Using Artificial Intelligence to Build Next Generation Solutions

Prof. Dr. Michael Feindt & Dr. Ulrich Kerzel
(Chief Scientific Advisor) | (Principal Data Scientist)

Abstract

Cutting edge artificial intelligence (AI) technology is the cornerstone of Blue Yonder's products. Machine learning is at the core of AI. Our products concentrate on optimization and automation of concrete processes, using data-driven predictive analytics and prescriptive analytics (optimized decisions), all of which are important value-generating subtopics of AI. We build upon and contribute to technologies developed in the open source community, while pushing the boundaries of predictive technologies further with our proprietary algorithms, own research and custom development. This article explores our research strategy and gives some insight as to why steady development is needed to build the next generation of machine learning applications that will help companies to transform into Predictive Enterprises.

Introduction

Blue Yonder's machine learning solutions are focussed on delivering decisions to our customers in the time and granularity required by their operational systems. The starting point of making these decisions are predicted probabilities or probability density distributions that are calculated by sophisticated machine learning algorithms. These predictions are then used with appropriate metrics, such as a cost function, to optimize the individual action, which needs to be taken at the appropriate moment. In a retail environment, such a decision could be setting the best price for a specific product at a specific time for a specific store, made with consideration to other constraints from inventory, supply chain and logistics. In an industrial setting, decisions might be made in regard to choosing the next best step while a specific item or material is being manufactured, based on real-time measurements of the quality of ingredients or pre-manufactured components, production capacity, etc.

The development of machine learning algorithms and artificial intelligence has been an active research area for decades, and the global academic community publishes hundreds of papers every day on the subject. Tech giants have managed to secure the support of the founding fathers of artificial intelligence including G. Hinton at Google or Y. LeCun at Facebook, driving the technology forward. It is worth noting that these companies are not traditionally rooted in a specific domain or are considered as IT specialists, such as IBM or Microsoft by their heritage. Instead, they are rigorously data-driven and apply their knowledge to enter a number of industry verticals, pushing traditional players aside.

While simple linear modeling is still in the majority of commercial applications, more sophisticated techniques are rapidly emerging. Generalized linear models, neural networks, decision trees, random forests, support vector machines, Bayesian belief networks, NeuroBayes, reinforcement learning and deep learning are just some of the buzzwords in this vibrant field.

Is there a single best method that can be applied in all cases? Are some methods always better than others? In this article we want to highlight how we here at Blue Yonder make finding solutions to each challenge the focus of our work and combine a variety of approaches to achieve our goals.

Different Problems Demand Different Solutions

It may seem like stating the obvious, but different problems demand different solutions. At Blue Yonder, we follow a strictly scientific approach to problem solving. In order to make accurate forecasts of future events, we extract knowledge from data and combine this with some *a priori* knowledge, to build better *a posteriori* knowledge. We then base our decisions on the resulting predictions. Modelling prior knowledge (in a Bayesian sense) allows us to include details such as domain knowledge or further information about a specific process.

Each application and use-case is unique: While the general description of the problem may be easy to describe and the same for a number of scenarios, the details are, in most cases, specific to the concrete implementation. For example, in an industrial setting concerned with Predictive Maintenance, one might ask: "When does this machine or robot need servicing?" This question is sufficiently generic to be applied to almost all production plants. However, a more specific (and useful) question would be: "When should the next service slot for this machine or robot be scheduled?" The two questions are related but the former is focused on the machine in question, while the latter takes context into account, such as what is the utilization level of the production plant? What is the order pipeline? Do I have to risk producing a specific good even in adverse circumstances because a penalty would be too high? Are spare machines available that could continue production while this specific machine is being serviced? What are the planned maintenance slots and the availability of service personnel, as well as spare parts?

Being able to make a good prediction allows us to address the question: "When will the machine break down?" But incorporating the relevant context allows us to derive a decision from it that can then be used in the operational decisions.

Each new implementation comes with its own unique challenges: Which data are available – and which are available in principle but have to be extracted from a number of "historic" systems or even printed documents (such as maintenance reports)? Is sufficient data available and what quality does the data have? How much time will be needed to clean the data? Are there specific processes or procedures unique to this new implementation that have to be taken into consideration? Which further constraints apply?

