Peaks2Tails

www.peaks2tails.com

A regression and survival hybrid LGD model for unsecured and secured loans

Satyapriya Ojha (FRM, CQF)

Karan Aggarwal(CFA L3 cleared,FRM,CQF)

Contents

Abstract

Loss given default(LGD) is one of the key component banks need to estimate for expected and unexpected loss calculations. The LGD modeling poses several challenges in practice on top of scarcity of recovery data. First, LGD is significantly more involved than PD or EAD as one has to incorporate both default definition and post default term structure of balance outstanding as and when recoveries come through. Second, LGD needs to be estimated for a loan under a multitude of plausible states i.e., Performing, Defaulted, Charged-off etc. Third, one also has to address the possibility of 'cure' where a defaulted loan turns back to performing. Fourth, it also becomes paramount to consider the time spent in the above states as a key determinant for future recoveries. Fifth, one also has to account for cost of recovery operations and suitable discount rate to calculate present value of the recovery stream to arrive at LGD estimate. Sixth, due to long drawn-out nature of recovery data, bank ends up with incomplete recovery cases in the development and validation samples which cannot be ignored as they contain partial recovery data.

This paper aims to address the above key challenges using a three-component model, one for each sub-population discussed above by using a hybrid methodology that combines traditional regression technique and survival analysis. The model estimates feed into one another to arrive at the final LGD estimate. With these models, a bank can estimate LGD for a loan in any of the above states. We will also discuss how to test calibration for the LGD estimates.

Data and Methodology

1. Data

For our study, we collected Fannie Mae single family loan data which has the following data points in panel form.

- Account Id
- Reporting Date
- Loan Age
- FICO
- DTI
- LTV

- Balance Outstanding
- Delinquency
- Sold property value

In Section A, we consider unsecured loans and in section B we focus on mortgage loans. Therefore, for unsecured loans, we drop the *LTV* and *sold property value* columns. Instead, we simulate the recovery data post default. For section B, that focusses on LGD for Mortgage loans, we retain the original data format.

We define our sample data from 20/1/2008 to 20/1/2016. We further split this data to *development* and *validation* sets as given in Table 1 below. Note that *In-Sample* and *Out-of-Sample* data sets utilize the same time window i.e., 20/1/2008 to 20/1/2014 that is split in 70 : 30 ratio whereas the *out-of-time* data set has a window from 20/2/2014 to 20/1/2016. The *In-sample* data set is used for development and the other two sets are used for validation.

Data Set	Start Date	End Date
In-Sample (70%)	20/1/2008	20/1/2014
Out-of-Sample (30%)	20/1/2008	20/1/2014
Out-of-Time	20/2/2014	20/1/2016

Table 1

2. Methodology Overview

LGD modeling is highly involved and requires meticulous handling of various stages of a loan life cycle (Refer to the Figure 1 below). At any given point in time a bank will have some arbitrary composition of loans in these various stages. We need to develop flexible models that can provide LGD estimate for any loan not only based on what stage it is but how long it has been in that stage.

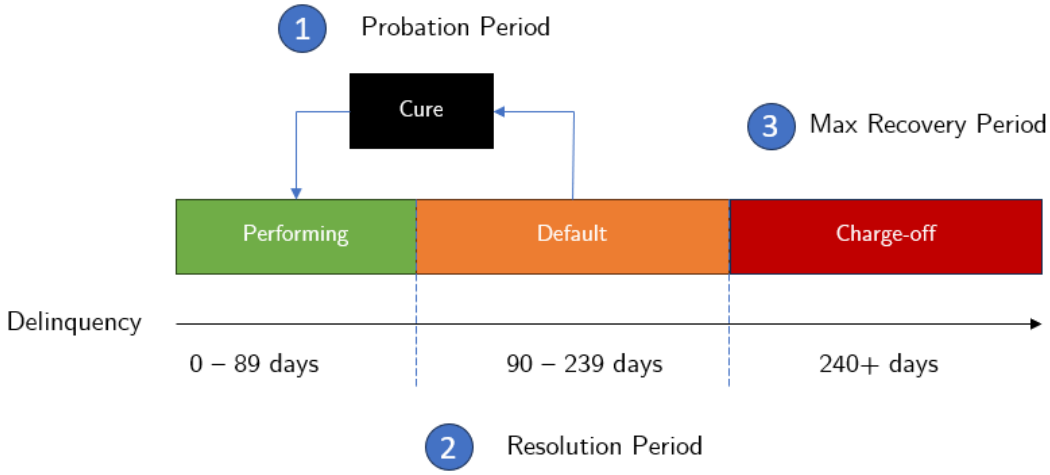

Figure 1

In order to model LGD, we need to first understand the interplay between various stages and define timelines of various events in the lifecycle.

