

**Operation Arjun - Analytics Documentation(SCV)**

First Draft |01-14-18

MAHINDRA AUTOMOTIVES

Contents

[Contents i](#_Toc467746893)

[1. Data Processing 1](#_Toc467746894)

[1.1 Data processing for propensity modelling 1](#_Toc467746895)

[1.2 Missing Value Imputation 4](#_Toc467746896)

[2. Propensity Modeling 9](#_Toc467746900)

[2.1 Why do we use a propensity model? 9](#_Toc467746901)

[2.2 What are the different model options? 9](#_Toc467746902)

[2.3 Logistic Regression 10](#_Toc467746903)

[2.4 Probabilty cutoff determination for P1/P2/P3 12](#_Toc467746904)

[2.5 Integration with the DMS 14](#_Toc467746905)

[3. Model Refresh & Addition of new fields/brands 9](#_Toc467746900)

[3.1 Why model refresh 9](#_Toc467746901)

[3.2 Model Refresh methodology 9](#_Toc467746902)

1. Data Processing

1.1 Data processing for propensity modeling

1.1.1 Data Sources for propensity modelling

* Data collected from the Dealership Management System
* Data sources cover enquiry generation, test drive and booking stages of enquiry
* DMS tables used are :

Table 1.1 - DMS data sources

|  |  |
| --- | --- |
| Data Source | DMS table name |
| Enquiry Data | AD\_LIST |
| Previously owned vehicle data | PS\_LEAD\_CURNT\_VEHCL\_DTL |
| Prospect Details (Usage\_Area, Load\_Type,  Application, KM\_PerDay, Load\_Carried) | PS\_LEAD\_PROSPCT\_DTL |
| Revised enquiry data pull | PS\_LEAD\_HDR |

* Reference key for merging these three tables – “Parnt\_Grop”||”Loctn\_cd”||”Enq\_no”||”Enquiry\_date”

1.1.2 Variables used for propensity modelling

* Fields captured in the CDIF form used for modeling
* Fields to be considered for modeling picked on basis of data accuracy, fill rates and hypothesis on relation with conversion rate :

Table 1.2 - DMS fields and fill rates

|  |  |  |  |
| --- | --- | --- | --- |
| Variable Name | Variable Description | Used in Logistic Regression | Fill Rate |
| AREA\_OFFICE | Area Office of dealership | No | 100% |
| COLR\_CD | Preferred car color | No | **100%** |
| CUSTMR\_TYPE | **Individual/Corporate** | **Yes** | **100%** |
| DEALER\_LOCATION | Identifier | No | 100% |
| DEALER\_NAME | Identifier | No | 100% |
| DISTRICT\_DESC | District of dealership | No | 100% |
| ENQ\_DATE | Date enquiry created | Filter | 100% |
| ENQ\_NO | Unique Identifier | No | 100% |
| ENQ\_SOURCE | Source of enquiry | No | 97% |
| ENQ\_SO\_CD | **Code assigned to enquiry source** | **Yes** | **100%** |
| ENQ\_STATUS | Status of enquiry | Creation of dependent variable | 100% |
| ENQ\_TYPE | **Type of enquiry** | **Yes** | **100%** |
| FOLW\_ACTL\_DATE | Date of follow up | No | 92% |
| LOCTN\_CD | Location code of dealership | No | 100% |
| MODL\_CD | Code of enquired model | No | 100% |
| MODL\_DESC | Detailed description of model enquired after | No | 100% |
| MODL\_GROP\_DESC | Model enquired after | No | 100% |
| MRC\_SORC | Source of MRC enquiry | No | 9% |
| OCCUPATION | Occupation | **No** | 20% |
| OEM\_MODL\_CD | Model code of car enquired after | No | 100% |
| PARNT\_GROP | Identifier for dealership | No | 100% |
| REFERENCE\_NAME | Name of referee | No | 0% |
| REF\_PHON | Phone number of referee | No | 0% |
| SALS\_MAN | Name of sales man | No | 100% |
| TEHSIL\_DESC | Tehsil name of enquirer | No | 98% |
| VARNT\_DESC | Description of variant enquired afar | **No** | 100% |
| ZONE | Dealership zone | No | 100% |
| Dealership | Dealership name | Identifier | 100% |
| ENQ\_AGE | **Age** | **Yes** | **98%** |
| Consideration Set | Data not available | No | NA |
| Intended date of purchase | Data not available | No | 100% |
| Exchange Model | Data not available | No | NA |
| Usage Area | Area in which vehicle will be used | Yes | 100% |
| Load Type | Type of load carried by vehicle | Yes | 100% |
| Application | Type of Application for which vehicle will be used | Yes | 100% |
| First Time User | Whether using vehicle for first time or not | No | 100% |
| Driver Cum Owner | Whether person will drive on his own | No | 70% |
| FINC\_IND | Whether person has taken finance or not | No | 100% |
| KM\_PER\_DAY | Distance travelled by a person | Yes | 100% |
| LOAD CARRIED | Load carried by enquired model | Yes | 100% |
| VEHCL\_MAK | Previous car ownership | **No** | 15% |
| VEHCL\_MODL | Previous car ownership | **No** | 15% |

