EXTENSION OF BANK APPLICATION SCORING MODEL WITH BIG DATA ANALYSIS

Dr. László Madar
Institute for Training and Consulting in Banking
H1011 Szalag u 19., Budapest, Hungary
E-mail: lmadar@bankarkepzo.hu

KEYWORDS
Big Data, Social Scoring, Scoring Model,

ABSTRACT

Application scoring models of credit institutions has been subject to research since the 1960s. Micro and SME lending has also been improving, however current application scoring models for smaller firms have still lower power statistics as models for private individuals or larger firms. This portfolio has not so strong financials and have less bureau collected behavioral information that makes individual assessment of these firms a hard job for credit institutions. This paper presents an actual example on how external unstructured information can be collected, assessed and used in order to increase performance of this portfolio segment.

INTRODUCTION

During the 1960s, several important advancements were made in the area of credit risk assessment. It was the time when rating and scoring systems for credit institutions were first written in a way that these institutions could use them for differentiation good and bad clients in a procedural way (e.g. Altman (1968), Orgler (1970)). Institutions were given a toolset to improve lending process and if data was collected and analyzed correctly, improve the quality of their credit portfolio. Extent of this improvement was different for each portfolio segment credit institutions have, as different amount of reliable data sources were available.

Application scoring systems, where the credit institution first meets its clients always bear the most risk in term of credit risk, as for already approved clients behavioral models can provide exact repayment probabilities at an individual basis, resulting in very high discriminatory power figures. Application information is usually pooled for a more or less homogeneous pools of clients, enabling to treat them as a portfolio. However, if institutions wish to be more precise, they can assess also clients on an individual basis.

In case of private individuals, institutions do not only assess their socio-demographic characteristics, they rather include their individual behavior. By using credit bureau data on how that individual client pays other loans, bills, how much is the debt burden, the credit risk of that individual can be assessed far more accurately than using only socio-demographic variables. This additional accuracy comes from the inclusion of more data that can be bound to the individual behavior.

In case of firms, inclusion of individual payment information is far more difficult. In case of established firms with existing loans bureau information can be accessed and included in the analysis. Also, for large corporation it is worth the bank to assess the corporation on an individual basis. In case of smaller firms, in the micro and SME segment, this gets trickier. Loan sizes are small, individual treatment of clients is expensive, financials are not as solid as is case of larger firms. However, if and institution can lock to itself prospering firms at an early stage, it might get before their competitors. Competitors can win over these companies only with significantly better conditions to make these companies to switch banks. The initiative and profit will remain at credit institutions who at an early stage can differentiate their clients better.

In this paper I introduce a logic that helps institutions to obtain an individual-basis information about SME clients from various data sources that can be applied in the process of lending and credit assessment. This can be used as an augmentation of the current business process as it processes additional information to that is already available. My method is that we collect unstructured information about clients available on the World Wide Web from various sources and process them in a way that it will predict the firm’s future creditworthiness, even in cases where the firm have not applied for a loan before. This is the part when we introduce the now trending big data analysis. In our interpretation, big data analysis helps us obtain information from unstructured data. As we include information on an individual basis, the discrimination of clients will be better as just looking at their mere financials.

Collected unstructured data is processed using text mining methods, and a targeted, simplified semantic analysis to determine the context of the piece of unstructured information collected. The complete process enhances the predictive power of the model as described in detail. This logic was performed using Hungarian language that belongs to a rather complicated family of languages that can be hard to crack with semantics. However, using the targeted approach described, this procedure can be used with any languages on the world.

This presented approach can be used by credit institutions that are willing to invest in unstructured information collection and analysis to improve their lending process for micro and SME corporations.

Proceedings 31st European Conference on Modelling and
Simulation ©ECMS Zita Zoltay Paprika, Péter Horák, Kata Váradi,
Péter Tamás Zwierczyk, Ágnes Vidovics-Dancs, János Péter Rádics (Editors)
ISBN: 978-0-9932440-4-9/ ISBN: 978-0-9932440-5-6 (CD)

MODEL EXTENSION LOGIC

This section presents the approach I followed by specifying, building and testing our approach.

