MIS 431 Data Mining

Model Fitting Process

David Svancer – George Mason University School of Business

GEORGE MASON UNIVERSITY
School of Business

Data Mining Steps

Cross-Industry Standard Process for Data Mining (CRISP-DM)

The CRISP-DM methodology was designed specifically for data mining but is used for most data science/analytical projects. The steps include:

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. Evaluation
6. Deployment

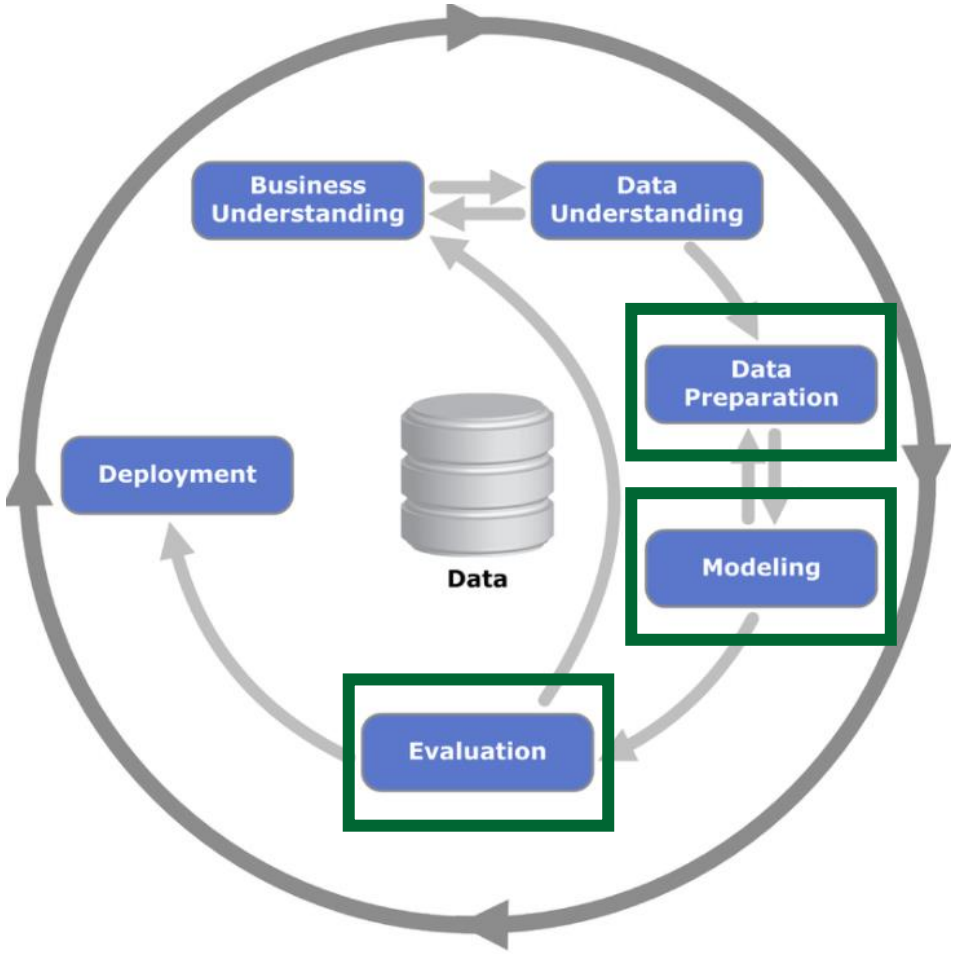
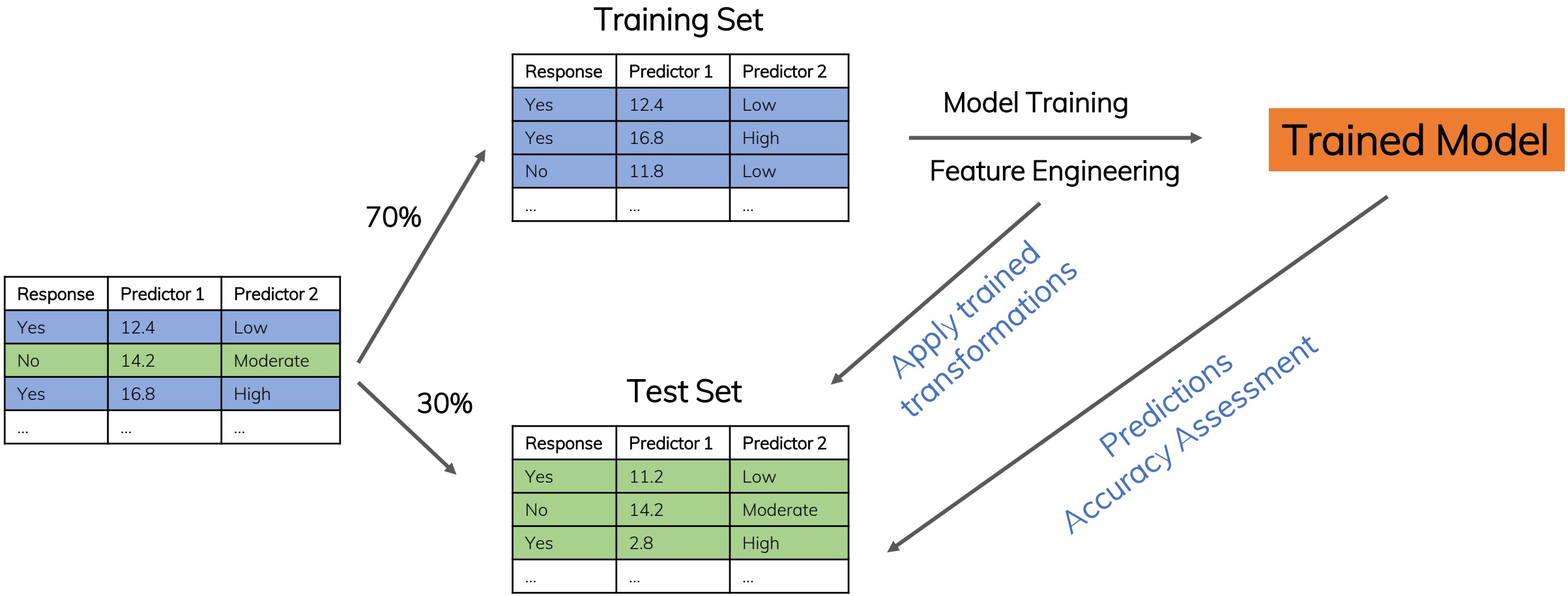


