**FINANCIAL CROSS-SELL PROJECT REPORT**

**OBJECTIVE :**

To find out the customers who are more likely to buy cross-sell product (permanent disability insurance with a low monthly premium).

**a) From Marketing point of view**, such strategies to be developed based on ideal customer to get a list of customers buying the insurance product.

**b) From Business point of view**, a segment of customers should be targeted having propensity to buy cross sell insurance products.

**POINTS TO BE CONSIDERED :-**

=> To understand the basic concept of CROSS-SELLING products and various strategies to implement it the right way.

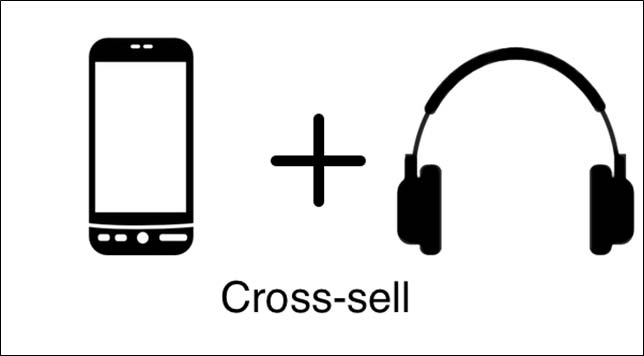
=> To extract such insightful information out of the data that guides towards the eligible group of customers to be targeted to buy cross-sell products.

=> To develop a clear understanding about how cross-sell products could be bought by a customer and be totally relevant at the same time.

**BRIEF ABOUT CROSS-SELL:-**

=> Cross-sell involves the sale of multiple products offered by a single product/service provider to a new or existing customer.

=> For example:- Suppose a customer plans to purchase a mobile phone worth Rs. 30,000(~$500), but the salesman offered him a charming deal of buying mobile phone with exclusive JBL headphones for Rs.40,000 (~$634) only and he again got swayed away. This is Cross-Selling (figure 1.)



**FIGURE 1:- Cross-Sell**

**APPROACH :-**

The objective of this project is accomplished by using supervised statistical techniques like **Logistic Regression** and **Multiple Linear Regression,** which are further elaborated.

|  |  |  |  |
| --- | --- | --- | --- |
| Statistical Technique | WHAT | WHEN | WHY |
| 1) Multiple Linear Regression | It is a tool to find the **association/ relation** between a dependent and a set of independent variables. | 1. Relation between dependent and independent is **linear** and in a **additive** form  2. Whenever the dependent is in a **continuous format/ numeric format.** | To find which independent variables **statistically explain** the dependent variable. |
| 2) Logistic Regression | **Logistic Regression is supervised classification technique which is used to predict the probability of the Dependent Variables (Categorical in nature)** given the independent variable | 1) Whenever the **Dependent is categorical in nature** -2) **Set of Statistical Assumptions( Classical Linear Regression) fail** to build the model 3) The objective is to **predict the probability of Dependent Variable as per the Independent Variables.** | To find which independent variables **statistically explain** the dependent variable. |

**PROJECT STEPS:**

**1) UNDERSTANDING THE BUSINESS PROBLEM AND INDUSTRY:-**

=> **BUSINESS PROBLEM :-**

**a)** **From Business point of view,** to target a group of customers to buy cross sell products which eventually arise the sales and becomes beneficial for both customer and as well as the firm.

**b) From Marketing point of view,**  to build strategies in order to appeal the customers to buy suggested cross-sell products as per the relevance.

=> **INDUSTRY APPROACH :-**

**a) Grounding few Business rules for financial cross-sell business problems:**

The objective of this part of the approach is to provide a framework of business

rules to guide the sales through service initiative at the bank. The inputs to this

framework were mainly experiences of other organizations (case studies of

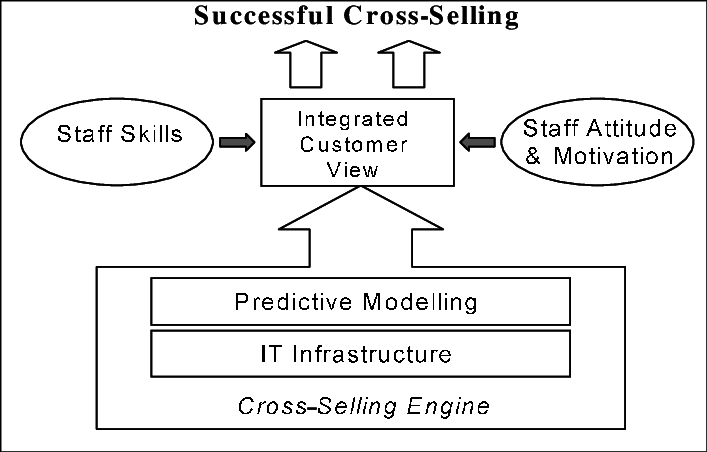
leading ﬁnancial institutions) and basic theories and research related to cross-selling in general and the ﬁnancial sector specifically. Hence, the resulting framework can be viewed as a generic best practice approach.

