Predikcia ročných platov na základe pracovných ponúk

Tomáš Baránek[1] and Martin Ilavský[2]

[1] FIIT STU, Bratislava, Slovensko xbaranek@stuba.sk
[2] FIIT STU, Bratislava, Slovensko xilavskym@stuba.sk

Abstrakt. Článok sa zaoberá predikciou ročných platov na základe popisu zverejnených pracovných ponúk. Využívame pritom atribúty, ako popis pracovnej ponuky, mesto, kategória, názov pozície a ďalšie. Vo fáze predspracovania sme vytvorili viaceré číselné atribúty a aplikovali na ne rôzne metódy strojového učenia. Konkrétne sme sa zamerali pri lineárnej regresii na Lasso regresiu, z regresných metód založených na stromoch sme použili Random Forest a Gradient boosting a taktiež sme vyskúšali aj viacvrstvové neurónové siete (MLP) a SVM. Vybrané metódy sme sa snažili vylepšiť úpravou parametrov jednotlivých algoritmov, čím sme dostali zlepšenie výsledkov v niektorých prípadoch až o 8%. V práci sme sa taktiež zaoberali módou učenia súborom metód, no vzhľadom na rozloženie výsledných dát, táto metóda nepriniesla zlepšenie. Podarilo sa nám dosiahnuť priemernú odchýlku pri predikovanom ročnom plate 5461.248, čo nás v súťaži na portáli Kaggle radilo na 25. miesto.

Kľúčové slová: regresia, predikcia platov, strojové učenie

1 Opis Problému

Cieľom našej práce bola predikcia výšky platov na základe ponúk zverejnených na rôznych pracovných portáloch. Pomocou použitia rôznych metód strojového učenia sme sa snažili odhadnúť ponúkaný plat na základe dostupných dát.

Zdrojom dát pre našu prácu bola jedna zo súťaží, ktoré sú usporadúvané na portáli Kaggle [1]. Konkrétne je zaradená do kategórie „Kaggle Startup program“. Súťaž bola vypísaná spoločnosťou Adzuna, ktorá ponúka údaje obsahujúce v jednotlivých atribútoch rôzne formátované dáta, nakoľko ide o dáta zozbierané z viacerých zdrojov.

Adzuna vytvorila súťaž za účelom vytvoriť prediktívny nástroj, ktorý bude predikovať ročnú mzdu pre jednotlivé inzeráty podľa dostupných atribútov v zadanom inzeráte. Takto vytvorený nástroj by mal pomôcť pri vytváraní nových inzerátov, ale taktiež by mohol pomôcť uchádzačom o prácu zistiť približné finančné ohodnotenie pracovných pozícií. Toto riešenie by malo byť nasadené hlavne vo Veľkej Británii, ale Adzuna plánuje toto riešenie rozšíriť i do celého sveta.

2 Opis dát

K dispozícii sme mali dva samostatné datasety na trénovanie a vyhodnotenie. Testovací dataset však neobsahoval predikované hodnoty, ktoré neboli zverejnené ani po skončení súťaže. Trénovaciu, validačnú a testovaciu množinu dát sme tak vytvorili z pôvodných trénovacích dát v pomere 60:20:20.

2.1 Štruktúra dát

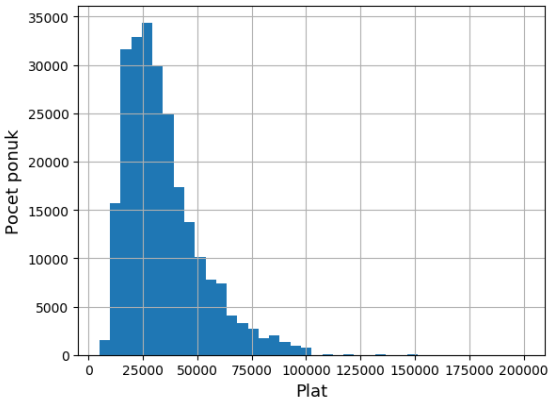
Hlavná množina údajov obsahovala 244 614 záznamov predstavujúcich jednotlivé ponuky pracovných miest. V stĺpcoch sa nachádzajú jednotlivé atribúty pracovnej pozície, ktoré sú opísané v Tab. 2.1. Všetky atribúty okrem predikovanej normalizovanej výšky platu sú textové.

