Regrese znečištění v Pekingu
MAD 3 projekt

Bc. Moravec Vojtěch

ZS 2019/2020

Vysoká škola báňská – Technická univerzita Ostrava

Cíl

- Provést explorační analýzu
- Připravit datasety pro regresi
- Provést a ohodnotit regresi

Informace ohledně datasetu

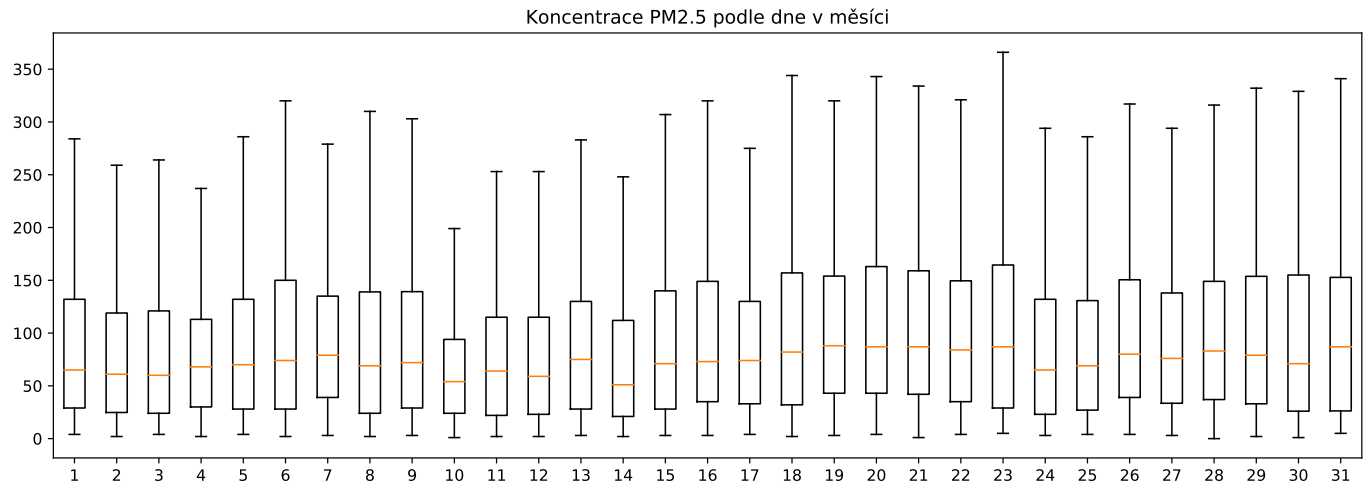
- 1 Transakce = Záznam o měření koncentrace pevných části $PM_{2.5}$ ve vzduchu ($\mu g/m^3$)
- Data od 1.1.2010 až do 31.12.2014
- 43 824 záznamů a 12 atributů
  - 6 numerických atributů + 1 cílový
  - zbylé atributy reprezentují datum a čas
  - z data (den, měsíc, rok) jsme vytvořili nový atribut "den v týdnu"

Časové atributy

- Jak pracovat s atributy dne, měsíce, roku, hodiny, dne v týdnu?
- Zpracovat jako numerický nebo kategoriální atribut?
- Podíváme se na závilost cílového atributu, vzhledem k těmto atributům
  - Nalezneme-li závislost (např. koncentrace roste s měsícem) - numerický
  - Jinak kategoriální a provedeme jejich binarizaci

