

### 3. zápočtová úloha z 01RAD

Jiří Franc

2021-12-16

### 3. zápočtová úloha z 01RAD

#### Popis úlohy

Datový soubor vychází z datasetu `House Sales in King County, USA`, který je k nalezení například na [kaggle.com](https://www.kaggle.com/datasets/kc_house), nebo v knihovně `library(moderndiver)` 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.

```
## [1] "M:/01RAD/asignments/2021"
```

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

## Podmínky a body

Úkol i protokol vypracujte samostatně. Pokud na řešení nějaké úlohy budete přesto s někým spolupracovat, radit se, nespomeňte 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 mohou případně individuálně udělit za řešení mini domácích úkolů z jednotlivých hodin.

## Odevzdání

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

## 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?

- 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?

- Otázka 03

Zkontrolujte pro 4 vybrané 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?

### Lineární model (použijte pouze trénovací data, tj. `split == "train"`):

- Otázka 04

Spočtete korelace mezi jednotlivými regressory a graficky je znázorníte. Dále spočtete číslo podmíněnosti matice regresorů Kappa a VIF. Pokud se v datech vyskytuje znatelná multicollinearita, rozhodněte jaké proměnné a proč použijete v následném lineárním modelu.

- Otázka 05

Pouze pomocí trénovacích dat (tj., `split == "train"`) a všech vybraných 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.

- Otázka 06

Pro Vámi vybraný model z předešlé otázky spočtete příslušné inflační míry. Uveďte 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?

- 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?

### Train, test, validation ...:

- Otázka 08

Pokud jste se rozhodli z dat odstranit nějaká pozorování, tak dále pracujte s vyfiltrovaným datasetem a přetrénujte model z otázky 5. A spočtete 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 omezení?