УНИВЕРЗИТЕТ У БЕОГРАДУ

ФАКУЛТЕТ ОРГАНИЗАЦИОНИХ НАУКА

ЗАВРШНИ РАД

Предвиђање цена половних аутомобила применом алгоритама машинског учења

Ментор

Др Сандро Радовановић,

Доцент

Студент

Алекса Радојичић, 2019/0165

Београд, 2024. године

Захваљујем се срдачно и професору Андрији Петровићу који је увек стајао на располагању за савете и усмерења у току израде овог завршног рада.

АПСТРАКТ

Тржиште половних аутомобила у Републици Србији јесте доста развијено и мало ко од возача није упознат у платформу *polovniautomobili*, где корисници оглашавају углавном половна возила за продају. Због великог броја нових огласа сваког дана увиђа се да би понудиоцима и потражиоцима била од помоћи информација да ли је цена за половна кола прецењена или потцењена у односу на остала слична кола са платформе. Тај податак би поспешио продају кола јер би се брже ускладила понудна и потражна цена, што би побољшало пословање странице и задовољство корисника.

Циљ завршног рада је што боље предвидети цену половних аутомобила са мрежне странице *polovniautomobili* применом алгоритама машинског учења. На основу више основних алгоритама, обученим над скупом за обуку (23.759 слогова), процењује се који ће дати вишу вредност R2 метрике над подацима за тестирање (5.940 слогова). За случајну шуму и модел градијентног појачавања су оптимизовани хиперпараметри унакрсним вредновањем, а такође су оптимизовани GBM модели за 90%-не интервале предвиђања. Програмски језик *Python* се користио за остварење пројекта.

Модел градијентног појачавања се показао најбољим, остваривши коефицијент детерминације 0,94027 над скупом за тестирање, а интервали предвиђања се нису показали употребљивим, обухвативши стварне вредности излаза само у 43% случаја. Оптимизовани модел тачније предвиђа цену јефтинијих кола због већег узорка, а најутицајније обележје за одлуку о излазу модела је било година производње аута.

Кључне речи: *машинско учење, предвиђање, регресија, случајна шума, модел градијентног појачавања, половни аутомобили, унакрсно вредновање, интервали предвиђања, коефицијент детерминације*.

САДРЖАЈ

1. Увод 1

2. Теоријске основе 3

2.1. Машинско учење 3

2.1.1. Надгледано машинско учење 4

2.1.2. Недаће приликом обуке модела МУ 4

2.2. Дрво одлучивања 6

2.3. Ансамбл модели 8

2.3.1. Случајна шума 8

2.3.2. Модел градијентног појачавања 10

2.4. Метрике за процену у регресији 11

2.5. Раздела података на скупове за обуку, вредновање и тестирање 14

2.6. Унакрсно вредновање са k преклопа 15

2.7. Решеткаста претрага 16

2.8. Квантилна регресија 17

3. Опис истраживања 19

3.1. Фазе пројекта 19

3.2. Коришћене технологије 20

3.3. Прикуп података 21

3.4. Предобрада података 22

3.4.1. Почетни чистач 24

3.4.2. Цевовод за предобраду 27

3.5. Увиди из једноваријабилне и вишеваријабилне анализе 29

3.5.1. Информације о ступцима 29

3.5.2. Значајност стубаца пре обуке 33

3.5.4. Пирсонови саодноси 36

4. Резултати и расправа 38

4.1. Модели са подразумеваним хиперпараметрима 38
4.2. Модели са оптимизованим хиперпараметрима 40
4.2.1. Оптимални хиперпараметри за RF 41
4.2.2. Оптимални хиперпараметри за GBM $Q0{,}50$ 42
4.2.3. Оптимални хиперпараметри за GBM $Q0{,}05$ 43
4.2.4. Оптимални хиперпараметри за GBM $Q0{,}95$ 44
4.3. Процена оптимизованих модела 45
4.4. Расправа 51
5. Закључак 52
Литература 54

СПИСАК ТАБЕЛА

Табела 1. Ступци из одељка „сигурност“ 24
Табела 2. Ступци из одељка „стање“ 24
Табела 3. Ступци из одељка „опрема“ 25
Табела 4. Описне статистике ознаке 29
Табела 5. Описне статистике ступца „година производње“ 31
Табела 6. Пет најучесталијих локација продаваца кола 31
Табела 7. Пет најучесталијих марки кола 32
Табела 8. Учесталост емисионих класа мотора 32
Табела 9. Учесталост врсти оштећења кола 33
Табела 10. Перформансе модела са подразумеваним хиперпараметрима модела над скупом за тестирање 39
Табела 11. Перформансе RF, GBM $Q0{,}50$, GBM $Q0{,}05$ и GBM $Q0{,}95$ са подразумеваним хиперпараметрима модела над скупом за обуку 39
Табела 12. Прва решеткаста претрага за RF и налази 41
Табела 13. Друга решеткаста претрага за RF и налази 41
Табела 14. Оптимални хиперпараметри за RF 41
Табела 15. Прва решеткаста претрага за GBM $Q0{,}50$ и налази 42
Табела 16. Друга решеткаста претрага за GBM $Q0{,}50$ и налази 42
Табела 17. Оптимални хиперпараметри за GBM $Q0{,}50$ 42
Табела 18. Прва решеткаста претрага за GBM $Q0{,}05$ и налази 43
Табела 19. Друга решеткаста претрага за GBM $Q0{,}05$ и налази 43
Табела 20. Оптимални хиперпараметри за GBM $Q0{,}05$ 43
Табела 21. Прва решеткаста претрага за GBM $Q0{,}95$ и налази 44
Табела 22. Друга решеткаста претрага за GBM $Q0{,}95$ и налази 44
Табела 23. Оптимални хиперпараметри за GBM $Q0{,}95$ 44
Табела 24. Перформансе модела са оптимизованим хиперпараметрима над скупом за тестирање 45
Табела 25. Перформансе модела са оптимизованим хиперпараметрима над скупом за обуку 45
Табела 26. Покривеност стварних излаза интервалима предвиђаја [%] 49

СПИСАК СЛИКА

Слика 1. Модел машинског учења као математичка функција 3
Слика 2. Поређење потприлагодбе, оптималног решења и преприлагодбе код регресије (Uhlig, Alkhasli, Schubert, Tschöpe, & Wolff, 2023) 5
Слика 3. Дрво одлучивања на примеру предвиђања цене половних кола 6
Слика 4. Схема модела случајне шуме 9
Слика 5. Упоредни приказ апсолутне линеарне и квадратне функције 13
Слика 6. Раздела матрице података на скуп за обуку, вредновање и тестирање (*Everything you need to know about AI model training*, 2023) 14
Слика 7. Унакрсно вредновање у пет преклопа (Patro, 2021) 15
Слика 8. Приказ решеткасте претраге за два хиперпараметра (Hien, Tien, & Van Hieu, 2020) 16
Слика 9. Квантилна функција цене за различне вредности тау 17
Слика 10. Фазе *CRISP-DM* процесног модела (Hotz, 2024) 19
Слика 11. Релације базе података у MySQL-у 21
Слика 12. Склоп цевовода од базе података до модела 23
Слика 13. Хистограм ознаке 30
Слика 14. Хистограм логаритмоване ознаке 30
Слика 15. Хистограм са функцијом густине и кутијасти графикон ступца „година производње“ 31
Слика 16. Узајамне информације између стубаца из одељка „стање“ и ознаке 33
Слика 17. Узајамне информације између стубаца из одељка „сигурност“ и ознаке . 34
Слика 18. Узајамне информације између бројних обележаја и ознаке 35
Слика 19. Пирсонов саоднос између бројних обележаја без ознаке 36
Слика 20. Пирсонов саоднос бројних обележаја са ознаком 37
Слика 21. Поређење резултата модела са подразумеваним и оптимизованим хиперпараметрима над скупом за тестирање 46
Слика 22. График везе предвиђених и стварних излаза за оптимизоване GBM и RF над скупом за тестирање 47
Слика 23. График везе предвиђених и стварних излаза за оптимизоване GBM и RF над скупом за обуку 47
Слика 24. График везе предвиђених и стварних излаза за оптимизовани GBM над скупом за тестирање са интервалима предвиђаја 48
Слика 25. Значајност стубаца за оптимизовани GBM $Q0{,}50$ и RF 49

1. УВОД

Окосница дипломског рада јесте предвиђање цене половних аутомобила са платформе *polovniautomobili* помоћу техника машинског учења. Замисао је обучити разноврсне моделе над скупом података за обуку и онда над скупом података за тестирање проверити да ли су кола прецењена, потцењена или им је цена (мерена у еврима) објективна. Поред цене кола се пружа и поузданост предвиђаја, односно колико је модел поуздан у своју одлуку, а ова информација умногоме може бити од помоћи крајњим корисницима.

Човек је од најранијег доба имао потребу за предвиђањем. Пољопривредна револуција доноси са собом потребу за планирањем сетве и жетве, што указује на потребу имања некоје врсте календара као начина дугорочног мерења времена. Данас човек не жели само предсказати какво ће му време бити идућег дана или месеца, већ се труди предвиђати што је више ствари могуће: од тога колики ће бити четвртински новчани прилив предузећа, па до предвиђања могућности сусрета човека са ванземаљским културама (Sandberg, Drexler, & Ord, 2018).

Предвиђање је један од врло учесталих проблема данашњице који се делотворно може решити алгоритмима машинског учења. Било да се предвиђа цена некретнина, деоница или половних аутомобила, може се применити слична породица алгоритама која ће на основу података о познатим некретнинама успети предвидети цену непознатих некретнина за које алгоритам (тачније модел) није знао.

Могло би се казати да је тржиште половних аутомобила у Републици Србији доста развијено са обзиром на број огласа путничких возила од 77.973, при чему је укупан број огласа који обухватају све категорије возила једнак чак 1.371.266 (Polovni automobili, 2024). Корисник платформе *polovniautomobili* може обелоданити оглас за своја половна кола и поставити цену која им заправо не приличи: или ће их потценити или преценити. У некојим случајима ће корисник желети што пре продати своје половно возило и неће имати времена истражити колико оно заиста вреди, те прети опасност да постави много мању цену од оне коју би заједница очекивала. Друга је пак крајност да продавац додели много већу цену него што би она требала бити (у просеку) и онда је мања вероватноћа да му се јаве потенцијални купци.

Предвиђање цене половних аутомобила је подједнако корисно и за саме купце. Наиме, купац за жељена кола може помоћу обученог модела машинског учења проверити колико се огласна цена подудара са оном која је излаз из модела. Уколико је продавац потценио свој аутомобил, онда је то савршена прилика за купца јер ће његовом куповином уштедети новац у односу на остала кола на платформи која су им блиска, то јесте слична.

Светско тржишно учешће машинског учења 2022. године износило је 19,20 милијарди долара, а до 2030. очекује се повећање до штовише 225,91 милијарди (*Machine learning market size, share, growth | Trends [2030]*, 2023). Иако највећи удео тржишта заузимају Сједињене Америчке Државе (*Machine learning - Worldwide | Statista market forecast*, 2024) примећује се пораст интересовања за ову област и у нашим крајевима. Воза и остали (2023) бавили су се предвиђањем износа раствореног кисеоника у води (тачније у реци Тиси) ради предупређења и смањења њезине загађености. Цвејић и остали (2023) су такође помоћу различних ових алгоритама покушали што боље предвидети принос сунцокретовог уља.

Машинско учење може се успешно употребити у малим и средњим предузећима, а већина нових стартапа заснива се управо на овим методама (Costa-Climent, Haftor, & Staniewski, 2023). Кључан предуслов да би се примениле ове технике јесте да се поседују извесни подаци који би били предмет предвиђања. То би могла бити нека бројна вредност, односно износ нечега или категорија / разред која се предсказује. Данас је много лакше добавити податке него раније, а притом многи алгоритми машинског учења не захтевају рачунаре снажног хардвера, као што су дрво одлучивања или линеарне регресија, а који и дан данас постижу добре резултате у појединим гранама и задацима.

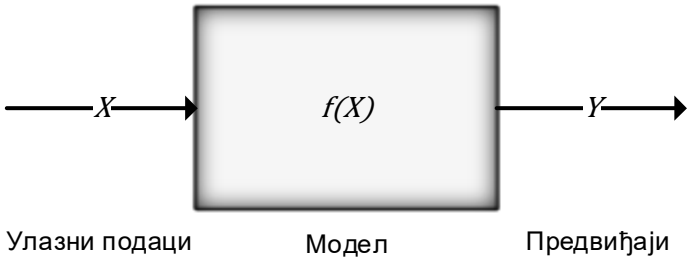
Циљ овога рада јесте пронаћи који од честих алгоритама машинског учења ће се добро показати на задатку процене цене половног аутомобила. Они који се добро покажу са подразумеваним хиперпараметрима ће бити додатно удешени и усклађени за још већи раст перформанси. За алгоритам градијентног појачавања ће се наћи интервали предвиђаја како би корисник имао податке о горњој и долњој граници у којој би се у 90% случаја нашла стварна цена кола.

2. ТЕОРИЈСКЕ ОСНОВЕ

У идућим одељцима прелазиће се теоријски концепти неопходни за разумевање и праћење осталих битних поглавља завршнога рада.

2.1. МАШИНСКО УЧЕЊЕ

Објашњење машинског учења [МУ] које се приписује Артуру Самуелу се врло често чује у овој области и гласи: „**Машинско учење** је научна област која оспособљава рачунаре да уче без да су непосредно програмирани" (Mahesh, 2020). Главна намена машинског учења јесте предвиђање излазних вредности (енгл. *outputs*) на основу улазних вредности (енгл. *inputs*) које се убацују у модел за машинско учење. Математички се модел машинског учења може исказати овако: потребно је наћи функцију $f(X)$ која ће улазе X успешно пресликати у излазе Y (Lotz, 2018). X представља матрицу података, коју можемо угрубо посматрати као табелу са ступцима и редовима, функција $f(X)$ чини модел МУ, а Y јесте матрица предвиђаја (енгл. *predictions*) (Слика 1).



Слика 1. Модел машинског учења као математичка функција

Машинско учење се може поделити на три велике подскупине према врсти реалних проблема који бивају решавани (Nikolić & Zečević, 2019):

1) Надгледано (енгл. *supervised*);
2) Ненадгледано (енгл. *unsupervised*);
3) Учење уз поткрепе (енгл. *reinforcement learning*).

Предвиђање цене половних аутомобила потпада у проблеме надгледаног машинског учења, па о преосталим двема скупинама неће бити речи у овому раду.

2.1.1. Надгледано машинско учење

Код **надгледаног машинског учења** модел треба обучити подскуп матрице података X_{train} тако што ће некако пронаћи пресликавање за већ познате и приложене излазне вредности Y_{train}. Напослетку ће модел предвидети излазне вредности Y_{test} за други подскуп матрице података X_{test} чији слогови (енгл. *instances*), односно редови нису коришћени за обуку (Shalev-Shwartz & Ben-David, 2014). X_{train} представља скуп за обуку (енгл. *training set*), а X_{test} скуп за тестирање (енгл. *testing set*). Битно је да сваки појединачни слог из скупа за обуку садржи и податак о својој излазној вредности који се назива **ознака** (енгл. *label*), док је ознака код свих редова скупа за тестирање непозната.

Две су најпознатије подгране надгледаног учења и оне су подељене према природи излазних вредности: уколико су излазне вредности бројне ($Y \in \mathbb{R}$) онда је у питању **регресија**, а ако су оне категорије (цели бројеви или знаковне вредности), онда се то назива **сврставање** (енгл. *classification*) (Deisenroth, Faisal, & Ong, 2020). Окосница у овом раду је регресија јер је цена бројна и ненегативна променљива и онда ће се у будуће подразумевати да се говори о регресији, а не о сврставању.

2.1.2. Недаће приликом обуке модела МУ

Једне од главних потешкоћа које се јављају током обуке модела за машинско учење јесу потприлагодба и преприлагодба.

