HEC MONTRÉAL

Theme 4

Predictions for binary variables
(with logistic regression)

Predictions for a binary variable

Often, the variable to predict may be binary:

Will this individual:

- buy or not?
- pay his loan back or default?
- leave the company in the next three months?
- contribute to our fundraising campaign?
- claim from his insurance contract?

In this context, the fundamental principles of data mining still apply:

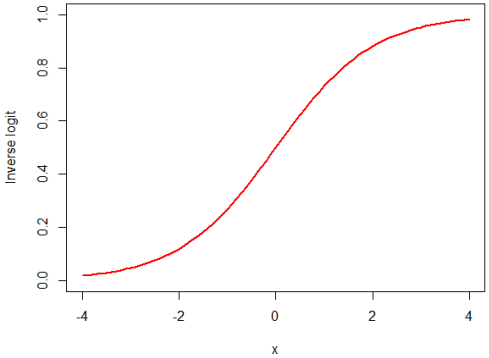
- Need historical data,
- Only use variables that will be available at the time of making predictions,
- Base model assessment on validation data, never used to train the model.

Logistic regression model

$$\text{logit}\{P(Y = 1|X_1, \ldots, X_p)\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

where $\text{logit}(p) = \ln\left(\frac{p}{1-p}\right)$.

or $\text{logit}^{-1}(y) = e^y/(1 + e^y)$



Right-hand side:

- Usual linear model.

Left-hand side:

- Model for $P(Y = 1|X_1, \ldots, X_p)$ rather than Y.
- A link function maps the values of the linear part (ranging from $-\infty$ to $+\infty$) to a probability scale (values between 0 and 1).

Other link functions are sometimes considered, but logit is the most popular.

The logit link is the most popular because of its interpretation in inference.

A probability p may also be expressed in terms of odds:

$$odds = \frac{p}{1-p}$$

Odds are between 0 and ∞ rather than $[0,1]$. This allows multiplicative effects

Consider the effect of a change in variable X_1 by comparing:

$$p = P\left(Y = 1 \middle| \boldsymbol{X_1 = x_1 + 1}, X_2 = x_2 \dots, X_p = x_p\right)$$

$$q = P(Y = 1 | \boldsymbol{X_1 = x_1}, X_2 = x_2 \dots, X_p = x_p)$$

The odds of q are $\frac{q}{1-q} = e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}$ and similarly for p.

the coefficients of a logistic regression can therefore be interpreted as odds ratios:

$$\frac{p(1-q)}{q(1-p)} = \frac{e^{\beta_0+\beta_1(X_1+1)+\cdots+\beta_p X_p}}{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}} = e^{\beta_1}$$

Estimates for the logistic regression model are found through maximum likelihood.

$$\ell(\beta_0, \ldots, \beta_p) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{(1-y_i)}$$

where y_i is the response variable for individual i and

$$p_i = \frac{e^{\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}}}{1 + e^{\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}}}$$

Contrarily to the linear regression, the equations do not have a closed-form solution, and numerical algorithms are used to find the optimal values.

In all cases, the prediction made by the model is an estimate of the probability that the response variable is equal to 1.

Example: Line of credit

Source: SAMPSIO library from SAS. Existing clients ($n = 5960$, of which we choose 2000 randomly for validation) of a bank ask for a line of credit on their house.

Variable	Description
PAID	1 = paid, 0 = defaulted
CLAGE	Age in months of the oldest credit line
CLNO	Number of credit lines
DEBTINC	Debt-to-income ratio
DELINQ	Number of delinquent credit lines
DEROG	Number of major derogatory reports
JOB	Job category
LOAN	Amount of current loan request
MORTDUE	Amount due on existing mortgage
NINQ	Number of recent credit inquiries
REASON	Home improvement or debt consolidation
VALUE	Value of current property
YOJ	Years on current job

The dataset has some missing values that are imputed.

Imputation steps must be determined on the training set, then applied to other sets (validation/test) when needed. Imputation is the topic of a later class.

Often, the bare fact that a value is missing may be informative. For that reason,

`MISS_xxxx` is an indicator that `xxxx` was missing.

To decide who should get a loan, we:

- Predict the probability that the client will repay,
- Identify the best clients based on that value.

On the right, the top probabilities when using all variables with imputed values, but without the `MISS_xxxx` variables.

