Predictive analytics

Predictive analytics encompasses a variety of statistical techniques from predictive modeling, machine learning, and data mining that analyze current and historical facts to make predictions about future, or otherwise unknown, events.[1][2]

In business, predictive models exploit patterns found in historical and transactional data to identify risks and opportunities. Models capture relationships among many factors to allow assessment of risk or potential associated with a particular set of conditions, guiding decision making for candidate transactions.[3]

The defining functional effect of these technical approaches is that predictive analytics provides a predictive score (probability) for each individual (customer, employee, healthcare patient, product SKU, vehicle, component, machine, or other organizational unit) in order to determine, inform, or influence organizational processes that pertain across large numbers of individuals, such as in marketing, credit risk assessment, fraud detection, manufacturing, healthcare, and government operations including law enforcement.

Predictive analytics is used in actuarial science,[4] marketing,[5] financial services,[6] insurance, telecommunications,[7] retail,[8] travel,[9] healthcare,[10] pharmaceuticals,[11] capacity planning and other fields.

One of the most well known applications is credit scoring,[1] which is used throughout financial services. Scoring models process a customer's credit history, loan application, customer data, etc., in order to rank-order individuals by their likelihood of making future credit payments on time.

1 Definition

Predictive analytics is an area of data mining that deals with extracting information from data and using it to predict trends and behavior patterns. Often the unknown event of interest is in the future, but predictive analytics can be applied to any type of unknown whether it be in the past, present or future. For example, identifying suspects after a crime has been committed, or credit card fraud as it occurs.[12] The core of predictive analytics relies on capturing relationships between explanatory variables and the predicted variables from past occurrences, and exploiting them to predict the unknown outcome. It is important to note, however, that the accuracy and usability of results will depend greatly on the level of data analysis and the quality of assumptions.

Predictive analytics is often defined as predicting at a more detailed level of granularity, i.e., generating predictive scores (probabilities) for each individual organizational element. This distinguishes it from forecasting. For example, "Predictive analytics—Technology that learns from experience (data) to predict the future behavior of individuals in order to drive better decisions."[13]

2 Types

Generally, the term predictive analytics is used to mean predictive modeling, "scoring" data with predictive models, and forecasting. However, people are increasingly using the term to refer to related analytical disciplines, such as descriptive modeling and decision modeling or optimization. These disciplines also involve rigorous data analysis, and are widely used in business for segmentation and decision making, but have different purposes and the statistical techniques underlying them vary.

2.1 Predictive models

Predictive models are models of the relation between the specific performance of a unit in a sample and one or more known attributes or features of the unit. The objective of the model is to assess the likelihood that a similar unit in a different sample will exhibit the specific performance. This category encompasses models in many areas, such as marketing, where they seek out subtle data patterns to answer questions about customer performance, or 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, in order to guide a decision. With advancements in computing speed, individual agent modeling systems have become capable of simulating human behaviour or reactions to given stimuli or scenarios.

The present predictive models are mainly targeted to one aspect of computing, namely network, memory, processor, or other such as database. This is mainly due to complicated behaviour of any application. As presented in the paper [14] one can see the network based prediction.

The available sample units with known attributes and known performances is referred to as the "training sample." The units in other samples, with known attributes but unknown performances, are referred to as "out of

[training] sample" units. The out of sample bear no chronological relation to the training sample units. For example, the training sample may consists of literary attributes of writings by Victorian authors, with known attribution, and the out-of sample unit may be newly found writing with unknown authorship; a predictive model may aid in attributing a work to a known author. Another example is given by analysis of blood splatter in simulated crime scenes in which the out of sample unit is the actual blood splatter pattern from a crime scene. The out of sample unit may be from the same time as the training units, from a previous time, or from a future time.

2.2 Descriptive models

Descriptive models quantify relationships in data in a way that is often used to classify customers or prospects into groups. Unlike predictive models that focus on predicting a single customer behavior (such as credit risk), descriptive models identify many different relationships between customers or products. Descriptive models do not rank-order customers by their likelihood of taking a particular action the way predictive models do. Instead, descriptive models can be used, for example, to categorize customers by their product preferences and life stage. Descriptive modeling tools can be utilized to develop further models that can simulate large number of individualized agents and make predictions.

2.3 Decision models

Decision models describe the relationship between all the elements of a decision — the known data (including results of predictive models), the decision, and the forecast results of the decision — in order to predict the results of decisions involving many variables. These models can be used in optimization, maximizing certain outcomes while minimizing others. Decision models are generally used to develop decision logic or a set of business rules that will produce the desired action for every customer or circumstance.

3 Applications

Although predictive analytics can be put to use in many applications, we outline a few examples where predictive analytics has shown positive impact in recent years.

3.1 Analytical customer relationship management (CRM)

Analytical Customer Relationship Management is a frequent commercial application of Predictive Analysis. Methods of predictive analysis are applied to customer data to pursue CRM objectives, which involve constructing a holistic view of the customer no matter where their information resides in the company or the department involved. CRM uses predictive analysis in applications for marketing campaigns, sales, and customer services to name a few. These tools are required in order for a company to posture and focus their efforts effectively across the breadth of their customer base. They must analyze and understand the products in demand or have the potential for high demand, predict customers' buying habits in order to promote relevant products at multiple touch points, and proactively identify and mitigate issues that have the potential to lose customers or reduce their ability to gain new ones. Analytical Customer Relationship Management can be applied throughout the customers lifecycle (acquisition, relationship growth, retention, and win-back). Several of the application areas described below (direct marketing, cross-sell, customer retention) are part of Customer Relationship Managements.

