Cigniti

Credit Scoring: Model Validation and Best Practices

A Tech Brief

World's Leading AI & IP-led Digital Assurance & Digital Engineering Services Company

Abstract

If you are a Credit Analyst, Risk Manager, Chief Financial Officer (CFO), or Chief Executive Officer (CEO) of a bank/financial institution, and do not answer these questions, it's time to reinvent your strategy.

Which prospects are on our best offer radar? How likely it is that a particular set of clients will default? What credit line and interest rate are suitable for a specific buyer or purchase? Which consumers are most likely to respond to collection efforts for past-due accounts? Does the overall level of risk in our lending portfolio fall within the bank/financial institution permitted range?

Customers now anticipate financial firms to make credit decisions quickly. The enterprise that takes the longest to respond will lose out to more nimble rivals.

This tech brief explores effective and efficient statistical credit scoring model validations and best practices that help CFOs & CEOs in devising the best credit services for their customers.

Why does Credit Scoring Model Validation Matter?

Innovations in Big Data, Digital, and Analytics brought a paradigm shift in banks' credit-decisioning models that underscore their lending processes. In fact, the banks (and fintech companies) that have leveraged high-performance models have already increased revenue, reduced credit-loss rates, and made significant efficiency gains thanks to more precise and automated decisions.

> **"Banks need to implement more automated credit-decisioning models that can tap new data sources, understand customer behaviors more precisely, open up new segments, and react faster to changes in the business environment."**
>
> -Mckinsey

But these automated credit risk scoring models have their fair share of risks. Model validation is a critical activity to verify that credit risk scorecards are working as intended and that model usage is in line with business objectives and expectations. It can also serve as an early warning system for identifying when a change may be necessary, whether it be an adjustment to a score cut-off strategy or a full model redevelopment.

A Roadmap to Model Building

Our objective is to build a classification model that will give some probabilistic output based on some input variables. We can convert this probabilistic output to a binary output based on some threshold value. Here, the input variables are independent variables and the output is a dependent variable. One of the most important steps in predictive model building is to establish the correct definition of the dependent variable. In the case of the credit decision model, the clients i.e. the dependent variable are mainly classified into two types-good or bad. There are a few more types of clients as well, but they are not taken into account for model building. However, we will define all of them in brief.

First, the two most important categories of clients are defined –**good** and **bad**. This definition is based on two important parameters:

[DPD – Days Past Due is the number of days after the due date.**]**

- **The client's number of days after the due date (days past due or DPD):** If we consider only DPD, the clients who are not delinquent are identified as good and clients with high DPD are identified as bad. We need to set a threshold for this DPD value. Sometimes the client may delay payment innocently because of some technical glitches, or he/she may have forgotten. In our case study, we have considered clients with DPDs less than 30 or equal to zero as good. On the other hand, clients with more than 90 days of DPD are considered as bad.

[APD– Amount Past Due is the due amount after the due date.**]**

- **The Amount Past Due:** In this case, also, we need to set some tolerance levels. This means we need to define what is considered debt and what is not. It makes little sense to regard a small amount as past due. In our case study, $5 is the threshold, more than that amount will be considered debt. This may vary depending on the product.

With the definition of good and bad clients, here comes another type of client which is the borderline between these two. Clients are indeterminate if they are delinquent but do not exceed the given DPD threshold. In our case when DPD is between 30 to 90 days it is considered indeterminate.

The clients with a very short credit history (less than a year) are called **insufficient.**

The clients with significantly misleading data (fraudsters) are called **excluded**.

The **rejected** clients are the clients whose credit application was rejected due to some documentation issue.

Analysis of Data

Now we need to analyze the inputs to the model i.e. the independent variables. This should be done before building the model but, in some cases, the model is already built. In this case, it makes sense to analyze the input features first before analyzing the prediction of the credit scoring models.

a. Check the correlation of the features:

