Solution Guide
for
Data Science and Big Data Analytics,
by EMC Education Services

Chapter 1

1) Big data is characterized by Volume, Variety, and Velocity each of which present unique and differing challenges.

Volume – Growing well beyond terabytes, big data can entail billions of rows and millions of columns. Data of this size cannot efficiently be accommodated by traditional infrastructure or RDBMS.

Variety – Data that comes in many forms, not just well-structured tables with rows and columns. Some unstructured data examples include: video files, audio files, XML, and free text. Traditional RDBMS provide little support for these data types.

Velocity – Data that is collected and analyzed in real time. Often, this type of data is time sensitive and its value diminishes with time. This type of data may require in-memory data grids to accommodate the real-time nature of this data.

The main considerations in processing Big Data are how to cost effectively store and analyze the data in an efficient manner. Often new tools and technologies (e.g. Hadoop) are necessary to accomplish these goals.

2) An analytics sandbox is a workspace that is typically isolated from production applications and warehouse environments. This separation facilitates data exploration that does not interfere with the production environment. An analytics sandbox is crucial to the data scientist in that it facilitates the analytics lifecycle activities.

3) The difference between Business Intelligence (BI) and Data Science can be explained by examining four common facets of each.

- **Statistical Category**
  BI – Descriptive Statistics
  Data Science – Inferential Statistics
- **Analytical Techniques**
  BI – Standard & ad hoc reporting, dashboards, alerts, queries, details on demand
  Data Science – Optimization, predictive modeling, forecasting
- **Data Types**

BI – Structured data, traditional data sources, manageable datasets
Data Science – Structured & unstructured data, many types of sources, very large data sets

- **Common Questions**
  BI – What happened last quarter? How many units sold? Where is the problem?
  Data Science – What if? What's the optimal scenario for our business? What will happen next?

4) Data science projects require workspaces that are purpose-built for experimenting with data, with flexible and agile data architectures.

Traditional analytical architecture is designed to support enterprise data warehouses which enforce rigorous validation and data structuring processes. They are designed for mission critical "operational" data and don't lend themselves to exploration and analysis by data scientists.

Moreover, the current environments:

- Don't support the analysis of unstructured data
- Limit the ability of performing in-memory analytics
- Are prioritized for operational reporting needs, not ad-hoc analysis

5) Data scientists are technically savvy, have strong analytical skills, and have a combination of skills that enable them to handle raw, unstructured data and to apply complex analytical techniques at scale. They tend to be trained in a quantitative discipline such as mathematics, statistics, economics, and machine learning.

Chapter 2

1) The data preparation phase is the most iterative one and the one that teams tend to underestimate the amount of effort involved. This is because the teams are anxious to begin analyzing the data, testing hypotheses, and getting answers to some of the questions posed in the discovery phase.

As long as the appropriate effort was conducted in the earlier phases to gather, understand, and prep the data, the teams are expected to spend the least time in model building phase.

2) GINA Team wanted to analyze and measure innovation, which is a difficult concept to measure. Team was undertaking this type of project for the first time. In the mini case conducting a pilot project before a full scale roll out is good because of the uncertainties that the team would have to face at multiple levels in the analytics life cycle prior to the completion of the project.

The benefits of doing this are:

1. Mitigate risk of project failure

   GINA project faced uncertainty and challenges in finding data sources, forming new teams in the data discovery phase and faced challenges with the quality of data collected in the data

preparation phase. A full scale roll out of the project would have resulted in unnecessary deadlocks and lead to the failure of the project.

2. Speed

GINA prototype was faster to build and test resulting in quicker delivery of results.

3. Learning and Improvement

GINA team was able to quickly learn from the performance at different phases of the analytics lifecycle and improve the results.

