JaylenBrownMod1_Final

Felix Liang

2024-07-23

```
# Clear environment
rm(list = ls())

# Load necessary libraries
library(tidymodels)
```

```
## Warning: package 'tidymodels' was built under R version 4.3.3
```

```
## ── Attaching packages ────────────────────────────────────── tidymodels 1.2.0 ──
```

```
## ✓ broom        1.0.5      ✓ recipes      1.0.10
## ✓ dials        1.2.1      ✓ rsample      1.2.1
## ✓ dplyr        1.1.4      ✓ tibble       3.2.1
## ✓ ggplot2      3.5.0      ✓ tidyr        1.3.1
## ✓ infer        1.0.7      ✓ tune         1.2.0
## ✓ modeldata    1.3.0      ✓ workflows    1.1.4
## ✓ parsnip      1.2.1      ✓ workflowsets 1.1.0
## ✓ purrr        1.0.2      ✓ yardstick    1.3.1
```

```
## Warning: package 'dials' was built under R version 4.3.3
```

```
## Warning: package 'ggplot2' was built under R version 4.3.3
```

```
## Warning: package 'infer' was built under R version 4.3.3
```

```
## Warning: package 'modeldata' was built under R version 4.3.3
```

```
## Warning: package 'parsnip' was built under R version 4.3.3
```

```
## Warning: package 'recipes' was built under R version 4.3.3
```

```
## Warning: package 'rsample' was built under R version 4.3.3
```

```
## Warning: package 'tidyr' was built under R version 4.3.3
```

```
## Warning: package 'tune' was built under R version 4.3.3
## Warning: package 'workflows' was built under R version 4.3.3
```

```
## Warning: package 'workflowsets' was built under R version 4.3.3
```

```
## Warning: package 'yardstick' was built under R version 4.3.3
```

```
## ── Conflicts ───────────────────────────────────────── tidymodels_conflicts() ──
## ✖ purrr::discard() masks scales::discard()
## ✖ dplyr::filter()  masks stats::filter()
## ✖ dplyr::lag()     masks stats::lag()
## ✖ recipes::step()  masks stats::step()
## • Dig deeper into tidy modeling with R at https://www.tmwr.org
```

```
library(tidytext)
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ forcats   1.0.0     ✔ readr     2.1.5
## ✔ lubridate 1.9.3     ✔ stringr   1.5.1
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ readr::col_factor() masks scales::col_factor()
## ✖ purrr::discard()    masks scales::discard()
## ✖ dplyr::filter()     masks stats::filter()
## ✖ stringr::fixed()    masks recipes::fixed()
## ✖ dplyr::lag()        masks stats::lag()
## ✖ readr::spec()       masks yardstick::spec()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
e errors
