Blue Yonder Research Papers

Why cutting edge technology matters for Blue Yonder solutions

Prof. Dr. Michael Feindt, Chief Scientific Advisor

ABSTRACT

This article gives an overview of the stack of predictive technologies developed and used by Blue Yonder and compares its position to other related technologies. It explores the reasons for this position and gives some insight as to why steady development is necessary and how we achieve this.

INTRODUCTION

Blue Yonder projects and products are focussed on making predictions and decisions based on data and expert knowledge of the relevant domain. There are many machine learning algorithms on the market, and new methods are constantly being developed. In the global academic community hundreds of papers are published every day in this area.

While simple linear modelling is still used in many areas of science and in the majority of commercial applications, more sophisticated techniques are emerging rapidly. Generalized linear models, neural networks, decision trees, random forests, support vector machines, Bayesian believe networks, NeuroBayes, Big Data, 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 I want to highlight how we 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. We at Blue Yonder 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 where appropriate, to build a better *a posteriori* knowledge. We then base our decisions on the resulting predictions.

Each project may come with its own unique challenges: we may have prior knowledge about the context in which the task is set, or not. We have a clear understanding of the underlying process, or not. Our customers may help us to understand the underlying process, or not. In some cases we have a large historical dataset, in others there is only a very limited amount of data available. The data itself may be structured or unstructured. Single observations in the data may be correlated to each other or uncorrelated. The data may be made available in a single large transfer or may enter our systems as a continuous stream. Is the problem rather deterministic (e.g. letter recognition) or almost completely stochastic (e.g. the prediction of future equity price changes)? In some projects we have many more observations or recorded historic events than number of variables available to describe those events. In other cases this may be the other way around. The data may contain errors, outliers or some values may even be missing – in rare cases the "raw" data we get is in pristine condition…

Beyond these more technical questions, the fast-pace of business leads to further considerations: How great is the time pressure for model building? What are the time constraints for the predictions once the model is built?

The stochastic component of the data is a key point that is often missed; most things we want to predict actually have a large stochastic component and only a limited deterministic component, but it is only this deterministic part which forms the core of an individualised prediction. The large stochastic component on the other hand leads to an uncertainty or volatility of the prediction, but that is also something we can quantify and predict. 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).

The event or "target" we wish to predict may be a yes/no decision – an event may either happen, or not happen. In this case the answer is a probability from zero to 100%. In other cases the prediction may be of a quantity (e.g. turnover), which in the simplest case is the expectation value for that quantity. To assess the risk of a decision, a measure for the uncertainty or even the whole probability density of the target value should be predicted. At Blue Yonder this is always calculated.

It may seem that making a single prediction in one area has nothing to do with making a different prediction in another, or even in the same area. But this is rarely the case.

...BUT ON A COMMON GROUND

There are many common ingredients in different prediction tasks that actually make them more similar than one would think. Meta analysis of the huge experience gained by Blue Yonder's scientists have 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. All this combined with the in-house Data Science Academy is the basis of Blue Yonder's way of creating excellent solutions for new and demanding problems.

SUPERB TECHNOLOGY IS MADE BY SUPERB PEOPLE

All data scientists at Blue Yonder have academic degrees in physics, information science or mathematics, with years of research experience in a variety of demanding areas of academia. Many have worked, for example at particle accelerator laboratories like CERN (Switzerland), Fermilab (USA) or KEK (Japan), at cosmic ray telescopes, in quantum information theory, nanotechnology, biophysics 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 brought together at 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 in Blue Yonder much stronger than sum of its individual members.

Blue Yonder has developed into a highly attractive employer for data scientists. Our people are always looking for the best solution, eagerly taking every chance to learn more and try new techniques to improve their and their team's abilities.

TO KNOW WHAT MAKES 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 astonishing that many professional predictions do not meet *any* of these requirements, and this is often disguised by the fact that evaluating forecasts and predictions requires a high degree of expertise and skill in its own right. Many people claim they can predict this or that, mainly after the event happened – and often they really believe this. This confusion of *a posteriori* knowledge with knowledge at the time the prediction was made is a serious cognitive bias (probably good for survival in our evolutionary past, but not optimal in our modern society[1]). Another common mistake is to overgeneralize using too small a sample of historic events. This is a crucial pit-fall to avoid and we routinely use methods like cross-validation, bootstrap, and out-of-sample tests, event-to-event correlations and Bayesian methods to avoid overtraining.

Non-Gaussian statistics. In standard university statistics courses (and actually most literature) the root mean square deviation is presented as the criterion for judging prediction quality. This is correct for Gaussian residuals, as occuring in many popular toy datasets used in academic literature. However, our experience shows that in many real life problems the residual distributions have non-Gaussian tails. Here this issue cannot be neglected, and one should ask questions like "do I really pay square Euros when my prediction is not perfect?", as in the case of the root mean square deviation. The correct answer may lead to very different solutions with a much better pay-off.

