AUTUMN RELEASE

Not long after being called the “sexiest job of the 21st century”, the job of data scientist may soon disappear as insiders foresee the end of the Human Data Scientist in less than 10 years.

Will Data Science be one day fully automated?

Do not look for an answer in the issue of the magazine... but you can expect some interesting insights from Sandro Saitta and Phong Nguyen on the subject.

In this release, you will also find some feedback on the Swiss Data Science conference by Olivier Zaech, as well as an article about inventory optimization from Varunraj Valsaraj and Sandro’s review of a book about forecasting.

Enjoy reading!

---

MEETUP BIG DATA ROMANDIE: GOING TOGETHER BEYOND THE BUZZ

“What does Big Data mean for you?” When asking an audience, answers rarely venture beyond volume considerations. Some will raise that large dataset have existed since the advent of scientific computing and nothing is radically new under the sky. We shall contest this latter opinion. “Big Data” is certainly a buzzword of the decade, but there are definitely some good reasons for the buzz.

Firstly, size does matter, but not only. Information systems grow in many dimensions.

**Big Data is fat data.** Once restricted to fields such as physics, bioinformatics and economy, more and more domains produce terabytes of data such as social media, IoT, web analytics etc.

**Big Data is fast data.** When data streams fast down the line, such as real time foreign exchange or public transport systems.

**Big Data is complex data.** Rich events models, evolving faster than requirements set in stone as often with classic RDBMs modeling. New approaches, such as data lakes, “schema on read” or NoSQL are coming to the rescue. Beyond the buzz. As shown in the figure below, the Big Data promise is to offer new architectural and technical paradigms to face those challenges and push the limits beyond classic RDBMS powered models.

This promise is mostly backed by fault tolerant solutions scaling out on commodity hardware with open source components.

However, the major (and frightening) obstacle faced by newcomers is the abundance of available solutions. Hadoop, to mention the most ubiquitous one, is not only a distributed redundant file system, but a complex multi purpose ecosystem. Beside Hadoop, many other database engines (columnar, graph), complex events processing frameworks, reactive models, make this jungle even richer.

**Why a Big Data Romandie meetup?**

What are those tools? Who use them around me? How to make complex visualizations? How to get my data scientist out of RStudio?

To share the knowledge, meet old timers and welcome newcomers, we have launched this monthly meetup, held in Lausanne or Geneva. Join us!

www.meetup.com/Big-Data-Romandie
WILL DATA SCIENTISTS BE REPLACED BY MACHINES?

Sandro Saitta, Data Scientist – Demand Planning, Nestlé Nespresso

Data Science automation is a hot topic recently, with several articles about it. Most of them discuss the so-called “automation” tools. Too often, editors claim that their tools can automate the Data Science process. This provides the feeling that combining these tools with a Big Data architecture can solve any business problems.

The misconception comes from the confusion between the whole Data Science process and the sub-tasks of data preparation (feature extraction, etc.) and modeling (algorithm selection, hyper-parameters tuning, etc.) which I call Machine Learning. This issue is amplified by the recent success of platforms such as Kaggle (www.kaggle.com) and DrivenData (www.drivendata.org). Competitors are provided with a clear problem to solve and clean data. Choosing and tuning a machine learning algorithm is the main task. Participants are evaluated using metrics such as test set accuracy. In industry, data scientists will be evaluated on the value added to the business, rather than algorithm accuracy. A project with 99 % classification accuracy, but that isn’t deployed in production, is bringing no value to the company.

I recently read how the winner of a Kaggle competition, Gert Jacobusse, spent his time on solving the challenge: “I spent 50 % on feature engineering, 40 % on feature selection plus model ensembling, and less than 10 % on model selection and tuning”. This is very far from what I have experienced in industry. It is usually more something like: data preparation and modeling (10 %) and the rest (90 %). I will explain below what I mean by “the rest”. When you read news about tools that automate Data Science and Data Science competitions, people with no industry experience may be confused and think that Data Science is only modeling and can be fully automated.

