9 – Big Data Analytics for managing Risk

# 1 – Big Data Characteristics and Sources

Objective: Categorize the various characteristics and sources of big data available for insurance and risk management applications

Insurers’ traditional data is organized into databases with defined fields. Insurance professionals can produce reports that show results from this data. Risk managers can usually access their insurers’ data to produce reports about claims they are handling. Big data has introduced different types and sources of data than those traditionally used.

It is important for risk management and insurance professionals to understand the different types of big data (sets of data that are too large to be gathered and analyzed by traditional methods). The varieties, volume, and sources of data are rapidly increasing. For example, risk managers can obtain new types of data on safety from sources such as sensors on employees. Underwriters can obtain new data on risks from drones and social media. Claim adjusters can better identify fraud by identifying pattern in internal and external data. Risk management and insurance professionals increasing work with data scientists to determine the types of data that are useful to make business decisions.

## Data Characteristics

The term “bid data” implies large quantities of data. Although this is true, big data is also different from traditional data in other respects.

**These are the characteristics that differentiate big data from traditional data:**

* **Volume** – there is enormous amount of data that is now available, and the amount continues to increase.
* **Variety** – traditionally, insurers used structured data. Big data also includes unstructured data in larger volume.
* **Velocity** – **this is the constantly increasing speed at which data arrives at an insurer. Velocity also includes the growing rate of change in types of data.**
* **Veracity** – this refers to the completeness and accuracy of data. Unstructured data is likely to have less veracity than structured data.
* **Value** – value is derived from the results of data analysis to help insurers make better decisions. Big data has great potential to add value, but it must be obtained and analyzed with techniques that provide meaningful results. This is the goal of data science.

**The velocity of traditional data and big data differs because big data has a constantly increasing velocity at which it arrives at an insurer compared to traditional data. The velocity of big data also includes the growing rate of change in the types of data compared to traditional data**.

## Internal and External Data

Data science allows insurers access to increasingly larger and more varied data, referred to as big data. Some of this data is received directly by the insurer of the risk manager’s organization. Other data is obtained from outside sources.

### Internal Data

**Internal data is data that is owned by an organization. Insurers have always relied on data, and they possess large quantities of it. Certain types of data, such as risk factors, losses, premium, rating factors, rates, and customer information, have been traditionally used to make business decision. Much of this data is also reported to state rating bureaus. Similarly, risk managers also have relied on data about losses and premium, in addition to safety statistics that are reported to OSHA.**

**discover**. Example; an insurer may not analyze computer claims codes about claimant’s preexisting medical conditions – traditional internal data that could be useful if data science techniques are applied to it

Data science also provides techniques to use **nontraditional internal data; for example, voice analysis can be applied to recorded statements of claimants to identify vocal characteristics associated with fraud, such as gaps in the claimant’s version of the accident or a defensive tone.**  Underwriters often obtain photos of property, and claim adjuster often obtain photos of an accident scene. Artificial intelligence can analyze these photos to find information that may be missed by human analysis, such as potential environmental hazards around a property, or through facial recognition, the identity of a witness to an accident.

### External Data

**There is sometimes a blurring of the boundary between internal and external data. For example, data from telematics is obtained from a device that is installed on a customer’s vehicle. The insurer owns the device, but the customer owns the vehicle. Because of the customer’s ownership of the vehicle and the nature of the data provided – the customer’s personal driving habits – telematics should probably be categorized as external data**.

Another source of data with a blurred boundary is a wearable sensor used by an employee while working. The employer provides the sensor, but it is placed on t he employee’s person and transmits information about the individual. This information could be considered internal because it is used in the workplace and because the sensor is owned by the employer. However, it could also be considered external data because it is obtained from and about an individual.

Other sources of external data include sensors in customers’ homes or commercial buildings, photos of a catastrophe obtained by drones, and the internet – especially social media. For insurers to access information from sensors on customers’ property, customers must first grant permission. Regulators have already approved the use of drones by insurers to inspect catastrophe scenes. However, there may be other regulations governing what information can be used and how it can be used.

