Предвиђање цена половних аутомобила применом алгоритама машинског учења

Ментор: Др Сандро Радовановић

Алекса Радојичић, 2019/0165

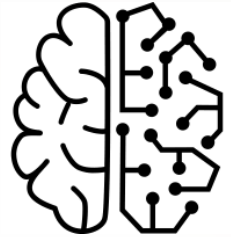
Садржај

1

Увод

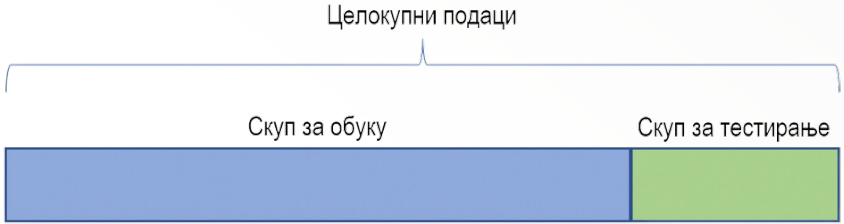
- 77.973 путничких возила на платформи *polovniautomobili*
- Прецењеност и потцењеност кола и 90%-ни интервали предвиђања

2 Теоријске основе



Машинско учење [МУ]

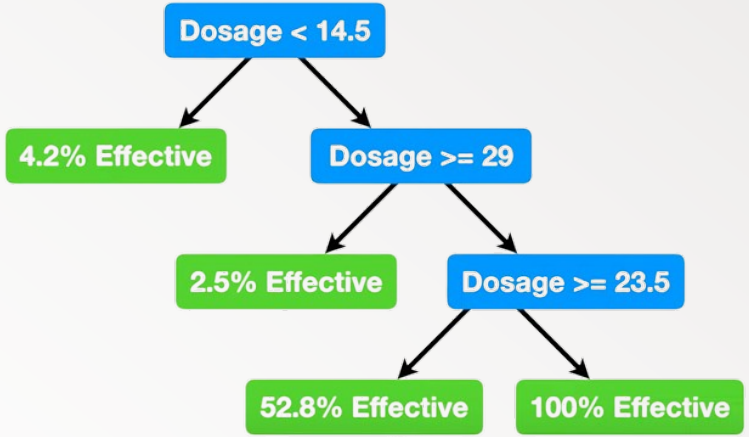
Направити математички модел који предвиђа излазне вредности на основу улазних вредности. Користити познате податке за пресказивање сродних и непознатих података.



Потприлагодба спрам **преприлагодбе**

2.1 Дрво одлучивања [DT]

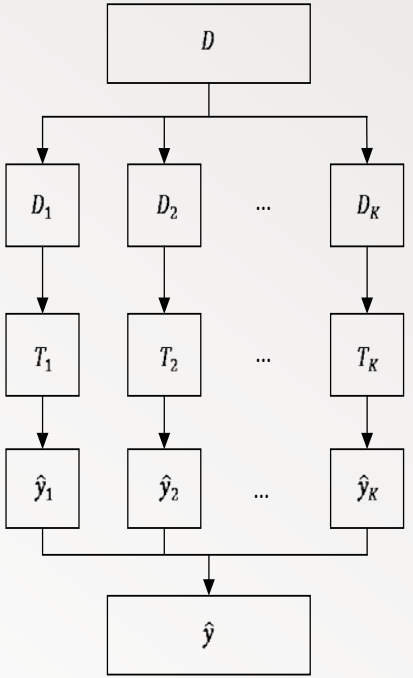
- Састоји се од чворова и грана. Чвор који се не грана је лист.
- Врши се гранање тако да се највише повећа информациона добит.
- Нелинеаран алгоритам.
- Тумачљив и временски и меморијски исплатив.



Преузета слика са StatQuest снимка

2.2 Случајна шума [RF]

- Начинити више подскупа са различитим редовима и / или ступцима (**багинг**) и над сваким обучити дрва одлучивања.
- Боље перформансе и честа употреба у пракси.
- Мање тумачљив алгоритам и меморијски скуп.
- Подржава паралелизацију.



