1. Introduction to Business Forecasting

Business analysts may choose from a wide range of forecasting techniques to support decision making. Selecting the appropriate method depends on the characteristics of the forecasting problem, such as the time horizon of the variable being forecast, as well as available information on which the forecast will be based.

Three major categories of forecasting approaches are *qualitative and judgmental techniques*, *statistical time-series models*, and *explanatory/causal methods*. In this chapter, we introduce forecasting techniques in each of these categories and use basic Excel tools, *XLMiner*, and linear regression to implement them in a spreadsheet environment.

Qualitative and Judgmental Forecasting

Qualitative and judgmental techniques rely on experience and intuition; they are necessary when historical data are not available or when the decision maker needs to forecast far into the future. For example, a forecast of when the next generation of a microprocessor will be available and what capabilities it might have will depend greatly on the opinions and expertise of individuals who understand the technology. Another use of judgmental methods is to incorporate nonquantitative information, such as the impact of government regulations or competitor behavior, in a quantitative forecast. Judgmental techniques range from such simple methods as a manager's opinion or a group-based jury of executive opinion to more structured approaches such as historical analogy and the Delphi method.

The Delphi Method

A popular judgmental forecasting approach, called the **Delphi method**, uses a panel of experts, whose identities are typically kept confidential from one another, to respond to a sequence of questionnaires. After each round of responses, individual opinions, edited to ensure anonymity, are shared, allowing each to see what the other experts think. Seeing other experts' opinions helps to reinforce those in agreement and to influence those who did not agree to possibly consider other factors. In the next round, the experts revise their estimates, and the process is repeated, usually for no more than two or three rounds. The Delphi method promotes unbiased exchanges of ideas and discussion and usually results in some convergence of opinion. It is one of the better approaches to forecasting long range trends and impacts.

Indicators and Indexes

Indicators and indexes generally play an important role in developing judgmental forecasts.

Indicators are measures that are believed to influence the behavior of a variable we wish to forecast. By monitoring changes in indicators, we expect to gain insight about the future behavior of the variable to help forecast the future.

Example 1 Leading Economic Indicators

The Department of Commerce initiated an Index of Leading Indicators to help predict future economic performance.
Components of the index include the following:

- average weekly hours, manufacturing
- average weekly initial claims, unemployment insurance
- new orders, consumer goods, and materials
- vendor performance—slower deliveries
- new orders, nondefense capital goods
- building permits, private housing
- stock prices, 500 common stocks (Standard & Poor)
- money supply
- interest rate spread
- index of consumer expectations (University of Michigan)

Business Conditions Digest included more than 100 time series in seven economic areas. This publication was discontinued in March 1990, but information related to the Index of Leading Indicators was continued in *Survey of Current Business*. In December 1995, the U.S. Department of Commerce sold this data source to The Conference Board, which now markets the information under the title *Business Cycle Indicators*; information can be obtained at its Web site (www.conference-board.org). The site includes excellent current information about the calculation of the index as well as its current components.

1.1 Statistical Forecasting Models

Statistical time-series models find greater applicability for short-range forecasting problems. **Time Series**

A **time series** is a stream of historical data, such as weekly sales. We characterize the values of a time series over T periods as A_t , t = 1, 2, c, T. Time-series models assume that whatever forces have influenced sales in the recent past will continue into the near future; thus, forecasts are developed by extrapolating these data into the future. Time series generally have one or more of the following components: random behavior, trends, seasonal effects, or cyclical effects.

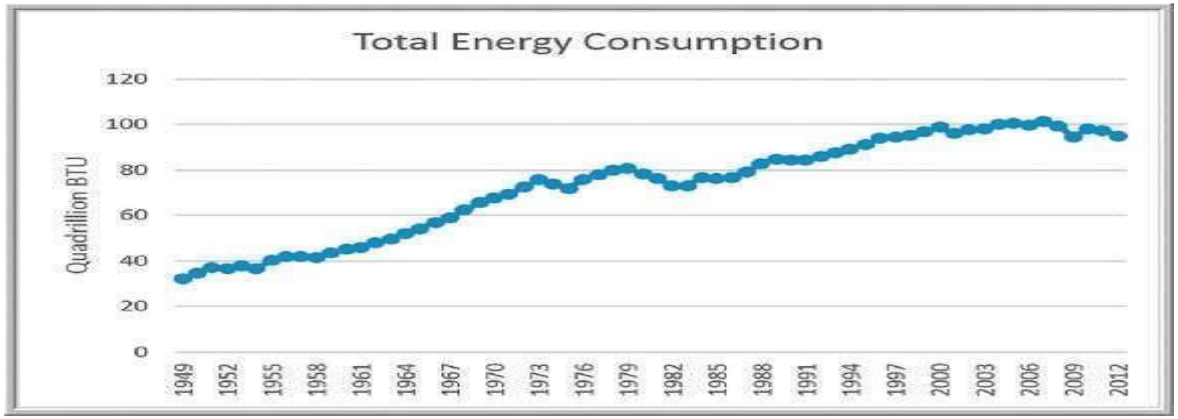
Stationary Time Series

Time series that do not have trend, seasonal, or cyclical effects but are relatively constant and exhibit only random behavior are called **stationary time series**.