Tab. 2.1 Štruktúra datasetu

Názov atribútu	Opis
Id	Jedinečný identifikátor pre každú pracovnú pozíciu.
Title	Názov inzerátu. Väčšinou názov pracovnej pozície.
FullDescription	Úplný popis pracovnej pozície. Všetky číselné hodnoty, ktoré mohli obsahovať výšku platu boli nahradené znakmi ***.
LocationRaw	Voľný text mesta, kde je pracovná pozícia ponúkaná.
LocationNormalized	Adzuna normalizovala polohu z jej vlastného lokalizačného stromu, ktorý vytvorili na základe poľa LocationRaw. Normalizátor však nie je dokonalý.
ContractType	„full_time“ – práca na plný úväzok „part_time“ – polovičný úväzok, brigáda Pole je prázdne v 79,85 % prípadoch.
Company	Meno zamestnávateľa.
Category	Kategória, ktorá je vybratá z 30 štandardných kategórií práce.
SalaryRaw	Voľný text, ktorý obsahuje plat, ktorý bol získaný z inzerátu.
SalaryNormalised	Ročný plat interpretovaný Adzunou z poľa SalaryRaw. Je vyrátaný zo strednej hodnoty. Toto je hodnota, ktorú sa v danej súťaži snažia súťažiaci predikovať.
SourceName	Názov webovej lokality, alebo inzerujúceho, z ktorej bol inzerát získaný.

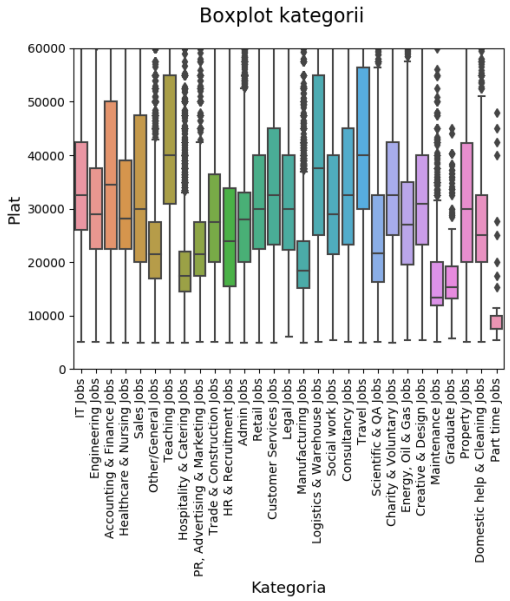
2.2 Analýza dát

Analýzou dát sme zistili nasledujúce skutočnosti. Rozdelenie platov je Gausove s dlhým chvostom platov od 100 000 do 200 000. Najčastejšie sa ponuky vyskytujú v rozmedzí 15 000 až 35 000 ročne, ako môžeme vidieť na obr. 2.1.



Obr. 2.1 Histogram platov

Z boxplotu platov na obr. 2.2 podľa kategórií, môžeme taktiež povedať, že bude mať tento atribút v niektorých prípadoch vplyv na ročnú mzdu. Rovnaké tvrdenie môžeme povedať aj o grafoch boxplotov pre zdroje ponúk a zamestnávateľov, ktoré neuvádzame z dôvodu veľkého počtu rôznych zdrojov (162) a zamestnávateľov (20 812).



Obr. 2.2 Boxplot kategórií ponúk

Všetky atribúty v našom datasete sú textové, a preto si vyžadovali rôzne typy predspracovania, aby ich bolo možné použiť v algoritmoch strojového učenia. Všetky úpravy sú bližšie popísané v časti 5.

