insideBIGDATA

InsideBIGDATA Guide to

Predictive Analytics

by Daniel D. Gutierrez

BROUGHT TO YOU BY TIBCO Spotfire®

Predictive Analytics Defined

Predictive analytics, sometimes called advanced analytics, is a term used to describe a range of analytical and statistical techniques to predict future actions or behaviors. In business, predictive analytics are used to make proactive decisions and determine actions, by using statistical models to discover patterns in historical and transactional data to uncover likely risks and opportunities.

Predictive analytics incorporates a range of activities which we will explore in this paper, including data access, exploratory data analysis and visualization, developing assumptions and data models, applying predictive models, then estimating and/or predicting future outcomes.

Contents

The History of Predictive Analytics

Modern day predictive analytics has its origin in the 1940s, when governments started using the first computational models — Monte Carlo simulations, computational models for neural networks, and linear programing — for decoding German messages in WWII, automating targeting of anti-aircraft weapons against enemy planes, and computer simulations to predict behavior of nuclear chain reactions for the Manhattan Project. In the 1960s, corporations and research institutions began the era of commercializing analytics with non-linear programming, and computer-based heuristic problem solving — for the first models to forecast the weather, solving the "shortest path problem" to improve air travel and logistics, and applying predictive modeling to credit risk decisions. Then in the 1970s – 1990s analytics was used more broadly in organizations, and tech start-ups made real-time and prescriptive analytics a reality. However, predictive analytics mainly remained in the hands of corporate statisticians, brought to the business only in static, batch-driven reports.

Today predictive analytics has finally arrived into the corporate mainstream, being used by everyday business users for a broadening set of use cases. This growing phenomenon has been driven by the realities of a global economy, i.e., organizations continually looking for competitive advantage, and enabled by strong technological innovations.

These technological innovations include more scalable computing power, relational databases, new Big Data technologies, such as Hadoop, and self-service analytics software that puts data, and predictive models in the hands of front-line decision makers. All of this has allowed organizations to compete based on analytic innovation.

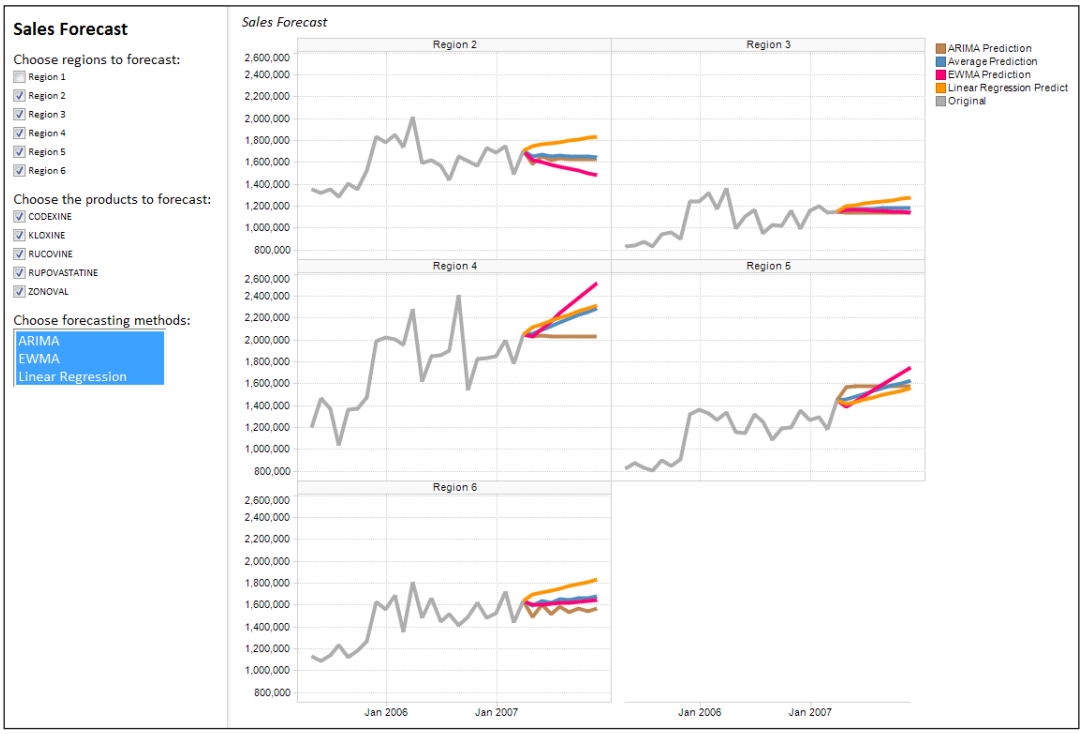
First, organizations embraced simple data discovery analytics to understand the state of their business, and to look deeply in the data to understand the "whys" behind the data. But as they become more comfortable with the data, they see the opportunity to further outperform their competitors using advanced analytics.

