LOGIN

Salman_Ahmed_asking_what_is_water

Personal Blog, Mind Rapture, Freedom of Ideas

« Back to blog

AIDA 181 - Big Data Analytics for Risk and Insurance

Chapter 1 - Exploring Big Data Analytics

Technology has transformed our lives in how we communicate, learn and do business. The convergence of **big data and technology has started a transformation of the property-casualty insurance business**.

Traditionally the ability of insurers to provide coverage for a variety of risk exposures is based on the **law of large numbers using loss histories, by which they could reasonably predict cost of future claims**.

Insurers and risk managers have vast quantities of internal data they have not used. The utilization rate is very low.

Big Data : Sets of data that are too large to be gathered and analysed by traditional methods.

Law of Large numbers : a mathematical principle stating that as the number of similar but independent exposure units increases, the relative accuracy of predictions about future outcomes (losses) also increases. *provided the condition wherein those future events occur remain same*.

Big Data and Technology are central to the future of the insurance industry. Communication between those who perform the daily work of an insurer and those designing data scientists data analytics is important for success.

Data-driven decision making has been proved to produce better business results that other types of decision making. ***Insurers can benefit from a framework they can use to approach problems through data analysis.***

!!! Claims or Underwriting professionals typically study data analytics to be able to communicate with data scientists in their organizations. and understand their results!!!

Big Data is mostly available in two forms:

1. Data from internal source : Data that is owned by an organization, can include traditional and non-traditional data. Example: **Text mining** and **Data Mining**
2. Data from external source : Data that belongs to an entity other than the organization that wishes to acquire and use it. Example: social media, statistical plan data, other aggregated insurance industry data, competitor's rate filings, **telematics (driving patterns)** and **economic data** and **geodemographic data** and **IOT (Internet of Things)**

Economic Data : Data regarding interest rates, asset prices, exchange rates, the Consumer Price Index, and other information about the global, the national, or a regional economy

Geodemographic data : Data regarding classification of a population.

Data science is especially useful to gather, categorize, and analyze unstructured data. new techniques to analyze, explore data. Even if the results may be useful or not

There is sometimes a burring of the boundary between internal and external data. Example, data from telematics is obtained from

a device that is installed on a customer's vehicle as the device is owned by an insurance company and installed in a vehicle owned by the customer. Or wearable sensor used by an employee used while working.

As the amount of data has increased exponentially referred to as **Big Data**. Two types of advanced analysis techniques have been made "**DATA MINING**" and "**TEXT MINING**".

- **Text Mining** : Obtaining information through language recognition. Example : can be used to analyse claims adjusters' notes to **identify fraud indicators**.
- **Data Mining** : The analysis of large amounts of data to find new amounts of data to find new relationships and patterns that will assist in developing business solutions.This technique can be used to identify previously unknown factors that are common to an insurer's most profitable auto insurance customers. Also find innovative ways to market to those customers.
- **Telematics** : Evolving source of big data. The use of technological devices in vehicles with wireless with wireless communication and GPS tracking that transmit data to businesses and GPS tracking that transmit data to businesses or government agencies; some return information for the driver. This technique can be used to find driving patterns and change premiums dynamically. *The information obtained from telematics can also be applied to an insurer's online application and selection process*. ***Telematics is also an example of structured external data***.

Internet of Things (IOT) : An emerging technology where a network of objects that transmit data to computers. It is similar to Telematics. Can be used for a new type of process lie Nano-Technology as its eventual effects are unknown. IOT can help insurers monitor risks associated with this process. the potential also involves **machine-to-machine communication** and **machine learning**. **It is also a source of rapidly growing data source.**

Drones is a technology that can assist adjusters in evaluating claims after catastrophes

Machine Learning : Artificial intelligence in which computers continually teach themselves to make better decisions based on previous results and new data. ***Machine learning can be applied over time to refine a model to better predict results.***

Artificial Intelligence (AI) : computer processing or output that simulates human reasoning or knowledge. For example: In claim fraud analysis claim adjusters can use AI techniques to recognize voice patterns that may indicate the possibility of fraud.

Data Science : An interdisciplinary field involving the design and use of techniques to process very large amounts of data from a variety of sources and to provide knowledge based on the data. Data scientists have interdisciplinary skills in mathematics, statistics, computer programming, and sometimes engineering that allows them to perform data analytics on big data. *Data science also provides techniques to use non-traditional internal data*. *Data science is a new field that arose from the need to link big data and technology to provide useful information.*

***A major concept of data science is that the data organization and analysis are automated rater that performed by an individual**. However **human evaluation is most critical**; because of the listed reason below;*

- First, analysis performed by computers is not always accurate.
- Second, the automated analysis may be correct but irrelevant to a given business problem.
- And third, just as the technology is rapidly evolving, so are the physical, political, economic, and business environments.

An example of valuable insight which can be found by human evaluation of prediction modelling results is that a well-constructed, well-maintained property in a high-risk area can be a better risk than a poorly maintained one in a moderate-risk area.

For data science to be useful to an insurer or to specific functional area, such as underwriting or claims, it is usually *important to define the business problem to be solved*, such as improving claims or underwriting results. ***In future, data analytics may be used to forecast unforeseen, or "black swan" events. However, this is still an area for exploration rather than application to business decision***.

An insurer's big data consists of both its own internal data and external data. It can be both quantitative and categorical.

In order to be useful it can be both structured or unstructured

The most important reason for risk management and insurance professionals to learn about data analytics is because big data and technology are central to the insurance industry. Insurers and risk managers have vast quantities of internal data they have not used.

Big Data 1.0 : Organization began using the Internet to conduct business and compile data about their customers. Insurers developed online insurance applications. they also used the data from application, as well as claims history, to improve underwriting efficiency and to provide customer information for product development and marketing. *Most insurance companies are at this stage*.

Big Data 2.0 : This stage allows organizations to obtain and aggregate vast amounts of data (such as vehicles, homes and wearable technology) very quickly and extract useful knowledge from it. *Only some companies are actively pursuing at this stage*.

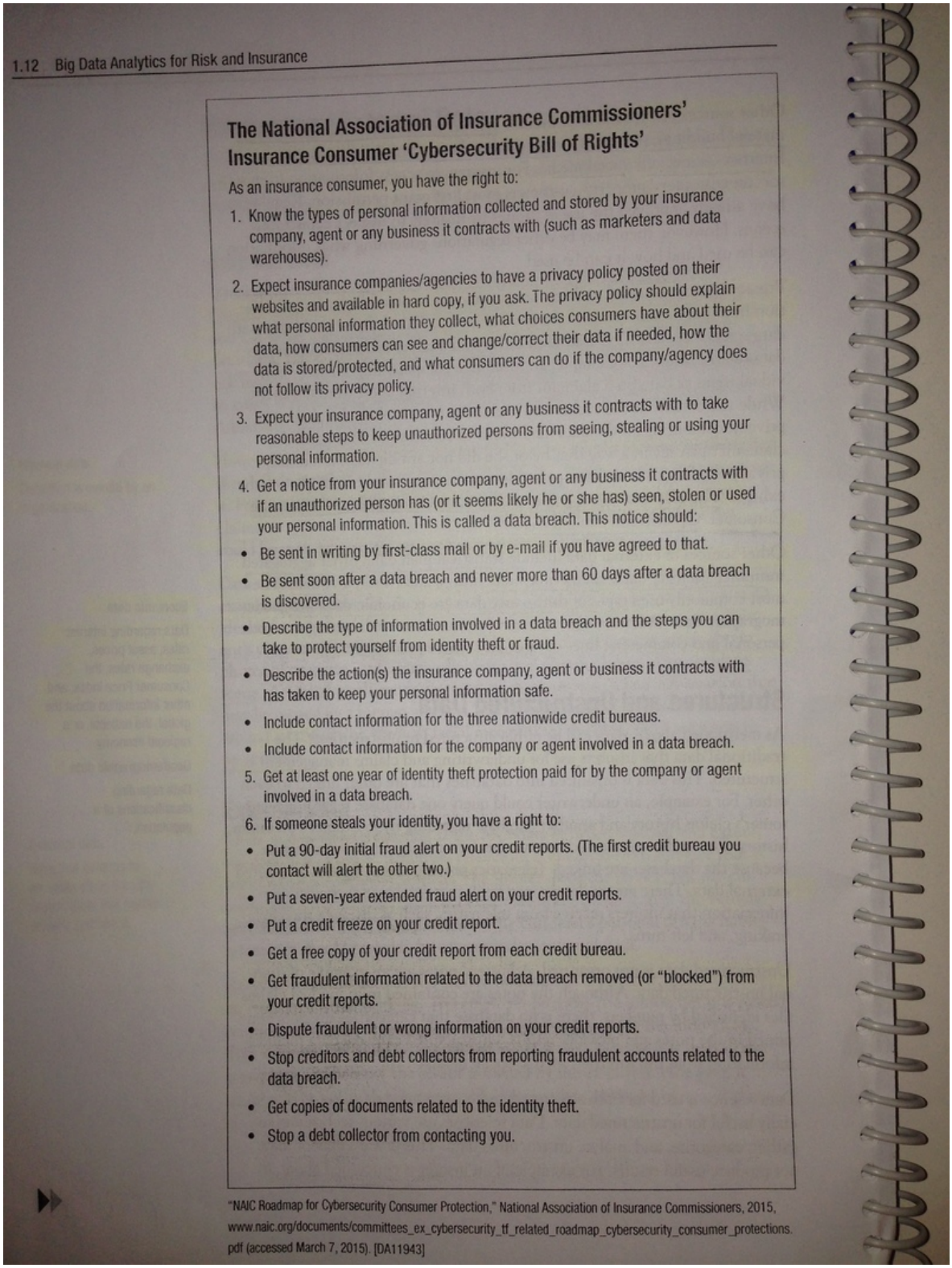
Big Data characteristics and sources

The varieties, volume, and sources of data are rapidly increasing. to better understand big data, these categories will be discussed:

- Data characteristics :
  - volume : size of the data
  - variety : structured and unstructured data
  - velocity : growing rate of change in the types of data
  - veracity : completeness and accuracy of data
  - value : goal of data science to derive value from the results of data analysis to help insurers make better decisions.
- Internal (example: risk factors, losses, premium, rating factors, rates and customer information) and External Data
- Structured data : data growing into databases with defined fields, including links between databases.
- Unstructured data : data that is not organized into predetermined formats, such as databases and often consists of text, images, or other non-traditional media, news reports

!!There is a sometimes a blurring of the boundary between internal and external data.!!

	Structured	**Unstructured**
	Telematics	Social media
External	Financial Data	News reports
	Labor Statistics, Geodemographic Data	Internet Videos
	Policy Information	Adjuster notes
Internal	Claims history	Customer voice records
	Customer Data	Surveillance videos



1.12 Big Data Analytics for Risk and Insurance

The National Association of Insurance Commissioners' Insurance Consumer 'Cybersecurity Bill of Rights'

As an insurance consumer, you have the right to:

1. Know the types of personal information collected and stored by your insurance company, agent or any business it contracts with (such as marketers and data warehouses).
2. Expect insurance companies/agencies to have a privacy policy posted on their websites and available in hard copy, if you ask. The privacy policy should explain what personal information they collect, what choices consumers have about their data, how consumers can see and change/correct their data if needed, how the data is stored/protected, and what consumers can do if the company/agency does not follow its privacy policy.
3. Expect your insurance company, agent or any business it contracts with to take reasonable steps to keep unauthorized persons from seeing, stealing or using your personal information.
4. Get a notice from your insurance company, agent or any business it contracts with if an unauthorized person has (or it seems likely he or she has) seen, stolen or used your personal information. This is called a data breach. This notice should:

- Be sent in writing by first-class mail or by e-mail if you have agreed to that.
- Be sent soon after a data breach and never more than 60 days after a data breach is discovered.
- Describe the type of information involved in a data breach and the steps you can take to protect yourself from identity theft or fraud.
- Describe the action(s) the insurance company, agent or business it contracts with has taken to keep your personal information safe.
- Include contact information for the three nationwide credit bureaus.
- Include contact information for the company or agent involved in a data breach.

5. Get at least one year of identity theft protection paid for by the company or agent involved in a data breach.
6. If someone steals your identity, you have a right to:

- Put a 90-day initial fraud alert on your credit reports. (The first credit bureau you contact will alert the other two.)
- Put a seven-year extended fraud alert on your credit reports.
- Put a credit freeze on your credit report.
- Get a free copy of your credit report from each credit bureau.
- Get fraudulent information related to the data breach removed (or "blocked") from your credit reports.
- Dispute fraudulent or wrong information on your credit reports.
- Stop creditors and debt collectors from reporting fraudulent accounts related to the data breach.
- Get copies of documents related to the identity theft.
- Stop a debt collector from contacting you.

"NAIC Roadmap for Cybersecurity Consumer Protection," National Association of Insurance Commissioners, 2015, www.naic.org/documents/committees_ex_cybersecurity_tf_related_roadmap_cybersecurity_consumer_protections.pdf (accessed March 7, 2015). [DA11943]

Data Producers : are claim adjusters, underwriters, and risk managers - data entered by the should be as accurate and complete as possible

Data Users : professionals who use reports based on data. Example : accident year loss costs for personal auto insurance.

Quality Data is **accurate, appropriate, reasonable, and comprehensive relative to a given use**. **Accurate data** is free from mistakes and can therefore be relied on to produce useful results, because sometimes insurance data is submitted to states and rating bureaus *and should be free from material limitations*.

Reasonable Data has been validated by comparison with outside sources or audited for consistency and accuracy

Comprehensive Data contains the full range of information needed to produce a reliable results or analysis.

Metadata : The data about data that provide context for analyzing transaction facts with efficient structures for grouping hierarchical information. documents the contents of a database, it contains information about business rules and data processing. Example of Metadata is a **Statistical Plan**

!! If an insurer has inaccurate data, it can affect the insurer's financial results and the entire insurance industry. !!

Statistical plan : a formal set of directions for recording and reporting insurance premiums, exposures losses, and sometimes loss expenses to a statistical agent.

An example of metadata for a three-year compilation of accident year incurred losses should address these factors:

- The accident years included in the report
- The range of dates of paid loss transactions
- The evaluation data of the case-based loss reserve transactions
- the definitions of incurred losses (such as paid plus latest loss reserve, inclusive or exclusive of allocated expenses,net of deductible or reinsurance, and so forth).

Metadata can provide criteria for edits or restrictions of data inputs. A best practice is to reject, or at least flag, inputs that fail to meet the criteria.

*!! **Meta data** can provide criteria for edits or restrictions of data inputs. a best practice is to reject, or at least flag, inputs that fail to meet the criteria. It is also useful to include a data quality matrix for each data element that describes the quality checks done on the data element, how frequently the checks are done, and where in the process the checks occur.*

Metadata can be enhanced by documentation.!!

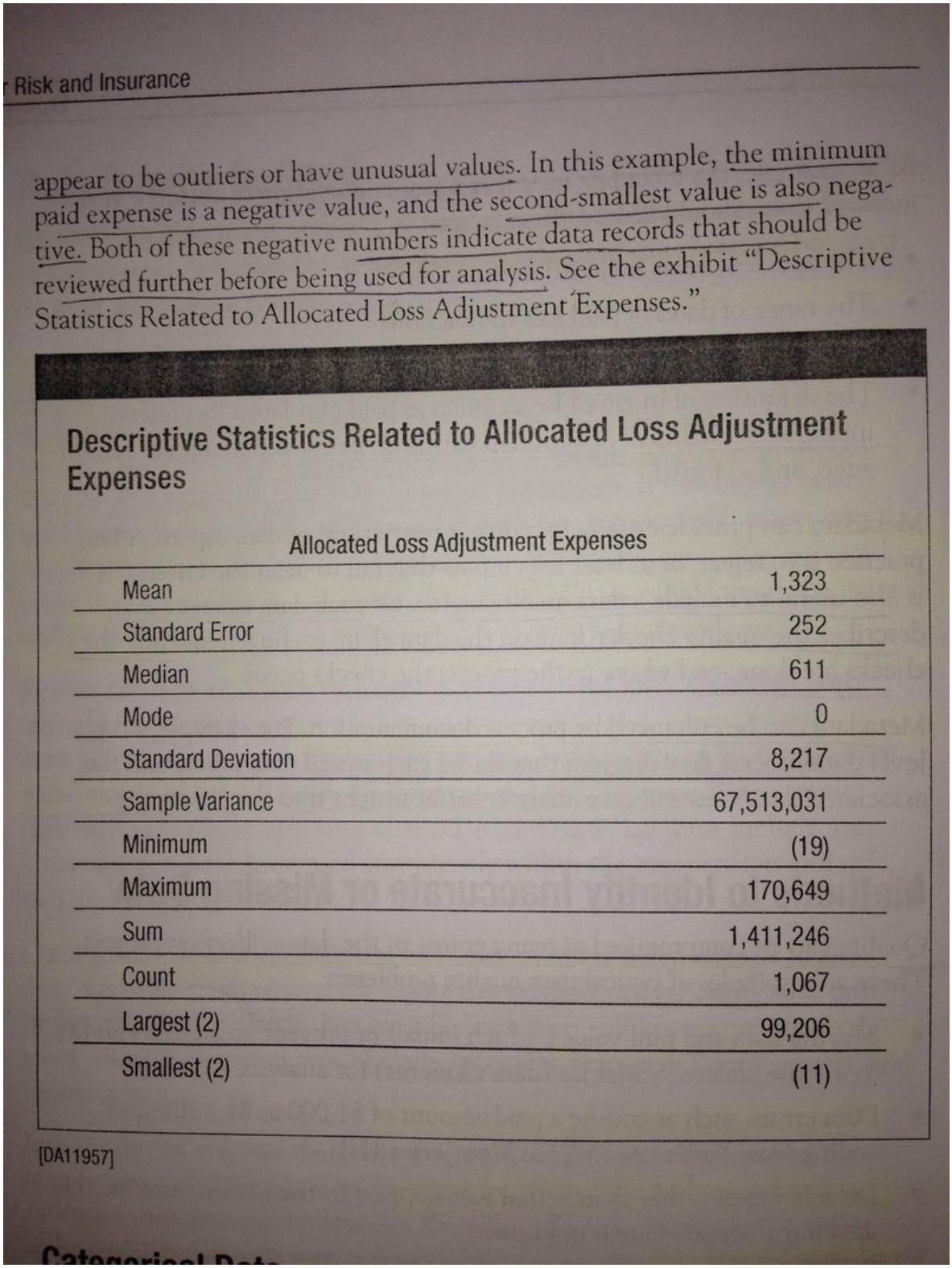
Examples of inaccurate data

1. Missing data and null values
2. Data errors in entering
3. Default values rather than actual values
4. Duplicate transactions

Data can take two forms

1. Quantitative or numerical
2. Categorical or alphabetic

Descriptive Statistics : quantitative summaries of the characteristics of a dataset, such as the total or average. *can also be used to identify missing or inaccurate quantitative data and also outliers*. *The descriptive approach is applied when an insurer or risk manager has a specific problem.*



Risk and Insurance

appear to be outliers or have unusual values. In this example, the minimum paid expense is a negative value, and the second-smallest value is also negative. Both of these negative numbers indicate data records that should be reviewed further before being used for analysis. See the exhibit "Descriptive Statistics Related to Allocated Loss Adjustment Expenses."

