REPUBLIC OF THE PHILIPPINES
POLYTECHNIC UNIVERSITY OF THE PHILIPPINES
STA. MESA, MANILA
COLLEGE OF ENGINEERING
COMPUTER ENGINEERING DEPARTMENT

CMPE 40113 – Machine Leaning Track Elective 1

PREDICTIVE ANALYTICS MODELLING, SIMULATION AND OPTIMIZATION

INSTRUCTIONAL MATERIAL

DR. ARVIN R. DE LA CRUZ
ENGR. ROMAN ANGELO C. TRIA
ENGR. JAN REUELLE P. TEÑA
ENGR. JAN LENNARD A. AUGUSTO

TABLE OF CONTENTS

MODULE 1
Introduction to Predictive Analytics

There are few technologies that have the ability to revolutionize how business operates. Predictive analytics is one of those technologies. Predictive analytics consists primarily of the "Big 3" techniques: regression analysis, decision trees, and neural networks. Several other techniques, such as random forests and ensemble models have become increasingly popular in their use. Predictive analytics focuses on building and evaluating predictive models resulting in key output fit statistics that are used to solve business problems. This module defines essential analytics terminology, walks through the nine-step process for building predictive analytics models, introduces the "Big 3" techniques, and discusses careers within business analytics.

Learning Outcomes:

1. Define business analytics, big data, and predictive analytics.

2. Differentiate Descriptive analytics from predictive analytics.

3. Describe the nine-step predictive analytics model.

4. Explain why a critical thinking mindset is necessary when solving business problems.

Introduction

1. Predictive Analytics in Action

   In an increasingly fierce competitive market, companies need to identify techniques that garner competitive advantages to thrive. Imagine a company with the ability to predict consumer behaviors, prevent fraud, mitigate risk, identify new customers, and improve operations. What if in real time an organization can:

   - ✓ Identify customer's spending behaviors and cross-sell efficiently or sell additional products to their customers.
   - ✓ Enhance customer satisfaction and customer loyalty.
   - ✓ Identify the most effective marketing campaign and communication channels.
   - ✓ Identify fraudulent payment transactions.
   - ✓ Flag potential fraudulent claims and pay legitimate insurance claims immediately.
   - ✓ Predict when machinery will fail.

   With the advancement in computer hardware (faster processing speeds, in-line memory, cheaper storage, and massively parallel processing (MPP) architectures) coupled with new technologies such as Hadoop, MapReduce, Hive, Pig, Spark, and MongoDB and analytics for processing data, companies are now positioned to collect and analyze enormous amounts of structured and unstructured data gaining valuable

insights from the data in real time (run predictive algorithms on streaming data) or near real time.

Amazon, the number one online retailer, uses predictive analytics for targeted marketing. Their algorithms analyze seemingly endless customer transactions searching for hidden purchasing patterns, relationships among products, customers, and purchases. Their collected data includes purchase transactions, information that is contained in their customers' wish lists, and products the customers reviewed and searched for the most.

Netflix also has a recommendation system ("You might want to watch this…") that uses member's past viewing behavior to recommend new movies or shows. Netflix was able to predict that "House of Cards" would be a hit before the show went into production. Using predictive analytics, Netflix determined that a series directed by David Fincher, starring a certain male actor, and based on a popular British series was worth a $100 million investment (Carr 2013). Netflix gathers huge amounts of data from their 130 million memberships. Netflix collects what the members are streaming, the ratings given by their members, their searches, what day of the week, how many hours per month they watch shows, as well as what members do when the credits roll. With all of this data, Netflix can see trends and membership behavior and make predictions about what new streaming content to offer, personalized recommendations to make with the intent of increasing and retaining memberships.

Predictive analytics is a vital part of Walmart's strategy. Walmart uses analytics to predict stores' checkout demand at certain hours to determine how many associates are needed at the checkout counters. Through their predictive analytics, Walmart is able to improve the customer experience and establish the best forms of checkout, i.e., self-checkout or facilitated checkout by store. Similarly, Walmart uses predictive analytics in their pharmacies by analyzing how many prescriptions are filled in a day, the busiest times during a day and month, and then optimizes their staffing and scheduling to ultimately reduce the amount of time it takes to get a prescription filled. Additionally, Walmart uses predictive analytics to fast-track decisions on how to stock store shelves, display merchandise, new products, discontinue products, and which brands to stock (Walmart Staff 2017).

In 2011, the Internal Revenue Service (IRS) created a new division called the Office of Compliance Analytics. The purpose of the new division is to create an advanced analytics program that will use predictive analytics algorithms to reduce fraud. The IRS collects commercial and public data, including information from Facebook, Instagram, and Twitter. Algorithms are run against their collected information combined with the IRS' proprietary databases to identify potential noncompliant taxpayers (Houser and Sanders 2017).

Police departments are using predictive analytics (referred to as predictive policing) to optimize crime prevention strategies. Predictive policing is proactive and can help answer questions such as where is violent gun crime likely to occur? Where is a serial

burglar likely to commit his next crime? Where are criminals located? Data sources include past crime data, population density, the presence of expensive cars, the presence of tourists, payday schedules, weather and seasonal patterns, traffic patterns, moon phases, time of day, weekend or weekday, escape routes (e.g., bridges, tunnels, dense foliage), and social networks (e.g., family, friends, affiliations). Applying predictive analytics helps police departments create crime maps which forecast criminal hot spots, areas where crime is likely to occur. Then, the police departments can allocate resources and increased focus in these areas. Predictive analytics is also employed in tracking criminals after the crime has been committed (Bachner 2013).

2. Analytics Landscape

The analytics landscape includes terms like business analytics, big data analytics, data mining, and big data. Gartner, a leading information technology research and advisory company, defines business analytics as follows: “Business Analytics is comprised of solutions used to build analysis models and simulations to create scenarios, understand realities and predict future states. Business analytics includes data mining, predictive analytics, applied analytics and statistics (Gartner n.d.a).” Today, business analytics is essential as it can improve an organization’s competitive edge thus enhancing revenues, profitability, and shareholder return.

IBM defines big data analytics as “the use of advanced analytic techniques against very large, diverse data sets that include structured, semi-structured and unstructured data, from different sources, and in different sizes from terabytes to zettabytes (IBM n.d.).”

Data mining is applying advanced statistical techniques to discover significant data patterns, trends, and correlations in large data sets with the overarching goal of gathering insights previously unknown about the data sets and transforming the data for future use.

Figure 1.1 illustrates the relationships between big data, data mining, and analytics. These three terms overlap and collectively are considered business analytics. Basically, business analytics is the use of advanced analytic techniques to discover meaningful insights from large, complex data sets in an opportunistic timeframe.

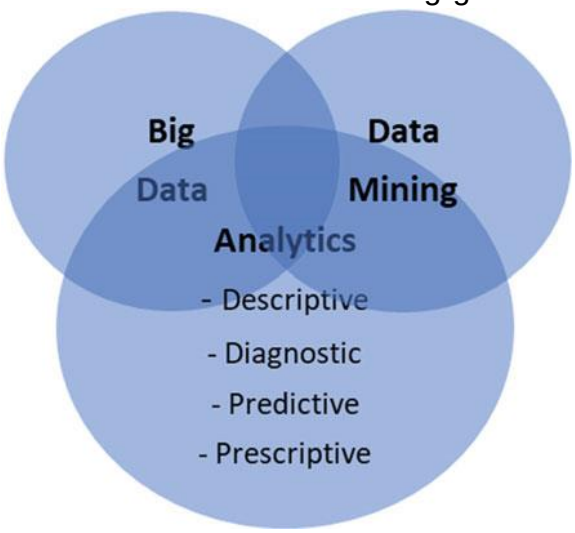


Figure 1.1 Business Analytics

Big Data

When individuals hear the term big data typically they think of large amounts of data and that it is just size that matters. Most individuals recognize megabytes, gigabytes, or even terabytes. Today, large companies are storing transactions in petabytes. In a SAS white paper, Big Data Meets Big Data Analytics, SAS states the following:

- ✓ "Wal-Mart handles more than a million customer transactions each hour and imports those into databases estimated to contain more than 2.5 petabytes of data.
- ✓ Radio frequency identification (RFID) systems used by retailers and others can generate 100–1000 times the data of conventional bar code systems.
- ✓ Facebook handles more than 250 million photograph uploads and the interactions of 800 million active users with more than 900 million objects (pages, groups, etc.) each day.
- ✓ More than 5 billion people are calling, texting, tweeting and browsing on mobile phones worldwide." (Troester n.d.)

Table 1.1 Large data sizes

Unit	Approximate size (decimal)	Examples
Bytes (B)	8 bits	One byte = one character; ten bytes = one word
Kilobyte (KB)	1,000 bytes	Two KBs = a typewritten page
Megabyte (MB)	1,000,000 bytes	One MB = a small novel; five MBs = complete work of Shakespeare
Gigabyte (GB)	1,000 megabytes	16 h of continuous music
Terabyte (TB)	1,000 gigabytes	130,000 digital photographs
Petabyte (PB)	1,000 terabytes	Two PBs = All US academic research libraries
Exabyte (EB)	1,000 petabytes	Five EBs = estimate of all words ever spoken by human beings
Zettabyte (ZB)	1,000 exabytes	36,000 years of high definition video
Yottabyte (YB)	1,000 zettabytes	

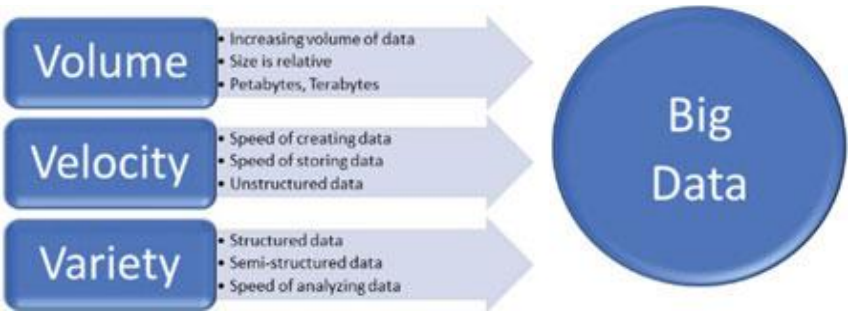


Figure 1.2 Big Data

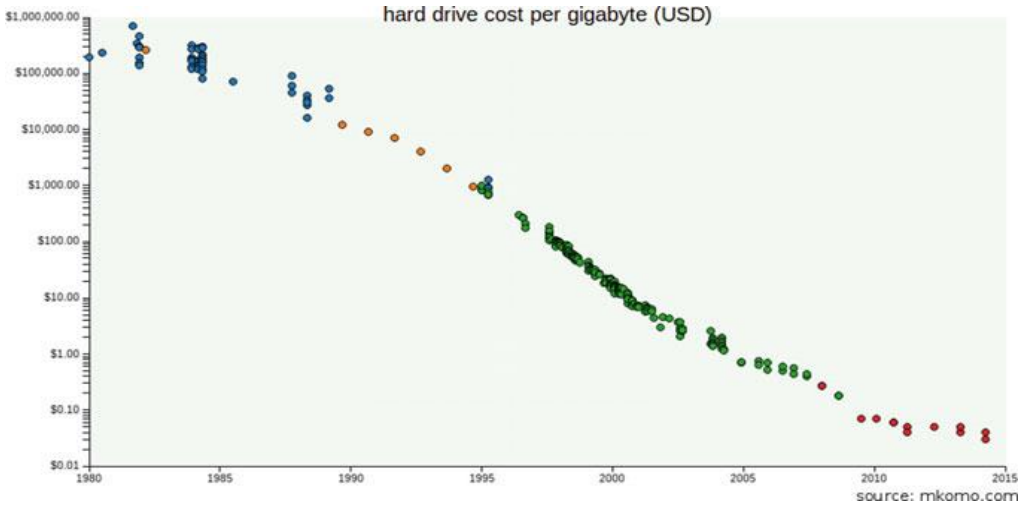


Figure 1.3 Hard drive cost per gigabyte (Komorowski 2014)

