**PROJECT REPORT ON CROSS SELL INSURANCE**

**BY: RINNY KAPOOR**

**INTRODUCTION**

Cross-selling is defined as "the action or practice of selling among or between established clients, markets, traders, etc." or "that of selling an additional product or service to an existing customer “It is a sales technique in which the salesperson recognizes what a customer is purchasing and will make suggestions or recommendations of other related merchandise the shopper may also be interested in purchasing. It is also known as suggestive selling. It also stands for being able to offer to the existing bank customers, some additional banking products, with a view to expand banking business, reduce the per customer cost of operations and provide more satisfaction and value to the customer. An existing customer provides an advantage to the seller over the competitor. The satisfied customer will always consider the same seller for other requirements before searching for other sellers. The seller can make use of this situation and the customer will be pleased as well.

**PROBLEM DEFINITION AND METHODOLOGY**

The project objective was to find out which customers are more likely to buy cross sell products (Permanent Disability Insurance with a Low Premium).The dataset consists of 33285 observations and 48 variables. A financial institution looking to cross sells their insurance product to existing customers. I have customer’s personal information with their assets, liabilities and transactions. For that I have used Logistic Regression technique.

**Steps Required for Modelling:**

**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.

2. **VARIABLES DESCRIPTION**

* Dataset used for modelling consists of 48 variables.
* INS=Insurance Product which is the dependent variable and it is binary in nature.
* Description of dataset is mentioned below:

1. Continuous Variables: 28

2. Binary Variables: 18

3. Categorical Variable: 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:** It means to check if our dataset contains missing values or any outlier. As there are missing values in our dataset as well as some outlier’s. Since in our dataset variables used are of continuous, categorical and binary type. So to remove missing values and to treat outliers for categorical and continuous variables I used fine classing and coarse classing technique. In this model I used SAS Macro to create Information Value (IV) and Weight of Evidence (WOE) Values for each variable in dataset. So to understand this we must be familiar with these terms which I am defining below:

****

**3. FINE CLASSING:** Fine classing is a variable reduction technique for both types of variables i.e. Continuous and Categorical.

Method: We check the IV value for each variable. The value of IV for each variable must be between 0.1 and 0.5 then only we will include these variables for analysis. So here what I did is I selected the variables whose IV Values is between 0.1-0.5.

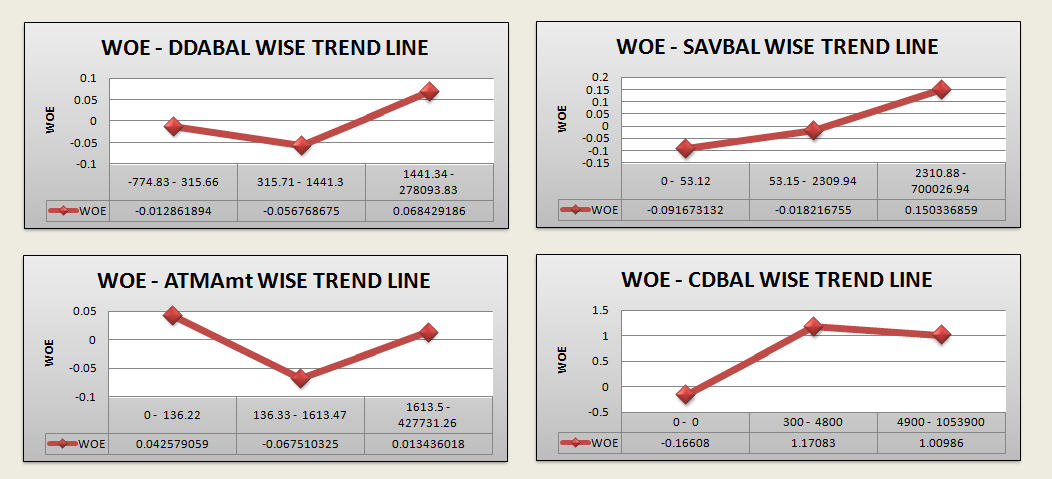
After doing fine classing I got 9 variables which are DDABAL, SAVBAL, ATMAMT, CDBAL, MMBAL, AGE, DEP, DEPAMT and PHONE.

**4. COARSE CLASSING:** It is the next step after fine classing. It is done on selected variables from fine classing process. After that we have to transform all the value of variables to WOE values by creating a new variable ex**: 'VAR\_WOE**.We will use these new variables for further analysis and it works for categorical and continuous variables only.

Method: Here we check for the kinks in the line graph of variables and line graph is plotted on WOE Values. We have to draw WOE diagram for selected variables and we should get diagram with less than 2 kinks and these kinks are in upward and downward direction.