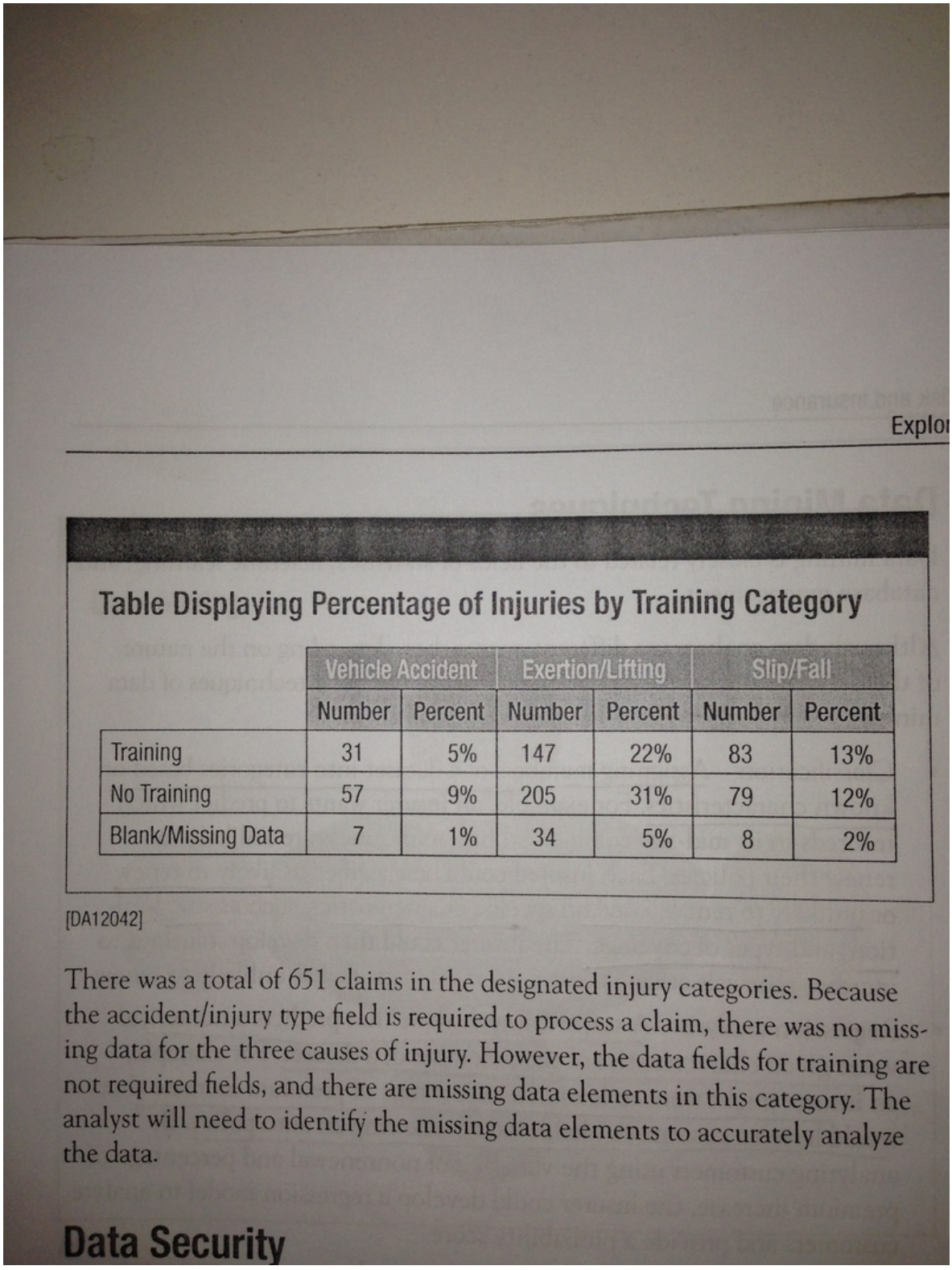
Descriptive Statistics Related to Allocated Loss Adjustment Expenses

	Allocated Loss Adjustment Expenses
Mean	1,323
Standard Error	252
Median	611
Mode	0
Standard Deviation	8,217
Sample Variance	67,513,031
Minimum	(19)
Maximum	170,649
Sum	1,411,246
Count	1,067
Largest (2)	99,206
Smallest (2)	(11)

[DA11957]

Categorical Data

Categorical Data : a multidimensional partitioning of data into two or more categories also called chunks or data cube. These chunks are then organized into tables that will allow the data to be analyzed for unusual values. A **data cube** is a multidimensional partitioning of data into to or more categories. It is sued to determine the percentage of each combination of injury type and training in the accident data for the past year. Data cubes can be used to identify inaccuracies and omissions in categorical data.



Explor

Table Displaying Percentage of Injuries by Training Category

	Vehicle Accident		Exertion/Lifting		Slip/Fall	
	Number	Percent	Number	Percent	Number	Percent
Training	31	5%	147	22%	83	13%
No Training	57	9%	205	31%	79	12%
Blank/Missing Data	7	1%	34	5%	8	2%

[DA12042]

There was a total of 651 claims in the designated injury categories. Because the accident/injury type field is required to process a claim, there was no missing data for the three causes of injury. However, the data fields for training are not required fields, and there are missing data elements in this category. The analyst will need to identify the missing data elements to accurately analyze the data.

Data Security

Data Security : Best practice to safeguard important data on which important decisions are based is to **restrict** the data so that no one can change the data. **Authorized** users' lack of access should be based on their responsibilities. and use cyber security techniques to reduce the likelihood that **malware** will corrupt data in their systems

Malware : Malicious software, such as a virus, that is transmitted from one computer to another to exploit system vulnerabilities in the targeted computer.

Data mining is closely related to the fields of **statistics**, **machine learning** and **database management**.

Statistics : A field of science that derives knowledge from data; it provides a root understanding of useful approaches to data analysis.

Database : A collection of information stored in discrete units for ease of retrieval manipulation, combination or other computer processing.

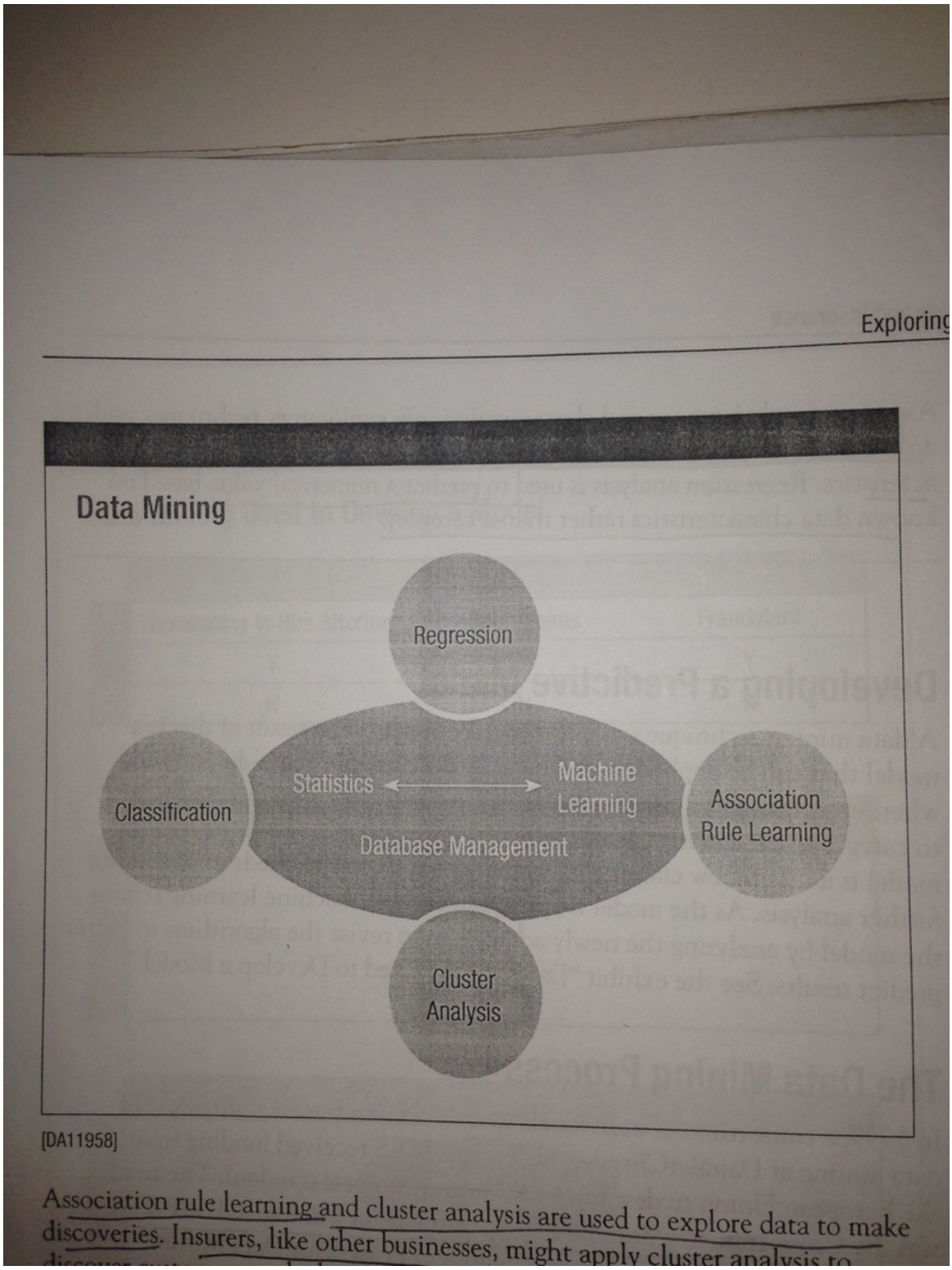
Algorithm : An operational sequence used to solve mathematical problems and to create computer programs.

Basic techniques of data mining :

- **Classification** : assigning members of a dataset into categories based on known characteristics.
- **Regression Analysis** : a statistical technique that predicts a numerical value given characteristics of each member of a dataset.
- **Association Rule learning** : examining how data to discover new and interesting relationships.for these relationships, **algorithms** are used to develop new rules to apply to new data.
- **Cluster Analysis** : using statistical methods, a computer program explores to find groups with common and previously unknown characteristics. *the results of the cluster analysis may or may not provide useful information*.

Association rule learning and cluster analysis (exploratory techniques)are used to explore data to make discoveries. Insurers, like other business, might apply cluster analysis to discover customer needs that could lead to new products. *Unlike classification and regression analysis, there are no known*

characteristics of the data beforehand. The purpose of association rule learning and cluster analysis to to discover relationships and patterns in the data and then determine if that information is useful for making business decisions.

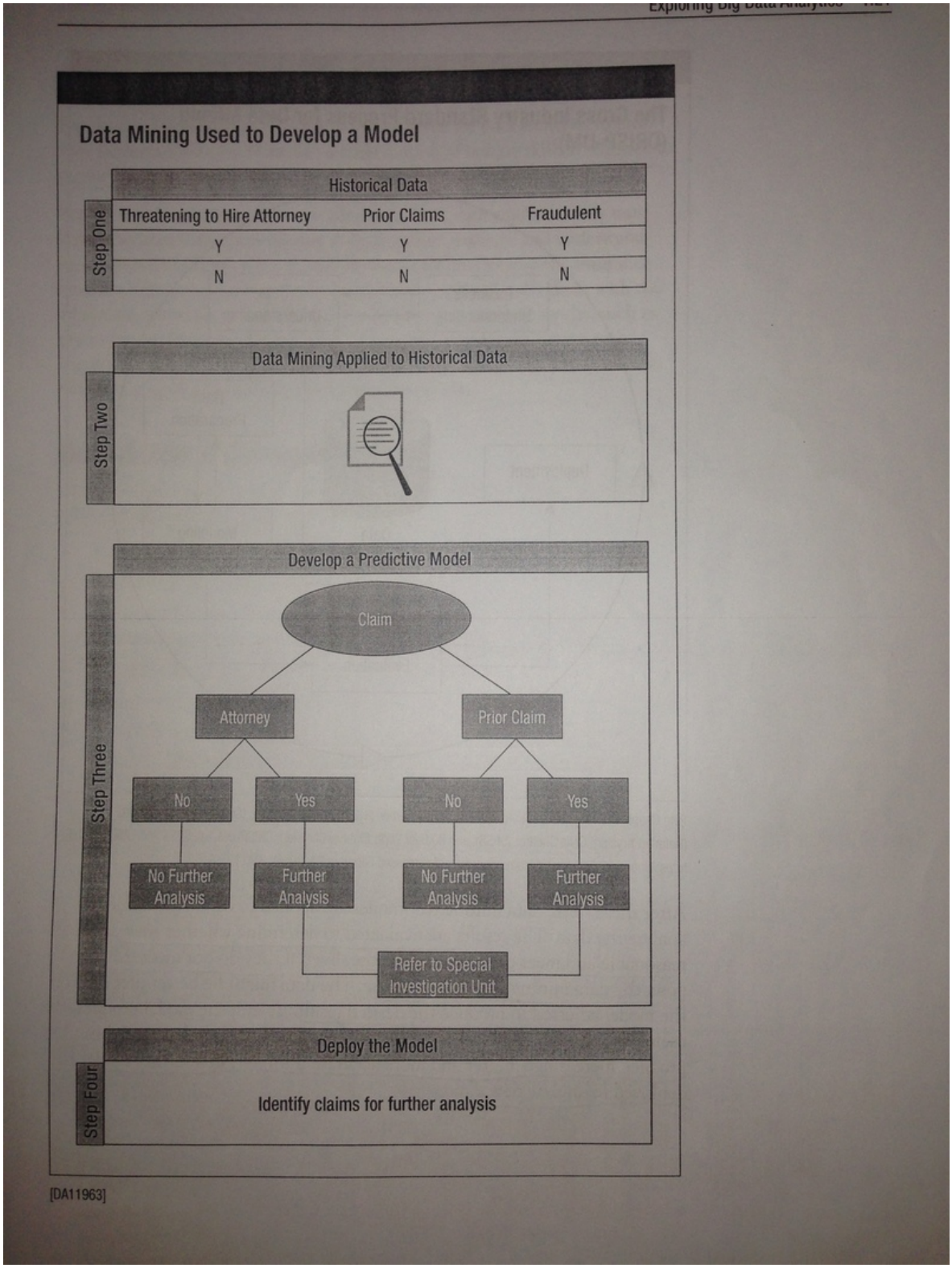


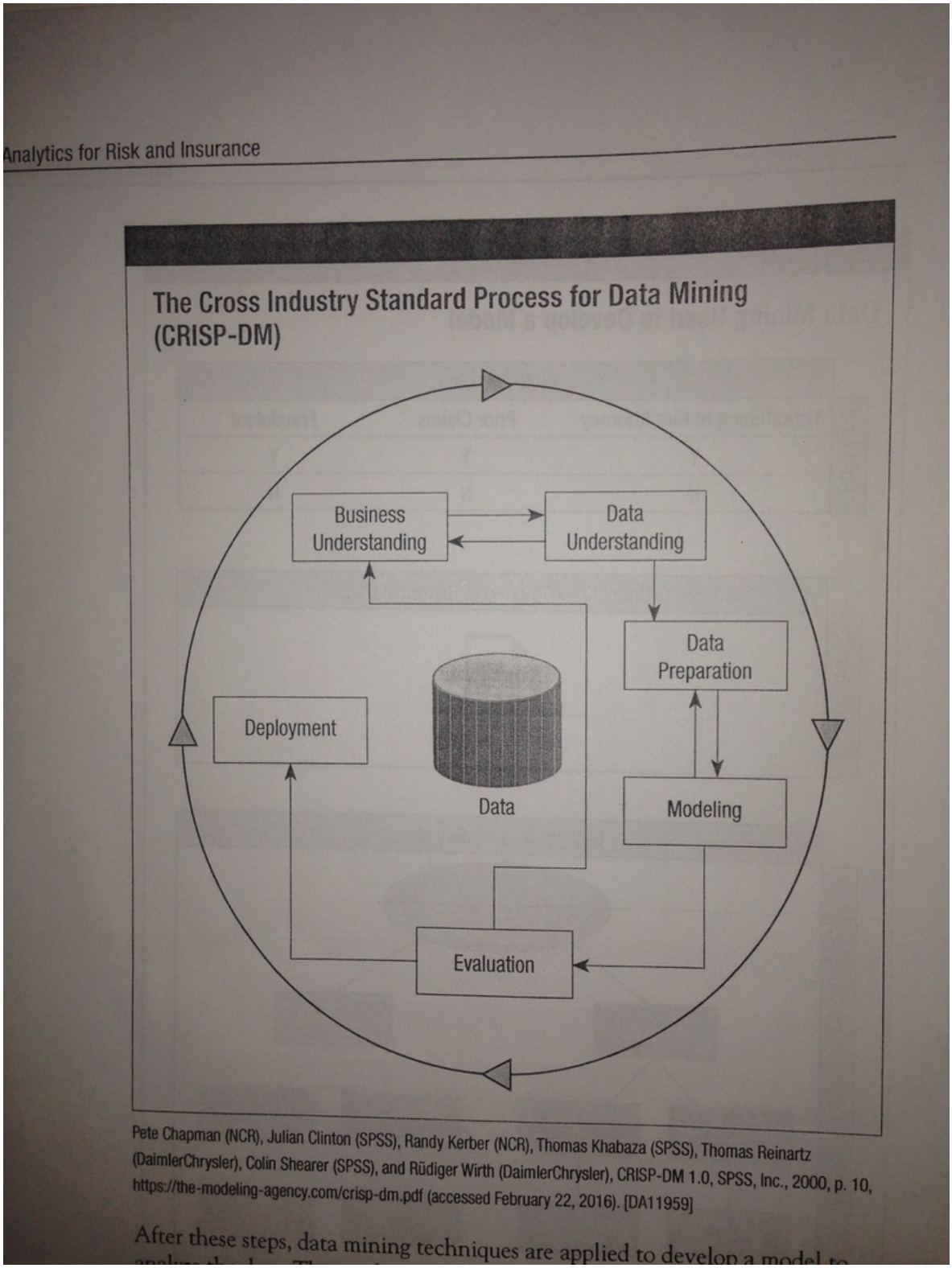
A predictive approach to data analytics involves providing a method to be used repeatedly to provide information.

Cross Industry Standard Process for Data mining (CRISP-DM) : An accepted standard for the steps in any data mining process used to provide business solutions. developed in 1999, a consortium of individuals who worked in the emerging field of data mining at DaimlerChrysler, NCR, and SPSS received funding from the European Union to develop a data mining process standard.

Steps in CRISP-DM

1. to understand what a business want to achieve by applying data mining to or more sets of data
2. types of data hat are being used - internal or external - structured or unstructured
3. cleaning or pre-processing of data
4. data mining techniques are applied to develop a model
5. results are evaluated to determine whether they are reasonable and meet the business objective
6. pre deployment or post deployment of the model, continuous refinement of the model is done to produce increasingly better and more accurate data. the circle the surrounds the diagram, indicate the data mining is a continuous process that involves continuously evaluating and refining the model.


Data Mining Used to Develop a Model
Step One
Historical Data
Threatening to Hire Attorney
Prior Claims
Fraudulent
Y
Y
Y
N
N
N
Step Two
Data Mining Applied to Historical Data
Step Three
Develop a Predictive Model
Claim
Attorney
Prior Claim
No
Yes
No
Yes
No Further Analysis
Further Analysis
No Further Analysis
Further Analysis
Refer to Special Investigation Unit
Step Four
Deploy the Model
Identify claims for further analysis
[DA11963]



A central feature of CRISP-DM is the circle that represents it is an ever evolving process

Data Science :

Is a new field at the frontier of data analytics. It is often experimental, and methods evolve rapidly. It uses the scientific method, which consists of these steps:

- a question or problem is raised
- Research is conducted regarding the subject
- a hypothesis is developed, based on the research, regarding the answer to the question or the cause of/solution to the problem.
- Experiments are performed to test the hypothesis
- Data from the experiments is analysed
- A conclusion is reached

The quest of data science, as with all the sciences, is to increase knowledge of the world and provide more advanced solutions to complex questions and problems.

!!Data science is especially useful for unstructured data.!!

