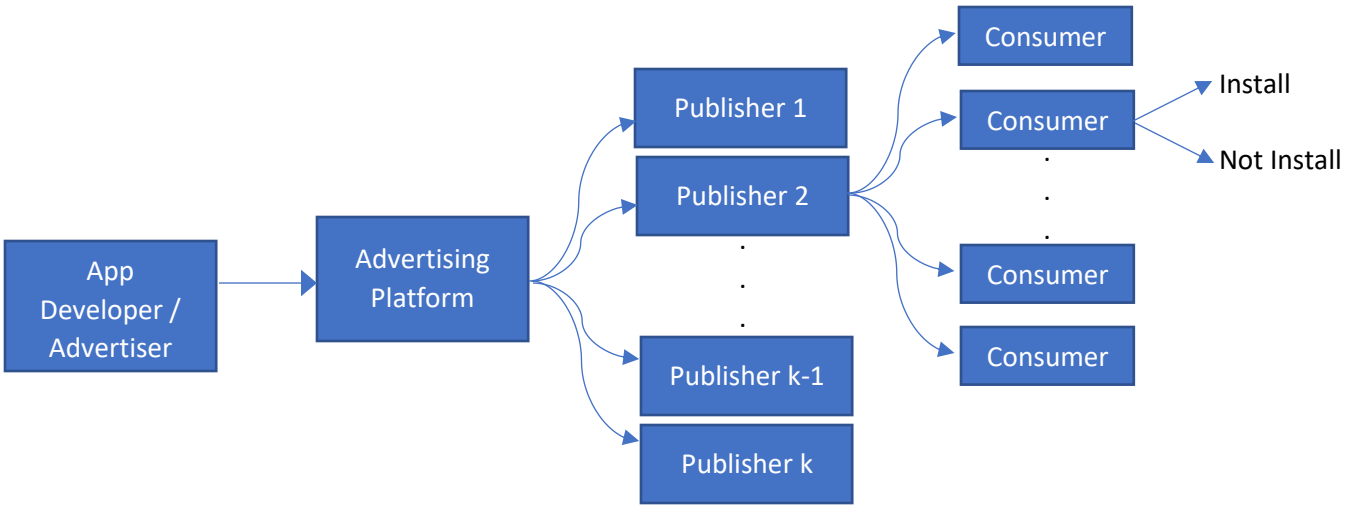
BUAN 6337 Predictive Analytics using SAS

Project 2

Predictive Models for Mobile Advertising

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*. These other apps are developed by other game *publishers*. Consumers who view the 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* and the other apps as *publishers*. 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.

The advertiser needs to determine how much to pay for placing an 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.

a) Develop a **linear probability model** to predict the probability of installing the ad based on publisher and consumer characteristics. Describe in detail your approach for model building, evaluation and selection. Present your final model and performance metrics.

   The description of your approach should include, for example, what variables to include in your model building process (and why), did you create new variables from existing ones (and why), how / what alternative models did you consider, how did you compare these alternative models and why did you compare these models in this way.

b) Develop a **logistic regression model** to estimate the probability of installing the ad based on publisher and consumer characteristics. Describe your approach as in part (a) above – elaborating only what is new or different than above. Present your final model and performance metrics.

   In particular, discuss whether you need to consider modeling of rare events in this case – why / why not? Compare the results with and without considering rare events - (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
   (Note: See lecture for how you can to calculate the correction after estimating the model. One approach is to implement this correction using a DATA step after estimating the model with the oversampling approach. Another approach is to directly implement through PROC LOGISTIC - see support.sas.com/kb/22/601.html for how to do this).

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.

a) For each of the above models you estimated in part I above, generate the ROC table using SAS, and plot the total cost for different threshold values. (question contd. next page)

   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

b) 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. Upload 1 pdf/word file for the entire project which includes your description for all the questions.

- 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. Make sure your codes can be executed properly when uploaded, as it is part of your project score.

Approach:

Part One

- **Linear probability models**

First, set the working directory using Libname and import the dataset 'ADS.DATA' using DATA.
The dataset:

	install	device_volume	wifi	resolution	device_height	device_width	publisher_id_class	device_os_class	device_make_class	device_platform_class
1	0	0.870000005	1	0.727039993	640	1136	3	1	1	iOS
2	0	0.860000014	1	1.000499964	750	1334	10	4	2	iOS
3	0	0.560000002	1	0.727039993	1136	640	10	1	5	iOS
4	0	1	1	0.727039993	640	1136	10	4	5	iOS
5	0	0.119999997	1	0.727039993	640	1136	6	3	1	iOS
6	0	1	1	0.727039993	640	1136	10	1	1	iOS
7	0	0.310000002	0	0.727039993	1136	640	9	10	3	iOS
8	0	0.469999999	0	2.251125097	2001	1125	10	1	2	iOS
9	0	0.119999997	0	0.727039993	1136	640	8	5	2	iOS
10	0	0.059999999	1	0.727039993	1136	640	3	1	3	iOS
11	0	0.189999998	1	0.727039993	1136	640	3	4	1	iOS
12	0	0.810000002	0	0.727039993	640	1136	6	1	1	iOS
13	0	0.439999998	1	0.727039993	1136	640	10	1	3	iOS
14	0	0.310000002	1	0.727039993	1136	640	7	1	2	iOS
15	0	0.119999997	1	0.727039993	640	1136	6	5	1	iOS
16	0	0.460000008	0	0.727039993	640	1136	3	5	2	iOS
17	0	0.430000007	0	0.727039993	640	1136	10	2	2	iOS
18	0	0.439999998	1	2.742336035	1242	2208	5	3	1	iOS
19	0	1	1	2.742336035	1242	2208	10	1	1	iOS
20	0	0.310000002	0	0.727039993	1136	640	10	1	1	iOS
21	0	0.870000005	1	3.145728111	2048	1536	10	4	4	iOS
22	0	1	1	1.000499964	1334	750	4	6	1	iOS
23	0	0.310000002	1	3.145728111	1536	2048	6	1	8	iOS
24	0	0.560000002	1	0.727039993	640	1136	2	1	1	iOS
25	0	0.560000002	1	0.727039993	640	1136	3	1	5	iOS
26	0	0.310000002	1	2.742336035	1242	2208	5	1	2	iOS
27	0	0.730000019	1	0.727039993	640	1136	2	6	2	iOS
28	0	0.689999998	1	3.145728111	2048	1536	8	4	8	iOS
29	0	1	1	0.727039993	1136	640	10	2	2	iOS
30	0	0.870000005	0	0.727039993	640	1136	6	8	2	iOS
31	0	0.25	1	0.614400029	640	960	3	10	10	iOS
32	0	0.310000002	1	3.145728111	1536	2048	5	10	9	iOS
33	0	1	1	0.727039993	640	1136	4	1	1	iOS
34	0	0.25	1	0.727039993	640	1136	2	4	1	iOS
35	0	0.620000005	1	1.000499964	750	1334	10	1	1	iOS
36	0	0.439999998	1	3.145728111	1536	2048	4	2	9	iOS
37	0	0.189999998	1	1.000499964	1334	750	10	4	7	iOS
38	0	1	1	1.000499964	750	1334	4	2	1	iOS
39	0	0.189999998	1	3.145728111	2048	1536	7	1	4	iOS
40	0	0.870000005	1	0.727039993	1136	640	10	5	3	iOS
41	0	0.930000007	0	2.073600054	1920	1080	10	10	10	android
42	0	0.310000002	1	3.145728111	1536	2048	10	2	4	iOS
43	0	0.439999998	1	2.742336035	2208	1242	10	1	2	iOS
44	0	0.689999998	1	0.614400029	640	960	10	2	10	iOS
45	0	1	0	0.727039993	640	1136	4	1	3	iOS
46	0	0.189999998	0	3.145728111	1536	2048	10	2	8	iOS
47	0	1	0	0.727039993	1136	640	10	2	3	iOS
48	0	0.280000001	1	0.727039993	1136	640	10	3	5	iOS
49	0	1	0	0.786432028	768	1024	6	8	6	iOS
50	0	0.189999998	1	3.145728111	1536	2048	10	1	10	iOS

Next we will split the dataset into Training and Test Dataset sample. We will split the dataset in the ratio 80% Training and 20% Test dataset
Training dataset:

	Selection Indicator	install	device_volume	wifi	resolution	device_height	device_width	publisher_id_class	device_os_class	device_make_class	device_platform_class
1	1	0	0.860000014	1	1.000499964	750	1334	10	4	2	iOS
2	1	0	0.560000002	1	0.727039993	1136	640	10	1	5	iOS
3	1	0	0.310000002	0	0.727039993	1136	640	9	10	3	iOS
4	1	0	0.469999999	0	2.251125097	2001	1125	10	1	2	iOS
5	1	0	0.119999997	0	0.727039993	1136	640	8	5	2	iOS
6	1	0	0.189999998	1	0.727039993	1136	640	3	4	1	iOS
7	1	0	0.810000002	0	0.727039993	640	1136	6	1	1	iOS
8	1	0	0.439999998	1	0.727039993	1136	640	10	1	3	iOS
9	1	0	0.310000002	1	0.727039993	1136	640	7	1	2	iOS
10	1	0	0.430000007	0	0.727039993	640	1136	10	2	2	iOS
11	1	0	0.439999998	1	2.742336035	1242	2208	5	3	1	iOS
12	1	0	1	1	2.742336035	1242	2208	10	1	1	iOS
13	1	0	0.310000002	0	0.727039993	1136	640	10	1	1	iOS
14	1	0	1	1	1.000499964	1334	750	4	6	1	iOS
15	1	0	0.310000002	1	3.145728111	1536	2048	6	1	8	iOS
16	1	0	0.560000002	1	0.727039993	640	1136	2	1	1	iOS
17	1	0	0.310000002	1	2.742336035	1242	2208	5	1	2	iOS
18	1	0	0.730000019	1	0.727039993	640	1136	2	6	2	iOS
19	1	0	0.689999998	1	3.145728111	2048	1536	8	4	8	iOS
20	1	0	1	1	0.727039993	1136	640	10	2	2	iOS
21	1	0	0.25	1	0.614400029	640	960	3	10	10	iOS
22	1	0	1	1	0.727039993	640	1136	4	1	1	iOS
23	1	0	0.25	1	0.727039993	640	1136	2	4	1	iOS
24	1	0	0.620000005	1	1.000499964	750	1334	10	1	1	iOS
25	1	0	0.439999998	1	3.145728111	1536	2048	4	2	9	iOS
26	1	0	0.189999998	1	1.000499964	1334	750	10	4	7	iOS
27	1	0	1	1	1.000499964	750	1334	4	2	1	iOS
28	1	0	0.189999998	1	3.145728111	2048	1536	7	1	4	iOS
29	1	0	0.870000005	1	0.727039993	1136	640	10	5	3	iOS
30	1	0	0.930000007	0	2.073600054	1920	1080	10	10	10	android
31	1	0	0.689999998	1	0.614400029	640	960	10	2	10	iOS
32	1	0	1	0	0.727039993	640	1136	4	1	3	iOS
33	1	0	0.189999998	0	3.145728111	1536	2048	10	2	8	iOS
34	1	0	1	0	0.727039993	1136	640	10	2	3	iOS
35	1	0	0.280000001	1	0.727039993	1136	640	10	3	5	iOS
36	1	0	1	0	0.786432028	768	1024	6	8	6	iOS
37	1	0	0.189999998	1	3.145728111	1536	2048	10	1	10	iOS
38	1	0	0.560000002	1	1.000499964	1334	750	10	10	1	iOS
39	1	0	1	1	0.727039993	640	1136	4	1	1	iOS
40	1	0	0.439999998	1	3.145728111	1536	2048	10	1	4	iOS
41	1	0	0.370000005	1	3.145728111	2048	1536	10	10	4	iOS
42	1	0	0.620000005	0	0.727039993	1136	640	10	3	1	iOS
43	1	0	0.689999998	1	0.786432028	768	1024	10	10	6	iOS
44	1	0	0.189999998	1	0.727039993	640	1136	3	1	7	iOS
45	1	0	0.439999998	0	2.742336035	1242	2208	10	3	2	iOS
46	1	0	1	1	0.727039993	640	1136	10	3	3	iOS
47	1	0	1	1	3.145728111	2048	1536	10	8	8	iOS
48	1	0	1	0	0.727039993	1136	640	10	1	2	iOS
49	1	0	0.5	1	0.786432028	768	1024	2	2	6	iOS

Testing Dataset:

	Selection Indicator	install	device_volume	wifi	resolution	device_height	device_width	publisher_id_class	device_os_class	device_make_class	device_platform_class
1	0	0	0.870000005	1	0.727039993	640	1136	3	1	1	iOS
2	0	0	1	1	0.727039993	640	1136	10	4	5	iOS
3	0	0	0.119999997	1	0.727039993	640	1136	6	3	1	iOS
4	0	0	1	1	0.727039993	640	1136	10	1	1	iOS
5	0	0	0.059999999	1	0.727039993	1136	640	3	1	3	iOS
6	0	0	0.119999997	1	0.727039993	640	1136	6	5	1	iOS
7	0	0	0.460000008	0	0.727039993	640	1136	3	5	2	iOS
8	0	0	0.870000005	1	3.145728111	2048	1536	10	4	4	iOS
9	0	0	0.560000002	1	0.727039993	640	1136	3	1	5	iOS
10	0	0	0.870000005	0	0.727039993	640	1136	6	8	2	iOS
11	0	0	0.310000002	1	3.145728111	1536	2048	5	10	9	iOS
12	0	0	0.310000002	1	3.145728111	1536	2048	10	2	4	iOS
13	0	0	0.439999998	1	2.742336035	2208	1242	10	1	2	iOS
14	0	0	0.939999998	0	0.727039993	640	1136	4	1	1	iOS
15	0	0	0.059999999	0	0.727039993	640	1136	10	1	5	iOS
16	0	0	0.620000005	0	0.727039993	640	1136	10	1	3	iOS
17	0	0	0.5	1	3.145728111	1536	2048	10	5	4	iOS
18	0	0	0.870000005	0	1.000499964	750	1334	10	3	2	iOS
19	0	0	0.5	0	1.000499964	750	1334	10	1	1	iOS
20	0	0	0.119999997	1	0.727039993	640	1136	4	9	3	iOS
21	0	0	0.310000002	0	1.000499964	1334	750	10	3	1	iOS
22	0	0	0.810000002	1	1.000499964	1334	750	10	1	2	iOS
23	0	0	1	1	0.727039993	640	1136	3	1	5	iOS
24	0	0	0.439999998	1	0.614400029	960	640	8	7	10	iOS
25	0	0	0.189999998	1	0.786432028	768	1024	10	2	6	iOS
26	0	0	0.209999993	1	2.742336035	1242	2208	10	2	2	iOS
27	0	0	0.209999993	1	1.000499964	1334	750	10	3	1	iOS
28	0	0	0.189999998	1	0.727039993	640	1136	10	1	1	iOS
29	0	0	0.689999998	0	0.786432028	1024	768	6	10	6	iOS
30	0	0	1	1	3.145728111	1536	2048	10	8	4	iOS
31	0	0	0.469999999	1	0.727039993	1136	640	8	1	1	iOS
32	0	0	1	1	3.145728111	1536	2048	10	10	4	iOS
33	0	0	1	1	1.000499964	750	1334	10	1	1	iOS
34	0	0	0.439999998	1	3.145728111	1536	2048	2	2	9	iOS
35	0	0	0.189999998	1	3.145728111	1536	2048	9	5	8	iOS
36	0	0	0.460000008	0	0.727039993	640	1136	4	5	2	iOS
37	0	0	1	1	1.000499964	750	1334	10	1	1	iOS
38	0	0	0.75	1	3.145728111	1536	2048	6	8	9	iOS
39	0	0	0.5	1	3.145728111	2048	1536	10	10	8	iOS
40	0	0	0.560000002	0	0.727039993	1136	640	8	10	2	iOS
41	0	0	0.810000002	1	3.145728111	1536	2048	6	5	9	iOS
42	0	0	0.439999998	1	0.727039993	1136	640	7	4	3	iOS
43	0	0	0.170000002	1	3.145728111	2048	1536	10	8	4	iOS
44	0	0	0.370000005	1	0.727039993	1136	640	10	10	3	iOS
45	0	0	0.059999999	1	0.727039993	640	1136	2	1	3	iOS
46	0	0	1	1	1.000499964	750	1334	10	1	7	iOS
47	0	0	0.059999999	1	5.595136166	2048	2732	10	4	10	iOS
48	0	0	1	0	0.727039993	640	1136	3	4	1	iOS
49	0	1	0.439999998	1	0.786432028	768	1024	10	2	6	iOS

As the dataset have categorical variables, so converting it into numerical variables using the proc glmmod.

The encoding for the initial dataset. Notice the device_platform_class_ prefixed columns:

	install	Selection Indicator	device_volume	wifi	resolution	device_height	device_width	publisher_id_class	device_os_class	device_make_class	device_platform_clas android	device_platform_clas iOS
10	0	0	0.059999999	1	0.727039993	1136	640	3	1	3	0	1
11	0	1	0.189999998	1	0.727039993	1136	640	3	4	1	0	1
12	0	1	0.810000002	0	0.727039993	640	1136	6	1	1	0	1
13	0	1	0.439999998	1	0.727039993	1136	640	10	1	3	0	1
14	0	1	0.310000002	1	0.727039993	1136	640	7	1	2	0	1
15	0	0	0.119999997	1	0.727039993	640	1136	6	5	1	0	1
16	0	0	0.460000008	0	0.727039993	640	1136	3	5	2	0	1
17	0	1	0.430000007	0	0.727039993	640	1136	10	2	2	0	1
18	0	1	0.439999998	1	2.742336035	1242	2208	5	3	1	0	1
19	0	1	1	1	2.742336035	1242	2208	10	1	1	0	1
20	0	1	0.310000002	0	0.727039993	1136	640	10	1	1	0	1
21	0	0	0.870000005	1	3.145728111	2048	1536	10	4	4	0	1
22	0	1	1	1	1.000499964	1334	750	4	6	1	0	1
23	0	1	0.310000002	1	3.145728111	1536	2048	6	1	8	0	1
24	0	1	0.560000002	1	0.727039993	640	1136	2	1	1	0	1
25	0	0	0.560000002	1	0.727039993	640	1136	3	1	5	0	1
26	0	1	0.310000002	1	2.742336035	1242	2208	5	1	2	0	1
27	0	1	0.730000019	1	0.727039993	640	1136	2	6	2	0	1
28	0	1	0.689999998	1	3.145728111	2048	1536	8	4	8	0	1
29	0	1	1	1	0.727039993	1136	640	10	2	2	0	1
30	0	0	0.870000005	0	0.727039993	640	1136	6	8	2	0	1
31	0	1	0.25	1	0.614400029	640	960	3	10	10	0	1
32	0	0	0.310000002	1	3.145728111	1536	2048	5	10	9	0	1
33	0	1	1	1	0.727039993	640	1136	4	1	1	0	1
34	0	1	0.25	1	0.727039993	640	1136	2	4	1	0	1
35	0	1	0.620000005	1	1.000499964	750	1334	10	1	1	0	1
36	0	1	0.439999998	1	3.145728111	1536	2048	4	2	9	0	1
37	0	1	0.189999998	1	1.000499964	1334	750	10	4	7	0	1
38	0	1	1	1	1.000499964	750	1334	4	2	1	0	1
39	0	1	0.189999998	1	3.145728111	2048	1536	7	1	4	0	1
40	0	1	0.870000005	1	0.727039993	1136	640	10	5	3	0	1
41	0	1	0.930000007	0	2.073600054	1920	1080	10	10	10	1	0
42	0	0	0.310000002	1	3.145728111	1536	2048	10	2	4	0	1
43	0	0	0.439999998	1	2.742336035	2208	1242	10	1	2	0	1
44	0	1	0.689999998	1	0.614400029	640	960	10	2	10	0	1
45	0	1	1	0	0.727039993	640	1136	4	1	3	0	1
46	0	1	0.189999998	0	3.145728111	1536	2048	10	2	8	0	1
47	0	1	1	0	0.727039993	1136	640	10	2	3	0	1
48	0	1	0.280000001	1	0.727039993	1136	640	10	3	5	0	1
49	0	1	1	0	0.786432028	768	1024	6	8	6	0	1
50	0	1	0.189999998	1	3.145728111	1536	2048	10	1	10	0	1
51	0	1	0.560000002	1	1.000499964	1334	750	10	10	1	0	1
52	0	1	1	1	0.727039993	640	1136	4	1	1	0	1
53	0	1	0.439999998	1	3.145728111	1536	2048	10	1	4	0	1
54	0	1	0.370000005	1	3.145728111	2048	1536	10	10	4	0	1
55	0	1	0.620000005	0	0.727039993	1136	640	10	3	1	0	1
56	0	1	0.689999998	1	0.786432028	768	1024	10	10	6	0	1
57	0	0	0.939999998	0	0.727039993	640	1136	4	1	1	0	1
58	0	0	0.059999999	0	0.727039993	640	1136	10	1	5	0	1

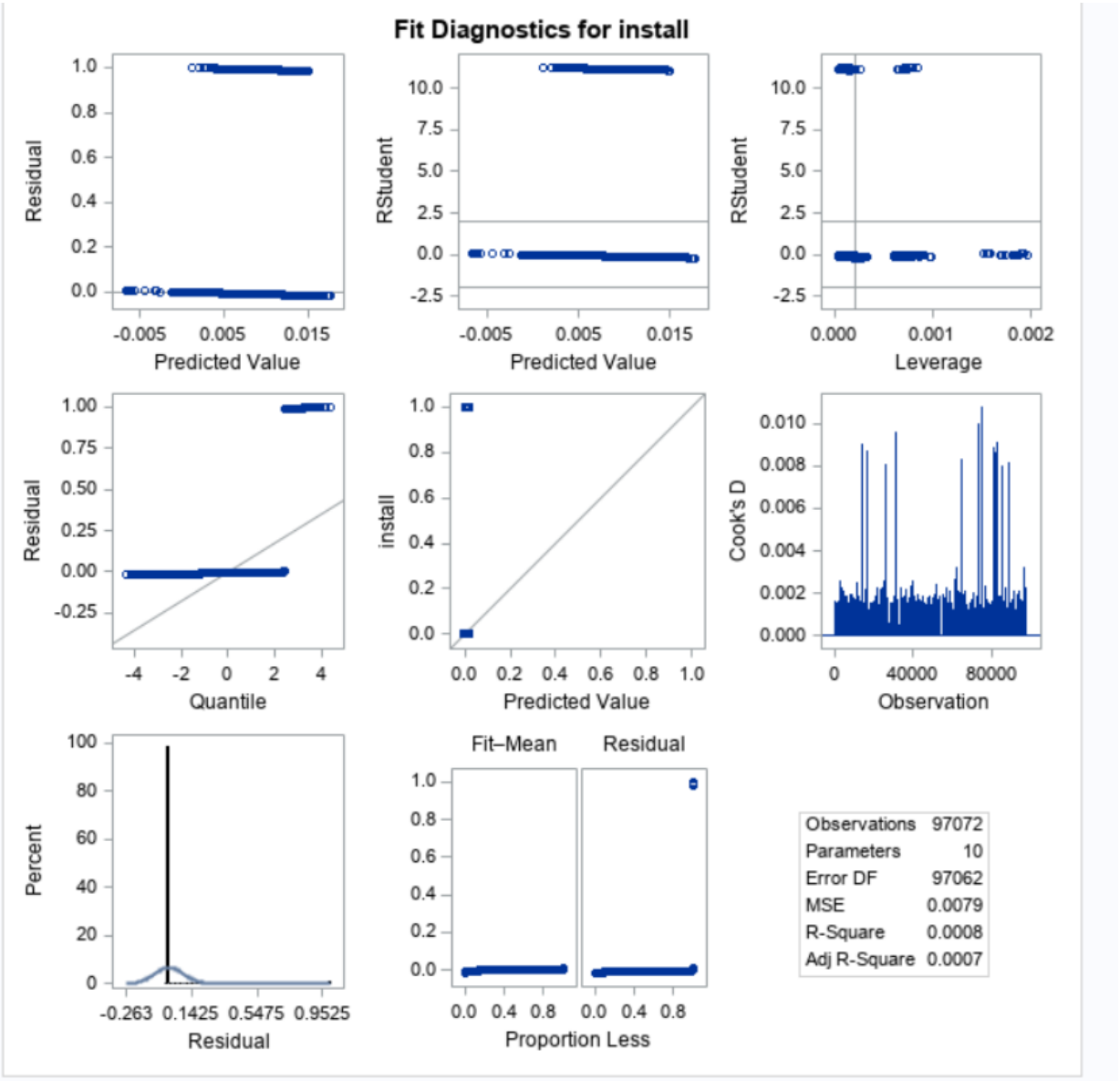
Using the proc contents statement to provide alias names as column numbers to the feature variables.

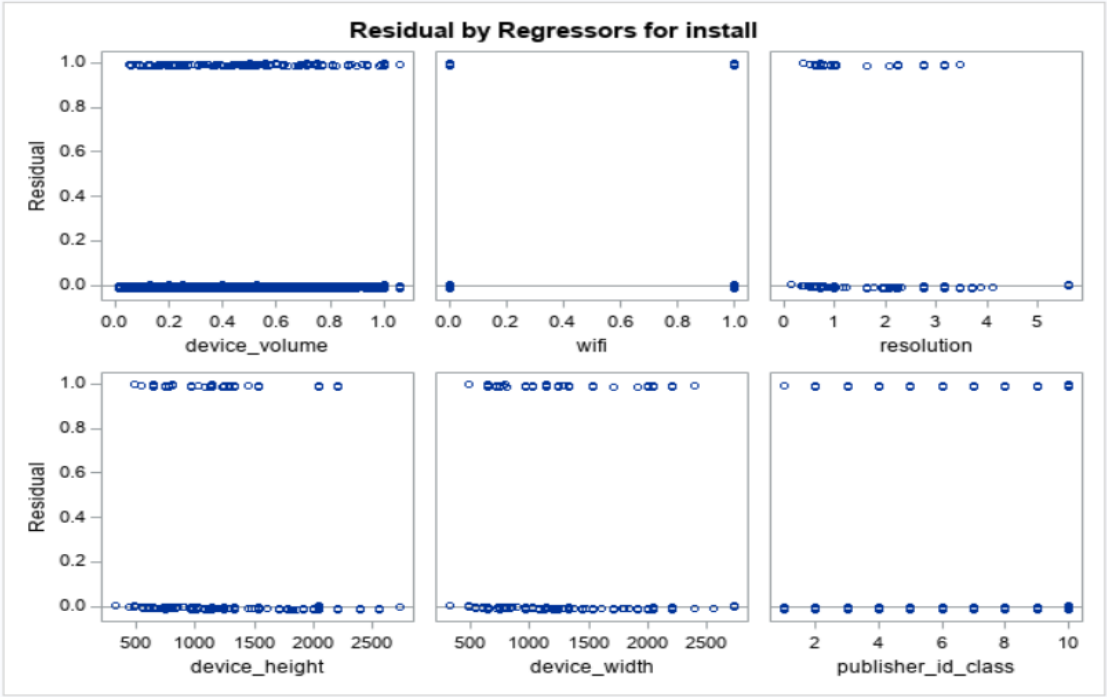
Alphabetic List of Variables and Attributes

