CREDIT SCORING BASED ON SOCIAL NETWORK DATA

Alexey A. MASYUTIN
Post-graduate student, School of Data Analysis and Artificial Intelligence,
Faculty of Computer Science, National Research University Higher School of Economics

Address: 20, Myasnitskaya Street, Moscow, 101000, Russian Federation
E-mail: alexey.masyutin@gmail.com

Social networks accumulate huge amounts of information, which can provide valuable insights on people's behavior. In this paper, we use social data from Vkontakte, Russia's most popular social network, to discriminate between the solvent and delinquent debtors of credit organizations. Firstly, we present the datacenter architecture for social data retrieval. It has several functions, such as client matching, user profile parsing, API communication and data storing. Secondly, we develop two credit scorecards based exclusively on social data. The first scorecard uses the classical default definition: 90 days delinquency within 12 months since the loan origination. The second scorecard uses the classical fraud definition as falling into default within the first 3 months. Both scorecards undertake WOE-transformation of the input data and run logistic regression afterwards. The findings are as follows: social data better predict fraudulent cases rather than ordinary defaults, social data may be used to enrich the classical application scorecards. The performance of the scorecards is at the acceptable level, even though the input data used were exclusively from the social network. As soon as credit history (which usually serves as input data in the classical scorecards) is not rich enough for young clients, we find that the social data can bring value to the scoring systems performance. The paper is in the area of interest of banks and microfinance organizations.

Key words: credit scoring, social networks, probability of default, social data, Vkontakte.

Citation: Masyutin A.A. (2015) Credit scoring based on social network data. *Business Informatics*, no. 3 (33), pp. 15–23.

1. Introduction

Social networks are an inexhaustible source of information about people. According to its financial filings at the time of its IPO which took place in 2012, Facebook stored around 111 megabytes of photos and videos for each of its users [1], whose number now exceeds a billion. That adds up to 100 petabytes of personal information.

The data are mostly not structured and chaotic by type, but they allow one to get to know their clients much deeper. Properly processed and then structured information about users brings value to online stores, recruitment agencies, banks and many other business-to-client and business-to-business companies. E-commerce is able to learn more about the behavior and preferences of thecustomers, and banks can determine the credit rating of the borrowermore accurately.

For example, when assessing a person's income, they can use information about the places they visit, as well as the countries theytravelto. The places and countries are classified according to price category. Further formed scorecards are lists of places and countries with ratios of their influence on the level of creditworthiness. For each country, the scorecard will be unique.

However, there are several issues concerning social data implementation. First, the companies have to construct solid business processes, which would adopt data-driven decision-making, before they can extract value from the flow of social data. The cornerstone of such processes is the data retrieval mechanism and architecture of data storage. Here companies face a dilemma: to build up the database within local IT department, or to outsource the social data retrieval to the third parties (SaaS). The latter case is very similar to credit history bureau query, when banks request information on the applicants.

The second issue is related to the legal aspects of personal data processing. Obviously, when dealing with personal data from social networks,companies should learn to do it legally, so as not to be subject to prosecution. The laws vary from country to country and can dramatically influence the ability to use the social data.

Thirdly, the company must have enough competencies to perform the data analysis. The techniques are versatile machine learning algorithms, ranging from logistic regression to sophisticated data mining techniques. Again, the company can outsource this process as soon as there are many consulting analytics (e.g. SAS).

We have to mention that there is not much academic literature covering the use of social data in credit scoring [5]. However, there is a number of papers analyzing the social networking from the sociological point of view [6]. There were also several attempts to bind the activities within social networks (such as Twitter) with movements in stock prices [8]. The scarcity of academic research is explained by the fact that the use of social data in scoring started no more than 3-4 years ago. Moreover, the accumulation of particular results concerning the use of social data takes place in the business area, rather than academic area. The companies that launch such «social data» projects aren't likely to focus on derivation of universal laws or theoretical results. Besides, the companies are running their activities in a highly competitive environment and are not interested in sharing the knowledge at instance, whichis actually implied by the academic style of research. At least, we can list some companies whose domain is social data retrieval, aggregation and customer analytics for credit organizations: Wonga (USA), Kreditech (Germany), Big Data Scoring (Estonia, Finland), Lenddo (Philippines, Columbia), SOCSCOR (Russia), Crediograph (Ukraine).

