House Prediction

Team

2024-12-01

R Markdown

```
knitr::opts_chunk$set(echo = TRUE)
# Load libraries
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(caret)
```

```
##      lattice
##
##    'caret'
##
## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(gbm)
```

```
## Loaded gbm 2.2.2
## This version of gbm is no longer under development. Consider transitioning to gbm3, https://github.co
```

```
library(xgboost)
```

```
##
##    'xgboost'
##
## The following object is masked from 'package:dplyr':
##
##     slice
```

```r
library(randomForest)
```

```
## randomForest 4.7-1.2
## Type rfNews() to see new features/changes/bug fixes.
##
##     'randomForest'
##
## The following object is masked from 'package:dplyr':
##
##     combine
##
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```r
library(ggplot2)

# Load data
housing_data <- read.csv("C:/Users/Bowen/Downloads/437/CPTS_437_data/whitman_property_details.csv")

# Data cleaning and feature engineering
clean_data <- housing_data %>%
  mutate(
    Total_Area = as.numeric(gsub(",", "", ifelse(Total_Area == "None", NA, Total_Area))),
    Year_Built = as.numeric(ifelse(Year_Built == "None", NA, Year_Built)),
    Total_Value = as.numeric(gsub(",", "", ifelse(Total_Value == "None", NA, Total_Value))),
    Bedrooms = as.numeric(ifelse(Bedrooms == "None", NA, Bedrooms)),
    Bathrooms = as.numeric(ifelse(Bathrooms == "None", NA, Bathrooms)),
    Garage_Stalls = as.numeric(ifelse(Garage_Stalls %in% c("None", "Block"), 0, Garage_Stalls))
  ) %>%
  filter(!is.na(Total_Value) & !is.na(Total_Area) & !is.na(Year_Built)) %>%
  mutate(
    log_value = log(Total_Value + 1),
    log_area = log(Total_Area + 1),
    age = 2024 - Year_Built,
    has_garage = ifelse(is.na(Garage_Stalls), 0, 1),
    rooms_per_area = (Bedrooms + Bathrooms) / log_area,
    condition_score = case_when(
      grepl("3.0", Condition) ~ 3.0,
      grepl("3.5", Condition) ~ 3.5,
      grepl("4.0", Condition) ~ 4.0,
      TRUE ~ 3.0
    ),
    age_condition_interaction = age * condition_score
  ) %>%
  filter(
    Total_Value > quantile(Total_Value, 0.03) & Total_Value < quantile(Total_Value, 0.97),
    Total_Area > quantile(Total_Area, 0.03) & Total_Area < quantile(Total_Area, 0.97)
  ) %>%
  select(log_value, log_area, age, has_garage, rooms_per_area, Bathrooms,
         condition_score, age_condition_interaction) %>%
  na.omit()
```

```r
# Metrics function
metrics <- function(predictions, actual) {
  rmse <- sqrt(mean((predictions - actual)^2))
  r2 <- 1 - sum((actual - predictions)^2) / sum((actual - mean(actual))^2)
  mae <- mean(abs(predictions - actual))
  return(list(RMSE = rmse, R2 = r2, MAE = mae))
}

# Loop over seeds 1 to 100
for (seed in 1:100) {
  # Set random seed
  set.seed(seed)

  # Split data
  train_index <- createDataPartition(clean_data$log_value, p = 0.8, list = FALSE)
  train_data <- clean_data[train_index, ]
  test_data <- clean_data[-train_index, ]

  # GBM Model
  gbm_model <- gbm(
    log_value ~ .,
    data = train_data,
    distribution = "gaussian",
    n.trees = 3000,
    interaction.depth = 8,
    shrinkage = 0.005,
    n.minobsinnode = 8,
    bag.fraction = 0.8,
    cv.folds = 5
  )
  best_iter_gbm <- gbm.perf(gbm_model, method = "cv", plot.it = FALSE)

  # XGBoost Model
  train_matrix <- xgb.DMatrix(data = as.matrix(train_data %>% select(-log_value)), label = train_data$l
  test_matrix <- xgb.DMatrix(data = as.matrix(test_data %>% select(-log_value)))