On my blog, I listed the different Data Science steps and discuss the ones that can be automated. Most complex and time consuming tasks such as defining the problem to solve, getting data, exploring data, deploying the project, debugging and monitoring can’t be fully automated. This is without mentioning the iterative aspect of the whole process (see the CRISP-DM figure). In a recent study from MIT, researchers said

3 See for example https://www.the-modeling-agency.com/crisp-dm.pdf
5 http://www.dataminingblog.com/can-we-automate-data-mining/
6 http://www.datanami.com/2015/10/19/machine-learning-tool-seeks-to-automate-data-science/
their tool bested more than 600 teams out of 900. What was the benchmark used? Clearly defined and closed world problem from Kaggle competitions. Such challenges don’t represent the heart of data scientists’ activities. It’s not that available tools are useless, on the contrary, they can free up time for the data scientist. Still, they don’t automate Data Science.

Don’t get me wrong: Kaggle and the likes are really good places to start learning about Machine Learning algorithms and it will certainly improve your feature engineering and modeling skills. However, you won’t learn the main aspects of Data Science within these competitions: business problem definition, data gathering and cleaning, deployment, stakeholder management, email communications, presentation skills... well, "the rest". A recent article mentions that Data Science will be automated within a few years7. Machine Learning, as defined above, can be automated to a certain extend. A good example is the meta-mining framework described by Phong Nguyen in the present issue of the Swiss Analytics Magazine. However, we are far from automating the whole Data Science process. Even for Machine Learning, we need specialists to develop new algorithms, adapted to our business challenges, people that will make the field progress. Here is an interesting metaphor, from Berry and Linoff8 relating Data Science to photography:

“The camera can relieve the photographer from having to set the shutter speed, aperture and other settings every time a picture is taken. This makes the process easier for expert photographers and makes better photography accessible to people who are not experts. But this is still automating only a small part of the process of producing a photograph. Choosing the subject, perspective and lighting, getting to the right place at the right time, printing and mounting, and many other aspects are all important in producing a good photograph.”

The main reason that makes Data Science difficult to automate is that business challenges are by definition ill-posed open world problems. To the often asked question “Will machines replace Data Scientists?”, my answer is “Yes, just after all the other jobs in the World”.

A first version of this article was published on KDnuggets (www.kdnuggets.com).

7 http://www.kdnuggets.com/2015/05/data-scientists-automated-2025.html
8 Explained by David S. Coppock
I have recently visited the 3rd Swiss Conference on Data Science and I would like to share some insights on the sessions I attended.

Please note that this post reflects my personal view on the event. For any mistakes or misunderstandings I am the one to blame ;)

The conference was organized by several members of the ZHAW Datalab organization and attracted more than 220 attendees this year. The SDS event turned from a small workshop to a real Swiss-wide data science conference this year.

Main sponsors were Google, SAP, Microsoft Partner, pwc, Zühlke.

Keynote on “So you think you have all the data ? Causes and consequences of selection bias”

David Hand, Imperial College

Very good and lively talk about the issue of selection bias when preparing data for analysis. He suggests to always watch carefully when you are selecting (sample) data for analysis. Check the data for consistency, bias by selection, completeness, asymmetric data selection -> selection distortion. Revealing question to ask a data scientist: “How do you handle missing data?” If there’s none – the data might not be complete or badly selected, i.e. it might not reflect the “real world”. Always be clear when looking at data coming from (web) polls, phone polls, twitter, etc. Do they really reflect a trustworthy sample? What was the whole population?

Links:
- Homepage: http://www.imperial.ac.uk/people/d.j.hand
- Interview about Open Data: https://www.theguardian.com/society/2012/jul/10/open-data-force-for-good-risks

Keynote on “Big Data Integration”

Philippe Cudré-Mauroux, University of Fribourg

How can structured and “semi-structured” data be combined for analysis, i.e. looked up / searched for together?

Answer: Use “entities” as “mediation”. Tag the entities with identifiers, then connect the entities using the identifiers. This will then result into a “Knowledge Graph”. This would then be identity-centric data/information management.

