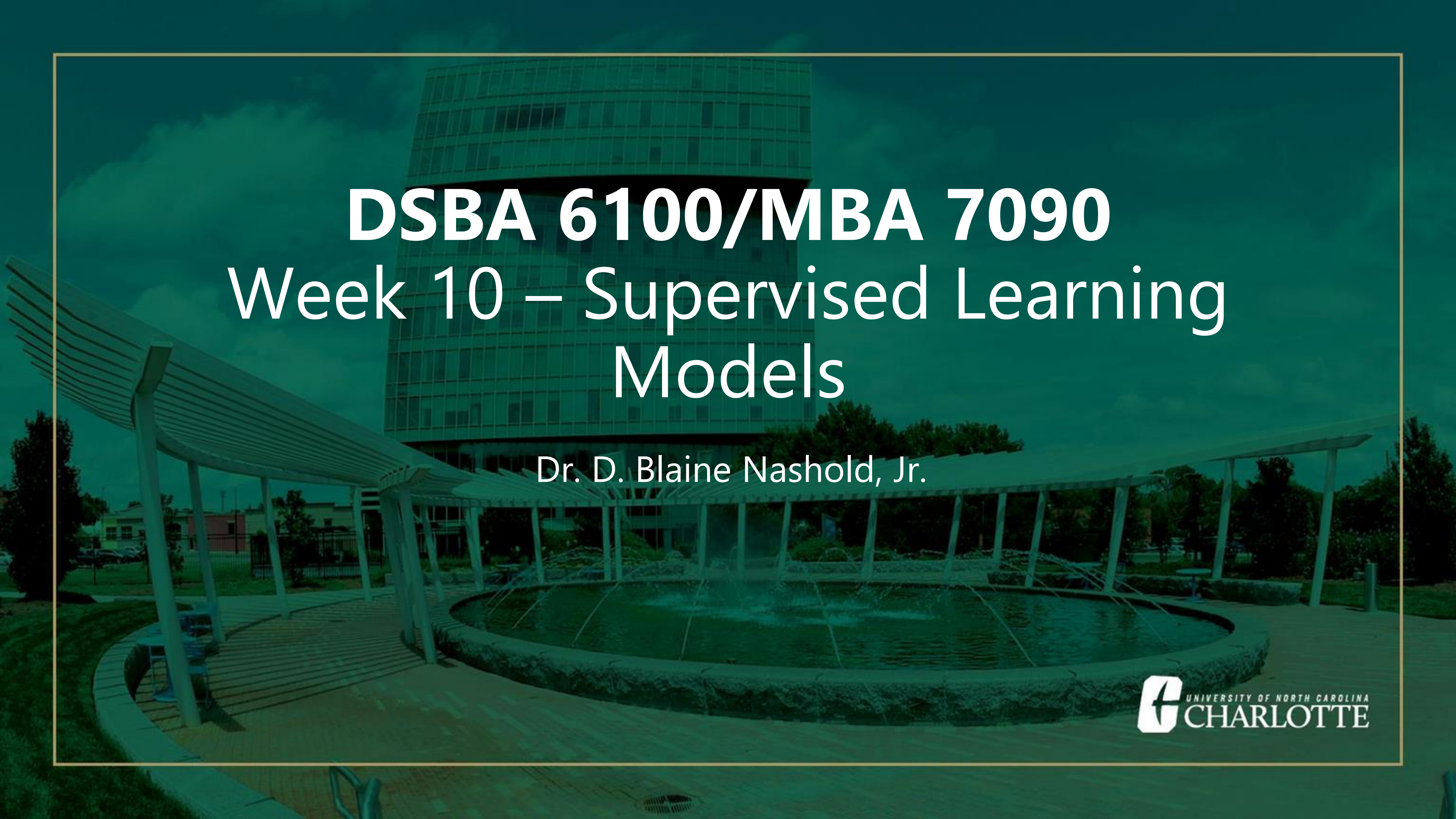


DSBA 6100/MBA 7090
Week 10 – Supervised Learning Models

Dr. D. Blaine Nashold, Jr.

Outline

- Machine Learning & Supervised Learning (or Predictive Modeling)
- Supervised Learning Techniques
- Discussion of select techniques
- Classification Matrix & ROC Curve
- Model Assessment Approaches

Machine Learning

- **Machine learning** is a field of computer science that often uses statistical techniques to give computers the ability to "learn" (i.e., progressively improve performance on a specific task) with data, without being explicitly programmed. (*Source: Arthur Samuel vis Wikipedia*)
- Evolved from “pattern recognition” and “computational learning theory” in artificial intelligence
- In data analytics, machine learning methods are used for predictions by learning from historical relationships and trends in the data
  - These methods are also called predictive analytics methods

Types of Machine Learning

- ***Supervised learning***
  - Each observation in the dataset has an outcome variable (dependent variable)
    - E.g., If customer responded to your promotion or how much did they buy from you last week
  - The learning technique observes the features or predictors (independent variables) of each observation and the outcome and attempts to learn the pattern
  - E.g., Based on data about customer and if she/he bought the product in the past, a model can predict the likelihood of a similar customer buying the product in the future
  - Logistic Regression, Decision Trees, and Random Forests are examples of supervised learning

Types of Machine Learning

- ***Unsupervised learning***
  - There is no outcome variable
  - All variables are features
  - The learning technique attempts to learn if there are underlying groups or relationships among the observations
  - E.g., Based on data about customers, a model can segment the customers into groups with similar characteristics
  - Eg., Based on purchases made by customers, a model to find the strongest and weakest association between the purchase of different products (e.g., how frequently are beer and diapers bought together)
  - K-means clustering and Association Mining are examples of unsupervised techniques

Types of Machine Learning

- ***Reinforcement learning***
  - A newer form of machine learning focused on adaptive learning and artificial intelligence
  - An automated agent starts by selecting a best possible path or answer given current information
  - Algorithm rewards or punishes the agent in each step depending on the outcome
  - The agent is supposed to take the steps to the final state by maximizing the cumulative rewards and/or minimize the cumulative punishments

Classic Statistical Techniques for Learning

- Some classic statistical techniques
  - Linear regression
  - Logistic regression
  - Non-linear regression
  - Time series
    - Autoregressive model
    - Moving averages model
  - Survival analysis

Machine Learning Techniques

- Supervised learning techniques more used now
  - Decision trees
  - Random Forests
  - Support Vector Machines
  - Neural networks
  - Deep learning
  - k-NN (k-nearest neighbor)
- Unsupervised learning techniques
  - Cluster analysis
  - Association mining (market-basket analysis)

Regression for Prediction

- What is a regression model?
  - Statistical model for estimating the relationship between a dependent variable (target) and one or more independent variables (predictors)
- Different types of relationships means different models
  - Linear regression – Simple and Multiple
  - Logistic regression and Multinomial regression
  - Non-linear regression

Logistic Regression Overview

- Logistic regression is a regression technique used when the dependent variable has binary outcomes (e.g., outage or no outage; 1 or 0)
- The logistic regression estimates the probability of the event occurring based on past data
- Logistic regression is considered “supervised learning technique” since the data includes actual outcome values from past observations

Logistic Regression Use Cases

- Medical
  - A model to determine the likelihood of a patient’s successful response to a specific medical treatment or procedure
  - Input variables could include age, weight, blood pressure, etc.
- Finance
  - Determine probability of applicant’s loan default using input variables from credit history and loan details
- Other examples
  - Predicting the probability of a system or part failing
  - Predicting the probability of a subscriber cancelling a service (churn)
  - Predicting the probability of a patient readmission
  - Predicting the probability of an outage