Image: https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining

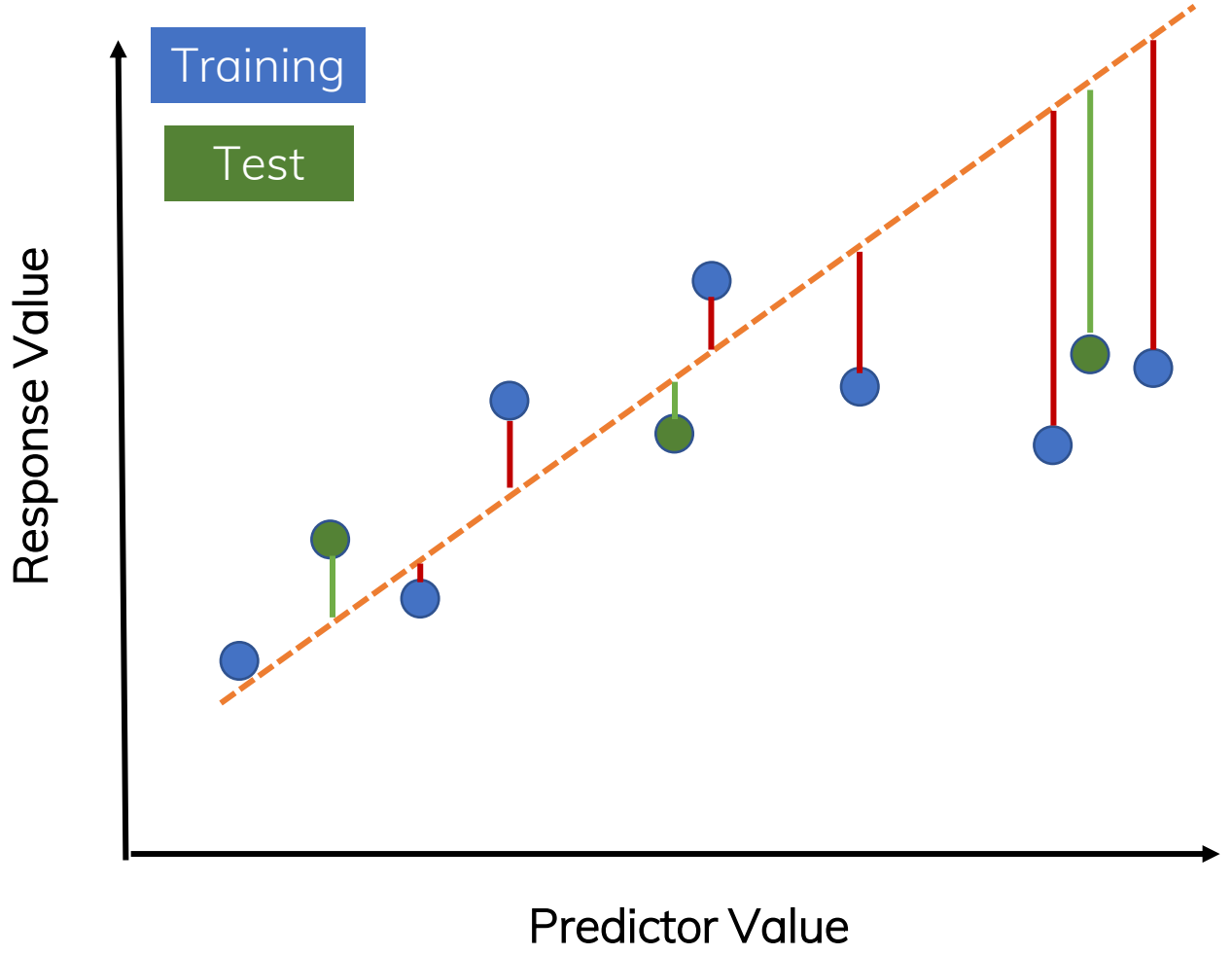
Machine Learning Process



Machine Learning Process

Training and Test Sets

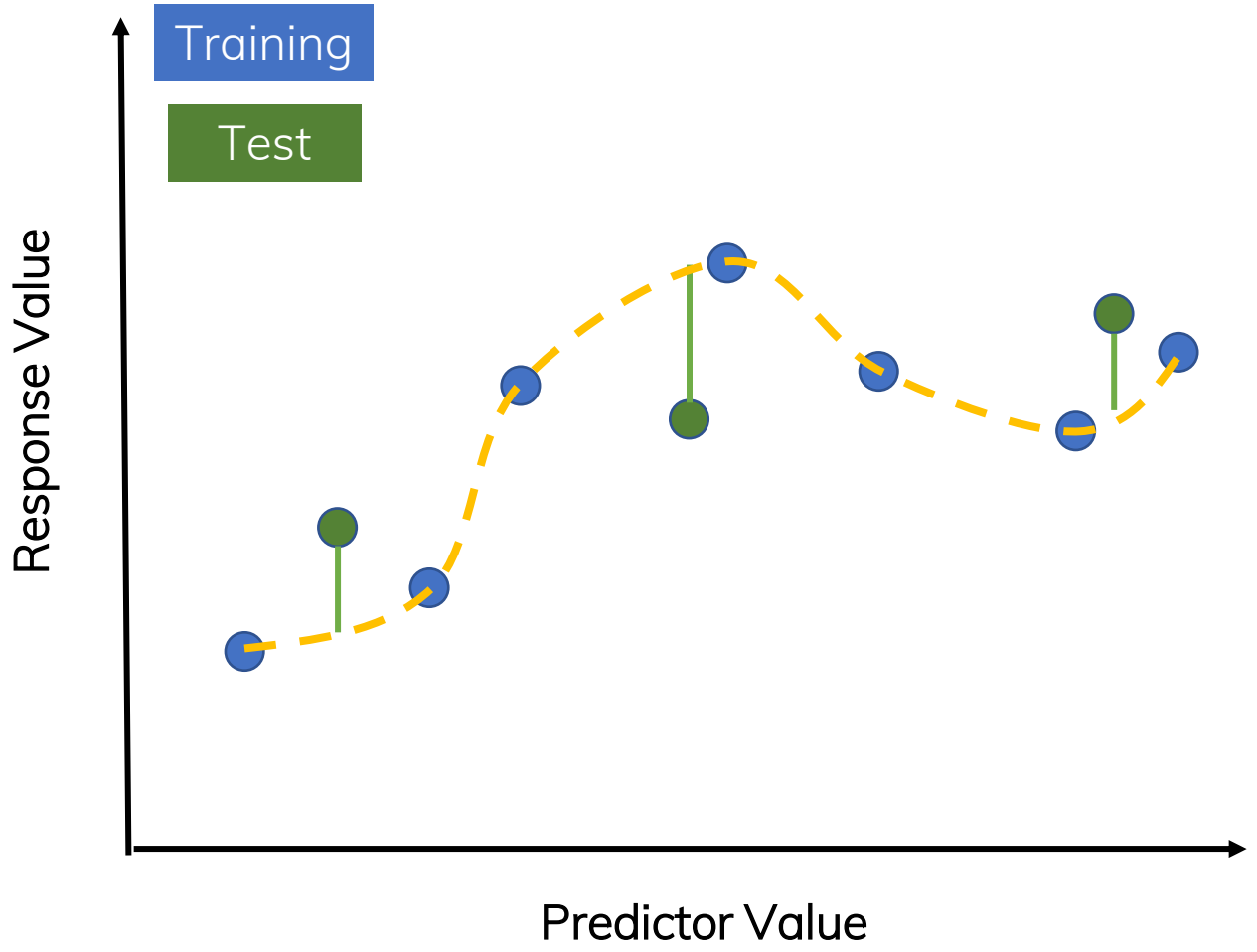
- Why split the data into a training set and test set?
  - Guard against
    - under-fitting
    - Give away – poor accuracy on both training and test sets



Machine Learning Process

Training and Test Sets

- Why split the data into a training set and test set?
  - Guard against
    - under-fitting
    - Give away – poor accuracy on both training and test sets
    - over-fitting
    - Give away – high accuracy on training data, poor accuracy on test data