Business Uses of Predictive Analytics

The need for predictive analytics in the enterprise is clear, as it can provide smarter analysis for better decision making, increased market competiveness, a direct path in taking advantage of market opportunity and threats, a way to reduce uncertainty and manage risk, an approach to proactively plan and act, discovery of meaningful patterns, and the means to anticipate and react to emerging trends.

Advanced quantitative analysis has demonstrated benefits in a wide cross section of industries and domains. Many classes of business problems can be solved with predictive analytics, here are just a few:

- **Sales forecasting –** predict what you should expect to book this month or this quarter, taking into account your historical conversion rates, i.e. what your sales team's winning percentage on similar opportunities has been like in the past, coupled with your current sales pipeline, i.e. the number of opportunities your team is working on within this time window.
- **Fraud detection –** find inaccurate credit applications, fraudulent transactions both done offline and online, identity thefts, false insurance claims, etc.
- **Retail campaign optimization –** allows marketers to model outcomes of campaigns based on a deep analysis of customer behaviors, preferences, and profile data.
- **Marketing & customer analytics –** collect data from digital marketing, social media, call centers, mobile apps, etc. and use the information on what customers have done in the past to gauge what they may do in the future.
- **HR analytics –** enable organizations to analyze the past and look forward to spot trends in key factors related to voluntary termination, absences and other sources of risk, as well as identifying trends in required skill sets vs. current resources.
- **Risk management –** predict the best portfolio to maximize return in capital asset pricing model and probabilistic risk assessment to yield accurate forecasts.



Predictive analytics for sales forecasting provides targeted, relevant predictive analytics to a broad spectrum of business users to improve decision making. Image courtesy of TIBCO Spotfire.

Classes of Predictive Analytics

At its core, predictive analytics relies on capturing relationships between past data points, and using those relationships to predict future outcomes. In order to make predictions based on a given data set, one or more predictor variables are used to predict a response variable. In its simplest form, predictive analytics assists with developing forecasts for business decision making. To handle more complex requirements, advanced predictive analytics techniques are applied to drive critical business processes. In this section we will provide a high-level view of the primary classes of predictive analytics: supervised learning and unsupervised learning.

Regression is the most common form of predictive analytics. With regression, there is a quantitative response variable (what you're trying to predict).

Supervised learning is divided into two broad categories: regression for responses that are quantitative (a numeric value), such as miles per gallon for a particular car, and classification for responses that can have just a few known values, such as 'true' or 'false'.

- **Regression –** Regression is the most common form of predictive analytics. With regression, there is a quantitative response variable (what you're trying to predict) like the sale price of a home, based on a series of predictor variables such as the number of square feet, number of bedrooms, and average income in the neighborhood according to census data. The relationship between sale price and the predictors in the training set would provide a predictive model.

  There are many types of regression methods including the above – multivariate linear regression, polynomial regression, and regression trees, to mention a few.

Classification is another popular type of predictive analytics. With classification, there is a response categorical variable. The classifier examines a data set where each observation contains information on the response variable as well as the predictor variables.