1.1.3 Additional feature generation

Coarse classing done for some variables:

v

**Dealer Category**

v

**Pin code**

v

v

**Km Travelled**

**Previously Owned Vehicle Brand**

v

**Previously Owned Vehicle Category**

v

**Load Carried**

***Attachments***

Annotated R-code for data processing and feature generation

1.2 Missing Value Imputation

Imputation is the process of replacing missing data with substituted values. There are three main problems that missing data causes:

Because missing data can create problems for analyzing data, imputation is seen as a way to avoid pitfalls involved with [list wise deletion](https://en.wikipedia.org/wiki/Listwise_deletion) of cases that have missing values.

1.2.1 Algorithm Leveraged for Imputation of Enquirer’s age (Enq\_Age)

Missing values of “Age group” variable has been imputed using Multivariate Imputation by Chained Equation (MICE) technique

MICE-The mice package implements a method to deal with missing data. The package creates multiple imputations (replacement values) for multivariate missing data. The method is based on Fully Conditional Specification, where each incomplete variable is imputed by a separate model. The MICE algorithm can impute mixes of continuous, binary, unordered categorical and ordered categorical data.

1.2.2 The mice package contains functions to

* Inspect the missing data pattern
* Impute the missing data *m* times, resulting in *m* completed data sets
* Diagnose the quality of the imputed values
* Analyze each completed data set
* Pool the results of the repeated analyses
* Store and export the imputed data in various formats
* Generate simulated incomplete data
* Incorporate custom imputation methods

*The univariate imputation algorithm leveraged for imputing Enquirer’s Age (ENQ\_AGE) under MICE package is Polytomous logistic regression.*

2. Propensity Modeling

2.1 Why do we use a propensity model?

The propensity model is needed to predict the likelihood of a customer buying a vehicle s/he has enquired for.

2.2 What are the different model options?

Since, we are predicting a binary outcome, we can use either:

* A decision tree based approach (non-parametric: CART of CHAID) OR
* Logit models (parametric).

2.2.1 Key advantages of each model

2.2.1.1Decision Tree Advantages

* **Easier to interpret** - Output can be essentially interpreted as a set of business rules segregating the population and provides an easy visual representation of the data
* **No issue with missing values** – Missing values get routed to separate branches and imputation is not necessary. Particularly helpful in cases where strong hypotheses for missing value imputation can’t be formulated
* **Provides a deep understanding of the data** in an easy and succinct manner

2.2.1.2 Logit model Advantages

* **Accuracy gain** - Prediction is in the form of a continuous variable with individual probabilities being predicted for each observation. Hence, it is more accurate than a CHAID model where observations are binned together in one probability bucket
* **Statistical robustness** : Being a parametric model, it is more statistically robust

2.3 Logistic Regression

**2**.**3.1 Why logistic regression?**

* Have higher predictability but relatively difficult to interpret
* Prediction is in the form of a continuous variable as opposed to discrete buckets in CHAID, making P1/P2/P3 incidence adjustment smoother
* Can be utilized as a second generation model to enhance results obtained via CHAID trees

**2.3.2 Logistic regression - model building**

***Software Used – R***

* Basic dataset prepared (methodology outlined in the data processing section)
* Dependent variable – Conversion (binary 1/0 variable; 1 – customer bought a car, 0 – customer didn’t buy the car)
* Independent Variables – All CDIF variables were examined and the following were picked based on hypotheses about ability to predict conversions and data accuracy (refer to the data processing section for data fill rates) :
  + - Pin code categories (categorization detailed in data preparation)
    - Dealer categories (categorization detailed in data preparation)
    - Enquiry Source
    - Enquiry Type
    - Enquirer’s age
    - Previously owned vehicle category (*Not Significant as of now*)
    - Previously owned vehicle brand (*Not Significant as of now*)
    - Customer type
    - Load Type (*Not Significant as of now*)
    - Application
    - Usage Area
    - Km travelled (interval classification detailed in data preparation)
    - Load Carried (interval classification detailed in data preparation)
* Training and Out of sample dataset created :

- Maxi truck, Jeeto and Supro Load –

1’st Nov 2015 to 31st March 2017 (Training)

