Prof. Dr. Jürgen Priemer: Projekte und Fallstudien

Semesterbegleitende Leistung, Teil: Data Mining

Sie dürfen diese Aufgabe in Gruppen mit max. 3 Studierenden bearbeiten.

Diese Aufgabe stammt aus dem Internet. Ausnahmsweise verschweige ich Ihnen vorläufig die Quelle. Aber auch wenn Sie die Quelle finden, nutzt Ihnen das reine Abschreiben der Ergebnisse nichts! Die Daten dürften übrigens aus dem Jahr 2005 stammen.

„Verglichen mit dem klassischen stationären Handel ist der Online-Handel ein Eldorado für kleine und große "Betrüger", deren Tätigkeitsfelder bis hin zum organisierten bandenmäßigen Betrug reichen. Die Regel "Ware gegen Geld" ist im Online-Handel, wie im ganzen Versandhandelsgeschäft, im direkten 1:1 Verhältnis nicht umsetzbar, mit Ausnahme des Nachnahmeversands. Die Nachnahme als einzige Zahlungsmethode anzubieten und damit ein recht betrugssicheres Online-Geschäft zu betreiben, ist jedoch für das ansonsten sehr flexible Online-Geschäft eher hinderlich.

Also stoßen wir in Online-Shops auf eine Vielzahl von Zahlungsmöglichkeiten - von Rechnungslegung (meist auf einen Betrag limitiert) über Lastschriftverfahren bis hin zu Kunden- oder Kreditkarten. Doch wie erkennt der Händler, ob es sich bei einer Bestellung um einen zahlungswilligen Kunden handelt, der letztendlich die Ware auch bezahlt?“

Als **Trainingsdaten** stehen 30.000 Datensätze zur Verfügung. Für sie ist das Zielattribut „TARGET_BETRUG“ bereits bekannt. Die Werte „ja“ und „nein“ sind möglich. Ist der Wert „ja“ gegeben, handelt es sich um eine betrügerische Bestellung, die nicht bezahlt wurde. Eine Beschreibung der Attribute finden Sie weiter hinten.

Für die **Klassifizierungsdaten** sollen Sie eine Vorhersage des Zielattributs vornehmen, d.h. es soll vorhergesagt werden, ob die Aufträge aus diesem Datentopf betrügerisch sind oder nicht.

Sie finden Trainingsdaten und Klassifizierungsdaten in Form von CSV-Dateien auf Trainex.

Teilaufgaben:

a) Analysieren Sie **vor** dem Durchführen des Data Minings die **Trainingsdaten** im Hinblick auf das Zielattribut „TARGET_BETRUG“. Fällt Ihnen etwas auf, was das Data Mining eventuell erschweren könnte.

b) Schauen Sie sich bei den unabhängigen Attributen die Spalten ANUMMER_01 bis ANUMMER_10 an. Was verbirgt sich hinter diesen Spalten und warum ist die in der Tabelle gewählte Darstellung ungünstig?
[Anmerkung: Sie können trotzdem mit dieser Darstellung weiter arbeiten. Es sei denn, Sie haben eine Idee, wie man die Daten umformen könnte und können diese auch praktisch umsetzen!]

c) Führen Sie das **Training mit drei unterschiedlichen und für die Problemstellung geeigneten Verfahren** durch, dabei können Sie das Werkzeug selbst auswählen. Beim Training behalten Sie bitte 20% der Trainingsdaten für den Test zurück. Stellen Sie die Testergebnisse der drei Verfahren jeweils als Konfusionsmatrix dar und berechnen Sie Precision, Recall und Accuracy.

Nehmen Sie Stellung zu den erzielten Ergebnissen, ggf. auch im Hinblick auf die bei Teilaufgabe a) gemachte Beobachtung. Was können Sie hier eventuell zur Verbesserung des Ergebnisses tun? [Wenn es für Sie technisch möglich ist, versuchen Sie die entsprechende Maßnahme durchzuführen!]

d) **Suchen Sie sich das Verfahren aus, das nach Ihrer Meinung am besten geeignet ist.** Nutzen Sie dieses Verfahren für die Vorhersage von TARGET_BETRUG für die neuen Bestellungen (Tabelle Klassifizierungsdaten). Wenn Ihr Werkzeug die Verbesserungsmaßnahme aus c) unterstützt, so nutzen Sie diese Verbesserungsmöglichkeit bitte.

e) **Beschreiben Sie in einer kurzen Ausarbeitung Ihre Vorgehensweise und geben Sie Antworten auf die in den Teilaufgaben a bis d gestellten Fragen. Erstellen Sie Screenshots zu den im Teil d) durchgeführten Prozess.** Senden Sie die erstellten Ergebnisdateien und -dokumente per Mail an mich (juergen.priemer@w-hs.de) oder erstellen Sie ein Github-Repository und laden Sie alles dort hoch.