3.2 Clinical decision support systems

Experts use predictive analysis in health care primarily to determine which patients are at risk of developing certain conditions, like diabetes, asthma, heart disease, and other lifetime illnesses. Additionally, sophisticated clinical decision support systems incorporate predictive analytics to support medical decision making at the point of care. A working definition has been proposed by Robert Hayward of the Centre for Health Evidence: "Clinical Decision Support Systems link health observations with health knowledge to influence health choices by clinicians for improved health care."

3.3 Collection analytics

Many portfolios have a set of delinquent customers who do not make their payments on time. The financial institution has to undertake collection activities on these customers to recover the amounts due. A lot of collection resources are wasted on customers who are difficult or impossible to recover. Predictive analytics can help optimize the allocation of collection resources by identifying the most effective collection agencies, contact strategies, legal actions and other strategies to each customer, thus significantly increasing recovery at the same time reducing collection costs.

3.4 Cross-sell

Often corporate organizations collect and maintain abundant data (e.g. customer records, sale transactions) as exploiting hidden relationships in the data can provide a competitive advantage. For an organization that offers multiple products, predictive analytics can help analyze customers' spending, usage and other behavior, leading to

efficient cross sales, or selling additional products to current customers.[2] This directly leads to higher profitability per customer and stronger customer relationships.

3.5 Customer retention

With the number of competing services available, businesses need to focus efforts on maintaining continuous consumer satisfaction, rewarding consumer loyalty and minimizing customer attrition. In addition, small increases in customer retention have been shown to increase profits disproportionately. One study concluded that a 5% increase in customer retention rates will increase profits by 25% to 95%.[15] Businesses tend to respond to customer attrition on a reactive basis, acting only after the customer has initiated the process to terminate service. At this stage, the chance of changing the customer's decision is almost impossible. Proper application of predictive analytics can lead to a more proactive retention strategy. By a frequent examination of a customer's past service usage, service performance, spending and other behavior patterns, predictive models can determine the likelihood of a customer terminating service sometime soon.[7] An intervention with lucrative offers can increase the chance of retaining the customer. Silent attrition, the behavior of a customer to slowly but steadily reduce usage, is another problem that many companies face. Predictive analytics can also predict this behavior, so that the company can take proper actions to increase customer activity.

3.6 Direct marketing

When marketing consumer products and services, there is the challenge of keeping up with competing products and consumer behavior. Apart from identifying prospects, predictive analytics can also help to identify the most effective combination of product versions, marketing material, communication channels and timing that should be used to target a given consumer. The goal of predictive analytics is typically to lower the cost per order or cost per action.

3.7 Fraud detection

Fraud is a big problem for many businesses and can be of various types: inaccurate credit applications, fraudulent transactions (both offline and online), identity thefts and false insurance claims. These problems plague firms of all sizes in many industries. Some examples of likely victims are credit card issuers, insurance companies,[16] retail merchants, manufacturers, business-to-business suppliers and even services providers. A predictive model can help weed out the "bads" and reduce a business's exposure to fraud.

Predictive modeling can also be used to identify high-risk fraud candidates in business or the public sector. Mark Nigrini developed a risk-scoring method to identify audit targets. He describes the use of this approach to detect fraud in the franchisee sales reports of an international fast-food chain. Each location is scored using 10 predictors. The 10 scores are then weighted to give one final overall risk score for each location. The same scoring approach was also used to identify high-risk check kiting accounts, potentially fraudulent travel agents, and questionable vendors. A reasonably complex model was used to identify fraudulent monthly reports submitted by divisional controllers.[17]

The Internal Revenue Service (IRS) of the United States also uses predictive analytics to mine tax returns and identify tax fraud.[16]

Recent advancements in technology have also introduced predictive behavior analysis for web fraud detection. This type of solution utilizes heuristics in order to study normal web user behavior and detect anomalies indicating fraud attempts.

3.8 Portfolio, product or economy-level prediction

Often the focus of analysis is not the consumer but the product, portfolio, firm, industry or even the economy. For example, a retailer might be interested in predicting store-level demand for inventory management purposes. Or the Federal Reserve Board might be interested in predicting the unemployment rate for the next year. These types of problems can be addressed by predictive analytics using time series techniques (see below). They can also be addressed via machine learning approaches which transform the original time series into a feature vector space, where the learning algorithm finds patterns that have predictive power.[18][19]

3.9 Risk management

When employing risk management techniques, the results are always to predict and benefit from a future scenario. The Capital asset pricing model (CAP-M) "predicts" the best portfolio to maximize return, Probabilistic Risk Assessment (PRA)--when combined with mini-Delphi Techniques and statistical approaches yields accurate forecasts and RiskAoA is a stand-alone predictive tool.[20] These are three examples of approaches that can extend from project to market, and from near to long term. Underwriting (see below) and other business approaches identify risk management as a predictive method.

