Budnikas, Germanas

Article

COMPUTERISED RECOMMENDATIONS ON E-TRANSACTION FINALISATION BY MEANS OF MACHINE LEARNING

Statistics in Transition New Series

Provided in Cooperation with:
Polish Statistical Association

Suggested Citation: Budnikas, Germanas (2015) : COMPUTERISED RECOMMENDATIONS ON E-TRANSACTION FINALISATION BY MEANS OF MACHINE LEARNING, Statistics in Transition New Series, ISSN 2450-0291, Exeley, New York, NY, Vol. 16, Iss. 2, pp. 309-322, https://doi.org/10.21307/stattrans-2015-017

This Version is available at:
https://hdl.handle.net/10419/207775

STATISTICS IN TRANSITION new series, Summer2015
Vol. 16, No. 2, pp. 309–322

COMPUTERISED RECOMMENDATIONS ON E-TRANSACTION FINALISATION BY MEANS OF MACHINE LEARNING

Germanas Budnikas[1]

ABSTRACT

Nowadays a vast majority of businesses are supported or executed online. Website-to-user interaction is extremely important and user browsing activity on a website is becoming important to analyse. This paper is devoted to the research on user online behaviour and making computerised advices. Several problems and their solutions are discussed: to know user behaviour online pattern with respect to business objectives and estimate a possible highest impact on user online activity. The approach suggested in the paper uses the following techniques: Business Process Modelling for formalisation of user online activity; Google Analytics tracking code function for gathering statistical data about user online activities; Naïve Bayes classifier and a feedforward neural network for a classification of online patterns of user behaviour as well as for an estimation of a website component that has the highest impact on a fulfilment of business objective by a user and which will be advised to be looked at. The technique is illustrated by an example.

Key words: online behaviour, Google Analytics, Naïve Bayes classifier, artificial neural network.

1. Introduction

Practically all of nowadays businesses rely on websites and web services. The structure of their interaction with a customer can be represented as a two-phase process. During the first phase a user gets some information about a service, during the second phase the user finalises his (her) transaction with a website and/or leaves the website. A transaction finalisation is a website content dependent process – it might be a service ordering, commenting, Facebook likes, etc. It is extremely important for business owners to know how website guests behave online and if it

[1] University of Bialystok, Faculty of Economics and Informatics in Vilnius. Kaunas University of Technology, Faculty of Informatics, Lithuania. E-mail: german.budnik@uwb.edu.pl.

is possible to influence their actions. This paper addresses these issues and presents results of the research.

The topic of the paper has a practical value. Analysis and understanding of web user behaviour is a key topic of a behavioural targeting. Behavioural targeting is an evolving area of a web mining that deals with optimisation of web online ads based on an analysis of web user behaviours. The research presented in the paper has some similarities to works in the considered field of the study. Methods of behavioural analysis investigate web surfing data gathered mainly from log files. The topic is actively investigated; examples of similar works include papers by (Angeletou, Rowe and Alani, 2011), (Dembczyński K., 2009), (Robinson D.J., 2008).

Approach by (Robinson D.J., 2008) suggests a method for monitoring user online behaviour. The method is implemented based on data pulled from log files where HTTP/GET requests are saved when a user clicks a hyperlink. These data are gathered using agent devices installed on a user's computer. The approach uses Open Directory Project (Xian, Chen and Wang, 2014) for a categorisation of visited websites. The research emphasises the creation of behaviour profiles with respect to web page visitation event, frequencies and probability distributions, and causality relations or time-dependencies.

The technique by (Dembczyński K., 2009) describes the problem of predicting behaviour of web users based on real historical data. The data are gathered from the user's cookie files. An analysis is performed using a statistical decision theory.

Paper by (Angeletou, Rowe and Alani, 2011) presents a method for modelling and analysis of user behaviour in online communities that include personal profiles, wiki, blogs, file sharing, and a forum. The approach implements behaviour modelling, role mining and role inference and is based on a statistical clustering.

