Pronalaženje skrivenog znanja

Projektni zadatak - Izveštaj za zadatke 4 i 5

Student: Vukašin Gligorijević 19/3128

Jul 2021.

Zadatak 4

Četvrti zadatak podrazumevao je implementaciju algoritma linearne regresije za predikciju cene nekretnine za prodaju na teritoriji grada Beograda. Kao ulazne promenljive korišćene su:

- Kvadratura nekretnine (kvadratni metri)
- Udaljenost od centra grada (vazdušna linija, kilometri)
- Broj soba
- Godina izgradnje koja je predstavljena sa tri nova atributa:
    - Nekretnine izgrađene nakon 2000. godine
    - Nekretnine izgrađene pre 2000. godine
    - Nekretnine za koje ne postoji podatak o godini izgradnje

Pre samog treniranja, set koji će se koristiti za treniranje i testiranja ručno je prečišćen. Finalni skup obuhvata nekretnine čija je prodajna cena između 10.000 eur i 500.000 eur. Takođe profiltrirane su nekretnine čija je veličina značajno odstupa (više hiljada kvadrata, kao i kvadratura manja od 10 metara kvadratnih), kao i nekretnine koje su od centra grada udaljene više od 10 kilometara vazdušnom linijom. Inicijalni set brojao je 9.724 podataka, a nakon čišćenja set za treniranje i testiranje sadrži 8.229 podataka.

Prilikom obučavanja korišćena je metoda stohastičkog gradijentnog spusta, a kako bi se izbeglo preobučavanje modela korišćena je L2 regularizacija. Treniranje se izvršavalo kroz 50 epoha. Pre samog obučavanja bilo je potrebno dopuniti podatak o broju soba, jer određeni deo podataka nije imao taj podatak. U tim slučajevima za broj soba uzeta je medijana broja soba celog skupa. Nakon toga, set podataka je podeljen na trening i test skup u odnosu 80% - 20%. Kako bi model bio što precizniji bilo je potrebno odrediti hiperparametre modela *learning rate*, *regularization coefficient.* Za odabir oprtimalnih hiperparametara korišćen je metod unakrsne validacije deljenjem trening skupa na 5 delova (*k-fold cross validation*). Nakon obučavanja dobijeni su optimalni koeficijenti linearne jednačine, koji su dalje korišćeni za predikciju cene nekretnine. Poređenjem prediktovane cene i cene koja je očekivana na osnovu stečenog znanja o cenama nekretnina u Beogradu, zaključujemo da ovaj model ne radi baš najbolju predikciju, tj. cene dobijene predikcijom su znatno više od očekivane. Ovo ponašanje može se tumačiti na više načina. Jedan je taj da se u Beogradu jasno izdvajaju delovi grada koji nisu toliko blizu centra grada a važe za “elitne” delove gde je cena nekretnine znatno viša. Drugo objašnjene bi bilo da udaljenost od centra grada nije baš najbolji parametar iz razloga što je Beograd kao i drugi evropski gradovi policentrčan, tj. moguće je izdvojiti više delova grada koji imaju svoje centre (Zemun, Novi Beograd (Fontana, Stari Merkator), Novi Beograd (Blokovi 72, 45)). Potencijalno poboljšanje podrazumevalo bi definisanje više gradskih celina sa njihovim centrima, a zatim i klasifikaciju podataka u definisane gradske celine.

Zadatak 5

Peti zadatak podrazumevao je implementaciju algoritma k najbližih suseda za određivanje opsega cene prodajne nekretnine na teritoriji grada Beograda. U okviru ovog zadatka korišćeni su isti atributi koji su navedeni u prethodnom zadatku, s tim što period izgradnje nije pretvoren u tri nova atributa. Dve funkcije rastojanja koje su realizovane su:

- Euklidska distanca
- *Manhattan* distanca

Set podataka koji se koristi za ovaj algoritam je isti kao i u prethonom zadatku. Prečišćeni su *outlier*-i u pogledu cene, veličine i udaljenosti od centra grada.

Kako je u pitanju problem klasifikacije set podataka je klasifikovan u 5 cenovnih rangova:

- Klasa 1 - cena manja od 49 999 EUR
- Klasa 2 - cena između od 50 000 - 99 999 EUR
- Klasa 3 - cena između od 100 000 - 149 999 EUR
- Klasa 4 - cena između od 150 000 - 199 999 EUR
- Klasa 5 - cena veća od 200 000 EUR

Algoritam kao ulazne podatke koristi skup podataka, novi podatak koji je potrebno klasifikovati, kao i proizvoljan broj K. Ukoliko se K ne definiše koristiće se optimalno K koje je definisano kao kvadratni koren ukupnog broja podataka. Ukoliko je dobijeno K paran broj, K će biti uvećano za 1.