Going from Modeling to Meta-Modeling

(Going from being a student of data science to a professional)

Christian Wypasek

2019-10-25

Going from Modeling to Meta-Modeling Discussion Direction

- Punch line: For individuals in the data sciences, an effective personal meta-modeling paradigm is an indicator of professional maturity.
- Model vs Meta-Model
- Speaker background and a few precepts
- Mathematics in asset pricing
  - A basic example: Building a credit score
  - More complex consumer asset-based cash flows
  - High level modeling process vs reality
- A Meta-Model as not just a coping mechanism
- Meta-Modeling in model specification

Model vs. Meta-Model

- A model is an abstraction of a phenomenon or process
- Meta-[-] represents a higher level of abstraction
  - Meta-Data is data about data
  - Meta-Models are models for building models
- Here, we will use *Meta-Model* to describe abstractions that can be either structured or personal. For example:
  - A model specification style is structured
  - One's individual Meta-Model is an abstraction of how one builds models
- For mentors, how should we help students recognize the value of good Meta-Modeling and integrating it into their craft?

Background: Christian Wypasek, PhD

- 2018 – Wypasek Data Science, Inc.,
  - WypasekDataScience.com or wdatasci.com
  - https://github.com/wdatasci
- 9+ Years – Centerbridge Partners, LP
- 3 Years – Merrill Lynch
- 6+ Years – Wachovia/First Union
- 2 Years – GE Capital Mortgage
- NCST – Post Doc at CRSC
- Clemson – PhD/Masters in Mathematical Sciences
- Baldwin Wallace – BS in Mathematics, Cleveland State, CCC
- Family Business – Commercial Art Studio

Some core precepts driving the philosophy of this discussion

- In business presentations, start with the punch line
- FW: “Use the right tool for the job, either buy it or make it”
- FW: “Now that you have shown you can do it, work on your craftsmanship”
- FW: “You would make a good inventor because you’re lazy”
- The “Wonder Bread” approach
- The “coffee cup and doughnut” topological approach
- Does it pass the “Sesame Street Test”?
- “Know your trade”, “Know your data”, “Know your craft”
- “No thought is original”, “No Absolutes”, “No Exaggerating”
- The difference between a confidence interval and a CYA interval

Basic Example: A Credit Score

The objective of a credit score to rank order future account risk based on historical performance:

Database of historical loan files with a reponse (good or bad)

Example
Balance: $100,000
LTV: 80%
FICO: 700
Channel: Retail
Response: good

→

Mathematical representations are used to compare loans.

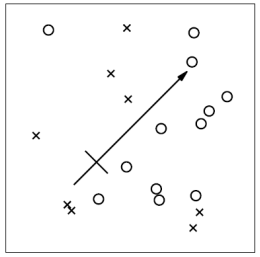
Example

$$\begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ .33 \\ .67 \\ 1 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} \text{Balance} \le \$80{,}000 \\ \text{Balance} > \$80{,}000 \\ \text{LTV} \le 70\% \\ 70\% < \text{LTV} \le 90\% \\ 90\% < \text{LTV} \\ \text{Nghd of FICO to 620} \\ \text{Nghd of FICO to 740} \\ \text{Retail Channel} \\ \text{Other Channel} \\ \text{Response = good} \end{pmatrix}$$

→

Statistical methods find the direction that maximizes the distance between goods and bads

2D Example



→

Calculation of the difference in the optimal direction between a loan and a reference loan is translated into a human-readable scorecard and coded into a delivery system.

Advanced techniques in model selection and structure focus on being predictive and usable.

Data determines the applicable space and the strength of the model.

What about ML or AI for credit scoring?

What about ML?

- Traditional model identification (with a logistic or other regression method) is a machine learning method.
- Would unsupervised methods be useful?
  - Since there is a specific goal, what do you do with unrelated discovered features?

What about AI?

- Neural networks were generally abandoned because of overfitting and difficulty in getting regulator approval

→ ***Any model worth using is worth defending***

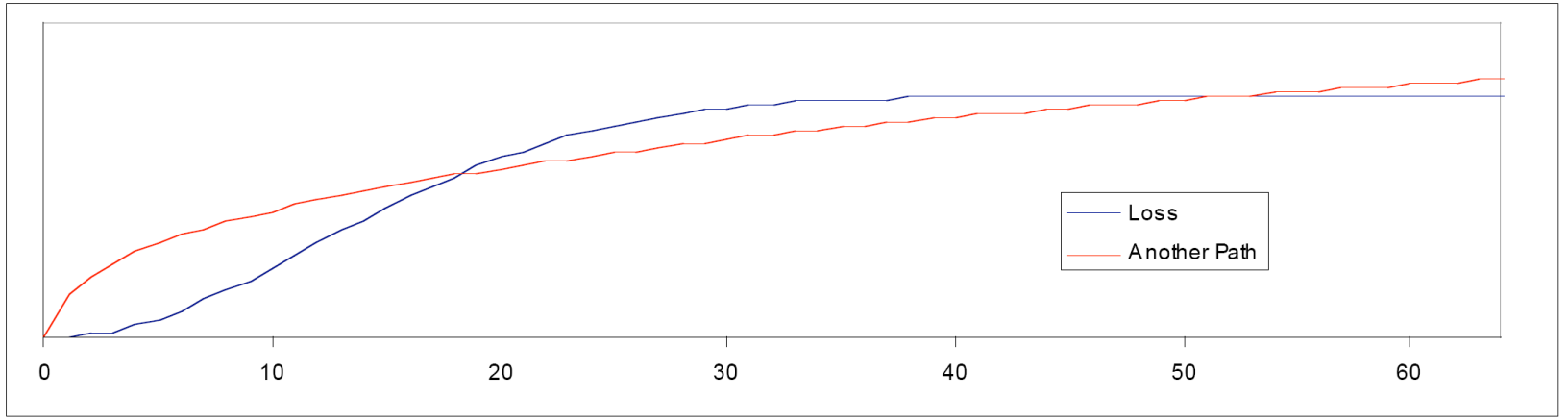
However, on one hand, the smart use of any ML or AI method could be used to inform variable selection and feature investigation in a more conventional regression