HEATMAP ON DEVELOPMENT DATA

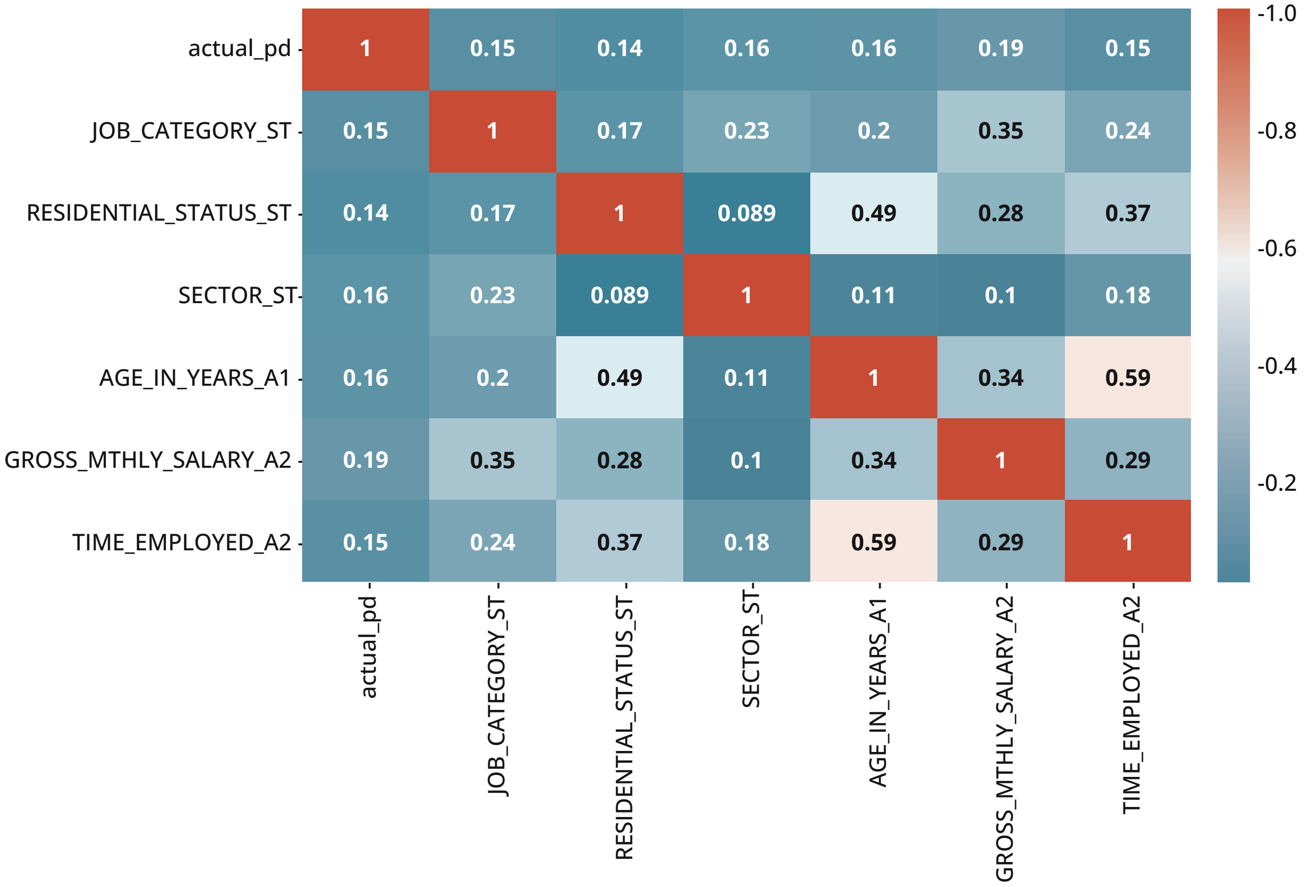


Fig: Heat-map of Pearson's Correlation Matrix for the development data of Demographic model

If high correlations are found between two variables, either we need to drop one of them or merge them into one combined variable to get rid of the multicollinearity effect.

Observation:

The correlation value between AGE_IN_YEEARS_A1 and TIME_EMPLOYED_A2 is 0.59 which is greater than average. Either one feature needs to be dropped or we should combine them into one variable.

b. Normalize the data:

The data should be normalized to bring all the variables to the same range to get faster convergence and better prediction accuracy.

A Min-Max scaling is typically done via the following equation:

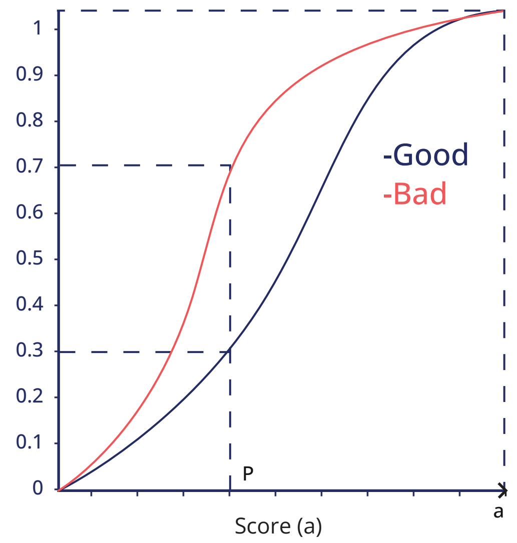
$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

We are not going to discuss the training of the model; the main objective of this blog is to validate the model which is already trained. Logistic Regression and Decision Tree models are widely used as credit decisioning models. In our case, we used a logistic regression model.

