# Table of Contents

1 Introduction ................................................................. 3

2 Corporate PD Models ....................................................... 3  
   2.1 Structural Models .................................................... 4  
   2.2 Regression Model ................................................... 4  
   2.3 Default Intensity Model ............................................ 4  

3 Performance Measure .................................................... 5  
   3.1 Likelihood ............................................................ 5  
   3.2 Goodness of Fit ..................................................... 5  
      3.2.1 Pearson's Chi-Squared Test ................................. 5  
      3.2.2 Deviance Test ................................................ 6  
      3.2.3 Hosmer-Lemeshow Test .................................... 6  
      3.2.4 Challenges .................................................... 6  
      3.2.5 Some Thoughts ............................................... 6  
   3.3 ROC and AR .......................................................... 7  
      3.3.1 ROC and AUC ................................................ 7  
      3.3.2 CAP and AR .................................................. 7  
      3.3.3 Challenges .................................................... 8  
      3.3.4 Some Thoughts ............................................... 8  

4 Research Questions ...................................................... 10
1 Introduction

In the Financial Services industry, credit risk refers to the risk that a borrower does not make timely payments of contractually obligated interest or principle. In essence, a borrower owns a default option, and credit risk analysis concerns assessing the likelihood that the option will be exercised and how to price the option. There are many kinds of borrowers, including corporations, financial institutions, insurance companies, municipalities, sovereign governments, and special purpose vehicles (entities that borrow from investors to finance the purchase a portfolio of credit risky instruments). A typical example is a corporation issuing a bond that promises to pay a fixed coupon (interest) on a regular basis until maturity, at which time the principle is repaid. If the firm runs into trouble, it may not be able to meet this obligation, or it may find it is more advantageous to default on the obligation and go through a restructuring or bankruptcy.

In quantitative finance, in particular for assessing corporate credit risk, it is common to build Probability of Default (PD) models. Based on historical data for attributes of a population of firms (debt levels, earnings, market cap, and volatility, etc.) and default experience, models are calibrated to assign a probability that a firm with specific characteristics will default over a given time horizon (typically one year). PD models are often used to monitor the risk of investment portfolios, set capital requirements for banks, validate qualitative credit risk assessments, or help analyze risk-return properties of potential investments. PD models may also be used for counterparty risk assessment and setting limits or margin requirements.

One of the key challenges associated with PD models is assessing their performance. For high quality credits, the PD over a one year horizon is generally less than 0.5% (often much less), so the number of observed defaults to test performance against may be fairly small. Even the riskier companies issuing new debt tend to have PDs less than 5%. For this reason it is also difficult to calibrate PD levels accurately. Therefore, most performance measures for PD models tend to focus on comparing PD to outcome (default or no default). A good model has the sample defaulters strongly associated with the higher PD scores, and the sample non-defaulters associated with the lower PD scores. Both likelihood methods (e.g. likelihood ratio test) and categorization tests (e.g. Receiver Operating Characteristic curves and Cumulative Accuracy Profiles/Accuracy Ratio measures) are essentially measures of the ability of the model to separate defaulters from non-defaulters, as opposed to a direct measure of whether the PD itself was accurate. For this reason, these performance measures tend to be strongly sample dependent, with models being assessed a superior when the sample includes a mix of highly risky and highly safe names, while for a homogenous (in PD) sample a model could be assessed as poor even if the PD model itself is perfect.

We would like to review the current methodologies for measuring PD model performances to assess their strengths and weaknesses. We would then like to develop new performance measures for PD models that incorporate several properties:

- This measure should not be skewed by the nature of data populations;
- Evaluation of true model would likely receive close to perfect score;
- This measure should reward models that consistently assign higher PD as a firm moves toward default.

