Katedra Geomatiky

Strojové učení v dálkovém průzkumu Země

155YUSU

Lekce 6: Návrh a řešení projektu strojového učení

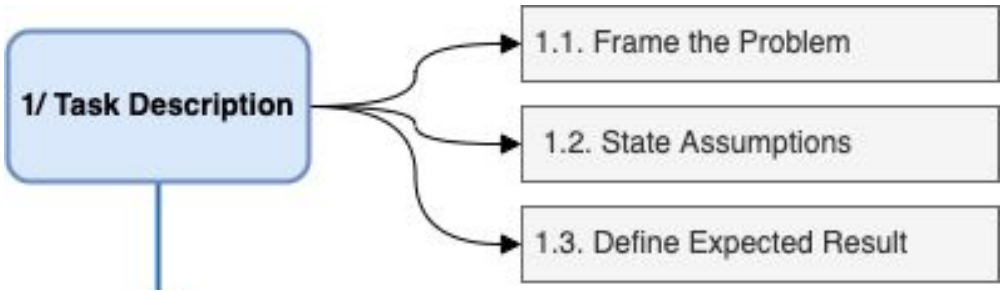
Přednášející: Lukáš Brodský lukas.brodsky@natur.cuni.cz , Martin Landa martin.landa@fsv.cvut.cz a Ondřej Pešek ondrej.pesek@fsv.cvut.cz

Pracovní postup

Pracovní postup

1. Popis problému
2. Explorační analýza dat (Exploratory Data Analysis – EDA)
3. Příprava dat
4. Výběr a trénování modelu
5. Ladění modelu (fine-tuning)
6. Interpretace výsledků

1. Popis problému



Problém

1. Popiš hlavní úkol
   Definuj předpoklady problému na základě dostupných informací;
   -> zkontroluj předpoklady po EDA
2. Definuj předpokládané výsledky
3. Zvol metriky přesnosti

Jsou-li data příliš velká, lze trénování provést jako dávkové (batch learning) na několika serverech

Problém

Příklad:

DEMO: Lesní požáry či ceny nemovitostí v Kalifornii

Metriky přesnosti

- Výběr metrik přesnosti ovlivňuje, jaké charakteristiky považujeme ve výsledcích za důležité
- Volba modelu se může odvíjet právě od zvolených metrik přesnosti

Metriky přesnosti – Regrese

- **RMSE** (střední kvadratická chyba – root mean square error/distance) – určuje směrodatnou odchylku;
- Ke zvážení také **MAE** (střední absolutní chyba – mean absolute error);

RMSE vs. MAE

- Obé popisuje vzdálenost mezi vektory;
- MAE ~ l_1 norma (manhattanská norma – vzdálenost mezi dvěma body ve městě, v němž se lze pohybovat pouze po pravoúhlých blocích)
- RMSE ~ l_2 norma (Euclidean norm)
- RMSE je citlivější na odlehlé hodnoty než MAE;
- Jsou-li odlehlé hodnoty vzácné, RMSE bývá preferováno;

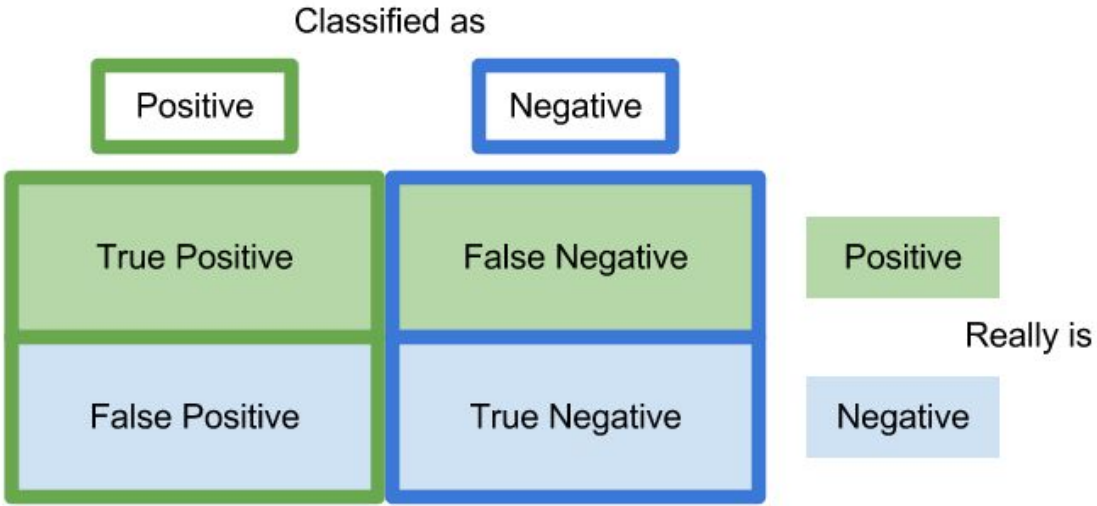
Klasifikace

- Proces kategorizace
- Proces, při kterém jsou objekty/pixely rozpoznávány, zařazovány a je jim porozuměno.