2.3 Модел градијентног појачавања [GBM]

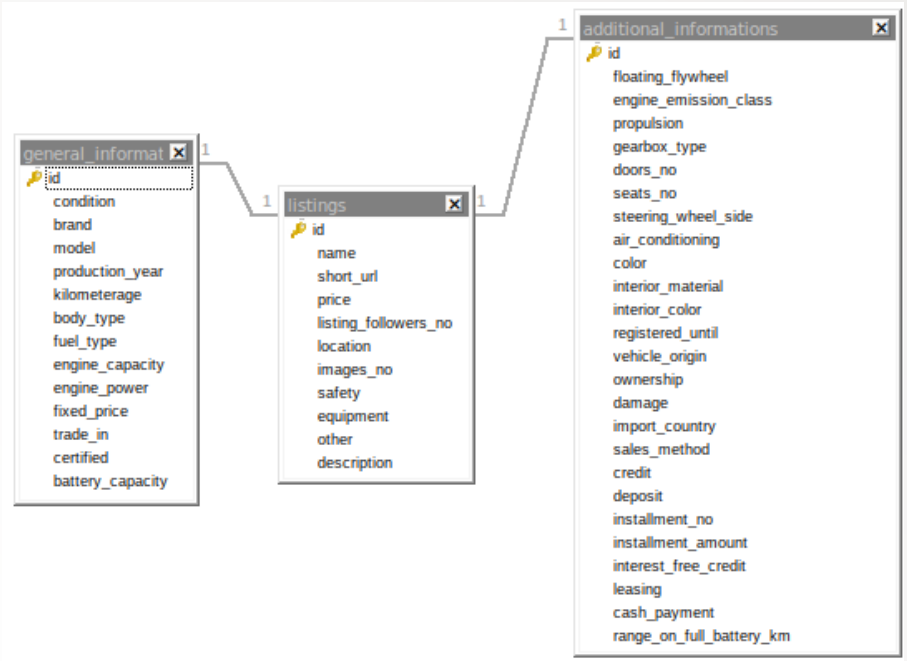
- Редно (итеративно) се творе дрва одлучивања тако да свако идуће буде боље од претходног (**појачавање**).
- Максимални саоднос (корелација) са негативним градијентом функције губитка.
- Омогућавају творење интервала предвиђаја.
- Не подржава паралелизацију.

3

Опис истраживања

1) Прикуп података
2) Увид у податке
3) Предобрада података
4) Обука модела
5) Оптимизација модела

3.1 Прикуп података



- Прикупљено 30.788 редова и 50 стубаца (колона).

BMW 116 i M paket 109 PS 2015. godište

13.500 €
+ dodatni troškovi kupovine

46 PRIJAVI OGLAS

Izračunaj ratu keš kredita

Svi oglasi oglašivača

Valjevo
ul.Jakova Nenadovica
55

KLIK DA VIDIŠ BROJ

POŠALJI PORUKU

Opšte informacije

Stanje:	Polovno vozilo	Kubikaža	1499 cm³
Marka	BMW	Snaga motora	80/109 (kW/KS)
Model	116	Fiksna cena	NE
Godište	2015.	Zamena:	NE
Kilometraža	109.000 km	Broj oglasa:	22855780
Karoserija	Hečbek	Broj šasije:	WBA2R11040V558470
Gorivo	Benzin		

Dodatne informacije

Emisiona klasa motora	Euro 6	Boja	Bela
Pogon	Zadnji	Materijal enterijera	Štof
Menjač	Manuelni 6 brzina	Boja enterijera	Crna
Broj vrata	4/5 vrata	Registrovan do	04.2025.
Broj sedišta	5 sedišta	Poreklo vozila	Domaće tablice
Strana volana	Levi volan	Vlasništvo	Vodi se na prodavca
Klima	Automatska klima	Oštećenje	Nije oštećen

