|  |
| --- |
|  |
| NBO Framework Development –Savings Account |
| Model Development Document |
|  |
| **Sneha Jha** |
| **4/8/2016** |

|  |
| --- |
|  |

[1. INTRODUCTION 3](#_Toc447707355)

[2. MODEL METHODOLOGY 4](#_Toc447707356)

[Modeling Approach 4](#_Toc447707357)

[Association Rules 5](#_Toc447707358)

[Recommendation Framework 6](#_Toc447707359)

[Target Variable Definition 6](#_Toc447707360)

[Choice of Products 6](#_Toc447707361)

[Covariates Considered 6](#_Toc447707362)

[Validation Approach 7](#_Toc447707363)

[3. DATA PULL/EXPLORATION 7](#_Toc447707364)

[Data Lineage and Quality 7](#_Toc447707365)

[Data Extraction, Match/Merge 8](#_Toc447707366)

[Target Variable Creation 8](#_Toc447707367)

[Sampling 9](#_Toc447707368)

[Univariate Analysis 9](#_Toc447707369)

[Missing, Caps and Floors 9](#_Toc447707370)

[4. MODEL DEVELOPMENT 10](#_Toc447707371)

[Segmentation 10](#_Toc447707372)

[Variable Transformations 11](#_Toc447707373)

[Variable Selection/Reduction 11](#_Toc447707374)

[Model Estimation 12](#_Toc447707375)

[Selection Criteria 12](#_Toc447707376)

[5. RECOMMENDATION FRAMEWORK 13](#_Toc447707377)

[Validation 13](#_Toc447707378)

[6. IMPLEMENTATION/SCORING 13](#_Toc447707379)

[Spec Sheet 13](#_Toc447707380)

[Data Pull for Scoring 14](#_Toc447707381)

[Scoring Code 14](#_Toc447707382)

[Missings, Caps/Floors & Transformations 14](#_Toc447707383)

[Implementation Environment 14](#_Toc447707384)

[7. Ongoing Model Risk Management 14](#_Toc447707385)

[Summary of Model Risks and Limitations 14](#_Toc447707386)

[8. References 16](#_Toc447707387)

# INTRODUCTION

IndusInd Banking Limited possesses an array of products that can be offered to customers based on their needs. Each product is unique and therefore should be offered effectively to the customers by studying their behavior and characteristics. The objective of this project is to develop a strategy to identify the products that can be offered to the customers for which the response will be high which will result in higher revenue per customer.

The development of recommendation framework primarily depends on two key factors - propensity of the customer to accept the product and revenue generated if the customer accepts the product. These factors are triangulated with the customer to generate a rank order of the products for recommendation for each customer.

The source of data is both internal and external; the number of events has been augmented using data from CIBIL and Equifax. Customers who don’t have product with IndusInd but with any other financial organization which is captured in Bureau have been considered as an event. Therefore the characteristics of the customers which are events as per external data source might not truly define the characteristics of the customers who are holding a product internally. It is assumed that all the events are exhibiting characteristics that will define before accepting a product.

The products that are offered to the customers have different periods of maturity. There are many products that are still growing and in future the event rate will increase which may become significantly different from the current event rate. The model performance might vary once the event rate changes significantly in future and might need to redevelop the models.

# MODEL METHODOLOGY

## Modeling Approach

Propensity of the customers to accept a product plays an important role in the recommendation framework. Therefore, propensity models are developed for each product based on historical data. The objective of the model is to calculate the probability of a customer to accept the product. Since the dependent variable is binary, regression method with logit link is used to develop the models.

**Theory about logistic regression**

In [statistics](https://en.wikipedia.org/wiki/Statistics), logistic regression, or logit regression, or logit model is a [regression](https://en.wikipedia.org/wiki/Regression_analysis) model where the [dependent variable (DV)](https://en.wikipedia.org/wiki/Dependent_and_independent_variables) is [categorical](https://en.wikipedia.org/wiki/Categorical_variable). Logistic regression is an estimation of Logit function. Logit function is simply a log of odds in favor of the event. This function creates an s-shaped curve. Binomial or binary logistic regression deals with situations in which the observed outcome for a [dependent variable](https://en.wikipedia.org/wiki/Dependent_variable) can have only two possible types. Logistic regression is used to predict the [odds](https://en.wikipedia.org/wiki/Odds) of being a case based on the values of the [independent variables](https://en.wikipedia.org/wiki/Independent_variable) (predictors). The odds are defined as the probability that a particular outcome is a case divided by the probability that it is a noncase.

Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). Logistic regression can be seen as a special case of generalized linear model. Logistic regression generated coefficients calculate the probability (p) of an event.





**Odds ratio**

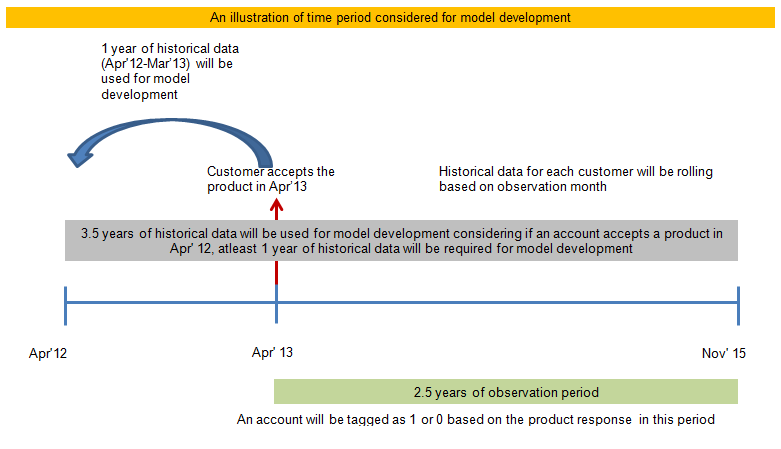
The Odds Ratio represents the odds that an outcome will occur given a particular exposure, compared to the odds of the outcome occurring in the absence of that exposure.

Everything starts with the concept of probability.  Let's say that the probability of success of some event is .8.  Then the probability of failure is 1- .8 = .2.  The odds of success are defined as the ratio of the probability of success over the probability of failure.  In our example, the odds of success are .8/.2 = 4.  That is to say that the odds of success are 4 to 1.  If the probability of success is .5, i.e., 50-50 percent chance, then the odds of success is 1 to 1.

The transformation from probability to odds is a monotonic transformation, meaning the odds increase as the probability increases or vice versa.  Probability ranges from 0 and 1.  Odds range from 0 and positive infinity.

The transformation from odds to log of odds is the log transformation.  Again this is a monotonic transformation.  That is to say, the greater the odds, the greater the log of odds and vice versa.

In this Project, depending on the number of events captured – time stamp model or look-alike model is developed for each product. The time stamp model captures the time varying attributes of the customer at the time of acceptance of the product. For the customers who do not hold the product, the latest 12 months of data is used. Below figure provides an illustration of the time stamp model.



However, for the look-alike model, all attributes are considered from the latest 12 months of data for all customers. In this approach, model is built based on the attributes of the customer who holds/ does not hold the product.

## Association Rules

Association rules was also considered for a smaller segment that did not have enough transaction information. It provides recommendation based only on the product rules and does not consider customer attributes. In this methodology, rules are formed using the current product holdings and based on the support, confidence and lift of the rules, a single product is recommended for each customer. However, for this method, it is required to have customers having atleast two product holdings to create rules. For the considered segment, only 27% of customers had more than two products so the results were not conclusive. Therefore, logistic regression models were developed for this segment as well.

## Recommendation Framework

The objective of the project is to identify the products that can be recommended to the customers based on their likelihood to accept the product and revenue of the product. Product recommendation is done based on the distance approach where each product is a coordinate of relative odds and revenue on a two-dimensional plane and distance is calculated from origin. Product having maximum distance would be most recommended. However, for cases where the revenue is very high and relative odds is very small and farthest from origin, the distance is being driven by high revenue and this will be misleading. Since offering a product to a customer having very low odds of accepting the product even though the product has very high revenue will probably lead to a decline in the offer. Hence for such cases, where the distance is maximum but the relative odds is not, product having higher relative odds and second best distance is considered.

## Target Variable Definition

The target variable for the model is binary – whether the customer has accepted a product or not. For the time stamp model, the condition is restricted to whether the customer has accepted a product in the observation period. For look-alike model, the target variable is if the customer holds the product ever or not.

## Choice of Products

Products have been selected for model development based on the event rate and volume of events such that when the data is divided into train and test, there is enough number of events. Each segment has varying product reach.



