tidymodels_notes

TidyModels Notes

What Ben is looking for

- Overall read through to get the gist
- Keep track of questions along the way (for me to answer or for us to answer as we both learn more)
- Be thinking about how we'll use TM in the cc_earn project (notes here too)

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

> The purpose of a descriptive model is to describe or illustrate characteristics of some data. The analysis might have no other purpose than to visually emphasize some trend or artifact in 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 used1. 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

> Sometimes data are modeled to produce the most accurate prediction possible for new data. Here, the primary goal is that the predicted values have the highest possible fidelity to the true value of the new 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"

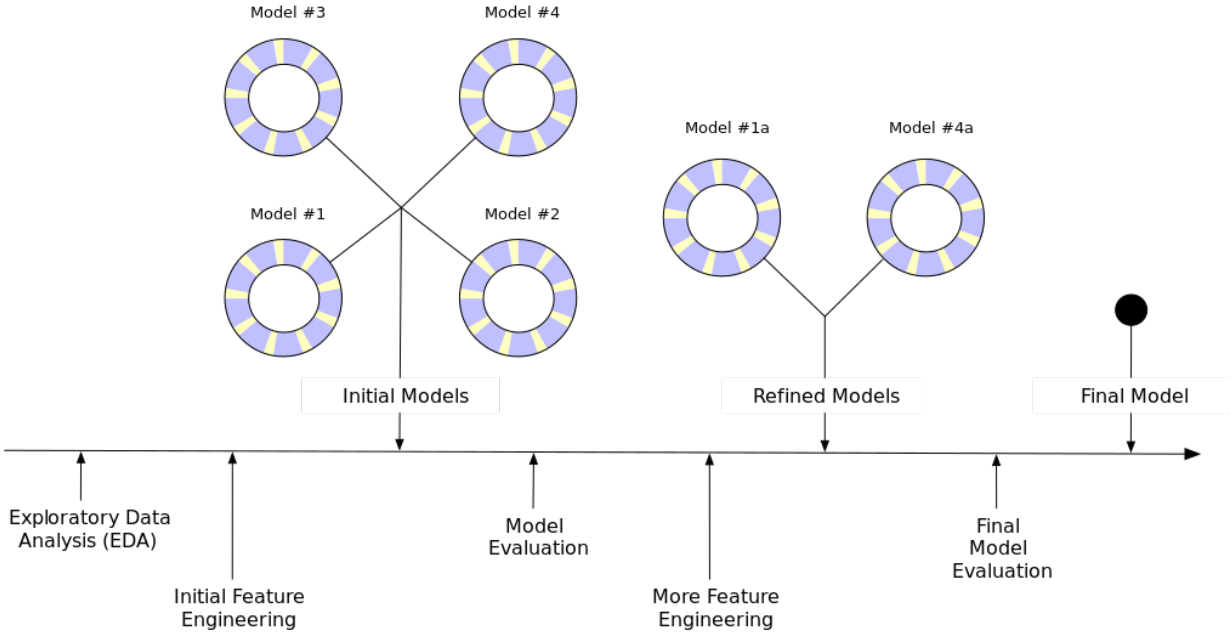
If a model has limited fidelity to the data, the inferences generated by the model should be highly suspect. In other words, statistical significance may not be sufficient proof that a model is appropriate.

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.