Read me

05 April 2024 06:39

A MODEL DEVELOPMENT :

1) Dataset : preprocessing < continuous Variables / Discrete Variables

2) Dependent Variable / Target :

Good → Non-default → 1

Bad → Default → 0

3) Splitting the data < train (80%) < targets (373028,) / inputs (373028, 207)
test (20%) < targets (93257,) / inputs (93257, 207)

4) More processing of data

↳ which is not needed for test data but needed for train data for modeling

Independent Variables < Continuous / Discrete } → Categorical Variables ↓ Dummies ✓

* Dummy Variables
* Binning on the basis of WOE

WOE = How much evidence does the independent variable has w.r.t Variation in the dependent Variable

$$WoE_i = \ln\left(\frac{\%(y=1)_i}{\%(y=0)_i}\right) = \ln\left(\frac{\%\,Good}{\%\,bad}\right)$$

$$= \ln\left(\frac{\%\ Non\text{-}default}{\%\ default}\right)$$

. . line

classing < fine / Coarse

$$IV = \sum_{i=1}^{K} \left[(\%.\text{good} - \%.\text{bad}) \times \ln\left(\frac{\%.\text{good}}{\%.\text{bad}} \right) \right]$$

(K-categories)

IV → Can be used to identify the variables which have better predictive power w.r.t dependent variable.

5) MODEL BUILDING :

Feature/inputs ↓ Co-efficients ↓ p-values ↓

filtering out the significant features based on p-value (< 0.05 → Rejecting the null hypothesis

(Co-efficient → significant)

6) Higher co-efficients → greater the odds of being a good borrower

observation → odds of categories → can be compared with the help of → Reference categories

7) As we have trained the model (using train)

Now we have to test (using test)

Now we have to test (using test)

VALIDATION :

i) probabilities will be estimated

Decide on the cut-off (0, 1) Range

ii) Confusion Matrix

		Predicted 1	Predicted 0
Actual	1	TP	FN
	0	FP	TN

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

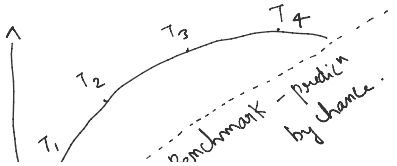
$$\text{Precision} = \frac{TP}{TP + FP}$$ [p^3 - precisely predicted positives out of Total positive predictions]

(TPR) $$\text{Recall} = \frac{TP}{TP + FN}$$ [Rightly Identified positives out of total positive]

4) Specificity / TNR $= \frac{TN}{TN + FP}$

5) F1 = Harmonic Mean of precision & Recall, $\rightarrow \left[\frac{2PR}{P + R}\right]$

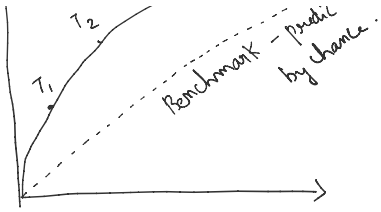
↳ ROC : 'threshold' → ***



AUROC - Area Under ROC

Prediction power of ?

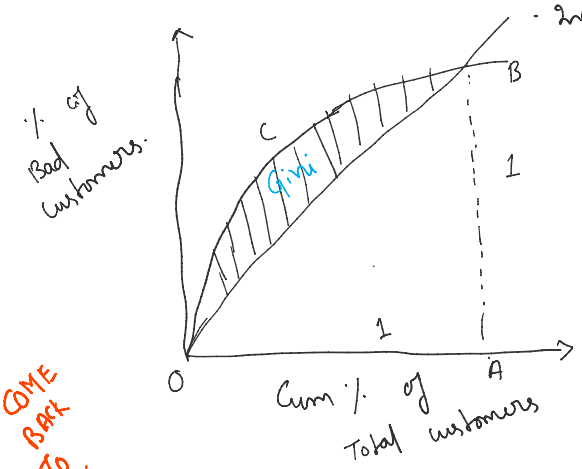
Prediction Power of Model ?



↳ Gini:

(Measure of inequality) (Originally Economics - Income Inequality)

Discriminatory power of model



COME BACK TO CHECK THIS →

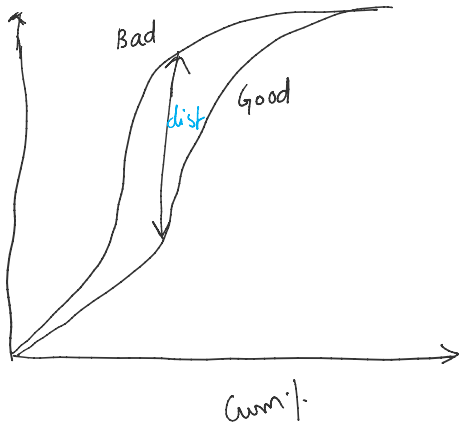
Area $\Delta OAB = \frac{1}{2} \times 1 \times 1 = \frac{1}{2}$

Gini = Shaded reg^n

AUROC = O C B area under

Gini = AUROC - 1/2.

Kolmogorov - Smirnov:



Perfect Model → Max. dist → KS = 1 (Very high)

predic^n by chance → Max. dist → KS = 0 (Very low)

↳ Applying Model:

Suppose new customer

falls under what all dummy variables → Consider only those co-efficients

(if my_cat → then Co-eff = 0)

Sum of all co-eff → prob of non-default

(1 - prob of non-default) → prob of default

↳ Scorecard_Development: (including my_categories) → for easier interpretations.

(Min = ? → a borrower falls into lowest co-eff categories (zeros/negatives) = -1.49

Scorecard_description (including rej_categories) → for easier interpretations.

Scores { Min = ? → a borrower falls into lowest co-eff categories (zeros/negatives) = -1.49
Max = ? → a borrower falls into highest co-eff categories = 5.6332

Rescaling co-efficients → Scores

$$Variables\_Score = Variables\_Coeff \times \frac{(max\_score - min\_score)}{(max\_sum\_Co\text{-}eff - min\_sum\ co\text{-}eff)}$$

scaling intercept to around min_score

Min Score - 300
Max Score - 850
scaling done ✓ scoring ✓

↳ Calculating Credit Scores: ✓

↳ Calculating PD from credit score: (reverse the process)

Banks → more business → lower cut-off of PD
less " → higher " " "

Setting the cut-offs:

from ROC calculations ← FPR 1) TPR 2) thresholds 3)

Derive 'Good' probs from scores

TABLE: thresholds FPR TPR Score

Now lets do the 1) Approval Rate 2) Rejection Rate

3) No. of Approved Applicants. ✓

4) No. of Rejected Applicants ✓

Based on our Risk appetite:

Export scorecard to CSV

↳ Model monitoring:

check whether the model is still relevant

Suppose → now data is 2007 to 2014

Suppose → now data is 2007 to 2014

we got new data → 2015

whether our model still relevant for this '2015' data

for that we need to calculate the **Population Stability Index** PSI

PSI = formula

'PSI' calculated for all features.

PSI ∈ [0, 1]

Interpretation of values

=0 No difference in population

=1 Absolute difference.

Generally PSI > 0.25 → indicating to rebuilt the model/changes to be made (action to be taken)

(check score only)

tion changed significantly.