MOBILE ADVERTISING PREDICTION

BUAN 6346

Onengiyeofori Harry

TABLE OF CONTENTS

- Exploratory Analysis
- Data Preparation
- Question 1: Model Building and Selection
- Question 2: Classification

EXPLORATORY ANALYSIS

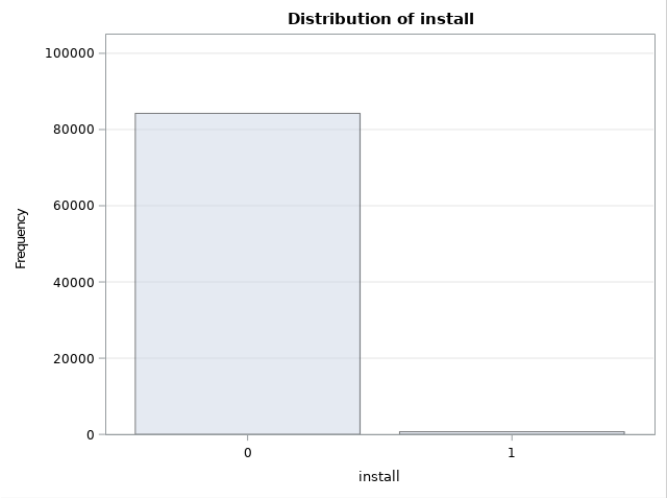
Assumptions:

- All exploratory analysis was carried out on the training data set. The test set is assumed to be "unseen" data.
- The model to be built should be generic to publishers, it was built to be applied to any set of publishers that have the same customer demographics as shown in the data set.

Below is a summary of the exploratory analysis performed, because the task is to predict probability consumer installs all the percentages reported are with respect to proportion of success to the data set and then proportion of successful events.

Proportion of Installations: The number of app installations in the training dataset is low at 0.81%

install	Frequency	Percent
0	84246	99.19
1	692	0.81

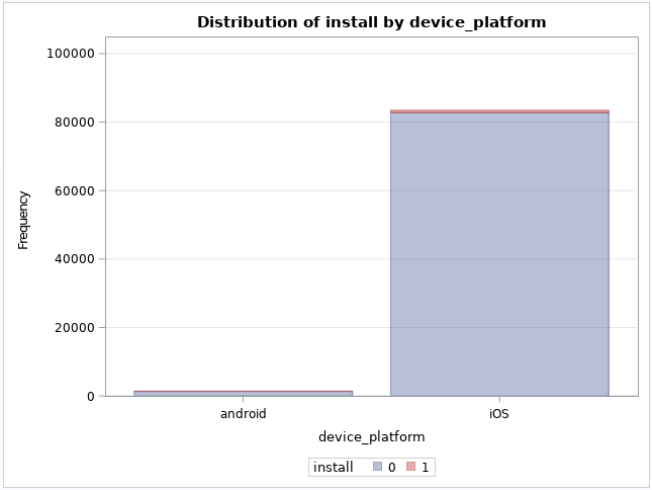


Number of Publishers: The dataset had 1304 publishers, only 57 or 0.044% had success of consumers installing the app.

install	Frequency
0	1247
1	57

Proportion of Operating Systems Recorded: The most operating system used was the Android. It accounts for 0.8% of all installations which is about 99% of the install group.