Sigurnost

Airbag za vozača	Airbag za suvozača	Bočni airbag	Child lock
ABS	ESP	ASR	Kodiran ključ
Centralno zaključavanje	Mehanička zaštita		

Oprema

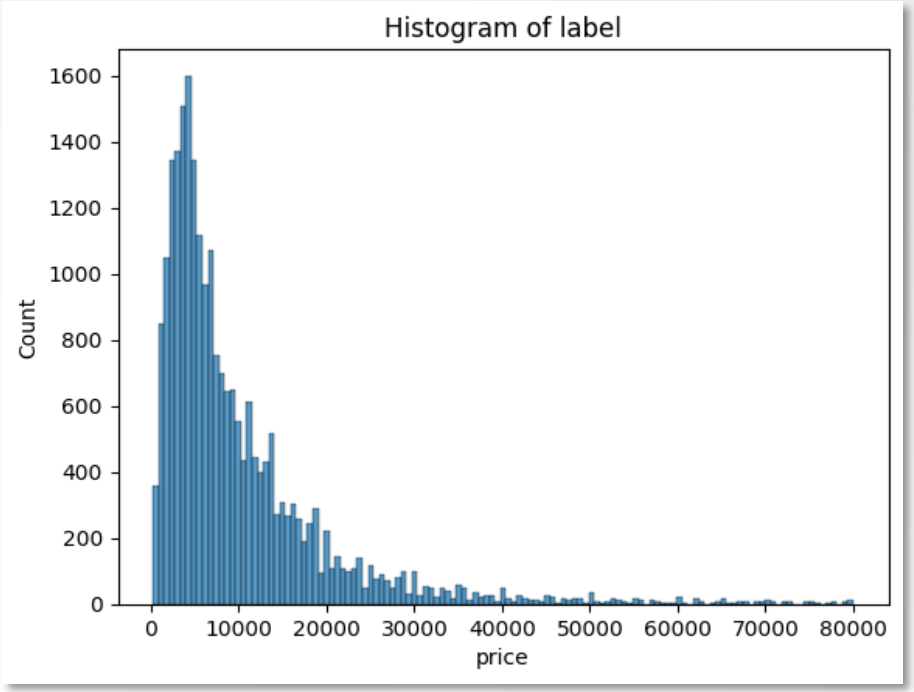
Branici u boji auta	Servo volan	Multifunkcionalni volan	Tempomat
Daljinsko zaključavanje	Putni računar	Tonirana stakla	Električni podizači
Električni retrovizori	Grejači retrovizora	Sedišta podesiva po visini	Grejanje sedišta

Saveti

3.2 Увид у податке

	фрекв.	удео [%]
„марка“		
Volkswagen	3734	15,716150
Audi	2529	10,644387
BMW	2151	9,053411
Opel	1862	7,837030
Peugeot	1582	6,658529

