Data Understanding

In this phase we would try to understand the data collected from our internal clients

i.e the marketers of the company like what are the different sources of data,the size of

the data set,the frequency at which the data is generated per interval like a hour day or

month ,whether the data is structured or unstructured.

Gain an understanding of customer related attributes like age,id,name,transaction

history,number of items purchased in a specific interval and type of customer whether

he is loyal,need based,discount ,wandering,high returned customers and other

personal data.Data needs to be explored and verified for quality,check if data is

complete.

The problems we may face while understanding data are data having missing

value,noise in the data,data not available for some customers or attributes.

Data Preparation

This phase is the most important and time consuming of the life cycle.After gaining

understanding of data we can handle the problems by either use a mean value for the

missing values,clean the data and handle noise by smoothening bin by mean or

median.We have to construct and integrate the data from multiple sources.For

example if we have information about the customer general characteristics and the

customer account details in separate tables so aggregation can be done here.Data

may also need to be formatted for example removing commas from within text fields

in comma separated data files.

Also we might do dimensionality reduction as a part of data reduction ,the algorithm

we can use here is PCA(Principal Component Analysis) so that we can get a reduced

representation of the dataset that is smaller in volume but still it preserves the

structure and information with the data.So we can remove unimportant attributes

like zip code from the customer data.In this phase ETL is perfomed for the data.Data

can be visualized using tools like R,GnuPlot,Tableau.

If new problems are encountered and solutions are not available then it can mean that

we need better understanding of data and can return back to the data understanding

phase.

Data Modelling

In this phase we can select modelling techniques,generate test design and build

predictive model for the loyal,need based,discount ,wandering,high returned

customers .

Labelled Historical data of customers and the present data of the customer

transactions is total data that we have.A predictive model can be generated that can

predict the type of customer using this data and based on the type of customer

marketers in the company can develop appropriate strategy.For example for loyal

customers the regular products can be shown.We separate the Labelled

historical data into train and test sets,build the model on test set.

Historical data is used to train the model to arrive at the most suitable modelling

technique.We can use modelling techniques like cluster analysis to find the pattern

of transaction details in the historical data for similar type of customers like if the

customers are buying lot of discounted product then prediction can be made about

discount customers .So if other customer buy those products they can also put in the

same type and marketing strategy can be developed . Modelling techniques might

require data in a specific format so it might also be possible that we need to go back

to data preparation phase.

Evaluation

This phase consists of evaluating the results with respect to the business success

criteria of the Project.Classifier will be run on the test data taken from

the labelled historical data and accuracy will be checked.

After it is run the results will be compared to the business objectives and

the results will be shown to internal clients.For example are we able to predict the

customer type for any customer by looking at his current transactions.We can use

cross validation here to find the subset of attributes of a customer that can help us

build the best predictive model. If we have customer data which is grouped by

observations, tools such as cluster analysis, association rules, and k-nearest

neighbors usually provide the best results.

Deployment

Creation of the model is generally not the end of the project. Even if the purpose of

the model is to increase knowledge of the data, the knowledge gained will need to

be organized and presented in a way that is useful to the customer.Based on the

evaluation results we need to develop deployment plan and document it.

Also develop a monitoring and maintenance plan to check that the model

performance does not deteriorate.

Prepare a final report which can be summary of the project or final and

comprehensive result of the project.