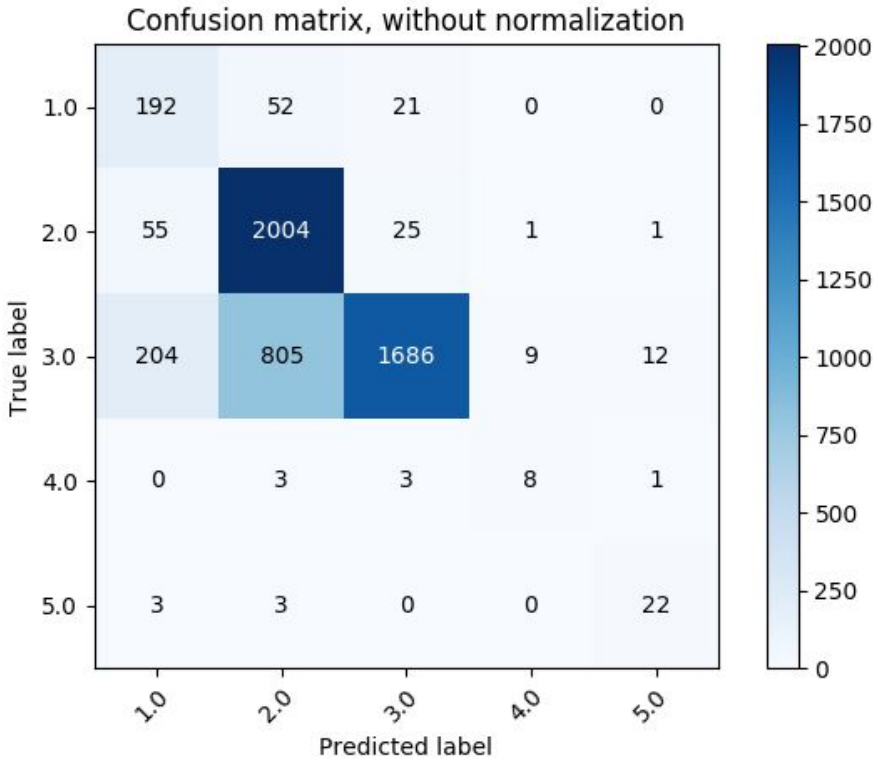
- Matice záměn
- Celková přesnost
- F1-skóre
- …

Matice záměn (confusion matrix)

Souhrn výsledků predikce ukazující, mezi kterými třídami dochází k záměnám



Víceprvková matice záměn



Metriky přesnosti

- Celková přesnost: podíl správně určených prvků z celkového množství provedených odhadů

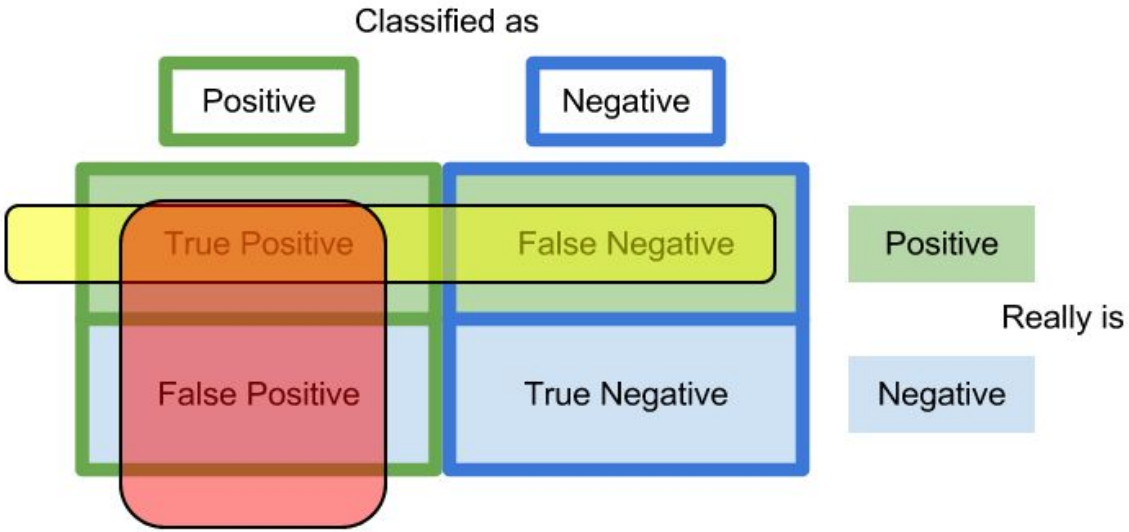
  celková_přesnost = (skutečně_pozitivní + skutečně_negativní) / vše
  celková_přesnost = (TP + TN) / (TP + FP + TN + FN)

  -> především v případě “šikmých” (skewed) datasetů nevhodné (šikmý dataset – některé třídy jsou mnohem častější než jiné)

Preciznost (precision)

Preciznost: přesnost pozitivních odhadů;

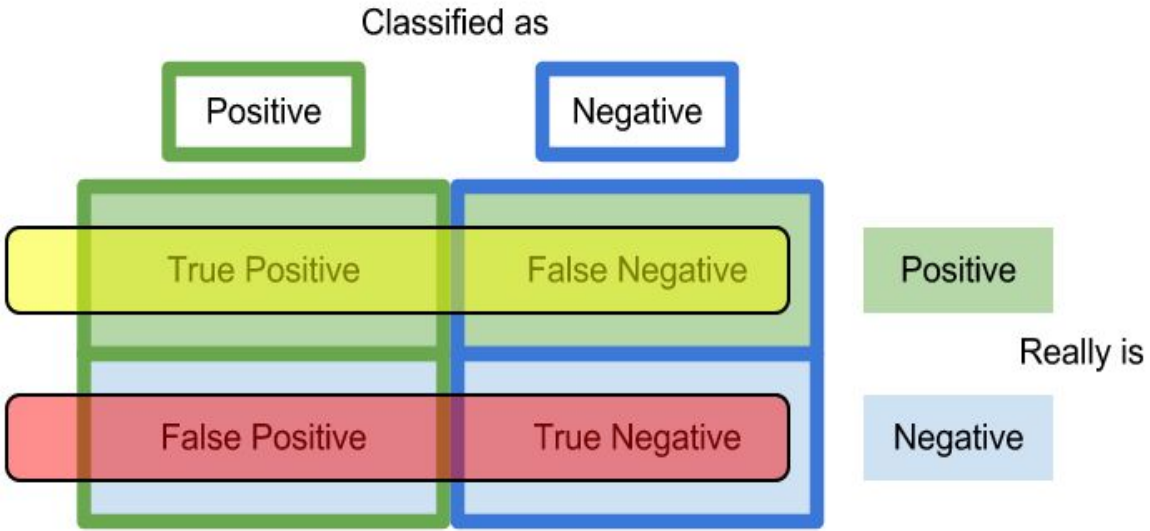
přesnost = skutečně_pozitivní / (všechny_pozitivní) = TP / (TP + FP)



