Exploratívna analýza 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 prvého odovzdania semestrálneho projektu z predmetu Objavovanie znalostí. Menovite sa v nej nachádza opis problému, motivácia riešenia tohto problému a opis nami vybraných dát.

1 Opis problému a motivácia

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.

Našou motiváciou je získanie cenných skúseností z metód dátovej analýzy a využitia algoritmov strojového učenia.

2 Opis dát spolu s charakteristikami dát

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 (link).

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.

2.1 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)
- **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

2.2 Výber z analýzy jednotlivých atribútov

Nasleduje výber najzaujímavejších atribútov. Niektoré atribúty obsahovali veľmi vysoký počet outlierov, odhalenie čoho viedlo k zváženiu normalizácie týchto stĺpcov.
Normalizované boli nasledovné atribúty: price, sqft_living, sqft_lot, sqft_above, sqft_lot15.

Stĺpec date obsahuje 358 unikátnych hodnôt, analýza tohto atribútu nám potvrdila hypotézu, že cez víkend budú predaje významne nižšie, ako počas týždňa. Taktiež sme zistili, že najvyšší počet predajov bol v mesiaci máj, čo je ale pochopiteľné, keďže je tento mesiac ako jediný v našom datasete dvakrát (pre rok 2014 aj 2015).

Stĺpec price je dôležitý, pretože v ďalších fázach projektu by sme hodnoty tohoto stĺpca chceli predikovať. Ide o cenu, za ktorú bola nehnuteľnosť predaná. Z grafu 1 je možné vidieť, že hodnoty cien sú veľmi naklonené doľava, čo dokazuje aj hodnota koeficientu asymetrie. Pre účel identifikácie outlierov sme použili box-cox transformáciu (viď 2).

Stĺpec sqft_living udáva rozlohu obytnej časti nehnuteľnosti v štvorcových stopách. Pri analýze tohto stĺpca bolo odhalené množstvo outlierov, no iba na pravej strane boxplotu. Dáta tohto atribútu sme teda normalizovali pomocou logaritmu (3).

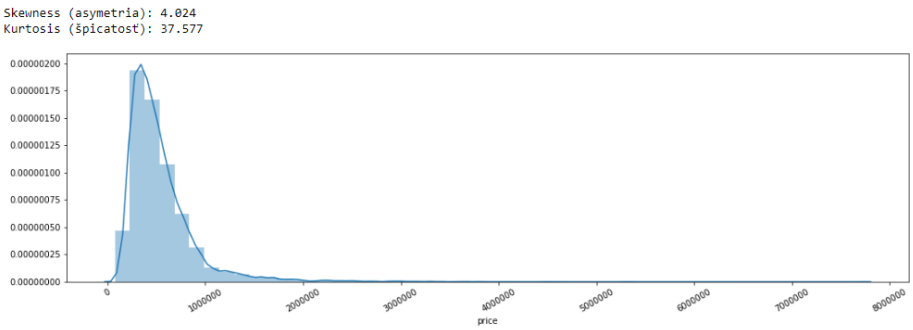

Fig. 1. Rozdelenie hodnôt stĺpca price pomocou grafu početnosti.

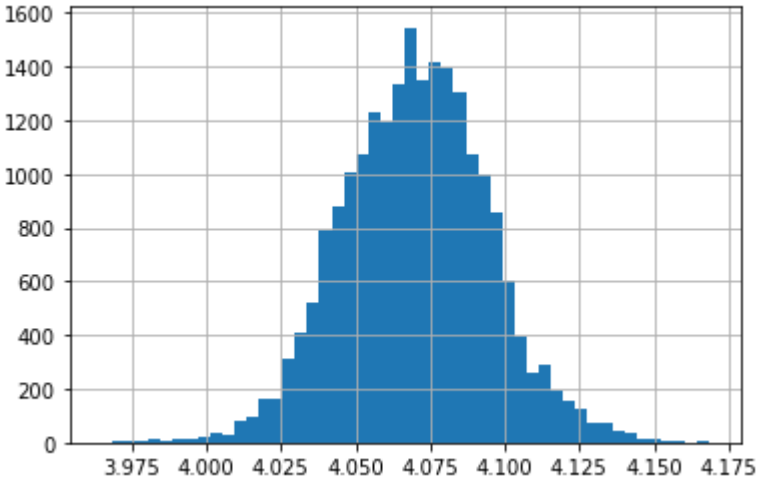

Fig. 2. Stĺpec price po použití box-cox normalizácie.

