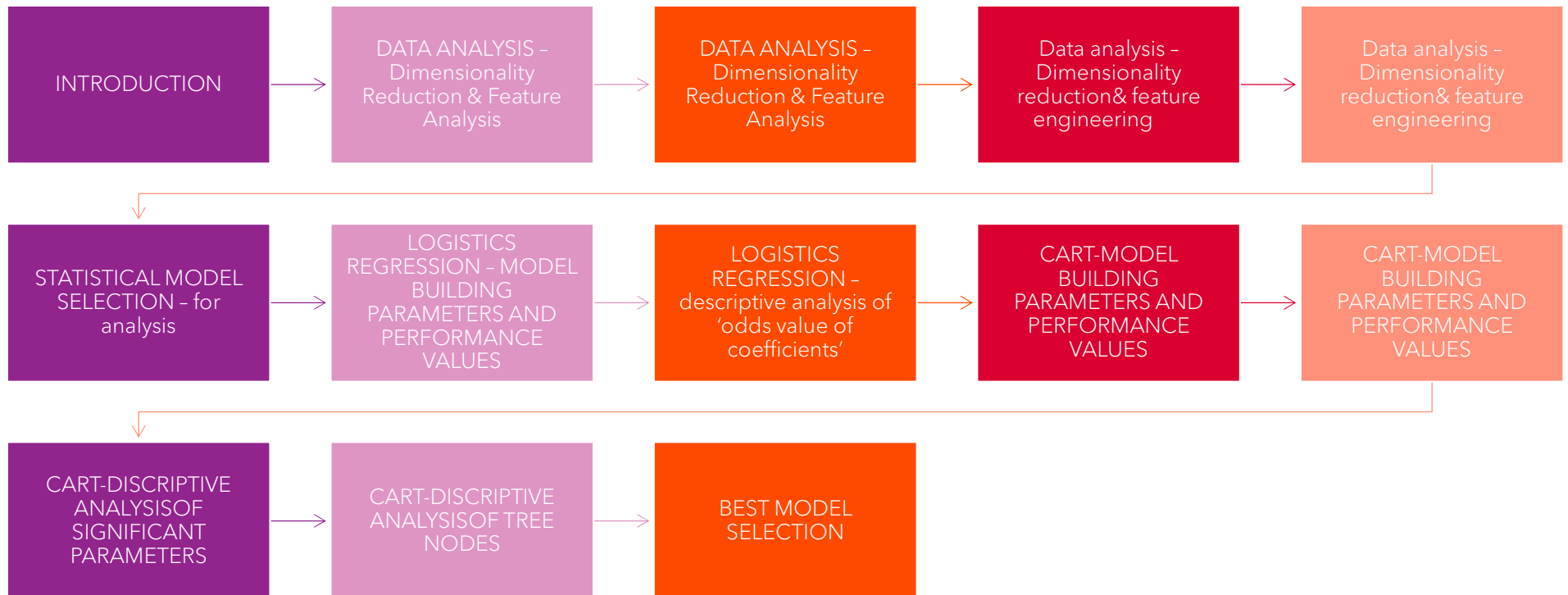


DELIVERY RISK ANALYSIS IN SUPPLY CHAIN- STATISTICAL MODEL BUILDING

- CAPSTONE PROJECT FOR
PG-BABI[PART 2]



INDEX




INTRODUCTION -

The dataset in hand - represents a Supply Chain management system wherein we found from the EDA that the company is in HIGH RISK of LATE DELIVERY.



With LATE DELIVERY in 54% of the orders placed - we found that Shipment Scheduling is an area to focus for Improvement.



We will now be attempting to analyze the problems, quantify the issues statistically and give Actionable Insights for improving SCM of the company - by applying Machine Learning techniques.

DATA ANALYSIS - DIMENSIONALITY REDUCTION & FEATURE ENGINEERING

- **FEATURE ENGINEERING -**

- The independent variables - 52 of them - given in the dataset must be verified for its inter-dependency.
- Having declared 'Late,Delivery.Risk' variable as the one upon which the statistical analysis will be made, all rest of the variables(51 of them known as 'Predictor' variables) must be checked for inter-dependency and significance of each 'Predictor Variable' upon the Target Variable', i.e. 'Late.Delivery.Risk.
- 'CHI.SQUARE Test for Significance' lets us know which all predictor variables are significant with respect to the target variable.
 - This test done on the basic dataset, eliminated 25 predictor variables which were found insignificant.
 - 'scm2' dataset was generated with 180511 observations and 30 variables. 08 NA's were also omitted in this process.
- EDA of the original dataset provided insights for evaluating the calculated field - ORDER PROCESSING TIME = ["Shipping. Date" - "Order. Date"] This predictor variable was created as an engineered variable and added to 'scm2'.
- Dataset 'scm6' is now generated from 'scm2' - ready for dimensionality reduction.
- Dimensionality reduction techniques are used to nullify the multi-collinearity between predictor variable - both Numerical & Categorical variables.

