|  |
| --- |
| Supply Chain  Analysis |
|  |
| June-2020  Great Learning  Authored by: Manika Gupta |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Aim of Project: To analyze the qualitative and quantitative company data and evaluate the problem areas to improve the business. Suggest a strategy or process to make data-driven decisions based on evidence and analytics.  A Dataset of Supply Chain used by the company Data Co Global has been used for the analysis. Areas of important registered activities: Provisioning, Production, Sales, Commercial Distribution. It also allows the correlation of Structured Data with Unstructured Data for knowledge generation.   * Problem Understanding * Identifying the Target Variable * Descriptive analysis * Perform EDA * Diagnostic analysis * Predictive analysis * Prescriptive analysis   Understanding the various variables  Univariant analysis  The variables provided in the dataset are both categorical and numerical.  List of variables and their type and range  'data.frame': 180516 obs. of 48 variables:  Type : chr "DEBIT" "TRANSFER" "CASH" "DEBIT" ...  Days.for.shipping..real. : int 3 5 4 3 2 6 2 2 3 2 ...  Days.for.shipment..scheduled.: int 4 4 4 4 4 4 1 1 2 1 ...  Benefit.per.order : num 91.2 -249.1 -247.8 22.9 134.2 ...  Sales.per.customer : num 315 311 310 305 298 ...  Delivery.Status : chr "Advance shipping" "Late delivery" "Shipping on time" "Advance shipping" ...  Late\_delivery\_risk : int 0 1 0 0 0 0 1 1 1 1 ...  Category.Id : int 73 73 73 73 73 73 73 73 73 73 ...  Category.Name : chr "Sporting Goods" "Sporting Goods" "Sporting Goods" "Sporting Goods" ...    Customer.City : chr "Caguas" "Caguas" "San Jose" "Los Angeles" ...  Customer.Country : chr "Puerto Rico" "Puerto Rico" "EE. UU." "EE. UU." ...  Customer.Fname : chr "Cally" "Irene" "Gillian" "Tana" ...  Customer.Id : int 20755 19492 19491 19490 19489 19488 19487 19486 19485 19484 ...    Customer.Lname : chr "Holloway" "Luna" "Maldonado" "Tate" ...    Customer.Segment : chr "Consumer" "Consumer" "Consumer" "Home Office" ...  Customer.State : chr "PR" "PR" "CA" "CA" ...  Customer.Street : chr "5365 Noble Nectar Island" "2679 Rustic Loop" "8510 Round Bear Gate" "3200 Amber Bend" ...  Customer.Zipcode : int 725 725 95125 90027 725 14150 725 33162 725 94583 ...  Department.Id : int 2 2 2 2 2 2 2 2 2 2 ...  Department.Name : chr "Fitness" "Fitness" "Fitness" "Fitness" ...  Latitude : num 18.3 18.3 37.3 34.1 18.3 ...  Longitude : num -66 -66 -122 -118 -66 ...  Market : chr "Pacific Asia" "Pacific Asia" "Pacific Asia" "Pacific Asia" ...  Order.City : chr "Bekasi" "Bikaner" "Bikaner" "Townsville" ...  Order.Country : chr "Indonesia" "India" "India" "Australia" ...  Order.Customer.Id : int 20755 19492 19491 19490 19489 19488 19487 19486 19485 19484 ...  order.date..DateOrders. : chr "1/31/2018 22:56" "1/13/2018 12:27" "1/13/2018 12:06" "1/13/2018 11:45" ...  Order.Id : int 77202 75939 75938 75937 75936 75935 75934 75933 75932 75931 ...  Order.Item.Cardprod.Id : int 1360 1360 1360 1360 1360 1360 1360 1360 1360 1360 ...  Order.Item.Discount : num 13.1 16.4 18 22.9 29.5 ...  Order.Item.Discount.Rate : num 0.04 0.05 0.06 0.07 0.09 ...  Order.Item.Id : int 180517 179254 179253 179252 179251 179250 179249 179248 179247 179246 ...  Order.Item.Product.Price : num 328 328 328 328 328 ...  Order.Item.Profit.Ratio : num 0.29 -0.8 -0.8 0.08 0.45 ...  Order.Item.Quantity : int 1 1 1 1 1 1 1 1 1 1 ...  Sales : num 328 328 328 328 328 ...  Order.Item.Total : num 315 311 310 305 298 ...  Order.Profit.Per.Order : num 91.2 -249.1 -247.8 22.9 134.2 ...  Order.Region : chr "Southeast Asia" "South Asia" "South Asia" "Oceania" ...  Order.State : chr "Java Occidental" "Rajastán" "Rajastán" "Queensland" ...  Order.Status : chr "COMPLETE" "PENDING" "CLOSED" "COMPLETE" ...  Product.Card.Id : int 1360 1360 1360 1360 1360 1360 1360 1360 1360 1360 ...  Product.Category.Id : int 73 73 73 73 73 73 73 73 73 73 ...  Product.Name : chr "Smart watch " "Smart watch " "Smart watch " "Smart watch " ...  