Prem24/25

2025-06-16

Loading Libraries and Data

```
library('tidyverse')
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library('dplyr')
library("tidymodels")
```

```
## Warning: package 'tidymodels' was built under R version 4.4.2
```

```
## -- Attaching packages -------------------------------------- tidymodels 1.2.0 --
## v broom        1.0.6     v rsample      1.2.1
## v dials        1.3.0     v tune         1.2.1
## v infer        1.0.7     v workflows    1.1.4
## v modeldata    1.4.0     v workflowsets 1.1.0
## v parsnip      1.2.1     v yardstick    1.3.1
## v recipes      1.1.0
```

```
## Warning: package 'dials' was built under R version 4.4.2
```

```
## Warning: package 'infer' was built under R version 4.4.2
```

```
## Warning: package 'modeldata' was built under R version 4.4.2
```

```
## Warning: package 'parsnip' was built under R version 4.4.2
```

```
## Warning: package 'recipes' was built under R version 4.4.2
```

```
## Warning: package 'rsample' was built under R version 4.4.2
## Warning: package ’tune’ was built under R version 4.4.2

## Warning: package ’workflows’ was built under R version 4.4.2

## Warning: package ’workflowsets’ was built under R version 4.4.2

## Warning: package ’yardstick’ was built under R version 4.4.2

## -- Conflicts ----------------------------------------- tidymodels_conflicts() --
## x scales::discard() masks purrr::discard()
## x dplyr::filter()   masks stats::filter()
## x recipes::fixed()  masks stringr::fixed()
## x dplyr::lag()      masks stats::lag()
## x yardstick::spec() masks readr::spec()
## x recipes::step()   masks stats::step()
## * Search for functions across packages at https://www.tidymodels.org/find/
```

```r
library('tidylog')
```

```
## Warning: package ’tidylog’ was built under R version 4.4.2

##
## Attaching package: ’tidylog’
##
## The following objects are masked from ’package:dplyr’:
##
##     add_count, add_tally, anti_join, count, distinct, distinct_all,
##     distinct_at, distinct_if, filter, filter_all, filter_at, filter_if,
##     full_join, group_by, group_by_all, group_by_at, group_by_if,
##     inner_join, left_join, mutate, mutate_all, mutate_at, mutate_if,
##     relocate, rename, rename_all, rename_at, rename_if, rename_with,
##     right_join, sample_frac, sample_n, select, select_all, select_at,
##     select_if, semi_join, slice, slice_head, slice_max, slice_min,
##     slice_sample, slice_tail, summarise, summarise_all, summarise_at,
##     summarise_if, summarize, summarize_all, summarize_at, summarize_if,
##     tally, top_frac, top_n, transmute, transmute_all, transmute_at,
##     transmute_if, ungroup
##
## The following objects are masked from ’package:tidyr’:
##
##     drop_na, fill, gather, pivot_longer, pivot_wider, replace_na,
##     separate_wider_delim, separate_wider_position,
##     separate_wider_regex, spread, uncount
##
## The following object is masked from ’package:stats’:
##
##     filter
```

```r
library("dplyr")
library("yardstick")
library("rsample")
library("stringr")
library("recipes")
library("kknn")
```

```
## Warning: package 'kknn' was built under R version 4.4.2
```

```r
library("zoo")
```

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

```r
prem_ou24 <- read.csv("C:/Users/tobyr/OneDrive/Desktop/Footy Analy/Prem/premier_stats24-25.csv")
```

Selecting Relevant Columns

```r
# 2024-25 Season Cleaned
prem_ou24.1 <- prem_ou24 %>% select(date_GMT, attendance, home_team_name, away_team_name, referee, Game
                                 home_ppg, away_ppg, home_team_goal_count, away_team_goal_count, total_g
```

```
## select: dropped 28 variables (timestamp, status, Pre.Match.PPG..Home.,
## Pre.Match.PPG..Away., home_team_goal_timings, ...)
```