Потприлагодба (енгл. *underfitting*) је појава да се модел не може успешно прилагодити скупу за обуку, па самим тим и скупом за тестирање (Grus, 2019). Нека D представља скуп података дефинисан као унија обележаја X и ознака Y, а функција $\text{accuracy}_D(\hat{Y})$ меру колико добро предвиђаји из некојег модела $\hat{Y}$ описују податке D. Предвиђаји $\hat{Y}$ се потприлагођавају скупу за обуку D_{train} ако постоји Y' такав да важи (Bilmes, 2020):

$$\begin{aligned} \text{accuracy}_{D_{train}}(\hat{Y}) &< \text{accuracy}_{D_{train}}(Y') \\ \text{accuracy}_{D_{test}}(\hat{Y}) &< \text{accuracy}_{D_{test}}(Y') \end{aligned} \tag{1}$$

Преприлагодба (енгл. *overfitting*) је појава кад се модел превише прилагоди скупом за обуку, па модел много боље објашњава податке за обуку него скуп за тестирање

(Winn, 2023). Обучени модел је у превеликој мери научио обрасце у скупу за обуку и тиме је изгубио могућност уопћавања за невиђене слогове. Предвиђаји $\hat{Y}$ се преприлагођавају скупу за обуку D_{train} ако постоји Y'' такав да важи (Bilmes, 2020):

$$\begin{aligned} \text{accuracy}_{D_{train}}(\hat{Y}) &> \text{accuracy}_{D_{train}}(Y'') \\ \text{accuracy}_{D_{\text{test}}}(\hat{Y}) &< \text{accuracy}_{D_{test}}(Y'') \end{aligned} \tag{2}$$

У продужетку је дат цртеж три графика која описују потприлагодбу, оптимално решење и преприлагодбу (Слика 2) за једну објашњавајућу променљиву и бројну вредност излаза (регресија). Предвиђавна линија модела који потприлагођава неће обухватити многе слогове; модел који преприлагођава биће доста сложенији и моћи ће савршено моделовати целокупан скуп за обуку али ће доста грешити над скупом за тестирање; оптималан модел ће постићи савршени склад како би се остварила највећа могућа тачност на невиђеним подацима.

Слика 2. Поређење потприлагодбе, оптималног решења и преприлагодбе код регресије (Uhlig, Alkhasli, Schubert, Tschöpe, & Wolff, 2023)

Преприлагодба се може смањити на следеће начине:

1) Проширењем скупа за обуку;

   Што више слогова има у подацима за обуку то је теже начинити хиперраван (кад је број димензија, односно обележаја n) која ће садржати све њих. Недостатак ова приступа је што ће време за обуку модела бити дуже, повећава се време анализе и чишћења података, увећава трошак прикупа података (Ying, 2019).

2) Смањењем броја стубаца скупа података;
3) Коришћењем једноставнијег модела са мањом сложеношћу (Hawkins, 2004);
4) Поједностављењем хиперпараметара модела;

   За сваки модел МУ који су коришћени ће се објаснити у наставку његови хиперпараметри и како се односе на преприлагодбу.

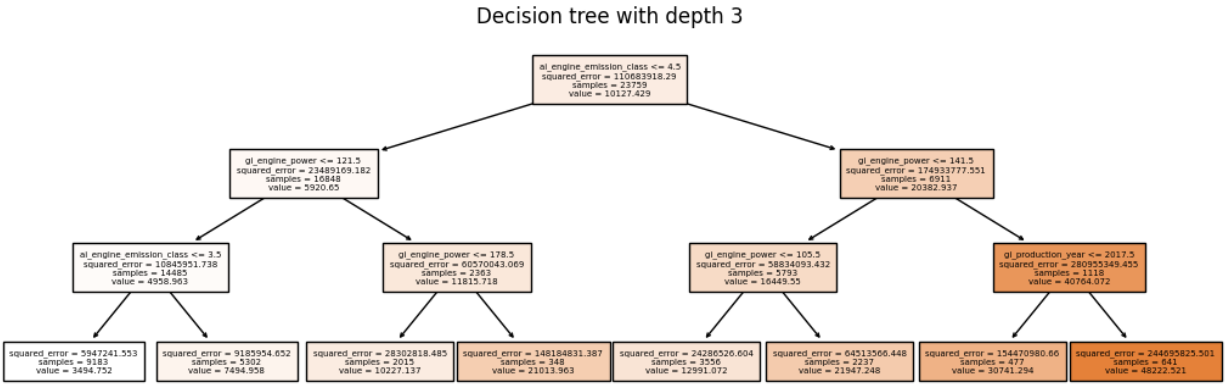
5) Коришћењем унакрсног вредновања (Rao, Fung, & Rosales, 2008).

   О овом приступу ће се више говорити у одељку 0.

2.2. ДРВО ОДЛУЧИВАЊА

Модел **дрва одлучивања** [DT] заснован на алгоритму *ID3* ради по следећем начелу (Delibašić, Suknović, & Jovanović, 2009):

1) Рачуна се информациона добит матрице података;
2) Бира се обележје (енгл. *feature*), односно стубац из матрице података за гранање који највише повећава информациону добит;
3) Поступак се понавља док се не истроше сва обележја или док није задати критеријум задовољен.



Слика 3. Дрво одлучивања на примеру предвиђања цене половних кола

Дрво одлучивања у регресији подржава више различитих хиперпараметара који утичу на структуру модела. Неки од њих јесу (*scikit-learn: Machine Learning in Python*, 2024):

1) Критеријум разделе (*criterion*);

   Функција која рачуна инфомрациону добит и подржава следећа четири критеријума: квадратна грешка, Фридманова просечна квадратна грешка, апсолутна грешка и Поасонов критеријум.

2) Максимална дубина дрва (*max_depth*);

   Дубља дрва показују склоност преприлагодби јер користе више обележаја, а плића дрва потприлагодби (Kroese, Botev, & Taimre, 2019). Што се иде више у дубину то се број слогова смањује који ће се наћи у листовима и онда се на основу мањег броја узорака доноси предвиђавна одлука која неће ваљати за нове податке.

3) Минимални број слогова за гранање (*min_samples_split*);

   Најмањи број редова неопходан да би се извршило следеће гранање приликом творбе дрвета. Већи број казује да је неопходан већи узорак да би дошло до разделе што смањује преприлагодбу. Минималан број чворова за гранање је један, што производи сложенија дрва.

4) Минимални број слогова у листу (*min_samples_leaf*);

   Уколико би се гранањем добила два листа таква да један од њих или оба обухватају број слогова мањи од вредности ова хиперпараметра, гранање се неће извршити.

5) Максимални број листова[1] (*max_leaf_nodes*);

   Остављају се они листови који највише доприноше повећању информационе добити.

6) Минимално смањење нечистоће (*min_impurity_decrease*).

   Чвор ће се гранати само уколико повећава информациону добит за вредност већу или једнаку *min_impurity_decrease.*

Предност дрва одлучивања јесте у њиховој једноставности, тумачљивости самога модела и отпорности на изнимке (енгл. *outliers*) (Song & Ying, 2015). Сложени модели (особито вештачке неуронске мреже) се теже могу похвалити својом тумачљивошћу за разлику од дрва одлучивања, код којега се јасно види графичким приказом модела која обележја су утицала на дати излаз. Слика 3 показује цртеж дрва одлучивања дубине три за предвиђање цене половних аутомобила.

[1] Листови су они чворови у дрву који немају децу, односно који се не гранају даље.

2.3. АНСАМБЛ МОДЕЛИ

Још 1996. године Брајман (1996) показује да се скупина модела за предвиђање обучених над истим матрицама података, које називамо **ансамблима**, може боље показати него појединачни предвиђачи. Ансамбл алгоритми машинског учења јесу у ствари мета-алгоритми јер су засновани и саткани од мноштва основних предвиђача (енгл. *base predictors*) који деле исти алгоритам (Delibašić, Suknović, & Jovanović, 2009). Основни предвиђачи углавном не требају бити сложени, већ простији модели, те отуда назив за њих слаби учиоци (енгл. *weak-learners*). У протеклих неколико година ансамбли модела победили су на многим такмичењима МУ над табеларним подацима, укључујући и она на *Kaggle* платформи (Stamp, Chandak, Wong, & Ye, 2021), а вероватно ће се такав тренд наставити.

Учење помоћу ансамбала се може поделити на две гране (Alzubi, Nayyar, & Kumar, 2018) са становишта приступа на којима се заснивају:

1) Редни приступи – творе се слаби учиоци једни за другим заредом. Напослетку ће се говорити о моделу градијентног појачавања, као једног од познатог представника и модела који је коришћен за предвиђање цена половних кола;
2) Упоредни (паралелни) приступи – основни предвиђачи су међусобно независни и творе се упоредо. Једна њихова предност у односу на моделе реднога приступа јесте што подржавају паралелизацију и дистрибуирано извршавање. У наставку ће бити речи о алгоритму случајној шуми.

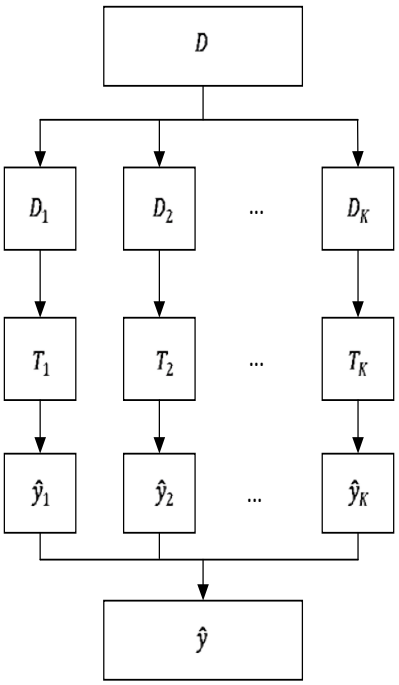
2.3.1. Случајна шума

Случајна шума [RF] је један од веома коришћених ансамбл алгоритама сачињен од више простих дрва одлучивања (Liu, Wang, & Zhang, 2012). Она користи **багинг** (енгл. *bagging*, од *bootstrap aggregating*) технику како би најпре створила више подскупа са различним редовима и / или ступцима (Alzubi, Nayyar, & Kumar, 2018), што постиже делимичну независност између сваког појединачног подскупа. У опћем случају се не мора као основни предвиђач користити дрво одлучивања, већ то могу бити и други алгоритми МУ. Над сваким подскупом се обуче дрва одлучивања и исходни предвиђаји ће се, у случају регресије, добити као аритметичка средина предвиђаја свих обучених дрва:

$$\hat{y} = \frac{1}{K}(\hat{y}_1 + \hat{y}_2 + \ldots + \hat{y}_K) = \frac{1}{K}\sum_{k=1}^{K} \hat{y}_k \quad (3)$$

где је $\hat{y}$ крајњи предвиђај, а K укупан број дрва.

У продужетку је схематски приказана случајна шума која предвиђа, на пример, цену једног појединачног слога из матрице података:



Слика 4. Схема модела случајне шуме

D означава главни скуп за обуку, док су $D_k,\ k = 1, \ldots, K$ подскупи за обуку који имају насумичне редове и насумичне подскупе атрибута изворног скупа D. T_k се односи на k-то дрво одлучивања, где $k = 1, \ldots, K$.

Случајне шума, уз све хиперпараметре које има дрво одлучивања, поседује и својствени хиперпараметар *n_estimators* („број процењивача“, односно број дрвећа у ансамблу). Већи износ дрвећа усложњава модел, уводи могућност преприлагодбе и продужава време обуке.

Предност алгоритма случајне шуме у односу на дрво одлучивања јесте што је мање подложан преприлагодби, али га прати мања тумачљивост и разумљивост због његове сложености (Stamp, Chandak, Wong, & Ye, 2021). Притом, захтева доста велику

меморијску потрошњу што се више процењивача користи за разлику од модела градијентног појачавања о којем ће сада бити реч.

2.3.2. Модел градијентног појачавања

Модели градијентног појачавања [GBM] утврђени су на појму и замисли **појачавања** (енгл. *boosting*) код којег основни предвиђачи предсказују мало боље од насумичног погађања (Alzubi, Nayyar, & Kumar, 2018). Сваки предвиђач надовезује се на претходног надопуњујући га (то јесте појачавајући га) и смањује његову грешку предвиђања како би се сачинио крепкији и моћнији модел. Појачавање омогућава градњу много истанчанијег модела који ће лакше предвидети оне слогове над којима су се простији модели машинског учења мучили, уз поштовање једног од предуслова да су ефикасно основни предвиђачи скројени (Shalev-Shwartz & Ben-David, 2014). Као и код RF и код GBM су најчешће у сржи дрва одлучивања као основни предвиђачи.

Начелно се код GBM итеративно творе основни предвиђачи такви да показују максимални саодност са негативним градијентом **функције губитка** / **цене** (енгл. *loss / cost function*) (Natekin & Knoll, 2013). На тај начин ће сваки идући основни предвиђач мало поправити вредност функције губитка што ће допринети бољем предвиђају. Пошто је потребно у свакој итерацији рачунати градијент, функција губитка би требала бити диференијабилна, односно да има први извод, мада то не мора нужно бити случај.

Пошто је дрво одлучивања основни предвиђач у GBM-у, онда поседује и хиперпараметре DT-а (види страну 6). Неки од битних хиперпараметара својствени за GBM јесу:

1) функција губитка (*loss*);

   Врло често употребљаване то функције губитка јесу функције квадратне (1) и апсолутне грешке[2] (2) (Friedman, 2001), а *scikit-learn* библиотека подржава и Хуберову (упарује квадратну и апсолутну грешку) и квантилну грешку (за квантилну регресију) (*scikit-learn: Machine Learning in Python*, 2024).

[2] Пошто функција цене $L(y, \hat{y})$ није диференцијабилна у тачки $y = \hat{y}$, њезин извод се не рачуна аналитички већ се налази апроксимацијом нумеричким методама.

$$L(y, \hat{y}) = \frac{1}{2}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2 \quad (4)$$

$$L(y, \hat{y}) = \sum_{i=1}^{m}|y_i - \hat{y}_i| \quad (5)$$

2) стопа учења (*learning_rate*);

   Представља ненегативан реалан број који умножава градијент функције губитка који се рачуна приликом сваке итерације, умањујући његово дејство (Natekin & Knoll, 2013). Обично је између 0 и 1, мада може бити и већи. Што је стопа учења већа, то ће се спорије мењати вредности излазне променљиве за време обуке. Мањи износ може довести до модела који је потприлагођен, а виши до преприлагодбе.

3) број процењивача (*n_estimators*);

   Аналогно броју процењивача у случајној шуми, с тим што је у једном временском тренутку потребно творити само једно дрво одлучивања, те овај хиперпараметар указује и на број итерација.

4) алфа (*alpha*).

   Утиче на број квантила уколико је за функцију губитка одабрана квантилна или Хуберова функција губитка. Креће се у распону од 0 до 1, не укључујући их.

Модели градијентног појачавања могу бити штовише тачнији од неуронских мрежа и тумачљивији од линеарних модела кад су основни предвиђачи дрва одлучивања (Konstantinov & Utkin, 2021). Показују се изузетно добро у пракси и спадају међу најбоље методе машинског учења уопће (Nikolić & Zečević, 2019) и то јесу разлози због чега ће се за предвиђање цене половних аутомобила укључити GBM у овом раду.

2.4. МЕТРИКЕ ЗА ПРОЦЕНУ У РЕГРЕСИЈИ

Приликом разматрања који модел нам више одговара неопходно је одабрати метрику која ће служити за поређење и процену (евалуацију) модела. У делу који следи ће се објаснити просечна апсолутна грешка, просечна квадратна грешка, корен просечне квадратне грешке и коефицијент детерминације.

Просечна апсолутна грешка [MAE] (енгл. *mean absolute error*) представља разлику између предсказаних и стварних вредности излаза, занемарујући знак (да ли је разлика позитивна или негативна). m у формули (3) и убудуће значи укупан број слогова, односно редова матрице података.

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|y_i - \hat{y}_i| \tag{6}$$

Просечна квадратна грешка [MSE] (енгл. *mean squared error*) је објашњена на следећи начин, аналогно апсолутној функцији губитка (2). Недостатак ове функције јесте што није отпорна на изнимке (Plevris, Solorzano, Bakas, & Ben Seghier, 2022), али може дати путоказе да су изнимци присутни у матрици података над којој се врши анализа и да се могу средити. Неретко се употребљава као функција грешке у алгоритмима машинског учења приликом обуке модела због лакоће добијања првог извода, на пример преко градијентног спуста (енгл. *gradient descent*).

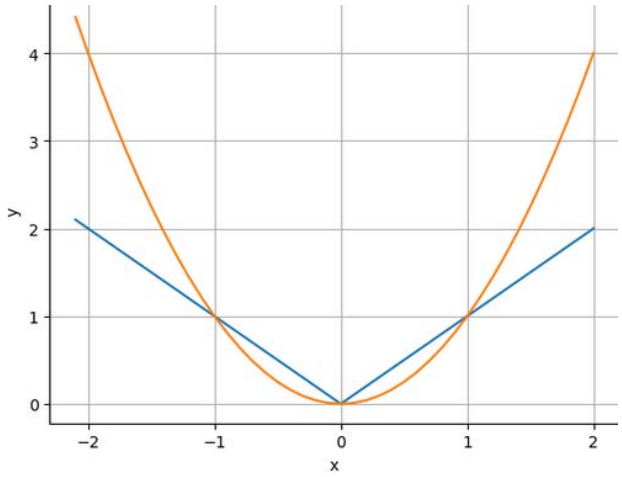
$$MSE = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2 \tag{7}$$

Варијација MSE метрике је **корен просечне квадратне грешке** [RMSE] (енгл. *root mean squared error*) која се добија као квадратни корен просечне квадратне грешке. Повољност у односу на MSE јесте што је тумачљивији јер је исте скале као и сами предвиђаји. У примеру предсказања цене половних кола мерна јединица RMSE метрике је евро, исто као и код излазног ступца.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2} \tag{8}$$

Пошто је у жижи MSE квадратна функција, онда је јасно да ће они слогови чија је квадратна разлика (резидуал) између стварних y и процењених $\hat{y}$ вредности велика (>> 1) много више повећати ову метрику, те тиме казнити модел. Са друге стране, познато је да ће производ два броја између нуле и један дати мањи број (Слика 5), те отуда се још̌те мање кажњавају они слогови чије се предвиђене и стварне вредности скоро поклапају.