The variables I got after transforming the data are DDABAL\_WOE, SAVBAL\_WOE, ATMAMT\_WOE, CDBAL\_WOE, MMBAL\_WOE, AGE\_WOE, DEP\_WOE, DEPAMT\_WOE and PHONE\_WOE.

**INTERPRETATIONS OF VARIABLES BASED ON WOE\_DIGRAMS**

****

**1.) DDABAL**

* Customers who have **checking balance** between **1441.34 and 278093.83** are **most likely to buy Insurance products because bank will look for the customers who have maintained enough balance in their account so that they can afford the premium for insurance.**
* Customers who have **checking balance** between **315.71 and 1441.3** are **more likely to buy Insurance products.**
* Customers who have **checking balance** between **-774.83and 315.66** are **least likely among all to buy Insurance products.**

**2.) SAVBAL**

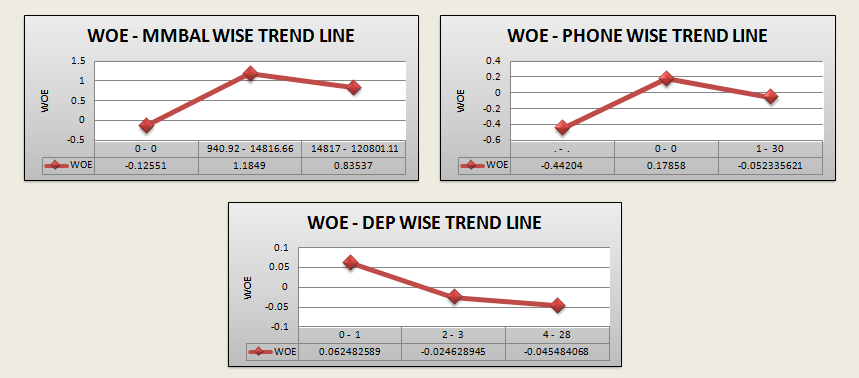
* Customers who have **saving balance** between **2310.88 and 700026.94** are **most likely to buy Insurance products because bank will look for the customers who have maintained enough balance in their account so that they can afford the premium for insurance.**
* Customers who have **saving balance** between **53.15 and 2309.94** are **more likely to buy Insurance products.**
* Customers who have **saving balance** between **0 and 53.12** are **least likely among all to buy Insurance products.**

**3.) ATMAMT**

* Customers whose **ATM withdrawal amount** lies between **1613.5 and 427731.26** are **most likely to buy Insurance products because the customers who have made more transactions will have more chances for bank to cross sell their product because of promotions givens at the time of transaction.**
* Customers whose **ATM withdrawal amount** lies between **136.33 and 1613.47** are **more likely to buy Insurance products.**
* Customers whose **ATM withdrawal amount** lies between **0 and 136.22** are **least likely among all to buy Insurance products.**

**4.) CDBAL**

* Customers who have **CD (certificate of Deposit) balance** between **4900 and 1053900** are **most likely to buy Insurance products because bank will look for the customers who have maintained balance in their account so that they can afford the premium for insurance.**
* Customers who have **CD (certificate of Deposit) balance** between **300 and 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.**

****

**5.) MMBAL**

* Customers who have **MM (Money Market) balance** between **14817 and 120801.11** are **most likely to buy Insurance products because bank will look for the customers who have maintained balance in their account so that they can afford the premium for insurance.**
* Customers who have **MM (Money Market) balance** between **940.92 and 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.**

**6.) DEP**

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

**7.) 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 and 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.**

Next step is to check missing values for binary variables which are as follows DIRDEP,NSF,DDA,SAV,ATM,CD,MM,MOVED,LOC,SDB,IRA,MTG,ILS,INAREA,CC,INS,HMOWN,INV.

|  |  |  |  |
| --- | --- | --- | --- |
| 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 | IINS | Insurance Product | 0 |

So we found that there are 3 binary variables which have missing values that are CC=4133, HMOWN=5133 and INV=4133.