**Privacy issues are a legal and regulatory concern in obtaining mass information from public sources, including social media. This is particularly true with data science because it is used to discover and analyze previously unknown data relationships. Example, an insurance company might access individual items of data on a claimant from both internal and external sources. While the data from each source by itself might not violate the claimant’s privacy, combining all the data might lead to a conclusion that harm the claimant’s privacy in a way that he or she did not anticipate**. As a result, privacy regulations have been passed by various states and countries. (Cybersecurity Bill of Rights).

Other sources of external data include statistical plan data, other aggregated insurance industry data, competitors’ rate filings, and third-party data. The most commonly used types of third-party data are economic data (interest rates, asset prices, exchange rates, the Consumer Price Index, and other info about the global, national or regional economy) and geodemographic data (classifications of a population). Additionally, insurers obtain and use credit rating for both personal and commercial lines of insurance where permitted by law.

## Structures and Unstructured Data

Data can be either structured or unstructured. **The traditional data that insurers use for underwriting and claims management is structured. This data is organized into databases that are often linked to each other**. Example; an underwriter could query one database regarding a customer’s claim history and another database regarding the customer’s premium history. The underwriter could receive this information in the same report because the databases are linked. Telematics provide an example of structured external data. There are predetermined fields in a database for the driving information that insurers receive from telematics, such as distance traveled, braking, and left turns.

**Unstructured data is not organized**. **Adjuster’s notes are an example of internal unstructured data. Although the notes are contained in individual claim files identified by numbers, there is no database that categorizes the data contained in notes and provides results.** Unstructured external data includes information from the Internet, such as social media.

Data science is used for both structured and unstructured data, but it is especially useful for unstructured data. Data scientists use various techniques to gather, categorize, and analyze unstructured data. This process may or may not produce useful results.

**Big Data Categories and Examples**

|  |  |  |
| --- | --- | --- |
|  | Structured | Unstructured |
| **External** | Telematics  Financial Data  Labor Statistics | **Social Media**  News Reports  Internet Videos |
| **Internal** | Policy Information  Claims History  Customer Data | Adjuster Notes  Customer Voice Records  Surveillance Videos |

# 2 – Data-Driven Decision Making

Objective: Explain how data-driven decision making applies to risk management and insurance

Data-driven decision making – an organizational process to gather and analyze relevant and verifiable data and then evaluate the results to guide business strategies.

Risk managers and insurers have traditionally made many decisions based on data. Insurers use their own and industry loss data to price insurance products. Risk managers use accident data to reduce losses. However, data analytics can improve the types of data, methods of analysis, and results by applying data driven decision making.

Risk management and insurance were previously dependent on human analysis of data. Computers increasingly were able to provide data, such as losses and premiums, that could be segmented based on the business need. Data science provides more advanced methods of aggregating and analyzing data.

## Data-Science and Data-Driven Decision Making

Data science can help insurers and risk managers improve their business results through data-driven decision making. These are several ways to achieve improvements:

* **Automating decision making for improved accuracy and efficiency** – online insurance quotes
* **Organizing large volumes of new data**  - organizing according to multiple characteristics (telematics – speed, braking patterns, distance traveled, and left turns)
* **Discovering new relationships in data** – characteristics of workers never filed a claim and how to improve safety for all workers
* **Exploring new sources of data** – text mining to analyze claim adjuster notes for various purposes such as developing an automated system to predict claim severity and assign the appropriate resources to those claims

Across all types of business, perceived improvement in companies using big data has increased by 26% and is expected to increase by 41% in the next few years. Currently insurers only use approximately 10-15% of the available structured data.

**Regarding the sources of big data and data science for decision making is compelling. Insurers and risk managers have access to a wealth of existing data they are not tapping. Additionally, there is a rapidly growing quantity of external data from various sources, such as the internet at large and the internet of things**.