The approach proposed in the current paper differs from the works listed above by its application area – it operates at Internet level, while (Angeletou, Rowe and Alani, 2011) and (Robinson D.J., 2008) approaches operate at Intranet level. The approach proposed is similar to (Angeletou, Rowe and Alani, 2011) because they both use a dynamical update of estimations with respect to new data.

The technique suggested in the paper consists of the following steps. At the beginning, in order to know the actual on-site user behaviour, user browsing activities should be formalised. The paper applies Business Process Modelling Notation (Drejewicz, 2012) for such formalisation. It enables a definition of data to be read off from a website during monitoring user browsing activities by means of Google Analytics (Clifton, 2012) tracking function. This permits to gather statistical data required for an analysis. The aim of such analysis is to build a model of user on-site behaviour (an earlier paper on that topic can be found in (Budnikas, 2015)) – whether a website guest is willing to finalise a transaction or not. The statistical data are handled during the second step. The model used for the analysis is based on classic machine learning techniques – Naïve Bayes classifier and a feedforward artificial neural network – Multi-layer Perceptron model. In the third step, Naïve Bayes classifier is applied to analyse the actual website user browsing

activities based on gathered statistical data. In the fourth step, the two already mentioned techniques are used together in order to classify actual user online behaviour with respect to gathered statistical data. Depending on an outcome of the classification, the website may recommend a visit to that web page to an online user, which has the biggest impact on the transaction finalisation to be defined using auxiliary classification.

The paper is structured in the following way. The second section gives a formalisation of browsing activities with respect to a website category as well as data needed to monitor a website. The third section presents a sketch of a procedure to gather statistical data from a website and to handle possible inconsistency cases. Machine learning data analysis methods used in the proposed technique are discussed in the fourth section, namely – Naïve Bayes classifier and Multi-layer Perceptron model, whose structure and parameters are given too. The fifth section illustrates the technique proposed. Conclusions summarise main results achieved and state future work directions.

2. Website formalisation

Surfing on websites usually differs with respect to types of these sites. Open Directory Project (ODP) differentiates the following website types: Arts, Business, Computers, and 13 more instances. These types generalise manually selected websites in different languages and are used in various kind of research including the suggested in this paper. Classification of websites into types helps in understanding possible kinds of behaviour. Specification of sub-types and its instances is actual for understanding behaviour cases. The paper considers an instance of the Consumer Goods and Services sub-type of a business type with respect to ODP classification. Each browsing activity on websites, especially on business sites, can be logically divided into two parts – *introductory* part, which usually includes list of services, descriptions, etc., and (transaction) *finalisation* part, which could be expressed by paying for services, commenting, Facebook likes and so on. According to Figure 1, the introductory logical part of a browsing activity may consist of Product Category Selection, Product Selection, Product Related Information Viewing, Delivery and Company Information Viewing; while Check-out and Payment browsing activity corresponds to the logical part – the transaction finalisation.

Formalisation of browsing activity makes it possible to understand user on-site behaviour that can be monitored by using various techniques, e.g., Google Analytics (Clifton, 2012) tracking function.