Steps of the approach:

Step1: named entity recognition: lemmatization > list of n-grams > n-gram index > inverted n-gram index

Step2: entity linking: using crowd sourcing (“human computation”) > sending small questions to the crowd and collect responses (tool: amazon’s “Mechanical Turk” platform)

Step3: entity typing: entities can have many types or “facets”, they use up to 5000 types in an entity

Step4: co-reference resolution: consolidate human responses using calculated models

Service using this approach & technology: http://sciencewise.info/ a search engine for searching scientific articles.
“Sentiment Analysis - State-of-the-Art in Research and Industry”

Mark Cieliebak, ZHAW School of Engineering

Text base categories: text groups (corpora), large texts, documents, single sentences

Sentiment tags used: +, -, neutral, mixed

A survey showed that sentiment analysis (SA)-tools can deliver 60%-70% of accuracy, depending on text base. Some tools are better in specific text types (news, tweets, large texts).

Insights from a SA-competition in 2013: mainly feature based approaches were used, best in class: Sentigem. In 2015 deep learning approach could improve results to 63%-75% accuracy. Approach that was used: 3-phase distant supervision + 2-layer convolutional neural network (CNN) combined with “word embedding”. Word embedding serves as a non-linguistic i.e. statistical approach to text analysis [see “semantic differential” (Osgood) for a linguistic approach]. It converts words into vectors that “preserve the contextual similarity of words” so they can be processed by algorithms, e.g. machine/deep learning. Cieliebak: “Of course adding a lexicon would improve accuracy a lot – but you don’t want to create a lexicon for every language.” Having a linguistic research background I was quite surprised how the results from linguistic research is simply dropped…

Just do it:

- Natural Language Toolkit NLTK (python library): http://www.nltk.org/
- Stanford NLP Group’s toolkits (Java): http://nlp.stanford.edu/software/
- Apache OpenNLP: https://opennlp.apache.org/
- The amazing power of word vectors
- Introduction to Distributed Word Vectors (kaggle.com course)
- Semantic Differential: http://www.semanticdifferential.com

“Predictive Analytics on Big Data platform with Native Spark Modeling”

Priti Mulchandani, SAP Switzerland AG
Andreas Forster, SAP Switzerland AG

A demo of SAP’s web frontend using native Spark modeling in the background was given.

See this blog entry that explains the concept behind the scenes: http://scn.sap.com/community/predictive-analytics/blog/2016/03/18/big-data-native-spark-modeling-in-sap-predictive-analytics-25

“Unlocking the business value of publicly available geo-data”

Sotiris Dimopoulos, Switzerland IT-Logix AG / ETH Zurich

How can publicly available geo-data and demographic data be used to develop business cases and tools?

Swiss public data sources:
- OpenData.swiss - Portal of the Swiss Open Government initiative
- Statistisches Lexikon der Schweiz
- Swiss Federal Statistical Office: Swiss Statistics (FSO, BFS)
- swisstopo
- geo.admin.ch
- Federal Roads Office (FEDRO, ASTRA)
- Bundesamt für Raumentwicklung ARE
- Federal Office of Public Health - The FOPH
“From Idea to Data Product”

Gian-Marco Baschera, Switzerland Zühlke Engineering AG
Sandro Strebel, Bring ! Labs AG

The presentation explained how the mobile shopping app “Bring!” is being developed further using analytics on customer data. Zühlke Consulting did some customer data analytics that now are being turned into new features of the app: shopping basket analysis, seasonal shopping preferences, next shopping suggestion, etc.

The project was done using R, R Shiny, Docker, MySQL, AWS.

Keynote on “Data Preparation - The Key to Successful Data Science”

Lars Grammel, Trifacta

The presentation was about Trifacta’s tool “Trifacta Wrangler” that eases the data preparation task in a new innovative way. A subset of the data is presented in an EXCEL-sheet like view with basic statistical measures attached at the top of each row. Then the data scientist can select and filter the rows that should go into the analysis. The selection directly generates the necessary code to bring the data into the desired format. This code is then submitted to the whole data set sitting on Hadoop.

