Research Paper *IJBARR*
Impact Factor: 3.072 *E- ISSN -2347-856X*
Peer Reviewed, Listed & Indexed *ISSN -2348-0653*

DEVELOPING RISK SCORECARD FOR APPLICATION SCORING AND OPERATIONAL EFFICIENCY

Avisek Kundu* Ms. Seeboli Ghosh Kundu**
**Senior consultant – Ernst and Young.*
***Senior Lecturer – ITM Business Schooland Research scholar- BharathiarUniversity.*

Introduction

The Commercial Tax Department of West Bengal has always been exposed to the riskiness of the different dealers in different sectors in incorporating malpractices in terms of financial abnormalities and de-regularities. The differentdealers working in this sector many a times take the advantage of the interstate suppliers of raw materials and try to evade the regularities through differencing the input and the output taxes. Only after the audit these irregularities were captured when there is substantial loss of resources and many a times when interventions were not possible. The continuous pressure on the operational departments in monitoring the transactional data & gauging the riskiness through their functional knowledge always creates operational bottlenecks and thus a statistical model for optimized revenues rather than the gut feeling is most sought after.

The desired requirement always was to create a predictive model or a scorecard which will predict dealers' likelihood of risk. Risk score generated can be used to scrutinize dealers for audit thus removing the continuous pressure on the operations. Predictive model gives impact of different risk parameters from returns, registration and other data modules. This traction time flag will safeguard significant resources and would result in actionable interventions at the transaction time only for optimal results and maximizing revenues.

Predictive Modeling (Logistic Regression): Risk Model for Operational Efficiency

Objective of Analysis

The Objective of this analysis it to develop a predictive risk model to predict dealers' likelihood of risk. Risk score generated can be used to scrutinize dealers for audit. Predictive model gives impact of different risk parameters from returns, registration and other data modules. It also helps to gauge the probability of default at the transaction time of the dealers based on the explanatory variables by predicting the outcome for right interventions for optimized outcome thus reducing the operational bottlenecks and increasing the efficiency.

Rationale for Analysis

The historical behaviour of dealers provides insight to predict future behaviour of dealers and therefore facilitate West Bengal Commercial Tax Department to categorize dealers into risky and non-risky dealers. Dealers risk profiling can be done by identifying the significant parameters from returns, registration and other internal data modules.

Sources of Data and Data Description

1. Dealer master
2. Registration data
3. Returns data
4. Audit risk output
5. A sample of approx. 1000 dealers was considered for the analysis for two financial years from 2012-13 to 2013-14.

Description of Technique used for model development:

Predictive analytics is used as a risk management tool which assists in determining centralized, uniform, more consistent and reliable decision management across business unit to meet defined business goals. Strategically, predictive analytics identifies precisely whom to target, how to contact, when to contact, and what message should be communicated thus creating an optimized strategy reducing operational bottleneck & increasing operational efficiency.

Data Understanding

1. Dealer data was obtained by compiling parameters from different data modules.
2. Dealer risk parameter was identified basis the audit risk output file. A dealer categorised as risk having been categories as risky in the past and otherwise non-risky.
3. 5 parameters were identified for the model post preliminary analysis and discussion with CTD officers

Data Preparation

1. Dealers data from return, registration and audit risk out was consolidated and imported to R statistical software for analysis

2. Inputs were collected from West Bengal CTD team for grouping of the variable and incorporating information in the predictive model
3. Data variables were grouped into broader categories based on the inputs
4. Derived new variables from existing variables

Model Development

1. Data split into model development and validation samples
2. Applied logistic regression on the development sample
3. Carried out several iteration of model and checked for model fit statistics
4. Choose the best model and generating risk scores for dealers

Predictive Model Output

Dependent Variable: A response variable was created as indicator of risk with value 1 and 0. 1 indicates dealer is risky whereas 0 indicates dealer is not risky.

