Predspracovanie k semestrálnemu projektu

Bc. Miroslav Čulík a Bc. Andrej Gáfrik

Fakulta informatiky a informačných technológií,
Slovenská technická univerzita v Bratislave,
Ilkovičova 2, 842 16 Bratislava
info@fiit.stuba.sk
https://www.fiit.stuba.sk

Abstract. V tejto priebežnej správe sú uvedené vybrané časti z druhého odovzdania semestrálneho projektu z predmetu Objavovanie znalostí. Po fáze exploratívnej analýzy sme podľa navrhnutých riešení pristúpili k predspracovaniu dát, zároveň sme vytvorili 2 nové atribúty, konkrétne recon_age a price_per_sqft, ktoré sú kombináciou pôvodných atribútov. Okrem toho sme experimentálne vyskúšali niekoľko modelov predikujúcich závislú premennú ceny nehnuteľnosti. Tieto modely sme vyhodnotili zvolenými metrikami a rovnako sme ich už v tejto fáze priebežne porovnali s modelmi, ktoré boli použité vo vedeckých štúdiách zaoberajúcimi sa podobnými úlohami predikcie cien nehnuteľností.

1 Opis problému a dát

V našom projekte by sme chceli sa chceli zaoberať problémom predikcie cien nehnuteľností na základe dostupných atribútov o týchto nehnuteľnostiach. Táto predikčná úloha môže pomôcť lepšie odhadnúť skutočnú cenu nehnuteľností pri ich predajoch a nákupoch.

Nami zvolený dataset obsahuje údaje o domoch, ktoré boli predané v USA (oblasť King County v štáte Washington) v období máj 2014 až máj 2015. Dáta sme získali z portálu kaggle.com vo formáte **.csv** súboru[1].

Po vykonaní exploratívnej analýzy bolo zistené, že naše dáta majú celkovo 21 atribútov (stĺpcov), z toho je 20 atribútov numerických a jeden atribút obsahuje dátumové hodnoty. Dokopy náš dataset obsahuje 21613 záznamov, pričom chýbajúce hodnoty v ňom nie sú prítomné.

1.1 Charakteristika dát

Ako bolo povedané vyššie, náš dataset má 21 stĺpcov, pričom okrem jedného prípadu ide o numerické atribúty. Charakteristika jednotlivých atribútov:

- **id** - index inštancie
- **date** - dátum predaja nehnuteľnosti
- **price** - cena nehnuteľnosti (pravdepodobne v USD)

[1] https://www.kaggle.com/harlfoxem/housesalesprediction

- **bedrooms** - počet spální v nehnuteľnosti
- **bathrooms** - počet kúpeľní v nehnuteľnosti
- **sqft_living** - rozloha obytného priestoru v stopách štvorcových (square feet)
- **sqft_lot** - rozloha celého pozemku v stopách štvorcových (square feet)
- **floors** - počet poschodí
- **waterfront** - či z nehnuteľnosti je výhľad na pobrežie
- **view** - index kvality výhľadu z nehnuteľnosti
- **condition** - stav nehnuteľnosti
- **grade** - stupeň vnútorného zariadenia
- **sqft_above** - rozloha obytného priestoru v nadzemnej časti budovy v stopách štvorcových (square feet)
- **sqft_basement** - rozloha obytného priestoru suterénu v stopách štvorcových (square feet)
- **yr_built** - rok postavenia nehnuteľnosti
- **yr_renovated** - rok renovovania nehnuteľnosti
- **zipcode** - poštové smerovacie číslo
- **lat** - zemepisná šírka (severná šírka)
- **long** - zemepisná dĺžka (západná dĺžka)
- **sqft_living15** - rozloha obytného priestoru pre najbližších 15 susedných nehnuteľností
- **sqft_lot15** - rozloha celého pozemku pre najbližších 15 susedných pozemkov

2 Stručný opis prác iných autorov

V rámci práce na druhom odovzdaní sme vyhľadali niekoľko vedeckých a nevedeckých článkov, ktoré pracovali s podobnými dátami ako my. Nasleduje ich stručný opis.