Мањак метрика MAE, MSE и RMSE јесте што не указују на свеукупне перформансе модела за регресију и пружају само апсолутну информацију о грешци.



Слика 5. Упоредни приказ апсолутне линеарне и квадратне функције

Коефицијент детерминације [R2] представља удео варијабилитета који је модел успешно објаснио од укупног варијабилитета (Вуковић & Булајић, 2014). Математичка формула је дата испод, а $\bar{y}$ је аритметичка средина стварних излаза:

$$R2 = \frac{SS_{model}}{SS_{total}} = \frac{SS_{total} - SS_{res}}{SS_{total}} = 1 - \frac{SS_{res}}{SS_{total}} = 1 - \frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{m}(y_i - \bar{y})^2} \qquad (9)$$

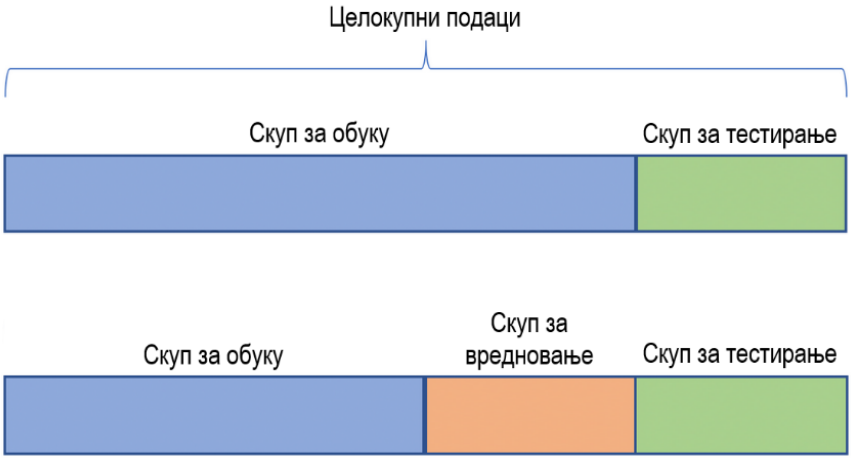
Савршен модел који успешно објашњава целокупан варијабилитет имаће $R2 = 1$ и то значи да се предвиђене вредности поклапају са стварним излазима. $R2 = 0$ исказује да модел не може боље описати везу између улаза и излаза од аритметичке средине предвиђаја (водоравне линије), чинећи модел практично бескорисним. Велика је заблуда многих да је опсег вредности R2 од 0 до 1, из разлога што се може догодити да модел горе описује податке од аритметичке средине (Chicco, Warrens, & Jurman, 2021) и онда је прави опсег коефицијента детерминације: $R2 \in (-\infty, 1]$.

Неповољност R2 метрике јесте кад је њезин износ негативан и тада се не може разазнати у коликој мери модел греши, због непостојања конкретне долње вредности опсега (односно $-\infty$) (Chicco, Warrens, & Jurman, 2021).

2.5. РАЗДЕЛА ПОДАТАКА НА СКУПОВЕ ЗА ОБУКУ, ВРЕДНОВАЊЕ И ТЕСТИРАЊЕ

Већ је раније поменуто у одељку за надгледано учење да се матрица података мора поделити на **скуп за обуку** и **скуп за тестирање**. Модел се обучи скупом за обуку и онда се испита његова делотворност и перформансе над скупом за тестирање. Пошто модел у току обуке установи обрасце над подацима за обуку како би исправно предвидео излаз, онда неће бити веродостојни и смислени резултати његове процене јер је суштина да се исправно предвиђају невиђени слогови (Deisenroth, Faisal, & Ong, 2020).

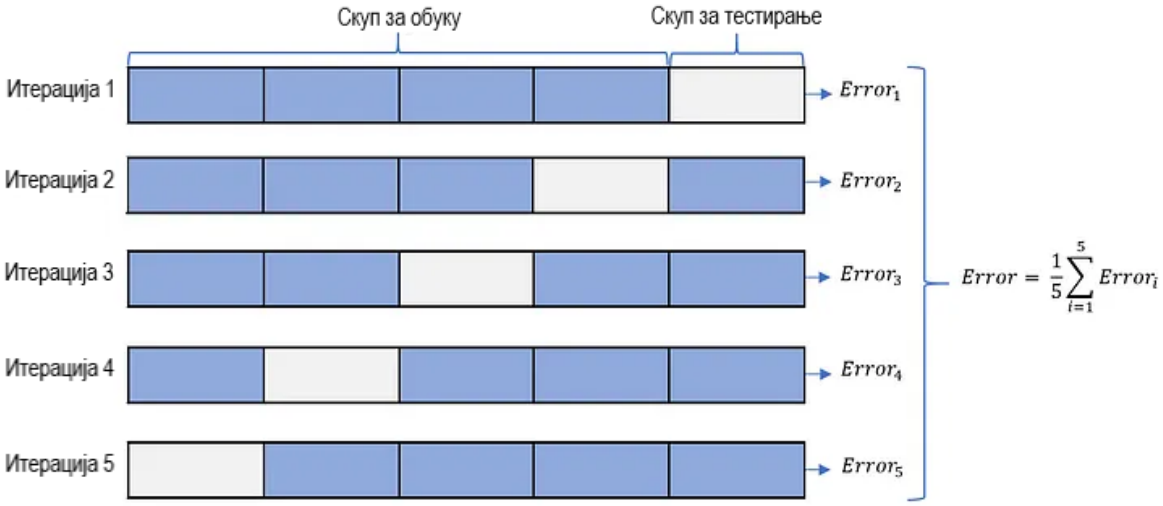
Када се врши оптимизација хиперпараметара модела машинског учења, онда није никако пожељно користити скуп за тестирање приликом њихова одабира. Уколико се ипак тако поступи, модел се преприлагођава подацима и тиме губи способност уопћавања, јер неће моћи исправно предвидети излазну променљиву нових слогова (Xu & Goodacre, 2018). Наведена потешкоћа се делотворно отклања уводом још једнога скупа података: скупа за **вредновање (валидацију)**.



Слика 6. Раздела матрице података на скуп за обуку, вредновање и тестирање (*Everything you need to know about AI model training*, 2023)

2.6. УНАКРСНО ВРЕДНОВАЊЕ СА *K* ПРЕКЛОПА

Унаксрсно вредновање са *k* преклопа (енгл. *k-fold cross validation*) је једна од чувених техника за вредновање обученог модела (Hawkins, Basak, & Mills, Assessing model fit by cross-validation, 2003), чија је тежња да доведе до бољих налаза него обука модела над само једним скупом за вредновање / тестирање. Подаци се најпре деле на *k* раставних (дисјунктних) покдскупа, који се зову преклопи (енгл. *folds*), приближно истоветне величине (Berrar, 2019). У свакој од укупно *k* итерација се по један различни преклоп користи за тестирање, а преосталих k - 1 преклопа за обуку (Nti, Nyarko-Boateng, Aning, & others, 2021). Уколико се као регресивне метрике процене модела користе RMSE, MAE и R2, онда ће коначна исходна метрика бити једнака упросеченим просечним квадратним грешкама добијеним у свим итерацијама (Слика 7). Предност оваквог унакрсног вредновања јесте што се сви слогови матрице података употребљавају за тестирање што може довести до мање преприлагодбе и непристраснијих резултата (Rao, Fung, & Rosales, 2008).

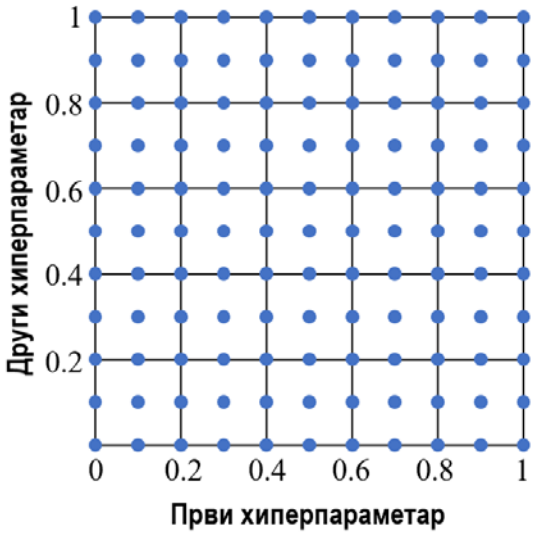


Слика 7. Унакрсно вредновање у пет преклопа (Patro, 2021)

У пракси се учестало користи $k = 10$ или $k = 5$ (Nti, Nyarko-Boateng, Aning, & others, 2021), с тим што се веће вредности *k* више преприлагођавају скупу и притом увећавају свеукупно време обуке.

2.7. РЕШЕТКАСТА ПРЕТРАГА

Један од можда најчувенијих техника оптимизовања хиперпараметара модела за машинско учење јесте **решеткаста претрага** (енгл. *grid search*). То је пријемчива и исцрпна метода која упарује разнородне хиперпараметре међусобно, стварајући решетку свих комбинација вредности за сваки хиперпараметар (Слика 8) (Zöller & Huber, 2021).



Слика 8. Приказ решеткасте претраге за два хиперпараметра (Hien, Tien, & Van Hieu, 2020)

Добре стране решеткасте претраге јесу што ју није тешко имплементовати због своје интуитивности, а поред тога подржава паралелизацију (Bergstra & Bengio, 2012). Недостатак решеткасте претраге огледа се у спорости извршавања и неистанчаном начину проналаска хиперпараметара јер се морају испитати све њихове комбинације. Примера ради, библиотека *scikit-optimize* нуди снажнију вештину претраге хиперпараметара под називом Бајесова претрага која користи Бајесово оптимизовање како би сузила хиперпараметарски простор претраге.

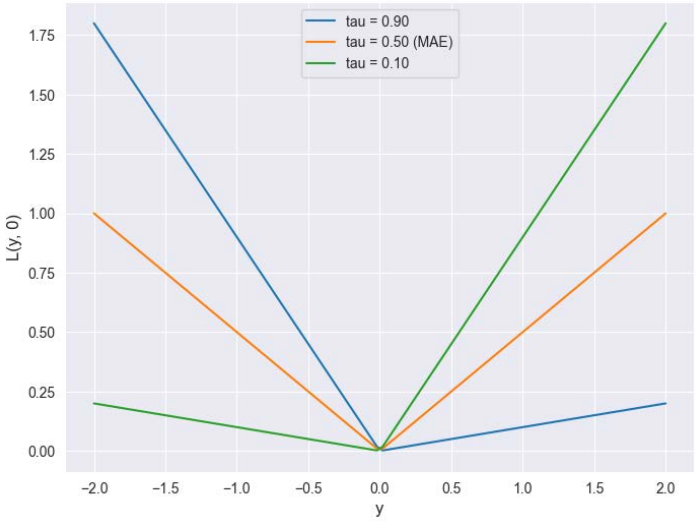
2.8. КВАНТИЛНА РЕГРЕСИЈА

Квантилна регресија, за разлику од уобичајене регресије, придодаје другачије тежинске множиоце разлици предвиђавних вредности регресивног модела и стварних вредности, да би се потом ове разлике минимизовале (Le Cook & Manning, 2013). Функција губитка за квантилну регресију захтева и износ квантилне вероватноће (представља хиперпараметар модела), а ево њезиног математичког објашњења (Kocherginsky, He, & Mu, 2005):

$$L_\tau(y, \hat{y}) = \begin{cases} \sum_{i=1}^{m} \tau \cdot (y_i - \hat{y}_i), & y_i \geq \hat{y}_i \\ \sum_{i=1}^{m} (1 - \tau) \cdot (y_i - \hat{y}_i), & y_i < \hat{y}_i \end{cases} \quad (10)$$

Квантил $Q_\tau(x)$ је вредност обележја x таква да τ (у постоцима) слогова x има мању вредност од $Q_\tau(x)$ (Вуковић, 2012), где $0 < \tau < 1$. Када је $\tau = 0{,}50$ (медијана), онда се функција губитка за квантилну регресију своди на просечну апсолутну грешку (MAE функцију губитка). Уколико је, на пример, $\tau = 0{,}90$ то значи да ће функција кажњавати девет пута више када су стварни излази већи од предвиђаја (модел потцењује). Другачије речено, више ће се уступати место предвиђајима који прецењују и то у овом случају 90% у просеку.

Испод се налази график квантилне функција губитка за $\tau \in \{0{,}10, 0{,}50, 0{,}90\}$, за зацртано $\hat{y} = 0$ (Слика 9).



Слика 9. Квантилна функција цене за различне вредности тау

Квантилна регресија може бити згодна за примену у случајима када је нелинеарна веза између улазних променљивих и излазне променљиве (ознаке) (Le Cook & Manning, 2013). Велика погодност је што се помоћу квантилне регресије могу наћи интервали предвиђаја за модел који то омогућава.

Интервал предвиђаја представља одсечак који има долњу и горњу границу где би се стварна вредност излаза из скупа за обуку требала налазити. 90% интервал предвиђаја би се добио тако што би се обучио модел за $\tau = 0{,}95$ који чини горњу међу, и модел за $\tau = 0{,}05$ који представља долњу границу. То значи да ће одсечак приближно у 90% случаја обухватити стварну вредност излазне променљиве, макар на скупу за обуку, што не мора бити тако и на скупу за тестирање (обично ће бити мање). Уколико модел МУ подржава квантилну грешку за функцију губитка онда се интервал предвиђаја математички дефинише на следећи начин.

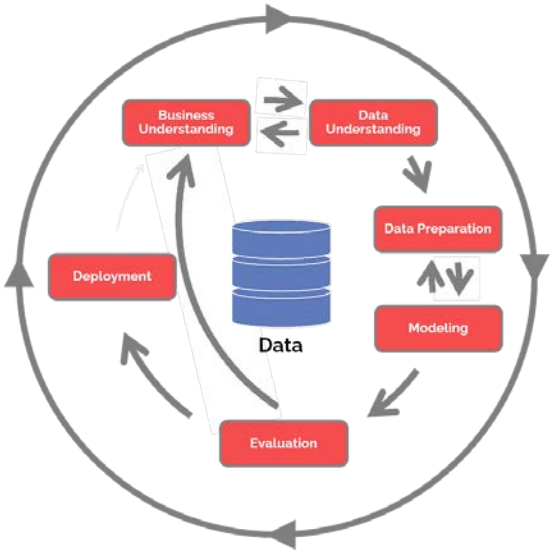
Модел градијентног појачавања подржава квантилну и Хуберову функцију губитка, што значи да подржава и израчунавање интервала предвиђаја. Да би се добили прецизнији интервали неопходно је моделе GBM $Q_{0,95}$ и GBM $Q_{0,05}$ оптимизовати, односно њихове хиперпараметре. Није свеједно да ли ће горњи износ одсечка износити 500 или 10.000€ ако је стварна цена половних кола 500 евра. Што су интервали ужи, то су од већег значаја крајњем кориснику модела јер су ближи тачној вредности.

3. ОПИС ИСТРАЖИВАЊА

У одељцима који предстоје ће се казивати о фазама и склопу пројекта, коришћеним технологијама, методологија прикупа и складиштења података. Биће речи о томе како је постепено обрађена матрица података, склоп цевовода за предобраду, као и какви су подаци над којима се моделовала цена половног аута (састав стубаца, значајност стубаца пре обуке и Пирсонов саоднос).

3.1. ФАЗЕ ПРОЈЕКТА

Пројекат је остварен по узору на ***CRISP-DM*** (*CRoss Industry Standard Process for Data Mining*) процесни модел за развој пројеката за откривање законитости у подацима који чини ток израде пројекта поузданијим, исплативијим, бржим, уједначенијим и лакшим за управљање (Wirth & Hipp, 2000). Кораци *CRISP-DM* су испод (Слика 10).



Слика 10. Фазе *CRISP-DM* процесног модела (Hotz, 2024)

Све фазе модела са модела *CRISP-DM* методологије (Слика 10) су урађене осим последњек корака *Deployment*, који се односи на пуштање модела крајњим корисницима.

Да би се добили ваљани резултати неопходно је да подаци буду добро припремљени и у оптималном облику (већина модела захтева да сва обележја буду бројног типа, а понеки подржавају и категорије као знаковне вредности). Не може се очекивати од упрљаних и несређених података да изнедре модел који ће добро предсказавати за

нове вредности. Отуда се у току фазе Оцене (*Evaluation*) (Слика 10) треба враћати на почетни корак Разумевања података (*Data Understanding*) и Разумевање пословног проблема (*Business Understanding*) уколико модел не доводи до жељених налаза.