3) Possible tools for phases 2 and 4 are:

a. Phase 2: Data Preparation

Hadoop – Massive parallel processing
Alpine Miner - GUI for analytic workflows
OpenRefine – Open source data preparation tool
Data Wrangler – Data cleaning and transformation tool

b. Phase 4: Model building

R – Open source data analytics tool
SAS Enterprise Miner – Predictive and descriptive models b
SPSS Modeler - Explore and analyze data
Alpine Miner - Analytic workflows
Statistica and Mathematica – Data mining tools
Octave – Computational modeling tool
Weka – Open source data mining software package
Python – Open source programming language
MADlib or other in-database machine learning library

Chapter 3

1) fdata contains three levels: 1 2 3

2) cbind() is used to combine variables column wise

```
cbind(v1,v2)

     v1 v2
[1,]  1  6
[2,]  2  5
[3,]  3  4
```

```
[4,]    4  3
[5,]    5  2
```

rbind() is used to combine datasets row wise.

rbind(v1,v2)

	[,1]	[,2]	[,3]	[,4]	[,5]
v1	1	2	3	4	5
v2	6	5	4	3	2

3) is.na() - provides test for missing values

na.exclude() - returns the object with incomplete cases removed (see page 86)

4) The function install.packages() is used to install a R package. For example, install.packages("ggplot2") would install the ggplot2 package

5) factor () is used to encode a vector as category

6) rug() function creates a one dimensional density plot on the bottom of the graph to emphasize the distribution of observation

7) Viewing the logarithm of data can help detect structures that might otherwise be overlooked in a graph with a regular, non-logarithmic scale.

8) The "box" of the box and whisker shows the range that contains the central 50% of the data, and the line inside the box is the location of the median value. The upper and lower hinges of the boxes correspond to the first and third quartiles of the data. Upper whisker extends from the hinge to the highest value that is within 1.5*IQR of the hinge. Lower whisker extends from the hinge to the lowest value within 1.5*IQR of the hinge. Points outside the whiskers are considered as possible outliers. (see Figure 3-16)

9) According to the scatterplot, within certain species, there is a high correlation between:

- sepal.length and sepal.width (setosa)
- sepal.length and petal.length (veriscolor and virginica)
- sepal.width and petal.length (veriscolor)
- sepal.width and petal.width (veriscolor)
- petal.width and petal.length (veriscolor and virginica)

The relationship between these attributes is a linear relationship. The correlations can be determined using the cor() function.

10) loess() function with the predict() function can be used to fit a nonlinear curve to data

11) If the data is skewed and positive, viewing the logarithm of data can help detect structures that might otherwise be overlooked in a graph with a non-logarithmic scale.

12) Type 1 error is the rejection of null hypothesis when the null hypothesis is true

Type 2 error is the acceptance of null hypothesis when the null hypothesis is false

Committing one error is not necessarily more serious than the other. Given the underlying assumptions, the type I error can be defined up front before any data is collected. For a given deviation from the null hypothesis, the Type 2 error can be obtained by using a large enough sample size.

13) Let's assume that the objective is to compare whether or not a person receiving an offer will spend more than someone who does not receive an offer. If normality of the purchase amount distribution is a reasonable assumption, the Student's t test could be used. Otherwise, a non-parametric test such as the Wilcoxon rank-sum test could be applied.

14) P value of 0.0000433 < 0.05. Therefore, the decision will be to reject null hypothesis

Chapter 4

1) The solution involves expressing the Euclidean distance between a point and its associated centroid and illustrating that the calculated distance is dependent on the units of measure. In this case, the choice is whether a person's height is expressed in centimeters or meters. Let (a_1, h_1) denote the observed age (years) and height (centimeters) of a particular individual. The distance, *d*, from this point to a possible centroid (a_c, h_c) can be expressed as:

$$d = \sqrt{(a_1 - a_c)^2 + (h_1 - h_c)^2} = \sqrt{(a_1 - a_c)^2 + (100h_1' - 100h_c')^2} = \sqrt{(a_1 - a_c)^2 + 100^2(h_1' - h_c')^2}$$

where h' denotes the heights expressed in meters

Thus, if the distance was calculated using height expressed in meters, the distance, *d*', would be:

$$d' = \sqrt{(a_1 - a_c)^2 + (h_1' - h_c')^2}$$

So, except in the rare case where $h_1 = h_c$, $d > d'$ and the contribution of height under the radical sign, will be 10,000 time greater when the height is expressed in centimeters than when expressed in meters. Thus, height expressed in centimeters will have a greater influence in determining the clusters.