The stochastic component of the data is a key point that is often missed. Most realistic scenarios are neither fully deterministic nor completely random. Predictions for deterministic systems such as a simple pendulum are (almost) trivial as the outcome can be calculated for any time in the future, just as a grandfather clock measures time. Predictions for random systems are pointless as there is nothing to predict, just as asking "What are the numbers that will be drawn in the next lottery?" cannot be answered. Almost all practical scenarios have a deterministic and a stochastic component, but only this deterministic part forms the core of an individualized prediction. On the other hand, the large stochastic component leads to an uncertain or volatile prediction. This means we cannot know exactly what will happen, but we can make probability statements that contain all the information about a specific prediction.

In some cases, the probability may be very near to 0% (the event will not happen) or 100% (the event will definitely happen). Being able to quantify volatility is a key aspect of predicting stochastic events as this allows us to optimize the subsequent decisions in the presence of uncertainty.

Finding Common Grounds

There are many common ingredients in different prediction tasks that actually make them more similar than initially appears. Meta analysis of Blue Yonder's scientists' experience has led to common problem solving strategies, common data organization, data set exploration strategies, analysis frameworks, template programs, visualization tools, the NeuroBayes and other machine learning libraries. This is the base Blue Yonder uses to create excellent solutions for new and demanding problems and products.

Same Problems Require Same Solutions

This is obvious and is the basis for Blue Yonder's product roadmap for retail, with world-leading replenishment and pricing solutions already existing with great operational success in real-world uses.

Superb Technology is Made by Superb People

All of our data scientists at Blue Yonder have a strong academic background in STEM research areas, most have a PhD and a significant number spent time pursuing fundamental and applied research topics prior to joining us. Many have worked, for example at particle accelerator experiments like CERN (Switzerland), Fermilab (USA) or KEK (Japan), at cosmic ray telescopes, in quantum information theory, nanotechnology, biophysics, computer science or in robotics. They all have profound statistical knowledge, are keen analytical thinkers and have excellent programming skills.

Specialists from different research areas and scientific communities with diverse backgrounds in different methods and research cultures are the heart of every Blue Yonder team. This diverse skill set, founded on a common ground of research excellence and paired with our creative, cooperative and communicative atmosphere makes the data science team at Blue Yonder much stronger than the sum of its individual members.

Blue Yonder has developed into a highly-attractive employer for data scientists, winning talent from across Europe. Our people are always looking for the best solution, eagerly taking every chance to learn more and try new techniques to improve.

What Makes a Good Prediction?

One of the key questions when dealing with predictions is how to quantify the quality of the prediction. Naïve expectations are often formulated as "The prediction should match the observed event". Although intuitive, this approach doesn't take the stochastic component of nature into account, nor does it reflect the fact that predictions are either probabilities or probability density distributions. A vast amount of quality metrics are known in the academic literature – but just applying any (or all) of them doesn't help much in evaluating the quality of the prediction.

What then makes a prediction a good prediction? Generally speaking, a good prediction is one that contains all the information about the problem in question while remaining insensitive to statistical fluctuations. It should also be well calibrated, and – most importantly – it must be accurate. The statements made in the prediction must turn out to be correct when the predicted event has taken place – the method must be generalizable. It is quite astonishing that many professional predictions do not meet *any* of these requirements. Many people claim they can make accurate predictions for specific use-cases, quite often after the event happened. How does this come about? The confusion stems from misinterpreting *a posteriori* knowledge with knowledge from the time the prediction was made. This is a serious cognitive bias (probably good for survival in our evolutionary past, not optimal in our modern society[1]). Another common mistake is to over-generalize using too small a sample of historic events. It is crucial to avoid this pitfall and we routinely use methods like cross-validation, bootstrap, out-of-sample tests, event-to-event correlations and Bayesian regularization methods to avoid overtraining.