3 Definovanie úlohy objavovania znalostí

Pomocou regresných metód strojového učenia sme predikovali ročnú mzdu na základe popisu pracovných ponúk. Presnosť našej metódy sme vyhodnocovali na základe priemernej odchýlky od skutočných ročných platov. Spracovaním textových atribútov ponúk sme vytvorili nové číselné atribúty, ktoré je možné v týchto metódach použiť. Vyskúšali sme rôzne metódy regresie a ladenia parametrov týchto metód. Na záver sme sa pokúsili skombinovať všetky použité metódy na dosiahnutie čo najlepších výsledkov, tzv. učením súborom metód.

4 Práce iných autorov

K problematike predikcie platov sme našli viacero zaujímavých článkov, ktoré sa z veľkej časti venovali priamo súboru dát, s ktorým pracujeme v našej práci.

Michael Salmon sa vo svojej práci [2] síce nevenuje nášmu súboru dát, ale detailne popisuje postup predikcie platov stiahnutých z portálu www.indeed.com. Postup sťahovania dát z toho portálu detailne popisuje vo svojom ďalšom článku [3], čo nás však pre účely tejto práce nezaujíma. Zaujímavejší je článok, kde sa venuje priamo analýze dát a samotnej predikcií platov. V krátkych a prehľadných krokoch ukazuje možnosti modulu scikit [4]. Salmon rozdeľuje pracovné inzeráty podľa mediánu a priemernej hodnoty platu, pričom sa zameriava na parametre, ako City, Job Title, Company Name, Location a Job Summary.

Z týchto parametrov boli vytvorené otázky, ktoré boli následne binárne ohodnotené <0,1>. A to, konkrétne:

- Bola táto práca umiestnená v Chicagu, vo Washingtone DC alebo v San Franciscu?
- Obsahuje názov pozície „dobré“ termíny, ako napríklad: „data scientist“, „machine“, „learning“ alebo „enginner“?
- Obsahuje názov pozície „zlé“ termíny, ako napríklad: „research“, „analyst“ alebo „associate“?
- Obsahuje názov spoločnosti „dobrý“ termín „associates“?
- Obsahuje názov spoločnosti „zlý“ termín „university“?
- Je inzerát uverejnený v meste uverejnenia alebo v blízkej oblasti?
- Obsahuje položka „job summary“ „dobré“ termíny, ako napríklad: „machine learning“, „data scientist“, „data science“, „senior“, „lead“, „big data“, „analytics“ alebo „years“?
- Obsahuje položka „job summary“ „zlé“ termíny, ako napríklad: „research“, „analyze“ alebo „analysis“?

Na takto pripravené dáta autor použil logickú regresiu z modulu scikit [4]. Výsledky boli vyhodnotené pomocou viacerých metrík, pozri Tab. 4.1. Táto tabuľka zobrazuje výsledky predpovedania, či bude daný inzerát pod alebo nad medián platu.

Tab. 4.1 Výsledky Salmon

	Precision	Recall	f1-score	Support
pod median	0.80	0.89	0.84	45
nad median	0.88	0.78	0.82	45
avg / total	0.84	0.83	0.83	90

Archit Khosta sa v práci Job Salary Prediction [5] priamo venuje súboru dát, ktorý používame aj my. Khosta rozdelil svoju prácu do „modelov“, pričom:

- Model 0 – Tento model pracuje čisto len s priemerným platom.
- Model 1 – Pri tomto modeli boli brané do úvahy položky Location, Name Of Company a Title of the Job.
- Model 2 – Z modelu 1 bol pridaný parameter Category.
- Model 3 – Boli použité rôzne regresné techniky. Atribúty, ktoré boli použité boli vybrané z „company name“, „location“, a „category“. Následne na tieto atribúty použil Label Binarizer a ten skúšal pomocou SGDRgressora, Linear Regressora a Random Forest Regressora. SGDRegressor priniesol najlepšie výsledky pre tento model.
- Model 4 – použitie strojového učenia, pričom tento model nepriniesol najlepšie výsledky a trval príliš veľa času. Nakoniec teda predikoval cenu pre každú kategóriu a tu následne použil ako predikovanú cenu.

