Bank Licenses Revocation Modeling

Jaroslav Bologov, Konstantin Kotik, Alexander Andreev, and Alexey Kozionov

Deloitte Analytics Institute, ZAO Deloitte & Touche CIS, Moscow, Russia
`{jbologov,kkotik,aandreev,akozionov}@deloitte.ru`

Abstract. This paper is devoted to developing the models of losing bank licenses by Russian banks. The factors of the models are drawn from banks' financial statements and macroeconomic reports. The algorithms proposed are capable to estimate both the probability and the exact time of license revocation. In order to do so multiple choice problem is formulated with the target variable represented the probabilities of revocation within a certain time period after the forecast date. The modeling was conducted using logistic regression model, ensemble of decision trees, gradient boosting and artificial neural network.
The results of this study have useful implications both in government organizations and in private companies. The regulators can adjust manageable macroeconomic indicators to control the intensity of bank licenses revocation. Companies can use estimated probabilities in solving funds distribution problems.

Keywords: Bank licenses revocation · Russian banks · Probability of default · Multi-target classification

1 Introduction

Currently Russian banking sector is becoming more concentrated. Table 1 illustrates the dynamic of a Herfindahl-Hirschman index calculated using the amounts of loans granted[1].

This process is supported by numerous cases of bank licenses revocations. Eighty-seven licenses were revoked in 2014, ninety-three—in 2015 and one hundred—in 2016[2]. That corresponds to an average frequency of two bank licenses revoked per week. In these circumstances the identification of the reasons, which lead to license revocation, and estimation of the revocation probabilities for Russian banks become problems of high importance.

The results of solving these problems may be used by the regulators in finding out candidates for revocation and in adjusting macroeconomic indicators they can manage in order to control the intensity of revocation process. On the other hand, banks provide deposit services for many commercial organizations and individuals, so license revocation leads to credit loss for them. Thus, solid probability estimates of this kind of risk plays an essential role in providing sustainability of their operating activities.

---

[1] Data is provided by `http://www.banki.ru/`.
[2] Data is provided by `http://kuap.ru/revoke/`.

Table 1. Mean Herfindahl-Hirschman index value of Russian banking sector

Year	2008	2009	2010	2011	2012	2013	2014	2015	2016
HHI, %	11.0	11.9	11.4	11.5	12.6	12.2	13.4	13.9	14.4

Due to its high relevance license revocation modeling (that is often called "banks' defaults modeling") had become a subject of many researches, the most notable of which are [1], [2], and [3]. First two papers are devoted to determining the factors which drive the revocation process. The purpose of the last one is to get accurate estimates of revocation probabilities that is quite similar to this paper goals. Current research extends approaches proposed in these articles aiming to increase the predictive performance of models.

Another goal is to show that data from public sources can be used to obtain solid revocation probabilities estimates.

2 Data acquisition and preprocessing

The dataset used in modeling covers a period from February, 2011 to March, 2017 containing 18812 monthly observations, which correspond to 746 distinct banks. Each observation contains set of financial indicators values of a given bank as of given period of time including capital adequacy ratio (also known as N1), instant liquidity ratio (N2), current liquidity ratio (N3), credit and deposit portfolios amounts, assets and equity volumes, ratings issued by Moody's, Standard & Poors, Fitch and Expert RA agencies, and set of macroeconomic indicators values, which are identical for all observations corresponding to the same period of time.

Set of banks' financial indicators was obtained by combining the data from two sources: website "banki.ru"[3] and the "CreditOrgInfo" web-service provided by Central Bank of Russia[4].

Banks' financial data contains many missing values. In particular, it is an acute problem for N1, N2 and N3 features. So if one decided to remove all the observations with missing values from the dataset one would end up with few percents of the initial number of observations left. This makes clear that the problem of missing values must be handled. In this research linear interpolation is used for filling these values, as it is a generally accepted robust method of time series preprocessing.

Target variable definition

The key aspect of proposed algorithms that allows for the revocation period estimating is the target variable. It was defined in a following way. For the banks which license were revoked during the modeling period observations were labeled

---

[3] http://www.banki.ru/banks/ratings/?LANG=en

[4] http://www.cbr.ru/CreditInfoWebServ/CreditOrgInfo.asmx

- with the class "4", if the difference in months between the revocation date and the observation date is not greater than 3,
- with the class "3", if the difference in months between the revocation date and the observation date is in the interval (3, 6],
- with the class "2", if the difference in months between the revocation date and the observation date is in the interval (6, 12],
- with the class "1", if the difference in months between the revocation date and the observation date is in the interval (12, 24],
- with the class "0", if the difference in months between the revocation date and the observation date is greater than 24.