Logistic Regression Prediction Formula

$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

The goal is to estimate the values of the parameters on the right-hand side, i.e., β_0 , β_1 , β_2

The Super-natural World of Natural Logarithm & e

Logarithm (**log**) is an inverse function of exponentiation of base 10.

log(100) = 2 → 100 = 10^{2}

log(150) = 2.1761 → 150 = $10^{2.1761}$

log(0.005) = -4.30103 → 0.0005 = $10^{-4.30103}$

Natural logarithm (**ln**) is an inverse function of exponentiation of base e (i.e., 2.7182818…)

ln(100) = 4.6052 → $e^{4.6052}$ = 100

Log and **ln** are used to work with very large and very small values, but **ln** is preferred in computational work due to its natural representation of compound growth & decay

The Super-natural World of Natural Logarithm & e

Some operations involving **ln** and base **e** that you must familiarize yourself with:

$\mathbf{ln}(ab) = \mathbf{ln}(a) + \mathbf{ln}(b)$

$\mathbf{ln}(a/b) = \mathbf{ln}(a) – \mathbf{ln}(b)$

$\mathbf{ln}(a^2) = 2*\mathbf{ln}(a)$

$\mathbf{ln}(1/a^2) = \mathbf{ln}(a^{-2}) = -2*\mathbf{ln}(a)$

$\mathbf{e^{-x} = (1/e^{x})}$

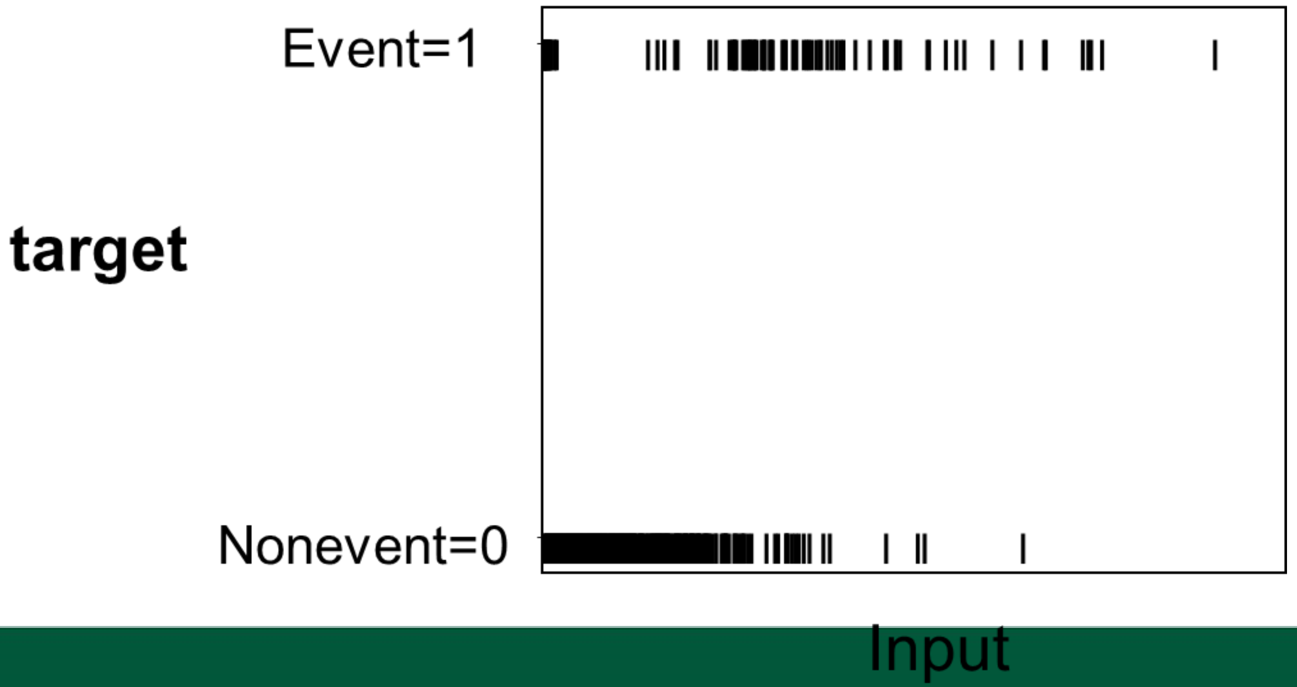
$\mathbf{e^{x} = (1/e^{-x})}$

$\mathbf{e^{x} * e^{y} = e^{x+y}}$

$\mathbf{e^{x} / e^{y} = e^{x-y}}$

Binary Dependent Variable

- When the dependent variable is binary (e.g., values 1 or 0), linear regression does not work directly, since the results are generally unbounded.
- Instead, we use the probability *p* that the event will occur rather than directly using 1 or 0 .



Odds and Probability

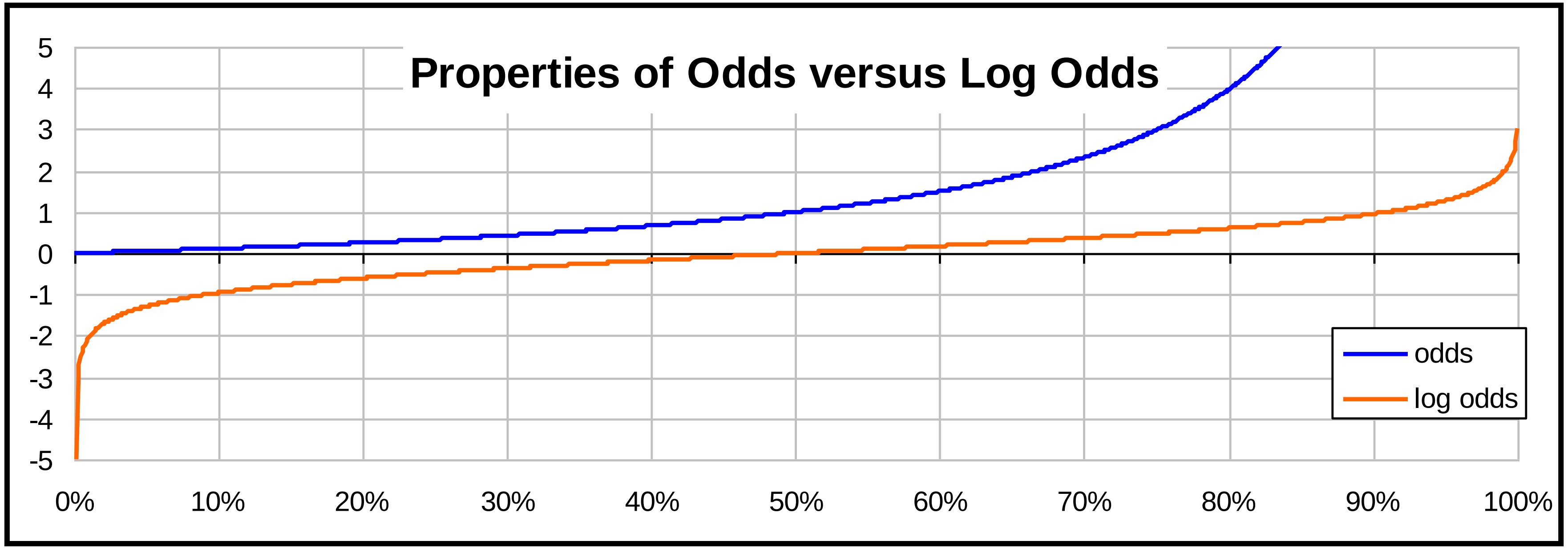
- Consider the probability *p* of an event (such as an outage) occurring.
- The probability of the event not occurring is *1-p*.
- The odds of the event happening are *p*:(1-*p*)

$$odds = \frac{p_{win}}{p_{loss}} = \frac{p}{1-p}$$

- For example, if the probability of outage p=0.2, the odds are $\frac{0.2}{0.8} = 0.25$
- Odds may also be expressed as integers such as 1:4, which means that in every 5 events, 1 event is an outage and 4 events are non-outages
- From the above odds, the probability of an outage is computed as

$$p(outage) = \frac{1}{(1+4)} = 0.2$$

Properties of Odds and Log Odds



- Odds is not symmetric, varying from 0 to infinity.
- Odds is 1 when the probability is 50%.

