Predicting TB Medication Adherence Using a Risk Score Model

November 2023

Data Preprocessing

In order to create a risk score model, we must have a binary outcome with integer (or near-integer) covariates. The raw medication adherence data includes two continuous outcome variables – `PCTadherence` and `PCT_adherence_sensi`. Both measure the percentage of doses taken on time, the former calculating this by recorded number of doses missed and the latter calculating this by treatment length and expected treatment length. To make these outcomes binary, we'll need to define a cutoff above which we'll consider the patient "adherent". Ideally, there should be approximately equal numbers of patients in the "adherent" and "non-adherent" groups. Table 1 explores the class imbalance when using a cutoff of 100% (where a patient must have 100% adherence to be considered "adherent"). We observe better balance when using the `PCTadherence_sensi` variable, with 39.1% of patients classified as "non-adherent" (n = 110) and 60.2% of patients classified as "adherent" (n = 154). We thus moved forward with the dichotomized `PCTadherence_sensi` variable as the outcome of interest.

Table 1: Patient Count by Dichotomized Outcome

	<100% Adherence	100% Adherence	Missing
PCTadherence	78	188	0
PCTadherence_sensi	110	154	2

The raw data includes 64 potential covariates. We dropped `PTID2`, `hunger_freq`, `health_ctr`, and `post_tb_pt_work` because they were either not listed in the data dictionary or were listed with a note that they should not be included in the model. We summarized the family support, evaluation of health services, motivation, and TB disinformation variables by taking the median value for each category.

The data dictionary noted that `pills` variable was coded as 1 for 0-3 pills, 2 for 4-6 pills, 3 for 7-9 pills, 4 for 10-11 pills, and 5 for 12+ pills. However, the values in the data were 0.25, 0.50, 0.75, and 1.00. We converted this variable to the scale in the data dictionary by multiplying each value by 4. However, note that this results in no values of 5 in the cleaned data (indicating no subjects taking 12+ pills). The `adr_freq` variable also needed to be multiplied by 4 for the same reason. In this variable's case, all values listed in the data dictionary (0, 1, 2, 3, and 4) are present in the transformed data.

Distributions of the continuous covariates are visualized in the appendix. The variables `self_eff`, `stig_tot`, and `phq9_tot` were dropped There were four continuous variables that were converted to categorical using the following cutoffs (selected based on class balance):

- `age_BLchart`: <16 years, 16-17 years, 18+ years
- `audit_cat`: 0, >0
- `ace_cat`: 0, 1, >1
- `tx_mos_cat`: $\leq$ 6 months, >6 months

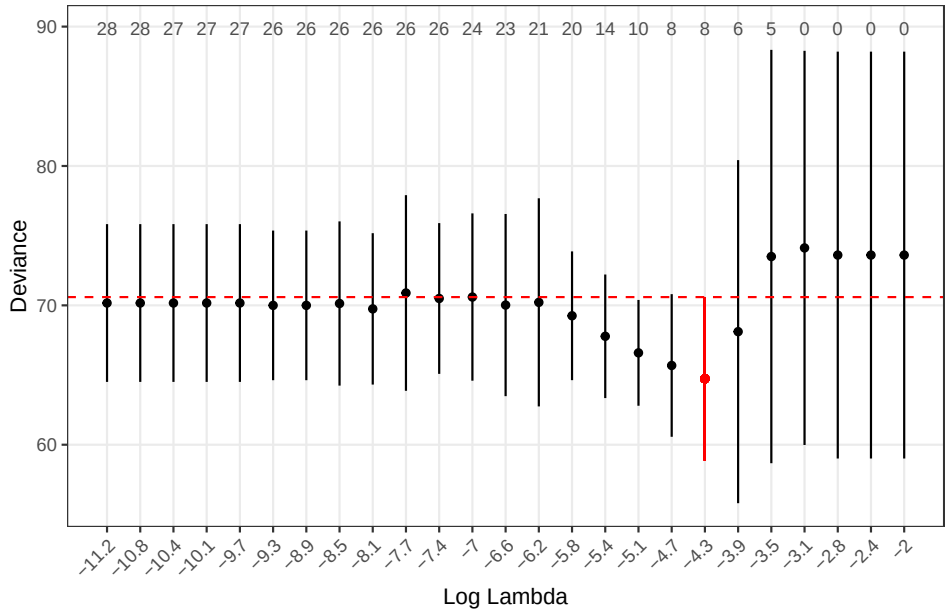
The categorical variables `ram` and `regular_drug` were dropped because they each had only 5 patients responding with "yes". The variables `edu_level_mom` and `edu_level_dad` could potentially be combined, but they were not included in the data dictionary so they were dropped. The `monitor1` variable was also dropped because it was not clear how to relevel it.

After processing, the dataset contained 30 covariates (listed in the Appendix).

