Modelling Deep Sleep data

Ian Petrus Tan

2024-06-07

Type of linear model - should we use GLM?

```
cat('Range of values of deep sleep:' ,range(cleaned_sleep_df$Deep.Sleep.duration))
```

```
## Range of values of deep sleep: 12 83
```

```
ggplot(cleaned_sleep_df, aes(x = Deep.Sleep.duration)) +
  geom_histogram(aes(y = after_stat(density)), binwidth = 0.5, fill = "blue", color = "black") +# Adjus
  geom_density(alpha = 0.2, fill = "#FF6666") +
  ggtitle("Distribution of Deep Sleep Duration") +
  xlab("Deep Sleep Duration (mins)") +
  ylab("Density") +
  xlim(c(0, max(cleaned_sleep_df$Deep.Sleep.duration)+10))
```

The outcome distribution looks somewhat normally distributed.

```
colnames(cleaned_sleep_df)
```

```
##  [1] "X"                         "Timestamp"
##  [3] "Sleep.quality"             "Deep.Sleep.duration"
##  [5] "Alcohol"                   "Alcohol.hours.before.bed"
##  [7] "Full.before.bed."          "Music.while.sleeping."
##  [9] "Type.of.music"             "Caffiene"
## [11] "Exercise"                  "Calories.count"
## [13] "Game"                      "Date.of.the.night"
## [15] "chamomile"                 "Coffee.Type"
## [17] "Other"                     "Amount.of.sleep"
## [19] "X.1"                       "blackout.eye.mask."
## [21] "Melatonin."                "REM.sleep"
## [23] "Other.1"                   "Date.of.night"
## [25] "day_of_week"               "Amount.of.sleep.hours"
## [27] "Sleep.quality.ordinal"     "Game.past.10pm"
## [29] "chamomile_after_10pm"      "chamomile_after_10pm_bin"
```

```
df_model <- cleaned_sleep_df %>% select('Deep.Sleep.duration', 'Alcohol', 'Game.past.10pm', 'Exercise',
'chamomile_after_10pm', 'Amount.of.sleep.hours', 'blackout.eye.mask.', 'day_of_week')
```