•Ln Odds is symmetric, going from minus infinity to positive infinity, like a line.

•Ln Odds is 0 when the probability is 50%.

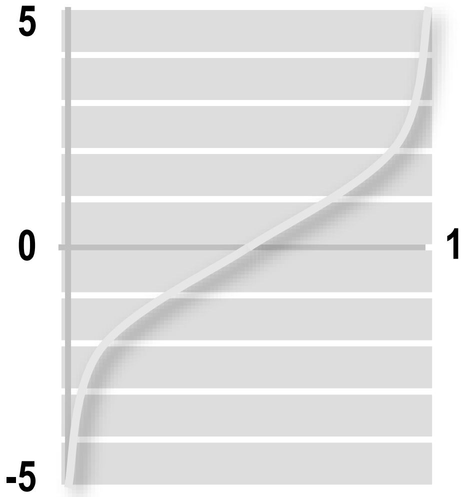
•It is highly negative for low probabilities and highly positive for high probabilities.

Logistic Regression Prediction Formula

$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

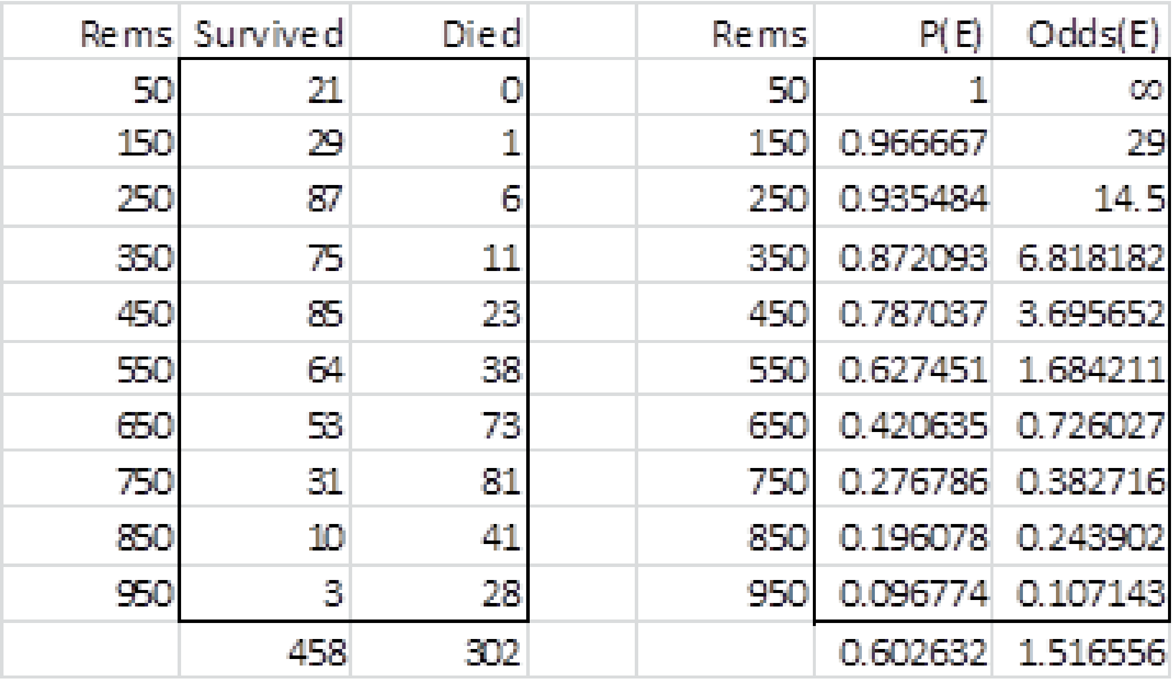
The left-hand side of the above expression is called the logit link function, also referred to as logit(p). The logit link function transforms probabilities (between 0 and 1) to logit scores (between −∞ and +∞).

logit link function



A Sample Dataset

- Following a nuclear accident, individual’s radiation exposure and survival were captured in a dataset. The objective is to estimate the relationship between the radiation exposure (in Rems) and survival.



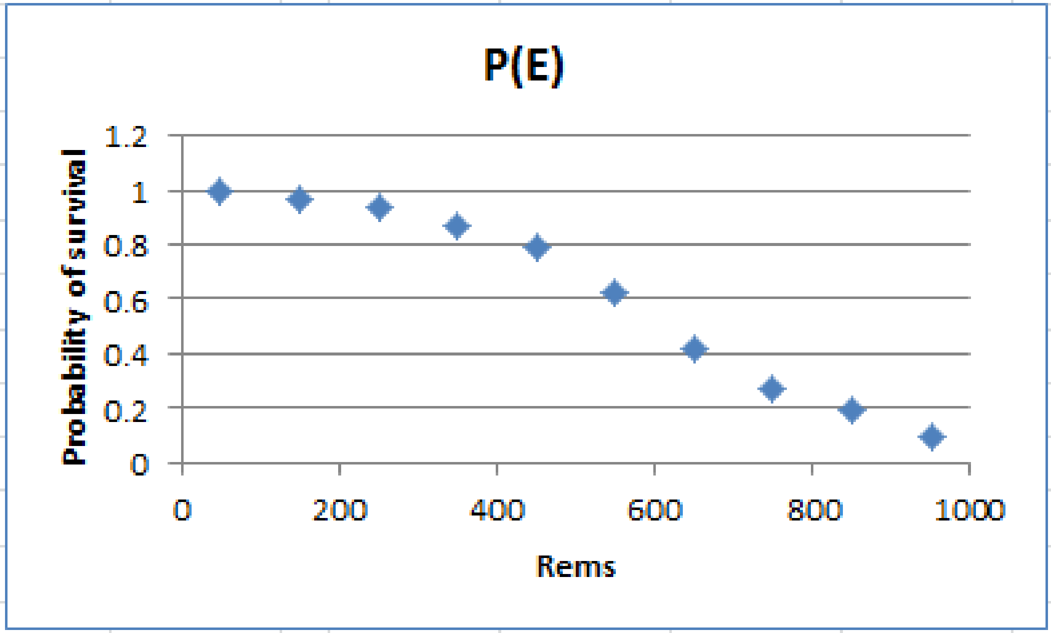
Rems	Survived	Died
50	21	0
150	29	1
250	87	6
350	75	11
450	85	23
550	64	38
650	53	73
750	31	81
850	10	41
950	3	28
	458	302