**Insurers, as well as other organizations, acquire data engineering and processing technology that can manage big data, which is too large for most conventional systems. Models can then be developed to gather and analyze data in the context of an insurer’s areas of interest, with results ultimately provided to the data analytics team or the manager who requested the data analysis. The appropriate person(s) can make data-decisions accordingly**.

Example of a Data-Driven Decision-Making Process: Automated for Underwriting personal auto

|  |  |  |
| --- | --- | --- |
| Data Engineering and Processing Technology: | Data Science: | Data-Driven Decision Making: |
| Data, computers, and software able to manage big data | Model for automated decision making.  Machine learning to continually improve the model.  Analysis for data-driven decision making | Managers evaluate the results of the data analytics and automated decision making to develop business strategies |

Data-driven decision making can be applied across an insurer’s or risk manger’s enterprise to solve a variety of business problems, achieve greater efficiency, and provide a competitive advantage. There are two basic approaches to data-driven decision making: descriptive and predictive.

## The Descriptive Approach

**The 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**.

Example; to increase market share, an insurer changed its underwriting guidelines for auto insurance one year ago. It now wants to obtain data regarding the accident rates for insures who were accepted for coverage during that period. The insurer also wants to compare those insureds with those accepted prior to the change in the guidelines. The insurer intends to use the results of this analysis to determine if it now three-year guideline is producing acceptable results. Using data driven decisions making, executive will decide whether to continue with the new guidelines.

## The 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**.

Example; automated underwriting for personal auto insurance is predictive approach that is used each time a person applies for insurance. The computer make the underwriting decision and issue a price quote.

Another example; a risk manager may receive data from sensors regarding which employees are using required safety equipment. The risk manager will make a decision about how to address the problem of employees who are not using the safety equipment.

Ain an additional example, predictive analytics can be used to assign claims to an adjuster. However, the claims manger may have authority to reassign a claim as a minor claim that the computer assigns as a minor claim if the claim turns out to be sever.

## A Model for Data-Driven Decision Making in Risk Management and Insurance

Following the process outlined in the decision-making model will help ensure the best results. **The important first step is to define the risk management or insurance problem. Without a business context, modeling and analyzing data is unlikely to be effective.** Example; a risk manager for a large national retailer observes that customer slip and falls are increasing in both frequency and severity. The risk manager wants to determine why this increase is occurring and develop a solution for the problem. This involves a descriptive approach because the problem requires a one-time analysis.

After defining the problem, quality data must be gathered to obtain information about the customer accidents. The risk manager who has access to the liability insurer’s claims system decides to use the insurer’s claim data. The variable to be used are date of accident; store location; area of store property where the accident occurred; weather data, and sales volume at the store on the date of the accident. The risk manager runs a report to identify missing or obviously inaccurate data, such as a date of loss stated as 1/1/1898. The risk manager eliminates claims with missing or inaccurate data from the data set.

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 falls at stores located in northern locations in December during snowstorms. Sales volume was high because of the holiday shopping season. Also, there was an increase in falls at the stored in Florida in August during tropical storms. Sales volumes was also higher in August because of back-to-school shipping.

The risk manager is able to make several decisions driven by this data.

**Risk Management and Insurance Data Analytics Decision-Making Model**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **\*** | **Feedback** | **\*** |  |
| Define Risk Management of Insurance Problem | Data | Analysis and Modeling | Insights | Actionable Decisions |
|  | Data Selection  Data Quality  Data Preparation  Data Understanding | Big data analysis techniques  Descriptive and predictive models | Trends  Relationships  Behaviors  Events | Risk pricing  Risk Selection  Claim Fraud Det  Claim assignment  Product develop  Risk identification  Risk assessment  Loss control |

**The first step in the data driven decision making model is to define the risk management or insurance problem. The next three steps are feedback steps from the data science approach: identifying the data and ensuring its quality; modeling the data using big data analysis techniques; and achieving insights. After the feedback steps, decisions are made.**

This example illustrates how the model for data-driven decision making can be applied in risk management and insurance context. **The model began with the important first step of clearly defining the problem. Appropriate data was selected and prepared by removing claims with missing or inaccurate data. Relevant variable were selected. There was interaction between the risk manager and the data analytics team to discuss the problem. That data analytics team developed a method to discover patterns and correlations in the data. The risk manger was able to use the results to make a data-driven solution**.

