DE GRUYTER OPEN

JAISCR, 2016, Vol. 6, No. 2, pp. 69 – 79
10.1515/jaiscr-2016-0007

PREDICTION OF THE SHOPPERS LOYALTY WITH AGGREGATED DATA STREAMS

Vladimir Nikulin

Department of Mathematical Methods in Economy,
Vyatka State University, Kirov, Russia

Abstract

Consumer brands often offer discounts to attract new shoppers to buy their products. The most valuable customers are those who return after this initial incentive purchase. With enough purchase history, it is possible to predict which shoppers, when presented an offer, will buy a new item. While dealing with Big Data and with data streams in particular, it is a common practice to summarize or aggregate customers' transaction history to the periods of few months. As an outcome, we compress the given huge volume of data, and transfer the data stream to the standard rectangular format. Consequently, we can explore a variety of practically or theoretically motivated tasks. For example, we can rank the given field of customers in accordance to their loyalty or intension to repurchase in the near future. This objective has very important practical application. It leads to preferential treatment of the right customers. We tested our model (with competitive results) online during Kaggle-based Acquire Valued Shoppers Challenge in 2014.

Keywords: Big Data, data streams, data aggregation, classification and regression, shopping, loyalty, churn, marketing, business informatics

1 Introduction

Customer loyalty is widely seen as a key determinant of a company's profitability. Loyalty to an object (e.g. a brand, store, service or company) is shown by favourable propensities towards that object. These propensities may be behavioural or attitudinal [5]. We can differentiate between behavioural and attitudinal loyalty, also referred to as share-of-wallet and share-of-heart, respectively. Behavioural loyalty refers to customers buying exclusively or mostly only one brand, whereas attitudinal loyalty is all about having an emotional attachment to a brand, liking it more than others, and even loving it. These two types of loyalty are not fully dependent.

Loyalty of the customers affect not only shops, but any other business: insurance, banking and telecommunication. For example, one of the major problems of mobile operators has been churning customers. Churning means that subscribers may move from one operator to another operator for some reasons such as the cost of services, corporate capability, credibility, customer communication, customer services. Therefore, churn management becomes an important issue for the mobile operators to deal with.

1.1 Customer Churn and Related Literature

Customer churn has become a significant problem for firms in publishing, financial services, insurance, electric utilities, health care, banking, internet, telephone, and cable service industries. In the cellular phone industry, annual churn rates range from 23.4% to 46% [14]. A way to manage customer churn is to predict which customers are most

likely to churn and then target incentives to those customers to induce them to stay. This approach enables the firm to focus its efforts on customers who are truly at risk to churn, and it potentially saves money that would be wasted in providing incentives to customers who do not need them.

It is a well known fact that the cost of retaining a subscriber is much cheaper than gaining a new subscriber from another mobile operator. When the unhappy subscribers are predicted before the churn, operators may retain subscribers by new offerings. In this situation, in order to implement efficient campaigns, subscribers have to be segmented into classes such as loyal, hopeless, and lost. This segmentation has advantages to define the customer intentions. Many segmentation methods have been introduced and discussed in the literature [6].

Predicting customer churn with the purpose of retaining customers is a popular scientific topic. Targeting the right customers for a specific retention campaign carries a high priority [21]. Special prediction models are developed by academics and practitioners to effectively manage and control customer churn in order to retain existing customers. As churn management represents an important activity for companies to retain loyal customers, the ability to correctly predict customer churn is necessary. Additionally, identifying the shopper who will become a loyal buyer – prior to the initial purchase - is a more challenging task [19].

Generally, customer churn has become a critical issue, especially in the competitive and mature credit card industry. From an economic and risk management perspective, it is important to understand customer characteristics in order to retain customers and differentiate high-quality credit customers from the bad ones. However, studies have not yet adequately developed churn model based on customer characteristics and related past history [7].

In the past decade, with the help of advanced computer technology, databases have been growing rapidly. Consequently, commercial banks, insurance companies and retail networks hold enormous amounts of their customers transaction data in customer relationship management databases, including data related to sales, servicing and marketing functions. However, the data is only as good as the system that turns it into usable information. In this paper we present one particular framework or approach, which was used in application to the real retail transaction data.

1.2 Data Mining Models

In order to survive in an increasingly competitive marketplace, many companies are turning to data mining techniques for churn analysis. To control effectively customer churn, it is important to build a more effective and accurate customer churn prediction model. Statistical and data mining techniques have been utilized to construct the churn prediction models. The data mining techniques can be used to discover interesting patterns or relationships in the data, and predict or classify the behavior by fitting a model based on available data.