Product.Price : num 328 328 328 328 328 ...  Product.Status : int 0 0 0 0 0 0 0 0 0 0 ...  shipping.date..DateOrders. : chr "1/18/2018 11:48" "1/18/2018 12:27" "1/17/2018 12:06" "1/16/2018 11:45" ...  Shipping.Mode : chr "Standard Class" "Standard Class" "Standard Class" "Standard Class" ...  So first analyzing the numerical variables.  The distribution of various variables across the dataset      The variable Benefit per order graph shows that its left skewed   * The Benefits on order w.r.t the status of order. One thing to be noticed that many orders with pending payment are high benefit orders. * Major Payment successfully completed orders are with high benefit.     Major sales below 500, so the graph is right skewed    Based on the sales this variable is also right skewed.    This variable is same as sales so can chose either one of them. This can further be confirmed after seeing the multicollinearity graph    The profit ratio is more towards positive side .    Mostly the orders have one product per order, it’s a categorical variable    Again this is same as sales data distribution    Total amount per order is same as the sales per order considering the orders mostly consist of single item.    Profit per order is same as benefit per order    Again, the distribution of this variable is same as order item total  Now, analyzing the categorical variables    Most of the payment made is through Debit    As per the above distribution Late delivery is highest 54% of the total data as compared to Advance shipping , shipping cancelled and shipping on time.    We can infer from the graph the most sold categories are: 17,18,24,46,45,48,43,9,29        Higher no of purchase made by customers from EE.UU    Consumers have purchased more online then corporate and the least no of purchase made by home office category customers      Department id is the id of the store from where the delivery of the product is made to the order region / location. Most of the deliveries are made from the department id 7 then 4.    Most late deliveries are from department id 7 i.e store whose id is 7 and most of them are shipped via second class.          Late delivery risk is more at destination country Central America    Most of the orders are complete.      Most of the deliveries are made through Standard class. Most no of late deliveries were made by second class shopping mode.  TARGET VARIABLE:  Late delivery risk is the target variable and all the other attributes are predictors.  Now carrying out a bivariant analysis  Distribution of type of customer and late delivery spread “1” for late delivery and “0” for not   |  |  |  | | --- | --- | --- | |  | 0 | 1 | | Consumer | 42255 | 51247 | | Corporate | 24806 | 29982 | | Home Office | 14479 | 17747 |   Highest late deliveries for consumers  Distribution based on the country the order was delivered   |  |  |  | | --- | --- | --- | |  | 0 | 1 | | EE. UU. | 50158 | 60985 | | Puerto Rico | 31382 | 37991 | |  |  |  |   Most late deliveries are made to the EE.UU country  Type Distribution    0 1  CASH 8507 11109  DEBIT 29645 39648  PAYMENT 17720 24004  TRANSFER 25668 24215  Most late deliveries were paid through debit card  Delivery status distribution    0 1  Advance shipping 41592 0  Late delivery 0 98976  Shipping canceled 7754 0  Shipping on time 32194 0  This is the same status as the late delivery status  Distribution of late delivery, based on target market   |  |  |  | | --- | --- | --- | |  | 0 | 1 | | Africa | 5274 | 6340 | | Europe | 22508 | 27742 | | LATAM | 23550 | 28044 | | Pacific Asia | 18547 | 22712 | | USCA | 11661 | 14138 |   Target Market where late deliveries are highest is LATAM and next is Europe.    