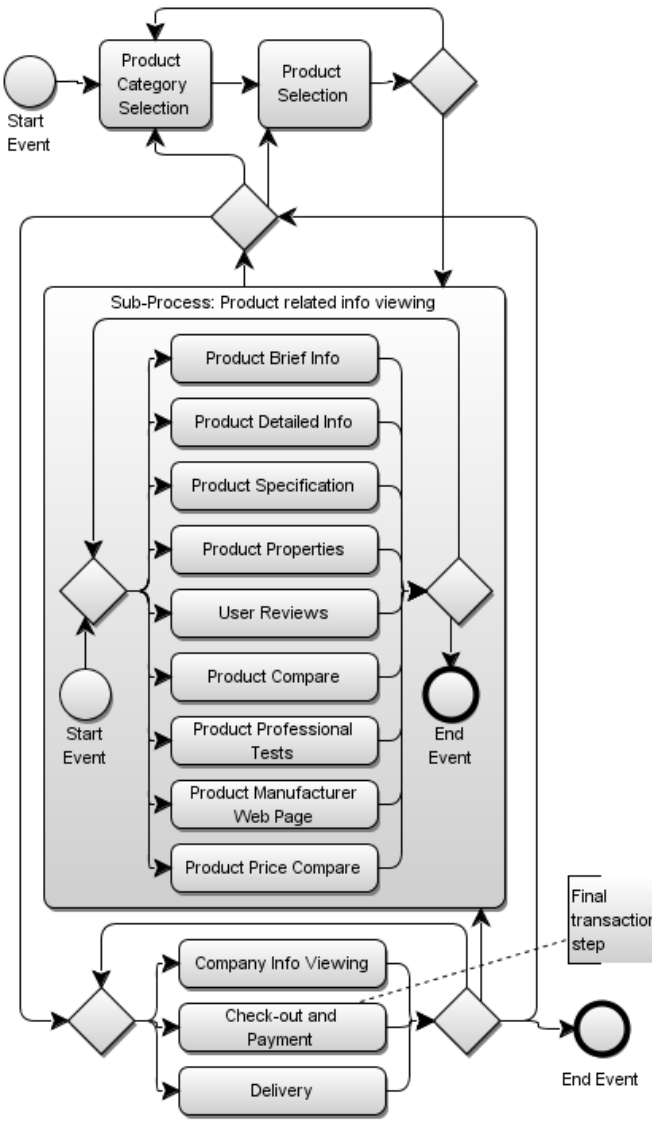

Figure 1. A generalised view of user behaviour on "Consumer Goods and Services" sites using Business Process Modelling Notation.

Source: (Budnikas, 2015).

A 5-tuple

<*e, y, u, t, m*>, where

e is user browsing *session* during which website pages are visited;

y is a *category* of a product viewed by a user. As e-commerce website may contain a huge number of products (even of the same category), products are differentiated only if they belong to the different categories;

u is a *user* that is identified by a cookie file. A cookie is a small text file that contains user visiting on-site specific information;

t is a kind of an activity or a *task* performed by a user on the website page like "Product Category Selection", "Viewing Product Price Comparison" (see Figure 1);

m is activity *t* start time *moment* which application is twofold. First, it is used to know a sequence number of a web page visit for the first time during a session. Second, it is used to count revisits to the same web page.

defines data needed for monitoring user online activities. These data also set requirements for database table where browsing activity statistical data are stored.

3. Gathering statistical data

In order to classify user actual on-site behaviour, a training data set should be collected from the site. The technique suggested in this paper uses Event Tracking method, which is a part of Google Analytics tracking code (Clifton, 2012). It enables recording user interactions with website elements, such as web page, embedded AJAX page element, page gadgets, and Flash-driven element and so on. Additionally to tracking function, a cookie file is used for unique user identification (Nikiforakis, Acar and Saelinger, 2014).

During a session of website browsing information about visited pages is collected and stored in the following form

$$\langle e, y, u, t_1, \dots, t_{n-1}, r_{t_1}, \dots, r_{t_{n-1}}, s_{t_1}, \dots, s_{t_{n-1}}, t_f \rangle,$$

where t_f corresponds the final task, s_{t_i} – means a sequence number of the t_i-th web page visit for the first time during the *e*-th session andr_{t_i} is a counter of revisits to the same t_i-th web page. For example, Table 1 record R_1 represents a situation that a user during his/her first session has visited Product Properties (t_4), Product Price Compare (t_9) and Delivery (t_{11}) web pages and has not finalised the transaction – Check-out and Payment task (t_f) has not been accomplished and web page t_4 has been visited first in a sequence (s_{t_4}=1) and was revisited twice. Task designations have the following meanings: 0 means a web page has not been visited (i.e., a task has not been accomplished) and 1 means that a web page has been visited. User

next session (see record R_2 of Table 1) consists of visits to the same pages (it is marked by grey background colour in the table) that resulted in the transaction finalisation.

As seen in Table 1, inconsistent data entries with respect to the visited web pages may exist in the gathered statistical data. An inconsistency case is when the same set of accomplished tasks in different data entries is followed by opposite finalisation tasks. A fragment of the pseudo-code of an algorithm used for the inconsistency case handling is presented next (see Figure 2). This fragment excludes variables $s_{t_i}$and r_{t_i}as they have no influence on inconsistency. Note also that if the same user visits a site repeatedly and his/her browsing activity is different, corresponding records do not join since separate records represent a real situation on browsing activities in a database. Such an approach also simplifies computations.