## Covariates Considered

Data was provided from multiple sources and the variables were related to bank holding information, transaction, demographics, product holdings and bureau. Bureau data was provided from two sources – CIBIL and Equifax at customer level. There were three kinds of information present in the bureau files – demographic, product enquiry related and loan related information. Both raw and derived variables were used for model development. Attached is the list of variables that have been used for model development.



## Validation Approach

Models were validated on the hold out sample. Modeling data was divided randomly into 70:30 for training and testing. Validation of the model was done on the test data.

# DATA PULL/EXPLORATION

## Data Lineage and Quality

Data was provided by IndusInd in csv and SAS format. The data was from multiple sources and related to bank holding, transaction, product holding and bureau related information. Listed below are the multiple files provided:

|  |  |  |
| --- | --- | --- |
| Type of Data | Data File | File Name |
| Base Data | SA Base File | NBO\_Sample\_SA\_CIF\_Acc\_Mask\_Final\_08Dec15.csv |
| Demographics Data | CIBIL Demographics | CIBIL\_Demographics\_v3.csv |
| Equifax Demographics File | EQUIFAX\_DEMOGRAPHICS.csv |
| IndusInd Product Data | Salutation File | CASA\_Salutation\_MaskID.csv |
| Life Insurance File | LI.csv |
| Fixed Deposit File | FD.csv |
| Personal Loan File | PL\_final.csv |
| Business Loan File | BL\_data\_31Oct\_CIF\_MaskID.csv |
| Accident Insurance File | Accident.csv |
| Card Protection Plan File | CPP\_Data.csv |
| Gold Loan File | GL.csv |
| General Insurance File | GI.csv |
| Home Loan File | HL.csv |
| Health Insurance File | Health.csv |
| Loan Against Property File | LAP.csv |
| Mutual Funds File | MF.csv |
| Brokerage File | BROK.csv |
| Transaction Scrub Data | Health Insurance Scrub File | wealthHI.csv |
| Home Loan Scrub File | wealthHL.csv |
| Life Insurance Scrub File | wealthLI.csv |
| Mutual Funds Scrub File | wealthMF.csv |
| Bureau Product Data | Cibil Loan File | CIBIL\_Loans\_v3.csv |
| Equifax Loan File | EQUIFAX\_LOAN\_INFO.csv |
| Credit Card Data | Credit Card CIF Mask Mapping | NBO\_Sample\_SA\_ENTCIF\_Mask |
| Credit Card File | Clm\_static\_sample\_cognylytics (SAS Data Set ) |
| CA Base Data | Current Account Data | NBO\_Sample\_CA\_Acc\_Mask\_Final\_08Dec15.csv |
| Transaction Data | Transaction Data (Oct- Dec'14) | transdt.trans\_oct14\_dec14\_CIFAC\_Mask\_Fin\_20Nov15.csv |
| Transaction Data (Jan- Mar'15) | trans\_jan15\_mar15\_CIFAC\_Mask\_Fin\_20Nov15.csv |
| Transaction Data (Apr - Sep'15) | trans\_apr15\_sep15\_CIFAC\_Mask\_Fin\_20Nov15.csv |
| Bureau Enquiry Data | CIBIL Enquiries File | CIBIL\_Enquiries\_v3.csv |



## Data Extraction, Match/Merge

Data was provided by IndusInd in the form of csv and SAS datasets. The base data was 25% random sample of the current SA portfolio. The customer ID (CIF) was masked and was the merging key for all data files.

Multiple exclusions were done on the base data.

* All records where the status of the account was ‘live’ were retained.
* Also, customers with account opening date in Oct 2015 or later were excluded from the base data. This was done because every customer should have at least 1 month of transaction information.
* Later, during data preparation for model development (for various products), exclusion was done based on months on books. Customers having MOB<4 were excluded since they did not have enough transaction information.

Transaction data was available from Jan 2013 to Sep 2015. In case of time stamp model, the observation period was considered from Jan 2014 to Sep 2015 since 12 months of information was used for historical data considered for the model development. For look-alike models, the latest 12 months data was considered for model development i.e. Oct 2014 to Sep 2015.

The demographic information like Gender, Date of Birth and Marital Status had high percentage of missing values. Gender and Date of birth were augmented using the bureau data whereas marital status was augmented using salutation file that was provided independently.

## Target Variable Creation

For each product the data is prepared independently. Depending on the product reach and the volume of events, it is decided whether a time stamp or a look-alike model will be developed. Hence the target variable is also created accordingly.