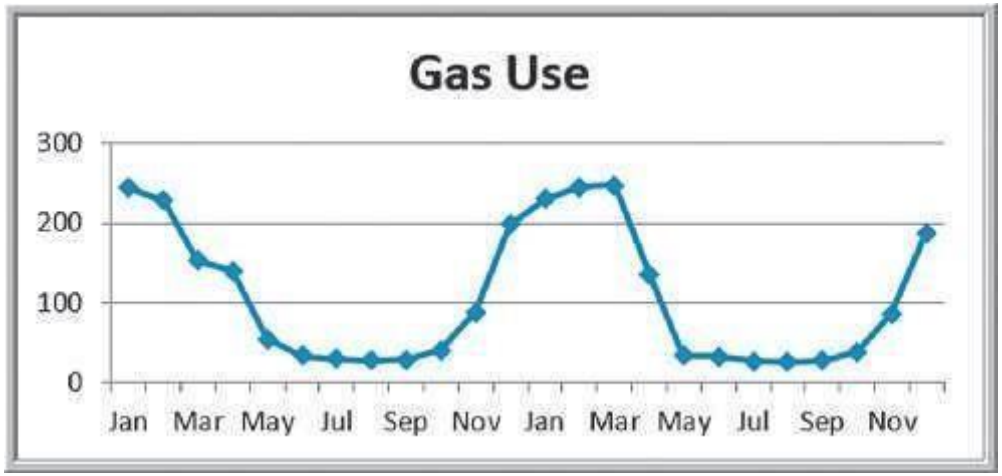
Many forecasts are based on analysis of historical time-series data and are predicated on the assumption that the future is an extrapolation of the past Statistical time-series models find greater applicability for short-range forecasting problems. A **trend** is a gradual upward or downward movement of a time series over time.

Time series may also exhibit short-term seasonal effects (over a year, month, week, or even a day) as well as longer-term cyclical effects, or nonlinear trends. A seasonal effect is one that repeats at fixed intervals of time, typically a year, month, week, or day.

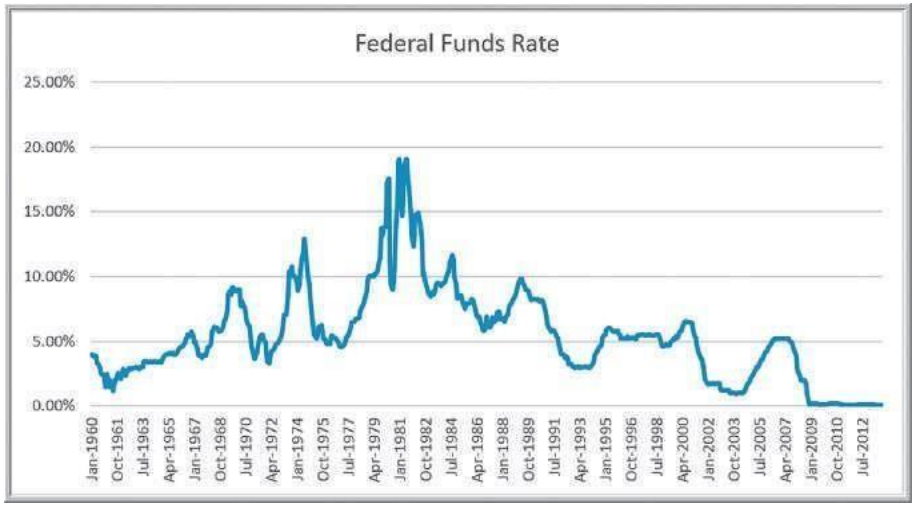
At a neighborhood grocery store, for instance, short-term seasonal patterns may occur over a week, with the heaviest volume of customers on weekends; seasonal patterns may also be evident during the course of a day, with higher volumes in the mornings and late afternoons. Figure shows seasonal changes in natural gas usage for a homeowner over the course of a year (Excel file Gas & Electric). Cyclical effects describe ups and downs over a much longer time frame, such as several years. shows a chart of the data in the Excel file Federal Funds Rates.



Total Energy Consumption Time Series



Seasonal Effects in Natural Gas Usage



Cyclical Effects in Federal Fund Rates

1.2 Moving Average Models

simple moving average method is a smoothing method based on the idea of averaging random fluctuations in the time series to identify the underlying direction in which the time series is changing.

Error Metrics and Forecast Accuracy

The quality of a forecast depends on how accurate it is in predicting future values of a time series. In the simple moving average model, different values for *k* will produce different forecasts.
To analyze the effectiveness of different forecasting models, we can define *error metrics*, which compare quantitatively the forecast with the actual observations. Three metrics that are commonly used are the *mean absolute deviation*, *mean square error*, and *mean absolute percentage error*.

1. Mean Absolute Deviation (MAD):

The **mean absolute deviation (MAD)** is the absolute difference between the actual value and the forecast, averaged over a range of forecasted values:

$$\text{MAD} = \frac{\sum_{t=1}^{n} |A_t - F_t|}{n}$$

where *At* is the actual value of the time series at time *t*, *Ft* is the forecast value for time *t*, and *n* is the number of forecast values (*not* the number of data points since we do not have a forecast value associated with the first *k* data points). MAD provides a robust measure of error and is less affected by extreme observations.

2. **Mean square error (MSE):**

Mean square error (MSE) is probably the most commonly used error metric. It penalizes larger errors because squaring larger numbers has a greater impact than squaring smaller numbers. The formula for MSE is

$$\text{MSE} = \frac{\sum_{t=1}^{n} (A_t - F_t)^2}{n}$$

n represents the number of forecast values used in computing the average.