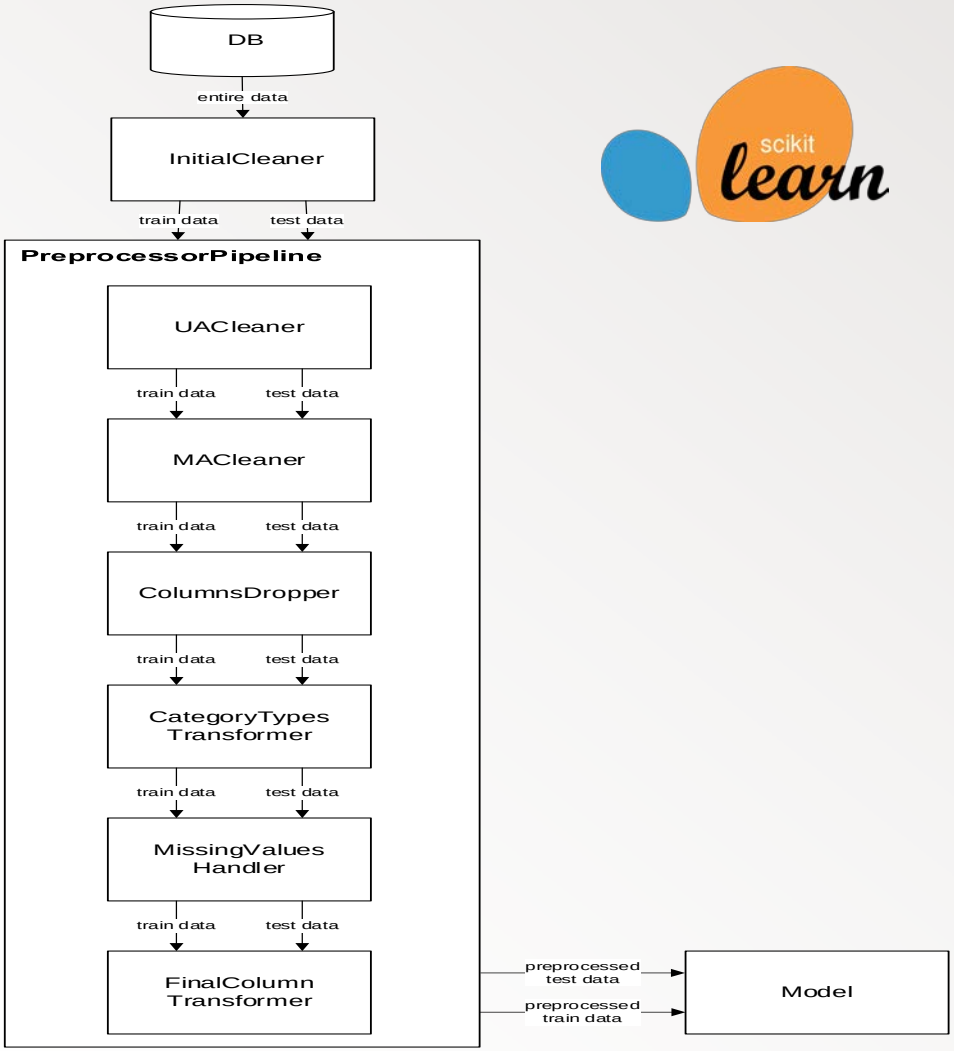
	фрекв.	удео [%]
„локација“		
Београд	3828	16,111789
Нови Сад	1709	7,193064
Ниш	954	4,015321
Крагујевац	883	3,716486
Чачак	730	3,072520



	фрекв.	mean	std	min	25%	50%	75%	max
„година производње“	23759	2010,59472	6,25557	1953	2007	2011	2015	2024

3.3 Предобрада података

- Димензије података после уводног чишћења: 29.699, 164.
- Цевовод за предобраду:
  уклања слабо употребљиве колоне, обрађује типове података, начин попуне недостајућих вредности...
- Обрађује засебно скупове за обуку и тестирање.



3.4

Обука модела – почетни модели –

- 80% кола (23.759) за обуку, 20% (5.940) за тестирање.
- Коришћено свих 153 стубаца.
- Случајна шума се најбоље показала над скупом за тестирање.

	Метрике над скупом за тестирање		
Модел	RMSE	MAE	R2
DT	4198,384	2050,968	0,84021
RF	**2786,725**	**1364,816**	**0,92960**
GBM $Q_{0,50}$	2921,088	1656,298	0,92265
GBM $Q_{0,05}$	6795,749	3460,504	0,58134
GBM $Q_{0,95}$	5720,118	3635,153	0,70339

	Метрике над скупом за обуку		
Модел	RMSE	MAE	R2
RF	**1045,089**	**510,826**	**0,99013**
GBM $Q_{0,50}$	2727,873	1583,053	0,93277
GBM $Q_{0,05}$	5593,659	3502,329	0,57025
GBM $Q_{0,95}$	6896,814	3615,054	0,71731

3.5 Оптимизација модела

- Употребљено 28 стубаца.
- Цена логаритмована.
- Решетказта претрага за добијање оптималних хиперпараметара (80% обука, 20% вредновање).

