

Social Analytics role in high-tech business

A Master's Thesis submitted for the degree of
“Master of Business Administration”

Supervised by
Prof. Robert D. Hisrich, PhD

Lucian Viorel Stoenescu

h1154082

Vienna, 27th of June 2013

Affidavit

II, **LUCIAN VIOREL STOENESCU**, hereby declare,

1. that I am the sole author of the present master's thesis "**Social Analytics role in high-tech business**", 73 (core) pages / out of 96, bound, and that I have not used any source or tool other than those referenced or any other illicit aid or tool, and
2. that I have not prior to this date submitted this master's thesis as an examination paper in any form in Austria or abroad.

Date

Signature

Preface

This paper has the following purposes:

- To explain the need of social analytical solutions in high-tech business and in particular in Telecom domain, to offer solutions providing a continuous contextual insight of customer behavior, usage patterns, circumstances (location, influencer circles) combined with predictive modeling in order to help Customer Service Providers / CSPs maximize their revenue by providing targeted services to the right subjects at the right moment
- Research and describe alternative analytics solutions based on predictive analytics (open source and commercial enterprise versions) as opposed to classical descriptive (business intelligence reporting types); conduct an empirical benchmarking test on an open-source solution (RapidMiner) to emphasize the criteria for selecting the best solution suitable for specific target users (professionals, business users)

In the *first chapter* we will find a description of both operational information as base for descriptive analytics and of strategic information type that is the outcome of predictive analytics solution. The need for predictive analytics based solutions is complemented with cross-industry implementation examples.

Second chapter presents the predictive tools spectrum and the PMML standardization used to split the predictive model creation and consumer part.

Third chapter describes a typical Telecom data flow, the Global Mobile Market Share as business market landscape and the overall NSN-Comptel solution analytics (including its unique selling proposition and deployment types).

Fourth chapter exhibits a brief business opportunity assessment (including value proposition for pre/post-paid mobile telecom cases).

Chapter five contains the conclusions (as well as a checklist for selecting the best fit solution analytics / predictive analytics based solutions) and further interesting topics where the above mentioned can play an important role.

Abstract

Social Analytics in high-tech business

Business in “high-tech” domains is especially recognized for its “marketing uncertainty” and “competitive volatility”. To overcome these problems, firms should use solutions that switch attention from the “operational information” to “strategic information” type, the outcome of predictive analytics solutions based on data mining technologies and algorithms.

Predictive analytics based tools (including social analytics) are soon-to-be a “must have” asset and moreover integrated in a Decision Management System, the key to an optimized business decision.

The current paper includes the presentation of predictive analytics tools spectrum, their main features/ add-ons including “R” environment integration, Predictive Model Markup Language Input / Output capabilities or their modular architecture platform independent and data processing capabilities (such as shared nothing parallel processing).

The paper highlights the benefit of integrating a Social Analytics solution based on predictive modeling applied in the high-tech Telecom domain. While open-source tools are mainly dedicated to professionals requesting data mining process knowledge (exemplified in a benchmarking test containing data from a Telecom prepaid Mobile Virtual Network Operator (MVNO)), the NSN-Comptel Social Analytics solution is addressed to business users, emphasizing the social networking role in mobile subscribers’ churn decision or in their adoption of new products / services.

Additionally, the paper encompasses a business opportunity assessment based on NSN-Comptel Social Analytics solution (embedded on NSN’s Customer Experience Management on Demand portal solution deployed on public Cloud as Software as a Service) for a MVNO prepaid provider.

Regarding future prospects I listed three important topics for predictive / social analytics:

- On-The-Fly computing concept having similarities with XaaS (anything as a Service) used in Marketplace portal creation for Telco, Enterprise, Machine-To-Machine Apps;
- Wire-line / fixed operator data mining to prepare customer oriented B2C applications;
- Photo and video data mining, with huge applicability even in governmental institutions

Keywords: *strategic information, data mining, predictive analytics, social analytics, decision management systems*

1 TABLE OF CONTENT

1. Literature part.....	1
1.1. Strategic information versus operational information	1
1.2. <i>WHY</i> the need for an appropriate analytics solution? Opportunities in different industries due to the use of strategic information; cross-industry market landscape	6
1.3. Social analytics solutions based on predictive analytics as important cross-industry business growth driver	14
2 Predictive analytics related (types, tools, data mining algorithms, PMML standardization for import / export predictive models).....	18
2.1 Types of analytics (demystifying <i>descriptive</i> and <i>predictive</i> analytics).....	18
2.2 Predictive Analytics tool and solution spectrum (in-house proprietary solutions developed using C++ / Java / .NET technologies versus open-source solutions; predictive model creation and consumer separation using PMML (Predictive Model Markup Language), an xml based markup language technology)	24
2.3 Brief description PMML components; PMML version 4.1 and supported data mining model (Decision Trees; Support Vector Machines; Neural Networks; Naïve Bayes; Regression; Scorecards; K-Nearest Neighbors (KNN); Clustering Models; Association Rules)	33
3 Implementations in Telecom domain (Social Analytics as stand-alone or in a complex Customer Experience Management solution).....	40
3.1 Brief mobile technology evolution description	40
3.2 Data flow in an End-to-End Telecom business (from raw <i>data collection</i> , including aggregation and transformation → <i>insight</i> (reports, dashboards) → <i>Analytics</i> (predictive modeling) → <i>Actions</i> (e.g. business decisions, marketing targeted promotions).....	42
3.3 <i>HOW</i> - Overall solution architecture presentation (e.g. NSN, IBM, KNIME, Rapid-I solutions – churn prediction MVNO /empirical tests); present NSN Cloud-based Social Analytics solution included in Customer Experience Management on Demand (CEMoD) portal –offered as SaaS (Software as a Service)..	46
3.4 Solution Description and Unique Selling Proposition (NSN-Comptel Social Analytics).....	54
3.5 Social Analytics Basic modules (use-cases).....	56
3.6 Measuring the prediction accuracy (some point in time after the new service launch in the market)...	58
3.7 Deliverables (Social Analytics outputs)	58
3.8 Solution deployment (cloud-based approach, Software as a Service / SaaS).....	59
4 Opportunity Assessment - “Who and How?”	61
4.1 Target Model	61
4.2 Business Model and Initial Customers	64
4.3 Future Expansion.....	68
5 Conclusion.....	68

5.1	Emphasize the Telecom / Enterprise Customer Service Provider benefits of investing in a social analytics (predictive analytics based) solution	68
5.2	Check-list for selecting the best fit predictive / social analytics tool	70
5.3	Future developments	71
6	Bibliography.....	74
	Appendix_A	78
	Appendix_B	79
	Appendix_C	80
	Appendix_D	81
	Appendix_E.....	82
	Appendix_F.....	83
	Appendix_G	84
	Appendix_H	85
	Appendix_I.....	86

Table of figures

Figure 1 Data Warehouse 2.0 concept.....	3
Figure 2 Corporate Information Factory	3
Figure 3 Business Intelligence evolution towards Analytics.....	4
Figure 4 Predictive Analytics tool spectrum	6
Figure 5 RapidMiner versions and pricing structure (open-source vs. Enterprise) w/o support and training	7
Figure 6 Need for Predictive Analytics solutions to strengthen Enterprises' business - as per IBM's concept.....	8
Figure 7 Predictive Analytics TDWI survey	8
Figure 8 Predictive Analytics World, Survey Results Jan 2009.....	13
Figure 9 TDWI Survey - Predictive Analytics tool adoption, Aug 2006	14
Figure 10 Social Analytics and Cloud Parallel Processing as big expectations in a 2-5 year time frame, Gartner July 2012	16
Figure 11 Big Data priority matrix - Gartner. July 2012.....	16
Figure 12 Big Data Opportunity Map by Industry, Gartner, July 2012	17
Figure 13 Informs - Business Intelligence / Business Analytics Venn diagram	20
Figure 14 CRISP-DM process diagram.....	20
Figure 15 PMML Components.....	34
Figure 16 Decision tree used to optimize an investment portfolio (bold lines mark the best selection)	37
Figure 17 SVM Hyper planes (H1 does not separate the classes. H2 does, but only with a small margin. H3 separates them with the maximum margin); Wikipedia SVM	38
Figure 18 Model ensemble - scores from all models are computed and the final prediction is determined by a voting mechanism or the average	39
Figure 19 3GPP mobile technology evolution	40
Figure 20 Global Mobile Market Shares; Informa Telecoms & Media, WCIS+, Q1 2013	42
Figure 21 High Level Definition of a Telco environment.....	43
Figure 22 Data Flow in NSN CEM solution	44
Figure 23 Data Mining methods / Spring 2005	45
Figure 24 IBM SPSS Modeler - Predictive in 20 min - training VM screenshot.....	47
Figure 25 KNIME Churn Prediction sample.....	48
Figure 26 KNIME decision tree learning and scoring.....	49
Figure 27 Subscriber churn prediction for an Italian MVNO - using RapidMiner	50
Figure 28 Process design - Churn Prediction model using Decision Tree.....	51
Figure 29 Validation operator change from split to cross type	51
Figure 30 Predictive model accuracy (using boosted SVM PSO algorithm)	53
Figure 31 Boosting the SVM PSO algorithm - Process design.....	53
Figure 32 EMC Greenplum MPP Shared-nothing architecture.....	55
Figure 33 Social Analytics - social network role in predictive modeling	56
Figure 34 NSN Customer Experience Management Portal.....	59
Figure 35 NSN Anything as a Service (XaaS) High Level Architecture using NSN Cloud GW for package virtualization and installation	60

Figure 36 Eurostat - Telecommunication services: Operators and service providers CEE.....	62
Figure 37 MVNO –MNO relationship; Source Nereo Consulting.....	62
Figure 38 MVNO classification - Nereo Consulting.....	63
Figure 39 Eurostat - Market Shares in Telecom / CEE	63
Figure 40 On-site Social Analytics deployment - Pricing Structure.....	64
Figure 41 NSN-Comptel Social Analytics modular use cases	65
Figure 42 Social Analytics extended use cases	65
Figure 43 CSP's revenue boost due to Churn Prediction and Campaign uptake.....	67
Figure 44 On-The-Fly Computing - Project development phases, Source HNI 2011.....	72
Figure 45 The Forrester Wave™: Big Data Predictive Analytics Solutions, Q1 2013	78
Figure 46 Open source vs. commercial solutions ranking (market presence and strategy) - The Forrester Wave™: Big Data Predictive Analytics Solutions, Q1 2013	79
Figure 47 SAP HANA - R integration	80
Figure 48 Orange data mining algorithms.....	81
Figure 49 KNIME workflow example for credit scoring	82
Figure 50 RapidMiner data mining process design example	83
Figure 51 The CRISP-DM User Guide from NCR Systems Engineering Copenhagen.....	84
Figure 52 Comptel Social Analytics Churn Statistics and Dashboard view	86

List of abbreviations

3GPP 3rd Generation Partnership Project

3G Third Generation

4G Fourth Generation

API Application Programming Interface

BA Business Analytics

BI Business Intelligence

CAPEX CAPital EXpenditure

CDMA Code Division Multiple Access

CEM Customer Experience Management

CN Core Network

CSP Customer Service Provider

DL DownLink

DSL Digital Subscriber Line

ETSI European Telecommunications Standards Institute

FDM Frequency Division Multiplexing

GGSN Gateway GPRS Support Node

GMSC Gateway Mobile Switching Center

GPRS General Packet Radio Service

GPS Global Positioning System

HSDPA High Speed Downlink Packet Access

HSPA High Speed Packet Access

HSUPA High Speed Uplink Packet Access

IEEE Institute of Electrical & Electronics Engineers

IETF Internet Engineering Task Force

IMEI International Mobile Equipment Identity

IMS IP Multimedia Subsystem

IMSI International Mobile Subscriber Identity

IMT International Mobile Telecommunication

IP Internet Protocol

ITU-R International Telecommunication Union – Radio communication

LTE Long Term Evolution

LTE-A Long Term Evolution-Advanced

M2M Machine-to-Machine

MIMO Multiple Input Multiple Output

MNO Mobile Network Operator

MSISDN Mobile Subscriber Integrated Services Digital Network Number

MU-MIMO Multi-User MIMO

MVNO Mobile Virtual Network Operator

OAM Operations and Maintenance

OLTP Online Transaction Processing

OFDM Orthogonal Frequency Division Multiplexing

OFDMA Orthogonal Frequency Division Multiple Access

OPEX OPERational EXpenditure

QoS Quality of Service

RAN Radio Access Network

SDM Service Device Management

SIM Subscriber Identity Module

SMS Short Message Service

UE User Equipment

UL UpLink

UMTS Universal Mobile Telecommunication System

VoIP Voice over IP

WCDMA Wideband Code Division Multiple Access

1. Literature part

1.1.Strategic information versus operational information

Nowadays business is growing in complexity, enterprises or corporations are spread globally and competition becomes tougher and tougher.

Therefore business executives need the relevant information in real-time to base their *strategic decisions* in order to increase the company's bottom line.

This type of information often referred to as ***strategic information*** is different from the one used for running the normal day-to-day operations (e.g. order processing, billing / invoicing, general ledger, human resources / payroll, materials / resource planning, financial, etc.). The strategic information helps the respective management structures from the company's business environment (e.g. *senior management*, *middle management* – scientists and knowledge workers, *operational management* – production and service workers; data workers¹) to define new strategies and objectives, to retrieve their status when needed for critical decisions or to monitor the result of their actions as per their responsibility and accountability.

In time, this recognized crossed-industrial need generated a huge IT evolution, from *batch applications* (ad-hoc or automated *host based query* and *reporting*), *data warehousing* to complete *business intelligence* solutions having as main objective to provide the answer and information to the key business questions as they were requested.

Currently this latest stage of business information systems, ***business intelligence*** is becoming a must have for every company and is evolving more to ***business analytics*** techniques (actually Gartner already renamed Business Intelligence to Analytics, underlying the change from data store and report related activities to more analytical ones that open the path towards new actionable dimensions in every organization).

In the business analytics solution spectrum an important and growing role is played by the ***predictive analytics*** and in particular ***social analytics***, the main subject of the current paper (as stand-alone or as core element in a complex Decision Management System).

Now let's briefly try to explain and position the predictive analytics solutions in the overall business information system environment.

From the beginning, businesses used the scattered application databases to track the basic transactions, but also the ones that helped the company ran in an efficient way through better decisions. In this mess of different applications and databases spread on a multitude of computer systems the necessity of a “data warehouse” emerged (*“A data warehouse or enterprise data warehouse (DW, DWH, or EDW) is a database used for reporting and data analysis. It is a central repository of data which is created by integrating data from one or more disparate sources. Data warehouses store current as well as historical data and are used for creating trending reports for senior management reporting such as annual and quarterly comparisons)”* (Wikipedia - Data Warehouse)).

Bill Inmon considered the “father of data warehouse” since the 70s (defining and discussing the Data Warehouse term) presented the continuous evolving stage of the data warehouse with respect to its form and structure (newest model DW 2.0 emphasize the importance of *unstructured*, textual data type, the need to incorporate in DWH the formal *metadata* infrastructure and the fact that exists a *data life cycle* from access point of view) as well as the complete infrastructure surrounding the data warehouse which leads to a complete architecture called “cif” – corporate information factory as depicted below (Inmon, 2010).

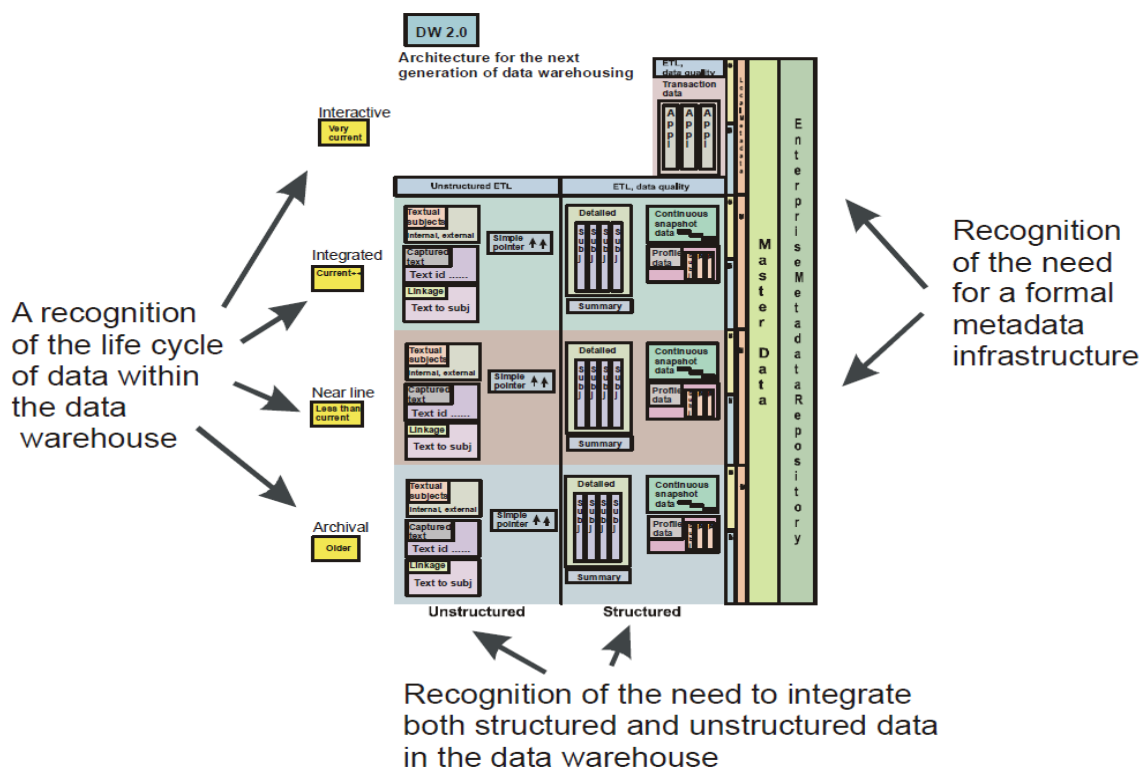


Figure 1 Data Warehouse 2.0 concept

As seen, enterprises can build a centralized data warehouse or create smaller, decentralized data warehouses called *data marts*, focused on a single subject area or line of business (“where different departments had their own version of the base data found in the data warehouse environment”³).

Here is to be mentioned that an alternative to the classical Bill Inmon’s model for DWH (top down approach) is Ralph Kimball’s model (bottom up approach) based on the view that a “data warehouse is nothing more than the union of all the data marts”⁴. Despite its fast development and reduced costs, this model is more suitable for small to medium corporations and can create an inconsistent data warehouse, especially in large organizations.

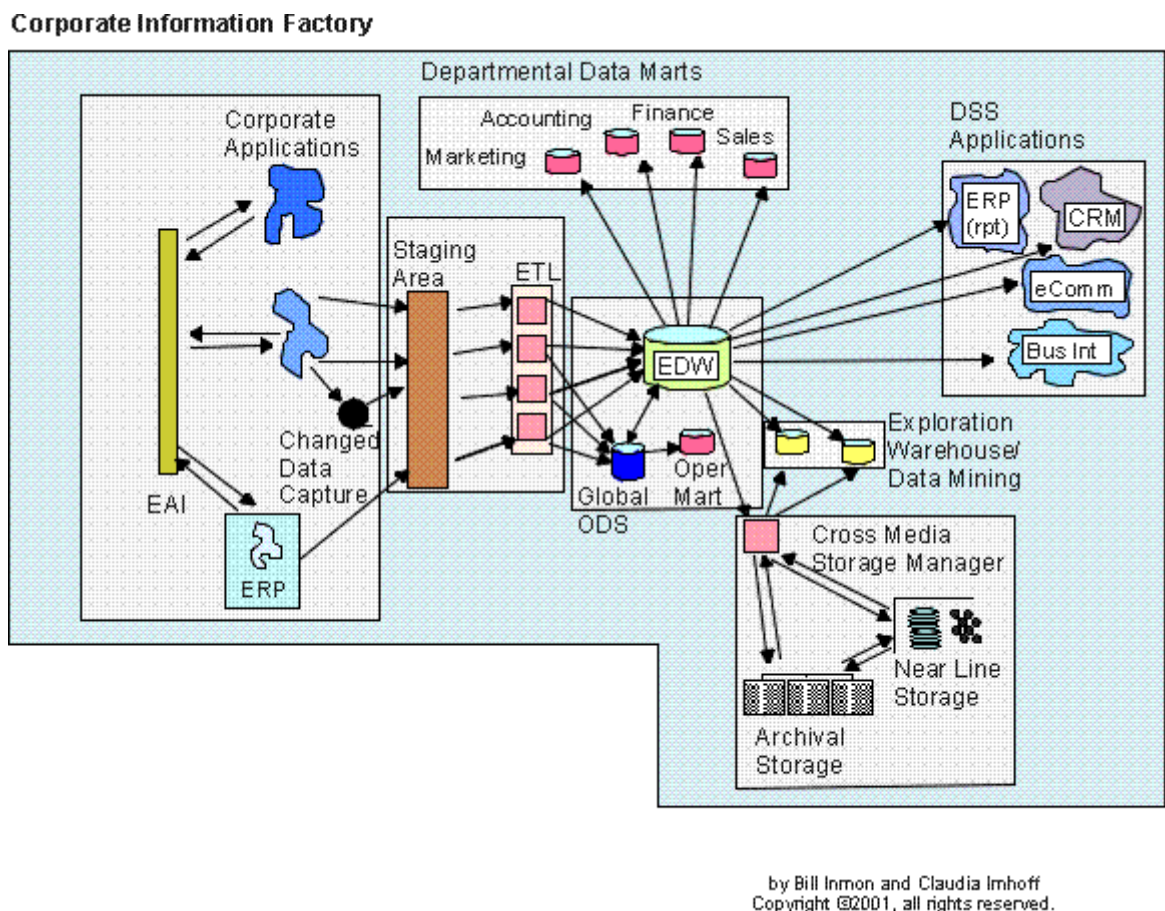


Figure 2 Corporate Information Factory

As soon as the data is stored and organized in data warehouses and data marts, can be subject for further *analysis* and / or interrogations using specific tools for business intelligence or analytics.

There are three types of *data analysis* ((Practical Analytics Wordpress, 2013)):

- *Predictive* (on the top of data mining used for forecasting future outcomes of events)
- *Descriptive* (business intelligence and data mining patterns)
- *Prescriptive* (optimization and simulation, suggesting actions to benefit from predictions together with the implications of every taken decision)

There is a lot of information and also room for interpretation related to *business intelligence* and *business analytics*, some specialists include the analytics in the business intelligence (e.g. from Wikipedia, (Davenport)“*Thomas Davenport argues that business intelligence should be divided into querying, reporting, OLAP, an "alerts" tool, and business analytics. In this definition, business analytics is the subset of BI based on statistics, prediction, and optimization*”

From application and tooling historical reasons I will consider the explanation from the Data Warehouse Institute emphasized also in the following graph:

“Business analytics allows users to examine and manipulate data to drive positive business actions. Armed with advanced analytics insights, business users can make well-informed, fact-based decisions to support their organizations’ tactical and strategic goals.

Business analytics includes advanced techniques such as spatial analytics, customer analytics, and enterprise decision management. Analytic applications bundle tools for data access, dashboard reporting, scorecards, and analytics into packages. Predictive analytics identify relationships and patterns in large volumes of data to create predictive models. Text mining parses unstructured data and merges it with structured data to support user queries, reports, and analyses. “Big data” analytics implement MapReduce, Hadoop, and specialized, non-SQL programming methods to speed insight from huge volumes of data drawn typically from online sources (TDWI Business Analytics).”

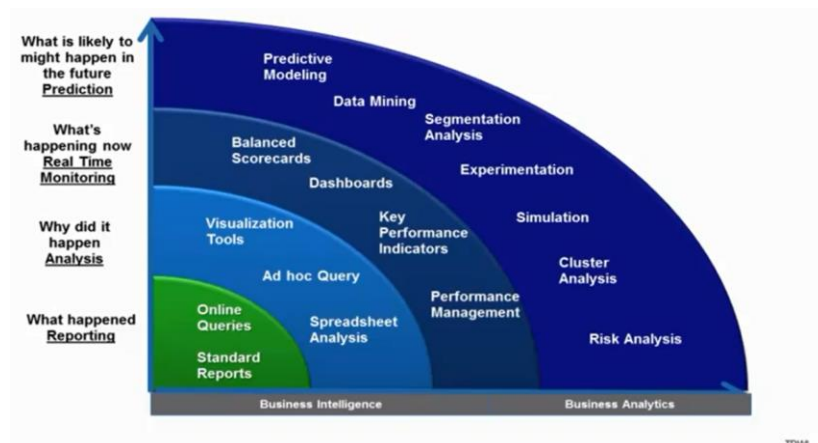


Figure 3 Business Intelligence evolution towards Analytics

“Business intelligence (BI) unites data, technology, analytics, and human knowledge to optimize business decisions and ultimately

drive an enterprise's success. BI programs usually combine an enterprise data warehouse and a BI platform or tool set to transform data into usable, actionable business information (TDWI Business Analytics)."

Business intelligence is a *decision support system* (DSS) where information is gathered for the purpose of predictive analysis and support for business decisions and generally can cover the current (e.g. real time monitoring) and historical enterprise data.

Predictive analytics as part of business analytics is relying on business intelligence data for forecasting and modeling. Usually predictive analytics data is correlated to *future* patterns; is using "data mining techniques, historical data, and assumptions about future conditions to predict outcomes of events"¹.

Prescriptive analytics automatically combines company internal data with external sources (social data like), mathematical sciences, business rules, and machine learning to make predictions and suggests decision options related to a future business opportunity emphasizing the risks in every scenario.

Data mining is the computational process of discovering hidden patterns and relationships in large databases involving methods that lay at the intersection of artificial intelligence, machine learning, statistics, and database systems. The outcome is information like:

- associations ("occurrences linked to a single event")
- sequences ("events linked over time")
- classifications ("recognizes patterns that describe the group to which an item belongs by examining existing items that have been classified and by inferring a set of rules")
- clusters ("similar to classification when no groups have yet been defined")
- forecasts (more statistic forecasts - "uses a series of existing values to forecast what other values will be – e.g. estimate the future value of continuous variables, such as sales figures")¹.

Data mining (DM) is also known as Knowledge Discovery in Databases. Following a formal definition by W. Frawley, G. Piatetsky-Shapiro and C. Matheus (in AI Magazine, Fall 1992, pp. 213–228), DM has been defined as "The nontrivial extraction of implicit, previously unknown, and potentially useful information from data." [D. Ruan, et al., 2005, Intelligent Data Mining]

1.2.WHY the need for an appropriate analytics solution? Opportunities in different industries due to the use of strategic information; cross-industry market landscape

“Drawing predictions from big data is at the heart of nearly everything, whether it's in science, business, finance, sports, or politics”- Stephen Baker, author, The Numerati and Final Jeopardy: Man vs. Machine and the Quest to Know Everything (from posts about Eric Siegel’s book – Predictive Analytics (Decisionstats))

Recognizing the importance of this phenomenon, Eric Siegel the founder of Predictive Analytics World is organizing events, conferences and workshops aiming to share cross-industrial applications experiences based on predictive analytics -“*delivering vendor-neutral sessions across verticals such as banking, financial services, e-commerce, education, government, healthcare, high technology, insurance, non-profits, publishing, social gaming, retail and telecommunications*” (Siegel)

Below are listed some of the best-known companies having Predictive Analytics toolsets in their portfolio (Practical Analytics Wordpress).



Figure 4 Predictive Analytics tool spectrum

The difference between these tools is often in the level of customization and the allowed volume of data, as well as the complete business suite incorporated (e.g. containing embedded business rules automation as crucial for Decision Management Systems).

