Abschlussprojekt: Gebrauchtwagenkäufe

Modul 3 | Kapitel 4 | Notebook 1

Du hast in den letzten Kapiteln und Modulen zahlreiche Data-Science-Methoden kennengelernt. Nachdem du im Übungsprojekt eine Regression auf Daten mit relativ wenigen Features durchgeführt hast, kannst du dich nun im Abschlussprojekt der Klassifizierung widmen. Dazu erhältst du ein Datenset, welches Gebrauchtwagenkäufe und unausgeglichene Zielkategorien beinhaltet.

Am Ende der Übung hast du:

- Daten eingelesen und bereinigt,
- Feature Engineering betrieben,
- ein Modell an die Daten angepasst,
- die einzelnen Schritte zu einer Datenpipeline kombiniert,
- dein Modell interpretiert.

In dieser Übung hast du noch wenige Hinweise und sollst so selbstständig wie möglich arbeiten.

Am Ende dieses Projekts wirst du Klassifizierungen für Datenpunkte erzeugen, deren Zielwerte nur wir kennen. Sie befinden sich in der Datei *features_aim.csv*. Du wirst also die meisten Schritte wiederholen müssen. Am einfachsten kannst du das umsetzen, indem du alle relevanten Arbeitsabschnitte in Funktionen oder in eine Pipeline verpackst. Diese Funktionen rufst du am Ende in der richtigen Reihenfolge auf und wendest sie auf das Zieldatenset *features_aim.csv* an.

Szenario: Du arbeitest als Data Scientist für einen US-amerikanischen Gebrauchtwagenhändler. Der Händler kauft gebrauchte Autos günstig in Onlineauktionen und von anderen Autoverkäufern, um sie dann auf der eigenen Plattform gewinnbringend weiterzuverkaufen. Es ist nicht immer einfach zu erkennen, ob sich der Kauf eines Gebrauchtwagens lohnt: Eine der größten Herausforderungen bei Gebrauchtwagenauktionen ist das Risiko, dass ein Auto solch schwerwiegende Probleme hat, die verhindern, dass es an Kunden weiterverkauft werden kann. Hierbei handelt es sich um sogenannte "Montagsautos" - also Autos, die von Hause aus erhebliche Mängel aufgrund von Produktionsfehlern aufweisen, welche die Sicherheit, die Verwendung oder den Wert dieses Autos erheblich beeinträchtigen und gleichzeitig nicht in einer angemessenen Anzahl an Reparaturen oder innerhalb eines bestimmten Zeitraums behoben werden können. Der Kunde hat in diesem Fall das Recht, sich den Kaufpreis zurückerstatten zu lassen. Neben den Anschaffungskosten führt der Fehleinkauf von solchen "Montagsautos" also zu erheblichen Folgekosten, wie z.B. der Einlagerung und Reparatur des Wagens, welche in Verlusten beim Weiterverkauf des Fahrzeugs resultieren können.

Deshalb ist es deiner Chefin wichtig, möglichst viele Fehlkäufe von "Montagsautos" auszuschließen. Um die Einkäufer im Unternehmen bei der riesigen Anzahl an Angeboten zu entlasten, sollst du ein Modell entwickeln, welches vorhersagt, ob ein Angebot ein Fehlkauf im Sinne eines Montagsautos wäre. Das darf allerdings nicht dazu führen, dass zu viele gute Käufe ausgeschlossen werden. Genauere Angaben zu den Kosten und Gewinnen der jeweiligen Käufe erhältst du für die Entwicklung des Prototyps noch nicht.