In our analysis we will examine the predictive power of the social data from Vkontakte social network (also VK). It is the number one in Russia by visitors per month. Its monthly audience makes a total of over 50 million users.

This paper consists of four parts. The first one is introduction. The second one describes the typical architecture for social data retrieval and integration with bank databases. It will be based on an example of one Russian bank whose name we cannot disclose due to the confidential reasons. The third part involves data description used for default prediction, discusses the variables relevant to the default event and considers prediction accuracy. The fourth part is conclusion.

2. Social Data Retrieval and Storage

2.1. Social network objects

First of all, we define the set of notions we use when talking about social data. A profile is a set of properties describing a user of a social networking service. It can include their name, patronymic name, last name, date of birth, place of residence, work and study, interests, communities, friends and feed.

A feed is a part of a user's profile which contains events describing the user: messages, what they are fond of ('likes'), communities, applications installed etc.

Open data are a part of a user's profile, which is accessible without loggingin and can be retrieved automatically.

Accessible data are, meanwhile, another part of a user's profile, which can be accessed from other users' accounts and which can be retrieved automatically on a regular basis using someone's else account. It does not require other users' personal involvement and there is no limit on the number of pattern messages sent from this user's account. In addition to this, accessible data should be structured, e.g. the feed should contain XML or JSONmarkings.

Parsing is the process of matching a linear sequence of lexemes (words, tokens) of a natural or formal language with its formal grammar.

2.2. Retrieval process and datacenter architecture

The basic tool for data retrieval is the so-called API (application programming interface). In effect, it is a set of ready-to-use classes, procedures, functions, structures and constants offered by an application (library, service) to be used in outer software products. A social network user must allow the access to their personal data within the social network. This usually happens when the user installs an API application, it can be a game or a discount offer. The application is granted rights to perform API requests to the user profile and activities at any time.

The architecture of the solution is shown below in the *Fig. 1.*

MDM stands for Master Data Management system, a core banking information system. Apparently, the credit applicant must be identified with Vkontakte user. The identification is carried out by a combination of parameters: first and last name, date of birth, current city, city of origination, e-mail, or exact social network id (if the applicant gives such information via the application).