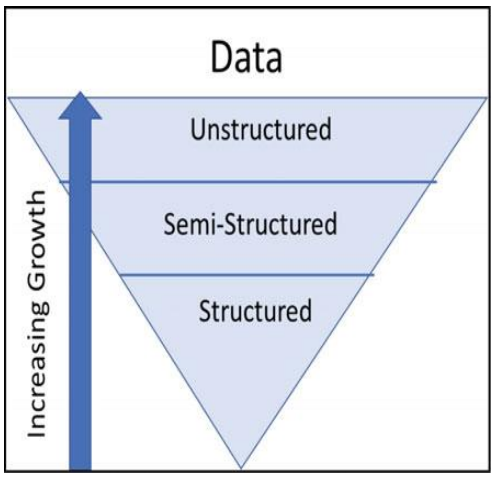


Figure 1.4 Type of Data

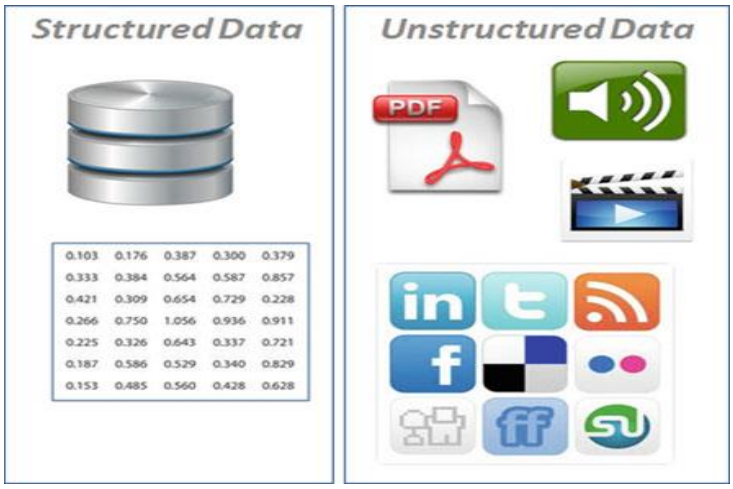


Figure 1.5 Examples of structured and unstructured data

3. Analytics

Analytics are applied to obtain useful information from the data. The fundamental goal of business analytics is to make better decisions based on the data. In this section, the two main types of analytics will be discussed—descriptive analytics and predictive analytics. Additionally, diagnostic and prescriptive analytics will be briefly mentioned. Descriptive analytics is usually the starting point of analysis performed on the data.

Descriptive analytics are a very basic form of analytics. Descriptive analytics summarize the raw data and describes the past. In other words, “What has happened?” Descriptive analytics is useful in providing information about past behaviors, patterns, or trends in the data. Descriptive analytics also aids in the preparation of data for future use in prescriptive analytics. Descriptive analytics include information like sums, averages, and percent changes. Business examples include the sum of sales by store, total profit by product or distribution channels, average inventory, number of complaints resolved in the past quarter, or classification by customer, or average amount spent per customer. Walmart uses descriptive analytics to uncover patterns in sales, determine what customers buy online versus in the store, and to see what is trending on Twitter.

Predictive analytics uses historical data to predict future events. The central question for predictive analytics is “What will happen?” Predictive analytics uses advanced statistics and other machine learning techniques. It is essential that historical data be representative of future trends for predictive analytics to be effective. Predictive analytics techniques can be used to predict a value—How long can this airplane engine run before requiring maintenance or to estimate a probability—How likely is it that a customer will default on a mortgage loan? Predictive analytics techniques can also be used to pick a category—What brand of sneakers will the customer buy? Nike, New Balance, or Adidas? Predictive analytics uses data-driven algorithms to generate models. Algorithms are step-by-step processes for solving problems. Algorithms take the data through a sequence of steps to calculate predictive results. Effectively, predictive analytics algorithms automate the process of discovering insights (e.g., patterns, trends) in the data.

Predictive analytics algorithms are split into supervised learning algorithms and unsupervised learning algorithms. Supervised learning algorithms involve an iterative process of learning from the training (historical) data set. The training data set contains labels and has the correct information (i.e., the information the model is learning to predict). The prediction output is referred to as the target variable.

As the predictive model is trained and reaches an optimal point, the model is now ready to produce predictive output.

Predictive analytics modeling techniques fall into two major categories: regression techniques and machine learning techniques.

4. Regression Analysis

Regression analysis is a widely used technique in predictive analytics. Regression analysis is a technique for measuring relationships between two or more variables and can be used to predict actual outcomes. The value that you want to predict the outcome is the dependent variable ("The effect of" and typically shown as Y in the regression equation). Also known as the target variable in predictive analytics. Independent variables and predictor variables are inputs into the regression equation that are assumed to have a direct effect on the target variable (dependent variable and typically shown as an X in the regression equation). Regression techniques provide a mathematical equation that describes the relationship between the target (predictor) variable and the other independent variables.

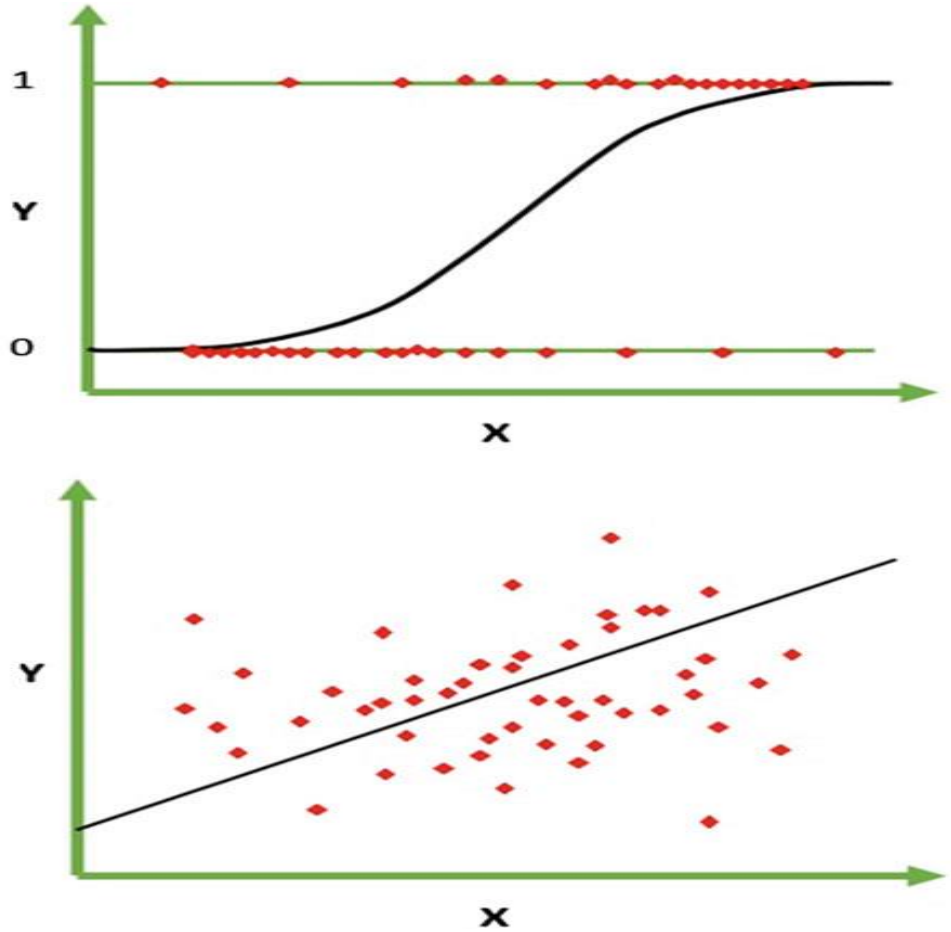


Figure 1.6 Linear regression graph and logistic regression graph

5. Machine Learning Techniques

Machine learning, a branch of artificial intelligence, uses computation algorithms to automatically learn insights from the data and make better decisions in the future with minimal intervention. Regression techniques may also be in machine learning techniques, e.g., neural networks. An artificial neural network (ANN) or neural network is a system, including algorithms and hardware, that strive to imitate the activity of neurons in the brain. A key element of the neural network is the ability to learn from the data. The neural network is composed of multiple nodes and is typically organized in layers. The layers are comprised of interconnected nodes.

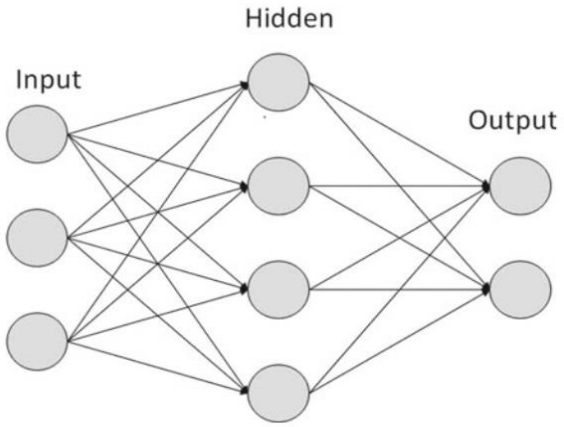


Figure 1.7 Neural Network

Decision trees are very popular in predictive modeling for several reasons—they are easy to understand, easy to build, can handle both nominal and continuous variables, can handle missing data automatically, and can handle a wide variety of data sets without applying transformations to the input. Decision trees are comprised of nodes and branches. The nodes contain questions (premises), and the response to the questions determine the path to the next node (conclusion). Branches link the nodes reflecting the path along the decision tree. The initial question on the decision tree is referred to as the root node. Figure 1.8 provides an example of a decision tree.

In addition to having knowledge of the statistical algorithms and machine learning tools, three other components are necessary to create value in predictive analytics (Fig. 1.9). The first component is strong business knowledge. Typically, a predictive analytics project will be team-based including members with in-depth knowledge of the organizations' industry and strategies as well as data science and analytics experts.