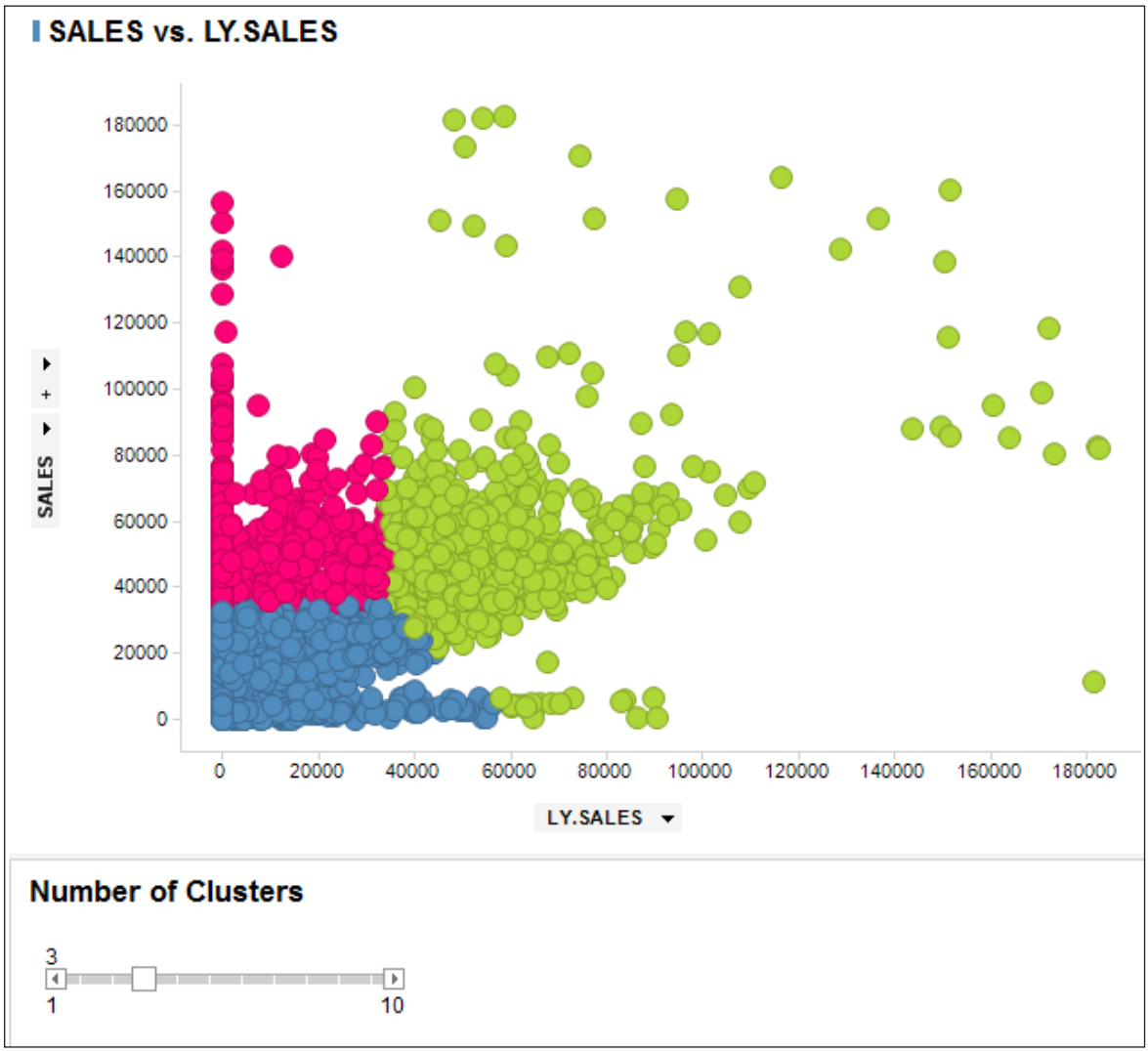
- **Classification –** Classification is another popular type of predictive analytics. With classification, there is a response categorical variable, such as income bracket, which could be partitioned into three classes or categories: high income, middle income, and low income. The classifier examines a data set where each observation contains information on the response variable as well as the predictor variables. For example, suppose an analyst would like to be able to classify the income brackets of persons not in the data set, based on characteristics associated with that person, such as age, gender, and occupation. This is a classification task that would proceed as follows: examine the data set containing both the predictor variables and the already classified response variable, income bracket. In this way, the algorithm learns about which combinations of variables are associated with which income brackets. This data set is called the training set. Then the algorithm would look at new observations for which no information about income bracket is available. Based on the classifications in the training set, the algorithm would assign classifications to the new observations. For example, a 51 year old female marketing director might be classified in the high-income bracket.

  There are many types of classification methods such as logistic regression, decision trees, support vector machines, Random Forests, k-Nearest Neighbors, naïve Bayes, to mention a few.

