Using `tidymodels`

Let's first split our data into training and testing datasets:

```r
set.seed(1)
split <- initial_split(data = gss_subset, prop = 3/4)
gss_train <- training(split)
gss_test <- testing(split)
```

Next, let's use 5-fold cross validation:

```r
folds <- rsample::vfold_cv(gss_train, v = 5)
```

Now, let's make our `recipe()` and `workflow()` that will be used for each of our models:

```r
# Create the recipe
gss_recipe <- recipe(partyid ~ ., data = gss_train) %>%
  step_rm(year) %>%
  step_dummy(all_nominal(), -all_outcomes()) %>%
  step_zv(all_predictors())

# Create the workflow
gss_workflow <- workflow() %>%
  add_recipe(gss_recipe)

# View the workflow
gss_workflow
```

```
## == Workflow ====================================================================
## Preprocessor: Recipe
## Model: None
##
## -- Preprocessor ----------------------------------------------------------------
## 3 Recipe Steps
##
## * step_rm()
## * step_dummy()
## * step_zv()
```

Now, we can begin to specify our models for our model stack:

```r
# Basic Logistic regression specification
basic_logreg_spec <- logistic_reg() %>%
  set_engine("glm")
# Add logistic regression to workflow
basic_logreg_workflow <- gss_workflow %>%
  add_model(basic_logreg_spec)
# Fit with our cross validation
set.seed(13)
basic_logreg_resamples <- fit_resamples(
  basic_logreg_workflow,
  resamples = folds,
  control = control_stack_resamples()
)
```

```
##
## Attaching package: 'rlang'

## The following objects are masked from 'package:purrr':
##
##     %@%, as_function, flatten, flatten_chr, flatten_dbl, flatten_int,
##     flatten_lgl, flatten_raw, invoke, list_along, modify, prepend,
##     splice

##
## Attaching package: 'vctrs'

## The following object is masked from 'package:dplyr':
##
##     data_frame

## The following object is masked from 'package:tibble':
##
##     data_frame

## The workflow being saved contains a recipe, which is 0.4 Mb in memory. If this was not intentional,
```

```
# Penalized Logistic regression specification
logreg_spec <- logistic_reg(penalty = tune(), mixture = 1) %>%
  set_engine("glmnet")
# add grid
lr_reg_grid <- tibble(penalty = 10^seq(-4, -1, length.out = 30))
# Add logistic regression to workflow
logreg_workflow <- gss_workflow %>%
  add_model(logreg_spec)
# Fit with our cross validation
set.seed(13)
logreg_resamples <- tune_grid(
  logreg_workflow,
  resamples = folds,
  grid = lr_reg_grid,
  control = control_stack_grid()
)
```

```
## Loading required package: Matrix

##
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack

## Loaded glmnet 4.0-2

## The workflow being saved contains a recipe, which is 0.4 Mb in memory. If this was not intentional,
```

```r
# LDA specification
lda_spec <- discrim_linear(penalty = tune()) %>%
  set_engine("mda")
# Add LDA to workflow
lda_workflow <- gss_workflow %>%
  add_model(lda_spec)
# Fit with our cross validation
set.seed(13)
lda_resamples <- tune_grid(
  lda_workflow,
  resamples = folds,
  control = control_stack_grid()
)
```

```
## Loading required package: class

## Loaded mda 0.5-2

##
## Attaching package: 'mda'

## The following object is masked from 'package:parsnip':
##
##     mars

## The workflow being saved contains a recipe, which is 0.4 Mb in memory. If this was not intentional, 
```

```r
# SVM specification (ooh fancy)
svm_spec <- svm_rbf(
    cost = tune(),
    rbf_sigma = tune()
  ) %>%
  set_engine("kernlab") %>%
  set_mode("classification")

# WF
svm_workflow <-
  gss_workflow %>%
  add_model(svm_spec)

# tuning
set.seed(13)
svm_res <-
  tune_grid(
    svm_workflow,
    resamples = folds,
    grid = 5,
    control = control_stack_grid()
  )
```

```
##
## Attaching package: 'kernlab'
## The following object is masked from 'package:scales':
##
##     alpha

## The following object is masked from 'package:purrr':
##
##     cross

## The following object is masked from 'package:ggplot2':
##
##     alpha

## The workflow being saved contains a recipe, which is 0.4 Mb in memory. If this was not intentional, p
```

Now, we can stack these models:

```
gss_stack <- stacks() %>%
  add_candidates(basic_logreg_resamples) %>%
  add_candidates(logreg_resamples) %>%
  add_candidates(lda_resamples) %>%
  add_candidates(svm_res)
```

```
## Warning: Predictions from the candidates c(".pred_dem_logreg_resamples_1_02",
## ".pred_rep_logreg_resamples_1_02") were identical to those from existing
## candidates and were removed from the data stack.
```

```
gss_stack <- gss_stack %>%
  blend_predictions()

gss_stack <- gss_stack %>%
  fit_members()

gss_preds <-
  gss_test %>%
  dplyr::select(partyid) %>%
  bind_cols(
    predict(
      gss_stack,
      gss_test,
      members = TRUE
    )
  )

colnames(gss_preds) %>%
  map_dfr(
    .f = accuracy,
    truth = partyid,
    data = gss_preds
  ) %>%
  mutate(member = colnames(gss_preds))
```

```
## # A tibble: 13 x 4
##    .metric  .estimator .estimate member
##    <chr>    <chr>          <dbl> <chr>
##  1 accuracy binary         1     partyid
##  2 accuracy binary         0.645 .pred_class
##  3 accuracy binary         0.639 .pred_class_logreg_resamples_1_09
##  4 accuracy binary         0.639 .pred_class_logreg_resamples_1_10
##  5 accuracy binary         0.642 .pred_class_logreg_resamples_1_21
##  6 accuracy binary         0.639 .pred_class_logreg_resamples_1_24
##  7 accuracy binary         0.636 .pred_class_logreg_resamples_1_25
##  8 accuracy binary         0.636 .pred_class_logreg_resamples_1_27
##  9 accuracy binary         0.570 .pred_class_logreg_resamples_1_29
## 10 accuracy binary         0.637 .pred_class_lda_resamples_1_04
## 11 accuracy binary         0.637 .pred_class_lda_resamples_1_07
## 12 accuracy binary         0.630 .pred_class_svm_res_1_3
## 13 accuracy binary         0.598 .pred_class_svm_res_1_4
```

For a base-line, let's fit a simple logistic regression:

```
glm_logreg <- glm(partyid ~ ., data = gss_train, family = "binomial")
probs_logreg <- predict(glm_logreg, gss_test, type = "response")
preds_logreg <- ifelse(probs_logreg >=.5, 1, 0)
confusion_logreg <- table(preds_logreg, gss_test$partyid)
confusion_logreg

accuracy_logreg <- 1 - (confusion_logreg[1,2] + confusion_logreg[2,1]) / nrow(gss_test)
accuracy_logreg
```