Multicollinearity    Highly correlated attributes: order item total and sales per customer and sales  Profit ratio and benefits per order  Order item profit and Product price  Order profit per order and Benefit per order  Order item product price and product price  Mildly Correlated attributes: Days of shipment scheduled and real  Order item total and order item profit  Order item profit ratio and benefit per order      Multicollinearity happens when one predictor variable in a multiple regression model can be linearly predicted from the others with a high degree of accuracy. This can lead to skewed or misleading results    **Zipcode with maximum late deliveries payment based (2)**  **Clustering**  Supply chain clustering brings a new industrial organization paradigm with strong logistics support. Customer clustering means rowbased clustering is done. The main findings establish that clustering the common links (activities, profits, delivery modes, types of product, target locations, shipping mode, processes and/or services) of supply chains from different sectors running operations, enables a swift answer to an everyday more volatile demand and delivery system, more resilient to the disruptions caused by men and nature, and reducing the total logistics cost.  Clustering to understand the data better   | Group.1  <int> | Benefit.per  .order  <dbl> | Sales.per.  customer  <dbl> | Latitude  <dbl> | Longitude  <dbl> | Order.Item.  Discount  <dbl> | Order.Item.  Discount.Rate  <dbl> | Order.Item.  Product.Price  <dbl> | Order.Item.  Profit.Ratio  <dbl> | Order.Item.  Quantity  <dbl> | order.lead  .time  <dbl> | Delivery  .lead.time  <dbl> | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | 1 | 86.57 | 346.52 | 29.77 | -84.56 | 45.56 | 0.118 | 373.65 | 0.25 | 1.18 | 3.52 | 6.56 | | 2 | 31.38 | 146.27 | 37.03 | -97.20 | 15.26 | 0.098 | 90.48 | 0.21 | 2.32 | 4.65 | 8.25 | | 3 | 31.30 | 147.40 | 18.52 | -66.39 | 15.44 | 0.099 | 91.87 | 0.20 | 2.31 | 3.67 | 6.78 | | 4 | 31.24 | 150.17 | 35.50 | -94.67 | 15.60 | 0.098 | 95.92 | 0.20 | 2.29 | 1.76 | 3.46 | | 5 | -275.44 | 220.30 | 29.73 | -84.49 | 24.45 | 0.101 | 173.70 | -1.348 | 2.18 | 3.54 | 6.54 |   Now based on clustering we make clusters (color grades for 5 clusters) based on the category id and the size of the rectangles is based on the benefits per order.  Sheet 1  Takeaway from the clusters diagram: In cluster one highest benefit is from cluster 45    Cluster 1 has highest benefits record  Cluster 5 has losses record | | | | | |
|  | Colors are the cluster markings and so the cluster 2 has the maximum no of late deliveries almost equivalent to the cluster 3. Cluster 5 has least late deliveries.      New Features derived: Feature / Variable creation is a process to generate a new variables / features based on existing variable(s). We can generate new variables that may have better relationship with target variable.  1: Order lead time  2: Delivery lead time  3: latitude and longitude of the Order delivery destination  4: Consecutively the distance between the Order delivery store and the Location where the order must be delivered  The above graph shows the order lead time and the Delivery lead time based on customer clustering. Clustering was done after normalizing the data and then the clusters where collaborated in the prior dataset and then the characteristics were drawn to understand the clusters of the dataset. Since it’s a big data, we must cluster the columns to get a better understanding of the data. This is done by dimensionality reduction using PCA ( and MCA (Multiple Correspondence analysis). Now, clustering the dataset based on columns. Since the dataset has numeric as well as categorical variables, we need to perform PCA and MCA  Multiple correspondence analysis (MCA) is a data analysis technique for nominal categorical data, used to detect and represent underlying structures in a data set. MCA can be viewed as an extension of simple correspondence analysis (CA) in that it is applicable to a large set of categorical variables.  