APPLICATION OF SURVIVAL MODELS TO ANALYSE DEFAULT RATES ON BANK LOANS

Fhatuwani Nemakhavhani  Karabo Mofomme
Liberty Holdings Pty(Ltd)  Financial Services Board (FSB)
AGENDA

1. Aim of research
2. Credit Scoring Model
3. Factors influencing default rates
4. Types of survival models
5. Data analysis
6. Application of models and results
7. Conclusion
INTRODUCTION

• The Credit Bureau Monitor (CBM) shows that 9.59m out of 20.21m credit active consumers had impaired records for the quarter ended in June 2013.

• Stricter measures are needed by financial providers (such as banks) to address the above credit problem.

• For example - tighter credit assessments processes when issuing credit could be implemented.
INTRODUCTION (CONTINUED)

• History of credit risk models:
  • First generation models:
    • Black-Scholes option model, single factor term structure models, the Jamshidian Bond Option model
  • Second generation models:
    • Merton Model of risky debts, the Monte Carlo simulation etc.
• The use of survival models:
  • Mortality investigations
  • Default rates on loans
AIM OF THE RESEARCH PROJECT

• Investigating factors which influence default rates

• Applying survival models:
  • Indicate good and bad-risk lenders
  • Calculate the probability of surviving to a specified duration
  • Calculate default rates on bank’s personal loans

• Projection of default rates
  • Using a survival model
CREDIT SCORE MODEL (CSM)

- Used to differentiate between good-risk and bad-risk lenders
- Data mining techniques are used in model
- SCM is used together with the Logistic Model
  - Define Random Variable $Y_i$: Random variable indicator for the $i$th borrower
  
  $Y_i = \begin{cases} 
  0 & \text{if no default} \\
  1 & \text{if default} 
  \end{cases}$

  - The probability of default is given by:

  $$P(Y_i = 1|X_{1i} = x_1, X_{2i} = x_2, \ldots, X_{ni} = x_n) = \frac{1}{1 + e^{-(\alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n)}}$$
FACTORS INFLUENCING DEFAULT RATES

• Age
• Occupation
• Marital status
• Amount borrowed
• Gender
• Number of dependents
• Residential area
• Education level
• Unemployment rate
• Method of payment
• Date since employment
• Loan term

*Source: http://dx.doi.org/10.1080/02642060903437535
SURVIVAL MODELS

The use of survival models in modelling human mortality:

• When applying survival analysis on loans, assumptions relating to the following need to be made:
  • Default as opposed to death
  • Uncertainty of and timing of a loan’s “death”
  • Survival time
  • Censoring a loan
SURVIVAL MODELS (CONTINUED)

• Generalised definitions used:
  
  • T: is the continuous random variable for the survival time of a loan
  
  • Default
    
    • An account with an arrear bucket of at least 3 in the first twelve month is considered to have defaulted. Where arrear bucket is defined as the arrear value divided by the original instalment value
  
  • H(t) Hazard function: is the instantaneous potential per unit time of a failure to occur given that the loan has not defaulted at time t
  
  • S(t)=P(T>t): is the survival function and F(t)=1-S(t): is the probability distribution function
TYPES OF SURVIVAL MODELS

- Parametric models:
  - Weibull proportional hazard model

- Non parametric models:
  - Kaplan Meier
  - Nelson Aalen

- Semi Parametric models:
  - General proportional hazard model
PARAMETRIC MODEL- WEIBULL PROPORTIONAL HAZARD MODEL

• Assume a hazard to follow a particular distribution:
  • Weibull proportional hazard model:

The hazard function is given by \( h(t) = \lambda t^{\gamma-1} \) where \( t > 0 \) and the two parameters \( \lambda \) and \( \gamma \) are also positive.

The corresponding density function will be given by \( f(t) = \lambda t^{\gamma-1}e^{-(\lambda t)^\gamma} \).

Using this form of the hazard function, we can now find the survivor function as follow:

\[
S(t) = 1 - \int_{0}^{t} f(u)\,du
\]

## NON PARAMETRIC MODEL: KAPLAN MEIER

- KM model does not assume any underlying distribution
- Default M: the number of defaults observed at each duration
- Ordered failure time t: the observed time until an event (censoring and/or default) occur