3.10 Underwriting

Many businesses have to account for risk exposure due to their different services and determine the cost needed to cover the risk. For example, auto insurance providers need to accurately determine the amount of premium to charge to cover each automobile and driver. A financial company needs to assess a borrower's potential and ability to pay before granting a loan. For a health insurance provider, predictive analytics can analyze a few years of past medical claims data, as well as lab, pharmacy and other records where available, to predict how expensive an enrollee is likely to be in the future. Predictive analytics can help underwrite these quantities by predicting the chances of illness, default, bankruptcy, etc. Predictive analytics can streamline the process of customer acquisition by predicting the future risk behavior of a customer using application level data.[4] Predictive analytics in the form of credit scores have reduced the amount of time it takes for loan approvals, especially in the mortgage market where lending decisions are now made in a matter of hours rather than days or even weeks. Proper predictive analytics can lead to proper pricing decisions, which can help mitigate future risk of default.

4 Technology and big data influences

Big data is a collection of data sets that are so large and complex that they become awkward to work with using traditional database management tools. The volume, variety and velocity of big data have introduced challenges across the board for capture, storage, search, sharing, analysis, and visualization. Examples of big data sources include web logs, RFID, sensor data, social networks, Internet search indexing, call detail records, military surveillance, and complex data in astronomic, biogeochemical, genomics, and atmospheric sciences. Big Data is the core of most predictive analytic services offered by IT organizations.[21] Thanks to technological advances in computer hardware — faster CPUs, cheaper memory, and MPP architectures — and new technologies such as Hadoop, MapReduce, and in-database and text analytics for processing big data, it is now feasible to collect, analyze, and mine massive amounts of structured and unstructured data for new insights.[16] Today, exploring big data and using predictive analytics is within reach of more organizations than ever before and new methods that are capable for handling such datasets are proposed [22] [23]

5 Analytical Techniques

The approaches and techniques used to conduct predictive analytics can broadly be grouped into regression techniques and machine learning techniques.

5.1 Regression techniques

Regression models are the mainstay of predictive analytics. The focus lies on establishing a mathematical equation as a model to represent the interactions between the different variables in consideration. Depending on the situation, there are a wide variety of models that can be applied while performing predictive analytics. Some of them are briefly discussed below.

5.1.1 Linear regression model

The linear regression model analyzes the relationship between the response or dependent variable and a set of independent or predictor variables. This relationship is expressed as an equation that predicts the response variable as a linear function of the parameters. These parameters are adjusted so that a measure of fit is optimized. Much of the effort in model fitting is focused on minimizing the size of the residual, as well as ensuring that it is randomly distributed with respect to the model predictions.

The goal of regression is to select the parameters of the model so as to minimize the sum of the squared residuals. This is referred to as **ordinary least squares** (OLS) estimation and results in best linear unbiased estimates (BLUE) of the parameters if and only if the Gauss-Markov assumptions are satisfied.

Once the model has been estimated we would be interested to know if the predictor variables belong in the model – i.e. is the estimate of each variable's contribution reliable? To do this we can check the statistical significance of the model's coefficients which can be measured using the t-statistic. This amounts to testing whether the coefficient is significantly different from zero. How well the model predicts the dependent variable based on the value of the independent variables can be assessed by using the R^2 statistic. It measures predictive power of the model i.e. the proportion of the total variation in the dependent variable that is "explained" (accounted for) by variation in the independent variables.

5.1.2 Discrete choice models

Multivariate regression (above) is generally used when the response variable is continuous and has an unbounded range. Often the response variable may not be continuous but rather discrete. While mathematically it is feasible to apply multivariate regression to discrete ordered dependent variables, some of the assumptions behind the theory of multivariate linear regression no longer hold, and there are other techniques such as discrete choice models which are better suited for this type of analysis. If the dependent variable is discrete, some of those superior methods are

logistic regression, multinomial logit and probit models. Logistic regression and probit models are used when the dependent variable is binary.

5.1.3 Logistic regression

For more details on this topic, see logistic regression.

In a classification setting, assigning outcome probabilities to observations can be achieved through the use of a logistic model, which is basically a method which transforms information about the binary dependent variable into an unbounded continuous variable and estimates a regular multivariate model (See Allison's Logistic Regression for more information on the theory of Logistic Regression).

The Wald and likelihood-ratio test are used to test the statistical significance of each coefficient b in the model (analogous to the t tests used in OLS regression; see above). A test assessing the goodness-of-fit of a classification model is the "percentage correctly predicted".

5.1.4 Multinomial logistic regression

An extension of the binary logit model to cases where the dependent variable has more than 2 categories is the multinomial logit model. In such cases collapsing the data into two categories might not make good sense or may lead to loss in the richness of the data. The multinomial logit model is the appropriate technique in these cases, especially when the dependent variable categories are not ordered (for examples colors like red, blue, green). Some authors have extended multinomial regression to include feature selection/importance methods such as Random multinomial logit.

5.1.5 Probit regression

Probit models offer an alternative to logistic regression for modeling categorical dependent variables. Even though the outcomes tend to be similar, the underlying distributions are different. Probit models are popular in social sciences like economics.

A good way to understand the key difference between probit and logit models is to assume that there is a latent variable z.

