Improving long run model performance using Deviance statistics

Matt Goward

August 2011

Objective of Presentation

Why model stability is important

Financial institutions are interested in long run model stability in terms of population and model performance.

Model performance volatility can:

- Impact operational performance, e.g. collections models.
- Impact consistency of decisions through time, e.g. application and behavioural models.
- Require a greater level of conservatism in implementation, e.g. Basel.

For the business re-building models can be:

- Time consuming
- Costly in terms of resource, retros, implementation, etc.
- Costly in terms of regulatory compliance

Objective of Presentation

Characteristic selection can improve model stability

Lots of things can affect the stability and long term performance of the model:

- Stability of scorecard characteristics over time
- Stability of the underlying portfolio in terms of marketing, growth and delinquency
- The consistency and quality of the data, etc.

Characteristic selection can also affect stability and long term model performance as:

- Too many model characteristics generates over-fitting, excess model variance and short lived models.
- Spurious characteristics can also introduce model bias and produce short lived models.
- Too few characteristics and the model is poorly specified and potentially biased.

Therefore, get characteristic selection right and your model lasts longer. Choose the maximum number of variables your dataset whilst limiting unnecessary model variance.

Objective of Presentation

MIV and significance testing working together

A number of methods are used across the industry to select characteristics in a binary target scorecard build. Use of Information Values and Marginal Information Values (MIVs) in conjunction with expert judgement, selecting one characteristic at a time based on the highest MIV, is one common method used to select characteristics.

With this in mind, for models of a binary target variable (using Logistic Regression):

- Does reliance on Marginal Information Values (MIVs) lead to over fitting models/ reduced long term model performance?
  It can do

- Can using a statistical test of significance in conjunction with MIVs improve long term model performance?
  Yes

- Which statistical test of significance is best?
  Deviance statistic

- Does expert judgement already ensure long term stability?
  Depends on the analyst but in my case . . . no

Statistics for characteristic selection

Marginal Information Values

LLOYDS BANKING GROUP

- Marginal information values give an index measure of how much information a characteristic would add to a model given all the other characteristics already within it. For example, if a model has 4 characteristics and you want to find out how much additional information 'Assets as a % of Liabilities' would add, its MIV can be calculated.

	Observed			Expected						
Assets as a % of Liabilities	Goods	Bads	WoE	Goods	Bads	WoE	Delta Score	Delta * G%	Delta * B%	Attr MIV
Default	536	174	-0.98	577.7	112.3	-0.47	-0.51	-0.037038386	-0.098812342	0.061774
<=25%	278	48	-0.35	194.6	31.4	-0.28	-0.07	-0.002536808	-0.003599637	0.001063
26 to 50%	94	17	-0.40	196.2	34.8	-0.38	-0.02	-0.000246188	-0.0003659	0.00012
51 to 75%	273	46	-0.33	273.1	45.9	-0.32	0.00	-9.35327E-05	-0.000129519	3.6E-05
76% +	6240	618	0.21	6179.4	678.6	0.10	0.10	0.086862547	0.070698651	0.016164
TOTAL	7421	903		7421	903					
									MIV	0.079156

- The expected figures in the example are derived from the average score (based on the existing 4 characteristic model) for each attribute (which has an associated Ln odds).

- The Weights of Evidence for each attribute can be calculated for the observed and actual figures (WoE = Ln(%Goods / %Bads)). The Delta score is then calculated as the observed WoE minus the expected WoE.

Statistics for characteristic selection

Marginal Information Values

	Observed			Expected						
Assets as a % of Liabilities	Goods	Bads	WoE	Goods	Bads	WoE	Delta Score	Delta * G%	Delta * B%	Attr MIV
Default	536	174	-0.981268	577.7	112.3	-0.468466	-0.51	-0.037038386	-0.098812342	0.061774
<=25%	278	48	-0.349926	194.6	31.4	-0.282208	-0.07	-0.002536808	-0.003599637	0.001063
26 to 50%	94	17	-0.396265	196.2	34.8	-0.376829	-0.02	-0.000246188	-0.0003659	0.00012
51 to 75%	273	46	-0.325516	273.1	45.9	-0.322974	0.00	-9.35327E-05	-0.000129519	3.6E-05
76% +	6240	618	0.2059	6179.4	678.6	0.102598	0.10	0.086862547	0.070698651	0.016164
TOTAL	7421	903		7421	903					
									MIV	0.079156

- Then for each attribute:

  Delta Score * (%Observed Goods - %Observed Bads) = MIV per attribute.

Statistics for characteristic selection

Marginal Information Values

	Observed			Expected						
Assets as a % of Liabilities	Goods	Bads	WoE	Goods	Bads	WoE	Delta Score	Delta * G%	Delta * B%	Attr MIV
Default	536	174	-0.981268	577.7	112.3	-0.468466	-0.51	-0.037038386	-0.098812342	0.061774
<=25%	278	48	-0.349926	194.6	31.4	-0.282208	-0.07	-0.002536808	-0.003599637	0.001063
26 to 50%	94	17	-0.396265	196.2	34.8	-0.376829	-0.02	-0.000246188	-0.0003659	0.00012
51 to 75%	273	46	-0.325516	273.1	45.9	-0.322974	0.00	-9.35327E-05	-0.000129519	3.6E-05
76% +	6240	618	0.2059	6179.4	678.6	0.102598	0.10	0.086862547	0.070698651	0.016164
TOTAL	7421	903		7421	903					
									MIV	0.079156

- Then for each attribute:

  Delta Score * (%Observed Goods - %Observed Bads) = MIV per attribute.

