric(df$Release_Clause)

#處理NA值
table(is.na(df))

mean_value<-mean(df$Stamina, na.rm = TRUE)
df[which(is.na(df[, "Stamina"])), "Stamina"] <- mean_value

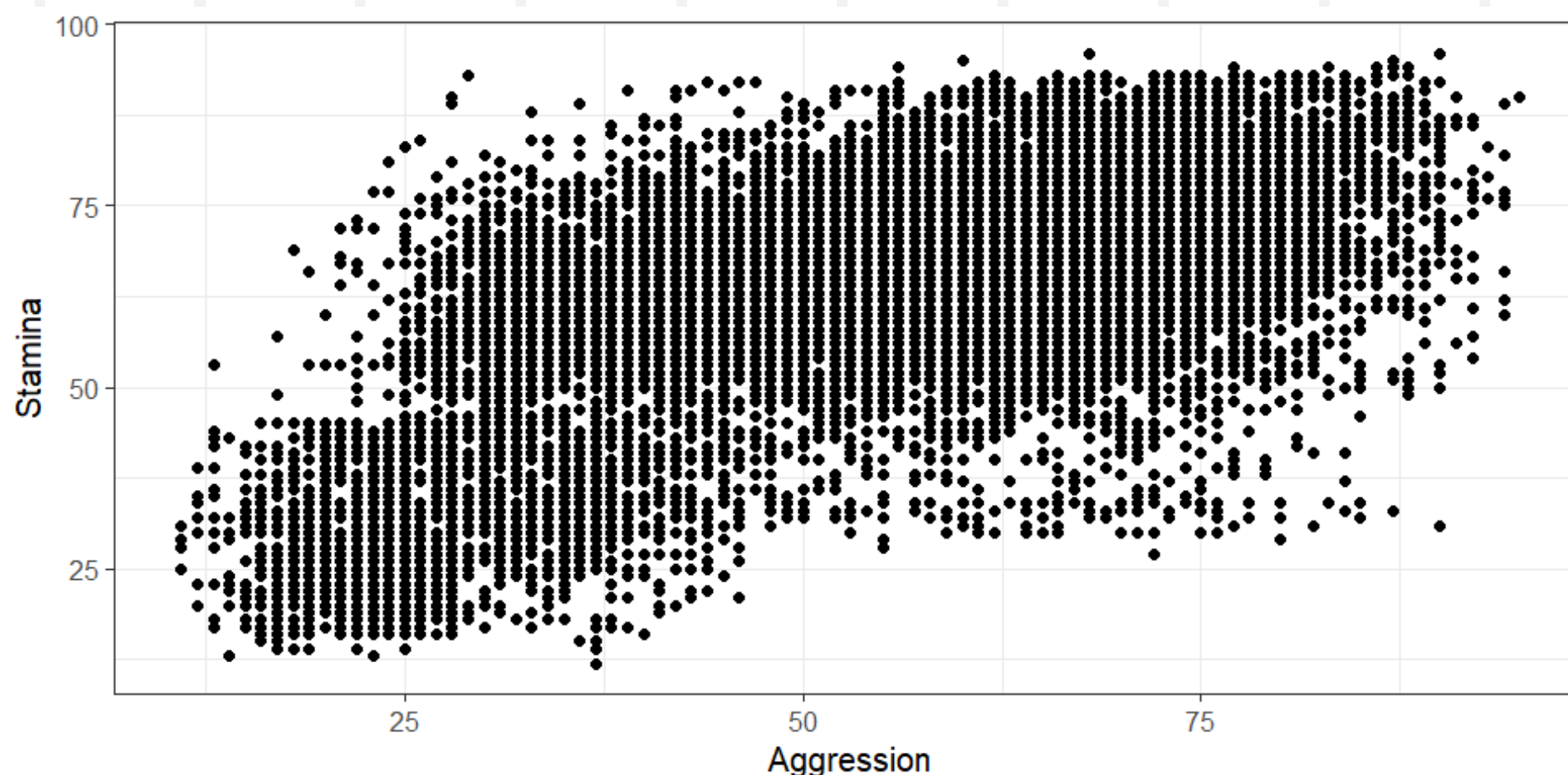
mean_value<-mean(df$Aggression, na.rm=TRUE)
df[which(is.na(df[, "Aggression"])), "Aggression"] <- mean_value

