Random Forest Regressor & Logistic Regression

Bushra Haque

2025-07-13

Random forrest:

```
library(randomForest)
```

```
## Warning: package 'randomForest' was built under R version 4.4.3
```

```
## randomForest 4.7-1.2
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
library(readr)

df <- read_csv("model_ready_dataset.csv")
```

```
## Rows: 61 Columns: 6
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): date, time_of_day
## dbl  (2): level, steps_last_30min
## dttm (1): timestamp
## time (1): time
##
## 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.
```

```
df$time_of_day <- as.factor(df$time_of_day)

rf_model <- randomForest(level ~ steps_last_30min + time_of_day, data = df, ntree = 500)


print(rf_model)
```

```
##
## Call:
##  randomForest(formula = level ~ steps_last_30min + time_of_day,      data = df, ntree = 500)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 1
##
##           Mean of squared residuals: 1883.001
##                     % Var explained: 11.07
```

```r
predicted_levels <- predict(rf_model, df)

rmse <- sqrt(mean((predicted_levels - df$level)^2))
print(paste("RMSE:", round(rmse, 2)))
```

```
## [1] "RMSE: 41.18"
```

This Random Forest regression model was trained to predict glucose level based on steps taken in the past 30 minutes and time of day. The model consisted of 500 trees, with one variable considered at each split. The root mean squared error (RMSE) is approximately 41 mg/dL, indicating that the typical prediction error was around 41 mg/dL. The model explained approximately 11.6% of the variance in glucose levels, suggesting that while walking activity and time of day provide some predictive signal, most of the variation in glucose levels remains unexplained by these two variables alone.

This relatively low explanatory power highlights the complexity of glucose regulation, which depends on multiple factors such as meal intake, medication timing, stress, and individual physiology factors not captured in this dataset. However, this model provides a useful foundation for understanding how walking and time of day contribute to short-term glucose variation.

Logistic Regression:

```r
df$glucose_drop <- ifelse(df$level < 70, 1, 0)

logit_model <- glm(glucose_drop ~ steps_last_30min, data = df, family = "binomial")

summary(logit_model)
```

```
##
## Call:
## glm(formula = glucose_drop ~ steps_last_30min, family = "binomial",
##     data = df)
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -2.657e+01  5.978e+04       0        1
## steps_last_30min  1.542e-18  5.536e+00       0        1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 0.000e+00  on 60  degrees of freedom
## Residual deviance: 3.539e-10  on 59  degrees of freedom
## AIC: 4
##
## Number of Fisher Scoring iterations: 25
```