Harshaanth Thiyagaraja Kumar

Use the file of summer products with rating and performance to build a model and find factors that could affect whether the product is popular.

Variable definitions could be found in the file.

Run the logistic regression, don't forget to **split** data into training and validation datasets, and use stepwise for **variable selection**. Choose the **best** model among models recommended by variable selection. Report the selected model's

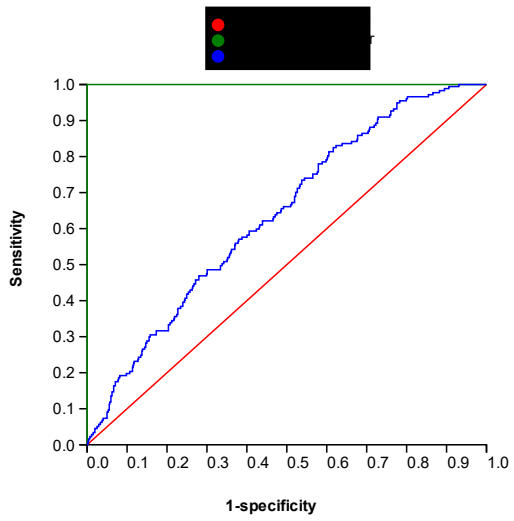
1. Regression Model (write down the model in **three** forms and paste the coefficients table)

Predictor	Estimate	Confidence Interval: Lower	Confidence Interval: Upper	Odds	Standard Error	Chi2-Statistic	P-Value
Intercept	-10.866959	-14.9674187	-6.766499919	1.91E-05	2.09210956	26.98032854	2.06E-07
merchant_has_profile	0.65805694	0.224159788	1.091954082	1.931037	0.221380163	8.835874396	0.002954
rating	0.52099081	0.098095774	0.943885846	1.683695	0.215766738	5.8303031	0.015752
merchant_rating	1.80936119	0.816950658	2.801771714	6.106545	0.506341206	12.76920901	0.000352

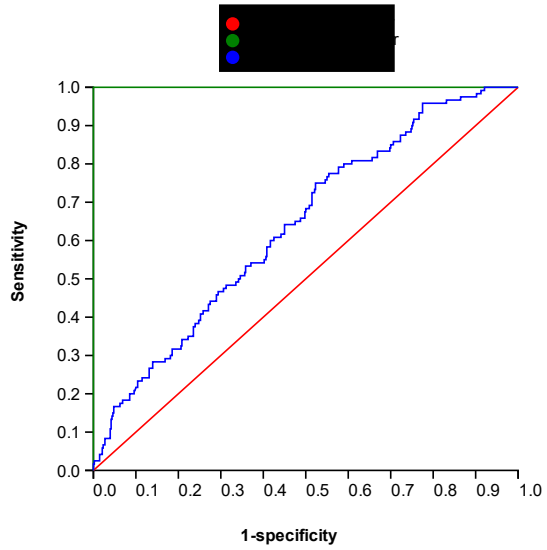
Logit(popular=1)=-10.87+0.66merchant_has_profilepicture+0.52rating+1.82merchant_rating
Odds(popular=1)=e^(-10.87+0.66merchant_has_profilepicture+0.52rating+1.82merchant_rating)
=1.91e-05*(1.93^merchant_has_profilepicture)*(1.68^rating)*(6.11^merchant_rating)
P(popular=1)=1/(1+e^-(-10.87+0.66merchant_has_profilepicture+0.52rating+1.82merchant_rating))

2. Explain the coefficients (**all** of them)
   If a product is sold by a merchant with a profile picture, the odds of this product to be popular would be 1.93 time the odds of a product sold by a merchant with no picture to be popular holding other variables constant.
   If the rating of the product increases by 1 star, the odds of the product to be popular would be multiplied by 1.68 holding other variables constant.
   If the merchant rating increases by 1 star, the odds of the product to be popular would be multiplied by 6.11 holding other variables constant.

3. How does this model work? (hint: check the ROC curve and compare ROC curve of training and validation)
   Training AUC=0.638



Validation AUC=0.636



This model works well. The auc is larger than 0.5, so the model is better than the benchmark. And the performances are pretty similar in the training and the validation data, so it will work well on new dataset too.
(some version of xlminer may not show auc, it's fine if they just compare the roc curve)

4. Using validation score, what's the error rate of class 0, the error rate of class 1, the precision, and recall. Please also show the calculation steps.

