Regrese koncentrace pevných látek v Pekingu
MAD 3 projekt

Bc. Moravec Vojtěch

ZS 2019/2020

Obsah

1 Úvod

V tomto projektu do předmětu Metody Analýzy Dat 3, se zaměříme na regresi, neboli předpověď numerického atributu. Námi vybraný dataset [1] se zaměřuje na znečištění vzduchu v Pekingu od 1. Ledna 2010 do 31. Prosince 2014. Tento dataset jsme získali z UCI Machine Learning Repository [2]. Znečištěním v tomto případě rozumíme koncentraci pevných částic ve vzduchu. Koncentrace se měří v mikrogramech na metr krychlový ($\mu g/m^3$). V našem případě se jedná o částice $PM_{2.5}$, jejíchž průměr je maximálně 2.5 µm. Našim cílem je tedy provést explorační analýzu datasetu, upravit dataset a následně vyzkoušet několik standardních algoritmu pro regresi, předpověď koncentrace znečištění vzhledem k času a přiloženým meteorologickým datům.

2 Popis datasetu

Originální dataset bez úprav obsahuje celkem 43 824 záznamů a 12 atributů, nepočítáme-li číslo řádku, které bude pro analýzu hned odstraněno. V celkem 2 067 řádcích chybí cílová koncentrace a proto, byly tyto řádky také odstraněny. Všechny záznamy obsahují informace o datu a čase, kdy byla koncentrace změřena. Z data jsme navíc vytvořili nový atribut dne v týdnu. Pro regresi máme tedy k dispozici celkem 12 atributů, kde většina je numerická. Statistické vlastnosti numerických atributů, spolu s cílovým atributem jsou shrnuty v Tabulce 1.

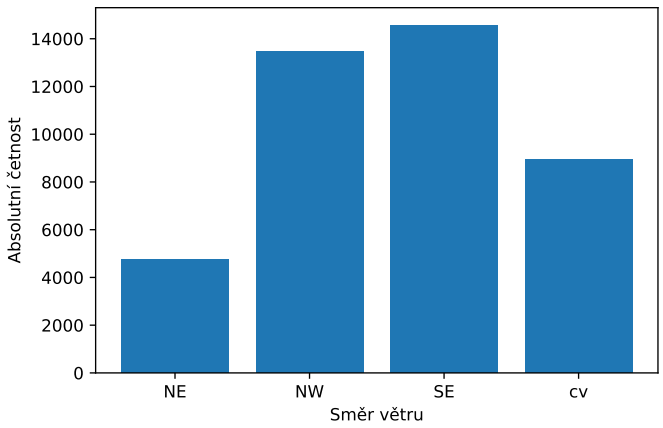
Název atributu	Průměr	σ	Min	Max
Rosný bod (°C)	1,7502	14,4337	-40,0000	28,0000
Teplota (°C)	12,4016	12,1752	-19,0000	42,0000
Tlak (hPa)	1016,4429	10,3007	991,0000	1046,0000
Rychlost větru (m/s)	23,8667	49,6175	0,4500	565,4900
Doba sněžení (hod.)	0,0553	0,7789	0,0000	27,0000
Doba deště (hod.)	0,1949	1,4182	0,0000	36,0000
Koncentrace $PM_{2.5}$ ($\mu g/m^3$)	98,6132	92.0504	0,0000	994,0000

Tabulka 1: Souhrn numerických atributů

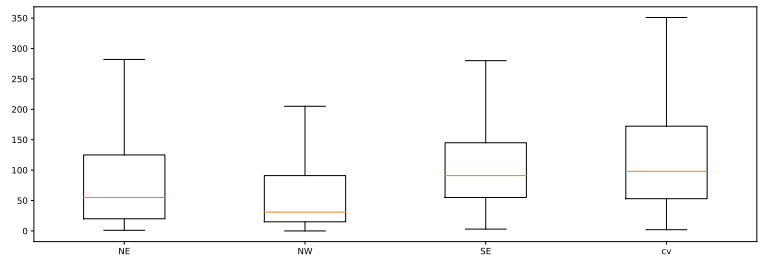
Dále si uvedeme informace o dalších atributech, které jsou dostupné pro každý záznam.

Směr větru je kategoriální atribut, nabývající čtyřech různých hodnot, jejich distribuci můžeme vidět v grafu na Obrázku 1. Originální měřená data obsahovala celkem 16 různých směrů větru, ale autoři je seskupili do 5, kde

4 nalezneme v námi zvoleném datasetu. Hodnota CV znamená klidný, proměnný směr. Na Obrázku 2 můžeme vidět závislost cílové proměnné právě na směru větru, všimneme si, že největší koncentrace pevných látek ve vzduchu je právě při klidném větru a naopak nejmenší, když vítr fouká severozápadně.



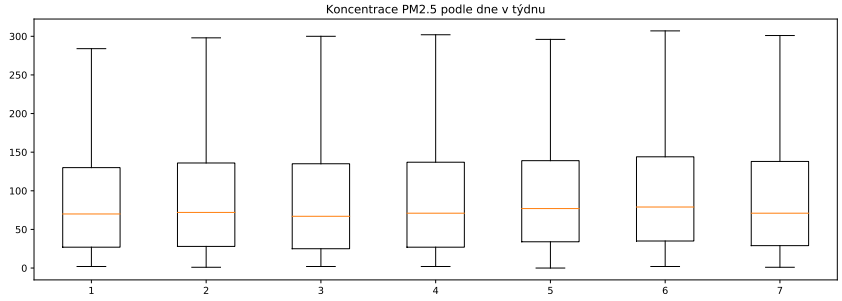
Obrázek 1: Absolutní četnosti směru větru