<table>
<thead>
<tr>
<th>Ordered Failure times((T))</th>
<th>Number of borrowers exposed to risk</th>
<th>Default ((M))</th>
<th>Censoring ((C))</th>
<th>Hazard estimate (h_r)</th>
<th>Estimated survival function (S(t))</th>
</tr>
</thead>
<tbody>
<tr>
<td>(t(0) = 0)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>(h_1 = \frac{m_1}{R(t(1))})</td>
<td>(S(t) = 1 - h_1)</td>
</tr>
<tr>
<td>(t(1))</td>
<td>(R(t(1)))</td>
<td>(m_1)</td>
<td>(c_1)</td>
<td>(h_2 = \frac{m_2}{R(t(2))})</td>
<td>(S(t) = \prod_{r \leq 2}(1 - h_r))</td>
</tr>
<tr>
<td>(t(2))</td>
<td>(R(t(2)))</td>
<td>(m_2)</td>
<td>(c_2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(t(k))</td>
<td>(R(t(k)))</td>
<td>(m_k)</td>
<td>(c_k)</td>
<td>(h_k = \frac{m_k}{R(t(k))})</td>
<td>(S(t) = \prod_{r \leq k}(1 - h_r))</td>
</tr>
</tbody>
</table>
Thus the survival distribution will be as follows,

\[
S(t) = \begin{cases} 
1 & \text{for } 0 \leq t < t_1 \\
S(t_1) & \text{for } t_1 \leq t < t_2 \\
\vdots \\
S(t_{k-1}) & \text{for } t_{k-1} \leq t < t_k \\
S(t_k) & \text{for } t \geq t_k
\end{cases}
\]

Note: \( t_k \) does not need to be the time of the last default of the loans in our investigation but the last default observed during the period of observation.

Source: By D.G. Kleinbaum and M. Klein, 1996, Survival Analysis
NON PARAMETRIC MODEL: NELSON ALEN

- Similar to the Kaplan Meier
- Only difference is that hazard function takes into account the continuous and the discrete probability of defaulting
- The survival distribution function of the Nelson Aalen is assumed to be exponentially distributed
- Survival distribution function is as follows:

\[
\hat{\Lambda}(f) = \sum_{t(j) \leq f} \hat{h}_j \quad \text{and} \quad \hat{S}(t) = \exp\{-\hat{\Lambda}(f)\} \quad \text{for} \quad t(f) \leq t \leq t(f+1)
\]
SEMI PARAMETRIC MODEL: GENERAL PROPORTIONAL HAZARD MODEL

- General proportional hazard model:
  - Define hazard function: the instantaneous potential per unit time of a failure to occur given that the loan has not defaulted at time t;

\[
h(t) = \lim_{\Delta t \to 0} \left\{ \frac{P\left( \left\{ T_i \in [t + t + \Delta t] | T_i \geq t \right\} \right)}{\Delta t} \right\}
\]

\[
h_i(t) = h_0(t) \cdot \phi(x_i) = h_0(t) \cdot \exp(\beta \cdot x_i) = h_0(t) \cdot \exp(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)
\]

- Distribution of survival function:

\[
S(t) = P(T \geq t) = \exp \left\{ - \int_0^t h(z)dz \right\} \quad \forall \ t > 0
\]
**DATA ANALYSIS (DATA EXTRACT)**

- Data for the study comes from the retail banking industry

- Based on short to long term loans (6 to 240 months)

<table>
<thead>
<tr>
<th>Obs</th>
<th>Loanref</th>
<th>applicationdate</th>
<th>bo_term</th>
<th>Org_Bond</th>
<th>r_calc_bal</th>
<th>Def_Within12</th>
<th>Def_Flag_Ever</th>
<th>gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1465616</td>
<td>20070601</td>
<td>12</td>
<td>2285</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>F</td>
</tr>
<tr>
<td>2</td>
<td>1465635</td>
<td>20070601</td>
<td>24</td>
<td>10790</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>F</td>
</tr>
<tr>
<td>3</td>
<td>1465636</td>
<td>20070601</td>
<td>36</td>
<td>18890</td>
<td>66.51</td>
<td>0</td>
<td>0</td>
<td>M</td>
</tr>
<tr>
<td>4</td>
<td>1465641</td>
<td>20070601</td>
<td>24</td>
<td>11140</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>M</td>
</tr>
<tr>
<td>5</td>
<td>1465656</td>
<td>20070601</td>
<td>24</td>
<td>5650</td>
<td>23.73</td>
<td>0</td>
<td>0</td>
<td>M</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Obs</th>
<th>occupation</th>
<th>Marital_Status</th>
<th>Age_Days</th>
<th>numberof_dependents</th>
<th>Annual_Net_Salary</th>
<th>Resid_Suburb</th>
<th>DaysSinceEmp</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CLERK</td>
<td>Married COP</td>
<td>11196</td>
<td>0</td>
<td>53688</td>
<td>SOWETO</td>
<td>1065</td>
</tr>
<tr>
<td>2</td>
<td>TEACHER</td>
<td>Single</td>
<td>9712</td>
<td>0</td>
<td>80329.08</td>
<td>BRAAMFISCHER</td>
<td>151</td>
</tr>
<tr>
<td>3</td>
<td>POLICEMAN</td>
<td>Single</td>
<td>15984</td>
<td>3</td>
<td>54612</td>
<td>TEMBISA</td>
<td>2677</td>
</tr>
<tr>
<td>4</td>
<td>CARETAKER</td>
<td>Single</td>
<td>21508</td>
<td>6</td>
<td>48395.28</td>
<td>FLORIDA</td>
<td>13952</td>
</tr>
<tr>
<td>5</td>
<td>OPERATOR</td>
<td>Married COP</td>
<td>13810</td>
<td>.</td>
<td>54012</td>
<td>MALELANE</td>
<td>5630</td>
</tr>
</tbody>
</table>
## DATA ANALYSIS - DESCRIPTIVE STATISTICS