**SINCE, THE MISSING VALUES CAN BE TREATED ONLY IF THEY REPRESENT LESS THAN 2-3% OF THE DATA.**

**HERE, IT IS MORE THAN 10%.**

**HENCE, THESE 3 VARIABLES CANNOT BE CONSIDERED FOR MODELING=> NO MISSING VALUE TREATMENT IS REQUIRED. \*/**

**5. DIVIDING THE DATASET INTO DEVELOPMENT AND VALIDATION:** Here I divided the dataset into two parts i.e. we use 70% of part as Development and 30% of data in Validation part. I have to develop the model in development sample and validate the model in validation sample.

|  |  |  |
| --- | --- | --- |
| 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 is 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** |

**#Since VIF<=2 so there is no multicollinearity problem and we can build the model.**

**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** |

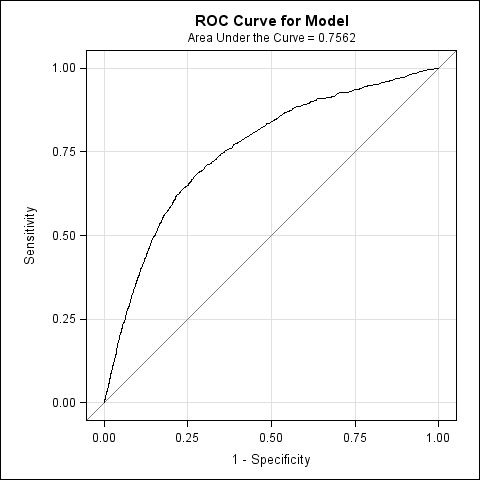
**\* WHY P-VALUE< APLHA (0.1 or 0.05 or 0.001)?**

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.

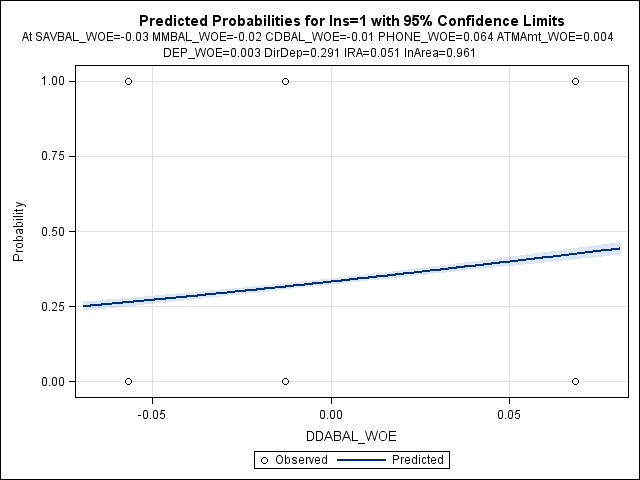
**The p-value cut-off should be decided in discussion with the business. Ideally the p-value<0.0001. However in case of smaller population size p-value could be <0.05 or p-value<0.1 or p-value<0.001.**

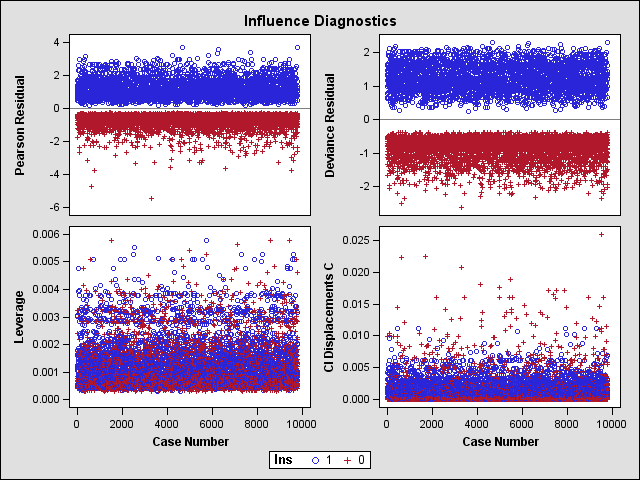
**8. MODEL DIAGNOSTICS:**

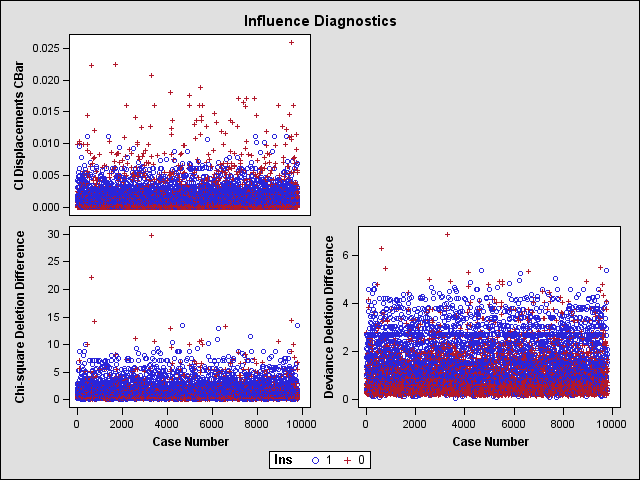
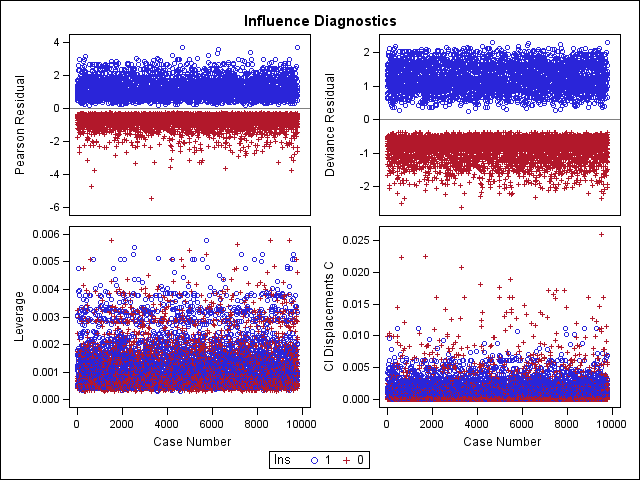
**1. ROC CURVE**