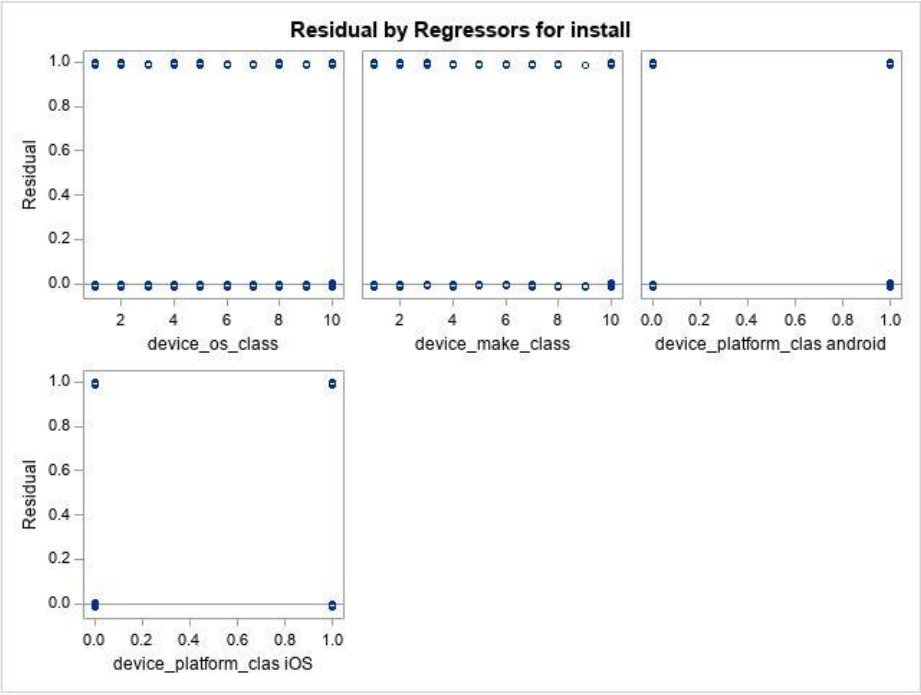
#	Variable	Type	Len	Format	Informat	Label
2	Col1	Num	8			device_volume
3	Col2	Num	8			wifi
4	Col3	Num	8			resolution
5	Col4	Num	8			device_height
6	Col5	Num	8			device_width
7	Col6	Num	8			publisher_id_class
8	Col7	Num	8			device_os_class
9	Col8	Num	8			device_make_class
10	Col9	Num	8			device_platform_clas android
11	Col10	Num	8			device_platform_clas iOS
1	install	Num	8	BEST12.	BEST32.	

(i) Linear Probability Model:

Ran a linear model (Kitchen sink regression) taking install as dependent and rest as independent variable using the Proc reg.







As we can observe from the Fit Diagnostic, there are lot of outliers in the dataset which can result into poor model. So we imputed those values to make sure the variance effect of the outliers does not affect the model.

Initial Linear Model :

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: install

Number of Observations Read	97072
Number of Observations Used	97072

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	0.60172	0.06686	8.45	<.0001
Error	97062	768.21086	0.00791		
Corrected Total	97071	768.81258			

Root MSE	0.08896	R-Square	0.0008
Dependent Mean	0.00798	Adj R-Sq	0.0007
Coeff Var	1114.31475		

Col10 = Intercept - 157E-13 * Col3 + 2E-14 * Col4 + 201E-16 * Col5 - Col9

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	B	-0.01931	0.00632	-3.05	0.0023
Col1	device_volume	1	0.00082070	0.00093070	0.88	0.3779
Col2	wifi	1	0.00201	0.00063472	3.17	0.0015
Col3	resolution	B	-0.01495	0.00386	-3.87	0.0001
Col4	device_height	B	0.00002219	0.00000497	4.46	<.0001
Col5	device_width	B	0.00001997	0.00000500	4.00	<.0001
Col6	publisher_id_class	1	-0.00049980	0.00010791	-4.63	<.0001
Col7	device_os_class	1	-0.00012328	0.00009633	-1.28	0.2006
Col8	device_make_class	1	0.00052335	0.00013412	3.90	<.0001
Col9	device_platform_clas android	B	-0.00091736	0.00241	-0.38	0.7040
Col10	device_platform_clas iOS	0	0	.	.	.

Here we can observe that all the variables have significant p – values at 1%, 5% and 10% levels.
Now we will run a log model to check if the significance of the variables becomes more significant.

The SAS System

The REG Procedure
Model: Logmodel
Dependent Variable: install

Number of Observations Read	121339
Number of Observations Used	97072

Weight: Selected Selection Indicator

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	0.47895	0.05987	7.56	<.0001
Error	97063	768.33363	0.00792		
Corrected Total	97071	768.81258			

Root MSE	0.08897	R-Square	0.0006
Dependent Mean	0.00798	Adj R-Sq	0.0005
Coeff Var	1114.39805		

Col5 =	Intercept - Col4
Col15 =	13.8155 * Intercept + 1.13E-9 * Col2 - 5.15E-9 * Col3 + 1.72E-8 * Col4 - 314E-12 * Col6 - 3.01E-9 * Col11 - 1 * Col12 - 1.03E-9 * Col13 + 1 * Col14

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	B	0.02400	0.00911	2.64	0.0084
Col2	publisher_id_class	B	-0.00040942	0.00010792	-3.79	0.0001
Col3	device_make_class	B	0.00024138	0.00011257	2.14	0.0320
Col4	device_platform_clas android	B	0.00052431	0.00239	0.22	0.8265
Col5	device_platform_clas iOS	0	0	.	.	.
Col6	device_os_class	B	-0.00011909	0.00009633	-1.24	0.2164
Col11	wifi	B	0.00182	0.00063306	2.88	0.0040
Col12	log_dev_width	B	-0.00211	0.00125	-1.69	0.0911
Col13	log_device_volume	B	0.00037884	0.00036509	1.04	0.2994
Col14	log_resolution	B	0.00347	0.00083353	4.17	<.0001
Col15	ldevice_height	0	0	.	.	.

We can see that linear model has more significant values than the log model so we will move forward with Linear model.

Next, selecting variables which are suitable for the model and removing other variables using various methods.

Forward Selection:

The SAS System

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 7).

Effects:	Intercept Col2 Col3 Col4 Col5 Col6 Col7 Col8

Note: The p-values for parameters and effects are not adjusted for the fact that the terms in the model have been selected and so are generally liberal.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	0.59453	0.08493	10.73	<.0001
Error	97064	768.21806	0.00791		
Corrected Total	97071	768.81258			

Root MSE	0.08896
Dependent Mean	0.00798
R-Square	0.0008
Adj R-Sq	0.0007
AIC	-372654
AICC	-372654
BIC	-469726
C(p)	6.90956

BIC	-469726
C(p)	6.90956
PRESS	768.35117
SBC	-469653
ASE	0.00791

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.018336	0.006191	-2.96	0.0031
Col2	1	0.001966	0.000632	3.11	0.0019
Col3	1	-0.014679	0.003814	-3.85	0.0001
Col4	1	0.000021870	0.000004925	4.44	<.0001
Col5	1	0.000019617	0.000004935	3.97	<.0001
Col6	1	-0.000503	0.000107	-4.69	<.0001
Col7	1	-0.000128	0.000094571	-1.35	0.1756
Col8	1	0.000500	0.000128	3.90	<.0001

Backward Selection:

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 3).

Effects:	Intercept Col2 Col3 Col4 Col5 Col6 Col8

Note: The p-values for parameters and effects are not adjusted for the fact that the terms in the model have been selected and so are generally liberal.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	0.58000	0.09667	12.21	<.0001
Error	97065	768.23258	0.00791		
Corrected Total	97071	768.81258			

Root MSE	0.08896
Dependent Mean	0.00798
R-Square	0.0008
Adj R-Sq	0.0007
AIC	-372655
AICC	-372655
BIC	-469727
C(p)	6.74438
PRESS	768.35096
SBC	-469662
ASE	0.00791

AIC	-372655
AICC	-372655
BIC	-469727
C(p)	6.74438
PRESS	768.35096
SBC	-469662
ASE	0.00791

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.018418	0.006191	-2.97	0.0029
Col2	1	0.002032	0.000631	3.22	0.0013
Col3	1	-0.014484	0.003811	-3.80	0.0001
Col4	1	0.000021656	0.000004922	4.40	<.0001
Col5	1	0.000019414	0.000004933	3.94	<.0001
Col6	1	-0.000507	0.000107	-4.73	<.0001
Col8	1	0.000452	0.000123	3.67	0.0002

Stepwise:

The SAS System

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 6).