The first sub-population that we need to calculate LGD estimate for are the loans in non-defaulted status (Performing) which can be either non-delinquent or delinquent up to 89 days. This is as per BASEL III guidelines of default definition as 90 days delinquent. Since BASEL III requires a bank to calculate 12-month PD for performing loans for capital calculations, it is important to have LGD estimate for loans that are currently performing but can default within next 12 months. When we take PD x LGD x EAD , the term PD x LGD will capture the conditional loss % of a performing loan should it default within next 12 months.

The second sub-population that we need to consider is the already defaulted population. For ex-ante prediction of loss, the PD is taken as 1, and so the LGD represents the loss% on of the exposure at default. LGD by definition is measured as the present value of %loss of the exposure at default. Now, since some loans would have already spent some time in the defaulted state, it's important to capture only the future losses (= 1 – future recoveries) as a percentage of the balance outstanding at *observation* (*not the balance at default*). This is to isolate the recoveries that would have already happened between default and observation. We also need to address the possibility of *Cure* where a loan returns back to performing status if it pays of all the arrears. Typically, banks define a *probation period* under which a default flag of a loan that pays off all the arrears is not lifted until the loan completes the probation period. We will perform the analysis to arrive at a *probation period* later in the paper. From the defaulted state, the loan can end up in one of the following states

- Closed (account pays off outstanding balance and gets closed)
- Charged – Off (No Cure is possible, start recovery)
- Cure
- Default (Unresolved)

We consider a *resolution time* period which is basically the time taken for most of the defaulted accounts to reach a final outcome other than unresolved. This improves the efficiency of modeling to a great extent as we only need to look ahead the *resolution period* to capture the recoveries till the final outcome (*charge-off* or *cure*) is reached. Based on the outcome we can plug in the LGD estimate from that onwards with suitable discounting. Now, if outcome is 'Cure' , we assume a recovery of 100% at the point of cure. On the other hand, if the outcome is 'charge-off', we substitute the LGD estimate for a charged-off state based in risk data set available at the point of charge-off discounted back to observation point. This saves us following the loan through the months post charge-off.

The third and final sub-population is the *charged-off* population. The charge-off state refers to a point from where no cure is possible. This is point where the long-drawn recovery process begins. Since at any point in time there will be loans in charge-off state already, it's important to consider the time already spent in this state and to only model the remaining losses as a fraction of the outstanding balance at *observation*. Also, since the recovery process is a long-drawn process, we need to define a

conservative time window that captures majority of the recoveries called as 'Max Recovery Period'. This is when the cumulative recovery curves saturate.

The following sections give a detailed account of the model development an validation followed by conclusion.

A. Unsecured Loans

1. Model

As discussed, we will be developing three separate models for the three sub-populations i.e. *Performing*, *Default*, *Charge-off* and chaining the models together to arrive at the LGD estimate. Figure 2 below shows a detailed schematic of the modelling approach.