3. **Root mean square error (RMSE):**

Sometimes the square root of MSE, called the **root mean square error (RMSE)**, is used. Note that unlike MSE, RMSE is expressed in the same units as the data (similar to the difference between a standard deviation and a variance), allowing for more practical comparisons.

$$\text{RMSE} = \sqrt{\frac{\sum_{t=1}^{n} (A_t - F_t)^2}{n}}$$

4. **Mean absolute percentage error (MAPE)**:

MAPE is the average of absolute errors divided by actual observation values.

$$\text{MAPE} = \frac{\sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right|}{n} \times 100$$

The values of MAD and MSE depend on the measurement scale of the time-series data. For example, forecasting profit in the range of millions of dollars would result in very large MAD and MSE values, even for very accurate forecasting models. On the other hand, market share is measured in proportions; therefore, even bad forecasting models will have small values of MAD and MSE. Thus, these measures have no meaning except in comparison with other models used to forecast the same data. Generally, MAD is less affected by extreme observations and is preferable to MSE if such extreme observations are considered rare events with no special meaning. MAPE is different in that the measurement scale is eliminated by dividing the absolute error by the time-series data value. This allows a better relative comparison. Although these comments provide some guidelines, there is no universal agreement on which measure is best.

1.3 Exponential Smoothing Models

Simple Exponential smoothing Model

A versatile, yet highly effective, approach for short-range forecasting is **simple exponential smoothing**. The basic simple exponential smoothing model is

$$F_{t+1} = (1 - \alpha)F_t + \alpha A_t$$
$$= F_t + \alpha(A_t - F_t)$$

where F_{t+1} is the forecast for time period $t + 1$, F_t is the forecast for period t, A_t is the observed value in period t, and a is a constant between 0 and 1 called the **smoothing constant**.

To begin, set $F1$ and $F2$ equal to the actual observation in period 1, $A1$.

Using the two forms of the forecast equation just given, we can interpret the simple exponential smoothing model in two ways. In the first model, the forecast for the next period, F_{t+1}, is a weighted average of the forecast made for period t, Ft, and the actual observation in period t, At. The second form of the model, obtained by simply rearranging terms, states that the forecast for the next period, F_{t+1}, equals the forecast for the last period, Ft, plus a fraction a of the forecast error made in period t, $At - Ft$. Thus, to make a forecast once we have selected the smoothing constant, we need to know only the previous forecast and the actual value. By repeated substitution for Ft in the equation, it is easy to demonstrate that F_{t+1} is a decreasingly weighted average of all past time-series data. Thus,the forecast actually reflects *all* the data, provided that a is strictly between 0 and 1.

Double Exponential Smoothing

In double exponential smoothing, the estimates of a_t and b_t are obtained from the following equations:

$$a_t = \alpha F_t + (1 - \alpha)(a_{t-1} + b_{t-1})$$
$$b_t = \beta(a_t - a_{t-1}) + (1 - \beta)b_{t-1}$$

In essence, we are smoothing both parameters of the linear trend model. From the first equation, the estimate of the level in period t is a weighted average of the observed value at time t and the predicted value at time t, $a_{t+1} + b_{t+1}$,based on simple exponential smoothing. For large values of a, more weight is placed on the observed value. Lower values of a put more weight on the smoothed predicted value. Similarly, from the second equation, the estimate of the trend in period t is a weighted average of the differences in the estimated levels in periods t and t - 1 and the estimate of the level in period t - 1.

Forecasting Time Series with Seasonality:

When time series exhibit seasonality, different techniques provide better forecasts,

- **Regression-Based Seasonal Forecasting Models**

One approach is to use linear regression. Multiple linear regression models with categorical variables can be used for time series with seasonality.

- **Holt-Winters Forecasting for Seasonal Time Series**

Holt-Winters models are similar to exponential smoothing models in that smoothing constants are used to smooth out variations in the level and seasonal patterns over time. For time series with seasonality but no trend, *XLMiner* supports a Holt-Winters method but does not have the ability to optimize the parameters

- **Holt-Winters Models for Forecasting Time Series with seasonality and Trend**

Many time series exhibit both trend and seasonality. Such might be the case for growing sales of a seasonal product. These models combine elements of both the trend and seasonal models. Two types of Holt-Winters smoothing models are often used. The **Holt-Winters additive model** is based on the equation

$$F_{t+1} = a_t + b_t + S_{t-s+1}$$

and the **Holt-Winters multiplicative model** is

$$F_{t+1} = (a_t + b_t)S_{t-s+1}$$

The additive model applies to time series with relatively stable seasonality, whereas the multiplicative model applies to time series whose amplitude increases or decreases over time. Therefore, a chart of the time series should be viewed first to identify the appropriate type of model to use. Three parameters,$\propto$,β,γ, are used to smooth the level, trend,and seasonal factors in the time series. *XLMiner* supports both models.

Selecting Appropriate Time-Series-Based Forecasting Models

The table summarizes the choice of forecasting approaches that can be implemented by *XLMiner* based on characteristics of the time series.

	No Seasonality	Seasonality
No trend	Simple moving average or simple exponential smoothing	Holt-Winters no-trend smoothing model or multiple regression
Trend	Double exponential smoothing	Holt-Winters additive or Holt-Winters multiplicative model

Regression Forecasting with Causal Variables

In many forecasting applications, other independent variables besides time, such as economic indexes or demographic factors, may influence the time series. For example, a manufacturer of hospital equipment might include such variables as hospital capital spending and changes in the proportion of people over the age of 65 in building models to forecast future sales. Explanatory/causal models, often called **econometric models**, seek to identify factors that explain statistically the patterns observed in the variable being forecast, usually with regression analysis.