Furthermore, students may consider the resulting units of d *or* d' when the units of measure are not removed by dividing through by the standard deviation.

2) The following five clustering algorithms will be compared and contrasted relative to:

- The shape of the resulting clusters
- How the attributes are selected and the order in which the selection occurs
- Characteristic of the identified centers

Characteristic	K-means	K-modes	Partitioning Around Medoids	Hierarchical Agglomerative	Density
Cluster Shape	Spherical*	NA**	Spherical*	Tree structure that combines smaller clusters	Irregular
Attribute selection	Equally weighted and considered simultaneously	Equally weighted and considered simultaneously	Equally weighted and considered simultaneously	Bottoms up (all observations start in their own clusters)	Equally weighted and considered simultaneously
Centers	Based on averages of the cluster members; typically not part of the dataset	Not necessarily part of the dataset	Appear in the dataset	NA***	Not necessarily part of the dataset

Notes:

* Spherical refers to the idea of a radius (Euclidean distance measurement) from the center of the cluster. The actual shape may not appear spherical depending on how close the centers are and the observations in the provided dataset.

** With categorical data, there is no meaningful difference between two values as is the case with interval data (see Chapter 3).

*** Clusters are combined with other clusters or individual observations to make larger clusters.

3) Using R to calculate and plot the Within Sum of Squares for k=1 to 10 clusters, k=4 provides the greatest cumulative reduction in the WSS. See attachment for example R code.


ruspini_kmeans.R

Chapter 5

1) The Apriori property states that if an itemset is considered frequent, then any subset of the frequent itemset must also be frequent.

For example, if an itemset {pencil, book} appears in 70% of all retail transactions, the itemset {pencil} as well as the itemset {book} will appear in at least 70% of the transactions

2) Given an itemset L, the "support" of L is the percentage of transactions containing L. To meet support criteria of 0.5 we need to find the sets of transactions that show up at least 50% of the time.

T1 : { A,B,C }
T2 : { A,C }
T3 : { B,C }
T4 : { A,D }
T5 : { A,C,D }

Itemset	Frequency	Support
A	**4**	**4/5**
B	2	2/5
C	**4**	**4/5**
D	2	2/5
AB	1	1/5
AC	**3**	**3/5**
AD	2	2/5
BC	2	2/5
CD	1	1/5
ABC	1	1/5
ACD	1	1/5

Note: the itemsets with frequency of zero are omitted from this table.

Itemsets **A, C, and AC** satisfy the minimum support of 0.5.

3) Interesting rules are identified by their measure of confidence.

$$\text{Confidence}(X \rightarrow Y) = \frac{\text{Support}(X \wedge Y)}{\text{Support}(X)}$$

A high confidence indicates that the rule $(X \rightarrow Y)$ is more interesting or more trustworthy, based on the sample dataset. When greater than a predefined threshold, known as the minimum confidence, a relationship is considered interesting.

Confidence is used to determine which rules are interesting; however, it cannot determine whether or not the rule is by coincidence.

Lift is used to determine how X and Y are related , that is whether their relationship is coincidental or not.

$$\text{Lift}(X \rightarrow Y) = \frac{\text{Support}(X \wedge Y)}{\text{Support}(X) * \text{Support}(Y)}$$

A Lift value is less than or close to 1 indicates that X and Y occur independently of each other and the relationship is thus considered coincidental.

Lift > 1 indicates there is a relationship; the larger the Lift (greater than 1) the more useful the rule.

Note that Lift(X→Y)=Lift(Y→X). So, lift is typically only applied to the rules that meet some minimum confidence level.