Pri prvých troch modeloch bol použitý nástroj Pandas [6], ktorý slúži na dátovú analýzu. Detailný popis jednotlivých modelov však v článku chýba. V Tab. 4.2 môžeme vidieť výsledky, ktoré jednotlivé modely dosiahli.

Tab. 4.2 Výsledky Khosta

	Absolútna chyba	Priemerná absolútna chyba	Stredná kvadratická chyba
Model 0	796966682	4341	321230055
Model 1	737971741	4020	261936172
Model 2	663712124	3615	230256255
Model 2.1	705694939	3844	251742591
Model 3	690242229	3760	237854119
Model 4	739674141	4029	269918358

Pre nás zaujímavý parameter je priemerná absolútna chyba, pričom výsledne hodnoty tohto autora radia na 2. miesto v súťaži, čo vzbudzuje podozrenie, že výsledky sú buď nesprávne interpretované alebo vymyslené. Vzhľadom na to, že pri modeli 0 sa mu podarilo len s priemernou hodnotou platu dosiahnuť chybu, ktorá by skončila na 10 mieste, myslíme si, že uvedené čísla sú nesprávne.

Z výslednej tabuľky vyplýva, že rozdelenie na základe kategórie pomocou nástroja Pandas dosahovalo najlepšie výsledky, no i tak je táto metóda diskutabilná.

V článku [7] autori uvádzajú, že sa im v súťaži Adzuna podarilo dostať do top 6%. Konkrétne, dosiahli priemernú chybu 4933, pričom použili rôzne lineárne modely, Random Forest a neurónové siete. Kolektív autorov tohto článku vychádzal z článku [8], kde sú dosiahnuté dva výsledky, a to:

- Pri použití priemerného platu – hodnota priemernej chyby 13253.
- Pri použití Random Forest - 7633 (parametre, ktoré vstupovali do učenia pre Random Forest, sú detailne popísané v práci [7]).

Vylepšenie výsledku na hodnotu 4933 sa im podarilo dosiahnuť lineárnou kombináciou štyroch najlepších metód pre Random Forest.

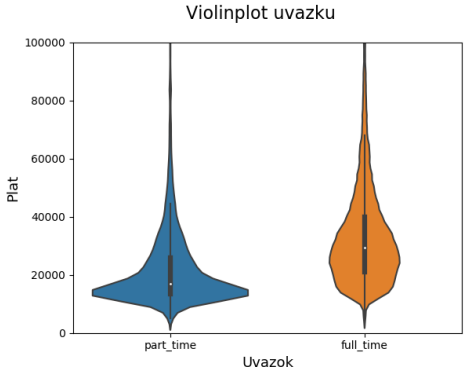
Z uvedených prác plynie, že je možné pomocou metódy náhodných stromov dosiahnuť výsledky, ktoré môžu dosahovať výsledky medzi 10% najlepších výsledkov súťaže.

5 Opis metód

V tejto časti uvedieme, aké metódy predspracovania sme aplikovali na existujúce atribúty za cieľom doplnenia chýbajúcich dát alebo vytvorenia úplne nových atribútov. Opíšeme zvolené metódy regresie a spôsob vyhodnocovania výsledkov.

5.1 Predspracovanie

Atribút typ úväzku obsahoval v 79,85 % prípadoch prázdnu hodnotu, a preto sme tieto hodnoty pri niektorých ponukách, pokiaľ to bolo možné, doplnili. Za brigádu sme považovali tú ponuku, ktorá obsahovala v popise slová part time, temporary a odd job. Ponuky so slovami full time, regular, permanent a stable zase za trvalé pracovné pomery. Po tomto doplnení zostalo prázdnych 58,75 % hodnôt. Tieto kategorické hodnoty boli ďalej reprezentované pomocou samostatných atribútov s hodnotami 0 a 1. Na obr. 5.1 môžeme vidieť, že tento atribút bude mať vplyv na výšku mzdy. Väčšina ponúk na brigády sa nachádza pod hranicou 20 000 dolárov, pričom ponuky na trvalý pracovný pomer bývajú očakávane platovo lepšie ohodnotené.



