We'll be using processed data from data prep practice assignment for this practice assignment. Our goal here is to build a predictive model for predicting our target Y with given characteristics. I have laid out steps to carry out operations. Few of these questions/steps will also include further details as and when required.

Split The Data : Train & Test

Split the data into two parts so that train contains 75% of the data and test contains rest 25%. Use seed `123`.

Note: We are not using validation in this exercise to save some effort. You should ideally use this to build a model which is better at generalising.

Remove Multi-collinearity

Run a simple linear regression model. Pass the model object to function `vif` [which is found in package `car`]. Remove variables which have VIF values higher than 10. [Note : DO NOT USE P VALUES FROM THIS TO DROP VARIABLES. WE ARE ONLY CONCERNED WITH VIF VALUES FROM HERE. If in doubt, take to QA forum for further discussion]

Note : You might get error relating to aliased coefficient. This can happen for couple of reasons

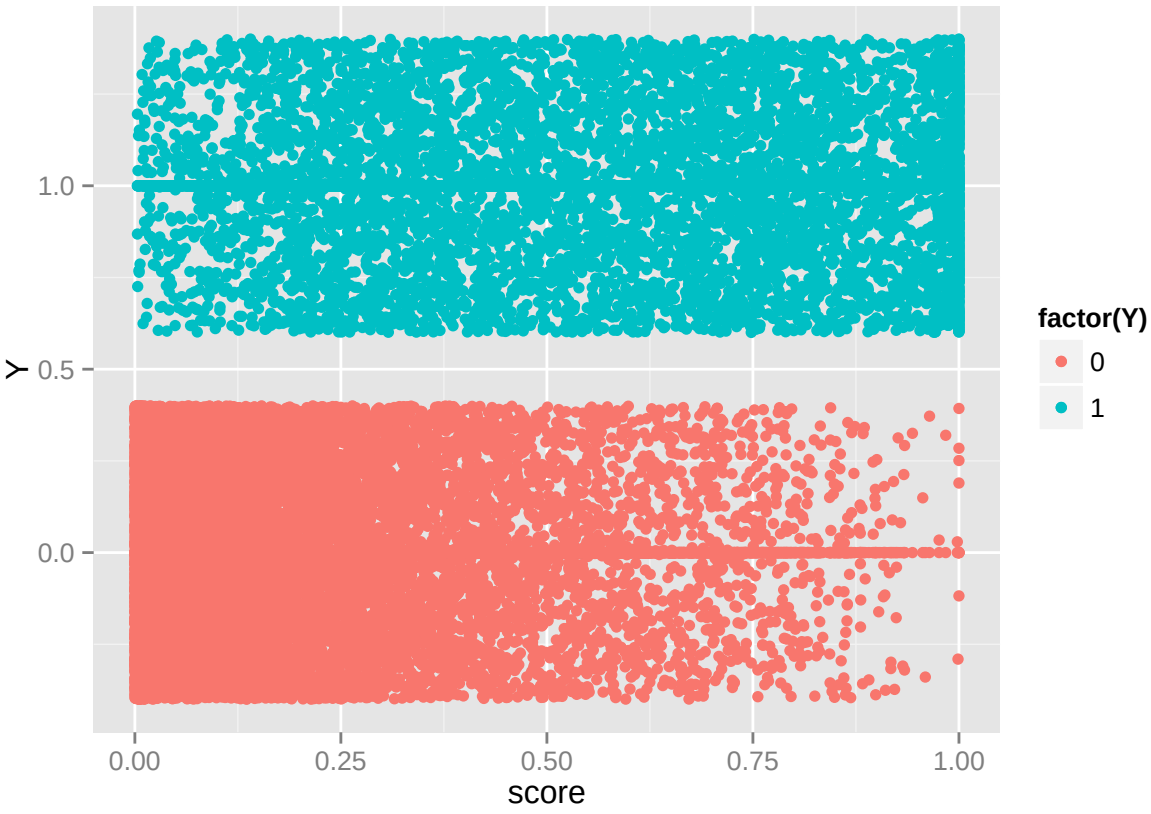
1. Some variable has a constant value for entire data
2. Some categorical variable has been left in the data along with the dummy variables which you created from it. This leads to duplication of data
3. One or more variable columns are identical

To check which variables are having these issues, do `summary(model_object)`. Where ever you see `NA` against a variable name [instead of coefficient values]; those are the variables with issues. Check if you made some mistake in data prep while creating those variables.

Build a logistic regression model

Once you are done with removing variables based on vif , build a logistic regression model. Pass it through step function to drop variables which do not meaningfully contribute to your model.Post that , save the probability score to train.

plot probability scores with outcome to see if score has been able to bring some differentiation. Your plot would be similar to this.



