# tidymodels_notes

## Software for Modeling

Models are very multifaceted & serve various purposes in statistical methods, but they are ultimately meant to disentangle complicated statistical phenomena.

Software used to create models should:

- have a easy-to-use interface
- protect users from making mistakes

## Types of Models

### Descriptive Models

Descriptive models are used generally to describe data attributes/traits, more of an opportunity to get to know your data better rather than make any kind of inference about a relationship related to the data

### Inferential Models

Models of this nature are use to test a particular claim of a relationship between variables, a particular hypothesis, etc. After this model is created, a state of statistical significance/insignificance is produced

Output from these models usually includes a p-value, confidence intervals, etc. to estimate probabilites (i.e., of committing Type 1 error, etc.)

### Predictive Models

*These models deal with questions of estimation as opposed to inference*

This type of model is used to provide accurate predictions utilizing previous/historical data

### Building Predictive Models

- Mechanistic models: using predetermined equations/assumption to derive a model equation, can make statements about how the model will perform based on how it performs with existing data
- Empirically driven models: machine learning models, example: K-nearest neighbor (given a dataset, a new sample is speculated using the K most similar data in the set.)
  - note: "the primary method of evaluating the appropriateness of the model is to assess its accuracy using existing data"

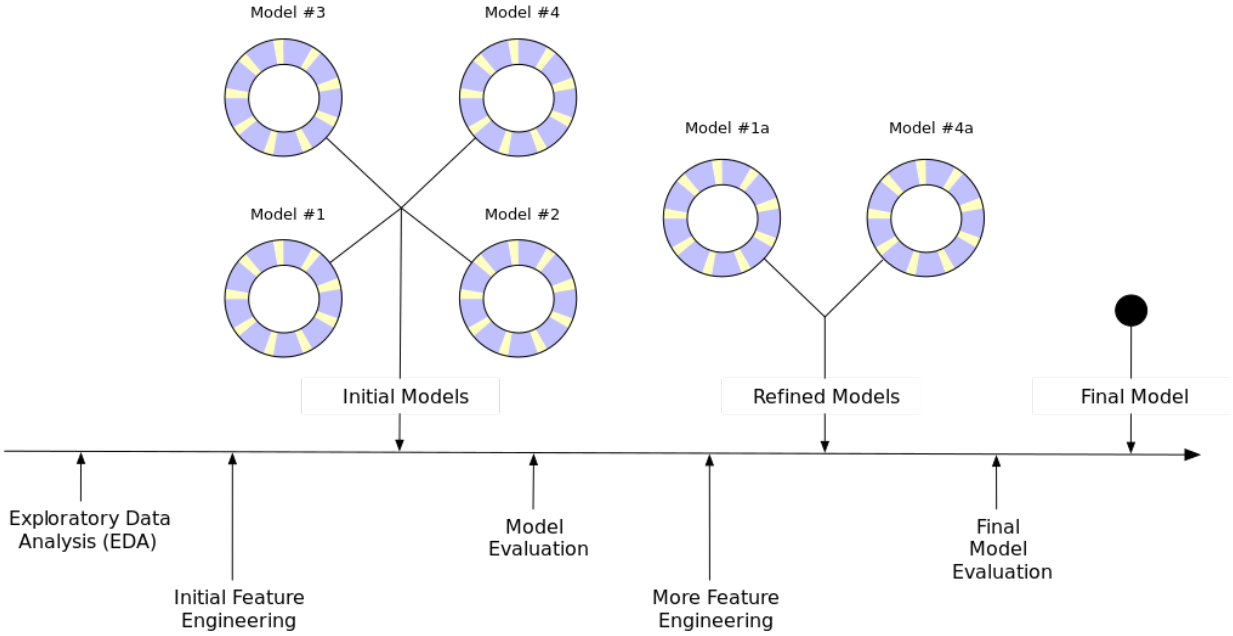
You may generate models that are statistically significant, but that doesn't necessarily mean its the best model for your data/represents your data appropriately

## Terminology

- Unsupervised models: those models that learn patterns, clusters or other characteristics of the data but lack an outcome
- Supervised models: those models that have an outcome variable
  - Regression: predicts numeric outcome
  - Classification: produces ordered/unordered qualitative values

### Phases of Data Analysis

- Cleaning data
  - involves looking closely at your data to see how it is presented/aligns with the research questions you are trying to answer
- Exploratory data analysis (EDA)
  - understanding how the data came about (sampling information, etc.); should also consider how appropriate/relevant the data are to your research
  - one should outline a clear goal for the model and how success with the model will be judged




Process for choosing appropriate model

- Exploratory data analysis (EDA)
  - oscillating between numerical analysis & data visualization
- Feature engineering
  - refiguring predictors to make them easier to model
- Model tuning and selection
  - creating initial models to assess & compare; usually involves parameter tuning, etc.
- Model evaluation
  - using formal evaluation techniques to compare model results, differences, etc.