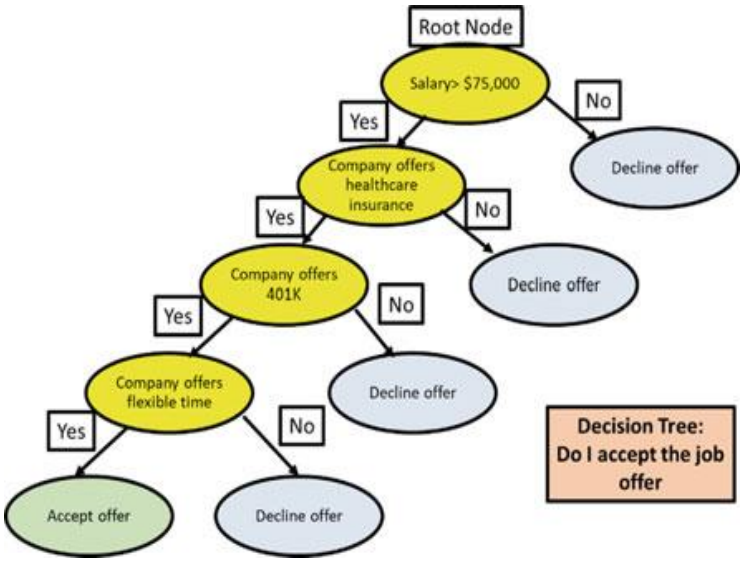


Figure 1.8 Decision Tree

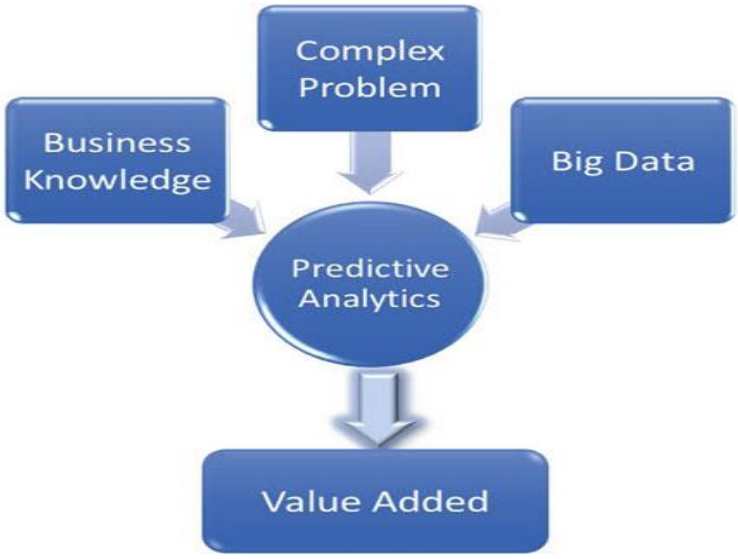


Figure 1.9 Decision Tree

Table 1.2 Common business problems

Detection of fraud	• Banking and financial service companies need to know how to detect and reduce fraudulent claims • Audit firms need to know how to identify fraudulent business transactions • Health and property-casualty insurance companies need to know which claims are fraudulent
Improve operational efficiency	• Manufacture and retail companies need to know how to create better inventory forecasts • Companies need to know how to improve resource management • Hotels need to know how to maximize occupancy rates and what room rate to charge per night • Airline companies need to optimize the timing of when aircraft engines require maintenance
Optimization of market campaigns/customer insights	• Retail companies need to determine customer responses, purchase behavior, cross-sell opportunities, and effectiveness of promotional campaigns • Customer loyalty programs target customers at the right time to maximize their purchases
Reduce risk	• Banking and financial service companies currently rely on credit scores to determine a customer's creditworthiness • Health insurance companies need to know which individuals are more at risk for chronic diseases
Enhance innovation	• Many companies need to constantly bring new products to the market • Pharmaceutical companies use for drug discovery

6. Predictive Analytics Model

Figure 1.10 provides a nine-step model for using predictive analytics. The first step in the model is to identify the business problem. Business-specific knowledge and problem identification are critical to developing a successful predictive analytics model. Additionally, knowledge of the data storage and infrastructure as well as proficiency in predictive modeling is required. Frequently, this will require a team of experts in each of these areas. The ability to solve these business problems rapidly may lead to higher revenues or lower expenses resulting in a competitive advantage. Step 2 in the predictive analytics model is to define the hypotheses. The hypotheses are developed from the business problem. The purpose of the hypotheses is to narrow down the business problem and make predictions about the relationships between two or more data variables. The hypotheses should clearly state what is going to be analyzed, i.e., predicted. Table 1.3 provides some examples of transitioning from a business problem to a hypothesis.

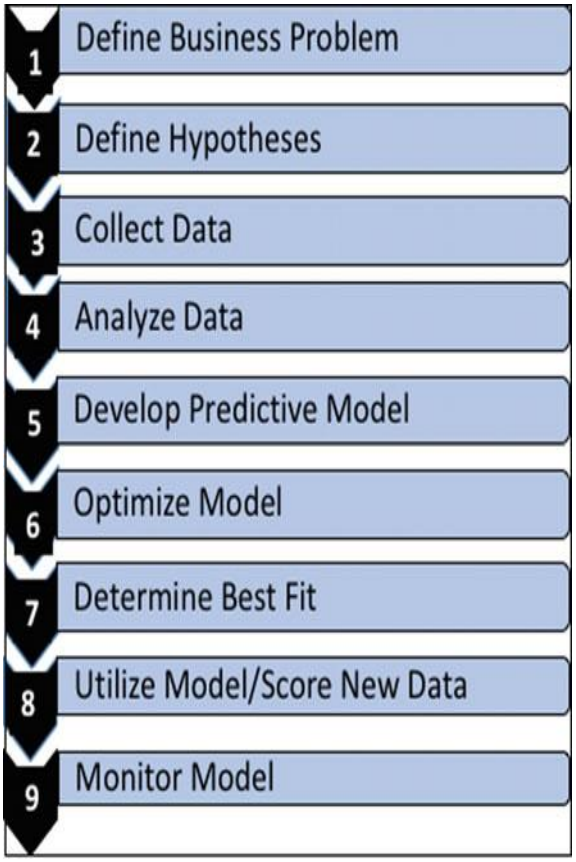


Figure 1.10 Predictive analytics model

Table 1.3 Business problem and hypothesis

Business problem	Hypothesis example
How to detect and reduce fraudulent claims	Gender, education, marital status, and income are indicators of individuals who would commit insurance fraud
Need to increase sales	Amount of time spent on company Web site determines purchase Facebook advertising generates more sales compared to LinkedIn advertising The word "call" in a marketing campaign results in increased leads
Filling a prescription takes too long; need to know how much staff to have available	The day social security checks are released are the busiest days The hours between 9 am and noon are the busiest
Need to know what individuals will purchase the next generation phone for marketing campaign to gain potential buyers	Age, prior product use, income, education, and location will indicate individuals who will purchase new phone

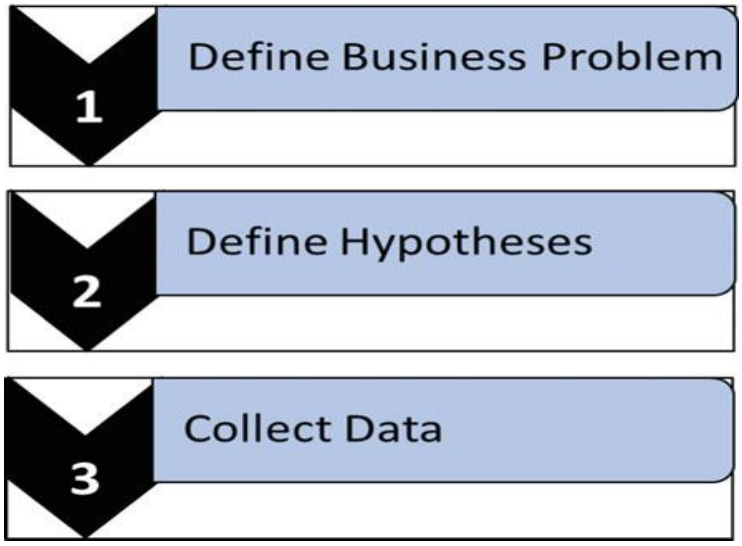


Figure 1.11 Decision Tree

Activity:

1. Define the requirements for data processing
2. Identify possible IDE to be used in data analytics
3. Create a tutorial video of data analytics using a certain IDE

Modules Test

1. Describe big analytics, big data, and predictive analytics. What do they have in common?
2. Describe an example of an organization that uses predictive analytics and how it has added value to the company.
3. Discuss the nine steps of the predictive analytics model.
4. Why is a critical thinking mindset important in a career in analytics?
5. Discuss at least three different possible career opportunities enabled by predictive analytics.

MODULE 2
Data Classification and Preparation

The module will begin with a description of the different categories of data followed by a review of the methods used for preparing the data.

Learning Outcomes:

1. Differentiate between different data types.
2. Identify and explain the four basic data roles.
3. Identify the methods used for handling missing values, outliers, and redundant data.
4. Prepare data for predictive modeling.

Introduction

Once the business problem is identified, the hypotheses are developed, and the data is collected, the next step in the process is to analyze the data and prepare the data for predictive modeling. Most raw data is considered “dirty” or “noisy” because the data may have incomplete information, redundant information, outliers, or errors.

Discussion

2.1 Classification of Data

Once the data is collected, the data scientist or analyst needs to first understand the data. For example, how many records were collected, how many variables are contained in the data? A variable or data item is an identifiable piece of data that can be measured or counted. The variable’s values can appear as a number or text, which can be converted into number. Variables can have different properties or characteristics.

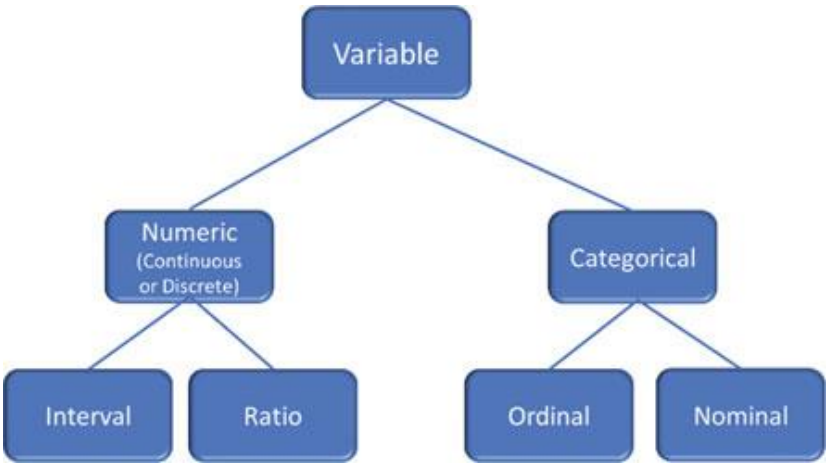


Figure 2.1 Variable Classifications

2.1.1 *Qualitative Versus Quantitative*

Before analyzing the data, it is important to distinguish between the two types of variables: qualitative and quantitative (or numeric). Qualitative or categorical variables are descriptive in nature, for example, the color of a sweater, the name of a city, or one's occupation. In SAS Enterprise Miner™, categorical variables are referred to as class variables (Fig. 2.1). Quantitative variables are numeric. They represent a measurable quantity. Quantitative data can also be further categorized as discrete or continuous. Variables are said to be discrete if the measurements are countable in a finite amount of time. They can only take on a limited number of values. For example, shoe size, the number of heads in a coin toss, and SAT scores are discrete variables.

Continuous variables can take on any value, usually within some range. When trying to determine if a variable is continuous versus discrete, check to see if the value can be divided and reduced to more exact measurements. For example, height, weight, and age are continuous variables. Age is considered continuous; one can always break it down into smaller and smaller increments: 5 years, 2 months, 3 days, 6 h, 4 s, 4 ms, 8 ns, etc. A person's age in years would be discrete.

2.1.2 *Scales of Measurement*

Scales of measurement represent the methods in which types of data are defined and categorized. The four scales of measurement are nominal, ordinal, interval, and ratio. The nominal and ordinal measurement scale is used to further describe class variables

Nominal scales are used to identify or name qualitative values without any order. Examples of nominal scales include gender (male/female), hair color (blonde/brunette), or a student's major (marketing/finance/accounting). Nominal scales may be represented by numbers (1-male, 0-female). If so, these numbers are used to identify the object and mathematical operations cannot be performed on the values. These values also do not represent a quantity or a measurement. For example, a zip code is a variable that would be a nominal scale measurement. Numbers represent its values. No meaningful data can be derived from adding, multiplying, or averaging zip codes.

Ordinal scales are where there is some natural relationship or rank order, such as a student's letter grade (A, B, C, D, F), or a shirt size (small, medium, large). Survey responses may range from strongly disagree to strongly agree which is another example of ordinal data. For values measured on an ordinal scale, the difference between the values may not be equal. For example, the difference

between a small beverage and a medium beverage may not be the same as the difference between a medium and a large beverage. So, it is the order of the values that is important, but the difference between each one is not known.

Interval scales characterize quantity, and the difference between the levels is equal. Examples of values that are measured with interval scales include the temperature in Fahrenheit or Celsius, a calendar year, an IQ score, or a student's SAT score. Let's look at temperature in Fahrenheit. The difference between 60- and 70-°F is the same as the difference between 50- and 60-°F. Interval values do not have a "true zero". For example, there is no such thing as having no temperature. A person cannot have a zero IQ or SAT score. Values measured on interval scales are ordered, and their differences may be meaningful; however, evaluating a ratio or doubling its value is not meaningful. A 100 °F is not twice as warm as 50-°F.