Model Fitting

Out model-fitting algorithm uses a parameter λ_0 that penalizes nonzero coefficients. In other words, a higher value of λ_0 will result in fewer covariates being included in the risk score model. To determine the best value for λ_0, we run cross-validation. Figure 1 plots the mean model deviance for a range of potential λ_0 values. The best fitting model will have the lowest deviance. The numbers at the top of the plot show the number of nonzero coefficients in the model fit with each value of λ_0. We can see that the λ_0 producing the lowest mean deviance results in a model with 4 features. Note that due to the relatively small size of this data set, these cross validation findings aren't robust. Changing the randomization seed will cause models with different numbers of nonzero coefficients to have the lowest deviance, including models with no nonzero coefficients. In fact, we found that a model with no nonzero coefficients was returns the majority of the time, suggesting that there isn't a high signal between the covariates and the outcome. However, for the purpose of this example, we chose a randomization seed that would return a model with nonzero coefficients.

Figure 1. Cross Validation Results



The score card for the model selected by cross-validation is presented in Table 2. A patient's total score can be calculated by multiplying each variable's response by the number of points shown on the right and then adding the points together. For example, if a patient's responses to the variables listed in Table 2 were 1, 2, 4, 2, 5, 0, 5, and 0, respectively, their total score would be $1(-4)+2(1)+4(-1)+2(1)+5(1)+0(-3)+5(-1)+0(5) = -4$. Table 3 can then be used to map this score to its associated risk of non-adherence. For this example patient, their risk of non-adherence would be 47%.

Table 2: Score Card

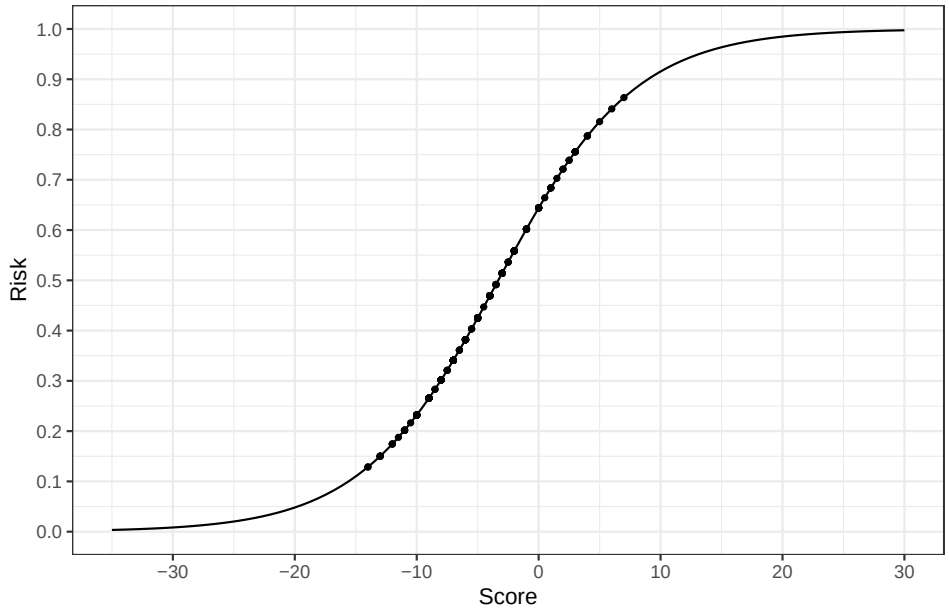
	Points
Lives with Mom (0/1)	-4
Pills Value (1-5)	1
Accompanied by Family (1-5)	-1
Family Dislikes Friends (1-5)	1
Feels Ashamed in Health Center (1-5)	1
Has Never Had Covid (0/1)	-3
Median Motivation Score (1-5)	-1
Treatment >6 months (0/1)	5

Table 3: Score-Risk Map

Score	-14	-12	-10	-8	-6	-4	-2	0	2	4	6	8	10	12	14	16	18
Risk	0.13	0.17	0.23	0.3	0.38	0.47	0.56	0.64	0.72	0.79	0.84	0.88	0.92	0.94	0.96	0.97	0.98

Figure 2 visualizes the logistic regression curve of this model. The observed scores from this dataset are plotted as points along the curve.