Table 1. Illustration of statistical data fragment read off from a website

Record number	Session	Product category	User ID	Brief Info	Detailed info	Specification	Properties	User Reviews	Compare	Professional Tests	Manufacturer Web page	Price Compare	Company Info	Delivery	Check-out and Payment	…	Revisits# to Properties page	Sequence# of web page 1st visit	…
	e	y	u	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}	t_f	…	r_{t_4}	s_{t_4}	
R_1	1	1	1	0	0	0	1	0	0	0	0	1	0	1	0	…	2	1	…
R_2	2	1	1	0	0	0	1	0	0	0	0	1	0	1	1	…	0	3	…

Source: own elaboration.

Algorithm Handling of inconsistency cases in gathered statistical data

1: $T^I = \emptyset; T = \emptyset$
2: $T = T \cup t^i | t^i = \{y^i, u^i, t_1^i, \dots, t_{n-1}^i, t_f^i\}$
3: If $t_f^i \neq t_f^j$, where $\forall i, j, k: i \neq j; y^i = y^j; t_k^i = t_k^j; t^{i,j} \in T \cup T^I$ then
4: If $u^i = u^j$, $t_f^i \neq t_f^j$, $t_f^j = 0$, $t_k^i = t_k^j$ then
5: $T = T \cup t^i \setminus t^j$
6: $T^I = T^I \setminus t^i \setminus t^j$
7: End If
8: If $u^i \neq u^j$, $t_f^i \neq t_f^j$, $t_f^j = 0$, $t_k^i = t_k^j$ then
9: $T^I = T^I \cup t^i \cup t^j$
10: $T = T \setminus t^i \setminus t^j$
11: End If
12: End If
13: GOTO 2

Figure 2. A fragment of the algorithm for inconsistent data handling
Source: own elaboration.

The algorithm initialises an inconsistent data set T^I and a statistical data set *T*. Further, the set *T* is supplemented with data about web page visits, a user, and a product category. If the transaction finalisation tasks t_f^i and t_f^j in *i* and *j* data entries from the all data sets are different while the rest of accomplished tasks are the same for the same product category, two inconsistency handling options are available – described in the steps 4-7 and 8-11 respectively. If inconsistency has arisen in the browsing sessions by the same users u^i and u^j, data entry t^i corresponding to the transaction finalisation is added to the statistical data set *T* and excluded from the inconsistent data set T^I, while opposite data entry t^j is excluded from all the sets. If inconsistency has arisen in browsing sessions by different users, inconsistent data entries $t^i$and t^j are added to the set T^I and excluded from the set *T*. The algorithm is repeated starting from the step 2 along with the arrival of data about next browsing session.

4. Machine learning data analysis methods

In spite of recent research in big data analysis that is common for well-known e-commerce sites, less known e-commerce websites still exist, whose customer visits and number of successful transaction finalisations are not so big. As statistical data are being gradually added to a database, the number of training data entries is not sufficient for some classification methods. This fact sets a premise to use a classification technique like Naïve Bayes classifier that works well with a

comparatively small set of training data. When the number of statistical data reaches the threshold corresponding to the minimal number of entries in a training data set that is sufficient for a classification with a predefined error level, Multi-layer Perceptron (MLP) technique is applied additionally to Naïve Bayes classifier. If outcomes of the two classification methods are different, a class that represents transaction non-finalisation is regarded as dominating. The threshold is calculated using the rule of thumb

$$threshold = number\ of\ weights\ /\ error\ level$$

where *number of weights* and *error level* are parameters of the MLP model.

If estimation of actual user browsing activity shows transaction non-finalisation possibility and a user has visited at least 30% of all pages as described in website activity formalisation step, a website-to-user interaction procedure starts (note that 30% level is set based on experiment outcomes). The purpose of the procedure is to estimate the web page with the highest impact on the transaction finalisation and suggest a user to visit that page. Estimation experiments use the classification technique to find maximal similarity to the desired class while considering distinct unvisited web pages.