There is a free (limited) single user version of the tool available.

Keynote on “Deep Learning RNNaisance”

Jürgen Schmidhuber, IDSIA

Probably one of the most prominent researchers in AI. Since 1987 Schmidhuber has published numerous articles on AI, neural networks, robotics, genetic programming, etc.

He heads The Swiss AI Lab IDSIA (Istituto Dalle Molle di Studi sull’Intelligenza Artificiale) a lab where many of today’s standard algorithms in deep learning have been developed, “with the funding of the Swiss tax payers” as Schmidhuber mentions. Companies such as IBM, Apple, Google, Microsoft and many others are using these algorithms for free in their AI offerings.

The inspiring talk ended with the outlook that in 2050 robots (autonomous machines) will be developed that surpass the capacity of the human brain. This will then be the starting point where robots (autonomous machines) will develop themselves further, leaving mankind behind in terms of brain capacity. They will populate the universe from then on. The place for mankind is not clear then, but safe as Schmidhuber notes, because robots will not be interested in “dumb” humans...

If you want to dig more into the topic of AI, have a look at Schmidhuber’s homepage and the articles below.

Links:
- Wikipedia entry: https://en.wikipedia.org/wiki/J%C3%BCrgen_Schmidhuber
- Homepage with all the articles and topics: http://www.idsia.ch/~juergen/
- Articles /Interviews (german):
  - « Der Mensch wird keine dominante Rolle mehr spielen »
  - Wie Google in Zürich Computern das Denken beibringt
  - Von Ameisen und Übermenschen
  - Hier spricht ein Roboter
  - DAS DUELL
- Also worth reading:
  - Computer Power and Human Reason: From Judgment to Calculation
  - Mar 1976, by Joseph Weizenbaum
  - German edition: Die Macht der Computer und die Ohnmacht der Vernunft
  - Understanding Weizenbaum: The Danger of Artificial Intelligence
  - Apr 1, 2015, by Hercules Bantas
A META-MINING SYSTEM TO SUPPORT DATA MINING WORKFLOW PLANNING AND OPTIMIZATION

In this article, we will describe a system which combines data mining (DM) workflow planning and meta-mining in view of designing DM workflows which are expected to achieve good performance on a given mining problem. Our meta-mining framework, together with an initial sketch of our planning system, were already described in the last issue of the Swiss Analytics Magazine. Here we will describe in more details the different components of the system and give experimental results.

The system with its components is illustrated in Figure 1. As first component, we have the Data Mining Experiment Repository (DMER) which contains all past DM experiments and meta-data (MD) from which we learn heterogeneous similarity measures, associations between dataset and workflow descriptors that lead to optimal performance, according to our method described in Nguyen et al. (2012b). Then we have as second component the Meta-Miner that exploits the learned associations to guide the AI-planner in the workflow construction during the planning of DM workflows for the input dataset. This last one is provided by the User Interface, our third component, together with a DM goal such as classification. The Meta-Miner and the AI-planner then tightly interact together to form our IDA (fourth component) where at each planning step the planner submits a set of partial candidate workflows that are ranked by the Meta-Miner in order to prioritize those workflows that are expected to deliver the best performance on the input dataset. Finally in the fifth component the system delivers after planning a set of optimal plans ranked by order of their expected performance from which the user can select which one to apply on her dataset.

HTN planning

In this section, we will describe the hierarchical task network (HTN) planner we use in our system. Given some goal $g \in G$,
the AI-planner will decompose this goal in a top down manner into elements of two sets, tasks $T$ and methods $M$. For each task $t \in T$ that can achieve $g$, there is a subset $M' \subseteq M$ of associated methods that share the same data input/output (I/O) object specification with $t$ and that can address it. In turn, each method $m \in M'$ defines a sequence of operators, and/or abstract operators (see below), and/or sub-tasks, which executed in that order can achieve $m$. By recursively expanding tasks, methods and operators for the given goal $g$, the AI-planner will sequentially construct an HTN plan in which terminal nodes will correspond to operators, and non-terminal nodes to HTN task or method decompositions, or to dominating operators ($X$-Validation for instance will dominate a training and a testing sub-workflows). An example of an HTN plan is given in Figure 2. This plan corresponds to a feature selection and classification workflow.