3.2. КОРИШЋЕНЕ ТЕХНОЛОГИЈЕ

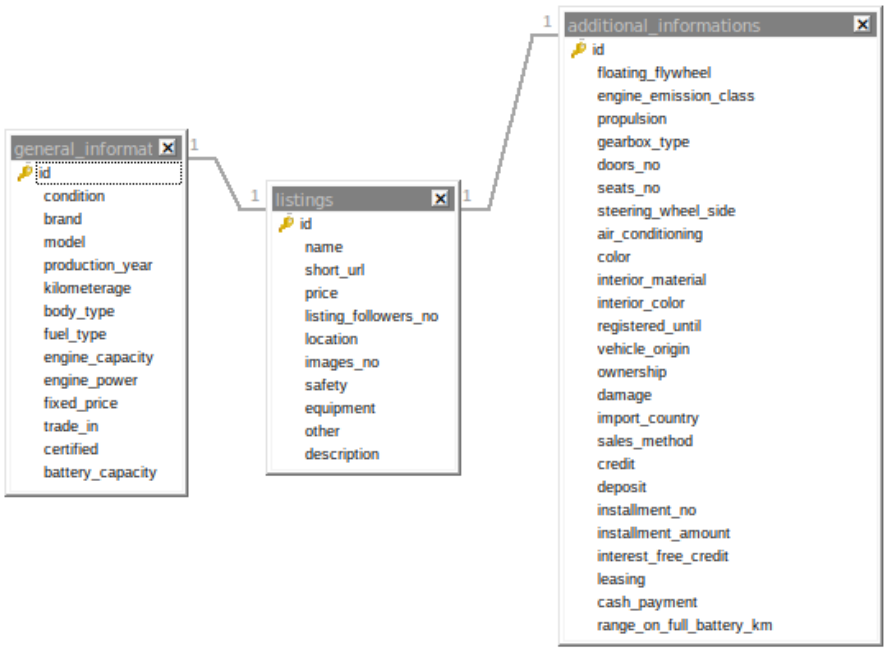
За остварење завршнога рада употребљене су софтверске технологије, оруђа и библиотеке пописане испод:

- *Linux Mint* 20.3 *Cinammon* оперативни стистем;
- *Python* 3.11.8 програмски језик;
- *Visual Studio Code* развојно окружење (IDE);
- *MySQL* систем за управљање базом података [СУБП];
- *SQLYog Ultimate* оруђе за развој и управљање *MySQL* СУБП-има;
- *SQLAlchemy ORM* (*Object Relational Mapper*) библиотека;
- *BeautifulSoup* за мрежно скрејпање;
- *Tor* мрежни прегледач за анонимно крстарење интернетом;
- *Stem* библиотека за рад са *Tor*-ом у програмском језику *Python*;
- *Selenium* за аутоматизовани рад са мрежним прегледачима;
- *tbselenium* као спона између *Selenium* библиотеке и *Tor* прегледача;
- *Pandas* за сређивање података;
- *NumPy* за операције над векторима и матрицама;
- *Matplotlib* и *Seaborn* за исцртавање графика и визуелизације;
- *Jupyter* за олакшано добијање увида из табела;
- *scikit-learn* за творење модела машинског учења и сродних ствари;
- *PyYaml* формат конфигурационих датотека;
- *Hydra* за управљање конфигурационим датотекама.

Софтверски код за израду завршног рада је доступан на *Github* платформи на вези https://github.com/aleksa-radojicic/second_hand_car_price_prediction.

3.3. ПРИКУП ПОДАТАКА

Подаци над којима је вршена анализа су повучени са *polovniautomobili* платформе помоћу *BeautifoulSoup* библиотеке за веб-скрејпање и преко *Tor* прегледача у фебруару 2024. године, а у обзир су улазила искључиво путничка возила. База података је створена помоћу *SQLAlchemy* библиотеке и *MySQL* системом за управљање базом података, а укупан број редова је био **30.788**, а стубаца **50**. Сви ступци су чувани као знаковни типови, са наканом подробног сређивања у одговарајући тип и чишћења приликом даље рашчлане. Слика 11 показује схему базе података створену у софтверу *SQLyog Ultimate*.



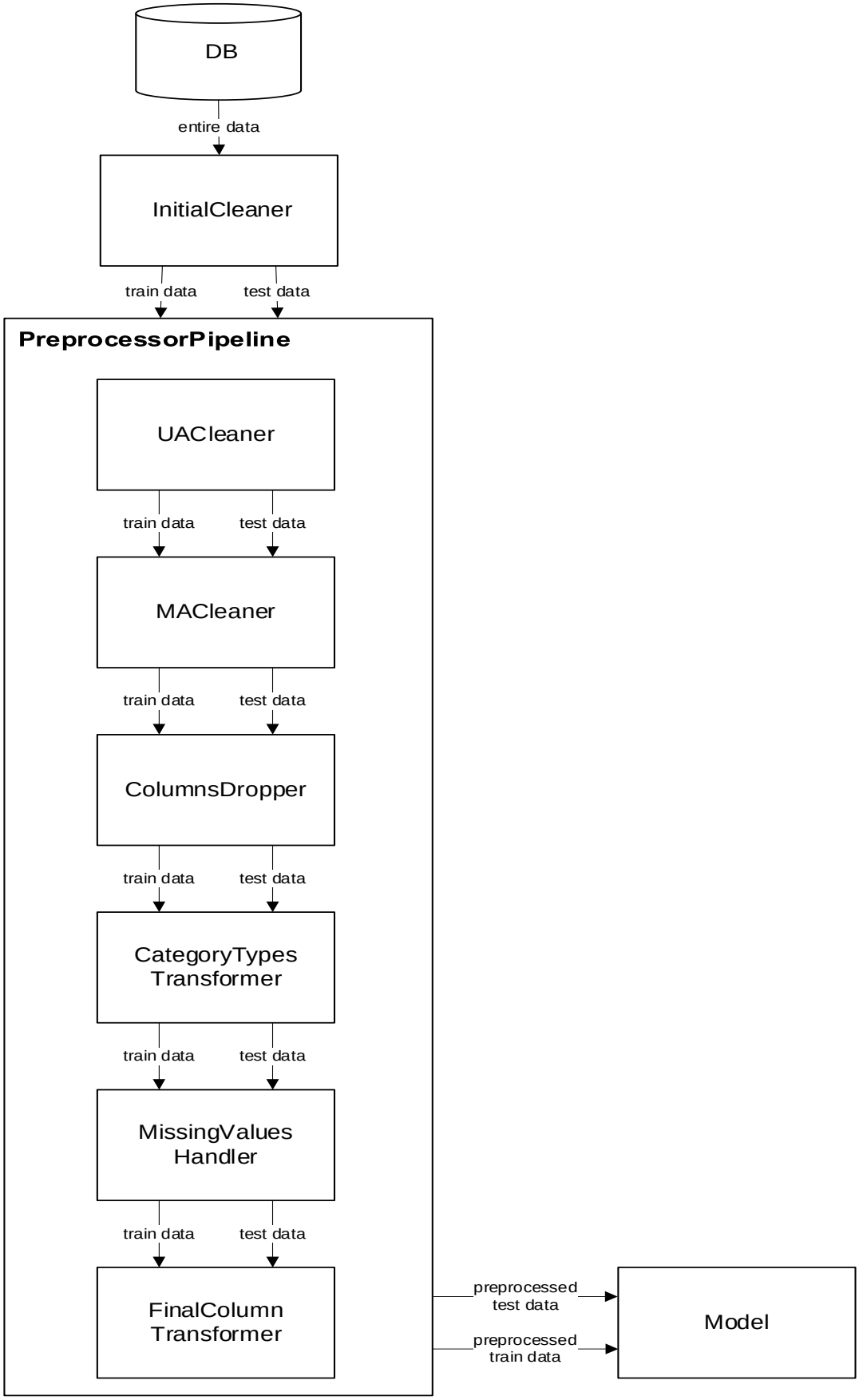
Слика 11. Релације базе података у MySQL-у

Повучене су следеће информације са сваког појединачног огласа са *polovniautomobili* странице: „Опште информације“, „Додатне информације“, „Сигурност“, „Опрема“, „Стање“, „Опис“, број пратиоца огласа, наслов, цена (ознака), број слика и место оглашавања. Треба напоменути да слике нису чуване, а стубац „опис“ је послужио само као метаподатак за дотично возило и није употребљен приликом моделовања. Није узета временска димензија, јер би имала више смисла да је би било доступних података кад је неко возило продато.

3.4. ПРЕДОБРАДА ПОДАТАКА

У овоме поглављу ће се говорити како је скуп података половних кола предобрађен (енгл. *preprocess*), значи све трансформације над ступцима и редовима матрице података које беху учињене. Биће дата схема претварача и структуре цевовода за предобраду података у програмском језику *Python*, који ће припремити податке непосредно пре сваког обучавања модела МУ.

Слика 12, то јесте дијаграм, показује како сирови подаци из базе података бивају сређивани кроз доста слојева да би на крају били улаз модела за машинско учење.



Слика 12. Склоп цевовода од базе података до модела

3.4.1. Почетни чистач

Почетни чистач (InitialCleaner) обрађује типове података свакојем ступцу и избацује кола која се нису уклапала на некоји начин са остатком узорка. Из стубаца који одговарају одељцима на интернет страници polovniautomobili „сигурност“, „стање“ и „опрема“ су извучена бинарна обележја која највише доприноше порасту почетног броја стубаца (редом Табела 1, Табела 2 и Табела 3). 18 је добијено из „сигурности“, 97 из „опреме“ и 13 стубаца из одељка „стање“.

Табела 1. Ступци из одељка „сигурност“

s_ABS	*s_ASR*	*s_Airbag_za_suvozača*
s_Airbag_za_vozača	*s_Alarm*	*s_Asistencija_praćenja_trake*
s_Automatsko_kočenje	*s_Blokada_motora*	*s_Bočni_airbag*
s_Centralno_zaključavanje	*s_Child_lock*	*s_ESP*
s_Kodiran_ključ	*s_Mehanička_zaštita*	*s_OBD_zaštita*
s_Senzor_mrtvog_ugla	*s_Ulazak_bez_ključa*	*s_Vazdušni_jastuci_za_kolena*

Табела 2. Ступци из одељка „стање“

o_Garancija	*o_Garažiran*	*o_Kupljen_nov_u_Srbiji*
o_Oldtimer	*o_Prilagođeno_invalidima*	*o_Prvi_vlasnik*
o_Restauriran	*o_Rezervni_ključ*	*o_Servisna_knjiga*
o_Taxi	*o_Test_vozilo*	*o_Tuning*
o_Vozilo_auto_škole		

Табела 3. Ступци из одељка „опрема“

e_360_ kamera	*e_AUX_ konekcija*	*e_Adaptivna_ svetla*	*e_Adaptivni_ tempomat*	*e_Aluminijumske_felne*	*e_Ambijentalno_osvetljenje*	*e_Android_Auto*
e_Apple_CarPlay	*e_Asistencija_za_kretanje_na_uzbrdici*	*e_Automatsko_parkiranje*	*e_Automatsko_zatamnjivanje_retrovizora*	*e_Autonomna_vožnja*	*e_Bluetooth*	*e_Branici_u_boji_auta*
e_Brisači_prednjih_farova	*e_CD_changer*	*e_DPF_filter*	*e_DVD_ili_TV*	*e_Daljinsko_zaključavanje*	*e_Digitalni_radio*	*e_Dnevna_svetla*
e_Držači_za_čaše	*e_Ekran_na_dodir*	*e_Električni_podizači*	*e_Električni_retrovizori*	*e_Elektro_otvaranje_prtljažnika*	*e_Elektro_podesiva_sedišta*	*e_Elektro_sklopivi_retrovizori*
e_Elektro_zatvaranje_prtljažnika	*e_Elektronska_ručna_kočnica*	*e_Fabrički_ugrađeno_dečije_sedište*	*e_Glasovne_komande*	*e_Grejanje_sedišta*	*e_Grejanje_volana*	*e_Grejači_retrovizora*
e_Grejači_vetrobranskog_stakla	*e_Hands_free*	*e_Hard_disk*	*e_Head_up_display*	*e_ISOFIX_sistem*	*e_Indikator_niskog_pritiska_u_gumama*	*e_Kamera*
e_Keramičke_kočnice	*e_Kožni_volan*	*e_Krovni_nosač*	*e_Kuka_za_vuču*	*e_LED_prednja_svetla*	*e_LED_zadnja_svetla*	*e_MP3*
e_Masažna_sedišta	*e_Matrix_farovi*	*e_Memorija_sedišta*	*e_Metalik_boja*	*e_Modovi_vožnje*	*e_Multifunkcionalni_volan*	*e_Multimedija*
e_Naslon_za_ruku	*e_Navigacija*	*e_Ostava_sa_hlađenjem*	*e_Otvor_za_skije*	*e_Paljenje_bez_ključa*	*e_Panorama_krov*	*e_Parking_senzori*
e_Podešavanje_volana_po_visini	*e_Postolje_za_bežično_punjenje_telefona*	*e_Prednja_noćna_kamera*	*e_Privlačenje_vrata_pri_zatvaranju*	*e_Putni_računar*	*e_Radio_CD*	*e_Radio_ili_Kasetofon*
e_Retrovizor_se_obara_pri_rikvercu	*e_Rezervni_točak*	*e_Ručice_za_menjanje_brzina_na_volanu*	*e_Sedišta_podesiva_po_visini*	*e_Senzori_za_kišu*	*e_Senzori_za_svetla*	*e_Servo_volan*
e_Sportska_sedišta	*e_Sportsko_vešanje*	*e_Start_stop_sistem*	*e_Subwoofer*	*e_Svetla_za_maglu*	*e_Tempomat*	*e_Tonirana_stakla*
e_Torba_za_skije	*e_USB*	*e_Upravljanje_na_sva_četiri_točka*	*e_Utičnica_od_12V*	*e_Vazdušno_vešanje*	*e_Ventilacija_sedišta*	*e_Virtuelna_tabla*
e_Volan_u_kombinaciji_drvo_ili_koža	*e_Webasto*	*e_Xenon_svetla*	*e_Zaključavanje_diferencijala*	*e_Zavesice_na_zadnjim_prozorima*	*e_Šiber*	

Следеће претворбе доведоше до смањења редова:

- Уклон нових кола (*gi_kilometerage* = 0);

  Избачено редова: 634.
- Уклон релативно нових кола (*gi_kilometerage* < 500);

  Избачено: 43.
- Брисање аутомобила без исправне вредности ознаке (возила која су се могла купити само на кредит);

  Избачено: 18.
- Уклон такси-возила (нетипични и мали узорак);

  Избачено: 16.
- Избачај олдтајмера са два седишта (нетипични);

  Избачено: 1.
- Избачај возила ауто-школе (нетипични и мали узорак);

  Избачено: 7.
- Уклон кола са бројем седишта већим од седам (мали узорак);

  Избачено: 21.
- Брисање редова који одговарају колима цене веће од 80.000€ (мали узорак);

  Избачено: 205.
- Брисање возила са воланом на десној страни.

  Избачено: 144.

У Почетном чистачу су уклоњени ступци који следе:

- *o_Taxi*;
- *o_Vozilo_auto_škole*;
- *gi_condition* („стање“);

  Обрисано је јер су само у жижи половним аутомобилима, а *gi_condition* је имало само категорије „Novo vozilo“ и „Polovno vozilo“.
- *ai_steering_wheel_side* („страна волана“);
- *ai_deposit* („учешће“);
- *ai_installment_no* („број рата“);
- *ai_installment_amount* („висина рате“);
- *ai_cash_payment* („готовинска уплата“).

Ступци „учешће“, „број рата“, „висина рате“ и „готовинска уплата“ су показале непосредни саоднос са цене возила зато што у себи садрже податке о њој. На пример, за кола која имају познате ове ступце важи: $ai\_deposit + ai\_installment\_amount \times ai\_installment\_no \geq price$. Да ове променљиве нису обрисане дошло би до цурења података, самим тим и невањаних налаза.

После уводног чишћења базе, исходни скуп података је имао димензије (**29.699**, **164**).

3.4.2. Цевовод за предобраду

Цевовод за предобраду (*PreprocessorPipeline*) је најважнија компонента за предобраду која се примењује над сваким скупом за обуку и тестирање који се моделују. Састављен је од наредних јединица:

1) **Чистач у једноваријабилној анализи** (*UACleaner*[3]);

   Обрађује појединачна обележја не узимајући у обзир међусобни утицај.

2) **Чистач у вишеваријабилној анализи** (*MACleaner*[4]);

   Обрађује обележја на основу њихова међусобна односа.

3) **Избацивач стубаца** (*ColumnsDropper*);

   Обележја која су се дојмила за брисање нису одмах уклањана већ су премештана у распоред за брисање, јер се могло догодити да се приликом даљег истраживања утврди да је ипак мудрије оставити их. Према томе, овај избацивач уклања из података ступце убачене у поменути распоред за брисање.