The Practice of Forecasting

Surveys of forecasting practices have shown that both judgmental and quantitative methods are used for forecasting sales of product lines or product families as well as for broad company and industry forecasts. Simple time-series models are used for short- and medium-range forecasts, whereas regression analysis is the most popular method for long range forecasting. However, many companies rely on judgmental methods far more than quantitative methods, and almost half judgmentally adjust quantitative forecasts.

In practice, managers use a variety of judgmental and quantitative forecasting techniques.
Statistical methods alone cannot account for such factors as sales promotions, unusual environmental disturbances, new product introductions, large one-time orders, and so on. Many managers begin with a statistical forecast and adjust it to account for intangible factors. Others may develop independent judgmental and statistical forecasts and then combine them, either objectively by averaging or in a subjective manner.

It is important to compare quantitatively generated forecasts to judgmental forecasts to see if the forecasting method is adding value in terms of an improved forecast. It is impossible to provide universal guidance as to which approaches are best, because they depend on a variety of factors, including the presence or absence of trends and seasonality, the number of data points available, length of the forecast time horizon, and the experience and knowledge of the forecaster. Often, quantitative approaches will miss significant changes in the data, such as reversal of trends, whereas qualitative forecasts may catch them, particularly when using indicators.

2. Logic and Data Driven Models

Predictive modeling means the developing models that can be used to forecast or predict future events. Models can be developed either through logic or data.

Logic driven models

Logic driven models are created on the basis of inferences and postulations which the sample space and existing conditions provide. Creating logical models require solid understanding of business functional areas, logical skills to evaluate the propositions better and knowledge of business practices and research.

To understand better, let us take an example of a customer who visits a restaurant around six times in a year and spends around ₹5000 per visit. The restaurant gets around 40% margin on per visit billing amount. The annual gross profit on that customer turns out to be 5000 × 6 × 0.40 = ₹12000. 30% of the customers do not return each year, while 70% do return to provide more business to the restaurant.

Assuming the average lifetime of a customer (time for which a consumer remains a customer) W 1/.3 = 3.33 years. So, the average gross profit for a typical customer turns out to be 12000 × 3.33 = ₹39,960.

Armed with all the above details, we can logically arrive at a conclusion and can derive the following model for the above problem statement:

Economic Value of each Customer (V) = (R × F × M)/D

Where,

R = Revenue generated per customer

F = Frequency of visits per year

M = Profit margin

D = Defection rate (non-returning customers each year)

Example –

A restaurant customer dines 6 times a year and spends an average of $50 per visit. The restaurant realizes a 40% margin on the average bill for food and drinks.

Annual gross profit on a customer = $50(6)(0.40) = $120

30% of customers do not return each year. Average lifetime of a customer = 1/.3 = 3.33 years.

Average gross profit for a customer = $120(3.33) = $400

OR Average gross profit for a customer = $120/.3 = $400

$$V = \frac{R \times F \times M}{D}$$

Thus, the economic value of a customer is

- V = value of a loyal customer
- R = revenue per purchase
- F = purchase frequency (number visits per year)
- M = gross profit margin
- D = defection rate (proportion customers not returning each year)

Data-driven Models

The main aim of data-driven model concept is to find links between the state system variables (input and output) without clear knowledge of the physical attributes and behaviour of the system. The data driven predictive modelling derives the modelling method based on the set of existing data and entails a predictive methodology to forecast the future outcomes.

It is data-driven only when there is no clear knowledge of the relationships among variables/system, though there is lot of data. Here, you are simply predicting the outcomes based on the data. The model is not based on hand-picked variables, but may contain unobserved, hidden combination of variables.

It refers to the models in which data is collected from many sources to qualitatively establish model relationships. Logic driven models is often used as a first step to establish relationships through data-driven models. Data driven models include sampling and estimation, regression analysis, correlation analysis, forecasting models and stimulation.

3. Data Mining and Predictive Analysis Modelling:

Data mining is a rapidly growing field of business analytics that is focused on better understanding characteristics and patterns among variables in large databases using a variety of statistical and analytical tools. Many of the tools that we have studied in previous chapters, such as data visualization, data summarization, PivotTables, correlation and regression analysis, and other techniques, are used extensively in data mining. However, as the amount of data has grown exponentially, many other statistical and analytical methods have been developed to identify relationships among variables in large data sets and understand hidden patterns that they may contain

Some common approaches in data mining include the following

Data Exploration and Reduction.

This often involves identifying groups in which the elements of the groups are in some way similar. This approach is often used to understand differences among customers and segment them into homogenous groups. For example, Macy's department stores identified four lifestyles of its customers: "Katherine," a traditional, classic dresser who doesn't take a lot of risks and likes quality; "Julie," neotraditional and slightly more edgy but still classic; "Erin," a contemporary customer who loves newness and shops by brand; and "Alex," the fashion customer who wants only the latest and greatest (they have male versions also).4 Such segmentation is useful in design and marketing activities to better target product offerings. These techniques have also been used to identify characteristics of successful employees and improve recruiting and hiring practices.

• **Classification**. Classification is the process of analyzing data to predict how to classify a new data element. An example of classification is spam filtering in an e-mail client. By examining textual characteristics of a message (subject header, key words, and so on), the message is classified as junk or not. Classification methods can help predict whether a credit-card transaction may be fraudulent, whether a loan applicant is high risk, or whether a consumer will respond to an advertisement.