  xgb_model <- xgboost(
    data = train_matrix,
    objective = "reg:squarederror",
    nrounds = 2000,
    max_depth = 6,
    eta = 0.01,
    subsample = 0.8,
    colsample_bytree = 0.8,
    verbose = 0
  )

  # Random Forest Model
  rf_model <- randomForest(
    log_value ~ .,
    data = train_data,
    ntree = 500,
    mtry = floor(sqrt(ncol(train_data))),
    importance = TRUE
```

```r
  )

  # Predictions
  gbm_predictions <- exp(predict(gbm_model, test_data, n.trees = best_iter_gbm)) - 1
  xgb_predictions <- exp(predict(xgb_model, test_matrix)) - 1
  rf_predictions <- exp(predict(rf_model, test_data)) - 1
  actual_values <- exp(test_data$log_value) - 1

  # Calculate metrics
  gbm_metrics <- metrics(gbm_predictions, actual_values)
  xgb_metrics <- metrics(xgb_predictions, actual_values)
  rf_metrics <- metrics(rf_predictions, actual_values)

  # Save results to a file
  output_file <- paste0("results_seed_", seed, ".txt")
  writeLines(
    c(
      paste("Seed:", seed),
      "\nGBM Metrics:",
      paste("  RMSE:", round(gbm_metrics$RMSE, 2)),
      paste("  R2:", round(gbm_metrics$R2, 3)),
      paste("  MAE:", round(gbm_metrics$MAE, 2)),
      "\nXGBoost Metrics:",
      paste("  RMSE:", round(xgb_metrics$RMSE, 2)),
      paste("  R2:", round(xgb_metrics$R2, 3)),
      paste("  MAE:", round(xgb_metrics$MAE, 2)),
      "\nRandom Forest Metrics:",
      paste("  RMSE:", round(rf_metrics$RMSE, 2)),
      paste("  R2:", round(rf_metrics$R2, 3)),
      paste("  MAE:", round(rf_metrics$MAE, 2))
    ),
    con = output_file
  )

  cat(paste("Results for seed", seed, "saved to", output_file, "\n"))
}
```

```
## Results for seed 1 saved to results_seed_1.txt
## Results for seed 2 saved to results_seed_2.txt
## Results for seed 3 saved to results_seed_3.txt
## Results for seed 4 saved to results_seed_4.txt
## Results for seed 5 saved to results_seed_5.txt
## Results for seed 6 saved to results_seed_6.txt
## Results for seed 7 saved to results_seed_7.txt
## Results for seed 8 saved to results_seed_8.txt
## Results for seed 9 saved to results_seed_9.txt
## Results for seed 10 saved to results_seed_10.txt
## Results for seed 11 saved to results_seed_11.txt
## Results for seed 12 saved to results_seed_12.txt
## Results for seed 13 saved to results_seed_13.txt
## Results for seed 14 saved to results_seed_14.txt
## Results for seed 15 saved to results_seed_15.txt
## Results for seed 16 saved to results_seed_16.txt
```

```
## Results for seed 17 saved to results_seed_17.txt
## Results for seed 18 saved to results_seed_18.txt
## Results for seed 19 saved to results_seed_19.txt
## Results for seed 20 saved to results_seed_20.txt
## Results for seed 21 saved to results_seed_21.txt
## Results for seed 22 saved to results_seed_22.txt
## Results for seed 23 saved to results_seed_23.txt
## Results for seed 24 saved to results_seed_24.txt
## Results for seed 25 saved to results_seed_25.txt
## Results for seed 26 saved to results_seed_26.txt
## Results for seed 27 saved to results_seed_27.txt
## Results for seed 28 saved to results_seed_28.txt
## Results for seed 29 saved to results_seed_29.txt
## Results for seed 30 saved to results_seed_30.txt
## Results for seed 31 saved to results_seed_31.txt
## Results for seed 32 saved to results_seed_32.txt
## Results for seed 33 saved to results_seed_33.txt
## Results for seed 34 saved to results_seed_34.txt
## Results