SAS® Modeling Best Practices

Using SAS® Enterprise Miner™

Presenter: Melodie Rush, Principal Data Scientist

Q&A: Twanda Baker, Data Scientist

Host: Dean Shaw, Global Webinar Strategist

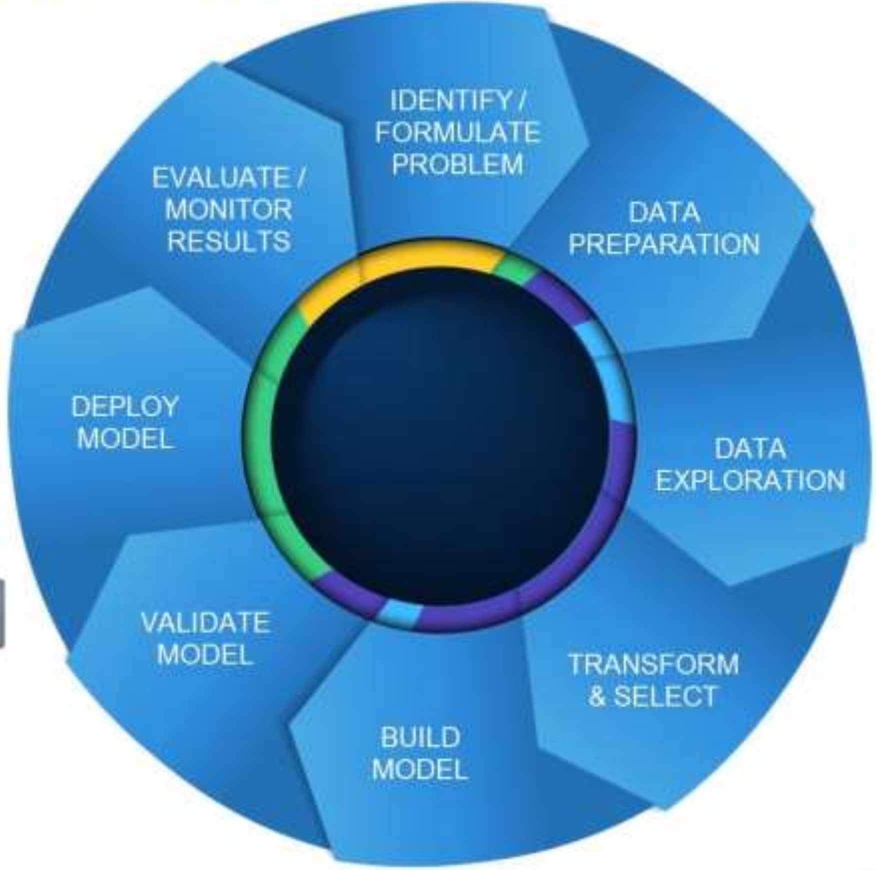
THE PREDICTIVE ANALYTICS LIFECYCLE

BUSINESS MANAGER

Domain Expert
Makes Decisions
Evaluates Processes and ROI

IT SYSTEMS / MANAGEMENT

Model Validation
Model Deployment
Model Monitoring
Data Preparation



BUSINESS ANALYST

Data Exploration
Data Visualization
Report Creation

DATA MINER / STATISTICIAN

Exploratory Analysis
Descriptive Segmentation
Predictive Modeling

SAS® Modeling Best Practices

Agenda

Business Purpose

- Problem definition
- Supervised vs. unsupervised learning

Data Understanding & Preparation

- Best model for available data?
  - Modeling assumptions
  - Objective
  - Target data available?
- Choosing & transforming features

Model Build & Evaluation

- Holdout & test samples
- Statistics

Modeling Best Practices Case Study

Predicting Credit Risk

Scenario: The loan officers of the bank are trying to decide what rate to offer loan applicants.

Available Data:

- 1000 observations (past applicants)
- Information on attributes & behavior of past applicants
  - Ex: property, age, savings
- Label indicating "good" or "bad" candidates
  - Based on loan result (i.e. whether the applicant was able to pay the loan while adhering to the terms of service)
  - 70% good, 30% bad

Considerations:

- Offering a "good" applicant a more favorable rate will result in a 35% profit, while offering a "bad" applicant the same rate will result in a total unit loss

