UNIVERSITY OF Southampton
School of Management

Modelling LGD for unsecured personal loans

Comparison of single and mixture distribution models

Jie Zhang, Lyn C. Thomas
School of Management University of Southampton
26-28 August 2009 Credit Scoring and Credit Control XI

UNIVERSITY OF Southampton
School of Management

Outline

- Loss Given Default and Recovery Rate
- Research Methods
- Data
- Single Distribution Models
- Mixture Distribution Models
- Model Comparison
- Conclusions and Further Research

Loss Given Default

- LGD is the final loss of an account as a percentage of the exposure, given that the account goes into default
- Recovery Rate = 1- LGD
- RR=Recovery Amount / Default Amount
- Recovery Amount = Default Amount – Write-off Amount

  OR Default Amount – Last Outstanding Balance

UNIVERSITY OF Southampton
School of Management

Research Methods (1)

Single distribution models

- Linear Regression
- Survival Analysis Models
  - Censored data
  - Fit various distributions
  - Quantile

UNIVERSITY OF Southampton
School of Management

Survival Analysis models

- Usually use survival analysis in time but here use it in money or percentage of debt recovered.
- F(t)= Probability recovery rate is no greater than t%
- S(t) =1-F(t) = Prob. Recovery rate above t
- Hazard function h(t) = F'(t)/(1-F(t))= density function recovery rate is t given it is at least t.
- For borrower with characteristics x
  - Accelerated life model , $S(t) = S_0(e^{c.x}t)$
  - Proportional Hazard model $h(t) = e^{c.x} h_0(t)$

Research Methods (1) cont.

- Survival Analysis Models
  - Accelerated failure time models
    - Logistic Regression first classify zero recoveries and non-zero recoveries
    - Set distribution type in model building
  - Cox proportional hazards models
    - Both ways were tried
    - Can fit any types of distribution

UNIVERSITY OF Southampton
School of Management

Research Methods (2)

- Mixture distribution models
  - Different Recovery Rates in segments
  - Debtors' Views
  - Different Distribution in segments

Classification Tree model to segment the whole population then to build linear regression models and survival analysis models on each segment

Data

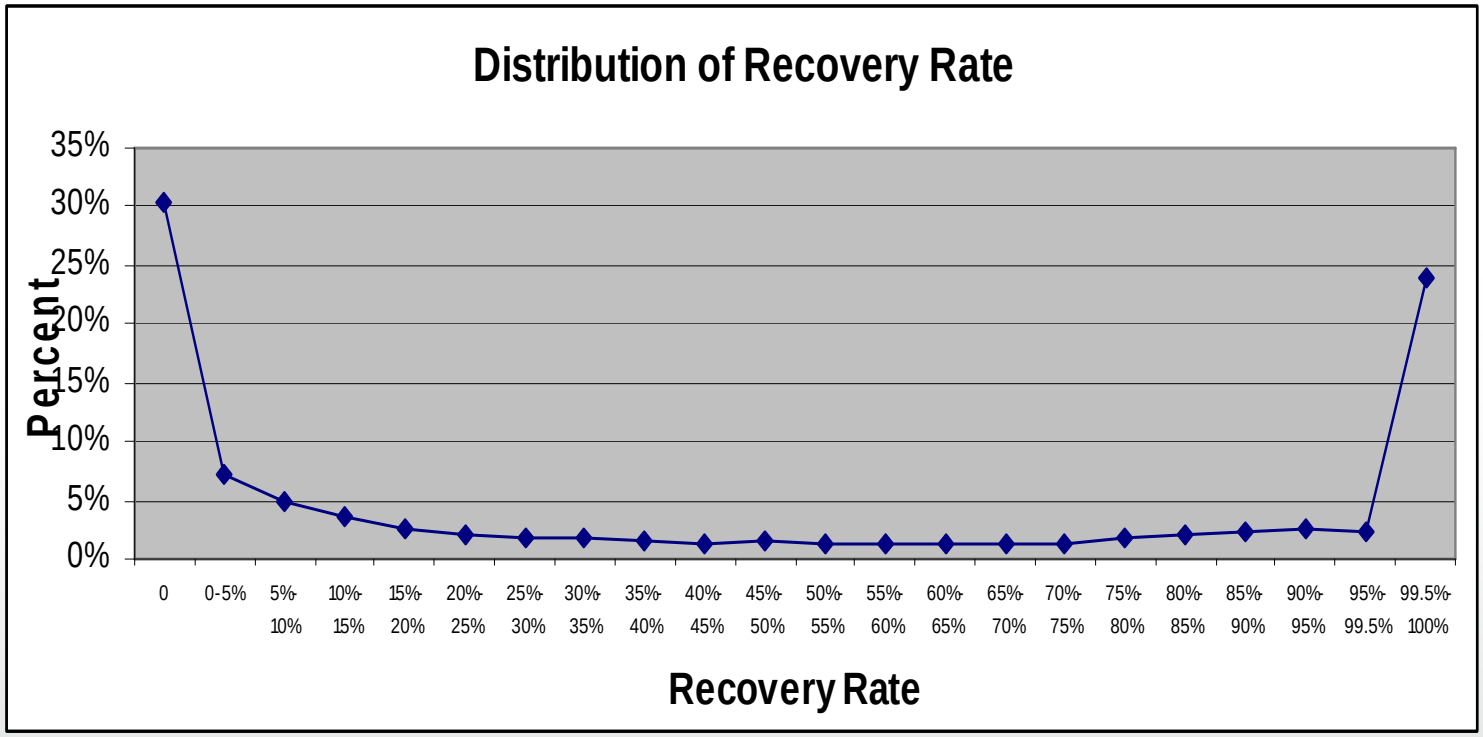
The data is a default personal loan data set from a UK bank. The loans were issued from 1987 to 1999, and repayment patterns were recorded until the end of 2003.