```

```
library(plotly)
```

```
## Warning: package 'plotly' was built under R version 4.3.3
```

```
##
## Attaching package: 'plotly'
##
## The following object is masked from 'package:ggplot2':
##
##     last_plot
##
## The following object is masked from 'package:stats':
##
##     filter
##
## The following object is masked from 'package:graphics':
##
##     layout
```

```r
library(scales)
library(ranger)
```

```
## Warning: package 'ranger' was built under R version 4.3.3
```

```r
library(zoo)
```

```
## Warning: package 'zoo' was built under R version 4.3.3
```

```
##
## Attaching package: 'zoo'
##
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```r
library(ggcorrplot)
```

```
## Warning: package 'ggcorrplot' was built under R version 4.3.3
```

```r
library(car)
```

```
## Warning: package 'car' was built under R version 4.3.3
```

```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 4.3.3
```

```
##
## Attaching package: 'car'
##
## The following object is masked from 'package:purrr':
##
##     some
##
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
library(optimx)
```

```
## Warning: package 'optimx' was built under R version 4.3.3
```

```
# Load dataset
jb <- read_csv("https://raw.githubusercontent.com/maliknyc/NBA-Prediction-Testing/main/JaylenBro
wnTest.csv")
```

```
## New names:
## Rows: 102 Columns: 35
## ── Column specification
## ──────────────────────────────────────────────────────── Delimiter: "," chr
## (6): Date, Age, Tm, ...6, Opp, ...8 dbl (28): Rk, G, GS, FG, FGA, FG%, 3P, 3PA,
## 3P%, FT, FTA, FT%, ORB, DRB, TR... time (1): MP
## ℹ Use `spec()` to retrieve the full column specification for this data. ℹ
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## • `` -> `...6`
## • `` -> `...8`
```

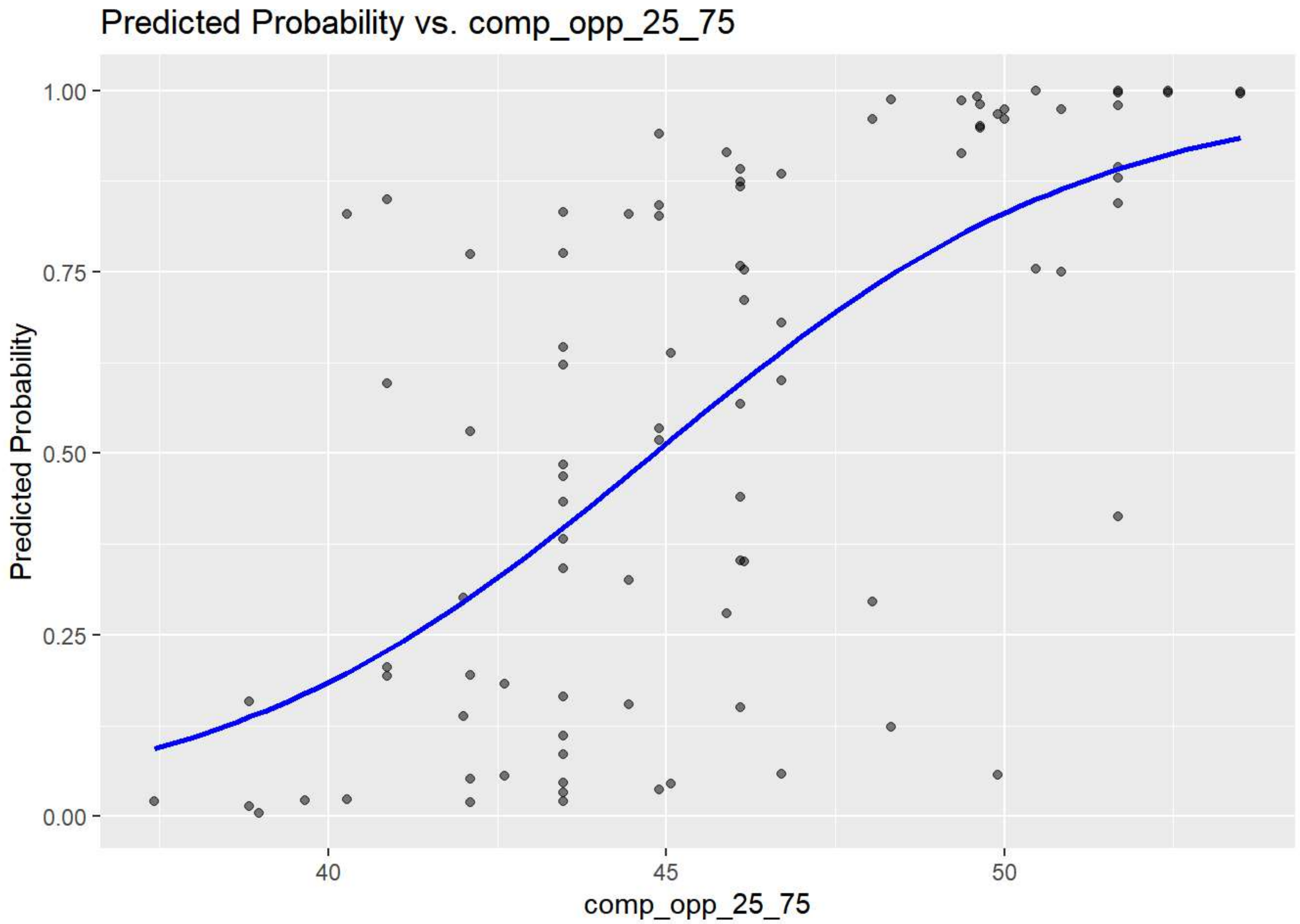
```
# Preprocess the dataset
jb_cleaned <- jb %>%
  rename(TPA = `3PA`, TPP = `3P%`) %>%
  mutate(Date = as.Date(Date, format = "%m/%d/%Y")) %>%
  arrange(Date) %>%
  select(Date, Opp, MP, FGA, TPA, TPP, TRB, AST, STL, PTS, PRA, PR, PA, RA, SB)

# Add average points vs team
avg_points_vs_team <- jb %>%
  group_by(Opp) %>%
  summarize(avgPTS_vteam = mean(PTS)) %>%