Independent Variables: Following independent parameter were considered for the analysis

1. Age of Account
2. Output Input ratio
3. Total Tax Paid
4. Business Status
5. Business Type

Variable Selection

Data was imported into R for analysis. Logistic regression model technique was used and significant variables were selected after running multiple model iteration. Variable were selected based on Chi-square test statistics and best model was selected as mathematical equation with combination of all significant variables. In current model only two variables were identified as significant (age of account and Out Input ratio ratio). Age of account contributed significantly in the model and has positive impact on the risk variable i.e. chances of risk of dealer increase with increase in number of years in the system. Out tax to input tax credit ration was grouped into 3 groups (IO ratio equal to zero, IO ratio less than 1 and IO ratio greater than 1). OI ratio category of where OI ratio less 1 has positive impact on the risk variable i.e. dealer having OI less than 1 has higher chances of being at risk as capered to dealer with OI ratio greater than 1.

Variable Selection Summary

Variables Considered for Model	Categories	Significant variables	Definition	Model Coefficient	Odds Ratio	Impact/ Interpretation
Age of Account	NA	yes	Total number of years in the system	0.245	1.27	one unit increase in age of account the odds of being a risky dealer increase a factor of 1.27
Output Input ratio	0	yes	Output tax to input tax credit ratio	-		
	less than 1	yes	Output tax to input tax credit ratio	2.24	9.44	Dealers have higher chances of being risky if they have OI ratio less than 1
	greater than 1	yes	Output tax to input tax credit ratio	2.05	7.8	Dealers have less chances of being risky if they have OI ratio greater than 1 as compared to dealer having OI ratio less than 1
Revenue/ Tax Paid	1	yes	paid tax is greater than 1 cr	1.39	4.02	Dealers with tax paid greater than 1 crore has higher chances of being risky

Model coefficient: Model coefficients are the coefficient of the parameters (e.g. age of account coefficient is 0.245) estimated while training the model from the historical data. Parameter coefficient demonstrates type of relation (positive or negative) between Risk status (dependent variable in the model) and independent variables (age of account, revenue, output to input ratio). A positive sign of coefficient means that there is a positive correlation between independent and dependent variables. In case of age of accounts it's positive which indicates that increase in age of account would increase chances of a being a risky dealer. One unit increase in age of account would increase 1.27 units increase in risk status.

Odds Ratio: it's way describing the effect of dependent variable on independent variable in relative terms. It explains the effect of one dependent variable on Risk status variable by keeping rest independent variable at constant.

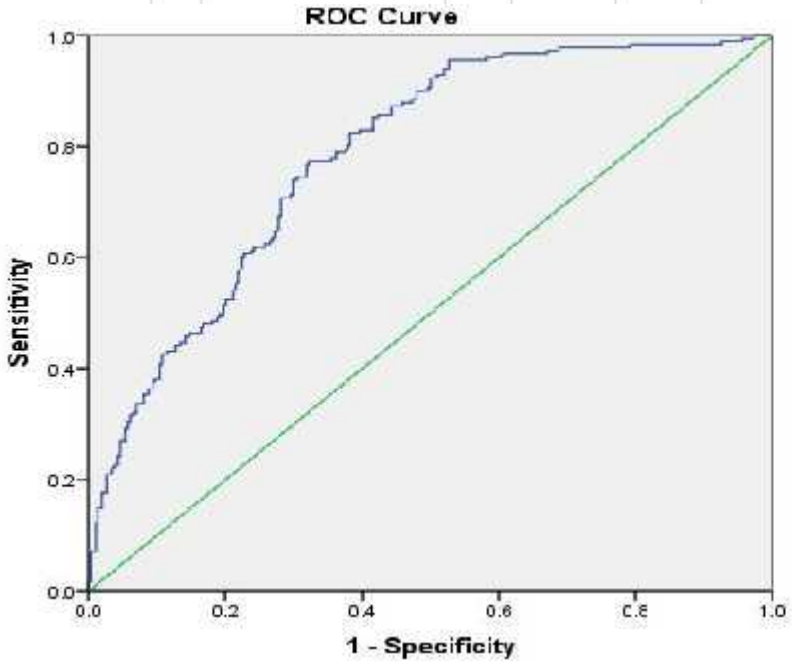
Risk Profiling is assigning risk scores to dealers in form of mathematical model represented as a set of weights of significant variable of model assigned to dealer's characteristics that affect tax paying loyalty of a dealer. Risk score were calculated from the model's mathematical equation and dealers were categorised as risky and non-risky.