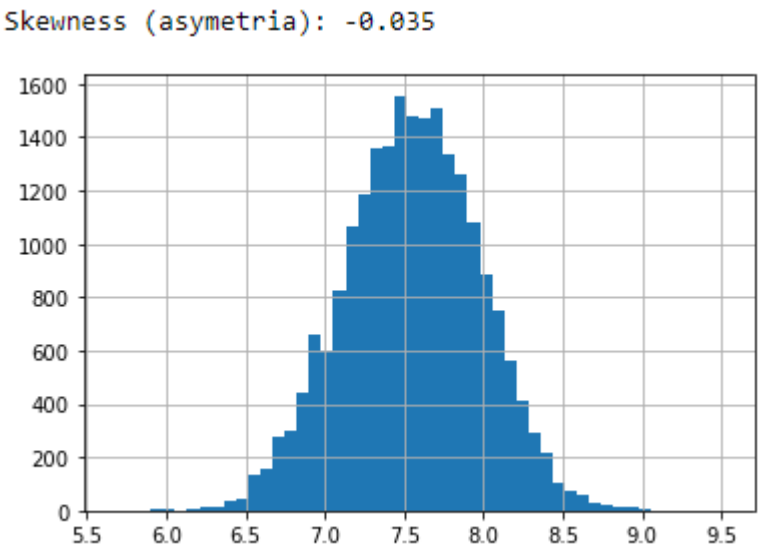

Fig. 3. Logaritmická normalizácia stĺpca sqft_living.

Stĺpec recon_age sme sa rozhodli pridať kvôli zisteniu počtu rokov od rekonštrukcie budovy. Pokiaľ dom zrekonštruovaný nikdy nebol, bude platiť rozdiel medzi rokom predaja a rokom postavenia, inak budeme robiť rozdiel roku predaja a roku rekonštrukcie. Tento atribút bol pridaný z dôvodu, že sme chceli preskúmať možnú koreláciu medzi vekom domu (časom od rekonštrukcie) a inými atribútmi (prioritne cenou). Takýto vzťah sa nepotvrdil, pomerne slabú koreláciu sme objavili s atribútmi opisujúcimi technický stav budovy.

2.3 Výber z párovej analýzy atribútov

Nasleduje výber toho najlepšieho, čo ponúka párová analýza nášho datasetu. V nami vybranom datasete je možné pozorovať viacero zaujímavých korelácii, napr. medzi atribútmi grade a sqft_above, sqft_above a sqft_living alebo price a sqft_living (viď. 4). Silná a zaujímavá korelácia bola identifikovaná medzi stĺpcami price a sqft_living. Tiež je možné na (5) vidieť, že domy s výhľadom na pobrežie majú vyššiu cenu pri porovnateľných hodnotách rozlohy obytnej časti. Taktiež môžeme sledovať vysoký vplyv rastúcej úrovne vybavenia domu na rastúci medián jeho ceny (6).

2.4 Opis chýbajúcich a vychýlených hodnôt

V datasete sme nezaznamenali žiadne chýbajúce hodnoty v žiadnom z atribútov. K vychýleným hodnotám sme pristúpili nasledovne:

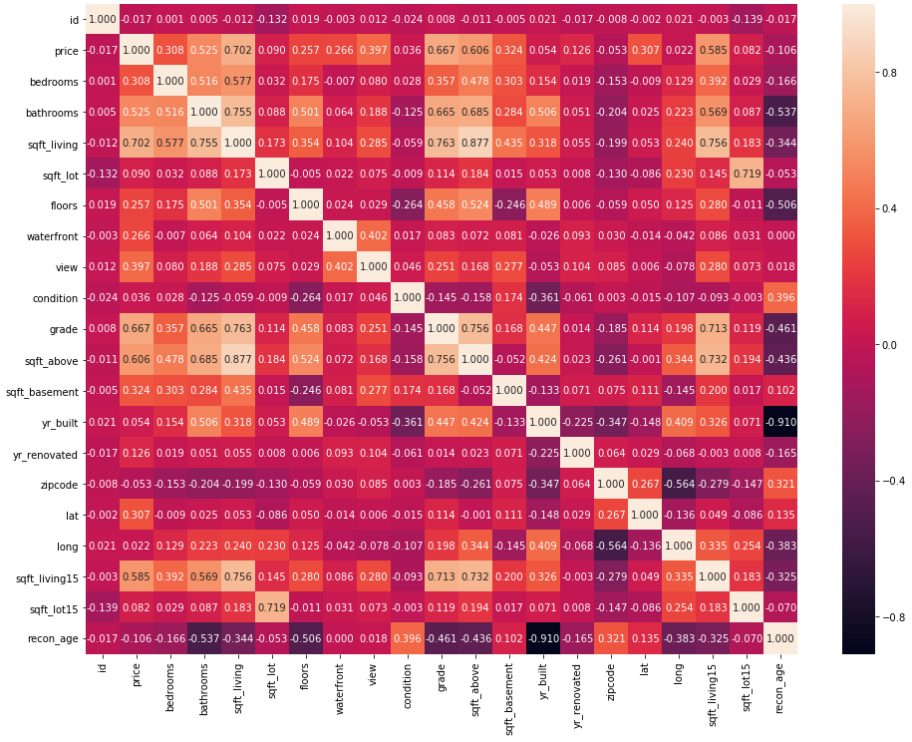

Fig. 4. Teplotná mapa korelácií jednotlivých atribútov.

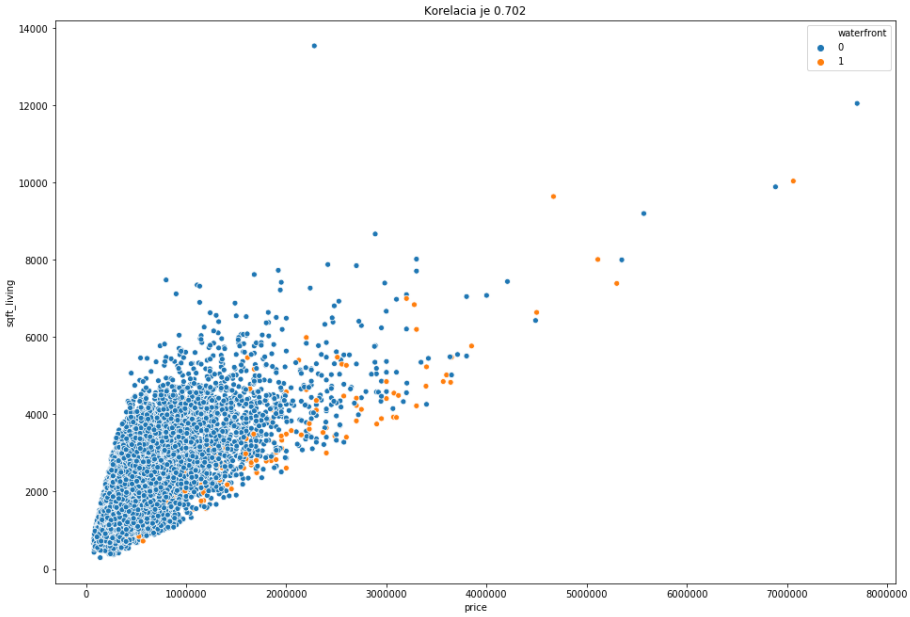

Fig. 5. Scatterplot vyjadrujúci koreláciu stĺpcov price a sqft_living.