Predicting Credit Risk

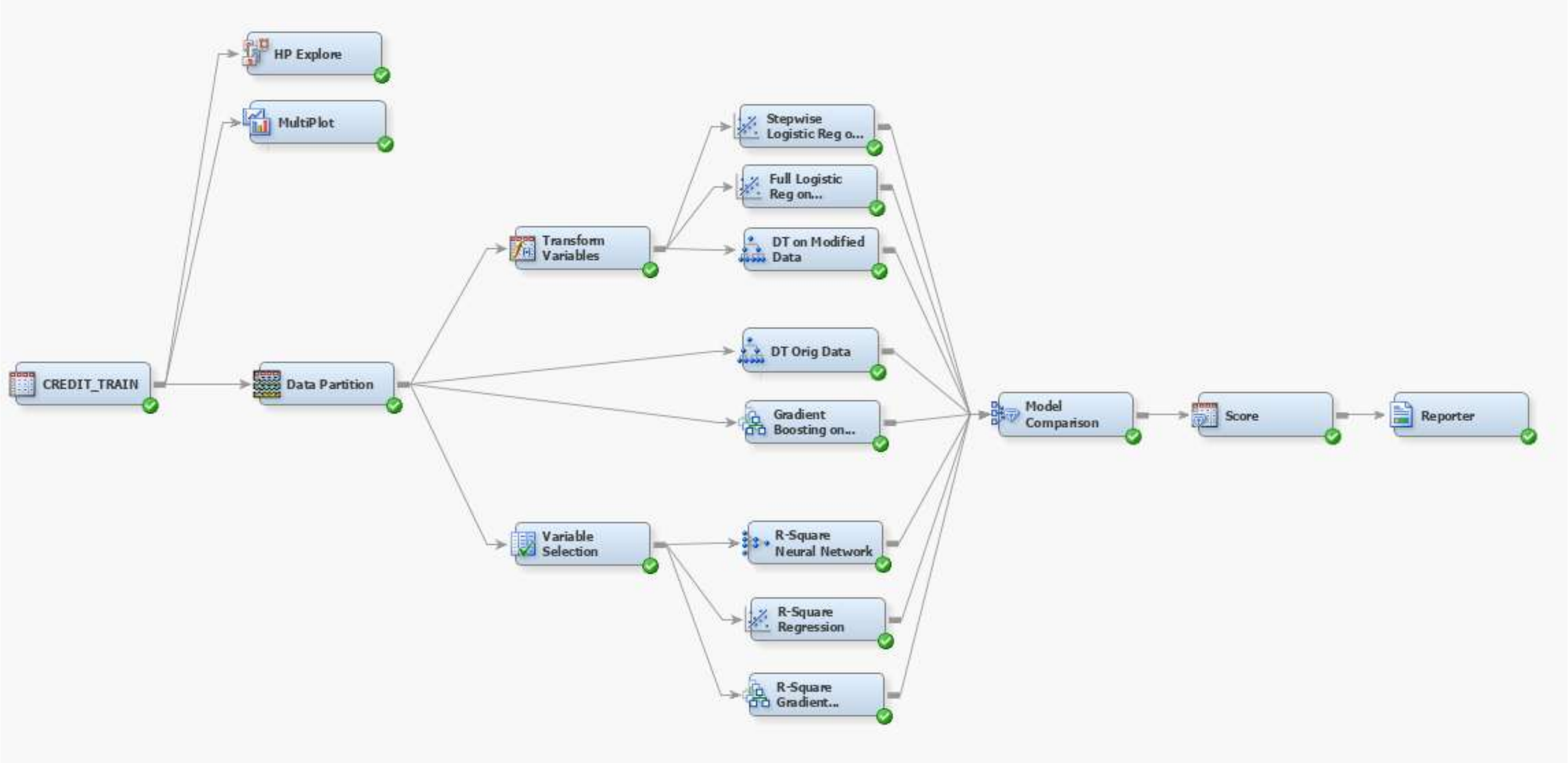
Data Columns

Name	Role	Level
Age_in_years	Input	Interval
Amount_of_credit_in_DM	Input	Interval
Balance_of_current_account	Input	Nominal
Creditability	Target	Binary
Duration_in_months	Input	Interval
Foreign_worker	Input	Binary
Further_debtors_Guarantors	Input	Nominal
Further_running_credits	Input	Nominal
Has_been_employed_by_current_emp	Input	Nominal
Installment_in___of_available_in	Input	Nominal
Living_in_current_household_for	Input	Nominal
Marital_Status_Sex	Input	Nominal
Most_valuable_available_assets	Input	Nominal
NewVar	Input	Interval
Number_of_persons_entitled_to_ma	Input	Ordinal
Number_of_previous_credits_at_th	Input	Ordinal
Occupation	Input	Nominal
Payment_of_previous_credits	Input	Nominal
Purpose_of_credit	Input	Nominal
Telephone	Input	Binary
Type_of_apartment	Input	Nominal
Value_of_savings_or_stocks	Input	Nominal

Predicting Credit Risk

Data

	Creditability	Occupation	Telephone	NewVar	Balance of current account	Duration in mon...	Payment of previous credits	Purpose of credit	Amount of credit in DM	Value c
1	bad	skilled worker/skilled employee/minor ci...	yes	2.960618...	no running account	36.0	no problems with current credits at thi...	retraining	2145.0	no savings
2	good	executive/self-employed/higher civil se...	yes	2.879810...	no balance	48.0	hesistant payment of previous credits	retraining	12204.0	greater
3	bad	executive/self-employed/higher civil se...	yes	2.592734...	>=200 DM	36.0	no previous credits or paid back	used car	10974.0	no savings
4	good	executive/self-employed/higher civil se...	yes	2.576882...	no running account	24.0	paid back previous credits at this bank	new car	6419.0	no saving
5	good	skilled worker/skilled employee/minor ci...	yes	2.527170...	>=200 DM	24.0	no previous credits or paid back	retraining	1258.0	no savin
6	good	unskilled with permanant residence	no	2.502577...	no balance	12.0	no previous credits or paid back	retraining	1037.0	less than 10
7	bad	unskilled with permanant residence	no	2.439117...	no running account	30.0	no previous credits or paid back	used car	3108.0	no savir
8	good	skilled worker/skilled employee/minor ci...	yes	2.375307...	no balance	15.0	paid back previous credits at this bank	items of furniture	1537.0	greater th
9	good	skilled worker/skilled employee/minor ci...	yes	2.182903...	>=200 DM	15.0	paid back previous credits at this bank	items of furniture	1471.0	no savings
10	bad	unskilled with permanant residence	no	2.156286...	no balance	27.0	paid back previous credits at this bank	items of furniture	2520.0	betweer
11	bad	skilled worker/skilled employee/minor ci...	yes	2.047069...	no balance	24.0	no previous credits or paid back	used car	4057.0	no savings
12	bad	unemployed/unskilled with no permaan...	no	2.004313...	no running account	18.0	no previous credits or paid back	repair	750.0	no savin
13	bad	executive/self-employed/higher civil se...	yes	2.003039...	no balance	36.0	no problems with current credits at thi...	retraining	4455.0	no saving
14	good	unskilled with permanant residence	no	1.990002...	>=200 DM	6.0	no problems with current credits at thi...	retraining	1743.0	less than 1
15	good	skilled worker/skilled employee/minor ci...	yes	1.989741...	no running account	12.0	no previous credits or paid back	other	1893.0	no savi
16	good	skilled worker/skilled employee/minor ci...	yes	1.939072...	>=200 DM	42.0	no previous credits or paid back	items of furniture	7166.0	greater tha
17	good	skilled worker/skilled employee/minor ci...	no	1.935472...	no running account	48.0	no previous credits or paid back	new car	4788.0	no saving
18	bad	unskilled with permanant residence	no	1.919844...	no balance	24.0	problematic running accounts	repair	1837.0	no saving
19	good	skilled worker/skilled employee/minor ci...	yes	1.90445[illegible]71	no balance	24.0	paid back previous credits at this bank	other	3878.0	less than 10


HP Explore
MultiPlot
CREDIT_TRAIN
Data Partition
Transform Variables
Variable Selection
Stepwise Logistic Reg o...
Full Logistic Reg on...
DT on Modified Data
DT Orig Data
Gradient Boosting on...
R-Square Neural Network
R-Square Regression
R-Square Gradient...
Model Comparison
Score
Reporter

Business Purpose

Understanding your Objective

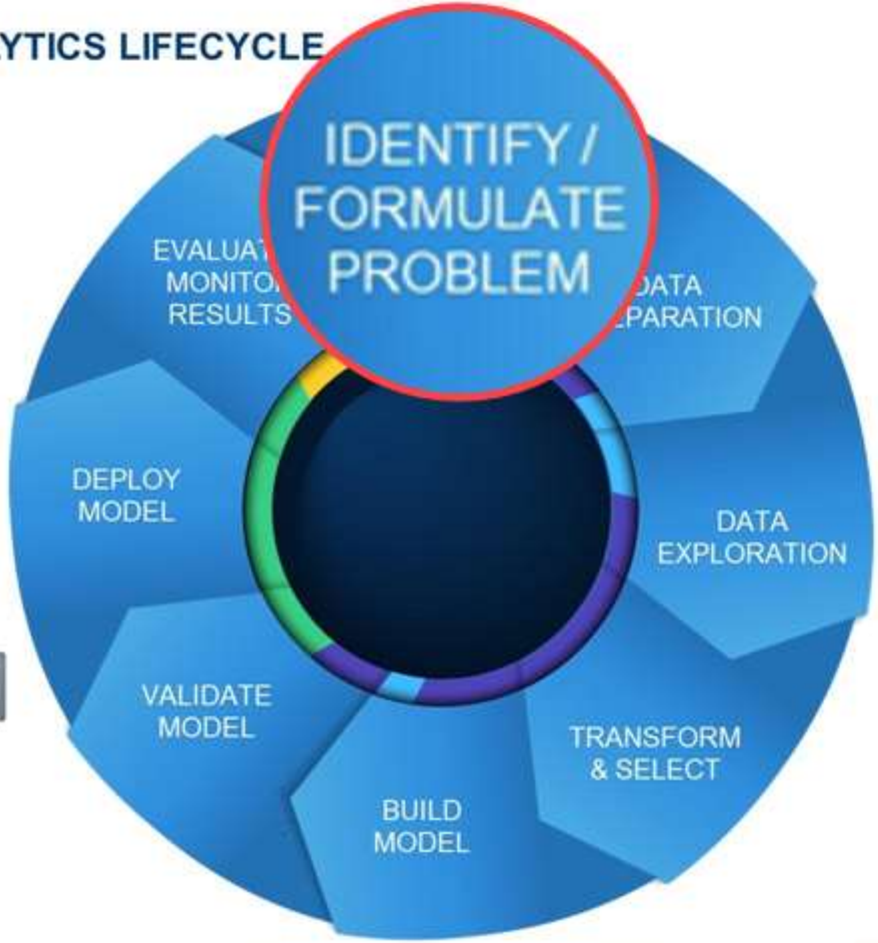
THE PREDICTIVE ANALYTICS LIFECYCLE

BUSINESS MANAGER

Domain Expert
Makes Decisions
Evaluates Processes and ROI

IT SYSTEMS / MANAGEMENT

Model Validation
Model Deployment
Model Monitoring
Data Preparation



BUSINESS ANALYST

Data Exploration
Data Visualization
Report Creation

DATA MINER / STATISTICIAN

Exploratory Analysis
Descriptive Segmentation
Predictive Modeling

Business Purpose

Common Questions to Ask

What is your objective about?

