3. zÃąpoÄŽtovÃą Ãžloha z 01RAD

Emanuel FrÃątrik

2021-12-16

1 3. zápočtová úloha z 01RAD

1.1 Popis úlohy

Datový soubor vychází z datasetu `House Sales in King County, USA`, který je k nalezení například na kaggle.com, nebo v knihovně `library(moderndive)` data `house_prices`. Původní dataset obsahuje prodejní ceny domů v oblasti King County, která obsahuje i město Seattle, a data byla nasbírána mezi květnem 2014 a květnem 2015. Pro naše potřeby bylo z datasetu vypuštěno jak několik proměnných, také byl dataset výrazně osekán a lehce modifikován.

Dále byl dataset již dopředu rozdělen na tři části, které všechny postupně v rámci 3. zápočtové úlohy využijete.

X	id	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront
1	1	2395000	4	3.25	3800	19798	2.0	0
2	2	679000	3	2.50	2770	9350	2.0	0
3	3	664000	2	1.75	1720	5785	1.0	0
4	4	915000	5	2.50	2750	5589	1.5	0
5	5	450000	5	2.50	2850	209523	1.0	0
6	6	305000	4	2.50	2320	4683	2.0	0

view	condition	grade	sqft_above	sqft_basement	yr_built	yr_renovated	sqft_living15	sqft_lot15	split
0	3	10	3800	0	1969	2009	3940	18975	train
3	3	8	2770	0	1957	2000	2660	9695	train
0	3	6	860	860	1948	2002	1680	5184	train
0	5	9	1840	910	1910	0	1460	4250	train
0	4	7	1930	920	1925	1968	2220	209523	train
0	3	7	2320	0	2007	0	2230	5750	train

Data celkem obsahují následujících 18 proměnných, přičemž naším cílem je prozkoumat vliv 12 z nich na cenu nemovitostí `price`. Přičemž anglický popis jednotlivých proměnných (sloupců) je následující:

Feature	Description
id	Our notation for a house
price	Price is prediction target
bedrooms	Number of Bedrooms/House
bathrooms	Number of Bathrooms/Bedrooms
sqft_living	Square footage of the home
sqft_lot	Square footage of the lot
floors	Total floors (levels) in house
waterfront	House which has a view to a waterfront
view	Has been viewed
condition	How good the condition is Overall
grade	Overall grade given to the housing unit
sqft_above	Square footage of house apart from basement
sqft_basement	Square footage of the basement
yr_built	Built Year
yr_renovated	Year when house was renovated
sqft_living15	Living room area in 2015 (implies– some renovations)
sqft_lot15	lotSize area in 2015 (implies– some renovations)
split	Splitting variable with train, test and validation sample

1.2 Podmínky a body

Úkol i protokol vypracujte samostatně. Pokud na řešení nějaké úlohy budete přesto s někým spolupracovat, radit se, nezapomeňte to u odpovědi na danou otázku uvést. Tato zápočtová úloha obsahuje 10 otázek po 1 bodu. Celkem za 3 zápočtové úlohy bude možné získat 30 bodů, přičemž pro získání zápočtu je potřeba více jak 20. Další dodatečné body mohu případně individuálně udělit za řešení mini domácích úkolů z jednotlivých hodin.

1.3 Odevzdání

Protokol ve formátu pdf (včetně příslušného Rmd souboru) odevzdejte prostřednictvím MS Teams, nejpozději do 31. 1. 2022.

1.4 Průzkumová a grafická část:

- Otázka 01

  Ověřte rozměry datového souboru, typy jednotlivých proměnných, a shrňte základní popisné charakteristiky všech proměnných. Vykreslete histogram a odhad hustoty pro odezvu `price`, dá se z toho již něco odvozovat pro budoucí analýzu?

