GBUS 738 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

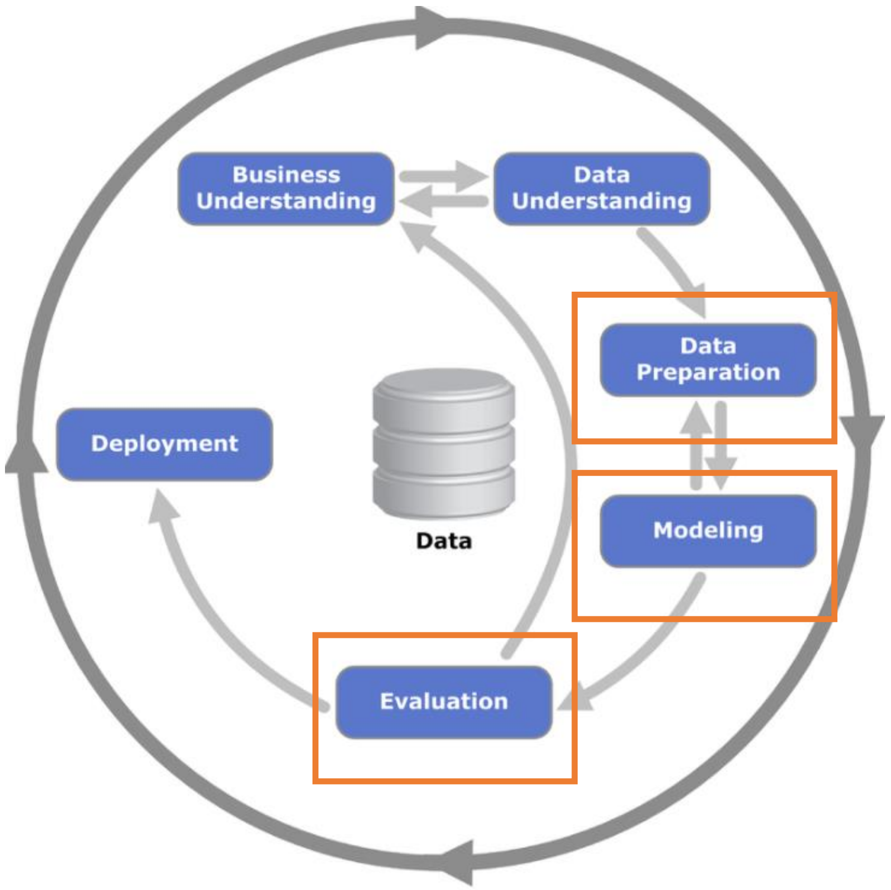
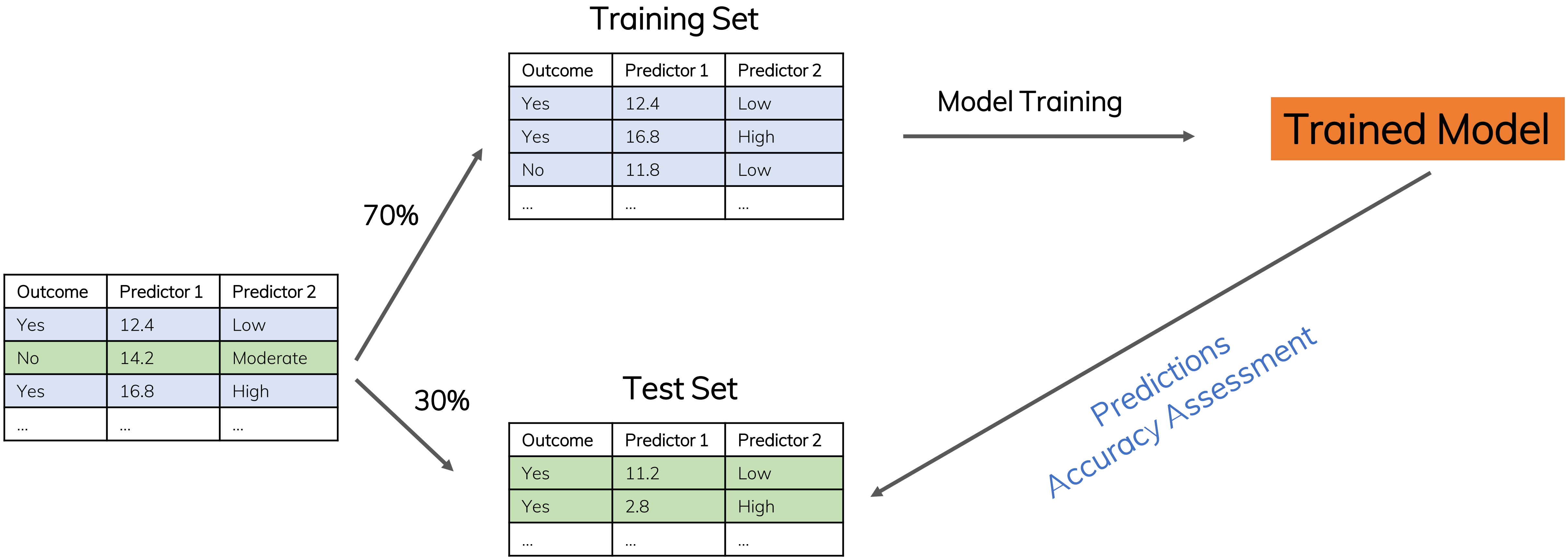


