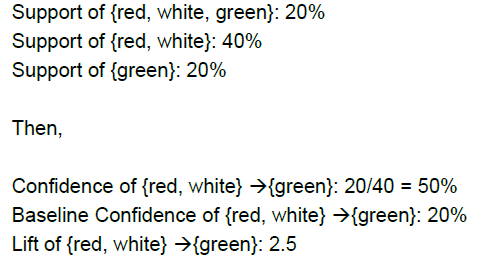
PA Study Guide

**Overview of Data Analytics:**

* **Official data analytics definition:** 
  + Iterative process of creating predictive and descriptive models, by uncovering previously unknown trends and patterns in a vast amount of data, in order to support decision making
  + Extracts useful information and knowledge from large volumes of data by following a process with reasonably well-defined steps
* **Sun definition: DIDA**
  + **D –** use **Data** to extractuseful **I – Information** to make better business **D- Decision**s in order to gain a competitive  **A – Advantage**
* **Analytics Steps:**
  + **Business Understanding –** understand or create the business problem
  + **Data Understanding –** Understand the strength and limitation of data
  + **Data Preparation –** Convert the data into the right format
  + **Modeling –** Apply data mining techniques to the data
  + **Evaluation –** assess the performance of the data mining results
  + **Deployment –** Put into real use to realize the ROI
* **Supervised Learning: classification and prediction**
  + Goal is to predict a single outcome known as the target
  + Training our data means to use historical data where target is already known to create, test and train model
  + Scoring data is next step which applies training data model in order to predict future or new results
  + Key property of supervised learning is that **historical data (input) MUST be present and already have known outcome** used in creating model ex must have past customer data (input) AND their purchases (outcome)
  + **Classification –** predict categorical target variable such as purchase or no purchase
    - target is often binary
  + **Prediction –** predict numerical target variable such as sales, revenue, and performance
* **Unsupervised Learning: association rules, variable reduction, and clustering**
  + Goal is to segment data into meaningful groups and detect patterns
  + Key property is that there **is no known target outcome variable is present** to predict or classify
  + **Association Rules (also affinity analysis and market basket analysis)**
    - Goal is to produce rules that define “if X was purchased, Y was also purchased”
  + **Data Reduction –** reducing number of variables with goal of making data set more concise, accurate and easier to work with, usually by clustering and targeting which groups will maximize outcome
* Inference is about explaining the data and answering the “why” whereas prediction is to directly impact metrics such as increase sales and revenue

**Association Rules: the study of what goes with what**

* “IF” = antecedent
* “THEN” = consequent
* Antecedent and consequent must be disjoint to perform analysis of association rules
* **Support** – percent of transactions that include both antecedent and consequent of interest that actually appear in given data set
  + Calculated by number of transactions containing both divided by total number of transactions
    - P(A and B)
  + **Ex** 10% support of red and blue marbles means that out of the entire sets of pairings, red and blue only appear together 10% of the time
* **Measure of performance: Confidence and Lift**
  + **Confidence –** Probability that the customer has item B given that they have item A
    - Calculated by number of transactions w/both antecedent and consequent item sets *divided by* number of transactions with antecedent sets
      * P(B given A) = P(A and B) / P(A)
  + **Benchmark Confidence** is consequent over all transactions
    - P(B)
  + **Lift –** Measure of strength of the association as compared to a random transaction
    - Calculated by confidence divided by expected confidence
      * P(B given A) / P(B)