4) There are 10,000 transactions in all with following statistics:

Itemset	Count
{battery}	6000
{sunscreen}	5000
{sandals}	4000
{bowls}	2000
{battery, sunscreen}	1500
{battery, sandals}	1000
{battery, bowls}	250
{battery, sunscreen, sandals}	600

a) Applying the definition of support:

Itemset	Count	Support
{battery}	6000	0.6
{sunscreen}	5000	0.5
{sandals}	4000	0.4
{bowls}	2000	0.2
{battery, sunscreen}	1500	0.15
{battery, sandals}	1000	0.1
{battery, bowls}	250	0.025
{battery, sunscreen, sandals}	600	0.06

b) Based on a minimum support of 0.05 each itemset is considered “frequent.”

c) Confidence is defined by:

$$\text{Confidence}(X \rightarrow Y) = \frac{\text{Support}(X \wedge Y)}{\text{Support}(X)}$$

Applying the formula yields:

Itemset	Confidence	
{battery} -> {sunscreen}	0.25	
{battery, sunscreen} -> {sandals}	**0.4**	**Highest Confidence = Most interesting**

d) To determine interesting rules we look at their confidence and to determine which rules are coincidental we look at their lifts:

Chap5_4d.xlsx

Chapter 6

1) The non-normality of the outcome variable does not necessarily violate any linear regression model assumptions. The only normality assumption applies to the distribution of the error terms. Depending on the choice of input variables and how well they help to estimate the expected value along the regression line, the normality assumption of the error terms may be justifiable. However, if the normality assumption of the error terms is not justified, transformations of the outcome variable or input variables may be prudent as well as a new linear model parameterization or the introduction of additional input variables.

2a) For a categorical variable with n values, only n-1 binary variables would need to be included in the regression model to account for the n possible values. The reason is that the n-1 binary variables would indicate which of the n-1 values corresponds for a given data record. When the remaining *n*th value is appropriate, the n-1 binary variables would be set to 0. Thus, the contribution of the *n*th value would be imbedded in the intercept term. Thus, the *n*th value is often called the reference case, since the impact of the other n-1 values in the regression model will adjust the intercept term appropriately. See the U.S. states example starting on page 170.

2b) Using n binary variables would be problematic since there would be no unique solution for the estimates of the corresponding coefficients and the intercept. For any change in the intercept, a corresponding change to the coefficient estimates for the binary terms would occur.

3) When using another state as the reference case, the contribution of that state to the expected value of the linear regression model would be included in the value of the intercept estimate. Of course, the intercept estimate would also be adjusted to account for the contribution of the new binary input variable for Wyoming.

4) Logistic regression can be used as a classifier by determining a probability threshold at which one of the binary outcomes is assigned. For example, if the outcome is either pass (y=1) or fail (y=0), a standard threshold would be 0.5, thus for a given set of values for the input variables, if the estimated probability is 0.5 or greater, a classification of pass would be assigned to the observation, otherwise a

classification of fail would be assigned. Of course, as illustrated in the churn example, a threshold other than 0.5 can be selected depending on which classification events (false positives or false negatives) are least desirable.

5) The ROC curve is based on the True Positive Rate and the False Positive Rate. However, these rates correspond to a threshold value in [0, 1]. By plotting these rates against various threshold values in [0,1], a tradeoff can be made between correctly identifying most positive outcomes vs. incorrectly classifying many negative outcomes as positive.

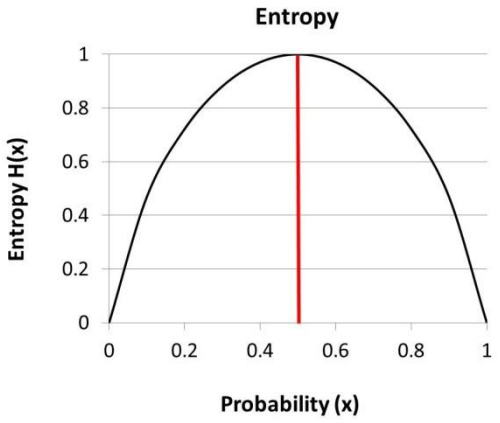
6a) If the probability of an outcome is 0.4, the odds ratio is: .4/(1-.4) = .4/.6 = 2/3

6b) If the probability of an outcome is 0.4, the logs odds ratio is: ln(.4/(1-.4)) = ln(.4/.6) = ln(2/3) = -0.41

7) Referencing equation (6-10), for a one unit increase in the value of x_3 , the log odds ratio would change by -0.5, the corresponding estimated coefficient value. Thus, the odds ratio would change by a multiplicative factor of exp(-0.5) or about 0.61.

