Pa.02 – Lab Assignment

The resources needed for tasks 1 and 2 can be found in this Google Drive folder. You might want to copy this folder to your own Google Drive or to your local storage. You can run the Google Colab Notebooks through Google Drive (once you've copied them) without having to install anything. If you have Jupyter Notebook installed and prefer, you can also edit and run the Notebooks locally.

1. Decision problem for a bank

Resources for this task

- Google Colab (Pa.02_1-DecisionProblemForABank.ipynb)
- Dataset of loan applicants (loan_applicants.csv)

Banks give out loans to applicants if they believe that the applicant is capable of repaying the loan that they applied for. Imagine that a bank only knows the probability with which someone can repay the loan. We'll now see how high this probability has to be to warrant the bank giving an applicant a loan if we assume that the bank wants to maximize its profit.
We will assume that every applicant asks for a 1-year loan of size H. The bank offers all these applicants an interest rate of z. Specifically, they're given a 10% interest rate, i.e., z = 0.1. For the following, we will assume that the bank knows the exact repayment probability p for each applicant, i.e., the probability with which they repay their loan. In real cases, a bank would use an ML model to estimate this probability, and act on these estimation - but for this exercise, we assume that the bank has a perfect prediction model.
The bank faces a binary decision problem with D=1 (grant the loan) or D=0 (deny the loan). The decision-critical variable Y has the values Y=1 (applicant will pay back) and Y=0 (applicant will not pay back).

(a) Write the utility matrix of the bank, quantifying the four elements u_{ij} for the four combinations of Y and D as the profit of the bank for these cases.
- What's the profit of the bank if an applicant is accepted and repays the loan?
- What's the profit of the bank if an applicant is accepted but defaults on the loan?
- What's the profit of the bank if an applicant is NOT accepted but would have repaid the loan?

- What's the profit of the bank if an applicant is NOT accepted and would have NOT repaid the loan?

(b) For which repayment probability p would D=1 be the better decision? For each decision option d=0/d=1 calculate the expected profit (as a function of p) and decide which decision would be the one leading to the highest expected utility of the bank.
  - Does this rule depend on H?

(c) For the following, we assume for simplicity that H=1 for all customers. What is the rule that the bank has to apply if it wants to create the maximum expected profit from a population of applicants, where this population contains applicants with different values for p? Does this rule depend on specifics of the chosen population of applicants (e.g., how many applicants with a low probability are in the population), or is this a rule that holds for all possible populations?

(d) Assume that you are given the population listed in the resources. What is the expected **total** profit of the bank if they apply the rule from (c)?

(e) Assume that you are given the population listed in the resources. What is the expected profit of the bank **per positive decision** (given in percent) if they apply the rule from (c)?

(f) Assume now that the bank wants to increase its expected profit **per positive decision**. It should be at least 10% more than under the rule that maximizes the total profit. What rule does the bank then have to apply? Check all fixed threshold rules at 0.01 increments, so $p > p_0$ where $p_0 \in \{0, 0.01, 0.02, 0.03, \dots, 0.98, 0.99, 1\}$.

2. Decision problem for different subpopulations

Resources for this task

- [Google Colab (Pa.02_2-DecisionProblemForDifferentSubpopulations.ipynb)]()
- [Dataset of loan applicants (loan_applicants_gender.csv)]()

The bank also collects some personal data about their applicants, such as their gender. We will assume a binary gender collection. The bank now wants to test which effect different decision rules have on the loan approval rate of women and men.

(a) Assume a threshold rule of $p > 0.7$, so only those with a repayment probability of above 0.7 receive a loan. What share of women receives a loan? What share of men receives a loan?

(b) Assume a threshold rule of $p > 0.8$. What share of women receives a loan? What share of men receives a loan?

(c) In both (a) and (b), there is a difference of the acceptance rates of men vs. women. What is the cause of this difference in acceptance rate?

(d) Generalize your insight: Under which circumstances would the share of accepted applicants be identical for women and men if the bank applied the rule $p > p_0$ for an arbitrary p_0?

(e) Assume again a threshold rule of $p > 0.7$. With respect to fairness, one might argue that a difference in acceptance rates is not a problem, for example if one has reason to believe that women are more financially responsible than men and only apply if they have a high probability of repaying their loan. In such a case, a higher acceptance rate of women would be natural. One could then restrict the analysis to those individuals who actually repay their loan. Of course, this is not known at the time of decision making. However, if the true probability is known as we assume here, we still can do the evaluation. One could argue that the decision rule is fair if women who do repay their loan have the same acceptance rate as men who do pay back.
Use the provided data set and calculate: What share of the women who actually repay their loan is accepted? What share of men who would be able to repay their loan actually receives a loan? (Hint: for each individual, the repayment probability is given. For each individual i, you can define a random variable Y_i which is $Y_i = 1$ for a repaying individual, and $Y_i = 0$ else. The expectation value of Y_i is a number between 0 and 1. The expected number of repaying customers is the sum over all these individual expectation values.)