Effects:	Intercept Col2 Col4 Col6 Col8

Note: The p-values for parameters and effects are not adjusted for the fact that the terms in the model have been selected and so are generally liberal.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	0.45249	0.11312	14.29	<.0001
Error	97067	768.36009	0.00792		
Corrected Total	97071	768.81258			

Root MSE	0.08897
Dependent Mean	0.00798
R-Square	0.0006
Adj R-Sq	0.0005
AIC	-372643
AICC	-372643
BIC	-469715
C(p)	18.85548

Dependent Mean	0.00798
R-Square	0.0006
Adj R-Sq	0.0005
AIC	-372643
AICC	-372643
BIC	-469715
C(p)	18.85548
PRESS	768.44404
SBC	-469669
ASE	0.00792

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.005507	0.001023	5.39	<.0001
Col2	1	0.001834	0.000629	2.92	0.0035
Col4	1	0.000003455	0.000000682	5.06	<.0001
Col6	1	-0.000453	0.000101	-4.50	<.0001
Col8	1	0.000214	0.000103	2.07	0.0382

Best Models:

The REG Procedure
Model: MODEL1
Dependent Variable: install

C(p) Selection Method

Number of Observations Read	97072
Number of Observations Used	97072

Number in Model	C(p)	R-Square	Adjusted R-Square	AIC	BIC	Variables in Model
6	6.7444	0.0008	0.0007	-469728.64	-469726.64	Col2 Col3 Col4 Col5 Col6 Col8
7	6.9096	0.0008	0.0007	-469728.48	-469726.48	Col2 Col3 Col4 Col5 Col6 Col7 Col8
7	8.0383	0.0008	0.0007	-469727.35	-469725.35	Col1 Col2 Col3 Col4 Col5 Col6 Col8
8	8.1444	0.0008	0.0007	-469727.24	-469725.24	Col1 Col2 Col3 Col4 Col5 Col6 Col7 Col8
7	8.3686	0.0008	0.0007	-469727.02	-469725.02	Col2 Col3 Col4 Col5 Col6 Col8 Col9
7	8.3686	0.0008	0.0007	-469727.02	-469725.02	Col2 Col3 Col4 Col5 Col6 Col8 Col10
8	8.7776	0.0008	0.0007	-469726.61	-469724.61	Col2 Col3 Col4 Col5 Col6 Col7 Col8 Col9
8	8.7776	0.0008	0.0007	-469726.61	-469724.61	Col2 Col3 Col4 Col5 Col6 Col7 Col8 Col10
8	9.6378	0.0008	0.0007	-469725.75	-469723.75	Col1 Col2 Col3 Col4 Col5 Col6 Col8 Col9
8	9.6378	0.0008	0.0007	-469725.75	-469723.75	Col1 Col2 Col3 Col4 Col5 Col6 Col8 Col10

From the above methods the best predictors and the final model which involves the following columns:

- Col 2
- Col 3
- Col 4
- Col 5
- Col 6
- Col 8

Finale Linear Model:

The REG Procedure
Model: MODEL1
Dependent Variable: install

Number of Observations Read	97072
Number of Observations Used	97072

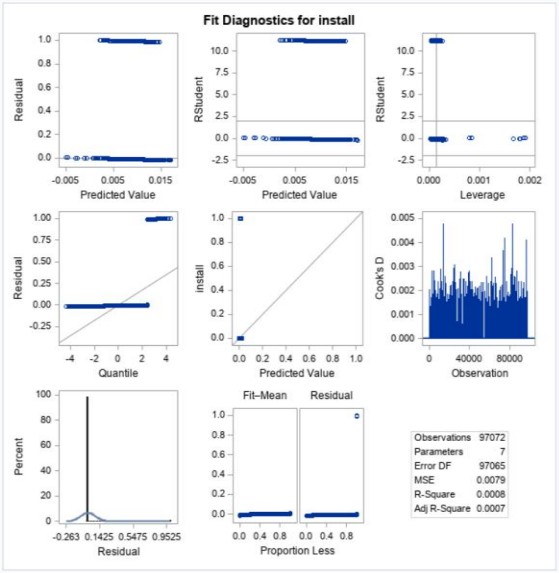
Analysis of Variance

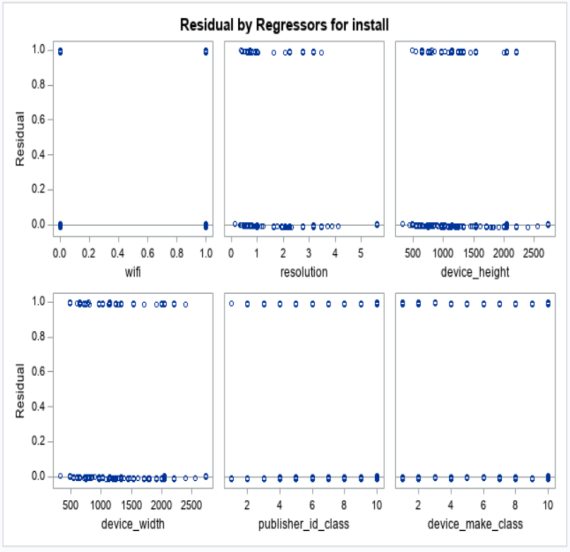
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	0.58000	0.09667	12.21	<.0001
Error	97065	768.23258	0.00791		
Corrected Total	97071	768.81258			

Root MSE	0.08896	R-Square	0.0008
Dependent Mean	0.00798	Adj R-Sq	0.0007
Coeff Var	1114.31328		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-0.01842	0.00619	-2.97	0.0029
Col2	wifi	1	0.00203	0.00063058	3.22	0.0013
Col3	resolution	1	-0.01448	0.00381	-3.80	0.0001
Col4	device_height	1	0.00002166	0.00000492	4.40	<.0001
Col5	device_width	1	0.00001941	0.00000493	3.94	<.0001
Col6	publisher_id_class	1	-0.00050680	0.00010721	-4.73	<.0001
Col8	device_make_class	1	0.00045190	0.00012309	3.67	0.0002





- Although the predicted variables are within probability range, the residuals have to normal which has to be proved using plots.
- The std error estimates will be invalid as the normality assumption is violated and therfore the hypothesis testing on predictor variables wouldn't be valid.

(ii) Logistic Model

Initial Logistic Model:

The SAS System

The LOGISTIC Procedure

Model Information	
Data Set	WORK.AD_TRAINING_WITH_INDICATORS
Response Variable	install
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	97072
Number of Observations Used	97072

Response Profile		
Ordered Value	install	Total Frequency
1	0	96297
2	1	775

Probability modeled is install='1'.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	9032.831	8976.013
SC	9042.314	9070.845
-2 Log L	9030.831	8956.013

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	74.8185	9	<.0001
Score	75.9751	9	<.0001
Wald	75.5739	9	<.0001

Col10 =	Intercept - Col9

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-8.3224	0.7855	112.2422	<.0001
Col1	1	0.1059	0.1176	0.8118	0.3676
Col2	1	0.2827	0.0856	10.9213	0.0010
Col3	1	-1.9060	0.4750	16.1028	<.0001
Col4	1	0.00280	0.000610	21.0548	<.0001
Col5	1	0.00253	0.000617	16.8276	<.0001
Col6	1	-0.0618	0.0135	20.8839	<.0001
Col7	1	-0.0155	0.0124	1.5478	0.2135
Col8	1	0.0632	0.0163	15.0319	0.0001
Col9	1	-0.1657	0.3113	0.2831	0.5947
Col10	0	0	.	.	.

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Col1	1.112	0.883	1.400
Col2	1.327	1.122	1.569
Col3	0.149	0.059	0.377
Col4	1.003	1.002	1.004
Col5	1.003	1.001	1.004
Col6	0.940	0.915	0.965
Col7	0.985	0.961	1.009
Col8	1.065	1.032	1.100
Col9	0.847	0.460	1.560

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	59.3	Somers' D	0.186
Percent Discordant	40.7	Gamma	0.187
Percent Tied	0.0	Tau-a	0.003
Pairs	74630175	c	0.593

Trial	Selection Method	Selection Criteria & Parameters	Number of Predictors	Log-likelihood
1	Stepwise Selection	Significance level: - Entry value: 0.25 Stay value: 0.35	8	8973.616

2	Forward Selection	Significance level: - Entry value: 0.25	8	8957.097
3	Backward Selection	Significance level: - Stay value: 0.35	8	8956.306

Stepwise:

Summary of Stepwise Selection

Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq	Variable Label
1	Col3		1	1	26.5971		<.0001	resolution
2	Col2		1	2	10.7420		0.0010	wifi
3	Col6		1	3	11.0515		0.0009	publisher_id_class
4	Col4		1	4	5.8497		0.0156	device_height
5		Col3	1	3		0.7046	0.4012	resolution
6	Col8		1	4	3.6922		0.0547	device_make_class
7	Col7		1	5	1.4768		0.2243	device_os_class

Partition for the Hosmer and Lemeshow Test

Group	Total	install = 1 Observed	install = 1 Expected	install = 0 Observed	install = 0 Expected
1	9699	46	48.53	9653	9650.47
2	9749	54	57.50	9695	9691.50
3	9701	63	61.68	9638	9639.32
4	9658	61	67.25	9597	9590.75
5	9573	67	70.96	9506	9502.04
6	9735	84	75.79	9651	9659.21
7	9601	91	79.53	9510	9521.47
8	9712	96	87.24	9616	9624.76
9	9390	84	96.20	9306	9293.80
10	10254	129	130.33	10125	10123.67

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
6.2089	8	0.6238

Final Logistic Model:

(i)Estimation of the model without considering rare event

The SAS System

The LOGISTIC Procedure

Model Information	
Data Set	WORK.AD_TRAINING_WITH_INDICATORS
Response Variable	install
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	84938
Number of Observations Used	84938

Response Profile		
Ordered Value	install	Total Frequency
1	0	84258
2	1	680

Probability modeled is install='1'.

Probability modeled is install='1'.