**b) Setting the goal of the sales** **through service system** is to understand customers so the organization can tailor any offering to them, at a profit :

* What are they buying ?
* What would they buy?
* What financial products are they buying from other companies?
* What market segments are they in ?
* What are the purchase–behavioral characteristics of that segment ?
* What is the composition of those households ?
* Where are they with respect to lifecycle need?
* How to best approach the customer (which channels, when, what to say) ?
* How to measure their reactions ?
* How to ensure satisfaction?
* What skills do the workforce need ?
* How to motivate them?
* What systems are required to support them?

**c) FOLLOW AN IT INFRASTRUCTURE:**

**Figure 2 given below**, provides an overview of the cornerstone ingredients that must be tackled successfully to develop such a system. What follows is an overview of these ingredients.



**Figure 2: A framework for a best practice IT-based cross-selling system**

**2) UNDERSTANDING THE DATA VARIABLES :-**

=> There are **32264** number of observations with **48** variables in the given dataset.

=>'**INS'- Insurance Product** is the **Dependent variable** and that is **Binary** in nature.

=> The descriptions of all the variables is given below:

|  |  |
| --- | --- |
| Data Type | Number of variables |
| Binary | **18** |
| Continuous | **28** |
| Categorical | **2** |

|  |  |  |  |
| --- | --- | --- | --- |
| S.No. | NAME | LABEL | VARIABLE TYPE |
| 1 | **AcctAge** | **Age of Oldest Account** | **Continuous** |
| 2 | **DDA** | **Checking Account** | **Binary** |
| 3 | **DDABal** | **Checking Balance** | **Continuous** |
| 4 | **CashBk** | **Number Cash Back** | **Continuous** |
| 5 | **Checks** | **Number of Checks** | **Continuous** |
| 6 | **DirDep** | **Direct Deposit** | **Binary** |
| 7 | **NSF** | **Number Insufficient Fund** | **Binary** |
| 8 | **NSFAmt** | **Amount NSF** | **Continuous** |
| 9 | **Phone** | **Number Telephone Banking** | **Continuous** |
| 10 | **Teller** | **Teller Visits** | **Continuous** |
| 11 | **Sav** | **Saving Account** | **Binary** |
| 12 | **SavBal** | **Saving Balance** | **Continuous** |
| 13 | **ATM** | **ATM** | **Binary** |
| 14 | **ATMAmt** | **ATM Withdrawal Amount** | **Continuous** |
| 15 | **POS** | **Number Point of Sale** | **Continuous** |
| 16 | **POSAmt** | **Amount Point of Sale** | **Continuous** |
| 17 | **CD** | **Certificate of Deposit** | **Binary** |
| 18 | **CDBal** | **CD Balance** | **Continuous** |
| 19 | **IRA** | **Retirement Account** | **Binary** |
| 20 | **IRABal** | **IRA Balance** | **Continuous** |
| 21 | **LOC** | **Line of Credit** | **Binary** |
| 22 | **LOCBal** | **Line of Credit Balance** | **Continuous** |
| 23 | **ILS** | **Installment Loan** | **Binary** |
| 24 | **ILSBal** | **Loan Balance** | **Continuous** |
| 25 | **MM** | **Money Market** | **Binary** |
| 26 | **MMBal** | **Money Market Balance** | **Continuous** |
| 27 | **MMCred** | **Money Market Credits** | **Continuous** |
| 28 | **MTG** | **Mortgage** | **Binary** |
| 29 | **MTGBal** | **Mortgage Balance** | **Continuous** |
| 30 | **CC** | **Credit Card** | **Binary** |
| 31 | **CCBal** | **Credit Card Balance** | **Continuous** |
| 32 | **CCPurc** | **Credit Card Purchases** | **Continuous** |
| 33 | **SDB** | **Safety Deposit Box** | **Binary** |
| 34 | **Income** | **Income** | **Continuous** |
| 35 | **HMOwn** | **Owns Home** | **Binary** |
| 36 | **LORes** | **Length of Residence** | **Continuous** |
| 37 | **HMVal** | **Home Value** | **Continuous** |
| 38 | **Age** | **Age** | **Continuous** |
| 39 | **CRScore** | **Credit Score** | **Continuous** |
| 40 | **Moved** | **Recent Address Change** | **Binary** |
| 41 | **InArea** | **Local Address** | **Binary** |
| 42 | **Ins** | **Insurance Product** | **Binary** |
| 43 | **Branch** | **Branch of Bank** | **Categorical** |
| 44 | **Res** | **Area Classification** | **Categorical** |
| 45 | **Dep** | **Checking Deposits** | **Continuous** |
| 46 | **DepAmt** | **Amount Deposited** | **Continuous** |
| 47 | **Inv** | **Investment** | **Binary** |
| 48 | **InvBal** | **Investment Balance** | **Continuous** |

**3) CLEANING THE DATA:-**

In order to clean the data special method called **WOE (Weight of Evidence) & IV** **(Information Value)** has been used.