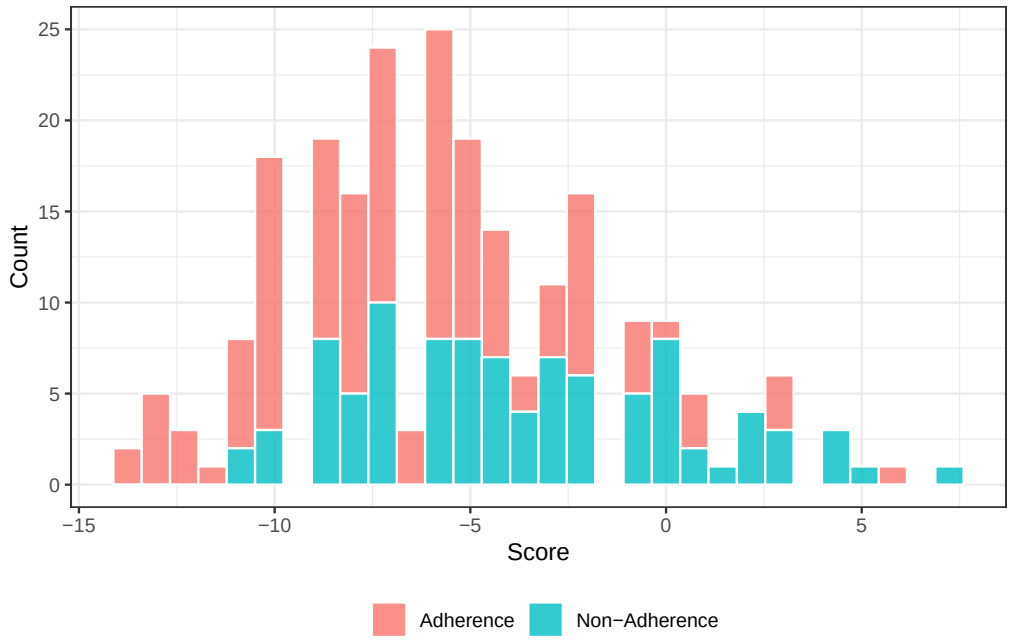
Figure 2. Risk Score Model



Model Evaluation

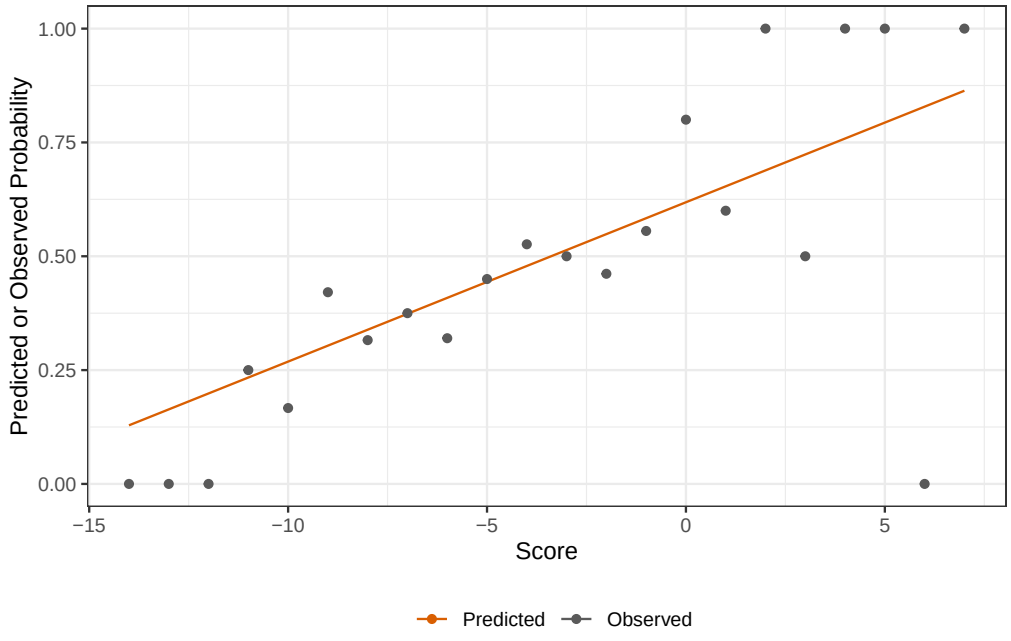
Figure 3 plots the distribution of scores by adherence outcome. We can see that those with lower scores tended to be adherent while those with higher score tended to be non-adherent.

Figure 3. Distribution of Scores by Outcome



We can also visualize the relationship between score and probability of adherence as in Figure 4. The grey points indicate the observed probability of adherence for each observed score. The orange line indicates the predicted probability of adherence associated with each score. We can see that this line tends to overestimate the probability with scores less than -10 and tends to underestimate the probability with scores greater than 3. We also observed an outlier with a score greater than 5 (predicted non-adherent) who did adhere to their medication.

Figure 4. Predicted vs. Observed Probabilities per Score



We compared the performance of our risk score model to logistic regression, lasso regression, and rounded logistic regression. The coefficients for each of these models are reported in Table 4. As expected, the logistic regression model assigns a non-zero coefficient value to each covariate, while the lasso model shrinks many coefficients to zero. In fact, the lasso model only selected one covariate: `tx_mos_cat> 6 mos` (an indicator

for treatment longer than 6 months).

Performance metrics for these models are reported in Table 5. Overall, we observed that variable selection reduced the performance of the models, as the logistic and rounded regression models had AUC values above 0.75, while the risk score and lasso models had AUC values of 0.69 and 0.63. However, these metrics were evaluated using the training data, where the logistic and rounded logistic models were likely highly overfit. Rerunning these models on a validation data set would help estimate their true performance.

We do observe that our risk score model had a higher AUC than the lasso model. Although the lasso model had higher specificity, it had much lower sensitivity than our model.