Ratio scales have similar characteristics to the interval scales but also have a true (absolute) zero; that is, no numbers exist below zero. The values have a rank order and the intervals between values are equal. Doubling the values or calculating its ratios is meaningful. For example, doubling an object's weight is meaningful. It is twice as heavy. Height, weight, age, and temperature in Kelvin are all examples of ratio scale variables.

2.2 Data Preparation Methods

Data in its raw, original form is typically not ready to be analyzed and modeled. Data sets are often merged and contain inconsistent formats, missing data, miscoded data, incorrect data, and duplicate data. The data needs to be analyzed, "cleansed," transformed, and validated before model creation. This step can take a significant amount of time in the process but is vital to the process.

2.2.1 *Inconsistent Formats*

Data in a single column must have consistent formats. When data sets are merged together, this can result in the same data with different formats. For example, dates can be problematic. A data column cannot have a date format as mm/dd/yyyy and mm/dd/yy. The data must be corrected to have consistent formats.

2.2.2 *Missing Data*

Missing data is a data value that is not stored for a variable in the observation of interest. There are many reasons that the value may be missing. The data may not have been available, or the value may have just been accidently omitted. When analyzing data, first determine the pattern of the missing data. There are three pattern types: missing

completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Missing completely at random occurs when there is no pattern in the missing data for any variable. Missing at random occurs when there is a pattern in the missing data but not on the primary dependent variables.

2.2.3 *Outliers*

An outlier is a data value that is an abnormal distance from the other data values in the data set. Outliers can be visually identified by constructing histograms, stemand-leaf plots or box plots and looking for values that are too high or too low. There are five common methods to manage the outliers:

1. Remove the outliers from the modeling data.
2. Separate the outliers and create separate models.
3. Transform the outliers so that they are no longer outliers
4. Bin the data.
5. Leave the outliers in the data.

2.2.4 *Other Data Cleansing Considerations*

Miscoded and incorrect values should be addressed. Identification of these errors can be done in a couple of ways depending on whether the variable is categorical or numeric. If categorical, frequency counts can aid in the identification of variables that are unusual and occur infrequently. In this case, categories can be combined. For numeric variables, they may be outliers. Duplicate values should be removed.

2.3 Data Sets and Data Partitioning

In predictive analytics to assess how well your model behaves when applied to new data, the original data set is divided into multiple partitions: training, validation, and optionally test. Partitioning is normally performed randomly to protect against any bias, but stratified sampling can be performed. The training partition is used to train or build the model. For example, in regression analysis, the training set is used to fit the linear regression model. In neural networks, the training set is used to obtain the model's network weights. After fitting the model on the training partition, the performance of the model is tested on the validation partition. The best-fitting model is most often chosen based on its accuracy with the validation data set. After selecting the best-fit model, it is a good idea to check the performance of the model against the test partition which was not used in either training or in validation. This is often the case when dealing with big data. It is important to find the correct level of model complexity. A model that is that is not complex enough, referred to as under fit, may lack the flexibility to accurately represent the data. This can be caused by a training set that is too small. When the model is too complex or overfit, it

can be influenced by random noise. This can be caused by a training set that is too large. Often analysts will partition the data set early in the data preparation process.

2.4 Model Components

To create a predictive model requires the creation of four components:

1. **Project**—This contains the diagrams, data sources, and the library for the data. Generally, a separate project is created for each problem that is trying to be solved.
2. **Diagram**—This is the worksheet where you build your model and determines the processing order and controls the sequence of events for preparing the predictive model.
3. **Library**—This is a pointer to the location of the stored data files.
4. **Data Source**—This contains both the data and the metadata that defines and configures an input data set.

The data preparation steps are:

1. Create the first three model components: Project, Diagram, and Library
2. Import the data
3. Create the Data Source
4. Partition the data
5. Explore the data
6. Address missing values
7. Address outliers
8. Address categorical variables with too many levels
9. Address skewed distributions
10. Transform variables

Activity:

1. Create a data model based on the define IDE, process, and simulate.

Modules Test

1. Describe the different variable types and the measurement scales. Give examples of each.
2. Discuss the different methods for handling missing values and outliers.
3. What are some of the data problems you should adjust for before modeling your data?
4. Why is data partitioned?
5. What are some methods for filtering data? Provide examples of each.

MODULE 3
Descriptive Statistics

This module focuses on a review of descriptive statistical analysis that is used to prepare and support predictive analytics. It reviews methods to ensure that the data is prepared for analysis as well as methods for combing or reducing variables to improve the results of predictive analytics.

Learning Outcomes:

1. Perform hypothesis testing, and identify the two types of errors that can occur.
2. Describe the importance of central tendency, variation, and shape for numerical variables and how it can affect your predictive model.
3. Compare and contrast data distributions and how to correct for skewness and kurtosis.
4. Calculate the covariance and the coefficient of correlation.
5. Evaluate ANOVA results, and describe when it is appropriate to use.
6. Evaluate Chi-square results, and describe when it is appropriate to use.
7. Identify the various methods for evaluating the fit of a predictive model.
8. Identify and describe the methods for variable reduction.
9. Describe stochastic models.

Introduction

Prior to analyzing any data set, it is important to first understand the data. Descriptive statistics presents data in a meaningful way for the purpose of understanding what if anything will need to be done to the data to prepare it for analysis. There are many statistical tests that can be utilized.

Discussion

3.1 Descriptive Analytics

Descriptive analytics is the first stage of data analysis. It provides a summary of historical data that may indicate the need for additional data preprocessing to better prepare the data for predictive modeling. For example, if a variable that is highly skewed, the variable may need to be normalized to produce a more accurate model.

Descriptive statistics are broken down into measures of central tendency and measures of variability and shape. Measures of central tendency include the mean, median, and mode. Measures of variability include the standard deviation, variance, range and the kurtosis and skewness.

3.2 The Role of the Mean, Median, and Mode

The central tendency is the extent to which the values of a numerical variable group around a typical or central value. The three measures we will look at are the arithmetic mean, the median and the mode. The mean is the most common measure of central tendency. It is the sum of all the values divided by the number of values. Although it is the preferred measure, extreme values or outliers affect it.

The median is another measure of central tendency. It is the middle value in an ordered set of observations. It is less sensitive to outliers or extreme values. If the number of observations is odd, the median is simply the middle number. If the number of observations is even, the median is the average of those two values on either side of the center. The median is not influenced by outliers. The third measure of central tendency is the mode. The mode is the value that occurs most often. Figure 3.1 shows an example of the calculation of the mean, median, and mode. Which measure should you choose? Unless extreme values or outliers exist, the mean is generally the preferred measure. The median is used when the data is skewed; there are a small number of observations or working with ordinal data. The mode is rarely used. The only situation in which the mode would be preferred is when describing categorical or class variables.

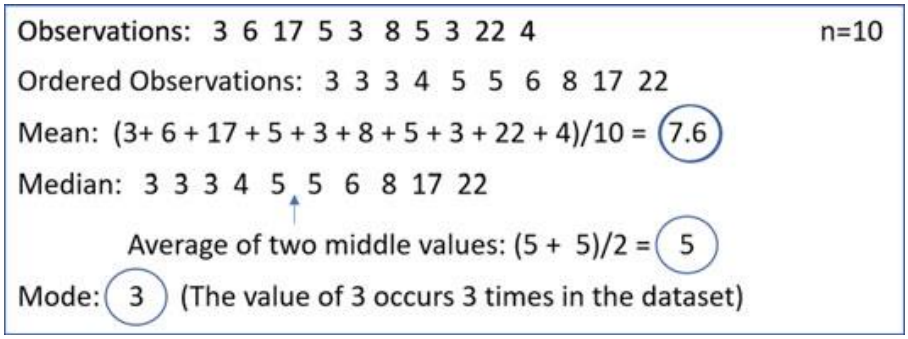


Figure 3.1 Calculating measures of central tendency example

3.3 Variance and Distribution

Measures of variation provide information on the spread or variability of the data set. Some of the measures of variation include the range, the sample variance, the sample standard deviation, and the coefficient of variation. The range is the easiest measure to calculate. It is the difference between the highest and the lowest value. The range does not provide any information about the distribution of the data. The range is also sensitive to outliers. A common measure of variation is the sample variance. The sample variance is the average of the squared deviations of each observation from the mean. Figure 3.2 shows the formulas for the sample standard deviation, usually denoted S2

The standard deviation is the square root of the variance and is in the same units of measurement as the original data. Figure 3.3 shows an example calculation. The more the data is spread out, the greater the range, variance, and standard deviation. The more the data is concentrated, the smaller the range, variance, and standard deviation.

$$S^2 = \frac{\Sigma(X-\overline{X})}{n-1}$$

Where X = the individual observation

$\overline{X}$ = the mean of the observations

n = the number of observations

Figure 3.2 Variance formula

Observations: 3 6 17 5 3 8 5 3 22 4 n=10

Range: 22-3 = 19

Mean = 7.6

Variance: $((3-7.6)^2 + (6-7.6)^2 + (17-7.6)^2 + (5-7.6)^2 + (3-7.6)^2 + (8-7.6)^2 + (5-7.6)^2 + (3-7.6)^2 + (22-7.6)^2 + (4-7.6)^2)/(10-1) = 43.16$

Standard Deviation: $\sqrt{43.16} = 6.57$

Figure 3.3 Calculating the variance and distribution

3.4 The Shape of the Distribution

Skewness measures the extent to which input variables are not symmetrical. It measures the relative size of the two tails. A distribution can be left skewed, symmetric, or right skewed. In a left-skewed distribution, the mean is less than the median, and the skewness value is negative. Notice in Fig. 3.4 that the peak is on the right for a left-skewed distribution. For a normal, symmetrical distribution, the mean and the median are equal and the skewness value is zero. For a right-skewed distribution, the mean is greater than the median, the peak is on the left, and the skewness value is positive. The formula to calculate the skewness is given in Fig. 3.5.

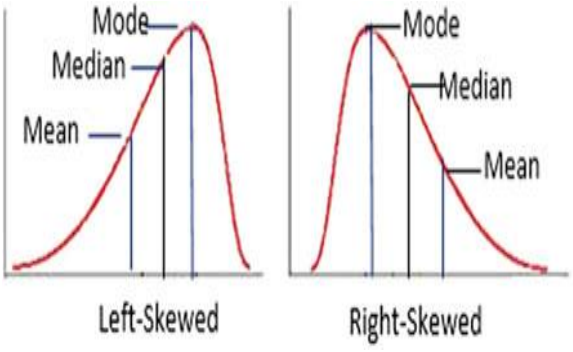


Figure 3.4 Skewness

$$\frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s}\right)$$

Where n is the sample size,
x_i is the i^{th} value of the variable,
$\bar{x}$ is the sample average, and
s is the sample standard deviation

Figure 3.5 Skewness formula

Kurtosis measures how peaked the curve of the distribution is. In other words, how sharply the curve rises approaching the center of the distribution. It measures the amount of probability in the tails.

When performing predictive modeling, the distribution of the variables should be evaluated. If it is highly skewed, a small percentage of the data points (i.e., those lying in the tails of the distribution) may have a great deal of influence on the predictive model. In addition, the number of variables available to predict the target variable can vary greatly. To counteract the impact of skewness or kurtosis, the variable can be transformed. There are three strategies to overcome these problems.

1. Use a transformation function, to stabilize the variance.
2. Use a binning transformation, which divides the variable values into groups to appropriately weight each range.
3. Use both a transformation function and a binning transformation. This transformation
can result in a better fitting model.