Chapter 7

1) In binary classification the possible values of entropy fall within [0,1]. The minimum value of 0 is achieved when the probability P(x) is either 0 or 1. The maximum entropy value of 1 is achieved when P(x) = .5. This can be demonstrated by expressing equation (7-1) in terms of P(X) and 1-P(X) and finding the root of its first derivative.



2) When deciding which attributes to split on, a decision tree algorithm will choose the most informative attributes which is determined by the attribute with the greatest information gain as defined by equation (7-3).

3) Per Bayes' Theorem:

$$\mathrm{P(C|A)} = \mathrm{P(A|C)} * \frac{P(C)}{\mathrm{P(A)}}$$

Where:

C = Having Swine Flu
A = Testing Positive for Swine Flu

Since
$\mathrm{P(A)} = \mathrm{P(A \cap C)} + \mathrm{P(A \cap \neg C)}$

$= \mathrm{P(C)} * \mathrm{P(A|C)} + \mathrm{P(\neg C)} * \mathrm{P(A|\neg C)}$

Using the data provided:

$$\mathrm{P(A)} = (.0002) * (.99) + (.9998) * (.01) = 0.010196$$

$$\mathrm{P(C|A)} = \mathrm{P(A|C)} * \frac{P(C)}{\mathrm{P(A)}}$$

$$\mathrm{P(C|A)} = (.99) * \frac{(.0002)}{(0.010196)} = 0.0194$$

The probability of John having Swine Flu given a positive result is 1.94%

4) The naïve Bayes assumption of the conditional independence of each a_i in equation (7-12) allows the probabilities in equation (7-14) to be calculated in a straightforward manner which is computationally efficient. The naïve Bayes classifier is simple to implement even without special libraries. The calculations are based on simply counting the occurrences of events, making the entire classifier efficient to run while handling high-dimensional data.

5) The data science team should consider using decisions trees. Decision trees are robust to redundant, correlated and non-linear variables and handle categorical variables with multiple levels.

6) Since the probabilities are of interest, naïve Bayes classifiers are excluded. Since most of the variables are continuous, a logistic regression model could be built by omitting the correlated variables or transforming the correlated input variables in some way.

7) Please note that the values are swapped in the Total column as provided in the textbook.

			Predicted Class	
		Good	Bad	Total
	Good	671 (TP)	29 (FN)	700
Actual Class	Bad	38 (FP)	262 (TN)	300
Total		709	291	1000

TPR = TP/(TP + FN) = 671/(671+29) = 95.9%

FPR = FP/(FP + TN) = 38 /(38 + 262) = 12.7%

FNR = FN/(TP + FN) = 29/(671 + 29) = 4.1%

Chapter 8

1) The use of autocorrelation eliminates the concerns with the choice of units and the magnitude of the values. The autocorrelation can be considered a normalized covariance where the resulting values will be between -1 and 1.

2) Suppose, for a random variable X, the distribution of X is symmetric around zero. Then E[X] = 0. Define a random variable Y as $Y = X^2$. Since the value of Y depends on X, the random variables X and Y are not independent. However,

$$cov(X,Y) = E[XY] - E[X]E[Y] = E[X \cdot X^2] - E[X]E[Y] = E[X^3] - 0 \cdot E[Y] = E[X^3] = 0$$

since the distribution of X^3 will also be symmetric around zero.

Alternatively, Let P(X=-1) = 0.25, P(X=0) = 0.50, and P(X=1) = 0.25. Define a random variable Y as $Y = X^2$. Then,

i. Determine the density of Y.
ii. Determine the joint density of (X,Y)
iii. Find at least one example to illustrate that X and Y are not independent, i.e. $P(X,Y) \neq P(X)P(Y)$
iv. Determine the density of XY.
v. Compute the covariance of X and Y.

3) Fitting appropriate ARIMA models to the respective datasets

a.

b.

c.