- Odpoveď 01

  Dataset obsahuje 2000 pozorovaní popísaných 19 premennými pričom premenná `X` je rovnaká ako `id` a preto ju vylúčim. Ostatné relevantné premenné sú reprezentované numerickými hodnotami. Niektoré z nich ako napr. `view` alebo `condition`, `waterfront`, `grade` popisujú akési triedy a preto ich je možné reprezentovať ako kategorické premenné čo ale nie je nutné a preto to nevykonám. Podobne premenné `yr_built` a `yr_renovated` popisujú roky vzniku a renovácie domu ale keďže neplánujeme predikciu ceny podľa časovej rady bude možno vhodnejšie popisovať roky v desaťročiach resp. obe premenné faktorizovať. Popisné charakteristiky numerických a kategorických premenných sú zhrnuté v tabuľkách

Table 1: Popisné štatistiky pre numerické premenné

PREMENNÁ	PRIEMER	ROZPTYL	MEDIÁN	MIN	MAX	POČET NA
price	6.31e+05	2.61e+11	5.01e+05	301	8.50e+06	0
bedrooms	3.41e+00	7.47e-01	3.00e+00	2	6.00e+00	0
bathrooms	2.18e+00	6.10e-01	2.25e+00	1	4.75e+00	0
sqft_living	2.15e+03	8.61e+05	1.98e+03	105	1.00e+04	0
sqft_lot	1.60e+04	1.47e+09	7.69e+03	690	6.41e+05	0
floors	1.50e+00	2.68e-01	1.50e+00	1	3.00e+00	0
waterfront	2.10e-02	2.10e-02	0.00e+00	0	1.00e+00	0
view	3.68e-01	9.31e-01	0.00e+00	0	4.00e+00	0
condition	3.33e+00	3.93e-01	3.00e+00	1	5.00e+00	0
grade	8.14e+00	7.80e+01	7.00e+00	4	2.32e+02	0
sqft_above	1.82e+03	6.56e+05	1.65e+03	580	7.68e+03	0
sqft_basement	3.34e+02	2.10e+05	0.00e+00	0	2.36e+03	0
yr_built	1.96e+03	9.44e+02	1.96e+03	1900	2.02e+03	0
yr_renovated	8.79e+02	9.83e+05	0.00e+00	0	2.02e+03	0
sqft_living15	1.97e+03	4.64e+05	1.83e+03	780	5.79e+03	0
sqft_lot15	1.31e+04	6.32e+08	7.63e+03	1023	3.11e+05	0

?? a 2. Histogram rozdelenia premennej `price` je zobrazený na obrázku 1. Vidíme, že rozdelenie je vpravo zošikmené a teda nie normálne. Toto by ale nemalo spôsobovať problémy keďže vyžadovaná je len normalita reziduii.

- Otázka 02

  Jsou všechny proměnné použitelné pro analýzu a predikci ceny nemovitostí? Pokud data obsahují chybějící hodnoty, (případně nesmyslné hodnoty), lze je nějak nahradit (upravit), nebo musíme data odstranit?

- Odpoveď 02

  Premenné relevantné na predikciu sú všetky okrem premennej `id` a `split`, ktoré nepopisujú nejaké vlastnosti domu ale sú to len pomocné premenné. Ostatné relevantné premenné bude treba preskúmať predovšetkým na multikolinearitu. Dataset neobsahuje žiadne chýbajúce hodnoty. Prediktor `grade` obsahuje 6 veľmi vysokých hodnôt oproti ostatným hodnotám. Tieto hodnoty sú pravdepodobne chybné. Takéto chybné pozorovania nemusíme nutne odstraňovať ale môžeme ich napr. nahradiť priemerom alebo v prípade kategorickej premennej skôr modusom teda najčastejšou hodnotou. Možné by bolo aj natrénovať model z ostatných dát a tento použiť na predikciu chýbajúcich/chybných dát.

- Otázka 03

  Zkontrolujte pro 4 vybranné proměnné (price, sqft_living, grade, yr_built) bylo-li rozdělení datasetu pomocí proměnné split náhodné. Tj mají zmíněné proměnné ve skupinách train, test a validation přibližně stejné rozdělení?