Validating Discriminatory Power of Model

Discriminating power is the ability to discriminate between defaulting and non-defaulting borrowers. The two most popular quality indices that are widely used to measure the discriminating power of the credit scoring model are**Kolmogorov-Smirnov statistics** and **Gini index**and both are based on **Cumulative Distribution Functions (CDFs).**

a. Kolmogorov-Smirnov (KS) statistics



CDF of Good and Bad clients

It looks at the maximum difference between the CDF of good clients and the CDF of bad clients.
Si is the score of individual clients and the class label is Dk.

$$D_k = \begin{cases} \mathit{1,\ client\ is\ good} \\ \mathit{0,\ otherwise} \end{cases}$$

Let there are n good and m bad clients in the dataset.

CDF of good client

$$F_{n.good}(a) = \frac{1}{n} \sum_{i=1}^{n} I\,(\,S_i \leq a \wedge D_k = 1\,)$$

CDF of bad client

$$F_{m.bad}(a) = \frac{1}{n} \sum_{i=1}^{n} I\,(\,S_i \leq a \wedge D_k = 0\,) \text{ where } a \in [\,L,H\,]$$

Where ⊠ ⊠[⊠,⊠] and I is the indicator function where I(true) = 1 and I(false) = 0. L and H are the minimum and maximum values of a given score respectively.

$$KS = \max_{a \in [\,L,H\,]} \left| F_{m.bad}(a) - F_{n.good}(a) \right|$$

In the above picture, when the KS score is ≤⌧, there are 30% good and 70% bad clients.
Mathematical Interpretation:

Decile	Min Score	Max Score	Bad	Good	Total	Good Rate%	Bad Rate%	Cum Good Rate%	Cum Bad Rate%	KS
1	0.016565	0.022967	2	514	516	99.61%	0.39%	11.02%	0.40%	10.61
2	0.02297	0.028197	5	511	516	99.03%	0.97%	21.97%	1.42%	20.56
3	0.028208	0.034067	16	500	516	96.90%	3.10%	32.69%	4.66%	28.03
4	0.034083	0.041497	12	504	516	97.67%	2.33%	43.49%	7.09%	36.41
5	0.041517	0.051581	19	497	516	96.32%	3.68%	54.15%	10.93%	43.22
6	0.051584	0.065372	29	486	515	94.37%	5.63%	64.57%	16.80%	47.76
7	0.065407	0.088412	52	464	516	89.92%	10.08%	74.51%	27.33%	47.18
8	0.088486	0.13294	86	430	516	83.33%	16.67%	83.73%	44.74%	38.99
9	0.132981	0.234112	94	422	516	81.78%	18.22%	92.78%	63.77%	29.01
10	0.234237	0.951121	179	337	516	65.31%	34.69%	100.00%	100.00%	0

KS statistics ranges between 0% (worst) and 100% (best). Better the KS, better the model. Practically KS value of 30% to 70% is considered as good for discriminating power of a model.

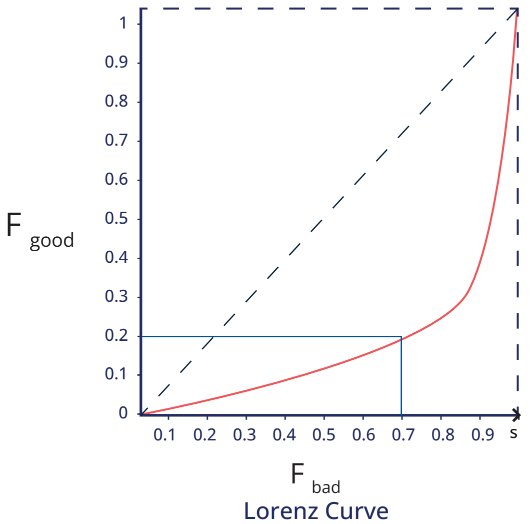
Observation:

From the above table, we can observe a KS value of 47.76% at decile 6 that indicates the good discriminating power of the credit scoring model.

b. Lorenz curve and Gini Index

Receiver Operating Characteristics (ROC) are used to measure the performance of a classification model. But in the financial industry, it is used to determine what proportion of bad and what proportion of good clients is rejected at every point on this curve or at every threshold. So, the problem that is faced in credit scoring is just the same as the classification problem. But instead of plotting FPR vs TPR, this ROC curve, also called Lorenz Curve is plotted between the percentage of bad customers and the percentage of good customers. These are just other names of FPR and TPR. There is another difference between the ROC curve of the classification model and the credit scoring model. Unlike the classification model, the score is ordered in ascending order, so the curve moves downward the diagonal line.