We do not observe z but instead observe y which takes the value 0 or 1. In the logit model we assume that y follows a logistic distribution. In the probit model we assume that y follows a standard normal distribution. Note that in social sciences (e.g. economics), probit is often used to model situations where the observed variable y is continuous but takes values between 0 and 1.

5.1.6 Logit versus probit

The Probit model has been around longer than the logit model. They behave similarly, except that the logistic distribution tends to be slightly flatter tailed. One of the reasons the logit model was formulated was that the probit model was computationally difficult due to the requirement of numerically calculating integrals. Modern computing however has made this computation fairly simple. The coefficients obtained from the logit and probit model are fairly close. However, the odds ratio is easier to interpret in the logit model.

Practical reasons for choosing the probit model over the logistic model would be:

- There is a strong belief that the underlying distribution is normal
- The actual event is not a binary outcome (*e.g.*, bankruptcy status) but a proportion (*e.g.*, proportion of population at different debt levels).

5.1.7 Time series models

Time series models are used for predicting or forecasting the future behavior of variables. These models account for the fact that data points taken over time may have an internal structure (such as autocorrelation, trend or seasonal variation) that should be accounted for. As a result, standard regression techniques cannot be applied to time series data and methodology has been developed to decompose the trend, seasonal and cyclical component of the series. Modeling the dynamic path of a variable can improve forecasts since the predictable component of the series can be projected into the future.

Time series models estimate difference equations containing stochastic components. Two commonly used forms of these models are autoregressive models (AR) and moving-average (MA) models. The Box–Jenkins methodology (1976) developed by George Box and G.M. Jenkins combines the AR and MA models to produce the ARMA (autoregressive moving average) model which is the cornerstone of stationary time series analysis. ARIMA (autoregressive integrated moving average models) on the other hand are used to describe non-stationary time series. Box and Jenkins suggest differencing a non stationary time series to obtain a stationary series to which an ARMA model can be applied. Non stationary time series have a pronounced trend and do not have a constant long-run mean or variance.

Box and Jenkins proposed a three-stage methodology which includes: model identification, estimation and validation. The identification stage involves identifying if the series is stationary or not and the presence of seasonality by examining plots of the series, autocorrelation and partial autocorrelation functions. In the estimation stage, models are estimated using non-linear time series

or maximum likelihood estimation procedures. Finally the validation stage involves diagnostic checking such as plotting the residuals to detect outliers and evidence of model fit.

In recent years time series models have become more sophisticated and attempt to model conditional heteroskedasticity with models such as ARCH (autoregressive conditional heteroskedasticity) and GARCH (generalized autoregressive conditional heteroskedasticity) models frequently used for financial time series. In addition time series models are also used to understand inter-relationships among economic variables represented by systems of equations using VAR (vector autoregression) and structural VAR models.

5.1.8 Survival or duration analysis

Survival analysis is another name for time to event analysis. These techniques were primarily developed in the medical and biological sciences, but they are also widely used in the social sciences like economics, as well as in engineering (reliability and failure time analysis).

Censoring and non-normality, which are characteristic of survival data, generate difficulty when trying to analyze the data using conventional statistical models such as multiple linear regression. The normal distribution, being a symmetric distribution, takes positive as well as negative values, but duration by its very nature cannot be negative and therefore normality cannot be assumed when dealing with duration/survival data. Hence the normality assumption of regression models is violated.

The assumption is that if the data were not censored it would be representative of the population of interest. In survival analysis, censored observations arise whenever the dependent variable of interest represents the time to a terminal event, and the duration of the study is limited in time.

An important concept in survival analysis is the hazard rate, defined as the probability that the event will occur at time t conditional on surviving until time t. Another concept related to the hazard rate is the survival function which can be defined as the probability of surviving to time t.

Most models try to model the hazard rate by choosing the underlying distribution depending on the shape of the hazard function. A distribution whose hazard function slopes upward is said to have positive duration dependence, a decreasing hazard shows negative duration dependence whereas constant hazard is a process with no memory usually characterized by the exponential distribution. Some of the distributional choices in survival models are: F, gamma, Weibull, log normal, inverse normal, exponential etc. All these distributions are for a non-negative random variable.

Duration models can be parametric, non-parametric or semi-parametric. Some of the models commonly used are Kaplan-Meier and Cox proportional hazard model (non parametric).

5.1.9 Classification and regression trees

Main article: decision tree learning

Globally-optimal classification tree analysis (GO-CTA) (also called hierarchical optimal discriminant analysis) is a generalization of optimal discriminant analysis that may be used to identify the statistical model that has maximum accuracy for predicting the value of a categorical dependent variable for a dataset consisting of categorical and continuous variables. The output of HODA is a non-orthogonal tree that combines categorical variables and cut points for continuous variables that yields maximum predictive accuracy, an assessment of the exact Type I error rate, and an evaluation of potential cross-generalizability of the statistical model. Hierarchical optimal discriminant analysis may be thought of as a generalization of Fisher's linear discriminant analysis. Optimal discriminant analysis is an alternative to ANOVA (analysis of variance) and regression analysis, which attempt to express one dependent variable as a linear combination of other features or measurements. However, ANOVA and regression analysis give a dependent variable that is a numerical variable, while hierarchical optimal discriminant analysis gives a dependent variable that is a class variable.