Besides these commercial tools, some important *open source* predictive analytics applications, statistical and data mining packages and programming languages are as follows:

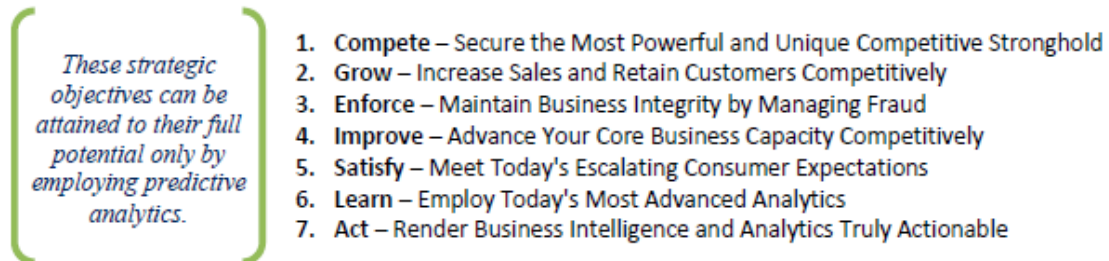
- KNIME (<http://www.knime.org> – applications like Churn Analysis, Credit Scoring ...)
- Orange (<http://orange.biolab.si/>, Python based)
- Python (<http://www.python.org/>)
- R (with statistical package and Rattle plug in, <http://www.r-project.org/>)
- PMML (<http://www.dmg.org/v4-1/GeneralStructure.html>, XML based)
- RapidMiner (<http://rapid-i.com/>, Java based)
- Weka (data mining package - <http://www.cs.waikato.ac.nz/ml/weka/>, Java based)

Overview	Features	Screenshots	Testimonials	Editions and Prices		
	Community Edition	Enterprise Edition		Big Data Edition	OEM Edition	
		Small	Standard			
General						
Number of Analysts		1 Analyst	5 Analysts	5 Analysts	Per OEM Agreement	
Number of Users (Access to Results)		Unlimited	Unlimited	Unlimited	Per OEM Agreement	
Extension Packs for Additional Users / CPU Cores			✓	✓	✓	
License	Open Source under AGPL3	Closed Source	Closed Source	Closed Source	OEM Contract	
Features						
RapidMiner Process Engine	✓	✓	✓	✓	✓	
MarketPlace Integration	✓	✓	✓	✓		
Extension Mechanism	✓	✓	✓	✓	✓	
RapidAnalytics Connector	✓	✓	✓	✓		
Cube Connector			✓	✓		
SAP Connector			✓	✓		
Hadoop Integration				✓		
PMML Support	✓	✓	✓	✓	✓	
In-Database Mining		✓	✓	✓	✓	
Connection to Action VectorWise		✓	✓	✓	✓	
Process Documentation Workbench		✓	✓	✓		
Operator Libraries		✓	✓	✓		
Process Profiler		✓	✓	✓		
Remote Troubleshooting		✓	✓	✓		
Process Preview		✓	✓	✓		
Process Optimization		✓	✓	✓		
Performance Tuning		✓	✓	✓		
Maintenance						
Software Maintenance	Community	Rapid-I	Rapid-I	Rapid-I	Rapid-I	
Patch Releases		✓	✓	✓	✓	
Fixes Included in Future Releases		✓	✓	✓	✓	
Stabilized and Certified Software Releases		✓	✓	✓	✓	
Managed Release Cycles		✓	✓	✓	✓	
Prices						
Prices	Free	Contact us	Contact us	Contact us	Contact us	

To be mentioned that most of open sources product variants are free of charge only for small to medium data traffic and they do not take into account the access part or major security aspects, documentation, maintenance (releases, bug fixes) or consultative support / training; as soon as those topics are becoming a must for the interested companies, the vendors are providing customized retail package versions. Example from RapidMiner:

Figure 5 RapidMiner versions and pricing structure (open-source vs. Enterprise) w/o support and training

Seven Reasons You Need Predictive Analytics — Key Strategic Objectives Attained:



Source: White Paper Eric Siegel, Ph.D sponsored by IBM

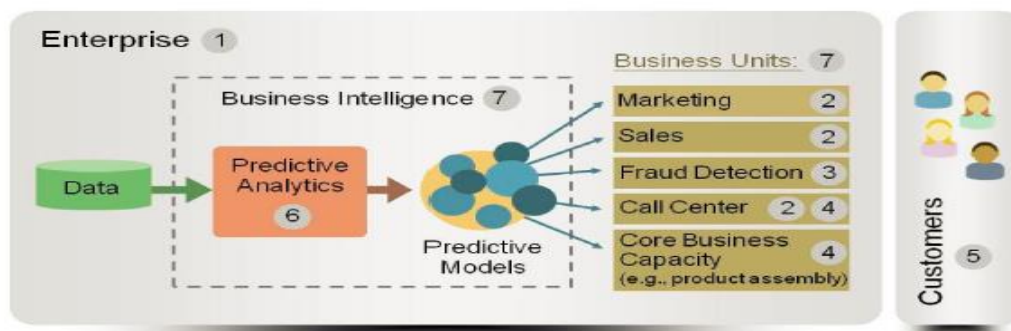


Figure 1. An enterprise deploying predictive analytics across business units. The circled digits 1 through 7 indicate where each strategic objective listed above is attained.

Figure 6 Need for Predictive Analytics solutions to strengthen Enterprises' business - as per IBM's concept

Also TDWI conducted several interviews with companies (167 in 2006) which applied predictive analytics and the most relevant areas are presented in the below graph:

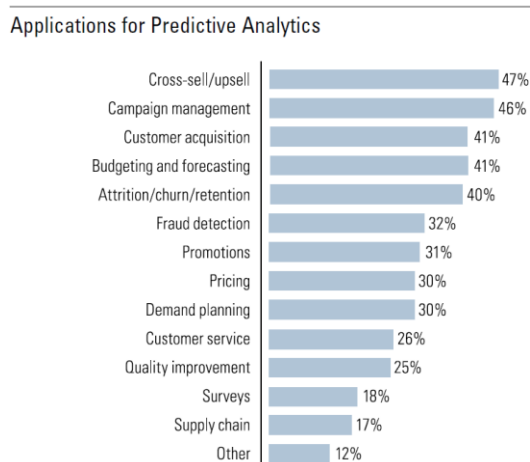


Figure 7 Predictive Analytics TDWI survey

Let's now present some successful implementation stories in order to underline the vast area of applicability.

IBM predictive analytics stories:

1. Business opportunity:

South Africa's largest short-term insurance (Santam) company uses predictive analytics to uncover a major insurance fraud syndicate, save millions on fraudulent claims and resolve legitimate claims 70 times faster than before.

- Like most insurers around the world, Santam was losing millions of dollars paying out fraudulent claims every year
- Expenses were being passed on to the customer in the form of higher premiums and longer waits to settle legitimate claims
- To improve its bottom line and enhance customer satisfaction, the company needed to detect and stop insurance fraud early in the claims process
- It also needed to find a way to isolate risky, fraudulent claims so that claims managers could more quickly process lower-risk claims

Solution:

- Gained the ability to spot fraud early with an advanced analytics solution that
- captures data from incoming claims, assesses each claim against identified risk factors and segments claims to five risk categories, separating higher-risk cases from low-risk claims
- Plans to use propensity modeling to enhance and refine segmentation process

2. IBM has successful stories in the "predictive policing"; together with agencies such as London's Metropolitan Police, the Polish National Police and a number of US and Canadian cities developed this predictive modeling to identify the criminal behavioral pattern, targeting "hot spots" and helping the police forces for an effective intervention, thus reducing the criminality index (CRUSH program – Criminal Reduction Using Statistical History).

3. Another IBM example "*Ultimately, business analytics helps our students succeed by matching them closely to the courses that we predict will go well for them, which is healthy for*

their careers and healthy for our long-term future." - David Wright, Assistant Vice-President for Strategic Planning and Business Intelligence, Wichita State University (Wright)

Business need:

Managing the business affairs of WSU requires attention to both academic standards and financial stability, and the two are closely linked. WSU needed to understand the costs of each course and the supporting faculty, allocate the fees generated, and ensure students complete successful academic careers.

Solution:

WSU implemented a suite of IBM business analytics software to collect data from multiple source systems, enable advanced data management using cross-platform technologies, and deliver consolidated information and predictive analysis to key decision-makers.

Results:

Eliminates the need to hire external analysts to score applicants, a saving of over \$10,000 in the first two years of implementation.

Benefits:

- Predicts the chances of success for potential students, enabling marketing teams to focus on high-quality applicants. This has boosted registration yields by 15 percent.
- Delivers more robust results: Wichita's recruitment model provides greater accuracy in identifying high-yield prospects than the best external providers' model (96 % versus 82 %).

4. IBM Content and Predictive Analytics for Healthcare—the first Ready for IBM Watson Solution—rapidly interprets complex and disparate types of structured and unstructured information, helping to eliminate clinical and operational blind spots by making insights available to care providers.

5. Body Shop International plc. used predictive analytics on the top of database catalogue, info from their Web page and retail store customers to identify the latter who were more likely to make catalogue purchases so that the company was able to build a more targeted marketing (mailing list for their catalogues), resulting a better response rate related to those mailings and catalogue revenues¹.

6. FedEx is using SAS Institute's Enterprise Miner and predictive analytics tools to develop models that predict how customers will respond to price changes and new services, which customers are most at risk of switching to competitors, and how much revenue will be generated by new storefront or drop-box locations¹.

7. Prediction of user interests in e-commerce, mass-media, and entertainment industries. The main source of data for those predictive analytics tasks is browsing and buying behavior of the visitors. Web analytics application is sensitive to context, a collection of external factors influencing visitor behavior (e.g. location, time, access device, weather and holidays).

8. Polling predictions in politics (is known that Obama's administration was pretty much using such kind of predictive modeling) and a great interest is confirmed even on the personal campaign site aiming the candidate's re-election. (*"The campaign seeks interns to join the department at headquarters in downtown Chicago, IL. Each intern will work with an experienced team of predictive modelers and other analytics professionals—with the goal of re-electing President Obama"*⁸)

A lot of implementations are having as destination the **high-tech** domains (e.g. aerospace, automotive, IT, biotechnology, telecommunications, etc. (Wikipedia - High Tech)), the ones with the most advanced technology, known as being very dynamic and therefore imposing frequent decision changes over the entire enterprise's business processes.

9. *"In the automotive industry, both advanced condition monitoring and warranty claims can benefit from predictive analytics. Through dealership and telematics information, customer sentiment expressed through social media, and customer driving patterns, manufacturers can better predict maintenance requirements for specific drivers and forecast the success of aftermarket parts and services. Manufacturers can better predict component failures and anticipate maintenance opportunities by analyzing warranty claims and defect trends and patterns.*

Predictive analytics can help the automotive industry:

- *Develop new maintenance and aftermarket offerings*
- *Accelerate vehicle launches*
- *Reduce warranty cost and recalls*
- *Improve future product quality*
- *Lower fraudulent warrant claims*
- *Identify the source of defects in the supply chain"* (IBM BigData Hub)

10. Comptel predictive analytics solutions for the Telecommunications industry - analysis, assess and act based on the operational data helping sales and marketing personnel discover data-driven patterns, relationships in their data that impact activations, other Key Performance Indicators (KPIs) or decision making. As a result, telecom providers are able to identify new market and customer segment opportunities or to apply targeted campaigns to prevent customers to churn, thus drive revenue growth.

Also predictive analytics can be used with respect to Call Centre activities to help customer service representatives identify customers with the highest level of dissatisfaction and take the corresponding actions to fulfill their expectations.

11. Pharmaceutical companies are seeking to reduce their direct/field sales forces and look for alternative solutions to improve sales and marketing efficacy. Therefore predictive analytics is becoming increasingly critical for business success, especially for sales and marketing managers who gain predictive insights that can improve sales force productivity.

12. Predictive analytics used for helping the IT security professionals in predicting future computer attacks (internal and external, combining information from Identity Management Systems, human resources / HR and other sources (e.g. network protocols related, etc.) with physical access control and video surveillance logs).

13. In *biotechnology* domain great efforts are made with respect to drug development process to prepare safe and efficient products and one of the most important phases (also from time and cost perspective) is the clinical trial which is gradually starting from animals to human subjects upon their step-wise proved progress. Significant results appeared due to entrepreneurial solutions in this area, leveraging the existing information with respect to biological materials and processes (e.g. *predictive analytics models* that are helping to forecast “human toxicities from in vitro datasets” (BioMap Systems)).

14. *Aviation* is well known for its complex broadened operations and a multitude of local governmental and international regulations. Predictive analytics solutions are a great support in avoiding the overbookings or empty seats, as well as for influencing the decisions that may affect the airlines, airports and additional area functionalities due to the various involved 3rd

party companies, institutions or agencies (including security, catering, maintenance services, communications, fuel delivery, luggage transportation, taxi services, etc.) especially in case of natural unwished events, accidents, congestions, delays or illegal traffic activities (including nuclear ones). Simulations are enforced in order to drive to the best overall management decisions to reallocate the resources as needed, mitigate the eventual crisis or to implement measures and controls in order to avoid causes that led in the past to flight incidents; all this in an effort to ensure the business continuity and keep customer's satisfaction as a priority due to its direct impact in revenue generation (Futron).

15. Also organizations as United Nations Office on Drug and Crime (UNODC) are implementing predictive based modeling solutions in order to help the investigators in their specific work or to guide common activities with country border control responsible authorities (automatic processed based on rules and predictive models having as result lists for border and custom control officers containing the top most likely individuals that can be subject to illegal activities depending on the location, date and time). Similar predictive solutions are part to Enterprise Resource Planning implementation projects or to detect the propensity for state (or other entities) *nuclear weaponry tests* and the sounding impact, having as input the collected data from probes in the likely affected areas (mainly radiation detection in geopolitical interest regions) that may indicate pre / post-launch nuclear tests and any related unstructured information from media or social networks.

Only for one kind of application type where the predictive models can be fitted (e.g. Fraud and Security Applications) there are several analogue domain areas⁷:

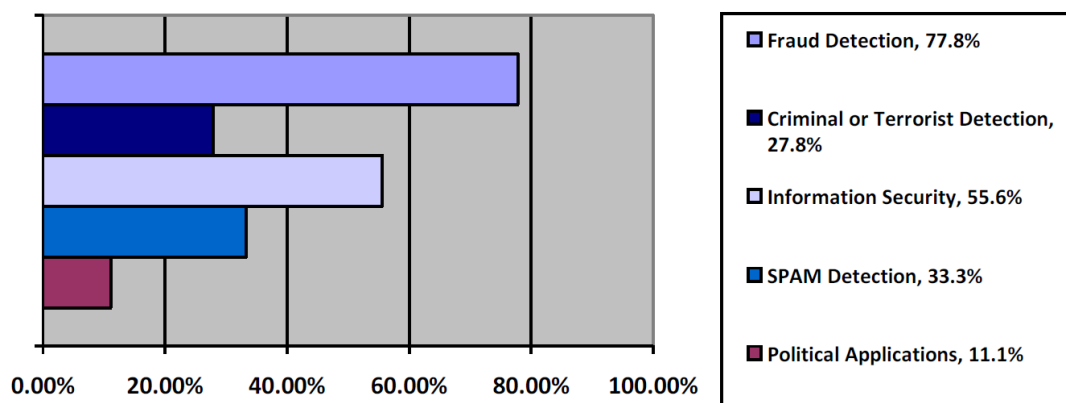


Figure 8 Predictive Analytics World, Survey Results Jan 2009

The above list is non-exhaustive and is just to present the variety of domains and possible implementations using solutions based on predictive analytics models and confirm that “predictive analytics” is not a hype but a tremendous source of growth especially for businesses where a fast and optimal decision creates a huge impact.

Further on I will try to emphasize that predictive analytics (and particularly social analytics) is also capturing the momentum of “big data” expansion and therefore is not just a coincidence that more and more businesses started to understand the necessity of achieving such capabilities in their portfolio.

1.3.Social analytics solutions based on predictive analytics as important cross-industry business growth driver

Fast reaction time, efficient and effective corporate strategy are crucial especially in high-tech industry and therefore the advanced, mathematically based approaches of the predictive analytics models are powerful tools for leveraging data.

Usability and functionality of both predictive analytics and big data technologies have increased in the last few years and there are no more technology incompatibilities. Lot of predictive analytic tools support access to a wide range of data sources, including those typically branded “big data,” such as unstructured text, or semi-structured Web logs and sensor data (e.g. RFID tags).

Forrester (a global research and advisory firm, www.forrester.com) defines “*big data as the frontier of a firm’s ability to store, process, and access (SPA) all the data it needs to operate effectively, make decisions, reduce risks, and serve customers*”. (Gualtieri).

At the level of year 2006 TDWI identified predictive analytics as being in “early-adopter phase” and stating that despite of its high value, the penetration degree in the organizations was quite low⁵ (Microstrategy - TDWI Report, 2006, p. 4)

Status of Predictive Analytics

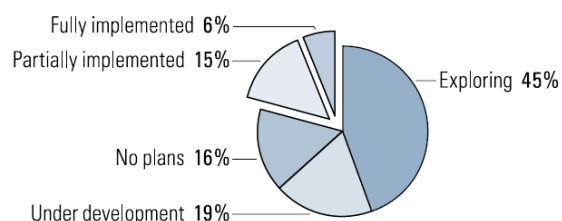


Figure 9 TDWI Survey - Predictive Analytics tool adoption, Aug 2006

Figure 1. Predictive analytics is still in an early-adopter phase. Based on 833 respondents to a TDWI survey conducted August 2006.

“The first annual Predictive Analytics Business Applications survey⁷ was open for four weeks” (Predictive Analytics World, 2009, p. 1) in January 2009 and promoted via the Predictive Analytics World web site and several blogs on predictive analytics and business intelligence showed already a big improvement compared to 2006 and denote that more companies were aware of the real predictive analytics power applied on the big data technologies.

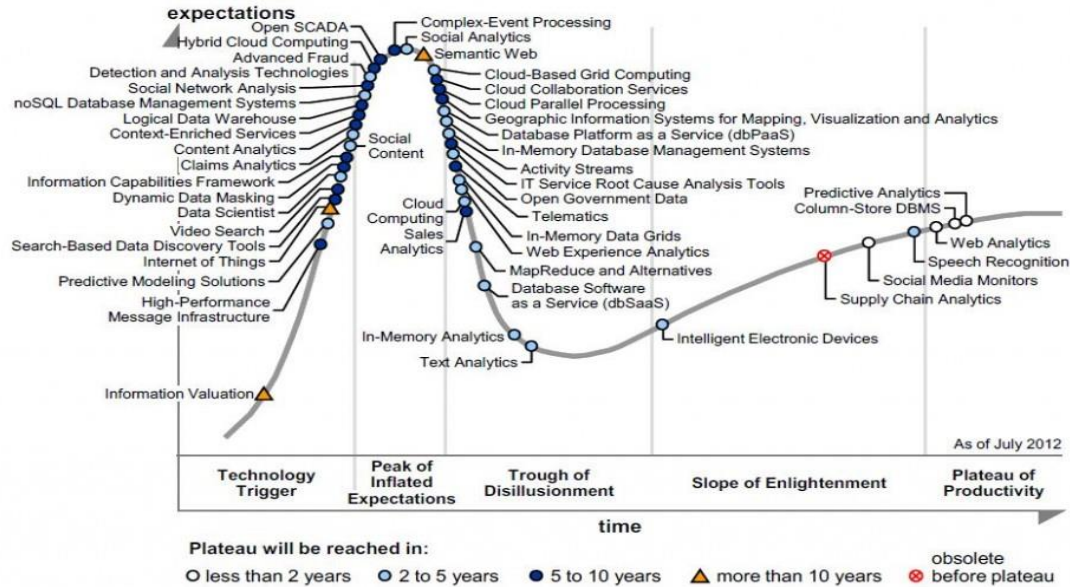
This survey *“was opened to all members of the community, including vendors, consultants and companies, whether actively employing predictive analytics or not”* (Predictive Analytics World, 2009, p. 1). From 94 valid responses to the survey (47 organizations - corporate, non-profit or government are included in the survey report) more than half do not provide predictive analytics software or services – i.e., they are the current and future users/consumers of the technology. Some of them (51.5%) have never deployed predictive analytics, but the vast majority - all but 15.2% - had plans to do so within the next five years, while 51.5% planning to do it in the next six months. *“The top three reasons for doing so are to obtain strategic insights (75%), achieve decision support (57.1%) and enact decision automation (46.4%). 90.1% of respondents who have deployed predictive analytics attained a positive ROI from their most successful initiative. All in all, the survey results promise strong growth for the predictive analytics industry. Predictive analytics vendors and consultants comprised 47 of the survey respondents (coincidentally the same count)* (Predictive Analytics World, 2009, p. 1).”

Now situation started to boost exponentially. From Gartner’s analyst firm report: IT to spend \$232B on Big Data over 5 years” – *“Big Data Drives Rapid Changes in Infrastructure and \$232 Billion in IT Spending Through 2016”* (Columbus, Forbes - Gartner’s 232B Big Data Forecast, 2012)

“Big data has become a major driver of IT spending. The benefits to organizations for adding big data to their information management and analytics infrastructure will force a more rapid cycle of replacing existing solutions (Mark A. Beyer, 2012).”

The Hype Cycle for Big Data according to Gartner is showing that in a period of 2-5 years (from 2012) a maximum will be reached by *social analytics* solutions (based on *predictive analytics*)⁶ (Columbus, Forbes - Roundup of big data forecasts and market estimates, 2012)

Figure 1. Hype Cycle for Big Data, 2012

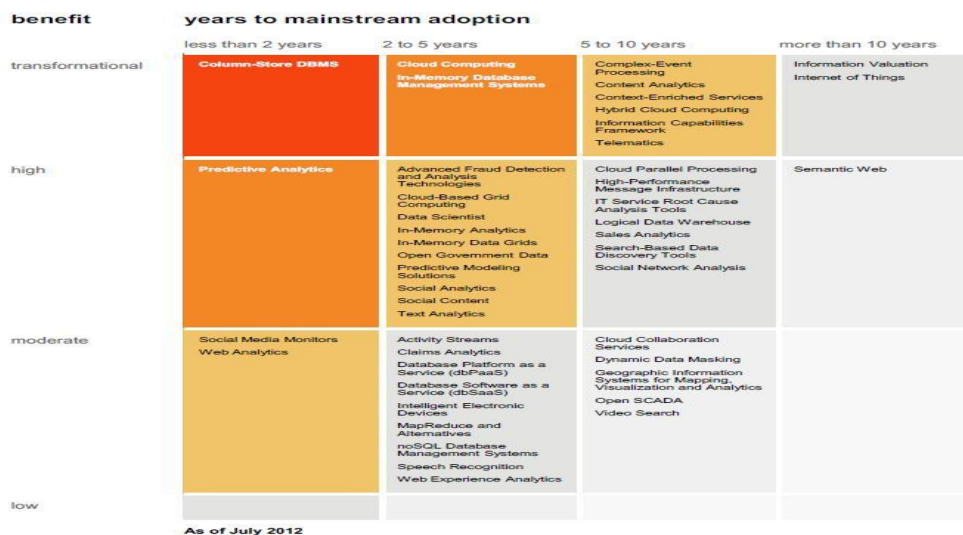


Source: Gartner (July 2012)

Figure 10 Social Analytics and Cloud Parallel Processing as big expectations in a 2-5 year time frame, Gartner July 2012

Predictive modeling algorithms are gaining momentum with enterprises which are using them to support claims analysis, Customer Relationship Manager / CRM, risk management, strategies for pricing optimization, Web-based quoting, etc. According to Gartner, the Priority Matrix for Big Data, 2012 is quoting *predictive analytics* as solutions returning a high benefit:

Figure 2. Priority Matrix for Big Data, 2012

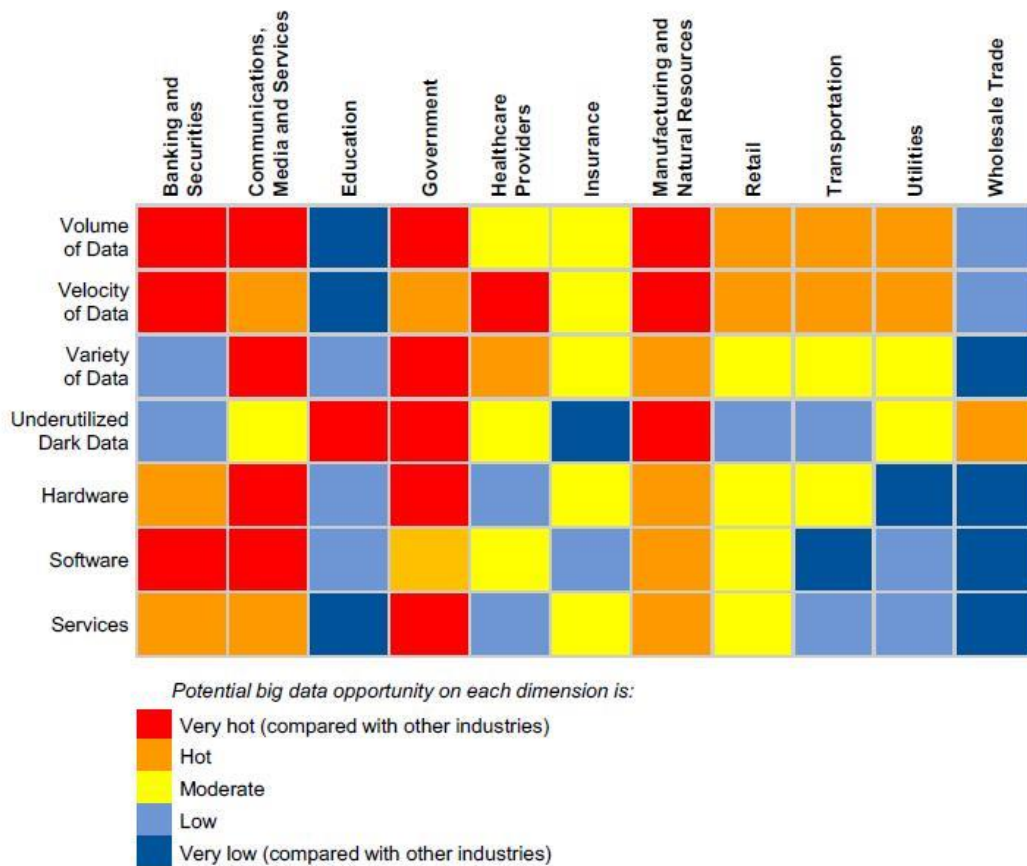


Source: Gartner (July 2012)

Figure 11 Big Data priority matrix - Gartner, July 2012

In Gartner's report "Market Trends: Big Data Opportunities in Vertical Industries" is presented the greatest cross-industry potential opportunity map for Big Data.

Figure 2. Big Data Opportunity Heat Map by Industry



Source: Gartner (July 2012)

Figure 12 Big Data Opportunity Map by Industry, Gartner, July 2012

If the buzz words were "data-driven" business decisions or actions till late 2010 [while the focus on enterprise business intelligence and analytics was centered on *structured*, semi-structured data (emails, logs and call records)], together with "big data" and evolution of technology the focus changed to *unstructured* data (like sensor data from RFID tags , multimedia, online web content), thus emerged the need for better *analytical* capabilities so that the newly emanated buzz words become "advanced analytics", "analytics-driven" or "big insights."

Leveraging on big data technology, the power of predictive analytics is getting a lot of coverage, and software vendors are struggling to include the latest related technologies and algorithms to help firms increase knowledge of their business, competitors, and customers by reducing the risks, making intelligent decisions and create differentiated and personalized customer experiences.

From marketing campaign analysis and social graph analysis to network monitoring, fraud detection and risk modeling, there's unquestionably a Big Data use in combination with advanced predictive analytics, predictive models meant to improve business outcomes.

Gartner Analyst Rita Sallam estimates that today about 30 percent of users deal with analytics, but said that the number will rise up to 50 percent by 2014 and to 75 percent by 2020; Gartner CIO surveys consistently rank business intelligence and analytics ahead of such areas as mobile and cloud in terms of overall priorities (Sallam, 2013)

“Accenture (ACN - news - people) research shows that high-performance businesses have a much more developed analytical orientation than other organizations. They are five times more likely than their low-performing competitors to view analytical capabilities as core to the business”⁹ (Harris, 2010).

By using predictive analytics all the key elements of a business: business strategy, business model, infrastructure and technology are in sync and through process improvement the revenue generation will be improved, cost structures better defined, thus leading to enhanced decision making.

2 Predictive analytics related (types, tools, data mining algorithms, PMML standardization for import / export predictive models)

2.1 Types of analytics (demystifying *descriptive* and *predictive* analytics)

Analytics in broader final purpose handles how an entity (e.g. business) arrives to an optimal or realistic decision based on existing data.

According to INFORMS (<https://www.informs.org/>, The Institute for Operations Research and the Management Sciences, which is the largest professional society in the world for professionals in the field of operations research (O.R.), management science, and business analytics) analytics are classified as we mentioned before in:

- a) *Descriptive analytics (use of data to determine what happened in the past or “now”)***
 - *“Prepares and analyzes **historical** data*
 - *Identifies patterns from samples for reporting of trend (Informs, Community - Analytics)”*

Here are included: *business intelligence (traditional - aiming post mortem reports for organization's business departments in order to seek for the reasons behind past success or failure), data mining*

Data Mining is an analytic process (analysis step in Knowledge Discovery in Databases process, KDD) designed to explore data (usually large amounts of data, typically business/internal or market/external related) in search of consistent trends, patterns and/or systematic relationships between variables. At a high level, data mining can be seen as *gathering* knowledge about relationships, while the resulting predictive analytics model is *applying* that knowledge.

Descriptive models determine relationships in data and are often used to classify the customers or prospects into groups. While predictive models are generally focusing on predicting a single customer behavior, descriptive models spot many different relationships between customers or products. Predictive models rank-order customers by their likelihood to behave or act in a specific way unlike the descriptive models.

Data Mining is the first step in the predictive analytics flow; the data identified as relevant by the mining process can be used to develop the predictive model. It catalogues all relationships or correlations that may be found in the data chunk, disrespecting the cause of those relationships. Sometimes there is a great degree of confusion between “machine learning” and “data mining”; data mining is using many machine learning methods (*machine learning* performance is due to the ability to reproduce the “**known**” knowledge, while *data mining* core task is to discover the previous “**unknown**” knowledge).

INFORMS presents the overall business intelligence phenomena (“*as a melding of technologies, models, techniques and practices*”) in an attempt to classify and explain their fuzzy terminology and overlapping, as per the below depicted framework: (“*The three circles of the Venn diagram each represent areas of study and application that had previously been considered quite distinct: 1. information systems and technology, 2. statistics, and 3. OR/MS. It serves to encapsulate the broadening definition of BI. With this new vision, we may now characterize BI from each of three viewpoints as: business information intelligence (BII), business statistical intelligence (BSI) and business modeling intelligence (BMI). Each of the viewpoints has particular business aspects, and academically speaking, courses that are independent of the other viewpoints. Conversely, each viewpoint can work together or utilize techniques/skills from one or possibly two of the other disciplines*” (Miori, 2010).