- The total Marginal Information Value for the characteristic is then the sum of these attribute level MIVs. Industry standards vary in determining a 'significant' MIV. MIV's greater than or equal to 0.03 are usually analysed for additional scorecard performance impact.

Statistics for characteristic selection

Marginal Information Values

- **Pros**:
  - Can compare the predictiveness of characteristics even across datasets in one common measure.
  - Computationally inexpensive.
  - Common across the industry.
  - The ‘amount of information’ in a characteristic is a measure of magnitude not significance.

- **Cons**:
  - Can be difficult to interpret as there are no associated statistical tests.
  - No trade off between number of bins for a characteristic and its difficulty in entering the model.
  - Cut-off for model entry is arbitrary and doesn’t account for sample size.

- Pros are very attractive but cons can lead to over fitting of models.

Statistics for characteristic selection

Tests for statistical significance

LLOYDS BANKING GROUP

- Using a measure of 'magnitude' for characteristic selection – i.e. how much additional information a characteristic holds, is very useful.
- To guard against over-fitting an analyst could use a statistical significance test to validate a characteristic selected by MIV.
- Using significance testing has its drawbacks but may improve an iterative MIV characteristic selection.
- Therefore, the statistic should be as powerful as possible, i.e. – maximise the probability of correctly identifying a significant characteristic.
- Here we discuss two choices in the context of an additional characteristic (e.g. Assets as a % of Liabilities) added to a model (e.g. with 4 existing characteristics):
  - Wald statistics – automatically outputted by PROC LOGISTIC
  - Likelihood Ratio (Deviance) statistics – produced in a PROC LOGISTIC macro loop or PROC GENMOD
- NB. Important to use unweighted counts for significance testing.

Statistics for characteristic selection

Wald Statistic

Wald Statistic = $(\hat{\beta} - \beta_0) / s.e$

- This statistic measures whether the parameter estimate for a given characteristic is significantly different from zero.

- The statistic relies upon an accurate estimate of the characteristic standard error. Standard error estimates are only reliable as the vector of parameter estimates (beta hat) gets very close to its population value (see Appendix B).
  - Wald is only 100% reliable when the parameter estimate is accurate. If other scorecard characteristics have been poorly selected then the Wald test for whether the next characteristic is significant won't be accurate.
  - Need to employ the standard error estimate from the model. This may not be accurate for very significant characteristics (helpful!!!).

- In addition, due to various undesirable properties the Wald statistic can exhibit aberrant behaviour (see Hauck and Donner 1977). In extreme cases an insignificant result may mean that the parameter estimate is relatively far or quite near to zero.

Statistics for characteristic selection

Wald Statistic

- **Pros**:
  - Automatically outputted by SAS in PROC LOGISTIC.
  - Computationally inexpensive as multiple models do not need to be estimated.
  - Automated model selection functionality in SAS.

- **Cons**:
  - The statistic relies upon correct estimation of variance which is used for the null hypotheses.
  - The statistic can exhibit aberrant behaviour (see Appendix C).

Statistics for characteristic selection

Deviance Statistic – for the model

Log Likelihood Function	→	log probability density function of the observed data as a function of unknown parameters.
Maximum Likelihood Estimation	→	set of parameter estimates where the observed data has the maximum probability of occurrence.

Model Deviance Statistic = -2*(Maximised Log Like (No variables) – Maximised Log Like (Model))

- The more of the Log Likelihood Function explained by the model the greater the Deviance.
- This statistic is consistent but not particularly powerful.
- However . . . . .

Statistics for characteristic selection

Deviance Statistic – for the characteristic

Characteristic Deviance Statistic = -2*(Maximised Log Like (Model without Char) – Maximised Log Like (Model with Char))

Characteristic Deviance Statistic isolates the additional explanatory power of the characteristic in question and is powerful.

- Unlike the Wald statistic any previous model mis-specification, noise or bias cancels out. Therefore just the significance of the additional explanatory power of the characteristic (e.g. Assets as a % of Liabilities) is analysed.
  - No dependence on estimating the variance.
  - No dependence on the base model being correct.
  - The statistic is consistent and powerful.

Statistics for characteristic selection

Deviance Statistic

- **Pros**:
  - Not reliant on a consistent estimate of variance for the null and alternative hypothesis.
  - Does not exhibit aberrant behaviour.
  - Uses the maximum available information with respect to the relationship between the log likelihood function and the betas (see Appendix C). As a result this statistic is more flexible and has proven more consistent in empirical research.

- **Cons**:
  - Can be computationally expensive as one has to run multiple models to evaluate each model characteristic. However, given much greater modern day processing power this is not such a drawback.

Comparing techniques in open competition

- 6 months of PCA collections data was extracted from several years ago along with customer level characteristics.
- From this dataset 1,500 Goods and 1,500 Bads were randomly sampled based on a 90 day 'return to order' definition.
- A sample of 1,500 Goods and 1,500 Bads was also taken from one month for Out of Time validation.
- Manual classing of the development sample based on WoE patterns was undertaken. On this sample 4 models were built:
  1. An 'MIV' model built using only MIV statistics to select characteristics.
  2. A 'Deviance' model built using Deviance Comparison statistics to validate characteristics identified using MIV.
  3. A 'Wald' model built using Wald statistics to validate characteristics identified using MIV.
  4. A 'Judgement' model taking account of full expert judgement akin to a formal collections scorecard build.
- All four models were analysed in full quarterly monitoring packs spanning 11 quarters.

