**Final Project**

The data for this project comes from the mobile advertising space. In order to encourage consumers to install its app (e.g. a game), an app developer advertises its app on other apps (e.g., other games) through a mobile advertising platform. Consumers viewing these ads on these other apps can click on the ad to install the app from the developer. We will refer to the advertising app developer as the advertiser. See figure below.

Advertiser

Advertising Platform

.

.

.

Publisher 1

Publisher 2

Publisher k-1

Publisher k

Consumer

Consumer

Consumer

Consumer

.

.

.

Install

Not Install

The dataset for this project contains data about ads from one particular advertiser through multiple publishers. Each observation corresponds to one ad shown to a consumer on a particular publisher app. The observation contains information about the publisher id, consumer’s device characteristics, and whether the advertiser’s app was installed or not. The description of the variables are given below.

|  |  |  |
| --- | --- | --- |
| **Variable** | **Type** | **Description** |
| publisher\_id\_class | Categorical | Publisher Id |
| device\_make\_class | Categorical | Device Manufacturer |
| device\_platform\_class | Categorical | Phone OS Type (iPhone / Android) |
| device\_os\_class | Categorical | Phone OS Version |
| device\_height | Numerical | Display Height (in pixels) |
| device\_width | Numerical | Display Width (in pixels) |
| resolution | Numerical | Display Resolution (pixels per inch) |
| device\_volume | Numerical | Device Volume when Ad was displayed |
| wifi | Binary | Whether WiFi was enabled when ad was displayed (Yes = 1, No = 0) |
| install | Binary | Whether Consumer Installed Advertiser’s App (Yes = 1, No = 0) |

**Part I.**

The advertiser needs to determine how much to pay for placing ad depending on the publisher and on the consumer characteristics. The optimal payment is proportional to the probability that a consumer seeing the ad will install the ad.

Develop a **linear probability model** to estimate the probability of installing the ad based on publisher and consumer characteristics. Present only the final model and explain the procedure and different measures you have used to come up with this model.

Develop a **logistic regression model** to estimate the probability of installing the ad based on publisher and consumer characteristics. Present only the final model and explain the procedure and different measures you have used to come up with this model. Do you need to consider modeling of rare events in this case – why / why not? Present the results of both approaches – that is (i) estimate the model without considering rare events, and (ii) estimate the model using oversampling approach for handling rare events and then applying the correction to obtain the corrected intercept (see lecture, also see <http://support.sas.com/kb/22/601.html> for how to directly handle this in SAS).

Plot the ROC curves for all models. (Hint: You can use PROC LOGISTIC to plot the ROC for the linear probability model without fitting the model – see lecture. Similarly, you can plot the ROC using the original dataset for the logistic regression model for rare outcomes after estimating it with the oversampled data. Using the oversampled data for the ROC will give you a wrong comparison).

Which of the above models has the highest AUC (area under the curve). At the 95% confidence level, is the AUC of this model higher than those of the other models? (Hint: You need to look at the confidence intervals reported for AUC).

**Part II**

The advertising platform would like to determine whether to show the ad from this advertiser depending on the publisher and consumer characteristics. In particular, the advertising platform needs to come up with a threshold such that if the probability of installing the ad is above that threshold, the ad is shown to the consumer.

Showing an ad to a consumer who would not install the app results in some inconvenience cost to the consumer which in turn leads to less participation and causes a loss of 1 cent to the platform. On the other hand, not showing an ad to a consumer who would have installed the app results in a missed opportunity cost of 100 cents to the platform. The platform would like to minimize the total expected cost.

For each of the above models you estimated, generate the ROC table using SAS, and plot the total cost for different threshold values. Note that for the linear probability model (unlike the logistic regression model), SAS does not generate the ROC table automatically. You will need to write a proc or data step to create the table yourself. To make your job easier, you can calculate the total cost at these thresholds:

0.001 0.005 0.010 0.015 0.020 0.025 0.030 0.035 0.040 0.045 0.050

Which of these models provide the lowest total cost? (For the logistic regression model for rare outcomes, you cannot use the oversampled data to calculate the cost since this is not representative of the actual distribution of outcomes.)

Deliverables

* Project Report: For each question above, describe the model building and selection process that you followed, along with suitable tables and graphs as necessary.
* SAS code: Include a SAS file with detailed comments to reproduce all the results, tables and figures in the report. The code must be clearly labeled so that it is straightforward to see how to reproduce a particular result / table / figure. If the code will not execute, then points will be deducted. The code should assume that it will be executed in the folder containing the dataset.