Оптимални хиперпараметри за RF

Хиперпараметри	Вредности
n_estimators	**245**
min_samples_split	**2**

Оптимални хиперпараметри за GBM $Q_{0,50}$

Хиперпараметри	Вредности
learning_rate	**0,05**
n_estimators	**500**
min_samples_split	**15**
max_depth	**12**
criterion	**„friedman_mse“**

3.5 Оптимизација модела – интервали предвиђаја –

Оптимални хиперпараметри за GBM $Q_{0,05}$ (долњи интервал предвиђаја)

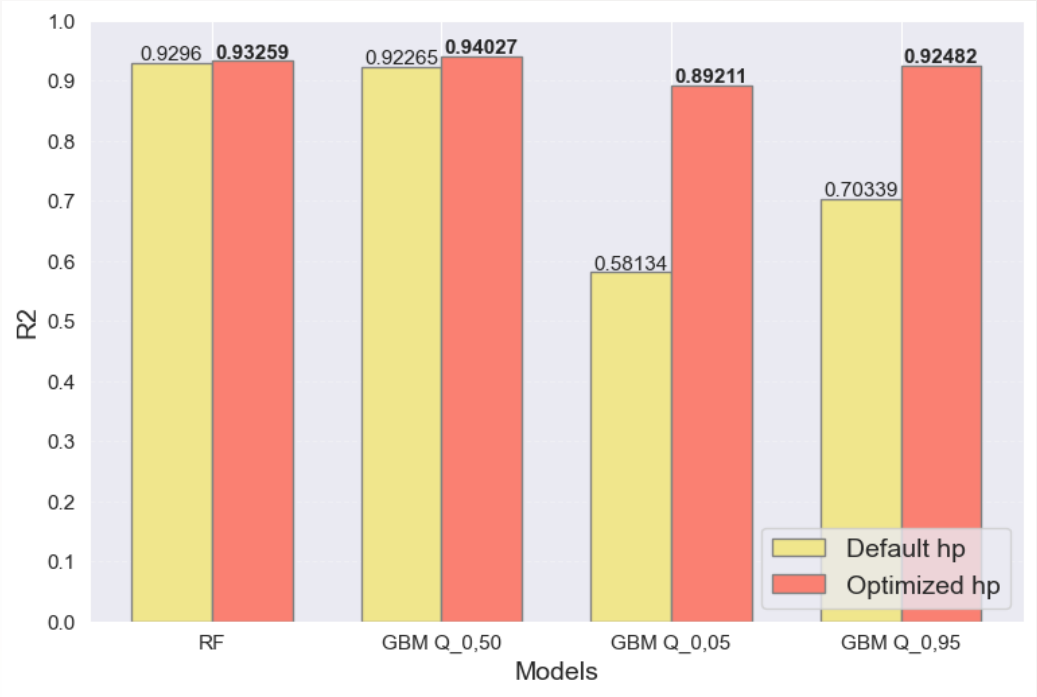
Хиперпараметри	Вредности
learning_rate	**0,1**
n_estimators	**2000**
max_depth	**28**
criterion	**„friedman_mse“**

Оптимални хиперпараметри за GBM $Q_{0,95}$ (горњи интервал предвиђаја)