There is another way of plotting this Lorenz curve or ROC curve. One can plot the CDF score of bad clients vs the CDF score of good clients. Both will result in the same plot.



Lorenz Curve

Each point on this curve represents some value of a given score. For any cut-off score, we can find what will be the percentage of rejection for bad and good clients. We can see from the curve that at a cut-off value g (let g=0.3) we reject 70% bad clients at the same time we reject 20% good clients as well.

The next quality measure is the Gini index which is calculated based on Lorenz Curve.

$$Gini = \frac{A}{A+B}$$

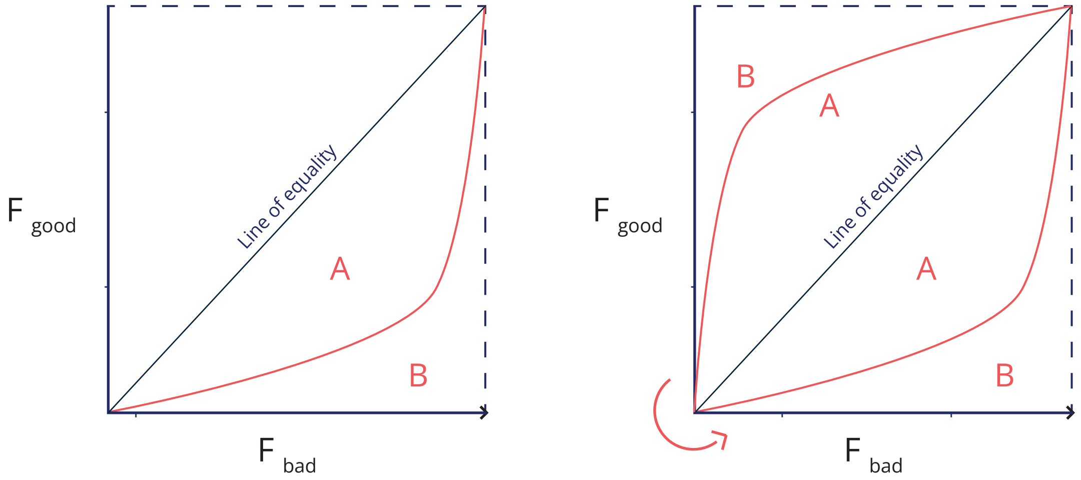
By inspection A + B = 0.5. So,

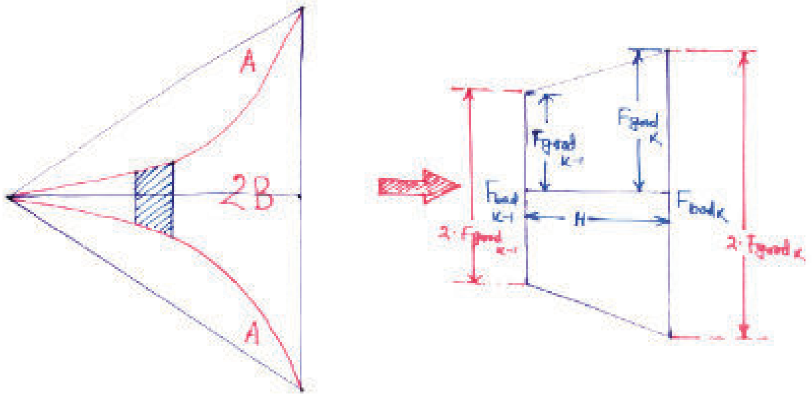
$$Gini = \frac{A}{A+B} = \frac{A}{0.5} = 2A$$

