Hochschule Fulda
University of Applied Sciences

Übung Data Mining – Wintersemester 2020/2021

Praxis-Challenge 2020: Vorhersage von Betrug bei Selbstbedienungskassen

Szenario

Verschiedene Kaufhäuser bieten alternativ zu den mit Personal besetzten Kassen Selbstbedienungskassen an. An diesen können die Kunden die gekauften Produkte selbst einscannen und so Zeit in der Warteschlange sparen. Das Kaufhaus spart sich so Personal an der Kasse. Ein wesentliches Problem dabei ist, dass Kunden an diesen Selbstbedienungskassen wissentlich oder unwissentliche falsche Eingaben machen können, z.B. ein einzelnes Produkt nicht einscannen, um Geld zu sparen. Um den Betrug zu kontrollieren, werden von vielen Kaufhäusern einzelne Einkäufe in Stichproben kontrolliert.

Daraus ergibt sich das folgende Problem: Welche Kunden sollen kontrolliert werden, welche nicht? Wenn alle Kunden kontrolliert werden, kann Betrug verhindert werden, aber der Aufwand ist dann höher als bei herkömmlichen Kassen. Auch wenn nicht alle Kunden kontrolliert werden, sind Kontrollen problematisch, da einzelne Kunden sich durch Nachkontrollen gestört fühlen könnten und zukünftig von dem Kaufhaus fernbleiben könnten. Eine qualitativ gute Identifikation möglicher Betrugsfälle ist also wesentlich.

Aufgabe

Ziel der diesjährigen Challenge ist es, auf Grundlage eines Trainings-Datensatzes ein Klassifikationsmodell zu erstellen, welches eine möglichst hohe Vorhersagequalität besitzt. Auf Basis dieses Klassifikationsmodells soll eine Vorhersage über den Betrug (1) bzw. ausbleibenden Betrug (0) von neuen Kunden erstellt werden. Für diese neuen Kunden fehlt die Klassifikationsvariable im classify-Datensatz.

Beachten Sie, dass ein aufgedeckter Betrug nach Abzug der Kontrollkosten im Schnitt 5 Euro einbringt, während ein nicht-aufgedeckter Betrug im Schnitt einen Verlust von 5 Euro kostet. Da Kunden, die fälschlicherweise kontrolliert werden, teilweise zukünftig nicht mehr einkaufen, ist dieser Fall mit durchschnittlich 25 Euro besonders teuer. Beachten Sie also die folgende Kostenmatrix:

Tabelle 1 Kostenmatrix

	Vorhersage		
Tatsächlicher Wert		Kein Betrug (0)	Betrug (1)
	Kein Betrug (0)	0 Euro	-25 Euro
	Betrug (1)	-5 Euro	5 Euro

Daten

Die Datei challenge2020_train.csv beinhaltet Daten von 1879 Einkäufen und dient als Train- als auch als Test-Datensatz. Ein zu klassifizierender Datensatz challenge2020_classify.csv ohne Zielvariable wird später zur Verfügung gestellt und enthält die gleichen Spalten mit Ausnahme der Zielvariablen. Es folgt eine Tabelle mit den enthaltenen Variablen:

Tabelle 2 Variablen im Datensatz

Spalte	Beschreibung
trustLevel	Vertrauenswürdigkeit des Kunden (6 ist maximal)
totalScanTimeInSeconds	Sekunden zwischen dem Scan des ersten und des letzten Produkts
grandTotal	Gesamtsumme des Einkaufs
lineItemsVoids	Anzahl leerer Scans
scansWithoutRegistration	Anzahl der Aktivierungen des Scanners ohne Scan
quantityModification	Anzahl einer Veränderung der Stückzahl
scannedLineItemsPerSecond	Durchschnittliche Anzahl von gescannten Produkten pro Sekunde
valuePerSecond	Durchschnittlicher Wert der gescannten Produkte pro Sekunde
lineItemsVoidsPerPosition	Durchschnittliche Anzahl leerer Scans durch Gesamtanzahl aller gescannten (und nicht stornierten) Produkte
fraud	Klassifikation als Betrug (1) oder keinen Betrug (0)

Abgabetermin

Das erstellte Klassifikationsmodell in Form des IBM SPSS Modeler Streams ist zusammen mit dem Klassifikationsergebnis (zum Klassifikationsergebnis folgt eine Beschreibung) bis zum **Freitag, 11. Dezember 2020, 23:55 Uhr** abzugeben.

Q: Wird am Ende der Challenge bekanntgegeben.