Part I: Logistic Regression

1.

```
result=mba263.logit(data['buyer_dummy'],data[ ['last','total_','female','child',
'youth','cook','do_it','refernce','art','geog'] ])
result.summary()
mba263.odds_ratios(result)
```

```
Optimization terminated successfully.
         Current function value: 0.241222
         Iterations 7
```

	Odds ratios	std err	z	P>|z|	[0.025	0.975]
last	0.909634	0.002540	35.575710	0.0	0.904707	0.914562
total_	1.001117	0.000198	5.627190	0.0	1.000732	1.001502
female	0.467330	0.016712	31.873279	0.0	0.434908	0.499751
child	0.830094	0.014346	11.843363	0.0	0.802263	0.857926
youth	0.893173	0.023320	4.580965	0.0	0.847933	0.938414
cook	0.763134	0.013071	18.121158	0.0	0.737776	0.788493
do_it	0.583235	0.015727	26.499322	0.0	0.552724	0.613746
refernce	1.264514	0.033583	7.876323	0.0	1.199362	1.329665
art	3.175878	0.070327	30.939604	0.0	3.039444	3.312311
geog	1.775845	0.033086	23.449340	0.0	1.711658	1.840032

2. Summarize and interpret the results (so that a marketing manager can understand them). Which variables are significant? Which seem to be 'important'? Interpret the odds-ratios for each of the predictors.

Odds of purchasing 'The Art History of Florence'

Predictors	Odds	Significance
Last	0.9096	A 1 month increase in recency decreases the odds of making a purchase by 100 – 91 = 9.0%.
total_	1.001	A dollar increase in mv increases the odds of buying by 100 - 100.01 = 1%.
Female	0. 4673	If customers are female, the odd of purchasing is reduced by 100 – 46.73 = 53.3%.

Child	0.8301	If customers buy children books, then the odds of purchasing are reduced by 100 – 83 = 17%.
Youth	0.8932	If customers that buy youth books, then the odds of purchasing are reduced by 100 – 89 = 11%.
Cook	0.7631	If customers purchase a cookbook, the odds of buying decreases by 100 – 76 = 24%
do_it	0.5832	If customers purchase a do-it yourself book, the odds of buying reduced by 100 – 58 = 42%
Reference	1.2645	If customers purchase a reference book, the odds of buying increase by 100 – 126.5 = 26.5%.
Art	3.1759	If customers purchase an art book, odds of buying increases by 100 – 317 = 217%.
Geog	1.7759	If customers purchase a geography book, the odds of buying increases by 100 – 177.6 = 77.6%

The features or variables in our data are significant because their p-values are less than 0.05 or 5%. If the p-values are less than 0.05, we can reject the hypothesis that the dependent and predicted value have little to no correlations. The most significant variables are art, geog, and reference, since those features or variables generate increases in future purchases. We also need to be cognizant of the fact that female customers are less likely to make future purchases.

Part II: Decile Analysis of Logistic Regression Results

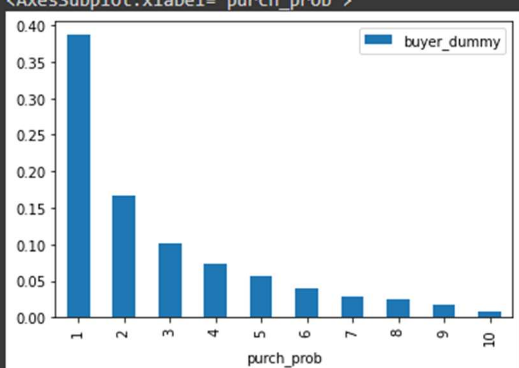
1. Assign each customer to a decile based on his or her predicted probability of purchase.

```
[24] data['predicted'] = result.predict()
     result = 10 - mba263.ntile(data['predicted'],10)
     data['purch_prob']=result
```

2. Create a bar chart plotting response rate by decile (as just defined above).

```
[25] data[['buyer_dummy','purch_prob']].groupby('purch_prob').mean().plot(kind='bar')

<AxesSubplot:xlabel='purch_prob'>
```



3. Generate a report showing number of customers, the number of buyers of "The Art History of Florence' and the response rate to the offer by decile for the random sample (i.e. the 50,000 customers) in the dataset.

```
data[['buyer_dummy','purch_prob'] ].groupby('purch_prob').describe()
```

	buyer_dummy							
	count	mean	std	min	25%	50%	75%	max
purch_prob								
1	5000.0	0.387000	0.487112	0.0	0.0	0.0	1.0	1.0
2	5000.0	0.167200	0.373192	0.0	0.0	0.0	0.0	1.0
3	5000.0	0.102200	0.302941	0.0	0.0	0.0	0.0	1.0
4	5000.0	0.073600	0.261145	0.0	0.0	0.0	0.0	1.0
5	5000.0	0.056800	0.231483	0.0	0.0	0.0	0.0	1.0
6	5000.0	0.039200	0.194090	0.0	0.0	0.0	0.0	1.0
7	4998.0	0.027811	0.164448	0.0	0.0	0.0	0.0	1.0
8	5002.0	0.024190	0.153655	0.0	0.0	0.0	0.0	1.0
9	5000.0	0.018000	0.132964	0.0	0.0	0.0	0.0	1.0
10	5000.0	0.008400	0.091275	0.0	0.0	0.0	0.0	1.0

```
data[ ['buyer_dummy','purch_prob'] ].groupby('purch_prob').sum()
```

	buyer_dummy
purch_prob	
1	1935
2	836
3	511
4	368
5	284
6	196
7	139
8	121
9	90
10	42