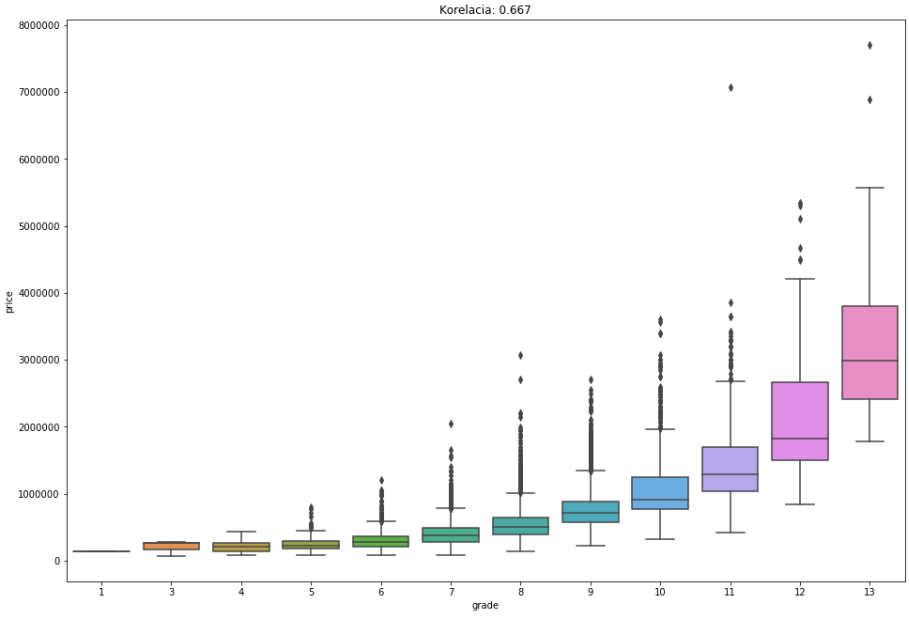

Fig. 6. Boxplot vyjadrujúci koreláciu stĺpcov grade a price.

- **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

Okrem zistení sme objavili 10 inštancií, ktoré majú hodnotu 0.0 pri počte kúpeľňového vybavenia v stĺpci **bathroom**, pričom až 7 z nich má hodnotu počtu spální v atribúte **bedrooms** rovnú 0, zvyšné 3 majú hodnotu 1. Navrhujeme teda týchto 10 inštancií zmazať.

3 Definovanie úlohy objavovania znalostí

Keďže chceme predikovať hodnotu spojitého numerického atribútu, identifikovali sme úlohu regresie. Pri riešení tejto úlohy by sme chceli vyskúšať viaceré algoritmy.

Ako prvé by sme chceli vyskúšať modely **lineárnej regresie**, konkrétnejšie jednoduchej a viacnásobnej lineárnej regresie, pomocou ktorých je možné pomocou jedného, resp. viacerých spojitých numerických atribútov predikovať hodnoty ďalšieho spojitého numerického atribútu. Ďalej by sme chceli preskúmať modely **polynomiálnej regresie**, ktorá používa väčšiu než prvú mocninu nezávislej premennej.

Support Vector Regression (SVR), ktorý je rozšírením Support Vector Machine(SVM), môže byť použitý pri regresii, keďže SVR regresia využíva necitlivú stratovú funkciu ε (insensitive loss function), ktorá má za úlohu minimalizovať chybovosť pomocou úpravy vektora váh a zavedených prídavných premenných ξ (angl. slack variables).

Podobne ako pri klasifikácii, tak aj pri úlohách regresnej analýzy môžeme použiť rozhodovacie stromy, celým názvom **Decision Tree Regression / Regression Trees**. Regresné stromy pracujú na princípe učenia sa lokálnych lineárnych regresií pomocou aproximácie sínusovej funkcie s prihliadnutím na šum v pozorovaniach.

4 Predpokladaný scenár riešenia (problémy)

Po vykonaní vyčerpávajúcej exploratívnej analýzy sme identifikovali vychýlené hodnoty v určených stĺpcoch a v nasledujúcej fáze by sme chceli pristúpiť k transformácii hodnôt v týchto stĺpcoch alebo k inému vhodnému nahradeniu, keďže viaceré algoritmy regresnej analýzy sú na tieto odľahlé hodnoty veľmi citlivé. Rovnako uvažujeme aj o vynechaní niektorých atribútov, ktoré nevykazovali veľkú koreláciu k cieľovému atribútu ceny, prípadne žiadnemu inému atribútu.

Z hľadiska nahrádzania chýbajúcich hodnôt nie je nutné pristupovať k žiadnemu riešeniu, keďže žiadne chýbajúce hodnoty sme neobjavili. Po fáze predspracovania by sme chceli siahnuť po konkrétnych implementáciách rôznych typov modelov, ktoré sme v predchádzajúcej kapitole stručne opísali a vhodnými metrikami by sme ich vyhodnotili.