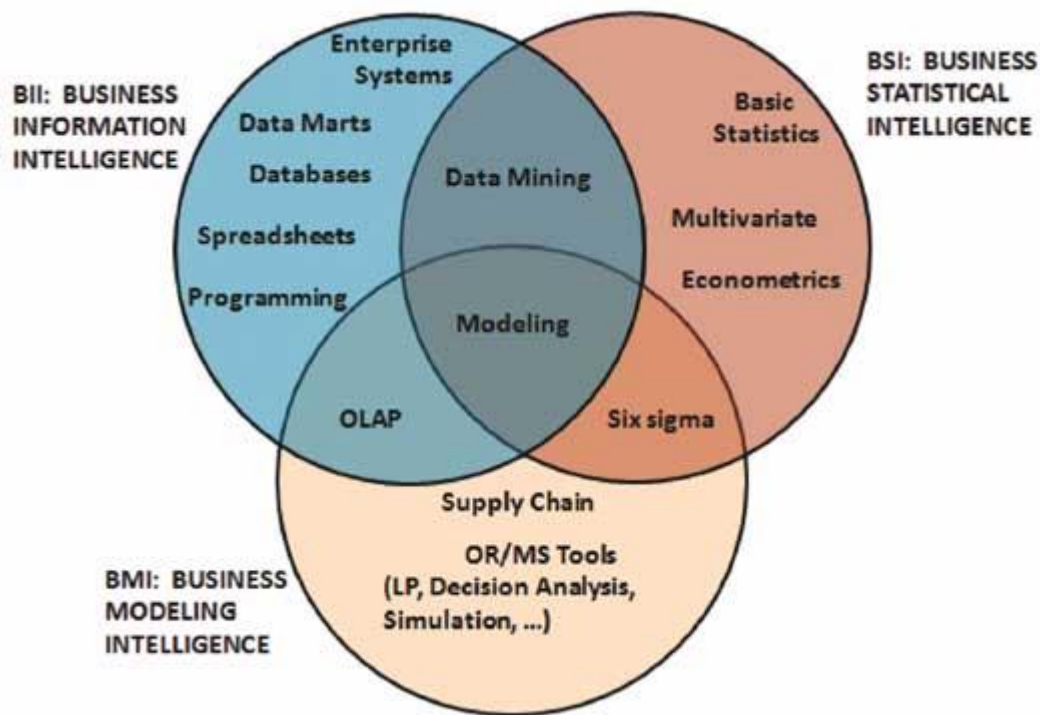


Figure 13 Informs - Business Intelligence / Business Analytics Venn diagram

b) Predictive analytics (what could happen in the future)

- “Predicts **future** probabilities and trends (identifying the risks and opportunities)
- Finds relationships in data that may not be readily apparent with descriptive analysis (Informs, Community - Analytics)”

Predictive modeling acts as an application of the learned data pattern knowledge.

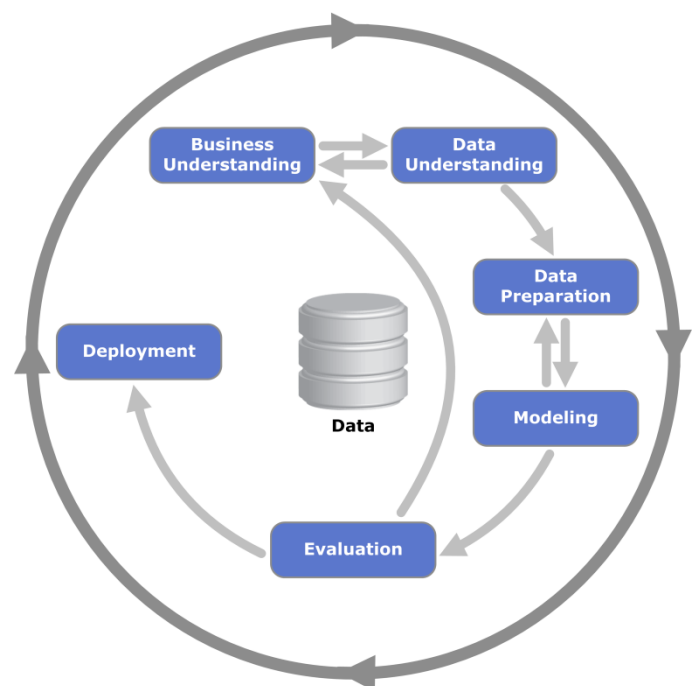


Figure 14 CRISP-DM process diagram

Used by “machine learning” to generate a predictive model

Used to evaluate the predictive model

Training Data

Testing Data

Predictive modeling is mentioned as a “mechanism that predicts a behavior of an individual”; as input has the characteristics (variables) of the individual and as output delivers the “*predictive score*” (the higher the score, the greater the likelihood that the individual will manifest the predicted “behavior”).

Predictive modeling is based on “machine learning” (a branch of artificial intelligence aiming the construction and study of systems that can learn from data). The core of machine learning deals with representation and generalization. Representation of data instances and functions evaluated on these instances are part of all machine learning systems. Generalization is the property that the system will perform well on unseen data instances (Wikipedia - Machine Learning)

In the past (and still existing in some open source tools) it requested extensive human interaction and skills (even the *machine learning* terminology supposed the full automation of learning, system designer had to specify the data representation and select which mechanism will be used to search through the data). Nowadays, the leaders in analytical tools are embedding in their software the so called “analytical workbenches”, which are automatically applying multiple models and algorithms allowing the non-specialists to be involved in the creation of analytical models that best fit the data sample and the desired outcome. Modeling methods implied in machine learning are starting with decision trees, artificial networks, logit / probit / linear regressions, support vector machines, TreeNet, etc. and the way towards a good *predictive model* is their incremental modification until the resulted prediction from the training cases is improved.

The challenge here is not to kill the learning (by *overlearning* / over-fitting, pitfall of mistaking the noise for information, assuming too much about what was shown in the analyzed data (algorithms finding patterns that are just a fit with essential random data) – “if you torture the data long enough, it will confess”, as mentioned by Eric Siegel in his “Predictive Analytics” book. As a prove to overlearning pitfall he mentions a famous funny and in the same time nonsense story reported in the Wall Street Journal of an attempt to use data mining in financial markets. During this attempt it was found that US stock returns could be predicted with 99% accuracy by using as inputs the US cheese production and the total population of sheep in the US and Bangladesh. ☺

Therefore the used trick against overlearning is to randomly select a “*testing data set*” (*out-of-sample set*, usually 20-30% of the initial training data) kept aside for validation / evaluation of the resulted predictive model derived from the remaining data (“*training data*”).

Usually for efficient and effective *predicting modeling* output, as prerequisite is a *data mining project* that rely on different frameworks (such as CRISP-DM, Six Sigma / DMAIC, SAP / SEMMA); most of times software tools for data mining are specifically designed and documented to fit into one of these specific frameworks (e.g. *IBM SPSS Modeler* is having CRISP-DM embedded).

CRISP-DM stands for *Cross Industry Standard Process for Data Mining* and is the leading data mining methodology / process model that describes commonly used approaches that expert data miners use in developing data mining and knowledge discovery projects.

DMAIC (*Six Sigma*) model which is pretty popular in American industries is a data-driven methodology aiming to eliminate the defects, waste or quality control problems for all kind of business activities from manufacturing, service delivery, management, etc.

Define → Measure → Analyze → Improve → Control

SEMMA (*SAP* Institute framework – focused on technical activities involved in a data mining project) which stands for:

Sample → Explore → Modify → Model → Assess

c) “Prescriptive analytics

- *Evaluates and determines **new** ways to operate*
- *Targets business objectives*
- *Balances all constraints (Informs, Community - Analytics)”*

Grouped around: *optimization, simulation*

“Prescriptive analytics automatically synthesizes big data, mathematical sciences, business rules, and machine learning to make predictions and then suggests decision options to take advantage of the predictions.

Prescriptive analytics goes beyond predicting future outcomes by also suggesting actions to benefit from the predictions and showing the decision maker the implications of each decision option.

*Prescriptive analytics not only anticipates what will happen and when it will happen, but also **why it will happen**.*

Further, prescriptive analytics can suggest decision options on how to take advantage of a future opportunity or mitigate a future risk and illustrate the implication of each decision option. In practice, prescriptive analytics can continually and automatically process new data to improve prediction accuracy and provide better decision options.

Prescriptive analytics synergistically combines data, business rules, and mathematical models. The data inputs to prescriptive analytics may come from multiple sources, internal (inside the organization) and external (social media, et al.). The data may also be structured, which includes numerical and categorical data, as well as unstructured data, such as text, images, audio, and video data, including big data. Business rules define the business process and include constraints, preferences, policies, best practices, and boundaries. Mathematical models are techniques derived from mathematical sciences and related disciplines including applied statistics, machine learning, operations research, and natural language processing (Analytic Bridge)”

While predictive analytics helps you model and forecast what might happen in the future, *prescriptive analytics* is beyond predictive analytics, a true help in Decision Management Systems, giving the indications meant to decide the best course of action with respect to the given objectives, requirements and constraints. From various choices, alternatives that directly influence the outcome, it seeks to find the optimal solution. Some prescriptive analytics are using *stochastic optimization* to also take into consideration the *uncertainty* that might exist in the data used in the analysis.

This processing task analyze the potential decisions, the interactions between them, the factors and constraints on each decision together with the business outcome of each scenario in part to derive the optimal solution.

This combination of *predictive* (simulation) and *prescriptive* (optimization) analytics in a complex Decision Management System can help any industrial domain (from strategic planning to operational issues) in boosting its efficiency and effectiveness.

The information given by the predictive analytics can help for example a retailer to understand the drivers behind its customer buying patterns, so that the top products customers want could be anticipated.

Prescriptive analytics can provide a great help in the whole supply chain design (including scheduling, production, inventory) to deliver the products customers want in the most optimized way. Nowadays, the increased computational power together with the progress of mathematical optimization algorithms makes possible the existence of real-time problem solving capabilities in order to support the operational, tactical or strategic decision.

2.2 Predictive Analytics tool and solution spectrum (in-house proprietary solutions developed using C++ / Java / .NET technologies versus open-source solutions; predictive model creation and consumer separation using PMML (Predictive Model Markup Language), an xml based markup language technology)

The previous chapters briefly presented the evolution of the analytical informational systems from *support* to *strategic* role and the importance of *predictive* analytics solutions in the current vast business spectrum as a key element in *decision* making.

More software companies and open-source communities foreseen the dramatic increase of this *need* on various market segments from small, medium size till large corporations, they all brought their contribution in this area. Generally the open source tools are for tech savvy people and they lack in graphical user interfaces or their usability is not recommended for non-professionals as they request intensive programming language and statistical know-how. The bigger software players and IT integrators are presenting a large solution portfolio in a fierce competition to grab the newer opportunity and spread their product popularity. Some of them (like IBM SPSS and SAS) are providing in their latest solution offerings, the support (free plug-ins) for the open source statistical programming language as “**R**” (which is mostly used by academics and university students, built on *C, Fortran and R*) by this recognizing the various customer needs especially statistics related and taking the advantage of the huge community of users who are constantly building for R new statistical methods or extensions.

“Forrester Research”, a global research and advisory firm evaluated 10 vendors of predictive analytics solutions (Angoss Software, IBM, KXEN, Oracle, Revolution Analytics, Salford Systems, SAP, SAS, StatSoft, and Tibco Software) and presented their ranking, grouped on three categories (current offerings, strategy and market presence), scoring done on a scale from 0 (weak) to 5 (strong) (Mike Gualtieri, 2013) – [Appendix A](#)

Forrester ranks SAS and IBM as strong leaders while SAP, a newcomer to Predictive Analytics, is holding the third place. Started mainly as open-source software predictive tools providers, KXEN and Rapid Analytics are presented as strong performers and respectively contenders taking into account their market presence and the adopted strategy by comparison with the mentioned commercial enterprise predictive analytics solutions - [Appendix B](#).

Now let's present the main commercial predictive tools and their open-source competitors, underlying the effort made to adapt their offerings to comply with a seamless integration of "R" open source programming language.

Enterprise data mining and predictive tools:

- **SAS Enterprise Miner** tool is easy to learn and able to run analysis *in-database* or on *distributed clusters* to handle big data ((*In-database* analytics technology allows data processing to be conducted within the database by building analytic logic into the database itself. Therefore the time and effort required to transform data and move it both ways between a database and a separate analytics application is completely eliminated). When combined with **SAS Rapid Predictive Model** will detect the best mining fitting algorithm and present the classification matrix which emphasizes the prediction accuracy.

SAS Rapid Predictive Modeler for **SAS Enterprise Miner** is considered the easiest of all tools that require the user to only specify the target variable. Moreover Rapid Predictive Modeler for SAS Enterprise Miner automatically treats the data to deal with outliers, missing values, skewed data, variable or model selection.

As there is a huge amount of free algorithms as add-on packages for "R", in SAS versions since 9.22 was implemented an interface to "R" created as a SAS procedure called PROC IML (Interactive Matrix Language). SAS IML programming allows the direct access of data sets (to edit or create new ones) or to develop new SAS / IML modules and integrate them in applications.

SAS is also providing vast vertical solutions that focus on specific domain outcome, particularly in Telco. Here I would like to mention **Customer Analytics for Communications** which is using customer-centric foundation data model (FDM) aligned with *Frameworkx* TM Forum Information Framework (SID). SID is a reference model and common vocabulary for all the information necessary to implement Business Process Framework (eTOM – enhanced Telecommunication Operation Map) processes. Offering off-the-shelf information model that can easily align all involved parties will result a decrease of the design, development, system integration and service complexity.

- **IBM's Smarter Planet** campaign (The vision of transportation, health care, cities, retailing, finance and other fields made more intelligent with digital technology, all relying on immense data collection and analysis) and acquisitions of SPSS, Netezza, and Vivisimo emphasize its strong interest to big data predictive analytics.

Predictive analytics products related I would like to mention is **IBM SPSS Modeler Professional** edition which is a powerful data mining workbench that helps in building predictive models of future events and interactions without having as prerequisites any programming skills.

Many features of statistical analysis programs were initially programmable via proprietary languages (e.g. SPSS Statistics – with 4GL syntax language / fourth-generation programming language), while their later version developments includes Python programmability extension (e.g. SPSS starting with version 14). Due to the Python integration plug-ins (Python being itself written in “C” programming language) and the fact that most of the standard modules are “C” code, was common to adapt existing libraries or code for use as Python extension modules (C, C++, VB.NET, Fortran, etc.) allowing SPSS to run any of the existing “R” open source statistic packages. As for the graphical user interface, front-end is Java written.

- **SAP** as a newcomer to big data predictive analytics is considered a “Leader” due to a strong architecture and strategy and its **SAP HANA in-memory** computing appliance (Accessing data *in-memory* eliminates seek time when querying the data, which provides faster and more predictable performance than disk access). To move the application logic into the database SAP HANA makes use of application functions (similar to database procedures written in C++ and called from outside in order to perform complex data operations). Functions related to a specific topic are grouped in an application function library (AFL), such as *Predictive Analysis Library* (PAL functions can be called within SQLScript procedures, an extension of SQL and can perform analytic algorithms) and Business Function Library (BFL). AFL comes as an archive that must be installed separately (not part of HANA appliance - see **SAP HANA Appliance Software SPS05** / PAL Reference) [Appendix C](#)

Coming to broaden SAP Business Objects (BO) suite, **SAP Predictive Analytics** supporting both open source R and SAP-written data mining algorithms is to be smoothly integrated with the other existing BO tools that helps in data preparation or model building.

SAP Predictive Analytics inherits data acquisition, manipulation and visualization from SAP Visual Intelligence and can work with locally available data (CSV file, Excel, ODBC connection to a database) or on SAP HANA empowered by HANA-R and HANA Predictive Analysis Library (PAL) algorithms (SAP Predictive Analytics - User Guide).

As per SAP information, their Predictive Analytics tool currently support 16 algorithms (including R open source library ones) and are already in the process of releasing a wrapper library (thin layer of code which translates a library's existing interface into a client compatible interface) to permit C++ code function library copy / pasting actions from the open source R community. It is worth

mentioning that in combination with HANA (Predictive and HANA) they support only 5 algorithms, but they have plans to extend it to 15. When data is transferred to HANA, everything is settled in HANA (logic, processing, predictive).

▪ **KXEN** (KXEN - InfiniteInsight) as a *strong performer* presents its **InfiniteInsight** suite solutions for:

- *Communication* (acquisition through cross-sell, up-sell, retention, churn prevention campaigns together with next best activity for every interaction through multiple customer channels)
- *Financial Services* (in addition credit scoring, reducing risk and fraud);
- *e-Business* (gathering the insight of e-Commerce, online retailers and auctions, media and subscription based content, social gaming, etc. to maximize the opportunity by deciding on the proper -personalized content, product and social recommendations, targeted ads, e-Marketing, e-Fraud)
- *Retail* (in addition customer segmentation, product assortment and forecasting, market basket analysis, social network analysis)

InfiniteInsight predictive modeling suite has seven elements of the product range:

- *Explorer* (preparing and transforming the data in order to be used by the analytical engines; uses the introduced semantic layer by which the power users are able to define a set of business components called analytical records, that can be reused within automatic creation of the analytical data sets involved in modeling);
- *Modeler* (automated building of enhanced predictive models using all data mining functions from classification, regression, attribute importance, segmentation /clustering, forecasting, and association rules);
- *Scorer* (automated scoring and optimized code generation – structure and sequence of calculation - as well as preparation for integration in the database production environment by supporting variety of formats: SAS, Java, C, PMML, etc.);
- *Factory* (provides an industrialized production-line modeling approach to support building, deploying and improving any predictive models in a browser based environment that permits customizations for different user profiles such as analysts, supervisors, administrators);

- *Social Network Analysis* (detects the hidden links in the data, identifies the customers with stronger network influence / “influencers” and integrates the detected social attributes in powerful predictive models for different industries and business scenarios spanned between customer lifecycle to risk & fraud);
- *Recommendation* (provides personalized recommendations to each unique visitor by analyzing the transactional data sources and integrate predictive modeling methods together with the existing business rules to weight between the automated recommendations and corporate know-how to prepare the recommendations to be deployed in the production apps (website, mobile))
- *Genius* (automated predictive analytics process – from preparation of the analytical data set, building predictive models, deploying the scores in production environment, executing periodic refreshes as per the latest provided data; deliver out-of-the-box pre-defined templates for marketing campaign optimization)

Lately KXEN offers Cloud Prediction solutions available also for smaller businesses which like to seize the various opportunities by having a thorough insight of their applications already deployed in the cloud environment (they seem to firstly target the users of Salesforce Apps as they provide a seamless integration).

KXEN (Knowledge Extraction Engine) solutions are the commercial version (KJDM) of the open-source Java Data Mining (JDM) standard, a Java “wrapper” built on the top of the KXEN Analytic Framework in an attempt to ease the integration of data mining within enterprise business environments (based on JDM Web services built with Apache Axis). Data access (engine using ODBC / open database connectivity API / Application Programming Interface) and computations are done in C++ (due to efficiency and memory allocation purposes), while Java wrapper is releasing the power of enterprise application development. As many of the predictive analytics suites, KXEN is able to import or export developed models using the PMML / Predictive Model Markup Language (XML based format) that ease the exchange of the models between the applications.

From the “Contenders” predictive analytics solutions spectrum worth to be mentioned are Revolution Analytics, Comptel, Microsoft, 11Ants Analytics:

- *Revolution Analytics* with its **R-Enterprise** production analytics software offers commercial versions of R with increased performances related to *big data* (remove the issue of having to fit everything into memory leading to memory-management problems), that speed computation times (up

to 50 times faster having the ability to leverage multithreading and processor capabilities on all x86 platforms to increase performance), provides an enterprise ready version with deployment options for business users (through web services interfaces to common Business Intelligence tools). They also provide a range of high speed database connectors and in-database engines.

Revolution R-Enterprise make use of its included powerful **RevoScaleR** predictive analytics library package which is designed to be fast and scalable, touching all components involved in the data analytics process: (Revolution Analytics - RevoScaleR)

- *data storage* with its own efficient file format storage (XDF optimized for block / *chunks* reads of data) which allows working with chunks of data instead all dataset to be resident in memory at a specific point in time;
- *computing infrastructure resources* (memory / RAM, processor's core / CPU's core and clustered computers) and the involved
- *streaming* (data from disk to memory and combining the partial results), *optimized* (one processor core dedicated to read/Input – write / Output data operations (I/O), remaining cores processing the buffered data) and *parallelized algorithms* (exploiting the multiple core, the *fastest algorithms* per core which are also the fastest when parallelized and coded using C++ templates plus *categorical data* / variables derived from either or both qualitative or quantitative data observations)

Revolution Analytics is also providing an open source version *R-Community* built with the Intel® Math Kernel Library (to run most computation-intensive programs significantly faster than the original *Base R*) on the final stable release of each R version (containing the critical updates and bug fixes) and relying on multi-core processor library “ParallelR Lite”.

- **Comptel** (Comptel - Social Links) a software company having as core products in the Operation Support Systems (OSS) / Business Support Systems (BSS) Telecom niche.

Lately, they are also providing analytical solutions (e.g. Social Links coming from the former Xtract company) that help transforming Customer Service Providers (CSPs) network data (event – gathering the usage data) into intelligent, automated *decisions* (predictive analytic driven decisions) and *actions* (campaigns towards the subscribers or ensuring better network QoS / quality of service by fine tuning of the relevant parameters). This solution was verified by IBM and acknowledged to be ready for IBM SmartCloud Services – PaaS / Platform as a Service, IaaS / Infrastructure as a Service, backup services. Analytics and modeling is mainly done using R, Java and SQL. This predictive

analytics tool can be integrated in a Customer Experience Management solution and is mainly addressed to business users due to its friendly graphical user interface and full automation features.

- **Microsoft** is also entering in the predictive analytics world with its SQL Server 2012 combining the optimized near-term predictions (ARTXP) and stable long-term predictions (ARIMA) with Better Time Series Support (recently they announced SQL Server 2014 which adds new in-memory capabilities for OLTP / online transaction processing and data warehouses plus a strong platform for hybrid cloud (private / on-premises database applications and public – extending the scalability and backup features with Azure Infrastructure Services). This solution is coming to help the developers in building performance applications in .NET, C/C++, Java and PHP (scripting language initially designed to maintain web /personal home pages) and it can be deployed on customer premises and in the hybrid cloud type (mixture of public and private).

- **11Ants Analytics** like many other software providers are offering vertical data mining and predictive tools based on Microsoft Excel (11Ants Model Builder is a friendly Excel add-on using data mining methods; its predictive models are prepared with the help of 11Ants Predictor and ready to be deployed onto enterprise databases (special support for Oracle, Microsoft SQL Server, Teradata databases).

Free data mining and predictive tools:

- Machine learning and statistical packages in **R** ([CRAN Task View](#))

R started as a statistical programming language and became a powerful tool opening the way to multiple opportunities deriving from its usage in data processing, manipulation, visualization and statistical analysis. Is it supported by a huge community of developers, users, academics and practitioners who are continuously releasing freely distributed add-on packages or contributing to bug fixes or improvements thus keeping it stable and reliable.

It was developed by Ross Ihaka and Robert Gentleman as a free software environment used for academic purposes when teaching at the University of Auckland, New Zealand (they started with a similar syntax as for the commercial “S” programming language for statistics). If initially R was based more on Fortran and C, was easily to call within R the code from these languages; lately the community helped connecting R with other programming languages as C++, Java, Python, .NET, etc. In the previous chapters we have seen that “R” attracted the bigger analytical software players as SAS, IBM, Revolution Analytics attention (based on enhanced R plus a graphical front-end), etc., all

offering add-ons to connect with it and exploring more than its powerful data mining environment. The base R installation comes with a great package library for linking R to file systems, databases and other applications (e.g. “foreign” – reading data from statistical packages SPSS, SAS, etc., RODB – read from databases using Open Database Connectivity protocol (ODBC)). R also includes a standard graphical user interface (Windows R editor, *RGui*) which has some tools such as the console window for writing the scripts, instructions, etc. You have the freedom to select your own editing tools to connect to R, such as the open source enhanced code editing and development environment *RStudio*. There are two additional items that are worth to be mentioned: R is a *vector-based* language (vector could be a row, column of numbers or text) which permits performing many operations through one step (without programming a loop as for other programming languages not vector based) and that R is an *interpreted language* (no need of a “compiler” as in C or Java that creates a program from the given code in order to be used; this easy development cycle has the usual downside as the code runs slower than the compiled codes)

- **Orange**, open source data analytics and mining through *visual programming* (by drag and drop of widgets (more than 100 wrappers / visual building blocks containing the data analysis code and interactions on this data that provide the graphical user interface) or Python *scripting* (Orange - data analytics))

Orange has embedded components for visualization (scatterplots, bar charts, trees, to dendrograms, networks and heatmaps), rule learning, clustering, model evaluation plus add-ons for bioinformatics (*bioorange*) and text mining. Also there is available a great collection of documents starting from Python tutorials to Orange widget development and extensions in C++ (using Python scripting or extension modules such as *orangeom*). It runs on Windows, Linux and Mac OS X operating systems [Appendix D](#).

- **Weka** - originally created in 1993, the Weka project was established in the same academic compound as 11AntsAnalytics data mining tool, University of Waikato -New Zealand having as purpose the research and testing of advanced machine learning algorithms (Weka Data mining)

It is Java written, runs on almost any platform and contains data mining tasks like data pre-processing, classification, regression, clustering, association rules, and visualization. Apart of the stand-alone version contained in the Weka project (which was acquired in 2006 by Pentaho, the creator of an open source Business Intelligence (BI) suite), its set of data mining tools is contained in other products (such as KNIME or RapidMiner). Supported data formats:

- *ARFF* (Attribute Relation File Format – has two sections: *Header* – define attribute name, type and relations; *Data* – lists the data records);
- *CSV*: Comma Separated Values (text file);
- Data read from a *Database* through Java Database Connector (JDBC) / Open Database Connector (ODBC)

▪ **KNIME** offers extensible open source data mining platform implementing the data pipelining paradigm (based on [eclipse](#) IDE / integrated development environment) (Eclipse). Eclipse not-for-profit Foundation (originally created by IBM in 2001 and supported by consortium of software vendors) declared its focus in enabling the software vendors to use the Eclipse technology to develop their commercial software products and services (KNIME data mining platform)

Its open-source version, *KNIME Desktop* the open source from KNIME is a powerful data mining, visualization and reporting workbench and supports R scripts to be run by installing an additional plug-in. *Weka* Input /Output (I/O), data mining algorithms, association rules and predictor are already included with *Weka* add-on installation package as shown in the screenshot ([Appendix E](#)). Related to I/O (Read / Write) operations, *KNIME Desktop* supports as well the *Predictive Model Markup Language* (PMML Read, respective PMML write.

▪ **RapidMiner**, a leading open-source system for knowledge discovery, data mining and predictive analysis contains also *Weka* data mining library. Supports a large variety of data formats including Excel, Access, Dbase, C4.5, ARFF, SPSS, SAS, Stata, Sparse, text files, etc. Contains analytical ETL (Extract/Transform/Load), data mining, predictive and evaluation operators (such as validation types, performance measurement, etc.) (Rapid-I - RapidMiner data mining and predictive analytics)

RapidMiner provides not only a great graphical user interface, but also during the workflow construction has the embedded on-the-fly error recognition and recommended quick fixes (right click on the red marked error appeared on the process area and select from the quick fixes the appropriate variant) [Appendix F](#)

One of the *RapidMiner* extensions is related to *PMML*. This *PMML* Extension is adding a new operator for writing of data mining models into *PMML* standard in order to be shared between various applications and platforms.

As presented in this chapter, the multitude of vendors creating enterprise or open-source predictive analytics vertical solutions or suite workbenches raised their tool integration challenge in customer's production environment that usually contains several applications and different databases types, especially because of their proprietary *predictive models* codes written in a collection of programming languages that may vary between R/C/C++/Python/Java/.NET. Therefore since 1997 Data Mining Group (DMG) was preoccupied to develop and release the open standard PMML (Predictive Model Markup Language) programming language in an effort to permit the sharing of the *data mining* and *predictive models* between PMML compliant applications.

2.3 Brief description PMML components; PMML version 4.1 and supported data mining model (Decision Trees; Support Vector Machines; Neural Networks; Naïve Bayes; Regression; Scorecards; K-Nearest Neighbors (KNN); Clustering Models; Association Rules)

The businesses driven market proved the need to develop predictive models to be used in broaden cross-industry opportunities. Therefore emerged the variety of powerful predictive analytical tools, many of them open-source as we have seen before, all in an attempt to build effective predictive models.

