# To Mail or Not to Mail

Direct mailings to a company's potential customers – "junk mail" to many – can be a very effective way for them to market a product or a service. However, as we all know, much of this junk mail is really of no interest to the people that receive it. Most of it ends up thrown away, not only wasting the money that the company spent on it, but also filling up landfill waste sites or needing to be recycled.

If the company had a better understanding of who their potential customers were, they would know more accurately who to send it to, so some of this waste and expense could be reduced.

## Data Files

Train Dataset = carvan_train.csv

Test Dataset = carvan_test.csv

## Formal Problem Statement

We want you to predict whether a customer is interested in a caravan insurance policy from other data about the customer. Information about customers consists of 86 variables and includes product usage data and socio-demographic data derived from zip area codes. The data was supplied based on a real world business problem. The training set contains over 5000 descriptions of customers, including the information of whether or not they have a caravan insurance policy. A test set contains 4000 customers of whom target variable is not shared with you.

Target Variable is V86.

## Submission

You need to use train data for building the model and then use that model to predict outcome for given test data. We expect outcomes to be either 0 or 1. [You'll need to find cutoff for converting your probability predictions to hard classes ]

In order to get a passing grade in this project you need to get Fbeta score greater than 0.26 [ beta =2 ] for your test data predictions .

## General Guidelines for the project

- Your submission will be a csv file with a single column containing your predictions for target. Order of these predicitons should be same as observation orders in the test data to which these predictions correspond to.
- you will find data details in 'data dictionary.txt' file
- You will notice that many variables which are numeric in the data but should have been categorical in reality. Handling those variables in proper fashion might improve your model
- Real catch in this problem is very low number of responses being 1. Simpler models might

not perform very well on this. You will have to focus on parameter tuning very well. Since the dataset is fairly small, it wouldnt be an issue.