try 2,5,10 fold cross validation

Chongjun Liao

12/1/2021

```
# original Rsquared
# table 2
## sat on income
library(caret)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

```
merged_df <- read.csv("merged_final.csv")
og_df <- merged_df[merged_df$FAMILY_INCOME > 10000, ]
og_df <- og_df[!is.na(og_df$FAMILY_INCOME),]
set.seed(1993)
merged_folds <- trainControl(method = "cv", number = 10)

merged_topics <- as.matrix(log(og_df[,6:75]))
# Drop min cor, topic 16
merged_topics <- merged_topics[,-16]
topic_total_df <- as.data.frame(cbind(og_df$RSAT_TOTAL_SCORE, merged_topics))
topic_total_mod <- train(V1 ~ ., method = "lm",
                    data = topic_total_df, trControl = merged_folds)
topic_ebrw_df <- as.data.frame(cbind(og_df$RSAT_EBRW, merged_topics))
topic_ebrw_mod <- train(V1 ~ ., method = "lm",
                    data = topic_ebrw_df, trControl = merged_folds)
topic_math_df <- as.data.frame(cbind(og_df$RSAT_MATH_SCORE, merged_topics))
topic_math_mod <- train(V1 ~ ., method = "lm",
                    data = topic_math_df, trControl = merged_folds)
# liwc
merged_liwc <- as.matrix(og_df[,76:167])
liwc_total_df <- as.data.frame(cbind(og_df$RSAT_TOTAL_SCORE, merged_liwc))
liwc_total_mod <- train(V1 ~ ., method = "lm",
                    data = liwc_total_df, trControl = merged_folds)
liwc_ebrw_df <- as.data.frame(cbind(og_df$RSAT_EBRW, merged_liwc))
liwc_ebrw_mod <- train(V1 ~ ., method = "lm",
                    data = liwc_ebrw_df, trControl = merged_folds)
liwc_math_df <- as.data.frame(cbind(og_df$RSAT_MATH_SCORE, merged_liwc))
liwc_math_mod <- train(V1 ~ ., method = "lm",
                    data = liwc_math_df, trControl = merged_folds)
rsquared_t1 = rbind(topic_total_mod$results$Rsquared,
                    topic_ebrw_mod$results$Rsquared,
                    topic_math_mod$results$Rsquared,
```

```
                    liwc_total_mod$results$Rsquared,
                    liwc_ebrw_mod$results$Rsquared,
                    liwc_math_mod$results$Rsquared)
RMSE_t1 = rbind(topic_total_mod$results$RMSE,
                    topic_ebrw_mod$results$RMSE,
                    topic_math_mod$results$RMSE,
                    liwc_total_mod$results$RMSE,
                    liwc_ebrw_mod$results$RMSE,
                    liwc_math_mod$results$RMSE)
```

```
merged_folds <- trainControl(method = "cv", number = 2)
#topics
topic_total_mod <- train(V1 ~ ., method = "lm",
                    data = topic_total_df, trControl = merged_folds)
topic_ebrw_mod <- train(V1 ~ ., method = "lm",
                    data = topic_ebrw_df, trControl = merged_folds)
topic_math_mod <- train(V1 ~ ., method = "lm",
                    data = topic_math_df, trControl = merged_folds)
#liwc
liwc_total_mod <- train(V1 ~ ., method = "lm",
                    data = liwc_total_df, trControl = merged_folds)
liwc_ebrw_mod <- train(V1 ~ ., method = "lm",
                    data = liwc_ebrw_df, trControl = merged_folds)
liwc_math_mod <- train(V1 ~ ., method = "lm",
                    data = liwc_math_df, trControl = merged_folds)
rsquared_t2 = rbind(topic_total_mod$results$Rsquared,
                    topic_ebrw_mod$results$Rsquared,
                    topic_math_mod$results$Rsquared,
                    liwc_total_mod$results$Rsquared,
                    liwc_ebrw_mod$results$Rsquared,
                    liwc_math_mod$results$Rsquared)
RMSE_t2 = rbind(topic_total_mod$results$RMSE,
                    topic_ebrw_mod$results$RMSE,
                    topic_math_mod$results$RMSE,
                    liwc_total_mod$results$RMSE,
                    liwc_ebrw_mod$results$RMSE,
                    liwc_math_mod$results$RMSE)
```