Obrázek 2: Koncentrace $PM_{2.5}$ vzhledem ke směru větru

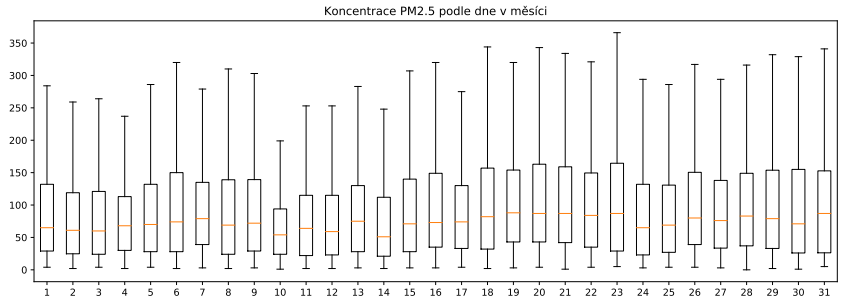
2.1 Atributy data a času

Každý záznam v datasetu obsahuje informaci o dnu, měsíci, roku a hodině, kdy byl vytvořen. My jsme dále přidali den v týdnu. Všechny tyto atributy mohou být reprezentovány jako numerické nebo jako kategoriální. Abychom se mohli rozhodnout, jak budeme s těmito atributy pracovat podíváme se na krabicové grafy cílového atributu v závislosti na dnu, měsíci atd. Nejprve se podíváme na závislost pro dny v týdnu, tu můžeme vidět na Obrázku 3.



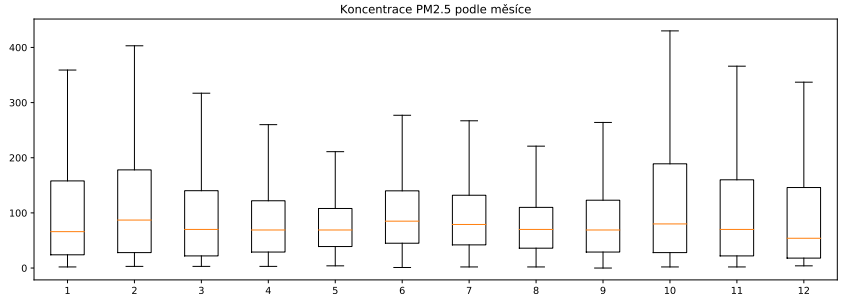
Obrázek 3: Koncentrace $PM_{2.5}$ vzhledem ke dnu v týdnu

Zde nepozorujeme žádnou velkou závislost. Podobný výsledek můžeme vidět na Obrázcích pro dny měsíce 4, měsíců 5 i roky 6. Koncentrace cílového atributu kolísá a neumíme najít žádnou závislost, která by potvrzovala, že se vzrůstajícím dnem či měsícem, koncentrace $PM_{2.5}$ roste či klesá. Rozhodli jsme se tedy, že s atributy budeme pracovat jako s kategoriálními. Navíc vytvoříme další dataset, kde den v měsíci zcela odstraníme.

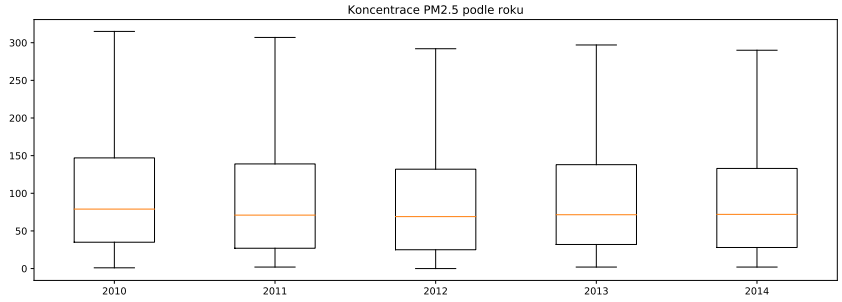


Obrázek 4: Koncentrace $PM_{2.5}$ vzhledem ke dnu v měsíci

V prílohách ještě najdeme koncentrace vzhledem ke dnu v týdnu 19 a měsíci 20 zvlášť pro každý rok.



Obrázek 5: Koncentrace $PM_{2.5}$ vzhledem k měsíci

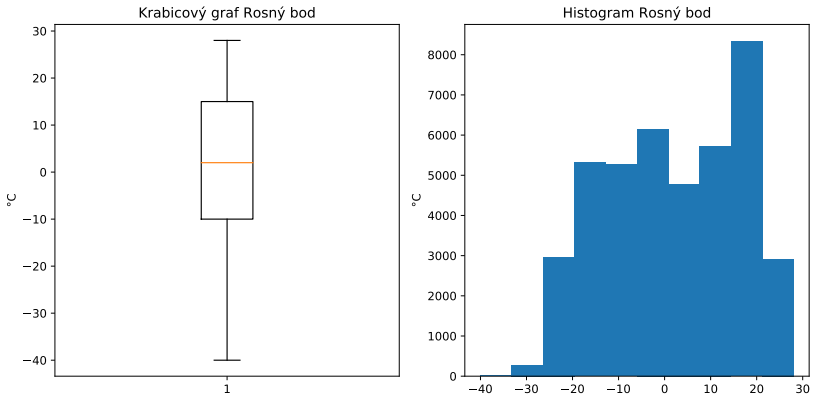


Obrázek 6: Koncentrace $PM_{2.5}$ vzhledem k roku

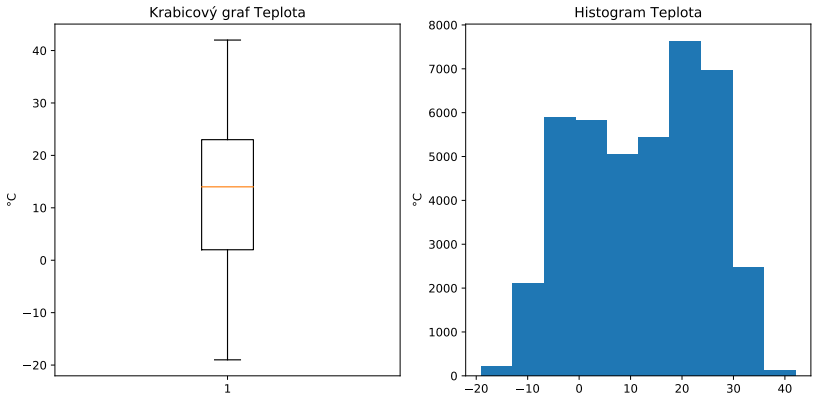
2.2 Distribuce hodnot a odlehlá pozorování numerických atributů