Classification and regression trees (CART) are a non-parametric decision tree learning technique that produces either classification or regression trees, depending on whether the dependent variable is categorical or numeric, respectively.

Decision trees are formed by a collection of rules based on variables in the modeling data set:

- Rules based on variables' values are selected to get the best split to differentiate observations based on the dependent variable
- Once a rule is selected and splits a node into two, the same process is applied to each "child" node (i.e. it is a recursive procedure)
- Splitting stops when CART detects no further gain can be made, or some pre-set stopping rules are met. (Alternatively, the data are split as much as possible and then the tree is later pruned.)

Each branch of the tree ends in a terminal node. Each observation falls into one and exactly one terminal node, and each terminal node is uniquely defined by a set of rules.

A very popular method for predictive analytics is Leo Breiman's Random forests.

5.1.10 Multivariate adaptive regression splines

Multivariate adaptive regression splines (MARS) is a non-parametric technique that builds flexible models by fitting piecewise linear regressions.

An important concept associated with regression splines is that of a knot. Knot is where one local regression model gives way to another and thus is the point of intersection between two splines.

In multivariate and adaptive regression splines, basis functions are the tool used for generalizing the search for knots. Basis functions are a set of functions used to represent the information contained in one or more variables. Multivariate and Adaptive Regression Splines model almost always creates the basis functions in pairs.

Multivariate and adaptive regression spline approach deliberately overfits the model and then prunes to get to the optimal model. The algorithm is computationally very intensive and in practice we are required to specify an upper limit on the number of basis functions.

5.2 Machine learning techniques

Machine learning, a branch of artificial intelligence, was originally employed to develop techniques to enable computers to learn. Today, since it includes a number of advanced statistical methods for regression and classification, it finds application in a wide variety of fields including medical diagnostics, credit card fraud detection, face and speech recognition and analysis of the stock market. In certain applications it is sufficient to directly predict the dependent variable without focusing on the underlying relationships between variables. In other cases, the underlying relationships can be very complex and the mathematical form of the dependencies unknown. For such cases, machine learning techniques emulate human cognition and learn from training examples to predict future events.

A brief discussion of some of these methods used commonly for predictive analytics is provided below. A detailed study of machine learning can be found in Mitchell (1997).

5.2.1 Neural networks

Neural networks are nonlinear sophisticated modeling techniques that are able to model complex functions. They can be applied to problems of prediction, classification or control in a wide spectrum of fields such as finance, cognitive psychology/neuroscience, medicine, engineering, and physics.

Neural networks are used when the exact nature of the relationship between inputs and output is not known. A key feature of neural networks is that they learn the relationship between inputs and output through training. There are three types of training in neural networks used by different networks, supervised and unsupervised training, reinforcement learning, with supervised being the most common one.

Some examples of neural network training techniques are backpropagation, quick propagation, conjugate gradient descent, projection operator, Delta-Bar-Delta etc. Some unsupervised network architectures are multilayer perceptrons, Kohonen networks, Hopfield networks, etc.

5.2.2 Multilayer Perceptron (MLP)

The Multilayer Perceptron (MLP) consists of an input and an output layer with one or more hidden layers of nonlinearly-activating nodes or sigmoid nodes. This is determined by the weight vector and it is necessary to adjust the weights of the network. The backpropagation employs gradient fall to minimize the squared error between the network output values and desired values for those outputs. The weights adjusted by an iterative process of repetitive present of attributes. Small changes in the weight to get the desired values are done by the process called training the net and is done by the training set (learning rule).

5.2.3 Radial basis functions

A radial basis function (RBF) is a function which has built into it a distance criterion with respect to a center. Such functions can be used very efficiently for interpolation and for smoothing of data. Radial basis functions have been applied in the area of neural networks where they are used as a replacement for the sigmoidal transfer function. Such networks have 3 layers, the input layer, the hidden layer with the RBF non-linearity and a linear output layer. The most popular choice for the non-linearity is the Gaussian. RBF networks have the advantage of not being locked into local minima as do the feed-forward networks such as the multilayer perceptron.

5.2.4 Support vector machines

Support Vector Machines (SVM) are used to detect and exploit complex patterns in data by clustering, classifying and ranking the data. They are learning machines that are used to perform binary classifications and regression estimations. They commonly use kernel based methods to apply linear classification techniques to non-linear classification problems. There are a number of types of SVM such as linear, polynomial, sigmoid etc.

5.2.5 Naïve Bayes

Naïve Bayes based on Bayes conditional probability rule is used for performing classification tasks. Naïve Bayes

assumes the predictors are statistically independent which makes it an effective classification tool that is easy to interpret. It is best employed when faced with the problem of 'curse of dimensionality' i.e. when the number of predictors is very high.