Non-Gaussian statistics: In standard university statistics courses, the root mean square deviation is presented as *the* criterion for judging prediction quality. This is correct for Gaussian residuals. This metric is popular in the academic literature and practice because Gaussian statistics are particularly easy to use, to understand and to calculate. However, our experience shows that in many real life problems, the residual distributions have significant non-Gaussian tails. We have written a book on how to evaluate predictions correctly[2], though it is currently only available in German.

Testing with billions of predictions: Testing the quality of all predictions at a large scale is an important step in ensuring that the stack of predictive technology always gives best results. At Blue Yonder, the quality of the prediction is tested by ongoing quality checks: Once the event has taken place and the "true" value or realization associated with a specific prediction becomes available, the quality of the prediction can be validated. In the meantime, billions of NeuroBayes predictions on many different research topics, projects and products have been tested *a posteriori* with a frequentist method. We continuously verify that classification probabilities are correct and mean values, as well as all credibility intervals of regression predictions, are correctly predicted.

[1] D. Kahneman, Thinking, Fast and Slow , Macmillan Us 2011
[2] M. Feindt, U. Kerzel, Prognosen bewerten, Springer Gabler 2015

Blue Yonder and Fundamental Research

The origins of Blue Yonder are rooted in fundamental research carried out at the leading high energy physics laboratories in the world including CERN (Geneva, Switzerland), Fermi National Accelerator Laboratory (Chicago, USA) and KEK (Japan). Many years of research by a large group of excellent researchers and students working with myself, Professor Dr. Michael Feindt led to the development of the NeuroBayes algorithm, which is now the cornerstone of the Blue Yonder technology stack. Once its potential for applications outside pure research was realized, a spin-off was founded at the Karlsruhe Institute of Technology (KIT), where Professor Feindt continues to hold a chair in the particle physics department. This spin-off eventually became Blue Yonder.

All rights on NeuroBayes belong to Blue Yonder. Professionalization and further development of NeuroBayes is done exclusively by Blue Yonder. However, CERN, Fermilab and KEK have been granted research licenses in order to provide further support.

Many of the data scientists at Blue Yonder have worked for CERN or other international particle physics laboratories. As a result, the stimulating international, creative, competitive and simultaneous collaborative culture of these institutions have also influenced the culture at Blue Yonder.

Blue Yonder and NeuroBayes

NeuroBayes was originally developed for analyzing data recorded by particle physics experiments. One of its earliest applications was the analysis of the b-quark-fragmentation function at the DELPHI experiment at CERN in 1999[3]. In the meantime, it has found hundreds of applications in particle physics experiments across the globe. Though it is not used by everyone in the scientific community, those who do use it find that they have an edge in the race between scientists to extract physics knowledge from the huge experimental data sets. NeuroBayes has been successfully used for improving the efficiency and resolution of complex particle detectors, for determining particle types, for optimizing reconstruction of decay chains, to distinguish quarks from antiquarks, to find top-quarks, among many other applications. Its use in many flagship analyses has led to the discovery of new particles and subtle quantum effects like the oscillation between B_s particles and their antiparticles[4].

The original NeuroBayes technology[5,6], is based on a second-generation neural network with sophisticated pre-processing of the input patterns, Bayesian regularization and pruning schemes. One of its unique features is the ability to predict complete probability density distributions for regression predictions. Unlike many other algorithms, NeuroBayes does not predict a single number (e.g. "5 apples are going to be sold tomorrow"), but rather a complete distribution that associates a probability with each possible outcome (e.g. the probability for selling 0 apples tomorrow is x, the probability for selling 1 apple tomorrow is y, the probability for selling 2 apples is z, etc). Having the full probability density distribution available allows the subsequent decision to be optimized — in the case of the apples: How many should be ordered from the wholesaler, keeping further constraints into account and optimizing the business objectives?

NeuroBayes also requires very little time for training new models and is able to discern even small effects. Since the statistical significance of each effect is tested in the training process, only relevant features are retained while the "noise" is discarded. As a consequence, NeuroBayes is immune to overtraining. Special attention has been paid to making the NeuroBayes suite very easy to work with. Features, such as the automatic determination of steering parameters and variable selection, allow the user to focus on the data and not on the tool.

