DATA WAREHOUSE WITH IBM CLOUD

Phase 2 : INNOVATION

Consider incorporating advanced analytics tools or machine learning models for predictive analysis within the data warehouse

Predictive analysis

Predictive analytics is a form of advanced analytics that uses current and historical data to forecast activity, and trends. It involves applying statistics analysis techniques, data queries and machine learning algorithms to data sets to create predictive models DEFINITION

Predictive analytics is a key discipline in the field of *data analytics*, an umbrella term for the use of quantitative methods and expert knowledge to derive meaning from data and answer fundamental questions about a business, the weather, healthcare, scientific research and other areas of inquiry. In the context of businesses, the main focus here, that process is often referred to as *business analytics*.

What are business analytics

There are three major types of business analytics.

- Descriptive analytics : The most common type is *descriptive analytics*, which gives an account of what has happened in a business.
- *Predictive analytics*: the subject of this guide, helps businesses predict what will likely happen. It looks for patterns in data and projects them forward to help businesses mitigate risks and capitalize on opportunities
- Prescriptive analytics : prescribes or automatically takes a next best course of action based on intelligence generated by the other two kinds of analytics.

- Diagnostics analytics: which explores why something happened
- Real time: which analyzes data as it's generated, collected or updated.

Analytical tools can be used for predictive analytics

- IBM SPSS Statistics : Best For Dashboard Capabilities.
- SAS Advanced Analytics : Best For Variety.
- SAP Predictive Analytics: Best For ERP Data.
- TIBCO Data Science : Best For Collaboration.
- Oracle Cloud Infrastructure (OCI) Data Science: Best For Cloud Management.

Predictive Analytics Models

Classification Model

The classification model is, in some ways, the simplest of the several types of predictive analytics models we're going to cover. It puts data in categories based on what it learns from historical data.

Classification models are best to answer yes or no questions, providing broad analysis that's helpful for guiding decisive action. These models can answer questions such as:

- For a retailer, "Is this customer about to churn?"
- For a loan provider, "Will this loan be approved?" or "Is this applicant likely to default?"
- For an online banking provider, "Is this a fraudulent transaction?"

The breadth of possibilities with the classification model—and the ease by which it can be retrained with new data—means it can be applied to many different industries.

• Clustering Model

The clustering model sorts data into separate, nested smart groups based on similar attributes. If an ecommerce shoe company is looking to implement targeted marketing campaigns for their customers, they could go through the hundreds of thousands of records to create a tailored strategy for each individual. But is this the most efficient use of time? Probably not. Using the clustering model, they can quickly separate customers into similar groups based on common characteristics and devise strategies for each group at a larger scale.

Other use cases of this predictive modeling technique might include grouping loan applicants into “smart buckets” based on loan attributes, identifying areas in a city with a high volume of crime, and benchmarking SaaS customer data into groups to identify global patterns of use.

• Forecast Model

One of the most widely used predictive analytics models, the forecast model deals in metric value prediction, estimating numeric value for new data based on learnings from historical data.

This model can be applied wherever historical numerical data is available. Scenarios include:

- A SaaS company can estimate how many customers they are likely to convert within a given week.
- A call center can predict how many support calls they will receive per hour.
- A shoe store can calculate how much inventory they should keep on hand in order to meet demand during a particular sales period.

The forecast model also considers multiple input parameters. If a restaurant owner wants to predict the number of customers she is likely to receive in the following week, the model will take into account factors that could impact this, such as: Is there an event close by? What is the weather forecast? Is there an illness going around?

• Outliers Model

The outliers model is oriented around anomalous data entries within a dataset. It can identify anomalous figures either by themselves or in conjunction with other numbers and categories.