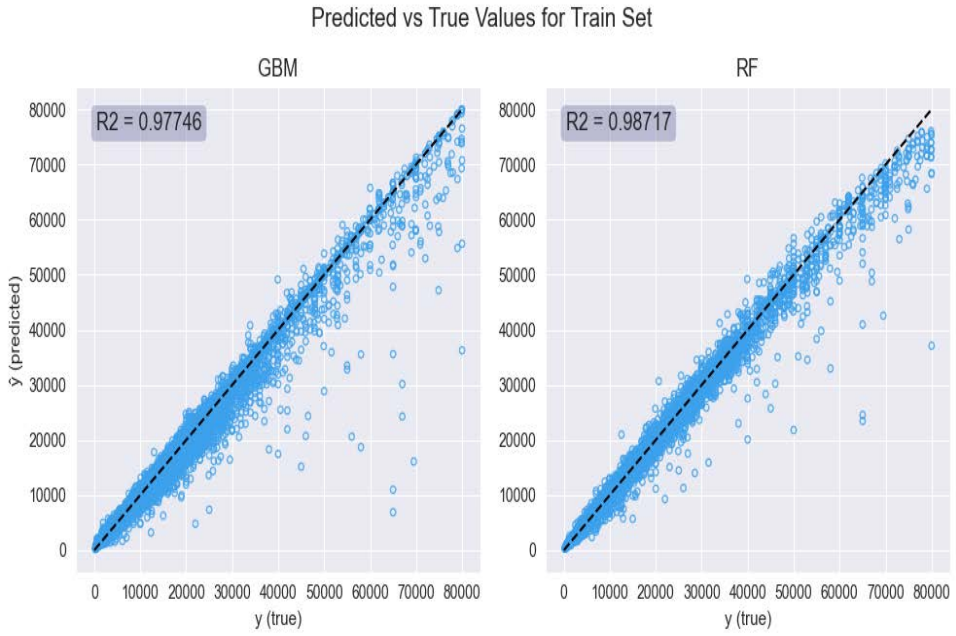
Хиперпараметри	Вредности
n_estimators	**1000**
min_samples_split	**5**
max_depth	**28**
criterion	**„friedman_mse“**

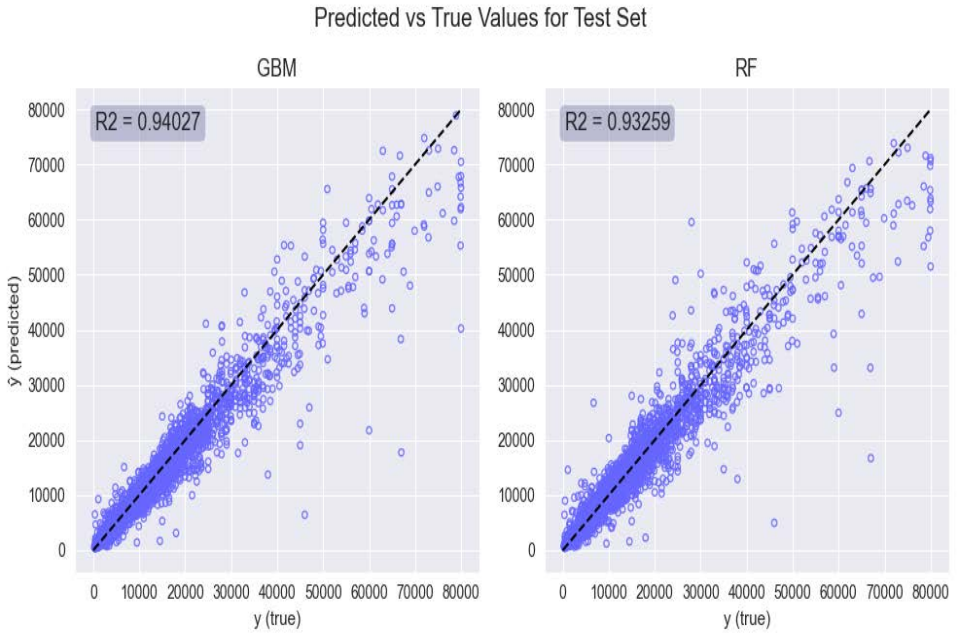
4 Резултати

	Метрике над скупом за тестирање		
Модел	RMSE	MAE	R2
RF	2726,884	1311,138	0,93259
GBM $Q_{0,50}$	**2566,822**	**1230,770**	**0,94027**
GBM $Q_{0,05}$	3449,857	1715,0888	0,89211
GBM $Q_{0,95}$	2879,736	1624,346	0,92482