- Odpoveď 03

  Rovnakosť resp. homogenitu rozdelenie v skupinách `train`, `test` a `validation` pre 4 vybrané premenné

Table 2: Popisné charakteristiky pre kategorické premenné

PREMENNÁ	N = 2,000
yr_built	
1900-20	282 / 2,000 (14%)
1921-40	282 / 2,000 (14%)
1941-60	522 / 2,000 (26%)
1961-80	404 / 2,000 (20%)
1981-2000	273 / 2,000 (14%)
2000-2015	237 / 2,000 (12%)
yr_renovated	
never	1,119 / 2,000 (56%)
1931-40	3 / 2,000 (0.1%)
1941-50	6 / 2,000 (0.3%)
1951-60	24 / 2,000 (1.2%)
1961-70	39 / 2,000 (2.0%)
1971-80	54 / 2,000 (2.7%)
1981-90	158 / 2,000 (7.9%)
1991-2000	186 / 2,000 (9.3%)
2001-2010	246 / 2,000 (12%)
2011-2015	165 / 2,000 (8.2%)

[1] n / N (%)

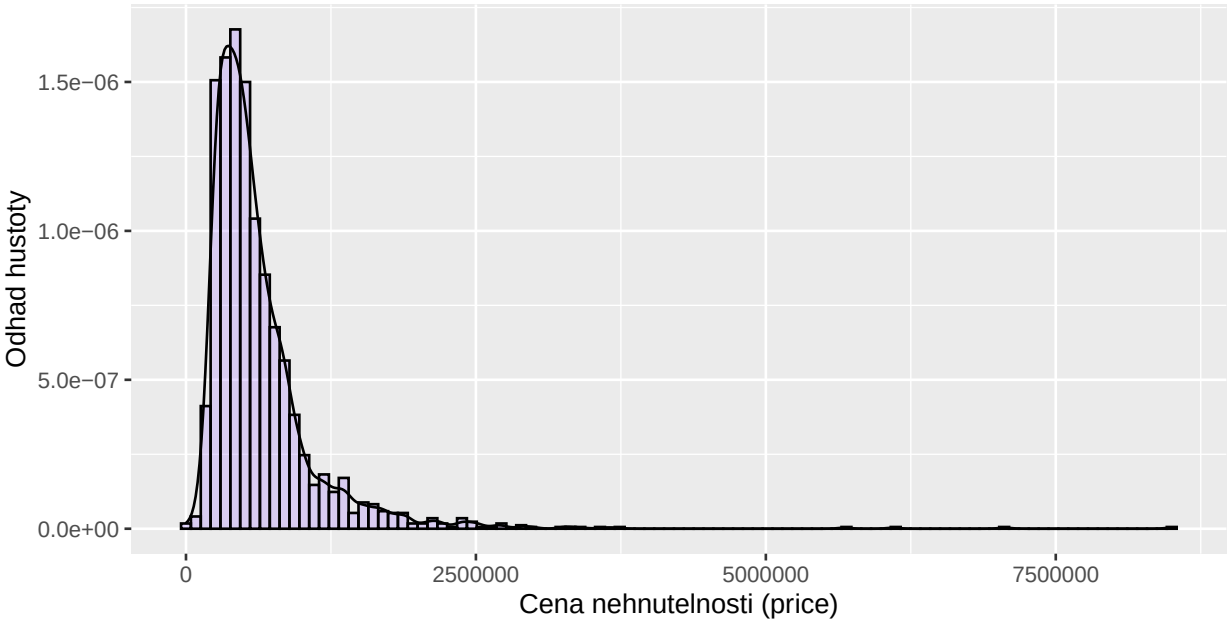

Figure 1: Histogramový a jadrový odhad hustoty premennej `price`

Table 3: Výsledky Chi squared testu homogenity rozdelenia premenných vybraných premenných v ‘train‘, ‘test‘ a ‘validation‘ kategoriach

premenná	chi2 statistic	p-value
price	70.1	0.822
sqft_living	93.4	0.898
grade	27.1	0.618
yr_built	7.4	0.687

môžme overiť pomocou `chi-squared` testu homogenity pričom spojité premenné je potrebné najskôr kategorizovať. Zhrnutie výsledkov testu môžme vidieť v tabuľke 3. Vo všetkých štyroch prípadoch vyšiel výsledok testu nesignifikantne na hladine 0.05 a preto nezamietame hypotézu o homogenite rozdelenia v kategóriach `train`, `test` a `validation` vo všetkých štyroch prípadoch.