2.1 Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia[1]

Autori tohto článku [1] sa zamerali na predikciu cien nehnuteľností v austrálskom veľkomeste Melbourne. Z počiatočného počtu 34000 záznamov sa im podarilo dôkladným čistením vyfiltrovať 14000 záznamov, výsledné modely boli teda trénované na korpuse cca 20000 záznamov. Autori k outlierom pristupovali individuálne podľa stĺpca, v ktorom sa nachádzali (viac informácií autori v článku neposkytli). Predikovaný stĺpec s cenou nehnuteľností bol transformovaný pomocou logaritmu. Pri redukcii dimenzionality boli použité techniky **Stepwise** a **Boosting** na získanie najsilnejších prediktorov, pre model SVM bola taktiež použitá technika PCA. Všetky výsledky boli porovnávané s východzím modelom lineárnej regresie. Najlepšie výsledky boli dosiahnuté pri SVM, avšak veľké rozdiely medzi train MSE a eval MSE naznačujú preučený model.

- **Použité modely:** Lineárna regresia, Polynomiálna regresia, Regression Trees, Neuronová sieť, SVM
- **Metriky:** MSE

TABLE V. PREDICTION RESULTS

Model	Train MSE	Eval. MSE	Eval. Ratio
Linear regression	0.0948	0.0994	**1.00**
Polynomial regression	0.0773	0.0832	**0.84**
Regression tree	0.0925	0.0985	**0.99**
Neural Network	0.2657	0.2749	**2.77**
Stepwise & SVM	0.0558	0.0615	**0.62**
Stepwise & tuned SVM	0.0480	0.0561	**0.56**
PCA & SVM	0.0721	0.0810	**0.82**
PCA & tuned SVM	0.0474	0.0728	**0.74**

Fig. 1. Výsledky predikcií pri použití rôznych modelov[1]

2.2 Research on the Sustainable Development of UrbanHousing Price Based on Transport Accessibility:A Case Study of Xi'an, China [2]

Autori [2] v článku skúmajú vplyv prístupnosti verejnej dopravy na predikciu reálnych cien nehnuteľností. Tento článok pracuje s dátami obsahujúcimi nehnuteľnosti a možnosti transportu(cestná sieť, autobusové linky a linky metra) v hlavnej mestskej časti čínskeho mesta Xi'an. Použitím tradičného hedonického cenového modelu (z angl. *traditional hedonic price model*) a algoritmom náhodného lesa (z angl. *Random Forest*) autori porovnávajú situácie s prítomným, resp. neprítomným atribútom znázorňujúcim prístupnosť verejnej dopravy(prístupnosť autobusovej verejnej dopravy a metra). Výsledkom tejto práce bolo zistenie, že schopnosť modelov predikovať reálnu cenu nehnuteľností po pridaní atribútu prístupnosti verejnej dopravy je štatisticky zlepšená oproti situácii, kedy atribút prístupnosti verejnej dopravy prítomný nie je.

- **Použité modely:** Random Forests(M3 - max_features=8, M4 - max_features=4)
- **Metriky:** R^2, RMSE

Table 3. Comparison of the models in the testing set.

Model	R^2	RMSE	Runtime(s)
M1	0.219	4801	0.006
M2	0.227	4735	0.006
M3	0.797	2415	0.417
M4	0.840	2142	0.421

Fig. 2. Výsledky predikcií pri použití rôznych modelov[2]

2.3 Regression using sklearn on KC Housing Dataset [3]

Tento nevedecký článok [3] sa zaoberal predikciou cien nehnuteľností na našom datasete. V článku sa autor šikovne vyhol predspracovaniu dát, avšak dokázal pri jednoduchších modeloch dosiahnuť zaujímavé výsledky.

Model	$\mathbf{R^2}$
Jednoduchá lineárna regresia (sqft_living)	0.496
Jednoduchá lineárna regresia (grade)	0.46
Viacnásobná lineárna regresia (features1)	0.555
Viacnásobná lineárna regresia (features2)	0.672
Polynomiálna regresia(features1) - stupeň 2	0.759
Polynomiálna regresia(features1) - stupeň 3	0.664

Table 1. Výsledky predikcií pri použití rôznych modelov[3]

3 Predspracovanie a výber atribútov

V našom datasete sme detekovali niekoľko stĺpcov, ktoré obsahovali outlierov. Na detekciu sme používali metódu tzv. fúzov (z angl. *whiskers*), ktorá považovala za outlier hodnoty mimo intervalu (25 percentil -1,5*IQR; 75 percentil + 1,5*IQR).