=> **METHOD:-** IV is used to reduce the number of variables at a manageable size. The value of IV of each variable must lie between (0.1-0.5) then only it can be included in the analysis. Then WOE diagram should be drawn for the selected variables which should not have more than 2 kinks. It may have upward or downward slope which again will be used to analyze data later on.

=> **OUTPUT:-** Selected 9 variables after applying the above method:

|  |  |
| --- | --- |
| S.NO | SELECTED VARIABLE |
| 1 | **DDABAL** |
| 2 | **SAVBAL** |
| 3 | **ATMAmt** |
| 4 | **CDBal** |
| 5 | **MMBal** |
| 6 | **AGE** |
| 7 | **DEP** |
| 8 | **DEPAmt** |
| 9 | **PHONE** |

**4) TRANSFORMING TO WOE VALUES TO TREAT MISSING VALUES/ OUTLIERS:**

=> **METHOD:** After treating the missing values & outliers, all the value of variables has been transformed to WOE values by creating a new variable ex**: 'VAR\_WOE'** ..so on these new variables will be used for further analysis. (**WORKS FOR ONLY CONTINUOUS AND CATEGORICAL VARIABLES).**

=> **OUTOUT:-**

|  |  |  |
| --- | --- | --- |
| S.NO | TRANSFORMED VARIABLE | NUMBER OF MISSING VALUES |
| 1 | **DDABAL\_WOE** | **0** |
| 2 | **SAVBAL\_WOE** | **0** |
| 3 | **ATMAmt\_WOE** | **0** |
| 4 | **CDBal\_WOE** | **0** |
| 5 | **MMBal\_WOE** | **0** |
| 6 | **AGE\_WOE** | **0** |
| 7 | **DEP\_WOE** | **0** |
| 8 | **DEPAmt\_WOE** | **0** |
| 9 | **PHONE\_WOE** | **0** |

**B) FOR BINARY VARIABLES:-**

Missing values found only in 3 variables as follows :

|  |  |  |  |
| --- | --- | --- | --- |
| S.No. | VARIABLE | VARIABLE DESCRIPTION | NUMBER OF MISSING VALUES |
| 1 | DDA | Checking Account | 0 |
| 2 | DirDep | Direct Deposit | 0 |
| 3 | NSF | Number Insufficient Fund | 0 |
| 4 | Sav | Saving Account | 0 |
| 5 | ATM | ATM | 0 |
| 6 | CD | Certificate of Deposit | 0 |
| 7 | IRA | Retirement Account | 0 |
| 8 | LOC | Line of Credit | 0 |
| 9 | ILS | Installment Loan | 0 |
| 10 | MM | Money Market | 0 |
| 11 | MTG | Mortgage | 0 |
| 12 | CC | Credit Card | 4133 |
| 13 | SDB | Safety Deposit Box | 0 |
| 14 | HMOwn | Owns Home | 5533 |
| 15 | Moved | Recent Address Change | 0 |
| 16 | InArea | Local Address | 0 |
| 17 | Inv | Investment | 4133 |
| 18 | Ins | Insurance Product | 0 |

**NOTE:-** The missing values can be treated only if they represent less than 2-3% of the data. Since here it is more than 5% Hence, these 3 variables cannot be considered for modelling. Therefore, no missing value treatment is required for binary variables.

**=> DATA HAS BEEN CLEANED AND READY FOR STATISTICAL MODELLING.**

**5) DEVELOPMENT AND VALIDATION SAMPLING:-**

**=> METHOD:** The data - set is divided into 2 parts: Development sample (70%) & Validation Sample (30%). In order to Develop the model in the Development Sample, and further validate it on the Validation Sample.