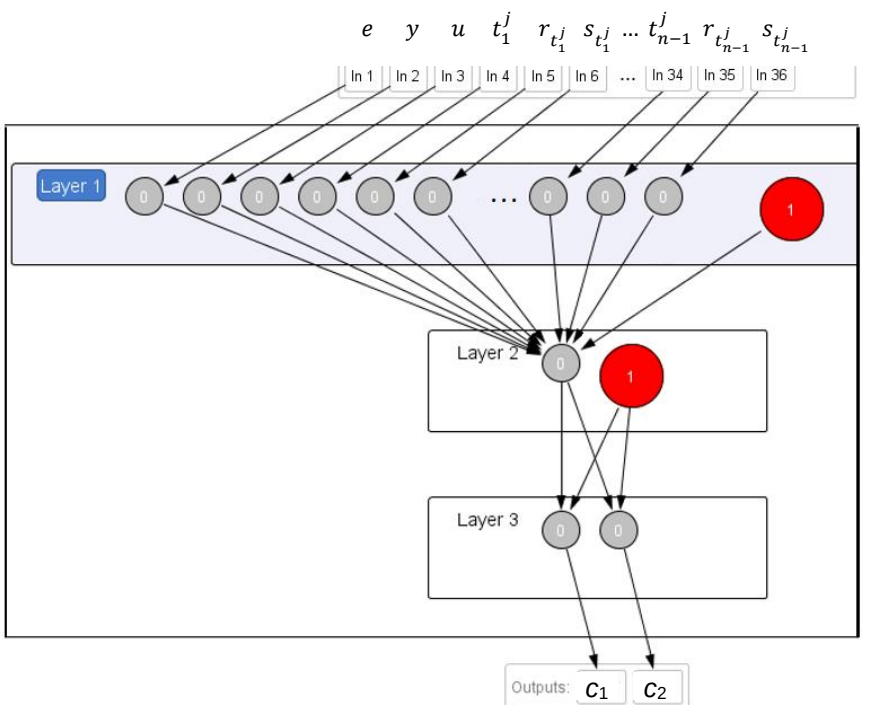

Figure 3. A structure of Multi-layer Perceptron model

Source: own elaboration using NeurophStudio.

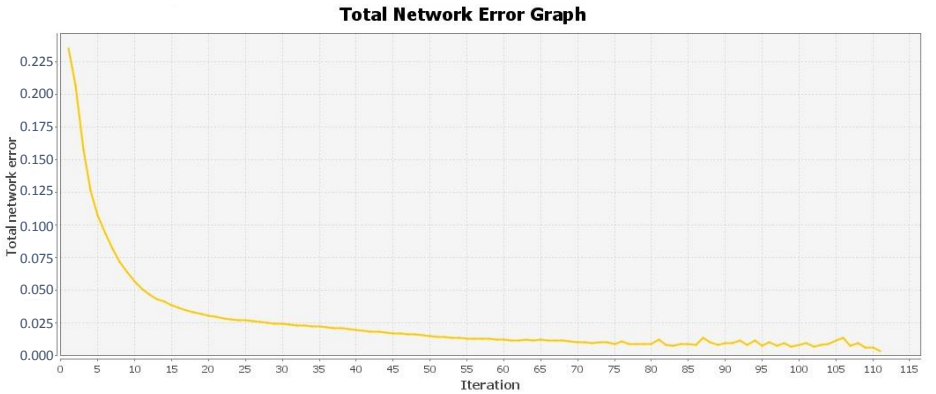

Figure 4. A total error network graph for the MLP model with respect to the number of training data set equals 7800, max error level equals 0.005, and learning rate equals 0.15

Source: NeurophStudio generated graph based on training process of the MLP.