Methodology

The entire data is divided into 2 parts namely the training and validation in the ratio of 70:30. The data is randomly divided with every observation given the equal chance of being picked thus removing bias. The Logistic Regression Model is created on the training data and the model validation is performed on the validation data. The three candidate models created using three different algorithms are full fit model, forward and backward stepwise regression. The candidate models are compared against each other in terms of misclassification. Further bucketing of the observations in terms of cumulative gain is calculated and after incorporating the profit matrix the optimized bucket is targeted and the threshold probability is zeroed for maximum impact.

The output from the Full Fit Model of Logistic Regression is as follows

							Sensitivity	1-Specificity	Specificity			
Row Labels	Risky	Total_Decile	Not Risky	Cu_Not Risky	Cu_Risky	Cu_Not Risky%	Cu_Risky%	Cu_Risky Avoided	Profit	P(0)_Threshold_min	P(1)_Threshold_max	
1	2	13	11	11	2	13.58%	3.92%	96.08%	100			
2	0	13	13	24	2	29.63%	3.92%	96.08%	1400			
3	2	13	11	35	4	43.21%	7.84%	92.16%	1900	0.784	0.216	
4	4	13	9	44	8	54.32%	15.69%	84.31%	400			
5	6	13	7	51	14	62.96%	27.45%	72.55%	-1900			
6	5	13	8	59	19	72.84%	37.25%	62.75%	-3600			
7	8	13	5	64	27	79.01%	52.94%	47.06%	-7100			
8	7	13	6	70	34	86.42%	66.67%	33.33%	-10000			
9	10	13	3	73	44	90.12%	86.27%	13.73%	-14700			
10	7	15	8	81	51	100.00%	100.00%	0.00%	-17400			
Grand Total	51	132	81	51	51	100.00%	100.00%	0.00%	-17400			



Area Under the Curve

Test Result Variable(s): Predicted probability

Area
.739

The test result variable(s): Predicted probability has at least one tie between the positive actual state group and the negative

	Pred_Risky	Pred_Non Risky	
Actual Risky	0	4	
Actual_Non risky	0	35	
		Classification Rate	0.897436

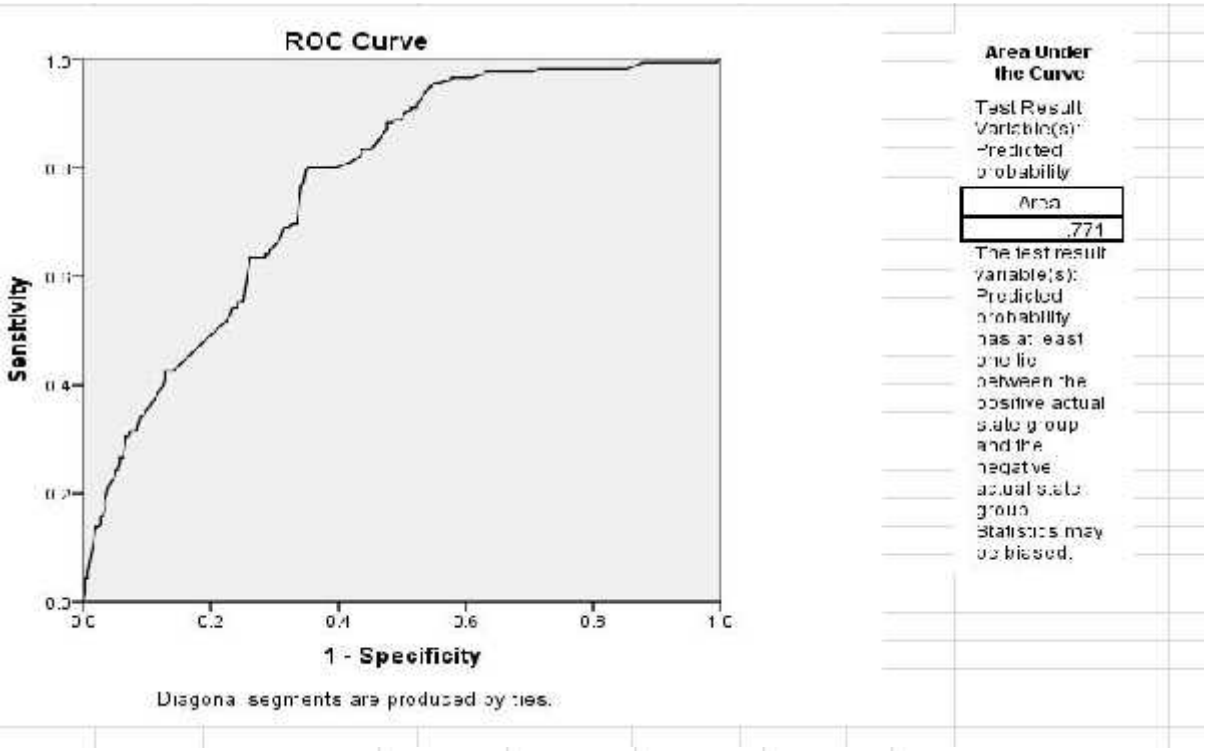
Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	11.424	8	.179