Jede Zeile des Datensets entspricht einem Auto, das zunächst ersteigert und anschließend weiterverkauft wurde. Das Datenwörterbuch sieht wie folgt aus:

Spaltennummer	Spaltenname	Datenniveau	Beschreibung
1	`'IsBadBuy'`	kategorisch (nominal)	Identifiziert, ob das ersteigerte Fahrzeug ein "Montagsauto" und somit ein Fehlkauf war (`0` = kein Montagsauto, `1` = Montagsauto)
2	`'PurchDate'`	kontinuierlich (`datetime`)	Das Datum, an dem das Fahrzeug bei der Auktion gekauft wurde
3	`'Auction'`	kategorisch (nominal)	Auktionsanbieter, bei dem das Fahrzeug gekauft wurde
4	`'VehYear'`	kontinuierlich (`int`)	Baujahr des Fahrzeugs
5	`'VehicleAge'`	kontinuierlich (`int`)	Alter des Autos zum Zeitpunkt der Auktion
6	`'Make'`	kategorisch (nominal)	Fahrzeughersteller
7	`'Model'`	kategorisch (nominal)	Fahrzeugmodell
8	`'Trim'`	kategorisch (nominal)	Fahrzeugausstattung
9	`'SubModel'`	kategorisch (nominal)	Fahrzeug-Submodell
10	`'Color'`	kategorisch (nominal)	Fahrzeugfarbe
11	`'Transmission'`	kategorisch (nominal)	Fahrzeug-Getriebeart (Automatik, Manuell)
12	`'WheelTypeID'`	kategorisch (nominal)	Die Typ-ID der Felgen
13	`'WheelType'`	kategorisch (nominal)	Art der Felgen
14	`'VehOdo'`	kontinuierlich (`int`)	Meilenstand des Fahrzeugs
15	`'Nationality'`	kategorisch (nominal)	Das Land des Herstellers
16	`'Size'`	kategorisch (nominal)	Die Größenklasse des Fahrzeugs (Kompakt,

Spaltennummer	Spaltenname	Datenniveau	Beschreibung
			SUV, etc.)
17	'TopThreeAmericanName'	kategorisch (nominal)	Identifiziert, ob der Hersteller einer der drei führenden amerikanischen Hersteller ist.
18	'MMRAcquisitionAuctionAveragePrice'	kontinuierlich (int)	Anschaffungspreis in US-Dollar für dieses Fahrzeug im durchschnittlichen Zustand zum Zeitpunkt des Kaufs
19	'MMRAcquisitionAuctionCleanPrice'	kontinuierlich (int)	Anschaffungspreis in US-Dollar für dieses Fahrzeug im überdurchschnittlichen Zustand zum Zeitpunkt des Kaufs
20	'MMRAcquisitionRetailAveragePrice'	kontinuierlich (int)	Anschaffungspreis in US-Dollar für dieses Fahrzeug im Einzelhandel im durchschnittlichen Zustand zum Zeitpunkt des Kaufs
21	'MMRAcquisitonRetailCleanPrice'	kontinuierlich (int)	Anschaffungspreis in US-Dollar für dieses Fahrzeug im Einzelhandel in überdurchschnittlichem Zustand zum Zeitpunkt des Kaufs
22	'MMRCurrentAuctionAveragePrice'	kontinuierlich (int)	Anschaffungspreis in US-Dollar für dieses Fahrzeug im Durchschnittszustand zum aktuellen Tag
23	'MMRCurrentAuctionCleanPrice'	kontinuierlich (int)	Anschaffungspreis in US-Dollar für dieses Fahrzeug im überdurchschnittlichen Zustand zum aktuellen Tag
24	'MMRCurrentRetailAveragePrice'	kontinuierlich (int)	Anschaffungspreis in US-Dollar für dieses Fahrzeug im Einzelhandel im durchschnittlichen Zustand zum aktuellen Tag
25	'MMRCurrentRetailCleanPrice'	kontinuierlich (int)	Anschaffungspreis in US-Dollar für dieses Fahrzeug im Einzelhandel im überdurchschnittlichen