4) **Претварач категоријских типова** (*CategoryTypesTransformer*);

   Претвара категоријска обележја из знаковног типа у бројеве коришћењем *OrdinalEncoder* објекта. Неопходност ова корака лежи у чињеници да идући кораци захтевају да категорије не буду знаковног типа, као и сам *API scikit-learn* библиотеке.

[3] *UA* је скраћено од *UnivariateAnalysis*
[4] *MA* је скраћено од *MultivariateAnalysis*

5) **Сређивач недостајућих вредности** (*MissingValuesHandler*);

   Сређује недостајуће вредности свакојег обележја према унапред задатој стратегији. Доступне стратегије: попуна аритметичком средином, модусом, медијаном, константом -1 за знаковне, константом 0 за бројне и константом 0 за тачно / нетачно типове. Последње три стратегије су употребљене онда када су недостајуће вредности представљале засебне вредности.

6) **Коначни претварач стубаца** (*FinalColumnTransformer*).

   Помоћу *ColumnTransformer* објекта се свака различна врста обележја претвара на следећи начин:
   бројна обележја → остављају се каква јесу, *StandardScaler*, *RobustScaler* или *MinMaxScaler*;
   ординална категоријска обележја → *OrdinalEncoder*;
   бинарна категоријска обележја → *OrdinalEncoder*
   номинална категоријска обележја → *OrdinalEncoder* или *OneHotEncoder*.
   Особито претварање за дату врсту обележја представља хиперпараметар који се може оптимизовати.

Свака класа која чини јединствени корак Цевовода за предобраду наслеђује класу ***CustomTransformer*** која наслеђује класе из *scikit-learn* библиотеке неопходне да би обрадна класа могла сарађивати и бити у складу са прописима *scikit-learn API*-а (*TransformerMixin* и *BaseEstimator*). Главна намена класе *CustomTransformer* јесте да прошири функционалност претварача да се поред матрице података X и излазног ступца y користи и кориснички уведена структура ***Metadata*** (метаподаци) која поседује информације о типу податка свакојег ступца и стратегије за уклон изнимака за свакоји стубац.

Потреба за оваквим склопом лежи у чињеници да се коришћењем уграђених претварачких класа *TransformerMixin* и *BaseEstimator* избегава преприлагодба током одабира хиперпараметара у унакрсном вредновању, јер се никад неће догодити да подаци из скупа за тестирање некако исцуре у скуп за обуку.

3.5. УВИДИ ИЗ ЈЕДНОВАРИЈАБИЛНЕ И ВИШЕВАРИЈАБИЛНЕ АНАЛИЗЕ

Да би уопће било могуће скројити саставне делове цевовода неопходно је било користити *Jupyter* свеске, ради ласнијег налажења значајних информација и увида. Према томе, код је из *Jupyter* свески после прилагођен и претворен у одговарајуће *Python* датотеке, односно у *scikit-learn* претвараче (енгл. *transformers*). Нужан је овај корак да би се избегло цурење података, јер поједине информације из скупа за тестирање могу изцурети у модел приликом обуке, чинећи предвиђаје слабијим и онда је то незгода преприлагодба уколико се не користе претварачи.

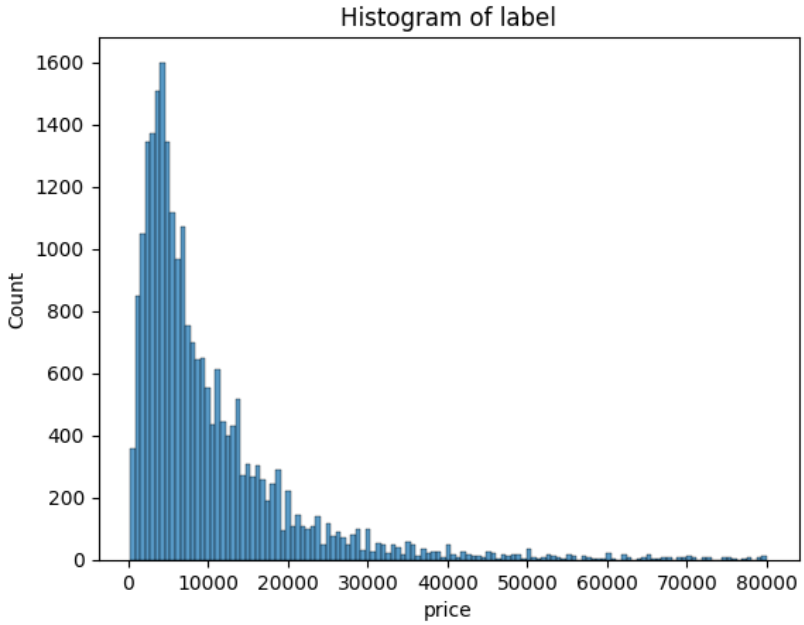
Важно је нагласити да су увиди рађени искључиво над скупом за обуку ради пристрасности, ваљаности и избегавања преприлагодбе, слично као у примеру о претварачима објашњеном изнад.

3.5.1. Информације о ступцима

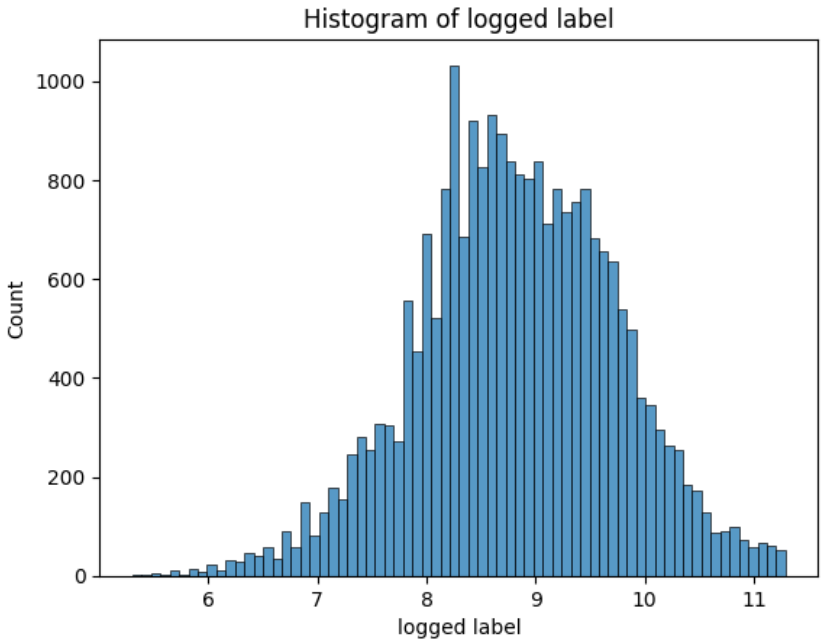
Хистограм за ознаку, односно променљиву „цена“ (Слика 13), јесте асиметричан у леву страну и уочава се да је највише возила усредсређено око цене од око 6.000€, што показује и медијана (Табела 4). Знатно је нижи број кола скупљих од 30.000€, што је и смислено јер је аритметичка средина већа од медијане. Најјефтинији ауто је имао цену од 200 евра, а најскупљи од 80.000€, при чему је раније речено да су у узорку остављана кола мање или једнаке цене од ове. Примењивањем трансформације логаритмовања ће ово обележје више наличити да има нормалну (Гаусову) расподелу (Слика 14).

Табела 4. Описне статистике ознаке

	фрекв.	mean	std	min	25%	50%	75%	max
„цена“	23759	10127,42944	10520,86329	200	3700	6600	12800	80000

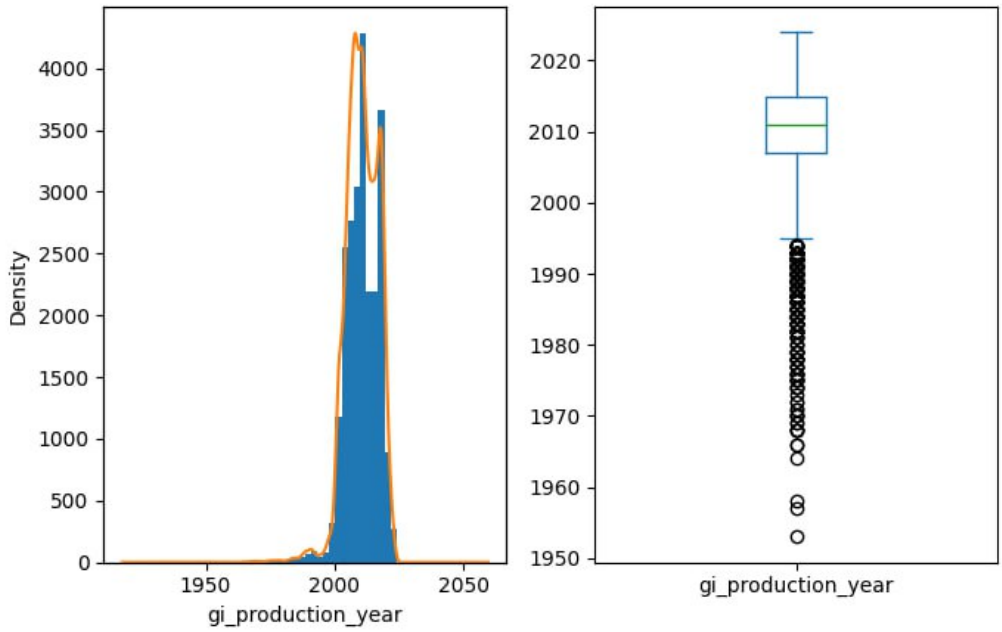


Слика 13. Хистограм ознаке



Слика 14. Хистограм логаритмоване ознаке

Већина половних аутомобила је из 21. столећа према табели описних статистика, тачније вредности 0,25. квантила (Табела 5), а хистограм са функцијом густине и кутијасти графикон (енгл. *boxplot*) (Слика 15) на то такође указују показујући их као изнимке. Расподела више личи на нормалну него што беше случај са ценом, а и мање је одступање аритметичке средине од медијане.



Слика 15. Хистограм са функцијом густине и кутијасти графикон ступца „година производње“

Табела 5. Описне статистике ступца „година производње“

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

Београд заузима прво место по заступљености локација продаваца половних кола, а затимм Нови Сад, Ниш, Крагујевац и Чачак (Табела 6).

Табела 6. Пет најучесталијих локација продаваца кола

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

Табела 7 указује на пет марки кола најзаступљенијих у пречишћеној матрици података скупа за обуку.

Табела 7. Пет најучесталијих марки кола

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

30% кола поседује вредност ступца „емисиона класа мотора“ поседује вредност четири, што износи 7140 редова табеле, а 0,9% има вредност 1 (Табела 8). Само једна кола су имала недостајућу вредност која је замењена модусом овог обележја, односно вредношћу 4.

Табела 8. Учесталост емисионих класа мотора

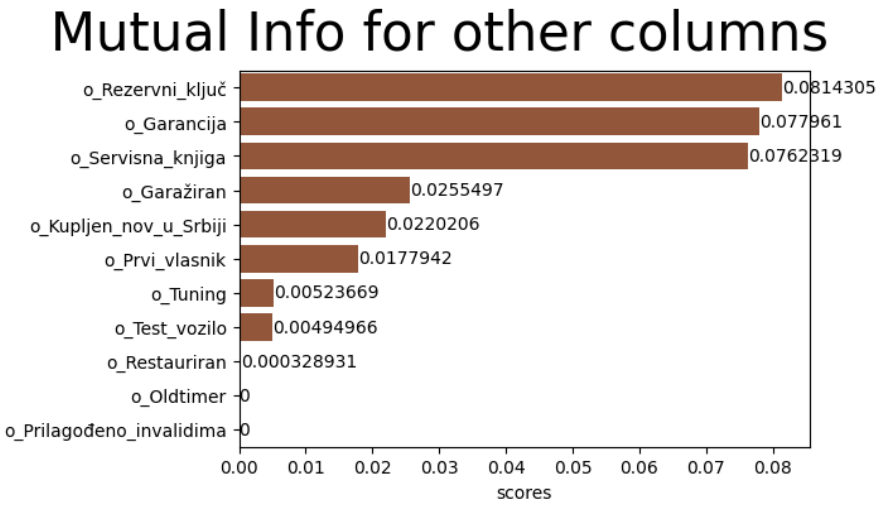
	фрекв.	удео [%]
„емисиона класа мотора“		
4	7140	30,051770
6	6911	29,087925
5	6632	27,913633
3	2559	10,770655
2	304	1,279515
1	212	0,892293
NaN	1	0,004209

Табела 9 даје до знања да је већина продаваца са платформе *polovniautomobili* била поправљала своја кола пре него што је окачила оглас, јер их је свега 1,24% оштећено. Две су недостајуће вредности које беху попуњене модусом.

Табела 9. Учесталост врсти оштећења кола

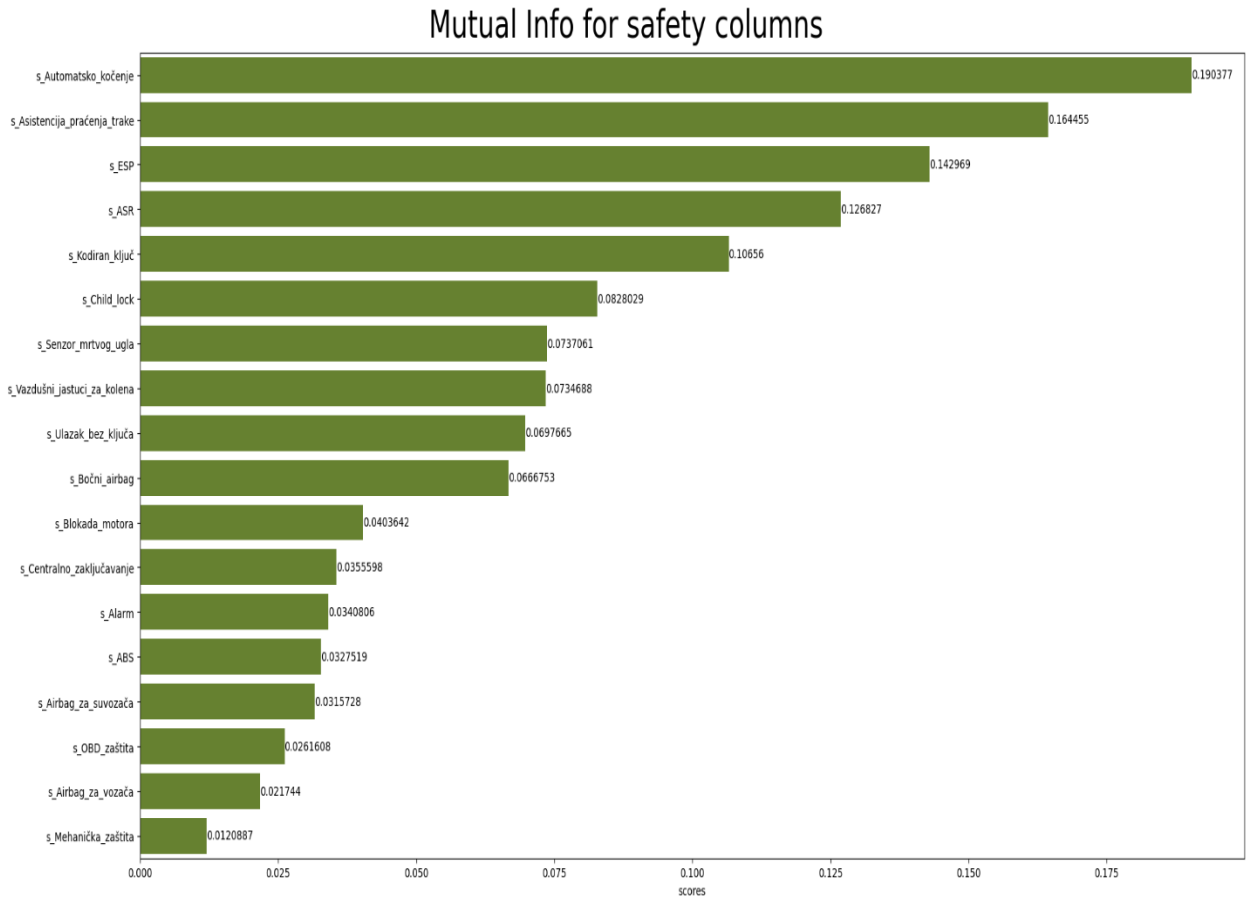
	фрекв.	удео [%]
„оштећење“		
Није оштећен	23465	98,762574
Оштећен – у возном стању	176	0,740772
Оштећен – није у возном стању	116	0,488236
NaN	2	0,008418

