Credit Scoring and Credit Control XV - 2017 Edingburg

LGD and RR Modeling - Comparison of models

Sukriye TUYSUZ

Yeditepe University

Uses of Loss Given Default (LGD) models : Regulatory capital, Economic capital, Pricing, Loan Loss Provisioning,...

Basel 2 ; determination of the minimum required capital for credit risk can be determined by using Advanced IRB model :
Required capital = $EAD.LGD.(WCDR - PD)$,
where : EAD : exposure at default, LGD : loss given default and PD=probability of default.
EAD, LGD and PD are determined internally.

- LGDs values are within the interval 0-1
- LGDs distribution : skewed and/or bimodal.

Dynamic of LGD can be determined by evaluating :

- the dynamic of LGDs,
- the dynamic of Recovery Rate (RR) : RR=1-LGD,
- the dynamic of the losses : losses = EAD*LGD.

- Earlier studies considered classical model : OLS, Fractional Response model, Tobit model
- Latter authors proposed models more appropriated to the features of LGDs (skewed and/or bimodal) :
  - Simple skewed regression : beta regression, gamma regression, inverse Gaussian regression, log-normal regression.
  - Adjusted regression : zero/one-adjusted regressions
  - Inflated regressions : Inflated Regression and Ordered Logistic Regression
- Validation and discriminatory powers of models are compared by using several classical and advanced metrics.

The aim of this paper :

- is to evaluate the dynamic of LGDs/RRs by using classical approaches as well as models taken into account the unusual form of LGD/RR (skewed models and mixture models).
- And compare models performance by using classical metrics as well as more recent metrics used mainly by practitioner.

Review of literature

1) Classical model

- Earlier studies have formalized LGDs'/RRs' dynamic with a standard statistical model (OLS).
  - Drawbacks : 1) Estimated LGDs/RRs maybe out of 0-1 interval. 2) LGDs/RRs are not normally distribution.
  - Solutions : Transform LGDs/RRs to make them normally distributed.
  - Transformation functions : inverse Gaussian (IG), Beta transformation (IGB), Normal (N), log (Log), logit (L), BoxCox, ArcSin, Power, and Reciprocal transformation.
  - Need adjusted LGDs/RRs :
    local adjustment ($LGD_0 = LGD + \epsilon$ if $LGD = 0$ and $LGD_1 = LGD - \epsilon$) if $LGD = 1$,
    and global adjustment ($LGD = b + (1 - 2b).LGD$).

- Fractional response regression (FR) - Papke and Wooldridge (1996)
  - $$E(LGD|x) = G(x\beta),$$
  - The function $G(.)$ is often specified as a (1) logistic function or (2) a log-log function (3) or a probit (4) or a cauchit function.

2) Unusual distribution of LGD : skewed distributions
2a) Skewed distributions - 0/1 Adjusted regressions

- Beta Regression (BE) - Moody's KMV Losscalc software package - Gupton and Stein (2005)
- Gamma Regression (GA) - Sigrist and Stahel (2011)
- Inverse Gamma regression (IGA)
- Inverse Gaussian regression (IG)
- Log-Normal regression (LogN)

According to Ospina and Ferrari (2010, 2011), a continuous-discreet distribution is more suitable for LGD. This mixed model is expressed as :

$$\begin{aligned} P(y;\mu,\theta) &= \pi && \text{if} \quad y = c, \\ &= (1-\pi) f(y;\mu,\theta) && \text{if} \quad y \in (0,1), \end{aligned}$$

where : $y = LGD, RR$
$f(y;\mu,\theta)$ is the beta density. Can be a beta, gamma, log-normal, inverse Gaussian and normal probability density function.
π represents the probability at point c, which can be considered as 0 or 1 (c=0 or c=1). In case $c = 0$, we have zero-adjusted distribution and a one-adjusted distribution in case $c = 1$. The zero-adjusted distribution is more appropriated for data set containing large number of 0 and a one-adjusted distribution for a dataset containing a large number of 1.