Většina hodnot rosného bodu se pohybuje v rozmezí od -20 °C do 20 °C. V rosném bodě nepozorujeme žádné odlehlé pozorování. Krabicový graf a histogram můžeme vidět na Obrázku 7. Dále můžeme vidět grafy pro teplotu a to na Obrázku 8. Pro teplotu jsme také nenalezli odlehlé pozorování, stejné platí i pro atmosférický tlak.

Nejvíce odlehlých pozorování nalezneme pro rychlost větru, jedná se o 4 893 záznamů z celkových 41 757. Tyto pozorování můžeme vidět na grafu v Obrázku 9. Na histogramu vedle si všimneme, že rychlost větru je většinou do 100 m/s. Odlehlé pozorování tedy odpovídají extrémům, kdy se do Pekingu dostal silný vítr. Tak stejně odlehlé pozorování najdeme u doby sněžení a deště. Obrázky 10 a 11. Drtivou většinu času v Pekingu nesněží ani neprší více jak hodinu v kuse, proto jsou hodnoty od 1 - 2 hodin označeny jako odlehlé pozorování. Dle našeho názoru by bylo špatné rozhodnutí, tyto hodnoty odstranit, neboť právě ony mohou značit určitý pokles či vzrůst koncentrace



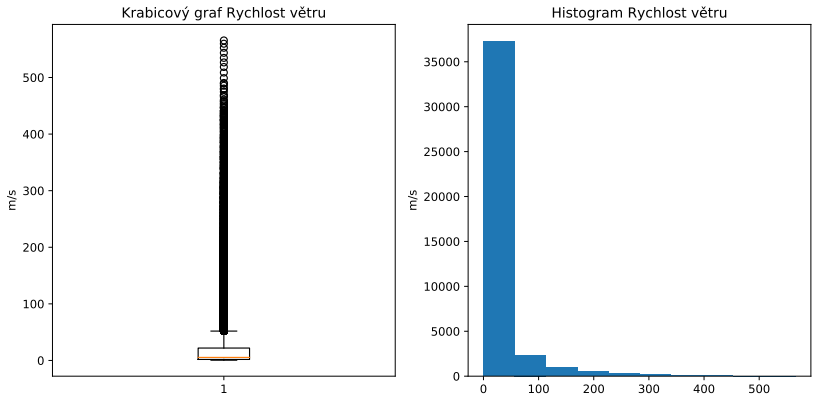
Obrázek 7: Krabicový graf, histogram rosného bodu



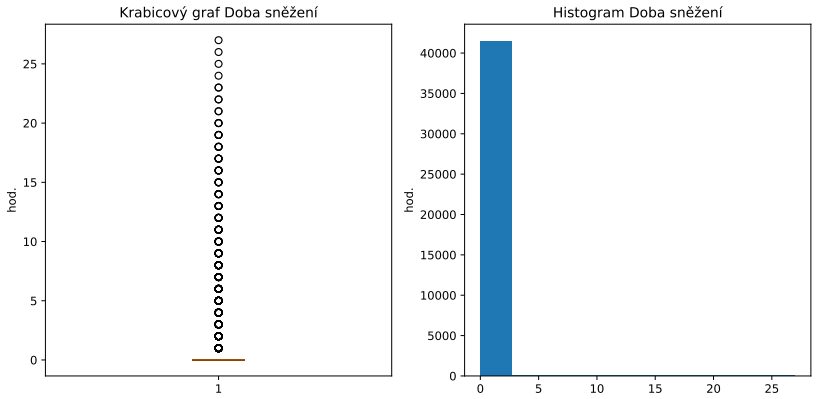
Obrázek 8: Krabicový graf, histogram teploty

$PM_{2.5}$.

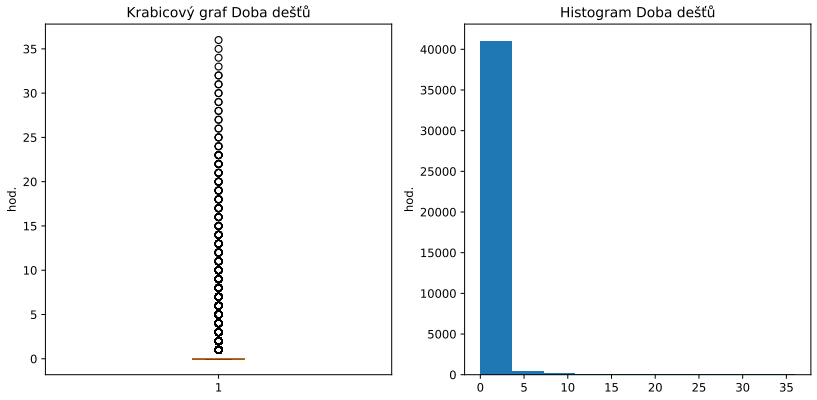
Pro cílový atribut koncentrace $PM_{2.5}$ bylo nalezeneo celkem 1773 odlehlých pozorování, jak můžeme vidět v Obrázku 12. Tyto odlehlé pozorování budou odstraněny pro všechny datasety, pro které budeme zkoušet regresi. Odlehlé pozorování z předchozích atributů odstraníme v nově vytvořeném datasetu a vyzkoušíme jaký vliv bude mít jejich přítomnost či absence.