3.5.2. Значајност стубаца пре обуке



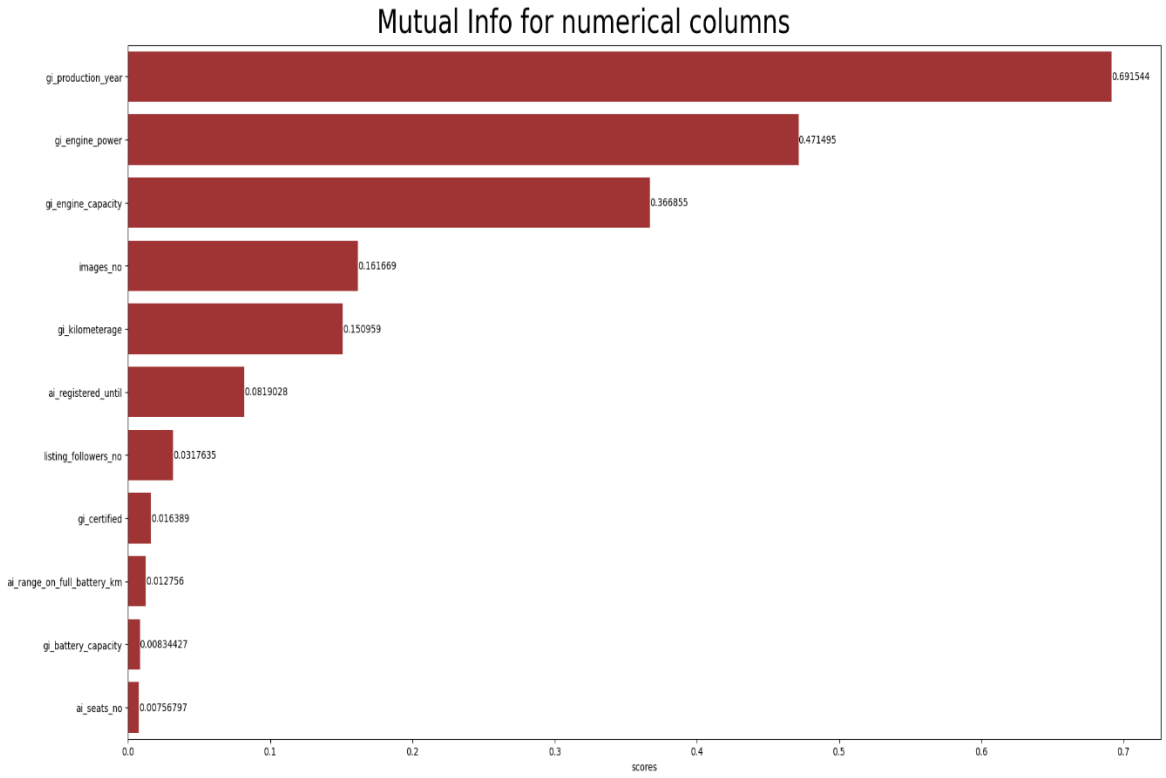
Слика 16. Узајамне информације између стубаца из одељка „стање“ и ознаке

Од стубаца из одељка „стање“ најупотребљивије су се, према узајамним информацијама (види *mutual_info_regression* из *scikit-learn* библиотеке), показали *o_Rezervni_ključ*, *o_Garancija* и *o_Servisna_knjiga* (Слика 16). Најгору везу са излазном променљивом имају *o_Prilagođeno_invalidima*, *o_Oldtimer* и *o_Restauriran*, из простога разлога зато што им је премала варијанса (нимали или мали број кола има ове особитости.



Слика 17. Узајамне информације између стубаца из одељка „сигурност“ и ознаке

Заједничке информације међу бројним ступцима и излазом се виде испод (Слика 18). Најјачу везу са излазном променљиве има обележје „година производње“ (*gi_production_year*) и то скоро 0,70. „Снага мотора“ и „кубикажа“ износе редом 0,472 и 0,367, што такође указује да су ови ступци важни за предвиђање цене половних кола и онда ће јамачно доспети у корак моделовања.



Слика 18. Узајамне информације између бројних обележаја и ознаке

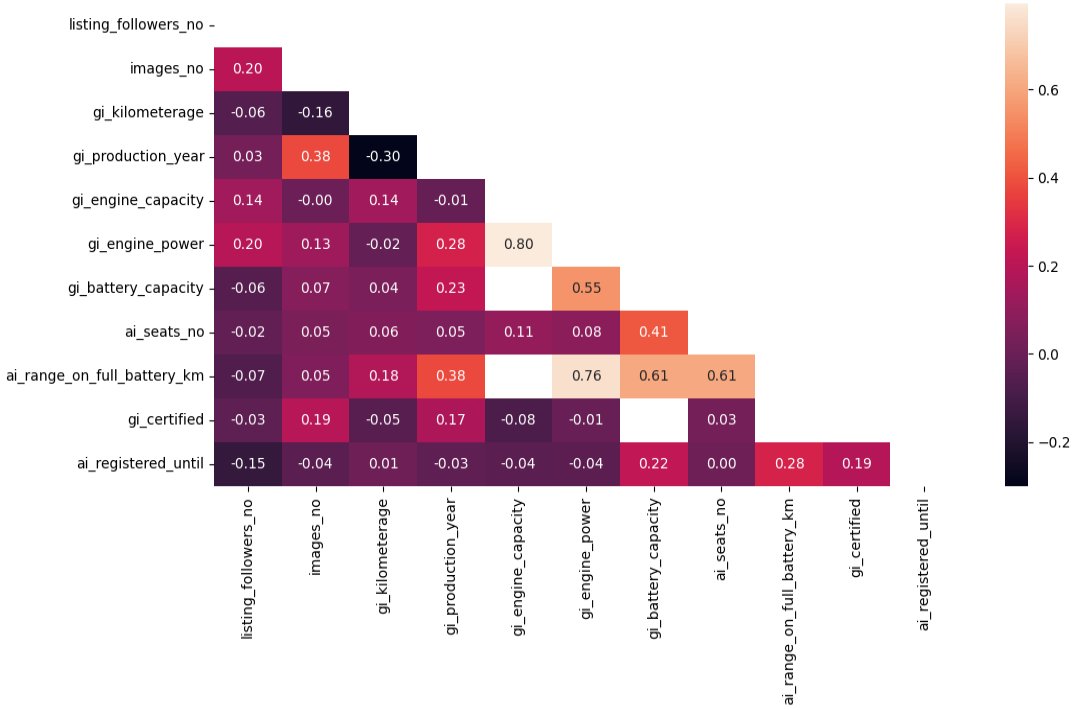
На основу увида из једноваријабилне и вишеваријабилне анализе, а понајвише на основу значајности обележаја, компонента Избацивач стубаца је обрисала следећа обележја из распореда за брисање:

- *e_Fabrički_ugrađeno_dečije_sedište* (ниска значајност);
- *e_Volan_u_kombinaciji_drvo_ili_koža* (ниска значајност);
- *o_Oldtimer*;
- *o_Prilagođeno_invalidima*;
- *o_Restauriran*;
- *o_Test_vozilo*;
- *o_Tuning*.

Ступци из одељка „опрема“ који су уклоњени показаше најнижу узајамну информацију за излазном променљивом „цена“ (Слика 16), што има смисла зато што је број *True* вредности за њих посве мали.

3.5.4. Пирсонови саодноси

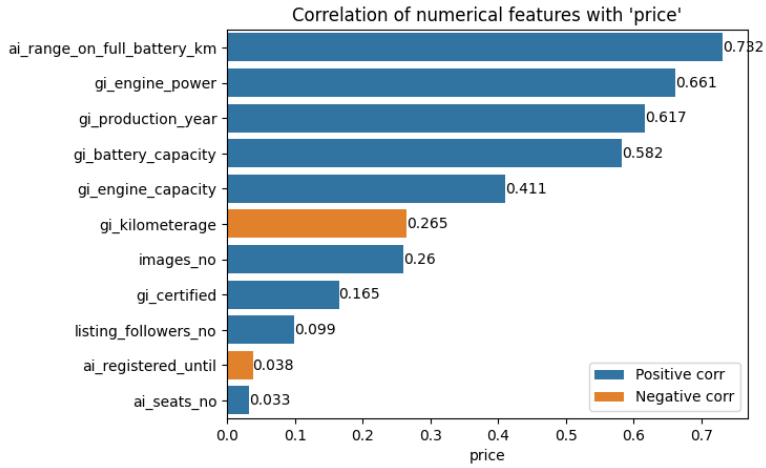
У продужетку је дат **Пирсонов саодност (корелација)**, као мера линеарне зависности бројних обележаја (Вуковић, 2012), изузимајући цену:



Слика 19. Пирсонов саоднос између бројних обележаја без ознаке

„Снага мотора“ (*gi_engine_power*) и „кубикажа“ (*gi_engine_capacity*) показују највећи апсолутни саоднос од свих бројних обележаја (чак +0,80). „Домет са пуном батеријом (km)” (*ai_range_on_full_battery_km*) и „снага мотора“ исто показују снажан саоднос и то +0,76 (Ratner, 2009). Ипак, нужно је истаћи да 99,6% возила има недостајућу вредност домета са пуном батеријом и онда је посве мали узорак ушао приликом израчунавања Пирсоновог саодноса, те је стварна вредност овога броја заправо мања.

„Капацитет батерије“ и „кубикажа” (7. ред, 5. стубац), као и „кубикажа“ и „домет са пуном батеријом (km)” (3. ред од доле, 7. стубац) имају непостојећи саоднос јер само електрична кола имају позитивне вредности за „капацитет батерије“ и „домет са пуном батеријом (km)”, док се кубикажа односи само на неелектрична возила.



Слика 20. Пирсонов саоднос бројних обележаја са ознаком

Од бројних обележаја највећи износ Пирсоновог саодноса са излазном променљивом јесте са капацитетом батерије (+0,707), потом са кубикажом мотора (+0,7) и годином производње (+0,532) (Слика 20). Уочава се да број седишта не може линеарно објаснити цену кола, као ни датум трајања регистрације (исказан као разлика у месецима између текућег датума када је вршена анализа и датума истека регистрације).

4. РЕЗУЛТАТИ И РАСПРАВА

Сврха овог рада је показати како ће се модели машинског учења показати приликом процене цене половних аутомобила са интернет странице *polovniautomobili*. Процениће се неколико познатих модела МУ са подразумеваним хиперпараметрима над скупом за обуку и обучити они за које се утврди да могу добро моделовати улазне податке у излаз. За GBM модел ће се наћи интервали предвиђања како би пружили додатну важну информацију о излазној вредности и њеној поузданости, што ће имати још више смисла уколико се GBM најбоље покаже. Главна метрика по којој ће се упоређивати модели јесте R2.

4.1. МОДЕЛИ СА ПОДРАЗУМЕВАНИМ ХИПЕРПАРАМЕТРИМА

Скуп за тестирање износио је 20% насумично изабраних аутомобила из полусређене табеле која је била излаз из Почетног чистача, што значи да је **23.759** кола уврштено у обуку, а **5.940** је служило за тестирање. Излазна променљива је, као што поменуто, цена возила (*price*) у противвредности евра. Предобрађени скупови за обуку и тестирање имали су **153** стубаца, изузимајући излазну променљиву, при чему су искључена обележја која послужише као метаподаци: *name* (назив огласа), *short_url* (URL аутомобила) и *description* („опис“).

Обучено је више модела машинског учења са подразумеваним хиперпараметрима над скупом за обуку, уз додавање параметра за постизање истоветних налаза (енгл. *random seed*) у вредности 2024. Такође су се обучавали и модели градијентног појачавања за долње (0,05) и горње (0,95) вредности квантила како би се нашао интервал предвиђања. Сваки пут су се полусређени подаци пуштали у Цевовод за предобраду кад се различити модел обучавао. Хиперпараметри Цевовода за предобраду су овако подешени:

- За моделе засноване на дрвима:
  - final_ct__numerical_encoder=”passthrough”
  - final_ct__nominal_encoder=”ordinal”
- За моделе засноване на недрвима:
  - final_ct__numerical_encoder=”standardscaler”
  - final_ct__nominal_encoder=”onehot”

У наставку су дате перформансе модела са почетним и неоптимизованим хиперпараметрима над подацима за тестирање (Табела 10), а испод над подацима за обуку (Табела 11). RMSE и MAE вредности су заокругљене на три, а R2 на пет децимала.

Табела 10. Перформансе модела са подразумеваним хиперпараметрима модела над скупом за тестирање

	Метрике над скупом за тестирање		
Модел	RMSE	MAE	R2
Dummy median[5]	11063,807	6339,458	-0,10966
Dummy mean[6]	10503,007	6979,143	0
Ridge регресија	4340,172	2626,473	0,82924
KNN[7]	3659,045	1906,868	0,87863
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

Табела 11. Перформансе RF, GBM $Q_{0,50}$, GBM $Q_{0,05}$ и GBM $Q_{0,95}$ са подразумеваним хиперпараметрима модела над скупом за обуку

	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

[5] Модел који предвиђа медијану излазне променљиве скупа за обуку.
[6] Модел који предвиђа аритметичку средину излазне променљиве скупа за обуку.
[7] Скраћеница за модел *K-Nearest Neighbours* (К најближих суседа).

Гледано према свим регресивним метрикама су се зацело најбоље показали ансамбл модели над непознатим подацима који су од ових најсложенији, поготово случајна шума која обухвата успешно 92,960% укупног варијабилитета. У свим моделима је већи износ RMSE у односу на MAE, па се из тога може извести закључак да највероватније постоје изнимци у скупу података и њихови предвиђаји много одступају од стварних и онда више повећавају RMSE. Апсолутна грешка код RF се тумачи овако: у просеку модел прецењује или потцењује кола у вредности од 1364,816 евра, што је боље када је цена кола велика. На пример, ако полован аутомобил кошта 20.000€ онда релативна грешка износи 6,82% укупне цене и тада би стварна цена требала припадати интервалу [18.635,184€; 21.364,816€], мада може и излазити и ван граница одсечка.

Пошто је познато да су ансамбл модели машинског учења подоста делотворни у пракси, биће оптимизовани хиперпараметри за њих, тачније за RF и GBM $Q_{0,05}$ као и за GBM $Q_{0,05}$ и $Q_{0,05}$ за налажење подеснијег интервала предвиђаја.

GBM модели за интервале предвиђаја показују слабе износе коефицијента детерминације, особито модел градијентног појачавања за 0,05. квантил (R2 = 0,58134). Оба модела показују потприлагодбу, што је у овоме тренутку битна информација која ће олакшати одабир опсега за оптимизацију хиперпараметара. На пример, за GBM $Q_{0,05}$ ће се покушати са доста већим вредностима за број процењивача и дубину појединачног дрвета него за GBM $Q_{0,95}$.

Кад се упореде метрике над скупом за обуку RF види се да се модел доста преприлагодио подацима за обуку (Табела 11), а GBM има много уравнотеженије метрике за обуку и тестирање (уз благу преприлагодбу).

4.2. МОДЕЛИ СА ОПТИМИЗОВАНИМ ХИПЕРПАРАМЕТРИМА

За случајну шуму и модел градијентног појачавања је рађена оптимизација хиперпараметара коришћењем решеткасте претраге у пет преклопа, такорећи 80% података за обуку (19.007), а остатак за вредновање (4752 редова). Као метрика за оптимизацију је узет R2.

Излазна променљива је логаритмована како би била ближа нормалној расподели. Остављено је 28 обележаја које показаше највећу значајност над обученом случајном шумом са подразумеваним хиперпараметрима.

4.2.1. Оптимални хиперпараметри за RF

Најпре је за случајну шуму рађена прва решеткаста претрага за комбинације хиперпараметара *n_estimators* и *min_samples_split*, са већим кораком, односно размаком између појединачних вредности (Табела 12). На крају се пустила друга решеткаста претрага за исте хиперпараметре али вредности са мањим кораком како би се нашли још повољнији хиперпараметри (Табела 13). Ради прегледности приказује се и Табела 14 са крајњим хиперпараметрима.

Табела 12. Прва решеткаста претрага за RF и налази

Хиперпараметри	Простор претраге	Оптималне вредности
n_estimators	[70; 100; 130; 160; 200; 230; 260; 300; 330; 360; 400]	245
min_samples_split	[2; 5; 8; 11; 14; 17; 20; 23; 26]	2 (подразумевана)

Табела 13. Друга решеткаста претрага за RF и налази

Хиперпараметри	Простор претраге	Оптималне вредности
n_estimators	[200; 205; 210; ...; 295; 300]	245
min_samples_split	[2; 3; 4; 5; 6; 7; 8; 9]	2

Табела 14. Оптимални хиперпараметри за RF

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

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

За модел градијентног појачавања $Q_{0,50}$ су извршене две решеткасте претраге једна за другом. Прва користи хиперпараметре *learning_rate, n_estimators, min_samples_split* и *max_depth* (Табела 15), а друга *criterion* (Табела 16).

Табела 17 даје податке о коначним хиперпараметрима који су дали најбољу вредност коефицијента детерминације.

