# Preparing Your Data for Successful Predictive Modeling

Don McCormack – Principal Enablement Engineer

JMP Division of SAS Institute

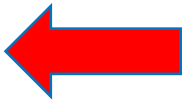
# Preparing Data for Predictive Modeling

## What is Data Mining?

- Evaluating future observations (vector or scalar) as correct or close.
- Devising a set of rules for classifying current or future observations into groups or as anomalous
  - Market segmentation
  - Market basket analysis
  - Outlier screening
  - Image recognition
- Optimizing a sequence of events
  - Traveling salesman problem
  - Logistics
  - Self driving cars

# Preparing Data for Predictive Modeling

## Predictive Modeling Steps

1. Curation, organization, and cleaning
2. Preparation 
3. Modeling
4. Evaluation

# Preparing Data for Predictive Modeling

## Preparation

- Our emphasis will be on predicting future observations.
- We will assume the data has been curated, organized, and cleaned.
- Five topics will be covered :
  - Imputation
  - Data Validation
  - Sampling
  - Feature Creation
  - Predictor Screening
- As with data cleaning, analysis of the data may cause you to reexamine preparation steps.

# Preparing Data for Predictive Modeling

## Data Used for Illustration: Pima Diabetes

- Incidence of diabetes in Pima culture women.
- Response: Individual has diabetes
  - Yes: diabetes was diagnosed between one and five years of observation
  - No: no diabetes seen in the five years after the observation
- Predictors: times pregnant, blood glucose, diastolic BP, tricep skinfold measure, blood insulin, BMI, diabetes pedigree (numerical measure from survey), Age
- Observations: 768
- Source

# Preparing Data for Predictive Modeling

## Data Used for Illustration: Direct Marketing

- Optimize donations to national veterans organization (less cost of mailing) from direct mailing. Part of a national data mining challenge.
- Responses: Donation (binary), donation amount (continuous)
- Predictors: 481
- Observations: 95,412 for training data, approximately the same for validation/testing data set

# Preparing Data for Predictive Modeling

## Imputation

- Missing values are problematic. They may lead to data reduction because of case-wise deletion.
- They may also cause model bias depending on why observations are missing.
- Types of missing data:
  - **Missing Completely at Random** (MCAR) – Missingness is independent of observed or missing cases.
    - Ex: Sometimes, doctors don't have time to perform the test (assuming time constraints affect doctors equally).
  - **Missing at Random** (MAR) – The distribution of missing values is a function of what is in the data.
    - Ex: Women who are pregnant more often are more likely to have an inconclusive test value.
  - **Not Missing at Random** (NMAR) – The distribution of missing values is a function of what isn't observed in the data.
    - Ex: Doctors who are less efficient are less likely to perform the test.

# Preparing Data for Predictive Modeling

## Imputation

- Imputation is a way to fill in the missing values. It can be applied to MCAR and MAR cases. NMAR cases are avoided unless there is additional information about the missing data mechanism.
- Simple Imputation:
  - Univariate: based on some function of the variable (e.g., mean or median)
  - Multivariate: based on some function of the correlation between variables.
  - Impute (in R)
- Model based:
  - Use a model built on the complete values to predict the missing values.
  - Use an iterative modeling technique robust to missing values (e.g., PLS, EM)

# Preparing Data for Predictive Modeling

## Imputation

- Nonparametric approaches: Random Forests, nearest neighbors, clustering.
- Multiple Imputation is iterative an iterative approach where multiple passes are made at the data with each iteration producing a different set of imputed values. Randomization, such as resampling the with replacement, may be used to facilitate the process. The results are collected and aggregated (e.g., averaged) to produce a set of final values.
- While most imputation is performed on continuous data, there are methods that can be used with categorical data (e.g., clustering, multiple correspondence analysis).
- Older techniques such as hot deck and cold deck have lost favor to more sophisticated routines.

