Project: Analyzing the effectiveness of Ads through app installation

<u>Mobile Ads dataset</u> (Source: mobilead.sas7bdat)

For this project, the above-mentioned dataset is used to develop two models namely – A logistic regression model and a linear probability model. The project aims to present the model evolution, factors considered, and the results obtained for each model. A brief overview of the dataset is presented before moving to the modelling part.

The source dataset is read into SAS. The dataset contains data about ads from an advertiser over various publishers (apps). Each observation represents an ad shown to a consumer on a publisher app. The dataset has 1,21,339 observations. The output variable is 'install' which is binary in nature. There are 9 predictor variables - 4 categorical, 4 numerical and 1 binary. All the categorical variables are considered for the analysis including the *publisher_id_class* as it doesn't truly represent an ID as it is not unique across all the observations.

The following are some exploratory analysis before the modelling process.

<u>CORRELATION ESTIMATION</u>

The correlations are determined prior to the modelling part as we wouldn't any highly correlated variables as it might hinder our modelling process giving us larger variances and incorrect estimates. Hence, using PROC CORR the correlations are estimated. The results are captured through the below screenshot. Please refer to the SAS code file for the coding.

The CORR Procedure

5 Variables:	device_volume wifi resolution device_height device_width

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
device_volume	121339	0.53488	0.30936	64902	0.01000	1.06000
wifi	121339	0.68106	0.46607	82639	0	1.00000
resolution	121339	1.37863	0.95955	167282	0.15360	5.59514
device_height	121339	1126	447.73521	136631037	320.00000	2732
device_width	121339	1174	484.02271	142449708	320.00000	2732

Pearson Correlation Coefficients, N = 121339
Prob > |r| under H0: Rho=0

	device_volume	wifi	resolution	device_height	device_width
device_volume	1.00000	-0.09792 <.0001	-0.01071 0.0002	-0.01301 <.0001	0.00199 0.4883
wifi	-0.09792 <.0001	1.00000	0.12840 <.0001	0.14062 <.0001	0.04552 <.0001
resolution	-0.01071 0.0002	0.12840 <.0001	1.00000	0.76585 <.0001	0.81404 <.0001
device_height	-0.01301 <.0001	0.14062 <.0001	0.76585 <.0001	1.00000	0.26162 <.0001
device_width	0.00199 0.4883	0.04552 <.0001	0.81404 <.0001	0.26162 <.0001	1.00000

It can be observed that significant correlations are: (*resolution, device_width)* – 81% (significant) , (*resolution, device_height)* – 76% (significant). The correlations aren't significantly high. Hence, no variables are omitted at this step.

TARGET VARIABLE ANALYSIS

Next, we analyze the target variable – *install* for checking the frequency of each class (class 0 and class 1) as we would want a balanced set of the output classes for effective modelling. This is done using PROC FREQ and the results are captured using below screenshot.

The FREQ Procedure

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

We see that the event=1 is present only for 0.83%. Hence, we don't have a balanced set of the output classes. And since, the desired event is occurring in very small percent, this is termed as rare event modelling.

MISSING OBSERVATIONS ANALYSIS

The dataset's predictor variables are analyzed for any missing values using PROC FREQ. Missing values might lead to a reduced no. of observations being considered for the regression and hence leads to a loss of predictive power and sample size. A snipper of the results is presented below as screenshots. It can be inferred that the predictors don't have any missing values.

The FREQ Procedure

publisher_id_class	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	135	0.11	135	0.11
2	9987	8.23	10122	8.34
3	9834	8.10	19956	16.45
4	8694	7.17	28650	23.61
5	6746	5.56	35396	29.17
6	5121	4.22	40517	33.39
7	5100	4.20	45617	37.59
8	4903	4.04	50520	41.64
9	3854	3.18	54374	44.81
10	66965	55.19	121339	100.00

device_os_class	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	51279	42.26	51279	42.26
2	14993	12.36	66272	54.62
3	10989	9.06	77261	63.67
4	10762	8.87	88023	72.54
5	5082	4.19	93105	76.73
6	3691	3.04	96796	79.77
7	3456	2.85	100252	82.62
8	3399	2.80	103651	85.42
9	2244	1.85	105895	87.27
10	15444	12.73	121339	100.00

device_make_class	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	34136	28.13	34136	28.13
2	25428	20.96	59564	49.09
3	12430	10.24	71994	59.33
4	9927	8.18	81921	67.51
5	8100	6.68	90021	74.19
6	7824	6.45	97845	80.64
7	7235	5.96	105080	86.60
8	4287	3.53	109367	90.13
9	3747	3.09	113114	93.22
10	8225	6.78	121339	100.00

wifi	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	38700	31.89	38700	31.89
1	82639	68.11	121339	100.00

platform	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	119274	98.30	119274	98.30
1	2065	1.70	121339	100.00

After all the preliminary analysis done above, we move forward with a sense of the data we are going to model and with a good hang of how we are going to model.

Part I. Logistic Regression Analysis

- Prior to the modelling step, the dataset is split into training and test set for model building, model evaluation purposes. The dataset us split using PROC SURVEYSELECT and the parameters used are seed = 10 and sampling ratio of 80%. Hence, the training set consists of 97072 (0.8*121339) observations. The result of this step is available through below screenshot.

The SURVEYSELECT Procedure

Selection Method	Simple Random Sampling

Input Data Set	MOBILEAD
Random Number Seed	10
Sampling Rate	0.8
Sample Size	97072
Selection Probability	0.800007
Sampling Weight	0
Output Data Set	MOBILEAD_SAMPLED

- One more important point to be noted is, since this is a case of rare-event model modelling, I am using FIRTH's penalized MLE which is available in SAS for my logistic regression analysis. The coding can be referred in SAS code file.

1. The final best model as per my analysis consists of a mix of interaction terms and quadratic terms. The model is presented below.

$$
\begin{aligned}
BX = {} & B_0 + B_1.(device\_volume) + B_2.(resolution) + B_3.(wifi) + B_4.(device\_width) \\
& + B_5.(device\_height) + B_6.p1 + B_7.p2 + \cdots + B_{14}.p9 + B_{15}.m1 + \cdots \\
& + B_{23}.m9 + B_{24}.o1 + ..+ B_{32}.o9 + B_{33}.(resolution * device\_height) \\
& + B_{34}.(resolution * device\_width) + B_{35}.(resolution * resolution) \\
& + B_{36}.(device\_width * device\_height) \\
& + B_{37}.(device\_width * device\_width)
\end{aligned}
$$

$$install = 1/(1 + e^{-BX})$$

This best model is arrived at through a series of experimentation and the evolution is presented below. For this model, as reported above, we have the following variables in use:

Output variable(1): install

Input variables(37):

- Original variables: device_volume, resolution, wifi, device_width, device_height
  p1 – p9, m1 – m9, o1 – o9 {p, m, o -> one hot encoded variable for publisher_id_class, device_make_class, device_os_class }
- Interaction terms: resolution*device_width, resolution*device_height
- Quadratic terms: resolution*resolution, device_height*device_height, device_weight*device_weight

For this model, the dataset is read into SAS. The categorical variables are specified as class variables and the interaction terms and quadratic terms are specified. The model is built using PROC LOGISTIC and the modelling results are presented using below screenshot. The best model is titled **'Model 4'.** Please refer to the SAS code file for the implementation process.

