Definition 1 (Predictive Modeling)**.** *The process of developing a mathematical tool or model that generates an accurate prediction.*

Why do predictive models fail?

1. Inadequate pre-processing of the data.
2. Inadequate model validation.
3. Unjustified extrapolation, e.g.:the application of the model -> data is in an application not seen by the model.
4. Over-fitting the model to the existing data.

0.1 Prediction Versus Interpretation

For applications that require accurate predictions:

- Historical Data — Interested in accurately projecting the chances of future events, not why they occurred.
- Pricing Homes on Zillow — Interested in accurate price estimates, not how they predicted them.
- Medical — Predict patient response to a treatment based on a significant number of factors.

Secondary considerations are given to the interpretation of the data, not primary. Higher accuracy models are more complex, and as a consequence reduce interpretability.

0.2 Ingredients of Predictive Models

- The best models are influenced and produced by modelers with expert knowledge and field context of the problem.
- These modelers can pre-filter irrelevant information and place more meaningful constraints on data sets.
- Modelers can also put personal biases on data.
- Predictive modeling is not a substitute for intuition, but a complement.
- Traditional experts make better decisions with results of statistical prediction.

0.3 Terminology

Definition 2 (Sample, Data Point, Observation, Instance)**.** *A single independent unit of data (A Customer), or a subset of data points (A Training Sample).*

Definition 3 (Training Set)**.** *Contains the data used to develop models.*

Definition 4 (Validation Set)**.** *Contains the data used to evaluate the performance of the final set of candidate models.*

Definition 5 (Predictors, independent Variables, Atrtributes, Descriptors)**.** *The data used as input for the prediction equation.*

Definition 6 (Outcome, Dependent Variables, Target, Class, Response)**.** *The outcome event, or quantity that is being predicted.*

Definition 7 (Continuous Data)**.** *Has natural, numeric scales. Ex: Blood pressure, cost, quantity.*

Definition 8 (Categorical, Nominal, Attribute, or Descrete Data)**.** *Has specific values without scale.*

Definition 9 (Model-Building, -Training, Parameter Estimation)**.** *The process of using data to determine values of model equations.*

0.4 Notation

1. n = the number of data points.
2. P = the number of predictors.
3. y_i = the ith observed value of the outcome, $i = 1 \dots n$.
4. $\hat{y}_i$ = the predicted outcome of the ith data point, $i = 1 \dots n$.
5. $\overline{y}$ = the average or sample mean of the n observed values of the outcome.
6. $\mathbf{y}$ = a vector of all n outcome values.
7. x_{ij} = the value of the jth predictor for the ith data point, $i = 1 \dots n$ and $j = 1 \dots P$.
8. $\mathbf{x}_i$ = a collection of the P predictors for the ith data point, $i = 1 \dots n$.
9. $\mathbf{X}$ = a matrix of P predictors for all data points; this matrix has n rows and P columns.
10. $\mathbf{X}'$ = the transpose of $\mathbf{X}$; this matrix has P rows and n columns.

0.5 Other Notational Guidelines

1. C = the number of classes in a categorical outcome.
2. C_l = the value of the lth class level.
3. p = the probability of an event.
4. p_l = the probability of the lth event.
5. $P_r[.]$ = the probability of event.
6. $\sum_{i=1}^{n}$ = the summation operator over the index i.
7. $\mathbf{\Sigma}$ = the theoretical covariance matrix.
8. $E[\cdot]$ = the expected value of $[\cdot]$.
9. $f(\cdot)$ = a function of $[\cdot]$; $g(\cdot)$ and $h(\cdot)$ also represent functions throughout the text.
10. β = an unknown or theoretical model coefficient.
11. b = an estimated model coefficient based on a sample of data points.