Four fundamental concepts of data science:

1. **Systematic process can be used to discover useful knowledge from data, this framework forms the foundation for data-analytical processes.**
2. **Information technology can be applied to big data to reveal the characteristics of groups of people or events of interest.**
3. **Analyzing the data too closely can result in interesting findings that are not generally applicable**
4. **The selection of data mining approaches and evaluation of the results must be thoughtfully considered in the context on which the results will be applied**

Data Scientist : Traditionally data is usually numerical or categorical. *Data Scientists must be able to analyze traditional data as well as new types*, including texts, geo-location based data, sensor data, images, and social network data. They must also have the skills to manage increasingly large amounts of data.

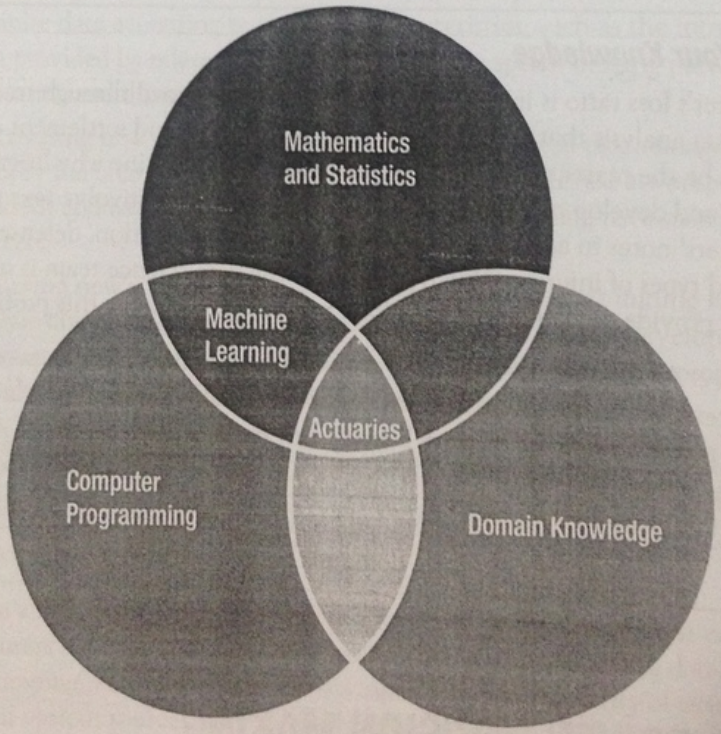
series of examinations in actuarial science. Using the p

analytics and physics, actuaries would be like the traditional physicists. They focus primarily on pricing, ratemaking, and claim reserving.

Data scientists are like the new physicists in this analogy. They are exploring previously undiscovered realms and forces, such as social networks and new technology. The knowledge they obtain may lead to new insurance products and risk management techniques, as well as the refinement of existing products. However, there is no clear division between the roles of actuary and data scientist. Many actuaries are acquiring advanced computer programming skills using new data analysis programming languages, such as R, to supplement their mathematical and statistical knowledge.

Information related to the context for data science is often referred to as domain knowledge. While data scientists may have advanced mathematics, statistics, and computer programming knowledge, they may not have knowledge about insurance or risk management. See the exhibit "Necessary Skills for a Data Scientist."

Necessary Skills for a Data Scientist



[DA11967]

Actuary : A person who uses mathematical methods to analyze insurance data for various purposes, such as to develop insurance rates or set claim reserve. focus primarily or pricing, rate-making or claim reserving.

Actuaries are the insurance professionals who have traditionally analyzed data and made predictions based on their analyses.

however there is no clear distinction between actuary and data scientist

Information related to the context for data science is often referred to as domain knowledge.

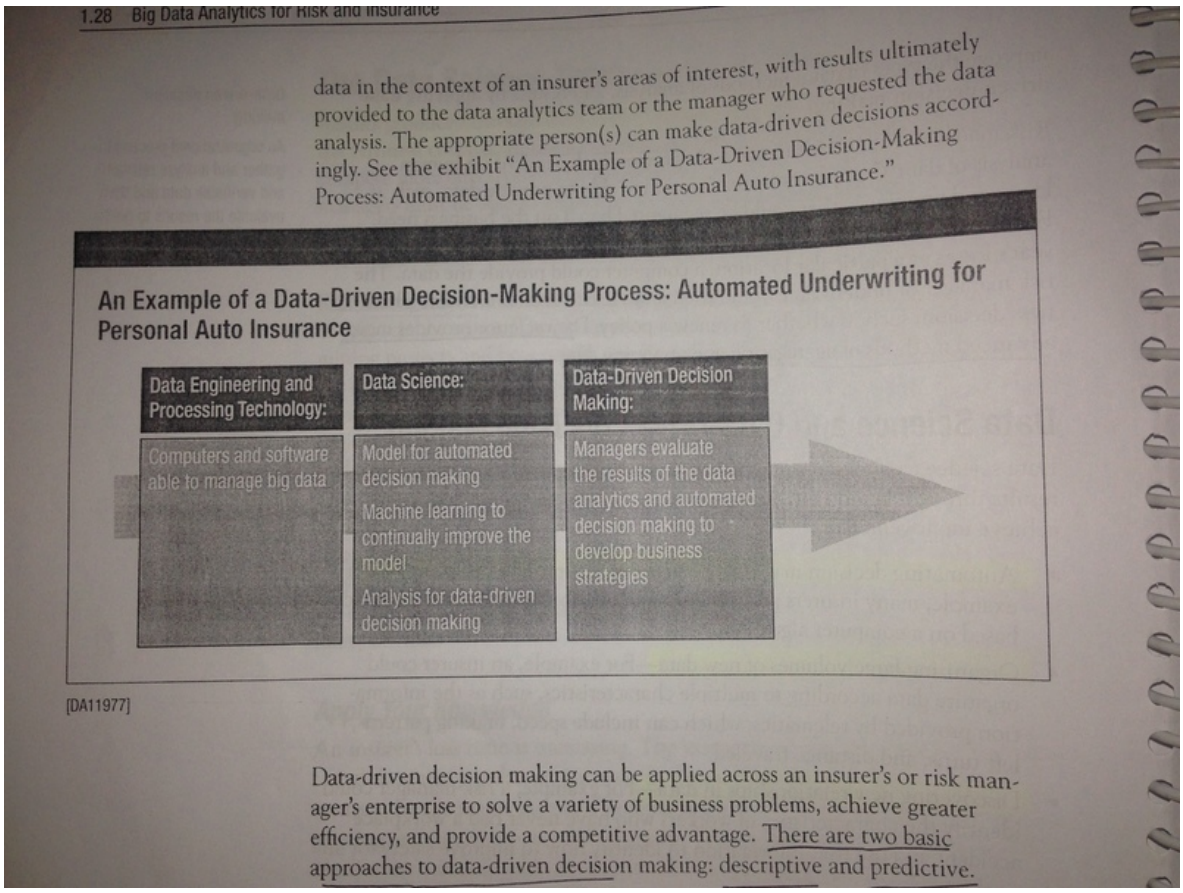
Risk management and insurance professionals can be valuable members of data science teams.

Insurance company uses data driven decision using traditional methods, however data analytics can improve the types of data, methods of analysis and results by applying **data-driven decision modeling**.***Discovering new relationships in data is a way insurers and risk managers can use data science to improve their results through data-driven decision making.***

Data-driven decision making : An organizational process to gather and analyze relevant and verifiable data and then evaluate the results to guide business strategies. there are two approaches predictive approach and descriptive approach

Improvements via data-decision making

1. Automating decision making for improved accuracy and efficiency
2. Organizing large volumes of new data
3. Discovering new relationships in data
4. Exploring new sources of data



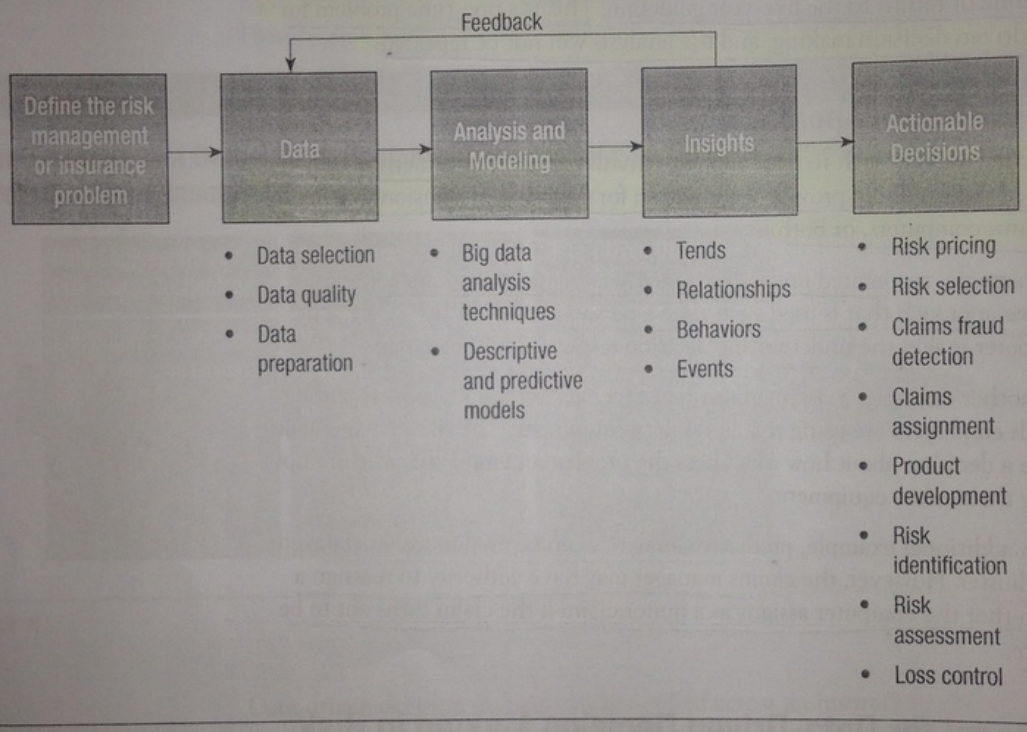
Descriptive Approach : is applied when an insurer or risk manager has a specific problem. Data Science is intended to be used to provide data that will help solve the problem. Insurers or Risk Managers do not continue to use data-driven decision making beyond the specific problem. This is a one-time problem for data-driven decision making, and the analysis will not be repeated.

Predictive Approach : A predictive approach to data analytics involves providing a method that can be used repeatedly to provide information for data-driven decision making by humans, computers or both.

Approach for Analytics

- Analyze the data
- Improve data quality to make it accurate, appropriate, reasonable and comprehensive.
- Select the analytical technique
- Make a data-driven decision

Risk Management and Insurance Data Analytics Decision-Making Model



[DA11978]

variables to be used are date of accident; store location; area of store property in which the accident occurred, such as a parking lot or entrance; weather data; and sales volume at the store on the date of accident. The risk manager runs a report to identify missing or obviously inaccurate data, such as a date of loss stated as January 1, 1898. The risk manager eliminates claims with missing or inaccurate data from the dataset.

The risk manager then meets with the data analytics team to discuss a process that will find any correlations or patterns that might be useful in analyzing the losses and the reason for the increase. Using a cluster technique, the analytics team discovers several patterns in the data: There was a spike in fall or

In a data - driven decision making the below steps are followed:

1. clearly define the problem
2. select appropriate data
3. prepare the data by removing missing or inaccurate data
4. select relevant variables
5. develop a model
6. to discover patterns and correlations in the data

Chapter 2 - Predictive Modeling Concepts

Insurers (underwriters and actuaries) rely on data mining techniques to help them make more informed underwriting, claims, and marketing decisions. Data modeling is the key to transforming raw data into something more useful.

Understanding the basic modeling terms and types of models can help insurance and risk management professionals effectively communicate with data scientists to create models that will benefit the insurance industry.

In the data mining process, modeling is the representation of data; this representation is then used to define and analyze the data. Modeling the data is complex process that uses concepts from both machine learning and statistics. Actuaries and data scientists select the appropriate methods for modeling based on the type of data available and their particular goal in analyzing it.

Below two types of data mining techniques, ***used to find pattern in large datasets***, ***come from the field of machine learning***

1. **Supervised Learning** : A type of model creation, derived from the field of machine learning, in which the target variable is defined, *challenge is that data should be there about the target*
2. **Unsupervised Learning** : A type of model creation, derived from the field of machine learning, that does not have a defined target variable. Can be used to pre-process data into groups before supervised learning is used. Can be used to provide the information needed to define an appropriate target for supervised learning.

    - *A disadvantage is that unsupervised learning can sometimes provide meaningless correlations*
    - *conducting unsupervised learning first may provide the information needed to define an appropriate target for supervised learning*

The below techniques are also used

1. **Predictive Model** : A model used to predict an unknown outcome by means of a defined target variable. *It can predict values in the future past and present*.
2. **Descriptive Model** : a model used to study and find relationships within data. It can help in gaining insight.

Other terms used in modeling techniques:

- **Attribute** : A variable that describes a characteristic of an instance within a model
- **Instance (example)** : the representation of a data point described by a set of attributes within a model's dataset.
- **Target Variable** : The predefined attribute whose value is being predicted in a data analytical/predictive model.
- **Class Label** : The value of the target variable in a model. For example is the variable binary

After an insurance selects the business objectives of its business model and the data that will be analyzed, an algorithm is selected.

Machine learning algorithms can take a number of forms, such as mathematical equations, classification trees and clustering techniques. Experienced actuaries and data scientists have the experience needed to select which algorithm is appropriate for a particular problem.

Information Gain : A measure of the predictive power of one or more attributes. can be thought of in terms of how much if affects the entropy of a given dataset, which is also a indication of how it provides about the target variable.

Entropy : A measure of disorder in a dataset, it is essentially a measure of how unpredictable some is.

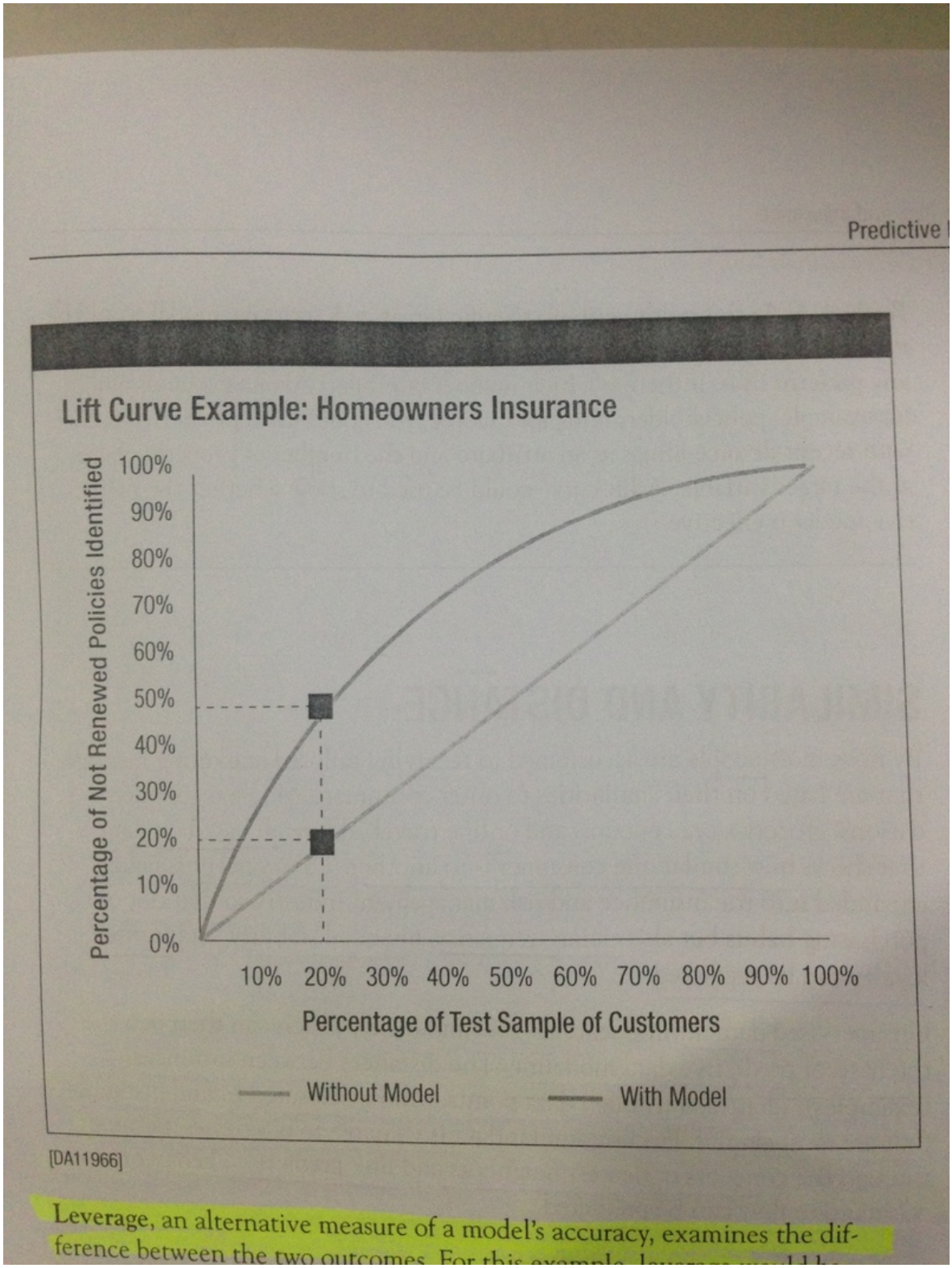
When a dataset is segmented based on informative attributes (those with high information gain). the entropy decreases.

Lift : In model performance evaluation(or value of a model), the **percentage** of positive predictions made by the model divided by the percentage of positive predictions that would be made in the absence of the model. *Helps in determining the value the model brings to the business*.

Example :

Percentage with model : 50 percentage
Percentage without model : 20 percentage

Lift = 0.50/0.20 = 2.5



Leverage : alternate measure of a model's accuracy, examines the difference between two outcomes. In model performance evaluation , the percentage of positive predictions made by the model minus the percentage of positive predictions that would be made in the absence of the model.

Example :

Percentage with model : 50 percentage
Percentage without model : 20 percentage

Leverage = 0.50-0.20 = 0.30

Lift and Leverage is an effective way for insurance and risk management professionals to evaluate predictive models and know how reliable they are.

Similarity and Distance

Unsupervised data mining searches for similarities which can then become the basis of predictive modeling. The distances between instances (examples), identified through data points, can be measured to find the instances that are most popular. Finding similar data is valuable to businesses, because through the concepts of nearest neighbors and link prediction, behaviors and information flow can be predicted.

The objective of data modeling is to find the similarity between data points (instances). **Determining which similarities can be significant can be subjective**.

In a data mining context, similarity is usually measured as the distance between two instances' data points. a small distance indicates a high degree of similarity and a large distance indicates a low degree of similarity.