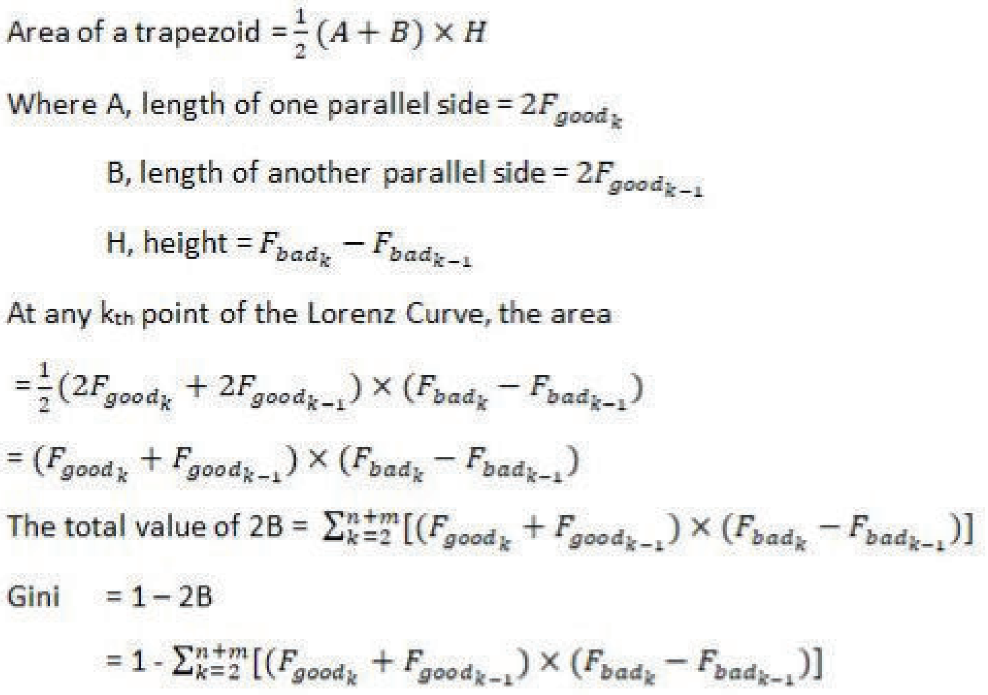
$$Gini = \frac{A}{A+B} = 1 - \frac{B}{A+B} = 1 - \frac{B}{0.5} = 1 - 2B$$

How to calculate the value of 2B?

First, move part2 below part1 as shown in the image. If we take a small section of the Lorenz curve, we get a trapezoid.







Area of a trapezoid $= \frac{1}{2}(A + B) \times H$

Where A, length of one parallel side = $2F_{good_k}$

B, length of another parallel side = $2F_{good_{k-1}}$

H, height = $F_{bad_k} - F_{bad_{k-1}}$

At any k_{th} point of the Lorenz Curve, the area

$$= \frac{1}{2}(2F_{good_k} + 2F_{good_{k-1}}) \times (F_{bad_k} - F_{bad_{k-1}})$$

$$= (F_{good_k} + F_{good_{k-1}}) \times (F_{bad_k} - F_{bad_{k-1}})$$

The total value of 2B = $\sum_{k=2}^{n+m}[(F_{good_k} + F_{good_{k-1}}) \times (F_{bad_k} - F_{bad_{k-1}})]$

Gini = 1 – 2B

$$= 1 - \sum_{k=2}^{n+m}[(F_{good_k} + F_{good_{k-1}}) \times (F_{bad_k} - F_{bad_{k-1}})]$$

It measures the global quality of a scoring function and ranges between -1 to +1. The ideal model which perfectly separates good and bad clients have a Gini index equal to 1. The Gini index value for a random model i.e. a model that assigns random scores to a client is 0. Negative values of the Gini index mean there is something wrong that's why the model is predicting good clients as bad and vice versa. In that case, we need to reverse the meaning of the scores.

Gini	AUC	Result
>=0.5	>=0.75	Strong
0.3 - 0.5	0.65 - 0.75	Acceptable
<0.3	< 0.65	Weak

Observation:

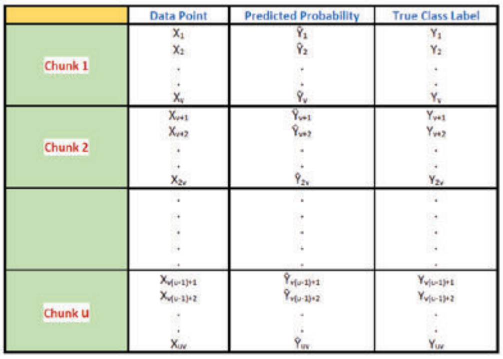
We found Gini score = 0.56 on OOT data which indicates a strong model

Validating calibration of the model

In credit scoring, we should be very careful about the probability estimate i.e. the confidence of a data point belonging to a particular class. Here the score generated from the classification model gives the predicted probabilistic interpretation of a point belonging to class 1 that should be in line with the observed probability distribution in the training data.