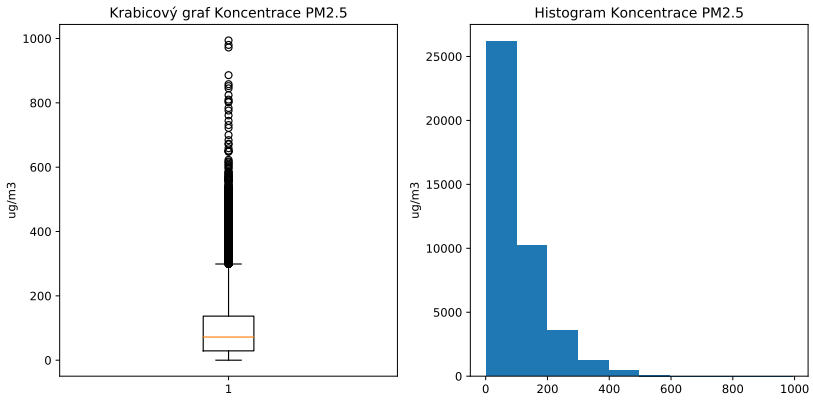
Obrázek 9: Krabicový graf, histogram rychlost větru



Obrázek 10: Krabicový graf, histogram doba sněžení



Obrázek 11: Krabicový graf, histogram doba deště



Obrázek 12: Krabicový graf, histogram koncentrace $PM_{2.5}$

3 Příprava datasetů

V této sekci popíšeme operace, které jsme provedli s originálním datasetem a uvedeme informace o všech datasetech, která jsme z něj vytvořili.

První dataset jsme vytvořili odstraněním odlehlých pozorování cílového atributu $PM_{2.5}$ a binarizací všech časových atributů (den v týdnu, den v měsíci, měsíc, rok a hodina). Odlehlé pozorování cílového atributu budou odstraněny pro všechny datasety.

Druhý dataset, jsme vytvořili z prvního odstraněním informace o dnu, toto jsme provedli, protože nám nedává smysl, aby měla informace o dnu vliv na znečištění.

Pro třetí dataset jsme odstranili úplně všechny odlehlé pozorování. Z třetího datasetu bez všech odlehlých pozorování, jsme vytvořili ještě 2 datasety pomocí redukce dimenze. První z těchto dvou obsahuje pouze 20 nejdůležitějších atributů a druhý obsahuje pouze 10 atributů. U všech vytvořených datasetů jsme provedli normalizaci hodnot do rozmezí 0,0 až 1,0. Souhrn datasetu nalezeneme v Tabulce 2.

Jméno	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

Tabulka 2: Souhrn datasetů pro regresi

4 Regrese

V této sekci vyzkoušíme regresi na námi připravených datasetech. Konkrétně vyzkoušíme několik algoritmu pro regresi a následně několik ansámbl metod. Všechny algoritmy pochází z Python nástroje scikit-learn [3].

Co se týče testování zvolili jsme dvě metriky:

1. R^2 skóre
2. Průměrná kvadratická chyba (MSE)

Jako hlavní metriku bereme R^2 skóre, která udává podíl rozptylu předvídané proměnné, předvídané z nezávislých atributů Nejlepší možné skóre je 1,0, které dostane regresor, který vždy predikuje správnou hodnotu. Naopak regresor, který vydá náhodnou hodnotu, zahazuje vstupní argumenty dostane skóre 0,0 anebo záporné. Průměrnou kvadratickou chybu se snažíme naopak minimalizovat co nejblíž nule. Problém této metriky je to, že nevidíme přímou korelaci s hodnotou cílové proměnné.

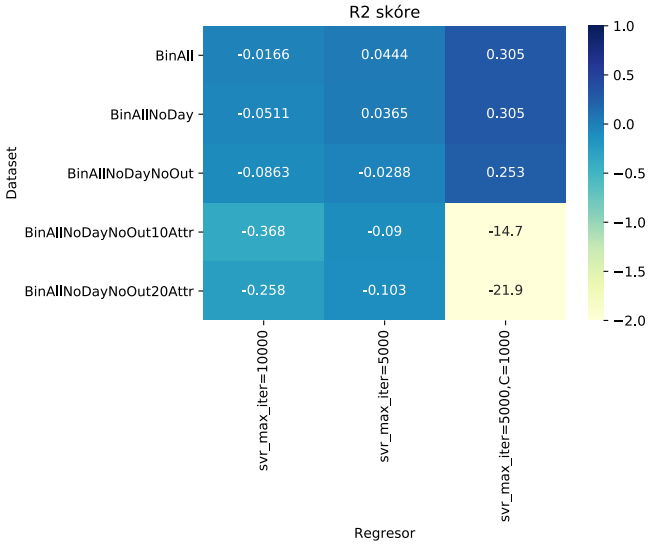
Testování jsme prováděli pomocí funkce *cross_validate* s $k = 3$ (3x byl proveden KFold.). Výsledné metriky, spolu s časem jsou průměrem, přes tyto 3 testování.