The basic micro and SME model used in this paper was a model I built previously, described in the PhD dissertation in Madar (2015). In this reference work, an SME/micro rating model was developed using financial information of small-medium firms. The developed model was a crisis-proof, stable rating system, where the rating logic itself did not cause cyclicality or movements in the capital requirement, the ordering of the model was stable during crisis years.

In this paper I will use the ROC-Curve and its index number, the AUC measure to assess discriminatory powers. Further information of this measure can be seen in BIS (2005). There might be other measures to be considered, however, all indices show roughly the same changes, as it is a real-word data.

Our overall starting AUC measure that we will use to show the discriminatory power was 0.718 (with a 95% confidence interval of 0.707-0.729), which is an average value for these types of firms, using only financial information to assess their creditworthiness. This model was built using stepwise logistic regression, containing six variables as described in the dissertation.

This basis model was extended by an additional module that contained the result of a big data analysis. The two models were combined with an ensemble approach to achieve optimal combination.

The unstructured data we collected came from various sources. My main source for unstructured data was the social media sites of SME/micro corporations – if they existed. Unlike private profiles of individuals that are well protected by privacy filters, feeds on pages of firms and corporations can be accessed easily, likes and comments, social media activity can be traced. From social media sites, Facebook was processed providing a view of the corporation. For further extension, other social media sites and

To deriving data from these sources, we depended heavily on the algorithms and tools provided by Russel (2015). Data queries used for the current work were based on processes and samples provided by this book.

Lack of unstructured data sources is also an information that adds to the assessment of the firm. It displays either that the firm is too small to have a social media responsible (micro companies with 1-2 employee) or they do not see that as a source of customer acquisition.

Overall, those firms that have web presence is mostly a good thing. Web articles of SME are mostly have a strong positive impact or a strong negative impact. A smaller company is news only when something big is happening around it – getting some investment, boosting sales, having a successful market entry and rocketing sales or when they are negatively impacted, have financial difficulties, have a legal dispute or have to close business just to name a few. Social media sites provide a more sophisticated picture of the corporations. Most SME companies have some social media presence, it is regularly a Facebook presence in Hungary. Activity on the Facebook page displays the SME's commitment to marketing, public relations and social management of clients. Generally SMEs with products for the public are available on the social media, however in some industries it is not so widespread (such as building industry, chemistry, etc., where most of the firms have a limited number of potential customers or customer recruitment is far more effective using non-electronic ways).

This information was analyzed using text mining approach. Word tokens were built from the articles, posts and comments and tokens were aggregated to obtain tokens of similar meaning. Most common tokens (tokens with the most number of counts) were classified manually into three categories: positive, negative and neutral meaning – i.e. a vocabulary was built reflecting the relation of a word to the state of the company in terms of creditworthiness. A positive token would indicate an increase, development, funding of the firm, a negative token would indicate a loss, drop or shutdown. This was the most tedious part of the model building, as in Hungarian language plenty of words can be stemmed from the same root. Unflagged tokens and many common word tokens were treated as neutral, having no impact on the meaning of the article, post or comment collected.

Using this approach, the text of each collected item were classified and could play part of a second tier analysis, where number of posts/likes/comments/articles and its positive/negative classification were assessed. Final assessment of the extension module came from a logistic regression, assessing all information came from the approach described above.

The logistic regression model using financial variables and the logistic regression model using the above analysis were combined using an ensemble logic.

Final model was assessed using the standard AUC measure that showed significant improvement of the discriminatory power. The power of the model improved well outside the 95% AUC confidence level. Based on the effort and sophistication of additional data collection, our results might be further improved.

DATA

In this section, data sources for all model components are described.

Financial data of firms

Financial data of firms

In Hungary all of the data for corporation are public. However, compiling a database from these data takes time, as it can be queried one firm at a time. Data of the complete Hungarian firm register can be queried from the firm service the Ministry of Justice online (http://www.e-cegjegyzek.hu/). This contains the basic information about the corporation, firm type, establishment date, owners, etc. Financial information for each financial year is available also on the homepage of Ministry of Justice (http://e-beszamolo.im.gov.hu/oldal/kezdolap). Data is available mostly in PDF format, however for firms providing their financials in an electronic format, data is available in table format online. Negative information (used to define the default of counterparties) can be accessed on the homepage of the Hungarian Firm Registry Court under liquidations, bankruptcy (http://www.cegkozlony.hu/gazdasagi_ugyszak), and additional negative information is available on the homepage of the tax authority for those with tax payment arrears, enforcement proceedings, suspended tax numbers (http://www.nav.gov.hu/nav/adatbazisok/adoslista).