In case of time stamp model, if the customer has accepted the product in the observation period, the earliest instance of the product of acceptance is considered and dependent flag is 1 for that customer. For all other customers the dependent flag is 0. Similarly, in case of look-alike model, any customer who has accepted a product ever dependent flag is 1 for that customer and for all other customers the dependent flag is 0. For products where the product acceptance date is not available, look-alike model will be developed.

## Sampling

Random sampling was done in the data dividing the whole data in 70%-30% as train and test dataset such that they have similar event rates. Model was built on the train dataset and validated on the test dataset.

## Univariate Analysis

The base data consisted of banking relation variables and the balance information from Apr 2012 to Sep 2015. Univariate analysis was done to understand the distribution of the data. In particular, percentile distribution and basic statistics was observed for all continuous variables and the distribution of categories was observed for categorical variables.



## Missing, Caps and Floors

Base data was treated by removing the duplicate customer information. It resulted in a unique row of information for each customer. All continuous variables were capped at 99.9 percentile. This was done to avoid any skewness in data from the outlier values. All the balance variables with value <1 were floored to 1.

It was observed that the demographic variables like date of birth, gender and marital status had high % of missing values. Gender and date of birth were augmented using bureau data whereas marital status was augmented using salutation file. After augmentation, the missing values of date of birth were imputed by the median value.

In case of categorical variables which had more than ~10 sub categories, the top categories that represented atleast 70% of population were kept as it is and the rest of the sub categories were grouped to create another category called ‘Others’.

# MODEL DEVELOPMENT

## Segmentation

After applying all exclusions, SA base contained about 7,03,147customers. Since the base is large, segmentation was done to create homogeneous population and get better model results. Methodology used for segmentation was k-means clustering.

K-means clustering is an algorithm to classify or to group the objects based on attributes/features into K number of group. K is positive integer number. The grouping is done by minimizing the sum of squares of distances between data and the corresponding cluster centroid.

There were about 400+ transaction variables so for segmentation these variables were shortlisted using varclus method. In varclus, clusters are generated with like variables in the same cluster and clusters are heterogeneous with each other. From each of these clusters, variables are selected based on R square ratio. Transaction, demographic and banking variables were input to segmentation procedure.

It was observed that there was a large group of customers that were banking inactive. Therefore two groups were created based on the transaction activity of the customers. Customers who did not make a single transaction (credit/ debit) and if their sum of balance in last 6 months was <100 then they were grouped as banking inactive while others were grouped under active customers.

Clustering was done on banking active customers. While clustering, it was important that the cluster size was big enough to build robust models. After several iterations, three clusters with defined features were created as follows:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Cluster | Count | MOB | Avg\_6\_DB\_AMT | Last\_6\_DB\_CNT | Avg\_6\_Bal | Last\_12\_CR\_CNT | Gender\_F | nbd\_lock\_holding |
| 1 | 107,774 | 65 | 3,310 | 3 | 16,291 | 6 | 11% | 1.22 |
| 2 | 267,666 | 17 | 13,162 | 6 | 26,318 | 8 | 34% | 1.90 |
| 3 | 64,480 | 39 | 124,927 | 54 | 81,029 | 35 | 15% | 2.57 |

Banking inactive customers were further divided into two groups – customers who did not have any product holding and others. Customers who had atleast one product holding were not very engaged but had atleast one product with either IndusInd or any other financial organization.

## Variable Transformations

After the creation of derived variables and segmentation process, binning was done for the continuous variables. This was done for all the products and segments independently. Variables were divided into 10 groups and log of odds and IV was calculated for each group. This was done to study the monotonicity of the variables or the behavior of the variables with the dependent variable. In case of categorical variables, dummy variables were created for each sub-category with value 1/ 0.

## Variable Selection/Reduction

After the creation of derived variables, there were a total of 400+ variables which were shortlisted based on the following:

1. IV Reduction: IV was calculated for each variable (continuous and categorical). For models with variables having high IV, all variables were considered having IV between 0.3 and 1.5. For models, where a lot of variables were having very low IV the threshold was lowered.
2. Variable Clustering: Remaining variables were passed on to be operated under Varclus procedure which divides a set of numeric variables into disjoint or hierarchical clusters depending on eigen values before and after splitting. It is based on divisive clustering technique.