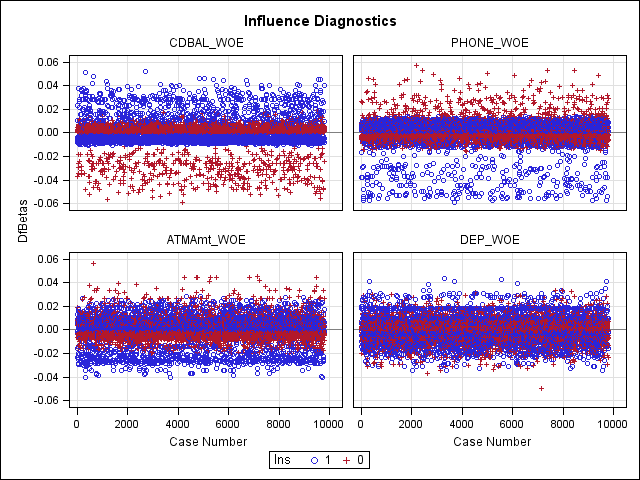
****

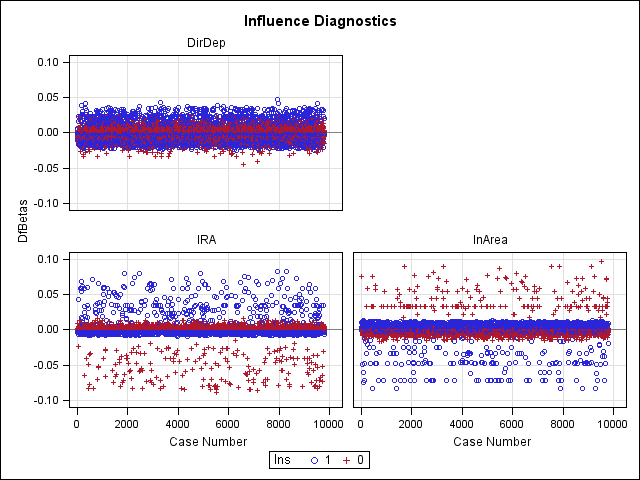
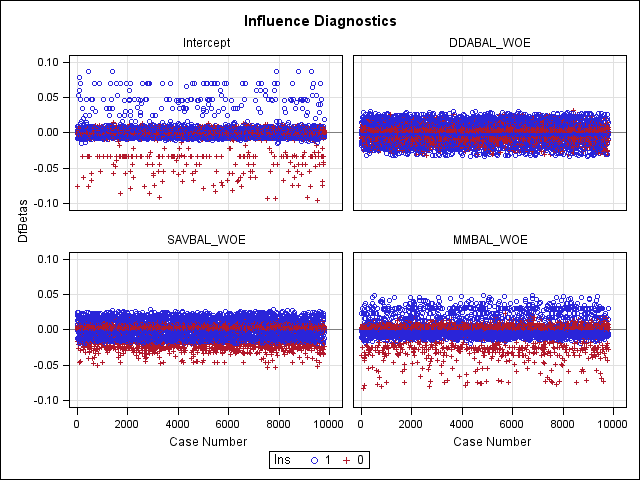
**2.EFFECT PLOT**