Výtěžnost / úplnost / senzitivita (recall / sensitivity)

Senzitivita: podíl skutečně pozitivních vůči všem pozitivním

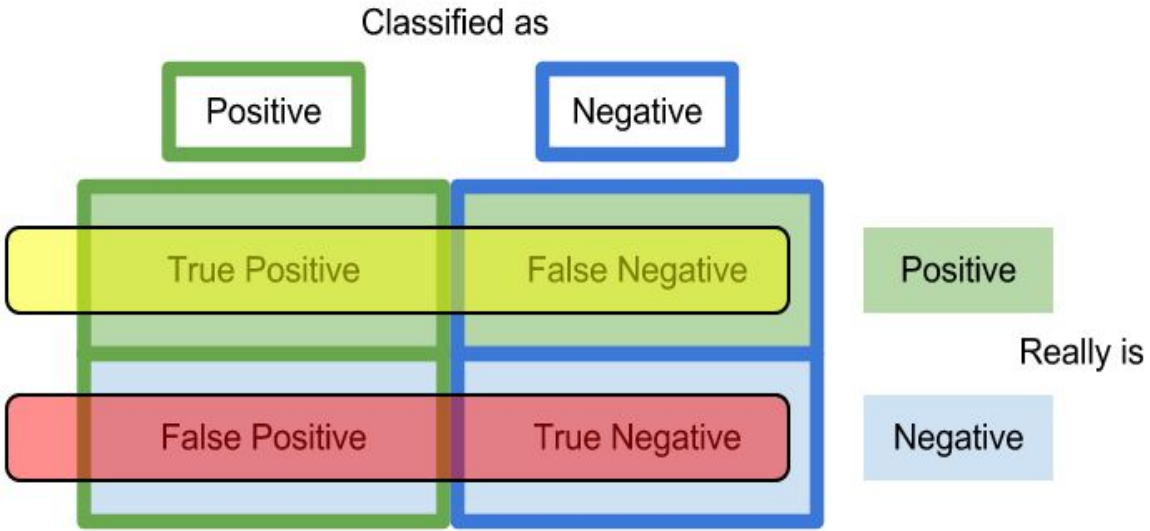
senzitivita = skutečně_pozitivní / (skutečně_pozitivní + falešně_negativní) = TP / (TP + FN)



Specificita (specificity)

Specificita: podíl skutečně negativních vůči všem negativním

senzitivita = skutečně_negativní / (skutečně_negativní + falešně_pozitivní) = TN / (TN + FP)



- **Preciznost - kolik z pozitivně určených prvků bylo skutečně pozitivních.** Test může být zkreslený a přehnaně optimistický, vrátí-li se pozitivní odhad pouze tam, kde si je model nejjistější
- **Senzitivita - jak dobrý je model v detekci pozitivních prvků.** Test může být zkreslený a přehnaně optimistický, vrátí-li se pozitivní odhad všude
- **Specificita - jak dobrý je model ve vyhýbání se falešným alarmům.** Test může být zkreslený a přehnaně optimistický, vrátí-li se negativní odhad všude

F1-skóre (F1-score)

F1-skóre je **vážený průměr preciznosti a senzitivity**

F1-skóre interpretuje, jak precizní, ale zároveň robustní model je

$$F_1 = \frac{2}{\frac{1}{\text{precision}} + \frac{1}{\text{recall}}} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} = \frac{TP}{TP + \frac{FN + FP}{2}}$$

Problémy!

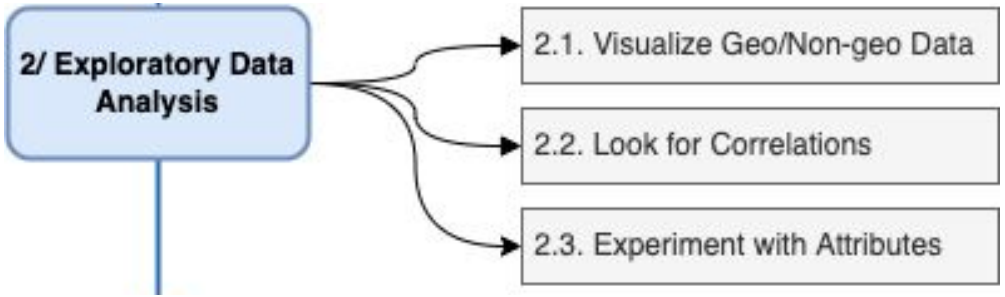
- Scénář 1: klasifikace zemského povrchu (N tříd) -> vyvážená preciznost a senzitivita
- Scénář 2: klasifikace videí vhodných pro děti -> preference vysoké preciznosti
- Scénář 3: klasifikátor detekce zlodějů na bezpečnostních záznamech -> prefetence vysoké senzitivity

Perspektiva přesnosti v DPZ

Matice záměn – všimněte si prohozených řádků a sloupců

		Reference Data			
		Water	Forest	Urban	Total
Classified Data	Water	21	6	0	**27**
	Forest	5	31	1	**37**
	Urban	7	2	22	**31**
	Total	**33**	**39**	**23**	***95***