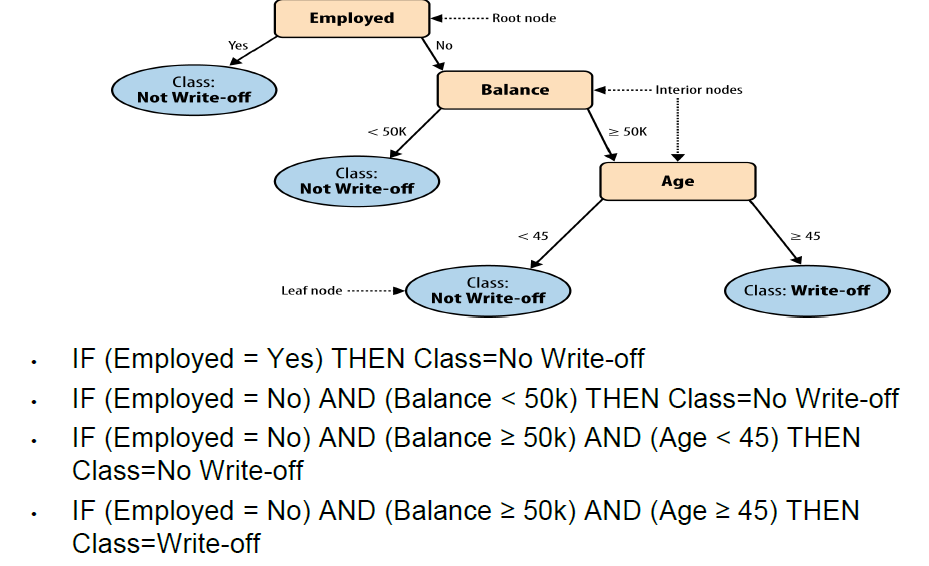


**Data Preparation and Pre-Processing**

* **Dependent variable (y-axis) –** what you want to predict, in SAS called Target variable
* **Independent variable (x-**axis) – variables that are relevant to the dependent variable, in SAS called predictors or input variables
* **Numerical Variable (interval in SAS) –** variable that assumes numeric value such as price
* **Categorical Variable –** limited and fixed number of values, usually the name of something such as fuel type, blood type, flight status
  + **Nominal –** usually the name of a variable with no particular order such as color
  + **Binary –** only two distinct values such as yes/no, on-time/delay, purchase/no purchase
* **Dummy Indicators –** assigning numerical values (usually binary) to categorical data
  + Number of dummies = number of categories -1
  + Ex delayed or not delayed is 2 categories so you only need 1 dummy variable
* **Rule of thumb number for making predictions:** n = 10 x *u*
  + **N = number of observations**
  + ***U*= number of predictors**
    - **Ex.** Trying to predict Price with age (1), horse power (2), mileage (3)  
      n = 10 x 3 = 30 observations recommended
* **Rule of thumb number for classification:** n = 6 x *m* x (*u* + 1)
  + ***M* = number of classes in the dependent variable (such as yes/no m=2)**
  + ***U* = number of predictors**
* **Normalizing Data** – used when variables are on different scales and will skew results
  + **Technique 1**: each observation minus mean and divide by standard deviation
  + **Technique 2**: scale 0-1 by subtracting minimum value and dividing by the range
    - most used when data contains both numeric and dummy variables
* **Missing Values**
  + Most algorithms default is to drop records with missing values
  + **Rule of thumb: do not drop more than 10% of records**
    - Recall a record is a full row
  + Can use regression to fill missing values as well, but run the risk of overfitting model
* **Imputation:** replacing missing values with reasonable substitutes, usually mean or median
* **Partitioning the Data –** separating data set into training and validation sets
  + **Training –** used to develop model
  + **Validation –** used to implement model and evaluate performance with new data
    - compares new predicted values with historical outcomes
    - compares accuracy and error rates to already known past results
  + Worded differently: the model is fit on the training data set and the model performance is evaluated on the validation data set
* **Overfitting –** incorporating outliers into model in order to account of every data point AVOID
  + only use typical data points and do not attempt to account of every single point
  + leads to less accurate model
* **Accuracy on validation data set is most important before deploying a model**

**Decision Trees**

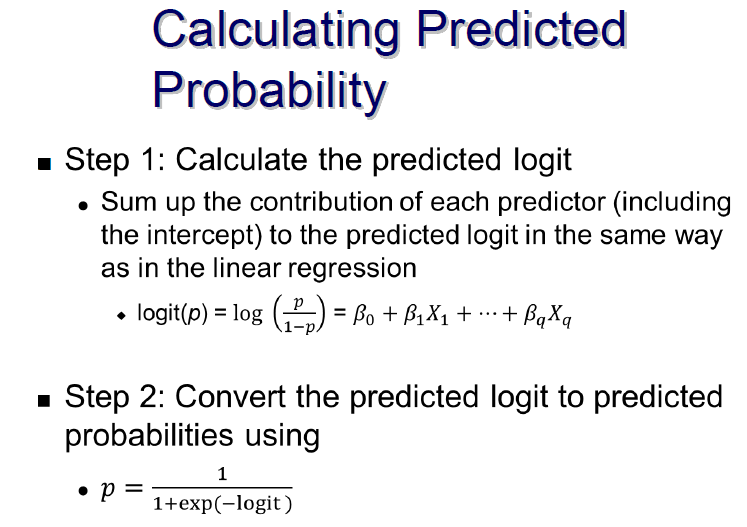
* **Goal** is to classify or predict an outcome based on a set of predictors
* **Process:** 
  + **Recursive partitioning –** repeatedly splitting the records into two parts to achieve maximum homogeneity within new parts
  + **Pruning the tree –** simplify the tree to avoid overfitting
* Determine where splits occur by maximum log worth
  + Logworth = -log(chi-square p-value)
  + If values are very different, chi-square test statistic will be very high and p-value will be low (significant) resulting in a high logworth score
* **Tree Structure:**
  + **Node –** split point with value in the center
  + **Leaves –** at the end of a split with no other node coming out of it
  + **Cases –** number of lines between nodes