Unsupervised learning is used to draw inferences from datasets consisting of input data without labeled responses. The most common unsupervised learning method is cluster analysis, which is used for exploratory data analysis to find hidden patterns or grouping in data.

- **Clustering –** Using unsupervised techniques like clustering, we can seek to understand the relationships between the variables or between the observations by determining whether observations fall into relatively distinct groups. For example, in a customer segmentation analysis we might observe multiple variables: gender, age, zip code, income, etc. Our belief may be that the customers fall in different groups like frequent shoppers and infrequent shoppers. A classification analysis would be possible if the customer shopping history were available, but this is not the case in unsupervised learning — we don't have response variables telling us whether a customer is a frequent shopper or not. Instead, we can attempt to cluster the customers on the basis of the variables in order to identify distinct customer groups.

There are other types of unsupervised statistical learning including k-means clustering, hierarchical clustering, principal component analysis, etc.



Clustering shows the relationships between the variables or observations by determining whether they fall into relatively distinct groups. Image courtesy of TIBCO Spotfire.

Predictive Analytics Software

There is a vast array of predictive analytics tools, but not all are created equal. Software differs widely in terms of capability and usability — not all solutions can address all types of advanced analytics needs. There are different classes of analytics users — some need to build statistical models, others just need to use them.

For the advanced user, the importance of tool selection centers on the ability to put proprietary models into the hands of business users (front line decision makers) so that they can act competitively with predictive analytics, hiding the complexity of these proprietary models under the hood.

Effective predictive analytics tools provide a wide variety of algorithms and methods to support all the data characteristics and business problems that users encounter, as well as the ability to flexibly apply these algorithms as needed.

Business users possess the domain knowledge necessary to understand the business answer they are looking for from predictive analytics, but at the same time they don't need, want to, or can't develop the models themselves. So the optimal tool provides an easy method of putting the data scientists' expertise in the hands of frontline decision makers, often times in the form of a guided analytic application, with the predictive model encapsulated under the covers. This enables best practices use of advanced analytics (i.e. taking the risk out of having business people try to develop their own models), and broad deployment of secret-sauce analytics.

When selecting the right tool for your organization, you need to ensure you choose a tool which has the depth and breadth of capability, from simple out-of-the-box functionality for the easiest problems to the most advanced statistic capability to support data scientists, so that competitive models can be embedded in business users' analytic dashboards for day-to-day use.

Effective predictive analytics tools provide a wide variety of algorithms and methods to support all the data characteristics and business problems that users encounter, as well as the ability to flexibly apply these algorithms as needed. The extensibility to easily integrate new analytic methods as they become available is also critical for maximizing competitive advantage. An important criteria when selecting the right tool, is to make sure the feature-set matches your business data characteristics and that the tool will benefit your data analysts. The right tool typically combines powerful data integration and transformation capabilities, exploratory features, analytic algorithms, all with an intuitive interface. In essence there are three important ingredients providing a recipe for success in utilizing predictive analytics: (i) the data scientist builds the most competitive model, (ii) the analytic application author embeds the competitive model into the analytic application, and (iii) the business user engages the competitive model as part of the regular flow of business.

Here is a short list of characteristics and considerations to focus on when evaluating a predictive analytics tool:

- Consider the processing capabilities of the analytics tool for addressing the needs of the predictive analytics cycle — data munging, exploratory data analysis, predictive modeling techniques such as forecasting, clustering, and scoring, as well as model evaluation.
- Find a tool that supports combining the analyst's business and data knowledge with predefined procedures and tools, and graphical workflows to simplify and streamline the path from preparation to prediction.
- A good tool must easily integrate with the data sources required to answer critical business questions.
- The tool should be readily usable by all classes of users: business users, business analysts, data analysts, data scientists, application developers and system administrators.
- Consider tools that serve to minimize the need for IT professionals and data scientists to set-up integration with multiple data sources.

The goal in selecting a robust tool is to secure a broad range of predictive analytics capabilities — from the simplest, such as trend lines and a forecast tool, all the way through to leveraging an entire ecosystem of statistical capabilities where you have the full depth of capability in creating and executing any type of statistical model or algorithm. Out-of-the-box/standard algorithms aren't going to gain you a competitive advantage once your competitors start using those same tools. You need the tools to create your own proprietary models that will allow you to build that competitive advantage by leveraging your enterprise data assets.