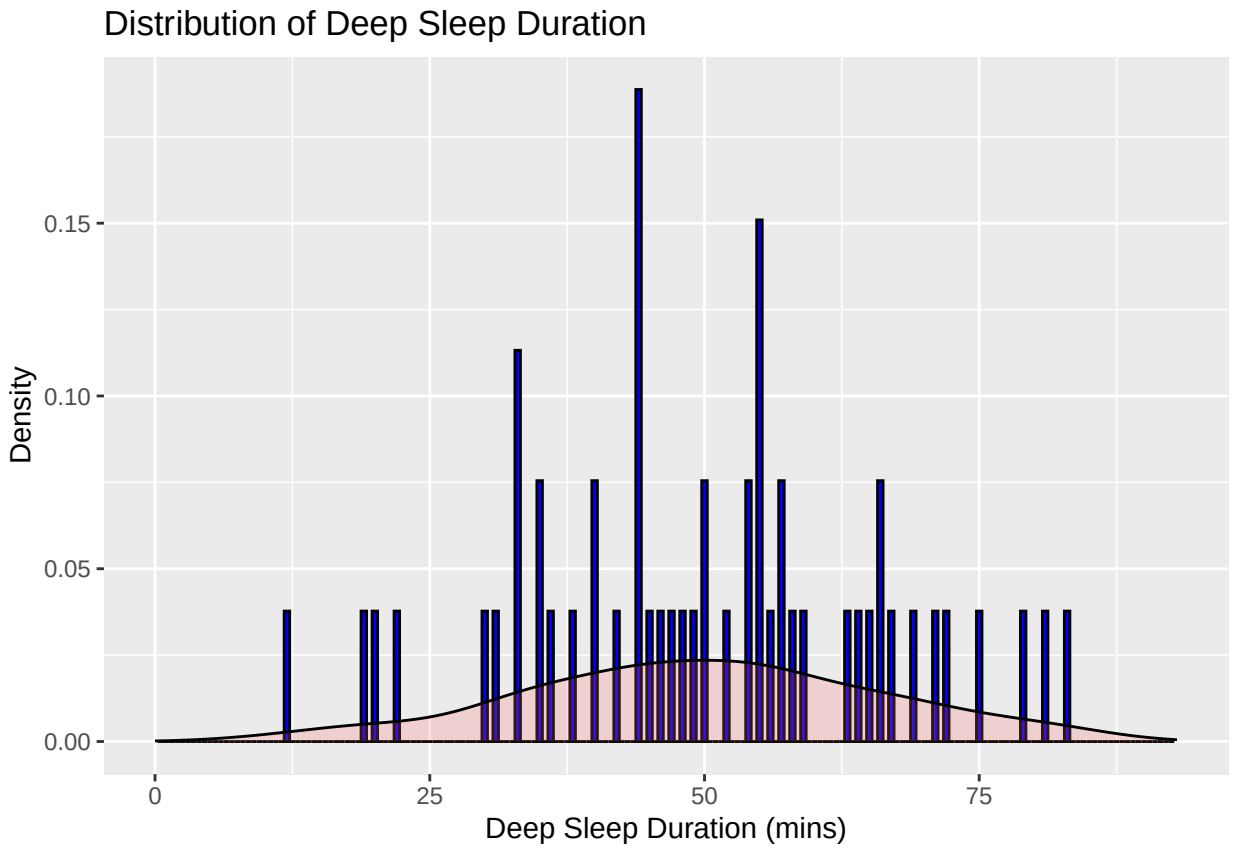


Figure 1: Distribution of deep sleep duration

```
kable(summary(df_model)) ## NA spotted in calories count
```

	Deep.Sleep.duration	Alcohol	Game.past.10pm	Exercise	Calories.count	chamomile_after_10
	Min. :12.00	Length:53	Length:53	Length:53	Min. : 300.0	Length:53
	1st Qu.:40.00	Class :character	Class :character	Class :character	1st Qu.: 478.2	Class :character
	Median :50.00	Mode :character	Mode :character	Mode :character	Median : 570.0	Mode :character
	Mean :49.85	NA	NA	NA	Mean : 706.0	NA
	3rd Qu.:59.00	NA	NA	NA	3rd Qu.:1000.0	NA
	Max. :83.00	NA	NA	NA	Max. :1500.0	NA
	NA	NA	NA	NA	NA's :1	NA

```
## Check for the missing row
missing_index <- which(is.na(df_model['Calories.count']))
kable(df_model[missing_index,]) ## Missing row is for the first obs with exercise = gym
```

Deep.Sleep.duration	Alcohol	Game.past.10pm	Exercise	Calories.count	chamomile_after_10pm	Amount.of.slee
19	No	False	Gym	NA	True	7

```
## Impute using Exercise information
df_model$Calories.count <- ave(df_model$Calories.count, df_model$Exercise, FUN=function(x)
  ifelse(is.na(x), mean(x, na.rm=TRUE), x))

## Check result
df_model[missing_index,]
```

```
##   Deep.Sleep.duration Alcohol Game.past.10pm Exercise Calories.count
## 1                  19      No          False      Gym       561.1364
##   chamomile_after_10pm Amount.of.sleep.hours blackout.eye.mask. day_of_week
## 1                 True              7.283333                  No         Tue
```

Model 1: Full Standard Linear Model (normal errors)

```
model_1 <- df_model %>% lm(Deep.Sleep.duration ~ ., data = .)
# (xtable(model_1, type = 'pdf'))
(xtable(summary(model_1), type = 'pdf'))
```

% latex table generated in R 4.3.1 by xtable 1.8-4 package % Fri Jun 7 18:37:06 2024

```
m1_sum <- summary(model_1)

## Model scores
m1_sum$r.squared
```

