DATA MINING TECHNIQUES FOR CUSTOMER CHURN PREDICTION

Author:

Patrick Schneider

Course:

MIRI – Algorithms for Data Mining

Date:

27.04.2017

University:

UPC FIB Barcelona, Spain

ABSTRACT

Customer churn prediction has developed to a core research for different kind of companies in recent years. With the past trend of big data, huge amount of data is gathered and generated. Combined with the steady development in data mining techniques, customer churn has emerged as one of the most popular business use cases for marketing prediction.

Studies showed that it is more expensive to gain a new customer than to retain an existing one. To retain existing customers, businesses need to know the reasons of churn, which can be realized through the knowledge extracted from data. This essay gives an overview of the business and marketing, as well as the data science perspective and introduces the proof of concept of a stream mining approach.

"It takes months to find a customer and only seconds to lose one." - Unknown

TABLE OF CONTENT

1. BUSINESS UNDERSTANDING - CUSTOMER CHURN

Businesses in the consumer market and in all enterprise sectors must deal with customer churn. The idea of tackling this issue is to identify customers that are likely to cancel a service or product subscription.

Key reasons for customers to Churn:

- *There are competitors / other companies offering related products and services*
- *Have a better pricing model and options*
- *Bankruptcy by companies*
- *Social media influence and sentiment - word of mouth in social circles*
- *Better customer connects and touch points to address the concerns*

One Example

A classic example can be found in the telecommunication industry where subscribers are known to frequently switch from one provider to another.

Handset or device choice is a driver of churn in the mobile phone business. A popular policy is to subsidize the price of a handset for new subscribers and charging a full price to existing customers for an upgrade. This policy has led to customers jumping from one provider to another to get a new discount. This lead providers to refine their strategies.

High volatility in handset offers is a factor that invalidates models of churn that are based on current handset models. Furthermore, mobile phones are not only telecommunication devices - They are also fashion statements. These social aspects of a buying decision are hard to find in data sets.

The net result for modeling is that you cannot devise a sound policy simply by eliminating known reasons for churn. A continuous modeling strategy contains classical models for the quantification of categorical variables, like for example decision trees.

Examples in other fields that have a use for churn optimization are:

- Internet service provider
- Pay TV
- Banks
- Insurance
- Software as a service(SaaS)

The analysis of customer churn rates is considered in many of those named fields as the key business metric. The mayor reason for this lays in the far less costs of retaining a customer compared to acquiring a new one. Depending on the industry, the cost for a new customer is anywhere between five to 25 times more expensive than retaining an existing one. [1]

However, individualized customer retention is difficult for companies with a high number of customers, where the arising cost would outweigh the generated revenue. The cost for those prevention strategies can be reduced by predicting and targeting the right churning customer.

To archive those opportunities, following churn preventions can be done:

- *Personalized offers based on customer behavior patterns, social network influence, customer usage patterns.*
- *Location-based data for location based advertising.*
- *Optimal Campaigns for up-sell, cross-sell, acquiring new customers, influential or viral marketing.*
- *Combining real-time network feeds, back office data/logs, network inventory, capacity planning, and monitoring service quality, along with subscriber information can provide opportunities to increase wallet share per customer and satisfaction.*

Now, thanks to prediction services and API's it is no longer withheld of company's smaller sizes, that where in the past not able to afford a data science team.

By predicting churning customers with a considerable risk of leaving, insights can be won, as well as a reduction in cost of the prevention strategies.

Benefits of churn prediction:

- *Reduce marketing costs - maximize profits*
- *Reduce churn thru predictive models*
- *Segment market into alike clusters - identify the customers that will generate most profits*
- *Understand customers & their behaviors*
- *Adversely impacts the profitability of organization*
- *Reduce the loss of referrals via the existing customers, if they churn out*
- *For making highly targeted and cost effective marketing strategies*

Using big data sets on their customers, organizations are performing churn detection analytics as an effective approach to the problem.

2. ROLE OF BIG DATA – CHURN PREDICTION

Nowadays, customer churn prediction is one of the most popular fields in business use cases for big data. Over the past years, providers have accumulated significant knowledge about churn drivers, which are the factors that drive customers to switch. The main idea is to identify the potential leaving customers before they decide to churn.