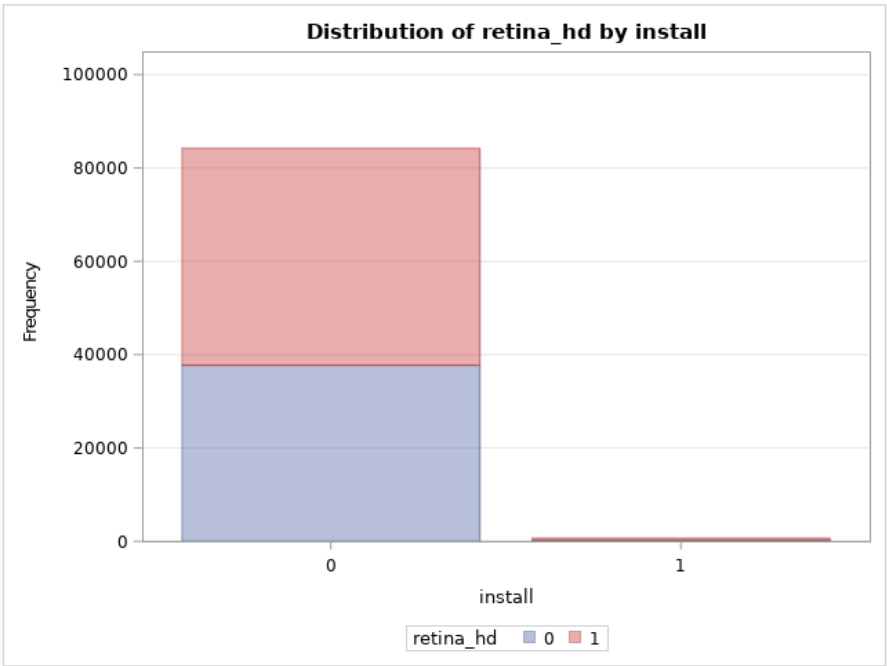
Table of install by device_platform			
	device_platform		
install	android	iOS	Total
0	1422 1.67	82824 97.51	84246 99.19
1	12 0.01	680 0.80	692 0.81
Total	1434 1.69	83504 98.31	84938 100.00



Proportion of installations by device make: From the above chart it can be seen that in the data set most of the consumers used android phones. The charts below show the proportions of installations from iPhones, iPads and Apple HD devices.

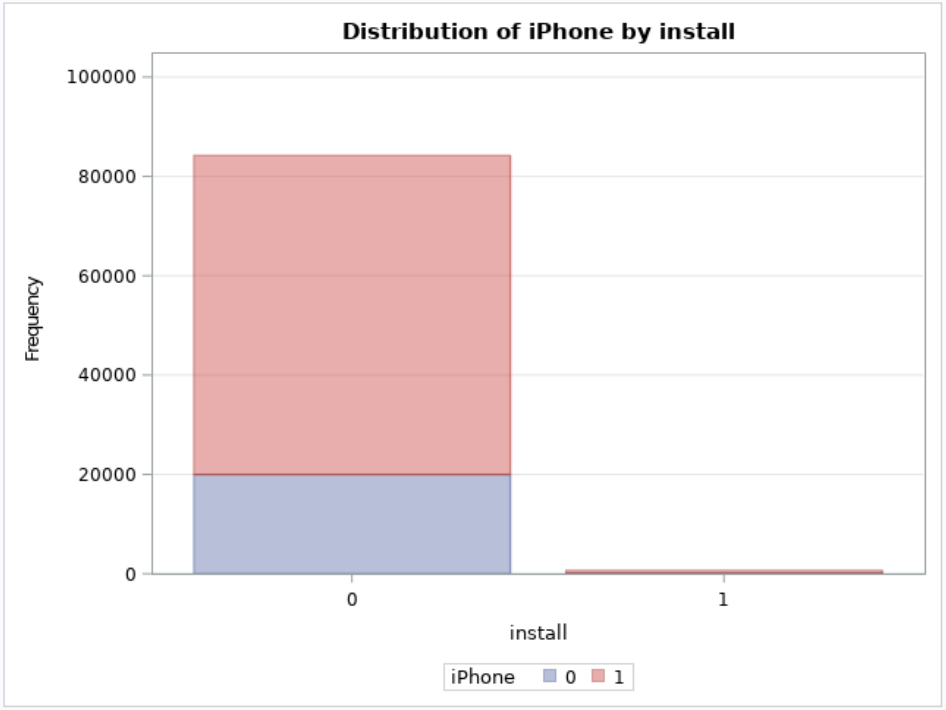
1. **By Apple HD Devices**

Table of retina_hd by install			
	install		
retina_hd	0	1	Total
0	37758 44.45	308 0.36	38066 44.82
1	46488 54.73	384 0.45	46872 55.18
Total	84246 99.19	692 0.81	84938 100.00



