Modeling Probability of Default and Credit Limits

Zala Herga
Jožef Stefan Institute
Jamova 39
Ljubljana, Slovenija
zala.herga@ijs.si

Jan Rupnik
Jožef Stefan Institute
Jamova 39
Ljubljana, Slovenija
jan.rupnik@ijs.si

Primoz Škraba
Jožef Stefan Institute
Jamova 39
Ljubljana, Slovenija
primoz.skraba@ijs.si

Blaž Fortuna
Jožef Stefan Institute
Jamova 39
Ljubljana, Slovenija
blaz.fortuna@ijs.si

ABSTRACT

Creditors carry the risk of their clients not meeting their debt obligations. In the literature, these events are often referred to as *default events*. These can be modeled for each company through a *probability of default* (PD). Measures can be taken to limit the default risk: in this paper we focused on credit limit. Firstly, we predict PD of a company using a logistic regression model, based on publicly available financial data. Secondly, we effectively find an optimal portfolio under risk aversion constraints and show how variation of inputs affects the results.

Categories and Subject Descriptors

Mathematics of computing [**Mathematical optimization**]: [Linear programming, Convex optimization]; Computing methodologies [**Machine learning**]: [Supervised learning by regression]

Keywords

PD model, logit, credit limit model, portfolio optimization, linear programming, risk management

1. INTRODUCTION

Payment defaults represent a key default risk (also credit risk) to creditors. Creditors can limit their risk by either insuring their claims or taking preventative measures before extending a credit. Standard tools to measure default risk include different kinds of credit ratings.

Our goal was to create a model that predicts a company's *probability of default (PD)* and provides credit limit suggestions based on the computed PD. One of our constraints was that the underlying PD model be simple and easy to understand. For credit limits, we implemented a linear programming based approach [5] which provides portfolio optimization with risk aversion constraints.

This paper presents the workflow and the methodology that we employed to build a portfolio credit allocation model. Due to privacy concerns it does not include any expiremental results and does not discuss any concrete results or aspects of real data.

The paper is organized as follows. Section 2 provides an overview of related work. In Section 3 the data that is used in modeling is described. Section 4 first describes the approach and computation of the PD model and then presents the results. Section 5 provides a short theoretical introduction to portfolio optimization and then presents our computation and results. Section 6 concludes the paper.

2. RELATED WORK

The Altman Z-score [1] is a widely used credit-scoring model. It is a linear combination of five commonly used financial indicators and it predicts company's degree of PD. Both [6] and [8] argue that Altman Z-score and distance-to-default ([7]) are not appropriate to use in the context of small businesses. The authors in [6] predict PD using delinquency data on French small businesses. They propose a scoring model with an accuracy ratio based solely on information about the past payment behavior of corporations. Similarly, [8] forecasts distress in European SME portfolios. They estimate the PD using a multi-period logit model. They found that the larger the SMEs, the less vulnerable they are to the macroeconomic situation. They also show that SMEs across Europe are sensitive to the same firm-specific factors.

[2] examine the accuracy of a default forecasting model based on Merton's bond pricing model [7] and show that it does not produce a sufficient statistic for the PD. [11] compare the predictive accuracy of PD among six data mining methods. They also present a novel "sorting smoothing method" for estimating the real PD. Using a simple linear regression, they show that artificial neural networks produce the best forecasting model. [3] used the Merton model to show that, on contrary to what theory suggests, the difference in returns between high and low PD stock is negative and that

returns almost monotonically decrease as the PD increases. However, they found a positive relationship systematic default risk exposure and returns.

On the problem of portfolio optimization, [9] showed that *Conditional Value at Risk* minimization with a minimum expected return can be computed using linear programming techniques. [5] built on this idea and showed that alternatively, one can maximize returns while not allowing large risks.

3. DATA

The dataset that we used covers several thousand companies from several European countries. Data for each company consists of two parts: financial data and trading data.

Financial data corresponds to publicly available data - balance sheets and income statements of a company.

Trading data consists of private information on trades between our data provider and his clients. It contains monthly data about the sum of trades, outstanding debts, disputed claims and delayed payments. The data is available for years between January 2010 and June 2016.

4. PD MODEL

All creditors carry the risk of their clients not paying the bills. We can be quite certain that some clients will not pay, however, we have difficulties identifying those clients. Hence, our first task was to compute the probability of default for each client company. The end model should also be simple and easy (intuitive) for domain experts to understand.

Default can be defined in multiple ways, considering the available data and how strict we want our model to be. Do we consider it a default if the client is only one day late on payments? Or do we let them be 30, 45, 60 days late before taking action? Are we going to consider a client defaulted if he owes 10€? What if the client didn't pay one bill but has been paying all the bills after? We are required to make a judgment call and choose a definition that meets the creditor's needs best.