Outlierov sme nahrádzali 3 rôznymi spôsobmi:

- **5-95 percentil** - hodnoty nachádzajúce sa pod ľavým "fúzom" boli nahradené 5 percentilom stĺpca, hodnoty nad pravým "fúzom" naopak 95 percentilom, takto boli nahradené vychýlené hodnoty v stĺpcoch price, sqft_living, sqft_above, sqft_living15

- **medián** - všetci outlieri boli nahradení mediánom daného stĺpca, napr. pri sqft_lot, sqft_lot15
- **priemer** - analogický postup, ako pri mediáne, len nahrádzanie priemernou hodnotou stĺpca, napr. sqft_basement

Po podrobnej analýze datasetu sme prišli k záveru, že vo väčšine prípadov sú outlieri reálne hodnoty a nie chyby (kontrola napr. výpisom rozlohy najdrahších nehnuteľností, počtu izieb, poschodí...). V stĺpcoch bathroom a bedrooms ale existovali záznamy, pri ktorých sme museli konštatovať, že vznikli chybou. Ide napr. o situáciu, kedy sme identifikovali záznam s počtom izieb 33. Keďže rozloha obytnej časti neodpovedala počtu izieb, rozhodli sme sa tento záznam vyhlásiť za chybu a hodnotu 33 nahradiť číslom 3. Na opačnej strane sme identifikovali 7 prípadov, pri ktorých bol počet izieb, ako aj kúpeľní 0. Keďže tieto záznamy mali aj podozrivo malú rozlohu, boli sme nútení pristúpiť k ich odstráneniu.

V minulej časti sme idetifikovali vychýlené hodnoty a navrhli sme ich nahradenie, ktoré sme zrealizovali nasledovne:

- **price** - transformácia pomocou boxcox transformácie a nahradenie outlierov 5-95 percentilom
- **bedrooms** - manuálna úprava inštancie s hodnotou 33 na 3
- **sqft_living** - transformácia pomocou logaritmu a nahradenie outlierov 5-95 percentilom
- **sqft_lot** - transformácia pomocou boxcox transformácie a nahradenie outlierov mediánom
- **sqft_above** - transformácia pomocou boxcox transformácie a nahradenie outlierov 5-95 percentilom
- **sqft_basement** - transformácia pomocou odmocniny a nahradenie outlierov priemerom alebo odstránenie stĺpca
- **sqft_living15** - transformácia pomocou logaritmu a nahradenie outlierov 5-95 percentilom
- **sqft_lot15** - transformácia pomocou boxcox transformácie a nahradenie outlierov mediánom

Pri stĺpci zipcode sme pristúpili k zakódovaniu jeho hodnôt binárnym kódovaním (binary_encoding), keďže sa jednalo o nominálny kategorický atribút. Pristúpili sme aj k vytvoreniu nových atribútov:

- **recon_age** - popisuje počet rokov od poslednej rekonštrukcie nehnuteľnosť po rok predaja. Ak nebola uskutočnená žiadna rekonštrukcia, ide o rozdiel roku predaja nehnuteľnosti a roku postavenia.
- **price_per_sqft** - popisuje strednú hodnotu ceny na 1 stopu štvorcovú (square feet2) ako pomer ceny nehnuteľnosti a súčet atribútu sqft_living a sqft_basement podľa danej lokality, ktorá je reprezentovaná pôvodnými hodnotami atribútu zipcode

Podľa atribútu **date** sme zoradili všetky inštancie a rozdelili sme ich na trénovaciu, validačnú a testovaciu množinu v pomere **70:20:10**. Keďže nám už tento atribút nebol viac potrebný, zmazali sme ho.