2 Corporate PD Models

Corporate PD models use quantitative inputs such as obligor financial, macroeconomic, and market information to estimate the probability of default over a given time horizon for corporate obligors. Some of the common models include structural, regression, and default intensity models. Ultimately, the performance of these models depends greatly on the quality of the data available for calibration, and in particular the number of default observed within the data sample.
2.1 Structural Models

For firms with publically traded equity, Merton (1974) assumes a simple structure for a credit obligor at time $t$ with a firm franchise asset value $A_t$ (i.e. the total value of the firm), market cap (equity) $E_t$ and total debt value $D_t$. Merton assumed the debt to be a single zero coupon bond which matures at time $T$ with face value $K$. At maturity, the equity holders will decide to payoff the debt if $A_T \geq K$ and default otherwise. So the equity can be viewed as a call option on the asset value with maturity coinciding with debt maturity and strike equal to the face value of the debt. If we assume a geometric brownian motion process for the underlying asset value process

$$\frac{dA_t}{A_t} = \mu dt + \sigma dW_t,$$

the probability of default can be written as

$$P(A_T < K) = N(-DD),$$

with

$$DD = \frac{\ln\frac{A_t}{K} + (\mu + \frac{1}{2}\sigma^2)(T-t)}{\sigma \sqrt{T-t}}.$$

Merton used risk neutral dynamics (i.e. $\mu = r$ the risk free interest rate) to price defaultable debt. This did not work particularly well. However, the statistic $DD$ has proved highly useful as a factor in predicting default. Many others have extended this model to incorporate a richer debt structure and methods for estimating asset value volatility. As the asset value and asset value volatility are not observable quantities, it is necessary to use the call option framework to estimate these quantities. Note also that it is necessary to have a large default data set to estimate actual probabilities of default as the assumption of normality for $DD$ is not accurate due to the complexity of the debt structure for more firms.

2.2 Regression Model

A regression model tries to estimate probability of default via

$$E(Y|X) = f(X, \beta),$$

for a given set of variables. When $Y$ is a dummy variable and takes value 1 if an observation defaults, then

$$P(Y = 1|X) = E(Y|X) = f(X, \beta).$$

In practice, logistic regression is often adopted for default modeling. Again the key factor to success in such models is a good calibration data set on which to estimate the parameters $\beta$ that best identify the defaulting firms. For firms with publicly traded equity, the factor $DD$ is often used as one of the variables in the set $X$.

2.3 Default Intensity Model

Assume random variable $\tau$ is the time when a company defaults. If we further assume that the conditional probability of default between infinitesimal time increment $t$ and $t + \delta t$ is a function of time $t$ and variable $x$ as following

$$\lim_{\delta t \downarrow 0} \frac{1}{\delta t} P(\tau \leq t + \delta t | \tau > t) = \lambda(t|x),$$

then the probability of default over time horizon $T$ can be written as

$$P(\tau < T) = 1 - \exp(-\int_0^T \lambda(t|x) dt).$$

A popular model is the Cox and D.Oakes (1984) proportional hazard rate model. This is similar to a regression model with an addition time component. Other approaches include estimating a stochastic process for $\lambda(t)$. 
3 Performance Meausre

In this section, we introduce some commonly used performance measures for corporate PD models and their challenges.

3.1 Likelihood

For probabilistic models, the likelihood measures the probability of the observed outcomes given the model and the estimated parameters. For a data with \( N_D \) defaulters and \( N_S \) non-defaulters, assume that we have model estimated PD for \( i \)-th obligor \( PD_i \). The log-likelihood can be written as

\[
L = \sum_{j=1}^{N_D} \log(PD_i(j)) + \sum_{j=1}^{N_S} \log(1 - PD_i(j)).
\]

When comparing the fit of two models to the default data, people often uses likelihood ratio test. The test statistics is defined as

\[
D = 2 \left( L_{\text{alternative}} - L_{\text{null}} \right),
\]

where \( D \) is approximately chi-squared distributed with degree of freedom \( df_{\text{alternative}} - df_{\text{null}} \) where \( df_{\text{alternative}} \) and \( df_{\text{null}} \) are the number of free parameters of alternative model and null model. We then calculate the one sided p-value and reject the null hypothesis that alternative model is not better than the null model if p-value is too small.