# Preparing Data for Predictive Modeling

## Imputation – Examples (simple)

| Obs | Dia BP | Tricep Skin Fold | Insulin | BMI |
|---|---|---|---|---|
| 1 | 72 | 35 | • | 33.6 |
| 2 | 66 | 29 | • | 26.6 |
| 3 | 64 | • | • | 23.3 |
| 4 | 66 | 23 | 94 | 28.1 |
| 5 | 40 | 35 | 168 | 43.1 |
| 6 | 74 | • | • | 25.6 |
| 7 | 50 | 32 | 88 | 31 |
| 8 | • | • | • | 35.3 |
| 9 | 70 | 45 | 543 | 30.5 |
| 10 | 96 | • | • | • |
| 11 | 92 | • | • | 37.6 |
| 12 | 74 | • | • | 38 |
| 13 | 80 | • | • | 27.1 |

| Obs | Dia BP | Tricep Skin Fold | Insulin | BMI |
|---|---|---|---|---|
| 1 | 72 | 35 | 125 | 33.6 |
| 2 | 66 | 29 | 125 | 26.6 |
| 3 | 64 | 29 | 125 | 23.3 |
| 4 | 66 | 23 | 94 | 28.1 |
| 5 | 40 | 35 | 168 | 43.1 |
| 6 | 74 | 29 | 125 | 25.6 |
| 7 | 50 | 32 | 88 | 31 |
| 8 | 72 | 29 | 125 | 35.3 |
| 9 | 70 | 45 | 543 | 30.5 |
| 10 | 96 | 29 | 125 | 32.3 |
| 11 | 92 | 29 | 125 | 37.6 |
| 12 | 74 | 29 | 125 | 38 |
| 13 | 80 | 29 | 125 | 27.1 |

Replace with median

| Obs | Dia BP | Tricep Skin Fold | Insulin | BMI |
|---|---|---|---|---|
| 1 | 72.00 | 35.00 | 225.90 | 33.60 |
| 2 | 66.00 | 29.00 | 68.95 | 26.60 |
| 3 | 64.00 | 22.18 | 270.36 | 23.30 |
| 4 | 66.00 | 23.00 | 94.00 | 28.10 |
| 5 | 40.00 | 35.00 | 168.00 | 43.10 |
| 6 | 74.00 | 22.40 | 121.89 | 25.60 |
| 7 | 50.00 | 32.00 | 88.00 | 31.00 |
| 8 | 74.08 | 31.45 | 134.52 | 35.30 |
| 9 | 70.00 | 45.00 | 543.00 | 30.50 |
| 10 | 96.00 | 32.92 | 159.06 | 34.20 |
| 11 | 92.00 | 33.19 | 121.10 | 37.60 |
| 12 | 74.00 | 35.60 | 259.80 | 38.00 |
| 13 | 80.00 | 27.71 | 204.47 | 27.10 |

Multivariate normal

Data shown: first 13 observations, columns with missing values.

# Preparing Data for Predictive Modeling

## Imputation – Examples (multiple)

| Obs | Dia BP | Tricep Skin Fold | Insulin | BMI |
|---|---|---|---|---|
| 1 | 72 | 35 | • | 33.6 |
| 2 | 66 | 29 | • | 26.6 |
| 3 | 64 | • | • | 23.3 |
| 4 | 66 | 23 | 94 | 28.1 |
| 5 | 40 | 35 | 168 | 43.1 |
| 6 | 74 | • | • | 25.6 |
| 7 | 50 | 32 | 88 | 31 |
| 8 | • | • | • | 35.3 |
| 9 | 70 | 45 | 543 | 30.5 |
| 10 | 96 | • | • | • |
| 11 | 92 | • | • | 37.6 |
| 12 | 74 | • | • | 38 |
| 13 | 80 | • | • | 27.1 |

| Obs | Dia BP | Tricep Skin Fold | Insulin | BMI |
|---|---|---|---|---|
| 1 | 72.00 | 35.00 | 199.65 | 33.60 |
| 2 | 66.00 | 29.00 | 100.50 | 26.60 |
| 3 | 64.00 | 30.08 | 166.24 | 23.30 |
| 4 | 66.00 | 23.00 | 94.00 | 28.10 |
| 5 | 40.00 | 35.00 | 168.00 | 43.10 |
| 6 | 74.00 | 22.71 | 102.76 | 25.60 |
| 7 | 50.00 | 32.00 | 88.00 | 31.00 |
| 8 | 77.24 | 27.49 | 145.14 | 35.30 |
| 9 | 70.00 | 45.00 | 543.00 | 30.50 |
| 10 | 96.00 | 28.91 | 160.37 | 30.07 |
| 11 | 92.00 | 30.63 | 170.54 | 37.60 |
| 12 | 74.00 | 36.78 | 224.34 | 38.00 |
| 13 | 80.00 | 35.86 | 218.14 | 27.10 |

