enginius
MARKETING ENGINEERING ONLINE

Enginius

Predictive Modeling

Ashutosh Jagdish Narvekar, Northeastern University

Table of Contents

Warnings

The following warnings were triggered during execution. Although they did not interrupt the analyses, they might indicate that there is an issue with the data or with the options chosen. Please review them carefully before going any further.

> Insample data has NA values. Rows with NA values have been removed for analysis.

Predictive options

Options selected

Option	Selection
Target	Choice between 2 alternatives (0/1)
Target variable	How likely are you to purchase life insurance in the next 2 years
Box-Cox transform the predictors	No
Log transform the target variable	No
Cross validation	10 fold
Out-of-sample prediction	No
Date and time	2024-06-25 23:32:00 UTC

Options selected.

Data description

	Data	Number of Rows	Number of columns	Column names
1	Calibration data	42	10	\, How likely are you to purchase life insurance in the next 2 years , Importance of easily customize plans , importance of insurance company to have a large market cap, Age,

Data description.

Data description

Calibration data

Calibration data contains 41 rows and 14 columns. It contains 13 predictors, to which we added an intercept.

	Importance of easily customize plans	importance of insurance company to have a large market cap	Age = 23-27	Age = Above 35	Gender = Female	Gender = Male	Current health status	Health conscious	Marital status = Single	Marital status = Married	No. of Dependents = 0	No. of Dependents = 1-2	No. of Dependents = 4+
Average	7.512	7.415	0.5122	0.3902	0.4390	0.5122	7.341	7.732	0.7317	0.1951	0.6098	0.1951	0.1220
Standard deviation	2.293	2.085	0.5061	0.4939	0.5024	0.5061	2.198	1.988	0.4486	0.4012	0.4939	0.4012	0.3313
Median	8.000	7.000	1.0000	0.0000	0.0000	1.0000	8.000	8.000	1.0000	0.0000	1.0000	0.0000	0.0000
Minimum	0.000	1.000	0.0000	0.0000	0.0000	0.0000	0.000	2.000	0.0000	0.0000	0.0000	0.0000	0.0000
Maximum	10.000	10.000	1.0000	1.0000	1.0000	1.0000	10.000	10.000	1.0000	1.0000	1.0000	1.0000	1.0000

Summary of predictors. These data have been used for model calibration.

	Alternative 0	Alternative 1
Count	16.0	25.0
Frequency	39.0%	61.0%

Frequency table for observed choices.

	choice	Importance of easily customize plans	importance of insurance company to have a large market cap	Age = 23-27	Age = Above 35	Gender = Female	Gender = Male	Current health status	Health conscious	Marital status = Single	Marital status = Married	No. of Dependents = 0	No. of Dependents = 1-2	No. of Dependents = 4+
Respondent 1	1	3	1	1	0	1	0	6	7	1	0	1	0	0
Respondent 2	1	6	7	0	0	0	1	8	7	1	0	0	1	0
Respondent 3	1	8	10	0	0	0	1	10	10	1	0	1	0	0
Respondent 4	1	8	10	1	0	0	1	9	8	1	0	1	0	0
Respondent 5	1	7	8	1	0	1	0	10	10	1	0	1	0	0
Respondent 6	1	8	7	1	0	0	1	7	3	1	0	0	0	1
Respondent 7	0	8	7	0	0	1	0	7	6	1	0	1	0	0
Respondent 8	1	8	7	1	0	0	1	0	7	1	0	1	0	0
Respondent 9	0	7	8	1	0	1	0	9	8	1	0	1	0	0
Respondent 10	0	5	5	1	0	0	0	3	2	1	0	1	0	0

Calibration data (excerpt).

Model results

Model parameters

Model parameters reported here have been estimated on the entire calibration data.

	Parameter	Standard deviation	P-value
Intercept	-6.8883	81.1724	0.9324
`Importance of easily customize plans `	0.4831	0.2377	0.0421
` importance of insurance company to have a large market cap`	-0.2461	0.2325	0.2899
`Age = 23-27`	-0.4967	1.3730	0.7175
`Age = Above 35`	-3.5733	1.9638	0.0688
`Gender = Female`	9.5664	81.0881	0.9061
`Gender = Male`	9.1420	81.0964	0.9102
`Current health status`	-0.6772	0.4059	0.0953
`Health conscious`	0.5234	0.3507	0.1356
`Marital status = Single`	2.9434	2.4393	0.2276
`Marital status = Married`	3.5140	2.4002	0.1432
`No. of Dependents = 0`	-4.8938	2.9568	0.0979
`No. of Dependents = 1-2`	-2.3239	2.6238	0.3758
`No. of Dependents = 4+`	-2.2144	3.1754	0.4856

Model statistics. For identification purposes, parameters for the alternative 0 have been fixed to 0. These are not reported here. P-value = probability that parameter estimate is different from zero only by chance.