Steady development: Although NeuroBayes has its origins as a "bare" neural network, continued improvements over the years have transformed it to a mature machine learning suite. It includes advanced features such as pre-processing, boosting schemes, meta-estimators and, in particular, using event weights to improve feature finding, generalization and causal inference. The Blue Yonder technology stack grows through the implementation of more and more algorithms from the literature, as well as a number of in-house developments. These are regularly benchmarked on real-world datasets, adapted for special tasks and serve as components for improving and extending the existing capabilities of NeuroBayes.

Blue Yonder Technology and Bayesian Statistics

In our opinion, Bayes' theorem is one of the most important formulae in the world. Very roughly, in one simple equation, it connects the probability of a model being correct given the observed data (the so-called posterior) to the probability of observing the data given the model (the likelihood) and the knowledge before the new data arrived (the prior). Scientific progress (progress in models) is essentially the repeated application of Bayes theorem with more experimental data (progress in likelihood). There has been an almost fundamentalist war between frequentist and Bayesian statisticians over centuries. Our approach at Blue Yonder is to take the best methods from both schools of thought and to know when to use them.

In NeuroBayes, Bayesian methods are the basis of the conditional density algorithm and important in regularization during pre-processing of the input patterns, automatic relevance determination of input variables and architecture pruning, among other processes. These methods also give rise to excellent generalization properties. When the level of statistics is too low to learn from, or when the training targets don't show statistically significant deviations from randomness, NeuroBayes tells us: "There is nothing to learn here".

[3] DELPHI Coll. (M. Feindt, U. Kerzel et al.) A study of the b-quark fragmentation function with the DELPHI detector at LEP I and an averaged distribution obtained at the Z Pole, Eur.Phys.Jour. C71:1557 2011
[4] An overview can be found at //www.neurobayes.de
[5] M.Feindt, A Neural Bayesian Estimator for Conditional Probability Densities, 2004, http://arxiv.org/abs/physics/040209
[6] M. Feindt, U. Kerzel, Nucl. Instrum. Methods A 559, 190 (2006)

A good example are next week's lottery numbers, which are completely random and cannot be predicted. NeuroBayes is not a crystal ball where we glimpse the future in bottomless swirls of white smoke. It is a scientific method and will only give a prediction, if, at the time, it is able to do better than a random guess.

This should not be compared to statements like "I knew it would happen" made after the event – laymen often mix up *posteriori* statements like this with our probability statements made before the event actually happened. NeuroBayes knows in advance the level of uncertainty associated with its prediction.

Blue Yonder Technology and Big Data

Volume: Big data, first and foremost means large volumes of data. Particle physicists have been dealing with huge amounts of data for decades, always at the edge of what the current technology would support – and sometimes a little beyond that edge. Long before the term "big data" was coined, CERN and Fermilab produced huge amounts of data that had to be read-out from sensors, reduced in volume online, reconstructed in hierarchical procedures, distributed into different channels, distributed worldwide on the GRID (from which commercial cloud computing emerged) for storage and distributed analysis by thousands of users simultaneously. Petabytes of collision data recorded by huge detectors are complemented by Monte Carlo simulations on a similar scale. We know how to manage such huge data sets efficiently and safely, and under reasonable cost-budget constraints, having developed many of the underlying technologies ourselves in our scientific careers.

Velocity: Velocity is a must-have in big data. Vertical database designs are imperative for efficient data analysis and parallelization is necessary in really big data. Both of these were around in the high energy physics community long before the advent of MapReduce. Efficient algorithms and speed-optimized programming are pivotal to the success of many projects, both in science and in business. At the LHCb experiment at CERN, 30,000 NeuroBayes instances were run in parallel and helped decide whether recorded events are interesting and stored or not interesting and discarded. For online big data prediction tasks, it may also become important for the calculation to be extremely fast.

In a cooperation between Blue Yonder and Karlsruhe Institute of Technology, the NeuroBayes expert algorithm was implemented on massively parallel hardware chips (FPGA). This will be used in the Japanese accelerator experiment Belle II, where so much raw data is produced in the sensors that it is not possible to read them out to computers at all. In total, 40 FPGA chips directly placed on the sensors each perform 2 billion intelligent decisions per second (!) to read out only those regions of the detector that contain interesting information.