• **Association**. Association is the process of analyzing databases to identify natural associations among variables and create rules for target marketing or buying recommendations.

For example, Netflix uses association to understand what types of movies a customer likes and provides recommendations based on the data. Amazon.com also makes recommendations based on past purchases. Supermarket loyalty cards collect data on customers' purchasing habits and print coupons at the point of purchase based on what was currently bought.

• **Cause-and-effect modeling**. Cause-and-effect modeling is the process of developing analytic models to describe the relationship between metrics that drive business performance—for instance, profitability, customer satisfaction, or employee satisfaction. Understanding the drivers of performance can lead to better decisions to improve performance. For example, the controls group of Johnson Controls, Inc., examined the relationship between satisfaction and contract-renewal rates. They found that 91% of contract renewals came from customers who were either satisfied or very satisfied, and customers who were not satisfied had a much higher defection rate. Their model predicted that a one-percentage point increase in the overall satisfaction score was worth $13 million in service contract renewals annually. As a result, they identified decisions that would improve customer satisfaction. Regression and correlation analysis are key tools for cause-and-effect modelling.

Predictive Modeling

Predictive modeling is a commonly used statistical technique to predict future behavior. Predictive modeling solutions are a form of data-mining technology that works by analyzing historical and current data and generating a model to help predict future outcomes.

In predictive modeling, data is collected, a statistical model is formulated, predictions are made, and the model is validated (or revised) as additional data becomes available. For example, risk models can be created to combine member information in complex ways with demographic and lifestyle information from external sources to improve underwriting accuracy. Predictive models analyze past performance to assess how likely a customer is to exhibit a specific behavior in the future. This category also encompasses models that seek out subtle data patterns to answer questions about customer performance, such as fraud detection models. Predictive models often perform calculations during live transactions—for example, to evaluate the risk or opportunity of a given customer or transaction to guide a decision. If health insurers could accurately predict secular trends (for example, utilization), premiums would be set appropriately, profit targets would be met with more consistency, and health insurers would be more competitive in the marketplace.

Predictive modeling is a method of predicting future outcomes by using data modeling. It's one of the premier ways a business can see its path forward and make plans accordingly. While not fool proof, this method tends to have high accuracy rates, which is why it is so commonly used. Predictive modelling uses statistics to predict outcomes. Most often the event one wants to predict is in the future, but predictive modelling can be applied to any type of unknown event, regardless of when it occurred. For example, predictive models are often used to detect crimes and identify suspects, after the crime has taken place.

In many cases the model is chosen on the basis of detection theory to try to guess the probability of an outcome given a set amount of input data, for example given an email determining how likely that it is spam. Models can use one or more classifiers in trying to determine the probability of a set of data belonging to another set.

For example, a model might be used to determine whether an email is spam or "ham" (non-spam). Depending on definitional boundaries, predictive modelling is synonymous with, or largely overlapping with, the field of machine learning, as it is more commonly referred to in academic or research and development contexts. When deployed commercially, predictive modelling is often referred to as predictive analytics.

Predictive modelling is often contrasted with causal modelling/analysis. In the former, one may be entirely satisfied to make use of indicators of, or proxies for, the outcome of interest. In the latter, one seeks to determine true cause-and-effect relationships. This distinction has given rise to a burgeoning literature in the fields of research methods and statistics and to the common statement that "correlation does not imply causation".

1.4 What Is Predictive Modeling?

In short, predictive modeling is a statistical technique using machine learning and data mining to predict and forecast likely future outcomes with the aid of historical and existing data. It works by analyzing current and historical data and projecting what it learns on a model generated to forecast likely outcomes.

Predictive modeling can be used to predict just about anything, from TV ratings and a customer's next purchase to credit risks and corporate earnings. A predictive model is not fixed; it is validated or revised regularly to incorporate changes in the underlying data. In other words, it's not a one-and-done prediction. Predictive models make assumptions based on what has happened in the past and what is happening now.

If incoming, new data shows changes in what is happening now, the impact on the likely future outcome must be recalculated, too. For example, a software company could model historical sales data against marketing expenditures across multiple regions to create a model for future revenue based on the impact of the marketing spend. Most predictive models work fast and often complete their calculations in real time. That's why banks and retailers can, for example, calculate the risk of an online mortgage or credit card application and accept or decline the request almost instantly based on that prediction. Some predictive models are more complex, such as those used in computational biology and quantum computing; the resulting outputs take longer to compute than a credit card application but are done much more quickly than was possible in the past thanks to advances in technological capabilities, including computing power.

Top 5 Types of Predictive Models

Fortunately, predictive models don't have to be created from scratch for every application. Predictive analytics tools use a variety of vetted models and algorithms that can be applied to a wide spread of use cases.

Predictive modeling techniques have been perfected over time. As we add more data, more muscular computing, AI and machine learning and see overall advancements in analytics, we're able to do more with these models.

The top five predictive analytics models are:

1. **Classification model**:
   Considered the simplest model, it categorizes data for simple and direct query response. An example use case would be to answer the question "Is this a fraudulent transaction?"