- Recording a spike in support calls, which could indicate a product failure that might lead to a recall
- Finding anomalous data within transactions, or in insurance claims, to identify fraud
- Finding unusual information in your NetOps logs and noticing the signs of impending unplanned downtime

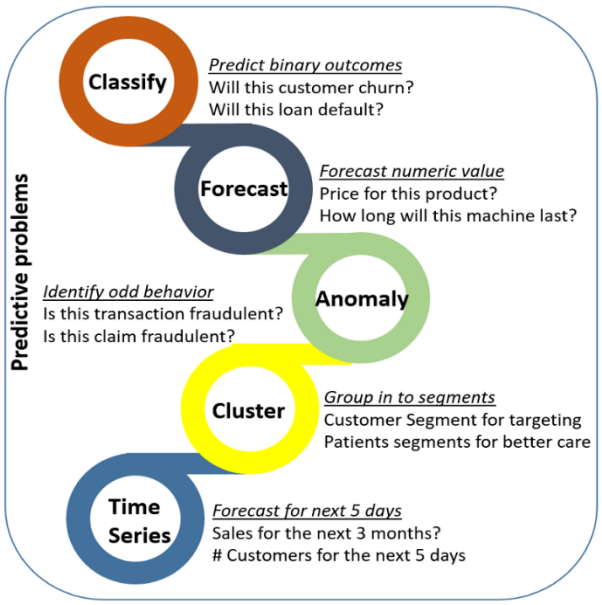
The outlier model is particularly useful for predictive analytics in retail and finance. For example, when identifying fraudulent transactions, the model can assess not only amount, but also location, time, purchase history and the nature of a purchase (i.e., a $1000 purchase on electronics is not as likely to be fraudulent as a purchase of the same amount on books or common utilities).

• Time Series Model

The time series model comprises a sequence of data points captured, using time as the input parameter. It uses the last year of data to develop a numerical metric and predicts the next three to six weeks of data using that metric. Use cases for this model includes the number of daily calls received in the past three months, sales for the past 20 quarters, or the number of patients who showed up at a given hospital in the past six weeks. It is a potent means of understanding the way a singular metric is developing over time with a level of accuracy beyond simple averages. It also takes into account seasons of the year or events that could impact the metric.

If the owner of a salon wishes to predict how many people are likely to visit his business, he might turn to the crude method of averaging the total number of visitors over the past 90 days. However, growth is not always static or linear, and the time series model can better model exponential growth and better align the model to a company's

trend. It can also forecast for multiple projects or multiple regions at the same time instead of just one at a time.



Common Predictive Algorithms

predictive analytics algorithms can be separated into two groups: machine learning and deep learning.

- **Machine learning** involves structural data that we see in a table. Algorithms for this comprise both linear and nonlinear varieties. Linear algorithms train more quickly, while nonlinear are better optimized for the problems they are likely to face (which are often nonlinear).
- **Deep learning** is a subset of machine learning that is more popular to deal with audio, video, text, and images.