What are you trying to learn?

Are we trying to predict, classify, or describe something?

Business Purpose

Problem Definition

Business Purpose	Problem Definition
Clients	• People or groups who benefit from the outcomes of the models
Objectives	• Goals to be achieved that serve the interests of the clients
Criteria	• Measures of success or failure
Decision Makers	• People who influence the achievement of objectives

Business Purpose

Problem Definition

Resources

- Available time, labor, capital for development & deployment of model

Constraints

- Limitations

Critical Assumptions

- Implicit & explicit assumptions about the world or industry in which the model/project is being developed

Modeling Best Practices Case Study

Predicting Credit Risk

Clients
- Loan officers
- Bank institution/organization

Objective(s)
- Accurate classification of applicant
- Maximum average profit

Criteria
- Initial: model evaluation statistics
- Long term: profit derived from results

Decision Makers
- Chief Financial Officer (CFO)
- Chief Information Officer (CIO)

Resources
- IT available to deploy model?
- Type of system used
- Analysts & programmers available to create and maintain model?

Constraints
- Data limitations (availability, amount of observations)
- System limitations

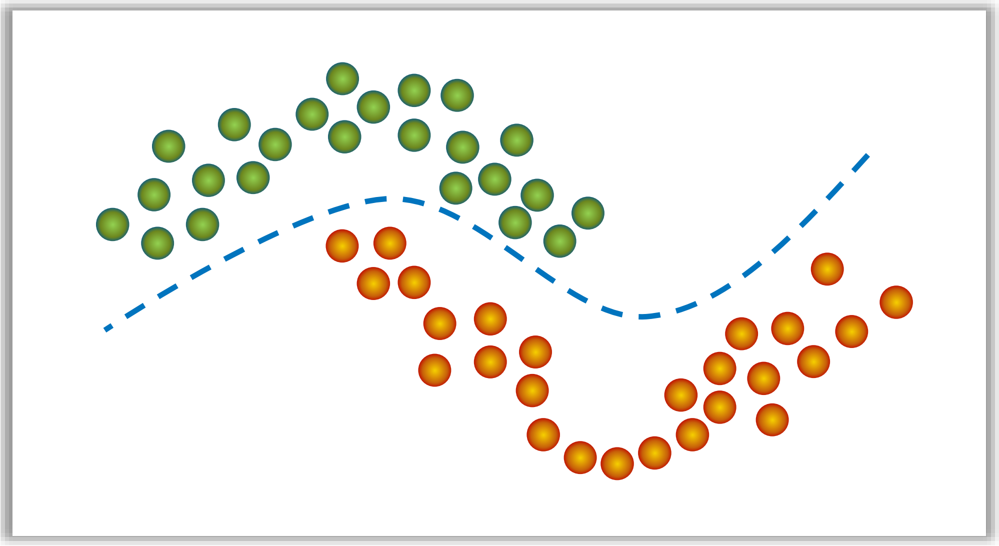
Critical Assumptions
- Higher interest rate for riskier applicants prevents large loss of $
- Reward less risky applicants with more lower interest rates in effort to attract more favorable business

Model Objective → Supervised vs. Unsupervised Learning?

Types of Learning

Supervised Learning

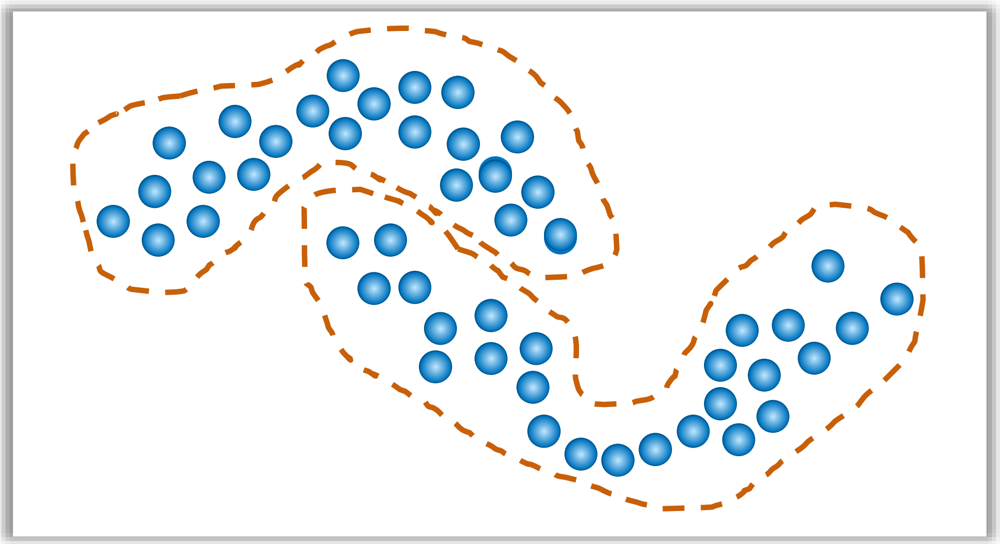
Trained on labeled examples



Types of Learning

Unsupervised Learning

Trained on unlabeled examples



	Supervised Learning	Unsupervised Learning
Common Questions Answered	• How much will prospect *x* spend? • Will customer *x* default on her loan?	• What items are commonly purchased together? • What other companies are like our best small business customers? • What does normal behavior look like? • Do my customers form natural groups?
Techniques	• Involves classification or regression • Random forests • Decision trees • Neural networks* • Linear regression • Logistic regression • Support vector machines • k-NN (k-nearest neighbors) • Gradient boosting • Ensembles	• Clustering (by observation or variable) • Anomaly detection • Principal component analysis (PCA) • Singular value decomposition (SVD) • Expectation-maximization algorithm • Multivariate analysis