- 1st April 2017- 31st May 2017 (Out of sample)

-Supro Passenger–

1’st Nov 2015 to 31’st May 20167(Training)

- Cross validation technique was used (Data size too small for OOS)

- Alpha Passenger and Pick up–

1’st Nov 2015 to 31st May 2017 (Training)

- 1’st June 2017 to 31’st August 2017 (Out of sample)

- Jeeto Passenger, Imperio and Alfa Load-

1st Nov 2015 to 31st August 2017(Training)

- Cross validation technique was used (Data size too small for OOS)

*Note that all enquiries must be given at least 8 weeks to convert, thus time periods should be adjusted accordingly to avoid under-reporting of conversion rates.*

* Training sample picked up and further divided into modeling and testing datasets (70% to 30% split has been done in this case).
* Note: Please set the seed before doing this step to ensure replicability of results
* All variables are converted to factors (we are currently not using any continuous variable)
* Run a stepwise regression for variable selection,
* Make a full model with all variables (including intercept)
* Make a null model with only intercept
* Run stepwise regression using the null model as lower bound and full model as the upper bound
* Following parameters can be adjusted to provide flexibility,
  + Direction of regression – Can be set as “both” for stepwise; “forward” for forward selection; “backward” for backward section
  + Steps – Maximum number of steps to be considered (default is 1000)
  + K – Used to adjust degrees of freedom for penalizing AIC criteria, higher degrees of freedom will penalize AIC more (set as 5 usually)
* The stepwise regression function adds and drops variables in each step such that it can achieve an improvement in the AIC measure
* The function stops when no further improvement can be achieved and ends up choosing the equation with the least AIC
* The output is formula based and final equation can be seen by calling formula()
* The model selection path can be seen using summary()
* From the selected model, we check the VIF of all variables using the CAR package
* In case there is VIF (in the initial logistic regression, VIF was present only between Previously Owned Car Brand and Previously Owned Vehicle Category),
* Calculate Wald chi-squared values of the variables with high VIF
* Remove variable with lowest wald chi-squaed and run a regression using glm(), make note of the AIC
* Remove variable with second highest wald chi-squared and run a regression using glm(), make note of the AIC
* Select model with the lowest AIC
* Re-run the car() package to check VIF with the dropped variables, we should have VIF < 2 for all variables

2.4 Probability cutoff determination for P1/P2/P3

2.4.1 Why it is needed

* Customers will be scored based on their likelihood to convert and assigned priorities according to the scores.
* Customers will be assigned three priorities – High, Medium and Low (P1/P2/P3)

2.4.2 How are thresholds determined?

* ROC curve is used to determine P1/P2/P3 thresholds

2.4.3 What is ROC?

* The receiver operating curve shows the tradeoff between True Positive Rate (Percentage of correctly predicted Positives) and false Negative Rate (percentage of falsely predicted negatives) for each possible cut off level (cut offs are determined by the terminal nodes in case of CHAID)

2.4.4 Rationale for determining thresholds

* Usually, ROC automatically predicts response using the cutoff corresponding to maximum sensitivity given minimum specificity
* For us, the cost of false negative is higher than the cost of a false positive. Thus, we want to customize ROC threshold such that sensitivity is optimized (given a minimum level of sensitivity or accuracy)
* Minimum percentage of P1 1s needed to make strategy feasible

2.4.5 Methodology for determining customized ROC threshold

*Software used – R*

* Divide training dataset into smaller datasets for each dealership
* For each of these datasets create ROC curves and determine the threshold which gives a stipulated number of P1s, captures majority of conversions in P1s and has minimal level of accuracy
* The specific loop run to achieve this is as follows –
* Build ROC curve using variable node probabilities for a dealership
* If ROC curve can't be built, then all further calculations are stopped
* If ROC curve can be built, find minimum probability at which Rate of Positive Predictions [(TP+FP)/all] is >= 30%
* Find maximum point at which Rate of Positive predictions [(TP+FP)/all] is < 30%
* Calculate the deviation of both the above points from 30% and take the point with the least deviation
* Calculate the accuracy and True Positive Rate at this point
* If True positive rate >= 60% and accuracy >= 40% then the point calculated above would be the P1 cutoff
* If either TPR < 60% or Accuracy < 60%, find the points on the ROC where TPR > =60% and Accuracy >=40%
* Take the minimum of these two points and set as threshold for P1
* Exclude all enquiries tagged as P1 based on above logic and make a P2+P3 dataset
* Build another ROC curve for this dataset
* If ROC curve can't be built, then all the enquiries become P2
* If ROC curve can be built, find the minimum point where Rate of Positive Predictions [(TP+FP)/all] >=50%
* Take this point as the P2 threshold
* Label all enquiries below this probability as P3
* Two outputs are produced by this code. These outputs can be pasted in the attached template for an easy understanding of model statistics
* Check if accuracy is above 40% and incidence of P1/P2/P3 is satisfactory
* For each dealership check if conversion of P1 > Conversion of P2 > Conversion of P3 (template for this check is attached)
* Save dealership thresholds

