Projektni prijedlog

Predviđanje cijene kuća

Maja Piskač
Lucija Puškarić
Ivona Raguž
Mia Tadić

travanj, 2020.

Sadržaj

1 Uvodni opis problema

1.1 Problem

Problem koji rješavamo je predviđanje cijena kuća u Amesu, Iowi u razdoblju 2006. – 2010. godine.
Kupci se obično pri kupnji kuće vode nekim osnovnim stavkama kao što su broj soba, izgled i stanje kuće. No, podaci koje koristimo pri stvaranju modela sadrže 79 atributa, tj. varijabli koje detaljno opisuju brojne aspekte kuća i pokazatelje njihove kvalitete, pa čak i one za koje na prvu mislimo da uopće nemaju utjecaja na ukupnu cijenu.

1.2 Skup podataka

Skup podataka koje koristimo pri rješavanju problema poznat je pod nazivom *The Ames Housing dataset*. Podaci su sastavljeni su od strane Deana De Cocka u svrhe *data science* obrazovanja. Skup podataka je nadogradnja na standardni *Boston housing dataset*, s većim brojem primjera i značajki.

2 Cilj i hipoteze istraživanja problema

Cijene nekretnina važan su odraz ekonomije. Rasponi cijena kuća od velikog su interesa kako za kupce, tako i za prodavatelje. Cilj ovog projekta je stvoriti regresijski model koji je u stanju precizno procijeniti cijenu kuće u Amesu, Iowi, obzirom na njene značajke, tj. aspekte. Nastojimo minimizirati razliku između predviđene i stvarne cijene.
Za početak se nameću intuitivne pretpostavke kao što su:

- Kuće s više soba vrijede više. Obično su kuće s više soba veće, pa time kuća može primiti više ljudi, što opravdava veću cijenu.
- Kuće s boljim vanjskim stanjem i izgledom koštat će više.
- Kuće s većom nadzemnom površinom bit će skuplje.
- Kuće s većim brojem soba bit će skuplje.

3 Pregled dosadašnjih istraživanja

U dosadašnjim istraživanjima, često se kao metoda koja rješava ovakav problem koristi linearna regresija i to regularizirani modeli jer se postižu bolji rezultati i izbjegava se *overfitting*. Naravno, dobiju se još bolji rezultati ako se pronađe pogodan parametar *alpha* koji se javlja u l_1-normi te l_2-normi.

1. **Lasso regresija**

   Ova regresija daje dobre rezultate uz odabir značajki koje su prema dosadašnjim eksploratornim analizama očekivano važne za predikciju cijena kuća. Ono što je bitno je da ovaj model može biti osjetljiv na podatke koji su izvan očekivanih okvira (engl. *outliers*) i stoga se mora paziti da se na podacima provede pogodna metoda skaliranja (npr. *RobustScaler*).

2. **Ridge regresija**

   Lasso regresija te potonja Elastic net regresija su popularnije nego Ridge regresija što se tiče primjene na ovaj problem. Ridge regresija daje malo slabije rezultate nego Lasso regresija.

3. **Elastic net regresija**

   Ova metoda, kao hibrid gornje dvije, je također bila relativno uspješna u predikciji kao i Lasso regresija. Uz odabir parametra *alpha*, potrebno je uložiti trud i u odabir parametra *l1_ratio* tj. omjera između Lasso i Ridge kazne (eng. *penalty*).

Osim linearne regresije, za rješavanje ovog problema pogodne su *Gradient Boosting* metode. Najčešće su primjenjivane:

1. **Gradient Boosting Regression**
2. **LightGBM**
3. **XGBoost**

Najpopularnija je XGBoost metoda, no sve su u dosadašnjim istraživanjima bile uspješne, a rezultati su podjednako dobri kao i kod metoda linearne regresije. Nedostatak ovih metoda je da uspješnost znatno ovisi o parametrima pa je važno izvršiti dobru prilagodbu podataka (eng. *parameter tuning*).

S ciljem dobivanja još boljeg modela, često su se koristili razni ansambli gore navedenih modela, npr. ansabli bazirani na *averaging*-u, *stacking*-u te ansambl *voting regressor*.