Model 4: Using quadratic terms

The LOGISTIC Procedure

Model Information	
Data Set	WORK.TRAIN_MOBILEAD
Response Variable	install
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring
Likelihood Penalty	Firth's bias correction

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'.

Class Level Information										
Class	Value	Design Variables								
publisher_id_class	1	1	0	0	0	0	0	0	0	0
	2	0	1	0	0	0	0	0	0	0
	3	0	0	1	0	0	0	0	0	0
	4	0	0	0	1	0	0	0	0	0
	5	0	0	0	0	1	0	0	0	0
	6	0	0	0	0	0	1	0	0	0
	7	0	0	0	0	0	0	1	0	0
	8	0	0	0	0	0	0	0	1	0
	9	0	0	0	0	0	0	0	0	1
	10	-1	-1	-1	-1	-1	-1	-1	-1	-1
device_make_class	1	1	0	0	0	0	0	0	0	0
	2	0	1	0	0	0	0	0	0	0
	3	0	0	1	0	0	0	0	0	0
	4	0	0	0	1	0	0	0	0	0
	5	0	0	0	0	1	0	0	0	0
	6	0	0	0	0	0	1	0	0	0
	7	0	0	0	0	0	0	1	0	0
	8	0	0	0	0	0	0	0	1	0
	9	0	0	0	0	0	0	0	0	1
	10	-1	-1	-1	-1	-1	-1	-1	-1	-1
device_os_class	1	1	0	0	0	0	0	0	0	0
	2	0	1	0	0	0	0	0	0	0
	3	0	0	1	0	0	0	0	0	0
	4	0	0	0	1	0	0	0	0	0
	5	0	0	0	0	1	0	0	0	0
	6	0	0	0	0	0	1	0	0	0
	7	0	0	0	0	0	0	1	0	0
	8	0	0	0	0	0	0	0	1	0
	9	0	0	0	0	0	0	0	0	1
	10	-1	-1	-1	-1	-1	-1	-1	-1	-1

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

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

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	8784.199	8655.864
SC	8793.682	9025.709
-2 Log L	8782.199	8577.864

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	204.3348	38	<.0001
Score	214.8087	38	<.0001
Wald	214.9565	38	<.0001

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
device_volume	1	1.0429	0.3071
wifi	1	7.9132	0.0049
resolution	1	2.7117	0.0996
device_height	1	3.6524	0.0560
device_width	1	1.6998	0.1923
publisher_id_class	9	100.1913	<.0001
device_os_class	9	20.4763	0.0152
device_make_class	9	41.3978	<.0001
platform	1	4.4319	0.0353
resolutio*device_hei	1	2.9720	0.0847
resolutio*device_wid	1	1.0107	0.3147
resolutio*resolution	1	1.5896	0.2074
device_he*device_hei	1	3.2892	0.0697
device_wi*device_wid	1	0.1563	0.6926

Interpreting the model results, firstly we look at the Chi-square test which tests if the model is statistically significant or not. The hypothesis is provided below:
Null hypothesis (H_0): All beta coefficients are equal to zero
Alternate hypothesis (H_1): At least one of the beta coefficients is not zero
The chi-square test records a high test-statistic of 204.33 and a p-value<0.0001 which is less than 0.05 for 95%CI. Hence, we fail to reject the null and conclude that **built regression model is statistically significant**. Going forward, we see that the model reports significantly lower values of **AIC = 8655, BIC = 9025**.
The coefficients of the model are interpreted in the following way.

- The *wifi* variable is significant at 95%CI. The presence of wifi increases log of odds of consumer installing the app by 0.2432 units.
- The *publisher_id_class* of 1 increases the odds of the app getting installed by e^1.4703 times as compared to *publisher_id_class* of 10.
- The *platform* variable suggests that Android platform has reduced probability of app getting installed as compared to iOS platform.
- The odds of the app getting installed when resolution is increased by 1 unit, depends on *device_height, device_width* etc.

All the above interpretations involve keeping the other variables that are not involved in the interpretation, constant. Since there are 37 variables, I have interpreted only a few. The interpretation of the other variables is similar.

Analysis of Penalized Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-24.1400	11.0913	4.7371	0.0295
device_volume		1	0.1192	0.1167	1.0429	0.3071
wifi		1	0.2432	0.0865	7.9132	0.0049
resolution		1	-41.5321	25.2209	2.7117	0.0996
device_height		1	0.0322	0.0168	3.6524	0.0560
device_width		1	0.0211	0.0162	1.6998	0.1923
publisher_id_class	1	1	1.4703	0.5806	6.4139	0.0113
publisher_id_class	2	1	0.3485	0.1464	5.6686	0.0173
publisher_id_class	3	1	0.8576	0.1265	45.9645	<.0001
publisher_id_class	4	1	-0.0340	0.1606	0.0448	0.8325
publisher_id_class	5	1	0.1470	0.1628	0.8160	0.3664
publisher_id_class	6	1	-0.2989	0.2274	1.7284	0.1886
publisher_id_class	7	1	-1.0186	0.2430	17.5759	<.0001
publisher_id_class	8	1	-0.9707	0.2579	14.1647	0.0002
publisher_id_class	9	1	-0.3486	0.2212	2.4833	0.1151
device_os_class	1	1	0.1648	0.0852	3.7456	0.0529
device_os_class	2	1	0.0992	0.1201	0.6833	0.4085
device_os_class	3	1	-0.0971	0.1433	0.4590	0.4981
device_os_class	4	1	0.2626	0.1181	4.9466	0.0261
device_os_class	5	1	0.3165	0.1659	3.6397	0.0564
device_os_class	6	1	0.2259	0.1812	1.5547	0.2124
device_os_class	7	1	-0.8110	0.2976	7.4287	0.0064
device_os_class	8	1	0.4088	0.1861	4.8250	0.0280
device_os_class	9	1	-0.5276	0.3546	2.2134	0.1368
device_make_class	1	1	-0.7524	0.2397	9.8546	0.0017
device_make_class	2	1	-0.9047	0.2447	13.6698	0.0002
device_make_class	3	1	-0.7462	0.2366	9.9442	0.0016
device_make_class	4	1	0.8376	0.3431	5.9612	0.0146
device_make_class	5	1	-0.5952	0.2456	5.8738	0.0154
device_make_class	6	1	-0.2506	0.1945	1.6611	0.1975
device_make_class	7	1	-0.2545	0.2584	0.9699	0.3247
device_make_class	8	1	0.8752	0.3583	5.9665	0.0146
device_make_class	9	1	1.0459	0.3642	8.2487	0.0041
platform		1	-0.9140	0.4342	4.4319	0.0353
resolutio*device_hei		1	0.0149	0.00862	2.9720	0.0847
resolutio*device_wid		1	0.00824	0.00820	1.0107	0.3147
resolutio*resolution		1	-3.6086	2.8622	1.5896	0.2074
device_he*device_hei		1	-0.00001	5.757E-6	3.2892	0.0697
device_wi*device_wid		1	-2.07E-6	5.226E-6	0.1563	0.6926