2. Explorační analýza dat (EDA)



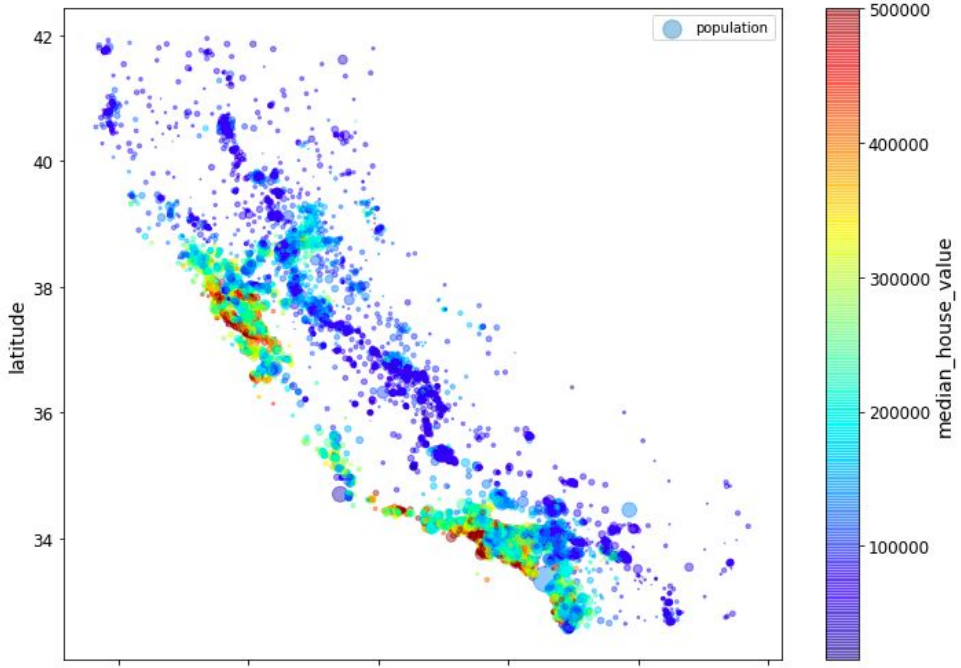
Příprava dat

- Podívejte se v rychlosti na strukturu dat
- Nechybí nějaká data?

Je současný dataset vhodný?

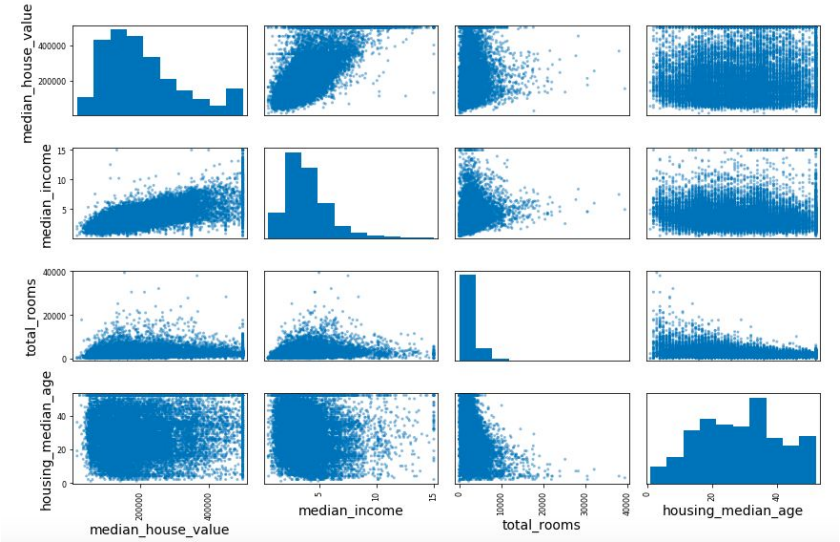
Prostorová vizualizace

Umístění?
Prostorová autokorelace?

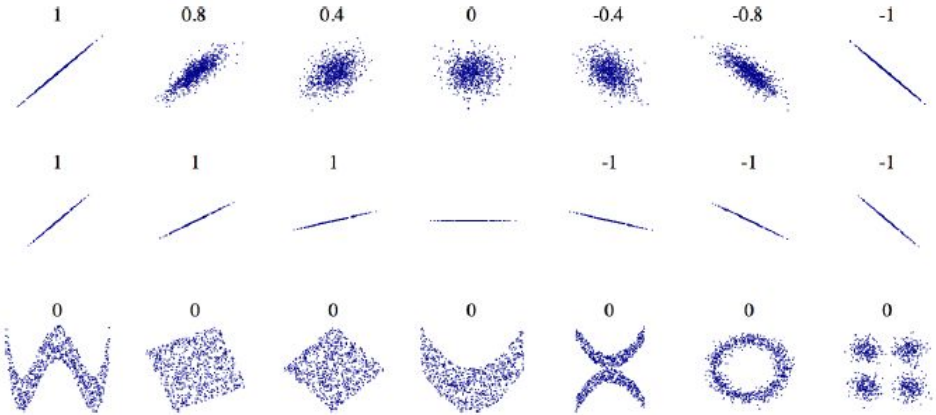


Nejsou v datech vzory (patterns)?

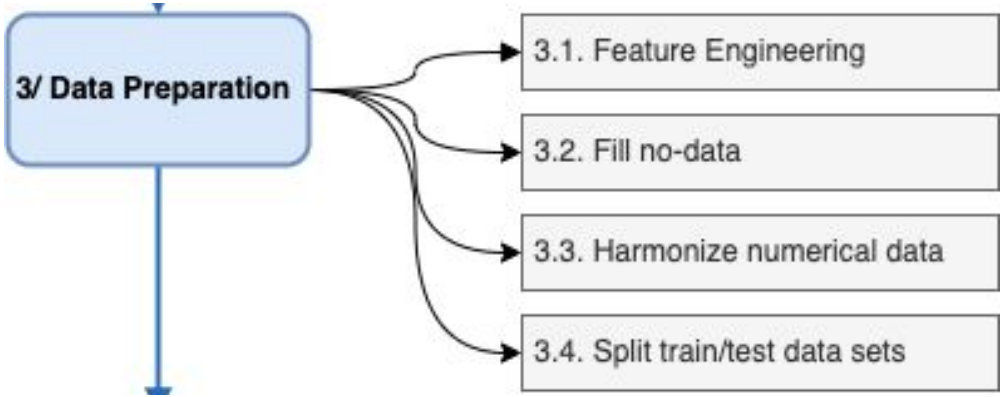
Korelace v atributech?
Co je korelace?