Spaltennummer	Spaltenname	Datenniveau	Beschreibung
			Zustand zum aktuellen Tag
26	`'PRIMEUNIT'`	kategorisch (nominal)	Identifiziert, ob das Fahrzeug eine höhere Nachfrage als ein Standardkauf haben würde
27	`'AUCGUART'`	kategorisch (nominal)	Das Garantielevel, der Versteigerungsplattform, das für das Fahrzeug gegeben wird (`'GREEN'` - Garantie vorhanden, `'YELLOW'` - Garantie unklar, `'RED'` - keine Garantie)
28	`'BYRNO'`	kategorisch (nominal)	Eindeutige Nummer, die dem Käufer zugewiesen wird, der das Fahrzeug gekauft hat
29	`'VNZIP1'`	kategorisch (nominal)	Postleitzahl, wo das Auto gekauft wurde
30	`'VNST'`	kategorisch (nominal)	Staat, in dem das Auto gekauft wurde
31	`'VehBCost'`	kontinuierlich (`int`)	Anschaffungskosten in US-Dollar, die für das Fahrzeug zum Zeitpunkt des Kaufs bezahlt wurden
32	`'IsOnlineSale'`	kategorisch (nominal)	Identifiziert, ob das Fahrzeug ursprünglich online gekauft wurde
33	`'WarrantyCost'`	kontinuierlich (`int`)	Kosten der Garantie für eine Laufzeit von 36 Monaten

Tipp: Das Datenwörterbuch befindet sich auch in der Datei *Datenwoerterbuch.ipynb*. Diese kannst du mit deinem *file browser* in einem eigenen Fenster öffnen.

Die folgenden Überschriften und Texte sollen dir eine grobe Richtlinie geben. Du kannst jederzeit weitere Codezellen einfügen oder dein Vorgehen ändern, wenn du das für richtig hältst.

Wir folgen hier wieder den Schritten des *Stackfuel-Way*. Du findest die PDF hierzu in *Der Data Science Workflow* (Modul 3, Kapitel 3).

Preparation

Ein Data-Science-Modell hat immer die Aufgabe, ein bestimmtes Problem zu lösen. Also befasse dich am besten kurz nochmal mit der Aufgabenstellung und überlege dir den

Kontext deines Modells.

- Welches Problem soll das Modell lösen?
- Welcher Art ist das Problem (z.B. Klassifikation, Regression, Clustering ...)
- Wie sähe eine Anwendung aus, die dein Modell benutzt?
- Welche Anforderungen stellt der Auftraggeber an dein Modell?
- Welche Daten benötigst du, um dein Modell zu bauen?

Define Metric

Anhand deines Verständnisses des vorliegenden Problems solltet du dir nun überlegen, welche Metrik(en) am besten geeignet sind, den Erfolg deines Modells zu beurteilen.

Gather Data

Die Daten befinden sich in der Datei *data_train.csv*. Der Zielvektor ist durch die Spalte `'IsBadBuy'` gegeben. Importiere die Module, die du typischerweise für das Einlesen und die Exploration benötigst und lies die Daten anschließend ein.

In [ ]:

In [1]:
```
# import modules
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt

pd.set_option('display.max_rows', None)
pd.set_option('display.max_columns', None)
```

In [2]:
```
# read data
df = pd.read_csv('data_train.csv')
df.head()
```

Out[2]:

	IsBadBuy	PurchDate	Auction	VehYear	VehicleAge	Make	Model	Trim	SubMod
0	0	1257897600	OTHER	2007	2	KIA	SPECTRA	NaN	4D SEDA E
1	0	1231286400	ADESA	2005	4	SUZUKI	FORENZA 2.0L I4 EFI	EX	4D WAGO
2	1	1288656000	OTHER	2006	4	CHEVROLET	COBALT	LT	2D COUF I
3	0	1236124800	MANHEIM	2004	5	CHEVROLET	VENTURE FWD V6 3.4L	LS	PASSENGE 3.4
4	0	1248307200	MANHEIM	2007	2	CHRYSLER	TOWN & COUNTRY 2WD V	Bas	MINIVA 3.3

EDA