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
device_volume	1.127	0.896	1.416
wifi	1.275	1.077	1.511
publisher_id_class 1 vs 10	5.068	1.426	18.004
publisher_id_class 2 vs 10	1.650	1.221	2.231
publisher_id_class 3 vs 10	2.746	2.148	3.511
publisher_id_class 4 vs 10	1.126	0.812	1.561
publisher_id_class 5 vs 10	1.349	0.977	1.863
publisher_id_class 6 vs 10	0.864	0.532	1.402
publisher_id_class 7 vs 10	0.421	0.256	0.691
publisher_id_class 8 vs 10	0.441	0.258	0.754
publisher_id_class 9 vs 10	0.822	0.528	1.280
device_os_class 1 vs 10	1.230	0.949	1.593
device_os_class 2 vs 10	1.152	0.838	1.582
device_os_class 3 vs 10	0.946	0.659	1.358
device_os_class 4 vs 10	1.356	0.992	1.853
device_os_class 5 vs 10	1.431	0.953	2.149
device_os_class 6 vs 10	1.307	0.850	2.011
device_os_class 7 vs 10	0.463	0.240	0.896
device_os_class 8 vs 10	1.570	1.005	2.453
device_os_class 9 vs 10	0.615	0.277	1.368
device_make_class 1 vs 10	0.224	0.103	0.488
device_make_class 2 vs 10	0.192	0.087	0.422
device_make_class 3 vs 10	0.225	0.107	0.476
device_make_class 4 vs 10	1.097	0.546	2.203
device_make_class 5 vs 10	0.262	0.122	0.561
device_make_class 6 vs 10	0.370	0.219	0.623
device_make_class 7 vs 10	0.368	0.164	0.827
device_make_class 8 vs 10	1.139	0.546	2.376
device_make_class 9 vs 10	1.351	0.646	2.826
platform	0.401	0.171	0.939

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	64.7	Somers' D	0.294
Percent Discordant	35.3	Gamma	0.294
Percent Tied	0.0	Tau-a	0.005
Pairs	74630175	c	0.647

This summarizes the presentation of the final best model. The evolution of the modelling process is covered in the below section.

MODEL EVOLUTION

My first model was a rudimentary model with all the input variables with categorical variables defined as class variables for one-hot key encoding and numerical variables in their original form without any processing. This model was titled **Model 1.** The model was built using PROC LOGISTIC. The main results are provided through the below screenshot. The model significance, AIC/BIC values and the estimates are reported below.

It can be inferred from the below screenshot that the built model is significant with a chi-square test statistic of 176.50 and p-value<0.0001. The AIC value is 8758 and BIC value is 9081. Since, this is the first model there is no basis for comparison, and I move forwarded to build a better model by eliminating the insignificant variables in this model.

Analysis of Penalized Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-8.4124	1.1816	50.6889	<.0001
device_volume		1	0.1328	0.1167	1.2959	0.2550
wifi		1	0.2402	0.0864	7.7262	0.0054
resolution		1	-2.0594	0.8389	6.0269	0.0141
device_height		1	0.00300	0.00103	8.4460	0.0037
device_width		1	0.00245	0.00103	5.6162	0.0178
publisher_id_class	1	1	1.3713	0.5799	5.5916	0.0180
publisher_id_class	2	1	0.2607	0.1439	3.2833	0.0700
publisher_id_class	3	1	0.7498	0.1238	36.6854	<.0001
publisher_id_class	4	1	-0.0585	0.1604	0.1331	0.7152
publisher_id_class	5	1	0.2282	0.1613	2.0012	0.1572
publisher_id_class	6	1	-0.4109	0.2262	3.2987	0.0693
publisher_id_class	7	1	-0.9328	0.2425	14.7968	0.0001
publisher_id_class	8	1	-0.7708	0.2551	9.1264	0.0025
publisher_id_class	9	1	-0.3510	0.2196	2.5552	0.1099
device_os_class	1	1	0.1221	0.0837	2.1303	0.1444
device_os_class	2	1	0.0771	0.1198	0.4137	0.5201
device_os_class	3	1	-0.1446	0.1424	1.0309	0.3099
device_os_class	4	1	0.2284	0.1169	3.8197	0.0507
device_os_class	5	1	0.2636	0.1655	2.5377	0.1112
device_os_class	6	1	0.1788	0.1810	0.9762	0.3231
device_os_class	7	1	-0.4769	0.2862	2.7776	0.0956
device_os_class	8	1	0.3707	0.1858	3.9782	0.0461
device_os_class	9	1	-0.5728	0.3549	2.6046	0.1066
device_make_class	1	1	-0.2831	0.1287	4.8378	0.0278
device_make_class	2	1	-0.4345	0.1396	9.6903	0.0019
device_make_class	3	1	-0.2821	0.1466	3.7009	0.0544
device_make_class	4	1	0.1808	0.1794	1.0163	0.3134
device_make_class	5	1	-0.1348	0.1607	0.7034	0.4016
device_make_class	6	1	-0.3577	0.1785	4.0142	0.0451
device_make_class	7	1	0.2187	0.1566	1.9518	0.1624
device_make_class	8	1	0.2044	0.2102	0.9460	0.3307
device_make_class	9	1	0.3868	0.2122	3.3216	0.0684
platform		1	-0.4122	0.3736	1.2168	0.2700

Model 1: Using existing variables

The LOGISTIC Procedure

Model Information	
Data Set	WORK.TRAIN_MOBILEAD
Response Variable	install
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring
Likelihood Penalty	Firth's bias correction

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 Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	8869.341	8758.834
SC	8878.825	9081.263
-2 Log L	8867.341	8690.834

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	176.5078	33	<.0001
Score	186.4715	33	<.0001
Wald	183.1242	33	<.0001

By eliminating variables such as *device_volume, device_os_class* and *platform*, I eliminated the insignificant variables and estimated a model using PROC LOGISTIC. This model is titled as **Model 2.** The results are reported through screenshots below.
Looking at the model fit, it can be inferred that the Model is significant with a good chi-square test statistic of 158.70 and p-value of <0.0001. The AIC is 8796 and BIC = 9014 which is higher than the original model estimated which is Model 1. Hence, it was inferred that Model 2 wasn't an improvement over Model 1 which had an AIC of 8758.