Stein (2007) showed that the focus on calibration exclusively can lead likelihood methods to incorrectly reject powerful, but imperfectly calibrated models in favor of weaker but better calibrated models. So likelihood based measure should be used in conjunction with methods that focus on the ability of the model to differentiate defaulters from non-defaulters, such as ROC described below, during validation.

3.2 Goodness of Fit

Likelihood measures, especially likelihood ratio tests, are designed for making relative comparison between models, not for evaluating if the model is close to being correctly calibrated or not. It means the best model could still be poorly calibrated to the data. To test how model calibrated to the data overall, goodness of fit is often implemented. There are three types of test as suggested by Hosmer and Lemeshow (2000): Peasron’s chi-squared test, Deviance test, and Hosmer and Lemeshow test.

3.2.1 Pearson’s Chi-Squared Test

Pearson’s chi-squared test uses a measure of goodness of fit which is the sum of differences between observed and expected outcome frequencies (that is, counts of observations), each squared and divided by the expectation:

\[
\chi^2 = \sum_{i=1}^{n} r(y_i, PD_i)^2,
\]

where Pearson residual is defined as

\[
r(y_i, PD_i) = \frac{y_i - PD_i}{\sqrt{PD_i(1 - PD_i)}},
\]

and \( PD_i \) is a estimated probability of default for \( i \)-th sample, \( y_i = 1 \) if \( i \)-th sample defaulted and \( y_i = 0 \) otherwise. \( \chi^2 \) is chi squared distributed with \( n - (p + 1) \) degrees of freedom for a dataset with \( n \) observations and the model has \( p \) free parameters. We calculate the p-value of the statistics and reject the null hypothesis that the model fits the data well if the p-value is too small.
3.2.2 Deviance Test

The Deviance statistic is

\[ D = 2L \]

where \( L \) is the log-likelihood defined above. It has a chi-square distribution. We calculate the p-value of the statistics and reject the null hypothesis that the model fits the data well if the p-value is too small.

3.2.3 Hosmer-Lemeshow Test

For data sets that can be grouped into \( G \) bins by PD, the Hosmer-Lemeshow test statistic is given by

\[ H = \sum_{g=1}^{G} \frac{(O_g - E_g)^2}{N_g \pi_g (1 - \pi_g)}, \]

where \( O_g, E_g, N_g, \) and \( \pi_g \) are the observed number of defaults, expected number of defaults \( = N_g \pi_g \), number of observations, and predicted average PD for the \( g \)-th risk group by PD. We usually choose the number of groups \( G \) to be 10 or 25. The test statistics is chi-squared distributed with \( G - 2 \) degree of freedom. Similarly, we calculate the p-value of the statistics and reject the null hypothesis that the model fits the data well if the p-value is too small.

3.2.4 Challenges

The goodness-of-fit test is able to reject or not reject the model but is less capable of differentiating the models. Also, the test results are heavily data dependent. Another challenge is that statistical tests of likelihood or even the likelihood formulas basically assume independent observations. However credit obligors usually have same drivers of defaults so almost all the data is not independent. This is even more important for practical model users such as portfolio manager who are trying to manage the risk of their credit portfolio.