Matrix completion

| Obs | Dia BP | Tricep Skin Fold | Insulin | BMI |
|---|---|---|---|---|
| 1 | 72 | 35 | 132 | 33.6 |
| 2 | 66 | 29 | 60 | 26.6 |
| 3 | 64 | 30 | 328 | 23.3 |
| 4 | 66 | 23 | 94 | 28.1 |
| 5 | 40 | 35 | 168 | 43.1 |
| 6 | 74 | 20 | 160 | 25.6 |
| 7 | 50 | 32 | 88 | 31 |
| 8 | 82 | 33 | 64 | 35.3 |
| 9 | 70 | 45 | 543 | 30.5 |
| 10 | 96 | 46 | 120 | 33.2 |
| 11 | 92 | 36 | 57 | 37.6 |
| 12 | 74 | 23 | 328 | 38 |
| 13 | 80 | 17 | 325 | 27.1 |

MICE (R package, using defaults)

Data shown: first 13 observations, columns with missing values.

# Preparing Data for Predictive Modeling

## Imputation

- Simple
  - Advantages: easy to implement, simple to understand, computational fast.
  - Disadvantages: may not capture the structure of the data sufficiently well and bias the results.
- Multiple
  - Advantages: Flexible and robust
  - Disadvantages: Can be difficult to implement and explain. Can be computationally slow for very large data sets.

# Preparing Data for Predictive Modeling

## Data Validation

- If the modelling algorithm is flexible what guards against overfitting (i.e., producing predictions that are too optimistic)?
  - Put another way, how do we protect from trying to model the noise as part of our model?
- Solution – Hold back part of the data, using it to check against overfitting. Break the data into two or three sets:
  - The model is ***built*** on the training set.
  - The validation (or tuning) set is used to ***select*** model features by comparing predictive quality on data not used to build the model.
  - The test set is often used to ***evaluate*** how well model predicts independent of training and validation sets.
  - Common methods include holdback, bootstrapping, and k-fold.

# Preparing Data for Predictive Modeling

## Data Validation – Holdback

- Holdback splits the data into two (train/tune) or three (train/tune/test) parts. The goal is that each part is representative of the same population to be studied.
- Random assignment is the simplest and assumes all the data comes from the same population.
- Stratified random sampling can be used if there is belief that within strata representation is different from across strata representation. For example, sampling may be done to assure equal representation within each lot of material or from observations coming from the same time period.
- Clustering can be used to determine sampling strata.

# Preparing Data for Predictive Modeling

## Data Validation – Bootstrapping

- Bootstrapping is an iterative technique where at each iteration, the data is repartitioned randomly (with or without stratification).
- A variation on this is to weight the observations, changing the weights at each iteration. This technique can be applied to very small data sets if care is exercised (see Wu, Boos, & Stafanski).

# Preparing Data for Predictive Modeling

## Data Validation – k fold validation

- K-fold crossvalidation is an iterative technique that splits the data into k equal or nearly equal parts.
  - At each iteration one of the parts is held out for validation and the remaining parts are used to build the model.
  - This continues until all parts are held out once.
  - The final model can be selected from the individual models or a function of all models.
- K-fold is useful when there are a small number of observations. It may have a tendency to overfit, however.
- The typical choice for k is between 5 and 10.

# Preparing Data for Predictive Modeling

## Sampling

- Sampling can be used insure representativeness of validation partitions. The tuning and testing data partitions (the folds if k-fold validation is used) should look about the same.
  - Techniques: clustering
- It can also be used to account for response class imbalance in the response:
  - A categorical response where the level of interest has a much smaller proportion than the other levels.
  - A continuous target when the focus is on a small proportion of the responses such as the first or last decile.