Different data sets can be used for conducting advanced analytical techniques:

- *Billing systems*
- *Customer Care - customer profiling*
- *Call detail records (CDR's) - usage pattern, behavior, geo profile*
- *Customer social media influence and network - sentiments, blogs / posts*
- *Product/Service portfolio*
- *Network Service, Costs*
- *Customer*

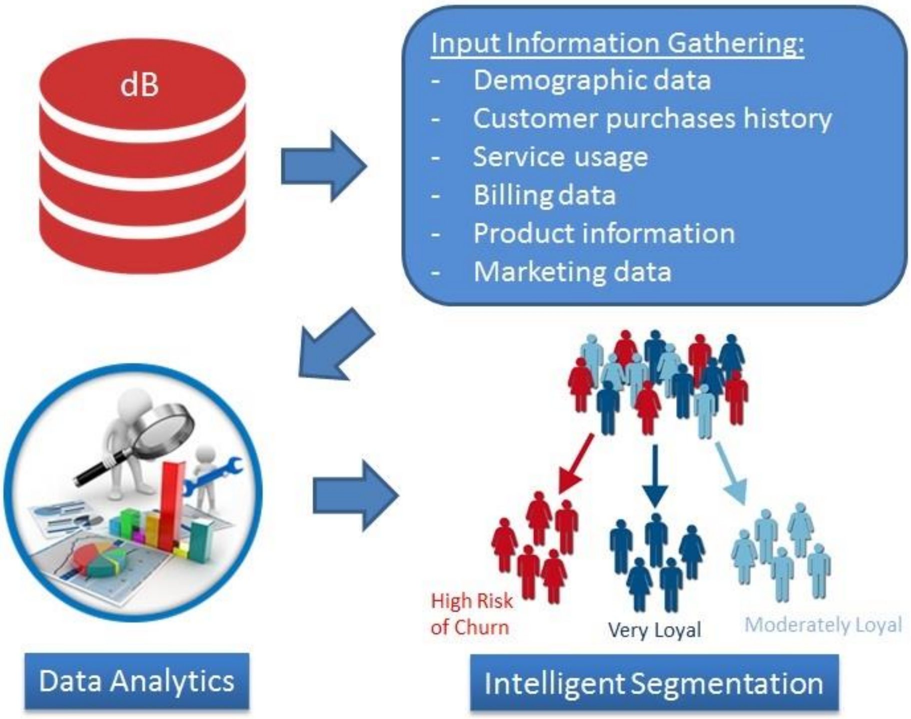


Figure 1: Big data analytics for customer segmentation [2]

3. METHODOLOGY FOR PREDICITON

A common problem-solving process to solve customer churn consists of:

1. *A **classification model** predicting the potential churner.*
2. *A **risk model** contains how actions affect probability and risk.*
3. *An **intervention model** allows to consider how the level of intervention could affect the probability of churn and the amount of customer lifetime value.*
4. *A **qualitative analysis** of the results leads to a proactive marketing campaign that targets customer segments to deliver the optimal offer.*

In this essay, the focus will be held on classification models.

The main goal of the classification prediction is to score every customer with the probability of churn and address the top N ones. In the following section the phases will be described in context of the CRISP-DM Methodology.