Model 2: Eliminating insignificant variables

The LOGISTIC Procedure

Model Information	
Data Set	WORK.TRAIN_MOBILEAD
Response Variable	install
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring
Likelihood Penalty	Firth's bias correction

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'.

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

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

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	8911.580	8796.874
SC	8921.063	9014.988
-2 Log L	8909.580	8750.874

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	158.7055	22	<.0001
Score	170.7616	22	<.0001
Wald	165.5040	22	<.0001

Analysis of Penalized Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-7.6911	1.0674	51.9140	<.0001
wifi		1	0.2469	0.0859	8.2558	0.0041
resolution		1	-1.5592	0.7570	4.2429	0.0394
device_height		1	0.00241	0.000935	6.6223	0.0101
device_width		1	0.00188	0.000930	4.0897	0.0431
publisher_id_class	1	1	1.3670	0.5818	5.5211	0.0188
publisher_id_class	2	1	0.2483	0.1441	2.9681	0.0849
publisher_id_class	3	1	0.8020	0.1214	43.6574	<.0001
publisher_id_class	4	1	-0.0696	0.1605	0.1877	0.6648
publisher_id_class	5	1	0.2162	0.1612	1.7997	0.1797
publisher_id_class	6	1	-0.4185	0.2269	3.4011	0.0652
publisher_id_class	7	1	-0.9010	0.2428	13.7739	0.0002
publisher_id_class	8	1	-0.8223	0.2547	10.4251	0.0012
publisher_id_class	9	1	-0.3421	0.2201	2.4171	0.1200
device_make_class	1	1	-0.1679	0.1132	2.1984	0.1382
device_make_class	2	1	-0.3264	0.1252	6.7985	0.0091
device_make_class	3	1	-0.1930	0.1403	1.8928	0.1689
device_make_class	4	1	0.1175	0.1745	0.4537	0.5006
device_make_class	5	1	-0.0498	0.1555	0.1023	0.7491
device_make_class	6	1	-0.3254	0.1673	3.7824	0.0518
device_make_class	7	1	0.3048	0.1418	4.6216	0.0316
device_make_class	8	1	0.1386	0.2063	0.4511	0.5018
device_make_class	9	1	0.2858	0.2051	1.9404	0.1636

Moving forward, I tried building a better model using Model 1 as the base. Looking at the input variables, I deduced that there are certain variables whose change might impact other variables. For instance, *device_width* and *device_height*. Many mobile devices have standard *device_width* and *device_height* values. The mobile dimensions have never been random and arbitrary. Also, I analyzed that the resolution of the screen is dependent on the *device_width* and *device_height*. So, I included 3 interaction terms to weigh my deductions. I built **Model 3** using PROC LOGISTIC and got the following results.
Interpreting the results, the model is significant with a high test statistic of 182.31 and p-value <0.0001. The model recorded lower AIC and BIC values of 8724 and 9066 as compared to **Model 1**'s AIC and BIC values of 8758, 9081.

Model 3: Using interaction terms

The LOGISTIC Procedure

Model Information	
Data Set	WORK.TRAIN_MOBILEAD
Response Variable	install
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring
Likelihood Penalty	Firth's bias correction

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'.

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

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

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	8836.933	8724.615
SC	8846.416	9066.011
-2 Log L	8834.933	8652.615

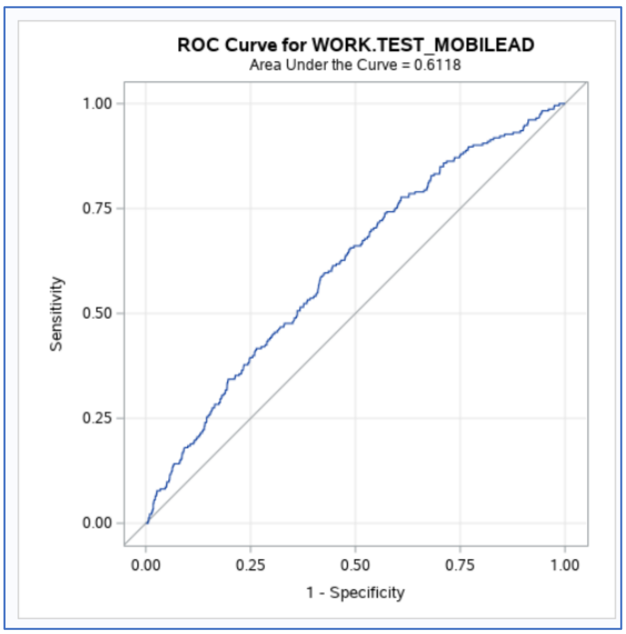
Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	182.3178	35	<.0001
Score	190.8962	35	<.0001
Wald	187.9845	35	<.0001

Analysis of Penalized Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-12.7685	2.3581	29.3206	<.0001
device_volume		1	0.1237	0.1166	1.1249	0.2889
wifi		1	0.2424	0.0864	7.8714	0.0050
resolution		1	-12.5367	4.6001	7.4272	0.0064
device_height		1	0.00875	0.00286	9.3584	0.0022
device_width		1	0.00851	0.00288	8.7565	0.0031
publisher_id_class	1	1	1.3723	0.5798	5.6020	0.0179
publisher_id_class	2	1	0.2573	0.1446	3.1673	0.0751
publisher_id_class	3	1	0.7714	0.1244	38.4617	<.0001
publisher_id_class	4	1	-0.0758	0.1605	0.2233	0.6365
publisher_id_class	5	1	0.2016	0.1616	1.5551	0.2124
publisher_id_class	6	1	-0.4062	0.2261	3.2275	0.0724
publisher_id_class	7	1	-0.9317	0.2425	14.7554	0.0001
publisher_id_class	8	1	-0.7526	0.2555	8.6771	0.0032
publisher_id_class	9	1	-0.3504	0.2207	2.5205	0.1124
device_os_class	1	1	0.1474	0.0844	3.0514	0.0807
device_os_class	2	1	0.0663	0.1199	0.3057	0.5803
device_os_class	3	1	-0.1208	0.1428	0.7159	0.3975
device_os_class	4	1	0.2543	0.1172	4.7087	0.0300
device_os_class	5	1	0.2818	0.1656	2.8959	0.0888
device_os_class	6	1	0.1935	0.1812	1.1404	0.2856
device_os_class	7	1	-0.6105	0.2897	4.4416	0.0351
device_os_class	8	1	0.3863	0.1860	4.3144	0.0378
device_os_class	9	1	-0.5562	0.3548	2.4581	0.1169
device_make_class	1	1	-0.6511	0.2096	9.6497	0.0019
device_make_class	2	1	-0.8030	0.2169	13.7053	0.0002
device_make_class	3	1	-0.6069	0.2043	8.8243	0.0030
device_make_class	4	1	0.6969	0.2984	5.4534	0.0195
device_make_class	5	1	-0.4610	0.2149	4.6037	0.0319
device_make_class	6	1	-0.3085	0.1788	2.9781	0.0844
device_make_class	7	1	-0.1600	0.2308	0.4803	0.4883
device_make_class	8	1	0.7190	0.3169	5.1472	0.0233
device_make_class	9	1	0.9110	0.3212	8.0453	0.0046
platform		1	-0.8133	0.4083	3.9679	0.0464
device_he*device_wid		0	0	.	.	.
resolutio*device_hei		1	0.00147	0.000574	6.5345	0.0106
resolutio*device_wid		1	0.00131	0.000568	5.2889	0.0215

Looking at the coefficients, I inferred that the interaction device_height and device_weight was insignificant, but others were indeed significant at 95% CI.

