2020 NFL Predictions Analysis

Your Name

2024-06-21

##A Description of the Data

The dataset I am working with contains game-by-game Elo ratings for the 2020 NFL season. This data IS from two sheets in an Excel workbook that we will combine into one. Our analysis aims to clean and explore the data, identify any errors, and produce visualizations to aid in understanding the data.

##Read the Data into R

##Combining Datafiles

##Cleaning the Data ##Data Inspection

# Display the first few rows of the combined dataset  
head(combined\_data)

## # A tibble: 6 × 30  
## date season neutral playoff team1 team2 elo1\_pre elo2\_pre elo\_prob1 elo\_prob2  
## <dbl> <dbl> <dbl> <chr> <chr> <chr> <chr> <dbl> <dbl> <dbl>  
## 1 44084 2020 0 <NA> KC Hous… 1664.84… 1528. 0.762 0.238  
## 2 44087 2020 0 <NA> MIN GB 1571.14… 1582. 0.577 0.423  
## 3 44087 2020 0 <NA> BUF NYJ 1511.35… 1458. 0.664 0.336  
## 4 44087 2020 0 <NA> ATL SEA 1534.51… 1547. 0.575 0.425  
## 5 44087 2020 0 <NA> JAX IND 1438.50… 1483. 0.530 0.470  
## 6 44087 2020 0 <NA> BAL CLE 1638.43… 1441. 0.820 0.180  
## # ℹ 20 more variables: elo1\_post <dbl>, elo2\_post <chr>, qbelo1\_pre <dbl>,  
## # qbelo2\_pre <dbl>, qb1 <chr>, qb2 <chr>, qb1\_value\_pre <chr>,  
## # qb2\_value\_pre <chr>, qb1\_adj <chr>, qb2\_adj <dbl>, qbelo\_prob1 <dbl>,  
## # qbelo\_prob2 <dbl>, qb1\_game\_value <dbl>, qb2\_game\_value <dbl>,  
## # qb1\_value\_post <dbl>, qb2\_value\_post <dbl>, qbelo1\_post <dbl>,  
## # qbelo2\_post <dbl>, score1 <chr>, score2 <chr>

# Check for any missing values  
missing\_values <- sum(is.na(combined\_data))  
missing\_values

## [1] 370

##Handling Missing values

# Identify columns with missing values  
cols\_with\_na <- colnames(combined\_data)[apply(combined\_data, 2, anyNA)]  
cols\_with\_na

## [1] "playoff" "team1" "team2" "elo1\_pre"   
## [5] "elo2\_pre" "elo\_prob1" "elo\_prob2" "elo1\_post"   
## [9] "elo2\_post" "qbelo1\_pre" "qbelo2\_pre" "qb1"   
## [13] "qb2" "qb1\_value\_pre" "qb2\_value\_pre" "qb1\_adj"   
## [17] "qb2\_adj" "qbelo\_prob1" "qbelo\_prob2" "qb1\_game\_value"  
## [21] "qb2\_game\_value" "qb1\_value\_post" "qb2\_value\_post" "qbelo1\_post"   
## [25] "qbelo2\_post" "score1" "score2"

# Impute missing values or remove rows/columns with missing values  
# For simplicity, we will remove the rows with missing values  
cleaned\_data <- combined\_data %>% drop\_na()  
  
# Verify that there are no missing values left  
sum(is.na(cleaned\_data))

## [1] 0

##Taking care of Duplicates

# Identify potential duplicate rows  
duplicates <- combined\_data[duplicated(combined\_data), ]  
duplicates

## # A tibble: 0 × 30  
## # ℹ 30 variables: date <dbl>, season <dbl>, neutral <dbl>, playoff <chr>,  
## # team1 <chr>, team2 <chr>, elo1\_pre <chr>, elo2\_pre <dbl>, elo\_prob1 <dbl>,  
## # elo\_prob2 <dbl>, elo1\_post <dbl>, elo2\_post <chr>, qbelo1\_pre <dbl>,  
## # qbelo2\_pre <dbl>, qb1 <chr>, qb2 <chr>, qb1\_value\_pre <chr>,  
## # qb2\_value\_pre <chr>, qb1\_adj <chr>, qb2\_adj <dbl>, qbelo\_prob1 <dbl>,  
## # qbelo\_prob2 <dbl>, qb1\_game\_value <dbl>, qb2\_game\_value <dbl>,  
## # qb1\_value\_post <dbl>, qb2\_value\_post <dbl>, qbelo1\_post <dbl>, …