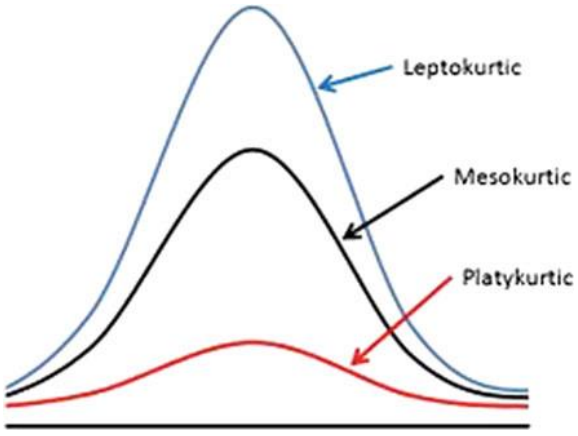


Figure 3.6 Kurtosis

$$\frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}\left(\frac{(x_i-\bar{X}}{S}\right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}$$

Where n is the sample size,
x_i is the i^{th} value of the variable,
$\bar{x}$ is the sample average, and
s is the sample standard deviation

Figure 3.7 Kurtosis Formula

3.5 Covariance and Correlation

It is important to investigate the relationship of the input variables to one another and to the target variable. The input variables should be independent of one another. If the input variables are too related (i.e., correlated) to one another, multicollinearity can occur which affects the accuracy of the model. Both the covariance and correlation describe how two variables are related. The covariance indicates whether two variables are positively or inversely related. A measure used to indicate the extent to which two random variables change in tandem is known as covariance.

The correlation does measure the relative strength of the relationship. It standardizes the measures so that two variables can be compared. If there is a strong correlation between the input variables, there can be some multicollinearity, which may negatively affect the predictive model. This undermines the assumption of independent input variables.

$$COV(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Where n is the sample size,
x_i is the i^{th} value of the variable x
$\bar{x}$ is the average of x,
v_i is the i^{th} value of the variable y,
$\bar{y}$ is the average of y

Figure 3.8 CovarianceFormula

$$r_{x,y} = \frac{cov(x,y)}{S_x S_y}$$

Where $S_x S_y$ are the sample standard deviation of x and y.

Figure 3.9 Correlation Formula

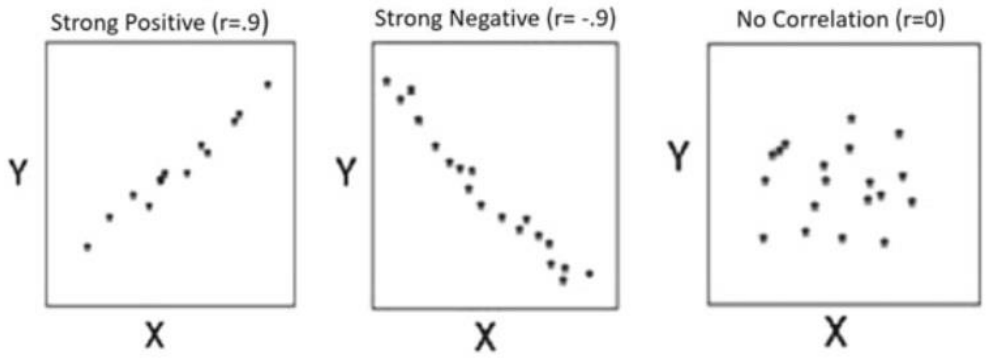


Figure 3.10 Correlation

3.6 Variable Reduction

Reducing the number of variables can reduce multicollinearity, redundancy, and irrelevancy and improve the processing time of the model. Two methods for variable reduction include variable clustering and principal component analysis.

Variable clustering identifies the correlations and covariances between the input variables and creates groups or clusters of similar variables. Clustering attempts to reduce the correlation within the groups. A few representative variables that are fairly independent of one another can then be selected from each cluster. The representative variables are used as input variables and the other input variables are rejected.

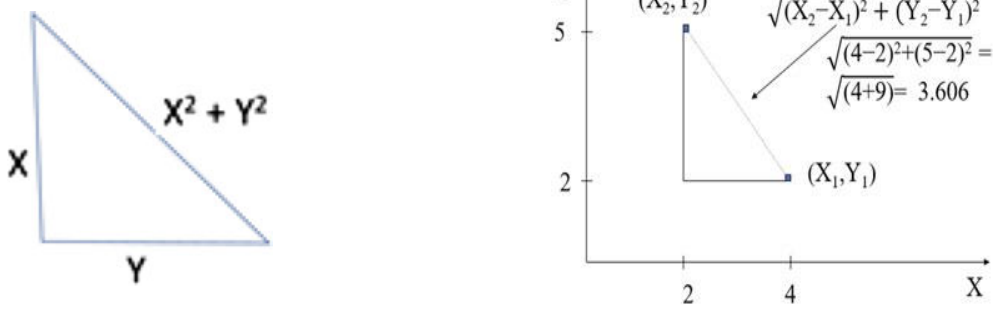


Figure 3.11 Illustration of Pythagorean theorem

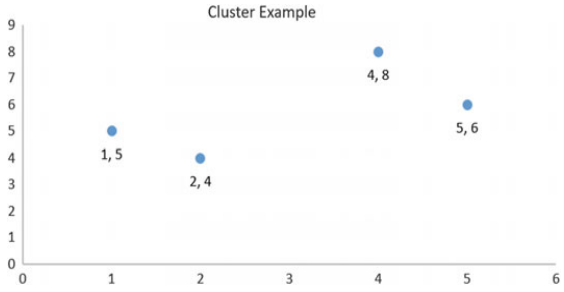


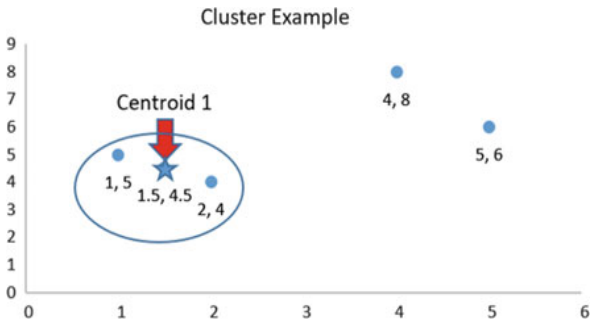
Figure 3.12 Euclidean distance in two-dimensional space

Figure 3.13 Scatter plot

Table 3.1 Iteration 1

Start (x, y)	End (x, y)	Euclidean distance
1, 5	2, 4	$1.41 = \sqrt{(2-1)^2 + (4-5)^2}$
1, 5	4, 8	4.24
1, 5	5, 6	4.12
2, 4	4, 8	4.47
2, 4	5, 6	3.61
4, 8	5, 6	2.24

Figure 3.14 Scatter plot- one cluster



Start (x, y)	End (x, y)	Euclidean distance
(1.5, 4.5)	4, 8	4.30
(1.5, 4.5)	5, 6	3.81
4, 8	5, 6	2.24

Table 3.1 Iteration 2

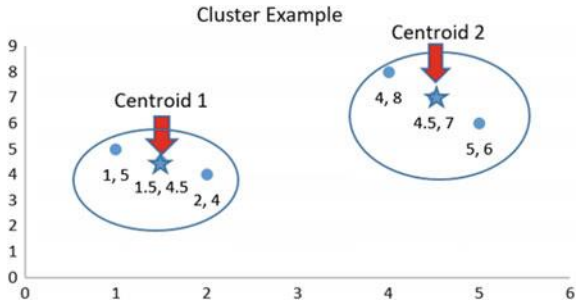


Figure 3.15 Scatter plot with two cluster

Principal component analysis (PCA) is another variable reduction strategy. It is used when there are several redundant variables or variables that are correlated with one another and may be measuring the same construct. Principal component analysis mathematically manipulates the input variables and develops a smaller number of artificial variables (called principal components). These components are then used in the subsequent nodes. The first principal component is created so that it captures as much of the variation in the input variables as possible.

The second principal component accounts for as much of the remaining variation as possible and so forth with each component maximizing the remaining variation. The number of principal components can be based upon the proportion of variance explained, the scree plot (Eigenvalue plot) or an Eigenvalue greater than 1. One major drawback of using PCA is that it is very difficult to interpret the principal components, and it is hard to determine how many components are needed (Brown 2017). A scree plot is a line segment graph that displays a decreasing function that shows the total variance explained by each principal component.

3.7 Hypothesis Testing

A hypothesis is a supposition or observation regarding the results of sample data. It is the second step in the scientific method (Fig. 3.45). In hypothesis testing, information is known about the sample. The purpose of hypothesis testing is to see if the hypothesis can be extended to describe the population

Figure 3.16 Scientific method

In hypothesis testing, there are two types of errors that can occur. A type I and type II error. A type I occurs when you reject a true null hypothesis. This is like finding an innocent person guilty. It is a false alarm. The probability of a type I error is referred to as alpha, α, and it is called the level of significance of the test. A type II error called beta is the failure to reject a false null hypothesis. It is equivalent to letting a guilty person go. A type II error represents a missed opportunity. The power of the test is the probability of rejecting a false null hypothesis. It is equivalent to finding a guilty person guilty. The power of the test is equal to 1 minus beta. The steps to hypothesis testing are as follows:

1. Specify the null hypothesis.
2. Specify the alternative hypothesis.
3. Select the appropriate test statistic and the significance level (alpha, α).
4. Calculate the test statistic and corresponding p-value.
5. Draw a conclusion.

$$t_{\text{statistic}} = \frac{X-\mu}{S/\sqrt{n}} = \frac{16-15}{3/\sqrt{25}} = 1.667, \text{ the critical } t_{\alpha/2} = .025$$

Figure 3.17 *t*-statistic formula

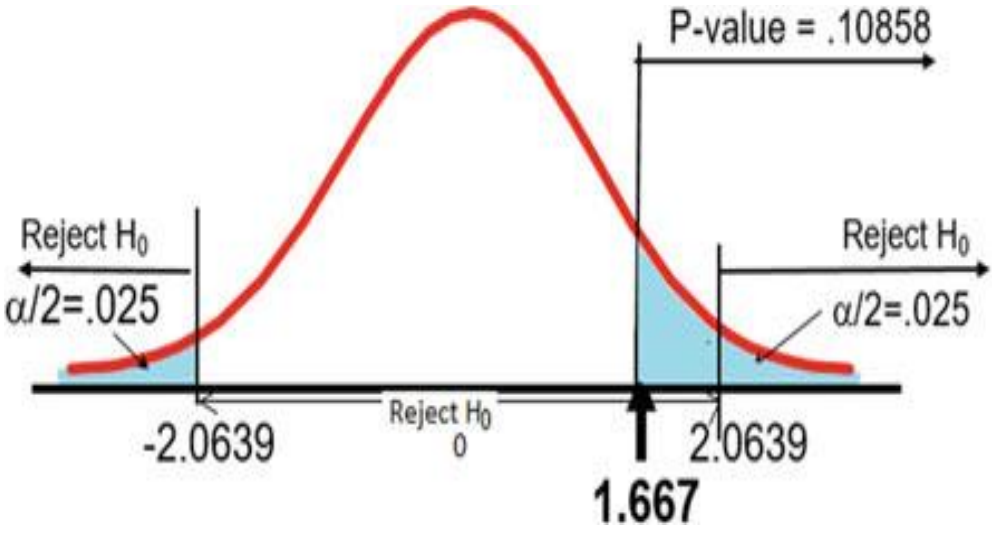


Figure 3.18 Hypothesis testing illustration

3.8 Analysis of Variance (ANOVA)

ANOVA is used to test the difference among the means of input variables. It can be broken into two categories: one-way ANOVA and two-way ANOVA. In one-way ANOVA, the difference among the means of three or more inputs is evaluated to see if any of them are equal. Two-way ANOVA examines the effect of two inputs on the target variable. In other words, it examines the interaction between two input variables on the target variable

The null hypothesis for one-way ANOVA is that the means for the different input variables are equal. The alternative hypothesis is not all input variable means are equal. There is a factor effect. At least one input mean is different. The test statistic is an F-statistic equal to the mean squared among the groups divided by the mean squared within groups.