	ID	Prob_Pay
1008	3680	0.9994608
1625	3929	0.9963288
1435	4418	0.9963280
1492	3869	0.9955650
685	2834	0.9941442
1913	3014	0.9938280
101	5730	0.9921831
1825	5687	0.9916714
826	3857	0.9913847
1658	5756	0.9912081
1068	5643	0.9893241
507	2388	0.9859544
749	3856	0.9858357
452	1856	0.9855210
150	2193	0.9850751
843	2114	0.9847823
1084	4258	0.9841383
1811	2967	0.9838700
1885	867	0.9830795

Suppose that anyone with a predicted chance of paying above 50% gets a loan.
The confusion matrix below summarizes the situation.

		Prediction		
		0	1	Total
Truth	0	123	279	402
	1	64	1534	1598
Total		187	1813	2000

Misclassification rate: (279 + 64) / 2000 = 17.15%
Precision = P(truth is 1 | prediction is 1) = 1534 / 1813 = 84.61%
Sensitivity = P(predict 1 | truth is 1) = 1534 / 1598 = 95.99%
Specificity = P(predict 0 | truth is 0) = 123 / 402 = 30.60%

- The matrix and the values shown are based on the **validation** set.
- Despite using a model, we reject only 30.60% of bad payers.
- Almost 16% of selected individuals will not repay…
- Should we choose a more conservative threshold? For instance, granting a loan only to those who show a probability of repaying of 90% or more?

With a 90% threshold:

		Prediction		
		0	1	Total
Truth	0	352	50	402
	1	878	720	1598
Total		1230	770	2000

Misclassification rate: (50 + 878) / 2000 = 46.40%

Precision = P(truth is 1 | predict 1) = 720 / 770 = 93.51%

Sensitivity = P(predict 1 | truth is 1) = 720 / 1598 = 45.06%

Specificity = P(predict 0 | truth is 0) = 352 / 402= 87.56%

- The matrix and the values shown are based on the **validation** set.
- The misclassification rate is going up.
- Specificity is going up.
- We now decline a loan to 87.56% of bad payers.
- But fewer loans were given (770 now vs. 1813 previously).

A large number of other statistics may also be relevant.

A long list is available on Wikipedia, for instance.

The first part of the list covers all conditional probabilities.

Most have multiple synonyms.

		Prediction		
		0	1	Total
Truth	0	**TN**	**FP**	**N**
	1	**FN**	**TP**	**P**

condition positive (P)
the number of real positive cases in the data
condition negative (N)
the number of real negative cases in the data

true positive (TP)
eqv. with hit
true negative (TN)
eqv. with correct rejection
false positive (FP)
eqv. with false alarm, Type I error
false negative (FN)
eqv. with miss, Type II error

sensitivity, recall, hit rate, or true positive rate (TPR)
$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{P}} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} = 1 - \mathrm{FNR}$$
specificity, selectivity or true negative rate (TNR)
$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{N}} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} = 1 - \mathrm{FPR}$$
precision or positive predictive value (PPV)
$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} = 1 - \mathrm{FDR}$$
negative predictive value (NPV)
$$\mathrm{NPV} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FN}} = 1 - \mathrm{FOR}$$
miss rate or false negative rate (FNR)
$$\mathrm{FNR} = \frac{\mathrm{FN}}{\mathrm{P}} = \frac{\mathrm{FN}}{\mathrm{FN} + \mathrm{TP}} = 1 - \mathrm{TPR}$$
fall-out or false positive rate (FPR)
$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{N}} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} = 1 - \mathrm{TNR}$$
false discovery rate (FDR)
$$\mathrm{FDR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TP}} = 1 - \mathrm{PPV}$$
false omission rate (FOR)
$$\mathrm{FOR} = \frac{\mathrm{FN}}{\mathrm{FN} + \mathrm{TN}} = 1 - \mathrm{NPV}$$
Threat score (TS) or Critical Success Index (CSI)
$$\mathrm{TS} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN} + \mathrm{FP}}$$

To tell the expressions:

$$\left\{\begin{matrix}\text{true}\\\text{false}\end{matrix}\right\}\ \left\{\begin{matrix}\text{positive}\\\text{negative}\end{matrix}\right\}\ \text{rate}$$

apart, remember that:

- Positive/negative refers to the prediction made
- The rates are conditional on the truth.