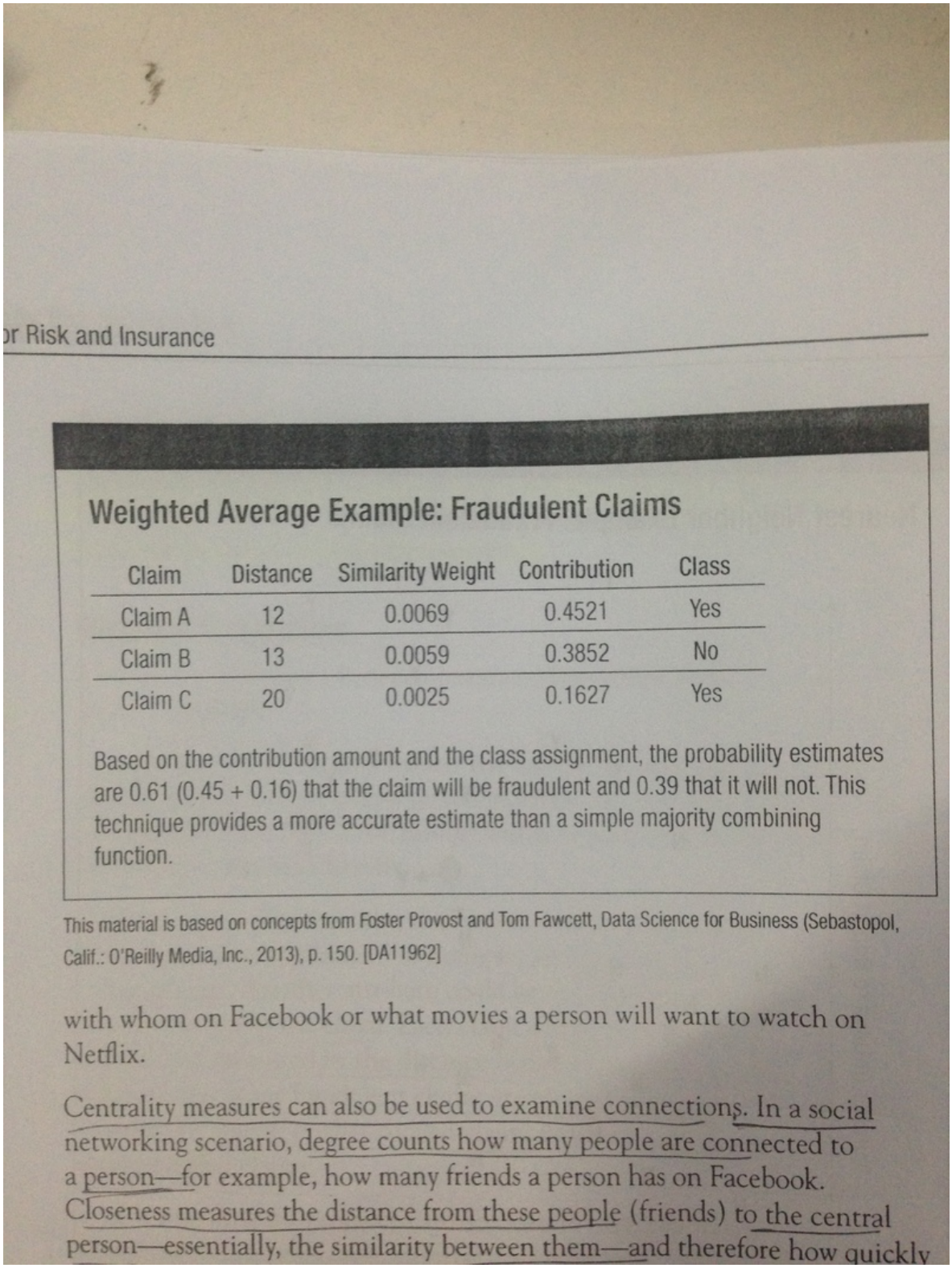
Businesses, including insurers, can apply the concept of similarity and distance to help them understand their customers and predict behaviors. These predictions can be valuable planning resources.

Nearest Neighbor : In data mining concepts, the most similar instances in a data model are called as nearest neighbors.

k nearest neighbor (K-NN) : an algorithm in which 'k' equals the number of nearest neighbors plotted on a graph.

Combining function : The combination of two functions to form a new function. Using a **majority class function**, the majority of the nearest neighbors' values predict the class label of the target variable. Using the majority combining function gives equal weight to all of the nearest neighbors. It does not consider their distance, even if the instance are not an equal distance from the average data point.

To calculate the contributions, each distance amount is squared, and the reciprocal of the square is calculated, resulting in a similar weight. the similarity weights are then weighted relative to each other so that they total 1.00, resulting in a contribution amount. To calculate a probability estimate for each claim, a score of 1.00 is assigned to "YES", and a score of 0.00 is assigned to "NO".



or Risk and Insurance

Weighted Average Example: Fraudulent Claims

Claim	Distance	Similarity Weight	Contribution	Class
Claim A	12	0.0069	0.4521	Yes
Claim B	13	0.0059	0.3852	No
Claim C	20	0.0025	0.1627	Yes

Based on the contribution amount and the class assignment, the probability estimates are 0.61 (0.45 + 0.16) that the claim will be fraudulent and 0.39 that it will not. This technique provides a more accurate estimate than a simple majority combining function.

This material is based on concepts from Foster Provost and Tom Fawcett, Data Science for Business (Sebastopol, Calif.: O'Reilly Media, Inc., 2013), p. 150. [DA11962]

with whom on Facebook or what movies a person will want to watch on Netflix.

Centrality measures can also be used to examine connections. In a social networking scenario, degree counts how many people are connected to a person—for example, how many friends a person has on Facebook. Closeness measures the distance from these people (friends) to the central person—essentially, the similarity between them—and therefore how quickly

Applications of Similarity and Distance
Steps for Data scientists and Insurers to apply Nearest Neighbors Algorithm

1. compile a list of general attributes which indicate a desirable risk
2. using a **majority combining function**, the majority of nearest neighbors's values predict the class label of the target variable

3. to calculate the weighted average

Majority Combining function : the majority of nearest neighbors' values predict the class label of the target variable, it gives equal weight to all of the nearest neighbors. It does not consider their distance.

A majority combining function gives equal weight to all of the nearest neighbors, while a weighted average weights the nearest neighbors' contributions by their distance.

Measuring similarity in Networks

Similarity does not always have to be measured as the distance on a graph. It can also be examined through network connections. Insurers can examine social networks and gain information about their customers based on their similarities.

In **Link Prediction**, a model attempts to predict a pair of instances. It does this by measuring the similarities between the two instances. It does this by measuring the similarities between the two instances.

Centrality measures can also be used to measure connections. **In a social networking scenario**, **degree counts** how many people are connected to a person.

Closeness measures the distance from these people (friends) to the central person - essentially, the similarity between them - and therefore how quickly information will travel between them.

Between measures the extent to which a person connects others. for example in social network scenario, between friends are considered to have a high degree of betweenness.

Training and Evaluating a Predictive Model

Implementation of predictive modeling can improve an insurer's consistency and efficiency in marketing, underwriting, and claims

services by helping to define target markets, increasing the number of policy price points and reducing claims fraud. Understandable, business are cautious about relying on predictive models. therefore, during the development of predictive models, business must be able to assess the model's **effectiveness (specificity) and reliability (sensitivity)** .

Organizations can use the models to determine the likelihood of risk.

Training Data : Data that is used to train a predictive model and that therefore must have known values for the target variable of the model. the selection of attributes to use in a predictive model determines the model's success, and the selection must often be fine-tuned several times. The model must be made complex enough to be accurate. However if too many attributes are used, a model can easily overfit the training data.

Overfitting : The process of fitting a model too closely to the training data for the model to be effective on other data.

Holdout Data : In the model training process, existing data with a known target variable that is not used as part of the training data.

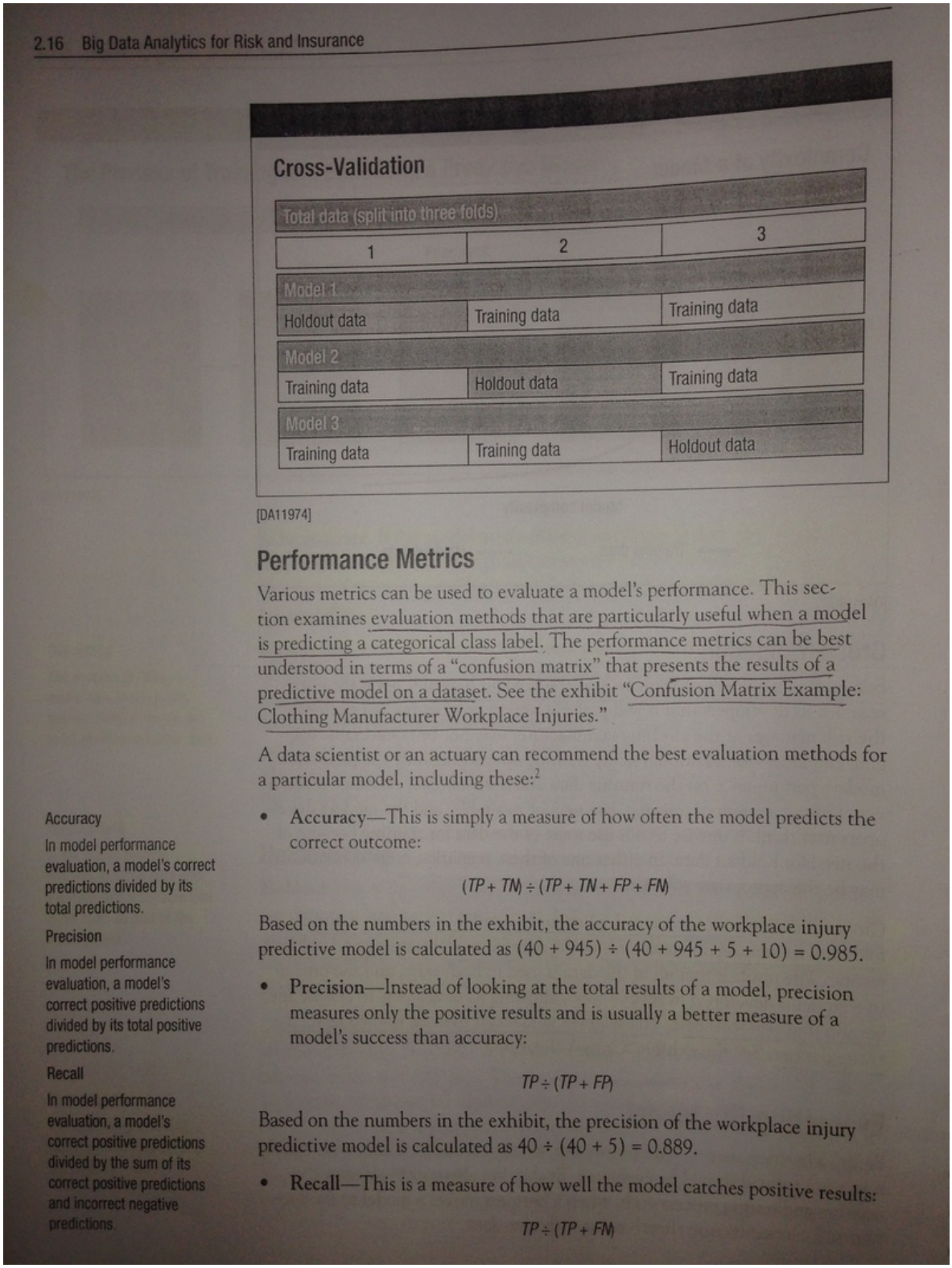
Generalization : The ability of a model to apply itself to data outside the training data. It should have some complexity.

Cross - Validation : is the process of splitting available data into multiple folds and then using different folds of the data for model training and holdout testing. The result is that all of the data is used in both ways, and the predictive model is developed in several ways, allowing its developers to choose the versions that performs best. the selected model can then be trained using the entire dataset.

***Cross - Validation is mostly used for the below reasons*:**

- A model's performance on the holdout data will not sufficiently reassure its developers it will perform well in production

- A very limited amount of training data may be available, and the model's developers think it unwise to not use some of the other data for training because of the need for holdout data.



Cross-Validation

Total data (split into three folds)		
1	2	3
Model 1		
Holdout data	Training data	Training data
Model 2		
Training data	Holdout data	Training data
Model 3		
Training data	Training data	Holdout data

[DA11974]

Performance Metrics

Various metrics can be used to evaluate a model's performance. This section examines evaluation methods that are particularly useful when a model is predicting a categorical class label. The performance metrics can be best understood in terms of a "confusion matrix" that presents the results of a predictive model on a dataset. See the exhibit "Confusion Matrix Example: Clothing Manufacturer Workplace Injuries."

A data scientist or an actuary can recommend the best evaluation methods for a particular model, including these:[2]

Accuracy
In model performance evaluation, a model's correct predictions divided by its total predictions.

Precision
In model performance evaluation, a model's correct positive predictions divided by its total positive predictions.

Recall
In model performance evaluation, a model's correct positive predictions divided by the sum of its correct positive predictions and incorrect negative predictions.

- **Accuracy**—This is simply a measure of how often the model predicts the correct outcome:

$$(TP + TN) \div (TP + TN + FP + FN)$$

Based on the numbers in the exhibit, the accuracy of the workplace injury predictive model is calculated as $(40 + 945) \div (40 + 945 + 5 + 10) = 0.985$.

- **Precision**—Instead of looking at the total results of a model, precision measures only the positive results and is usually a better measure of a model's success than accuracy:

$$TP \div (TP + FP)$$

Based on the numbers in the exhibit, the precision of the workplace injury predictive model is calculated as $40 \div (40 + 5) = 0.889$.

- **Recall**—This is a measure of how well the model catches positive results:

$$TP \div (TP + FN)$$

Evaluating the Data

Before a business uses a predictive model, its effectiveness on data with which it was not trained should be evaluated. ***Evaluation also continues after the training and testing process ends, when a business moves a model to production and can see its true effectiveness with new data***.

Performance Metrics for Categorical class label

Accuracy : In model performance evaluation, a model's correct predictions divided by its total predictions. It is a measure of how often he model predicts the correct outcome.

Formula : (TP+TN) / (TP+TN+FP+FN)

Precision : In model performance evaluation, a model's correct positive predictions divided by its total positive predictions. **It is better measure of a model's success than accuracy**.

Formula : (TP) / (TP+FP)

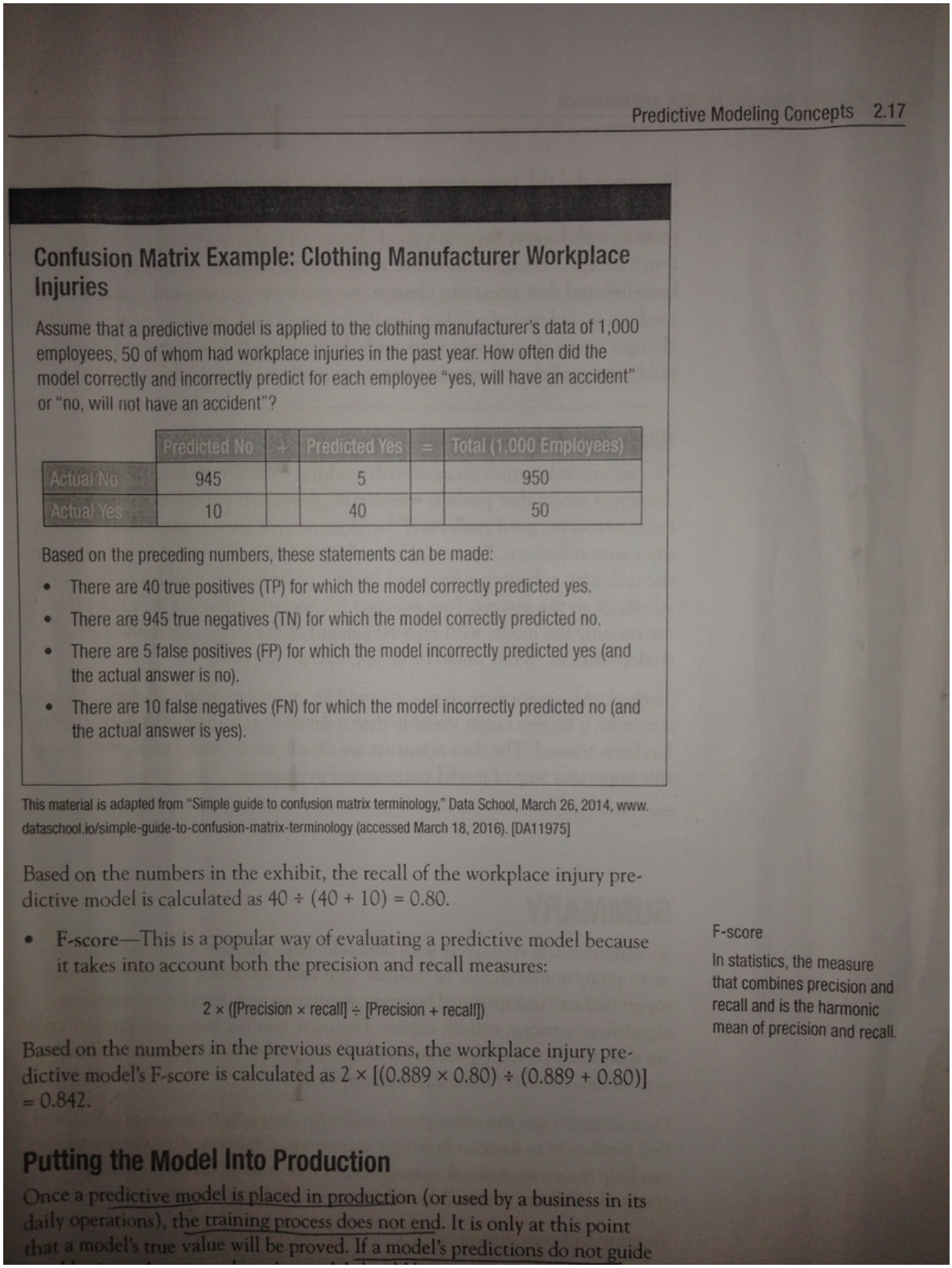
Recall : In model performance evaluation, a model's correct positive predictions divided by the sum of its correct positive predictions and negative predictions. it is a measure of how well a model catches positive results.

Formula : (TP) / (TP+FN)

F-Score : In statistics, the measure that combines precision and recall and and is the harmonic mean of precision and recall.

Formula : 2 * ([Precision*Recall] / [Precision + Recall])

The F-score is popular way of evaluating predictive model because it takes into account both the precision and recall measures



Confusion Matrix Example: Clothing Manufacturer Workplace Injuries

Assume that a predictive model is applied to the clothing manufacturer's data of 1,000 employees, 50 of whom had workplace injuries in the past year. How often did the model correctly and incorrectly predict for each employee "yes, will have an accident" or "no, will not have an accident"?

	Predicted No	+	Predicted Yes	=	Total (1,000 Employees)
Actual No	945		5		950
Actual Yes	10		40		50

Based on the preceding numbers, these statements can be made:

- There are 40 true positives (TP) for which the model correctly predicted yes.
- There are 945 true negatives (TN) for which the model correctly predicted no.
- There are 5 false positives (FP) for which the model incorrectly predicted yes (and the actual answer is no).
- There are 10 false negatives (FN) for which the model incorrectly predicted no (and the actual answer is yes).

This material is adapted from "Simple guide to confusion matrix terminology," Data School, March 26, 2014, www.dataschool.io/simple-guide-to-confusion-matrix-terminology (accessed March 18, 2016). [DA11975]

Based on the numbers in the exhibit, the recall of the workplace injury predictive model is calculated as 40 ÷ (40 + 10) = 0.80.

- **F-score**—This is a popular way of evaluating a predictive model because it takes into account both the precision and recall measures:

$$2 \times ([\text{Precision} \times \text{recall}] \div [\text{Precision} + \text{recall}])$$

F-score
In statistics, the measure that combines precision and recall and is the harmonic mean of precision and recall.

Based on the numbers in the previous equations, the workplace injury predictive model's F-score is calculated as 2 × [(0.889 × 0.80) ÷ (0.889 + 0.80)] = 0.842.