5.2.6 *k*-nearest neighbours

The nearest neighbour algorithm (KNN) belongs to the class of pattern recognition statistical methods. The method does not impose a priori any assumptions about the distribution from which the modeling sample is drawn. It involves a training set with both positive and negative values. A new sample is classified by calculating the distance to the nearest neighbouring training case. The sign of that point will determine the classification of the sample. In the k-nearest neighbour classifier, the k nearest points are considered and the sign of the majority is used to classify the sample. The performance of the kNN algorithm is influenced by three main factors: (1) the distance measure used to locate the nearest neighbours; (2) the decision rule used to derive a classification from the k-nearest neighbours; and (3) the number of neighbours used to classify the new sample. It can be proved that, unlike other methods, this method is universally asymptotically convergent, i.e.: as the size of the training set increases, if the observations are independent and identically distributed (i.i.d.), regardless of the distribution from which the sample is drawn, the predicted class will converge to the class assignment that minimizes misclassification error. See Devroy et al.

5.2.7 Geospatial predictive modeling

Conceptually, geospatial predictive modeling is rooted in the principle that the occurrences of events being modeled are limited in distribution. Occurrences of events are neither uniform nor random in distribution – there are spatial environment factors (infrastructure, sociocultural, topographic, etc.) that constrain and influence where the locations of events occur. Geospatial predictive modeling attempts to describe those constraints and influences by spatially correlating occurrences of historical geospatial locations with environmental factors that represent those constraints and influences. Geospatial predictive modeling is a process for analyzing events through a geographic filter in order to make statements of likelihood for event occurrence or emergence.

6 Tools

Historically, using predictive analytics tools—as well as understanding the results they delivered—required advanced skills. However, modern predictive analytics tools are no longer restricted to IT specialists. As more organizations adopt predictive analytics into decision-making processes and integrate it into their operations, they are creating a shift in the market toward business users as the primary consumers of the information. Business users want tools they can use on their own. Vendors are responding by creating new software that removes the mathematical complexity, provides user-friendly graphic interfaces and/or builds in short cuts that can, for example, recognize the kind of data available and suggest an appropriate predictive model.[24] Predictive analytics tools have become sophisticated enough to adequately present and dissect data problems, so that any data-savvy information worker can utilize them to analyze data and retrieve meaningful, useful results.[2] For example, modern tools present findings using simple charts, graphs, and scores that indicate the likelihood of possible outcomes.[25]

There are numerous tools available in the marketplace that help with the execution of predictive analytics. These range from those that need very little user sophistication to those that are designed for the expert practitioner. The difference between these tools is often in the level of customization and heavy data lifting allowed.

Notable open source predictive analytic tools include:

- scikit-learn
- KNIME
- OpenNN
- Orange
- R
- Weka
- GNU Octave
- Apache Mahout

Notable commercial predictive analytic tools include:

- Alpine Data Labs
- BIRT Analytics
- Axibase Time Series Database
- Angoss KnowledgeSTUDIO
- IBM SPSS Statistics and IBM SPSS Modeler
- KXEN Modeler
- Mathematica
- MATLAB
- Minitab
- Neural Designer
- Oracle Advanced Analytics

- Pervasive
- Predikto
- Predixion Software
- RapidMiner
- RCASE
- Revolution Analytics
- SAP
- SAS and SAS Enterprise Miner
- STATA
- Statgraphics
- STATISTICA
- TeleRetail
- TIBCO

The most popular commercial predictive analytics software packages according to the Rexer Analytics Survey for 2013 are IBM SPSS Modeler, SAS Enterprise Miner, and Dell Statistica <http://www.rexeranalytics.com/Data-Miner-Survey-2013-Intro.html>

6.1 PMML

In an attempt to provide a standard language for expressing predictive models, the Predictive Model Markup Language (PMML) has been proposed. Such an XML-based language provides a way for the different tools to define predictive models and to share these between PMML compliant applications. PMML 4.0 was released in June, 2009.

7 Criticism

There are plenty of skeptics when it comes to computers and algorithms abilities to predict the future, including Gary King, a professor from Harvard University and the director of the Institute for Quantitative Social Science. [26] People are influenced by their environment in innumerable ways. Trying to understand what people will do next assumes that all the influential variables can be known and measured accurately. "People's environments change even more quickly than they themselves do. Everything from the weather to their relationship with their mother can change the way people think and act. All of those variables are unpredictable. How they will impact a person is even less predictable. If put in the exact same situation tomorrow, they may make a completely different decision. This means that a statistical prediction is only valid in sterile laboratory conditions, which suddenly isn't as useful as it seemed before." [27]

8 See also

- Criminal Reduction Utilising Statistical History
- Data mining
- Learning analytics
- Odds algorithm
- Pattern recognition
- Prescriptive analytics
- Predictive modeling
- RiskAoA a predictive tool for discriminating future decisions.

9 References

[1] Nyce, Charles (2007), *Predictive Analytics White Paper* (PDF), American Institute for Chartered Property Casualty Underwriters/Insurance Institute of America, p. 1

[2] Eckerson, Wayne (May 10, 2007), *Extending the Value of Your Data Warehousing Investment*, The Data Warehouse Institute

[3] Coker, Frank (2014). *Pulse: Understanding the Vital Signs of Your Business* (1st ed.). Bellevue, WA: Ambient Light Publishing. pp. 30, 39, 42,more. ISBN 978-0-9893086-0-1.