Performance Metrics for A particular Cutoff

A cutoff can be decided with multiple considerations or business requirement. All of these requirements are generally based on confusion matrix. Confusion matrix is nothing but cross table of real 1/0[response] against predicted 1/0 [predicted response].

Consider an arbitrary cutoff 0.3. Get predicted response for this cutoff. Using the predicted response column, calculate Following metrics . [Consider 1= Positive , 0=Negative]

- TP (True Positive), FP (False Positive), TN (True Negative), FN(False Negative)
- Accuracy { Defined as (TP+TN)/(P+N), where P= TP+FN and N=TN+FP}
- S_n (Sensitivity) { Defined as TP/P}
- S_p (Sepecificity) {Defined as TN/N}
- Dist { Defined as $\sqrt{(1-S_n)^2+(1-S_p)^2}$}
- KS { Defined as $\frac{TP}{P}-\frac{FP}{N}$}
- M { A hypothetical metric defined as $\frac{9*FN+0.6*FP}{1.9*(P+N)}$}

Getting optimal cutoff for each metric

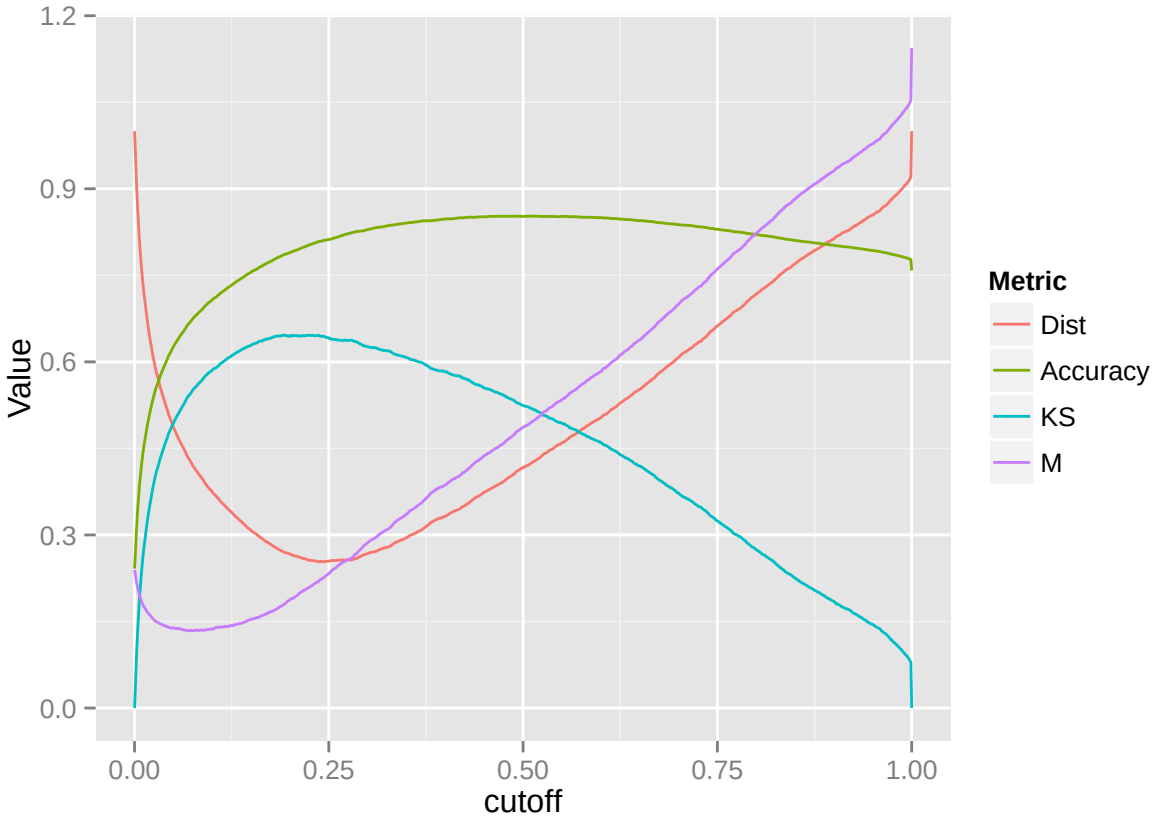
Consider 1000 cutoffs between 0 to 1 [Equally spaced , not random]. { Hint : Use function seq to generate these}. Using a for loop calculate Dist, Accuracy, KS and M for all these cutoffs. Find cutoffs for which

- Dist is minimum
- Accuracy is maximum

- KS is maximum
- M is minimum

Note: optimal cutoffs for all four condition will not be same [if they are , it will be a coincidence]

Also plot these values in a single plot to show how these metrics vary across cutoff range. Your plot will look like these



Performance on Test Data

Get scores for test data also . Apply respective cutoffs and evaluate the related metrics as well for each of those cutoffs.