But then again, on another hand, point estimates may not be applicable to cash flow forecasts.

From credit scoring to competing risks

Different behaviors may result in similar binomial response rates

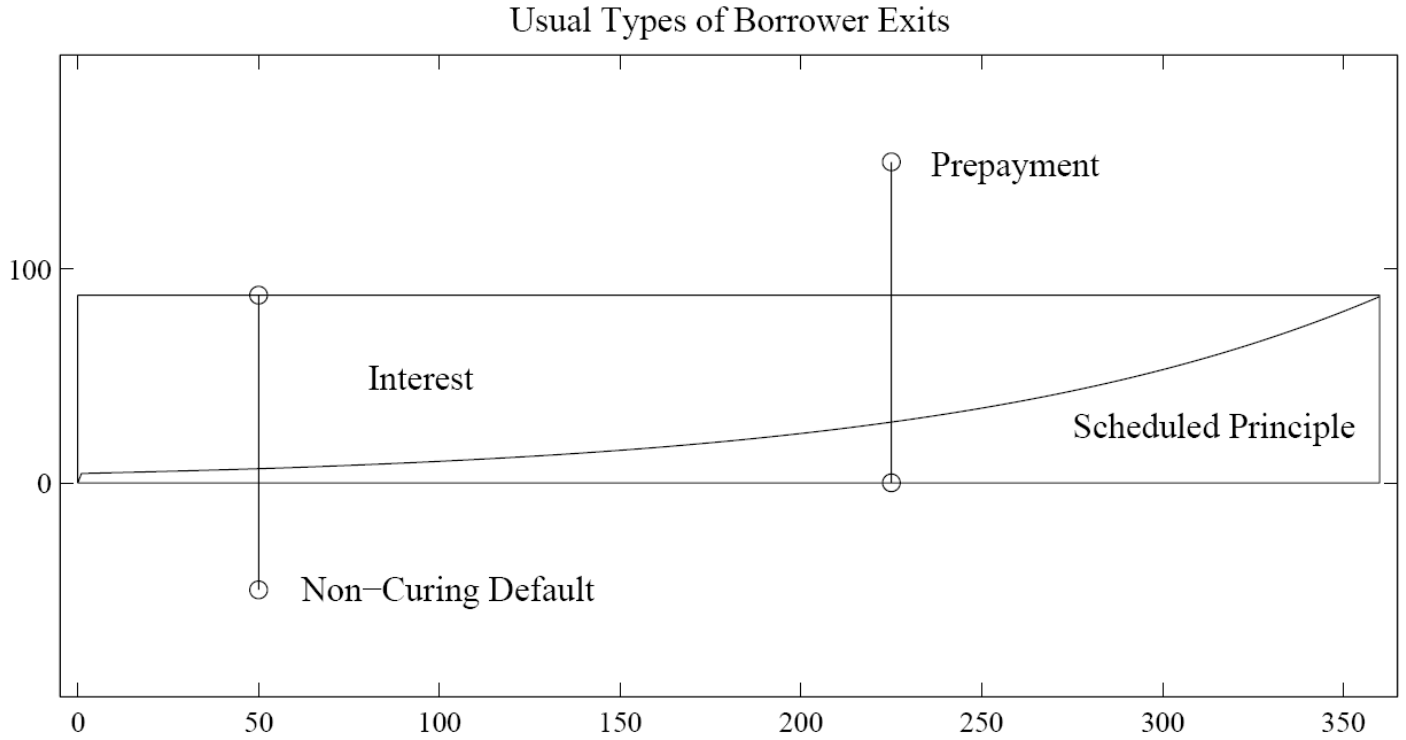
- Consider two populations with the same 18 month cumulative response rate:



- The time value of money also requires a more detailed modeling of the timing of the response. For example, losses 10 years out are less valuable than losses 1 year out.