[1] 0.3801823

```
m1_sum$adj.r.squared
```

[1] 0.1047078

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	12.6780	34.5744	0.37	0.7160
AlcoholYes	-11.5853	7.5974	-1.52	0.1360
Game.past.10pmTrue	-12.9153	7.4819	-1.73	0.0929
ExerciseFootball + Gym	-6.7560	11.7616	-0.57	0.5693
ExerciseGym	-3.3666	11.8150	-0.28	0.7773
ExerciseNone	1.7747	12.8199	0.14	0.8907
ExerciseSwim	3.2462	21.5626	0.15	0.8812
Calories.count	0.0059	0.0192	0.31	0.7599
chamomile_after_10pmTrue	-4.6502	6.3781	-0.73	0.4707
Amount.of.sleep.hours	4.9851	3.1834	1.57	0.1261
blackout.eye.mask.Yes	2.7443	5.8313	0.47	0.6408
day_of_weekMon	-2.1055	10.1031	-0.21	0.8361
day_of_weekSat	-1.7546	11.6069	-0.15	0.8807
day_of_weekSun	6.9350	10.7808	0.64	0.5241
day_of_weekThu	6.9494	9.1080	0.76	0.4504
day_of_weekTue	-3.2598	9.1856	-0.35	0.7247
day_of_weekWed	-1.2664	8.6384	-0.15	0.8843

Table 1: VIF for model 1 predictors

	GVIF	Df	GVIF^(1/(2*Df))
Alcohol	1.496252	1	1.223214
Game.past.10pm	1.937812	1	1.392053
Exercise	18.918500	4	1.444145
Calories.count	6.764168	1	2.600801
chamomile_after_10pm	2.115609	1	1.454513
Amount.of.sleep.hours	1.907979	1	1.381296
blackout.eye.mask.	1.346811	1	1.160522
day_of_week	17.590731	6	1.269912

```
## Evaluate model
library(car)
kable(vif(model_1), caption = 'VIF for model 1 predictors')
```

We can see that the VIF of calories is really high, suggesting that it can be linearly predicted by the rest of the variables. We can try a second model that excludes

Feature selection

```
model_2 <- df_model %>% lm(Deep.Sleep.duration ~ . -Exercise, data = .)
# print(model_2)
xtable(summary(model_2), type = 'pdf')
```

% latex table generated in R 4.3.1 by xtable 1.8-4 package % Fri Jun 7 18:37:06 2024

```
m2_sum <- summary(model_2)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	10.5689	23.9271	0.44	0.6611
AlcoholYes	-12.7020	6.7730	-1.88	0.0681
Game.past.10pmTrue	-12.6710	6.8546	-1.85	0.0719
Calories.count	0.0032	0.0091	0.35	0.7255
chamomile_after_10pmTrue	-6.1453	5.7387	-1.07	0.2907
Amount.of.sleep.hours	5.5932	2.9041	1.93	0.0612
blackout.eye.mask.Yes	2.5166	5.1865	0.49	0.6302
day_of_weekMon	-3.9045	9.2073	-0.42	0.6738
day_of_weekSat	-5.1098	9.2226	-0.55	0.5826
day_of_weekSun	6.9274	10.1968	0.68	0.5008
day_of_weekThu	4.7138	8.2509	0.57	0.5710
day_of_weekTue	-2.6605	8.7835	-0.30	0.7635
day_of_weekWed	-3.0348	7.6609	-0.40	0.6941

Table 2: VIF for model 2 predictors

	GVIF	Df	GVIF^(1/(2*Df))
Alcohol	1.286442	1	1.134214
Game.past.10pm	1.759564	1	1.326486
Calories.count	1.644623	1	1.282428
chamomile_after_10pm	1.852860	1	1.361198
Amount.of.sleep.hours	1.717795	1	1.310647
blackout.eye.mask.	1.152626	1	1.073604
day_of_week	6.811961	6	1.173382

```
## Model scores
m2_sum$r.squared
```

[1] 0.36341

```
m2_sum$adj.r.squared
```

[1] 0.172433

```
## Evaluate model
kable(vif(model_2), caption = 'VIF for model 2 predictors')
```

The adjusted R^2 has significantly improved.

Random forests

