tidymodels_notes

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

> The goal of an inferential model is to produce a decision for a research question or to test a specific hypothesis, in much the way that statistical tests are used. The goal is to make some statement of truth regarding a predefined conjecture or idea. In many (but not all) cases, a qualitative statement is produced (e.g., a difference was "statistically significant").
>
> Inferential modeling techniques typically produce some type of probabilistic output, such as a p-value, confidence interval, or posterior probability. Generally, to compute such a quantity, formal probabilistic assumptions must be made about the data and the underlying processes that generated the data

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"

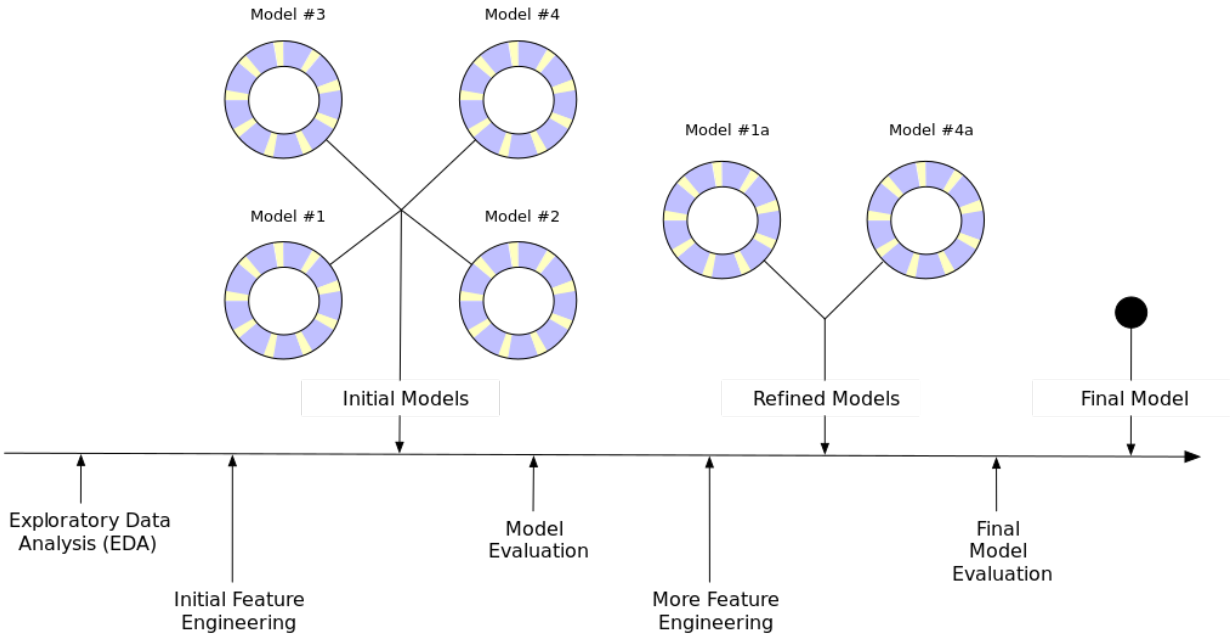
You may generate models that are statistically significant, but that doesn't necessarily mean its the best model for your data/represents your data appropriately

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