=> **OUTPUT:-**

|  |  |  |
| --- | --- | --- |
| SAMPLE (PERCENTAGE) | SAMPLE DATASET NAME USED IN THE MODEL | NUMBER OF OBSERVATIONS |
| DEVELOPMENT (70%) | **DEV** | **22473** |
| VALIDATION (30%) | **VAL** | **9791** |

**6) CHECKING THE MULTI COLLINEARITY PROBLEM:-**

**=> METHOD:** If VIF <=2, Then the variable can be used for analysis.

**=> OUTPUT:**  All variables are checked and the variables with VIF<=2 are selected for further analysis and given below:

|  |  |  |
| --- | --- | --- |
| S.NO | SELECTED VARIABLE | VIF |
| 1 | **DDABAL\_WOE** | **1.05384** |
| 2 | **SAVBAL\_WOE** | **1.05226** |
| 3 | **ATMAmt\_WOE** | **1.08953** |
| 4 | **CDBal\_WOE** | **1.04683** |
| 5 | **MMBal\_WOE** | **1.00407** |
| 6 | **AGE\_WOE** | **1.04216** |
| 7 | **DEP\_WOE** | **1.16299** |
| 8 | **DEPAmt\_WOE** | **1.39913** |
| 9 | **PHONE\_WOE** | **1.16102** |
| 10 | **NSF** | **1.07622** |
| 11 | **MOVED** | **1.00237** |
| 12 | **LOC** | **1.09316** |
| 13 | **MTG** | **1.0511** |
| 14 | **ILS** | **1.03789** |
| 15 | **SDB** | **1.03471** |
| 16 | **IRA** | **1.037** |
| 17 | **INAREA** | **1.10152** |

**NOTE:- ALL VIF<=2, Hence, NO MULTICOLLINEARITY.**

**7) MODEL DEVELOPMENT- BUILD MODEL-**

=> **Logistic Regression** is used to build the model.

=> Significant Variables having **p-value < alpha** are chosen and given as follows:

|  |  |  |
| --- | --- | --- |
| SELECTED VARIABLE | ESTIMATE | Pr>ChiSq |
| DDABAL\_WOE | **5.9484** | **<0.0001** |
| SAVBAL\_WOE | **6.6345** | **<0.0001** |
| ATMAmt\_WOE | **2.9407** | **<0.0001** |
| CDBal\_WOE | **0.7479** | **<0.0001** |
| MMBal\_WOE | **0.9356** | **<0.0001** |
| DEP\_WOE | **2.2344** | **<0.0001** |
| PHONE\_WOE | **1.0835** | **<0.0001** |
| DIRDEP | **-0.2354** | **<0.0001** |
| IRA | **0.4928** | **<0.0001** |
| INAREA | **-0.2882** | **0.0002** |

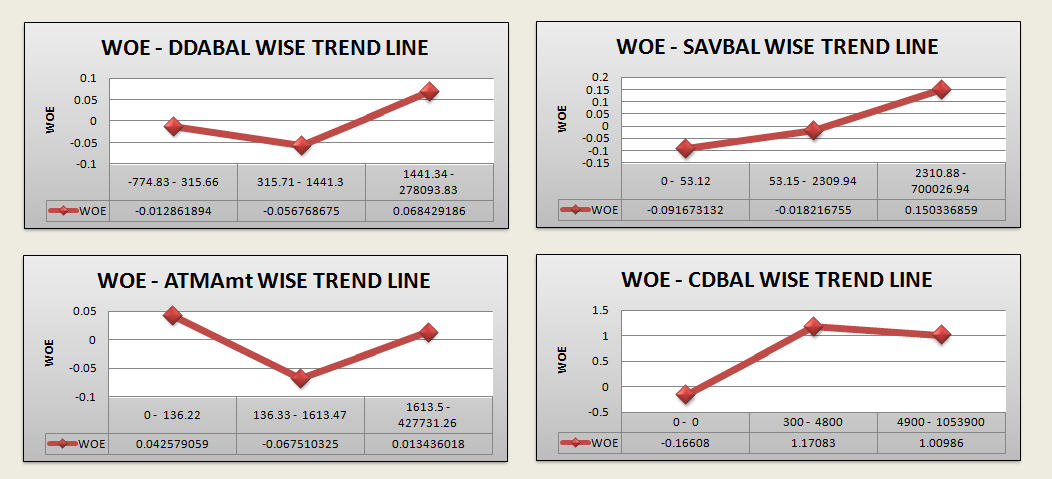
**\* WHY P-VALUE< APLHA (0.1 or 0.05) ?**

**=>** Chi-Square value for each explanatory variable- the Chi-Square value indicates that the level of significance, i.e- the impact of independent(explanatory) variable on the dependent variable.