However the biggest challenge appeared when trying to integrate and deploy the predictive solution within a production IT infrastructure. Due to custom codes or in-house proprietary processes the integration and deployment can be spread over many months and with great chances of early failures. Therefore *Predictive Model Markup Language* (PMML) *standard* based on XML (Extended Markup Language) gained broad industry support due to its vendor agnostic property (platform and application independent), the users being able to seamlessly develop *predictive models* within one application (commercial or open-source) and use another one for their execution in the production environment (can be seen as an interface between model producers and model consumers, PMML producers and PMML consumers). Moreover the benefit increased due to the possibility to apply the predictive models directly in database systems on large amounts of data (Big Data phenomena) that can be available in a hybrid cloud-based system (including the *public* one like Amazon Elastic Compute Cloud / EC2 or the *private* one as for example based on distributed clusters / Apache Hadoop farm on a shared infrastructure). With respect to CRISP-DM (Cross Industry Standard for Data Mining)

processes, PMML permits a clear task split especially related to *model development* and *model deployment* and eliminates the need of custom code or proprietary model deployment. [Appendix G](#)

PMML is owned by Data Mining Group, a consortium led by IBM, SAS, Microstrategy, FICO, Equifax, NASA, Salford Systems, Zementis, KNIME, Open Data Group, Rapid-I and others and its specification comes in an XSD form (XML Schema Definition - a set of rules to which an XML document has to comply in order to be acknowledged as “valid” in relation to the schema (Wikipedia -

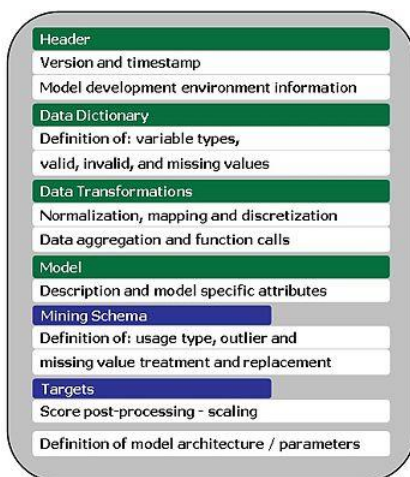


Figure 15 PMML Components

Predictive Model Markup Language).

The purpose of an XML Schema is to define the valid building blocks of an XML document, just like a Document Type Definition – DTD (Wikipedia - Document Type Definition (DTD))

An XML Schema:

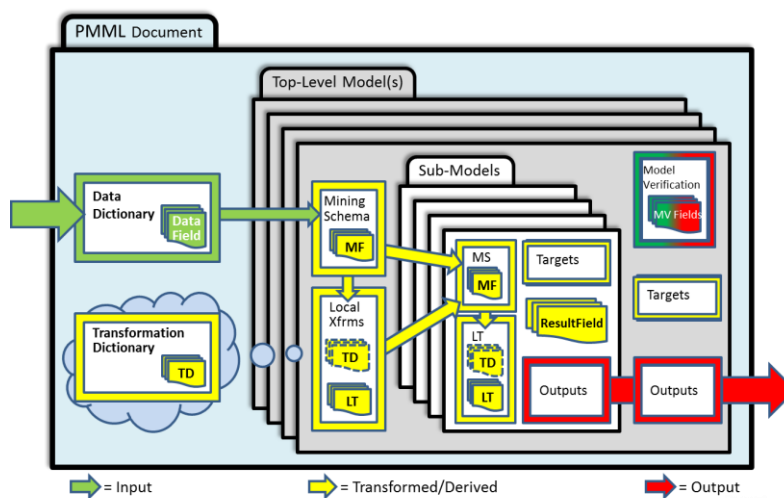
- defines *elements* that can appear in a document
- defines *attributes* that can appear in a document
- defines which elements are child elements
- defines the order of child elements
- defines the number of child elements
- defines whether an element is empty or can include text
- defines *data types* for elements and attributes
- defines default and fixed values for elements and attributes

Its current version, PMML 4.1 has the following component structure, briefly explained in the above photo. (More information under the official link <http://www.dmg.org/v4-1/GeneralStructure.html>)

New PMML 4.1 added three new model elements and enhanced the language with: (Guazzelli, 2012)

- New model elements for representing *Scorecards* (by this having the ability to represent *reason codes* for explaining any adverse actions derived from a scorecard – e.g. case of credit-related decisions or risk of fraud in case of online purchases), *k-Nearest Neighbors* (KNN - instance-based learning algorithm; prediction is based on the K training instances closest to the case being scored) and *Baseline Models* (used for defining a change detection model).
- Simplification of multiple models- the same element is used to represent model segmentation, ensemble, and composition.

- Overall definition of *field scope* and *field names* (<http://www.dmg.org/v4-1/FieldScope.html> - context used to define the visibility and accessibility of *variables* in different parts of the program. This is important when the same variable name is used in different places / module elements within the program, so that name conflicts are resolved without side-effects and ensuring the module independence).
- A new attribute (“isScorable”) that identifies for each model element if the model is ready or not for production deployment.
- Three new built-in functions (thru logical, arithmetic and string operators adding value in data pre-processing steps)
- Enhanced *post-processing* capabilities (apart of the *Targets* element introduced with previous versions - important for scaling implementation, the *Output/s* element has now enhanced functionality - apart of the scaling, now useful in data manipulations - allows transformations and built-in functions to be applied to the output variables). *Outputs* are features of the *predicted field* and so are typically the predicted value itself, the probability, cluster affinity (for clustering models), standard error, etc.; all the built-in and custom functions that were originally available for *pre-processing* only are now also available for *post-processing* (What is supported in every version can be found under the <http://www.dmg.org/coverage/>).



This is the PMML Scope Diagram (part of the Field Scope document) representing the flow of data in a PMML 4.1 document. Field Scope document gathers the *rules* around *field scope* / variable and *field names* that were previously scattered over several documents (fields in PMML are fixed into collections with *pre-defined* scope,

opposed to the declarative software languages like C and Java, which allow any variable to take a variety of scopes (e.g., global, friend and local)).

Legend:

TD = Transformation Data; LT = Local Transformation; MF = Mining Fields; MS= Mining Schema

As data mining is the most important activity in the process of predictive modeling creation, some of the implied ***data mining techniques*** (especially *supervised* and *unsupervised* learning techniques – will keep apart *reinforcement learning* and *game theory*) need their succinct description.

Supervised learning – *modeler specifies what to predict* (named also classification or inductive learning) is a machine learning technique for creating a function from *training data* (pairs of input objects and desired outputs – data set collected in the past, analogue to human learning from past experiences). The *output* of the function can be a continuous value (called *regression*), or can *predict* a class label of the input object (called *classification*). The task of the supervised learner is to predict the value of the function for any valid input object after having seen a number of training examples (i.e. pairs of input and target output).

Or much easier said in *supervised learning* the machine receives a sequence of *inputs* and is given a sequence of *desired outputs*; the final goal of the machine is *to learn* to produce the *correct output* given a *new input*. This output could be a class label (in classification) or a real number (in regression) (Data mining articles - Data mining)

- ***Bayes Classifiers***

Bayesian classifiers use a probabilistic approach to classifying data and are mostly used in case of large numbers of input attributes. Most often used technique is *Naive Bayes* (“naïve” is coming from the fact that input attributes should be independent of each other (no correlations between them, which often is not true).

Bayes create an overall probability of an event to be true by combining the conditional probabilities. Bayesian factor can be used to determine the best fit algorithm related to a given data set (and does not depend on the parameters used by each model). Additionally includes a penalty against too much model structure, which works against one of the mentioned challenges, the over-fitting.

- ***Decision Trees***

Is literally a tree of decisions (flow-chart like structure) and it creates rules which are easy to understand and code. The internal node represents test on an attribute, each branch represents outcome of test and each leaf node represents class label (decision taken after computing all attributes). A path from root to leaf (the very last nodes where the target variable is categorized) represents a classification rule.

Important is how we order the decisions (the order in which we apply attributes by choosing a variable at each step that best splits the set of items) to create the tree. To help in creation of the best tree are coming algorithms such as *Gini impurity* (probability calculations used to determine tree

quality - (Wikipedia - Decision tree learning)) and *information gain* (which uses entropy calculation concept from information theory).

In large data sets it may happen that the leaf nodes become barely populated with just a few entries in each leaf which can create problems for generalization (for the predictive case, the predictive capability decrease when the leaves contain few records). Therefore most data mining tools support *pruning*, removing sections of the tree that provide little power to classify instances. Usually the level of pruning is determined by trial and error to calculate the best predictive capability.

A decision tree consists of 3 types of nodes:

1. Decision nodes - commonly represented by squares;
2. Chance nodes - represented by circles;
3. End nodes - represented by triangles (Wikipedia - Decision tree learning)

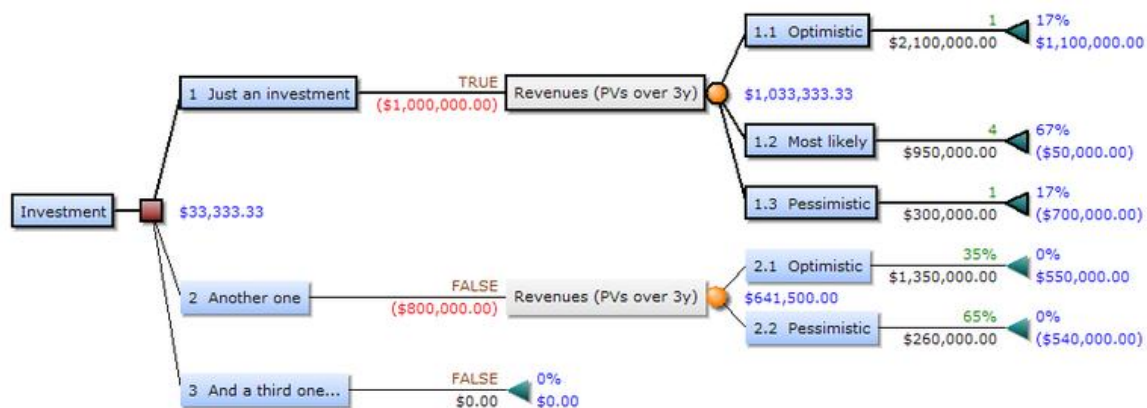


Figure 16 Decision tree used to optimize an investment portfolio (bold lines mark the best selection)

- **Regression** – predicts a numeric outcome based on a set of specific inputs (inputs and output are pre-defined). Used in Marketing Campaign response rate.
- **Scorecards**

Based on regression models, scorecards are a popular technique used by financial institutions to assess risk. With scorecards, all data fields in an input record are associated with specific reason codes. During processing, data fields are weighted against a baseline risk score. After the fields with the highest influence on the final output are identified, their associated reason codes are then returned together with the output (Guazzelli, IBM - Predicting the future, Part 2: Predictive modeling techniques, 2012).

- **Nearest Neighbors (k-NN)**

KNN is a mechanism for data / object classification by establishing the closest neighborhood a particular record (training examples) lives in.

An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its k nearest neighbors (k is a positive integer, typically small). If $k = 1$, then the object is simply assigned to the class of its nearest neighbor.

- **Neural Networks**

Neural networks are used for prediction, classification or clustering (development of self-organizing maps (SOM)). They mimic the behavior of neurons within the brain in a complex nervous system, with inputs from the environment, processing the inputs in order to create the output. The neurons are nodes in a neural network. Also in this data mining technique is important the proper selection of relevant inputs (features to be used as input), availability of training data (and size of the training set) and understanding of the target output.

Usually the neural networks have three layers - the *input* / *hidden* / and the *output* layer. The hidden layer has no direct contact to inputs or outputs. Important is how big should be the hidden layer (number of hidden layers and nodes per hidden layer - too large will make the network to memorize the whole training data set, so will destroy the predictive capability; too small will result in missing patterns). Also relevant are parameters such as combination and transfer functions.

- **Support Vector Machines**

Support Vector Machines (SVMs) is one of the powerful classes of predictive analytics technologies. They classify the data by separating it into regions (by hyper-planes in multi-dimensional spaces; the best hyper-plane is the one that represents the largest separation, or margin, between the two classes – in the below example H3).

SVMs are an important component in any machine learning toolkit (almost all vendors are including it – e.g. Weka (Java), Spider (Matlab), Torch (C++)) (Wikipedia - Support vector machine).

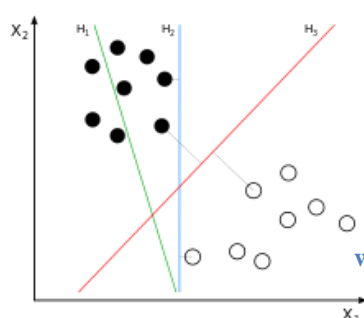


Figure 17 SVM Hyper planes (H1 does not separate the classes. H2 does, but only with a small margin. H3 separates them with the maximum margin); Wikipedia SVM

Unsupervised learning – (inputs, outputs are not pre-defined; outcome is to find a pattern); is a method of machine learning where a model is fit to observations. Unlike supervised learning there is

no a priori output (there are *no supervised* target outputs, *nor rewards / punishments* as result of the actions that affect the state of the environment such as for reinforcement learning). In unsupervised learning, a data set of input objects is gathered; the input objects are treated as a set of random variables and a joint density model is then built for the data set.

- **Clustering**

Clustering is very similar to the k-NN technique but without specifying a particular attribute which should to be classified. Data is fed into the clustering algorithm, which based on some techniques will group the set of objects ; the objects in the same group (called cluster) are more similar to each other than to those in other groups (clusters). Typical application is in customer segmentation / profiling.

- **Association Rule Mining**

Association rule learning is concerned with the discovery of strong rules which might exist between data attributes and does not consider the order of items neither within a transaction nor across transactions. A typical application is for *market basket analysis* where the input is gathered from the retail chains devices (Point-Of-Sale transactions, deducting regularities between the purchased products).

The ultimate application in the last Knowledge Data Discovery process in the *post-processing* part is the *predictive analytics* having as base the predictive model that can also rely on *modeling techniques* at the same time (combined together – *model ensemble* as described in the following article (Guazzelli, IBM - Predicting the future, Part 2: Predictive modeling techniques, 2012).

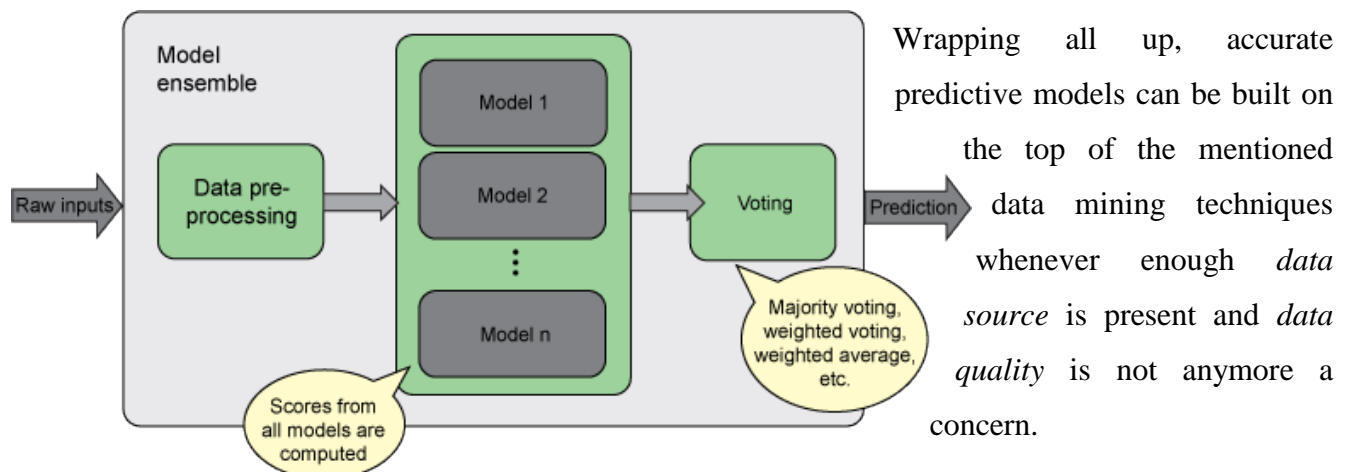


Figure 18 Model ensemble - scores from all models are computed and the final prediction is determined by a voting mechanism or the average

3 Implementations in Telecom domain (Social Analytics as stand-alone or in a complex Customer Experience Management solution)

3.1 Brief mobile technology evolution description

Telecommunication is one of the vast and most demanding domains which changed the way people received the news. In the recent years we assist to the continuous development of internet based services including content delivery networks (CDNs) serving most of the internet content web objects (text, graphics, URLs and scripts), downloadable objects (media files, software, documents), applications (e-commerce, portals), live streaming media, on-demand streaming media, and social networks as well to the booming of newer generation of smarter wireless user equipment / devices with lot of resources for various always-on-always connected applications devices (most of the wireless devices, smartphones, tablets, laptops, USB dongles having 3G capabilities / High Speed Downlink Packet Access (HSDPA) – e.g. Downlink (DL) 42 Mbps, Uplink (UL) 12 Mbps). In the mobile communication this dramatic increase of data exchange brought the need for newer generation of systems more oriented to the faster *packet-switch* data ((network resources are consumed only when users are transferring the data) and networking is Internet Protocol (IP) based) as opposed to the classical one merely based on *circuit-switch* data (open data connection must be maintained, the network resources even when idle (e.g. standard voice connections)).

3GPP (Third Generation Partnership Project) was founded in 1998 by standardization bodies from Europe, Japan, South Korea, USA and China (<http://www.3gpp.org/About-3GPP>) to create the technical specification for the third-generation of mobile systems (Universal Mobile Telecommunication System / UMTS – known as **3G**; the initial work was inherited from the European Telecommunications Standards Institute (ETSI) which defined during 80s/90s the Global System for

Mobile Communications (*GSM*) standard – known as **2G**).

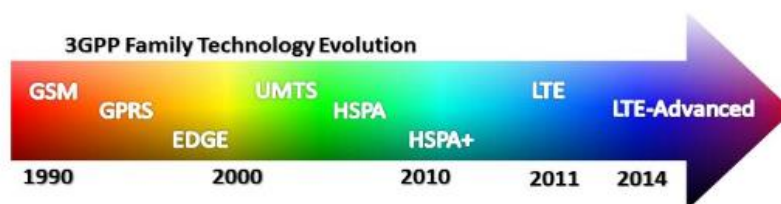


Figure 19 3GPP mobile technology evolution

When the specifications of Long Term Evolution (*LTE*) contained in 3rd Generation Partnership Project (3GPP,

Release 8) were just finished, then the newer Long Term Evolution Advanced (*LTE-A*) standard (starting with Release 9) has shown up.

LTE-A exceeds the requirements imposed by International Telecommunication Union (ITU) to Fourth Generation (**4G**) mobile systems, called International Mobile Telecommunication Advanced (IMT-A). Some of the IMT-A 4G standard requirements are coming to differentiate the *peak speed* requirements for 4G service in case of **high mobility** communications (from train or car) at 100 Mbit/s from the **low mobility** communication (stationary users, pedestrians) at 1 Gbps and DL bandwidth of 100 MHz. (Guillaume de la Roche, 2012)

Some of the requirements are welcomed by the mobile operators:

- *Spectrum flexibility:*
 - ❖ Use of new, re-farmed or unused spectrum
 - ❖ FDD and TDD
 - ❖ Variable channel bandwidth
- *Performance:*
 - ❖ Higher peak rates
 - ❖ Higher bandwidth
 - ❖ From start designed for "always on applications"
- *Cost:*
 - ❖ IP-based flat Network Architecture - (no circuit switched domain)

(Traditionally the mobile communication system contained three parts - *User Equipment* / Terminals, Radio / *Access Network* and *Core Network* (containing circuit-switched and packet-switched domains))

- ❖ Low OPEX
- ❖ Simplified operation (less configuration, higher degree of self-configuration)

Users will benefit of high quality standards as for optical fiber, very low latency in the *user* and *control* plane, very high capacity in mobility conditions and global roaming capabilities.

The latency at the *control plane* is the time in the transition between two connection states (e.g. from the idle state to the active state); *user plane* latency is defined as the time elapsed since the Internet Protocol (IP) packet is available at the base station (BS) until this packet is properly received by the IP layer of the end user).

To solve the capacity and quality challenges *LTE-A* is coming with technological improvements (based on *multiple antennas* at the transmitter and receiver, advanced Multiple Input Multiple Output

(MIMO) radio channels techniques based on multiple antennas, new transmission schemes with multiple carriers, like Orthogonal Frequency Division-Multiplexing (OFDM) in the DownLink (DL), machine-type communications, etc.) (Guillaume de la Roche, 2012)

From historical point of view every *technology generation* lasted at least 6 years and as we can see from the latest research information, the older technology based on GSM/HSPA is still occupying about 90% of the global mobile market shares with about 6.5 billion cellular connections (4G Americas).

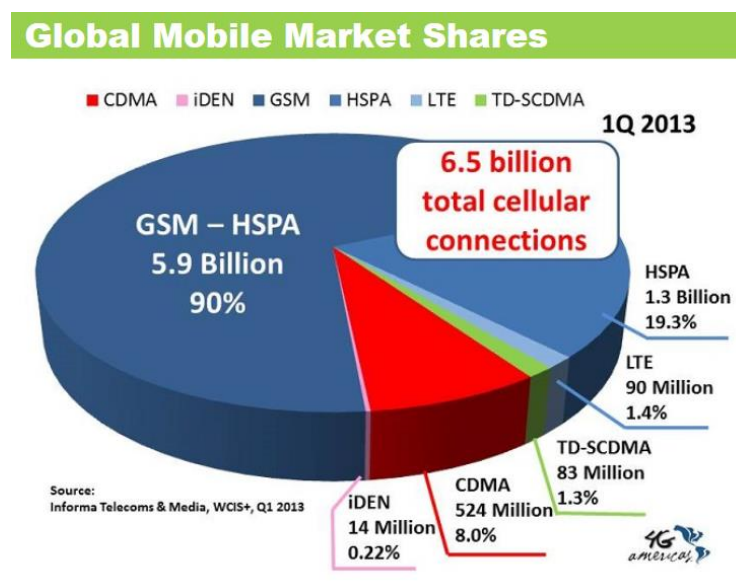


Figure 20 Global Mobile Market Shares; Informa Telecoms & Media, WCIS+, Q1 2013

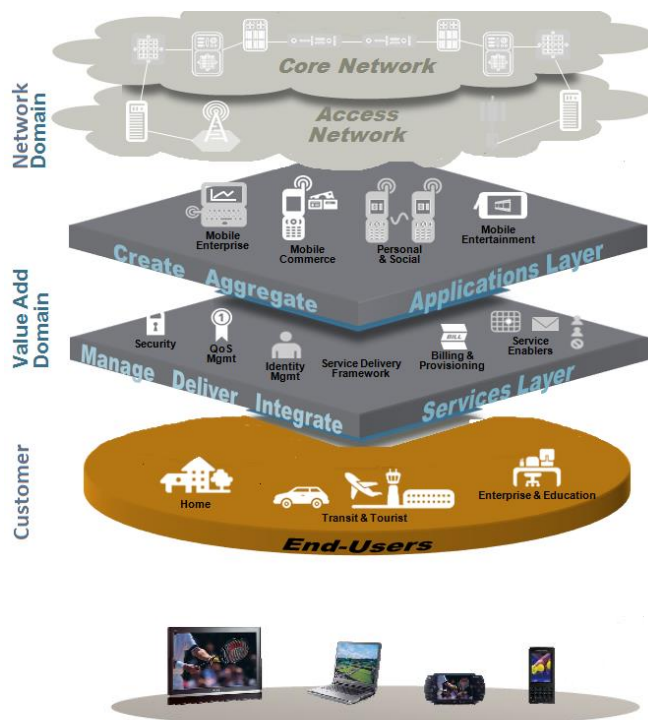
Therefore in the current paper I will consider the current field reality in the mobile telecom world with data gathered from mixed network, service, products and application types.

3.2 Data flow in an End-to-End Telecom business (from raw *data collection*, including aggregation and transformation → *insight* (reports, dashboards) → *Analytics* (predictive modeling) → *Actions* (e.g. business decisions, marketing targeted promotions))

The below picture is a high level description of a typical mobile communication system (user / customer equipment or terminals, access network (radio or non-radio) and core network) and of a business related value added domain as a wrap-up of the application and service layers. Both value added layers rely on the communication system infrastructure to create, aggregate vendor agnostic

applications or to support, manage, deliver and integrate various services (including Cloud based). As for the users they can have different mobility profiles (more static or fixed to dynamic ones) and can be part of the public (institutions) or private environment (individual and enterprises).

Figure 21 High Level Definition of a Telco environment



For any vendor commercializing vertical or end-to-end solutions for Telecom domain, this mixed customer landscape is doubtless presenting numerous *technical* challenges (with respect to data transmission, size and location, its storage and backup / restore, processing, maintenance, security aspects – due to various expectations, technological capacity limitations, global or internal regulatory issues or policies) as well as *business* challenges (business model related) in order to make the value added solutions profitable for mobile operators, Customer Service Providers (CSPs) or themselves (overcoming the competition). Needless to say a strong cooperation should exist between CSPs and Telecom solution vendors to provide a higher level of quality of services (QoS) at affordable and customizable tariff plans (including subscriptions) that will ensure an increased end-user satisfaction.

From commercial point of view, business flow in a Telco environment is grouped as expressed in the below picture (extracted from NSN's Customer Experience Management executive slides) which follows a KDD process relying on the gathered mixed data (*data collection and aggregation* - network, services, subscriber, devices) to prepare it for reports (business intelligence / *descriptive* analytics related) – task of *insight –reporting* block or take it to a more dynamic and *predictive / prescriptive* step thru the *Insight – Analytics* block which is responsible for the next business decision management

step (here exemplified by the *Action* block) in an effort to utilize the resulted knowledge. In a Customer Experience Management (CEM) solution, all dashboards, reports and actionable parts are gathered in a user-friendly *portal* and all use cases (e.g. High Value Customer Insight, Churn Prediction, Device Configuration / self-service, Service Quality, etc.) are presented as CEM package subscriptions that are available as per CSP's business request. The *Action* block will take further the responsibility to transpose the business decisions into actions, enabling new services, applications, applying marketing targeted promotions (e.g. learn from data and predictive capabilities will trigger re-charging / including bonuses, cross-sell, up-sell service / device campaigns, etc.) or lead the operation and customer care support activity based on various service / Apps enablers (e.g. using Software Delivery Platforms (SDF) framework)).

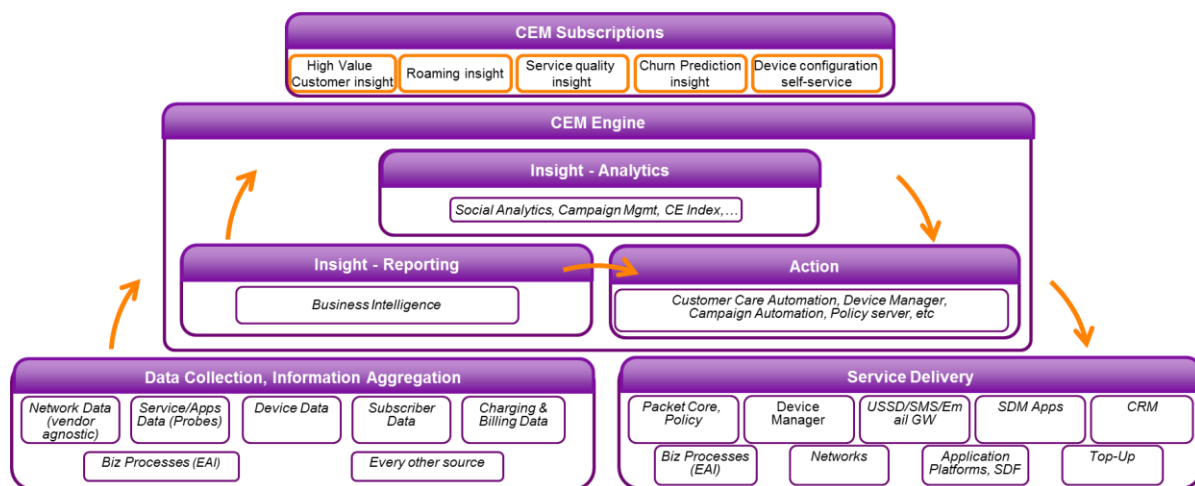


Figure 22 Data Flow in NSN CEM solution

The previous chapter shown the solution based on *predictive analytics* as core element in a Telco business decision management flow meant to capitalize the resulted relevant information in customized actions on the path to CSP's revenue maximization.

It was already stated that the predictive analytics are following the *Knowledge Data Discovery* (KDD) process to generate a *prediction model* to be applied to the production environment to sustain the decision maker's activities and improve their success rate (more efficient business processes, network faults / location and identification, quality of service / QoS issues, customer behavior, customer needs, supply of customer-oriented services, etc.)

Now let's succinctly describe the KDD process (as depicted below) involved throughout a predictive analytics solution.

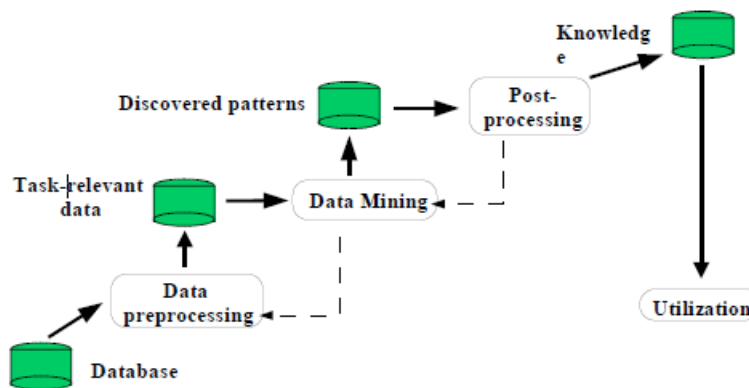


Figure 23 Data Mining methods / Spring 2005

KDD as iterative process is composed by:

- *Data pre-processing*
 - *Data cleaning* (solving the missing values, smooth noisy data (solving the errors, identify or remove the outliers (data objects very different or inconsistent with the remaining set of data), solving inconsistencies (code or name discrepancies));
 - *Data integration* (source coming from multiple database types, raw files, etc.);
 - *Data transformation* (smoothing – remove noise from data, aggregation (summarize, data cube construction), generalization (hierarchy climbing concept) and normalization – to fit in a smaller specified range (min-max, decimal scaling))
 - *Data reduction* (dimensionality reduction – feature selection (attribute subset), heuristic method (decision tree induction, etc.), attribute discretization, concept hierarchies (reduce the data by representing it from low-level to high-level / e.g. age attribute can have instead of numeric values – young, middle-aged, senior));
- *Data mining* (pattern discovery using data mining algorithms)
- *Data post-processing* (pattern evaluation (interestingness (evidence / statistical significance, confidence, etc.), interpretation, visualization, *utilization* of the discovered *knowledge*)

It is said “knowledge is power”, thus the benefit of predictive analytics solutions that are coming to understand the provided data and to better predict the future events.