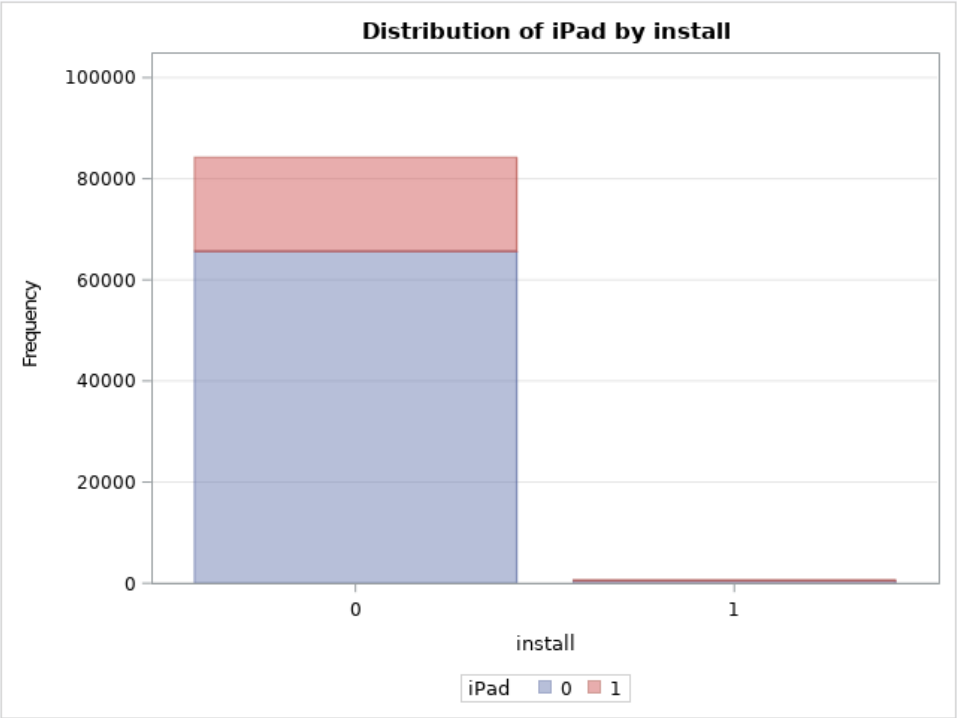
2. By iPhones

Table of iPhone by install			
	install		
iPhone	0	1	Total
0	20022 23.57	184 0.22	20206 23.79
1	64224 75.61	508 0.60	64732 76.21
Total	84246 99.19	692 0.81	84938 100.00



3. By iPads

Table of iPad by install			
	install		
iPad	0	1	Total
0	65683 77.33	520 0.61	66203 77.94
1	18563 21.85	172 0.20	18735 22.06
Total	84246 99.19	692 0.81	84938 100.00

