MUNI
ECON

Artificial Intelligence in Finance

Supervised learning - discrete outcomes part A

Štefan Lyócsa

Department of Finance, Faculty of Economics and Administration

October 30, 2024

Outline for Section 1

Introduction

Instead of predicting a specific value (on an interval) for a **continuous target** variable, we might want to predict a **qualitative variable** (e.g. color, social status, ...). More broadly, we are interested in **classification problems**:

- The patient is: i) healthy, ii) has a common cold, iii) flue, iv) COVID-19 or v) something else?
- The respondent is willing to vote for candidate: i) A, ii) B, ...
- Is it likely that the company will have financial distress (1 - yes, 0 - no)?
- Is the customer going to buy the product (1 - yes, 0 - no)?
- Is the borrower going to repay the loan (1 - yes, 0 - no)?
- Is the price going up (1 - yes, 0 - no)?

Introduction

A specific case of a classification problem is related to a **binary decision** (1 - Yes, 0 - No).
Classification **is distinct** from continuous outcome prediction. We have **different models** and a different concepts of **what constitutes a good prediction**. Some methods:

- Logistic regression.
- Penalized logistic regressions:
  - LASSO.
  - RIDGE.
  - Elastic net.
- Tree based methods.
- Support Vector Machines.
- K-Means clustering (sort of).
- Neural networks and other methods...

Outline for Section 2

Probability linear model

Let $Y_i, i = 1, 2, ..., n$ denote a bi-variate outcome $1 - survived, 0 - not\ survived$ sinking of the Titanic and X_i the age of the person. The following model is the **probability linear model** and is estimated via OLS:

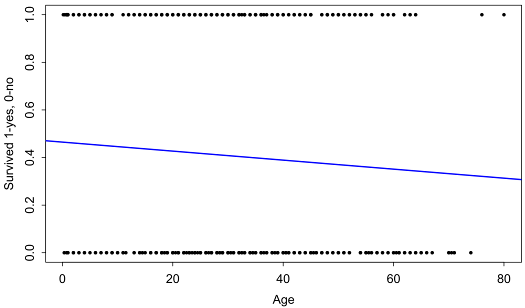
$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{1}$$

with following estimates:

$$Y_i = 0.46 - 0.001894 X_i + \hat{\epsilon}_i \tag{2}$$

Probability linear model

The estimated regression line shows you why such a model **might not be the best idea**:



Issues:

- The model is **heteroscedastic almost by design**.
- Predicted values might **fall below 0** and **exceed 1**.

Binary logistic regression

Ideally you might want to model the probability directly. Let $h(\boldsymbol{X}_i; \beta)$ be a **link-function** that includes k features in vector $\boldsymbol{X}$ and corresponding k parameters in vector β. The probability that event happens $Y_i = 1$ is:

$$P(Y_i = 1|\boldsymbol{X}_i; \beta) = h(\boldsymbol{X}_i; \beta) \in [0, 1] \tag{3}$$

For probability that event will **not** happen we have:

$$P(Y_i = 0|\boldsymbol{X}_i; \beta) = 1 - P(Y_i = 1|\boldsymbol{X}_i; \beta) = 1 - h(\boldsymbol{X}_i; \beta) \tag{4}$$

We can combine both equations into:

$$P(Y_i|\boldsymbol{X}_i; \beta) = h(\boldsymbol{X}_i; \beta)^{Y_i}(1 - h(\boldsymbol{X}_i; \beta))^{(1-Y_i)} \tag{5}$$

This is a Bernoulli trial.

Binary logistic regression

The Bernoulli trial:

$$P(Y_i|\boldsymbol{X}_i;\beta) = h(\boldsymbol{X}_i;\beta)^{Y_i}(1-h(\boldsymbol{X}_i;\beta))^{(1-Y_i)} \tag{6}$$

assuming **independence** between outcomes, leads to a **Binomial process** and we can combine multiple observations of the outcome into a **likelihood function**:

$$L(\beta) = P(Y|\boldsymbol{X};\beta) = \prod_{i=1}^{n} h(\boldsymbol{X}_i;\beta)^{Y_i}(1-h(\boldsymbol{X}_i;\beta))^{(1-Y_i)} \tag{7}$$

The goal is to find such parameters of β that lead to the highest possible value of the $L(\beta)$. Why?