3.3 *HOW* - Overall solution architecture presentation (e.g. NSN, IBM, KNIME, Rapid-I solutions – churn prediction MVNO /empirical tests); present NSN Cloud-based Social Analytics solution included in Customer Experience Management on Demand (CEMoD) portal –offered as SaaS (Software as a Service)

In the second chapter were presented the open-source predictive analytics tools / workbenches (such as KNIME, RapidMiner) versus the ones offered by domain leaders (IBM – SPSS) and contenders (NSN - Comptel) with respect to their coding (in-house / open-source), statistical R environment integration or the predictive model creator or consumer (PMML input / output format) feasibility.

With the risk that I may be biased by my latest assignment in NSN Austria I would still consider and present our joint (NSN – Xtract / now acquired by Comptel) predictive modeling based solution (Social Network Analytics – based on Comptel Social Links), emphasize its leadership position from *technical* and *business* point of view as per the current situation (fact confirmed by customer testimonials in the delivered projects containing the solution or even by domain related researchers - Frost & Sullivan in their report “Exploring the Use of Social Network Analysis (SNA) in the Telecommunications Industry” (January, 2010)).

Social Links / Social Network Analysis solution was rewarded as winner the second consecutive year (14th of May 2013 - *Innovation in CEM* – *Comptel* (runner-up: Orga Systems); *Innovation in Cloud and Virtualization* - Microsoft (runner-up: Nokia Siemens Networks)) (Pipelinepub - 2013 Competitive Communications Landscape)

Three possible competitors will be exemplified from practical point of view (a high end solution such as IBM SPSS Modeler and two open-source tools namely KNIME and RapidMiner – here a test was performed to check the tool capabilities and its business users friendliness approach); NSN-Comptel Social Links / Social Analytics as a strong niche newcomer will be presented as last.

- From the predictive analytics tool environment **IBM SPSS Modeler** (a powerful data mining and text analytics workbench) is a *high-end* solution that requires analytical know-how and expert intervention (not an automated solution that can be easily operationalized) and is addressed to big customer businesses due to its high project specific effort and costs (following a CRISP-DM methodology approach).

Therefore as per the current status I would not consider it as a threat for NSN's Social Analytic solution (as part of Customer Experience Management on Demand - CEMoD) due to its business target that will be presented later on. The below print-screen from a virtual machine (VM) used for IBM SPSS Modeler trainings will show the extensive possibilities included in this workbench (from the CRISP-DM project / task approach as per Data Mining processes, broad pre-defined data source and flat file formats, till text analytics features, predictive model export in PMML format). Have to mention that even the VM solution was named – “Predictive in 20 Min Solution”, this is true (or even for lesser time) if the data *pre-processing* is already done (depends on the data size, quality, transformation, etc. - normally this counts for minimum 60% of the total predictive modeling process); also very important steps are *variable selection criteria* and the correct *data mining training algorithm* / *multi-model training* (e.g. in classification and regression - bagging, boosting, meta-learning).

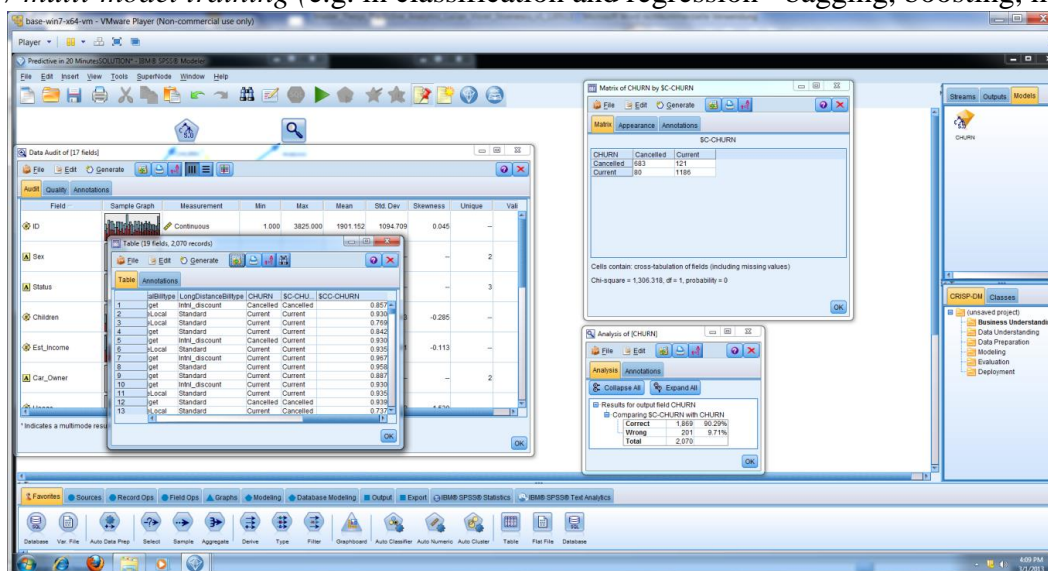


Figure 24 IBM SPSS Modeler - Predictive in 20 min - training VM screenshot

Most of the existing predictive tools managed to solve a principal old issue - to avoid the usage of vendor specific programming languages in order to create

predictive models. The battle is still related to *flexibility* (architecture / technology types and scaling, deployment on the production environment), *efficiency* (“doing the things right” in order to support the operational or business activities and lower the CSP’s costs in B2B technical and business model attractive proposals) and *effectiveness* (“doing the right thing” – here with emphasis on *predictive model* accuracy by selecting the proper techniques and wrap-up the entire solution in an attractive and simple business-user presentation layer).

- Also the main *open-source* competitors are offering predictive tools and workbenches that are still requesting extensive data mining know-how in order to prepare the whole predictive model process. Below an example from **KNIME** workbench related to mobile users churn prediction modeling use case, which speaks for itself related to the *business-user* friendly approach. Nevertheless the latest versions are coming with the R, PMML related extensions which increase its usage in the tech-savvy data analysts world.

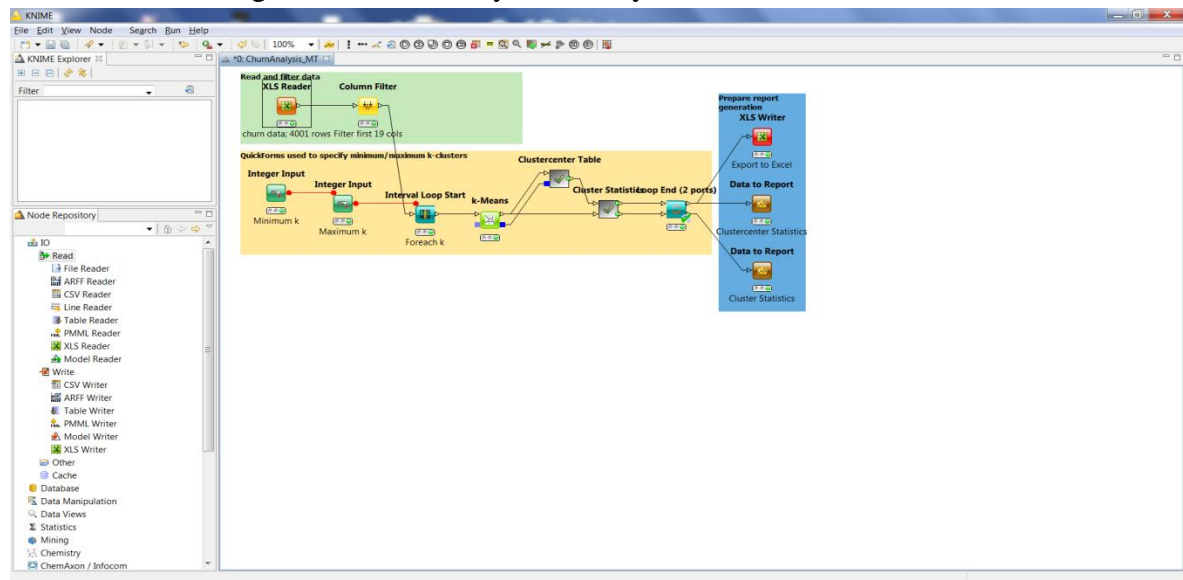
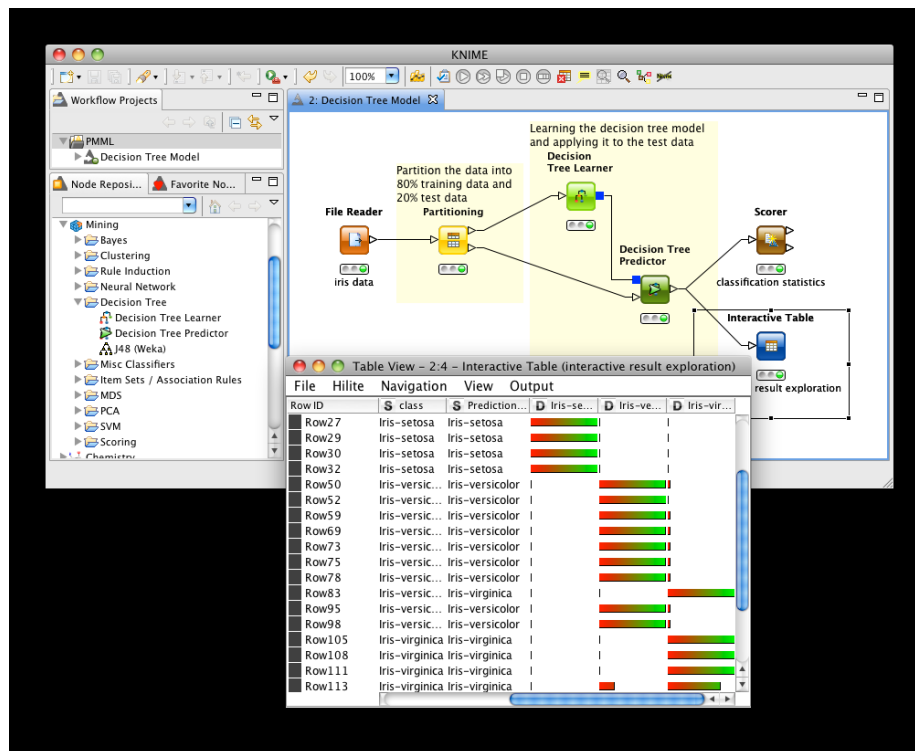


Figure 25 KNIME Churn Prediction sample

The overall process flow to create a *prediction score* is:

- data *pre-processing* – the gathered source data is usually *partitioned* (most used ratio is 70/30 in the favor of *training set* against the *test set*);
- predictive model creation* (here we assume a feature base modeling – which states that data objects are described by a set of features / attributes and the models will have to find dependencies between the features / attributes or predict the final unknown target variable value (highly correlated attributes should be removed as are redundant); the reserved *training data* amount to be used by machine learning based on data mining algorithms has as outcome a prediction model;
- model evaluation* - the quality of the previous detected prediction model is verified by applying it to the out-of-sampling historical *test reserved data* to check for the target variable (label) prediction hypothesis (e.g. determine the accuracy, false positive / false negative hypothesis versus real values / what actually happened)

- *scoring* – applying the model on a new data set and creating a propensity / probability score for a specified target variable



(Dominik
Morent)

Figure 26 KNIME decision tree learning and scoring

- Let's now spend some time analyzing a basic effort to prepare a predictive mobile subscribers churn model using the **RapidMiner** (from Rapid-I) open-source with medium data mining skill-knowledge. This exercise could count for a benchmarking when some of the CSP's operations people will be requested to express their opinion (before or during proof-of-concept (PoC) project) with respect to a specific vendor tool based on predictive analytics and related to the overall solution friendliness from business-users point of view.

This test using the **RapidMiner** has as outcome the predictive model validation of the outcome variable **CHURN_Status** (value – Cancelled, means that the subscriber moved to another operator, thus churned and the value – Current / means that the subscriber is still belonging to the original operator / source of the data records input).

For performing the open-source *predictive analytics test* (using RapidMiner 5.0) the following *input* was used: 4000 Customer Relationship Manager (CRM) *subscriber data* entries saved as MS Excel format (data collected from an Italian Mobile Virtual Network Operator (MVNO) where the MSISDN (the real mobile subscriber number) was mapped to an ID due to data privacy regulations. A number of

17 variables were selected as important for predictive model determination. Data is available as excel format in [Appendix H](#)

The following main steps as per the KDD process approach were performed:

1. Data read (selection of 17 attributes; no other special treatment as data is fully consistent);
2. A validation operator (split - with relative 0.7 shuffled ratio and cross – 10 shuffled validations) was firstly used to partition the data set (in *training* and *test* data sets);
3. Selection in the *training* child process (validation operator is a nested operator containing child processes) of a proper data mining algorithm (the best performing ones for the given training data set was SVM PSO (Support Vector Machine Particle Swarm Optimization) in order to generate the predictive model (whenever needed some attribute conversions are performed – e.g. SVM supporting numerical type attributes);
4. Applying the model against the *test data* kept aside – (part of validation testing child, performance evaluation (accuracy, probability, etc.))
5. Alternatively checking to boost a modeling algorithm/s (here SVM PSO) using *correlation matrix* (correlations between attributes are weighted; highly correlated attributes can be removed as are similar in behavior, thus similar impact in prediction calculation, making them redundant; this will save space and time for complicated algorithms) and *sampling* (from the given data set a sample is extracted; could be *absolute* – sampling size having relevant role, *relative* – ratio important or *probability* type – e.g. probability per class).

(Rapid-I - RapidMiner Operator Reference)

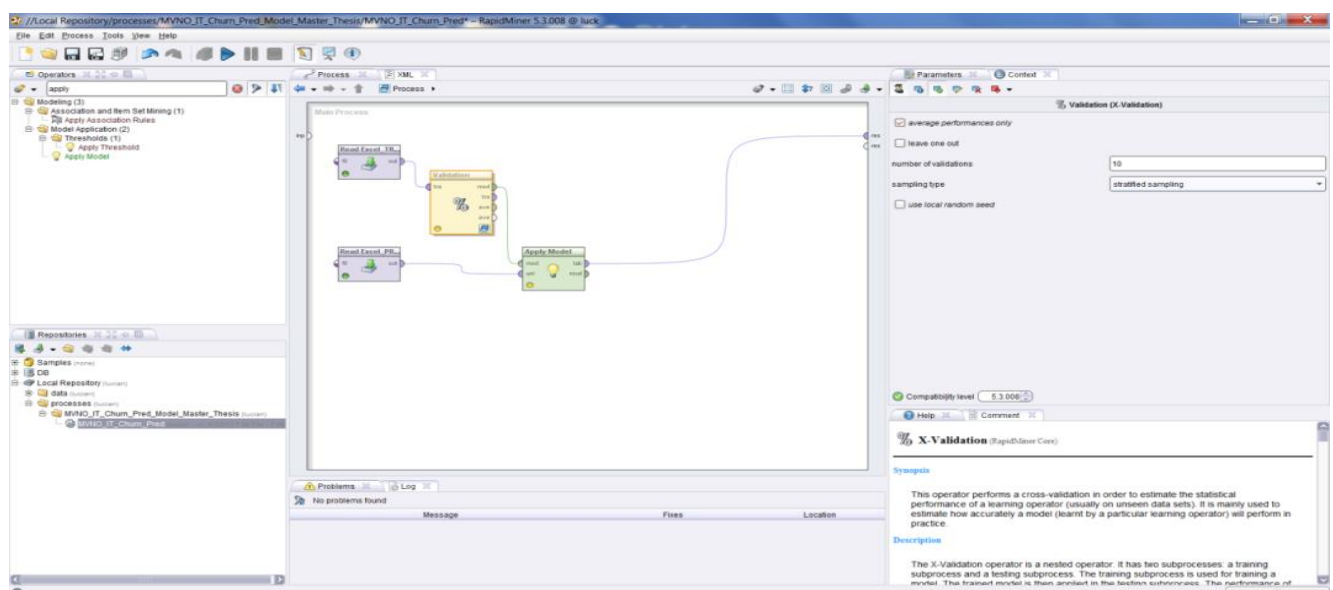


Figure 27 Subscriber churn prediction for an Italian MVNO - using RapidMiner

Some important process print screens together with the final predicted outcome and prediction accuracy are shown below. Here was used a cross validation operator (10 numbers, stratified sampling). Using the cross validation operator two data sets were created (2516 data entries for model training using first the *decision tree* algorithm – gain ratio criterion, 3 pre-pruning alternatives to allow the leafs which are not adding discriminative power to the algorithm to be removed; 1484 data entries were kept for the **testing part – predictive model performance evaluation**).

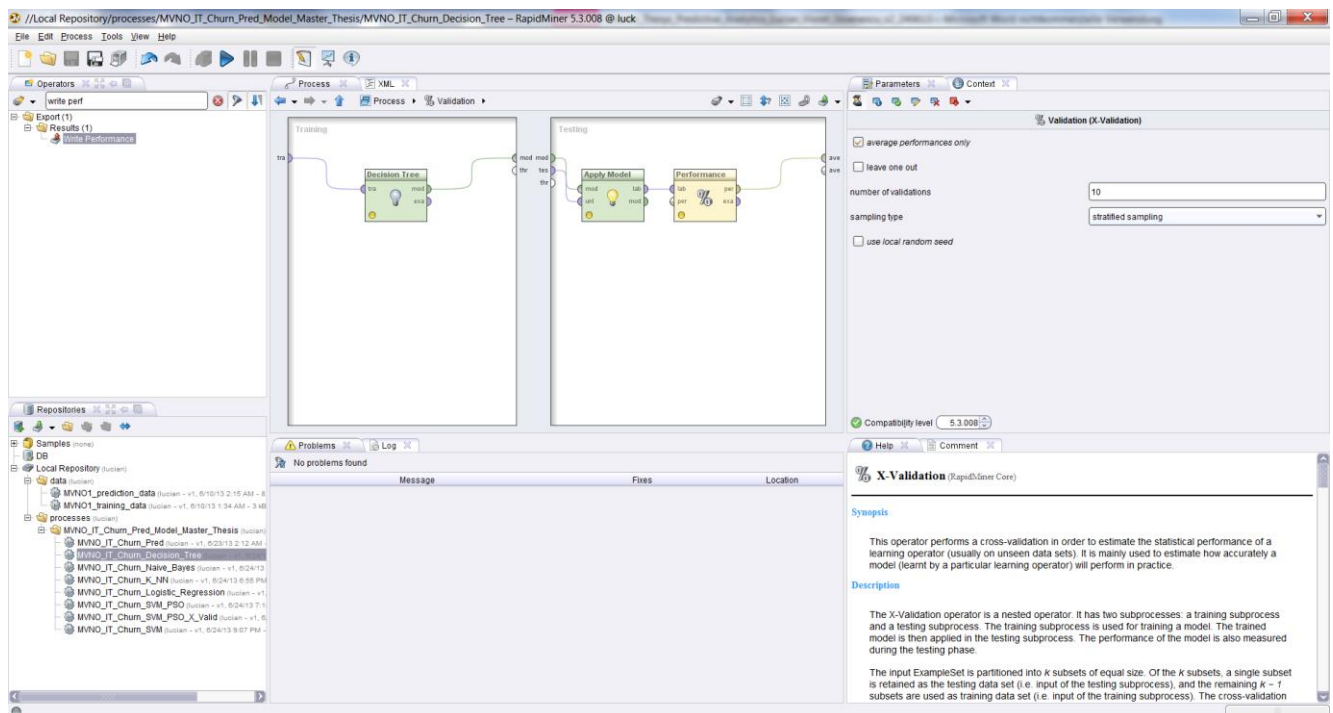


Figure 28 Process design - Churn Prediction model using Decision Tree

Model evaluation performance (when applied to the 1484 test data entries) shown that this type of data mining algorithm is not describing the best the available CRM input data (the targeted / label

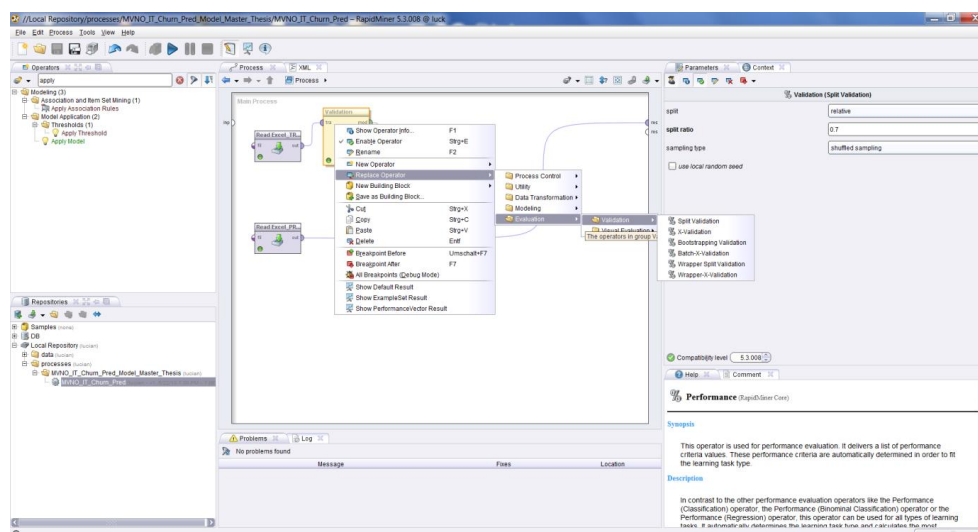


Figure 29 Validation operator change from split to cross type

CHURN_Status binominal variable can have the values Cancelled or Current).

Confidence_Cancelled avg = 0.435 +/- 0.496; Confidence_Current avg = 0.565 +/- 0.496

Important to notice in this Rapid Miner version 5.0 that are semi-automated facilities which offer a great help during the process design phase (in case of missing connections or incorrect connectivity issues between the block operators, some “quick fixes” advices are provided; also the drag-and-drop feature is quite excellent). Moreover the operator change in the process design phase is proving to be great; is enough to select the respective operator and with a click of the mouse you can browse through the targeted operator classes to find the wished one to be automatically changed to. In general this operation will automatically include the link reconnection part; if not the quick fixes will help you thru until all the blocks semaphores will switch to the yellow color meaning that are *ready to be executed* (upon the *correct execution* their color will change to *green*).

Above is also a print-screen example on how easy is to change the validation operator from split mode to cross validation type.

As stated earlier, for this particular data set, the best fitting data mining algorithm was Support Vector Machine Particle Swarm Optimization (SVM PSO) which somehow is understandable due to the fact that the input data was only collected from the Customer Relationship Manager and appear as a loosely structured collection of subscriber attributes (if the input data will contain other samples – e.g. services, network related) it could be that other data mining algorithms will fit as well (e.g. neural networks). Particle Swarm Optimization algorithm is about optimizing a solution to a problem (based on particle swarm, a simulation of a simplified social system, initially started from bird flocking behavior (avoid to collide with the neighbors, fly with the same speed as the neighbors and move to the center of the flock) or ant algorithms for food finding, path construction, nest building, etc. <http://www.swarmintelligence.org/tutorials.php>)

Boosting SVM PSO - overall accuracy 94% (from a sampling size of 100, target variable / label *CHURN_Status*);

- ***pred.Cancelled*** – 35 (31 correct out of the predicted 35 resulting a class precision of 88,57%), true Cancelled 33 (31 correct predicted)→ class recall 93,94%;
- ***pred.Current*** – 65 (63 correct out of 65 → class precision 96,92%), true Current 67 (correct 63 out of 67) → class recall 94,03%

Here the predictive model using *SVM PSO* algorithm is using as described in the action step 5 the sampling (absolute type, sampling size of 100 out of 2516 entries) and correlation matrix operator (check the correlation between all 17 variables to remove the highly correlated ones).

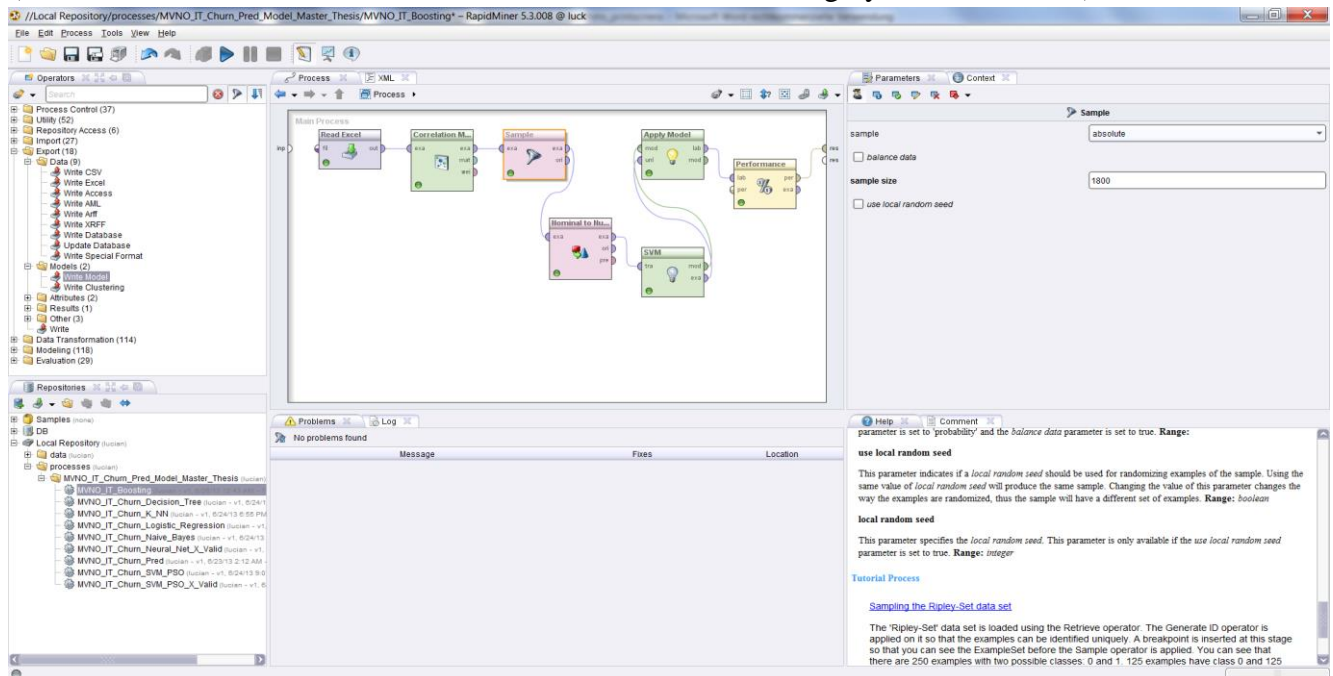


Figure 31 Boosting the SVM PSO algorithm - Process design

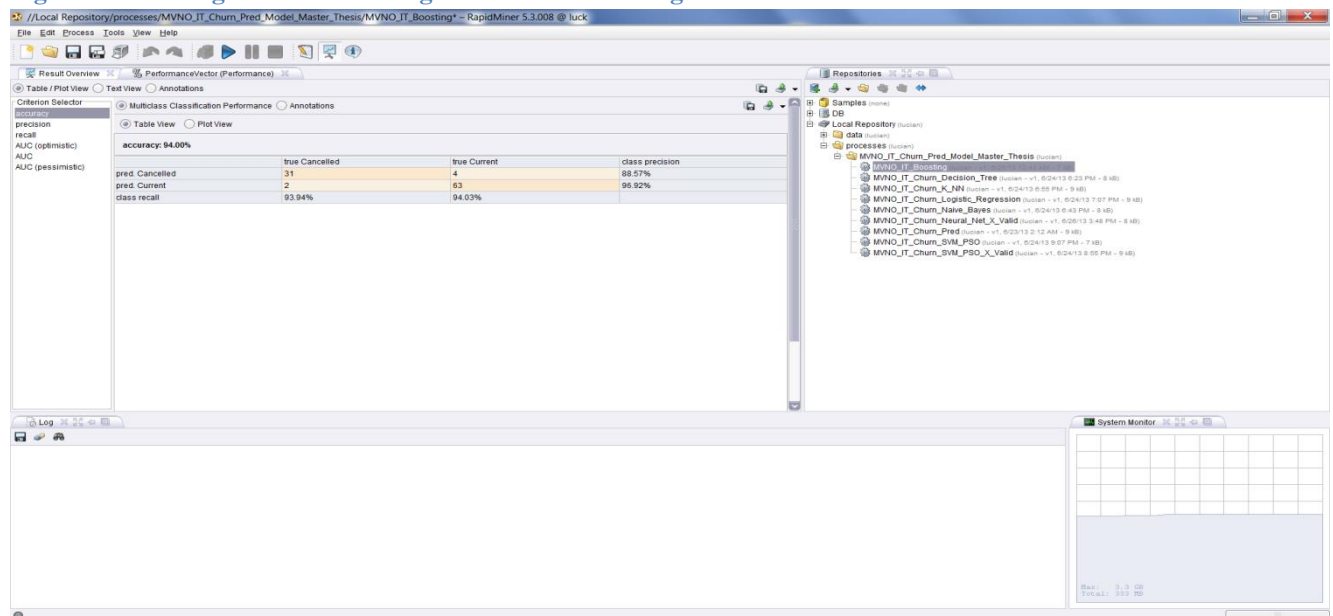


Figure 30 Predictive model accuracy (using boosted SVM PSO algorithm)

The predictive model can be exported in PMML 4.0 format if the PMML extension (for RapidMiner tool) was previously installed from the Marketplace.