Putting the Model Into Production

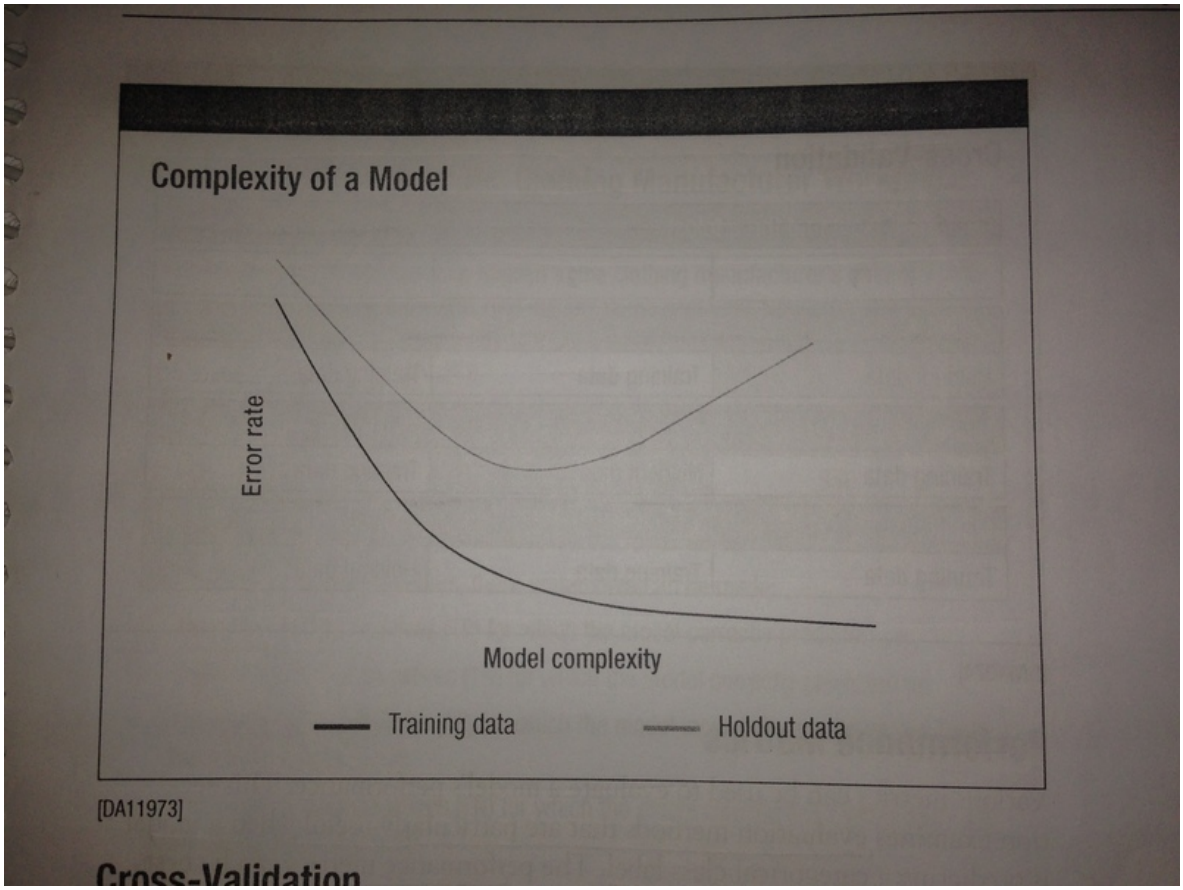
Once a predictive model is placed in production (or used by a business in its daily operations), the training process does not end. It is only at this point that a model's true value will be proved. If a model's predictions do not guide

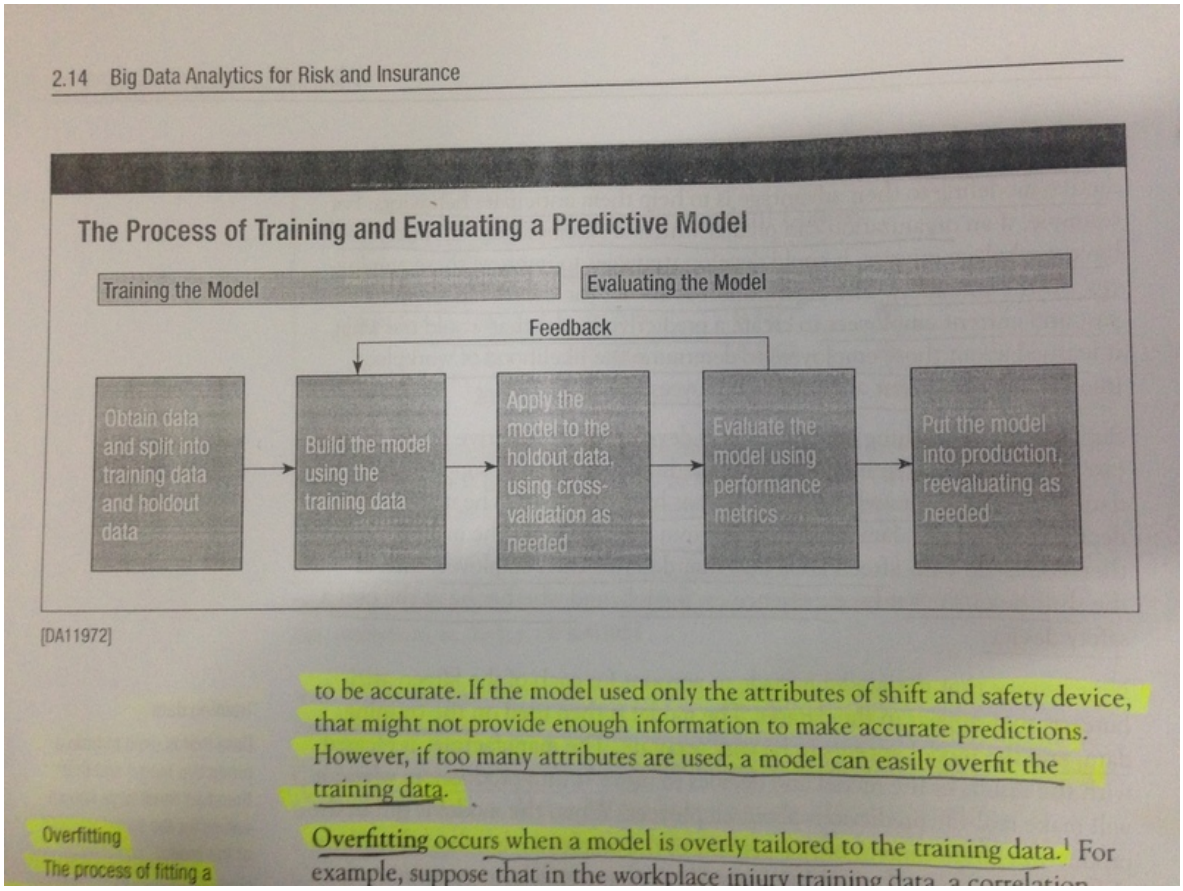
Putting the model into production

Once a model is moved into production, the training process does not end. If a model's predictions do not guide good business decisions, then the model should be re-evaluated. ***Keep in mind, no predictive model will make accurate decision indefinitely, because situations change***.

Understanding these modelling concepts and terms can help insurance and risk management professionals use data modelling across the industry, rather than solely in a marketing context.

Insurance and risk management professional should also defer to their professional experience when examining a model's results. If the results do not make sense there is a chance that the model was overfitted or is not complex enough.





Data scientists use the concepts of similarity, distance, nearest neighbors, and link prediction to forecast behavior and trends. Understanding these concepts can help insurance and risk management professionals use data modeling across the industry, rather than solely in a marketing context. Insurance and Risk management professional should be aware of the limitations of models - they can be too complex or not complex enough, and they must be reevaluated frequently.

Chapter 3 - Big Data Analysis Techniques

Most insurance and risk management professional will not directly apply data analysis techniques. Knowledge and fundamental understanding of how big data interacts with technology and how it can provide new information that will offer a competitive advantage to insurers and to insurance professionals who want to advance their careers.

Traditional data analysis techniques continues to be important in insurance and risk management functions for below reasons :

1. These techniques remain important in providing information and solving business problems.
2. They form the foundation for the newer techniques. It is also important for professionals to have some understanding of the newer techniques to be able to work with data analysts in teams to better understand the future of insurance.
3. Also, these techniques can often be adapted or used in various combinations to analyse large volumes of data or *data that comes from nontraditional sources, such as social media*.

Unsupervised Learning : Is a model that explores data to find relationships that can be further analyzed. Example : Text mining of Social Media for new insurance products.

There is often an interesting relationship; for example after information is obtained through unsupervised learning, a predictive model could be developed with that information for supervised learning.

Exploratory Data Analysis : Is a valuable approach to a business problem that can be used before developing and testing a model. The analyst can also obtain information about missing or inaccurate data. The techniques involve charts, and graphs that show data patterns and correlations among data.

Examples of charts and graphs

1. Scatter Plot : which is a two-dimensional plot of point values and show the relationship between two attributes.
2. Bubble Plot : Is a type of scatter plot in which the size of the bubble represents a third attribute for which there is data.
3. Correlation Matrix : similar to a 2-dimensional matrix with cells, the stronger the correlation the darker the the cell color shade in the matrix

The above graphs and plots are used to examine relationships before developing a model

Data Analysis Technique : After exploratory data analysis is complete and a decision is made to develop a predictive model, the analyst selects the most appropriate technique for a model, that will fit the business context and the type and source of the data. *Like data and technology in general, the techniques continually evolve*.

Example of some techniques

1. **Segmentation** : An analytical technique in which data is divided into categories, **can be used for both supervised and unsupervised learning**.
2. **Association Rule Learning** : Examining data to discover new and interesting relationships among attributes that can be stated as business rules, involves unsupervised learning in which a computer uses various algorithms to find potentially useful relationships among large amounts of data and to develop a rule that can be applied to new data. *Association rule learning is used to discover relationships between variables in order to develop a set of rules to make recommendations.*

> **Use Case for association rule learning** : Insurers may use association rule learning to recommend an umbrella policy with Auto and Homeowner's policy.

Traditional Data Analysis techniques are still used to solve business problems, such as determining rates for insurance products and reserves for unpaid or future claims. ***These techniques are usually applies to structured data, such as loss cost by claimant and type of injury***. sometimes external structured data is also used, such as gross domestic product by month.

Use Case : One of the following types of outcomes through traditional data analysis

1. A non-numerical category into which data will belongs such as buckets or bins
2. A numerical answer based on data, a model which give a value such as linear regression
3. A probability score based on historical data, example would be catastrophe loss for commercial insurance , example model as classification tree or logistic regression
4. A prediction of future results based on current and past data, for example an insurer or a risk manager wants to predict the cost of employees' back injuries, we can use decision tree and event tree analysis

Traditional data analysis techniques include classification trees, various types of statistical regression models, and cluster analysis. Moreover these types of techniques have been applied using machine learning to improve the accuracy of results.

Classification Trees : A *supervised learning technique* that uses a structure similar to a tree segment data according to known attributes to determine the value of categorical target variable.

- **Node** : A representation of a data attribute.
- **Arrow** : A pathway in a classification tree.
- **Leaf node** : A terminal node in a classification tree that is used to classify an instance based on its attributes. the leaf node determines the value of the target or output variable or probability or classification. ***The part of a classification tree that indicates the classification of the target variable is the leaf node***.
-

It is important to understand that these classifications are not necessarily what the actual outcomes will be

Regression Models : To determine numerical value for a target variable

Linear Regression : A statistical method (or algorithm) to predict the numerical value of a target variable (**which can also be a ratio of two attributes**) based on the values of explanatory variables. Using averaging method to predict average output value of target variable. *The working of linear regression algorithm id developed using a method that minimizes the errors represented by the difference between the actual target variable values in the data and those forecast by the model.*

Generalized Linear Model : (should not be confused with general linear model, which is a broad group of different types of linear models). A statistical technique that ***increases the flexibility of a linear model by linking it with a nonlinear function***. *It is used for more complex data classification and is widely used in Property-Casualty insurance business*. *A generalized linear model contains three components* :

- *the **first is a random component**, which refers to the probability distribution of the response variable*.
- *The **second is a systematic component**, which describes the linear combination of explanatory variable (attributes)*.
- *The **third component is a link function**, which relates the results of the random and systematic components*.
    - **Link function** : A mathematical function that describes how the random values of a target variable depend on the mean value generated by a linear combination of the explanatory variables (attributes).

Cluster Analysis : is unsupervised learning. A model that determine previously unknown groupings of data

Data analytics has to be handle not only large volumes of data but also rapidly increasing velocity of data. *Data generated from new technologies are often unstructured and therefore require new techniques for analysis*.

Association Rule Learning : is used to discover relationships between variables in order to develop a set of rules to make recommendations.

Data generated from new sources technologies are often unstructured and therefore require new techniques for analysis.

Text Mining : Modeling approach to search for words that are neither rare nor too common. *Another modelling approach is to look for adjacent words*. Example : the words "serious" combines with "injury: would likely be indicators of claim severity.

Social Network Analysis : The study of the connections and relationships among people in a network. Analysis helps in finding preferences or associations in a social network (Email chain is also a social network) . Helps is determining claims fraud rings or probability of fraud.

Neural Network : Technique can be used for both supervised and unsupervised learning. A data analysis technique composed of three layers, including an input layer, a hidden layer with non-linear functions, and an output layer, that is used for complex problems. The neural network is able to model complex relationships between the input and output variables.

The combination of analytics and machine learning has enhanced insurers' predictive modeling capabilities.

Classification Tree model : A classification tree is then built, through an induction process, where the computer **recursively** analyzes different splits on the value of the attributes in various sequences based on training data. the tree is then applied to holdout data to see whether it generalizes so as to make accurate predictions using data that was not included in its development. ***There is a risk that the tree might overfit the training data. In such cases, it will not produce useful results and will need to be pruned***.

When developing a classification tree model, various attributes are independently analysed for the information gain they provide in classifying

!! Note : the sequence of attributes does not follow the relative information gain ranking of the attributes, which is often the case when the attributes are analyzed recursively to construct a tree !!

Recursively : Successfully applying a model

Root Node : The first node in a classification tree that provides the greatest information gain for determining the value of the target variable.

Each classification rule is represented by a leaf (attribute) node of the classification tree.

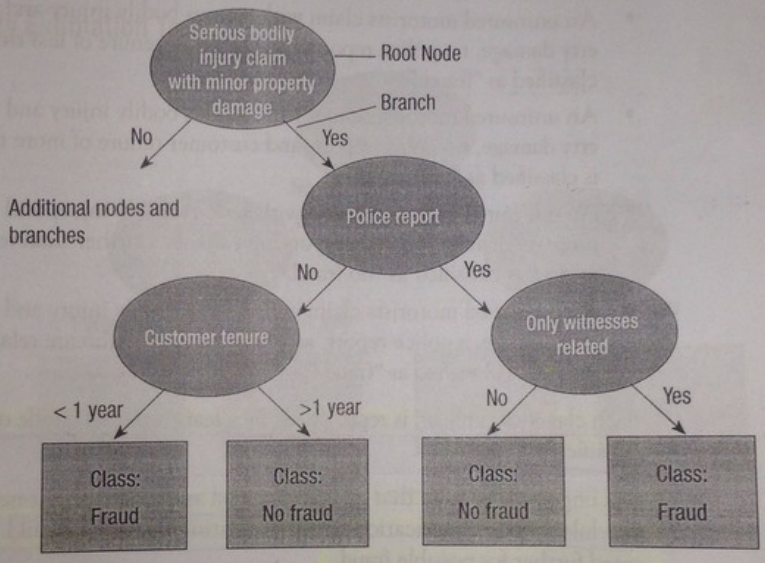
Probability Estimation Tree : For more information users of predictive models want to know the predicted class for a new instance but also the **likelihood, or probability, that it is correctly classified**. This information **enables them to do a cost-benefit analysis**. the data to construct a classification tree can be used to estimate a probability. **This is known as tree-based class probability estimation**. **Tree-based probabilities are calculated by dividing the number of times the model correctly predicted the value of the target variables by the number of total predictions for each class at each leaf node**.

A classification tree is then built, through an induction process whereby the computer **recursively** analyzes different splits on the value of the attributes in various sequences based on training data. The tree is then applied to holdout data to see whether it generalizes so as to make accurate predictions using data that was not included in its development. There is a risk that the tree might overfit the training data. In such cases, it will not produce useful results and will need to be pruned.

Recursively
Successively applying a model.

The tree shown is simplified for illustration purposes. In reality, it would be much more complex. Note that the sequence of attributes does not follow the relative information gain ranking of the attributes, which is often the case when the attributes are analyzed recursively to construct a tree. See the exhibit "Classification Tree for Detecting Uninsured Motorists Claims Fraud."

Classification Tree for Detecting Uninsured Motorists Claims Fraud



This illustration is a simplified version of a classification tree. It does not include all the attributes in the chart illustrating information gain. In actual practice, a classification tree model would be much more complex and include many more nodes.

[DA12078]

In the exhibit, the **root node** represents the attribute (serious bodily injury claim with minor property damage) that provides the greatest information gain for determining the value of the target variable. The data is further

Root node
The first node in a classification tree.

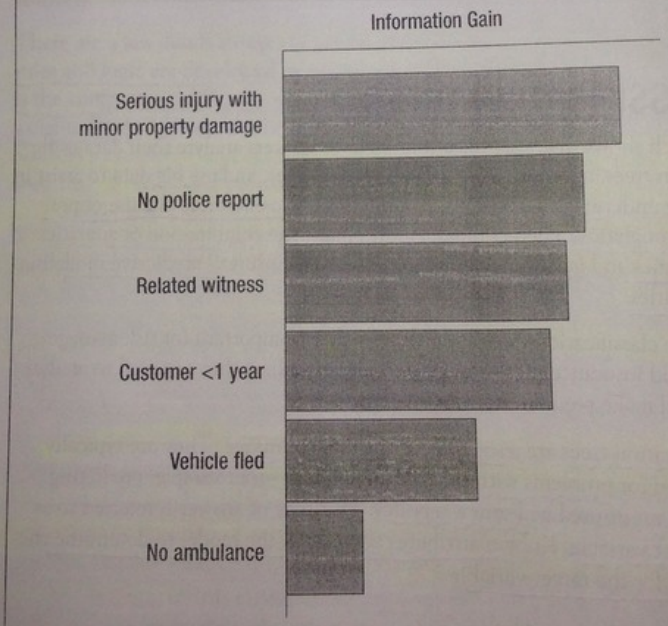
Classification Tree Model

To illustrate the use of a classification tree, assume that an insurer would like to develop a model to determine at time of first report whether an uninsured motorists claim is fraudulent. When developing a classification tree model, various attributes are independently analyzed for the information gain they provide in classifying an uninsured motorists claim as fraudulent or not fraudulent. The insurer used these attributes in developing the model:

- Serious bodily injury but minor property damage to the vehicle—Yes/No
- Police report—Yes/No
- Customer tenure—Less than one year/More than one year
- Only witnesses are relatives or friends of the insured—Yes/No
- Other vehicle fled the scene—Yes/No
- Ambulance called—Yes/No

The attributes were then ranked according to their information gain. See the exhibit "Information Gain From Attributes for Detecting Uninsured Motorists Claims Fraud."

Information Gain From Attributes for Detecting Uninsured Motorists Claims Fraud



[DA12077]

Linear functions to make business decisions

Linear function : A mathematical model in which the value of one variable changes in direct proportion to the change in the value of another variable.

Linear Discriminant : A data analysis technique that uses a line to separate data into two classes. Used to separate and classify the

data. *Miss-classifications can be evaluated by accuracy and precision of the model*.

Regression Analysis : A statistical technique that is used to estimate relationships between variables.

Logistic Regression : A type of generalized linear model in which the predicted values are probabilities.

Instance space : An area of two or more dimensions that illustrates the distribution of instances.*If two dimensions are used them the dimensions are plotted as an X-Y graph, if more than two dimensions are used then the instance space will be a three-dimensional cube and the discriminant a plane rather than a line*

Support Vector Machine (SVM) : is a technique often used to increase the accuracy of a linear discriminant. A linear discriminant that includes a margin on each side of the line for classifying data. The linear discriminant line can be drawn at any angle to separate the classes, *An SVM add a margin on each side of the discriminant line. It them applies mathematical equations to maximize the margins while selecting the optimal angle for the linear discriminant. the technique improves the accuracy of the linear discriminant to separate the classes*. ***The insurer uses the optimized linear discriminant to create rule to predict***.