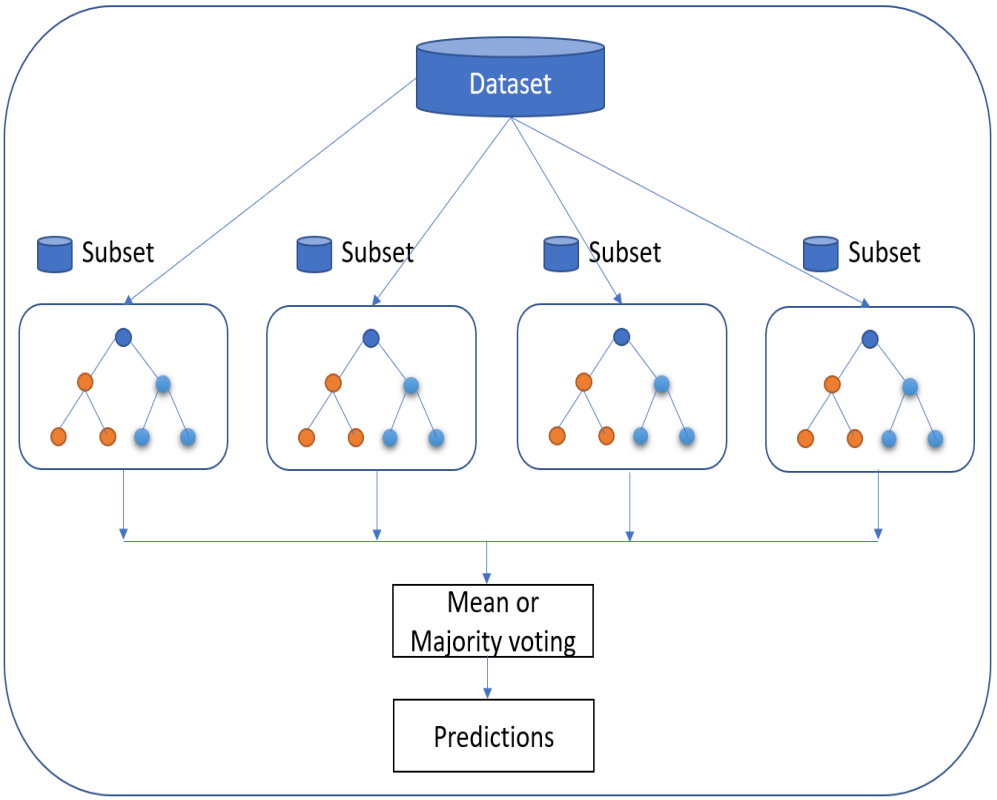
With machine learning predictive modelling there are several different algorithms that can be applied. Below are some of the most common algorithms

• Random Forest

Random Forest is perhaps the most popular classification algorithm, capable of both classification and regression. It can accurately classify large volumes of data.

The name “Random Forest” is derived from the fact that the algorithm is a combination of decision trees. Each tree depends on the values of a random vector sampled independently with the same distribution for all trees in the “forest.” Each one is grown to the largest extent possible.

Predictive analytics algorithms try to achieve the lowest error possible by either using “boosting” (a technique which adjusts the weight of an observation based on the last classification) or “bagging” (which creates subsets of data from training samples, chosen randomly with replacement). Random Forest uses bagging. If you have a lot of sample data, instead of training with all of them, you can take a subset and train on that, and take another subset and train on that (overlap is allowed). All of this can be done in parallel. Multiple samples are taken from your data to create an average.



• Generalized Linear Model (GLM) for Two Values

The Generalized Linear Model (GLM) is a more complex variant of the General Linear Model. It takes the latter model's comparison of the effects of multiple variables on continuous variables before drawing from an array of different distributions to find the "best fit" model.

Let's say you are interested in learning customer purchase behavior for winter coats. A regular linear regression might reveal that for every negative degree difference in temperature, an additional 300 winter coats are purchased. While it seems logical that another 2,100 coats might be sold if the temperature goes from 9 degrees to 3, it seems less logical that if it goes down to -20, we'll see the number increase to the exact same degree.