Table 4: Model Coefficients

	Risk	Logistic	Lasso	Rounded
gendermale	0	-0.187	0.000	-8
case_findingpassive case finding	0	0.337	0.000	15
concomitant_tbyes	0	0.195	0.000	9
lives_w_momyes	-4	-1.287	0.000	-57
lives_w_parentsno parents	0	-1.485	0.000	-65
lives_w_parentssingle dad	0	-0.434	0.000	-19
lives_w_parentssingle mom	0	0.272	0.000	12
current_sx_none1	0	0.521	0.000	23
pills	1	0.182	0.000	8
dosis_fijasyes	0	0.116	0.000	5
daily_contyes	0	1.107	0.000	49
regimensubsequent	0	-0.468	0.000	-21
monoRyes	0	-0.420	0.000	-19
adr_freq	0	0.193	0.000	8
fam_accompany_dot	-1	-0.190	0.000	-8
fam_dislikefriends	1	0.186	0.000	8
autonomy_obedient	0	-0.009	0.000	0
tobacco_freq	0	-0.102	0.000	-5
drug_useyes	0	-0.531	0.000	-23
stig_health_ctr	1	0.171	0.000	8
tto_anterior_tb1	0	0.108	0.000	5
prior_covidno	-3	-1.045	0.000	-46
prior_covidsuspected (unconfirmed)	0	-0.952	0.000	-42
covid_es	0	-0.202	0.000	-9
psych_interventionno intervention needed	0	0.074	0.000	3
psych_interventionnot evaluated	0	0.593	0.000	26
psych_interventionSAME	0	0.481	0.000	21
family_median	0	0.116	0.000	5
health_svc_median	0	0.123	0.000	5
motiv_median	-1	-0.361	0.000	-16
knowledge_median	0	-0.023	0.000	-1
age_cat16-17	0	-0.100	0.000	-4
age_cat18+	0	-0.322	0.000	-14
audit_cat0	0	-0.619	0.000	-27
ace_cat0	0	-0.461	0.000	-20
ace_cat1	0	-0.243	0.000	-11
tx_mos_cat> 6 mos	5	1.150	0.139	51

Table 5: Model Performance

	Threshold	Specificity	Sensitivity	AUC
Risk	0.414	0.657	0.625	0.690
Logistic	0.403	0.724	0.719	0.751
Lasso	0.442	0.784	0.479	0.631
Rounded	0.396	0.716	0.729	0.752

Conclusion

Although our risk score algorithm usually selected models with no nonzero coefficients, we were able to set the randomization seed to produce a model that could be used as an example. The fact that both our algorithm and the lasso algorithm tended to choose no covariates suggest that these variables do not have a strong relationship to the adherence outcome or there isn't a high enough sample size to detect this relationship.

Despite these issues, our example model output made sense intuitively. It makes sense that a patient living with their mom, being accompanied by family to the health clinic, and having high motivation would increase the chance of adherence. Likewise, it makes sense that having more pills to take, having a family who dislikes their friends, feeling ashamed at the clinic, and having a longer treatment would increase the risk of non-adherence. Although the performance of this model was fairly low (AUC = 0.69), it may still help with identifying which factors are the most important in predicting non-adherence.