Over 27,000 debts, 20% had been paid off, 14% were still being paid, 66% were written off.

Key characteristics about debtors and debts includes:

Residential status, Employment status, Marital status, Time at address, Time in occupation, Time at the bank, Second applicant status, Loan purpose, Age, Whether have mortgage, Loan term, Monthly income, Monthly expenditure, and so on…

Data


Distribution of Recovery Rate
Percent
35%
30%
25%
20%
15%
10%
5%
0%
0
0-5%
5%-10%
10%-15%
15%-20%
20%-25%
25%-30%
30%-35%
35%-40%
40%-45%
45%-50%
50%-55%
55%-60%
60%-65%
65%-70%
70%-75%
75%-80%
80%-85%
85%-90%
90%-95%
95%-99.5%
99.5%-100%
Recovery Rate

Results from single distribution models (Recovery Rate)

The whole population was spilt to 2 parts:

- 70% as training sample for model building
- 30% as test sample for model test

Results are based on test sample

	Optimal Quantile	R square	Spearmen Rank Coefficient	MAE	MSE
Linear regression		0.0904	0.29593	0.3682	0.1675
Accelerated Weibull	34%	0.0598	0.25306	0.3586	0.2042
Accelerated Log-logistic	34%	0.0638	0.25990	0.3560	0.2060
Accelerated Gamma	36%	0.0527	0.23496	0.3635	0.2015
Cox-including 0	46%	0.0673	0.27261	0.3546	0.2006
Cox-excluding 0	30%	0.0609	0.25506	0.3564	0.2072

Results from single distribution models (Recovery Amount)

	Optimal Quantile	R square	Spearmen Rank Coefficient	MAE	MSE
Linear regression		0.1807	0.28930	1212.1	2634270
Accelerated Weibull	34%	0.1341	0.30594	1123.5	3026908
Accelerated Log-logistic	34%	0.1318	0.31178	1111.7	3047317
Cox-including 0	46%	0.1572	0.31788	1138.9	2887499
Cox-excluding 0	30%	0.1400	0.30437	1125.3	3017661

Another way to get predictions

- To get RR from Recovery Amount models

$$\frac{\text{Predicted Recovery Amount}}{\text{Default Amount}} \longrightarrow \text{Predicted RR}$$

- To get Recovery Amount from RR models

$$\text{Predicted RR} \times \text{Default Amount} \longrightarrow \text{Predicted Recovery Amount}$$

UNIVERSITY OF Southampton
School of Management

Results from single distribution models (Two ways for Recovery Rate)

	from Recovery Amount model				from Recovery Rate model			
	R square	Spearmen Rank Coefficient	MAE	MSE	R square	Spearmen Rank Coefficient	MAE	MSE
Linear regression	0.0292	0.22837	0.4077	0.2432	0.0904	0.29593	0.3682	0.1675
Accelerated Weibull	0.0544	0.24410	0.3606	0.2070	0.0598	0.25306	0.3586	0.2042
Accelerated Log-logistic	0.0591	0.25315	0.3575	0.2077	0.0638	0.25990	0.3560	0.2060
Accelerated Gamma					0.0527	0.23496	0.3635	0.2015
Cox-including 0	0.0425	0.22646	0.3693	0.2216	0.0673	0.27261	0.3546	0.2006
Cox-excluding 0	0.0504	0.23269	0.3624	0.2108	0.0609	0.25506	0.3564	0.2072