# 3 – Training and Evaluating A Predictive Model

Objective: Explain how to train and evaluate a predictive model

Many organizations, including insurers use predictive models (a model used to predict an unknown outcome by means of a defined target variable) to guide future business decisions. Example; proper implementation of a 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. Therefore, during the development of predictive models, businesses must be able to assess the models’ effectiveness and reliability.

One way in which insurance professionals and risk managers can employ predictive modeling to their advantage is to help them anticipate behaviors. Example; if an organization can predict which of its employees are likely to have workplace injuries, it could develop strategies to improve these employees on the job safety.

The first step in building the predictive model is to supply it with data about employees who did and did not have accidents. The manufacture decides to limit the data it uses to that from the past year. The manufacturer then comes up with fifteen attributes that describe each employee, such as shift, experience, whether they use safety device.

Now assume that the model is made to account for each of the fifteen attributes as they appear in the dataset. The model is then used on the 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) and the model appears to be very accurate. The manufacture is pleased with the results of the model and decides to move it into production, where it will make real-life predictions about employees. After time they determined the information the model was providing was not accurate flagging only 1% of the employees, when the employer new according to historical data roughly 5%.

Most likely the model was tailored too closely to the training data. The mistake could have been avoided by initially saving some of the training data for evaluation and by applying performance metrics to the model before moving it into production.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Training the Model | | | Evaluating the Model | | |
|  | **\*** | **Feedback** | | **\*** |  |
| Obtain data and split into training data and holdout data | Build the model using the training data | Apply the model to the holdout data, suing cross validation as needed | | Evaluate the model using performance metrics | Put the model into production, reevaluating as needed |

## Training the Model

A predictive model is trained using data that has known values for the target variable (the predefined attribute whose value is being predicted in a data analytical model. In the example of the workplace injury predictive model, the class labels (the value of the target variable in a model) of the target variable are “yes, workplace accident”, and “no, no workplace accident”.

## Training Data

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. For the workplace injury model, the insurer selected 15 attributes to describe each instance (the representation of a data point described by a set of attributes within a model’s dataset). **The model must be made complex enough to be accurate. If the model used only the attributes of shift and safety device, that might not provide enough information to make accurate predictions. However, if too many attributes are used, a model can easily overfit (the process of fitting a model too closely to the training data for the model to be effective on other data) the training data**.

## Holdout Data

Holdout data is data that is held out from the training data**. Like the training data, the holdout data must have known class labels for the model’s target variable. The holdout data is used to test the model to make sure it performs well on data that was not used in its development and therefore, helps to ensure that the model is not overfitted to the training data**.

**Using the holdout data, the predictive model indicates a correlation between second-shift employees and workplace injuries that is similar to that it discovered in the training data. This is a good indication that the model can apply generalization (the ability of a model to apply itself to data outside the training data). The model must have some complexity to be accurate, but if it has too much complexity (in other words, it is overfit), it will not be accurate beyond the training data**.

## Cross-Validation

Sometimes, a model’s performance on the holdout data will not sufficiently reassure its developers that it will perform well in production. Example, the risk manager at the clothing manufacturer may not be comfortable flagging employees for additional safety training after seeing only the predictive model’s performance on the training data and holdout data. At other times, a very limited amount of training data may be available, and the model’s developers may think it unwise to not use some of the data for training because of the need for holdout data. In either one of these scenarios, cross-validation may be the appropriate solution.

