Chair of Decision Sciences and Systems
TUM School of Computation, Information and Technology
Technical University of Munich

Business Analytics & Machine Learning
Tutorial sheet 7: Model Evaluation and Selection

Prof. Dr. Martin Bichler
Julius Durmann, Markus Ewert, Yutong Chao, Dr. Mete Ahunbay

Exercise T7.1 *Metrics*

Assume your Business Analytics team has trained a binary classifier with the following results on unseen test data:

True Class	0	0	1	1	0	1	0	1	0	1
Predicted Class	0	1	1	0	0	0	0	1	1	0

a) State the confusion matrix for the ground truth and predicted labels given in the following table and interpret the results.

b) With these results, calculate recall, false alarm rate, precision, specificity and accuracy.

Exercise T7.2 *Imbalanced Data*

You are developing a binary classifier. During the test-phase of the model, you obtain the following confusion matrix (0 = negative, 1 = positive):

	Predicted 0	Predicted 1
Actual 0	TN = 85	FP = 5
Actual 1	FN = 8	TP = 2

a) What could be a problem with the data set?

b) Compute the accuracy of the classifier.

c) Compute the balanced accuracy (BAC) of the classifier:

$$BAC = \frac{\text{Sensitivity + Specificity}}{2} \tag{1}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{2}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{3}$$

d) What can you conclude?

Exercise T7.3 *Gain Curve, Lift Curve, ROC Curve*

Working for a car rental company, you have trained a classification model that predicts whether a car requires maintenance based on several input features. You have reserved some validation data, on which to you now apply the model and receive the following results:

	True Class	Probability of Classifier
Car 1001	-	0.24
Car 1002	-	0.02
Car 1003	+	0.82
Car 1004	-	0.67
Car 1005	+	0.94
Car 1006	-	0.78
Car 1007	-	0.37
Car 1008	+	0.63
Car 1009	+	0.17
Car 1010	+	0.73

Table 1 Your classifier results

Each row represents a car, which either requires maintenance (positive label) or not (negative label). The probabilities indicate the likelihood of being classified as requiring maintenance. The default cutoff value is 0.65.
Draw the

a) Gain Curve

b) Lift Curve

c) ROC Curve

of the classifier. Name the axes and write down the (x,y)-coordinates for every point. Mark the point corresponding to the default cutoff value. Would you argue that this cutoff value is appropriate? Explain your reasoning.

d) Your colleague has trained an alternative model on the entire dataset and evaluates the model on a subset thereof:

   Your colleague argues:
   "*This is the best possible classifier. Its gain curve is above the gain curve of a random classifier, and therefore it clearly dominates a random classifier. Moreover, the lift curve of my classifier will be a constant function at 2, whereas the lift curve of your classfier is monotonically decreasing. We should employ my classifier instead of yours.*"

   Explain three reasons why you disagree with your colleague's statement.

Exercise T7.4 *Churn prediction*

Working for a large e-commerce enterprise, you are given a dataset for customer churn prediction. The file `e-commerce-dataset.xlsx` contains a description of all variables in the sheet `Data Dict` as well as

	True Class	Probability of Classifier
Car 501	+	0.78
Car 502	+	0.81
Car 503	-	0.38
Car 504	-	0.27
Car 505	-	0.55
Car 506	-	0.01
Car 507	-	0.15
Car 508	+	0.62
Car 509	+	0.93
Car 510	-	0.49

Table 2 Your colleague's classifier results

the raw data itself in sheet `E_Comm`.

You intend to find a good random forest model to predict future customer churns.

Note: Use the py-script `churn.py`.

You begin with some data preparations. You remove all cases with missing data and factorize all variables where necessary.

a) Your colleague proposes to train the model on the entire dataset and argues to tune the `n_estimators` and `max_features` parameters of `sklearn.ensemble.RandomForestClassifier` until the training accuracy is maximized. Do you agree? If not, which issues can you identify with this approach?

Your colleague has little understanding of the issues that you raised. You decide to illustrate your ideas by means of the first twenty instances in your data $\{1, 2, \ldots, 20\}$.

b) You decide to split your dataset into training and test sets with an 80-20 % split and perform a 4-fold cross-validation. Using the small dataset, design an exemplary split of the data. Show how you would partition the data for the 4-fold cross validation. Explain the purposes of each subset and which operations/actions you perform with each subset.

You now turn back to the original dataset.

c) Perform training, 4-fold cross-validation, and testing with a 60-20-20 % split in Python. Use the precision as metric for model selection. Build a confusion matrix for the test set and report precision, accuracy, and recall.

d) On another dataset, you notice missing values for some numeric attributes in the test set. Your colleague suggests imputing these missing values by the mean of this attribute across all test instances. Do you agree? Explain your reasons.