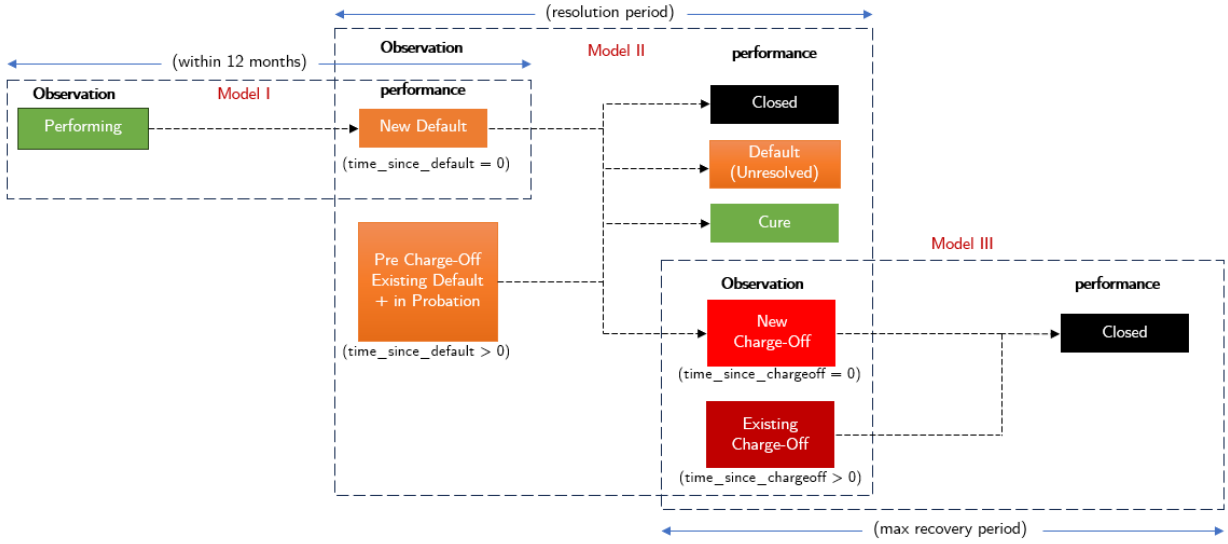

Figure 2

For the above 3 models given in the schematic, a monthly frequency was chosen for observation and a performance window as applicable for each of the model.

1.1 Model I (Performing at observation, defaulting within 12 months)

The first model (Model I) refers to the LGD estimates for loans which are performing at observation but default within 12 months. During Model development, we considered 12 month rolling performance window at monthly intervals and included only those loans which have performing status at observation but has defaulted within 12 months from observation.

Let's define the following variables

$t_0 = observation\ point$

$t^* = default\ time\ ,\ t_0 < t^* < t_0 + 12$

$X(t_0) = risk\ set\ at\ observation\ (\ input\ for\ Model\ I)$

$B(t_0) = Balance\ at\ observation$

$y(t_0) = LGD\ estimate\ (\ target\ for\ Model\ I)$

$X(t^*) = risk\ set\ at\ Default\ (\ input\ for\ Model\ II)$

$B(t^*) = Balance\ at\ Default$

$y(t^*) = LGD\ estimate\ (\ target\ for\ Model\ II)$

$r\ = suitable\ discount\ rate$

The risk data $X(t_0)$ for Model I consisted of *Loan Age*, *FICO*, *DTI* available at the observation point (t_0). The target $y(t_0)$ is the LGD estimate applicable which is derived from Model II estimate of LGD as per the following logic.(Refer to *Figure 3*)