Confusion matrix and hit rate

Model performance is assessed using 10-fold cross-validation. The model is first estimated on 9/10 of the calibration data, and model performance is assessed on the remaining 1/10 of the data. The process is repeated 10 times, with perfect replacement. The results are then combined and reported here.

	Predicted 0	Predicted 1	Total
Actual 0	6	10	16
Actual 1	8	17	25
Total	14	27	41

Confusion matrix (count). The model has correctly classified 23 of the 41 observations. The off-diagonal elements are classification errors.

	Predicted 0	Predicted 1
Actual 0	38%	63%
Actual 1	32%	68%

Confusion matrix (%). The global hit rate of the model is 56%. The diagonal elements represent alternative-specific hit rates.

Model predictions

	Prob. 0	Prob. 1	Predicted	Actual	Correct
1	81%	19%	0	1	no
2	8%	92%	1	1	yes
3	78%	22%	0	1	no
4	91%	9%	0	1	no
5	41%	59%	1	1	yes
6	39%	61%	1	1	yes
7	0%	100%	1	0	no
8	0%	100%	1	1	yes
9	53%	47%	0	0	yes
10	100%	0%	0	0	yes

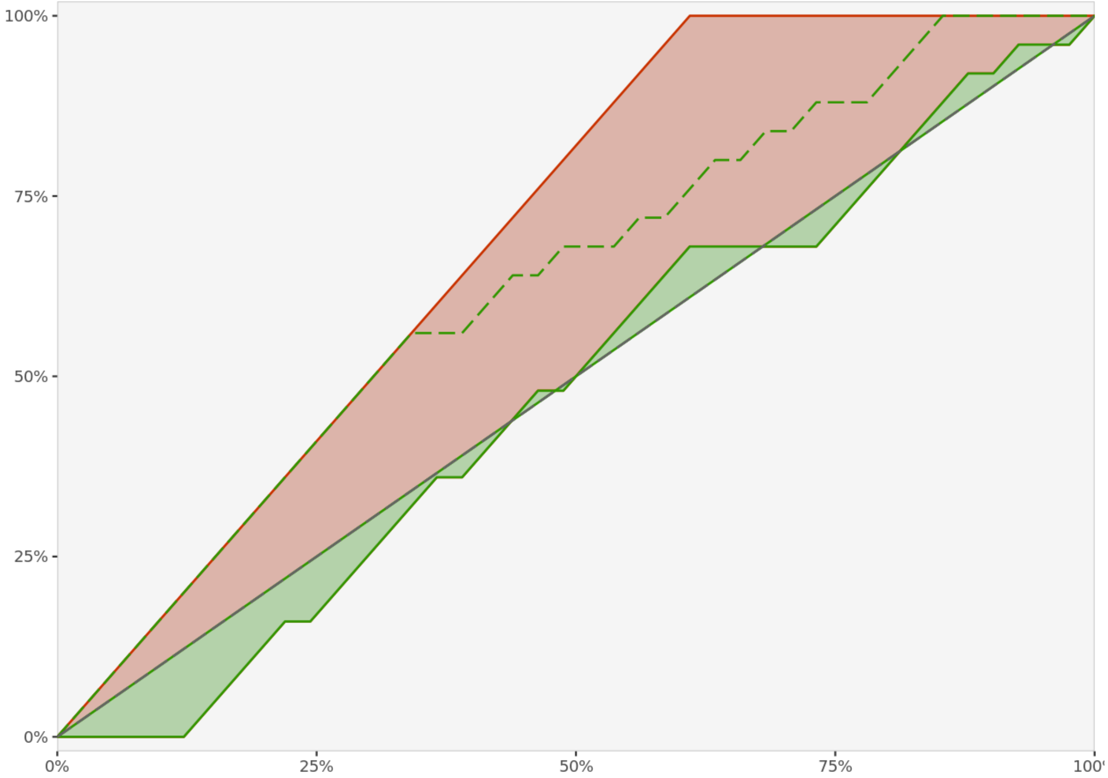
Model predictions (in-sample) (excerpt).

Gain chart and lift

A gain chart is a representation of how good a predictive model is at identifying the most favorable responses.