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>Std Dev</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>numberofdependents</td>
<td>1972373</td>
<td>1.35</td>
<td>1.30</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>Def_Within12</td>
<td>1972373</td>
<td>0.11</td>
<td>0.31</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Def_Flag_Ever</td>
<td>1972373</td>
<td>0.18</td>
<td>0.39</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Age</td>
<td>1972373</td>
<td>39.13</td>
<td>10.22</td>
<td>17</td>
<td>80</td>
</tr>
<tr>
<td>Loan_Term</td>
<td>1972373</td>
<td>37.11</td>
<td>15.32</td>
<td>6</td>
<td>240</td>
</tr>
<tr>
<td>genders</td>
<td>1972373</td>
<td>1.58</td>
<td>0.49</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Event_Flags</td>
<td>1972373</td>
<td>2.11</td>
<td>1.01</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Provinces</td>
<td>1915367</td>
<td>4.77</td>
<td>2.49</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>Annual_Net_Salaries</td>
<td>1972372</td>
<td>2.42</td>
<td>1.54</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>Loan_Amounts</td>
<td>1972373</td>
<td>1.20</td>
<td>0.49</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>Outstanding_Balances</td>
<td>1972372</td>
<td>1.11</td>
<td>0.36</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Inception_Dates</td>
<td>1972373</td>
<td>2010.18</td>
<td>1.74</td>
<td>2007</td>
<td>2013</td>
</tr>
<tr>
<td>YearSinceEmps</td>
<td>1921006</td>
<td>2.67</td>
<td>1.95</td>
<td>1</td>
<td>7</td>
</tr>
</tbody>
</table>
Default frequency by inception date

Inception Date

Number of Default
Crude default rates
Default rate by age
APPLICATION OF LOGISTIC MODEL

• A logistic regression model was used

• The following factors were used
  • Inception date, Age, Gender, Loan term, Loan Amount, Outstanding balance, Year since Employment, Province, Annual Salary, Marital State, Number of Dependents, Annual salary

\[
\text{Def}_{\text{Within12}} = \beta_0 + \beta_1 \text{Inception\_Date} + \beta_2 \text{Age} + \beta_3 \text{Gender} + \beta_4 \text{Loan\_Term} + \beta_5 \text{Loan\_Amount} + \beta_6 \text{Outstanding\_Balance} + \beta_7 \text{YearsSinceEmp} + \beta_8 \text{Province} + \beta_9 \text{Marital\_Status} + \beta_{10} \text{NumberOfDependents} + \beta_{11} \text{Annual\_Salary} + \beta_{12} \text{Event\_Flag} + \varepsilon
\]

Where $\beta_0$ is the intercept and $\beta_0 \ldots \beta_{12}$ are the parameters of the regression model and $\varepsilon$ is the error term.
### APPLICATION OF LOGISTIC MODEL (RESULTS)