The output from Forward Logistic Regression is as follows:

	Predicted Risky	Predicted Non Risky
Actual Risky	0	6
Actual Non Risky	0	46
	Classification%	0.884615385

						Sensitivity	1- Specificity	Specificity	
Row Labels	Risky	Count of Decile	Non Risky	Cu Risky	Cu Non Risky	Cu Risky %	Cu Non Risky %	Cu Risky Avoided%	Profit
1	2	13	11	2	11	3.92%	13.25%	96.08%	100
2	1	13	12	3	23	5.88%	27.71%	94.12%	800
3	1	13	12	4	35	7.84%	42.17%	92.16%	1500
4	2	13	11	6	46	11.76%	55.42%	88.24%	1600
5	6	13	7	12	53	23.53%	63.86%	76.47%	-700
6	6	13	7	18	60	35.29%	72.29%	64.71%	-3000
7	9	13	4	27	64	52.94%	77.11%	47.06%	-7100
8	8	13	5	35	69	68.63%	83.13%	31.37%	-10600
9	7	13	6	42	75	82.35%	90.36%	17.65%	-13500
10	9	17	8	51	83	100.00%	100.00%	0.00%	-17200
Grand Total	51	134	83	51	83	100.00%	100.00%	0.00%	-17200



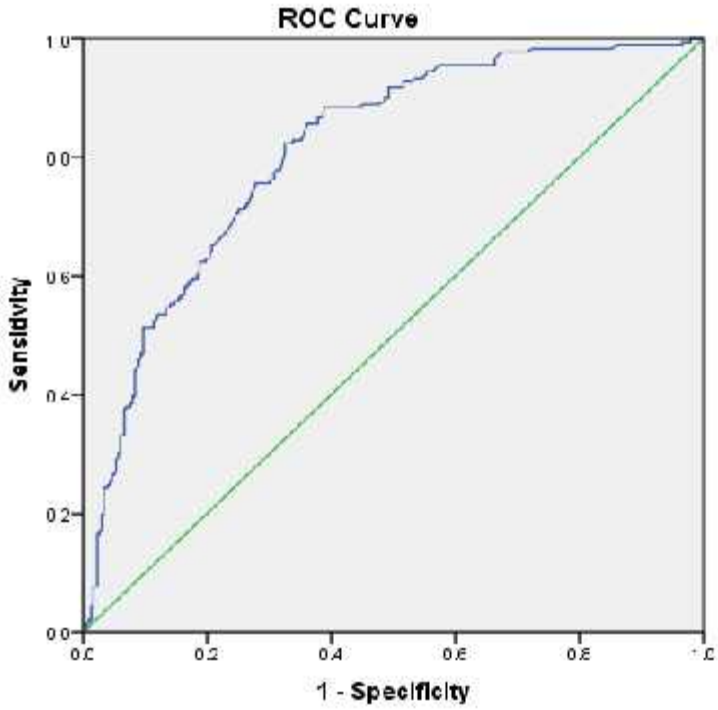
Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
6	7.724	8	.461