## Tidyverse notes

Gives broad explanation as to the differences between the tidyverse & base R, essentially providing evidence as to how tidyverse is more intutive & user-friendly for folks that are not super familiar with coding/programming. Goes over syntax, the importance of tibbles & pipes in the tidyverse and provides examples of how the tidyverse is super helpful in the modeling process

## Review of R Modeling Fundamentals

Summarizes the workings of the R formula and typical functions that are utilized in model creation. As general principle, tidymodels relies heavily on the strength of R's existing functions (object-oriented programming) and emphasizes understandable defaults to function arguments, consistency across function names & arguments and user-friendliness (not restricting functions to specific data structures)

Briefly talks about combining base R w/ tidyverse that might be helpful in creating your models.

Tidymodels is essentially considered a "metapackage" with a core set of tidymodels & packages. It splits packages by their function (ex: data splitting/resampling w/ rsample, measuring performance w/ yardstick)

## Basics (using Ames housing data)

models sale prices of homes in Ames, IA using a histograms (with transformed/not transformed data; i.e. log data)

conducting EDA (exploratory data analysis) to determine distributions of predictors, correlations between predictors or possible associations between predictors & outcomes

## Spending our data

Steps to creating useful model:

- parameter estimation
- model selection
- tuning
- performance assessment

data spending: allocating specific subsets of data to do different things

common ways to split/"spend" data:

- split the larger dataset into two, using one as a training set to develop & optimize the model
  - the second set is the test set, used after 1-2 models are selected and the test set is used to produce these models & determine their efficacy
- r function initial_split() will perform this action for you, takes the dataset & desired ratio as arguments
  - training() and testing() will produce the necessary datasets with the object you just created with initial_split()

can either use simple random sampling/stratified sampling to contribute to the model (depends on the level of imbalance in the data) - the strata argument can be added to the initial_split() function to conduct stratified random sampling

**for time series data: initial_time_split() will save most recent data for the test set

validation sets can also be part of this data splitting process, getting a sense of model performance before using the test set

with multi-level data, have to keep in mind that data splitting should be happening according to the independent unit level of the data

## Fitting models with parsnip

Parsnip is a tidymodels package that can be used to produce a variety of different models - the functions fit() and predict() are the most commonly used with a parsnip object

example: modeling linear regression

- base R: lm() for ordinary linear regression and stan_glm() for regularized linear regression (Bayesian); for non-Bayesian approach & regularized regression –> glmnet()
- tidymodels unifies this process, eliminates need to memorized different syntax for functions
  1) identify model type (linear regression, random forest, etc.)
  2) identify engine for fitting the model (i.e., software package)
  3) identify the mode of the model (regression/classification)

```
linear_reg() %>% set_engine("lm")
# linear regression model specification, computational engine: lm

linear_reg() %>% set_engine("glmnet")
# linear regression model specification, computational engine: glmnet
```

you can use fit() or fit_xy() to begin model estimation; fit() is more appropriately used with a formula, fit_xy() is more appropriate if data is pre-processed

translate() will explain the process of the user's code being converted to package syntax

```
linear_reg() %>% set_engine("lm") %>% translate()
# Linear Regression Model Specification (regression)
# Computational engine: lm
# Model fit template:
# stats::lm(formula = missing_arg(), data = missing_arg(), weights = missing_arg())

lm_model <-
  linear_reg() %>%
  set_engine("lm")

lm_form_fit <-
  lm_model %>%

fit(Sale_Price ~ Longitude + Latitude, data = ames_train)

lm_xy_fit <-
  lm_model %>%
  fit_xy(
      x = ames_train %>% select(Longitude, Latitude),
      y = ames_train %>% pull(Sale_Price)
    )

lm_form_fit

# Call:
# stats::lm(formula = Sale_Price ~ Longitude + Latitude, data = data)

# Coefficients:
# (Intercept)    Longitude     Latitude
#     -302.97        -2.07         2.71
#

lm_xy_fit

# parsnip model object
#
#
# Call:
# stats::lm(formula = ..y ~ ., data = data)
#
# Coefficients:
# (Intercept)    Longitude     Latitude
#     -302.97        -2.07         2.71
```

parsnip also provides consistency in model arguments ex: random forest function w/ parsnip (arguments)