For example, rough sets theory [12], [13], [4] maybe efficient to discover hidden information in data and to explore the rules and characteristics of customer churn. The decision rules can be transferred into a flow network graph to represent the connections of pathways and the degrees of their interdependency [20].

Note that the historical data is usually imbalanced. That is, the number of “bad” or churn customers constitutes only a small minority of the data [23]. Consequently, the classifier has tendency to ignore minority class, which is the most important. Using principles of the homogeneous ensembling we can transfer consideration from original imbalanced dataset to many randomly sampled balanced subsets [16]. Homogeneous ensemble represents an average of many single learners, each of which is based on a particular subset. As far as the quality of ensemble is an increasing function of the number of single or base learners, the whole computation process maybe very expensive in the terms of required computation time. However, by splitting the main big task into sequence of smaller subtasks, we shall be reducing required computer memory. In addition, we can use the remaining part of data for validation. Thus, in line with main computation, we can, also, compute validation trajectory (which is defined in Section 2.4 as cross-validation (CV) passport) as an average of base validators. In difference to the standard cross-validation technique, CV-passport, which maybe regarded as a homogeneous ensemble of base validators, is constructed against all training data, and, by definition, mimic closely the main solution, see Section 2.4.

Essentially, we cannot apply learning model directly to the transaction data or data stream. According to the data mining literature, it is a common practice to summarize customers' past behavior in terms of their Recency (i.e. the elapsed time since last purchase or renewal), Frequency (i.e. the number of prior purchases or renewals) and Monetary value (i.e. the total amount of purchases) or their RFM characteristics.

Data preprocessing represents a key-step to transfer the data to the standard rectangular format with rows as samples and columns as a secondary features. After that we can apply the most suitable classification or regression machine learning model. Among the popular techniques to predict customer churn are: random forests, neural networks, support vector machines or logistic regression models [22].

1.3 Mining Data Streams

Mining Data Streams is the process of extracting knowledge structures from continuous, rapidly growing data records. A data stream is a timely ordered sequence of instances that can be read only once or a small number of times using limited computing and storage capabilities. Examples of data streams include computer network traffic, phone conversations, ATM transactions, web searches, and retail (shopping) transactions, which are the subject of this paper. In many data stream mining applications, the goal is to predict behaviour of the customers in the future (well related to the time-series analysis).

A typical industrial model must be updated in a continuous basis and may be required to provide predictions on billions of events per day [8]. Many organizations today have more than very large databases; they have databases that grow without limit at a rate of several million records per day. Mining these continuous data streams brings unique opportunities, but also new challenges [3].

In the last decades, the emerging computing technologies led to a deep evolution in the ability of companies to collect, store and analyze large datasets. For each customer, thousands, or even millions of data objects are stored, enabling the analysis of the complete purchasing history. Moreover, the changes in the relationship between companies and customers, due to the recent economic and social changes, has made companies change from transaction marketing to relationship marketing [9].

1.3.1 Structure of the Paper

This paper is structured as follows. In Section 2 we describe the structure of data and the problem. In Section 2.2 we explain feature engineering as the most essential part of our method. In Section 2.3 we present the most important experimental results. Finally, Section 3 concludes the paper.

2 Acquire Valued Shoppers Challenge

This International challenge on the popular internet platform Kaggle[1] continued for 95 days from 10th April 2014 to 14th July 2014, and attracted 952 active participants. Area under receiver operating curve (AUC) was used for the evaluation. Our final result (best out of two selected submissions) was 0.61172 or, formally, 13th place in the Contest.

The Acquire Valued Shoppers Challenge asked participants to predict which shoppers are most likely to repeat purchase. To help with algorithmic development, the organizers provided complete, basket-level, pre-offer shopping history (data stream) for a large set of shoppers, who were targeted for an acquisition campaign. Training and test sets included details regarding incentive, see Table 1, where the label or indicator of loyalty was presented in the training set, and was required to be predicted in the test set.

2.1 Data

This data captures the process of offering "incentives" (or, in other words, coupons, see Table 1) to a large number of customers and forecasting those who will become loyal to the product. Let's say 100 customers are offered a discount to purchase two bottles of water. Out of the 100 customers, 30 choose to redeem the offer. These 30 customers are the focus of this competition. The participants were asked to rank the field of 100 customers under offer according to the probability to