MIV Model Results

- The MIV model was built iteratively one characteristic at a time based on the highest MIV in the dataset. Spurious characteristics and characteristics with illogical Weight of Evidence patterns were omitted from the modelling process. Each characteristic added had to produce an uplift in KS, Divergence and Gini for both the development and hold out validation samples.

- **The final model contained 10 characteristics and achieved a development Gini of 40.8%.**

- Over the 11 quarters of monitoring all stability and alignment measures were within tolerance. However there was a deterioration in model performance from Q8. As a rough industry standard the Gini tolerances applied for this project were Gini <95% Development = AMBER, Gini <85% Development = RED.

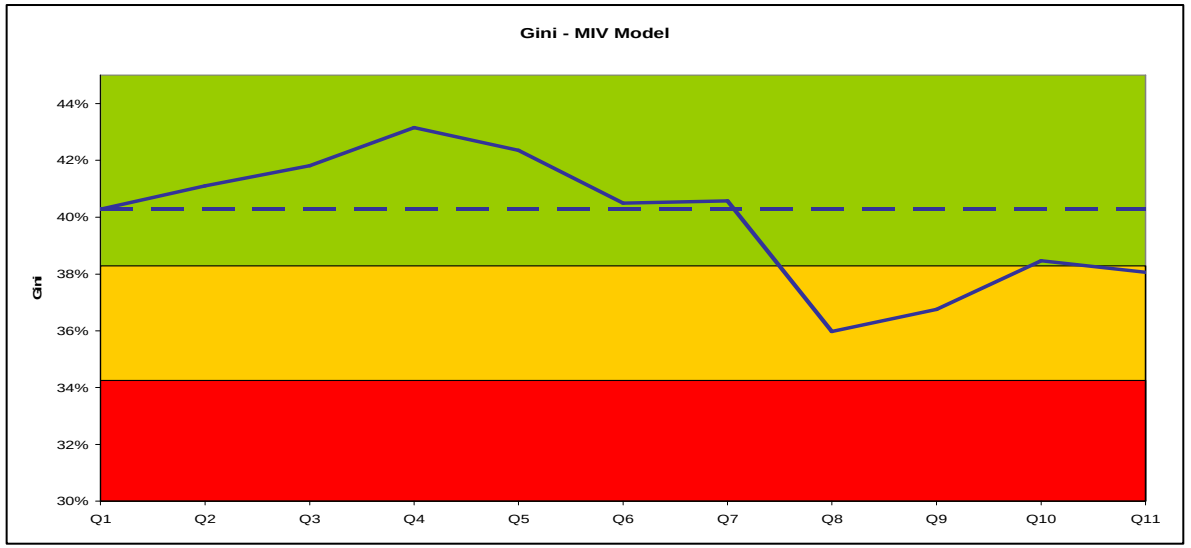


Figure 2 – Gini – MIV Model

Deviance Model Results

- The Deviance model was built in the same way as the MIV model but with all model characteristics validated by the Deviance statistics at each modelling stage.
- **The final model contained 6 characteristics and achieved a development Gini of 37.1%.**
- Over the 11 quarters of monitoring all stability and alignment measures were within tolerance. There was no deterioration in model performance relative to development.

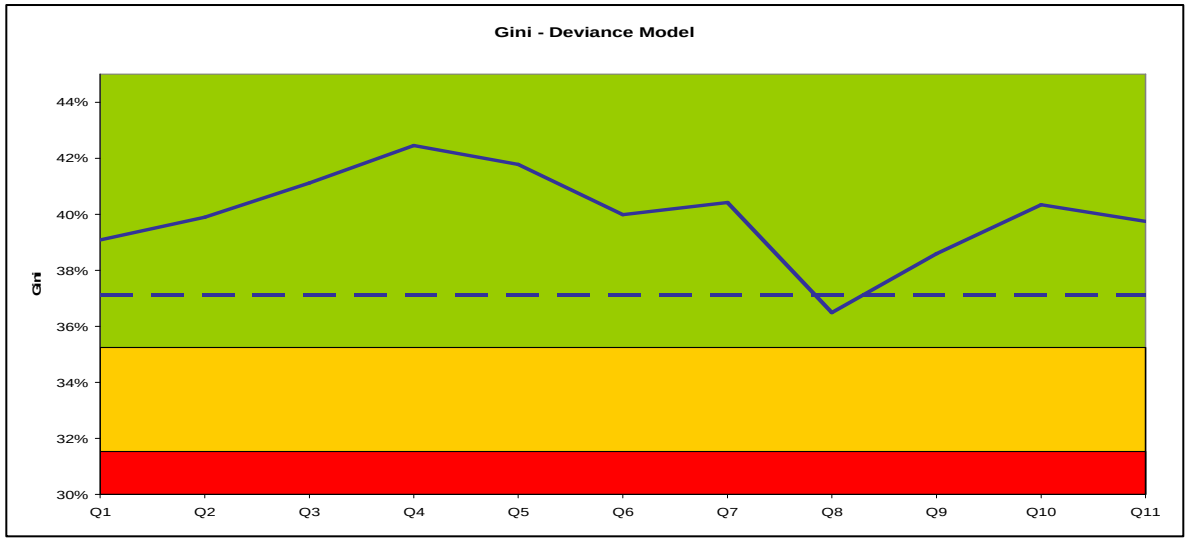


Figure 3 – Gini – Deviance Model

Deviance – MIV model comparison

- Figure 5 shows that the 3.7 Gini point difference between the models in development does not translate to live performance in the months shortly after development. Over time the Deviance model outperforms the MIV model with far fewer characteristics. This is also demonstrated by the monthly Gini differences in Figure 4.