Principal component analysis (PCA) is a technique for reducing the dimensionality of datasets, increasing interpretability but at the same time minimizing information loss. It does so by creating new uncorrelated variables that successively maximize variance  MCA plot: showing cluster based on variables    In the indicator matrix approach, associations between variables are uncovered by calculating the chi-square distance between different categories of the variables and between the individuals. These associations are then represented graphically which eases the interpretation of the structures in the data. Oppositions between rows and columns are then maximized, in order to uncover the underlying dimensions best able to describe the central oppositions in the data. As in factor analysis or principal component analysis, the first axis is the most important dimension, the second axis the second most important, and so on, in terms of the amount of variance accounted for. The number of axes to be retained for analysis is determined by calculating modified eigenvalues.  Based on scree plot 5 clusters were selected to demonstrate the variability and reliability of the dataset.                Categorical variable dimensional reduction : dimension allocation:   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | |  | Dim 1 | Dim 2 | Dim 3 | Dim 4 | Dim 5 | | Type | 0.001 | 0.001 | 0.002 | 0.000 | 0.660 | | Delivery.Status | 0.000 | 0.000 | 0.000 | 0.000 | 0.661 | | Category.Id | 0.312 | 0.146 | 0.004 | 0.000 | 0.003 | | Customer.Country | 0.000 | 0.001 | 0.001 | 0.001 | 0.001 | | Customer.Segment | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | | Customer.State | 0.001 | 0.002 | 0.004 | 0.003 | 0.003 | | Department.Id | 0.227 | 0.116 | 0.001 | 0.000 | 0.002 | | Market | 0.903 | 0.948 | 0.998 | 0.999 | 0.001 | | Order.Country | 0.916 | 0.950 | 0.998 | 0.999 | 0.022 | | Order.Region | 0.914 | 0.950 | 0.998 | 0.999 | 0.016 | | Order.Status | 0.001 | 0.001 | 0.002 | 0.001 | 0.984 | | Shipping.Mode | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |   Dimensions based on allocations were named as follows:  Dimension 1= Product Category  Dimension 2= Destination location  Dimension 3= D3 category and chain logistics  Dimension 4= Target Market  Dimension 5= Status of order and type of transaction  Similarly, for numerical variables, after correcting the skewness the PCA was applied and to further demarcate the variables they were rotated and dimensionality was reduced.  Unrotated scores:    Rotated scores:     |  | RC2  <S3: AsIs> | RC1  <S3: AsIs> | RC4  <S3: AsIs> | RC3  <S3: AsIs> | RC5  <S3: AsIs> | h2  <dbl> | u2  <dbl> | com  <dbl> | | --- | --- | --- | --- | --- | --- | --- | --- | --- | | Benefit.per.order | -0.0023 | 0.0691 | 0.0039 | -0.0010 | 0.9973 | 0.9994153 | 0.0005847042 | 1.009636 | | Sales.per.customer | 0.0006 | 0.9297 | 0.0345 | 0.0015 | 0.0832 | 0.8723687 | 0.1276312775 | 1.018809 | | Latitude | -0.0030 | 0.0002 | -0.0032 | -0.8732 | -0.0021 | 0.7625673 | 0.2374327461 | 1.000061 | | Longitude | 0.0027 | 0.0004 | -0.0008 | 0.8732 | -0.0038 | 0.7625728 | 0.2374272184 | 1.000058 | | Order.Item.Discount | 0.0003 | 0.3696 | 0.8972 | 0.0008 | 0.0273 | 0.9422421 | 0.0577578790 | 1.332000 | | Order.Item.Discount.Rate | 0.0009 | -0.1475 | 0.9677 | 0.0012 | -0.0159 | 0.9583981 | 0.0416019196 | 1.047010 | | Order.Item.Product.Price | 0.0015 | 0.8994 | 0.0920 | -0.0015 | 0.0235 | 0.8179385 | 0.1820614559 | 1.022325 | | order.lead.time | 0.9725 | -0.0002 | 0.0007 | 0.0021 | -0.0017 | 0.9456947 | 0.0543053365 | 1.000017 | | Delivery.lead.time | 0.9725 | 0.0019 | 0.0004 | 0.0024 | -0.0014 | 0.9456948 | 0.0543052046 | 1.000024 |   \*\*PCA Technical information for better understanding of the dimensionality reduction: for ref  Principal Components Analysis  Call: principal(r = mydata\_normalised, nfactors = 5, rotate = "Varimax")  Standardized loadings (pattern matrix) based upon correlation matrix  RC2 RC1 RC4 RC3 RC5  SS loadings 1.8914 1.8363 1.7510 1.5251 1.0031  Proportion Var 0.2102 0.2040 0.1946 0.1695 0.1115  Cumulative Var 0.2102 0.4142 0.6087 0.7782 0.8897  Proportion Explained 0.2362 0.2293 0.2187 0.1905 0.1253  Cumulative Proportion 0.2362 0.4656 0.6842 0.8747 1.0000  Mean item complexity = 1  Test of the hypothesis that 5 components are sufficient.  