  ungroup()

jb_cleaned <- jb_cleaned %>%
  left_join(avg_points_vs_team, by = "Opp")

# Add defensive ratings
def_ratings <- data.frame(
  Opp = c("BOS", "DEN", "OKC", "MIN", "LAC", "DAL", "NYK", "MIL", "NOP", "PHO", "CLE", "IND", "L
AL", "ORL", "PHI", "GSW", "MIA", "SAC", "HOU", "CHI", "ATL", "BRK", "UTA", "MEM", "TOR", "SAS",
"CHO", "POR", "WAS", "DET"),
  DEF_RTG = c(110.6, 112.3, 111.0, 110.4, 113.1, 114.9, 111.4, 110.2, 111.9, 113.7, 115.0, 117.
6, 114.8, 110.8, 113.0, 114.5, 111.5, 114.4, 112.8, 115.7, 115.4, 114.6, 119.6, 113.7, 115.6, 11
6.1, 116.9, 118.0, 118.6, 118.0)
)

jb_cleaned <- jb_cleaned %>%
  left_join(def_ratings, by = "Opp") %>%
  drop_na()

# Calculate days of rest and cap at 30 days
jb_cleaned <- jb_cleaned %>%
  mutate(days_rest = as.numeric(difftime(Date, lag(Date), units = 'days')),
         days_rest = ifelse(days_rest > 30, 30, days_rest)) %>%
  drop_na()

# Calculate moving averages and other features
jb_ra <- jb_cleaned %>%
  mutate(Opp = as.factor(Opp), over_21_5 = ifelse(PTS > 21.5, 1, 0),
         avg_PTS_5 = rollapply(PTS, width = 5, FUN = mean, fill = NA, align = "right"),
         avg_PTS_10 = rollapply(PTS, width = 10, FUN = mean, fill = NA, align = "right"),
         avg_TPA_5 = rollapply(TPA, width = 5, FUN = mean, fill = NA, align = "right"),
         trend_PTS = rollapply(PTS, width = 5, FUN = function(x) coef(lm(x ~ seq_along(x)))[2],
fill = NA, align = "right"),
        perc_over = rollapply(over_21_5, width = 10, FUN = mean, fill = NA, align = "right") * 1
00) %>%
  drop_na()

# Select relevant variables for the model
jb_vars <- jb_ra %>%
  mutate(int_PTS5_TPA5 = avg_PTS_5 * avg_TPA_5,
         comp_opp_25_75 = 0.25 * DEF_RTG + 0.75 * avgPTS_vteam) %>%
  select(Date, Opp, avgPTS_vteam, days_rest, avg_PTS_5, avg_TPA_5, DEF_RTG, over_21_5, comp_opp_
25_75, trend_PTS, int_PTS5_TPA5, perc_over)

# Fit logistic regression model
form.vars_cleaned1 <- "over_21_5 ~ comp_opp_25_75 + avg_PTS_5 + trend_PTS + perc_over"
glm_cleaned <- glm(as.formula(form.vars_cleaned1), jb_vars, family = binomial)

# Add predicted probabilities to jb_vars
jb_vars <- jb_vars %>%
  mutate(predicted_prob = predict(glm_cleaned, newdata = jb_vars, type = "response"))

# View the updated dataframe with the new column
jb_vars %>%
  select(Date, Opp, over_21_5, predicted_prob) %>%