As soon as a client defaulted we removed it from the dataset. Meaning, each client can only default once and after that we assume there was no more trading done with them.

There were some companies that were already defaulted at the beginning of the time-span of our dataset. We removed these companies from the dataset since they provide no useful information. We also filtered out clients with low sales since their financial data can be very unreliable and their impact on our model is big compared to the trading volume that they generate.

Due to the constraint that the model should be simple and easy to understand we chose to model the PD with logistic regression.

From the available financial data we calculated 45 financial indicators for each company that cover aspects of solvency, liquidity, debt, profitability and operative effectiveness status of a company. We transformed each financial indicator vector into a feature vector by binning and assigning *Weight of Evidence (WOE)* [10] to it. The idea is as follows: create n bins in range from min to max indicator value and assign each company to the corresponding bin. Then count the number of 'bad' (defaulted) and 'good' companies in each bin. Then assign WOE to a bin as

$$\log \frac{\mathbb{P}(company = good)}{\mathbb{P}(company = bad)}.$$

Since WOE scores will be used as inputs to a linear model, they should be a monotonous function over bins, meaning, the higher the financial indicator value the better the company is (if WOE is increasing) or the higher the indicator the worse the company is (if WOE is decreasing). In Figure 1 we show an example of binning and WOE transformation on simulated data.

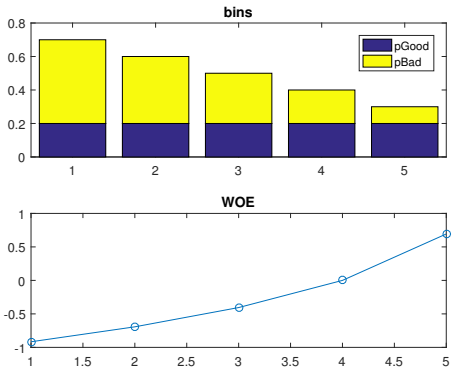

Figure 1: Bins and WOE on simulated data. The greater the value of the simulated data the greater the evidence that a company is 'good'.

Thus, we obtained 40 out of 45 features. Another feature was the size of the company (based on income) and four of them were country features (dummy variables - one for each country).

As described above, we mapped "raw" features into WOE features:

$$(x_1, x_2, ...) \rightarrow (woe(x_1), woe(x_2)..)$$

PD of a company is then predicted through logistic distribution function:

$$F(x) = \frac{1}{1 + \exp^{-(\beta_0 + \beta_1 \cdot woe(x_1) + \beta_2 \cdot woe(x_2) + ... + \beta_n \cdot woe(x_n))}}$$

where β_i denotes linear regression coefficients.

4.1 Computation

Since we have the response variable observations on monthly level, we interpolated features to obtain the same frequency of explanatory variables. We also took into account offset of the financial and trading data: financial data is only made available sometime around June each year for the previous year (e.g. in June 2016 we only know financial statuses of 2015).

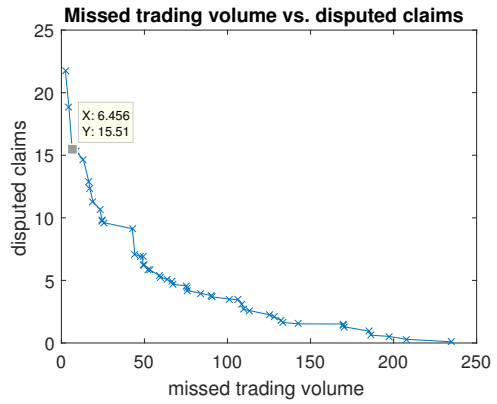

Figure 2: Missed trading volume vs. disputed claims on simulated data. The info box marks 6%-level: if trading was stopped with the worst 6% companies, approx. 7.5 could be saved on disputed claims and $6.5 \cdot margin$ profit would be lost on trading volume. The figure is included for illustrative purposes and is not based on real data for privacy concerns.

We then filtered out features that:

- had high percentage of missing values (We kept threshold as a parameter. in this paper, threshold of 0.7 was used.)
- were highly correlated (In this paper, threshold of 0.9 was used).

After that, 27 features were left in our feature set.

We used Matlab for computation. The data was trained on 42 months and tested on 19 months. Two models were trained: standard logistic model and stepwise logistic model. In stepwise regression, explanatory variables are added to a model by an automatic procedure [4]. Regression coefficients are estimated by maximum likelihood estimation.