[1] http://www.kaggle.com

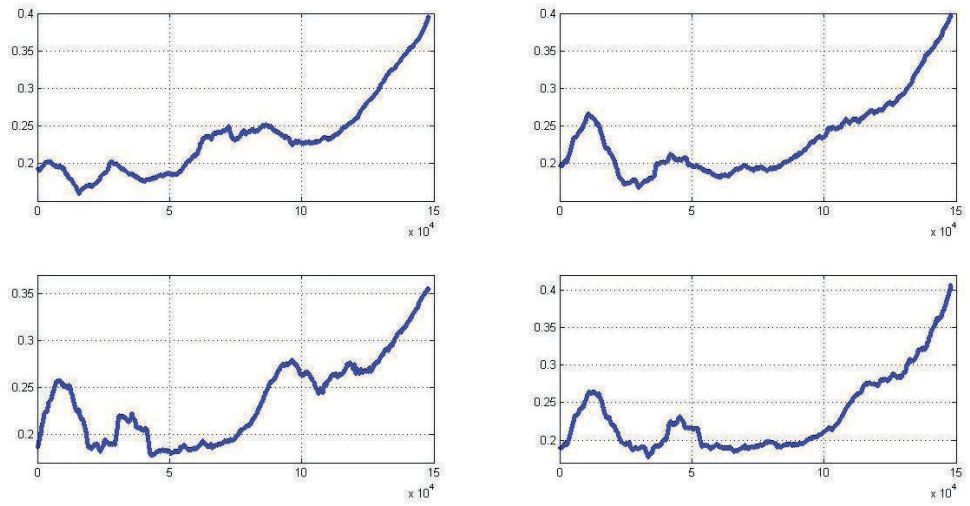

Figure 1. Moving averages in the terms of labels: top row, from left to right: $x_{cat}(a)$ and $x_{cat}(a,60)$; bottom row, from left to right: $x_b(a)$ and $x_b(a,60)$. Horizontal axis is defined in (1a), vertical axis is defined in (1b), used smoothing parameter $\Delta = 12000$, see Section 2.3

purchase the same item again in the near future. The database is binary: in the training dataset loyal customers have label one, and, otherwise, the value of the label is zero.

To create this prediction, it was given a minimum of a year of shopping history prior to each customer's incentive, as well as the purchase histories of many other shoppers (some of whom received the same offer). The transaction history contains all available items purchased, not just items related to the offer. Only one offer per customer is included in the data. The training set is comprised of offers issued from 1st March to 30th April 2013. The test set includes offers issued from 1st May to 31st July 2013.

This challenge provides 349,655,789 rows of completely anonymized transactional data (data stream) from over 300,000 shoppers: 160,057 shoppers in the training set (among them 43455 or about 27.15% are loyal) and 151,484 shoppers in the test set, see Table 2.

Table 1. List of the defining characteristics for an incentive offer. The total number of offers is 37

0	offer (index)
1	category
2	quantity
3	company
4	offervalue
5	brand

The file containing all given transactions has size of more than 22GB, see Table 3. To increase speed of computations, we reduced it to 1.66GB (27,764,694 records) by filtering those records which do not contain any of the companies, brands or chains compared to the training and test sets.

Table 2. Training data: list of features or defining characteristics with known outcome (fields N4 and N5 - are labels). The structure of the test data is the same, where fields N4 and N5 were removed

0	shopper id
1	chain
2	offer (index)
3	market
4	repeattrips
5	repeater
6	**date of offer**

Table 3. Transaction data: list of fields per one record

0	shopper id
1	chain
2	department
3	category
4	company
5	brand
6	**date of purchase**
7	productsize
8	productmeasure
9	purchasequantity
10	purchaseamount

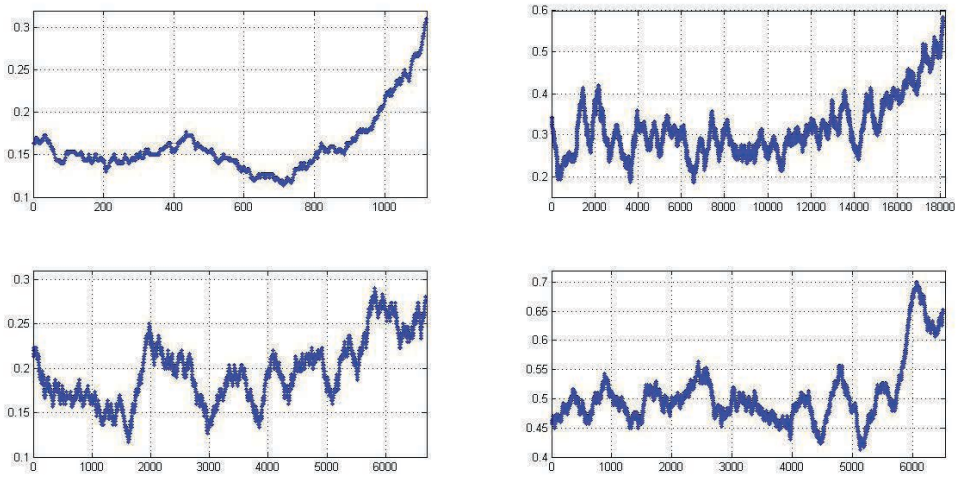

Figure 2. Part 1: particular moving averages (corresponding to different offers) with $x_b(a,60)$. Horizontal axis is defined in (1a), vertical axis is defined in (1b), used smoothing parameter $\Delta = 300$, see Section 2.3

2.2 Feature Engineering

Many organizations have collected and stored a wealth of data about their current and past and potential customers, suppliers and business partners. However, the inability to discover valuable information hidden in the data prevents the organizations from transforming these data into valuable and useful knowledge. Data mining tools could help these organizations to discover the hidden knowledge in the enormous amount of data [15].

