Pilot Study Proposal

Email marketing campaigns are an important way of reaching out to potential customers and it is very likely that machine learning techniques can be helpful in increasing the effectiveness of such campaigns. Given that there is data available about the success/failure of previous emails sent to customers, this is a supervised learning task in which a classifier can be trained using the result of previous campaigns and their outcomes.

Choosing the right information on which to train our classifier is likely to have a big impact on its effectiveness. We would need as much relevant information about each email as possible from the following:

* Recipient features (personal customers): Age, gender, occupation, purchase history, marital status etc.
* Recipient features (business customers): Type of business, turnover, staff strength, purchase history etc.
* Product features: Type, price, size, discount on usual price, colour etc.
* Email features: Words, style, design

Given that this is a supervised classification problem, we could use the following algorithms:

* Decision Tree: A big advantage of decision trees is that rules generated by decision trees are easily comprehensible, which makes them a very useful tool for exploration of a problem. A potential problem with decision trees is that they can over-fit the data, hence care needs to be taken by employing appropriate pruning methods where needed.
* Naïve Bayes: Although Naïve Bayes make a strong assumption of conditional independence in attributes given the classification variable, they can be very powerful in certain situations. The main advantage of Naïve Bayes is their computational efficiency and the fact that they need much less data then would be needed for a complete Bayes classifier.
* K-nearest neighbour (KNN): KNN can be used for the given situation as it assigns patterns to the majority class amongst K nearest neighbours.
* Support Vector Machine (SVM): SVMs can overcome a major limitation of some other techniques such as KNN and Decision Trees that these techniques do not give good results for problems which are not linearly separable. SVMs can efficiently classify non-linearly separable data with use of appropriate kernels such as a polynomial or Gaussian kernel.