4.1 Algoritmy regrese

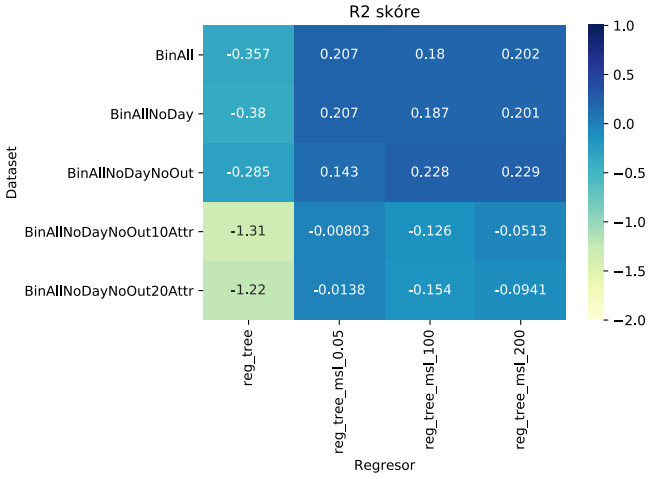
Pro regresi jsme zvolili 3 různé algoritmy (názvy dle tříd ve scikit-learn), jedná se o SVR, DecisionTreeRegressor a MLPRegressor. U SVR jsme použili RBF kernel a dále vyzkoušeli různý počet iterací a větší hodnotu penalizačního argumentu C. Výsledky SVR s R^2 skóre můžeme vidět na Obrázku 13. Všimneme si, že predikce nefunguje pro datasety s redukovaným počtem atributů a nejlepší dosažené skóre je 0,305. Obecně můžeme říct, že SVR potřebuje co nejvíc iterací a vyšší hodnota penalizačního argumentu C pomohla.

Dále jsme vyzkoušeli algoritmus rozhodovacího stromu, u kterého jsme vyzkoušeli vliv minimální velikosti listu na kvalitě predikce. Výsledky predikce můžeme vidět na Obrázku 14. Nejlepšího skóre 0,229 jsme dosáhli při minimální velikosti listu 200. Celkově rozhodovací strom dosáhnul horších výsledků než SVR.

Poslední vyzkoušená metoda, byla regrese založena na neuronových sítích. Zde jsme se snažili najít co nejlepší strukturu sítě (počet skrytých vrstev, počet neuronů ve vrstvě). Výsledky regrese najdeme v Obrázku 15. Nejlepších výsledku jsme dosáhli se základní neuronovou sítí, která má 100 neuronů v



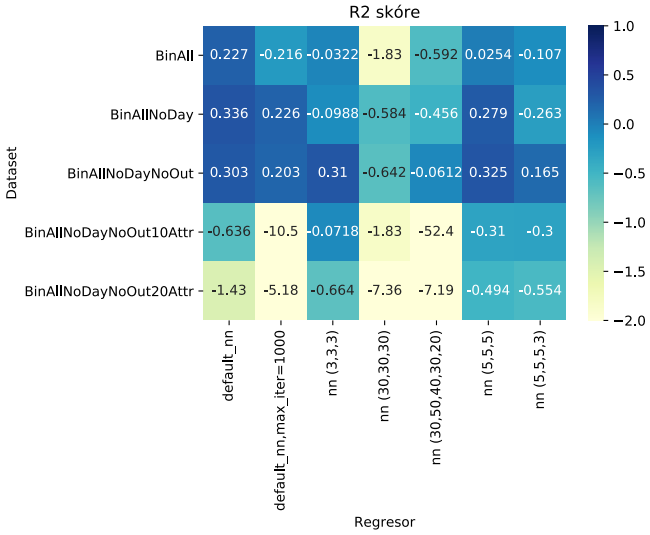
Obrázek 13: Výsledky regrese pro SVR



Obrázek 14: Výsledky regrese pro DecisionTreeRegressor

jedné skryté vrstvě. Dále fungovali dobře sítě se třemi skrytými vrstvami se třemi nebo pěti neurony. Zvýšení počtu iterací nevedlo k lepším výsledkům.

V Tabulce 3 můžeme vidět shrnutí všech 3 algoritmů s jejich nejlepšími výsledky. Všimneme si, že DecisionTreeRegressor vydává nejhorší výsledky ze tří testovaných, ale za to je mnohonásobně rychlejší v natrénování. Obecně



Obrázek 15: Výsledky regrese pro MLPRegressor

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

Tabulka 3: Shrnutí nejpších výsledků algoritmů

jsme doufali v lepší výsledky s R^2 skóre alespoň 0,5 čehož bohužel nedosáhnul žádný algoritmus. Toto může být následek toho, co jsme viděli již v explorační analýze, že cílový atribut nebyl příliš závislý na žádném atributu data a času, kterých je poměrně hodně.

4.2 Ansámbl metody

Dále jsme se rozhodli, vyzkoušet ansámbl metody pro zlepšení výsledků algoritmu DecisionTreeRegressor. Tento algoritmus pracoval velice rychle, ale jeho výsledky byly nejslabší, je tedy dobrým adeptem pro tyto metody. Navíc některé umí pracovat pouze s rozhodovacími stromy. Vyzkoušeli jsme jeden průměrovací algoritmus RandomForestRegressor a tři boostovací algoritmy GradientBoostingRegressor, HistGradientBoostingRegressor a AdaBoostRegressor.

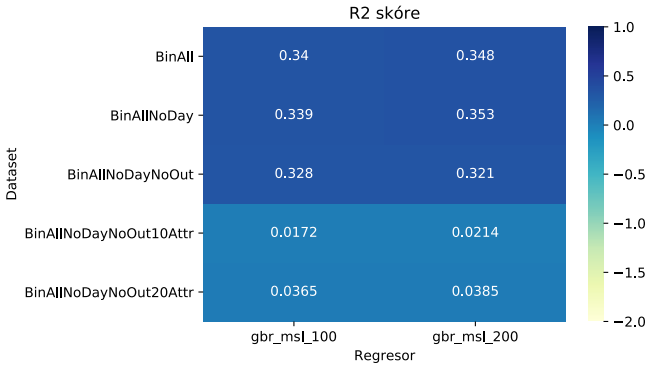
RandomForestRegressor využívá několika stromů k získání lepších výsledků, v našem případě jsme využili 100 stromů. RandomForestRegressor

