Prognozy kursu dolara- sprawozdanie

Metody sztucznej inteligencji w inżynierii oprogramowania 24Z
Ćwiczenia 3,4,5
Filip Horst 311257

1 Wstęp

Badanie dotyczy prognozowania kursu dolara z horyzontem T+1 oraz T+7 z użyciem ręcznie konfigurowanych modeli KNN, MLP, drzewa wzmacniane gradientowo, lasy losowe, SVR oraz modeli zespołowych z nich utworzonych. Badanie odbywa się w środowisku Statistica oraz Excel.

2 Analiza statystyczna

Zmienna	F-value
USD t	22521,83
USD t-1	19659,61
USD t-2	17301,32
USD t-3	15165,30
USD t-4	13609,17
USD t-5	12448,19
USD t-6	11478,32
USD t-7	10891,60
CHF t	2540,37
CHF t-1	2538,56
CHF t-2	2520,99
CHF t-3	2504,45
CHF t-4	2476,28
CHF t-5	2454,65
CHF t-6	2447,34
CHF t-7	2437,34
ROPA t	1094,93
ROPA t-1	1090,34
ROPA t-2	1086,39
ROPA t-3	1081,67
ROPA t-4	1077,54
ROPA t-5	1073,36
ROPA t-6	1071,10
ROPA t-7	1068,17
SREBRO t	605,07
SREBRO t-1	604,34
SREBRO t-2	602,90
SREBRO t-3	600,41
SREBRO t-4	598,44
SREBRO t-5	596,62
SREBRO t-6	594,37
SREBRO t-7	593,17
EUR t	564,51
EUR t-1	561,15

EUR t-2	558,81
EUR t-3	556,37
EUR t-4	554,03
EUR t-5	552,37
EUR t-6	550,60
EUR t-7	548,77
ZŁOTO t	87,81
ZŁOTO t-1	87,79
ZŁOTO t-2	87,48
ZŁOTO t-3	86,94
ZŁOTO t-5	86,84
ZŁOTO t-6	86,71
ZŁOTO t-4	86,61
ZŁOTO t-7	85,99
YEN t	50,15
YEN t-1	49,64
YEN t-2	49,04
YEN t-3	48,66
YEN t-4	48,34
YEN t-5	47,96
YEN t-6	47,71
YEN t-7	47,61

Tabela z wartościami F-value dla każdej zmiennej objaśniającej

Prosta analiza wykorzystująca miarę statystyczną F pozwoliła wstępnie ustalić najważniejsze zmienne w zbiorze. Zgodnie z oczekiwaniami były to poprzednie kursy dolara. Jak się okazuje drugą najważniejszą grupą są dane o frankach szwajcarskim, a co ciekawe kolejnymi: ropa i srebro. Najgorsze okazały się być chińskie yen-y: dla nich korelacja z wyjściem T+1 i T+7 wyniosła odpowiednio w przybliżeniu 0.057 oraz 0.045. Trzeba jednak pamiętać, że nie jest to jednoznaczna informacja o ich bezużyteczności lub tym, że wręcz pogorszą wyniki – jest to tylko sygnał, że są potencjalnie mało użyteczne i trzeba na nie zwrócić uwagę.

3 Prognozy KNN

	błąd		miara	optymalna liczba	liczba podzbiorów	zakres min-maks	
	MAPE%	RMSE	odległości	najbliższych sąsiadów	w sprawdzianie krzyżowym	liczby sąsiadów	lista numerow wejsc
	HORYZONT t+1						
1	0,8616607	22,17414	euclidean	30	2	10, 30	5-60
2	0,8996451	22,9166	euclidean	34	2	20, 60 (increment 2)	5-60
3	0,8550037	22,30732	manhattan	30	2	10,50(4)	5-60
4	0,8679972	22,14679	chebyshev	24	2	20,40(2)	5-60
5	0,921938	23,34951	chebyshev	28	6	20,28	5-60
6	0,9502615	23,99488	euclidean (dist.weight)	45	2	25,45	5-60
7	0,538012	16,35342	chebyshev	2	2	wniosek z wielokrotnych b	5-60
HORYZONT t+7							
1	1,1953736	30,29146	euclidean	50	2	20, 60 (3)	5-60
2	1,1892019	30,20414	chebyshev	1	2	1,15	5-60
3	0,7810762	19,10899	manhattan	15	2	1,15	5-60
4	1,13174	28,26433	manhattan	40	2	10, 40 (2)	5-60
5	1,2500683	31,26025	manhattan	56	2	20,80 (4)	5-60
6	0,7369399	17,88297	manhattan (dist. Weight)	20	2	12,20	5-60
7	0,9526579	23,56394	manhattan (dist. Weight)	56	3	20, 60 (3)	5-60
8	0,6688498	16,18207	manhattan (dist. Weight)	15	2	14,15	5-60

W komórce zakres min-maks dla 7 modelu t+1 jest informacja, że zakres został wybadany na podstawie wielu eksperymentów.

Wartości w nawiasach to interwał przy przeszukiwaniu.

Wejścia 5-60 można utożsamiać z „wszystkie”.

3.1 T+1

W celu wyszukania ogólnych zależności pierwsze przeprowadzane testy polegały na przeszukaniu dużej przestrzeni K z dużym interwałem. Dla miar euklidesowych i Manhattan wyniki optymalna liczba sąsiadów okazała się bardzo podobna. Przy włączeniu ważenia wagami liczba sąsiadów dla miary euklidesowej znacznie wzrosła, a wyniki zaczęły się pogarszać.