Summing up, we have seen a great tool (RapidMiner open source) which is mostly addressed to data analysts and to the eventual CSP's operational team that are trying to benchmark some vendor specific solutions in case of a predictive analytics tender. RapidMiner is not intended for business user

segment, but is very powerful software for data mining. Their *Enterprise* commercial edition (4.5) comes with additional features including reporting engine and the support of multi-core systems (<http://rapid-i.com/content/view/147/1/>). Contrary to the presented predictive analytics tools and workbenches, NSN Social Analytics solution which will be presented below has other *target segment* and is flexible from technical (different deployment types) and business point of view (modular software approach which permits several business models for its acquisition or usage).

3.4 Solution Description and Unique Selling Proposition (NSN-Comptel Social Analytics)

Social Analytics as a predictive analytics cornerstone solution based on predictive modeling complies with the innovation definition as presented by Peter F. Drucker:

“Innovation is the specific tool of entrepreneurs, the means by which they exploit change as an opportunity for a different business or a different service. It is capable of being presented as a discipline, capable of being learned, capable of being practiced.” (Drucker, 2011, p. 17)

Social Analytics uniqueness at this time is ensured by the combined NSN-Comptel’s analytical solution (based on Social Links) that encompasses the best capabilities of the two companies as explained below:

- *On-the-fly* collection and decoding of Telecom specific source data: *network, subscriber profile, customer device* and *value-added services* (from classical SMS, MMS, video streaming to location based services, mobile browsing, charging and billing, Top-Up, etc.). This is in addition to customer related data available in Customer Relationship Manager (CRM) databases, text information available in help-desk tickets from Customer Center databases or business related information. Social Analytic solution has already *embedded adaptor libraries* for an extensive *network probe* domain. This is mainly due to NSN’s strength, know-how and product portfolio including the Telco Operations Support Systems (OSS) and Business Support Systems (BSS). The extensive usage of multiple data sources will increase the prediction model accuracy;
- *Fully automated* predictive tool special dedicated for *C-Level business users* (CxO) to help them in their business decisions without requesting domain related skills (programming or data mining analytics / algorithms related). This is ensured by a *friendly graphical user*

interface (GUI) that can be included in a composed Customer Experience Management on Demand solution (CEMoD) / portal, which permits to be deployed as stand-alone or in hybrid clouds in the form of Software as a Service SaaS (e.g. private clouds such as NSN's, CSP's or public Clouds as Amazon EC2);

- When part of the CEMoD, Social Analytics is also *modular* and presented as a *package* to be installed platform independent, allowing end-to-end solutions and integration with Campaign Automation Systems in the area of churn prevention, promotion & loyalty management, target advertisement, etc. use cases. This presents a *technical* (easy to be installed, maintained, updated especially on Clouds via NSN's Cloud Framework / Gateway) as well as a *business* advantage (CSP's will purchase only the relevant use case for their business strategy development);
- *Predictive Model* monitoring and reporting as part of the model lifecycle (including self-tuning mechanisms);
- Comptel Social Links product as the core of the Social Analytics solution is a recognized Social Network Analytics *leader* with respect to the *prediction accuracy*. Social Links relies on the best in class EMC / *Greenplum Database Architecture* (*Share-Nothing Massively Parallel Processing Architecture*) available also in the product virtualization form to be deployed on the Cloud environment. This avoids the traffic bottlenecks (due to input / output operations in correlation with the bandwidth) or the inefficient traffic between the data source nodes.

<http://www.emc.com/collateral/hardware/white-papers/h8072-greenplum-database-wp.pdf>

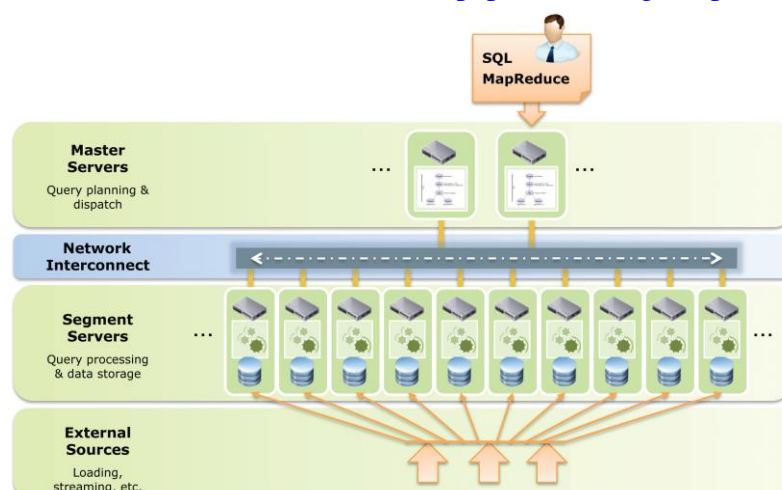


Figure 32 EMC Greenplum MPP Shared-nothing architecture

On the top of the mentioned unique capabilities Nokia Siemens Networks (NSN) has a strong *business consultancy team* (Professional Services including technical part) to help the CSP's if needed

in their offerings, based on the deep understanding of markets, trends and end-user needs, based on studies & surveys conducted on regular basis. Needless to say that NSN has huge operational capability due to its best-shore Managed Service organization which has as base the initial Global Network Operation Centers (GNOC).

Solution Description Social Analytics solution with Social Links as core application is based on the so called *social intelligence*. This stands for a new approach to subscriber profiling and segmentation as a combined measure of both *personal*/psychological and *social*/cultural factors. *Personal factors* are typically related to *demographics* (e.g. gender, age, profession) and *behavior* (e.g. purchase history, usage preferences, experience with quality of services, etc.).

Social factors relate to the social neighborhood, role and connection behavior. Social Analytics generates customer's holistic view based on his behavior, demographics and social network influence.

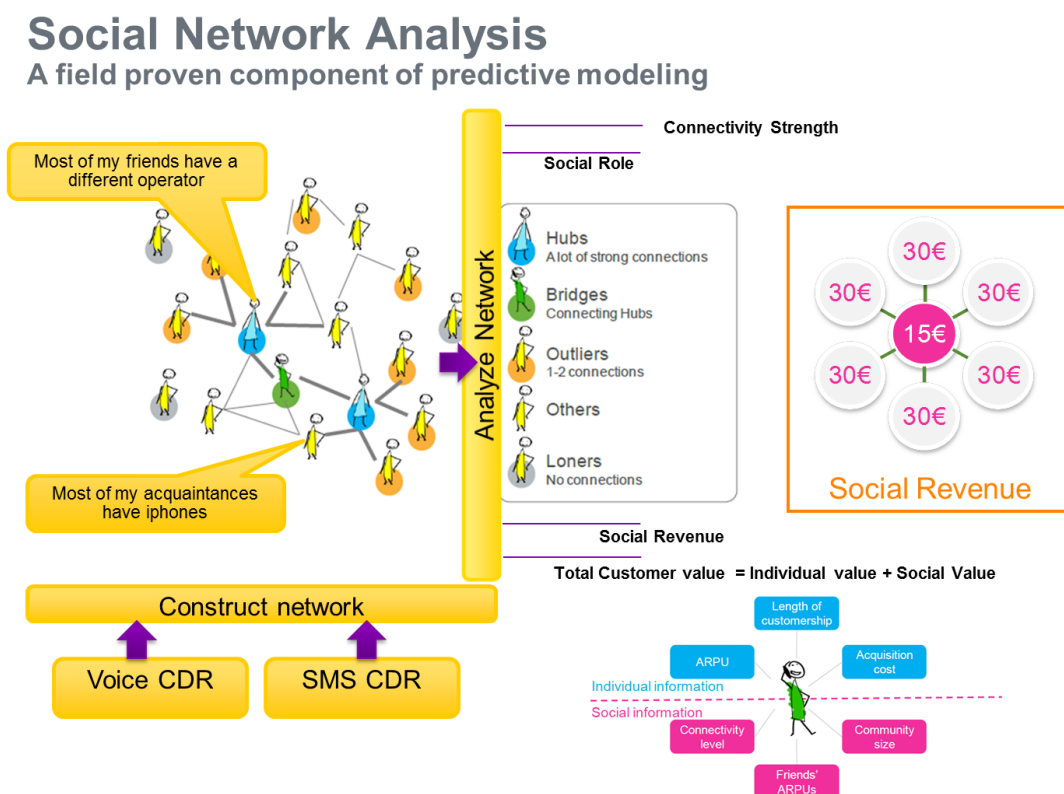


Figure 33 Social Analytics - social network role in predictive modeling

3.5 Social Analytics Basic modules (use-cases)

Below are the main customer challenges tackled by the respective *use-case*:

SOCIAL NETWORK INSIGHT (understand the social phenomena in customer base)

- How are my customers connected
- Who influences whom, who are my social relevant customers
- Who are the right customers to start viral marketing campaigns

Hubs are customers with many connections to neighboring nodes

Bridges are customers connecting two or more Hubs that are not directly connected

Outliers are customers with one or two connections

CHURN PREDICTION (How to reduce the churn of high value customers)

The module scores all the customers based on their probability to churn and influence others to churn. The score is called *Churn Alpha Score* and the customers with the highest scores are called Churn Alphas

- How likely is a customer to churn
- Who are multi-SIM users and rotational churners
- What is the impact if a customer churns on his community
- What is the right offer to retain customers
- Which are the customers that will positively react on the retention offers

CAMPAIGN OPTIMIZATION (How to optimize direct marketing campaigns)

- How likely is a customer to accept an offer
- What's the impact of this customer on his community
- What is the right incentive for individual customers
- How to make sure that marketing budget is spent on the right customers

TARGETED ADVERTISEMENT (How to use customer data for targeted advertisement)

- How can I make sure that my customers' demographics data are correct
- How can I enrich my customer profiles

- How can I find the best fitting targets for advertisements

3.6 Measuring the prediction accuracy (some point in time after the new service launch in the market)

First the number of churners is to be identified

Number of churners calculated -> $X_{churners}$ (Social Analytics calculate and deliver top most likely churners)

Number of churners calculated -> $R_{churners}$

Solution Improvement Reference model is used to select top, e.g., 10k most likely churners

(Reference model: random or operator's churn model); $(X_{churners} - R_{churners}) / R_{churners}$

Success evaluation (Key Performance Indicator (KPI) Solution Analytics Improvement > 10% → Success)

3.7 Deliverables (Social Analytics outputs)

Churn Alpha List containing

Subscriber ID	Identifier (number) of the subscriber
Churn propensity	Categorization of the individual propensity to churn
Churn Alpha	Categorization of the calculated Churn Alpha score

Product Alpha List

Subscriber ID	Identifier (number) of the subscriber
Product ID	Product identifier
Product propensity	Categorization of the individual propensity to take the product
Product Alpha	Categorization of the calculated Product Alpha score

Acquisition Alpha List

Subscriber ID	Identifier (number) of the subscriber
Acquisition Alpha	Categorization of the <i>calculated Acquisition Alpha score</i>

As for input data (minimum of 6 months historical data):

- Subscriber *usage data* (Call Data Records / CDR files) – containing service session details (e.g. voice call, SMS, etc.) from all subscribers that have been active during the specified time period. Each session should be represented in the data file by a single line.
- Subscriber *profile data* (Customer Relationship Manager / CRM files) – containing descriptive subscriber profile data from all subscribers that have been active during the specified time period, including new and churned subscribers. Data should be grouped in monthly files for the specified time period with at least the subscriber identifier and the date when has churned. By default, when no churn information is provided, it is considered a subscriber to have churned when no activity has been initiated by the subscriber (sessions or top-ups for pre-paid) for one month.
- Top-up records – containing top-up (recharging) credit actions for pre-paid subscribers.
- Black list (optional) – containing list of numbers to be excluded from the social network analysis (e.g. call center numbers, etc.).
- Campaign profile – containing the characteristics of the advertisement campaign.
- Campaign results – containing the result of past advertisement campaigns

Some prediction tool outputs (churn statistic view and dashboard are posted in [Appendix I](#))

3.8 Solution deployment (cloud-based approach, Software as a Service / SaaS)

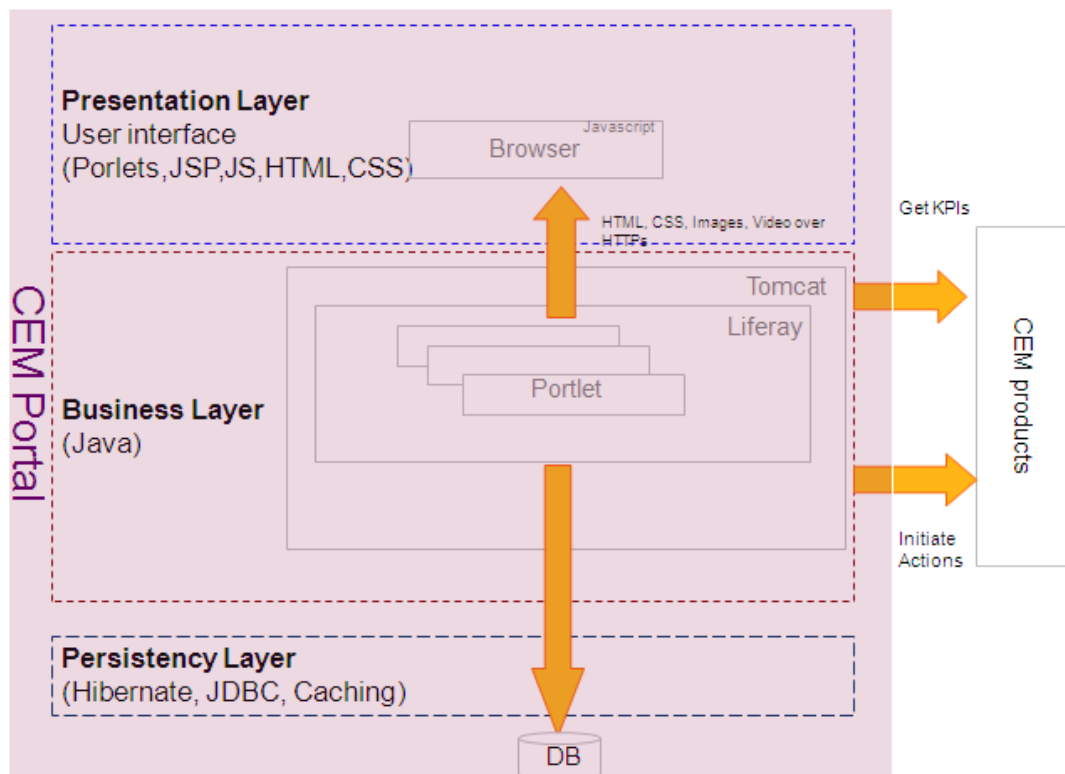


Figure 34 NSN Customer Experience Management Portal

Social Analytics – Software as a Service / its use-cases are available as installable packages in the Customer Experience

Management (CEM) on Demand solution (Cloud based solution – is a CEM portal dedicated especially for business level users).

The CEM Portal is java enterprise web application. Therefore, the content is presented in various technology standards to web applications such as:

- HTML / CSS / Java Script / JSP / Servlet

The user interactions are captured by HTML forms and also JavaScript. Social Analytics as part of CEMoD is following the XaaS High Level Architecture approach (any possible combination can be achieved using the NSN Cloud Gateway to permit the virtualized package installation on any infrastructure, any Product as a Service (database types related, in-memory cache, structural / relational databases (RDBMS), etc.) or as Software as a Service (SaaS – e.g. Telco Apps, Enterprise, Machine-to-Machine / M2M Apps (telematics, tracking vehicles / inventory, vending machine status, technical parameter collection from Energy or Automotive fields, etc.)).

The ultimate *scope for SaaS is the GUI portal* which will act as a *Software Marketplace* from where the specific applications (in our case Telco Apps such as Social Analytics) can be accessed in a package / subscription model.

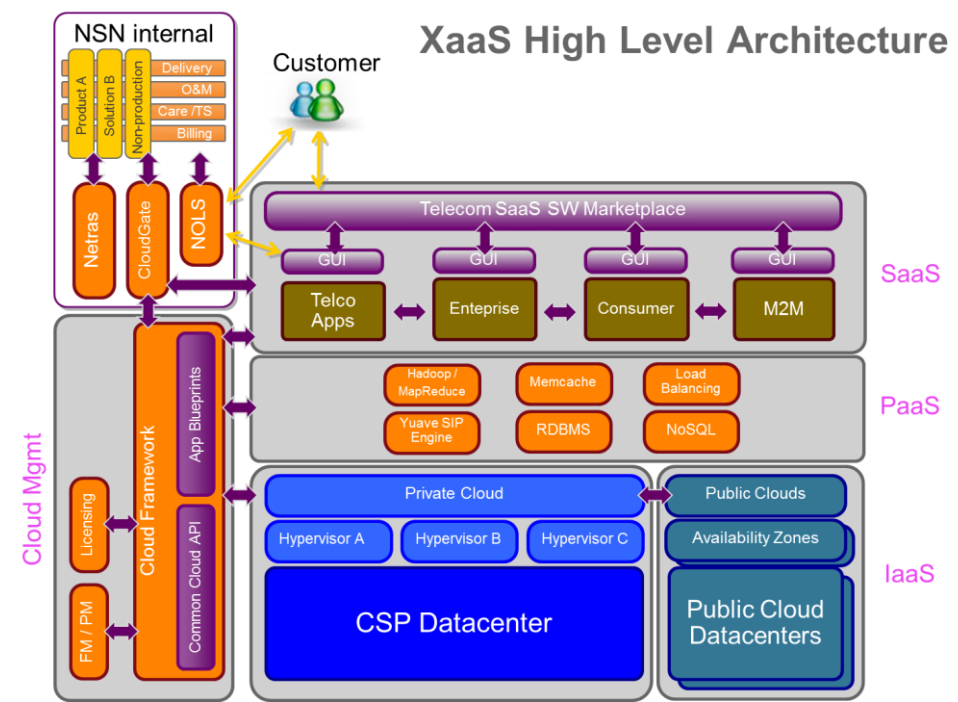


Figure 35 NSN Anything as a Service (XaaS) High Level Architecture using NSN Cloud GW for package virtualization and installation

We have seen until now NSN – Comptel’s Social Analytical solution from technical point of view, we have underlined its mixed *uniqueness* capabilities and its place in solution offerings towards B2B Telecom Customer Service Providers (CSPs), which *need* proper marketing campaigns launches, at the proper time, with the proper message, to the proper subscribers, thus maximizing the campaign take-up rate and implicit their bottom line.

Now let’s investigate this *business opportunity* from corporate entrepreneurial perspective and wrap-it up in a *business opportunity assessment*.

4 Opportunity Assessment - “Who and How?”

4.1 Target Model

The opportunity assessment plan is based on the framework extracted from Prof. Dr. Robert D. Hisrich presentation during Entrepreneurial Leadership module held at WU Wien, October 2012. “(Hisrich, 2012)”

First chapter presented the *need* (WHY) for predictive analytics based solutions (such as social analytics) especially in high-tech cross-industry domains, where business decisions have to comply with the fast changing environment in which they operate.

Later on, chapter 3.4 included NSN-Comptel’s social analytic *solution description* underlying its *unique selling proposition* as per the current multi-vendor tool spectrum.

NSN-Comptel Social Analytics is a fully automated tool offered in a Customer Experience Management on Demand portal deployed on public Cloud (as an installable SaaS package on Amazon Elastic Computing Cloud EC2), tuned for enterprise *business usage* auditorium. Now let’s see our target strategy (*Who* should be our target customers and *How* should look like our business model offering).

The strategy here is to start as a B2B *niche player* in mobile Telecom domain (Telco Applications), where mobile network competencies will act as a high entry barrier, thus avoiding the eventual competitors, as this area perfectly fits Nokia Siemens Networks’ background, experience wrapped-up in its “*market knowledge*” (“*possession of information, technology, know-how, and skills that provide insight into a market and its customers*”) emphasized by its long history in this respect. (Robert D. Hisrich, 2010, p. 69)

From the huge global telecom *markets landscape* (of 5.9 Billion of mobile connections as presented in the 3rd chapter) the initial target (WHO) will be the MVNO (Mobile Virtual Network Operators) *prepaid* customers, firstly choosing to start in Austria.

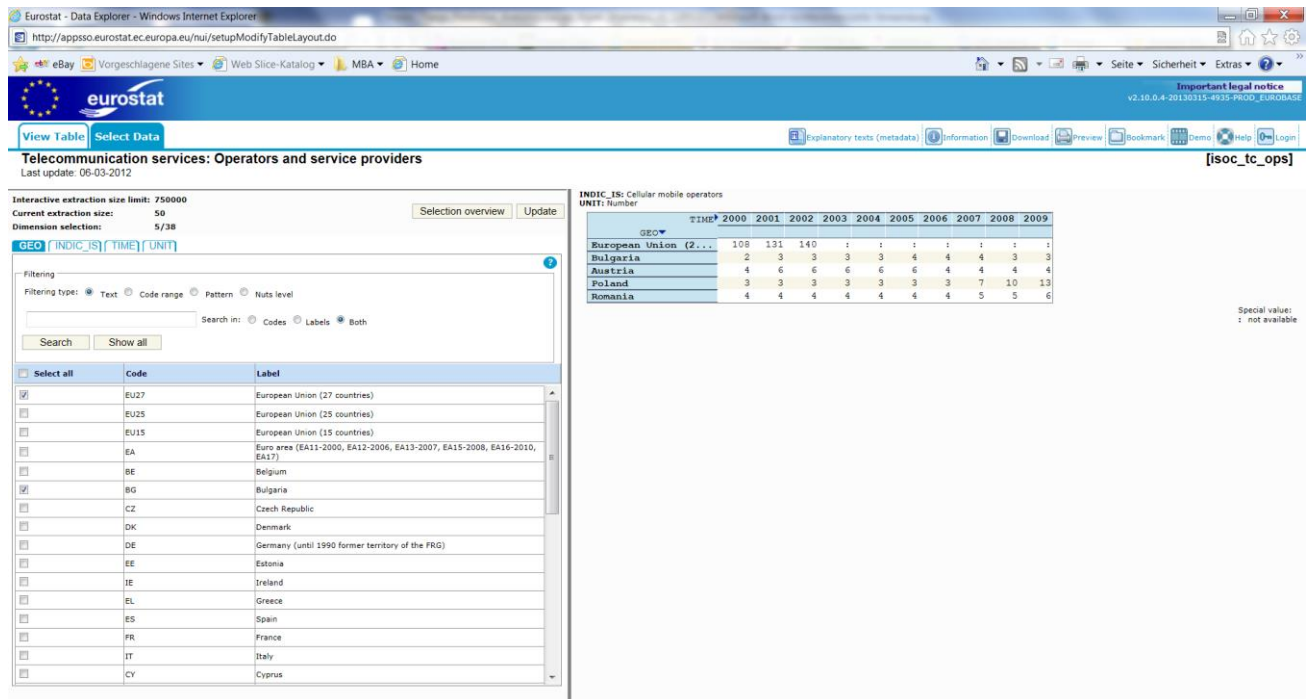


Figure 36 Eurostat - Telecommunication services: Operators and service providers CEE

NSN has strong ties with Tele2 Austria, part of the Tele2 Group due to previous consultancy and service delivery activities. In Austria are 17 MVNOs grouped related to the 3 Mobile Network Operators (A1, Orange, T-Mobile)

<http://www.prepaidmvno.com/mvno-companies/eu-mvno-companies/austria-mvno-companies/#0149>

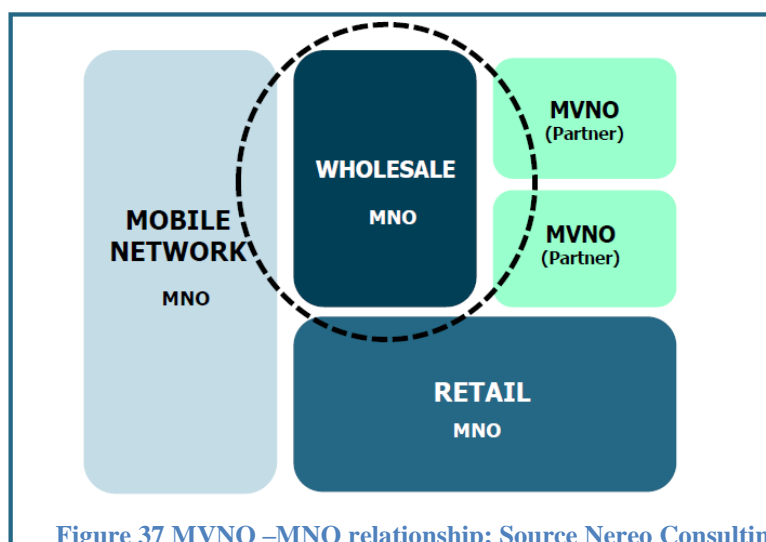
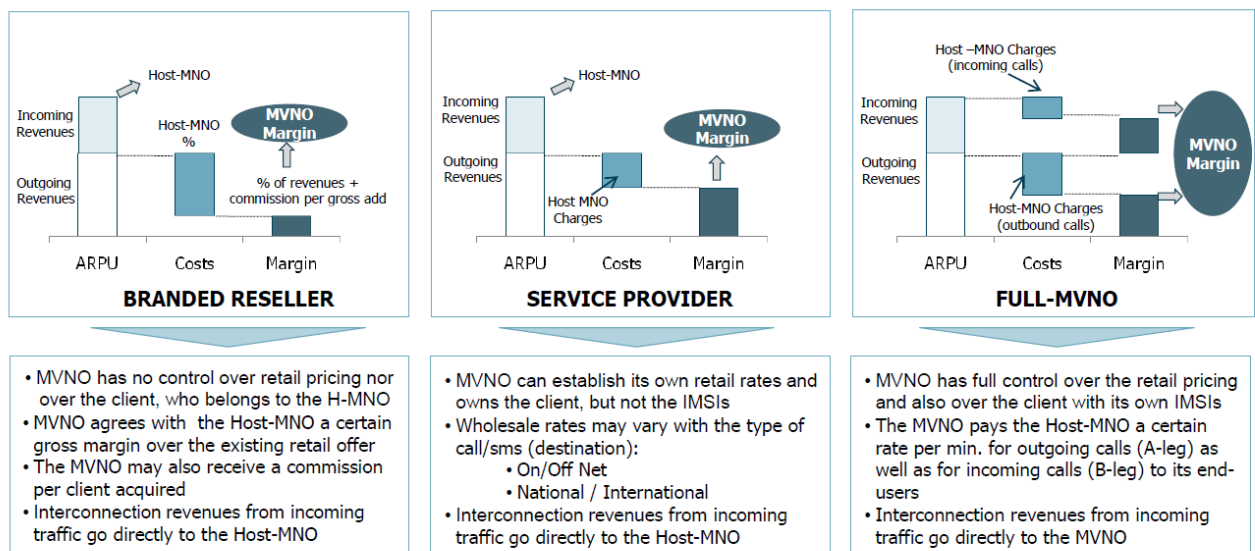


Figure 37 MVNO –MNO relationship; Source Nereo Consulting

Tele2 is a *full MVNO*, has full control with respect to retail prices and the end-users (International Mobile Subscriber Identity - IMSI); additionally its latest license agreements in Russia opens the various business opportunities for 60 mi.

potential subscribers and its impressive business history, as we can see from the latest 2012 Annual Report (38 million subscribers – presence in 11 countries, 7.7% from the net profit due to mobile telephony, 3264 mi. SEK net profit), recommend it as a very good target segment). Eurostat presents, at the level of 2010, Austria as accountable for 41% of market shares - leading operator in mobile Telecommunications (http://www.tele2.com/TL2_AR12_ENG.PDF).