Linear Regression : Is a type of linear function which predicts a numerical value for the target variable. It assumes that the predicted value changes proportionately (a linear relationship) with the attribute value. It is used to calculate a line that best fits in the observations in order to estimate the mean relationship between the attribute and the target variable.

The vertical distances between the dots and the line are known are "errors" produced by the model. they are measured as the difference between the observed value of the target variable (the dot) and the mean value as estimated by the line.

The objective in calculating the regression line is to minimize total errors a common algorithm for determining the line minimizes the sum of the squared errors and is known as **least squares regression**. Squaring the errors increases the penalty for values further from the line compared with those closer to the line.

!!Unlike a linear discriminant, which is used to predict a class for the target variable, linear regression predicts a numerical value for the target variable as a functions of one or more attributes (explanatory variables). Just as with a linear discriminant, linear regression can be show as a line on a two-dimensional graph.!!

Least Squares Regression : A type of linear regression that minimizes the sum of the squared vales of the errors based on the differences between the actual observed and estimated (mean) values.

Multiple Regression : A variation of linear regression whereby two or more attributes (explanatory) are used for prediction. The values of each attributes would be weighted based on its influence in estimating the value of the target variable.

Disadvantages of Linear Regression :

- the assume that the variability (as measured by the model error), or variance, around the estimated mean of the target variable is normally distributed and therefore follows a standard symmetrical pattern. **But with property-casualty insurance data, variability around the mean is not normally distributed**.
- Second assumption is that the variability around the estimated mean of the target variable is the same regardless of its size. This is known as **homoskedasticity**. **But with property-casualty insurance data, variability around an estimated mean of the target variable usually varies with its size**.

Homoskedasticity : In linear modeling, variance of the model errors around the estimated mean of the target variables the same regardless of its size.

Generalized Linear Models :

There are three characteristics of GLMs :

- GLMs use a link function that defines the relationship between the mean values estimated by a linear model model (called the systematic component) and the variability of the target variable around those means (called the random component).
- The random component is usually specified as an exponential distribution, which is not normally distributed. that overcomes the problem previously discussed that the random component of a linear model is normally distributed.
- The variability (variance) of an exponential distribution is a function of its mean overcoming the problem of homoskedasticity.

Explanation of GLMs :

When estimating auto frequency, it is common to use a linear model of various explanatory variables, such as miles driven and driver experience, as the systematic component. the mean values estimated by the linear model are connected with a log link function to a **Poission distribution**, as exponential distribution that serves as random component. A Poisson distribution is commonly used to model a probability distribution of claims frequency.

Poisson distribution : a distribution providing the probability of a number of independent events happening in a fixed time period.

Logistic Regression : It is a classification analysis model which not only tells the classification for a new instance but also the likelihood, probability, that it falls into that class. this is known as **class probability distribution**, which is an estimate of the probability that an instance will fall into its predicted class.

Logistic regression is a type of generalized linear model that uses a link function based on logarithm of the odds (log odds, or logit) of the event occurring. (The odds of an event occurring is the ratio of

the probability of the event occurring to the probability of the event not occurring.)

Cluster Analysis

Is a unsupervised learning technique that can explore data to discover previously unknown information or relationships. Several iterations of cluster analysis can be applied to subdivide clusters and provide more granular information.

Hierarchical clustering

A modeling technique in which multiple clusters are grouped according to their similarities.

Dendogram : A diagram that illustrates the groupings from hierarchical clustering. The distance between clusters represents how similar each cluster is to another cluster. F clusters are grouped at one level of the hierarchy, they remain grouped as each higher level. The distance between clusters represents how similar each cluster is to another cluster.

K nearest Neighbors algorithm :

Voting Method, the prediction of the instance of the future target variable based on the majority of historic data points which are grouped as similar. the nearest neighbors that are most similar would have the greatest weight. We can make a decision on Voting Method.

There is another method used along with Voting Method, which is the average method, where the average or probability is calculated for likelihood

K-means : An algorithm in which 'k' indicates the number of clusters and "means" represents the clusters' centroids.

Centroid : the center of a cluster is the average of the values for the instances in each cluster.

Clusters : can be described in two ways - **differential and characteristic**

- A characteristic description describes the typical attributes of the cluster.
- A differential description describes the differences between the instances of one cluster and those of other clusters.

Difference between K-means and Hierarchical clustering :

- K-means clustering presented a differential description. The clustering around different characteristic points provided attributes that were different for each cluster in order to distinguish those instances.
- Hierarchical clustering is in groups according to similar attributes

Text Mining

Sources of text (unstructured data both internal and external): claims files, social media posts, news stories, and consumer reviews.

Bot traditional and newer modeling techniques can be applied to the results of text mining.

Text is unstructured and must be cleaned up and turned into structures data before it can be used in a model. To understand the challenges and benefits of text mining, it is important to understand its process.

To apply a modeling algorithm to text, the text must be put into a form that the algorithm can recognize, such as by giving it numerical values.

Steps in text mining process:

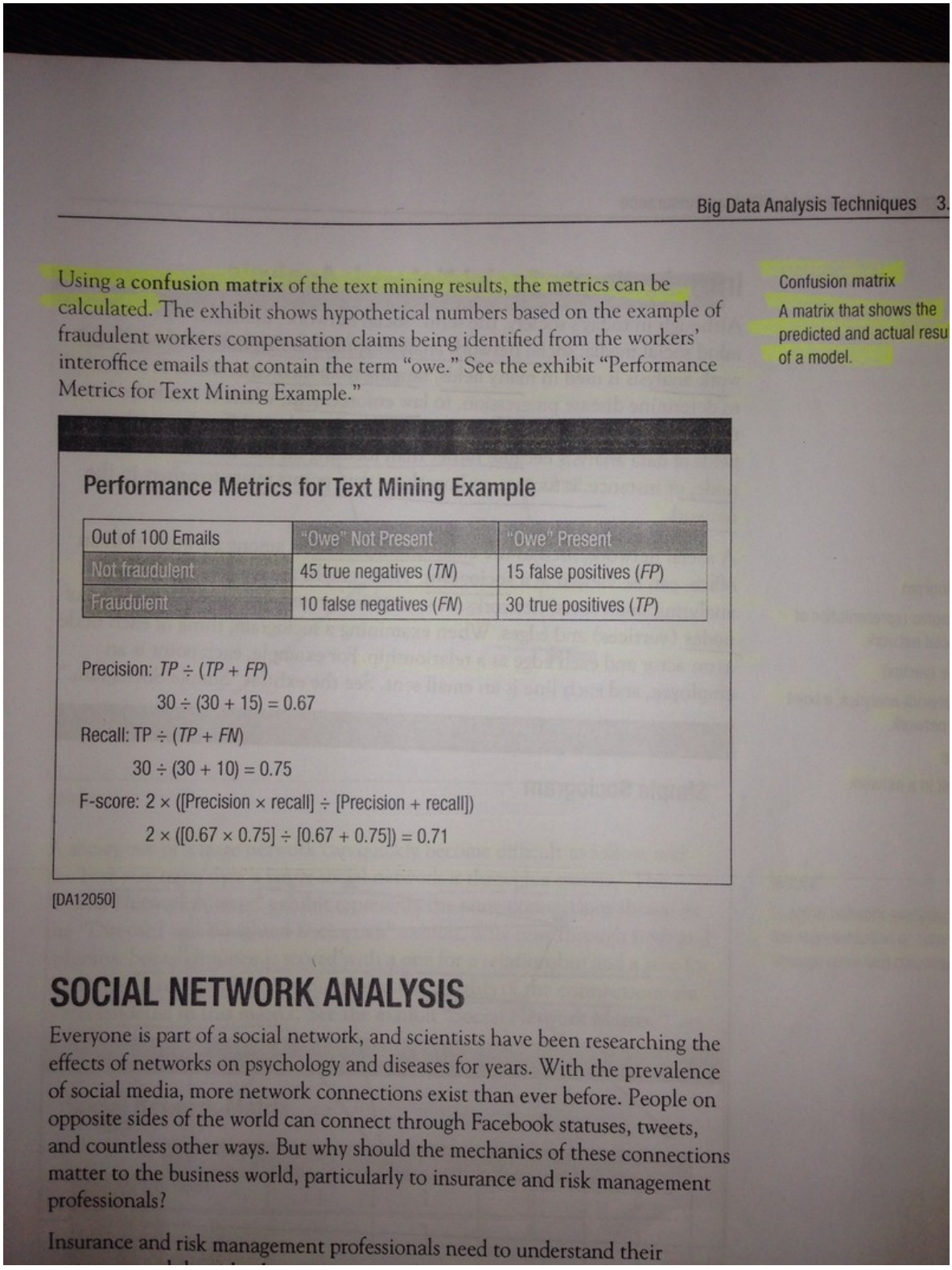
1. ***Retrieve and Prepare text***
    1. collecting a set of writings or documents called **corpus**, where each document is made of various words, also called **terms** or **tokens**. (In a text mining context, a meaningful term or group of terms are called **tokens**)
    2. **Preprocessing** of corpus, such as cleaning of text, removing punctuation, spaces, and numbers (depending on context) called as **stopwords** (are common words that offers little meaning and is filtered out in a text mining context).
    3. Remove abbreviations
    4. **Stemming** the removal of a suffix from a term for the purpose of text mining.
2. Create a Structured Data from Unstructured Data
    1. key terms can be extracted and represented in a table or spreadsheet as structured data
    2. **Term Frequency** : a measurement of how often a term appears in a document. we generally set a upper bound and lower bound of the term frequency and analyse only those terms which fall in that inner bound.
    3. **Inverse document Frequency (IDF)** : the measurement of the significance of a term (or rarity of a term) within a document of test in a corpus based on how many documents within which it appears in that corpus.
    Formula : **IDF(t) = 1 + log(Total number of documents / number of documents containing it),** ***the higher the value of IDF, the rarer the term and higher the significance of the term and the document it is mentioned it***
3. Create a model using Data Mining Techniques,
    1. ***To apply a modeling algorithm to text, the text must be put into a form that the algorithm can recognize, such as by giving it numerical values.***
    2. using nearest neighbors or classification learning technique, rules can be created and applied to new documents.
    3. Drawback : over-fitting with too many list of words used in model, different variations of the same word.

    4. After probability has been defined using classification technique, a linear scoring method provides weight to each term, based on its probability, which appears in a document. The document is then a final score, which is the probability that it indicates fraud.
4. Evaluate the Text Mining Model
    1. Using a **confusion matrix** of the text mining results, the metrics can be calculated. a matrix that shows the predicted and actual results of a model.

Sentiment Analysis : is a way of measuring the opinion of something (often a product), be it positive, negative, or neutral. In sentiment analysis, each term is given a positive, negative or neutral value. The values are added to produce a sentiment score.

This type of analysis is known as sentiment analysis. Sentiment analysis is a way of measuring the opinion or emotion of something be it positive, negative, or neutral in unstructured social data

Semantic Network : a network that shows the logical relationships between concepts. In some text mining situations, particularly those involving large amounts of data and limited amounts of time-it can be beneficial to skip some of the pre-processing steps and have the computer itself find the meaningful words in unstructured data. One approach is to create a set of linguistic rules that can automatically process text, the results are represented as a **semantic network**. ***A second approach to text mining unstructured data is the use of neural networks***. ***Neural network attempt to replicate the thought process of the human brain; however they have limitations in linguistic settings***.



Using a **confusion matrix** of the text mining results, the metrics can be calculated. The exhibit shows hypothetical numbers based on the example of fraudulent workers compensation claims being identified from the workers' interoffice emails that contain the term "owe." See the exhibit "Performance Metrics for Text Mining Example."

Confusion matrix
A matrix that shows the predicted and actual resu of a model.

Performance Metrics for Text Mining Example

Out of 100 Emails	"Owe" Not Present	"Owe" Present
Not fraudulent	45 true negatives (*TN*)	15 false positives (*FP*)
Fraudulent	10 false negatives (*FN*)	30 true positives (*TP*)

Precision: $TP \div (TP + FP)$

$30 \div (30 + 15) = 0.67$

Recall: $TP \div (TP + FN)$

$30 \div (30 + 10) = 0.75$

F-score: $2 \times ([\text{Precision} \times \text{recall}] \div [\text{Precision} + \text{recall}])$

$2 \times ([0.67 \times 0.75] \div [0.67 + 0.75]) = 0.71$

[DA12050]

SOCIAL NETWORK ANALYSIS

Everyone is part of a social network, and scientists have been researching the effects of networks on psychology and diseases for years. With the prevalence of social media, more network connections exist than ever before. People on opposite sides of the world can connect through Facebook statuses, tweets, and countless other ways. But why should the mechanics of these connections matter to the business world, particularly to insurance and risk management professionals?

Insurance and risk management professionals need to understand their

Social Network Analysis (or Network analysis)

The study of the nodes (vertices) and edges(lines) in a network. or in business words the links and influence of individuals and others.

These methods offer ways to observe social network connections:

- Link Analysis
- social network metrics
- Network classification

It helps in study rather the attributes of the nodes or points of information but how those data points relate to each other in the network.

"Claim Adjusters using social network analysis to trace links among the various individuals."

A social network, such as the email communication among co-workers in an office can be viewed as a **sociogram**. A graphic representation of a social network where ***each node (vertex or point in a network) is an actor and each edge (link) has a relationship***.

Some graphs use arrows to distinguish that some connections flowing in only one direction, and some use weighted edges(thicker or thinner lines) to show the strength or frequency of the connections. A **directed tie** is an edge within a network graph that has direction, represented by an arrow, may or may not be reciprocated.

A **sociogram** of a large network can quickly become difficult to follow, and the best way to **analyze a larger social network** is through a **matrix which is the representation of data through rows and columns of a social network analysis**.

Small social network - Sociogram

Large Social Network - Matrix

Data mining techniques can be used to analyse there networks, both supervised and unsupervised learning techniques such as clustering can be applied.

Sentiment Analysis : is a way of measuring the opinion or emotion of something or behind a selection of text.

Link Analysis : is a key concept of network analysis. ***Link prediction is the process of trying to predict a pair of links***. Measuring similarity, the basis of link prediction, allows insurance and risk management professionals to analyze groups of customers and/or potential customers and observe trends. **"Similarity is the basis of link prediction."**

Insurers must be careful in the use of social media, for privacy concerns.

Social Network Metrics

The term "path" is used to describe how something flows through a network, passing each node and edge only once. The efficiency of the flow between social network connections can be determined through **centrality measures**

centrality measures In a social network context the quantification of a node's relationship to to other nodes **in the same network**

Degree : a measure of the connections each node has.

Closeness : the measure of the average distance, or path length, between a particular node and the other node in a network.

Betweenness : The measure of how many times a particular node is part of the shortest path between two other nodes in a network. The node serves as a connection between otherwise separated nodes in the network

Network classification : Nearest neighbor and data mining algorithms, such as logistic regression, can also be used to analyze social network.

Egonet : A network composed of an ego (personality or trait of an individual) and the nodes (similar people) directly connected to it.

Bigraphs : is a network that has two types of nodes. This can provide a complete picture of a network

the tendency of people to connect to others who are similar to then is called *homophily*

Local Variable : In a social network analysis an attribute that relates to the characteristics of a specific node.

Network Variable : In a social network analysis, an attribute that relates to the node's connection within the network.

the above two variable are used in a modeling context where each node is viewed as an instance, describes by its local and network variables. This format allows the data gleaned from social network analysis to be used in various data mining algorithm, such as classification, clustering and logistic regression.

Neural Network

Most important ability is to analyze large quantities of data. ***Can only operate on numerical data.***

!!IMPORTANT - In order for neural networks to predict the success of a project, the factors that caused the success or failure of previous projects must be understood. !!!

Predictive Analytics : Statistical and analytical techniques used to develop models that predict future events or behaviors.

Hidden Layer : A layer in a neural network in which neurons learn and recode information received from input data. the hidden layer performs various calculation using mathematical functions to match input and output. *These functions may employ cluster analysis, classification analysis, or linear functions*. A hidden layer is made up of **neurons which is mathematical function ins a neural network that receives input**. **The hidden layer of a neural network performs various mathematical functions to match inputs to outputs.**

Classification analysis : a supervised learning technique to segment data according to the values of known attributes to determine the value of a categorical target variable.

Application of Neural Network

1. customer retention after premium increase
2. renewal after changes to premium
3. produce more accurate rates tailored to the individual customers
4. can pick up small nuances, that means small patterns in data
5. find factors using machine learning and cluster analysis to determine probability of failure or success of a project
6. It develops rules to make predictions as it performs various mathematical functions.

Drawbacks of Neural Network

1. hidden layer is too opaque, optimization of the hidden layer(s) must be precisely determined
2. Network must be able to learn from errors, however there is a risk of over fitting
3. It is more difficult to identify over-fitting and correct it
4. it cannot work on non-numerical data

Optimization in a neural network is achieved when the prediction at the output layer is evaluated to actual values. work of a data scientist to precisely determine the optimization function.

Data scientists can use clustering with neural networks to produce more accurate results. Unlike other types of data analysis techniques, neural networks can make observations.

Chapter 4 - Underwriting Applications of Big Data Analytics

Generally automobile rates are based on a class basis. that is, similar exposures are grouped and the rates of the group are charged to each member of that group. *Underwriters and Actuaries identify attributes and additional attributes that reflect potential frequency and severity of loss*.

The use of telematics can help gather information ***(additional attributes such as accelaration of car and hard braking)*** so that underwriter can better understand personal and commercial policyholders and to develop more accurate automobile insurance rates or *it can be used as simple as a risk management tool*.

It is difficult to rely solely on traditional underwriting guidelines for newly introduced products and emerging technologies, therefore advance data analytics is important

Telematics can be used through temporary or permanent installation of tracking devices, through embedded telematics equipment and through smartphone applications. these devices track driving habits, regarding braking, acceleration, speed, cornering, lane shifting, left turn versus right turns, and the time of the day the vehicle is driven.

privacy and regulatory considerations for vehicle telematics, vehicle telematics data does not typically fall under states' definitions of protected consumer information.

Purpose of Telematics devices is to track driving habits and then transmit that information wirelessly to a computer.

Traditional Attributes Rating Attributes

- Territory - which include road conditions, state safety laws, and the extent of traffic regulation are territorial factors
- Use of the Auto
- Age
- Gender
- Marital Status
- Vehicle Weight and Type

- Vehicle Use : heavy vehicles more likely to cause severe damage
- Radius of Operation : long distance increases the changes of accident
- Special Industry Classifications : food delivery, waste disposal, farmers
- Driving Record
- Driver Education
- Years of Driving Experience
- Credit Score

Usage Based Insurance : A type of auto insurance in which the premium is based on the policyholder's driving behavior.

Advantages of Usage Base Insurance :

- Allows insurers to better segment drivers in rating classifications between preferred, standard and non-standard

Disadvantage of Usage Base Insurance :

- People who participate are likely already safe drivers.
- Another challenge is the need to evaluate the data in context so that the data is not painting the wrong picture. For example, a driver in a congested area will brake harder and more frequently than a driver in a rural area.

Loss Ratio : A ratio that measures loss and loss adjustment expenses against earned premiums and that reflects the percentage of premiums being consumed by losses.

By using telematics loss ratios will decrease and retention ratios will increase

Telematics can be also used at risk-management level by organizations.

Vehicle telematics can function as ongoing driver education for those who participate.

It takes sophisticated modelling techniques to make that information relevant and to determine how much the information should affect rates.