Headed in the same direction, I reasoned out that there could be certain variables which can have negative returns as they increase. For example, *device_height* is one variable which when increased will give a better space for advertising and hence can impact consumers to install the app. But after a point, increasing *device_height* could lead to more ads to be displayed and hence leading to a negative impact. To capture these negative returns, I included 3 quadratic terms each for *device_weight, device_height* and *resolution* as they follow a similar reasoning.
I built a model using PROC LOGISTIC after removing the insignificant interaction term identified in previous **Model 3.** This model is titled as **Model 4**.
The results of this model are already presented at the top as this is the **best model.** This model was significant with a very high chi-square test statistic and significantly lower values of AIC and BIC as compared to all other models.
The screenshots, results and interpretation can be found above. This concludes the find for my bets model, which is **Model 4.**
All the above models were modelled using FIRTH's penalized MLE as mentioned at the start.

2. The best model reported above was used to perform predictions on the test set. Using the test predictions, a ROC curve was plotted using PROC LOGISTIC. Please refer to the SAS code file for the coding part. The results are provided below as screenshots. The Area under the ROC curve is 0.6118. The 95%CI is also reported below.



ROC Association Statistics

ROC Model	Mann-Whitney Area	Mann-Whitney Standard Error	Mann-Whitney 95% Wald Confidence Limits		Somers' D	Gamma	Tau-a
ROC1	0.6118	0.0181	0.5763	0.6474	0.2237	0.2237	0.00425

3. In this question, the success event is event=1, i.e. the consumer installing the app. Hence, the False positives are the ones who are predicted to install the app but really don't. False negatives are the ones who are predicted as people who won't install the app but in actual sense, they do.
The costs incurred by the advertiser for each of the above misclassifications is given below:
False positive Cost = \$0.01
False negative Cost = \$1
Total Cost -> (False negatives * \$1) + (False positives * \$0.01)

A ROC table is generated using the above analogy and the cost calculations are performed on this ROC table. A screenshot of the resultant table is reported below.

The threshold at which the cost is minimized is 0.007. The cost associated with that threshold is $196.36.

	PROB	_FALPOS_	_FALNEG_	False_positive_cost	False_negative_cost	Total_cost
1	0.0077100818	10236	94	102.36	94	196.36
2	0.0077096636	10239	94	102.39	94	196.39
3	0.0077096066	10241	94	102.41	94	196.41
4	0.0077092015	10242	94	102.42	94	196.42
5	0.0077084305	10243	94	102.43	94	196.43
6	0.0077083875	10244	94	102.44	94	196.44
7	0.0077736925	10044	96	100.44	96	196.44
8	0.0077726123	10045	96	100.45	96	196.45
9	0.0077083145	10246	94	102.46	94	196.46
10	0.0077723444	10046	96	100.46	96	196.46
11	0.0077719032	10047	96	100.47	96	196.47
12	0.0077082695	10247	94	102.47	94	196.47
13	0.007707431	10249	94	102.49	94	196.49
14	0.007771499	10050	96	100.5	96	196.5
15	0.0077686419	10051	96	100.51	96	196.51
16	0.0077053882	10252	94	102.52	94	196.52
17	0.0077668341	10053	96	100.53	96	196.53
18	0.0077050758	10253	94	102.53	94	196.53
19	0.0077667764	10054	96	100.54	96	196.54

Part II. Linear Probability Model

- Prior to modelling the Linear Probability Model, the 4 categorical variables – *device_os_class, device_make_class, device_platform_class, publisher_id_class* are manually one-hot key encoded, since PROC REG doesn't allow the specification of categorical variables as class variables like PROC LOGISTIC.
- The dataset is split into training and test sets using PROC SURVEYSELECT. The parameters passed are seed=10, sampling ratio=0.8. Hence, the training test is 80% of the dataset in size and the remaining 20% constitutes the test set. A screenshot of the results is pasted below for reference. Please refer the code file for the SAS codes for the same.

The SURVEYSELECT Procedure

Selection Method	Simple Random Sampling

Input Data Set	MOBILEAD_LPM
Random Number Seed	10
Sampling Rate	0.8
Sample Size	97072
Selection Probability	0.800007
Sampling Weight	0
Output Data Set	MOBILEAD_LPM_SAMPLED

1. The final best Linear Probability model is presented below. The model is a compact, heteroskedasticity-free model with logarithmic predictors. It is a result of a weighted regression of the logarithmic predictors.

$$install = B_0 + B_1.log(device\_height) + B_2.log(resolution) + B_3.(wifi) + B_4.p1 + \cdots + B_{12}.p9 + B_{13}.m1 + \cdots + B_{21}.m9 + B_{22}.o1 + .. + B_{30}.o9$$

This best model is arrived at through a series of experimentation and the evolution is presented below. For this model, as reported above, we have the following variables in use:
Output variable(1): install
Input variables(30):

- Original variables: wifi, p1 – p9, m1 – m9, o1 – o9 {p, m, o -> one hot encoded variables for publisher_id_class, device_make_class, device_os_class }
- Transformed variables: log(device_height), log(resolution)

For this model, the dataset is read into SAS. The categorical variables are one-hot key coded manually and some of the input variables are log transformed. The model is built using PROC REG and the modelling results are presented using screenshots. The best model is titled **'Model 4'.** Please refer to the SAS code file for the implementation process.