Calculating Running Average of Metrics for Each Team

```r
# Define the function
calculate_avg <- function(data, team_column, stat_column, new_column) {
  data %>%
    group_by_at(team_column) %>%
    mutate(!!new_column := rollapplyr(!!sym(stat_column), width = seq_along(!!sym(stat_column)), FUN =
    ungroup()
}

# Apply function to data
prem_ou24.1 <- calculate_avg(prem_ou24.1, "home_team_name", "home_team_shots", "HT_avgShots")
```

```
## group_by_at: one grouping variable (home_team_name)
## mutate (grouped): new variable 'HT_avgShots' (double) with 270 unique values and 0% NA
## ungroup: no grouping variables remain
```

```r
prem_ou24.1 <- calculate_avg(prem_ou24.1, "away_team_name", "away_team_shots", "AT_avgShots")
```

```
## group_by_at: one grouping variable (away_team_name)
## mutate (grouped): new variable 'AT_avgShots' (double) with 258 unique values and 0% NA
## ungroup: no grouping variables remain
```

```
prem_ou24.1 <- calculate_avg(prem_ou24.1, "home_team_name", "home_team_shots_on_target", "HT_avgTarget"
```

```
## group_by_at: one grouping variable (home_team_name)
## mutate (grouped): new variable 'HT_avgTarget' (double) with 212 unique values and 0% NA
## ungroup: no grouping variables remain
```

```
prem_ou24.1 <- calculate_avg(prem_ou24.1, "away_team_name", "away_team_shots_on_target", "AT_avgTarget"
```

```
## group_by_at: one grouping variable (away_team_name)
## mutate (grouped): new variable 'AT_avgTarget' (double) with 190 unique values and 0% NA
## ungroup: no grouping variables remain
```

```
prem_ou24.1 <- calculate_avg(prem_ou24.1, "home_team_name", "home_team_possession", "HT_Possess")
```

```
## group_by_at: one grouping variable (home_team_name)
## mutate (grouped): new variable 'HT_Possess' (double) with 314 unique values and 0% NA
## ungroup: no grouping variables remain
```

```
prem_ou24.1 <- calculate_avg(prem_ou24.1, "away_team_name", "away_team_possession", "AT_Possess")
```

```
## group_by_at: one grouping variable (away_team_name)
## mutate (grouped): new variable 'AT_Possess' (double) with 316 unique values and 0% NA
## ungroup: no grouping variables remain
```

Adding the Target Variable (Two or More Goals, Three or More Goals, etc…)

```
prem_ou24.1 <- prem_ou24.1 %>% mutate(TwoOrMore = ifelse(total_goal_count >= 2, 1, 0))
```

```
## mutate: new variable 'TwoOrMore' (double) with 2 unique values and 0% NA
```

```
prem_ou24.1$TwoOrMore <- as.factor(prem_ou24.1$TwoOrMore)
prem_ou24.1 <- prem_ou24.1 %>% mutate(ThreeOrMore = ifelse(total_goal_count >= 3, 1, 0))
```

```
## mutate: new variable 'ThreeOrMore' (double) with 2 unique values and 0% NA
```

```
prem_ou24.1$ThreeOrMore <- as.factor(prem_ou24.1$ThreeOrMore)
prem_ou24.1 <- prem_ou24.1 %>% mutate(FourOrMore = ifelse(total_goal_count >= 4, 1, 0))
```

```
## mutate: new variable 'FourOrMore' (double) with 2 unique values and 0% NA
```

```
prem_ou24.1$FourOrMore <- as.factor(prem_ou24.1$FourOrMore)
```

Adding the Odds of the Under (Based on the Over Provided in Data)

```
# THIS IS NOT USED TO CALCULATE PROFIT!

# Adding Odds of Under (Used for quick analysis not the actual profit/loss)
prem_ou24.1 <- prem_ou24.1 %>% mutate(PercentOver35 = 1 / odds_ft_over35)
```

```
## mutate: new variable 'PercentOver35' (double) with 101 unique values and 0% NA
```

```
prem_ou24.1 <- prem_ou24.1 %>% mutate(PercentUnder35 = 1 - PercentOver35)
```

```
## mutate: new variable 'PercentUnder35' (double) with 101 unique values and 0% NA
```

```
prem_ou24.1 <- prem_ou24.1 %>% mutate(OddsUnder35 = 1 / PercentUnder35)
```

```
## mutate: new variable 'OddsUnder35' (double) with 101 unique values and 0% NA
```

```
prem_ou24.1 <- prem_ou24.1 %>% mutate(PercentOver25 = 1 / odds_ft_over25)
```

```
## mutate: new variable 'PercentOver25' (double) with 62 unique values and 0% NA
```

```
prem_ou24.1 <- prem_ou24.1 %>% mutate(PercentUnder25 = 1 - PercentOver25)
```

```
## mutate: new variable 'PercentUnder25' (double) with 62 unique values and 0% NA
```

```
prem_ou24.1 <- prem_ou24.1 %>% mutate(OddsUnder25 = 1 / PercentUnder25)
```

```
## mutate: new variable 'OddsUnder25' (double) with 62 unique values and 0% NA
```

