Everything on DM Project

Start with data preparation and remove rows containing :

1) Missing values
2) Data entry errors
3) Minority values
4) Flat and Wide variables
5) Outliers

Also recode any data entry inconsistencies

Proceed with training a model using around 70% of data for training and the rest for testing. Lastly, use 10 fold cross validation to verify the results

Cross Validation (10-Fold)

Very accurate and it reduces the risk of a lucky test set

Splits the data into 10 subsets. Trains 10 models each using 9 of the 10 subsets as training data and the 10th for testing. The result score is the average of all 10 models

The success of a model can be assessed by the :

1) Mean Squared Error - MSE
2) Percentage of correct classifications (Only for classifications)
3) Confusion Matrix (Only for classifications)
4) Correlation Coefficient (Only for prediction)

ROC Curves

An ROC curve displays how many false positives and true positives you get for each possible threshold. The threshold for a positive is varied from 0 to 1

The idea is to optimise the trade-off between finding as many positives as possible, while wrongly including as few negatives as possible