Табела 15. Прва решеткаста претрага за GBM $Q_{0,50}$ и налази

Хиперпараметри	Простор претраге	Оптималне вредности
learning_rate	[1; 0,75; 0,5; 0,25; 0,1; 0,05; 0,01; 0,005; 0,001]	0,05
n_estimators	[50; 100; …; 450; 500]	500
min_samples_split	[2; 5; 10; 15; 20]	15
max_depth	[3; 6; 9; 12]	12

Табела 16. Друга решеткаста претрага за GBM $Q_{0,50}$ и налази

Хиперпараметри	Простор претраге	Оптималне вредности
criterion	[„friedman_mse“; „squared_error”]	„friedman_mse“ (подразумевана)

Табела 17. 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“**

4.2.3. Оптимални хиперпараметри за GBM $Q_{0,05}$

Поступак добијања хиперпараметара за GBM $Q_{0,05}$ је текао у два корака, слично као код GBM $Q_{0,50}$. Утврђено је да је потребно направити сложенији модел јер је грешка на тренингу била ниска на подразумеваној вредности хиперпараметра од 100 дрвећа, па је долња граница постављена на 600, а горња на 2000 процењивача. У продужетку су налази двеју претрага (редом Табела 18 и Табела 19) и оптимални хиперпараметри (Табела 20).

Табела 18. Прва решеткаста претрага за GBM $Q_{0,05}$ и налази

Хиперпараметри	Простор претраге	Оптималне вредности
learning_rate	[0,25; 0,1; 0,01; 0,001]	0,1
n_estimators	[600; 900; … ; 1700; 2000]	2000
max_depth	[4; 8; 12; 16; 20; 24; 28]	28

Табела 19. Друга решеткаста претрага за GBM $Q_{0,05}$ и налази

Хиперпараметри	Простор претраге	Оптималне вредности
criterion	[„friedman_mse“; „squared_error”]	„friedman_mse“

Табела 20. Оптимални хиперпараметри за GBM $Q_{0,05}$

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

4.2.4. Оптимални хиперпараметри за GBM $Q_{0,95}$

Табела 21 даје приказ прве решеткасте претраге за модел градијентног појачавања за 0,95. квантил која ће представљати горње границе интервала предвиђаја. Друга Решеткаста претрага проверава који је бољи критеријум за гранање (Табела 22). Табела 23 даје обједињене налазе двеју претрага, то јесте оптималан спој хиперпараметара за GBM $Q_{0,95}$.

Табела 21. Прва решеткаста претрага за GBM $Q_{0,95}$ и налази

Хиперпараметри	Простор претраге	Оптималне вредности
n_estimators	[400; 500; … ; 900; 1000]	1000
min_samples_split	[2; 5; 10; 20]	5
max_depth	[4; 8; 12; 16; 20; 24; 28]	28

Табела 22. Друга решеткаста претрага за GBM $Q_{0,95}$ и налази

Хиперпараметри	Простор претраге	Оптималне вредности
criterion	[„friedman_mse“; „squared_error”]	„friedman_mse“

Табела 23. Оптимални хиперпараметри за GBM $Q_{0,95}$

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

4.3. ПРОЦЕНА ОПТИМИЗОВАНИХ МОДЕЛА

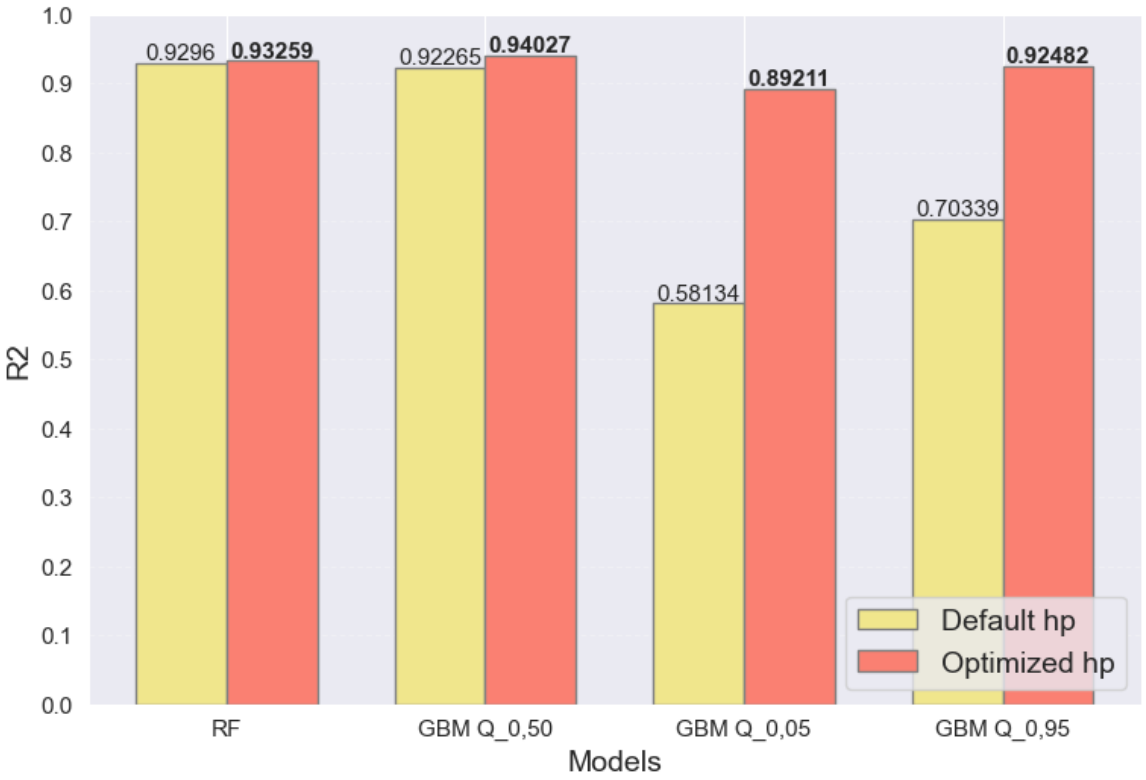
Да би се проценили модели RF, GBM са 0,05., 0,50. и 0,95. квантилом за које су нађене најбоље вредности хиперпараметара потребно је обучити моделе над целокупним подацима за обуку и израчунати метрике за процену над скупом за тестирање. У табелама које предстоје дате су перформансе модела над скупом за тестирање и над скупом за обуку (Табела 24, Табела 25), као и стубасти графикон поређења перформанси подразумеваних и неоптимизованих модела за R2 метрику (Слика 21).

Табела 24. Перформансе модела са оптимизованим хиперпараметрима над скупом за тестирање

	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

Табела 25. Перформансе модела са оптимизованим хиперпараметрима над скупом за обуку

	Метрике над скупом за обуку		
Модел	RMSE	MAE	R2
RF	**1191,538**	**499,685**	**0,98717**
GBM $Q_{0,50}$	1579,571	547,138	0,97746
GBM $Q_{0,05}$	2470,072	974,428	0,94488
GBM $Q_{0,95}$	1435,852	818,770	0,98137

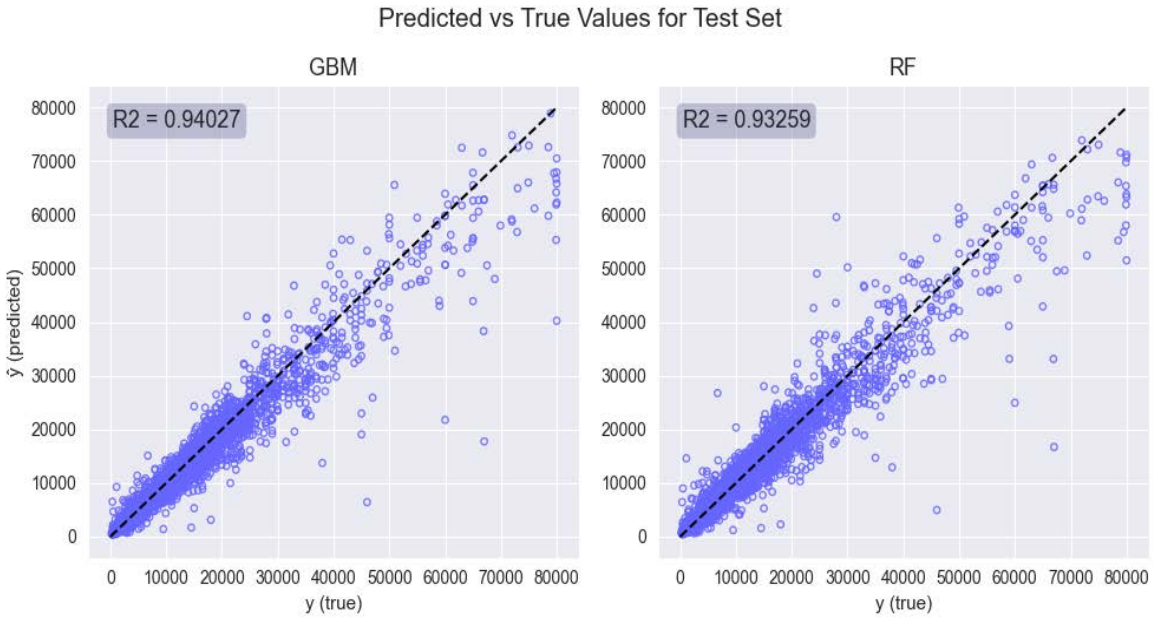


Слика 21. Поређење резултата модела са подразумеваним и оптимизованим хиперпараметрима над скупом за тестирање

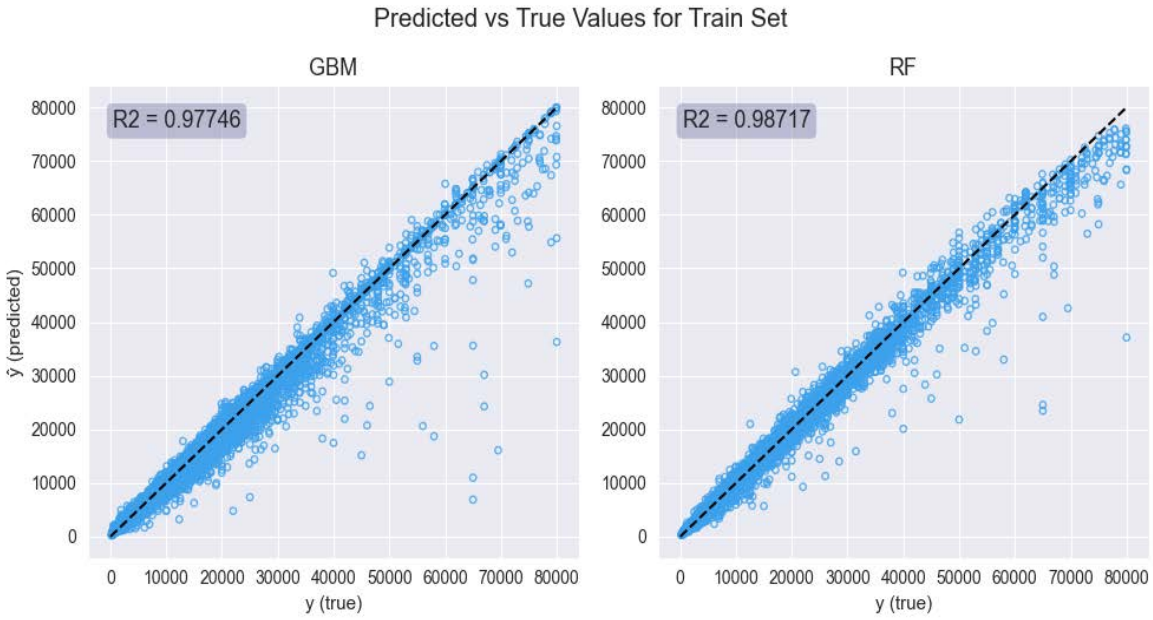
Повећај коефицијента детерминације случајне шуме је ситан у односу на почетни модел. Са друге стране, модел градијентног појачава са 0,50. перцентилом показује убедљиво најбоље резултате, побољшавши R2, MAE и RMSE у односу на неоптимизовани модел. GBM $Q_{0,50}$ се мање преприлагодио подацима него RF јер му је мањи износ R2 над подацима за обуку него што је то случај код случајне шуме.

GBM са долњим квантилом највише је повећао R2 од свих модела, што се одражава на прецизније одређивање долње границе интервала за предвиђаје. Неопходно је било доста усложнити модел постављајући број процењивача на 2.000 и користити све доступних ступце да би боље успео мапирати улазне податке у излазе, што је уследило доста бољем налазом коефицијента детерминације. GBM за горњи квантил исто је поприлично повећао R2 и то за 31%, чинећи горњу границу предвиђаја ужом и тиме употребљивијом.

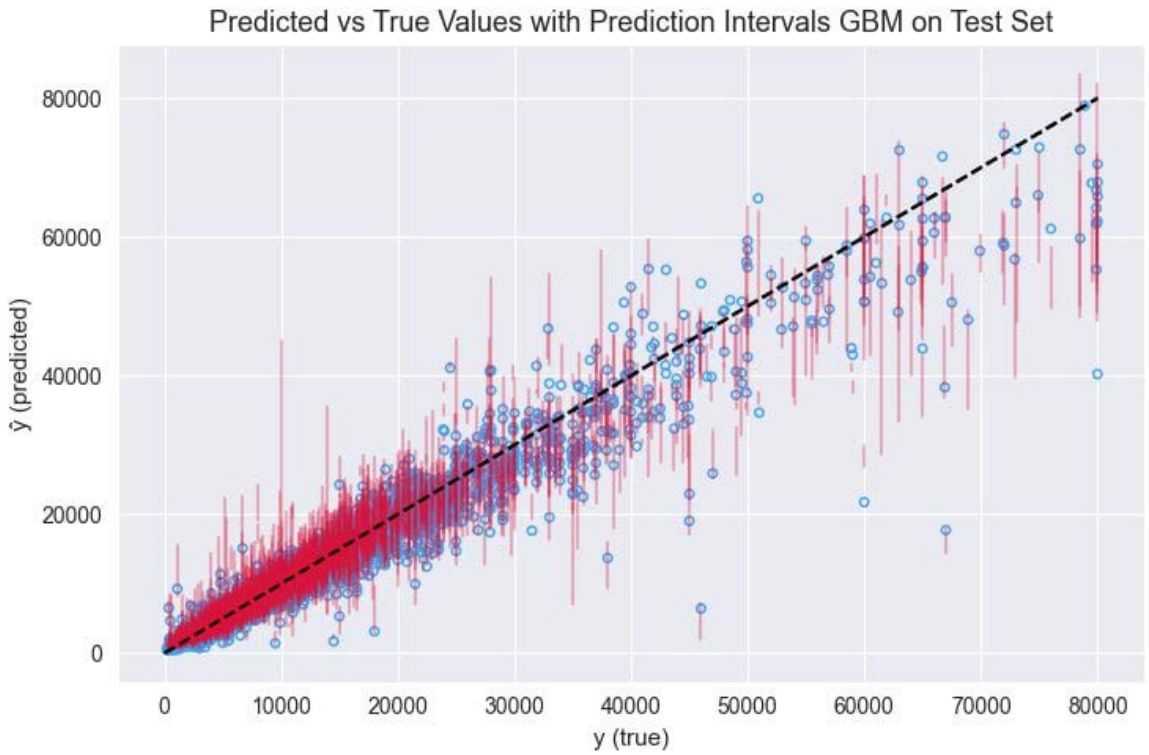
У наставку дати су дати графици предвиђених и стварних вредности случајне шуме и модела градијентног појачавања над скупом за тестирање (Слика 22), затим над скупом за обуку (Слика 23), а Слика 24 представља график предвиђених и стварних излаза над скупом за тестирање са посебним осврtом на интервале предвиђаја само за GBM $Q_{0,50}$.