EDA Appendix

Figure A1. Outcome Distributions

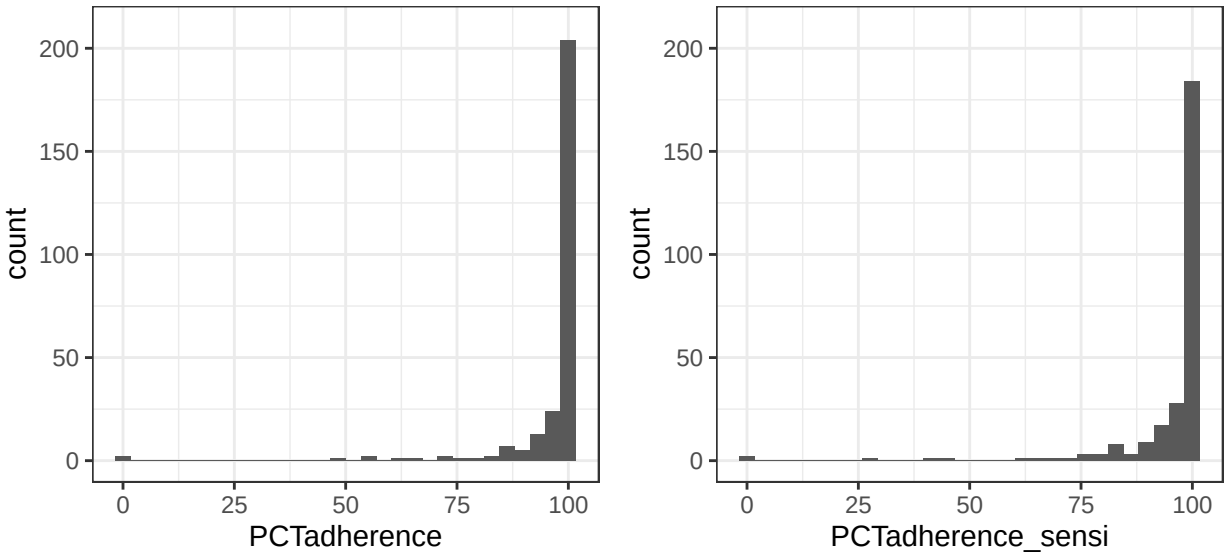


Figure A2. Continuous Covariate Distributions

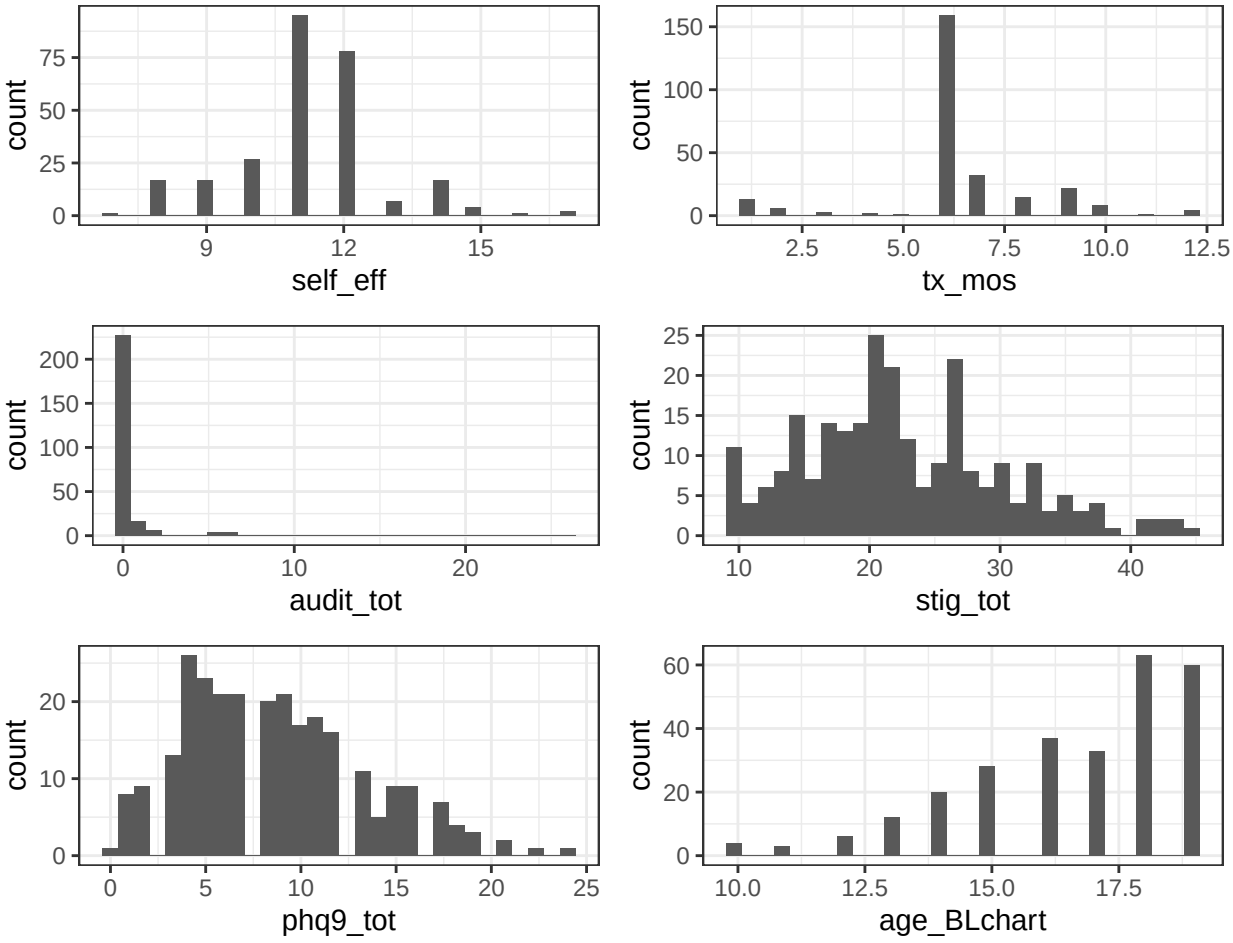


Table A1. Processed Variable Summary

Characteristic	N = 230
gender	
female	84 (37%)
male	146 (63%)
case_finding	
active case finding	7 (3.0%)
passive case finding	223 (97%)
concomitant_tb	23 (10%)
lives_w_mom	199 (87%)
lives_w_parents	
2 parents	133 (58%)
no parents	14 (6.1%)
single dad	15 (6.5%)
single mom	68 (30%)
current_sx_none	
0	179 (78%)
1	51 (22%)
pills	
1	36 (16%)
2	46 (20%)
3	23 (10%)
4	125 (54%)
dosis_fijas	77 (33%)
daily_cont	25 (11%)
regimen	
first	216 (94%)
subsequent	14 (6.1%)
monoR	23 (10%)
adr_freq	
0	93 (40%)
1	76 (33%)
2	40 (17%)
3	13 (5.7%)
4	8 (3.5%)
fam_accompany_dot	
1	37 (16%)
2	35 (15%)
3	45 (20%)
4	18 (7.8%)
5	95 (41%)
fam_dislikefriends	
1	81 (35%)
2	76 (33%)
3	47 (20%)
4	13 (5.7%)
5	13 (5.7%)
autonomy_obedient	
1	1 (0.4%)
2	6 (2.6%)
3	43 (19%)
4	76 (33%)