3.2.5 Some Thoughts

In order to account for the default correlation among the firms in a sample (resulting from shared sentivity to macro-economic and sector factors), we adopt a similar setting as Vasicek (2002) and assume a portfolio with \( M \) credit obligors. Each firm’s asset returns can be represented as

\[ r_i = \sqrt{\rho} F + \sqrt{1 - \rho} \epsilon_i, \]

where \( F \) is a single common factor and \( \epsilon_i \) is a Gaussian distributed idiosyncratic factor for each firm. If we further assume that an obligor defaults when \( r_i < c_i \), then

\[ PD_i = E[P(r_i < c_i|F)] = E[N\left(\frac{c_i - \sqrt{\rho} F}{\sqrt{1 - \rho}}|F\right)] = N(c_i). \]

If \( L \) is the portfolio loss due to defaults, then conditional on the systemict factor \( F \) the

\[ E(L|F) = \frac{1}{M} \sum_{i=1}^{M} N\left(\frac{N^{-1}(PD_i) - \sqrt{\rho} F}{\sqrt{1 - \rho}}\right). \]

If we assume a homogeneous portfolio with \( PD_i = p \) and take the limit as \( M \to \infty \), then the probability of portfolio loss bigger than \( \alpha \) is

\[ P(L > \alpha) = N\left(\frac{N^{-1}(p) - N^{-1}(\alpha) \sqrt{1 - \rho}}{\sqrt{\rho}}\right). \]
For $\alpha = 20\%$, $p = 5\%$ and $\rho = 50\%$, $P(L > \alpha) = 0.0688$. In comparison, for the special case if we ignore the correlation and apply $\rho = 0$, the probability of portfolio loss exceeding $\alpha$ will be 0. We clearly see the impact of correlation on the performance statistics.

**Question 1.** For the aforementioned goodness of fit tests based on likelihood, is there alternative version which takes into account the correlation of observations.

### 3.3 ROC and AR

Another school of thought is to measure how well the model performs in differentiating credits that eventually default from survivors based on rankings by PDs. There are two most commonly used measure: area under the ROC curve and accuracy ratios.

#### 3.3.1 ROC and AUC

The receiver operating characteristic (ROC), is a graphical plot which illustrates the performance of a binary classifier system as its discrimination threshold is varied. It is created by plotting the fraction of true positives out of the total actual positives (TPR = true positive rate) vs. the fraction of false positives out of the total actual negatives (FPR = false positive rate), at various threshold settings. In the PD setting, $TPR(C)$ means the percentage of defaulters in the data set with PDs greater than the cut-off threshold $C$ and $FPR(C)$ means the percentage of non-defaulters in the data set with with PDs greater than $C$. The ROC curve is then a plot of $TPR(C)$ vs $FPR(C)$ as $C$ goes from 1 to 0.

People often use a single summary statistic of the ROC curve, the area under the ROC curve (AUC). It is shown by Fawcett (2006) to be equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one (assuming 'positive' ranks higher than 'negative'). For a score $s$ that indicates higher default risk when the score is low\(^1\), the AUC can be calculated as

$$AUC = \int_{-\infty}^{\infty} TPR(s)FPR'(s) ds.$$ 

The larger the AUC is, the better the model is at ranking observations. As a general rule that is often used by market practice, if $AUC < 0.5$, this indicates no discrimination (i.e. we might as well flip a coin) if $0.7 \leq AUC < 0.8$, this is considered acceptable discrimination, if $0.8 \leq AUC < 0.9$, this is considered excellent discrimination, if $AUC \geq 0.9$, this is considered outstanding discrimination. However, AUC statistics is a statistic that depend on data and should not always be used in absolute value. These ranges for performance are only a rough guide that assumes a specific range of 'true' PDs but in general these numbers aren’t very meaningful and can be gamed or at least be misleading about the performance of the model as we will show later.

However as shown by Bamber (1975), the sample variance of AUC is $\frac{A(1-A)}{D}$ where $A$ is the AUC, $D$ is the number of defaults in the data sample. So the number of defaults instead of the total number of observations in the validation sample is critical for a robust test result. When we are testing the performance on the real data, default count is often limited to about 1000, assuming a model with AUC of 80%, the 95% confidence interval for AUC will be (77.5%, 82.5%) which is a really wide range. If we have a further limited data with only 100 defaults (for example in some industries or regions), the 95% confidence interval for AUC will be (72%, 88%). So caution must be applied when comparing model performance based on AUC for a small sample.