**Can be used as an unsupervised learning technique as well*

Data Understanding

Choosing the Best Technique

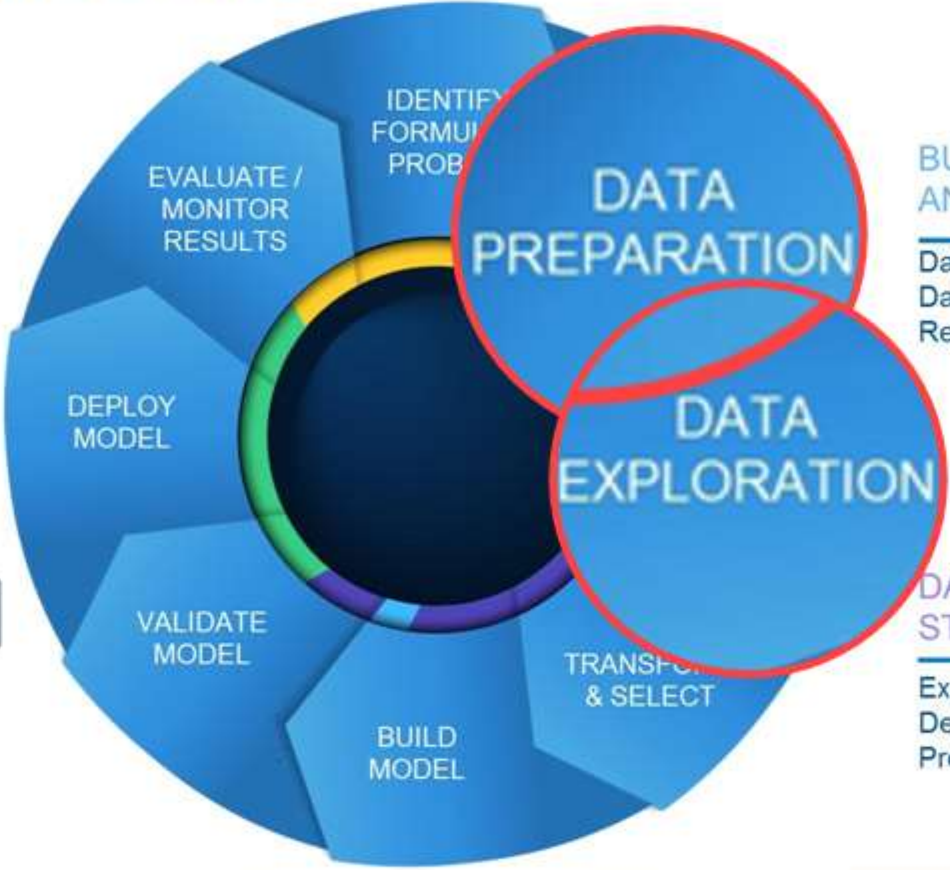
THE PREDICTIVE ANALYTICS LIFECYCLE

BUSINESS MANAGER

Domain Expert
Makes Decisions
Evaluates Processes and ROI

IT SYSTEMS / MANAGEMENT

Model Validation
Model Deployment
Model Monitoring
Data Preparation



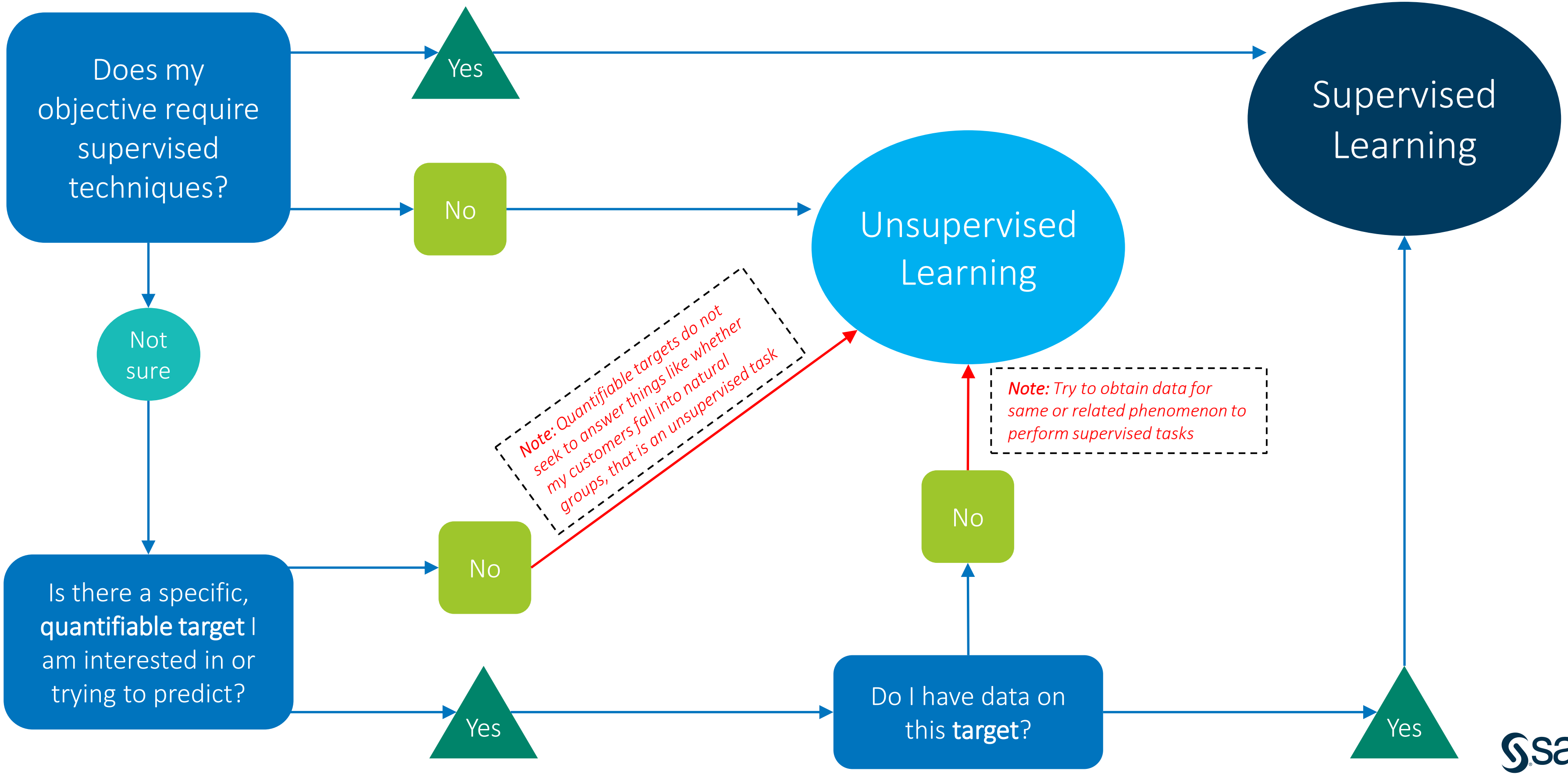
BUSINESS ANALYST

Data Exploration
Data Visualization
Report Creation

DATA MINER / STATISTICIAN

Exploratory Analysis
Descriptive Segmentation
Predictive Modeling

Data Understanding



Modeling Best Practices Case Study

Predicting Credit Risk

Is there a specific, ***quantifiable*** target I am interested in or trying to predict?

- Yes → we are interested in the credit risk of loan applicants (either “bad” or “good”)

What question am I seeking to answer?

- Does applicant *X* have “good” or “bad” credit?

Do I have ***data*** on the target?

- Yes → 1000 observations of previous applicants & resulting behavior (labels that indicate as “good” or “bad”)

Supervised Learning

Supervised Learning Techniques

Data Understanding

Supervised Learning → Classification or Regression?

- Classification → categorical target
  - Target has discrete, NON-ordinal values
  - Most common case = binary classification
  - Probability estimation or ranking
    - Exception wherein classification model predicts continuous values such as probabilities or ranks/scores
    - Probability estimation → model predicts a score b/w 0 & 1 for each available class
      - Use: cost or benefit is known relatively precisely & may not be constant across instances
    - Ranking → model predicts a score wherein a higher score indicates higher likelihood of being in given class (in case of binary classification)
      - Use: cost or benefit is constant across instances & is unknown or difficult to calculate
- Regression → numeric target

Data Understanding

Decision Tree vs. Linear Models

- Questions to Consider:
  - *What is more comprehensible to stakeholders? Rules or a numeric function?*
  - *How "smooth" is the underlying phenomenon being modeled?*
  - *How "non-linear" is the underlying phenomenon being modeled?*
  - *How much data do you have?*
  - *What are the characteristics of the data?*

Data Understanding
When to apply Machine Learning ?