For the banks which licenses were not revoked during the modeling period observations were labeled only with marks "0" when class "0" condition (the last of specified above) was met. Observations for which the condition was not met were deleted from the set, because in that case it is not known whether license would be revoked in 24 months or not.

Splitting all the observations into 5 classes is one of the many ways how banks' licenses revocation modeling may be turned into multi-target classification problem. The decision on how many classes a sample should be divided into has to come from the specific problem that is needed to be solved. For example, if one need to estimate short-term, mid-term and long-term risk then 3-class problem would be probably the most suitable kind of classification model with the time intervals in class assignment conditions determined by what one considers to be long, short and mid terms. General pattern of changing quantity of classes is that the more classes are in the model, the less accurate probability estimates this model produces. So, one may balance between high resolution of time intervals provided by large number of classes and high precision of revocation predictions on the other hand.

One may use two approaches to determine the revocation date. The first one is to look at the date stamp of the last observation that corresponds to a certain bank. If the value of this date is equal to the maximum date in entire data set then it means that license of this certain bank was not revoked during the modeling period. Otherwise, the date of revocation may be determined as the next period after the date stamp. The second approach is to follow press-releases issued by Central Bank of Russia, which contain information about revocation date and the reasons of revocation itself.

Both approaches have their advantages and weaknesses. Latter does not take into account that license revocation may occur several month after the bank is declared insolvent. During this period bank in being run by interim managers (usually appointed by Deposit Insurance Agency) who put bank's operations on hold and generally do not publish regular financial reports. In that case date stamp comparison method will treat insolvency declaration date as the revocation date, and this is the right way for economic reasons because from this date onwards bank becomes nonfunctional. On the other hand, the former method confuses merge and acquisition processes (which are always accompanied

by purchased bank's license revocation) with the real cases of banks' defaults which are the subject of this paper.

For the purpose of modeling the second approach was chosen because M&A cases are more frequent than divergence between revocation and insolvency declaration dates in the dataset.

Counts of target class labels are shown in the Table 2.

Table 2. Class labels distribution

Class label	"0"	"1"	"2"	"3"	"4"
Frequency	14496	1834	1183	639	660

The distribution of class labels is biased towards the "early" classes, which correspond to long periods until the revocation. It is an implication of the fact that most of the banks represented in the data set did not lose their licenses during the modeling period, and corresponding observations were labeled with class "0" or removed from the dataset according to the rules described earlier. For each of other banks only 3 observations were labeled with the classes "4" and "3", 6 and 12 observations were labeled with the classes "2" and "1" respectively[5], the rest were labeled with "0".

Adjusting for this bias and other preprocessing methods are discussed in the following subsection.

Data deskewing and augmentation

To make class distribution more uniform a dropout procedure was performed on observations with class "0", i. e. a certain fraction of these observations was randomly removed from the dataset. As the result of this procedure new dataset is formed with less skewed classes depending on the specified fraction.

As the dataset contains more than 18 thousands observations and list of predictive indicators (both banks' financial indicators and macroeconomic features) includes several dozens of factors, it is possible to extend the number of features without fear of deteriorating the quality of model parameters estimates. Thus, the further prepossessing was a data augmentation. New features representing pair-wise ratios of financial indicators and increments of macro-factors were added to predictors set in order to improve the quality of models predictions.

---

[5] The number of labeled observations is determined by the length of classes' time intervals.

3   Models and evaluation

Problem definition

The problem of predicting banks' failures was set in the following way. Probability of license revocation is calculated as a result of preforming the multi-target classification with the target variable representing the class label and predictors representing banks' financial indicators and macroeconomic factors. Thus, the main step in accomplishing this problem is developing of algorithm that could transform historical values of predictors into actual class labels.

In order to test algorithm's predictions, the data was randomly split into train and validation sets[6]. Former was used in fitting the models, latter—in evaluating its outcomes. The result of model fitting is a matrix of probabilities with the number of rows equals to the number of observations in validation set and the number of columns equals to the number of class labels.

The probability of observation belonging to class "4" is interpreted as the probability of license revocation within 3 month after the observation date. Similarly, class "3", "2" and "1" probabilities are interpreted as the probability of license revocation within the intervals from 3 to 6, from 6 to 12 and from 12 to 24 months respectively. Class "0" probability corresponds to the revocation date in the interval from 24 to basically infinity and may be regarded as a probability of non-revocation or one minus probability of default. Observations are considered to belong to the class with the highest probability value in a row.