4.2 Results

Test data consisted of 20% of the data (not used in training). There were 16 features chosen in stepwise model. Both of the models have 0 p-values and return very similar results in predicted values.

We evaluated models by comparing the amount of disputed claims (true negatives) to the amount of missed trading volume (false positives) given ceased trading with companies with PD exceeding some threshold (Figure 2). Note, that expenses based on missed trading volume cannot be directly compared to disputed claims; disputed claims are a direct expense, whereas missed trading volume number consist largely of expenses (that a company in that case did not have). Hence, one needs to multiply the trading volume with company's (average) margin to obtain actual opportunity costs.

5. CREDIT LIMITS MODEL

Naturally, a question arises once we identify risky clients: how to handle them? Client should have set a credit limit, but how to set the limit? If the limit is to high, the client might not be able to pay the bills, but if the limit is too low, profit is lost on trading volume. The model that we created is inspired by [5], is based on PD calculation presented in first part of this paper and takes into account the level of risk that creditor is willing to take.

Let us introduce some standard financial risk-related terms. *Value at Risk* (VaR) is an upper percentile of loss distribution. Probability level is denoted as α. E.g. $95\% - VaR$ of $1,000,000€$ means that there is a 0.05 probability that loss will exceed $1,000,000€$. *Conditional Value at Risk* (CVaR) is the conditional expected loss under the condition that it exceeds VaR. CVaR at $\alpha = 95\%$ level is the expected loss in the 5% worst scenarios. ω denotes the maximum allowed CVaR of the portfolio at level α.

We will denote loss associated with the portfolio x and random vector y (with density $p(y)$) as $f(x, y)$; $CVaR$ will be noted as $\phi_\alpha(x)$, which is given by

$$\phi_\alpha(x) = (1-\alpha)^{-1} \int_{f(x,y)>VaR_\alpha} f(x,y)p(y)dy.$$

It has been established [9] that $\phi_\alpha(x)$ can be computed by minimizing the following function:

$$F_\alpha(x,\zeta) = \zeta + (1-\alpha)^{-1} \int_{y\in\mathbb{R}^n} \max\{f(x,y)-\zeta, 0\}p(y)dy,$$

and that the value ζ which attains the minimum is equal to VaR_α.

Finding credit allocations $x \in X$ that maximize the expected profit under $CVaR$ is equivalent to the following optimization problem [5]:

$$\begin{aligned} \min_{x\in X, \zeta\in\mathbb{R}} \quad & -R(x) \\ \text{subject to} \quad & F_\alpha(x,\zeta) \le \omega \end{aligned} \tag{1}$$

where $R(x)$ is the expected profit and the set X is given by a set of box constraints (lower and upper bounds on each component of x).

5.1 Computation

By using the PDs from the first part of the paper, we can simulate the default events and compute several random scenarios for our portfolio. Since each company is assigned a probability of default we can generate random scenarios (where certain companies default) over the full portfolio by sampling from independent (with different weight) Bernoulli random variables.

By generating a set of sample scenario vectors $y_1, \ldots, y_J$ with their corresponding probabilities $\pi_1, \ldots, \pi_J$ we can approximate $F(x, y)$ by a finite sum:

$$\tilde{F_\alpha}(x,\zeta) = \zeta + (1-\alpha)^{-1} \sum_{j=1}^{J} \pi_j \max\{f(x,y)-\zeta, 0\}$$

Using

$$z_j \ge f(x,y_j) - \zeta, \quad z_j \ge 0, \quad j = 1,\ldots,J, \quad \zeta \in \mathbb{R}$$

the constraints in (1) can be reduced to a system of linear constraints:

$$\zeta + (1-\alpha)^{-1} \sum_{j=1}^{J} \pi_j z_j \leq \omega \tag{2}$$

$$f(x, y_j) - \zeta - z_j \leq 0, \quad z_j \geq 0, \quad j = 1, \ldots, J, \quad \zeta \in \mathbb{R} \tag{3}$$

In our case, the expected profit is

$$R(x) = (1 - pd) \cdot x \cdot margin - pd \cdot x.$$

As for PD modeling, we used Matlab for computation. We combined left-side part of constraints from (2) and (3) in a matrix A; the right-hand side was combined in a vector b. We also added additional constraints on x, that are specific to our problem: we set an upper and lower bound (ub and lb correspondingly) to the credit limit.