Obr. 5.1 Boxplot kategórií ponúk

Ďalšie nové atribúty sme vytvorili na základe kľúčových slov vypovedajúcich o vzdelaní v popise pracovnej pozície. Vznikli tak binárne atribúty vzdelania, pričom pracovná ponuka mohla obsahovať kľúčové slová v rôznych tvaroch z viacerých definovaných skupín, ako PHD (phd, doctor, senior), Master (master, ing, mgr, meng) a Graduate (graduate, bachelor, junior). Celkovo bolo ponúk s PhD. 2425, Ing./Mgr. 20873 a s absolventskými zručnosťami 78108.

Kategorické atribúty zamestnávateľa typu pracovnej pozície a zdroja ponuky sme sa rozhodli neinterpretovať vektorovo z dôvodu veľmi dlhých vektorov (21 003 stĺpcov) už v samo o sebe veľkom datasete. Každý kategorický atribút sme reprezentovali len troma stĺpcami obsahujúcimi minimálnu, maximálnu a mediánovú mzdu danej kategórie, v ktorej je pracovná ponuka zaradená. Vytvorili sme tak 9 nových čŕt.

Rovnaký postup sme zvolili aj pri analýze textu. Popisy a názvy pracovných ponúk sme vložili do nástroja Elasticsearch. Následne sme pre každú ponuku našli 5 najpodobnejších a opäť z nich vyrátali minimum, maximum a experimentovali sme s priemerom, mediánom a váženým priemerom. Pri hľadaní najpodobnejších ponúk sme použili dopyt „more like this“, pričom bola názvu pracovnej pozície priradená 5-násobne vyššia váha, ako jej popisu.

V tejto časti práce sme implementovali aj klastrovanie ponúk na základe parametra „LocationRaw“. Nakoľko Adzura a aj článok [5] poukazujú na to, že parameter „LocationNormalized“ nie je dokonalý, rozhodli sme sa vytvoriť vlastný klastrovací nástroj pre polohu inzerátu. Najskôr sme pomocou Google Maps Geolocation API získali GPS súradnice jednotlivých ponúk. Tieto sme sa potom snažili zatriediť do skupín pomocou DBSCAN algoritmu. Pomocou Google API sa nám podarilo získať GPS súradnice k 97.47% inzerátov, čo predstavuje 6199 inzerátov bez GPS súradníc.

Posledné pridané binárne atribúty boli vytvorené z kľúčových slov samostatne pre názvy a popisy pracovných ponúk, ktorých ročná mzda bola pod a nad hodnotou mediánu (30 000) všetkých ponúk. Pre každú kombináciu (názov alebo popis ponuky a ponuky pod mediánom a nad mediánom) sme zistili 10 kľúčových slov pomocou elasticsearch dopytu „significant term“ hľadajúci slová, ktoré sa vo zvolenej množine

nachádzajú vo vyššej frekvencii ako v ostatných záznamoch. Vytvorených tak bolo 40 nových atribútov obsahujúcich hodnoty 0 alebo 1.

5.2 Metódy dolovania v dátach

Pri našich experimentoch sme z dôvodu predikcie číselných nekategorických hodnôt použili regresné metódy. Pri inšpirovaní sa možnými metódami sme čerpali najmä z dokumentácie k modulu scikit [4].

Vyskúšané boli nasledovné metódy. Pre lineárnu regresiu sme použili Lasso regresiu. Z regresných metód založených na stromoch sme použili Random Forest a Gradient boosting. Poslednými testovanými metódami, ktoré vyžadovali normalizáciu vstupných hodnôt do intervalu <0,1> boli viacvrstvové neurónové siete (MLP) a SVM.

Dosiahnuté výsledky sme porovnávali na základe priemernej odchýlky od skutočnej ročnej mzdy, ktorá bola taktiež použitá aj pri vyhodnocovaní súťaže. Skrátený rebríček súťaže s dosiahnutými odchýlkami môžeme vidieť v tab. 5.1.