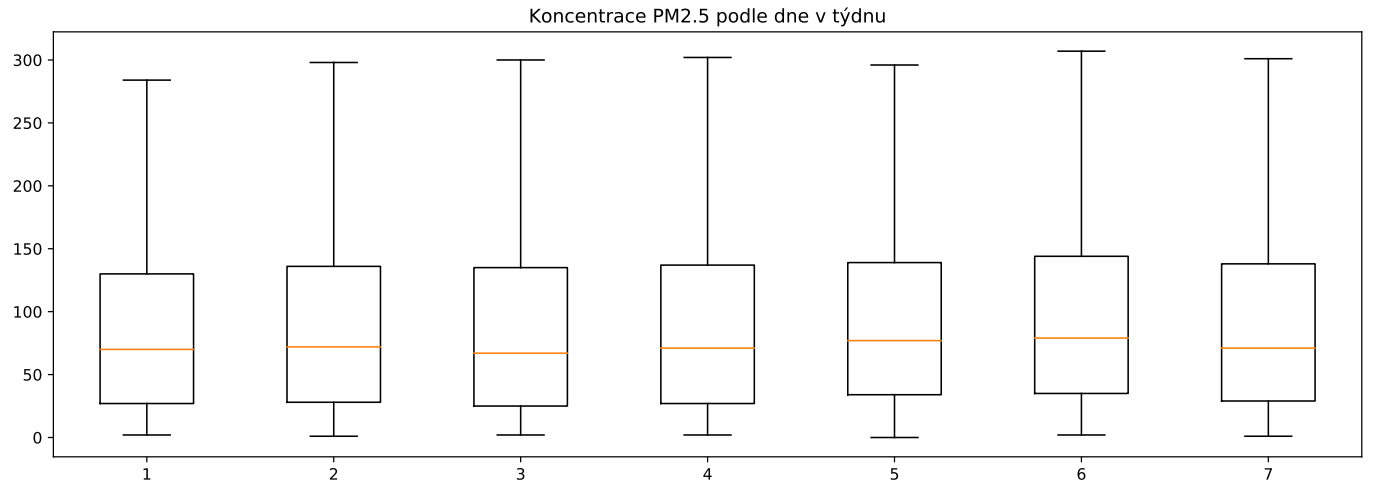
Koncentrace vzhledem k měsíci

Nevidíme žádnou závislost, že koncentrace klesá nebo roste spolu s měsícem.



Koncentrace vzhledem k dnu týdne

Koncentrace $PM_{2.5}$ je nezávislá na dnu týdne.



Odlehlé pozorování

Odlehlé pozorování nalezené pro:

- Rychlost větru (4 893) [1]
- Doba deště (1 739) [1]
- Doba sněžení (368) [1]
- Koncentrace $PM_{2.5}$ (1 773) - všechny odstraněny

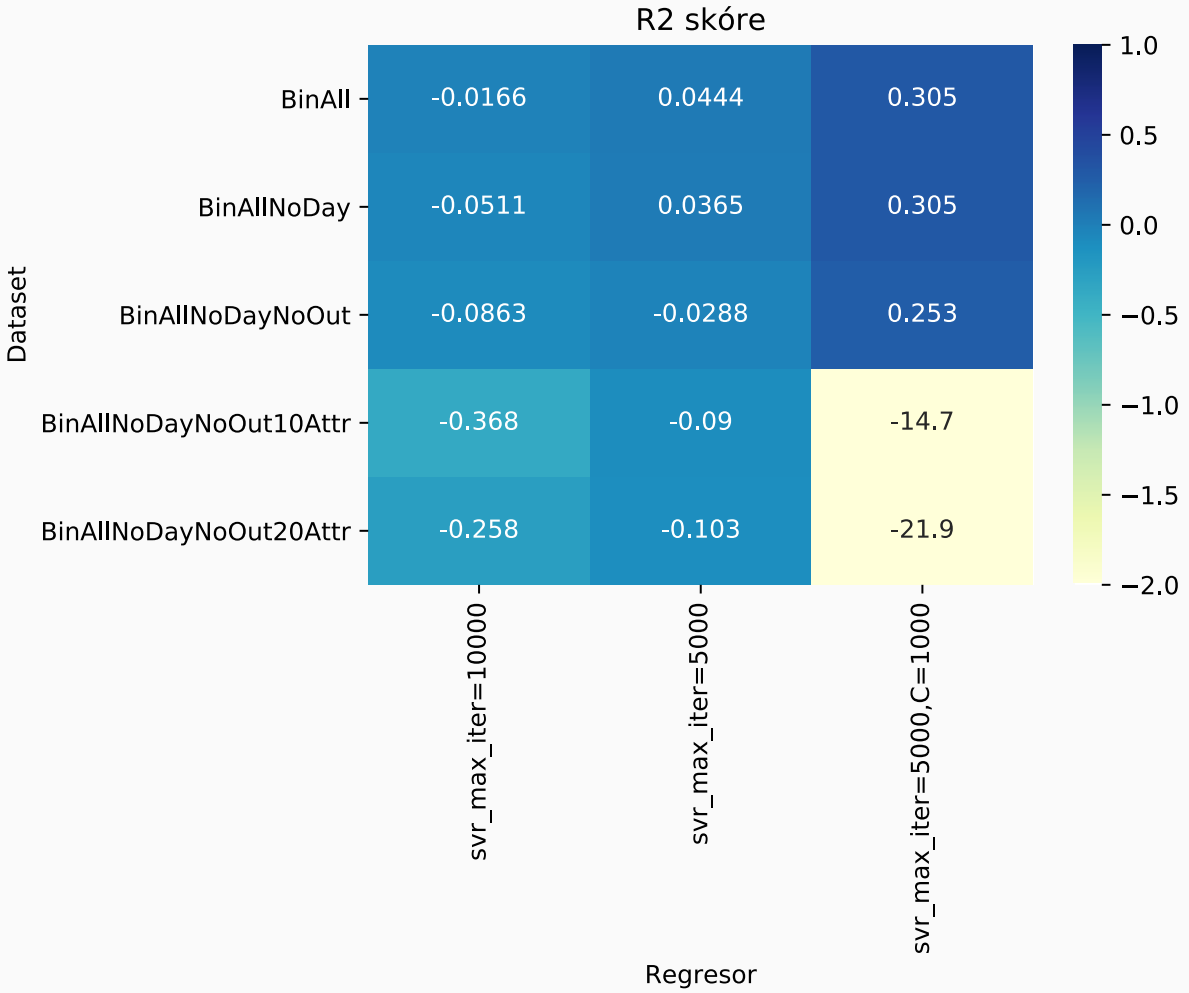
[1] Odlehlé pozorování byly ponechýny v některých datasetech