For instance, a "false positive rate" is

$$P(\text{predict } 1|\text{truth } 0)$$

Which is also $1 - \text{specificity}$.

Similarly, since sensitivity is

$$P(\text{predict } 1|\text{truth } 1)$$

It is also a "true positive rate."

Are there statistics that do not require a threshold to change probabilities into classification?

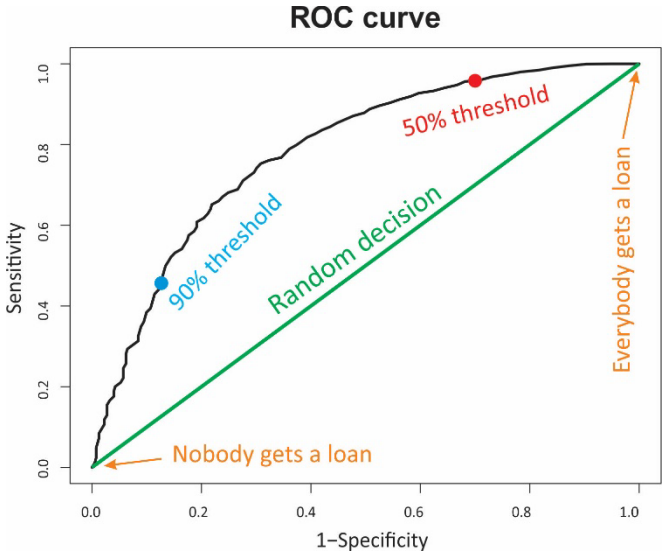
The ROC curve

Varying the threshold makes sensitivity and specificity move in opposite directions.

All possible thresholds are considered in a ROC curve.

Better models get closer to the point (0%, 100%).

The area under the ROC curve (AUC) is used to summarize the ability of a model to predict a binary variable. AUC is typically between 0.5 and 1.

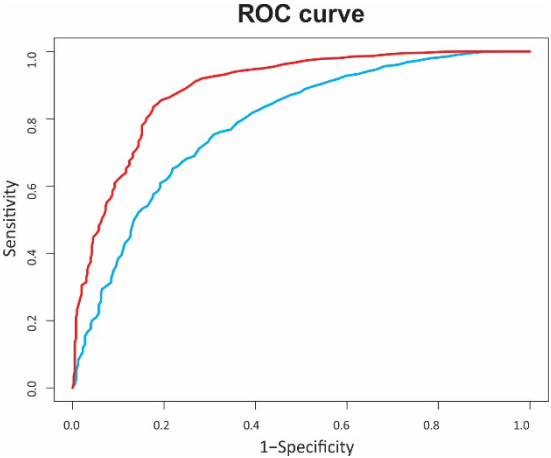


A larger AUC means a better model.

The ROC curve on the right shows the sensitivity and specificity on the **validation** set.

Let us compare two logistic models based on the ROC curve:

- Model 1: Predict PAID with all variables
- Model 2: Predict PAID with all variables as well as `MISS_xxxx`



The model is fitted on the training set; The validation set is used for the curve.

The AUC goes from 0.7848 to 0.8913.

For all thresholds, sensitivity and specificity are better under Model 2.

AUC is used to compare many models similarly to RMSE for linear regression.

Here, adding the indicator variables improves the performance of the model a lot.

Cumulative lift chart

This cumulative lift is the ratio:

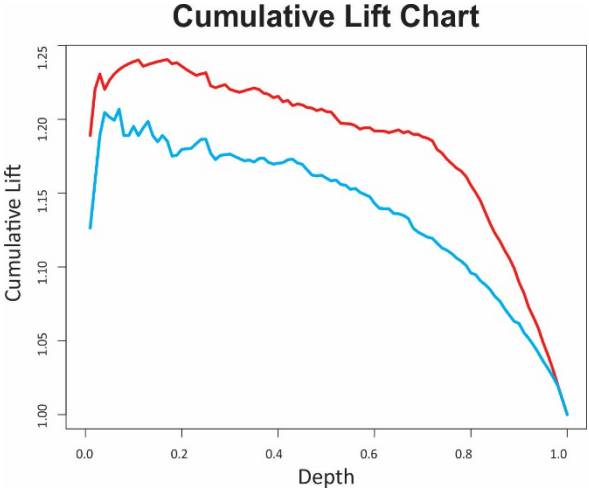
$$\frac{P(\text{truth = 1} \mid \text{in top x\%})}{P(\text{truth} = 1 \text{ in the population})} = \frac{P(\text{good payer} \mid \text{in top x\%})}{P(\text{good payer in the population})}$$

`Depth` is the percentage of people who get the offer.