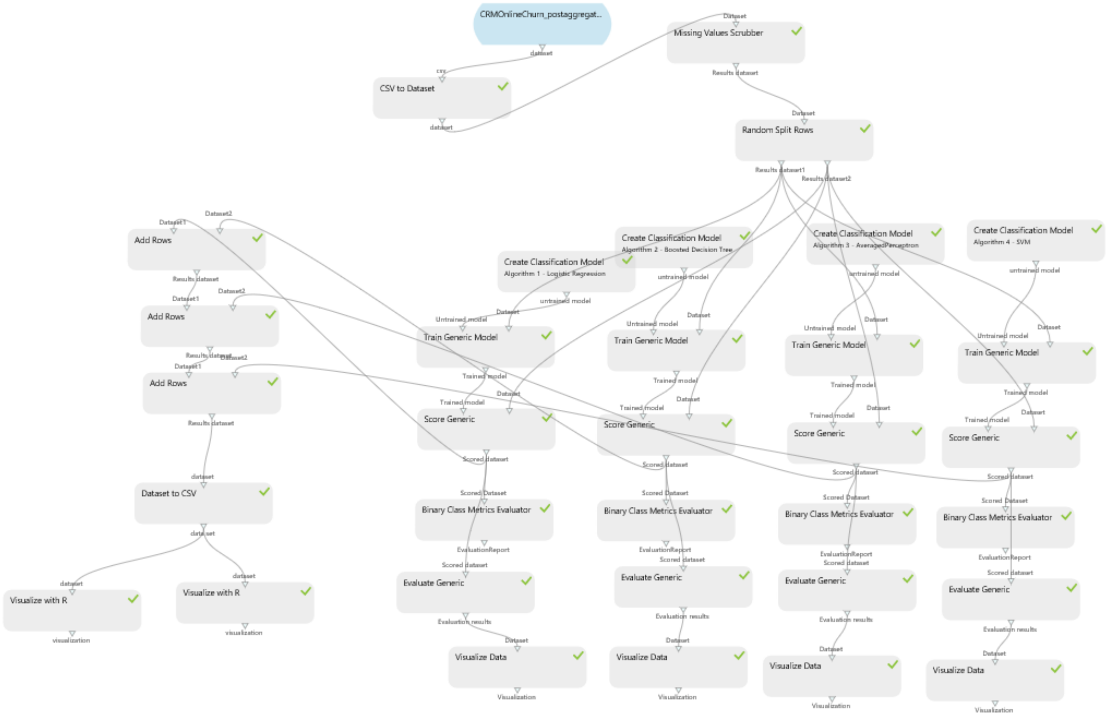


Figure 2: Analyzing Customer Churn by using Azure Machine Learning Exmaple

3.1 Data Understanding

The primary task to build a data base for prediction is to find all relevant information that could give us an insight into the current customer situation as well as historical data. An example data set that can be used to predict the churn rate includes information about:

- *Leaving customers of the last month – the Churn*
- *The assigned services of a customer – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies*
- *Customer account information – since when are they customer, contract, payment method, paperless billing, monthly charges, and total charges*
- *Demographic info about customers – gender, age range, family status*
- *Service consumption stats – how frequently do they use the service or specific features*

3.2 Data pre-Processing

In the real world, data is mostly incomplete, noisy, and inconsistent for example due to human or computer error at data entry, errors in data transmission or from different data sources. This results in the major tasks of data pre-processing includes data cleaning, data integration, data transformation, data reduction, and data discretization. [3]

Data cleaning is one of the three biggest problems in data warehousing. In the data cleaning process, some tasks may be to fill in missing values, identify outliers, smooth out noisy data, correct inconsistent data, and resolve redundancy caused by data integration. Missing and noisy data are resolved by using attribute mean to fill in, or employing a regression function to find a fitted value generally.

In this phase following issues must be solved:

- *Imputation of the missing values*
- *Discretization of numerical variables*
- *Transformation from one set of discrete values to another;*
- *Feature selection of the most informative variables;*
- *New variable derivation.*

3.3 Classification Techniques

After data are pre-processed, knowledge discovery algorithms can be applied to the processed data. The type of algorithms used, depends on the nature of the problem. If the problem can be viewed as a problem of classification or prediction (and a complete set of training data is available) then the problem is well structured. Supervised learning algorithms like multilayer neural networks, regression, support vector machine or decision trees can be used to learn the relationship between variables and correct decisions. The followings describe these four well-known algorithms in churn prediction:

3.3.1 Neural Networks

Neural networks are the popular and widely used algorithm in data mining. Neural Networks is the attempt to simulate biological neural systems which learns by changing the strength of the synaptic connection between neurons by repeated stimulation by the same impulse. [4] Neural networks can be distinguished into single-layer perceptron and multilayer perceptron (MLP). The multilayer perceptron consists of multiple layers of simple, two taste, sigmoid processing nodes or neurons that interact by using weighted connections. The MLP network may contain several intermediary layers between input and output layers. Such intermediary layers are called as hidden layers and composed of several nodes embedded in these layers, which are called as hidden nodes. Multilayer perception is a relatively accurate neural network model. [5]

3.3.2 (Boosted) Decision Tree

A decision tree is constructed by many nodes and branches on distinct stages and various conditions. It is a very popular and powerful tool for many prediction and classification problems It can produce several decision rules. Several algorithms of decision trees have been created, such as C4.5 and C5.0. Among them, classification and regression trees (CART) [6] is a non-parametric statistical method to construct a decision tree to solve classification and regression problems. Boosted decision trees incrementally build an ensemble by training each new instance to emphasize the training instances previously miss-modeled.

