Elements of ML and DS
WS 2023
Exercise Sheet 3

Exercise 3: Evaluation of Machine Learning Models

due **on** 2023-12-15

Important information regarding the exercises:

- Use the Moodle system to submit your solution. You will also find your corrections there.
- Upload your pen & paper solutions as PDF. If your solutions are handwritten, ensure that your handwriting is legible and that the pictures are not blurred and are taken under appropriate lighting conditions. All non-readable submissions will be discarded immediately.
- Submit coding exercises in a separate Python file.

Elements of ML and DS
WS 2023
Exercise Sheet 3

General Questions (2.5 Points)

Please answer the following multiple-choice questions by simply giving the letters of the correct answer. Multiple answers can be correct. Partial points will be deducted for incorrect answers, but each question will at worst count as 0 points.

1. Why do we need a train-test split? (**0.5 Points**)

   A: To detect overfitting after training.

   B: To prevent overfitting during training.

   C: Because more data is always better for training machine learning models.

   D: Evaluating both train and test data gives a better comparison.

2. Why do we sometimes need an additional validation dataset to do testing during model training? (**0.5 Points**)

   A: To detect overfitting after training.

   B: To prevent overfitting on the train data.

   C: To validate that the model achieves a low variance on the test data.

   D: To prevent overfitting on the test data.

3. Why is it sometimes necessary to consider profit matrices together with the confusion matrix? (**0.5 Points**)

   A: To evaluate the profit as the difference between the income of model prediction and cost to do inference with the model.

   B: We live in a capitalistic world and everything needs to be measured in money.

   C: The costs of wrong misclassifications can vary between false positives and false negatives.

   D: Profit matrices are essential for understanding how each prediction contributes to customer satisfaction.

4. The point on the ROC curve closest to the top-left is considered the optimal threshold. Optimizing towards this point is equivalent to optimizing . . . (**0.5 Points**)

   A: the F1-measure.

   B: the Accuracy.

   C: the Precision.

   D: none of the above.

5. The t-test is used for . . . (**0.5 Points**)

   A: comparing the variances of two groups.

   B: assessing the statistical significance of differences in the means of two groups.

   C: assessing the statistical significance of differences in the medians of two groups.

   D: comparing the standard deviations of two groups.

Problem 1 (Confusion Matrices, 3.5 Points)

Consider the table of flight delays below. For every ID you see the target label of the flight being on time or delayed and the corresponding prediction of a machine learning model.

a) Compute the corresponding confusion matrix. (**0.5 Points**)

b) Calculate the accuracy, misclassification, recall, precision and F1-measure corresponding to the confusion matrix. (**0.5 Points**)

c) Considering the metrics you just calculated, fill in the following blanks with one word each. (**2 Points**)

If the model predicts a positive class it is likely to be ______________. However, for classifying positive instances correctly, the model performs similarly to ______________ ______________. We can conclude the first insight from the ______________ and the second from the ______________.

Flight ID	Target	Predicted
1	Delayed	On Time
2	On Time	On Time
3	On Time	On Time
4	On Time	Delayed
5	Delayed	Delayed
6	Delayed	Delayed
7	On Time	Delayed
8	Delayed	Delayed
9	On Time	On Time
10	On Time	On Time
11	On Time	Delayed
12	On Time	Delayed
13	Delayed	Delayed
14	Delayed	Delayed
15	On Time	Delayed
16	On Time	On Time

Problem 2 (Profit Matrices, 1 Point)

You are given the following profit matrix

		Predicted	
		On Time	Delayed
Target	On Time	0	-80
	Delayed	-10	-20

to evaluate the cost of the machine learning model from Problem 1.

a) Compute the corresponding profit. **(0.5 Points)**

b) Would you recommend using this model in practice? Answer in no more than two sentences. **(0.5 Points)**

Problem 3 (Confusion Matrices with Mulitnomial Targets, 1 Point)

A new model has been trained to predict the targets on time, delayed and cancelled. Using some unseen test data the confusion matrix below was derived. Calculate the precision and recall for each label.

		Predicted		
		On Time	Delayed	Canceled
Target	On Time	40	9	1
	Delayed	5	20	3
	Canceled	3	8	8

Problem 4 (ROC Curves, 2.5 Points)

The data tuples of the table on the right are sorted by decreasing probability value, as returned by a classifier. The probability represents the threshold to be classified as a positive instance.

a) For each tuple, compute the values for the number of true positives, false positives, true negatives, and false negatives. Thereby, you always lower the threshold to the probability value of the corresponding tuple. In addition, compute the true positive rate and false positive rate.
(**2 Points**)

b) Use your calculations from part a) to plot the ROC curve for the data. (**0.5 Points**)

Flight ID	Target	Probability
1	Positive	0.9
2	Positive	0.85
3	Negative	0.7
4	Positive	0.65
5	Positive	0.55
6	Negative	0.53
7	Positive	0.48
8	Negative	0.43
9	Negative	0.40
10	Negative	0.38

Problem 5 (AUC, 2 Points)

Another model has been trained in addition to the model from Problem 4. An evaluation yields the following true positive and false positive rates for 11 different thresholds:

TPR	0.0	0.2	0.4	0.6	0.6	0.6	0.8	0.8	0.8	1.0	1.0
FPR	0.0	0.0	0.0	0.0	0.2	0.4	0.4	0.6	0.8	0.8	1.0

a) Plot the ROC curve in your graph from Problem 4. (**0.5 Points**)

b) Compute the AUC for both models. (**1 Point**)

b) How does the new model compare to the old model? Answer in one sentence. (**0.5 Points**)

Problem 6 (k-Fold Cross-validation, 3 Points)

In this problem, you are tasked with performing 3-fold cross-validation to compare the effectiveness of a logistic regression model and a decision tree. The dataset consists of information on Titanic passengers to predict whether a passenger survived or not. To complete this problem, you will need to have the Python packages ’pandas’ and ’scikit-learn’ installed.

a) Complete the provided Python file ’cross_valid.py’ by filling in the missing lines of code necessary to execute the cross-validation. This file and the dataset can be found in the zip file of the exercise. The dataset and Python file need to be in the same directory. Once completed, submit your Python file and state the outputs of your program in your submitted PDF solution. (**2.5 Points**)

b) What model performs better in your opinion? Answer in a single sentence. (**0.5 Points**)

Problem 7 (Assessing the Training Procedure, 7 Points)

You are using a randomised supervised ML procedure to train a predictive model. How to assess the training procedure? What could go wrong? Fill in each blank with a single word in the following text:

Due to the ________________ in training machine learning models, we must train multiple ________________ models to obtain statistically ________________ results. Otherwise, our evaluation is very sensitive to ________________ and training results with a high ________________. Once we have trained multiple models we need to analyse the ________________ of performance metrics across ________________ trainings and compare different ________________. We should not just pick the best model, since this skews the performance distribution to high-performing ________________, might ________________ on the test data and hinders ________________. Furthermore, we should not simply aggregate the performance of all models, since this might overlook the model's ________________. The model might perform well ________________ ________________ but have a high ________________.

Problem 8 (Evaluation after Deployment, 2.5 Points)

You have trained a predictive model using supervised machine learning, carefully assessed its performance and deployed it in practice. What could happen to invalidate earlier performance assessments? Fill in each blank with a single word in the following text:

When deploying machine learning models, especially over longer periods, it can happen that the underlying ________________ of the data we wish to model changes and thus differs from the data we ________________ our model on. This is a violation of the ________________ ________________ assumption. We often notice this, because of a performance ________________ of the model. This phenomenon is called ________________ ________________.