Let's explain this in detail. First, we sort the data points based on the predicted score in increasing order. Then split all the data points into k chunks, m points each. The average of the predicted score of m points in every chunk gives thepredicted probabilityfor that particular chunk. On the other hand, the fraction of positive points in every chunk gives theobserved probabilityfor that particular chunk. If the predicted probability and observed probability values for every chunk are close to each other, we can say the model is calibrated.

Let there are u chunks and in every chunk there are v points.



	Data Point	Predicted Probability	True Class Label
Chunk 1	X_1 X_2 . . X_v	$\hat{Y}_1$ $\hat{Y}_2$ . . $\hat{Y}_v$	Y_1 Y_2 . . Y_v
Chunk 2	X_{v+1} X_{v+2} . . X_{2v}	$\hat{Y}_{v+1}$ $\hat{Y}_{v+2}$ . . $\hat{Y}_{2v}$	Y_{v+1} Y_{v+2} . . Y_{2v}
	. . . . .	. . . . .	. . . . .
Chunk u	$X_{v(u-1)+1}$ $X_{v(u-1)+2}$ . . X_{uv}	$\hat{Y}_{v(u-1)+1}$ $\hat{Y}_{v(u-1)+2}$ . . $\hat{Y}_{uv}$	$Y_{v(u-1)+1}$ $Y_{v(u-1)+2}$ . . Y_{uv}

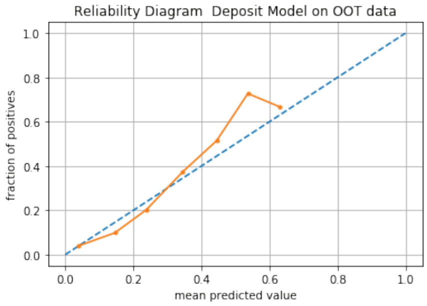
Predicted Probability for every chunk = avg-ŷi= $\frac{1}{v}\sum_{i=0}^{v} y\hat{}i$

Observed Probability for every chunk = avg-yi= $\frac{1}{v}\sum_{i=0}^{v} yi$

These avg-ŷi and avg-yi values should be close for a well-calibrated model.

Observation:

We plot predicted probability vs observed probability of every chunk for visual representation.



- Between 0.0 and 0.3 we can see that the model is under-reporting i.e. predicted score is below the observed score.
- Between 0.3 and 0.6 the model is over-reporting i.e. predicted score is above the observed score.

How a global Bank has benefitted from our Model Validation

Model validation for a Caribbean Bank

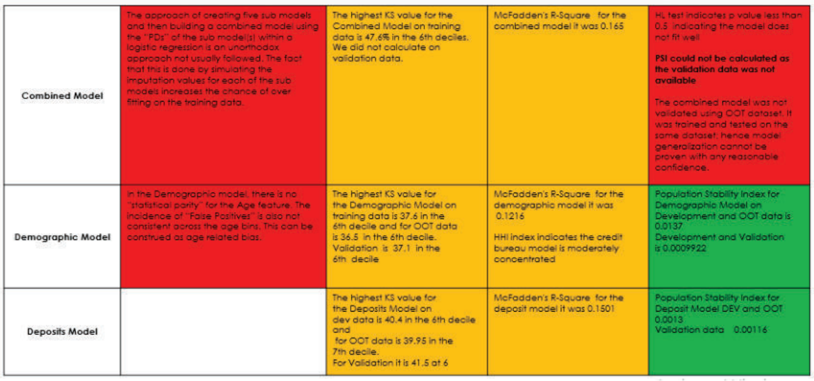
Background

- The credit risk model for a bank has been developed over a period of time by various parties.
- The bank needs to understand the model usage across consumer, non-consumer, secure, unsecured credits, and get it validated, for productionizing.

Approach

- Divided into categories like consumer, non consumer, secured and non-secured.
- The model has been validated at various levels like the data, discriminatory power and calibration, based on Basel II and III, IFRS standards.
- Both Point in time and Through the cycle ratings have been considered.