4) When data needs to have any trends removed, differencing is applied. For simple linear trends, differencing once (d=1) is typically sufficient to provide a stationary series to fit an appropriate ARMA model. For non-linear trends additional differencing (d>1) may be necessary.

Chapter 9

1) The main challenges of text analysis include: high dimensionality due to the number of possible words, the various structures and formats in which the text may be provided, determining the meaning of words, and determining when to treat similar or variations of a word as the same word.

2) A corpus is a large collection of text documents.

3) Common words such as *a, and, of* are called **stop words**. Such words are often ignored when performing text analysis such as TFIDF.

4) Term Frequency (TF) tells us how frequently a word may be used in a document, but does not provide any indication of how unique that word is across a set of documents.

5) The caveat of Inverse Document Frequency (IDF) is that a word appearing once or 100 times in a document, would still have the same IDF. The TFIDF accounts for the frequency of a word in a document and weights it by the IDF.

6) Three benefits of TFIDF are:

- It is a measure that considers the prevalence of a term within a document (TF) and the scarcity of the term over the entire corpus (IDF).
- TFIDF is easy and straightforward to calculate
- TFIDF does not requiring any understanding of the meaning of the text

7) Classification methods such as naïve Bayes, maximum entropy, and support vector machines are often applied to sentiment analyses.

8) A topic is formally defined as a distribution over a fixed vocabulary of words. (see page 274)

9) The tradeoff between precision and recall is that, in general, by returning more documents in order to improve recall, precision may suffer. Conversely, precision could be 1.0, but only a few (say < 10%) of the relevant documents could have been identified. (see page 281)

10) Please see the following link for performing LDA topic modeling:

https://shuyo.wordpress.com/2013/07/24/python-implementation-of-labeled-lda-ramage-emnlp2009/

11) Please see the following link for performing sentiment analysis on Tweets in R:

https://sites.google.com/site/miningtwitter/questions/sentiment/sentiment

Chapter 10

1) There are many resources for finding uses cases and implementations of Hadoop. Here are a couple of relevant links:
http://conferences.oreilly.com/strata/big-data-conference-ca-2015/public/schedule/proceedings

http://wiki.apache.org/hadoop/PoweredBy

2) Use cases can be found on multiple internet resources. Here are some aspects that may be compared and contrasted.

Aspect	Hadoop (MapReduce)	Pig	Hive	HBase
Access	Batch	Batch	Batch	Real-time
Storage	HDFS	HDFS	HDFS	HDFS and RAM
Programming Style	Object Oriented	Script	SQL	CRUD
Relative Ease of Programming	Difficult	Medium	Easy	Medium
Key "Item"	Map & Reduce functions	Relations	Tables	A table

3) MapReduce:
Build jar file and submit M/R job similar to
hadoop jar wc.jar WordCount input/lab1 output/WCoutput

4) Pig:

5) Hive:

Chapter 11

1) For an ARIMA(0,1,1),with no constant term, applied to time series $\{x_t\}$,

$$x_t - x_{t-1} = \varepsilon_t + \theta\varepsilon_{t-1}$$

Given x_t and $E[\varepsilon_t] = 0$, the forecast for x_{t+1} will be recursively defined based on the observed residual as:

$$\hat{x}_{t+1} = x_t + \hat{\varepsilon}_{t+1} + \theta\hat{\varepsilon}_t = x_t + 0 + \theta(x_t - \hat{x}_t) = (1+\theta)x_t - \theta\hat{x}_t$$

Let $\alpha = 1 + \theta$. Then

$$\hat{x}_{t+1} = \lambda x_t - (\alpha - 1)\hat{x}_t = \alpha x_t + (1-\alpha)\hat{x}_t$$