  print(n = 90)
```

```
## # A tibble: 92 × 4
##    Date       Opp   over_21_5 predicted_prob
##    <date>     <fct>     <dbl>          <dbl>
##  1 2023-05-09 PHI           1        0.842
##  2 2023-05-11 PHI           0        0.518
##  3 2023-05-14 PHI           1        0.827
##  4 2023-05-17 MIA           1        0.775
##  5 2023-05-19 MIA           0        0.382
##  6 2023-05-21 MIA           0        0.0857
##  7 2023-05-23 MIA           0        0.0330
##  8 2023-05-25 MIA           0        0.111
##  9 2023-05-27 MIA           1        0.646
## 10 2023-05-29 MIA           0        0.468
## 11 2023-10-25 NYK           0        0.0190
## 12 2023-10-27 MIA           1        0.165
## 13 2023-10-30 WAS           1        0.968
## 14 2023-11-01 IND           0        0.880
## 15 2023-11-04 BRK           1        0.711
## 16 2023-11-06 MIN           1        0.754
## 17 2023-11-08 PHI           0        0.0362
## 18 2023-11-10 BRK           1        0.753
## 19 2023-11-11 TOR           1        0.951
## 20 2023-11-13 NYK           1        0.774
## 21 2023-11-17 TOR           1        0.981
## 22 2023-11-19 MEM           0        0.0204
## 23 2023-11-20 CHO           0        0.00512
## 24 2023-11-22 MIL           1        0.205
## 25 2023-11-24 ORL           0        0.325
## 26 2023-11-26 ATL           0        0.680
## 27 2023-11-28 CHI           1        0.961
## 28 2023-12-01 PHI           0        0.534
## 29 2023-12-04 IND           1        0.980
## 30 2023-12-08 NYK           0        0.194
## 31 2023-12-12 CLE           1        0.353
## 32 2023-12-14 CLE           1        0.568
## 33 2023-12-15 ORL           0        0.155
## 34 2023-12-17 ORL           1        0.830
## 35 2023-12-19 GSW           1        0.973
## 36 2023-12-20 SAC           1        0.991
## 37 2023-12-23 LAC           1        0.830
## 38 2023-12-25 LAL           0        0.158
## 39 2023-12-29 TOR           1        0.949
## 40 2023-12-31 SAS           1        0.914
## 41 2024-01-02 OKC           0        0.302
## 42 2024-01-05 UTA           0        0.0222
## 43 2024-01-06 IND           1        0.845
## 44 2024-01-08 IND           1        0.999
## 45 2024-01-10 MIN           1        1.00
## 46 2024-01-11 MIL           0        0.597
## 47 2024-01-13 HOU           