2. **Clustering model**:
   This model nests data together by common attributes. It works by grouping things or people with shared characteristics or behaviors and plans strategies for each group at a larger scale. An example is in determining credit risk for a loan applicant based on what other people in the same or a similar situation did in the past.
3. **Forecast model**:
   This is a very popular model, and it works on anything with a numerical value based on learning from historical data. For example, in answering how much lettuce a restaurant should order next week or how many calls a customer support agent should be able to handle per day or week, the system looks back to historical data.
4. **Outliers model**:
   This model works by analyzing abnormal or outlying data points. For example, a bank might use an outlier model to identify fraud by asking whether a transaction is outside of the customer's normal buying habits or whether an expense in a given category is normal or not. For example, a $1,000 credit card charge for a washer and dryer in the cardholder's preferred big box store would not be alarming, but $1,000 spent on designer clothing in a location where the customer has never charged other items might be indicative of a breached account.
5. **Time series model:**
   This model evaluates a sequence of data points based on time. For example, the number of stroke patients admitted to the hospital in the last four months is used to predict how many patients the hospital might expect to admit next week, next month or the rest of the year. A single metric measured and compared over time is thus more meaningful than a simple average.

Predictive Algorithms:

Some of the more common predictive algorithms are:

1. **Random Forest**: This algorithm is derived from a combination of decision trees, none of which are related, and can use both classification and regression to classify vast amounts of data.
2. **Generalized Linear Model (GLM) for Two Values**: This algorithm narrows down the list of variables to find "best fit." It can work out tipping points and change data capture and other influences, such as categorical predictors, to determine the "best fit" outcome, thereby overcoming drawbacks in other models, such as a regular linear regression.
3. **Gradient Boosted Model:** This algorithm also uses several combined decision trees, but unlike Random Forest, the trees are related. It builds out one tree at a time, thus enabling the next tree to correct flaws in the previous tree. It's often used in rankings, such as on search engine outputs.
4. **K-Means**: A popular and fast algorithm, K-Means groups data points by similarities and so is often used for the clustering model. It can quickly render things like personalized retail offers to individuals within a huge group, such as a million or more customers with a similar liking of lined red wool coats.

5. **Prophet**: This algorithm is used in time-series or forecast models for capacity planning, such as for inventory needs, sales quotas and resource allocations. It is highly flexible and can easily accommodate heuristics and an array of useful assumptions.

Predictive modeling is often performed using curve and surface fitting, time series regression, or machine learning approaches. Regardless of the approach used, the process of creating a predictive model is the same across methods.

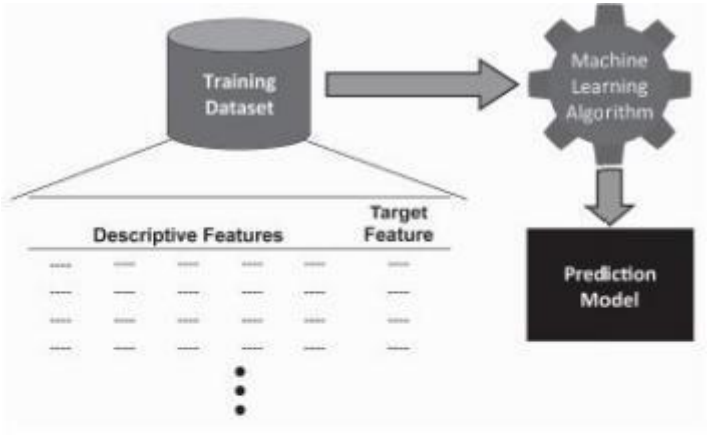
Steps for Predictive Modeling:

The steps are:

1. Clean the data by removing outliers and treating missing data.
2. Identify a parametric or nonparametric predictive modeling approach to use.
3. Preprocess the data into a form suitable for the chosen modeling algorithm.
4. Specify a subset of the data to be used for training the model.
5. Train, or estimate, model parameters from the training data set.
6. Conduct model performance or goodness-of-fit tests to check model adequacy. 7. Validate predictive modeling accuracy on data not used for calibrating the model.
8. Use the model for prediction if satisfied with its performance.

4. Machine Learning for Predictive Analytics

Machine learning is defined as an automated process that extracts patterns from data. To build the models used in predictive data analytics applications, we use supervised machine learning. Supervised machine learning techniques automatically learn a model of the relationship between a set of descriptive features and a target feature based on a set of historical examples, or instances. We can then use this model to make predictions for new instances. These two separate steps are shown in figure,



(a) Learning a model from a set of historical instances

(b) Using a model to make predictions

The two steps in supervised machine learning. Table1.1 lists a set of historical instances, or dataset, of mortgages that a bank has granted in the past. This dataset includes descriptive features that describe the mortgage, and a target feature that indicates whether the mortgage applicant ultimately defaulted on the loan or paid it back in full. The descriptive features tell us three pieces of information about the mortgage: the OCCUPATION (which can be professional or industrial) and AGE of the applicant and the ratio between the applicant's salary and the amount borrowed (LOANSALARY RATIO). The target feature, OUTCOME, is set to either default or repay. In machine learning terms, each row in the dataset is referred to as a training instance, and the overall dataset is referred to as a training data sets.

Table 1.1

A credit scoring dataset.

