Quicken Quickbooks Upgrade Assignment

The purpose of this exercise is to provide you with the opportunity to try your hand at modeling the response to an upsell campaign. The "intuit75.dta" file contains data on 75,000 individuals who were selected at random from the 801,821 who were sent the wave-one mailing.[1] Variable "res1" denotes which of these individuals responded to the mailing. The remaining variables are available to help you predict who will respond to a wave-two mailing.

I have simplified, added, and deleted some of the variables listed in Exhibit 3 of the case.
Please ignore the variable description in exhibit 3 of the case.

Instead, please see the variable description in the "intuit75.dta" file (type "describe"). Here are the variables at your disposal:

id	Customer ID.
sex	1=Male, 2=Female, 3=Unknown.
bizflag	Business Flag. Address contains a Business name (1=yes, 0=no or unknown).
zip	5-Digit ZIP Code (0=unknown, 99999=international ZIPs).
zip_bins	Zip-code bins (20 approx. equal sized bins from lowest to highest zip)
sincepurch	Time (in months) since original (not upgrade) purchase of Quickbooks
version1	Is 1 if customer's current Quickbooks is version 1, 0 if version 2
upgraded	Is 1 if customer upgraded from Quickbooks vs. 1 to vs. 2
owntaxprod	Is 1 if customer purchased tax software, 0 otherwise
last	Time (in months) since last order from Intuit direct in last 36 months
numords	Number of orders from Intuit direct in last 36 months
dollars	Total $ ordered from Intuit in last 36 months
res1	Response to wave 1 mailing (1=responded, 0=did not respond)
training	70/30 split, 1 for training sample, 0 for validation sample

Assignment Instructions:

1. Decide to which of the 22,500 individuals in the validation sample you would choose to mail wave two. Completely specify which individuals should receive the mailing. 1/3 of your grade on this assignment will be based on the profit (not ROI) achieved on your mailing by comparing your prediction with the actual findings from the second mailing (which I know because in reality all consumers who did not respond in wave 1 were mailed in wave 2).

2. In your write-up, please scale the profit estimate derived from your best model's performance in the validation sample to the full set of customers to whom wave 2 would go. To do this you need to know that of the 801,821 customers who were sent the wave-one mailing 38,487 responded in wave 1 and should not be mailed again.

[1] A part of this data is disguised and has been artificially added for teaching purposes. As a result, the specific findings from this case analysis should not be carried over to real situations.

3. For the purposes of this exercise, each mail piece costs $1.41, and the profit from each responder is $60. Thus the response rate needed to break even is 2.35%.
***Please ignore any other number in the case that relates to profits and costs*.**

4. Please note the following statement in the case (page 4): "Usual practice would be to assume a 50 percent drop off in response from wave 1 to wave 2." For your analysis this means that when you decide whom to mail to in wave 2, you should assume that **every individual's** response probability in wave 2 is only 50% of the response probability you predict for that individual based on wave 1 responses.

5. Two thirds of your grade (65 points) for this assignment will be based on your write-up (with at least 1.5 line spacing) that should explain how you came up with the list of IDs. 35 points will be based on model performance.

Answer the following questions:

- Describe how you developed your response model (You should explain which variables you used in your model clearly), and discuss its expected predictive performance. If you created new variables to include into the response model, please describe these as well.
- What criteria did you use to decide who should receive the wave 2 mailing?
- How much profit do you anticipate with your wave 2 mailing? (see instruction #2)
- What did you learn about the type of consumers who are likely to upgrade?
- You should include your do-file at the end of your write-up as an appendix for TA to understand how you ran your model. Appendix is not counted for the word count limit.

6. By 6 p.m. on Monday , please send me by e-mail the list of customer you want to mail (no need to send the assignment early). The reason is that I want to compile all the results before class and ask some groups during class to discuss what they did. As a result, in the evening before class I will need to compile about many data files, line up all the IDs and evaluate the results. This takes quite a bit of time and you can help me a lot if you stick to the following instructions **exactly**!

Please include in the e-mail:

- The first and last names of your group members
- A Stata dataset (not excel) with exactly 3 variables
  1. The original **"id"** variable from the intuit75.dta dataset (lower case, please don't rename the variable). The column should contain all customer IDs from the validation sample (do NOT delete the customers you don't want to target from the dataset, i.e. please do not just send me a list of those you plan to mail; I want 22,500 rows in the data!).
  2. Please sort the data by the id variable.
  3. A variable named **"mailto_wave2"** (lower case) that contains a "1" if you want to target the customer for wave 2, and "0" if you don't.
  4. A variable named **"group"** (lower case) that contains in every row the **first** names of **all** your group members separated by underscores "_" (e.g. Nancy_Sam_Manuel). You create this variable by typing, for example:

```
gen group="Nancy_Sam_Manuel"
```

  - Example:
    Please ignore the section variable in the below example because I am teaching only one section at UCL.