The lift says how much more "ones" we get compared to picking at random.

Here again, Model 2 is better since it has a higher lift at all depth.

If we give loans to the best 20% (depth), we find a lift of 1.18 or 1.24 depending on the model.



Cumulative lift charts are often preferred in a marketing context.

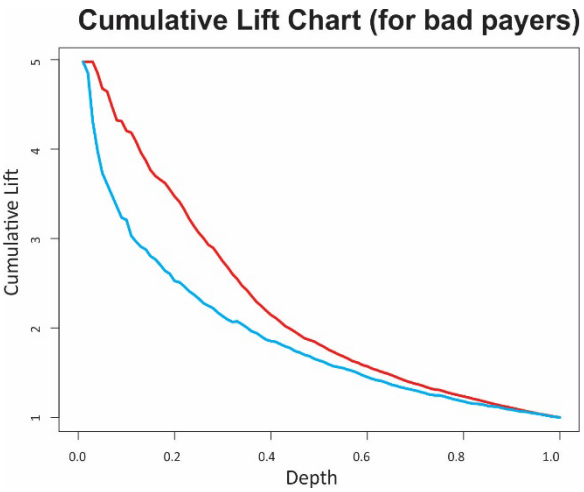
Be careful about their interpretation:

In this case, 80% of clients are good payers.
This means that a perfect model may not exceed a lift of 1.25.

With the same data, if we predict clients that are bad payers, a perfect model would have a lift of 5: there is only 20% of bad payers.

To appreciate a lift value, you need to know the prevalence of ones in the population.

Model 2 is still better, but the face value of the lift needs to be considered carefully.



Cost of errors

If the cost (or gain) associated with each possible decision is known, they can be used to determine the ideal threshold. The measure of performance is then the cost itself.

		Prediction	
		0	1
Truth	0	**A**	**C**
	1	**B**	**D**

The cost of each option may be entered directly. A gain should be negative.
For any threshold, the cost in each cell is multiplied by the number of individuals.

Cost can be determined for al thresholds, and the optimal solution prevails.

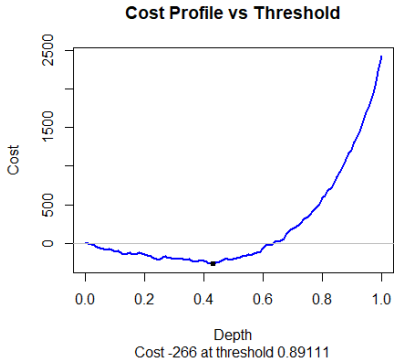
As usual:

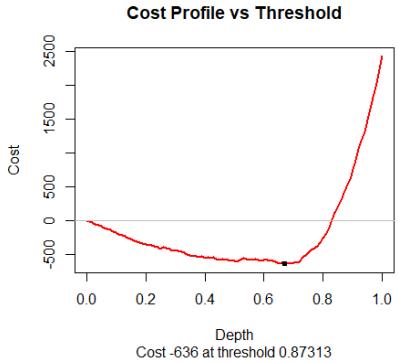
- The model is determined on the training set
- The costs are based on the validation set
- If needed, a test set could be used to confirm the cost of the final decision.

For instance, consider that:

- Not giving a loan: no cost, no gain
- A loan to a good payer is -1 (gain 1 dollar)
- A loan to a bad payer is 10 (lose some capital)

We get the following cost profiles:





Model 2 allows us to make more profit (more negative cost) by giving more (larger depth) better-targeted loans.

Alternative coding of the cost

The cost of errors are sometimes given only through:

- false positives
- false negatives

The cost must then include the unrealized profits of not sending an offer.

For the example above,

- a false positive costs $10 because we lend to a bad payer,
- a false negative costs $1 because we missed an opportunity to make a profit.