The sets of goals $G$, tasks $T$, methods $M$ and operators $O$, and their relations, are described in the DMWF ontology (Kietz et al., 2009). There, methods and operators are annotated with their pre- and post-conditions so that they can be used by the AI-planner. Additionally, the set of operators $O$ has been enriched with a shallow taxonomic view in which operators that share the same I/O object specification are grouped under a common ancestor called an abstract operator, i.e. an operator choice points among a set of syntactically similar operators. For example, the abstract AttributeWeighting operator given in Figure 2 will contain any feature weighting algorithms such as InformationGain or ReliefF, and similarly the abstract Predictive Supervised Learner operator will contain any classification algorithms such as NaiveBayes or a linear SVM. Overall the HTN grammar contains descriptions of 16 different tasks and more than 100 operators, over which the planner can plan.

**Workflow Selection Task and Planning Strategies**

To support the user in the selection of DM workflows, we adopt a heuristic hill climbing approach to guide the planner. At each abstract operator we need to determine which of the $n$ candidate operators are expected to achieve the best performance on the given dataset. This amounts to determine the performance of the partial workflows build so far with the candidate operators concatenated at the end of these workflows. Thus, at each planning step $s$, we can build the set $S_s$ of $k$ selected candidate workflows according to:

$$S_s = \arg\max_{x} w | g$$

![Figure 2: HTN plan of a feature selection plus classification workflow](image-url)
where \( x \) is a vector description of the input dataset according to various dataset characteristics typically used in meta-learning, \( ws \) is a binary vector that provides a propositional representation from to the DMOP ontology (Hilario et. al, 2011) of each of the \( k \) candidate partial workflows, and \( g \) is the data mining goal we want to address. Finally, at the end of planning, the system delivers the \( k \) candidate workflows that are expected to deliver the best performance on the input dataset.

To estimate the expected performance \( y \) of the candidate workflows, we use the homogeneous and heterogeneous dataset/workflow similarity measures of Nguyen et. al, 2012, from which we define two planning strategies. The first planning strategy which we call P1 uses before planning the dataset homogeneous metric to determine the similarity of the input dataset \( x \) to each of the training datasets in our DMER. Then, during planning we determine at each planning step the similarity of each candidate workflow \( w \) to each of the training workflows in our DMER using the homogenous workflow metric. Finally, we estimate the expected performance \( y \) of each candidate workflow \( w \), through a weighted average of the performances of past similar DM experiments in a similar manner to the works in practical reinforcement learning in continuous spaces (Forbes and Andre, 2000). The second planning strategy which we call P2 is more straightforward as it can directly use the heterogeneous metric between dataset and workflow to estimate the expected performance \( y \) of each candidate workflow \( w \).

**Experimental Results**

We evaluate our approach on the data mining task of classification. To train and evaluate our approach we have collected 65 benchmark classification datasets representing genomic or proteomic microarray data and coming from the same problem type, namely cancer diagnosis and/or prognosis. We have applied on them 35 data mining workflows composed of 28 feature selection plus classification workflows, additionally with seven classification-only workflows, and we evaluated their performance with ten-fold cross-validation. We used four feature selection algorithms: Information Gain, Chi-square, Relieff, and recursive feature elimination with SVM; we fixed the number of selected features to ten. For classification, we used seven algorithms: one-nearest-neighbor, the C4.5 and CART decision tree algorithms, Naive Bayes, logistic regression and SVM with the linear and the RBF kernels.