The root mean square of the residuals (RMSR) is 0.05  with the empirical chi square 32445.84 with prob < 0  Fit based upon off diagonal values = 0.9637  With factor =5  Naming the Rotated factors / clusters based on the variables:  2,1,4,3,5  RC1 = Sales.data  RC2 = Lead.time  RC3 = Location. Destination of order  RC4 = Discounts  RC5 = Benefits (profit / loss)  Now various modelling techniques were performed and the model performance measures were considered to find out the best fit model. The basis on which the modelling techniques were weighed was based on the Confusion matrix.  Since the target variable is a factor variable with output 1 and 0. So its also called a classifier.  A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known.  The number of correct and incorrect predictions are summarized with count values and broken down by each class. This is the key to the confusion matrix.  The confusion matrix shows the ways in which your classification model  is confused when it makes predictions.  It gives you insight not only into the errors being made by your classifier but more importantly the types of errors that are being made.  Demystifying the Confusion Matrix Using a Business Example  • the accuracy: (TP+TN)/(TP+TN+FP+FN)  • the sensitivity (recall, TP rate): TP/(TP+FN)  • the specificity: TN/(TN+FP)  • positive predictive value (precision): TP/(TP+FP)  • negative predictive value: TN/(TN+FN)  • FP rate (fall-out): FP/(FP+TN)  First modelling technique: Logistic regression  The modelling is done on the normally distributed data, hence the scores obtained from PCA and MCA were considered for modelling the data  Odds ratio  (Intercept) Product.Category Destination.location D3 Market  1.2197391 1.0385308 1.0567618 1.0518296 0.9875753  Status Lead.time Sales.data Discount Location.destination  0.3903917 1.6339443 0.9877800 1.0006404 1.0021178  Benefits  0.9941997  Probability ratio  (Intercept) Product.Category Destination.location D3 Market  0.5494966 0.5094506 0.5137988 0.5126301 0.4968744  Status Lead.time Sales.data Discount Location.destination  0.2807782 0.6203412 0.4969262 0.5001601 0.5005289  Benefits  0.4985457  VIF: Variable importance factor  Overall  Product.Category 3.4499081  Destination.location 4.8956191  D3 4.4304092  Market 1.0938538  Status 65.3869312  Lead.time 83.4003709  Sales.data 2.1426740  Discount 0.1120319  Location.destination 0.3695832  Benefits 1.0175322  So the importance of variable is from highest VIF to lowest.  Most important identifier of late delivery is:   * Lead time * Status of the delivery * Destination.Location * Category of product ordered * D3 / Supply chain Logistics   These variables need to be critically monitored and would play a major role in determining that whether the delivery would be delayed or in time.  After calibrating and tuning the model:  Calibrating and tuning to increase accuracy  Confusion Matrix and Statistics  Reference  Prediction 0 1  0 9456 6202  1 51699 68030    Accuracy : 0.5723  95% CI : (0.5697, 0.575)  No Information Rate : 0.5483  P-Value [Acc > NIR] : < 2.2e-16    Kappa : 0.0761    Mcnemar's Test P-Value : < 2.2e-16    Sensitivity : 0.9165  Specificity : 0.1546  Pos Pred Value : 0.5682  Neg Pred Value : 0.6039  Prevalence : 0.5483  Detection Rate : 0.5025  Detection Prevalence : 0.8843  Balanced Accuracy : 0.5355    'Positive' Class : 1  Test data with same tuning  Confusion Matrix and Statistics  Reference  Prediction 0 1  0 3213 2083  1 17172 22661    Accuracy : 0.5733  95% CI : (0.5688, 0.5779)  No Information Rate : 0.5483  P-Value [Acc > NIR] : < 2.2e-16    Kappa : 0.0786    Mcnemar's Test P-Value : < 2.2e-16    Sensitivity : 0.9158  Specificity : 0.1576  Pos Pred Value : 0.5689  Neg Pred Value : 0.6067  Prevalence : 0.5483  Detection Rate : 0.5021  Detection Prevalence : 0.8826  Balanced Accuracy : 0.5367      AUC area = 0.6595126  Ks = 0.3728878  Gini= 0.3190252  Second Modelling technique:  KNN  Output of