Variety: Variety is important for many purposes, but not for all of our projects. Unstructured often means a complicated and dynamic structure. All machine learning algorithms need data in an interpretable form with a clear meaning, so that in the end, the complexity of the unstructured data must be understood and transformed into a simple structure, which can be managed throughout the project.

Value: It's said that there are three Vs to big data, but we like to add a fourth — Value. Among the big data hype, it is important to quickly determine the value of a proposition. We only work on big data projects where we can clearly see a value. Big, in and of itself, is not valuable.

Blue Yonder, Neural Networks and Deep Learning

Neural Networks are at the core of the NeuroBayes machine learning suite. In general, neural networks mimic the basic principles by which the human brain functions, though this does not necessarily imply that neural networks are used to mimic the brain as a whole or are used to understand consciousness and higher-level thinking. In fact, a large variety of different architectures and learning methods have been proposed over the years. After a period of hype, the last two decades saw neural networks fall in popularity compared with other machine learning methods, mainly because some of the first generation networks were highly over-trained and the predictions failed to live up to expectations. Although this was solved partly by using Bayesian methods (which are extensively used in NeuroBayes), neural networks were only rediscovered by the "mainstream research" recently, especially in the context of deep learning and artificial intelligence (AI).

Recent advances in the development of deep learning techniques have caught everyone's attention for its potential. From fully automatic speech recognition to object identification in pictures to classification of handwritten letters from pixels to defeating the world-champion in Go by an AI system developed by Google's DeepMind. However, it is important to keep in mind that "deep learning" refers to a set of methods in the broader research area of neural networks and not the recent development of a new academic field. Famously, Ronan Collobert said that "deep learning is just a buzzword for neural nets"[7]. The recent successes were made possible by solving several issues that previously made training large and multi-layered neural networks cumbersome, which often led to overtraining and poor performance of the predictions.

The NeuroBayes machine learning suite shares many characteristics of the techniques employed in deep learning applications. It is deliberately designed not to be the most general neural network that tries to mimic the human brain, but it is specialized to do numeric predictions better, faster, more reliably and with less bias than a human brain.

[7] R. Collobert (May 6, 2011). "Deep Learning for Efficient Discriminative Parsing". videolectures.net.

One example is the ability to learn hierarchical structures. For example, an image is presented to a deep learning network, which receives it as a set of pixels of varying color and intensity in its input layer. From this raw input, more complex features such as lines and then faces, followed by facial expressions need to be extracted, building a hierarchical description of the image showing a human face.

A recent and very successful application of our NeuroBayes in a scientific setting was done as an analysis of particles, so-called B mesons, which were measured at the Japanese electron-positron collider, KEKB. We used neural networks for identifying (reconstructing) the underlying signatures of these particles from the equivalent of the single pixels in a picture: Pi-0 mesons are identified from their associated photons, D mesons from "tracks" they leave behind as they traverse the complex sensors upon which a particle physics experiment is built. The next level (D* mesons) were identified from these D mesons and further tracks and finally, the B mesons were recognized from D or D* mesons that were identified in the previous step and further tracks associated to particles as they traverse the detector. All in all, 72 NeuroBayes networks were combined, automating the work previously done by highly-skilled scientists. Fully deployed, the system found twice as many "interesting" events than had been found by 400 researchers in the previous 10 years[8]. Data collected over 10 years at an international large scale experiment cost about €700 million ($756 million), so the scientific and economic value of doubling the outcome is obvious.

For the next generation experiment Belle II, we extended the system to a level that it performs the complete reconstruction completely automatically, in real-time, directly at data-taking, with a much smaller code base, faster and with even better performance.

Sometimes, though not always, it is useful to let a deep neural network learn the complete hierarchical structure starting from raw data. This has been especially successful for AI-problems that are easily manageable for humans, like image recognition and text or audio understanding. The main aim is to teach machines things humans can easily do. This will be important to improve man-machine interfaces and for automation. Usually huge CPU (or GPU) resources are needed to gain important progress when applying deep learning in these areas.