- **DIMENSIONALITY Reduction through PCA [Principle Component Analysis]-**

- I. The most popular method used for dimensionality reduction is "PCA"
- II. With PCA, a new dataset is created wherein the VARIABLES with similarity are clubbed together and a new predictor variable is created.

DATA ANALYSIS - DIMENSIONALITY REDUCTION & FEATURE ENGINEERING

- TRANSFORMED VARIABLES -

	BEFORE ' PCA' - Numerical variables	AFTER ' PCA'
1	Days.for.shipping..real	Shipment.Scheduling
2	Days.for.shipment..scheduled	
3	orderprocessingtime	
4	Benefit.per.Order	Profitability
5	Sales.per.Customer	Sales.Order.Value
6	Order.Item.Product.Price	

	BEFORE ' PCA' - Categorical variables	AFTER ' PCA'
1	Customer.Id	Customer.ID
2	Order.Customer.Id	
3	Type	Order.Fullfilment
4	Order.Status	
5	Customer.State	Customer.Order.Origin
6	Order. City	Delivery.Destination
7	Order.State	
8	Market	Delivery.Region
9	Order.Region	
10	Customer.Fname	Customer. Name
11	Customer. Street	
12	Shipping. Mode	Shipping.Class

DATASET 'scm7' was generated with the above transformations - for MODEL BUILDING.



STATISTICAL MODEL SELECTION - FOR ANALYSIS

- PURPOSE of the MODEL-

Our priority is to understand - What is wrong in the system that is making the ORDER's DELIVERED LATE.

This , thereby is a case where we need to get in-depth analytics of significant factors that delays the ORDER DELIVERY, rather than predicting an outcome of whether the Delivery of ORDER will be late or NOT.

A pre-emptive correction can always make the SCM robust and thereby nullify LATE DELIVERY.

- MODEL SELECTION -

With a descriptive analysis in mind, LOGISTICS REGRESSION & CART models will be best suitable for the given scenario.

LOGISTICS Regression- enables us to carry out Prediction as well as find out explicitly and quantifiably - which all factors contribute to LATE DELIVERY.

CART - Classifies the significant variables and STEP-WISE each leaf of the CART TREE defines and quantifies the Loss of Delivery made due to Each predictor variable.

LOGISTICS REGRESSION - MODEL BUILDING PARAMETERS AND PERFORMANCE VALUES

SPLITTING DATASET to 'TRAIN' & TEST'-

- Dataset is split into TRAIN & TEST Dataset with a 70:30 ratio.
- Proportionality is checked w.r.t. Target Variable presence - to confirm uniformity in split.

MODEL BUILDING -

- Initial model is built utilizing all the predictor variables of the dataset.
- The 'summary' of the model is inspected to find out whether all coefficients are significant or not.
- The model is pruned again by removing the in-significant variable - 'Delivery. Destination'.
- Model 'logitmodel1' is generated with less AIC value than 'logitmodel' - further taken up for performance analysis.
- For checking the Accuracy and Sensitivity, the prediction scores are generated.

	ACCURACY	SENSITIVITY	AUC Value	THRESHHOLD Value
TRAIN Dataset	0.7956	0.8289	0.7921	0.5
TEST Dataset	0.795	0.8287	0.7915	0.5

LOGISTICS REGRESSION - DESCRIPTIVE ANALYSIS OF 'ODDS VALUE OF COEFFICIENTS'

ODDs VALUE - DISCRIPTIVE ANALYSIS-

TEST DATASET	Intercept	Coeff. Value	ODDs Value	1-ODDs Value is used for Interpretation	Probability
Customer ID	1.3242499	1.0393562	2.3636061	1.3636061	0.702700028
ORDER Fullfillment	1.3242499	0.9249505	2.2492004	1.2492004	0.692231972
CUSTOMER Order Origin	1.3242499	0.9788104	2.3030603	1.3030603	0.697250456
Delivery Region	1.3242499	0.9730317	2.2972816	1.2972816	0.696719868
Customer Name	1.3242499	0.586046	1.9102959	0.9102959	0.656392328
Shipping Class	1.3242499	0.1108482	1.4350981	0.4350981	0.589338926
Shipment Scheduling	1.3242499	7.4216534	8.7459033	7.7459033	0.897392785
Profitability	1.3242499	0.9681491	2.292399	1.292399	0.696270106

- Shipment. Scheduling - The [1-ODDs] value indicates that one-point increase will ensure SEVEN Times, the LATE DELIVERY of the order. Shipment Scheduling comprises of 'Order Processing Time'; Shipping Date & Order Date.
- Customer ID, Order Fulfillment, Customer Order Origin, Delivery Region & Profitability - have shown equal importance , such as - these variables will ensure TWO Times More the LATE Delivery of the order.
- Customer Name & Shipping Class - an increase in these variable values reduces the LATE DELIVERY RISK.