"Category", "company" and "brand" represent the most important characteristics in the transaction data. These data are very noisy, and must be aggregated to the specific periods, which must be large enough. But not too large in order to prevent oversmoothing.

We generated the following ten main secondary features:

1. x_c - "has bought from the company": the number of times a shopper has bought from the company on offer;

2. $x_c(a)$ - "has bought amount from the company": the total amount a shopper has bought from the company on offer;

3. $x_c(q)$ - "has bought the quantity from the company": the quantity of items a shopper has bought from the company on offer;

4. $x_c(30)$ - "has bought from the company": the number of times a shopper has bought from the company on offer in 30 days (smoothing period) before the date the coupon was offered;

5. $x_c(60)$ - the same as above: 60 days before the date the coupon was offered;

6. $x_c(90)$ - the same as above: 90 days before the date the coupon was offered;

7. $x_c(120)$ - the same as above: 120 days before the date the coupon was offered;

8. $x_c(150)$ - the same as above: 150 days before the date the coupon was offered;

9. $x_c(180)$ - the same as above: 180 days before the date the coupon was offered;

10. $x_c(n)$ - "has never bought from the company": a negative feature indicating that the shopper has never bought from the company on offer before.

Remark 1 *Notations: as an illustration only, above secondary features were defined in the terms of "company". The remaining cases of "category" and "brand" are supposed to be considered similarly.*

Remark 2 *In our experiments we used six periods of smoothing: 30, 60, 90, 120, 150 and 180. It follows from Table 5 that the importance of the feature is an increasing function of the period of smoothing in the cases of "company" and "category". But, in the case of "brand" it is increasing to the level of 60, and is declining after that.*

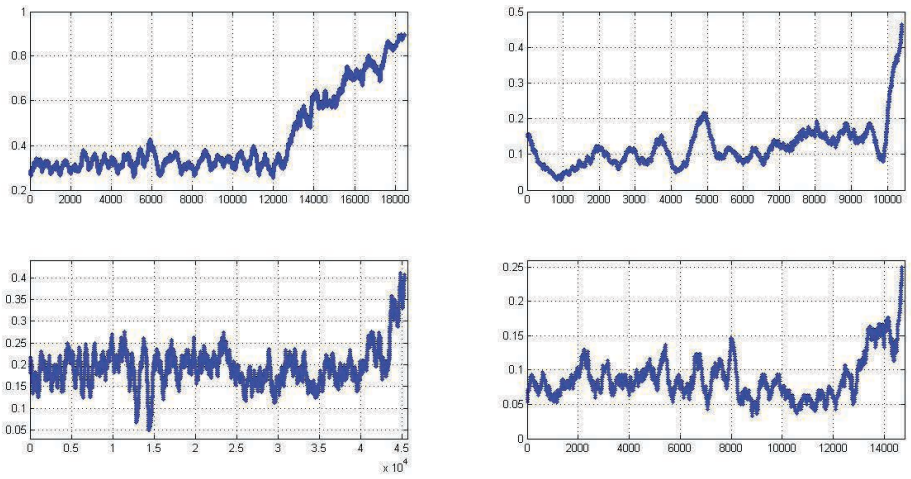

Figure 3. Part 2: particular moving averages with $x_b(a,60)$. Horizontal axis is defined in (1a), vertical axis is defined in (1b), used smoothing parameter $\Delta = 300$, see Section 2.3

Table 4. First block of 11 features, where importance ratings were computed using randomForest function in R, see, also, Table 5

N	name	rating
1	never bought company	2.28
2	never bought category	3.19
3	never bought brand	1.77
4	bought company, category and brand	1.64
5	bought category and brand	1.38
6	bought company and brand	2.3
7	category	86.22
8	company	99.4
9	offer value	22.78
10	brand	48.2
11	purchase amount	48.62

In total, we have 56 features in the main database: $11+3\cdot 15$, where first 11 features are given in Table 4, and another three blocks of features (with 15 features each) are given in Table 5. In the terms of electronic memory the size of the training dataset with 56 features is about 42MB (test dataset is about 41MB). We note that there are many ways how to extend this database further, see Section 2.5. All importance ratings for the features given in Tables 4 and 5 were computed using special R-based function Random Forests.