Class Level Information		
Class	Value	Design Variables
device_platform_class	android	1
	iOS	-1

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	11667.493	11594.683
SC	11677.199	11691.746
-2 Log L	11665.493	11574.683

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	90.8101	9	<.0001
Score	91.2357	9	<.0001
Wald	90.8005	9	<.0001

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
device_volume	1	1.6476	0.1993
wifi	1	18.4964	<.0001
resolution	1	20.8792	<.0001
device_height	1	26.2606	<.0001
device_width	1	22.1841	<.0001
publisher_id_class	1	17.0330	<.0001
device_os_class	1	2.3397	0.1261
device_make_class	1	19.3637	<.0001
device_platform_clas	1	0.8347	0.3609

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-8.5095	0.7386	132.7411	<.0001
device_volume		1	0.1323	0.1030	1.6476	0.1993
wifi		1	0.3255	0.0757	18.4964	<.0001
resolution		1	-1.9094	0.4179	20.8792	<.0001
device_height		1	0.00275	0.000537	26.2606	<.0001
device_width		1	0.00255	0.000542	22.1841	<.0001
publisher_id_class		1	-0.0490	0.0119	17.0330	<.0001
device_os_class		1	-0.0167	0.0109	2.3397	0.1261
device_make_class		1	0.0632	0.0144	19.3637	<.0001
device_platform_clas	android	1	-0.1276	0.1396	0.8347	0.3609

device_width		1	0.00255	0.000542	22.1841	<.0001
publisher_id_class		1	-0.0490	0.0119	17.0330	<.0001
device_os_class		1	-0.0167	0.0109	2.3397	0.1261
device_make_class		1	0.0632	0.0144	19.3637	<.0001
device_platform_clas	android	1	-0.1276	0.1396	0.8347	0.3609

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
device_volume	1.141	0.933	1.397
wifi	1.385	1.194	1.606
resolution	0.148	0.065	0.336
device_height	1.003	1.002	1.004
device_width	1.003	1.001	1.004
publisher_id_class	0.952	0.930	0.975
device_os_class	0.983	0.963	1.005
device_make_class	1.065	1.036	1.096
device_platform_clas android vs iOS	0.775	0.448	1.339

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	59.1	Somers' D	0.181
Percent Discordant	40.9	Gamma	0.181
Percent Tied	0.0	Tau-a	0.003
Pairs	121293648	c	0.591

Most of the variables are significant to the 0.01% level.

We do not need to compare the number of rare events in this case, because the number of rare events is 1008 in the full sample and 680 in the training data which is high.

The expected number of rare events should be around 20 for each independent variable which is 10 in this case.

Therefore 200 rare events would have been the optimal count. Since 200<680, the modelling of rare events is not required.

(ii) Estimation of the model considering rare events using oversampling approach to handle rare events and also applying correction mechanism to correct for intercept values.

Using the proc freq statement to get the response counts in the full and over sampled dataset which is the number of rare events.

response counts in full data set

The FREQ Procedure

install	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	120331	99.17	120331	99.17
1	1008	0.83	121339	100.00

Response counts in oversampled, subset data set

The FREQ Procedure

install	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	1070	51.49	1070	51.49
1	1008	48.51	2078	100.00

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	2880.870	2833.721
SC	2886.509	2890.112
-2 Log L	2878.870	2813.721

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	65.1487	9	<.0001
Score	64.2118	9	<.0001
Wald	62.3173	9	<.0001

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
device_volume	1	1.9548	0.1621
wifi	1	13.5132	0.0002
resolution	1	4.1789	0.0409
device_height	1	5.8619	0.0155
device_width	1	5.3726	0.0205
publisher_id_class	1	15.3111	<.0001
device_os_class	1	2.5091	0.1132
device_make_class	1	10.2265	0.0014
device_platform_clas	1	0.0563	0.8124

Applying the formula off=log((r1*(1-p1)) / ((1-r1)*p1)). to correct for the intercept values.

Unadjusted model:

In the next step, we run the logistic procedure on the oversampled dataset if the model remains unadjusted, i.e. how the intercept and co-efficient of the predictors change when the model is adjusted to handle the rare events but without performing the necessary corrections. The results are as follows,

Here we don't apply the correction, our intercept values deviate from each other.

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.6886	1.0322	6.7847	0.0092
device_volume		1	0.2032	0.1453	1.9548	0.1621
wifi		1	0.3747	0.1019	13.5132	0.0002
resolution		1	-1.2245	0.5990	4.1789	0.0409
device_height		1	0.00188	0.000776	5.8619	0.0155
device_width		1	0.00179	0.000773	5.3726	0.0205
publisher_id_class		1	-0.0665	0.0170	15.3111	<.0001
device_os_class		1	-0.0247	0.0156	2.5091	0.1132
device_make_class		1	0.0654	0.0205	10.2265	0.0014
device_platform_clas	android	1	-0.0462	0.1945	0.0563	0.8124

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
device_volume	1.225	0.922	1.629
wifi	1.455	1.191	1.776
resolution	0.294	0.091	0.951
device_height	1.002	1.000	1.003
device_width	1.002	1.000	1.003
publisher_id_class	0.936	0.905	0.967
device_os_class	0.976	0.946	1.006
device_make_class	1.068	1.026	1.111
device_platform_clas android vs iOS	0.912	0.425	1.955

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	60.2	Somers' D	0.204
Percent Discordant	39.8	Gamma	0.204
Percent Tied	0.0	Tau-a	0.102
Pairs	1078560	c	0.602

From the screenshot above, we observe that without the necessary corrections, the intercept differs significantly from the intercept of the original model. Hence, the unadjusted model should not be considered for the final model selection.

Weight adjusted model:

In the next step, we run the oversampled dataset using the weight adjusted model which yields better results compared to the unadjusted model as shown below,

The weight- adjusted model applies the correction which brings the intercept values closer.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	201.778	217.613
SC	207.417	274.005
-2 Log L	199.778	197.613

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2.1652	9	0.9886
Score	2.2225	9	0.9874
Wald	2.1953	9	0.9880

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
device_volume	1	0.0304	0.8616
wifi	1	0.4673	0.4942
resolution	1	0.1500	0.6985
device_height	1	0.2032	0.6521
device_width	1	0.1894	0.6634
publisher_id_class	1	0.5071	0.4764
device_os_class	1	0.0801	0.7772
device_make_class	1	0.3692	0.5434

device_make_class	1	0.3692	0.5434
device_platform_clas	1	0.0140	0.9060

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-7.5213	5.8428	1.6571	0.1980
device_volume		1	0.1351	0.7747	0.0304	0.8616
wifi		1	0.3933	0.5753	0.4673	0.4942
resolution		1	-1.2739	3.2893	0.1500	0.6985
device_height		1	0.00190	0.00422	0.2032	0.6521
device_width		1	0.00186	0.00427	0.1894	0.6634
publisher_id_class		1	-0.0646	0.0907	0.5071	0.4764
device_os_class		1	-0.0242	0.0854	0.0801	0.7772
device_make_class		1	0.0669	0.1101	0.3692	0.5434
device_platform_clas	android	1	-0.1283	1.0862	0.0140	0.9060

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
device_volume	1.145	0.251	5.225
wifi	1.482	0.480	4.577

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
device_volume	1.145	0.251	5.225
wifi	1.482	0.480	4.577
resolution	0.280	<0.001	176.429
device_height	1.002	0.994	1.010
device_width	1.002	0.994	1.010
publisher_id_class	0.937	0.785	1.120
device_os_class	0.976	0.826	1.154
device_make_class	1.069	0.862	1.327
device_platform_clas android vs iOS	0.774	0.011	54.667

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	60.1	Somers' D	0.203
Percent Discordant	39.8	Gamma	0.203
Percent Tied	0.0	Tau-a	0.101
Pairs	1078560	c	0.601

Offset adjusted model:
As a secondary approach, we also run the offset adjusted model which yielded the following results,

Criterion	Intercept Only	Covariates
AIC	2880.870	2833.721
SC	2886.509	2890.112
-2 Log L	2878.870	2813.721

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	65.1487	9	<.0001
Score	64.2117	9	<.0001
Wald	62.3245	9	<.0001

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
device_volume	1	1.9549	0.1621
wifi	1	13.5149	0.0002
resolution	1	4.1795	0.0409
device_height	1	5.8627	0.0155
device_width	1	5.3734	0.0204
publisher_id_class	1	15.3132	<.0001
device_os_class	1	2.5096	0.1132
device_make_class	1	10.2279	0.0014
device_platform_clas	1	0.0563	0.8124

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-7.4113	1.0322	51.5558	<.0001
device_volume		1	0.2032	0.1453	1.9549	0.1621
wifi		1	0.3748	0.1019	13.5149	0.0002
resolution		1	-1.2246	0.5990	4.1795	0.0409
device_height		1	0.00188	0.000776	5.8627	0.0155
device_width		1	0.00179	0.000773	5.3734	0.0204
publisher_id_class		1	-0.0665	0.0170	15.3132	<.0001
device_os_class		1	-0.0247	0.0156	2.5096	0.1132
device_make_class		1	0.0654	0.0205	10.2279	0.0014
device_platform_clas	android	1	-0.0462	0.1945	0.0563	0.8124
off		0	1.0000	0	.	.