Najlepszy okazał się model z miarą Czebyszewa i **tylko** 2 sąsiadami.

3.2 T+7

Pierwszym etapem analizy drugiego horyzontu było sprawdzenie, czy zachowanie modelu jest takie samo jak przy t+1. Zupełnie inne wyniki pierwszych trzech modeli (testy różnych miar w różnych zakresach) pokazały, że nie występuje taka sytuacja i należy dokładniej zbadać też inne możliwości.

Ostatecznie okazało się, że najskuteczniejsze są modele z metryką miejską i ważeniem odległościami. W tej konfiguracji najlepsze wyniki udało się osiągnąć dla K = 15.

Bardzo ciekawą obserwacją jest to, że błąd RMSE uzyskane dla t+7 jest **mniejszy** niż dla t+1.

Innym ważnym wnioskiem, który wpływa na badanie również innych modeli jest to, że zmiana horyzontu może znacznie zmienić parametry najlepszych modeli.

4 Prognozy SVR

	błąd				zakresy min-maks		optymalne				
	MAPE%	RMSE	SVM dla regr	SVM dla regr	stałe uczenia		stałe uczenia		typ		
	HORYZONT t+1		typ-1	typ-2	pojemność	epsilon	pojemność	epsilon	jądra	liczba iteracji	lista numerow wejsc
1	0,57292	14,7748	x		1,20	.05, .5 (.01)	18	0,05	rbf	2000	5-60
2	0,77248	18,773	x		10,30(2)	.001, .1 (.005)	28	0,011	rbf	2000	5-60
3	0,59782	15,0788	x		10,25(1,5)	.01, .5 (.01)	14,5	0,01	rbf	3000	5-60
4	1,52554	39,6835	x		1,2	.01,.5 (.01)	18	0,03	poly(3,0)	2000	5-60
5	0,64197	16,2509		x	1,30 (3)	.01, .5, (.02)	10	0,23	rbf	2000	5-60
6	0,64563	16,0874	x		15,25(2)	.001, .2 (.005)	19	0,036	rbf	2000	5-60
7											
HORYZONT t+7											
1	1,26782	30,4868	x		1,20	.01, .5 (.01)	11	0,07	rbf	2000	5-60
2	1,658	38,4946		x	1,20(2)	.01, .5 (.01)	15	0,34	rbf	2000	5-60
3	1,89202	46,7788	x		1,20(4)	.01, .5 (.05)	17	0,01	poly(3,0)	2000	5-60
4	1,75466	39,7205	x		10,18(2)	.01, .2 (.01)	12	0,2	rbf	3000	5-60
5											
6											
7											

4.1 T+1

W pierwszym horyzoncie najlepszy okazał się pierwszy utworzony model. Podjęto próby zmiany zakresów pojemności, czy epsilon, zmianę typu SVM, typu jądra a nawet liczby iteracji – nic nie dało lepszych wyników, choć model 3 uzyskał bardzo podobne wartości błędów.

Interesujące jest zestawienie modeli 1, 3, 5, 6 (może też 2), które mają stosunkowo podobne wyniki, ale ich optymalne (spośród przeszukanych) parametry znacznie się od siebie różnią. Zgadza się to z moimi poprzednimi doświadczeniami SVR, które można krótko podsumować tak, że SVR albo działa dla danego problemu albo nie i manipulacja parametrami daje podobne wyniki w wielu konfiguracjach.

Model SVR w tym eksperymencie pokazał również to jak trudno jest dostosowywać jego parametry. W przypadku KNN można łatwo sprawdzić wartość optymalną K i ją po prostu wybrać. W modelu SVR zarówno C, jak i eps mają duży wpływ na działanie modelu i same siebie (np. zmiana zakresu C sprawia, że inny zakres eps zaczyna działać lepiej). Sprawia to, że analiza konfiguracji bez użycia

GridSearch jest bardzo monotonna, a nawet z automatycznym przeszukaniem ciężko wysnuć jednoznaczne wnioski.

Ostatecznie najlepszym modelem jest model 1.

4.2 T+7

Niestety, dla t+7 uzyskano bardzo słabe wyniki, jednak podjęte próby modyfikacji różnych parametrów nie dawały prawie żadnej poprawy, co pozwala podejrzewać że model SVR może sobie po prostu nie radzić z tym problemem. Takie założenie należy traktować jednak jako ostateczność, ale jednocześnie nie warto upierać się przy jednym modelu, bo może się okazać że inny poradzi sobie z zadaniem lepiej i dużo szybciej.

Badany problem nie jest krytyczny (medycyna, wojsko), dlatego odpuszczenie bardzo trudnego w konfiguracji modelu może być bardziej opłacalne nawet jeśli utraci się potencjalnie parę procent skuteczności, które może udało by się wydobyć z SVR po czasochłonnych badaniach.

Ponadto, jeśli to badanie byłoby wykonane jako pierwsze można by też było podejrzewać że wyniki są gorsze bo ten horyzont jest o tyle trudniejszy do przewidzenia – to można wykluczyć, ponieważ akurat w tym przypadku model KNN był badany wcześniej i dał dobre wyniki.