```
merged_folds <- trainControl(method = "cv", number = 5)
#topics
topic_total_mod <- train(V1 ~ ., method = "lm",
                    data = topic_total_df, trControl = merged_folds)
topic_ebrw_mod <- train(V1 ~ ., method = "lm",
                    data = topic_ebrw_df, trControl = merged_folds)
topic_math_mod <- train(V1 ~ ., method = "lm",
                    data = topic_math_df, trControl = merged_folds)
#liwc
liwc_total_mod <- train(V1 ~ ., method = "lm",
                    data = liwc_total_df, trControl = merged_folds)
liwc_ebrw_mod <- train(V1 ~ ., method = "lm",
                    data = liwc_ebrw_df, trControl = merged_folds)
liwc_math_mod <- train(V1 ~ ., method = "lm",
                    data = liwc_math_df, trControl = merged_folds)
```

```
rsquared_t3 = rbind(topic_total_mod$results$Rsquared,
                    topic_ebrw_mod$results$Rsquared,
                    topic_math_mod$results$Rsquared,
                    liwc_total_mod$results$Rsquared,
                    liwc_ebrw_mod$results$Rsquared,
                    liwc_math_mod$results$Rsquared)
RMSE_t3 = rbind(topic_total_mod$results$RMSE,
                    topic_ebrw_mod$results$RMSE,
                    topic_math_mod$results$RMSE,
                    liwc_total_mod$results$RMSE,
                    liwc_ebrw_mod$results$RMSE,
                    liwc_math_mod$results$RMSE)
```

```
merged_folds <- trainControl(method = "cv", number = 20)
#topics
topic_total_mod <- train(V1 ~ ., method = "lm",
                    data = topic_total_df, trControl = merged_folds)
topic_ebrw_mod <- train(V1 ~ ., method = "lm",
                    data = topic_ebrw_df, trControl = merged_folds)
topic_math_mod <- train(V1 ~ ., method = "lm",
                    data = topic_math_df, trControl = merged_folds)
#liwc
liwc_total_mod <- train(V1 ~ ., method = "lm",
                    data = liwc_total_df, trControl = merged_folds)
liwc_ebrw_mod <- train(V1 ~ ., method = "lm",
                    data = liwc_ebrw_df, trControl = merged_folds)
liwc_math_mod <- train(V1 ~ ., method = "lm",
                    data = liwc_math_df, trControl = merged_folds)
rsquared_t4 = rbind(topic_total_mod$results$Rsquared,
                    topic_ebrw_mod$results$Rsquared,
                    topic_math_mod$results$Rsquared,
                    liwc_total_mod$results$Rsquared,
                    liwc_ebrw_mod$results$Rsquared,
                    liwc_math_mod$results$Rsquared)
RMSE_t4 = rbind(topic_total_mod$results$RMSE,
                    topic_ebrw_mod$results$RMSE,
                    topic_math_mod$results$RMSE,
                    liwc_total_mod$results$RMSE,
                    liwc_ebrw_mod$results$RMSE,
                    liwc_math_mod$results$RMSE)
```

```
table = data.frame(rsquared_t1,rsquared_t2,rsquared_t3,rsquared_t4)
rownames(table) = c("topics SAT composite","topics SAT EBRW","topics SAT Math",
                    "liwc SAT composite","liwc SAT EBRW","liwc SAT Math")
colnames(table) = c("k=10","k=2","k=5","k=20")
knitr::kable(table,digits = 3,caption = "Rsquared 10,2,5,20-folds cross validation")
```

Table 1: Rsquared 10,2,5,20-folds cross validation

	k=10	k=2	k=5	k=20
topics SAT composite	0.482	0.482	0.482	0.483

	k=10	k=2	k=5	k=20
topics SAT EBRW	0.423	0.423	0.423	0.423
topics SAT Math	0.470	0.470	0.470	0.470
liwc SAT composite	0.432	0.431	0.432	0.432
liwc SAT EBRW	0.365	0.364	0.366	0.365
liwc SAT Math	0.402	0.401	0.402	0.402

```
table = data.frame(RMSE_t1,RMSE_t2,RMSE_t3,RMSE_t4)
rownames(table) = c("topics SAT composite","topics SAT EBRW","topics SAT Math",
                    "liwc SAT composite","liwc SAT EBRW","liwc SAT Math")
colnames(table) = c("k=10","k=2","k=5","k=20")
knitr::kable(table,digits = 3,caption = "RMSE 10,2,5,20-folds cross validation")
```

Table 2: RMSE 10,2,5,20-folds cross validation

	k=10	k=2	k=5	k=20
topics SAT composite	124.401	124.460	124.395	124.403
topics SAT EBRW	64.635	64.681	64.644	64.634
topics SAT Math	74.104	74.142	74.120	74.099
liwc SAT composite	130.306	130.426	130.357	130.292
liwc SAT EBRW	67.800	67.879	67.797	67.804
liwc SAT Math	78.715	78.774	78.715	78.703