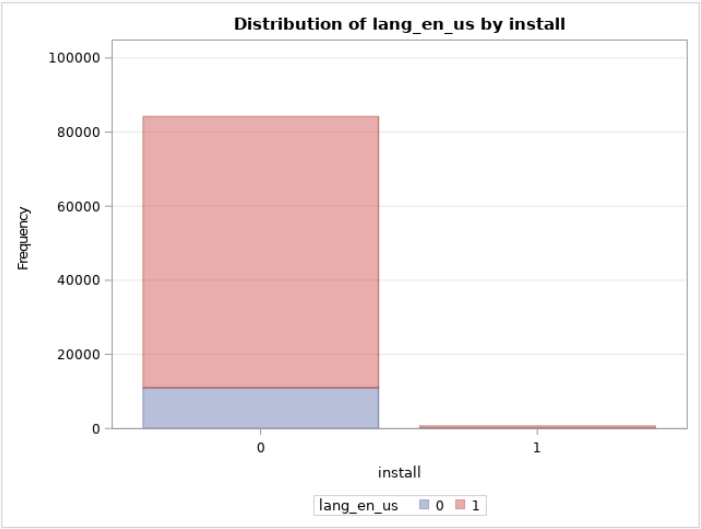


WIFI Use: 0.61% of the users which is about 75% of the installation group had WIFI enabled when they installed the app

Table of install by wifi			
	wifi		
install	0	1	Total
0	31.72	67.47	99.19
1	0.20	0.61	0.81
Total	27113 31.92	57825 68.08	84938 100.00

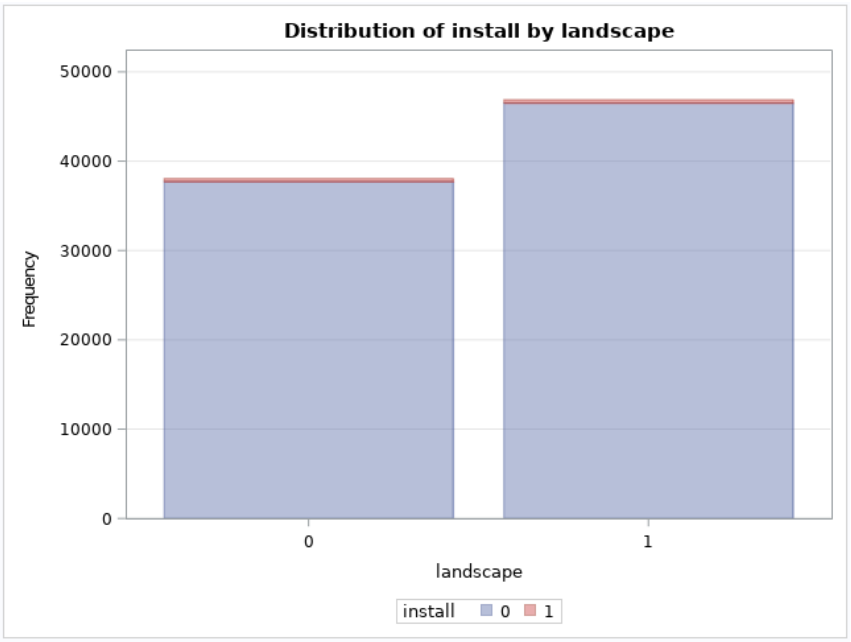
Language Settings: 0.73% of them which is about 90% of the installation group had their phone language settings on US English

Table of lang_en_us by install			
	install		
lang_en_us	0	1	Total
0	11013 12.97	71 0.08	11084 13.05
1	73233 86.22	621 0.73	73854 86.95
Total	84246 99.19	692 0.81	84938 100.00



Screen Orientation: A screen orientation variable called portrait was created to capture the position of the screen when the advert was viewed; when the value is 1 screen orientation was portrait at 0 it was landscape. 0.43% of the users which comes down to 53% in the install group had the phone in landscape mode while 0.39% of users which is 47% in the install group had the phone in portrait mode.

Table of install by landscape			
	landscape		
install	0	1	Total
0	37724 44.41	46522 54.77	84246 99.19
1	330 0.39	362 0.43	692 0.81
Total	38054 44.80	46884 55.20	84938 100.00



DATA PREPARATION

To prepare the training and test datasets, new variables were created by either setting indicator variables from prior variables or calculating new statistics to measure characteristics. Redundant variables or variables that could be inferred from other variables were dropped:

- The portrait variable was created by comparing the device_height and device_width variables. I dropped both variables because they can be determined from the resolution and portrait variable.