3.3.3 Logistic Regression

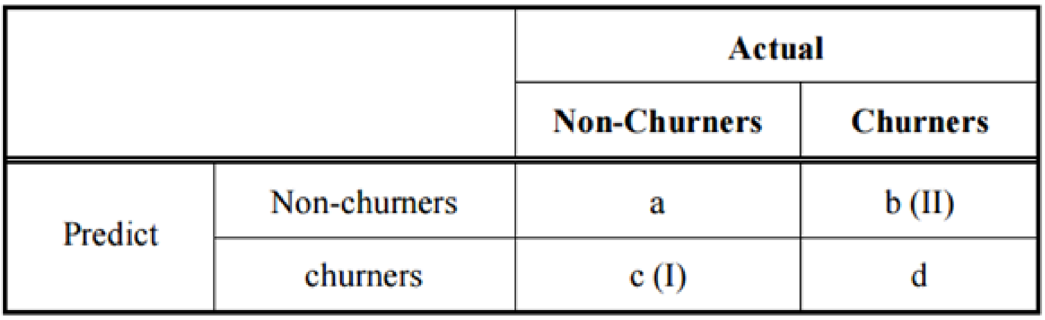
Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable, in which there are only two possible outcomes. It is used to forecast the value of two class labels or sequence variables. Even though it is one of the traditional statistical techniques, the logistic regression model does not necessarily require the assumptions of discriminant analysis. The model is as efficient and accurate as discriminant analysis. [8, 9]

3.3.4 Support Vector Machine

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples.

3.4 Evaluation

To evaluate the performance of churn prediction models, the average prediction accuracy and Type I and Type II errors are usually examined. Table 1 shows a confusion matrix used for obtaining the performance measures.



		Actual	
		Non-Churners	**Churners**
Predict	Non-churners	a	b (II)
	churners	c (I)	d

Table 1: Confusion Matrix

Type I error means the error of not rejecting a null hypothesis when the alternative hypothesis is the true state of nature. In customer churn prediction, Type I error means the event occurred when the model predicts the non-churners group as the churners group.

Type II error represents the error of rejecting a null hypothesis when it is the truth. In customer churn prediction, the Type II error means the event occurred when the model predicts the churners group as the non-churners group. In addition, to enhance the reliability of the evaluation result, N-fold cross-validation is usually used. It is based on dividing N equal parts of a given dataset, in which N-1/N of the dataset performs model training, and the rest for model testing. Every subset will be trained and tested N times, and the average prediction performance can be obtained consequently.

The rate of **prediction accuracy** is based on: $\frac{a+d}{a+b+c+d}$

4 TRADITIONAL DATA MINING VS STREAM MINING APPROACH

The traditional approach for data mining, uses data from the past to predict future behavior. For example: Find pattern and subscriber profiles that are churning.

The past prediction models were used to predict which customers would be most likely to churn based on their historical behavior. The behavior that was analyzed could cover weeks, months, or even longer. These models were implemented through campaigns at times of the choosing, not necessarily the time when the customer was most likely to churn.

Event Stream Processing (ESP) changes that by offering fine tune and implement predictive models based on specific customer behavior or events in near real time. ESP is the (near-)real time approach of stream mining. In ESP, knowledge is extracted by analyzing continuous and rapid data records (so called stream). With stream mining, a high volatile customer pattern can be analyzed and fast reactions performed. In the following are factors that can be integrated in the analysis:

- *Moves of the own company and competitors*
- *Effects of marketing campaigns*
- *Billing changes*
- *External events*
- *Market changes*
- *Social trends*

A continuous online construction of the predictive models is performed, instead of building the models offline and applying them online.

4.1 Ericsson and UPC Architecture Proof of Concept [12]

In 2012, a Proof of Concept was created for a platform for adaptive real-time churn prediction using Stream Mining by Ericsson in cooperation with UPC Barcelona.