The ROCR package computes the costs this way.

If you are given a full matrix of costs, do not make the mistake of counting a cost twice by having simultaneously the profit and the unrealized profit – you would be counting the same cost twice.

Targeted marketing dataset

Let us consider one more dataset where clients from a streaming company are offered a promo to buy “extra” programs (such as sports).

Variable	Description
sale	The client bought an “extra” program with the offer
amount	The amount of that “extra” sale
female	0=male, 1=female
age	Age of client in years
revenue	Categories: 1= < \$35,000, 2= between \$35,000 and \$75,000, 3= > \$75,000
region	Categories from 1 to 5
spouse	Does the client have a significant other (0=no, 1=yes)
yearclient	Number of years that the individual has been a client of the company
weekslast	Number of weeks since last “extra” purchase
amountlast	Amount of that last “extra” purchase
amountyear	Amount spent by the client in the last 12 months
salesyear	Number of purchases in the last 12 months
train	Indicator that the client was in the pilot study (training set)
test	Indicator that the client received the offer after the pilot study (test set)

A group of 1,000 clients are contacted as a pilot study.
The company has 100,000 additional clients.

In this case, the company decided to extend the offer to all clients after the pilot, so the 100,000 clients become a large test set for our purpose.

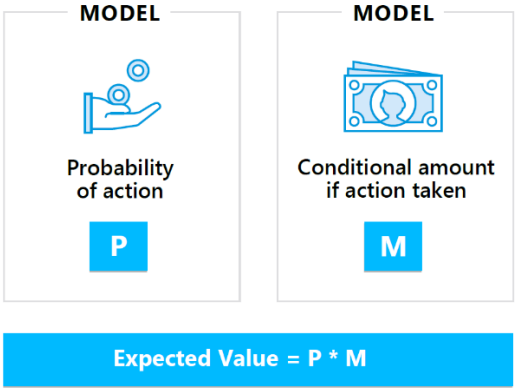
In this dataset, there are two target variables:

- Sale: did the client buy or not?
- Amount: For those who bought, how much did they spend?

Logistic regression may be used on sale, but it does not represent the value of the sale, just the probability that it happened.

Two-stage modeling

When the value of an action is known, two-stage modeling is an option:



We build two models with the usual methodology:

- For the probability of the action (on all the data)
- For the amount when the action is taken (only on those who took action)

Both models are used on all individuals who need a prediction.
The product of the estimates is the expected value generated.

Two-stage modeling implies that two models are built.

- Use the same training (or same cross-validation folds) for both models,
- You may develop both models separately using the methodology discussed,
- The best model for P and the best model for M are combined,

This means that performances are evaluated on P and M, not on their product.

Since we do not measure the real PM, it is not possible to use a measure of performance that compares the prediction of the combined models to their true value.

Instead of choosing the individual with higher probabilities, we can now select those with the highest value.

	M	p	ExpectedSale
2	105.12514	1.238064e-03	1.301517e-01
4	73.03445	9.276374e-01	6.774949e+01
7	80.80522	1.200138e-04	9.697745e-03
8	51.82198	1.189068e-01	6.161983e+00
11	103.05571	5.912387e-01	6.093053e+01
13	85.10802	1.426742e-01	1.214272e+01
18	77.53900	4.870717e-01	3.776706e+01
23	94.73731	1.033291e-02	9.789118e-01
25	77.43768	2.165793e-01	1.677140e+01
27	88.34334	1.233666e-04	1.089861e-02
29	57.61103	6.479182e-04	3.732723e-02
32	53.34413	5.430038e-01	2.896607e+01
34	70.25287	9.682783e-02	6.802433e+00
39	65.07016	2.700422e-04	1.757169e-02
45	81.84774	5.554404e-02	4.546155e+00
55	57.89081	1.350211e-01	7.816484e+00
62	85.22932	3.450822e-04	2.941112e-02
66	65.37518	8.892922e-01	5.813763e+01
78	74.08307	8.890905e-02	6.586655e+00
80	101.90933	4.770037e-04	4.861112e-02
85	83.07768	3.212491e-03	2.668863e-01
89	78.58323	2.066424e-02	1.623863e+00
93	67.87645	3.787482e-03	2.570809e-01
98	70.30380	1.404978e-03	9.877530e-02