4. For the 50,000 customers in the dataset, generate a report showing the mean values of the following variables by probability of purchase decile:
Total $ spent, Months since last purchase, and Number of books purchased for each of the seven categories (i.e., children, youth, cookbooks, do-it-yourself, reference, art, and geography).

```
data[['last','total_','female','child','youth','cook','do_it','refernce','art','geog','purch_prob']].groupby('purch_prob').mean()
```

purch_prob	last	total_	female	child	youth	cook	do_it	refernce	art	geog
1	7.194400	257.352600	0.418800	1.064800	0.513800	1.066800	0.471400	0.562800	1.500600	1.330800
2	7.958000	224.869200	0.491000	0.836400	0.392800	0.848200	0.393400	0.404600	0.753000	0.890800
3	8.618800	214.228400	0.548800	0.791000	0.365400	0.796000	0.369800	0.383200	0.480200	0.701000
4	8.782800	207.643000	0.631800	0.752600	0.362600	0.796600	0.340400	0.308200	0.302400	0.540400
5	9.573200	199.111800	0.697800	0.758000	0.333800	0.820800	0.369800	0.272400	0.216800	0.463800
6	10.937600	199.130200	0.728200	0.748000	0.364800	0.864800	0.394200	0.258800	0.163400	0.386200
7	12.372149	191.297319	0.778711	0.761104	0.348139	0.836134	0.420968	0.227491	0.132053	0.294718
8	14.417833	191.598161	0.813075	0.804678	0.360256	0.909036	0.447821	0.204918	0.113954	0.254298
9	17.857600	193.610800	0.770400	0.960600	0.405200	1.118200	0.650600	0.252400	0.127600	0.316000
10	25.868400	204.341600	0.781800	1.067400	0.463000	1.309400	0.772200	0.247600	0.069200	0.291600

5. Summarize and interpret the decile analysis results. Are the patterns in the decile analysis consistent with your conclusions from the logistic regression?

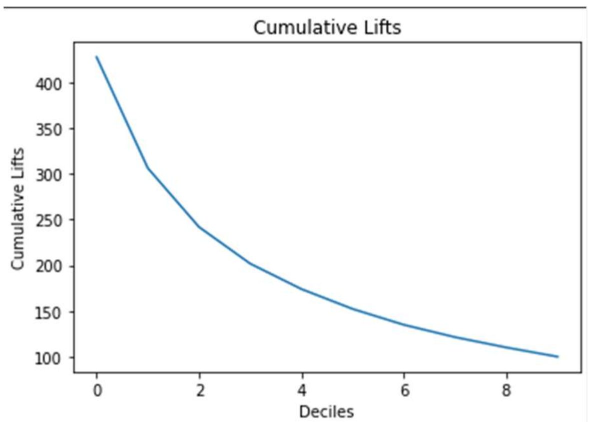
The patterns in the decile analysis are mostly consistent with our conclusions from the logistic regression. The few differences between the decile analysis and logistic regression are that the child variable is predicted to reduce the chance to make future purchases, however, we can observe from our decile analysis that it is not consistent with our logistic regression. Another difference is that features such as art and geog have small jumps or spikes in decile values, more specifically art has a small spike on decile 9 and geog has a small spike on deciles 9 and 10. Also we observe that the decile analysis for cook is not consistent with out logistic regression as the values for cook book show increases in purchases in a few deciles.

Part III: Lifts and Gains

1. Use the information from the report in II.3 above to create a table showing the lift and cumulative lift for each decile. You may want to use Excel for these calculations.

A	B	C	D	E	F	G	H	I
Decile	Customers	Tot Cust	Buyers	Tot Buyer:	RR (%)	Lift	Tot RR (%)	Tot Lift
1	5000	5000	1935	1935	38.7	427.91	38.7	427.91
2	5000	10000	836	2771	16.7	184.9	27.7	306.39
3	5000	15000	511	3282	10.2	113	21.88	241.93
4	5000	20000	368	3650	7.4	81.4	18.25	201.79
5	5000	25000	284	3934	5.7	62.8	15.74	173.99
6	5000	30000	196	4130	4	43.3	13.77	152.22
7	4998	34998	139	4269	2.8	30.75	12.2	134.87
8	5002	40000	121	4390	2.4	26.75	10.98	121.35
9	5000	45000	90	4480	2	19.9	9.96	110.08
10	5000	50000	42	4522	1	9.29	9.04	100
	50000		4522					