****

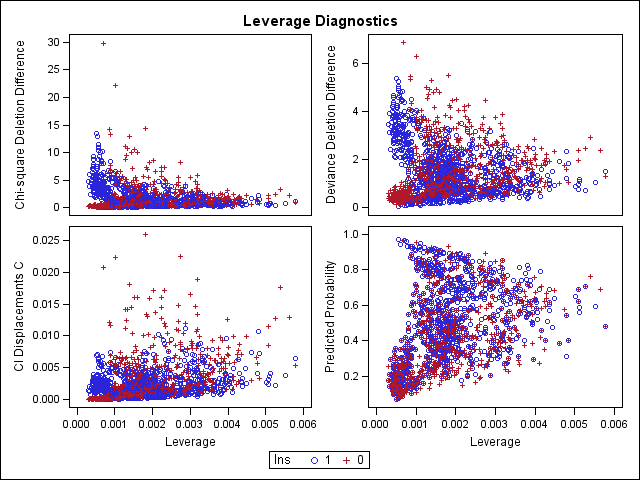
3. INFLUENCE PLOTS



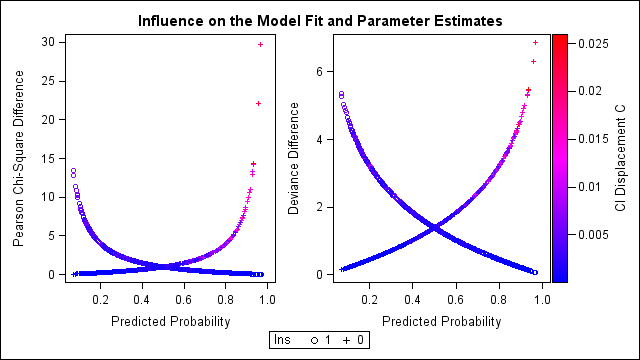


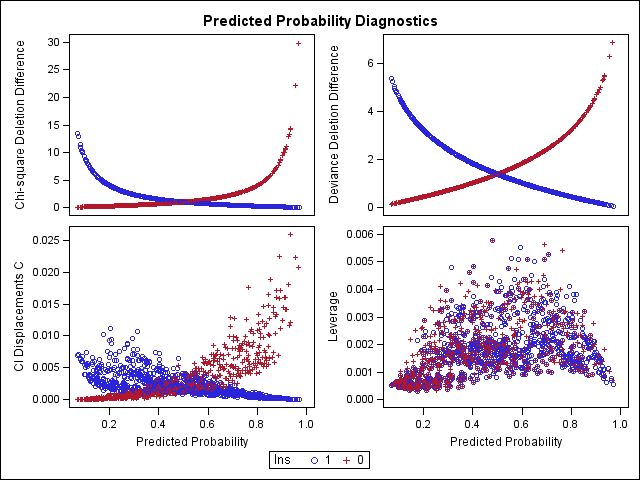


**5. LIVERAGE PLOTS**



**6. 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 variable 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 because bank will look for the customers who have maintained balance in their account so that they can afford the premium for insurance.** |
| 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 because bank will look for the customers who have maintained balance in their account so that they can afford the premium for insurance.** |
| 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 because bank will look for the customers who have maintained balance in their account so that they can afford the premium for insurance.** |
| 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 because bank will look for the customers who have maintained balance in their account so that they can afford the premium for insurance.** |
| 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 opts for telephone banking, he is 1.08 more times likely to buy the insurance product because executive of concerned institute has chance to push the product to a particular customer.** |
| 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 because in this case customer makes payment without the intervention of bank so it’s become difficult for bank to reconcile the transactions because reconciliation is done either quarterly or on closing.** |
| IRA | **+VE** | **0.4928** | **=> If the customer has Retirement Account, he is 0.49 more times likely to buy the insurance product because customer has enough funds so bank can cross sell insurance.** |
| INAREA | **-VE** | **-0.2882** | **=> If the customer has given Local address, he is 0.28 times less likely to buy the insurance product because a bank can lose his customer as customer will transfer his account to location which is near to customer.** |

**BUSINESS INSIGHTS:**

* To develop market strategies Bank should target the customers who have higher
* **ATM Withdrawal amount.**
* **Money Market Balance.**
* **Saving Balance.**
* **Checking Balance.**
* There are very less customers who are adopting the services of Direct Deposits and Phone Banking so Strategies should be appealing to a customer in such a way they adopt services like Direct Deposits and Phone Banking.
* Customers who have retirement account can be targeted to cross sell because as they have lump sum retirement payment and loss of medical protection from company so bank can cross sell its insurance product to that customer which is useful for bank as well as for customer.
* Customers with **Certified Deposit Balance** can be targeted for secured payments and purchases.
* If a customer is changing his local address or moving to a bigger place then a bank should upgrade his existing product and cross sell wealth management products.
* Cross-sell products should be offered to a segment of customers as per their history of purchases as it will be easier for bank to prioritize those customers who are in need of that product.
* 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 Report**-----------------------------