Implementáciu predspracovania ako aj ďalších potrebných častí v rámci tejto fázy máme uloženú v nasledujúcich python scriptoch:

- **analysis.py**- funkcie použité hlavne vo fáze exploratívnej analýzy, okrajovo tieto funkcie používame aj v tejto fáze pri sledovaní zmien po predspracovaní jednotlivých atribútov.
- **preprocessing2.py** - funkcie realizujúce predspracovanie jednotlivých atribútov ako pipeline jednotlivých funkcií.
- **feature_selection2.py** - obsahuje funkcie realizujúce výber atribútov. Pri výbere atribútov sme použili metódy filter(funkcia feature_filter) a wrapper(funkcie select_features_SFS a select_features_RFE).
- **metrics2.py** - obsahuje funkcie realizujúce výpočty pri jednoduchých modeloch

4 DM metódy

Na riešenie problému predikcie ceny nehnuteľností sme v prvotných experimentoch použili nasledovné prístupy strojového učenia:

- **Jednoduchá lineárna regresia** - pomocou jednej nezávislej premennej chceme predikovať závislú premennú ceny
- **Viacnásobná lineárna regresia** - pomocou viacerých nezávislých premenných chceme predikovať závislú premennú ceny, pričom každá nezávislá premenná je rádu 1
- **Polynomiálna regresia** - pomocou viacerých nezávislých premenných rádu vyššieho ako 1 predikujeme závislú premennú ceny. Počas našich experimentov vykonaných v tejto fáze sme skúsili polynomiálnu regresiu 2. a 3. stupňa.
- **Regresný rozhodovací strom** - Ide o upravenie rozhodovacích stromov, ktoré sa primárne používajú v úlohách klasifikácie. Regresné stromy pracujú na princípe učenia sa lokálnych lineárnych regresií pomocou aproximácie sínusovej funkcie s prihliadnutím na posun v pozorovaniach. Algoritmus si pri veľkom počte atribútov dokáže sám určiť, podľa ktorých bude modelovať závislú premennú (v našom prípade atribút **price**)

Pri viacnásobnej lineárnej regresii, ako aj polynomiálnej regresii sme využili algoritmy na výber čŕt (*feature selection*), konkrétne *filter* **pomocou hodnoty korelácie s predikovaným atribútom** a *wrapper* (**Sequential Feature Selection - SFS** a **Recursive Forward Elimination - RFE**)

5 Prvotné experimenty

Po predspracovaní dát nami vytvorenou pipeline, sme mali dáta rozdelené v pomere 70:20:10 (trénovacia : validačná : testovacia množina). Použili sme implementácie algoritmov z modulu sklearn. Najlepšie výsledky podľa metriky R^2 dosiahol model polynomiálnej regresie 3. stupňa s použitím selekcie čŕt RFE

(0.916 na trénovacej a 0.828 na testovacej množine), druhý najlepší výsledok podľa tejto metriky dosiahol model Regresného rozhodovacieho stromu(0.861 na trénovacej a 0.802 na testovacej množine). V prípade modelu jednoduchej lineárnej regresie sme vyskúšali aj 5, resp. 10 - násobnú krížovú validáciu, na porovnanie výsledkov s modelom jednoduchej lineárnej regresie trénovaného tradičným spôsobom, avšak tento prístup nám nepreukázal žiaden signifikantný rozdiel vo výsledkoch pri modeloch jednoduchej lineárnej regresie.

6 Vyhodnocovanie

Pri vyhodnocovaní vyššie opísaných modelov sme použili tieto metriky:

- Mean Square Error (MSE)
- Root Mean Square Error (RMSE)
- Root Mean Squared Log Error (RMSLE)
- R^2
- Adjusted R^2

6.1 Opis jednotlivých metrík

Mean squared error (MSE) je základná metrika, ktorá zohľadňuje absolútny rozdiel medzi predikovanými hodnotami a skutočnými hodnotami. Vypočíta sa ako priemer rozdielov predikovaných a reálnych hodnôt umocnených na druhú. Táto metrika preferuje modely uprednostňujúce väčšinové výsledky, jej použitie môže viesť k zavádzajúcim alebo nepresným záverom, preto je ideálne ju používať v kombinácii s inými metrikami. Vzťah na výpočet MSE je nasledovný:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2$$

Root mean squared error (RMSE) je metrika odvodená od MSE. Ako už z názvu vyplýva, vypočíta sa odmocnením výsledku MSE. Nasleduje vzťah výpočtu.