1.5 Lineární model (použijte pouze trénovací data, tj. split == “train”):

Table 4: VIF faktor pre dizajnovu maticu všetkych regresorov

	GVIF	Df	GVIF^(1/(2*Df))
bedrooms	1.84	1	1.36
bathrooms	3.01	1	1.73
sqft_living	18188.68	1	134.87
sqft_lot	2.45	1	1.57
floors	2.19	1	1.48
waterfront	1.41	1	1.19
view	1.76	1	1.33
condition	1.38	1	1.17
grade	308.15	1	17.55
sqft_above	14667.14	1	121.11
sqft_basement	4487.23	1	66.99
yr_built	3.55	5	1.14
yr_renovated	2.84	9	1.06
sqft_living15	2.75	1	1.66
sqft_lot15	2.59	1	1.61

- Otázka 04

  Spočtěte korelace mezi jednotlivými regressory a graficky je znázorněte. Dále spočtěte číso podmíněnosti matice regresorů Kappa a VIF. Pokud se v datech vyskytuje znatelná multicollinearita, rozhodněte jaké proměnné a proč použijete v následném linárním modelu.

- Odpoveď 04

  Korelačné koeficienty ako aj scatterploty resp. boxploty sú znázornené na obrázku 2. Vysoká korelácia medzi prediktormi môže naznačovať kolinearitu medzi danými prediktormi. Z obrázka vidíme, že takýmito kolineárnymi prediktormi môžu byť napríklad `sqft_living` a `bathrooms` ďalej `sqft_living` a `sqft_above` čo je v zhode s popisom daných dvoch regresorov. Vysokú mieru korelácie vykazujú aj regresory `sqft_living` a `sqft_living15` a podobne aj `sqft_lot` a `sqft_lot15`. Čo sa týka `VIF` faktora tak tento je zobrazený v tabuľke 4 pričom v prvom stĺpci vidíme klasickú hodnotu `VIF` a v poslednom stĺpci tabuľku zobecnený `VIF`, ktorý je použiteľný aj pre faktorové premenné keďže zahŕňa aj počet stupňov voľnosti. `VIF` faktor poukazuje na multikolinearitu pravdepodobne medzi regresormi `sqft_living`, `sqft_above` a `sqft_basement`. Po odstránení regresoru `sqft_above` sa hodnoty `VIF`