Tab.5.1 Skrátená výsledková listina súťaže

Poradie	Priemerná odchýlka (MAE)
1	3464.55935
2	3612.65985
3	3862.80973
10	4346.28819
20	5154.85622
30	5664.55000
40	6131.98944
50	6545.92354
80	10062.704

6 Experimenty

V tejto časti opíšeme vykonané experimenty. V prvej časti sa zaoberáme relevantnosťou jednotlivých vytvorených atribútov. Druhá časť sa venuje ladeniu parametrov jednotlivých vybraných metód a v poslednej časti zhrnieme a zosumarizujeme získané poznatky. Všetky experimenty boli z výkonnostných a časových dôvodov vykonané nad 20 % dát. Tri najlepšie vyladené metódy boli následne otestované nad všetkými dátami.

6.1 Vplyv atribútov

Relevantnosť jednotlivých parametrov sme vyhodnocovali pomocou regresie Random Forest s nastavením 10-tich stromov s maximálnou hĺbkou 8. Konkrétne pridané črty k predchádzajúcim použité pri jednotlivých testoch môžeme vidieť v Tab. 6.1.

Tab.6.1 Testovacie scenáre relevantnosti atribútov

Číslo testu	Pridané črty oproti predchádzajúcemu testu	Dosiahnutá odchýlka
1	Typ úväzku	12944.62
2	Vzdelanie	13154.74
3	Medián, Min, Max kategórií	13127.26
4	Medián, Min, Max zdrojov	11139.43
5	Medián, Min, Max zamestnávateľov	9407.27
6	Priemer, Min, Max podobných ponúk	6915.29
7	Medián, Min, Max podobných ponúk	6666.25
8	Vážený priemer, Min, Max podobných ponúk	6634.65
9	Odstránenie dlhého chvosta. Len ponuky pod 100 000 (99,62 %)	6404.45
10	Klastrovanie polohy do 30km	6366.45
11	Kľúčové slová názvov aj popisov pracovných ponúk	5901.53

Na základe testov vieme povedať, že najväčšiu výpovednú hodnotu mali očakávane črty vytvorené na základe podobnosti textov pracovných ponúk. Črty kategórií nemali príliš veľký vplyv, pretože ich je najmenší rôzny počet (19). Najväčší vplyv kategorických čŕt bol z tohto dôvodu pri zamestnávateľoch, kde sme dosiahli zlepšenie o 1 732 oproti predchádzajúcemu testu.

Rozdiely pri použití priemerov kategorických čŕt namiesto mediánov neuvádzame, pretože boli v tomto prípade z dôvodu použitého algoritmu zanedbateľné.

Vplyv priemeru, mediánu a váženého priemeru sme zaznamenali pri podobných ponukách získaných pomocou Elasticsearchu. Najlepšie výsledky dosiahol vážený priemer s výsledkom o 32 bodov lepším oproti mediánu a 280 bodov lepším oproti klasickému priemeru.

6.2 Ladenie parametrov

Ladenie parametrov sme aplikovali na všetky vybrané metódy okrem lineárnej regresie Lasso a SVM z dôvodov dlhých učiacich sa časov (viac ako 30 minút resp. viac ako 10 minút) a vysokej odchýlke s prednastavenými hodnotami (11 000 a 8000). Konkrétne rozsahy a hodnoty parametrov pri ladení pomocou RandomizedSearchCV môžeme vidieť v tabuľkách 6.2 až 6.5. Grafické znázornenie výsledkov pred a po ladení je možné vidieť na obr 6.1. Na základe výsledkov môžeme povedať, že najúspešnejšia bola metóda GradientBoosting, nasledovaná metódami Random Forest a MLP.