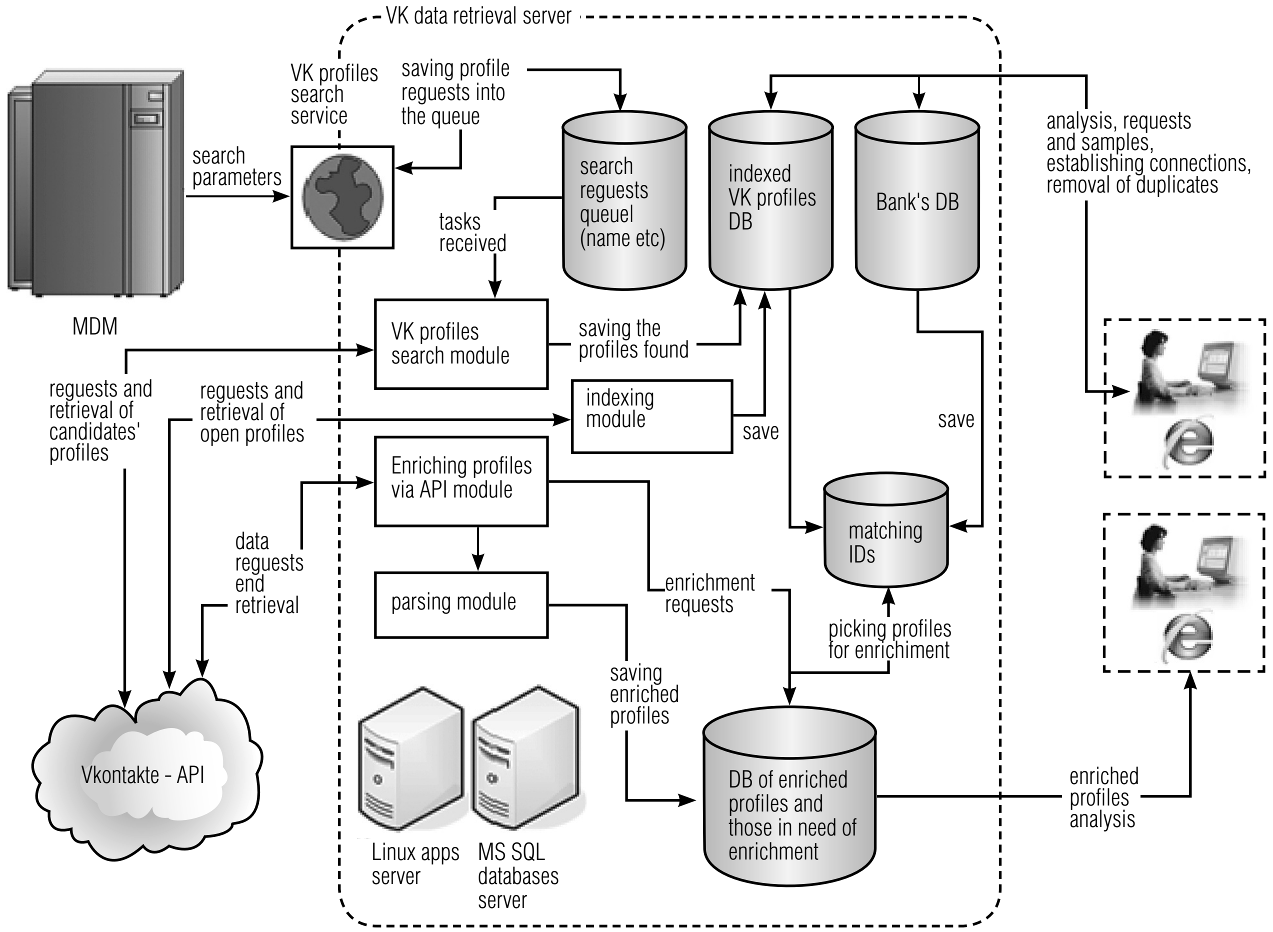
VK data retrieval server
VK profiles search service
saving profile reguests into the queue
search parameters
MDM
search reguests queuel (name etc)
indexed VK profiles DB
Bank's DB
tasks received
analysis, requests and samples, establishing connections, removal of duplicates
VK profiles search module
saving the profiles found
reguests and retrieval of candidates' profiles
reguests and retrieval of open profiles
indexing module
save
save
Enriching profiles via API module
matching IDs
data reguests end retrieval
parsing module
enrichment requests
picking profiles for enrichiment
saving enriched profiles
Vkontakte - API
DB of enriched profiles and those in need of enrichment
enriched profiles analysis
Linux apps server
MS SQL databases server

Fig. 1. The datacenter architecture for social data retrieval and storage

The architecure of the datacenter is aimed to accomplish the following tasks:

✦ Downloading all Vkontakte profiles,

✦ Appending newly created profiles on a regular basis,

✦ Downloading detailed data (subscriptions, communities, friends, wallposts) of Vkontakte profiles with the span of 3 years,

✦ Regular downloading detailed data of VK profiles (subscriptions, communities, friends, new posts on the wall since the last upload).

Additionaly, it provides an interface to browse, search and compare the downloaded data to clients' data from the core banking system. Downloading and updating is performed in astreaming mode according to the chosen set of profiles, the updating period can vary from several days to several weeks, i.e. the downloading is not performed online.

The solution has following volume indicators:

✧ Downloading and storage of all Vkontakte profiles for indexing purposes (about 250 million),

✧ Downloading detailed data on 300,000 Vkontakte profiles.

The system is able to adjust to scale for downloading a 5 times larger number of profiles without significant changes in its architecture and functions.

3. Data Description and Default Prediction

The event of default in retail banking is defined as more than 90 days of delinquency within the first 12 months after the loan origination. Defaults are divided into fraudulent cases and ordinary defaults. The default is told to be a fraudulent case when delinquency starts at one of the three first months. It means that when submitting a credit application, the borrower did not even intend to payback. Otherwise, the default is ordinary when the delinquency starts after the first three months on book.