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}$$

Root Mean Squared Log Error (RMSLE) sa používa na výpočet pomeru predikovaných a reálnych hodnôt. Jej najlepšie využitie je v prípade, kedy pracujeme s veľmi veľkými hodnotami, teda aj MSE a RMSE narastajú do veľkých rozmerov. RMSLE taktiež viac penalizuje menšie predikcie, ako väčšie pri rovnakej vzdialenosti od reálnej hodnoty.

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(log(y_i) - log(\widehat{y_i}))^2}$$

$\mathbf{R}^2$ je štatistická metrika, ktorá znázorňuje mieru kvality regresného modelu. Vyjadruje, aký podiel variability závislej premennej model vysvetľuje. Nadobúda hodnoty od 0 po 1, pričom model s hodnotou 1 dokonalo predikuje hodnoty závislej premennej, pri hodnote 0 hovoríme o neužitočnom modeli. Vypočítame ju nasledovne:

$$\frac{\sum(\widehat{y_i} - \overline{y_i})^2}{\sum(y_i - \overline{y_i})^2}$$

Adjusted $\mathbf{R}^2$ je modifikáciou vyššie opísanej metriky. Taktiež vypovedá o schopnosti modelu vysvetliť variáciu závislej premennej, no narozdiel od R^2 berie do úvahy aj počet čŕt v modeli. Pri pridaní zbytočnej črty do modelu sa Adjusted R^2 zníži. Adjusted R^2 je vždy menšie alebo rovnaké ako R^2.

Model	$\mathbf{R}^2$
Jednoduchá lineárna regresia (sqft_living)	0.402
Jednoduchá lineárna regresia (grade)	0.432
Viacnásobná lineárna regresia (features1) - filter	0.501
Viacnásobná lineárna regresia (features2) - SFS	0.744
Viacnásobná lineárna regresia (features3) - RFE	0.754
Polynomiálna regresia(features1) - stupeň 2	0.789
Polynomiálna regresia(features1) - stupeň 3	0.828
Regresný rozhodovací strom - max. hĺbka 8	0.802

Table 2. Výsledky použitých modelov

Modely, s ktorými sme experimentovali sme primárne porovnávali podľa metriky R^2. Pri metrikách MSE, resp. RMSE sme odhalili príliš veľké hodnoty chybovosti, ktoré sme si pôvodne interpretovali ako neschopnosť modelu naučiť sa predikovať ceny nehnuteľností, ktoré majú vysoké hodnoty (viď obrázok). Tieto vysoké hodnoty metrík MSE a RMSE sa nám podarilo znížiť experimentálnym vyfiltrovaním inštancií, ktorých skutočná hodnota ceny bola vyššia než zvolená hodnota 815 000, avšak týmto krokom sa zároveň značne znížila aj hodnota metriky R^2, čo našu pôvodnú hypotézu vyvrátilo.