**The p-value cut-off should be decided in discussion with the business. Ideally the p-value<0.0001. However in case of smaller population size p-value could be <0.05 or p-value<0.1.**

**\* INTERPRETATION OF VARIABLES IN TERMS OF WOE DIAGRAMS:**

=>

****

**A) DDABal -**

* Customers who have **checking balance** between **1441.34-278093.83** are **most likely to buy Insurance products.**
* Customers who have **checking balance** between **315.71-1441.3** are **more likely to buy Insurance products.**
* Customers who have **checking balance** between **-774.83-315.66** are **least likely among all to buy Insurance products.**

**B) SAVBal -**

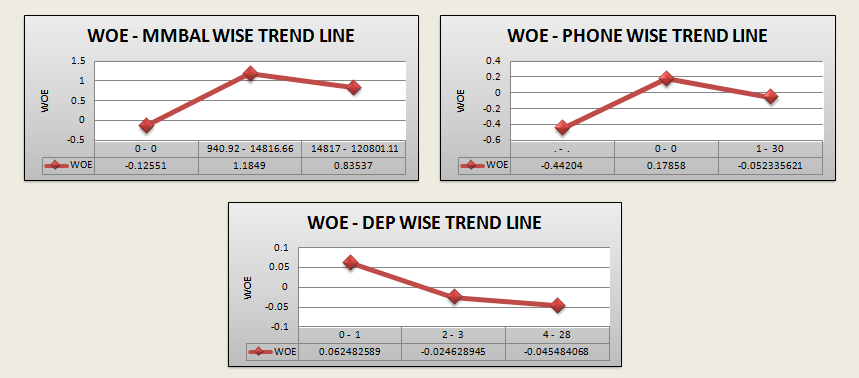
* Customers who have **saving balance** between **2310.88 - 700026.94** are **most likely to buy Insurance products.**
* Customers who have **saving balance** between **53.15 - 2309.94** are **more likely to buy Insurance products.**
* Customers who have **saving balance** between **0 - 53.12** are **least likely among all to buy Insurance products.**

**C) ATMAmt -**

* Customers whose **ATM withdrawal amount** lies between **1613.5 - 427731.26** are **most likely to buy Insurance products.**
* Customers whose **ATM withdrawal amount** lies between **136.33 - 1613.47** are **more likely to buy Insurance products.**
* Customers whose **ATM withdrawal amount** lies between **0 - 136.22** are **least likely among all to buy Insurance products.**

**D) CDBal -**

* Customers who have **CD(certificate of Deposit) balance** between **4900 - 1053900** are **most likely to buy Insurance products.**
* Customers who have **CD(certificate of Deposit) balance** between **300 - 4800** are **more likely to buy Insurance products.**
* Customers who have **ZERO** **(0) CD(certificate of Deposit) balance** are **unlikely among all to buy Insurance products.**

****

**E) MMBal -**

* Customers who have **MM(Money Market) balance** between **14817 - 120801.11** are **most likely to buy Insurance products.**
* Customers who have **MM(Money Market) balance** between **940.92 - 14186.66** are **more likely to buy Insurance products.**
* Customers who have **ZERO** **(0) MM(Money Market) balance** are **unlikely among all to buy Insurance products.**

**F) DEP -**

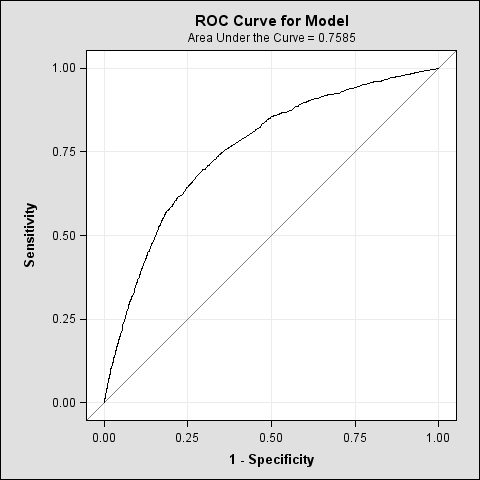
* Customers who have **Checking deposits** **0 & 1** are **most likely to buy Insurance products.**
* Customers who have **Checking deposits 2 & 3** are **more likely to buy Insurance products.**
* Customers who have **Checking deposits between 4-28** are  **least likely among all to buy Insurance products.**

**G) PHONE -**

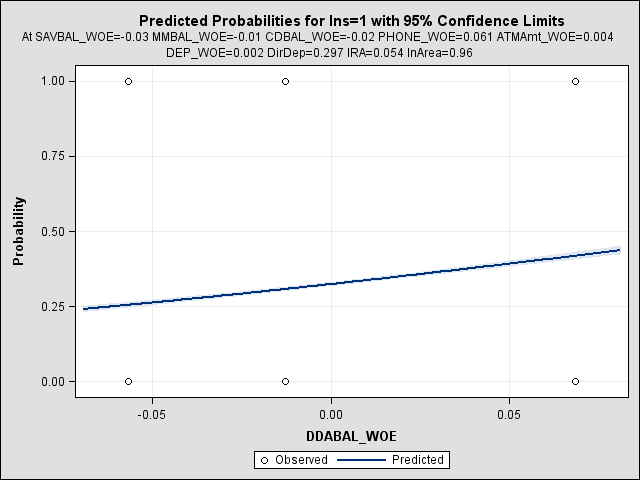
* Customers who have **ZERO** as **Telephone number for Banking** are **most likely to buy Insurance products.**
* Customers who have **Telephone number for Banking** between **1-30** are **more likely to buy Insurance products.**
* Customers whose **Telephone number is Missing for Banking** are **least likely among all to buy Insurance products.**