Слика 22. График везе предвиђених и стварних излаза за оптимизоване GBM и RF над скупом за тестирање



Слика 23. График везе предвиђених и стварних излаза за оптимизоване GBM и RF над скупом за обуку



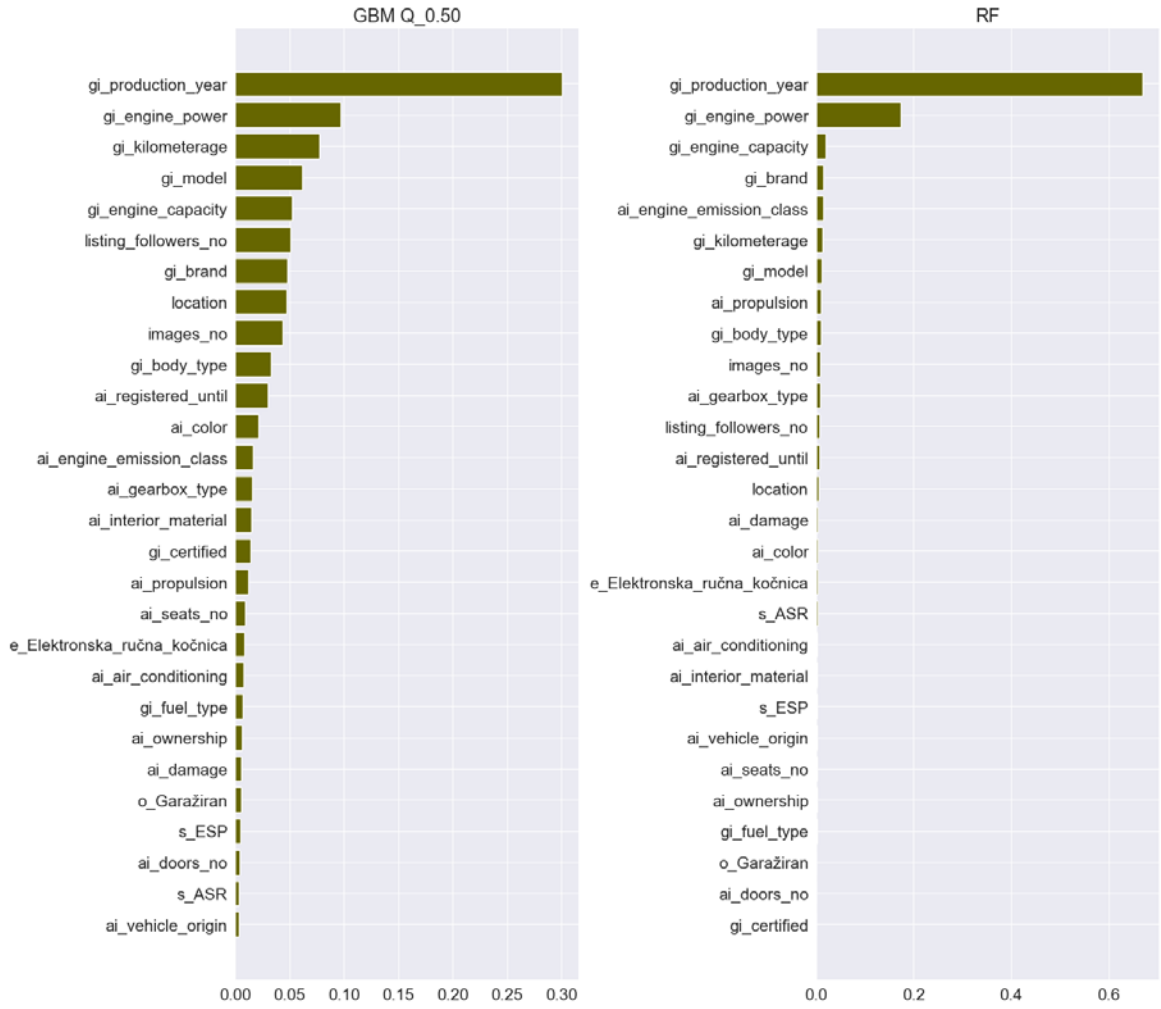
Слика 24. График везе предвиђених и стварних излаза за оптимизовани GBM над скупом за тестирање са интервалима предвиђаја

Могу се установити да има немали број кола која имају велику апсолутну разлику између $\hat{y}$ и y, како над подацима за тестирање тако и скупом за обуку. Код RF је мање таквих случаја, али треба имати на уму да модел лошије уопћава у односу на GBM. Ово указује да модели не могу успешно пресликати улазне податке у цену за кола веће цене.

Табела 26 даје увид у постотак обухваћености GBM обученим на редом долњом и горњом вредношћу квантила, као и то колики је удео слогова који припадају интервалу предвиђаја над скупом за обуку и тестирање. Уочено је са поменуте табеле да је нижи број слогова чије је y мање од 0,05. квантила него број слогова чији је 0,95. квантил већи од y. Скуп за обуку приближно обухвата 90% редова матрице података, док су перформансе на скупу за тестирање доста слабе и износе 43,38%, чинећи интервале предвиђаја недовољно поузданим за практичну примену, што се закључује и из графика везе предвиђаја и стварних излаза за интервале предвиђаја GBM-а (Слика 24).

Табела 26. Покривеност стварних излаза интервалима предвиђаја [%]

	Скуп за обуку	Скуп за тестирање
GBM $Q_{0,05}$	94.40633	**68,97306**
GBM $Q_{0,95}$	94.54944	**73.55219**
Укупно	89.056778	**43.38384**



Слика 25. Значајност стубаца за оптимизовани GBM $Q_{0,50}$ и RF

На одлуку који ће бити износ цене половних кола највише је утицало обележје „година производње“ за оба модела (Слика 25), мада у знатној мери више код RF него код GBM-а. Према Пирсоновом саодносу најснажнију је везу са ознаком имала „снага мотора“, а на другом место стубац „година производње“, док код оба модела су рангови замењени, то јесте „снага мотора“ је мање битна. Иако су „снага мотора“ и „кубикажа“ показала значајан саоднос, ипак је „снага мотора“ однела превласт над

значајношћу. Модел градијентног појачавања знатно боље користи информације из свих стубаца за предвиђај него случајна шума, јер сва обележја испод „снаге мотора“ код RF имају изразито ниску важност и висину стуба. Ступци „марка“, „модел“, и „локација“ захтевају подробније чишћење и сређивање како би се показала кориснијим за модел, те уопће побољшала моделе.

Што је колима већа година производње и снага мотора, то им је цена виша у просеку. Разумно је да ће се очекивати да кола са вишом километражом коштају просечно више јер то значи да су дуже рабљена и да је већа извесност да се неки део поквари.

4.4. РАСПРАВА

Кола са идентификационим кључем 22491542 су имала најшири интервал предвиђања [5.525€; 44.977€], са предсказаном вредношћу 8.386€ и 10.000€ стварним излазом. У питању је Opel Rekord аутомобил из 1961, који је и даље присутан био на сајту *polovniautomobili* (20. априла) и његова цена је тада била привремено спуштена на 6.500€. Разлог разлике између предвиђене и стварне цене више од 10.000 и огромног интервала јесте у великој мери због тога што је било трунка половних аута старијих од 1980. године (свега 51 кола у скупу за обуку, а 8 у скупу за тестирање). Уколико би се у тесту за тестирање узела у обзир само она кола годишта 1990. или млађа, онда би R2 порастао на 0,94258 за GBM $Q_{0,50}$, побољшавши и остале метрике за процену.

Слог са највећом разликом од чак 49.254 евра између стварне (67.000) и предвиђене цене (17.746) био је Mercedes Benz G 350. То је дизел џип из 2012., са 155 kW / 211 коњских снага, са 2987 кубика и пет седишта. У овом случају се слабе предвиђавне перформансе могу протумачити тако да је мањи узорак возила скупљих од 60.000€, па модел није имао довољно података на основу којих би разлучио њихова својства и тако боље моделовао цену.

Треба свакако узети у обзир да су поједини продавци кола могли намерно ставити прецењену цену за кола у огласу, што онда прави искривљенију слику приликом уопћавања за нова кола. Мада, уколико су за исти модел сва кола на сајту прецењена у односу на цену новог аутомобила, онда ће корисник који објави оглас са мањом просечном ценом за тај модел бити најисплативији на платформи *polovniautomobili*. Питање је да ли ће се сви ти аутомобили продати по оној цени исказаној на огласу или мањој, што се не може знати на основу података са мрежне странице, и то значи да би сам податак о успешној продајној цени много допринео моделу.

5. ЗАКЉУЧАК

Над сировим подацима сакупљених са мрежне странице *polovniautomobili* је спроведено основно чишћење и обрада стубаца и редова, што је исходило полуприпемљеним скупом за обуку и тестирање. Обучени су модели МУ, као и GBM $Q_{0,05}$ и GBM $Q_{0,85}$ за интервале предвиђања, са подразумеваним хиперпараметрима над скупом за обуку и процењени над скупом за тестирање. Одабрани су оптимални хиперпараметри кроз више унакрсних вредновања за RF, GBM $Q_{0,50}$, GBM $Q_{0,05}$ и GBM $Q_{0,95}$, над искључиво скупом за обуку. Потом су обучени модели са оптималним хиперпараметрима над скупом за обуку и процењени над скупом за тестирање. На крају се тумачило која обележја највише утицаху на одлуку модела да додели половним колима одређену цену.

Модел градијентног појачавања са 0,50. квантилом се испоставио бољим избором него случајна шума, због јачег пораста главне метрике R2, а притом је мање рачунарских ресурса захтевао за обуку и удешавање хиперпараметара. Уз то сам GBM нуди и могућност налажења интервала превдиђаја, што би додатно помогло крајњим корисницима модела. Добијени модели са оптимизованим хиперпараметрима за GBM $Q_{0,05}$ и GBM $Q_{0,95}$ нису успели начинити поуздан и употребљив интервал предвиђаја због ниске покривености над подацима за тестирање.

Један од корака који би било мудро спровести као идући степеник истраживања јесте пронаћи изнимке помоћу различитих метода и тестирати моделе обучене са и без тих изнимака. Могло би се – поред осталих стубаца – укључити и текстуално обележје „опис“ за предвиђање цене аутомобила који би захтевао посебну обраду и пажњу. Осим табеларних података свакако би се приликом мрежног скрејпања нових половних кола могле преузимати и слике које би такође биле део склопа за предвиђање. Боља употребљивост саме матрице података уследила би и побољшању одсечака предвиђаја, то јесте повећају прецизности (ужи интервал) и порасту покривености. Не би одмогло ни испробавање просечне апсолутне постотне грешке (енгл. *mean absolute percentage error*) као метрике за праћење, као и анализа оних кола који имају велики износ ове метрике како би се открили и изрудариле информације за потенцијално побољшање модела.

Добра ствар јесте што мрежна страница *polovniautomobili* поседује доста огласа и сваки дан се ажурирају тако да не би било тешко сакупити преко 100.000 слогова, чиме би се оснажио узорак старијих кола и оних које су скупље цене, с тим што овај корак дакако усложњава обраду података. То значи да би се смањила преприлагодба, која је била присутна и код модела случајне шуме и градијентног појачавања, што би уследило порасту метрика за процену.

ЛИТЕРАТУРА

Alzubi, J., Nayyar, A., & Kumar, A. (2018). Machine learning from theory to algorithms: an overview. *Journal of physics: conference series, 1142*, p. 012012.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of machine learning research, 13*.

Berrar, D. (2019). Cross-validation. *Cross-validation*.

Bilmes, J. (2020). Underfitting and overfitting in machine learning. *UW ECE course notes, 5*.

Breiman, L. (1996). Bagging predictors. *Machine learning, 24*, 123–140.

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *Peerj computer science, 7*, e623.

Costa-Climent, R., Haftor, D., & Staniewski, M. (2023, December). Using machine learning to create and capture value in the business models of small and medium-sized enterprises. *International Journal of Information Management, 73*, 102637. doi:10.1016/j.ijinfomgt.2023.102637

Cvejić, S., Hrnjaković, O., Jocković, M., Kupusinac, A., Doroslovački, K., Gvozdenac, S., . . . Miladinović, D. (2023, October 17). Oil yield prediction for sunflower hybrid selection using different machine learning algorithms. *Scientific Reports, 13*, 17611. doi:10.1038/s41598-023-44999-3

Deisenroth, M. P., Faisal, A. A., & Ong, C. S. (2020). *Mathematics for machine learning*. Cambridge University Press.

Delibašić, B., Suknović, M., & Jovanović, M. (2009). *Algoritmi mašinskog učenja za otkrivanje zakonitosti u podacima*. Fakultet organizacionih nauka.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.

Grus, J. (2019). *Data science from scratch: first principles with python*. O'Reilly Media.

Hawkins, D. M. (2004). The problem of overfitting. *Journal of chemical information and computer sciences, 44*, 1–12.

Hawkins, D. M., Basak, S. C., & Mills, D. (2003). Assessing model fit by cross-validation. *Journal of chemical information and computer sciences, 43*, 579–586.

Hien, N. L., Tien, T. Q., & Van Hieu, N. (2020). Web crawler: Design and implementation for extracting article-like contents. *Cybernetics and Physics, 9*, 144–151.

Kocherginsky, M., He, X., & Mu, Y. (2005). Practical confidence intervals for regression quantiles. *Journal of Computational and Graphical Statistics, 14*, 41–55.

Konstantinov, A. V., & Utkin, L. V. (2021). Interpretable machine learning with an ensemble of gradient boosting machines. *Knowledge-Based Systems, 222*, 106993.

Kroese, D. P., Botev, Z., & Taimre, T. (2019). *Data science and machine learning: mathematical and statistical methods.* Chapman and Hall/CRC.

Le Cook, B., & Manning, W. G. (2013). Thinking beyond the mean: a practical guide for using quantile regression methods for health services research. *Shanghai archives of psychiatry, 25*, 55.

Liu, Y., Wang, Y., & Zhang, J. (2012). New machine learning algorithm: Random forest. *Information Computing and Applications: Third International Conference, ICICA 2012, Chengde, China, September 14-16, 2012. Proceedings 3*, (pp. 246–252).

Lotz, M. (2018). Mathematics of machine learning. *Lecture notes (accessed 2021-03-30)*.

Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet], 9*, 381–386.

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in neurorobotics, 7*, 21.

Nikolić, M., & Zečević, A. (2019). Mašinsko učenje. *Beograd: Matematički fakultet*.

Nti, I. K., Nyarko-Boateng, O., Aning, J., & others. (2021). Performance of machine learning algorithms with different K values in K-fold cross-validation. *International Journal of Information Technology and Computer Science, 13*, 61–71.

Plevris, V., Solorzano, G., Bakas, N. P., & Ben Seghier, M. E. (2022). Investigation of performance metrics in regression analysis and machine learning-based prediction models. *8th European Congress on Computational Methods in Applied Sciences and Engineering (ECCOMAS Congress 2022)*.

Rao, R. B., Fung, G., & Rosales, R. (2008). On the dangers of cross-validation. An experimental evaluation. *Proceedings of the 2008 SIAM international conference on data mining*, (стр. 588–596).

Ratner, B. (2009). The correlation coefficient: Its values range between+ 1/- 1, or do they? *Journal of targeting, measurement and analysis for marketing, 17*, 139–142.

Sandberg, A., Drexler, E., & Ord, T. (2018). Dissolving the Fermi paradox. *arXiv preprint arXiv:1806.02404*.

Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms.* Cambridge university press.

Song, Y.-Y., & Ying, L. U. (2015). Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry, 27*, 130.

Stamp, M., Chandak, A., Wong, G., & Ye, A. (2021). On ensemble learning. *Malware analysis using artificial intelligence and deep learning*, 223–246.

Uhlig, S., Alkhasli, I., Schubert, F., Tschöpe, C., & Wolff, M. (2023). A review of synthetic and augmented training data for machine learning in ultrasonic non-destructive evaluation. *Ultrasonics*, 107041.

Voza, D., Dehghani, H., & Veličković, M. (2023). THE DISSOLVED OXYGEN PREDICTION BASED ON THE MACHINE LEARNING TECHNIQUES.

Winn, J. (2023). *Model-based machine learning.* CRC Press.

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining, 1*, стр. 29–39.

Xu, Y., & Goodacre, R. (2018). On splitting training and validation set: a comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. *Journal of analysis and testing, 2*, 249–262.

Ying, X. (2019). An overview of overfitting and its solutions. *Journal of physics: Conference series*, *1168*, стр. 022022.

Zöller, M.-A., & Huber, M. F. (2021). Benchmark and survey of automated machine learning frameworks. *Journal of artificial intelligence research, 70*, 409–472.

Вуковић, Н. (2012). Основе вероватноће. Београд: Факултет организационих наука.

Вуковић, Н., & Булајић, М. (2014). Основе статистике. Београд: Факултет организационих наука.

Everything you need to know about AI model training. (2023, December 13). Labellerr. Преузето 14. априла 2024, са https://www.labellerr.com/blog/everything-you-need-to-know-about-ai-model-training.

Hotz, N. (2024, March). *What is CRISP DM?* Data Science Process Alliance. Преузето 11. априла 2024, са https://www.datascience-pm.com/crisp-dm-2.

Machine learning market size, share, growth | Trends [2030]. (2023, May). Fortune Business Insights™ | Global Market Research Reports & Consulting. Преузето 12. априла 2024, са https://www.fortunebusinessinsights.com/machine-learning-market-102226.

Machine learning - Worldwide | Statista market forecast. (2024, March). Statista. Преузето 15. априла 2024, са https://www.statista.com/outlook/tmo/artificial-intelligence/machine-learning/worldwide.

Patro, R. (2021, February 1). Cross validation: K fold vs Monte Carlo. Medium. Преузето 13. априла 2024, са https://towardsdatascience.com/cross-validation-k-fold-vs-monte-carlo-e54df2fc179b.

(n.d.). Polovni automobili. Преузето 17. априла 2024, са https://www.polovniautomobili.com.

(n.d.). *scikit-learn: Machine Learning in Python*, ver. 0.16.1. Преузето 12. марта 2024, са https://scikit-learn.org.