Hledejte nelineární vzory



3. Příprava dat



Připravte data

3.1 Inženýrství příznaků (feature engineering): transformace surových dat do formy datasetu;

- scikit-Learn poskytuje mnoho užitečných nástrojů (PCA, …);
- Uživatelé mohou napsat vlastní funkce, které data čistí nebo kombinují (např. NDVI);

- Cíl: vytvořit informativní příznaky (příznaky hodnotné pro predikci);

Připravte data

Inženýrství příznaků

- **normalizace:** min-max škálování (scaling) přeškáluje hodnoty do rozsahu 0 až 1 (*MinMaxScaler*);

$$\bar{x}^{(j)} = \frac{x^{(j)} - min^{(j)}}{max^{(j)} - min^{(j)}},$$

- **standardizace**: průměr 0, rozptyl 1 (*StandardScaler*);

$$\hat{x}^{(j)} = \frac{x^{(j)} - \mu^{(j)}}{\sigma^{(j)}}.$$

Připravte data

3.2 Čištění dat / zaplnění mezer

- **Mezery**: odstraňte *None* hodnoty, zaplňte *medianem* nebo nalezněte lepší řešení (např. geostatistika);

Připravte data

3.3 Práce s datovými typy: Textové a kategorické atributy

Odstraňte textové atributy užitím objektu **OrdinalEncoder** nebo **OneHotEncoder**

red = [1, 0, 0]
yellow = [0, 1, 0]
green = [0, 0, 1]

Připravte data - tvorba validační a testovací sady

- Mimo trénovací dataset: 30 %
  - validační set (15%) -> ladění hyperparametrů!
  - testovací set (15 %) -> test generalizační chyby!

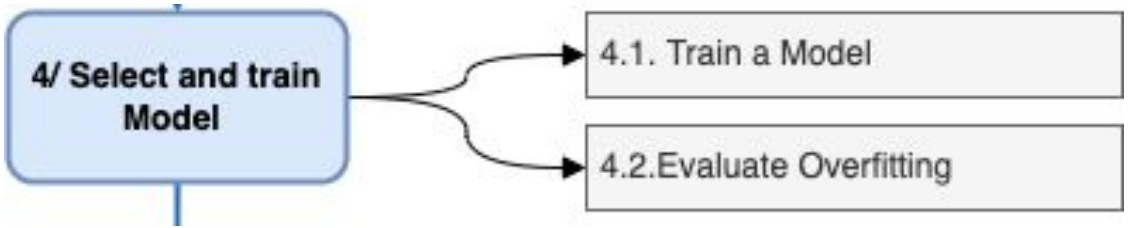
1/ Náhodné vzorkování (*jaký je risk?*)

2/ Složené (stratifikované) vzorkování (*stratified sampling* – stratifikace tvaru příznaků, ale také prostorového umístění);

3/ Vzorkování podle latinské hyperkrychle (*latin hypercube sampling – LHS – N-rooks sampling* – vícerozměrné problémy);

4/ Prostorové vzorkování (*spatial sampling*)

4. Výběr a trénování modelu



Výběr učicího algoritmu

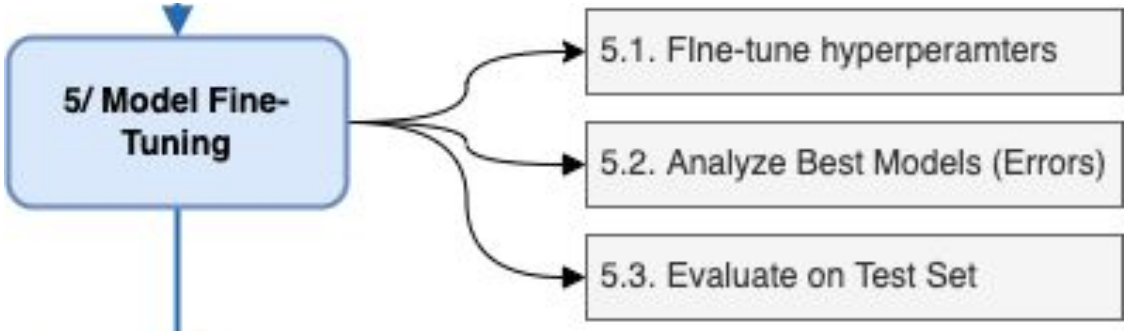
1. Nelinearita dat;
2. kategorické vs. numerické příznaky;
3. vysvětlitelnost (*explainability*; černé vs. bílé skříňky; důležitost příznaků, jednoduchost, …);
4. počet příznaků vs. počet trénovacích příkladů;
5. rychlost učení;
6. rychlost predikce;

Křížová validace (cross-validation)

- rozdělte trénovací dataset na **k** (*k-folds*) částí
- vyzkoušejte různé kombinace těchto částí
  - 1 část je v testovacím datasetu, zbytek (**k-1**) v trénovacím


Training set
Training folds
Test fold
1st iteration
2nd iteration
3rd iteration
10th iteration
E1
E2
E3
E10