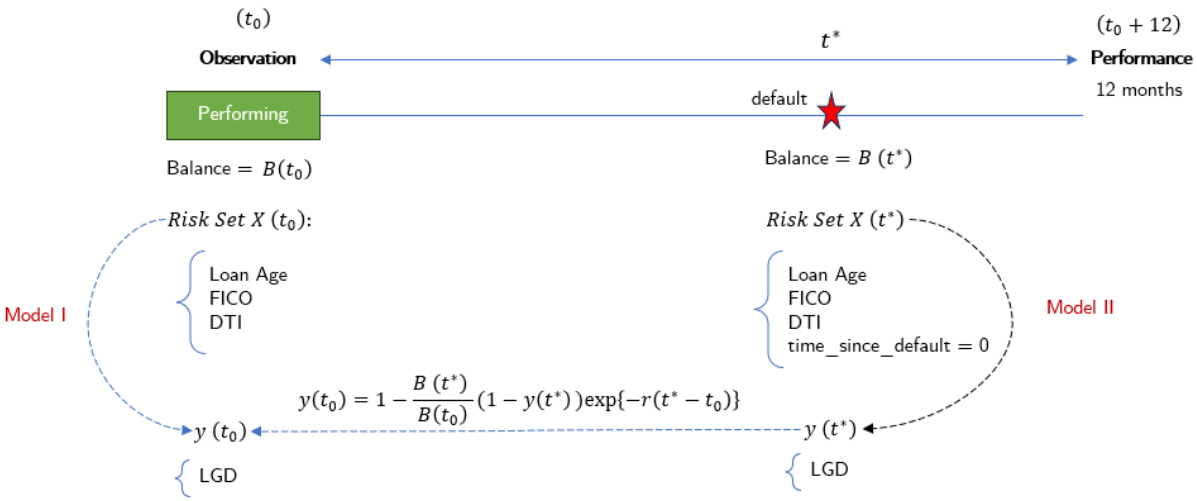

Figure 3

Use Model II with the risk set $X(t^*)$ with time_since_default $= 0$ to get the LGD estimate $y(t^*)$. Calculate the PV of future recoveries from t^* as $B\ (t^*)(1 - y(t^*)$. Discount this further back to the point t_0 that is $B\ (t^*)(1 - y(t^*)\ )exp\{-r(t^* - t_0)\}$. Finally, the LGD estimate

$$y(t_0) = 1 - \frac{B\ (t^*)}{B(t_0)}(1 - y(t^*)\ )\exp\{-r(t^* - t_0)\}$$

The above process allows us to only follow the account till default and not beyond as we could substitute ready estimates from Model II for recoveries beyond this point. We will discuss the estimation procedure for Model II subsequently.

For Model I we chose multiple linear regression

$$y(t_0) = \beta_0 + \beta_1 * Loan\ Age + \beta_2 * DTI + \beta_3 * FICO + \epsilon$$

1.2 Model II (pre-Charge-off default)

Model II allows us to estimate LGD for already defaulted loans that are not yet charged off. Accordingly, we include only such accounts that have default flag turned on at observation. This includes newly defaulted accounts and existing defaulted accounts not yet charged off. A defaulted loan can be closed, remain in default(unresolved), get cured or enter into charge off state. Of these, we consider the status to be resolved if it gets either closed, gets cured or gets charged off. To increase efficiency of the procedure, a *resolution window* was defined and estimated so that we only follow the account till the resolution period. If the Outcome is 'Charge Off' then similar to Model I, we substitute the estimate from Model III by setting (time_since_charge_off = 0). If the outcome is 'Cure', we assume 100% recovery at the point of cure. Figure 4 (Charge Off) and Figure 5 (Cure) illustrate the same.

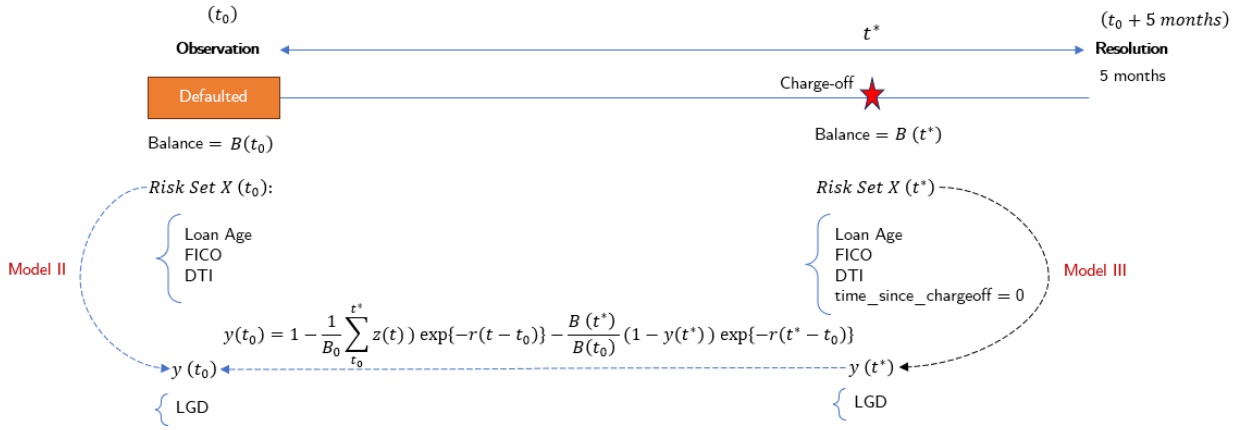

Figure 4

The variables have similar meanings as described in Model I. Note that the for LGD estimate to be used as target for Model II, the present value of all collections in between observation and charge off needs to be incorporated as well.

