Minh D. Vo

Non-Technical Report

1. Problem description:

As a media company, helping local small and medium sized business through digital advertising is one of our main revenue streams. The advertising system is operated on its platforms such as Google, Facebook … The objective is to predict if a client will stop running advertising campaigns (churn).

2. Dataset:

The data (907 KB) in csv file includes 10,000 observations which cover some information (10 variables). They include cost per lead with respect to business category, client location, duration of advertising campaign, number of distinct advertising products, average monthly budge … The description and name of each variables are provided in two text files (dict.txt and header.txt).

3. Approach

I have treated this study as a supervised learning classification problem. First, I visualized data and explored any abnormal values in dataset. There are some missing values. The appropriate assumption was considered to fill these missing values.

Secondly, all data were scaled and converted into numerical values for further analysis. I don't use all data for building model. Only 80% of data were used for training the model. The other data (20%) were kept for evaluating the model. The target variable is churn. It can have only two values (1 for churn – positive class and 0 for retention – negative class)

Different classification models were investigated and tuned their parameters.

In this problem, the sensitivity and precision were selected as a criterion for model evaluation. For sensitivity term, it should to answer the question of how often the prediction is correct when the actual value is positive. For precision aspect, it can describe how 'precise' the model is when predicting positive instances.

Finally, the adjustment of prediction is also considered for improving the performance of the model. The objective is to maximize the sensitivity and precision.

4. Results

From this dataset, I can draw some findings:

- 80% of clients are retention with advertising program (target variable is 0). The change in client's cost per lead in the past three months (CPL_wrt_self) is often higher than change in cost per lead with respect to business category (CPL_wrt_BC).
- The average time of advertising campaign is around 27.7 months. This could imply that most of clients will consider the effect of advertisement after around 2 years. Clients (nearly 80%) prefer to use only one distinct advertising product.

- Advertisement affects clearly on the number of clicks received more than calls received. Apparently, because this is digital advertisement. Most of customers can see the advertisement when they are using the internet (via google, facebook).
- Average monthly budget spent on advertising campaigns can vary in the large range from 9 to 148555.6 (the mean is 1512.6). In business, all clients always try to reduce this cost.
- More than 26% of clients are from 'Home & Home Improvement' section. Top 5 states with highest clients are CA, TX, FL, NY, PA.

From the model evaluation, it can be seen that:

- Logistic regression has a poor performance.
- Decision tree, random forest and gradient boosting are the better model.
- With a suitable adjustment, gradient boosting becomes the good choice (compared with other model in this study) for this problem.
- In addition, it also presented that duration, CPL_wrt_self, CPL_wrt_BC and avg_budget are important features. They are strongly influenced on the target variable.

5. Recommendation

The performance of the classification model is not optimized yet. I think there are some ways to improve it:

- It should consider handling the abnormal values (extremely high values)
- Different classification models can be tried.
- All the important features can be examined for further correlations/patterns.