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
device_volume	1.225	0.922	1.629
wifi	1.455	1.191	1.776
resolution	0.294	0.091	0.951
device_height	1.002	1.000	1.003
device_width	1.002	1.000	1.003
publisher_id_class	0.936	0.905	0.967
device_os_class	0.976	0.946	1.006
device_make_class	1.068	1.026	1.111

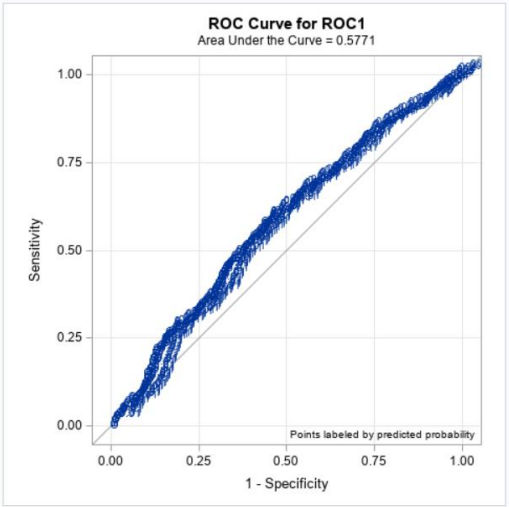
wifi	1.455	1.191	1.776
resolution	0.294	0.091	0.951
device_height	1.002	1.000	1.003
device_width	1.002	1.000	1.003
publisher_id_class	0.936	0.905	0.967
device_os_class	0.976	0.946	1.006
device_make_class	1.068	1.026	1.111
device_platform_clas android vs iOS	0.912	0.425	1.955

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	60.2	Somers' D	0.204
Percent Discordant	39.8	Gamma	0.204
Percent Tied	0.0	Tau-a	0.102
Pairs	1078560	c	0.602

The next step is to plot the ROC curves for the initial and final linear models and logistic regression models.

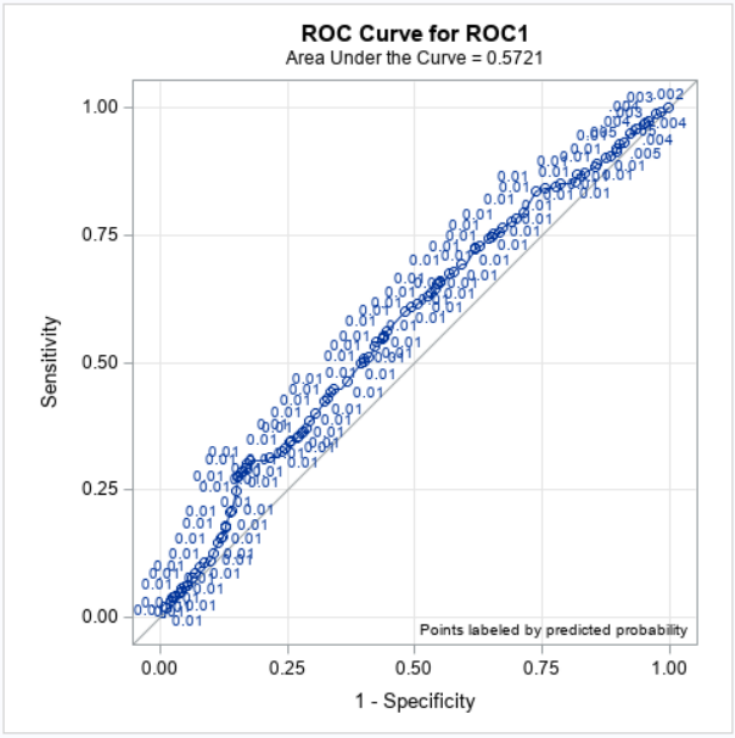
Initial Linear Model:

All the 10 predictors were used to get the below ROC.



ROC Association Statistics							
	Mann-Whitney						
ROC Model	Area	Standard Error	95% Wald Confidence Limits		Somers' D	Gamma	Tau-a
ROC1	0.5771	0.0186	0.5407	0.6135	0.1542	0.1542	0.00293

Final Linear Model:
Here getting the ROC with 6 predictors

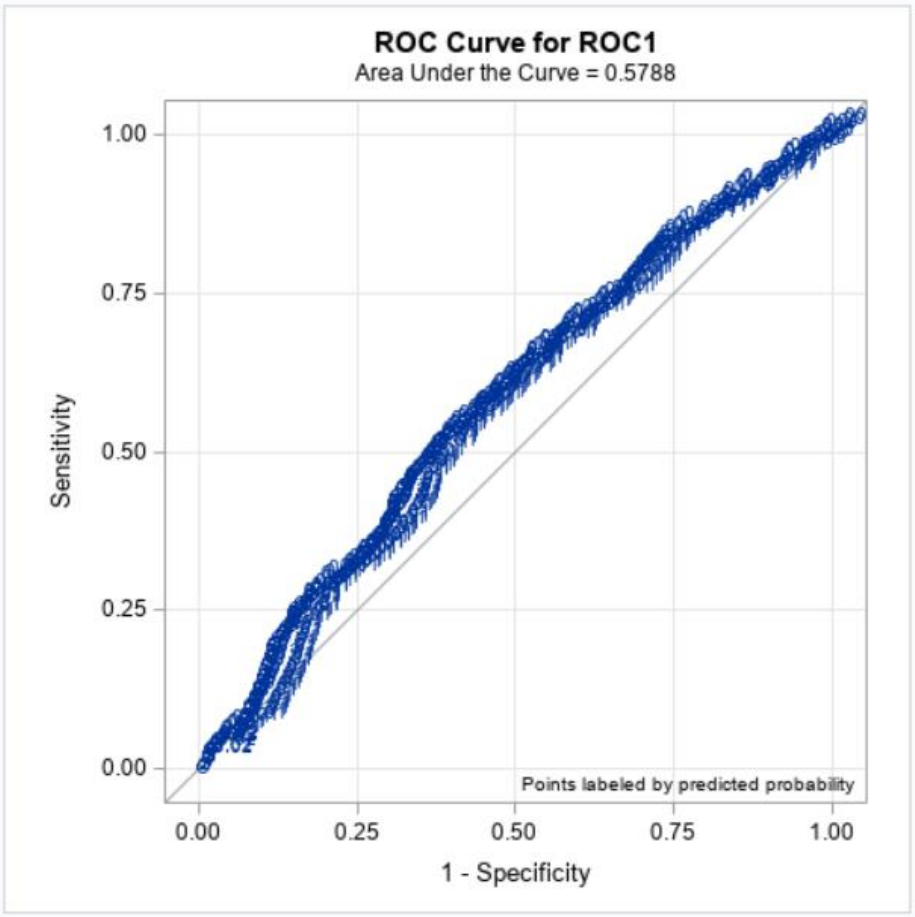


ROC Association Statistics							
	Mann-Whitney						
ROC Model	Area	Standard Error	95% Wald Confidence Limits		Somers' D	Gamma	Tau-a
ROC1	0.5721	0.0189	0.5350	0.6091	0.1441	0.1456	0.00274

From both these results we see that ROC curve looks similar and the area under the curve remains almost the same. However, it should be noted that the final model was able of reach the same Area under the curve value inspite of using less predictors.

Initial Logistic Model:

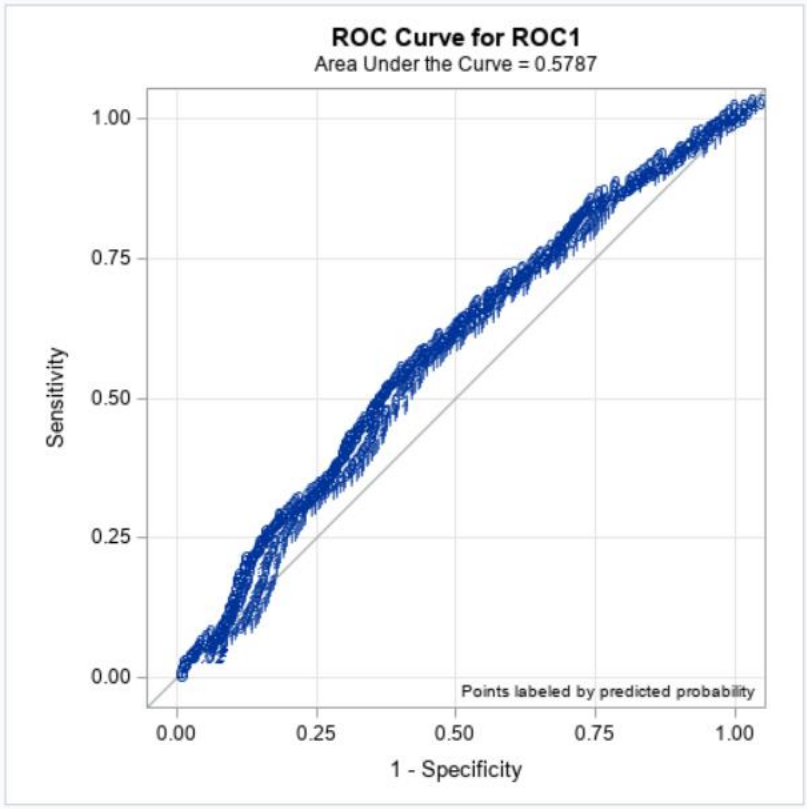
Using 10 predictors to get the below ROC Curve



ROC Association Statistics							
	Mann-Whitney						
ROC Model	Area	Standard Error	95% Wald Confidence Limits		Somers' D	Gamma	Tau-a
ROC1	0.5788	0.0186	0.5424	0.6152	0.1576	0.1576	0.00300

Final Logistic Model:

Using 8 predictors to get the below ROC Curve



ROC Association Statistics							
	Mann-Whitney						
ROC Model	Area	Standard Error	95% Wald Confidence Limits		Somers' D	Gamma	Tau-a
ROC1	0.5787	0.0185	0.5424	0.6151	0.1574	0.1575	0.00299

Here, the initial and the final logistic models both have ROC curve that looks similar and the area under the curve remains the same.