[4] Conz, Nathan (September 2, 2008), "Insurers Shift to Customer-focused Predictive Analytics Technologies", *Insurance & Technology*

[5] Fletcher, Heather (March 2, 2011), "The 7 Best Uses for Predictive Analytics in Multichannel Marketing", *Target Marketing*

[6] Korn, Sue (April 21, 2011), "The Opportunity for Predictive Analytics in Finance", *HPC Wire*

[7] Barkin, Eric (May 2011), "CRM + Predictive Analytics: Why It All Adds Up", *Destination CRM*

[8] Das, Krantik; Vidyashankar, G.S. (July 1, 2006), "Competitive Advantage in Retail Through Analytics: Developing Insights, Creating Value", *Information Management*

[9] McDonald, Michèle (September 2, 2010), "New Technology Taps 'Predictive Analytics' to Target Travel Recommendations", *Travel Market Report*

[10] Stevenson, Erin (December 16, 2011), "Tech Beat: Can you pronounce health care predictive analytics?", *Times-Standard*

[11] McKay, Lauren (August 2009), "The New Prescription for Pharma", *Destination CRM*

[12] Finlay, Steven (2014). *Predictive Analytics, Data Mining and Big Data. Myths, Misconceptions and Methods* (1st ed.). Basingstoke: Palgrave Macmillan. p. 237. ISBN 1137379278.

[13] Siegel, Eric (2013). *Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die* (1st ed.). Wiley. ISBN 978-1-1183-5685-2.

[14] "Web Application Framework" (PDF). H Karlapaudi ,J Martin.

[15] Reichheld, Frederick; Schefter, Phil. "The Economics of E-Loyalty". *Havard Business School Working Knowledge*. Retrieved 10 November 2014.

[16] Schiff, Mike (March 6, 2012), *BI Experts: Why Predictive Analytics Will Continue to Grow*, The Data Warehouse Institute

[17] Nigrini, Mark (June 2011). "Forensic Analytics: Methods and Techniques for Forensic Accounting Investigations". Hoboken, NJ: John Wiley & Sons Inc. ISBN 978-0-470-89046-2.

[18] Dhar, Vasant (April 2011). "Prediction in Financial Markets: The Case for Small Disjuncts". *ACM Transactions on Intelligent Systems and Technologies* **2** (3).

[19] Dhar, Vasant; Chou, Dashin; Provost Foster (October 2000). "Discovering Interesting Patterns in Investment Decision Making with GLOWER – A Genetic Learning Algorithm Overlaid With Entropy Reduction". *Data Mining and Knowledge Discovery* **4** (4).

[20] https://acc.dau.mil/CommunityBrowser.aspx?id=126070

[21] http://www.hcltech.com/sites/default/files/key_to_monetizing_big_data_via_predictive_analytics.pdf

[22] Ben-Gal I. Dana A., Shkolnik N. and Singer (2014). "Efficient Construction of Decision Trees by the Dual Information Distance Method" (PDF). Quality Technology & Quantitative Management (QTQM), 11(1), 133-147.

[23] Ben-Gal I., Shavitt Y., Weinsberg E., Weinsberg U. (2014). "Peer-to-peer information retrieval using shared-content clustering" (PDF). Knowl Inf Syst. doi:10.1007/s10115-013-0619-9.

[24] Halper, Fern (November 1, 2011), "The Top 5 Trends in Predictive Analytics", *Information Management*

[25] MacLennan, Jamie (May 1, 2012), *5 Myths about Predictive Analytics*, The Data Warehouse Institute

[26] Temple-Raston, Dina (Oct 8, 2012), *Predicting The Future: Fantasy Or A Good Algorithm?*, NPR

[27] Alverson, Cameron (Sep 2012), *Polling and Statistical Models Can't Predict the Future*, Cameron Alverson

10 Further reading

- Agresti, Alan (2002). *Categorical Data Analysis*. Hoboken: John Wiley and Sons. ISBN 0-471-36093-7.
- Coggeshall, Stephen, Davies, John, Jones, Roger., and Schutzer, Daniel, "Intelligent Security Systems," in Freedman, Roy S., Flein, Robert A., and Lederman, Jess, Editors (1995). *Artificial Intelligence in the Capital Markets*. Chicago: Irwin. ISBN 1-55738-811-3.
- L. Devroye, L. Györfi, G. Lugosi (1996). *A Probabilistic Theory of Pattern Recognition*. New York: Springer-Verlag.
- Enders, Walter (2004). *Applied Time Series Econometrics*. Hoboken: John Wiley and Sons. ISBN 0-521-83919-X.
- Greene, William (2012). *Econometric Analysis, 7th Ed*. London: Prentice Hall. ISBN 978-0-13-139538-1.
- Guidère, Mathieu; Howard N, Sh. Argamon (2009). *Rich Language Analysis for Counterterrrorism*. Berlin, London, New York: Springer-Verlag. ISBN 978-3-642-01140-5.
- Mitchell, Tom (1997). *Machine Learning*. New York: McGraw-Hill. ISBN 0-07-042807-7.
- Siegel, Eric (2013). *Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die*. John Wiley. ISBN 978-1-1183-5685-2.
- Tukey, John (1977). *Exploratory Data Analysis*. New York: Addison-Wesley. ISBN 0-201-07616-0.
- Finlay, Steven (2014). *Predictive Analytics, Data Mining and Big Data. Myths, Misconceptions and Methods*. Basingstoke: Palgrave Macmillan. ISBN 978-1-137-37927-6.
- Coker, Frank (2014). *Pulse: Understanding the Vital Signs of Your Business*. Bellevue, WA: Ambient Light Publishing. ISBN 978-0-9893086-0-1.