(continued)

Characteristic	**N = 230**
5	104 (45%)
tobacco_freq	
0	196 (85%)
1	27 (12%)
2	6 (2.6%)
3	1 (0.4%)
drug_use	29 (13%)
stig_health_ctr	
1	146 (63%)
2	39 (17%)
3	26 (11%)
4	10 (4.3%)
5	9 (3.9%)
tto_anterior_tb	
0	226 (98%)
1	4 (1.7%)
prior_covid	
confirmed	26 (11%)
no	166 (72%)
suspected (unconfirmed)	38 (17%)
covid_es	
0	72 (31%)
1	128 (56%)
2	29 (13%)
3	1 (0.4%)
psych_intervention	
MINSA referral	28 (12%)
no intervention needed	52 (23%)
not evaluated	98 (43%)
SAME	52 (23%)
adherence_outcome	96 (42%)
family_median	
1	2 (0.9%)
1.5	1 (0.4%)
2	4 (1.7%)
2.5	2 (0.9%)
3	31 (13%)
3.5	13 (5.7%)
4	57 (25%)
4.5	9 (3.9%)
5	111 (48%)
health_svc_median	
3	5 (2.2%)
3.5	9 (3.9%)
4	68 (30%)
4.5	31 (13%)
5	117 (51%)
motiv_median	
2	1 (0.4%)
2.5	1 (0.4%)
3.5	1 (0.4%)

(continued)

Characteristic	**N = 230**
4	35 (15%)
4.5	25 (11%)
5	167 (73%)
knowledge_median	
2	2 (0.9%)
3	57 (25%)
3.5	3 (1.3%)
4	168 (73%)
age_cat	
< 16	65 (28%)
16-17	62 (27%)
18+	103 (45%)
audit_cat	
>0	37 (16%)
0	193 (84%)
ace_cat	
> 1	107 (47%)
0	63 (27%)
1	60 (26%)
tx_mos_cat	
<= 6 mos	155 (67%)
> 6 mos	75 (33%)

[1] n (%)