- The output from SAS

<table>
<thead>
<tr>
<th>Parameter</th>
<th>DF</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>Wald Chi-Square</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>570.5</td>
<td>3.6449</td>
<td>24499.6275</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Inception_Dates</td>
<td>1</td>
<td>-0.2816</td>
<td>0.00181</td>
<td>24150.4567</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Age</td>
<td>1</td>
<td>0.0154</td>
<td>0.000325</td>
<td>2238.2394</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Loan_Term</td>
<td>1</td>
<td>0.000215</td>
<td>0.000209</td>
<td>1.0521</td>
<td>0.305</td>
</tr>
<tr>
<td>Loan_Amounts</td>
<td>1</td>
<td>1.8357</td>
<td>0.0215</td>
<td>7300.0197</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Outstanding_Balances</td>
<td>1</td>
<td>-2.4686</td>
<td>0.0219</td>
<td>12687.1201</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>YearSinceEmps</td>
<td>1</td>
<td>0.1759</td>
<td>0.00204</td>
<td>7414.535</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Event_Flags</td>
<td>1</td>
<td>-1.2211</td>
<td>0.00281</td>
<td>188469.882</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Provinces</td>
<td>1</td>
<td>-0.00949</td>
<td>0.00106</td>
<td>80.3726</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Annual_Net_Salaries</td>
<td>1</td>
<td>0.204</td>
<td>0.00258</td>
<td>6261.7283</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Marital_State</td>
<td>1</td>
<td>-0.0536</td>
<td>0.00165</td>
<td>1050.2116</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>numberofdependents</td>
<td>1</td>
<td>0.0895</td>
<td>0.00232</td>
<td>1490.0764</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>
APPLICATION OF LOGISTIC MODEL (RESULTS)

• The following factors were found to be significant at a 1% level of confidence:
  • Inception date
  • Age
  • Gender
  • Loan amount
  • Outstanding balance
  • Years since employment
  • Province
  • Annual Net salary
  • Marital State
  • Number of dependents

• The above result confirms our literature study
APPLICATION OF LOGISTIC MODEL (RESULTS)

• The following factor was found to be insignificant at a 1% and 5% level of significance:
  • Loan term

• The above conclusion is not supported by our literature study.
  • Which shows that loan term significantly affects default rates.
APPLICATION OF LOGISTIC MODEL
(LOAN TERM)

Number of borrowers by Loan Term

Number of default by Loan Term

Frequency of default
APPLICATION OF LOGISTIC MODEL (LOAN TERM)

• Possible reasons for Loan term to not have an effect on default:
  
  • Majority of borrowers in this study have a loan with a term which is around 37 months
  
  • Majority of defaults are from people with loans of term which is around 36 months
  
  • This results in Loan term having no significant effect on default since the average population Loan term contributes to the average defaults
APPLICATION OF KAPLAN MEIER MODEL (ASSUMPTIONS)

• Assumptions used:
  • We have modeled two areas using the KM model
  • Investigation 1:
    • Investigation period is 1 Jan 08 to 1 Jan 11
  • Investigation 2:
    • Investigation period is from 1 Jan 2011 to 31 May 13
  • We do not allow re-entering the population once you have exited the population
APPLICATION OF KAPLAN MEIER MODEL (ASSUMPTIONS CONTINUED)

• Assumptions used:
  
  • Right censoring is used in our investigations
  
  • Censoring occurs as follows: Written Off, completion of loan term and end of investigation

• Result from using the above assumption:
  
  • We had over 2 million data entries recorded from 1 June 2007 to 31 May 2013
  
  • Investigation 1: Resulted with roughly 195,000 entries, i.e. roughly 9.75% of the entire dataset
  
  • Investigation 2: Data entries were 480,000 entries, i.e. roughly 24% of the data
• Using section 1 of our data to model default rates, using the KM method
• Period of investigation is 36 months (3 years)
• Goodness of fit test showed us that we cannot reject the hypothesis that the KM curve from historic data can be used as estimate of future default patterns.
CONCLUSION

• Usefulness of results in the industry:
  • Credit scoring when granting a loan
  • Pricing for a new loan
  • Setting loan conditions such as what security to call for and what term to grant the loan
  • For raising impairment charges and when reserving for capital adequacy purposes to protect the bank

• Implementation of survival models by banks:
  • As an actuarial technique to assess probability of defaults when granting bank loans
  • Suitable to use to estimate population default rate patterns
CONCLUSION (CONTINUED)

• Challenges faced:
  • Sorting the data for modelling, both for the Kaplan Meier and the Regression model
  • Importing data into SAS and Excel
  • Computer running speed too slow
  • Some variable were too broad to group
    – e.g. Occupation (Over 200 different occupation types in the data).
  
• Further research can be done on the application of survival models on bank loans.
ACKNOWLEDGMENTS

• Michael Tichareva from Banking and Finance Committee of Actuarial Society
• Dr Conrad Beyers from the University Of Pretoria
• Thehan Claassen from the banking industry
• Antonie Jagga from PWC
• Quinton Lancaster from the banking industry
QUESTIONS

Fhatuwani Nemakhavhani (Liberty Holdings Pty(Ltd))

Karabo Mofomme (Financial Services Board (FSB))