Key Observations



Combined Model	🟥 The approach of creating five sub models and then building a combined model using the "PDs" of the sub model(s) within a logistic regression is an unorthodox approach not usually followed. The fact that this is done by simulating the imputation values for each of the sub models increases the chance of over fitting on the training data.	🟨 The highest KS value for the Combined Model on training data is 47.6% in the 6th deciles. We did not calculate on validation data.	🟨 McFadden's R-Square for the combined model it was 0.165	🟥 HL test indicates p value less than 0.5 indicating the model does not fit well **PSI could not be calculated as the validation data was not available** The combined model was not validated using OOT dataset. It was trained and tested on the same dataset; hence model generalization cannot be proven with any reasonable confidence.
Demographic Model	🟥 In the Demographic model, there is no "statistical parity" for the Age feature. The incidence of "False Positives" is also not consistent across the age bins. This can be construed as age related bias.	🟨 The highest KS value for the Demographic Model on training data is 37.6 in the 6th decile and for OOT data is 36.5 in the 6th decile. Validation is 37.1 in the 6th decile	🟨 McFadden's R-Square for the demographic model it was 0.1216 HHI index indicates the credit bureau model is moderately concentrated	🟩 Population Stability Index for Demographic Model on Development and OOT data is 0.0137 Development and Validation is 0.0009922
Deposits Model		🟨 The highest KS value for the Deposits Model on dev data is 40.4 in the 6th decile and for OOT data is 39.95 in the 7th decile. For Validation it is 41.5 at 6	🟨 McFadden's R-Square for the deposit model it was 0.1501	🟩 Population Stability Index for Deposit Model DEV and OOT 0.0013 Validation data 0.00116

Value Delivered

- Shared validation report with observations around Model Development methodology, Data Analysis, Discriminatory Analysis and Model Calibration.
- Recommendations for potential improvements.

RoundSqr (Part of Cigniti)

Methods to calibrate a model

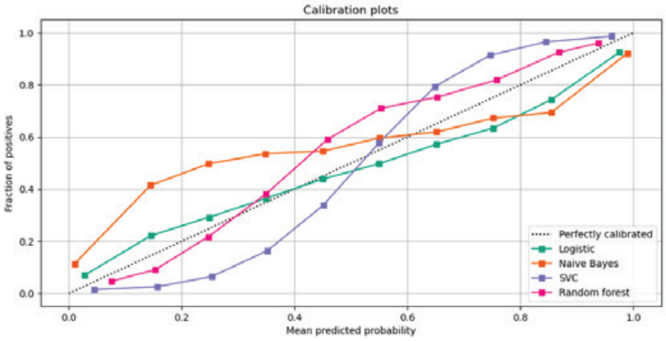
There are two popular methods to calibrate the model, Platt's Scaling/Sigmoid calibration and Isotonic calibration.

a. Platt's Scaling/Sigmoid calibration

Platt scaling is a way of transforming classification output into a probability distribution. The formula is mentioned below:

$$P(y = 1|x) = \frac{1}{1 + \exp(Af(x) + B}$$

f(x) is the predicted probability score from the model i.e. ŷi. A and B are the hyper parameters. Platt Scaling is most effective when the distortion in the predicted probabilities is sigmoid-shaped.

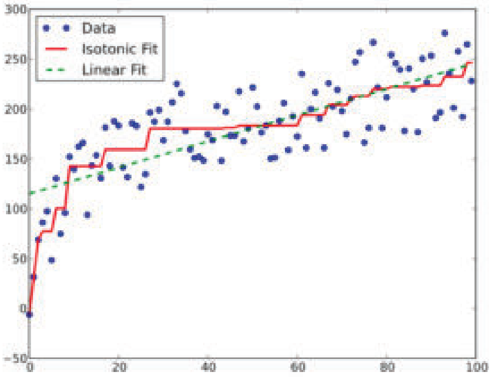


Source: https://en.wikipedia.org/wiki/Probabilistic_classification

b. Isotonic calibration

Isotonic Regression is a more powerful calibration method that can correct any monotonic distortion. This technique requires a huge quantity of data points for generalization because it is more prone to overfitting otherwise it will perform worse than Platt's calibration.

In calibration, we are trying to find a function that maps ŷi to yi i.e. f(ŷi) = yi



Source: https://en.wikipedia.org/wiki/Isotonic_regression

It breaks up the whole predicted curve into multiple (k) linear parts and for every part, it tries to fit into a linear regression model i.e. find the slope and intercept term for every section to match with the observed curve. This accounts for why isotonic regression requires many data points. The number of parts that we divide into should not be too large or too small, this is a hyper parameter.

Best Practices:

The PD model must be back-tested periodically on historical data in order to make solid predictions.

The behavior Score carding model i.e. Probability of Default of customers should be tracked even after approving the loans. This is critical to understand the exposure to credit risk on an ongoing basis.

Model interpretability is very crucial in credit scoring. It will be good to use a single PD model that is developed using all the relevant features rather than building separate models and combining them.

The credit decisioning model must be tested on separate Out of Time (OOT) data for testing the generalization capabilities.

To understand the discriminating power of the model, we should measure both the Gini score and KS value to take a decision on it.