4 Materijali, metodologija i plan istraživanja

Koristit ćemo podatke dostupne na kaggle challengeu, na linku [1]. Dosad smo se upoznale s podacima, analizirale varijablu *SalePrice* koju želimo predvidjeti, analizirale korelacije među podacima, osobito koreliranost između *SalePrice* i ostalih varijabli, pronašle nedostajuće podatke (engl. *missing data*) i podatke koji su izvan očekivanih okvira, te analizirale jesu li varijable normalno distribuirane. U ovoj fazi smo samo analizirale koje bi podatke mogle ukloniti te uočile da primjena log transformacije nad podacima pozitivno asimetrične (engl. *skewed*) varijable, rezultira normalnom distribucijom te varijable i zadovoljavanjem homoscedastičnosti, ali ćemo modele testirati sa i bez tih promjena kako bismo vidjele koliko one pridonose boljem predviđanju. Prilikom obrade podataka najviše smo se služile kaggle notebookom dostupnim na linku [2].

Problem ćemo pokušati riješiti sljedećim metodama:

1. **Lasso regresija**
   Kako je navedeno u prethodnom poglavlju, Lasso regresija je osjetljiva na outliere, pa će za ovaj model biti jako važno transformirati podatke i maknuti outliere. Budući da imamo velik broj atributa, koristit ćemo Lasso regresiju jer je ona pogodnija kod reduciranja kompleksnosti modela i prevencije overfittinga do koje bi moglo doći linearnom regresijom, te također koristi kod odabira atributa (engl. *feature selection*).

2. **XGBoost**
   Analiziramo podatke koji nedostaju u trening skupu, odnosno atribute čije vrijednosti nisu zabilježene za primjere trening skupa. Provest

ćemo analizu rješavanja *missing values* problema. Rješavamo i problem kategoričkih varijabli (*Label Encoding*, *One Hot Encoding*). Eliminiranje outliera je obvezan posao za precizniju analizu podataka, što ćemo i izvršiti. Promotrit ćemo i analizirati XGBoost parametre. Bit će vrlo bitno prilagoditi parametre (npr. *colsample_bytree*, *subsample*) jer previsoke ili preniske vrijednosti nekih parametara mogu dovesti do *overfittinga* ili *underfittinga*. Izmjerit ćemo grešku procijenjenih vrijednosti za primjere iz trening skupa.

3. **Linearna regresija**
   Pri analizi podataka, uočili smo atribute čija vrijednost nedostaje u nekim primjerima unutar trening skupa. Potrebno je vidjeti kako riješiti taj problem poznat pod nazivom *missing values*. Neke od atributa čije vrijednosti nedostaju nećemo uzimati u obzir pri stvaranju modela. Za neke atribute ćemo na temelju drugih atributa moći zaključiti koja vrijednost bi trebala biti upisana, a za neke atribute ćemo upisivati njegovu prosječnu vrijednost. Pozabavit ćemo se rješavanjem problema kategoričkih varijabli (*Label Encoding*, *One Hot Encoding*). Već pri analizi podataka uočile smo stupce, tj. atribute koji su najviše korelirani s ciljnom varijablom *Sales Price*. Te atribute ćemo koristiti pri stvaranju modela, a ostale zanemariti. U obzir će se uzimati i korelacije između ulaznih varijabli (atributa), što je prikazano i vizualizirano u analizi podataka. Pri stvaranju modela nećemo uključivati međusobno korelirane varijable. Nakon izvršenih manipulacija nad podacima, istrenirat ćemo model na gore izabranim podacima. Model je spreman za previđanje cijene na test podacima.

Cilj projekta je minimizirati pogreške između predviđenih cijena primjera iz trening skupa i njihovih stvarnih cijena, a zatim ćemo modele primijeniti na testnom skupu te rezultate usporediti s rezultatima na kaggleu. Također ćemo usporediti pogreške između pojedinih algoritama.

5 Očekivani rezultati predloženog projekta

Po završetku projekta očekujemo modele koji će, uslijed analize svih gore navedenih komponenti, dobro previđati cijene kuća koje se nalaze u testnom skupu. Očekujemo da izbacivanje nekih atributa iz promatranja u cilju smanjenja složenosti modela neće previše utjecati na razliku između stvarne i

predviđene cijene.

Literatura

[1] *Dataset*. URL: https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data.

[2] *Comprehensive data exploration with Python*. URL: https://www.kaggle.com/pmarcelino/comprehensive-data-exploration-with-python.