- **Therefore, the Deviance model provides a more stable and long term predictive solution compared to the MIV model.**

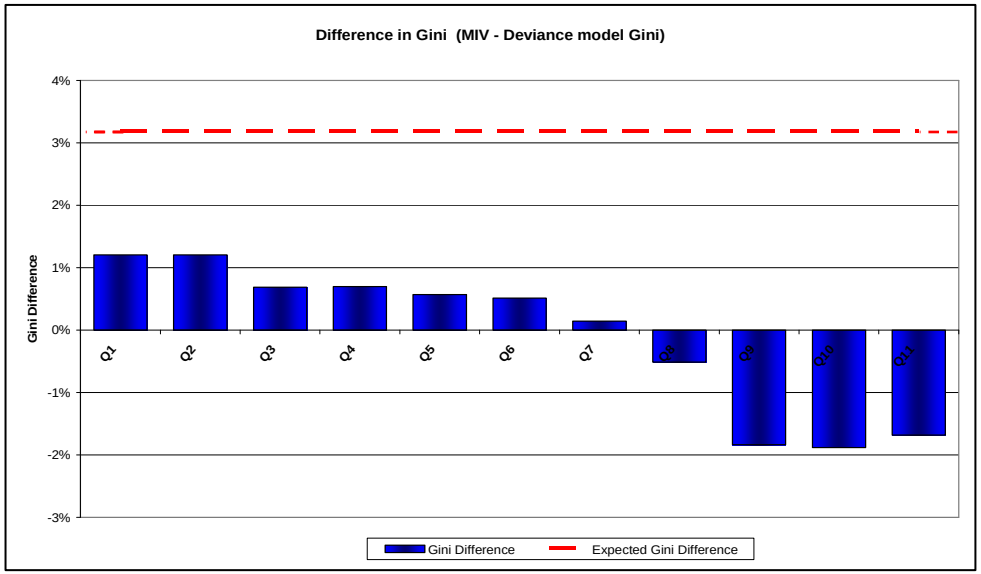


Figure 4 – Difference in Gini (MIV – Deviance model Gini)

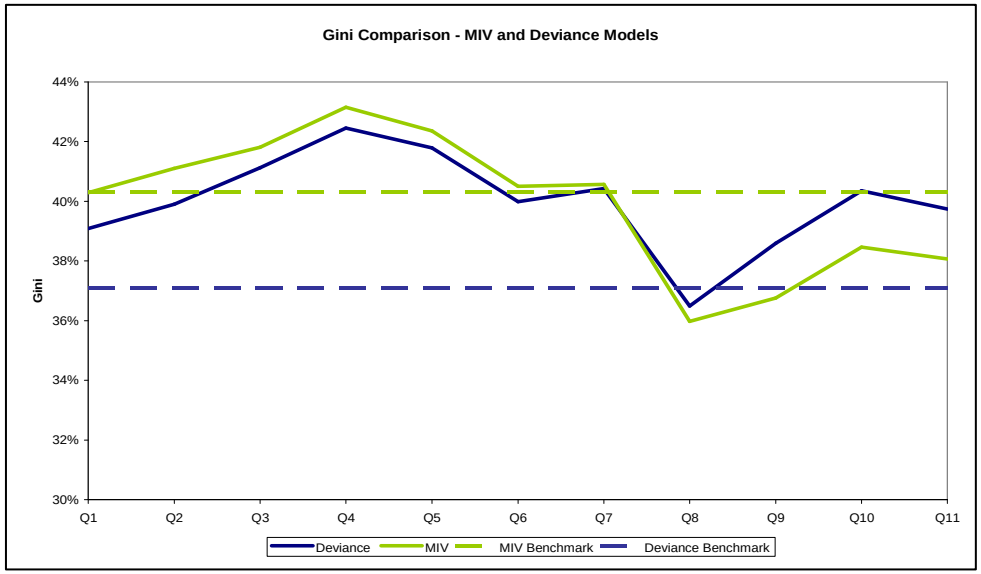


Figure 5 – Gini Comparison – MIV and Deviance Models

Wald Model Results

- The Wald model was built in the same way as the MIV model but with all model characteristics validated by the Wald statistics at each modelling stage.
- **The final model contained 4 characteristics and achieved a development Gini of 36.0%.**
- Over the 11 quarters of monitoring all stability and alignment measures were within tolerance. There was no deterioration in model performance relative to development.

