Presentazione Dataset e analisi per progetti di Gruppo

Rodolfo Metulini

2023-11-30

Come scrivere il report (consigli generali validi anche per il futuro):

1. Descrizione del problema e dei dati: descrivere qual è lo scopo dell'analisi e descrivere quante e quali variabili si hanno.
2. EDA (Analisi Esplorativa dei Dati): boxplot, grafico delle correlazioni, barplot, tabelle.
3. Data Engineering (opzionale): creare o aggiungere nuove variabili.
4. Missing imputation (non richiesto per questo Progetto) e outlier detection: spesso gli outlier vengono rimossi o tenuti da conto per l'analisi di un dataset.
5. Feature selection: decidere quali variabili entrano in gioco nell'analisi e perché, giustificare le proprie scelte o utilizzare modelli che aiutino a fare queste scelte.
6. Model selection: fare fitting di diversi modelli e confrontarli per scegliere il modello migliore per i vostri dati.
7. Conclusioni: concludere esplicitando a cosa è servito il modello finale e se avete risposto alla domanda di ricerca presentata al primo punto.
8. Bibliografia: è buona norma includere le fonti utilizzate per produrre il vostro report (paper di riferimento, ecc.).

BASKETBALL TEAMS' DATASET

```
# setwd("G:\\Il mio Drive\\01.Bergamo\\Didattica\\Statistica LT ING INF\\
# AA2324\\Parte 9 - Modello di regressione\\Progetti")
ds = read.delim("basketball_teams.txt")
head(ds)
```

Analisi delle variabili del dataset

year: stagione del torneo

lgID: nome della lega

franchID: nome squadra 3 cifre

confID: Conference di appartenenza della squadra

divID: Division di appartenenza della squadra

rank: classifica fine stagione regolare nella division

confRank: classifica fine stagione regolare nella conference

playoff: qualifica ai playoff (come da legenda)

name: nome squadra

o_: offensive

d_: defensive

fgm: tiri realizzati su azione

fga: tiri tentati su azione

ftm: tiri liberi realizzati

fta: tiri liberi tentati

3pm: tiri da 3 realizzati

3pa: tiri da 3 tentati

oreb: rimbalzi offensivi

dreb: rimbalzi difensivi

reb: rimbalzi totali (offensivi + difensivi)

asts: assists

pf: falli commessi

stl: palle recuperate

to: palle perse

blk: palle stoppate

tmRebound: rimbalzi di squadra

homeWon: vittore in casa

homeLost: sconfitte in casa

awayWon: vittorie in trasferta

awayLost: sconfitte in trasferta

neutWon: vittorie in campo neutro

neutLoss: sconfitte in campo neutro

confWon: vittorie contro squadre della stessa conference

confLoss: sconfitte contro squadre della stessa conference

divWon: vittorie con squadre stessa division

divLoss: sconfitte con squadre stessa division

won: vittorie totali

lost: sconfitte totali

games: partite giocate totali in stagione (won + lost)

min: minuti giocati in stagione

arena: nome stadio

attendance: numero spettatori

Progetto 1:

ABA league (1967-1975): VARIABILE DIPENDENTE: qualificazione ai playoff (SI/NO) e/o la vittoria di almeno la metà delle partite

COVARIATE: tutte le altre (o uno specifico insieme di queste, in base all'obiettivo di analisi)

attenzione che rimbalzi offensivi e difensivi, recuperi e stoppate, sono state rilevate solo a partire da un certo anno

Progetto 2:

NBA moderna (1976-2011): VARIABILE DIPENDENTE: qualificazione ai playoff (SI/NO)

COVARIATE: tutte le altre (o uno specifico insieme di queste, in base all'obiettivo di analisi)

Progetto 3:

NBA moderna (1976-2011): VARIABILE DIPENDENTE: qualificazione alle finali (SI/NO)

COVARIATE: tutte le altre (o uno specifico insieme di queste, in base all'obiettivo di analisi)

Progetto 4:

NBA moderna (1976-2011): VARIABILE DIPENDENTE: vittoria del titolo (SI/NO)

COVARIATE: tutte le altre (o uno specifico insieme di queste, in base all'obiettivo di analisi)

Progetto 5:

NBA moderna (1976-2011): VARIABILE DIPENDENTE: numero di vittorie in stagione

COVARIATE: tutte le altre (o uno specifico insieme di queste, in base all'obiettivo di analisi)

attenzione: qui considerare solo le squadre che hanno giocato 82 partite (dataset$games==82)

Progetto 6:

Confronto NBA moderna (1976-2011) e ABA (1967-1975): VARIABILE DIPENDENTE: qualificazione ai playoff (SI/NO)

COVARIATE: tutte le altre.

QUESITO: Le variabili che determinano la qualificazione ai playoff sono diverse nelle 2 leghe?