modelling:  K=3  Confusion Matrix and Statistics  Reference  Prediction 0 1  0 17211 3174  1 4445 20299    Accuracy : 0.8312  95% CI : (0.8277, 0.8346)  No Information Rate : 0.5201  P-Value [Acc > NIR] : < 2.2e-16    Kappa : 0.661    Mcnemar's Test P-Value : < 2.2e-16  K=4  Confusion Matrix and Statistics  Reference  Prediction 0 1  0 17007 3378  1 4588 20156    Accuracy : 0.8235  95% CI : (0.8199, 0.827)  No Information Rate : 0.5215  P-Value [Acc > NIR] : < 2.2e-16    Kappa : 0.6455    Mcnemar's Test P-Value : < 2.2e-16    Sensitivity : 0.7875  Specificity : 0.8565  Pos Pred Value : 0.8343  Neg Pred Value : 0.8146  Prevalence : 0.4785  Detection Rate : 0.3769  Detection Prevalence : 0.4517  Balanced Accuracy : 0.8220    K=5  Confusion Matrix and Statistics  Reference  Prediction 0 1  0 17486 2899  1 4684 20060    Accuracy : 0.832  95% CI : (0.8285, 0.8354)  No Information Rate : 0.5087  P-Value [Acc > NIR] : < 2.2e-16    Kappa : 0.6634    Mcnemar's Test P-Value : < 2.2e-16    Sensitivity : 0.7887  Specificity : 0.8737  Pos Pred Value : 0.8578  Neg Pred Value : 0.8107  Prevalence : 0.4913  Detection Rate : 0.3875  Detection Prevalence : 0.4517  Balanced Accuracy : 0.8312    'K' in KNN is the number of nearest neighbours used to classify or (predict in case of continuous variable/regression) a test sample. So, the value of k indicates the number of training samples that are needed to classify the test sample. The value of k is non-parametric and a general rule of thumb in choosing the value of k is k = sqrt(N)/2, where N stands for the number of samples in your training dataset. KNN algorithm is one of the simplest classification algorithms and it is one of the most used learning algorithms. KNN is a non-parametric, lazy learning algorithm. Its purpose is to use a database in which the data points are separated into several classes to predict the classification of a new sample point.  K=6  Confusion Matrix and Statistics  Reference  Prediction 0 1  0 17326 3059  1 4694 20050    Accuracy : 0.8282  95% CI : (0.8247, 0.8317)  No Information Rate : 0.5121  P-Value [Acc > NIR] : < 2.2e-16    Kappa : 0.6556    Mcnemar's Test P-Value : < 2.2e-16    Sensitivity : 0.7868  Specificity : 0.8676  Pos Pred Value : 0.8499  Neg Pred Value : 0.8103  Prevalence : 0.4879  Detection Rate : 0.3839  Detection Prevalence : 0.4517  Balanced Accuracy : 0.8272      Ks.train= 0.2911009  AUC.train.area= 0.6465595  Gini = 0.293119  Third modelling technique: Decision Tree  Decision tree modelling  With a Collinearity, removing a column does not affect results. Finally, since these issues affect the interpretability of the models, or the ability to make inferences based on the results, we can safely say that a multicollinearity or collinearity will not affect the results of predictions from decision trees. Luckily, decision trees and boosted trees algorithms are immune to multicollinearity by nature  We can use the categorical and numerical variables without binning or normalizing the numerical variable.  Output of the decision tree:      Pruning the tree and calibrating the tree input parameters    train\_CM  0 1  0 59363 0  1 0 71921  Confusion Matrix and Statistics    0 1  0 59363 0  1 0 71921    Accuracy : 1  95% CI : (1, 1)  No Information Rate : 0.5478  P-Value [Acc > NIR] : < 2.2e-16    Kappa : 1    Mcnemar's Test P-Value : NA    Sensitivity : 1.0000  Specificity : 1.0000  Pos Pred Value : 1.0000  Neg Pred Value : 1.0000  Prevalence : 0.4522  Detection Rate : 0.4522  Detection Prevalence : 0.4522  Balanced Accuracy : 1.0000    CART\_test\_CM    0 1  0 22176 1  1 0 27055  Confusion Matrix and Statistics    0 1  0 22176 1  1 0 27055    Accuracy : 1  95% CI : (0.9999, 1)  No Information Rate : 0.5496  P-Value [Acc > NIR] : <2e-16    Kappa : 1    Mcnemar's Test P-Value : 1    Sensitivity : 1.0000  Specificity : 1.0000  Pos Pred Value : 1.0000  Neg Pred Value : 1.0000  Prevalence : 0.4504  Detection Rate : 0.4504  Detection Prevalence : 0.4505  Balanced Accuracy : 1.0000    Fourth