We also think that the other aspect of AI, namely to become better than the best humans in deciding in complex situations (see the AlphaGo example above), is even more attractive. Our R&D in deep learning concentrates on this area. The idea that deep networks can — in principle, with lots of computing time — learn from unrefined raw data instead of cleverly engineered features is not important here, as domain knowledge, experience and expert data scientists are in no short supply at Blue Yonder. Take, for example, our initiative OR-by-AI (Operations Research by Artificial Intelligence)[9].

Recent successes included learning a sort of "gut feeling" for the results of complicated mathematical calculations, leading to speed-up factors by several ten thousand compared to classical dynamic programming methods, even learning strategies for solving problems yet unsolvable with reasonable resources.

Blue Yonder and Reinforcement Learning

In the last decade, reinforcement learning has become another very active field of research, with impressive progress being made. The idea here is to learn the optimum course of action through repeated interactions with the environment and observation of the results. Most of the progress has been in learning deterministic relations (e.g. for robot walking), but we think about further development of these techniques in highly stochastic areas. By integrating NeuroBayes prediction technologies into agents and initialization with off-policy historical data, we expect to optimize long-term success, as opposed to short-term optimization of business processes. Combinations of NeuroBayes and deep learning ideas with modern reinforcement learning techniques is, in our view, the most promising route toward super-human capabilities in optimizing complex stochastic decision tasks, such as the OR-by-AI mentioned above.

Blue Yonder and Domain Expertise

Experience is pivotal to success. Some may hope or believe that the recent advances in machine learning and artificial intelligence will allow machines to learn everything from data itself. However, history and our own experiences tell us differently. Let's look a bit deeper into the success of Google's AlphaGo, considered one of the most advanced AI systems right now, made for the specific task of playing the ancient game of Go. Its success is attributed to several components: Development of deep learning networks, recent advances in computer hardware to power these networks, combining several deep learning networks with more traditional approaches – and human expertise of professional Go players, which has guided the development of AlphaGo.

Blue Yonder has extensive experience from past projects and implementations in many different verticals. We have collected substantial expert-level knowledge in a range of sectors, focusing on retail in particular. This allows us to build the best predictive application solutions for specific applications and implementations, tailored to our customer's individual needs. Another key to our success is our effective cooperation with sector experts on the client side, as they know the specifics and "quirks" of the systems and processes that can leverage the competitive advantage our technology offers.

[8] M. Feindt et al., A hierarchical NeuroBayes-based algorithm for full reconstruction of B mesons at B factories, Nuclear Instruments and Methods in Physics Research A 654 (2011) 432–440

[9] M. Feindt, https://www.linkedin.com/pulse/pushing-forward-artificial-intelligence-learn-how-make-michael-feindt

Resume: Why Cutting Edge AI Technology Matters

Artificial intelligence technologies today become ubiquitous "tomorrow" and often obsolete soon after — this truism of the innovation cycle is accelerated by today's omnipresence of computers, which pervade almost every aspect of our lives. Building the next generation of products to serve our customers' needs, sometimes even anticipating them, requires constant and steady progress in the development of new machine learning techniques and in the field of artificial intelligence.

In pure science – and also in enterprises – we often find that progress is rarely steady. Our in-depth experience has taught us that research communities often stick with the same standard picture and hold onto familiar tools for a long time. At Blue Yonder, we believe in constant innovation and strive for academic and operational excellence in all we do. We don't rely on mainstream consensus and instead constantly improve our skills and technology stack to meet tomorrow's challenges. Our strong scientific heritage allows us to draw the right conclusions from the data and our academic excellence is the best foundation to develop new techniques and technologies to continuously advance our products.

The aim in all of our endeavours is to create measurable and reliable value for our customers and to give them a competitive advantage far ahead of the competition – and it's tremendous fun to constantly push the boundaries of what is possible.

Blue Yonder GmbH
Ohiostraße 8
76149 Karlsruhe
Germany
+49 721 383117 0

Blue Yonder Software Limited
19 Eastbourne Terrace
London, W2 6LG
United Kingdom
+44 20 3626 0360

Blue Yonder Analytics, Inc.
5048 Tennyson Parkway
Suite 250
Plano, Texas 75024
USA