Dataset	Čas učení (s)	R^2	MSE
df_binAll	22,923	0,260	3533,602
df_binNoDay	15,468	0,242	3618,715
df_noOut_binNoDay	12,429	0,282	3471,420
df_noOut_binNoDay_20attr	32,458	-0,046	5043,728
df_noOut_binNoDay_10attr	16,515	-0,204	5810,074

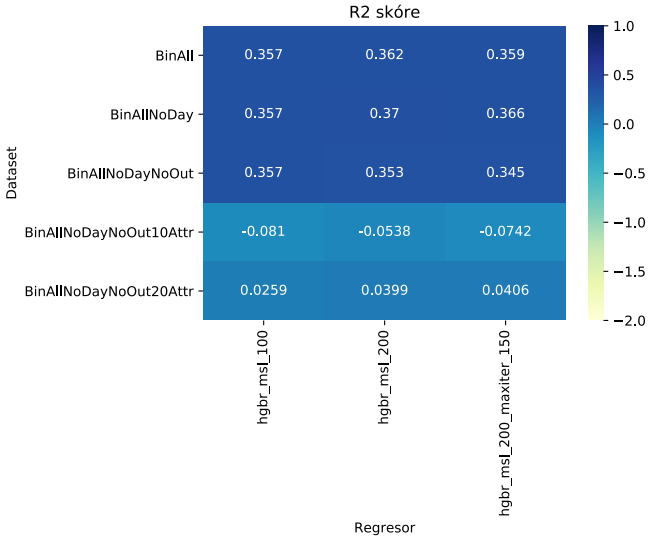
Tabulka 4: Výsledky regrese pro RandomForestRegressor

dokázal zlepšít R^2 skóre v nejlepším případě o 0,051. Výsledky můžeme vidět v Tabulce 4.

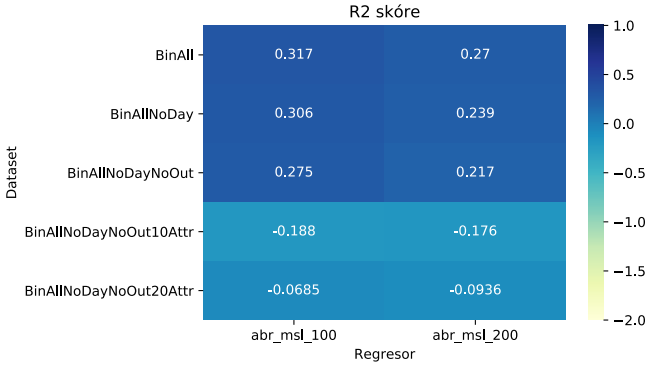
Oproti RandomForestRegressoru jsou na tom boostovací metody lépe, výsledky můžeme vidět na Obrázcích 16, 17 a 18. Největší zlepšení pozorujeme u HistGradientBoostingRegressor, poté GradientBoostingRegressor a nakonec AdaBoostRegressor. Vylepšené výsledky už dosahují většího skóre než SVR a MLPRegressor, avšak pořád se nejedná o moc dobré výsledky. Souhrnné výsledky R^2 skóre najdeme v přílohách na Obrázku 21, taktéž pro průměrnou kvadratickou chybu na Obrázku 22 a dobu učení na 23.



Obrázek 16: Výsledky regrese pro GradientBoostingRegressor



Obrázek 17: Výsledky regrese pro HistGradientBoostingRegressor



Obrázek 18: Výsledky regrese pro AdaBoostRegressor

5 Závěr

V rámci této práce jsme se zaměřili na data, které pojednávají o koncentraci znečištění v Pekingu. Nad těmito daty jsme provedli explorační analýzu, pomocí jejíž výsledků jsme připravili několik datasetů, které vstupovali do

regrese. V rámci predikce znečištění jsme vyzkoušeli několik algoritmu pro regresi, všechny z knihovny scikit-learn [3]. Poté jsme uvedli zjištěné výsledky regrese, které jsme se dále rozhodli vylepšit pomocí Ansámbl metod. Ansámbl metody vedly k vylepšení skóre predikce. Avšak ano po vylepšení nejsou výsledky příliš přívětivé. Nemůžeme tedy doporučit žádnou metodu pro reálnou predikci koncentrace znečištění.

