Hochschule Furtwangen
Fakultät Informatik
Prof. Dr. Lothar Piepmeyer

Praktikum ‚Informationssysteme'

Aufgabenblatt 1

Abgabe bis zum 22.10.2018

In der Datei `creditCards.csv` finden Sie Daten über Kreditkarten. Es gibt die folgenden Attribute
`income` - Das Einkommen des Karteninhabers
`balance` - Die aktuellen Kreditkartenschulden
`default` - Ein Prädikat, das anzeigt, ob der Karteninhaber in Zahlungsverzug gekommen ist
`isStudent` - Ein Prädikat, das anzeigt, ob der Karteninhaber ein Student ist.

Aufgabe 1:
Untersuchen Sie die Daten und versuchen Sie

- Die Verteilung der Daten zu verstehen
- Zusammenhänge zwischen den Spalten zu ermitteln.

In den folgenden Aufgaben schätzen Sie das Risiko eines Zahlungsverzugs ab, indem Sie eine Vorhersage für den Wert von `default` machen. Beachten Sie bitte: In der Vorlesung ging es darum, eine Größe (`salary`) vorherzusagen. Probleme dieser Art werden als *Regressionsprobleme* bezeichnet. Im vorliegenden Fall haben wir keine stetige Größe, sondern die beiden Klassen 'Yes' und 'No'. Probleme dieser Art werden als *Klassifikationsprobleme* bezeichnet.

Aufgabe 2:

a. Entwickeln Sie ein erstes ganz einfaches Modell für die Vorhersage von `default`.

b. Bei Regressionsproblemen arbeitet man mit dem RMSE als Fehlermaß. Überlegen Sie sich, wie Sie im vorliegenden Fall den Fehler messen können.

c. Ermitteln Sie mit der Metrik, die Sie im vorhergehenden Aufgabenteil gefunden haben, den Fehler für das Modell aus a.

Aufgabe 3:

a. Arbeiten Sie sich in das Verfahren ‚Logistische Regression' ein. Entwickeln Sie mit Hilfe der logistischen Regression ein Modell für den gesamten Kreditkartendatensatz und berechnen Sie den Fehler.

b. Teilen Sie die Daten in eine Trainings- und eine Testmenge. Entwickeln Sie ihr Modell aus a. nur für die Trainingsdaten. und wenden Sie dieses Modell dann auf die Testdaten an. Welcher Fehler ergibt sich? Vergleichen Sie diesen Fehler mit dem Fehler aus c.

c. Die logistische Regression liefert vielleicht nicht die besten Vorhersagen, ist aber interpretierbar. Interpretieren Sie das Modell, das Sie im vorhergehenden Modell entwickelt haben.

Aufgabe 4:

Führen Sie die Vorhersagen mit einem weiteren Verfahren Ihrer Wahl, wie etwa dem Extreme-Gradient-Boosting durch und vergleichen Sie die Ergebnisse mit denen aus vorhergehenden Aufgabenteilen.

Aufgabe 5:

Gibt es eigentlich einen Zusammenhang zwischen den Attributen default und isStudent? Erläutern Sie!