Jason “Scott” Person

Learning Response 4.1: Feature Engineering

20250429

# Response Questions

Summarize, explain, and illustrate the following concepts. Add screenshots to help illustrate.

### From [McCormick, The Essential Elements of Predictive Analytics and Data Mining](https://www.linkedin.com/learning/the-essential-elements-of-predictive-analytics-and-data-mining/)

* What is data construction?

The transformation (using this term broadly) of incoming data to increase its value in the modeling process. This can include summarizing, combining, extraction, derivation, and other techniques to create additional, higher value columns, or new records.

* What are relevant terms and synonyms?
  + Feature Engineering
  + Transformation
  + Derived Attributes (this seems more like a product of data construction)
* Throughout the video, McCormick stresses the importance of data construction in several different ways. Summarize and list those.

Note: I’m unsure if this is asking us why data construction is important or the method McCormick uses to express its importance. Seems to be the latter.

* + Quoting from CRISP-DM
  + He worked on a data churn project with over 800 variables. Almost all of those variables were constructed.
  + McCormick states that 9 or 10 out of 10 of his most important variables, over his career, have been constructed.

Here are some reasons that it is important:

* + Most variables come from data construction
  + Clarifies for the algorithm the insight that we want to get from the data
  + Increases the fidelity of some attributes
  + Provides a mechanism and process for subject matter experts to add value directly to the process
* What are three examples he gives of constructed variables?
  + Ratios and formulas (BMI etc.)
  + Date derivations (length of time as a customer [derived from date of first transaction])
  + Extracted value from smart key (VIN number contains letter indicating model year)

### From [McCormick, Data Assessment for Predictive Modeling](https://www.linkedin.com/learning/data-science-foundations-data-assessment-for-predictive-modeling/)

* Define and illustrate the following basic levels of measurement (aka data types):
  + Nominal (aka Categorical) – separate and distinct categories (divisions of the data) that are not meaningfully ranked. Examples include sex, race, vehicle type
  + Ordinal – separate and distinct categories that are meaningfully ranked. Ticket class (from the Titanic dataset) and bucketed ages are examples.
  + Continuous (aka Scale) – variables where taking an average makes sense or are measured on a continuum. These include temperature, weight, speed etc.
* Dummy Coding (aka One-Hot Encoding)
  + Define and give examples to illustrate what dummy coding is.

Dummy coding takes a categorical variable and turns it into n binary variables where n is the number of categories for the variable. As an example, sex has two categories – male and female. These would be dummy coded as “is a” variables with binary indicators as true or false. An individual would have is a male and is a female with true and false set as appropriate. By definition only one of these variables could be set to true.

Of note is that a variable with many categories would result in many variables when dummy coded.

* + Notes:
    - One-hot encoding is a frequently used alternate term for this.
    - Python / pandas does not perform this work automatically. But it has handy methods to implement it.
* Expanded levels of measurement, including
  + Potential ID Fields – think about the potential to link to other data where it would be a possible key in the join condition.
  + Nominal variable sub-types:
    - One value – zero value for modeling
    - Binary – true/false is classic example.
    - Low order
    - High order – lots of categories results in lots of variables when dummy coded (12 to 15 is when you start thinking about this)
    - Very high order – product code/SKU, these are “deadly” to supervised machine learning models. Don’t include them.
  + Dates – and extracting meaningful values – Rarely useful in predictive models; typically use date arithmetic to transform into useful metrics (dates since x or day of week)
  + Text — and extracting meaningful values – turn text into something the supervising model can use. Scan unstructured text to see if there is useful, discrete data that can be extracted.

**Now from this group of McCormick videos (combined) ...**

* *Taking an initial look at possible key variables*
* *Dealing with duplicate IDs and transaction data*
* *How many potential variables (columns) will I have?*

**Create bulleted lists of key insights and recommendations from the entire group of three videos, categorized under these key headings:**

* Examples of likely relevant and useful variables
  + ID variables when linking with same or separate dataset or for deriving metrics like counts
  + Variables that contain metrics for aggregation then binning
  + Application installed (Has\_Ap)
  + TV\_Customer
  + Spend
  + Number of Features
  + Churn
* Examples of usually useless variables
  + ID variables – when modeling
  + Title
  + Name
  + Address
  + City, State (questionable; may have usage in aggregating/grouping)
* Ideas for constructing useful features by extracting data from otherwise useless fields
  + Identifying multiple occurrences of a given customer (CC number), household (address), phone lines
  + Lat/Long – when calculating distance, or when grouping geographically (requires heavy transformation)
  + Deriving customer duration from month field (ostensibly when less than 12)
  + Aggregation across months or count of months
  + Time since first date give duration as customer
  + Deriving country of assembly, model year, manufacturer name from VIN (not in this video set)
  + DOB 🡪 Age

### From [Berkun, Python: Working with Predictive Analytics](https://www.linkedin.com/learning/python-working-with-predictive-analytics/)

First, note that “categorical” is a frequently used synonym for “nominal.” (McCormick favors “nominal.” Pandas and therefore Python users tend to use “categorical.”)

Define and give examples to illustrate these terms:

* Label Encoding – encode a variable with two categories as 0 and 1 values. We don’t need to add an additional column as a single column can encode this. The most common example is True/False data, but we’ve also seen this with gender in our feature engineering on the Titanic dataset.
* One-Hot Encoding (aka “dummy coding”) – referred to as dummy encoding elsewhere. This takes the distinct set of categories for a variable and creates a variable for every set member. These are then encoded as 0 or 1 for each record. Examples of this is for color where a column is created for each possible color and each record has a one in the appropriate color’s column with zeros for all other colors.

***Additional note: Label encoding is the preferred option when:***

* **the data is ordinal**, such as Pclass 1, 2, or 3 — where there the classes exist in a hierarchy (like 1st, 2nd, 3rd)
* **the data has more than approx. 12 or 15 categories** (what McCormick calls “high order nominals”) — in which case creating a new column for each category becomes counterproducive.