**Cross validation is the process of splitting the 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 version that performs best. The selected model can then be trained using the entire data set**.

|  |  |  |
| --- | --- | --- |
| Total Data (split into 3 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 |

**Evaluating the Model**

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

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. Exhibit “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 EEs** |
| **Actual No** | 945 = TN |  | 5 = FP |  | 950 |
| **Actual Yes** | 10 = FN | Recall TP/TP+FN | 40 = TP |  | 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).

A data scientist or an actuary can recommend the best evaluation methods for a particular model, including these;

* **Accuracy – this is simply a measure of how often the model predicts the correct outcome:**
  + **(TP) + (TN) / (TP + TN + FP + FN)**

Based on the numbers in the exhibit, the accuracy of the workplace injury predictive model is calculated as

(40 + 945) / (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 / (TP + FP)**

Based on the numbers in the exhibit, the precision of the workplace injury predictive model is calculated as

40 / (40 + 5) = 0.889

* **Recall – this is a measure of how well the model catches positive results:**
  + **TP / (TP + FN)**

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 x (precision x recall) / (precision + recall)

Based on the numbers in the previous equations, the workplace injury predive model’s F-Score is calculated

2 x (0.889 x 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 good business decisions, then the model should be reevaluated. Returning to the workplace injury predictive model example, if one year after the model has been used in production, the manufacturer’s risk manager finds that it has been flagging the wrong employees as being likely to have an accident and that the company has wasted resources providing additional safety training for these employees the model should be retrained and reevaluated.

Insurance and risk management professionals should also defer to their professional experience when examining a model’s results. If the results do not seem to make sense, there is a good chance that the model was overfitted or is not complex enough. Insurance professionals and risk managers should also keep in mind that situations change. No predictive model will make accurate predictions indefinitely. When significant economic, technological, or cultural changes occur, they should work with data scientists or actuaries to update predictive models and retrain them to provide the best results.

# 4 – Overview of Big Data Analysis Techniques

Objective: Summarize big data analysis techniques and their application to risk management and Insurance

**There are two main methods used for data analysis: Supervised learning and unsupervised learning. In supervised learning, both the attributes and the target variable are defined. In unsupervised learning, there is not a defined target variable**.

For example, determining the premium for an applicant for personal auto insurance involves supervised machine learning (artificial intelligence in which computers continually teach themselves to make better decisions based on previous results and new data). Using Machine learning, a model is created to apply defined variables, such as age, driving experience, and accident record, to determine a defined target variable – in this case, the premium.

In a different example, creating a model to discover ideas for new insurance products using data from social networks involves unsupervised learning. The attributes are unknown, and there is not target variable. The model explores data to find relationships that can be further analyzed.

Certain techniques are used mainly for supervised learning, while others are used mainly for unsupervised learning. There is often an interrelationship between the techniques. For example, after information is obtain through unsupervised learning a predictive model could be developed using that information for supervised learning.

## Exploratory Data

Exploratory data analysis is a valuable approach to a business problem that can be used before developing and testing a model. The information from exploratory analysis provides a basic understanding of the data. The analyst can also obtain information about missing or inaccurate data.

The techniques used for exploratory data analysis are relatively simple. The involve charts and graphs that show data patterns and correlations among data. Example, a scatter plot, which is a two-dimensional plot of point values, could show the relationship between two attributes, such as driver age and average accident frequency. A bubble plot is a type of scatter plot in which the size of the bubble represents a third attribute for which there is data. An insurer, for example, may want to determine average accident severity in addition to average accident frequency by driver age.

A correlation matrix is another type of exploratory data analysis technique. Example, a risk manager identified a few worker attributes and would like to know whether they indicate that an employee is likely to work safely. The risk manager would like to develop a predictive model to identify the attributes of employees who are most likely to have an accident. These attributes can be explored using a correlation matrix. The stronger the correlation, the darker the cell color shade in the matrix.