df<-na.omit(df)
```

程式碼-繪圖

畫圖找出兩者的關聯性，找尋隱藏的資訊。看適不適合做為線性迴歸的自變數和應變數。

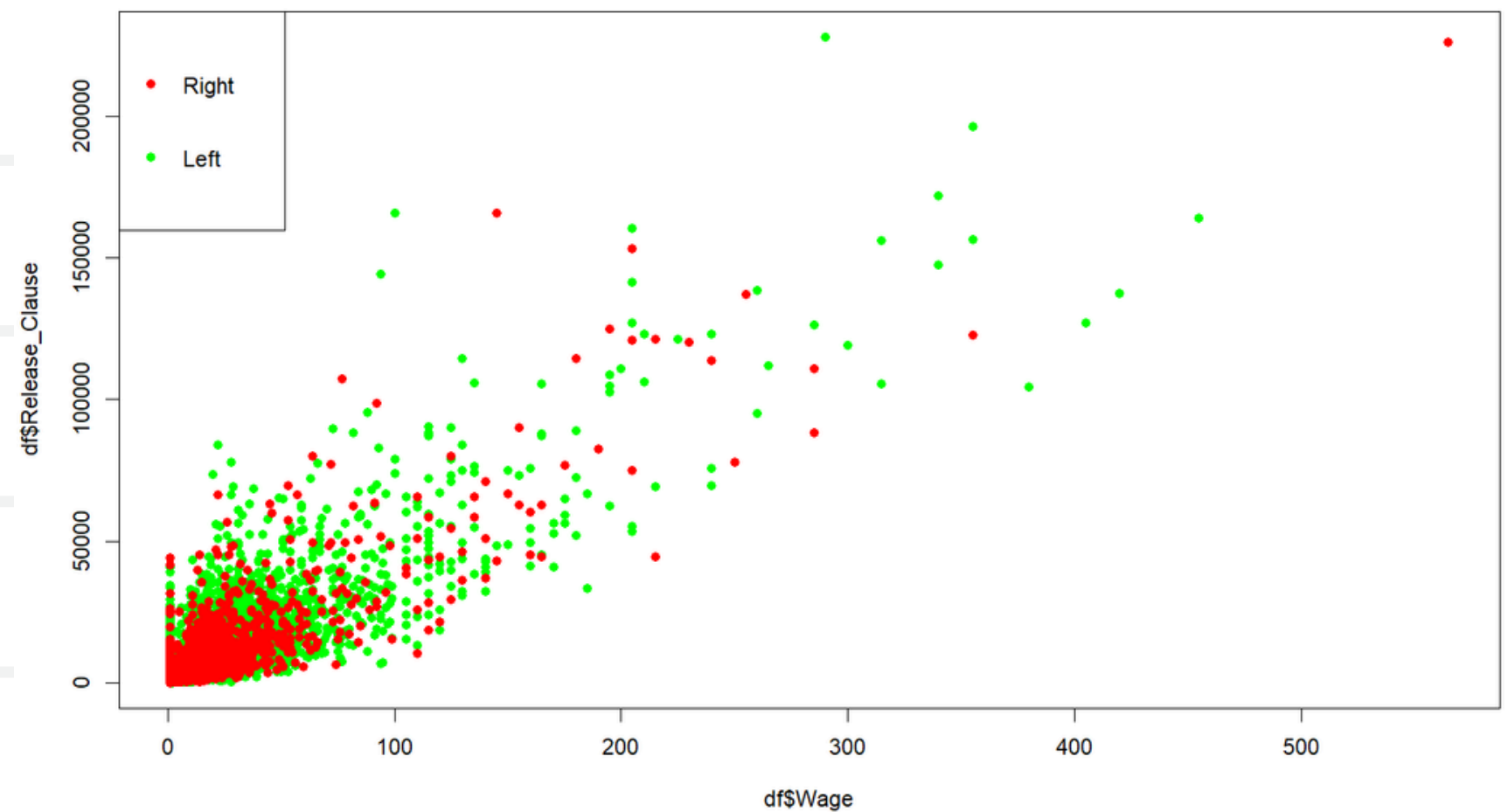
右圖是耐力和攻擊力，可看出兩者有線性關係大致呈現正相關，資料量大且分佈廣無法得到有用資訊。



程式碼-繪圖

右圖是薪資和違約金的散布圖，可以看出兩者存在線性關係。

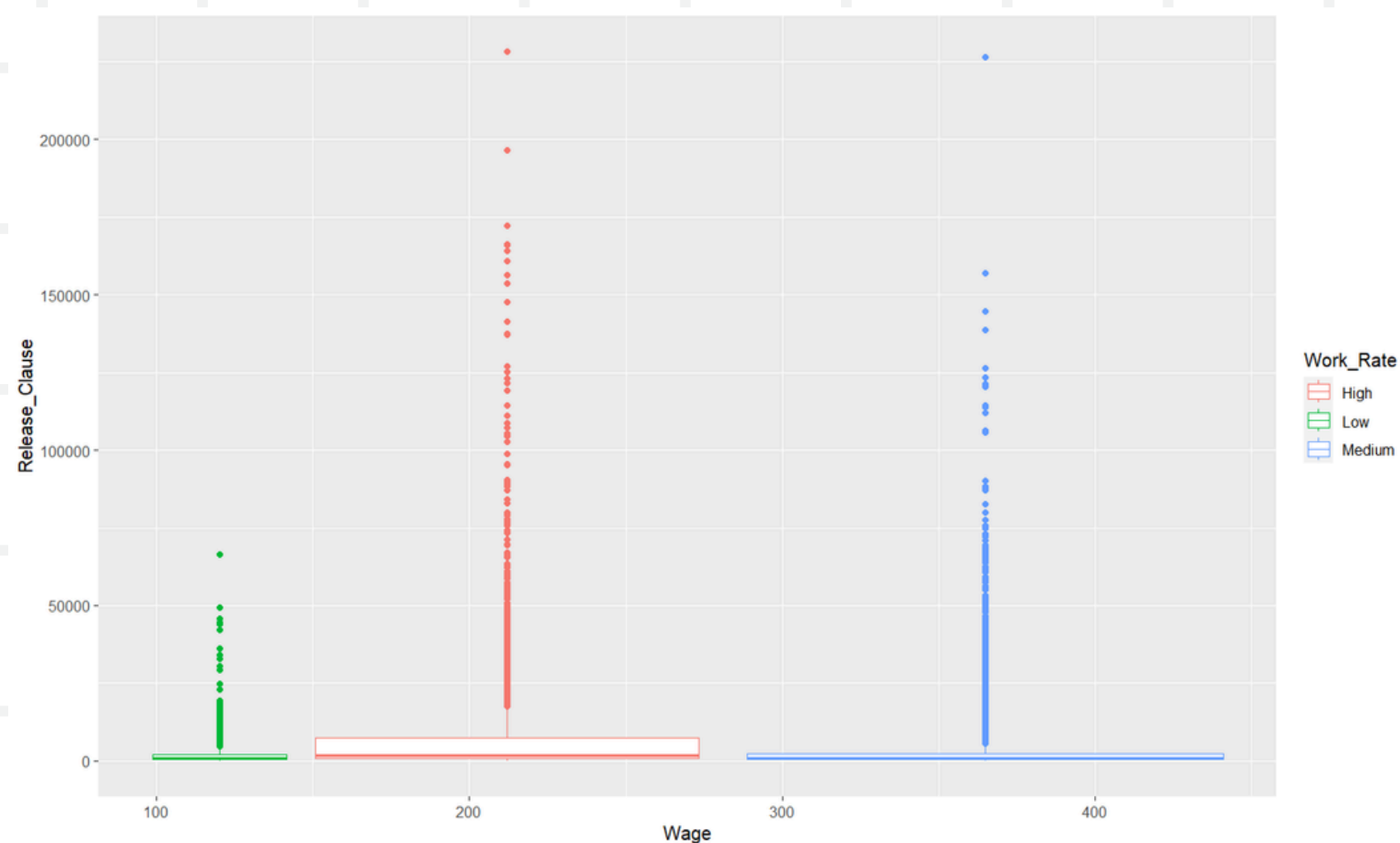
再分別以紅色和綠色作為左右腳分群，發現以右腳為慣用腳的選手較密集分佈在低薪資、低違約金的左下角。



程式碼-繪圖

右圖是不同努力程度下，薪資和違約金的盒鬚圖。違約金最大值出現在努力程度高的選手且努力程度中等的選手數量最多。

雖然三者之中位數皆相差不大，但努力程度越高，偏斜程度愈大，資料分散的程度也越大。



程式碼-迴歸模型

以違約金為依變數(Y)，薪資、年齡、耐力為自變數(X)建立模型進行迴歸分析。

三個自變數(X)的p-value皆小於0.05表示對依變數(Y)都達到顯著。R-squared=0.7404代表此模型的預測能力佳。