- Questions to Consider:
  - *How large is your data set?*
    - *Speaks to scalability → may be easy to classify a few hundred emails as spam or not but this problem becomes more tedious & difficult as the size of the emails increases to the millions*
  - *How easily can you outline the underlying phenomenon?*
    - *Large # of factors could influence answer to specific classification or prediction problem*
    - *Rules overlap or need to be finely tuned*
    - *Ex: classify email as spam or not*
      - *What constitutes spam?*
      - *What affects whether an email is spam?*
      - *Is this specific to the person or organization?*

Unsupervised Learning Techniques

Data Understanding

Clustering

PCA or EFA

Descriptive Statistics & Dimension Reduction

Unsupervised learning techniques can be used in conjunction with supervised techniques in an effort to improve model performance. Additionally, it can be used on it's own when there is a lack of target data.

- Observation or Variable Clustering
  - Obs. clustering provides description of data (ex: do your consumers fall naturally in to specific groups? → regionally, financially, etc.)
  - Variable clustering reduces # of variables for use in supervised modeling technique → improves performance by minimizing modeling complexity
- Additional dimension reduction techniques

Data Understanding

Descriptive Statistics & Dimension Reduction

Unsupervised learning techniques can be used in conjunction with supervised techniques in an effort to improve model performance. Additionally, it can be used on it's own when there is a lack of target data.

Anomaly Detection

- Identifies items, events or observations which do not conform to an expected pattern or other items in dataset → descriptive

Multivariate Data Analysis

- Analyzing data from more than one variable
- ANOVA or MANOVA
  - ANOVA tests for difference in means b/w 2 or more groups
  - MANOVA tests for difference in 2 or more vectors of means

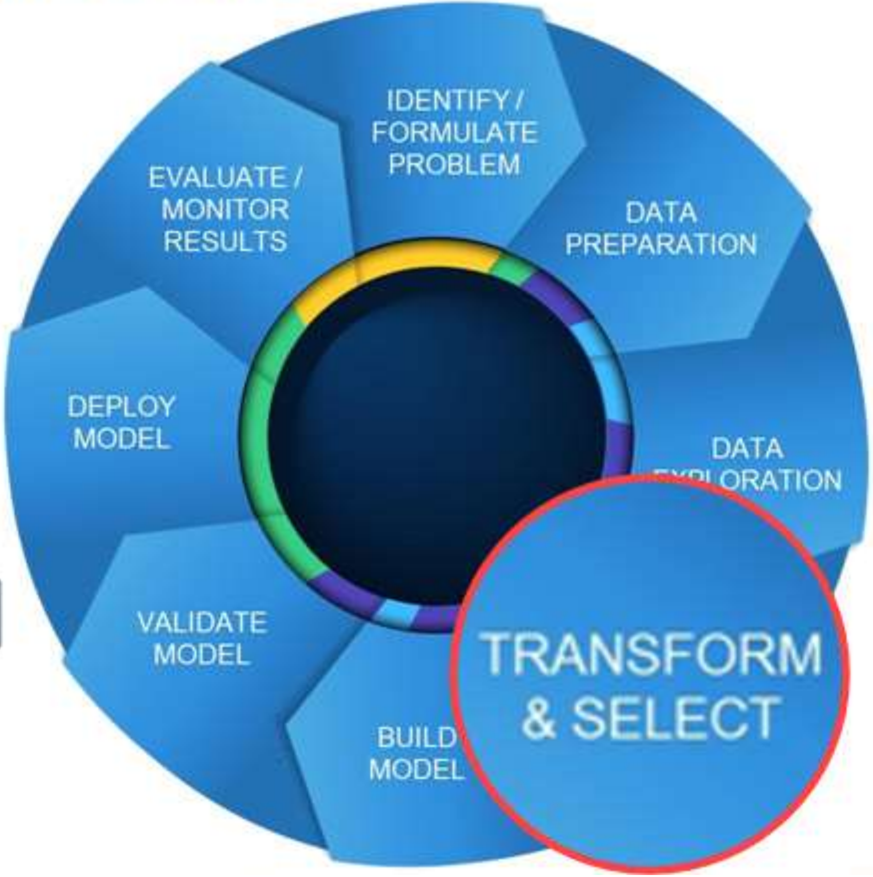
THE PREDICTIVE ANALYTICS LIFECYCLE

BUSINESS MANAGER

Domain Expert
Makes Decisions
Evaluates Processes and ROI

IT SYSTEMS / MANAGEMENT

Model Validation
Model Deployment
Model Monitoring
Data Preparation



BUSINESS ANALYST

Data Exploration
Data Visualization
Report Creation

DATA MINER / STATISTICIAN

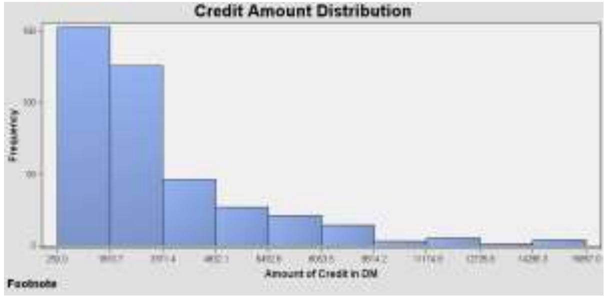
Exploratory Analysis
Descriptive Segmentation
Predictive Modeling

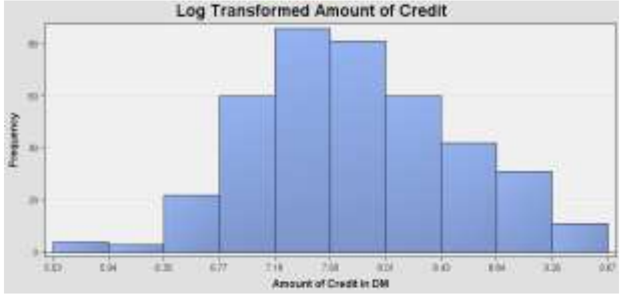
Transforming & Selecting Variables

- Reasons to transform
  - Force variable distribution to be normal
  - Standardize all inputs to make sure all are on same scale
  - Remove bias
- Methods
  - Nominal → dummy indicators, group rare levels
  - Interval → bucket, center, equalize, exponential, inverse, log, optimal binning, quantile, square, square root, standardize (normalize)

Transforming & Selecting Variables

- Transform variables
  - Modeling assumptions → for models such a linear regression, there are certain assumptions that need to be met to ensure the accuracy of the model
    - Linearity, normality, heteroscedasticity
    - Adherence to assumptions looser for logistic regression
  - Normality assumes all inputs have normal distribution (skewed distribution can be normalized by applying log transformation, exponential)



