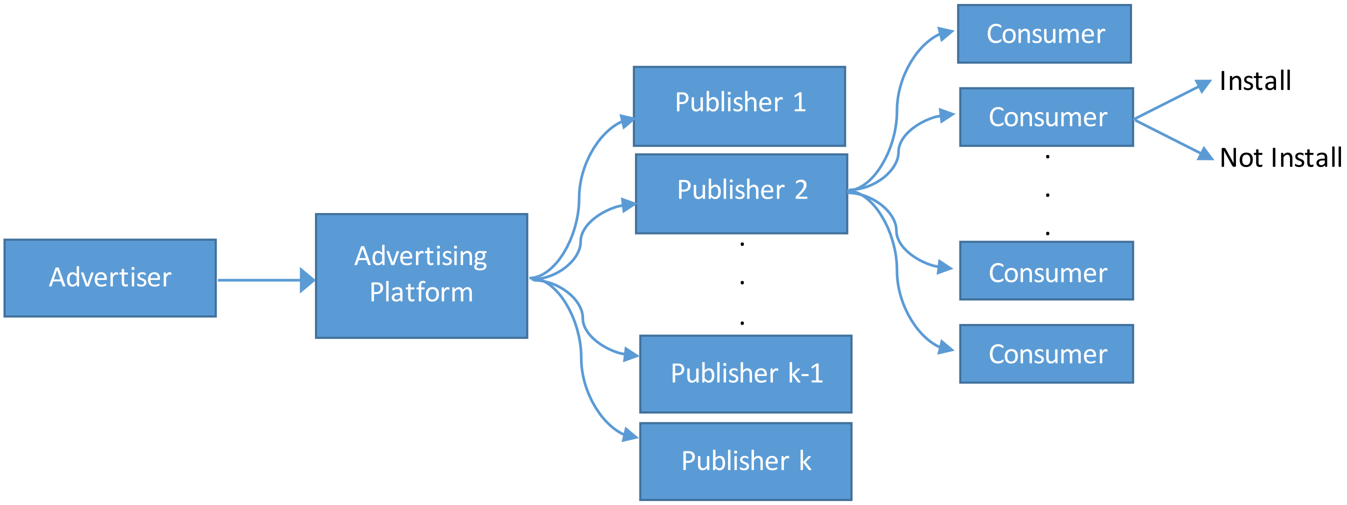
Final Project

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.



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	Numerical	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. Logistic Regression Analysis

1. 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 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.
2. Plot the ROC curve for this model, and report the area under the ROC curve.
3. 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. Using the ROC table generated by SAS, plot the total cost for different threshold values. Calculate the threshold at which the cost is minimized and report the cost at this threshold.

Part II. Linear Probability Model

1. 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.
2. Plot the ROC curve for this model, and report the area under the ROC curve. Note that in order to this, you have to use the *Proc logistic* command, but without fitting the model. (Refer to the lecture). At the 95% confidence level, is the area below this graph higher than the one you obtained from the logistics regression? (Hint: You need to look at the confidence intervals)
3. Repeat the analysis you have done in part I to determine the optimal threshold assuming the same numbers for costs. Is the optimal threshold different in two cases? How about the optimal expected cost?
   Note that unlike the logistic regression case, SAS does not generate the ROC table automatically. 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

<u>Deliverables</u>

- 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.