Insight - Interesting to note that 'Shipping Class' is NOT CONTRIBUTING TO LATE DELIVERY.

CART-MODEL BUILDING PARAMETERS AND PERFORMANCE VALUES

SPLITTING DATASET to 'TRAIN' & TEST'-

- The same dataset used for Logistics Regression is used for CART Modelling.

MODEL BUILDING -

- Initial model is built to allow maximum nodes to the TREE. Control parameters are kept to its minimum.
- Thereafter the CP Value - & CROSS-Validation error is derived .
- It is found that THERE IN NO OVER-FITTING Observed - by which the model developed is suggested to be robust.
 - Overfitting observed when the 'xerror' value reduces and then tends to increase.

	CP VALUE	n split	Rel. Error	x error	xstd
1	0.423057	0	1	1	0.0031021
2	0.321274	1	0.576943	0.576943	0.0027356
3	0.046806	2	0.255669	0.255669	0.0019916
4	0.016835	3	0.208863	0.208863	0.0018214
5	0.015774	8	0.106624	0.106624	0.0013341
6	0.014889	10	0.075075	0.075075	0.0011278
7	0.01	11	0.060186	0.060186	0.0010133

CART-MODEL BUILDING PARAMETERS AND PERFORMANCE VALUES

TREE PRUNING -

- Now the control parameters for tree building are evaluated , such as CP value and XVAL [cross validation value / xerror value] from the above table.
- The values are input into the model building and the FINAL MODEL is generated - 'prunedtraincarttree'.
- The same model is used to generate the prediction scores in the TRAIN & TEST Dataset.
- The Accuracy & Sensitivity parameters are evaluated from the confusion matrix generated for the target variable.

	ACCURACY	SENSITIVITY
TRAIN Dataset	0.9728	0.985966
TEST Dataset	0.9719	0.98674

CART- DISSCRIPTIVE ANALYSIS OF SIGNIFICANT PARAMETERS

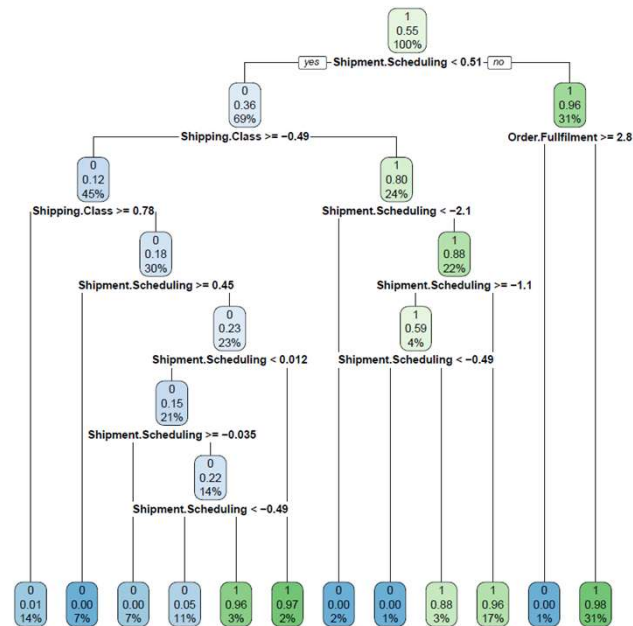
SIGNIFICANT PARAMETERS-



Interesting to note that BOTH - 'Shipping Schedule' as well as 'Shipping Class' has gained importance is deciding the LATE DELIVERY RISK of the orders.

CART- DISSCRIPTIVE ANALYSISOF TREE NODES

CART TREE & NODES ANALYSIS-



- NODE 1 - branch is created with the most important variable - Shipment.Scheduling(S.S).
- NODE 2 & 3 bifurcates S.S with parameter values below 0.51 & Above| Equal to 0.51. Here , it indicates how many orders are Late Delivered due to Shipment Scheduling.
- NODE 4,5,6 & 7 bifurcates with respect to 'Shipment Class(S.C)- indicating the effect of S.C over Late Delivery orders. This nodes sits over the S.S factor and further regulates the Delivery Order status.

From the CART Tree, we get a very explicit and quantified output as to at which stage the LATE DELIVERY RISKS are prominent and DUE to Which all combined Reasons.

BEST MODEL SELECTION



BEST MODEL SELECTION



LOOKING AT THE PROBLEM IN HAND AND THE ANALYTICAL PLAYFIELD EACH MODEL DISPLAYED, ' CART - Classification AND Regression Tree' TECHNIQUE IS THE BEST SUITED.

THE SENSITIVITY WHICH THE MODEL DISPLAYED SHOWS THAT 'CART' MODEL IS ROBUST IN ITS PERFORMANCE.

THANK
YOU!