Datasety pro regresi

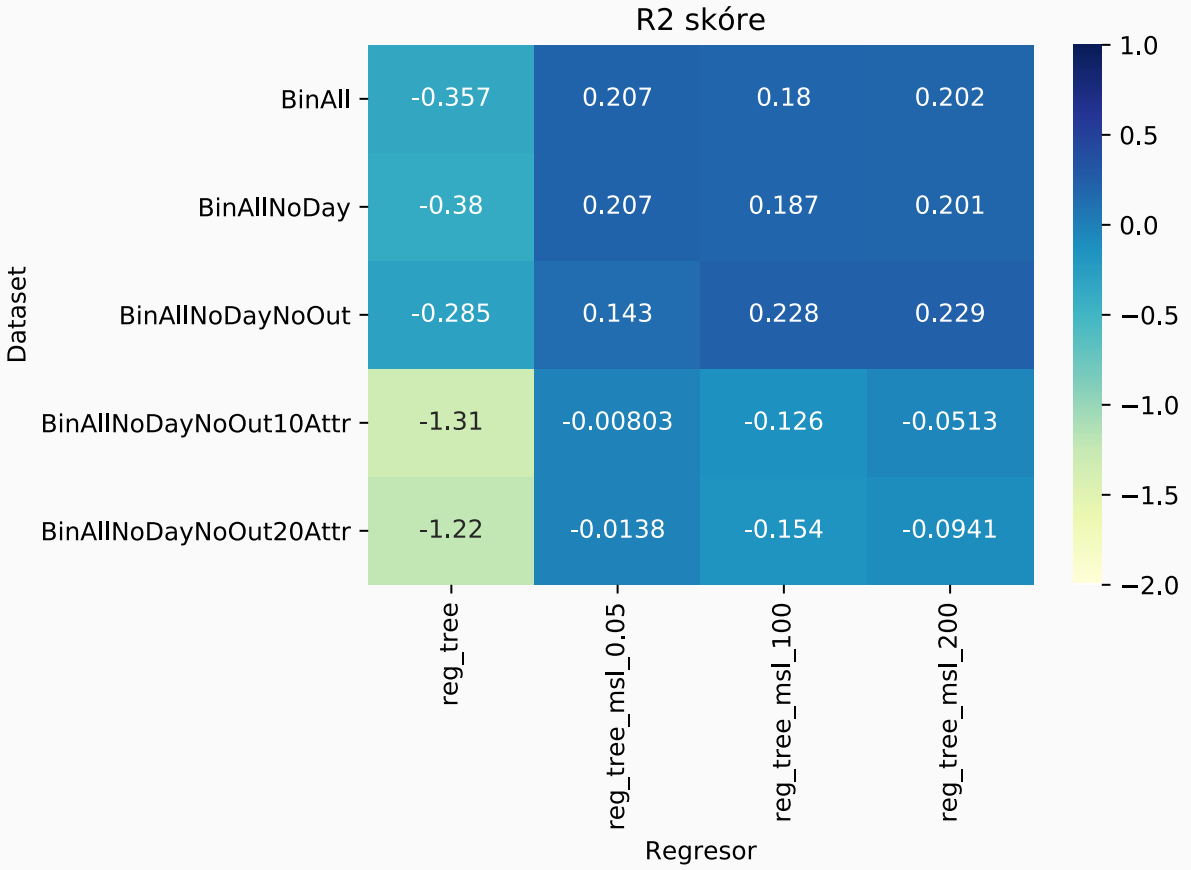
- Ve všech datasetech byla provedena normalizace hodnot do rozmezí 0,0 až 1,0