We explore two distinct evaluation scenarios. In the first one, we constrain the system so that it plans DM workflows by selecting operators from a restricted operator pool, namely operators with which we have experimented in the base-level experiments. In the second scenario we allow the system to also choose from operators with which we have never experimented but which are described in the DMOP ontology. These additional algorithms are one feature selection algorithm: Information Gain Ratio and four classification algorithms: a Linear Discriminant Analysis algorithm, a Rule Induction algorithm, Ripper, a Random Tree algorithm, and a Neural Network algorithm. The total number of possible workflows in this setting is 62.

As baseline methods, we have used a standard meta-learning approach that uses the Euclidean similarity metric over the dataset characteristics – we call it \( \text{Eucl} \) – to select the N most similar datasets to the input dataset \( x \) and simply ranks the workflows according to the average performance they achieve in this neighborhood. The second meta-learning method that we call \( \text{Metric} \) uses the homogeneous dataset metric of Nguyen et. al, 2012, and ranks the workflows similarly to the \( \text{Eucl} \) baseline method.
Finally, as third baseline, we have used a default recommendation strategy, Def, that simply averages the performance of the workflows in the training datasets of our DMER.

To estimate the performance of the planned workflows in both evaluation scenarios, we use leave-one-dataset-out. We evaluate each method by measuring how well the estimated list of top-k ranked workflows that the method delivers on a given dataset correlates with the “true” list of top-k ranked workflows. We use the Kendall distance with penalty $p=1/2$ of Fagin, Kumar, and Sivakumar, 2003. This metric gives the number of exchanges needed in a bubble sort to convert one list to the other. We furthermore normalize and inverse this metric to get the normalized Kendall similarity metric and finally we compute the Kendall gain $Kg$ of each method by dividing the Kendall similarity metric of a given method by the Kendall similarity metric of the default recommendation strategy.

Figure 3 gives the results of the Kendall gain $Kg$ for the two scenarios describe before where $k$ is the number of selected workflows. In the first scenario (Figure 3, left), we can see that the P2 method which uses the heterogeneous metric between datasets and workflows is the best-performing method with the main improvement being on the top-10 selected workflows. The P1 and Metric methods deliver similar performance as both use the homogeneous similarity metrics. Finally the standard meta-learning approach performs correctly on the top-10 selected workflows but its performance degrades significantly after. In the second scenario where we evaluate the use of unseen operators (Figure 3, right), we can see that the two proposed approaches P1 and P2 both outperform the default baseline. Recall that this baseline amounts to execute all the possible workflows and average their performance, i.e. we have executed all the 62 possible workflows on the 65 datasets to get this baseline, while the P1 and P2 strategies solely rely on the execution of the original set of 35 workflows on our training dataset. Thus, our planning strategies are able to deliver better performances with less workflow executions than the default brute-force approach.

**Conclusion**

In this article, we have described a novel planning system able to automatically plan new data mining workflows for a given dataset. The system can design workflows that deliver good performance, at least better than the standard meta-learning approach and the default brute-force approach of executing all possible workflows and ranking them by their overall performance. More details on the system together with the proposed Meta-Mining framework can be found in Nguyen et. al, 2014.

**About the Author**

Phong Nguyen holds a PhD in machine learning and data mining from the University of Geneva. He is currently...
Senior Data Scientist in Expedia working on sort algorithms and recommender systems.

Bibliography


Governments and businesses are expecting the Internet of Things (IoT) to hit the mainstream by 2020.

Early adopters have already gained some experience. 75 teams were interviewed on recent experiences across all industries.

4 KEY LESSONS

- **SCOPE** is almost always **UNDERESTIMATED**
- **USER EXPERIENCE** defines **RESULTS**
- **DATA & GOVERNANCE** skills are **CRITICAL**
- **MATURITY** dictates **VALUE**

DIVERSE EXPECTATIONS

- **43%** Improved operational efficiency
- **36%** Improve user experience
- **29%** New product or service design
- **25%** Improve resource management

3 MOST POPULAR LESSONS

- **APPLY DESIGN THINKING**
- **TEST IN ADVANCE**
- **PLAN FOR SCALABILITY**

CLOSING THE SKILLS GAP

- Storytelling/data flow visualisation
- Interpreting results
- Analysing data
- Critical thinking/pragmatic data scientist