With exponentially weighted moving average (EWMA), the forecast for time t+1 will be the smoothed value at time t. Thus, rewriting the above equation in terms of smoothed EWMA values:

$$EWMA_t = \alpha x_t + (1-\alpha)EWMA_{t-1}$$

2) To demonstrate that the weights decay exponentially in time, expand the recursive representation of EWMA

$$EWMA_t = \alpha y_t + (1-\alpha)EWMA_{t-1}$$

as

$$EWMA_t = \alpha y_t + (1-\alpha)[\alpha y_{t-1} + (1-\alpha)EWMA_{t-2}] = \alpha y_t + \alpha(1-\alpha)y_{t-1} + (1-\alpha)^2 EWMA_{t-3}$$

$$= \cdots = \sum_{i=1}^{t} \alpha(1-\alpha)^{t-i}\, y_t$$

Thus, the weights can be expressed as:

$$w_i = \sum_{i=1}^{t} \alpha(1-\alpha)^{t-i}$$

For $0 < \alpha < 1$, the weights will decay exponentially, in a geometric progression, as the previous weight is multiplied by the value,$1 - \alpha$, which is also between 0 and 1, to obtain the next weight. As the length of the series increases, the weights on the oldest terms in the series will asymptotically approach zero.

3) Using Greenplum,

4) Records from a table could be selected for several reasons. First, a small subset of records can be selected to minimize the about of data that must be processed during development and testing. Of course, this may result in some performance issues during production. Second, a dataset could be randomly split into a training set and a test set. Similarly, some machine learning techniques, such as random forests, require repeated random selection from a dataset to train a model. Two approaches to randomly select records are provided. The use of the ORDER BY clause may be impacted by large dataset. However, the use of random() in Greenplum in the WHERE clause will only result in a variable number of records in repeated random sampling.

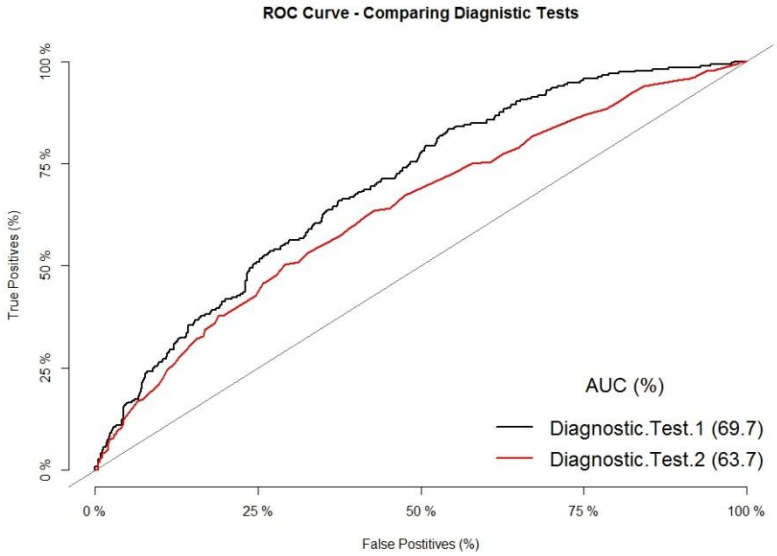
Chapter 12

1) The four common deliverables for an analytics project include:

- **Presentation for Project Sponsors** contains high-level takeaways for executive-level stakeholders, with a few key messages to aid their decision-making process. Focus on clean, easy visuals for the presenter to explain and for the viewer to grasp.
- **Presentation for Analysts** describes changes to business processes and reports. Data scientists reading this presentation are comfortable with technical graphs (such as Receiver Operating Characteristic [ROC] curves, density plots, and histograms) and will be interested in the details.
- **Code** is for technical people, such as engineers and others managing the production environment
- **Technical specifications** for implementing the code