The null hypothesis for two-way ANOVA is the interaction effect of two input variables being equal to zero. The alternative hypothesis is the interaction effect of two input variables is not equal to zero. This means there is an interaction effect. The test statistic is an F-statistic equal to the mean square interaction divided by the mean squared error.

3.9 Chi Square

The chi-square test is used to determine if two categorical (class) variables are independent. The null hypothesis is that the two categorical variables are independent. The alternative is that they are not independent. The chi-square test statistic

approximately follows a chi-square distribution with degrees of freedom, where r is the number of levels for one categorical variable, and c is the number of levels for the other categorical variable.

For the Formula of ANOVA and Chi Square, please search and see online resources.

Activity:

1. From the Create data in module 2, please apply the following descriptive statistics
2. Create a new research problem and applies the statistical treatment.

Modules Test

1. Discuss the three measures of tendency.
2. Discuss the measures of variance.
3. Describe skewness and kurtosis? What actions should be taken if the data is skewed?
4. Describe methods for variable reduction.
5. What statistics can be used to evaluate the fit of the model?
6. Describe when is it appropriate to use ANOVA and chi-square analysis?

MODULE 4
Predictive Model Using Regression

This module defines these techniques and when it is appropriate to use the various regression models. Regression assumptions for each type are discussed. Evaluation metrics to determine model fit including R2, adjusted R2, and p-values are examined. Variable selection techniques (forward, backward, and stepwise) and examination of model coefficients are also discussed.

Learning Outcomes:

1. Compare and contrast the different types of regression methods.
2. Explain the classical assumptions required in linear regression.
3. Explain how to validate these assumptions and discuss what to do if the assumption is violated.
4. Identify and discuss the components of the multiple linear regression equation.
5. Distinguish the various metrics used to determine the strength of the regression line.
6. Compare and contrast three common variable selection methods.
7. Create regression models
8. Evaluate the regression output.

Introduction

After performing descriptive analysis and data preparation, the next step is to build the predictive model. Regression models can be used as a predictive model. Popular regression models include linear regression, logistic regression, principal component regression, and partial least squares.

Discussion

4.1 Regression analysis techniques are one of the most common, fundamental statistical techniques used in predictive analytics. The goal of regression analysis is to select a random sample from a population and use the random sample to predict other properties of the population or future events. Regression analysis examines the degree of relationship that exists between a set of input (independent, predictor) variables and a target (dependent) variable. Regression analysis aids organizations in understanding how the target variable (what is being predicted) changes when any one of the input variables changes, while the other input variables are held constant. For example, an organization may want to predict their sales revenue in a new market in the southeast (target variable) next year. The organization hypothesizes that population, income level, and weather patterns (input/independent/predictor variables) may influence sales revenue in the new southeast market.

Regression analysis is also known as line-fitting or curve-fitting since the regression equation can be used in fitting a line to data points or in fitting a curve to data points. The regression technique is approached so that the differences in the distance between data points from the line or curve are minimized, referred to as the line of best fit. It is important to note that the relationships between the target variable and the input variables are associative only and any cause-effect is merely subjective.

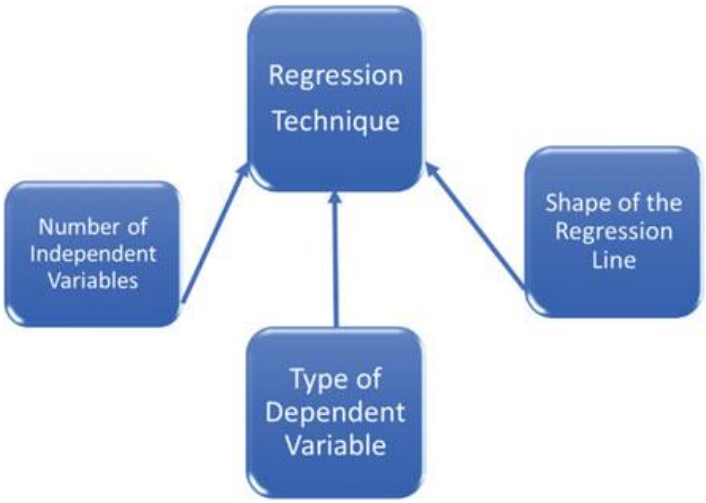


Figure 4.1 Determination of regression techniques

Classical Assumptions

Regression analysis is parametric; that is, it comes from a population that follows a probability distribution based on a fixed set of parameters and therefore makes certain assumptions. If these assumptions are violated, then the regression model results will not be suitable for adequate prediction. Therefore, these assumptions must be validated to produce an effective model. Below are five common assumptions that must be validated for a model to generate good results:

1. Linearity—There should be a linear and additive relationship between the target (dependent) variable and the input (predictor) variables.
2. Independence—Multicollinearity does not exist. Multicollinearity is when two or more input variables are moderately or highly correlated. In linear regression, the input variables should not be correlated.
3. Constant variance or Homoskedasticity—An error term (another variable) exists when the predictive model does not precisely portray the relationship between the input and the target variables. The error term must have the same variance across all values of the input variables. If the variance is not constant it is referred to as heteroskedasticity.
4. Autocorrelation—There should be no correlation between the residual (error) terms. This can occur where the one data point in a variable is dependent on another data point within

the same variable. This is frequently seen in time series models. When the error terms are correlated, the estimated standard errors tend to underestimate the standard error.

5. Normality—The error terms must be normally distributed for any given value of the input variables with a mean of zero.

4.2 Ordinary Least Squares

Linear regression aims to fit a straight line (or hyperplane for multiple regression) through a data set. Ordinary least squares or least squares is a linear regression method that seeks to find out the slope and intercept of a straight line between input variables. It is called least squares as it aims to find out the slope and intercept in such a way as to minimize the sum of the squares of the differences between actual and estimated values of the input variables.

4.3 Simple Linear Regression

Recall the purpose of linear regression is to model the relationship between variables by fitting a linear equation to the collected data. Simple linear regression and multiple linear regression are two basic and well-known predictive modeling techniques. Simple linear regression estimates the relationships between two numeric (continuous/interval) variables—the target variable (Y) and the input variable (X). For example, an organization may want to relate the number of clicks on their Web site to purchasing a product. Can the regression equation predict how many clicks on average it takes before the customer purchases the product?

Determining Relationship Between Two Variables

It is imperative to ensure that a relationship between the variables is determined prior to fitting a linear model to the collected data. However, as mentioned earlier, it is important to note that just because a relationship exists, does not mean that one variable causes the other; rather, some significant association between the two variables exists. A common method to determine the strength of the relationship between two variables is a scatter plot.

Line of Best Fit and Simple Linear Regression Equation

The goal of simple linear regression is to develop an equation that provides the line of best fit. The line of best fit is a straight-line equation that minimizes the distance between the predicted value and the actual value. Figure 4.2 shows a scatter plot with a line of best fit also known as the regression line or trend line. Software programs

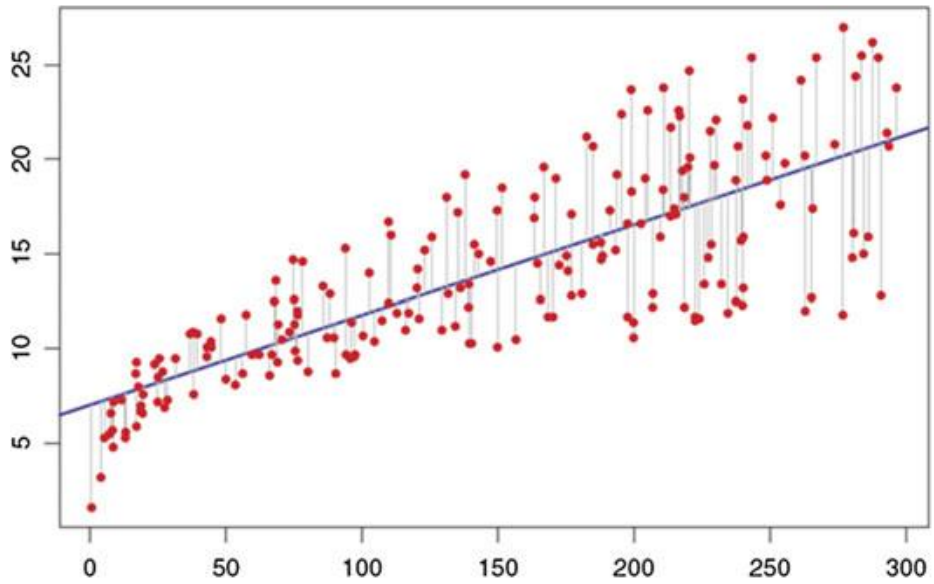


Figure 4.2 Scatter plot with line of best fit

4.4 Multiple Linear Regression

Multiple linear regression analysis is used to predict trends and future values. Like simple linear regression, multiple linear regression estimates the relationship between variables. The key difference is multiple linear regression has two or more input variables. Therefore, multiple linear regression estimates the relationship between a numeric (continuous/interval) target variable and two or more input variables.

Selection of Variables in Regression

Forward selection is a simple method for variable selection. In this approach, variables are added to the model one at a time based on a preset significance level (pvalue or “alpha to enter”). The process begins by adding the most significant variable to the model. The process then continues to add variables that are significant until none of the remaining variables are significant. The preset p-value is typically greater than the standard 0.05 level. Many software packages are less restrictive and have default values set at 0.10 or 0.15. The forward selection method may be useful when multicollinearity may be a problem. However, a limitation of the forward selection process is that each addition of a new variable may cause one or more of the included variables to be nonsignificant.

Backward selection is the opposite of forward selection and may help overcome the limitation noted above. The backward selection process starts with fitting the model with all identified input variables. In this approach, variables are removed from the model one at a time based on a preset significance level (p-value or “alpha to remove”). The process begins by removing the least significant variable to the model. The process

then continues to remove variables that are not significant until all remaining variables in the model are significant. A limitation of this method is the model selected may include unnecessary variables.

Stepwise selection incorporates both forward and backward selection processes. The basic idea behind the stepwise selection method is that the model is developed from input variables that are added or removed in a sequential manner into the model until there is no arguable reason to add or remove any more variables. The stepwise selection method starts with no input variables in the model. After each step where a variable is added, a step is performed to see if any of the variables added to the model have had their significance reduced below the preset significance level. Two significant levels are determined. A significance level ("alpha to enter") is set for identifying when to add a variable into the model, and a significance level ("alpha to remove") is set for identifying when to remove a variable from the model.

4.5 Principal Component Regression

PCA applies a method referred to as feature extraction. Feature extraction creates "new" input variables. The "new" input variables are a combination of each of the "old" input variables. The new input variables are created through a process known as orthogonal transformation (a linear transformation process). The new input variables (referred to as principal components) are now linearly uncorrelated variables. Orthogonal transformation orders the new input variables such that the first principal component explains the largest amount of variability in the data set as possible and each subsequent principal component in sequence has the highest variance possible under the limitation that it is orthogonal (perpendicular) to the preceding components. Since the new input variables are ordered from highest variance to smallest variance (least important), a decision can be made which variables to exclude from the prediction model. The results of PCA are normally defined as component scores or factor scores.

Principal component regression is an extension of PCA. Principal component regression takes the untransformed target variable and regresses it on the subset of the transformed input variables (principal components) that were not removed. This is possible since the newly transformed input variables are independent (uncorrelated) variables.

Multiple linear regression is a good regression technique to use when data sets contain easy-to-deal with input variables to explain the predictor variables, the input variables are not significantly redundant (collinear), the input variables have an implied relationship to the predictor variable, and there are not too many input variables (e.g., the number of observations does not exceed the number of input variables). Conversely, if these conditions fail, multiple linear regression would be inappropriate and would result in a useless, overfit predictive model.

