Универзитет у Београду

Електротехнички факултет



**Проналажење скривеног знања**

Пројектни задатак  
2020/2021

|  |  |
| --- | --- |
| Професор: | Студент: |
| проф. др. Бошко Николић | Нина Трифуновић 2019/3281 |

Београд, јул 2021.

# Увод

Пројектни задатак на предмету Проналажење скривеног знања подразумева практичну примену стеченог теоријског знања на примеру једног конкретног проблема, а циљ овог пројекта јесте анализа понуде некретнина у Србији.

Пројекат се састоји из пет целина. Прва целина односи се на прикупљање података – имплементацију веб индексера и веб парсера, друга целина се односи на анализу прикупљених података, трећа целина на њихову визуелизацију, док се последње две целине односе на имплементацију конкретних алгоритама машинског учења, конкретно линеарне регресије и класификације коришћењем к најближих суседа.

# Прикупљање података

Подаци који су коришћени у овом пројектном задатку прикупљени су са сајта [nekretnine.rs](https://nekretnine.rs/). Реализована база података садржи информације о идентификацији понуде (адреса странице), типу некретнине, типу понуде, затим информације о локацији (град, део града, улица), величини некретнине, години изградње, површини земљишта, укупној спратности и спрату стана, податке о укњижености, типу грејања, броју соба и купатила, да ли постоји паркинг, да ли постоји лифт, да ли некретнина има терасу, податке о стању некретнине и цени исте.

За имплементацију веб индексера коришћен је *scrapy* пакет, уз коришћење ротирајућих проксија, док је за парсер коришћена техника *HTML* парсера.

# Претпроцесирање података

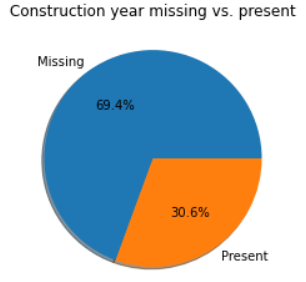
Како би се подаци припремили за коришћење у моделима машинског учења, најпре је извршена њихова експлоративна анализа и претпроцесирање.

Према захтевима пројектног задатка, издвојени су само станови који се продају у Београду.

Колоне које нису од користи при даљем моделирању су избачене ('offer\_id', 'estate\_type', 'offer\_type', 'city', 'street', 'land\_area', 'total\_floors', 'elevator'). Такође, избачени су они подаци о некретнинама код којих фале информације о делу града, обзиром на то да их није могуће допунити аутоматски.

Затим, визуелизована је дистрибуција *size:price*, и избачени станови који имају преко 400 квадрата, обзиром на то да су потенцијални *outlier*-и.

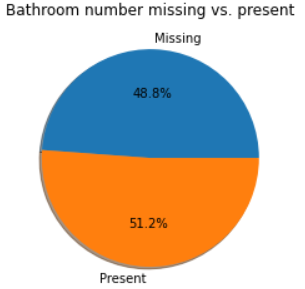
Обзиром на то да већина података нема информацију о години изградње, ова колона у потпуности је елиминисана.



Слика 1: Присуство/одсуство године изградње

За податке код којих фали информација о спрату коришћена је средња вредност, док су они подаци код којих недостаје информација о типу грејања избачени. За број соба такође је коришћена средња вредност.

Како ни већина података нема информацију о броју купатила, и та колона у потпуности је елиминисана.



Слика 2: Присуство/одсуство броја купатила

На крају, елиминисани су они редови код којих је недостајала информација о стању некретнине.

Овако добијен скуп података затим је коришћен у изградњи модела машинског учења.

# Регресија

Коришћењем претходно генерисаних података, имплементиран је алгоритам линеарне регресије. У питању је вишеструка линеарна регресија, а испробавањем различитих комбинација параметара, закључено је да следећа комбинација даје најбоље резултате: 'district', 'size', 'floor', 'registration', 'rooms', 'parking', 'balcony', 'state', 'price'.

Обучавање је реализовано коришћењем методе градијентног спуста, циљна променљива јесте цена некретнине за продају, а подаци су подељени на скуп за тренирање и скуп за тестирање.

Како би се упоредили резултати, коришћен је алгоритам линеарне регресије из *sklearn* библиотеке. За евалуацију коришћени су *Root Mean Squared Error* и *R squared*. На следећој табели дато је поређење добијених резултата:

|  |  |  |
| --- | --- | --- |
|  | *RMSE* | *R2* |
| *Sklearn* | 52173.01733639259 | 0.7328485950994026 |
| *Custom* | 52315.13716712996 | 0.7313911664770196 |

# Класификација

За реализацију класификације имплементиран је алгоритам к најближих суседа, при чему је одабир параметра к дозвољен за мануелну промену, а као функције растојања имплементиране су Еуклидова и Менхетн дистанца.

И у класификацији коришћена је иста комбинација улазних параметара, као и за линеарну регресију, према захтеву пројектног задатка.

Циљна променљива јесте ценовна класа некретнине за продају, а подаци су подељени на скуп за тренирање и скуп за тестирање.

Како би се упоредили резултати, коришћен је алгоритам к најближих суседа из *sklearn* библиотеке. За евалуацију коришћени су *Accuracy* и *Weighted avg F1 score*. На следећој табели дато је поређење добијених резултата:

|  |  |  |
| --- | --- | --- |
|  | *Accuracy* | *Weighted avg* |
| *Sklearn* | 0.57 | 0.55 |
| *Custom* | 0.57 | 0.55 |

# Апликација

Реализована је једноставна *streamlit* апликација. У менију са леве стране кориснику се даје могућност избора методе која ће бити коришћена (линеарна регресија или класификација помоћу к најближих суседа).

Неопходно је да корисник унесе податке о стану за који жели да се изврши предикција, након чега се помоћу раније имплементираних модела машинског учења даје процена о цени, односно ценовном опсегу, у зависности од изабраног алгоритма машинског учења.

# Закључак

Поређењем резултата имплементираних алгоритама машинског учења са већ постојећим имплементацијама из *sklearn* библиотеке, може се приметити да алгоритми дају веома сличне резултате.

Алгоритам линеарне регресије има нешто већу грешку у односу на већ постојећи регресор, док алгоритми к најближих суседа постижу исте оцене.

Оно што се такође може приметити јесте да су имплементирани алгоритми значајно спорији од алгоритама из *sklearn* библиотеке, и када је у питању класификација и када је у питању регресија, те се ту може даље радити на побољшању перформанси напреднијим методама.

Такође, посматрањем добијених резултата, може се закључити да ни једни ни други алгоритми не постижу сјајне резултате, међутим прикупљени подаци не дозвољавају претерано побољшање и уз већи квалитетније одабран скуп података резултати би могли бити значајно бољи.