Moneyball: DATA621 Project #1

Team 2.

03/07/2021

Contents

File to submit.

0.1 Introduction

[add text intro here]

0.2 Setup

This analysis requires installation of `tidyverse`,`corrplot`, and `reshape2` [add other pckgs gere]

0.2.1 Load Raw Data

The data has been divided in advance into a training set, `training_raw`, and an evaluation set, `eval_raw`. The evaluation set does *not* contain the target information.

```
url1 = "https://raw.githubusercontent.com/Jagdish16/CUNY_DATA_621/main/project_1/moneyball-trai
training_raw = read_csv(url1)

url2 = "https://raw.githubusercontent.com/Jagdish16/CUNY_DATA_621/main/project_1/moneyball-eval
eval_raw=read_csv(url2)
```

0.2.2 Rename Columns

Rename the columns to be more human-readable.

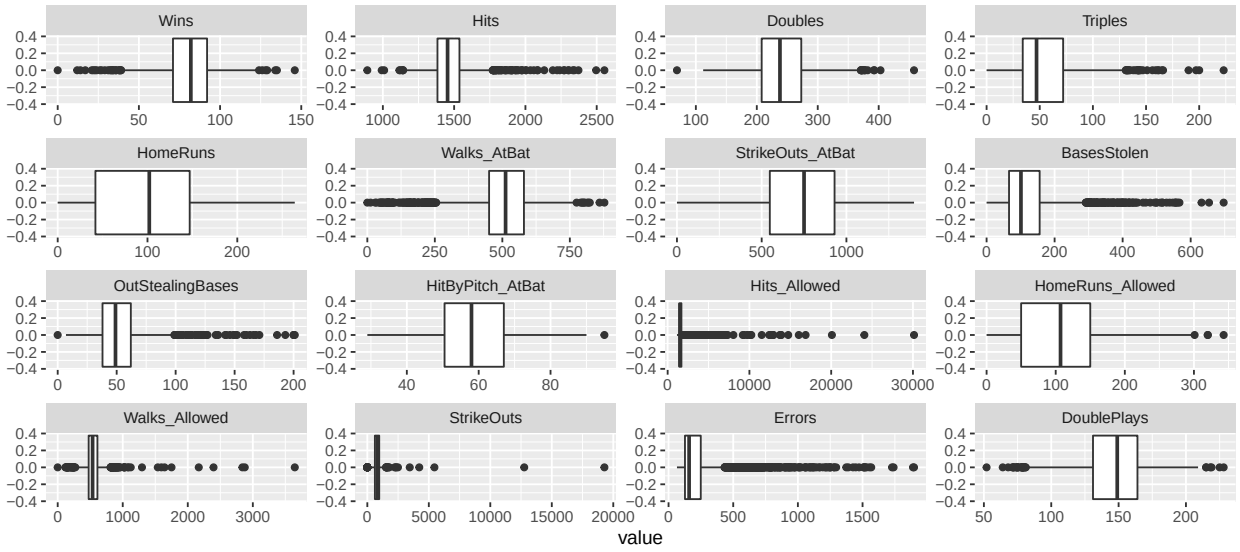
```
training <- training_raw %>%
  select(-INDEX) %>%
  rename_with(~ gsub("TEAM_", "", .x)) %>%
  rename_with(stringr::str_to_title) %>%
  dplyr::rename(
    Wins = Target_wins,
    Hits = Batting_h,
    Doubles = Batting_2b,
    Triples = Batting_3b,
    HomeRuns = Batting_hr,
    Walks_AtBat = Batting_bb,
    StrikeOuts_AtBat = Batting_so,
    BasesStolen = Baserun_sb,
    OutStealingBases = Baserun_cs,
    Hits_Allowed = Pitching_h,
    HitByPitch_AtBat = Batting_hbp,
    Errors = Fielding_e,
    HomeRuns_Allowed = Pitching_hr,
    Walks_Allowed = Pitching_bb,
    StrikeOuts = Pitching_so,
    DoublePlays = Fielding_dp
  )
```

0.3 Data Exploration

What does the data look like?

```
long <- training %>% as.data.frame() %>% melt

long %>%
  ggplot(aes(x=value)) +
  geom_boxplot() +
  facet_wrap(~variable, scales='free')
```



0.3.1 Missing cases

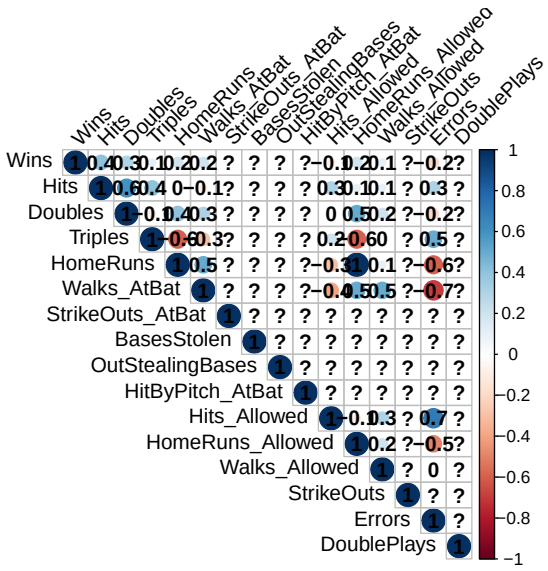
```
missing <- sapply(training, FUN = function(x) sum(is.na(x)))
missing
```

```
##             Wins             Hits          Doubles          Triples
##                0                0                0                0
##         HomeRuns      Walks_AtBat StrikeOuts_AtBat      BasesStolen
##                0                0              102              131
## OutStealingBases HitByPitch_AtBat     Hits_Allowed HomeRuns_Allowed
##              772             2085                0                0
##    Walks_Allowed       StrikeOuts           Errors      DoublePlays
##                0              102                0              286
```

0.3.2 Feature Correlation

```
# computing correlation matrix
corr <- round(cor(training), digits = 1)

corrplot(corr, type = 'upper', addCoef.col = 'black', tl.col = 'black', tl.srt = 45)
```



0.4 Model Testing

0.4.1 Handling of NA values

0.4.1.1 All features

0.4.1.2 All features with imputed values for NAs

0.4.2 Box-Cox Transform

0.4.3 Remove features

0.5 Model Selection & Conclusions

0.6 References

Sellmair, Reinhard. "How to handle correlated Features?" June 25, 2018. https://www.kaggle.com/reisel/how-to-handle-correlated-features

Xie, Yihui, J. J. Allaire, and Garrett Grolemund, *R Markdown: The Definitive Guide*, CRC Press-December 14, 2020 https://bookdown.org/yihui/rmarkdown/r-code.html.

https://rstatisticsblog.com/data-science-in-action/data-preprocessing/six-amazing-function-to-create-train-test-split-in-r/