The main novelty and contribution of our paper is successful application of the above method to one important Big Data problem. We do believe, this method maybe extended to a variety of different problems, but any particular case should be considered specially, and any problem related to Big Data cannot be straightforward and simple.

Table 5. Second, third and fourth blocks with 15 features each. The index "c", see column "name", corresponds to "company". Indexes "cat" and "b" are used for "category" and "brand", respectively

N	name	company	category	brand
1	x_c	11.74	29.18	9.38
2	$x_c(q)$	14.14	25.47	10.68
3	$x_c(a)$	21.58	42.6	19.5
4	$x_c(30)$	4	7.68	9.02
5	$x_c(q,30)$	4.49	7.97	9.1
6	$x_c(a,30)$	11.22	17.63	17.55
7	$x_c(60)$	5.3	8.29	15.8
8	$x_c(q,60)$	6.58	8.48	15.34
9	$x_c(a,60)$	16.01	23.38	32.03
10	$x_c(90)$	5.46	10.51	9.12
11	$x_c(q,90)$	6.92	10.96	9.25
12	$x_c(a,90)$	17.43	30.37	26.79
13	$x_c(180)$	8.04	17.97	8.37
14	$x_c(q,180)$	9.94	16.5	8.6
15	$x_c(a,180)$	20.4	38.66	22.78

2.2.1 Data Compression

In the above section we described method to aggregate the given data stream to a standard rectangular format. This data-table will be suitable as an input for many data mining regression and classification packages available in R, Python or Matlab. From the other point of view, data storage facilities are limited, and as a very significant advantage

of our method, we highlight the fact that the new data maybe stored or compressed to a required format immediately after collection. There are no any need to store all the huge transaction data as a time-series sequence as described in Section 2. The compressed data maybe used as an input (to compute predictions for the selected data samples) immediately as requested.

2.3 Experiments

Table 6. Experimental results in terms of AUC, where last column "RS" indicates the number of random sets which were used in the experiment, see Section 2.4.2

Model	CV	LB	CV1	LB1	RS
RF	0.702	0.5921	0.7072	0.5938	40
gbR	0.7087	0.5797	0.7101	0.5805	50
glmNet	0.6188	0.5882	0.6612	0.591	200
Nnets	0.635	0.5318	0.638	0.5467	100

The total number of offers under consideration is 37. The structure of any offer is given in Table 1. We grouped offers with the same brands. Consequently, we reduced the number of offers to 19. After that we filtered those groups, which are not big enough (smaller that 1000). As a result, we have got 12 groups. Moving averages in the terms of lables with $x_b(a,60)$ are presented in Figures 2 - 4, where we used $\Delta = 300$ as a smoothing parameter, see equations (1a - 1b), defining horizontal and vertical axes.

2.3.1 Computation of the moving averages

The objective of this section is to explain in full details the computation procedure, which was employed in order to prepare important illustrations Figures 1 - 4.

First, we shall select Δ - smoothing parameter. Let us consider matrix A with 2 columns: 1) $x_b(a)$ and 2) y - target variable. We sort A according to the first column in an increasing order. As a result, we shall obtain matrix B as a sorted matrix A.

Figures 1 - 4 illustrate moving averages, which are defined by the following formula

$$MovAv_x(t) = \frac{1}{\Delta}\sum_{i=t}^{t+\Delta-1} B_{i,1}, t = 1,\ldots,n-\Delta+1; \tag{1a}$$

$$MovAv_y(t) = \frac{1}{\Delta}\sum_{i=t}^{t+\Delta-1} B_{i,2}, t = 1,\ldots,n-\Delta+1, \tag{1b}$$

where n is the number of rows in matrix A; $MovAv_x$ - horizontal axis and $MovAv_y$ - vertical axis.

Remark 3 *We used $x_b(a)$ as an example, any other feature is to be considered absolutely similarly.*

2.3.2 Vowpal Wabbit

In addition to the models, presented in Table 6, we, also, conducted experiments with well known software Vowpal Wabbit [1], and had observed quite competitive results: 0.59104 in public (preliminary test score on the LeaderBoard, which was observed during life of the Contest), and 0.58538 in private (final test score).

Remark 4 *About the same results were reported by other participants on the forum of the Contest.*

2.3.3 Regulation parameters

The following parameters (as base points) were used in our experiments.