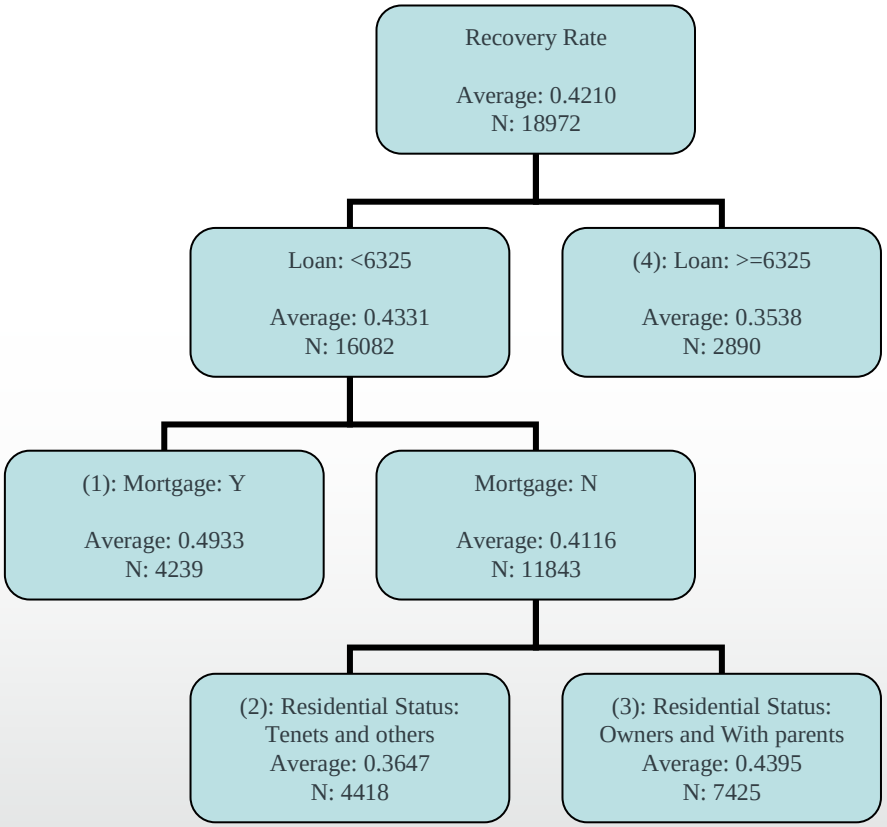
UNIVERSITY OF Southampton
School of Management

Results from single distribution models (Two ways for Recovery Amount)

	from Recovery Amount model				from Recovery Rate model			
	R square	Spearmen Rank Coefficient	MAE	MSE	R square	Spearmen Rank Coefficient	MAE	MSE
Linear regression	0.1807	0.28930	1212.1	2634270	0.2068	0.31522	1162.4	2549591
Accelerated Weibull	0.1341	0.30594	1123.5	3026908	0.1424	0.31149	1116.1	2982477
Accelerated Log-logistic	0.1318	0.31178	1111.7	3047317	0.1396	0.31697	1105.9	3014320
Accelerated Gamma					0.1413	0.30139	1141.5	2972807
Cox-including 0	0.1572	0.31788	1138.9	2887499	0.1628	0.34619	1101.9	2906821
Cox-excluding 0	0.1400	0.30437	1125.3	3017661	0.1377	0.31246	1107.4	3028183