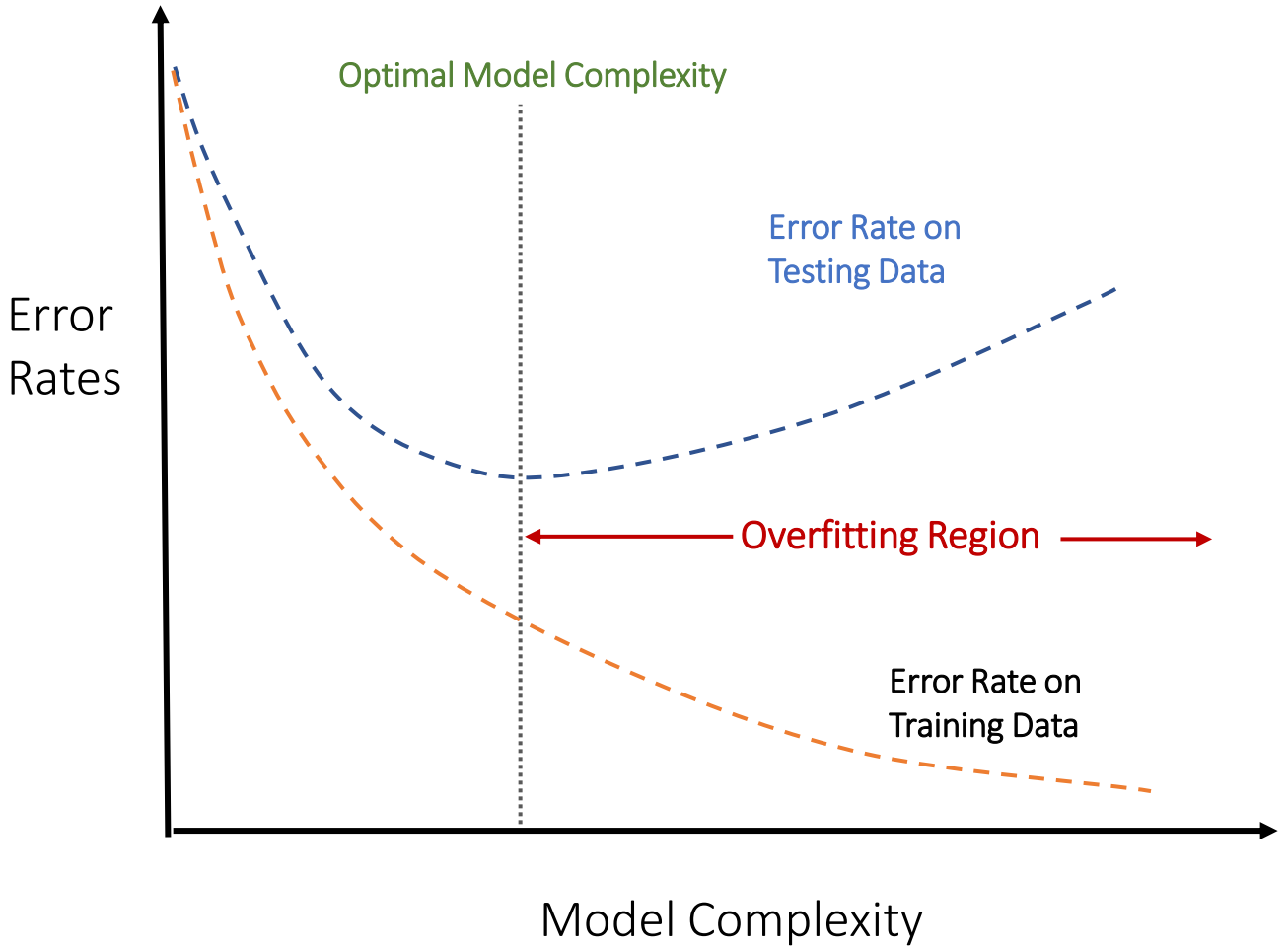


Machine Learning Process

Training and Test Sets

Generally, as we go from simple models to more complex

- Training error constantly decreases
- Test error decreases initially, but increases when we are over-fitting
- Goal is to find the optimal model complexity to ensure good performance on new data



Machine Learning Process

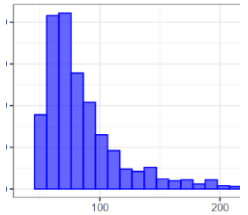
Feature Engineering

- Removing Skewness
- Center and Scale (Z-Transform)
- Dummy Variables
- Impute Missing Data
- …

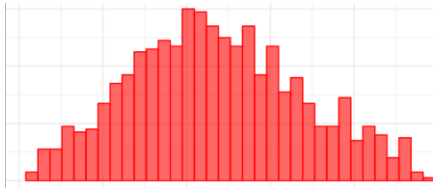
Response	Predictor 1	Predictor 2
Yes	12.4	Low
No	14.2	Moderate
Yes	16.8	High
…	…	…

Response	Predictor 1	Predictor_2_Moderate	Predictor_2_High
Yes	0.2	0	0
No	0.75	1	0
Yes	1.3	0	1
…	…	…	…

Predictor 1



Predictor 1
Skewness Transformation



Machine Learning Process