Naïve Bayes classifier is calculated using a classical formula (Russell, 2010):

$$\text{classify}\left(e, y, u, t_1^j, \ldots, t_{n-1}^j, r_{t_1^j}, \ldots, r_{t_{n-1}^j}, s_{t_1^j}, \ldots, s_{t_{n-1}^j}\right) =$$

$$\underset{c \in \{c_1, c_2\}}{\text{argmax}}\, p(C = c) \prod_{i=1}^{n-1} p\left(\langle e, y, u, t_i^j, r_{t_i^j}, s_{t_i^j}\rangle \,|\, C = c\right), \text{ where}$$

C denotes one of the possible classes representing the transaction finalisation (c_1) or non-finalisation (c_2). Note that t_f is not used in the formula as defined in the classical approach because it corresponds the final task, which occurrence probability is evaluated.

MLP uses data about browsing activity $e, y, t_1^j, \ldots, t_{n-1}^j, r_{t_1^j}, \ldots, r_{t_{n-1}^j}, s_{t_1^j}, \ldots, s_{t_{n-1}^j}$ as inputs and classify them into two opposite classes – c_1 or c_2. A structure of a feedforward neural network corresponding to a general website, which browsing activity is depicted in Figure 1, is presented further (see Figure 3).

The MLP model presented in Figure 3 consists of one hidden layer with a neuron. Input and hidden layers have a bias (denoted by a bigger red circle). MLP uses back-propagation learning algorithm and hyperbolic tangent transfer function. A total error network graph for the considered MLP model (see Figure 4) shows an ability of the neural network to perform classification experiments at the predefined error level.

5. Experiment: recommendations based on analysis of user online behaviour

An abstract website, which browsing activity diagram is presented in Figure 1, was used for an illustration of the proposed technique.

Let us consider the situation when a statistical database contains 30 entries and user online activities form the following new data entry – see Table 2.

Table 2. An example of a fragment of actual browsing activity by a user (record R_{31})

	e	y	u	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}	…
…	…	…		…	…	…	…	…	…	…	…	…	…	…	…
R_{31}	1	1	x	0	0	0	1	0	0	0	0	0	1	1	…

Source: own elaboration.

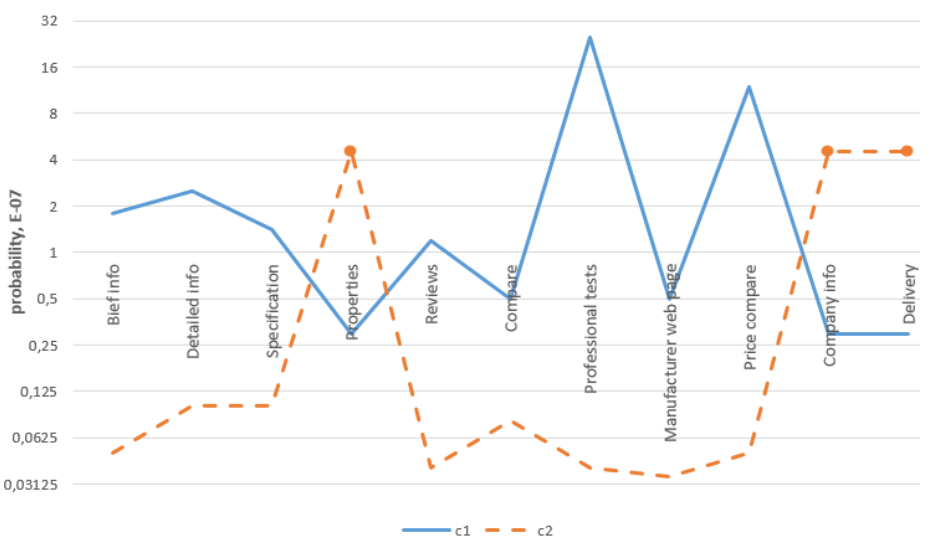

Figure 5. Results of experimental estimations of probabilities in order to forecast and recommend a web page with the highest impact on the transaction finalisation. Data series denoted by markers correspond to existing probability of the actual browsing activity

Source: own elaboration.