- number of sampled predictors (mtry)
- number of trees (trees)
- number of data points to split (min_n)

model arguments in parsnip are separated into two different categories: - main arguments: commonly used & available across engines - engine arguments: specific to a particular engine, can be specified in set_engine()

### Using Model Results

Examining model output using extract_fit_engine(), adding vcov() will return model output as matrix

However, matrices are not super malleable/easy to change & manipulate

- the broom package in tidymodels convert all different kinds of objects to a tidy structure
  - column names are standardized across models
  - row names converted to columns

### Make predictions

predict() function

- returns results in tibble format
- column names are always predictable
- as many rows in the tibble as there are in the input data

these features make the merging of predictions & original data much easier

Table 6.4: The tidymodels mapping of prediction types and column names.

| type value | column name(s) |
|---|---|
| numeric | .pred |
| class | .pred_class |
| prob | .pred_{class levels} |
| conf_int | .pred_lower, .pred_upper |
| pred_int | .pred_lower, .pred_upper |

other packages can be used with parsnip, list found here

parsnip_addin() will provide a list of possible models for each model mode

## A model workflow

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_p x_{ip}$$

Figure 1: Model equation

Above is the typical structural equation for a linear model, including predictors (x) and outcome variable (y).

**a feature selection algorithm is mentioned to determine the minimum predictors needed for the model, wish i knew what that was lmao**

model workflow: pre-processing steps, model fit and potential post-processing activities

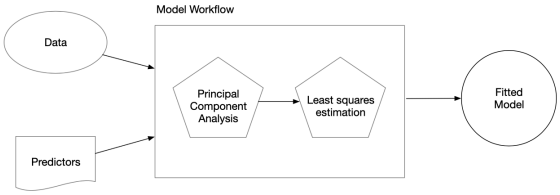



Figure 2: Correct mental model of where model estimation occurs in the data analysis process.

workflows package in R:

```
lm_wflow <-
  workflow() %>%
  add_model(lm_model)

lm_wflow
#>   Workflow
#> Preprocessor: None
#> Model: linear_reg()
#>
#>   Model
#> Linear Regression Model Specification (regression)
#>
#> Computational engine: lm
```

*with a simple model, a formula would be added as a preprocessor using add_formula()

- fit() can be used to create the model; takes created workflow & data as arguments, use predict() after on the fitted workflow
- use update_formula() to update/remove model and/or preprocessor

you can add raw variables to your workflow by using add_variables(); takes outcomes & predictors as arguments - any outcome variables accidentally used in the predictors will be removed; fit() then takes the ready workflow & training set data

## Formulas & Workflow

- different kinds of encoding procedures for formulas are necessary for different statistical methods
- example: tree-based models
  - when fitting a random forest model using ranger/randomForest, the workflow will know to leave the predictors columns that are factors alone
- determinations regarding the need to create dummy variables, etc. is made for each model & engine combination
  - ex: add_model() can take a formula argument to specify model formula

## Multiple Workflows

can be helpful when trying to build predictive models (evaluate different model types)

good for sequential testing (beginning with full model and gradually taking out predictors)