A best-practice choice is a solution that integrates predictive analytics within the entire analytic decision-making process, allowing it to be incorporated where appropriate into self-service dashboards and exploratory data discovery. This orientation provides advanced analytics access to all analytic users, giving them the tools necessary to spot new opportunities, manage risks, and swiftly react to unforeseen events. Further, professionals managing mission-critical departments and global processes have the ability to immediately and intuitively ask questions and get answers from their data — anticipating what's next, taking quick and educated actions.

R as the Choice for Predictive Analytics

Although there are many choices for performing tasks related to data analysis, data modeling, and predictive analytics, R has become the overwhelming favorite today. This is due to the widespread use of R in academia over commercial products like SAS and SPSS, where new graduates enter industry with a firm knowledge of R.

There are currently spirited debates between the R user community and both the SAS and Python communities as to what is the best tool for data science. R has compelling justifications including the availability of free open source R, a widely used extensible analytical environment, over 5,000 packages available on CRAN to extend the functionality of R, and top-rated visualization capabilities

The only issue with the open source R engine is its inherent limitation as a scalable production environment. R is notoriously memory-based, meaning it can only run on the confines of its compute environment.

using ggplot2. In addition, R enjoys a thriving user community flush with local Meetup groups, online courses, and specialty blogs (see top blogs via consolidator: r-bloggers.com).

Open source R is a logical first choice for predictive analytics modeling as the statistical environment contains a number of algorithms in the base R package as well as additional packages that have extended functionality.

- Linear regression using lm()
- Logistic regression using glm()
- Regression with regularization using the glmnet package
- Neural networks using nnet()
- Support vector machines using the e1071 package
- Naïve Bayes models using the e1071 package
- K-nearest-neighbors classification using the knn() function from the class package
- Decision trees using tree()
- Ensembles of trees using the randomForest package
- Gradient boosting using the gbm package
- Clustering using kmeans(), hclust()

The only issue with the open source R engine is its inherent limitation as a scalable production environment. R is notoriously memory-based, meaning it can only run on the confines of its compute environment. A good best practices policy for implementing R in production would be to leverage a commercial, enterprise-grade platform for running the R language, such as TERR (TIBCO Enterprise Runtime for R), in order to get the great benefits of R, while avoiding the scalability challenges.

Data Access for Predictive Analytics

Enterprise data assets are what feed the predictive analytic process, and any tool must facilitate easy integration with all the different types data sources required to answer critical business questions. Robust predictive analytics needs to access analytical and relational databases, OLAP cubes, flat files, and enterprise applications. The following data integration areas may be required by predictive analytics:

- Structured data sources such as traditional SQL databases and data warehouses already in use by the enterprise.
- Unstructured data sources such as social media, e-mail, etc.
- External third-party data included from vendors such as Salesforce.

Tools for predictive analytics should make integration with multiple data sources quick and straightforward without the need for exhaustive work by IT professionals and data scientists.

Users should have the flexibility to quickly combine their own private data stores, such as Excel spreadsheets or Access databases, with corporate data stores, such as Hadoop or cloud application connectors (e.g. Hadoop/Hive, Netezza, HANA, Teradata, and many others). Support for in-database, in-memory and on-demand analytics via direct connectors are all features gaining steam in the predictive analytics arena.

Open source R offers a low cost of entry for ample enterprise data access as it possesses many packages providing access to a wide range of data sources including ODBC databases, Excel, CSV, Twitter, Google Analytics just to name a few.

Best practice is to ensure your predictive analytics solution provides access to all types of data sources so you can combine and mashup data in any variety of ways in order to get a holistic view of the business — preferably without coding or without requiring IT involvement. This capability will empower users to derive powerful insights and make educated business decisions in real time.

An integral step in preparing for predictive analytics is to become intimately familiar with the data, a process known as exploratory data analysis (EDA).

Exploratory Data Analysis (EDA)

An integral step in preparing for predictive analytics is to become intimately familiar with the data, a process known as exploratory data analysis (EDA). A clear understanding of the data provides the foundation for model selection, i.e. choosing the appropriate predictive analytics algorithm to solve your business problem. Various types of software can be used by different users for an initial exploration of data.

