BUSINESS ANALYSIS OF CHARITABLE GIVING

Business Analytics Project Description

IST 5420

SP18

Due: 11 May 2018

Original Version: Posted 11 April 2018
Modified: Posted 18 April 2018

MODIFICATIONS APRIL 18, 2018

I've been reconsidering this project. It is worth 200 points, or about twice a "normal" homework problem. Hence, I feel the need to scale it down a bit.

Also, I want you to have all the of the analysis tools needed ASAP.

Hence, I like to make the follow changes:

1. Focus on the binary response variable of "Donate/Don't Donate".
2. You do not need to also build a model for the amount donated.
3. Furthermore, you can limit the focus on explanatory variable by considering only categorical variables.

MODIFICATIONS – BUSINESS QUESTIONS

- Your primary accomplishment with this restricted problem will be to limit the size of the distribution. That is, mail only to those the model predicts will denote. Did you make a significant reduction?
- Keep in mind, the mean of a variable is a perfectly reasonable estimate. If you would like to keep "amount donated" in your model, you might consider calculating the average donation over the restricted pool of potential donors. Then use that number to estimate revenue generated.
- If you want more details, you might consider classifying the amount donated by using a discrete variable: [0,100),[100,50),[500,100), etc. This would create two categorical response variables. This is not required.

IDEA BEHIND THIS PROJECT

It is the purpose of this project is to give you a summary experience in big data business analytics

You should be able to demonstrate an understanding of mathematical modeling, parametric analysis/estimation, basic statistical analysis, visualization, regression analysis, logistic modeling, and applications in R.

Beyond the data science, however, the big data is embedded in business. This requires communication skills, which you will have an opportunity to develop.

CONTENTS

- The "Big Data" Business Problem
  - A "real-world" business situation
  - Big Data overview
  - Fundamental Analysis
  - Measures of Success
- Project Basics
  - Requirements
  - Paper
  - Presentation
  - Point Value
- Strategy for Attack
- Origin of idea & data

6 THE BUSINESS PROBLEM

Charitable Direct Mailings

BUSINESS PROBLEM

In 1997, the Paralyzed Veterans of America (PVA) used a direct mailing to appeal to "lapsed" donors. Here is a brief summary of the numbers. On the Valuation data set:

96367 people were mailed

This cost was $0.68 per mailing

This totaled $65,280.

The total amount of donation: $76,100

Hence, the profit: $10,560.

Q: How can we target a smaller group and yet produce greater profit?

DATA OVERVIEW

- There are two data sets. The “learning” data set used to develop a predictive model and the “valuation” data set that can be used to measure the performance of the predictions against actual outcomes.
- Each set is large. About 480 characteristics of 100,000 donors are recorded.
- (Required) One target variable is binary, whether or not a person will donate.
- (Not Required) The other target variable is continuous. It reports the amount donated.

ANALYTICS OVERVIEW

For each person in the population, we must predict whether he/she will donate (not required: and how much might he/she give.)

Modeling the binary variable is a logistic regression.

(Of course, these are coupled. To be specific, we might include an individual in our reduced population only if the predicted donation exceeds $0.68.)

COMMENTS ON THE MEASURES

- Your main contribution will be in reducing the size of the pool. You don't need to work beyond that.
- If you want to include amount donated in these measures, consider what I mentioned earlier, such as using the mean donation among those within your reduced pool.

Profit Measure

We introduce the follow notation to express the profit formula:

Q_{∞} – The total set of people who might be solicited

N_{∞} – The number of people in the whole set who receive mail. Note: $N_{\infty} = \|Q_{\infty}\|$

n – The nth member of the population

μ - The marketing cost (0.68/person)

A_n - The amount contributed by the nth person

$V(Q_{\infty})$- Value (profit) produced by mailing to Q_{∞}

$$V(Q_{\infty}) = \sum_{n \in Q_{\infty}} (A_n - \mu) = \left(\sum_{n \in Q_{\infty}} A_n \right) - \mu N_{\infty}$$

PREDICTION PROCESS

In this problem, we develop a model to predict the nature of giving a person might have based on his/her characteristics. To do this, we formulate a model and use the learning data set to estimate parameters.