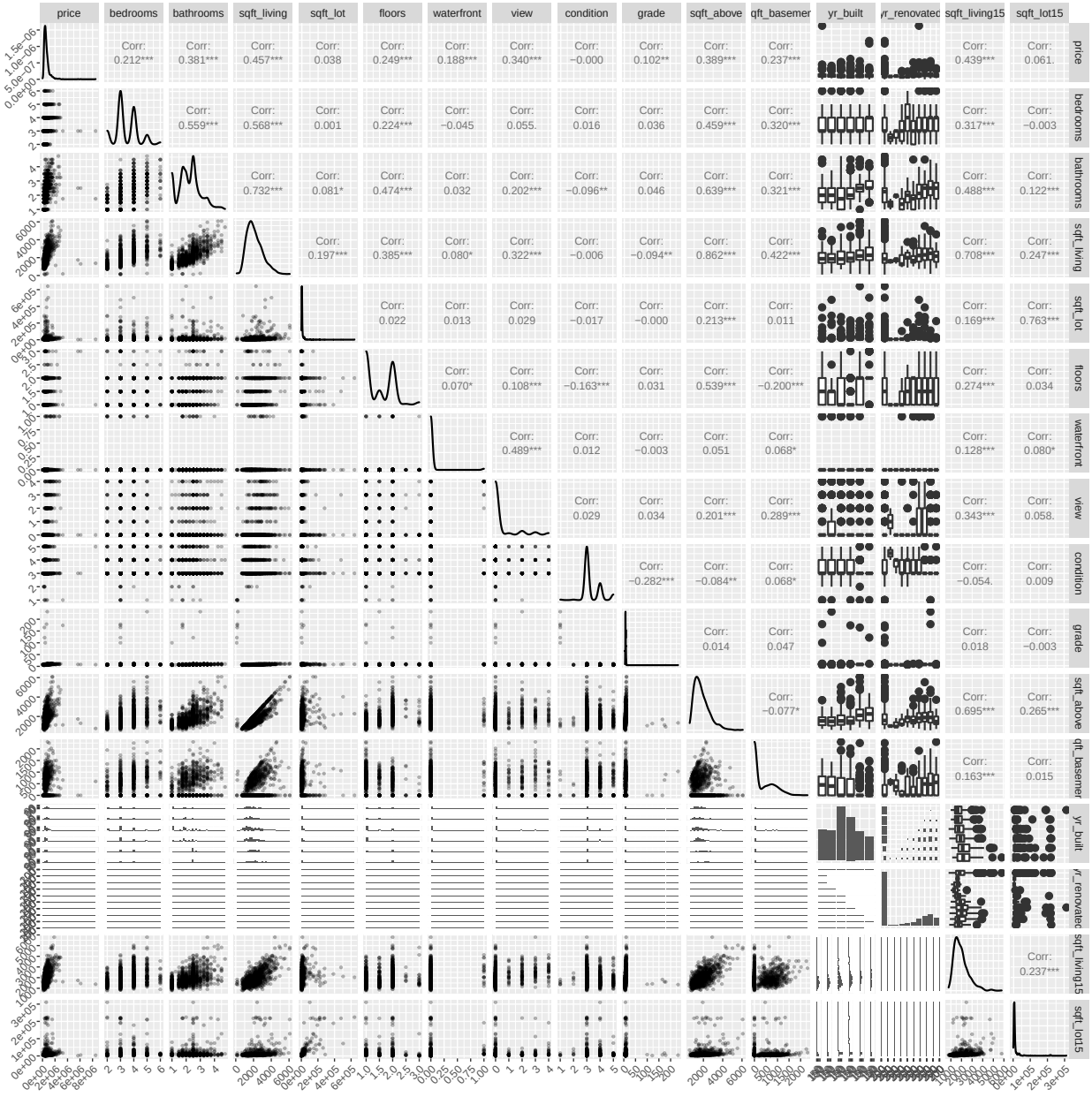

Figure 2: Scatter plot medzi numerickými premennými

Table 5: VIF faktor pre dizajnovu maticu po vynechani niektorych regresorov

	GVIF	Df	GVIF^(1/(2*Df))
bedrooms	1.65	1	1.28
bathrooms	2.59	1	1.61
floors	2.04	1	1.43
waterfront	1.40	1	1.18
view	1.73	1	1.32
condition	1.37	1	1.17
grade	1.11	1	1.06
sqft_basement	1.68	1	1.30
yr_built	3.36	5	1.13
yr_renovated	2.66	9	1.06
sqft_living15	1.66	1	1.29
sqft_lot15	1.10	1	1.05

znížili pod hodnotu 5. Ďalej predpokladám, že cena jednotlivých domov bola stanovená práve v roku 2014-15 a preto na ňu mal vplyv aktuálny stav. Na základe tejto úvahy ale aj koeficientov korelácie som sa rozhodol ďalej z datasetu vylúčiť regresory `sqft_livng` a `sqft_lot` a ponechať len korešpondujúce regresory pre rok 2015. `VIF` faktor po odstránení troch regresorov je zobrazený v tabuľke 5 pričom už nedetekujem výraznú multikolinearitu. Pre 2 kategorické premenné je lepšie analyzovať zobecnený `VIF` pričom tento ešte umocniť na druhú. Takto prepočítaný zobecnený `VIF` pre `yr_built` a `yr_renovated` nepoukazujú na možnú multikolinearitu. Kappa koeficient pre plnú dizajnovú maticu bez faktorových premenných bol odhadnutý na 606.914 čo sa po odstránení troch spomínaných regresorov znížilo na 3.313.

- Otázka 05

  Pouze pomocí trénovacích dat (tj., split == "train") a všech vybranných proměnných najděte vhodný lineární regresní model, který má za úkol predikovat co nejlépe cenu, tj. minimalizovat střední kvadratickou chybu reziduí (MSE). Jakou jinou metriku pro výběr modelu byste případně navrhovali a proč? U výsledného modelu porovnejte VIF a Kappa s původní celkovou maticí regresorů.