$$Z(t) = collections\ at\ time\ t\ , t_0 < t < t^*$$

For Charge off case, we have

$$y(t_0) = 1 - \frac{1}{B_0}\sum_{t_0}^{t^*} z(t)\,)\,\exp\{-r(t-t_0)\} - \frac{B\,(t^*)}{B(t_0)}(1 - y(t^*)\,)\,\exp\{-r(t^*-t_0)\}$$

For Cure case, we have

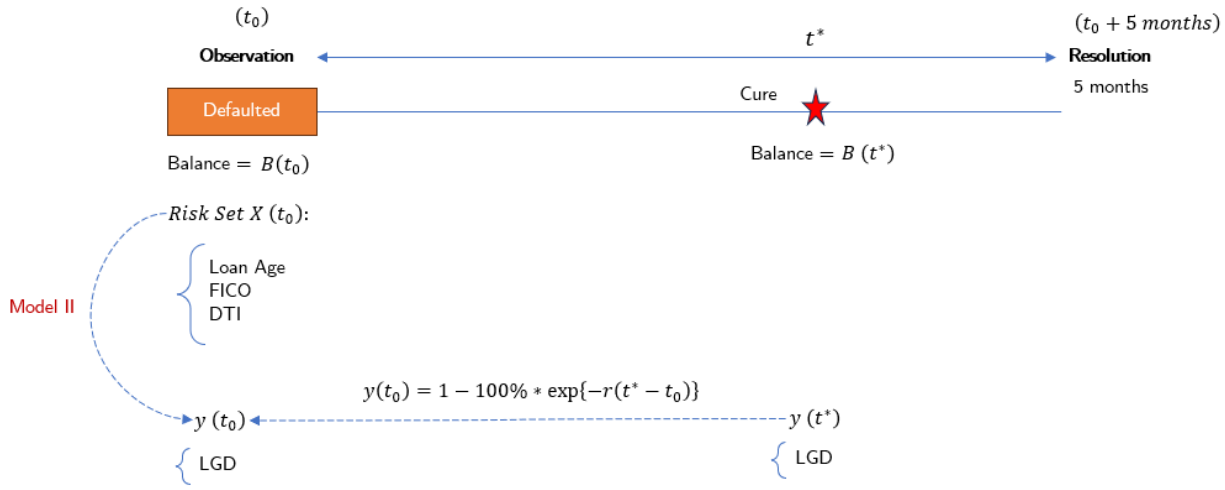

Figure 5

1.2.1 Definition of Cure

Banks practices for cure vary from 'instant' to 'probationary'. The idea of instant cure is that as soon as a defaulted loan pays off all the arrears, the default status is lifted instantly. Probationary cure on the other hand involves a monitoring period where the loan is observed for a certain period and the default status is retained even if the loan has paid all the arrears. If the account does not default again till the end of the period, it's considered as cured and it returns back to performing status.

For our analysis, we have followed the probationary cure approach. Now, to determine the appropriate length of the probation period, we try out with different length of probation period (3m, 6m, 9m and 12m) and examined the re-default rate of each. We chose the minimum length that results in a lower re-default rate as compared to the delinquent non-defaulted population (which is the riskiest performing segment). The idea is to what can be called as Cure must exhibit the properties of performing segment and therefore conservatively it should at least be better than riskiest segment of the performing population i.e., accounts with delinquency <= 89 days.

In our data, we found the 3m definition as adequate for cure as the re-default rate is lower than the delinquent segment. (Refer to Table 2 and Figure 6)

Segment	2008	2009	2010	2011	2012	2013	2014	2015	2016
Delinquent	0.565123	0.672269	0.662877	0.684667	0.67837	0.629814	0.583589	0.565439	0.488355
3m(probation)	0.468085	0.385965	0.111111	0.157303	0.183673	0.102564	0.033333	0.111111	0.133333
6m(probation)	0.285714	0.333333	0.13913	0.150943	0.111111	0.125	0.054054	0.111111	0.076923
9m(probation)	0.272727	0.333333	0.119403	0.132813	0.08	0.081081	0.078947	0.166667	0
12m(probation)	0.066667	0.384615	0.113636	0.150376	0.068966	0.058824	0.081081	0.1	0.066667