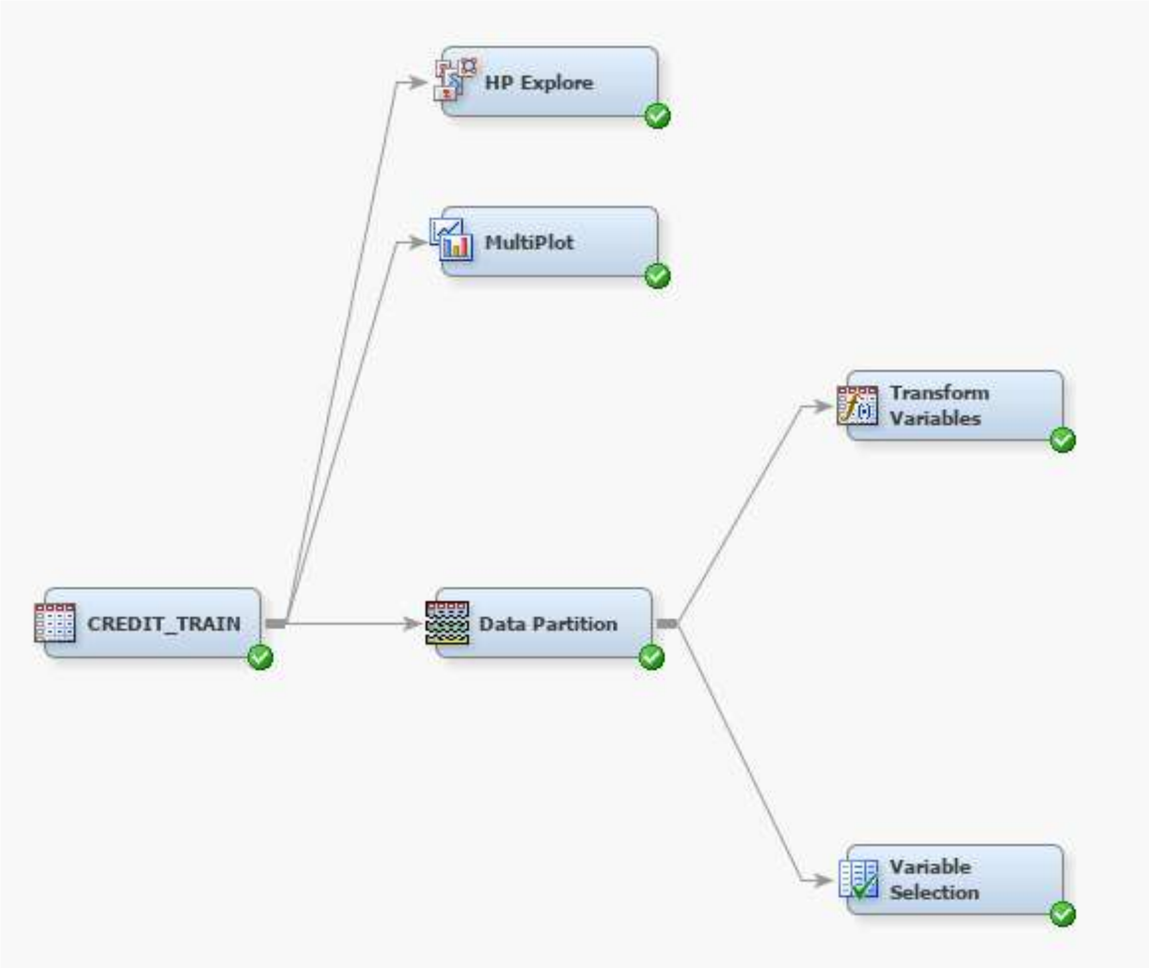


Transforming & Selecting Variables

- Selecting variables
  - Many modeling methods choose inputs as a part of the building process
  - Linear or logistic regression employs stepwise, backward or forward selection (can also choose to just include all available inputs)
  - Tree models
    - Decision → builds tree based on variable importance
    - Random forests → builds multiple trees, each with different sampling of observations & inputs
  - Prior to applying model
    - Chi square or R square
    - LASSO or LAR
    - Unsupervised → correlation, covariance, sum of squares or cross product

Predicting Credit Risk Case Study

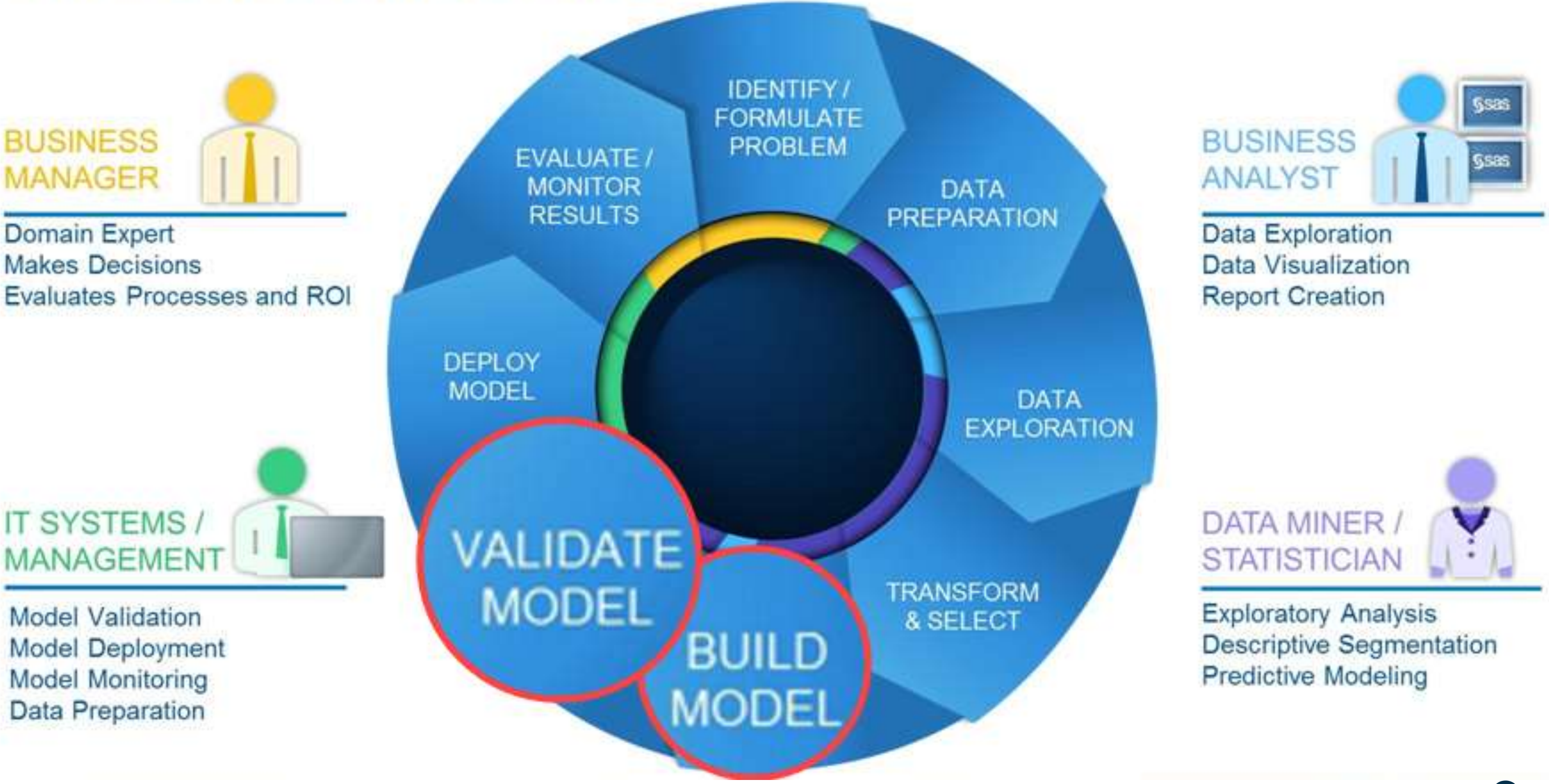
Applying Techniques in Enterprise Miner™


HP Explore
MultiPlot
Transform Variables
CREDIT_TRAIN
Data Partition
Variable Selection

Modeling & Evaluation

Meeting your Objective

THE PREDICTIVE ANALYTICS LIFECYCLE



Modeling & Evaluation

Measuring Accuracy

- Partition data
  - Train, validate & test (holdout) samples
  - Validate normally used to choose model (technique, features, complexity parameters) while test confirms accuracy
  - 40-30-30 split is default
  - Additional technique → cross validation
    - Randomly partition data into *k* folds, run training/test evaluation *k* times
  - Be aware of overfitting or underfitting
    - Validation set helps to prevent overfitting
    - Overfitting → model fits data well but is not generalizable