Binary logistic regression

The maximization process is over parameters $\boldsymbol{\beta}$:

$$\max_{\boldsymbol{\beta}} \rightarrow L(\boldsymbol{\beta}) = \prod_{i=1}^{n} h(\boldsymbol{X}_i; \boldsymbol{\beta})^{Y_i}(1 - h(\boldsymbol{X}_i; \boldsymbol{\beta}))^{(1-Y_i)} \tag{8}$$

Instead of working with the product a more convenient method is to use **log-likelihood**:

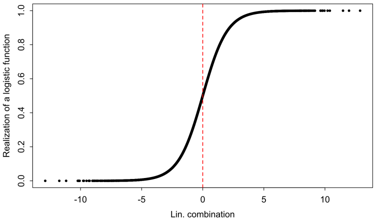
$$\max_{\boldsymbol{\beta}} \rightarrow LL(\boldsymbol{\beta}) = \sum_{i=1}^{n} Y_i log[h(\boldsymbol{X}_i; \boldsymbol{\beta})] + (1 - Y_i) log[(1 - h(\boldsymbol{X}_i; \boldsymbol{\beta}))] \tag{9}$$

We have to figure out, how should the $h(.)$ function look like. A popular option is a form of a **sigmoid function**.

Binary logistic regression

Specifically, a popular option is the **logistic function**; hence the **logistic regression**. Let denote $\sum_{j=1}^{k} \beta_j X_{i,j}$ simply as x. The logistic function has a form:

$$P_i = P(Y_i = 1|\boldsymbol{X}_i; \boldsymbol{\beta}) = h(.) = \frac{1}{1+e^{-x}} = \frac{e^x}{1+e^x} \tag{10}$$



$$\max_{\boldsymbol{\beta}} \rightarrow LL(\boldsymbol{\beta}) = \sum_{i=1}^{n} Y_i \left(\sum_{j=1}^{k} \beta_j X_{i,j} \right) - log\left(1 + e^{\sum_{j=1}^{k} \beta_j X_{i,j}}\right) \tag{11}$$

Binary logistic regression

Recall: In machine learning applications we do not care so much about parameter estimates. Still if you want to interpret coefficients, remember that from $h(.)$ you can express P_i. A popular approach is to look at **odds**:

$$O_i = \frac{P_i}{1 - P_i} = e^x \tag{12}$$

This looks better, now taking the (natural) log leads to the **logit**:

$$log\left(\frac{P_i}{1 - P_i}\right) = log\left(e^x\right) = x \tag{13}$$

and it looks similar to a linear regression.

Outline for Section 3

Prediction model

Turning back to the survivors of the sinking of the titanic (yR0lWICH3rY). We use a training sample and consider the following specification (with estimates):

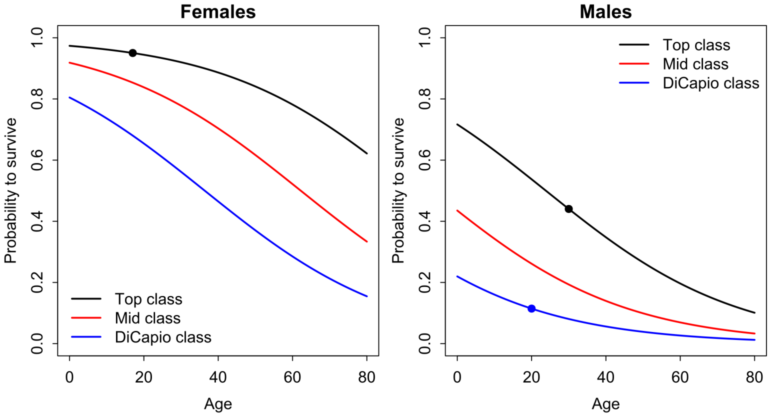
$$\begin{aligned}\sum_{j=1}^{k} \hat{\beta}_j X_j &= -1.27 + 2.19 Top_i + 1.01 Mid_i + 2.68 Female_i \\ &\quad -0.04 Age_i + 0.94 Parent_i\end{aligned}$$

- How would you estimate the effect of Age on the probability of surviving?
  - Use logistic function.
  - The effect is **non-linear** and depends on other variables!