Table 2 *(re-default rates for various segments)*

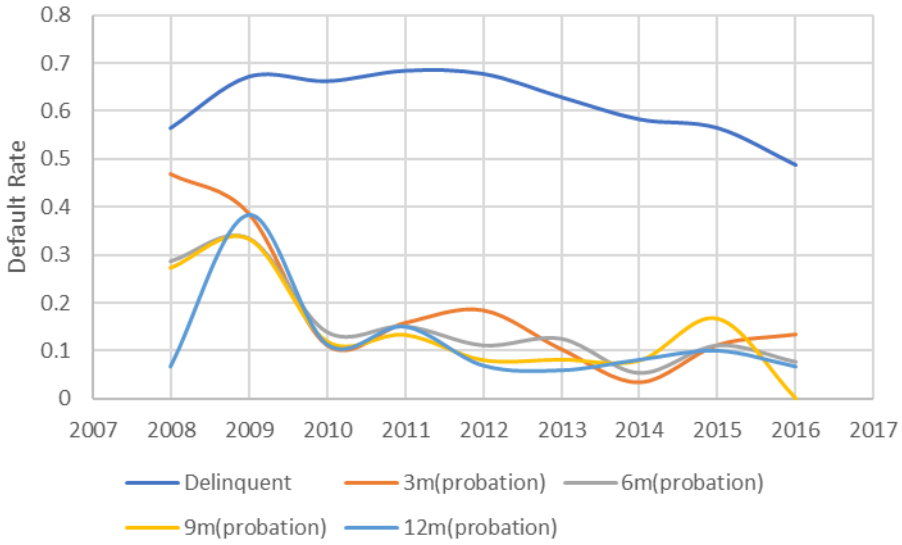

Figure 6

1.2.2 Resolution Period

To determine the resolution period, we monitored the cumulative count of defaulted accounts that move into a resolved state i.e. closed, charged off or cured on a month by month basis for various vintages. We found the resolution period of 5 months to be adequate for most loans to resolve.(Refer to Figure 7). This definition is important for two reasons; 1) we can define the delinquency to charge off as 90 days + 150 days = 240 days , 2) we need to follow through the recoveries / collections of a defaulted loan maximum till the resolution period, so we most definitely encounter a resolved state. Accordingly, we can use estimates from Model III if the outcome is charged off as targets for Model II.

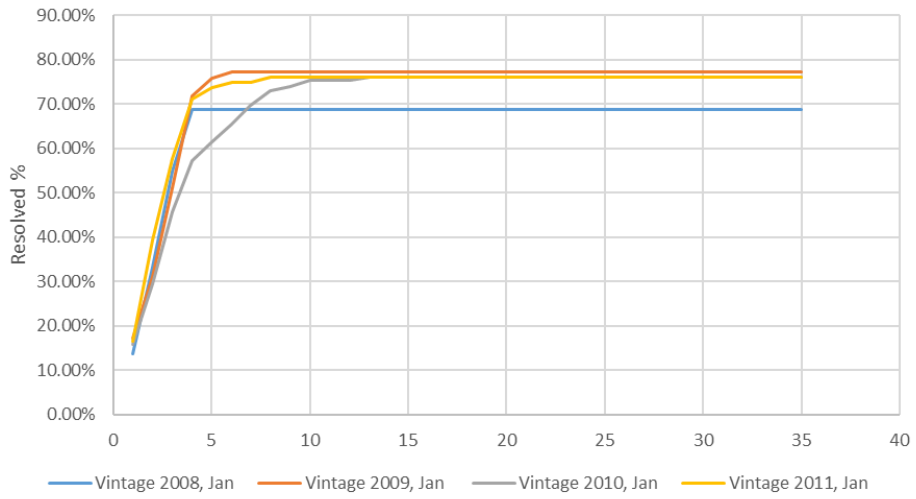

Figure 7

As far as Model II is concerned, we have used decision tree regression. A critical input variable in Model II is *time_since_default*. Decision tree regression in contrast to simple linear regression can accommodate non-linearities in the data and provides natural segmentation based on time_since_default. Refer to Results section for estimates.

1.3 Model III (post Charge-off)

Model III concerns with long drawn recoveries post charge off. Because of the long-drawn nature, this stage brings significant modeling challenges because the training and validation sets often consists of incomplete recovery cases. Banks that use traditional regression methods often utilize chain and ladder mechanism to extrapolate incomplete recovery cases before they can be used in the model.

In this paper, we chose to use Survival Analysis to model recoveries post charge-off event. Now, survival analysis has several advantages over traditional regression-based models.