Dataset	Počet transakcí	Počet attributů
df_binAll	39984	89
df_binNoDay	39984	58
df_noOut_binNoDay	33512	58
df_noOut_binNoDay_20attr	33512	20
df_noOut_binNoDay_10attr	33512	10

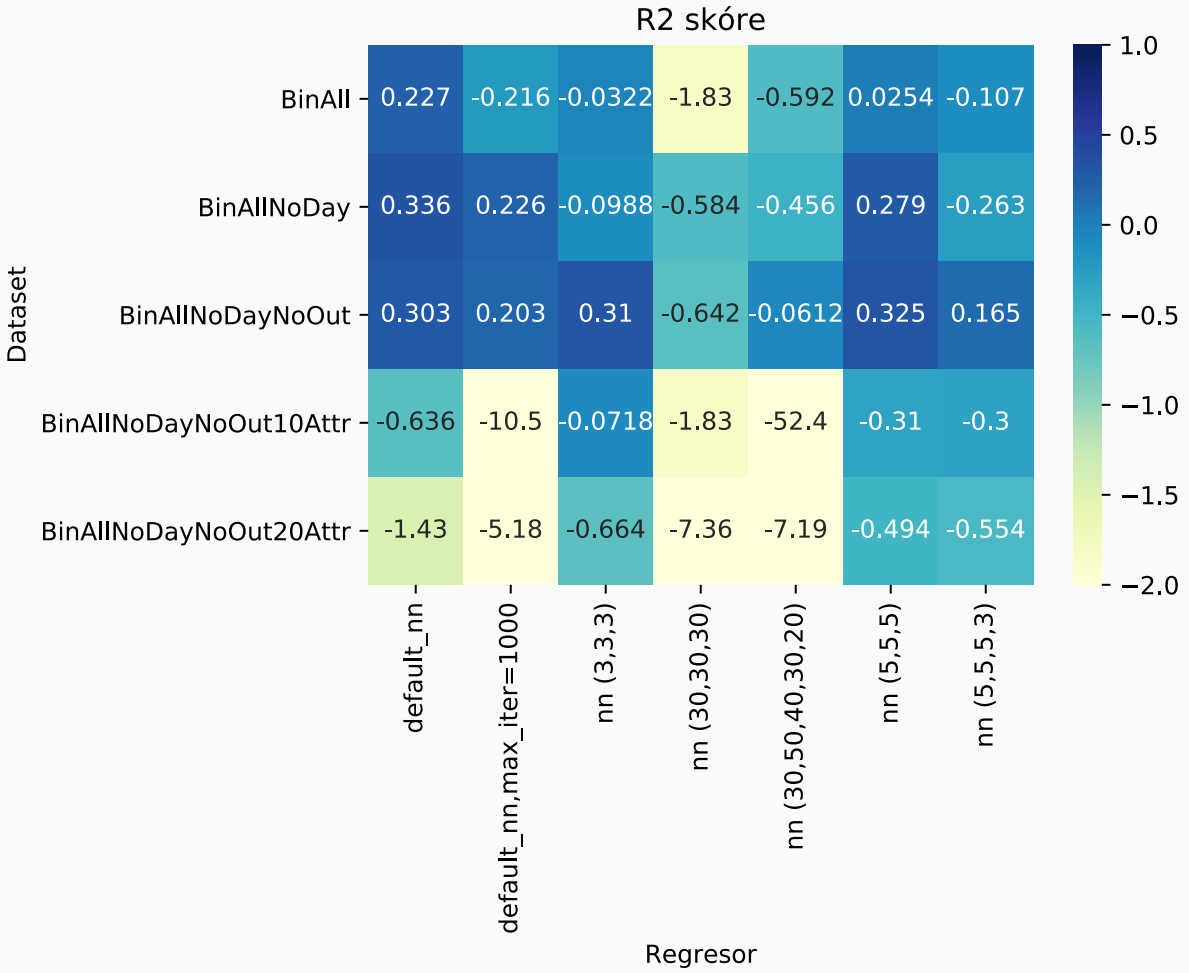
Regrese pomocí SVR