# Preparing Data for Predictive Modeling

## Sampling

- Handling class imbalance:
  - Oversampling – sampling more observations from the level with the target value. This may require sampling with replacement or creating similar cases via simulation.
    - A disadvantage of this approach is that it artificially inflates the sample size or makes assumptions about the structure of the data.
  - Undersampling – sampling fewer observations from the non target levels.
    - A disadvantage of this approach is not adequately capturing the structure of the non-target levels.
  - Weighting – Putting a higher weight on the target level. One way of doing this is by associating a cost (or profit) for misclassification.

# Preparing Data for Predictive Modeling

## Feature Creation

- Feature creation focuses on capturing relationships between the predictors and responses that don’t exist in the data and are not recoverable by modeling technique used.
  - Adding features using existing or external data
  - Aggregation/disaggregation
  - Transformation/standardization
  - Principal components/latent structures

# Preparing Data for Predictive Modeling

## Feature Creation – Adding Features

- The RAMNT variables in the direct marketing data contain information on how much was given for each of the mailings in the previous two years (24 mailings). The following features may not be captured solely through modeling:
  - A person who donated *n* of the last *p* mailings
  - A person who donated at least $*x* for one mailing
  - A person who made a large donation more than *p* mailings ago
  - The number of mailings a person received over this period
- Focusing on the Zip variable, external information such as median income by zip code may be helpful.

# Preparing Data for Predictive Modeling

## Feature Creation – Aggregation/Disaggregation

- Aggregation can occur across multiple variables or within multilevel variables.
  - Combining binary variables to get counts or continuous variables to get totals.
  - Combining levels of a categorical variable can be based on:
    - Convenience: combine all levels with fewer than *n* observations into a single category
    - Intuitive appeal: combine all 5 digit zips into 3 digit zips
    - Perceived relationship with the response: combine high income zips. This can be accomplished through techniques such as clustering or regression trees.
- Disaggregation occurs where a single variable is split into multiple variables. For example, the RFA variables can be split into their individual components (recency, frequency, amount).

# Preparing Data for Predictive Modeling

## Feature Creation - Transformations

- Transformations can used to account for asymmetry or skewness when dealing with linear models. Two common general transformations are Box-Cox and Johnson Family.
- Techniques such as neural networks and tree based methods are less reliant on transformation because of their underlying approach to modeling data.
- Standardization centers continuous data at zero and scales by the standard deviation. This is often done to put continuous variables on the same scale.
- Outliers may effect the results and should be investigated before this step.

# Preparing Data for Predictive Modeling

## Feature Creation

- Several multivariate techniques can be used to find latent structures in the data. These include:
  - Principal components analysis (PCA) and partial least squares (PLS).
    - PCA requires continuous inputs. Focus is on capturing variabilility across a multidimensional space.
    - PLS looks at the covariation between a set of inputs and outputs. Inputs and outputs are usually continuous though versions of PLS have be developed for categorical data.
  - Multidimensional scaling (MDS), factor analysis (FA), latent class analysis (LCA), multiple correspondence analysis (MCA). These techniques can be thought of as categorical versions of PCA
- These techniques can often be used for variable reduction and aggregation (clustering) as well.
- Caution should be applied if using these techniques as they may be computationally time consuming for large data sets.

# Preparing Data for Predictive Modeling

## Predictor Screening

- Predictor (feature) screening can be used for two purposes:
  - Finding good predictors
  - Removing unhelpful predictors
- In general, you want screening to be fast and flexible. Some ideas:
  - Removal of predictors with a large proportion of missing values or a variability close to zero.
  - Random forests are well suited for this.

# Preparing Data for Predictive Modeling

## Other Resources


- Wu, Boos, D.D., & Stefanski, L.A. (2007). Controlling Variable Selection by the Addition of Pseudovariables. Journal of the American Statistical Association, 102, 477, pp. 235-243.
- Elements of Statistical Learning
- Torgo, L. (2011). Data Mining with R, Learning with Case Studies. CRC Press. Boca Raton.
- Kuhn, M. & Johnson, K. (2013). Applied Predictive Modeling. Springer. New York
- KDNuggets
- University of California Irvine Machine Learning Repository

# Thank You!

jmp.com