That is why scorecards are usually divided into fraud and application scorecards. In fact the only difference is the target variable definition, while the sets of predictors and the data mining techniques remain the same. The default cases are said to be «bad», and the non-default cases are said to be «good».

We consider a dataset of 27540 microfinance loans, originated in 2012. This segment is characterized by small-cash high-margin loans with short credit term (3-6 months). The further details of the data source are not presented due to the non-disclosure agreement. We develop both scorecards and examine their accuracy via out-of-sample validation. The validation process requires calculation of performance metrics (ROC-curve and Gini coefficient) of the model based on the data sample that was retrieved from the same parent population but was not used to develop the model itself (validation set). This approach allows the user to check for accuracy and stability of the model. The size of the validation set we used was 30% from the parent population.

The analytics was carried out using SAS Enterprise Miner, the most spread analytical software in the banking sphere.

The mathematical architecture of the scorecard is based on a logistic regression, which takes the transformed variables as input. The transformation of the initial variables we use is WOE-transformation [3]. It is wide-spread in credit scoring, to apply such a transformation to the input variables as soon as it accounts for non-linear dependencies and provides certain robustness coping with potential outliers. The aim of the transformation is to divide each variable into no more than k categories. At step 0, all the continuous variables are binned into 20 quantiles, the nominal and ordinal variables are left as they are. Now, when all the variables are categorized, we compute the odds ratio for each category. Then for each predictor variable X_i $(i = 1 \ldots n)$ we merge non-significant (chi-square statistics based on differences in odds) categories.

1. If X_i has 1 category only, stop and set the adjusted p-value to be 1.

2. If X_i has k categories, go to step 7.

3. Else, find the allowable pair of categories of X_i (an allowable pair of categories for ordinal predictor is two adjacent categories, and for nominal predictor is any two categories) that is least significantly different (i.e., most similar). The most similar pair is the pair whose test statistic gives the largest p-value with respect to the dependent variable Y.

4. For the pair having the largest p-value, check if its p-value is larger than a user-specified alpha-level α merge. If it does, this pair is merged into a single compound category. Then a new set of categories of X_i is formed. If it does not, then if the number of categories is less or equal to k go to step 6, else merge two categories with highest p-value.

5. Go to step 2.

6. (Optional) Any category having too few observations (as compared with a user-specified minimum segment size) is merged with the most similar other category as measured by the largest of the p-values.

7. The adjusted p-value is computed for the merged categories by applying Bonferroni adjustments [4].

Having accomplished the abovementioned steps, we acquire categorized variables instead of the continuous ones. When each variable X_i $(i = 1 \ldots n)$ is finally binned into a certain number of categories (k_i), we are able to calculate the odds for each category j ($j = 1 \ldots k_i$), the weight of evidence for each category and, as a result, information value for each variable:

$$odds_{ij} = \frac{\% goods_{ij}}{\% bads_{ij}}$$

$$WOE_{ij} = ln(odds_{ij}) \qquad (1)$$

$$IV_i = \sum_{j=1}^{k_i} (\% goods_{ij} - \% bads_{ij}) \cdot WOE_{ij} .$$

Information value can also be used in feature selection (e.g. variables with information value less than 0.02 are often rejected, as soon as they do not reveal predictive power). More details on WOE-transformation are provided in [7].

The role of the WOE-transformation is that, instead of initial variables, the logistic regression receives WOE as input. So, each input variable is a discrete transformed variable, which takes values of WOE. When estimating the logistic regression, the usual maximum likelihood was applied.

3.1. Ordinary default scorecard

A scorecard is a set of rules. Each rule gives a number of scorepoints to the applicant. Then the scorepointsare summed up to form the final score. The higher the score, the less the probability of default. The variables were selected from a set of about 100 calculated parameters. The criteria for includinga variable into the scorecard was Information value no less than 0.2 and correlation analysis within variables themselves (so as to rule out multi-collinearity).

Table 1.