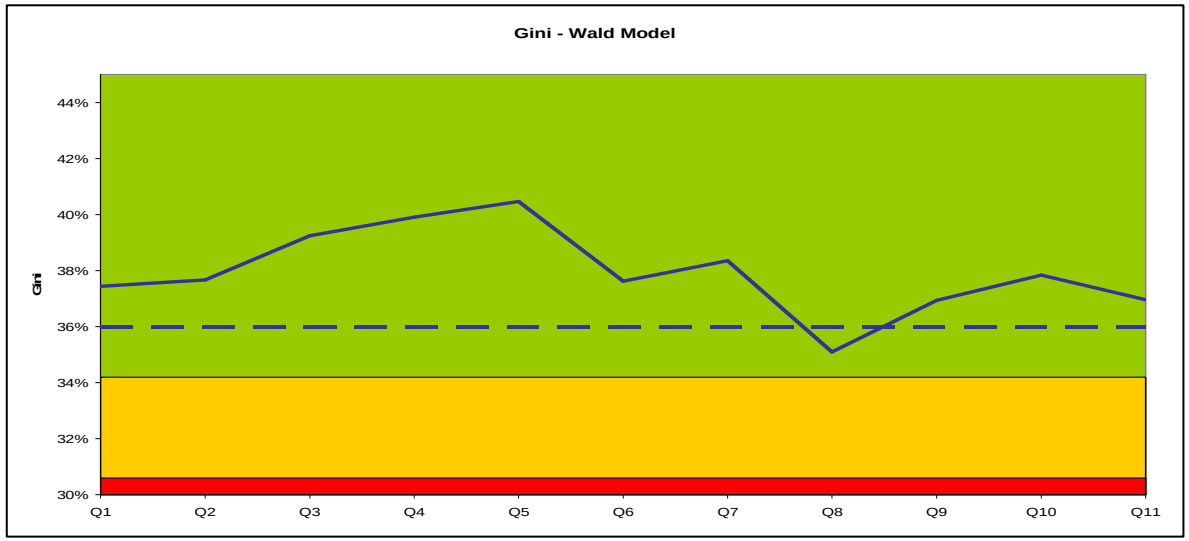


Figure 6 – Gini – Wald Model

Wald v Deviance model comparison

- Figure 7 shows that the 1.1 Gini point difference between the models in development is always exceeded in the monitoring period. Over time the performance uplift of the Deviance model over the Wald model grows. This is supported in Figure 8.

- **Therefore, the Deviance model provides a efficient long term predictive solution compared to the Wald model.**

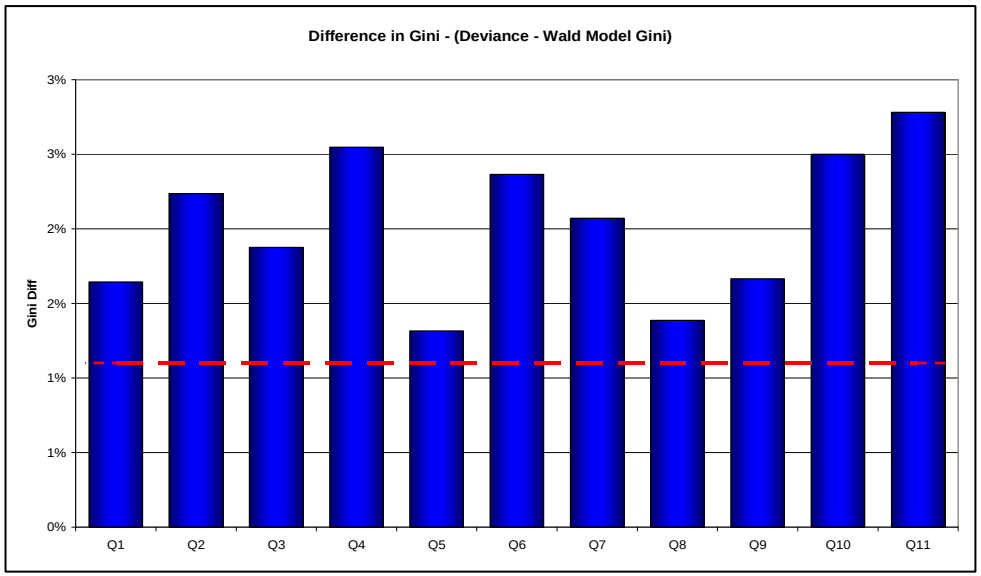


Figure 7 – Difference in Gini (MIV – Deviance model Gini)

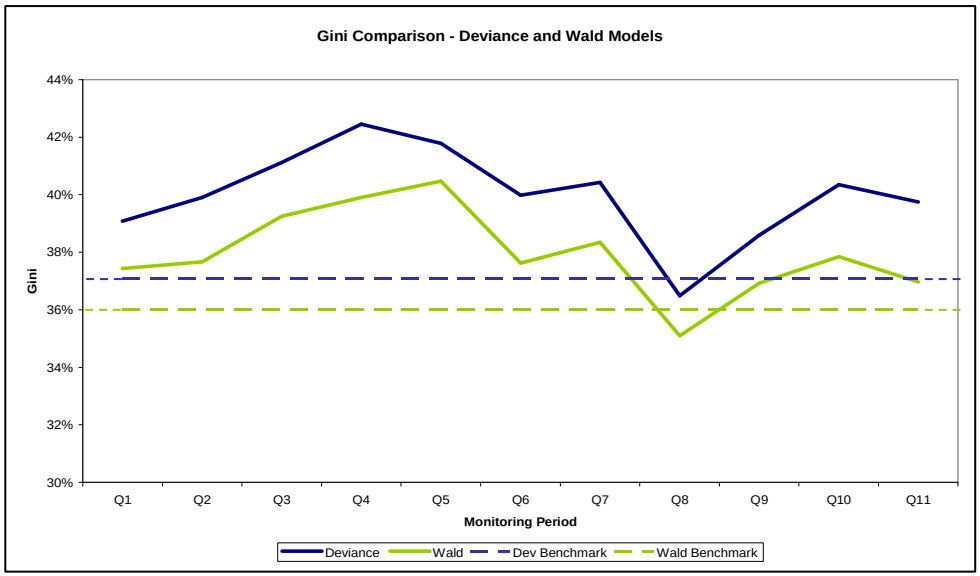


Figure 8 – Gini Comparison – MIV and Deviance Models

Judgement Model results

- For the Judgement model MIVs were used to indicate predictive characteristics. However, characteristics were also chosen to generate a good variable mix in terms of product type (PCA, loans, etc.), degree of historical data used in characteristic calculation (3m, 6m and 12m), and characteristic type (credits, debits, instalments, etc.). Many iterations were run to arrive at the final model.