Billions of predicition quality tests. At Blue Yonder the quality of the prediction is tested by ongoing quality checks: once the truth of a prediction is known after the predicted event has occurred, the quality of the prediction can be checked. In the meantime, billions of NeuroBayes predictions on many different research topics, projects and products have been tested a posteriori with a frequentist method, and not only have classification probabilities been shown to be correct, but also mean values and all credibility intervals of real valued targets.

THE CONNECTION BETWEEN BLUE YONDER TECHNOLOGY AND DOMAIN KNOWLEDGE IN DIFFERENT INDUSTRIES

Experience is pivotal to success. Since Blue Yonder already has successfully completed many different projects in many different verticals, we have built substantial expert knowledge in a range of sectors. Another important ingredient is the close cooperation with sector experts on the client side. Here we make use of the knowledge and experience of data scientists specialized in this role, who know how to ask the right questions and to communicate with sector experts without a mathematical background. Testing and possibly including existing expert models may or may not be a good starting point. We have the technology to verify existing models very quickly, and are almost always able to improve them.

THE CONNECTION BETWEEN BLUE YONDER TECHNOLOGY AND NEUROBAYES

The NeuroBayes algorithm[2][3] is the basis of Blue Yonder's technology. NeuroBayes is based on a 2nd generation neural network with intelligent and ever-developing pre-processing of the input patterns, Bayesian regularization and pruning schemes. It trains rapidly 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 against overtraining. The Expertise which holds the "essence" of the data, in the form of numerical constants, is small, and the generalization ability is outstanding.

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. Fast execution speeds for training and making prediction calculations were and are our guiding principles during its development, as well as the ability to predict more complicated quantities (such as probability densities) than almost all machine learning algorithms, which can deliver only classification or least square regression. Although there are many competing classification algorithms, NeuroBayes is still unique in its ability to predict conditional probability densities. This is most probably because too much expert knowledge and practical experience are required and the value of conditional density estimates has not yet been appreciated by the majority of scientists. Blue Yonder has many extensions up its sleeve, ready to be included as the solution enters the next stage of development.

Steady development. Many advanced improvements, boosting

[1] D. Kahneman, *Thinking, Fast and Slow* , Macmillan Us 2011

[2] M. Feindt, *A Neural Bayesian Estimator for Conditional Probability Densities*, 2004, http://arxiv.org/abs/physics/0402093

[3] M. Feindt, U. Kerzel, *Nucl. Instrum. Methods A 559*, 190 (2006)

schemes and meta-estimators that make use of NeuroBayes have been developed, in particular using event weights to improve feature finding and generalization. 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.

THE CONNECTION BETWEEN BLUE YONDER TECHNOLOGY AND CERN

NeuroBayes was originally developed for analyzing the b-quark-fragmentation function at the DELPHI experiment a CERN in 1999 (the final version is published here[4]). In the meantime it has found hundreds of applications in particle physics experiments across the globe. Though of course it is not used by everyone in the community, those who do 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 and so on. Its use has led to the discovery of new particles and subtle quantum effects like the oscillation between Bs particles and their antiparticles[5].

All rights on NeuroBayes belong to Blue Yonder, and 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 support the field further.

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

THE CONNECTION BETWEEN BLUE YONDER TECHNOLOGY AND NEURAL NETWORKS

Neural networks are the "heart" of NeuroBayes. These algorithms mimic the basic principles on which the human brain functions. 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 their expectations. Although this was solved, for example by using Bayesian methods (as in NeuroBayes), neural networks were only rediscovered by the mainstream recently, when large scale projects to understand the human brain at a much deeper level were undertaken.

NeuroBayes is not the most general neural network that tries to mimic the human brain, but one which is specialized to do numeric predictions better, faster, more reliably and with less bias than a human brain could achieve.

THE CONNECTION BETWEEN BLUE YONDER TECHNOLOGY AND BAYESIAN STATISTICS

To my mind 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) essentially is the repeated application of Bayes theorem with more experimental data (progress in likelihood). There has been an almost religious 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, for example, in regularization during pre-processing of the input patterns, automatic relevance determination of input variables and architecture pruning. 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". 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.

THE CONNECTION BETWEEN BLUE YONDER TECHNOLOGY AND DEEP LEARNING

It has recently been recognised that Deep Learning is very successful when applied to tasks like fully automatic speech recognition, object identification in pictures and classification of handwritten letters from pixels.

NeuroBayes and the surrounding Blue Yonder technology stack, as well as the term Deep Learning, all refer to sets of related methods, rather than to strictly defined algorithms. Both are under continuous development. Analysis shows that NeuroBayes and Deep Learning have many features in common, as outlined in the following:

Neural networks, not other machine learning algorithms. Ronan Collobert has said that „deep learning is just a buzzword for neural nets" [6]. Deep Learning has recently drawn great attention as a new set of methods for better training deep (many-layered) neural net-

[4] 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 20115

[5] An overview can be found at //www.neurobayes.de

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

works by pre-training weights between two adjacent layers with stochastic methods, e.g. restricted Boltzmann machines, to avoid problems that are usually associated with neural networks, such as local minima and overtraining.