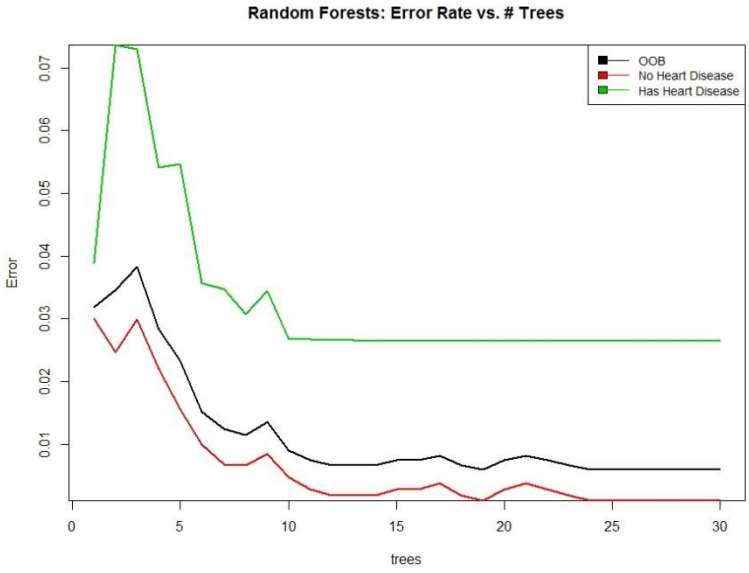
2) For a project sponsor, show that the team met the project goals. Focus on what was done, what the team accomplished, what ROI can be anticipated, and what business value can be realized. Give the project sponsor talking points to evangelize the work. Remember that the sponsor needs to relay the story to others, so make this person's job easy, and help ensure the messaging is straightforward by providing a few talking points. Find ways to emphasize ROI and business value, and mention whether the models can be deployed within performance constraints of the sponsor's production environment. The technical aspects of the modeling and the diagnostics should not be included in the presentation for a project sponsor; the focus should be on business outcomes.

3)



The chart shown can be used by analysts to demonstrate the predictive power of two diagnostic tests. Appropriate labels are present along with a legend that clearly demonstrates that one test outperforms another as indicated by the Area Under the Curve (AUC). The AUC can be used to explain how the better performing diagnostic test can be leveraged to produce a better outcome.

A data scientist would not have an issue understanding what this chart is saying; in fact Receiver Operating Characteristic curves (ROC) are used quite extensively by the data scientist to measure the performance of models.



The chart shown demonstrates the error rate of a Random Forests* model as a function of the categorical levels (**Has Heart Disease**, **No Heart Disease**). An analyst could leverage this graph to demonstrate the accuracy of the Random Forests model broken out by the categorical levels. If presenting to other analysts, focus more time on the methodology and findings. Analysts can be more expansive in describing the outcomes, methodology and the analytical experiment with a peer group, as they will be more interested in the techniques, especially if you developed a new way of processing or analyzing data that can be reused in the future or applied to similar problems.

The data scientist would intuitively understand what this graphic is saying without the need for an analyst to explain.

* Random Forests extends the decision tree approach by creating a "forest" of decision trees. With this approach many decision trees are used to predict an outcome

4) For data that is changing over time the best graphical representation is the line chart. Line charts make it easier to spot trends in data. Also, the line chart works well because time data tends to have a lot of data points and a line connecting the successive points is often the best representation. Line charts are right up there with bars and pies as one of the most frequently used chart types.

5) The Business Intelligence Analyst would receive the presentation that falls into the "Presentation for Analysts" bucket. This form of presentation describes changes to business processes and reports which could have an impact on the dashboard he manages. The infographic below demonstrates some fundamental facets of a presentation for both the Analyst and the Project sponsor.

Presentation Component	Project Sponsor Presentation	Analyst Presentation
Project Goals	List top 3–5 agreed-upon goals.	
Main Findings	Emphasize key messages.	
Approach	High-level methodology	High-level methodology Relevant details on modeling techniques and technology
Model Description	Overview of the modeling technique	
Key Points Supported with Data	Support key points with simple charts and graphics (example: bar charts).	Show details to support the key points. Analyst-oriented charts and graphs, such as ROC curves and histograms Visuals of key variables and significance of each
Model Details	Omit this section, or discuss only at a high level.	Show the code or main logic of the model, and include model type, variables, and technology used to execute the model and score data. Identify key variables and impact of each. Describe expected model performance and any caveats. Detailed description of the modeling technique Discuss variables, scope, and predictive power.
Recommendations	Focus on business impact, including risks and ROI. Give the sponsor salient points to help her evangelize work within the organization.	Supplement recommendations with implications for the modeling or for deploying in a production environment.