With model in hand, we focus on the validation data set and select individuals whom we predict to give more than the marketing cost ($0.68). (We will call this a "sufficient giver".) This selection process forms a new population $\hat{Q}$ where

and $\hat{Q} \subseteq Q_{\infty}$

$$\hat{N} = \left\| \hat{Q} \right\| \leq N_{\infty}$$

QUALITY OF PREDICTION MODEL

Using the notation of the previous two slides we denote:

$$V(\hat{Q}) = \sum_{n \in \hat{Q}} (A_n - \mu) = \left(\sum_{n \in \hat{Q}} A_n \right) - \mu \hat{N}$$

Recall that in the validation set, the amount donated by an individual is known A_n so $V(\hat{Q})$ tell us exactly how much would have been made if using the subpopulation $\hat{Q}$ rather than

the entire population Q_∞

14

PROJECT BASICS

Summary of Requirements

PROJECT COMPONENTS

- A significant data set will be supplied.
- Analysis of the data will center around particular target variables and with specific questions to be answered
- You will need to show you that you performed a serious investigation. You need to proposed several models, motivated by visualizations and other techniques, and draw conclusions about their applicability. In the end, you will need to offer answers to the business questions and assess the quality of your answer.
- The final deliverable is a pdf document. (You might consider use Rmarkdown.)

OF FUNDAMENTAL IMPORTANCE

What I expect to see from you:

The formation of a model to explore.

The use of R, visualization, and statistics to demonstrate a complete analysis of the proposed model.

Proper testing of the model on the validation data. I very much want you to make it to this point.

Finally, I want a business judgment. Has this quick initial effort established there is merit in the model and it should receive a more complete study with associated expenditures? Or has it demonstrated we should drop this model? (This is a perfectly fine conclusion… and perhaps the most likely…)

POINT VALUE

As per the syllabus, this project is worth 200:

- Quality of Analysis (100): Demonstrate
  - a thorough exploration of the data,
  - maturity in model selection,
  - adequate statistical analysis of the model,
  - Appropriate application to the business questions
- Business Analysis (50): Ultimately, your work is to predict and prescribe action. Make these conclusion clear to non-technical people
- Professional presentation of the work in written form in both maturity of prose and visual (50)

PLAN OF ATTACK

Thoughts on how to proceed

AVAILABLE SUPPORT

There will be more lectures and hw on analysis techniques.

All homework will be due on the last day of the semester, Friday, 4 May 2018. This will give you time to balance your work on the project while learning various modeling, graphical, and statistical techniques.

You are welcome to discuss you analysis with others. (Of course, you must do and submit your own work. I will notice.) Brain-storming with each other is useful.

I am going to have the remaining Friday's of the semester be Open Lab/Work Days. (April 13th, 20th,27th, and May 4th) You can ask questions on any work I have assigned (Reading/HW/Project). If you want to just bounce ideas off of me, that is great. If you want to just absorb.

For the distance students, we are not bound to the Friday time. We can arrange a phone call or skype. Email me to arrange a time and I can send you my person phone number. (This is a "big" offer. I am terrified of the phone and rarely use it. I almost didn't get my wife because she is a phone person and I avoid them like the plague. ☺)

GETTING/KEEPING THE BALL ROLLING…

- The rest of the semester is being self governed. If I were you, would write down a timeline for the completion of the HW and Project milestones.
- I will continue to lecture to supply you a rich collection of resources and problems to push your understanding.
- You will need to do some self-teaching. I cannot possibly lecture on every possible situation you might encountered.
- You need to be aggressive. If you can do all work in a week, great! If you leave everything until you the last minute, you will be crushed.

21

ORIGINS OF IDEA AND DATA

Web Resources

PROJECT ORIGIN - KDD-CUP-98

The Association for Computing Machinery (ACM) has a special interest group (SIG) in Knowledge Discovery and Data Mining (KDD).

They have an annual competition on discovery tools and methods… It is call the KDD-Cup.

The project herein is the 1998 competition

WEB RESOURCES

You can find all of the materials on the KDD-CUP 1998 are found at:

http://www.sigkdd.org/kdd-cup-1998-direct-marketing-profit-optimization

The original competition website is

http://www.kdnuggets.com/meetings/kdd98/kdd-cup-98.html

CREDIT FOR DATA

- The data set was *"provided by the Paralyzed Veterans of America (PVA). PVA is a not-for-profit organization that provides programs and services for US veterans with spinal cord injuries or disease."*
- *"This mailing was dropped in June 1997 to a total of 3.5 million PVA donors. It included a gift "premium" of personalized name & address labels plus an assortment of 10 note cards and envelopes. All of the donors who received this mailing were acquired by PVA through premium-oriented appeals like this."*