Possibili quesiti di ricerca

1. Quanto è importante avere una buona precisione nel realizzare (da 2, da 3, ai liberi)?
2. Quanto è importante che gli avversari tirino male?
3. Qual'è l'effetto di prendere più rimbalzi degli avversari e di recuperare molti palloni (stl, blk)
4. è importante saper vincere anche fuori casa per qualificarsi ai playoff?
5. è importante vincere contro le squadre delle propria division per vincere il titolo?
6. l'effetto delle variabili esplicative è cambiato nel corso dsel tempo? (dummy temporale)
7. ci sono arene percui giocare li aiuta a vincere o a qualificarsi ai playoff?

Operazioni utili

```
# select league
nba = ds[ds$lgID=="NBA",]
aba = ds[ds$lgID=="ABA",]
nbl = ds[ds$lgID=="NBL",]

# define the variable: "qualified for the playoff"
ds$DP_playoff = 0
ds$DP_playoff[ds$playoff !=""] = 1
```

```
# define the variable: "went to the finals"
ds$DP_playoff = 0
ds$DP_playoff[ds$playoff =="F"] = 1

# define the variable: "won the title"
ds$DP_playoff = 0
ds$DP_playoff[ds$playoff =="Won NBA championship"] = 1

# choose selected years
first = 2000 # first year
last = 2009 # last year
dsy = ds[ds$year >= first & ds$year <= last,]
```

Analisi preliminari

```
# Vediamo se ci sono valori mancanti nel dataset:
sum(is.na(ds))

# usiamo il summary pervedere in quali variabili
summary(ds)
# risulta minuti giocati
ds2 = ds[!is.na(ds$min),]
# usare quest'ultimo nel caso in cui la
# variabile "min" venga usata come esplicativa


# Dato che alcune variabili (lgID, franchisIDM CONFid)
# vengono considerate come character e non factor le trasformiamo:
str(ds)

ds$lgID <- as.factor(ds$lgID)
ds$franchID <- as.factor(ds$franchID)

summary(ds)
# questo ci permette di avere una migliore rappresentazione col summary
# e anche di poter generare variabili dummy

# sintesi dati fattori
table(ds$lgID)
table(ds$year, ds$lgID)



# Vediamo un pochino meglio il dataset ora:
summary(ds)
```

Analisi sulla variabile risposta

```
# Supponiamo la variabile risposta essere "won",
# per le stagione in cui le gare sono state 82

nba82 = ds[ds$lgID=="NBA" & ds$games==82,]

# histogram
hist(nba82$won)
#plot density
plot(density(nba82$won))

# buon adattamento normale
# altrimenti log trasfomation

plot(density(log(nba82$won)))
# meglio di no

library("corrplot")
M <- cor(as.matrix(nba82[,c(11:25,54)]))
         # correlation matrix
corrplot(M, method = 'number')
```

Analisi distribuzione variabili esplicative

```
# variabili fattori
boxplot(nba82$won ~ nba82$tmID, las=2)
```

Modello di regressione lineare

```
# stimo il modello
(res1 = lm(won ~ o_fgm, data = nba82))
(res2 = lm(won ~ o_fgm + o_fga, data = nba82))

# creo nuova variabile (percentuale di tiro dal campo)
nba82$o_fgpc = nba82$o_fgm / nba82$o_fga * 100
(res3 = lm(won ~ o_fgpc, data = nba82))

# dipende anche da percentuale tiro dal campo squadra avversaria
nba82$d_fgpc = nba82$d_fgm / nba82$d_fga * 100
(res4 = lm(won ~ o_fgpc + d_fgpc, data = nba82))

# dipende anche dai rimbalzi totali di dquadra e degli avversari
(res5 = lm(won ~ o_fgpc + d_fgpc + o_reb + d_reb, data = nba82))

# variabili fattori (dummy)
table(nba82$tmID)
#media vittorie in più rispetto ad Atlanta
```

```
(res6 = lm(won ~ o_fgpc + d_fgpc + tmID, data = nba82))

# Dean's 4 factors
nba82$o_df1 = (nba82$o_fgm + 0.5*nba82$o_3pm) / nba82$o_fga
nba82$d_df1 = (nba82$d_fgm + 0.5*nba82$d_3pm) / nba82$d_fga
nba82$o_df2 = nba82$o_to / (nba82$o_to + nba82$o_fga + 0.44*nba82$o_fta)
nba82$d_df2 = nba82$d_to / (nba82$d_to + nba82$d_fga + 0.44*nba82$d_fta)
nba82$o_df3 = nba82$o_oreb / (nba82$o_oreb + nba82$d_dreb)
nba82$d_df3 = nba82$d_oreb / (nba82$d_oreb + nba82$o_dreb)
nba82$o_df4 = nba82$o_ftm / nba82$o_fga
nba82$d_df4 = nba82$d_ftm / nba82$d_fga

(res7 = lm(won ~ o_df1 + d_df1 + o_df2 + d_df2 + o_df3 + d_df3 +
 o_df4 + d_df4, data= nba82))
summary(res7)
```