W badaniach nie pomagał również fakt, że przeszukiwanie parametrów dla SVR zajmowało bardzo dużo czasu (w kontekście obliczeń komputerowych, nie tylko liczby kombinacji), szczególnie w porównaniu do innych modeli.

5 Prognozy las losowy

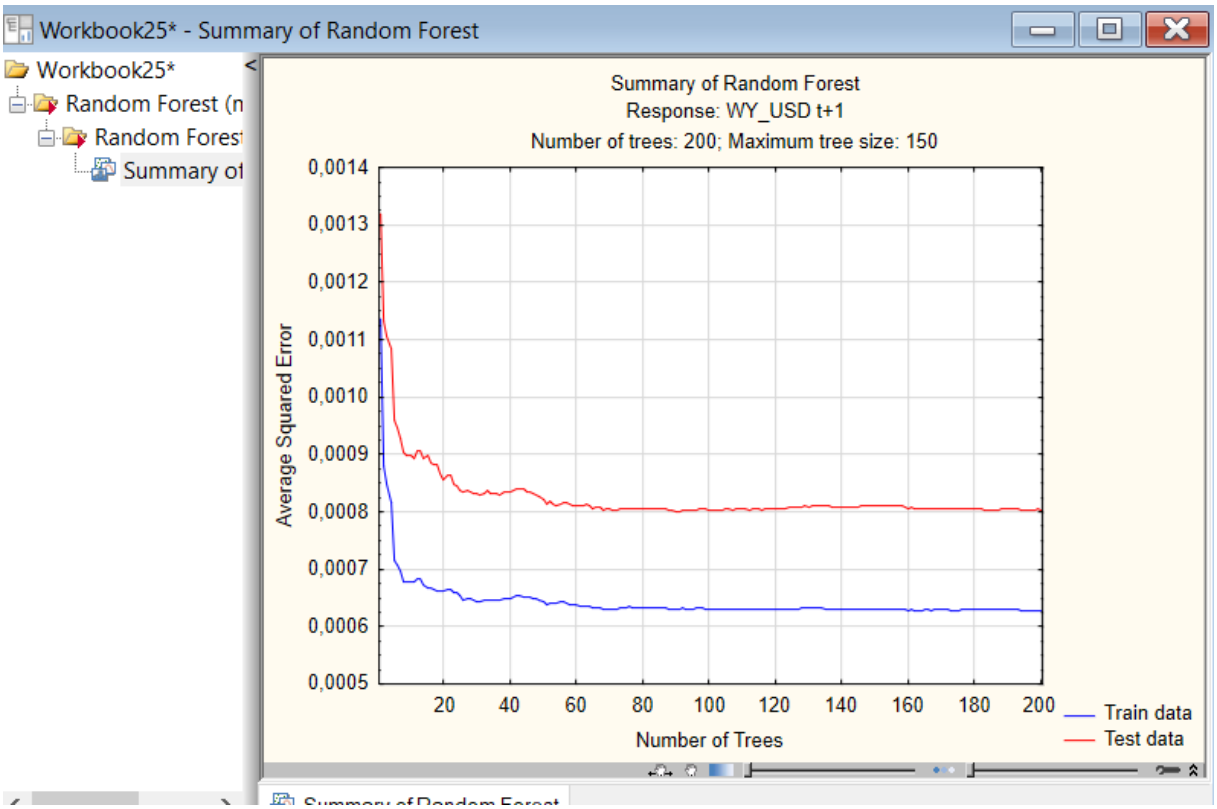
	błąd			parametry zatrzymania					
	MAPE%	RMSE	liczba drzew	min licznosc	maks liczba poziomow	min licznosc potomstwa	maks liczb wezlow	liczba predyktorów w pojedynczym drzewie	lista numerow wejsc (całkowita pula predyktorów do
	HORYZONT t+1								
1	0,56025	16,0009	200	91	10	5	150	40	5-60
2	0,55126	15,7751	150	91	10	5	150	55	5-60
3	0,56599	16,1152	150	100	15	5	150	55	5-60
4	0,52913	14,8975	150	70	7	5	150	55	5-60
5	0,48862	13,5192	25	50	6	5	150	55	5-60
6	0,55972	16,1176	25	50	6	30	80	55	5-60
7	0,49171	13,5681	110	50	6	5	150	55	5-60
	HORYZONT t+7								
1	1,16973	28,6868	200	91	10	5	150	55	5-60
2	1,13669	27,651	150	70	7	5	150	55	5-60
3	1,11817	27,3127	110	50	6	5	150	55	5-60
4	1,23028	30,0122	50	50	6	30	100	55	5-60
5	1,15328	28,3057	50	91	10	2	200	55	5-60
6	1,20816	29,1678	40 (early stop)	40	4	5	150	55	5-60
7	1,11162	27,2137	150	40	6	5	150	55	5-60

5.1 T+1

Las losowy zgodnie z przewidywaniami od razu zaczął dawać dobre wyniki. Eksperymenty dotyczące ustawień skupiły się głównie na liczbie drzew i ich rozmiarach. Ostatecznie okazało się, że mniejsza liczba drzew, które dodatkowo są płytsze pozwoliło uzyskać lepsze wyniki. Jest to ciekawa obserwacja, ponieważ podobnie jak w KNN okazało się, że te bardzo proste metody najlepiej generalizują.

Model 7 był sprawdzeniem, czy zwiększenie liczby drzew poprawi najlepszy dotychczas model 5, ale okazało się że dochodzi najprawdopodobniej do przeuczenia i nie jest to skuteczne.