Ordinary Default

VAR CODE	ORDINARY DEFAULT SCORECARD		
	Variable name	**Value range**	**Scorepoint**
Grouped Levels for x_{11}	Marital status	'In love', 'Engaged'	11
		'Single'	19
		Missing	23
		'Married'	27
Grouped Levels for x_{69}	Political views	'Communistic', 'Monarchic', 'Socialistic'	16
		'Indifferent', 'Liberal', Missing	24
Age	Age (in years)	age< 25	5
		25<= age< 28	12
		28<= age< 37	21
		37<= age< 52	44
		52<= age	66
Sex	Sex	Female	27
		Male	15
x_1	Number of days since last visit	x1< 1	26
		1<= x1< 3	30
		3<= x1< 37	20
		37<= x1< 149	17
		149<= x1	11
x_{30}	Number of days since the first post	x30< 265	20
		265<= x30< 399	15
		399<= x30< 1330.5	23
		1330.5<= x30	36
x_{76}	Number of job places	0, Missing	21
		>=1	29

We use the ROC analysis to evaluate the performance of the scorecard. The ROC analysis provides the set of feasible accuracy measures, which form an ROC curve and reflect the accuracy of the classifier. In terms of classification, the scorecard labels the applicant as positive (high probability of default) when its score is less than the cutoff (decision boundary). The accuracy is defined by the probability of the two possible errors. The first one is to approve the loan for a bad applicant (false negative), and the second one is to reject a good applicant (false positive). This gives us the two dimensions for the ROC curve: sensivity (vertical axis) and one minus specificity (horizontal axis):

$$Sensivity = \frac{TruePositive}{Positive}; \qquad (2)$$

$$Specificity = \frac{TrueNegative}{Negative} \Rightarrow$$

$$\Rightarrow 1 - Specificity = \frac{FalseNegative}{Negative}. \qquad (3)$$

The ROC curves for train and validation set are presented below in *Figures 2* and *3*:

Fraudulent case scorecard can be found below in *Table 2.*

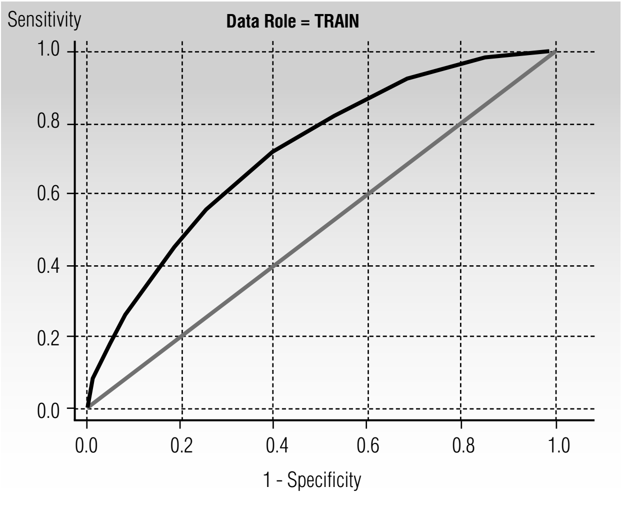

Fig. 2. ROC-curve for ordinary default (training)

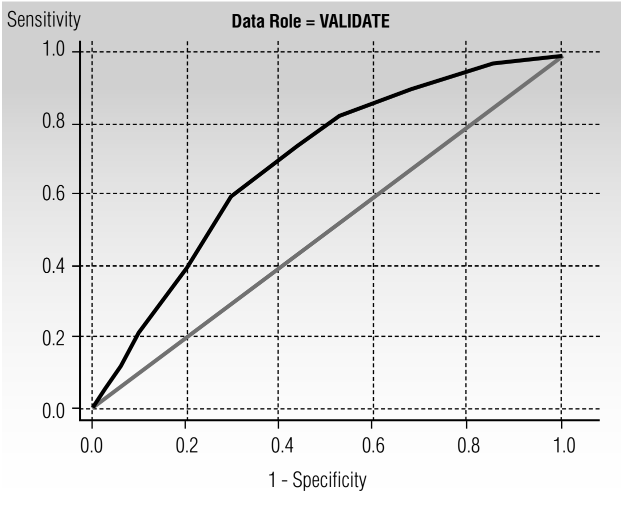

Fig. 3. ROC-curve for ordinary default (validation)

Table 2.