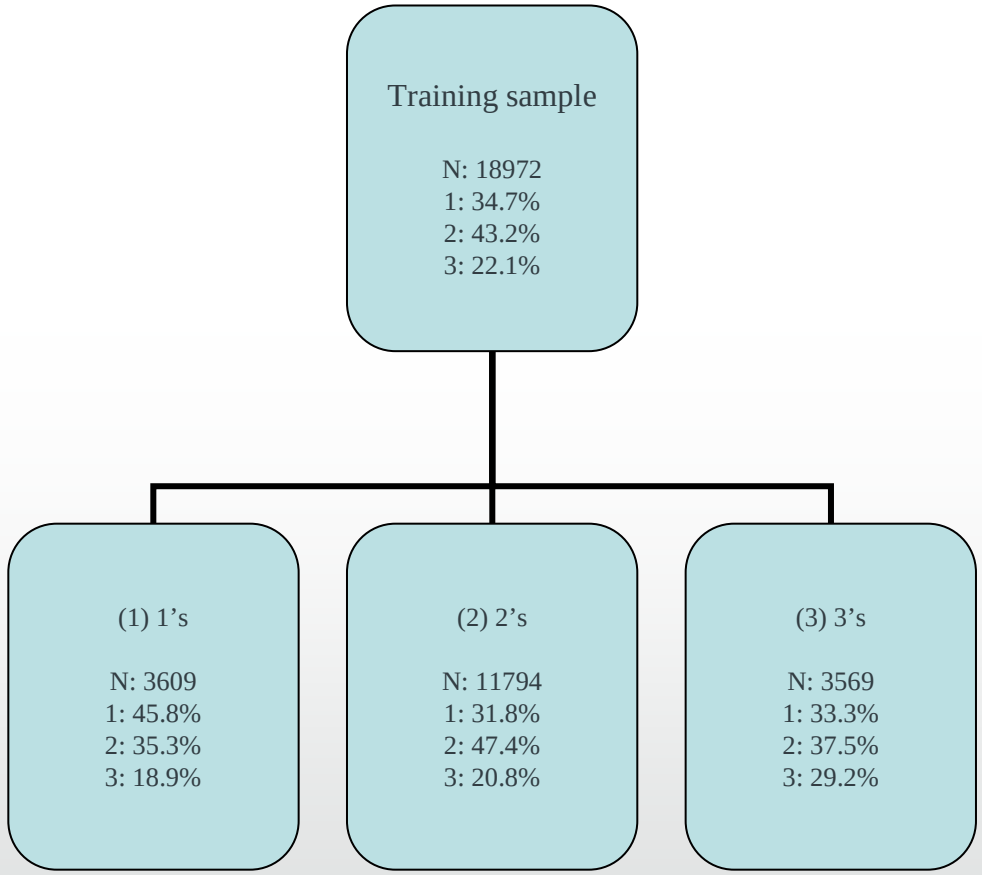
Mixture distribution models

- Method 1: to maximise the distance of average RR between segments
- The classification tree is built on training sample and 4 segments are created. Linear regression and survival analysis models are built for each 4 segment. 4 test samples are combined to form the whole test sample the same as before.



Mixture distribution models

- Method 2 : to split the whole population into 3 segments.
  - (1) No recovery (RR<0.05)
  - (2) Partial recovery (0.05<RR<0.95)
  - (3) Full recovery (RR>0.95)



(1): have mortgage, term=<12; OR have mortgage, time at address<78 months, have current account

(2): others

(3): loan<4320, insurance accepted

Results from mixture distribution models (Recovery Rate)

Method 1	R square	Spearmen Rank Coefficient	MAE	MSE
Linear regression	0.0840	0.28544	0.3693	0.1688
Accelerated	0.0660	0.26625	0.3549	0.2055
Cox-including 0	0.0752	0.28581	0.3518	0.1967
Cox-excluding 0	0.0636	0.26236	0.3549	0.2067

Method 2	R square	Spearmen Rank Coefficient	MAE	MSE
Linear regression	0.0734	0.26453	0.3695	0.1688
Accelerated	-	-	-	-
Cox-including 0	0.0570	0.25869	0.3588	0.2051
Cox-excluding 0	-	-	-	-

UNIVERSITY OF Southampton
School of Management

Results from mixture distribution models (Recovery Amount)

Method 1	R square	Spearmen Rank Coefficient	MAE	MSE
Linear regression	0.1942	0.31824	1166.7	2593870
Accelerated	0.1346	0.31820	1102.3	3030185
Cox-including 0	0.1574	0.35314	1100.5	2976283
Cox-excluding 0	0.1357	0.31564	1105.8	3068188

Method 2	R square	Spearmen Rank Coefficient	MAE	MSE
Linear regression	0.2054	0.31356	1169.4	2564149
Accelerated	-	-	-	-
Cox-including 0	0.1669	0.33888	1125.7	2930725
Cox-excluding 0	-	-	-	-

Model comparison (Recovery Rate)

		R square	Spearmen Rank Coefficient	MAE	MSE
Single distribution model	Linear regression	0.0904	0.29593	0.3682	0.1675
	Cox-including 0	0.0673	0.27261	0.3546	0.2006
Mixture distribution model Method 1	Linear regression	0.0840	0.28544	0.3693	0.1688
	Cox-including 0	0.0752	0.28581	0.3518	0.1967
Mixture distribution model Method 2	Linear regression	0.0734	0.26453	0.3695	0.1688
	Cox-including 0	0.0570	0.25869	0.3588	0.2051

UNIVERSITY OF Southampton
School of Management

Model comparison (Recovery Amount)

		R square	Spearmen Rank Coefficient	MAE	MSE
Single distribution model	Linear regression	0.2068	0.32522	1162.4	2549591
	Cox-including 0	0.1628	0.34619	1101.9	2906821
Mixture distribution model Method 1	Linear regression	0.1942	0.31824	1166.7	2593870
	Cox-including 0	0.1574	0.35314	1100.5	2976283
Mixture distribution model Method 2	Linear regression	0.2054	0.31356	1169.4	2564149
	Cox-including 0	0.1669	0.33888	1125.7	2930725

Conclusion and Further Research

- Linear regression is better than Survival analysis models for modelling LGD for unsecured consumer loans
- The prediction of recovery amount from RR model is better than that from recovery amount model
- Mixture distribution model does not improve prediction accuracy
- But linear regression is still poor, some magical variables are missing?
- How to segment the population? Cluster analysis?