Calibration is very important for the credit decision model to get a proper probability estimate. If the model is not calibrated properly it may underreport or overreport risk for customer segments in various deciles. Validating model calibration should be done at least once a quarter and it can be done using the Chi-Square test. If the $p\text{-value} < 0.05$ then we can say that the model is not fitting the data well.

Closing Thoughts

Applying analytics, process automation, and improved governance to credit risk modeling and decision-making are outline the following benefits:

- Increased efficiency of credit risk models: Create fresh machine-learning models to increase the accuracy of predictions.
- More information has led to wiser credit judgments: Utilize decision logic and analytics to address consumer demands with the appropriate credit offer at the appropriate moment.
- Decreased time in the decision-making process: Automate credit decision-making procedures to cut down on decision-making time.

Why Cigniti

Both experience and efficiency can be positively impacted with AI/ML interventions that are done responsibly. This capability of AI/ML is no longer just a differentiator for an organization but is a critical cog in the wheel. Combined with the right data & insights capability and RPA expertise, AI/ML is one area that is a slingshot for enterprises, especially for those that are geared towards rapid digital transformation.

Success Stories

Senior IT leaders share how our services helped them win in the platform age.

CTO Speak

“Data pipeline buildout, Software development, Salesforce development, AWS System admin/DevOps, BI/Dashboard. The execution has been very good.

- Satyadeep “Bobby” Patnaik, CTO

Lafayette Square

CEO Speak

“They understood that product development was iterative and they patiently worked through our requirements even as they rapidly evolved.

- Dr. Ganesh Naidoo, CEO

medmate

CTO Speak

“Proven technical ability in both web and mobile development; strong project/product management expertise; the ability to become part of the extended BA365 team.

- Graeme Dollar, CTO

BUSINESS ACCELERATOR 365

COO Speak

“I have worked with hundreds of service providers and consultants, RoundSqr (Part of Cigniti) is absolutely one of the best. The people are highly skilled, very hard-working, and have a “can do” attitude.

- Mark Mortimer, COO

Adelphoi

Analyst Recognitions

NelsonHall recognized Cigniti as a “LEADER” i n its 2022 NEAT evaluation for Overall **Quality** Engineering, Continuous Testing, Application **Security Testing, and AI &** Cognitive

Cigniti is mentioned as a “Pure Play Testing Vendor” in Gartner’s Market Guide for Application Testing Services, 2022

Cigniti is mentioned as “API Testing Vendor” in Gartner’s Hype Cycle for Managed IT Services and APIs, 2022

Recognized as “LEADER” in ISG IPL for Next-Gen ADM Services under Continuous Testing Specialists category for the US region for 2021 and 2022

Recognized as “RISING STAR” in UK Region in ISG IPL for Next-Gen ADM Services 2022

Cigniti is recognized as a **Strong Performer** in the Forrester Wave: **Continuous Automation and Testing** Services, Q3 2021.

One of the top 3 leaders in Agile Testing and DevOps in the Forrester Wave: Continuous Automation and Testing Services, Q3 2017.

Provides Cigniti with **“Best in Class”** rating for Buyer satisfaction.

This tech brief is issued for information only. Cigniti declines all responsibility for any errors and any loss or damage resulting from use of the contents of this tech brief. Cigniti also declines responsibility for any infringement of any third party's Intellectual Property rights but will gladly acknowledge any IPR and correct any infringement of which it is informed.

About Cigniti

Cigniti Technologies Limited (NSE: CIGNITITEC; BSE: 534758) is the World’s Leading AI & IP-led Digital Assurance and Digital Engineering Services Company providing software quality engineering, software testing, automation, and consulting services. 4100+ Cignitians worldwide help Fortune 500 & Global 2000 enterprises across 24 countries accelerate their digital transformation journey across various stages of digital adoption and help them achieve market leadership by providing transformation services leveraging IP & Platform-led innovation with expertise across multiple verticals and domains.

Our global customers have benefitted with measurable outcomes, millions of dollars of savings, significant ROI, and delightful, frictionless experiences utilizing our flagship digital assurance and full cycle software quality engineering services. Our AI-led digital engineering services cover Data engineering services, softwareplatform, and digital product engineering, AI/ML engineering services, intelligent automation, big data analytics, and Blockchain development.

We are headquartered in Hyderabad, India, with global offices spread across the USA, Canada, UK, UAE, Australia, South Africa, Czech Republic, and Singapore.

To learn more, visit www.cigniti.com

Website LinkedIn YouTube Blog Facebook Twitter