Fraudulent Case

VAR CODE	FRAUDULENT CASE SCORECARD		
	Variable name	Value range	Scorepoint
Age	Age (in years)	age< 29	3
		29<= age< 33	11
		33<= age< 39	25
		39<= age< 47	42
		47<= age	62
Sex	Sex	Male	4
		Female	26
x_1	Number of days since last visit	x1< 2	21
		2<= x1< 3	0
		3<= x1< 10	22
		10<= x1< 1565	10
		1565<= x1	-4
x_{11}	Marital status	'In love', 'All is difficult', 'In couple'	4
		'Single'	11
		'Engaged'	14
		'Married'	24
x_{21}	Number of subscriptions	x21< 2, _MISSING_	19
		2<= x21< 3	17
		3<= x21< 7	20
		7<= x21< 16	11
		16<= x21	4
		1330.5<= x30	8
x_{30}	Number of days since the first post	x30< 497	26
		497<= x30< 710	21
		710<= x30< 1324	14
		1324<= x30	13
x_{39}	Number of user's posts with photos	x39< 1	13
		1<= x39< 2	18
		2<= x39< 4	15
		4<= x39< 44	11
		44<= x39	19
x_{41}	Number of user's posts with video	x41< 1, Missing	15
		1<= x41< 2	17
		2<= x41< 3	13
		3<= x41< 16	12
		16<= x41	5
x_{51}	Number of children	x51< 1, Missing	12
		1<= x51< 2	13
		2<= x51	15
x_{65}	Major things in life	'Career and Money', 'Entertainment', 'Fame and Influence'	2
		'Beauty and Art', 'Research and Science', Missing	16
x_{66}	Major qualities in people	'Kindness and Honesty', 'Humor and Lust', 'Health and Beauty'	19
		'Courage and Perseverance', 'Mind and Creativity'	14

The corresponding ROC analysis can be found below in *Figures 4* and *5*.

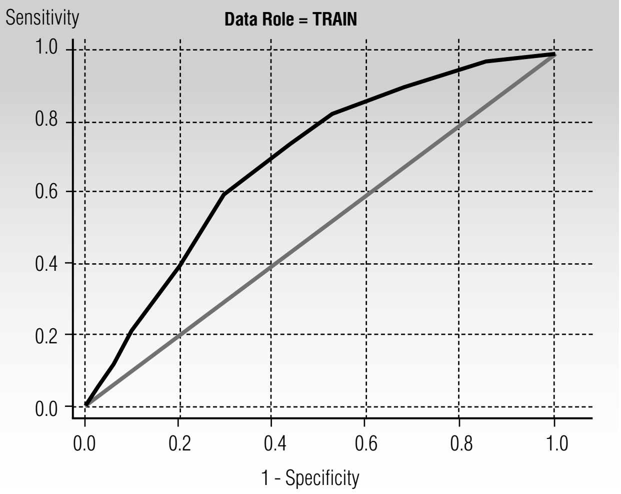

Fig. 4. ROC-curve for fraudulent cases (training)

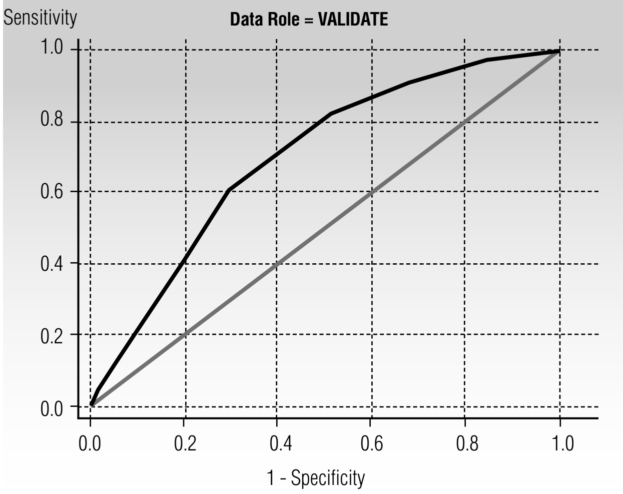

Fig. 5. ROC-curve for fraudulent cases (validation)

The ROC analysis show that both models are of mild predictive power and of high stability. Gini coefficients are 36% (32%) for train (validation) sample for ordinary default scorecard. The corresponding Gini coefficients for fraudulent case scorecard are 47% (39%). Of course, the main contribution was made by Age and Sex variables, which are known even without social network. However, if other variables are excluded, the Gini coefficients fall by 7-8%.