Modeling & Evaluation
Measuring Accuracy

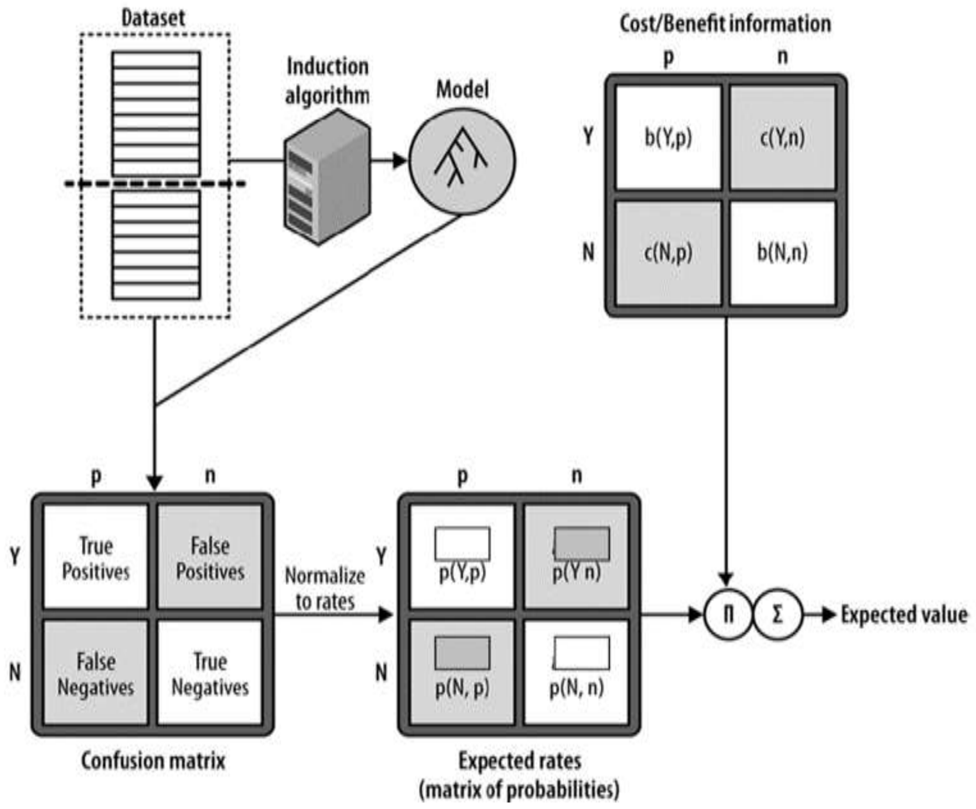
- Fit statistics
  - Depends on many factors including objective & available information
  - Regression → Average square error
  - Classification
    - Misclassification/error rate = percentage of incorrect classifications
    - Confusion matrix
      - True positive rate (sensitivity or recall) = $\frac{a}{a+c}$
      - True negative rate (specificity) = $\frac{d}{b+d}$
      - Positive predictive value (precision) = $\frac{a}{a+b}$

		Actual	
		+	-
Predicted	Y	a	b
	N	c	d

Classifier Evaluation

Business Costs & Benefits

- Taking into account business objective
- Example:
  - Objective → maximize profit
  - Target → binary, yes or no
  - Need to combine accurate classification with profit & losses

