How To Identify Potential Donors

Machine Learning Project Proposal

Riccardo Paciello, Darryl Abraham

April 1, 2024

1 The Problem

The problem we have chosen is a regression problem. The goal is to build a model to accurately predict what type of people are likely to donate and how much in response to a direct mailing campaign. The problem focuses on a dataset from the 1998 KDD Cup Data Mining Competition, which features a real data collected by Paralyzed Veterans of America. This dataset contains a record for every donor that has received the donation mail, but has not donated in the past 12 months (only includes people that have donated previously). For each person it is given whether they have donated, a binary classification task, and how much, a regression task. For each person they are described by characteristics ranging from geographical to socioeconomic indicators. Each of these models can be used to predict whether a certain individual is likely to make a donation.

2 Motivation

We believe that this is a interesting problem to investigate as further understanding donor behaviour and identifying the key predictors to donations can help optimise resource allocation, improve campaign effectiveness, develop better and more targeted fundraising strategies, and enhance donor engagement for non-profit organisations that help the most marginalised members of our community. Additionally, from a purely research perspective it can help us glean insights on the societal outlook on humanitarian campaigns, in terms of understanding human nature and how a communal identity plays a role in generosity.

3 Dataset

The dataset contains 191779 records, of which 95412 are training cases and 96367 test cases. The split has already been made by the competition organizers (however, this can be reversed, or simply only the training dataset taken to be the whole for the purposes of this project). The data set contains 481 variables. There is a wide range of different numerical and categorical variables to choose from. We understand that this dataset is much larger in terms of features than expected, however we believe we can narrow it down with a rigorous feature selection process. Additionally, the training set is very large but only contains 5% positive records. Thus, we believe that a lot of our solution and modelling will be defined by a robust feature selection, and case selection process. In this way we can size down the dataset drastically to make the problem more manageable.

Link to dataset: https://kdd.ics.uci.edu/databases/kddcup98/kddcup98.html
Link to feature dictionary: https://kdd.ics.uci.edu/databases/kddcup98/epsilon_mirror/cup98dic.txt

References

[1] Chris Fleizach and Satoru Fukushima. "A naive bayes classifier on 1998 kdd cup". In: *Dept. Comput. Sci. Eng., University of California, Los Angeles, CA, USA, Tech. Rep* (1998).