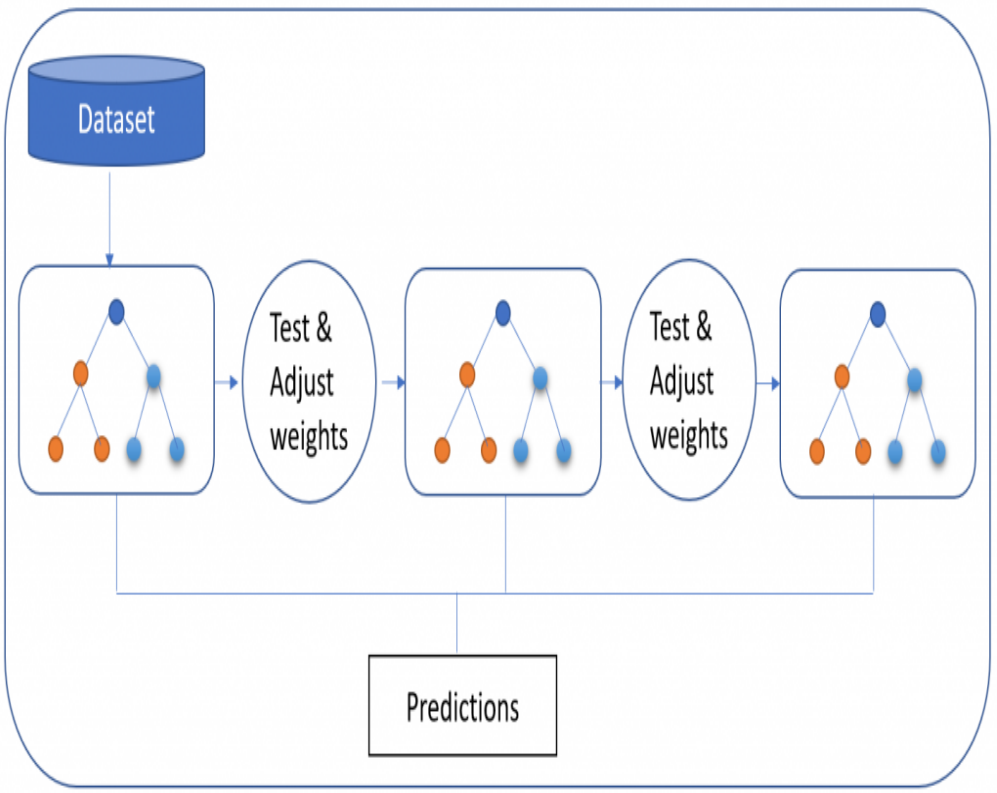
The Generalized Linear Model would narrow down the list of variables, likely suggesting that there is an increase in sales beyond a certain temperature and a decrease or flattening in sales once another temperature is reached.

The advantage of this algorithm is that it trains very quickly. The response variable can have any form of exponential distribution type. The Generalized Linear Model is also able to deal with categorical predictors, while being relatively straightforward to interpret. On top of this, it provides a clear understanding of how each of the predictors is influencing the outcome, and is fairly resistant to overfitting. However, it requires relatively large data sets and is susceptible to outliers

• Gradient Boosted Model (GBM)

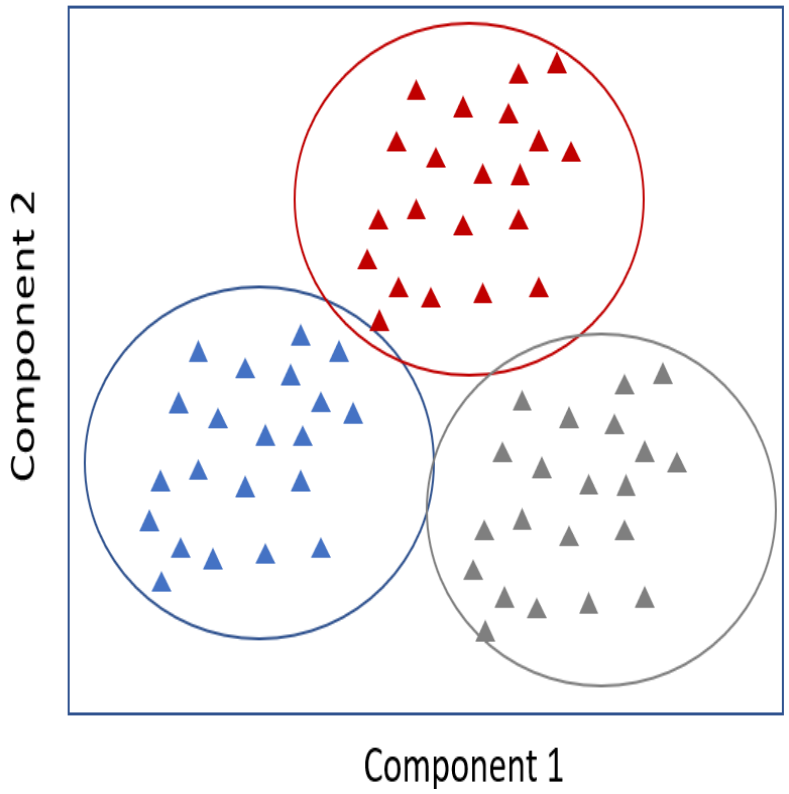
The Gradient Boosted Model produces a prediction model composed of an ensemble of decision trees (each one of them a "weak learner," as was the case with Random Forest), before generalizing. As its name suggests, it uses the "boosted" machine learning technique, as opposed to the bagging used by Random Forest. It is used for the classification model.