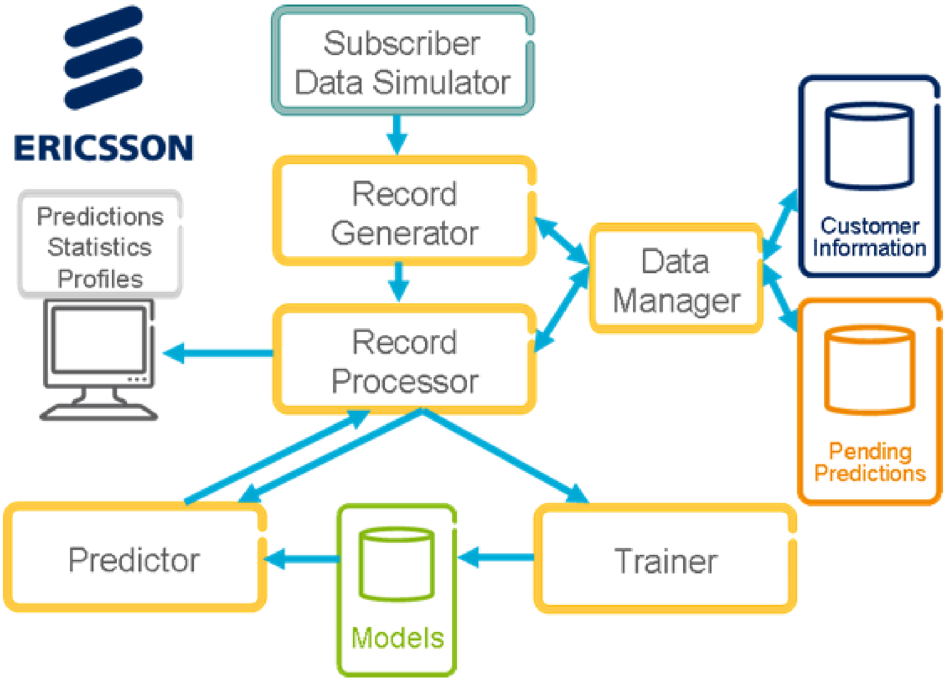


Figure 3: Architecture and Design of a Platform for Adaptive, Real-time Churn Prediction using Stream Mining [12]

Explanation of the Architecture:

Example for **stream data** (in Figure 2 Subscriber Data Simulator) can be call records, billing actions by company, bill payments by the subscribers, social media activity connected to the customer, etc.

The **customer Information data base** contains the customer master data as well as dynamic data. An example for dynamic data are numbers called, data usage, etc.

The **record generator** receives the stream data and updates in the first step the customer information data base. In the second step, it creates records out of each event and enriches them with information of the customer information data base. This record is a feature vector, that identifies the subscriber and passes all relevant information for prediction in a stream to the record processor. Record feature example: “data usage the last week”, “average internet usage time the last week”, etc.

In the **record processor,** the predictive models are build, maintained and applied. When a “churn not indicating” record gets processed, it passes through the current churn prediction model. The record with its prediction is queued into a

“Pending Predictions” queue. When a “churn indicating” record gets processed, related records are fetched of the “pending prediction” queue and passed to the model builder as positive instances of churning. Expired records in “Pending Predictions” that did not churn within a specified time are passed to the model builder as negative instances of churn. All subscriber state records are passed to the clustering method to build subscriber profiles.

The actions of each user are governed by a dynamic markovian model whose **current churn status** determines the user’s “mood”, which is in one of four states: {happy, neutral, angry, churn}. The dynamics of this model are as follows:

1. *Time between state changes is larger for smaller values of I*
2. *The more time spent in “angry” state, the higher probability of churning*
3. *A high bill or an unresolved complaint makes the customer angrier and increases the churn probability*
4. *Resolved complaints set the status back to “happy”*

This internal mood state affects the behavior of the user in multiple ways. In the following is the mood effects that are defined by UPC and Ericsson:

1. *User only complains if ”Angry”. Time of state “Angry” not measurable.*
2. *The longer time in ”Angry”, the less he calls.*
3. *The longer time in ”Happy”, the more he calls.*
4. *When user goes back to ”Neutral”, the rate of calls per day goes back slowly towards the default value.*

After identifying the churner, proper measurements to counter churn can be assessed.

4.2 Application Fields of Real Time Churn Analytics [13]

There are several fields in which real time churn prediction finds it application.

Real-time Targeted Offers and Campaign Management: Streaming analytics offer service providers to deliver individual real-time offers and focused campaigns based on “right in time” data. This can be enabled with network data, location information, customer profile data, events and rules. Creating innovative and highly relevant offers increases revenue, customer lifetime value, affinity and loyalty.

