# 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.

You need to use train data for building the model and then use that model to predict outcome for given test data. Test dataset does not have a response column; you need to predict those values and submit it in a csv format. We expect outcomes to be either 0 or 1.

## Evaluation Criterion

**Part 1:**

You will first attempt Part 1 of this project which is a quiz. You can access it through LMS. This quiz needs to be answered based on exploration of the dataset given and some generic questions about algorithms discussed in the course. Consider only the training dataset for data cleaning and exploration to answer the quiz questions. There will be 10 questions of which you need to get at least 7 correct in order to pass the project.

**Part 2:**

Here you work on creating the machine learning models and choosing the one which gives the best performance. You can refer to the Project Process Guides provided in LMS to understand how to approach and work on a project.

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.

**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. Your submission will be a csv file with a single column containing your predictions for target. Order of these predicitons should be same as order of the observations in the test data to which these predictions correspond.

You can make as many submissions you want if you want. [We might ask you to submit the script which was used to generate the submission at any time].

## General Guidelines for the project

- Since its a small dataset and you can quickly run many experiments, we are not providing any benchmark script for you to get started.
- One more reason for not providing a benchmark script is that, entire data is conveniently numeric and you need to spend very less time in preparing the data.
- 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 will perform very poorly on this data. You will have to focus on parameter tuning very well. Since the dataset is fairly small, it wouldnt be an issue.
- As mentioned in the project 1, do break your train data into two parts; use one part to build your model and use another to asses its performance, so that while submitting your results, you know how your model performs rather than wait for our evaluations.
- While you are breaking your data into two parts, make sure that you stratified sampling so that both part have same percentages of 0/1 as in the original data. This way you'll avoid falling in trap of severe over/underfit while assesing performance of your model.
- In case of any doubt , feel free to reach out to us.

In order to clear this project, you are required to clear both, Part 1 as well as Part 2 of this assignment.

Wish you all the best!