- Publisher_install_rate is the installation success rate for each publisher. It is the sum of installs a publisher has divided by the number of visits for that publisher. I dropped the publisher id because the install rate is constant for each publisher and so to apply the model to a new dataset what would be required will be the publishers install rate instead of their id.

- An indicator variable for device_platform was created – for the various iOS levels from iOS10 to iOS 7.

- Based on Apple's marketing brand names and the percentage of installations the devices that fall under the class of Apple's retinaHD where grouped under a variable retina_HD, iPhones and iPads were put in the variabes iPhone and iPad respectively

- Also indicator variables where created for the language settings based on percentage of successful installs. The US english setting lang_en_us was significant in a chi square test.

Question 1: MODEL BUILDING & SELECTION

There are 692 events in a dataset of 84246 observations, because of this, the models were estimated using the Firth's penalized maximum likelihood approach using proc logistic.

Three models were estimated by iteratively testing the effects of different variables and their interactions in the model.

Based on the AUC of the the model on test data, model 2 was chosen as the appropriate model. Below are the model summaries.

Model 1 – main effects model: I estimated this model to confirm which variables had influence on the dependent variable.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	7986.908	7846.365
SC	7996.258	7986.611
-2 Log L	7984.908	7816.365

Analysis of Penalized Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-6.1097	0.4517	182.9175	<.0001
resolution	1	0.1799	0.0518	12.0616	0.0005
publisher_install_ra	1	33.9120	5.3304	40.4743	<.0001
wifi	1	0.2733	0.0920	8.8172	0.0030
landscape	1	-0.1106	0.0800	1.9148	0.1664
device_volume	1	0.0482	0.1248	0.1489	0.6996
retina_hd	1	-0.0241	0.1115	0.0467	0.8289
iOS10	1	1.1557	1.4993	0.5942	0.4408
iOS9	1	1.0872	1.4986	0.5263	0.4682
iOS8	1	0.6983	1.5229	0.2103	0.6466
iOS7	1	0.8009	1.5158	0.2792	0.5972
iPhone	1	-0.7359	1.4427	0.2602	0.6100
iPad	1	-0.8193	1.4441	0.3219	0.5705
lang_en_us	1	0.2524	0.3555	0.5042	0.4776
lang_en_other	1	0.1733	0.3786	0.2096	0.6471

Model 2 – main effects and interaction effects for generic devices. I estimated this model to measure the effect of the general characteristics on the dependent variable.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	8030.188	7750.528
SC	8039.538	7844.025
-2 Log L	8028.188	7730.528

Analysis of Penalized Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-6.0647	0.1708	1260.7186	<.0001
resolution	1	0.0514	0.0451	1.3003	0.2542
publisher_install_ra	1	53.0485	4.6870	128.1045	<.0001
lang_en_us	1	0.3113	0.1326	5.5136	0.0189
landscape	1	0.1669	0.0933	3.1996	0.0737
wifi	1	0.2370	0.0898	6.9633	0.0083
install_rate_sq	1	-108.8	9.1595	141.1873	<.0001
resolutio*publisher_	1	19.9239	2.3832	69.8947	<.0001
publisher_*landscape	1	2.1194	5.6754	0.1395	0.7088
resolu*publis*landsc	1	-17.0206	4.0510	17.6534	<.0001