The output from the Backward Stepwise Logistic Regression is as follows:

	Predicted Risky	Predicted Non-Risky
Actual Risky	0	6
Actual Non Risky	0	46
Classification		0.884615385

						Sensitivity	1 Specificity	Specificity	
Row Labels	Risky	Total_each decile	Non Risky	Cu Risky	Cu Non Risky	Cu Risky%	Cu Non Risky%	Cu Risky Avoided%	Profit
1	2	13	11	2	11	3.92%	13.25%	96.08%	100
2	1	13	12	3	23	5.88%	27.71%	94.12%	800
3	1	13	12	4	35	7.84%	42.17%	92.16%	1500
4	2	13	11	6	46	11.76%	55.42%	88.24%	1600
5	6	13	7	12	53	23.53%	63.86%	76.47%	-700
6	6	13	7	18	60	35.29%	72.29%	64.71%	3000
7	9	13	4	27	64	52.94%	77.11%	47.06%	-7100
8	8	13	5	35	69	68.63%	83.13%	31.37%	-10600
9	7	13	6	42	75	82.35%	90.36%	17.65%	-13500
10	9	17	8	51	83	100.00%	100.00%	0.00%	-17200
Grand Total	51	134	83	51	83	100.00%	100.00%	0.00%	17200



Area Under the Curve

Test Result Variable(s): Predicted probability

Area
.811

The test result variable(s): Predicted probability has at least one tie between the positive actual state group and the negative actual state group. Statistics may

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	4.187	8	.840
6	6.734	8	.566

Predicted probability:

Using the logistic regression fitted model to do predictions for the validation data:

TIN_NO	BUS_TYPE	BUS_STAT	REG_TYPE	EXM_SALE	Risk_A	Agecode	Bus_risk	OI_code	Ave_C	SumofTOT	RevSum	SumofPAI	Business	Business	Bus_Type	CIRatio	Random	PRE_1	Risk_A	Dodic
2.93E+10	3	1	VAT	Y	0	9	1	3	0.788548	8.15E+08	0	0	AGENCY	SOLE PRO	A_Retiler	1.207236	10	0	0	1
2.93E+10	11	1	VAT	N	1	9	1	3	0.890079	4.77E+08	0	55802	SERVICES	SOLE PRO	Others	1.19139	10	0.00109	1	1
2.93E+10	5	1	VAT	N	0	1	1	1	0	37003482	0	0	RETAILER	SOLE PRO	A_Retiler	0	8	0.015286	0	1
2.95E+10	5	1	VAT	N	0	2	1	1	0	91181695	0	0	RETAILER	SOLE PRO	A_Retiler	0	9	0.018947	0	1
2.94E+10	1	8	VAT	N	0	1	2	1	0	0	0	0	MANUFAC	PRIVATE L	Manufact	0	10	0.019903	0	1
2.94E+10	5	1	VAT	N	0	3	1	1	0	0	0	0	RETAILER	SOLE PRO	A_Retiler	0	10	0.022333	0	1
2.93E+10	4	8	VAT	N	0	2	2	1	0	0	0	0	WHOLESA	PRIVATE L	Dist_WS	0	8	0.025434	0	2
3E+10	5	1	VAT	Y	0	1	1	1	0	38510488	0	0	RETAILER	SOLE PRO	A_Retiler	0	9	0.044881	0	2
2.94E+10	5	2	VAT	Y	0	1	1	1	0	160841	0	0	RETAILER	UNREGIST	A_Retiler	0	9	0.044881	0	2

Risk Status	
1	Risky
0	Non-Risky

- Table shows that actual risk status and predicted risk status based on the probability of risky dealer from the model.

