# Understanding Predictive Modeling

## 1.Predictive Modeling Fundamentals

### Macros

#### %global inputs;

#### %let inputs=variable\_1 variable\_2 variable\_n (&inputs)

### EDA

#### Proc means n nmiss mean min max

#### Proc freq tables outcome\_var nominal\_var

## 2.Predictive Modeling Challenges

### Challenges

#### Observational data

##### Often massive

##### Contain errors and/or missing values

#### Mixed measurement scales

##### Dummy variables

##### Collapse levels

#### High dimensionality

##### Curse of dimensionality

#### Rare target events

##### Target event <1% of data set

##### Non-representative sample

###### -disproportionality over-represents target event cases-equal # events & non-events (oversampling)

### Analytic Challenges

#### Nonlinearity & interactions

#### Model selection

### Honest assessment

#### Splitting the data proc surveyselect, proc freq(tables outcome\_var\*selected), created training & validation data sets: if…then ouput else output

##### 2/3 training,1/3 testing

##### Stratified random sampling

#### k-fold validation

#### bootstrapping

# Fitting the Model

## Understanding the Logistic Regression Model

### Proc logistic

#### Using maximum likelihood method

#### Binary, ordinal, nominal

### Goodness of fit

#### Concordant/discordant

##### Higher % concordant pairs, lower % discordant, lower % tied pairs

##### Somer’s D-higher better

##### Gamma-higher better

##### Tau-a-higher better

##### c-higher better

#### AIC-smaller the better (smaller penalty vs SC)

#### SC-smaller the better (favor more parsimonious model)

### Global null hypothesis

#### Likelihood ratio

#### Score

#### Wald

### Which variables are sig.

#### Wald Chi square

### Analysis of Maximum Likelihood Estimates- relative importance

#### Standardized estimates-rank

### Score data-PROC LOGISTIC

#### Add the SCORE statement

#### Add the outmodel= and inmodel= option

#### Add the CODE statement

## Correcting for Oversampling

### Offset equation

#### Ln\*(proportion of non-events in pop\*proportion of events in sample/ proportion of events in pop\*proportion of non-events in sample)

#### incorporate sampling weights

### Scoring with correcting for oversampling

#### Tell SAS proportion of target event in population by defining a macro

#### Score new data set and adjust predicted probabilities back to pop scale

#### PRIOREVENT= in SCORE Statement, CODE statement

# Preparing the Input Variables

## Handling Missing Values-PROC LOGISTIC

### Missing completely at random (MCAR)?

#### Value of unobserved variables-lurking inputs?

#### Depend on value of other observed inputs or a combination of other inputs

#### Values of the target variable

### Default complete case analysis-uses only cases with no missing values

#### Good with only small number missing & MCAR

#### Not good for predictive

#### Can’t SCORE with missing values

### Imputation

#### Median (missing values <=50%, if >50% omit from model)

#### Mode

#### Mean

#### Binary variable-median

#### Regression

#### make sure makes sense based on Subject-matter knowledge

#### hot-decking (KNN) (sorts all cases by several variables-case closes to it)

#### new category

### Handling missing values

#### Numeric values

##### Create a missing value indicator

##### Impute a value

#### Categorical variables

##### Create a missing value level

##### Impute a value

### Coding steps-PROC STDIZE

#### Find variables with missing values: proc print, nummiss

#### Create array of variables

#### Method=median | mean | midrange

#### Cluster mean imputation-PROC FASTCLUS

## Working with Categorical Inputs

### Nominal variables

#### small # levels-dummy coding-reference cell coding-CLASS variable variable-effectcoding | referencecoding

#### large # levels-high dimensionality, i.e. zip-code-PROC CLUSTER, PROC FREQ tables variable / chisq;, data logpvalue=logsdf(‘CHISQ’, chisquare,degfree);, PROC SGPLOT, PROC SQL, PROC TREE, PROC SORT, PROC PRINT, DATA \_null\_ scoring code to assign branches to a cluster

##### smarter variables-map do other data-census, i.e. Urban, Rural, etc..

##### GreenAcre’s method-hierarchically clusters the levels based on reduction in the chi-square test of association between input & target: when min chi-square min loss of info