Modelling Technique: NB (Naïve Bayes)  Naïve bayes  pred\_nb  0 1  0 18204 42951  1 8754 65478  Confusion Matrix and Statistics  pred\_nb  0 1  0 18204 42951  1 8754 65478    Accuracy : 0.6181  95% CI : (0.6155, 0.6207)  No Information Rate : 0.8009  P-Value [Acc > NIR] : 1    Kappa : 0.1891    Mcnemar's Test P-Value : <2e-16    Sensitivity : 0.6753  Specificity : 0.6039  Pos Pred Value : 0.2977  Neg Pred Value : 0.8821  Prevalence : 0.1991  Detection Rate : 0.1345  Detection Prevalence : 0.4517  Balanced Accuracy : 0.6396    Fifth Modelling Technique: Random Forest  Confusion Matrix and Statistics    0 1  0 56873 5  1 2 69476    Accuracy : 0.9999  95% CI : (0.9999, 1)  No Information Rate : 0.5499  P-Value [Acc > NIR] : <2e-16    Kappa : 0.9999    Mcnemar's Test P-Value : 0.4497    Sensitivity : 1.0000  Specificity : 0.9999  Pos Pred Value : 0.9999  Neg Pred Value : 1.0000  Prevalence : 0.4501  Detection Rate : 0.4501  Detection Prevalence : 0.4501  Balanced Accuracy : 0.9999  RF CM Test  Confusion Matrix and Statistics  0 1  0 24660 0  1 0 29495    Accuracy : 1  95% CI : (0.9999, 1)  No Information Rate : 0.5446  P-Value [Acc > NIR] : < 2.2e-16    Kappa : 1    Mcnemar's Test P-Value : NA    Sensitivity : 1.0000  Specificity : 1.0000  Pos Pred Value : 1.0000  Neg Pred Value : 1.0000  Prevalence : 0.4554  Detection Rate : 0.4554  Detection Prevalence : 0.4554  Balanced Accuracy : 1.0000    Variable importance based on RF  Delivery.lead.time order.lead.time Shipping.Mode  51283.070406 38719.225817 28125.756658  Order.City Order.Status Order.State  9805.361260 5712.398218 3286.564251  Order.Country Customer.State Longitude  568.956976 21.230040 6.839369  Final Inference from the analysis:  Comparative data:  Key observations  As per the above comparative data obtained of various key parameters considered for comparison, the Decision tree methods are giving a better model performance.  KNN is also better than Naïve Bayes  Other Key Observations from ETA have been mentioned alongwith the graphical representations.  F1 is not considered here since it is more useful with uneven distribution of the target variables.  Since it is the weighted average of the precision and recall absed on classifier distribution  We have gathered information regarding the important variables which will help in identifying the late delivery and therefore controlling them can help reduce the late deliveries.  Based on the model comparison Decision tree comes out to be the best predictor and modelling technique. | | | | | |
| Prediction Summary | | | | |  |
| Techniques | | Accuracy | Sensitivity | Specificity | Recall |
| Logistic Regression | | 57.23% | 91.65% | 15.46% | 56.82% |
| KNN | | 83.2% | 78.87% | 87.37% | 85.78 |
| Decision tree | | 100% | 100% | 100% | 100% |
| Naïve Bayes | | 61.81% | 67.53% | 60.39% | 29.77% |
| Random Forest | | 99% | 99.98% | 99.25% | 99% |
|  | Thanks | | | | | |

References:

EDA

<https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/>

<https://www.researchgate.net/publication/334794986_Heuristic_dynamic_approach_to_perishable_products_in_presence_of_deterioration_effect>

<https://www.researchgate.net/publication/281904707_Using_big_Data_in_the_Supply_Chain_Context_Opportunities_and_Challenges>

<https://www.mckinsey.com/business-functions/operations/our-insights/big-data-and-the-supply-chain-the-big-supply-chain-analytics-landscape-part-1>

for creating dummy variables

<https://www.analyticsvidhya.com/blog/2015/11/easy-methods-deal-categorical-variables-predictive-modeling/>

MCA

<http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/114-mca-multiple-correspondence-analysis-in-r-essentials/#:~:text=The%20Multiple%20correspondence%20analysis%20(MCA,more%20than%20two%20categorical%20variables.>

Model evaluation

<https://lgatto.github.io/IntroMachineLearningWithR/supervised-learning.html>

odds and probability ratio

<https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Confidence_Intervals/BS704_Confidence_Intervals10.html>