ID	OCCUPATION	AGE	LOAN-SALARY RATIO	OUTCOME
1	industrial	34	2.96	repay
2	professional	41	4.64	default
3	professional	36	3.22	default
4	professional	41	3.11	default
5	industrial	48	3.80	default
6	industrial	61	2.52	repay
7	professional	37	1.50	repay
8	professional	40	1.93	repay
9	industrial	33	5.25	default
10	industrial	32	4.15	default

An example of a very simple prediction model for this domain would be

if LOAN-SALARY RATIO > 3 then OUTCOME = default

else

OUTCOME = repay

We can say that this model is consistent with the dataset as there are no instances in the dataset for which the model does not make a correct prediction. When new mortgage applications are made, we can use this model to predict whether the applicant will repay the mortgage

or default on it and make lending decisions based on this prediction. Machine learning algorithms automate the process of learning a model that captures the relationship between the descriptive features and the target feature in a dataset. For simple datasets like the one in Table , we may be able to manually create a prediction model, and in an example of this scale, machine learning has little to offer us. Consider, however, the dataset in Table, which shows a more complete representation of the same problem. This dataset lists more instances, and there are extra descriptive features describing the AMOUNT that a mortgage holder borrows, the mortgage holder's SALARY, the type of PROPERTY that the mortgage relates to (which can be farm,house, or apartment) and the TYPE of mortgage (which can be ftp for first-time buyers or stb for second-time buyers).

The simple prediction model using only the loan-salary ratio feature is no longer consistent with the dataset. It turns out, however, that there is at least one prediction model that is consistent with the dataset; it is just a little harder to find than the previous one:

if LOAN-SALARY RATIO < 1.5 then OUTCOME = repay

else if LOAN-SALARY RATIO > 4 then OUTCOME = default

else if AGE < 40 and OCCUPATION =industrial then OUTCOME = default

else

OUTCOME = repay

To manually learn this model by examining the data is almost impossible. For a machine learning algorithm, however, this is simple. When we want to build prediction models from large datasets with multiple features, machine learning is the solution.

How does Machine Learning Work?

Machine learning algorithms work by searching through a set of possible prediction models for the model that best captures the relationship between the descriptive features and target feature in a dataset. An obvious criteria for driving this search is to look for models that are consistent with the data.

There are, however, at least two reasons why just searching for consistent models is not sufficient in order to learn useful prediction models.

First, when we are dealing with large datasets, it is likely that there will be noise in the data, and prediction models that are consistent with noisy data will make incorrect predictions.

Second, in the vast majority of machine learning projects, the training set represents only a small sample of the possible set of instances in the domain. As a result, machine learning is an ill-posed problem. An ill-posed problem is a problem for which a unique solution cannot be determined using only the information that is available. Table 1.2

A more complex credit scoring dataset.

ID	AMOUNT	SALARY	LOAN-SALARY RATIO	AGE	OCCUPATION	PROPERTY	TYPE	OUTCOME
1	245,100	66,400	3.69	44	industrial	farm	stb	repay
2	90,600	75,300	1.20	41	industrial	farm	stb	repay
3	195,600	52,100	3.75	37	industrial	farm	ftb	default
4	157,800	67,600	2.33	44	industrial	apartment	ftb	repay
5	150,800	35,800	4.21	39	professional	apartment	stb	default
6	133,000	45,300	2.94	29	industrial	farm	ftb	default
7	193,100	73,200	2.64	38	professional	house	ftb	repay
8	215,000	77,600	2.77	17	professional	farm	ftb	repay
9	83,000	62,500	1.33	30	professional	house	ftb	repay
10	186,100	49,200	3.78	30	industrial	house	ftb	default
11	161,500	53,300	3.03	28	professional	apartment	stb	repay
12	157,400	63,900	2.46	30	professional	farm	stb	repay
13	210,000	54,200	3.87	43	professional	apartment	ftb	repay
14	209,700	53,000	3.96	39	industrial	farm	ftb	default
15	143,200	65,300	2.19	32	industrial	apartment	ftb	default
16	203,000	64,400	3.15	44	industrial	farm	ftb	repay
17	247,800	63,800	3.88	46	industrial	house	stb	repay
18	162,700	77,400	2.10	37	professional	house	ftb	repay
19	213,300	61,100	3.49	21	industrial	apartment	ftb	default
20	284,100	32,300	8.80	51	industrial	farm	ftb	default
21	154,000	48,900	3.15	49	professional	house	stb	repay
22	112,800	79,700	1.42	41	professional	house	ftb	repay
23	252,000	59,700	4.22	27	professional	house	stb	default
24	175,200	39,900	4.39	37	professional	apartment	stb	default
25	149,700	58,600	2.55	35	industrial	farm	stb	default

Table 1.3

A simple retail dataset

ID	BBY	ALC	ORG	GRP
1	no	no	no	couple
2	yes	no	yes	family
3	yes	yes	no	family
4	no	no	yes	couple
5	no	yes	yes	single

We can illustrate how machine learning is an ill-posed problem using an example in which the analytics team at a supermarket chain wants to be able to classify customer households into the demographic groups single, couple, or family, based solely on their shopping habits.

The dataset in Table 1.3 contains descriptive features describing the shopping habits of 5 customers. The descriptive features measure whether a customer buys baby food, BBY, alcohol, ALC, or organic vegetable products, ORG. Each feature can take one of the two values: yes or no. Alongside these descriptive features is a target feature, GRP, that describes the demographic group for each customer (single, couple, or family). The dataset in Table 1.3 is referred to as a labeled dataset because it includes values for the target feature.