Ważną obserwacją jest to, że Statistica tworzy tyle drzew ile jej się poda, co wskazuje na to że warunki stopu mogły być zbyt luźne. Nie zmieniałem ich jednak, ponieważ badania nie były na tyle obszerne żeby tego wymagać. Zamiast tego wykorzystany został wykres przebiegu uczenia przedstawiony poniżej:



Przykładowy wykres uczenia lasu losowego (model 7)

Z wykresu najważniejszym wnioskiem jest, to że błąd szybko maleje osiągając minimum lokalne w zakresie 20-30, następnie nieznacznie rośnie i zaczyna bardzo wolno spadać. Oznacza to, że model uczy się bardzo dobrze w krótkim czasie, a później wprowadza tylko małe poprawki, co można wykorzystać do szybszej analizy różnych konfiguracji. Można wykonać badania dla niewielu drzew, a później zbudować większy las dla jednego z nich (to właśnie miało miejsce w modelach 5 i 7).
W trakcie badań występowały również inne wykresy gdzie błąd przy 20-30 drzewach był wręcz **mniejszy** od dalszych osiągnięć. Oczywiście wnioski nie zostały sformułowane na bazie jednego załączonego wykresu, lecz dla wielu różnych konfiguracji.

5.2 T+7

Model dla drugiego horyzontu zaczął dawać znacznie słabsze wyniki. Podobnie jak w pierwszym przypadku udało się je trochę poprawić poprzez spłycanie drzew, ale wartości błędów wciąż były znacznie większe.

W jednym z przypadków doszło nawet do naruszenia warunków zatrzymania i budowie lasu z 40 drzew, gdy docelowa liczba było ustawiona na ponad 100.

Interesujące jest zestawienie modeli 2 oraz 7, które mimo bardzo odmiennych konfiguracji pod kątem rozmiaru drzew w zespole dały bardzo podobne wyniki. Biorąc pod uwagę tę obserwację, fakt, że pogłębianie drzew nie dawało lepszych rezultatów oraz podobieństwo błędów do SVR zaprzestano dalszych badań lasu losowego, ponieważ brakowało przesłanek że lepszy wynik jest osiągalny. Jest to jednak bardzo zaskakujące, ponieważ lasy losowe są bardzo skutecznym modelem, a znacząco przegrały w tym zadaniu z trywialnym KNN – jest to bardzo nietypowe i wskazuje na błąd lub niewystarczająco dogłębne badania, jednak jak już wspomniałem brakowało przesłanek o możliwości uzyskania lepszych rezultatów, a do badań zawsze można wrócić po analizie innych algorytmów.

6 Sieci MLP

	błąd		liczba neur	funkcja akt	funkcja akt	inicjalizacja	algorytm	liczba	
	MAPE%	RMSE	ukrytych	ukryta	wyjsciowa	wag	uczacy	epok	lista numerow wejsc
	HORYZONT t+1								
1	0,4639	12,5438	20	tanh	identity	normal	BFGS	59	4-59
2	0,46237	12,5568	25	tanh	identity	normal	BFGS	53	4-59
3	0,46365	12,5852	17	tanh	identity	normal	BFGS	52	4-59
4	0,47184	12,8174	15	identity	identity	uniform	CG	196	4-27 44-59 (bez yen I zloto)
5	0,45558	12,4174	15	identity	identity	normal	BFGS	46	4-59
6	0,45469	12,4081	15	identity	identity	normal	CG	325	4-59
7	0,45061	12,3532	15	identity	identity	uniform	CG	254	4-59
	HORYZONT t+7								
1	1,20578	29,0536	15	identity	identity	normal	BFGS	40	4-59
2	1,07355	26,0153	15	tanh	identity	normal	BFGS	130	4-59
3	1,20166	28,8995	15	tanh	identity	normal	CG	148	4-59
4	1,05903	25,4895	15	tanh	identity	uniform	BFGS	220	4-59

Przy sieciach MLP najlepsze (te zapisywane w excel) były wybierane wstępnie m.in. na podstawie wyników Test w Statistica. Zbiór Val wg. Statistica, czyli testowy – do porównywania modeli nie brał w ogóle udziału w decyzji! Nie zostały użyte przesiewy automatyczne, tylko badania ręczne. W ramach każdej konfiguracji uczone było kilka sieci i wybierana ta najlepsza, aby zminimalizować wpływ czynników losowych na wyniki.

6.1 T+1

W porównaniu do pierwszego domyślnego zestawu ustawień zaproponowanego przez prowadzącego udało się ustalić kilka elementów poprawiających wyniki. Pierwszym z nich była zamiana funkcji aktywacji w warstwie ukrytej z tanh na identity, drugim było zmniejszenie liczby neuronów ukrytych, a dwoma kolejnymi – najbardziej zaskakującymi – zmiana algorytmu uczenia na gradienty sprzężone (CG) oraz losowanie wag metodą uniform. W trakcie badań zweryfikowane zostały również scenariusze zwiększania rozmiaru warstwy ukrytej, innych funkcji aktywacji oraz uczenia spadkiem gradientu (zaobserwowano, że wymaga zwiększenia LR oraz Momentum). Podjęto również próbę usunięcia najmniej znaczących według analizy statystycznej (nie analizy wrażliwości!) cech, ale nie poprawiło to wyników, a wręcz je pogorszyło – co ciekawe w przeciwieństwie do oczekiwań po usunięciu części cech nie powstał wymóg zmiany rozmiaru sieci. Analiza wrażliwości wykazała, że tylko czasami występują cechy mające znaczenie <1, a nawet wtedy są ekstremalnie blisko wartości 1, dlatego nie usunięto na tej podstawie żadnej z nich.