1. Vowpal Wabbit: 0.85 - learning rate, 40 - number of passes or iterations, 0.6 - quantile parameter;

2. randomForest in python: n.estimators = 200, criterion = 'mse', max.depth=12, min.samples.split= 10, max.features = 'sqrt', min.samples.leaf=3, min.density =0.1;

3. gbR (GradientBoostingRegressor) in python: n.estimators =500, learning.rate=0.08, max.depth=5, loss='ls';

4. glmNet in R: family = 'binomial', alpha = 0, $lambda = 2^{12}$;

5. nnet in R: size=5, maxit=200.

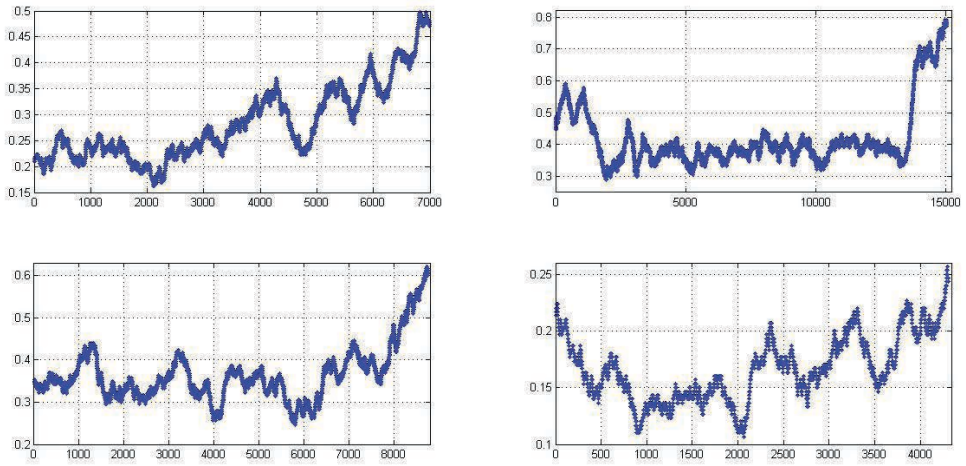

Figure 4. Part 3: particular moving averages with $x_b(a,60)$. Horizontal axis is defined in (1a), vertical axis is defined in (1b), used smoothing parameter $\Delta = 300$, see Section 2.3

Remark 5 *Above parameters cannot be regarded as an optimal in any case. Based on the CV-experiments applied to the particular dataset (we considered datasets with up to 120 secondary features), we can conduct optimisation of the parameters, which may affect performance of the model significantly.*

2.4 Calculation of the CV-passports as a validation trajectories against all training data

The effectiveness of an ensemble is highly dependent on the quality (accuracy and diversity) of the individual classifiers. Diversity in ensembles can be reached when the individual classifiers were built with different training subsets. In this case, we can use the remaining labelled data for calculation of the individual validators. In machine learning, the ensembles of classifiers represent a very efficient barrier against traps by a local minimum.

We can expect that the ensemble of validators will, also, perform as a method of validation at a more advanced level. In our earlier papers [17] and [18], we introduced cross-validation (CV) passport of homogeneous ensemble as unified validation trajectory against all available training data. In general terms, to ensure a high quality of the ensemble, we have to consider a large number of the single learners, and, accordingly, the related computational process maybe very expensive. The proposed definition of the CV-passports naturally fits the mechanism of the construction of the ensemble and doesn't require any extra computation time. Based on the fundamental principles of our approach, CV-passport will naturally accompany any homogeneous ensemble.

Assuming that CV-passports mimic closely the corresponding test solutions, we can use them for the consideration of many tasks including optimizations of blends and heterogeneous ensembles, calculation of the biases and cut-off parameters, construction of novel features and many other statistical characteristics as required.

In accordance with the principles of homogeneous ensembling, it appears to be natural to consider calculation of the decision function as an average of the large number of the single learners (or base classifiers), where any single learner is based on the randomly selected subsamples of observations and features.

2.4.1 Definition of CV-passports

Let us describe the proposed method in more details. Suppose that we are using about 30% of the available data for training during one global iteration. Then, remaining 70% of the data maybe used for the validation control. In line with the principles of cross-validation, we shall test stability of the validation results by considering a sequence of the random splittings. It will take only 5-10 global iteration to fill all the gaps in the training data. Accordingly, we shall be having averaged validation results against all the samples in the training data.

Definition 1 *We are proposing to accumulate the validation results against all training samples in line with construction of the homogeneous ensemble. We shall call averaged validation trajectory as a CV-passport of the corresponding homogeneous ensemble.*

2.4.2 Experiments with CV-passports

In Table 6 the columns "CV" and "LB" correspond to the solutions, which were computed using specified models applied to the data with 56 features as described in Section 2.2. All solutions were computed using principles of homogeneous ensembling. Accordingly, we were able compute CV-passports corresponding to the test predictions. Combined together, CV-passport and corresponding test prediction maybe used as a new feature. We added this feature to the original data. After that, we computed new solution, see columns "CV1" and "LB1", and observed some improvement.