   Ihr Ergebnis ist ebenfalls im Rahmen einer kurzen Präsentation am letzten Vorlesungstermin vorzustellen.

Beschreibung der Attribute

Datenfeld	Beschreibung
BESTELLIDENT	ID Nummer der Online-Bestellung
TARGET_BETRUG	Zielmerkmal der Aufgabenstellung (nur in der Trainingsdatei enthalten)
B_EMAIL	wurde bei der Bestellung eine E-Mail Adresse angegeben (ja/nein)
B_TELEFON	wurde bei der Bestellung eine Telefonnummer angegeben (ja/nein)
B_GEBDATUM	Geburtsdatum, wenn bei Bestellung angegeben
FLAG_LRIDENTISCH	wenn Liefer- und Rechnungsanschrift identisch dann "ja", sonst "nein"
FLAG_NEWSLETTER	wurde bei der Bestellung der E-Newsletter bestellt (ja/nein)
Z_METHODE	ausgewählte Zahlungsmethode
Z_CARD_ART	ausgewählter Kartentyp, wenn Zahlung per Karte erfolgen soll
Z_CARD_VALID	die Gültigkeitsinformation der Karte in Monat.Jahr (d.h. 05.2006, gültig bis Monat Mai 2006)
Z_LAST_NAME	ist der Name des Konto- oder Karteninhabers identisch mit dem Namen aus der Lieferadresse (ja/nein)
WERT_BEST	Bestellwert in Euro
TAG_BEST	Wochentag der Bestellung
TIME_BEST	Uhrzeit der Bestellung
ANZ_BEST	Anzahl der bestellten Artikel
ANUMMER_01	Datenfeld für Artikelnummer bestellter Artikel
ANUMMER_02	Datenfeld für Artikelnummer bestellter Artikel
ANUMMER_03	Datenfeld für Artikelnummer bestellter Artikel
ANUMMER_04	Datenfeld für Artikelnummer bestellter Artikel
ANUMMER_05	Datenfeld für Artikelnummer bestellter Artikel
ANUMMER_06	Datenfeld für Artikelnummer bestellter Artikel
ANUMMER_07	Datenfeld für Artikelnummer bestellter Artikel
ANUMMER_08	Datenfeld für Artikelnummer bestellter Artikel
ANUMMER_09	Datenfeld für Artikelnummer bestellter Artikel
ANUMMER_10	Datenfeld für Artikelnummer bestellter Artikel
CHK_LADR	Bestellungen mit gleicher Lieferanschrift im Zeitraum 3 Tage (ja/nein)
CHK_RADR	Bestellungen mit gleicher Rechnungsanschrift im Zeitraum 3 Tage (ja/nein)
CHK_KTO	Bestellungen mit gleicher Kontoverbindung im Zeitraum 3 Tage (ja/nein)
CHK_CARD	Bestellungen mit gleicher Kartennummer im Zeitraum 3 Tage (ja/nein)
CHK_COOKIE	Bestellungen mit gleichem Browser-Cookie im Zeitraum 3 Tage (ja/nein)
CHK_IP	Bestellungen mit gleicher Browser-IP im Zeitraum 3 Tage (ja/nein)
FAIL_LPLZ	Lieferadresse - PLZ nicht bekannt (ja/nein)
FAIL_LORT	Lieferadresse - Ort nicht bekannt (ja/nein)
FAIL_LPLZORTMATCH	Lieferadresse - PLZ und Ort passen nicht zusammen (ja/nein)
FAIL_RPLZ	Rechnungsadresse - PLZ nicht bekannt (ja/nein)
FAIL_RORT	Rechnungsadresse - Ort nicht bekannt (ja/nein)
FAIL_RPLZORTMATCH	Rechnungsadresse - PLZ und Ort passen nicht zusammen (ja/nein)
SESSION_TIME	Dauer der Online-Session zur Bestellung in Minuten
NEUKUNDE	ist der Besteller Neukunde (ja/nein)
ANZ_BEST_GES	Anzahl (Gesamt) der bestellten Artikel aus früheren Bestellungen, wenn vorhanden
WERT_BEST_GES	Bestellwert (Gesamt) aus früheren Bestellungen, wenn vorhanden
DATUM_LBEST	Datum der letzten Bestellung, wenn vorhanden
MAHN_AKT	aktuelle Mahnstufe des Kunden, wenn vorhanden
MAHN_HOECHST	höchste bislang aufgetretene Mahnstufe des Kunden, wenn vorhanden

HINWEIS zu TIME_BEST: Hier ist nur die Uhrzeit, nicht das Datum relevant!