Co do algorytmów uczenia, CG faktycznie pozwolił zwiększyć jakość sieci ale wydłużył czas uczenia. Nie tylko pod kątem liczby epok, ale również czasu ich obliczania. W przypadku tego zadania wszystko działo się na tyle szybko, że nie miało to znaczenia ale warto zapamiętać tę cechę.

Ostatecznie wyniki po wielu próbach (nie wszystkie zostały zapisane w tabeli, ponieważ brakło miejsca i nie dawały poprawy) udało się poprawić o zaledwie 1-2%. W tym zadaniu okazuje się więc, że wpływ hiperparametrów jest bardzo niski. Na szczęście okazało się, że wartość błędu na której utknęła sieć jest bardzo dobra – przynajmniej w porównaniu do innych modeli.

6.2 T+7

Dla horyzontu t+7 okazało się, że hiperparametry dają zupełnie inne wyniki. Tym razem najlepsza okazała się sieć z f. aktywacji tanh, algorytmem uczenia BFGS (aż 220 iteracji!) oraz losowaniem wag metodą uniform. Dla tego zakresu analiza wrażliwości podobnie jak przy t+1 nie pokazała wyraźnie przeszkadzających zmiennych – pokazała jednak, że wpływ najważniejszej zmiennej USD t jest znacznie niższy (ok. 45 w porównaniu do ok.188 dla t+1, spadek o ponad 75%), co pozwala się spodziewać znacznie gorszych wyników. Ostatecznie najlepsza sieć osiągnęła wynik RMSE 25,49, co stawia sieć na środku dotychczas zbadanej stawki modeli.

Dla tego horyzontu doszło do interesującego zjawiska – model znacznie dłużej się uczył, co widać po liczbie epok BFGS (wzrost z 60 do 200+), co teoretycznie mogłoby wskazywać na zbyt małe rozmiary sieci – ma za małą strukturę na douczenie się, więc potrzebuje więcej czasu. Okazało się jednak, że zwiększanie liczby neuronów nie poprawiała jakości sieci.

Inną obserwacją jest to, że po raz kolejny inicjalizacja uniform okazała się tą lepszą, co jest świetnym przykładem, że każdy nawet pozornie błahy hiperparametr może odgrywać zauważalną rolę.

7 Drzewa wzmacniane gradientowo

wariant	MAPE%	RMSE	wsp. uczenia	liczba drzew	min licznosc	maks liczba poziomow	min licznosc potomstwa	maks liczb wezlow	lista numerow wejsc (całkowita pula predyktorów do wykorzystania
	HORYZONT t+1								
1	0,54397	14,5966	0,1	279	91	6	1	10	4-59
2	0,56087	14,6993	0,2	190	91	4	1	10	4-59
3	0,53466	14,3676	0,1	257	91	10	1	15	4-59
4	0,51936	14,0408	0,1	283	80	10	1	15	4-59
5	0,52368	13,8663	0,1	290	70	10	1	15	4-59
6	0,51978	13,942	0,1	270	80	8	1	12	4-59
7	0,51368	13,6749	0,1	260	80	5	1	15	4-59
	HORYZONT t+7								
1	0,77314	19,4385	0,1	2396	91	6	1	10	4-59
2	0,76715	19,306	0,2	2764	91	10	1	15	4-59
3	0,93068	23,091	0,4	~1960	70	10	1	15	4-59
4	0,7672	19,7199	0,15	2434	80	12	1	15	4-59
5	0,76661	19,0378	0,1	2482	130	5	1	10	4-59
6	0,75113	19,6154	0,1	2499	150	10	1	15	4-59
7	0,71564	18,1094	0,13	4997	130	4	1	10	4-59

Wejścia 4-59 to „wszystkie”.

7.1 T+1

Sprawdzone zostały dwa kierunki zmian – spłycanie oraz pogłębianie drzew. Lepsze w tym problemie okazały się modele z płytkimi drzewami jak chodzi o liczbę poziomów, ale z dużą dopuszczoną liczbą węzłów. Dodatkowo pomogło zmniejszenie minimalnej liczności. Ciekawym zjawiskiem było to, że learning rate równy 0.1 był praktycznie idealny. Nawet takie zmiany jak 0.11, czy 0.9 sprawiały, że model zaczynał dawać gorsze wyniki. Oczywiście w pierwszym przypadku był to prawdopodobnie skutek zbyt dużych „skoków” po funkcji błędu (podobnie do uczenia MLP mimo różnic w obu metodach), ponieważ jeśli problemem byłby overfitting to statistica po prostu zatrzymała by model na mniejszej liczbie drzew. Dla 0.9 przeciwnie problemem było niedouczenie modelu.