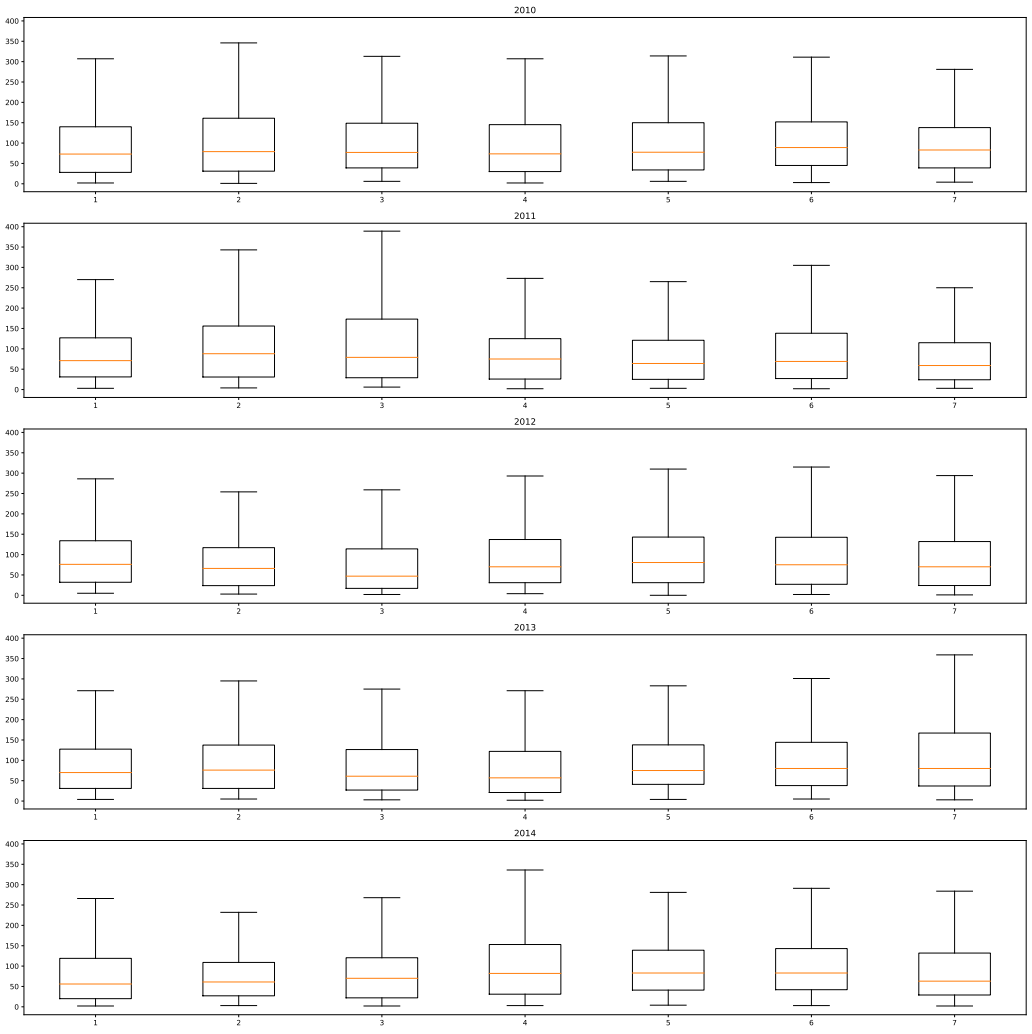
Reference

[1] Liang X., Zou T., Guo B., Li S., Zhang H., Zhang S., Huang H., and Chen S. X., "Assessing beijing's pm2.5 pollution: severity, weather impact, apec and winter heating," 2015.

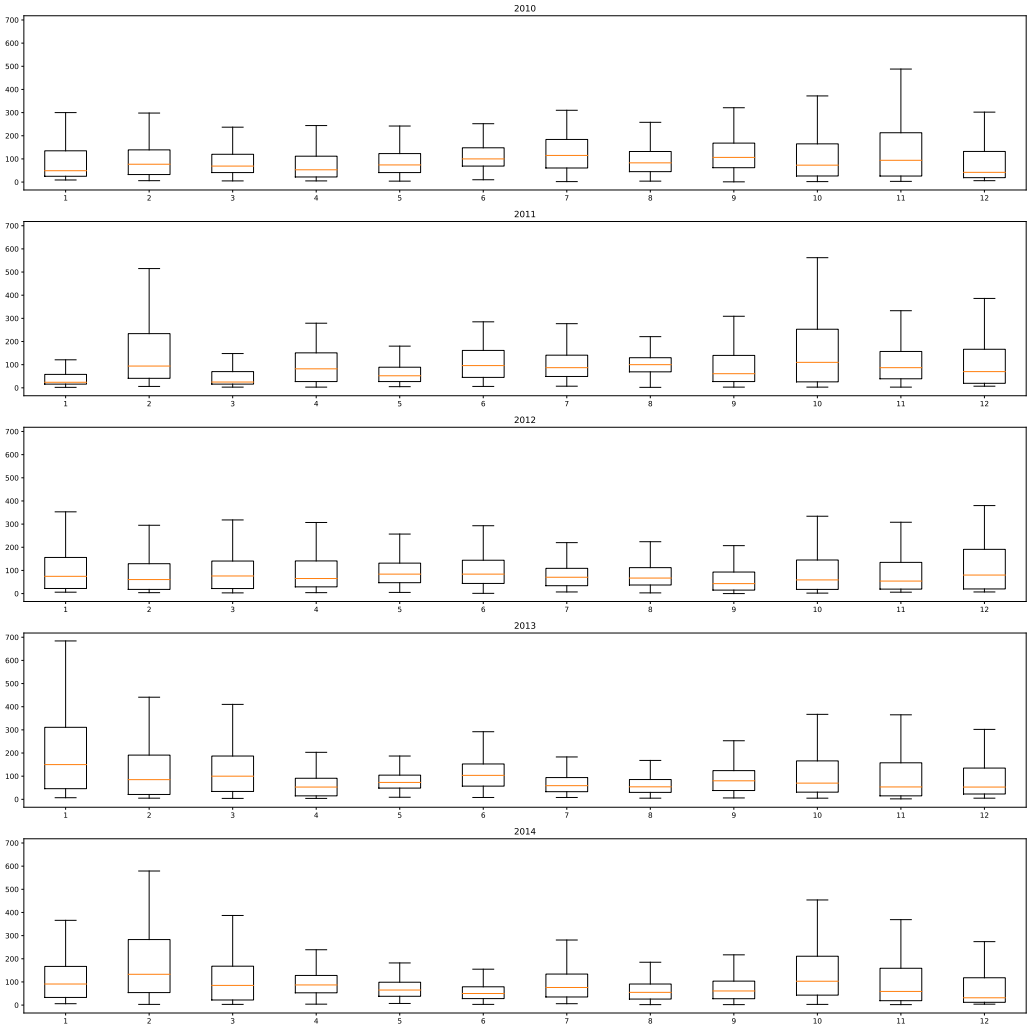
[2] D. Dua and C. Graff, "UCI machine learning repository," 2017.