From credit scoring to competing risks

Loan cash flows can also have multiple risks:

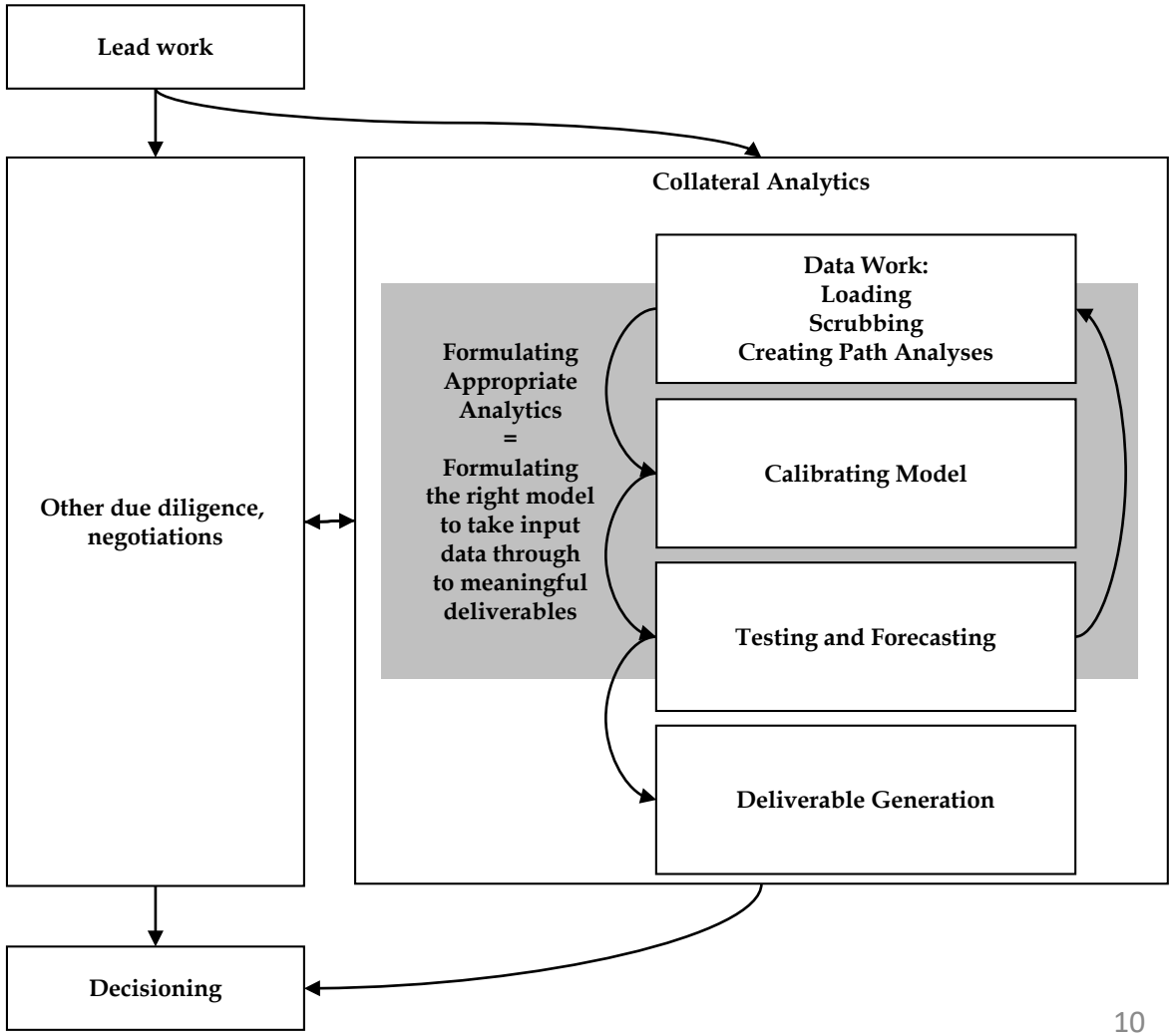


Example: amortizing cash flows in a usual constant payment loan

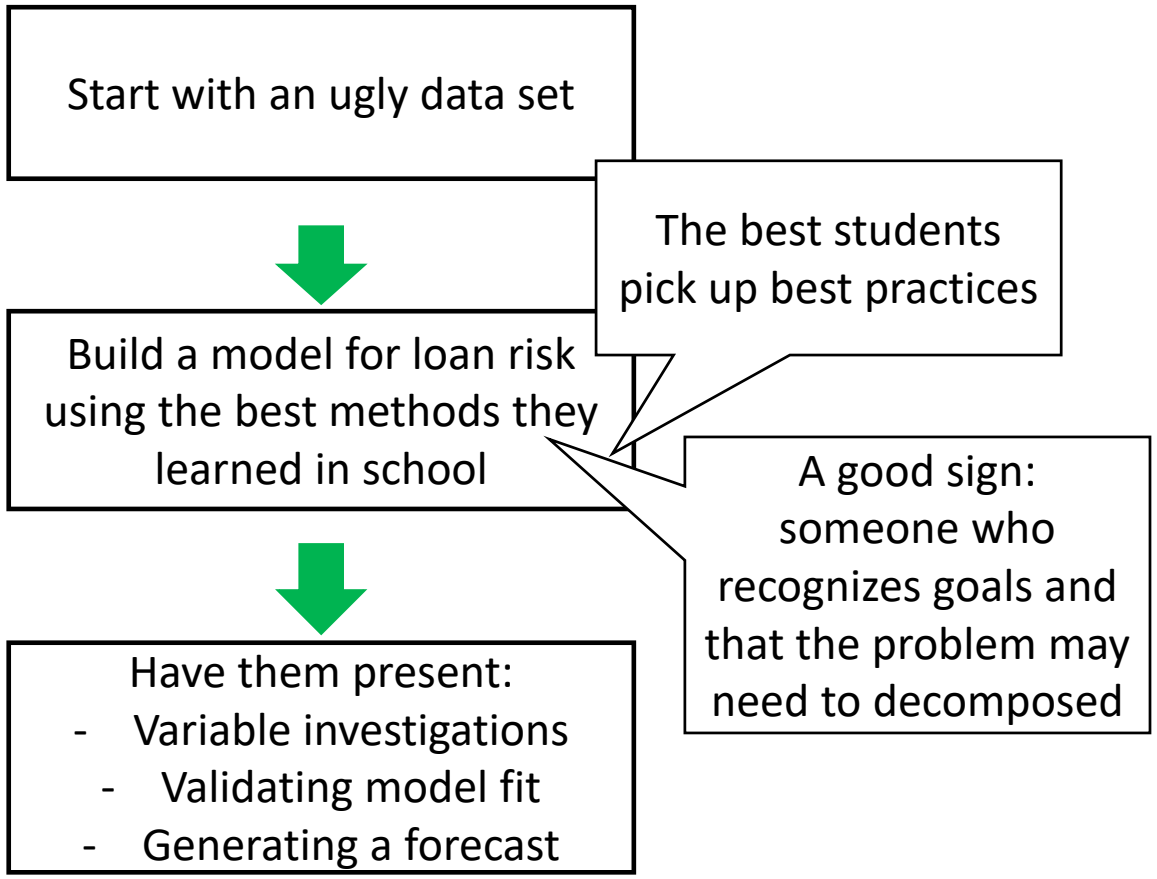
Mathematics in asset pricing
Collateral analytics in a basic deal