From our point of view, the result that fraudulent case scorecard showed better performance can be explained in the following way. When the borrower initially intends to receive a loan and not to pay any installments right from the start, of course, he or she tries to disclose less information in the social network. Missing values of predictors' parameters tend to receive lower scorepoints. As far as the interests in life are concerned, the person who is obsessed with 'Entertainment' or 'Fame' is, to our mind, more likely to try to get easy money by receiving a loan with no intent to pay it back.

If we compare the Gini coefficients of our models to the ones of the classical application scoring models (usually 45-55%, see [2]), we find that the latter are more accurate. That brings the understanding that the social data indeed brings value and increases discriminative power,however, it should not be used standalone. The social data should enrich the usual set of application variables and contribute to the higher discriminative power of banking scoring. Such performance metrics as Gini coefficient are not the only way to estimate the efficiency of social data implementation. In fact, the most important criterion is the profit increase due to the model implementation. However, the final effect depends on the particular bank or microfinance organization that is considering the possibility to use social data. Basically, a 7-8%Gini increase may lead to 0.5-1% decrease in default frequency, which is the key performance indicator of the risk management system in the bank. That is why we have to understand when the project concerning social data implementation is going to bring payoff, and when it will become just a burden. At the moment, we see two situations when the credit scoring based on social data is applicable. The first one is the microfinance segment. The application flow is traditionally very risky in this segment, and classical scorecards tend to reject all the applications. However, many microfinance organizations compensate that risk with extraordinarily high rates. That is why the companies still have to obtain additional information to provide at least any discrimination between «bad» and, say, «very bad» applicants. Moreover, given that young people tend to be riskier and they do not have sufficient credit history, the social data starts being the only data source except for the loan application form. The second situation that can be favorable for social data implementation is when the bank has a large client base and huge amount loans in its portfolio. The reasoning is quite straightforward. Given that fixed costs of social data center launching and its support are high, the decrease in default frequency will be sufficient to compensate those costs only when the loan portfolio volume is large.

The real life examples of successful implementations, from our point of view, are still to come, as soon as many banks and microfinance organizations start pilot social dataprojects. As far as Russia is concerned, there are several consultant agencies such as Digital Society Laboratory, SOCSCOR, Double Data, SAS, SocioHub etc. that provide scoring service for banks and the microfinance segment. However, the details of such projects are still rarely disclosed, and the discussions that take place within professional banking conferences, as a rule, are held in a more general way.

4. Conclusion

In this paper we described the schema of social data retrieval in banking sphere. The social network we considered is Vkontakte, which is the largest in Russia. Then we examined the value that the data can bring to the credit scoring. We developed two scorecards (fraudulent case and ordinary default) based solely on the social data. The findings are the social data can indeed show acceptable discriminative power, especially in the case of fraud scoring.

As an area for further research, one can test whether there is an increase in Gini coefficient, when the social data predictors are added to the bureau of credit history scoring (e.g. Equifax) with the information on behavioral variables of the applicant. ■

References

1. Tucker P. (2013) *Has Big Data made anonymity impossible?* MIT Technology Review. Available at: http://www.technologyreview.com/news/514351/has-big-data-made-anonymity-impossible/ (accessed 5 November 2014).

2. Thomas L., Edelman D., Crook J. (2002) Credit scoring and its applications. *Monographs on Mathematical Modeling and Computation*, SIAM: Pliladelphia, pp. 107–117.
3. Wu J., Coggeshall St. (2012) *Foundations of predictive analytics*, CRC Press.
4. Hochberg Y. (1988) A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, vol. 75, no. 4, pp. 800–802.
5. Skiba S.A., Loiko V.I. (2013) Social'nyj skoring [Socialscoring]. *Scientific Journal of KubGAU*, no.7 (91), pp. 1258–1275. Available at: http://ej.kubagro.ru/2013/07/pdf/89.pdf(accessed 05 November 2014). (in Russian)
6. D'Andrea, Alessia et al. (2009) An overview of methods for virtual social network analysis. *Computational Social Network Analysis: Trends, Tools and Research Advances*, Springer, pp. 3–25.
7. SAS Institute Inc. (2012).*Developing credit scorecards using credit scoring for SAS® Enterprise Miner*™ 12.1. Cary, NC: SAS Institute Inc.
8. Porshnev A., Redkin I. (2014) Analysis of Twitter users' mood for prediction of gold and silver prices in the stock market. *Communications in Computer and Information Science*, no. 436, pp. 190–197.