One way to gain this level of familiarity is to utilize the many features of the R statistical environment to support this effort — numeric summaries, plots, aggregations, distributions, densities, reviewing all the levels of factor variables and applying general statistical methods. Other tools can also be used effectively for EDA including: TIBCO Spotfire, SAS, SPSS, Statistica, Matlab among many others. Statistical software, such as R, enables a user to very flexibly explore and visualize data, but requires a high level of knowledge of scripting.

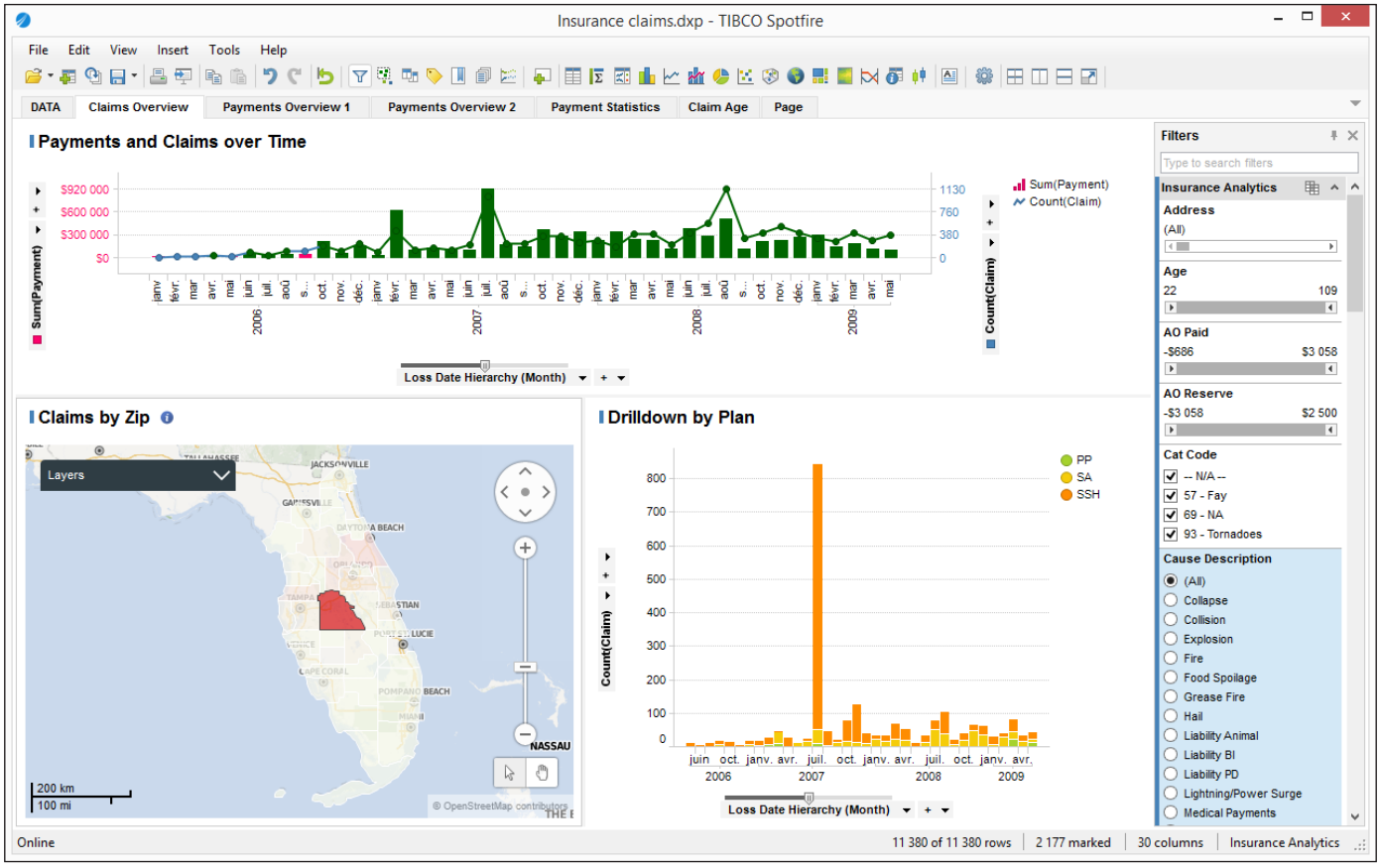
With thorough EDA you can gain important insights into the story your data is telling and how best to utilize the data to make accurate predictions.