1        0.638
## 48 2024-01-17 SAS           0        0.279
## 49 2024-01-19 DEN           0        0.123
## 50 2024-01-21 HOU           0        0.0455
## 51 2024-01-22 DAL           1        0.750
## 52 2024-01-25 MIA           0        0.485
## 53 2024-01-27 LAC           0        0.0231
## 54 2024-01-29 NOP           1        0.0556
## 55 2024-01-30 IND           1        0.413
## 56 2024-02-01 LAL           0        0.0139
## 57 2024-02-07 ATL           0        0.0579
## 58 2024-02-09 WAS           0        0.0568
## 59 2024-02-11 MIA           0        0.0463
## 60 2024-02-13 BRK           0        0.351
## 61 2024-02-22 CHI           0        0.295
## 62 2024-02-24 NYK           1        0.530
## 63 2024-02-27 PHI           1        0.939
## 64 2024-03-01 DAL           1        0.974
## 65 2024-03-03 GSW           1        0.960
## 66 2024-03-05 CLE           0        0.441
## 67 2024-03-07 DEN           1        0.987
## 68 2024-03-09 PHO           1        0.996
## 69 2024-03-11 POR           1        0.986
## 70 2024-03-14 PHO           1        0.999
## 71 2024-03-18 DET           1        0.998
## 72 2024-03-20 MIL           0        0.850
## 73 2024-03-22 DET           1        0.996
## 74 2024-03-25 ATL           1        0.884
## 75 2024-03-28 ATL           0        0.600
## 76 2024-03-30 NOP           0        0.182
## 77 2024-04-03 OKC           1        0.138
## 78 2024-04-07 POR           1        0.913
## 79 2024-04-09 MIL           0        0.193
## 80 2024-04-11 NYK           0        0.0518
## 81 2024-04-21 MIA           0        0.0213
## 82 2024-04-24 MIA           1        0.622
## 83 2024-04-27 MIA           1        0.832
## 84 2024-04-29 MIA           0        0.341
## 85 2024-05-01 MIA           1        0.433
## 86 2024-05-07 CLE           1        0.867
## 87 2024-05-09 CLE           0        0.758
## 88 2024-05-11 CLE           1        0.891
## 89 2024-05-13 CLE           1        0.873
## 90 2024-05-15 CLE           0        0.150
## # ℹ 2 more rows
```

```r
# Export the updated dataframe to CSV
write.csv(jb_vars %>% select(Date, Opp, over_21_5, predicted_prob), "predictions_jb.csv", row.names = FALSE)

# Plot predicted probabilities vs. comp_opp_25_75
plot_comp_opp_25_75 <- ggplot(jb_vars, aes(x = comp_opp_25_75, y = predicted_prob)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "glm", method.args = list(family = "binomial"), se = FALSE, color = "blue") +
