Übungsblatt

Grundlagen der Künstlichen Intelligenz

17.11.2023, DHBW Lörrach

- Klassifikation mit logistischer Regression -

Der folgende Datensatz mit ca. 40.000 Aufzeichnungen *(engl. Records)* und 21 Spalten stammt aus einer direkten Marketing Kampagne mit Telefonanrufen und beinhaltet Informationen über Bankkunden:

	age	job	marital	education	default	housing	loan	contact	month	day_of_week	...	y
0	44	blue-collar	married	basic.4y	unknown	yes	no	cellular	aug	thu	...	0
1	53	technician	married	unknown	no	no	no	cellular	nov	fri	...	0
2	28	management	single	university.degree	no	yes	no	cellular	jun	thu	...	1

Beispiele aus dem Datensatz

Mithilfe logistischer Regression soll vorhergesagt werden, ob der Kunde das Angebot für eine Termineinlage annimmt (y=1) oder nicht (y=0).

1. Weshalb stellt die Imbalance zwischen den Klassifikationsergebnissen der gelabelten Daten (y=1: 4640 und y=0: 36.548) ein Problem für den Algorithmus dar?

Lösung
Das Modell würde für Testdaten fast ausschließlich die mehrheitlich vorhandene Klasse vorhersagen.

2. Wählen Sie diejenigen Prädiktorvariablen aus der linken Spalte der Übersicht aus, die Sie entfernen können, um die folgende Null-Hypothese zu verwerfen und formulieren Sie die entsprechende Arbeitshypothese.

   - Es besteht keine Beziehung zwischen den Prädiktorvariablen (z.B. job_blue-collar) und der abhängigen Variablen y (Termineinlage 1/0).

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
euribor3m	-0.4634	0.0091	-50.9471	0.0000	-0.4813	-0.4456
job_blue-collar	-0.1736	0.0283	-6.1230	0.0000	-0.2291	-0.1180
job_housemaid	-0.3260	0.0778	-4.1912	0.0000	-0.4784	-0.1735
marital_unknown	0.7454	0.2253	3.3082	0.0009	0.3038	1.1870
education_illiterate	1.3156	0.4373	3.0084	0.0026	0.4585	2.1727
default_no	16.1521	5414.0744	0.0030	0.9976	-10595.2387	10627.5429
default_unknown	15.8945	5414.0744	0.0029	0.9977	-10595.4963	10627.2853
contact_cellular	-13.9393	5414.0744	-0.0026	0.9979	-10625.3302	10597.4515
contact_telephone	-14.0065	5414.0744	-0.0026	0.9979	-10625.3973	10597.3843
month_apr	-0.8356	0.0913	-9.1490	0.0000	-1.0145	-0.6566
month_aug	-0.6882	0.0929	-7.4053	0.0000	-0.8703	-0.5061
month_dec	-0.4233	0.1655	-2.5579	0.0105	-0.7477	-0.0990

Auszug aus der Zusammenfassung der Trainingsergebnisse des Logit-Modells. Ein P-Wert kleiner als 0.05 gibt an, dass ein Zusammenhang (bzw. Korrelation) zwischen zwei Variablen besteht.

Lösung

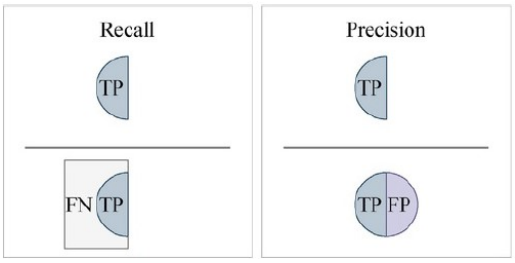
- default_no, default_unknown, contact_cellular, contact_telephone, da P-Werte größer 0.05

- Arbeitshypothese: Es besteht eine Beziehung zwischen den Prädiktorvariablen und der abhängigen Variablen y.

3. Würden Sie der Bank die Optimierung der Metrik *Präzision* oder *Trefferquote* empfehlen, um die Ressourcen für die Marketingkampagne möglichst sparsam einzusetzen? Begründen Sie Ihre Wahl.

Lösung

- Trefferquote (Recall) um falsche Negative zu vermeiden:



s. Vorlesung

- Die falschen Positiven (vgl. Präzision) machen prozentual einen kleineren Anteil der Beispiele aus, weshalb weniger Anrufe notwendig sind.