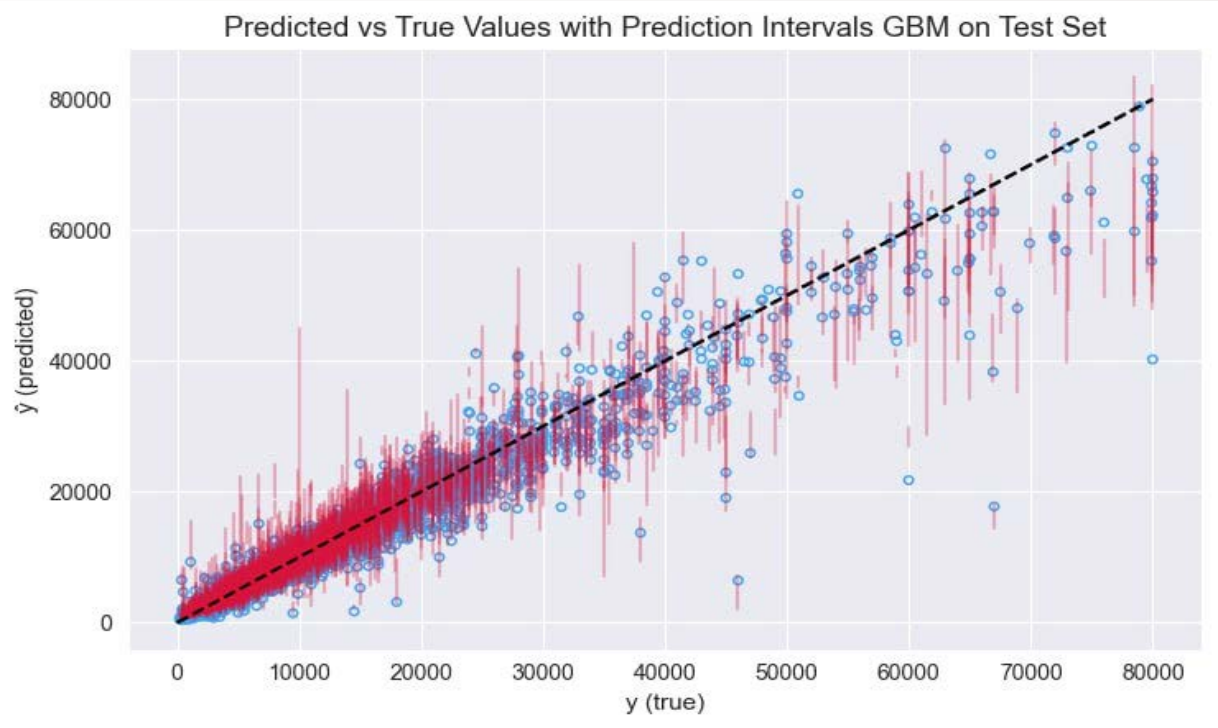
4 Резултати

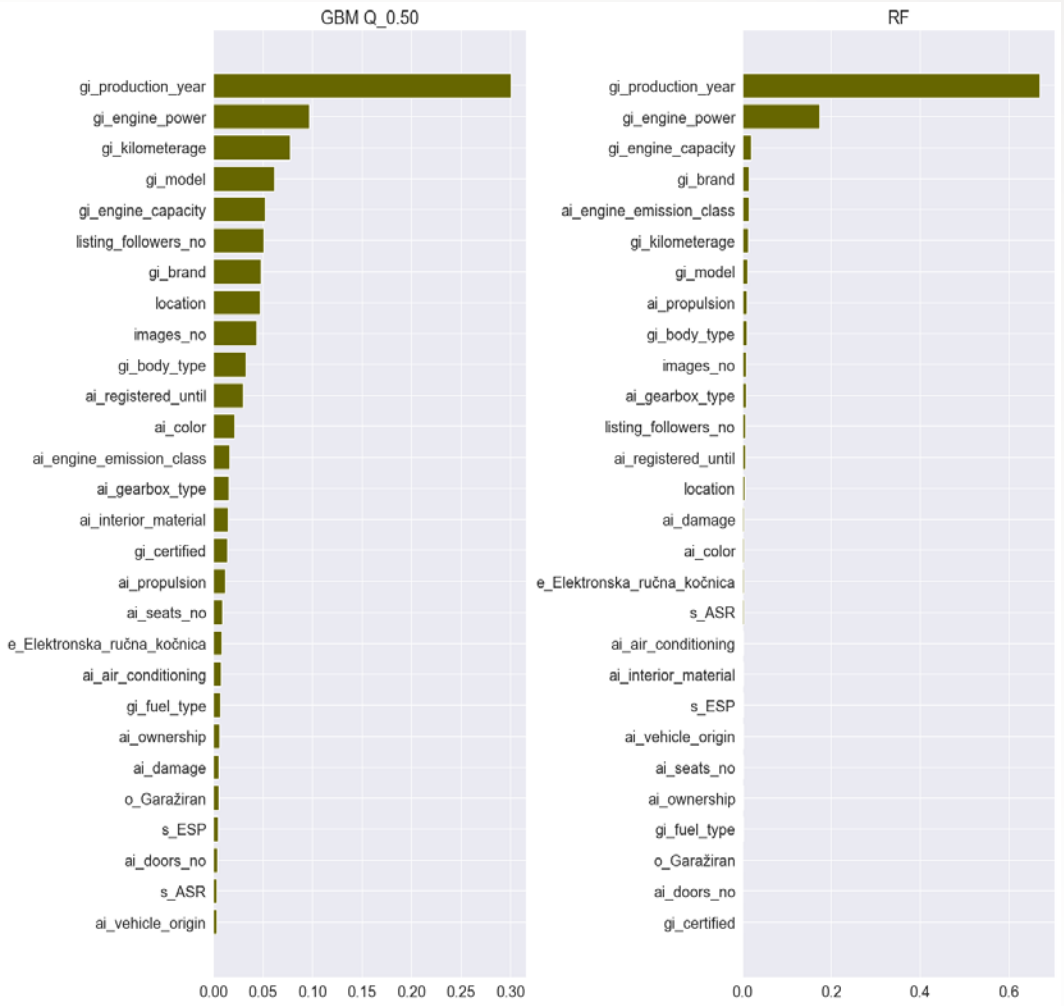

Predicted vs True Values for Train Set
GBM
R2 = 0.97746
RF
R2 = 0.98717
ŷ (predicted)
y (true)
0
10000
20000
30000
40000
50000
60000
70000
80000


Predicted vs True Values for Test Set
GBM
R2 = 0.94027
RF
R2 = 0.93259
ŷ (predicted)
y (true)
0
10000
20000
30000
40000
50000
60000
70000
80000

4

Резултати – приказ интервала предвиђаја





4

Резултати Значајности стубаца

- Обележје „година производње“ јесте најзначајније, а затим „снага мотора“ (KS).
- GBM знатно боље користи информације из свих стубаца за предвиђај него RF.

5

Расправа

- Мали узорак за кола скупља од 60.000€.
- Да ли су некоји продавци намерно дигли цену својих кола?
- Недаћа: непостојање података о тачном износу цене по којој је возило продато.

6 Закључак

- Интервали предвиђаја слабо употребљиви (свега 43,38384% покривеност од очекиваних ~90%).
- Користити текст и слике приликом предвиђања (вештачке неуронске мреже), искористити више стубаца приликом обуке, сажети обележја у мањи број...
- Велики потенцијал метода за машинско учење уз доста података.