* All variables start in one cluster. Then, a principal components analysis is done on the variables in the cluster to determine whether the cluster should be split into two subsets of variables.
* If the second eigenvalue for the cluster is greater than the specified cutoff, then the initial cluster is split into two clusters. If the second eigenvalue is large, it means that at least two principal components account for a large amount of variation among the inputs.
* To determine which inputs are included in each cluster, the principal component scores are rotated obliquely to maximize the correlation within a cluster and minimize the correlation between clusters.
* This process ends when the second eigenvalues of all current clusters fall below the cutoff.
* If a cluster has only 1 variable in it, it means that this variable has only one principal component and hence, second eigenvalue of this variable is 0.

1. Rank and Plot: After varclus, rank and plots of remaining variables were created by dividing them into deciles and obtaining their individual distribution of target variable. Variables having very erratic distribution of log of odds value were dropped and those with minor deflections were binned.
2. Correlation: Remaining variables were refined using correlation check. One variable out of each pair having more than 60% correlation was dropped. Variable having greater IV and making more business sense was kept while the other one was dropped.
3. Variance Inflation Factor: Finally VIFs for continuous variables were checked for the final set of variables and those having VIF>5 were dropped.

## Model Estimation

The shortlisted variables were provided as input to the model development process. The final model comprised of a set of variables which were selected by the logistic model keeping in mind the p-value, point estimate and Chi-Square value of the variables.

Attached is the model summary along with the list of variables along with their coefficients.



## Selection Criteria

Model selection and evaluation criteria comprised of the following parameters:

1. Concordance – It defines model’s predictive power. A concordant pair is where the probability of response is higher than non-response. The model is considered good if it has concordance ratio of more than 60%.
2. C-value represents area under ROC curve. It is used to compare the goodness of fit of logistic regression models. It ranges from 0.5 to 1. Models are typically considered reasonable when the C-value is higher than 0.7
3. K-S Statistic measures the degree of separation between positive and negative distributions. For a model to be good, K-S Statistic should be in top 3 deciles and a value of more than 40%.
4. Gini is a measure of inequality of a distribution. It can be derived from AUC curve. It has value equals (2 \* Area under AOC curve) – 1. Gini above 50% is a good model.
5. Sensitivity & Specificity - Sensitivity measures the proportion of positive that are correctly identified as such i.e. true positive rate whereas Specificity measures the proportion of negatives that are correctly identified as such i.e. true negative rate. These are defined as follows in terms of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) for classification tasks

Sensitivity =

Specificity =

1. Maximum Probability value predicted for customers should not be very high i.e. close to 1
2. Probability score clustering – Clustering of the probability should not be very high since high clustering means model is not able to discriminate the population very well between event and non-event.

Also, along with the above selection criteria, the model performance was compared with test model results to check the consistency of the performance. All the above criteria were evaluated for test results.

# RECOMMENDATION FRAMEWORK

## Validation

After developing propensity models for each product and segment, the approximate revenue was calculated for each product and applied to the recommendation framework. A rank order by recommendation of products was generated for each customer. The validation of the framework was done by dividing the population into deciles by their average balance in last 6 months and studying the pattern of the product recommendation by these deciles.

The recommendation framework results are attached in reference section.

# IMPLEMENTATION/SCORING

## Spec Sheet

Attached is the user guide used for Scoring code.

****

## Scoring Code



## Missings, Caps/Floors & Transformations

Missing values and outlier treatment is incorporated in the scoring code.

## Implementation Environment

The codes have been implemented in IndusInd SAS environment.

# Ongoing Model Risk Management

## Summary of Model Risks and Limitations

IndusInd is a growing bank and some of the products that are a part of recommendation framework are evolving. With time the event rate on which the models have been developed will change and this can impact the model results significantly. Hence, model performance should be monitored strictly going forward. Also, for many products the events have been augmented using bureau data. This means that the behavior of customers who are holding products from other organizations are considered as event and their behavior is considered to be similar to the customer who are holding products with IndusInd. This can cause a conflict in behavior and may not provide a true representation of event behavior. Models also use data from bureau for product holding and enquiry information. These have proved to be significant and therefore, for the models to perform consistently in future, these models have to be updated in future for latest data.

# References

* Segment profiles



* Model estimation results



* Recommendation Framework results

\*The above recommendation results are based on the full data (CA as well as SA) provided by IndusInd.