Even though this chart is ordered correctly, the high-level process goes backwards:

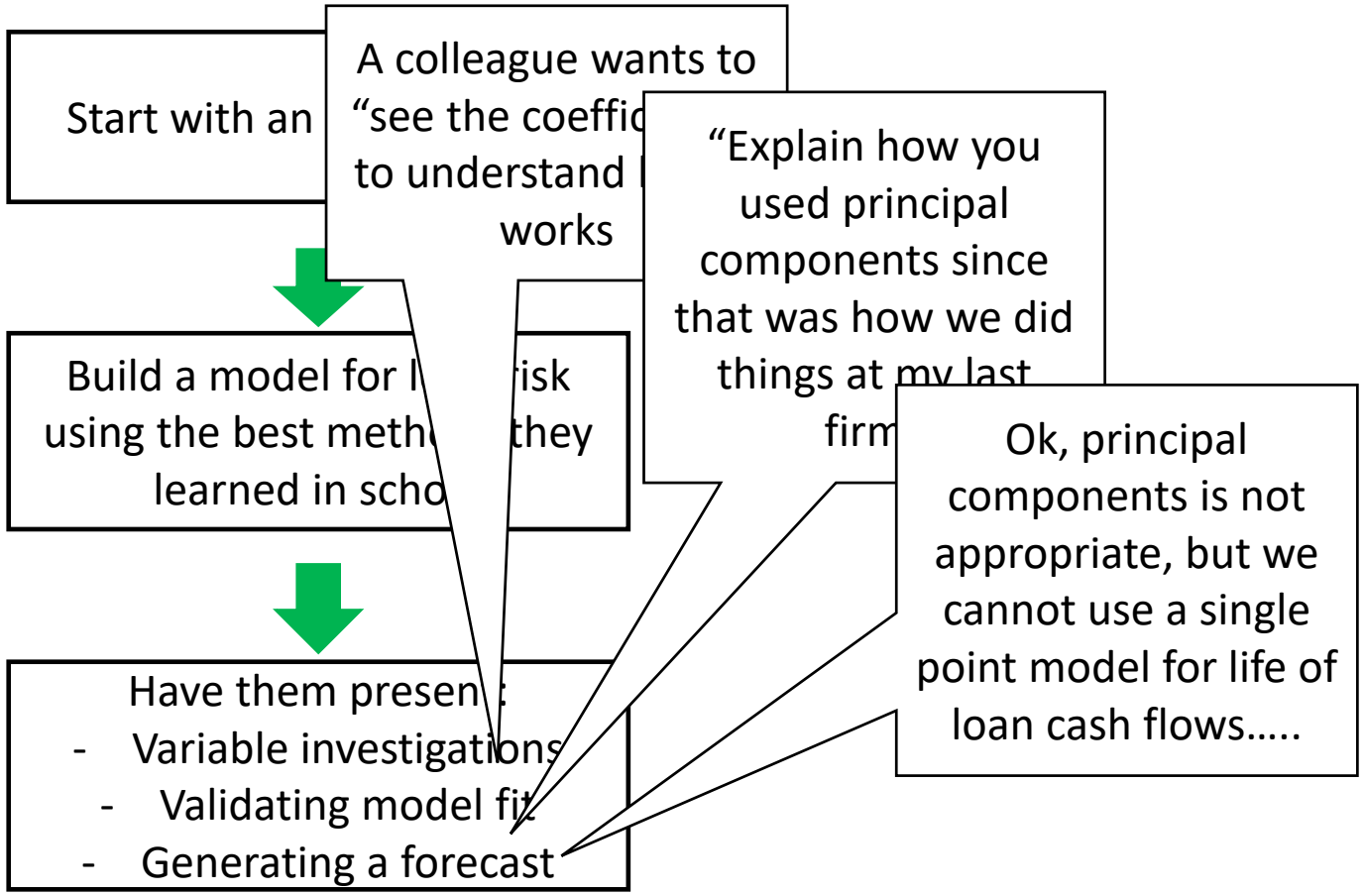
- What do I need?
  - What are the deliverables and the clients' critical-to-quality points?
- What can I use?
  - The right tool for the job
  - Find a model that facilitates the deliverables
- What do I have?
  - Does data provide (or can it be transformed to provide) the required information?