Read the full report at [www.sas.com/iotebook](http://www.sas.com/iotebook) to learn more.
MULTI-ECHelon INVENTORY OPTIMIZATION AT A MAJOR DURABLE GOODS COMPANY

Varunraj Valsaraj, Sr. Operations Research Specialist at SAS

Multi-echelon inventory is ever more a requirement in this era of globalization, which is both a boon and bane for manufacturing companies. Global reach allows these companies to expand to new territories, but at the same time increases the competition on their home turf. Consider General Motors, which faces new competition at home from Asian manufacturers such as Hyundai and Kia but has also benefited by global expansion. In 2015, they sold 63% of their vehicles outside North America. In fact, GM and their joint ventures sold as many vehicles in China as in North America. The supply chain of most manufacturing companies has also expanded globally, because most of them source their raw materials from different parts of the world, making their supply chain more complex, which is why a holistic approach to optimize the entire supply chain is critical. However, customers still expect the same great service despite the complexity, which is why multi-echelon inventory optimization can provide real benefits to global companies.

A major durables goods company in this situation partnered with SAS to leverage our advanced analytics capability to right-size their inventory. SAS® Demand-Driven Planning and Optimization provided this company with a structured and efficient process to right-size the inventory in their complex, multi-echelon supply chain network. The SAS inventory optimization process consists of four steps:

In the first step, we identify the probability distribution that best fits the forecasted demand. Estimating the right probability distribution playing a critical role in determining the amount of inventory investment required to cover uncertainties in demand. SAS Inventory Optimization Workbench (IOW) support a number of continuous distributions, such as normal, log normal and discrete distributions like Poisson and binomial. One of the key differentiators of SAS IOW is its ability to handle intermittent demand. In most supply chains, 80% of items account only for 20% of the total demand. There are lot of items with low demand and high variance. IOW uses a specialized technique called Croston’s method to handle these intermittent demand items. In Croston’s method, demand is estimated through exponential smoothing and the interval between demands is estimated separately. In addition we utilize historical performance to estimate the right amount of safety stock that would satisfy the service level requirements.

The second step is to calculate the inventory target or order-up-to-level. The inventory target consists of three components: pipeline stock, cycle stock, and safety stock. Pipeline stock is the amount of inventory required to cover the demand during lead time while cycle stock is the inventory required to cover the period between replenishment. Safety stock is the amount of inventory required to cover the uncertainty in demand. The variance of forecast data and the desired

[Diagram: Automated Demand Estimation, Inventory Target Calculation, Monte Carlo Simulation, Visual Report Update]
service level for a given product-location pair are the main drivers for the safety stock calculation.

In the third step, we simulate 200 scenarios of forecasted demand over the horizon and calculate KPIs, such as on hand inventory, service level, lost sales, and replenishment orders using Monte Carlo Simulation. In the final step, we update the Visual Analytics report using the output from the Inventory Optimization Workbench. Users can easily visualize inventory target, forecasted demand, and projected orders over time and use these reports to spot any outliers.

Inventory is a significant investment for every company. So it is critical for supply chain team to estimate the benefits from inventory optimization so that they can set the right expectation with their management team. In addition there are many parameters that need to be fine-tune periodically to optimize the system performance. SAS team developed a new automated simulation-optimization approach called Tuning and Validation to overcome these challenges. In the first step, called tuning, we use historical data to automatically calibrate the parameters of the inventory optimization. In the second step, we simulate using the optimized policy and compare it to the historical performance to quantify the improvements in KPIs such as inventory cost, lost sales, service level etc. as a result of inventory optimization.