Model 4: Correcting heteroskedasticity

The REG Procedure
Model: linear
Dependent Variable: install

Number of Observations Read	121339
Number of Observations Used	125

Weight: log_device_volume

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	28	0.01931	0.00068961	1.72	0.0277
Error	96	0.03849	0.00040098		
Corrected Total	124	0.05780			

Root MSE	0.02002	R-Square	0.3341
Dependent Mean	0.00800	Adj R-Sq	0.1398
Coeff Var	250.30475		

Interpreting the results obtained, we first look at the F-test of model significance. The F-test has the below hypothesis:
Null hypothesis (H_0): All beta coefficients are equal to zero
Alternate hypothesis (H_1): At least one of the beta coefficients is not zero
Looking at the results, we see that the model is significant at 95% CI with an F-statistic of 1.72 and p-value=0.0277 which is less than 0.05(alpha). The **adj-R^2 value** is **0.1398**, which is significantly higher than all the other previous models. Although it has increased manifold, it is still lesser than 15%.
A screenshot of the coefficient estimates is pasted below. Interpretations of some of the estimates are given below for **Model 4**:

- The estimates for *publisher_id_class = 9* and *device_os_class = 9* is zero, hence, they don't have any effect on the installation of the app by a consumer.
- The presence of *wifi* increases the chances of the app getting installed by the consumer by 0.01321 units
- Increasing resolution by 1% increases the chances of the app installation by 0.0000146 units. It has a very minor effect.
- On the other hand, increase device_height by 1% increases chances of app installation by 0.00006 units.
- It can be inferred that *publisher_id_class = 1*has a negative impact on the installation of the app with respect to the base group of *publisher_id_class = 10.*
- It is important to note that almost all the coefficient estimates are insignificant at 95%CI. This is because heteroskedasticity-consistent estimators have a higher standard error. That being said, the estimator is still unbiased and consistent and hence, our estimates are right.

Parameter Estimates								
						Heteroscedasticity Consistent		
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standard Error	t Value	Pr > |t|
Intercept	1	-0.05162	0.43824	-0.12	0.9065	0.15747	-0.33	0.7438
p1	1	-0.06676	0.09419	-0.71	0.4802	0.05264	-1.27	0.2078
p2	1	-0.00015059	0.02978	-0.01	0.9960	0.01364	-0.01	0.9912
p3	1	-0.01412	0.02407	-0.59	0.5590	0.01167	-1.21	0.2292
p4	1	-0.02655	0.03248	-0.82	0.4157	0.02097	-1.27	0.2085
p5	1	0.11377	0.03751	3.03	0.0031	0.08666	1.31	0.1923
p6	1	-0.00326	0.04558	-0.07	0.9431	0.01824	-0.18	0.8586
p7	1	-0.04656	0.04110	-1.13	0.2601	0.04837	-0.96	0.3382
p8	1	0.03941	0.05740	0.69	0.4941	0.03347	1.18	0.2420
p9	0	0	.	.	.	.	.	.
o1	1	0.01048	0.02978	0.35	0.7256	0.01546	0.68	0.4995
o2	1	-0.04483	0.03513	-1.28	0.2050	0.03727	-1.20	0.2320
o3	1	0.03120	0.04727	0.66	0.5108	0.02629	1.19	0.2382
o4	1	0.01757	0.03679	0.48	0.6340	0.01735	1.01	0.3138
o5	1	-0.02048	0.04747	-0.43	0.6672	0.02982	-0.69	0.4939
o6	1	0.01354	0.04973	0.27	0.7860	0.01766	0.77	0.4449
o7	1	-0.04559	0.11594	-0.39	0.6950	0.04301	-1.06	0.2919
o8	1	0.14563	0.04521	3.22	0.0017	0.10941	1.33	0.1863
o9	0	0	.	.	.	.	.	.
m1	1	-0.01384	0.05404	-0.26	0.7984	0.01987	-0.70	0.4878
m2	1	-0.00564	0.05805	-0.10	0.9228	0.02027	-0.28	0.7813
m3	1	-0.03340	0.06523	-0.51	0.6098	0.03444	-0.97	0.3346
m4	1	-0.02857	0.04726	-0.60	0.5470	0.02332	-1.23	0.2236
m5	1	-0.06699	0.07236	-0.93	0.3569	0.05593	-1.20	0.2340
m6	1	0.07491	0.05835	1.28	0.2023	0.05570	1.34	0.1818
m7	1	-0.02418	0.06860	-0.35	0.7252	0.02584	-0.94	0.3516
m8	1	-0.01799	0.05157	-0.35	0.7279	0.02199	-0.82	0.4153
m9	1	0.00274	0.05088	0.05	0.9571	0.01428	0.19	0.8481
wifi	1	0.01321	0.02125	0.62	0.5355	0.01250	1.06	0.2930
log_resolution	1	0.00146	0.04996	0.03	0.9768	0.01705	0.09	0.9320
log_device_height	1	0.00660	0.06464	0.10	0.9189	0.02241	0.29	0.7691

The above interpretations are made assuming all other variables are held constant except for the variable of interest. This summarizes the presentation of the best model **(Model 4)**.

MODEL EVOLUTION

I started out with the most basic model with all the input variables in their original form. The categorical variables were one-hot key encoded manually. They were – *publisher_id_class, device_make_lass, device_os_class* and *device_platform_class.* The other numerical variables were kept in their original form. The model was built using PROC REG. The output is captured through below screenshots. This model was titled **Model 1.**

The interpretations are given below the screenshots. Please refer to the SAS code file for the coding part.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.01615	0.00858	-1.88	0.0598
p1	1	0.01410	0.00878	1.61	0.1085
p2	1	0.00267	0.00116	2.31	0.0211
p3	1	0.00890	0.00116	7.68	<.0001
p4	1	0.00028993	0.00119	0.24	0.8075
p5	1	0.00237	0.00133	1.78	0.0754
p6	1	-0.00158	0.00150	-1.06	0.2908
p7	1	-0.00544	0.00150	-3.64	0.0003
p8	1	-0.00380	0.00155	-2.45	0.0144
p9	1	-0.00233	0.00169	-1.38	0.1683
o1	1	0.00138	0.00101	1.36	0.1743
o2	1	0.00090089	0.00124	0.73	0.4658
o3	1	-0.00048243	0.00132	-0.37	0.7148
o4	1	0.00256	0.00134	1.91	0.0557
o5	1	0.00233	0.00166	1.40	0.1601
o6	1	0.00191	0.00187	1.02	0.3083
o7	1	-0.00358	0.00243	-1.47	0.1405
o8	1	0.00326	0.00192	1.70	0.0900
o9	1	-0.00255	0.00230	-1.11	0.2674
m1	1	-0.00654	0.00218	-3.00	0.0027
m2	1	-0.00765	0.00223	-3.44	0.0006
m3	1	-0.00632	0.00216	-2.93	0.0034
m4	1	-0.00325	0.00217	-1.50	0.1334
m5	1	-0.00535	0.00225	-2.38	0.0172
m6	1	-0.00665	0.00213	-3.12	0.0018
m7	1	-0.00134	0.00244	-0.55	0.5839
m8	1	-0.00304	0.00246	-1.23	0.2175
m9	1	-0.00102	0.00250	-0.41	0.6825
device_volume	1	0.00099056	0.00093631	1.06	0.2901
wifi	1	0.00172	0.00064967	2.65	0.0082
resolution	1	-0.01544	0.00629	-2.45	0.0141
device_height	1	0.00002301	0.00000773	2.98	0.0029
device_width	1	0.00001845	0.00000774	2.38	0.0172
platform	1	-0.00336	0.00298	-1.13	0.2591