КРЕДИТНЫЙ СКОРИНГ НА ОСНОВЕ ДАННЫХ СОЦИАЛЬНЫХ СЕТЕЙ

А.А. МАСЮТИН

аспирант, департамент анализа данных и искусственного интеллекта,
факультет компьютерных наук, Национальный исследовательский
университет «Высшая школа экономики»

Адрес: 101000, г. Москва, ул. Мясницкая, д. 20
E-mail: alexey.masyutin@gmail.com

Социальные сети аккумулируют значительное количество информации, которая позволяет получать дополнительные сведения о поведении людей. В данной работе мы используем данные наиболее посещаемой социальной сети «Вконтакте», чтобы выделять сегменты неплатежеспособных клиентов банка. Во-первых, мы представляем архитектуру центра хранения и обработки данных из социальных сетей. Он включает в себя инструменты для соотнесения реального клиента и его виртуального профиля в социальной сети, парсинг профилей социальной сети, получение данных об активности пользователя через API, и наконец, само хранилище данных. Во-вторых, на исторических данных мы разрабатываем две скоринговые карты, основанные исключительно на данных активности клиента в социальных сетях. Первая карта прогнозирует событие обычного дефолта – выхода на просрочку по ссуде более 90 дней за первые 12 месяцев с момента получения кредита. Вторая скоринговая карта прогнозирует событие мошеннического дефолта. Обе карты используют WOE-трансформацию входящих данных и затемприменяют логистическую регрессию по преобразованным данным. В результате данные социальных сетей лучше прогнозируют случаи мошеннических дефолтов, в отличие от обычных случаев просрочки. Качество скоринговых карт находится на приемлемом уровне, что подтверждается ROC-анализом и коэффициентами Джини. Поскольку классические скоринговые системы во многом опираются на кредитную историю клиента, которая зачастую отсутствует у молодых заемщиков, мы считаем, что данные социальных сетей могут служить их заменой. Таким образом, данные социальных сетей могут быть использованы для обогащения классических скоринговых систем банков и микрофинансовых организаций.

Ключевые слова: кредитный скоринг, социальные сети, вероятность дефолта, социальные данные, Вконтакте.

Цитирование: Masyutin A.A. Credit scoring based on social network data // Business Informatics. 2015.
No. 3 (33). P. 15–23.

Литература

1. Tucker P. Has Big Data made anonymity impossible? MITTechnologyReview, 2003. [Электронный ресурс]: http://www.technologyreview.com/news/514351/has-big-data-made-anonymity-impossible/ (дата обращения 05.11.2014).
2. Thomas L., Edelman D., Crook J. Credit scoring and its applications / Monographs on Mathematical Modeling and Computation. SIAM: Pliladelphia, 2002. P. 107–117.
3. Wu J., Coggeshall St. Foundations of predictive analytics, CRC Press, 2012.
4. Hochberg Y.A sharper bonferroni procedure for multiple tests of significance // Biometrika. 1988. Vol.75, No. 4. P. 800–802.
5. Скиба С.А., Лойко В.И. Социальный скоринг// Научный журнал КубГАУ. 2013. № 7 (91). С. 1258–1275) [Электронный ресурс]: http://ej.kubagro.ru/2013/07/pdf/89.pdf (дата обращения 05.11.2014).
6. D'Andrea, Alessia et al.An overview of methods for virtual social network analysis // Computational Social Network Analysis: Trends, Tools and Research Advances. Springer, 2009. P. 3–25.
7. Developing credit scorecards using credit scoring for SAS® Enterprise Miner™ 12.1. Cary, NC: SAS Institute Inc., 2012.
8. Porshnev A., Redkin I. Analysis of Twitter users' mood for prediction of gold and silver prices in the stock market // Communications in Computer and Information Science. 2014. No. 436. P. 190–197.