Prediction model

Assuming that the person has following characteristics, Top = 1, Mid = 0, Female = 1, Age = 30, Parent = 0, the probability tu survive is given by:

$$0.92 = \left(1 + e^{-1\times(-1.27+2.19+2.68-0.04\times 30)}\right)^{-1} \tag{14}$$



How accurate are forecasts?

Consider $Y_i = 1$ to be a positive and $Y_i = 0$ a negative outcome. Let's the prediction of the probability be $\hat{p}_i$ and assume to have a given threshold $p_T \in (0, 1)$ such, that if $\hat{p}_i > p_T \rightarrow \hat{Y}_i = 1$. Given a sample of observations in the testing sample, $i = 1, 2, ..., n$ we can construct the following **confusion matrix**, predictions from *plm* and $p_T = 0.5$:

	Observed $Y_i = 0$	Observed $Y_i = 1$
Predicted $\hat{Y}_i = 0$	118	76
Predicted $\hat{Y}_i = 1$	4	12

- True positives? $TP = 12$.
- True negatives? $TN = 118$.
- False positives? $FP = 4$.
- False negatives? $FN = 76$.

How accurate are forecasts?

	Observed $Y_i = 0$	Observed $Y_i = 1$
Predicted $\hat{Y}_i = 0$	118	76
Predicted $\hat{Y}_i = 1$	4	12

- **Accuracy** $= \frac{TP+TN}{TP+TN+FP+FN} = \frac{118+12}{118+12+4+76} = 0.62$
- **Sensitivity** $=$ Recall $=$ TPR $= \frac{TP}{TP+FN} = \frac{12}{12+76} = 0.14$
- **Specificity** $=$ TNR $= \frac{TN}{TN+FP} = 0.97$
- **Precision** $= \frac{TP}{TP+FP} = \frac{12}{12+4} = 0.75$
- **Balanced accuracy** $= \frac{\text{Sensitivity}+\text{Specificity}}{2} = \frac{0.14+0.97}{2} = 0.55$
- **F1** $= 2 \times \frac{\text{Precision}\times\text{Sensitivity}}{\text{Precision}+\text{Sensitivity}} = 2 \times \frac{0.75\times0.14}{0.75+0.14} = 0.24$

How accurate are forecasts?

Compare the confusion matrix from the *plm* model:

	Observed $Y_i = 0$	Observed $Y_i = 1$
Predicted $\hat{Y}_i = 0$	118	76
Predicted $\hat{Y}_i = 1$	4	12

To the confusion matrix from the logistic regression:

	Observed $Y_i = 0$	Observed $Y_i = 1$
Predicted $\hat{Y}_i = 0$	102	29
Predicted $\hat{Y}_i = 1$	20	59

Which model leads to **better** predictions?

How accurate are forecasts?

Which model leads to **better** predictions? It depends right? Still the differences appear to be substantial:

Models	PLM	LR
Accuracy	0.62	**0.77**
Sensitivity	0.14	**0.67**
Specificity	0.97	0.83
Precision	0.75	**0.74**
Balanced accuracy	0.55	**0.75**
F1 score	0.24	**0.70**

Note, that the threshold $p_T = 0.5$ was set arbitrarily. In fact, it might be considered to be a **hyperparameter** that you need to tune using **cross-validation**.

Changing threshold

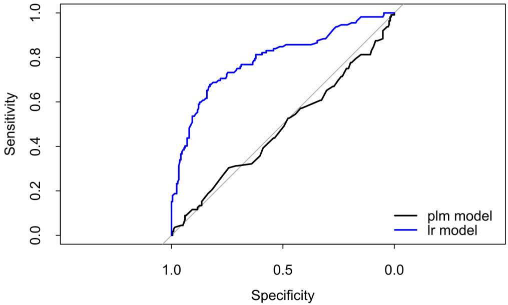
Let's change the threshold to $p_T = 0.45$.

Models	PLM	LR
Accuracy	0.54	**0.76**
Sensitivity	0.30	**0.69**
Specificity	0.72	0.80
Precision	0.43	**0.72**
Balanced accuracy	0.50	**0.75**
F1 score	0.35	**0.69**

Not an improvement for LR! The model is the same, only the threshold changed.