Model 1: Using existing variables

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	33	1.49539	0.04531	5.73	<.0001
Error	97038	767.31719	0.00791		
Corrected Total	97071	768.81258			

Root MSE	0.08892	R-Square	0.0019
Dependent Mean	0.00798	Adj R-Sq	0.0016
Coeff Var	1113.80412		

The most basic model is significant as it records a good F-statistic of 5.73 and a p-value which is <0.0001. Although, this is the case, the R^2 value is 0.0019. The adj-R^2 value is 0.0016. When the R^2 and adj-R^2 values are within 5%, then we can use R itself as a model evaluation metric. The low values are indicative that Linear Probability model is performing poorly because it is the case of rare-event modelling. And also, a simple relationship among the independent and dependent variables is not enough to explain the variation in the output variable.

Looking at the coefficient estimates, though it can be concluded that many of them are significant at 95%CI, their estimates and signs aren't truly indicative of the theory.

Going forward, as I had reasoned earlier in PART I (Logistic Regression Analysis), I wanted to include interaction effects among the variables *device_height, resolution* and *device_weight*. The interaction term – *device_ height*device_width* deems to biased estimates when included in the model. Hence, that term is excluded. For this model, **Model 2**, I included two interaction terms – *device_weight*resolution* and *device_height*resolution.* The model is built using the PROC REG procedure and the results are captured below through screenshots. Please refer to the SAS code for the coding part. The interpretation are given below the screenshots.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.04299	0.01891	-2.27	0.0230
p1	1	0.01399	0.00879	1.59	0.1114
p2	1	0.00257	0.00118	2.19	0.0285
p3	1	0.00896	0.00117	7.65	<.0001
p4	1	0.00015886	0.00120	0.13	0.8944
p5	1	0.00219	0.00134	1.64	0.1007
p6	1	-0.00159	0.00150	-1.06	0.2899
p7	1	-0.00543	0.00150	-3.63	0.0003
p8	1	-0.00365	0.00156	-2.34	0.0191
p9	1	-0.00231	0.00169	-1.37	0.1722
o1	1	0.00147	0.00101	1.45	0.1471
o2	1	0.00085411	0.00124	0.69	0.4894
o3	1	-0.00039963	0.00132	-0.30	0.7622
o4	1	0.00266	0.00134	1.99	0.0467
o5	1	0.00239	0.00166	1.44	0.1493
o6	1	0.00196	0.00187	1.05	0.2945
o7	1	-0.00425	0.00248	-1.72	0.0859
o8	1	0.00331	0.00192	1.72	0.0855
o9	1	-0.00250	0.00230	-1.09	0.2776
m1	1	-0.00962	0.00285	-3.37	0.0007
m2	1	-0.01074	0.00289	-3.72	0.0002
m3	1	-0.00917	0.00275	-3.33	0.0009
m4	1	-0.00049806	0.00276	-0.18	0.8567
m5	1	-0.00820	0.00282	-2.90	0.0037
m6	1	-0.00717	0.00215	-3.34	0.0008
m7	1	-0.00448	0.00307	-1.46	0.1447
m8	1	-0.00026916	0.00299	-0.09	0.9283
m9	1	0.00177	0.00304	0.58	0.5597
device_volume	1	0.00096012	0.00093646	1.03	0.3052
wifi	1	0.00172	0.00064993	2.64	0.0083
resolution	1	-0.08156	0.04212	-1.94	0.0528
device_height	1	0.00005926	0.00002471	2.40	0.0165
device_width	1	0.00005710	0.00002479	2.30	0.0213
platform	1	-0.00538	0.00320	-1.68	0.0922
r_dh	1	0.00000936	0.00000554	1.69	0.0909
r_dw	1	0.00000803	0.00000550	1.46	0.1447

Model 2: Using interaction terms

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	35	1.52659	0.04362	5.52	<.0001
Error	97036	767.28599	0.00791		
Corrected Total	97071	768.81258			

Root MSE	0.08892	R-Square	0.0020
Dependent Mean	0.00798	Adj R-Sq	0.0016
Coeff Var	1113.79295		

The model is significant with an F-statistic of 5.52 and a p-value <0.0001. Although the adj-R^2 value is 0.0016, the R^2 has increased to 0.0020 as compared to Model 1.

Looking at the estimates of the interaction terms, we see that one of them is significant at 90%CI. Introducing them has bettered the model, although not significantly.

Going forward, I pursued my same reasoning as in Logistic Regression analysis and introduced quadratic terms to my linear probability model in order to try recording higher R^2 values. So, I introduced three quadratic terms – *resolution*resolution, resolution*device_width, resolution*device_height* to my linear probability model. This is mainly done to capture the negative effects of the variables. This model is titled **Model 3.** I used PROC REG to build the model and the obtained results are reported below through screenshots. The interpretations are given below the screenshots.

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.03440	0.06553	-0.52	0.5997
p1	1	0.01512	0.00879	1.72	0.0853
p2	1	0.00365	0.00120	3.04	0.0024
p3	1	0.00997	0.00119	8.37	<.0001
p4	1	0.00087975	0.00121	0.73	0.4661
p5	1	0.00219	0.00134	1.64	0.1008
p6	1	-0.00040022	0.00152	-0.26	0.7927
p7	1	-0.00558	0.00150	-3.73	0.0002
p8	1	-0.00489	0.00158	-3.10	0.0019
p9	1	-0.00174	0.00169	-1.03	0.3030
o1	1	0.00163	0.00102	1.60	0.1091
o2	1	0.00103	0.00124	0.83	0.4045
o3	1	-0.00018106	0.00132	-0.14	0.8911
o4	1	0.00277	0.00134	2.06	0.0390
o5	1	0.00265	0.00166	1.60	0.1105
o6	1	0.00219	0.00187	1.17	0.2430
o7	1	-0.00628	0.00258	-2.43	0.0151
o8	1	0.00347	0.00192	1.80	0.0712
o9	1	-0.00224	0.00230	-0.98	0.3292
m1	1	-0.01041	0.00302	-3.45	0.0006
m2	1	-0.01155	0.00304	-3.80	0.0001
m3	1	-0.01025	0.00296	-3.46	0.0005
m4	1	0.00138	0.00319	0.43	0.6657
m5	1	-0.00924	0.00302	-3.06	0.0022
m6	1	-0.00715	0.00216	-3.31	0.0009
m7	1	-0.00517	0.00322	-1.60	0.1087
m8	1	0.00175	0.00339	0.52	0.6053
m9	1	0.00366	0.00345	1.06	0.2888
device_volume	1	0.00088345	0.00093710	0.94	0.3458
wifi	1	0.00175	0.00065028	2.69	0.0071
resolution	1	-0.07982	0.15859	-0.50	0.6148
device_height	1	0.00009668	0.00010212	0.95	0.3438
device_width	1	0.00000114	0.00010113	0.01	0.9910
platform	1	-0.00605	0.00329	-1.84	0.0655
r_dh	1	0.00003341	0.00005516	0.61	0.5448
r_dw	1	-0.00002518	0.00005480	-0.46	0.6459
r2	1	0.00235	0.01924	0.12	0.9026
dh2	1	-2.96904E-8	3.639803E-8	-0.82	0.4147
dw2	1	4.379821E-8	3.583509E-8	1.22	0.2216