There are many ways how to compute secondary features. Some of those methods are presented in the following Section 2.5. We note, also, that according to the study [21], Random Forests performs best (in terms of AUC) in distinguishing churners from non-churners. This conclusion coincides with our results, see Table 6.

2.5 Some alternative methods

First of all, we can consider aggregation, applied not only to the pure characteristics: "company", "category" and "brand" (see Section 2.2), but to their combinations by two and by three. In total, there are four such combinations.

2.5.1 Bayesian approach

As it was mentioned in Section 2.1, the given transaction data contains only 7.55% of the records with relevant characteristics in the terms of "companies", "categories" and "brands". Using historical transaction data, we can establish relationships between relevant and remaining characteristics. In total, there are nine relationships: $c \to c, c \to cat, c \to b; cat \to c, cat \to cat, cat \to b; b \to c, b \to cat, b \to b$. We computed number of occurrences for any pair of items when both items were seen together for the same customer. After that we can compute Bayesian probabilities, which establish relation between observed and relevant characteristic. With this approach we can form new secondary features.

2.5.2 Matrix factorisation with stochastic gradient descent

Note that training data is perfectly fit to apply stochastic gradient descent (SGD) to conduct matrix factorisation applied to the pairs

$$\{id, c\}, \{id, cat\}, \{id, b\},$$

where *id* - is a customer index. Consequently, we computed vectors of k factors for any *id*, *c*, *cat* and *b*. We note that customers in the training and test sets are different, but "companies", "categories" and "brands" are the same. Accordingly, we can use k-dimensional vectors of factors as a secondary features.

Remark 6 *We observed some reasonable improvement in performance with SGD-technique as described in this section.*

3 Concluding Remarks

Customer relationship management (CRM) is one of the most significant managerial tasks in organizations. CRM as a combination of systems and techniques that supports building strategic relationships with customers in a long term and profitable fashion [10].

One of the managerial concerns in any business is to understand the risks and also the opportunities that the organization is dealing with. Predictive models are used to identify these challenges through investigating historical and transactional customer data. Thus, predictive analytics are considerably different from traditional business intelligence tools. Traditional *Business Informatics* tools are only useful to depict and explain past trends and performance of organizations based on historical data, while predictive analytics are capable of making predictions, inform decisions and forecast future movements of the customers and industries.

The organizations may have an ocean of data but still they are starving for information or more specifically for valuable knowledge. Data mining tools are necessary in order to help these organizations to extract hidden patterns of useful information [2]-[11].

The data pre-processing stage in data mining is a very important step for the final model performance in the terms of prediction quality. With proposed in the paper method of aggregation or timely moving sums (see Section 2.2), we can transfer huge dataset of transactions to the traditional format of rectangular matrix, where any row corresponds to the particular customer and column corresponds to the secondary aggregated feature. Note, also, that the proposed method maybe used for data compression: we can store the data in a required format directly, and without any intermediate steps. According to our calculations, the data in a new rectangular format will require about 50 times less space in the terms of electronic memory.

The main finding in this paper: the artificially created novel features (see Section 2.2) are very informative: Figures 1 - 4 illustrate strong correspondence between intensity of the transactions during a few months immediately before offer and loyalty of the customer.

References

[1] A. Agarwal, O. Chapelle, M. Dudik and J. Langford. A Reliable Effective Terascale Linear Learning System. Journal of Machine Learning Research, 15, 2014, pp. 1111 – 1133.

[2] V. Bhambri. Data Mining as a Tool to Predict Churn Behaviour of Customers. International Journal of Management Research, April 2013, pp. 59 – 69.

[3] P. Domingos and G. Hulten. Mining high-speed data streams. KDD 2000, pp. 71 – 80.

[4] P. Dhandayudam and I. Krishnamurthi. Customer Behavior Analysis Using Rough Set Approach. Journal of Theoretical and Applied Electronic Commerce Research. ISSN 0718-1876 Electronic Version, Universidad de Talca - Chile, Vol. 8(2), 2013, pp. 21-33.

[5] R. East, P. Gendall, K. Hammond and W. Lomax. Consumer Loyalty: Singular, Additive or Interactive? Australasian Marketing Journal, 13(2), 2005, pp. 10 – 26.

[6] A. Karahoca, D. Karahoca and N. Aydin. Benchmarking the Data Mining Algorithms with Adaptive Neuro-Fuzzy Inference System in GSM Churn Management. Data Mining and Knowledge Discovery in Real Life Applications, Book edited by: Julio Ponce and Adem Karahoca, 2009, Vienna, Austria, pp. 229 – 242.