Conclusion:

From the above results of the ROC Curve; we conclude that the AUC value of Final Logistic Regression = 0.5787 and 95% Conf Interval values (0.5424, 0.6151) is higher than the Final Linear Probability Model with AUC = 0.5721 and Conf Interval values (0.5350, 0.6091). So we will use Logistic Regression Model for prediction.

Part 2:

The objective is to decide on a threshold based on the ROC table such that if the probability of installing the ad is above that threshold, the ad is shown to the consumer.

First, we need to calculate the total expected cost as follows:
Total expected cost = # False positives*False positive cost + # False negatives*False negative cost

The two possible situations where the ad company can incur a loss is as follows:

False positive- The platform shows an ad to a consumer but the consumer ends up not installing an app. The loss is estimated to be 1 cent (0.01$)

False negative- The platform fails to show an ad where the consumer actually would have installed the app. The loss here is 1$.

Logistic regression models

The proc logistic statement is used to create a ROC table for both the initial and final models. The total cost column is created using the above formula.

The ROC table for the Initial logistic model is as follows:

	Probability Level	No. of Correctly Predicted Events	No. of Correctly Predicted Nonevents	No. of Nonevents Predicted as Events	No. of Events Predicted as Nonevents	Sensitivity	1 - Specificity	total_cost1	min_cost
1	0.0252531503	0	24033	1	233	0	0.0000416077	233.01	202.95
2	0.0236611529	0	24032	2	233	0	0.0000832154	233.02	202.95
3	0.0232245689	0	24031	3	233	0	0.0001248232	233.03	202.95
4	0.0230569184	0	24030	4	233	0	0.0001664309	233.04	202.95
5	0.023033065	0	24029	5	233	0	0.0002080386	233.05	202.95
6	0.0211677607	0	24028	6	233	0	0.0002496463	233.06	202.95
7	0.0210364463	0	24027	7	233	0	0.0002912541	233.07	202.95
8	0.0209452671	0	24026	8	233	0	0.0003328618	233.08	202.95
9	0.0207593635	0	24025	9	233	0	0.0003744695	233.09	202.95
10	0.0204641685	0	24024	10	233	0	0.0004160772	233.1	202.95
11	0.0201020943	0	24023	11	233	0	0.0004576849	233.11	202.95
12	0.0200334027	0	24022	12	233	0	0.0004992927	233.12	202.95
13	0.0198860642	0	24021	13	233	0	0.0005409004	233.13	202.95
14	0.0198199204	0	24020	14	233	0	0.0005825081	233.14	202.95
15	0.0197582658	0	24019	15	233	0	0.0006241158	233.15	202.95
16	0.0194450099	0	24018	16	233	0	0.0006657236	233.16	202.95
17	0.0193241716	0	24016	18	233	0	0.000748939	233.18	202.95
18	0.0192923979	0	24015	19	233	0	0.0007905467	233.19	202.95
19	0.019251734	0	24014	20	233	0	0.0008321544	233.2	202.95
20	0.0192040696	0	24013	21	233	0	0.0008737622	233.21	202.95
21	0.0191529398	0	24012	22	233	0	0.0009153699	233.22	202.95
22	0.0190846995	0	24011	23	233	0	0.0009569776	233.23	202.95
23	0.0189463537	0	24008	26	233	0	0.0010818008	233.26	202.95
24	0.0188208048	0	24007	27	233	0	0.0011234085	233.27	202.95
25	0.0187882403	0	24006	28	233	0	0.0011650162	233.28	202.95
26	0.01864659	0	24005	29	233	0	0.0012066239	233.29	202.95
27	0.0185306196	0	24004	30	233	0	0.0012482317	233.3	202.95
28	0.0184035198	0	24003	31	233	0	0.0012898394	233.31	202.95
29	0.0183140996	0	24002	32	233	0	0.0013314471	233.32	202.95
30	0.0183007979	0	24001	33	233	0	0.0013730548	233.33	202.95
31	0.0183004455	0	24000	34	233	0	0.0014146626	233.34	202.95
32	0.0182055159	0	23999	35	233	0	0.0014562703	233.35	202.95
33	0.0181680295	0	23998	36	233	0	0.001497878	233.36	202.95
34	0.0180587312	0	23997	37	233	0	0.0015394857	233.37	202.95
35	0.018047752	0	23996	38	233	0	0.0015810935	233.38	202.95
36	0.0180041274	0	23995	39	233	0	0.0016227012	233.39	202.95

Min Cost: $202.95
The probability threshold: 0.0081

The ROC table for the Final logistic model is as follows:

	Probability Level	No. of Correctly Predicted Events	No. of Correctly Predicted Nonevents	No. of Nonevents Predicted as Events	No. of Events Predicted as Nonevents	Sensitivity	1 - Specificity	total_cost2	min_cost
1	0.0248156736	0	24033	1	233	0	0.0000416077	233.01	202
2	0.0232689367	0	24032	2	233	0	0.0000832154	233.02	202
3	0.0228445682	0	24031	3	233	0	0.0001248232	233.03	202
4	0.0226815853	0	24030	4	233	0	0.0001664309	233.04	202
5	0.0226583949	0	24029	5	233	0	0.0002080386	233.05	202
6	0.0207567506	0	24028	6	233	0	0.0002496463	233.06	202
7	0.0206294853	0	24027	7	233	0	0.0002912541	233.07	202
8	0.0204772813	0	24026	8	233	0	0.0003328618	233.08	202
9	0.0202660656	0	24025	9	233	0	0.0003744695	233.09	202
10	0.0202412209	0	24024	10	233	0	0.0004160772	233.1	202
11	0.0200567633	0	24023	11	233	0	0.0004576849	233.11	202
12	0.019672627	0	24022	12	233	0	0.0004992927	233.12	202
13	0.0196317452	0	24021	13	233	0	0.0005409004	233.13	202
14	0.0196121605	0	24020	14	233	0	0.0005825081	233.14	202
15	0.0194534912	0	24019	15	233	0	0.0006241158	233.15	202
16	0.0191153066	0	24018	16	233	0	0.0006657236	233.16	202
17	0.0190065185	0	24017	17	233	0	0.0007073313	233.17	202
18	0.0189973771	0	24016	18	233	0	0.000748939	233.18	202
19	0.0188806911	0	24014	20	233	0	0.0008321544	233.2	202
20	0.018838171	0	24013	21	233	0	0.0008737622	233.21	202
21	0.0188242218	0	24010	24	233	0	0.0009985853	233.24	202
22	0.0187952651	0	24009	25	233	0	0.0010401931	233.25	202
23	0.0187647081	0	24008	26	233	0	0.0010818008	233.26	202
24	0.018649424	0	24007	27	233	0	0.0011234085	233.27	202
25	0.0186052642	0	24006	28	233	0	0.0011650162	233.28	202
26	0.0185636056	0	24005	29	233	0	0.0012066239	233.29	202
27	0.0185158039	0	24002	32	233	0	0.0013314471	233.32	202
28	0.0184495338	0	24001	33	233	0	0.0013730548	233.33	202
29	0.0182234498	0	24000	34	233	0	0.0014146626	233.34	202
30	0.0181967537	0	23998	36	233	0	0.001497878	233.36	202
31	0.0181263387	0	23997	37	233	0	0.0015394857	233.37	202
32	0.0180928254	0	23996	38	233	0	0.0015810935	233.38	202
33	0.0180413772	0	23995	39	233	0	0.0016227012	233.39	202
34	0.0180334301	0	23994	40	233	0	0.0016643089	233.4	202

The min cost= $202
The probability threshold= 0.00821

Linear Probability Model:

We have to create ROC tables manually for each threshold values.

False positive- if install=0 and predicted=1 then false_pos=1

False negative- if install=1 and predicted=0 then false_neg=1

Initial Linear Model:

ROC table for initial linear probability model

	probability	false_positive	false_negative	total_cost
1	0.001	23995	0	239.95
2	0.005	21428	0	214.28
3	0.01	4776	0	47.76
4	0.015	45	0	0.45
5	0.02	0	0	0
6	0.025	0	0	0
7	0.03	0	0	0
8	0.035	0	0	0
9	0.04	0	0	0
10	0.045	0	0	0
11	0.05	0	0	0

Min cost= 214.28$

The probability threshold= 0.005

Final Linear Probability Model:

We follow the same procedure as we did for the Initial Linear Model:

	probability	false_positive	false_negative	total_cost
1	0.001	24021	0	240.21
2	0.005	31428	0	314.28
3	0.01	4776	0	47.76
4	0.015	45	0	0.45
5	0.02	0	0	0
6	0.025	0	0	0
7	0.03	0	0	0
8	0.035	0	0	0
9	0.04	0	0	0
10	0.045	0	0	0
11	0.05	0	0	0

The min cost= 240.21$

The probability threshold= 0.001

From the above results we can draw the following table:

Model	**Probability Threshold**	**Minimum total cost**
Initial logistic regression model	0.0081	$202.95
Initial linear probability model	0.005	$214.28
Final logistic regression model	0.00821	$202
Final linear probability model	0.001	$240.21

Therefore, from the results above we can say Logistic Model provide the lowest cost at a probability level 0.0081.