11 Text and image sources, contributors, and licenses

11.1 Text

- **Predictive analytics** *Source:* https://en.wikipedia.org/wiki/Predictive_analytics?oldid=694377046 *Contributors:* The Anome, Edward, Michael Hardy, Kku, Karada, Ronz, Avani~enwiki, Giftlite, SWAdair, Thorwald, Bender235, Rubicon, Calair, Rajah, Mdd, Andrewpmk, Stephen Turner, Dominic, Oleg Alexandrov, OwenX, Qwertyus, HughJorgan, RussBot, Leighblackall, Aeusoes1, Gloumouth1, DeadEyeArrow, EAderhold, Zzuuzz, Talyian, Allens, Zvika, Yvwv, SmackBot, CommodiCast, Mcld, Gilliam, EncMstr, Eudaemonic3, Colonies Chris, S2rikanth, Onorem, JonHarder, Krexer, Lpgeffen, Doug Bell, Kuru, IronGargoyle, JHunterJ, Ralf Klinkenberg, CmdrObot, Van helsing, Requestion, Pgr94, Myasuda, LeoHeska, Dancter, Talgalili, Scientio, Marek69, JustAGal, Batra, Mr. Darcy, AntiVandalBot, MER-C, Fbooth, VoABot II, Baccyak4H, Bellemichelle, Sweet2, Gregheth, Apdevries, Sudheervaishnav, Ekotkie, Dontdoit, Jfroelich, Trusilver, Dvdpwiki, Ramkumar.krishnan, Atama, Bonadea, BernardZ, Cyricx, GuyRo, Vishal0soni, Deanabb, Dherman652, Arpabr, Selain03, Hherbert, Cuttysc, Vikx1, Ralftgehrig, Rpm698, MaynardClark, Drakedirect, Chrisguyot, Melcombe, Maralia, Into The Fray, Kai-Hendrik, Ahyeek, Howie Goodell, Sterdeus, SpikeToronto, Stephen Milborrow, Jlamro, Isthisthingworking, SchreiberBike, Bateni, Cookiehead, Qwfp, Angoss, Sunsetsky, Vianello, MystBot, Vaheterdu, BizAnalyst, Addbot, Fgnievinski, MrOllie, Download, Yobot, SOMart, Sorgina.13, AnomieBOT, IRP, Nosperantos, Citation bot, Jtamad, BlaineKohl, Phy31277, CorporateM, GESICC, FrescoBot, Boxplot, I dream of horses, Triplestop, Dmitry St, SpaceFlight89, Jackverr, Σ, Peter.borissow, Ethansdad, Trappist the monk, Cambridgeblue1971, Pamparam, Kmettler, Glenn Maddox, Vrenator, Crysb, DARTH SIDIOUS 2, Onel5969, WikitanvirBot, Sugarfoot1001, Josve05a, Jssgator, Chire, MainFrame, Synecticsgroup, ClueBot NG, Thirdharringtonskier, Stefanomazzalai, Ricekido, Widr, Luke145, Mikeono, Helpful Pixie Bot, WhartonCAI, JonasJSchreiber, Wbm1058, BG19bot, Rafaelgmonteiro, Lisasolomonsalford, TLAN38, Lynnlangit, Flaticida, MrBill3, MC Wapiti, HHinman, BattyBot, Raspabill, Jeremy Kolb, TwoTwoHello, Andrux, Mkhambaty, Cmdima, HeatherMKCampbell, Tommycarney, Tentinator, Sgolestanian, Gbtodd29, Brishtikonna, Thisaccountisbs, Mitchki.nj, Jvn mht, JaconaFrere, Pablodim91, Cbuyle, Monkbot, AmandaJohnson2014, JSHorwitz, Thomas Speidel, HappyVDZ, Wikiperson99, Justincahoon, Femiolajiga, Stevenfinlay, Rlm1188, Bildn, Nivedita1414, Frankecoker, Annaelison, Olosko, Vedanga Kumar, Gary2015, HelpUsStopSpam, Heinrichvk, Rodionos, Olavlaudy, Mullingar89, Brunod7374, Jsmayank, NeilP777 and Anonymous: 225

11.2 Images

- **File:Question_book-new.svg** *Source:* https://upload.wikimedia.org/wikipedia/en/9/99/Question_book-new.svg *License:* Cc-by-sa-3.0 *Contributors:*

  Created from scratch in Adobe Illustrator. Based on Image:Question book.png created by User:Equazcion *Original artist:*

  Tkgd2007
- **File:Text_document_with_red_question_mark.svg** *Source:* https://upload.wikimedia.org/wikipedia/commons/a/a4/Text_document_with_red_question_mark.svg *License:* Public domain *Contributors:* Created by bdesham with Inkscape; based upon Text-x-generic.svg from the Tango project. *Original artist:* Benjamin D. Esham (bdesham)

11.3 Content license

- Creative Commons Attribution-Share Alike 3.0