Partial Least Squares (PLS) is a flexible regression technique that has features of principal components analysis and extends multiple linear regression. PLS is best used when the data set contains fewer observations than input variables and high collinearity exists. PLS can also be used as an exploratory analysis tool to identify input variables to include in the model and outliers. Like multiple linear regression, PLS' overarching goal is to create a linear predictive model.

Activity:

1. Apply Predictive Modelling using Regression based on the dataset in the previous module

Modules Test

1. Describe a business problem where it would be appropriate to use regression for analysis.
2. Describe conditions which may cause significant differences between the backward, forward, and stepwise methods.
3. There are many different types of regression. Research one other regression type, not covered in this chapter. Discuss how it compares to linear or logistic regression.
4. Explain what is meant by underfitting and overfitting, and describe an example of each.
5. Present a checklist for all of the steps that should be performed to complete a multiple linear or logistic regression.

MODULE 5
Predictive Model Using Decision Tree

In this module, decision trees are defined and then demonstrated to show how they can be used as an important predictive modeling tool. Both classification and regression decision trees will be considered.

Learning Outcomes:

1. Describe decision trees.
2. Create and utilize decision trees.
3. Evaluate the key properties that define a decision tree.
4. Evaluate the results of a decision tree output.

Introduction

Decision trees are an important predictive modeling tool, not because of their complexity but because of their simplicity. They are often used to be able to provide an easy method to determine which input variables have an important impact on a target variable. Decision trees can be system generated or built interactively; both will be demonstrated. Focus will be given to interpreting the output of a decision tree. Decision trees are useful when it is important to understand and even control which variables impact a target.

Discussion

What Is a Decision Tree?

Decision trees are one of the most widely used predictive and descriptive analytic tools. One reason for their popularity is that they are easy to create, and the output is easy to understand. They are particularly useful in situations where you want to know specifically how you arrived at an outcome. Decision trees require at least one target variable which can be continuous or categorical. Decision trees use algorithms to determine splits within variables that create branches. This forms a tree like structure. The algorithms create a series of IF-THEN -ELSE rules that split the data into successively smaller segments. Each rule evaluates the value of a single variable and based upon its value splits it into one of two or more segments. The segments are called nodes. If a node does not result in a subsequent split (i.e., it has no successor nodes), it is referred to as a leaf . The first node contains all of the data and is referred to as the root node. A node with all of its successor nodes is referred to as a branch of the decision tree. The taller and wider the decision tree is, the more data splits that will have occurred. Decision trees can be a great place to begin predictive modeling to gain a deeper understanding of the impact input variables have on a target variable; however, they rarely produce the best-fit model. A decision tree can be used to perform one of the following tasks:

1. Classify observations based upon a target that is binary, ordinal, or nominal.
2. Predict an outcome for an interval target.
3. Predict a decision when you identify the specific decision alternatives.

Decision trees are frequently used for market or customer segmentation. Consider the following simple example of a decision tree (Fig. 5.1). A bank wants to make a mortgage decision based upon credit rating. The root node is the mortgage decision. The first branch is the credit rating. The node for a low credit rating is a leaf because there is no subsequent node from that branch. The node for a high credit rating is a branch as it has subsequently been split on the income variable creating two additional leaves.

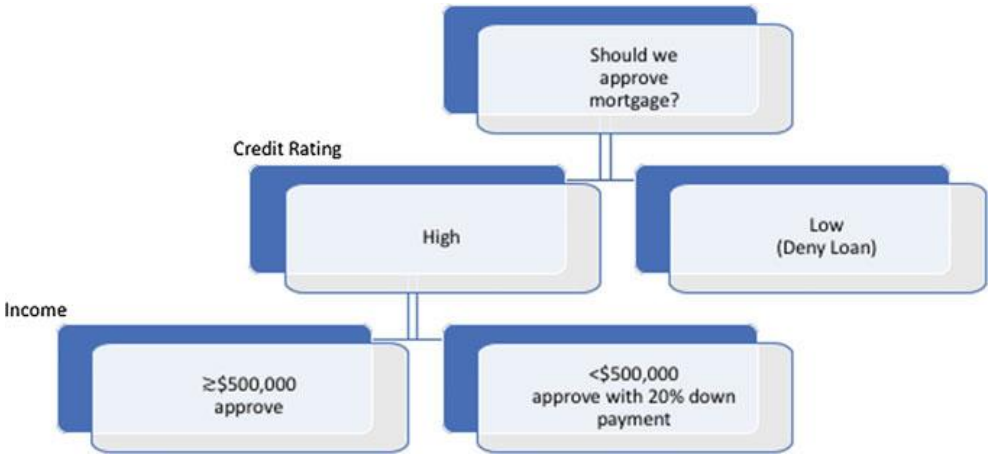


Figure 5.1 Decision Tree Example

Table 5.1 Confidence values

Confidence	Strength of the relationship
0.001	Extremely good
0.01	Good
0.05	Pretty good
0.10	Not so good
0.15	Extremely weak

Creating a Decision Tree

To create a decision tree, a default algorithm is used to split the data. The algorithm begins with a split search. The split search begins by selecting an input variable to be split (i.e., create subsets). If the input variable has an interval data type, each unique value within the test partition serves as a potential split. If the input variable has a categorical data type, the average value of the target within each category is computed.

For the selected input variable, two groups are created. The left branch includes all the cases less than the split value; the right branch contains all the cases greater than the split value.

After the nodes are split, they can then be evaluated to determine if they can be split further. This process continues until a stop condition is reached. A stop condition is reached when one of the following occurs:

1. The maximum depth of the tree has been reached (i.e., the number of nodes between the root node and the given node).
2. The nodes can no longer be split in a meaningful way (i.e., the threshold worth).
3. The nodes contain too few observations to be meaningful.

Classification and Regression Trees (CART)

Decision trees are categorized as classification tree models or regression tree models. This is determined by the data type of the target variable. Categorical target variables result in a classification tree. Interval target variables result in a regression tree. The goal in both cases is to build a model that predicts a target variable based upon the input variables. The major difference between classification and regression trees is the test statistics used to evaluate the splits. For a regression tree, the test statistic used has an F-distribution (i.e., ANOVA) and the assessment measure is the average squared error.

Building a decision tree has three main goals:

1. Decrease the entropy (unpredictability of the target variable—also referred to as the information gain).

2. Ensure that data set is reliable and consistent.

3. Have the smallest number of nodes.

Decision trees can be helpful in creating segments of a larger group. For example, many retail organizations have loyalty cards that are capable of collecting large amounts of data about customer spending patterns. A decision tree can help identify the characteristics of subgroups of customers that have similar spending patterns.

Data Partitions and Decision Trees

The training data set is used to generate the initial decision tree and to develop the splitting rules to assign data to each of the nodes. The training data set assigns each node to a target level to be able to determine a decision. The proportion of cases within the training data set that are assigned to each node (referred to as the posterior

probabilities) is calculated. As is the case with other predictive modeling techniques, the general rule of thumb is the smaller the data set, the larger the percentage that should be allocated to the training data set. In general, the posterior possibilities of the node will be more reliable from a larger training data set.

The training data set will produce an initial or maximal tree. The maximal tree is the whole tree or largest possible tree. However, the optimal tree is the smallest tree that yields the best possible fit. Optimizing a decision tree is referred to as pruning the tree. The validation data set is used to prune a decision tree. Pruning a tree is a process of evaluating the results of subtrees by removing branches from the maximal

tree. Once the subtrees are built, the system provides three possible methods to determine which subtree to use. These are:

1. Assessment (i.e., the best assessment value)

2. Largest (i.e., the subtree with the most leaves)

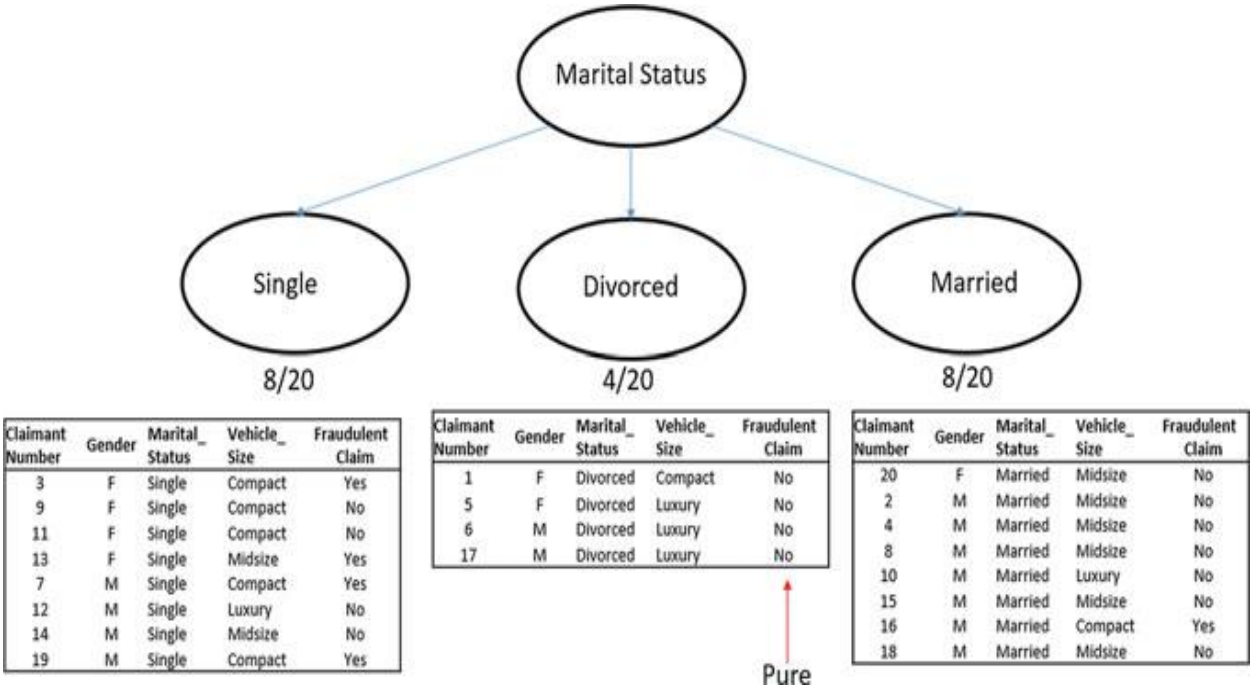
3. N (i.e., the subtree that contains the number of leaves indicated by the user).

Table 5.2 Sample training data set—Fraudulent Claims

Claimant_Number	Gender	Marital_Status	Vehicle_Size	Fraudulent_Claim
1	F	Divorced	Compact	No
2	M	Married	Midsize	No
3	F	Single	Compact	Yes
4	M	Married	Midsize	No
5	F	Divorced	Luxury	No
6	M	Divorced	Luxury	No
7	M	Single	Compact	Yes
8	M	Married	Midsize	No
9	F	Single	Compact	No
10	M	Married	Luxury	No
11	F	Single	Compact	No
12	M	Single	Luxury	No
13	F	Single	Midsize	Yes
14	M	Single	Midsize	No
15	M	Married	Midsize	No
16	M	Married	Compact	Yes
17	M	Divorced	Luxury	No
18	M	Married	Midsize	No
19	M	Single	Compact	Yes
20	F	Married	Midsize	No

Creating a Decision Tree

Let's examine the use of a decision tree as a predictive analytics tool. Two decision trees will be applied to the claim fraud case, and we will examine the results to determine which decision tree provides the best predictive capability.