Code Appendix

```
knitr::opts_chunk$set(echo = FALSE)

library(knitr)
library(xtable)
library(glmnet)
library(pROC)
library(tableone)
library(glmnet)
library(caret)
library(gridExtra)
library(kableExtra)
library(gtsummary)
library(riskscores)
#source('risk.R')

raw_data <- read.csv("../data/Peru_TB_data.csv")

data.frame(some_missed = c(sum(raw_data$PCTadherence < 100, na.rm = TRUE),
                           sum(raw_data$PCTadherence_sensi < 100, na.rm  = TRUE)),
           none_missed = c(sum(raw_data$PCTadherence == 100, na.rm = TRUE),
                           sum(raw_data$PCTadherence_sensi == 100, na.rm  = TRUE)),
           missing = c(sum(is.na(raw_data$PCTadherence)),
                       sum(is.na(raw_data$PCTadherence_sensi))),
           row.names = c("PCTadherence", "PCTadherence_sensi"))  %>%

  kableExtra::kbl(format = "latex", caption = "Patient Count by Dichotomized Outcome",
                  col.names = c("<100% Adherence", "100% Adherence", "Missing"),
                  booktabs = TRUE)  %>%

  kableExtra::kable_styling(latex_options = c("HOLD_position"), font_size = 10)



source('tb_preprocessing.R')
tb_df <- tb_preprocessing(raw_data)
tb_mat <- tb_as_matrix(tb_df)


X <- as.matrix(tb_mat[,-ncol(tb_mat)])
y <- tb_mat[,ncol(tb_mat)]

# CV
# get folds
folds <- stratify_folds(y, nfolds = 5, seed = 5)

cv_results <- cv_risk_mod(X, y, foldids = folds, a = -5, b = 5)

plot(cv_results, lambda_text = FALSE) +
  labs(title = "Figure 1. Cross Validation Results")
```

```r
mod <- risk_mod(X, y, lambda0 = cv_results$lambda_min, a = -5, b = 5)


data.frame(mod$model_card, row.names = c("Lives with Mom (0/1)",
                    "Pills Value (1-5)",
                    "Accompanied by Family (1-5)",
                    "Family Dislikes Friends (1-5)",
                    "Feels Ashamed in Health Center (1-5)",
                    "Has Never Had Covid (0/1)",
                    "Median Motivation Score (1-5)",
                    "Treatment >6 months (0/1)")) %>%

  kableExtra::kbl(format = "latex", caption = "Score Card" ,
                  booktabs = TRUE, linesep = "")  %>%
  kableExtra::kable_styling(latex_options = c("HOLD_position"), font_size = 10)

score_range <- seq(-14, 18, 2)

data.frame(Score = as.character(score_range),
           Risk = as.character(round(get_risk(mod, score_range), 2))) %>%
  t() %>%

  kableExtra::kbl(align = "c", format = "latex", caption = "Score-Risk Map" ,
                  booktabs = TRUE)  %>%
  kableExtra::kable_styling(latex_options = c("HOLD_position"), font_size = 8) %>%
  row_spec(row = 1, hline_after = TRUE)



predict_df <- data.frame(score = predict(mod, type = "score"),
                         response = predict(mod, type = "response"))

ggplot() +
  geom_point(data = predict_df, aes(x = score, y = response), size = 1) +
  geom_function(data = data.frame(x = seq(-35,30)), aes(x),
                fun = function(x) get_risk(mod, x)) +
  labs(x = "Score", y = "Risk", title = "Figure 2. Risk Score Model") +
  scale_x_continuous(breaks = seq(-60, 50, 10)) +
  scale_y_continuous(breaks = seq(0, 1, 0.10)) +
  #geom_point(aes(x = -14, y = get_risk(mod, -14)), color = "blue",
  #           shape = 3, size = 3) +
  #geom_point(aes(x = 18, y = get_risk(mod, 18)), color = "blue",
  #           shape = 3, size = 3) +

  theme_bw()
ggplot() +
  geom_histogram(aes(x = predict_df$score, fill = tb_df$adherence_outcome),
                 alpha = 0.8, color = "white") +
  labs(x = "Score", y = "Count", fill = "", title = "Figure 3. Distribution of Scores by Outcome") +
  scale_fill_discrete(labels = c("Adherence", "Non-Adherence")) +
  theme_bw() +
  theme(legend.position = "bottom")
```

```r
# for each score find predicted probability and also find percent class 1
range_scores <- range(predict_df$score)
all_scores <- unique(predict_df$score)
vals <- mod$gamma*(mod$beta[1]+range_scores)
probs <- exp(vals)/(1+exp(vals))
props <- tb_df %>%
  mutate(rnd_scores = floor(predict_df$score)) %>%
  group_by(rnd_scores) %>%
  summarize(prop = sum(adherence_outcome)/n())

ggplot() +
  geom_line(aes(x=range_scores, y=probs, color = "Predicted")) +
  geom_point(aes(x=props$rnd_scores,y=props$prop, color = "Observed")) +
  scale_color_manual(name = "",
                     breaks = c("Predicted", "Observed"),
                     values = c("Predicted" = "#d95f02", "Observed" = "#5A5A5A")) +
  labs(x="Score", y="Predicted or Observed Probability",
       title = "Figure 4. Predicted vs. Observed Probabilities per Score") +
  theme_bw() +
  theme(legend.position = "bottom")

coef_vals <- matrix(0, ncol=4, nrow=ncol(X)-1)

# risk model prediction
coef_vals[, 1] <- coef(mod)[-1]
risk_probs <- predict_df$response
risk_pred <- as.factor(ifelse(risk_probs < 0.5, FALSE, TRUE))

# glm prediction
glm_mod <- glm(y~X-1, family = "binomial")
coef_vals[, 2] <- coef(glm_mod)[-1]
glm_probs<- predict(glm_mod, type="response")
glm_pred <- as.factor(ifelse(glm_probs < 0.5, FALSE, TRUE))

# lasso prediction
lasso_res <- cv.glmnet(x=X[,-1], y=y, alpha=1)
lasso_mod <- glmnet(x=X[,-1], y=y, lambda=lasso_res$lambda.min, alpha=1)