Image: https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining

Machine Learning Process



Outcome	Predictor 1	Predictor 2
Yes	12.4	Low
No	14.2	Moderate
Yes	16.8	High
...	...	...

70%

30%

Training Set

Outcome	Predictor 1	Predictor 2
Yes	12.4	Low
Yes	16.8	High
No	11.8	Low
...	...	...

Test Set

Outcome	Predictor 1	Predictor 2
Yes	11.2	Low
Yes	2.8	High
...	...	...

Model Training

Trained Model

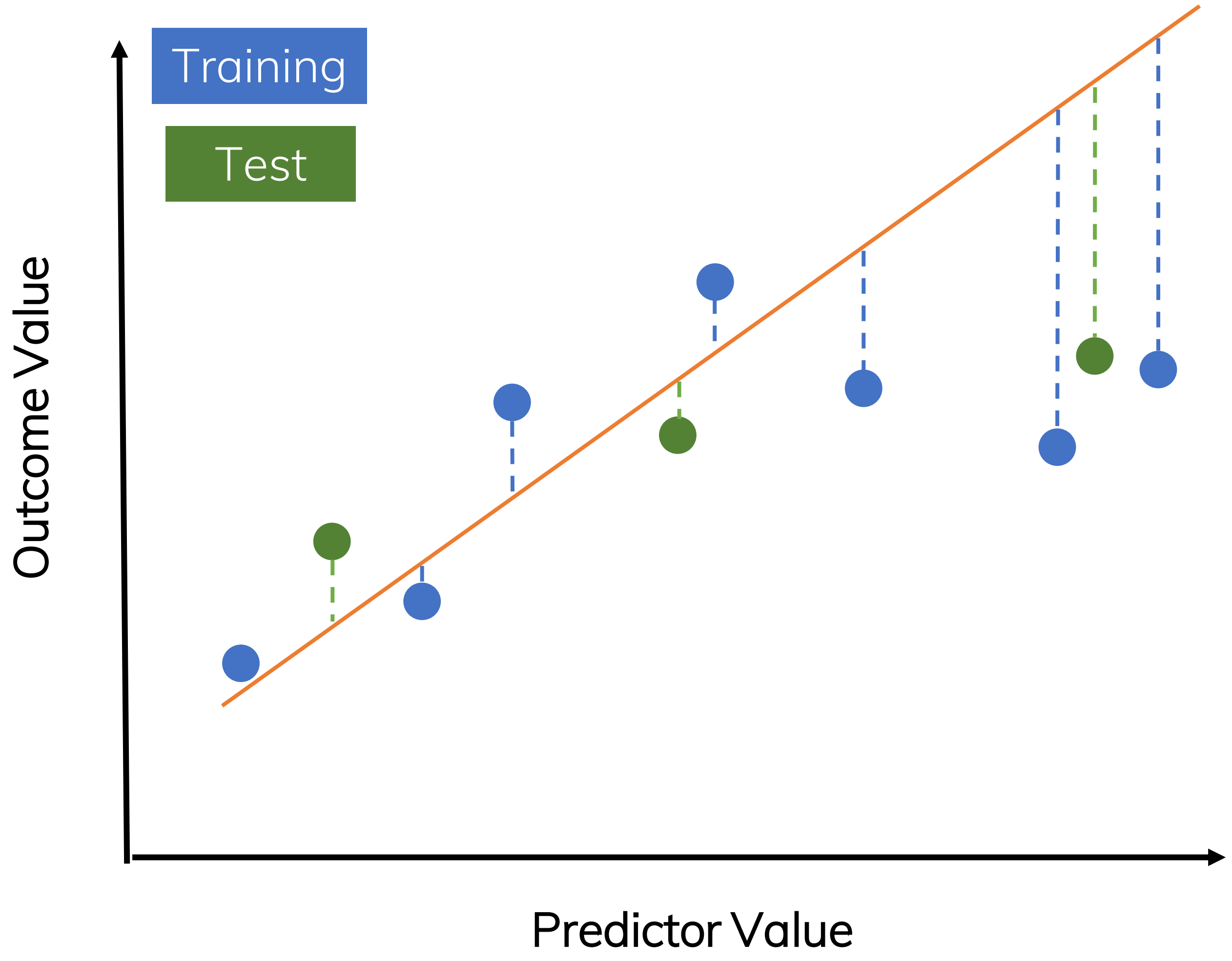
Predictions
Accuracy Assessment

Machine Learning Process

Training and Test Sets

Why split the data?

- Guard against
  - **Under-fitting**
    - Model can't capture complex trends in the data
    - Give away – poor accuracy on both training and test sets

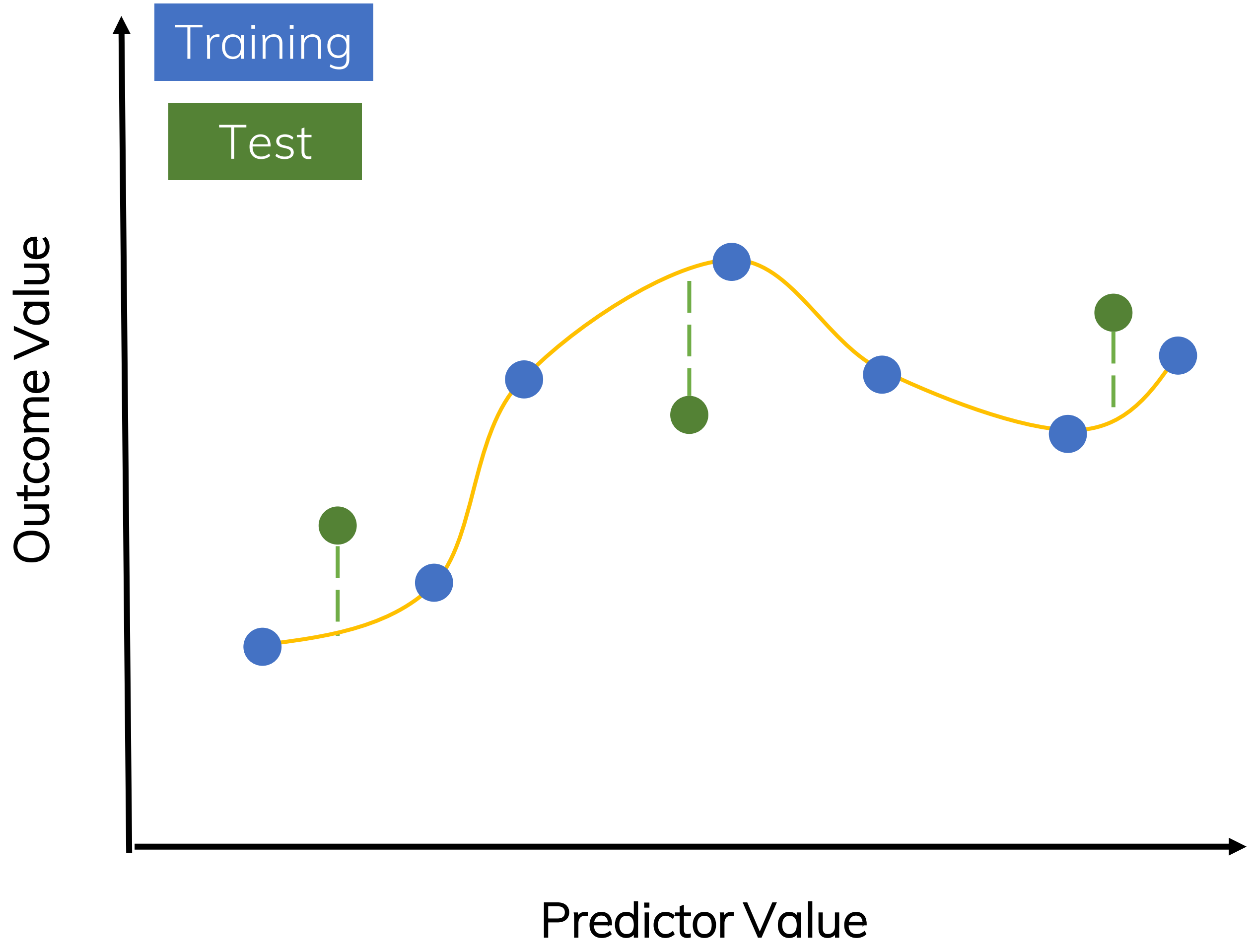


Machine Learning Process

Training and Test Sets

Why split the data?