5. Ladění modelu (fine-tuning)



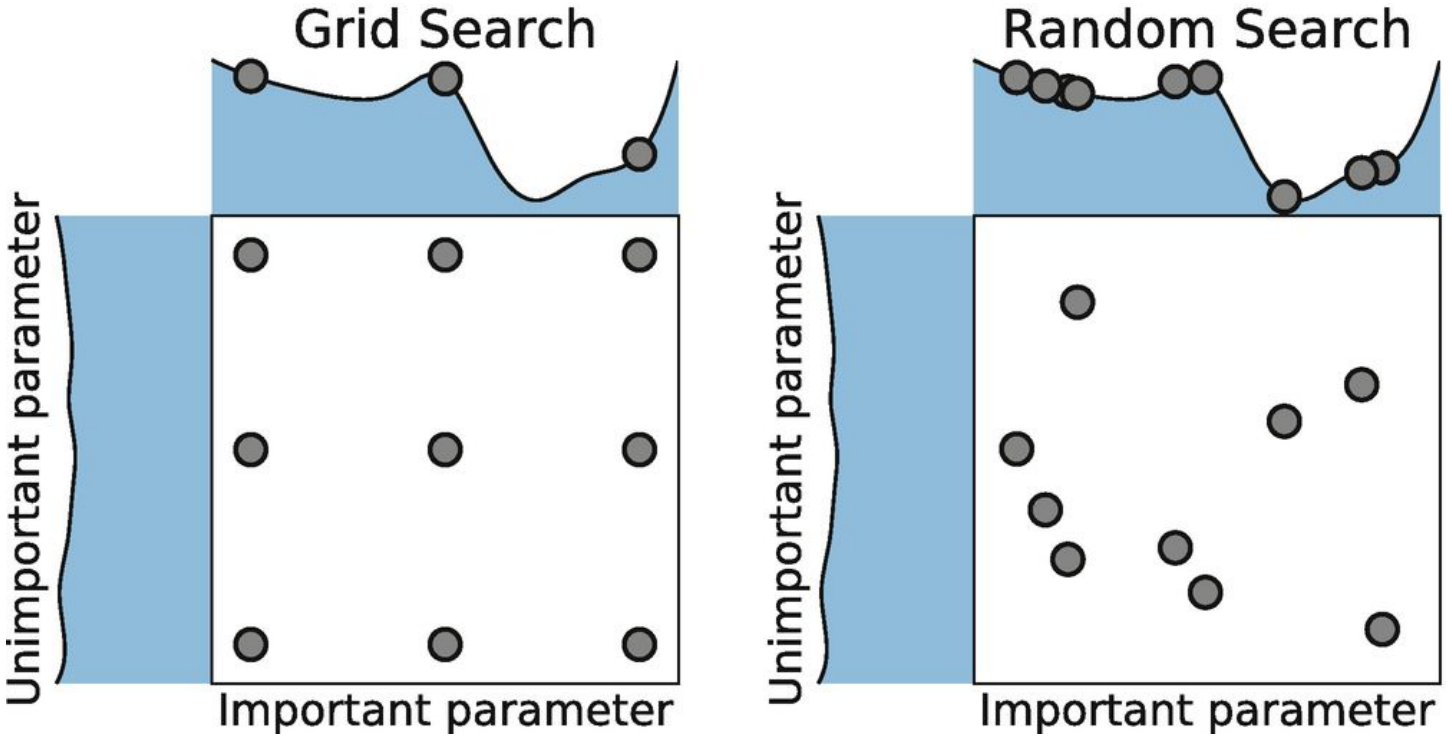
Laďte model

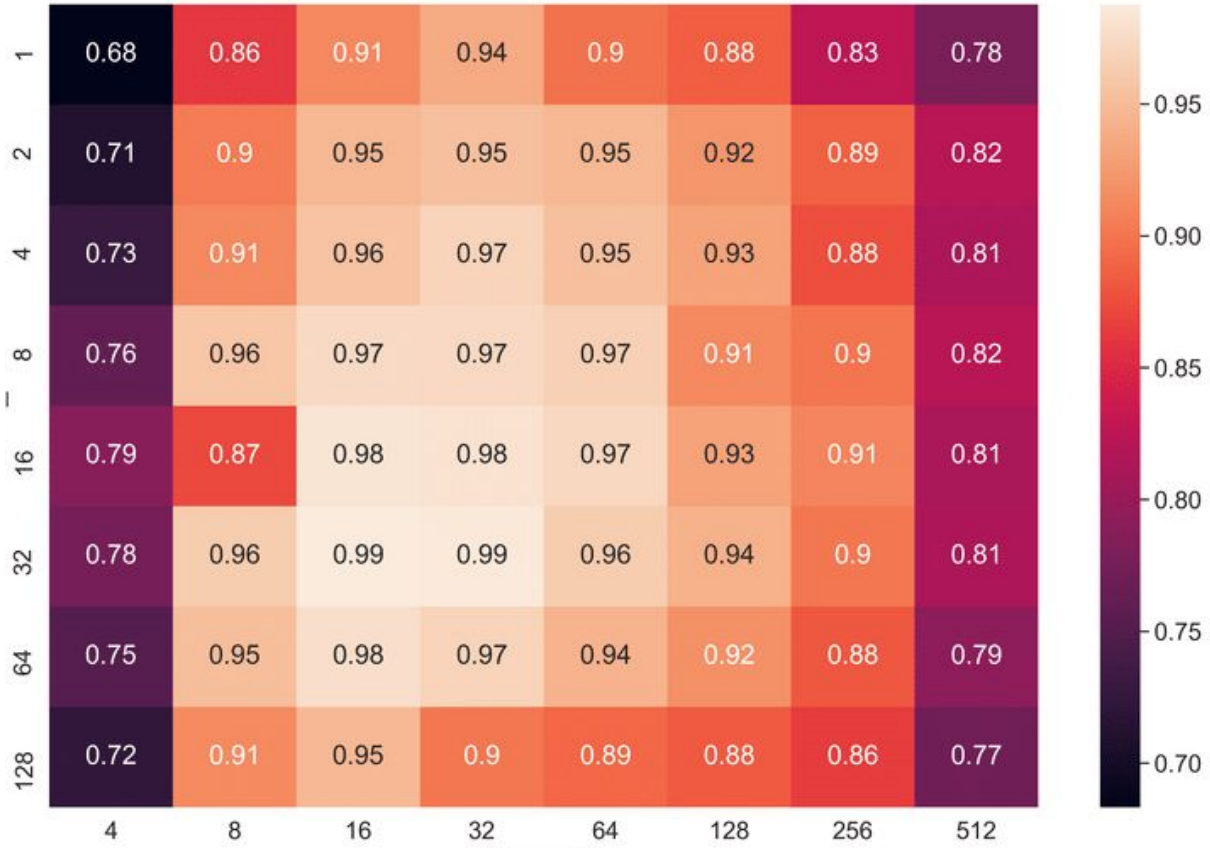
5.1 Laďte hyperparametry modelu:

- Systematické prohledávání (*grid search*)
- náhodnostní prohledávání (*randomized search*)

nebo spusťte ansámblové metody

Systematické vs. náhodnostní prohledávání

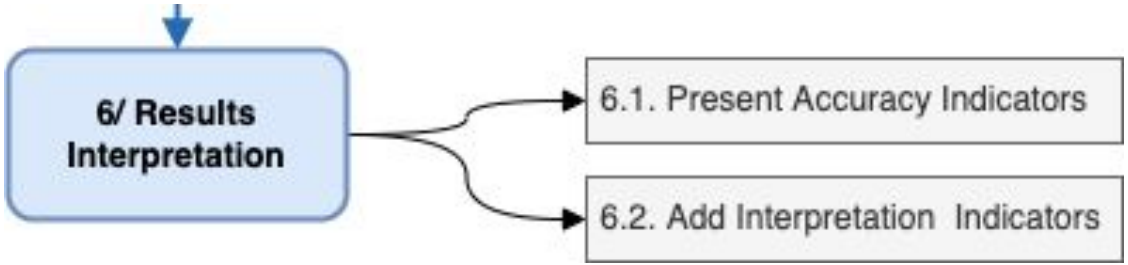



1
2
4
8
16
32
64
128
0.68 0.86 0.91 0.94 0.9 0.88 0.83 0.78
0.71 0.9 0.95 0.95 0.95 0.92 0.89 0.82
0.73 0.91 0.96 0.97 0.95 0.93 0.88 0.81
0.76 0.96 0.97 0.97 0.97 0.91 0.9 0.82
0.79 0.87 0.98 0.98 0.97 0.93 0.91 0.81
0.78 0.96 0.99 0.99 0.96 0.94 0.9 0.81
0.75 0.95 0.98 0.97 0.94 0.92 0.88 0.79
0.72 0.91 0.95 0.9 0.89 0.88 0.86 0.77
4 8 16 32 64 128 256 512
0.95
0.90
0.85
0.80
0.75
0.70

Laďte model

1. Předefinujte vhodnou metriku přesnosti!
2. Analyzujte nejlepší modely a jejich chyby
3. Zhodnoťte váš systém na testovacím setu, abyste předešli přeučení (*overfitting*)