The distinguishing characteristic of the GBM is that it builds its trees one tree at a time. Each new tree helps to correct errors made by the previously trained tree—unlike in the Random Forest model, in which the trees bear no relation. It is very often used in machine-learned ranking, as in the search engines Yahoo and Yandex.



- K-Means

A highly popular, high-speed algorithm, K-means involves placing unlabeled data points in separate groups based on similarities. This algorithm is used for the clustering model. For example, Tom and Rebecca are in group one and John and Henry are in group two. Tom and Rebecca have very similar characteristics but Rebecca and John have very different characteristics. K-means tries to figure out what the common characteristics are for individuals and groups them together. This is particularly helpful when you have a large data set and are looking to implement a personalized plan—this is very difficult to do with one million people.

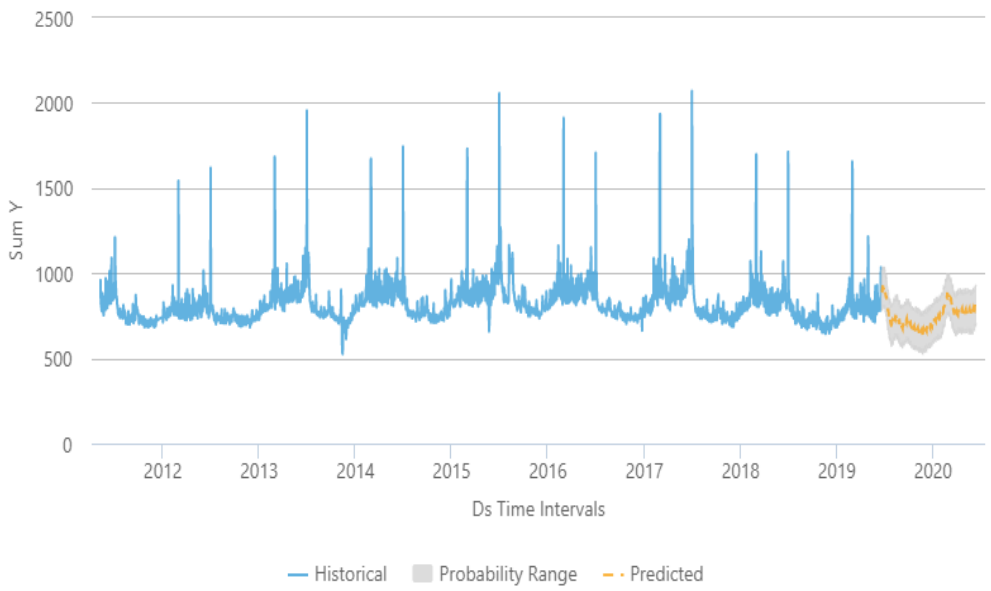


In the context of a sample size of patients might be placed into five separate clusters by the algorithm. One particular group shares multiple characteristics: they don't exercise, they have an increasing hospital attendance record (three times one year and then ten times the next year), and they are all at risk for diabetes. Based on the similarities, we can proactively recommend a diet and exercise plan for this group.

- Prophet

The Prophet algorithm is used in the time series and forecast models. It is an open-source algorithm developed by Facebook, used internally by the company for forecasting.

The Prophet algorithm is of great use in capacity planning, such as allocating resources and setting sales goals. Owing to the inconsistent level of performance of fully automated forecasting algorithms, and their inflexibility, successfully automating this process has been difficult. On the other hand, manual forecasting requires hours of labor by highly experienced analysts.