	id	mailto_wave2	group	section
1	2	0	Nancy_Sam_Manuel	62
2	3	1	Nancy_Sam_Manuel	62
3	9	0	Nancy_Sam_Manuel	62
4	15	1	Nancy_Sam_Manuel	62
5	18	0	Nancy_Sam_Manuel	62
6	22	0	Nancy_Sam_Manuel	62

  - Please name the dataset descriptively with the same name as the group variable in the dataset (e.g. Nancy_Sam_Manuel.dta)
- Put in the subject line of the e-mail "Intuit:" followed by the same name as group variable (e.g. "Intuit: Nancy_Sam_Manuel ")
- Sorry to be so particular about this, but you help me out big-time if you follow these instructions!

Hints

1. This case is a bit of a journey of discovery. An important part of that discovery is in the zip-code information (don't ignore zip-codes!).

    a. In case you want to use zip codes as predictors, I have created zip code buckets ("zip_bins") for you from 1-20, each of which contains about 3750 consumers sorted by the zip code they are in. Please remember that, although this variable that contains numbers 1-20, it really is a categorical variable, not a metric variable.

    b. Note that the data contains the actual zip-code of consumers. You can use these to create more/different bins than those I defined in "zip_bins".

    c. The zip codes in the "zip" variable don't all have 5 digits. The reason is that "zip" is a numerical variable and therefore leadings zeros are not displayed. For example, 2138 is actually 02138.

2. In addition to using Stata you might also consider using the neural networks in Azure ML to inform your choice of response model. We will have done a demo in class on how to use Azure ML.

3. There is a useful command for quickly transforming categorical variables into dummies. The command is `"xi"`. For example, assume you want to regress X on Y, Z, and T. Suppose that Z is a categorical (non-metric) variable that you decide you would like to put in as a series of dummy variables. Then you could type:

```
xi i.Z, prefix(d_)
```

    The idea is that any variable ahead of which you place `"i."` when you execute the "xi" command will be automatically be converted by Stata into a series of dummy variables, the first of which is automatically dropped since you always have to leave out one of each set of dummies in any regression (if it is not clear to you why this needs to be true, please go back to reading the dummy variable discussion in "Tips for Using Statistics in Marketing Analytics".) Hence, suppose that Z contained 5 values. Then this will create 4 dummy variables:

```
d_Z_2, d_Z_3, d_Z_4, d_Z_5
```

    Now you can run your regression as follows:

```
regress X Y d_Z_2-d_Z_5 T
```

where "–" tells Stata to include all variables from `d_Z_2` through `d_Z_5`, i.e. `d_Z_2, d_Z_3, d_Z_4,` and `d_Z_5.`

The effect of each `d_Z` dummy will now be measured relative to the omitted `d_Z_1`.

Now, suppose, however, that in the logistic regression you want to measure everything against `d_Z_3`, i.e. you want to omit a dummy other than the first one. To do this use the "noomit" option in xi:

```
xi i.Z, prefix(d_) noomit
```

Then this will create 5 dummy variables:

```
d_Z_1, d_Z_2, d_Z_3, d_Z_4, d_Z_5
```

Now you can run your regression as follows:

```
regress X Y d_Z_1-d_Z_4 T
```

or:

```
regress X Y d_Z_1-d_Z_2 d_Z_4-d_Z_5 T
```

The effect of each dummy will now be measured relative to the omitted "`d_Z_5`" in the first logistic command or "`d_Z_3`" in the second logistic command above.

*Overall, remember that the odds ratio for each dummy measures the effect relative to the **dummy (or the average of multiple dummies)** that is (are) left out.*

NOTE: Every time you issue the "xi" command Stata will delete all variables with the prefix which you specify in the command (e.g. in the prior example the prefix is "d_"). To leave previously "xi" created variables intact, use a different prefix when using "xi" a second time (e.g. "d2_").

4. There is a nice way to have Stata automatically determine which variables should be left in the regression and which ones do not matter.

   - First, you will need to give Stata more room to include variable in the regression. Type "set matsize 400" (this is the maximum number of variables you want to include in your regression, any number up to 800 is accepted)

   - Now use the "sw" stepwise command. For example say that you want to run the logistic regression of res1 on the sex categories. Then you would type:

     ```
     xi i.sex, prefix(d_)
     sw logistic res1 d_sex*, pr(.2)
     ```

   - This makes Stata start with all specified variables and recursively drop those with a significance level of below 0.2. This is a way to quickly tell you which of the variables have a predictive value. Then you can run the logistic regression again (without sw which can take a long time), but only with the important variables.

   - Using pr(.2) is a good guideline. Please don't use pr(.05). If you did, this would correspond to dropping all variables that are not significant. This is not a good idea because even when a

variable is not significant, if it has a low enough p-value, it can still cause omitted variable bias when left out of the regression.

5. Contact me if you have problems or questions!