Event-Based and Personalized Marketing: Telecommunication providers can take advantage of location-based data and movement-over-time patterns that provide insight into how to better target users through geo-fencing or location-based advertising. When subscribers enter certain geographical zones, they receive a free and time limited SMS or a targeted ad banner through their social media accounts from a nearby merchant. This can be realized based on profile and behavioral characteristics.

Real-time Churn Prediction and Prevention: Telecommunication providers can identify customers with a higher chance to churn with other subscribers within their social circle. The ability to process information about all interactions that impact the customer experience in real-time is the key to identify churning.

Real-time Subscriber Experience Management: By capturing real-time geo-location data from subscriber devices, telecommunication providers can monitor individual subscribers as well as corporate customers and their activity. The retrieved data can be enriched with profitability, service status and customer profile information to get a better decision base.

4.3 Precise Use Case Example

Example of a major telecommunication company found that a specific upsell model could be made more accurate and effective by binding it to the time their customers were recharging their prepaid cards. Through ESP, the model can be realized so that when the prepaid recharge transaction was detected, an SMS promotion is sent to the customer while the transaction is still under way.[14]

Separately the effectiveness can be enhanced with other scoring models by joining cellphone usage patterns. When those patterns are detected in the ESP stream triggers are used to generate individualized offers.

5. CONCLUSION

The field of customer churn can develop into a complex field in the prediction domain and requires quantitative analysis skills in marketing.

An implementation or deployment of a customer churn model needs following things to consider:

> **First**, for the data pre-processing step, it is unknown that which feature selection method performs the best by selecting the most representative features to make prediction models provide the highest rate of accuracy. That is, there are several feature selection methods, which can be applied for churn prediction, such as principal component analysis, genetic algorithms, decision trees, stepwise, etc. [10]. Besides feature selection, outlier detection and removal is another important pre-processing task, which aims at filtering out bad/noisy data that can degrade the prediction performances.
>
> **Second,** some advanced machine learning techniques can be constructed to provide better prediction performances, for example, classifier ensembles (or multiple classifiers), hybrid classifiers, stacked generalization, etc. In short, they combine several different classifiers rather than single ones as used in the literature [11].

Personal feedback: Exploring this topic showed me an important business related prediction algorithm approach. With the won understanding about the theory, my next step of researching is the implementation of a prediction model for customer churn in an analytics platform (e.g. KNIME, Azure). My personal interest in further studies leads to the field of the implementation of real time stream processing algorithms and evaluation methods.

6 REFERENCES

[1] Amy Gallo, The Value of Keeping the Right Customers, Harvard business review, 2014

[2] About Zarema Plaksij, 12 Ways to Prevent Customer Churn, SuperOffice, 2015

[3] Han J, Kamber M. Data mining: Concepts and techniques. 2nd ed. Morgan Kaufman: USA 2006.

[4] West D, Dellana S, Qian J. Neural network ensemble strategies for financial decision applications. Comp Oper Res 2005; 32(10): 2543-2559.

[5] Zhang G, Patuwo B, Hu M. Forecasting with artifiical neural networks: the state of the art. Int J Forecast 1998; 14(1): 35-62.

[6] Yang, Q., Gupta, Y., Wilson, K., Sedukhin, I.: EP1520237A2 (2005).

[7] Breiman L, Friedman JH, Olshen RA, Stone PJ. Classification and regression trees. Wadsworth International Group 1984.

[8] Cox DR, Snell EJ. Analysis of binary data. 2nd ed. Chapman and Hall: UK 1989.

[9] Hosmer DW, Lemeshow S. Applied logistic regression. Wiley: USA 1989.

[10] Tsai C-F. Feature selection in bankruptcy prediction. Knowl-Based Syst 2009

[11] Wolpert DH. Stacked generalization. Neural Netw 1992

[12] Borja Balle, Bernardino Casas, Alex Catarineu, Ricard Gavald, David Manzano-Macho, The Architecture of a Churn Prediction System Based on Stream Mining, UPC Barcelona 2012

[13] Ari Banerjee, Big Data & Advanced Analytics in Telecom, Heavy reading, 2013

[14] Bill Vorhies, Stream Processing – What Is It and Who Needs It, Data Magnum, 2015