DATA 621 HW 1

Business Analytics and Data Mining

Homework #1 Assignment Requirements

Overview

In this homework assignment, you will explore, analyze and model a data set containing approximately 2200 records. Each record represents a professional baseball team from the years 1871 to 2006 inclusive. Each record has the performance of the team for the given year, with all of the statistics adjusted to match the performance of a 162 game season.

Your objective is to build a multiple linear regression model on the training data to predict the number of wins for the team. You can only use the variables given to you (or variables that you derive from the variables provided).

Below is a short description of the variables of interest in the data set:

Variable Names	Definition	Theoretical Effect
INDEX	Identification Variable (do not use)	None
TARGET_WINS	Number of wins	$12
TEAM_BATTING_H	Base Hits by batters (1B,2B,3B,HR)	Positive Impact on Wins
TEAM_BATTING_2B	Doubles by batters (2B)	Positive Impact on Wins
TEAM_BATTING_3B	Triples by batters (3B)	Positive Impact on Wins
TEAM_BATTING_HR	Homeruns by batters (4B)	Positive Impact on Wins
TEAM_BATTING_BB	Walks by batters	Positive Impact on Wins
TEAM_BATTING_HBP	Batters hit by pitch (get a free base)	Positive Impact on Wins
TEAM_BATTING_SO	Strikeouts by batters	Negative Impact on Wins
TEAM_BASERUN_SB	Stolen bases	Positive Impact on Wins
TEAM_BASERUN_CS	Caught stealing	Negative Impact on Wins
TEAM_FIELDING_E	Errors	Negative Impact on Wins
TEAM_FIELDING_DP	Double Plays	Positive Impact on Wins
TEAM_PITCHING_BB	Walks allowed	Negative Impact on Wins
TEAM_PITCHING_H	Hits allowed	Negative Impact on Wins
TEAM_PITCHING_HR	Homeruns allowed	Negative Impact on Wins
TEAM_PITCHING_SO	Strikeouts by pitchers	Positive Impact on Wins

Deliverable:

- A write-up submitted in PDF format. Your write-up should have four sections. Each one is described below. You may assume you are addressing me as a fellow data scientist, so do not need to shy away from technical details.
- Assigned predictions (the number of wins for the team) for the evaluation data set.
- Include your R statistical programming code in an Appendix.

Write Up:

1. **DATA EXPLORATION (25 Points)** Describe the size and the variables in the moneyball training data set. Consider that too much detail will cause a manager to lose interest while too little detail will make the manager consider that you aren't doing your job. Some suggestions are given below. Please do NOT treat this as a check list of things to do to complete the assignment. You should have your own thoughts on what to tell the boss. These are just ideas.

a. Mean / Standard Deviation / Median
b. Bar Chart or Box Plot of the data
c. Is the data correlated to the target variable (or to other variables?)
d. Are any of the variables missing and need to be imputed "fixed"?

2. **DATA PREPARATION (25 Points)** Describe how you have transformed the data by changing the original variables or creating new variables. If you did transform the data or create new variables, discuss why you did this. Here are some possible transformations.

a. Fix missing values (maybe with a Mean or Median value)
b. Create flags to suggest if a variable was missing
c. Transform data by putting it into buckets
d. Mathematical transforms such as log or square root (or use Box-Cox)
e. Combine variables (such as ratios or adding or multiplying) to create new variables

3. **BUILD MODELS (25 Points)** Using the training data set, build at least three different multiple linear regression models, using different variables (or the same variables with different transformations). Since we have not yet covered automated variable selection methods, you should select the variables manually (unless you previously learned Forward or Stepwise selection, etc.). Since you manually selected a variable for inclusion into the model or exclusion into the model, indicate why this was done. Discuss the coefficients in the models, do they make sense? For example, if a team hits a lot of Home Runs, it would be reasonably expected that such a team would win more games. However, if the coefficient is negative (suggesting that the team would lose more games), then that needs to be discussed. Are you keeping the model even though it is counter intuitive? Why? The boss needs to know.
4. **SELECT MODELS (25 Points)** Decide on the criteria for selecting the best multiple linear regression model. Will you select a model with slightly worse performance if it makes more sense or is more parsimonious? Discuss why you selected your model. For the multiple linear regression model, will you use a metric such as Adjusted R2 , RMSE, etc.? Be sure to explain how you can make inferences from the model, discuss multi-collinearity issues (if any), and discuss other relevant model output. Using the training data set, evaluate the multiple linear regression model based on (a) mean squared error, (b) R2, (c) F-statistic, and (d) residual plots. Make predictions using the evaluation data set.