workflowset creates combos of components of workflow (list of preprocessors & model specifications that be matched)

```
  location <- list(
    longitude = Sale_Price ~ Longitude,
    latitude = Sale_Price ~ Latitude,
    coords = Sale_Price ~ Longitude + Latitude,
    neighborhood = Sale_Price ~ Neighborhood
    )

  library(workflowsets)
  location_models <- workflow_set(preproc = location, models = list(lm = lm_model))

  location_models
#> # A workflow set/tibble: 4 × 4
#>   wflow_id        info             option    result
#>   <chr>           <list>           <list>    <list>
#> 1 longitude_lm    <tibble [1 × 4]> <opts[0]> <list [0]>
#> 2 latitude_lm     <tibble [1 × 4]> <opts[0]> <list [0]>
#> 3 coords_lm       <tibble [1 × 4]> <opts[0]> <list [0]>
#> 4 neighborhood_lm <tibble [1 × 4]> <opts[0]> <list [0]>

location_models$info[[1]]
#> # A tibble: 1 × 4
#>   workflow   preproc model      comment
#>   <list>     <chr>   <chr>      <chr>
#> 1 <workflow> formula linear_reg ""
extract_workflow(location_models, id = "coords_lm")

#>   Workflow
#> Preprocessor: Formula
#> Model: linear_reg()
#>
#>   Preprocessor
#> Sale_Price ~ Longitude + Latitude
#>
#>   Model
#> Linear Regression Model Specification (regression)
#>
#> Computational engine: lm
```

# Feature engineering with recipes

- feature engineering involves reformatting predictors to make them easier to use in a model
  - can be done with recipes package; combines preprocessing/feature engineering tasks into one object to then apply to different data sets
- a recipe essentially works like a formula without actually executing anything
  - it lists out what should be done

## Formula

```
lm(Sale_Price ~ Neighborhood + log10(Gr_Liv_Area) + Year_Built + Bldg_Type, data = ames)
```

## Recipe

```
library(tidymodels) # Includes the recipes package
tidymodels_prefer()

simple_ames <-
  recipe(Sale_Price ~ Neighborhood + Gr_Liv_Area + Year_Built + Bldg_Type,
         data = ames_train) %>%
  step_log(Gr_Liv_Area, base = 10) %>%
  step_dummy(all_nominal_predictors())

simple_ames

#> Recipe
#>
#> Inputs:
#>
#>       role #variables
#>    outcome          1
#>  predictor          4
#>
#> Operations:
#>
#> Log transformation on Gr_Liv_Area
#> Dummy variables from all_nominal_predictors()
```

step_dummy() converts qualitative data to dummy/indicator variables; all_nominal_predictors() is a function that captures predictor columns that are nominal in nature

advantages of recipes:

- computations can be reused across models
- recipes offer broader set of choices for data processing not available with formulas
- compact syntax
- data processing can be captured in single R object

if adding recipes to existing workflow: - remember to remove variables if previous preprocessing method was used: remove_variables(), then use add_recipe()

use new workflow & training data to call fit() and use the predict() function to test data; extract_recipe() can be used to retrieve recipe details after it's been estimated

```
  lm_fit %>%
    extract_recipe(estimated = TRUE)
#> Recipe
#>
#> Inputs:
#>
#>       role #variables
#>    outcome          1
#>  predictor          4
#>
#> Training data contained 2342 data points and no missing data.
#>
#> Operations:
#>
#> Log transformation on Gr_Liv_Area [trained]
```

```
#> Dummy variables from Neighborhood, Bldg_Type [trained]

lm_fit %>%
  # This returns the parsnip object:
  extract_fit_parsnip() %>%
  # Now tidy the linear model object:
  tidy() %>%
  slice(1:5)
#> # A tibble: 5 × 5
#>   term                         estimate std.error statistic   p.value
#>   <chr>                           <dbl>     <dbl>     <dbl>     <dbl>
#> 1 (Intercept)                  -0.669    0.231        -2.90 3.80e-  3
#> 2 Gr_Liv_Area                   0.620    0.0143       43.2  2.63e-299
#> 3 Year_Built                    0.00200  0.000117     17.1  6.16e- 62
#> 4 Neighborhood_College_Creek    0.0178   0.00819       2.17 3.02e-  2
#> 5 Neighborhood_Old_Town        -0.0330   0.00838      -3.93 8.66e-  5
```

## How data are used by the recipe

1) when calling recipe(), data serves as an argument of the function to determine the data types for each column and facilitate future use of selectors like all_numeric(), etc.
2) when calling fit(), the training data are used for estimation operations
3) when calling predict(), takes new test data and values are standardized based on training data values

## Recipes package capabilities

-encoding qualitative data (ex: step_other() can be used to group infrequent values into a catch-all level) - step_dummy() to create dummy/indicator variables for factor predictors - also allows for user to customize the naming of the dummy variable created