Creating the Model Recipes

```
# Premier League 2024-25 Recipes
prem_ourecipe24.2 <-
    recipe(TwoOrMore ~ odds_ft_home_team_win + odds_ft_away_team_win + odds_ft_draw + home_ppg + away_pp
             Home.Team.Pre.Match.xG + Away.Team.Pre.Match.xG + average_goals_per_match_pre_match + avera
prem_ourecipe24.3 <-
    recipe(ThreeOrMore ~ odds_ft_home_team_win + odds_ft_away_team_win + odds_ft_draw + home_ppg + away_
             Home.Team.Pre.Match.xG + Away.Team.Pre.Match.xG + average_goals_per_match_pre_match + avera
prem_ourecipe24.4 <-
    recipe(FourOrMore ~ odds_ft_home_team_win + odds_ft_away_team_win + odds_ft_draw + home_ppg + away_p
              average_goals_per_match_pre_match + average_corners_per_match_pre_match + average_cards_pe
prem_ourecipe24.3.2 <-
    recipe(ThreeOrMore ~ Home.Team.Pre.Match.xG + Away.Team.Pre.Match.xG + average_goals_per_match_pre_m
```

Splitting Test and Train

```
train_ou24.1 <- prem_ou24.1[1:271,]
test_ou24.1 <- prem_ou24.1[272:380,]
```

Defining the Model and Extracting Results

```
# This is the model for Over/Under 3.5 Goals
knn_model <-
  nearest_neighbor(neighbors = tune("K")) %>% # Define K as a hyperparameter to tune
  set_engine("kknn") %>% # Define the method as KNN
  set_mode("classification")

knn_workflow <-
  workflow() %>%
  add_recipe(prem_ourecipe24.3) %>%
  add_model(knn_model)

knn_grid <-
  parameters(knn_workflow) %>% # Refer to tuning parameters in the method object
  update(K = neighbors(c(1, 15))) %>% # Define a test range of K between 1 and 15
  grid_regular(levels = 15) %>% # Capture all values of K between 1 and 15
  filter(K %% 2 == 1)
```

```
## Warning: ‘parameters.workflow()‘ was deprecated in tune 0.1.6.9003.
## i Please use ‘hardhat::extract_parameter_set_dials()‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.

## filter: removed 7 rows (47%), 8 rows remaining
```

```
best_k <- tune_grid(
  knn_workflow,
  resamples = vfold_cv(train_ou24.1, v = 10),
  grid = knn_grid,
  metrics = metric_set(yardstick::accuracy) # Ensure accuracy is included in metrics
) %>%
  select_best(metric = "accuracy")  # Explicitly name the metric argument
 #Select the K that leads to the highest accuracy for KNN

knn_workflow_final <- finalize_workflow(knn_workflow, best_k) # Finalize workflow using best k
fit_knn <- fit(knn_workflow_final, data = train_ou24.1)

predicted_results_knn <-
  predict(fit_knn, new_data = test_ou24.1, type = "prob") %>%
  pluck(2)

results_knn <-
  predicted_results_knn %>%
  bind_cols(test_ou24.1, predictedProbability = .) %>%
  mutate(predictedClass = as.factor(ifelse(predictedProbability > 0.6, 1, 0)))
```

```
## mutate: new variable 'predictedClass' (factor) with 2 unique values and 0% NA
```

```
results_knn %>% filter(Game.Week == c(28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38)) %>% select(home_team_
```

```
## Warning: There was 1 warning in `.fun()`.
## i In argument: `Game.Week == c(28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38)`.
## Caused by warning in `Game.Week == c(28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38)`:
## ! longer object length is not a multiple of shorter object length
```

```
## filter: removed 100 rows (92%), 9 rows remaining
## select: dropped 52 variables (date_GMT, attendance, referee, Game.Week, home_ppg, ...)
```

```
## # A tibble: 9 x 3
##   home_team_name         away_team_name         predictedClass
##   <chr>                  <chr>                  <fct>
## 1 Nottingham Forest      Manchester City        0
## 2 Manchester City        Brighton & Hove Albion 1
## 3 Manchester City        Leicester City         0
## 4 Tottenham Hotspur      Southampton            1
## 5 Aston Villa            Newcastle United       1
## 6 Brighton & Hove Albion West Ham United        0
## 7 Chelsea                Liverpool              1
## 8 Nottingham Forest      Leicester City         0
## 9 Manchester City        AFC Bournemouth        1
```