- Guard against
  - **Over-fitting**
    - Model finds trends that don't exist
    - Give away – high accuracy on training data, poor accuracy on test data

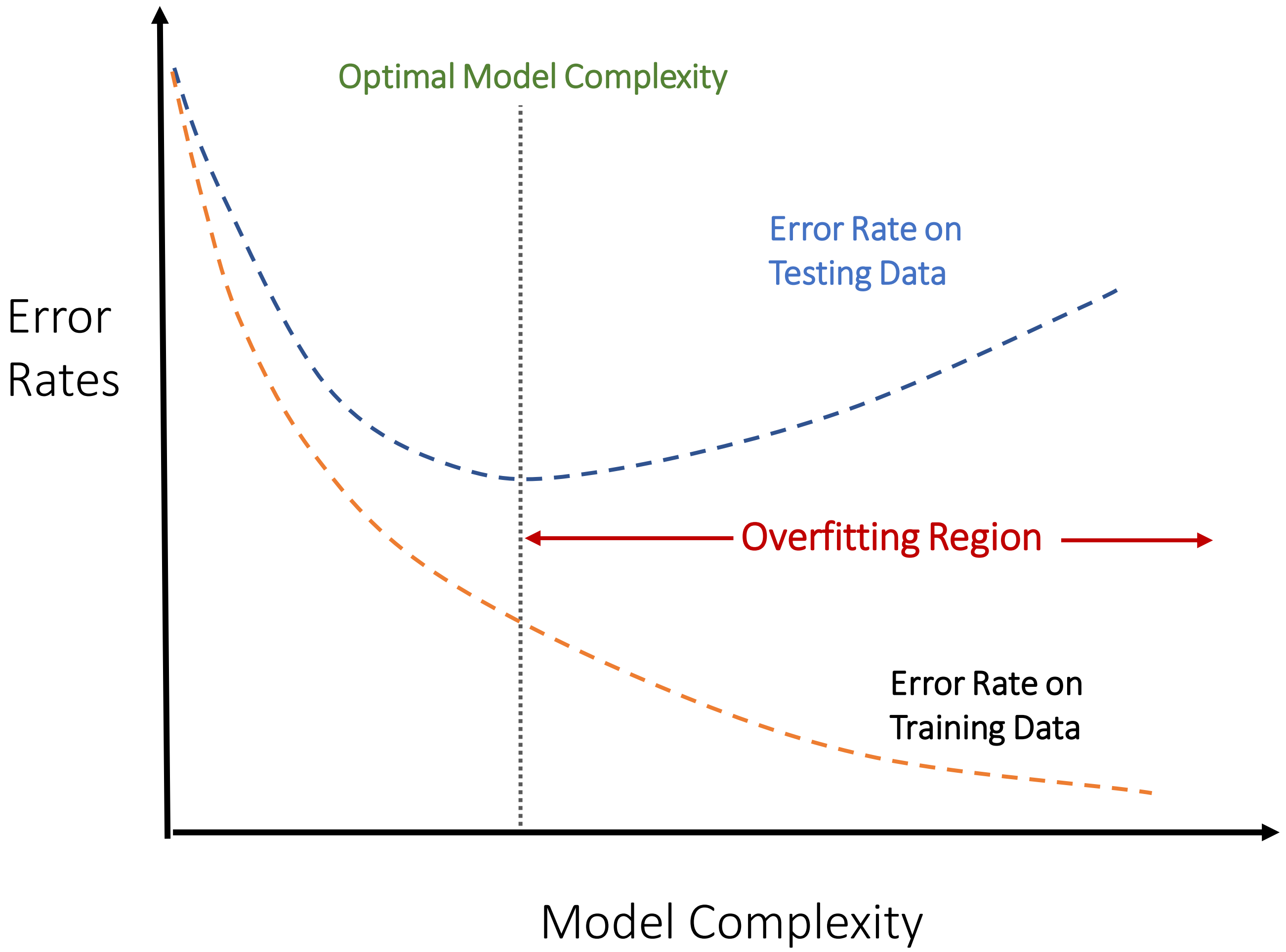


Machine Learning Process

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



School of Business

Machine Learning Process

Feature Engineering

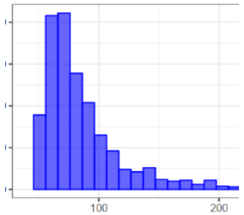
- Removing Skewness
- Center and Scale (Z-Transform)
- Dummy Variables
- Impute Missing Data
- …

outcome	predictor_1	predictor_2
yes	12.4	low
no	14.2	moderate
yes	16.8	high
…	…	…