for seed 35 saved to results_seed_35.txt
## Results for seed 36 saved to results_seed_36.txt
## Results for seed 37 saved to results_seed_37.txt
## Results for seed 38 saved to results_seed_38.txt
## Results for seed 39 saved to results_seed_39.txt
## Results for seed 40 saved to results_seed_40.txt
## Results for seed 41 saved to results_seed_41.txt
## Results for seed 42 saved to results_seed_42.txt
## Results for seed 43 saved to results_seed_43.txt
## Results for seed 44 saved to results_seed_44.txt
## Results for seed 45 saved to results_seed_45.txt
## Results for seed 46 saved to results_seed_46.txt
## Results for seed 47 saved to results_seed_47.txt
## Results for seed 48 saved to results_seed_48.txt
## Results for seed 49 saved to results_seed_49.txt
## Results for seed 50 saved to results_seed_50.txt
## Results for seed 51 saved to results_seed_51.txt
## Results for seed 52 saved to results_seed_52.txt
## Results for seed 53 saved to results_seed_53.txt
## Results for seed 54 saved to results_seed_54.txt
## Results for seed 55 saved to results_seed_55.txt
## Results for seed 56 saved to results_seed_56.txt
## Results for seed 57 saved to results_seed_57.txt
## Results for seed 58 saved to results_seed_58.txt
## Results for seed 59 saved to results_seed_59.txt
## Results for seed 60 saved to results_seed_60.txt
## Results for seed 61 saved to results_seed_61.txt
## Results for seed 62 saved to results_seed_62.txt
## Results for seed 63 saved to results_seed_63.txt
## Results for seed 64 saved to results_seed_64.txt
## Results for seed 65 saved to results_seed_65.txt
## Results for seed 66 saved to results_seed_66.txt
## Results for seed 67 saved to results_seed_67.txt
## Results for seed 68 saved to results_seed_68.txt
## Results for seed 69 saved to results_seed_69.txt
## Results for seed 70 saved to results_seed_70.txt
```

```
## Results for seed 71 saved to results_seed_71.txt
## Results for seed 72 saved to results_seed_72.txt
## Results for seed 73 saved to results_seed_73.txt
## Results for seed 74 saved to results_seed_74.txt
## Results for seed 75 saved to results_seed_75.txt
## Results for seed 76 saved to results_seed_76.txt
## Results for seed 77 saved to results_seed_77.txt
## Results for seed 78 saved to results_seed_78.txt
## Results for seed 79 saved to results_seed_79.txt
## Results for seed 80 saved to results_seed_80.txt
## Results for seed 81 saved to results_seed_81.txt
## Results for seed 82 saved to results_seed_82.txt
## Results for seed 83 saved to results_seed_83.txt
## Results for seed 84 saved to results_seed_84.txt
## Results for seed 85 saved to results_seed_85.txt
## Results for seed 86 saved to results_seed_86.txt
## Results for seed 87 saved to results_seed_87.txt
## Results for seed 88 saved to results_seed_88.txt
## Results for seed 89 saved to results_seed_89.txt
## Results for seed 90 saved to results_seed_90.txt
## Results for seed 91 saved to results_seed_91.txt
## Results for seed 92 saved to results_seed_92.txt
## Results for seed 93 saved to results_seed_93.txt
## Results for seed 94 saved to results_seed_94.txt
## Results for seed 95 saved to results_seed_95.txt
## Results for seed 96 saved to results_seed_96.txt
## Results for seed 97 saved to results_seed_97.txt
## Results for seed 98 saved to results_seed_98.txt
## Results for seed 99 saved to results_seed_99.txt
## Results for seed 100 saved to results_seed_100.txt
```