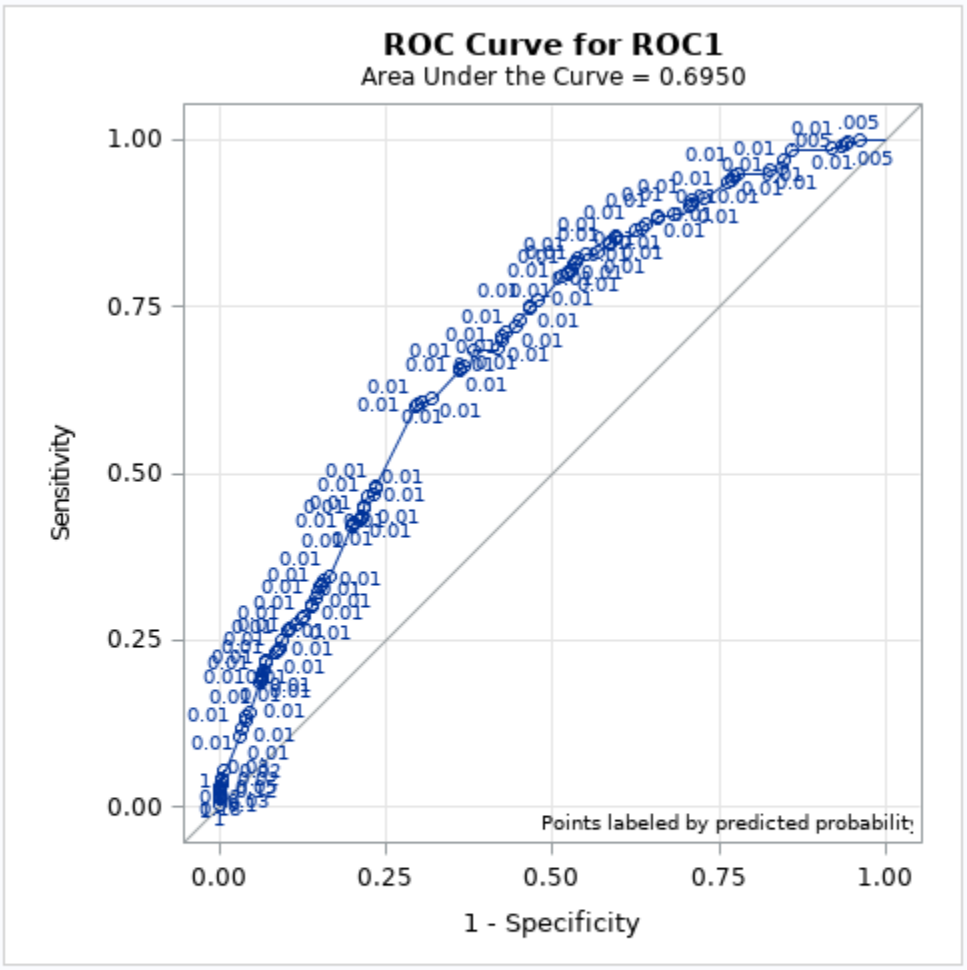
Model 3 - main effects and interaction effects for device specific. Measures the effect of individual device categories. Only the iPad had significant effect on the model

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	8023.425	7735.135
SC	8032.775	7875.381
-2 Log L	8021.425	7705.135

Analysis of Penalized Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-6.2793	0.2266	768.2267	<.0001
resolution	1	0.1381	0.0622	4.9323	0.0264
publisher_install_ra	1	61.8441	5.3376	134.2484	<.0001
lang_en_us	1	0.4110	0.1868	4.8413	0.0278
landscape	1	1.0117	0.3689	7.5223	0.0061
wifi	1	0.3357	0.1401	5.7394	0.0166
install_rate_sq	1	-112.4	8.7249	165.8374	<.0001
resolutio*publisher_	1	9.3705	3.7350	6.2942	0.0121
publisher_*landscape	1	1.1028	5.6031	0.0387	0.8440
resolu*publis*landsc	1	-12.6099	3.9452	10.2160	0.0014
landscape*wifi	1	-0.9554	0.3954	5.8373	0.0157
lang_en_us*landscape	1	-0.8350	0.3659	5.2085	0.0225
lang_en*landsca*wifi	1	0.8360	0.3890	4.6200	0.0316
iPad	1	-0.2826	0.1534	3.3945	0.0654
publisher_insta*iPad	1	28.2951	8.5385	10.9815	0.0009