Although this information is freely available, composing a large enough database takes time and effort, so Hungarian credit institutions rather delegate this task for specialized data provider companies that collect and manage these types of information.

We used the databased compiled for Madar (2015), to further improve with data described in the next section.

Unstructured data for firms

In case of social media sites, data extraction from Facebook followed the logic decribed here. Firms could be in most cases identified by running a search query through the API and select the page with the most number of likes/comments/posts. In case of Facebook that is the most widespread social app in Hungary, a Facebook search query had to be run first and the page IDs and total number of fans had to be queried using the Facebook graph API.

Data from sites were available at structured JSON format that could be deserialized (decompiled) into ordinary SQL database table content to be used in modelling. The complete page feed could be accessed generally by getting first all post id-s belonging to the page and then querying all posts and then querying all comments to these posts. For example in Facebook we used the Graph API to access this information (an access token (denoted as <TOKEN> from now on) is needed for it. I created a web application (available at http:///www.bankarkepzo.hu:15555) to get the queries with its own type of application token. At first step, only page id and the related posts were queried (for example https://graph.facebook.com/ 100878286637188/feed?limit=1000&fields=id,message, created_time&access_token=<TOKEN>). There are a lot of posts, we limited the first 1000 occurrences. However, at the end of the obtained JSON data a link is provided by Facebook for querying the next 1000 occurrences, and so on, until the data block becomes empty. At the second step, we queried 1) the post text themselves and its 2) likes and 3) comments using three different query (e.g. for one post: https://graph.facebook.com/100878286637188_1031983890193285?access_token=<TOKEN>;likes:https://graph.facebook.com/100878286637188_1031983890193285/likes?limit=1000&access_token=<TOKEN>;comments:https://graph.facebook.com/100878286637188_1031983890193285/comments?limit=1000&access_token=<TOKEN>). Text from post and comments are stored in unstructured format, number of likes, number of comments, etc. are available for all feed items, showing a time series of information about the page.

Unstructured data storage was simple, text information from all various sources were stored in the same result table. Text was saved in a long character object and flagged, where it was coming from (web or any of the social media sites), what type of text it was (article, post, comment) and which corporation it belonged to and what weight it got, how relevant was the match of the company name.

Structured data was stored in a timely basis, for each day a summary was generated about the number of contents found (posts, news, likes and comments).

All these compiled dataset enabled us to mine a large amount of data, and extend the scoring model using information from these unstructured data source.

The following table provides an overview of some database totals of the data collected:

Table 1: Collected data sample size

Data source	Type	Count
Facebook	Post	141 688
Facebook	Comment	2 408 190
Facebook	Like	5 298 308

The table above shows that for each data source a reasonable amount of information could have been collected.

However, for about 80% of all Hungarian micro and SME firms, not a single entry was available. This could be a sign of weakness of the approach, however, a number of these firms are dormant, serve as a privately held company controlled by a few people, having no active side bank contact (only marginal interest expense can be observed).

DATA

To address the information hidden within the unstructured data, the collected information had to be processed by using text analytic approach. To uncover some meaning of the text, we had to introduce a quasi-natural language processing logic that could interpret the general context of the article/comment.

Processing of free-form text can be done using two different approaches: rule-based text mining, where statistics are generated from the text for modelling, and linguistics-based text mining, where – using a simple or complex dictionary – some interpretation of the text is made before the modelling starts. As simple rule-based mining can be misleading due to the fact that free-form text tends to use different words and/or phrasing for the same concept. Linguistic methods have an advantage of mapping synonyms or even similar concepts to a common reduced dictionary, and using this common set of words, modelling can be more precise as using the natural form of the free text.