MVNOs can be classified broadly into the following 3 models, each with their specific economic implications for the business



Full-MVNO operational model provides higher margins and total independence from the Host-MNO, and it requires also the lowest effort to be implemented by the MNO

Figure 38 MVNO classification - Nereo Consulting

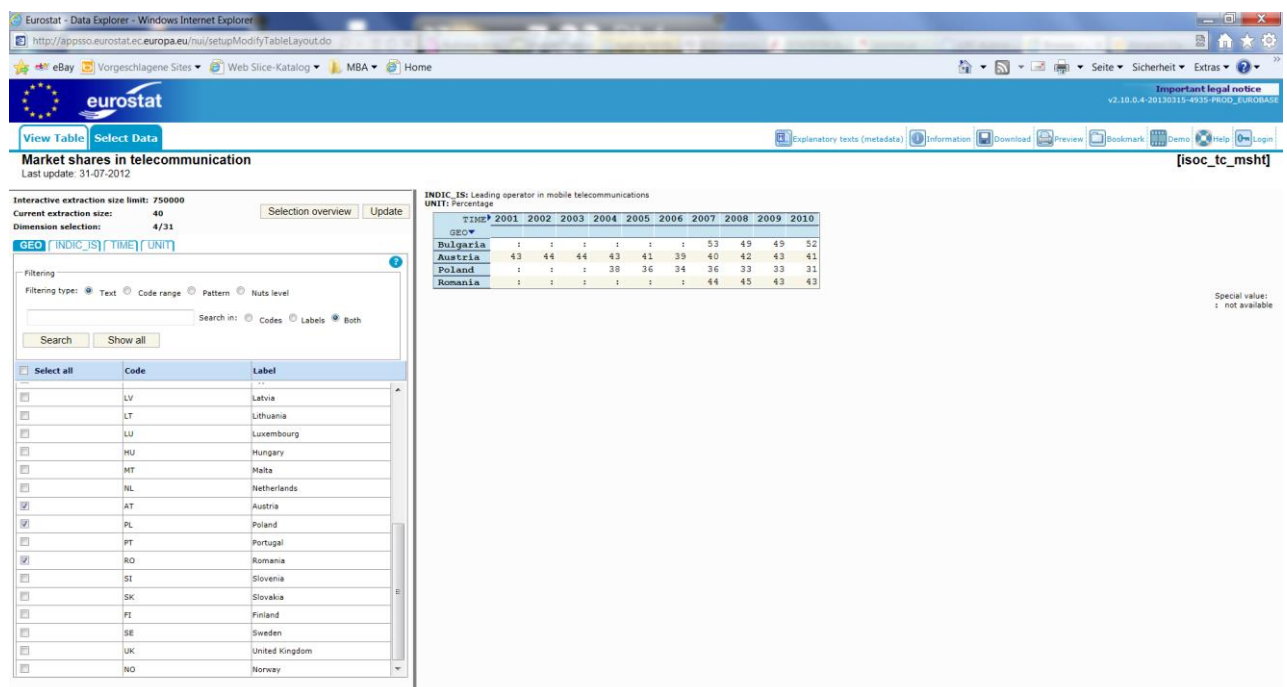


Figure 39 Eurostat - Market Shares in Telecom / CEE

4.2 Business Model and Initial Customers

Solution Analytics will be sold via *direct channel* as NSN has already the sales Customer Team (Account Management, F&C) in contact with the targeted customer, due to older delivered services.

Therefore no need of marketing budget, the eventual announcements will be in the professional network groups (such as Xing, LinkedIn) to spread the news related to NSN-Comptel Social Analytics solution implementation in Europe and publish the Airtel's success story from India. The interested parties may join those professional group discussions and presentations may be available on request.

Pricing parameters to be taken into account for determining the business model are:

- For the presented Social Analytics embedded in CEMoD public Cloud deployment (SaaS on Amazon EC2) -->ARPU (Average Revenue Per User), subscriber number, number of acquired use cases (e.g. social network insight, churn prediction, targeted advertisement support, campaign optimization, customer acquisition)
- If the customer will require private cloud installation (solution hosted in NSN or Customer's premises – offered as SaaS) or if it will be on customer premises (classical installation) then pricing structure will be as in the below figure depending on the selected deployment model

Model	perpetual license	1 year right-to-use license	1 year managed service	1 year hosted service
Description	CSP buys SW, HW and related services	CSP buys time-limited SW license, HW and related services	CSP buys Social Analytics as a Service provided on CSP owned HW	CSP buys Social Analytics as a Service provided on NSN hosted HW
Pricing Elements	SW license, HW, system set up & integration (incl. training), project mgmt, SW & HW maintenance	SW license (incl. maintenance), HW, system set up & integration (incl. training), project mgmt, HW maintenance	service fee, HW (optional), service set up & integration, project mgmt, HW maint. (optional)	service fee, service set up & integration, project mgmt

Figure 40 On-site Social Analytics deployment - Pricing Structure

Growth

Social Analytics, due to its *modular* structure and *integration* in CEMoD *portal* as a package / subscription (SaaS) it can easily benefit of *up-sell* (until all initially supported 4 use cases as presented below or even to the newer extended ones such as Zero Day Insight mentioned in the Attribute Packs figure) or *cross-sell* (other type of relevant packages part of the CEMoD offering, such as Radio Cell related, etc.); some discounts may appear due to overall sales volume (depending also on the CSP's subscriber number – *value-added growth* paradigm).

Overview of Supported Use Cases

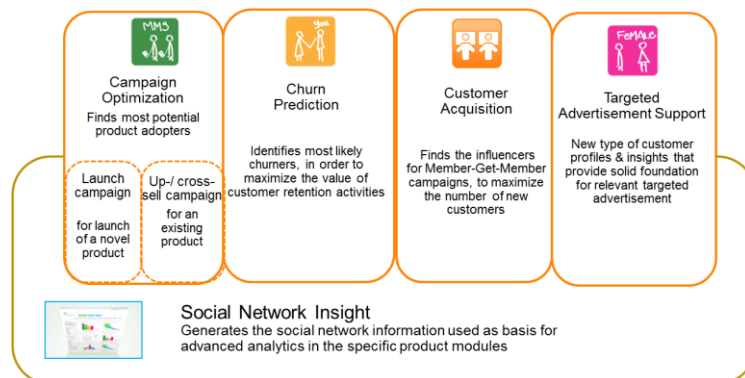


Figure 41 NSN-Comptel Social Analytics modular use cases

Attribute Packs and its Derived Attributes

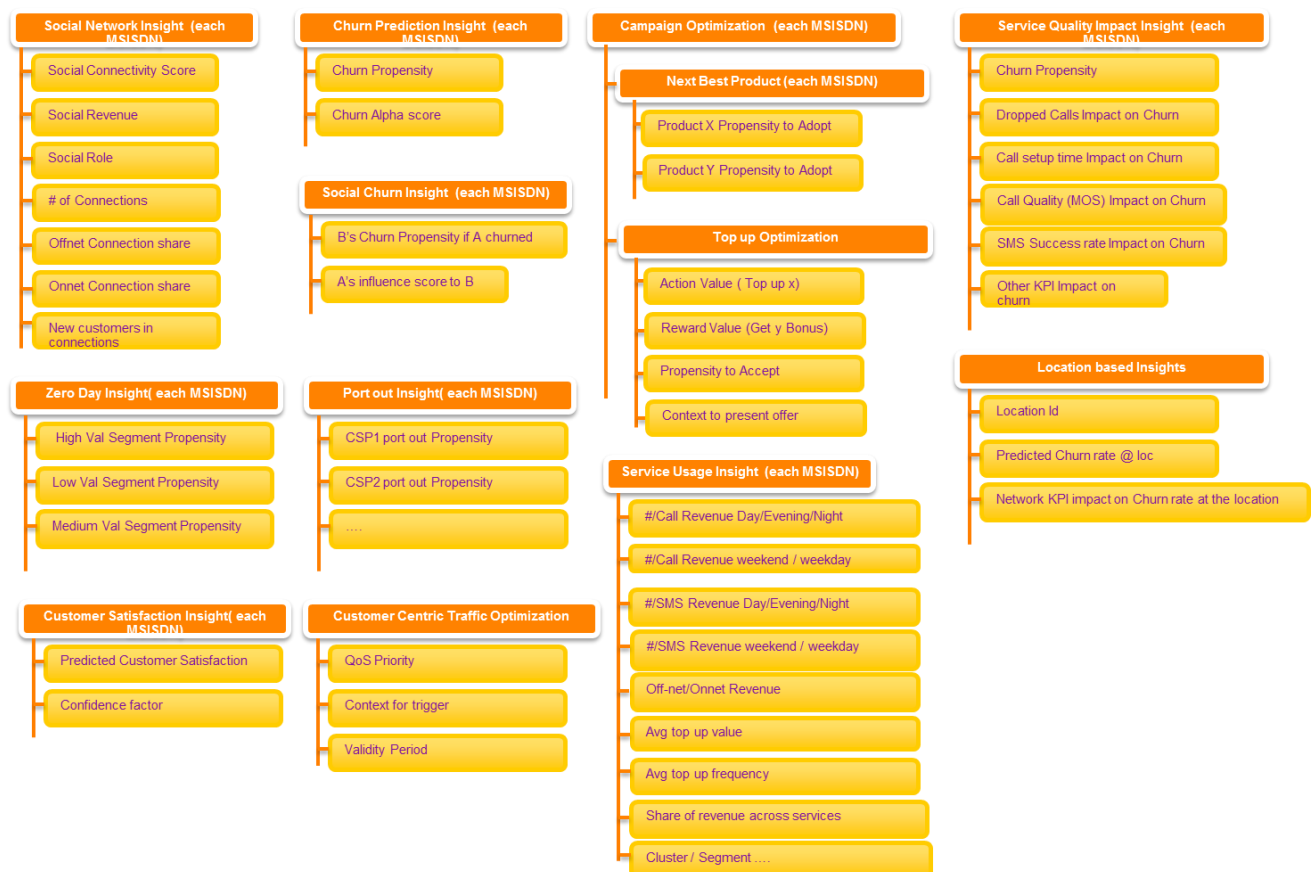


Figure 42 Social Analytics extended use cases

Pricing Scenario description - CSP pays quarterly about 242K EUR to benefit of the Social Analytics solution SaaS deployed on Amazon EC2

This scenario has taken into account the following input variables:

- Social Analytics – Software as a Service, included in Customer Experience Management on Demand (CEMoD) - (related to the CSP's 2 million subscriber base, 1 use case (churn prediction) and 10 EUR ARPU in a 2 year (8 quarters) hosted service contract (solution hosted on public Amazon EC2 Cloud).

Assumption is that CSP paying at the beginning of every quarter; all calculations related to Cloud costs, solution deployment, use case, Cloud change management & maintenance support as per the consultancy cost (SPC) of the involved NSN business lines (Operations (OPR), Consultancy and System Integration (CSI), etc.) are contained in the below embedded excel – CEM_SA_pricing_sheet.xlsx).

Cash Flow calculation for two years

Note: Contract for 2 Years / 8 quarters (The 3rd Year or following ones are just as example).

Contract details		Pricing Summary	
Duration of contract	8 Quarters	Setup fee	10.595
Payment Cycle	1 Times per Quarter	Periodic fee	241.811
Cash Flow			
	Year 1	Year 2	Year 3
Revenue	977.837	967.242	967.242
Costs	666.820	610.320	610.320
Net Cash	311.018	356.923	356.923

	Year 1				Year 2			
	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
Upfront cost	56.500	0	0	0	0	0	0	0
Recurring costs	152.580	152.580	152.580	152.580	152.580	152.580	152.580	152.580
Total cost incurred within the quarter	209.080	152.580	152.580	152.580	152.580	152.580	152.580	152.580
Revenue	252.406	241.811	241.811	241.811	241.811	241.811	241.811	241.811
Net Cash	43.326	89.231	89.231	89.231	89.231	89.231	89.231	89.231
Cumulative cash flow	43.326	132.557	221.787	311.018	400.249	489.479	578.710	667.941



CEM_SA_pricing
sheet.xlsx

(For further details click on the excel icon)

Even if the initial selected opportunity (hosted Amazon EC2 Cloud based solution, 1 subscription package / use-case, the easiest implementation, no significant implementation effort) looks like less promising (about 350K EUR net profit / year), as mentioned before in the growth description part, it will be very easy to reach at least 1 million EUR net / year thru an eventual *up-sell* (for the additional 4 *modules*, keeping the same inputs as subscriber number and ARPU).

A calculation for all 4 use cases cannot be performed without a clear case and due diligence on CSP's premises (cause it may contain relevant technology gaps that may be translated in consistent development and integration effort – at minimum the estimation could be done by multiplying with the additional use case numbers, as Churn Prediction was the cheapest and easier to be implemented → resulting a yearly profit of 1.4 mi. EUR)

Then the net profit will be exponentially increased due to the subscriber numbers and ARPU (here a *shared-gain* contractual binding can be performed ensuring the customer's long-term commitment – case of Managed Services for Hosted solutions in case of huge subscriber numbers).

Sales Pitch Documents (Prepaid & PostPaid CSP's revenue calculation and Social Analytics improvement related to the reference model – in-house analytics based)

Churn Prediction - Reference comparison. Value Based Argumentation Prepaid

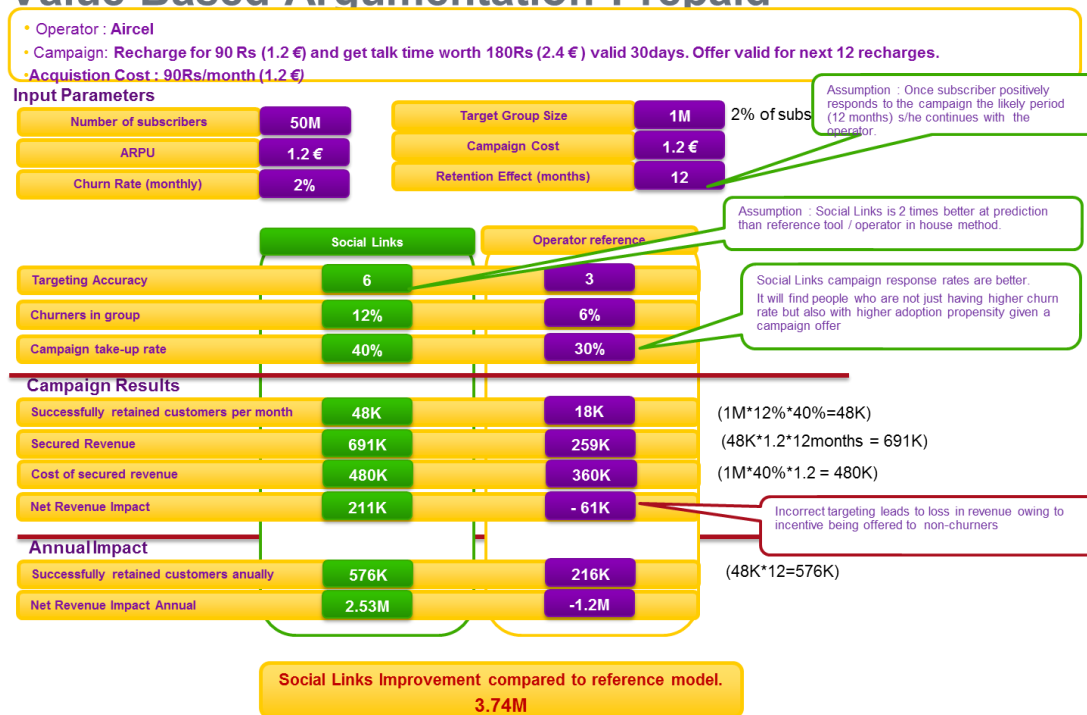


Figure 43 CSP's revenue boost due to Churn Prediction and Campaign uptake



Solution_Analytics_Sa
les_Pitch_PostPaid.xls

(For further details click on the excel icon)

4.3 Future Expansion

Phase 1 & 2 – Geo-Market expansion

Phase 3 & 4 – Portfolio and use case diversification

- Phase 1

Continue with MVNO business types for the operator (Tele2) with the aim to go to the emerging markets (in special Russia – potential of 60 million users due to Tele2 recent acquisition and licenses)

- Phase 2

Supported by the success stories in Asia and Europe (MVNO related) to penetrate in other Telecom markets global-wide (offering *Turn-Key* solutions especially in Asia, South America) (Robert D. Hisrich, 2010, p. 148)

- Phase 3

Extend the domain area – other Telco / IT opportunities

Machine-to-Machine (telematics, vending machines, energy sector related)

Airport telecommunication traffic

IT – banks (ATM, credit card requests, etc.); fixed / landline service providers (unexplored area probably due to privacy regulations in some countries, e.g. USA related)

- Phase 4

Any kind of related B2B activity which request such solution, domain independent

5 Conclusion

5.1 Emphasize the Telecom / Enterprise Customer Service Provider benefits of investing in a social analytics (predictive analytics based) solution

Even if the predictive analytics phenomena and its cutting-edge social analytics iceberg are considered to be in an incipient phase, we had seen their great potential due to various implementations in many sectors of our society and with emphasis on high-tech business.

High-tech business environment is known for its “market uncertainty”, “technological uncertainty” and “competitive volatility” (Jakki Mohr, 2010, p. 11), which foster the adoption of predictive analytics based tools that aim to prepare the enterprises in their journey towards a “competitive advantage” as part of their ongoing “business strategy” (Grant, 2011, p. 19)

Last but not least, *Telecom domain* - the main high-tech selected area of interest for the current paper, benefits of predictive analytics immixture in its business deliverable decisions. Moreover *social analytics* as a branch of predictive analytics emphasize its role in effective and efficient mobile customer behavior predictions, detecting subscribers’ potential *influential circles* and important aspects (service or non-service related) that all converge towards their final decision *to churn or not* to other Mobile Network Operators (MNOs) or to adopt new products or services.

From sales point of view the general attitude changed in the latest years from the “*spray and pray*” technique (predominant at the end of the 90s, where sales responsible personnel were presenting (“spraying”) vendors specific technologies and were hoping (“praying”) to find an interested customer to perfect the contract with) to *customer oriented* techniques, where the main driver lies in the business outcome in a win-win situation and has as input *customer challenges* that have to be solved with the help of the offered solution. Therefore I will now summarize the main mobile Telecom Customer Service Provider’s (CSPs) challenges / pain-points (from B2B / B2B2C perspective) that are to be solved thru a *social analytics* solution:

- Despite having an “ocean of customer data”, CSPs struggle in targeting right customers with right messages at the right time, due to improper “market segmentation” (“process of dividing the market into small homogeneous groups”, which allows “the entrepreneur to more effectively respond to the needs of more homogeneous consumers”) (Robert D. Hisrich, 2010, p. 238); additionally, relevant strategic data is hard to be found (data scattered through many silo IT / Telecom specific platforms, hard to be “digested” with in-house empiric analytics or based on someone’s “guts”);
- Their marketing budget (inclusive retention and loyalty budget) is wasted and campaigns show sub-optimal results (lower adoption rate, non-decrease of the churn rate)
- CSPs struggle to understand how Quality of Service affects its subscriber base which is diverse in terms of its expectations and requirements

Social Analytics solution is coming to tackle this win-win business opportunity offering to CSPs:

- Highly accurate predictions of customers’ near-future behavior regarding loyalty & churn or product responsiveness / affinity in an effort to enable operators for:

- Improving their marketing campaign effectiveness (including cross-sell, up-sell campaigns)
- Drive up overall customer lifetime value
- Boost customer satisfaction and experience (monitoring the Key Performance Indicators (KPI) / Key Quality Indicators (KQI) / Customer Experience Index (CEI) respectively related to *network / services / customer satisfaction / perception*).

With the help of Software Delivery Platform (Service Oriented Architecture), Social Analytics plays an important triggering role for various actionable activities (such as Top-Up automated campaigns, Value Added Services, retention campaigns, etc.). Moreover this incorporation of Social Analytics in an automated Decision Management System facilitates emphasize its strategic analytical role “as decision support” at the enterprise organization level (Grant, 2011, p. 25) as is well-known that decisions based on “guts” are subject to “bounded rationality” (“cognitive limitation that constrain all human beings”) (Grant, 2011, p. 25).

As a corollary, by transposing all Social Analytics knowledge deliverables (lists, reports, charts) into decision making actionable, the path to CSP’s revenue maximization is strongly secured.

5.2 Check-list for selecting the best fit predictive / social analytics tool

- CSP should make its own in-house due diligence for determining the eventual analytical skills know-how (if not clearly part of some employees daily assignments) to be used for supporting the process from operational point of view (to find out if this task should be outsourced or included in the vendor’s managed services contract);
- The predictive / social analytics tool should match the CSP’s business strategy goal with respect to revenue figures, targeted and friendly usage (e.g. GUI portal for CxO, operations, marketing, financial departments, etc.);
- The presented solution should have flexible pricing structure and deployment capabilities (e.g. public, hybrid deployments);
- Should ensure software (SW) modularity (use-case selection, scalability, independent and fast development), its platform architecture independence and hardware (HW) scalability (eventual Virtual Machine / Cloud applicability);

- The tool should contain a comprehensive spectrum of data mining algorithms or meta learning capability (automatic learning mechanism, possibility to select, modify or combine different learning algorithms to effectively fit the input data);
- Important to contain features related to open-community library integration and benefit of the large open-source development communities (e.g. R environment, Weka)
- Fast input / output processing capability (e.g. non-shared parallel processing architecture); import / export of PMML based predictive models
- Identity management / user access policy management to the presented solution
- Fully encrypted / data secrecy protection during all prediction modeling related phases and its final storage (e.g. scoring if requested – to be further used in Decision Management Systems)
- To have a proper predictive modeling life-cycle monitoring and reporting including a fine-tuning mechanism (e.g. part of a “learning mechanism” to fine-tune the predictive models if needed after targeted promotion observation in a specific time)
- Eventual to present some success stories or value proposition for their use cases
- To offer a Proof-Of-Concept (PoC) trial and commit to extensive documentation sharing or present training / supporting service offers – long-term commitment
- To present the product roadmap (e.g. to tackle semantic, web / text mining, etc.) and its governance strategy

5.3 Future developments

- Solutions based on predictive analytics and their implementation in complex *decision systems* are proved to be a hot topic also in various academic media – e.g. Heinz Nixdorf Institute finances the Project CRC 901 (HNI Jahresbericht 2011) - containing 3 phases:

Project Area A “Algorithmic and Economic Foundations for Organizing Large Dynamic Markets”.

Project Area B “Modeling, Composition and Quality Analysis for On-the-Fly Computing”.

Project Area C “Reliable Execution Environments and Application Scenarios for On-the-Fly Computing”

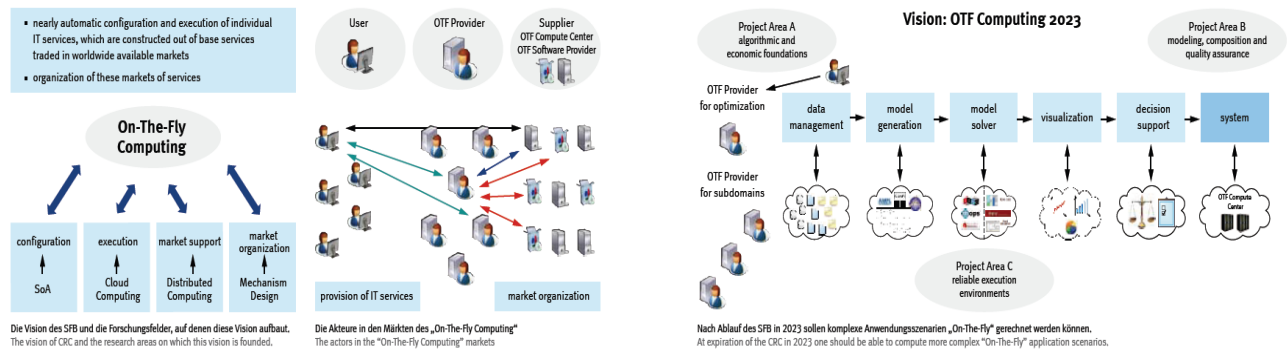


Figure 44 On-The-Fly Computing - Project development phases, Source HNI 2011

Integration of a predictive analytics solution in a complex *Decision Management System* is the ultimate way towards the already mentioned prescriptive analytics, where the focus is on the fully automated business decisions. “*Decision Management Systems adapt to constantly changing circumstances, continually trying new approaches to see what will work best as situation change*” (Taylor, *Decision Management Systems - A Practical Guide to Using Business Rules and Predictive Analytics*, 2013, p. 20) ¹⁰. There are two major take-away specified in the above mentioned book:

- with respect to **predictive analytics model** techniques to be **up-to-date**: ***self-learning*** capability (automatic build of predictive models upon new data arrival – usually the initial deployed model is done with the help of an analyst); ***model refresh*** (manual or automatically done when new data is available – in order to improve the prediction accuracy or “corrects for declining accuracy” (Taylor, *Decision Management Systems - A Practical Guide to Using Business Rules and Predictive Analytics*, 2013, p. 180), with great results when “the predictive model is an ensemble of multiple predictive analytic models”, final output resulted “by using voting techniques across all models (Taylor, *Decision Management Systems - A Practical Guide to Using Business Rules and Predictive Analytics*, 2013, p. 180)”; decision performance due to new predictive model change, ***champion-challenger*** approach (“the challenger alternative” on iterative mode tested on a “randomly selected small percentage of the production population”; results are compared to the ones of the existing “champion” tested on the bulk population; challenger will be promoted as new “champion” if better results are achieved, then the process continues in the same manner to test other challengers till an “optimal point” is reached – some optimization techniques can be used to “design better challengers speeds progress toward optimal” (Taylor, *Decision Management Solutions*, p. 10));

- three steps in building an ongoing process towards **building a Decision Management System (DMS)**: “***discover and model***” the decisions that will be part of the DMS (“repeatable operational and tactical decisions”; “***design and implement these decisions***” as “service oriented components or Decision Services” (Service Oriented Architecture), “each Decision Service being the “right combination of business rules, predictive analytic models, and optimization technology to deliver accurate, repeatable decisions”; “***create the processes and infrastructure to continually monitor and improve the way your decisions are made***” (Taylor, Decision Management Systems - A Practical Guide to Using Business Rules and Predictive Analytics, 2013, p. 264)
- What would be of a great interest is one of the Telecom’s untouched fields for collecting the relevant subscriber data such as coming from *wire-line / fixed operators* (may be subject to specific regulations). One possible idea comprised in some local projects is to provide the end-users with specific tools that allow them to act as Internet Service Providers and even to receive some revenues (generally based on shared internet connectivity provided by the fixed Telecom operators; is known that few users need the maximum of their bandwidth, so that they may share portions of unused quota to other potential customers; here Telecom operators are providing the proper tools and act as mediators in helping their primarily high-profile customers that benefit of quality of services (QoS) and reserved bandwidth, to collect the revenue from the other “sharing buddies” based on their location, peak rate, time of the day, etc.). Predictive analytics based solutions may play an important role in studying the user behavior, predicting their interest area, usage, preparing the customer profiling selection, campaign targeting towards low-end (sharing buddies) / high-end customers (QoS), etc;
- *Photo* and *video* decomposition using data mining techniques (e.g. from Computer Tomography and Magnetic Resonance Imaging (MRI)) could be used as predictive analytics opportunities in medical field (by performing a fast scan thru the existing structured and unstructured data in case of various cancer / tumors types with the aim to predict the subject’s health in this respect and help in adoption of the right medical decisions to diminish or even to avoid in a timely manner the spread of the disease); another distant market could be related to security based services by studying the surveillance data and predicting the most likely locations, date and time window/s for the unwished criminal activities.