Modeling was performed using four types of classification methods: logistic regression, random forest, gradient boosting and feed-forward artificial neural network. Predictions quality of models was evaluation by calculation of F1 scores for each class label based on the "one-vs-all" approach. All these models have exogenous parameters that were optimized by grid search. The quality of predictions is measured by testing the model on validation set with the best combination of it's exogenous parameters.

Validation results

In Table 3 below the results of models validation are shown. It contains the values of F1 score calculated for predictions of models that use augmented data.

Table 3 clearly shows the difference in predictions quality between linear logistic regression model and non-linear techniques. Logistic regression in widely used in predicting banks' solvency as its output can be easily interpreted in terms of factors' influence on bank financial performance and in terms of factors' relative importances. However, the predictive score of this model is quite low, and using it for testing hypothesis concerning factors and predictors means putting oneself at risk of making conclusions which are based on unreliable results.

---

[6] All the observations belonging to a certain bank must be either in train or in validation set. If these observation appeared both in train and validation sets then validation procedure will overestimate the quality of models predictions.

Table 3. F1 validation scores of predictions of models with augmented data

Method	Dropout	Class "0"	Class "1"	Class "2"	Class "3"	Class "4"	Average
Log. Reg.	0	0.891	0.01	0.288	0.018	0.157	0.273
	0.2	0.891	0.012	0.286	0.016	0.152	0.271
	0.4	0.890	0.012	0.288	0.017	0.152	0.272
	0.6	0.792	0.072	0.326	0.017	0.162	0.274
	0.8	0.692	0.163	0.348	0.019	0.179	0.280
Rand. For.	0	0.959	0.677	0.602	0.438	0.528	0.641
	0.2	0.961	0.695	0.610	0.447	0.536	0.650
	0.4	0.963	0.710	0.611	0.443	0.536	0.653
	0.6	0.964	0.717	0.620	0.464	0.552	0.663
	0.8	0.951	0.667	0.613	0.465	0.554	0.650
Gr. Boost.	0	0.965	0.735	0.623	0.409	0.564	0.659
	0.2	0.961	0.752	0.640	0.441	0.574	0.674
	0.4	0.951	0.757	0.646	0.433	0.575	0.672
	0.6	0.942	0.778	0.653	0.436	0.586	0.679
	0.8	0.912	0.799	0.669	0.447	0.594	0.684
Neur. Net.	0	0.911	0.722	0.519	0.398	0.535	0.617
	0.2	0.962	0.684	0.702	0.531	0.555	0.687
	0.4	0.931	0.647	0.610	0.384	0.645	0.643
	0.6	0.898	0.716	0.638	0.458	0.492	0.641
	0.8	0.898	0.812	0.691	0.515	0.628	0.709

The differences between non-linear models are far less significant presumably due to the fact that these methods did the best what can be done with this data (although it is hard to say for neural network as it's configuration may vary a lot even if restricted only to dense feed-forward layers). Overall (average) scores of these models are in range [0.65, 0.709] with the individual scores of classes—between 0.384 and 0.965. These results allow us to conclude that a solid predictive model can be build using only banks' public financial reports—without any insiders' information—and public macroeconomic data.

An interesting property of the proposed algorithm of target variable definition is that mean absolute error (MAE) metric can be used for evaluating the quality of estimates due to class orderliness. If the value of this metric is low, it indicates that even when model misclassifies an observation the predicted class is next to actual one, i.e. predicted time of revocation is close to actual revocation date.

Table 4 shows the values of MAE metric for different methods of classification. The exact formula of the metrics is

$$MAE = \frac{1}{\sum_{i=1}^{N} I(y_i \neq \hat{y}_i)} \sum_{\substack{i=1 \\ I(y_i \neq \hat{y}_i)=1}}^{N} |y_i - \hat{y}_i| \qquad (1)$$

where N is the total number of observations, y is actual class number, $\hat{y}$ is predicted class number. Metric is designed to evaluate the mean difference between actual and predicted class numbers over the misclassified observations.