- A dealer with predicted probability greater than 0.50 has been classified as risky and non-risky dealer otherwise.

Model Comparison among the candidate models across different criterions or parameters:

Thus it emerged that the classification rate at the optimized bucket (3rd decile) for full fit model is 89.74%, forward model is 88.46% at the optimized (4th decile) and 88.46% at the optimized (4th decile) for backward logistic regression. The hosmer lemeshow value for all the three candidate models are more than 0.05 (the default 5% significance) showcasing stable models. The ROC value for all the three models are more than 0.7 stating very strong and robust. The profit loss ratio is 1:5 stating that profit from 5 units of good dealers are neutralised by 1 bad dealer. Incorporating this profit matrix the optimized bucket in terms of profit is incorporated as showcased above.

Strategic Choices:

i) Conservative Approach:
This approach leads in selection of the full fit logistic regression with the classification rate of 89.74% at the optimized 3rd bucket (30th percentile) with the targeted profit being 1500 units with 43.21% of the entire good dealers are captured and 92.16% of the risky dealers avoided thus incorporating only 7.84% of the risky dealers.

ii) More Aggressive Approach:
This approach leads in selection of the backward stepwise logistic regression with the classification rate of 88.46% at the optimized 4thbucket (40th percentile) with the targeted profit being 1600 units with 55.42% of the entire good dealers are captured and 88.24% of the risky dealers avoided thus incorporating around 11.76% of the risky dealers. Though the return is high it incorporates more risk than the previous strategy.

Thus based on the intended approach the right strategy using this model would result in optimized profit and would incorporate interventions for the riskier dealers predicted at the transaction time rather than at the audit after the transactions ends reducing the operational bottlenecks and thus increasing the efficiency.

Recommendations and Benefits

- Risk rating model would help in identifying good and bad dealer based upon the key parameter and that would further help department easing the process of tax payment for good dealer and developing strategy for dealing with bad dealers through an optimized mechanism removing operational bottlenecks & increasing efficiency.
- Dealer can be sorted based upon the risk score generated by model for selecting the dealers for audit purpose again removing the operational pressure
- Risk rating of customers would help in reducing the efforts of auditing randomly without any prior analysis using the historical data analysis.
- Risk model would help West Bengal CTD in understanding and identifying the key significant parameter for measuring dealer's behaviour again streamlining the operations for optimized impact.

Thus based on the intended approach the right strategy using this model would result in optimized profit and would incorporate interventions for the riskier dealers predicted at the transaction time rather than at the audit after the transactions ends.

Bibliography

1. Bryman, A (2006) Integrating quantitative and qualitative research: how is it done?' *Qualitative research*, Vol.6, No. 1, pp. 97 – 113 Sage.
2. Saunders, M, Lewis, P Thornhill, A, 2007; *Research methods for business students*, 4th Edition, Prentice Hall
3. Creswell John W., (2003) *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. 2nd Ed. Sage Publications.
4. Creswell. J. W. & Miller. D. L. (2000): Determining validity in qualitative inquiry. Theory into Practice, 39(3), 124-131.
5. Charles, C. M. (1995). Introduction to educational research (2nd ed.). San Diego, Longman Churchill.G.A, Jr. &Lacobucci.D: Marketing Research, *methodological foundation, Tenth Edition* (2009)
6. Lee, L., & Billington, C. (2007). *The Evolution of Supply chain Management Models and practices at Hewlett-Packard.* Stanford: Depertment of Industrial Engineering and Engineering Management, Stanfor University.
7. Lun, V., & lAI, k. (2010). *Shipping and Logistic Management.* New-York: Springer.
8. Magee, F. (2008). *Modern Logistic Management:Integrated Marketing, Manufacturing and logistic system.* Canada: Jhon Willy and sons.
9. Mentzer, T. (2005). *Supply Chain Management.* United Kingdom: Sage Publication.