```
library(randomForest)
library(varImp)
# df_model <- df_model %>%
#   mutate(across(where(is.character), as.factor))

## Split data into predictors and response
predictors_df <- df_model %>% select(-Deep.Sleep.duration)
```

Table 3: Feature importance from RF

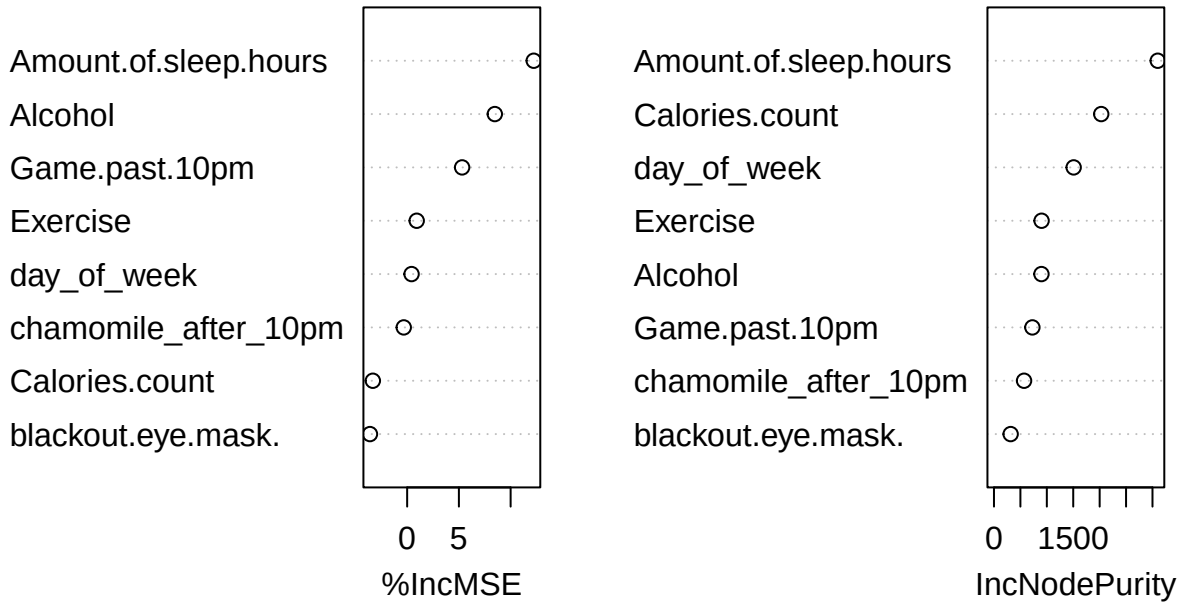
	%IncMSE	IncNodePurity
Alcohol	8.4722409	898.0518
Game.past.10pm	5.3110176	726.8358
Exercise	0.9225890	900.9267
Calories.count	-3.3170714	2029.7863
chamomile_after_10pm	-0.3289814	569.1291
Amount.of.sleep.hours	12.2331179	3099.5732
blackout.eye.mask.	-3.5946757	315.4897
day_of_week	0.4255392	1507.9677

```
response_df <- df_model$Deep.Sleep.duration

## Fit Random Forest model
rf_model <- randomForest(x = predictors_df, y = response_df, ntree = 500, mtry = floor(sqrt(ncol(predic

## Show results
print(rf_model)
```

```
##
## Call:
##  randomForest(x = predictors_df, y = response_df, ntree = 500,      mtry = floor(sqrt(ncol(predictors
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 2
##
##           Mean of squared residuals: 232.0709
##                     % Var explained: 9.65
```

```
## importance
kable(importance(rf_model), caption = 'Feature importance from RF')
```

```
varImpPlot(rf_model)
```




```
?importance
```

Compared to the second linear model, there is a drop in the percentage of variance explained.

On both accounts of importance (increase in error, and increase in node purity), amount of sleep edges out above the rest. It proves to be important to not skimp on sleep, i.e. it is hard to control what kind of sleep you get from the total duration. Best is just to sleep more. Alcohol is agreeably detrimental to sleep.