  labs(title = "Predicted Probability vs. comp_opp_25_75", x = "comp_opp_25_75", y = "Predicted Probability")

# Plot predicted probabilities vs. avg_PTS_5
plot_avg_PTS_5 <- ggplot(jb_vars, aes(x = avg_PTS_5, y = predicted_prob)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "glm", method.args = list(family = "binomial"), se = FALSE, color = "blue") +
  labs(title = "Predicted Probability vs. avg_PTS_5", x = "avg_PTS_5", y = "Predicted Probability")

# Plot predicted probabilities vs. trend_PTS
plot_trend_PTS <- ggplot(jb_vars, aes(x = trend_PTS, y = predicted_prob)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "glm", method.args = list(family = "binomial"), se = FALSE, color = "blue") +
  labs(title = "Predicted Probability vs. trend_PTS", x = "trend_PTS", y = "Predicted Probability")

# Display the plots
print(plot_comp_opp_25_75)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

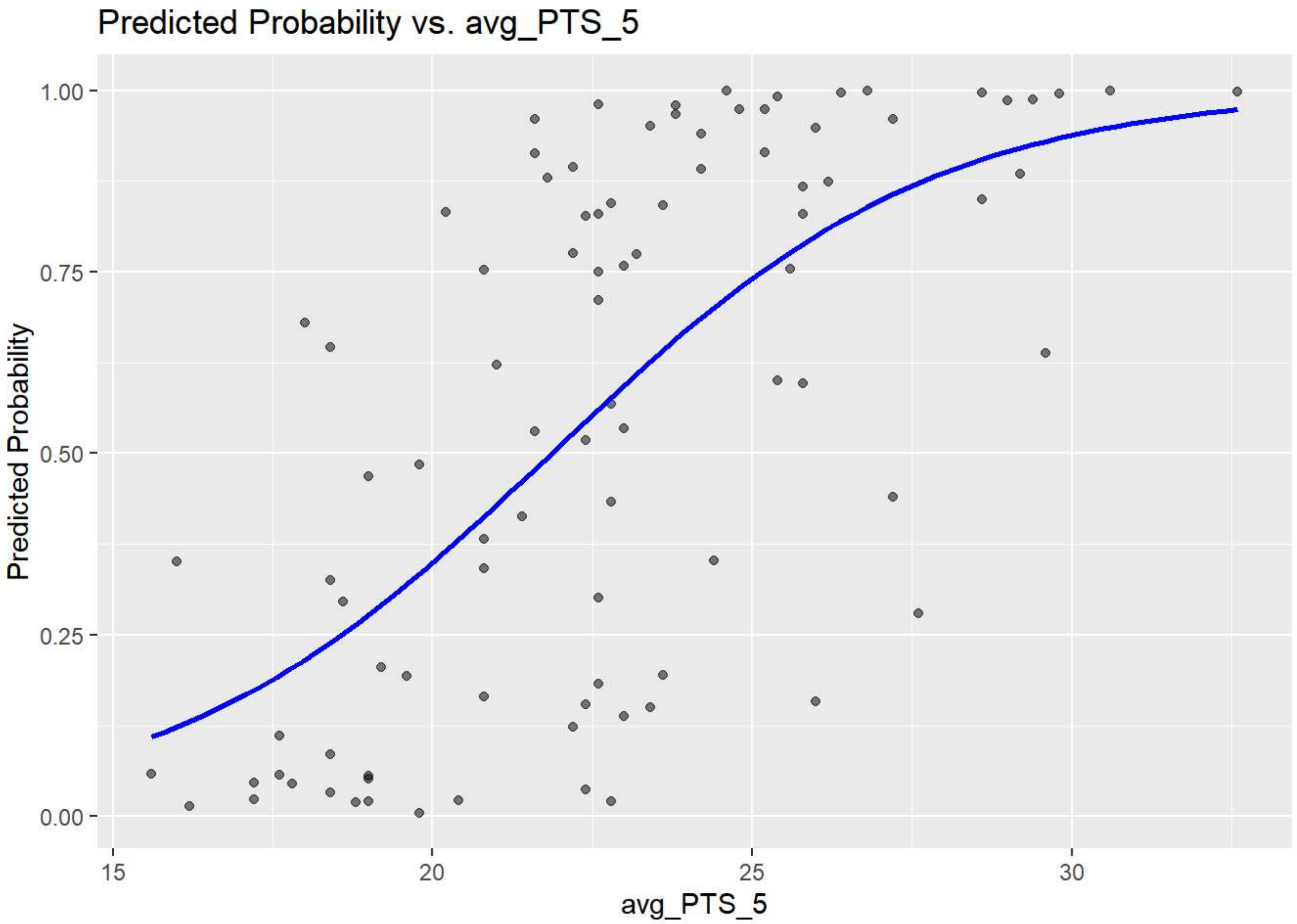
```
## Warning in eval(family$initialize): non-integer #successes in a binomial glm!
```



```
print(plot_avg_PTS_5)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

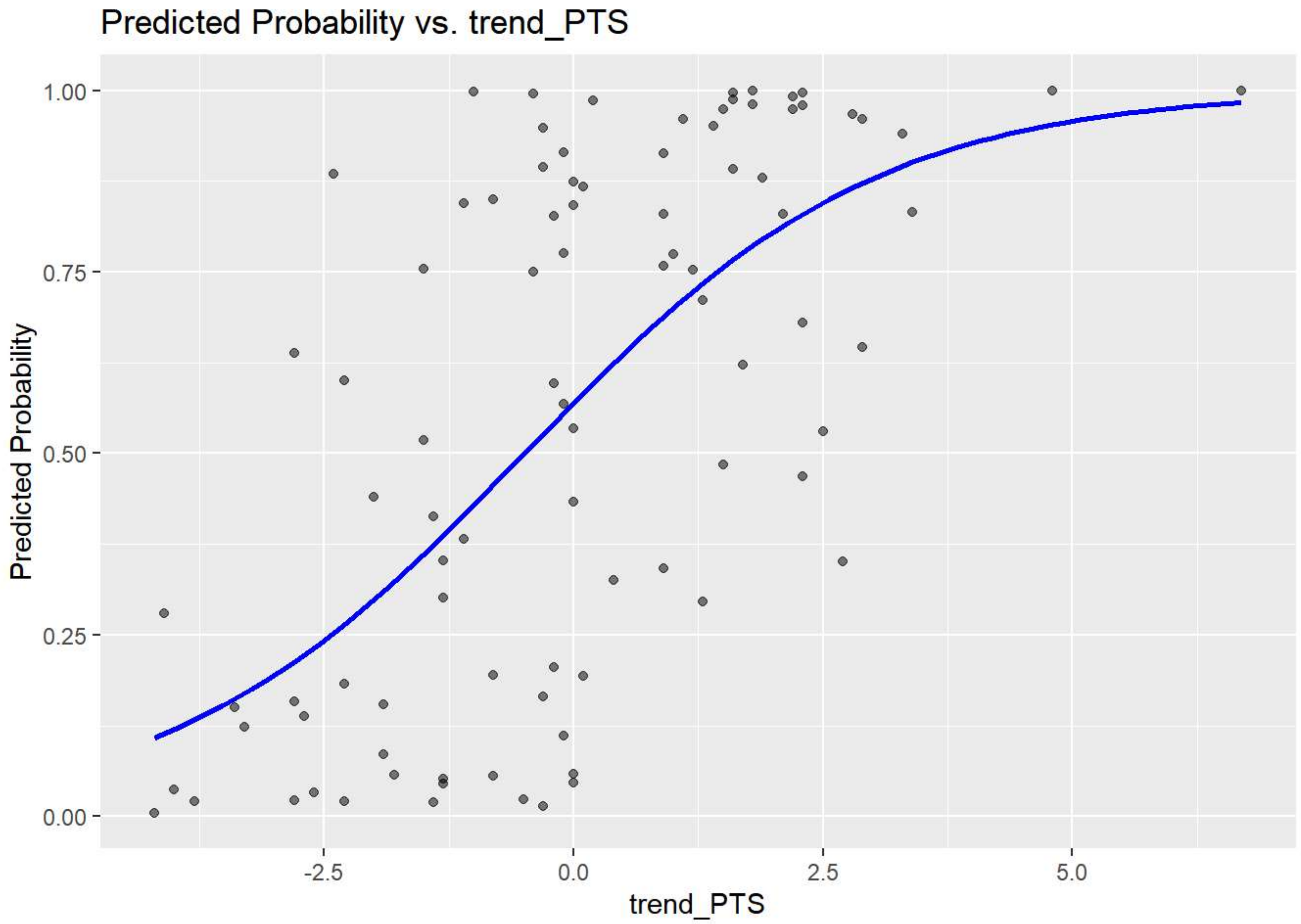
```
## Warning in eval(family$initialize): non-integer #successes in a binomial glm!
```



```
print(plot_trend_PTS)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning in eval(family$initialize): non-integer #successes in a binomial glm!
```



```
# Cross-validation
cvRes <- NULL
for(i in 1:1000) {
  inds <- sample(1:nrow(jb_vars), size = round(nrow(jb_vars)*.8), replace = F)
  train <- jb_vars %>% slice(inds)
  test <- jb_vars %>% slice(-inds)

  # Train the model on the training set
  model_clean <- glm(as.formula(form.vars_cleaned1), train, family = binomial(link = 'logit'))

  # Predict on the test set
  toEval_CV <- test %>%
    mutate(prob_over21_5 = predict(model_clean, newdata = test, type = 'response'),
           over_21_5 = factor(over_21_5, levels = c('1', '0')))

  cvRes <- cvRes %>%
    bind_rows(yardstick::roc_auc(toEval_CV, over_21_5, prob_over21_5) %>%
                mutate(cvInd = i))
}

cvRes %>%
  summarise(mean_auc = mean(.estimate))
```

```
## # A tibble: 1 × 1
##   mean_auc
##      <dbl>
## 1    0.894
```