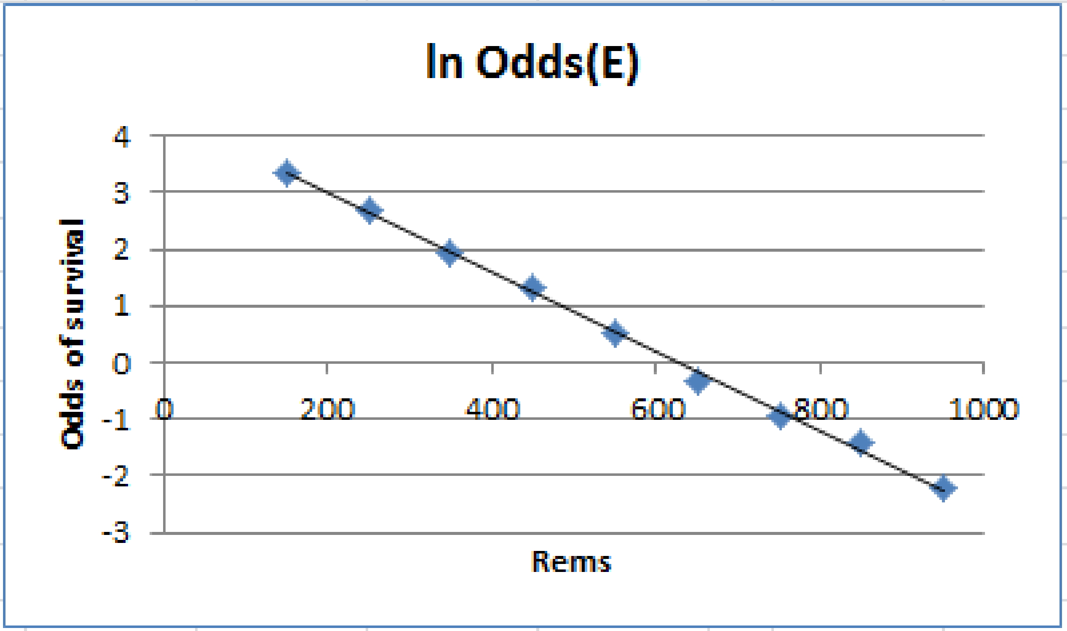
Rems	P(E)	Odds(E)
50	1	∞
150	0.966667	29
250	0.935484	14.5
350	0.872093	6.818182
450	0.787037	3.695652
550	0.627451	1.684211
650	0.420635	0.726027
750	0.276786	0.382716
850	0.196078	0.243902
950	0.096774	0.107143
	0.602632	1.516556

Source: http://www.real-statistics.com/logistic-regression/basic-concepts-logistic-regression/

Comparing p(E) vs ln(odds(E))

E = event, which in this case is the individual survived





Source: http://www.real-statistics.com/logistic-regression/basic-concepts-logistic-regression/

Logistic Regression Prediction Formula

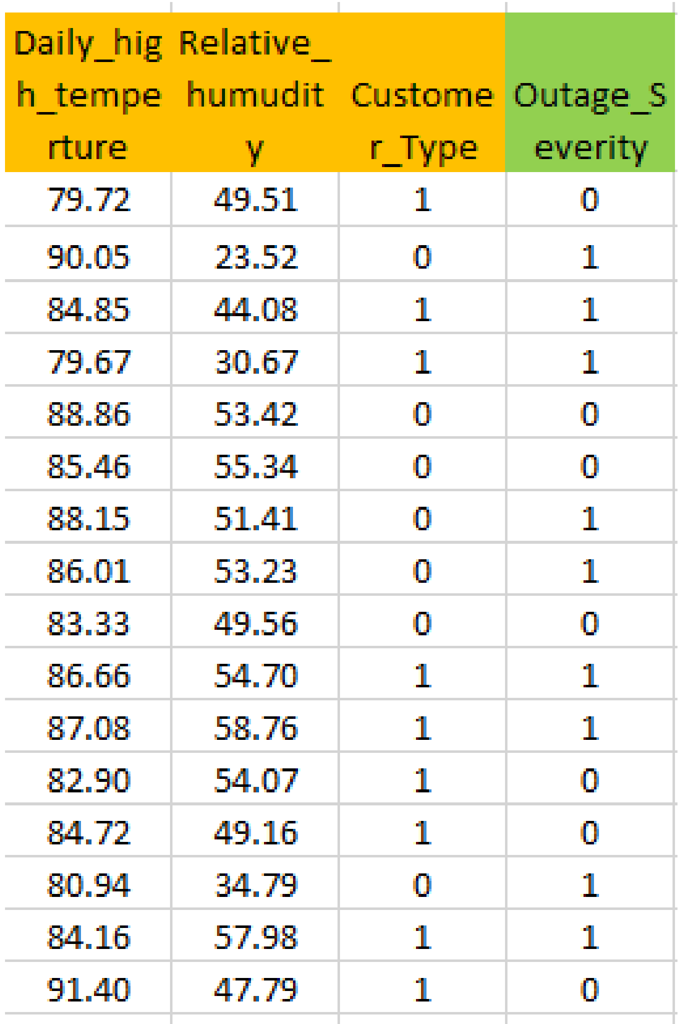
$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 = \text{logit}(\,p\,)$$

Once the parameters are estimated, logit values for given x_1 and x_2 can be computed and the probability of the event is estimated as below

$$\hat{p} = \frac{e^{\text{logit}(p)}}{1 + e^{\text{logit}(p)}} = \frac{1}{1 + e^{-\text{logit}(p)}}$$

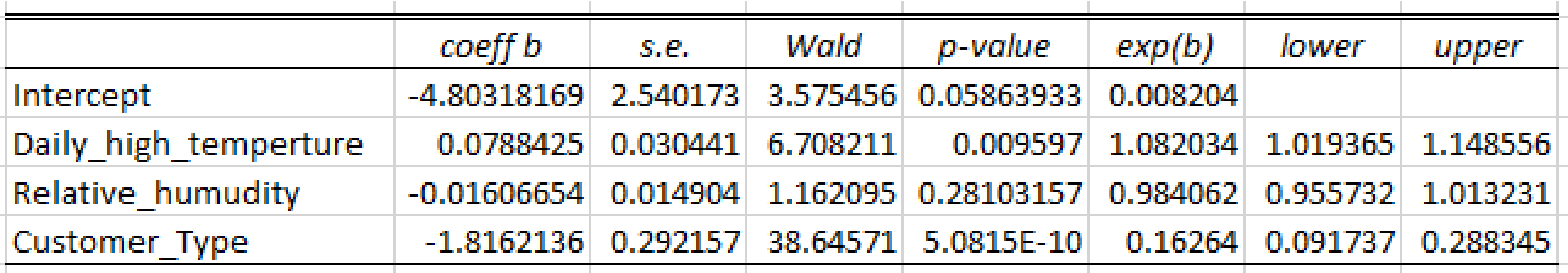
A logistic regression example

- The table on the right is a sample data of power outage:
- Independent variables
  - Daily high temperature (F)
  - Relative humidity (%)
  - Customer type (1=business; 0=residential)
- Dependent variable
  - Outage severity (1=outage lasted for 5 hours or more; 0= otherwise)



Daily_high_temperture	Relative_humudity	Customer_Type	Outage_Severity
79.72	49.51	1	0
90.05	23.52	0	1
84.85	44.08	1	1
79.67	30.67	1	1
88.86	53.42	0	0
85.46	55.34	0	0
88.15	51.41	0	1
86.01	53.23	0	1
83.33	49.56	0	0
86.66	54.70	1	1
87.08	58.76	1	1
82.90	54.07	1	0
84.72	49.16	1	0
80.94	34.79	0	1
84.16	57.98	1	1
91.40	47.79	1	0

Logistic Regression Estimates



	coeff b	*s.e.*	*Wald*	*p-value*	*exp(b)*	*lower*	*upper*
Intercept	-4.80318169	2.540173	3.575456	0.05863933	0.008204		
Daily_high_temperture	0.0788425	0.030441	6.708211	0.009597	1.082034	1.019365	1.148556
Relative_humudity	-0.01606654	0.014904	1.162095	0.28103157	0.984062	0.955732	1.013231
Customer_Type	-1.8162136	0.292157	38.64571	5.0815E-10	0.16264	0.091737	0.288345