The matrix indicated, not surprisingly, that certain combinations of attributes correlated strongly with worker accidents. Having a prior accident in combination with certain other attributes, such as variable shift, wage rate, and overtime, is particularly correlated with worker accidents. The risk manager then observed the correlations between the attribute of the work experience and the attributes of shift and overtime separately. The more experienced employees usually work rotating shifts to ensure that an experienced employee was always present. Additionally, they are given preference for overtime work. Because experienced employees work variable shifts and long hours, they may be tired and therefore prone to having an accident. As a result of this exploratory analysis, the risk manger realized that consulting with HR about worker overtime and shift assignment would be advisable before attempting to develop a predictive model.

## Data Analysis Techniques

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 evolve.

However, a few techniques are typically used, which can be adapted for each situation. Example, segmentation (data is divided into categories) is used in both supervised and unsupervised learning techniques.

More techniques employ machine learning in which a computer will continually refine the model as additional data and results are received. Example, **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**.

## Traditional Data Analysis Techniques

Insurers and risk managers have relied on traditional data analysis techniques to solve business problems, such as determining rates for insurance products and reserves for unpaid claims. These techniques are usually applied 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.

Usually and insurer or a risk manager wants to determine one of the following types of outcomes through *traditional data analysis*:

* A non-numerical category into which data belongs – whether an accident involves injuries only, whether employees have had an accident
* A numerical answer based on data – analyze the data an applicant enters online to determine the premium for auto insurance
* A probability score based on historical data – determine the probability that a safety solution will reduce accident frequency, or determine the probability of a commercial insured having a catastrophic loss
* A prediction of future results based on current and past data – determine the ultimate cost of claims for a line of insurance, or predict the cost of employees’ back injuries

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

## Classification Trees

**When using data analysis to solve business problems, the data usually must have certain characteristics to be useful.** Example, if a risk manager want to determine whether an employee is likely to be able to return to work after an accident, the attributes might include type of occupation, age, qualifications for retraining, and available positions with lighter physical requirements.

**A classification tree is a supervised learning technique that uses a structure similar to a tree to segment data according to known attributes to determine the value of a categorical target variable**. In this example, a classification tree could be used to segment workers compensation data. **The tree would contain nodes, arrows, and leaf nodes. The leaf nodes of the tree indicate the values of the target variable. Each worker’s attributes would follow the arrows of the tree through leaf nodes that segment the data by attribute to an ultimate leaf node labeled “return to work,”, or ‘”not return to work”** **– the target variable**. It is important to understand that these classifications are not necessarily what the actual outcomes will be.

## Regression Models

**In many cases, insurance or risk management professionals want to determine a numerical value for a target variable rather than, as in the above, a categorical value. In these cases, mathematical functions, such as those used in statistical regression techniques, are used to create algorithms. One of these methods is linear regression (a statistical method to predict the numerical value of a target variable based on the values of explanatory variables).**

Example, a claims manger wants to determine the relationship between two attributes, the number of claim in a claims representatives caseload and the number of years of experience as a claims representative – and the target variable, which is the ratio of closed claims during a specified period of time.

Training and holdout data, in which the value of the attributes and the target variable are known, is used to develop and test the model. Since the target variable is numerical, linear regression is used as the algorithm. The linear regression algorithm is developed using a method that minimizes the errors, represented by the differences between the actual target variable values in the data and those forecast by the model.

Another type of linear model that can be used for more complex data classification is a generalized linear model (a statistical technique that increase the flexibility of a linear model by linking it with a nonlinear function). This should not be confused with a general linear model, which is broad group of different types of linear models. 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 explanator variable (attributes). The third component is a 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), which relates the results of the random and systematic components.

## Cluster Analysis

**Cluster analysis is a technique for unsupervised learning. The technique is commonly used when an insurer knows a general problem it wants to solve but does not know the variable it must analyze to do so. The model results result in the clustering of data around particular values of explanatory variables.**

Example, an insurer might want to determine the characteristics that increase the severity of a claim so that it can assign claims that are likely to become sever to the most appropriate resources as soon as possible. The data analyst would develop a model to analyze a dataset of the insurer’s severe claims. The results could be the clustering of claims around particular values of explanatory variables. Such as a certain age, certain month of the year. The information gained through cluster analysis could the be applied to develop a predictive model using one of the other data analysis techniques that employs supervised learning.