**8) MODEL DIAGNOSTICS:-**

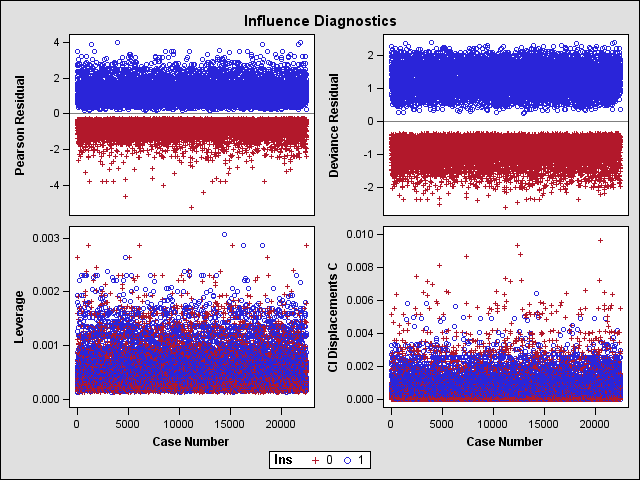
**- ROC CURVE:**

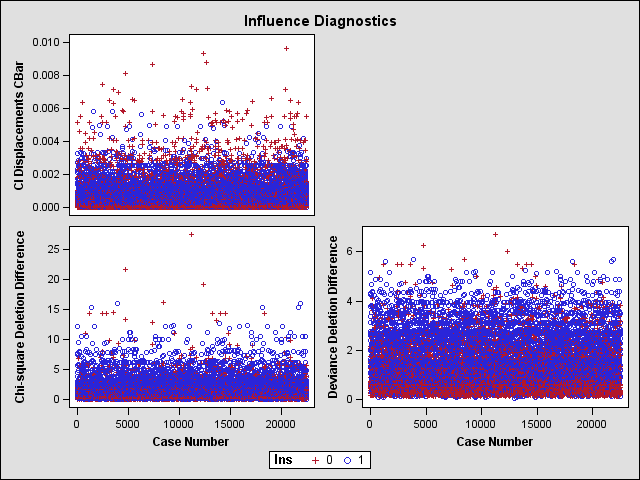
****

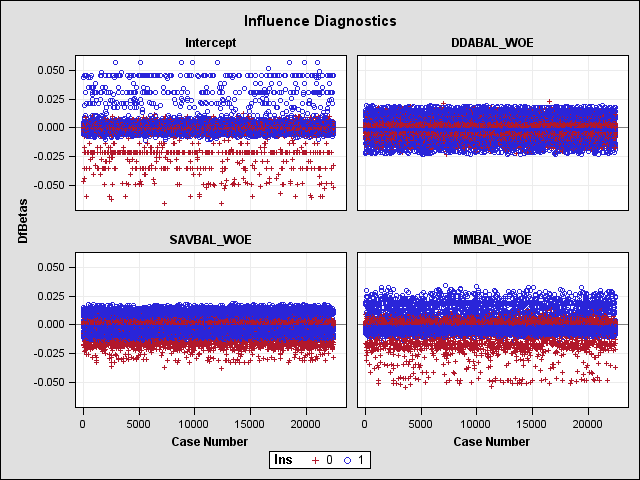
**-EFFECT PLOT:**

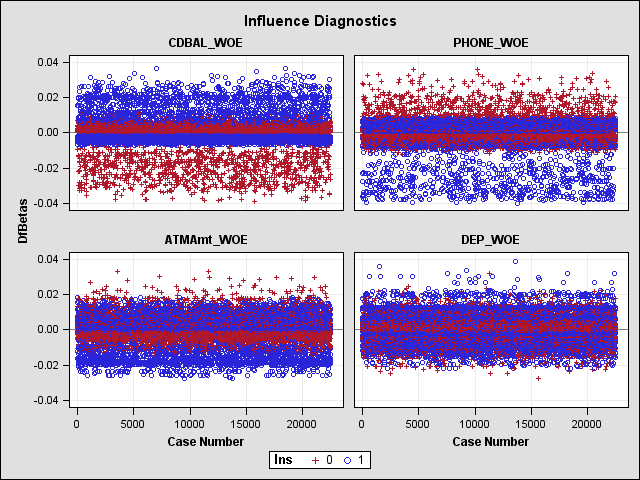
****

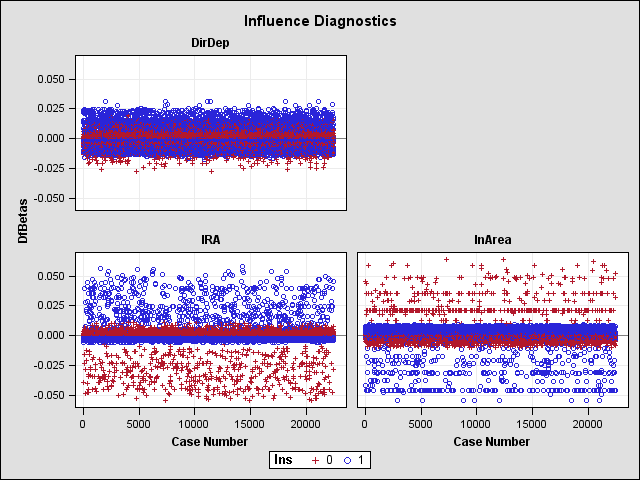
**-INFLUENCE PLOTS:**

****

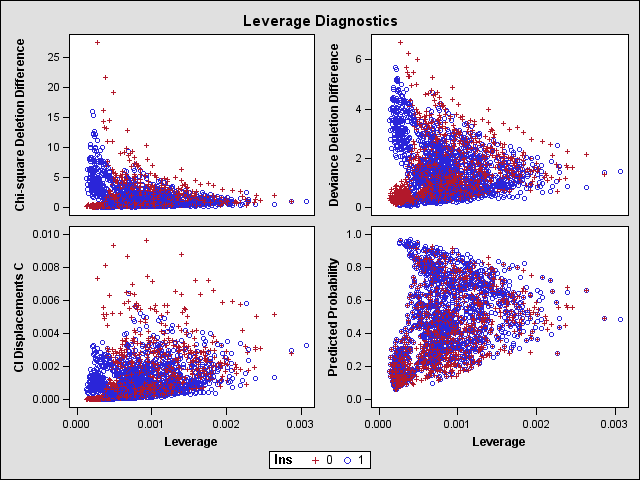
****

****

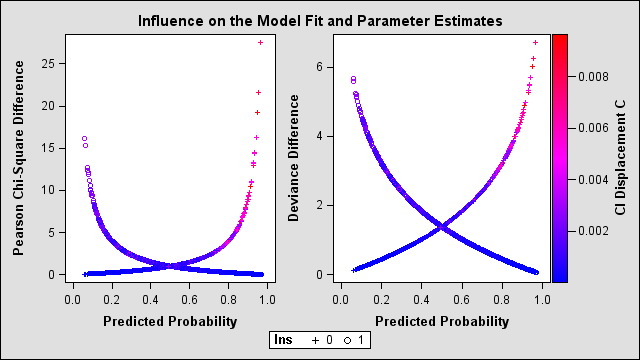
****

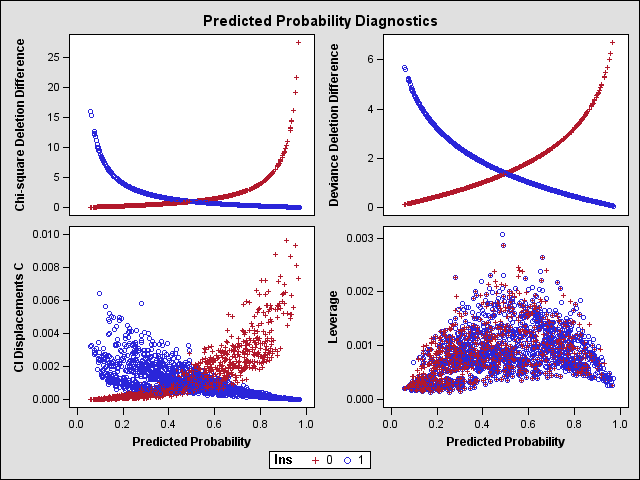
****

**- LIVERAGE PLOTS:**

****

**- PREDICTED PROBABILITY DIGNOSTICS:**

****

****

**9) TESTING THE MODEL:-**

|  |  |  |  |
| --- | --- | --- | --- |
| TEST NAME | P-VALUE/ STATISTIC MEASURE | | INTERPRETATION |
| DEV | VAL |
| CONCORDANCE | **75.4** | **75.1** | **=> The concordance is higher in both samples , the separation of scores between good & bad accounts is larger.** |
| GINI COEFFICIENTS | **0.2567** | **0.253** | **=> Represents the area covered under Lorenz curve which lies between 0.2-0.35 in both the sample. Hence, model is good.** |
| KS STATISTICS | **39.7** | **40.2** | **=>The absolute differences between cumulative % of goods & cumulative % of bads is greater than 20.** |
| RANK ORDERING | **All Satisfactory** | **All Satisfactory** | **=> The model is able to differentiate the Goods from the Bads across the population breakup is validated 'Satisfactory'.** |