ROC

The Receiver Operating Characteristic curve displays two types of errors for all possible thresholds (James et al. 2018, [3]).



The overall performance of a classifier across all possible thresholds is the **area under the ROC**, denoted as **AUC**. In cases above $AUC_{plm} = 0.51$ and $AUC_{lr} = 0.84$.

Brier score

There are several popular alternatives to evaluate classification forecasts that can be used in the model confidence set framework as well. The **Brier** (1950, [1]) **score** for two class problems is given by:

$$S_B = n^{-1} \sum_{i=1}^{n} (\hat{p}_i - Y_i)^2 \tag{15}$$

, which is the mean squared error between the predicted probability ($\hat{p}_i$) and the observed outcome (Y_i).

In our examples above we have $S_{B,plm} = 0.24$ and $S_{B,lr} = 0.16$ and only the logistic regression model is in the set of superior models.

Cross entropy

The **Cross-entropy** is a quite popular measure for classification purposes. For two class problems it is given by (the lower the value the more accurate the model):

$$S_E = n^{-1} \sum_{i=1}^{n} - \left[log(\hat{p}_i) Y_i + log(1 - \hat{p}_i)(1 - Y_i) \right] \tag{16}$$

The two terms are switched on/off depending on whether the observed event happened or not. **You get penalized if you are confident and wrong**.

In our examples above we have $S_{E,plm} = 0.68$ and $S_{E,lr} = 0.48$ and only the logistic regression model is in the set of superior models.

Finance related cost functions

A threshold and loss functions should be driven by the **domain knowledge**. The mapping $D : \hat{p}_i \rightarrow \hat{Y}_i, \hat{Y}_i \in \{0, 1\}$ should not be driven by purely statistical measures.

Consider a loan market with three participants, lender, borrower and investor. Lender and investor are designing credit-scoring models.

- What should be the criterion for the lender?
- What should be the criterion for the investor?

Outline for Section 4

Intuition

In the titanic dataset, 40.82% survived. This is **not overly imbalanced**. However, using the Zopa dataset ('zsnew.csv'), we have only 8.155% of defaulted loans. This is a severaly imbalanced dataset, where the majority class (good loans) has significant representation in the data.

Imbalanced data might lead to **accuracy paradox**. Say you predict a stock to default in the next year. You have 99.5% of firms that have not defaulted (**majority** class) and only 0.05% that have (**minority** class):

- How accurate is a prediction that will unconditionally always predict a non-default (i.e. 0)?
- Your model will have a tendency to learn from mostly successful companies that are over-represented in the sample.
- The accuracy of the model is likely to reflect the underlying distribution imbalance.

Intuition

Possible solutions:

- **Under-sampling** the majority class.
- **Over-sampling** the minority class.
- **Under**-sampling the majority **and Over**-sampling the minority class.
- Use **cost weighted learning** - more weight given to the minority class.
- Use synthetic minority over-sampling technique (SMOTE) of Chawla et al., (2002, [2]).
- Appropriate **adjustment** of the decision threshold p_T (use cross-validation).
- Instance hardness threshold of Smith et al., (2014, [5]).
- Balance cascade of Liu et al., (2009, [4]).

[1] Glenn W Brier et al. "Verification of forecasts expressed in terms of probability". In: *Monthly weather review* 78.1 (1950), pp. 1–3.

[2] Nitesh V Chawla et al. "SMOTE: synthetic minority over-sampling technique". In: *Journal of artificial intelligence research* 16 (2002), pp. 321–357.

[3] Gareth James et al. *An introduction to statistical learning*. Springer, 2013.

[4] Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou. "Exploratory undersampling for class-imbalance learning". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 39.2 (2008), pp. 539–550.

[5] Michael R Smith, Tony Martinez, and Christophe Giraud-Carrier. "An instance level analysis of data complexity". In: *Machine learning* 95.2 (2014), pp. 225–256.

Artificial Intelligence in Finance

Supervised learning - discrete outcomes part A

Štefan Lyócsa

Department of Finance, Faculty of Economics and Administration

October 30, 2024

MASARYK
UNIVERSITY