Both approach starts with the tokenization of text. By creating text tokens (i.e. 1-2 word elements without punctuations), the text is formalized and can be processed by text mining approaches. Normally, words are selected as tokens from the text and reduced to their roots. Our source language, Hungarian, is an agglutinative language, i.e. it uses affixes (attached to the word) to change the meaning of the word root, and show tense, plural, etc. Tokenization itself therefore needed a further step to derive etymons, the root meaning of the word that could be classified further. This was made using a statistical approach: the typical affixes were searched within the word (either from the beginning or from the end), and by purging them, if a meaningful word remained that was found within the existing token list, the affix could be deleted, and the token could be changed to the purged token. A token mapping was created that enabled us to classify only the etymons, and not a definitely larger variety of word tokens.

As there is no Hungarian text mining dictionary available at the time, or own mapping had to be created using the derived tokens. We only categorized the most used text tokens, as they were relevant in the business language. The text mining needed a dictionary, and we created a very simplified dictionary of terms. Adjectives, nouns and verbs were classified having either positive or negative meaning. We kept our text context target simple so that the model we build would be more robust. This also enabled us to avoid a creation of a complete linguistic dictionary.

Using this approach, all of the words were classified into a positive, negative or neutral context. Economic articles were the best fit for text mining as they use a type of language that is more or less free from negations (e.g. by stating "not increasing" instead of "decreasing"), sarcasm, jokes, bad grammar and other characteristics of the human language that obfuscate the true meaning of the words. On the other hand, comments, tweets, and sometimes posts are using a language that are sometimes very hard even to tokenize, because they use sometimes abbreviations, have grammatical errors, accent shortage (typical in Hungarian language), sarcastic phrases. This lead to our approach of classification where tokenization and classification of simplified tokens were conducted on web articles, and extended to other collected unstructured data. This resulted that the proportion of positive/negative terms found in non-article sources were lower than in case of articles, but only comments/tweets with proper phrasings were taken into account (however, we had no secret weapon against sarcasm, if well-formulated, it can blur the results of the analysis).

After the assessment of text entries were made, the database was extended a count of negative and positive items within the text record. It was now time to generate explanatory variables for scoring.

As articles/posts/feed have a temporal distribution, beside the general positive/negative state of the client it was also important how the perception about this firm has changed over time. Therefore we built temporal variables measuring the change of comments and their trends in positive and negative relation. Numbers were not absolute as firms might get an increased social/web activity, therefore the change in the ratio of positive and negative comments were measures (from 100% of all text, how many was positive or negative in a given timeframe and with how many percentage points did it increase/decrease).

Generating explanatory variables for defaulted firms was not easy. Normally, social media activity and even the social media pages are closed when a firm declares bankruptcy and data that was available prior to the proceedings is no more available. However, web articles here are also insightful, and in addition to that, social media pages of news portals and news aggregators still have references to defaulted firms. This might lead to a positive bias for firms that are still alive, as they have more positive content over the social media sites as defaulted ones.

Our final list of analyzed variables from the free form text and social media data are the following:

- Overall positive proportion of text
- Overall negative proportion of text
- Overall net proportion of text
- Change in positive proportion over last 3/3 months
- Change in negative proportion over last 3/3 months
- Change in number of likes/+1s over last 3/3 months

It shall be emphasized that we did not only search for correlation within the unstructured data, I tried to build predictive variables that can foresee the future of the firm at a predefined period of time. Similar to behavioral scoring developments at credit institutions, where the payment arrears are the best performing variables, having also a short term predictive power, we expected the same with the collected individual unstructured data. If the firm gets into trouble, customer satisfaction falls right before bankruptcy, and this can provide an additional information to predict the firm's default status. However, predictive power might be also as short lived as payment arrears, as the firm can go bankrupt in a very short amount of time, therefore – also as in case of behavioral scoring – these variables might be used only as a warning signal in credit monitoring, rather than showing the general riskiness of a client.

MODEL CREATION

After the variables have been created, they were processed similarly as any other variable in the logistic regression described in Madar (2015). To avoid the drop in the discriminatory power due to non-linear effects, variables were categorized along their general level of riskiness and a WOE (Weight of Evidence) transformation was performed, to calculate the log odds for each category created. This allowed also the use of logistic regression in the model.