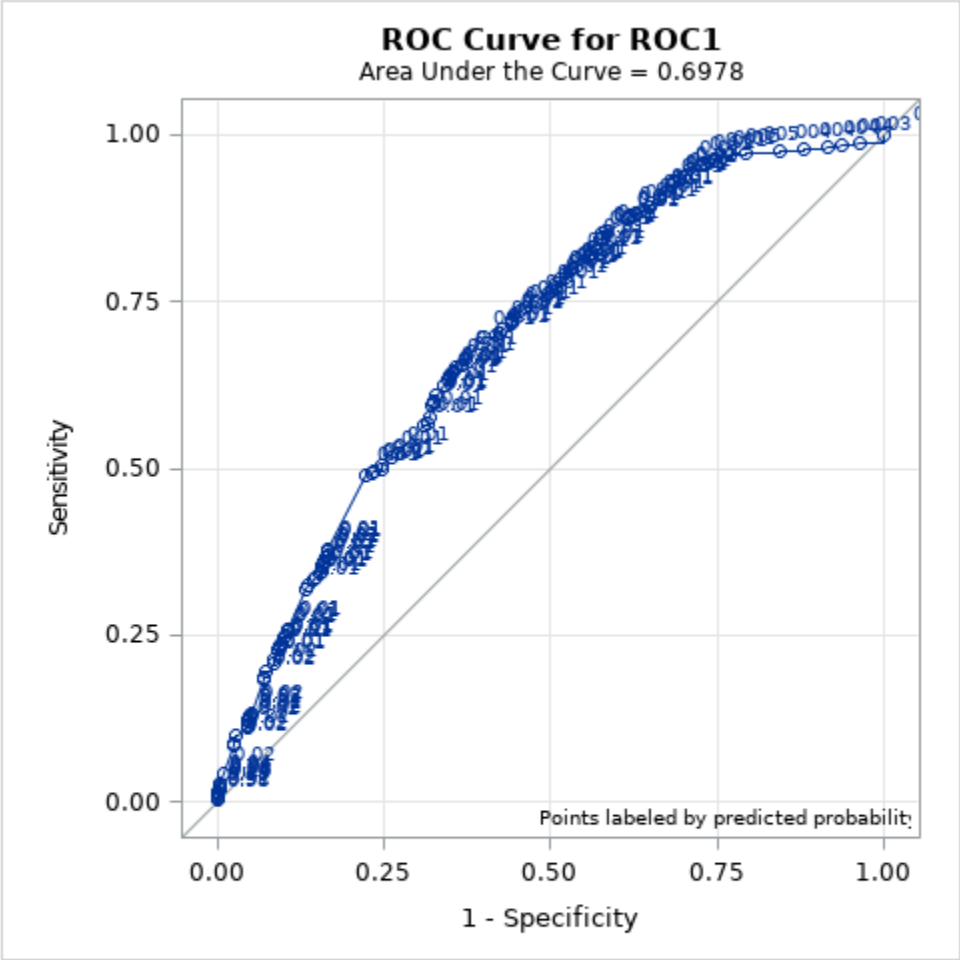
Below are the ROC curves from estimating the models on the test data set