Naïve Bayes classifier estimates the probabilities for the class c_1 and c_2:

$$\text{classify}\left(e, y, t_1^{31}, \dots, t_{11}^{31}, r_{t_1^{31}}, \dots, r_{t_{11}^{31}}, s_{t_1^{31}}, \dots, s_{t_{11}^{31}}\right) =$$

$$= \underset{c \in \{c_1, c_2\}}{\operatorname{argmax}} p(C = c) \prod_{i=1}^{11} p\left(\langle e, y, u, t_i^{31}, r_{t_i^{31}}, s_{t_i^{31}} \rangle | C = c\right) = c_2$$

(0.29E-07 < 4.54E-07)

Next, the procedure is being activated that experimentally estimates a web page to offer a user a visit, which has a maximal impact on the transaction finalisation. Figure 5 presents results of experimental estimations of probabilities of class 1 (solid line) and class 2 (dashed line). Figure 5 vividly shows – website-to-user interaction will advise visiting Professional tests web page as it has a maximal impact on the transaction finalisation:

$$\text{classify}(1,1,\text{x},0,0,0,1,0,0,\mathbf{1},0,0,1,1,\dots) = c_1$$

(24.73E-07 > 0.04E-07)

Let us consider the situation when statistical database contains 7800 entries and user online activities form the following new data entry – see Table 3.

Results of estimations by Naïve Bayes classifier:

$$\text{classify}\left(e, y, t_1^{7801}, \dots, t_{11}^{7801}, r_{t_1^{7801}}, \dots, r_{t_{11}^{7801}}, s_{t_1^{7801}}, \dots, s_{t_{11}^{7801}}\right) =$$

$$= \underset{c \in \{c_1, c_2\}}{\operatorname{argmax}} p(C = c) \prod_{i=1}^{11} p\left(\langle e, y, u, t_i^{7801}, r_{t_i^{7801}}, s_{t_i^{7801}} \rangle | C = c\right)$$

$$= c_2$$

(4.61E-05 < 14.33E-05)

MLP model has classified given data as class 2 with a score 0.967.

Next, the procedure is being activated that experimentally estimates a web page to offer the user a visit, which has a maximal impact on the transaction finalisation. Figure 6 presents results of experimental estimations of data classification using MLP to class 1 (denoted as c1 (MLP)) or class 2 (denoted as c2 (MLP)). Several unvisited web pages were analysed – Brief info, Price compare and Company info. For comparative purposes, results of the experimental estimations using Naïve Bayes classification are presented too (classes 1 and 2 are denoted as c1 (NB) and c2 (NB) respectively). Figure 6 clearly shows that both classification methods work well and estimate the same outcome. A website-to-user interaction will advise visiting Price compare web page.

Table 3. An example of actual browsing activity by a user (record R_{7801})

	e	y	u	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}	…
…	…	…		…	…	…	…	…	…	…	…	…	…	…	…
R_{7801}	1	2	y	0	1	1	1	1	1	1	1	0	0	1	…

Source: own elaboration.

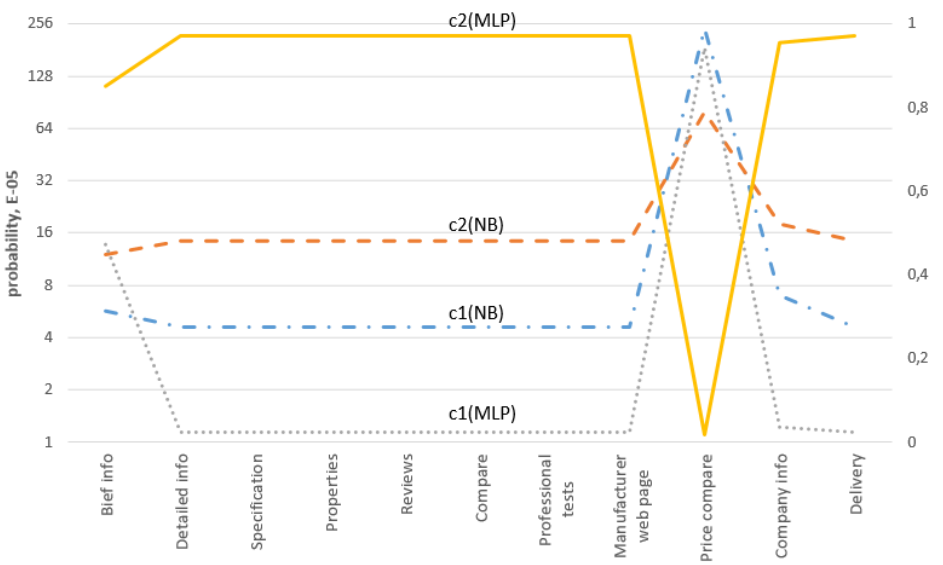

Figure 6. Results of experimental estimations of data classification using Multi-layer Perceptron and Naïve Bayes classifier

Source: own elaboration.