Claimant Number	Gender	Marital_ Status	Vehicle_ Size	Fraudulent Claim
3	F	Single	Compact	Yes
9	F	Single	Compact	No
11	F	Single	Compact	No
13	F	Single	Midsize	Yes
7	M	Single	Compact	Yes
12	M	Single	Luxury	No
14	M	Single	Midsize	No
19	M	Single	Compact	Yes

Claimant Number	Gender	Marital_ Status	Vehicle_ Size	Fraudulent Claim
1	F	Divorced	Compact	No
5	F	Divorced	Luxury	No
6	M	Divorced	Luxury	No
17	M	Divorced	Luxury	No

Claimant Number	Gender	Marital_ Status	Vehicle_ Size	Fraudulent Claim
20	F	Married	Midsize	No
2	M	Married	Midsize	No
4	M	Married	Midsize	No
8	M	Married	Midsize	No
10	M	Married	Luxury	No
15	M	Married	Midsize	No
16	M	Married	Compact	Yes
18	M	Married	Midsize	No

Figure 5.2 Marital Status Split

Figure 5.3 Decision Tree Node

There are several properties that are essential to controlling the decision tree that will be generated. The following summarizes critical properties of the Decision Tree node:

Train Properties

Variables—This specifies the properties for each variable. The variable properties can be changed for the decision tree only; however, in most cases, it is advisable to change them on the data source so that all other nodes use the same properties.
Interactive—This is used to create a user defined customized decision tree (an example is described below).
Frozen Tree—When set to Yes, it prevents the maximal decision tree from

being changed by other settings. The default is No.

Splitting Rule Properties

Interval Target Criterion—This identifies one of two methods used for split-ting when the target is an interval variable.
Nominal Target Criterion—This identifies one of three methods used for splitting when the target is a nominal variable.
Ordinal Target Criterion—This identifies one of two methods used for splitting when the target is an ordinal variable.

Node Properties

Leaf Size—This specifies the minimum number of training observations. The splitting rules try to create leaves that are as alike as possible within the same branch but different when compared to leaves of other branches at the same level. This property therefore specifies a minimum variability within the tree. The default setting is five.
Number of Rules—A decision tree uses only one rule to determine the split of a node, but others may have been attempted before the final split. This property determines how many rules will be saved for comparison. A splitting rule statistic is computed. LOGWORTH is used for ProbChisq or ProbF statistics; WORTH (variance) is used for entropy or Gini. The default value for the Number of Rules property is five.
Number of Surrogate Rules—A surrogate rule is a back-up to the primary splitting rule. It is used to split non-leaf nodes and invoked when the primary splitting rule relies on a variable with a missing value. The default value for the Number of Surrogate Rules property is zero.
Split Size—This specifies the minimum number of training observations to be considered for a split within a node.

Activity:

1. Apply Predictive Modelling using Decision Tree based on the dataset in the previous module

Modules Test

1. Discuss an example where a decision tree can be used to explain the results of a business problem/issue.
2. Describe three properties that are part of the decision tree node.
3. Explain what is meant by pruning a tree.
4. Using the claim fraud example, how can you modify the maximal tree to improve the maximal tree model?
5. Create a decision tree to solve a problem. Report your assumptions and findings.

MODULE 6
Predictive Models Using Neural Networks

In this module, a variety of different neural network architectures will be described. Next, an analysis of how to optimize and evaluate neural networks will be presented, followed by using a decision tree to show how to describe a neural network. Finally, multiple neural networks will be applied to the automobile insurance data set to determine which neural network provides the best-fit model.

Learning Outcomes:

1. Create and utilize neural networks.
2. Differentiate between different neural network architectures.
3. Evaluate the key properties that define and optimize a neural network.
4. Utilize a decision tree to explain a neural network.

Introduction

One of the most powerful predictive analytics techniques is neural network. The concept of the neural network is over fifty years old, but it is recent advance in computing speed, memory, and data storage that have enabled their more current widespread use.

Discussion

What Is a Neural Network?

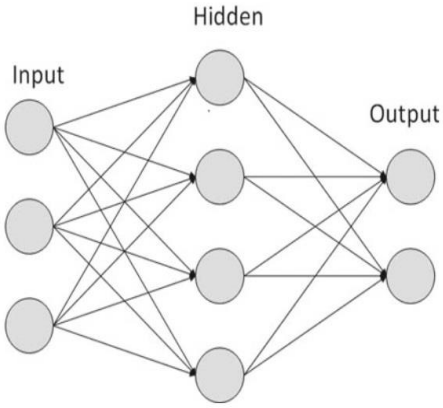
Neural networks are a powerful analytic tool that seeks to mimic functions of the human brain. A key component is their ability to learn from experience. Within the brain, neurons are the component that enables cognition and intelligence. The brain is comprised of a system (network) of neurons that work together to form one cohesive unit. Inputs arrive to each neuron through a connection called a dendrite.

Dendrites transmit their information to the neuron by sending neurotransmitters across a synaptic gap. These neurotransmitters either excite or inhibit the receiving neuron. If they excite the receiving neuron, this is referred to as firing the neuron. If they inhibit the neuron, it does not become active. It is also important to note that the amount of the neurotransmitters that are transmitted across the synaptic gap determines the relative strength of each dendrite's connection to their corresponding neuron. This relative strength can be expressed as the weight of the input.

The weights of the inputs are summed. If the sum of the weights exceeds an acceptable threshold level, then it causes the neuron to fire, sending a pulse to other neurons. This threshold level is referred to as a bias. The more active the synapse, the stronger the

connections, the weaker the synapse, the more likely the neuron atrophies from lack of use.

Neural networks use this same concept to mirror the biological functions of the brain. A neural network consists of neurons with input, hidden layer(s), and output connections. Figure 6.1 presents a simple model of a neural network.



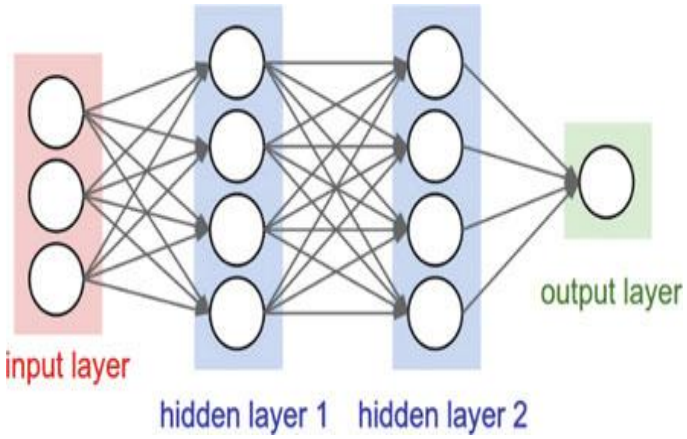


Figure 6.1 NN One hidden Layer

Figure 6.2 NN Two hidden Layer

Neural networks learn from experience through categorization-based machine learning and are useful for pattern recognition-based problems. To accomplish this, they must be trained. Like the human brain, they learn from experience. Training takes place through the use of iterations of data being processed within the network. The more factors that are involved for the network to process, the more nodes (or units) that are needed (and usually the longer it will take to train the network).

As is the case with decisions made by the human brain, some factors have more influence (weight) than others. Increasing the number of nodes and adding weights to those nodes increase the complexity of the network and add to the computational requirements (see Fig. 6.2). It can also increase the amount of time it takes for the network to train. Though training may take time, once a neural network is trained it can be one of the fastest predictive modeling techniques. Therefore, they are frequently used today in applications such as voice response systems and fraud detection.

Neural networks have many applications in business today. They are used in sales forecasting, customer retention and identification of potential new customers, risk management, and industrial process control. Neural networks are used in credit scoring and credit card fraud applications. Neural networks are used widely in applications that make recommendations to customers on complimentary products. They have even begun to move into the healthcare market in applications such as personal heath monitoring. Neural networks work best in situations where prior data is both available and indicative of future events.

History of Neural Networks

The concept of an artificial neural network predates our ability to automate them through specialized computer software. The concept was first put forth by McCulloch and Pitts (1943). They provided theorems to prove that networks of neurons could take any set of binary inputs and transform them into a set of binary outputs. They demonstrated that artificial neural networks are able to model input–output relationships to any degree of precision. McCulloch and Pitts (1943) also purported that the overall structure of the neural network does not change over time; however, the interaction between the neuron's changes. They based their assumptions on the following:

- ✓ The activity of a neuron is an "all-or-none" process.
- ✓ A fixed number of synapses must be excited to excite a neuron. The synapses are
- ✓ independent of the previous position of the neuron and its activity.
- ✓ Synaptic delay is the significant delay.
- ✓ Neuron excitation can be prevented by the presence of inhibitory synapses.
- ✓ Neural network structures do not change over time (McCulloch and Pitts 1943).

In 1949, Canadian neuropsychologist Donald Hebb put forth his theory of Hebbian learning in his book The Organization of Behavior. Hebb (1949) stated that

When an axon of cell A is near enough to excite cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased

Widrow and Hoff (1960) developed their delta rule to address this instability in the weights of the neurons. The delta rule minimizes the sum of the squared error by replacing the receiving neuron's output value with an error value.

Rosenblatt (1958) implemented a variation of the delta rule in a computer model that he referred to as a perceptron. Though his initial model had input layers linked directly to output layers, he experimented with multilayer perceptrons (i.e., the introduction of hidden layers). Hidden layers transform the network from a simple linear model to one that is capable of generating nonlinear relationships, thus more closely mirroring how the brain functions.

Explaining a Neural Network

Neural networks can be a complex and powerful tool, and as a result one of their criticisms is that they are difficult to understand. They are often thought of as a black box, or their output was produced by "the system." In many business problems however, it is not enough to merely produce an answer, and the management of an organization may also

want to know what were the underlying factors (i.e., what were the significant input variables) that led to the result. Using SAS Enterprise Miner™, a neural network can be explained using a decision tree. This provides a reasonable explanation of the most important factors that were utilized by the neural network.

Let's see this in action by examining the multilayer perceptron neural network that contained three hidden units. This node was chosen because it was the best performing model that was created using the claim fraud example. To examine this neural network with a decision tree, two nodes will be utilized (see Fig. 6.25):

1. Metadata node

2. Decision Tree node.

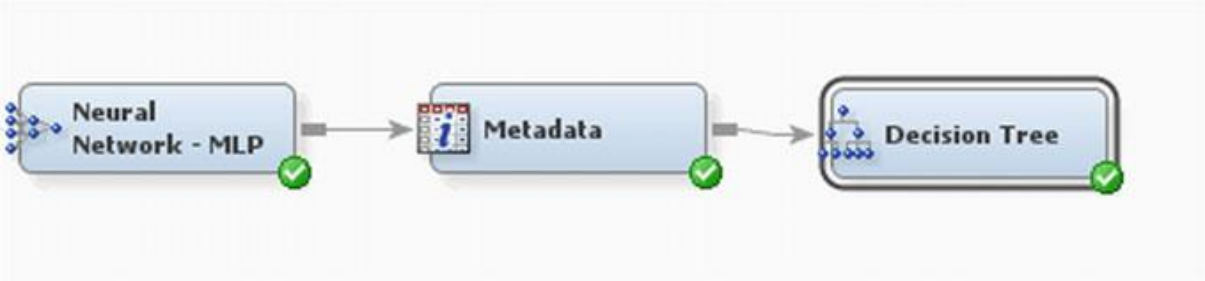


Figure 6.3 Model for decision tree to explain a neural network

Activity:

1. Apply Predictive Modelling using Neural Networks based on the dataset in the previous module

Modules Test

1. What is meant by the term deep learning?
2. There are many applications of neural networks used in businesses today. Discuss one example where an organization is using neural networks. What is the problem or issue they are addressing? How has the neural network helped them? What is the architecture of the network?
3. Discuss how a neural network can be optimized to improve its performance.
4. Compare the use of neural networks for predictive analytics with linear and logistic regression. What are the advantages and disadvantages of each technique?
5. SAS Enterprise Miner™ supports several different neural network architectures and permits a wide variety of options, based on the setting of properties for these networks. Describe three properties that are part of the neural network node. What are their values? When is it appropriate to use them?
6. Explain what is meant by backpropagation.