- The above results give the estimated model as:

  logit(*p*) = –4.803 + 0.079*(daily high temp) – 0.016*(rel humidity) – 1.816*(customer type)

- To find the probability of a severe outage given
  - **Daily high temp = 85F**
  - **Rel humidity = 50%**
  - **Customer type = Business**
- Compute **logit(p)** value with the numbers above and substitute in the following formula

$$\hat{p} = \frac{e^{\mathrm{logit}(\,p\,)}}{1 + e^{\mathrm{logit}(\,p\,)}} = \frac{1}{1 + e^{-\mathrm{logit}(\,p\,)}}$$

Logistic Regression – Interpreting Effects of Independent Variables

$$ln(odds(Y_{x_1,x_2})) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$ for x_1 & x_2

$$ln(odds(Y_{x_1,(x_2+1)})) = \beta_0 + \beta_1 x_1 + \beta_2 (x_2+1)$$ for x_1 & (x_2+1)

$$ln(odds(Y_{x_1,(x_2+1)})) - ln(odds(Y_{x_1,x_2})) = \beta_2$$

$$ln(odds_{new}) - ln(odds_{old}) = \beta_2$$

$$ln(odds_{new} / odds_{old}) = \beta_2$$

$$odds\ ratio = e^{\beta 2}$$

Interpreting Logistic Regression Estimates

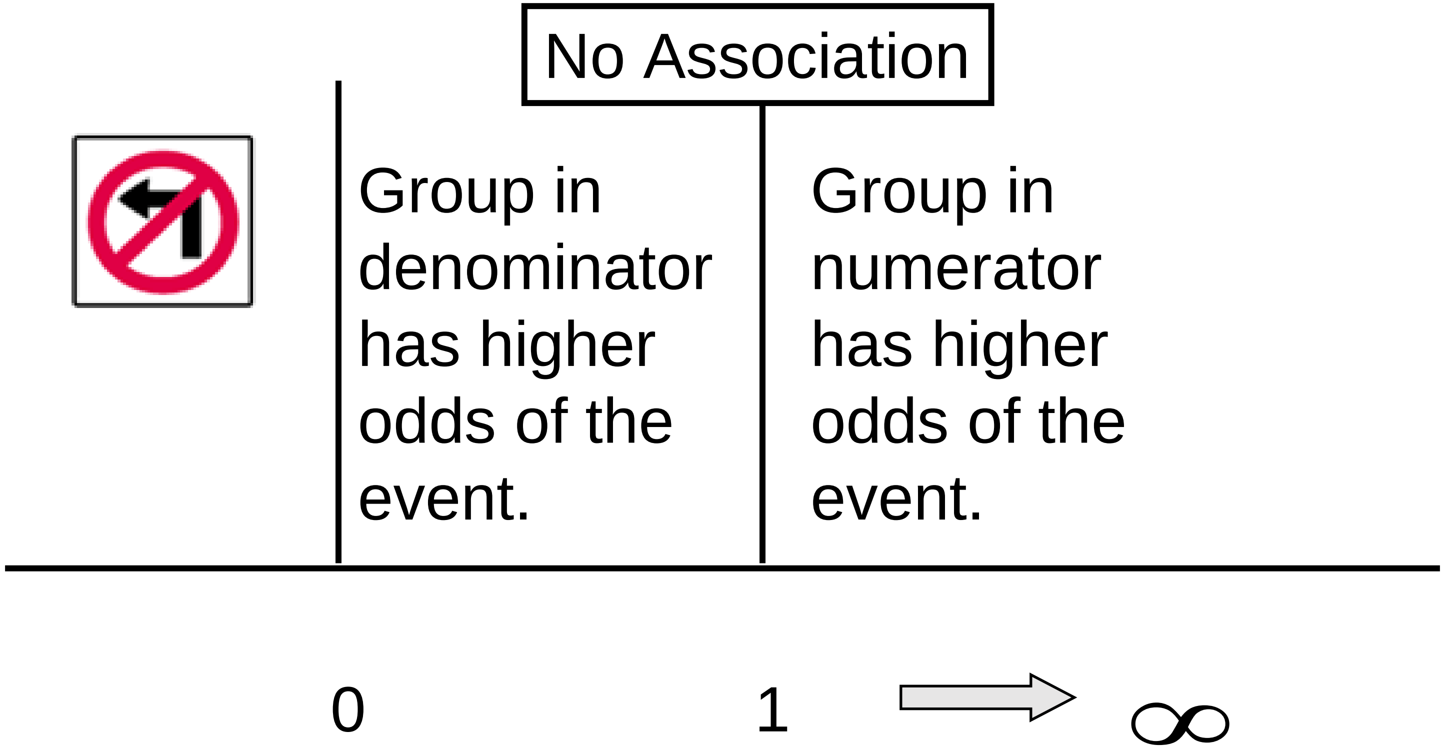
- Estimated logistic regression model is as follows:

logit(p) = –4.803 + 0.079*(**daily high temp**) – 0.016*(**rel humidity**) – 1.816*(**customer type**)

- Since we converted the **odds** to ***ln(odds)*** to estimate the coefficients, we need to convert the coefficient of each independent variable by raising it to the power of ***e*** to get **odds ratio**
- The estimated odds ratio for variable **daily high temp** = $\boldsymbol{e^{0.079}}$= **1.082**
- The odds ratio of 1.082 means that for every **<u>unit increase in daily high temp</u>**, the **odds of a severe outage increases by 1.082 times**
- In the above example, estimated odds ratio for variable **customer** =

  $\boldsymbol{e^{-1.816}}$**= 0.1627**
- The odds ratio of 0.1627 means that a **<u>business type customer</u>** has **0.1627 times the odds** of having a severe outage compared to **<u>residential type customer</u>** (i.e., business has lower odds of outage).

Properties of the Odds Ratio



No Association

Group in denominator has higher odds of the event.

Group in numerator has higher odds of the event.

0

1

∞

Logistic Regression Fit Measures

- The logistic regression coefficients are estimated using maximum likelihood estimation (MLE).
- Unlike the Least Squares Estimation used for linear regression models, the MLE begins with a tentative solution, revised it slightly to see if it can be improved, and repeats until the results have converged.
- Logistic regression's goodness of fit is measured as (-2*(log likelihood of the fitted model)).
- Other measures are likelihood ratio, Cox and Snell R^2 and Nagelkerke R^2.
- Wald Statistic, which is the ratio of the square of the regression coefficient to the square of the standard error of the coefficient, is used to assess the significance of each coefficient.