**2.4.6 Validation – Testing Sample**

* Pass the testing sample (30% holdout) through the final modeling object obtained in step 2.3.2 above
* Merge the resultant scored dataset with the thresholds data and prioritize all enquiries as P1/P2/P3
* Compare incidence and conversion rate of P1/P2/P3 with the training sample
* Compare AUC, Accuracy and Specificity with the training sample as well

**2.4.7 Validation – Out of sample**

* Pass the out of sample dataset through the final modeling object obtained in step 2.3.2 above
* Merge the resultant scored dataset with the thresholds data and prioritize all enquiries as P1/P2/P3
* Compare incidence and conversion rate of P1/P2/P3 with the training sample
* Compare AUC, Accuracy and Specificity with the training sample as well

**2.4.8 K-fold cross validation**

**2.4.8.1 Why cross validation**

* Cross validation is done to ensure that the model in not over fitted
* Used when out of sample dataset is small

**2.4.8.2 Cross validation methodology**

* Train Data is divided into 10 samples.
* Each sample is picked up and labeled as test dataset iteratively.
* The rest of the observations are used for training and the model equation used post stepwise regression is run on the training dataset.
* Model obtained above is applied on the test dataset.
* Both files are merged with the ROC thresholds and each enquiry is prioritized as P1/P2/P3
* Conversion and incidence of P1 is documented along with accuracy and specificity of the training and testing models
* Process repeated for all 10 samples
* All metrics compared, in subsequent samples, P1 incidence and conversion rate should be very similar

2.4 Integration with the DMS

* The algorithm has been integrated with the dealership management system to provide real time prioritization of incoming enquiries
* Model produced in R is translated into simple linear equation as follows:

Log (p/1-p) = β0 + ∑βiXi

p/1-p= exp (β0 + ∑βiXi)

p= exp (β0 + ∑βiXi) \*(1-p)

p\*(1+ exp (β0 + ∑βiXi)) =exp (β0 + ∑βiXi)

p= exp (β0 + ∑βiXi)/ (1+ exp (β0 + ∑βiXi)

p=1/1+exp (-(β0 + ∑βiXi))

* In the DMS, an enquiry has to go through the following steps,
* Enquiry generated
* Fields relevant to algo picked up (same as current algo implementation)
* Fields converted to algorithm values using lookups (same as current algo implementation)
* Converted into binary encoding based on presence of different field values
* Pick up the regression coefficient table based on cluster + car model combination
* Score the enquiry
* Match with P1/P2/P3 threshold table and prioritize enquiry as P1/P2/P3
* DMS team uploads these files in the DMS (process takes ~1 day) and prioritizes all open enquiries
* Any new enquiry created using app or DMS will also get prioritized post the integration

3. Model Refresh & Addition of new fields/brands

3.1 Why refresh the model

* Auto industry is fast paced with constant changes in,
  + Competitive offering
  + Own product offerings
  + Macroeconomic conditions (pay commissions etc.)

Thus customer profile for a car brand is constantly changing

* Model has to comply with **CDIF form**. Whenever there is a change in CDIF form model updation is required.
* For newly opened dealerships **or** Dealerships with high P1 **or** Dealerships whose ROCs cannot be built (due to fewer enquiries) while initial level model building can be rebuilt later (in latest refresh) to get accurate/latest priority scores and thresholds.

In the initial 1-2 years of model implementation, as enquiry generation is simplified using the app, more accurate data will flow in. Models butyl on historical data will have to be changed to incorporate recent more accurate data

3.2 Model refresh methodology

3.2.1 Data for model refresh

* Model refresh should capture new data being routed through the app. Since this is more accurate data, more weightage should be given to this data while creating the model
* To do this, consider making model on 6 months instead of 12 months (this would ensure that the months of Operation Arjun rollout have relatively more weightage)
* For model refresh, following types of data will be needed,

1. Enquiry data

2. Test drive data

3. Previous vehicle ownership data

4. Demographic data

* The variables asked for will remain the same as outlined in Section 1 (Data Processing)
* Data can be obtained from,
  + DMS Team – Nilesh Borole
  + Bristelcone Team – Vishal Shah
* Data preparation will be same as section 1 above