We then have to solve linear program:

$$\min_x -R(x) \quad \text{such that} \quad \begin{cases} Ax \leq b \\ lb \leq x \leq ub \end{cases}$$

5.2 Results

Our method provides an optimal portfolio based on α, ω, margin, PDs, credit limit upper- and lower bounds. By optimal portfolio we mean monthly credit limit for each company, which takes value between the provided credit limit bounds. In Figure 3 we present some scatter-plots based on results; the default values used for these graphs are $\alpha = 0.95$, $lb = 0$, $ub = max\ trading\ volume$ (based on historical data) and $margin = 0.01$. Most companies get either zero or maximum credit approved; there are only some companies the get part of the max credit approved. In 3b we decreased ω by a factor of 10. This is a lot stricter constraint and consequently there are more companies that get zero credit approved and many companies that get only a part of max credit approved. In 3c we increased margin from 0.01 to 0.1. In practice this means, that the credit-giver is making profit on trading volume. In 3d, we moved credit lower bound from zero to $0.1 \cdot ub$. This makes the PD threshold stricter.

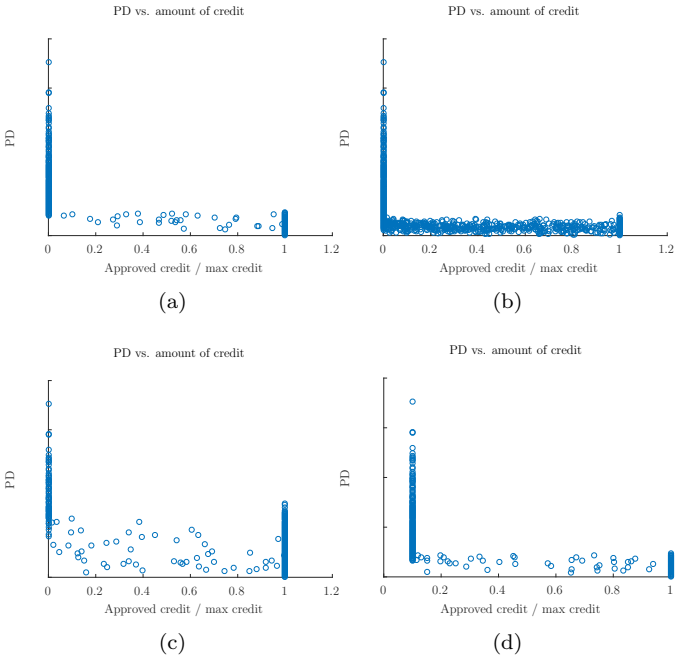

Figure 3: Each circle in these scatter-plots represents one company. The x-axis represents the relative amount of approved credit. Y-axis represents predicted PD of a company. In 3a default values are used, while Figure (b) is based on reducing ω by a factor of 10, figure (c) is based on increasing the margin by a factor of 10 and figure (d) is based on lifting the lower bound to 10% of the upper bound. The scales of PDs are omitted on the graphs due to privacy concerns.

6. CONCLUSION

We presented a logit model based on *weight of evidence* features to predict a company's *probability of default*. Standard and stepwise methods were used to train the data. Both methods provide similar results.

In second part of the paper we introduce an efficient portfolio optimization technique that was used to determine credit limits for creditor's clients. We presented the results and showed how variation of inputs impacts the results.

7. REFERENCES

[1] E. I. Altman. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The journal of finance*, 23(4):589–609, 1968.

[2] S. T. Bharath and T. Shumway. Forecasting default with the kmv-merton model. In *AFA 2006 Boston Meetings Paper*, 2004.

[3] S. Ferreira Filipe, T. Grammatikos, and D. Michala. Pricing default risk: The good, the bad, and the anomaly. *Theoharry and Michala, Dimitra, Pricing Default Risk: The Good, The Bad, and The Anomaly (March 2015)*, 2015.

[4] R. R. Hocking. A biometrics invited paper. the analysis and selection of variables in linear regression. *Biometrics*, 32(1):1–49, 1976.

[5] P. Krokhmal, J. Palmquist, and S. Uryasev. Portfolio optimization with conditional value-at-risk objective and constraints. *Journal of risk*, 4:43–68, 2002.

[6] A. Marouani. Predicting default probability using delinquency: The case of french small businesses. *Available at SSRN 2395803*, 2014.

[7] R. C. Merton. On the pricing of corporate debt: The risk structure of interest rates. *The Journal of finance*, 29(2):449–470, 1974.

[8] D. Michala, T. Grammatikos, S. F. Filipe, et al. Forecasting distress in european sme portfolios. *EIF Research & Market Analysis Working Paper*, 17, 2013.

[9] R. T. Rockafellar and S. Uryasev. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000.

[10] D. Sharma. Evidence in favor of weight of evidence and binning transformations for predictive modeling. *available at: http://ssrn.com/abstract=1925510*, September 2011.

[11] I.-C. Yeh and C.-h. Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2):2473–2480, 2009.