7.2 T+7

Dla drugiego horyzontu ponownie zbadano zachowanie modelu po pogłębieniu i spłyceniu drzew. Początkowe wyniki nawet przy ponad 2000 drzew wciąż nie były „nasycone”. Taka obserwacja wskazuje na niedouczenie drzew, więc naturalnie sprawdzono co się stanie, kiedy zwiększony zostanie parametr uczenia oraz złożoność drzew. Faktycznie pomogło to trochę z ich liczbą, ale znacznie pogorszyło wyniki. Po dalszych badaniach okazało się, że największą poprawę uzyskano po spłyceniu drzew i lekkim podniesieniu learning rate. Tym, co doprowadziło do poprawy była zmiana minimalnej liczności na większą. Po znalezieniu tej zależności ponownie sprawdzone zostało zachowanie modelu dla głębszych drzew i ponownie uzyskano gorsze wyniki. Ostatecznie najlepszy model to skutek głównie spłycenia drzew, zwiększenia minimalnej liczności i budowy ich ogromnej liczby. Wynik RMSE ~18.1, choć uzyskany w dziwny sposób to wciąż wartość, którą można uznać za dobrą ponieważ jest lepsza od wszystkich zbadanych modeli oprócz, co ciekawe, KNN. W przeciwieństwie do innych, temu modelowi udało się jednak chociaż zbliżyć do tego prymitywnego, ale jak widać skutecznego, algorytmu.

8 Modele zespołowe

8.1 T+1

Model	RMSE
MLP	12,3531597
RF	13,568
GB	13,675
SVR	15,07878169
KNN	16,35341713

Tabela porównująca najlepsze modele dla każdego algorytmu (T+1)

	Horyzont t+1				Lista modeli		
Nr zespołu	Typ wag	Wagi	MAPE%	RMSE	Numer	Nazwa	RMSE samotnie
1	Równe	1, 1	0,48327	13,24793924	1	GB7	13,67491292
	Biliniowe	0,49714; 0,50286	0,48317	13,24749347	2	RF5	13,51923245
					3		
					4		
					5		
2	Równe	1, 1, 1	0,45992	12,68660386	1	GB7	13,67491292
	Biliniowe	0,32067; 0,32436; 0,35498	0,45876	12,65851313	2	RF5	13,51923245
	Ręczne	1, 1, 12	0,44844	12,33121081	3	MLP7	12,3531597
					4		
					5		

*- przy ważeniu suma była dzielona przez sumę wag, dlatego to że ich suma przekracza 1 to nie problem

Dla T+1 MLP uzyskało wyraźnie najlepsze wyniki, dlatego zbadano dwa warianty modeli zespołowych: zespół 1, czyli para dwóch najlepszych modeli poza MLP oraz zespół 2, czyli poprzednia para w połączeniu z wyraźnie lepszym MLP.

W pierwszym zespole użycie połączenia zespołowego pozwoliło uzyskać wyraźnie lepszy wynik co oznacza, że modele całkiem dobrze się dopełniają. Co ciekawe wyniki RMSE były na tyle podobne dla pojedynczych modeli, że zastosowanie wag biliniowych praktycznie nic nie dało.

W drugim zespole zgodnie z oczekiwaniami doszło do wyraźnej poprawy dzięki dodaniu najlepszego modelu MLP. Zgodnie z oczekiwaniami również zastosowanie wag biliniowych poprawiło jakość zespołu, ponieważ skupiło zespół na predykcjach tego najlepszego. To co jest jednak ważne to fakt, że żaden zespół nie pokonał pojedynczego modelu MLP – sieć neuronowa świetnie sobie radzi w tym problemie.

Ze względu na własną ciekawość postanowiłem sprawdzić, co by było gdyby wagi wykorzystać jako parametr i modyfikować go ręcznie. Jest to oczywiście zupełnie niepoprawne, ponieważ uwzględnia dopasowywanie parametrów na zbiorze testowym, czego się absolutnie nie powinno robić i dlatego te wyniki nie zostały dalej wykorzystane. Udało się jednak uzyskać jeszcze lepsze wyniki, co pokazuje że różnorodność modeli jest dużą zaletą i dlatego modele zespołowe są tak skuteczne – po prostu ze względu na dużą dysproporcje osobistych rekordów w tym problemie zespoły sobie nie poradziły. Innym pytaniem jest: co by się stało gdyby wagi dodać na danych uczących? Lub jeszcze lepiej na walidacyjnych?

8.2 T+7

Model	RMSE
KNN	16,18207376
GB	18,1094034
MLP	25,48945357

RF	27,21372218
SVR	30,48675194

Tabela porównująca najlepsze modele dla każdego algorytmu (T+7)

	Horyzont t+7				Lista modeli		
Nr zespołu	Typ wag	Wagi	MAPE%	RMSE	Numer	Nazwa	RMSE samotnie
1	Równe	1, 1	0,61898	15,21609102	1	GB7	18,1094034
	Biliniowe	0,4719; 0581	0,61709	15,16154098	2	KNN8	16,18207376
	Ręczne	1; 1,65	0,61427	15,07437697	3		
					4		
					5		
2	Równe	1, 1, 1	1,0279	24,76606761	1	RF7	27,21372218
	Biliniowe	0,33781; 0,30153; 0,36066	1,01742	24,54705069	2	SVR1	30,48675194
	Ręczne	2, 1, 4	0,98106	23,70794993	3	MLP4	25,48945357
					4		
					5		

W problemie T+7 ponownie pojawił się problem rozbieżności wyników paru najlepszych modeli. Tym razem okazało się jednak, że para najlepszych wyników bardzo dużo zyskała na zestawieniu. Hipotezą tego jest to, że modele dobrze się dopełniają – tam gdzie jeden popełnia błąd, tam drugi radzi sobie świetnie. Jest to bardzo nieprzewidywalne zjawisko, więc można powiedzieć, że taki wynik jest po prostu szczęśliwy. To, że jest nieprzewidywalne nie oznacza jednak że nie można do niego dążyć – dobrą opcją wydaje się stosowanie różnorodnych modeli, ponieważ to zwiększa szanse że skupią się na innych cechach procesu. Zespół 1 jest pewnym potwierdzeniem (choć nie jest to dowód!) takiej hipotezy, ponieważ nie da się chyba wybrać bardziej różnorodnych podejść niż skomplikowane drzewa wzmacniane gradientowo i trywialny KNN.