In the last 15 years the majority of the machine learning community and lots of applied areas turned away from neural networks to Support Vector Machines (SVM) or Decision Trees, since they are thought to be more easily trainable and less prone to overtraining. Like the small Deep Learning community, we at Blue Yonder also did not abandon neural networks, because for most of our purposes (largely stochastic problems with small to medium predictable components in contrast to deterministically separable problems) we had already shown in 1999 that neural networks are better suited[7]. For the Blue Yonder NeuroBayes suite we have developed a large number of alternative ways to circumvent the problems commonly attributed to neural networks, e.g. intelligent automatic pre-processing, Bayesian regularization and pruning, weight boosting and inverse quantile boosting.

Learning of hierarchical structures. A second ingredient in Deep Learning is hierarchical mappings/transformations from low-level raw input to more and more abstract features (from single pixels via edges and lines to more complex ones like faces, facial expressions and so on).

This idea is not new to Blue Yonder technology; we have routinely dealt with the challenges of working with deep hierarchies of neural networks for many years. A recent and very successful application of NeuroBayes for Science was the full reconstruction of B mesons at the Belle experiment at the Japanese electron-positron collider KEKB.

Although the method is not fully automatic and no feedback loop has yet been implemented, we used neural networks for reconstructing pi-0 mesons from photons, D mesons from tracks, D* mesons from D mesons and tracks, and B mesons from D or D* mesons and tracks. In this case, the work of research physicists was automated - all necessary decisions for analyzing about 1100 different reactions were taken by 72 NeuroBayes networks. This system found twice as many fully reconstructed B-meson events than had been found by 400 physicists in the previous ten years[8].

Some other hierarchical ideas have been included in recent NeuroBayes extensions, which make the prediction model easier to understand and less like a black box.

The connection between Blue Yonder technology and Reinforcement Learning

In the last decade, reinforcement learning has become another very active field of research, with impressive progress being made. Here the idea is to learn the optimum course of action by repeated interactions with the environment and observation of the results. Whereas most of the progress is in learning deterministic relations (e.g. for robot walking), we think that further development of these techniques in highly stochastic areas, and integration of NeuroBayes prediction technologies into agents and initialisation with off-policy historical data, will open up very interesting new avenues for optimizing long term success, as opposed to short term optimisation of business processes.

THE CONNECTION BETWEEN 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 cloud computing emerged commercially) 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 parallelisation is necessary in really big data. Both of these were around in the high energy physics community long before the advent of Map Reduce. Efficient algorithms and speed-optimised programming are pivotal to the success of many projects, both in science and in the business world. At the LHCb experiment at CERN 30,000 NeuroBayes instances are run in parallel and help to 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. At Blue Yonder we are working on a R&D project to implement the NeuroBayes expert on massively parallel hardware chips.

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. We like to add a fourth V to Big Data: Value. Amongst 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.

RESUME: WHY CUTTING EDGE TECHNOLOGY MATTERS

Scientific progress is rarely steady. Our long experience has taught us that research communities often stick with the same standard picture and hold onto familiar tools for a long time, with almost no progress in the mainstream. Even in the best international research institutes, most researchers just apply what they have learned during university

[7] M. Feindt, C. Haag, *Support Vector Machines for Classification Problems in.High Energy Physics*, *CERN-DELPHI-Report 99-151*, (1999)

[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

education or – even worse – the methods taught to them by older colleagues when they first started, saying "we've always done it like that". Then, from time to time, there are sudden, disruptive changes and everybody has to learn the new dogma, which sometimes is itself not the only true approach.

No single truth. One such new dogma is that in big data you don't have to care for causality (which is very difficult), correlation is enough for prediction. While we think that this statement is largely correct for many problems (and we have been applying it successfully for long time), it is only true for systems that are not influenced by your actions. But with applications such as dynamic pricing, financial market strategies or strategies with regular customer interaction, this simple picture is wrong. Here causality matters, and in stochastic and/or complex situations with many interacting influences it is very difficult to predict how your own interaction could affect your prediction.

Sensible steady progress instead of hype jumps. At Blue Yonder we do not share the attitude described above. We never relied on mainstream or community consensus but are strong enough to break the rules and to judge for ourselves what is best for our projects – which are usually not in the mainstream. To be really successful in a competitive environment both in research as well as business, to win technology comparisons regularly and be fastest in finding hitherto unknown effects in data sets available to many competitors, one has to know about the latest technology and consider it in developing the best solutions. But one should not blindly follow hype or each new fashion. We have seen too many fads that turn out to be just that: all hype with no real value. But be assured, that at Blue Yonder you can hear buzzwords that cannot be found on Wikipedia.