Open source R has many visualization mechanisms for EDA including histograms, boxplots, barplots, scatterplots, heatmaps, and many others using the ggplot2 library. Using these tools allows for a deep understanding of the data being employed for predictive analytics. With thorough EDA you can gain important insights into the story your data is telling and how best to utilize the data to make accurate predictions.

Another method for EDA is data discovery software. Data discovery, such as TIBCO Spotfire, enables a wide variety of users to visually explore and understand their data, without requiring deep statistical knowledge. Users can perform enhanced EDA tasks that can add an additional layer of insights without having to request assistance from their IT departments or data scientists.

Combining data discovery and predictive analytics capabilities on the same analytics platform is a best practice to give analytics users a seamless experience as they move from one analytics task to another, in addition to providing a more sound total cost of ownership.

Data discovery, such as TIBCO Spotfire, enables a wide variety of users to visually explore and understand their data, without requiring deep statistical knowledge.



Data discovery software offers a rich, interactive analytic interface for EDA including accessing and manipulating data, and composing analyses. Image courtesy of TIBCO Spotfire.

Predictive Modeling

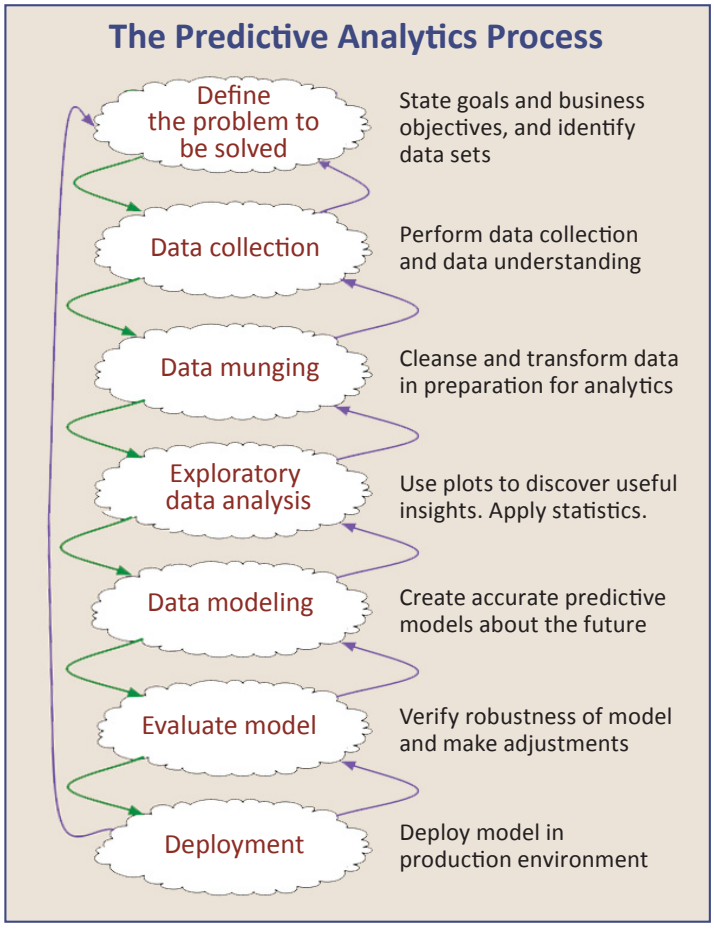
Using predictive analytics involves understanding and preparing the data, defining the predictive model, and following the predictive process. Predictive models can assume many shapes and sizes, depending on their complexity and the application for which they are designed. The first step is to understand what questions you are trying to answer for your organization. The level of detail and complexity of your questions will increase as you become more comfortable with the analytic process. The most important steps in the predictive analytics process are as follows:

- Define the project outcomes and deliverables, state the scope of the effort, establish business objectives, and identify the data sets to be used.
- Undertake data collection and data understanding.
- Perform data munging – the process of inspecting, cleaning, and transforming the data.
- Utilize exploratory data analysis (EDA) – use graphical techniques with the objective of discovering useful information, arriving at conclusions. Apply statistics to validate the assumptions, hypothesis and test using standard statistical techniques.
- Apply modeling principles to provide the ability to automatically create accurate predictive models about the future.
- Evaluate the model allowing you to verify the robustness of the chosen model and make mid-course corrections. Test models on existing data and apply predictions to new data.
- Select a deployment option to open up the analytical results to every day decision making and to get results by automating the decisions based on the modeling.