```
Call:
lm(formula = Release_Clause ~ Wage + Age + Stamina, data = df)

Residuals:
    Min       1Q   Median       3Q      Max
-58026  -1454   -683     361  121007

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2227.023    284.791     7.82  5.6e-15 ***
Wage          427.545      2.024    211.26 < 2e-16 ***
Age         -170.764      9.445    -18.08 < 2e-16 ***
Stamina       40.431      2.789     14.50 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5666 on 16639 degrees of freedom
Multiple R-squared:  0.7404,    Adjusted R-squared:  0.7403
F-statistic: 1.582e+04 on 3 and 16639 DF,  p-value: < 2.2e-16
```


程式碼-迴歸模型

利用殘差值檢驗是否符合三個假設：常態性、獨立性、同質性。由於母體資料大，因此隨機抽取500個樣本檢驗常態性。

可看出三者的p-value皆小於0.05，拒絕H0，符合以上三個性質。

```
> #常態性(Normality)
> #虛無假設H0:殘差服從常態分配
> #母體太大，隨機抽樣取500個樣本
> set.seed(123)
> sampled_data <- sample(mod$residual, size = 500)
> shapiro.test(sampled_data)
```

Shapiro-Wilk normality test

```
data:  sampled_data
W = 0.51924, p-value < 2.2e-16
```

```
>
> #獨立性(Independence)
> #虛無假設H0:殘差間相互獨立
> require(car)
> durbinWatsonTest(mod)
lag Autocorrelation D-W Statistic p-value
1 0.2337331 1.532113 0
Alternative hypothesis: rho != 0
>
> #變異數同質性(Homogeneity of Variance)
> #虛無假設H0:殘差變異數具有同質性
> require(car)
> ncvTest(mod)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 70166.3, Df = 1, p = < 2.22e-16
```

程式碼-迴歸模型

利用假設檢定檢驗慣用右腳和慣用左腳之選手的違約金是否有差異，設 H_0 為兩者違約金無差異。

分別建立迴歸模型做變異數分析，從Anova Table可以看出兩者的p-value皆小於0.05拒絕 H_0 ，兩者之間存在顯著差異。

```
> #變異數分析(ANOVA)Release_Clause有所差異
> #H0:  $\mu(\text{Right}) = \mu(\text{Left})$ 
> R.lm <- lm(Release_Clause~Preferred_Foot, data=df)
> anova(R.lm)
Analysis of Variance Table

Response: Release_Clause
              Df      Sum Sq   Mean Sq F value    Pr(>F)
Preferred_Foot    1 5.9246e+08 592455994   4.7934 0.02858 *
Residuals      16641 2.0568e+12 123597712
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> L.lm <- lm(Release_Clause~Preferred_Foot, data=df)
> anova(L.lm)
Analysis of Variance Table

Response: Release_Clause
              Df      Sum Sq   Mean Sq F value    Pr(>F)
Preferred_Foot    1 5.9246e+08 592455994   4.7934 0.02858 *
Residuals      16641 2.0568e+12 123597712
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

結論



用增加變數的方式逐步分析迴歸模型，可以看出有越多自變數，AIC值越低，則模型的預測能力越好。

從AIC值也可以看出解釋變數依序為薪資最佳，其次是耐力，最後為年齡。

```
> #逐步回歸
> mod <- lm(formula= Release_Clause ~ Wage + Age + Stamina,
+           data=df)
> full_mod<-formula(mod)
> mod1<-lm(Release_Clause~1,data=df)
> step(mod1,direction = "forward",scope = full_mod)
Start:  AIC=310106.2
Release_Clause ~ 1

      Df Sum of Sq  RSS   AIC
+ Wage   1 1.5072e+12 5.5014e+11 288156
+ Stamina 1 8.7386e+10 1.9700e+12 309386
+ Age     1 7.5955e+09 2.0498e+12 310047
<none>                 2.0574e+12 310106

Step:  AIC=288155.8
Release_Clause ~ Wage

      Df Sum of Sq  RSS   AIC
+ Age     1 9265991496 5.4087e+11 287875
+ Stamina 1 5521494314 5.4462e+11 287990
<none>                 5.5014e+11 288156

Step:  AIC=287875.1
Release_Clause ~ Wage + Age

      Df Sum of Sq  RSS   AIC
+ Stamina 1 6747650640 5.3413e+11 287668
<none>                 5.4087e+11 287875

Step:  AIC=287668.2
Release_Clause ~ Wage + Age + Stamina
```

結論



利用VIF(方差膨脹因子)找出自變數之間的相關程度，發現三者的VIF值皆大於1有共線性，但問題不算嚴重。

從前幾張的散布圖跟盒鬚圖可以看出此資料集存在極端值。若要加強模型的準確度，可以先處理極端資料再建立模型。

```
> #vif解釋變數之間相關程度
> vif(mod)
      Wage      Age  Stamina
1.052494 1.028941 1.040427
```

資料來源



🔍 **KAGGLE**
Sports Data Analysis

🔍 上課檔案