μ, π and θ can be time dependent with the following dynamics :

$$\begin{aligned} g_1(\pi) &= g_1(x^T\beta_1), \\ g_2(\mu) &= g_2(x^T\beta_2), \\ g_3(\theta) &= g_3(x^T\beta_3), \end{aligned}$$

where x is the set of explanatory variables and β represents the vector of coefficients. The functions $h(.)$ are the link functions. For μ and σ, Ospina and Ferrari (2011) propose : a logit $[h_2(\mu) = log(\frac{\mu}{1-\mu})]$, or a probit link function $[h_2(\mu) = \Phi(\mu)$, where $\Phi(.)$ is a cumulative normal distribution], or a complementary log-log $[log(-log(1-\mu))]$ or a log-log link function $[log(-log(\mu))]$.
As for $h_3(\theta)$, a log link $[h_3(\theta) = log(\theta)]$ or a square-root link $[h_3(\theta) = \sqrt{(\theta)}]$ are recommended. Coefficients of the model are estimated by maximizing the likelihood function.

2b) Inflated regression :

A general form of Adjusted regression is the Inflated regression, expressed as :

$$\begin{aligned} P &= P_0(\nu) && \text{if} \quad y = 0, \\ &= (1 - P_0 - P_1) f(y, \mu, \phi) && \text{if} \quad 0 < y < 0, \\ &= P_1(\tau) && \text{if} \quad y = 1, \end{aligned}$$

$y = LGD, RR$

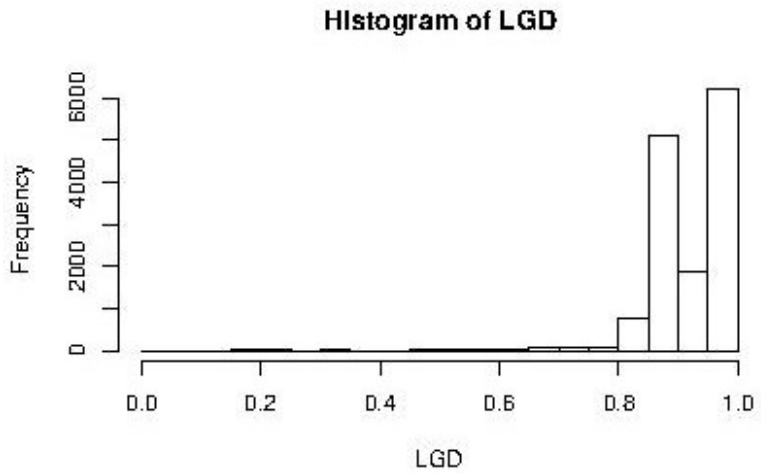
$f(.)$ is a probability density function (PDF). Can be a beta, gamma,... density function.

The mean μ and the probabilities P_0^i and P_1^i are defined as :

$g_1(\mu) = g_1(x^T \beta_1), \quad g_2(\phi) = g_2(x^T \beta_2), \quad g_3(\nu) = g_3(x^T \beta_3),$ and $\quad g_4(\tau) = g_4(x^T \beta_3)$

Data

- Data extracted online from the Lending Club.
- This article uses workout LGD.
- Weighted-balanced sample : random sample of 400 applications are selected randomly (300 for in-sample, 100 for out-sample) from every month from 01/2012 to 12/2015.
- In total, the in-sample is composed with 14400 applications and the out-sample with 4800 applications.



Explanatory variables

- Bellotti and Crook (2012) : five types of factors are important explanatory variables : 1) individual details, 2) account information at default, 3) changes in the personal/obligor situation over time, 4) macroeconomic situation and 5) decisions of the bank on the level of risk.
- 59 variables are assumed relevant among the 116 provided information (variables) for each customer.
- Transformation of variables : Weight of Evidence (WOE) values of Categorical variables and Natural logarithm transformation for certain continuous variables.
- Filtering process : 1) Cramer's V is considered in this article + our analysis and decisions and 2) Pearson correlations : 13 explanatory variables are finally retained.
- A final variables selection is done during the model estimation stage.