Drugi zespół miał na celu zweryfikować, co by się stało gdyby nie udało się zbudować dwóch najlepszych modeli, których wyniki bardzo znacząco odstawały od reszty. Tym razem dla trzech modeli udało się osiągnąć tylko trochę lepsze wyniki niż najlepszego z nich.

W obu zespołach skuteczne okazało się być podejście biliniowe do ustalania wag. W utworzonych zespołach mogło to być takie skuteczne, ponieważ wyniki nie były od siebie bardzo odległe, dzięki czemu ważenie biliniowe tylko trochę nakierowywało prognozy na ten lepszy model i nie miało miejsce silne upodabnianie zespołu do najlepszego członka. Oferowany był całkiem dobry kompromis.

W tym badaniu ponownie sprawdziłem, co dałyby wagi ustalane ręcznie. Ponownie uzyskane wyniki były by znacznie lepsze. Dalsze badanie nie zostały podjęte ze względu na obszerność projektu. W celu ich wykonania należałoby skorzystać ze zbioru, który był użyty do MLP, wyciąć dane testowe, wgrać pozostałe tak, aby dane walidacyjne były interpretowane jako testowe przez modele np. GB i RF i wykonać predykcje z odpowiednimi parametrami, a następnie wykorzystać wyniki walidacyjne do dobrania wag i sprawdzenie zachowania na danych testowych. Z drugiej strony, dane nazwane w tym arkuszu testowymi były również używane do wyboru najlepszych modeli (poza MLP gdzie był zbiór walidacyjny), więc trzeba by było poznać kontekst podziału danych na zbiory i rozważyć, czy użycie ich do dobrania w taki sposób wag jest poprawne. Ponadto, taki manewr byłby ryzykowny, ponieważ mógłby skutkować nadmiernym dopasowaniem do danych. Ostatecznie bezpieczniej jest zastosować wagi biliniowe dające dobry kompromis.

9 Analiza biznesowa

Analiza biznesowa odbyła przy następujących założeniach:

1. Sprawdzamy prognozę na T+1
2. Jeśli prognozowany jest wzrost to kupujemy USD za 100 PLN

3. Po upływie jednego dnia sprzedajemy wszystkie zakupione USD

Dla T+7 to samo, tylko sprzedaż po tygodniu.

Przykład działania utworzonej tabeli:

	A	B	C	D	E	F	G	H	I	J	K	L
1			Wartości rzeczywiste			Prognozy					Konto	Zysk
2	data	Kod	USD t	USD t+1	USD t+7	prognoza t+1	prognoza t+7		Prognoza T+1	Akcja		27,9591
3	2010-01-08	3	2,8752	2,8384	2,7992	2,874926	2,82536189		2,874926	0	-1	0
4	2010-01-11	3	2,8214	2,7953	2,8152	2,824258	2,83481722		2,824258	1	99,07493	-0,9251
5	2010-01-16	3	2,8115	2,8152	2,89	2,815490	2,85981713		2,815490	1	100,1316	0,1316
6	2010-01-18	3	2,8152	2,7929	2,896	2,818920	2,87077724		2,818920	1	99,20787	-0,7921
7	2010-01-20	3	2,8112	2,8759	2,8932	2,827316	2,8789142		2,827316	1	102,3015	2,30151
8	2010-01-25	3	2,896	2,879	2,9215	2,891092	2,9152131		2,891092	0	-1	0
9	2010-01-29	3	2,9186	2,9136	2,963	2,923110	2,94534742		2,923110	1	99,82868	-0,1713
10	2010-02-05	3	2,963	3,0151	2,9448	2,979938	2,96781893		2,979938	1	101,7584	1,75835
11	2010-02-15	3	2,953	2,9603	2,9589	2,951952	2,96291683		2,951952	0	-1	0
12	2010-02-23	3	2,9157	2,9442	2,907	2,924174	2,86380112		2,924174	1	100,9775	0,97747
13	2010-02-25	3	2,9319	2,9707	2,8454	2,941542	2,83992792		2,941542	1	101,3234	1,32337
14	2010-02-27	3	2,8893	2,908	2,848	2,894454	2,86433164		2,894454	1	100,6472	0,64722
15	2010-03-07	3	2,8501	2,8531	2,8219	2,858267	2,85218958		2,858267	1	100,1053	0,10526
16	2010-03-08	3	2,8531	2,8437	2,8242	2,860901	2,86802421		2,860901	1	99,67053	-0,3295

Na przykład dla wiersza 5: prognoza T+1 wynosząca 2,815 jest wyższa od dzisiejszego kursu 2,8115, więc kupujemy USD za 100 PLN i kolejnego dnia sprzedajemy. Tego dnia t+1 kurs rzeczywisty wyniósł 2,8152 – więcej niż w dniu zakupu więc sprzedajemy za 100,1316 z zyskiem 0,1316.