**Conclusion**

The durable goods manufacturer saw immediate benefits after implementing the Inventory Optimization Workbench. The service level improved from 65 % to 92 % and lost sales reduced from 8 % to 2 % in just 10 months after implementation. They can monitor their performance in real-time with visualization. Implementing the Demand-Driven Planning and Optimization solution they have an enhanced, integrated platform for demand planning and multi-echelon inventory optimization, eliminating reliance on multiple Excel-based processes.
BOOK REVIEW:
“SUPERFORECASTING: THE ART AND SCIENCE OF PREDICTION”, SUPERFORECASTING

Book review by Sandro Saitta, Data Scientist – Demand Planning, Nestlé Nespresso

Superforecasting – by Tetlock and Gartner – explains the huge study performed by Tetlock about the ability of people to predict future events (mainly geo-political). The closed questions (i.e. choose between yes/no) are far from real numbers you will predict in business forecasting. Tetlock discusses skills that have been identified as driving accurate forecasts. The point of the authors is that forecasting is a skill which can be improved. Superforecasters are accurate thanks to various good practices: they split the problem, update their forecast frequently, are objective, learn from their past errors, etc.

To be fair, the book title should be “Superforecasters”. Indeed, the book is about these common people that have the ability to forecast more accurately than the rest of the crowd, and even experts. This also means that the book is not about forecasting. It doesn’t cover any topic related to forecasting, in the sense of time series analysis. (ESM, ARIMA, etc.) nor business forecasting (forecast horizon, update frequency, etc.)

Tetlock and Gardner cover a wide range of topics in their book. They discuss the imprecision in language when discussing forecasts, which make them ambiguous to verify. Even when using precise numbers, forecasts are often misinterpreted. An excellent example is the “70 % chance of rain tomorrow”. When it is not raining the day after, people believe the forecast was wrong. Such conclusions are not valid when looking at only one event.

I wasn’t a big fan of Chapter 10 which was too historical. Beside this chapter, the book is a real pleasure to read. In conclusions, Tetlock and Gartner book is an excellent summary of Tetlock’s comprehensive research on the people ability to forecast. Not related to business forecasting, the book explains the skills you need to have to correctly answer geo-political questions. Superforecasting is a must have to understand psychological aspects behind (good) decision making and issues related to sharing and understanding forecasts.
Big Data Analytics Forum 2016

November 30th 2016, Zurich
SUPPORT THE SWISS ASSOCIATION FOR ANALYTICS!

If you like our association and would like to contribute, you can write an article, be a speaker at one of our events or start a discussion our LinkedIn group.

https://www.linkedin.com/groups/4586163/profile
E-mail: info@swiss-analytics.com
www.meetup.com/swiss-analytics

As a non-profit organization, we would also welcome any donation. Your participation will help us cover the costs related to event management or magazine printing.

Our bank account is:

- **Name:** Swiss Association for Analytics
- **Address:** Chemin des Fontannins 12, 1066 Epalinges
- **Bank:** PostFinance
- **Account:** 85-122140-9
- **IBAN:** CH65 0900 0000 8512 2140 9
- **BIC:** POFICHBEXXX

Thanks for your support!

---

**IMPRESSUM**

The Swiss Analytics Magazine is proposed by the Swiss Association for Analytics (www.swiss-analytics.com). The magazine is printed by Gasser Media (www.gasser-media.ch). Articles are written by experts in analytics. The views expressed in this magazine do not necessarily represent the views of the Swiss Association for Analytics. It is not allowed to reproduce, duplicate, copy, sell, resell or exploit any portion of the magazine without agreement of the Swiss Association for Analytics (info@swiss-analytics.com).

Printed in October 2016, Le Locle, Switzerland
LE MEILLEUR DE VOS DONNÉES!
COLLECTER, ORGANISER, DIFFUSER.

ÉDITION
GESTION DE PROJET
DEPUIS 2001

IMPRESSION
PRODUCTION À LA DEMANDE
DEPUIS 1947

PUBLICATION
WEB, APP, CROSS-MEDIA
DEPUIS 2004

ADRESSE
Gasser Media SA
Jambe-Ducommun 6a
CH-2400 Le Locle

CONTACT
Tél. +41 32 933 00 33
info@gasser-media.ch
www.gasser-media.ch

Pour suivre nos actualités :
www.gasser-media.ch/blog