- Selected variables : 1) The WOE of the LC assigned loan subgrade (*grade*), 2) The logarithm of self-reported annual income provided by the borrower during registration (*income*), 3) The WOE of the state provided by the borrower in the loan application (*address*), 4) The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years (*delinq*), 5) Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit. ($revol - util$), 6) The logarithm of the total current balance of all accounts ($cur - bal$), 7) Number of trades opened in the past 24 months ($acc - open$), 8) Number of mortgage accounts. ($mort - acc$), 9) Number of bankcard accounts ($num - bc$), 10) Number of installment accounts ($num - il$), 11) Number of revolving trades with balance > 0 ($num - rev$), 12) The logarithm of the total bankcard high credit/credit limit ($bc - limit$), and 13) $Max - months$.

Test of validation and Discriminatory power of models

Check the goodness-of-fit, prevision power and discriminatory power of models.
Classical approaches : error-based metrics (MSE, RMSE, MAE, MAPE), or/and correlation-based statistics (Pearson's r, Kendall's τ, Spearman's ρ, coefficient of determination R^2) or/and classification-based metrics (Area Under the Curve (AUC)).

Advanced approaches : Loterman et al. (2014) :
- Central Tendency metrics and Errors Dispersion Tests : T test and the Wilcoxon signed rank test - Test whether mean errors (T test) equals zero and whether median error in the Wilcoxon test equals zero.
- Error dispersion tests - Fisher test - enables to verify whether the error dispersion is getting wider.

Empirical Results

- In OLS regressions using transformed LGDs/RRs, calculated R^2 between transformed back fitted/predicted LGDs (and RRs) and observed LGDs (and RRs) are negative in case local adjustment is applied. In OLS regressions using globally adjusted and transformed LGDs, some R^2 are negative and some are positive.
- According to the error-based metrics, OLS regressions using transformed LGDs/RRs at power lower than 0 (0.5, 1/3 or 0/4) out-perform in term of validation and classification models using transformed LGDs/RRs at power higher than 0 in-sample as well as out-sample.
- Fractional response model based on probit and cauchit link functions present slightly higher predictive and discriminatory power than the two other retained link functions (logit and loglog) in both samples as well as for both retained response variables (LGD and RR).

- Our results indicate that fractional response model presents higher performance in term of prediction than Tobit model in both samples.
- Results of classical models indicate that performing OLS regression based models (simple OLS, *IG.G*, *Norm.G*, *Logit.G*) have higher predictive and discriminatory power than other retained classical models.
- Simple beta regression using locally adjusted LGDs (BE.L) out-performs other retained models modeling the skewed nature of LGDs in-sample as well as out-sample.
- As for the dynamic of RRs, our results reveal that the simple beta regression using locally adjusted RRs and the simple gamma regression have higher predictive and discriminatory power than the other retained models modeling the skewed nature of RRs.
- The Ordered Logistic regression (2-step approach) has higher predictive and discriminatory power than inflated regressions in both samples.

- All obtained results for LGDs suggest that zero-adjusted beta regression and simple beta regression have higher goodness of fit in both samples.
- As for our RRs, simple beta regression based on locally adjusted RRs and simple gamma regression are the most performing models followed by the one-adjusted beta regression.
- Is it better to evaluate directly the dynamic of LGDs or evaluate the dynamic of RRs and then draw the dynamic of LGD ($= 1 - RR$) ?
  It is better to evaluate the dynamic of RRs with the beta regression based and/or the gamma regression based model and then draw the fitted/predicted values of LGD.
- In all retained classical models, the mean LGD and the mean RR are influenced significantly by the annual income (*income*), the rating (*grade*) and the address (*address*).

- Compared to the classical models, the mean of LGD and the mean of RR react significant to less variables in models taking account the unusual nature of our dependent variables (adjusted regressions, inflated regression and ordered logistic regression).
- As our LGD dataset contains only 4 observations having 0 value, the probability of $LGD = 0$ (P_0) does not react significantly to retained explanatory variables as expected. By contrast, the probability of $LGD = 1$ (P_1) reacts significantly almost all retained independent variables.
- Similarly, the probability of $RR = 1$ does not react significantly to retained explanatory variables as our RRs dataset contains only few observation. Whereas the probability of $RR = 0$ is influenced by almost all retained covariates as the probability of $LGD = 1$.

Conclusion

- The simple beta regression using locally adjusted LGDs/RRs have higher predictive and discriminatory power.
- It is better to evaluate the dynamic of RRs with the beta regression based and/or the gamma regression based model and then draw the fitted/predicted values of LGD.