## Interactions

specified using step_interact(~ interaction term) *can use selectors in interaction term*

```
simple_ames <-
  recipe(Sale_Price ~ Neighborhood + Gr_Liv_Area + Year_Built + Bldg_Type,
         data = ames_train) %>%
  step_log(Gr_Liv_Area, base = 10) %>%
  step_other(Neighborhood, threshold = 0.01) %>%
  step_dummy(all_nominal_predictors()) %>%

  # Gr_Liv_Area is on the log scale from a previous step

  step_interact( ~ Gr_Liv_Area:starts_with("Bldg_Type_") )
```

## Spline functions

spline terms replace existing numeric predictors with a set of columns that allow the model to emulate a non-linear relationship

```
recipe(Sale_Price ~ Neighborhood + Gr_Liv_Area + Year_Built + Bldg_Type + Latitude,
         data = ames_train) %>%
  step_log(Gr_Liv_Area, base = 10) %>%
  step_other(Neighborhood, threshold = 0.01) %>%
  step_dummy(all_nominal_predictors()) %>%
  step_interact( ~ Gr_Liv_Area:starts_with("Bldg_Type_") ) %>%
```

```
  step_ns(Latitude, deg_free = 20)
```

## Feature extraction

- if creating new features from existing predictors (a la PCA), step_pca() can be used in recipe
  - step_normalize() might be needed before if the columns used are not all measured the same

## Row sampling steps

- downsampling: keeps minority class & takes random sample of majority so class frequencies are balanced
- upsampling: replicates samples from minority class to balance the classes
- hybrid: combination of both

*themis package has recipe steps for this, ex. from it: step_downsample(outcome_column_name)*

## More

- step_mutate() can be used for general transformations
- textrecipes package can be used to apply natural language processing methods to data (ex: split string of text into separate words, etc.)
- subsampling processes, etc. should not be applied to the data being predicted
  - each step function mentioned has a skip option; set it to TRUE so that it will be ignored by the predict function
  - all steps are then applied when using fit()

## Tidy a recipe

when using tidy() along with your recipe object, a summary of the recipe steps will appear

```
tidy(ames_rec)
#> # A tibble: 5 × 6
#>   number operation type     trained skip  id
#>    <int> <chr>     <chr>    <lgl>   <lgl> <chr>
#> 1      1 step      log      FALSE   FALSE log_66JTU
#> 2      2 step      other    FALSE   FALSE other_ePfcw
#> 3      3 step      dummy    FALSE   FALSE dummy_Z18Cl
#> 4      4 step      interact FALSE   FALSE interact_JLU36
#> 5      5 step      ns       FALSE   FALSE ns_rvsqQ
```

can specific id column if wanting to tidy() it

```
ames_rec <-
  recipe(Sale_Price ~ Neighborhood + Gr_Liv_Area + Year_Built + Bldg_Type +
         Latitude + Longitude, data = ames_train) %>%
  step_log(Gr_Liv_Area, base = 10) %>%
  step_other(Neighborhood, threshold = 0.01, id = "my_id") %>%
  step_dummy(all_nominal_predictors()) %>%
  step_interact( ~ Gr_Liv_Area:starts_with("Bldg_Type_") ) %>%
  step_ns(Latitude, Longitude, deg_free = 20)
```

refit workflow with new recipe, call tidy() with new object

```
estimated_recipe <-
  lm_fit %>%
  extract_recipe(estimated = TRUE)
```

```
tidy(estimated_recipe, id = "my_id")
```

```
#> # A tibble: 22 × 3
#>   terms        retained           id
#>   <chr>        <chr>              <chr>
#> 1 Neighborhood North_Ames         my_id
#> 2 Neighborhood College_Creek      my_id
#> 3 Neighborhood Old_Town           my_id
#> 4 Neighborhood Edwards            my_id
#> 5 Neighborhood Somerset           my_id
#> 6 Neighborhood Northridge_Heights my_id
#> # … with 16 more rows
```

step numbers can also be used in the tidy() call

### Column Roles

- columns used in the formula in recipe() are usually assigned a "predictor" or "outcome" depending on which side of the tilde they're on
  - however, if there's a column you'd like to keep that does not fall into one of those categories, you can use add_roles(), update_roles(), remove_roles()
  - can be helpful when data are resampled
    ames_rec %>% update_role(address, new_role = "street address")

all possible recipe steps found here

## Judging model effectiveness