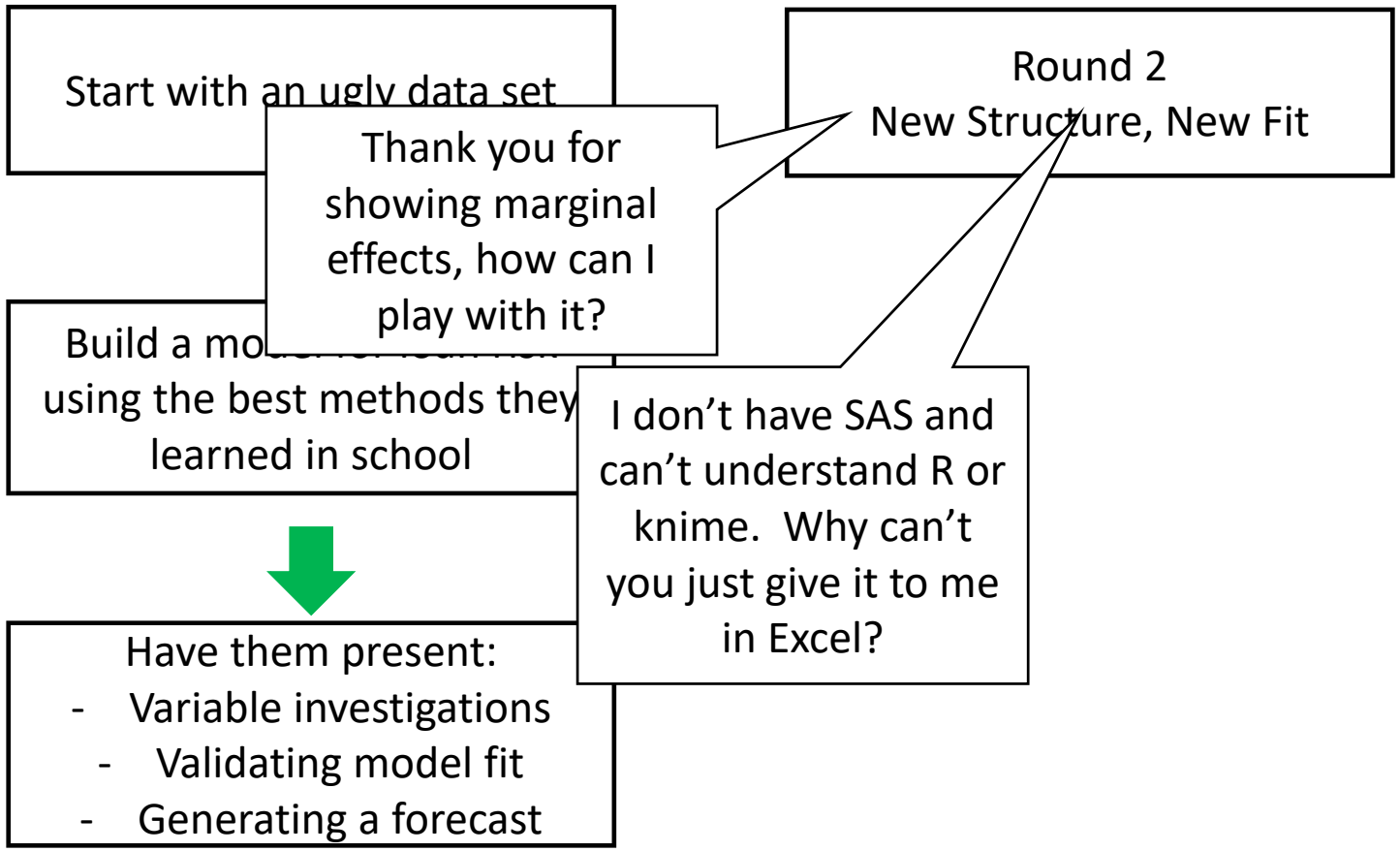
An exercise for new modelers



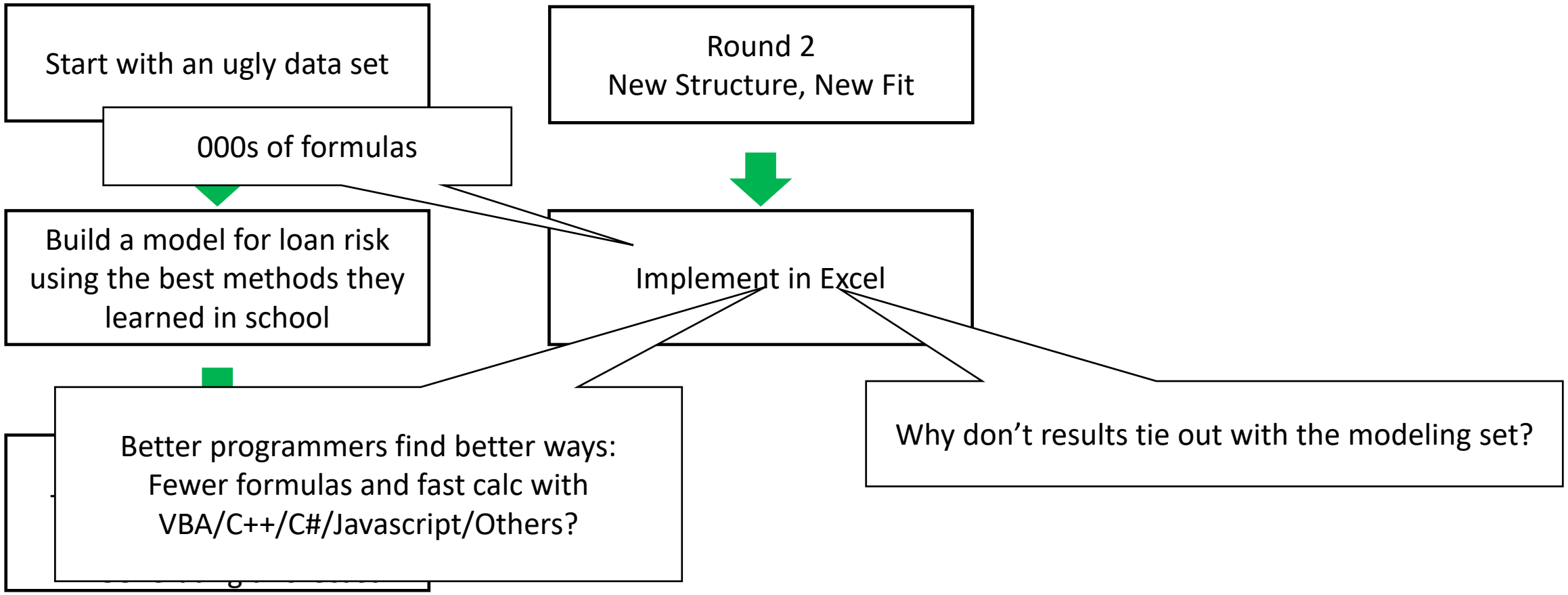
An exercise for new modelers



An exercise for new modelers



An exercise for new modelers



An exercise for new modelers