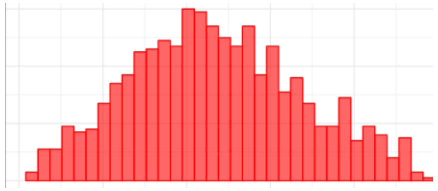
↓

outcome	predictor_1	predictor_2_moderate	predictor_2_high
yes	0.2	0	0
no	0.75	1	0
yes	1.3	0	1
…	…	…	…

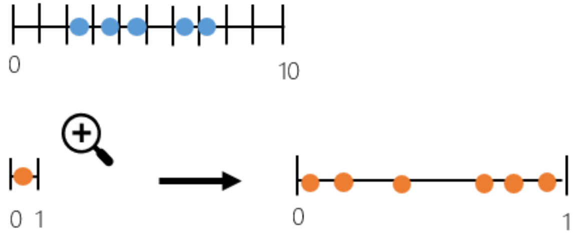
Predictor 1



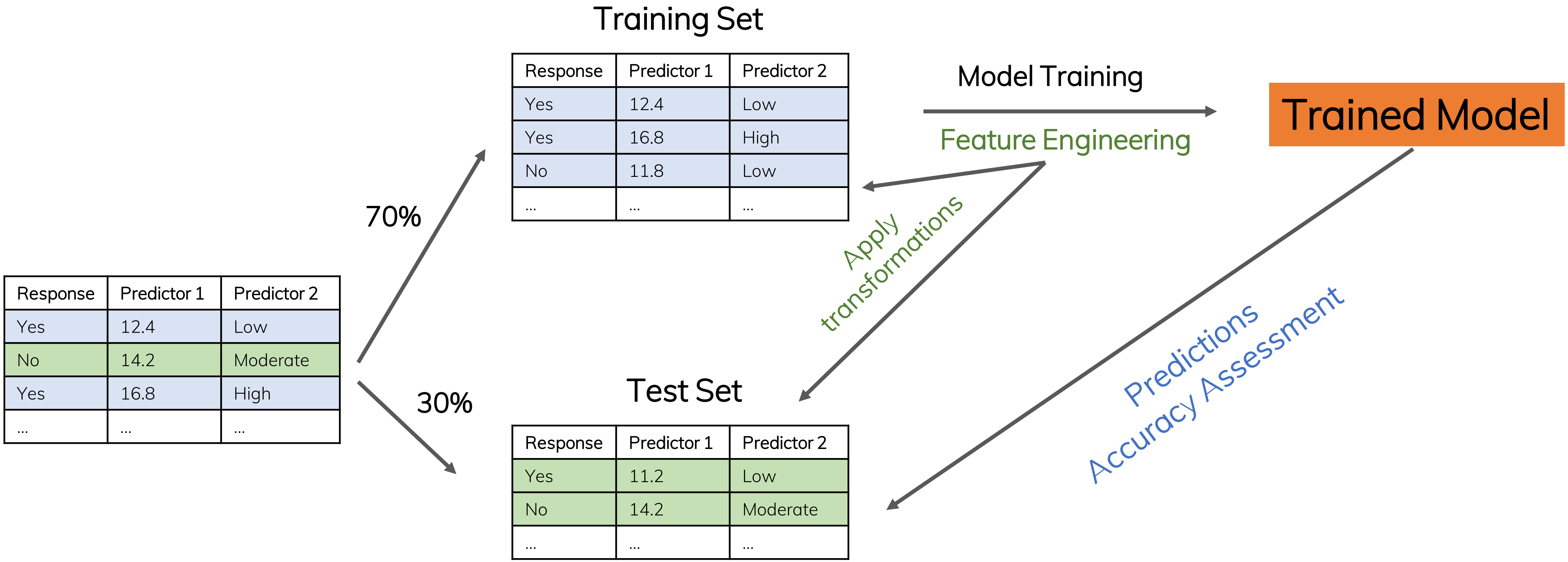
Predictor 1
Skewness Transformation



Scaling Predictor Values



Machine Learning Process



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

70%

30%

Training Set

Response	Predictor 1	Predictor 2
Yes	12.4	Low
Yes	16.8	High
No	11.8	Low
…	…	…

Test Set

Response	Predictor 1	Predictor 2
Yes	11.2	Low
No	14.2	Moderate
…	…	…

Model Training

Feature Engineering

Apply transformations

Trained Model

Predictions

Accuracy Assessment