Imagine we attempt to learn a prediction model for this retail scenario by searching for a model that is consistent with the dataset. The first thing we need to do is figure out many different possible models actually exist for the scenario. This defines the set of prediction models the machine learning algorithm will search. From the perspective of searching for a consistent model, the most important property of a prediction model is that it defines a mapping from every possible combination of descriptive feature values to a prediction for the target feature. For the retail scenario, there are only three binary descriptive features, so there are $2^3 = 8$ possible combinations of descriptive feature values. However, for each of these 8 possible descriptive feature value combinations, there are 3 possible target feature values, so this means that there are $3^8 =$ 6,561 possible prediction models that could be used. Table illustrates the relationship between descriptive feature value combinations and prediction models for the retail scenario. The descriptive feature combinations are listed on the left hand side of the table and the set of potential models for this domain are shown as 1 to 6,561 on the right hand side of the table. Using the training dataset from Table1.3 , a machine learning

algorithm will reduce the full set of 6,561 possible prediction models for this scenario down to just those that are consistent with the training instances. Table 1.4(b) illustrates this; the blanked out columns in the table indicate the models that are not consistent with the training data.

Table 1.4

Potential prediction models (a) before and (b) after training data becomes available.

BBY	ALC	ORG	GRP	M_1	M_2	M_3	M_4	M_5	...	M_{6561}
no	no	no	?	couple	couple	single	couple	couple	...	couple
no	no	yes	?	single	couple	single	couple	couple	...	single
no	yes	no	?	family	family	single	single	single	...	family
no	yes	yes	?	single	single	single	single	single	...	couple
yes	no	no	?	couple	couple	family	family	family	...	family
yes	no	yes	?	couple	family	family	family	family	...	couple
yes	yes	no	?	single	family	family	family	family	...	single
yes	yes	yes	?	single	single	family	family	couple	...	family

(b)

BBY	ALC	ORG	GRP	M_2	M_4	M_5	...
no	no	no	couple	couple	couple	couple	...
no	no	yes	couple	couple	couple	couple	...
no	yes	no	?	family	single	single	...
no	yes	yes	single	single	single	single	...
yes	no	no	?	couple	family	family	...
yes	no	yes	family	family	family	family	...
yes	yes	no	family	family	family	family	...
yes	yes	yes	?	single	family	couple	...

Table1.4(b) also illustrates the fact that the training dataset does not contain an instance for every possible descriptive feature value combination and that there are still a large number of potential prediction models that remain consistent with the training dataset after the inconsistent models have been excluded. Specifically, there are three remaining descriptive feature value combinations for which the correct target feature value is not known, and therefore there are $3^3 = 27$ potential models that remain consistent with the training data. Three of these- M_2,M_4,M_5- shown in Table1.4(b). Because a single consistent model cannot be found based on the sample training dataset alone, we say that machine learning is fundamentally an ill-posed problem.

We might be tempted to think that having multiple models that are consistent with the data is a good thing. The problem is, however, that although these models agree on what predictions should be made for the instances in the training dataset, they disagree with regard to what predictions should be returned for instances that are not in the training dataset. For example, if a new customer starts shopping at the supermarket and buys baby food, alcohol, and organic vegetables, our set of consistent models will contradict each other with respect to what prediction should be returned for this customer, for example, M_2 will return GRP = single, M_4 will return GRP = family, and M_5 will return GRP = couple. The criterion of consistency with the training data doesn't provide any guidance with regard to which of the consistent models to prefer when dealing with queries that are outside the training dataset. As a result, we cannot use the set of consistent models to make predictions for these queries. In fact, searching for predictive models that are consistent with the dataset is equivalent to just memorizing the dataset. As a result, no learning is taking place because the set of consistent models tells us nothing about the underlying relationship between the descriptive and target features beyond what a simple look-up of the training dataset would provide.

If a predictive model is to be useful, it must be able to make predictions for queries that are not present in the data. A prediction model that makes the correct predictions for these queries captures the underlying relationship between the descriptive and target features and is said to generalize well. Indeed, the goal of machine learning is to find the predictive model that generalizes best. In order to find this single best model, a machine learning algorithm must use some criteria for choosing among the candidate models it considers during its search.

Given that consistency with the dataset is not an adequate criterion to select the best prediction model, what criteria should we use? There are a lot of potential answers to this question, and that is why there are a lot of different machine learning algorithms. Each machine learning algorithm uses different model selection criteria to drive its search for the best predictive model. So, when we choose to use one machine learning algorithm instead of another, we are, in effect, choosing to use one model selection criterion instead of another.

All the different model selection criteria consist of a set of assumptions about the characteristics of the model that we would like the algorithm to induce. The set of assumptions that defines the model selection criteria of a machine learning algorithm is known as the inductive bias 6 of the machine learning algorithm.

There are two types of inductive bias that a machine learning algorithm can use, a restriction bias and a preference bias. A restriction bias constrains the set of models that the algorithm will consider during the learning process. A preference bias guides the learning algorithm to prefer certain models over others.

For example, we introduce a machine learning algorithm called multivariable linear regression with gradient descent, which implements the restriction bias of only considering prediction models that produce predictions based on a linear combination of the descriptive feature values and applies a preference bias over the order of the linear models it considers in terms of a gradient descent approach through a weight space. As a second example, we introduce the Iterative Dichotomize 3 (ID3) machine learning algorithm, which uses a restriction bias of only considering tree prediction models where each branch encodes a sequence of checks on individual descriptive features but also utilizes a preference bias by

considering shallower (less complex) trees over larger trees. It is important to recognize that using an inductive bias is a necessary prerequisite for learning to occur; without inductive bias, a machine learning algorithm cannot learn anything beyond what is in the data.

In summary, machine learning works by searching through a set of potential models to find the prediction model that best generalizes beyond the dataset. Machine learning algorithms use two sources of information to guide this search, the training dataset and the inductive bias assumed by the algorithm.