Model 3: Using quadratic terms

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	38	1.71917	0.04524	5.72	<.0001
Error	97033	767.09341	0.00791		
Corrected Total	97071	768.81258			

Root MSE	0.08891	R-Square	0.0022
Dependent Mean	0.00798	Adj R-Sq	0.0018
Coeff Var	1113.67039		

The model is significant with an F-statistic of 5.72 and a p-value <0.001. We see that the R^2 and adj-R^2 , both record higher values of 0.0022, 0.0018 than the previous model. Hence, with the introduction of the quadratic terms, we are able to better explain the variation in the output variable.

In the next step, I wanted to check for Heteroskedasticity as 'Linear probability models' usually suffer from heteroskedasticity because actual values are binary in nature but predicted values are probabilities between 0 and 1. The check for heteroskedasticity is done through PROC REG by passing 'hcc spec' options. The results are shown in below screenshot. It can be inferred that heteroskedasticity is indeed present in the model as the p-value <0.0001 and we reject the null hypothesis that the model is free of heteroskedasticity.

Checking heteroskedasticity

The REG Procedure
Model: MODEL1
Dependent Variable: install

Test of First and Second Moment Specification		
DF	Chi-Square	Pr > ChiSq
521	751.60	<.0001

An attempt is made to remove the heteroskedasticity by transforming the input variables to a logarithmic version and by performing a weighted regression using the PROC REG procedure. The variable used as weight is found to be *device_volume* though trial and error. The results are pasted below as screenshots. Please to the SAS code file for the coding part.

Model 4: Correcting heteroskedasticity

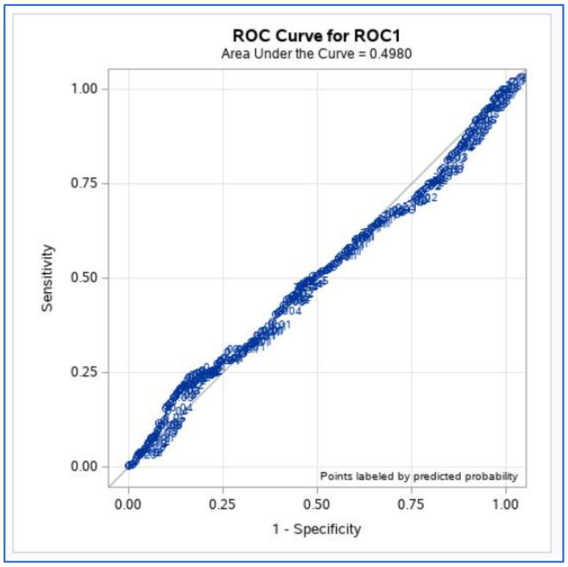
The REG Procedure
Model: linear
Dependent Variable: install

Test of First and Second Moment Specification		
DF	Chi-Square	Pr > ChiSq
91	91.23	0.4735

It can be inferred that the model is free of heteroskedasticity as we fail to reject the null at 95%CI. This model is rendered as the best model titled **Model 4.** The model fit and coefficient estimates are reported in the above sections as screenshots.
<u>This best model has the highest **adj-R^2** value of **0.1398** and **R^2** value of **0.3341**</u>. This tells us that 13.98% of the variation in the output variable is explained by the variation in input variables.

2. The ROC curve is plotted for the above best model by first predicting the output for the test set. The test set predictions are used to generate the ROC curve using PROC LOGISTIC and 'nofit' option. The curve is reported through the below screenshot.
   It can be seen that the <u>Area under the curve is 0.4980</u>. The Confidence interval is recorded as (0.4633, 0.5327).
   Contrasting with the results obtained for Logistic Regression case, the **Area under curve for the Logistic Regression is more – 0.6118**.

ROC Association Statistics							
	Mann-Whitney						
ROC Model	Area	Standard Error	95% Wald Confidence Limits		Somers' D	Gamma	Tau-a
ROC1	0.4980	0.0177	0.4633	0.5327	-0.00398	-0.00400	-0.00008



Hence, we conclude that at 95% Confidence level, Logistic Regression records a higher Area under curve of 0.6118 as compared to 0.4980 of Linear probability model.

3. The cost calculations done for the Logistic Regression model are replicated for the Linear probability model as well. The cost calculations follow below equations.
   False positive Cost = $0.01
   False negative Cost = $1
   Total Cost -> (False negatives * $1) + (False positives * $0.01)

   The ROC table is plotted and the cost for the below thresholds are noted.
   [0.001 0.005 0.010 0.015 0.020 0.025 0.030 0.035 0.040 0.045 0.050]
   The predictions are captured in a table and they are used to calculate the False negatives and False positives for each threshold using PROC SQL. Please refer to the SAS code for the coding part.
   A screenshot of the resultant table is posted below:

threshold	False_Positives	False_Negatives	FPCost	FNCost	Total_Cost
0.001	52014	585	520.14	585	1105.14
0.005	45330	626	453.3	626	1079.3
0.01	35517	709	355.17	709	1064.17
0.015	26979	816	269.79	816	1085.79
0.02	23888	847	238.88	847	1085.88
0.025	20467	870	204.67	870	1074.67
0.03	18828	881	188.28	881	1069.28
0.035	16266	892	162.66	892	1054.66
0.04	13285	912	132.85	912	1044.85
0.045	12833	914	128.33	914	1042.33
0.05	12372	917	123.72	917	1040.72

It can be inferred that the lowest cost of $1040.72 occurs at a threshold of 0.05.

This threshold is different from the value obtained for Logistic Regression model which was **0.007**.
I calculated the average cost for both the models. They are reported below:
Logistic Regression: Average Cost -> $211
Linear Probability Model: Average Cost -> $1067

Conclusively, it can be concluded that Linear Probability model isn't as effectively able to fit the data as Logistic regression. The associated cost for Linear probability is evidencing that. Also, considering the rare event modelling, Logistic Regression effects a penalized MLE algorithm to fit the data effectively and better than Linear Probability model.