Logistic Regression Estimation Tools

- There are different software options to estimate logistic regression
- The following are examples of add-ins available to solve logistic regression in Excel
  - **Real Statistics Resource Pack**: An Excel add-in available for free download from http://www.real-statistics.com
  - **XLSTAT-Base**: An Excel add-in available from https://www.xlstat.com/en/solutions/base with paid license
- **SAS**, **Stata**, and **SPSS Modeler** are examples of non-Excel software that has logistic regression modeling capabilities
- **R** is an open-source and popular statistical computing software that has logistic regression modeling capabilities

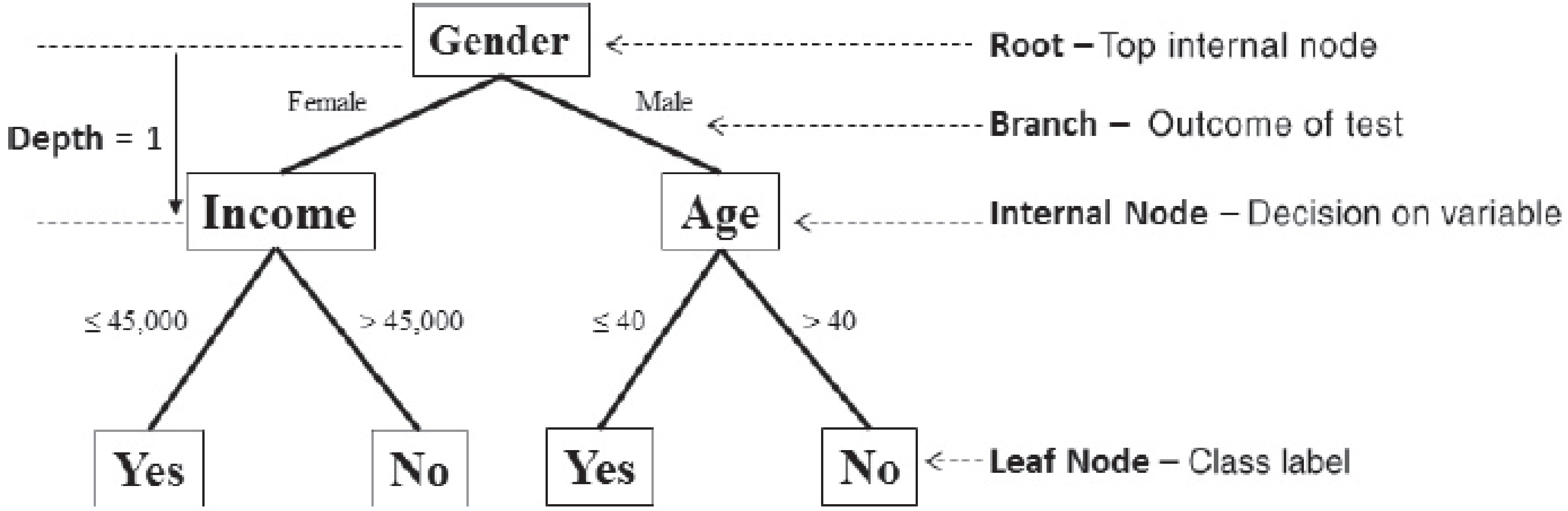
Decision Trees

- Decision tree is a classification technique
- When presented with examples that are already classified, the decision tree learns how likely the attributes of the observations contribute to the specific class of each observation
- A **decision tree** algorithm induces a **tree**-like graph or model of **decisions** rules and the possible consequences of following a set of rules
- Decision tree is a supervised learning technique since the training observations are already classified

Decision Trees - Terminology

- Each test point in a tree is called a node
- Branches represent the decisions or consequences of the test point
- The first test point (or the top node) is called the root node
- A node without further branches (or where testing stops) is called leaf node
- The leaf node shows the most likely class in that branch, and sometimes show the probability of being in that class

An example decision tree

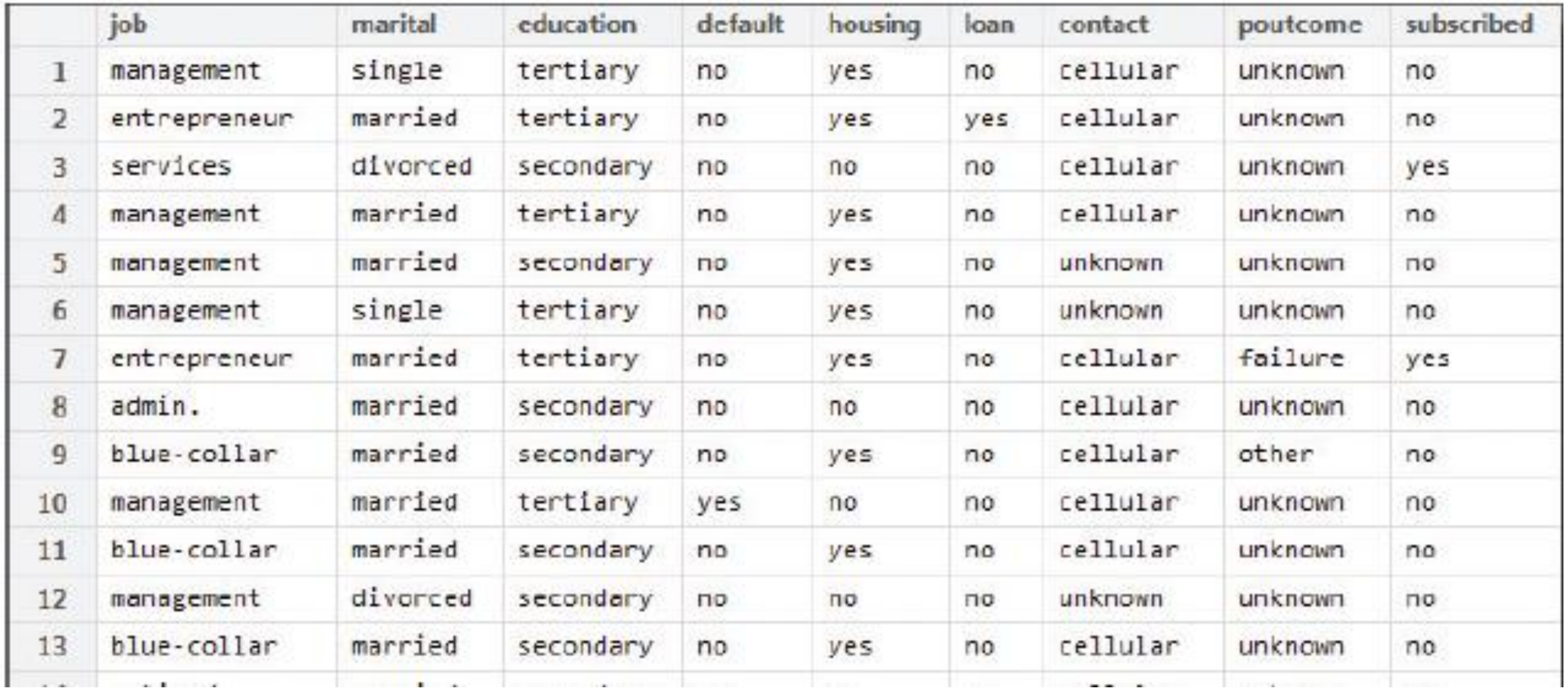

Gender
Female
Male
Depth = 1
Income
Age
≤ 45,000
> 45,000
≤ 40
> 40
Yes
No
Yes
No
Root – Top internal node
Branch – Outcome of test
Internal Node – Decision on variable
Leaf Node – Class label

Source: Chapter 7 of required textbook, page 193

Decision Tree Algorithms

- Employs the divide and conquer method
- Steps in building a decision tree:
  1. Create a root node and assign all of the training data to it.
  2. Select the best splitting attribute (i.e., the split data should be as dissimilar with respect to the outcome variable as possible).
  3. Add a branch to the root node for each value of the split. Split the data into mutually exclusive subsets along the lines of the specific split.
  4. Repeat steps 2 and 3 for each node until the node can be split no more or some stopping criteria is reached.