Regrese pomocí DecisionTreeRegressor



Regrese pomocí MLPRegressor



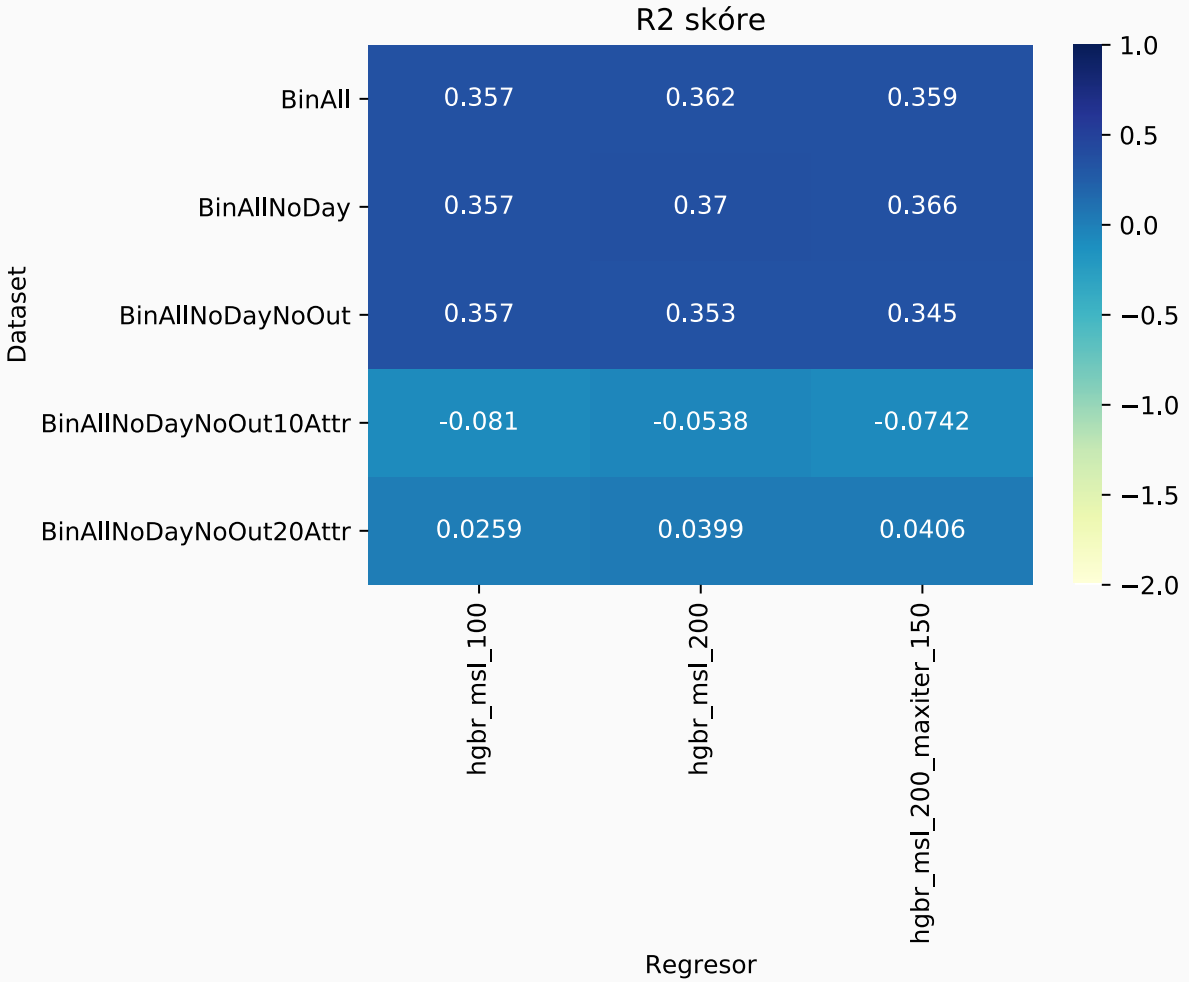
Sourhn výsledků

Algoritmus	Čas učení (s)	R^2	MSE
SVR	27,209	0,305	3326,722
DecisionTreeRegressor	0,078	0,229	3731,770
MLPRegressor	18,880	0,336	3172,276