Table 4. MAE values over misclassified observations

Method	Log. Reg					Rand. For.				
Dropuot	0	0.2	0.4	0.6	0.8	0	0.2	0.4	0.6	0.8
MAE	1.767	1.735	1.701	1.687	1.608	1.525	1.521	1.508	1.517	1.503
Method	Gr. Boost					Neur. Net.				
Dropuot	0	0.2	0.4	0.6	0.8	0	0.2	0.4	0.6	0.8
MAE	1.539	1.520	1.511	1.502	1.460	1.605	1.512	1.582	1.565	1.500

According to the results in the Table 4 non-linear methods of classification give more accurate estimates of revocation date than logistic regression. The best MAE scores are in proximity of 1.5, which corresponds to relatively small mistakes in revocation date predictions.

The general pattern of dropout influence is that the bigger its value is, the more accurate predictions are produced by models. Models overfit the most numerous class and underfit the others—less numerous—when one uses unbalanced dataset. The dropout procedure helps to reduce this overfitting and, despite the fact that scores of class "0" decrease, the average score increases due to more accurate fitting of the others classes.

4   Conclusion

In this paper a new approach to bank licenses revocation modeling was introduced. Proposed method of target variable definition allows to estimate probability of revocation and revocation date by setting a multi-target classification problem. The results of the modeling confirm the possibility of predicting licenses revocations (which may also considered as banks' defaults) by using the data from public sources only and advanced non-linear classification techniques.

In the course of the study, it was found that banks' financial indicators data is fairly incomplete and contains a considerable amount of missing values. In order to get representative and consistent dataset one has to use interpolation methods, which can fill the blanks. Most of resulting observations correspond to "normal" situation when banks do not have any visible problems with solvency. This leads to class imbalance with "normal" class heavily outnumbering the others. It was shown that correction for this skew helps to improve the quality of revocation probabilities estimates.

Banks' defaults modeling is usually held using logistic regression classifier. However, validation scores of models based on this routine are low compared to more sophisticated methods like random forest, gradient boosting and feed-forward neural networks. Although, it is easier to conduct factor analysis with logistic classification as one can get coefficients estimates with p-values and it is more difficult to perform this kind of analysis using advanced classifiers but these difficulties are not fundamental and resolved through the development of custom algorithm. Once it is done there are no reasons not to use, say, random forest as standard algorithm of banks' defaults prediction.

The results of modeling described in this paper can be interesting for Deposit Insurance Agency whose responsibility is to provide payments to bankrupt bank's depositors, as well as for Central Bank of Russia which controls Russian bank sector. Commercial companies can benefit from conducting analogous research in order to solve funds distribution problem.

References

1. Peresetsky, A.A., Karminsky, A.M., Golovan, S.V.: Probability of Default Models of Russian Banks. Economic change and restructuring, **44**(4), 297–334 (2011)
2. Karminsky, A.M., Kostrov, A.V.: Comparison of bank financial stability factors in CIS countries. Procedia Computer Science, **31**, 766–772 (2014)
3. Bortell, J.A., Giancola, M.J., Harding, E.J., Patias, P.: Predicting Bank License Revocation (2016). `https://web.wpi.edu/Pubs/E-project/Available/E-project-101716-093448/unrestricted/Final_Report.pdf`
4. Lanine, G., Vennet, R.V.: Failure prediction in the Russian bank sector with logit and trait recognition models. Expert Systems with Applications, **30**(3), 463–478 (2006)
5. Soest, van A.H.O., Peresetsky, A.A., Karminsky, A.M.: An analysis of ratings of Russian banks. Tiburg University CentER Discussion Paper Series, **85** (2003)
6. Boyacioglu, M.A., Kara, Y., Baykan, Ö.K.: Predicting bank financial failures using neural networks, support vector machines and multivariate statistical methods: A comparative analysis in the sample of savings deposit insurance fund (SDIF) transferred banks in turkey. Expert Systems with Applications, **36**(2), 3355–3366 (2009)
7. He, H., Edwardo, A.: Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering, **21**(9), 1263–1284 (2009)
8. Godlewski, C.J.: Are Ratings Consistent with Default Probabilities?: Empirical Evidence on Banks in Emerging Markets Economies. Emerging Markets Finance and Trade, **43**(4), 5–23 (2007)
9. Kolari, J., Glennon, D., Shin, H., Caputo, M.: Predicting large US commercial banks failures. Journal of Economics and Business, **54**(4), 361–387 (2002)
10. Lin, T.: A cross model study of corporate financial distress prediction in Taiwan: Multiple discriminant analysis, logit, probit and neural networks models. Neurocomputing, **72**(16–18), 3507–3516 (2009). doi:10.1016/j.neucom.2009.02.018