Evaluation

Load the data

```
df_train <- read.csv("https://raw.githubusercontent.com/melbow2424/Data621_HW1/main/moneyball-training-d

df_evaluation <- read.csv("https://raw.githubusercontent.com/melbow2424/Data621_HW1/main/moneyball-evalu
```

Review Data

```
skim(df_train)
```

Table 2: Data summary

Name	df_train
Number of rows	2276
Number of columns	17
Column type frequency:	
numeric	17
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
INDEX	0	1.00	1268.46	736.35	1	630.75	1270.5	1915.50	2535	
TARGET_WINS	0	1.00	80.79	15.75	0	71.00	82.0	92.00	146	
TEAM_BATTING_H	0	1.00	1469.27	144.59	891	1383.00	1454.0	1537.25	2554	
TEAM_BATTING_2B	0	1.00	241.25	46.80	69	208.00	238.0	273.00	458	
TEAM_BATTING_3B	0	1.00	55.25	27.94	0	34.00	47.0	72.00	223	
TEAM_BATTING_HR	0	1.00	99.61	60.55	0	42.00	102.0	147.00	264	
TEAM_BATTING_BB	0	1.00	501.56	122.67	0	451.00	512.0	580.00	878	
TEAM_BATTING_SO	102	0.96	735.61	248.53	0	548.00	750.0	930.00	1399	
TEAM_BASERUN_SB	131	0.94	124.76	87.79	0	66.00	101.0	156.00	697	
TEAM_BASERUN_CS	772	0.66	52.80	22.96	0	38.00	49.0	62.00	201	
TEAM_BATTING_HBP	2085	0.08	59.36	12.97	29	50.50	58.0	67.00	95	
TEAM_PITCHING_H	0	1.00	1779.21	1406.84	1137	1419.00	1518.0	1682.50	30132	
TEAM_PITCHING_HR	0	1.00	105.70	61.30	0	50.00	107.0	150.00	343	
TEAM_PITCHING_BB	0	1.00	553.01	166.36	0	476.00	536.5	611.00	3645	
TEAM_PITCHING_SO	102	0.96	817.73	553.09	0	615.00	813.5	968.00	19278	
TEAM_FIELDING_E	0	1.00	246.48	227.77	65	127.00	159.0	249.25	1898	
TEAM_FIELDING_DP	286	0.87	146.39	26.23	52	131.00	149.0	164.00	228	

Get the Means of columns in Training Data

```
train_means<-sapply(df_train, function(x) round(mean(x, na.rm = TRUE)))
train_means
```

```
##            INDEX      TARGET_WINS   TEAM_BATTING_H  TEAM_BATTING_2B
##             1268               81             1469              241
##   TEAM_BATTING_3B  TEAM_BATTING_HR  TEAM_BATTING_BB  TEAM_BATTING_SO
##               55              100              502              736
##   TEAM_BASERUN_SB  TEAM_BASERUN_CS TEAM_BATTING_HBP  TEAM_PITCHING_H
##              125               53               59             1779
## TEAM_PITCHING_HR TEAM_PITCHING_BB TEAM_PITCHING_SO  TEAM_FIELDING_E
##              106              553              818              246
## TEAM_FIELDING_DP
##              146
```

Get the Medians of columns in training data

```
train_medians<-sapply(df_train, function(x) round(median(x, na.rm = TRUE)))
train_medians
```

```
##            INDEX      TARGET_WINS   TEAM_BATTING_H  TEAM_BATTING_2B
##             1270               82             1454              238
##  TEAM_BATTING_3B  TEAM_BATTING_HR  TEAM_BATTING_BB  TEAM_BATTING_SO
##               47              102              512              750
##  TEAM_BASERUN_SB  TEAM_BASERUN_CS TEAM_BATTING_HBP  TEAM_PITCHING_H
##              101               49               58             1518
## TEAM_PITCHING_HR TEAM_PITCHING_BB TEAM_PITCHING_SO  TEAM_FIELDING_E
##              107              536              814              159
## TEAM_FIELDING_DP
##              149
```