- Incorporates time_since_charge_off naturally in the model which is the key input variable. This saves us from worrying about partial recovery cases or any need for extrapolation for that matter as we can match observed partial recovery data with predicted partially recovery based on length of time involved
- Readily available survival regression packages such as COX or AFT incorporate obligor attributes easily in the model

Typical survival analysis concerns primarily with time of study and covariates. In our case, we need to address cost of recovery and discounting the recoveries back to observation date. The cost (c) of

recovery typically is proportional to the amount recovered which is amount paid to the collection agency. For the discount rate (r), we have assumed WACC = 9%. We propose the following dynamics

λ_0: base hazard rate

$\lambda_i = \lambda_0\, exp(\beta_1(Loan\ Age) + \beta_2(DTI) + \beta_3(FICO))$: hazard rate for 'ith' account

$t_{0,i}$ = time since charge off

T_i = last observation point of an account (the data may be censored due to sample window, which will result in partial recovery data which is perfectly acceptable in the model)

We consider an infinitesimal time period dt in which the balance falls from $B(t)$ to $B(t+dt)$ as the consequence of recovery $dR(t) = -dB(t)$. (Refer to Figure 8)

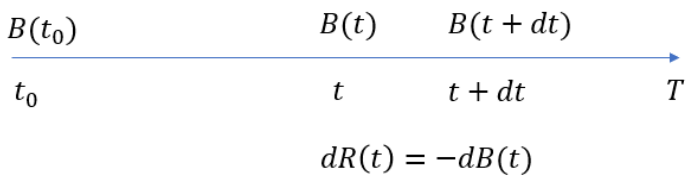

Figure 8

We have the following set of equations

$$\frac{dB}{dt} = -\lambda B$$

$$B(t) = B(t_0)\exp\big(-\lambda(t-t_0)\big)$$

$$dB(t) = -\lambda B(t_0)\exp\big(-\lambda(t-t_0)\big)$$

$$dR(t) = -dB = \lambda B(t_0)\exp\big(-\lambda(t-t_0)\big)$$

$$PV\ of\ net\ dR\%(t) = (1-c)\frac{dR(t) * \exp\big(-r(t-t_0)\big)}{B(t_0)} = (1-c)\lambda\exp\big(-(\lambda+r)(t-t_0)\big)$$

The PV of total recovery can be found by integrating above

$$R(t_0) = (1-c)\frac{\lambda}{\lambda+r}\big(1-\exp\big(-(\lambda+r)(t-t_0)\big)\big)$$

Therefore,

$$LGD(t_0) = 1-(1-c)\frac{\lambda}{\lambda+r}\big(1-\exp\big(-(\lambda+r)(t-t_0)\big)\big)$$

The above gives us a closed form formula that can be readily applied to estimate LGD of an account post charge-off.

For ex-ante prediction of LGD for loans which are already in charge-off state, we need to incorporate the notion of a *maximum recovery period* t_{max} , so that our estimate takes the form

$$LGD(t_0) = 1 - (1 - c)\frac{\lambda}{\lambda + r}\left(1 - \exp\left(-(\lambda + r)(t_{max} - t_0)\right)\right)$$

1.3.1 Maximum Recovery Period

In order to determine maximum recovery period, we observe month by month recovery of charged-off accounts and plot the cumulative recovery. The maximum recovery period is estimated as the recovery period where 95% of the recovery is completed.

In our dataset, we found 40 months to be the maximum recovery period. (Refer to Figure 9). The overall recoveries to stabilize at 40%.