Insurers use loss exposure data generated through telematics to supplement the data they have traditionally used for rate-making. thy analyze the data using sophisticated algorithms, such as generalized linear models, to determine the correlations and interactions among all the rating attributes being considered. these includes traditional rating attributes, such as vehicle use and type, and new attributes generated through telematics, such as braking, acceleration, and time of day. Increasingly insurers are also using machine learning to aid in the discovery of variable interactions that may not be evident when using a predicted linear model

One of the areas where the telematics information will be distorted by the fact that older autos have weaker brakes and are owned by a higher proportion of younger drivers that are newer autos. Therefore younger drivers tend to experience more accidents than older drivers. There a machine learning algorithm can analyse and pick these interactions among these rating attributes.

Privacy and Regulatory Considerations for Vehicle Telematics

- Customers who do not opt to use UBI have privacy concerns regarding how their personal information is protected when it is transmitted and who owns the data about a driver's behavior.
- Vehicle telematics data does not typically fall under states' definitions of protected consumer information.
- Insurers must also ensure that their use of vehicle telematics to make rate changes is transparent and not discriminatory.

Segmenting Home Owner Policies

Historically Insurers, have struggled to make a profit in selling home owner's policies, By using more data, new correlations between granular data between their customers and loss exposures.

Before advances in machine learning and predictive analytics, homeowners policy underwriters were Limited in their ability to incorporate the many attributes needed to accurately predict losses.

To develop the sophisticated model needed to refine its homeowners classifications, the insurer must choose the criteria it will use to segment its homeowner policies.

Traditional rating variables :

- year built
- construction type
- location
- **age of home electrical's system (additional attribute for risk grouping)**

Geo-location information (can be obtained from governmental wildfire agency) :

- underbrush
- relative wind speeds
- distance between home and fuel could help to better understand the increasing wild-fire related claim

Customers with a lower loss ratio are more profitable. Using analytics like machine learning, the insurer has managed to achieve its goal, it has increased rates, lowered overall ratios, and become more competitive in the marketplace.

Using traditional flat rate increase across the homeowners book, the customers with higher loss ratios would be more likely to stay, leaving the insurer with the least profitable segment of its book. The overall loss ratio would most likely not improve.

Personal lines insurers are investigating ways to incorporate discounts for homeowners insurance policies based

on homeowners' use of smart home devices. Personal lines insurers may even develop usage-based homeowners insurance, similar to that now offered for auto insurance, in which rates would be partially based on wireless data sent from devices within the home.

Relative wind speed is an attribute that could help an insurer better understand wildfire-related claims.

By using machine learning to segment its policies, the insurer see the effect of attributes and interactions it had not previously considered in its traditional rating model. The insurer notices that loss ratios are inconsistent across the segments based on the interactions between traditional and new variables. ***Producers' and claim representatives' files may contain a significant amount of data that could be made available to machine learning.***

Customers with a lower loss ratios - who are therefore more profitable to insurer - seek a competitive rate; the challenge for an insurer is to provide this while still challenging sufficient premiums for customers with higher loss ratios, who are unlikely to leave.

Without machine learning , finding variables would be really time-consuming

Underwriting Products Liability Risks using Data Mining

The data science team will apply cluster analysis, as unsupervised learning technique to find out whether there are any patterns in the recent large claims sometimes called as K-Means or nearest neighbor clustering. Supervised learning and predictive modelling require known attributes and a known target variable.

Use social media data and use text mining to search social media for references to the products manufactured by the insurer's customers. These can be equated to a positive sentiment or a negative sentiment which will give a sentiment score to the product.

Using this information along with insurance professionals knowledge, the data science team has the attributes and the target variable with which to build a predictive model.

- *Predicting Liability risks is important, because retailers, manufacturers, distributors can become liable when products cause injuries to consumers.*
- *Social philosophy favors consumers,*
- *and laws that hold businesses liable for their products have increasingly expanded.*

Data Mining can help an insurer properly evaluate an account's products liability loss exposures. Particularly when evaluating products with uncertain risk characteristics, underwriters can benefit for this approach, rather than rely on traditional underwriting guidelines and also find out new trends in risks which they may not have factored it in during underwriting.

Predictive Modeling for Emerging Risks

The insurer will also use the predictive model to underwrite new accounts. by more accurately predicting the likelihood of claims, the insurer will be able to determine which accounts it should accept to insure and what pricing levels are required to maintain profitability for the line of business.

Monitoring reactions to products and its liability can be monitored through text mining of social media. and it can also mine its historic liability claims to more accurately price similar exposures.

Through text mining, insurers can identify attributes and the target variable with which to build a predictive model.

In a more realistic and complex scenario, the insurer, working with the data scientists to analyse the data mining and using its insurance professionals' knowledge and past experience with liability claims, could determine which one of the products' attributes are important in terms of product liability.

In terms of risk management perspective, the insurer can use the valuable information it has gained from data mining to help its customers.

Machine learning allows insurers to incorporate more attributes in their underwriting decisions. A fully trained segmentation model can find meaningful patterns in data and automatically segment policies.

Data Mining can help an insurer properly evaluate an account's products liability loss exposures.

Cluster Analysis , text mining of social media and predictive modelling can help to predict claims, and therefore, an account;s loss exposures.

Chapter 5 - Claims Applications of Big Data Analytics

Mainly used in Fraud Detection and claim adjuster assignment for claims predicted as severe, **about 10 percent** of the property-casualty insurance industry's incurred losses and loss adjustment expenses stem from fraud. Insurer's attempt to detect fraud by identifying patterns, controlling underwriting at the points of sale and using **special investigation unit (SLU)**

Advances in data mining to more effectively identify patterns in fraudulent claims activity. Insurance and Risk Management professionals therefore benefit from understanding how to analyse links in a social network and clusters of data points to capture new trends in claims fraud.

Claims representatives can often find evidence that someone may be lying by comparing his or her social media posts with his or her statements in a claim

Insurer's data science team mainly use cluster analysis and classification trees.

Cluster analysis *can be used to create cluster of similar claims according to various attributes. These attributes are not known in advance because cluster analysis is an explanatory technique without any predefined variables. Clusters of claims instance would be expected to group around attributes associated with claim severity. Cluster analysis would continue to be applied on individual clusters until the data scientist has a good understanding of the relationships between significant variables.*

After cluster analysis is complete, the characteristics of the clusters can be used as attributes to develop classification trees with which to assign new claims.

Logistic Regression *could then be used to estimate the probability of each of these outcomes to enable the insurer to determine the appropriate resources for each complex claim.*

Predictive modeling seems to be the long term solution to detect fraud claims, without increasing auto rates and spend on resources to investigate claim of fraud in nature

Steps taken to identify fraudulent claims :

1. Detect claims fraud through traditional fraud indicators (lists of claims fraud indicators are published by National Insurance Crime Bureau) plus insurance company can make their own list of fraud indicators and also through mining social media data.
2. Apply network analysis by examining links and suspicious connections.
3. Apply cluster analysis to discover claims characteristics that might indicate fraud.

Examples of fraud indicators an insured or a claimant pushes for quick settlement or has too much or too little documentation

Although the insurer is taking steps to identify fraudulent claims, some fraud still goes undetected, the reason is that the traditional

fraud indicators is based on historical data. *Intelligent and innovative fraudsters will change their approaches and patterns, limiting the usefulness of these indicators*. ***The traditional approach is highly subjective and depends on claims representatives' experience in the field***. A more automated approach would allow for greater objectivity and enable new claims representatives to be more effective in less time.

In analyzing social media to detect fraud claims practices, requires not only investigating social media posts but also the connections within a network as well. ***The connections in a network would help in identifying a fraud ring***

Because for predictive modelling an insurance may have very less historical data to indicate fraud claim and since fraud is ever evolving, the model may become quickly outdated. Clustering techniques such as K-means are unsupervised learning techniques to identify new fraud indicators before predictive modeling is applied. ***Essentially the fraudulent claims are outliers within the already outlying cluster***

Using classification tree analysis in claims assignment :

In classification analysis, A computer recursively applies a model to analyze different splits in the values of attributes.

A relatively small percentage of claims account for much of the costs to insurers. some of the high-cost claims can be easily identified at the time of first report, such as the total loss of a building from fire, a major auto accident, or a plant explosion. *However many claims develop gradually into significant losses*.

Lift is the percentage of a model's positive predictions divided by the percentage expected by chance.

Example Workers Compensation Claims, where the majority or more than half costs to be paid are medical costs. Potentially

complex claims are the most difficult for insurers to identify at the time of first report.

To use the resources effectively, Insurers or Risk Managers can use their own data, or they an use their insurers' or their parties' data to identify potentially serious claim.

Interest in Modeling is also used to determine claimant characteristics to determine if the person will likely have more chronic problem or develop opoids problem.

Classification tree analysis can be used to assign new claims according to target variables.

Complex Claim : A claim that contains one or more characteristics that cause it to cost more than the average claim.

Procedure to create a model for complex claims

1. An important step after identifying the attributes of complex claims is ranking the attributes according to their relative information gain.
2. Determine list of attributes (input variables) and ranked by the relative importance using **Information Gain** by complex claims attributes. *Including a claims professional on the data science team can help with selection of the most important attributes*.
3. Using a classification tree to illustrate attributes of complex claims to build a predictive model. The computer recursively built a tree by analyzing different splits in the values of attributes in various sequences.
4. Validating the complex chain model. Through the machine learning, the model adjusts the weights assigned to each of the attributes to better predict accuracy. This is done to keep the model effective based on new incoming data.

!! Note : The sequence of attributes in the tree does not strictly follow the relative information-gain ranking for the attributes was developed independently of other attributes. that is usually the case, because the information-gain ranking for the

attributes was developed independently of the other attributes. However, the attributes Representative by each node in the tree depend on the attributes (and their values) that sit above them !!

Combination of Nodes : A representation of a data attribute in a classification tree.

Complex Claim Reporting :

1. Claim reported online or by telephone
2. Direct the claim to the appropriate claim intake based on geographical location
3. If identified as catastrophic claims, it is assigned to senior claims adjuster
4. Machine learning algorithm is used to model and predict the type of claim as "complex" or "not complex"
    1. It should be decided based on attributes on what type of modelling technique to use
5. through machine learning, the model adjusts weights assigned to each of the attributes to better predict complexity.

Claims payment are the biggest financial obligations for the insurer

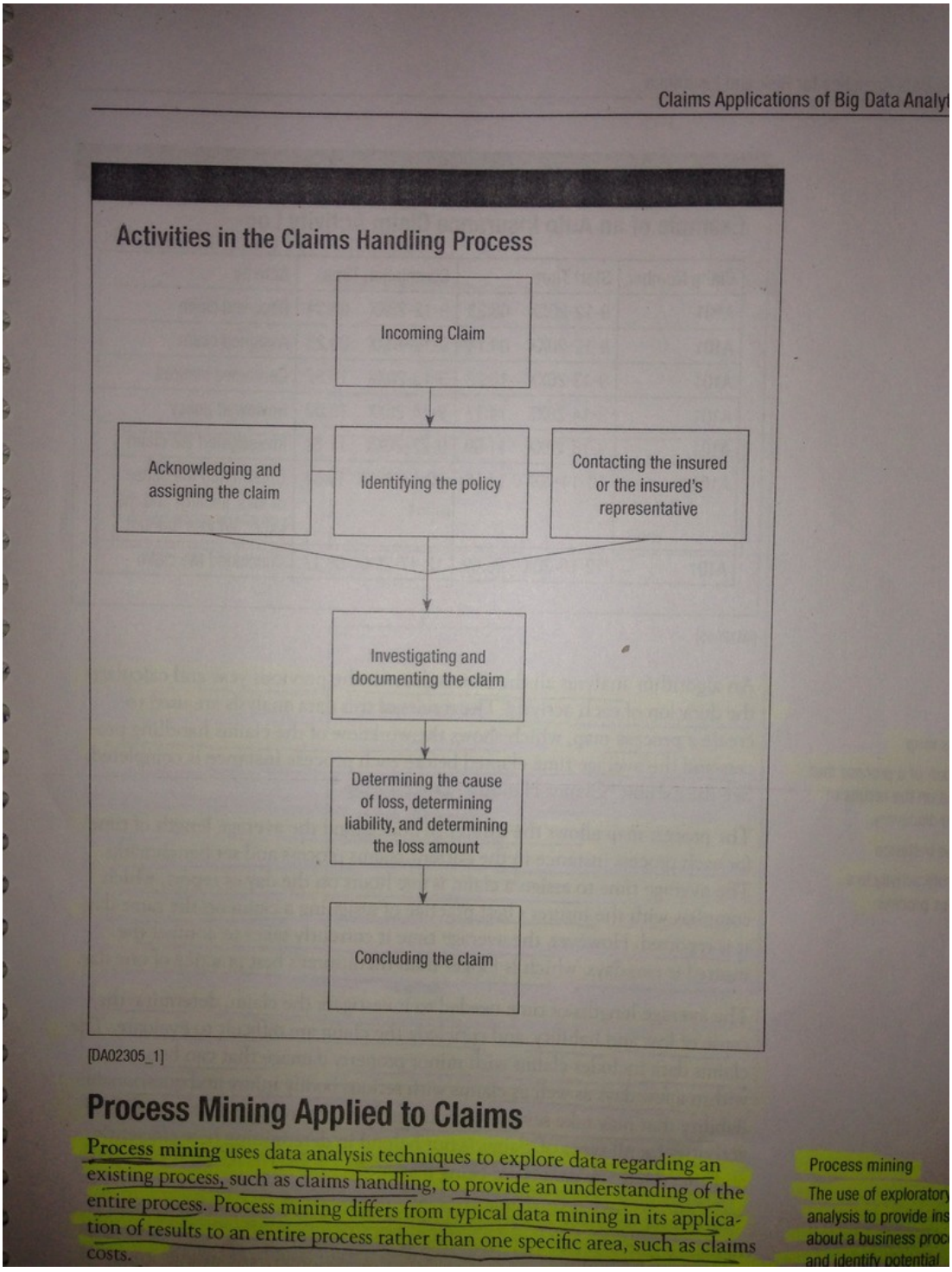
To illustrate how to improve claims processes using data analytics

1. Diagram the claims handling process, there are six major activities in the claims handling process:
    - acknowledging and assigning the claims
    - identifying the policy
    - contacting the insured or the insured's representative
    - investigating and documenting the claim
    - determining the cause of loss, liability, and the loss amount
    - concluding the claim

    - identify how the loss occurred, witness, potential fraud or subrogation opportunities, and the loss amount

1.
2. Decide how business process analytics can be applied to this process

    - to improve customer service, efficiency, and cost-effectiveness
    - BPM (Business Process Management) : A systemic, iterative plan to analyze and improve business processes through life-cycle phases to achieve long-term goals and client satisfaction
    - **Data analysis is used in gathering intelligence as well as developing models**

1. 

2. Process Mining applied to claims
    - starts with data analysis of claims activity logs
    - uses data analysis technique to explore data regarding an existing process, such as claims handling, to provide an understanding of the entire process. **Process mining differs from typical data mining on its application of results to an entire process rather than one specific area, such as claims costs**.
    - mining of the claims activity logs to identify claims handling activities and the timing of them. the first phase

in the process mining of the insurer's auto claims involves **process discovery** of claims activities. which shows the date and time when each activity began and ended - **an example indicator is Reviewed policy**
    - This information is then used to design classification tree models, for claims activities, such as claims assignment, fraud investigation, and subrogation potential

Process Mining : The use of exploratory data analysis to provide insights about a business process and identify potential improvements. Process is like data mining. It differs from typical data mining in its applications of results to an entire process rather than one specific area, such as claims costs

Process Discovery : The use of data analysis to identify the activities in a current process. it is sued to develop a process map of current claims activities, referred to as process instances, and the average duration of each process instance.

Process Map : **A diagram or workflow of a process that is based on the results of process discovery. which allows the insurer to know the average length of time for each process instance in a process and set benchmarks, made from claims logs.The process map would help insurance set the initial benchmarks for each activity in the current claims handling process based on internal metrics. The process map is a diagram of the process that is based on the results of the process discovery.**

Process instance : A discrete activity of a business process.

After process map, The insurer will need to combine process mining with data analysis techniques to gain a better understanding of what is involved in the investigation and resolution of different types of claims. Data analysis techniques include - classification techniques and clustering

Classification tree models can be used to design models for claims activities, such as claims assignment, fraud detection, and subrogation potential.

Business Process Management (BPM) : A systematic, iterative plan to analyse and improve business processes through life-cycle phases to achieve long-term goals and client satisfaction. The goal of BPM in claims handling is to improve claims efficiency, customer service and cost-effectiveness. BPM begins with explanatory analysis of the current process and then develops models for improvement. Data analysis is used on gathering intelligence as well as developing model.

Cluster analysis is an exploratory technique without predefined variables that can be used when attributes are not known.

An insurer should be prepared to Reevaluate the attributes in a predictive model.

To improve claims handling process, using of Cluster analysis is an exploratory technique without predefined variables that can be used when attributes are not known.

Chapter 6 - Risk Management Application of Data Analytics

In Workers Compensation Claims, through past accident data, it is often found that accident causation focuses on either single unsafe act or condition. *However, even the simplest accidents are the byproduct of multiple actions and complex interactions of conditions*.

- Using sensors is a powerful way to collect data about biological condition of workers and physical conditions of workplace environment.
- Using this data we can run advance predictive modelling or Machine Learning to improve the accuracy of forecasting accidents.

After data from sensors is categorized to develop attributes that help predict workplace accidents, these attributes are combined

with traditional workplace accident attributes and then analyzed independently for their information gain.

An organization's safety program may combine elements of these accident analysis techniques. For example, Fault Tree Analysis (FTA),which incorporates RCA concepts and FTA Techniques, can be used to identify potential accidents and predict the most likely system failures. Categories of Workplace accident analysis technique include :

- **System Safety** : A safety engineering technique also used as an approach to accident causation that considers the mutual effects of the interrelated elements of a system on one another throughout the system's life cycle. It analysis hazards and causes of hazards and estimate the probability of particular kinds of breakdowns and suggest cost-effective ways to prevent these system failures. It examines the organizations as a whole, and relies on specific techniques for determining how these hazards can lead to a system failures and accidents.
- **Root Cause Analysis (RCA)** : A systematics procedure that uses the results of the other analysis techniques to identify the predominant cause of the accident. it is used to identify the root cause of the accident and mitigate future actions, inactions, conditions or behaviour of such events.
- **Failure Mode and Effects analysis (FMEA)** : An analysis technique the reverses the direction of reasoning by starting with causes and branching out to consequences. The goal is to prevent or reduce the severity of a harmful event and address the event based on priority. ***The goal is is to prevent or reduce the severity of a harmful event by identifying the order in which critical failures should be addressed***. The specific actions (consequences) include eliminating the failure mode, minimizing, the severity (consequences). reducing the occurrence, and improving detection.
- **Criticality Analysis** : An analysis the identifies the critical components of a system and ranks the severity of losing each component. ***The goal is to prevent or reduce the severity of a harmful event and address the event based on priority***.

The goal is to prevent or reduce the severity of a harmful event by identifying the order in which critical failures should be addressed. The specific actions (consequences) include eliminating the failure mode, minimizing, the severity (consequences). reducing the occurrence, and improving detection.

- **Fault Tree Analysis (FTA)** : An analysis that takes a particular system failure and traces the events leading to the system failure backwards in time.Incorporates RCA concepts and FMEA techniques to identify various ways of "breaking" the fault tree; that is, it interrupts the sequence of events leading to system failure so that the failure itself can be prevented.

Failure mode and effects analysis is a traditional analysis technique that, when paired with criticality analysis, helps an organization prevent or reduce the severity of a harmful event by identifying the order in which critical failures should be addressed is failure mode and effects analysis.