**New Data Analysis Techniques**

The convergence of big data and new technology has resulted in advances that enable data scientist to analyze data and provide meaningful information that can guide business strategies. Data analytics has to be able to handle not only large volumes of data but also rapidly increasing velocity of data. Additionally, new analytical techniques are required to analyze textual and social network data. In the future, artificial intelligence will evolve further to provide better analysis of data from visual images and auditory sources. Big data and data generated from new technologies are often unstructured and therefore require new techniques for analysis.

## Text Mining

Text mining is becoming more common, especially with the growth of the internet and social media. Developing models to analyze text is more difficult than developing models to analyze other types of data. There are no organization fields and usually no numerical values. However, to continue the example of determining claims severity, the insurer may want to have a model perform text mining of adjuster notes for potentially sever claims. One modeling approach is to search for words that are neither rare nor too common. Another modeling approach is to look for adjacent words; the words “serious” and “severe” combined with “injury” would likely be indicators of claim severity.

## Social Network Analysis

Social network analysis is increased used by organizations. People use social media are familiar with product advertisements popping up while they view social media sites. These advertisements are based on data analysis algorithms that predict a person’s interest in a product based on preferences or associations in social networks. Social network analysis, sometimes referred to simply network analysis, does not apply only to social media sites such as Facebook and LinkedIn. It can be applied to any type of network in a business setting, such as email traffic between people and links between web pages.

This type of analysis is useful in detecting claims fraud rings. Many claims departments are already using social media analysis to identify fraud in individual claims.

## Neural Networks

**Neural networks are relatively new technique in data mining. They operate similarly to the human brain in their ability to infer rules from data patterns and then construct logic to use for analysis. A neural network technique produces a model that consists of layers of nodes with weighted links between the nodes. This technique can be used for both supervised an unsupervised learning**.

There are a few disadvantages of neural networks. One is that the way the rules and logic are developed by neural algorithms can be opaque. Another is the comparative difficulty in interpreting the information gained from the weighted links. However, the unlimited possibilities of neural networks have resulted in academic and business endeavors to improve the computer software that develops and processes them.

A team of actuaries used neural networks to develop a model to predict the duration of workers compensation indemnity claims that had been open for 6 months or more since the data of accident and that were not determined to be permanent total disability claims (which, theoretically, would remain open until the claimant’s death). The data used to train the model were closed indemnity claims that had been open more than six months before the date of closure. The input data formed the first layer of the neural network. The input then activated nodes in the hidden layer. The hidden layer generated output to the output layer. The neural network was able to model complex relationships between the input and output variables. Predicting the duration of workers compensation claims involves many input variables, and the relationship of these variables and their weight in predictions are complex. In this example, the input variables were accident code, gender, weekly wage, zip code, injury types, class code, body part, nature of injury, and age at the time of injury. The algorithms in the neural network were able to make accurate predictions of claims duration.

# 5 - Using Classification tree Analysis in Claims Assignment

Objective: Explain how classification tree analysis can be used to identify and assign complex claims

A relatively small percentage of claims account for mush of the costs to insurers. Some of the high-cost claims can be easily identified at the time of first report. However, many claims develop gradually into significant losses.

When an insurer identifies a serious injury, it usually assigns an experienced claims adjuster to handle the claim. It also often assigns a nurse case manager to help manage the medical treatment of the injured worker.

To use these resources effectively, the claims or risk management professional must identify the claim as serious or potentially serious. Many insurers are using predictive analytics to do this as soon as a claim is reported.

An insurer has been experiencing an increase in its workers compensation loss ratio for several years. An actuarial review of the claims data indicates that there has been an increase in the cost of complex claims (a claim that contains one or more characteristics that cause it to cost more than the average claim).