6. Interpretace výsledků

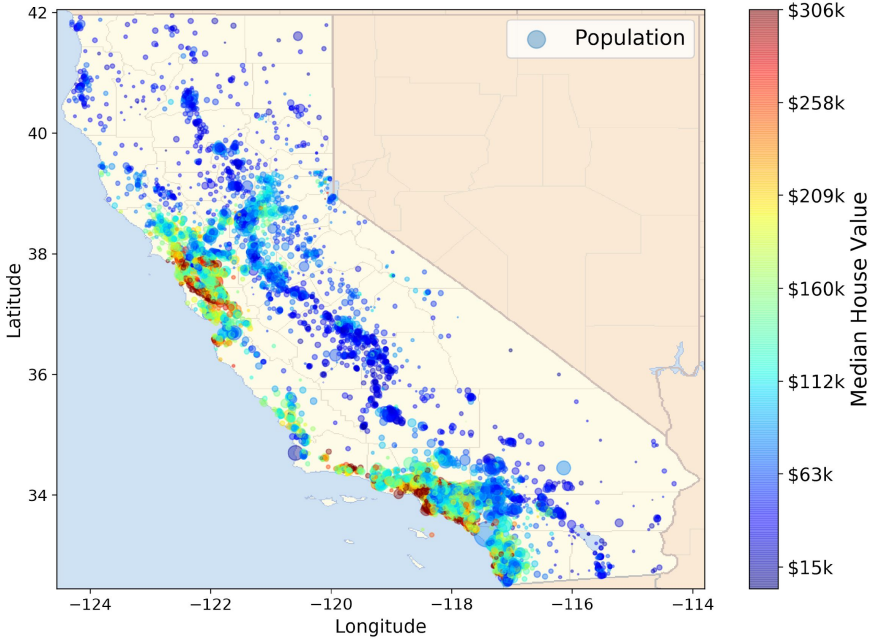


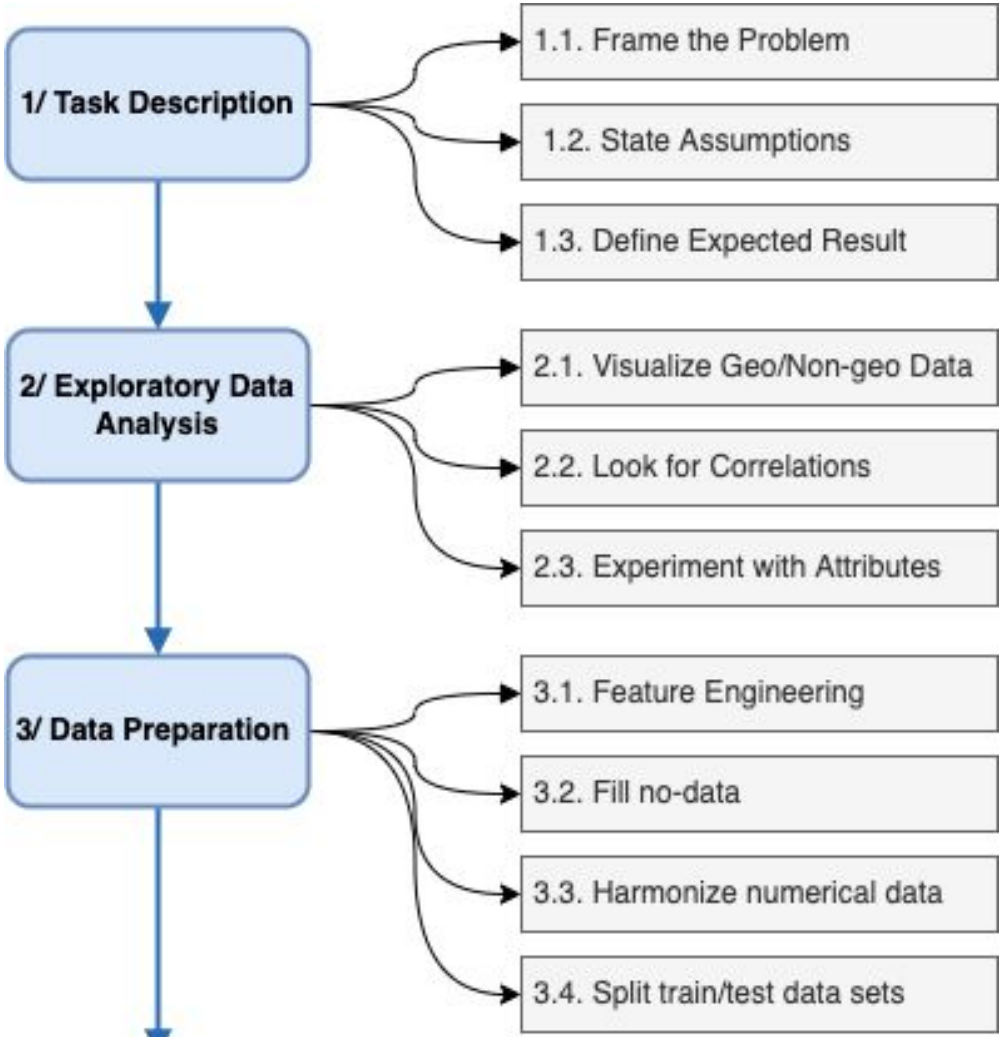
Interpretace výsledků

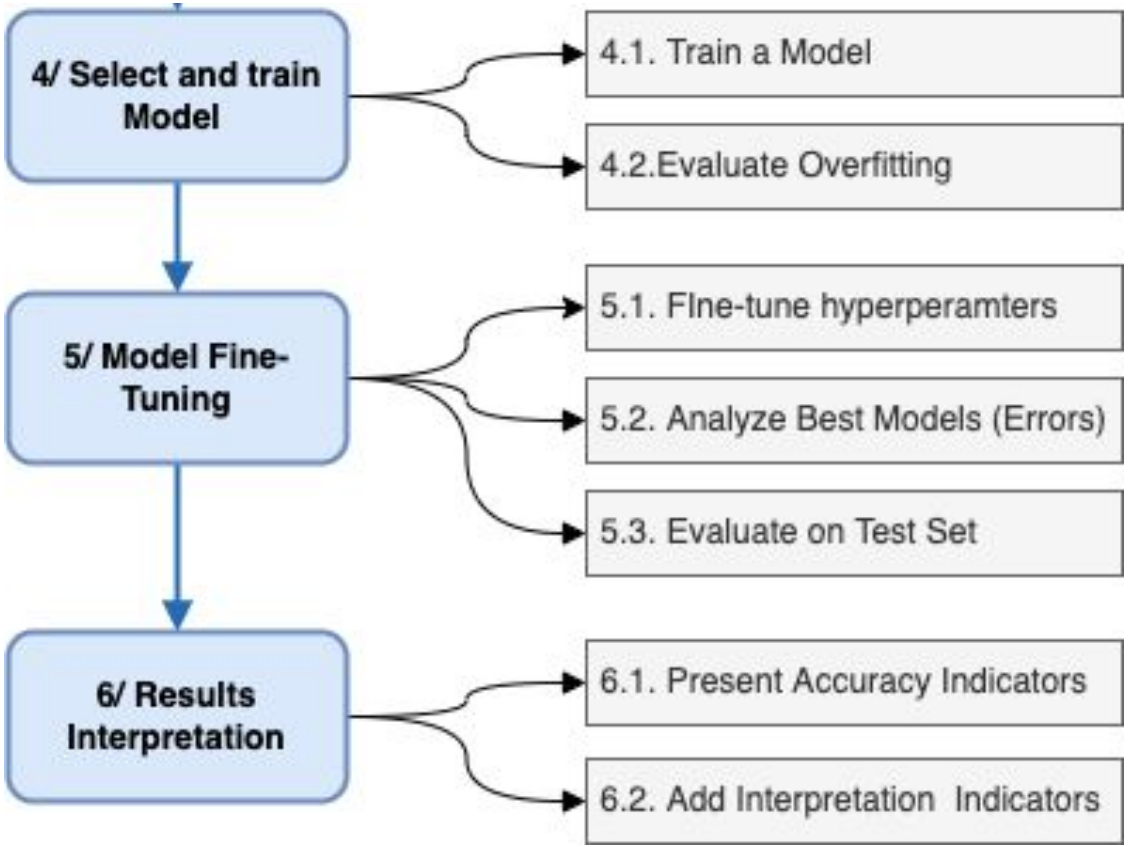
- Indikátory a vizualizace

RMSE = … (jednotka)

Chyba = 17 %




1/ Task Description
1.1. Frame the Problem
1.2. State Assumptions
1.3. Define Expected Result
2/ Exploratory Data Analysis
2.1. Visualize Geo/Non-geo Data
2.2. Look for Correlations
2.3. Experiment with Attributes
3/ Data Preparation
3.1. Feature Engineering
3.2. Fill no-data
3.3. Harmonize numerical data
3.4. Split train/test data sets


4/ Select and train Model
4.1. Train a Model
4.2.Evaluate Overfitting
5/ Model Fine-Tuning
5.1. FIne-tune hyperperamters
5.2. Analyze Best Models (Errors)
5.3. Evaluate on Test Set
6/ Results Interpretation
6.1. Present Accuracy Indicators
6.2. Add Interpretation Indicators

OTÁZKY?