In **Fault Tree Analysis (FTA)** the risk manager will examine the series of events and conditions that led to an accident. The accident appears at the op of the fault tree, and the events necessary to produce it appear as branches. This is known as fault tree "**TOP EVENT**"

- The tree's branches are connected by "and" gates and "or" gates. These gates represent the casual relationship between events, which are depicted as rectangles within the tree.
- A fault tree also can be used to calculate the probability of an accident if the probabilities of the casual events are known.

Limitations of FTA : If there is a high degree of uncertainty with underlying **or base events**, the probability of the accident may be ascertain. Additionally, important pathways to the accident might

not be explored if some casual events are not included in the fault tree.

*!! traditional analysis such as **FTA** cannot easily account for human error and conditional failures (which cause one another in succession, culminating in an accident) !!*

Holter monitors to detect cardiac arrests is an example of sensors. Limitations of sensors, is that if we cannot find the relationships among accumulated data points from all the sensors, which would explain the environmental conditions, worker attributes, and other relevant factors which affect the probability of an accident occurring. Its use will be not be fully realized

Steps to do Data Analytics

1. Collect the data and categorize.
2. The categorize are used to develop into similar attributes
3. Each attribute is then analyzed independently for its information gain
4. Supervised learning technique such as Classification tree algorithm is applied to training data. The algorithm executes recursively and shows how various combinations of attributes can be used to predict the probability of an accident occurring
5. Finally, holdout data is used to test the model and determine its predictive power on data that was not used in model development
6. Classification results from the combination of attributes connected to the rectangle through the sequence of arrows, each of which depicts the actual value of the attribute to which it is connected.
7. Probabilities are added to each of the rectangles based on the percentage of time the model correctly predicts the specified class when applied to the training data.

The Goal of the Classification tree algorithm is to determine the probability of an accident occurring or not occurring.

Assessing Reputation Risk through text mining and social network analysis

Reputation is an intangible asset the relates to an organization's goals and values, results from the behaviors and opinions of its stakeholders (stakeholder perception) and grows over time. It involves a comparison of stakeholder's experience and their expectations and is the pillar of the organization's legitimacy, or social license to operate.

Data analytics offers organization a way to keep up to date on the shifting perceptions regarding their services and products. Text mining and social network analysis will be used to come up with a plan of action.

An organization's risk manager can better target a response to a crisis by considering the combination of local variables and network variables in a social network.

Reputation Risk : The risk the negative publicity, whether true or not will damage a company's reputation and its ability to operate its business, it involves managing the organization's interactions with the general public, its stakeholders an its employees. Risk managers must use a systemic approach, to carefully manage the organizations interactions with the general public, its stakeholers and its employees.

Steps in Text Mining :

- find attributes to perform sentiment analysis one
- clean up by elimination spaces, punctuation's, and words that provide little information (stop words)
- reducing and words to their stems by removing prefixes and suffixes (stemming)
- create two dictionaries one with favorable words and another with unfavorable words
- Perform sentiment analysis term frequency (TF) and inverse document frequency(IDF) of certain terms

For example : One term is used multiple times within only one of the blog posts. It has a high TF within blog post and high IDF, which is a measure of its rarity within the corpus of blogs. T***he combination of high TF within the blog post and high IDF indicates that the term is worthy of consideration.***

Social Network Analysis

- Sociogram is for a small social network
- Matrix is for very large social network

Data scientists will examine social network connections through a sociogram to understand how far a negative sentiment that originates with one person or a few people can spread. The influence of the negative sentiment significantly depends on its posters' metrics given below:-

- Closeness: The measure of the average distance, or path length, between a particular node and the other node in a network
- Degree: The measure of the number of connections each node has
- Betweenness: The measure of how many times a particular node is part of the shortest path between two other nodes in a network.

!! Using Text Mining and Social Network Analysis, targeted messages can be sent to the correct group with more influence and effectively respond to reputation risk. !!

The centrality measures of degree, closeness, and betweenness matter not only in the context of one network, but also in the context of how one person or a few people can connect multiple social media networks.

An organization's risk manager can better target a response to a crisis by considering the combination of local variables and network variables in a social network.

A sociogram is particularly useful for analyzing small social networks

!! Organization should take precautions not to specifically identify individual customers, by combining pieces of information of the user. **This in itself is a reputation risk .** ***Data scientists continue to work on developing new ways to ensure individuals' privacy while still allowing organizations to reap the benefits of network analysis***.!!

Performing a network analysis after performing text mining on social data can provide information about how far and how quickly a sentiment has and could spread.

Using Clustering and Linear Modelling for Loss Development

It is important for risk management professionals to determine the ultimate value of losses as accurately as possible, as this is the biggest obligation of the insurer, the money left can be invested for insurance operations.

Long-Tail Claim : A claim in which there is a duration of more than one year from the date of loss to closure. *These claims are difficult to predict due to multiple factors*.

Ultimate Loss : The final paid amount for all losses in an accident year.

Adverse Development : Increasing claims costs for which the reserves are inadequate. If factors of adverse is not known then unsupervised learning like K-Means clustering can be sued to find patterns. Examples of Outliers can be Claim Size and Ratio of Incurred Loss to incurred loss at the eighteen month evaluation point (ratio > 3,5)

Excess Liability Insurance : Insurance coverage for losses that exceed the limits of underlying insurance coverage or a retention amount.

Cluster analysis is an exploratory technique without predefined variables that can be used when attributes are not known.: The sum of the paid losses and loss reserves and loss adjustment expense reserves. ***General liability losses are typically assumed to reach their ultimate loss value within 66 months.***

Loss Development factors, usually based on previous accident year's loss development data, are multiplied by the total incurred losses for each accident year to reach an estimate of the ultimate loss for that year.

"A closed claim is assumed to have reached the ultimate loss value."

Typical attributes of high severity claims

- Respond more than one week after the date of accident.
- Liability denied by the claim representative.
- Claimant represented by an attorney.
- Large Jury Awards

With the above factors of high severity claims, Data scientist can use a Generalized Linear Model (GLM) to project the average ultimate severity as the target variable for these types of claims based on their attributes. The claims reserves should be based on the results of the GLM.

Predictions of ultimate losses for specific accident years are based on total incurred losses. After a predictive model is developed to evaluate estimates of ultimate losses, the analysis should be repeated to determine accuracy.

Summary :

Workplace accidents and their causes are ideal subjects for the application of data analytics because traditional methods of workplace accidents analysis rely on human interpretation of limited amounts of information and cannot always account for how contributing factors combine to cause an accident.

Chapter 7 - Implementing a Data Analytics Strategy

In order to implement a sound and successful data analytics strategy - SWOT analysis is needed to be done. ***Especially for P&C insurers when the change entails a fundamental shift in philosophy, such as the one from reliance on traditional organizational infrastructures to data-driven analysis. The P &C Insurance industry, rooted in risk aversion and predicated on long-term strategy, has traditionally been low to embrace change***.

The emergence of data analytics-can therefore potentially gain a marketplace edge by adopting new practices or efficiently than its competitors. An organization's integration of data analytics begins with the acceptance of the fundamental concept that more decision-making information, used intelligently, is always better. One way an insurer can determine this is through SWOT analysis.

Underwriters can effectively analyze policy segmentation to make more-informed rate changes and improve loss ratios and customer retention when using a fully trained segmentation model through data analytics to find meaningful patterns in data and automatically segment policies.

An insurer could use predictive modeling to prioritize a claim for investigation based on the probability that it is fraudulent. Data mining techniques, such as text mining, social network analysis, and cluster analysis, can be used to extract the data.

SWOT (strength, weakness, opportunities and threats) Analysis : A method of evaluating the internal and external environments by assessing an organization's internal strengths and weakness and its external opportunities and threats. It allows an organization to consider both the general environment and its immediate environment. The method was devised by Albert S. Humphrey, a

business scholar and management consultant who specialized in business planning and change. The approach used by organizations varies based on each company's needs. **In SWOT Analysis a company should also thoroughly analyze how they affect its strategic plan**.

Because the SWOT list can be extensive , Understanding the business issue that has prompted the SWOT evaluation is very important. In the case of potential use of Data Analytics enables use of management to develop two general perspectives:

- **Internal (organizational)** strengths and weakness can be identified as financial, physical, human and organizational assets. based on factors like managerial expertise, available product lines, staff competencies, current strategies, customer loyalty, growth levels, organizational structure and distribution channels organizational structure and distribution channels. In a data analytics project, the risk manager's consideration of the internal environment begins with clarifying the objective of the project.
- **External factors**, which can be identified through trend analysis, it can be presented by new opportunities in new markets, possible acquisition targets, or a reduction in competition, economic downturns, or changes in customer preferences.

SWOT Analysis Table

	Strengths	Weaknesses
Internal	List assets, competencies, or attributes that enhance competitiveness Prioritize based on the quality of the strength and the relative importance of the strength	List lacking assets, competencies, or attributes that diminish competitiveness Prioritize based on the seriousness of the weakness and the relative importance of the weakness
	Opportunities	**Threats**
External	List conditions that could be exploited to create a competitive advantage Prioritize based on the potential of exploiting the opportunities	List conditions that diminish competitive advantage Prioritize based on the seriousness and probability of occurrence
	Note strengths that can be paired with opportunities as areas of competitive advantage	Note weaknesses that can be paired with threats as risks to be avoided

[DA03626]

Because an organization's strategies can be extensive, it is helpful to narrow the focus of a SWOT analysis by understanding the business issue that has prompted the evaluation. The strategies relevant to the issue can then be targeted for evaluation. In this case, a SWOT analysis focused specifically on the potential use of data analytics enables management to develop two general perspectives:

- Internal (organizational) strengths and weaknesses that could potentially be affected by a data analytics initiative—Identifying internal strengths and weaknesses involves consideration of financial, physical, human, and organizational assets.[1] Managers use SWOT analysis to determine the current state of their organizations based on factors like managerial expertise, available product lines, staff competencies, current strategies, customer loyalty, growth levels, organizational structure, and distribution channels.
- External factors that present opportunities for growth or threats to the organization's survival—One way to achieve this perspective is through trend analysis, which identifies patterns related to specific factors in the past and then projects those patterns into the future to determine potential threats or opportunities. Opportunities might be presented by new markets, possible acquisition targets, or a reduction in competition, while threats might include new competitors, an increase in competition levels, economic downturns, or changes in customer preferences.

A SWOT analysis focused specifically on the potential use of data analytics enables management to identify internal (organizational) strengths and weaknesses that could potentially be affected by a data analytics initiative and external factors that present opportunities for growth or threats to the organization's survival.

SWOT Analysis Process

Is done through a group activity that involves an organization's managers and is organized by a facilitator

- **Brainstorming**
  - If the group is large, the facilitator may divide the managers into smaller groups to encourage participation
  - through brainstorming, factors are listed under each of the SWOT headings.
  - This activity will produce many factors organized under each SWOT heading.
- **Refining**
  - To make the list easier to examine, similar items are clustered together.
  - Items of high importance are noted under each of the SWOT headings.
- **Prioritizing**
  - Strengths are ordered by quality and relative importance.
  - Weakness are ordered by the degree to which they affect performance and by their relative importance.
  - Opportunities are ordered by degree and probability of success.
  - Threats are ordered by degree and probability of occurrence.
  - Strengths that require little or no operational changes and can be paired with opportunities are designated for potential action to maximize competitive advantages that will entail low risk
  - Weaknesses that can be paired with threats are designated in the prioritized list for potential action to minimize consequences that entail high risk

Through Data Analysis a business can do below three

- *Reinforce Strengths by increasing its profitable business, although some opportunities may lie in reducing in costs as well.*

- Mitigate Weakness by using predictive modelling techniques like social network analysis, text mining and cluster analysis to better identify fraudulent claims. Use Machine Learning to identify patterns in underwriter guidelines, loss ratio to improve policy segmentation methods. Help in data driven decision making by intelligently deployed information
    - Inability to identify fraudulent claims
    - Diminishing effectiveness of existing policy segmentation methods
    - Pervasive conventional thinking

- Exploit Opportunities by hiring better talent using predictive modelling, analyze employee turnover and incentivize employees to remain engaged. Use text mining and social network analysis to accumulate as much publicly available intelligence about competitors. Make use of sentiment analysis algorithms, social network analysis, and other modelling techniques to measure the public attitudes about companies insurance policies.
    - Scarcity of new workers willing to enter the insurance industry : *Recruitment and retention challenges posed by the scarce entrance of new workers, customers' increased access to competitive information, and cynical public attitudes about insurers that create a more permissive environment for claims fraud are some of the technological, sociological, and industry trends that could fundamentally change the way insurers do business.*
    - Customers' increased access to information about competitors' prices
    - Cynical public attitudes about insurers : ***An insurer can use sentiment analysis algorithms, social network analysis, and other means of measuring public views contributing to claims fraud to address cynical public attitudes about insurers.***

Implementation of data analytics project and Management of inherent risks to implement a data analytics project is very important as the initiative itself

One way to manage risk is through project management. *Project Risk Management aims to optimize risk levels to achieve the project goals. Its underlying structured process allows risk managers to identify and assess a project's risks and to respond appropriately.*

In a data analytics project, the risk manager's consideration of the internal environment begins with clarifying the objective of the project.

Another way to manage risks is through Enterprise Risk Management (ERM) model where activities is done in five steps

- **Scan Environment** : Can be accomplished with an analysis of an organization's internal and external environments (external include Technological, Legal & regulatory and Political). ***In scanning the environment, the risk manger for an organization's data analytics initiative first needs to develop a sense of how the initiative will merge with its existing infrastructure and how the project could affect the organization's stakeholders. To hone the project's goal, the risk manager could interview various internal stakeholders and develop a series of quantifiable objectives, that, together constitute the project's overall goal.In a data analytics project, the risk manager's consideration of the internal environment begins with clarifying the objective of the project.***
    - ***The technological aspect of a data analytics initiative that involves the use of electronic monitoring and reporting equipment that must be installed in remote locations and tested and maintained for accurate reporting, security, proper location, and dependability is part of scanning the external environment.***
- **Identify Risks** : This is to determine which potential risks will require treatment.
- **Analyse Risks** : It is done to identify the components activities of the project and then examine (***by judging the likelihood and severity***) the risk for each activity. ***The risk manager can the prioritize the risk based on the highest priority given to risk which will extend the project beyond its timelines***.

- **Treat Risks** : Is done by avoidance, modification, transfer , retention or exploitation. *However the opportunities that accompany the risk should not be overlooked.* In addition project managers will involve contingency planning, which involves establishing alternative procedures for possible events thereby keeping the entire project on track within the project constraints.
- **Monitor and Assure** : To ensure each activity is within variance acceptable limits and in the budget allocation. ***Monitoring and assuring the data analytics project as it progresses involves the project manager comparing the quality, time, budget, and other constraints established for the project with the project status and determining whether the project goal will be achieved or whether resources should be reallocated to achieve the goal.***

ERM is done to ensure the risk management is in sync with the strategic goals and operational objectives of the specific project. This project envision real time data to make decision on the fly.

Risk avoidance, modification, transfer, retention, or exploitation, as well as contingency planning, are some of the major options for treating risk and keeping the entire project on track within variance requirements.

Data analytics change Management

The change management process entails is particularly difficult for a company when the change entails a fundamental shift in philosophy, such as the one from reliance on traditional organizational infrastructures to data-driven analysis.

An insurer's data analytics initiative can succeed only if the entire organization understands the value of intelligently deployed information and how to apply data-driven decision making.

Alignment of the data analytics project's goals with upper management's long-term organizational objectives is an

example of component activities associated with the data analytics initiative.

An organization's integration of data analytics begins with its acceptance of the fundamental concept that, when used intelligently, more decision-making information is always better.

To effectively articulate throughout the organization the need for a shift to data-driven processes as part of a data analytics initiative, top management should frame it as being urgent and crucial to the insurer's ability to acquire and retain business because competitors use data analytics techniques to improve their underwriting, pricing, fraud detection, and marketing.

An insurer's data analytics initiative should be guided by a vision statement that succinctly aligns, directs, and inspires collective action toward a common end point that improves the organization's measurable results.

The costs to implement machine learning techniques and to recruit, train, and retain personnel capable of executing data analytics strategies are costs related to the investment required to launch a data analytics technique that could undermine an insurer's cost-reduction strategy.

The leadership team for a major transformation in an organization is usually made up of one highly visible change agent and a collaborative team of sponsors of the change.

Analyzing the risks associated with the data analytics project includes establishing time estimates for crucial activities and assigning acceptable variances to be referenced in the event that they could cause results to fall outside the acceptable parameters.

- **Articulate the need for change** : **Sense of Urgency** helps persuade key individuals to invest in the change and creates the momentum helps persuade key individuals to invest in the

change and creates the momentum required to spur the organization to action.

Establishing a sense of urgency helps managers persuade key individuals to invest in the change and creates the momentum required to spur the organization to action.

- **Appoint a leadership team** : team of carefully selected individuals who recognize the need for change and provide solid, broad-based support for accomplishing it and create a **Center of Excellence (COE)**
- **Develop a written statement of the vision and strategies** : that is *realistic, desirable, feasible, motivational, and focused* enough to guide decision making. **A common error is to mistake plans and programs for an appropriate vision**.
    - It provides a general direction for improving the organization's products, services, cost control, and/or relationships with customers and stakeholders
    - It can be clearly described in five minutes or less
    - It considers trends in technology and the market
    - It is stated in positive terms
- **Communicate the vision and strategies** : can be in the form or arguments and eliminating signals against the message. the overarching vision and the mail supporting strategies to all employers using multiple fourms to ensure that everyone sees and hears the message repeatedly and consistently
- **Eliminate the barriers of change** : in the form of technological limitations, timelines, skill level and organization culture. The leadership team can try to eliminate barriers to change by establishing timelines that are reasonable and not inappropriately ambitious.
- **Recognize incremental successes** : Identify milestones through published goals and team charters. split goals into smaller goals this ensures the team communicate the relationship and relevance of the assigned actions to the overarching goal. ***Management and the COE may use milestones in a data analytics project to represent visible improvements or short-term accomplishments that are part of the progress toward the project goal.***

- **Entrench the change** : New systems, process and structures should be created so that employees can get the benefit of the new system in place and the positive outcome. Shared attitudes, values, goals, and persistent norms of behavior must be aligned with the change to data-driven philosophies.

A Center of Excellence (COE) address the need of strong business leadership and data analytics leadership. A COE consists of personnel dedicated solely to developing data analytics strategies, formulating department-specific data analytic goals, implementing plans to encourage organization-wide collaboration on data analytics initiatives, and overseeing the evolution of a data analytics model from concept to adoption.

The members of COE are responsible for creating the vision of the change and selling it to others. also members of the COE should have the power to eliminate any obstacles to progress, expertise relevant to the change, the credibility to convince others that the need for change should be taken seriously, and leadership to catalyse the change process.

Because it is neither practical nor feasible for a risk manager to identify all of the risks associated with a data analytics project, it is important to identify key and emerging risks so they can be analyzed for their effect on the project.

Upon completion of the data analytics project, individuals may not be able to see the broader perspective or effect on the change while it is taking place, so management may need to reinforce the vision.

Risk avoidance, modification, transfer, retention, or exploitation, as well as contingency planning, are some of the major options for treating risk and keeping the entire project on track within variance requirements.

Posted 3 years ago
March 16, 2017
11:17 AM
1978 views

3 responses

Do we need to have understanding for r-programming/python for doing Aida 181

— Amit

Do you have the AIDA book to sell?

— rajesh goyal

1 visitor upvoted this post.

We'll email you when there are new posts here.

Salman Ahmed

Lives and Works in Hyderabad, India.
Will love to answer to your
comments over a cup of tea. If you
don't hate me then you can find me
on other social networks
https://in.linkedin.com/in/salmanahmed84
-- @Linkedin
https://www.facebook.com/salmanahmed.84
-- @Facebook milleniumsalman --
@Skype (970) 344-1949 --
@Whatsapp
milleniumsalman@gmail.com -- @
Gmail

Search this site...

Browse the Archive »