[7] C.-S. Lin, G.-H. Tzeng, Y.-C. Chin. Combined rough set theory and flow network graph to predict customer churn in credit card accounts. Expert Systems with Applications, 38, 2011, pp. 8 – 15.

[8] H. McMahan, G. Holt, D. Sculley, M. Young, D. Ebner, J. Grady, L. Nie, T. Phillips, E. Davydov, D.Golovin, S. Chikkerur, D. LiuM. Wattenberg, A.Hrafnkelsson, T. Boulos, J. Kubica, Ad Click Prediction:a View from the Trenches, KDD '13 Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, Pp. 1222-1230, ACM New York, NY, USA 2013

[9] V. Migueis, D. den Poel, A. Camanho, J. Cunha. Modeling partial customer churn: On the value of first product-category purchase sequences. Expert Systems with Applications, 39, 2012, pp. 11250 – 11256.

[10] T. Mirzaei and L. Lyer. Application of Predictive Analytics in Customer Relationship Management: a Literature Review and Classification. Proceedings of the Southern Association for Information Systems Conference, Macon, GA, USA March 21st - 22nd, 2014.

[11] N. Hashmi, N. Butt and M. Iqbal. Customer Churn Prediction in Telecommunication A Decade Review and Classification. International Journal of Computer Science Vol. 10(5), 2013.

[12] W. Hu. Developing pertubation rate of the rough set theory to evaluate the electronic transaction quality of on-line shopping. Pakistan Journal of Statistics, 2012 Vol. 28(5), pp. 581-596.

[13] J. Liou, G.-H. Tzeng. A Dominance-based Rough Set Approach to customer behavior in the airline market . Information Sciences, 2010, 180, pp. 2230-2238.

[14] S. Neslin, S. Gupta, W. Kamakura, J. Lu, and C. Mason. Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models. Journal of Marketing Research, 43, 2006, pp. 204 - 211.

[15] E.W.T. Ngai, L. Xiu, D.C.K. Chau. Application of data mining techniques in customer relationship management: a literature review and classification. Expert Systems with Applications, 36, 2009, pp. 2592 - 2602.

[16] V. Nikulin. Classification of Imbalanced Data with Random Sets and Mean-Variance Filtering. International Journal of Data Warehousing and Mining, 4(2), 2008, pp. 63 – 78.

[17] V. Nikulin, A. Bakharia and T.-H. Huang. On the Evaluation of the Homogeneous Ensembles with CV-passports. LNCS 7867, Springer, J.Li et al. (Eds.), PAKDD 2013 Workshops, pp. 109 – 120.

[18] V. Nikulin. Hybrid Recommender System for Prediction of the Yelp Users Preferences. ICDM 2014, St.Petersburg, Russia. LNAI 8557, Springer, P. Perner (Eds.), pp. 85 – 99.

[19] B. Pal, R. Sinha, A. Saha, P. Jaumann and S. Misra. Customer Targeting Framework: Scalable Repeat Purchase Scoring Algorithm for Large Databases. Proceedings of 2012 4th International Conference on Machine Learning and Computing IPCSIT, vol. 25, 2012 IACSIT Press, Singapore, pp. 143 – 146.

[20] Z. Pawlak. Rough set. International Journal of Computer and Information Sciences, 11(1), 1982, pp. 341 – 356.

[21] K. Coussement, D. Van den Poel. Improving customer attrition prediction by integrating emotions from client/company interaction emails and evaluating multiple classifiers. Expert Systems with Applications, 36, 2009, pp. 6127 – 6134.

[22] A. Sharma and P. K. Panigrahi. A Neural Network based Approach for Predicting Customer Churn in Cellular Network Services. International Journal of Computer Applications, 27(11), 2011, pp. 26 – 31.

[23] Y. Xie, Xiu Li, E.W.T. Ngai, W. Ying. Customer churn prediction using improved balanced random forests. Expert Systems with Application, 36, 2009, pp. 5445 - 5449.

Vladimir Nikulin received his PhD from the Moscow State University in February 1986. He was working as a Scientific Programmer at the Perm State University before joining Vyatka State University in December 1986. From 1993 to September 2011, he worked in a commercial and academic sectors in Australia. He published over 66 papers in refereed journals and conferences, and, currently, serves as an Associate Professor in Mathematics and Computer Science. His research interests include machine learning, data mining and bioinformatics. Vladimir Nikulin received formal awards for successful participation in data mining competitions from the following conferences: IJCNN-2007, AI-2009, ICDM-2009, JSM-2010, RSCTC-2010, PAKDD-2010, PKDD-2011, RecSys-2013.