Tab.6.2 Ladené atribúty regresie Random Forest

Random Forest		
Parameter	**Testované hodnoty**	**Najlepšie hodnoty**
n_estimators	<10,80> veľkosť kroku: 5	65
max_depth	<4,12> veľkosť kroku: 1	11
min_samples_leaf	<1,10> veľkosť kroku: 1	3
min_samples_split	<2,6> veľkosť kroku: 1	5
max_features	auto, sqrt, log	auto

Tab.6.3 Ladené atribúty regresie Gradient Boosting

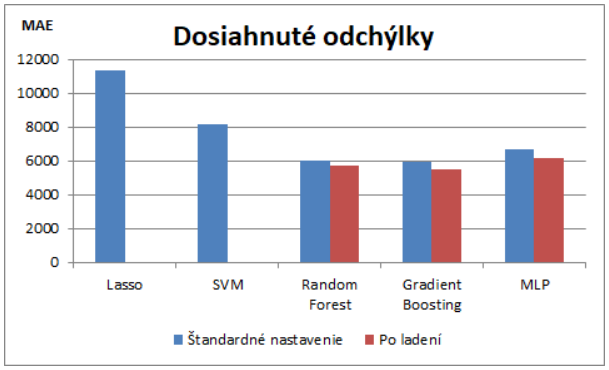
Gradient Boosting		
Parameter	**Testované hodnoty**	**Najlepšie hodnoty**
n_estimators	<10,80> veľkosť kroku: 5	65
max_depth	<4,12> veľkosť kroku: 1	8
min_samples_leaf	<1,10> veľkosť kroku: 1	7
min_samples_split	<2,6> veľkosť kroku: 1	6
loss	ls, lad, huber, quantile	huber
alpha	<0.5, 1.7> veľkosť kroku: 1	0.6

Tab.6.4 Ladené atribúty regresie MLP

MLP Neurónová sieť		
Parameter	**Testované hodnoty**	**Najlepšie hodnoty**
activation	identity, logistic, tanh, relu	relu
solver	lbfgs, sgd, adam	adam
max_iter	<200,500> veľkosť kroku: 20	300
hidden_layer_count	<1,5> veľkosť kroku: 1	4
hidden_Layer_size	<10,60> veľkosť kroku: 10	20,60,10,60

Tab.6.5 Dosiahnuté výsledky

Metóda	Štandardné nastavenie	Po ladení
Lasso	11345.65	-
SVM	8124.56	-
Random Forest	6046.413	5711.703
Gradient Boosting	5940.377	5461.248
MLP	6685.222	6163.577



Obr. 6.1 Dosiahnuté odchýlky po ladení parametrov

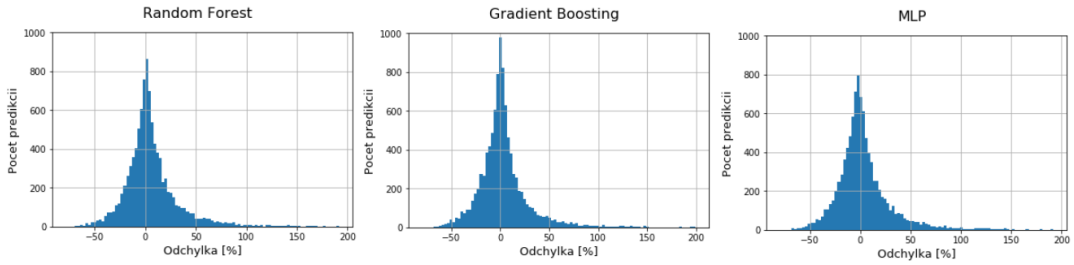
6.3 Učenie súborom metód

Posledným pokusom o zlepšenie výsledkov bolo spojenie vyladených metód strojového učenia s predpokladom, že ak niektorá z metód predikuje hodnoty väčšie ako je reálna hodnota platu a iná z metód zase menšie hodnoty, tak by sa mali tieto nepresnosti navzájom vyrušiť a dosiahnuť tak lepší výsledok. Pri predikcii