The X-axis represents the percentile level of the population ordered in decreasing order of choice likelihood, and the Y-axis represents the percentage of the actual favorable choices recovered by the model.The diagonal line represents performance of a model that randomly assigns the choice an individual makes from the set of choice alternatives. The red line represents a true model that predicts the choices perfectly. The green line represents the performance of the focal model.

The dashed green line represents the gain chart obtained on the entire calibration data, without cross-validation, whereas the green area represents the same obtained by cross-validation. The latter sometimes provides degraded but more realistic performance results.



Gain chart. The gain chart represents the expected performance of the model in predicting the favorable choice.

Lift is defined as the improvement in model performance at different percentile levels. If by selecting the top 10% of the ordered list, we can reach 0% of the individuals who make the appropriate choice (i.e., respond favorably), the focal model performs 0 times better than a model that makes random assignments. In that case, the lift at the 10-percentile level is 0.

The 'truth' is the true number of favorable responses in the ordered list. Improvement defines how well the truth is recovered by the model. An improvement of 100% means that all the favorable responses were recovered perfectly.

	Top 5%	Top 10%	Top 25%
Random	5.0%	10.0%	25.0%
Truth	8.2%	16.4%	41.0%
Model	0.0%	0.0%	17.0%
Observed lift	0.000	0.000	0.680
Improvement	-156.3%	-156.3%	-50.0%

Predicted lift and improvement ratios.

Elasticities

Elasticity is a measure of how responsive a target variable is to a change in the value of a predictor. Specifically, elasticity is defined as a ratio of percentage change in the target variable (Y) in response to a specified % change in predictor (X), such that Elasticity = (% change in Y) / (% change in X).

To compute the elasticities, Enginius follows these steps:

- Predict the target variable Y at the individual level for the current values of X. Average Y across respondents to obtain Y0.
- Increase the values of X (for each observation) by 1%, and predict the target variable Y at these new values. Average across respondents to obtain Y1.
- Compute elasticities as (Y1 – Y0) / Y0.

Keep in mind that, when X is discrete, an increase of 1% in X might be meaningless. For instance, if X = 1 means that the color is red, X = 1.01 has no useful interpretation, in which case elasticity computations do not lead to interpretable results.

	0	1
Importance of easily customize plans	-1.4729%	0.9428%
importance of insurance company to have a large market cap	0.7476%	-0.4785%
Age = 23-27	0.0992%	-0.0635%
Age = Above 35	0.5656%	-0.3620%
Gender = Female	-1.7700%	1.1329%
Gender = Male	-1.9381%	1.2405%
Current health status	2.0811%	-1.3320%
Health conscious	-1.6529%	1.0579%
Marital status = Single	-0.8763%	0.5609%
Marital status = Married	-0.2468%	0.1580%
No. of Dependents = 0	1.3449%	-0.8608%
No. of Dependents = 1-2	0.1750%	-0.1120%
No. of Dependents = 4+	0.0825%	-0.0528%

Elasticities. Changes in the relative likelihood of choosing each alternative after a 1% increase in the current value of the predictors. Elasticities larger than 1% in absolute value are color-coded.

Impact of changes in predictors

Because elasticities are expressed in relative terms to a baseline, we report here the changes in probabilities of choosing each alternative after a 1% increase in the current value of each predictor.

	Prob 0	Prob 1	Change in 0	Change in 1
Initial	0.390264	0.609736	N/A	N/A
Change in Importance of easily customize plans	0.384516	0.615484	-0.005748	0.005748
Change in importance of insurance company to have a large market cap	0.393182	0.606818	0.002918	-0.002918
Change in Age = 23-27	0.390651	0.609349	0.000387	-0.000387
Change in Age = Above 35	0.392471	0.607529	0.002207	-0.002207
Change in Gender = Female	0.383356	0.616644	-0.006908	0.006908
Change in Gender = Male	0.382700	0.617300	-0.007564	0.007564
Change in Current health status	0.398386	0.601614	0.008122	-0.008122
Change in Health conscious	0.383813	0.616187	-0.006451	0.006451
Change in Marital status = Single	0.386844	0.613156	-0.003420	0.003420
Change in Marital status = Married	0.389301	0.610699	-0.000963	0.000963
Change in No. of Dependents = 0	0.395513	0.604487	0.005249	-0.005249
Change in No. of Dependents = 1-2	0.390947	0.609053	0.000683	-0.000683
Change in No. of Dependents = 4+	0.390586	0.609414	0.000322	-0.000322

Absolute changes. Changes in the probability of choosing each alternative after a 1% increase in the value of each predictor.