References

1. Kenneth C. Laudon, Jane P. Laudon - Essentials of Management Information Systems, Pearson 2013
2. Henschen, Doug (4 January 2010). Analytics at Work: Q&A with Tom Davenport. (Interview).
3. Bill Inmon – Data Warehousing 2.0, April 2010 (White Paper)
4. Paulraj Ponniah - Data Warehousing Fundamentals for IT Professionals. Wiley, 2010
5. WAYNE W. ECKERSON - The Data Warehousing Institute (TDWI) Best Practices Report 2007 - PREDICTIVE ANALYTICS Extending the Value of Your Data Warehousing Investment
6. Louis Columbus - Roundup of Big Data Forecasts and Market Estimates, 2012 (<http://www.forbes.com/sites/louiscolumbus/2012/08/16/roundup-of-big-data-forecasts-and-market-estimates-2012/>)
7. Predictive Analytics World, Survey Results - Predictive Analytics Business Applications, January 2009
8. Barack Obama - <https://my.barackobama.com/page/signup/predictive-analytics-intern-application>
9. <http://www.forbes.com/2010/04/01/analytics-best-buy-technology-data-companies-10-accenture.html>
10. James Taylor – Decision Management Systems, A Practical Guide to Using Business Rules and Predictive Analytics, ISBN-13: 978-0-13-288438-9

6 Bibliography

- Microstrategy - TDWI Report.* (2006). Retrieved 12 05, 2012, from <http://www.microstrategy.com/Strategy/media/downloads/products/MicroStrategy-TDWI-Best-Practices-Report-Predictive-Analytics-Data-Warehousing.pdf>
- Predictive Analytics World.* (2009, Feb). Retrieved 12 05, 2012, from <http://www.predictiveanalyticsworld.com/Predictive-Analytics-World-Survey-Report-Feb-2009.pdf>
- Practical Analytics Wordpress.* (2013, 02 10). Retrieved from <http://practicalanalytics.wordpress.com/predictive-analytics-101/>
- 4G Americas.* (n.d.). Retrieved 05 10, 2013, from <http://www.4gamericas.org/index.cfm?fuseaction=page§ionid=242>

- Analytic Bridge*. (n.d.). Retrieved 02 20, 2013, from <http://www.analyticbridge.com/profiles/blogs/predictive-descriptive-prescriptive-analytics>
- BioMap Systems*. (n.d.). Retrieved 02 10, 2013, from <http://biomapsystems.com/what-is-biomap/predictive-analysis>
- Columbus, L. (2012, 10 15). *Forbes - Gartners 232B Big Data Forecast*. Retrieved 02 12, 2013, from Gartner - <http://www.gartner.com/id=2195915>:
<http://www.forbes.com/sites/louiscolumbus/2012/10/15/using-search-analytics-to-see-into-gartners-232b-big-data-forecast/>
- Columbus, L. (2012, 08 16). *Forbes - Roundup of big data forecasts and market estimates*. Retrieved 02 12, 2013, from <http://www.forbes.com/sites/louiscolumbus/2012/08/16/roundup-of-big-data-forecasts-and-market-estimates-2012/>
- Comptel - Social Links*. (n.d.). Retrieved 03 03, 2013, from <http://www.comptel.com/>
- CRAN Task View*. (n.d.). Retrieved 03 03, 2013, from <http://cran.r-project.org/web/views/MachineLearning.html>
- Data mining articles - Data mining*. (n.d.). Retrieved 03 12, 2013, from <http://www.dataminingarticles.com/info/data-mining-introduction/>
- Davenport, T. (n.d.). *Wikipedia - Business Intelligence*. Retrieved 02 10, 2013, from http://en.wikipedia.org/wiki/Business_intelligence
- Decisionstats*. (n.d.). Retrieved 02 10, 2013, from <http://decisionstats.com/tag/numerati/>
- Dominik Morent, e. a. (n.d.). *KNIME - Comprehensive PMML Preprocessing in KNIME*. Retrieved 06 15, 2013, from http://tech.knime.org/files/knime_pmml_kdd2011.pdf
- Drucker, P. F. (2011). *Innovation and Entrepreneurship*. New York: Routledge.
- Eclipse*. (n.d.). Retrieved 03 05, 2013, from <http://www.eclipse.org/org/>
- Futron*. (n.d.). Retrieved 2 10, 2013, from http://www.futron.com/upload/wysiwyg/Resources/Whitepapers/APF_Enhanced_Analytics_0808.pdf
- Grant, R. M. (2011). *Contemporary Strategy Analysis*. West Sussex: John Wiley & Sons, Ltd.
- Gualtieri, M. (n.d.). *Forrester - Blogs / Pragmatic Definition Of Big Data*. Retrieved 12 05, 2012, from http://blogs.forrester.com/mike_gualtieri/12-12-05-the_pragmatic_definition_of_big_data
- Guazzelli, A. (2012, 01 11). Retrieved 03 10, 2013, from <http://www.predictive-analytics.info/2012/01/pmml-41-is-here-mature-standard-for.html#!/2012/01/pmml-41-is-here-mature-standard-for.html>
- Guazzelli, A. (2012, Jun 19). *IBM - Predicting the future, Part 2: Predictive modeling techniques*. Retrieved 03 15, 2013, from <http://www.ibm.com/developerworks/library/ba-predictive-analytics2/>

- (2012). In A. A. Guillaume de la Roche, *LTE – Advanced and Next Generation Wireless Networks / Channel Modeling and Propagation* (pp. 4, 10). Chennai, India: Wiley.
- Harris, D. R. (2010, 01 04). *Forbes*, *Analytics best buy technology data companies*. Retrieved 02 15, 2013, from <http://www.forbes.com/2010/04/01/analytics-best-buy-technology-data-companies-10-accenture.html>
- Hisrich, P. D. (2012, Oct 14). *Identifying Opportunities and the Opportunity Analysis Plan*. Wien, Wien.
- IBM BigData Hub*. (n.d.). Retrieved 2 10, 2013, from <http://www.ibmbigdatahub.com/blog/using-predictive-analytics-improve-decision-making-and-business-outcomes-part-2>
- Informa, Community - Analytics*. (n.d.). Retrieved 02 17, 2013, from <https://www.informs.org/Community/Analytics>
- Inmon, B. H. (2010, April). *Data Warehousing 2.0 - Modeling and Metadata Strategies for Next Generation* (White Paper).
- Jakki Mohr, S. S. (2010). *Marketing of High-Technology Products and Innovations*. New Jersey: Pearson.
- KNIME data mining platform*. (n.d.). Retrieved 03 05, 2013, from <http://www.knime.org/>
- KXEN - InfiniteInsight*. (n.d.). Retrieved 03 01, 2013, from <http://www.kxen.com/>
- Mark A. Beyer, J.-D. L. (2012, Oct 12). *Gartner*. Retrieved 02 12, 2013, from <http://www.gartner.com/id=2195915>
- Mike Gualtieri, J. 3. (2013). *The Forrester Wave™: Big Data Predictive Analytics Solutions, Q1 2013*. Acorn Park Drive, Cambridge, MA 02140 USA: Forrester Research, Inc.
- Miori, R. K. (2010, Oct). *Informa - Back in Business, OR/MS Today*. Retrieved 02 17, 2013, from <https://www.informs.org/ORMS-Today/Public-Articles/October-Volume-37-Number-5/Back-in-Business>
- Orange - data analytics*. (n.d.). Retrieved 03 03, 2013, from <http://orange.biolab.si/>
- Orange Biolab*. (n.d.). Retrieved 03 03, 2013, from <http://orange.biolab.si/screenshots/>
- Pipelinepub - 2013 Competitive Communications Landscape*. (n.d.). Retrieved 06 10, 2013, from http://www.pipelinepub.com/2013_Competitive_Communications_Landscape/455/3
- Practical Analytics Wordpress*. (n.d.). Retrieved 02 10, 2013, from <https://practicalanalytics.files.wordpress.com/2012/03/makingmoney.jpg>
- Rapid-I - RapidMiner data mining and predictive analytics*. (n.d.). Retrieved 03 05, 2013, from <http://rapid-i.com/content/view/181/190/>
- Rapid-I - RapidMiner Operator Reference*. (n.d.). Retrieved 06 18, 2013, from http://docs.rapid-i.com/files/rapidminer/RapidMiner_OperatorReference_en.pdf

- Revolution Analytics - RevoScaleR*. (n.d.). Retrieved 03 03, 2013, from <http://www.revolutionanalytics.com/why-revolution-r/whitepapers/RevoScaleR-Speed-Scalability.pdf>
- Robert D. Hisrich, M. P. (2010). Entrepreneurship. In *Entrepreneurship* (p. 148). Singapore: McGraw - Hill International Edition.
- Sallam, R. (2013, 02 11). *Gartner, Press Release - STAMFORD, Conn.* Retrieved 02 15, 2013, from <http://www.gartner.com/newsroom/id/2332515>
- SAP - HANA*. (n.d.). Retrieved 03 01, 2013, from http://help.sap.com/hana/hana_dev_r_emb_en.pdf
- SAP Predictive Analytics - User Guide*. (n.d.). Retrieved 03 01, 2013, from http://help.sap.com/businessobject/product_guides/SAPpa10/en/pa1_0_7_user_en.pdf
- Siegel, E. (n.d.). *Predictive Analytics World*. Retrieved 02 10, 2013, from <http://www.predictiveanalyticsworld.com/>
- (2013). Decision Management Systems - A Practical Guide to Using Business Rules and Predictive Analytics. In J. Taylor, *Decision Management Systems - A Practical Guide to Using Business Rules and Predictive Analytics* (p. 20). US, Indiana: Pearson plc.
- Taylor, J. (n.d.). *Decision Management Solutions*. Retrieved 07 07, 2013, from http://decisionmanagementsolutions.com/attachments/080_PuttingPredictiveAnalyticsToWork.pdf
- TDWI Business Analytics*. (n.d.). Retrieved 02 10, 2013, from <http://tdwi.org/portals/business-analytics.aspx>
- Weka Data mining*. (n.d.). Retrieved 03 05, 2013, from <http://www.cs.waikato.ac.nz/ml/weka/index.html>
- Wikipedia - Data Warehouse*. (n.d.). Retrieved 2 10, 2013, from http://en.wikipedia.org/wiki/Data_warehouse
- Wikipedia - Decision tree learning*. (n.d.). Retrieved 03 12, 2013, from http://en.wikipedia.org/wiki/Decision_tree_learning
- Wikipedia - Document Type Definition (DTD)*. (n.d.). Retrieved 03 07, 2013, from https://en.wikipedia.org/wiki/Document_Type_Definition
- Wikipedia - High Tech*. (n.d.). Retrieved 02 10, 2013, from http://en.wikipedia.org/wiki/High_tech
- Wikipedia - Machine Learning*. (n.d.). Retrieved 02 18, 2013, from http://en.wikipedia.org/wiki/Machine_learning
- Wikipedia - Predictive Model Markup Language*. (n.d.). Retrieved 03 07, 2013, from http://en.wikipedia.org/wiki/Predictive_Model_Markup_Language
- Wikipedia - Support vector machine*. (n.d.). Retrieved 03 15, 2013, from http://en.wikipedia.org/wiki/Support_vector_machine
- Wright, D. (n.d.). *IBM - Software success stories*. Retrieved 02 10, 2013, from http://www-01.ibm.com/software/success/cssdb.nsf/CS/STRD-8NVMCD?OpenDocument&Site=default&cty=en_us

Appendix_A (To return please use Alt+Left Arrow Key)

The close “fight” is between SAS, IBM as Strong Leaders and SAP a newcomer. Important are Forester’s weighting with respect to current offering and strategy and for each of them the sub-levels (e.g. architecture, supported data formats, discovery, tools / licensing, commitment and roadmap)

Figure 3 Forrester Wave™: Big Data Predictive Analytics Solutions, Q1 '13 (Cont.)

	Forrester's Weighting	Angoss Software	IBM	KXEN	Revolution Analytics	Salford Systems	SAP	SAS	StatSoft	Tibco Software
CURRENT OFFERING	50%	2.74	4.31	2.68	2.68	1.81	3.57	4.59	2.94	3.21
Architecture	20%	2.82	5.00	2.50	3.07	1.10	4.96	5.00	1.30	2.10
Data	15%	3.00	4.00	4.00	2.50	2.50	4.50	4.50	4.00	3.00
Discovery	25%	1.70	3.85	2.90	2.85	1.00	3.05	4.40	2.85	3.10
Evaluation and optimization	5%	2.00	5.00	2.20	2.00	1.60	1.20	4.20	3.80	4.60
Deployment	5%	4.60	5.00	4.60	3.40	3.40	2.60	5.00	4.60	3.80
Tools	25%	3.60	4.00	1.60	2.50	2.80	3.40	4.40	3.40	3.80
Standards, integration, solutions, and extensibility	5%	1.50	5.00	2.25	1.75	0.25	2.00	5.00	2.00	4.00
STRATEGY	50%	2.48	3.58	3.08	1.98	1.98	3.58	4.18	3.28	3.58
Licensing and pricing	25%	2.50	2.50	2.50	2.50	2.50	2.50	2.50	2.50	2.50
Commitment	50%	2.20	3.40	3.40	2.20	2.20	3.40	4.60	3.80	3.40
Product road map	25%	3.00	5.00	3.00	1.00	1.00	5.00	5.00	3.00	5.00
MARKET PRESENCE	0%	2.09	4.21	2.94	2.20	2.15	1.93	4.46	2.66	2.09
Company financials	30%	1.00	3.50	2.00	3.00	1.00	1.00	4.00	2.00	1.00
Global presence and installed base	60%	2.68	4.60	3.24	1.80	2.68	2.28	4.60	3.24	2.28
Partnerships	10%	1.80	4.00	4.00	2.20	2.40	2.60	5.00	1.20	4.20

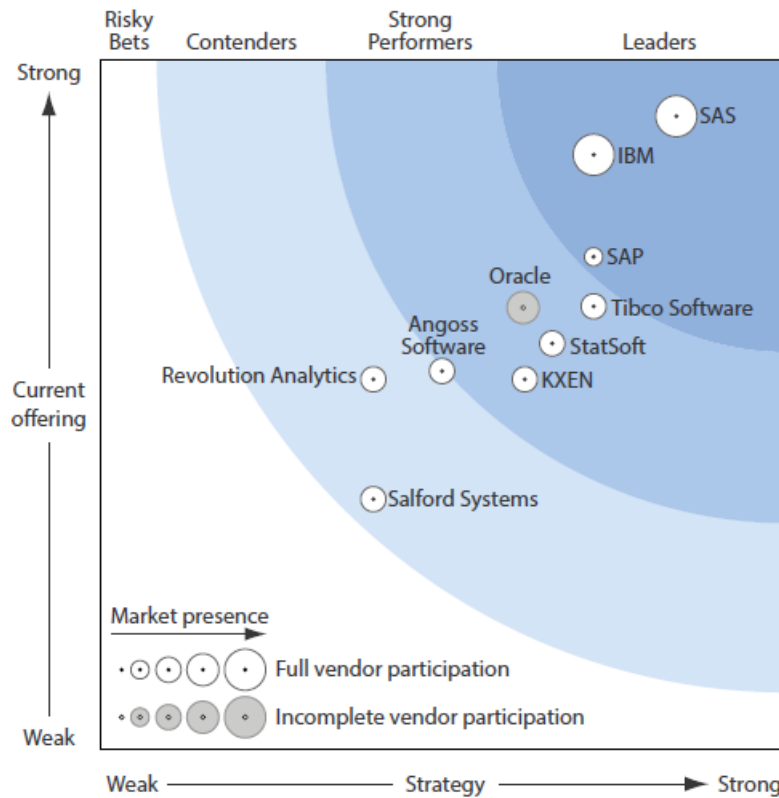
All scores are based on a scale of 0 (weak) to 5 (strong).

Source: Forrester Research, Inc.

Figure 45 The Forrester Wave™: Big Data Predictive Analytics Solutions, Q1 2013

Appendix_B (To return please use Alt+Left Arrow Key)

Figure 3 Forrester Wave™: Big Data Predictive Analytics Solutions, Q1 '13



Source: Forrester Research, Inc.

Figure 46 Open source vs. commercial solutions ranking (market presence and strategy) - The Forrester Wave™: Big Data Predictive Analytics Solutions, Q1 2013

Appendix_C (To return please use Alt+Left Arrow Key)

To enable the use of “R” open source programming language and software environment, SAP HANA database context has an embedded R code which is processed in-line (belong to the query execution plan) whenever an application is making use of R environments for statistical functions as you can see from the below picture of the integrated solution. Solution components: the SAP HANA based application, the SAP HANA database, and the R environment. It is worth mentioning that R / Rserve environments have to be installed on a separate host and are not shipped with SAP HANA database).(SAP - HANA)

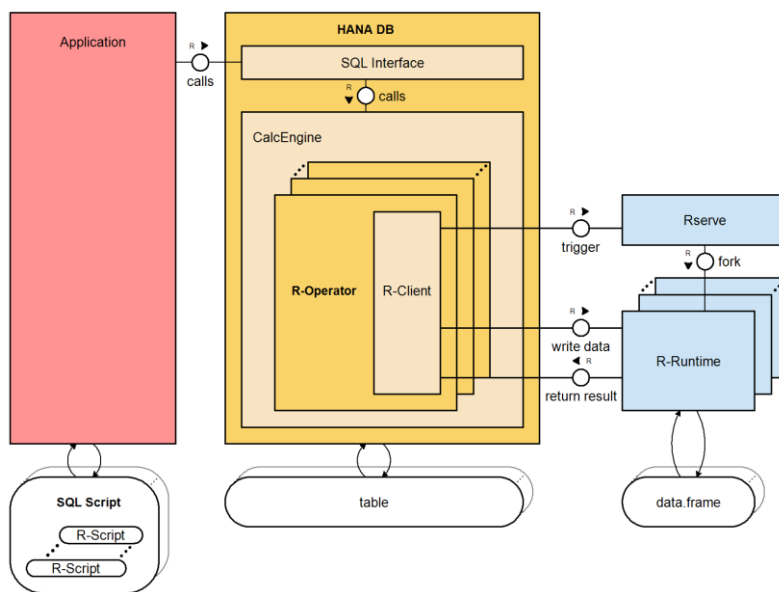


Figure 47 SAP HANA - R integration

Appendix_D (To return please use Alt+Left Arrow Key)

Orange open-source data mining process screenshot (test different learner types that fit the best the input data) (Orange Biolab)

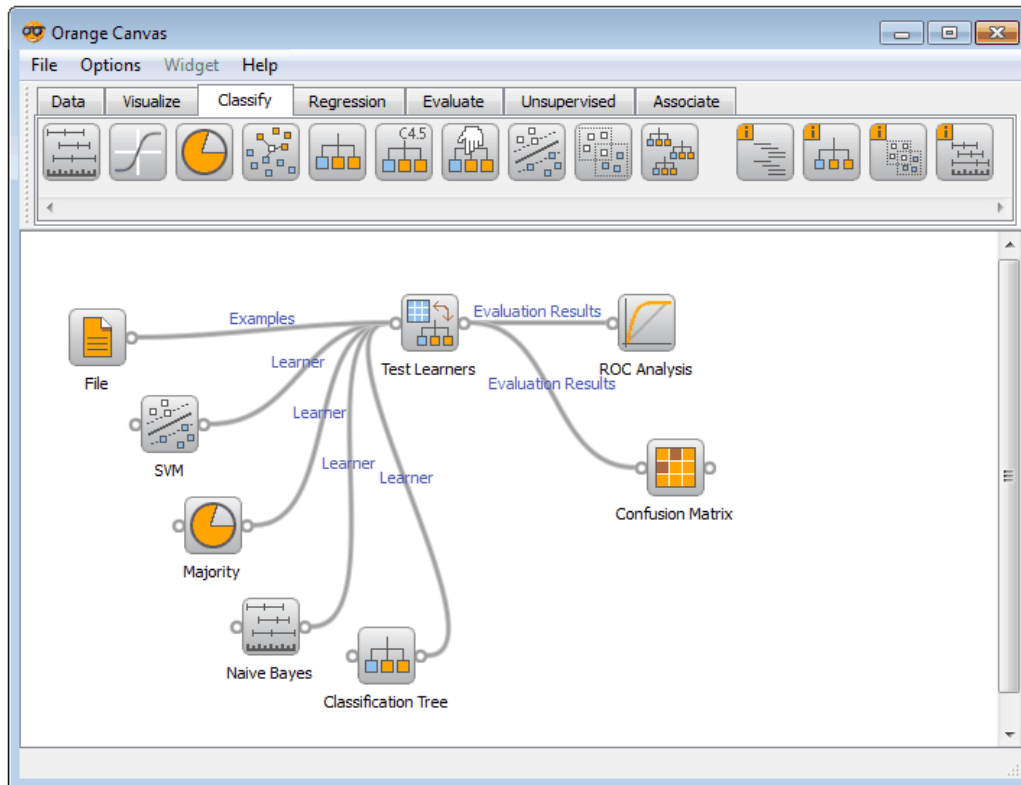


Figure 48 Orange data mining algorithms

Appendix_E (To return please use Alt+Left Arrow Key)

KNIME with pre-integrated Weka algorithms

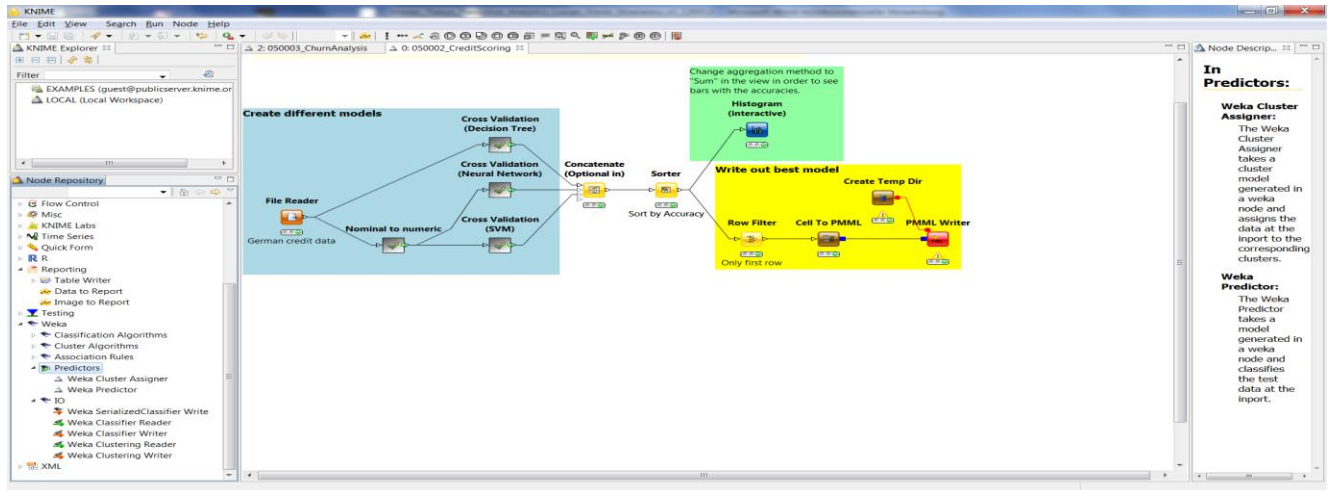


Figure 49 KNIME workflow example for credit scoring

Appendix_F (To return please use Alt+Left Arrow Key)

RapidMiner open source data mining and predictive analytics tool with a great graphical user interface and block elements connectivity correction capabilities. The predictive modeling workflow is simplified due to its drag and drop design options.

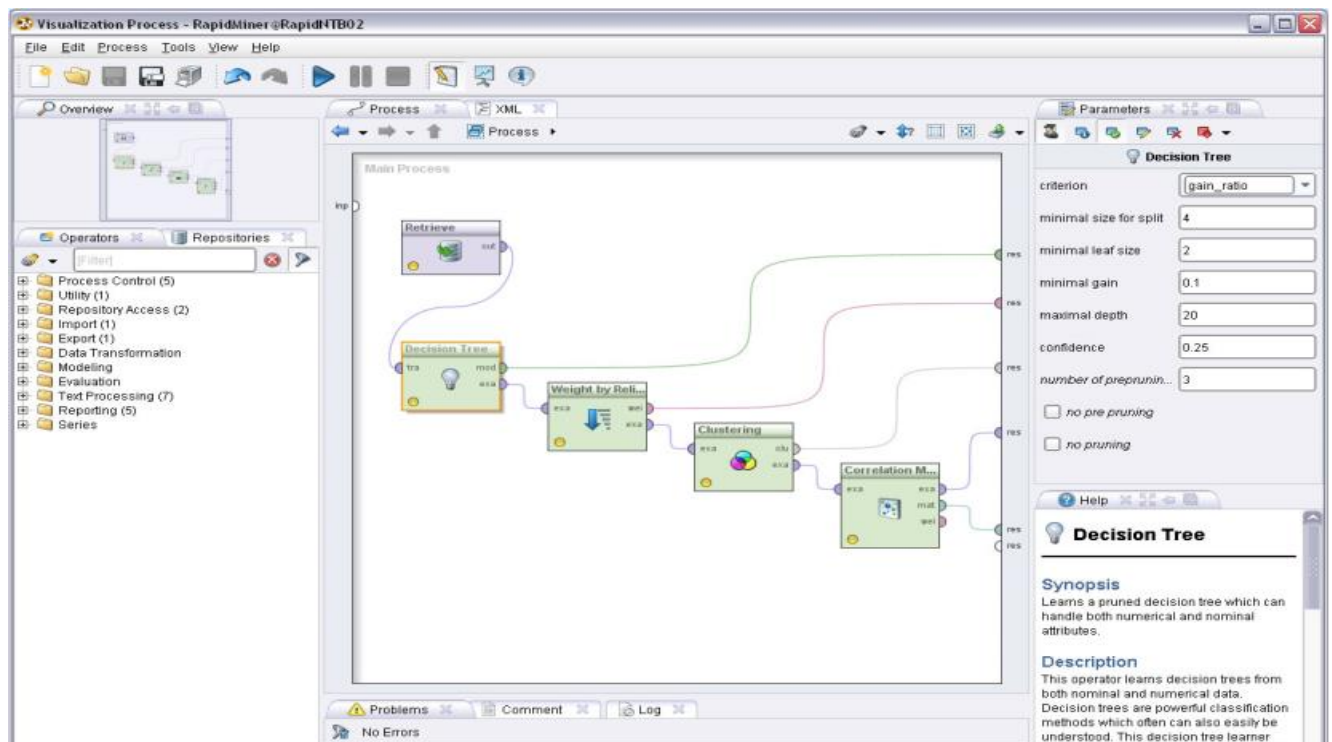


Figure 50 RapidMiner data mining process design example

Appendix_G (To return please use Alt+Left Arrow Key)

With respect to CRISP-DM (Cross Industry Standard for Data Mining) processes, PMML permits a clear task split especially related to *model development* and *model deployment* and eliminates the need of custom code or proprietary model deployment.

Initial phases are very important to correctly setup the data mining process (starting as it can be seen from business and data understanding) and continuing with the most consuming one (from quality and time perspective) – data preparation.


 Phases and Tasks					
Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives <i>Background</i> <i>Business Objectives</i> <i>Business Success Criteria</i>	Collect Initial Data <i>Initial Data Collection Report</i>	<i>Data Set</i> <i>Data Set Description</i>	Select Modeling Technique <i>Modeling Technique</i> <i>Modeling Assumptions</i>	Evaluate Results <i>Assessment of Data Mining Results w.r.t. Business Success Criteria</i> <i>Approved Models</i>	Plan Deployment <i>Deployment Plan</i>
Situation Assessment <i>Inventory of Resources</i> <i>Requirements, Assumptions, and Constraints</i> <i>Risks and Contingencies</i> <i>Terminology</i> <i>Costs and Benefits</i>	Describe Data <i>Data Description Report</i>	Select Data <i>Rationale for Inclusion / Exclusion</i>	Generate Test Design <i>Test Design</i>	Review Process <i>Review of Process</i>	Plan Monitoring and Maintenance <i>Monitoring and Maintenance Plan</i>
Determine Data Mining Goal <i>Data Mining Goals</i> <i>Data Mining Success Criteria</i>	Explore Data <i>Data Exploration Report</i>	Clean Data <i>Data Cleaning Report</i>	Build Model <i>Parameter Settings</i> <i>Models</i> <i>Model Description</i>	Determine Next Steps <i>List of Possible Actions</i> <i>Decision</i>	Produce Final Report <i>Final Report</i> <i>Final Presentation</i>
Produce Project Plan <i>Project Plan</i> <i>Initial Assessment of Tools and Techniques</i>	Verify Data Quality <i>Data Quality Report</i>	Construct Data <i>Derived Attributes</i> <i>Generated Records</i>	Assess Model <i>Model Assessment</i> <i>Revised Parameter Settings</i>		Review Project Experience <i>Documentation</i>
		Integrate Data <i>Merged Data</i>			
		Format Data <i>Reformatted Data</i>			

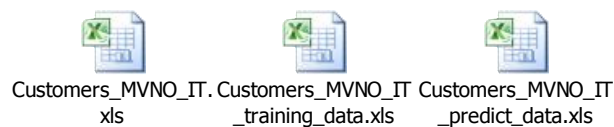
Figure 51 The CRISP-DM User Guide from NCR Systems Engineering Copenhagen

Appendix_H (To return please use Alt+Left Arrow Key)

MVNO Data input for performing the predictive analytics empirical test (using open-source RapidMiner):

As input were used 4000 Customer Relationship Manager (CRM) *subscriber data* entries (Customer Data Records / CDRs) saved as MS Excel format (data collected from an Italian Mobile Virtual Network Operator (MVNO) where the MSISDN (the real mobile subscriber number) was mapped to an ID due to data privacy regulations. A number of 17 variables were selected as considered important for predictive model determination.

From a total of 4000 entries – 2016 were used for “*data training*” against the best fit algorithm and 1484 data records were used to *validate* the predictive model found in the training phase.



Appendix_I (To return please use Alt+Left Arrow Key)

NSN-Comptel Social Analytics (Churn Statistics views and Dashboard view)

Social Analytics – Churn Statistics views

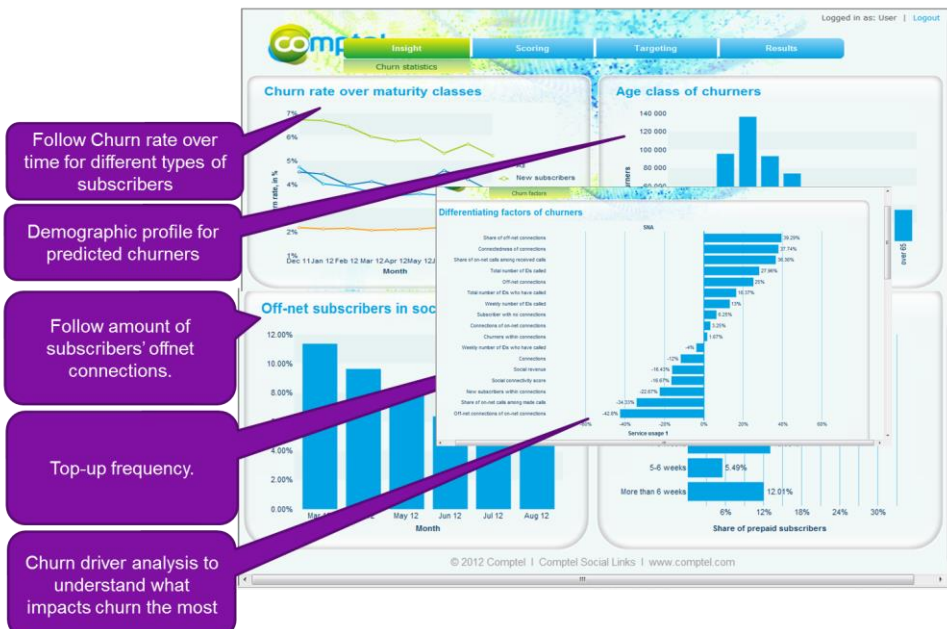


Figure 52 Comptel Social Analytics Churn Statistics and Dashboard view

Social Analytics – Dashboard view