Single factors had low discriminatory power overall, due to the fact that not many firms had observable web activity from the selected sample. Even firms with larger revenue were affected in less web-intensive sectors (e.g. in Hungary building industry, blue-collar services), most of them lacking social media presence and no articles were written about them. The most powerful single factor was the overall negative proportion and change in negative proportion over the last 3 months. These two can predict the defaults well – if only whose corporations are selected that have social media presence, single factor power is in the 30-40% GINI region, depending on categorization. This seems good, however, in relation to the total population the GINI is in the 5-10% region, adding only a bit to our knowledge about the clients.

After the single factor several models were run, bankruptcy as the target variable. Modelling was made using IBM Modeler and models were compared using standard scoring comparison methods. Beside logistic regression, decision tree and neural network models were put together to analyze that the variables are selected and treated appropriately. All models were built on the same dataset, and so, a separate model was created beside the one used as our starting point.

The results of the model was less than expected, the final text model using logistic regression provided a GINI of 14.3%, however, about 80% of the clients were classified in the "unknown" category having no web presence.

Using the ensemble logic of the IBM Modeler, a joint model was created using both logistic regression models, the financial and the unstructured big data source, and the resulting model was also tested. Test showed that discriminatory power increased in a small extent, it is outside the 95% significance limit, so we can say that the extension model brought in additional knowledge of these firms. It provides more insight, however discriminatory power is not so much elevated for a complete bank portfolio. However, if we again reduce our scope to those SME and micro firms that have web presence, the rise of the discriminatory power is in all terms significant, especially negative articles and web posts about firms predict in most cases their downfall.

As a result of the model extension, simple models can be created from big data, and using the approach described above, even further variables might be built to get a more detailed view of the portfolio. The final model has an extended statistical performance, and therefore provides and additional information source for credit institution to profile their corporate clients according to their expected riskiness and provide credit to those companies that have on the one hand better financials, on the other hand have better web perception, good online satisfaction.

This model extension adds to the financials as it is most likely measures the service quality and customer relations quality of a company. Limits of these approach are seen in that not all firms are active online, and they are in the grey zone of this extended analysis. However, the trend is clear, the proportion of clients with online presence is growing year by year..

CONCLUSION

Extension of a financial model with big data analysis is not a simple task. Data has to be accessed, queried and stored in a semi-structured format so that data mining (text mining) is possible on them. Text mining itself has its own challenges as languages might be tokenized with a different efficiency.

The paper showed a process that can be followed to collect and analyze unstructured data about firms, and use this data to assess the quality of the company using their web media content, articles, posts and comments about that firm. This information is out there and can be processed by interested third parties, such as credit institution to get an insight about the firms' online perception.

Power of the model is somewhat low, however we shall note that having a correlation is not always equal to having a prediction (i.e. a random variable n is perfectly correlated with 1-n, however we cannot predict their future state). Big data model in this case seems to extend our knowledge a bit, by introducing new variables that have some relation to the future performance of the company.

We see that there is more room in this topic to create better performing models. There are four points were performance could be increased. First, we can focus on a simpler language and country with deeper web penetration. This would enable to classify more clients from the grey zone. Second, a more detailed token classification can be made, instead of positive/negative, a more differentiated target can be set (e.g 'decreasing' and 'catastrophic' can be separated within the negative section). Third, more explanatory variables can be built that might add to the power of the model. Fourth, the time factor is important, if we collect data in an ongoing manner, we can also collect information about pages where history is not available or pages that ceased to exist due to their bankruptcy.

M. A. Russell, "Mining the Social Web, 2nd Edition". O'Reilly Media, 2013.

L. Madar, "Effects of scoring systems to the results of economic capital models in the institutional capital calculation". PhD Dissertation, University of Kaposvár, 2015.

E. I. Altman, "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy", The Journal of Finance Volume 23, Issue 4, September 1968, pp 589–609

Y. E. Orgler, "A credit scoring modell for commercial loans", Journal of Money, Credit and Banking , Vol. 2, No. 4, Nov., 1970, pp. 435-445

LÁSZLÓ MADAR was born in Budapest, Hungary and attended the Corvinus University of Budapest. He completed his PhD at University of Kaposvár in the field of economics. He is the partner consultant at the Institute for Training and Consulting in Banking (Bankárképző) and a part-time lecturer at Corvinus University of Budapest, Faculty of Finance. He researches big data and social scoring since 2015. His e-mail address is: lmadar@bankarkepzo.hu