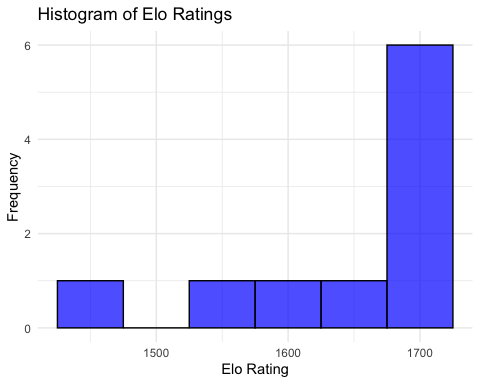
# Remove duplicate rows  
cleaned\_data <- cleaned\_data %>% distinct()  
  
# Display the first few rows of the cleaned dataset  
head(cleaned\_data)

## # A tibble: 6 × 30  
## date season neutral playoff team1 team2 elo1\_pre elo2\_pre elo\_prob1 elo\_prob2  
## <dbl> <dbl> <dbl> <chr> <chr> <chr> <chr> <dbl> <dbl> <dbl>  
## 1 44205 2020 0 w BUF IND 1693.21… 1597. 0.717 0.283  
## 2 44205 2020 0 w SEA LAR 1628.76… 1588. 0.648 0.352  
## 3 44205 2020 0 w WSH TB 1456.95… 1630. 0.349 0.651  
## 4 44206 2020 0 w TEN BAL 1599.07… 1654. 0.514 0.486  
## 5 44206 2020 0 w NO CHI 1695.68… 1500. 0.818 0.182  
## 6 44206 2020 0 w PIT CLE 1572.16… 1517. 0.666 0.334  
## # ℹ 20 more variables: elo1\_post <dbl>, elo2\_post <chr>, qbelo1\_pre <dbl>,  
## # qbelo2\_pre <dbl>, qb1 <chr>, qb2 <chr>, qb1\_value\_pre <chr>,  
## # qb2\_value\_pre <chr>, qb1\_adj <chr>, qb2\_adj <dbl>, qbelo\_prob1 <dbl>,  
## # qbelo\_prob2 <dbl>, qb1\_game\_value <dbl>, qb2\_game\_value <dbl>,  
## # qb1\_value\_post <dbl>, qb2\_value\_post <dbl>, qbelo1\_post <dbl>,  
## # qbelo2\_post <dbl>, score1 <chr>, score2 <chr>

##Ensuring ’elo\_pre is Numeric

##Data visualization ##Histogram of Elo Ratings

# Create a histogram of Elo ratings (using 'elo1\_pre' column)  
ggplot(cleaned\_data, aes(x = elo1\_pre)) +  
 geom\_histogram(binwidth = 50, fill = "blue", color = "black", alpha = 0.7) +  
 labs(title = "Histogram of Elo Ratings",  
 x = "Elo Rating",  
 y = "Frequency") +  
 theme\_minimal()



##Scatter Plot of Elo Ratings vs. Probability

# Create a scatter plot of Elo ratings vs. probability of winning (using 'elo1\_pre' and 'elo\_prob1' columns)  
ggplot(cleaned\_data, aes(x = elo1\_pre, y = elo\_prob1)) +  
 geom\_point(color = "blue", alpha = 0.5) +  
 labs(title = "Scatter Plot of Elo Ratings vs. Probability of Winning",  
 x = "Elo Rating",  
 y = "Probability of Winning") +  
 theme\_minimal()