Predictive analytics vs. machine learning

Machine learning lends itself to various applications, while predictive analytics focuses on forecasting specific variables and scenarios.

Benefits of Predictive Modeling

In a nutshell, predictive analytics reduce time, effort and costs in forecasting business outcomes. Variables such as environmental factors, competitive intelligence, regulation changes and market conditions can be factored into the mathematical calculation to render more complete views at relatively low costs.

Examples of specific types of forecasting that can benefit businesses include demand forecasting, headcount planning, churn analysis, external

factors, competitive analysis, fleet and IT hardware maintenance and financial risks.

Challenges of Predictive Modeling

It's essential to keep predictive analytics focused on producing useful business insights because not everything this technology digs up is useful. Some mined information is of value only in satisfying a curious mind and has few or no business implications. Getting side-tracked is a distraction few businesses can afford.

Also, being able to use more data in predictive modeling is an advantage only to a point. Too much data can skew the calculation and lead to a meaningless or an erroneous outcome. For example, more coats are sold as the outside temperature drops. But only to a point. People do not buy more coats when it's -20 degrees Fahrenheit outside than they do when it's -5 degrees below freezing. At a certain point, cold is cold enough to spur the purchase of coats and more frigid temps no longer appreciably change that pattern.

And with the massive volumes of data involved in predictive modeling, maintaining security and privacy will also be a challenge. Further challenges rest in machine learning's limitations.

Limitations of Predictive Modeling

According to a McKinsey report, common limitations and their "best fixes" include:

1. **Errors in data labeling:** These can be overcome with reinforcement learning or generative adversarial networks (GANs).
2. **Shortage of massive data sets needed to train machine learning:** Apossible fix is "one-shot learning," wherein a machine learns from a small number of demonstrations rather than on a massive data set.
3. **The machine's inability to explain what and why it did what it did:** Machines do not "think" or "learn" like humans. Likewise, their computations can be so exceptionally complex that humans have trouble finding, let alone following, the logic. All this makes it difficult

for a machine to explain its work, or for humans to do so. Yet model transparency is necessary for a number of reasons, with human safety chief among them. Promising potential fixes: local-interpretable-model-agnostic explanations (LIME) and attention techniques.

4. **Generalizability of learning, or rather lack thereof:** Unlike humans, machines have difficulty carrying what they've learned forward. In other words, they have trouble applying what they've learned to a new set of circumstances. Whatever it has learned is applicable to one use case only. This is largely why we need not worry about the rise of AI overlords anytime soon. For predictive modeling using machine learning to be reusable—that is, useful in more than one use case—a possible fix is transfer learning.
5. **Bias in data and algorithms:** Non-representation can skew outcomes and lead to mistreatment of large groups of humans. Further, baked-in biases are difficult to find and purge later. In other words, biases tend to self-perpetuate. This is a moving target, and no clear fix has yet been identified.

The Future of Predictive Modeling

Predictive modeling, also known as predictive analytics, and machine learning are still young and developing technologies, meaning there is much more to come. As techniques, methods, tools and technologies improve, so will the benefits to businesses and societies.

However, these are not technologies that businesses can afford to adopt later, after the tech reaches maturity and all the kinks are worked out. The near-term advantages are simply too strong for a late adopter to overcome and remain competitive.

Our advice: Understand and deploy the technology now and then grow the business benefits alongside subsequent advances in the technologies.

Predictive Modelling in Platforms

For all but the largest companies, reaping the benefits of predictive analytics is most easily achieved by using ERP systems that have the technologies built-in and contain pretrained machine learning. For example,

planning, forecasting and budgeting features may provide a statistical model engine to rapidly model multiple scenarios that deal with changing market conditions.

As another example, a supply planning or supply capacity function can similarly predict potentially late deliveries, purchase or sales orders and other risks or impacts. Alternate suppliers can also be represented on the dashboard to enable companies to pivot to meet manufacturing or distribution requirements.

Financial modeling and planning and budgeting are key areas to reap the many benefits of using these advanced technologies without overwhelming your team.