* **Maximal tree –** the tree with the maximum number of splits
* **Optimal tree –** the tree with the most concise, fewest splits after pruning leading to most efficient tree with least amount of error before it begins to rise

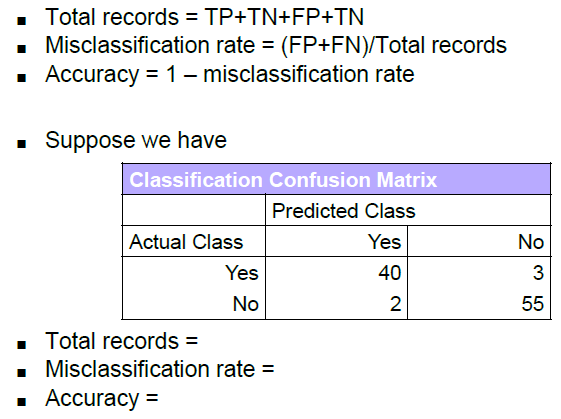
**Logistic Regression**

* **Goal is to find a function of the predictor variables that relates them to a 0/1 outcome**
  + Logistic regression extends to categorical outcomes
  + Unlike linear regression that could be any value, logistic is standardized and bound
  + Uses concepts of odds and probability
* **Probability vs Odds**
  + **Probability –** chance of an event occurring compared to total outcomes
    - ex a horse winning a race P
    - inversely the probability of an event not occurring is 1-P
  + **Odds –** chance of winning over chance of losing
    - **P(loss)/P(win) = P/(1-P)**
  + **Example:**
    - P(w) Probability to win = 1/20 = 5%
    - P(l) Probability to lose = 19/20 = 95% or 1-P(w)
    - Odds for this is P(w)/P(l) = **P/1-P** = 1/19
    - This translates to being shown as 19:1 odds meaning that the horse has 19 chances to loses for every 1 change to win
    - To move back from odds to probability = **Odds/1+odds** = (1/19)/(1+1/19) = (1/19)/(20/19) = 5%
    - Odds are 1 when probability is 50%
* **Odds Ratio:** 
  + the amount odds change with a unit change in an input variable
  + ex: if an odds ratio is 1.059 this means that for each additional unit increase in the input, the odds change by 5.9% increase



**Logistic Regression Part 2**

* **Reminder –** the best model is the one that can maximize predictive accuracy on the VALIDATION data set
* **Classification rule –** threshold percent probability in log regression to classify outcome as 1 or 0, yes or no. Usually binary and 50% threshold
  + links predicted probability w/predicted binary outcome
* **Confusion Matrix –** performance measure for classification
  + **Type 1 error –** false positive
  + **Type 2 error –** false negative
* **Example:**

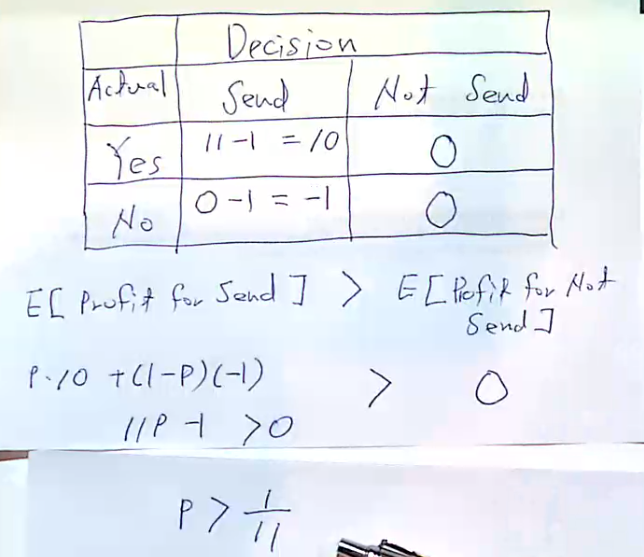


Total records = 40 + 2 + 3 + 55 = 100

Misclassification rate = 3 + 2 = 5/100 = 5%

Accuracy = 1 - 5% = 95%

* **Decision Rule is not the same as classification rule and % changes for asymmetry**



Decision Rule: P > 1/11 = 9.1% meaning that if probability of acceptance is greater than 9.1% then send, else do not send

**Neural Networks**

* **Neural Networks** are a set of connected Input/Output units where each connection has a weight associated with it
* The process is able to “learn” and adjust input weights each time it is run and tested to correctly classify training data
  + Has high tolerance for noise and incomplete data
  + Able to capture complex relationships where other models might fail
  + No need to specify curve or fit
  + “black box” in nature between the input and output, calculations are complex
* General process
  + Input 🡪 random assigning weights to variables 🡪 hidden layer where input variables overlap are developed 🡪 output layer with prediction 🡪 repeat to reduce error
* **Structure:**
  + Multiple layers – input, hidden, output
  + Nodes
  + Weights – coefficients
  + Bias values – intercepts