- **The final model contained 7 characteristics and achieved a development Gini of 34.2%.**

- Over the 11 quarters of monitoring all stability and alignment measures were within tolerance. There was only one quarter where the performance of the model breached the AMBER threshold.

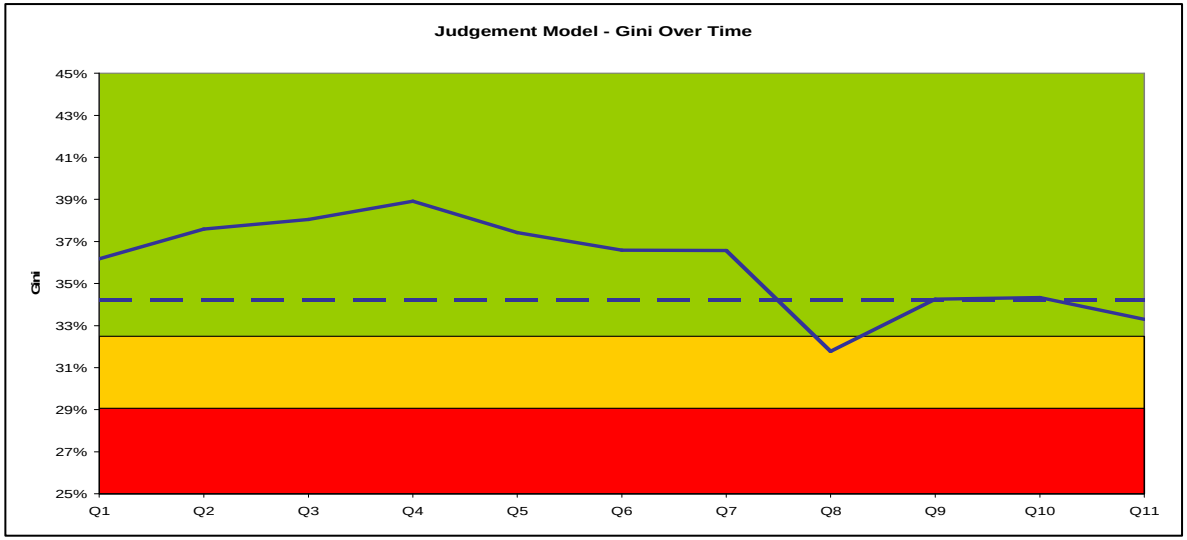


Figure 9 – Judgement Model Gini over time

Judgement v Deviance Model Comparison

- The Judgement model has a lower development Gini than the Deviance model with no benefit in terms of stability over the monitoring period (Figure 11). Indeed, model performance is slightly more volatile for the Judgement model. This is supported by the difference in Gini between these two models increasing over time (Figure 10 – the red line is the development difference in Gini).

- **Therefore, the Deviance model provides a more stable and long term predictive solution compared to the Judgement model.**

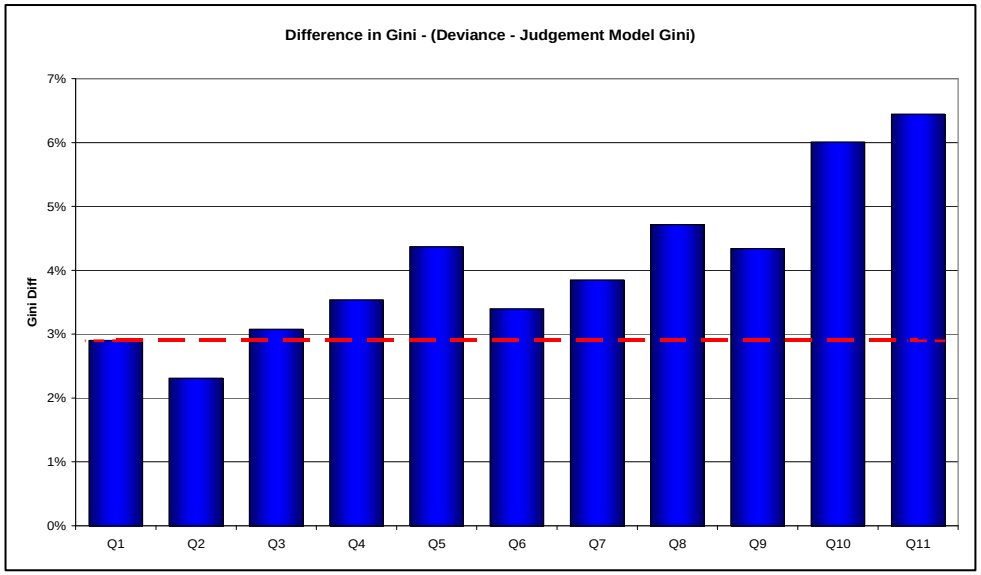


Figure 10 –Deviance Gini – (Deviance – Judgement Model Gini)