Start with an ugly data set

Build a model for loan risk using the best methods they learned in school

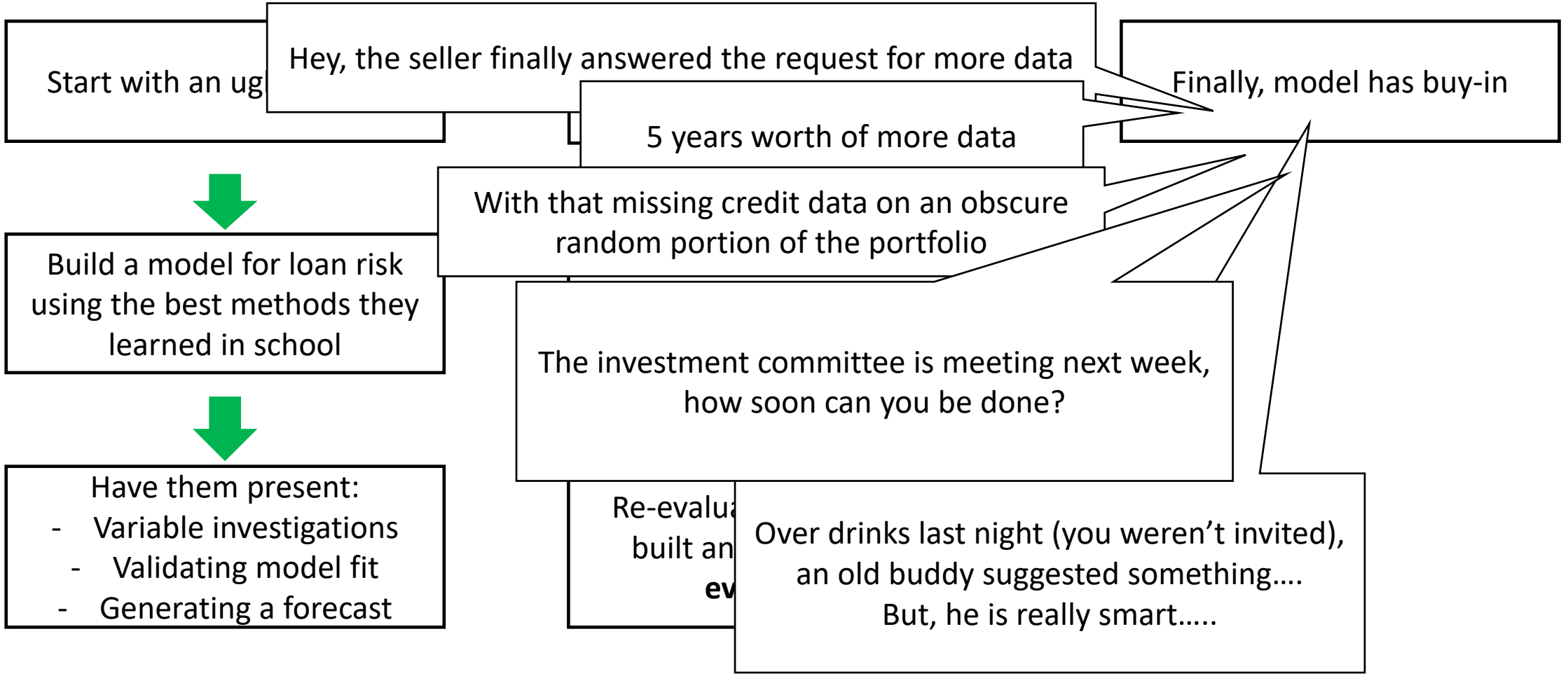
Have them present:
- Variable investigations
- Validating model fit
- Generating a forecast