## Overview of the procedure

The data science team included a claims manger, along with a data scientist and an actuary from the review team. This group began by examining data from the actuarial review to identify characteristics of claims that gradually developed into severe losses.

After identifying these characteristics, the team decided on the type of analytical technique to use to develop a predictive model.

## Information Gain By Complex Claims Attributes

The data science team looked for attributes that were known within a week of first report in claims data. for gradually developing complex claims.

**Each attribute was independently analyzed for the information gain (measure of the predictive power of one or more attributes) it provides in classifying a claim within a week of first report as “complex” or “not complex”. These attributes were then ranked according to their relative information gain**.

## Using a classification Tree to illustrate Attributes of Complex Claims

**The data science team used a classification tree technique to develop a predictive model based on the attributes of complex claims. With a machine learning algorithm, the computer recursively (successively apply a model) built a tree by analyzing different splits in the values of attributes in various sequences**.

Note that the sequence of attributes in the tree does not strictly follow the relative information-gain ranking of the attributes. That is usually the case, because the information-gain ranking for the attributes was developed independently of other attributes. However, the attributes represented by each node in the tree depend on the attributes (and their values) that sit above them.

**Because whether the claim was reported more than five days after the date of injury is determined to be the most informative attribute, it is listed as the root node (the first node in a classification tree). The most informative attribute is listed as the root node. This first node in a classification tree. For example, in a in a classification tree developed to predict complex claims, the root node may be “late reporting of a claim”. A leaf node is a terminal node in a classification tree that classifies a claim as complex or not complex based on its attributes**.

Note combination of nodes and branches that lead to a “complex claim” classification at the leaf nodes.

## Validating the Complex Claim Model

The data science team used a portion of the claims data to **develop the predictive model. This data was training data. The target variable is categorical, indicating whether or not a claim is classified as “complex” or “not complex”. Holdout data was used to test the classification tree model to determine its predictive accuracy on data that was not used in its development.**

**A key measure of the predictive model’s effectiveness is lift (in a model performance evaluation, 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). Lift is based on a model’s accuracy and precision**.

The insurer used 10,000 claims with known outcomes to evaluate the model’s lift, of which 3,000, or 30% are complex. The model predicts that 2,000, or 20% of the total claims are complex (positive predictions) and the model’s precision is 60% meaning that 1,200 of the predicted complex claims are truly complex (true positives) and 800 are not false positives. Then the claims are ordered with those predicted as “complex” listed before ones predicted as “not complex”

Assume that the insurer’s experienced adjusters have the capacity to handle only an aggregate of 2,000 claims. If the first 2,000 claims in the order specified by the model (therefore, all 2,000 are predicted as “complex”) were assigned to experienced adjusters, then on average, 1,200 (40% of the 3,000 complex claims) would actually turn out to be complex. This is a significant improvement over randomly assigning 2,000 (20%) of the total claims) to the experience adjusters, for which on average, only 500 (20% of the 3,000 complex claims) would be complex**. The lift is calculated by dividing the model percentage by the non-model percentage. Therefore, 40% is divided by 20% resulting in a lift of 2**.

## Complex Claims Reporting

The insurer’s claims reporting procedure begins with a claim reported online or by telephone to its claim intake unit. The intake unit then directs claims to the appropriate office based on the geographical location of the accident. Each office has a claims assignment team that assigns claims based on whether the first report indicates that the injured employee is losing time from work. Catastrophic claims, such as severe burns are assigned to a senior claim adjuster.

**The data science team recommended that one office test the predictive model on new claims. Each new claim determined likely to be complex by the model is assigned to an experienced claim adjuster. At the end of a month, the data science team reviews the results to determine the effectiveness of the model. Through machine learning, the model adjusts the weights assigned to each of the attributes to better predict complexity**.

As claims are better identified as complex at first report, an insurer can more effectively use claims resources.