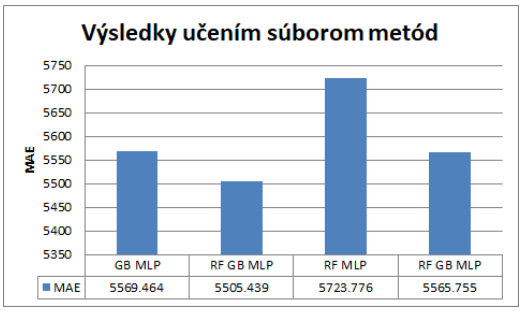
zohľadňujeme presnosť jednotlivých metód. Čím bola metóda i v našich predchádzajúcich testoch úspešnejšia, tým mala predikovaná hodnota h_i väčšiu váhu w_i pri rozhodovaní o finálnom výsledku. Váhy teda reprezentovali dosiahnuté odchýlky. Finálnu predikovanú hodnotu sme tak vypočítali podľa nasledujúceho vzorca:

$$Predikovaná\ hodnota = \sum_{i=1}^{n} \frac{h_i \frac{1}{w_i}}{\frac{1}{w_i}}$$

Metódu učenia súborom metód sme sa rozhodli vyskúšať na 3 najúspešnejších algoritmoch, a to konkrétne: Random Forest, Gradient Boosting a MLP. Veľkosti odchýlok jednotlivých metód je vidieť na obr. 6.2, odkiaľ môžeme usúdiť, že všetky použité metódy najčastejšie nadhodnocujú predikovanú hodnotu a ich spojením sme teda nedosiahli lepšie výsledky, ako to môžeme vidieť v grafe na obr. 6.3. Na tomto grafe je jasne vidieť, že priemerná hodnota akejkoľvek kombinácie týchto algoritmov mierne zhoršila výsledok najlepšej metódy.



Obr. 6.2 Odchýlky jednotlivých metód



Obr. 6.3 Výsledky kombinácií vyladených metód

6.4 Vyhodnotenie

Pri predikcii platov sme vyskúšali viaceré metódy strojového učenia. Po vyladení parametrov sme sa pokúsili o vylepšenie výsledkov pomocou učenia súborov metód, ktoré nebolo úspešne z dôvodu podobných rozložení predikovaných hodnôt jednotlivých metód. Najlepšie výsledky sme dosiahli s metódou Gradient Boosting,

ktorá mala po vyladení parametrov priemernú odchýlku 5461.248 zaraďujúca nás na 25. miesto v súťaži. Z výsledkov jasne plynie, že nami vyskúšané algoritmy nadhodnocovali predikované platy, preto vidíme možné zlepšenie našej predikcie v budúcom hľadaní dôvodov nadhodnocovania našich predikovaných hodnôt. Výsledný GitHub repozitár je možné nájsť na: https://github.com/bardzo12/salary_prediction.

7 Literatúra

1. Adzuna.: „Job Salary Prediction“, Kaggle, 2013, [ONLINE 3.4.2018]: https://www.kaggle.com/c/job-salary-prediction/data.
2. Salmon, M.: „Analysis of Web Scraped Job Data to Predict Relative Salaries“, 2017, [ONLINE 4.4.2018]: https://medium.com/@msalmon00/analysis-of-web-scraped-job-data-to-predict-relative-salaries-c7237954184a
3. Salmon, M.: „Web Scraping Job Postings from Indeed“, 2017, [ONLINE 4.4.2018]: https://medium.com/@msalmon00/web-scraping-job-postings-from-indeed-96bd588dcb4b
4. Scikit-learn.: „Machine learning in Python“, 2007, [ONLINE 3.4.2018]: http://scikit-learn.org/stable/documentation.html.
5. Khosla A.: „Job Salary Prediction“, [ONLINE 4.4.2018]: https://cseweb.ucsd.edu/~jmcauley/cse190/reports/sp15/012.pdf
6. Pandas: „Python Data Analysis Library“, [ONLINE 4.4.2018]: https://pandas.pydata.org/
7. Anonymous: „Job Salary Prediction“, [ONLINE 4.4.2018]: http://www.cs.ubc.ca/~nando/540-2013/projects/p58.pdf
8. Foxtrot: „Predicting advertised salaries“, [ONLINE 4.4.2018]: http://fastml.com/predicting-advertised-salaries/