Round 2
New Structure, New Fit

Implement in Excel

Re-evaluate how model is built and implemented **everywhere**

An exercise for new modelers



An exercise for new modelers

Start with an ugly data set

Build a model for loan risk using the best methods they learned in school

Have them present:
- Variable investigations
- Validating model fit
- Generating a forecast

Round 2
New Structure, New Fit

Implement in Excel

Re-evaluate how model is built and implemented **everywhere**

Finally, model has buy-in

Bid was lost

New deal next week

A Meta-Model is not just a coping mechanism

Effective Meta-Models are not just a good idea, it is a necessity.

- The “Wonder Bread” approach reduces operational risk.
- The “coffee cup and doughnut” topological approach recognizes similarities and encourages reusable code.
- Efficient and repeatable data processing (with efficient Meta-Data) can minimize differences between modeling and forecasting datasets.
- Efficient model specification facilitates both model building and implementation.
- Repeatable processes facilitate documentation and validation.

Core elements of a collateral analytics Meta-Model for a consumer finance project

- Business interaction
  - Determining deliverables, client CTQs, data sources, timeframes, etc.
- Project prototypes
  - Encapsulating past best practices
  - Sets up management of data, scripts, docs, deliverables, etc.
- An excellent and consistent naming convention used globally
  - Data work is easily 80% of the total time. It needs to be efficient.
- A model specification that:
  - Minimizes hand-written risk
  - Facilitates both model fitting and implementation and ensures consistency
- A continually evolving reusable cross-language code library
- Project post-mortem

Naming Conventions

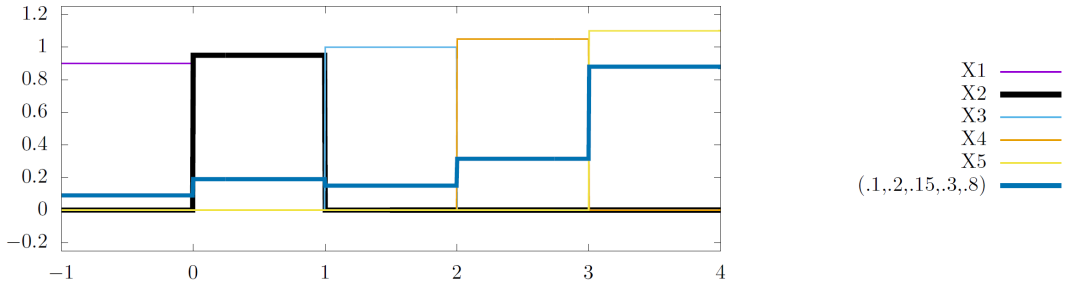
- Ideally, everyone would use a well-informed Meta-Data structure, such as ISO/IED 11179.19. The basic premise is a **three-part naming convention**:
  - **Object** [Qualifier] **Property Representation** [_Modifier]
- Object examples: Loan, Borr[ower], CoBorr[ower], Prin[cipal], Pmt [Payment], Int[erest]Rate, Collat[eral], Prop[erty]
- Property examples: Orig[ination], Cur[rent], Prin[cipal], Int[erest]
  - *Note,* Prin *might be an object or a property, as in the case of* PrinOrigBal *for an original principal balance and* PmtPrinAmt *for the payment amount applied to principal.*
- Representation examples: Name, Number/Nbr, Rate, Pct, Date, DateTime, Status, Code, Ind[icator]
- *Note: Either full words are used everywhere or a fixed set of abbreviations is used everywhere, but there is no mixing of whole words and abbreviations for the same topics. Redundant terms can be dropped.*

Naming Conventions (motivation with bad vs better examples)