Replace NA values in columns with their respective Mean

```
# Replace NA values in 'column_name' with 'mean'
df_train <- df_train %>%
  mutate(TEAM_BATTING_SO =
           ifelse(is.na(TEAM_BATTING_SO),
                  train_means[8],TEAM_BATTING_SO))%>%
  mutate(TEAM_BASERUN_SB =
           ifelse(is.na(TEAM_BASERUN_SB),
                  train_means[9], TEAM_BASERUN_SB))%>%
  mutate(TEAM_BASERUN_CS =
           ifelse(is.na(TEAM_BASERUN_CS),
                  train_means[10], TEAM_BASERUN_CS))%>%
  mutate(TEAM_BATTING_HBP =
           ifelse(is.na(TEAM_BATTING_HBP),
                  train_means[11],TEAM_BATTING_HBP))%>%
  mutate(TEAM_PITCHING_SO =
           ifelse(is.na(TEAM_PITCHING_SO),
                  train_means[15], TEAM_PITCHING_SO))%>%
  mutate(TEAM_FIELDING_DP =
           ifelse(is.na(TEAM_FIELDING_DP),
                  train_means[17], TEAM_FIELDING_DP))
```

Replace NA values with their respective Medians

```
# Replace NA values in 'column_name' with 'median'
# df_train <- df_train %>%
#   mutate(TEAM_BATTING_SO =
#            ifelse(is.na(TEAM_BATTING_SO),
#                   train_medians[8],TEAM_BATTING_SO))%>%
#   mutate(TEAM_BASERUN_SB =
#            ifelse(is.na(TEAM_BASERUN_SB),
#                   train_medians[9], TEAM_BASERUN_SB))%>%
#   mutate(TEAM_BASERUN_CS =
#            ifelse(is.na(TEAM_BASERUN_CS),
#                   train_medians[10], TEAM_BASERUN_CS))%>%
#   mutate(TEAM_BATTING_HBP =
#            ifelse(is.na(TEAM_BATTING_HBP),
#                    train_medians[11],TEAM_BATTING_HBP))%>%
#   mutate(TEAM_PITCHING_SO =
#            ifelse(is.na(TEAM_PITCHING_SO),
#                    train_medians[15], TEAM_PITCHING_SO))%>%
#   mutate(TEAM_FIELDING_DP =
#            ifelse(is.na(TEAM_FIELDING_DP),
#                    train_medians[17], TEAM_FIELDING_DP))
```

Note:

```
While deciding on whether to use 'Mean' or 'Median' both
codes were generated. Unused replacement is left
commented out, since only one can be applied at a time.
```

```
skim(df_train)
```

Table 4: Data summary

Name	df_train
Number of rows	2276
Number of columns	17
Column type frequency:	
numeric	17
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
INDEX	0	1	1268.46	736.35	1	630.75	1270.5	1915.50	2535	
TARGET_WINS	0	1	80.79	15.75	0	71.00	82.0	92.00	146	
TEAM_BATTING_H	0	1	1469.27	144.59	891	1383.00	1454.0	1537.25	2554	
TEAM_BATTING_2B	0	1	241.25	46.80	69	208.00	238.0	273.00	458	
TEAM_BATTING_3B	0	1	55.25	27.94	0	34.00	47.0	72.00	223	
TEAM_BATTING_HR	0	1	99.61	60.55	0	42.00	102.0	147.00	264	
TEAM_BATTING_BB	0	1	501.56	122.67	0	451.00	512.0	580.00	878	
TEAM_BATTING_SO	0	1	735.62	242.89	0	556.75	736.0	925.00	1399	
TEAM_BASERUN_SB	0	1	124.78	85.23	0	67.00	106.0	151.00	697	
TEAM_BASERUN_CS	0	1	52.87	18.66	0	44.00	53.0	54.25	201	
TEAM_BATTING_HBP	0	1	59.03	3.75	29	59.00	59.0	59.00	95	
TEAM_PITCHING_H	0	1	1779.21	1406.84	1137	1419.00	1518.0	1682.50	30132	
TEAM_PITCHING_HR	0	1	105.70	61.30	0	50.00	107.0	150.00	343	
TEAM_PITCHING_BB	0	1	553.01	166.36	0	476.00	536.5	611.00	3645	
TEAM_PITCHING_SO	0	1	817.74	540.54	0	626.00	818.0	957.00	19278	
TEAM_FIELDING_E	0	1	246.48	227.77	65	127.00	159.0	249.25	1898	
TEAM_FIELDING_DP	0	1	146.34	24.52	52	134.00	146.0	161.25	228	

Reference

- "Pythagorean Theorem of Baseball." Baseball Reference, https://www.baseball-reference.com/bullpen/Pythagorean_Theorem_of_Baseball. Accessed 11 September 2023.
- No author listed. "Pythagorean Expectation in Major League Baseball." Digital Commons @ Cal Poly, https://digitalcommons.calpoly.edu/cgi/viewcontent.cgi?article=1067&context=statsp. Accessed 11 September 2023.