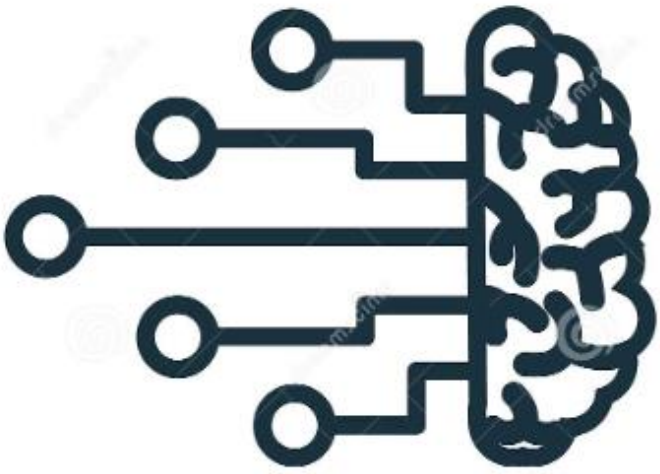

Used Cars Prediction

SELEKCIJSKI ZADATAK - CROZ

Martina Radenić | 05.05.2023.

Analiza, obrada i priprema podatak

IZBACIVANJE PODATAKA

U prvom koraku vidimo da postoje podaci koji nam isu previše važni za model, odnosno smatramo da ne utječu na cijenu automobila. Takve podatke izbacujemo, a oni su: 'id', 'url', 'region_url', 'VIN', 'image_url', 'description', 'lat', 'long', 'county', 'region'.

NUMERIČKI PODACI

Drugi korak se bavi proučavanjem numeričkih podataka, odnosno cijene, kilometraže i godine automobila.

```
df.describe()
```

	price	year	odometer
count	4.268800e+05	425675.000000	4.224800e+05
mean	7.519903e+04	2011.235191	9.804333e+04
std	1.218228e+07	9.452120	2.138815e+05
min	0.000000e+00	1900.000000	0.000000e+00
25%	5.900000e+03	2008.000000	3.770400e+04
50%	1.395000e+04	2013.000000	8.554800e+04
75%	2.648575e+04	2017.000000	1.335425e+05
max	3.736929e+09	2022.000000	1.000000e+07

Cijena automobila

Budući je razlika između 75% i max vrijednosti dosta velika za outliere uzimamo vanjskih 10%. Ovo činimo jer ne želimo izgubiti previše podataka.

Kilometraža

Za početak vidimo 0.28% podataka ima vrijednost kilometraže 0. To ne želimo budući da tražimo rabljene automobile. Te podatke izbacujemo. Isto tako, iz grafičkog prikaza zaključujemo da su sve vrijednosti iznad 4000000 outlieri. Te podatke također izbacujemo.

Godina

Kod ovog podatka, null vrijednosti čine samo 0.17% pa ih možemo sve izbaciti iz skupa.

NULL VRIJEDNOSTI

U ovom dijelu promatramo sve kategorije kojima nedostaju vrijednosti.

	null	percent
size	23720500	70.091
cylinders	13607000	40.207
condition	12318300	36.399
drive	10423400	30.800
paint_color	9813500	28.998
type	7470500	22.074
manufacturer	1225400	3.621
title_status	569800	1.684
model	376300	1.112
fuel	193200	0.571
transmission	128800	0.381
price	0	0.000
year	0	0.000
odometer	0	0.000
state	0	0.000
posting_date	0	0.000

Stanje automobila

Možemo zaključiti da je stanje automobila povezano sa prijeđenom kilometražom. Što se više auto vozio, to je u lošijem stanju. Stoga vrijednosti stanja koje nam nedostaju možemo nadopuniti koristeći podatke o kilometraži. Pronađemo mean kilometraže po stanjima i pomoću tih vrijednosti nadopunimo stanja koja nam nedostaju.

Titel_status, fuel, transmission, model, manufacturer

U ovim kategorijama postotak null vrijednosti je manji od 5% pa te vrijednosti izbacujemo.

Size

U kategoriji size postotak null vrijednosti je čak 70% pa tu kategoriju izbacujemo u potpunosti jer nedostaje previše podataka.

Broj cilindara

Broj cilindara utječe na cijenu automobila, stoga za početak izračunamo medijan cijena null vrijednosti. Vrijednost usporedimo sa medijanima cijena po cilindrima i vidimo da je najbliže kategoriji '8 cilindara'. Stoga vrijednosti koje nedostaju nadopunimo sa 8 cilindara.

Pogon

Za pogon ponovimo postupak kao i za broj cilindara. U ovom slučaju zaključujemo da nedostajuće vrijednosti trebamo nadopuniti sa fwd pogonom.

Tip vozila

Ponovno ponovimo postupak. Medijan je sada najbliži vrijednosti 'sedan' pa nadopunjujemo null vrijednosti tom informacijom.

Boja automobila

Koristimo forward fill za nadopunjavanje boje kako ne bismo izgubili taj podatak a smatramo da nam ipak ima utjecaja na cijenu.

Datum objave

Datum objave pretvaramo u broj dana koji je protekao od objave do trenutnog datuma. Na ovaj način dobivamo numeričku vrijednost.

Odabir, treniranje i evaluacija modela

U prvom koraku na svim ne-numeričkim podacima koristimo LabelEncoder kako bi dobili numeričke. Pri izradi train i test seta koristimo test_size = 0.2, a za random_state vrijednost 42.

Za model uzimamo Random Forest model. Model je postojan protiv outlierea, dobro radi sa nelinearnim podacima te dobro funkcionira i sa velikom količinom podataka. Preciznost modela je 89.92%.

Na samom kraju model spremamo kao pickle za korištenje u implementaciji API-a.