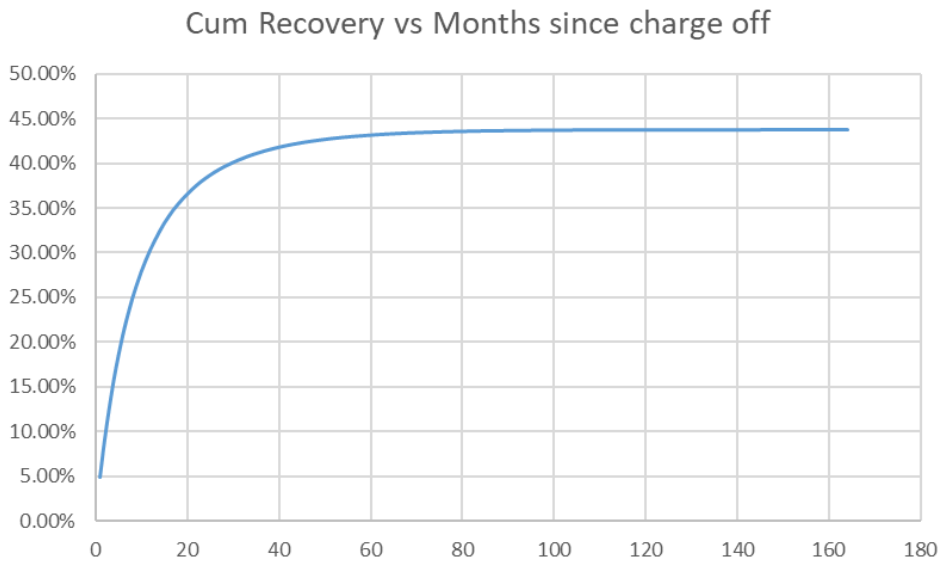

Figure 9

2. Estimation

The estimated parameters for each model are as follows

Model III : Estimates (recovery cost factor c was taken as 5%)

Parameter	estimate
λ_0	49%
DTI	-0.0179
FICO	0.00094

Table 3

For estimation of above β coefficients which make up λ, we have used OLS where we minimize sum of squared errors between observed LGD and fitted $\widehat{LGD}$ figures.

Model II : Estimates

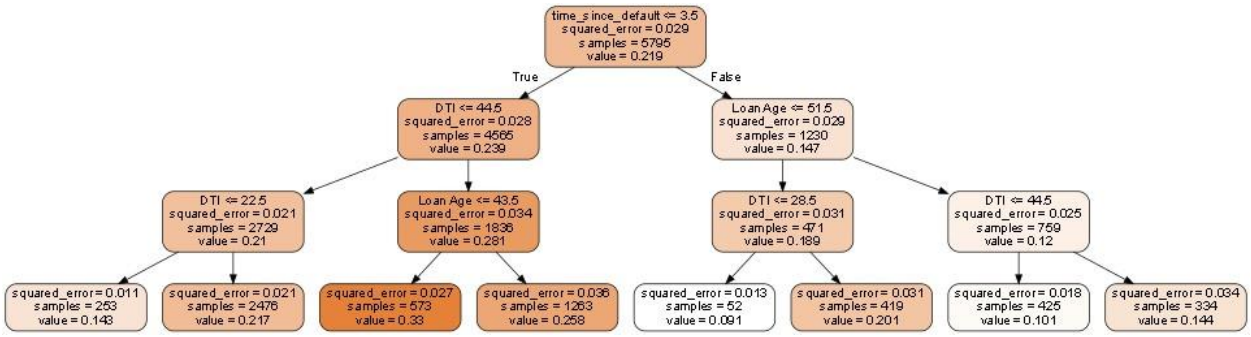

Figure 10

We can see the predominant variable is time_since_default and the least important is FICO score which does not even appear in the tree above.

Model I : Estimates

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.358105052
R Square	0.128239228
Adjusted R Square	0.128042295
Standard Error	0.040926664
Observations	13284

ANOVA

	df	*SS*	*MS*	*F*
Regression	3	3.272158571	1.09072	651.179
Residual	13280	22.24389148	0.001675	
Total	13283	25.51605005		

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	0.199332076	0.004598804	43.34433	0
Loan Age	-0.000381212	1.99522E-05	-19.1063	2.67E-80
DTI	0.001188566	3.11629E-05	38.14038	3.9E-302
FICO	-1.0878E-05	6.44552E-06	-1.68769	0.091494

Table 4

We can see that the overall R^2 of the model is low which means the attributes selected are not enough to explain the variance in LGD. Although, the β coefficients for '*Loan Age*' and '*DTI*' are extremely significant whereas that of FICO is moderately significant.

3. Validation

We perform calibration tests for LGD estimates with out of sample data and out of time data. For calibration we use spearman rank correlation between observed and fitted values of LGD. (Refer to table 5)

	In Sample	Out of Sample	Out of time
Spearman rank correlation	0.89	0.87	0.83
No of observations	838	457	526
t-stat	58.37	38.37	34.56

Table 5