Basic Foundations

Predictive Modeling Goals

- Predictive modeling is not just about making predictions
- There 3 main goals in predictive analytics
- Understanding the analytics question(s) you are trying to answer will give you clarity about the analysis goals
- Defining the specific goal(s) of your analysis will help you decide which modeling method is most appropriate.
- Your analysis goal could be one, many or all of these 3:

1. **Inference**
2. **Interpretation**
3. **Prediction**

Modeling Goal 1: Inference

- Once the analytic problem is formulated, and based on business domain knowledge and data mining exploration, one formulates **hypothesis** or **predictions**, e.g.:
  - Increased advertising leads to higher sales
  - Increased minimum wage leads to less unemployment
  - More years of college education leads to higher income
  - Low amounts of aspirin reduces heart disease
  - Use predictive models to test the hypotheses
- **Testing** hypotheses, e.g.:
  - H_0: What we are hypothesizing
  - H_0 can never be accepted, only failed to reject it
  - H_A: Alternative hypothesis if H_0 is rejected
  - It is better to set H_A as what we are trying to prove, e.g.,
    - ✓ H_0: $\beta=0$
    - ✓ **If rejected, then H_A: $\beta\neq0$ → effect is significant**

Modeling Goal 2: Interpretation

- Inference and interpretation are related
- Inference is about testing specific effects, whereas
- Interpretation is about **explaining** what the model results are telling us
- For example:
  - Holding weight, size and cylinder size constant (i.e., controlling for), adding one more vehicle cylinder reduces gas mileage by 2 mpg
  - Holding body weight, exercise activity and cholesterol level constant (i.e., controlling for), smoking one additional cigarette per day increases the chances of a heart attack by 0.5%
- **Some** methods are **great** for **interpretation** (e.g., OLS regression, logistic regression); **others** are **not** (e.g., regression trees, support vector machines)

Modeling Goal 3: Prediction

- Which model is the most **accurate** at predicting outcomes for new observations? e.g.,
  - What are the chances that a new credit card transaction is fraudulent?
  - What are the chances that a new email is spam?
  - How much income is a college graduate from Kogod is expected to earn 5 years after graduation?
- **Machine learning** methods are important to evaluate **predictive accuracy**:
  - **Train** the model using part of the data
  - **Test** it on data not used to train the model
  - **Re-sample** train and test sets and **re-train** and **re-test**
  - **Cross-validation:** testing the mean squared error (**MSE**) of the **trained model** using the **test data**
  - Select the **model** with the **lowest MSE**
  - As **new data** arrives, re-train and re-evaluate the model

KOGOD SCHOOL
of
BUSINESS