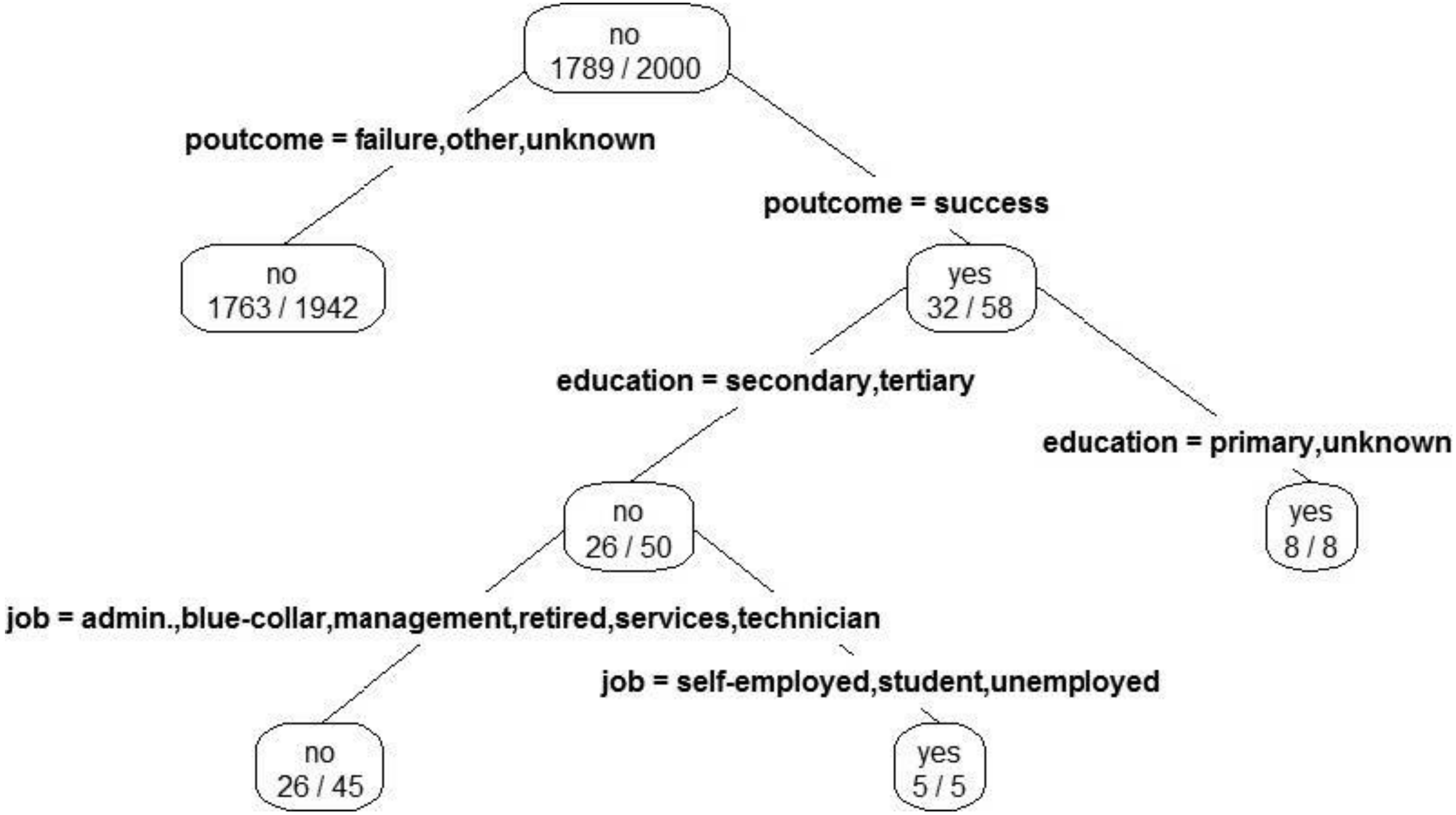
A Bank Dataset Example



	job	marital	education	default	housing	loan	contact	poutcome	subscribed
1	management	single	tertiary	no	yes	no	cellular	unknown	no
2	entrepreneur	married	tertiary	no	yes	yes	cellular	unknown	no
3	services	divorced	secondary	no	no	no	cellular	unknown	yes
4	management	married	tertiary	no	yes	no	cellular	unknown	no
5	management	married	secondary	no	yes	no	unknown	unknown	no
6	management	single	tertiary	no	yes	no	unknown	unknown	no
7	entrepreneur	married	tertiary	no	yes	no	cellular	failure	yes
8	admin.	married	secondary	no	no	no	cellular	unknown	no
9	blue-collar	married	secondary	no	yes	no	cellular	other	no
10	management	married	tertiary	yes	no	no	cellular	unknown	no
11	blue-collar	married	secondary	no	yes	no	cellular	unknown	no
12	management	divorced	secondary	no	no	no	unknown	unknown	no
13	blue-collar	married	secondary	no	yes	no	cellular	unknown	no

Select variable details: ***default*** (if credit is in default); ***housing*** (if has housing loan); ***loan*** (if has personal loan); ***poutcome*** (outcome of previous marketing contact); ***subscribed*** (subscribed or not to term deposit)

A Bank Dataset Example Decision Tree


no
1789 / 2000
poutcome = failure,other,unknown
poutcome = success
no
1763 / 1942
yes
32 / 58
education = secondary,tertiary
education = primary,unknown
no
26 / 50
yes
8 / 8
job = admin.,blue-collar,management,retired,services,technician
job = self-employed,student,unemployed
no
26 / 45
yes
5 / 5

Note: Each node shows the predicted class (“yes” or “no” if credit is in default). Also, the number on left in each node is the number of observations with the selected class and the number on the right is the total observations in that node.

Decision Trees – Splitting Criteria

- Gini index determines the purity of a specific class as a result of a decision to branch along a particular attribute/value
  - Used in CART decision tree algorithm
- Information gain uses entropy to measure the extent of uncertainty or randomness of a particular attribute/value split and compares it with base entropy
  - Used in ID3, C4.5, C5 decision tree algorithms
- Chi-square statistics(uses Chi-square test to determine the extent to which the split groups are different from each other with respect to the dependent variable
  - Used in CHAID decision tree algorithm

Decision Tree Pruning

- Pruning is a process used by some decision tree algorithms to avoid overfitting
- Pruning involves the following steps
  1. First, the nodes are repeatedly split (using the best split criteria at each node) without any stopping criteria to build the “maximal tree” and the performance of the tree is obtained from the validation data
     - The maximal tree is the most complex tree
  2. The lowest branch is removed and the performance of the tree is obtained from the validation data
  3. Each time, the lowest branch is removed and the performance of the tree is obtained from the validation data until there are no more branches to remove
  4. The tree with the best performance and smallest number of branches is selected as the optimal tree

Decision Trees

- Decision Tree algorithms mainly differ on
  - Splitting criteria
    - Which variable, what value, etc.
  - Stopping criteria
    - When to stop building the tree
  - Pruning (generalization method)
    - Pre-pruning versus post-pruning
- Most popular Decision Tree algorithms include
  - ID3, C4.5, C5; CART; CHAID; M5

Decision Tree Modeling Tools

- There are different software options to model decision trees
- **Simple Decision Tree** is an open-source Excel add-in available for download from https://sites.google.com/site/simpledecisiontree/
- **SAS** and **SPSS Modeler** are examples of non-Excel software that have decision tree modeling capabilities
- **r-part** is a R package available for download from https://cran.r-project.org/web/packages/rpart/index.html and can be used for decision trees
- **tree** is another R package available for download from https://cran.r-project.org/web/packages/tree/index.html and can be used for decision trees

Diagnostics of Classifiers

- For predictive models with binary dependent variable, the primary source for accuracy estimation is the decision matrix.