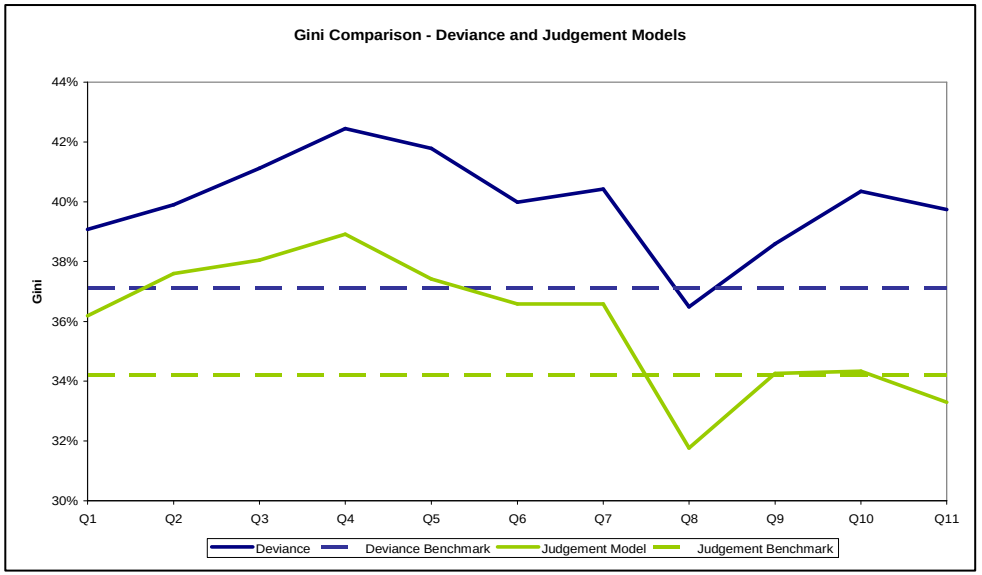


Figure 11 – Gini Comparison –Deviance and Judgement Models

Critique

LLOYDS BANKING GROUP

- This project was based on a relatively small sample by modern standards. The performance of all the statistics will converge with higher volumes.
- A longer monitoring period would have been preferable as all the models could have been tracked until performance was significantly degraded. This would have achieved an appreciation of model lifespan.
- In terms of the statistics themselves, they are used in a stepwise procedure. This has a number of drawbacks particularly around conducting multiple significance tests on the same dataset. It should always be borne in mind that p-values are conditional on other characteristics being included in the model and that multiple testing error has not been controlled for.
- Deviance statistics can only be used to compare 'nested models', i.e. model 2 is the same as model 1 plus one additional characteristic. When several models have been developed they can be compared using Akaike's Information Criterion (AIC) or Bayes' Information Criterion (BIC). Analysis of these statistics was outside the scope of this project.

Conclusions

- The evidence of this project demonstrates that using a significance test is beneficial to long term model stability. Also, the results indicate that the Deviance statistic is preferable to the Wald statistic for this purpose.

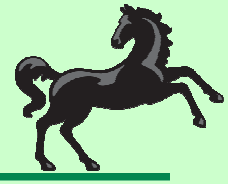
- Also in this project, using a statistics based modelling method produced superior outcomes compared to using expert judgement.

- An important caveat to this is that spurious or illogical characteristics were excluded from development – in itself a form of expert judgement. In addition, expert judgement will always be required to feed in knowledge of the data into the final model, gain stakeholder 'buy-in' for the new model , and ensure regulatory compliance in terms of 'treating customers fairly'.

- MIVs are needed as they give a measure of magnitude in terms of the relationship between the characteristic and the dependant variable. It is also sensible to not base characteristic selection purely on significance testing.

- Nevertheless, it appears that combining a significance measure (Deviance) with a magnitude measure (MIV) can work as a useful tool to modelling analysts – feeding benefits back to the business in the form of longer lasting and more stable models.

LLOYDS BANKING GROUP



Appendices

Appendix A - Deviance explained

- A Deviance statistic is the difference in log likelihood between a 'perfect' model where each observation is a characteristic, and the actual model. To get the Deviance of a characteristic you subtract the Deviance of the model with the characteristic omitted from the Deviance of the model with the characteristic included.

$G(I)^2 =$ Deviance for the independent model (no variables modelled)

$G(L)^2 =$ Deviance for the saturated model (each observation has its own coefficient = perfect fit)

$G(M)^2 =$ Deviance for the built model

$G(M_2 \mid M_1)^2 =$ Deviance of model 2 given the deviance of model 1. In this context model 2 has one additional variable over model 1.

$$G(M_2 \mid M_1)^2 = -2(L_2 - L_S) - [-2(L_1 - L_2)]$$

Where L is the log likelihood function and S is the saturated model

$$G(M_2 \mid M_1)^2 = G(M_2)^2 - G(M_1)^2$$

Appendix A - Deviance explained

- Below is a worked example. A model with one variable has been built. The figures below demonstrate the investigation as to whether an additional characteristic (char 2) significantly adds to the model.

Model Fit Statistics **Model1**

Criterion	Intercept Only	Intercept and Covariates
AIC	7992.888	7360.196
SC	7999.875	7388.145
-2 Log L	7990.888	7469.890

Model Fit Statistics **Model2 (Char 2 added)**

Criterion	Intercept Only	Intercept and Covariates
AIC	7992.888	7360.196
SC	7999.875	7388.145
-2 Log L	7990.888	7352.672

$G(I)^2 =$ 7990.888 with N-1 degrees of freedom [here 7999]

$G(M_1)^2 =$ 7468.890 with N-K1-1 degrees for freedom (K1= number of factors of varibales in model 1) [here 7994]

$G(M_2)^2 =$ 7352.672 with N-K2-1 degrees for freedom (K2= number of factors variables in model 2) [here 7993]

$G(M_2 \mid M_1)^2 =$ 7468.890 – 7352.672 = **116.28** [(N-K1-1)-(N-K2-1)] = **4 degrees of freedom**

Associated p-value <0.001 – Char 2 significantly adds to the model

Appendix B - Problems with Wald explained

Bad standard error estimates

- The Wald statistic is calculated to produce a distribution of the test statistic assuming that the parameter of interest in the 'beta-hat' vector is equal to zero. From this distribution the associated p-value is derived. Therefore, you have to employ the variance estimate from the maximised log likelihood function used to derive 'beta-hat'. This variance estimate is not necessarily the same for if the parameter of interest is hypothesised to be 0 (this is applicable to maximum likelihood estimation but not ordinary least squares in linear regression).