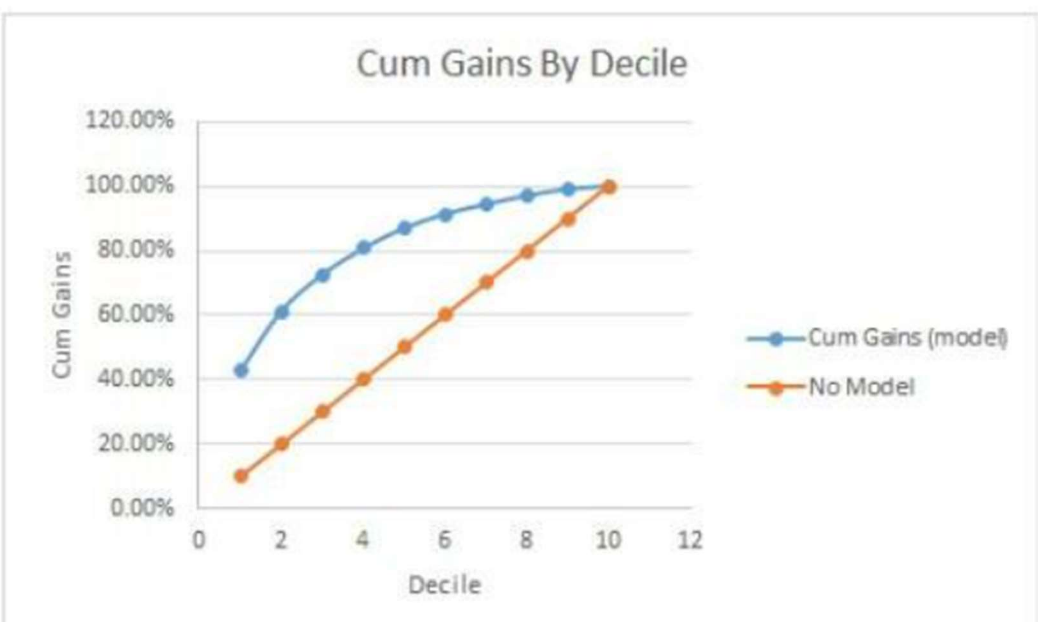
2. Create a chart showing the cumulative lift by decile.



3. Use the information from the report in II.3 above to create a table showing the gains and cumulative gains for each decile. You may want to use Excel for these calculations.

	A	B	C	D	E	F	G
1	Decile	Customer	Tot Cust	Buyers	Tot Buyer:	Gain	Tot Gain
2	1	5000	5000	1935	1935	42.79%	42.79%
3	2	5000	10000	836	2771	18.49%	61.28%
4	3	5000	15000	511	3282	11.30%	72.58%
5	4	5000	20000	368	3650	8.14%	80.72%
6	5	5000	25000	284	3934	6.28%	87.00%
7	6	5000	30000	196	4130	4.33%	91.33%
8	7	4998	34998	139	4269	3.07%	94.41%
9	8	5002	40000	121	4390	2.68%	97.08%
10	9	5000	45000	90	4480	1.99%	99.07%
11	10	5000	50000	42	4522	0.93%	100%
12		50000		4522			
13							
14							
15							

4. Create a chart showing the cumulative gains by decile along with a reference line corresponding to 'no model'.



Part IV: Profitability Analysis

1. What is the breakeven response rate?

Profit per sale: 18-9-3 = $6

Breakeven response rate = Cost to mail/Profit per sale = 0.50/6 = 8.3%

2. For the customers in the dataset, create a new variable (call it "target") with a value of 1

if the customer's predicted probability is greater than or equal to the breakeven response rate and 0 otherwise.

```
[17] data['target']=(data['predicted']>0.083)*1

[21] data['target']

0        0
1        0
2        0
3        0
4        0
        ..
49995    0
49996    1
49997    1
49998    1
49999    1
Name: target, Length: 50000, dtype: int64
```

3. Considering that there are 500,000 remaining customers, generate a report summarizing the number of customers, the expected number of buyers of 'The Art History of Florence' and the expected response rate to the offer by the "target" variable.

```
mba263.tabulate(data['target'])
```

	Name	Count	Frequency
0	0	34390	0.6878
1	1	15610	0.3122

4. For the 500,000 remaining customers, what would the expected gross profit (in dollars, and also as a percentage of gross sales) and the expected return on marketing expenditures have been if BookBinders had mailed the offer to buy "The Art History of Florence" only to customers with a predicted probability of buying that was greater than or equal to the breakeven rate?

Profit per sale = $6

Cost to mail = $0.50

Sample: 15610/50000 = 31.22%

Target: 500,000*.3122 = 156,100

Response rate = 21.36% = .2136 * 156,100 = 33,342.96 customers

Gross profit = ($6*33,342.96) - (0.50*156,100) = $122,007.76

Gross profit/Sales = 122,007.76/ (18*.2136*156,100) = 20.32%

Expected return on marketing expenditure = 122,007.76/ (0.5*156,100) = 156.32%