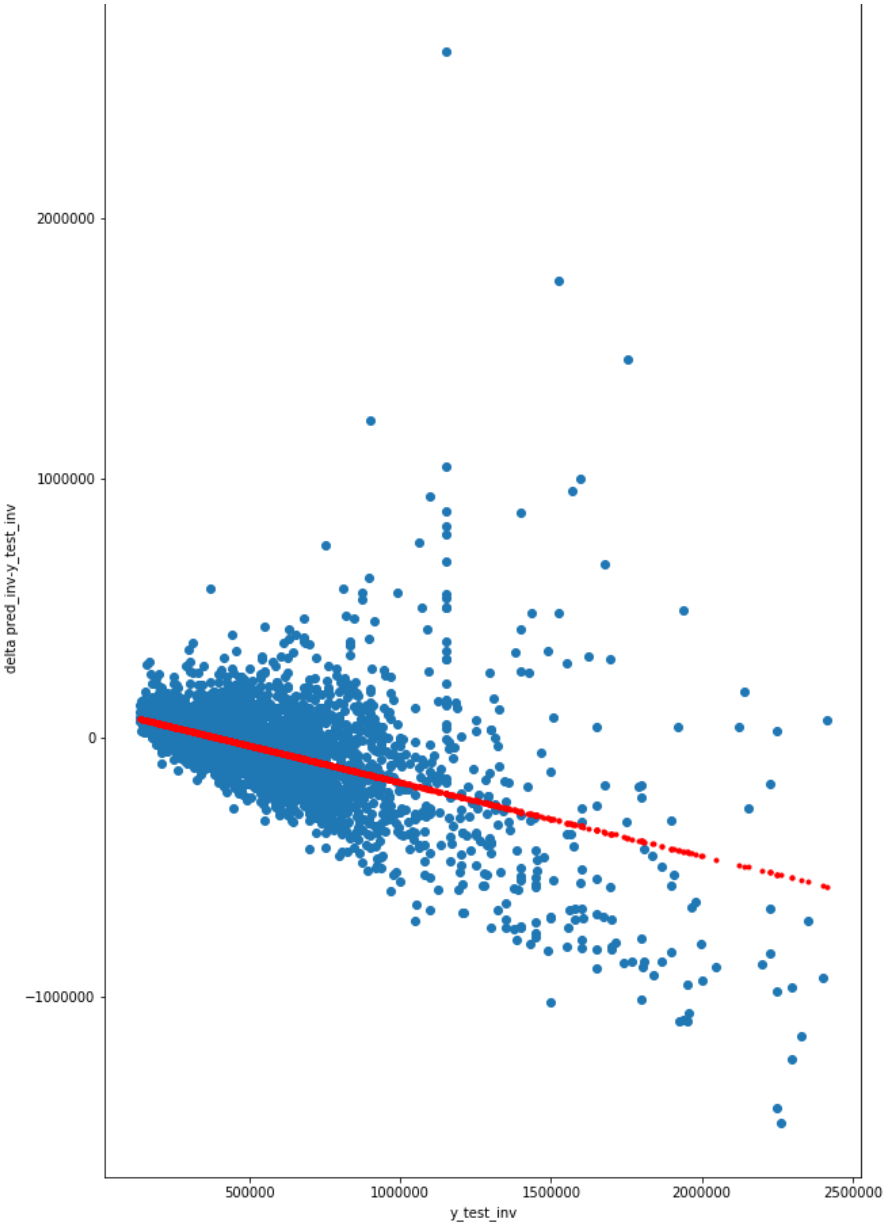

Fig. 3. Grafické zobrazenie závislosti skutočnej hodnoty a rozdielu predikovanej a skutočnej hodnoty ceny. Z grafu si môžeme všimnúť rastúci rozptyl rozdielov predikovaných cien so zvyšujúcou sa skutočnou cenou nehnuteľností

7 Predpokladaný scenár riešenia

V ďalšej fáze by sme sa chceli primárne sústrediť na výber modelu, ktorý by mohol predikovať výsledky s čo najvyššou úspešnosťou primárne s ohľadom na metriku R^2 a zároveň by sme chceli nájsť vhodný spôsob zníženia chybovosti, ktorá je reprezentovaná metrikami MSE a RMSE, prípadne RMSLE. Ako riešenia sa nám ponúka ladenie hyperparametrov (napr. pri regresnom rozhodovacom strome nastavenie maximálneho počtu atribútov pri rozhodovaní alebo hĺbka pri ktorej už začneme orezávať), prípadne drobné úpravy počas predspracovania. Ďalšou možnosťou, ktorú by sme chceli zrealizovať je vyskúšanie ďalších typov prístupov strojového učenia, konkrétnejšie Náhodný les (z angl. *Random Forest*) alebo SVR (z angl. *Support Vector Regression*). Všetky tieto modely plánujeme vyhodnotiť vyššie opísanými metrikami a porovnať ich s riešeniami, ktoré sme uviedli v časti s prácami iných autorov.

References

[1] Phan, The Danh: Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia. In: 2018 International Conference on Machine Learning and Data Engineering (iCMLDE), pp. 35–42., IEEE (2018)

[2] Xue, Chao and Ju, Yongfeng and Li, Shuguang and Zhou, Qilong: Research on the Sustainable Development of Urban Housing Price Based on Transport Accessibility: A Case Study of Xi'an, China. Sustainability **12**(4), 1497–1513 (2020)

[3] Mutyala, Nikhil Kumar, `https://towardsdatascience.com/regression-using-sklearn-on-kc-housing-dataset-1ac80ca3d6d4`. Last accessed 17 Apr 2020