###### Collapses redundant categories

###### Collapses row with very small cell counts

###### Use with nominal, but not ordinal

###### steps

Perform cluster analysis

Compute log of p-value for each cluster & plot log of p-value by # of clusters

Create output data shows which branches were assigned to each cluster

Assign branches to dummy variables

#### High dimensionality

#### Quasi-complete separation-target event rate= 0% or 100%

#### Solution

##### Create smarter variables that link to other data sets

##### Collapsing categories based on # of obs in a category (thresholding)

##### Collapsing the categories based on reduction in chi-square test of association bet categorical input and target

##### Using smoothed weight of evidence coding to convert the categorical input into a continuous input

###### PROC SQL, PROC MEANS, Data \_null\_, Data

###### WOE (weight of evidence) typically overfits

###### SWOE (smooth weight of evidence) avoid overfitting-takes into account sampling variability uses adj logodds

## Reducing Redundancy by Clustering Variables

### Negative effect

#### Destabilize the parameter estimates

#### Increase risk of overfitting

#### Confound interpretation

#### Increase computation time

#### Increase scoring effort

#### Increase cost of data collection and augmentation

### Find

#### Correlation matrix

### Methods

#### Variable clustering (id clusters of vars, select a var from each cluster) (PROC VARCLUS-divisive clustering-2nd Eigenvalue, cutoff-maxeigen=n; iterative principal component analysis)

## Performing Variable Screening

### Plot logits by input values

### Variable screening compare Spearman Correlation Stat vs Hoeffding D Statistic (-.5 to 1) (non-monotonic)

### PROC CORR -> PROC SORT -> Data merge -> PROC RANK -> PROC SORT -> PROC PRINT -> PROC SQL -> PROC SGPLOT

### Empirical Logit Plot (continuous vars-binning)

#### PROC RANK -> PROC PRINT -> PROC MEANS -> PROC PRINT -> DATA -> PROC SGPLOT

### Remedies for Nonlinear Relationships

#### Create new variables

#### Use a polynomial model

#### Use a flexible multivariate function estimator

##### Classification tree

##### Generalized additive models

##### Neural networks

#### Do nothing

## Selecting Variables Sequentially-PROC LOGISTIC MODEL response=<effects></options>;

### Options

#### Backward | B

#### Forward | F (detecting interactions)

#### None | N

#### Stepwise | S

#### Score

### Best-subset selection (Score)(rank order based on Chi-Square)(Best for up to 60 vars)

### Stepwise selection (Stepwise)

### Backward selection (Backward)(Fast)

### Significance level

#### BIC

# Measuring Model Performance

## Honest Assessment of the Model

### Data splitting

### Bootstrapping

#### Repeated sampling with replacement -> average assessment stat

### k-fold cross-validation

#### generate n folds -> n validation statistics -> average

## Common Metrics for Model Performance

### Confusion Matrix

#### Accuracy=true pos+true neg/(total # of cases)

#### Error rate=false pos + false neg/(total # of cases)

#### Sensitivity=true pos/total actual pos

#### Positive predicted value=true pos/total predicted positives

#### Specificity=true negatives/total actual negatives

#### Negative predicted value=true negatives/total predicted negatives

### ROC (receiver operating characteristic curve)

#### Sensitivity vs 1-Specificity

#### Gains Chart-PV+ vs Depth (total % of cases allocated to class 1)

##### Good if shape like a steep ski slope

#### Lift chart-PV+/marginal rate

##### Good if shape like a steep ski slope

### Oversampling

#### does not affect sensitivity or specificity

#### does affect PV+ and PV-

## Kolmogorov-Smirnov Statistic (PROC NPAR1WAY-D Statistic-higher better)

### Overall predictive power of a model

### Kolmogorov-Smirnov two sample test

#### Distance between the empirical distribution functions

#### K-S Statistic

##### Shape

##### \*Central tendency (more important in prediction)

##### Variance

### Wilcoxon-Mann Whitney test statistic-more predictive power regarding differences in central tendency)- equal to c statistic

## 5.Model Selection Plots

### Produce model of increasing complexity & choose one that generalizes best

### Compare ROC for multiple models

#### ROC & ROCCONTRAST statements in PROC LOGISTIC

#### Assess & FitandScore macros