(f) Assume again a threshold rule of $p > 0.8$. What share of women who would be able to repay their loan actually receives a loan? What share of men who would be able to repay their loan actually receives a loan?

3. Video recommendations

Marc likes watching movies. He uses of a special video streaming platform which offers exactly one selected movie each night – and each night, he has to decide whether or not to watch this movie. He either enjoys a movie or hates it – it is never the case that he is neutral with respect to the movie.
Fortunately, the video channel comes with a state-of-the-art and fully personalized movie recommendation system. For each offered movie, this system provides the probability p ($0<p<1$) that Marc would enjoy a movie. Marc's experience with this system is quite good, and, based on the past, he trusts its recommendations.
Marc assigns the following utilities to watching a movie:

- The reference case is not watching the movie, with a utility value of 0.
- If he enjoys a movie, he assigns this a utility of +1.
- If he hates a movie, he is so annoyed that this is definitively worse than not having watched the movie at all. He assigns a value of -1 to this case.

Marc does not only like to watch movies – he also likes reading books. Here, he is not that picky – he enjoys pretty much every book, but not as much as a good movie. He assigns reading a book a utility of +0.3.
The two options for Marc each night are: D=0: read a book. D=1: watch the movie. The unknown, but decision-relevant variable Y is whether or not Marc actually enjoys the movie.
We assume that the sequence of movies provided by the channel has very different movies, and all p-values between 0 and 1 are being realized. More specifically: each value of p in the interval [0,1] appears with the same frequency. For simplification, we assume that only the values p=0,0.1, 0.2, …,0.9,1 appear. About 1/11 of all movies have p=0, 1/11 have p=0.1, etc.

(a) What is the utility matrix of Marc?

Utility matrix W	Y=0	Y=1
D=0		
D=1		

(b) What is the optimum decision rule for Marc? For deriving this, draw the expectation value E(U|D) for both decision options D=0 and D=1, as a function of p. For each p, find the decision that maximizes the utility.

(c) What is the average utility that Marc receives per night? (use the discrete set of probabilities given above)

(d) Now assume that the prediction model is biased, in that its communicated probabilities are not the true probabilities of Marc. Specifically, the model provides the following predictions $\hat{p}$:

True p	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
$\hat{p}$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.95	1.0

Marc still assumes that the given probabilities are correct and bases his decision on them. Show that this reduces his expected utility and calculate the reduced utility numerically.

4. Properties of predictors

Consider a population with 300 individuals. There are three different types of individuals A, B, and C, which differ by their (true) probability p=P[Y=1]:

Type	Number of individuals	p
A	100	0
B	100	0.5
C	100	1

Of course, we never have access to the “true probability”. All we have access to is data about the actual outcomes, so whether someone was able to repay their loan or not. Such data is then used for training a prediction model. The prediction model then assigns a probability to an individual based on the data which is available about the individual at the time of decision making. Here, we assume that we have data for each of the individuals i=1,…,300. In the following, we assume that this data has been used to train three different prediction models, and we study their properties.

(a) First, we study a point predictor P1 which predicts the value of Y for each individual. We assume that P1 yields the following output (the numbers denote the number of individuals for which the specific prediction has been made):

Type	Prediction: $\hat{Y} = 0$	Prediction: $\hat{Y} = 1$
A	80	20
B	50	50
C	30	70

What is the accuracy of this predictor?

(b) A probabilistic predictor would assign a score r to each individual. What would be the scores that a <u>perfect</u> probabilistic predictor P2 would assign to the individuals of the three types?

Type	r = ???	r = ???	r = ???
A			
B			
C			

Fill in the correct scores in the top row, and the correct numbers in the empty cells.

(c) Now consider the following probabilistic predictor P3:

Type	r = 0	r = 0.5	r = 1
A	40	60	0
B	0	100	0
C	0	60	40

One way of testing whether or not a prediction model gives good results is to take our predictions for many individuals and compare them with their actual outcomes to see if our predictions work well on average. For example, what do we expect of individuals with an assigned repayment probability of 60%? If our predictor is good, then we would expect 60% of these individuals to have the actual outcome Y=1. This is similar to weather forecasts: For all days with a 60% chance of rain, we expect it to rain on 60% of those days. If this is fulfilled, we speak of a calibrated score. If this is fulfilled for all scores of the predictor, then the predictor is **calibrated**.

Show that P3 is calibrated.

(d) Explain why P3 would lead to suboptimal decisions if a decision maker would interpret the score as the individual probability of each individual. Give a concrete use case (invent one).