Oczywiście taka analiza jest lekko naiwna głównie ze względu na to, że kurs zakupu i sprzedaży są czasem zupełnie inne i skutecznie utrudniają takie transakcje. Pokazuje ona jednak potencjalny kontekst wykorzystywanych prognoz, co nie raz mówi dużo więcej niż wszelkie miary błędów matematycznych.

Badanie można by było zmodyfikować poprzez dodanie mechanizmu inwestowania kwoty zależnej od wzrostu kursu – np. jeśli wzrost o 0.01% to tylko 10 PLN, a jeśli wzrost o 0.1% to 100 PLN.

9.1 T+1

Metoda	Zysk PLN
Optimum	122,1869
MLP	27,95914

Zysk w zależności od użytej prognozy T+7

Prognoza optimum to po prostu poprawna wartość. Użyty najlepszy model, czyli MLP dał niestety znacznie mniejszy zysk, bo zaledwie ok. 23% potencjału. Jest to jednak zysk, więc można uznać model za skuteczny.

9.2 T+7

Metoda	Zysk PLN
Optimum	354,7304373
GB+KNN	311,7822592

Zysk w zależności od użytej prognozy T+7

Do badania użyto modelu zespołowego GB + KNN i wyniki okazały się być świetne. Potencjał został zrealizowany w aż prawie 88%. Jest to jednak bardzo ciekawe, ponieważ według matematyki błąd RMSE w horyzoncie T+7 był większy. Te różnice są prawdopodobnie spowodowane tym, że w praktycznym problemie błąd może być pozytywny – tego użyte miary nie brały pod uwagę. Poprzez to rozumiem, że błąd prognozy polegającym na niedoszacowaniu wzrostu jest niczym złym, podczas gdy

błąd polegający na niedoszacowaniu spadku jest czymś złym. Widocznie na zadanych danych model t+7 częściej się mylił, ale pomyłki te częściej były mniej kosztowne jak chodzi o inwestycje.

Inną hipotezą lepszego zysku jest stabilny wzrost wartości dolara. Wartość tej waluty mimo codziennych drgań w okresie zawartym w zbiorze posiadała ogólny trend, który był wzrostem. Użycie prognoz tygodniowych sprawia, że cała transakcja jest bardziej odporna na losowe wahania i jest przybliżona do długoterminowego zysku, który jest wyznaczany przez uśredniony trend. Oczywiście wadą takich prognoz jest to, że im dalej w czasie jest odsunięte wydarzenie tym ciężej je prognozować (na ogół, a przynajmniej dla tego zadania – są wyjątki jak zawsze).

Na podstawie tych obserwacji można powiedzieć, że wykonując takie badanie po raz kolejny na pewno warto rozważyć personalizowaną funkcję błędu, która będzie karać dodatkowo za pomyłki będące przewidzeniem wzrostu, gdy w rzeczywistości jest spadek, natomiast będzie mniej karać za pomyłki będące przewidzeniem spadku, gdy wystąpi wzrost. Być może takie podejście pozwoliłoby utworzyć model dający prognozy, które po wykreśleniu nie znajdowały by się przy prostej wartości rzeczywistych po obu stronach tylko lepiej – zawsze w pobliżu, ale nigdy ponad. Można również powiedzieć, że taki model i aktualny to zupełnie dwa różne podejścia do inwestycji. Aktualny jest bardziej ryzykowny, ale potencjalnie da większe zyski, natomiast teoretyczny model, który opisuję jest bardziej bezpieczny – sugeruje wpłatę tylko gdy jest pewien, co sprawia że inwestycje będą statystycznie częściej udane, ale dużo rzadsze, co może doprowadzić do bardzo powolnego zysku. Oczywiście ważny jest kompromis i decyzja użytkownika podjęta w oparciu o własne zasoby.

10 Wnioski ogólne

Z badania jako całości można wywnioskować, że prognozowanie jest bardzo czasochłonne i wymaga dużego doświadczenia lub zasobów, żeby je przyśpieszyć. Doświadczony pracownik wie jak zmieniać parametry, natomiast przy mojej pracy na niektórych modelach – głównie mniej znanym mi modelu drzew wzmacnianych gradientowo musiałem stosować podejście zachłanne, które jak wiadomo nie zawsze jest optymalne.

Niewątpliwie podejście pracy zespołowej było by tutaj skuteczniejsze. Rozpoczęcie prac od „strzału” eksperta opartego o doświadczenie, a następnie poszukiwanie w różnych kierunkach przez kilku pracowników – wtedy jeden mógłby zająć się badaniem konfiguracji o większej złożoności (większe drzewa, więcej neuronów), drugi przeciwnie mniejszej złożoności, a kolejni np. mniej czy bardziej losową eksploracją.

Analiza statystyczna cech była w tym projekcie zrobiona trochę nie po kolei – powinna być pierwszym etapem prac. Na szczęście okazało się, że co prawda są zmienne słabe tj. Yen, ale nie ma zmiennych przeszkadzających.

Ostatecznie zadanie można uznać za rozwiązane, ponieważ udało się zyskać w przypadku obu prognoz, co pokazuje analiza biznesowa.