#### 3.3.2 CAP and AR

Similar to ROC curve, we can obtain the cumulative accuracy profile (CAP) by sorting firms/assets/obligors from riskiest to safest as predicted by the credit model (x-axis) and plot against fraction of all defaulted obligors as

\(^1\)If instead a higher score indicates higher default risk, like PD, we need to integrate from $\infty$ to $-\infty$. 

the following figure. A similar statistics is gini coefficient or accuracy ratio which can be calculated as $B/(A + B)$ in the following plot\(^2\).

![Graph showing Area A and Area B](image)

It can be shown that $AR = 2AUC - 1$ if the same weight is attributed to Type I vs. Type II errors so it faces the same challenges as $AUC$. AUC is however a more general measure as different weights may be given to Type I and II errors. For credit risk models, for example, it is far more damaging to mistakenly classify a defaulter as non-defaulters than the other way around.

### 3.3.3 Challenges

AUC and AR are often used interchangeably to measure PD model ranking ability. They both suffer when the number of defaults are limited and the estimates can be of high variance. The statistics depend on the nature of data population, less so on the quality of model as we will show in the next section.

### 3.3.4 Some Thoughts

The previous calculation of AUC and AR is often based on the sample population and sample distribution of PD. If we were to extend it to a general distribution of PD as following, we will be able to have the following interesting result.

Assume $Y$ is a dummy variable that have value 1 when an obligor defaults and 0 otherwise. Assume the distribution of PD is $F(x)$ with density $f(x)$ and the associated conditional PD distribution on defaulters is $P(PD \leq x|Y = 1) = F_D(x)$ with density $f_D(x)$ and conditional PD distribution on non-defaulters is $P(PD \leq x|Y = 0) = F_N(x)$ with density $f_N(x)$, the true positive rate as a function of $x$ can be written as

$$TPR(x) = \int_{s=x}^{1} f_D(s)ds,$$

$$FPR(x) = \int_{s=x}^{1} f_N(s)ds,$$

\(^2\)Here the perfect model means perfect in differentiating defaulters and non-defaulters. However, as we will show later this is not equivalent to a perfect model that produce correct PDs.
and AUC can be written as

\[ AUC = \int_0^1 TPR(x) FPR'(x) dx = \int_0^1 \int_{s=x}^1 f_D(s) ds f_N(x) dx. \]

Alternatively, as shown by Calabrese (2012), the ROC curve as a function of PD \( x \) can be found as

\[ ROC(x) = F_N(F_D^{-1}(x)), \]

which will result in the same AUC formula:

\[ AUC = \int_0^1 ROC(x) dx = \int_0^1 F_N(F_D^{-1}(x)) dx = \int_0^1 F_N(x)d(F_D(x)) \]

\[ = \int_0^1 \int_{s=x}^1 f_N(s) ds f_D(x) dx = \int_0^1 \int_{x=s}^1 f_D(x) dx f_N(s) ds \quad (1) \]

If we further assume the distribution is the true distribution and denote \( P(Y = 1) = \overline{PD} \), by Bayes’ rule, the conditional default probability is

\[ P(Y = 1|PD = s) = \frac{P(Y = 1 \cap PD = s)}{P(Y = 1 \cap PD = s) + P(Y = 0 \cap PD = s)} = \frac{f_D(s)\overline{PD}}{f_D(s)\overline{PD} + f_N(s)(1-\overline{PD})} = s. \]

Notice

\[ f(s) = f_D(s)\overline{PD} + f_N(s)(1-\overline{PD}), \]

we have \( f_D(s) \propto s f(s) \). Similarly, \( f_N(s) \propto (1-s)f(s) \). Furthermore,

\[ \overline{PD} = \int_0^1 P(Y = 1|s) f(s) ds = \int_0^1 s f(s) ds, \]

and

\[ f_D(s) = \frac{s f(s)}{\overline{PD}} = \frac{s f(s)}{\int_0^1 s f(s) ds}, \]

\[ F_D(x) = \int_0^x \frac{s f(s)}{\overline{PD}} ds. \]