**NOTE:- The model for 'Development' sample has been re-ran over the 'Validation' sample and conclusions are as follows :**

* **Chi-sq values and level of significances and p-value for each explanatory variables are satisfied as 'development' sample .**
* **The p-values are not changed from the development sample to the validation sample.**
* **The sign of parameter Estimates are not changed.**
* **Rank order is SATISFACTORY for both the samples as mentioned above.**

**INTERPRETATIONS OF SIGNIFICANT VARIABLES:-**

|  |  |  |  |
| --- | --- | --- | --- |
| SELECTED VARIABLE | SIGN OF EFFECT | ESTIMATE | INTERPRETATION |
| DDABAL\_WOE | **+VE** | **5.9484** | **=> If the customer increases his Checking balance by one unit, he is 5.9 times more likely to buy insurance product.** |
| SAVBAL\_WOE | **+VE** | **6.6345** | **=> If the customer increases his Saving balance by one unit , he is 6.63 more times likely to buy the insurance product.** |
| ATMAmt\_WOE | **+VE** | **2.9407** | **=> If the customer increases his ATM withdrawal amount by one unit, he is 2.94 more times likely to buy the insurance product.** |
| CDBal\_WOE | **+VE** | **0.7479** | **=> If the customer increases his certified Deposit balance by one unit, he is 0.74 more times likely to buy the insurance product.** |
| MMBal\_WOE | **+VE** | **0.9356** | **=> If the customer increases his Money Market balance by one unit, he is 0.93 more times likely to buy the insurance product.** |
| DEP\_WOE | **+VE** | **2.2344** | **=> If the customer increases his deposit balance by one unit, he is 2.23 more times likely to buy the insurance product.** |
| PHONE\_WOE | **+VE** | **1.0835** | **=> If the customer changes telephone number for banking, he is 1.08 more times likely to buy the insurance product.** |
| DIRDEP | **-VE** | **-0.2354** | **=> If the customer adopts the service of Direct Deposit, he is 0.23 times less likely to buy the insurance product and vice versa.** |
| IRA | **+VE** | **0.4928** | **=> If the customer has Retirement Account, he is 0.49 more times likely to buy the insurance product.** |
| INAREA | **-VE** | **-0.2882** | **=> If the customer has given Local address, he is 0.28 times less likely to buy the insurance product and vice versa.** |

**MODEL COMPARISONS:-**

**MODELS WHICH ARE HIGHLY PREFERRED IN THE INDUSRTY AND COULD HAVE BEEN INCLUDED FOR CROSS-SELL DATASET :**

* **Factor Analysis =>** cross-selling can attain benefits at 3 important levels such as Insurance, Retail Banking , Telecom, which is further achieved at factors like customer Trust, Life Cycle information etc.
* **Cluster Analysis =>** Customer segmentation could be done for making effective Marketing strategies in order to offer credit cards and loans by using cluster analysis.

=> Hence, other analytical models can also be used here in order to extract insightful information as per asked.

**BUSINESS INSIGHTS:-**

* To develop market strategies targeting the customers who have higher:

**\* ATM Withdrawal amount.**

**\* Money Market Balance.**

**\* Saving Balance.**

**\* Checking Balance.**

* Customers with **'Retirement Accounts'** can also be targeted.
* Strategies should be appealing to a customer in order to adopt services like Direct Deposits and Telephone number for Banking.
* Great deals and offers can drive customers to buy Insurance products.
* Services like Internet or Phone Banking can also be proved beneficial for buying Insurance products.
* Relevant and discounted suggested products can be appealing for buying to a customer.
* Customers with **Certified Deposit Balance** can be targeted for secured payments and purchases.
* Cross -sell products should be chosen/suggested for a segment of customers as per their history of purchases.
* It can be targeted to the customers who use ATM cards and have active and secured transactional logs.
* Cross-sell Products should be available within the given local address of a customer in order to make his experience more better and satisfactory.
* E-commerce could be helpful to reaching out to the targeted customers with more relevant products at affordable prices.
* Business Strategy in to order to increase the revenue is possible by providing various Payment Modes such as:

\* E-wallets.

\* Cash on Delivery.

\* Debit Card/ Credit Card.

**-----------------END OF THE REPORT---------------**