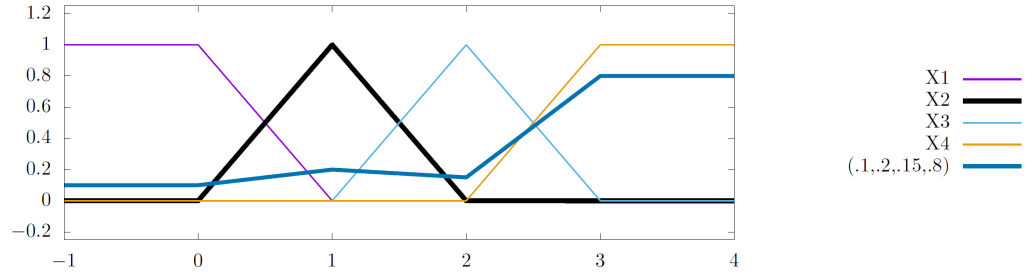
- Bad (but painfully real examples)
  - **For Original Loan Amount**: LOAN, OA, LOAN_AMOUNT, LOAN_AMT, ORIGINAL_BALANCE
  - **For Unpaid Balance:** UPB, BALANCE, CURRENT_BALANCE, CURBAL, TRIAL_BALANCE
  - **For Interest Rate:** IR, INTRATE, INTEREST_RATE, CUR_INT, CURRENT_INT_RATE, INT_RT, COUPON
- Better
  - **For Original Loan Amount**: PrincipalOriginalBalance or PrinOrigBal
  - **For Unpaid Balance:** PrincipalBalance or PrinBal
  - **For Interest Rate:** InterestRate or IntRate

Meta-Modeling in model specification Variable Treatments

- In practice, individual variables in a model are often replaced with a family of artificial variables allowing transformations and local effects.
- Furthermore, there are only a few types generally used, such as:
  - Levels or discretized



  - Hats or BZ1



Meta-Modeling in model specification Variable Treatments

- Conventions:
  - A variable, X, is replaced with artificials X0, X1, X2, …, XN.
  - X0 is always a missing indicator (if applicable).
  - N depends on the treatment and the number of specified critical values.
- Common Treatments:
  - None, X is used directly. N=1, X1=X.
  - CodedMissings, X is replaced by X0 and X1. X1=X if valid otherwise 0.
  - Levels (BZ0), Hats (BZ1), iHats, BZ2.
  - Categorical for character variables.
  - NCategorical for numeric variables.

Meta-Modeling in model specification Variable Treatments

- Simple modeling dataset SAS code for a discretized variable:

```
data modeldata;
        set inputdata;
        X=FICO;
        if X eq . then do;  X0=1; X1=0; X2=0; X3=0; X4=0;
        end; else if X lt 620 then do; X0=0; X1=1; X2=0; X3=0; X4=0;
        end; else if X lt 660 then do; X0=0; X1=0; X2=1; X3=0; X4=0;
        end; else if X lt 720 then do; X0=0; X1=0; X2=0; X3=1; X4=0;
        end; else /*if X ge 720 then*/ do;
                X0=0; X1=0; X2=0; X3=0; X4=1;
        end;
        run;
 proc reg data=modeldata;
        model Y=X0 X1 /* X2 */ X3 X4;
 run;
```

- The encapsulated information for creating a set of artificials to replace X:

  Variable Name: FICO

  Treatment for FICO: Levels

  Critical Values for FICO: 620, 660, 720

Meta-Modeling in model specification Variable Treatments

- SAS example of implementation:

```
MarginalScoreForFICO=0.0;
if ( FICO eq . ) then MarginalScoreForFICO=1.5;
else if ( FICO lt 620 ) then MarginalScoreForFICO=3;
else if ( FICO lt 660 ) then MarginalScoreForFICO=0;
else if ( FICO lt 720 ) then MarginalScoreForFICO=-1;
else MarginalScoreForFICO=-2;
FinalScore=FinalScore+MarginalScoreForFICO;
```

- C++ example of implementation:

```
double MarginalScoreForFICO=0.0;
if ( isnan(FICO) ) MarginalScoreForFICO=1.5;
else if ( FICO < 620 ) MarginalScoreForFICO=3;
else if ( FICO < 660 ) MarginalScoreForFICO=0;
else if ( FICO < 720 ) MarginalScoreForFICO=-1;
else MarginalScoreForFICO=-2;
FinalScore+=MarginalScoreForFICO;
```

- Python example of implementation:

```
MarginalScoreForFICO=0.0
if ( FICO is None or numpy.isnan(FICO) ) :
        MarginalScoreForFICO=1.5
elif ( FICO < 620 ) : MarginalScoreForFICO=3
elif ( FICO < 660 ) : MarginalScoreForFICO=0
elif ( FICO < 720 ) : MarginalScoreForFICO=-1
else : MarginalScoreForFICO=-2
FinalScore+=MarginalScoreForFICO
```

- Despite the implementation language, the information and general structure is the same. The difference is just “mark-up”. The encapsulated information is:
  - Variable Name: FICO
  - Treatment for FICO: Levels
  - Critical Values for FICO: 620, 660, 720
  - Coefficients for FICO: 1.5, 3, 0, 1, -2

Meta-Modeling in model specification Variable Treatments

- A Meta-Model for a variable treatment can be used during fitting and implementation.
  - Specify once during development
  - Fit model
  - Pull coefficients back into model specification format
- In XML, this information can look like:

```
<Variable Name="FICO">
    <Treatment>Levels</Treatment>
    <CriticalValues>620 660 720</CriticalValues>
    <Coefficients>1.5 3 0 1 -2</Coefficients>
</Variable>
```

Wrap-Up

- There still many facets of model specification and developing one's own Meta-Modeling process not discussed here.
- However, emphasizing process and Meta-Modeling from data modeling through model fitting and implementation can reduce risk, improve repeatability, and create multiple implementation avenues.