**Initial data analysis for classification problem using R.**

1. To understand whether a variable is numeric, categorical or discrete variable. Also, to understand level of factors in categorical variable.

str(tablename)

you will get value as follows:

$ Year : int 2014 2014 2014 2014 2014 2014 2014 2014 2014 2014 ...

$ Month\_End\_date : Factor w/ 12 levels "2014-01-31","2014-02-28",..: 1 1 1 1 1 1 1 1 1 1 ...

$ MonthStart : Factor w/ 12 levels "2014-01-01","2014-02-01",..: 1 1 1 1 1 1 1 1 1 1 ...

$ loc\_cd : Factor w/ 2327 levels "00021","00022",..: 362 215 215 286 362 58 1049 123 220 189 ...

$ Revenue : num 680 211 73246 -535 0 ...

This will give a basic idea about type of variables, level of factors, and possible values for numeric and discrete variable.

1. To understand central tendency such as mean, median, dispersion such as min, 1q, 3q, and max, and missing values.

For categorical variable, result will be as follows:

summary(table\_2016$Known\_CompanyGrowthClues\_Flag)

N Y NA's

422481 619588 275666

For numeric variable, result will be as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| summary(table\_2016$LastRental\_LocLevel\_CSIScore)  Min. 1st Qu. Median Mean 3rd Qu. Max. NA's  10.0 86.2 96.9 88.9 100.0 100.0 1251470  This will trigger further investigation to understand importance of a predictor from business prospective to see if missing values can imputed.  If a variable has missing values beyond certain % then that variable should be ignored because missing value imputation can lead to bias in sample. % varies from business to business and depends on how many observations are present in sample and whether characteristics of sample changes after removal of observation that has missing values.   1. To understand basic relation between target variable and predictors.   First, we need to create contingency table with target variable and predictor.   |  | | --- | | tbl1 <- table(table\_2014$NewProductLine,table\_2014$Known\_CompanyGrowthClues\_Flag, exclude = NULL)    N Y <NA>  Target valu1 10189 28393 10726  Target valu2 17576 34168 13850  Target valu3 27546 45169 21765  Target valu4 5345 10663 6689  Target valu5 18128 29337 14419  Target valu6 671 2420 2445  Target valu6 82546 136462 66482  Target valu6 106452 157905 76279  Target valu7 212 481 298  Target valu8 43 94 12  Target valu9 152336 254054 179724  Target valu10 687 2811 4661  Target valu11 113 130 11054  <NA> 0 0 0  This analysis will tell about importance of predictors and whether it can be ignored.  In summary, This analysis will give not only give basic idea of overall dataset but also insight about each variables and importance of predictors. In addition, it will bulid the foundation for further analysis and help in shaping solution design. | |  | | |  | | --- | |  | | |
|  |
| |  | | --- | |  | |