So the AUC can be found as

\[ AUC = \int_0^1 \int_{s=x}^1 f_D(s) ds f_N(x) dx \]

\[ = \frac{1}{\overline{PD}} \int_0^1 \int_{s=x}^1 f(s) ds f(x) (1-x) dx \]

\[ = \frac{1}{2\overline{PD}(1-\overline{PD})} \left( 1 - \int_0^1 F^2(x) dx - \overline{PD}^2 \right). \]

We notice the fact that \( f(s) \) is the true density of PD. So the AUC for a true model would be less than 100%. For example, if the density \( f(x) = 1 \) is uniform, then

\[ AUC = 2 \int_0^1 (1-x^2)(1-x) dx = \frac{5}{6}. \]
This result shows that the AUC depends only on the distribution of the PDs for the sample when defaults occur with true probability PD. In general, choosing a sample with highly risky firms mixed with very safe firms will lead to a higher AUC. Thus AUC does not necessarily reflect the quality of the PD model.

We can statistically test if the model PD follows the true PD distribution using the relationship between the true PD distribution and conditional distribution. If the model PD follows the true PD distribution, we know the theoretical distribution from (2) and (3). We can adopt a two-sample Kolmogorov-Smirnov test of sample distribution of PD conditional on default against the theoretical conditional distribution (3). We reject that the model PD follows true PD distribution if KS statistics is greater than a threshold.

Similarly, we can test if the model PD follows the true PD distribution by comparing the sample ROC curve against the theoretical ROC curve (1).

**Question 2.** What ROC level will imply a better PD model? Does a model ROC close to true PD distribution ROC imply that the model is good?

**Question 3.** Suppose we have two model PD distributions $F_{iD}^1$, $F_{iD}^2$, and $F_{iN}^j$ for two distribution $i = 1, 2$ and true PD distribution $F$. If for any $x$

$$|ROC_{F1}(x) - ROC_F(x)| < |ROC_{F2}(x) - ROC_F(x)|,$$

will this imply

$$\|F_{D}^1 - F_{D}\|_{L^1} < \|F_{D}^2 - F_{D}\|_{L^1}, \|F_{N}^1 - F_{N}\|_{L^1} < \|F_{N}^2 - F_{N}\|_{L^1}?$$

**Question 4.** Explore the properties and relationships of true PD distribution and conditional PD distribution. Design other statistical test of model PDs based on the PD distribution.

4 Research Questions

A number of interesting questions around credit risk models and how to measure their performance remain to be addressed. For the 2014 Mathematical Problems in Industry Workshop we pose the following open problems.

- Review existing methodologies for measuring PD model performance and assess strengths and weaknesses.

- Develop a new performance measure for PD models that incorporates the following properties:

  1. Measure should focus on the correctness of the PD estimate relative to true PD, and not be skewed by the nature of the sample population.
  2. The evaluation of a true model (firms default with exactly modeled PD frequency) would receive close to perfect score.
  3. Worst score would require significant mislabeling of (almost) guaranteed defaulters and survivors.
  4. The measure should reward PD models that consistently show an increase in PD as a firm moves to default.
  5. Ideally, correlation of defaulters in sample would be taken into account.
  6. Effects of finite sample size would be incorporated into measure.

- Consider how to incorporate differing misspecification costs into the performance measure to account for the fact that investing in a defaulter is generally more costly than not investing in a firm that doesn’t default.

- Consider how to differentiate two PD models on the same sample in order to identify the more powerful model.
References

D Bamber. The area above the ordinal dominance graph and the area below the receiver operating graph. 12, 387-415, Journal of Mathematical Psychology, 1975.