		Predicted Class (1 or 0)	
		Positive (1)	Negative (0)
Actual Class (1 or 0)	Positive (1)	True Positive (TP)	False Negative (FN)
	Negative (0)	False Positive (FP)	True Negative (TN)

$$Accuracy = \frac{TP + TN}{All\ Cases} * 100\%$$

$$Recall\ or\ True\ Positive\ Rate = \frac{TP}{TP + FN}$$

$$False\ Positive\ Rate\ or\ Type\ I\ Error\ Rate = \frac{FP}{FP + TN}$$

$$False\ Negative\ Rate\ or\ Type\ II\ Error\ Rate = \frac{FN}{TP + FN}$$

Decision Matrix - Example

- For a dataset of 1000 bank customers, a decision tree was used to classify them as “High” or “Low” risk for loan default.

		Predicted Risk Class	
		High	Low
Actual Risk Class	High	TP = 262	FN = 29
	Low	FP = 38	TN = 671

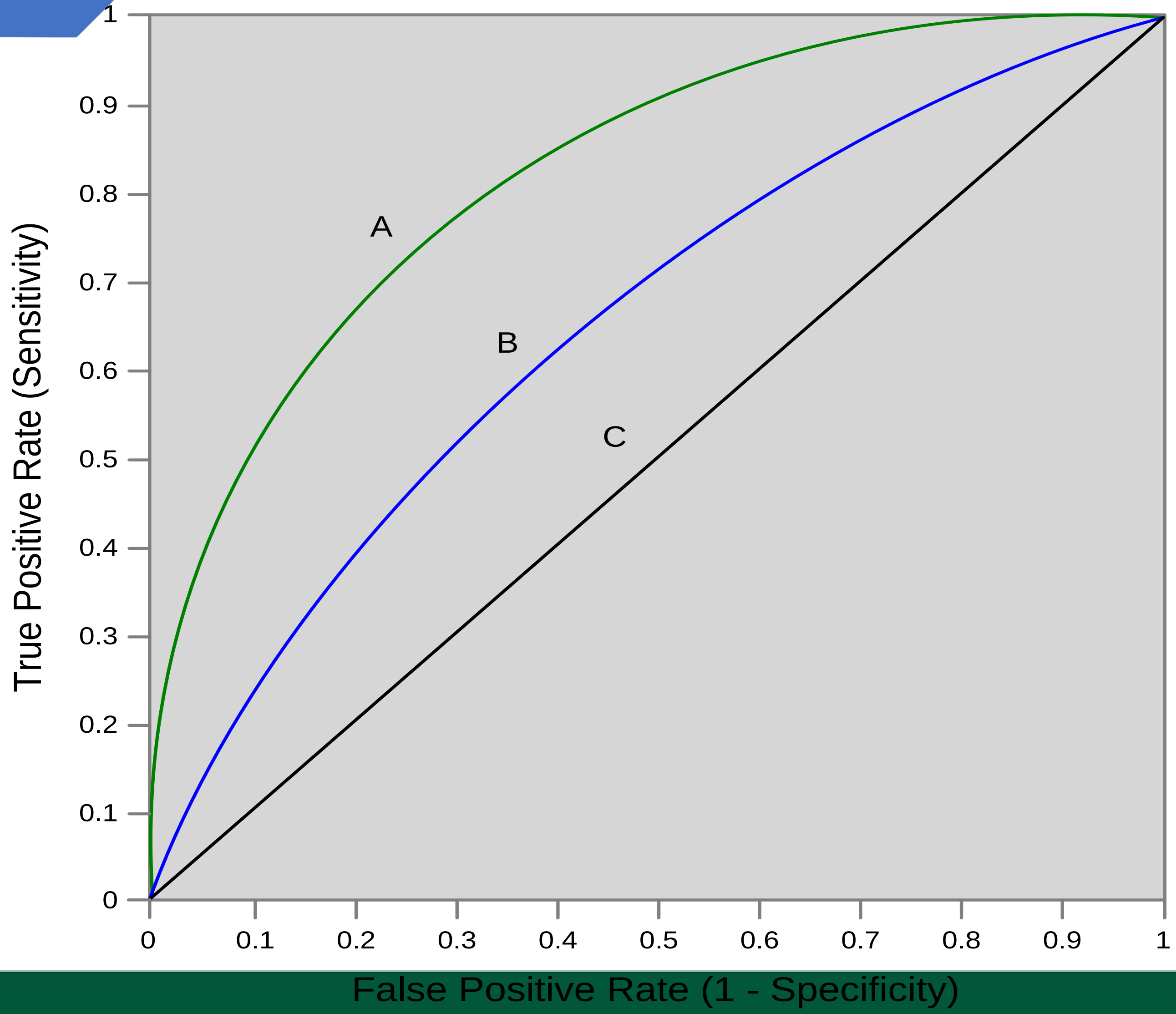
$Accuracy =$

$Recall\ or\ True\ Positive\ Rate =$

$False\ Positive\ Rate\ or\ Type\ I\ Error\ Rate =$

$False\ Negative\ Rate\ or\ Type\ II\ Error\ Rate =$

ROC Curve



- **Receiver Operating Characteristic curve**
- **Plots True Positive Rate (i.e., Sensitivity) against False Positive Rate (1-Specificity)**
- **A good classifier is supposed to have a high TPR and low FPR for a range of cut-off values**
- **As the curve traces from left to right, preferable for TRP to rapidly approach 1 with small changes in FPR**
- **The metric used is area under the curve, with preference given to classifier with more area under the curve**

Model Assessment

- When learning from data, there is a danger of over-fitting the model to the learning sample (training data)
- Overfitting: Model has very high prediction accuracy for the training data, but fails on new data
- Model assessment approaches
  - Partitioning data set
    - Training and Validation/Testing
    - Training data used to build model, but model performance obtained from Validation data
  - Multi-fold cross validation
  - Bootstrap procedures

Model Assessment Approaches

- Partitioning data set
  - This involves dividing the input data set into a minimum of two groups – training and validation
  - The training set is used to train the model (i.e., build the model)
  - The model is applied on the validation set to assess its performance (e.g., model accuracy)
  - There are no clear rules for partitioning, but usually the split is 60% for training and 40% for validation
  - Sometimes a dataset is partitioned into three groups – training, validation, and testing
  - When different types of models are to be tried (e.g., a decision tree, a neural network, a logistic regression), the training and validation sets are used to select the best in each type
  - Each selected model type is then run with the test set to select the final model type

Model Assessment Approaches

- Multi-fold cross validation
  - The data set is divided into *k* subsets
  - Each time, one of the *k* subsets is used as the validation set and the other *k-1* subsets are put together to form a training set
  - Then the average performance across all *k* trials is computed
  - The advantage of this method is that it matters less how the data gets divided
  - Every data point gets to be in a test set exactly once, and gets to be in a training set *k-1* times.

Model Assessment Approaches

- Bootstrapping
  - When using the training-validation partition, the trained model’s performance is obtained only once on the validation set
  - Instead, in bootstrapping, the dataset is sampled multiple times to create many validation sets
  - The sampled data is put back into the original set, so some of the observations may be used multiple times
  - The performance measure (e.g., accuracy) is now available for each sample and can be plotted as a distribution
  - Conclusions made about the performance of the model based on this distribution is more reliable than a single measure from one validation set

Sources

- “Business Intelligence and Analytics” by Sharda et al., Pearson Education, 2015
- “Advanced Business Analytics” Educator Training by SAS Inc.
- http://www.real-statistics.com
- http://docs.statwing.com/interpreting-residual-plots-to-improve-your-regression/
- Miller, “Modeling Techniques in Predictive Analytics with R” FT Press