   Error rate of class 0: 0/480=0
   Error rate of class 1:120/120=1
   Precision:0/0
   Recall: 0/120=0

Confusion Matrix		
Actual\Predicted	0	1

0	480	0
1	120	0

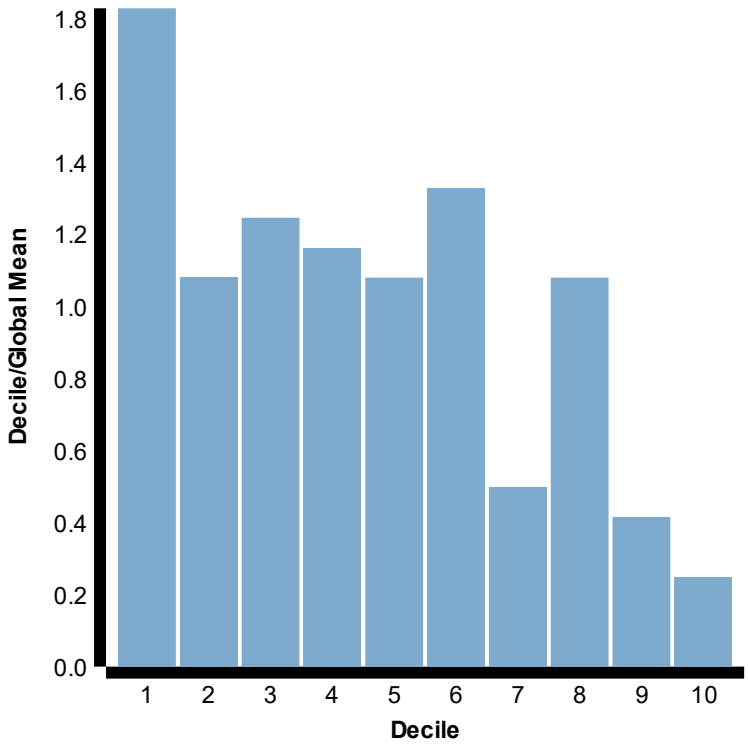
Error Report			
Class	# Cases	# Errors	% Error
0	480	0	0
1	120	120	100
Overall	600	120	20

Metrics	
Metric	Value
Accuracy (#correct)	480
Accuracy (%correct)	80
Specificity	1
Sensitivity (Recall)	0
Precision	N/A
F1 score	N/A
Success Class	1
Success Probability	0.5

5. Do the results help solve the problem to identify which products are more likely to be popular? If not, why?
   No, all the records are classified as 0.

6. Using the validation decile chart to choose a new cutoff value. (hint: consecutively higher than 1). Run the regression again using the new cutoff value. Using validation score, report the new error rate of class 0, error rate of class 1, precision, and recall. Please also show the calculation steps.

Harshaanth Thiyagaraja Kumar



Based on the decile chart, I choose use decile 6 as the cutoff. In total, we have 600 records in the validation data. After order predicted probability of belonging to class 1, I choose the 360th largest probability which is in row 396 (600*0.6+36 (the starting row)). I use 0.163 as the cutoff value for the new regression.
Error rate of class 0: 269/480=0.56
Error rate of class 1:27/120=0.225
Precision:93/(269+93)=0.257
Recall: 93/120=0.775

Confusion Matrix

Actual\Predicted	0	1
0	211	269
1	27	93

Error Report

Class	# Cases	# Errors	% Error
0	480	269	56.04166667
1	120	27	22.5
Overall	600	296	49.33333333

Metrics

Metric	Value

Harshaanth Thiyagaraja Kumar

Accuracy (#correct)	304
Accuracy (%correct)	50.66667
Specificity	0.439583
Sensitivity (Recall)	0.775
Precision	0.256906
F1 score	0.385892
Success Class	1
Success Probability	0.163

7. Using the results of question 6, explain precision and recall in your own words.

In total, we classify 362 products would be popular, out of which, 93 products were truly popular. 25.7% of predicted popular products are correctly classified.
77.5% of popular products are correctly classified.