```
# Final AUC
mean_auc <- cvRes %>%
  summarise(mean_auc = mean(.estimate)) %>%
  pull(mean_auc)

print(mean_auc)  # AUC: 0.892
```

```
## [1] 0.8937853
```

```
# View updated dataset with predicted probabilities
jb_vars %>%
  select(Date, Opp, over_21_5, predicted_prob) %>%
  print(n = 90)
```

```
## # A tibble: 92 × 4
##    Date       Opp   over_21_5 predicted_prob
##    <date>     <fct>     <dbl>          <dbl>
##  1 2023-05-09 PHI           1        0.842
##  2 2023-05-11 PHI           0        0.518
##  3 2023-05-14 PHI           1        0.827
##  4 2023-05-17 MIA           1        0.775
##  5 2023-05-19 MIA           0        0.382
##  6 2023-05-21 MIA           0        0.0857
##  7 2023-05-23 MIA           0        0.0330
##  8 2023-05-25 MIA           0        0.111
##  9 2023-05-27 MIA           1        0.646
## 10 2023-05-29 MIA           0        0.468
## 11 2023-10-25 NYK           0        0.0190
## 12 2023-10-27 MIA           1        0.165
## 13 2023-10-30 WAS           1        0.968
## 14 2023-11-01 IND           0        0.880
## 15 2023-11-04 BRK           1        0.711
## 16 2023-11-06 MIN           1        0.754
## 17 2023-11-08 PHI           0        0.0362
## 18 2023-11-10 BRK           1        0.753
## 19 2023-11-11 TOR           1        0.951
## 20 2023-11-13 NYK           1        0.774
## 21 2023-11-17 TOR           1        0.981
## 22 2023-11-19 MEM           0        0.0204
## 23 2023-11-20 CHO           0        0.00512
## 24 2023-11-22 MIL           1        0.205
## 25 2023-11-24 ORL           0        0.325
## 26 2023-11-26 ATL           0        0.680
## 27 2023-11-28 CHI           1        0.961
## 28 2023-12-01 PHI           0        0.534
## 29 2023-12-04 IND           1        0.980
## 30 2023-12-08 NYK           0        0.194
## 31 2023-12-12 CLE           1        0.353
## 32 2023-12-14 CLE           1        0.568
## 33 2023-12-15 ORL           0        0.155
## 34 2023-12-17 ORL           1        0.830
## 35 2023-12-19 GSW           1        0.973
## 36 2023-12-20 SAC           1        0.991
## 37 2023-12-23 LAC           1        0.830
## 38 2023-12-25 LAL           0        0.158
## 39 2023-12-29 TOR           1        0.949
## 40 2023-12-31 SAS           1        0.914
## 41 2024-01-02 