Each of the above steps can be considered iterative and may be revisited as needed. It should be noted that the data munging step often is very time-consuming depending on the cleanliness of the incoming data and can take up to 70% of the overall project timeline.

Characteristics of the data can often help you determine what predictive modeling techniques might best meet the data analyst's needs. Here are a number of points to consider when determining which technique to use based on your data and the problem you wish to solve.

- When the data is grouped by observations, tools such as cluster analysis, association rules, and k-nearest neighbors usually provide the best results.
- Use classification to separate the data into classes based on the response variable – both binary classes like True or False, as well as multi-class situations.
- Use single, multiple and polynomial regression when attempting to make a prediction rather than a classification.
- In poor quality or limited data situations, A/B testing is appropriate. As an example, A/B tests are statistical experiments that help you decide whether a change is actually making a significant impact on your product.



Production Deployment

The final step in the predictive analytics project timeline is to determine how best to deploy the solution to a production environment. Of primary concern is using open source R on larger data sets where performance is important. The open source R engine was not built for enterprise usage. Deploying open source R can problematic for the following reasons:

- **Poor memory management –** R does not reclaim memory well, so memory use can grow faster, leading to out-of-memory crashes, as well as non-linear performance due to increased garbage collection requests, and increased swapping.
- **Risk of deploying open source with GPL license –** software vendors are forbidden to embed or redistribute open source R as a part of any commercial closed-source software.

In order to avoid these issues, analysts often will opt to convert their working R solution to a different programming environment like C++ or Python. This path, however, is far from optimal since it requires recoding and significant retesting.

Best practice would be to use a commercial, enterprise-grade R solution, like TIBCO Software's Enterprise Runtime for R (TERR) to resolve the above limitations and to yield a robust production environment. Because many corporations already have legacy predictive models in house, it is also recommended that you ensure your analytics platform supports TERR, open source R, S+, MATLAB and SAS models, in order to take advantage of an ecosystems of predictive analytics.

Conclusion

In this Guide we have reviewed how predictive analytics helps your organization predict with confidence what will happen next so that you can make smarter decisions and improve business outcomes. It is important to adopt a predictive analytics solution that meets the specific needs of different users and skill sets from beginners, to experienced analysts, to data scientists.

With predictive analytics software you can:

- Transform data into predictive insights to guide front-line decisions and interactions.
- Predict what customers want and will do next to increase profitability and retention.
- Maximize the productivity of your people and processes.
- Increase the value of your data assets.
- Detect and avoid security threats and fraud before they affect your organization.
- Perform statistical analysis including regression analysis, classification, and cluster analysis.
- Measure the social media impact of your products, services and marketing campaigns.

About TIBCO Spotfire

TIBCO Spotfire® is the analytics solution from infrastructure and business intelligence giant, TIBCO Software. From interactive dashboards and data discovery to predictive and real-time analytics, Spotfire's intuitive software provides an astonishingly fast and flexible environment for visualizing and analyzing your data. As your analytics needs increase, our enterprise-class capabilities can be seamlessly layered on, helping you to be first to insight — and first to action.

TIBCO Spotfire has a long, rich history in predictive analytics. With Spotfire you can develop your own proprietary models and leverage your investments in R, S+, SAS, MATLAB, and in-database analytics of Big Data sources, such as Teradata Aster. Spotfire also offers a commercial-grade R environment, TERR (TIBCO Enterprise Runtime for R), which was built from the ground up to extend the reach of R to the enterprise, making R faster, more scalable, and able to handle memory much more efficiently than the open source R engine.

TIBCO regularly contributes to the R community, including feedback to the R Core team, and offers broad compatibility with R functions and a growing number of CRAN packages, currently 1800+. The company regularly tests TERR with a wide variety of R packages, and continues to extend TERR to greater R coverage. TERR can be used in RStudio, the popular R IDE and also integrates fully with TIBCO Spotfire, as well as TIBCO Complex Event Processing products, such as TIBCO Streambase.

Learn more about TIBCO Spotfire and TERR at spotfire.com