- Proof:

$$g(b;x) = \frac{\partial L(b;x)}{\partial b} \qquad \text{var}(g) = -E(H)$$

H is the Hessian matrix of second order partial derivatives , L is the log likelihood function, and b is the sample vector of parameters. g is a function of b (the estimated parameter vector) therefore:

$$\text{var}(g) = H\ \text{var}(b) H$$

$$\text{var}(b) = -H^{-1}$$

The variance estimate $\text{var}(b)$ is outputted by the model is for the estimated value $\hat{\beta}$. This is not the same as for the hypothesised value of b which is 0. The variance estimate will be sufficient for the Wald statistic if the log likelihood function for the hypothesised value (0) is of the form below and b and 'beta-hat' are relatively close:

$$L(b;x) \approx L(\hat{\beta};x) + g(\hat{\beta};x)(b-\hat{\beta}) + (b-\hat{\beta})' H (b-\hat{\beta})$$

Appendix B - Problems with Wald explained

Bad standard error estimates

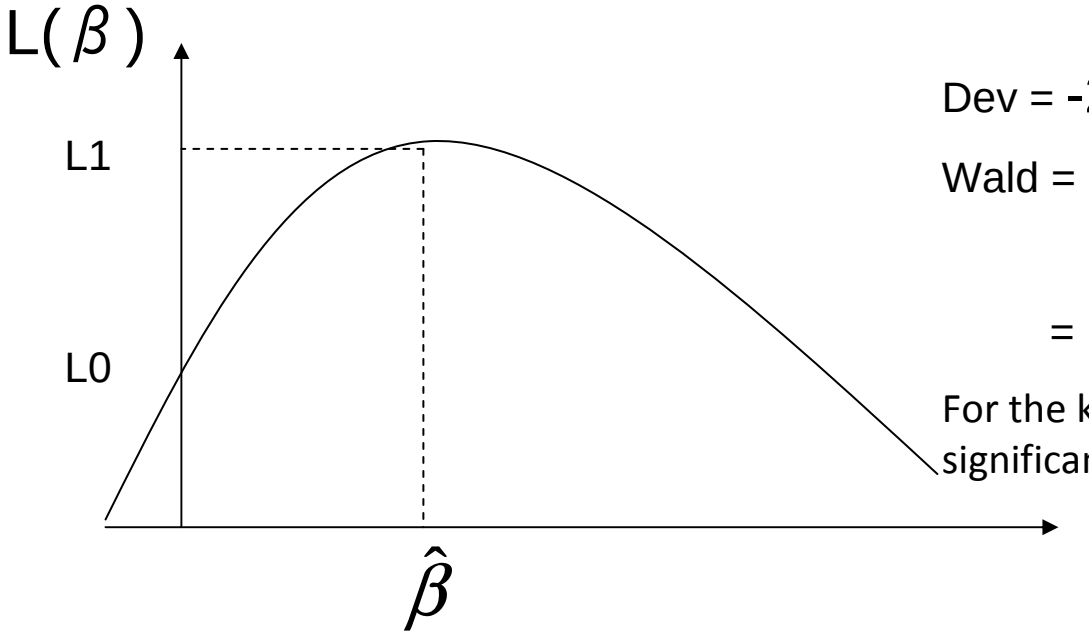
$$L(b;x) \approx L(\hat{\beta};x) + g(\hat{\beta};x)(b-\hat{\beta}) + (b-\hat{\beta})' H (b-\hat{\beta})$$

This is a second order Taylor expansion (g being FOC, H being SOC). If the log likelihood function deviates from this and the b vector becomes far away from the 'beta-hat' vector then the approximation will be poor. This issue can be seen in two contexts:

1. H is only accurate in the approximation above if b and 'beta-hat' are close together. Therefore if the existing model (of four variables in the example given throughout) is poorly specified b and 'beta-hat' may be far apart.
2. In hypothesis testing if you assume that 'beta-hat' is close to the true value 'beta' (which is itself non-zero), then setting b = 0 also leads to the above approximation being less likely to hold. This means that if a characteristic is very significant the associated p-value will not be consistent with this as the test statistic is based on an inaccurate variance estimate.

For reasons 1 and 2 the Wald statistic may be inaccurate.

Appendix C - Wald and Deviance compared



Dev = -2(L1 – L0)

Wald = $\hat{\beta} / (d^2 L(\hat{\beta})) d\hat{\beta}^2$

= $\hat{\beta} / H_{kk}$

For the k'th parameter in the model (the significance of which we are testing).

- As is now evident, the Wald statistic relies on correctly estimating the variance of 'beta-hat' with the assumption that this is sufficient for the null hypothesis. This is the same as estimating the curvature of the log likelihood function in the graph above. As has been stated, the variance estimate for 'beta-hat' will only be accurate as 'beta hat' converges on its (true) population value. Also, as stated in Appendix B, this variance estimate may not be correct for the hypothesised value of 'beta', i.e. 0.

- In contrast, the Deviance statistic does not require a variance estimate. It is simply calculated as the twice the vertical distance between L0 and L1 – thereby using all the available information to provide a significance estimate.

- It is for these reasons that the p-value for the Deviance statistic is usually less than the Wald equivalent – meaning the Deviance statistic has the best chance of correctly identifying a significant characteristic.