Model1



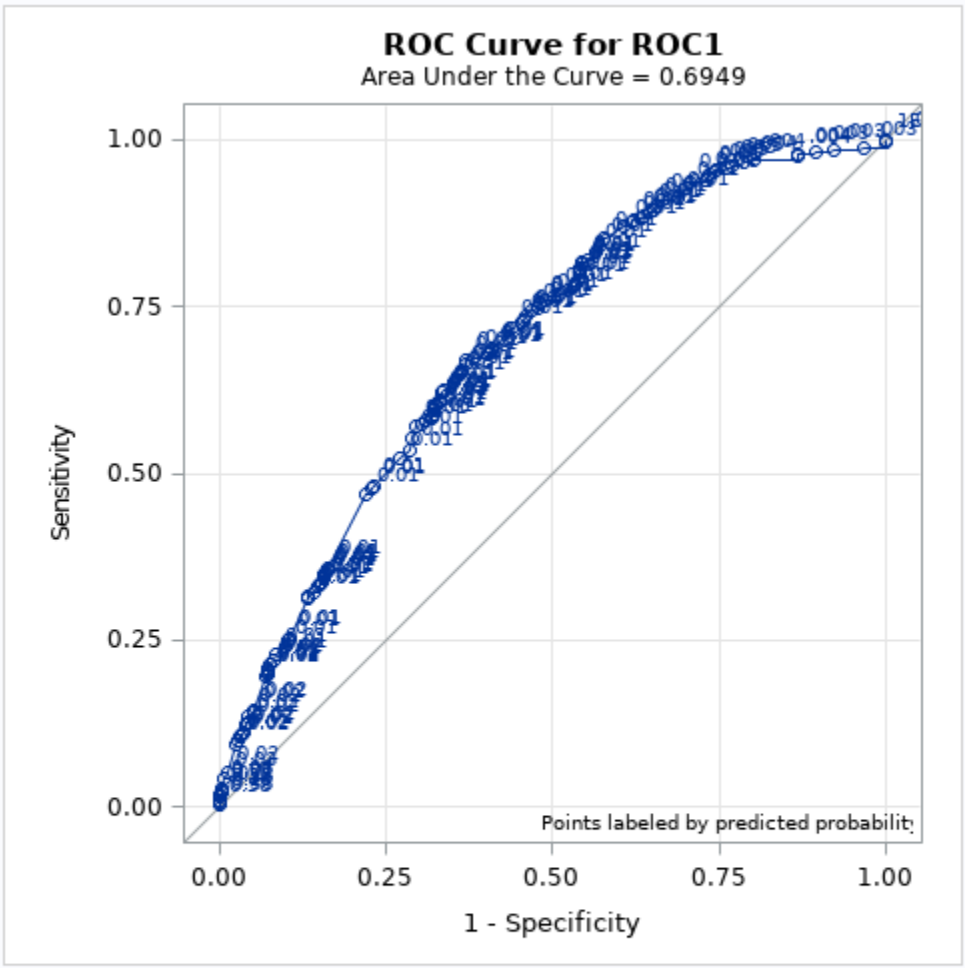
ROC Association Statistics							
	Mann-Whitney						
ROC Model	Area	Standard Error	95% Wald Confidence Limits		Somers' D	Gamma	Tau-a
ROC1	0.6950	0.0140	0.6676	0.7224	0.3901	0.3972	0.00671

Model2



ROC Association Statistics							
	Mann-Whitney						
ROC Model	Area	Standard Error	95% Wald Confidence Limits		Somers' D	Gamma	Tau-a
ROC1	0.6978	0.0136	0.6712	0.7243	0.3955	0.4015	0.00681

Model3



ROC Association Statistics							
	Mann-Whitney						
ROC Model	Area	Standard Error	95% Wald Confidence Limits		Somers' D	Gamma	Tau-a
ROC1	0.6949	0.0137	0.6680	0.7219	0.3899	0.3958	0.00671

Question 2: CLASSIFICATION

The following process was used to classification

1. Sort the probabilities in the classification table by ascending order
2. Calculate the total cost for cost ratios 25, 50, 100 and 200 for each of the probabilities using the formula below

   TC = C_1 * FP + C_2 *FN

   where TC is Total Cost, FP is False Positive and FN is False Negative

   Let cost ratio be defined as R, where R = C_2/C_1, then C_2 = C_1 * R

   Therefore TC = C_1*FP + C_1*R*FN

3. Choose the probability associated with the minimum cost. If there where probabilities with the same minimum cost choose the minimum probability to reduce the false negatives

Proabilities and Minimum Costs At 25, 50, 100, 200

Obs	probability	mincost
1	0.340	98.032
2	0.340	118.286
3	0.340	158.795
4	0.540	239.813

4. Classify the data based on this new probability