Note that Price compare functionality of a website is quite sensitive to any business. Moreover, as stated in Shopping Cart Abandonment report (Mulpuru, Hult and McGowan, 2010), 27% of e-customers cancel their purchases due to a comparison of prices from different retailers. According to the author's view, to overcome this issue, it is better to deal with this challenge on own website by applying the following policy. In order to offer a customer a product or a service at the lowest price, the mentioned Price compare functionality usually acts as follows. It offers an instant discount in case of an existence of a vendor that offers a lower price for the same product (if such a discount is possible with respect to a company price policy) or it includes vendors with poor customer ratings in the price compare list (if a discount cannot be applied). Obviously, giving a Price compare information, which is not profitable for a company − without any correction with regard to a website company − usually leads to a purchase cancellation, and should be avoided.

6. Conclusions

1. The proposed technique permits one to estimate website user actual on-site behaviour with respect to the transaction finalisation using Naïve Bayes classification and feedforward neural network – Multi-layer Perceptron model that are based on previous visitors' browsing activities.
2. The technique permits one to define actions to recommend a website user who theoretically has an impact on his/her decision to finalise a transaction.

Future works in this direction include widening the application of the technique to other areas as well as deepening the technique by applying other methods of multivariate analysis.

REFERENCES

ANGELETOU, S., ROWE, M., ALANI, H., (2011). Modelling and Analysis of User Behaviour in Online Communities. The Semantic Web – ISWC 2011 (pp. 35−50). Lecture Notes in Computer Science Volume 7031.

BUDNIKAS, G., (2015). Creation of user online behaviour analysis model for increase of an enterprise competitiveness. Rzeszów: In proceedings of VI Ogólnopolska Konferencja Naukowa „Społeczeństwo Informacyjne. Stan i kierunki rozwoju w świetle uwarunkowań regionalnych" (in press).

CLIFTON, B., (2012). Advanced Web Metrics with Google Analytics (3rd Edition ed.). Indianapolis: John Wiley & Sons.

DEMBCZYŃSKI, K., KOTŁOWSKI, W., SYDOW, M., (2009). Effective Prediction of Web User Behaviour with User-Level Models. Journal Fundamenta Informaticae, 89(2−3), 189−206.

DREJEWICZ, S., (2012). Zrozumieć BPMN. Modelowanie procesów biznesowych. Helion.

MULPURU, S., HULT, P., MCGOWAN, B., (2010, May 20). Understanding Shopping Cart Abandonment. Retrieved June 25, 2015, from https://www.forrester.com/Understanding+Shopping+Cart+Abandonment/fulltext/-/E-RES56827

NIKIFORAKIS, N., ACAR, G., SAELINGER, D., (2014). Browse at your own risk. Spectrum, IEEE, 51(8), 30−35.

ROBINSON, D. J. B. V., (2008). Online Behavioural Analysis and Modeling Methodology (OBAMM). Social Computing, Behavioural Modeling, and Prediction, 100−109.

RUSSELL, S. A., (2010). Artificial Intelligence: International Version: A Modern Approach (3 ed.). Pearson.

WHITE, R. W., CHU, W., HASSAN, A., HE, X., SONG, Y., WANG, H., (2013). Enhancing personalized search by mining and modeling task behavior. Proceedings of the 22nd International Conference on World Wide Web (pp. 1411–1420). ACM.

XIAN, X., CHEN, F., WANG, J., (2014). An Insight into Campus Network User Behavior Analysis Decision System. (pp. 537−540). Taichung: IEEE.