☆ Bank Case Study

☆ Predicting Bank Loan Default for Bking Industry

☆ Sales Strategy to influence existing / new customers

Cross Selling	Up-Selling
To buy complementary products / services in addition to org. purchase	To purchase a higher valued product or services of same product line.

Increasing Revenue

Aim for effectiveness

☆

Minimizing Customer Churn [Attrition]

When customers leave your product / service due to several reasons.

Aim to (↑) Promotional Offers

Acquisition Customer

Creating or acquiring customers by luring them with attractive offers.

Aim for customer experience

✱ Predicting Fraudulent Activity

Illegal activities in order to receive money, funds, credits from bk. / Institution.

- Phising
- Cr./Db Scam
- Check Scams
- Accounting fraud
- Identity fraud
- Money laundering

✱ Predicting Bank Defaults with Log Regr.

Logit model demonstrates a regrn. model where, resp. var. is binary / dichotomous & indept. var. can be binary / continuous/ordinal

✱ Logit types:
- Bi-Nominal = Default / Not-Df.
- Multi-Nomial = Yes / No / Maybe [No. Res]
- Ordered = Ratings

Eg:
$$\ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_n x_n$$

where, P = Probl. of Binary outcome.

Odds

★ Odds Ratio = $\dfrac{\text{Prob. of occurrence of events}}{\text{Prob. of occurrence of non events}}$

Table	D	ND	Total
U	50	10	60
E	40	20	60

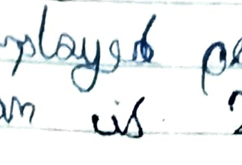
$$\text{Odds Ratio}_U = \frac{\left(\frac{50}{60}\right)}{\left(1-\frac{50}{60}\right)} = \frac{(0.8)}{(0.2)} = 4$$

$$\text{Odds Ratio}_E = \frac{\left(\frac{40}{60}\right)}{\left(1-\frac{40}{60}\right)} = \frac{(0.6)}{(0.4)} = 1.5$$

$$\text{Odds Ratios} = \frac{OR_U}{OR_E} = \frac{4}{1.5} = 2.66$$

★ The odds for an unemployed person to default on a loan is 2.66 times higher than employed

★ Logit Model Curve = S – Shape

= 

* Assump.ⁿ

- Resp. Var. must be binary
- Logit model estimates $P(Y=1)$, so resp. var. must abide by this assump.ⁿ & outcome must be in line with this code
- Model should be as perfect fit as possible nor over neither under
- Indept. var.'s must not be correlated to each other.
- Indept. var.'s must be linearly correlated to log odds.
- Sample size must be large

* Model Evaluation

- Likelihood ratio test : Used to compare (2) nested glm's & if p-value for the full model < 0.05, then its a better fit.

  Hyp : H_0 : Full model better, against
  H_a : Reduced model better

$$\chi^2 = (-2)\ln(\text{Reduced}) - (-2\ln(\text{full model}))$$

Telco Dt.: Pg.:

★ Not recommended

- Hosmer Lemeshow Test: Use Pearson-Chi sq. test to check whether obs. prop. of events are same to predicted probs. in model popl. subgroup (10 groups)

$$H = \sum_{g=1}^{10} \frac{(O_g - E_g)^2}{(E_g)}$$

where,

O_g = No. of obs. events.
E_g = No. of expected events.

Hyp: H_0: Model is good, against
H_a: Model is poor

★ Sig. of Indv. Indept. Var. [Tests to check]

★ Wald Statistic Test: Helps us determine the importance of an individual dept. var. by logistic regn. coeff.

$$W_j = \left(\frac{B_j}{SE_{B_j}}\right)^2$$

Hyp: H_0: Coeff. of Interest = 0, against
H_a: " " " $\neq 0$

If WST accepts H_0 ⇒ Indept. var. will not impact model's fit

Drawback - Large coeff. could get removed becoz of (↑) SE, as a result, we could underfit the model.

2) Likelihood Ratio : $G = (-2)(\ln(\text{reduced}) - \ln(\text{full}))$

The smaller the deviance b/w both models the better is correl. b/w dept. & Indept. var

Predictive Value Validation [Model Accuracy]

Few measurements like Confusion Matrix & Receiver operating characteristic helps us see, how accurately the model is predicting dept. var.

- Confusion Matrix: Technique used to evaluate predictive accuracy of the model

Conf. Mat.

		Predicted class x	Predicted class y
Actual class	X	TN	FP
	Y	FN	TP

True / False - Actual Outcomes
+ve / -ve - Predicted Outcomes

(+ve) = Predicted yes , (−ve) = Predicted No.

= H_0 = Customers would not churn, against

= H_a = Customers would churn

		Predicted	
		(+ve)	(−ve)
Actual	T	Correct	Correct
	F	Type I Error.	Type-II Error

FP = Not occuring predicted as occurred

FN = Occurring predicted as not occurred

		Actual (Disease) +	Actual (Disease) −	
Predicted (claim)	+	True (+)	False (+)	Type-I Error
	−	False (−)	True (−)	
		Type-II Error		

☆ Accuracy Rate = $\frac{\text{Accurately predicted}}{\text{Total Outcomes}}$

☆ Error (Misclassification) Rate = 1 − Accuracy Rate

(Sensitivity)

★ $T(+ve)$ Rate = $\frac{\text{Accurately predicted } (+ve)}{\text{Total actual } (+ve)} = \frac{TP}{TP + FN}$

★ $F(+ve)$ Rate = $\frac{\text{Inaccurately predicted } (+ve)}{\text{Total actual } (-ve)} = \frac{FP}{FP + TN}$

★ Specificity = $\frac{\text{Accurately predicted } (-ve)}{\text{Total actual } (-ve)} = 1 - FPR$

★ Precision = $\frac{\text{Accurately predicted } (+ve)}{\text{Total } (+ve)} = \frac{TP}{TP + FP}$

★ Prevalence = $\frac{\text{Actual } (+ve)}{\text{Total Outcomes}} = \frac{TP + FN}{TN + FN + TP + FP}$

★ Receiver Operating Curve : Used to measure binary classifier performance visually & AUC is used to quantify model performance.

★ Plots TPR on y axis, against FPR on x axis

★ Generally, AUC > 70% = Accurate Model
So, the more AUC ~ 100%, the more accuracy

★ ROC = Above diag = Better
ROC = Below diagonal = Worse
ROC = ≈ 1 = Best