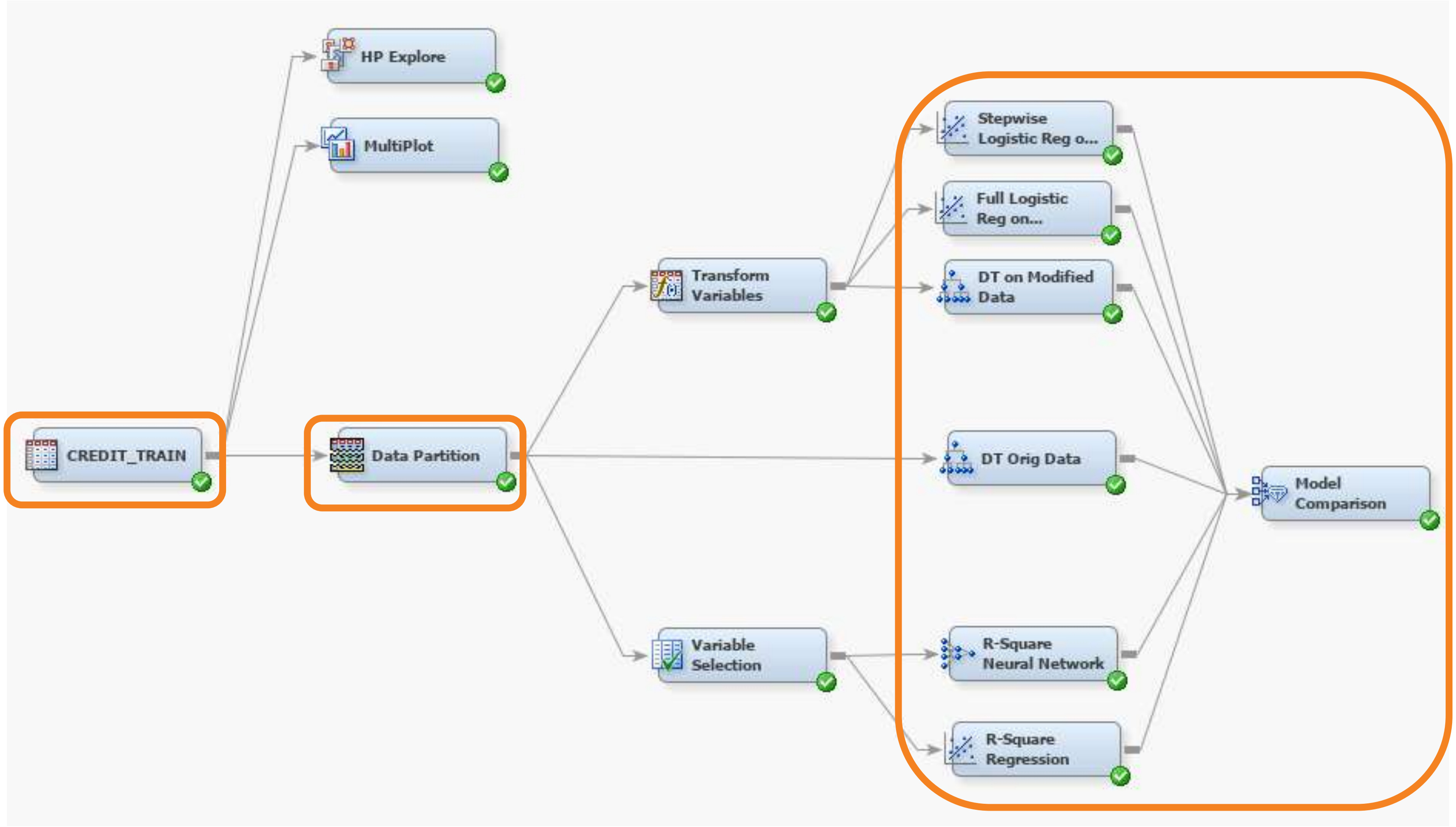


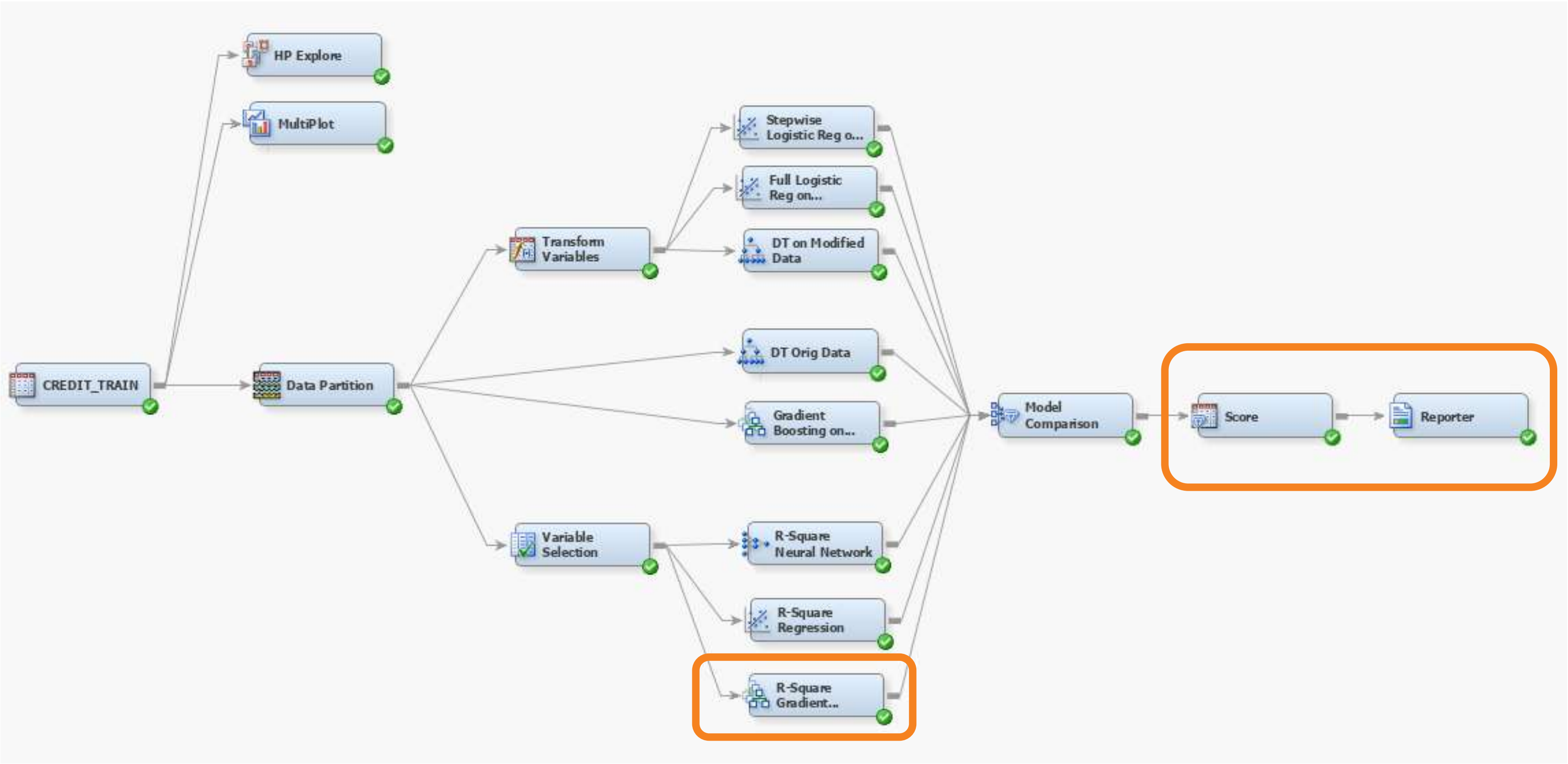
Modeling & Evaluation
Measuring Accuracy

- Visual evaluation
  - Works for both classification & regression models
    - ROC chart, AUC (area under ROC curve)
      - For classifier, gives probability that model will rank a positive case higher than negative case
      - Fair measure of quality of probability estimates
    - Lift chart
      - Measures effectiveness of predictive model calculated as ratio b/w results obtained w/ & w/o predictive model

Predicting Credit Risk Case Study

Applying Techniques in Enterprise Miner™


HP Explore
MultiPlot
CREDIT_TRAIN
Data Partition
Transform Variables
Variable Selection
Stepwise Logistic Reg o...
Full Logistic Reg on...
DT on Modified Data
DT Orig Data
R-Square Neural Network
R-Square Regression
Model Comparison


HP Explore
MultiPlot
CREDIT_TRAIN
Data Partition
Transform Variables
Stepwise Logistic Reg o...
Full Logistic Reg on...
DT on Modified Data
DT Orig Data
Gradient Boosting on...
Variable Selection
R-Square Neural Network
R-Square Regression
R-Square Gradient...
Model Comparison
Score
Reporter

Summary

- Business objective & available data is key to choosing the best model
- Modeling is cyclical
  - Questions to consider along the way are helpful in determining what methodologies to apply but you may have to make changes or tweak things along the way as you learn more about your data & the underlying phenomenon
- Try multiple methodologies to obtain the best possible model
  - Enterprise Miner™ is especially good for this (can easily evaluate multiple models at once)
  - EM™ is also good at making quick changes that will affect the rest of the process

Resources

Where to learn more

Ready to Get on the Fast Track with Enterprise Miner?

Visit sas.com/learn-em

and sign up to receive EM technical resources, tips & tricks delivered directly from Brett Wujek, Sr. Data Scientist from SAS R&D

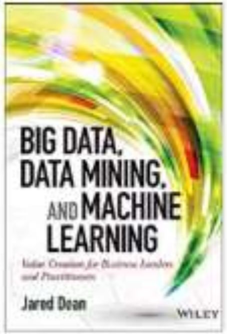
Further Reading

Papers

- Identifying and Overcoming Common Data Mining Mistakes by Doug Wielenga, SAS Institute Inc., Cary, NC
- Best Practices for Managing Predictive Models in a Production Environment by Robert Chu, David Duling, Wayne Thompson , SAS Institute Cary, NC
- From Soup to Nuts: Practices in Data Management for Analytical Performance by David Duling, Howard Plemmons, Nancy Rausch, SAS Institute Cary, NC

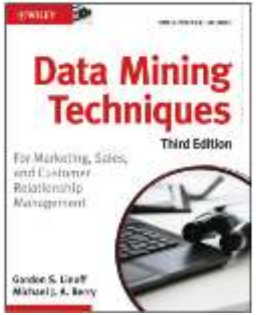
- (All available on support.sas.com)

Resources
Suggested Reading



Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners
By Jared Dean

Available on Amazon

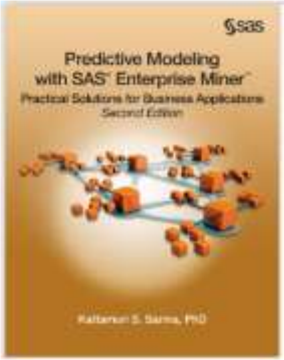


Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management
by Gordon S. Linoff and Michael J. A. Berry

Available on Amazon

Resources

Suggested Reading



Predictive Modeling with SAS Enterprise Miner: Practical Solutions for Business Applications, Second Edition, Edition 2

By Kattamuri S. Sarma, PhD

Available on [Amazon]

Applied Analytics Using SAS Enterprise Miner

By: SAS

Available on [Amazon]

Questions?

Thank you for your time and attention!

Connect with me:

LinkedIn: https://www.linkedin.com/in/melodierush

Twitter: @Melodie_Rush

sas.com