OKC           0        0.302
## 42 2024-01-05 UTA           0        0.0222
## 43 2024-01-06 IND           1        0.845
## 44 2024-01-08 IND           1        0.999
## 45 2024-01-10 MIN           1        1.00
## 46 2024-01-11 MIL           0        0.597
## 47 2024-01-13 HOU           1        0.638
## 48 2024-01-17 SAS           0        0.279
## 49 2024-01-19 DEN           0        0.123
## 50 2024-01-21 HOU           0        0.0455
## 51 2024-01-22 DAL           1        0.750
## 52 2024-01-25 MIA           0        0.485
## 53 2024-01-27 LAC           0        0.0231
## 54 2024-01-29 NOP           1        0.0556
## 55 2024-01-30 IND           1        0.413
## 56 2024-02-01 LAL           0        0.0139
## 57 2024-02-07 ATL           0        0.0579
## 58 2024-02-09 WAS           0        0.0568
## 59 2024-02-11 MIA           0        0.0463
## 60 2024-02-13 BRK           0        0.351
## 61 2024-02-22 CHI           0        0.295
## 62 2024-02-24 NYK           1        0.530
## 63 2024-02-27 PHI           1        0.939
## 64 2024-03-01 DAL           1        0.974
## 65 2024-03-03 GSW           1        0.960
## 66 2024-03-05 CLE           0        0.441
## 67 2024-03-07 DEN           1        0.987
## 68 2024-03-09 PHO           1        0.996
## 69 2024-03-11 POR           1        0.986
## 70 2024-03-14 PHO           1        0.999
## 71 2024-03-18 DET           1        0.998
## 72 2024-03-20 MIL           0        0.850
## 73 2024-03-22 DET           1        0.996
## 74 2024-03-25 ATL           1        0.884
## 75 2024-03-28 ATL           0        0.600
## 76 2024-03-30 NOP           0        0.182
## 77 2024-04-03 OKC           1        0.138
## 78 2024-04-07 POR           1        0.913
## 79 2024-04-09 MIL           0        0.193
## 80 2024-04-11 NYK           0        0.0518
## 81 2024-04-21 MIA           0        0.0213
## 82 2024-04-24 MIA           1        0.622
## 83 2024-04-27 MIA           1        0.832
## 84 2024-04-29 MIA           0        0.341
## 85 2024-05-01 MIA           1        0.433
## 86 2024-05-07 CLE           1        0.867
## 87 2024-05-09 CLE           0        0.758
## 88 2024-05-11 CLE           1        0.891
## 89 2024-05-13 CLE           1        0.873
## 90 2024-05-15 CLE           0        0.150
## # ℹ 2 more rows
```