Variable Types

Variable Types

- Understanding variable types is the **first step** in pre-processing
- Dependent (predicted or response) vs. Independent (predictor) variables

Ex: Income = f (Age, Education Level, School, City)

- **Variable Content Types:**
  - **Numeric:**
    - **Continuous** – e.g., 4.32, 0.48 – can do math
    - **Discrete** – e.g., 10, 20, 30, etc. – no decimals, can do math
    - **Binary** – 0, 1 (e.g., no/yes; approved/declined; sold/not)
  - **Character:**
    - **Ordinal** – e.g., 1-7 survey scale (can't do math, e.g., SET's)
    - **Categorical** – e.g., Foreign/Domestic; Urban/Suburban/Rural
    - **Text** – e.g., E-mail subject or content, speech transcriptions

Key Modeling Issues: Independent Variables

- **Independent** (predictor) **variables** can be continuous, discrete, binary or ordinal – it is OK if they are skewed or not normally distributed.
- **Categorical variables** cannot be used in most predictive models (e.g., regression) without some transformation, but can be used with some methods (e.g., frequencies, ANOVA, Chi-Square); but
- **Categorical (or Factor) variables** can also be transformed into counts, aggregates, or groups of binary variables, which can be used as quantitative predictors
- **Text data** often needs to be pre-processed to create meta data (e.g., word counts, character counts, event counts, matching text, synonyms, etc.

Key Modeling Issues: Dependent Variables

- **Dependent** variables **don't** need to be **normally distributed** if the resulting **errors** are normally distributed, but it is often desirable to transform non-normal dependent variables into normally distributed variables, which will provide better distribution of errors and more robust models.
- The type of **dependent** (predictor) **variable** has a substantial impact on the **modeling methods** that can be used:
  - **Continuous** – Regression and decision tree models
  - **Binary** – Classification models like, Logistic and Probit regressions, classification trees.
  - **Discrete** – Multinomial Logistic regression, Discriminant Analysis
  - **Ordinal** – Ordered Logistic, Ordered Probit regressions
  - **Time Sensitive** – Forecasting regression models

KOGOD SCHOOL
of
BUSINESS