[3] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

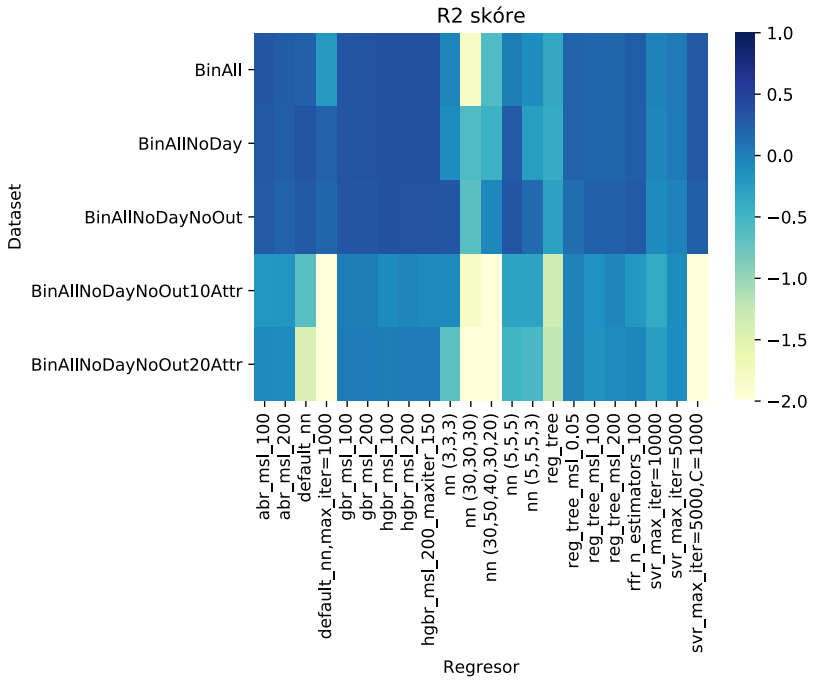
6 Přílohy



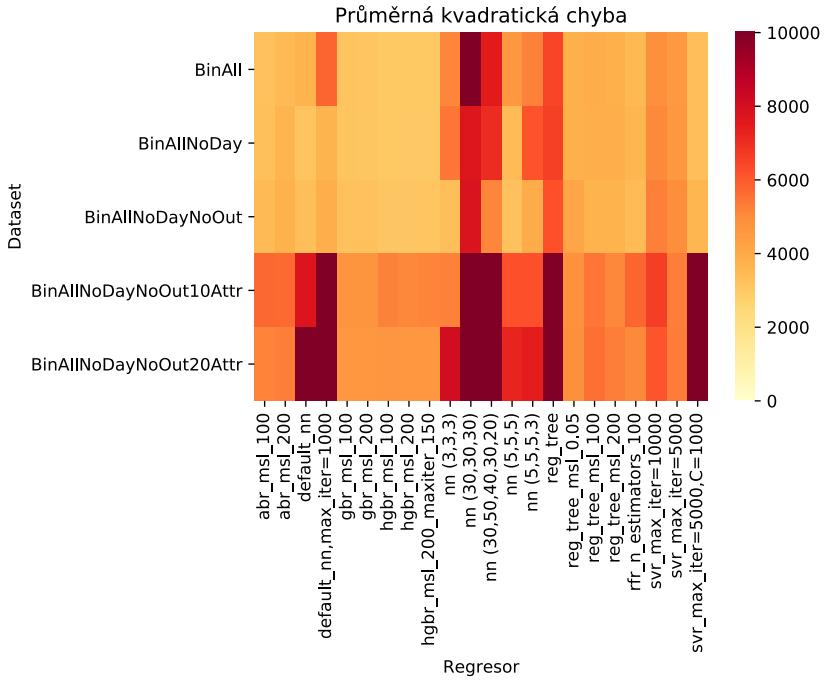
Obrázek 19: Koncentrace $PM_{2.5}$ vzhledem ke dnu v týdnu pro každý rok



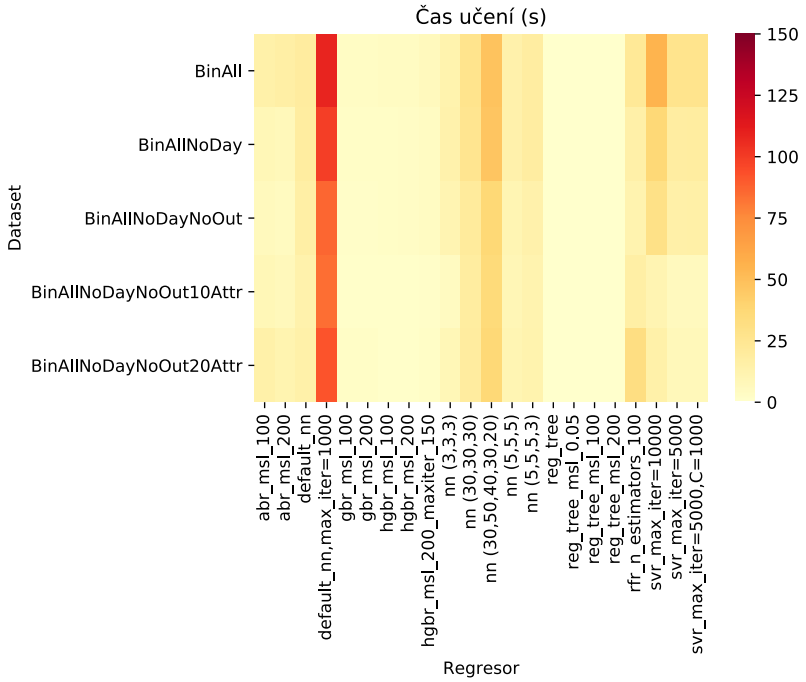
Obrázek 20: Koncentrace $PM_{2.5}$ vzhledem k měsíci pro každý rok



Obrázek 21: R^2 skóre pro všechny regresory



Obrázek 22: Průměrná kvadratická chyba pro všechny regresory



Obrázek 23: Čas učení pro všechny regresory