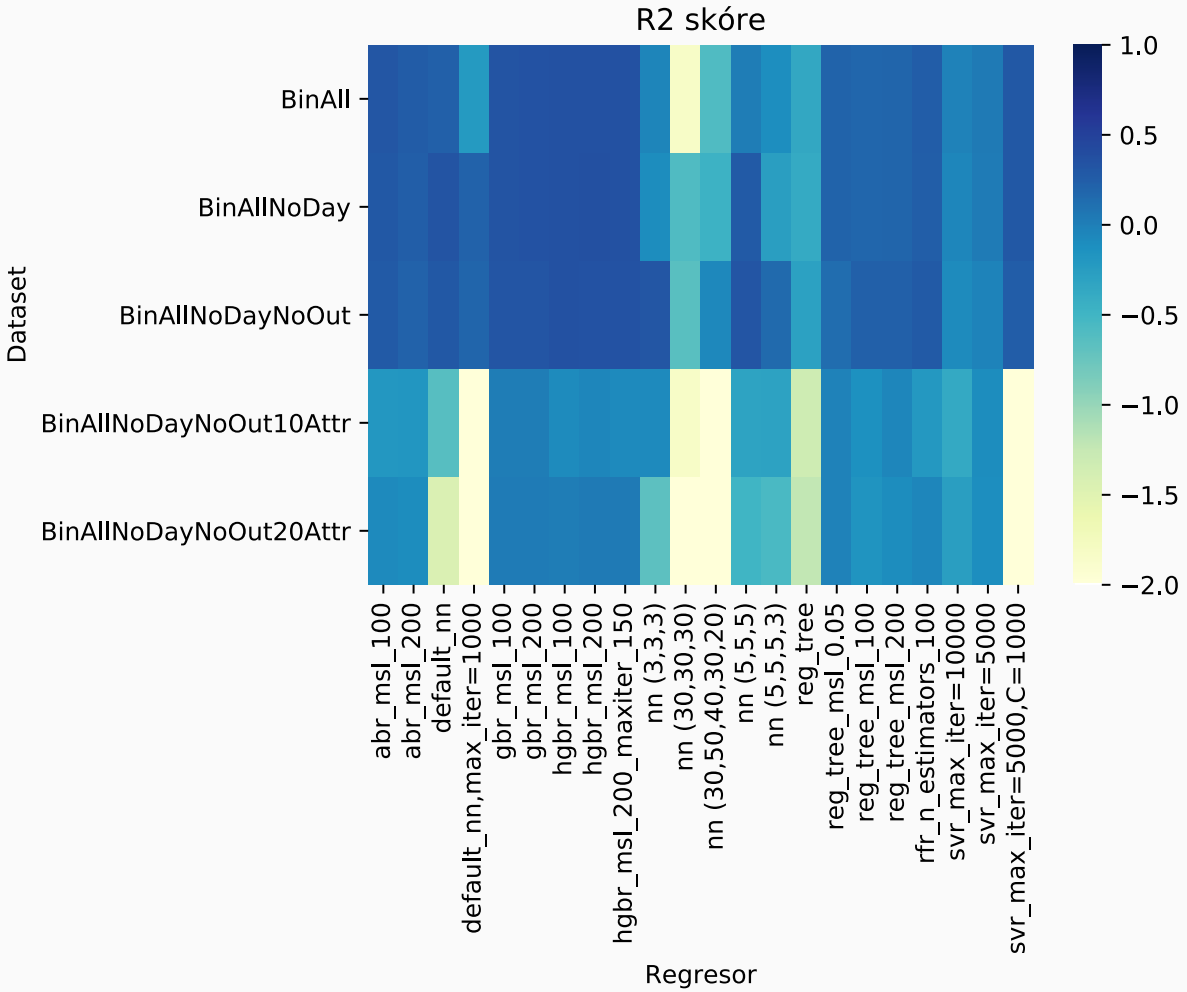
Ansámbl metody

- RandomForestRegressor vedl pouze k minimálnímu zlepšení
- Nejlepší boostovací metody seřazeny podle zlepšení (od nejlepšího):
  1. HistGradientBoostingRegressor
  2. GradientBoostingRegressor
  3. AdaBoostRegressor

Boostovací metody - HistGradientBoostingRegressor



Souhrn všech regresorů



Závěr

- Provedli jsme explorační analýzu
- Vytvořili jsme datasety
- Otestovali jsme regresory a zhodnotili jsme je
- Výsledky regresorů jsou podprůměrné
- Koncentrace $PM_{2.5}$ je těžko předpovidatelná

Děkuji za pozornost

Otázky?