coef_vals[,3] <- coef(lasso_mod)[-1]
lasso_probs <- as.vector(predict(lasso_mod, newx=X[,-1]))
lasso_pred <- as.factor(ifelse(lasso_probs < 0.5, FALSE, TRUE))

# rounded logistic
coef_vals[,4] <- round(coef_vals[,2] / abs(median(coef_vals[,2])),0)
x_vars <- tb_mat[,-c(1,ncol(tb_mat))]
round_score <- x_vars %*% coef_vals[,4]
mod_round <- glm(tb_mat[,ncol(tb_mat)] ~ round_score, family = "binomial")
round_probs <- predict(mod_round, type="response")
round_pred <- as.factor(ifelse(round_probs < 0.5, FALSE, TRUE))



# confusion matrices
confusionMatrix(factor(tb_df$adherence_outcome), risk_pred)
```

```r
confusionMatrix(factor(tb_df$adherence_outcome), glm_pred)
confusionMatrix(factor(tb_df$adherence_outcome), lasso_pred)
confusionMatrix(factor(tb_df$adherence_outcome), round_pred)

# discrimination
risk_roc <- roc(factor(tb_df$adherence_outcome), risk_probs)
glm_roc <- roc(factor(tb_df$adherence_outcome), glm_probs)
lasso_roc <- roc(factor(tb_df$adherence_outcome), lasso_probs)
round_roc <- roc(factor(tb_df$adherence_outcome), round_probs)


data.frame(coef_vals,
           row.names = dimnames(X)[[2]][-1]) %>%
  kbl(digits = 3, col.names = c("Risk", "Logistic", "Lasso", "Rounded"),
      booktabs = T, caption = "Model Coefficients") %>%

  kableExtra::kable_styling(font_size = 9,
                            latex_options = c("repeat_header", "HOLD_position"))
rbind(coords(risk_roc, "best"), coords(glm_roc, "best"),
      coords(lasso_roc, "best"), coords(round_roc, "best")) %>%
  data.frame(auc = c(risk_roc$auc[[1]], glm_roc$auc[[1]], lasso_roc$auc[[1]],
           round_roc$auc[[1]]),
           row.names = c("Risk", "Logistic", "Lasso", "Rounded")) %>%
  kbl(digit = 3,
      col.names = c("Threshold", "Specificity", "Sensitivity", "AUC"),
      caption = "Model Performance",
      booktabs = T) %>%

  kableExtra::kable_styling(font_size = 9,
                            latex_options = c("repeat_header", "HOLD_position"))
# outcome distributions
p1 <- ggplot(raw_data) +
  geom_histogram(aes(x = PCTadherence)) +
  lims(y = c(0, 210)) +
  theme_bw()

p2 <- ggplot(raw_data) +
  geom_histogram(aes(x = PCTadherence_sensi)) +
  lims(y = c(0, 210)) +
  theme_bw()

grid.arrange(p1, p2, ncol = 2)
p3 <- ggplot(raw_data) +
  geom_histogram(aes(x = self_eff)) +
  theme_bw()

p4 <- ggplot(raw_data) +
  geom_histogram(aes(x = tx_mos)) +
  theme_bw()

p5 <- ggplot(raw_data) +
  geom_histogram(aes(x = audit_tot)) +
  theme_bw()
```

```r
p6 <- ggplot(raw_data) +
  geom_histogram(aes(x = stig_tot)) +
  theme_bw()

p7 <- ggplot(raw_data) +
  geom_histogram(aes(x = phq9_tot)) +
  theme_bw()

p8 <- ggplot(raw_data) +
  geom_histogram(aes(x = age_BLchart)) +
  theme_bw()

p9 <- ggplot(raw_data) +
  geom_histogram(aes(x = ace_score)) +
  theme_bw()

grid.arrange(p3, p4, p5, p6, p7, p8, ncol = 2)
tbl_summary(tb_df)%>%
   as_kable_extra(booktabs = TRUE,
                  longtable = TRUE) %>%
  kableExtra::kable_styling(font_size = 9,
                            latex_options = c("repeat_header", "HOLD_position"))



# load risk model files
X <- tb_matrix[, -ncol(tb_matrix)]
y <- tb_matrix[, ncol(tb_matrix)]

coef_vals <- matrix(0, ncol=3, nrow=ncol(X)-1)

# risk model prediction
risk_output_cv <- cv_risk_mod(X, y, a=-5, b=5)
risk_output <- risk_mod(X, y, a=-5, b=5, lambda0=risk_output_cv$lambda_min)
coef_vals[,1] <- risk_output$beta[-1]
risk_probs <- predict(risk_output$glm_mod, type="response")
risk_pred <- as.factor(ifelse(risk_probs < 0.5, FALSE, TRUE))

# glm prediction
glm_mod <- glm(y~X-1, family = "binomial")
coef_vals[, 2] <- coef(glm_mod)[-1]
glm_probs<- predict(glm_mod, type="response")
glm_pred <- as.factor(ifelse(glm_probs < 0.5, FALSE, TRUE))

# lasso prediction

lasso_res <- cv.glmnet(x=X[,-1], y=y, alpha=1)
lasso_mod <- glmnet(x=X[,-1], y=y, lambda=lasso_res$lambda.min, alpha=1)
coef_vals[,3] <- coef(lasso_mod)[-1]
lasso_probs <- as.vector(predict(lasso_mod, newx=X[,-1]))
lasso_pred <- as.factor(ifelse(lasso_probs < 0.5, FALSE, TRUE))

print(coef_vals)
```

```r
# confusion matrices
confusionMatrix(as.factor(tb_df$adherence_outcome), risk_pred)
confusionMatrix(as.factor(tb_df$adherence_outcome), glm_pred)
confusionMatrix(as.factor(tb_df$adherence_outcome), lasso_pred)

# discrimination
roc(as.factor(tb_df$adherence_outcome), risk_probs)
roc(as.factor(tb_df$adherence_outcome), glm_probs)
roc(as.factor(tb_df$adherence_outcome), lasso_probs)

# find risk score probs to summarize
tb_df$scores <- (X[,-1] %*% risk_output$beta[-1])
ggplot(tb_df)+geom_histogram(aes(x=scores, fill=adherence_outcome), alpha=0.5)

# for each score find predicted probability and also find percent class 1
range_scores <- range(tb_df$scores)
all_scores <- seq(range_scores[1], range_scores[2])
vals <- risk_output$gamma*(risk_output$beta[1]+range_scores)
probs <- exp(vals)/(1+exp(vals))
props <- tb_df %>%
  mutate(rnd_scores = floor(scores)) %>%
  group_by(rnd_scores) %>%
  summarize(prop = sum(adherence_outcome)/n())

ggplot()+geom_line(aes(x=range_scores, y=probs)) +
  geom_point(aes(x=props$rnd_scores,y=props$prop)) +
  labs(x="Score", y="Predicted or Observed Probability")
```