- Odpoveď 05

  Zhrnutie vybraného modelu je v tabuľke **??**. Podľa môjho názoru sa ako metrika pre výber môže použiť aj štatistika R^2 nakoľko tiež hodnotí relatívnu pridikčnú kvalitu modelu podobne ako `MSE`. Namiesto strátovej funkcie L_2 v prípade `MSE` môžme použiť napr. L_1 a získame tak `MAE` metriku.

- Otázka 06

  Pro Vámi vybraný model z předešlé otázky spočtěte příslušné infulenční míry. Uveďtě id pro 20 pozorování s největší hodnotou DIFF, největší hodnotou leverage (hatvalues) a největší hodnotou Cookovy vzdálenosti. (tj, 3 krát 20 hodnot). Jaká pozorování považujete za vlivná a odlehlá pozorování a proč?

- Odpoveď 06

  ... TODO

- Otázka 07

  Validujte model pomocí grafického znázornění reziduí (Residual vs Fitted, QQ-plot, Cookova vzdálenost, Leverages, ...). Identifikovali jste na základě této a předchozí otázky v datech nějaká podezřelá pozorování, která mohla vzniknout při úpravě datasetu? Doporučili byste tyto pozorování z dat odstranit?

Table 6:

	Vybraný model
(Intercept)	−471 088.4749
	p-val = 0.0001
bedrooms	−15 782.4815
	p-val = 0.4487
bathrooms	94 222.6010
	p-val = 0.0011
floors	130 906.5066
	p-val = 0.0007
waterfront	223 481.2875
	p-val = 0.0589
view	52 479.7635
	p-val = 0.0048
condition	72 288.1313
	p-val = 0.0045
grade	4348.3487
	p-val = 0.0003
sqft_basement	90.6289
	p-val = 0.0233
yr_built1921-40	−69 354.5940
	p-val = 0.1968
yr_built1941-60	−108 447.9493
	p-val = 0.0264
yr_built1961-80	−144 040.1727
	p-val = 0.0083
yr_built1981-2000	−235 007.0260
	p-val = 0.0004
yr_built2000-2015	−160 536.8253
	p-val = 0.0271
yr_renovated1931-40	−175 442.8847
	p-val = 0.5814
yr_renovated1941-50	−24 304.3635
	p-val = 0.9258
yr_renovated1951-60	−17 893.6784
	p-val = 0.8865
yr_renovated1961-70	−104 574.8393
	p-val = 0.3323
yr_renovated1971-80	−49 115.4737
	p-val = 0.5779
yr_renovated1981-90	126 771.0459
	p-val = 0.0448
yr_renovated1991-2000	100 378.2751
	p-val = 0.0721
yr_renovated2001-2010	137 684.5440
	p-val = 0.0100
yr_renovated2011-2015	88 127.0416
	p-val = 0.1392
sqft_living15	260.7181
	p-val = 0.0000
sqft_lot15	−0.7729
	p-val = 0.1350
Num.Obs.	1000
R2	0.328
RMSE	437 360.16

8

1.6 Train, test, validation . . . :

- Otázka 08

  Pokud jste se rozhodli z dat odstranit nějaká pozorování, tak dále pracujtes s vyfiltrovaným datasetem a přetrénujte model z otázky 5. A spočtěte pro tento model R^2 statistiku a MSE jak na trénovacích tak testovacích datech (split == "test").

- Otázka 09

  Pomocí hřebenové regrese (případně pomocí LASSO a Elastic Net) zkuste najít nejlepší hyperparametr(y) tak, aby výsledný model měl co nejmenší MSE na testovacích datech. K odhadu regresních koeficientů použijte ale pouze trénovací data.

- Otázka 10

  Vyberte výsledný model a porovnejte MSE a R^2 na trénovacích, testovacích a validačních datech. Co z těchto hodnot usuzujete o kvalitě modelu a případném přetrénování? Je váš model vhodný pro predikci cen nemovitostí v okolí King County? Pokud ano, má tato predikce nějaká omezeni?